CUSTOMER SEGMENTATION BASED ON THE RFM ANALYSIS MODEL USING K-MEANS CLUSTERING TECHNIQUE: A CASE OF IT SOLUTION AND SERVICE PROVIDER IN THAILAND

CUSTOMER SEGMENTATION BASED ON THE RFM ANALYSIS MODEL USING K-MEANS CLUSTERING TECHNIQUE: A CASE OF IT SOLUTION AND SERVICE PROVIDER IN THAILAND

Ponlacha Rojlertjanya

This Independent Study Manuscript Presented to

The Graduate School of Bangkok University

in Partial Fulfillment

of the Requirements for the Degree

Master of Management (Business Innovation)

2019

This manuscript has been approved by

the Graduate School

Bangkok University

Title: Customer Segmentation Based on The RFM Analysis Model Using K-Means Clustering Technique: A Case of IT Solution and Service Provider in Thailand

Author: Mr. Ponlacha Rojlertjanya

Independent Study Committee:

Advisor [redacted]

(Dr. Ronald Vatananan-Thesenvitz)

Field Specialist [redacted]

(Dr. Xavier Parisot)

[redacted]

(Dr. Suchada Chareanpunsirikul)

Dean, Graduate School

August 16, 2019

Rojlertjanya, P. Master of Management (Business Innovation), August 2019, Graduate School, Bangkok University.

Customer Segmentation Based on The RFM Analysis Model Using K-Means Clustering Technique: A Case of IT Solution and Service Provider in Thailand (91 pp.)

Advisor: Ronald Vatananan-Thesenvitz, Ph.D.

ABSTRACT

Customer segmentation is crucial for every business to better understand their customers, to keep customers satisfied, and to develop personalized products and services. In this research, a case study of using data mining techniques to segment customers for an Information Technology (IT) solution and service provider in Thailand is presented. The objectives of this research are to construct a customer segmentation model based on customer demographics and purchase behaviours and to help business better understand its customers and support their customer-centric marketing strategy. The proposed segmentation model is regarding to the customers demographic data and Recency, Frequency, and Monetary (RFM) values generated from purchase behaviours, customers have been segmented using the K-means clustering technique into numerous groups based on their similarity, and the profile for each group is identified based on their characteristics. Accordingly, recommendations are provided to the business on marketing strategy and further analysis. RapidMiner Studio, data mining tool, is used in this research.

Keywords: IT Solution and Service Provider, Customer Segmentation, RFM Analysis, K-Means Clustering, Data Mining

ACKNOWLEDGEMENT

This independent study would not have been possible without the support of a number of fantastic individuals. My sincere appreciation to all of them for being part of my study and making this independent study possible.

I would like to express my gratitude to Dr. Ronald Vatananan-Thesenvitz, my advisor, for his patient guidance, enthusiastic encouragement, useful comments, and assistance in keeping my progress on schedule.

Besides my advisor, I would like to thank all staffs of a case study's company, who gave the permission to use all required information and necessary materials to complete this independent study.

Finally, my deep and sincere gratitude to my family for their continuous and exceptional love, help, and support. I am forever indebted to my dad and mom for giving me the wonderful opportunities and experiences that have made me who I am. This journey would not have been possible if not for them, and I dedicate this work to them.

Ponlacha Rojlertjanya

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (CONTINUED)

LIST OF TABLES

Page

LIST OF TABLES (CONTINUED)

Page

LIST OF FIGURES

Page

LIST OF FIGURES (CONTINUED)

Page

CHAPTER 1

INTRODUCTION

1.1 Research Background

In the past, most of the businesses used the product-centric as a marketing strategy, which mainly focused on the manufacturing to develop a better quality of products and to reduce the cost of goods manufactured rather than pay attention to the customers who purchase and use their, because they can gain a lot of profits from the market share, according to the principles of economies of scale and scope (Shah, T. Rust, Parasuraman, Staelin, & S. Day, 2006; Kenton, 2019a, 2019b).

The third industrial revolution, also known as the digital revolution or information technology (IT) revolution in the latter half of the 20th century introduced the new way of collecting, storing, processing, and transmitting the digital data and information (Forester, 1986). At that time, the customer-centric marketing was introduced as a new marketing strategy, which shifting from placing the products at the center of marketing design and delivery, to be an individual customer (Rygielski, Wang, & Yen, 2002; Custora, 2018). Eventually, the marketers had adopted the technology such as direct call and email, to reach their customer and tried to satisfy the needs of each customer with the customized and personalized products and services, which corresponding to the principle of mass customization, as Hart (1995) defined that, the mass customization is “the ability to provide your customers with anything they want profitably, any time they want it, anywhere they want it, any way they want it”.

Knowledge of customer behaviour is crucial to provide them the high quality products and services, to keep them royal to the business and to make big profits to the business. For this reason, the customer relationship management (CRM) was introduced to achieve the needs of customers while enhancing the strength of sales and marketing for the business (Cheng & Chen, 2009). Regarding to Swift (2001) and Parvatiyar & Sheth, 2001; Ngai, Xiu, & Chau (2009), CRM consists of four dimensions: Customer identification, Customer attraction, Customer retention, and Customer development. The customer segmentation was arranged in the first dimension of CRM, customer identification, that provide the ability to identify groups of customers that share common characteristics and behaviours. For the customer retention and customer development dimension, Demographic variables and Recency, Frequency, and Monetary (RFM) were utilized (Namvar, Gholamian, & Khakabi, 2010).

As CRM could help the business to better understand their customers and effectively allocation resources to the most potential group of customers. However, the incompetence to discover the meaningful information out of the collected raw data inhibit the business benefiting from the valuable knowledge from their customer (Berson, Smith, & Thearling, 2000). As the meaningful information is generally sophisticated and overlooked, the data mining techniques like clustering, classification, and association could help to finding meaningful information and patterns from the customer data (Rygielski et al., 2002).

In this research, a case study of using data mining techniques to segment customers for an IT solution and service provider in Thailand is presented. The objectives of this research are to construct a customer segmentation model based on customer demographics and purchase behaviours and to help business better understand its customers and support their customer-centric marketing strategy. Regarding to the customers demographic data and RFM values generated from purchase behaviours, customers have been segmented using the K-means clustering technique into numerous groups based on their similarity, and the profile for each group is identified based on their characteristics. Accordingly, recommendations are provided to the business on marketing strategy and further analysis. RapidMiner Studio is used in this research.

1.2 Statement of the Problems

For a successful business today, identification of your customers behaviours, their requirements and needs, and keeping them royal is a key task for marketers, sales, and business owner. It is not just about building better relationships with your customers, but it is also a source of lasting competitive advantage for your business. Many businesses are having tons of customer's data, but they cannot make use of it, because lacking skills and knowledge on how to manage and transform data into meaningful knowledge. To overcome this issue, the simple yet effective analytic model is required.

1.3 Research Objectives

This research proposes a case study of using data mining techniques, K-means clustering, and RFM analysis to construct a customer segmentation model, to segment customers for an IT solution and service provider business in Thailand based on their customer demographics and purchase behaviours. The findings, customer segments, are intended to be helpful for business to better understand its customers and can be supported their customer-centric marketing strategy.

1.4 Scope of Research

1. Construct the customer segmentation model using K-means clustering technique and RFM analysis.

2. Conduct the customer segmentation for an IT solution and service provider business in Thailand based on their customer demographics and purchase behaviours.

3. Analyze results and provide recommendations to the business on marketing strategy for each customer segment.

4. Conduct a face-to-face semi-structured interview with a company representative to discuss about research findings and recommendation for further analysis.

1.5 Research Goals and Questions

Table 1.1: Research Goals and Questions

Research Goal	**Research Question**
RG1: Construct the customer segmentation model using K-means clustering technique and RFM analysis.	RQ1.1: How data mining is used in customer segmentation?
	RQ1.2: How to apply RFM analysis to customer segmentation?
	RQ1.3: How to analyze RFM data using K-means clustering?

(Continued)

Table 1.1 (Continuted): Research Goals and Questions

Research Goal	Research Question
RG2: Conduct the customer segmentation based on customer demographics and purchase behaviours.	RQ2.1: How to segment customers with demographics?
	RQ2.2: How to segment customers with purchase behaviours?
	RQ2.3: How to combine or manage two types of segmentations models?
RG3: Identify the characteristic of customers in each segment and provide recommendations to the business.	RQ3.1: How to interpret and transform the findings, customer segments, into marketing strategy?
	RQ3.2: How to measure the results?

1.6 Significance of the Research

Academic Outcomes - The customer segmentation model constructed in this research is a demonstration of how data mining can be applied to a marketing area and how customer segments can be supported business in marketing strategy.

Practitioner Outcomes - The customer segments resulted in this research are used to enhance the business to better understand its customers in term of characteristics and behaviours, as well as, the recommendations of each segment provided are contemplated to support business for developing and improving the customer-centric marketing strategy, to define tactical marketing campaigns, to personalize products and services for the up-selling and cross-selling programs.

1.7 Definition of Terms

Product-Centric Marketing	An approach that places a focus on products and services and try to sell that individual products and services to as many customers as it possibly can.
Customer-Centric Marketing	An approach that places the individual customer at the center of marketing design and delivery for creating a positive customer experience.
Unsupervised Learning	A method used to enable machines to classify both tangible and intangible objects, patterns, and relationships without providing the machines any prior information about the objects.
Centroid	The middle of a cluster. A centroid is a vector that contains the mean of a variable for the observations in that cluster.
Cluster Distance	The distances between cluster centroids measures how far apart the centroids of the clusters from one another.

CHAPTER 2

LITERATURE REVIEW

In this chapter, a general review of related theories and studies will be discussed. Besides, referring to previous researches about a case study of RFM model-based customer segmentation, a two phase clustering method, and data mining techniques for customer relationship management. Lastly, the conceptual framework will be presented.

2.1 Customer Relationship Management (CRM)

Customer Relationship Management (CRM) is refers to practices, strategies, and technologies that companies use to better understand its markets and customers with the goals of enhancing relationship with their customer and to maximize the customer value (Rouse, Ehrens, & Kiwak, 2019). Generally, CRM is a two-stage concept. Firstly, the company need to build a customer focus and move away from product orientation, which follows the principle of mass customization and customer-centric marketing. The marketing strategy should be defined based on its customer needs rather than product features and functions. In the second stage, companies need to integrate their systems with CRM across entire customer journey and experience, by adopt and leverage IT and technology to support monitoring and managing relationship tasks between them and its customer (Rygielski et al., 2002).

Regarding to Swift (2001), Parvatiyar & Sheth (2001), and Ngai et al. (2009), CRM consists of four dimensions (see Figure 2.1: Customer Relationship Management Four Dimensions).

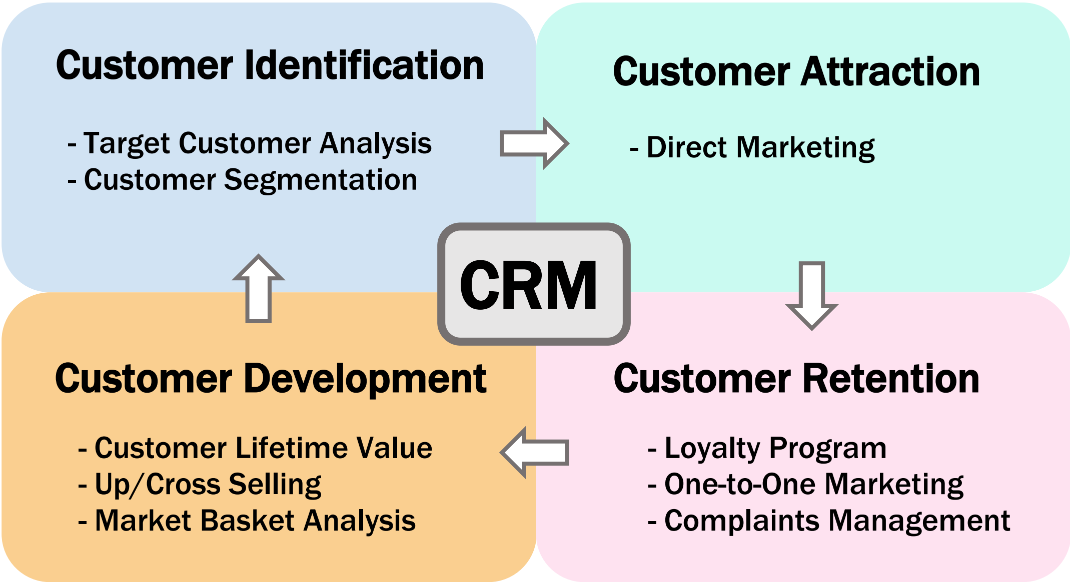


Figure 2.1: Customer Relationship Management Four Dimensions (Adapted from Ngai, Xiu, & Chau, 2009)

1) Customer Identification: CRM begins with the customer identification. This phase involves detailed customer intelligence (CI), the process of gathering and analyze customer's data, activities, and behavior Shaw & Reed (1999), to identify who are the most likely to become customer or most profitable customers and also can be used to identify which products to sell to which customers based on historical purchasing behavior. Woo, Bae, & Park (2005) state that, there are two elements for customer identification: target customer analysis and customer segmentation. The target customer analysis is used for seeking the specific group like the most profitable group or loyalty group through analysis of customer data, in comparison of customer segmentation that trying to grouping customers into smaller group based on its similarity.

2) Customer Attraction: This is a phase following the customer identification. Since, the target groups were identified by target customer analysis or customer segmentation, companies can now direct attracting the target customers. The direct

marketing described as the promotion process, which motivates and drives customers to place orders through various channels such as direct call, direct email, and coupon distribution (Thanuja, Venkateswarlu, & Anjaneyulu, 2011).

3) Customer Retention: Satisfaction, loyalty, and commitment are the essential components of an effective customer retention in CRM (Terblanche, 2006). Customer retention identifies and helps retain potential customers, leverages customer base, and also support companies to encourage less profitable customers to become loyal and more profitable (Nataraj, 2010). The one-to-one marketing, loyalty programs, and complaints management are important elements of the customer retention, refers to customized and personalized marketing campaigns that supported by results of customer behaviour analysis, and predicting in behaviour change (Chen, Chiu, & Chang, 2005).

4) Customer Development: This CRM phase is aim to maximize the transaction intensity, transaction value, and customer profitability. The customer lifetime value (LTV) analysis, up selling and cross selling, and market basket analysis are involving in this phase to achieve the goals. Thanuja, et al. (2011) state that a customer lifetime value analysis is defined as the pridiction of the net profit that companies can expect from the entire future relationship with a customer. Up and cross selling refers to method aiming to rise the value of a single transaction. Cross selling is provide existing customers the opportunity to purchase additional services and products offered, while up selling is to selling a more expensive and better services and products to the customers who already considering purchasing (Kubiak & Weichbroth, 2010).

2.2 Recency, Frequency, and Monetary (RFM) Analytic Model

The Recency, Frequency, and Monetary (RFM) analytic model is proposed by Hughes, Strategic database marketing (1994), is one of the important model for companies to formulate marketing strategies (Hughes, 2012). The RFM model represents customers' consumption behaviours based on the transaction database, which is simplified into three variables (attributes) as follows:

1) Recency (R): R represents recency, which refers to the period of time starting from recent consuming behaviour happens (last purchasing) and present. The date closer to present, the higher possibility customers will make another purchase. Thus, it will has a higher value in recency variable.

2) Frequency (F): F represents frequency, which refers to the number of transactions in a particular period. It is expect that the higher purchase frequency of customers, the higher loyalty customers are, and the higher customer value to the company. The many the frequency is, the higher value in frequency variable.

3) Monetary (M): M represents monetary, which refers to the total consumption money amount in a particular period. It is expect that the higher the value of monetary, the higher the profit contributions made by customer to the company, and the higher customer value.

According to Wu & Lin (2005) research, the bigger the value of recency (R) and frequency (F) is, the opportunity customers are place new orders with the company. Furthermore, the bigger the value of monetary (M) is, the more likely the corresponding customers are purchase products or services with company again.

Many researches showed that the RFM analysis model is a good method to segment target customers from large data into three essential variables. Nonetheless, there are two different concepts were indicated in researches about how to setup the weight of three variables. In the original RFM analysis model developed by Hughes (1994), he defined the three variables as equally important. Thus, the weight of variables are identical. However, the different in the important concept was issued by Stone (1995), he indicated that the important of each variable is depends on the characteristic of industry. Thus, the weight of variables are not equal in the Stone's concept.

2.3 Data Mining

Since we are entered into the information age, companies generate gigantic of datasets such as transactions, products information, customers information, and customer feedback. In order to uncover the knowledge out of the enormously amounts of data, the automatic and powerful tools are needed. This is a reason why data mining was invented (Han, Kamber, & Pei, 2012).

Data mining, also known as knowledge discovery in database (KDD) is coming into prominence in 1994, as the process of discovering interesting relationships and patterns from gigantic amounts of data. In the same way, SAS Institue (1998) defines data mining as the process of selecting, exploring, and modeling extensive of data to expose previously unknown patterns of data. Other KDD techniques are Online Analytical Processing (OLAP), Statistical Analysis, Data Warehouse, Data Visualization, and Ad hoc queries, which require a human to ask specific questions Rajagopal (2011). In contrast, data mining does not require any human intervention.

There are three steps involved in data mining (Ramageri, 2010):

1) Exploration: The initial step of data exploration is cleaned and transformed into appropriate variables and form, then defines data based on the problems.

2) Pattern identification: Once data is explored, cleaned, and defined, the next step is to form pattern identification. Identify, determine, and select the best patterns for further analysis.

3) Deployment: Patterns are analyzed to uncover the knowledge out of the sophisticated data patterns.

Han et al. (2012) have observed various types of data which data mining can be performed. For example, characterization and discrimination; frequent patterns; , associations and correlations; classification and regression; clustering analysis; and outlier analysis. In this research, clustering analysis is focused.

2.4 K-means Clustering

Clustering is the process of grouping various objects into groups of similar objects. A cluster is a term describes a collection of data that are similar in attributes to one another within the same cluster, and are dissimilar to the objects in other clusters (Han et al., 2012). K-means is one of the most well-known clustering technique, presented by Forgy (1965). K-means algorithm has been used in many research fields and analysis, such as data mining, statistical data analysis, customer and market segmentation, and other business applications. In this research, the customer segmentation using K-means clustering technique with RFM analysis and customer demographics is proposed, that is one of business application example. Based on Han et al. (2012) research, the simplest and most essential version of clustering analysis is partitioning, which segment the objects of a set into sub-groups or clusters. The K-means procedure is summarized by (Han et al., 2012) as follows:

Algorithm: k-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

k: the number of clusters,

D: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) Arbitrarily choose *k* objects from *D* as the initial cluster centers;

(2) **Repeat**

(3) (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) Update the cluster means, that is, calculate the mean value of the objects for each cluster;

(5) **Until** no change;

For an example of K-means algorithm, Figure 2.2: Example of K-means algorithm is provided. Dataset *D* is containing 27 objects without any data patterns provided, and assume that the number of clusters, *k,* is defined as 3 clusters. After each object is assigned to the cluster based on its similarity characteristics. In this case, colour is used as a significant pattern. Thus, 3 clusters are presented, and each cluster is represent the objects with the same colour.

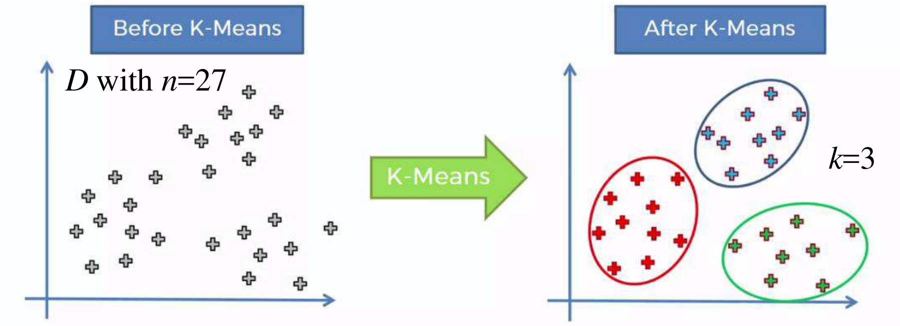


Figure 2.2: Example of K-means algorithm (Adapted from Patil, 2018)

Determining the optimal number of clusters, *k,* in dataset is a major issue in partitioning clustering, especially in K-means clustering. A variety of measures have been proposed in the literatures for evaluating number of clusters. The term clustering performance or clustering validation is used to define the procedures of evaluating clustering results (Oldach, 2019). There are several methods of clustering validation as follows:

1) The elbow method: This method probably the most well-known method, this method looks at the graph of the sum of squares at each number of clusters is calculated. The optimal number of cluster, *k*, will be determined whenever the slope of graph is changed from steep to shallow, this point called “elbow”.

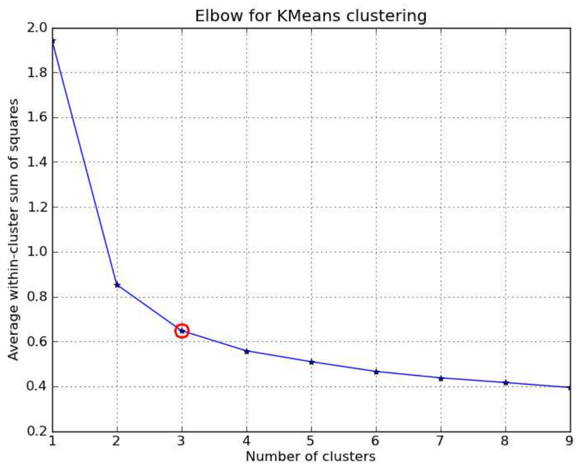


Figure 2.3: Elbow Method for K-Means Clustering (Adapted from Jordan, 2016)

The elbow method is very useful for K-means clustering technique because it shows how increasing the number of clusters, *k*, contribute grouping the instances in data set.

2) The gap statistic method: As study of Tibshirani, Walther, & Hastie (2001), the gap statistic method compares the total within intra-cluster variation for different the number of cluster, *k*, with the expected values under null reference distribution of the data. The value that maximize the gap statistic will be a *k* value.

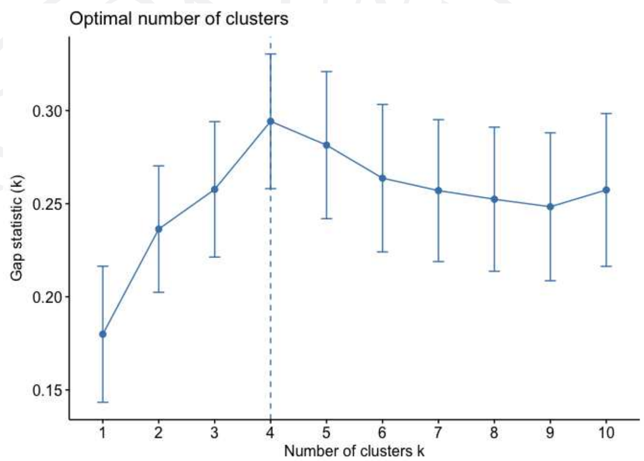


Figure 2.4: Gap Statistic Method for K-Means Clustering (Adapted from UC R Programming, 2017)

3) The average silhouette method: This method approach measures the quality of a clustering by determining how well each instance in dataset lie within the cluster. A higher value of average silhouette indicates better clustering. The optimal number of cluster, *k*, is the one that maximize the average silhouette over a range of *k* (UC R Programming, 2017).

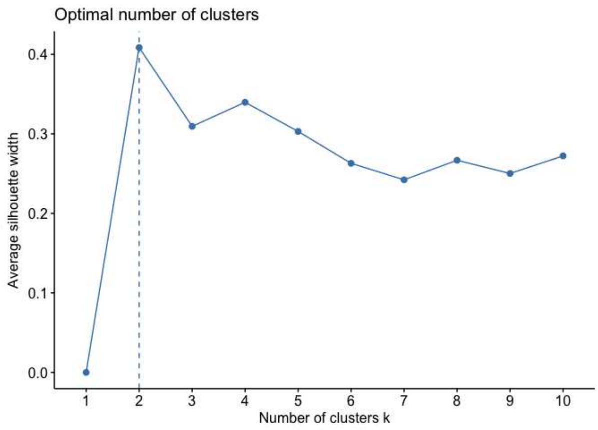


Figure 2.5: Average Silhouette Method for K-Means Clustering (Adapted from UC R Programming, 2017)

4) Davies-Bouldin index method: This method was introduced by Davies & Bouldin (1979), which can be measured the quality of clustering that has been performed by calculating the ratio between the within cluster distances and the between cluster distances. The Davies-Bouldin index's ratio is bounded from -0 to 1, lower score indicates the better performance of clustering. The limitation of Davies-Bouldin index method is the clustering technique is restricted to using only Euclidean distance function.

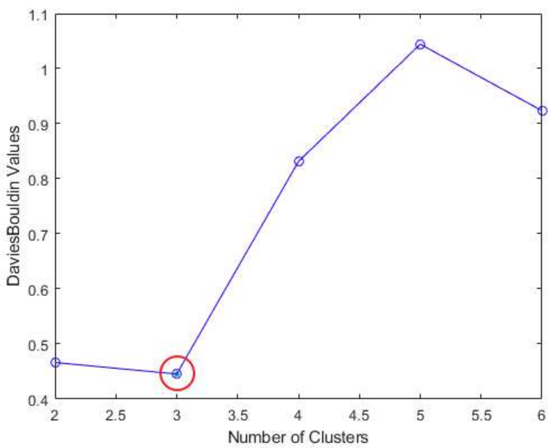


Figure 2.6: Davies-Bouldin Index Method for K-Means Clustering (Adapted from MathWorks, 2019)

Naeem & Wumaier (2018) conclude that there is no proper solution to find the true *k* value because each method determines in different dimensions and variables as illustrated in Table 2.1, it is trustworthy but there are some heuristic rules used to determine the value of *k*. However, there are some dimensions in common that highly effected to the number of clusters in K-means clustering. For example, initial centroid selection, outliers and noise, preprocessing, and high dimensionality (large spare data).

Table 2.1: Comparison of K Value Selection Methods

Methods	**Focused Dimensions**
Elbow Method	1. Ratio of variance outcome 2. Sum of Squared Errors (SSE) or Distortion
Gap Statistic Method	1. Cluster dispersion change
Average Silhouette Method	1. How closely it is matches to data point within its own cluster 2. How loosely it is match to data of the neighboring cluster
Davies-Bouldin Index Method	1. Intra-cluster distance (with-in cluster scatter distance) 2. Inter-cluster distance (between cluster separations)

2.5 Customer Segmentation

The concept of customer segmentation, also known as market segmentation, is introduced by Smith (1956) as a process of dividing the market into sub-groups or segments based on several factors such as demographic, geographic, behavioural, and psychographic. Customer segmentations provides ability to have a better understanding of customers for marketers and also make company's marketing strategy more effective (Gunter & Furnham, 1992).

Regarding to Smith (1956) research, there are four main types of customer segmentation: (1) Demographic, (2) Geographic, (3) Behavioural, and (4) Psychographic.

1) Demographic segmentation: The demographic segmentation divides customers into groups regarding to demographic variables such as age, gender, education, income, occupation, religion, social class, nationality, family size (Armstrong & Kotler, 2005). The demographic segmentation is one of the most popular type for customer segmentation because customer information are easy to identify and measure (Gunter & Furnham, 1992; Armstrong & Kotler, 2005). For example, customer information like age and gender are easily to gathered via online account registration, member registration, questionnaire, and others. This make easier, effortless, and costly for business to gather and make use of information about its customer (Gunter & Furnham, 1992; Armstrong & Kotler, 2005).

Demographic segmentation can help a business identify and pinpoint what product and services to target and what type of advertise or promotion to use with particular segment (3X 2Y Digital , 2013). Another advantage is customer loyalty and retention. When a business takes time focusing on each customer segment and develops products and services to match with its customer's needs, customers will come back or recommend their friends and family. However, similar demographics do not always signify similar needs within a particular segment. The effectiveness of the marketing campaign or message may be frustrated by a one-size-fits-all approach (Walker, 2017).

2) Geographic segmentation: This customer segmentation type is using customer's geographical areas such as nations, states, regions, cities as criteria to perform a segmentation. Many companies benefit from geographic segmentation to customize their products, services, advertising, promotion campaigns and marketing approaches to fit and appropriate with different need from different geographical variables (Armstrong & Kotler, 2005). Geodemographics was mentioned by Gunter & Furnham (1992) as a result of an increase in the globalization today, which represents a combination of geographic and demographic segmentation. Thus, the geodemographics classify and divides customers according to where they live or where they make a transaction in comparison to the social class and customer profiles (Gunter & Furnham, 1992).

3X 2Y Digital (2013) states that geographic segmentation can target areas with specific customer demographics, can target specific regions depending on business location, can identify the possible areas or locations customer can be found and also helping to figure out how to reach them. This segmentation is quite limited as it expects that all customers in a geographic area are similar in needs (McDonald & Dunbar, 2004).

3) Behavioural segmentation: The behavioural variables such as occasions, benefits, user status, usage rate, buyer-readiness stage, loyalty status, and attitude toward products and services are the good starting point for constructing and performing a customer segmentation (Kotler & Keller, 2009), seven related variables are described as follows:

- Occasions: Occasions are when the customers are grouped into segments based on period of time such as day, week, month, and year on which they think about or decide to make their purchase or use products and services (Armstrong & Kotler, 2005; Kotler & Keller, 2009). This is very useful for companies to understand its targeted customers when they construct the marketing strategies for special occasions like Christmas, New Year, Valentine’s day, etc.

- Benefits: The benefit segmentation divides customers into segments according to what are benefits customers may seek from products and services (Armstrong & Kotler, 2005). In other words, benefit segmentation results represent customer’s expectations to products and services, which company can use this to construct the up-selling and cross-selling strategy to match with customer’s expectation.

- User status: To customize marketing strategy for each group of customers, the customer status should be defined to each customer such as non-user, ex-user, potential user, first-time user, and regular user (Armstrong & Kotler, 2005).

- Usage rate: The usage rate segmentation segments customers regarding to how much they use products and services. Armstrong & Kotler (2005) defined four groups of customer *Heavy users*, *Medium users*, *Light user*, and *Non-users*. Companies seek to target a heavy users, because they generate the high percentage of the total purchasing (Gunter & Furnham, 1992). However, it is importance to not exclude the medium users, light user, and non-users due to they may provide a positive prospect for further expansion (Larsen, 2010).

- Buyer-readiness stage: The company should design their marketing strategy according to customer's awareness and interest of the products and services, because some of customers are aware or unaware of products and services, some are interested or may not interested, to lead the customer to purchase products and services at the end (Larsen, 2010).

- Loyalty status: Keeping the potential and profitable customers, loyal customer, is essential for every business, because they generate a lot of revenue and they are continually in need of company's products and services. Kotler & Keller (2009) and Larsen (2010) stated that there are four types of loyalty. *Hard-core loyals* is represents customers who are always loyal by purchase the same product or service. For customers that randomly purchase two or three brands are marked as *Split loyals.* The customers that moving from one brand to another, then moving to another brands again and again, there referred to a *Shifting loyals*. Last type is *Switches,* which represents customers who are not show any loyalty to one specific brand, but purchase product or brand at any available occasion.

- Attitude: Customer are grouped based on their attitude toward products and services such as positive, negative, enthusiastic, indifferent, and others. Understanding what customers think about products and services assists marketers to create the advertising and marketing messages that appeal to distinctive points of product and service (Hubbard).

This segmentation can build segments based on responsiveness of customers to products and services, promotion types, or payment methods from historical transactions. With today's data collection and tracking technology behavioural segmentation can provide more precise and granular insights of customer and also can help business to activate a non-user, light user, medium user to heavy user (Walker, 2017). Although the customer behaviours can be monitored and tracked, but it is not easy to determine the motivations behind the bahaviours, as the segmentation is based on complex data constructs and data can vary from person to person (Walker, 2017).

4) Psychographic segmentation: Psychographic segmentation is customer segmentation on the basis of the attitudes, values, interests, opinions and lifestyles of customers (Pickton & Broderick, 2005). There are two principles related to the psychological variable acquisition; personality profiles and lifestyle profile (Larsen, 2010). Psychological profiles, the results of psychographic segmentation, are used as an additional information to define the customer profiles, when geographic and demographics does not provide a sufficient information, and also enhance the understanding of the behavior of customers (Gunter & Furnham, 1992).

In order to define that the segments created will be useful, six criteria have provided by (Kotler & Keller, 2009) as follows.

1) Measurable: The characteristics of each segment should be able to easily identified and measured.

2) Substantial: The target segments are large enough in terms of sales and profitability.

3) Accessible: The target segments should be reachable through promotional and distributional efforts.

4) Differentiable: Each segment is homogenous and unique in characteristics and ways in response to the marketing approaches.

5) Actionable: Marketing approached can be designed and applied to targeted segments.

6) Responsive: Segments should response the marketing strategies applied.

Psychographic segmentation is typically used when a product or service is offered to a heterogeneous market, where customers have very different from each other (Marketing Tutor). As Pickton & Broderick (2005) state that the different variables or factors of this type of segmentation are divided into several different types such as personality traits, attitudes, interests, and activities. Business need to decide which factor will benefit their marketing campaign. Some of the commonly used psychographic segmentation variables are as follows:

1) Lifestyle Segmentation: This is the most common and popular type of psychographic segmentation business used, especially for retail and fashion sectors, because every person has different lifestyles and tastes. For example, many clothing brands present their collections based on which stage of lifestyle they are in, such as kids, college, professional, housewives, and others (Marketing Tutor).

2) Social Class Segmentation: Social class of a customer is determined by the buying power, which affected by background of the individual (Marketing Tutor). So, this is the most straightforward segmentation due to customers will be segmented by classes which they are usually divided into (Anastasia, 2018). For example, most of the premium brands target and maintain only high social class segment as they have buying power to buy their products or services.

3) Personality Segmentation: Personality is a variable that highly dependent on two factors: (1) Lifestyle, and (2) Social class of a customer. It is used to break up

customer into similar group that has the same habits and preferences (Chris, 2016). Here are the psychographic personality types to consider.

- The Belongers: This personality type tends to be conservative and religious. Individuals in this group do not like to be isolated and do not like to experience change, but they are fairy loyal (Chris, 2016).

- The Achievers: The achievers are always busy and aim to moving to a higher social class. Individuals in this group hate everything that is perceived as a waste of their time such as shopping and marketing, but they will make large purchases to symbolize their success (Anastasia, 2018).

- The Emulators: People in this group try to imitate the achievers group, but they do not have work ethic or skill to get there. Therefore, they will make feel like they are already successful and they will make large purchases that they cannot afford (Chris, 2016).

- The Saviors: Saviors do not work for themselves, but for the world. They go out to save the planet, volunteer their time for others, help the homeless. Money is not the most important thing in their life (Chris, 2016).

- The Doomsdayers: This group are the combination of belongers and saviors. They have an own way of life and tend to not rely on anyone else, but they will amazingly loyal to any or person they do trust (Anastasia, 2018).

- The Integrators: This group are the combination of achievers and saviors. People in this group are hardworking and giving back to their community, country, and the world (Anastasia, 2018).

- The Survivalists: People who are struggling to earn a living wage for some reason such as they get money to prove that they can earn it because someone said they could not earn it or they want to break the cycle of poverty. They are often spend money when it is absolutely necessary, because they are terrified of losing everything (Anastasia, 2018).

Psychographic segmentation can help business to understand much better insight of its customers, which can lead to clarification and identification of each segment's needs and motives (Walker, 2017). On the other hand, psychographic data are difficult to obtain than other segmentation base, also collecting and storing personal data of customers without permission may violate privacy (3X 2Y Digital , 2013).

Table 2.2: Advantages and Limitations of the Segmentation Bases

Segmentations	**Advantages**	**Limitations**
Demographic	1. Variables are easier to obtain and measure, compared to others segmentation. 2. It easier to point out what products and services to target to who and what type of advertising and promotion to use. 3. Customer retention and loyalty.	1. Not all customers the same profile have the same needs. 2. May lead to the proliferation of products and services, this could reduce the development of broad-brand equity. 3. Competitors may try to use the same marketing techniques and take away your customers.

(Continued)

Table 2.2 (Continued): Advantages and Limitations of the Segmentation Bases

Segmentations	**Advantages**	**Limitations**
Geographic	1. Can target specific regions based on business location. 2. Can target areas with specific customer information. 3. Can identify the most likely locations customers can be found and how to reach them.	1. It assumes that all customers in a geographic area are similar in needs 2. Usually needs to be used in conjunction with another segmentation base.
Behavioural	1. Can target customers based on historical transactions. 2. Can segment customers based on responsiveness to certain promotion, channel, and product category. 3. Can help to activate a non-user, light user, medium user to heavy user.	1. Customer's behaviour differs depending upon products and services, and their psychographics. It may be difficult to grouping. 2. Data-driven segments are more complex constructs, it is not easy to understand.
Psychographic	1. Gives a much better insight into the customers as a person, which lead to the identification of underlying needs and motives.	1. This approach requires the business have detailed data and research on the customer, which is probably beyond the scope of a small business.

(Continued)

Table 2.2 (Continued): Advantages and Limitations of the Segmentation Bases

Segmentations	**Advantages**	**Limitations**
Psychographic	2. Can help to create more valid and responsive segments and subsequent marketing programs.	2. Collecting and storing personal data of customers may violate privacy.

2.6 Related Literature and Previous Researches

Although market segmentation and customer segmentation have been researched and discussed similarity in many literatures, there are some differences in objectives and data availability for their clustering processes. Market segmentation generally proposed at acquiring new customers process, first stage of CRM, which using customer's demographic data. While customer segmentation can be held at all stages of CRM using both customer's demographic data and transaction data to analyze behaviour of customers (Namvar et al., 2010).

Same as Lefait & Kechadi (2010) research, they stated that there are two types of data available to perform clustering: the *priori* knowledge or demographic data, such as age or gender of customers, and another type is *post hoc* knowledge, which represent the information derived from purchase behaviour. Dennis, Marsland, Cockett, & Hlupic (2003) discussed that the customer's behaviour information are more correlated with the spending than descriptive variables like demographic data or customer's profiles. Moreover, an identified segments based on *post hoc* knowledge might remain originally and essentially the same over time, unlike the demographic characteristics and profile of segment that will be changed over time (Lefait & Kechadi, 2010).

Cheng & Chen (2009) mentioned in their paper that, customer value analysis is a method for discovery customer's characteristic to make companies know about who are the target customers which contribution is outstanding. In many researches, there are two customer value analysis methods were mentioned: Life Time Value (LTV) and Recency, Frequency, and Monetary (RFM).

LTV model considering three main things: past profit contribution, potential benefit, and customer defection and future purchase propability. For example, a case study of Hwang, Jung, & Suh (2004) which using LTV model to analyze the customers of wireless telecommunication company to find out the customer value and offered strategies based on customer segments, and another example is a customer segmentation that considering customer defection and cross-selling opportunity, which determined by LTV model (Kim, Jung, Suh, & Hwang, 2006).

Kaymak (2001) deemed that the RFM model is one of the well-known customer value analysis methods because its help to exact critical characteristic of customers by using fewer criterions (three variables: R, F, and M), that reduce the complexity of customer value analysis. Schijns & Schröder (1996) also pointed out that the RFM model is a method that help to measure the strength of customer relationship in consuming behaviour view. For example, Mccarty & Hastak (2007) used the combination of RFM model and others model such as CHAID and logistic regression as analytical methods for direct marketing segmentation with two different datasets. Another example is a Chen, Sain, & Guo (2012) research that proposed a case study of using RFM model with data mining technique such as K-means clustering and decision tree induction to segment retail's customers.

Besides, many researchers proposed a combination of input variables for customer segmentation. For example, Chan (2008) introduced a combination of customer targeting and customer segmentation for campaign strategies, which identified customer behaviour with RFM model and evaluated segmented customer using LTV model.

Table 2.3: Summarization of Input Variables for Customer Segmentation

Input Variables Used	**References**
Demographic	Lefait & Kechadi (2010)
LTV	Kim et al. (2006)
RFM	Kaymak (2001) and Cheng & Chen (2009)
Demographic + LTV	Hwang et al. (2004)
Demographic + RFM	Mccarty & Hastak (2007), Rajagopal (2011), and Chen et al. (2012)
LTV + RFM	Chan (2008)
Demographic + LTV + RFM	Namvar et al. (2010)

Many related works mentioned about the data mining models and techniques used in customer segmentation processes, which generally include association, classification, clustering, forecasting, regression, sequence discovery, and visualization. According to Ngai et al. (2009) research, each of CRM stage and customer segmentation can be supported by different data mining models as follows:

- Association: Association tries to discovering and establishing relationships between items which exist together in dataset (Ahmed, 2004). Market basket analysis and cross selling are an example of association model, which based on statistics and priori algorithms.

- Classification: Classification is one of the most common models in data mining (Ahmed, 2004). There are many methods regarding to classification model, such as neural networks, decision trees, and if-then-else rules. The objective of classification is to predict customer behaviour in the future, which resulted from classifying process that assigns each item into predefined categories or classes according to certain criteria.

- Clustering: Clustering is the process of grouping various objects into groups of similar objects (Ahmed, 2004; Han et al., 2012). The difference between clustering and classification is in clustering there are no predefined clusters (categories or classes), resulted clusters will be created at the time algorithm starts. Example of clustering tools are neural networks, discrimination analysis, K-means clustering, and hierarchical clustering.

- Forecasting: Forecasting estimates the future values based on historical data relationships and patterns (Ahmed, 2004). Neural networks and survival analysis are a typical example of forecasting model.

- Regression: Regression used to predict the future values, given a particular dataset (Ngai et al., 2009). For CRM, regression model can be help to predict profit and sales based on historical transactions. Common tools for regression are linear regression and logistic regression.

- Sequence discovery: Sequence discovery or Sequential pattern mining is technique that discovers statistically relevant patterns between each item in dataset over time. Statistics and set theory are used as a sequence discovery tools.
- Visualization: Visualization is a presentation of data, which provides clearer and easier understanding of the discovered relationships or patterns of data (Ngai et al., 2009). Example tools for data visualization is 3D graphs, interactive graphs and dashboard.

As Ngai et al. (2009) stated in their research, a combination of data mining models is usually required to support the objectives of CRM. For instance, in the case of up selling and cross selling program need to be implemented, customers need to be segmented into clusters before an association models are applied. In such case that business needs to retain a group of loyalty and potential customers, the customer segmentation is required to know which is a target group to be retained, before classification models are applied.

Table 2.4: Summarization of Data Mining Techniques for Customer Segmentation

Data Mining Models	**Data Mining Techniques**	**References**
Classification	Decision tree	Kim et al. (2006)
	Self-organizing Map	Ha, Bae, & Park (2002)
Clustering	K-means	Dennis et al. (2003), Tsai & Chiu (2004), Cheng & Chen (2009), Namvar et al. (2010),

(Continued)

Table 2.4 (Continued): Summarization of Data Mining Techniques for Customer Segmentation

Data Mining Models	**Data Mining Techniques**	**References**
Clustering	K-means	Lefait & Kechadi (2010), Rajagopal (2011), Chen et al. (2012)
	Pattern based cluster	Yang & Padmanabhan (2005)
Regression	Logistic regression	Hwang et al. (2004)

2.7 Conceptual Framework

According to the objectives of this research are to construct a customer segmentation model and perform a segmentation for customers of IT solution and service provider business in Thailand, this conceptual framework was presented.

In order to segment the customers and develop marketing strategies, a combination of demographic segmentation and behavioural segmentation model was constructed based on data availability from a case company.

Firstly, the raw data of purchasing transactions from 2016 to 2018 are elicited from the company in Microsoft Excel format. In the data pre-processing process, customer data and transaction data need to separate for further analysis. The appropriate variables such as *customer name*, *business size*, *industry*, and *company type* are selected to be used as customer demographics. For the purchasing transactions data, a transformation is needed according to the RFM analysis model, which transaction data need to be transformed into three essential variables, which are *Recency*, *Frequency*, and *Monetary*. In the processing step, RFM data are used in the first phase of segmentation as the input of K-means clustering, which *k* value will be chosen according to the results of cluster distance performance tested by average within cluster distance and Davies-Bouldin index. In the second phase, each cluster resulted from a previous phase is again clustered into new clusters using customer demographics data, which *k* value will be chosen according to the results of cluster distance performance tested by average within cluster distance and Davies-Bouldin index. Finally, each clusters need to be interpreted and identified the main characteristics, and then the recommendations are provided to the business in the marketing perspective. In the last step, a face-to-face semi-structured interview with a company representative was held to discuss about research findings and recommendation for further analysis. (see Figure 2.7: Conceptual Framework)

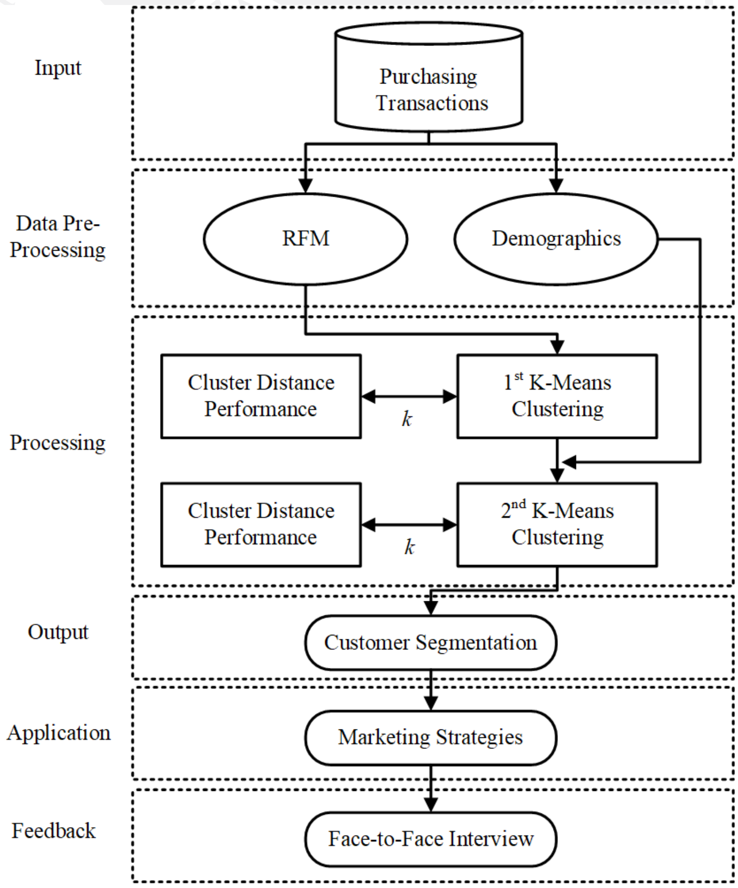


Figure 2.7: Conceptual Framework

CHAPTER 3

METHODOLOGY

3.1 Research Design

In order to satisfy the objectives of this research, the quantitative case study research methodology was held. According to Sturman (1977), “case study is a general term for the exploration of an individual, group or phenomenon”. Hence, a case study provides a comprehensive description of an individual case and its analysis. Case study typically combine data collecting methods such as observations, interviews, and questionnaires (Eisenhardt, 1989). The collected data can be qualitative, quantitative, or both. Case study can be used to achieve various proposes such as provide description, test theory, or generate theory (Eisenhardt, 1989).
Yin (2014) has mentioned in his book that, there are four types of design that researchers can make use of, which include single holistic design, single embedded design, multiple holistic design, and multiple embedded design. Holistic designs require one unit of analysis, whereas embedded designs require multiple units of analysis. Yin (2014) also shows that, there are five components of a case study design: (1) a study’s questions, (2) study’s propositions, (3) unit of analysis, (4) the logic linking the data to propositions, and (5) criteria for interpreting the findings. The researchers should review the related study and literature and include theoretical propositions regarding to case before starting to conduct any data collection or analysis and also measure the quality of the design against four criteria, which include construct validity, internal validity, external (Yin, 2014). Same as the eight processes

of building theory from case study research proposed by (Eisenhardt, 1989), which summarized as Table 3.1.

Table 3.1: Process of Building Theory from Case Study Research

Step	Activity
Getting Started	Situate problem or phenomenon: level, scope, focus Define research questions
Selecting Cases	Specify population using theoretical or analytical
Crafting Instruments and Protocols	Multiple data collection methods Qualitative and quantitative data combined
Entering the Field	Overlap data collection and analysis to sharpen concepts Flexible and opportunistic data collection methods
Analyzing Data	Within-case analysis Cross-case pattern search
Shaping Hypotheses	Iterative tabulation of evidence for each construct Replication, not sampling, logic across cases Search evidence for "why" behind relationships
Enfolding Literature	Compare with similar literature Compare with conflicting literature
Reaching Closure	Theoretical saturation on research question

For the advantage of the case study, it allow a researcher to identify and measure the indicators that best represent the theoretical concepts the researcher intends to measure, called the conceptual validity (George & Bennett, 2005). In this research, the IT solution and service provider business in Thailand was chosen as a case to implement the customer segmentation based on their customer demographics and their customer's historical purchasing transactions data from year 2016 to 2018.

3.2 Population and Sample Selection

As this research is proposing a case study of the IT solution and service provider business in Thailand. Hence, the sample datasets are customer demographics and historical purchasing transactions data from 2016 to 2018, collected from one of the leading Total Enterprise IT Solution and Service Provider company in Thailand, which serves IT solutions and services over 1,000 customers in various industries with world class standard quality ISO 20000 (IT Service Management) and ISO 27001 (Information Security Management) over 33 years.

3.3 Research Instrument

In this research, the Microsoft Excel is used to manipulate the raw dataset in term of normalizing, calculating, and transforming raw data into an appropriate variable and format in a data pre-processing stage. For the processing stage, data mining technique, K-means clustering, is performed using a data mining tools named RapidMiner Studio.

RapidMiner Studio is a software platform which has been developed by a RapidMiner Company. It provides a centralized solution for machine learning, predictive analysis, business analytics, and text mining (Dwivedi, Kasliwal, & Soni, 2016). RapidMiner Studio has been placed as a leader in Gartner Magic Quadrant for data science and machine learning platform, that represents the best balance between data science sophistication and ease of use compared with others (Idoine, Krensky, Brethenoux, & Linden, 2019

3.4 Data Collection Procedure

To address the research questions, a mixture of quantitative data collecting was used. The quantitative method is represented by collecting the customer demographics and historical purchasing transactions data from IT solution and service provider company, that researcher sent out an email with attached letter of recommendation for data collection and research work along with the non-disclosure agreement (NDA) to the company representative, and face-to-face semi-structured interview with a company representative was held to discuss about research findings and recommendation for further analysis.

3.5 Summary of Dataset

The customer transaction dataset held by one division of IT solution and service provider company from year 2016 to 2018. There are 1,971 record rows with four variables. (see Table 3.2: Variables in the Customer Transaction Dataset). Each record represents a transaction made by each customer for particular products. (see Table 3.3: Example of Customer Transaction Dataset).

Table 3.2: Variables in the Customer Transaction Dataset

Variable Name	**Data Type**	**Description**
TransactionDate	Date	The month and year when each transaction was generated in MM/YYYY format

(Continued)

Table 3.3 (Continued): Variables in the Customer Transaction Dataset

Variable Name	Data Type	Description
CustomerName	Nominal	Customer name in English
ProductName	Nominal	Product name in English
Price	Numeric	Product and service price in Thai Baht (THB)

Table 3.4: Example of Customer Transaction Dataset

TransactionDate	CustomerName	ProductName	Price
01/2016	Customer Name 1	Product Name 1	200000.00
01/2016	Customer Name 1	Product Name 20	25000.00
03/2017	Customer Name 2	Product Name 15	3000000.00
…	…	…	…
12/2018	Customer Name 3	Product Name 3	75600.00

According to conceptual framework (see Figure 2.7: Conceptual Framework), the customer demographic information and RFM information are needed for clustering in processing stage. Thus, the customer information and transaction need to be separated into two tables for further processing.

First, demographic table, this will be used as customer demographic information for clustering in processing stage. There are 242 customer name listed in a table. Regarding to the non-disclosure agreement (NDA) signed with a company, the identifiable of their customer, customer name variable, need to be replaced with a *CustomerID* variable. Moreover, appropriate variables are added to transform the raw customer data into customer demographic information. In this case, *BusinessSize* variable, *Industry* variable, and *Company*Type variable are added (see Table 3.4:

Variables in the Demographic Table). For an example data in demographic table (see Table 3.5: Example data in Demographic Table).

Table 3.5: Variables in the Demographic Table

Variable Name	Data Type	Description
CustomerID	Numeric	Customer code; a 4 digit number uniquely assigned for each customer name
BusinessSize	Nominal	Size of company according to the number of employees; S (2-50), M (51-500), L (501-5000), and XL (5001 and above)
Industry	Nominal	Industries name in English
CompanyType	Nominal	Types of company; Company Limited, Public Company Limited, Educational Institution, Foundation, Government Agency, Nonprofit

Table 3.6: Example data in Demographic Table

CustomerID	BusinessSize	Industry	CompanyType
0001	XL	Banking	Public Company Limited
0002	M	Food and Beverage	Company Limited
0003	L	Hospitals	Nonprofit
…	…	…	…
0242	S	Government	Government Agency

For the RFM table, raw transaction is need to be transformed according to the RFM analysis model, which three essential variables, *Recency*, *Frequency*, and *Monetary* are needed. (see Table 3.6: Variables in the RFM Table).

Table 3.7: Variables in the RFM Table

Variable Name	Data Type	Description
CustomerID	Numeric	Customer code; a 4 digit number uniquely assigned for each customer name
Recency	Numeric	Latest purchase month for each customer; 01-36 (01/2016-12/2018)
Frequency	Numeric	Frequency of purchase for each customer
Monetary	Numeric	Total amount spent for each customer

Recency variables can be discovered by picking up the latest month that each customer made a transaction from *TransactionDate* variable. However, for ease of calculation and limitation of K-means clustering operation in RapidMiner Studio, which is only supported the data in numeric type. Thus, *Recency* variable needs to be converted into numeric data type. (see Table 3.7: Conversion of Transaction Date Variable).

Table 3.8: Conversion of Transaction Date Variable

Transaction Date	Month [01-36]
01/2016	01
02/2016	02
03/2016	03
…	…
10/2018	34
11/2018	35
12/2018	36

Frequency variable can be calculated by counting numbers of transaction made by particular customer. *Monetary* variable can be calculated by summing all data values from *Price* variable which associated with particular customer.

Finally, RFM table with 242 records regarding to number of customers was created. (see Table 3.8: Example data in RFM Table).

Table 3.9: Example data in RFM Table

CustomerID	**Recency**	**Frequency**	**Monetary**
0001	31	3	74,025.00
0002	24	2	599,000.00
…	…	…	…
0242	35	6	414,460.00

The summary of the demographic dataset and RFM dataset are given in Table 3.9 and Table 3.10.

Table 3.10: Summary of the Demographic Dataset (242 Instances)

Variables	**Least**	**Most**
Dataset (242 instances)		
BusinessSize	XL (23)	M (112)
Industry	27 Industries (1)	Financial Services (31)
CompanyType	Foundation (1)	Company Limited (143)

Table 3.11: Summary of the RFM Dataset (242 Instances)

Variables	Minimum	Median	Maximum
Dataset (242 instances)			
Recency	1	27.90	36
Frequency	1	8.15	72
Monetary	760.00	3,499,257.51	45,426,800.00
FirstPurchase	1	9.86	36

CHAPTER 4

FINDINGS

4.1 Case Description

X-company, founded in 1986, located in Bangkok, Thailand. X-Company is leading IT solutions and services with the strong intention to position itself as a Total Enterprise Solution and Service Provider, which offers consulting, support, and technical services to enhance and accelerate digital business for customers. Nowadays, its employees are about 500 persons. Based on the partnership with large global companies and the world-class standard quality guaranteed, X-company has been success to serves IT solutions and services over 1,000 customers in various industries. Due to the rapid growing of IT, X-company needs to increase the revenue of sales and enhance a seamless relationship with customers.

4.2 Data Analysis

For the data analysis in processing stage regarding to conceptual framework (see Figure 2.7: Conceptual Framework), the process diagram has been set up in RapidMiner Studio for RFM clustering. There are three main processes: (1) Read data from file, (2) Select attributes, and (3) K-means clustering, as depicted in Figure 4.1.

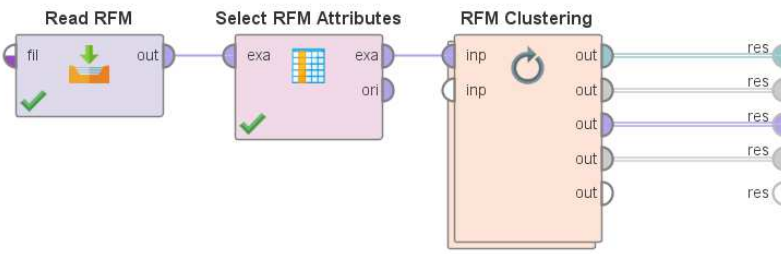


Figure 4.1: Process Diagram of RFM Clustering (Main Process)

In the K-means clustering process, there are 3 sub processes; (1) K-means clustering, (2) Cluster distance performance, and (3) Cluster model visualizer, as illustrated in Figure 4.2. For the clustering distance performance, there are 2 methods supported, (1) Average within cluster distance, and (2) Davies-Bouldin index, which every clusters will be tested with both methods for the best *k* value selection.

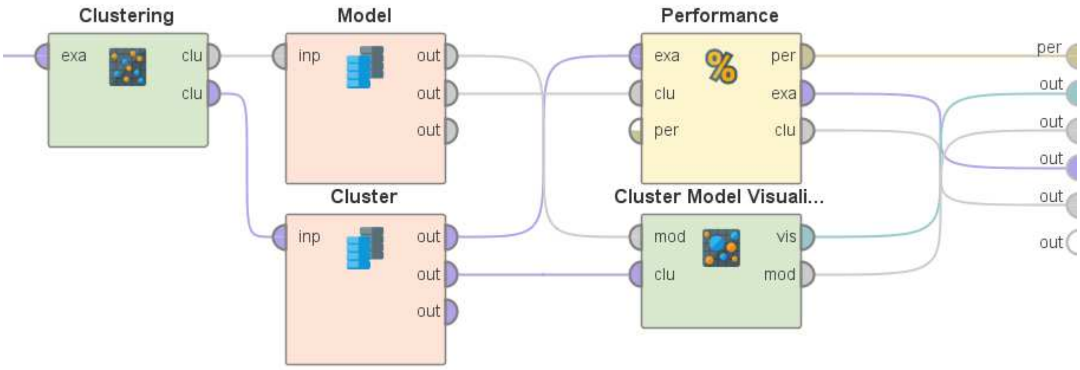


Figure 4.2: Process Diagram of RFM Clustering (Sub Process)

Part of RFM input is shown in Figure 4.3, and the histograms of the variables *Recency*, *Frequency*, and *Monetary* are illustrated in Figure 4.4. It is evident that the majority of dataset has high recency, low frequency, and low monetary.

	CustomerID integer CustomerID	Recency integer	Frequency integer	Monetary real
1	1	31	3	74025.000
2	2	24	2	599000.000
3	3	27	1	76500.000
4	4	35	2	1930000.000
5	5	27	7	5469100.020
6	6	27	3	821250.000
7	7	33	18	10150450.000
8	8	36	15	6603150.000
9	9	33	1	348500.000
10	10	31	3	1140000.000
11	11	23	1	291840.000
12	12	36	6	216848.000

no problems.

Previous Finish Cancel

Figure 4.3: Sample of RFM Input

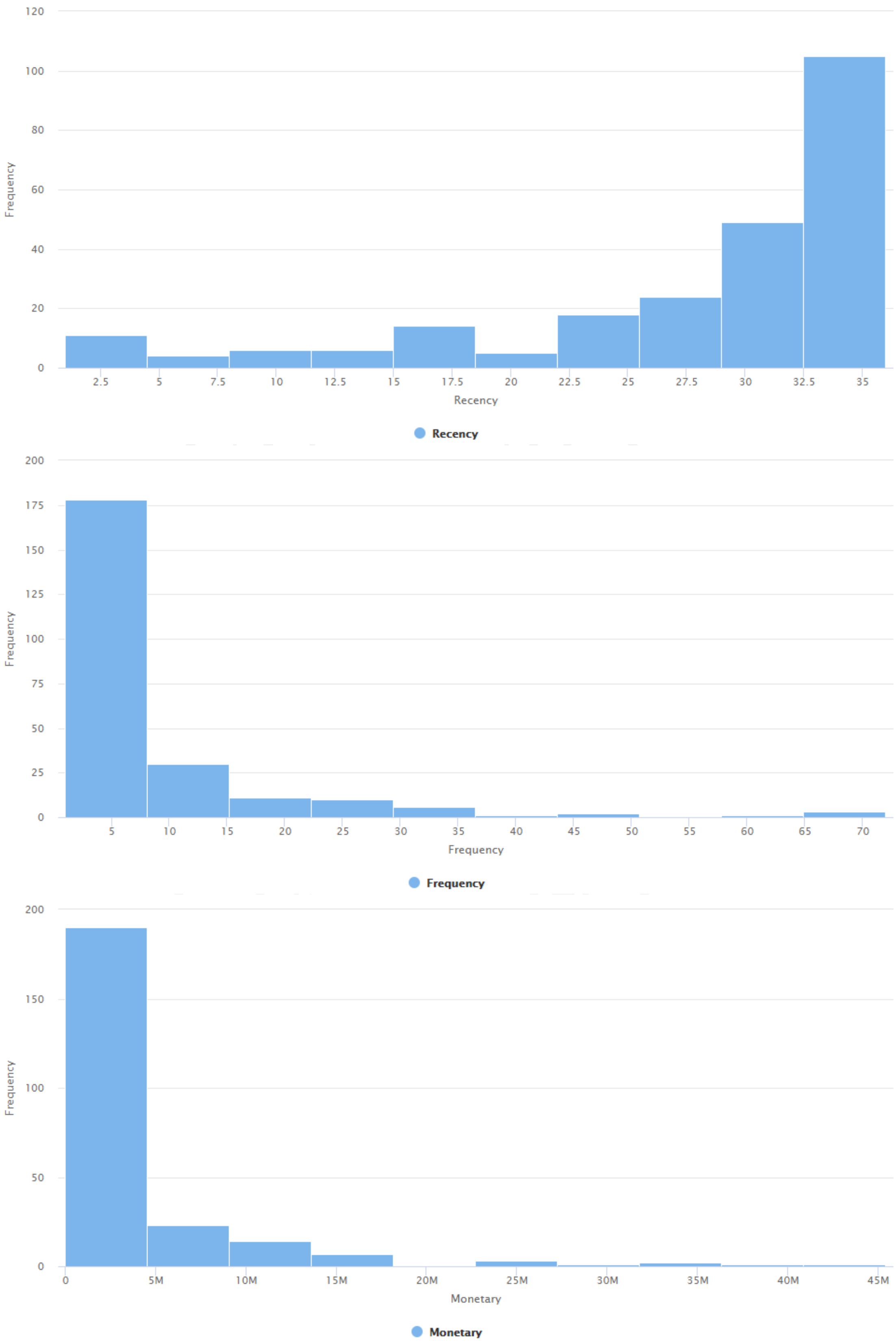


Figure 4.4: Distribution of the Recency, Frequency, and Monetary

According to K-means procedure proposed by (Han et al., 2012), the first step is choosing *k* objects from dataset as the initial duster centers. The appropriate *k* is selected by evaluating the performance of the clustering based on the cluster centroids. There are two performance measures are supported in RapidMiner Studio: Average within cluster distance and Davies-Bouldin index. Thus, range of *k* from 1 to 10 was selected for evaluating and the best *k* performance is defined as a number of cluster.

Results of evaluating with Average within cluster distance and Davies-Bouldin index shown as Figure 4.5 and Figure 4.6.

RFM Clustering (10 rows, 3 columns)

iteration ↑	Clustering.k	Avg. within centroid distance
1	1	43625951384339.910
2	2	13949951842165.830
3	3	4897018312589.233
4	4	3134368065245.969
5	5	1882925431533.773
6	6	1240226145520.309
7	7	893482289040.991
8	8	576125151971.187
9	9	433624338067.404
10	10	328101402832.790

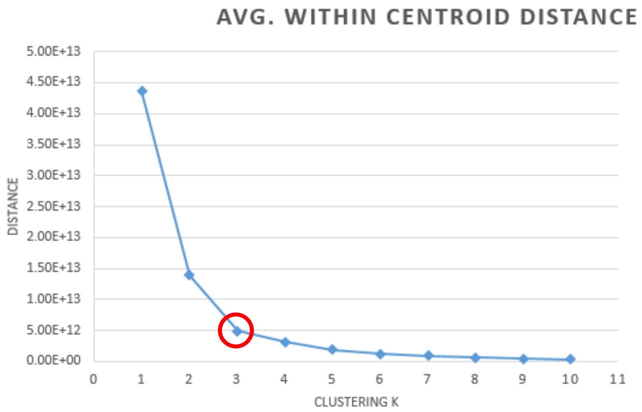


Figure 4.5: Result of Average within Cluster Distance for RFM Input

RFM Clustering (10 rows, 3 columns)

iteration ↑	Clustering.k	Davies Bouldin
1	1	∞
2	2	-0.118
3	3	-0.102
4	4	-0.116
5	5	-0.122
6	6	-0.115
7	7	-0.093
8	8	-0.092
9	9	-0.098
10	10	-0.099

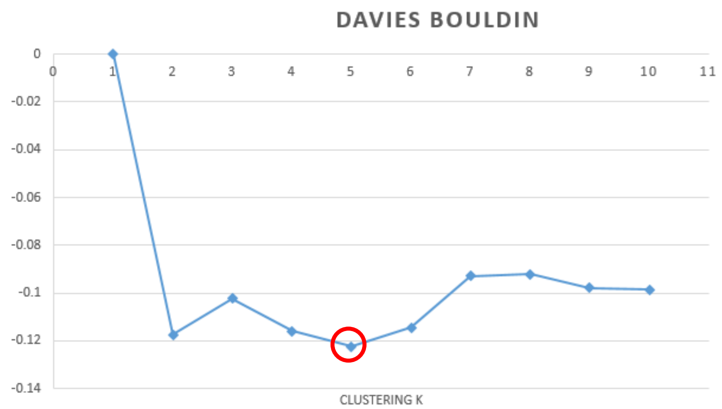


Figure 4.6: Result of Davies-Bouldin Index for RFM Input

Regarding to average within cluster distance graph, elbow method was chosen to determine the optimal number of cluster, which elbow point is located at k = 3. On the other hand, the optimal number cluster in Davies-Bouldin method is k = 5, as it is a minimize value. In comparison, k = 3 on average within cluster distance graph is more obviously than the Davies-Bouldin value. Therefore, the best k value is 3.

The clustering and segment results with 3 clusters are shown in Table 4.1 and Table 4.2, and the distribution of the instances within each cluster is detailed in Figure 4.7 to Figure 4.11.

Table 4.1: Instance in Each Cluster for RFM Input

Cluster	Frequency of Cluster	Percentage
Cluster_0	203	83.88
Cluster_1	31	12.81
Cluster_2	8	3.31

Table 4.2: Statistics of Each Cluster for RFM Input

Clusters/Variables	Minimum	Median	Maximum
Cluster_0			
Recency	1	26.67	36
Frequency	1	4.83	34
Monetary	760	1,223,702.85	5,767,000
FirstPurchase	1	10.82	36

(Continued)

Table 4.2 (Continued): Statistics of Each Cluster for RFM Input

Clusters/Variables	Minimum	Median	Maximum
Cluster_1			
Recency	26	34.42	36
Frequency	3	24.61	72
Monetary	6,471,250	11,085,242.54	17,970,009.25
FirstPurchase	1	5.23	18
Cluster_2			
Recency	27	33.75	36
Frequency	2	29	64
Monetary	23,396,800	32,067,563	45,426,800
FirstPurchase	1	4.5	16

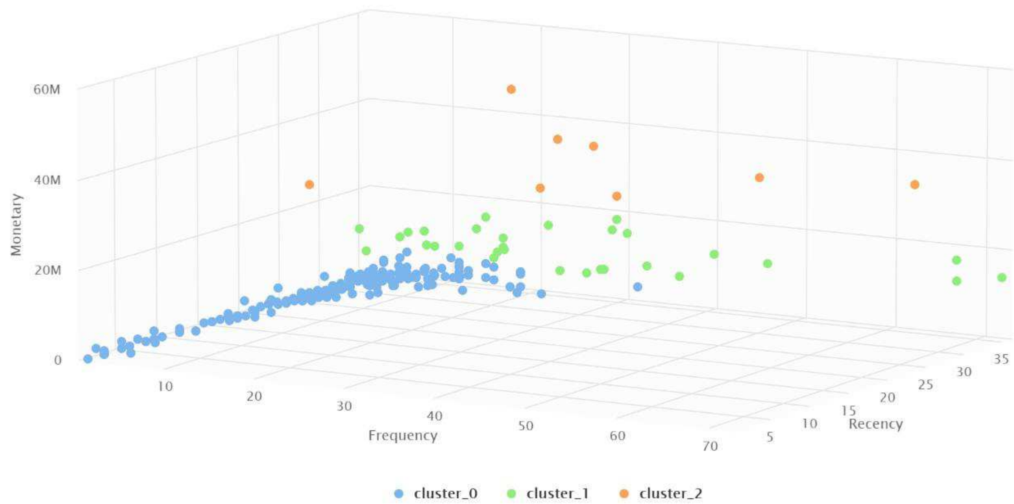


Figure 4.7: Distribution of All Clusters for RFM Input

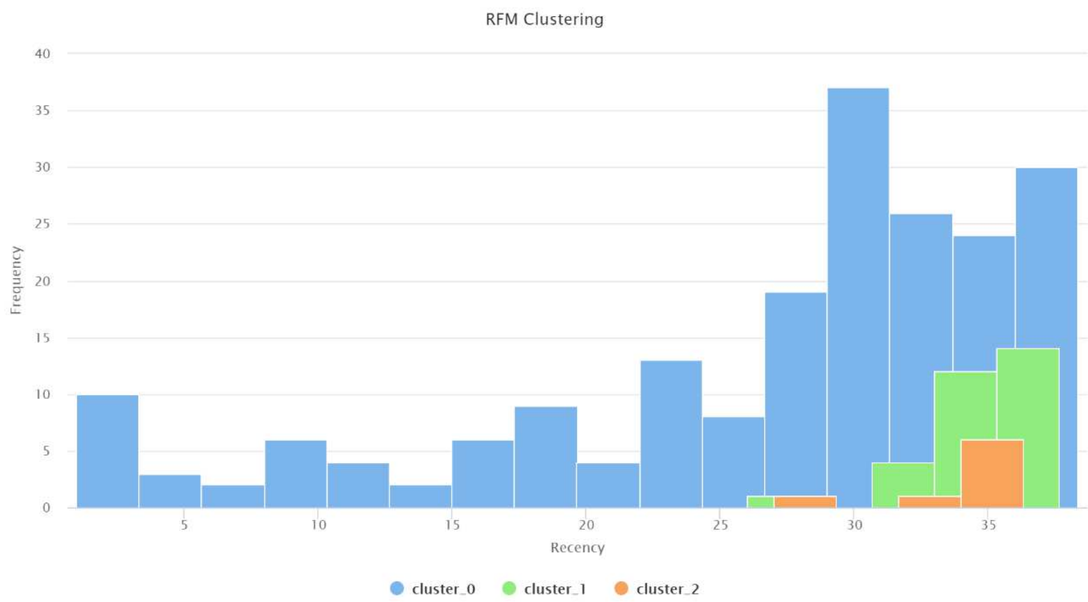


Figure 4.8: Distribution of Recency by Cluster

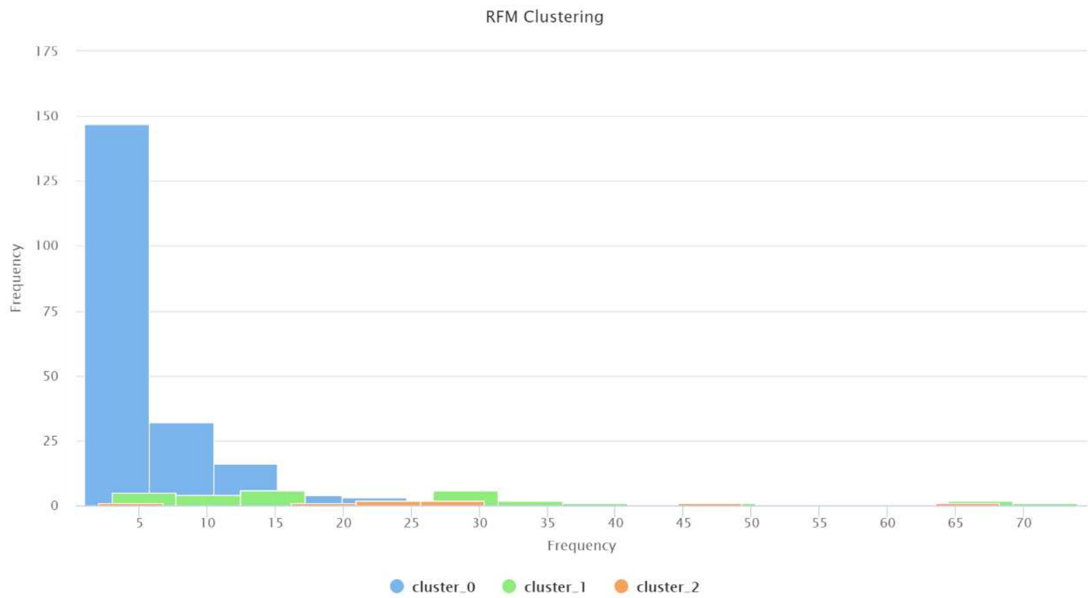


Figure 4.9: Distribution of Frequency by Cluster

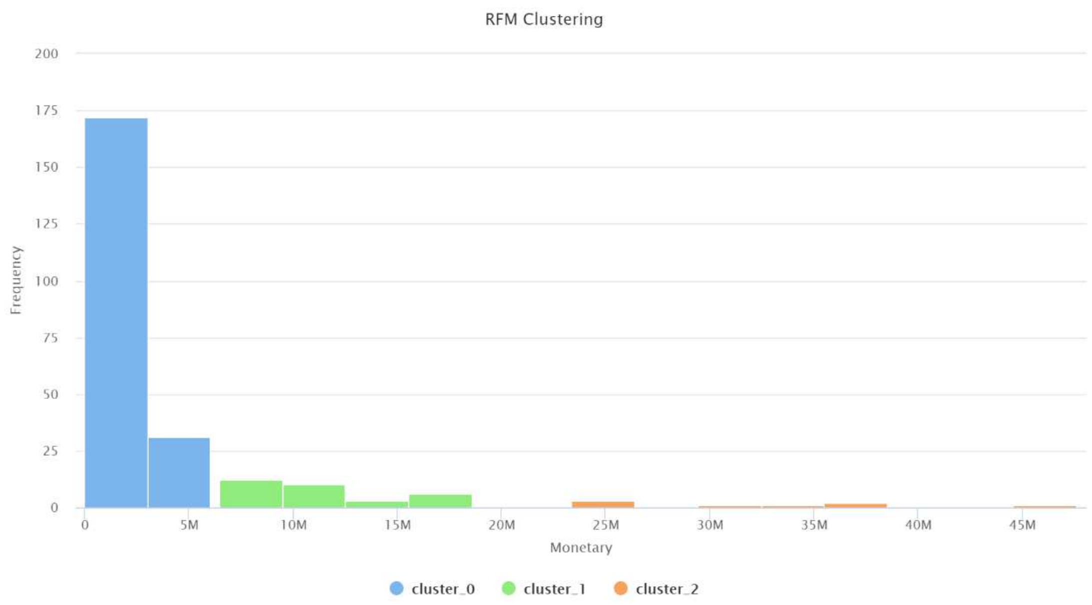


Figure 4.10: Distribution of Monetary by Cluster

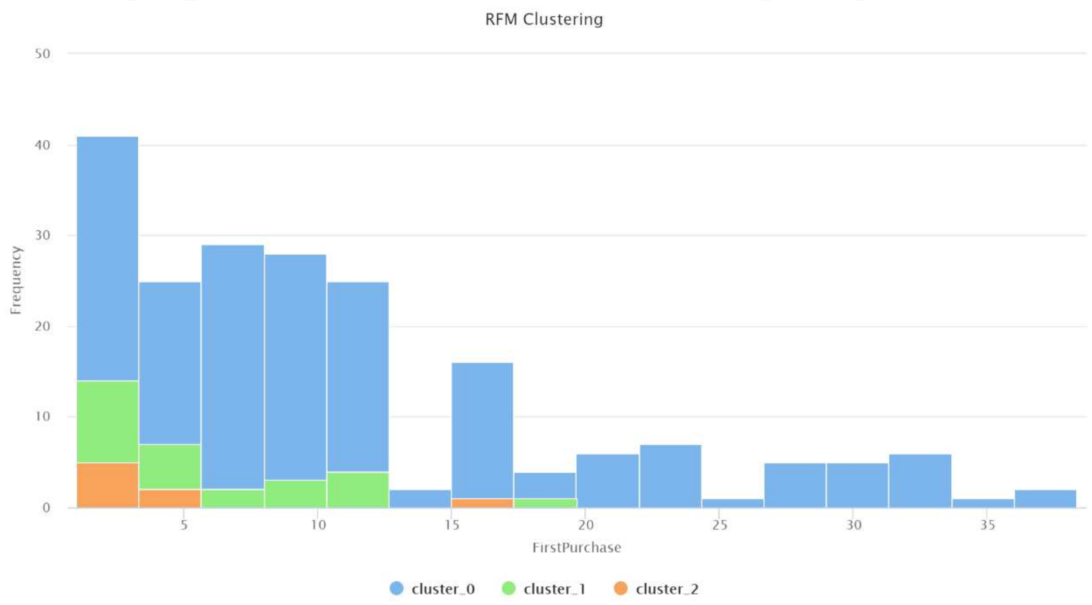


Figure 4.11: Distribution of First Purchase by Cluster

Cluster_0 relates to some 203 customers, composed of 83.8 percent of whole population. This group seem to be the least profitable group as it has a very low value of monetary and frequency. Contrasted with the customers in cluster_0, the 8 customers in cluster_2 started purchasing at the second quarter of 2016 and continued to the end of 2018 with an extremely high value of monetary and frequency. This group of customers can be categorized as very high recency, very high frequency, and very high monetary with a high spending per customer. In point of fact, those 8 customers contributed 30.3 percent of the total sales, while they are only composed of 3.3 percent of the whole population. This group seem to be the most profitable group. Cluster_1 is a second high profit group, which contain 31 customers with a very high value of recency, frequency, and monetary, although lower than the cluster_2.

For the second K-means clustering, demographic clustering, the process diagram again has been set up in RapidMiner Studio. There are 4 main processes: (1) Read data from file, (2) Select attributes, (3) Transform nominal data type into numeric type, and (4) K-means clustering, as illustrated in Figure 4.12 and Figure 4.13. Regarding to results from RFM clustering (first clustering), the input datasets are separated into 3 datasets based on RFM cluster; (1) cluster_0, (2) cluster_1, and (3) cluster_2.

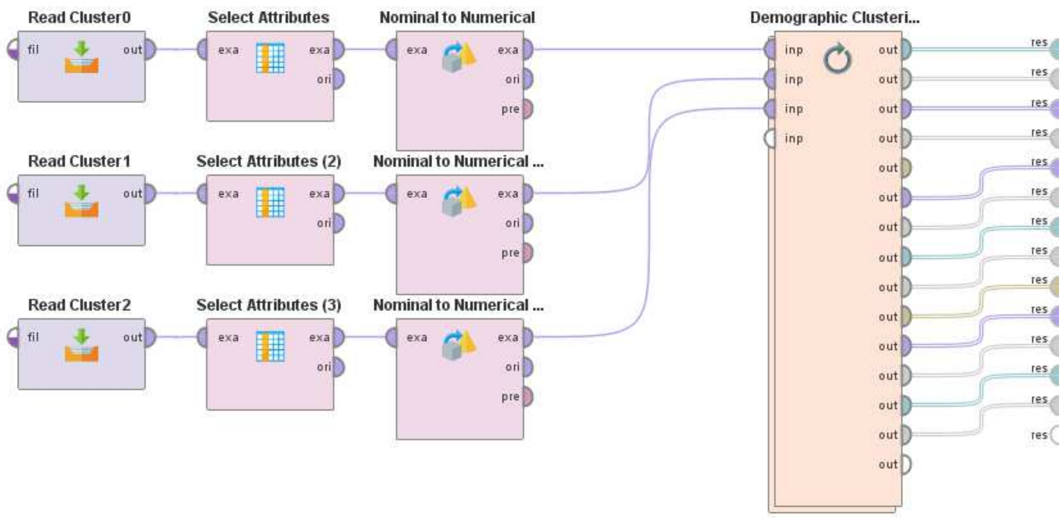


Figure 4.12: Process Diagram of Demographic Clustering (Main Process)

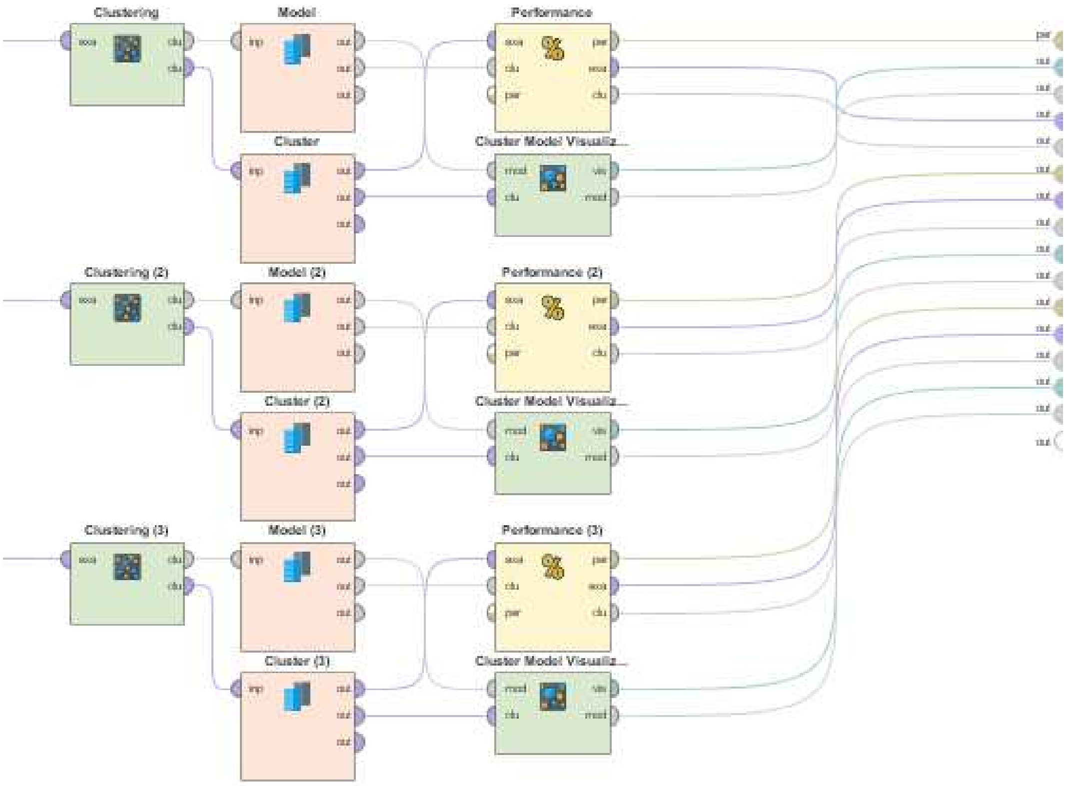
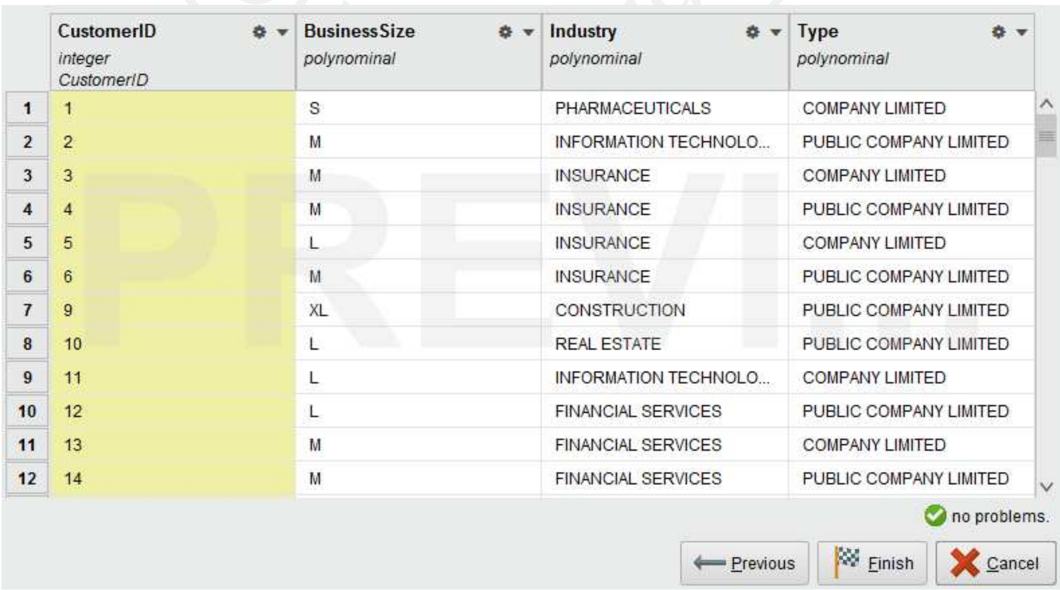


Figure 4.13: Process Diagram of Demographic Clustering (Sub Process)

Part of cluster_0 demographic input is shown in Figure 4.14, and the histograms of the variables *BusinessSize*, *Industry*, and *CompanyType* are illustrated in Figure 4.15.



	CustomerID *integer* *CustomerID*	BusinessSize *polynominal*	Industry *polynominal*	Type *polynominal*
1	1	S	PHARMACEUTICALS	COMPANY LIMITED
2	2	M	INFORMATION TECHNOLO...	PUBLIC COMPANY LIMITED
3	3	M	INSURANCE	COMPANY LIMITED
4	4	M	INSURANCE	PUBLIC COMPANY LIMITED
5	5	L	INSURANCE	COMPANY LIMITED
6	6	M	INSURANCE	PUBLIC COMPANY LIMITED
7	9	XL	CONSTRUCTION	PUBLIC COMPANY LIMITED
8	10	L	REAL ESTATE	PUBLIC COMPANY LIMITED
9	11	L	INFORMATION TECHNOLO...	COMPANY LIMITED
10	12	L	FINANCIAL SERVICES	PUBLIC COMPANY LIMITED
11	13	M	FINANCIAL SERVICES	COMPANY LIMITED
12	14	M	FINANCIAL SERVICES	PUBLIC COMPANY LIMITED

no problems.

Previous Finish Cancel

Figure 4.14: Sample of Cluster_0 Demographic Input

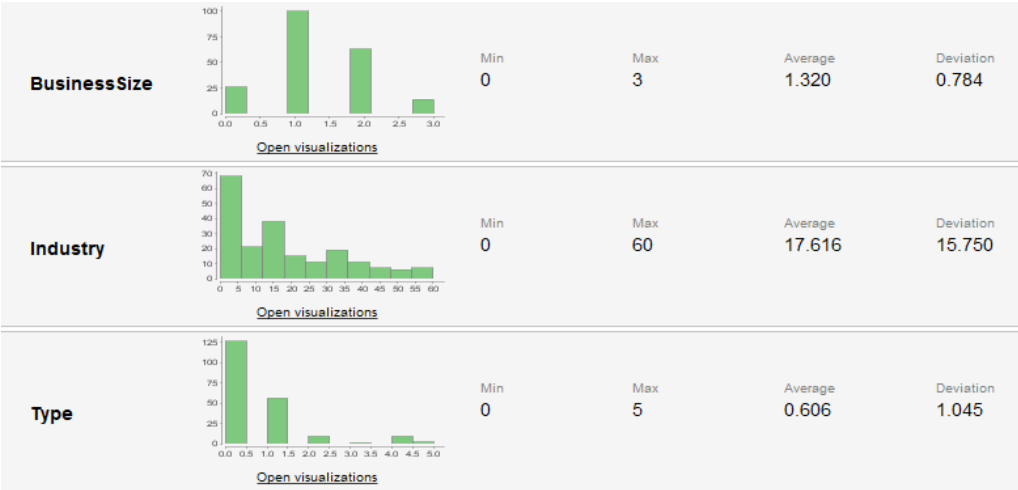


Figure 4.15: Distribution of All Variables for Cluster_0 Input

Results of evaluating with Average within cluster distance and Davies-Bouldin index for cluster_0 demographic input shown as Figure 4.16 and Figure 4.17.

Demographic Clustering (10 rows, 3 columns)

itera... ↑	Clustering.k	Avg. within centroid distance
1	1	82.848
2	2	20.929
3	3	10.564
4	4	4.859
5	5	3.305
6	6	2.475
7	7	1.974
8	8	1.618
9	9	1.302
10	10	1.118

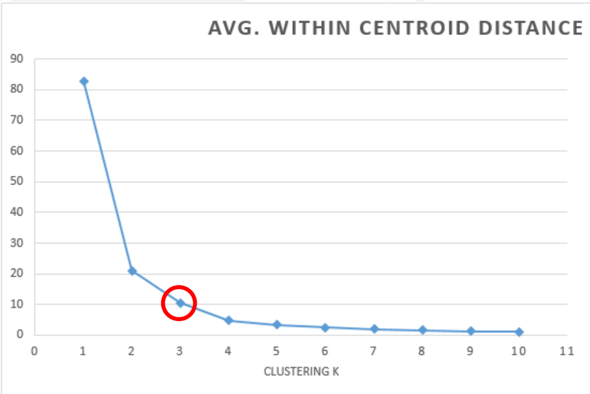


Figure 4.16: Result of Average within Cluster Distance for Cluster_0 Input

Demographic Clustering (10 rows, 3 columns)

itera... ↑	Clustering.k	Davies Bouldin
1	1	∞
2	2	-0.164
3	3	-0.170
4	4	-0.159
5	5	-0.175
6	6	-0.181
7	7	-0.193
8	8	-0.206
9	9	-0.213
10	10	-0.214

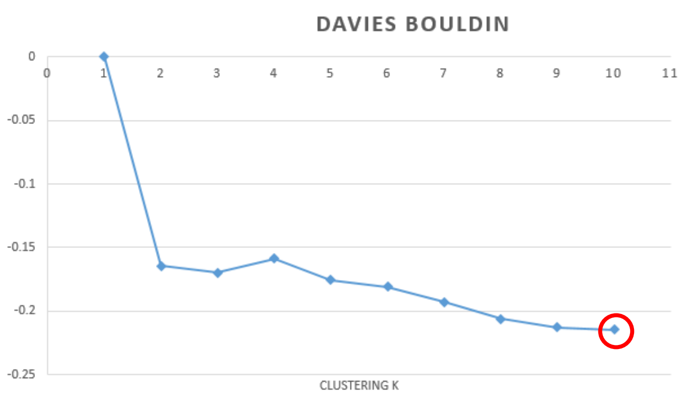


Figure 4.17: Result of Davies-Bouldin Index for Cluster_0 Input

Regarding to average within cluster distance graph, elbow method was chosen to determine the optimal number of cluster, which elbow point is located at $k = 3$. On the other hand, the optimal number cluster in Davies-Bouldin method is $k = 10$, as it is a minimize value. In comparison, $k = 3$ on average within cluster distance graph is more obviously than the Davies-Bouldin value. Therefore, the best k value is 3.

The clustering and segment results with 3 clusters are shown in Table 4.3, and the distribution of the instances within each cluster is detailed in Figure 4.18.

Table 4.3: Instance in Each Cluster for Cluster_0 Input

Cluster	Frequency of Cluster	Percentage
Cluster_0 (Cluster_0)	68	33.50
Cluster_1 (Cluster_1)	50	24.63
Cluster_2 (Cluster_2)	85	41.87

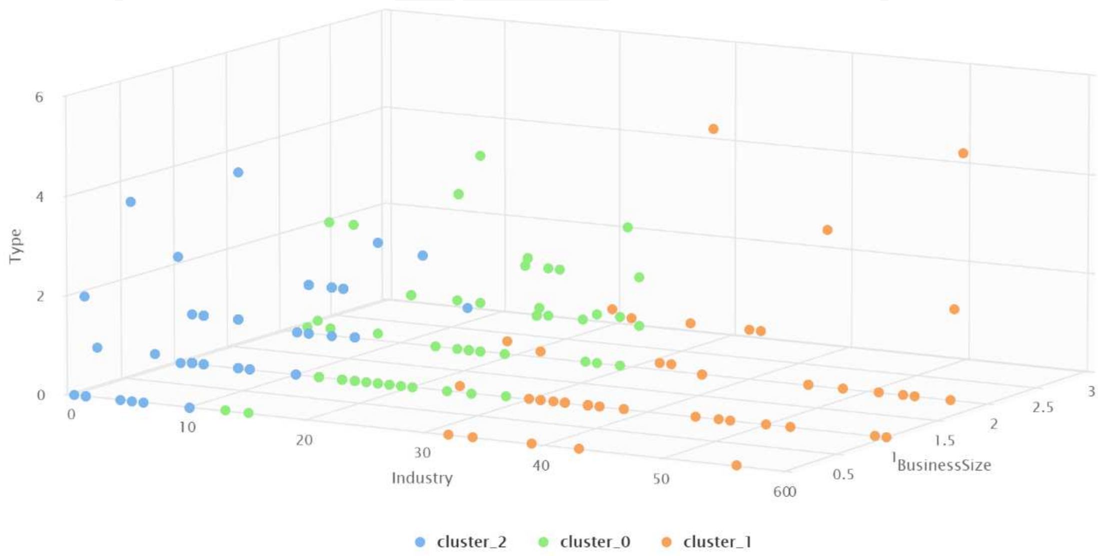
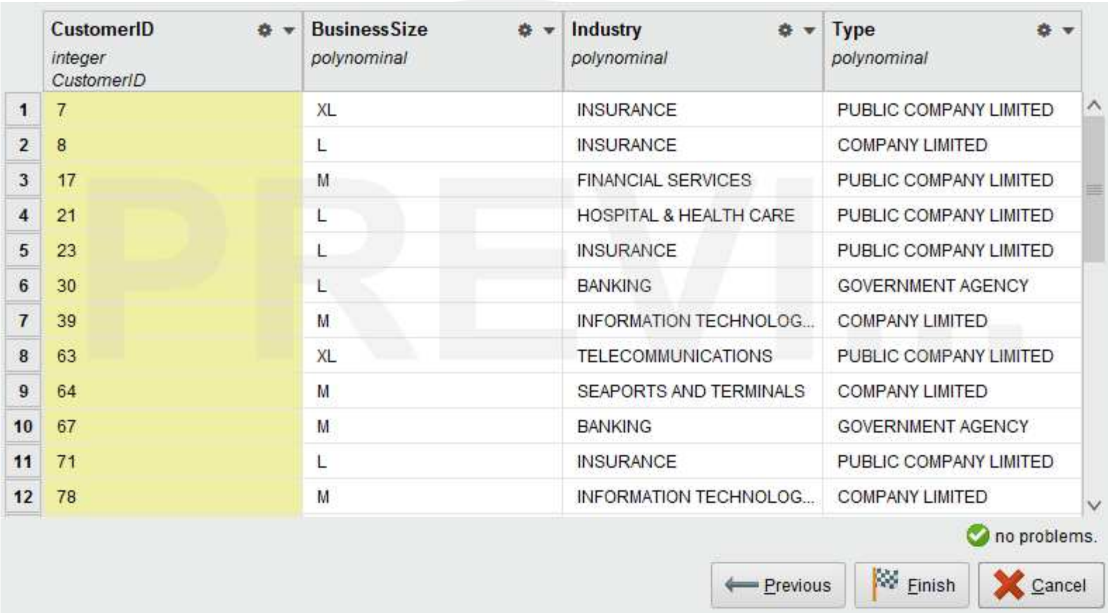


Figure 4.18: Distribution of All Clusters for Cluster_0 Input

Most of the customers in each cluster have some common characteristics, which is the majority of them are company limited type with medium business size.

However, the key difference between three clusters is an industry. Most of customers in cluster_0 are hospital and healthcare, while cluster_1 represents the hospitality and hotel industry, and the financial services is the largest industry group in cluster_2.

Part of cluster_1 demographic input is shown in Figure 4.19, and the histograms of the variables *BusinessSize*, *Industry*, and *CompanyType* are illustrated in Figure 4.20.



	CustomerID integer CustomerID	BusinessSize polynominal	Industry polynominal	Type polynominal
1	7	XL	INSURANCE	PUBLIC COMPANY LIMITED
2	8	L	INSURANCE	COMPANY LIMITED
3	17	M	FINANCIAL SERVICES	PUBLIC COMPANY LIMITED
4	21	L	HOSPITAL & HEALTH CARE	PUBLIC COMPANY LIMITED
5	23	L	INSURANCE	PUBLIC COMPANY LIMITED
6	30	L	BANKING	GOVERNMENT AGENCY
7	39	M	INFORMATION TECHNOLOG...	COMPANY LIMITED
8	63	XL	TELECOMMUNICATIONS	PUBLIC COMPANY LIMITED
9	64	M	SEAPORTS AND TERMINALS	COMPANY LIMITED
10	67	M	BANKING	GOVERNMENT AGENCY
11	71	L	INSURANCE	PUBLIC COMPANY LIMITED
12	78	M	INFORMATION TECHNOLOG...	COMPANY LIMITED

Figure 4.19: Sample of Cluster_1 Demographic Input

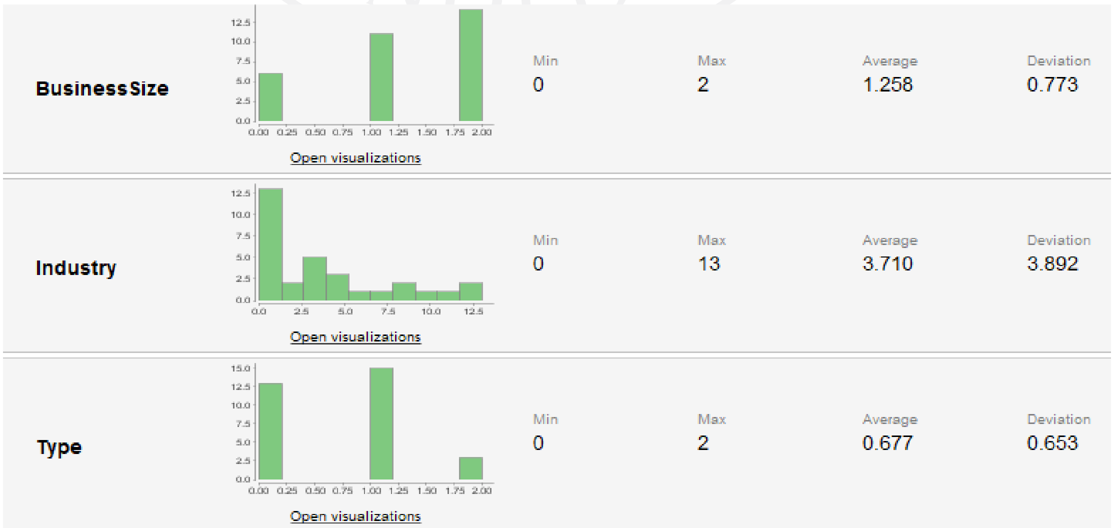


Figure 4.20: Distribution of All Variables for Cluster_1 Input

Results of evaluating with Average within cluster distance and Davies-Bouldin index for cluster_1 demographic input shown as Figure 4.21 and Figure 4.22.

Demographic Clustering (10 rows, 3 columns)

itera...	Clustering.k	Avg. within centroid distance
1	1	5.216
2	2	1.289
3	3	0.725
4	4	0.488
5	5	0.385
6	6	0.308
7	7	0.255
8	8	0.224
9	9	0.198
10	10	0.162

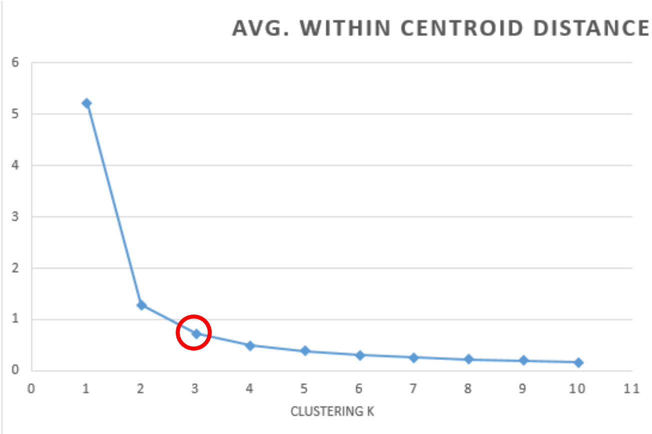


Figure 4.21: Result of Average within Cluster Distance for Cluster_1 Input

Demographic Clustering (10 rows, 3 columns)

itera...	Clustering.k	Davies Bouldin
1	1	∞
2	2	-0.162
3	3	-0.231
4	4	-0.245
5	5	-0.246
6	6	-0.253
7	7	-0.285
8	8	-0.297
9	9	-0.253
10	10	-0.237

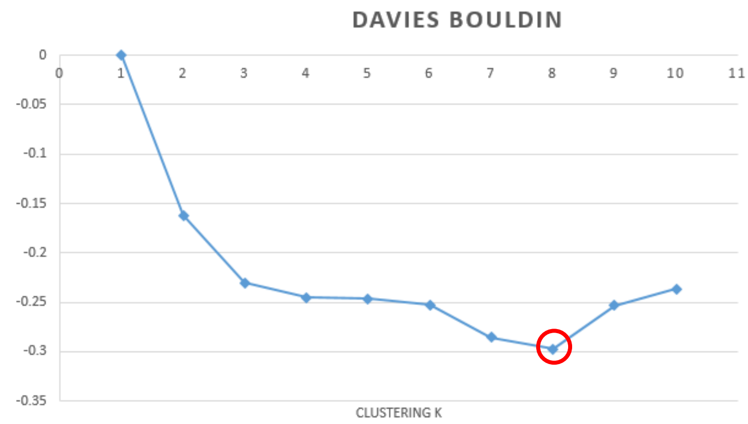


Figure 4.22: Result of Davies-Bouldin Index for Cluster_1 Input

Regarding to average within cluster distance graph, elbow method was chosen to determine the optimal number of cluster, which elbow point is located at $k = 3$. On the other hand, the optimal number cluster in Davies-Bouldin method is $k = 8$, as it is a minimize value. In comparison, $k = 3$ on average within cluster distance graph is more obviously than the Davies-Bouldin value. Therefore, the best k value is 3.

The clustering and segment results with 3 clusters are shown in Table 4.4, and the distribution of the instances within each cluster is detailed in Figure 4.23.

Table 4.4: Instance in Each Cluster for Cluster_1 Input

Cluster	Frequency of Cluster	Percentage
Cluster_0 (Cluster_3)	10	32.26
Cluster_1 (Cluster_4)	6	19.35
Cluster_2 (Cluster_5)	15	48.39

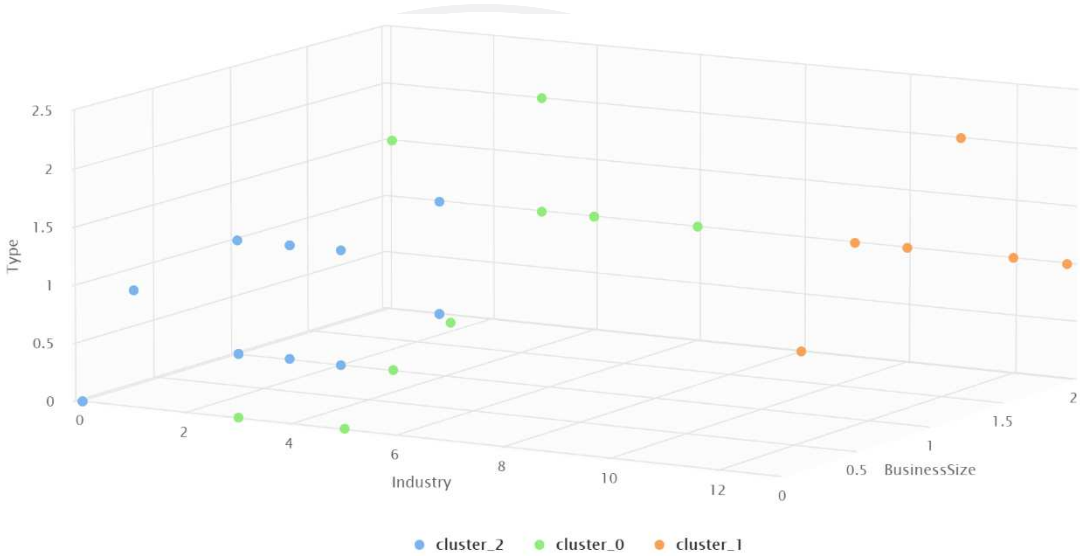


Figure 4.23: Distribution of All Clusters for Cluster_1 Input

Cluster_3 relates to 10 customers, which the majority of this group is the medium and large bank. Similarly, cluster_5 is represents the large financial services company. On the other hand, the industry of customer in cluster_4 is not identical, each customer has different industries.

Part of cluster_1 demographic input is shown in Figure 4.24, and the histograms of the variables *BusinessSize*, *Industry*, and *CompanyType* are illustrated in Figure 4.25.

	CustomerID integer CustomerID	BusinessSize polynominal	Industry polynominal	Type polynominal
1	18	L	FINANCIAL SERVICES	COMPANY LIMITED
2	29	XL	BANKING	PUBLIC COMPANY LIMITED
3	38	XL	FOOD & BEVERAGES	COMPANY LIMITED
4	46	L	FINANCIAL SERVICES	PUBLIC COMPANY LIMITED
5	131	L	INFORMATION TECHNOLOG...	COMPANY LIMITED
6	204	XL	ELECTRICAL/ELECTRONIC	COMPANY LIMITED
7	217	L	FINANCIAL SERVICES	NONPROFIT
8	228	S	INTERNET	COMPANY LIMITED

Figure 4.24: Sample of Cluster_2 Demographic Input

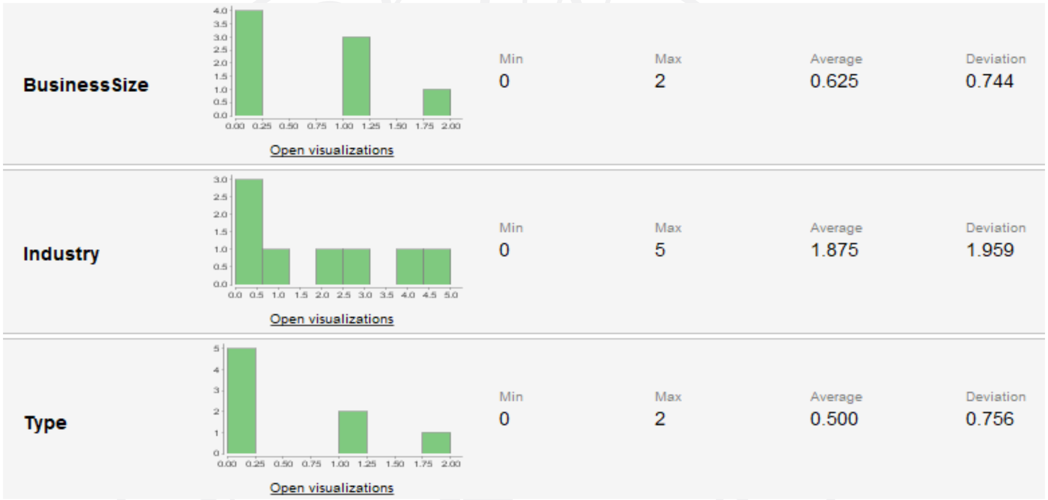


Figure 4.25: Distribution of All Variables for Cluster_2 Input

Results of evaluating with Average within cluster distance and Davies-Bouldin index for cluster_1 demographic input shown as Figure 4.26 and Figure 4.27.

Demographic Clustering (7 rows, 3 columns)

itera...	Clustering.k	Avg. within centroid distance
1	1	4.344
2	2	1.312
3	3	0.688
4	4	0.500
5	5	0.312
6	6	0.188
7	7	0.062

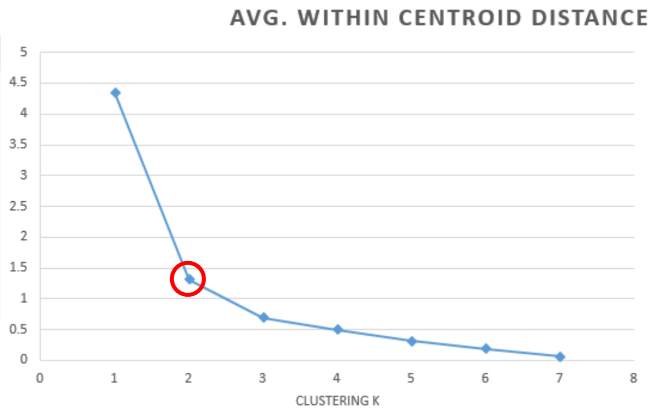


Figure 4.26: Result of Average within Cluster Distance for Cluster_2 Input

Demographic Clustering (7 rows, 3 columns)

iteration	Clustering.k	Davies Bouldin
1	1	∞
2	2	-0.607
3	3	-0.639
4	4	-0.874
5	5	-0.477
6	6	-0.372
7	7	-0.220

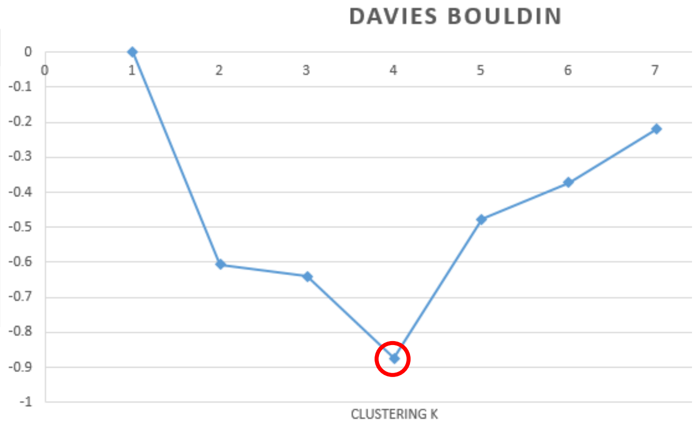


Figure 4.27: Result of Davies-Bouldin Index for Cluster_2 Input

Regarding to average within cluster distance graph, elbow method was chosen to determine the optimal number of cluster, which elbow point is located at $k = 2$. On the other hand, the optimal number cluster in Davies-Bouldin method is $k = 4$, as it is a minimize value. In comparison, $k = 4$ on Davies-Bouldin graph is more clearer. However, the input dataset is contains only 8 instances, in this case it is not suitable to segment into 4 clusters. Therefore, the best k value is 2.

The clustering and segment results with 2 clusters are shown in Table 4.5, and the distribution of the instances within each cluster is detailed in Figure 4.28.

Table 4.5: Instance in Each Cluster for Cluster_2 Input

Cluster	**Frequency of Cluster**	**Percentage**
Cluster_0 (Cluster_6)	4	50
Cluster_1 (Cluster_7)	4	50

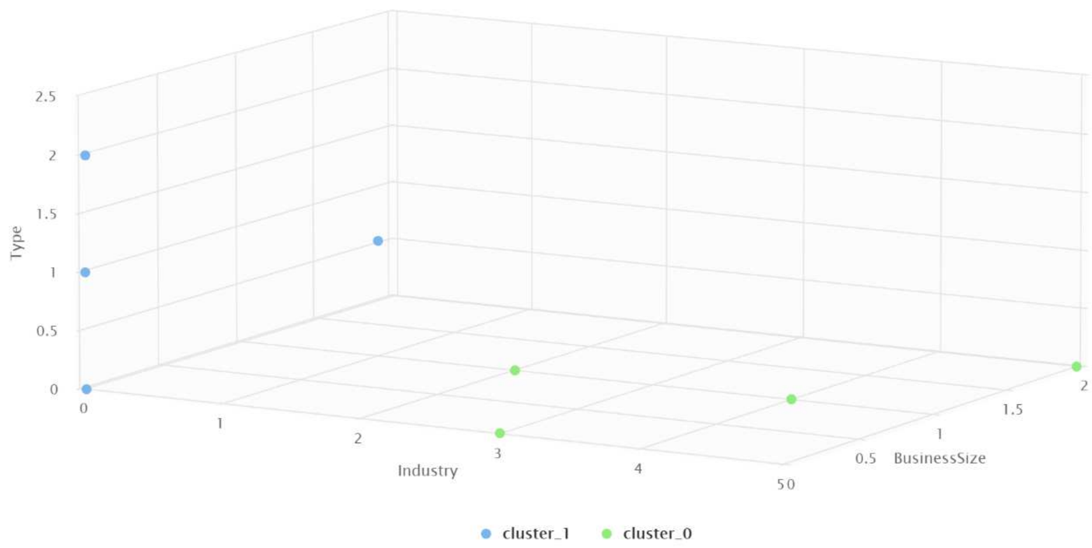


Figure 4.28: Distribution of All Clusters for Cluster_2 Input

Cluster_6 represents three large financial services customers, and one extra-large bank. In contrast, the industry of customer in cluster_7 is not identical, each customer has different industries such as food/beverages, Information Technology, electrical, and Internet industry.

4.3 Finding

The clustering and segment results with 8 clusters are shown in Table 4.6, the statistics of each cluster regarding to demographic and RFM variables are detailed in Table 4.7 and 4.8, and the distribution of 8 clusters is detailed in Figure 4.29.

Table 4.6: Instances in Each Cluster

Cluster	**Frequency of Cluster**	**Percentage**
Cluster_0	68	28.10
Cluster_1	50	20.66
Cluster_2	85	35.12

(Continued)

Table 4.6 (Continued): Instances in Each Cluster

Cluster	Frequency of Cluster	Percentage
Cluster_3	10	4.13
Cluster_4	6	2.48
Cluster_5	15	6.20
Cluster_6	4	1.65
Cluster_7	4	1.65

Table 4.7: Statistic of Each Cluster (Demographic)

Clusters/Variables	Least	Most
Cluster_0 (68)		
BusinessSize	S (5)	M (29)
Industry	Law (1), Business Supplies (1), Advertising (1), Outsourcing (1), Internet (1)	Hospital/Health Care (9)
CompanyType	Educational Institution (4)	Company Limited (42)
Cluster_1 (50)		
BusinessSize	XL (1)	M (27)
Industry	22 Industries (1)	Hospitality (6)
CompanyType	Government Agency (1)	Company Limited (36)
Cluster_2 (85)		
BusinessSize	XL (3)	M (44)
Industry	Foundation (1)	Financial Services (21)
CompanyType	Foundation (1)	Company Limited (48)

(Continued)

Table 4.7 (Continued): Statistic of Each Cluster (Demographic)

Clusters/Variables	**Least**	**Most**
Cluster_3 (10)		
BusinessSize	L (2)	M (5)
Industry	Telecommunications (1), Seaports/Terminals (1), Oil and Energy (1)	Banking (5)
CompanyType	Government Agency (2)	Company Limited (5)
Cluster_4 (6)		
BusinessSize	M (6)	M (6)
Industry	Computer Networking (1), Automotive (1), Food Production (1), Government (1), Construction (1), Cargo Service (1), Government Agency (1)	Computer Networking (1), Automotive (1), Food Production (1), Government (1), Construction (1), Cargo Service (1), Government Agency (1)
CompanyType	Public Company Limited (1)	Company Limited (4)
Cluster_5 (15)		
BusinessSize	M (3), XL (3)	L (9)
Industry	Hospital/Health Care (2)	Financial Services (7)
CompanyType	Company Limited (6)	Public Company Limited (9)

(Continued)

Table 4.7 (Continued): Statistic of Each Cluster (Demographic)

Clusters/Variables	Least	Most
Cluster_6 (4)		
BusinessSize	XL (1)	L (3)
Industry	Banking (1)	Financial Services (3)
CompanyType	Nonprofit (1), Company Limited (1)	Public Company Limited (2)
Cluster_7 (4)		
BusinessSize	S (1), L (1)	XL (1)
Industry	Food/Beverages (1), Information Technology (1), Electrical/Electronic (1), Internet (1)	Food/Beverages (1), Information Technology (1), Electrical/Electronic (1), Internet (1)
CompanyType	Company Limited (4)	Company Limited (4)

Table 4.8: Statistic of Each Cluster (RFM)

Clusters/Variables	Minimum	Median	Maximum
Cluster_0 (68)			
Recency	3	27.82	36
Frequency	1	5.09	23
Monetary	3,040.00	1,194,571.48	5,455,000.00
FirstPurchase	1	11.65	36

(Continued)

Table 4.8 (Continued): Statistic of Each Cluster (RFM)

Clusters/Variables	Minimum	Median	Maximum
Cluster_1 (50)			
Recency	1	25.40	36
Frequency	1	4.22	15
Monetary	5,000.00	898,657.92	4,710,000.00
FirstPurchase	1	9.08	35
Cluster_2 (85)			
Recency	2	26.51	36
Frequency	1	4.93	34
Monetary	760.00	1,417,335.78	5,767,000.00
FirstPurchase	1	11.08	33
Cluster_3 (10)			
Recency	31	34.50	36
Frequency	4	21.30	40
Monetary	7,137,901.02	11,729,828.99	17,970,009.25
FirstPurchase	1	5	11
Cluster_4 (6)			
Recency	26	33.33	36
Frequency	3	23.83	46
Monetary	6,471,250.00	9,661,067.14	15,688,545.00
FirstPurchase	2	4.167	12

(Continued)

Table 4.8 (Continued): Statistic of Each Cluster (RFM)

Clusters/Variables	Minimum	Median	Maximum
Cluster_5 (15)			
Recency	32	34.80	36
Frequency	5	27.13	72
Monetary	6,603,150.00	11,225,188.40	17,773,950.00
FirstPurchase	1	5.80	18
Cluster_6 (4)			
Recency	34	35	36
Frequency	17	27	45
Monetary	24,930,380.00	34,376,270.00	45,426,800.00
FirstPurchase	1	2	3
Cluster_7 (4)			
Recency	27	32.50	36
Frequency	2	31	64
Monetary	23,396,800.00	29,758,855.03	36,214,160.12
FirstPurchase	3	7	16

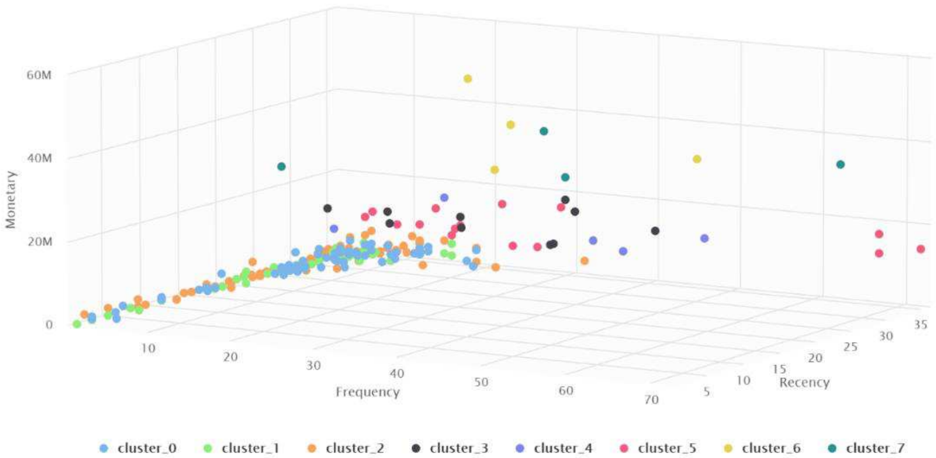


Figure 4.29: Distribution of All Clusters (RFM)

Interpreting and understanding each cluster is essential in developing marketing strategies. Considering Table 4.7, 4.8 and Figure 4.29, each cluster seems to be a group of customers that has some identify features as detailed as follows.

Cluster 1 relates to some 50 customers, composed of 20.66 percent of the whole population. This group seems to be the least profitable group as the customer in this group did not purchase often, the average value of frequency was only 4.22 times over past three years. Cluster 1 can be categorized as very low recency, very low frequency, and very low monetary. For the demographic of this cluster, most of the customers are the medium sized hospitality, customer goods, electronics, transportations, and educational institute respectively.

Contrasted with the customer in cluster 1, the 4 customers in cluster 6 started purchasing products and services at the beginning of 2016 and continued to the end of 2018 with an average of recency 35 (November of 2018). They spent an extremely high amount of money compared with other customers as a result. This group of customers can be classified as very high recency, very high frequency, and very high monetary with a high spending per customer. Actually, they are contributed 16.2 percent of the total sales over the past three years. This group seems to be the most profitable group. The customers in this group are the large sized of financial services and banking industry and most of them are public company.

Cluster 7 with four customers has a very high value for frequency and very high monetary, although lower than the cluster 6. This cluster should be the second high profitable group. The customer in this group have different types of business industries: food and beverage, information technology services, electronics, and internet industry.

Same as cluster 6 and 7, cluster 5 has a very high value of recency, high frequency, and high monetary. Moreover, the customers in this cluster have contributed about 20 percent over 2016-2018, it is a number one contributor as illustrated in Figure 4.30. Therefore, this cluster should be classify as the most profitable group same as the cluster 6. For the demographic, the majority of customers in this cluster are large financial services and insurance companies.

Cluster 3 and 4 have a high recency, medium frequency, and medium monetary, this indicating a smaller amount of spending per customer compared with cluster 6 and cluster 7. The customers in these clusters mainly started purchasing at the first quarter of 2016 and continued to the end of 2018, this point out that customers should be impressed with the products and services of the company. For the demographic, the medium sized bank was represented by cluster 3. Six customers in cluster 4 have different types of industries, which are computer networking, automotive, food production, government, construction, cargo services, and government agency.

Conversely, cluster 2 is the largest-sized group with 85 customers. Customers in this group have a low-medium value of recency and very small value of frequency. Compared with cluster 3, 4, 5, 6, and 7, the average value of monetary of this group is much lower than others, there are 10 times lower than cluster 3, 4, and 5, and 30 times lower than cluster 6 and 7, but the customers in this group have contributed over 14.2 percent of total sales as illustrated in Figure 4.30. Hence, this cluster seems to be another potential group. Cluster 2 represents the medium and large sized of financial services, information technology services, and insurance. Most of customers in this group are limited company and some of them are public company.

Lastly, Cluster 0, which relates to 68 customers, composed of 28.10 percent of the whole population has many features similar to cluster 2 such as the low-medium of recency, very small frequency, and low monetary. In terms of the first purchasing, both cluster 0 and 2 show that the customers started purchasing at the last quarter of 2016. These groups seem to be the new customers, and therefore have a certain level of uncertainty for the profitability in the long term. They might be a potential customer or unprofitable at all. The customers in this group have totally different industries, but there are three main industries, hospital and healthcare, food and beverage, and food production.

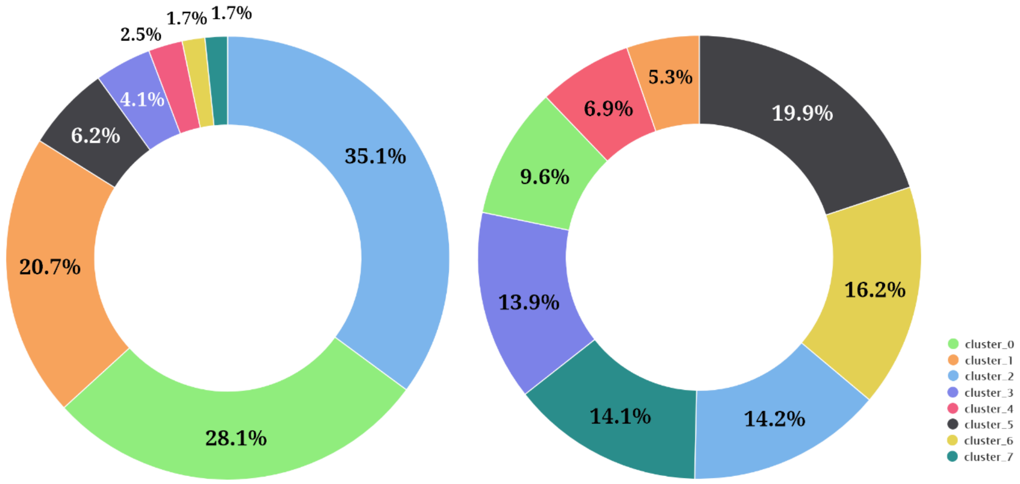


Figure 4.30: Customer Segmentation (Left) and Associated Sales (Right) by Cluster

Figure 4.30 shows that, in the whole population of the customers, 9.6 percent (cluster 5, 6, 7) were the extremely highly profit, 35 percent (cluster 2) were medium to high profit with the highly spending. Another 6.6 percent (cluster 3, 4) were the medium profit with high recency, and the remaining 48.8 percent (cluster 0, 1) were low profit group with lowest value of recency, frequency, and monetary.

The summary of all clusters and their characteristic both RFM and demographic is detailed in Table 4.9.

Table 4.9: Summary of Customer Segment Profiles

Profile Group	**Cluster**	**Behaviors**		**Demographics**	
“Luxury” Extremely High Value	5, 6, 7	R F M	High High High	Size Industry Type	L and XL Financial, Banking, IT, Food, Electric, Internet, Insurance, Hospital Public Company Limited, Company Limited, Nonprofit
“Premium” High Value	2	R F M	Low Low Low	Size Industry Type	M and L Financial, IT, Insurance, Automotive, Real Estate, Computer Software Public Company Limited, Company Limited
“Compact” Medium Value	3, 4	R F M	High Medium Medium	Size Industry Type	M, L and XL Banking, IT, Automotive, Oil and Energy, Food, Telecommunication, Government, Cargo Service Company Limited, Public Company Limited, Government Agency

(Continued)

Table 4.9 (Continued): Summary of Customer Segment Profiles

Profile Group	Cluster	Behaviors		Demographics	
"Eco" Low Value	0, 1	R	Low	Size	S, M, and L
		F	Low	Industry	Hospital, Financial, Food and Beverage, IT, Insurance, Chemical
		M	Low	Type	Public Company Limited, Company Limited, Government Agency, Educational Institution, Nonprofit

The most valuable customers of X-company, "Luxury group", have contributed more than 50 percent of the total sales over past three years (2016-2018), whereas the least valuable group made up only 14.9 percent of the total sales. For each of these customers profile, it is requisite to find out which products and services that customers in each profiles have purchased, which depicted in Figure 4.31, 4.32, 4.33, and 4.34, and the summary of the most and least products and services purchased by each customer group is detailed in Table 4.10.

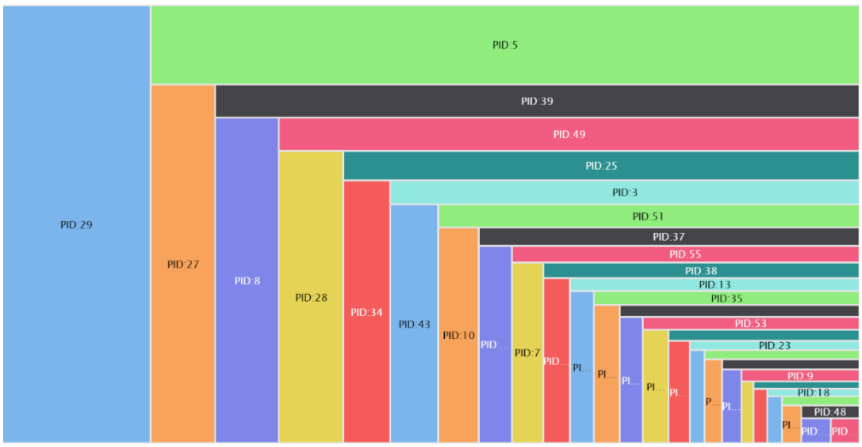


Figure 4.31: Products and Services Purchased by Luxury Group

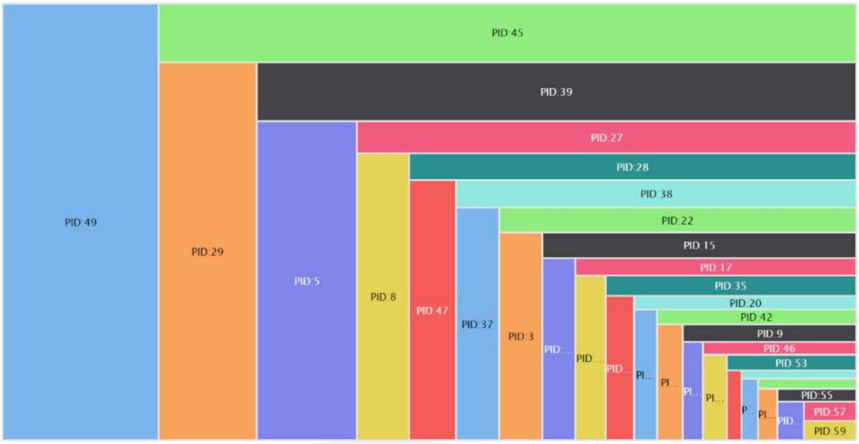


Figure 4.32: Products and Services Purchased by Premium Group

Figure 4.33: Products and Services Purchased by Compact Group

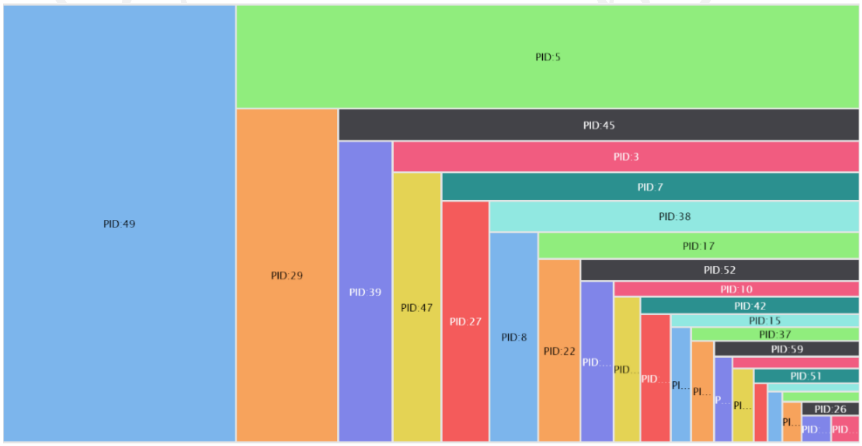


Figure 4.34: Products and Services Purchased by Eco Group

Table 4.10: Summary of the Products Purchased by Each Group

Profile Group	Most Purchased	Least Purchased	
Luxury	0029 (90)	0011 (1)	0041 (1)
	0005 (78)	0014 (1)	0044 (1)
	0027 (32)	0016 (1)	0048 (1)
	0008 (29)	0018 (1)	0054 (1)
	0039 (29)	0020 (1)	0056 (1)
Premium	0049 (68)	0013 (1)	0055 (1)
	0045 (41)	0021 (1)	0056 (1)
	0029 (37)	0036 (1)	0057 (1)
	0039 (35)	0041 (1)	0059 (1)
	0005 (32)	0043 (1)	
Compact	0049 (40)	0002 (1)	0035 (1)
	0005 (39)	0010 (1)	0037 (1)
	0027 (25)	0018 (1)	0040 (1)
	0039 (24)	0022 (1)	0044 (1)
	0003 (15)	0033 (1)	0053 (1)
		0034 (1)	0057 (1)
Eco	0049 (131)	0002 (1)	0026 (1)
	0005 (83)	0004 (1)	0032 (1)
	0029 (44)	0018 (1)	0035 (1)
	0045 (22)	0020 (1)	0051 (1)
	0039 (21)	0024 (1)	

Table 4.10 shows that the most common products that every customer groups have purchased are 0005 and 0039, the common products that three out of four

customer groups have purchased are 0029 and 0049, the products that only two customer groups have purchased are 0027 and 0045, and the products that only one group has purchased are 0003 and 0008.

Table 4.11: Summary of Spending on the Most Purchased Products

Products	**Minimum Spending**	**Average Spending**	**Maximum Spending**
0003	123,600.00	767,421.51	4,250,000.00
0005	20,000.00	546,883.59	9,080,000.00
0008	124,500.00	1,134,703.67	20,096,000.00
0027	12,500.00	228,761.61	2,080,000.00
0029	7,980.00	139,215.54	1,644,000.00
0039	81,425.00	1,275,990.13	13,900,000.00
0045	3,456.00	93,408.13	580,000.00
0049	18,000.00	246,630.92	4,300,000.00

Table 4.11 shows that customers have an extremely high spent on product 0008 and 0039 compared with other products. These products seem to be the potential products for X-company. In contrast with product 0029 and 0045, the customers have a low spent on these two products, as resulted in minimum and average spending values. The remaining products seem to be the common products that customers purchased, the product 0003 and 0005 have much higher spent compared with product 0027 and 0049.

Regarding all the findings mentioned above, the X-company could gain a better understanding of its customer in many perspectives. Moreover, as stated in the research objectives about the recommendation by the researcher to the X-company, there are some possible marketing strategies as follows.

- A retention strategy for the best profitable customs (Luxury group) such as providing them an amazing proactive customer service, loyalty and rewards programs, follow up them more closely, and personalized marketing.
- Up Sell and Cross Sell can be implanted to Luxury, Premium, and Compact group as they have purchasing power. X-company should lead the customers to find out their purchasing intention such as products demo, sharing sessions, or proof of concept (POC) the profitable products.
- Promotion can be applied to the premium group as they have high purchasing power but they did not purchase much as resulted in RFM values. Promotion should motivate them to purchase more. This can be applied together with up-selling and cross-selling strategies.
- The strategy for the eco group would be wait and see because it seems like they are regular customers with small spending and not have much intention to make big deals. Some potential customers in this group can be applied the promotions, up-selling, and cross-selling strategies to boost them up.

Table 4.12 shows the summary of strategies mentioned above for each customer group and include the detail of each strategy and product.

Table 4.12: Summary of Marketing Strategies

Up-Selling	
Luxury	X-company should convince the customers to purchase more on product 0005, 0008, and 0039, which already in most purchased list.
Premium	X-company should convince the customers to purchase more on product 0005, and 0039, which already in most purchased list.
Compact	X-company should convince the customers to purchase more on product 0003, 0005, and 0039, which already in most purchased list.
Eco	X-company should convince the customers to purchase more on every products, especially for the product 0005 and 0039 that already have a high spending in other groups.
Cross-Selling	
Luxury	X-company may try to sell other products and services to the customers based on their existing. From the top 10 most purchased products, product 0003, 0025, 0028, and 0034 should be considered.
Premium	X-company may try to sell other products and services to the customers based on their existing. From the top 10 most purchased products, product 0003, 0008, and 0028 should be considered.
Compact	X-company may try to sell other products and services to the customers based on their existing. From the top 10 most purchased products, product 0008, 0025, and 0028 should be considered.
Eco	X-company may try to sell other products and services to the customers based on their existing. From the top 10 most purchased products, product 0003, 0008, and 0027 should be considered.

(Continued)

Table 4.12 (Continued): Summary of Marketing Strategies

Promotion	
Premium	The sales promotion like discounted products, Buy More Save More, Branded Gifts should be implemented to motivate them to purchase more. This can be applied together with up-selling and cross-selling strategies.
Eco	The sales promotion like discounted products, buy more save more, branded gifts, training course giveaways should be implemented on the potential customers.
Retention	
Luxury	The personalized marketing like targeted emails, product recommendations, social media marketing, custom video message, and etc. should be applied to this customer group to improve their experience, increase company royalty, and create consistency across channels. Providing an amazing proactive customer service to support them 24/7 and follow up them closely for the any inconvenient caused.

4.4 Conclusion

A case study of X-Company, the IT solutions and services provider in Thailand, has been presented in this research to demonstrate how to construct the customer segmentation model using K-means clustering technique with the RFM analysis using RapidMiner Studio. It has been shown in the analysis that there are three steps in the whole segmentation processes that are very crucial and the most resources consuming. Data pre-processing, 2-tier cluster processing, and segments interpretation and profiling.

From the results of customer segmentation, eight clusters have been presented. Regarding to the customer's demographic and RFM values, eights clusters were categorized into four main profiles: (1) Luxury, (2) Premium, (3) Compact, and (4) Eco. Luxury represents the extremely high value customer group as it has been contributed over 50 percent of the total sales over past three years, although the population is only 9.6 percent. Most of the customers in luxury group are large and extra large sized of banking and financial industry with very high values of recency, frequency, and monetary. The premium group was identified as a second high value, due to the customers in this group have been contributed about 14.2 percent, even though they have a very low values of RFM. The third group is compact, which the customer in this group have medium to high values of RFM, and most of them are banking, information technology, government, and telecommunication industry that have purchasing power. The last group is eco, this group was categorized as a low value customer group because the customers have very low recency, frequency, and monetary, and has very low contribution for the past three years.

Therefore, the results show that the proposed approach can help to identify the values of the customers based on its demographic and RFM values. In the view of X-Company's representative, she said that the distinct customer groups segmented in the case study can help her and X-Company had better understand its customers in fields of profitability, correspondingly, adopt the marketing strategies for different types of customers to retain and boost the customers from lower value group to upper value group. She also pointed out that the suggested marketing strategies are reasonable for each customer group, further analysis and in-depth study will help it possible and practical

CHAPTER 5

DISCUSSION

5.1 Discussion

From the results of the customer segmentation constructed in this research, the main characteristics of each customer group are provided to the X-Company along with the marketing strategies for each particular group. X-Company's representative said that the eight clusters and four customer groups help her a lot to understand the customers in the sense of customers demographic and the behaviours in the ways that RFM analysis was presented, which just only the purchased date, frequency, and amount of spending. She also mentioned that the suggested marketing strategies are reasonable based on the analysis; nevertheless, if the customer's needs and motivations can be defined, it will help a lot for developing the marketing strategy.

According to the segmentation model in this research was composed with the basis of the demographic and behavioural segmentation. Hence, it did not provide the insights of customer in terms of needs and motivations behind the purchase behaviour, which it is crucial for the marketing strategy development. Therefore, the results from this research can fully support the X-Company to understand more about its customers, but might not help them a lot in the marketing strategy development.

5.2 Recommendation for Further Research

According to this case study research, the first thing to consider when performing the customer segmentation is the structure of data, because if the collected data was stored in the non-structured format, it will consume a lot of time in the data processing process. The second thing is the quality and the number of instances in the sample data is very important for the clustering technique, because clustering technique is an unsupervised learning algorithm that means the algorithm classify both tangible and intangible patterns and relationships between the data without any prior information. Therefore, the accuracy of clustering is depends on the quality and the amount of data. Lastly, the researchers must clearly understand the business of your case, because in the interpretation process the in-depth customer information must be used to analyze and interpret along with the segmentation results, especially when the marketing strategy needs to be developed, everything about products, customers, and sales information need to be understood.

BIBLIOGRAPHY

3X 2Y Digital . (2013). *Digital and Social Media Marketing Communications Management 2013 – 2014.* Retrieved from 3X 2Y Digital DSMMCM1314: https://3x2ydsmmcm1314.wordpress.com/2013/12/15/segmentation-pros-and-cons/

Ahmed, S. (2004). Applications of data mining in retail business. *International Conference on Information Technology: Coding and Computing,2*, 455-459.

Anastasia. (2018). *Psychographic Segmentation in Marketing: Definition & Examples*. Retrieved from Cleverism: https://www.cleverism.com/psychographic-segmentation-in-marketing-definition-examples/?fbclid=IwAR1YsviCh6-axzrUbIS0IyfUSnHOaoNSqhU0Dos-amiHNFNuH_pTfYircGQ

Armstrong, G., & Kotler, P. (2005). *Marketing: An Introduction.* Upper Saddle River, N.J.: Prentice Hall.

Berson, A., Smith, S., & Thearling, K. (2000). *Building data mining applications for CRM*. New York, NY: McGraw-Hill.

Chan, C. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications, 34*(4), 2754-2762.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management, 19*(3), 197-208.

Chen, M.-C., Chiu, A.-L., & Chang, H.-H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications, 28*(4), 773-781.

Chen, Y., Zhang, G., Hu, D., & Wang, S. (2006). Customer Segmentation in Customer Relationship Management Based on Data Mining. *IFIP International Federation for Information Processing Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management*, 288-293.

Cheng, C.-H., & Chen, Y.-S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36*(3), 4176-4184.

Chris, J. (2016). *Psychographic Personality Types*. Retrieved from Joseph Chris: http://www.josephchris.com/psychographic-personality-types

Custora. (2018). *Customer-Centric Marketing*. Retrieved from Custora U: https://university.custora.com/for-marketers/customer-centric-marketing/basic/the-history-of-customer-centric-marketing

Davies, D. L., & Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224-227.

Dennis, C., Marsland, D., Cockett, T., & Hlupic, V. (2003). Market segmentation and customer knowledge for shopping centers. *IEEE Transactions on Image Processing*.

Dwivedi, S., Kasliwal, P., & Soni, S. (2016). Comprehensive Study of Data Analytics Tools (RapidMiner, Weka, R tool, Knime). *Symposium on Colossal Data Analysis and Networking (CDAN)*, 1-8.

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review, 14*(4), 532-550.

Forester, T. (1986). *The information technology revolution.* Cambridge, MA: MIT.

Forgy, E. W. (1965). Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics, 21*, 768.

George, A. L., & Bennett, A. (2005). *Case Studies and Theory Development in the Social Sciences.* Cambridge: MIT Press.

Gunter, B., & Furnham, A. (1992). Consumer Profiles: an introduction to psychographics.

Ha, S., Bae, S., & Park, S. (2002). Customers time-variant purchase behavior and corresponding marketing strategies: an online retailers case. *Computers & Industrial Engineering,43*(4), 801-820.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* Amsterdam: Elsevier, Morgan Kaufmann.

Hart, W. C. (1995). Mass customization: conceptual underpinnings, opportunities and limits. *International Journal of Service Industry Management, 6*(2), 36-45.

Hubbard, L. (n.d.). *Attitude Factor in Market Segmentation.* Retrieved from azcentral: https://yourbusiness.azcentral.com/attitude-factor-market-segmentation-28315.html

Hughes, A. M. (1994). *Strategic database marketing.* Chicago: Probus Publishing Company.

Hughes, A. M. (2012). *Strategic database marketing*. New York: McGraw-Hill.

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications, 26*(2), 181-188.

Idoine, C., Krensky, P., Brethenoux, E., & Linden, A. (2019). *Magic Quadrant for Data Science and Machine Learning Platforms*. Retrieved from Gartner: https://www.gartner.com/doc/reprints?id=1-65WC0O1&ct=190128&st=sb

Jordan, J. (2016). *Grouping data points with k-means clustering*. Retrieved from Jeremy Jordan: https://www.jeremyjordan.me/grouping-data-points-with-k-means-clustering/

Kaymak, U. (2001). Fuzzy target selection using RFM variables. *9th IFSA World Congress and 20th NAFIPS International Conference,2*, 1038-1043.

Kazemia, A., & Babaeib, M. (2011). Modelling Customer Attraction Prediction in Customer Relation Management using Decision Tree: A Data Mining Approach. *Journal of Optimization in Industrial Engineering, 9*, 37-45.

Kenton, W. (2019a). *Economies of Scope*. Retrieved from Investopedia: https://www.investopedia.com/terms/e/economiesofscope.asp

Kenton, W. (2019b). *Economies of Scale*. Retrieved from Investopedia: https://www.investopedia.com/terms/e/economiesofscale.asp

Kim, S.-Y., Jung, T.-S., Suh, E.-H., & Hwang, H.-S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications, 31*(1), 101-107.

Kotler, P., & Keller, K. L. (2009). *Marketing Management (12th Edition).*

Kubiak, B. F., & Weichbroth, P. (2010). Cross- And Up-selling Techniques In E-Commerce Activities. *Journal of Internet Banking and Commerce, 15*(3).

Larsen, N. (2010). *Market Segmentation - A framework for determining the right target customers.* Aarhus School of Business.

Lefait, G., & Kechadi, T. (2010). Customer Segmentation Architecture Based on Clustering Techniques. *2010 Fourth International Conference on Digital Society*.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281-297.

Marketing Tutor. (n.d.). *Psychographic Segmentation Definition with Examples.* Retrieved from Marketing Tutor: https://www.marketingtutor.net/psychographic-segmentation-definition-examples/?fbclid=IwAR1ttuxxanuxZRo82dbH2Zc-Bz1yivtPnipay0DqlDQRMrZbaBGNkb74vLE

MathWorks. (2019). *Davies Bouldin Evaluation.* Retrieved from MathWorks: https://www.mathworks.com/help/stats/clustering.evaluation.daviesbouldinevaluation-class.html

Mccarty, J., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research,60*(6), 656-662.

McDonald, M., & Dunbar, I. (2004). *Market segmentation: how to do it, how to profit from it.* Oxford: Butterworth-Heinemann.

Naeem, S., & Wumaier, A. (2018). Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K. *International Journal of Computer Applications, 182*(31), 7-14.

Namvar, M., Gholamian, M. R., & Khakabi, S. (2010). A Two Phase Clustering Method for Intelligent Customer Segmentation. *2010 International Conference on Intelligent Systems, Modelling and Simulation*.

Nataraj, S. (2010). Customer Retention – CRM Application. *Issues in Information Systems, 6*(2), 44-47.

Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592-2602.

Oldach, M. (2019). *Towards Data Science*. Retrieved from Medium: https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92

Parvatiyar, A., & Sheth, J. N. (2001). Customer Relationship Management: Emerging Practice, Process, and Discipline. *Journal of Economic and Social Research 3*(2), 1-34.

Patil, P. (2018). *K Means Clustering : Identifying F.R.I.E.N.D.S in the World of Strangers*. Retrieved from Towards Data Science: https://towardsdatascience.com/k-means-clustering-identifying-f-r-i-e-n-d-s-in-the-world-of-strangers-695537505d

Pickton, D., & Broderick, A. (2005). *Integrated marketing communications*. Harlow: Prentice Hall.

Rajagopal, S. (2011). Customer Data Clustering using Data Mining Technique. *International Journal of Database Management Systems (IJDMS), 3*(4).

Ramageri, B. M. (2010). DATA MINING TECHNIQUES AND APPLICATIONS. *Indian Journal of Computer Science and Engineering, 1*(4), 301-305.

Rouse, M., Ehrens, T., & Kiwak, K. (2019). *CRM (customer relationship management).* Retrieved from SearchCustomerExperience: https://searchcustomerexperience.techtarget.com/definition/CRM-customer-relationship-management

Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society, 24*, 483-502.

SAS Institue. (1998). *From Data to Business Advantages: Data Mining, the SEMMA Methodology and SAS Software.* Cary, North Carolina: SAS Institue.

Schijns, J. M., & Schröder, G. (1996). Segment selection by relationship strength. *Journal of Direct Marketing, 10*(3), 69-79.

Shah, D., T. Rust, R., Parasuraman, A. P., Staelin, R., & S. Day, G. (2006). The Path to Customer Centricity. *Journal of Service Research, 9*(2), 113-124.

Shaw, R., & Reed, D. (1999). *Measuring and valuing customer relationships: how to develop measures that drive profitable CRM strategies.* London: Business Intelligence.

Smith, W. (1956). Product Differentiation and Market Segmentation As Alterna-tive Marketing Strategies. *Journal of Marketing*.

Stone, B. (1995). *Successful Direct Marketing Methods.* Lincolnwood, IL: NTC Business Books.

Sturman, A. (1977). Case study methods. In: J. P. Keeves (ed.). *Educational research, methodology and measurement: an international handbook (2nd ed.)*, 61-66.

Swift, R. S. (2001). *Accelerating customer relationships using CRM and relationship technologies*. Upper Saddle River: Prentice Hall.

Terblanche, N. S. (2006). The relationship between customer satisfaction and loyalty: an application of the American Customer Satisfaction Index in the South African fast food industry. *Management Dynamics, 15*(2), 31-43.

Thanuja, V., Venkateswarlu, B., & Anjaneyulu, G. S. (2011). Applications of Data Mining in Customer Relationship Management. *Journal of Computer and Mathematical Sciences, 2*(3), 399-580.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411-423.

Tsai, C.-Y., & Chiu, C.-C. (2004). A purchase-based market segmentation methodology. *Expert Systems with Applications, 27*(2), 265-276.

UC R Programming. (2017). *UC Business Analytics R Programming Guide.* Retrieved from K-means Cluster Analysis: https://uc-r.github.io/kmeans_clustering

Walker, B. (2017). *Insights on Today's Healthcare Consumer: The Pros and Cons of Consumer Segmentation.* Retrieved from Insights C2B Solutions: https://insights.c2bsolutions.com/blog/the-pros-and-cons-of-consumer-segmentation

Woo, J., Bae, S., & Park, S. (2005). Visualization method for customer targeting using customer map. *Expert Systems with Applications. 28*(4), 763-772.

Wu, J., & Lin, Z. (2005). Research on customer segmentation model by clustering. *ACM International Conference Proceeding Series*, 113.

Yang, Y., & Padmanabhan, B. (2005). A hierarchical pattern-based clustering algorithm for grouping web transactions. *IEEE Intelligent Systems,20*, 55-59.

Yin, R. K. (2014). *Case study research: design and methods.* Los Angeles: Sage.

BIODATA

Name: Mr. Ponlacha Rojlertjanya

Position: Senior Business Representative

Company: Datapro Computer Systems Company Limited

Nationality: Thai

Education Qualification: Bachelor of Engineering in Computer Engineering,

King Mongkut's University of Technology Thonburi

Expertise: Information Technology, Business Innovation

Email ID: ponlacha.rj@gmail.com

Bangkok University

License Agreement of Dissertation/Thesis/ Report of Senior Project

Day 16 Month August Year 2019

Mr./ Mrs./ Ms. Ponlacha Rojlertjanya now living at 105/64
Soi Phutthamonthon Sai 2 Soi 25 Street Phutthamonthon Sai 2
Sub-district Sala Thammasop District Thawi Watthana
Province Bangkok Postal Code 10170 being a Bangkok
University student, student ID 7610201738

Degree level ☐ Bachelor ■ Master ☐ Doctorate

Program MBI Department IKI-SEA School Graduate School
hereafter referred to as "the licensor"

Bangkok University 119 Rama 4 Road, Klong-Toey, Bangkok 10110 hereafter referred to as "the licensee"

Both parties have agreed on the following terms and conditions:

1. The licensor certifies that he/she is the author and possesses the exclusive rights of dissertation/thesis/report of senior project entitled CUSTOMER SEGMENTATION BASED ON THE RFM ANALYSIS MODEL USING K-MEANS CLUSTERING TECHNIQUE: A CASE OF IT SOLUTION AND SERVICE PROVIDER IN THAILAND submitted in partial fulfillment of the requirement for Master of Management (Business Innovation) of Bangkok University (hereafter referred to as "dissertation/thesis/report of senior project").
2. The licensor grants to the licensee an indefinite and royalty free license of his/her dissertation/thesis/report of senior project to reproduce, adapt, distribute, rent out the original or copy of the manuscript.
3. In case of any dispute in the copyright of the dissertation/thesis/report of senior project between the licensor and others, or between the licensee and others, or any other inconveniences in regard to the copyright that prevent the licensee from reproducing, adapting or distributing the manuscript, the licensor agrees to indemnify the licensee against any damage incurred.

This agreement is prepared in duplicate identical wording for two copies. Both parties have read and fully understand its contents and agree to comply with the above terms and conditions. Each party shall retain one signed copy of the agreement.

[redacted] Licensor

(Ponlacha Rojlertjanya)

[redacted] Licensee

([redacted])

Director, Library

[redacted] Witness

(Suchada Chareanpunsirikul, D.B.A.)

Dean, Graduate School

[redacted] Witness

([redacted])