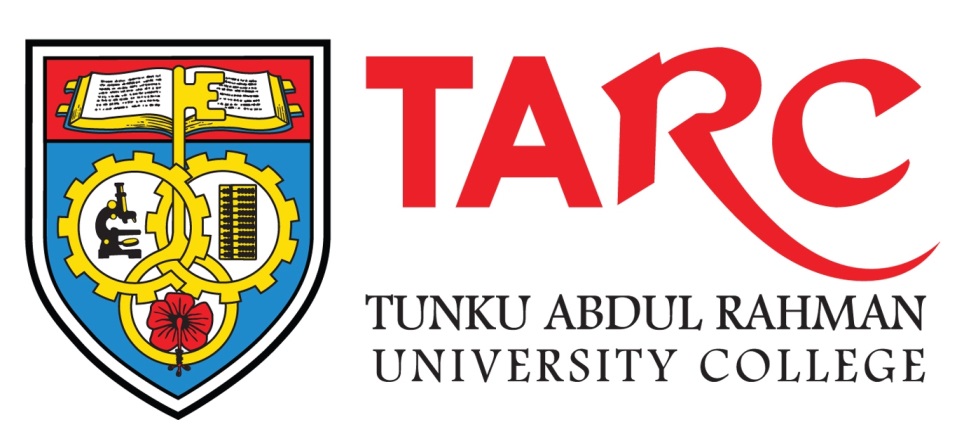
Customer Behavioral Analytics Using RFM

By

Lee Kang Wenn



FACULTY OF COMPUTING AND

INFORMATION TECHNOLOGY

TUNKU ABDUL RAHMAN UNIVERSITY COLLEGE

KUALA LUMPUR

ACADEMIC YEAR

2018

Customer Behavioral Analytics Using RFM

By

Lee Kang Wenn

Supervisor: Dr Chaw Jun Kit

A project report submitted to the

Faculty of Computing and Information Technology

in partial fulfillment of the requirement for the

Bachelor of Computer Science (Honours)

**Department of Information and Communication Technology**

Faculty of Computingand Information Technology

Tunku Abdul Rahman University College

Kuala Lumpur

2018

Copyright by Tunku Abdul Rahman University College.

All rights reserved. No part of this project documentation may be reproduced, stored in retrieval system, or transmitted in any form or by any means without prior permission of Tunku Abdul Rahman University College.

# 

Declaration

The project submitted herewith is a result of my own efforts in totality and in every aspect of the project works. All information that has been obtained from other sources had been fully acknowledged. I understand that any plagiarism, cheating or collusion or any sorts constitutes a breach of TAR University College rules and regulations and would be subjected to disciplinary actions.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Lee Kang Wenn

Bachelor of Computer Science (Honours) in Software Engineering

Abstract

The main objective of this project is to extract semantic customers behavioral information through data collected from a food delivery company, Running Man. The project aims to explore customer behavior by analyzing transaction data. One of the methods applied in this project is RFM. RFM (Recency, Frequency, Monetary) is a method to profile customer into groups. RFM score calculated will be then clustered to gain more insights into the pattern. In this project, K-means clustering will be used to analyses customer purchasing behavior, and to cluster customers’ RFM score. Visuals such as RFM heat map, RFM score binning, RFM scatterplot can be used to visualize customers purchasing behavior. RFM scoring is a simple and efficient model that can be used to analyse customers purchasing behavior. Clustering RFM score can also show majority of the customers belong to one big group. The final model is able to cluster customer’s RFM score into 4 groups of customers, with each group having their unique characteristics and show insights related to the business. However, the accuracy of unsupervised learning algorithm cannot be validated. In future, more unsupervised methods can be evaluated to cluster RFM scoring. There are more unsupervised methods than k-means clustering such as mixture models, hierarchical clustering can be tested out. Binning of customer into 5 bins for every score might not produce the best accuracy thus optimization is required.

Acknowledgement

First and foremost, I have to thank my research supervisor, Dr Chaw Jun Kit. Without his assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. I would like to thank you very much for your support and understanding throughout the whole project.

I would also like to show gratitude to my teammates, including Ms Yang Poh Yee, Ms Chan Peck Hui, and Ms Shim Hiu Yin.

Getting through my dissertation required more than academic support, and I have many, many people to thank for listening to and, at times, having to tolerate me over the past three years. I cannot begin to express my gratitude and appreciation for their friendship.

Most importantly, none of this could have happened without my family. To my parents – it would be an understatement to say that, as a family, we have experienced some ups and downs in the past three years. Every time I was ready to quit, you did not let me and I am forever grateful. This dissertation stands as a testament to your unconditional love and encouragement.

**Table of Contents**

[Declaration ii](#_Toc522546948)

[Abstract iii](#_Toc522546949)

[Acknowledgement iv](#_Toc522546950)

[Table of Contents v](#_Toc522546951)

[1 Introduction 2](#_Toc522546952)

[1.1 Objectives 2](#_Toc522546953)

[1.2 Research Question and Hypothesis 3](#_Toc522546954)

[1.3 Background 3](#_Toc522546955)

[1.4 Advantages and Contributions 3](#_Toc522546956)

[1.5 Project Plan 4](#_Toc522546957)

[1.5.1 Proposed Features 4](#_Toc522546958)

[1.5.2 Milestone 4](#_Toc522546959)

[1.5.3 Research Development Plan 5](#_Toc522546960)

[1.5.4 Testing plan 5](#_Toc522546961)

[1.6 Project Team and Organization 5](#_Toc522546962)

[1.7 Thesis Outline 6](#_Toc522546963)

[2 Research Background and Related Work 8](#_Toc522546964)

[2.1 Company Background 8](#_Toc522546965)

[2.2 Project Background 8](#_Toc522546966)

[2.3 Literature Review 10](#_Toc522546967)

[2.3.1 Consumer Behavioral Analytics 10](#_Toc522546968)

[2.3.2 Analytic Process Model 11](#_Toc522546969)

[2.3.3 Supervised and Unsupervised Machine Learning Algorithm 12](#_Toc522546970)

[2.4 Feasibility Study 20](#_Toc522546971)

[2.4.1 Technical Feasibility 20](#_Toc522546972)

[2.4.2 Economic Feasibility 20](#_Toc522546973)

[2.4.3 Operational Feasibility 21](#_Toc522546974)

[2.4.4 Organizational Feasibility 21](#_Toc522546975)

[2.5 Chapter Summary and Evaluation 21](#_Toc522546976)

[3 Methodology and Requirements Analysis 24](#_Toc522546977)

[3.1 Methodology 24](#_Toc522546978)

[3.1.1 Fact Gathering 24](#_Toc522546979)

[3.1.2 Interview 24](#_Toc522546980)

[3.1.3 Background Study 25](#_Toc522546981)

[3.1.4 Literature Review 25](#_Toc522546982)

[3.1.5 Fact Recording 25](#_Toc522546983)

[3.1.6 Fact Analysis 25](#_Toc522546984)

[3.2 Requirements Analysis 26](#_Toc522546985)

[3.2.1 Project Scope 26](#_Toc522546986)

[3.2.2 Development Environment 26](#_Toc522546987)

[3.2.3 Operation Environment 26](#_Toc522546988)

[3.2.4 External Interface Requirements 27](#_Toc522546989)

[3.2.5 Non-functional Requirements 27](#_Toc522546990)

[3.2.6 Functional Requirements 27](#_Toc522546991)

[3.3 Chapter Summary and Evaluation 27](#_Toc522546992)

[4 System Design 29](#_Toc522546993)

[4.1 Process 29](#_Toc522546994)

[4.2 K-means Clustering 30](#_Toc522546995)

[4.3 Data Preparation 33](#_Toc522546996)

[4.3.1 Data Description 33](#_Toc522546997)

[4.4 Data Preprocessing 34](#_Toc522546998)

[4.5 Modelling 35](#_Toc522546999)

[4.5.1 Customer Lifetime Value 35](#_Toc522547000)

[4.6 Chapter Summary and Evaluation 35](#_Toc522547001)

[5 Results 38](#_Toc522547002)

[5.1 Descriptive statistics of transaction data 38](#_Toc522547003)

[5.2 Analytical statistics of transaction data 41](#_Toc522547004)

[5.3 Chapter Summary and Evaluation 53](#_Toc522547005)

[6 System Deployment Error! Bookmark not defined.](#_Toc522547006)

[6.1 Conclusion 55](#_Toc522547007)

[6.2 Limitation and Future Improvements 55](#_Toc522547008)

[7 References 56](#_Toc522547009)

[8 Bibliography 58](#_Toc522547010)

[9 Appendices 61](#_Toc522547011)

[9.1 Developer Guide 61](#_Toc522547012)

[9.1.1 Software used 61](#_Toc522547013)

[9.2 User Manual 61](#_Toc522547014)

Chapter 1

Introduction

# Introduction

As penetration of smartphones and mobile apps increases, the technology is going to trigger a massive influx of big data. Consumer analytics is at the epicenter of a big data revolution. Technology helps capture rich and plentiful data on consumer behavior in real time.

The massive increase in the amount of data collected and stored by organizations around the world over the past few decades is evident. In conjunction with this, the ability to access and analyze this data is quickly becoming more and more important. However, some firms do not have the resources to perform consumer analytics and it’s often inaccurate (Erevelles et al., 2016). Analyzing consumer data often requires huge amount of time, manpower, and professionals, which leads to high cost for small organization to carry out data analysis.

Therefore, this project aims to use machine learning to build a more efficient and simpler consumer analytics model for small and medium-sized enterprise that does not have much resources to carry out data analysis. Data extracted that are trained by machine learning will be examined in terms of predictive accuracy in order to be applied in real-life situation. By examining information such as customers’ purchasing behavior, such information can help the management to make better decision in terms of marketing and planning, hence increases sales performance (Raorane and Kulkarni, 2011).

## Objectives

This project aims to achieve several objectives. The main objective of this project is to extract semantic customers behavioral information through data collected from a food delivery company, Running Man. Customer behavioral refers to actions done by customers on the website and number of customers visiting the website. The data is then processed by going through features extraction, pre-processing and input into machine learning algorithm for output. We will also explore the most accurate prediction model for the project. We will then provide suggestions such as cross-selling products, peak hour prediction and conducting promotions events based on the output from the prediction model.

Investigating customer behavior in retail contexts is essential to obtain various formal indicators that are interesting from the marketing research viewpoint such as the conversion rates, to further improve the food delivery experiences. It is hard for small and medium-sized enterprises (SMEs) to process big data in order to analyze and predict market and customer behavior. With good data analytics model being set up, it can help the company to meet customers’ need through fulfilling niche. It also helps the company in making better decisions that will benefit to the company and the customer (Sen et al., 2016).

## Research Question and Hypothesis

The experts or experienced businessman can predict the customers' needs, what they are going to buy by using intuition. However, for the ordinary marketer could not estimate the right offers for the right clients which may cause wrong prediction, resulting in high purchase of low demanding products, hence affect the company's revenue. With the use of customer behavior analytics by utilizing the knowledge of IoT and machine learning, we can analyze the past data patterns and trends to generate likely products and outcomes that can help the business to run smoothly and successfully. How small and medium-sized enterprises benefit from customer behavioral analytics without allocating majority of the resources to data analysis? What are the challenges faced when conducting customer behavioral analytics? Which machine learning algorithm is most appropriate in predicting customer behavior?

Due to the problems stated, we wish to investigate on the various methodology in predicting customer purchasing patterns and behaviors that may lead to revenue growth. To do so, we aim to further study on the most effective approach to study and analyze customers trend, so that we are able to achieve a win-win situation, where company’s revenue grows while increasing the customer satisfaction towards the services and products provided.

## Background

More and more companies embrace Artificial Intelligence to assist them in decision making. For example, Ford Motors uses consumer analytics to innovate its product and design (Erevelles et al., 2016). Ford uses insights from big data instead of traditional research method, such as vehicle engine statistics and accident statistics to speed up their innovation process (Satell, 2014).

## Advantages and Contributions

The amount of unstructured data collected and stored by business is evident. Furthermore, the business model for business has evolved significantly during recent years and the concept is now used in the context of e-commerce, strategy, and innovation management. Data-driven business model in the start-up world is growing exponentially (Hartmann et al., 2014). Thus, it is important for small and medium-sized enterprise to follow up and gain benefits from this business model. However, most of the businesses are struggling to organize unstructured data and turning them into useful information. Therefore, a much simpler and low cost solution is required to fulfill these needs. In this project, the main concern is to find out what information the existing data can provide, and the prediction or output the system can provide to the business owner for decision making.

## Project Plan

### Proposed Features

Sub-section numbering should be limited to a maximum of 3 levels (e.g. 2.3.1) in order to avoid confusion.

The system has several features.

Functional requirements of the system are as follow:

1. The system shall be able to process the data inputted by the user.
2. The system shall be able to perform training and learning on the data inputted.
3. The system shall also be able to perform calculation on the data.
4. The system shall be able to identify customers’ purchasing pattern.
5. The system shall be able to predict the likelihood of customer to buy a particular product.
6. The system shall be able to provide suggestions such as cross-selling product to the retailers.

Nonfunctional requirements of the system are as follow:

1. The result generated by the system should be accurate and reliable.
2. The system should be economical - the running cost of the system is not high.
3. The system should also be functional and able to perform the functions that it is required to.

### Milestone

The proposed project milestone is as follows:

|  |  |  |
| --- | --- | --- |
| Milestone | Milestone Goal | Deadline |
| Concept approval | Proposed concepts and development of the system is approved | 24/11/2017 |
| (Chapter 1) Introduction | Project scope and specification is approved | 01/12/2017 |
| (Chapter 2) Research / Background study | To carry out background research, including literature review, preliminary research, system background. | 29/12/2017 |
| (Chapter 3) Requirement gathering & analysis | Gather requirement in the system and perform analysis. | 05/03/2018 |
| (Chapter 4) System Design | System design such as UML Diagrams is completed | 23/03/2018 |
| Data Collection | Collect data needed for the system, unstructured data such as transaction detail. | 23/03/2018 |
| Data processing | Study the data pattern and perform features extraction and normalization of the data to make sure the data is suitable to proceed with machine learning. | 23/04/2018 |
| System development | Develop the predictive model with machine learning algorithm. | 28/05/2018 |
| System testing | Test the system with testing dataset to measure prediction accuracy. | 08/06/2018 |
| (Chapter 5)  Results and System Preview | To check whether the expected output is achieved, based on the test cases designed. | 08/06/2018 |
| Final Testing | Final testing of the system | 22/06/2018 |
| (Chapter 6)  Conclusion and Recommendations | Documentation of supportive document, user guide and discussion of further improvement | 20/07/2018 |

### Research Development Plan

The software development model in this project is Extreme programming software development methodology. Extreme programming is similar to Agile development that focuses on short development cycles and close interaction with customers. Incremental feature refactoring is the main activity in this software lifecycle. The period for implementation of this project is short, which takes only 4 months, which is from March 2018 to June 2018. Therefore, a flexible development model is needed. TensorFlow is the main machine learning library in this project. Another supported library such as scikit learn library will also be used in this project. Features in the system will be altered or added based on the time frame given.

### Testing plan

This system will be tested out by carrying out user testing in real scenario where real data is collected and inputted into the system.

## Project Team and Organization

The members in this project are Lee Kang Wenn and Chan Peck Hui. The initial distribution of workload within the team is shown in table 1.2.

Table 1.2: Project Team and Organization

|  |  |  |
| --- | --- | --- |
| **Customer Analytics system** | **Lee Kang Wenn** | **Chan Peck Hui** |
| Data collection | x |  |
| Data pre-processing |  | x |
| Features extraction |  | x |
| Machine learning and training | x |  |
| Customer purchasing prediction | x |  |
| System testing |  | x |

## Thesis Outline

Chapter 2 will discuss the background of the company, project’s background, literature review for this project, and feasibility study of the project.

Chapter 3 revisit the user-oriented problems and requirements in general. The chapter also describes the software development model in the subsequent sections.

Chapter 4 describes the various areas of design for the new proposed system. Areas of design include user interfaces, security, databases, processes, reports. This chapter describes the various tools and techniques used, the rationale for using them, any problems faced and how they resort to solve or reduce the impact of the problems faced.

Chapter 5 discusses a test plan for the system that utilizes various software testing strategies and methods. The results of the data collected from the research, such as experiment, simulation, etc. and various test cases will be included in this chapter.

Chapter 2

Research Background and Related Work

# Research Background and Related Work

In this chapter, relevant researches related to the tasks will be represented. This chapter will look into Consumer behavioural analytics, analytic process model and Artificial Neural Network.

## Company Background

The company we are associated is Running Man Sdn Bhd, a Food Delivery company. It provides delivery services of food from many food stores within the delivery area, where their main customers are students, young adults, adults. This company delivers to popular locations such as Kuala Lumpur, Mahkota Cheras, Cheras, Bangsar, Petaling Jaya, Setapak, Ampang. This company has been operated for few years, and has a stable sale.

Recently, they plan to create new marketing strategies and learn about their customers more in order to provide better services and improve customer satisfaction. By doing so, the owner requires various information regarding the customer’s behaviours and trend. Thus, our project is to solve this problem by analysing the data for insights.

## Project Background

More and more companies embrace Artificial Intelligence to assist them in decision making. For example, Ford Motors uses consumer analytics to innovate its product and design [(Erevelles et al., 2016)](https://paperpile.com/c/LOabYZ/cQwG). Ford uses insights from big data instead of traditional research method, such as vehicle engine statistics and accident statistics to speed up their innovation process [(Satell, 2014)](https://paperpile.com/c/LOabYZ/F9OaK).

This solution benefits two parties: the consumer and the merchant. The system can be implemented to provide “personalized shopping experience or services to the consumer”. Furthermore, the system can also be implemented to provide sophisticated insights to assist top management in decision making, product perfecting, and more importantly, customer satisfaction. When a firm embraces new technology and constantly seeks for transformation of knowledge, the knowledge creation process grows exponentially. With these knowledges in hand, a firm can innovate and enhance their products and services [(Kozlenkova et al., 2014)](https://paperpile.com/c/5k36rQ/vV70D).

Before a consumer prediction analytic model is created, an iteration of process is needed.



Figure 2.3.1: Process of modelling consumer prediction analytics

The first step is business understanding. We need to analyse and understand the business logic and the policy of the business in business perspective. Next up is data understanding. We need to study all the available data and investigate the relationship between all available data in order to perform data mining correctly. Data preparation is then carried out to perform ensure all the data is valid for the modelling. In modelling, various techniques are applied into the model and tested. The most suitable technique is then proceeded to evaluation. In evaluation, the model is evaluated in different measures such as accuracy, performance and efficiency. Overfitting of model might occur therefore an evaluation of the model is needed to make sure the model can be then be deployed. Lastly, the model will be deployed into real system to start servicing the user.

## Literature Review

### Consumer Behavioral Analytics

Consumer Behavioural Analytics is the analysis of the past data on what the customers do and how they act in a manner that will reflect in what they do and how they will react in the future. Consumer analytics is important to a company, it brings benefits such as understanding customers’ need, customers’ purchasing power and many more [(Erevelles et al., 2016)](https://paperpile.com/c/5k36rQ/YZDAx). If there is strong behavioural analytics exists, it can be widely used in helping to build a smarter business with social commerce so that retailers can now record and track the customers on how do they buy, what are their choice criteria, when and how frequent do they buy and their pathing channels when checking out products.

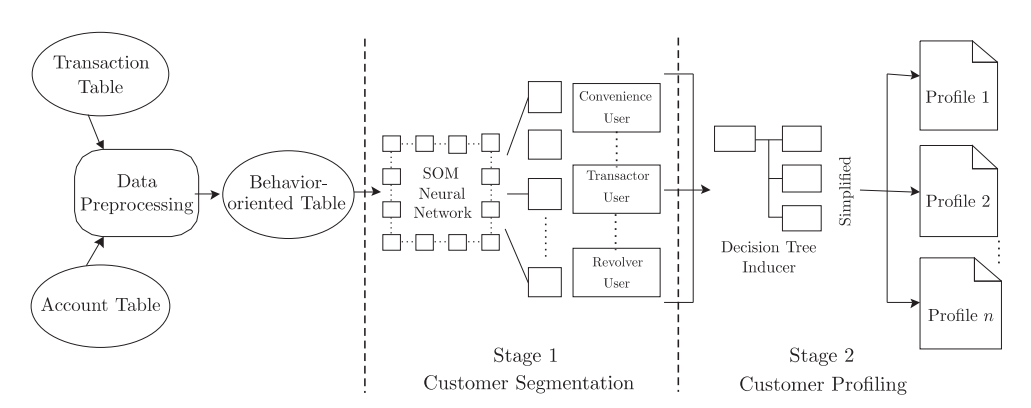


Figure 2.3.2 The two-stage framework of consumer behaviour analysis [(Hsieh and Chu, 2009)](https://paperpile.com/c/5k36rQ/YaX9).

The study of consumer analytics lies at the junction of Big Data and consumer behaviour. Big data is a hot issue in today’s world. 4.4 Zettabytes of data exist in the digital universe today, by 2020, the digital universe is expected to reach 44 zettabytes [(IDC, 2014)](https://paperpile.com/c/5k36rQ/HSXDo). Since data provide behavioural insights about consumers; marketers are able to translate those insights into market advantage. Big data is a top business priority and drives enormous opportunity for business improvement [(Kennedy, 2011)](https://paperpile.com/c/5k36rQ/iZcjP). Nevertheless, the first problem is that manually analyses the conglomeration of raw data to gain insights is inefficient and ineffective.

Consumer analytics can be performed by deductive or inductive approaches, where deductive approaches interpret consumer behaviour based on existing theories and model, while inductive approaches do not make any assumptions or hypothesis before the interpretation. Deductive approaches have been widely used, providing good results. However, the need to obtain even more insights has directed marketers’ interest towards inductive prediction approaches. Studies have shown that using inductive approaches consumer analytics can advance the understanding of marketing phenomena more compared to using deductive approaches [(Erevelles et al., 2016)](https://paperpile.com/c/5k36rQ/YZDAx). Without interconnecting the relationship among consumers’ purchases, customers’ flow and path on the web, deductive approaches would be inaccurate.

### Analytic Process Model

Prediction model is part of analytics process in data science. It is a model which study historical data to forecasting and make prediction for future. The prediction model is using statistics number to show the likelihood of that particular prediction will happen in future. Table 2.3.1 below with diagram and description will describe the whole process from data collection stage until prediction making stage.

Table 2.3.1: The Analytics Process Model and explanation

|  |  |
| --- | --- |
|  | |
| Source Data | unstructured data from various places such as: excel files, log book |
| Data Mining | Retrieve the data from source data and discover their correlation and pattern. |
| Pre-process Data | Data filtering process to ensure data is having minimal error before proceed to the next step. For example, deciding what to do with missing data - discard outlier or replace with values. |
| Transform Data | Because data is stored in various form and format, therefore data transformation is needed to transform the data into suitable data format for machine learning. |
| Analytic | Visualizing the information so that it’s easier to understand, in terms of diagrams such as histogram, bar chart. |
| Analytic Application | Extract knowledge from the visualized information and apply the knowledge into application. |

### Supervised and Unsupervised Machine Learning Algorithm

**Ordinary Least Square**

In statistical modelling, regression is a statistical method which estimates the relationship between variables and the outcome of output for this statistical model is continuous real value, also known as **function approximation** [(Pedregosa et al., 2011)](https://paperpile.com/c/5k36rQ/MVEn). There are many regression models that exists, such as Ridge Regression (impose penalty on size of coefficient), Lasso Regression (estimates sparse coefficients).

Ordinary Least Squares, also known as Linear Regression, is one of the simplest regression models in regression models. Linear Regression fits data with the best line which “goes through” the data points. The differences between the predicted point and the actual observation is the residue.

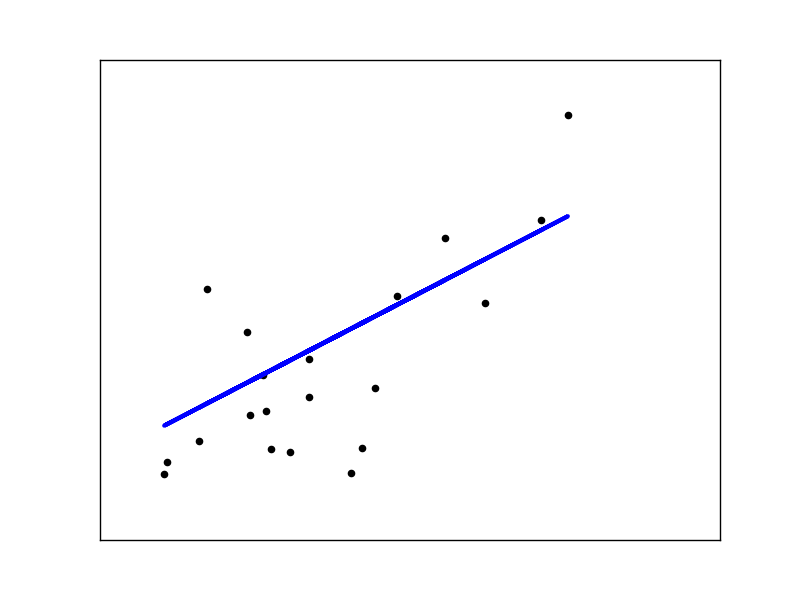


Figure 2.3.3: A linear regression visualization

Linear regression can find out the sum of error in the dataset. Gradient descent is an iterative method to find the minimum of a function. On the other hand, gradient ascent can be used to determine the maximum of a function.

Gradient descent starts with an initial set of data; in each iteration, it decreases each data in proportion to its partial derivative. There will be gradient descent step size, also known as learning rate, along with features and the data is given as input to the learning algorithm. The partial derivative specifies how much a small change in the data would change the error.

Linear regression computes the least squares solution using a singular value decomposition of X.

However, Linear regression only limited to linear relationship, which means it only looks at dependent and independent variables. Some of the data might not have straight relationship between two of the variables. Also, linear regression is sensitive to outliers. Outliers will affect the linear relationship plotted thus affect the accuracy of the relationship [(Sciencing, 2017)](https://paperpile.com/c/5k36rQ/uZ7T).

Linear Regression is a very powerful statistical technique and can be used to generate insight on the data that we, human usually could not understand without plotting the graph. For example, linear regression can be used in business sector to generate insight on consumer behaviour, understanding business and factors influencing profitability. By conducting a linear analysis on the sales data of the company, the company could forecast their sales in the future [(Xu et al., 2016)](https://paperpile.com/c/5k36rQ/QiDB).

**Support Vector Machine**

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier’s detection. It is mostly used in classification problems. In SVM, each data item is plotted as a point in n-dimensional space, where n is the number of features in the dataset) with the value of each feature being the value of a particular coordinate. Then, classification will be performed by finding the hyperplane that differentiate the two classes very well [(Ray et al., 2017)](https://paperpile.com/c/5k36rQ/5aza).

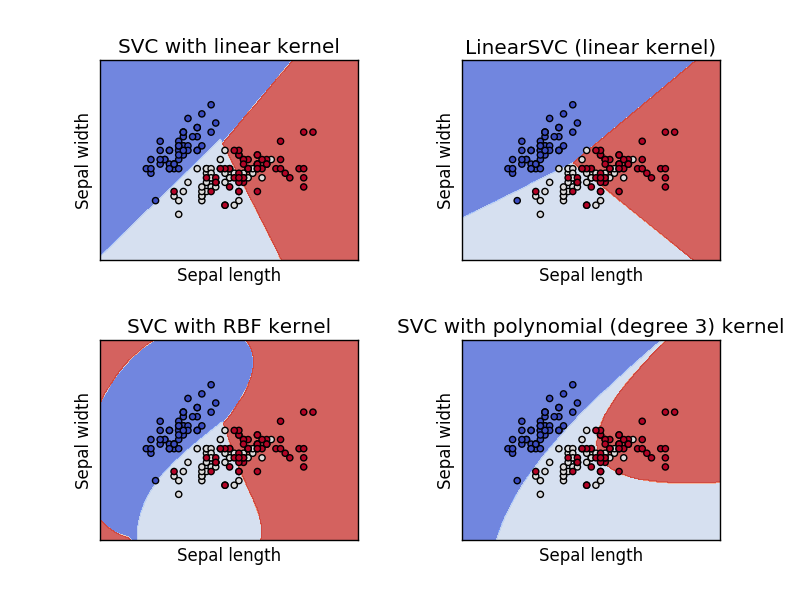
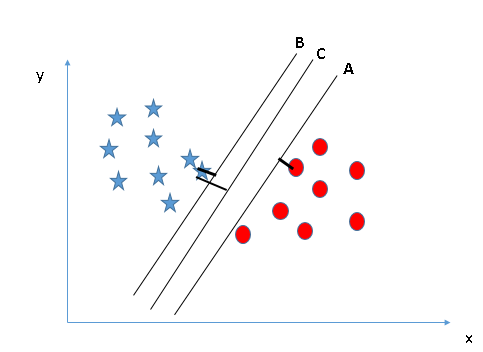


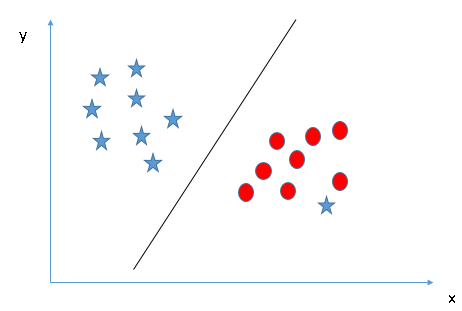
Figure 2.3.4 Support Vector Machine visualization

Support Vectors makes up of coordinates of individual observation. Support Vector Machine is best at segregating two classes with hyperplane or line.

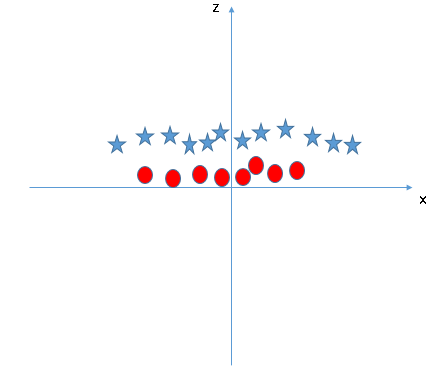
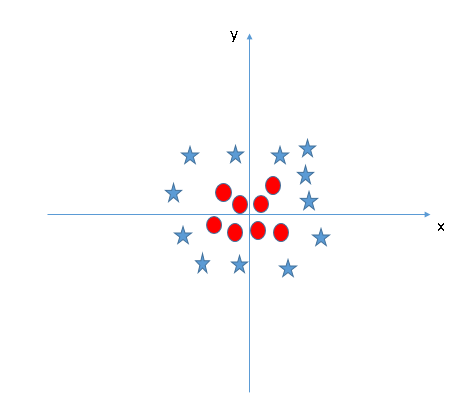
SVM can identify the right hyperplane by maximizing the distances (margin) between nearest data point (either class).



SVM also has a feature to ignore outliers and find the hyperplane that has maximum margin. Hence, we can say, SVM is robust to outliers.



SVM can also solve problem where the dataset could not have linear hyperplane between the classes. It solves this problem by introducing an additional feature. The machine replot the plane based on the new feature - equation on axis x and z, for example.



Support Vector Machine has the following advantages and disadvantages:

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| High-Dimensionality: Effective in high dimensional spaces, in cases where number of dimensions is greater than the number of samples | If the number of features is much greater than the number of samples, the method is likely to give poor performances |
| Memory Efficient: Uses a subset of training points in the decision function (support vectors) | SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation |
| Versatility: different Kernel functions (linear, polynomial, rbf, sigmoid) can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels |  |

Support Vector Machine, has been used successfully in many real-world problems. SVM has been used for text / hypertext categorization, image classification, bioinformatics (Protein classification, Cancer classification), handwritten character recognition [(Yu et al., 2010)](https://paperpile.com/c/5k36rQ/q2N2).

**Random Forest**

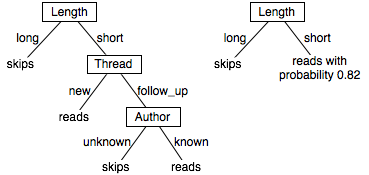
Decision Trees are a non-parametric supervised learning method used for classification and regression. Decision tree learning is one of the most successful techniques for supervised classification learning. Decision tree model aims to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. 

Figure 2.3.5: A decision tree visualization

A decision trees, also known as classification tree, is a flowchart like tree structure, which each internal (non-leaf) node is labelled an input feature. The branch (arcs coming from a node) represents an outcome of the test. Each leaf of the tree is labelled with a class name or class distribution. Random forest contains a forest of decision tree [(Pedregosa et al., 2011)](https://paperpile.com/c/5k36rQ/MVEn).

Random Forest belongs to a larger class of machine learning algorithms called ensemble methods. Ensemble learning involves the combination of several models to solve a single prediction problem.

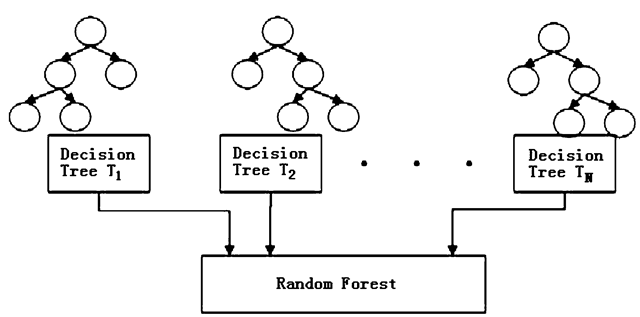


Figure 2.3.6: A random forest visualization

In random forests, each tree in the ensemble is built from a sample drawn with replacement (e.g. bootstrap sample) from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.

Random forest has several features:

* It is unexcelled in accuracy among current algorithms.
* It runs efficiently on large databases.
* It can process thousands of input variables without variable deletion.
* It gives predictions of what variables are important in the classification.
* It has an impressive method for determining missing data and maintains accuracy when a large proportion of the data are missing.
* It has methods for correcting error in class population unbalanced data sets.
* Generated forests can be saved for future use on other data.
* It offers an investigational method for detecting variable interactions.

However, random forest might overfit some of the dataset with noisy classification.

Decision tree is another powerful tool that is used for data prediction. In Astronomy, astronomy has been an active domain for using automated classification techniques. Decision trees is used for filtering noise from Hubble Space Telescope images [(Salzberg et al., 1995)](https://paperpile.com/c/5k36rQ/DtaP). Decision tree also have helped in star-galaxy classification, to determine galaxy counts and discovering quasars in the Second Palomar Sky Survey.

**Artificial Neural Network**

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information [(Specht, 1990)](https://paperpile.com/c/5k36rQ/AaWl). The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well.

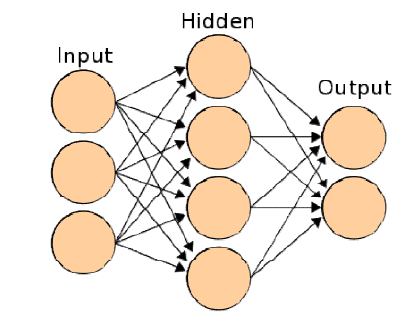


Figure 2.3.7 Artificial Neural Network visualization.

The input is the number of inputs, the data inputted will be push into hidden layer, and the hidden layer will process the information and output the data. In neural network, the input layer, hidden layer and output layer are similar as our human brain, where nerves is nodes in this case.

Advantages of using Artificial Neural Network:

* Work best when answering “what if” situation because it can generate multi possible path.
* Hidden layer will be adjusted based on the data and results will be more reasonable and rational.

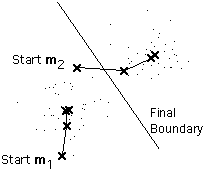
Disadvantages of using Artificial Neural Network

* Take longer time to train the algorithm (involves many iteration)
* Higher computational resources (Requires fast processing unit and big memory)
* Might overfitting the data.

**K-means clustering**

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

It is typically used for scenarios like understanding the population demographics, market segmentation, social media trends, anomaly detection, etc. where the clusters are unknown to begin with.



In training phase of K-Means, K observations are arbitrarily selected (known as centroids). Each point in the vector space is assigned to a cluster represented by nearest (Euclidean distance) centroid. Once the clusters are formed, for each cluster the centroid is updated to the mean of all cluster members. And the cluster formation restarts with new centroids. This repeats until the centroids themselves become mean of clusters, i.e., when updating centroids to mean doesn’t change them. The prediction of a test observation is done based on nearest centroid.

Business Uses

The K-means clustering algorithm is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

This is a versatile algorithm that can be used for any type of grouping. Some examples of use cases are:

* Behavioural segmentation:
  + Segment by purchase history
  + Segment by activities on application, website, or platform
  + Define personas based on interests
  + Create profiles based on activity monitoring
* Inventory categorization:
  + Group inventory by sales activity
  + Group inventory by manufacturing metrics
* Sorting sensor measurements:
  + Detect activity types in motion sensors
  + Group images
  + Separate audio
  + Identify groups in health monitoring
* Detecting bots or anomalies:
  + Separate valid activity groups from bots
  + Group valid activity to clean up outlier detection

Disadvantage of K-means

* The way to initialize the means was not specified. One popular way to start is to randomly choose k of the samples.
* The results produced depend on the initial values for the means, and it frequently happens that suboptimal partitions are found. The standard solution is to try a number of different starting points.
* It can happen that the set of samples closest to **m**i is empty, so that **m**i cannot be updated. This is an annoyance that must be handled in an implementation, but that we shall ignore.
* The results depend on the metric used to measure || **x** - **m**i ||. A popular solution is to normalize each variable by its standard deviation, though this is not always desirable.
* The results depend on the value of k.

## Feasibility Study

### Technical Feasibility

There are many machine learning algorithms to create the consumer prediction model. The machine learning algorithm must be able to fit into the data correctly and produce a accurate prediction.

### Economic Feasibility

Since this project involves mostly software components, most of the costs are the time taken to produce the software application itself. The costs and benefits summary of this project are shown in Table 2.3 and Table 2.4 respectively. It can be seen that the total benefits gained (RM16,000) are higher than the total costs invested (RM6,500) after 4 years upon system implementation.

Table 2.3.2: Total cost used for system implementation in 4 years

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cost | Year 0 | Year 1 | Year 2 | Year 3 |
| Tangible Costs (RM): |  |  |  |  |
| Software Application | 5000 |  |  |  |
| User training | 100 |  |  |  |
| Maintenance | 1,000 | 500 | 500 | 500 |
| Intangible Costs: |  |  |  |  |
| Learning curve cost |  |  |  |  |
| Total Cost: | 6500 | 500 | 500 | 500 |

Table 2.3.3: Benefits gained upon system implementation in 4 years

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Benefits | Year 0 | Year 1 | Year 2 | Year 3 |
| Tangible Benefits (RM): |  |  |  |  |
| Increase number of users | 4,000 | 4,000 | 4,000 | 4,000 |
| Intangible Benefits: |  |  |  |  |
| Increase company reputation |  |  |  |  |
| Increase customer satisfaction |  |  |  |  |
| Total Benefits: | 4,000 | 4,000 | 4,000 | 4,000 |

### Operational Feasibility

The stakeholder of this system will be the top management of the company. Top management will be the people that interact directly with the system while customer will be the indirect viewpoint of the system. Lastly, the proposed system has no confrontations with the company policies and user interface standards. All viewpoints of the system are shown in Figure 2.3.8.

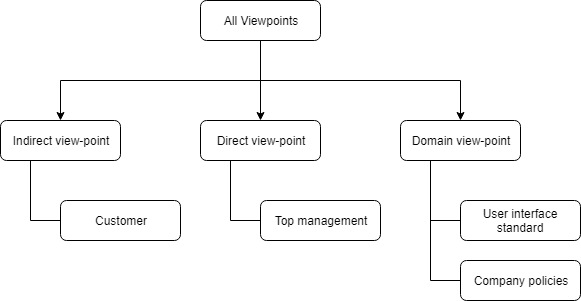


Figure 2.3.8: Viewpoint Hierarchy of the system

### Organizational Feasibility

The objectives of the company is to knowing more information regarding their customer behaviours, improving customer satisfactions and shopping experiences and increase in sales figure. And this project does support all the current objectives of the company.

## Chapter Summary and Evaluation

Company Background

* a Food Delivery company started of in Setapak.

Project Background

* Before a consumer prediction analytic model is created, an iteration of process to study the organisation is needed.

Literature Review

* Customer Behavioural Activity
* Analytic Process Model
* Supervised and Unsupervised Machine learning algorithm
  + Linear regression
  + Support Vector Machine
  + Random Forest
  + Artificial Neural Network
  + K-means Clustering

Feasibility Study

* Technical feasibility: The proposed system is workable with current software and hardware technologies.
* Economic Feasibility. The proposed system has higher benefits than cost in long run.
* Operational Feasibility. The proposed system is accepted by all stakeholders such as top management, customers and company policies.
* Organizational Feasibility. The proposed system does support the company objectives.

Chapter 3

Methodology and Requirements Analysis

# Methodology and Requirements Analysis

The project aims to explore customer behaviour by analysing transaction data. One of the methods applied in this project is RFM. RFM (Recency, Frequency, Monetary) is a method to profile customer into groups.

## Methodology

### Fact Gathering

Few requirements gathering technique were used in this project. Elicitation techniques were used to gather facts about existing system, company background, similar system developed by others and more. Elicitation techniques used are: Interview, background study, and literature review.

### Interview

Interview, as one of the stakeholder-driven elicitation techniques had been used in order to gather and understand users’ requirement in this project.  
Interview is a face-to-face conversation hold between interviewer and interviewee. The interviewer will question the interviewee to discover their opinion or experience on how the current system works and about requirements.

We had found out that we need more information on the current system, therefore we hold an interview session with the founder, Mr Andrew, to understand the company background, their daily operation, business analytics and also to seek requirement from him. Face-to-face interview enable us to probe the company more deeply on the constraints Mr Andrew faced while using the current system.

We have prepared a set of interview questions:

1. How the system works? (Process from customer order to delivery)
2. What data is stored and how it is stored?
3. If two delivery is made at the same time, at nearby location, how would this be handled?
4. Business Canvas (Partnership, key activities, key resources, cost, target customers, revenue, services provided, channel, competitors)
5. What data can you provide? (E.g., Database details)
6. Do you have any current business goals?
7. Have you done any analytics before? (e.g., any queries made to database?)
8. Do you face any problems?
9. What is your future plan?
10. Do you plan to provide additional services or promotions in the future?
11. Have you done any advertisement currently?
12. What are the requirements or requests for us to achieve?
13. Have you done any survey before? If yes, can you provide us the results you obtain (the data/statistical value)?

### Background Study

Background study is another stakeholder-driven technique to acquire, study and analyse the organization, the domain and the system-as-is.

Organizational charts, business plans, financial reports and meeting minutes are documents related to organization are studied to gather information for the system-to-be.

In terms of domain, we can study similar system used in other company such as Walmart, Google to understand more into customer analytics.

Background study is useful in this case as it prepares the fundamental knowledge about the organization and the current system for further analysis and development. We can study the effectiveness of this system in terms of the accuracy of output.  
Besides, this technique also acts as a prerequisite to other elicitation techniques as it can help us to get prepared before carry out other elicitation technique such as interview. For example, we have to know what is the background of the company before preparing questions for interview.  
Furthermore, background study is necessary in this case because some information may not be easily found through other elicitation techniques. For example, using questionnaire we could not questions the user how the system works exactly to understand the workflow of the system.

### Literature Review

Literature review is also another technique used in this project. Journals that are related to customer behaviour analytics were accessed to help in generating the model for customer behavioural analytics. Books on business analytics also used in generating techniques to explore customer behaviour through transaction data.

### Fact Recording

Software that will be used to record facts for this project is Google docs, draw.io. This software will be used to document all the requirements from the user and also to generate diagrams such as flow chart.

### Fact Analysis

User requirement will be analysed using scientific approach such as content analysis and transform to system functions.

## Requirements Analysis

### Project Scope

In this project, I am only focusing on one sub-module in the system - Segmentation of customer’s RFM value.

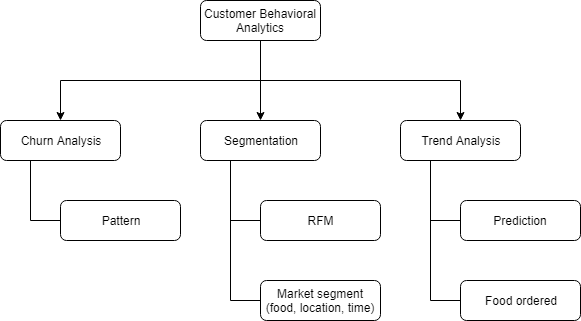


Figure 1.1 Hierarchical Chart showing Project Sub-Systems

### Development Environment

Spyder Integrated Development Environment will be used in this project. TensorFlow python library will be used for k-means clustering. Pandas python library will be used for data pre-processing and normalization. Besides, IBM SPSS will also be used to perform data exploration and visualization. In this project, the dataset will be obtained from the company's database, which consist of transaction data of their daily operation within the Klang Valley.

### Operation Environment

The target customers will be companies which they did not performed customer behavioural analysis before and would like to try out a basic system. The size of the company should be small as this is a rather simple analysis of customer data. The minimum system hardware requirements to run the software are:

* Processor: 1 gigahertz (GHz) or faster processor or SoC
* RAM: 4 gigabyte (GB)
* Hard disk space: 30 GB
* Graphics card: DirectX 9 or later with WDDM 1.0 driver
* Display: 800x600

The minimum system software requirements to run the software are:

* WinPython 3.6.5.0Qt5-64bit
* TensorFlow python library package

### External Interface Requirements

#### System Output

The system output will be the visualization of the result of the machine learning model.

### Non-functional Requirements

#### Accuracy

The system shall have at least 85% of output accuracy.

#### Economic

The running cost for the system should be low.

#### Reliable

The system should also be functional and able to perform the functions that it is required to.

### Functional Requirements

The system shall be able to process the data inputted by the user.

The system shall be able to perform training and learning on the data inputted.

The system shall also be able to perform calculation and clustering on the data using k-means clustering.

The system shall be able to identify customers’ purchasing power in terms of Recency, Frequency and Monetary.

The system shall be able to provide suggestions such as cross-selling product to the retailers.

## Chapter Summary and Evaluation

* Method used for fact gathering: Interview, background study and literature review.
* This project will be focusing on calculation of customer’s RFM value and segmentation of the RFM value.
* The functional and non-functional requirement of the system is listed out.

Chapter 4

System Design

# System Design

## Process

The process of the customer analysis is shown below:



Figure 4.1 Process of Analysing Customer Transaction Data

All the process of the analysis will be done using Spyder IDE, the programming language is python and coding will be needed for each phase of the process. Also, some python library such as NumPy, pandas, TensorFlow, sklearn will also be used to build the process. K-means clustering will be used to cluster customers’ RFM score. Customers that achieve top 20% of RFM score are most likely to respond to personalized marketing such as promotional campaign, vouchers, and so on.

## K-means Clustering

K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyse the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

The K-means clustering algorithm is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

K-means clustering is selected in this project because the purpose of this project is to discover new insights on the customer transaction data which we are not labelling and classifying but to identify new groups in the dataset. Thus, k-means clustering fits best to the requirement and therefore is selected to be used in this project.

#### Algorithm

The Κ-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters Κ and the data set. The data set is a collection of features for each data point. The algorithms start with initial estimates for the Κ centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

1. Data assignment step:

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if ci is the collection of centroids in set C, then each data point x is assigned to a cluster based on



where dist ( · ) is the standard (L2) Euclidean distance. Let the set of data point assignments for each ith cluster centroid be Si.

2. Centroid update step:

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.



The algorithm iterates between steps one and two until a stopping criterion is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

This algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

#### Choosing K

The algorithm described above finds the clusters and data set labels for a particular pre-chosen K. To find the number of clusters in the data, the user needs to run the K-means clustering algorithm for a range of K values and compare the results. In general, there is no method for determining exact value of K, but an accurate estimate can be obtained using the following techniques.

One of the metrics that is commonly used to compare results across different values of K is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing K will always decrease this metric, to the extreme of reaching zero when K is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K.

A number of other techniques exist for validating K, including cross-validation, information criteria, the information theoretic jump method, the silhouette method, and the G-means algorithm. In addition, monitoring the distribution of data points across groups provides insight into how the algorithm is splitting the data for each K.

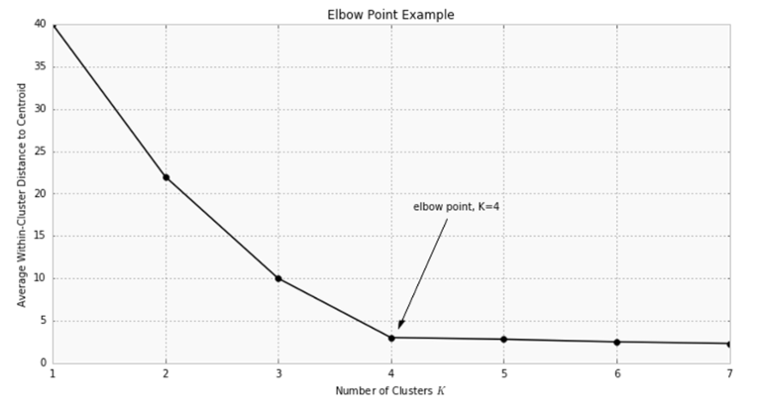


Figure 4.2 Elbow Point Example

## Data Preparation

### Data Description

The dataset consists of the following columns:

|  |  |  |
| --- | --- | --- |
| Column Name | Data type | Description |
| \_id | String |  |
| address.city | String | User’s city |
| address.street | String | User’s location address |
| comments | String | User’s comment when checking out |
| contact.email | String | User’s email |
| contact.fname | String | User’s first name |
| contact.lname | String | User’s last name |
| contact.number | String | User’s contact number |
| coordinate.lat | Double | User’s location in coordinate latitude |
| coordinate.lng | Double | User’s location in coordinate longitude |
| delivery\_time.date | Date | User’s order date |
| delivery\_time.schedule | Time | When to deliver the food (ASAP or later) |
| delivery\_time.time | Time | User’s order delivery time |
| payment.amount | Double | User’s order amount |
| transaction\_date | Date | User’s order date |
| transaction\_id | String | User’s order id |
| task | String | Items to be delivered, to send to another part of the system. |

## Data Preprocessing

We will be using user’s email to identify each unique customer. In order to calculate RFM, we need these three dimensions from the dataset:

* **Recency** – How recently did the customer purchase?
* **Frequency** – How often do they purchase?
* **Monetary Value** – How much do they spend?

Next, we will decide on number of categories to be assigned to each dimension. The number of categories will depend on the type of business, how long the business has been operating and other factors. For example:

**Recency** can be categories into 5 categories:

* score of 5 for last purchase within a week
* score of 4 for within a month
* score of 3 for a quarter month,
* score of 2 for six months
* score of 1 for a year.

**Frequency** can also be categories into 5 categories:

* score of 5 for 10 purchases in a year
* score of 4 for 8 purchases
* score of 3 for 6 purchases
* score of 2 for 4 purchases
* score of 1 for 3 purchases or under

**Monetary Value:** Over RM10,000 (5), RM8,000-RM9,999 (4), RM6,000-RM7,999 (3), RM4,000-RM5,999 (2) RM3,999 or less (1).

Thus, for this dataset, we will identify:

* The most recent transaction date of the customer (**transaction\_date**) and assign a **Recency** score based on the categories above
* Number of purchases done throughout the year (**Order Count**) and assign a **Frequency** score based on the categories above.
* The sum of all the purchase done throughout the year (**Order Sum**) and assign a **Monetary** score based on the categories above.

RFM score is calculated by concatenating Recency, Frequency and Monetary Score together.  
By refering to above scale, a person who spend RM3,000 twice during last six months with their latest purchase on Monday this week would have an Recency score of 5 (within a week), Frequency score of 1 (3 purchases or under) and Monetary value of 3 (between RM6,000-7,999), concatenating 3 of the score directly the customer RFM score will be 513.

With this scoring system for each dimension, we will have 125 possible scores for each customer. The higher the score, the more valuable the customer.

The RFM scores is then normalized by using min-max normalization method. Min-max normalization method is used for normalization step. This method performs a linear transformation on the original data. Suppose that *minA* and *maxA* are the minimum and maximum values of an attribute, A. Then Minmax normalization maps a value, v, of A to v’ in the range of [*newminA*, *newmaxA*] by computing in equation:

## Modelling

RFM parameters are included in clustering. In K-means clustering technique, the number of clusters should be determined by decision maker. In order to find the best number of cluster, we need to run the K-means clustering algorithm for a range of K values and compare the results.

### Customer Lifetime Value

After RFM score for customer is calculated, Customer Lifetime Value (CLV) will be calculated. To calculate CLV for each cluster, weighted RFM method is used. Analytic hierarchy process (AHP) will be needed to identify relative weights of the RFM variables. An average CLV value of each cluster can be calculated with the equation:

refers to normal Recency of cluster ci, is weighted Recency, is normal Frequency, is weighted Frequency, is normal Monetary, and is weighted Monetary.

## Chapter Summary and Evaluation

* K-means clustering is used to segment the customer based on the RFM score assigned to them.
* The data needs to be pre-processed before proceed with segmentation.
* Therefore, the data is analysed in terms of the structure, each columns definition and so on.
* Coding will be done on each phase of the process and will be done in python programming language, with the use of python libraries such as NumPy, pandas, TensorFlow.

Chapter 5

Results

# Results

This chapter review the results produced from the previous chapter. The result are the outputs collected from the machine learning algorithm. These results are then related to business operations to provide business insights.

## Descriptive statistics of transaction data

Descriptive statistics describes the basic structure of the data. Descriptive statistics such as mean, median, standard deviation provide simple summaries about the data. Descriptive statistic can be shown together with simple graphics analysis. For transaction data, several descriptive statistics are used.

1. **Total Number of transactions (per day)**

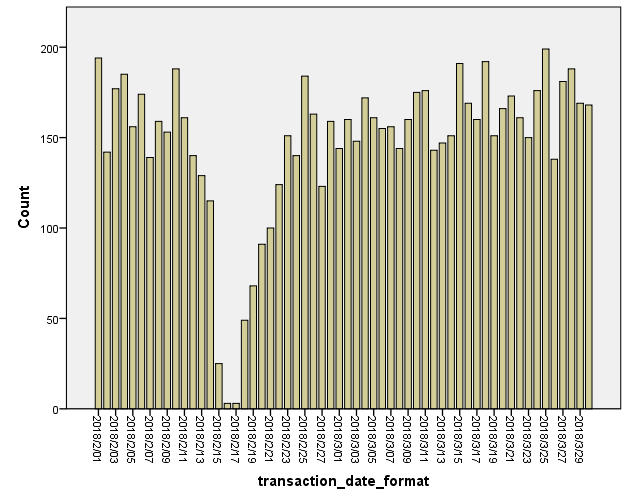


Figure 5.1: Total number of transaction per day

The dataset consists of 8519 transactions made from 1 February 2018 to 30 March 2018. There are average 150 transactions everyday except during 17 February, which is a public holiday (Chinese New Year).

1. **Spending amount**

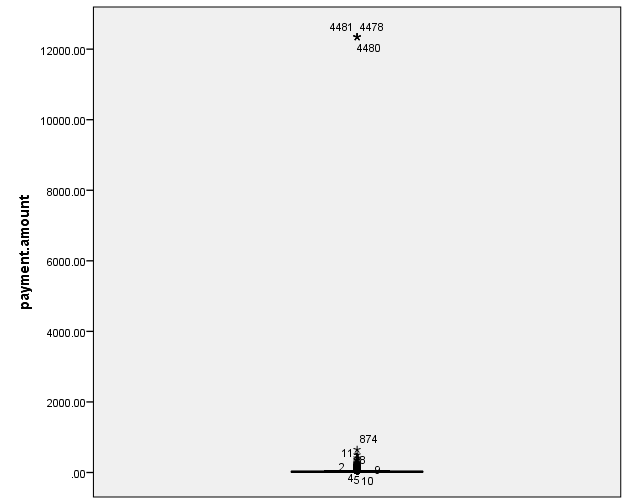


Figure 5.2 Box plot for payment amount

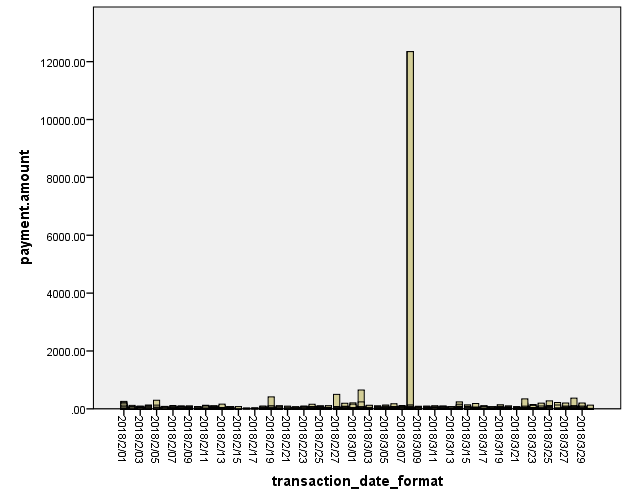


Figure 5.3 Histogram for payment amount

The dataset contains outlier, as shown in figure 5.2, the outlier will be removed later during the preprocessing of the data for k-means clustering. The sum is 285447.53, maximum payment amount is 12345, minimum payment amount is 0, and has a mean of 33.5072, with standard deviation of 299.21449. These values are highly affected by the outlier.

1. **Number of transactions done by per user**

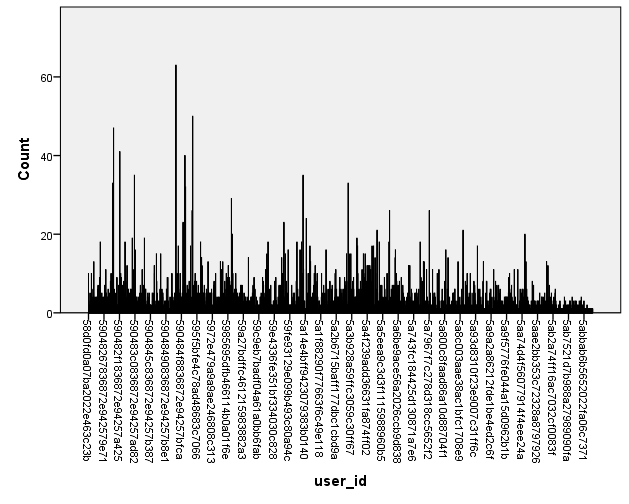


Figure 5.4 Number of transaction done by user

Figure 5.4 shows that most of the customer performed less than 20 transactions in 2 months. Only some of the customer performed more than 20 transaction in 2 months.

## Analytical statistics of transaction data

1. **RFM Bin Counts**

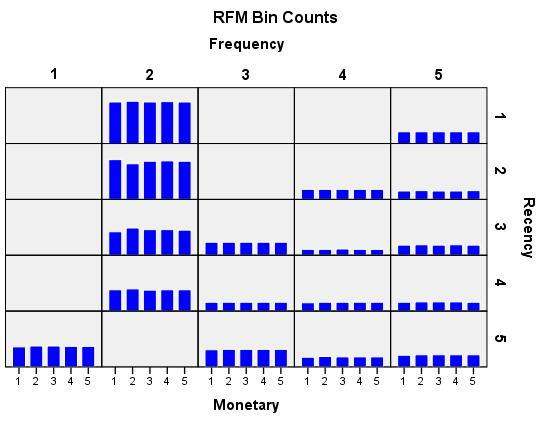


Figure 5.5: RFM Bin Counts

The RFM bin counts chart displays the data distribution for each bin. Each bar represents the number of customers that is assigned to that dimension score.

From the Frequency perspective, Figure 5.1 shows most of the customers only have the score of 2 in Frequency within two months (two months of transaction data). Score of 2 in this bin counts has a transaction count of 1. This means, most customers only purchased once in 2 months. However, There are still an amount of customer (around 100+ customers) who has a frequency score of 5.

Table 5.1: Count table for Frequency score

|  |  |
| --- | --- |
| Frequency score | Counts |
| 1 | 246 |
| 2 | 1511 |
| 3 | 469 |
| 4 | 415 |
| 5 | 615 |

Next, from the Recency perspective, the data is fairly distributed, with most of the customers having score of 3 to 5.

On the other side, from the Monetary Frequency, the data is also fairly distributed, with most of the score of roughly the same height, which means the data is following the normal distribution.

RunningMan can target customer of low frequency score by marketing techniques such as offering them discount vouchers or monthly package for delivery in order to increase their number of purchase with the company. One of reason the customer did not purchase frequently from the food delivery service is due to high spending. One delivery meal usually will cost more than RM10.

Although the data has fairly even distribution, with all (or most) bars of roughly the same height, a certain amount of variance should be expected when using the default binning method that assigns tied values to the same bin.

Extreme fluctuations in bin distribution and/or many empty bins may indicate that we should try another binning method (fewer bins and/or random assignment of ties) or reconsider the suitability of RFM analysis.

1. **RFM Heat Map**

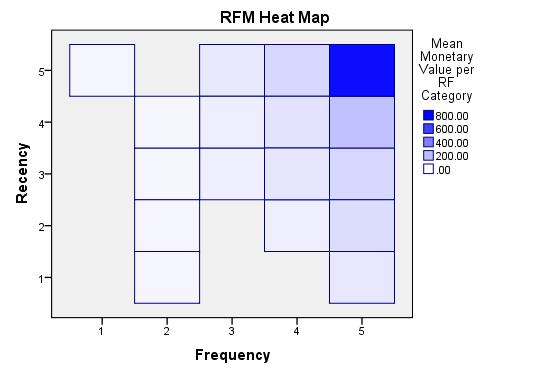


Figure 5.6: RFM Heat Map

The heat map shows the average monetary value for recency score and frequency score. Darker areas indicate a higher average monetary value. In other words, customers with recency and frequency scores in the darker areas tend to spend more on average than those with recency and frequency scores in the lighter areas.

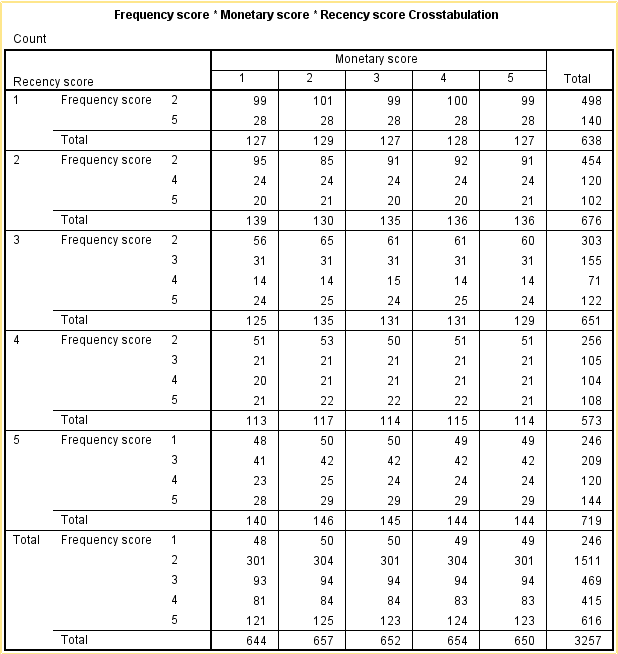
Figure 5.6 shows that customer with Recency and Frequency score of 5 have an average spending of 761.78, while customer with Recency score of 4 and Frequency score of 5 have an average spending of 193.62.

Table 5.2 Average Monetary Value for each score category

|  |  |  |
| --- | --- | --- |
| Frequency | Recency | Average Monetary value |
| 1 | 5 | 31.35 |
| 2 | 1 | 29.81 |
| 2 | 30.72 |
| 3 | 27.64 |
| 4 | 28.61 |
| 3 | 3 | 51.53 |
| 4 | 53.92 |
| 5 | 72.43 |
| 4 | 2 | 54.58 |
| 3 | 77.91 |
| 4 | 87.83 |
| 5 | 127.02 |
| 5 | 1 | 76.21 |
| 2 | 109.99 |
| 3 | 127.05 |
| 4 | 193.62 |
| 5 | 761.78 |

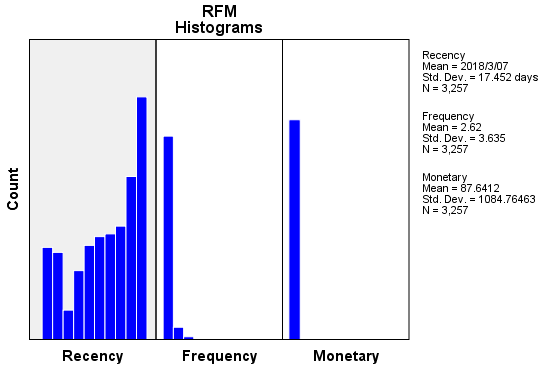
One of the group that will bring most benefit to the RunningMan is customer with recency and frequency score of 5, which has an average monetary value of 761.78. The next group of customer is customer with frequency score of 4 and recency score of 5, which has an average monetary value of 127.02. These 2 groups of customers will benefit the company in terms of profitability.

1. **Frequency score \* Monetary score \* Recency score Crosstabulation**



The table of bin counts displays the bin distribution for the selected binning method. Each cell represents the number of customers that will be assigned each combined RFM score. Although we typically want a fairly even distribution, with all (or most) cells containing a similar number of customers, a certain amount of variance should be expected when using the default binning method that assigns tied values to the same bin. Extreme fluctuations in cell counts and/or many cells with a count of 0 may indicate that we should try another binning method (fewer bins and/or random assignment of ties) or reconsider the suitability of RFM analysis.

1. **RFM Histogram**

****

The histograms show the relative distribution of values for the three fields used to calculate recency, frequency, and monetary scores. It is not unusual for these histograms to indicate somewhat skewed distributions rather than a normal or symmetrical distribution.   
  
The horizontal axis of each histogram is always ordered from low values on the left to high values on the right. With recency, however, the interpretation of the chart depends on the type of recency measure: date or time interval. For dates, the bars on the left represent values further in the past (a less recent date has a lower value than a more recent date). For time intervals, the bars on the left represent more recent values (the smaller the time interval, the more recent the transaction).

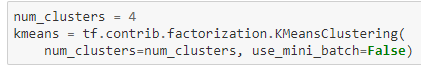
1. **RFM Scatterplots**

****

These scatterplots show the relationships between the three variables used to calculate recency, frequency, and monetary scores.   
  
It's common to see noticeable linear groupings of points on the frequency scale, since frequency often represents a relatively small range of discrete values. For example, if the total number of transactions doesn't exceed 15, then there are only 15 possible frequency values (unless you count fractional transactions), whereas there could be hundreds of possible recency values and thousands of monetary values.   
  
The interpretation of the recency axis depends on the type of recency measure: date or time interval. For dates, points closer to the origin represent dates further in the past. For time intervals, points closer to the origin represent more recent values.

1. **Clustering of RFM score**

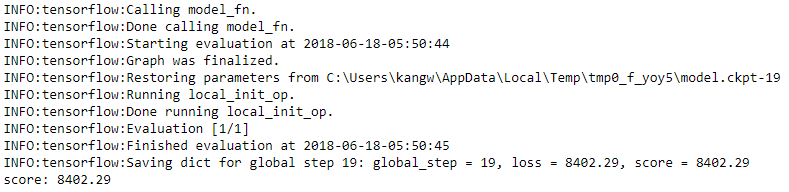
K-means clustering is used to segment the customer based on the RFM score assigned to them. Tensorflow library, ***tensorflow.contrib.factorization.KMeansClustering*** is used in this project. RFM parameters are included in clustering. After consideration, The number of cluster is assigned to 4.



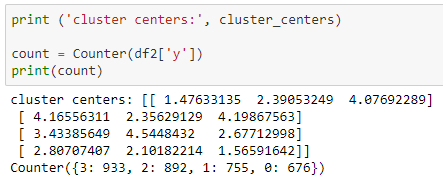
There are several reason the number of cluster is assigned to 4.

1. Size of business
   * Splitting the customers into 4 groups are just right for RunningMan to focus on with their limited resources. The number of customer are also just enough to split into 4 group with each group contain almost the same number of customers.
2. Interval of score
   * Since we only have 5 ranking score for Recency, Frequency and Monetary, it would be perfect to split the customer into 4 group as there’s only limited combination from the RFM score.

The model is then being trained using the dataset for 10 iterations. The model has a score of 8402.29.



The input points are mapped to their clusters and tabulated.



Based on the cluster centers, we can observe the following:

* The first group of customer (676 of them) has very low recency score, medium frequency score and high monetary score. This means some customer only order food for delivery when they needed to, with each order having high purchasing amount. For example, office staff order food for a group of staff or an event providing food for its participants.
* The second group of customers (755 of them) has high recency score, medium frequency score, and high monetary score. This group of customers are their frequent customer. They purchase on a basis. This group of customers can be students, workers, and so on.
* The third group of customers (892 of them) has slightly higher recency score, high frequency score, but low monetary score. This group of customers doesn’t spend much on each order but does order frequently. This might be students who just want a quick bite to fill their stomach.
* The forth group of customers (933 of them) has low recency score, low frequency score, and low monetary score. This group of customers most probably are students who only spend when there’s a promotion going on, for example, based on the dataset, one of the most popular food sold during February to March is Braised Pork Rice, and during this period they might have done promotion with the food stall to promote their business.

These cluster centers are then plotted onto a 3D graph, using Bokeh library.

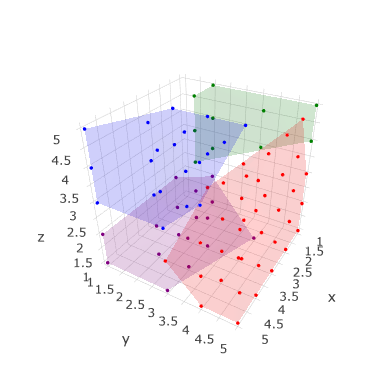


Figure 5.7: RFM Cluster in 3D

Figure 5.7 shows that most of the customer are in the red cluster. The label are: x-axis is Recency score, y-axis is Frequency score and z-axis is monetary score. The red cluster has RFM score of 2.81, 2.10, 1.57. We can assume that most of the customer in this cluster has a low RFM score and are the group of customer RunningMan should pay attention to.

1. **Cluster 3 visualized in map**

Cluster 3 which has the most number of customer, is plotted on map.



Figure 5.8: Cluster 3 of RFM plotted on map

Figure 5.8 shows that most of the customers are located in areas such as Wangsa Maju, Bandar Mahkota Cheras. These areas are populated with students with low spending. Therefore we can assume that RunningMan mostly target students with low spending power. Thus, in order to increase revenue in these areas, the company has to come up a service that is affordable and good to all the students in these areas. Also, the company can allocate more runner in these areas to provide a better services to their customer in these areas.

1. Relating Churner with RFM

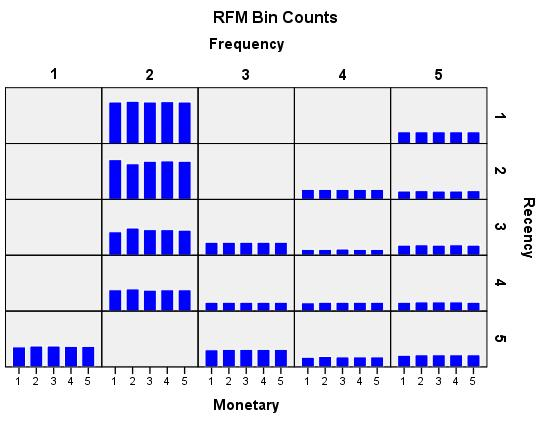
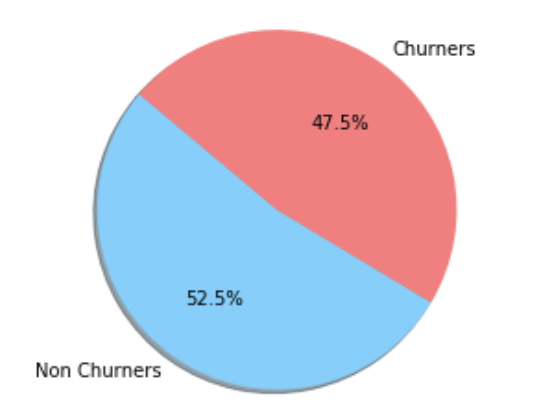


Figure 5.9: Relating Churner with RFM

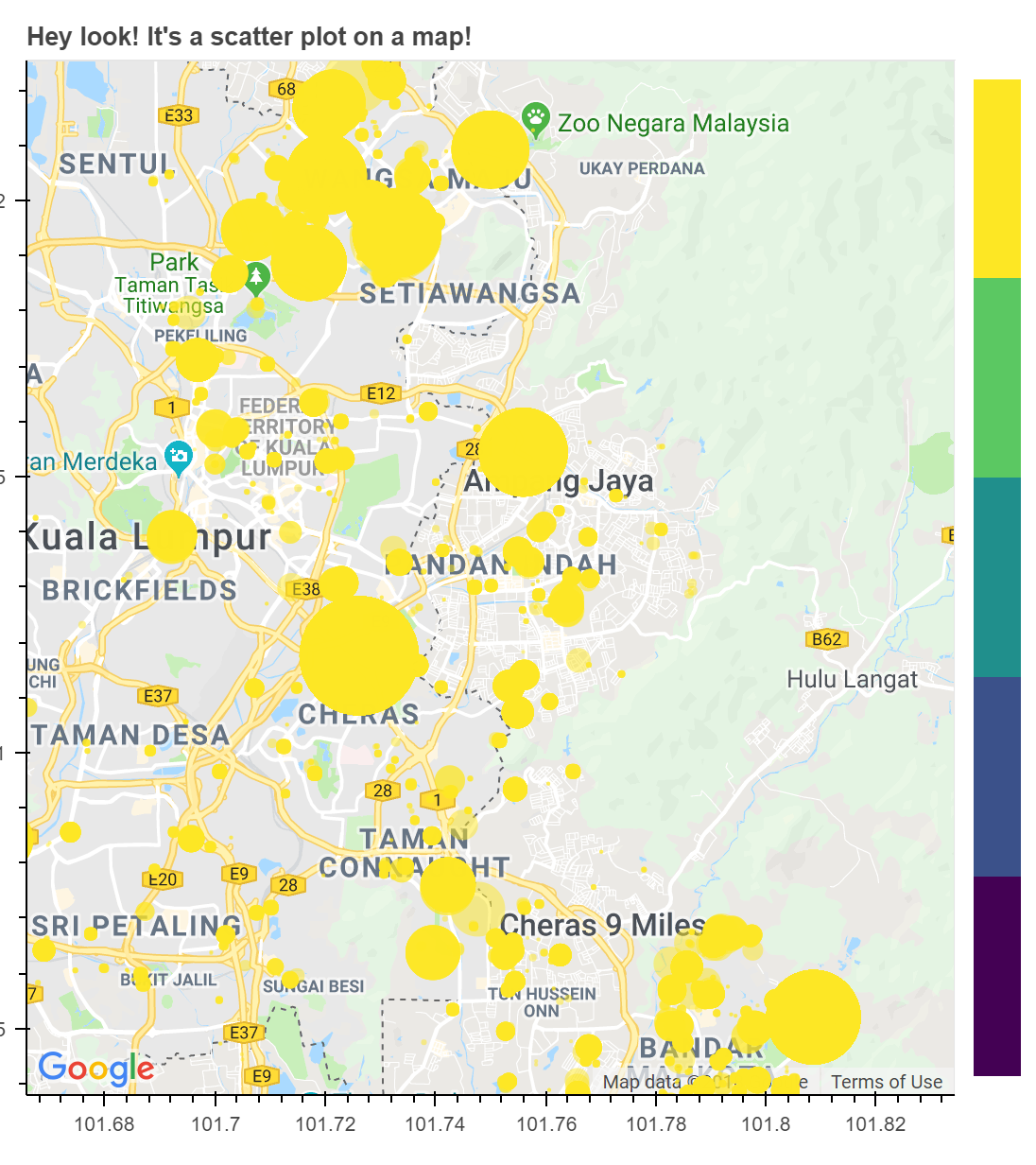


Figure 5.10: Relating cluster with churner

Figure 5.9 shows that 47.5% of the users are churners can be verified by relating it to RFM. Most of the users churn after purchasing only once. This causes their Frequency score to be low (2).

Figure 5.10 shows that the location of churner matches the RFM cluster 3 which has a RFM score of RFM score of - 2.81, 2.10, 1.57 and plotted on map in compare. The scatter plot in both maps are almost identical. This proves that customer churns because they either, could not afford to deliver food on a daily basis.

## Chapter Summary and Evaluation

* Several analysis are carried out to provide business insights to the company.
* RFM score binning shows customer are having low purchase count.
* RFM heat map shows customer tends to spend more when their score are higher.
* RFM clustering shows most of the customer has low RFM score.

Plotting cluster of transaction on map shows customer with low RFM score are students.

Chapter 6

Discussions and Conclusion

# Discussions and Conclusion

This chapter discuss the results to see whether the end product achieves the project objectives. This chapter also provides recommendation and discusses future improvements

## Conclusion

Consumer analytics is at the epicentre of a big data revolution. Technology helps capture rich and plentiful data on consumer behaviour in real time. Food delivery service provider, RunningMan has provided us dataset to explore business opportunities that can be concluded from analysing the dataset. This project aims to use machine learning to build a more efficient and simpler consumer analytics model for small and medium-sized enterprise that does not have much resources to carry out data analysis. RFM scoring is a simple and efficient model that can be used to analyse customers purchasing behaviour. The final model is able to accomplish all the objectives and functional requirements that stated in previous chapter.

## Limitation and Future Improvements

There are a few limitations and obstacles encountered in this project. First, using scikit-learn library, the data could not be constantly clustered into the same group. The cluster centres of each group are fluctuating. Therefore, TensorFlow library is used instead. TensorFlow library is able to cluster the data into the same group every time the code is executed.

Furthermore, since we are using k-means clustering to cluster our RFM scoring, there is no validation that can be done to validate the accuracy of the machine learning algorithm, because this is an unsupervised learning method, and we need to observe the result generated from the output and produce business conclusion from the results.

In future, more unsupervised methods can be evaluated to cluster RFM scoring. There are more unsupervised methods than k-means clustering such as mixture models, hierarchical clustering can be tested out.

Also, to find out the best binning count for the dataset. Binning of customer into 5 bins for every score might not produce the best accuracy. The number of bins per score should be searched using optimization techniques. For example, Monetary score can have 8 bins in order to further cluster customer into detailed groups while Recency and Frequency score can remain to be 5.

# References

1. Erevelles, S., Fukawa, N. and Swayne, L., 2016. Big Data consumer analytics and the transformation of marketing. Journal of business research, 69(2), pp.897–904. Available at: <http://dx.doi.org/10.1016/j.jbusres.2015.07.001>.
2. Hartmann, P.M., Zaki, M. and Feldmann, N., 2014. Big data for big business? A taxonomy of data-driven business models used by start-up firms. -Driven Business Models …. Available at: <http://www.nsuchaud.fr/wp-content/uploads/2014/08/Big-Data-for-Big-Business-A-Taxonomy-of-Data-driven-Business-Models-used-by-Start-up-Firm.pdf>.
3. Kozlenkova, I.V., Samaha, S.A. and Palmatier, R.W., 2014. Resource-based theory in marketing. Journal of the Academy of Marketing Science, 42(1), pp.1–21. Available at: <http://link.springer.com/10.1007/s11747-013-0336-7>.
4. Raorane, A. and Kulkarni, R.V., 2011. Data Mining Techniques: A Source for Consumer Behavior Analysis. arXiv [cs.DB]. Available at: <http://arxiv.org/abs/1109.1202>.
5. Satell, G., 2014, 5 Things Managers Should Know About The Big Data Economy [Online]. Available at: https://www.forbes.com/sites/gregsatell/2014/01/26/5-things-managers-should-know-about-the-big-data-economy/ [Accessed: 15 November 2017].
6. Sen, D., Ozturk, M. and Vayvay, O., 2016. An Overview of Big Data for Growth in SMEs. Procedia - Social and Behavioral Sciences. Available at: <http://dx.doi.org/10.1016/j.sbspro.2016.11.011>.
7. Hsieh, N.C. and Chu, K.C., 2009. Enhancing consumer behavior analysis by data mining techniques. International Journal of Information and Management Sciences. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.579.9503&rep=rep1&type=pdf>.
8. IDC, 2014, The digital universe of opportunities: Rich data & the increasing value of the internet of things. EMC2 [Online]. Available at: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf [Accessed: 13 November 2017].
9. Kennedy, by J., 2011, Amount of data in 2011 equal to 57.5bn 32GB Apple iPads - Enterprise | siliconrepublic.com - Ireland’s Technology News Service [Online]. Available at: https://www.siliconrepublic.com/enterprise/amount-of-data-in-2011-equal-to-57-5bn-32gb-apple-ipads [Accessed: 13 November 2017].
10. Kozlenkova, I.V., Samaha, S.A. and Palmatier, R.W., 2014. Resource-based theory in marketing. Journal of the Academy of Marketing Science, 42(1), pp.1–21.
11. Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. Journal of machine learning research: JMLR, 12(Oct), pp.2825–2830.
12. Ray, S. et al., 2017, Understanding Support Vector Machine algorithm from examples (along with code) [Online]. Available at: https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code [Accessed: 13 January 2018].
13. Salzberg, S. et al., 1995. Decision Trees for Automated Identification of Cosmic-Ray Hits in Hubble Space Images. Publications of the Astronomical Society of the Pacific, 107, pp.279–288.
14. Satell, G., 2014, 5 Things Managers Should Know About The Big Data Economy [Online]. Available at: https://www.forbes.com/sites/gregsatell/2014/01/26/5-things-managers-should-know-about-the-big-data-economy/ [Accessed: 15 November 2017].
15. Sciencing, 2017, The Disadvantages of Linear Regression [Online]. Available at: http://sciencing.com/disadvantages-linear-regression-8562780.html [Accessed: 13 January 2018].
16. Specht, D.F., 1990. Probabilistic neural networks. Neural networks: the official journal of the International Neural Network Society, 3(1), pp.109–118.
17. Xu, Z., Frankwick, G.L. and Ramirez, E., 2016. Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. Journal of business research. Available at: <http://dx.doi.org/10.1016/j.jbusres.2015.10.017>.
18. Yu, W. et al., 2010. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making, 10(1), p.16.

# Bibliography

1. Sismeiro, C. and Bucklin, R.E., 2004. Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach. Journal of Marketing Research, 41(3), pp.306–323.
2. (Stroie), L.M.B., 2014. Predicting Consumer Behavior with Artificial Neural Networks. Procedia Economics and Finance, 15, pp.238–246.
3. Gupta, R. and Pathak, C., 2014. A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing. Procedia Computer Science, 36, pp.599–605.
4. Zuo, Y., Ali, A.S. and Yada, K., 2014. Consumer Purchasing Behavior Extraction Using Statistical Learning Theory. Procedia Computer Science, 35, pp.1464–1473.
5. Anon, Big data and the Internet of Things: Two sides of the same coin? [Online]. Available at: https://www.sas.com/en\_us/insights/articles/big-data/big-data-and-iot-two-sides-of-the-same-coin.html.
6. Anon, 2017, Six Ways to Create Better Customer Behavior Analytics[Online]. Available at: <https://www.datameer.com/company/datameer-blog/six-ways-create-better-customer-behavior-analytics/>.
7. Callahan, L., 2017, Top Marketers Look to Predictive Analytics as the Future of Marketing [Online]. Available at: <https://martechexec.com/article/predictive-analytics-future-marketing.html>.
8. McLellan, C., 2015, The internet of things and big data: Unlocking the power [Online]. Available at: <http://www.zdnet.com/article/the-internet-of-things-and-big-data-unlocking-the-power/>.
9. Anon, 2014, Predictive modeling, supervised machine learning, and pattern classification [Online]. Available at: <http://sebastianraschka.com/Articles/2014_intro_supervised_learning.html>.
10. Prettenhofer, P. (2017). Ordinary Least Squares in Python - DataRobot. [online] DataRobot. Available at: https://www.datarobot.com/blog/ordinary-least-squares-in-python/ [Accessed 10 Jan. 2018].
11. code), U., code), U. and Ray, S. (2017). Understanding Support Vector Machine algorithm from examples (along with code). [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/ [Accessed 10 Jan. 2018].
12. Blog.hackerearth.com. (2017). Simple Tutorial on SVM and Parameter Tuning in Python and R | HackerEarth Blog. [online] Available at: http://blog.hackerearth.com/simple-tutorial-svm-parameter-tuning-python-r [Accessed 10 Jan. 2018].
13. Sciencing.com. (2017). Cite a Website - Cite This For Me. [online] Available at: http://sciencing.com/disadvantages-linear-regression-8562780.html [Accessed 10 Jan. 2018].
14. AGCross.com. (2017). Random Forests in Python with scikit-learn. [online] Available at: http://www.agcross.com/2015/02/random-forests-in-python-with-scikit-learn/ [Accessed 10 Jan. 2018].
15. J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.
16. Christian Albright, S. and Winston, W.L., 2014. Business Analytics: Data Analysis & Decision Making, Cengage Learning.
17. Evans, J.R., 2013. Statistics, Data Analysis, and Decision Modeling: International Edition, Pearson Education Limited.
18. Zhang, A., 2017. Data Analytics: Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life, CreateSpace Independent Publishing Platform.
19. Khajvand, M., Zolfaghar, K., Ashoori, S. and Alizadeh, S., 2011. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia computer science, 3, pp.57–63. Available at: http://www.sciencedirect.com/science/article/pii/S1877050910003868.
20. Khajvand, M., Zolfaghar, K., Ashoori, S. and Alizadeh, S., 2011. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia computer science, 3, pp.57–63. Available at: http://www.sciencedirect.com/science/article/pii/S1877050910003868.
21. Christian Albright, S. and Winston, W.L., 2014. Business Analytics: Data Analysis & Decision Making, Cengage Learning.
22. Evans, J.R., 2013. Statistics, Data Analysis, and Decision Modeling: International Edition, Pearson Education Limited.
23. Zhang, A., 2017. Data Analytics: Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life, CreateSpace Independent Publishing Platform.
24. Khajvand, M., Zolfaghar, K., Ashoori, S. and Alizadeh, S., 2011. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia computer science, 3, pp.57–63. Available at: http://www.sciencedirect.com/science/article/pii/S1877050910003868.
25. Trevino, A., Introduction to K-means Clustering [Online]. Available at: https://www.datascience.com/blog/k-means-clustering [Accessed: 12 April 2018].
26. Anon, Clustering - K-means [Online]. Available at: http://home.deib.polimi.it/matteucc/Clustering/tutorial\_html/kmeans.html [Accessed: 30 July 2018a].
27. Anon, 2016, Model evaluation, model selection, and algorithm selection in machine learning [Online]. Available at: https://sebastianraschka.com/blog/2016/model-evaluation-selection-part1.html [Accessed: 30 July 2018].
28. Anon, RFM Analysis Tutorial | Kaggle [Online]. Available at: https://www.kaggle.com/regivm/rfm-analysis-tutorial [Accessed: 30 July 2018b].
29. Brownlee, J., 2018, A Gentle Introduction to Data Visualization Methods in Python [Online]. Available at: https://machinelearningmastery.com/data-visualization-methods-in-python/ [Accessed: 30 July 2018].
30. Downard, I., How to plot data on maps in Jupyter using Matplotlib, Plotly, and Bokeh [Online]. Available at: http://www.bigendiandata.com/2017-06-27-Mapping\_in\_Jupyter/ [Accessed: 30 July 2018].
31. Kovalev, S., 2016, Using k-means Clustering in TensorFlow | Altoros [Online]. Available at: https://blog.altoros.com/using-k-means-clustering-in-tensorflow.html [Accessed: 30 July 2018].
32. VanderPlas, J., In Depth: k-Means Clustering | Python Data Science Handbook [Online]. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html [Accessed: 30 July 2018].
33. Anon, Install TensorFlow | TensorFlow [Online]. Available at: https://www.tensorflow.org/install/ [Accessed: 31 July 2018a].
34. Anon, User guide | Anaconda: Documentation [Online]. Available at: https://docs.anaconda.com/anaconda-cloud/user-guide/ [Accessed: 31 July 2018b].
35. Wikipedia contributors, 2018, Unsupervised learning [Online]. Available at: https://en.wikipedia.org/w/index.php?title=Unsupervised\_learning&oldid=843833133 [Accessed: 31 July 2018].

# Appendices

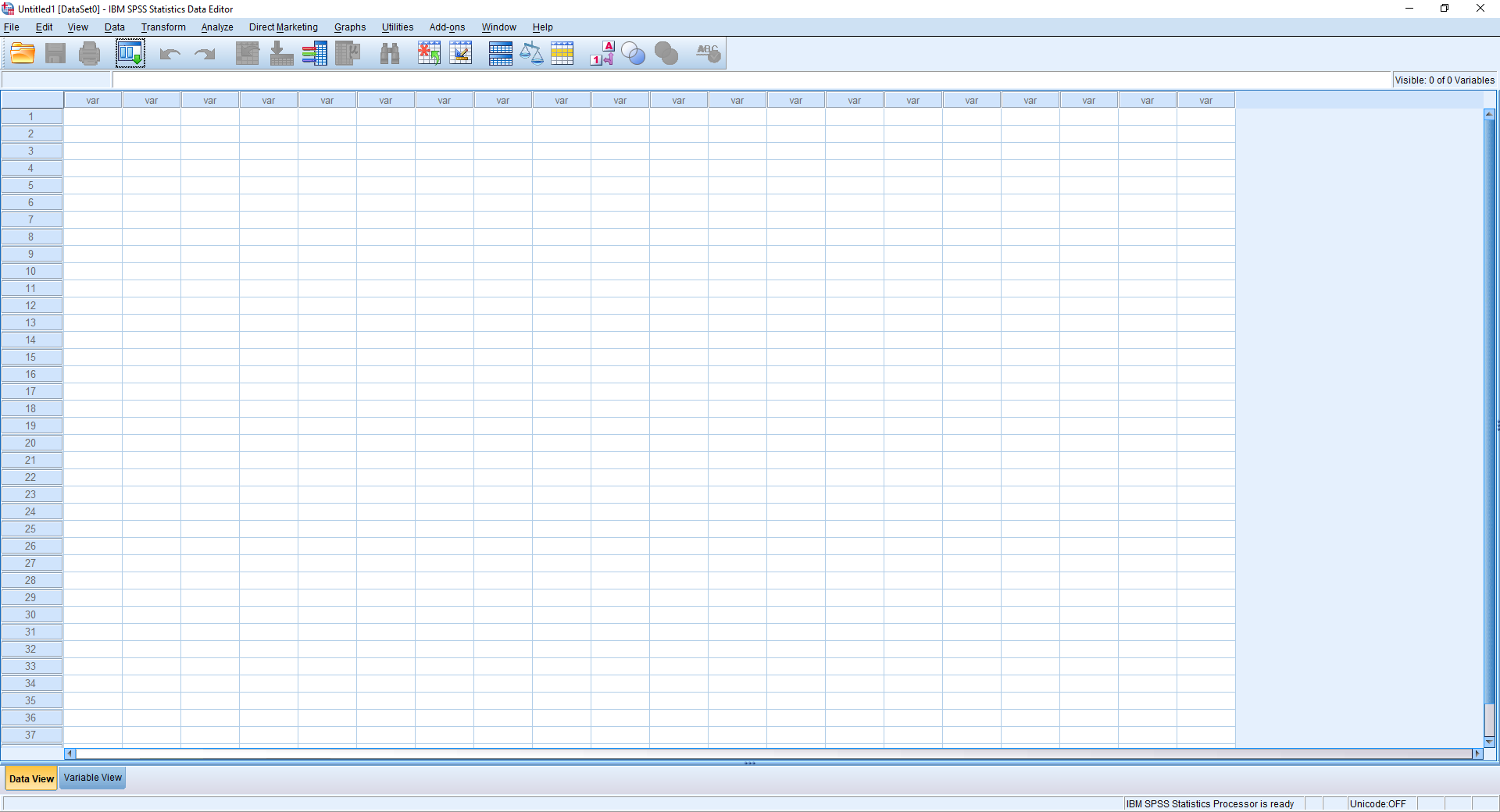
## Developer Guide

### Software used

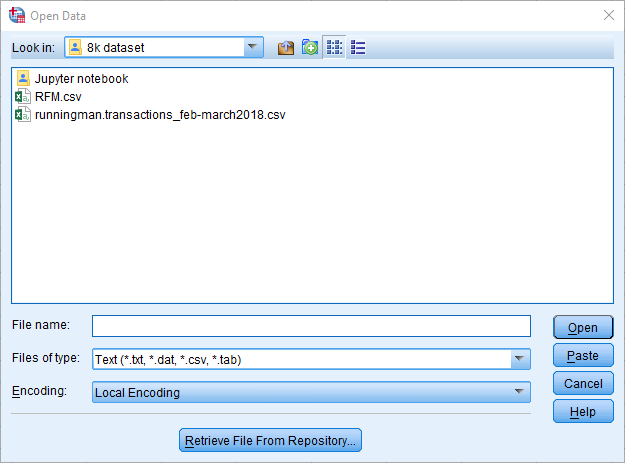
1. Anaconda
   1. Anaconda can be downloaded here: <https://anaconda.org/>
   2. Tensorflow environment is installed. For more information on how to create tensorflow environment, please refer to: <https://www.tensorflow.org/install/>
   3. Library used: pandas, numpy, plotly, matplotlib, bokeh
   4. Jupyter notebook has to be installed under the tensorflow environment.
   5. All the code will be ran in Jupyter notebook in tensorflow environment.
2. IBM SPSS Statistics 22
   1. SPSS is used for calculating RFM score.
   2. RFM score can be accessed under Direct Marketing menu.

## User Manual

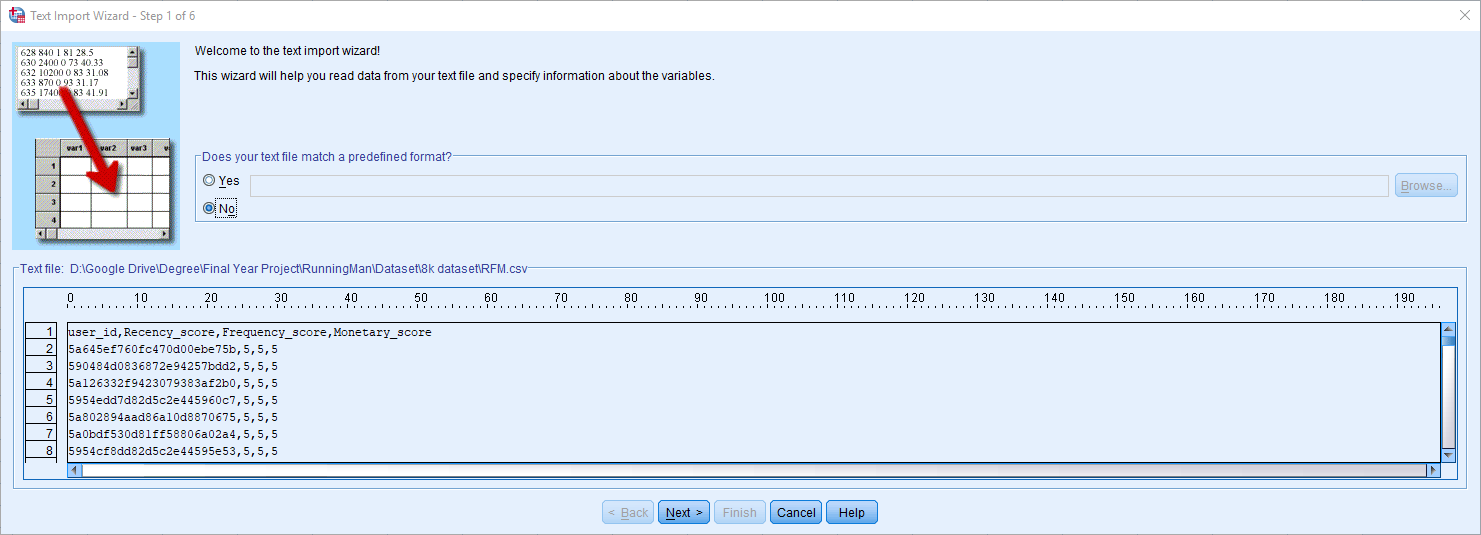
1. Install the required software as stated in Developer Guide - software used.
2. Start IBM SPSS Statistics 22

****

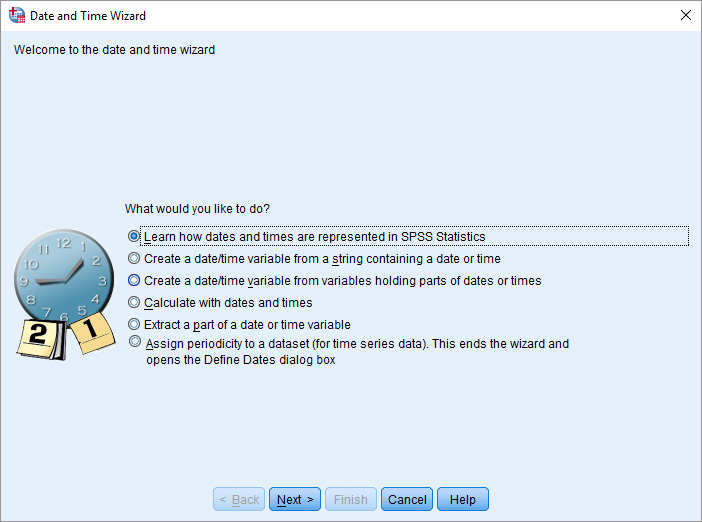
1. Go to File >> Open >> Data, under files of type, select Text (\*.txt, \*.dat, \*.csv, \*.tab), and direct to your dataset and open.



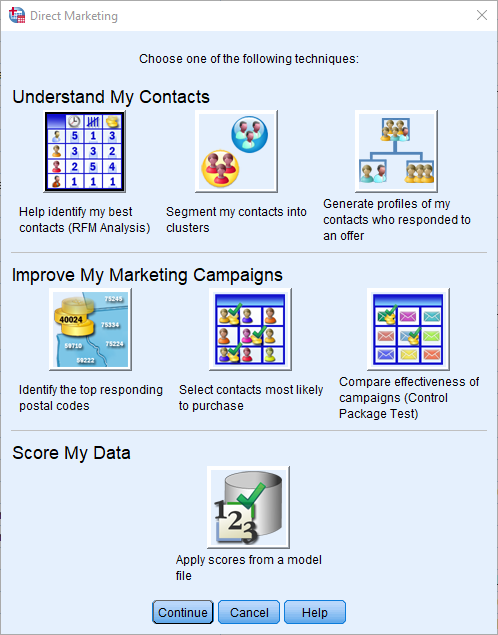
1. A Text Import Wizard will appear. Complete the steps until step 4, select Tab as the delimiters between variables. Complete the rest of the steps.



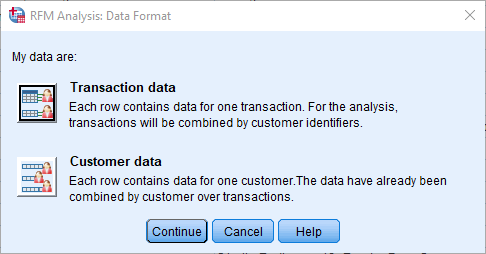
1. The dataset’s transaction date has to be formatted, so that it can be passed into RFM scoring calculation by going to Transform >> Date and Time Wizard.



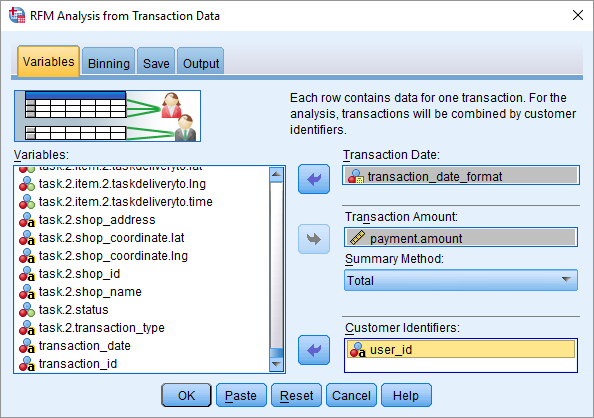
1. After formatting, to calculate RFM score, go to Direct Marketing >> Choose Technique >> Understand My Contacts >> Help identify my best contacts (RFM Analysis) and continue.



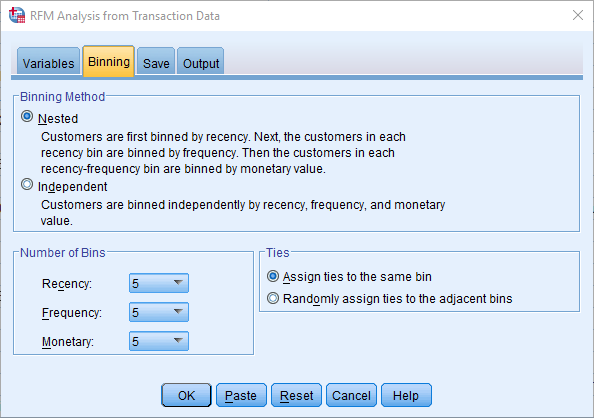
1. The data in this project is transaction data, therefore we will select transaction data.

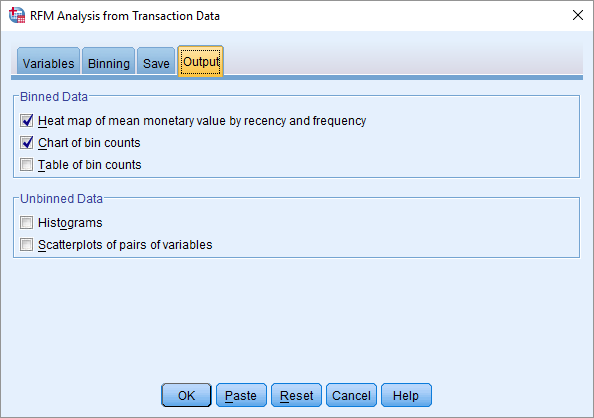


1. Drag the corresponding column in the dataset into the box.

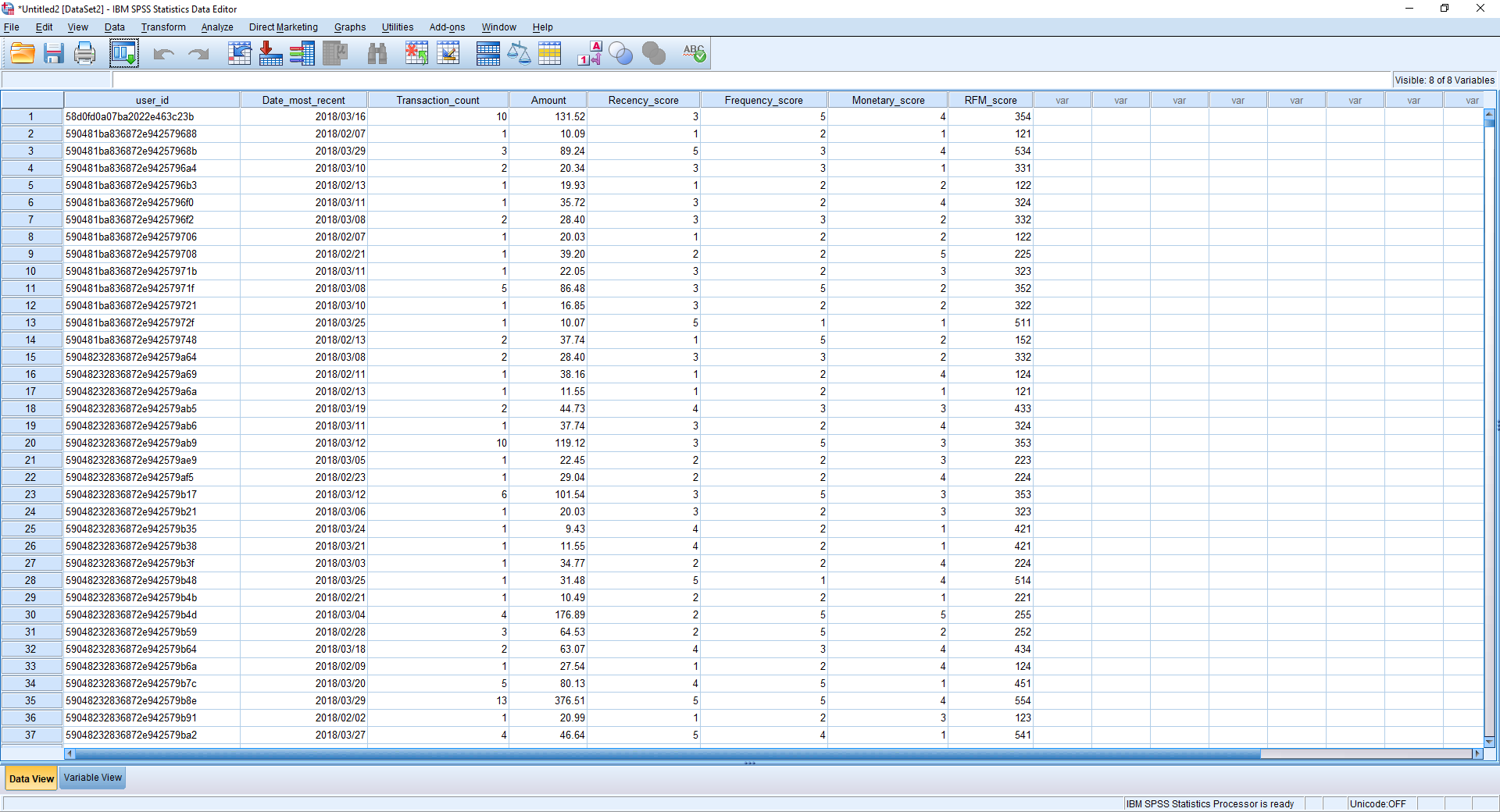


1. You can have more customization under binning, to select how many binning you would like to group, and also output more information like RFM histogram and RFM scatterplot.

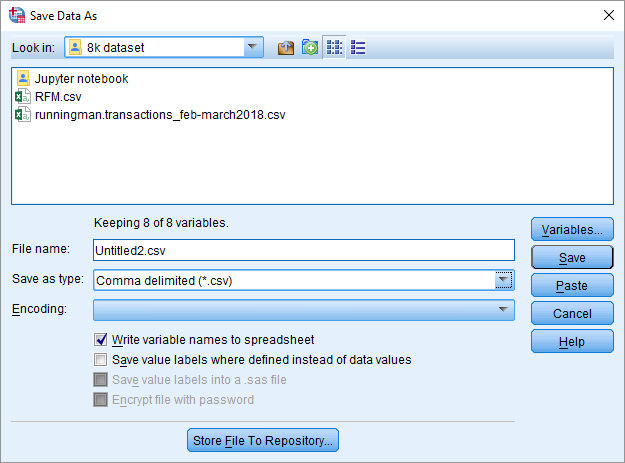




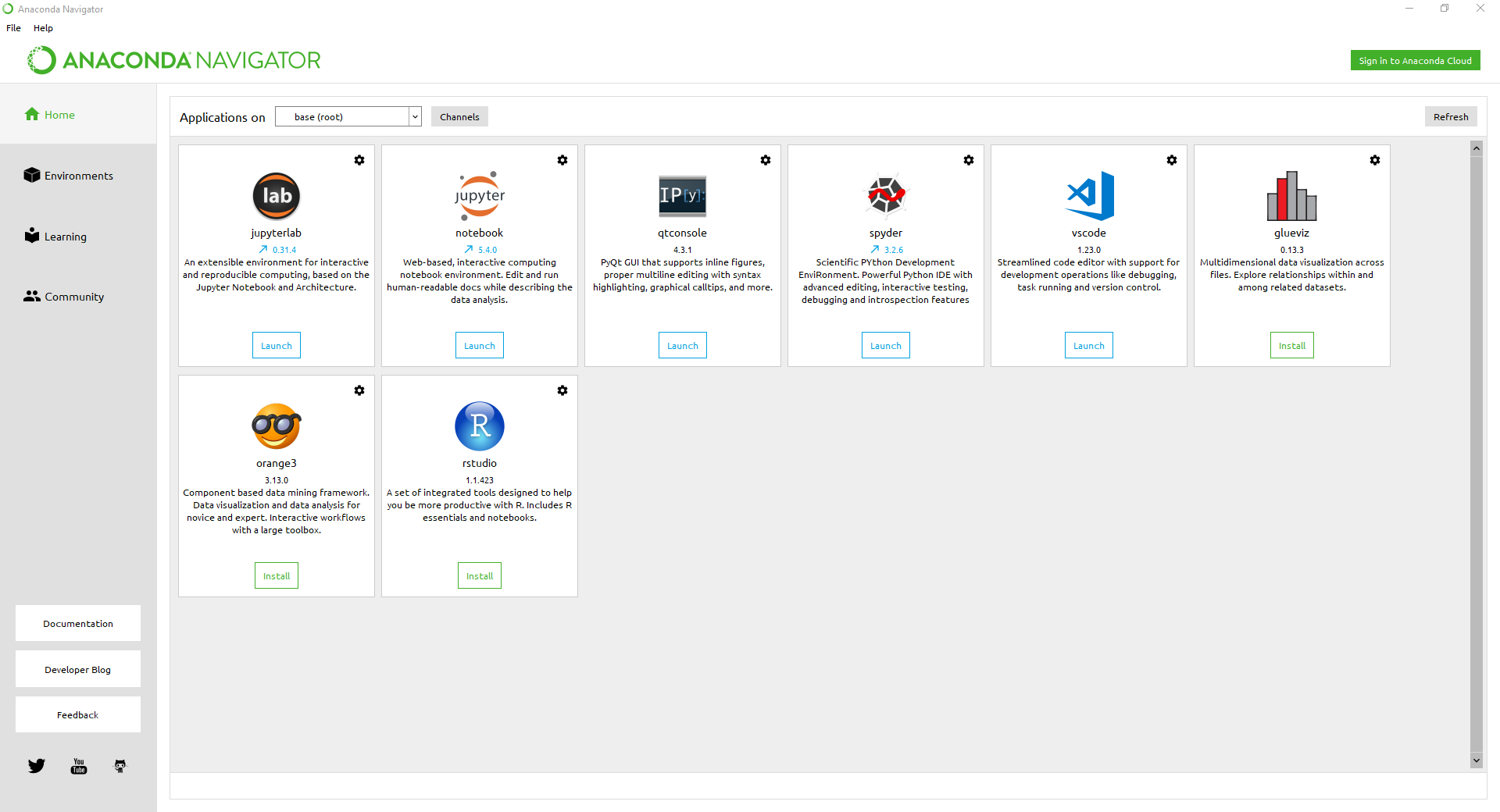
1. Press OK, and a new set of Dataset with RFM score will be produced.



1. Save the dataset as csv file.



1. Start Anaconda.



1. Under Applications on tab, change from base (root) to tensorflow, and start Jupyter notebook.
2. Open the notebooks provided and run the code.
3. Observe the result generated by the code.