Financial Investment Behaviour between Hong Kong and Mainland Chinese Investors and Predicting Investors' Preferences

By

Mark Kwong Yiu MAK

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

University of Warwick, WMG

September, 2018

Abstract

Behavioural finance has been popular in the literature pertaining to investment behaviour in recent years. However, the number of applications for exploring individual investment behaviour in Hong Kong and mainland China are limited. This research investigates investor behaviour to identify the major determining factors influencing investor behaviour between Hong Kong and mainland China, and derive customers' investment preferences from the behavioural patterns identified. The approach developed generates two new models, namely PSYC Model and Financial Data Mining Model (FDMM).

Data from 142,496 mainland Chinese and Hong Kong investors of Convoy Financial Service Limited ("Convoy"), one of the largest financial service providers in Hong Kong, were used to identify major influencing factors and examine differences of mainland Chinese and Hong Kong investors. Statistical analyses, including descriptive analysis, factor analysis, correlation analysis and regression analysis, were adopted. Six major factors were statistically supported to have impact on investment behaviours. More importantly, investment decisions made by investors from mainland China and Hong Kong are mainly affected by (i) age (demographic factor) (ii) investment experience (psychological factor) and (iii) annual income (sociological factor).

The newly developed PSYC Model on the basis of the statistical results generalized two major perspectives of investor behaviours, namely investing involvement and risk appetite, in response to the three factors of Hong Kong and mainland China investors. Following this, the FDMM was created to assist financial institutions in predicting customer behaviour. Clustering analysis and association rules were applied. Ten experts with at least 15 years of experience in banking, insurance and stock markets in Hong Kong and mainland China were consulted in form of group meetings, telephone interviews and teleconference to justify the two new models with their experience and practical knowledge in the industry.

The PSYC Model and FDMM were implemented within Convoy for validation. The former model helped in design and identification of appropriate product and marketing strategies for different classes of products while the latter model generated eight specific rules for Convoy to implement product offerings to Hong Kong investors which greatly improved business workflow and efficiency. As a result, the success rate selling financial products to its customers and the customer satisfaction level were boosted by 66% and 8.7% respectively.

With these promising results, the research objectives were met, and the outcomes of this research can assist financial institutions to gain better insights into the behaviour of their customers from Hong Kong and mainland Chinese, and offer the most suitable financial products for fulfilment. The model developed could be further generalized for adoption by other financial institutions after further evaluation and case studies with the support of a wider range of customer profiles in diversified financial institutions.

Acknowledgments

I would like to thank my supervisor, Professor Jay Bal, for his professional and patient guidance in my project in the past years. With his advance and insightful feedback on my paper, my knowledge has been broadened tremendously and my paper has been enriched to a huge extent. Such advance in my knowledge would help me have a better and wonderful life and definitely develop a greater career and business.

Special thanks to Dr. W.H. Ip, for his guidance and support throughout my project. With his excellent knowledge in the areas of modelling, customer relationship management and big data analytics, he provided me useful knowledge and inspired me in the areas of investigating and modelling the investors' behaviours.

Finally, I would like to take this opportunity to thank my ex-employer, Convoy Financial Services Limited, which has fully supported me to finish this paper. Convoy had provided me a great platform to learn and develop my career and business for 14 years. I am proud of Convoy's success to develop from 10 persons to the largest IFA in Hong Kong with thousands of professionals in Hong Kong as well as China.

Declaration

I declare that, except where acknowledged, the material contained in this thesis is my own work and that is has neither been previously published nor submitted elsewhere for the purpose of obtaining an academic degree.

MAK Kwong Yiu, Mark

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Association Rule	AR
Anti-Money Laundering	AML
Convoy Financial Services Holdings Limited	CFS
Clustering Module	CM
Customer Relationship Management	CRM
Data Mining	DM
Data Selection and Pre-processing Module	DSPM
Data Selection Process	DSP
Financial Data Mining Model	FDMM
Fuzzy Logic	FL
Hong Kong Dollar	HKD
Hong Kong Monetary Authority	HKMA
Kaiser-Meyer-Olkin	KMO
Know Your Customers	KYC
Korea Composite Stock Price Index	KOSPI
Mandatory provident Fund	MPF
Neural Network	NN
Rules Discovery Module	RDM
Severe Acute Respiratory Syndrome	SARS
Self-directed and Comfort-driven	SC
Securities and Futures Commission	SFC
Self-directed and Yield-driven	SY

Professional-directed and Comfort-driven	PC
Professional-directed, Self-directed, Yield-driven and Comfort-driven	PSYC
Professional-directed and Yield-driven	PY

Publication Arising from the Thesis

Refereed Journal Papers

Mak, M.K.Y., Ho, G.T.S. and Ting, S.L. (2011). A financial data mining model for extracting customer behaviour. *International Journal of Engineering Business Management*, 3(3), 59-72.

Mak, M.K.Y. and Ip, W.H. (2017). An exploratory study of investment behaviour of investors. *International Journal of Engineering Business Management*, 9, 1-12.

Chapter 1 Introduction

This is an investigation into behavioural economics to help explore investment behaviour with a focus on the differences of mainland Chinese and Hong Kong investors. This chapter introduces the domain of the research, identifies the research gap and the resulting research question with the associated research objectives. The structure of the thesis to investigate the research question is described in detail.

1.1 Background of study and research gap

Traditional theories assumed that investment behaviours are rational (Rizvi and Fatima, 2014). However, defining events, such as the collapse of several renowned hedge funds in 1998 and the financial tsunami between 2007 and 2008, have caused a rethink in the domain, directing attention to apparently irrational human behaviour. In an attempt to explain this irrational behaviour, the field of behavioural finance has gained popularity. Under the theory of behavioural finance, researchers suggest that in real life, investment behaviours of individual investors are influenced by a combination of specific psychological factors, such as:

i) Overconfidence (Jain *et al.* 2015; Tekce *et al.,* 2016) that refers to investors who over-estimate the probabilities of success and are over-confident in their own abilities and the accuracy of their judgments.

ii) Representativeness (Jain *et al.* 2015) means the tendency that investors make subjective judgments based on the similarity of events and outcomes.

iii) Herding behaviour (Shweta, 2014; Spyrou, 2013; Chang and Lin, 2015; Fred van Raaij, 2016) is defined as "irrational behaviour which illustrates the psychology of investors imitating others' investment decisions and over-relying on public opinion without consideration of their own information" (Zhang and Zheng, 2015, p.2).

The underlying assumption is that emotions and psychological factors have a strong influence on investment decisions (Loewenstein *et al.,* 2001; Simon, 1987; Sahi *et al.,* 2013; Baumeister *et al.,* 2007). However, research on psychological investment behaviour ignores sociological factors and personality traits, while it seems that investment behaviour is a more complex domain combining both rational and emotional elements. According to research in investment behaviour, behavioural finance is not purely based on psychological but also sociological factors (Zhang and Zheng, 2015). In addition, Fung and Durand (2014) suggested that, demographic factors such as age and gender are also important in explaining investors' behaviour. Including psychological, sociological and demographic factors into account, behavioural finance appears to provide the most comprehensive framework to explain investment behaviour (Zhang and Zheng, 2015; Iqbal, 2011; Jureviciene and Jermakova, 2012). However, from the literature reviewed in chapter 2.2 and supported by the Finance Industry expert panels conducted in this study (see chapter 3.2.3), there is still a knowledge gap regarding the major contributing factors to investment behaviour within the framework of behavioural finance. This is also consistent to the discussion results among the experts consulted (the details of the expert panel conducted in this study are in chapter 3.2.3) who claimed that investment behaviour could be explained by demographic, psychological and sociological factors nowadays. Thus, this research addresses this gap by identifying the major attributes to explain investment behaviour by leveraging psychological, sociological and demographic factors.

As an international financial centre, Hong Kong provides a variety of financial products, such as venture capital, private equity, mutual funds, stocks, and bonds for people to invest globally due to its proximity to mainland China, similarity in language and culture, low tax rates and global access. Hong Kong remains the top offshore investment destination for mainland Chinese investors, and in recent years, wealthy mainland investors have increased their holdings in stocks and other financial products sold in Hong Kong, from 19% in 2011, 33% in 2013 to 37% in 2015 (Sun, 2015). Hong Kong-domiciled fund products are more popular on the mainland than in Hong Kong, and mainlanders have been enthusiastic buyers of Hong Kong fund products. According to sales figures released by the China State Administration of Foreign Exchange, the value of Hong Kong-domiciled funds sold in the mainland was almost 70 times that sold in Hong Kong (Yu and Yiu, 2016). These facts encourage financial institutions in Hong Kong to review their marketing strategy for targeting this fast-growing market segment, i.e. Mainland investors investing in the offshore Hong Kong market. As discussed by the expert panel conducted for this study (see chapter 3.2.3), neither expert from Hong Kong nor China expressed that they have read or been aware any literature, articles and publications in relation to the comparison between Hong Kong and Chinese investors' behaviours.

Therefore, our research question is:
RQ: What are the key differences in the factors that affect financial behaviour between mainland Chinese and Hong Kong investors?

The outcomes of this research will greatly assist financial institutions to gain better insights into the behaviour of their customers. It is of paramount importance for financial institutions to recognize and predict investors' preferences towards various financial products. In a highly competitive environment, financial institutions need to be able to identify their customers' needs, create effective investment portfolios and offer the most suitable financial products for each individual.

The review of the literature showed that researchers focus mainly on what factors influence investors' behaviour and/or their impact in investment decision (Masini and Menichetti, 2012; Phan and Zhou 2014; Charles and Kasilingam, 2013; Sultana and Pardhasaradhi, 2012) as well as how investors act in the financial market (Cashman *et al.,* 2012; Hoffmann *et al.,* 2013). Studies rarely investigated how to predict investors' preferences based on the factors influencing their behaviours. This gap is probably due to researchers lacking access to the huge volumes of financial transaction data required to draw such conclusions from studying real-life behaviour. Access to this strictly confidential data held by financial institutions is the biggest obstacle to researchers attempting to conduct this kind of research.

Other major challenges for financial institutions include:

i) the right methodology to identify hidden relationships influencing customers' choice of financial products and

ii) the difficulty to efficiently process massive amount of historical data related to customers and financial products (Zhang and Zhou, 2004).

In failing to overcome these challenges, financial institutions may miss opportunities to attract and satisfy new investors. In order to unveil new business opportunities and enhance their competitiveness, discovery of hidden relationships, unexpected patterns and useful rules from the large databases of customer transactions may provide a robust solution. With the massively growing volumes of data in the business environment in recent years, knowledge discovery has become a new and effective tool (Vazirgiannis *et al.,* 2003; Hesham *et al.,* 2015), while data mining has been widely applied to handle massive amounts of data for uncovering business opportunities (Ravisankar, *et al.,* 2011). After conducting an extensive literature review of the domain into findings from the 176,000 articles published since 2010, it can be concluded that while a wide range of data mining models exists, none that would facilitate financial institutions to efficiently predict customers' investment preferences appear to exist.

In this research a new Financial Data Mining Model (FDMM) is proposed to help provide financial institutions with greater insights into customers' needs and offer them the right products to build their investment portfolio. This development from this research addresses research gap derived from the literature review in chapter 2.4.

1.2 Research objectives

In order to address the research question elaborated above, this study has the following objectives:

i) Identifying the major attributes that have been identified through previous work;

ii) Identifying the attributes that define mainland Chinese and Hong Kong investors;

iii) Weighing the attributes identified in the research context; and

iv) Studying how financial institutions predict investment preferences.

The results of this study provide an alternative model to explain customer preferences for financial products. In addition, the findings will help give financial institutions insights into strategic and marketing planning to enhance both the degree of customisation and suitability of financial investment products. These are essential aspects in efficiently and systematically meeting changing financial investment behaviours and preferences, especially for mainland and Hong Kong investors.

1.3 Structure of the thesis

Figure 1.1 shows the structure of this thesis.

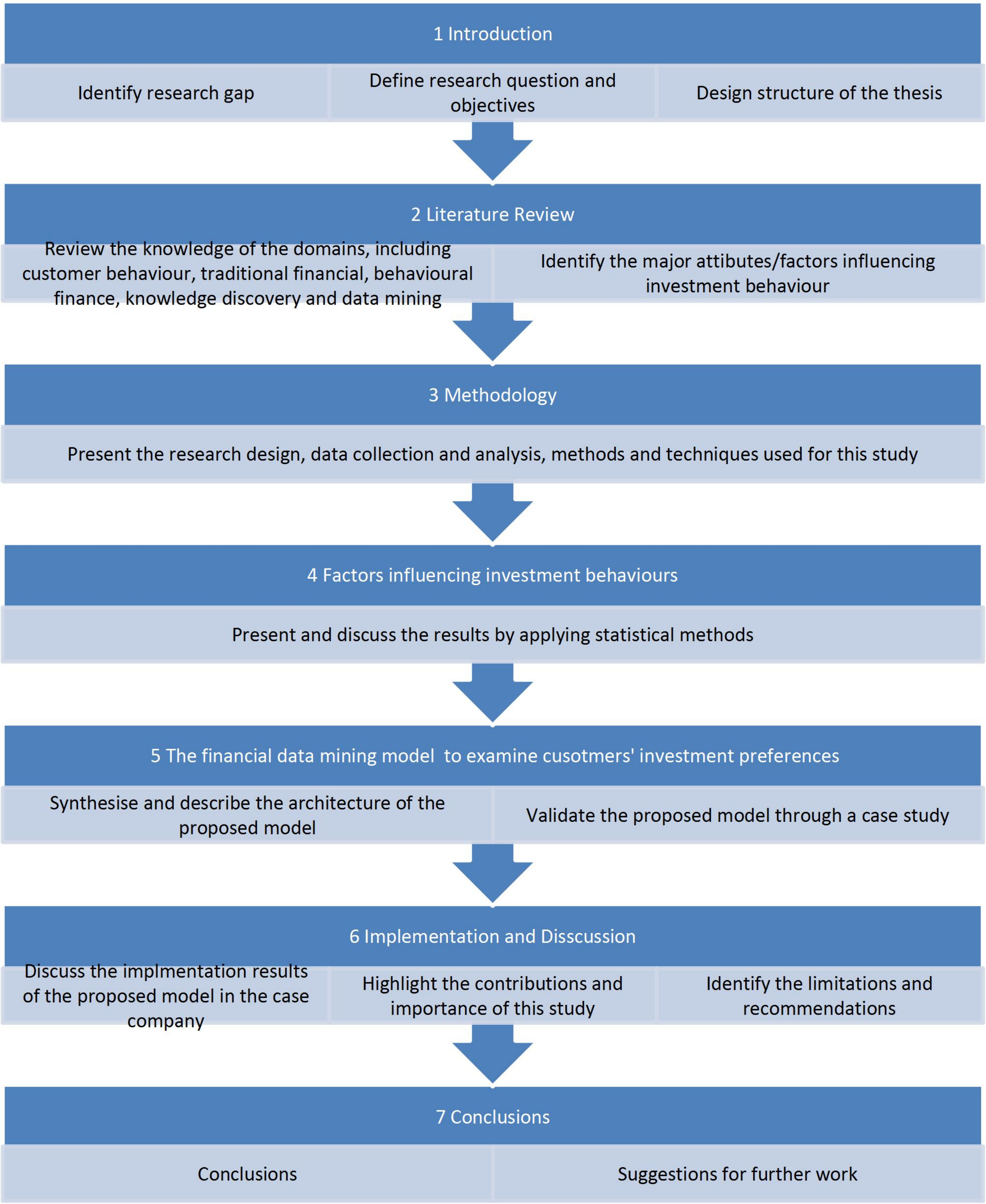

1 Introduction
Identify research gap
Define research question and objectives
Design structure of the thesis
2 Literature Review
Review the knowledge of the domains, including customer behaviour, traditional financial, behavioural finance, knowledge discovery and data mining
Identify the major attibutes/factors influencing investment behaviour
3 Methodology
Present the research design, data collection and analysis, methods and techniques used for this study
4 Factors influencing investment behaviours
Present and discuss the results by applying statistical methods
5 The financial data mining model to examine cusotmers' investment preferences
Synthesise and describe the architecture of the proposed model
Validate the proposed model through a case study
6 Implementation and Disscussion
Discuss the implmentation results of the proposed model in the case company
Highlight the contributions and importance of this study
Identify the limitations and recommendations
7 Conclusions
Conclusions
Suggestions for further work

Figure 1.1 The thesis structure

Chapter 2 Literature review

In order to address the research question this chapter is divided into three parts. First, this chapter reviews current issues discussed in the literature on consumer behaviour, behavioural finance and financial investment behaviour in general. Second, this chapter proceeds to a review of the attributes influencing investment behaviour, so as to identify the major ones in terms of psychological, sociological and demographic factors. Third, studies focusing on existing financial models and data mining techniques to process datasets to analyse customer behaviour are reviewed. This provides guidance for this research by studying how financial institutions predict investment preferences.

2.1 Importance of understanding consumer behaviour in the financial market

In today's increasingly competitive business environment, a clear understanding of the behaviour of consumers is a key element for ensuring business success. Many scholars have attempted to define consumer behaviour. Loudon (2001) and Solomon *et al.* (2012) define consumer behaviour as the study of customers and the processes they use to choose, consume and dispose of products and services that satisfy their needs and influence their experience.

Consumer behaviour defines how customers decide what to buy and what not to buy. Understanding the underlying mechanisms leading to certain customer responses therefore helps business organisations to make better managerial decisions and to provide the right product or service to their customers (East *et al.,* 2013). An in-depth understanding of consumer behaviour further helps business organisations to plan for future buying behaviour patterns by customers and formulate appropriate marketing strategies in order to build long-term customer relationships.

In financial markets, the customers or consumers are investors. Exploring the behaviour of investors is therefore important for financial institutions to devise appropriate strategies and to market appropriate financial products, or offer new financial products to investors in order to better satisfy their needs. During the last decade, defining events such as the collapse of several renowned hedge funds in 1998 and the financial tsunami between 2007 and 2008, have caused a rethink in the domain, directing attention to apparently irrational human behaviour. To study investor behaviour and explain the irrational behaviour, researchers have largely adopted the concept of behavioural finance (Rizvi and Fatima, 2014).

Although behavioural finance has raised the interest of researchers, relevant studies regarding Hong Kong investor behaviour are limited. The majority of the research conducted has focused on the behaviour of Hong Kong stock prices (Wong and Kwong, 1984; Lima and Tabak, 2004; Chan and Chui, 2016). Little research has been devoted to the behaviour of individual investors (Collard, 2009; Fidelity Investments Management (Hong Kong) Limited, 2004). One study conducted by Hon (2013) identified factors, such as age, personal income and investment experience as influencing Hong Kong investors' financial investment behaviour, and provided little managerial insights for financial institutions to utilise.

Hong Kong, as an international financial centre, provides a variety of financial products, such as private equity, venture capital, bonds and stocks for people to invest globally. Due to the proximity, low tax rate, similarities of language and culture as well as global access, Hong Kong remains the top offshore investment destination for mainland Chinese investors. In recent years, wealthy mainland investors have increased their holdings in stocks and other financial products, from 19% in 2011, 33% in 2013 to 37% in 2015 (Sun, 2015). In addition, Hong Kong-domiciled fund products are much more popular on the mainland than in Hong Kong, and mainlanders have been enthusiastic buyers of Hong Kong fund products. According to sales figures released by the China State Administration of Foreign Exchange, the value of Hong Kong-domiciled funds sold in the mainland was almost 70 times that sold in Hong Kong (Yu and Yiu, 2016). These facts, as a result, encourage financial institutions in Hong Kong to review their marketing strategy for targeting this fast-growing market segment, i.e. the mainland Chinese investors buying into the offshore Hong Kong market.

In order to address the research question identified in chapter one, an investigation into the major attributes influencing investment behaviours was conducted and is described in chapter 2.3, this is followed by an analysis of how the major attributes identified affect investors in Hong Kong and mainland China and is discussed in chapter four. A derived model, based on the major attributes identified to classify customer behaviour and applied in a validation case study, are discussed in chapter five.

2.2 From traditional finance to behavioural finance

In traditional finance, the dominant paradigm to understand investor behaviour is that investors make rational choices. Thus, according to traditional finance, investors are rational decision makers and evaluate all possible outcomes to identify the best or optimal choice (Baker and Ricciardi, 2014). Prominent theories such as efficient market hypothesis, capital asset pricing model, model portfolio theory, and more have been formulated. However, since the 2000s, researchers have focused their analysis on various levels within the financial market, such as the aggregate market level, cross-section of average returns, corporate finance, and individual investor behaviour (Barberis and Thaler, 2002). Ackert (2014) conducted a study to compare traditional and behavioural finance and indicated that the application of the concept of behavioural finance in terms of psychology to explain how investors reach their decisions is very popular. This is mainly because the fundamental issues of traditional finance are no longer invalid and thus researchers turned to observe how investors behave in behaviour finance. Bikas et al. (2013) confirmed that investment behaviour is influenced by many factors including psychological factors. This is one of the studies to advocate the emergence and trend of the concept of behavioural finance as they found traditional finance cannot explain the emotional factor on investment behaviour but a limited number of investor rationality. Behavioural finance is the application of psychology to finance, according to Hirshleifer (2014) while Sherfrin (2001) defined behavioural finance as the study of how psychology affects investors' decision making processes. Hussein and Al-Tamimi (2005) added that behavioural finance explains the impact of psychological principles on the behaviour of financial market participants. This emerging concept has fundamentally changed the research context and today many researchers are engaged in behavioural finance investigation.

Behavioural finance offers an alternative tool to study investor behaviour and the causing of market anomalies. Scholars have applied behavioural finance to explain financial market anomalies such as stock market bubbles, over-reaction and under-reaction to new information (Cooper *et al.,* 2001; Zhou and Sornette, 2006) that do not conform to the traditional finance theory. For example, Shleifer (2000) adopted behavioural finance to explain the collapse of several well-known hedge funds in 1998 and discovered that investor rationality was contradicted by psychological evidence. Shefrin and Statman (2011) found that excessive optimism creates speculative bubbles in financial markets. Ritter (2003) posited that behavioural finance has overcome the inability of the traditional finance theory to explain many empirical patterns, such as stock market bubbles in Japan, Taiwan and the U.S. Studies in the field of behavioural finance to investigate the financial crises of 2007 and 2008 were developed by researchers, such as Thaler and Sunstein (2008), Wong and Quesada (2009), and Akerlof and Shiller (2009). These studies were from the perspective of psychology and confirmed that confidence is a factor driving investors' decision-making. Researchers also applied behavioural finance to explain emotional investor behaviour. Through an exploratory study using survey data from more than 350 individual investors, Chandra and Kumar (2012) found that individual investor behaviour is influenced by psychological heuristics, such as overconfidence. Zhang and Zheng (2015) conducted a survey study on investment psychology of Chinese investors in terms of behavioural-psychological characteristics, such as overconfidence, over-reaction and herd effect, to explain market anomalies. In short, behavioural finance has been popular in the literature pertaining to investment behaviour in recent years.

Other researchers have also suggested that sociological and demographic factors are important to explain investors' behaviour (Zhang and Zheng, 2015; Fung and Durand, 2014; Conlin *et al.,* 2015). Though some researchers have studied the impacts of other factors such as gender or age difference on investment behaviour, these studies only explored the influences with regards to investor behaviour, but did not analyse the financial decision-making process of investors or predict their preference regarding financial products. For example, Jiang and Zhu (2006) found that the gender difference in income and employment opportunities affects investors' confidence towards financial investment. Through case studies, Ansari and Moid (2013) investigated the factors influencing investing activities among young professional and concluded that income and age are the dependent factors. These studies within the field of behavioural finance provide evidence that demographic factors such as age and gender should be considered when studying investor behaviour.

While the psychological principle dominates the behavioural finance literature, behavioural finance should be interdisciplinary and incorporate three elements – psychology, sociology and finance (Ricciardi and Simon, 2000). Its core purpose is to provide the framework, so as to gain a better understanding of the investment patterns of investors including what to invest, why to invest, and how to invest. Bikas *et al.* (2013) further elaborated that, in behavioural finance:

i) Psychology is to analyse how the physical, psychical and external environment of human beings affects the processes of behaviour and mind;

ii) Sociology is to analyse how social relations of human beings individually or as a group influence investors' attitudes and behaviour; and

iii) Finance is about the formation and use of financial resources to make investment decisions.

Frankfurter and McGoun (2000) indicated that psychology and sociology are the essence of behavioural finance. However, according to the literature identified researchers have emphasized the importance of psychological factors in the concept of behavioural finance and overlooked the importance of the other factors influencing investment behaviour.

Overall, in order to make the research representative of reality and to better comprehend the way investors behave, this study takes psychological, sociological and demographic factors into account to explore the major attributes of how investors behave. By identifying these major attributes in terms of psychological, sociological and demographic factors this study makes an important contribution towards closing the existing knowledge gap.

2.3 Factors influencing financial investment behaviour

In general, research in behavioural finance provides evidence that investors' decisions are affected by behavioural factors (Jagongo and Mutswenje, 2014). Researchers found that investors do not behave in a merely rational manner across financial markets and that there are a variety of factors influencing their decision-making in investment; among those factors,

i) psychological factors,

ii) social factors, and

iii) demographic factors,

are major elements (Kumar and Lee, 2006; Baker and Wurgler, 2007; Gärling *et al.,* 2009; Barnea *et al.,* 2010). Throughout the relevant literature, factors influencing investors' behaviour are generally classified into these three groups. In fact, it seems investor behaviours are always a mix of the rational and irrational (Dreman and Berry, 1995); there is little evidence in the literature that investors' decisions are completely made of either rationality or irrationality. To study behaviour under more realistic conditions and to better categorise the way investors behave, this study identifies and evaluates the major attributes explaining investment behaviour under three constructs – psychological factors, sociological factors and demographic factors, and how these factors impact investors' decision-making. This understanding may help financial institutions have better strategies for their business.

Reviewing the major literature on factors influencing investor behaviour, a range of psychological, sociological and demographic factors has been identified, and is discussed below:

2.3.1 Key psychological factors

Regarding psychological factors, research suggests that individual investors are driven by experience or through an investment appraisal process to make investment decisions (Kaustia and Knüpfer, 2007, 2008; Korniotis and Kumar, 2006; Feng and Seasholes, 2005; Malmendier and Nagel, 2011; Seru *et al.,* 2010). Past experience, as a consequence, affects investors' risk perception in terms of attitude to risk and risk tolerance (Corter and Chen, 2006). This is also supported by Byrne (2005), indicating the positive correlation between investment experience and risk. Chen et al (2007) further pointed out that accumulated investment experience significantly affects investment decisions of individual investors in terms of anchoring bias, and over-confidence (Wong and Lai, 2009). According to Korniotis and Kumar (2006), investors who are more experienced in investments tend to hold less risky portfolios, trade less frequently and have better diversification skill when compared with less experienced investors. Thaler and Johnson (1990) found that investor experience in terms of overall gains and losses influence investors' risk attitude and future decisions on the overall portfolio. This indicates that the investment experience of individual investors forms a stronger basis for investment decision rather than other psychological factors, such as attitude to risk (Fellner and Maciejovsky, 2007), risk tolerance (Grable and Lytton, 1999; Wang and Hanna, 1997), anchoring bias (Wong and Lai, 2009) and overconfidence (Wong and Lai, 2009), and is therefore included in this study by considering it as an important psychological factor that influences financial investment behaviour.

2.3.2 Key demographic factors

Demographic factors play a significant role in determining the behaviour of investors (Sadi *et al.,* 2011 and Maditinos *et al.,* 2007) and influence their choice regarding investment products (Charles and Kasilingam, 2013; Fellner and Maciejovsky, 2007; Kahneman and Tversky, 1979; Mittal and Vyas, 2008; Weber *et al.,* 2002). According to Watson and McNaughton (2007), age and choice of risky financial investment products tend to indicate a negative correlation. Huberman and Jiang (2006) found that older investors are likely to select a small number of funds, and choice is independent from the number of funds offered, as they are not willing to take more risk. Similarly, Speelman *et al.* (2013) pointed out return-chasing behaviour consistently increases with age and thus investors prefer financial products at lower risk level when getting older. Thus, age can be considered an essential factor and has a significant relationship with investment behaviour according to the literature.

Age and gender are often closely interacting. For example, Byrnes *et al.* (1999) showed that male investors have significantly higher propensity for risk taking than female investors, while this tendency decreases with age. In general, research found that gender plays an important role in influencing an individual's risk attitude and subsequently affecting decision making regarding investments (Donkers *et al.,* 2001; Kabra *et al.* (2010). Gender differences have been found regarding risk attitude and thus in the selection of financial investment products (Fellner and Maciejovsky, 2007; Hartog *et al.,* 2002; Weber *et al.,* 2002). Many existing studies support that women are more conservative than men when investing and are reluctant to take risks (Hartog *et al.,* 2002; Agnew *et al.,* 2008). Thus, female investors tend to show greater risk aversion than male investors (Speelman *et al.,* 2013).

Gunay and Demirel (2011) carried out a study about investment decisions and indicated that demographic factors, e.g. gender, impact investors' behaviour. For example, women prefer to "wait it out" when their investment does not produce an expected return. Men, on the contrary, are more likely to make investment changes than women when an investment does not produce expected return (Hira and Loibl, 2008). For financial institutions to offer financial products which are best suited for investors of different genders, understanding the gender difference in the investment behaviour of individuals is crucial and thus taken into account in this study.

2.3.3 Key sociological factors

According to the extant literature, level of income, marital status and level of education are significant factors determining investors' behaviour and their investment decisions (Kaleem *et al.,* 2009; Geetha and Ramesh, 2012; Fares and Khamis, 2011; Obamuyi, 2013). Shaikh *et al.* (2011) conducted an exploratory study and confirmed that these three factors highly influence investors' behaviour and decision making. Rizvi and Fatima (2015) found a significant correlation between income and investment, specifically that investors having higher income invest more frequently in the financial market. Batemany *et al.* (2008) revealed that higher income investors respond positively to increasing variance in returns when risk is presented for maximising the possible investment outcomes. Al-Ajmi (2008) added that in addition to income level, level of education is highly related to investors' decisions, namely that more wealthy and more educated investors are more risk tolerant.

This finding is supported by Fares and Khami (2011), who identified that education level of investors is highly relevant to investment decisions. Through a national survey, ACNielsen Research (2005) found that the lowest level of financial literacy was associated with people who are less educated, have lower income and are not married. Based on these literature review findings, the three factors of level of income, education level and marital status are included in this study.

2.4 A summary of factors influencing investment behaviour

The key attributes influencing investment behaviours and the key findings derived from the literature are summarised in Table 2.1. The six attributes, including investment experience, age, gender, economic status/income, marital status and education level identified from the literature review, two additional attributes, household net worth and nature of employment, should be considered in this study. The rationale and critical analysis for the attributes selection is discussed in Table 2.1.

Table 2.1 List of attributes influencing investor behaviours

Attributes identified	**Key findings from the literature**	**Critical analysis**
Investment experience	• Experience affects investors' risk perception in terms of attitude to risk and risk tolerance (Corter and Chen, 2006; Thaler and Johnson, 1990; Byrne, 2005) • Investors who are more experienced in investments tend to hold less risky portfolios, trade less frequently and have better diversification skill when compared with less experienced investors (Korniotis and Kumar, 2006) • Accumulated investment experience significantly affects investment decisions of individual investors (Chen *et al.*, 2007) • Investors always learn from their experience that has an impact on their investment expectation and choice (Seru *et al.*, 2010; Malmendier and Nagel, 2011)	The findings in the literature relating to investment experience are consistent and confirm factors that significantly affects investment behaviours. Investment experience of investors forms a strong basis for investment decision, and is therefore included in this study.
Age	• Age and choice of risky financial investment products tend to indicate a negative correlation (Watson and McNaughton, 2007) • Older investors are likely to select a small number of funds, and choice is independent from the number of funds offered, as they are not willing to take more risk (Huberman and Jiang, 2006) • Return-chasing behaviour consistently increases with age and thus investors prefer financial products at lower risk level when getting older (Speelman et al., 2013) • Age plays a crucial role on their behavioural biases and success of their investment decisions (Charles and Kasilingam, 2013) • Investing activity of young professionals is dependent on income (Ansari and Moid, 2013)	Age is an essential factor and has a significant relationship with investment behaviour, thus is also included in this study model.

Attributes identified	Key findings from the literature	Comments
Economic status / income	• A significant correlation between income and investment, specifically that investors having higher income invest more frequently in the financial market (Rizvi and Fatima, 2015) • Higher income investors respond positively to increasing variance in returns when risk is presented for maximising the possible investment outcomes (Batemany *et al.*, 2008) • Investors with lower income level preferred to take more safety in investment (Manish and Vyas, 2008) • Investing activity of young professionals is dependent on income and age (Ansari and Moid, 2013)	These three factors are inter-related and found to have crucial implication on investment behaviors, thus they are considered in this study. In the current literature, annual income of individual investor is an indicator commonly used to measure the economic status of investors. With reference to economic/financial reports (Johnson, 2016; Zumbrun, 2017; OECD, 2017), economic status in term of household net worth is commonly used. Is household net worth a better indicator to be considered in this study? Besides, investors' economic status depends on the nature of employment, as suggested by Appuhami (2007). In order to identify the major attributes influencing investment behaviours, annual income, household net worthy and nature of employment are taken into account for further investigation.
Marital status	• Married investors are more proactive investors (Iqbal, 2011) • A national survey (ACNielsen Research, 2005) indicated that the lowest level of financial literacy was associated with people who are less educated, have lower income and are not married	
Education level	• Level of education is highly related to investors' decisions (Reitan and Sorheim, 2010; Fares and Khami, 2011), namely that wealthier and more educated investors are more risk tolerant (Al-Ajmi, 2008)	

Attributes identified	Key findings from the literature	Comments
Gender	• Gender impacts investors' behaviour, for example, women prefer to "wait it out" when their investment does not produce an expected return (Gunay and Demirel, 2011) while men are more likely to make investment changes than women when an investment does not produce expected return (Hira and Loibl, 2008) • Gender plays an important role influencing individual's risk attitude subsequently affecting decision making regarding investments (Donkers et al., 2001; Kabra et al., 2010) and the selection of financial investment products (Fellner and Maciejovsky, 2007; Hartog et al., 2002; Weber et al., 2002) • Women are more conservative than men when investing and are reluctant to take risks (Hartog et al., 2002; Agnew et al., 2008; Speelman et al., 2013) • Male investors have significantly higher propensity for risk taking than female investors, while this tendency decreases with age (Byrnes et al., 1999)	For financial institutions to offer customised products which are better suited for investors of different genders, understanding the gender difference in the investment behaviour is crucial.

2.5 Difficulties in understanding financial investment behaviour

It is of paramount importance for financial institutions to recognize and predict investors' preferences towards various financial products. They can thus meet their customers' needs by creating the best investment portfolio for them. This understanding helps effectively market the most suitable financial products to particular customers, in a highly competitive environment. However, many studies rarely identify factors influencing actual investor behaviours and then use these factors to predict investors' preferences. This is because financial investment behaviours are complex and influences are difficult to isolate, in particular the hidden relationships regarding investors' behaviour toward their investment preference. This is due to the lack of consensus concerning the validity of the various hypotheses or factors identified (Narayan *et al.,* 2015). Another reason is the rocketing growth of financial data, which makes the understanding of financial investment behaviour very difficult (Vazirgiannis *et al.,* 2003) and increases the complexity of efficiently processing massive amounts of accumulated data related to customers and financial products (Zhang and Zhou, 2004). At the same time, the rapid growth of financial data could also be viewed as an opportunity providing huge datasets supporting deductive research and making findings more meaningful. A literature review of the domain shows 176,000 articles published since 2010 and suggests a lack of access to the huge volumes of financial transaction data, which are kept strictly confidential by financial institutions, to be the biggest obstacle for researchers to conduct this kind of research. With reference to the discussion of expert panel conducted in this study (see chapter 3.2.3), the general consensus from HK experts is that database owned by each financial institution containing sensitive data about clients and transactions is a critical asset and cannot be disclosed to external parties in view of the stringent privacy law. Expert from China expressed that privacy concern is arising in China and generally they will not disclose information to external parties unless those parties have a very good and strong relationship with their top management. Therefore, financial institutions in Hong Kong and China would only analyse investment behaviour and discuss key findings internally.

Turning data into information and then into knowledge for understanding investors' behaviour in order to offer investors what they want at the right time is a conversion process that often involves manipulation and analysis of large volume of data about the customers. For instance, in the financial service industry, financial specialists analyse the current trends and latest stock prices in the stock market. Financial specialists then generate a report for analysis and investment decision making. This kind of manual analysis is time consuming and is subject to highly subjective decision making. With technological advancement, researchers try to identify unknown and valuable knowledge from historical financial data using techniques such as data mining (Enke and Thawornwong, 2005; Liao and Chou, 2013). Enke and Thawornwong (2005) adopted neural network models for the estimation and classification of levels to provide a forecast of future stock market returns, while Liao and Chou (2013) adopted the data mining approach for investigating the co-movements on the Taiwan and mainland China stock markets. Even though there are various data mining models identified in the literature review, a model designed for facilitating financial institutions to predict customers' investment preferences appears not to exist in the literature. This statement was discussed at the expert panel conducted in this study (see chapter 3.2.3) and got full support from the finance experts in Hong Kong and China, who agreed that there was a necessity to develop an applied model to predict investment preferences but they could not find any in the literature nor from the industry. In the discussion, some experts criticised that most of the current models for studying investment behaviour are theoretic rather practical and realistic. Experts from Hong Kong indicated that their financial institutions have no models or just use simple customer relationship management analysis for the various product types purchased by customers. They have not found any data mining model on investment behaviour for Hong Kong and China but they do need such a model for evaluation and promotion of business.

In light of these difficulties in understanding financial investment behaviour, how can financial institutions better customize investment portfolio to satisfy customers? This study addresses this research gap by investigating the hidden relationships in investor behaviour based on the major determinants identified (in chapters 2.3.1 to 2.3.3) to extract likely customer behaviour by transforming investor and financial data into knowledge. The findings may give additional insights for financial institutions regarding the selection of investment strategy and the design of financial products best suited for specific investors.

2.6 Knowledge discovery and data mining

Knowledge discovery techniques aid in identifying correlations, relationships and hidden patterns in business data, and have been widely accepted as an important approach for innovation in business (Vazirgiannis *et al.,* 2003). In recent decades, data mining has been one of the most widely applied techniques in the business sector due to its efficiency in handling large numbers of records and data (Ravisankar, *et al.,* 2011). Data mining refers to the method and technique used for identifying and extracting hidden relationships or significant patterns from large data sets (Berry and Linnoff 2004; Chen *et al.,* 1996; Vazirgiannis *et al.,* 2003). In the field of behavioural finance, data mining technique could be practical and useful for discovering investment behaviours, and converting data into knowledge for the enhancement of customer satisfaction.

Many companies apply data mining techniques to identify potential problems and allow managers to make corresponding strategic decisions to outperform their business competition (Pivk *et al.,* 2013; Rygielski *et al.* 2002; Wu *et al.,* 2012). Once data such as the characteristics of customers are collected, companies can utilise it to improve workflow (Herbst and Karagiannis, 2004; Ho *et al.,* 2009), while deepening the understanding of customer behaviour (Batra *et al.,* 2012; Huang and Hsueh, 2010; Mehta and Dang, 2012; Nanda *et al.,* 2010). Data mining has been applied to a broad range of areas from customer relationship management (Hosseini *et al.,* 2010; Ngai *et al.,* 2008), biomedicine (Ting *et* *al.,* 2009), detection of fraudulent financial statements (Kirkos *et al.,* 2007) to the examination of investor behaviour (Mehta and Dang, 2012). While there are a wide range of data mining models, one specifically for financial institutions to efficiently predict customers' investment preferences is not readily available. To develop such a model for predicting and explaining investment behaviour, this study utilises data mining, as it is a proven and viable information extraction and analysis method. The proposed model should support financial institutions in extracting hidden patterns in investor behaviour, and guide the corresponding investment decisions for satisfying customers through the design of customised products and services.

2.6.1 Data mining techniques for financial sector

In order for financial institutions to improve their market analysis and customer relationship management, the utilization of data mining technique is becoming increasingly significant (Batra *et al.,* 2012; Mehta and Dang, 2012; Nanda *et al.,* 2010). These help financial institutions enhance competence and sustain the continual development of the company. The four most common data mining techniques in the financial industry are:

i) clustering analysis (Berry and Linoff, 2004),

ii) association rules (ARs) (Kim *et al.,* 2004),

iii) fuzzy logic (Novak, 2012; Ordoobadi, 2009), and

iv) neural network (NN) (Olson and Shi, 2007).

These four techniques are reviewed to identify the most appropriate method for this study. The technique must support the development of a financial data mining model to explore customers' investment preferences.

2.6.1.1 Clustering analysis

Cluster analysis is a process commonly used to identify significant distributions and patterns in datasets with respect to the data's characteristics on the basis of self-similarity (Berry and Linoff, 2004). Objects such as events and people in a cluster are similar and different from objects in other clusters (Chen *et al.,* 2013). Clustering analysis is "one of the most widely-adopted key tools in handling data information" (Chen *et al.,* 2013, p.2198) and has a wide range of applications in the fields of information retrieval (Mizutani *et al.,* 2008; Miyamoto, 2003) and pattern recognition (Bezdek, 1998) to resolve real world problems. For example, Kuo *et al.* (2007; 2009) adopted clustering analysis for market segmentation by identifying the characteristics of customers' buying behaviour based on different clusters, such as income, age, and marital status. Then customer satisfaction can be enhanced by formulating more appropriate marketing strategies, providing tailor-made services, and making recommendations according to the preferences of different customer clusters.

Clustering analysis has been applied for various purposes in the financial sector. For instance, Kou *et al.* (2014) used the clustering method in financial risk analysis (Phua *et al.,* 2005). Munnix *et al.* (2011) studied characteristic correlation structure patterns using the daily data of S&P 500 stocks in the 1992-2010 period to examine market similarity and financial market states. Chaudhuri and Ghosh (2016) used clustering method to assess daily data from the Indian stock market for two years to understand stock market volatility. Batra *et al.* (2012) conducted a data mining analysis to identify investors' perceptions regarding different investment options, while many other studies analysed investors' preferences in the stock market using clustering analysis (Kascelan *et al.,* 2014; Kerby and Lawrence, 2003; Someswar *et al.,* 2012; Lee *et al.,* 2010; Kopeti, 2010; Suliman and

Obaid, 2013). Nanda *et al.* (2010) categorized stocks on investment criteria for building investment portfolio using k-means cluster analysis.

Among different clustering algorithms, such as K-means algorithm, expectation-maximization algorithm and hierarchical clustering, So and Yoon (2008) claimed that other clustering algorithms generate too many clusters and are complicated to operate, compared with K-means algorithm which can practically derive a set of desirable clusters. K-means is the simplest and easiest algorithm to classify a given data set that solve known clustering problems in different scenarios (Wu and Kumar, 2009), which makes it the most widely used clustering algorithm in practice, supported by MacQueen (1967), Jain (2010), Gan and Wu (2007). In the financial field, Nanda *et al.* (2010) supported that K-means analysis helps build the most compact clusters and can reduce the time of stock selection. Using K-means clustering analysis can provide investors with insightful information to make better decision (Momeni *et al.*, 2015). Overall, the K-means clustering method being a simple and easy way to cluster a data set was found useful in the financial sector, and therefore suitable for this study to cluster investor behaviour.

2.6.1.2 Association rules (ARs)

ARs are one of the most widely used data mining techniques for discovering correlations between items in a database. Kim *et al.* (2004) commented that ARs are similar to if-then rules, in which a condition clause (if) triggers a result clause (then). In ARs, three thresholds, namely, support, confidence and lift ratio, can be used to describe an association rule (Berry and Linoff, 2004). Among these, support and confidence are defined as the measurement standards (Tan, *et al.*, 2005; Lai and Cerpa, 2001; Laio, et al., 2011).

i) Support: indicates the percentages of records containing an item or combination of items to the total number of records.

ii) Confidence: reflects the certainty that when the 'if' part is true, the 'then' part is also true under a particular condition.

iii) Lift ratio: shows how much the quality of the rule estimating the 'then' part is superior compared to having no rule.

The Apriori Algorithm proposed by Agrawal and Srikant (1994) is one of the most widely used AR methods (Kuo *et al.,* 2011) and is considered as a classical algorithm for effectively generating association rules between items in large databases (Chung and Tseng, 2012; Lim *et al.,* 2012). It is used for discovering frequent "item-set" in a database, then calculating the support for each "item-set" in order to determine whether they can be identified in the database for association rule group. According to Agrawal and Srikant (1994), an association rule is in the form of X→Y, where X and Y indicate a combination (item-set). They also define the following equations:

The minimum support (X → Y): $\frac{\text{No.of transactions which contain X } and\ Y}{\text{No.of transactions in the database}}$ Equation (2.1)

The minimum confidence (X → Y): $\frac{\text{No.of transactions which contain X } and\ Y}{\text{No.of transactions which contain X}}$ Equation (2.2)

Both the minimum support and confidence are parameters of association rules. The determination of the parameters will affect the results from rules extracted. Liao *et al.* (2008) suggested that the confidence level above 70 and support of itemset above 8 can be considered meaningful.

ARs are getting increased use in different areas. The typical application of ARs is the market basket analysis which discovers customer purchasing patterns by “extracting associations or co-occurrences from stores' transactional databases” (Chen *et al.,* 2005, p.1). The market basket analysis enables the retailers quickly and easily to discover the buying patterns of their customers so that they can increase the size and value of the basket of purchases. With the understanding of the purchasing patterns of customers, the company can effectively formulate customised marketing strategies as well as enhance customer value and extend the customer life cycle by employing the association rules (Chiang, 2011). For example, a study using association rule analysis discovered that male customers in Wal-Mart supermarkets who purchase baby diapers would also buy several bottles of beer. Wal-Mart then launched a promotion selling diapers and beer as a pack with surprising results: Sales of both improved significantly (Xu, 2016).

In addition to marketing analysis, application of ARs in the finance industry is also gaining popularity. Li *et al.* (2008) demonstrated how to use ARs to discover financial investment behaviours in the Shanghai stock market and apply ARs to an actual securities clearing dataset. Their study helped financial institutions identify patterns regarding how a financial investment portfolio can be built and to learn more about behavioural finance in general. Sung and So (2011) also used ARs for predicting changes in the Korea Composite Stock Price Index (KOSPI) based on the data of global stock market indices. The rules generated are expected to facilitate the decision making on buying or selling particular types of stock. Their study also revealed that using large sample sizes of raw data not only make the results more applicable, they can also be useful for finding unexpected patterns and rules. Kuo *et al.* (2009) proposed a novel algorithm, applying particle swarm optimisation to generate ARs. Their study discovered correlations between industrial categories in Taiwan by applying this algorithm to stock selection behaviour. The mining results can provide an insight into customers' transaction behaviour and information on decision making. Kumar and Kalia (2011) adopted ARs in obtaining frequent item-sets. For example, it was discovered that a portfolio with one finance company and one steel company would be built and get better returns in long term. The patterns obtained helped investors learn more about investment planning and build investment portfolios.

2.6.1.3 Fuzzy logic

Fuzzy logic (FL) theory builds on the basis of fuzzy sets, which aims at modelling two remarkable human abilities:

i) an ability to make rational decisions in an environment with incomplete information, and

ii) an ability to perform various physical and mental tasks without any measurements or any computation (Zadeh, 2008).

FL is an effective tool for managing imprecise attributes by offering a mathematical model (Novak, 2012; Ordoobadi, 2009). The fundamentals of FL are linguistic terms. FL mimics human decision making by performing approximate reasoning with linguistic terms so as to generate solutions (Liu and Lai, 2009; Wong and Lai, 2011). In FL, linguistic terms are represented by fuzzy sets which are employed to develop causal relationships between input and output variables (Tahera *et al.,* 2008). Each of the fuzzy sets is associated with a membership function, which allows variables to carry a degree of membership in a fuzzy set within a range of 0 to 1 (Azadegan *et al.,* 2011; Otero and Otero, 2012). There are three main components in a fuzzy system; they are:

i) fuzzification,

ii) inference engine and

iii) defuzzification (Lau *et al.,* 2009).

Fuzzification is responsible for converting crisp input values into fuzzy sets. These fuzzy sets are then transferred to an inference engine for converting input fuzzy sets into output fuzzy sets on a basis of a collection of fuzzy rules. Each fuzzy rule, in the form of if–then–else rules (Hajek, 2012; Lin and Lee, 1991), implies a fuzzy relationship between an antecedent and a consequence (Otero and Otero, 2012). Defuzzification is carried out to convert these output fuzzy sets into crisp values, as only exact numerical values are needed in actual control operations.

Applied in the fields of production priority management (Díaz *et al.,* 2004), quality management (Lau *et al.,* 2009) and production scheduling (Petrovic and Duenas, 2006), FL has also been applied in the financial and insurance industry. An expert opinion-based study (Hellman, 1995) revealed that the fuzzy expert system can identify and classify economic and insurance factors, so that insurance companies can evaluate and rank customers based on these two factors for bonus tariff premium decisions. Many researchers employed FL to evaluate the weight of a qualitative variable on the stock market trend (Dourra and Siy, 2002), examine credit risk (Lahsasna, 2009; Sreekantha and Kulkarni, 2008; 2012), analyse financial safety of monetary organisation (Khovrak and Petchenko, 2015) and predict arbitrage opportunities in the stock markets (Bernardo *et al.,* 2013). Dourra and Siy (2002) further suggested that using FL through technical analysis can facilitate investors' decision-making on trading financial products. Different strategies can therefore be implemented using the fuzzy indicator to match investors' preferences and industry conditions.

2.6.1.4 Neural network (NN)

A "neural network (NN)" can generate learning from a collection of historical data. Such a set of data are called a learning set. Each variable of input data has a node in the first layer. The last layer has one node for each classification category. In general, NN have at least one hidden layer of nodes, adding complexity to the model. Each node is connected by an arc to nodes in the next layer. The desired outputs can be adapted by modifying the connections between the nodes. Figure 2.1 illustrates the multi-layers of NN between input and output variables.

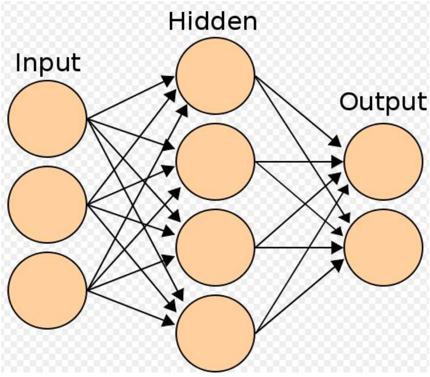


Figure 2.1 A multi-layer percepton (Rajola, 2003)

Arcs between the nodes have weights which are multiplied by the value of incoming nodes and summed. The values of the variables in the data set determine the values of the input nodes. The values of the middle layer nodes are the sum of the values of the incoming nodes, multiplied by the arc weights. In turn, the middle layer nodes are multiplied by the outgoing arc weight to successor nodes (Olson and Shi, 2007). The output for starting weights can be calculated based on a given input value. The resulting output is compared to the target values. When there is difference between the obtained output and the target output, the target output is fed back to the system to adjust the weights on arcs. This process is repeated until the network correctly identifies the proportion of learning data specified by the user. This is a simplified explanation; much further work has been done in exploring applications and models of neural networks (Maier and Dandy, 2000; Tang *et al.*, 2007).

NN is mainly used for classification or prediction in data mining. They can be fed data without a starting model estimation. Thus, they are widely applied in the financial (Chen and Leung, 2004), manufacturing (Wu, 2009), telecommunications (Chen *et al.,* 2011; Sohn and Kim, 2008) and many other sectors.

In the financial sector, NN has been applied to predict future stock indices (Chen *et al.,* 2003; Chenoweth and Obradovic, 1996; Desai and Bharati, 1998; Enke and Thawornwong, 2005; Motiwalla and Wahab; 2000; Qi and Maddala, 1999), to forecast foreign exchange rates (Chen and Leung, 2004), to identify trading preferences (Tsai *et al.,* 2009), to predict bank performance (Sharma and Shebalkoy, 2013) as well as financial distress (Chen and Du, 2009). The majority of researchers applied NN for future stock prediction (Chenoweth and Obradovic, 1996; Desai and Bharati, 1998; Enke and Thawornwong, 2005; Motiwalla and Wahab; 2000; Qi and Maddala,1999), while only limited prior research used NN in other financial areas (Chen and Du, 2009; Chen and Leung, 2004; Sharma and Shebalkov, 2013; Tsai *et al.,* 2009). Chen and Du (2009) proposed a neural network and data mining based financial distress prediction model. Empirical experimental results found that the usage of data mining and NN could be a more suitable methodology than statistical methods for financial distress prediction. Chen and Leung (2004) proposed a general regression neutral network to correct the errors of exchange rate forecasting. Results indicated that the proposed model can correct the errors in forecasting as compared to the previous single-stage model. Their proposed model also helped increase investment returns due to improved performance of overall forecasting. Sharma and Shebalkov (2013) presented an application of NN and simulation modelling to predict bank performance. Tsai *et al.* (2009) used neural networks and decision trees to identify trading preferences on the Taiwan stock market and found that characteristics of individual investors do affect their investment decisions.

2.6.2 Comparison of data mining techniques

The literature review of different data mining techniques shows that discovering critical hidden patterns in databases and converting the data into knowledge is a widely-used way to create business opportunities and facilitate effective marketing strategies. Data mining techniques help improve data management and the understanding of customer behaviour. Table 2.2 summarises the four commonly used data mining techniques in the financial industry for comparison.

Table 2.2 Comparison of data mining techniques

Techniques	**Strengths**	**Weaknesses**	**Common Financial-related Applications**
Clustering analysis	• Data grouped by self-similarity (Berry and Linoff, 2004) • Effective in information retrieval (Mizutani *et al.,* 2008; Miyamoto, 2003) and pattern recognition (Bezdek, 1998) • K-means clustering algorithm is simple, and can be easily understood; no technical knowledge is required (Wu and Kumar, 2009)	• Relationship among data relies on user decision (Berry and Linoff, 2004)	• Analyse financial risk (Phua *et al.,* 2005) • Explore stock characteristics (Nanda *et al.,* 2010) • Identify investors' perceptions about different investment options (Batra *et al.,* 2012) • Analyse investors' preferences (Kascelan *et al.,* 2014)
Association rules	• Effective for generating association rules between items in large databases (Chung and Tseng, 2012; Lim *et al.,* 2012) • Allows laymen to understand and implement results easily (Na and Sohn, 2011) • Discover customer purchasing patterns (Chen *et al.,* 2005)	• Determination of parameters affect the results of rules extracted, but it is acceptable and meaningful if the confidence level is set above 70 and the support is above 8 (Liao *et al.,* 2008)	• Discover financial investment behaviour (Li *et al.,* 2008) • Analyse stock characteristics (Kumar and Kalia, 2011) • Predict changes in the Korea Composite Stock Price Index (Sung and So, 2011) • Discover stock selection behaviour (Kuo *et al.,* 2009)
Fuzzy logic	• Effective in managing imprecise attributes (Ma *et al.,* 2006; Novák, 2012; Ordoobadi, 2009) • Able to deal with vague data (Novák, 2012)	• Difficult to understand (Shapiro, 2002) • Requires high levels of technical expertise (Prato, 2005)	• Evaluate credit risk (Lahsasna, 2009; Sreekantha and Kulkarni, 2008; 2012) • Classify economic and insurance factors (Hellman, 1995)
Neural network	• Pre-specification is not required during the modelling process (Enke and Thawornwong, 2005) • Able to solve problems involving non-linear modelling (Enke and Thawornwong, 2005)	• Training is very time consuming and requires practise (Pal and Mitra) • Not suitable for large data sets (Vazirgiannis *et al.,* 2003).	• Predict future stock index (Chen *et al.,* 2003; Chenoweth and Obradovic, 1996; Desai and Bharati, 1998; Motiwalla and Wahab; 2000; Qi and Maddala, 1999).

The literature reviewed above illustrates that studies relevant to financial aspects using data mining techniques focused on overall market trends, such as composite stock price indices, financial market status's and stock characteristics, but not individual investor behaviours and investment products not directly related to stock. This suggests the use of data mining techniques in the present study to explore how individual investors behave and how they build their investment portfolio would be a novel application of them. A financial data mining model to study customer behaviour is therefore proposed for this study in order to achieve the associated research objective (iv) - studying how financial institutions predict investment preferences - identified in chapter 1.2.

A significant financial data mining model requires a large dataset because there inevitably real data always contains a lot of noise. Comparing the strengths and weaknesses of the four commonly used data mining techniques (see Table 2.1), the most appropriate ones for the purposes of the present research are clustering analysis and ARs, for they are effective in handling large database and are easily manipulated and implemented to track the changing preferences of investors.

The selection of clustering analysis is supported by extensive literature showing that by applying clustering analysis in the financial industry, specifically to analyse investors' perceptions, how effective for information retrieval and how simple in application (Phua *et al.,* 2005; Nanda *et al.,* 2010; Batra *et al.,* 2012; Kascelan *et al.,* 2014) the approach is. Using an industry tested and approved approach in the analysis of investors' behaviour and their characteristics builds on a sound foundation before going beyond the present state of the art.

ARs have been proven to have the advantage of being effective in generating associations between items in large databases (Chung and Tseng, 2012; Lim *et al.,* 2012). This study adopted ARs to discover hidden relationships regarding investors' choice of financial products, as ARs are capable of dealing with massive amounts of investors' data.

Clustering analysis and ARs are sometimes used together to find hidden patterns in datasets (Kuo *et al.,* 2007; Sohn and Kim, 2008); clustering analysis verified suspected associations between factors and ARs identified unexpected associations between factors. For example, Sohn and Kim (2008) used the combination of clustering analysis and ARs to examine hidden patterns and concluded that the integrated approach can show more specific marketing implications. Kuo *et al.* (2007) applied clustering algorithms to discover clusters in the database, followed by ARs to discover correlations between the clusters. Results revealed that the integration of clustering and ARs can help extract the rules much faster and find more useful rules. Thus, both clustering analysis and ARs are applied in this study, so that:

i) Clustering analysis is used to segment investors, and

ii) ARs are then applied to find relationships and rules in the segmented groups, thereby effectively providing a basis for forecasting and decision making.

This integrated approach has the potential to minimise the mining time and increase the accuracy of the rules, while building a more customised financial investment portfolio. The results can help financial institutions formulate customised products and appropriate marketing strategies according to different characteristics of the clusters for investors to build their portfolio and to optimise returns.

Chapter 3 Methodology

The literature review in chapter two and the expert panel sessions conducted (see chapter 3.2.3) identified the following key parameters for an investigation on the impacts on investors' behaviour.

i) Investment experience

ii) Age

iii) Gender

iv) Economic status/income

v) Marital status

vi) Education level

vii) Household net worth

viii) Nature of employment

This chapter describes the research design and methodology used to address the research question and associated objectives identified in chapter one. The dataset specifically collected for this study are discussed, followed by a discussion on the various methods for data analysis. Methods such as descriptive analysis, factor analysis, correlation analysis, regression analysis and data mining techniques, with their advantages and disadvantages are assessed. The details of the case study data used to derive and validate the outcomes of this research are also discussed in this chapter.

3.1 Research design

The literature review in chapter two highlighted that financial investment behaviour is commonly influenced by demographic, psychological and sociological factors; however, little research has been devoted to the behaviour of individual investors (Collard, 2009; Fidelity Investments Management (Hong Kong) Limited, 2004). Hong Kong is the top offshore investment destination for mainland Chinese investors. Investors in Hong Kong appear to display different characteristics between locals and mainland Chinese customers. This study is devoted to the examination of the influences of those different characteristics. In order to achieve the research objectives identified in chapter 1.2, the research design was formulated as shown in Figure 3.1.

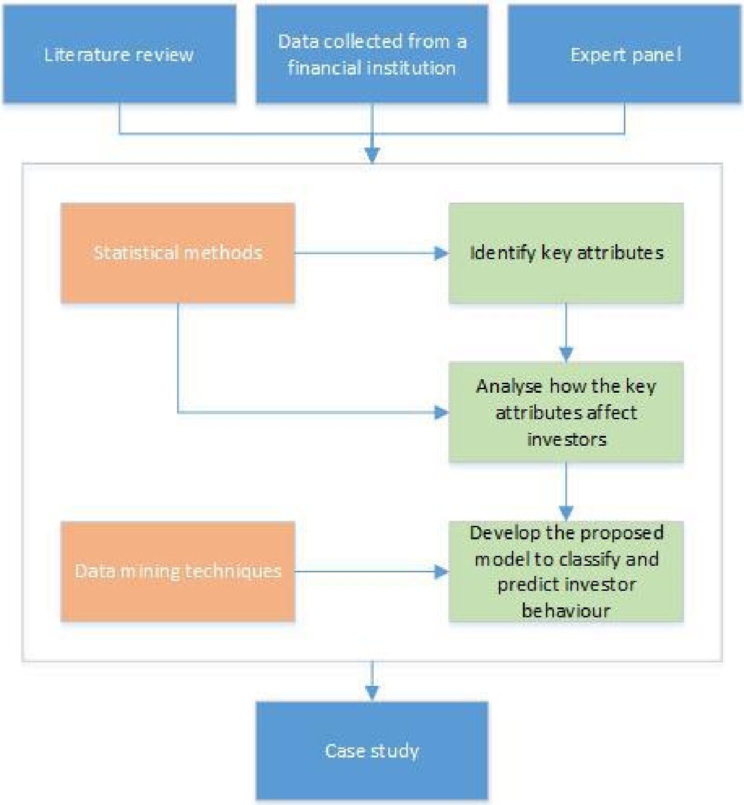

Literature review
Data collected from a financial institution
Expert panel
Statistical methods
Identify key attributes
Analyse how the key attributes affect investors
Data mining techniques
Develop the proposed model to classify and predict investor behaviour
Case study

Figure 3.1 Research design

As the variables in terms of investor attributes and investment preferences are qualitative, but the impact of the attributes on investment preferences as well as the prediction of investor needs are quantitative, an integrated approach using both qualitative and quantitative methods was adopted for this study.

To understand financial investment behaviour as well as to identify the variables in terms of investor attributes, the literature review was conducted to serve as a basis for designing further experimentation. A set of attributes, such as age, marital status, investment experience and annual income, were selected for this research derived from the literature review provided in chapter two.

Large volumes of financial transaction data and investors' descriptive information were collected and analysed for this study. As discussed in chapter two, access to this type of customer data has restricted research into investment behaviour for many researchers. The study data available was analysed using two methods, namely statistical analysis and data mining. The details of data source, its collection and structure are presented in chapter 3.2.

In general, there are two types of statistical analysis - descriptive and inferential statistics. Descriptive statistics are applied to analyse data by describing it in a meaningful way, such as in patterns emerging from the analysis; however, no conclusions beyond the analysed data and hypotheses can be drawn (Laerd Statistics, 2013). Inferential analysis uses small samples to make generalisations about the whole population we are interested in investigating, and thus is able to draw conclusions regarding statistical hypotheses (Laerd Statistics, 2013). Therefore, a combination of descriptive statistics and inferential statistics, such as factor and regression analysis, were applied to identify meaningful patterns, and then establish statistical differences and/or correlations. For example, key attributes influencing investment behaviours were filtered and classified using factor analysis and their impacts on investment preferences were studied using regression analysis. Statistical analysis was employed as the empirical validation approach for this study. A guide to the statistical analysis conducted are presented in chapter 3.3, while the results of the statistical analysis are provided in chapter four.

In order to turn the data into information and discover hidden patterns, data mining techniques were applied for knowledge discovery in this research. Based on the discussions in chapter 2.5.1, an integration of clustering analysis and association rules was applied. The former would help segment individual investors from mainland China and Hong Kong, and the latter would discover rules for each desired cluster. This approach enables financial institutions to get a better understanding of correlations between investors' characteristics and investment preferences with a significantly shortened processing time for rules mining. Financial institutions can then define more specific marketing strategies, such as formulating tailor-made financial investment portfolios for their customers based on the relationship rules extracted. In order for financial institutions to convert data into business knowledge in an effective and efficient manner, a Financial Data Mining Model (FDMM) is proposed. The most popular data mining techniques were described in chapter 2.5, while the development of the proposed model with an illustrative example, including the interpretation and evaluation results as well as the knowledge generated, is presented in chapter five. Further validation of the findings through a case study method are described in chapter six.

3.2 Data collection

3.2.1 Data source

A data set was obtained from Convoy Financial Service Limited, a subsidiary of Convoy Financial Holdings Limited (Stock code: 1019.HK), to explore the influence of the psychological, sociological and demographic factors, and to analyse how the major attributes identified under these constructs affect investors in both Hong Kong and mainland China. Established in 1993, and listed on Hong Kong Stock Exchange in 2010, Convoy grew rapidly from a small ten man office to a listed corporation with 1,600 professional financial consultants. Possessing over 20-year experience in financial services, Convoy is a good representative financial institution in Hong Kong and Convoy data should in practice represents the Hong Kong financial industry sufficiently.

Convoy Financial Service Limited is not only the largest listed independent financial institution in Hong Kong but also one of the Asia's leading financial service providers with financial planning, insurance, asset management, Mandatory Provident Fund (MPF), and money-leading business and operations in Hong Kong, Macau and China. Convoy provides a broad portfolio of financial solutions to its clients by offering over 1,000 diversified financial and insurance products. Its comprehensive financial investment product portfolio and dataset about Hong Kong and mainland Chinese investors made the data from this institution suitable for analysing and predicting investors' preferences. They are also a useful partner for applying and testing the methodology and findings over a longer term.

In recent years, the number of Convoy's customers from mainland China has increased considerably and thus the institution desires to learn about the investment behaviour of mainlanders, and understand the differences in investment preference between mainland Chinese and Hong Kong investors. For that reason the institution, in which the author was the Chief Executive Officer, supported this research by providing confidential data about its customers, allowing this research to overcome a major obstacle – a researchers' inability to access huge volumes of confidential financial transaction data – normally a major problem as highlighted in chapter 1.1.

3.2.2 Data sampling

Data covering the period 2002 and 2014 about customer characteristics and their financial transactions was obtained and compiled from Convoy. Such a period covered two major economic cycles, from 2002 to 2007 (short-term regional social crisis SARS occurred in Hong Kong and China in 2003) and from 2008 to 2014 (medium-term globally financial tsunami occurred in 2008 and 2009) which would be long enough to include various factors and situations to represent the normalized investors behaviours. To ensure the data purity, dataset that fulfilled due diligence and compliance requirements, including Anti-Money Laundering (AML) and Know Your Customers (KYC) regulations was considered valid and taken into account. Data from mainland Chinese and Hong Kong investors that was relating to the key attributes, including:

i) investment experience,

ii) gender,

iii) education level,

iv) age,

v) economic status/income,

vi) marital status, and

vii) education level,

identified in the literature review in chapter 2.4 and two additional attributes,

i) household net worth, and

ii) nature of employment.

These were chosen as inputs based on discussions by the expert panel sessions conducted. These parameters defined the dataset parameters used for this study in order to address the research gap question.

To gain more reliable insights and have representative results, a large sample size was adopted in the quantitative analysis for this research. As noted by Saunders et al. (2009), larger sample sizes can help produce more reliable results as the samples can be more representative of the whole population. Thus, data from 142,496 customers of Convoy were used to support this study, of which 87,057 were mainland Chinese investors and 55,439 were Hong Kong investors.

3.2.3 Expert Panels

In order to further discuss possible causations for the correlations found in the data as well by analysis, an expert panel was formed. The panel helped define and verify the chosen data set to be analysed and processed. They also evaluated and commented on veracity and possible causation mechanisms in the findings from the analysis. In essence the expert panels were formed to critically review the findings and comment on their reliability. The structure and operation of the expert panels are described below.

3.2.3.1 Structure and operation of expert panels

Fifty finance professionals, who possessed at least 15 years of experience, worked in Hong Kong or mainland Chinese financial institutions, and were proficient in different fields in the financial industry including private and retail banking, insurance, investment banking and stock brokerage, were invited by email randomly to participate in this study. Ten out fifty of the invited professionals indicated their willingness to join the expert panels for adding value to generalize the proposed models and the findings in this study.

The ten professionals joining the expert panels were knowledgeable and experienced in investment behaviours. The panel members were in senior positions in various financial institutions and dealt with numerous investors from Hong Kong and/or mainland China every day. Playing a senior managerial role in the institutions enabled these experts to have the big picture of customer profiles and to study their customer data and behaviours for strategic planning. Having sound understanding and knowledge of the strategies and internal practices in different companies, these experts can critically comment the methodology employed and the findings from the development of the investment preference models. Table 3.1 shows the background of experts involved in this study.

Table 3.1 List of experts

Group	Expert	Position, Company	Experience
1a	Mr. G	Executive Director, LGT Bank Hong Kong	20-year experienced private banker
	Mr. V	Fund Manager, China Everbright Limited	21-year experienced investment manager
	Mr. T	Branch Manager, DBS Bank Hong Kong	20-year experience in wealth management retail banking
1b	Mr. W	Chairman, Huagui Life Insurance Co. Ltd	18-year experience in wealth management and insurance
	Mr. C	General Manager, China Investment Securities Ltd	25-year experience in equity and debt market
	Mr. H	Senior Manager, China Merchant Bank	15-year experience in private banking
2	Mr. S	General Manager, Securities Brokerage KGI Asia Ltd.	25-year experience in investment brokerage in Hong Kong
	Mr. E	Senior VP, AXA Insurance HK Ltd.	22-year experience in pension and insurance business in Hong Kong
	Mr. L	GM, Private Banking, Mingsheng Bank	23-year experience in retail and private banking in China
	Mr. Z	Branch Manager, Bank of China	25-year experience in banking in China

In this study, two rounds of expert panel meetings were arranged. The format and content of the panel meetings were as follows:

The first round of expert panel meetings with the following arrangements were primarily to collect professionals' advice on possible causations for the correlations found in the data and the chosen data set to be analysed and processed.

i) Two face-to-face group meetings – one was conducted with three professionals working in Hong Kong's financial institutions (Group 1a indicated in Table 3.1) while another was with three professionals working in mainland Chinese financial institutions (Group 1b indicated in Table 3.1). By arranging face-to-face group meetings, communicative interactions were encouraged and stimulated to result in more in-depth discussions compared with survey and individual interview while explanations of responses could be easily probed for. However, cost is a major disadvantage for arranging face-to-face group meetings that it required more resources and consumed more time for meeting arrangement, particularly for professionals from mainland China to travel to Hong Kong for the meeting or vice versa. In the light of this disadvantage, two separate face-to-face group meetings for professionals from Hong Kong and mainland Chinese were arranged.

ii) Telephone interviews with four professionals (Group 2 indicated in Table 3.1), of which two from Hong Kong and two from mainland Chinese financial institutions were conducted, as it was challenging to assemble all the professionals to attend the face-to-face group meetings. The main advantage of conducting telephone interviews are that it reduced effort to assemble all the professionals from different locations – Hong Kong and mainland China. Another advantage is that telephone interviews allowed interviewees to feel less inhibited and provide honest answers and open responses without facing other people compared with panel meetings; this in fact overcame the possible disadvantage of face-to-face group meetings. The disadvantages of telephone interviews were that the body language could not be seen and it was very hard to make a judgement on how attentive the interviewees were.

The second round of expert panel meeting was to further clarify the findings of PSYC Model and FDMM as well as conclude the study. Teleconference with all the professionals except two (i.e. Mr. H and Mr. E) who were unavailable to participate, were conducted.

Semi-structured interviews were used in the two rounds of expert panel meetings to enhance the richness of data to be collected. The professionals discussed their awareness of possible models and studies about investment behaviours of mainland Chinese and Hong Kong investors, challenges faced by different financial institutions, the key factors influencing investment behaviours, results and findings of this study and the value of proposed model. Tables 3.2 and 3.3 demonstrated the questions and topics discussed in the two rounds of expert panel meetings. The relevant results of panel discussions were incorporated and highlighted throughout this report.

Table 3.2 Summary of the first round of expert panel meetings

Question/topic discussed	**Resulting point of view**	**Conclusion**
1. What is your customer mix in terms of Hong Kong, Mainland China and others?	China experts expressed that their customers were mainly mainland Chinese with Hong Kong customers. Most of their Hong Kong customers stayed in Shenzhen and some in Beijing. Hong Kong experts confirmed that mainland the number of mainland Chinese customers were increasing rapidly in the past ten years and they were the source of growth of customer base. Three experts further confirmed that they had over 50% of customers from mainland China while most the rest were from Hong Kong.	China experts may not be familiar with Hong Kong investors' behaviours and characteristics as they do not have much experience in dealing with them. Hong Kong experts are familiar with both Hong Kong and Chinese investors.
2. Have you read or been aware of any literature, articles, books or other documents about investment behaviour of mainland China and Hong Kong?	All experts generally confirmed that they paid attention to the topic of investment behaviours when they were doing newspaper reading, experience sharing with colleagues and training courses from the companies. They expressed that there was very limited academic study on investment behaviours which was practical and applicable for their companies and most of their knowledge in fact came from experience and practice.	There is a need for experts in the financial industry to acquire knowledge about investment behaviour of mainland China and Hong Kong. However, in the academic and business world, there is a lack of systematic, consistent and reliable approach to understand investors' behaviours.
3. How do your companies study investment behaviour of mainland Chinese and Hong Kong customers?	Both China and Hong Kong experts confirmed that their companies had not officially launched any project nor built any model to study investment behaviours of mainland Chinese and Hong Kong investors. All Hong Kong experts confirmed that they had team training and official internal courses about mainland Chinese investors' behaviours in relation to their respective businesses, such as credit appraisal, investment advising, anti-money laundering (AML) procedure, etc.	

Question/topic discussed	Resulting point of view	Conclusion
4. Have you found any useful models available for exploring and predicting investment behaviours?	All experts expressed that they desired to have such a model to help their businesses but they could not find any proficient tools and models.	There is a need of all-in-one model exploring and predicting investment behaviours.
5. What difficulties do your companies have regarding customer behaviour and investment needs?	All experts expressed that it was vague and difficult to transfer skills and knowledge to colleagues and subordinates in relation to customer behaviours and investment needs. Since experts gained skills and knowledge based on their experience and daily practice, there was no evidence and clear guidelines available for subordinates, particularly newly joined employees and those with less experience, to execute jobs. Employees had to spend time and effort to deal with customers in person in order to grow their skills and knowledge about customer behaviour and their investment needs. In addition, different people would have different views or approaches to handle customer behaviour issues. Their approaches were subjective based on individual personal experience and knowledge and this hindered the top management to create stable policies and guidelines.	There is a need of a systematic and reliable model or approach which should be developed from objective data and information to provide a stable environment for business development and knowledge sharing.
6. Are your companies willing to disclose internal and customer information for external research?	China experts expressed that it was difficult but may be possible. This greatly depended on the personal relationship with their top management. At that moment, the privacy law in China was not as stringent as Hong Kong but it was tightening continuously. Hong Kong experts opined that it would not be quite possible to ask their companies to disclose clients' information to external researchers due to the regulations of the Securities and Futures Commission (SFC) and Hong Kong Monetary Authority (HKMA) and the privacy laws in Hong Kong. Besides, their companies were seldom to cooperate with academic institutions in the past. Majority of experts expressed that their companies seemed to be not interested in dealing with academic institutions.	It is not easy for academic institutions to get support and real data from commercial institutions for their research.

Question/topic discussed	Resulting point of view	Conclusion
7. Do your companies have any models for exploring customer behaviour and predicting customer investment preferences?	Experts from banks expressed their banks implemented a customer relationship management (CRM) model to handle this but the model was a simple application. The model could only analyse client needs by their income level, age and/or other demographic factors. The model neglected psychological factors and was unable to predict investment behaviours. Experts from investment banks and brokerage houses indicated that their companies did not implement any systems to study customer behaviour and needs but they had a risk monitoring system for the on-going risk management purpose on clients' credit, margin and dealing behaviours. Experts from insurance companies expressed that they only had internal systems for their actuary and product development teams to understand well the companies' risk and customer profile.	Obviously, most of financial institutions do not have well-established model or system to analyse customer behaviour for their investment preferences prediction.
8. Six attributes including investment experience, age, gender, economic status/income, marital status and education level identified from the literature review were considered in this study. Do you support these factors? Any other factors influencing investment behaviour should also be considered?	All experts from China and Hong Kong agreed that the six attributes were common and worthy of study when exploring investment behaviours of mainland Chinese and Hong Kong investors. An expert from China suggested that household net worth and nature of employment might have impacts on investors' behaviour. These two factors were supported by the majority of experts from Hong Kong and China, as experts described that both factors could be relating to investors' economic status which should be a crucial factor affecting investment behaviour. One expert from Hong Kong suggested to consider the branding of financial institutions. However, it was not supported by other experts as they opined that the branding financial institutions might only affect clients' selection of financial institutions rather than investment products. Another expert from Hong Kong thought reading habits might affect investors' behaviour but other experts disagreed and claimed it was irrelevant and vague.	The six attributes identified from the literature and two additional attributes, household net worth and nature of employment supported by the experts are taken into account as the factors influencing investors from mainland China and Hong Kong.

Table 3.3 Summary of the second round of expert panel meeting

Question/topic discussed	Resulting point of view	Conclusion
1. Do you agree with the findings in chapter four? Are the results discussed in chapter four logical to you?	Experts from China and Hong Kong supported the results of factor analysis and confirmed that the classification of six selected attributes under demographic, psychological and sociological factors were realistic. Regarding the results of descriptive analysis, in general experts had no particular comments, except about the attributes – education level and age. • In terms of education level, mainland Chinese investors who had received poorer education, on average, held more fund units than those who had received higher education. Experts from China explained that higher education level allowed investors from mainland China to assess risk and thus investors often prefer investment products at lower risk level and diversify their investments to preserve capital. In case of Hong Kong investors, investors with different education levels did not show a significant difference in the quantity of fund unit held. Experts from China found this phenomenon interesting. Experts from Hong Kong explained that education level might not have a significant relationship with Hong Kong investors' decisions as Hong Kong investors were often open and confident about their investment decisions and made investment decisions based on their own experience. • In terms of age, when taken together, mainland Chinese and Hong Kong investors aged 55 and above held a smaller quantity of fund units. Experts from China and Hong Kong confirmed that investors aged 55 and above always take their retirement plan into consideration and instinctively invest for wealth preservation, therefore they will reduce the quantity of fund holdings.	The characteristics of and differences between mainland Chinese and Hong Kong investors described in chapter four were rational meaningful.

Question/topic discussed	Resulting point of view	Conclusion
2. Do you think it is necessary to develop the FDMM on top of the PSYC Model to generate specific rules for predicting customers' investment needs?	All experts supported that both FDMM and PSYC Model were unique and they added value to each other to fully describe and explain investment behaviours and predict investors' needs. An expert expressed that FDMM is necessary as "it is an action-oriented model which can exactly tell what actions the company should take to satisfy customers' investment needs". Two experts affirmed that the FDMM makes use of the big data from the financial institution and thus is practical that the results would be good for them to understand how to strategically deal with Hong Kong and mainland Chinese investors. Another expert stated that though it may not reflect the rationale behind the generated rules, PSYC is good to be the basis to rationalize the rules generated by the FDMM. Besides, an expert indicated that a centralised database was often critical for financial institutions to store and retrieve data for analysis, thus it was useful to develop the FDMM. He further confirmed that Data Selection and Pre-processing Module (DSPM) of the proposed FDMM would be very important to ensure the rules to be generated are relevant and significant and the data set used in this study for the model development and validation was reliable and typical. Another expert added that predicting customers' investment preferences was the key goal of financial institutions to improve their product and marketing strategies. The PSYC model allows financial institutions to understand the rationale of investment behaviour at the top level while the FDMM at the bottom level is designed to achieve the ultimate goal by predicting customers' investment preferences.	The FDMM and PSYC Model should be developed together for financial institutions to predict and rationalise investors' preference. Furthermore, DSPM is an important element to the reliability of the FDMM.

Question/topic discussed	Resulting point of view	Conclusion
3. Do you think the proposed PSYC Model can fairly reflect the characteristics of investors' behaviours in mainland China and Hong Kong and help the investment industry better understand their customers?	All experts agreed that age, income level and investment experience were the most important factors to account for their clients' investment behaviours. They indicated that they would consider these three factors in anticipating their clients' responses when facing investment exercise or advice. An expert indicated that "the PSYC Model summarising the common characteristics of mainland China and Hong Kong investors' behaviour in terms of risk appetite and personal involvement is unique. It provides financial institutions with directions to market their products". Another expert added that the PSYC Model further elaborated investment behaviours between mainland Chinese and Hong Kong investors by focusing on the three most important factors (age, income level and investment experience) could help financial institutions be more focused in their business consideration. Regarding the findings - Hong Kong investors have mixed behaviours, two experts from China and three experts from Hong Kong jointly commented that this might be due to the weaker crowd effect in the community of Hong Kong investors, compared with the mainland Chinese investors' community. An expert from China said this was also the reason why mainland Chinese investors have a clearer pattern of behaviour suggested by the PSYC Model. Experts from China and Hong Kong confirmed that the findings of PSYC Model about mainland Chinese investors' clear behaviour pattern were so realistic – those at/with low-age/high-income/less-experience would more prefer having professionals to help their investment and pursue higher risk investment for higher return while those are at high-age/low-income/more-experience would be happier to do investment by themselves for less risk investment. An expert added that mainland Chinese investors were less mature than Hong Kong investors and had less experience in dealing with different financial products and markets as there were less investment choices in the mainland Chinese market. This induced mainland Chinese investors to behave in a more straight-forward way, either SC or PY, as suggested by the PSYC Model, while	Experts supported the development of the PSYC Model which summarizes the investment behaviours between mainland Chinese and Hong Kong investors. Experts also agreed that the findings of the PSYC Model are logic and realistic, which provide directional information to financial institutions for business development.

Hong Kong investors behave in a more diversified way in terms of PY, PC, SC and SY. Other experts supported these findings as well.

For the factor of age, mainland Chinese and Hong Kong investors behave oppositely, i.e. SC/PY and PY/SC respectively. Experts opined that entrepreneurship was popular in China and youngsters would start their own business even at secondary school or university. Thus, younger mainland Chinese investors may focus more on their own business development and prefer more to leave their investment matters to professionals, and be more aggressive in investment return (i.e. PY). For younger Hong Kong investors, they grow up in a wealthier society, compared with mainland Chinese investors. They would look for a more stable environment which can be reflected from a lot of graduates applying for government positions. They would behave SC.

For factor of income, mainland Chinese and Hong Kong investors behave the same in the perspective of personal involvement and opposite in the perspective of risk appetite, i.e. PY/SC and PC/SY respectively. Experts commented that it was common as high income investors regardless from either mainland China or Hong Kong would be busy and reply on professionals to invest. For risk appetite, China is a developing country while Hong Kong is a developed district. So investors with higher income from mainland China would pursue opportunities with high risk appetite while the lower income investors would worry about the social protection issues and behave more conservatively. This situation is opposite in Hong Kong.

For factor of investment experience, mainland Chinese and Hong Kong investors behave the same, i.e. SC/PY. All experts agreed that most investment veterans would be confident in investment, therefore they would spend more time and involve more in handling investments directly themselves. In the meantime, they would realize importance of capital preservation and behave more conservative. Investors with less investment experience would more rely on professionals and be more aggressive. Experts from both China and Hong Kong confirmed that such behaviours would be the same no matter the investors were from Hong Kong or mainland China.

Question/topic discussed	Resulting point of view	Conclusion
4. Do you think it is necessary to develop the FDMM on top of the PSYC Model to generate specific rules for predicting customers' investment needs?	All experts supported that both FDMM and PSYC Model were unique and they added value to each other to fully describe and explain investment behaviours and predict investors' needs. An expert expressed that FDMM is necessary as "it is an action-oriented model which can exactly tell what actions the company should take to satisfy customers' investment needs". Two experts affirmed that the FDMM makes use of the big data from the financial institution and thus is practical that the results would be good for them to understand how to strategically deal with Hong Kong and mainland Chinese investors. Another expert stated that though it may not reflect the rationale behind the generated rules, PSYC is good to be the basis to rationalize the rules generated by the FDMM. Besides, an expert indicated that a centralised database was often critical for financial institutions to store and retrieve data for analysis, thus it was useful to develop the FDMM. He further confirmed that Data Selection and Pre-processing Module (DSPM) of the proposed FDMM would be very important to ensure the rules to be generated are relevant and significant and the data set used in this study for the model development and validation was reliable and typical. Another expert added that predicting customers' investment preferences was the key goal of financial institutions to improve their product and marketing strategies. The PSYC model allows financial institutions to understand the rationale of investment behaviour at the top level while the FDMM at the bottom level is designed to achieve the ultimate goal by predicting customers' investment preferences.	The FDMM and PSYC Model should be developed together for financial institutions to predict and rationalise investors' preference. Furthermore, DSPM is an important element to the reliability of the FDMM.

Question/topic discussed	Resulting point of view	Conclusion
5. A case study was conducted using the FDMM and eight rules were generated by the FDMM. Do the rules make sense to you?	An expert confirmed that "the eight rules are sensible and provide a good direction for sales to have a clear instruction to approach customers". Another expert added that such rules could be applied to validate the developed PSYC Model in return. Experts from Hong Kong further confirmed that the rules are applicable. The resulting discussions are as follows: • Six out of the eight rules indicated that the group of investors studied would usually choose technology stock as their choice of investment (i.e. rules 2, 4-8). An expert explained that given that this group of Hong Kong investors were relatively younger, had lower income and less investment experience, they had relatively less capital for investment and less averse for risk and thus preferred technology stock which is high growth to have greater return with smaller capital. Another expert added that "young investors in Hong Kong would like to have more involvement in their investment activities and stock investment could in fact offer them more personal involvement". • An expert supported that the characteristics of these group of investors – aimed for high return with small capital – explained why they would usually invest in global emerging markets equity fund (i.e. rules 1 and 3); the resulting rules generated by FDMM were consistent. Another expert added that owing to a lack of knowledge in investing in emerging markets' stocks, these investors would invest through a fund in bid to pursue a high-risk-high-return preference with small capital. • An expert mentioned that "these specific rules generated by FDMM are truly effective and practical for business practitioners." • An expert from Hong Kong stressed that neither the FDMM nor PSYC model would be valid without the available and completed data set provided by the financial institution. He appreciated the contribution of this study and said the FDMM would be the only available and reliable model to date for them to have a clear action plan to deal with customers.	Experts supported and confirmed that the FDMM is valid and the rules generated are reasonable according to their experience and know-how about the market.

Question/topic discussed	Resulting point of view	Conclusion
6. Other comments on this study	Experts had the following comments: • It was not easy to understand the rationale of the rules generated by FDMM. • Product definition system could be developed to enhance the practical usage of the PSYC Model by matching corresponding products. • They were not aware similar models in the market. This study would be the first step for establishing a complete practical model for business use. • Completeness of data would be a great challenge for this kind of research as the commercial world may not be keen to support. This study with the support of a large reasonable data set from a renowned typical financial institution made it exceptional and valuable. • This study could provide a convincing and scientific way to summarize the key factors affecting investment behaviour in terms of demographic, psychological and sociological factors.	This study was meaningful and useful to financial institutions and further study to enhance its practicality was suggested.
7. Changes of financial markets in recent years and resulting challenges faced by financial institutions in relation to investment behaviour	In China, financial markets undergone numerous crisis in the past five years, including A shares collapse, shadow banking, internet finance, de-leveraging, etc. The financial institutions in China were facing below challenges which would relate to investor behaviours: • Yield and investors' return expectation mismatching; • Short term risk aversion from investors due to frequent crisis; • Credibility of investors getting worse; and • Rapid and frequent regulation change. • In Hong Kong, financial markets had been increasingly correlated to the mainland Chinese markets since its sovereignty return. Financial institutions in Hong Kong were facing below challenges which would relate to investor behaviours: • Influx of Chinese investors has changed the investors behaviours landscape; • Shortage of relationship managers who well understand new investors' behaviours; and • Rapid and frequent regulation changes.	Market changes and regulation changes would have immediate impacts on investors' behaviours. Yet, such short-term impacts could not be reflected in the Models. Experts commented that it is difficult to consider and quantify the impact of ad hoc factors but appreciated if further study would be carried out.

3.3 Statistical methods

The collected data were processed and analysed using SPSS software and a combination of statistical methods, including descriptive analysis, factor analysis, correlation analysis and regression analysis, to achieve the research objectives. The reasons to use these statistical methods are as follows.

3.3.1 Descriptive analysis

Descriptive analysis, such as means and variances, was used to describe individual investors' personal information. To better understand the differences in financial investment behaviour between mainland Chinese and Hong Kong investors, descriptive analysis, which describes the basic features of the high volumes of data, can provide summaries about the given data set collected in this study and was thus used to evaluate differences in the quantity of fund unit held between mainland Chinese and Hong Kong investors in terms of variables such as age, gender, education level, marital and economic status.

In this research, economic status was expressed in terms of annual income and household's net worth, and converted to Hong Kong Dollar (HKD) while the education levels of investors were categorised by elementary (primary school or below), junior (secondary school in Hong Kong or high school in mainland China), and senior and above (Associate Degree/High Diploma/Bachelor's Degree and equivalent or above), for comparison purposes. The details of the descriptive analysis are elaborated in chapter 4.1.

3.3.2 Factor analysis

Factor analysis is always a common practice for factor filtering identified in the literature review (Hayton *et al.,* 2004; Sakar *et al.,* 2011). It is useful to test and reduce the number of explanatory variables, and at the same time to provide a validated data construct for further analysis. Factor analysis in this research statistically evaluated the major attributes influencing investor behaviours and classified the major attributes into three constructs – psychological, demographic and sociological factors, followed by validating the major attributes explaining investment behaviour. The details of the procedure and results of the factor analysis are discussed in chapter 4.2.

3.3.3 Correlation analysis

Correlation analysis was used to determine if there is a statistically significant association between variables and to study the strength of the relationship. In this research, only behavioural variables that remained and were found valid after the factor analysis were considered for correlation analysis.

i) The associations between the variables of mainland Chinese and Hong Kong investors and the quantity of fund unit held,

ii) The associations between the variables of mainland Chinese and Hong Kong investors and their choice of country-specific financial investment options.

iii) The strength of the associations among the behavioural variables.

The details of the correlation analysis are discussed in chapter 4.3.

3.3.4 Regression analysis

Linear regression analysis was applied to identify relationships among variables. In order to predict the investment behaviour of mainland Chinese and Hong Kong investors, the quantity of fund unit held and choice of country-specific financial investment options were used as dependent variables, while a set of variables verified by factor analysis were used as independent variables for regression analysis. The details of the regression analysis are discussed in chapter 4.4.

3.4 Data mining

Based on the comparison of the strengths and weaknesses of the four commonly used data mining techniques shown in Table 2.1 in chapter 2.5, the author found that

i) this study does not require fuzzy logic to deal with imprecise or vague qualitative data;

ii) this study involved the manipulation of large volumes of data that using neural network for such a data analysis would be a time-consuming process but association rules would be effective in handling large database and generate rules about investors' behaviour while clustering analysis would be helpful in retrieving information and recognising similarity and pattern of investors' behaviour; and

iii) clustering analysis and ARs were appropriate and feasible to be applied in this study.

3.4.1 Clustering analysis

Clustering analysis was conducted to segment individual investors from both mainland Chinese and Hong Kong investors into clusters based on their characteristics, including age, annual income (converted to HKD), gender, marital status, education level, and economic status (converted to HKD). The rationale of these attributes chosen was discussed in chapter 2.3 and chapter 2.4.

As explained in chapter 2.6.1, K-means algorithm was selected for clustering analysis in this research. According to Wu and Kumar (2009), K-means algorithm is simple, easy to be understood, and easily adjusted to deal with different scenarios. Similarly, So and Yoon (2008) pointed out that some of clustering algorithms generated too many clusters, while the K-means algorithm is a relatively practical clustering algorithm to derive a manageable set of clusters.

The K-means algorithm has also been proven to be suitable for the financial sector and has been increasingly applied in the financial area (Batra *et al.,* 2012; Kašćelan *et al.,* 2014; Kuo *et al.,* 2007; Nanda *et al.,* 2010). It can help financial institutions effectively manage their portfolio and asset selection. According to Nanda *et al.* (2010), K-means algorithm can group similar categories into a cluster, so as to select the best performing stock from those groups for building a better financial investment portfolio. Financial institutions can formulate their portfolio based on the characteristics of the clusters. It results in optimising the investor returns and minimising portfolio risk. Thus, applying the K-means algorithm was deemed suitable for this research.

3.4.2 Association rules (ARs)

ARs were applied on the clusters segmented by clustering analysis, so as to discover the interesting patterns of financial investment products purchased by each of the specific, targeted cluster, i.e. investment preferences. Thus, the rules extracted can effectively provide a basis for forecasting financial investment preferences of clusters and their decision making. The Apriori Algorithm proposed by Agrawal and Srikant (1994) is a classic algorithm for effectively generating association rules between items in large databases (Kuo *et al.,* 2011; Chung and Tseng, 2012; Lim *et al.,* 2012). It can help financial institutions find the frequent item-sets in a database in this study.

3.5 Design of a Financial Data Mining Model (FDMM)

Although prior research studied the variables influencing financial investment behaviour (Charles and Kasilingam, 2013; Kaustia & and Knüpfer, 2007, 2008; Feng & and Seasholes, 2005; Malmendier and Nagel, 2011; Manish and Vyas, 2008; Seru *et al.,* 2010), the research seldom examined how and to what extent these variables can help predict the financial investment preference of mainland Chinese and Hong Kong investors, as discussed in chapter 2.

With the enormous growth of stored data, better data management seems to help business organisations, especially in the financial sector, to sustain their businesses. Financial institutions analyse the large amount of data in order to maximise their customers' return on investment, but in addition they can also gain a better understanding of financial investment behaviours of mainland Chinese and Hong Kong investors regarding what their customers need. Manipulating and analysing such a large volume of data would be time-consuming and complicated. Therefore, it is essential to design special tools for financial institutions to find the required information from huge amounts of data more efficiently. If one can understand how to convert raw data into useful knowledge, this knowledge will lead them to have better understanding of investors' behaviour and investment preferences.

The FDMM was therefore designed to fill the research gap and help financial institutions in converting large amounts of data into actionable intelligence of business value in an effective and efficient manner. The FDMM model must enable financial institutions to devise the most appropriate financial investment strategies and design financial investment products/portfolios for their customers. The FDMM model uses the results obtained from the statistical analysis of the core data set, such as the major attributes identified influencing financial investment behaviour, as data input for further analysis. The design and architecture of the proposed model in addition to a realistic case study with the FDMM model are presented in chapter five.

3.6 Case study

To support and validate the proposed model, including the results generated from factor analysis, a case study was conducted using data collected from Convoy Financial Service Limited, and by applying the FDMM model to Convoy Financial services. With reference to prior research (Al-Hassan *et al.,* 2013; Geng, *et al.,* 2015), it is a common practice for researchers to use case studies to illustrate the effectiveness and feasibility of their approaches to financial data mining. According to Yin (2009), case studies can help apply the meaningful characteristics of situations in the real world. Therefore, it is appropriate to conduct a case study to validate the feasibility and test the user-friendliness of the FDMM model in this study.

Convoy is a market leader but also very representative of Hong Kong investment businesses. The case company's operations, such as the existing workflow and business strategy, were studied, so as to identify the challenges it faced. There follows, in chapter

3.6.1 is a brief introduction to Convoy Financial services but the details of the case study are provided in chapter five.

3.6.1 Company background

Convoy Financial Services Limited (“Convoy”) was founded in 1993. It is wholly owned by Convoy Financial Services Holdings Limited (CFS), and is an independent insurance and Mandatory Provident Fund (MPF) schemes brokerage firm in Hong Kong. Convoy provides a wide range of financial services and financial investment products, such as financial planning, insurance, asset management and MPF. To provide the customised financial investment products and services, Convoy insists on communicating with its customers and business partners. Convoy also offers a variety of tailor-made financial services and products for its clients. In the future, Convoy would like to expand its financial investment product offerings beyond Hong Kong.

3.6.2 Existing workflow of the case company

In the existing workflow process of Convoy, customer enquiries are collected from different sources, including the online customer zone, walk-in customers and the sales department. Convoy has a CRM system for its divisions to store and retrieve data. Each division of Convoy, namely Financial Services, Asset Management and Investment Services as shown in Figure 3.2, only stores its desired data in its own local database. There is no centralised database in Convoy and there is a lack of information sharing between divisions of the company. Customer service representatives will handle basic customer enquiries before transferring the query to financial consultants. Financial consultants will then prepare relevant information based on the enquiry information collected by the customer service representatives, and meet with the customers to understand their actual needs. Often the prepared information is not always useful, as the customers' needs can be different from their original enquiry. Financial consultants thus need to communicate carefully with the customers to gain a better understanding of their needs, before formulating a tailor-made investment plan. When the customers are satisfied with the investment plan, financial consultants create a customer profile. This profile is stored on the company's database and information concerning customers' options is forwarded to the product provider(s). If the customers are dissatisfied with the investment plan, financial consultants will follow-up each case. Financial consultants will review the performance of the investment portfolio periodically with the customers and give appropriate advice. They may also need to request customer information from product providers for periodic review and planning. The existing workflow of Convoy, together with the challenges in the existing workflow, is shown in Figure 3.3.

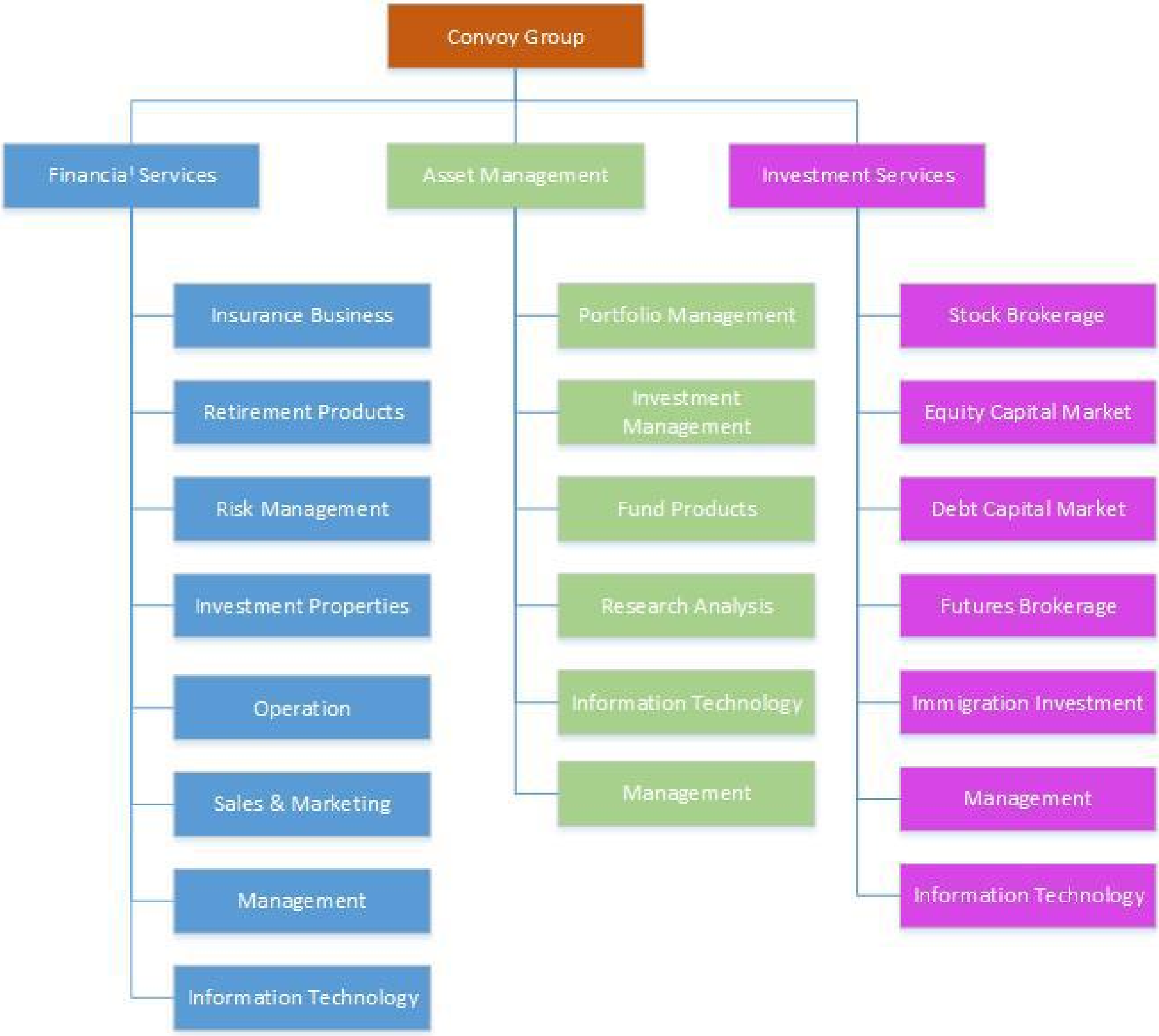


Figure 3.2 Organization chart of Convoy Group

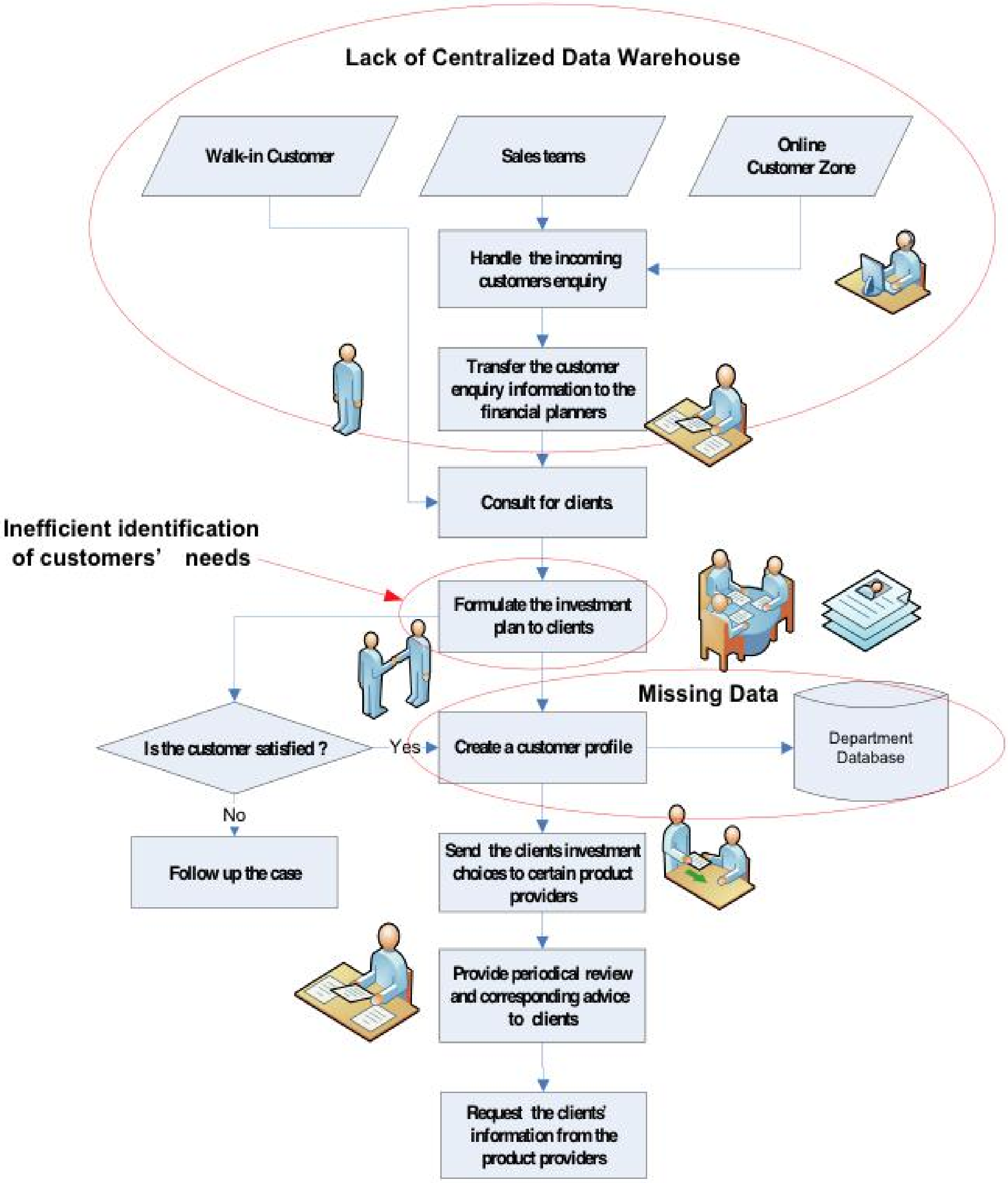

Lack of Centralized Data Warehouse
Walk-in Customer
Sales teams
Online Customer Zone
Handle the incoming customers enquiry
Transfer the customer enquiry information to the financial planners
Consult for clients
Inefficient identification of customers' needs
Formulate the investment plan to clients
Missing Data
Is the customer satisfied ?
Yes
No
Create a customer profile
Department Database
Follow up the case
Send the clients investment choices to certain product providers
Provide periodical review and corresponding advice to clients
Request the clients' information from the product providers

Figure 3.3 Existing workflow of case company

3.6.3 Challenges faced by the Convoy Financial Services

As shown in Figure 3.3, there are several challenges identified by a group of investment advisors from Convoy and senior management from its three main divisions as shown in Figure 3.2, hidden in the existing workflow, faced by Convoy. These challenges were also discussed in the expert panels conducted for this study (see chapter 3.2.3) and experts from other financial institutions opined that they were always spending effort to cope with such similar situations but no breakthrough solutions have been applied. Experts also complained that their respective customer relationship management systems were not smart and powerful enough to tackle those challenges

1. Lack of a centralised data warehouse

The first challenge is the lack of a centralised data warehouse within the case company. Each department only stores its desired data in its own database. Although each department has its own database, it cannot capture and link all the relevant information to other departments. Thus, each departmental database is isolated and it is difficult to share information. For instance, the sales department needs to send emails to request specific information from the customer services department. The lack of a centralised data warehouse not only results in difficulties regarding information sharing, but also increases the handling time of customer enquiries.

2. Inefficient identification of customer needs

Convoy also faces the challenge of inefficient identification of customer needs. In the existing workflow, Convoy stores the customer data in the database without any prior data mining or data handling. The database is only about the storage of data. Customers are not segmented into appropriate clusters to better understand their needs.

In addition, financial consultants make recommendations on the investment portfolio only based on the perceived customer needs collected during the time-consuming consultation or their experience. Financial consultants cannot identify or evaluate hidden patterns and relationships in the stored data.

Facing with the overwhelming amounts of data available from different sources, it is becoming increasingly difficult and time-consuming for Convoy to analyse the stored data without any assistance from data mining tools. Thus, often the identification of customer needs is inefficient in the existing workflow and this can result in a low level of customer satisfaction.

3. <u>Unsystematic identification of investment preferences</u>

In the existing workflow, financial consultants formulate the investment plan mainly based on the perceived customer needs during consultation and their experience. The identification of customers' investment preferences therefore relies heavily on the subjective judgement of financial consultants. The subjective identification of investment preferences may cause harm to the satisfaction of customers, who may then choose not to return, and thus damage the image of Convoy. The situation is worse when there is a wide range of investment products available, or when financial consultants are lacking in investment experience or distracted from a clear rational analysis.

For Convoy to increase the return on investment (ROI) for its customers and to maintain its professional reputation, it is desirable for Convoy to develop a systematic and effective approach for identifying customers' investment preferences based on more objective data regarding customer characteristics. Such a methodology would also benefit other financial institutions interested in a better understanding of their customers.

Chapter 4 Factors influencing investment behaviour

This chapter discusses the detailed steps taken to analyse how the major attributes identified in chapters 2.3 and 2.4 influence the behaviour of Hong Kong and mainland investors. Firstly, the results from descriptive analysis of the differences between mainland Chinese and Hong Kong investors are presented. Then factor analysis was conducted to validate and classify the key attributes influencing investors' behaviour. The correlations for the key attributes extracted and the investment preferences of both mainland Chinese and Hong Kong investors are further examined. Finally, this chapter discusses the regression analysis that assessed the relationships between the attributes and the investment preferences of mainland Chinese and Hong Kong investors. Recommendations about how financial institutions could maximise their business opportunities are made to conclude this chapter.

4.1 Descriptive analysis of Mainland Chinese and Hong Kong investors

To help categorise mainland Chinese and Hong Kong investment behaviour, descriptive analysis was adopted to segment the characteristics of individual investors from mainland China and Hong Kong in terms of age, education level, gender, household's net worth and marital status.

From the data on 142,496 Convoy clients, the characteristics of individual investors and their investment behaviours regarding fund holdings held by mainland Chinese and Hong Kong investors were analysed. The results from the analysis process are shown in Table 4.1, and are summarised as follows:

i) Gender: The bulk of individual investors (66%) from mainland China were women and on average they held more fund units than all men. This probably reflects the phenomenon of “big mother” investors – middle-aged women in mainland China who are active and dominant in the Hong Kong financial market. Hong Kong, unlike mainland China, does not have such significant gender domination for its own citizens when it comes to investments; the ratio of female and male investors in Hong Kong is 44:56.

ii) Education level: Almost 90% of mainland Chinese investors reported that they had attained senior or above education while only about half of Hong Kong investors reported that they had attained the same level. Mainland Chinese investors who had received poorer education, on average, held more fund units than those who had received higher education, while Hong Kong investors with different education levels did not show a significant difference in the quantity of fund unit held. This appears that investors from mainland China with a higher education level prefer more conservative investments or higher educated investors have a more diversified portfolio and use other financial service providers as well to reduce risk. As discussed in the expert panel with details provided in chapter 3.2.3, investors from Hong Kong are being open and confident about their investment decisions, and their education level has no significant impact to influence their investment behaviour.

iii) Age: Mainland China investors aged 50-54 held the largest quantity of fund units, while Hong Kong investors with the largest quantity of fund units tended to be younger, aged 45-49. Overall, investors from mainland China owned almost double the quantity of funds compared to those from Hong Kong. When taken together, mainland Chinese and Hong Kong investors aged 55 and above held a smaller quantity of fund units. The professionals in the expert panel detailed in chapter 3.2.3 rationalised that investors aged 55 and above always take their retirement plan into consideration and instinctively invest for wealth preservation, therefore they will reduce the quantity of fund holdings.

iv) Marital status: In Hong Kong, married investors (26.5%) purchased the largest quantity of fund units (mean=51.09), which is almost double the quantity compared to units held (mean=29.08) by single investors (71.4%). This may be because married investors have a higher household net worth and are hence willing and able to invest more. Similar phenomenon was found in mainland China that the quantity of fund units (mean=121.76) held by married mainland Chinese investors was obviously much more than that (mean=29.81) held by single investors.

v) Household net worth: The number of Hong Kong investors was inversely proportional to their household net worth; 68.1% of investors had a household net worth of less than HKD100K while 1% investors had a household net worth of HKD1M or above. Those investors with higher household net worth held a larger quantity of fund units on average, for instance, those investors with a household net worth of HKD1M and above held fund units at mean=555.36 while those with less than HKD100K held fund units at mean=24.11. Compared with the number of Hong Kong investors with net worth of HKD1M (1%), there was a significantly larger portion of mainland Chinese investors with a household net worth over HKD1M; they accounted for 17.1% of the total. Those mainland Chinese investors with a household net worth between HKD500,000 and HKD1M held the largest quantity of fund units on average (mean=110.23).

Table 4.1 Demographic characteristics of mainland Chinese and Hong Kong investors

		Mainland China			Hong Kong		
		Frequency	**Percent**	**Quantity of fund unit held**	**Frequency**	**Percent**	**Quantity of fund unit held**
				Mean			**Mean**
Gender	Female	57294	65.8	90.61	24337	43.9	27.71
	Male	29763	34.2	79.29	31102	56.1	40.07
	Total	87057	100	86.74	55439	100	34.64
Marital status	Divorced	1736	2	320.20	1156	2.1	15.82
	Married	48423	55.6	121.76	14703	26.5	51.09
	Single	36898	42.4	29.81	39580	71.4	29.08
	Total	87057	100	86.74	55439	100	34.64
Age	0-24	3991	4.6	20.96	4564	8.2	16.74
	25-29	24607	28.3	31.96	19558	35.3	23.04
	30-34	20773	23.9	47.04	14085	25.4	31.00
	35-39	15977	18.4	73.83	8144	14.7	38.87
	40-44	10148	11.7	126.30	4340	7.8	55.78
	45-49	6420	7.4	220.76	1886	3.4	158.18
	50-54	3651	4.2	460.34	2467	4.4	30.64
	55 and over	1490	1.7	97.32	395	0.7	61.51
	Total	87057	100	86.74	55439	100	34.64

		Mainland China			Hong Kong		
		Frequency	Percent	Quantity of fund unit held	Frequency	Percent	Quantity of fund unit held
				Mean			Mean
Educational Level	Elementary	191	0.2	657.67	442	0.8	40.64
	Junior	9763	11.2	144.14	25639	46.2	27.36
	Senior and above	77103	88.6	78.06	29358	53	40.91
	Total	87057	100	86.74	55439	100	34.64
Household's net worth	< HKD100K	25091	28.8	27.31	37751	68.1	24.11
	HKD100K-300K	33214	38.2	35.57	12508	22.6	41.66
	HKD300K-500K	8019	9.2	58.43	4150	7.5	38
	HKD500,000-1M	5813	6.7	110.23	491	0.9	66.01
	≥HKD1M	14920	17.1	86.74	539	1	555.36
	Total	87057	100	306.67	55439	100	34.641713

4.2 Using factor analysis for construct validation

The descriptive analysis shows that mainland Chinese and Hong Kong investors of different ages, education levels, genders, household net worth and marital status showed significant differences in investment preferences. To further study how these factors relate to and how they affect investors' behaviour, factor analysis was applied and is described.

In the literature review in chapter 2.3, a range of psychological, sociological and demographic factors influencing investor behaviour was identified, such as:

i) psychological factors: investment experience (Feng and Seasholes, 2005; Kaustia and Knüpfer, 2008; Malmendier and Nagel, 2011; Seru *et al.,* 2010);

ii) sociological factors: marital status (Iqbal, 2011), education (Reitan and Sorheim, 2010), and economic status/income (Lewellen *et al.,* 1977; Manish and Vyas, 2008); and

iii) demographic factors: age (Charles and Kasilingam, 2013; Lewellen *et al.,* 1977), and gender (Fellner and Maciejovsky, 2007; Weber *et al.,* 2002; Suleyman Gokhan Gunay, 2011).

The factors identified by the literature review and supported/suggested by the Convoy data in Table 2.1 were considered for the factor analysis. These factors include:

i) age

ii) gender

iii) investment experience

iv) education level

v) household net worth

vi) marital status

vii) annual income

viii) nature of employment

These eight factors were discussed by the expert panels (see chapter 3.2.3) and all experts agreed that they are rational to be considered according to their experience and professional judgement. Branding of financial institutions and reading habits were discussed by the expert panel but finally rejected by them. Experts opined that the branding of financial institutions might only affect clients' selection of financial institutions rather than investment products, while reading habits were irrelevant and vague as insufficient data was available to show any correlation with it.

The procedure to conduct the factor analysis is summarised with reference to these studies (Haitovsky, 1969; Field, 2000, 2005; Yong and Pearce, 2013; Cattell, 1978) and shown in Figure 4.1.

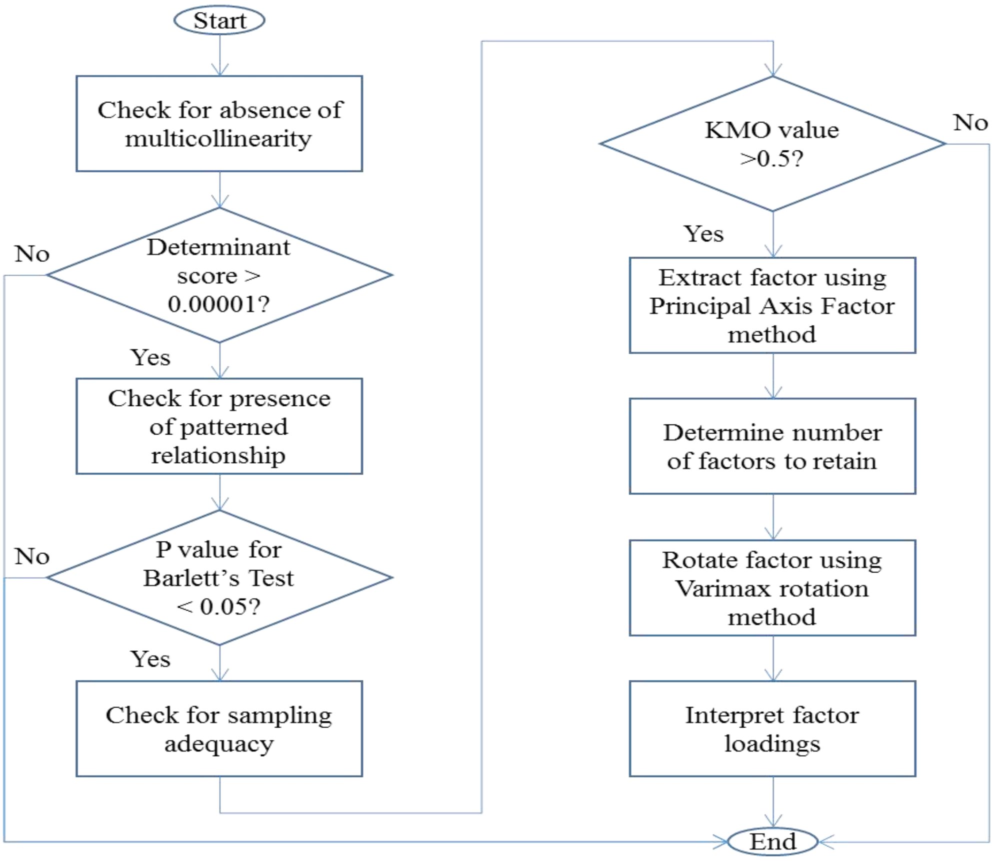

Start
Check for absence of multicollinearity
Determinant score > 0.00001?
No
Yes
Check for presence of patterned relationship
P value for Barlett's Test < 0.05?
No
Yes
Check for sampling adequacy
KMO value >0.5?
No
Yes
Extract factor using Principal Axis Factor method
Determine number of factors to retain
Rotate factor using Varimax rotation method
Interpret factor loadings
End

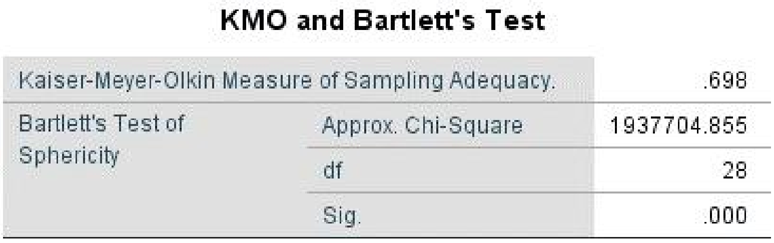
Figure 4.1 Flowchart for factor analysis

To determine if the data collected from Convoy were suitable for factor analysis, absence of multicollinearity, presence of patterned relationship amongst variables and sampling adequacy were tested, assuming the variables are ordinal and linear.

Multicollinearity refers to a phenomenon that two or more variables are highly correlated. Absence of multicollinearity is a crucial requirement to be met for factor analysis (Field, 2000). Haitovsky's test (1969) is the method for factor analysis to check if there is a problem of multicollinearity. It shows whether the determinant score is significantly different from zero, this indicates an absence of multicollinearity. As a rule of thumb, a value of a determinant score greater than 0.00001 indicates an absence of multicollinearity (Field, 2000). For the data collected from Convoy, a determinant score of 0.567 was found, which is greater than the necessary cut-off value of 0.00001, indicating an absence of multicollinearity.

To ensure the validity of variables, Kaiser-Meyer-Olkin Measure (KMO) and Bartlett's Test of Sphericity were performed. The presence of patterned relationships amongst variables was checked by Bartlett's Test of Sphericity (Field, 2000, 2005) while sampling adequacy was checked by the KMO of Sampling Adequacy (Field, 2000, 2005). A value of KMO greater than 0.5 indicates sufficient items for each factor.

The results of the KMO and Bartlett's tests are shown in Figure 4.2, the resulting KMO value of 0.698 exceeded 0.50 and thus confirms that sampling adequacy was acceptable. The Bartlett's Test of Sphericity with a level of significance at $p < 0.01$ confirms that patterned relationships amongst the variables exist and provides a reasonable basis for factor analysis in this case.



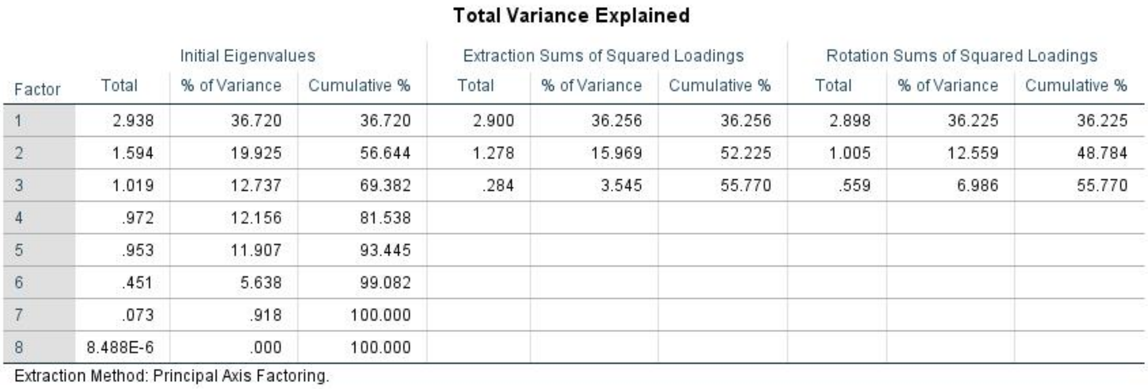
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.698
Bartlett's Test of Sphericity	Approx. Chi-Square	1937704.855
	df	28
	Sig.	.000

Figure 4.2 SPSS output for KMO and Bartlett's Test

These tests fulfilled the requirements for factor analysis. Then factor extraction was conducted using the principal axis factor method and Varimax rotation, which were parts of the factor analysis. These can help extract factors successively to attain an optimal structure with each variable loading on as few factors as possible, while maximizing factor loadings on each variable (Yong and Pearce, 2013). As indicated by Kaiser and Rice (1974), factors with eigenvalues (a measure of explained variance) greater than 1 should be retained, this is a common criterion for a factor to be useful. However, some researchers argue that Kaiser’s criterion may result in overestimation of the number of factors extracted (Costello and Osborne, 2005; Field, 2009). Yong and Pearce (2013) therefore suggested using the scree test method considering both eigenvalues and factor numbers (Cattell, 1978) in conjunction with the Kaiser’s criterion to determine the number of factors to retain. These criteria were thus employed for determining the number of factors to retain in this case.

An initial analysis was conducted to obtain the eigenvalues for each component and the results are shown in Figure 4.3. Three factors had eigenvalues greater than Kaiser’s criterion of 1 and together they explained 55.77% of the variance. To confirm that these three factors should be retained, a scree plot graphing the eigenvalue against the factor number was conducted (see Figure 4.4). This indicated that three factors should be extracted, as their data points exceeded eigenvalues of 1. From the third factor on, the eigenvalues are less than 1, meaning that each successive factor was accounting for smaller and smaller amounts of the total variance.



Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.938	36.720	36.720	2.900	36.256	36.256	2.898	36.225	36.225
2	1.594	19.925	56.644	1.278	15.969	52.225	1.005	12.559	48.784
3	1.019	12.737	69.382	.284	3.545	55.770	.559	6.986	55.770
4	.972	12.156	81.538						
5	.953	11.907	93.445						
6	.451	5.638	99.082						
7	.073	.918	100.000						
8	8.488E-6	.000	100.000						

Extraction Method: Principal Axis Factoring.

Figure 4.3 SPSS output for the total variance explained (extraction method: principal axis factoring)

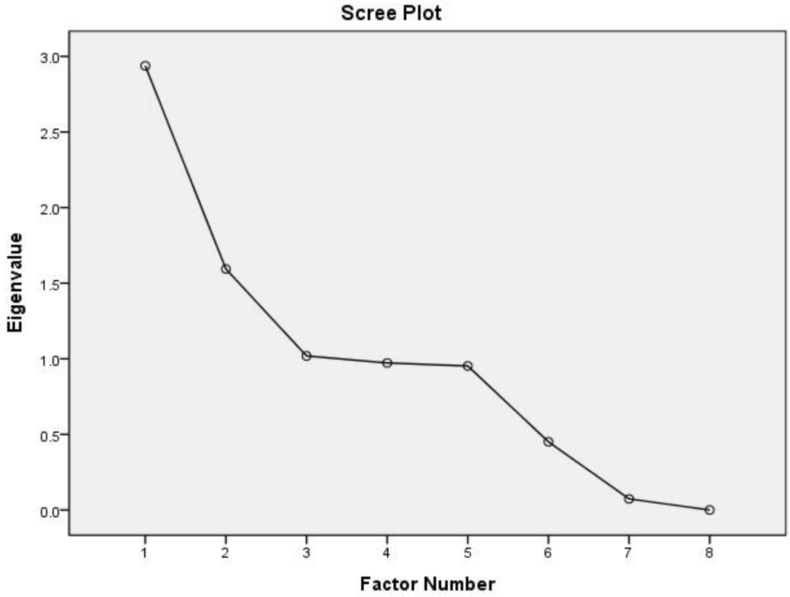


Figure 4.4 SPSS output for scree plot

According to a rule of thumb, the model is considered a good fit if there are less than 50% of non-redundant residuals with absolute values greater than 0.05 (Yong and Pearce, 2013). As indicated by the summary of non-redundant residuals shown in Figure 4.5, there were

3.0% non-redundant residuals with absolute values greater than 0.05, thus, the model was deemed a good fit.

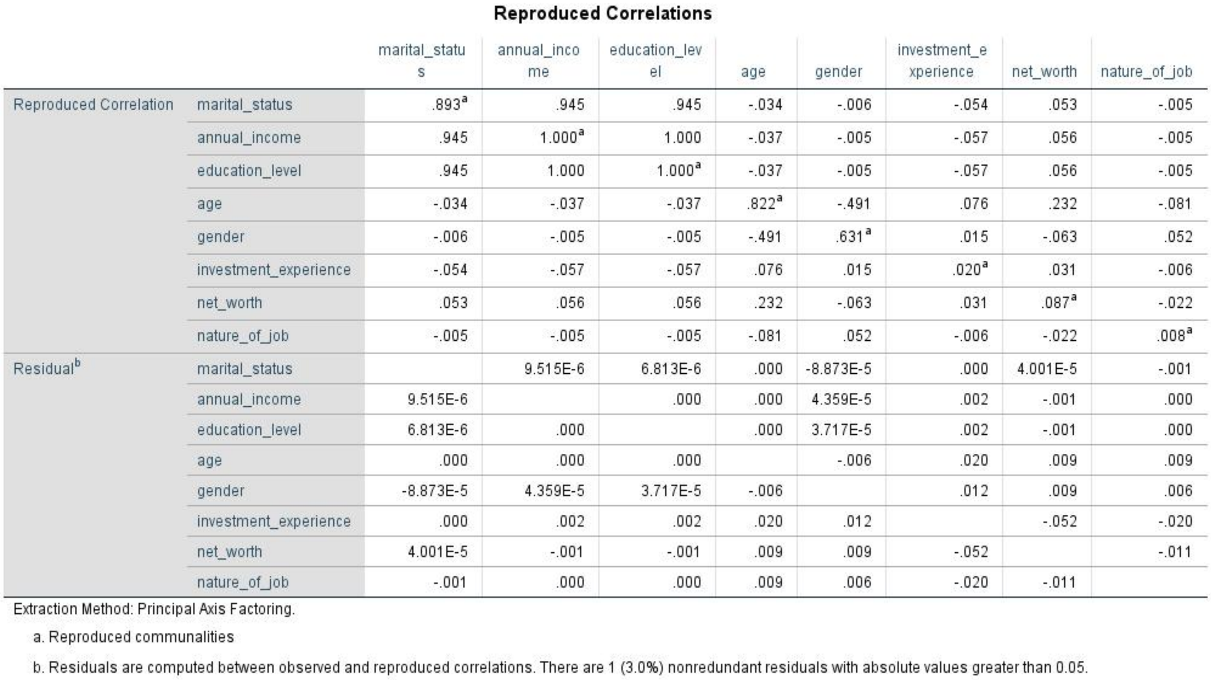
To filter statistically significant factors, a rotated factor matrix containing the rotated factor loadings was then generated and is shown in Figure 4.6. According to Field (2005), it is always helpful to increase the default value of 0.1 to 0.4 to ensure that factor loadings within plus or minus 0.4 are not displayed in the output, for easier interpretation. Stevens (1992) also recommended to interpret only factor loadings with an absolute value greater than 0.4. Thus the cut-off was set at 0.4 in the case study analysis. Referring to Figure 4.6, the factor loadings of “nature of employment” and “household net worth” are not displayed as the loadings are within plus or minus 0.4, meaning that these two variables should be excluded from further study. There were no cross-loadings on factors. The remaining variables have factor loadings above the cut-off value of 0.4 and were included. It was found that the variables cluster into three factors defined by the highest factor loading on each variable. Summarizing the factor structure of variables influencing investor behaviour.

i) Factor 1 is regarded as the sociological factor including marital status, education level, and annual income;

ii) Factor 2 is regarded as demographic factor including age and gender; and

iii) Factor 3 is regarded as psychological factors including investment experience.

Thus, the results of the factor analysis are in line with the findings from the literature reviewed in chapter 2.3.



Reproduced Correlations

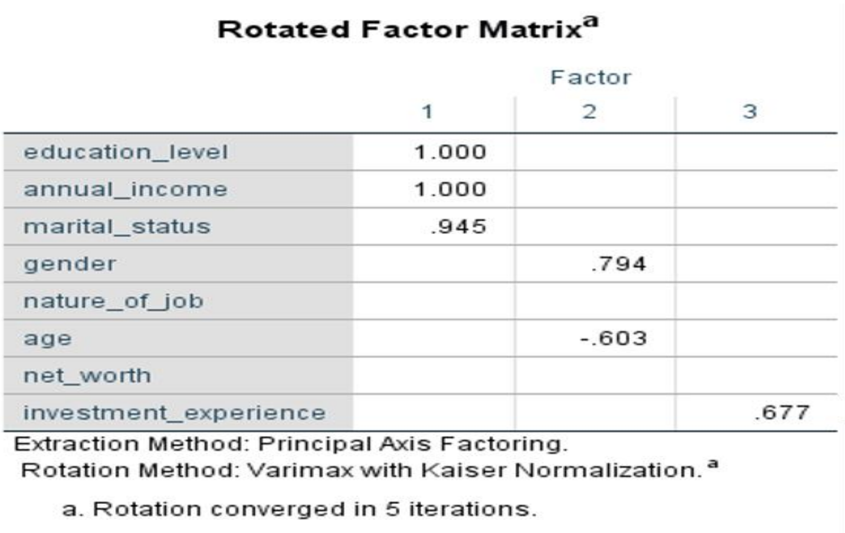
		marital_status	annual_income	education_level	age	gender	investment_experience	net_worth	nature_of_job
Reproduced Correlation	marital_status	.893[a]	.945	.945	-.034	-.006	-.054	.053	-.005
	annual_income	.945	1.000[a]	1.000	-.037	-.005	-.057	.056	-.005
	education_level	.945	1.000	1.000[a]	-.037	-.005	-.057	.056	-.005
	age	-.034	-.037	-.037	.822[a]	-.491	.076	.232	-.081
	gender	-.006	-.005	-.005	-.491	.631[a]	.015	-.063	.052
	investment_experience	-.054	-.057	-.057	.076	.015	.020[a]	.031	-.006
	net_worth	.053	.056	.056	.232	-.063	.031	.087[a]	-.022
	nature_of_job	-.005	-.005	-.005	-.081	.052	-.006	-.022	.008[a]
Residual[b]	marital_status		9.515E-6	6.813E-6	.000	-8.873E-5	.000	4.001E-5	-.001
	annual_income	9.515E-6		.000	.000	4.359E-5	.002	-.001	.000
	education_level	6.813E-6	.000		.000	3.717E-5	.002	-.001	.000
	age	.000	.000	.000		-.006	.020	.009	.009
	gender	-8.873E-5	4.359E-5	3.717E-5	-.006		.012	.009	.006
	investment_experience	.000	.002	.002	.020	.012		-.052	-.020
	net_worth	4.001E-5	-.001	-.001	.009	.009	-.052		-.011
	nature_of_job	-.001	.000	.000	.009	.006	-.020	-.011	

Extraction Method: Principal Axis Factoring.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 1 (3.0%) nonredundant residuals with absolute values greater than 0.05.

Figure 4.5 Summary of residuals



Rotated Factor Matrix[a]

	Factor 1	2	3
education_level	1.000		
annual_income	1.000		
marital_status	.945		
gender		.794	
nature_of_job			
age		-.603	
net_worth			
investment_experience			.677

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 5 iterations.

Figure 4.6 SPSS output for Rotated Factor Matrix

In sum, the eight variables influencing investor behaviour as suggested by the literature review were reduced to six as a result of factor analysis. This was supported by the expert panels. These were then further clustered into three factors, namely demographic, psychological and sociological factors. The structures of the factors are:

i) Demographic factors: age and gender

ii) Psychological factor: investment experience

iii) Sociological factors: annual income, education level and marital status

From the result of factor analysis, we can assume that financial investment behaviour is significantly affected by these demographic, psychological and sociological factors comprising altogether six variables, namely age, annual income, educational level, gender, investment experience and marital status. These six variables are significant predictors of financial investment behaviour for the Convoy data.

4.3 Correlation analysis to identify and measure associations between attributes and investment preferences

From the results presented in chapter 4.1, mainland Chinese and Hong Kong investors with different characteristics have different investment preferences. To further analyse how the six factors identified influence the investment behaviour of mainland Chinese and Hong Kong investors, correlation analysis was conducted to determine if correlations exist between the six variables and financial investment behaviours, and their respective impact on these relationships.

The six variables related to investors' characteristics are:

i) age

ii) annual income

iii) educational level

iv) gender

v) investment experience

vi) marital status

The two variables related to investment preferences are:

i) the quantity of fund unit held (Fund is managed by professional manager and would require less involvement from investors. Investors' behaviours would be greatly affected by their intention of investing involvement, namely passive or pro-active investors. Quantity of fund unit held by investors reflects their intention of personal involvement in investing exercise. Professionals from the expert panel expressed similar views.)

ii) the choice of country-specific financial investment options (Risk appetite definitely has great impact on investors' investing behaviours. Country-specific choice would indicate the risk level preferred from the macro-economic perspective. For example, the U.S. is generally perceived as low risk country for investment because of its largest economics and mature capital markets. This variable adopted as risk appetite measurement would be more reliable and broader than asset classes, such as stock and bonds which would have high risk and low risk respectively and would vary a lot from different judges. Experts from the panel agreed on these explanations.)

While the variables in this study were mixed with nominal and ordinal ones, normality may not actually apply to nominal data. As suggested by scholars (Kim, 2015; Sanani, 2012; Li et al., 2012; Caballero-Morales and Rahim, 2015; Mordkoff, 2016), it is not necessary to always transform the variables to make them normally distributed before running any analyses, especially with sufficiently large sample size. In this regard, it is assumed that the variables are ordinal, linear and normally-distributed.

It is common to check normality before analysis. With the support of the literature (Kim, 2015; Sanani, 2012; Li et al., 2012; Caballero-Morales and Rahim, 2015; Mordkoff, 2016), there is nothing inherently right or wrong to assume variables are normally distributed. Any assumption about normal variables or whether variables are normally distributed is indeed to ensure reliable results and draw valid conclusions. Normal distribution is a means, but not the results. In this study, whether the variables are assumed or confirmed to be normally distributed indeed does not affect the conclusions because:

i) The only concern on the effect of non-normality distributed variables is the sample size. With reference to the literature (Sanani, 2012; Caballero-Morales and Rahim, 2015; Mordkoff, 2016), normality assumption is relevant and critical only if the sample size is small. Sanani (2012) further indicated that "there is nothing inherently wrong with non-normal data" and "when it comes to statistical testing, normality is less critical" and irrelevant with sufficiently large sample size. Mordkoff (2016) suggested that "as long as the sample is based on 30 or more observations, the sampling distribution of the mean can be safely assumed to be normal". When the sample size is not large enough, the errors applied using non-normality distribution will have impacts on the p-values of the test on coefficients. Although in this study there is a non-normal distributed variable, gender, which is in fact considered as a negligible factor, so the distribution is not too grossly non-normal distributed. More importantly, in this study, the sample size is 142,496, which is sufficient to support the testing and address the concern of non-normality in the population.

ii) All the results generated by statistical analyses conducted in this study were diagnosed and validated by different tests such as Kaiser-Meyer-Olkin Measure, strength of correlation, Homogeneity of variance, and p-values. Their significance is confirmed statistically as well. Additionally, these statistical tools do not in fact require normally distributed variables.

iii) After undertaking the analyses, it is found that age, investment experience and annual income are most important factors (see table 4.3). However, the only non-normal variable, gender, among the data set is a negligible factor correlating to the investment behaviour. Therefore, this non-normal variable doesn't matter to the conclusions.

iv) According to Kim (2015) and Li *et al.* (2012), it is acceptable to include non-normal variables for analyses and they have given examples to illustrate. As suggested by Kim (2015), “in fact, linear regression analysis works well, even with non-normal errors. But, the problem is with p-values for hypothesis testing.” In this study, p-values were adopted for testing and validation of the results, this fulfilled what Kim advocated. According to Li *et al.* (2012), there are two key findings – (a) “there is a common misconception of the need to meet the normality assumption in linear regression techniques, and the validity of performing linear regression is compromised when this assumption is violated.” and (b) “in a large sample, the use of a linear regression technique, even if the dependent variable violates the normality assumption rule, remains valid.” The example discussed by Li *et al.* (2012) confirmed that “when a dependent variable is not distributed normally, linear regression remains a statistically sound technique in studies of large sample sizes (>3000).” Given that the sample size in this study is over 3000 and diagnostic checking was conducted, even if the normality assumption is violated or there is no such normality assumption, the relationships among the factors and their analysis results are still valid.

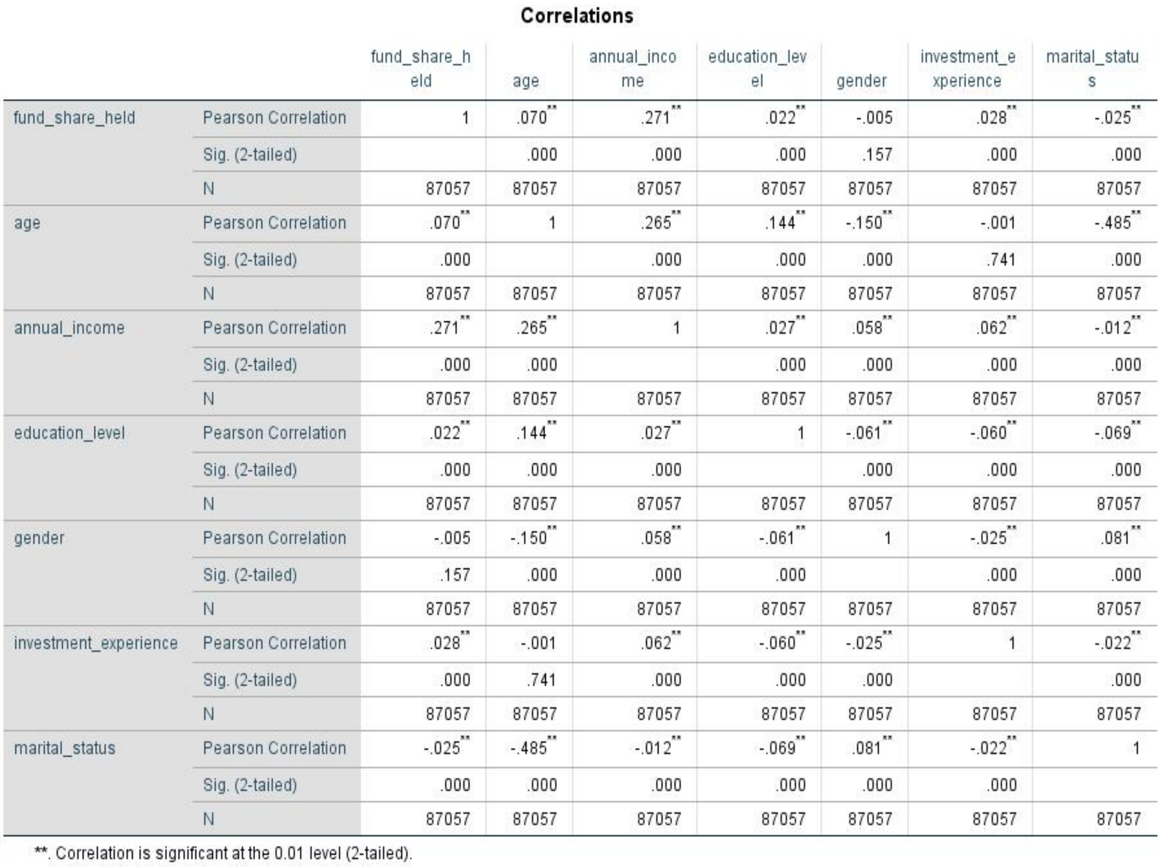
Figures 4.7 and 4.8 show the results of the correlation analysis between the six variables extracted by factor analysis and the quantity of fund unit held by mainland Chinese and by Hong Kong investors, respectively. Results are reported as Pearson correlation coefficients, r, which is a range of values from +1 to -1. As shown in Table 4.2, a value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association while a value less than 0 indicates a negative association. The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to having an absolute value of 1 (Thomas and Nelson, 2005).

Table 4.2 Description of Pearson correlation coefficients

Value of coefficient	**Description**
r<0	A negative association
r=0	No association between the two variables
r>0	A positive association

i) For mainland Chinese investors (see Figure 4.7), the correlations between the quantity of fund unit held and other variables, except gender with p > 0.001 was insignificant, existed (*p* < 0.001). Specifically, annual income (r = .271) had the strongest correlation with the quantity of fund unit held, followed by age (r = .070) and investment experience (r = .028).

ii) For Hong Kong investors (see Figure 4.8), correlations existed between all six variables and the quantity of fund unit held (*p* < 0.001). Among the six variables, the strength of the relationship was the strongest between annual income and quantity of fund unit held (r = .241), followed by age (r = .071) and investment experience (r = -.044).

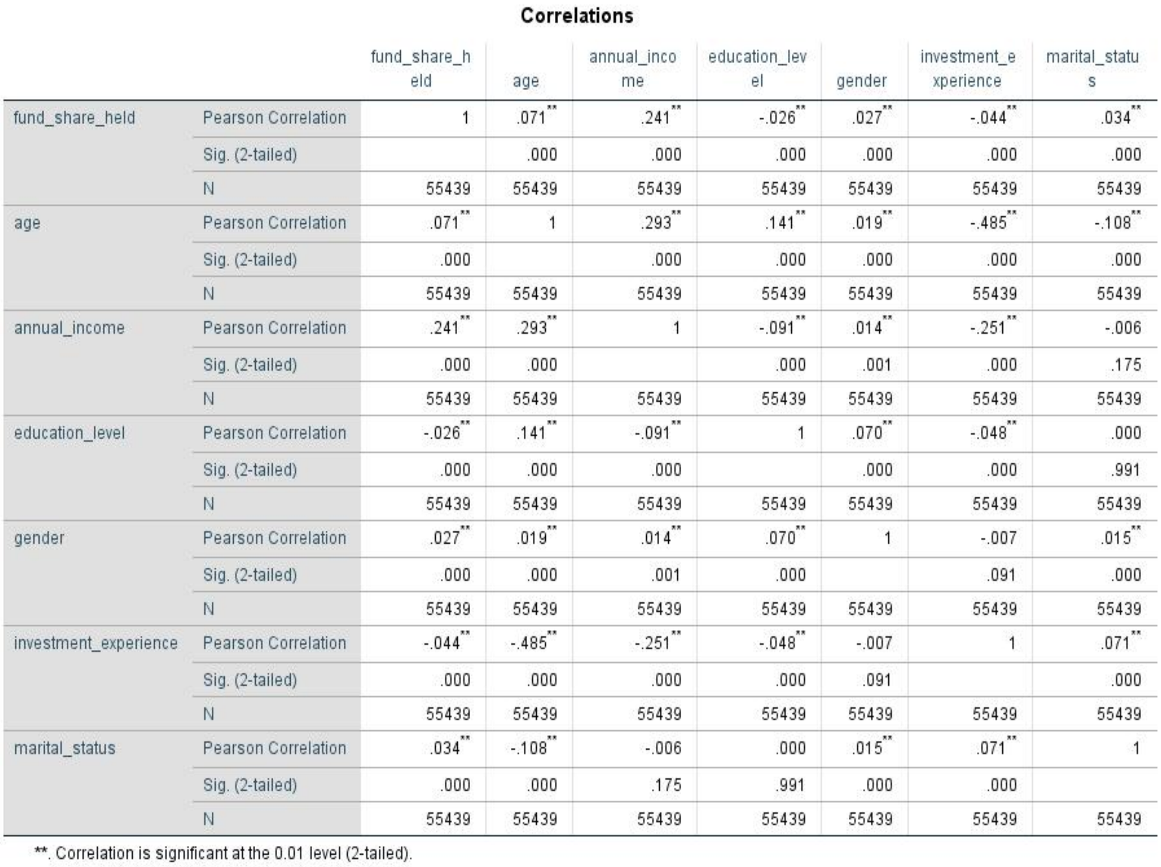


Correlations

		fund_share_held	age	annual_income	education_level	gender	investment_experience	marital_status
fund_share_held	Pearson Correlation	1	.070**	.271**	.022**	-.005	.028**	-.025**
	Sig. (2-tailed)		.000	.000	.000	.157	.000	.000
	N	87057	87057	87057	87057	87057	87057	87057
age	Pearson Correlation	.070**	1	.265**	.144**	-.150**	-.001	-.485**
	Sig. (2-tailed)	.000		.000	.000	.000	.741	.000
	N	87057	87057	87057	87057	87057	87057	87057
annual_income	Pearson Correlation	.271**	.265**	1	.027**	.058**	.062**	-.012**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000
	N	87057	87057	87057	87057	87057	87057	87057
education_level	Pearson Correlation	.022**	.144**	.027**	1	-.061**	-.060**	-.069**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
	N	87057	87057	87057	87057	87057	87057	87057
gender	Pearson Correlation	-.005	-.150**	.058**	-.061**	1	-.025**	.081**
	Sig. (2-tailed)	.157	.000	.000	.000		.000	.000
	N	87057	87057	87057	87057	87057	87057	87057
investment_experience	Pearson Correlation	.028**	-.001	.062**	-.060**	-.025**	1	-.022**
	Sig. (2-tailed)	.000	.741	.000	.000	.000		.000
	N	87057	87057	87057	87057	87057	87057	87057
marital_status	Pearson Correlation	-.025**	-.485**	-.012**	-.069**	.081**	-.022**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	
	N	87057	87057	87057	87057	87057	87057	87057

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 4.7 SPSS output for correlations between the quantity of fund unit held by mainland Chinese investors and six variables



Correlations

		fund_share_held	age	annual_income	education_level	gender	investment_experience	marital_status
fund_share_held	Pearson Correlation	1	.071**	.241**	-.026**	.027**	-.044**	.034**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	55439	55439	55439	55439	55439	55439	55439
age	Pearson Correlation	.071**	1	.293**	.141**	.019**	-.485**	-.108**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000
	N	55439	55439	55439	55439	55439	55439	55439
annual_income	Pearson Correlation	.241**	.293**	1	-.091**	.014**	-.251**	-.006
	Sig. (2-tailed)	.000	.000		.000	.001	.000	.175
	N	55439	55439	55439	55439	55439	55439	55439
education_level	Pearson Correlation	-.026**	.141**	-.091**	1	.070**	-.048**	.000
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.991
	N	55439	55439	55439	55439	55439	55439	55439
gender	Pearson Correlation	.027**	.019**	.014**	.070**	1	-.007	.015**
	Sig. (2-tailed)	.000	.000	.001	.000		.091	.000
	N	55439	55439	55439	55439	55439	55439	55439
investment_experience	Pearson Correlation	-.044**	-.485**	-.251**	-.048**	-.007	1	.071**
	Sig. (2-tailed)	.000	.000	.000	.000	.091		.000
	N	55439	55439	55439	55439	55439	55439	55439
marital_status	Pearson Correlation	.034**	-.108**	-.006	.000	.015**	.071**	1
	Sig. (2-tailed)	.000	.000	.175	.991	.000	.000	
	N	55439	55439	55439	55439	55439	55439	55439

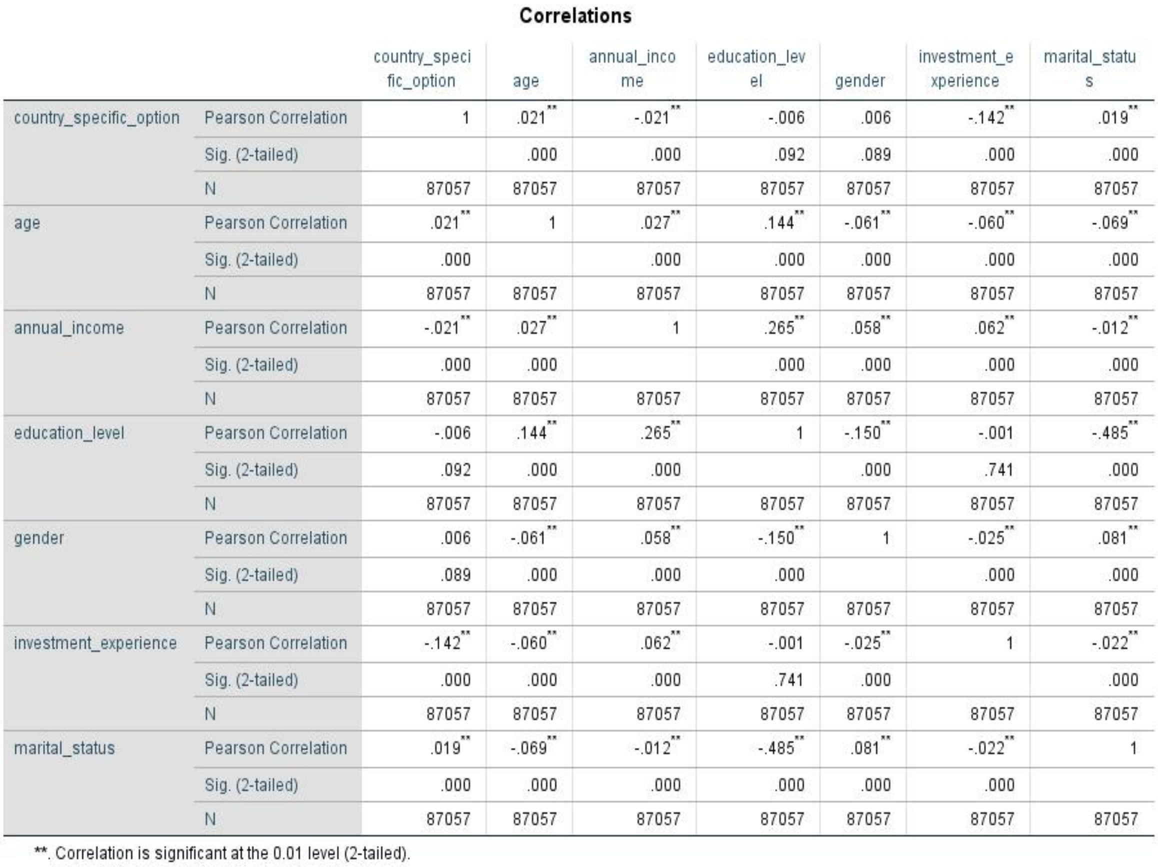
**. Correlation is significant at the 0.01 level (2-tailed).

Figure 4.8 SPSS output for correlations between quantity of fund unit held by Hong Kong investors and six variables

Figures 4.9 and 4.10 show results from the correlation analysis between the six variables extracted by factor analysis and the choice of country-specific financial investment options by mainland Chinese investors and by Hong Kong investors, respectively.

For mainland Chinese investors (see Figure 4.9), the correlations between the choice of country-specific financial investment options and the remaining four significant variables, which were age, annual income, investment experience and marital status, existed ($p < 0.001$). The strongest correlation with the choice of country-specific financial investment options was investment experience (r = .142), followed by age (r = .021) and annual income (r = -.021). Correlations between education level and the choice of country-specific financial investment options, and between gender and the choice of country-specific financial investment were not significant.

For Hong Kong investors (see Figure 4.10), significant correlations were found between age ($p < 0.05$), annual income ($p < 0.05$), investment experience ($p < 0.001$) and choice of country-specific financial investment options, respectively. The strength of the correlation was the strongest between investment experience and the choice of country-specific financial investment option (r = .075), followed by age (r = .012) and annual income (r = -.010). No correlations were found between education level, gender and marital status and the choice of country-specific financial investment options, respectively.

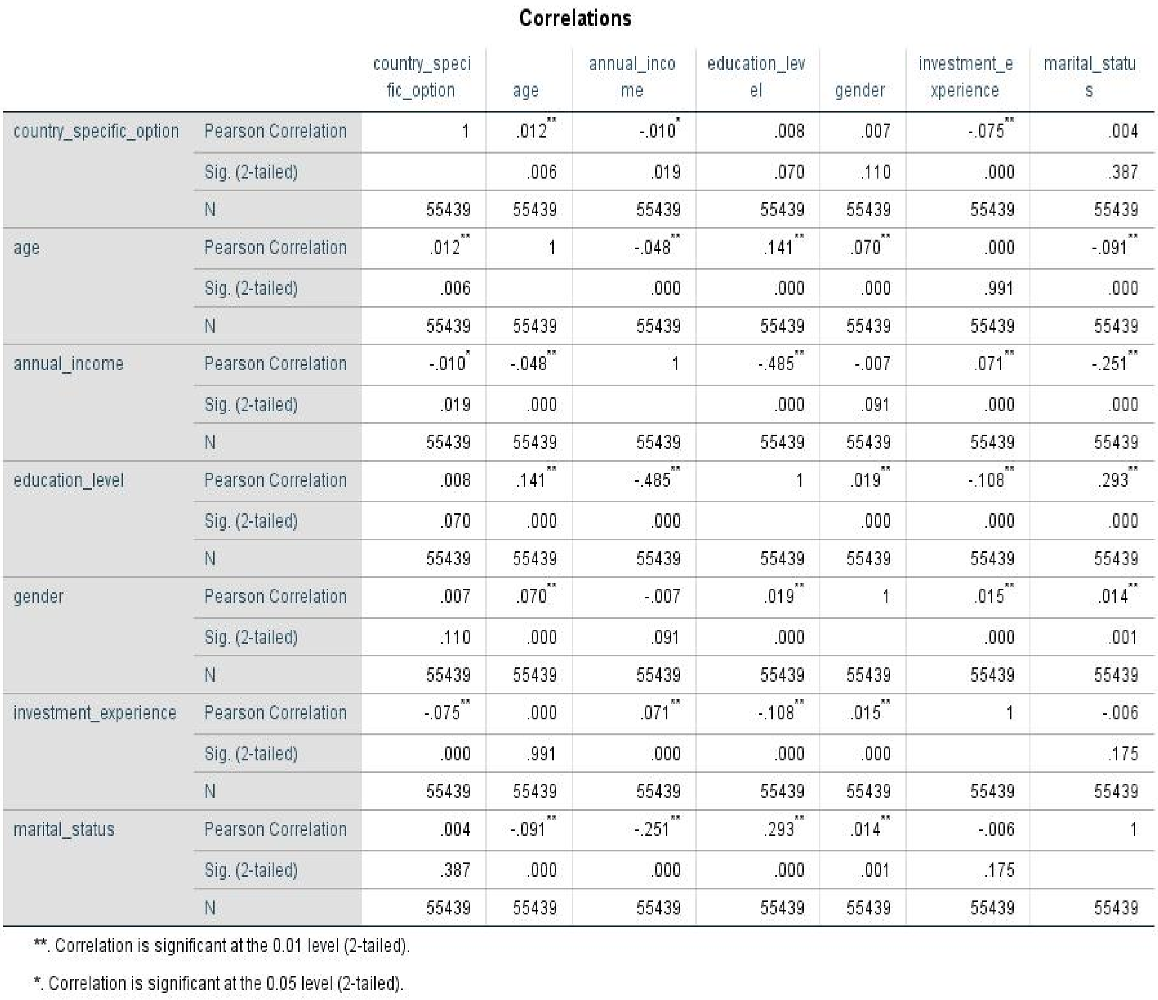


Correlations

		country_specific_option	age	annual_income	education_level	gender	investment_experience	marital_status
country_specific_option	Pearson Correlation	1	.021**	-.021**	-.006	.006	-.142**	.019**
	Sig. (2-tailed)		.000	.000	.092	.089	.000	.000
	N	87057	87057	87057	87057	87057	87057	87057
age	Pearson Correlation	.021**	1	.027**	.144**	-.061**	-.060**	-.069**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000
	N	87057	87057	87057	87057	87057	87057	87057
annual_income	Pearson Correlation	-.021**	.027**	1	.265**	.058**	.062**	-.012**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000
	N	87057	87057	87057	87057	87057	87057	87057
education_level	Pearson Correlation	-.006	.144**	.265**	1	-.150**	-.001	-.485**
	Sig. (2-tailed)	.092	.000	.000		.000	.741	.000
	N	87057	87057	87057	87057	87057	87057	87057
gender	Pearson Correlation	.006	-.061**	.058**	-.150**	1	-.025**	.081**
	Sig. (2-tailed)	.089	.000	.000	.000		.000	.000
	N	87057	87057	87057	87057	87057	87057	87057
investment_experience	Pearson Correlation	-.142**	-.060**	.062**	-.001	-.025**	1	-.022**
	Sig. (2-tailed)	.000	.000	.000	.741	.000		.000
	N	87057	87057	87057	87057	87057	87057	87057
marital_status	Pearson Correlation	.019**	-.069**	-.012**	-.485**	.081**	-.022**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	
	N	87057	87057	87057	87057	87057	87057	87057

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 4.9 SPSS output for correlations between the country-specific financial investment option selected by mainland Chinese investors and six variables



Correlations

		country_specific_option	age	annual_income	education_level	gender	investment_experience	marital_status
country_specific_option	Pearson Correlation	1	.012**	-.010*	.008	.007	-.075**	.004
	Sig. (2-tailed)		.006	.019	.070	.110	.000	.387
	N	55439	55439	55439	55439	55439	55439	55439
age	Pearson Correlation	.012**	1	-.048**	.141**	.070**	.000	-.091**
	Sig. (2-tailed)	.006		.000	.000	.000	.991	.000
	N	55439	55439	55439	55439	55439	55439	55439
annual_income	Pearson Correlation	-.010*	-.048**	1	-.485**	-.007	.071**	-.251**
	Sig. (2-tailed)	.019	.000		.000	.091	.000	.000
	N	55439	55439	55439	55439	55439	55439	55439
education_level	Pearson Correlation	.008	.141**	-.485**	1	.019**	-.108**	.293**
	Sig. (2-tailed)	.070	.000	.000		.000	.000	.000
	N	55439	55439	55439	55439	55439	55439	55439
gender	Pearson Correlation	.007	.070**	-.007	.019**	1	.015**	.014**
	Sig. (2-tailed)	.110	.000	.091	.000		.000	.001
	N	55439	55439	55439	55439	55439	55439	55439
investment_experience	Pearson Correlation	-.075**	.000	.071**	-.108**	.015**	1	-.006
	Sig. (2-tailed)	.000	.991	.000	.000	.000		.175
	N	55439	55439	55439	55439	55439	55439	55439
marital_status	Pearson Correlation	.004	-.091**	-.251**	.293**	.014**	-.006	1
	Sig. (2-tailed)	.387	.000	.000	.000	.001	.175	
	N	55439	55439	55439	55439	55439	55439	55439

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 4.10 SPSS output for correlations between the country-specific financial investment option selected by Hong Kong investors and six variables

Table 4.2 summarises the correlation coefficients of the correlations between the quantity of fund unit held, choice of country-specific financial investment options and the six variables for mainland Chinese and Hong Kong investors, and the order of the relationship strengths. The strengths of the relationships were measured by the magnitude of the correlation coefficient while the sign (+/-) measured the direction of the relationship. 'N/A' denotes the absence of a significant correlation or no correlation.

From Table 4.2, it is noted that annual income, age, investment experience, and marital status were the top four variables closely associating with the quantity of fund unit held by mainland Chinese and Hong Kong investors. The only difference regarding the correlations of the quantity of fund unit held by Mainland Chinese and Hong Kong investors was gender. The significant correlation between gender and the quantity of fund unit held only existed among Hong Kong investors but not mainland Chinese investors.

From Table 4.2, regarding the country-specific financial investment options selected by investors, it was shown that the choice by both mainland Chinese and Hong Kong investors was correlated with the most highly correlated variables – investment experience, age and annual income. No correlations were found between education level and the choice of country-specific financial investment options, nor between gender and the choice of country-specific financial investment options, selected by mainland Chinese and Hong Kong investors. Marital status showed the weakest correlation with the choice of country-specific financial investment options selected by mainland Chinese investors, and no correlation with options selected by Hong Kong investors.

Most importantly, it was shown that the investment behaviours, in terms of the quantity of fund unit held and the choice of country-specific financial investment options by both mainland Chinese and Hong Kong investors were correlated with age, annual income and investment experience.

Table 4.3 Summary of correlation analyses

	Quantity of fund unit held		Choice of country-specific financial investment option	
	Mainland China	Hong Kong	Mainland China	Hong Kong
Age	2 (r = .070)	2 (r = .071)	2 (r = .021)	2 (r = .012)
Annual income	1 (r = .271)	1 (r = .241)	2 (r = -.021)	3 (r = -.010)
Education level	5 (r = .022)	6 (r = -.026)	N/A	N/A
Gender	N/A	5 (r = .027)	N/A	N/A
Investment experience	3 (r = .028)	3 (r = -.044)	1 (r = .142)	1 (r = -.075)
Marital status	4 (r = .025)	4 (r = .034)	4 (r = .019)	N/A

Note. 1 meaning the strongest correlation; 6 meaning the weakest correlation

Based on the results of the correlation analysis conducted regarding financial investment behaviour, in terms of both the quantity of fund unit held and the choice of country-specific financial investment options, and the six variables, as summarised in Table 4.2, it was concluded that the financial investment decisions made by investors from mainland China and Hong Kong were mainly affected by the demographic factor age, psychological factor investment experience, and the sociological factor annual income.

4.4 Linear regression analysis to assess the relationships between the attributes and investment preferences

As discussed in chapter 4.2, the six variables including age, annual income, educational level, gender, investment experience and marital status, were statistically significant predictors of financial investment behaviour. Furthermore, age, investment experience and annual income were found to be the highly-associated factors with mainland Chinese and Hong Kong investors' preferences, as discussed in chapter 4.3. These results thus serve as basis for the next step, linear regression analysis, to further assess the relationships between the six variables and both mainland Chinese and Hong Kong investors' investment preferences. Linear regression is the most commonly used in regression and often works for analysing multi-factor data (Keith, 2006; Osborne and Waters, 2002; Montgomery, *et al.,* 2012). Polynomial regression as one of the linear regression models can be useful when a curvilinear relationship exists (Liu, *et al.*, 2005). Polynomial regression is a relatively simple but it can allow for both linear and non-linear relationship (The Pennsylvania State University, 2018).

Three regression models were formulated in this study to examine:

i) if the investment behaviour of mainland Chinese and Hong Kong investors, when considered together, can be predicted by the six variables - age, annual income, educational level, gender, investment experience and marital status;

ii) if the quantity of fund unit held by the mainland Chinese and Hong Kong investors, when considered separately, can be predicted by the six variables; and

iii) if the choice of country-specific financial investment options selected by the mainland Chinese and Hong Kong investors, when considered separately, can be predicted by the six variables.

The first regression model provides a general picture for understanding whether and how the six variables identified affect investors in both mainland China and Hong Kong. The second and third regression models are specified to obtain a more in-depth examination on the differences in investment behaviour between mainland Chinese and Hong Kong investors.

In these linear regression models, the variables are assumed ordinal, linear and normally-distributed, as justified in chapter 4.3. This assumption for regression analysis is further strengthened by Kim (2015), who conducted a study to examine the necessity of normality for variables through case studies and proven that none of the variables have to be normal in linear regression analysis to draw valid outcomes.

4.4.1 Regression model 1

The first regression model concerns the investment behaviour of both Hong Kong and mainland Chinese investors; the model is described below:

<u>Model 1a</u>

Dependent variable: the quantity of fund unit held ($\mathrm{fundhold}_T$)

Independent variables: age (age_T), annual income ($\mathrm{annualincome}_T$), gender (gender_T), educational level ($\mathrm{educationallevel}_T$), investment experience ($\mathrm{investmentexperience}_T$) and marital status ($\mathrm{maritalstatus}_T$)

Considering the dependent and independent variables, the below equation is formed, where α is the regression constant and $\beta_1, \beta_2, \ldots, \beta_6$ are regression coefficients for each independent variable:

$$\mathrm{fundhold}_T = \alpha + \beta_1 \mathrm{age}_T + \beta_2 \mathrm{annualincome}_T + \beta_3 \mathrm{educationlevel}_T + \beta_4 \mathrm{gender}_T + \beta_5 \mathrm{investmentexperience}_T + \beta_6 \mathrm{maritalstatus}_T \qquad (1)$$

<u>Model 1b</u>

Dependent variable: choice of country-specific financial investment options ($\mathrm{fundcurrency}_T$)

Independent variables: age (age_T), annual income ($\mathrm{annualincome}_T$), gender (gender_T), educational level ($\mathrm{educationallevel}_T$), investment experience ($\mathrm{investmentexperience}_T$) and marital status ($\mathrm{maritalstatus}_T$)

Considering the dependent and independent variables, the following equation is formed, where α is the regression constant and $\beta_1, \beta_2, \ldots, \beta_6$ are regression coefficients for each independent variable:

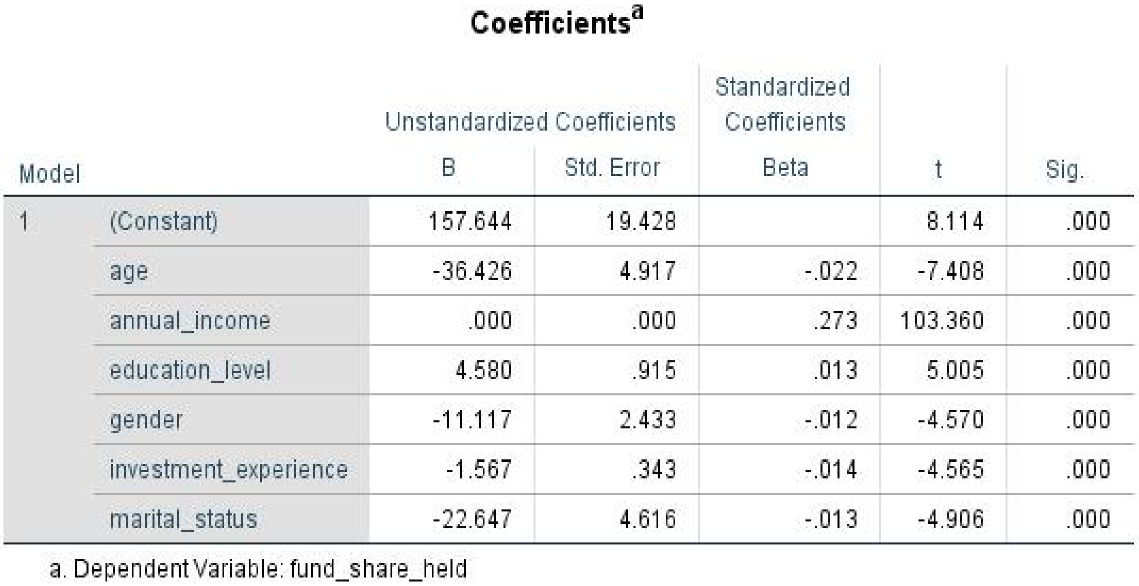
$$\mathrm{Fundcurrency}_T = \alpha \quad + \beta_1 \mathrm{age}_T + \beta_2 \mathrm{annualincome}_T + \beta_3 \mathrm{educationlevel}_T + \beta_4 \mathrm{gender}_T + \beta_5 \mathrm{investmentexperience}_T + \beta_6 \mathrm{maritalstatus}_T \qquad (2)$$

To ensure the validity of the regression model, several requirements for applying multiple regression, including linearity, multivariate normality, homogeneity of variance and multicollinearity, were tested. As discussed by Poole and O'Farrell (1971) and Antonakis and Dietz (2011), the models are only valid when these requirements are tested and satisfied. Before conducting the actual regression analysis, preliminary analyses were conducted to ensure the requirements for the regression models are fulfilled.

Linearity test

Linearity assumes that the dependent variable is a linear function of the independent variables (Darlington, 1968). Keith (2006) indicated that the assumption of linearity is the most important, as it relates directly to potential bias of results. Osborne and Waters (2002) further supported that multiple regression can accurately estimate the relationship between dependent and independent variables only when the relationship is linear in nature.

Results of the significance of the linear relationship are shown in Figure 4.11. The p-values for all variables were <0.0001, indicating that significant linear relationships between all independent variables and dependent variable existed. Thus, the assumption of linearity for regression was fulfilled.



Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	157.644	19.428		8.114	.000
	age	-36.426	4.917	-.022	-7.408	.000
	annual_income	.000	.000	.273	103.360	.000
	education_level	4.580	.915	.013	5.005	.000
	gender	-11.117	2.433	-.012	-4.570	.000
	investment_experience	-1.567	.343	-.014	-4.565	.000
	marital_status	-22.647	4.616	-.013	-4.906	.000

a. Dependent Variable: fund_share_held

Figure 4.11 SPSS output for multiple correlation coefficients

Multivariate normality test

Multivariate normality means that variables are normally distributed. According to Osborne and Waters (2002), non-normally distributed variables can distort relationships and significance tests. The assumption of multivariate normality can be checked by visual inspection of histograms of the standardised residuals (Stevens, 2009).

Referring to Figure 4.12, the histogram of residuals of model 1a showed a symmetrical bell-shape and fairly normal distribution. Thus, the assumption of multivariate normality for regression model 1 appeared fulfilled.

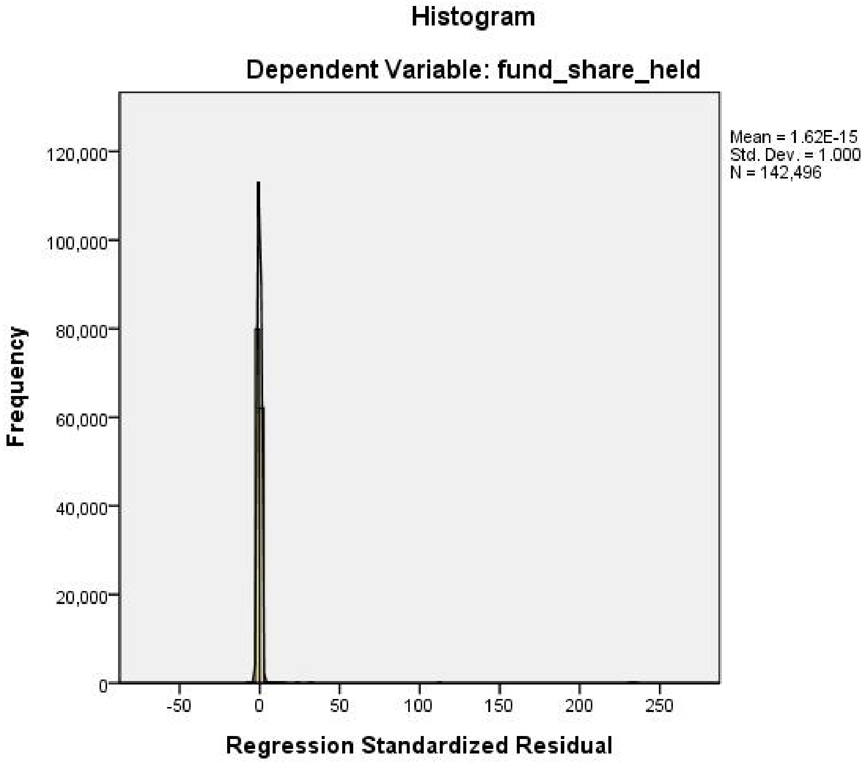


Figure 4.12 SPSS output for histogram of residuals

Homogeneity of variance

Homogeneity of variance means that the variance of the residuals is homogeneous for all values of the independent variable (Keith, 2006). When this assumption is violated, it leads to distortion of the findings, weakens the overall analysis and weakens its statistical power (Aguinis *et al.,* 1999; Osborne and Waters, 2002). To test the assumption of homogeneity of variance, a visual examination of the scatter plot of residuals should be conducted (Osborne and Waters, 2002). The assumption is fulfilled if the residuals are scattered randomly and close to the zero axis.

Referring to Figure 4.13, the data of model 1a showed almost a horizontal band of points, scattered around and close to the zero axis. Thus, the assumption of homogeneity of variance for regression model 1 was fulfilled.

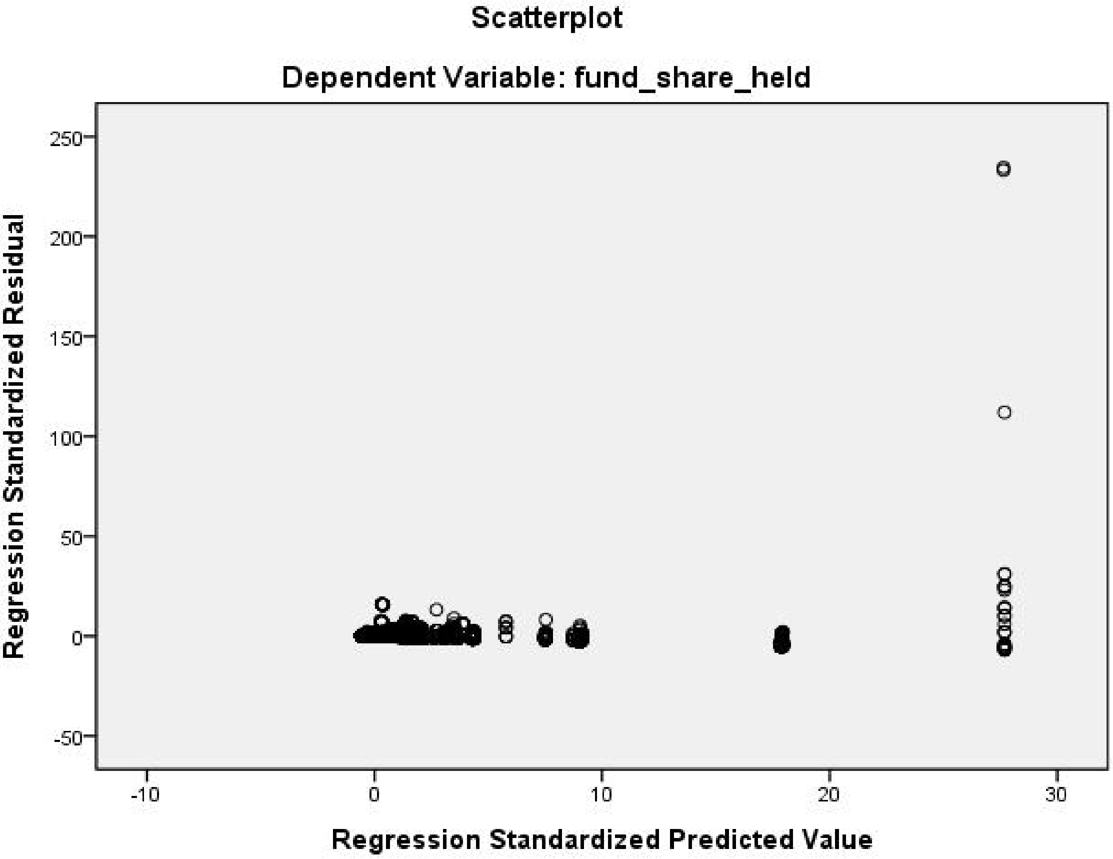


Figure 4.13 SPSS output for analysis of residuals

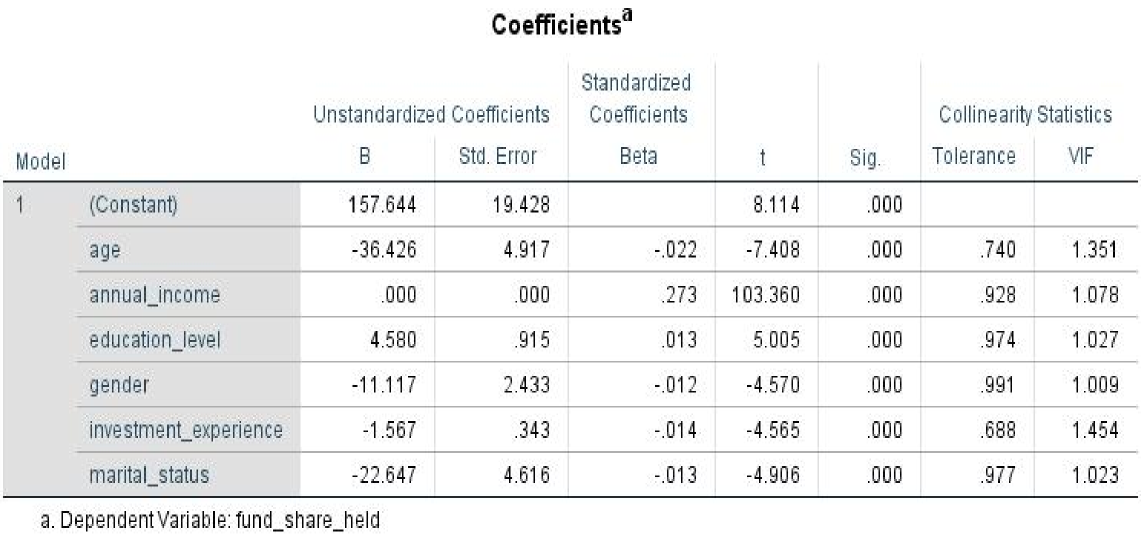
Multicollinearity

Multicollinearity assumes that the independent variables are uncorrelated, and contribute a unique part of the total explanation of the variance in the dependent variable (Darlington, 1968; Keith, 2006). The regression coefficients can be interpreted as the effects of the independent variables on the dependent variable only when collinearity is low (Keith, 2006; Poole and O'Farrell, 1971; Mason and Perreault Jr., 1991). To test the assumption of multicollinearity for regression model 1a, the coefficients of determination (R^2) and variance inflation factors (VIF) were calculated.

Tolerance measures the influence of one independent variable on all other independent variables. Tolerance is defined as $T = 1 - R^2$ and tolerance levels for correlations range from zero (no independence) to one (completely independent) (Keith, 2006). As a rule of thumb, a value of tolerance less than 0.2 indicates a problem with multicollinearity (Hart and Sailor, 2009; Rawlings, 1988).

VIF measures the amount of variance of each regression coefficient that are inflated as compared to the uncorrelated independent variables in a regression analysis (Keith, 2006). If the value of VIF is equal to 1, the predictors are not correlated. If the value of VIF is between 1 and 5, the predictors are moderately correlated. If the value of VIF is larger than 5 or equal to 10, the predictors are highly correlated (Keith, 2006; Shieh, 2010). The smaller the value of VIF is, the lower the probability of multicollinearity.

Referring to Figure 4.14, the values for tolerance for the six independent variables were greater than 0.20 and the VIF values were almost equal to 1. These indicate low multicollinearity and thus the assumption of multicollinearity for regression model 1a was fulfilled.



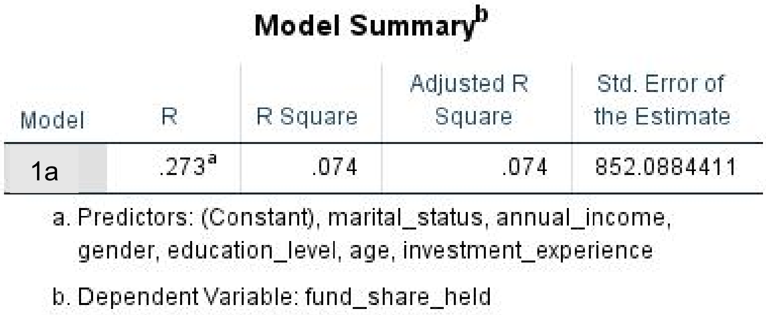
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	157.644	19.428		8.114	.000		
	age	-36.426	4.917	-.022	-7.408	.000	.740	1.351
	annual_income	.000	.000	.273	103.360	.000	.928	1.078
	education_level	4.580	.915	.013	5.005	.000	.974	1.027
	gender	-11.117	2.433	-.012	-4.570	.000	.991	1.009
	investment_experience	-1.567	.343	-.014	-4.565	.000	.688	1.454
	marital_status	-22.647	4.616	-.013	-4.906	.000	.977	1.023

a. Dependent Variable: fund_share_held

Figure 4.14 SPSS output for the measure of tolerance

Since it has been established that model 1a fulfilled all assumptions, the regression analysis was conducted. Results are shown in Figures 4.15 and 4.16. As can be seen in Figure 4.15, the R-squared value was 0.074 or 7.4%, indicating that 7.4% variability of the quantity of fund unit held could be explained by age, annual income, education level, gender, investment experience and marital status. According to Shieh (2010), it is typical to have a low R-squared value in case of predicting human behaviour. From Figure 4.17, p-values of age, annual income, education level, gender, investment experience and marital status were <.001. Thus, age, annual income, education level, gender, investment experience and marital status were statistically significant predictors of the quantity of fund unit held by mainland Chinese and Hong Kong investors. By leveraging the results shown in Figures 4.15 and 4.16, it could be concluded that the changes in the predictor values were correlated with changes in the value of the dependent variable.



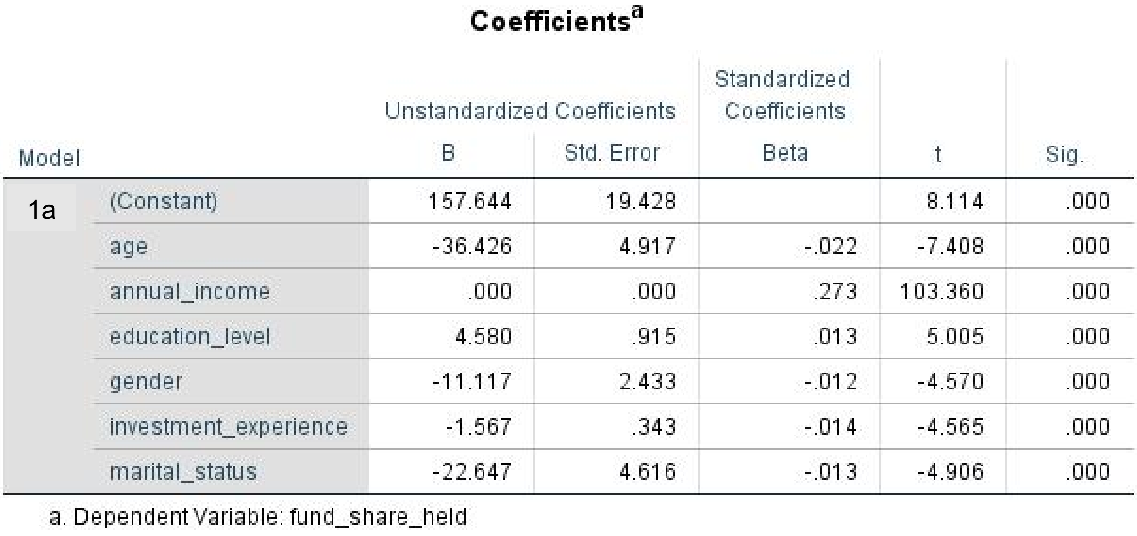
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1a	.273[a]	.074	.074	852.0884411

a. Predictors: (Constant), marital_status, annual_income, gender, education_level, age, investment_experience

b. Dependent Variable: fund_share_held

Figure 4.15 SPSS output for model summary of Model 1a



Coefficients[a]

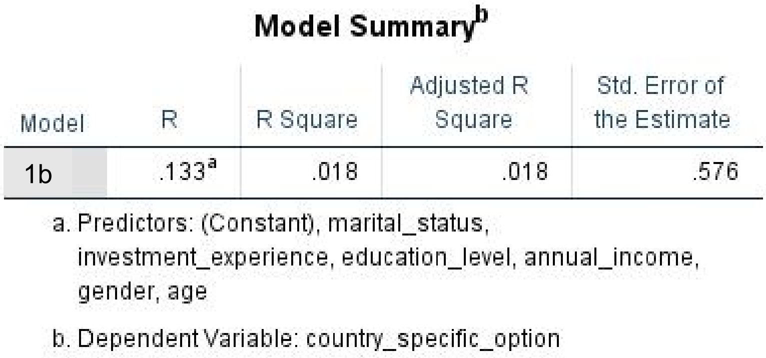
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1a	(Constant)	157.644	19.428		8.114	.000
	age	-36.426	4.917	-.022	-7.408	.000
	annual_income	.000	.000	.273	103.360	.000
	education_level	4.580	.915	.013	5.005	.000
	gender	-11.117	2.433	-.012	-4.570	.000
	investment_experience	-1.567	.343	-.014	-4.565	.000
	marital_status	-22.647	4.616	-.013	-4.906	.000

a. Dependent Variable: fund_share_held

Figure 4.16 SPSS output for regression coefficients of Model 1a

The same preliminary analyses of the checking assumptions as discussed in Model 1a were conducted for Model 1b, and all requirements were fulfilled. The actual regression results are shown in Figures 4.17 and 4.18.

From Figure 4.17, the R-squared value was 0.018 or 1.8%. This indicates that 1.8% variability of the choice of country-specific financial investment options by mainland Chinese and Hong Kong investors could be explained by age, annual income, education level, gender, investment experience and marital status. From Figure 4.18, p-values of the former variables were less than 0.0001, while the p-value of gender was 0.223 and hence greater than 0.05. Thus, age, annual income, education level, investment experience and marital status were statistically significant predictors of the choice of country-specific financial investment option selected by mainland Chinese and Hong Kong investors, while gender was not.



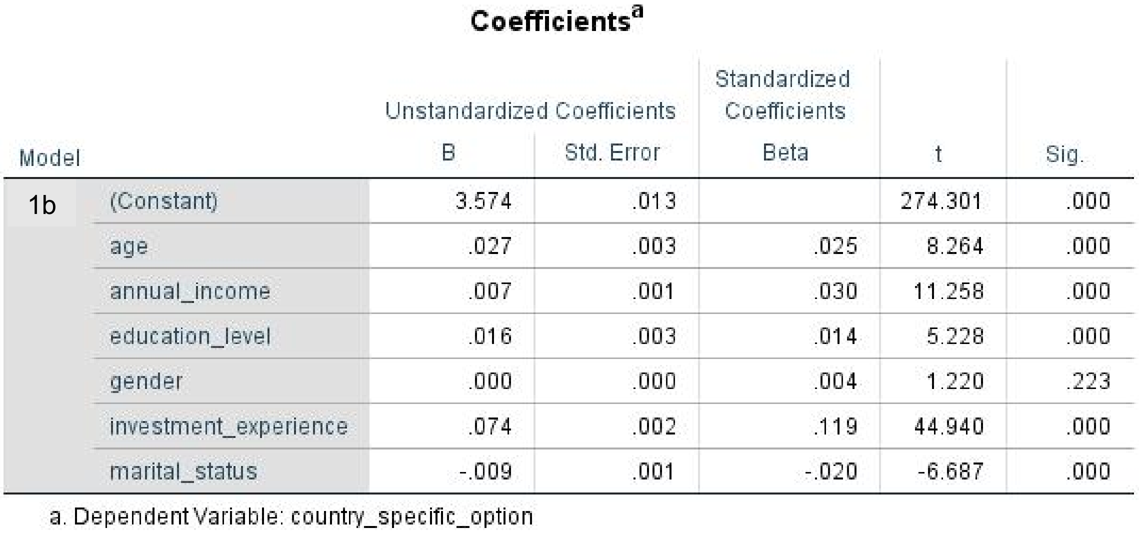
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1b	.133[a]	.018	.018	.576

a. Predictors: (Constant), marital_status, investment_experience, education_level, annual_income, gender, age

b. Dependent Variable: country_specific_option

Figure 4.17 SPSS output for model summary of Model 1b



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1b	(Constant)	3.574	.013		274.301	.000
	age	.027	.003	.025	8.264	.000
	annual_income	.007	.001	.030	11.258	.000
	education_level	.016	.003	.014	5.228	.000
	gender	.000	.000	.004	1.220	.223
	investment_experience	.074	.002	.119	44.940	.000
	marital_status	-.009	.001	-.020	-6.687	.000

a. Dependent Variable: country_specific_option

Figure 4.18 SPSS output for regression coefficients of Model 1b

The results of regression analyses for Models 1a and 1b are summarised in Table 4.3.

Table 4.4 Outputs of regression analyses for Model 1a and 1b

	Sig. (Standardised coefficient)	
	Model 1a	**Model 1b**
Age	<.0001 (-.022)	<.0001 (.025)
Annual income	<.0001 (.273)	<.0001 (.030)
Education level	<.0001 (.013)	<.0001 (.014)
Gender	<.0001 (-.012)	.223 (N/A)
Investment experience	<.0001 (-.014)	<.0001 (.119)
Marital status	<.0001 (-.013)	<.0001 (-.020)

Based on the results of regression Models 1a and 1b, it could be concluded that the financial investment behaviour of mainland Chinese and Hong Kong investors, when considered together, was inseparable with their demographic factor (i.e. age), psychological factor (i.e. investment experience) and sociological factors (i.e. annual income, education level and marital status). The top three variables influencing investor preference were age, annual income and investment experience.

4.4.2 Regression model 2

The second regression model, concerning the quantity of fund unit held by mainland Chinese and Hong Kong investors, is as follows:

Model 2a

Dependent variable: quantity of fund unit held by mainland Chinese investors ($\mathrm{fundhold}_{CN}$)

Independent variables: age (age_{CN}), annual income ($\mathrm{annualincome}_{CN}$), gender ($\mathrm{gender}_{CN}$), educational level ($\mathrm{educationallevel}_{CN}$), investment experience ($\mathrm{investmentexperience}_{CN}$) and marital status ($\mathrm{maritalstatus}_{CN}$)

Considering the dependent and independent variables, the below equation is formed, where α is the regression constant and $\beta_1, \beta_2, \ldots, \beta_6$ are regression coefficients for each independent variable:

$$\mathrm{fundhold}_{CN} = \alpha + \beta_1\mathrm{age}_{CN} + \beta_2\mathrm{annualincome}_{CN} + \beta_3\mathrm{educationlevel}_{CN} + \beta_4\mathrm{gender}_{CN} + \beta_5\mathrm{investmentexperience}_{CN} + \beta_6\mathrm{maritalstatus}_{CN} \quad (3)$$

Model 2b

Dependent variable: quantity of fund unit held by Hong Kong investors ($\mathrm{fundhold}_{HK}$)

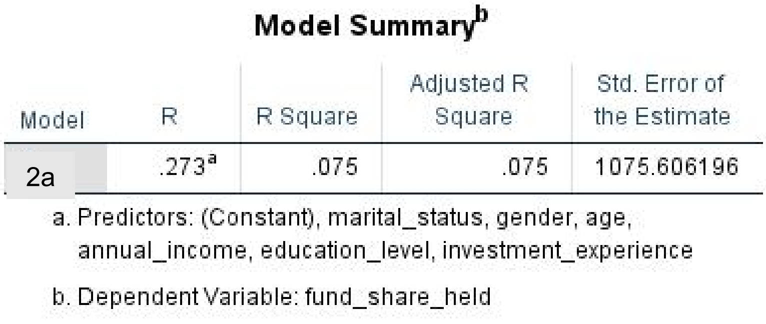
Independent variables: age (age_{HK}), annual income ($\mathrm{annualincome}_{HK}$), gender (gender_{HK}), educational level ($\mathrm{educationallevel}_{HK}$), investment experience ($\mathrm{investmentexperience}_{HK}$) and marital status ($\mathrm{maritalstatus}_{HK}$)

Considering the dependent and independent variables, the below equation is formed, where α is the regression constant and $\beta_1, \beta_2, \ldots, \beta_6$ are regression coefficients for each independent variable:

$$\mathrm{fundhold}_{HK} = \alpha + \beta_1\mathrm{age}_{HK} + \beta_2\mathrm{annualincome}_{HK} + \beta_3\mathrm{educationlevel}_{HK} + \beta_4\mathrm{gender}_{HK} + \beta_5\mathrm{investmentexperience}_{HK} + \beta_6\mathrm{maritalstatus}_{HK} \quad (4)$$

The same preliminary analyses for checking assumptions as discussed for Model 1a were conducted for Models 2a and 2b, and it was found that all the requirements were fulfilled. The regression results for Model 2a are shown in Figures 4.19 and 4.20, while the results for Model 2b are shown in Figures 4.21 and 4.22.

From Figure 4.19, the R-squared value was .075 or 7.5%. It indicated that 7.5% variability of the quantity of fund unit held by mainland Chinese investors could be explained by age, annual income, education level, gender, investment experience and marital status. From Figure 4.20, p-values of age, annual income, education level, investment experience and marital status were <.0001 while the p-value of gender was significant at <.05. Thus, it was confirmed that age, annual income, education level, gender, investment experience and marital status were statistically significant predictors of the quantity of fund unit held by mainland Chinese investors.



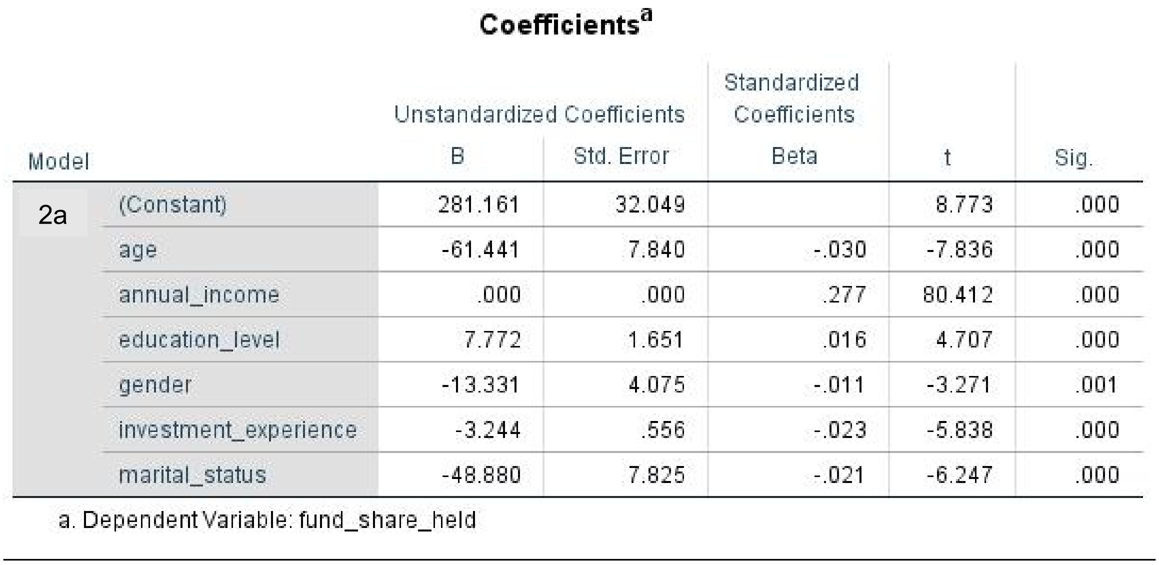
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
2a	.273[a]	.075	.075	1075.606196

a. Predictors: (Constant), marital_status, gender, age, annual_income, education_level, investment_experience

b. Dependent Variable: fund_share_held

Figure 4.19 SPSS output for model summary of Model 2a



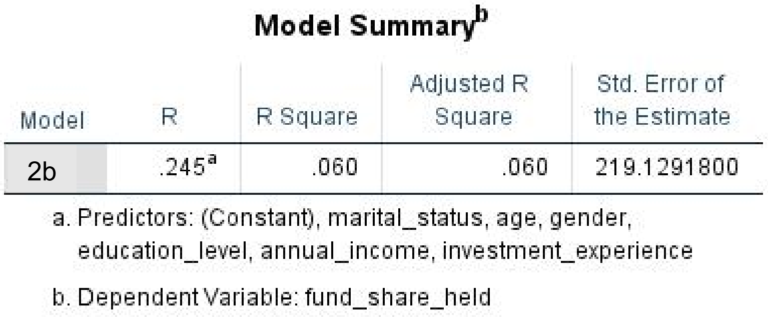
Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
2a	(Constant)	281.161	32.049		8.773	.000
	age	-61.441	7.840	-.030	-7.836	.000
	annual_income	.000	.000	.277	80.412	.000
	education_level	7.772	1.651	.016	4.707	.000
	gender	-13.331	4.075	-.011	-3.271	.001
	investment_experience	-3.244	.556	-.023	-5.838	.000
	marital_status	-48.880	7.825	-.021	-6.247	.000

a. Dependent Variable: fund_share_held

Figure 4.20 SPSS output for regression coefficients of Model 2a

From Figure 4.21, the R-squared value was .060 or 6.0%, which indicated that 6.0% variability of the quantity of fund unit held by Hong Kong investors could be explained by age, annual income, education level, gender, investment experience and marital status. From Figure 4.22, p-values of age, annual income, gender and investment experience were <.0001; p-value of marital status is <.05. Thus, age, annual income, gender, investment experience and marital status were significant predictors of the quantity of fund unit held by Hong Kong investors.



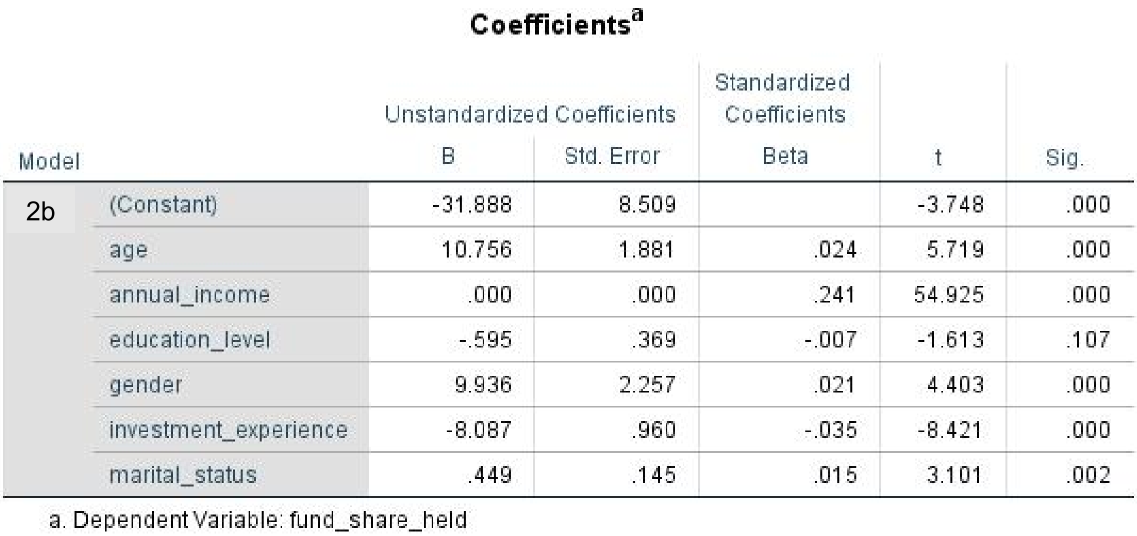
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
2b	.245[a]	.060	.060	219.1291800

a. Predictors: (Constant), marital_status, age, gender, education_level, annual_income, investment_experience

b. Dependent Variable: fund_share_held

Figure 4.21 SPSS output for model summary of Model 2b



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
2b	(Constant)	-31.888	8.509		-3.748	.000
	age	10.756	1.881	.024	5.719	.000
	annual_income	.000	.000	.241	54.925	.000
	education_level	-.595	.369	-.007	-1.613	.107
	gender	9.936	2.257	.021	4.403	.000
	investment_experience	-8.087	.960	-.035	-8.421	.000
	marital_status	.449	.145	.015	3.101	.002

a. Dependent Variable: fund_share_held

Figure 4.22 SPSS output for regression coefficients of Model 2b

The results of regression analyses for Model 2 are summarised in Table 4.4.

Table 4.5 Output of regression analyses for Model 2

	Sig. (Standardised coefficient)	
	Model 2a	**Model 2b**
Age	<.0001 (-.030)	<.0001 (.024)
Annual income	<.0001 (.277)	<.0001 (.241)
Education level	<.0001 (.016)	.107 (N/A)
Gender	.001 (-.011)	<.0001 (.021)
Investment experience	<.0001 (-.023)	<.0001 (-.035)
Marital status	<.0001 (-.021)	.002 (.015)

Based on regression Models 2a and 2b, it could be concluded that the quantity of fund unit held by mainland Chinese and Hong Kong investors, when considered separately, were correlated with their demographic factors (i.e. age and gender), psychological factor (i.e. investment experience) and sociological characteristics (i.e. annual income and marital status). The top three significant variables influencing investment behaviour were the age, annual income and investment experience. The standardised coefficient of age was -0.030 for mainland Chinese investors and 0.024 for Hong Kong investors. The standardised coefficient of annual income was 0.277 for mainland Chinese investors and 0.241 for Hong Kong investors. The standardised coefficient of investment experience was -0.023 for mainland Chinese investors and -0.235 for Hong Kong investors. In spite of the differences in the magnitude of the three most significant variables, the directions of the relationships were similar, except for age. For example, annual income of both mainland Chinese and Hong Kong investors had a positive impact on the quantity of fund unit held by both. However, age of mainland Chinese investors had a negative impact on the quantity of fund unit held, but a positive impact among Hong Kong investors. In other words, younger mainland Chinese investors held more fund units compared to older mainland Chinese investors, while older Hong Kong investors held more fund units compared to younger ones.

4.4.3 Regression model 3

The third regression model, concerning the choice of country-specific financial investment options of mainland Chinese and Hong Kong investors, is defined follows:

Model 3a

Dependent variable: choice of country-specific financial investment option by mainland Chinese investors ($\mathrm{fundcurrency}_{CN}$)

Independent variables: age (age_{CN}), annual income ($\mathrm{annualincome}_{CN}$), gender ($\mathrm{gender}_{CN}$), educational level ($\mathrm{educationallevel}_{CN}$), investment experience ($\mathrm{investmentexperience}_{CN}$) and marital status ($\mathrm{maritalstatus}_{CN}$)

Considering the dependent and independent variables, the below equation is formed, where α is the regression constant and $\beta_1, \beta_2,\ldots, \beta_6$ are regression coefficients for each independent variable:

$$\mathrm{fundcurrency}_{CN} = \alpha + \beta_1\mathrm{age}_{CN} + \beta_2\mathrm{annualincome}_{CN} + \beta_3\mathrm{educationlevel}_{CN} + \beta_4\mathrm{gender}_{CN} + \beta_5\mathrm{investmentexperience}_{CN} + \beta_6\mathrm{maritalstatus}_{CN} \qquad (5)$$

<u>Model 3b</u>

Dependent variable: choice of country-specific financial investment option by Hong Kong investors ($\mathrm{fundcurrency}_{HK}$)

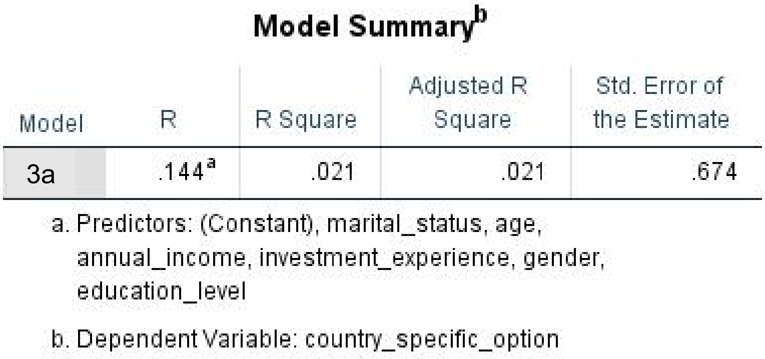
Independent variables: age (age_{HK}), annual income ($\mathrm{annualincome}_{HK}$), gender (gender_{HK}), educational level ($\mathrm{educationallevel}_{HK}$), investment experience ($\mathrm{investmentexperience}_{HK}$) and marital status ($\mathrm{maritalstatus}_{HK}$)

Considering the dependent and independent variables, the below equation is formed, where α is the regression constant and $\beta_1, \beta_2,\ldots, \beta_6$ are regression coefficients for each independent variable:

$$\mathrm{fundcurrency}_{HK} = \alpha + \beta_1\mathrm{age}_{HK} + \beta_2\mathrm{annualincome}_{HK} + \beta_3\mathrm{educationlevel}_{HK} + \beta_4\mathrm{gender}_{HK} + \beta_5\mathrm{investmentexperience}_{HK} + \beta_6\mathrm{maritalstatus}_{HK} \qquad (6)$$

The same preliminary analyses for checking assumptions as discussed in Model 1a were conducted for Model 3a and 3b, and all requirements were fulfilled. The regression results for Model 3a are shown in Figures 4.23 and 4.24, while the results for Model 3b are shown in Figures 4.25 and 4.26.

From Figure 4.23, the R-squared value was .021 or 2.1%, indicating that 2.1% variability of the choice of country-specific financial investment option by mainland Chinese investors could be explained by their age, annual income, education level, gender, investment experience and marital status. From Figure 4.24, p-values of age, annual income, investment experience and marital status variables were <.0001, meaning that they were significant predictors of choice of country-specific financial investment option by mainland Chinese investors.



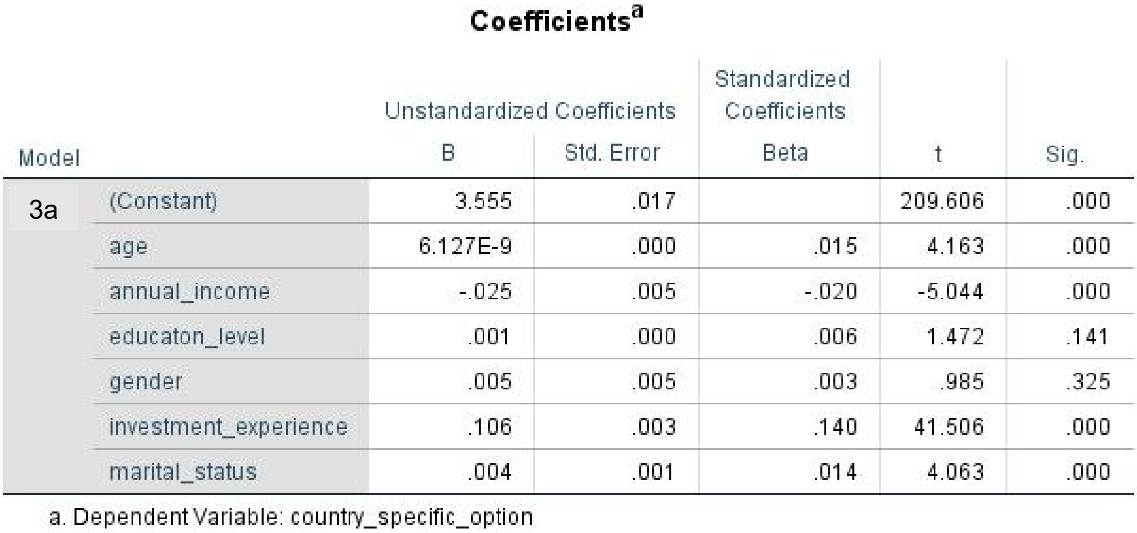
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
3a	.144[a]	.021	.021	.674

a. Predictors: (Constant), marital_status, age, annual_income, investment_experience, gender, education_level

b. Dependent Variable: country_specific_option

Figure 4.23 SPSS output for model summary of Model 3a



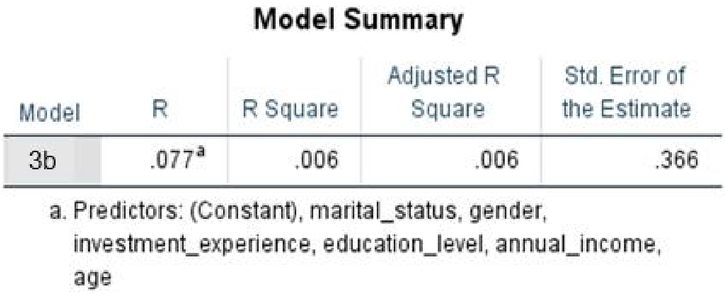
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
3a	(Constant)	3.555	.017		209.606	.000
	age	6.127E-9	.000	.015	4.163	.000
	annual_income	-.025	.005	-.020	-5.044	.000
	educaton_level	.001	.000	.006	1.472	.141
	gender	.005	.005	.003	.985	.325
	investment_experience	.106	.003	.140	41.506	.000
	marital_status	.004	.001	.014	4.063	.000

a. Dependent Variable: country_specific_option

Figure 4.24 SPSS output for regression coefficients of Model 3a

From Figure 4.25, the R-square was .006 or 0.6%, which indicated that 0.6% variability of the choice of country-specific financial investment option by Hong Kong investors could be explained by their age, annual income, education level, gender, investment experience and marital status. From Figure 4.26, p-values of age, annual income, education level, gender, and investment experience were <0.0001; while the p-value of marital status was .127 (>0.05). Thus, only age, annual income, education level, gender and investment experience were significant predictors of the choice of country-specific financial investment option by Hong Kong investors.

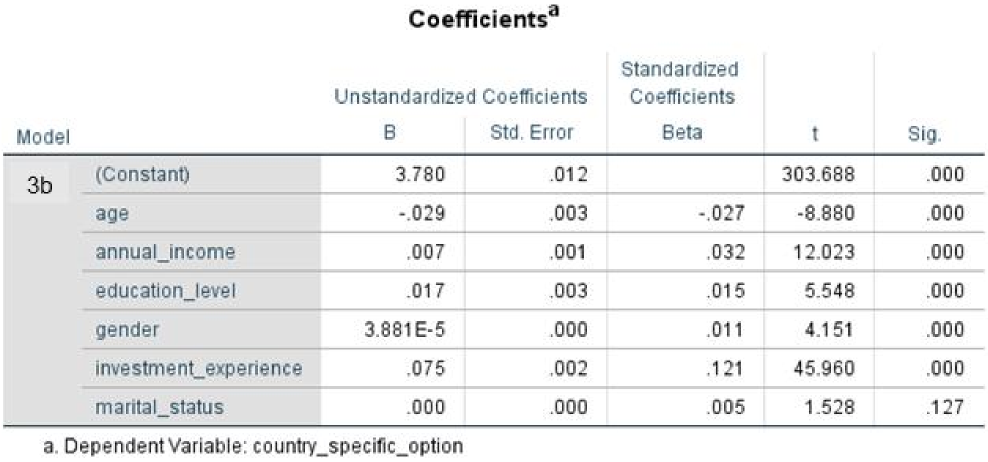


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
3b	.077[a]	.006	.006	.366

a. Predictors: (Constant), marital_status, gender, investment_experience, education_level, annual_income, age

Figure 4.25 SPSS output for model summary of Model 3b



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
3b	(Constant)	3.780	.012		303.688	.000
	age	-.029	.003	-.027	-8.880	.000
	annual_income	.007	.001	.032	12.023	.000
	education_level	.017	.003	.015	5.548	.000
	gender	3.881E-5	.000	.011	4.151	.000
	investment_experience	.075	.002	.121	45.960	.000
	marital_status	.000	.000	.005	1.528	.127

a. Dependent Variable: country_specific_option

Figure 4.26 SPSS output for regression coefficients of Model 3b

The results of regression analyses for Model 3 are summarised in Table 4.6.

Table 4.6 Output of regression analyses for Model 3

	Sig. (Standardised coefficient)	
	Model 3a	**Model 3b**
Age	<.0001 (.015)	<.0001 (-.027)
Annual income	<.0001 (-.020)	<.0001 (.032)
Education level	.141 (N/A)	<.0001 (.015)
Gender	.325 (N/A)	<.0001 (.011)
Investment experience	<.0001 (.140)	<.0001 (.121)
Marital status	<.0001 (.014)	<.0001 (.005)

Based on the results of regression models 3a and 3b as summarised in Table 4.6, it can be concluded that the choice of country-specific financial investment options by mainland Chinese and Hong Kong investors, when considered separately, were related to a demographical factor (i.e. age), a psychological factor (i.e. investment experience) and sociological factors (i.e. annual income and marital status). The three most significant variables influencing investment behaviour were age, annual income and investment experience. The standardised coefficient of age was 0.015 for mainland Chinese investors and -0.027 for Hong Kong investors. The higher the value of the coefficients, the stronger the effect. This means that age positively affected the choice of country-specific financial investment options of mainland Chinese, but negatively affected options selected by Hong Kong investors. In other words, younger mainland Chinese investors and older Hong Kong investors tended to have the same choice of country-specific financial investment options, meaning that they were willing to take similar amount of risk. This could probably be due to cultural difference. The standardised coefficient of annual income was -0.020 for mainland Chinese investors and 0.032 for Hong Kong investors. This indicates that annual income negatively affected the choice of country-specific financial investment option of mainland Chinese investors, but positively affected those selected by Hong Kong investors. The standardised coefficient of investment experience was 0.140 for mainland Chinese investors and 0.121 for Hong Kong investors. Investment experience had a positive impact on the choice of country-specific financial investment options of both mainland China and Hong Kong investors.

4.4.4. Summary of regression analyses results

From the results of regression models 1, 2 and 3, significant differences existed in financial investment behaviour in terms of the quantity of fund unit held and their choice of country-specific financial investment options between mainland China and Hong Kong investors.

i) The impact of age on the quantity of fund unit held by mainland Chinese and Hong Kong investors was reversed: Age negatively affected the quantity of fund unit held by mainland Chinese investors but positively affected the amount held by Hong Kong investors. This reflects that mainland Chinese investors would decreasingly rely on investment professional from the financial institution, such as fund managers, when their age increases while Hong Kong investors demonstrated the opposite. As suggested by the expert panel, the most likely causes are:

   - mainland Chinese investors would be less confident with investment professional such as fund managers as fund management in China is less professional and investors have had more bad experience with the fund management industry in mainland China as they age; and

- Hong Kong investors seem more confident with such professionals and more willing to allocate higher funding for investment as they age. Such phenomenon may also reflect the investors' differing confidence level of investment professional due to different market regulation regimes.

ii) Age positively affected the choice of country-specific financial investment options of mainland Chinese, but negatively affected those selected by Hong Kong investors. This implies that mainland Chinese would be more likely to allocate assets to currency of major economies like US and Japan when they are getting older. Hong Kong investors demonstrate the opposite direction. Both implement the concept of diversification as they age.

iii) Annual income negatively affected the choice of country-specific financial investment options of mainland Chinese investors, but positively affected those selected by Hong Kong investors. The results show higher income investors may be more interested in high yield currency, like the Australian dollar, in mainland China. Hong Kong investors are of the opposite thought.

Similarities between the investment behaviour of mainland Chinese and Hong Kong investors were also found:

i) The three most significant factors, namely age, annual income and investment experience, influencing investment behaviours for both mainland Chinese and Hong Kong investors were the same, though the influence may have been in the opposite direction.

ii) Annual income had a positive impact on the quantity of fund unit held by mainland Chinese and Hong Kong investors.

iii) Investment experience had a positive impact on the choice of country-specific financial investment options of both mainland Chinese and Hong Kong investors.

iv) Investment experience had a negative impact on the quantity of fund unit held by both mainland Chinese and Hong Kong investors.

4.5 Recommendations and the PSYC Model

From the findings, financial institutions could attempt to maximise their business opportunities by the following means:

i) designing different financial investment products based on demographic, psychological and sociological factors, such as age, annual income and investment experience, of individual investors from mainland China and Hong Kong;

ii) design more accurate and focused marketing campaign and strategies in targeting mainland Chinese investors; and

iii) offering more specific and professional trainings to their sales representatives in dealing with mainland Chinese and Hong Kong investors.

Mainland Chinese investors are becoming more and more important in the financial services industry in Hong Kong and in the world. An effective marketing strategy to attract these investors could be a critical factor for a successful financial services company. Based on the results of the regression analyses and the three most significant predictors identified, the following targeted marketing strategies can be formulated for financial services companies in Hong Kong to attract mainland Chinese and Hong Kong investors in a more precise and systematic fashion.

The expert panels reviewed and discussed the findings from the data analysis in chapter four and concluded that the industry did need a top-down model to discover the rationale of investment behaviour at the top level and then predict investors' preferences at the bottom level in order to get insights into the big picture of investing activities in the market. This chapter correlates to the top level, at which four common characteristics of investment behaviours can be generalized, namely:

i) Professional-directed (“P”),

ii) Self-directed (“S”),

iii) Yield-driven (“Y”), and

iv) Comfort-driven (“C”);

collectively the “PSYC Model”. With the knowledge of the investors’ preferences on these four dimensions, effective and appropriate marketing strategies for products and services design can be formulated to help companies to achieve their business objectives.

The “quantity of fund unit held” indicates the level of preference for reliance on professional investment managers. Higher levels point to a more Professional-directed preference. Conversely there is more Self-directed behaviour if they hold a smaller “quantity of fund unit held”.

The “country specific” financial investment option reflects the investors’ choice in the investment currency. Such a choice indicates the investors’ behaviour in relation to yield, i.e. Yield-driven, and leading economies, thus Comfort-driven.

Chinese Investors	Age		Annual Income		Investment Experience	
High	S	C	P	Y	S	C
Low	P	Y	S	C	P	Y

HK Investors	Age		Annual Income		Investment Experience	
High	P	Y	P	C	S	C
Low	S	C	S	Y	P	Y

Figure 4.27 Investors’ behaviour in relation to four common characteristics

Thus a high-yield currency fund would be the most appealing for young, high-income and less-experienced Chinese investors. Conversely major currency investment products with a more personal involvement would be better for senior, low-income and experienced Chinese investors. Product design and marketing strategies can be more effective with the guidance from Figure 4.27. For example, an Australian Dollar fund (“PY Product”) can be marketed to IT industry employees who usually are young, high income and with less-investment-experience in China. A public seminar is more suitable to promote stocks (“SC Product”) targeted at active investors and such a seminar would attract more senior participants with greater investment experience.

For Hong Kong investors, the product design and marketing strategies need to be more sophisticated than those for mainland Chinese investors. It may be due to Hong Kong being a more mature investment market with more experienced investors. It is much more difficult to construct a simple investment product to target a specific investor group with consideration of age, income and investment experience. For example, PY Product would be better marketed to senior and less-investment-experience investors in Hong Kong but not necessary for high- or low-income investors. Similarly, SC Product would not have obvious appeal for investors with different income level. As a consequence, a wider product spectrum is necessary for Hong Kong investors. For instance, high-income investors would prefer a USD fund ("PC Product") while low-income investors would like Euro stocks ("SY Product") for example. These suggestions are the correlations evident from the data analysis and supported by the expert panel.

Besides, the data from Figure 4.27 shows that Chinese investors have a more consistent investment behaviour pattern. This may be due to the market conditions in China, such as:

i) the history of investment market is relatively short;

ii) the diversity of investment products is relatively small; and

iii) investment experience and investment knowledge are relatively less.

Having understood the PSYC Model, experts from the panel conducted in this study (see chapter 3.2.3) agreed on the intuitive rationale and reckoned that this model is consistent with their practice in their respective industries and experience. Some experts expected a model in a more detailed form to provide more practical use for daily business operation. In addition, all experts agreed that financial product type in terms of PSYC and risk level are the paramount consideration for ordinary business operation. PS is reflecting product type direction while YC is reflecting risk level. Thus, the model is good to generalize the direction of investment product design for a specific group of investors, i.e. a segment of investment population.

The newly developed PSYC Model was derived to be a generalized and simple model for investment product design and marketing based on the behavioural finance concepts and the major factors influencing investment behaviour identified in this chapter. To further help financial institutions to manage investment preferences of their customers, a data mining system could be useful and thus is described in the next chapter.

Chapter 5 The Financial Data Mining Model

The major attributes influencing investment behaviour were identified in chapter four. Following this, an understanding of how financial institutions predict customer behaviour is required to be able to define the architecture as well as the implementation of a Financial Data Mining Model (FDMM). This model at the base level needs to be able to extract changing customer behaviour patterns and derive their investment preferences from the data. The architecture of the proposed model is discussed and a case study illustrative how the proposed model may work is presented in this chapter.

5.1 Introduction and architecture of the Financial Data Mining Model (FDMM)

The results of the statistical analyses in chapter four showed significant differences in financial investment behaviour between mainland Chinese and Hong Kong investors. Data showed their behaviours were affected by demographic, sociological and psychological factors. For financial institutions to market the most appropriate products to individual mainland Chinese and Hong Kong investors, a better understanding of their changing financial investment behaviour needs to be derived. This must be in an effective and efficient manner so that the knowledge generated can be applied promptly. The FDMM can be a tool to help identify financial investment preferences for potential customers. Using the FDMM financial institutions can identify their customers' preferences, devise the most appropriate financial investment strategies and design financial investment products/portfolios targeted at their customers. Referring to the discussion in chapter 2.6, a proposed architecture of FDMM requiring different modules in terms of data selection, customer segmentation, and rules generation is desired. The batch processing approach of the FMDD is illustrated in Figure 5.1 and its modules are described below.

i) Data Selection and Pre-processing Module (DSPM): A centralised database is critical for financial institutions to store and retrieve data for analysis as discussed in the expert panel with details given in chapter three. As identified in chapter 3.6.3, a lack of a centralised data warehouse is one of the challenges faced by Convoy. The DSPM is thus designed to handle individual datasets in Convoy by selecting and formatting relevant data for the mining process.

ii) Clustering Module (CM): Clustering helps create better customer segmentation. To better understand the needs of different customer groups and to better identify customer needs – a problem faced by Convoy, identified in chapter 3.6.3. The CM is designed to identify the key influencing factors that affect the customers' financial investment behaviour and then segment customers into different groups based on the factor(s) identified. Segmentation of the target group(s) among all mainland Chinese and Hong Kong customers in the database allows useful cluster-based rules to be discovered at subsequent stage.

iii) Rules Discovery Module (RDM): As discussed in the expert panel conducted for this study, predicting customers' investment preferences is regarded as a key goal, as this helps improve financial institutions' product and marketing strategies. The PSYC model discussed in chapter 4.5 is capable of understanding the rationale of investment behaviour at the top level while the FDMM at the bottom level is designed to achieve the ultimate goal. With the support of DSPM and CM, the RDM is specifically designed to discover "useful" rules for each desired cluster. The "useful" association rules identified, help financial planners understand the relationships to different financial investment products and help convert them into knowledge concerning the development of financial investment portfolios. The results of RDM provide insights for financial institutions into investors' preference.

Financial institutions can then formulate marketing strategies, such as price, products and promotion strategies, based on the rules uncovered.

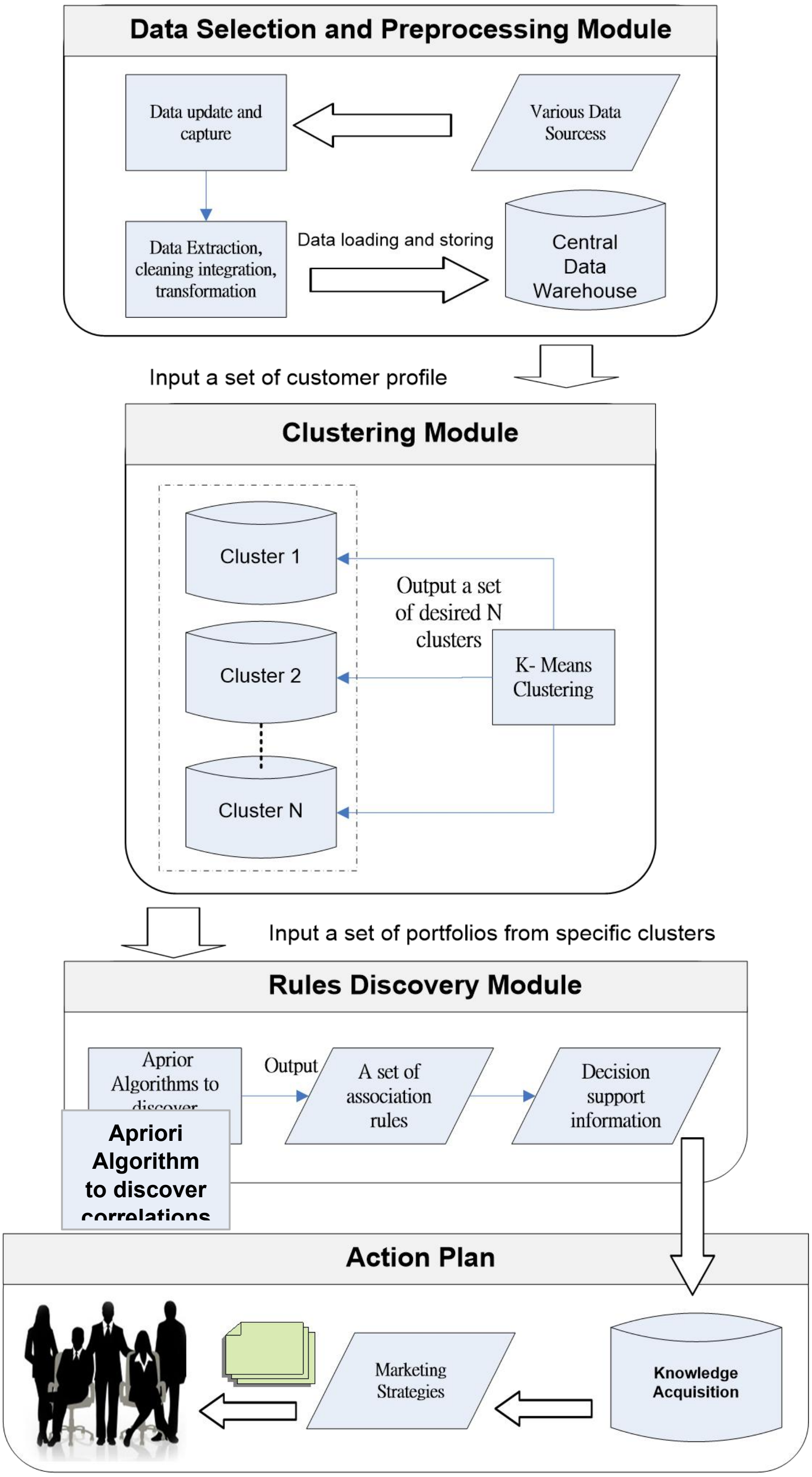

Data Selection and Preprocessing Module
Data update and capture
Various Data Sourcess
Data Extraction, cleaning integration, transformation
Data loading and storing
Central Data Warehouse
Input a set of customer profile
Clustering Module
Cluster 1
Cluster 2
Cluster N
Output a set of desired N clusters
K- Means Clustering
Input a set of portfolios from specific clusters
Rules Discovery Module
Aprior Algorithms to discover
Apriori Algorithm to discover correlations
Output
A set of association rules
Decision support information
Action Plan
Marketing Strategies
Knowledge Acquisition

Figure 5.1 Architecture of FDMM

5.1.1 Data Selection and Pre-processing Module (DSPM)

The Data Selection and Pre-processing Module (DSPM) builds a centralised data warehouse for supporting subsequent data mining tasks and quality information sharing. DSPM can be connected to various data sources, including departmental databases, online server and customer files. The departmental database is a major component of the DSPM and holds a wide range of potentially valuable data including customer profiles, financial investment portfolios and past investment records. The DSPM performs two key processes:

i) The data selection process, and

ii) The data pre-processing (DP).

Possible workflow within the DSPM is illustrated in Figure 5.2.

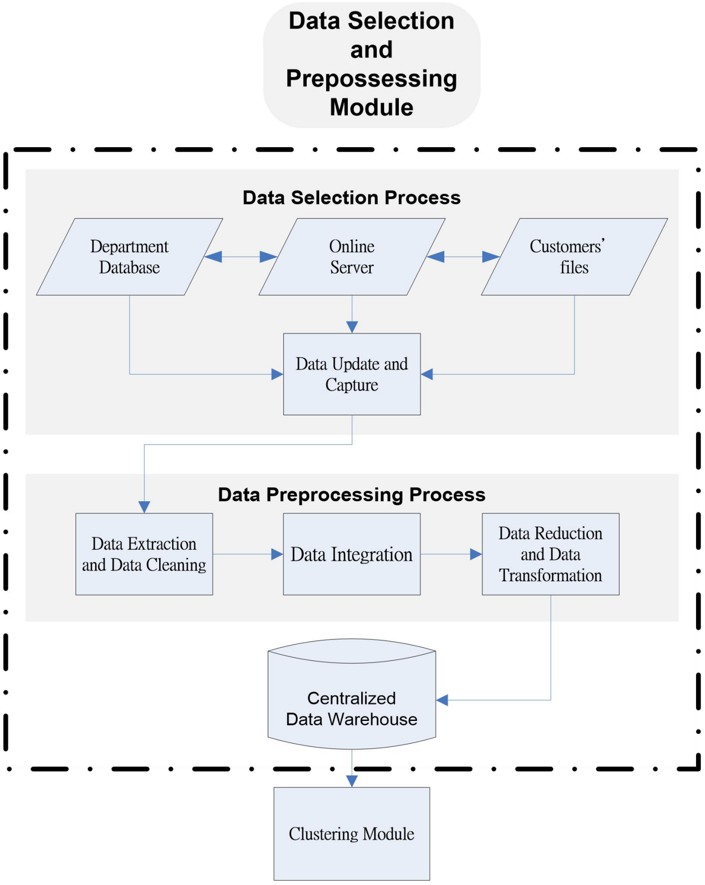

Data Selection and Prepossessing Module
Data Selection Process
Department Database
Online Server
Customers' files
Data Update and Capture
Data Preprocessing Process
Data Extraction and Data Cleaning
Data Integration
Data Reduction and Data Transformation
Centralized Data Warehouse
Clustering Module

Figure 5.2 Workflow of DSPM

5.1.1.1 Data Selection Process

The Data Selection Process (DSP) is used for selecting data from heterogeneous data sources, including departmental databases, online servers of the financial service provider and customer files.

Relevant data is updated in different data sources according to a daily process. Then, the DSP records the type and location of data to be collected. For instance, the movement in the stock market can be collected from the product and research department, while the market information such as market trend, and competitors' analysis can be collected from the marketing department. All updated data are captured and collected. After gathering relevant data from various data sources, the FDMM will proceed to the DP process which plays a significant role in the entire data mining process. It helps ensure the quality of model data.

5.1.1.2 Data Pre-processing (DP)

Data Pre-processing (DP) is an essential part of data mining. The purpose of DP is to ensure the quality of data by 'cleaning' before conducting data mining. In general, data cleaning is conducted by filtering, aggregating and handling missing values. Data cleaning reduces the presence of erroneous and thus harmful data. For example, noisy data, inconsistent data, and missing data can adversely affect the results of data mining.

Data is then integrated from different sources into the centralised data warehouse. Data integration can detect and reduce the presence of data value conflict or inconsistency. Data integration also improves the accuracy and the speed of the data mining processes. For example, all monetary data can be converted into the same currency.

Then all the data is parsed to convert the data into appropriate forms and normalises them into a specific range for mining. Data transformation can enhance the capability of interpreting different data by enabling different types of normalization. Normalization ensures data are converted into the specified data ranges.

Lastly data reduction is conducted in the DP. Data reduction ensures the quality and quantity of the data set, especially when a very large number of transactions are generated in the daily operations of the financial institutions. Only “qualified” data should be retained as it enables the data mining algorithms to be more effective. After data transformation and reduction, data are available in the centralised data warehouse for analysis. The centralised data warehouse can integrate data from multiple data sources, it increases the volume and variety data that could be used for data mining.

5.1.2 Clustering Module (CM)

Data selection and DP prepare data for mining. The purpose of CM is to shorten the processing time for the Rules Discovery Module (RDM). CM should not only improve the efficiency of rule discovery but also improve the ease of rule discovery. The K-means algorithm is applied to segment the aggregated and prepared data into different clusters. The results from the clustering analysis are passed to the RDM for the discovery of cluster-based rules. The detailed workflow of CM is shown in Figure 5.3.

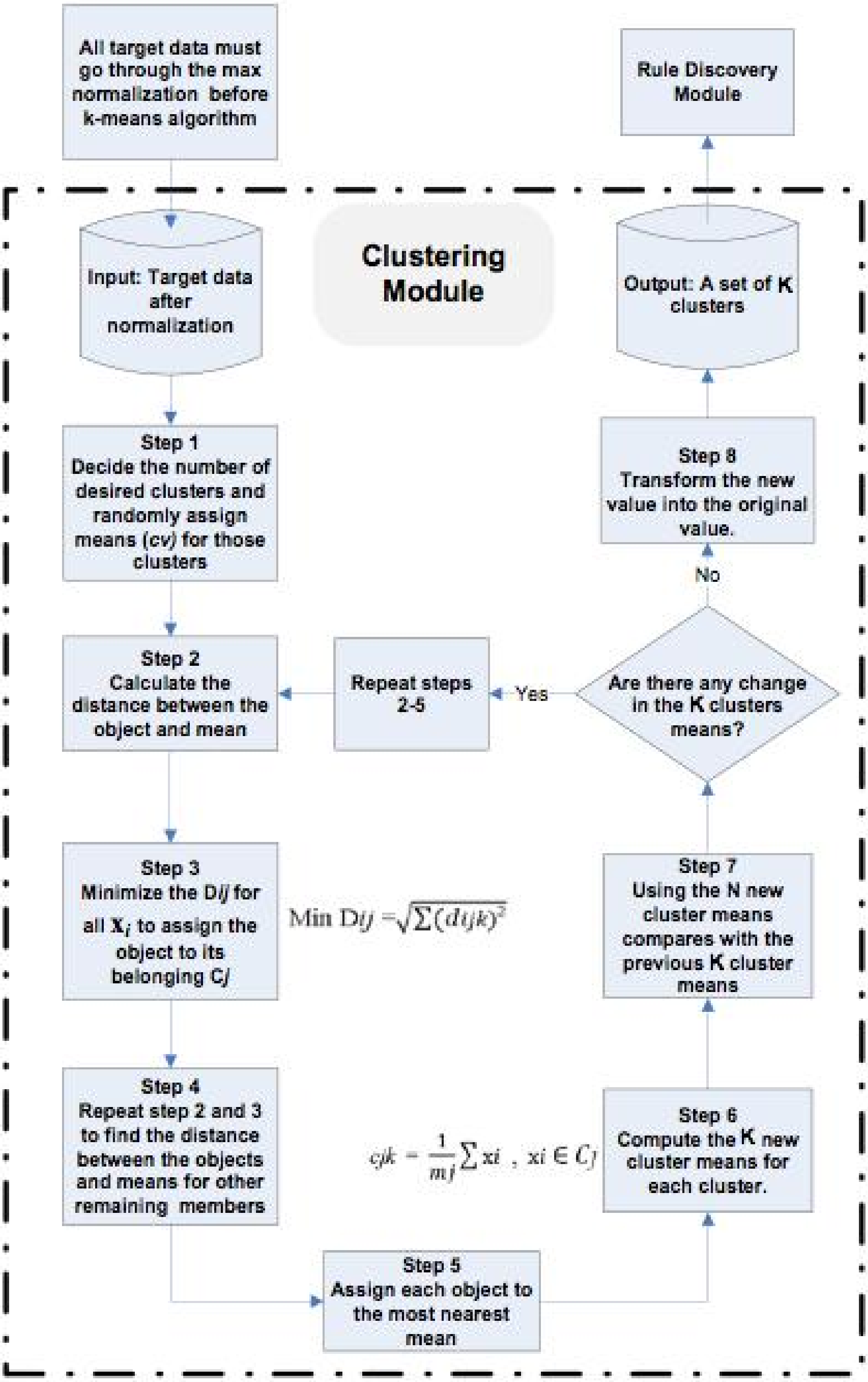

All target data must go through the max normalization before k-means algorithm
Rule Discovery Module
Clustering Module
Input: Target data after normalization
Output: A set of K clusters
Step 1 Decide the number of desired clusters and randomly assign means (cv) for those clusters
Step 8 Transform the new value into the original value.
No
Step 2 Calculate the distance between the object and mean
Repeat steps 2-5
Yes
Are there any change in the K clusters means?
Step 3 Minimize the Dij for all $\mathbf{X}_i$ to assign the object to its belonging Cj
$\text{Min } Dij = \sqrt{\sum (dijk)^2}$
Step 7 Using the N new cluster means compares with the previous K cluster means
Step 4 Repeat step 2 and 3 to find the distance between the objects and means for other remaining members
$cjk = \frac{1}{mj} \sum xi \ , \ xi \in Cj$
Step 6 Compute the K new cluster means for each cluster.
Step 5 Assign each object to the most nearest mean

Figure 5.3 Workflow of CM

The following present details concerning how the K-means algorithm is calculated are presented. Table 5.1 lists the notations involved in the K-means algorithm adopted in the CM.

Table 5.1 Notations for K-means algorithm in CM

Notation	Description
x’	The original value of an object before normalisation
x'_j	The original value of an object in the j^{th} cluster before normalisation
X	The new value of an object after normalisation, where $x = \{x_i \mid i = 1, 2\ldots u\}$
x_i	The i^{th} new value of an object, where $i = \{1, 2\ldots u\}$
x_{ij}	The i^{th} new value of an object in the j^{th} cluster after normalisation
x'_{ij}	The original i^{th} value of an object in the j^{th} cluster before normalisation
U	The number of objects in the data set
C_j	The j^{th} cluster
$\hat{c}$	The initially assigned means, where $\hat{c} = \{\hat{c}_k \mid k = 1, 2\ldots v\}$
c	The mean of all objects
c_{jk}	The k^{th} mean of the j^{th} cluster
v	The number of cluster means
m_j	The number of objects in the j^{th} cluster.
K	The number of clusters
d_{ijk}	Distance between the i^{th} new value of an object in the j^{th} cluster, and the k^{th} mean of the j^{th} cluster
D_{ij}	The minimised Euclidean distance between the i^{th} new value of an object in the j^{th} cluster, and the mean of j^{th} cluster

To reduce redundancy of data in the database, data as input variables that are multi-dimensional are transformed into appropriate mining formats through the max normalisation process before performing the K-means algorithm. The new value of the object is defined by Equation (5.1):

The new value of an object (x_{ij}) = $\frac{\text{original value (x'ij)}}{\text{Max value of x'j}}$ Equation (5.1)

where x' is the original value of the objects, x is the new value of the objects after the max normalization, x_i represents the i^{th} new value of the objects and the number of objects in the data set is represented as u, where x = {x_i ∣ i = 1, 2…u}.

As discussed in chapters 2.6 and 3.4, the K-means algorithm is a suitable choice for the model development and is applied in this study. Nanda *et al.*, 2010 and others such as Jain, 2010; Gan and Wu, 2007, suggest the analysis is conducted in eight steps as follows. (A case study using data collected from Convoy to illustrate how the CM works using K-mean algorithm is described in chapter 5.2).

Step 1:

The first step is to assign the initial N cluster "means" randomly, one for each cluster. The *C* represents the clusters, where *Cj* denotes j^{th} cluster. The set of initial N cluster "means" is indicated as $\hat{c}$ = {$\hat{c}_k$ ∣ k = 1, 2…v}, where k is an input of the base algorithm.

Step 2:

The next step is to identify the distance between the object and the 'mean'. Distance is indicated as d, where d_{ijk} = ∣ x_i —c_{jk} ∣ . The distance for one-dimensional data is defined by Equation (5.2). Equation (5.3) defines the k^{th} mean of the j^{th} cluster (c_{jk}). The distance for multi-dimensional data is measured by the Euclidean distance to minimize the squared distance of each point to its closest centroid, which is expressed as Equation (5.4).

$$d_{ijk} = \mid x_{ij} - c_{jk} \mid \qquad \text{Equation (5.2)}$$

$$c_{ijk} = \frac{1}{m_j} \sum x_{ij} \qquad \text{Equation (5.3)}$$

$$Min\ D_{ij} = \sqrt{\sum (d_{ijk})^2} \qquad \text{Equation (5.4)}$$

Step 3:

In this step, the distance needs to be minimized, i.e. D_{ij} for all x_i in order to assign the objects to their belonging C_j.

Step 4:

Repeat steps 2 - 3 to find the distance between the objects and the means for the remaining objects.

Step 5:

Assign each object to the cluster with the mean closest to that object. When all data points fall in a belonging cluster, this step is completed and an early clustering is done.

Step 6:

In this step, the K new cluster means need to be computed. The updated means are set to be c. The m represents the number of objects within C_j. Here, it must be noted that $\hat{c}$ is not equal to c. Table 3.4, the mean of the j^{th} cluster is defined by Equation 5.5.

$$c_{jk} = \frac{1}{mj} \sum xi \ , xi \in C \qquad \text{Equation (5.5)}$$

Step 7:

Scan the N cluster means and compare them with the previous means. If the means of K clusters are changed, steps 2-5 need to be repeated. In contrast, if the means of K clusters do not change, the algorithm is done.

Step 8:

Convert the results (new value after normalisation) into original values by using the conversion expressed below:

$$x'_{ij} = \text{Max. } x'_{j} \times x_{ij} \qquad \text{Equation (5.6)}$$

After the computation of K-means algorithm, *j* clusters are classified. The objects within a cluster tend to be more similar compared to those belonging to different clusters. Clustering groups data on the basis of self-similarity. The set of clusters will then be used for RDM processing. To enhance the precision of the clustering, attributes about investment behaviour were evaluated and filtered by the statistical model discussed in chapter four and only significant attributes were considered in this model for clustering.

5.1.3 Rules Discovery Module (RDM)

In this stage, the RDM aims at discovering relationships in a specific cluster and at generating useful rules. In this module, the Apriori algorithm, as discussed in chapters 2.5.1 and 3.4.2, was appropriate for model development and applied to find frequency patterns, correlations and associations. Using the rules identified, financial institutions can identify a set of financial investment products that customers in a specific cluster are likely to prefer. After the generation of rules, those rules were extracted for evaluation and strategy formulation by senior management. For example, the marketing department can use such rules for designing cluster (customer)-specific promotion strategies. The detailed workflow for RDM is shown in Figure 5.4.

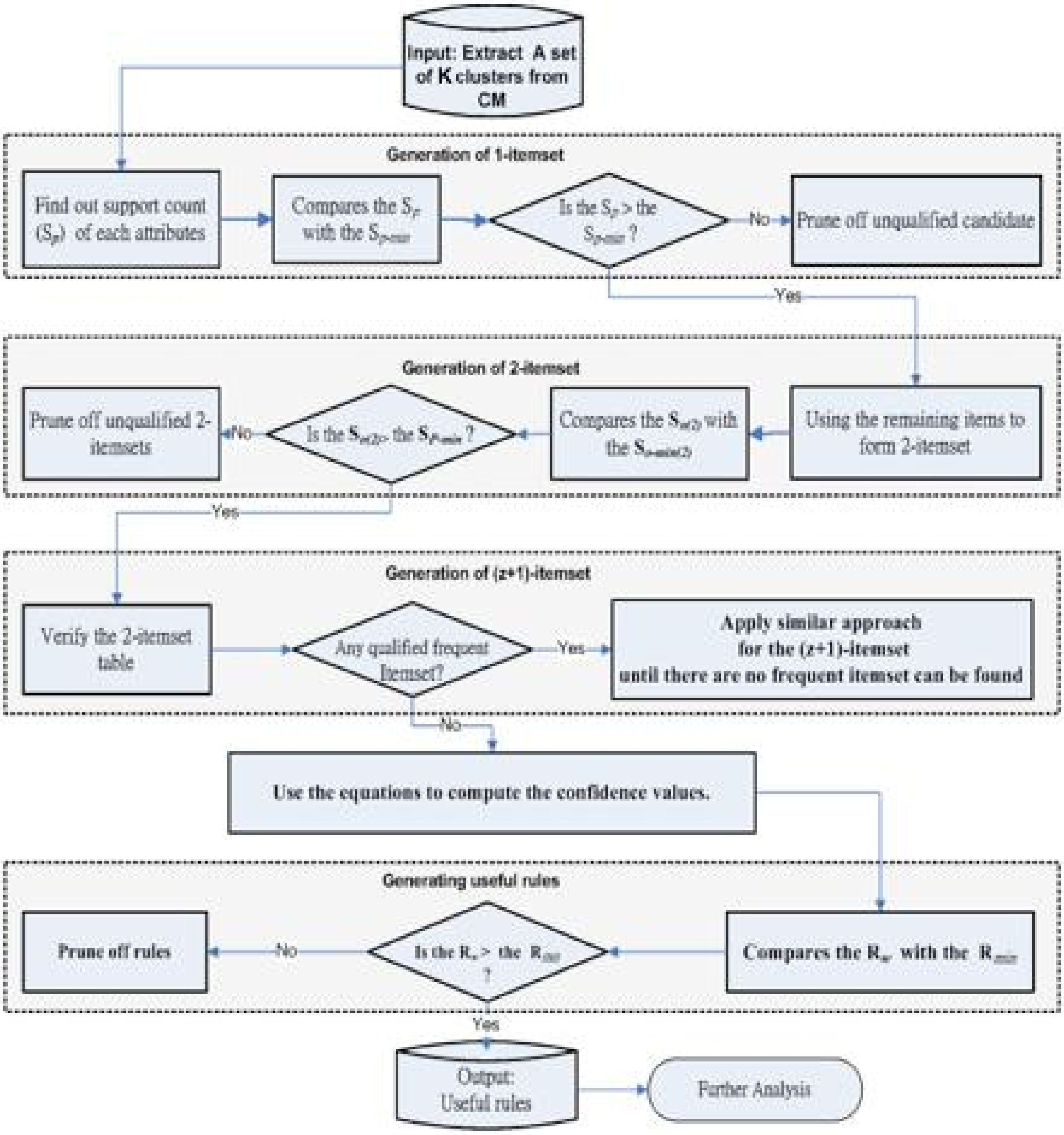

Input: Extract A set of K clusters from CM
Generation of 1-itemset
Find out support count (S_p) of each attributes
Compares the S_p with the $S_{p\text{-}min}$
Is the S_p > the $S_{p\text{-}min}$?
No
Prune off unqualified candidate
Yes
Generation of 2-itemset
Using the remaining items to form 2-itemset
Compares the $S_{u(2)}$ with the $S_{p\text{-}min(2)}$
Is the $S_{u(2)}$> the $S_{p\text{-}min}$?
No
Prune off unqualified 2-itemsets
Yes
Generation of (z+1)-itemset
Verify the 2-itemset table
Any qualified frequent Itemset?
Yes
Apply similar approach for the (z+1)-itemset until there are no frequent itemset can be found
No
Use the equations to compute the confidence values.
Generating useful rules
Compares the R_w with the R_{min}
Is the R_w > the R_{min} ?
No
Prune off rules
Yes
Output: Useful rules
Further Analysis

Figure 5.4 Workflow of RDM

5.1.3.1 Apriori algorithm

The Apriori algorithm consists of two phases: mining of frequent item-sets and generation of association rules. Table 5.2 lists the notations involved in the Apriori algorithm adopted in the RDM.

Table 5.2 Notations for Apriori Algorithm in RDM

Notation	Description
T	A set of transactions, where $T = \{ Jp \mid p = 1, 2 \ldots q\}$
T_a	The a^{th} transaction, where $Ta = \{Ta \mid a = 1, 2 \ldots b\}$
b	The number of the transactions
J	Subset of T
A	The attributes of the data set.
Jp	The p^{th} attribute ($p = p'$), where $J_p = \{ J_p \mid p = 1, 2 \ldots q\}$
q	The number of attributes
S_p	The support count of each attribute
$S_{ap'}$	Indicator for whether an attribute present in a transaction, where $S_{ap'} = \begin{cases} 0, if\ absence \\ 1, if\ presence \end{cases}$
$S_{p\text{-}min}$	The minimum support count of all attributes and potential frequent itemsets
$S_{(2)}$	The support count of 2-itemset
L_o	The o^{th} 2-itemset
$S_{o(2)}$	The support count of the o^{th} 2-itemset
z_a	The number of candidates of the next itemset table, where $z_a = z_{p+1}$
Z_p	The number of candidates in the present itemset table
$S_{g(z+1)}$	The support count of the g^{th} (z+1)-itemset combination (A_{ap}, A_{ap}…)
$S_{p\text{-}condition}$	The support count of the 1-itemset condition of the rules for 2-itemset
$S_{ap\text{-}condition}$	The support count of the 1-itemset condition of the rule for (z+1)-itemset
$S_{o(2)\text{-}condition}$	The support count of the 2-itemset condition of the rule for (z+1)-itemset

$S_{g(z)\text{-}condition}$	The support count of the z-itemset condition of the rule for (z+1)-itemset
$S_{p\text{-}result}$	The support count of the 1-itemset result of the rules for 2-itemset
$S_{ap\text{-}result}$	The support count of the 1-itemset result of the rule for (z+1)-itemset
$S_{o(2)\text{-}result}$	The support count of the 2-itemset result of the rule for (z+1)-itemset
$S_{g(z)\text{-}result}$	The support count of the z-itemset result of the rule for (z+1)-itemset
R	The confidence value of rule
R_w	The confidence value of the w^{th} association rule
$R_{\text{-}min}$	The threshold confidence value of all association rules.
I	The lift ratio of association rule
I_{mp}	The lift ratio of rules of the mp^{th} association rule

The Apriori algorithm is applied as follows:

Step 1a

Transform a set of transactions from a cluster: This set of transactions denotes as T, where T_a represents the a^{th} transaction. The number of transactions indicated as b, where $\text{T}a = \{\text{T}a \mid a = 1, 2 \ldots b\}$. Each transaction consists of different attributes. Here J denotes attributes, where p represents the p^{th} attribute. The number of attributes indicates as q, where $J_p = \{ J_p \mid p = 1, 2 \ldots q\}$. J is a subset of T. Hence, $\text{T} = \{ \text{J}p \mid p = 1, 2 \ldots q\}$. Note that if an itemset is frequent, any of its subsets is frequent as well.

The first step is to use the T to form a table to find out the frequency of occurrences (support count) of different attributes in transactions. The support count of each attribute denotes as S_p, where $S_p = \sum S_{ap'}$, $p = p'$. The $S_{ap'}$ is used to indicate the absence (0) or presence (1) of an attribute in a transaction. The predetermined threshold support count of all attributes and potential itemset are set to be $S_{p\text{-}min}$.

Step 1b:

S_p are compared with the $S_{p\text{-}min.}$ A candidate is retained only when it is equal or greater than the predefined threshold support count ($S_{p\text{-}min}$). If the S_p is smaller than the $S_{p\text{-}min}$, the corresponding candidates will be removed.

$$S_p \geq (S_{p\text{-}min}) \qquad \text{Equation (5.7)}$$

Step 2a:

Merge the remaining candidates to form an itemset with two items: The combination of these two itemsets is called L where L_o represents o^{th} 2-itemset. Create a 2-itemset table and find the support count of these 2-itemset. The support count of 2-itemset is $S_{(2)}$, where $S_{o(2)}$ indicates the support count of the o^{th} 2-itemset combination. Scan for support counts of these 2-itemsets ($S_{(2)}$) in the 2-itemset table.

Step 2b:

Compare the support count of the 2-itemset ($S_{(2)}$) with the predetermined threshold support count ($S_{p\text{-}min}$) of the candidates and prune off the unqualified 2-itemset candidates: The 2-itemset table contains only the 2-itemsets which have the support count ($S_{o(2)}$) equal to or greater than the threshold support($S_{p\text{-}min}$), i.e. the o^{th} 2-itemset is retained only when ($S_{o(2)}$) $\geq$ ($S_{p\text{-}min}$).

Step 3a:

Verify the 2-itemset table whether there are any qualified combinations.

Step 3b:

If there are still qualified combinations, the algorithm needs to be continued. Apply a similar approach as in step 2 to form a table for (z + 1)-itemset, where z is set to be 2 initial that is (2+1) as 3 itemset, (3+1) as 4-itemset etc. Then $z_a = z_{p+1}$ for further combinations. The z_a is the number of candidates of the next itemset table and z_p is the number of present candidates. Here, the support count of the g^{th} (z + 1)-itemset combination is set to be $S_{g(z+1).}$ The algorithm is said to be continued until no more frequent itemsets can be found.

Step 4:

Verify if the 2-itemset or (z+1) combination is valid. The 'condition' is the cause of the rule and it can be any segment of the itemset. It is necessary that the number of itemset combinations with condition must be smaller than the number of itemset combinations, equal to or greater than 1. The confidence value of each rule denotes as R, where R_w indicates the w^{th} confidence value of a rule. The equation of confidence values (R_w) are defined in Table 5.3.

Table 5.3 Equations of confidence value

Confidence Value (R_w) for 2-itemset candidates.	
(with 1-itemset condition)	Confidence(R_w) = $S_{o(2)}$ / $S_{p\text{-}condition}$
Confidence Value (R_w) for (z + 1)-itemset candidates.	
(with 1-itemset condition)	Confidence(R_w) = $S_{g(z+1)}$ / $S_{p\text{-}condition}$
(with 2-itemset condition)	Confidence(R_w) = $S_{g(z+1)}$ / $S_{o(2)\text{-}condition}$
(with z-itemset condition)	Confidence(R_w) = $S_{g(z+1)}$ / $S_{g(z)\text{-}condition}$

Step 5:

After computation, the confidence value (R_w) is compared with the threshold confidence value ($R_{\text{-}min}$). The confidence value which is smaller than the threshold confidence value. The remaining itemsets are the interesting rules with the desired level of quality.

Confidence value (R_w) $\geq$ threshold confidence value (R_{-min}) Equation (5.8)

In addition to the confidence value, the lift ratio of the rules is used to show how well the rules generated predict the results as compared to a single itemset. The equation is shown in Table 5.4.

Table 5.4 Equations of lift ratio

<table>
<tr><td colspan="2">Lift Ratio of the Rules (I_{mp}) for 2-itemset candidates.</td></tr>
<tr><td>(with 1-itemset result)</td><td>Lift ratio (I_{mp}) = $S_{o(2)}$ / $S_{p\text{-}result}$</td></tr>
<tr><td colspan="2">Lift Ratio of the Rules (I_{mp}) for (z + 1)-itemset candidates.</td></tr>
<tr><td>(with 1-itemset result)</td><td>Lift ratio (I_{mp}) = $S_{g(z+1)}$ / $S_{ap\text{-}result}$</td></tr>
<tr><td>(with 2-itemset result)</td><td>Lift ratio (I_{mp}) = $S_{g(z+1)}$ / $S_{o(2)\text{-}result}$</td></tr>
<tr><td>(with z-itemset result)</td><td>Lift ratio (I_{mp})) = $S_{g(z+1)}$ / $S_{g(z)\text{-}result}$</td></tr>
</table>

This is not the end of the algorithm, as the incremental data are continuously recorded and entered to the algorithm of RDM. The existing data and the newly formed data are loaded into the algorithm for rules mining. After the calculation of the algorithm, new rules may be found as the incremental data may show some preferences, customers' requirements etc. These implications point out rules are needed and need to be evaluated.

5.1.3.2 Rules evaluation

All extracted rules go through the rules evaluation process to assess their feasibility. The rules evaluation process helps the RDM define more knowledge about customer behaviour with reference to extracted rules. Users can adjust the rules on the basis of customers' requirements, latest market trends, professional industry knowledge and experience. After the modification of the rules, data are gathered and loaded into the model continuously. Hence, the entire proposed model is a continuous process without termination. This allows the knowledge to be continuously improved and to create higher customer satisfaction. To better illustrate the value of this model, a case study was conducted and is discussed in chapter 5.2.

5.1.3.3 Rules Classification

The purpose of the RDM is to discover useful rules that will help improve the understanding of investment behaviour in the financial industry. For example, a rule regarding the investment behaviour of younger customers may be, “that they would be more willing to buy riskier financial products compared to older investors”. Such information might, for example, be useful for managers of product and research departments to devise respective portfolios and to know to what types of customers they should be promoted to. Some derived rules may be trivial, they appear as common sense and their influence is minimal, for instance, the higher the risk of a financial product, the higher its return. Thus this known rule is less likely to generate new value.

Some data may not be easily interpreted by rules. Users may not understand the correlations from certain results by common sense alone. Such rules present a fact to users but cannot provide any insight into customer behaviour. It is sometimes not easy to understand why investors select a specific product. However, managers and financial consultants can uncover the area requiring improvement in the portfolio based on the rules extracted. This is because such rules enable managers to develop new concepts of portfolio coverage and show improvement in these areas.

5.2 Case Study: Implementation of FDMM in the case company

After identifying and defining the challenges faced by Convoy as presented in chapter 3.6, it has become clear that Convoy needs to find feasible solutions to tackle these challenges. A prototype based on the system architecture of FDMM was developed in XLMiner^{TM} and implemented at Convoy. XLMiner^{TM} is a simple and user-friendly data mining add-in for Excel to effectively manipulate the proposed algorithms. Based on the case information of Convoy, major problems for the company are poor information sharing and data management. The company does not apply any data mining tools to manage their business data. Therefore, it is difficult to identify any patterns and relationships in the data sets. Faced with overwhelming amounts of business data, it needs a discovery-driven data analysis technology to improve data management. In order to address these problems, the company needs to:

i) improve information sharing in the organization;

ii) enhance data management; and

iii) increase customer satisfaction.

The company thus adopted the proposed model and implemented changes based on the findings from this research to identify the best solutions for improving its existing workflow. The details of the implementation of the proposed FDMM are presented below.

5.2.1 Data preparation and transformation

The objectives of FDMM are to identify the influencing factors on investments and gain further understanding of investor behaviour. Referring to the results discussed in chapter 4.3, it was found that age, annual income and investment experience are the most important attributes affecting Hong Kong and mainland Chinese investors. These major attributes are therefore considered in the proposed model to predict customers' preferences for financial investment portfolios. In this case study, 11,700 sets of data relating to Convoy's most active customers were obtained from Convoy and chosen for simulation. A sample of the collected data is shown in Figure 5.5, in which each row represents a customer profile. Table 5.5 shows the range of the major attributes influencing investment behaviours.

Client ID	Age	Investment	Investment experience	Portfolio
1	26	30000	4	Life Insurance, MPF, China Growth Fund
2	30	63000	0	Fixed Deposit, Currency Linked Deposit, Technology Stock
3	40	45000	3	Currency Linked Deposit, Global Equity Fund, Life Insurance
4	24	23000	0	Technology Stock, Korean Equity Fund, MPF
5	33	32000	8	Emerging Market Equity Fund, Currency Linked Deposit
6	52	66000	3	Fixed Deposit, Material Stock, MPF, Life Insurance
7	38	38000	5	Energy Stock, China Growth Equity Fund, Medical Insurance, MPF
8	25	28000	1	Stocks, Currency-linked Deposit, Mutual Funds, MPF
9	46	55000	6	Fixed Deposit, Stocks, Insurance, MPF
10	29	26000	1	Stocks, Insurance, Currency-linked deposit, MPF
11	30	29000	9	Stocks, Mutual funds, Insurance, MPF
12	44	60000	7	Mutual Funds, Insurance, Currency-linked deposit
13	57	49000	13	Fixed Deposit, Cash Reserve, Insurance
14	48	56000	7	Fixed Deposits, Currency linked deposit, Insurance, Malaysia Equity Fund, MPF
15	21	14000	0	Cash Reserve, Insurance. Pacific Technology Equity Fund, MPF
16	29	36000	10	Fixed Deposited, Investment Linked Insurance, Energy Stock
17	50	49000	7	Cash Reserve, Russia Equity Fund, MPF, Life Insurance
18	35	29000	5	Material Stock, Emerging Europe Fund, Global Equity Find
19	36	60000	8	Investment Linked Insurance, Technology Stock
20	36	98000	13	China Growth Fund, Korea Equity Fund, Technology Stock

Figure 5.5 Raw data collected from case company for FDMM implementation

Table 5.5 Range of the major attributes

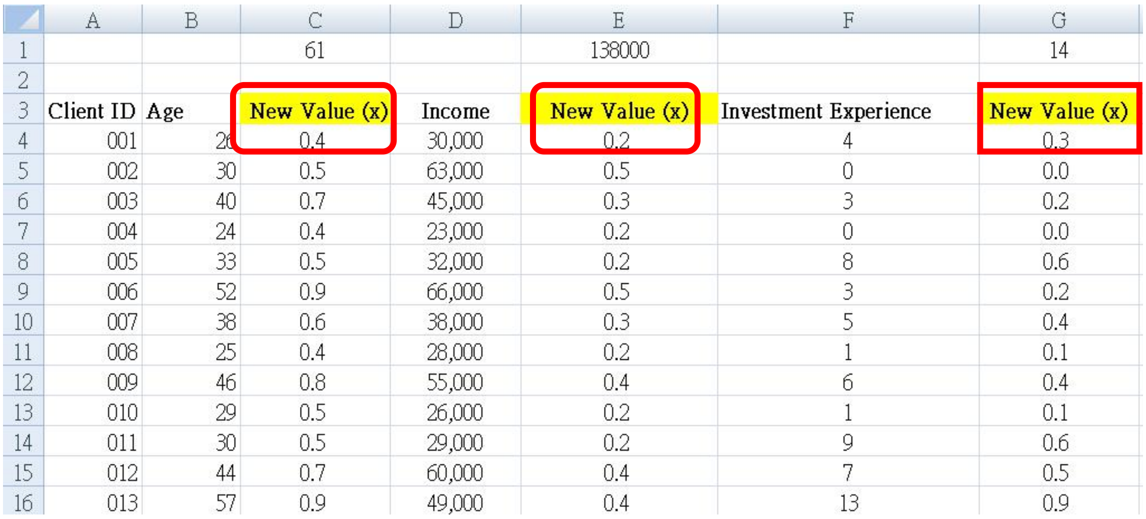
Range		
Variables	**Maximum**	**Minimum**
Age	61	20
Average Monthly Income (HKD)	138,000	12,700
Investment Experience	14	0

Among the 11,700 cases average monthly income ranges from HKD12,700 to HKD138,000; investment experience ranges from 0 to 14 years and finally age ranges from 20 to 61 years. Income has a larger range than investment experience and age. To ensure all variables are converted into the same mining form, all variables must go through max normalization. The values of data after max normalisation are shown in Figure 5.6. To take Client ID 001 as an example, max normalisation was conducted using the following equations:

The new value of age for Client ID 001 (x_{11}) = $\frac{\text{original value (x'11)}}{\text{Max value of x'1}} = \frac{26}{61} = 0.4$

The new value of average monthly income for Client ID 001 (x_{12}) = $\frac{\text{original value (x'12)}}{\text{Max value of x'2}} = \frac{30{,}000}{138{,}000} = 0.2$

The new value of investment experience for Client ID 001 (x_{13}) = $\frac{\text{original value (x'13)}}{\text{Max value of x'3}} = \frac{4}{14} = 0.3$



	A	B	C	D	E	F	G
1			61		138000		14
2							
3	Client ID	Age	New Value (x)	Income	New Value (x)	Investment Experience	New Value (x)
4	001	26	0.4	30,000	0.2	4	0.3
5	002	30	0.5	63,000	0.5	0	0.0
6	003	40	0.7	45,000	0.3	3	0.2
7	004	24	0.4	23,000	0.2	0	0.0
8	005	33	0.5	32,000	0.2	8	0.6
9	006	52	0.9	66,000	0.5	3	0.2
10	007	38	0.6	38,000	0.3	5	0.4
11	008	25	0.4	28,000	0.2	1	0.1
12	009	46	0.8	55,000	0.4	6	0.4
13	010	29	0.5	26,000	0.2	1	0.1
14	011	30	0.5	29,000	0.2	9	0.6
15	012	44	0.7	60,000	0.4	7	0.5
16	013	57	0.9	49,000	0.4	13	0.9

Figure 5.6 Variables after normalisation

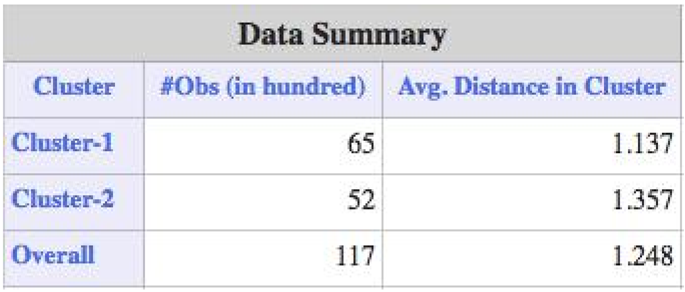
5.2.2 Clustering Module

This module segments customers into appropriate clusters based on the three variables, namely age (see Column B in Figure 5.5), income (see Column C in Figure 5.5) and investment experience (see Column D in Figure 5.5). Information concerning portfolio (see Column E in Figure 5.5) was used for generating rules in the Rules Discovery Module afterwards.

As discussed in chapters 2.6.1.1 and 3.4.1 on clustering analysis in the literature and the selection of K-means for this study, as well as in chapter 5.1.2 regarding its application to build the FDMM, step-by-step procedures were undergone for the Clustering Module and the results are presented below:

5.2.2.1 Results of clustering module

To begin the algorithm, the number of K clusters and the size of the dataset needs to be decided. To run the K-means clustering algorithm for a range of K values, from two to seven in this study and compare the results, in the dataset (*i* = 1, 2…11,700), the number of desired clusters (K) is determined as two (see Figure 5.7) based on the distance feature and results of elbow method. To determine the optimal number of clusters, elbow method as advocated by scholars (Kodinariya and Makwana, 2013; Madhulatha, 2012; Bholowalia and Kumar, 2014) was adopted that the sum of squared errors (SSE) for each value of k was calculated and a line chart of the SSE was plotted in Figure 5.8. Elbow method is a visual method. According to Figure 5.8, the best number of clusters for this dataset is two as at which the SSE decreases abruptly and then goes down very slowly after that, and also the graph has a clear elbow at k=2.



Data Summary

Cluster	#Obs (in hundred)	Avg. Distance in Cluster
Cluster-1	65	1.137
Cluster-2	52	1.357
Overall	117	1.248

Figure 5.7 Data summary of K-means clustering (K=2)

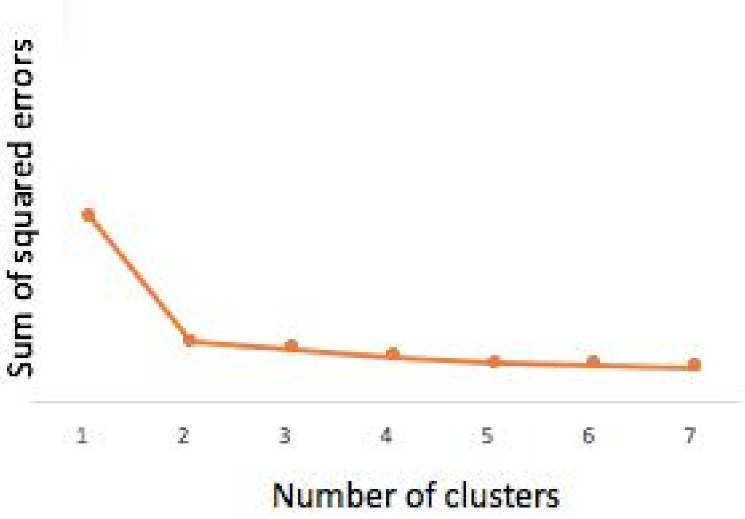
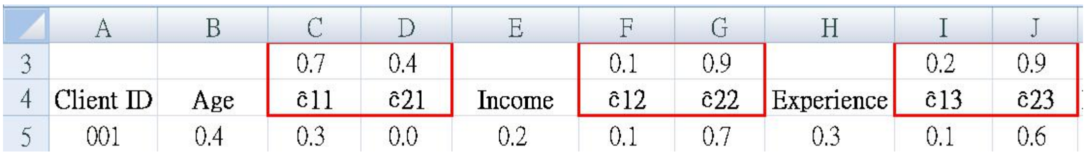


Figure 5.8 The results of Elbow method

To demonstrate the working process of k-means algorithm, customer 001 is taken as an example. Initial cluster means are assigned as set of $\hat{c} = \{\hat{c}k \mid k = 1, 2\ldots6\}$. Figure 5.9 shows the randomly assigned means for the two clusters:

C1 $\{\hat{c}11: 0.7;\ \hat{c}12: 0.1;\ \hat{c}13: 0.2\}$

C2 $\{\hat{c}21: 0.4;\ \hat{c}22: 0.9;\ \hat{c}23: 0.9\}$



	A	B	C	D	E	F	G	H	I	J
3			0.7	0.4		0.1	0.9		0.2	0.9
4	Client ID	Age	ĉ11	ĉ21	Income	ĉ12	ĉ22	Experience	ĉ13	ĉ23
5	001	0.4	0.3	0.0	0.2	0.1	0.7	0.3	0.1	0.6

Figure 5.9 Randomly assigned means for clusters

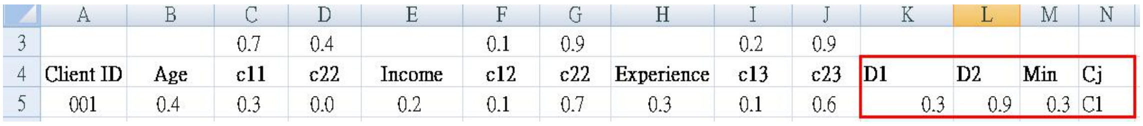
The Euclidean distance between the cluster means and the objects can be computed as follows using Equation (5.4).

The minimised distance of D11 = $\sqrt{\sum (0.7 - 0.4\,)^2 + (0.1 - 0.2\,)^2 + (0.2 - 0.3\,)^2}$ = 0.3

The minimised distance of D12 = $\sqrt{\sum (0.4 - 0.4\,)^2 + (0.9 - 0.2\,)^2 + (0.9 - 0.3\,)^2}$ = 0.9

After the calculation of distance of both clusters, distance needs to be minimized for assigning the ID 001 to the smallest one.

To assign ID 001 to the most minimized distance, the ID 001 is assigned to C1 as shown in Figure 5.10. Steps 2 – 5 mentioned in chapter 5.1.2 were repeated to find the distance between the objects and the means for the other 116 data.



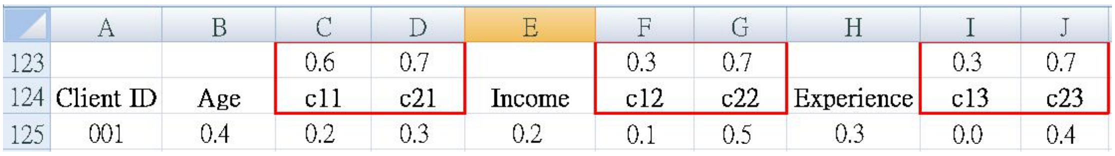
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
3			0.7	0.4		0.1	0.9		0.2	0.9				
4	Client ID	Age	c11	c22	Income	c12	c22	Experience	c13	c23	D1	D2	Min	Cj
5	001	0.4	0.3	0.0	0.2	0.1	0.7	0.3	0.1	0.6	0.3	0.9	0.3	C1

Figure 5.10 Assign object to the cluster with minimum distance

The two cluster means were computed by Equation (5.3) with the results below. The updated means of the two clusters are shown in Figure 5.11.

C1 {c11: 0.6; c12: 0.3; c13: 0.3}

C2 {c21: 0.7; c22: 0.7; c23: 0.7}



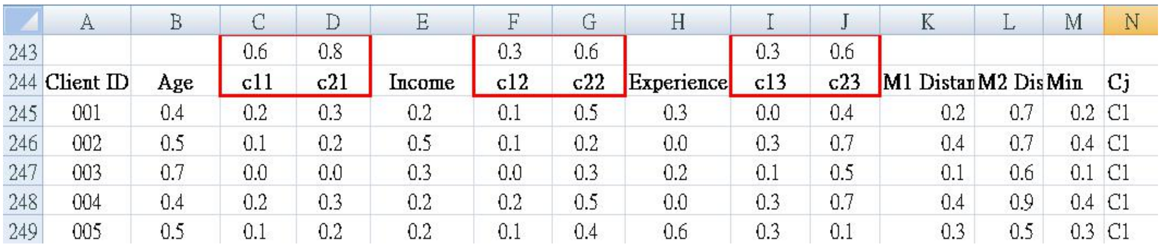
	A	B	C	D	E	F	G	H	I	J
123			0.6	0.7		0.3	0.7		0.3	0.7
124	Client ID	Age	c11	c21	Income	c12	c22	Experience	c13	c23
125	001	0.4	0.2	0.3	0.2	0.1	0.5	0.3	0.0	0.4

Figure 5.11 The updated means of clusters

By comparing the two new cluster means with their previous ones, it can be seen that the means of the 2 clusters changed, and thus steps 2－5 were repeated another two times. The means of the clusters after repeating steps 2 – 5 one more time are shown below (see Figure 5.12).

C1 {c11: 0.6; c12: 0.3; c13: 0.3}

C2 {c21: 0.8; c22: 0.6; c23: 0.6}



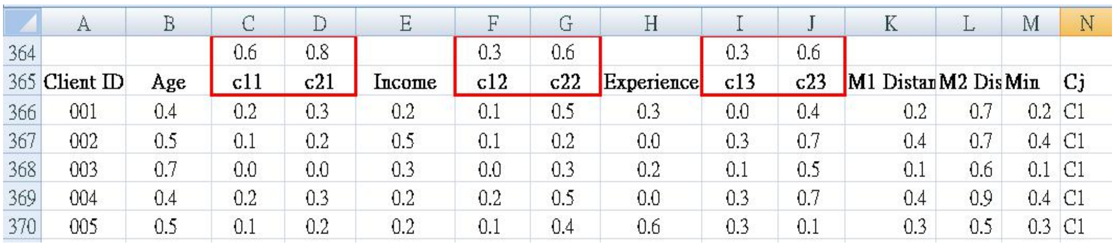
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
243			0.6	0.8		0.3	0.6		0.3	0.6				
244	Client ID	Age	c11	c21	Income	c12	c22	Experience	c13	c23	M1 Distan	M2 Dis	Min	Cj
245	001	0.4	0.2	0.3	0.2	0.1	0.5	0.3	0.0	0.4	0.2	0.7	0.2	C1
246	002	0.5	0.1	0.2	0.5	0.1	0.2	0.0	0.3	0.7	0.4	0.7	0.4	C1
247	003	0.7	0.0	0.0	0.3	0.0	0.3	0.2	0.1	0.5	0.1	0.6	0.1	C1
248	004	0.4	0.2	0.3	0.2	0.2	0.5	0.0	0.3	0.7	0.4	0.9	0.4	C1
249	005	0.5	0.1	0.2	0.2	0.1	0.4	0.6	0.3	0.1	0.3	0.5	0.3	C1

Figure 5.12 Means of clusters after repeating steps 2-4 once

The means of the clusters after repeating steps 2 - 5 a second time are shown below and in Figure 5.13.

C1 {c11: 0.6; c12: 0.3; c13: 0.3}

C2 {c21: 0.8; c22: 0.6; c23: 0.6}



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
364			0.6	0.8		0.3	0.6		0.3	0.6				
365	Client ID	Age	c11	c21	Income	c12	c22	Experience	c13	c23	M1 Distan	M2 Dis	Min	Cj
366	001	0.4	0.2	0.3	0.2	0.1	0.5	0.3	0.0	0.4	0.2	0.7	0.2	C1
367	002	0.5	0.1	0.2	0.5	0.1	0.2	0.0	0.3	0.7	0.4	0.7	0.4	C1
368	003	0.7	0.0	0.0	0.3	0.0	0.3	0.2	0.1	0.5	0.1	0.6	0.1	C1
369	004	0.4	0.2	0.3	0.2	0.2	0.5	0.0	0.3	0.7	0.4	0.9	0.4	C1
370	005	0.5	0.1	0.2	0.2	0.1	0.4	0.6	0.3	0.1	0.3	0.5	0.3	C1

Figure 5.13 Means of clusters after repeating steps 2-5 twice

After transformation, the original values of the cluster were:

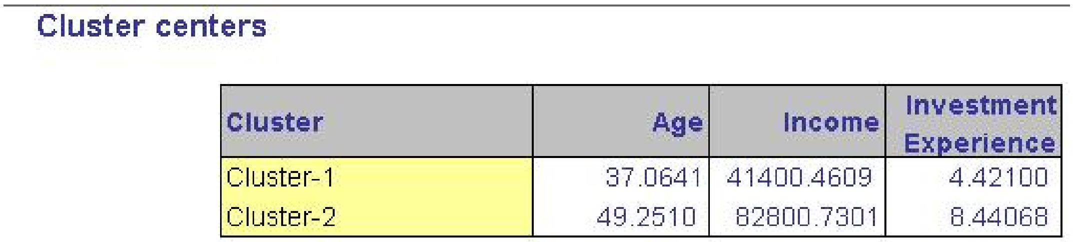
C1 {c11: 0.6×61 = 36.6; c12: 0.3×1,380,000 = 414,000; c13: 0.3 × 14 = 4.2}

C2 {c21: 0.8×61 = 48.8; c22: 0.6×1,380,000 = 828,000; c23: 0.6 × 14 = 8.4}

5.2.2.2 Interpretation of the resulting clusters

After clustering, there are 65 hundred observations (customers) in Cluster 1 and 52 hundred observations (customers) in Cluster 2 (see Figure 5.7). Figure 5.7 shows the number of observations (customers) in each cluster and the average distance between observations and the cluster’s mean. The average distance in Cluster 1 (i.e. 1.137) is smaller than that in Cluster 2 (i.e. 1.357), which means that observations (customers) in Cluster 1 have more similar characteristics when compared with Cluster 2.

The characteristics of each cluster in terms of age, annual income (in HKD) and investment experience (in years), of each cluster can be interpreted as described below. The cluster means of two clusters are shown in Figure 5.14. Table 5.6 shows the investment horizon of both clusters.



Cluster centers

Cluster	Age	Income	Investment Experience
Cluster-1	37.0641	41400.4609	4.42100
Cluster-2	49.2510	82800.7301	8.44068

Figure 5.14 Results of two cluster means

Table 5.6 Investment horizon

Period of investment horizon	**Total number of customers (Cluster 1)** *(in hundreds)*	**Total number of customers (Cluster 2)** *(in hundreds)*
1-3 years	2	8
3-5 years	4	14
5-10 years	7	21
10-20 years	13	5
20 years or above	39	4

Cluster 1:

From Figure 5.7, the results indicate that the total average distance of cluster 1 is smaller than that of cluster 2. This implies that the behaviour of variables in cluster 1 is more similar than in cluster 2. Thus, cluster 1 is selected to generate rules in the later stage.

With reference to Figure 5.14, cluster 1 has a relatively low value for income compared to cluster 2. The mean of income in this group is HKD414,000. The means for age and investment experience are 37 years and 4 years, respectively. Age ranges from 20 to 58 years. Based on the clustering results, it can be concluded that customers of cluster 1 are generally younger than those of cluster 2 and are less experienced in investment.

Over half the customers in this cluster 1 are more likely to seek a long-term investment (20 years or above) (see Table 5.6). Table 5.7 shows the description of investment linked assurance schemes offered by the company. There are a significant proportion of customers who prefer the S2 investment and almost half the customers selected S4. These two types of investment styles are riskier compared with S1. Thus, this cluster can be considered more aggressive than cluster 2. The results provide an insight into the marketing implications for this group. The company can identify such a group as potentially valued customers.

Table 5.7 Descriptions of investment linked assurance scheme

Code	**Name**	**Description**
S1	Dynamic Evergreen	Medium risk
S2	Dynamic Growth	Medium to high risk
S4	Global Opportunity	High risk

Cluster 2:

There are 52 hundred observations in cluster 2 (see Figure 5.7). This cluster has a remarkably high mean income, namely HKD 828,000, which is twice as high as the mean income in cluster 1 (see Figure 5.14). Besides, the means of age and investment experience of this cluster are 49 years and 8 years, respectively. The higher mean age of customers comes with higher investment experience on average. Previous investment experience might provide such customers with more risk awareness. Thus, this group tends to diversify their portfolios. The main bulk of the portfolios consist of more than 75% of bond funds, such as global and US bonds, which are expected to seek steady but slow returns. Therefore, the main portfolio type selected is S1. Thus, the company can offer a conservative portfolio to this group of customers.

5.2.3 Rules Discovery Module

This module aims at applying association rules to discover hidden patterns in the segmented clusters. Cluster 1 is selected to generate useful association rules because the mean distance to the centroid in Cluster 1 (i.e. 1.137) is shorter than that in Cluster 2 (i.e. 1.357), and Cluster 1 indicates more similarity regarding behaviour variables as compared with Cluster 2. By extracting the useful rules from this cluster, the company can gain a better understanding of customers' investment preferences in an efficient, effective as well as systematic manner, followed by customizing the best suited portfolios for such customers, as well as achieving high customer satisfaction.

Table 5.8 shows the list of investment products selected in Cluster 1. There are seven types of investment products in Cluster 1, including equity funds, bond funds, stocks, fixed deposits, currency linked deposits, MPF and insurance. The level of risk associated with each product is indicated. 9 out of 22 products are rated as high risk while 6 out of 22 products are rated as low risk. Over half the products are rated as higher risk, with about a fifth of products rated as low risk. This means that this group tends to mainly select higher risk products for their portfolios. As mentioned above, this implies that cluster 1 is aggressive in taking more risk to achieve higher returns in medium - to long-term capital growth. To simplify the calculation, the financial investment products contained in the portfolio are represented by letters (A-V), and are arranged in alphabetical order.

Table 5.8 List of financial investment options selected by Cluster 1

Types of Attributes Extracted from Cluster 1's Portfolio (Input)			
Name of Attribute	**Symbol**	**Risks level**	**No of Customer Selected (in hundreds)**
Global Equity Fund	A	Medium	2
Korea Equity Fund	B	Medium	24
Russia Equity Fund	C	High	2
Asia Pacific Equity Fund (Exclude Japan)	D	High	1
Emerging Europe Fund	E	High	3
Global Emerging Markets Equity Fund	F	High	32
Material Stock	G	Medium	28
Saving Insurance	H	Low	3
Malaysia Equity Fund	I	High	9
Global Technology Fund	J	High	12
Taiwan Equity Fund	K	High	25
Emerging Markets Bond Fund	L	Low	11
Global Bond Fund	M	Low	5
Fixed Deposit	N	Low	5
Currency Linked Deposit	O	Low	3
Life Insurance	P	N/A	6
China Growth Fund	Q	Low	24
Medical Insurance	R	N/A	2
Investment Linked Insurance	S	Medium	2
MPF	T	N/A	49
Technology Stock	U	High	53
Energy Stock	V	High	37

Source: Convoy Financial Service Limited

In the following, the steps of Apriori algorithm are undergone for Cluster 1.

Step 1 a:

Here, T = {Ta∣a = 1, 2…65} as extracted from cluster 1 for rules mining. The number of attributes are J_p = { J_p ∣ p = 1, 2…22}. A table to record the support count of each attribute is created. There are 22 types of financial investment products t selected by cluster 1. But the support counts of attributes are very different. Some attributes are frequent but unapparent. The unapparent items may generate unexpected rules. To allow more unapparent items to be considered, the most average support count (nearest integer) is taken as the threshold support, i.e. 15 (338/22 = 15.36). The predetermined threshold support count ($S_{ap\text{-}min}$) and threshold confidence level ($R_{\text{-}min}$) are set as 25% of total T and 95%, respectively. The threshold support count is 16 (65 × 25% = 16). The threshold confidence is relatively high, which can ensure the mining efficiency. Table 5.9 shows the support count of each financial investment product selected by cluster 1.

Table 5.9 Support count of financial investment options selected by cluster 1

Symbol	Support Count	Symbol	Support Count
A	2	L	11
B	24	M	5
C	2	N	5
D	1	O	3
E	3	P	6
F	32	Q	24
G	28	R	2
H	3	S	2
I	9	T	49
J	12	U	53
K	25	V	37

Step 1b:

After creating a table, the support count and the threshold support count (i.e. 16) were compared. If the support count is smaller than the threshold support count, the corresponding attributes are pruned off. The attributes are found in Table 5.10.

Table 5.10 List of qualified attributes

Symbol	Support Count (S_p)
B	24
F	32
G	28
K	25
Q	24
T	49
U	53
V	37

Step 2a:

The qualified attributes are then merged to form 2-itemset.

Step 2b:

By scanning the 2-itemset table and comparing the support count with the threshold support count, the threshold support count of 2-itemset is shown in Table 5.11.

Table 5.11 The 2-itemset table

Symbol	Support Count ($S_{p)}$)	Symbol	Support Count ($S_{p)}$)
BF	19	GT	16
BG	17	KT	23
BT	17	KU	25
BU	18	QT	21
FG	27	QU	22
FT	20	TU	49
FU	25	TV	24
FV	16	UV	28
GU	23	UK	21

Step 3

The threshold support count of 3-itemset is found in Table 5.12 by repeating step 2.

Table 5.12 The 3-itemset table

Symbol	Support Count (S_g)
BFG	17
BUT	17
FGU	22
FTU	20
GTU	16
KTU	23
QTU	20
TUV	37

Step 4:

After generating the 3-itemset table, no frequent itemsets can be found. Thus, the algorithm will terminate here. The minimum confidence value ($R_{\text{-}min}$) is set at 95% when the 3-itemsets are used for generating rules. The confidence and lift ratio of the possible rule using 3-itemset are shown in Table 5.13.

Itemset	**IF**	**Support**	**THEN**	**Support**	**Support**	**Confidence**	**Lift Ratio**
	(Condition)	**(Condition)**	**(Result)**	**(Result)**	**(Condition and Results)**	**(R_w)**	**(I_{-mp})**
		(1)		**(2)**	**(3)**	**(4) = (3)/(1)**	**(5) = (4)/(2)**
BUT	B and U	18/65 = 27.69%	T	49/65 = 75.38%	17/65 =26.15%	94.43%	1.25
	B and T	17/65 = 26.15%	U	53/65 = 81.54%	17/65 =26.15%	100%	1.23
	U and T	49/65 = 75.38%	B	24/65 = 36.92%	17/65 =26.15%	34.69%	0.94
	B	24/65 = 36.92%	U and T	49/65 = 75.38%	17/65 =26.15%	70.83%	0.93
	U	53/65 = 81.54%	B and T	17/65 = 26.15%	17/65 =26.15%	32.07%	1.23
	T	49/65 = 75.38%	B and U	18/65 = 27.69%	17/65 =26.15%	34.69%	1.25

Table 5.13 Confidence and lift ratio for 3-itemset

Itemset	IF (Condition)	Support (Condition) (1)	THEN (Result)	Support (Result) (2)	Support (Condition and Results) (3)	Confidence (R_w) (4) = (3)/(1)	Lift Ratio (I_{-mp}) (5) = (4)/(2)
BFG	B and F	19/65 = 29.23%	G	28/65 = 43.08%	17/65 = 26.15%	89.46%	2.08
	B and G	17/65 =26.15%	F	32/65 = 49.23%	17/65 = 26.15%	100%	2.03
	F and G	27/65 = 41.54%	B	24/65 = 36.92%	17/65 = 26.15%	62.95%	1.70
	B	24/65 = 36.92%	F and G	27/65 = 41.54%	17/65 = 26.15%	70.83%	1.71
	F	32/65 = 49.23%	B and G	17/65 = 26.15%	17/65 = 26.15%	53.12%	2.03
	G	28/65 = 43.08%	B and F	19/65 = 29.23%	17/65 = 26.15%	60.70%	2.08

Itemset	**IF** **(Condition)**	**Support** **(Condition)** **(1)**	**THEN** **(Result)**	**Support** **(Result)** **(2)**	**Support** **(Condition and Results)** **(3)**	**Confidence** **(R_{w})** **(4) = (3)/(1)**	**Lift Ratio** **(I_{-mp})** **(5) = (4)/(2)**
FGU	F and G	27/65 = 41.54%	U	53/65 = 81.54%	22/65 =33.85%	81.49%	1.00
	F and U	25/65 = 38.46%	G	28/65 = 43.08%	22/65 =33.85%	88.01%	2.04
	G and U	23/65 = 35.38%	F	32/65 = 49.23%	22/65 =33.85%	95.68%	1.94
	F	32/65 = 49.23%	G and U	23/65 = 35.38%	22/65 =33.85%	68.76%	1.94
	G	28/65 = 43.08%	F and U	25/65 = 38.46%	22/65 =33.85%	78.58%	2.04
	U	53/65 = 81.54%	F and G	27/65 = 41.54%	22/65 =33.85%	41.53%	1.00

Itemset	**IF (Condition)**	**Support (Condition) (1)**	**THEN (Result)**	**Support (Result) (2)**	**Support (Condition and Results) (3)**	**Confidence (R_{w}) (4) = (3)/(1)**	**Lift Ratio (I_{-mp}) (5) = (4)/(2)**
FTU	F and T	20/65 = 30.77%	U	53/65 = 81.54%	20/65 = 30.77%	100%	1.23
	F and U	25/65 = 38.46%	T	49/65 = 75.38%	20/65 = 30.77%	80%	1.06
	T and U	49/65 = 75.38%	F	32/65 = 49.23%	20/65 = 30.77%	40.81%	0.83
	F	32/65 = 49.23%	T and U	49/65 = 75.38%	20/65 = 30.77%	62.5%	0.83
	T	49/65 = 75.38	F and U	25/65 = 38.46%	20/65 = 30.77%	40.82%	1.06
	U	53/65 = 81.54%	F and T	20/65 = 30.77%	20/65 = 30.77%	37.74%	1.23

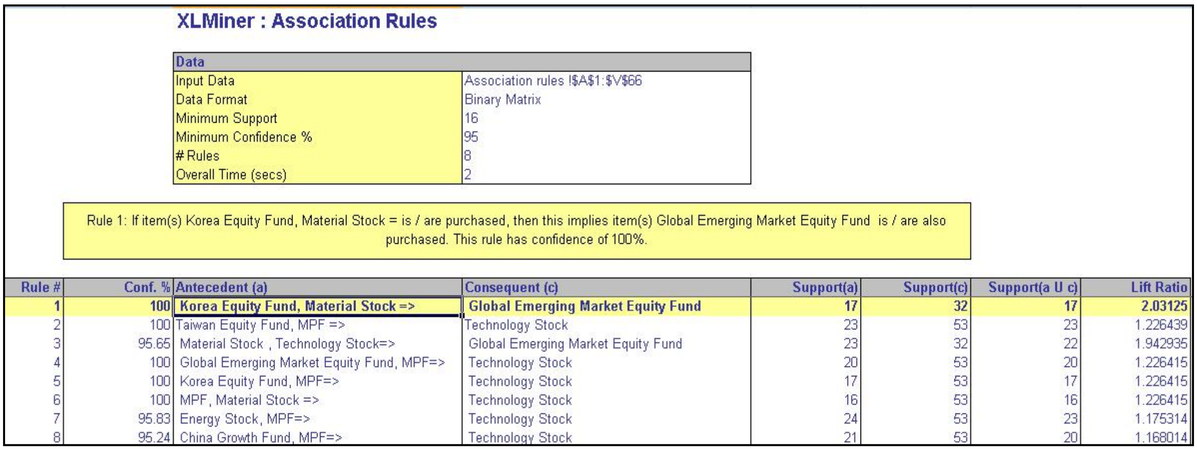
Itemset	**IF** **(Condition)**	**Support** **(Condition)** **(1)**	**THEN** **(Result)**	**Support** **(Result)** **(2)**	**Support** **(Condition and Results)** **(3)**	**Confidence** **(R_w)** **(4) = (3)/(1)**	**Lift Ratio** **(I_{-mp})** **(5) = (4)/(2)**
GTU	GT	16/65 = 24.62%	U	53/65 = 81.54%	16/65 = 24.62%	100%	1.23
	GU	23/65 = 35.38%	T	49/65 = 75.38%	16/65 = 24.62%	69.59%	0.87
	TU	49/65 = 75.38%	G	28/65 = 43.08%	16/65 = 24.62%	32.66%	0.76
	G	28/65 = 43.08%	TU	49/65 = 75.38%	16/65 = 24.62%	57.15%	0.76
	T	49/65 = 75.38%	GU	23/65 = 35.38%	16/65 = 24.62%	33.66%	0.95
	U	53/65 = 81.54%	GT	16/65 = 24.62%	16/65 = 24.62%	30.19%	1.23

Itemset	**IF (Condition)**	**Support (Condition) (1)**	**THEN (Result)**	**Support (Result) (2)**	**Support (Condition and Results) (3)**	**Confidence (R_w) (4) = (3)/(1)**	**Lift Ratio (I_{mp}) (5) = (4)/(2)**
KTU	K and T	23/65 = 35.38%	U	53/65 = 81.54%	23/65 = 35.38%	100%	1.23
	K and U	25/65 = 38.46%	T	49/65 = 75.38%	23/65 = 35.38%	92%	1.22
	T and U	49/65 = 75.38%	K	25/65 = 36.92%	23/65 = 35.38%	46.94%	1.27
	K	25/65 = 38.46%	T and U	49/65 = 75.38%	23/65 = 35.38%	92%	1.22
	T	49/65 = 75.38%	K and U	25/65 = 38.46%	23/65 = 35.38%	46.94%	1.22
	U	53/65 = 81.54%	K and T	23/65 = 35.38%	23/65 = 35.38%	43.39%	1.22

Itemset	**IF (Condition)**	**Support (Condition) (1)**	**THEN (Result)**	**Support (Result) (2)**	**Support (Condition and Results) (3)**	**Confidence (R_w) (4) = (3)/(1)**	**Lift Ratio (I_{-mp}) (5) = (4)/(2)**
TUV	T and U	49/65 = 75.38%	V	37/65 = 56.92%	23/65 = 35.38%	46.94%	0.82
	T and V	24/65 = 36.92%	U	53/65 = 81.54%	23/65 =35.38%	95.82%	1.18
	U and V	28/65 = 43.08%	T	49/65 = 75.38%	23/65 = 35.38%	82.13%	1.09
	T	49/65 = 75.38%	U and V	28/65 = 43.08%	23/65 =35.38%	46.94%	1.09
	U	53/65 = 81.54%	T and V	24/65 = 36.92%	23/65 = 35.38%	43.38%	1.17
	V	37/65 = 56.92%	T and U	49/65 = 75.38%	23/65 =35.38%	62.16%	0.82

Step 5:

The rules met the requirement of the threshold confidence value. The confidence value smaller than the threshold confidence value is pruned off. As highlighted above, the minimum confidence threshold is set at 95%. This means that the rules generated should not be less than 95%. Confidences of eight rules exceed the required minimum confidence level (≥95). Eight rules are retained. Figure 5.15 shows that there are eight useful rules generated in XLMiner^{TM}.



XLMiner : Association Rules

Data	
Input Data	Association rules !$A$1:$V$66
Data Format	Binary Matrix
Minimum Support	16
Minimum Confidence %	95
# Rules	8
Overall Time (secs)	2

Rule 1: If item(s) Korea Equity Fund, Material Stock = is / are purchased, then this implies item(s) Global Emerging Market Equity Fund is / are also purchased. This rule has confidence of 100%.

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	**100**	**Korea Equity Fund, Material Stock =>**	**Global Emerging Market Equity Fund**	**17**	**32**	**17**	**2.03125**
2	100	Taiwan Equity Fund, MPF =>	Technology Stock	23	53	23	1.226439
3	95.65	Material Stock , Technology Stock=>	Global Emerging Market Equity Fund	23	32	22	1.942935
4	100	Global Emerging Market Equity Fund, MPF=>	Technology Stock	20	53	20	1.226415
5	100	Korea Equity Fund, MPF=>	Technology Stock	17	53	17	1.226415
6	100	MPF, Material Stock =>	Technology Stock	16	53	16	1.226415
7	95.83	Energy Stock, MPF=>	Technology Stock	24	53	23	1.175314
8	95.24	China Growth Fund, MPF=>	Technology Stock	21	53	20	1.168014

Figure 5.15 Summary report of association rules

5.2.3.2 Results of the associated rules generated

As shown in Figure 5.16, five rules (rule 1, 2, 4, 5, and 6) have a confidence level of 100%. This implies that these rules are highly reliable in indicating the success rate of investment decision making. All rules' lift ratios are greater than 1, which provides an insight into the prediction, to increase the probability of the "THEN" (result) and the "IF (condition) parts.

It also indicates that all items in the generated rules are positively correlated with other items. Take rule 1 as an example (see Figure 5.16), the Korea equity fund and material stock have a positive correlation with the global emerging markets equity fund.

The results shown in Figure 5.16 also provide marketing implications. There are 53 hundred observations of technology stocks in cluster 1, which is 81.54% of total observations. As can be seen from the results, all rules contain technology stock. This means that cluster 1 will always buy technology stock together with other investment products. A strong relationship between MPF and technology stock is explained by rules 2, 4, 5, 6, 7, and 8 (IF the clients select MPF, THEN they will select technology stocks as well). The rules extracted provide information to the case company to facilitate decision making based on these rules. They also enable Convoy to have a better understanding of investment preferences of its customers from mainland China and Hong Kong, to enable better strategic and marketing planning. The results of these associated rules generated by the FDMM were discussed in the expert panel detailed in chapter 3.2.3, and experts confirmed that these rules are sensible and applicable to provide a good direction execution plan for sales to approach customers for marketing investment products.

Rule 1: Korea Equity Fund and Material Stock → Global Emerging Markets Equity Fund (Confidence 100%)

Rule 1 has a confidence level of 100%. This means that IF the investor purchases Korea equity fund and material stock, THEN they will purchase global emerging markets equity fund with a 100% confidence. Also, the rule is suitable for prediction (lift ratio = 2.03 > 1).

Rule 2: Taiwan Equity Fund and MPF → Technology Stock (Confidence = 100%)

Rule 2 has a confidence level of 100%. This means that IF the investor purchases Taiwan equity fund and MPF, THEN they will purchase technology stock as well with 100% confidence. This rule is also suitable for prediction (lift ratio = 1.23 > 1).

Rule 3: Material Stock and Technology Stock → Global Emerging Markets Equity Fund (Confidence 95.43%)

Rule 3 has a confidence level of 95.43%. This means that IF the investor purchases material stock and technology stock, THEN they will purchase global emerging markets equity fund as well with 95.43% confidence. This rule is also suitable for prediction (lift ratio = 1.94 > 1).

Rule 4: Global Emerging Markets Equity Fund and MPF→ Technology Stock (Confidence = 100%)

Rule 4 has a confidence level of 100%. This means that IF the investor purchases global emerging markets equity fund and MPF, THEN we have 100% confidence that they will purchase technology stock as well. Also, the rule is suitable for prediction (lift ratio = 1.23 > 1).

Rule 5: Korea Equity Fund and MPF → Technology Stock (Confidence = 100%)

Rule 5 has a confidence level of 100%, meaning that IF the investor purchases Korea equity fund and MPF, THEN we have 100% confidence that they will purchase technology stock as well. The lift ratio of 1.23 is greater than 1, hence such a rule is good and suitable for prediction.

Rule 6: MPF and Material Stock → Technology Stock (Confidence = 100%)

Rule 6 has a confidence level of 100%. This means that IF the investor purchases MPF and material stock, THEN they will purchase the technology stock as well with 100% confidence. Also, the rule is suitable and good for prediction (lift ratio = 1.23 > 1).

Rule 7: Energy and MPF → Technology Stock (Confidence 95.82%)

Rule 7 has a confidence level of 95.82%. This means that IF the investor purchases energy stock and MPF, THEN at 95.82% confidence they will purchase technology stock as well. Also, the rule is suitable for prediction (lift ratio = 1.18 > 1).

Rule 8: China Growth Equity Fund and MPF → Technology Stock (Confidence 95.24%)

Rule 8 has a confidence level of 95.24%. This means that IF the investor purchases China growth fund and MPF, THEN with 95.24% confidence they will purchase technology stock as well. Also, the rule is suitable for prediction (lift ratio = 1.17 > 1).

5.2.3.3 Rule analysis

After the generation of association rules using Rules Discovery Module, the rules can be used for formulating marketing strategies. Rule 2 is selected as an example (Taiwan Equity Fund and MPF → Technology Stock) to demonstrate how the rules identified can facilitate decision-making and marketing strategies in the real world (see Figure 5.16).

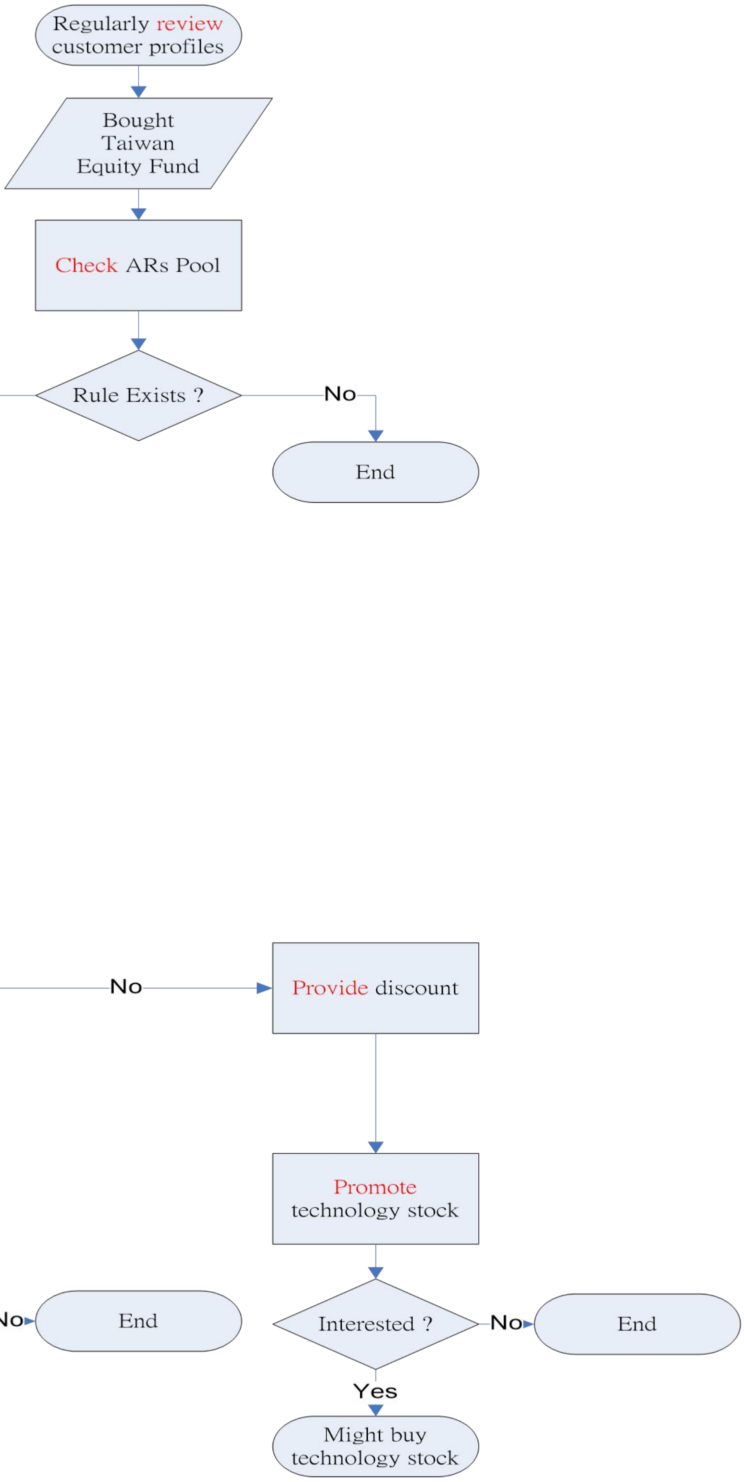


Figure 5.16 Example to support decision making based on Rule 2 generated by the proposed model

As indicated in Figure 5.16, the existing workflow for identifying customers' needs and identification of investment preferences can be improved with the aid of FDMM. Convoy can now identify customers' needs and investment preferences efficiently, effectively and systematically with the aid of FDMM.

For example, financial consultants can regularly review their customer profiles and provide recommendations in a proactive, efficient, effective and systematic manner with the aid of FDMM. The financial consultant can check the buying history for a customer. If the customer bought Taiwan Equity Fund, the financial consultant can then check whether any relevant rule exists in the ARs pool. If the rule (e.g., Rule 2) exists, the financial consultant can formulate the appropriate promotion plan for that customer. After that, the financial consultant can contact the customer to follow up on their existing financial plan and then promote the MPF. If the customer is not interested in buying MPF, the financial consultant could offer discounts or initiate an offer of regular smaller payment investment. It can increase the interest of customers to buy additional investment products and highlight the proactivity of Convoy in providing financial planning advices to its customer. On the other hand, if the customer is interested in buying or has bought MPF, the financial consultant can promote the Technology Stock later. The use of FDMM can increase the success rate of promotion as well as help retain customers.

Before the implementation of the FDMM, Convoy marketed financial products randomly to all kinds of customers without being based on any rules, leading to a relatively low sale success rate (which is measured by dividing the number of successful sales by the total number of sales leads and multiplying the outcome by 100), i.e. 32%. During 2015-2016, Convoy made use of the eight rules generated by the FDMM as supportive evidence to expand its financial product variety and product offers. Convoy's financial consultants marketed the financial products to 400 customers based on these rules. The success rate, i.e. customers decided to invest in the recommended financial products to enrich their investment portfolio, reached 98% (i.e. was enhanced by 66% over the period). Convoy also found that customer satisfaction towards its products and services was enhanced by 8.7% over the period. This outcome confirmed the significance and reliability of this study, and Convoy will keep using this model.

5.3 Using FDMM result of the case study for PSYC Model Validation

There are eight rules generated by FDMM for the case study. According to the experience of the experts involved in the panels and interviews, "these rules are sensible and provide a good direction for sales to have a clear instruction to approach customers" and "these specific rules generated by FDMM are truly effective and practical for business practitioners”. Such rules can in return be applied to validate PSYC Model developed in chapter 4.5.

The case study was performed with two clusters, namely cluster 1 and cluster 2, as mentioned in chapter 5.2.2.1 and chapter 5.2.2.2. Cluster 1 was adopted for FDMM which was lower age, lower income and less investment experience for Hong Kong investors. With reference to the results of PSYC Model and FDMM (see Table 5.15), it is discovered that investors in this cluster prefer investment in SC/SY/PY products (from the results of PYSC Model) and they would invest more in PY/SY products and invest at least one PY/SY product in their portfolio (from the results of FDMM), such technology stock (SY product; see Table 5.14) and global emerging markets equity fund (PY). Taking rule 3 generated by the FDMM as an example, the financial institution making use of the FMDD learns that if younger investors with lower income and investment experience own material stock (SY) and technology stock (PY) in their investment portfolio, they will probably invest in global emerging market fund (PY); this fits the conclusion drawn from the results of PYSC that this kind of investors will invest more in PY/SY products and at least one PY/SY products is included in their portfolio. Taking rule 7 resulting from the FDMM as another example, investors who invested in energy stock (SC) and MPF (PC) and would prefer to add technology stock (SY) in their investment portfolio; this also supports the findings of the PYSC that at least one PY/SY product would be invested by this kind of investors. These findings were reliable and valid by reviewing the eight rules resulting from the FDMM and the results of PSYC Model.

Therefore, financial institutions can make good use of the FDMM and PYSC Model to maximize the possibility of fulfilling investor needs in different clusters by identifying their investment behaviours and predicting investor preference in terms of PY/SY/PC/SC products.

Table 5.14 Product and classification for PSYC Model

Product Name	PSYC Model Classification
Korea Equity Fund	PY Product
Material Stock	SY Product
Global Emerging Markets Equity Fund	PY Product
Taiwan Equity Fund	PY Product
MPF	PC Product
Technology Stock	SY Product
Energy Stock	SC Product
China Growth Equity Fund	PY Product

Table 5.15 Predicted result of the FDMM versus PSYC Model for case study

Predicted result from FDMM	Predicted result from PSYC Model		
	SC	SY	PY
Rule 1: PY			√
Rule 2: SY		√	
Rule 3: PY			√
Rule 4: SY		√	
Rule 5: SY		√	
Rule 6: SY		√	
Rule 7: SY		√	
Rule 8: SY		√	

Note: This case study was for HK investors of low age, low income and less investment experience.
"√" represented the predicted result from FDMM in line with that from PSYC Model.

5.4 Summary

In this chapter, the Financial Data Mining Model (FDMM) was proposed and a case study to implement the proposed model was conducted. The FDMM is a data analysis system, aimed at segmenting customers, e.g. mainland Chinese and Hong Kong investors, into the most appropriate clusters, followed by finding useful rules regarding their investment behaviour, such as examining their investment preferences. Faced with rapid growth of data, it becomes increasingly difficult to handle data without proper data mining systems. Improvement in the situation is mandatory to prevent loss of customers. The proposed FDMM and implemented at Convoy helps financial institutions manage the investment preferences of their customers more efficiently with increased customer satisfaction. The generated rules can uncover hidden patterns in business data, to help analyse the investment behaviour of a specific group of customers. With increased understanding of investment preferences, financial institutions can formulate the most suitable tailor-made financial investment products/portfolios for individual customers.

The case study of FDMM illustrated in this chapter validated the PSYC Model developed in chapter four. The PSYC Model can provide a generalized but simple model for investment product design and marketing strategies formation. In addition, FDMM can enhance the efficiency and effectiveness of sales and marketing efforts with intelligent insights supported by continuously-improved specific rules created by endless cycles of data collection and analysis.

In chapter six the improved workflow after implementing FDMM in Convoy is described in more detail, with a focus on how the improved workflow can help Convoy to deal with the challenges identified in chapter 3.6.2.

Chapter 6 Implementation and discussion

From the results generated by the statistical analysis described in chapter four and data mining modelling in chapter five, the PSYC Model and FDMM were developed to address the problems of personal investment recommendation. FDMM was implemented in the case company (Convoy) and the results help validate the PSYC Model. The PSYC Model provides a generalized but simple model for investment product design and marketing for mainland Chinese and Hong Kong investors. In this chapter, the author further discusses how the FDMM improved the workflow in the case company after implementing the proposed model. The contributions of this research to knowledge, the practical and strategic importance of the FDMM as well as its limitations are described.

6.1 Improved workflow after implementing FDMM in the case company

The workflow of the case company after implementing FDMM is improved and shown in Figure 6.1. When all data are captured by the centralised data warehouse (figure 6.1), information sharing between departments can be enhanced. When the online server receives a customer enquiry, it will be automatically sent to the centralised data warehouse. The sales department will then be notified about the customer's enquiry. The centralised data warehouse reduces processing time to three hours on average (from two days on average before implementing the FDMM) by simplifying the process of transferring customer’s enquiry from the customer services department to the sales department.

Once the enquiry has been received, the financial consultant will check the customer profile and identify the respective cluster for the customer using the Clustering Module (CM). Then the financial consultant can formulate an investment plan for the customer based on the rules extracted by Rules Discovery Module (RDM). After the customer agrees and is satisfied with the investment portfolio recommended, the customer profile will be updated. Both the new data (new portfolio) together with the existing data (existing portfolio) will load into the centralised data warehouse, and go through CM and RDM for continuous generation of useful and updated rules.

The following paragraphs discuss the details of how Convoy can further utilise the FDMM to deal with their current challenges.

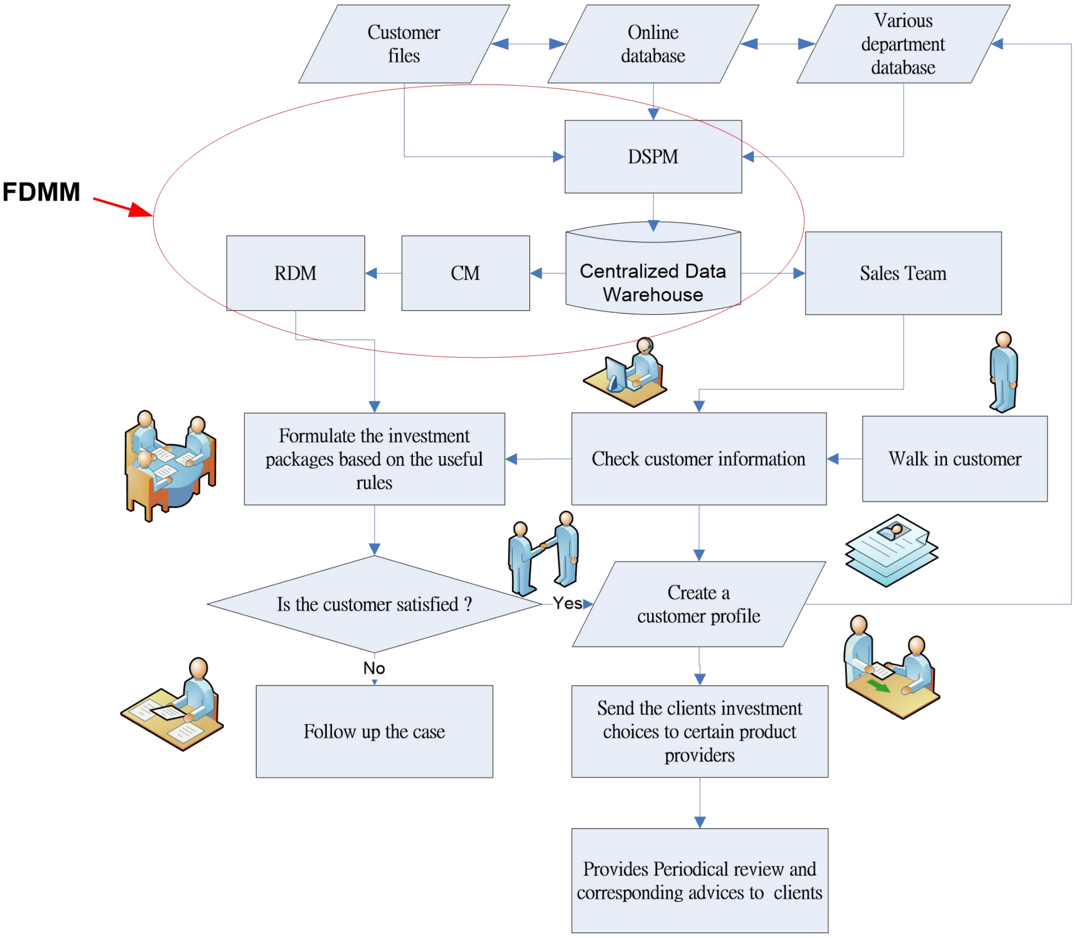

Customer files
Online database
Various department database
DSPM
FDMM
RDM
CM
Centralized Data Warehouse
Sales Team
Formulate the investment packages based on the useful rules
Check customer information
Walk in customer
Is the customer satisfied ?
Yes
No
Create a customer profile
Follow up the case
Send the clients investment choices to certain product providers
Provides Periodical review and corresponding advices to clients

Figure 6.1 Improved workflow of case company using FDMM

6.1.1 Centralised data warehouse

Implementation of FDMM enabled Convoy to have a centralised data warehouse. The Data Selection and Pre-processing Module (DSPM) helps Convoy to link all available data sources, including databases from different departments and the company's online database, to provide a centralised data warehouse which captures all relevant information for data pre-processing. The DSPM also helps to clean data, such as noise and missing data, before it is passed to subsequent processing via the Clustering Module (CM) and Rules Discovery Module (RDM).

The centralised data warehouse also compensates for the weaknesses from using information from a single data source, while increasing the availability of required data. When FDMM was implemented in Convoy, information sharing between departments was improved from one to two days to real time and each department could obtain the desired information from the centralised data warehouse instantly for analysis, such as monthly sales reviews.

The centralised data warehouse also enhanced the efficiency of data management by updating data automatically while preserving old data. The centralised database of FDMM was very different from the operational database. The operational databases usually purged old data periodically in order to improve the system's performance. In contrast, the centralised data warehouse of FDMM kept the valuable old data, such as past investment preference of customers, for further data analysis while updating the data. The centralised data warehouse could optimise the speed for data analysis regardless of the amount of data stored in the data sources.

6.1.2 Efficient identification of customers' needs and better customer experience

FDMM also helped Convoy identify customers' needs efficiently. First, the centralized data warehouse made all necessary customer and investment data available in real time. This enables Convoy to sort out customer requests and provide a prompt response while shortening the long on-hold periods when customers contacted Convoy for enquiry. Second, the CM enabled Convoy to have an efficient way to segment customers into specific target groups based on their characteristics, such as investment experience and income. Customers' potential investment preferences could then be identified effectively and systematically by referring to the PSYC model and the rules generated by FDMM. Since customer needs could be met with a quick and proactive solution by Convoy while customers received personalized and attentive service, the customer experience can be further improved. The communication between customers and Convoy become more effective and efficient as well. After the model implementation, the customer complaint rate on service responsiveness was lowered by 70%.

Convoy could also make use of CM to identify the general needs of specific clusters of customers by referring to the PSYC model and FDMM, as well as to gain a general understanding and prediction of investment needs for new customers once they have been classified into a cluster.

Moreover, Convoy could formulate cluster-based marketing strategies based on the clustering results provided by FDMM, such as recommending the most suitable financial products, such as PY/SY/PC/SC, for its customers, and designing promotional campaigns for a particular cluster. The customer satisfaction level was boosted by 8.7% with the support of the model, which was a leap over the past five years with an average of 2.6%, ranging from 1.1% to 3.2%. The increase in customer satisfaction level is also attributed to personalized and attentive service received by customers, and better customer experience delivered as explained in the first paragraph of this section.

6.1.3 Systematic identification of investment preferences

Last but not least, FDMM helps Convoy identify customers' investment preferences systematically and efficiently. By segmenting customers into different clusters, RDM can discover hidden patterns concerning the investment preferences of each cluster. Financial consultants can thus gain further understanding of the correlations between investment products and systematically formulate investment plans for customers based on the rules identified, instead of their subjective judgement or experience. Before implementing the model, financial consultants had to spend usually 2 days on average to digest loads of customer and product information for designing investment plans for customers. With the FDMM, financial consultants can easily compile and extract useful data for customer's investment plans, which saves time by 50% on average.

The RDM further assists financial consultants to examine the investment preferences of customers periodically, i.e. six months, so that they can formulate investment plans for their customers in a proactive manner. This review exercise was not conducted previously before model implementation.

Most importantly, RDM enables financial consultants to identify the investment preferences of customers systematically and efficiently, even among the wide range of investment products available at Convoy, when the financial consultant is lacking investment experience, or when customers change their investment portfolio frequently.

6.2 Innovations and contributions to knowledge

Apart from facilitating Convoy to formulate more appropriate marketing strategies by improving its workflow, this research has the following innovation and contributions to knowledge:

6.2.1 Developing the FDMM derived from the major attributes to predict customer behaviours

The application of the existing methods, including clustering analysis, apriori algorithms and association rules, as well as combining models, such as DSPM, CM, and RDM, to achieve FDMM are the innovations of this research. Apart from improving the business workflow of the case company, this research provides financial institutions with a good framework of FDMM for identifying hidden investment patterns of their customers. Financial institutions can refer to the framework and case study of the FDMM to customize the proposed model based on their background, so as to

i) efficiently process the massive amounts of data related to customers and financial products;

ii) predict customers' investment preferences for buying financial products;

iii) get more reliable customer information to better understand customer behaviour, thereby formulating more suitable marketing strategies;

iv) create artificial intelligence supported by continuously-improved specific rules which are created by endless cycles of data feeding and outputs to automatically produce sales and marketing rules.

6.2.2 Identifying the major influencing factors in investment behaviours and creating the PSYC Model

As discussed in chapters one and two, the field of behavioural finance has gained popularity due to the complexity of investor behaviours such as irrational/emotional behaviours behaviour (Zhang and Zheng, 2015; Iqbal, 2011; Jureviciene and Jermakova, 2012). Scholars (Ricciardi and Simon, 2000; Bikas *et al.*, 2013) advocated that while the psychological principle dominates the behavioural finance literature, behavioural finance should be interdisciplinary and incorporate three elements – psychology, sociology and finance. However, study focusing on these elements to predict investor preferences is limited due to the unavailability of confidential financial transaction data. This knowledge gap regarding the major contributing factors to investment behaviour within the framework of behavioural finance is identified, from the literature reviewed in chapter 2.2 and supported by the finance industry expert panels conducted in chapter 3.2.3. In this study, the creation of the PSYC Model fulfilled the knowledge gap by identifying the major attributes to explain investment behaviour by leveraging psychological, sociological and demographic factors. The PYSC Model created in chapter 4.5 serves as a generalized and simple model for financial institutions to understand investor preference toward different kinds of financial products in terms of PYSC characteristics. To the financial industry, this research provides a good reference regarding the investment behaviour of mainland Chinese and Hong Kong investors. The major determining factors influencing both investors include age, annual income and investment experience. The PSYC Model summarizes the results and provides a generalized and simple model for the industry to have a guidance to design investment products and formulate marketing strategies based on the behavioural finance concepts and the major factor influencing investment behaviour identified in this study. Financial institutions can make use of this model to analyse their customer segments and corresponding behaviour in order to sell the right investment products to the right customers. This model supported by the experts is the only one available in the market that helps the industry discover the rational of investment behaviour at the top level and then predict investors' preference at the bottom level in order to get insights into the big picture of investing activities.

6.2.3 Contributions to the literature

Behavioural finance is applied widely to explore the stock-market trends of the developed security markets. However, the number of applications for exploring individual investment behaviour in Hong Kong and mainland China are limited. From an economic perspective, understanding individual investment behaviours in Hong Kong and mainland China is important as China becomes more international and economically important to more regions of the world. This study is conducted to fill the gap and contribute to the study of using behavioural finance for all kinds of investment behaviour. It creates two reference models that can be easily adapted to different preferences by user businesses.

This project has empirically examined the differences in financial investment behaviour and investment preferences between mainland Chinese and Hong Kong investors. It was identified that the age, investment experience and annual income are the top three factors significantly affecting mainland Chinese and Hong Kong investors' behaviour and investment preferences. This deepens the financial sector's understanding of investors' financial investment behaviour and preferences. Such research is summarized in the PSYC Model for easy application by even layman.

This research also addresses the opportunities of turning data into business knowledge by proposing the FDMM. The application feasibility of the proposed model has been validated via a case study conducted in a representative investment management business. Results indicated that the proposed FDMM is effective not only in facilitating information sharing within the company and improving the workflow of the case company, but also in supporting the formulation of marketing strategies in an efficient, effective and user-friendly way.

6.3 Practical and strategic importance of PSYC Model and FDMM

Corporations today are increasingly concerned with customer satisfaction. Therefore, they analyse customer behaviour, attempt to predict customers' preferences and offer customised services/products. Better data management and knowledge discovery are feasible solutions to help organisations gain insights from data, so as to better understand customers' behaviour and customers' preferences.

Data selection and capture is the first issue needed to be addressed to help improve customer satisfaction. Despite the fact that data is available from numerous sources, such as transaction systems and organisational databases, data capture and selection might be prohibited in some cases (Chan and Zhang, 2014). Some companies resist exchanging information internally (Boohene and Williams, 2012), while some of them lack basic data mining knowledge. The proposed PSYC Model was derived to be a generalized and simple model for investment product design and marketing while the proposed FDMM is also proven to be effective for data mining and also acts as a road map for developing a data mining system in financial institutions.

The PSYC Model enables companies to have a simple model to apply in designing and marketing new investment products based on behavioural finance concepts. It does not require too much technical knowledge and is easy to understand for marketing to clients with limited financial investment knowledge and experience.

The FDMM enables companies to justify a centralised data warehouses for information sharing and data mining by linking all available data sources and capturing relevant information for data pre-processing. FDMM also enables companies to segment customers into desirable clusters. Then useful cluster-specific rules, such as correlations among different financial investment products, can be generated using RDM, to guide marketing and strategic decisions.

The proposed FDMM utilises all data chosen to discover hidden patterns on a continual basis. Historic and current data are continuously merged into RDM for rules mining. This thus tracks trends and increases the availability of knowledge, and thus agility in marketing strategy formulation.

Overall, the development of PSYC Model and FDMM was evaluated and supported by the expert panels created to monitor and guide the experimentation (see chapter 3.2.3). With reference to the panel discussions and interviews, the PSYC Model structure and operation is consistent with the experts' experience and understanding of Hong Kong and China investors' behaviour. However, the knowledge generated is still based on historic information and thus cannot, and should not provide clear direction for daily business operation. To enable a more reliable prediction output the selection rules need to be modified with daily market and news information. Techniques using linked-data could help provide such a prediction function.

The panel experts agreed that FDMM has derived from big data concepts, real operational knowledge to help understand how to respond to Hong Kong and mainland Chinese investors in a period of turbulence and change. However, they questioned the availability and completeness of available data. Nonetheless, currently FDMM would be the most available and reliable model for them to have clear action plan. From their past experience and phenomena in the financial market they concluded that FDMM generates actionable rules which are effective for business practitioners.

Though the proposed PSYC Model and FDMM offers practical and strategic benefits to the financial sector, there are some limitations of the proposed PSYC Model and FDMM. The following sections discuss these.

6.4 Limitations and recommendations

6.4.1 Limitations of XLMiner^{TM}

As an initial attempt, XLMiner^{TM} was applied for prototype building. The prototype built was then applied in the case company. Though XLMiner^{TM} has the ability to provide data mining solutions to users, it has several limitations. Due to the limitation of Excel, the worksheet would be processed very slowly for a large amount of data. Therefore, when a large amount of customer records is involved in future implementations, the proposed FDMM needs to be constructed using an appropriate database language and/or software. As the generic algorithms required by experimenting with XLMiner was established, this study could be extended by looking for more efficient applications of those algorithms.

6.4.2 Limitations of dataset

In the case study, 11,700 sets of data relating to Convoy's most active customers (at least one transaction a year) were obtained from the case company for simulation. Given the limited number of datasets, the empirical results may not fully represent the financial investment behaviours or investment preferences of all customers of the case company. In the future, more customer profiles should be collected from companies to achieve more accurate results. Furthermore, the proposed FDMM should be applied in more financial institutions to test its generalization and feasibility.

In addition, as the dataset about mainland Chinese customers possessed by Convoy included only those Chinese investors who invested in Hong Kong. So, the findings about investment behaviour of mainland Chinese investors studied in this research only applied to and was valid for mainland Chinese investor who invest offshore rather than domestically.

The data collected from Convoy and used for statistical analyses in this study are mixed with ordinal, interval and nominal. Although the validity of variables and analytical models were tested and their results are statistically significant as well as supported by the case company and expert panels, the analyses carried the assumption that the variables are ordinal, linear and normally-distributed. It is noted the non-normality and nominal variables may have impact on the analytical results, such as bias or inefficiency in regression models. There may also be a concern about the distortion of the results by undergoing the same analytical models with the mixed data. To further valid the generalization and application of this study, this study could be enhanced and extended by analysing categorical data using other software packages such as FACTOR as suggested by Baglin (2014), transforming data to fulfil the normality assumption of regression models by performing such as arbitrary outcome transformations (Schmidt, 2017), or excluding the nominal data, such as gender for conducting relevant analytical analyses for comparison.

6.4.3 FDMM implementation

Concerning the implementation of FDMM, successful implementation of the proposed FDMM relies heavily on the staff, and specifically the collaboration of staff from different departments in the company. As customer data is essential for the proposed FDMM to discover knowledge, the FDMM can never be implemented if staff resists the storage of customer data in the centralised data warehouse. Furthermore, the performance of the proposed FDMM will be greatly affected if staff is resistant to change. It is recommended highly that management communicates the implementation objectives and processes well to staff members. The implementation schedule as well as pros and cons of implementing FDMM and process consultation will help to motivate them to participate before FDMM implementation.

6.4.4 Limitation of FDMM

This proposed FDMM has been developed for the general situation but not a special event that occurs in a short period of time. The proposed FDMM in this study cannot predict investor behaviour in a changing environment including the impact of political and other events such as wars, natural disasters, and scandals. Once an event happens, financial institutions have to assess the impact of the event and adjust their product and marketing strategies manually. This study could be extended to explore how the changing environment affects investment behaviour and for example how political factors could be incorporated into the proposed model. FDMM could be further developed with applications in different environment scenarios.

Chapter 7 Conclusions

The financial industry plays a significant role in the economy of mainland China and Hong Kong. This realistic study of the determining factors contributing to behavioural finance and the application of the proposed PSYC Model and FDMM has generated an approach which appears to produce useful results using an academically sound methodology and commercially viable. To manage the massive amounts of customer data and to understand customers' financial investment behaviour and investment preferences, are essential so financial institutions can provide superior service with appropriate financial products for customers. This chapter provides a conclusion by describing how the objectives were achieved the achievements and major findings of this study and by suggesting further research.

7.1 Key findings and achievements

This study explored financial investment behaviours and predicted investors' preferences using an integrated approach. It addressed the research question raised in chapter 1.1 that the key differences in factors influencing financial behaviour between mainland Chinese and Hong Kong investors are age, annual income and investment experience, while four objectives identified in chapter 1.2 were achieved:

i) The major attributes explaining investment behaviours under three constructs (i.e. psychological, sociological and demographic factors) were identified through literature review discussed in chapter two and primary factor analysis described in chapter 4.2; the results were as follows:

a) Demographic factor: age and gender

b) Psychological factor: investment experience

c) Sociological factor: annual income, education level and marital status

Such results meet the research objective (i) described in chapter 1.2.

ii) Through the correlation analysis described in chapter 4.3, it was found that the investment behaviour of both mainland Chinese and Hong Kong investors are highly correlated with the attributes: age, annual income and investment experience. Such results fulfil the research objective (ii) identified in chapter 1.2

iii) The impacts of these major attributes on the investment behaviours of investors in Hong Kong and mainland China were further analysed through the three regression models discussed in chapter 4.4. The significant similarities and differences in investment behaviours of mainland Chinese and Hong Kong investors extracted by the generated process were as follows:

a) Similarity 1: Annual income had a positive impact on the quantity of fund unit held by mainland Chinese and Hong Kong investors.

b) Similarity 2: Investment experience had a positive impact on the choice of country-specific financial investment options selected by both mainland Chinese and Hong Kong investors.

c) Similarity 3: Investment experience had a negative impact on the quantity of fund unit held by both mainland Chinese and Hong Kong investors.

d) Difference 1: The impact of age on the amounts of fund units held by mainland Chinese and Hong Kong investors was reversed. Age negatively affected the quantity of fund unit held by mainland Chinese investors, but positively affected the amounts held by Hong Kong investors.

e) Difference 2: Age positively affected the choice of country-specific financial investment options selected by mainland Chinese but negatively affected the options selected by Hong Kong investors.

f) Difference 3: Annual income negatively affected the choice of country-specific financial investment options selected by mainland Chinese investors but positively affected those selected by Hong Kong investors.

These results regarding the major attributes influencing mainland Chinese and Hong Kong investors were summarized in the developed PSYC Model and fulfil the research objective (iii) identified in chapter 1.2.

iv) With reference to chapters five and six, the development and implementation of the FDMM in the case company demonstrated and confirmed that the proposed model could help financial institutions formulate marketing strategies by efficiently processing massive amounts of historical and new data related to customers and financial products, identifying the hidden relationships influencing the customers' choices of financial products, and predicting their investment needs and preferences.

Eight rules were generated by the FDMM and served as the basis for the case company to customize the most suitable products for individual customers and to formulate the most appropriate marketing strategies for increased customer satisfaction. Six out of the eight rules were at 100% confidence level, while the other two were at over 95% confidence level. The eight rules are as follows:

a) Rule 1: Investors who invest in Korea Equity Fund and Material Stock will also buy Global Emerging Markets Equity Fund (Confidence 100%)

b) Rule 2: Investors who invest in Taiwan Equity Fund and MPF will also purchase Technology Stock (Confidence = 100%).

c) Rule 3: Investors who buy Material Stock and Technology Stock will also invest in Global Emerging Markets Equity Fund (Confidence = 100%).

d) Rule 4: Investors who purchase Global Emerging Markets Equity Fund and MPF will also buy Technology Stock (Confidence = 100%).

e) Rule 5: Investors who invest in Korea Equity Fund and MPF will also purchase Technology Stock (Confidence = 100%).

f) Rule 6: Investors who buy MPF and Material Stock will invest in Technology Stock (Confidence = 100%).

g) Rule 7: Investors who buy Energy and MPF will also purchase Technology Stock (Confidence 95.82%).

h) Rule 8: Investors who purchase China Growth Equity Fund and MPF will also buy Technology Stock (Confidence 95.24%).

This intelligent FDMM utilises the availability of data to support decision making, but also helps financial institutions improve their business workflow to address the challenges identified in chapter 3.6.3. The application of FDMM achieved the research objective (iv) stated in chapter 1.2.

The fundamentals of FDMM are the following:

i) Firstly, all relevant quality data are collected and pre-processed through DSPM, so that information can be shared with each department across the organization.

ii) Secondly, customers can be classified into specific clusters using CM. The segmented groups provide marketing implications to sales managers, so as to develop customer values for highly profitable customers.

iii) Thirdly, the developed RDM can be used to generate useful rules for the target clusters.

The unique characteristic of RDM is that it will keep updating as newly formed data are continually loaded into RDM for rules mining. The new data can be of great interest for the company as knowledge discovery is an evolving process, which allows the central data warehouse to become richer and richer. The useful rules can be integrated with industry knowledge, experience, customer needs, and more. By making use of the FDMM, financial institutions can learn more about investors' behaviour in financial markets and discover the patterns on how to build a customized portfolio based on the rules extracted, for sustainable long-term success.

To conclude, despite the proximity between Hong Kong and mainland China as well as the similarity in languages and cultures, investors in mainland China and Hong Kong behave differently from each other mainly by the influence of age and annual income. The outcomes of this research could assist financial institutions to gain better insights into the behaviour of their customers from Hong Kong and mainland Chinese, and then offer the most suitable financial products for fulfilment.

7.2 Further work

To further improve this study and address the limitations discussed in chapter 6.4, the following future research is recommended.

i) First, the proposed FDMM could be further enhanced by using more efficient applications rather than XLMiner^{TM} to enhance its capability for handling larger amounts of customer data. A comparative analysis of data mining tools could be conducted.

ii) Second, the PSYC Model and FDMM developed could be applied and further validated by carrying out additional case studies in other financial institutions. By using a variety of customer profiles from a wider range of financial institutions, the models' applicability and capability could be examined and tested. Moreover, this study providing a framework to study financial investment behaviour can be further generalised by employing data from other nationalities. Since the proposed FDMM can be tailored based on the background of the financial institution, carrying out more case studies in other financial institutions or from other nationalities would strengthen its reliability and demonstrate its generalizability and scope.

iii) Other external factors, such as the general financial climate, government rules or initiatives and political events, which are not considered in this study but may have influences on investment behaviours of mainland Chinese and Hong Kong investors, could be incorporated into the developed models. Furthermore, it would be interesting to use the personality profiles generated by possibly Facebook, WeChat, Weibo, so forth and their correlations with investment preferences to produce better knowledge. This could probably make the models developed more dynamic and enable them to be operated in a changing environment.

References

Ackert, L. F. (2014). *Traditional and behavioral finance* (pp. 25-41). John Wiley & Sons, Inc.

ACNielsen Research (2005). ANZ Survey of Adult Financial Literacy in Australia: Final Report, ACNielsen Research, Melbourne, November.

Agnew, J.R., Anderson, L.R., Gerlach, J.R., & Szykman, L.R. (2008). Who chooses annuities? An experimental investigation of the role of gender, framing, and defaults. *American Economic Review*, 98, 418-442.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In: *Proceedings of 20th International Conference on Very Large Databases*, Santiago de Chile, 487–489.

Aguinis, H., Petersen, S., & Pierce, C. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods*, 2, 315-339.

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2), 503–527.

Al-Ajmi, J.Y. (2008). Risk tolerance of individual investors in an emerging market. *International Research Journal of Finance and Economics*, 17(2), 15-26.

Al-Hassan, A.A.; Alshameri, F., & Sibley, E.H. (2013). A research case study: Difficulties and recommendations when using a textual data mining tool. *Information & Management*, 50(7), 540-552.

Allen, F., Qian, J., & Qian, M. (2005). Law, finance, and economic growth in China. *Journal of Financial. Economics*, 77(1), 57–116.

Ansari, L. & Moid, S. (2013). Factors affecting investment behaviour among young professionals. *International Journal of Technical Research and Applications,* 1(2), 27-32.

Antonakis, J., & Dietz, J. (2011). Looking for validity or testing it? The perils of stepwise regression, extreme-score analysis, heteroscedasticity, and measurement error. *Personality and Individual Differences*, 50, 409-415.

Azadegan, A., Porobic, L., Ghazinoory, S., Samouei, P., & Kheirkhah, A.S. (2011). Fuzzy logic in manufacturing: A review of literature and a specialized application. *International Journal of Production Economics*, 132(2), 258-270.

Aziz, J., & Cui, L. (2007). Explaining China's low consumption: The neglected role of household income. *IMF Working Papers, 07(181),* 1-36.

Baker, H., & Ricciardi, V. (2014). *Investor Behavior: The psychology of financial planning and investing (*1st ed.). New York: John Wiley & Sons.

Baglin, J. (2014). Improving your exploratory factor analysis for ordinal data: A demonstration using FACTOR. *Practical Assessment, Research & Evaluation*, 19(5), 1-15.

Barberis, N., & R. Thaler. (2002). A survey of behavioral finance. In *Handbook of the Economics of Finance*, G. Constantinides, M. Harris, & R. Stulz(eds.). Forthcoming.

Batemany, H., Louviere, J., Thorp, S., Islam, T., & Satchel, S. (2008). An Experimental Survey of Investment Decisions for Retirement Savings. *SSRN Electronic Journal.*

Batra, G., Vijaylaxmi, & Gupta, A. (2012). Mining the investor's perception about different investment options using clustering analysis. *International Journal on Computer Science and Engineering*, 4(9), 1513 -1516.

Baumeister, R., Vohs, K., DeWall, N.C. & Zhang. L. (2007). How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, Rather Than Direct Causation. *Personality and Social Psychology Review*, 11(2), 167-203.

Bernardo, D., Hagras, H., & Tsang, E. (2013). A genetic type-2 fuzzy logic based system for the generation of summarised linguistic predictive models for financial applications. *Soft Computing*, 17(12), 2185-2201.

Bernstein, P.L. (1996). *Against the gods: The remarkable story of risk*. New York: John Wiley & Sons.

Berry, M.J., & Linoff, G. (2004). *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*. Indianapolis, IN: Wiley Pub.

Bezdek, J.C. (1998). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9), 17-24.

Bikas, E., Jurevičienė, D., Dubinskas, P., & Novickytė, L. (2013). Behavioural finance: The emergence and development trends. *Procedia-social and behavioral sciences*, 82, 870-876.

Boohene, R., & Williams, A.A. (2012). Resistance to organisational change: A case study of Oti Yeboah Complex Limited. *International Business and Management*, 4(1), 135-145.

Byrne, J.P., Miller, D.C., & Schafer W.D. (1999). Gender Differences in Risk Taking: A Meta-Analysis. *Psychological Bulletin*, 125(3), 367-383.

Byrne, K. (2005). How do consumers evaluate risk in financial products? *Journal of Financial Services Marketing*, 10(1), 21-36.

Phua, C., Lee, V., Smith, K., & Gayler R. (2005). A comprehensive survey of data mining-based fraud detection research. Clayton School of Information Technology, Monash University.

Caballero-Morales, S.O. & Rahim, A. (2015). Analysing the effect of non-normality on the solution space for the economic statistical design of X-bar control charts. In: *Proceeding of 2015 International Conference on Industrial Engineering and Operations Management (IEOM)*, Dubai, 1-6.

Capon, N., Fitzsimons, G.J., & Prince, R.A. (1996). An individual level analysis of the mutual fund investment decision. *Journal of Financial Services Research*, 10, 59-82.

Cattell, R.B. (1979). *The scientific use of factor analysis: in behavioral and life sciences*. New York: Plenum Press.

Chai, J., He, W., & Yu, H. (2013). Individual investors' lifestyles and its influence to investment behaviour. *Management Review*, 25(10), 147-156.

Chan, C.L.P., & Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314–347.

Chan, Y.C., & Chui, A.C.W. (2016). Gambling in the Hong Kong stock market. *International Review of Economics & Finance*, 44, 204-218.

Chandra, A., & Kumar, R. (2012). Factors Influencing Indian Individual Investor Behaviour: Survey Evidence. *SSRN Electronic Journal*.

Chang, C., & Lin, S. (2015). The effects of national culture and behavioral pitfalls on investors' decision-making: Herding behavior in international stock markets. *International Review of Economics & Finance*, 37, 380-392.

Chang, S., & Gan, Y.Y. (2015). A study on the impact of internet finance on Guangxi's economic development. *Journal of Guangxi Economic Management Cadre College*, 27(2), 45-55.

Charles, A., & Kasilingam, R. (2013). Does the investor's age influence their investment behaviour? *Paradigm*, 17(1-2), 11-24.

Chaudhuri, T. & Ghosh, I. (2016). Using Clustering Method to Understand Indian Stock Market Volatility. *Communications on Applied Electronics*, 2(6), 35-44.

Chen, A. & Leung, M. (2004). Regression neural network for error correction in foreign exchange forecasting and trading. *Computers & Operations Research*, *31*(7), 1049-1068.

Chen, A.S., Leung, M.T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers & Operations Research*, 30(6), 901–923.

Chen, C., Chiang, T., & So, M. (2003). Asymmetrical reaction to US stock-return news: evidence from major stock markets based on a double-threshold model. *Journal of Economics and Business*, 55(5-6), 487-502.

Chen, C.J.P., Li, Z.Q., Su, X.J., & Sun, Z. (2011b). Rent-seeking incentives, corporate political connections, and the control structure of private firms: Chinese evidence. *Journal of Corporate Finance*, 17(2), 229–243.

Chen, G., Kim, K., Nofsinger, J., & Rui, O. (2007). Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investors. *Journal of Behavioral Decision Making*, 20(4), 425-451.

Chen, M.S., Han, J., & Yu, P.S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-883.

Chen, N., Xu, Z. & Xia, M. (2013). Correlation coefficients of hesitant fuzzy sets and their applications to clustering analysis. *Applied Mathematical Modelling*, 37(4), 2197-2211.

Chen, Q.Z., Ou, Y.Q., & Sun, H. (2011). Design and implement of customer communication behavior analysis system. *Journal of Software*, 6(8), 1484-1491.

Chen, S.M., Sun, Z., Tang, S., & Wu, D.H. (2011a). Government intervention and investment efficiency: Evidence from China. *Journal of Corporate Finance*, 17(2), 259–271.

Chen, W.S., & Du, Y.K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications*, 36(2), 4075–4086.

Chen, Y. L., Tang, K., Shen, R. J., & Hu, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, 40, 339-354.

Chenoweth, T., and Obradovic, Z. (1996). A multi-component nonlinear prediction system for the S&P 500 Index. *Neurocomputing*, 10, 275–290.

Cheung, K.C., & Coutts, J.A. (2001). A note on weak form market efficiency in security prices: evidence from the Hong Kong stock exchange. *Applied Economics Letter*s, 8, 407-410.

Chiang, T.C., & Zheng, D.Z. (2010). An empirical analysis of herd behavior in global stock markets. *Journal of Banking & Finance*, 34(8), 1911-1921.

Chiang, T.C., Li, J.D., & Tan, L. (2010). Empirical investigation of herding behavior in Chinese stock markets: Evidence from quantile regression analysis. *Global Finance Journal*, 21(1), 111-124.

Chiang, W.Y. (2011). To mine association rules of customer values via a data mining procedure with improved model: An empirical case study. *Expert Systems with Applications*, 38, 1716-1722.

Chung, W., & Tseng, T.L. (2012). Discovering business intelligence from online product reviews: A rule-induction framework. *Expert Systems with Applications*, 39(12), 11870-11879.

Cochrane, J.H. (1994). Permanent and transitory components of GNP and stock prices. *Quarterly Journal of Economics*, 109, 241–65.

Collard, S. (2009). *Individual investment behaviour: A brief review of research*. Personal Accounts Delivery Authority, 1-32.

Cooper, M.J., Dimitrov, O., & Rau, P.R. (2001). A rose.com by any other name. *The Journal of Finance*, 56(6), 2371-2388.

Corter, J. & Chen, Y. (2006). Do Investment Risk Tolerance Attitudes Predict Portfolio Risk? *Journal of Business and Psychology*, 20(3), 369-381.

Corter, J.E., & Chen, Y.J. (2006). Do investment risk tolerance attitude predict portfolio risk? *Journal of Business and Psychology*, 29, 369-384.

Costello, A. & Osborne, J. (2005). Exploratory Factor Analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10(7), 1-9.

Dandapani, K. (2008). Growth of e-financial services: Introduction to the special issue. *Managerial Finance*, 34(6), 361-364.

Darlington, R. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69(3), 161-182.

De Bondt, W.F.M., & Thaler, R.H. (1987). Further evidence on investor overreaction and stock market seasonality. *The Journal of Finance*, 42(3), 557-581.

Demirer, R., & Kutan, A.M. (2006). Does herding behavior exist in Chinese stock markets? *Journal of International Financial Markets Institutions and Money*, 16(2), 123–142.

Demiriz, A., Ertek, G., Atan, T., & Kula, U. (2011). Re-mining item associations: Methodology and a case study in apparel retailing. *Decision Support Systems*, 52(1), 284-293.

Desai, V.S., & Bharati, R. (1998). The efficiency of neural networks in predicting returns on stock and bond indices. *Decision Sciences*, 29, 405–425.

Diacon, S., & Ennew, C. (2001). Consumer perceptions of financial risk. *The Geneva Papers on Risk and Insurance*, 26(3), 389-409.

Díaz, B.A., Gonzálezb, I., & Tuya, J. (2004). Incorporating fuzzy approaches for production planning in complex industrial environments: the roll shop case. *Engineering Applications of Artificial Intelligence*, 17(1), 73–81.

Donkers, B., B. Melenberg, & A.V. Soest (2001). Estimating Risk Attitudes Using Lotteries: A Large Sample Approach. *Journal of Risk and Uncertainty*, 22(2), 165-195.

Dourra, H., & Siy, H. (2002). Investment using technical analysis and fuzzy logic. *Fuzzy Sets and Systems*, 127, 221-240.

Dreman, D. & Berry, M. (1995). Overreaction, Underreaction, and the Low-P/E Effect. *Financial Analysts Journal*, 51(4), 21-30.

East, R., Wright, M. & Vanhuele, M. (2013). *Consumer Behaviour: Applications in Marketing*. 1st ed.

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, 29(4), 927–940.

Eoma, C., Jung, W.S., Choi, S. Oh, G., & Kim, S. (2008). Effects of time dependency and efficiency on information flow in financial markets. *Physica A: Statistical Mechanics and its Applications*, 387(21), 5219–5224.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2), 383–417.

Fama, E., & French, K. (1988) Permanent and temporary components of stock prices. *Journal of Political Economy*, 96, 246–273.

Fares, A. & Khamis, F. (2011). Individual Investors' Stock Trading Behavior at Amman Stock Exchange. *International Journal of Economics And Finance*, 3(6), 128-134.

Fayyad, U.M, Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.

Fellner, G., & Maciejovsky, B. (2007). Risk attitude and market behavior: Evidence from experimental asset markets. *Journal of Economic Psychology*, 28(3), 338-350.

Feng, L., & Seasholes, M.S. (2004). Correlated trading and location. *The Journal of Finance*, 5995, 2117-2144.

Fidelity Investments Management (Hong Kong) Limited. (2004). *Personal Investment Behaviour in Hong Kong*. Retrieved from https://www.hkupop.hku.hk/english/report/fidel05/pr.pdfhttps://www.hkupop.hku.hk/english/report/fidel05/pr.pdf

Field, A. (2000). *Discovering Statistics using SPSS for Windows*. London – Thousand Oaks – New Delhi: Sage publications.

Field, A. (2009). *Discovering statistics using SPSS*. Los Angeles [i.e. Thousand Oaks, Calif.]: SAGE Publications.

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd edition). London: Sage.

Financial Times (2008). *Storm clouds gather for China's airlines*. December 30.

Frankfurter, G. & McGoun, E. (2000). Market Efficiency or Behavioral Finance: The Nature of the Debate. *Journal of Psychology and Financial Markets*, *1*(3-4), 200-210.

Fred van Raaij, W. (2016). *Financial Literacy and Financial Behaviour*. Be'er Sheva, Israel: Economic Psychology.

Friedmann, R., Sanddorf-Köhle, W. (2002). Volatility clustering and nontrading days in Chinese stock markets. *Journal of Economics and Business*, 54, 193–217.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Probability*. Philadelphia, Pa.: SIAM, Society for Industrial and Applied Mathematics.

Gärling, T., Erich K., Alan L., & Fred van Raaji. (2009). Psychology, financial decision making, and financial crisis. *Psychological Sciences in the Public Interest*, 10(1), 1-47.

Geetha, N. & Ramesh, M. (2012). A study on relevance of demographic factors in investment decisions. *International Journal of Financial Management IJFM*, *1*(1), 39-56. Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, 241(1), 236-247.

Gerrans, P., & Murphy, M.C. (2004). Gender differences in retirement savings decision. *Journal of Pension Economics and Finance*, 3(1), 145-164.

Grable, J.E., & Lytton, R.H. (1999). Assessing financial risk tolerance: Do demographic socioeconomic and attitudinal factors work? *Family Relations and Human Development/Family Economics and Resource Management Biennial,* 3, 80-88.

Grossman, S.J., & Stiglitz, J.E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3), 393-408.

Gunay, S.G., & Demirel, E. (2011). Interaction between demographic and financial behavior factors in terms of investment decision making. *International Research Journal of Finance and Economics*, 66, 147-156

Haitovsky, Y. (1969). Multicollinearity in Regression Analysis: Comment. *Review of Economics And Statistics*, 51(4), 486-489.

Hájek, P. (2012). Credit rating analysis using adaptive fuzzy rule-based systems: An industry-specific approach. *Central European Journal of Operations Research*, 20(3), 421-434.

Hanafizadeh, P., Keating, B.W., & Khedmatgozar, H.R. (2014). A systematic review of Internet banking adoption. *Telematics and Informatics*, 31(3), 492–510.

Hart, M.A. & Sailor, D.J. (2009). Quantifying the influence of land-use and surface characteristics on spatial variability in the urban heat island. *Theoretical and Applied Climatology*, 95(3), 397–406.

Hartog, J., Ferrer-i-Carbonell, A., & Jonker, N. (2002). Linking measured risk aversion to individual characteristics. *Kyklos*, 55, 3-26.

Hellman, A. (1995). A fuzzy expert system for evaluation of municipalities: An application. In: *Proceedings of the 25th TICA*, 1, 159-187.

Herbst, A.F. (2001). E-finance: Promises kept, promises unfulfilled, and implications for policy and research. *Global Finance Journal*, 12(2), 205–215.

Herbst, J., & Karagiannis, D. (2004). Workflow mining with InWoLvE. *Computers in Industry*, 53(3), 245–264.

Hira, T.K., & Loibl, C. (2008). Gender differences in investment behavior. *Handbook of Consumer Finance Research, Part III*, 253-270.

Hirshleifer, D. (2014). Behavioral Finance. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2480892

Ho, G.T.S., Lau, H.C.W., Kwok, S.K., Lee, C.K.M., and Ho, W. (2009). Development of a co-operative distributed process mining system for quality assurance. *International Journal of Production Research*, 47(4), 883-918.

Hoffmann, A., Pennings, J., & Post, T. (2013). Individual Investor Perceptions and Behavior during the Financial Crisis. *Journal of Banking & Finance,* 37(1), 60-74.

Hon, T.Y. (2012). The behaviour of small investors in the Hong Kong derivatives markets: A factor analysis. *Journal of Risk and Financial Management*, 5, 59-77.

Hon, T.Y. (2013). The investment behaviour of small investors in stock market: A survey in Hong Kong. *International Journal of Financial Management*, 3(3), 8-25.

Hosseini, S.M.S., Maleki, A., & Gholamian, M.R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, 37(7), 5259 - 5264.

Huang, C.F., & Hsueh, S.L. (2010). Customer behavior and decision making in the refurbishment industry - a data mining approach. *Journal of Civil Engineering and Management*, 16(1), 75-84.

Huang, J., & Gao, F. (2005). An empirical study of investment fund behaviour and performance. *Securities Market Herald*, 15(2), 23-27.

Huang, J., & Gao, F. (2005). An empirical study of investment fund behaviour and performance. *Application of Statistics and Management*, 215(5), 581-587.

Huang, J.B., & Fu, F.L. (2006). Analysis of electricity generat firm's investment behaviour and environment regulation. *Chinese Journal of Management Science*. Z1, 713-718.

Huberman, G., & Jiang, W. (2006). Offering versus choice in 401(k) plans: Equity exposure and number of funds. *The Journal of Finance*, 56(2), 763-801.

Hussein, A. & Al-Tamimi, H. (2005). Financial risk and islamic banks' performance in the gulf cooperation council. *The International Journal of Business and Finance Research*, *9*(5), 103-112.

Iqbal, M.H.A. (2011). Behavioural implications of investors for investments in the stock market. *European Journal of Social Sciences*, 20(2), 240-247.

Jagongo, A. & Mutswenje, V. (2014). A survey of the factors influencing investment decisions: the case of individual investors at the NSE. *International Journal of Humanities And Social Science*, 4(4), 92-102.

Jain, A.K. (2010). Data clustering: 50 years beyond k -means. *Pattern Recognition Letters*, 31, 651 –666.

Jain, R., Jain, P., & Jain, C. (2015). Behavioral biases in the decision making of individual investors. *IUP Journal of Management Research*, 14(3), 7-27.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 281–297.

Jiang, X.D., & Zhu, X.B. (2006). 中国股市高换手率的成因. 经济导刊, 15(8), 75-79. (in Chinese).

Johnson, H. (2016). Our average net worth by age: How do you compare? Building Financial Independence & Retiring Early. Retrieved 25 March 2017 from http://www.investmentzen.com/blog/average-net-worth-by-age-american-households/

Jureviciene, D. & Jermakova, K. (2012). The Impact of Individuals' Financial Behaviour on Investment Decisions. *Electronic International Interdisciplinary Conference*.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica*. 47(2), 263-292.

Kaiser, H.F. & Rice, J. (1974). Little Jiffy, Mark IV. *Journal of Educational and Psychological Measurement*, 3(1), 111-117.

Kaleem S., Hassan F. & Saleem A. (2009). Influence of environmental variations on physiological attributes of sunflower. *African Journal of Biotechnology*, 8, 3531-3539.

Kašćelan, L., Kašćelan, V., & Jovanović, M. (2014). Analysis of investors' preferences in the Montenegro stock market using data mining techniques. *Economic Research-Ekonomska Istraživanja*, 27(1), 463-482.

Kaustia, M., & Knüpfer, S. (2008). Do investors overweight personal experience? Evidence from IPO subscriptions. *The Journal of Finance*, 63(6), 2679-2702.

Ke, D.M., Ng, L., & Wang, Q.H. (2010). Home bias in foreign investment decisions. *Journal of International Business Studies*, 41(6), 960-979.

Kerby A., & Lawrence, J. (2003). A Multivariate Statistical Analysis of Stock Trends, Alma College Miami University, Alma, MI Oxford,OH.

Keith, T. (2006). *Multiple regression and beyond* (1st ed.). Boston, Mass.: Pearson Education.

Khovrak, I., and Petchenko, M. (2015). Estimating the level of financial safety in banking institutions. *Actual Problems in Economics*, 164, 347-354.

Kim, B. (2015). Should I always transform my variables to make them normal? *Research Data Services and Sciences*. University Virginia Library.

Kim, B.J., Kim, I.K., & Kim, K.B. (2004). *Feature extraction and classification system for nonlinear and online data* (1st ed., 171-180). Sydney, Australia: Proceedings of the 8th Pacifc-Asia Conference on Knowledge Discovery and Data Mining.

Kim, M. J., Nelson, C., & Startz, R. (1991). Mean reversion in stock prices? A reappraisal of the empirical evidence. *The Review of Economic Studies,* 58, 515–528.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32(4), 995 - 1003.

Knight, F.H. (1921). *Risk, Uncertainty and Profit.* Boston: Houghton-Mifflin, 1921; reprinted New York: Sentry Press, 1964.

Kodinariya, T.M., & Makwana, P.R. (2013). Review on determining number of clusters in K-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), 90-95.

Kopeti H. (2010), Pattern classification of stock moving, a dissertation submitted to the University of Manchester, School of Computer Science 7 HARINATH KOPETI 7562567.

Korniotis, G., & Kumar, A. (2006). *Does investment skill decline due to cognitive aging or improve with experience?* Working paper. University of Notre Dame, Indiana.

Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, 275, 1-12.

Kumar, M., & Kalia, A. (2011). Mining of emerging pattern: discovering frequent itemsets in a stock data. *International Journal of Computer Technology and Applications*, 2(6), 3008-3014.

Kumar, N. (2012). Data Mining for Business Intelligence-Concepts, Techniques, and Applications in Microsoft Office Excel® with XLMiner®. *Journal of Quality Technology*, 44(1), 81-83.

Kuo, J., Han, X., Hsiao, C., Yates III, J., & Waterman, C. (2011). Analysis of the myosin-II-responsive focal adhesion proteome reveals a role for β-Pix in negative regulation of focal adhesion maturation. *Nature Cell Biology*, 13(4), 383-393.

Kuo, R.J., Chao, C.M., & Chiu, Y.T. (2009). Application of particle swarm optimization to association rule mining. *Applied Soft Computing*, 11(1), 326-336.

Kuo, R.J., Lin. S.Y., & Shih, C.W. (2007). Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan. *Expert Systems with Applications*, 33, 794-808.

Kwan, I.S.Y., Fong, J., & Wong, H.K. (2005). An e-customer behavior model with online analytical mining for internet marketing planning. *Decision Support Systems*, 41(1), 189 - 204.

Kwon, O. & Yanga, J.S. (2008). Information flow between composite stock index and individual stocks. *Physica A: Statistical Mechanics and its Applications*, 387(12), 2851–2856.

Laerd Statistics (2013). *Descriptive and Inferential Statistics*. Retrieved from https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php

Lahsasna, A. (2009). *Evaluation of credit risk using evolutionary-fuzzy logic scheme*. MS Thesis, University of Malaya.

Lai, K., & Cerpa, N. (2001). *Support vs Confidence in Association Rule Algorithms*. In Proceedings of the OPTIMA Conference, Curicó, October 10-12, 2001.

Laio, S.H., & Chen, Y.J., and Lin, Y.T. (2011). Mining customer knowledge to implement online shopping and home delivery for hypermarkets. *Expert Systems with Applications*, 38, 3982-3991.

Lam, K.S.K., & Qiao, Z. (2015). Herding and fundamental factors: The Hong Kong experience. *Pacific-Basin Finance Journal*, 32, 160-188.

Lau, H.C.W., Ho, G. T. S., Chu, K. F., Ho, W., & Lee, C. K. M. (2009). Development of an intelligent quality management system using fuzzy association rules. *Expert Systems with Applications*, 36(2), 1801-1815.

Lee, A. (2015). China begins internet finance clampdown. *International Financial Law Review* (Aug 3, 2015).

Lee, A.J.T., Lin, M., Kao, R., & Chen, K. (2010). An effective clustering approach to stock market prediction. *AIS Electronic Library*.

Lee, B.S. (1995). The response of stock prices to permanent and temporary shocks to dividends. *The Journal of Financial and Quantitative Analysis*, 30, 1–22.

Lee, S.S. (2012). Jumps and information flow in financial markets. *The Review of Financial Studies*, 25(2), 439-479.

Lewellen, W.G., Lease, R.C., & Schlarbaum, G.G. (1977). Patterns of investment strategy and behavior among individual investors. *Journal of Finance*, 50, 296-333.

Li, L. (1999). Behavioral finance theory to the challenges of the efficient market hypothesis. *Economic Science*, 3, 63-71.

Li, X., Wong, W., Lamoureux, E.L., and Wong, T.Y. (2012). Are linear regression techniques appropriate for analysis when the dependent (outcome) variable is not normally distributed? *ARVO Journal of Investigative Ophthalmology & Visual Science*, 53(6), 3082-3083.

Li, Y., Cai, H.J., & Tan. H. (2008). Frequent patterns of investment behaviors in Shanghai stock market. In: *Proceedings of 2008 International Conference on Computer Science and Software Engineering*, 4(12-14 December), 325-328.

Liao, S., Ho, H. and Lin, H. (2008). Mining stock category association and cluster on Taiwan stock market. Expert Systems with Applications, 35, 19-29.

Liao, S.H., & Chou, S.Y. (2013). Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio. *Expert Systems with Applications*. 40(5), 1542–1554.

Liao, S.H., Chen, C.M., & Wu, C.H. (2008). Mining customer knowledge for product line and brand extension in retailing. *Expert Systems with Applications*, 34, 1763-1776.

Lim, A. H. L., Lee, C. S., & Raman, M. (2012). Hybrid genetic algorithm and association rules for mining workflow best practices. *Expert Systems with Applications*, 39(12), 10544-10551.

Lima, E.J., & Tabak, B.M. (2004). Tests of the random walk hypothesis for equity markets: Evidence from China, Hong Kong and Singapore. *Applied Economics Letters*, 11(4), 255-258.

Lin, C.T., & Lee, C.S.G. (1991). Neural-network-based fuzzy logic control and decision system. *IEEE Transactions on Computers*, 40(2), 1320-1336.

Lin, C.Y., Ho, P.H., Shen, C.H., & Wang, Y.C. (2016). Political connection, government policy, and investor trading: Evidence from an emerging market. *International Review of Economics & Finance*, 42, 153–166.

Liu, H., Tarima, S., Borders, A.S., Getchell, T.V., Getchell, M.L., & Stromberg, A.J. (2005). Quadratic regression analysis for gen discovery and pattern recognition for non-cyclic short time-course microarray experiments. *BMC Bioinformatics*, 6(106).

Lin, J.H., & Jou, R. (2005). Financial e-commerce under capital regulation and deposit insurance. *International Review of Economics & Finance*, 14(2), 115–128.

Liu, K.F.R., & Lai, J.H. (2009). Decision-support for environmental impact assessment: A hybrid approach using fuzzy logic and fuzzy analytic network process. *Expert Systems with Applications*, 36, 5119-5136.

Lo, A.W., & MacKinlay, A.C. (1988). Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies*, 1, 41–66.

Lo, A.W., & MacKinlay, A.C. (1990). When are contrarian profits due to stock market over-reaction?, *Review of Financial Studies*, 3, 175–205.

Loewenstein, G., Weber, E., Hsee, C., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127(2), 267-286.

Loudon,. (2001). *Consumer Behavior: Concepts and Applications*. Tata McGraw Hill, New Delhi.

Macgregor, D.G., Slovic, P., Berry, M., & Evensky, H.R. (1999). Perception of financial risk: A survey study of advisors and Planners. *Journal of Financial Planning*, 12(8), 68-86.

Maditinos, D., Šević, Ž., & Theriou, N. (2007). Investors' behaviour in the Athens Stock Exchange (ASE). *Studies in Economics and Finance*, 24(1), 32-50.

Maier, H.R., & Dandy, G.C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling & Software*, 15(1), 101-124.

Mallkiel, B.G., & Fama, E.F. (1970). Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25(2), 383-417.

Malmendier, U., & Nagel, S. (2009). Depression babies: Do macroeconomic experiences affect risk-taking? *NBER Paper 14813*, National Bureau of Economic Research, Inc.

Malmendier, U., & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk-taking? *Journal of Economics*, 126(1), 373-416.

Mordkoff, T.J. (2016). The assumption(s) of normality. *Quantitative Methods in Psychology*, 1-6.

Hesham M.A; Duchamp, D.; Krapp, C.A. (2015). Proceedings of the International Conference on Data Mining (DMIN): 107-115. Athens: Athens The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Madhulatha, T.S. (2012). An overview of clustering methods. *IOSR Journal of Engineering*, 2(4), 719-725.

Masini, A. & Menichetti, E. (2012). Investment decisions in the renewable energy sector: An analysis of non-financial drivers. *Technological Forescasting and Social Change,* 80(3), 510-524.

Mason, C., & Perreault Jr., W. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, 28(3), 268-280.

Mehta, N., & Dang, S. (2012). Data mining techniques for identifying the customer behaviour of investment in stock market in India. *International Journal of Marketing, Financial Services & Management Research*, 1(11), 35-55.

Misal, D.M. (2013). A study of behavioural finance and investor's emotion in Indian capital market. *International Journal of Economics and Business Modeling*, 4(1), 206-208.

Mittal, M., & Vyas, R.K. (2008). Personality type and investment choice: an empirical study. *The ICFAI University Journal of Behavioral Finance*, 5(3), 7-22.

Miyamoto, S. (2003). Information clustering based on fuzzy multisets. *Information Processing & Management*, 39(2), 195-213.

Mizutani, K., Inokuchi, R., & Miyamoto, S. (2008). Algorithms of nonlinear document clustering based on fuzzy multiset model. *International Journal of Intelligent Systems,* 23(2), 176–198.

Momeni, M., Mohseni, M., & Soofi, M. (2015). Clustering Stock Market Companies via K-Means Algorithm. *KCAJBMR*, 4(5), 1-10.

Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012). *Introduction to Linear Regression Analysis*. USA: John Wiley & Sons.

Moreno, M.N., Ramos, I., García, F. J., & Toro, M. (2008). An association rule mining method for estimating the impact of project management policies on software quality, development time and effort. *Expert Systems with Applications*, 34(1), 522-529.

Motiwalla, L., & Wahab, M. (2000). Predictable variation and profitable trading of US equities: a trading simulation using neural networks. *Computer & Operations Research*, 27, 1111–1129.

Münnix, M., Shimada, T., Schäfer, R., Leyvraz, F., Seligman, T., Guhr, T. & Stanley, H. (2011). Identifying States of a Financial Market. *Scientific Reports*, 2, 644.

Na, S.H., & Sohn, S.Y. (2011). Forecasting changes in Korea Composite Stock Price Index (KOSPI) using association rules. *Expert Systems with Applications*, 38(7), 9046–9049.

Nanda, S.R., Mahanty, B., & Tiwari. M.K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications*, 37, 8793-8798.

Narayan, P.K., Narayan, S., Popp, S., & Ahmed, H.A. (2015). Is the efficient market hypothesis day-of-the-week dependent? Evidence from the banking sector. *Applied Economics*, 47(23), 2359-2378.

Ngai, E.W.T., Xiu, L., & Chau, D.C.K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2), 2592 ‑ 2602.

Nomura Research Institute, Ltd. (2014), Internet finance growing rapidly in China. *Iakyara*, 189. 1-4.

Novák, V. (2012). Reasoning about mathematical fuzzy logic and its future. *Fuzzy Sets and Systems*, 192, 25-44.

Obamuyi, T.M. (2013). An analysis of the deposits and lending behaviours of banks in Nigeria. *International Journal of Engineering and Management Sciences*, 4, 46-54.

OECD (Organization for Economic Co-operation and Development). 2017. OECD Data. Retrieved 25 March 2017 from https://data.oecd.org/hha/household-net-worth.htm

Olson, D. & Shi, Y. (2007). *Introduction to business data mining* (1st ed.). Boston: McGraw-Hill/Irwin.

Ordoobadi, S. M. (2009). Development of a supplier selection model using fuzzy logic. *Supply Chain Management: An International Journal*, 14, 314-427.

Osborne, J.W. & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2).

Otero, L.D., & Otero, C.E. (2012). A fuzzy expert system architecture for capability assessments in skill-based environments. *Expert Systems with Application*, 39, 654-662.

Pang, J., & Wang, K. (2009). Individual investors overconfidence behaviour regarding analysis of Chinese stock market. *Times Finance*, 1, 22-23.

Petrovic, D. & Duenas, A. (2006). A fuzzy logic based production scheduling/rescheduling in the presence of uncertain disruptions. *Fuzzy Sets and Systems*, *157*(16), 2273-2285.

Pick, T., Brautigam, A., Schulz, M., Obata, T., Fernie, A., & Weber, A. (2013). PLGG1, a plastidic glycolate glycerate transporter, is required for photorespiration and defines a unique class of metabolite transporters. *Proceedings of the National Academy of Sciences*, *110*(8), 3185-3190.

Pivk, A., Vasilecas, O., Kalibatiene, D., & and Rupnik, R. (2013). On approach for the implementation of data mining to business process optimisation in commercial companies. *Technological and Economic Development of Economy*, 19(2), 237-256.

Poole, M., & O'Farrell, P. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, 52, 145-158.

Prato, T. (2005). A fuzzy logic approach for evaluating ecosystem sustainability. *Ecological Modelling*, 187(2–3), 361–368.

Qi, M., & Maddala, G.S. (1999). Economic factors and the stock market: a new perspective. *Journal of Forecasting*, 18, 151–166.

Rajola, F. (2003). *Customer Relationship Management: Organizational and Technological Perspective*s. New York: Springer.

Ravisankar, P., Ravi, V., Rao, G.R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, 50, 491–500.

Rawlings, J.O. (1988) Applied regression analysis: a research tool. Wadsworth, Pacific Grove, CA, USA.

Reitan, B. & Sorheim, R. (2010). The informal venture capital market in Norway? Investor characteristics, behaviour and investment preferences. *Venture Capital*, *2*(2), 129-141.

Ricciardi, V. & Simon, H.K. (2000). What is behavioural finance?. *Business, Education & Technology Journal*, 2(2), 1-9.

Ritter, J. (2003). Editorial Board. Pacific-Basin Finance Journal, 11(4), Pages 429–437.

Rizvi, S. & Fatima, A. (2015). Behavioural Finance: A Study of Correlation between Personality Traits with the Investment Patterns in the Stock Market. In *Managing in Recovering Markets* (1st ed., 143-155). India: Springer India.

Ronay, R., & Kim, D.Y. (2006). Gender differences in explicit and implicit risk attitudes: A socially facilitated phenomenon. *British Journal of Social Psychology*, 45, 397-419.

Rygielski, C., Wang, J.C., & Yen, D.C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, 24(4), 483–502.

Sadi, R., Asl, H., Rostami, M., Gholipour, A. & Gholipour, F. (2011). Behavioral Finance: The Explanation of Investors' Personality and Perceptual Biases Effects on Financial Decisions. *International Journal of Economics and Finance*, 3(5).

Sahi, S., Arora, A. & Dhameja, N. (2013). An Exploratory Inquiry into the Psychological Biases in Financial Investment Behavior. *Journal of Behavioral Finance*, 14(2), 94-103.

Sakar, E., Keskin, S., & Unver, H. (2011). Using of factor analysis scores in multiple linear regression model for prediction of kernel weight in Ankara Walnuts. *The Journal of Animal & Plant Sciences*, 21(2), 182-185.

Sanani, K.L. (2012). Dealing with non-normal data. The American Academy of Physical *Medicine and Rehabilitation*, 4, 1001-1005.

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students* (5th ed.). Italy: Pearson Education Limited.

Schmidt, A., & Finan, C. (2017). Linear regression and the normality. *Journal of Clinical Epidemiology*, In Press.

Seasholes, M.A., & Zhu, N. (2010). Individual investors and local bias. *The Journal of Finance*, LXV (5), 1987-2010

Seru, A., Shumway, T., & Stoffman, N. (2010). Learning by trading. *Review of Financial Studies*, 23, 705-739.

Shaikh, ARH, & Kalkundrikar, A.B. (2011). Impact of Demographic Factors on Retail Investors' Investment Decisions- An Exploratory Study. *Indian Journal of Finance*, *5*(9), 35-44.

Shameli, G., Patel, N., & Bruce, P. (2007). *Data Mining for Business Intelligence*. Canada: John Wiley & Sons, Inc.

Shapiro, A.F. (2002). The merging of neural networks, fuzzy logic, and genetic algorithms Insurance. *Mathematics and Economics*, 31(1), 115–131.

Shapiro, A.F. (2004). Fuzzy logic in insurance. *Insurance: Mathematics and Economics*, 35(2), 399-424.

Sharma, S., & Shebalkov, M. (2013). Application of neural network and simulation modeling to evaluate Russian banks' performance. *Journal of Applied Finance & Banking*, 3(5), 19-37.

Shefrin, H. (2001). Behavioural corporate finance. *Journal of Applied Corporate Finance*, 14(3), 113-126. doi:10.1111/j.1745-6622.2001.tb00443.x

Shefrin, H., & Statman, M. (2011, November). Behavioral finance in the financial crisis: Market efficienct, Minsky, and Keynes (Working Paper). Santa Clara University

Shieh, G. (2010). On the misconception of multicollinearity in detection of moderating effects: Multicollinearity is not always detrimental. Multivariate Behavioral Research, 45, 483-507.

Shiller, R.J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17(1), 83-104.

Shleifer, A. (2000). Inefficient Markets: An Introduction to Behavioral Finance (1st ed.). New York: Oxford University Press Inc.

Shmueli, G., Patel, N.R., & Bruce, P.C. (2007). *Data mining for business intelligence concepts, techniques, and applications in Microsoft Office Excel with XLMiner*, N.J.: Wiley Interscience, Hoboken.

Shweta, G. (2014). Investor's Herding Behavior and Investment Performance: An Empirical Evidence from Delhi. *The International Journal of Business & Management*, 2(12), 56-59.

Simon, H.A. (1987). Satisfying. *In:* Newman P. (ed.). *The New Palgrave: A Dictionary of Economics*. London, UK: Macmillan, 243-245.

Sohn, S.Y., & Kim, Y. (2008). Searching customer patterns of mobile service using clustering and quantitative association rule. *Expert Systems with Applications*, 34, 1070-1077.

Solomon, M.R., Russell-Bennett R, & Previte, J. (2012). *Consumer Behaviour: Buying, having, being. 3rd Edition*. Australia: Pearson

Someswar, G.M., Satheesh, B., & Vivekanand, G. (2012). Finance Mining – Analysis of Stock Market Exchange for Foreign Using Classification Techniques. *International Journal of Engineering Research and Applications*, 2(4), 717-723

Speelman, C.P., Clark-Murphy, M., & Gerrans, P. (2013). Decision making clusters in retirement savings: Gender differences dominate. *Journal of Family and Economic Issues*, 34(3), 329-339.

Spyrou, S. (2013). Herding in financial markets: a review of the literature. Review of Behavioral Finance, 5(2), 175-194.

Sreekantha, D.K., & Kulkarni, R.V. (2008). Industrial loan processing using neuro fuzzy logic. *International Journal of Intelligent Information Processing*, 2, 305–318.

Sreekantha, D.K., & Kulkarni, R.V. (2012). Expert system design for credit risk evaluation using neuro-fuzzy logic. *Expert Systems*, 29(1), 56-69.

Steen, G.J. (1991). The empirical study of literary reading: Methods of data collection. *Poetics*, 20(5-6), 559-575.

Stevens, J.P. (1992). *Applied Multivariate Statistics for the Social Science (2nd edition)*. Hillsdale, NJ: Erlbaum.

Suliman, A. & Obaid, S. (2013). Application of Principal Component Method and k-means clustering algorithm for Khartoum stock Market. Nat Sci, 11(7), 108-112.

Sun, C. (2015). Hong Kong ‘still top choice for China’s rich’ investing outside the Mainland and overseas. *South China Morning Post*. Retrieved 10 November 2016, from http://www.scmp.com/news/china/money-wealth/article/1808983/hong-kong-still-top-choice-chinas-rich-investing-outside

Sung, H.N, & So, Y.S. (2011). Forecasting changes in Korea composite stock price index (KOSPI) using association rules. *Expert Systems with Applications*, 38, 9046-9049.

Tahera, K., Ibrahim, R.N., & Lochert, P.B. (2008). A fuzzy logic approach for dealing with qualitative quality characteristics of a process. *Expert Systems with Applications*, 34(4), 2630–2638.

Tan, L., Chiang, T.C., Mason, J.R., & Nelling, E. (2008). Herding behavior in Chinese stock markets: An examination of A and B shares. *Pacific-Basin Finance Journal*, 16(1-2), 61-77.

Tan, P.N., Steinbach, M., & Kumar, V. (2005). Association analysis: Basic concepts and algorithms. *Introduction to Data Mining*. Boston: Addison-Wesley.

Tang, H., Tan, K.C., & Yi, Z. (2007). *Neural Networks: Computational Models and Applications*. Springer-Verlag Berlin Heidelberg.

Tekçe, B., Yılmaz, N., & Bildik, R. (2016). What factors affect behavioral biases? Evidence from Turkish individual stock investors. Research in International Business and Finance, 37, 515-526.

Thaler, R. & Johnson, E. (1990). Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. *Management Science*, 36(6), 643-660.

Thaler, R.H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, 12(3), 183-206.

Thaler, R.H. (1999). The end of behavioural finance. *Financial Analysts Journal*, 55(6), 12-17.

Thaler, R.H. & Sunstein, C.R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven and London: Yale University Press.

Thomas J. R. & Nelson J. K. (2005). *Research Methods in Physical Activity* (5th ed.). Champaign, Illinois: Human Kinetics.

Townsend, R. (1994). Risk and insurance in village India. *Econometrica,* 62(3), 539-591.

Ting, S.L., Shum, C.C., Kwok, S.K., Tsang, A.H.C., & Lee, W.B. (2009). Data mining in biomedicine: Current applications and further directions for research. *Journal of Software Engineering & Applications*, 2(3), 150 - 159.

Tsai, C.F., Lin, Y.C., & Wang, Y.T. (2009). Discovering stock trading preferences by self-organizing maps and decision trees. *International Journal on Artificial Intelligence Tools*, 18, 603–611.

Vazirgiannis, M., Halkidi, M., & Gunopulos, D. (2003). *Uncertainty Handling and Quality Assessment in Data Mining*. Hong Kong: Springer.

Velmurugan, T., & Santhanam, T. (2011). A survey of partition based clustering algorithms in data mining: An experimental approach. *Information Technology Journal*, 10(3), 478-484.

Wang, H., & Hanna, S. (1997). Does risk tolerance decrease with age? *Financial Counseling and Planning*, 8(2), 27–32.

Watson, J., & McNaughton, M. (2007). Gender differences in risk aversion and expected retirement benefits. *Financial Analysts Journal*, 63(4), 52-62.

Weber, E.U., Blais, A., & Betz, N.E. (2002). A domain - specific risk - attitude scale: measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4), 263-290.

Wen, L., & Hao, Q. (2013). Consumer investment preferences and the Chinese real estate market. *International Journal of Housing Markets and Analysis*, 6(2), 231-243.

Wong, A. & Quesada, J.A. (2009). *El Comportamiento Humano en las Finanzas PriceWaterHouseCoopers IMEF*. M´exico, D.F.: Instituto Mexicano de Ejecutivos de Finanzas.

Wong, B.K., & Lai, V.S. (2011). A survey of the application of fuzzy set theory in production and operations management: 1998-2009. *International Journal of Production Economics*, 129(1), 157-168.

Wong, K.A., & Kwong, K.S. (1984). The behaviour of Hong Kong stock prices. *Applied Economics*, 16(6), 905-917.

Wong, W. & Dr Lai, M. (2009). *Investor Behaviour and Decision-Making Style: A Malaysian Perspective (1st ed.).*

Wu, D. (2009). Supplier selection: a hybrid model using DEA, decision tree and neural network. *Expert Systems with Applications*, 36, 9105–9112.

Wu, R.S., Ou, C.S., Lin, H.Y., Chang, S.I., & Yen, D.C. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, 39(10), 8769–8777.

Wu, X. & Kumar, V. (2009). *The top ten algorithms in data mining* (1st ed.). Boca Raton, New York: CRC Press.

Xu, Y. (2016). Research of Association Rules Algorithm in Data Mining. *International Journal of Database Theory and Application,* 9(6), 119-130.

Yin, R.K. (2009). *Case Study Research: Design and Methods*. SAGE Publications, Inc.

Yong, A.G., & Pearce, S. (2013). A Beginners Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79-94.

Yu, X. & Yiu, E. (2016). Hong Kong funds welcomed by Mainland investors. *South China Morning Post*. Retrieved 10 November 2016, from http://www.scmp.com/business/markets/article/1939572/hong-kong-funds-welcomed-mainland-investors

Zadeh, L. (2008). Is there a need for fuzzy logic?. *Information Sciences*, 178(13), 2751-2779.

Zhang, Y. & Zheng, X. (2015). A study of the investment behaviour based on behavioural finance. *European Journal of Business and Economics*, 10(1), 1-5.

Zhang, D. & Zhou, L. (2004). Discovering Golden Nuggets: Data Mining in Financial Application. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 34(4), 513-522.

Zhou, W.X., & Sornette, D. (2006). Is there a real-estate bubble in the U.S.? *Physica A*, 361(1), 297-308.

Zumbrun, J. (2017). U.S. household net worth reaches record $92.8 trillion. *The Wall Street Journal*. Retrieved 25 March 2017 from https://www.wsj.com/articles/u-s-household-net-worth-reaches-record-92-8-trillion-1489078918