****

**Data Mining Approach in Sales and Marketing Department in Retail Industry**

**By**

**HAZIM HASAN BAJAMAL**

**TP043071**

**UC3F2111CS(DA)**

A report submitted in partial fulfilment of the requirements of Asia Pacific University of Technology and Innovation for the degree of

**B.Sc. (Hons) Computer Science with Specialism in Data Analytics**

Supervised by Mr. MOHAMMAD NAMAZEE MOHD NIZAM

2nd Marker: Dr. PREETHI SUBRAMANIAN

July-2022

Acknowledgement

First and foremost, I’d like to thank to my parents for always believe me, for their constant support, in term of emotionally and mentally also act as ice breaker to help me not to be stressed out and overwhelmed with the project although we have been separated for a long time due to the pandemic.

Also, I would like to thank my supervisor, Mr. Mohammad Namazee Mohd Nizam for the suggestions during the idea generation that helped me to accomplish this project.

I would like to thank my friends for their guidance, support and motivations that helped me to accomplish this final year project.

Lastly, I’d also would like to thank APU for giving me the opportunity to test myself and made me prepared for the real-world life that I am going to face upon the completion of the study.

Table of Contents

[CHAPTER 1: Introduction to The Study 8](#_Toc109240873)

[1.1 Background to the project 8](#_Toc109240874)

[1.2 Problem statements 8](#_Toc109240875)

[1.3 Rationale 10](#_Toc109240876)

[1.4 Potential benefits 10](#_Toc109240877)

[1.4. 1 Tangible benefits 10](#_Toc109240878)

[1.4. 2 Intangible benefits 10](#_Toc109240879)

[1.5 Target users 10](#_Toc109240880)

[1.6 Scope and Objectives 11](#_Toc109240881)

[1.6. 1 Aim 11](#_Toc109240882)

[1.6. 2 Objectives 11](#_Toc109240883)

[1.6. 3 Deliverables 11](#_Toc109240884)

[1.6. 4 Nature of Challenges 11](#_Toc109240885)

[1.7 Overview of this report 11](#_Toc109240886)

[1.8 Project Plan 13](#_Toc109240887)

[CHAPTER 2: Literature Review 14](#_Toc109240888)

[2.1: Introduction 14](#_Toc109240889)

[2.2: Domain research 14](#_Toc109240890)

[2.2.1 Data Mining 14](#_Toc109240891)

[2.2.3. Marketing 15](#_Toc109240892)

[2.2.4 Sales 15](#_Toc109240893)

[2.2.5 Clustering 16](#_Toc109240894)

[2.2.6 RFM Analysis 17](#_Toc109240895)

[2.3 Similar System(s) 18](#_Toc109240896)

[CHAPTER 3: Technical Research 19](#_Toc109240897)

[3.1: Programming language chosen 19](#_Toc109240898)

[3.1.1 Introduction 19](#_Toc109240899)

[3.1.2 Python 20](#_Toc109240900)

[3.2: IDE (Interactive Development Environment) 25](#_Toc109240901)

[3.3: libraries chosen / Tools chosen 25](#_Toc109240902)

[NumPy 25](#_Toc109240903)

[Pandas 25](#_Toc109240904)

[Matplotlib 26](#_Toc109240905)

[Seaborn 26](#_Toc109240906)

[Datetime 26](#_Toc109240907)

[Scikit-Learn 26](#_Toc109240908)

[3.4 Database Management System 26](#_Toc109240909)

[3.5: Operating System chosen 27](#_Toc109240910)

[3.7: Web browser chosen - (this section is optional) 27](#_Toc109240911)

[3.8: Summary 27](#_Toc109240912)

[CHAPTER 4: Methodology 28](#_Toc109240913)

[4.1 Introduction 28](#_Toc109240914)

[4.2 Methods 28](#_Toc109240915)

[4.2.1 Business understanding 29](#_Toc109240916)

[4.2.2 Data Understanding 30](#_Toc109240917)

[4.2.3 Data Preparation 32](#_Toc109240918)

[4.2.4 Modelling 34](#_Toc109240919)

[4.2.5 Evaluation 35](#_Toc109240920)

[4.2.6 Deployment: 36](#_Toc109240921)

[4.3 Summary 36](#_Toc109240922)

[CHAPTER 5: Data Analysis 37](#_Toc109240923)

[5.1 Introduction 37](#_Toc109240924)

[5.2 Initial data exploration 38](#_Toc109240925)

[5.2.1 Import Dataset 38](#_Toc109240926)

[5.2.2 Data Types 39](#_Toc109240927)

[5.2.3 Descriptive Statistic 40](#_Toc109240928)

[5.2.4 Missing Value 42](#_Toc109240929)

[5.3 Data Cleansing 43](#_Toc109240930)

[5.3.1 Drop Missing Values 43](#_Toc109240931)

[5.3.2 Data Reduction 43](#_Toc109240932)

[5.3.3 Data Transformation 45](#_Toc109240933)

[5.4 Visualization 55](#_Toc109240934)

[5.4.1 Boxplot RFM variables 55](#_Toc109240935)

[5.4.2 Country 57](#_Toc109240936)

[5.4.3 CustomerID 58](#_Toc109240937)

[5.4.4 Stock Code 59](#_Toc109240938)

[5.4.5 Product Sold 59](#_Toc109240939)

[5.4.6 Historical Demand 60](#_Toc109240940)

[5.5 Hypothesis 60](#_Toc109240941)

[5.6 Reports and Dashboard 60](#_Toc109240942)

[5.7 Modelling 61](#_Toc109240943)

[5.8 Summary 64](#_Toc109240944)

[CHAPTER 6: Results and Discussion 65](#_Toc109240945)

[6.1 Introduction 65](#_Toc109240946)

[6.2 Results and discussion 65](#_Toc109240947)

[CHAPTER 7: Conclusions and Reflections 67](#_Toc109240948)

[REFERENCES 68](#_Toc109240949)

[APPENDICES 72](#_Toc109240950)

[Figure 1: CRISP-DM Life Cycle 29](#_Toc109240951)

[Figure 2: Importing Dataset and Required Libraries 38](#_Toc109240952)

[Figure 3: First 5 Rows of the Dataset 39](#_Toc109240953)

[Figure 4: Data Types 39](#_Toc109240954)

[Figure 5: Descriptive Statistics 40](#_Toc109240955)

[Figure 6: Checking Missing Values 42](#_Toc109240956)

[Figure 7: Removing Missing Values 43](#_Toc109240957)

[Figure 8: Remove Negative Value in Quantity and Unit Price 44](#_Toc109240958)

[Figure 9: Selecting Country to Only United Kingdom 45](#_Toc109240959)

[Figure 10: Change CustomerID Data Type into Object 45](#_Toc109240960)

[Figure 11: Transform Date Variable 46](#_Toc109240961)

[Figure 12: New Variable 'daysAgo' 47](#_Toc109240962)

[Figure 13: Create New Variable 'Monetary' 47](#_Toc109240963)

[Figure 14: Create Recency Data Frame 48](#_Toc109240964)

[Figure 15: Create Frequency Data Frame 49](#_Toc109240965)

[Figure 16: Create Monetary Data Frame 50](#_Toc109240966)

[Figure 17: Merging Multiple Data Frames 51](#_Toc109240967)

[Figure 18: For Loop to Remove Outliers 52](#_Toc109240968)

[Figure 19: Validating Outliers has been Removed 53](#_Toc109240969)

[Figure 20: Variables Scaled 54](#_Toc109240970)

[Figure 21: Boxplot with Outliers 55](#_Toc109240971)

[Figure 22: Boxplot without Outliers 55](#_Toc109240972)

[Figure 23: Top 10 Country 57](#_Toc109240973)

[Figure 24: Top 10 Customer based on Number of Transactions 58](#_Toc109240974)

[Figure 25: Top 10 Used Stock Code 59](#_Toc109240975)

[Figure 26: Top 10 Product Sold 59](#_Toc109240976)

[Figure 27: Historical Demand 60](#_Toc109240977)

[Figure 28: Initialize K-Means Model 61](#_Toc109240978)

[Figure 29: Elbow Method 61](#_Toc109240979)

[Figure 30: Visualization Elbow Method 62](#_Toc109240980)

[Figure 31: Silhouette Analysis 62](#_Toc109240981)

[Figure 32: Silhouette Score Output 63](#_Toc109240982)

[Figure 33: Declaring Final K-Means Model 63](#_Toc109240983)

[Figure 34: Adding Cluster Label to the Data Frame 64](#_Toc109240984)

[Figure 35: Average of Recency, Frequency, and Monetary in Each Clusters 65](#_Toc109240985)

[Figure 36: Cluster Distribution 66](#_Toc109240986)

[Figure 37: First Page of Turnitin Report 72](#_Toc109240987)

[Figure 38: Second Page of Turnitin Report 73](#_Toc109240988)

[Figure 39: FYP Poster 75](file:///C:\Users\User\Desktop\FYP\MR-HAZIM%20HASAN%20BAJAMAL-TP043071-UC3F2111CS(DA)-FYPDOC.docx#_Toc109240989)

[Figure 40: Confidentiality Document 76](#_Toc109240990)

[Figure 41: Library Cataloguing details 77](#_Toc109240991)

[Figure 42: Project log 1 78](#_Toc109240992)

[Figure 43: Project Log 2 79](#_Toc109240993)

[Figure 44: Page 1 PPF 80](#_Toc109240994)

[Figure 45: Page 2 PPF 81](#_Toc109240995)

[Figure 46: Page 3 PPF 82](#_Toc109240996)

[Figure 47: Page 4 PPF 83](#_Toc109240997)

[Figure 48: Page 1 PSF 84](file:///C:\Users\User\Desktop\FYP\MR-HAZIM%20HASAN%20BAJAMAL-TP043071-UC3F2111CS(DA)-FYPDOC.docx#_Toc109240998)

[Figure 49: Page 2 PSF 85](#_Toc109240999)

[Figure 50: Page 3 PSF 86](#_Toc109241000)

[Figure 51: Page 4 PSF 87](#_Toc109241001)

[Figure 52: Page 5 PSF 88](#_Toc109241002)

[Figure 53: Page 6 PSF 89](#_Toc109241003)

[Figure 54: Page 7 PSF 90](#_Toc109241004)

[Figure 55: Page 8 PSF 92](#_Toc109241005)

[Figure 56: Page 9 PSF 93](#_Toc109241006)

[Figure 57: Ethic Form Page 1 94](#_Toc109241007)

[Figure 58: Page 2 Ethic Form 95](#_Toc109241008)

[Figure 59: Page 3 Ethic Form 96](#_Toc109241009)

[Figure 60: Page 4 Ethic Form 97](#_Toc109241010)

[Figure 61: Page 5 Ethic Form 98](#_Toc109241011)

[Figure 62: Gantt Chart for the Whole Final Year Project 100](#_Toc109241012)

# CHAPTER 1: Introduction to The Study

## 1.1 Background to the project

Data mining's popularity has risen dramatically in recent years as a result of the massive amounts of data that are being collected on a daily basis as the retail industry shifts to online sales. Small amounts of electronic evidence or commonly known as data are generated by nearly everything we do, from opening an email to taking an extended walk in the hills. When properly analysed, this evidence can provide a competitive advantage in the global market by uncovering hidden trends and overt relationships within large data sets. Existing data and historical information are utilised to build statistical model, which are then used to segment the customers based on the characteristics of consumers. Therefore, the main purpose of this paper is to study how to maximize retail sales through effective advertising strategy by gaining a deeper understanding of customers through customer segmentation which then come up with promotional recommendations.

Data mining is a word used to describe the process of extracting meaningful information from large quantities of unstructured data by analysing data patterns in large datasets using one or more software tools. In order to find trends, generate new marketing tactics, and increase income, data mining is used to acquire and analyse data.

An organization's data mining efforts should include sales and marketing analyses. In order to maximise return on investment (ROI), improve customer relationship management (CRM) and market analysis, decrease marketing campaign expenses, facilitate fraud detection and client retention, firms utilise sales and marketing analysis as a process (Kanth, 2015)

## 1.2 Problem statements

Retail sales have increased dramatically in recent years, from 2.1 percent of yearly revenue in 2010 (Belascu, 2010) to over 6 percent in 2015. (NationalRetailFederation, 2020). This fast development over the next 11 years forces retail stores to adapt to new technologies in order to stay current with market trends. Data mining's popularity has risen significantly in recent times due to current technological industry trends because of the volume of data collected every day, which, when correctly analysed, may raise an organization's sales by around 60% (techvidvan.com, 2020), giving it a competitive advantage in the global market.

It can take up to eight touchpoints in the typical marketing plan before a consumer is first interacted with (Schultz, 2020), which can lead to sales targets not being met. According to (Krogue, 2018) on Forbes.com, over 64% of sales representative activities are non-revenue activities as a result of spending excessive time on administrative tasks such as data entry and report generation (Lieberman, 2020). Thus, the primary goal of using data mining in the marketing and sales sector of a retail shop is to maximise sales, boost efficiency, and reduce communication costs.

Moreover, the availability and precision of data, particularly for small retail stores, are not clean and well organised for every algorithm (Andrey, 2020), and it takes 51% of development time to gather and label data as well as clean and arrange the data (Andrey, 2020). (CrowdFlower, 2020).

Meanwhile, 38% of participants questioned by Bazaar Voice in 2018 in the United States of America expressed a willingness to churn from a store that does not give or makes bad recommendations (Kapoort, 2020). Thus, implementing an effective data mining strategy for the sales and marketing sectors of retail would increase client retention in the firm by minimising churn, which results in revenue generation.

Recently, the COVID-19 pandemic outbreaks had a significant impact on the retail industry, which resulted in a drop of 5% in overall retail in the United Kingdom. Additionally, fashion-related retail stores were significantly impacted by the pandemic, which resulted in a drop of 35 percent in sales in March 2020 (Russell, 2020). Furthermore, a recent study conducted by Statista.com revealed that the value of retail sales significantly drops in April 2020, falling by almost 25 percent in comparison to April 2019 sales (Sabanoglu, 2022).

One of the specific business issues at hand was that several small retail store owners lacked of strategies for implementing internet marketing in order to boost derived sales (Mukherjee et al., 2016).

## 1.3 Rationale

According to the problem stated above, data mining approach will help retail leverage the sales and elevate customer satisfaction. The proposed solution would help retailer to understand the similarity of the customers, and what differentiates the type of customer from each other by implementing the fundamental marketing principles segmentation and targeting. Additionally, it enables retailers to identify crucial client segments such as churn-risk consumers and develop marketing strategies aimed at converting such customers into active customers. Thus, it will assist retailers in effectively, efficiently, and prudently allocating marketing resources, while also boosting the impact of marketing on the firm.

## 1.4 Potential benefits

### 1.4. 1 Tangible benefits

* Potentially decrease the number of churn-rate that would lead to leverage sales.
* It helps retailer to make use of the marketing budgets effectively, efficiently and wisely.
* Efficiency in allocating resource for future growth.

### 1.4. 2 Intangible benefits

* Provide characters of customers that will be segmented in proper way that potentially elevate customer retention
* It helps retailer to design marketing campaign based on customer behaviour
* Allow retailer to understand the current product demand
* Increase the overall impact of marketing campaign

## 1.5 Target users

The target user of the project would be retailer industry staff. The overall function of the program will be built for retail industry to come up with business strategy, specifically sales and marketing department

## 1.6 Scope and Objectives

### 1.6. 1 Aim

To develop data mining program for retail industry to understand customer behaviour and segmenting the customers into distinct groups.

### 1.6. 2 Objectives

* To collect, pre-process, and analyze the dataset
* To segment customer based on Recency, Frequency, and Monetary (RFM)
* To give suggestions based on the analysis result

### 1.6. 3 Deliverables

The core functions of the project that should be achieved upon the completion are :

* Allow retail to segment customers to come up with their characteristics
* Report of customer behaviour in the form of chart visualization

### 1.6. 4 Nature of Challenges

In order to build the data mining approach programs, the developer is required to have good practical and theoretical knowledge in Python programming as the chosen programming language, which makes continuous learning and research is also required to understand how to find and implement the right algorithm to provide the right solution for the problems as there is no algorithm that can be used to provide solution for any problem. Additionally, A deep understanding of sales and marketing sectors and finding the right dataset is also another challenge as there is no all-in-one dataset that can be used to solve any problem.

## 1.7 Overview of this report

There are 7 chapters are included in this final year project report, each chapter is having its own purposes to study about the subject that matters with regards to this project.

The first chapter is the brief introduction about the project, which include problem statements, aim & objectives, nature of challenges of the project, the target user also the deliverables that will be given upon completion of this project.

The research proposal will describe the context of the problem describing investigated papers and the data relevant to the subject can be found in chapter 2.

Chapter 3 focus on the technical side of the research, which include the chosen operating system by the developer, libraries that will be used during the development phase, and comparison between the chosen programming language with other similar programming languages.

Chapter 4 is focusing on how to accomplish the project by using crisp-dm methodology. Moreover, this section describes the phases of crisp-dm, also task that involve in each phases.

Chapter 5 is data analysis part in which focus on analysing the dataset, cleanse the dataset with statistical approach, visualized the dataset with appropriate chart, and model building for K-Means clustering algorithm.

Chapter 6 is where the result of chapter 5 shown and discussed. In this chapter, the analyst also provides suggestions for the retail store with regards to the information gained through the analysis.

Chapter 7 as the last chapter of this research paper is the conclusion, which is the summary of the report, which includes what was achieved, limitation of the project, and future enhancement of the project.

## 1.8 Project Plan

Project Plan for FYP Semester 1



Figure 2: Project Gantt Chart

# CHAPTER 2: Literature Review

## 2.1: Introduction

This section is mainly focusing on studying Data mining, Sales, Marketing, Clustering, RFM, and Market Basket Analysis. Without proper understanding on how the situation with regards to the project is, the outcome also may not be proper.

## 2.2: Domain research

According to Forbes, 57% of enterprise executives believe that the most important growth benefit of AI and machine learning will be improving customer experiences and support. Hence, the demand of utilizing data to maximize sales and marketing strategy is high (Columbus, 2018) .

### 2.2.1 Data Mining

The term "data mining" refers to the process of extracting hidden predictive information from massive data sets (DBs). DM employs advanced statistical analysis and modelling approaches to reveal patterns and linkages buried in organisational databases through the automated discovery of implicit knowledge. Over the last four decades, the tools and methodologies for processing structured data have evolved from databases through data warehousing (DW) to data mining (DM). Data warehousing systems have evolved into mission-critical components of businesses. DM may extract even more value from these massive data warehouses.

The approaches to data mining are many and frequently perplexing. Data mining is gaining traction for a variety of reasons: firms are amassing more data about their operations, storage prices have decreased dramatically, and competitive business pressures have intensified. Other considerations include the advent of demands to rein in existing IT investments and, finally, the dramatic decrease in the cost or performance ratio of computer systems. Various mining approaches enable four fundamental mining operations: predictive model construction via supervised induction techniques; link analysis via association and sequence discovery techniques; database segmentation via clustering techniques; and deviation detection via statistical techniques.

businesses across a variety of industries, including retail, banking and finance, healthcare, manufacturing, telecommunication services, and aviation, as well as government entities, are already utilising data mining tools and methods to capitalize on historical data. By sifting through warehoused data using pattern recognition technology and statistical and mathematical methodologies, DM enables analysts to identify crucial facts, relationships, trends, patterns, exceptions, and anomalies that could otherwise go overlooked.

### 2.2.3. Marketing

The practise of marketing as a science involves conducting research and development on a product while simultaneously promoting its sale and distribution to the broader audience. Marketing has developed over time and continues to do so in response to shifting consumer preferences and purchasing habits. Therefore, marketing has evolved significantly from a few decades ago, owing in large part to a quickly growing global economy and technology breakthroughs that enabled open and speedy information exchange and commerce. With the deregulation of economies, the manner and speed with which information is delivered has evolved considerably (Kanth, 2015).

Marketing is defined as the act of putting the right product in the right area, at the right time, and at the right price. It is a process of conceptualising, pricing, advertising, and disseminating concepts, products, and services in order to generate exchanges that benefit both individuals and organisations (Kotler & Armstrong, 2006).

Much effort must be expended in understanding what people desire and where they purchase. Thus, it is a matter of determining how to manufacture the product at a price that represents its perceived value to clients and assembling it at the right moment.

### 2.2.4 Sales

According to (Alamsyah., 2021), sales analysis is the process of examining sales data in order to discover which items and services performed very well and which did not. The results of the study can be used to make decisions about how to keep inventory, how to evaluate the efficiency of a sales force, how to set production capabilities, and how the company is performing in respect to its objectives.

A sales analysis is crucial since it enables senior management in making judgments about inventory management, marketing efforts, strategies to implement, and any necessary process changes. Furthermore, by conducting sales analysis, the organisation gains insight into new industry dynamics, which supports in the development of future growth initiatives. According to Mckinsey, adopting Big Data Analytics in a retail chain can increase a retailer's operating profit by approximately 60%. (techvidvan.com, 2020)

### 2.2.5 Clustering

The process by which a collection of physical objects are assigned "classification" attributes and then grouped together into groupings of identical objects is referred to as "clustering" (Kanth, 2015). The data objects in a member of the same cluster are dissimilar to those in the other cluster, however they are not identical to those in other clusters. Data partitioning is used to find groups of data that share a common characteristic, such as location or category, and then groups are allocated various identifiers, dividing the whole data collection into a manageable number of subsets (Kanth, 2015).

The insight gained from purchasing trends, marketing population segmentation, and profiling enables marketers to identify different groups of customers and categorise them accordingly. This data allows a company to cross-sell its products, and thus aids management in implementing the 80/20 marketing philosophy, also known as the pareto principle, which states that 20% of your consumers generate 80% of your sales (Tardi, 2020). The problem is identifying the 20%, which can be accomplished through the use of clustering algorithms.

#### 2.2.5.1 K-Means

K-means algorithm is one of the clustering algorithms which considered as one of the oldest, commonly used, and most approachable algorithm for clustering method which was proposed by J.B. MacQueen (Li & Wu, 2012). K-means algorithm divides data objects into non-overlapping segment, which means that no data point can be a part of 2 or more clusters (Python, 2021)

Furthermore, k-means clustering algorithms feature two iterative phases which are initialization phase and centroid update phase. The initialization phase is an iterative process in which each data point is assigned to its nearest centroid using Euclidean metric (Oti et al., 2021) and the centroid update phase, in which cluster centroids are updated based on the partition achieved in the initialization phase. Moreover, the iterative procedure also comes to an end when no data point change clusters or a predetermined maximum number of iterations are achieved (Slonim et al., 2013).

Moreover, every machine learning algorithm have evaluation metrics to evaluate the performance of the algorithm, which in the case of clustering, the evaluation metrics are elbow method through the sum of squared error and silhouette analysis.

##### 2.2.5.1.1 Elbow Method

Elbow method or Elbow analysis is used to evaluate how successfully K-Means clustered a dataset. To determine it, first the distance between each data point and its centroid is measured, then that distance is squared, and finally the resulting squares are added together for each cluster. A good model has both low of sum squared error and a low cluster count (K). However, as the number of cluster increases, the score of the sum squared error reduces (Codeacademy.com, 2021).

##### 2.2.5.1.2 Silhouette Analysis

Silhouette analysis quantify on how well a data point fits to its assigned cluster through the calculation of the mean of the Silhouette Coefficient for each sample in each cluster. Moreover, by the used of python Scikit-Learn library, it allows the analyst to retrieve the silhouette score through the silhouette\_score() function in which it will summarized all the mean of silhouette coefficient into a score that can be used to evaluate the performance of the K-Means clustering algorithm (Python, 2021)

### 2.2.6 RFM Analysis

The RFM (recency, frequency, and monetary) model is a behaviour-based model which is used to assess a consumers’ behaviour and then create assumptions based on the dataset's behaviour (Wei et al., 2010).

The RFM model is the most often used segmentation approach since it combines three metrics (recency, frequency, and monetary) into a three-digit RFM cell code. Recency is sometimes viewed as one of the most crucial of the three RFM criteria (Wei et al., 2010)

The term "recency" refers to the time frame since the last transaction, "frequency" to the number of purchases made within a specific period of time, and "monetary" to the amount of money spent during this particular time period.

Consumers are grouped according to their purchase dates for recency. Recency is sometimes characterised in terms of the number of periods since the last purchase, which indicates the time gap amongst the most recent transaction and the analysis time (days or months), with the less days indicating a better recency rating (Wei et al., 2010). Hence, a high recency score indicates that a consumer is more likely to churn.

When it comes to frequency, the database is organised according to the number of purchases made in a certain time period. The term "frequency" is sometimes oversimplified to include two states: single and recurrent purchases. The top quintile is allocated a value of five, while the remaining quintiles are awarded values of four, three, two, and one. A higher frequency score, on the other hand, suggests more client loyalty (Wei et al., 2010). A consumer with a high frequency score indicates that he or she has a strong desire for the product and is more likely to purchase it often.

For monetary transactions, consumers are classified according to the total amount of money spent during a specific time period. The term "monetary" refers to the amount spent by the consumer during this time period, the average amount spent each purchase, or the total amount spent to date.

## 2.3 Similar System(s)

A similar system was discovered by the developer, this system used SAS Enterprise Miner to construct the system and was also applying clustering with the k-means algorithm. However, this system enhances the cluster into sub cluster for the best cluster performance. To obtain the sub-cluster with the best cluster performance, the system used clustering analysis, which was then improved by employing a decision tree to refine the segmentation of the customer (Chen et al., 2012).

Another study discovered that a similar system was built for the retail industry as well, which primarily focuses on mailing campaigns by employing the method of clustering to group customers who share similar characteristics, as well as another classification and regression algorithm for the individualization, which means one single target user, which then used for the personal mailing campaign (Garcke et al., 2010).

# CHAPTER 3: Technical Research

## 3.1: Programming language chosen

### 3.1.1 Introduction

Programming language in our use case is the core element for the purpose of the development of the data mining approach program. This also become one of the challenge due to variety of choices that can be chosen. Although there are a lot of programming languages available, that does not mean any programming language can be used for any type of project. therefore, developer need to spend some time to consider and choose the most suitable programming language to accomplish the objectives based on some factors.

Lastly, the availability of resources and supports of each programming language also take part in considering the right programming language since a programming language with less resources and support can be one of the challenges for the developer to accomplish the project.

### 3.1.2 Python

Python is an English-like syntax programming language which means that it uses English phrases and words that can be grasped by the users easily. Python is simple, and easy to learn syntax emphasizing readability and therefore reduces the cost of program maintenance (Python.org.2019).

Python is one of the most popular programming languages among data science field community, and according to (Carraz, 2022), a massive statistic in 4th quartile of 2018 shows that majority of machine learning developers are using python as their programming language with proportion of 69% are using it, compare to R programming with the proportion of 24% as the other similar programming language for machine learning (Carraz, 2022).

The chosen programming language by the developer is Python, specifically version 3.7.3 as the current python version installed in the developer’s computer. Python is a programming language that can be used for many types of purposes namely like data analysis, machine learning, web development, and many more. Python is found to be easy to use thanks to its simplicity and easy to learn which also supported with good and active ecosystem providing the libraries from data analysis, to machine learning which is well suited for the case of this project.

The table below will show some of python strengths and weaknesses, also comparison between python and the other 2 programming languages, R programming and SAS Programming.

Some of the Strengths and Weaknesses of python programming language are

|  |  |
| --- | --- |
| Strength | Weakness |
| Easy to learn, read, and write the code with English-like syntax. | Speed |
| Free and Open Source | Memory consumption |
| Varieties library supports available | Slow processing power |

Table 1: Python Strengths & Weaknesses

Comparison between Python, R programming, and SAS Programming.

|  |  |  |  |
| --- | --- | --- | --- |
| Comparison Factors | Python | R | SAS |
| Overview | Is an open-source scripting language that can be used for verities of development purposes namely like web development, mobile application development, including data mining tasks. | An open-source programming language to perform complex statistical and mathematical calculations | a software tool for statistical analytics |
| Cost | Both Python and R is cost-free programming language which can be accessed and used by the developer | | The most expensive option among the similar tool, as it is a commercial software. However, it provides a free university edition available (techdivan) |
| Ease of Learning & Coding Wise | Although python is a high-level programming, it is easier in term of learning compared to R and also lesser line of codes compare to R. | It is a low-level language; thus, it requires extra code for simpler tasks. | Easiest to learn due to its drag-and-drop feature also It requires minimum amount of coding in SQL programming language, and it is only if needed. |
| Community Support | Large Community and lot of documents available with regards to the functions, library, etc. | It has enormous size of community that consist from many different industries | Due to its cost, the SAS community is pretty small compared to the other 2, however in term of documentation, SAS provides comprehensive documentation |
| Application Advancement | the development of new features and techniques are fast. However, the risk of error is there due to the fact it is an open-source contribution (Ravindra Savaram, 2018) | | It is less prone to errors compared to R and Python due to SAS releases updates in controlled environment, which made them well tested beforehand. . |
| Data Handling Capability | All the similar programming languages / tools are similarly can handle huge amount of data | | |

Table 2: Programming Languages Comparisons

After the comparison in many factors between Python, R, and SAS programming language, it comes to decide which programming language will be used based on the considerations mentioned above. The developer chose python as the right programming language for the development of data mining program although the other 2 programming languages, R and SAS are also sufficient to develop the project. One of the reason python is the chosen programming language it is because the knowledge that the developer have with regards to its’ syntax, libraries, and structure are sufficient to perform the development of this project.

Each programming language's resource and support availability should be considered while making a comparison in order to choose the best programming language to use for the project. Therefore, when programming language implemented in the project, the developer able to use the benefit of python’s varieties of library, supports, and resources that are available to support the developer encountering issues during the development of data mining program.

Lastly, python can be used to compress huge amount of data that can turn complex data into an actionable insight. Which in this project, it can be used to perform RFM analysis and Market Basket Analysis to accomplish the project objectives.

## 3.2: IDE (Interactive Development Environment)

The Jupyter Notebook is an open-source web application that permits data scientists to create and share documents that combine live code, equations, computational output, visualisations, and other multimedia elements with explanatory text (Science, 2020).

The name of jupyter was establish from the acronym of the programming languages that jupyter notebook supports, which are Julia (JU), Python (Py) and R (Science, 2020)

The version of jupyter notebook that will be used is 5.7.8 that is installed in the computer of the developer.

The reason Jupyter Notebooks is the chosen IDE by the developer is because the fact that it can be used to perform data mining tasks including exploratory data analysis, data transformation, visualization and cleansing, statistical modelling (Science, 2020) , which are the core tasks to achieve the objective of the project.

## 3.3: libraries chosen / Tools chosen

### NumPy

NumPy is the foundational Python library for scientific computing. Multidimensional arrays, masked arrays, and matrices are just some of the things this Python library can do. It also has an array object, as well as a lot of other things that can be done with arrays quickly. This includes things like mathematical and logical operations, as well as things like sorting and selecting.

### Pandas

Pandas is a Python module that provides quick, versatile, and powerful data structures that make it simple and natural to work with "relational" or "labelled" data. It is intended to serve as the foundational high-level building block for carrying out the necessary data analysis task in Python.

### Matplotlib

Matplotlib is a Python package that enables the creation of static, animated, and interactive visualisations and It is a cross-platform data visualisation package built on NumPy arrays.

### Seaborn

Seaborn is a Python package for creating statistical visualizations. Seaborn based on matplotlib and tightly integrated with pandas’ data structures (Seaborn, 2022). Although Matplotlib is already selected for visualization, seaborn works by providing deeper visualization which in this project, seaborn will be used to generate graph from the dataset to retrieve actionable insight from the hidden pattern of the dataset with an appealing visualization to make statistical graph more attractive.

### Datetime

The Python Datetime package includes classes for manipulating date and time. These classes include various features for manipulating dates, times, and time intervals. Because date and datetime are objects in Python, handling them involves manipulating objects rather than strings or timestamps. Which in this project, the package will be used to separate the date and time of the transaction to retrieve valuable information with regards to marketing.

### Scikit-Learn

Scikit-learn is a free Python library for machine learning. The scikit-learn library offers a variety of fast machine learning and statistical modelling methods, such as classification, regression, clustering, and dimensionality reduction. Which in this project, Scikit-learn will be used to perform the clustering technique and to evaluate the appropriate number of cluster through elbow method and silhouette analysis

## 3.4 Database Management System

The developer does not require any DMBS for the development of the project.

## 3.5: Operating System chosen

Windows 10 Education 64-bit, version 1803 will be the operating system used by the developer for the design and development of the data mining program. Due to the Operating System is already installed way before the planning of this project, this become the reason the particular operating system, Windows 10 Education 64-bit, version 1803 is the selected operating system for the development of the project.

Hardware Requirements:

Intel® Core™ i5-6600 CPU @3.30 GHz

8.0GB RAM

## 3.7: Web browser chosen - (this section is optional)

The developer does require a web browser for the development of the project. The reason it is required because the developer is going design and build the data mining program in a web-based interactive development environment.

The chosen web browser is Google Chrome version 99.0.4844.52(64-bit). The reason that google chrome is the chosen web-browser although the objective of this project is not developing a web-based program, it is due to Jupyterlab that will redirect to a web-browser when launched made web-browser required for the development of the data mining program.

## 

## 3.8: Summary

To summarize the technical research chapter, the developer has chose python programming language version 3.7.3 as the only programming language that will be used for the purpose of development of the data mining program. Moreover, the IDE that will be used is jupyter notebook due to its functionalities that can support data mining task from data importing until the model building. Moreover, the python libraries that will be used throughout the development of the project are NumPy, Pandas, Matplotlib, Datetime, Sklearn.

Lastly, there is no computer without operating system. In this project, the chosen operating system by the developer is Windows 10 Education 64-bit, version 1803 as the installed OS in the developers’ computer. with the support of inter i5 6th generation and 8 of RAM, this such system is considerably sufficient to develop the project in Python programming language.

# CHAPTER 4: Methodology

## 4.1 Introduction

There are a variety of methodologies available on the market nowadays, making it more difficult for a project to pick the best appropriate methodology for the development of the project. each methodology has its own type of project that can be done, which including the size of the project that will be developed.

## 4.2 Methods

CRISP DM

Crisp-dm is an acronym for Cross Industry Standard Process for Data Mining. It is a data mining technique that was created in 1996 by five companies: Integral Solution Ltd (ISL), Teradata, Daimler AG, NCR Corporation, and OHRA, a large insurance business (Kumar, 2020). Another aspect of the Crisp-dm data mining life cycle is that it contains sequential steps, and the sequence is not strict, which means that if any of the stages fails to fulfil a need, one may simply go backwards to the previous stage without having to restart the entire process.

Diagram

Description automatically generated

Figure 1: CRISP-DM Life Cycle

As shown in the above figure, The crisp-dm methodology's lifecycle consists of six phases. The outcome of each stage determines the next phase or particular task within a phase that has to be accomplished. The arrows indicate the crucial and regular relationships between the stages. Additionally, the outer circle symbolises the cyclical nature of data mining. However, after the solution is built, the data mining process is not complete; it must be assessed by the team to ensure that it meets the company's objectives before being implemented (Chapman, et al., 2021). The crisp-dm lifecycle that involve in the development of data mining program for retails are:

### 4.2.1 Business understanding

As part of the CRISP-DM process, the first stage is to determine the company's goals in the business aspect. A detailed examination of the organization's goals and priorities is necessary. This phase's goal is to identify factors that might have a substantial impact on the project's outcome. However, if this stage is skipped, a lot of effort will be wasted trying to come up with the right answers to the wrong questions (sv-europe.com, 2021).

This phase necessitates the initialization of the company's business goal in order to meet its stated goal. As a result, in order to set a clear goal and schedule for the data analysis, the first step is to thoroughly research all of the relevant tools, assumptions, limits, and other factors.

### 4.2.2 Data Understanding

It is necessary for a developer to comprehend the data that has been collected from the data source by investigating data relationships, distributions, rudimentary statistics, and some aggregation during this stage of the development process. Furthermore, the outcome will be documented in a report in a professional manner to accurately reflect the outcome. However, it is not only restricted to examining the data, but it also validates the quality of the data by determining whether there are any missing or incorrect values, as well as determining if there is any mistake with the data format.

During the second phase of CRISP-DM cycle it comes to part where developer needs to understand the summary information of the data in the project resources and provide metadata such as variable name, data type, format of the data including description of the variable. In this project, dataset will be collected from Kaggle.com as there is no ample time to collect huge amount of data for the purpose of developing the data mining program.

Moreover, data understanding phase broken down into 3 tasks that needs to be performed which are:

#### Data Gathering

This process entails gathering information in order to develop a plan and set objectives in term of the data mining aspect. This procedure entails importing the data, which aids in the process of comprehending the information.

Moreover, before the data is finalised to be the chosen dataset for the development of the project, the quality and availability of the data must be verified.

#### Data Description

Among other things, this section outlines the data structure, quantity of data (e.g., the number of records),  and other aspects that may be used to determine if the data retrieved fits relevant requirements.

The dataset is a public Kaggle retail dataset that will be used to model RFM-based segmentation. Invoice number, stock code, time stamp of transaction, nation, and many more variables are included in this comprehensive retail detail dataset. The collected dataset is a CSV file containing 532,621 entries and 8 variables (Kaggle.com, 2017).

Below is a table containing information about the dataset's metadata.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data Type | Data Format | Description |
| Invoice No | Object | 536366 | Unique sequential code assigned to each invoice |
| Stock code | Object |  | A unique code to identify the product category |
| Description | Object | Knitted union flag | The name of the item |
| Quantity | Int64 | 1 | Number of quantities of a particular item that has been bought |
| Invoice Date | Object | 12/1/2010 8:26 | The date and time of the transaction take place |
| Unit Price | Float64 | 2.55 | The price per item |
| CustomerID | Float64 | 17850 | The id number of a specific customer |
| Country | Object | United Kingdom | Name of Country |

Table 3: Metadata

### 4.2.3 Data Preparation

The dataset must be cleaned and standardised as part of this phase in the data analysis process. As a result, the data preparation must precisely capture each record's invoice, value, and owners if the data is to be used for analysis.

The CRISP-DM approach relies heavily on this step. ' As key phase in the process of data mining, pre-processing occurs during the data preparation stage. It is common for raw data to contain inaccurate or meaningless information, due to the chaotic nature of the dataset format that is not 100% well suited for every objective.

Data variables that aren't relevant will be removed from further consideration in the pre-processing stage of the algorithm. In addition, it is critical to identify and highlight inconsistent data points in the dataset. Human mistake, for example, might lead to the mixing or erroneous filtration of information between columns.

Once this formal procedure is developed, it will be possible to determine if data is lost spontaneously or systematically.

Data preparation tasks are likely to be performed several times and in a variety of ways. This phase on the other hand is where the data reduction, transformation, cleansing will be done in order to come up with reliable model as statistical model mostly sensitive to some problems like outliers, missing or invalid value, and the consistency of the data.

Data preparation can be boken down into several task which are

#### Data cleansing

Preparing raw data for analysis by eliminating errors, arranging the data, and filling in blanks is called data cleaning. Almost every piece of data that is gathered has a missing value or a value that is not applicable. Imputation is one method for dealing with a missing value. As for categorical variables, the most frequent value is utilised to fill in any gaps left by missing values, while for continuous variables, imputation refers to assigning an average or mean value to replace the omitted information. Nevertheless, if the missing value has beyond a certain threshold, it is advised that the specific variable be eliminated in order to prevent erroneous data. In most cases, the threshold for missing values cannot exceed 30%.

#### Variable Selection

Data relevant to a certain analytic activity is obtained from a dataset via variable selection. Also, Data transformation and data consolidation may be necessary before the selection of data can be done. This makes it clear that the information in each area should be used or excluded based on the importance of the objectives and the information gathered.

#### Data Construction

Proactive data processing procedures, such as the production of extracted attributes or new records, or the transformation of information, are part of the data construction function. Which in this project, constructing data will take place in order to come up with the Recency, Frequency, Monetary value which it can only be generated by doing feature-engineering, which come up with additional records and variable for the dataset that was extracted from the collected dataset.

#### Data Transformation

Which in this project, when customer decided to return the item due to some reason, the dataset shows record of negative value in quantity, which does not make sense. This negative value can be assumed happened due to item return. Therefore, to get valuable information over inconsistent data, the negative values must be transformed. This phenomenon is known as data inconsistency in data mining term, which value that is distinct that the others. Thus, after the treatment of inconsistency, the data can now be stored.

#### Storing Data

Upon the completion of data transformation, cleansing, and selection has take place, the data is then will be stored which also allows the developer to have fully right to read or write with regards to the data.

### 4.2.4 Modelling

#### Selecting Modelling Technique

First and foremost, to select which algorithm will be used for the project. There are several modelling techniques but not all modelling technique can be used to perform a particular statistical model for a particular dataset. The objective of the project will be the core element of consideration in which model should be used to accomplish the objectives.

The chosen modelling technique to accomplish the project objective is

• K-means clustering

#### Generate the test design

It is necessary to divide the data into two datasets, one for training the model and the other for testing the model's predictions based on the training dataset. However, when it comes to clustering technique, separating data into 2 is not necessary as splitting data is commonly used for supervised learning while clustering is part of unsupervised learning. Instead of splitting data into 2 sets of datasets, clustering can straight perform build the model due to the fact that unsupervised learning means no labels can be used as the target variable.

#### Model(s) Building

While common model building required test set and training set, it is not the same when it comes to unsupervised learning. Unsupervised learning technique can straight perform model building upon the completion of the data preparation

#### Asses the model(s)

In this sub-task, the model will be reviewed from technical, also business domain perspective. The result of the assessing model sub-task will be a detail report with regards to the model that has been developed should be presented to the business domain. Moreover, a fine-tune revision might also take place if it is requested by the business domain

### 4.2.5 Evaluation

In this phase of CRISP-DM cycle, once the modelling results have been obtained, a comprehensive evaluation of the findings should be done. Formal evaluation might include evaluating the model's capabilities on the observed data to evaluate how successful and efficient the algorithm is. The evaluation phase is broken into three tasks, which are

#### Result Evaluation

In this stage, the model's alignment with the company's goals is assessed, and an attempt is made to determine if the model's disadvantage has a business justification. This step also includes a review of any additional data mining results that may have been generated. These models are important for the original business goals to be achieved, but any other observations that are not necessary but may reveal new challenges or facts might be indications for future path.

Upon the completion of this sub-task, the output will be a review of the outcomes for the business performance indicators defined during the business understanding phase. These findings would demonstrate unambiguously that the business objectives established at the beginning of the project were met.

in the evaluation phase, the data mining result will be checked with its respective evaluation metrics. Which in the case of this project, the evaluation matrices are

1. **Elbow Method**
2. **Silhouette technique**

to determine the appropriate number of clusters can be given for the model.

#### Reviewing the process of the development

Thus, after executing the initial evaluation, the review procedure begins. During this evaluation process, models that appear to be appropriate and match business needs are identified; consequently, more extensive research is required to discover whether any key component or activity was overlooked.

#### Determining the next action of the cycle

The evaluation phase ends with suggestions for the next step of the process. Moreover, during this sub-task, the model can be considered deployed-ready. Also, during this sub-task, the developer can go back to previous phase which are building another model or re-transform the data in-case it is needed without any prohibition

### 4.2.6 Deployment:

The final cycle of crisp dm can be a deployment stage, or presentation depends on objective of the project which in this project, the developer does not require to deploy the developed model, instead the developer should come with a comprehensive documentation report. The deliverables will allow the retail industry to understand the behavior of their customers, characteristics of the customers, also to help retailer to determine which products are often purchased together among their customers.

## 4.3 Summary

This section contains a description of the experiments conducted and the results achieved over the course of six CRISP-DM cycle:

Begin with the Business Understanding phase, it was apparent that the goal was to boost retail efficiency by enabling businesses to develop recommendation approaches and marketing strategies for various sorts of customers.

The Data Understanding phases required the generation of an overview of the data's essential features and a summary of the data, as well as an assessment of their consistency. This is where the retail transaction history is created, complete with the relevant variables such as the date and time of the transaction, the product purchased, and the country in which the transaction occurred.

Following this investigation, some basic data pre-processing will be performed in order to generate a rough prototype model. The procedure consists of locating measurement fields, cleaning the data.

A modelling approach can be used during the modelling phase to choose, seek for, and apply a suitable modelling technique to data that was generated during the Data modelling phase.

Clustering from unsupervised learning method will be utilized to ascertain the retail customer's behavior. Finally, an evaluation is conducted to establish the optimal number of customer clusters through the elbow method and silhouette analysis.

# CHAPTER 5: Data Analysis

## 5.1 Introduction

The dataset is a public Kaggle retail dataset that contains transactions that occurred between 01/12/2010 and 09/12/2011 for a UK-based online retail. The collected dataset is a CSV file containing 541,909 numbers of observation and 8 variables which are Invoice No, Stock Code, Description, Quantity, Invoice Date, UnitPrice, CustomerID, and Country.

## 5.2 Initial data exploration

### 5.2.1 Import Dataset

Graphical user interface, text, application

Description automatically generated

Figure 2: Importing Dataset and Required Libraries

First and foremost, in order for the analyst to start the project in Python, Pandas library from python must be imported at first. The reason pandas must be imported at first is due to its capability to read the dataset. Pandas is not limited to read dataset, it can also perform creation of a data frame, manipulating the dataset, counting value of data, and even a graph visualization.

Moreover, Pandas library can be used to read various types of dataset file types such as Column Separated Value (CSV), JavaScript Object Notation (JSON), Hypertext Markup Language (HTML), Excel, and Standard Query Language (SQL). In this project, the analyst is using a CSV file which then required the analyst to run read\_csv() function in order to import the dataset to the python notebook file. Moreover, in this project, the analyst required to use the encoding parameter in order to import the dataset, and type of encoding used by the analyst is latin1. It is a type of encoding which used to solve UnicodeDecodeError upon importing the dataset.

Graphical user interface, text, application

Description automatically generated

Figure 3: First 5 Rows of the Dataset

The figure 3 above shows table upon importing the dataset into the python notebook. Furthermore, the figure above shows that the analyst is printing the first 5 rows of the dataset by using pandas head() function. By default, the number of rows that will be displayed are 5, thus the corresponding cell output shows the first 5 rows of the data.

Moreover, the analyst also printing the shape of the data by using pandas shape which will retrieve the dimensions of the data frame. Which in this case, the shape function has been used to clarify that the dataset contains 541,909 observations with 8 variables as per the description of the dataset.

### 5.2.2 Data Types

Text

Description automatically generated with medium confidence

Figure 4: Data Types

The figure 4 shows the data types of each column that the dataset contains through the use of dtypes function which to return the data types in a data frame. Upon returning the data types, it shows that the there are 4 objects data types which are InvoiceNo, StockCode, Description, InvoiceDate, and Country, followed with 2 float dtypes which are UnitPrice, and CustomerID, and a single int data type which is Quantity. Object data type are column that contain value mixture of string, integers and float.

Whereas float64 data type is a column that contain floating point number with 64 bit of memory allocated to store the data and integer64 data type belongs to column that contain integer value with 64 bit of memory allocated to store the data.

Upon checking the data types, the analyst came up with a plan to change some column data type for the purpose of developing the program. The column data type that will be converted in the data cleansing stage are CustomerID into Object, InvoiceDate into date time data type.

### 5.2.3 Descriptive Statistic

Text, table

Description automatically generated

Figure 5: Descriptive Statistics

The above figure 5 shows the analyst retrieve the descriptive statistic of each column in the dataset by using describe function. Describe function is used to retrieve a descriptive statistic of a data frame which will return the value of frequency, mean, standard deviation, etc.

Upon analysing the descriptive statistic of the dataset, it shows that in average customer bought 10 quantity of product, and the max quantity in a single transaction are 80,995 quantities of a product which can be assumed it was a wholesaler customer who made that transaction. In term of UnitPrice, the average price of a product is £5, and the most expensive product is £38,970.

Furthermore, the analyst notice that the value of standard deviation is higher than the mean, which indicate that there are outliers in Quantity and UnitPrice columns. Not limited to that, the statistic also shows that there are some negative values in Quantity and UnitPrice which needs to be cleaned for the development of the clustering. The negative values that occurred in Quantity and UnitPrice can be assumed due to transaction that has been cancelled, therefore the value of Quantity and UnitPrice below 0 will be removed from the dataset in the data cleansing stage.

In order to retrieve the descriptive statistic for categorical variables like InvoiceNo, StockCode, Description, InvoiceDate, and Country, the analyst used the include parameter and set it to object which tells pandas to retrieve the descriptive statistic for object data types.

Upon analysing the descriptive statistic, it shows that there are 25,900 unique invoice number which indicate that there were 25,900 transactions between 01/12/2010 and 09/12/2011, also the highest frequency of an invoice number belong to InvoiceNo. 573585 with 1114 frequencies, which can be assumed that this customer is a wholesaler customer.

Moreover, there are 4223 of unique products sold by the retail store with product named ‘WHITE HANGING HEART T-LIGHT HOLDER’ as the most product sold with 2369 frequencies of product sold within the time range. In addition, the country variable is where the customer located and among 38 countries, United Kingdom is the country where most of customers located with 495,478 frequencies of customers located in this country.

Lastly, there are 4070 stock codes available in the retail store, and the most used stock code is StockCode 85123A with 2313 frequencies.

### 5.2.4 Missing Value

Table

Description automatically generated

Figure 6: Checking Missing Values

The above figure 6 shows how the analyst retrieve the number of missing values. The number of missing values are retrieved by the used of isnull() method from pandas library which returns Boolean value for the corresponding row, if it returns TRUE it mean there is missing value, otherwise FALSE. In order to get the total missing value, the analyst used sum() function to return the sum of missing value in each column. Next, the analyst decided to divide the sum by the number of rows in the data frame by using len() function to retrieve the percentage of missing values.

The output of the code shows there are 2 variables contains missing values which are Description with 0.27% of missing value, and CustomerID with almost a quarter of the whole dataset, specifically 24.93%. Hence, these missing values must be cleaned in the data cleansing section.

## 5.3 Data Cleansing

### 5.3.1 Drop Missing Values

Text

Description automatically generated

Figure 7: Removing Missing Values

The above figure shows how the analyst dealt with the dataset. The analyst used dropna() function which will drop or remove the row if there missing value detected, and update the data frame to remove index that contain missing value in the row. As a result, the data frame now contain 0 missing values.

### 5.3.2 Data Reduction

#### 5.3.2.1 Removing Negative Value in Quantity and UnitPrice column

Table

Description automatically generated

Figure 8: Remove Negative Value in Quantity and Unit Price

The above figure shows how the analyst dealt with the negative value in the Quantity and UnitPrice column that was detected during the initial data exploration. To remove the negative values, the analyst decided to update the data frame and put a conditional statement to update the data frame with only value above 0 will be stored in the data frame, otherwise it will be dropped. As a result, the analyst can gain proper insight through the descriptive statistic of the Quantity and UnitPrice column. The mean of quantity that was previously shown as 10 quantities per item, into 13 quantities on average of each product bought. Also, the minimum quantity of product sold became more realistic as previously was negative value.

Same goes to the UnitPrice, the average price of a product was shown £5 for a product during the initial data exploration, became £3 for a product. Hence, the insight gained from the descriptive statistics upon dealing with the negative values are more realistic.

#### 5.3.2.2 Reducing Country to Only United Kingdom

Graphical user interface, text, application

Description automatically generated

Figure 9: Selecting Country to Only United Kingdom

As per the result of descriptive statistic of the country column in initial data exploration section, the analyst reduced the country to only select United Kingdom as it is where most of the transactions located.

### 5.3.3 Data Transformation

#### 5.3.3.1 Transform CustomerID Data Type into Object

Table

Description automatically generated

Figure 10: Change CustomerID Data Type into Object

The above figure 10 shows the step of how the analyst changed the CustomerID data type into an object data type. The analyst used astype() function to cast or in simple term, to change or transform the CustomerID data type into an object data type. As a result, the output of dtypes function to show the data types in the data frame, shows that the CustomerID turned into an object data type.

#### 5.3.3.2 Transform Date Variable

Graphical user interface, text, application

Description automatically generated

Figure 11: Transform Date Variable

The analyst decided to come up with a feature engineering to came up with new column named Date which to store only the date time of the transaction instead of timestamp. In order to create the date time, the analyst must use Pandas to\_datetime() function to transform the data type of InvoiceDate that was previously as an object into datetime and update the new Date column data type.

Upon transforming the data type into datetime, the analyst used date function through the python datetime library to extract only the date of the transaction.

#### 5.3.2.3 Create New Variable ‘daysAgo’

Graphical user interface, text, application

Description automatically generated

Figure 12: New Variable 'daysAgo'

The figure 12 above shows how the analyst come up with new variable named as daysAgo to extract the difference between the date of transaction with the last date in the dataset. To calculate the difference, a variable named today declared to store the date return from the max() function which to retrieve the maximum date in the Date column.

Moreover, daysAgo variable created through the calculation of variable today, which stored the last day of the transaction, subtracted by the date of the transaction in the Date column. As a result, the column daysAgo is created containing value of how many days ago the transaction took place.

#### 5.3.2.4 Create New Variable ‘Monetary’

Graphical user interface, application

Description automatically generated

Figure 13: Create New Variable 'Monetary'

Figure (New Variable ‘Monetary’)

The figure 13 shows how the analyst come up with new variable, named as Monetary which is used to calculate the amount of monetary in a single transaction. Monetary variable was generated through the python multiplication operator to multiply the Quantity of a product bought with the UnitPrice of the product.

#### 5.2.3.5 Create Recency

Table

Description automatically generated

Figure 14: Create Recency Data Frame

Upon the completion of preparing some additional information needed from the dataset, the analyst now can calculate Recency variable from RFM. In cell 19 in the above figure, the analyst created a data frame named **rfm\_r,** which used to store the result of grouping the customerID with daysAgo column. The groupby() function allowed the analyst to group the data of a specified column with some function in it. Which in this case, the groupby() function will return the value of customerID and the value of daysAgo. In addition, the reset\_index() function is used to reset the indexing number of the data frame from 0 to number of rows – 1.

Furthermore, the analyst used days function from the python datetime library to retrieve only the number of days. Hence, the Recency variable is now ready to be used for further analysis.

#### 5.2.3.6 Create Frequency

Table

Description automatically generated with low confidence

Figure 15: Create Frequency Data Frame

The above figure 15 shows how the analyst count the Frequency variable for RFM. Same as creating the Monetary data frame, here the analyst also used groupby function which group the customerID with the count of Invoice number to retrieve information on how many times a customer make transaction with the store.

Moreover, the InvoiceNo column named changed into Frequency in the **rfm\_f** data frame by using columns function which used to label the name of column of the data frame.

#### 5.2.3.7 Create Monetary

Graphical user interface, table

Description automatically generated

Figure 16: Create Monetary Data Frame

The above figure 16 shows how the analyst count the Monetary variable from RFM for each customer. In order to calculate the total monetary value for each customer, the analyst used the Python sum() function to retrieve the summation amount of Monetary for each customer which was calculated through the calculation of Quantity multiplied by UnitPrice.

#### 5.2.3.8 Merge Multiple Data Frames

Table

Description automatically generated

Figure 17: Merging Multiple Data Frames

Upon the completion of creating data frames for each RFM variable, which are Recency, Frequency, and Monetary value, all three data frames are now ready to be merged from multiple data frames, into one single data frame.

In order to merge data frame, python does not support to merge more than 2 data frames at a time, hence the analyst merged the first two data frames, which are **rfm\_r** for Recency**,** and **rfm\_f** for Frequency variable by using Pandas merge function to merge the data frames, and then merge it with the last variable which is **rfm\_m** for Monetary.

The merge function has multiple parameters that can be defined, in this case, the analyst defined how parameter to use **inner** type of merge, which is similar to SQL inner join, in which it will combine the data frames when there is matching value found.

Moreover, the **on** parameter is used to define that the join is on column level. Since all the three data frames are depending on the CustomerID, hence CustomerID is defined in the **on** parameter which return value based on the CustomerID that can be found in the 3 data frames.

#### 5.2.3.9 Removing Outliers from the Data Frame

Text

Description automatically generated

Figure 18: For Loop to Remove Outliers

Upon the completion of merging data frames that contain only Recency, Frequency, and Monetary variables, the outliers must be treated with statistical approach to avoid bias result that could happened if outliers are not treated well.

The statistical approach to remove the outliers is through inter-quartile range (IQR) that can be retrieved through calculating the 3rd quartile subtracted by 1st quartile.

Moreover, the analyst can calculate IQR through finding the 1st and 3rd quartiles by using NumPy library percentile function to retrieve the 25th percentile which refer to 1st quartile, and 75th percentile as the 3rd quartile.

As a result, the IQR value will be used to calculate the upper bound and lower bound of the values which then will be used as an indicator where if any values fall outside of this range is considered as outlier. The upper bound can be calculated through Q3 + (1.5 \* IQR), whereas the formula for lower bound is Q1 - (1.5 \* IQR).

In the above figure, it shows the analyst created for loop for each variable in the rfm data frame which will calculate the Q1 and Q3 of each variable through percentile function to get the IQR range, max which refers to upper bound, and min which refers to lower bound.

Moreover, the for loop will check the value based on a conditional statement where if the value is lower than min or greater than max, the value will be converted into NaN or missing value by using NumPy nan function.

As a result, there are new missing values detected now in the rfm which indicate that variable contain missing values meaning that there are outliers detected by the for loop. Therefore, the value that converted into missing values will be dropped or removed from the rfm dataset which will resulting a clean dataset that can be used for developing the clustering algorithm.

Table

Description automatically generated

Figure 19: Validating Outliers has been Removed

#### 5.2.3.10 Rescaling the Variables

Text, table

Description automatically generated with medium confidence

Figure 20: Variables Scaled

In order to avoid bias result in terms of accuracy rate, attribute scaling is important for developing clustering algorithm like K-Means. Scaling is needed to make sure that all the data are at the same level during k-means modelling calculation.

In this step, the analyst created new data frame named **rfm\_df** which is used to store Receny, Frequency, and Monetary attribute without including the CustomerID as the scaled data frame is for model development. The data frame was then standardized by using StandardScaler() function from Scikit-learn library that is declared as **scaler**.

Upon initialized the StandardScaler(), the analyst can use the fit\_transform function from StandardScaler() to calculate the standardized value and transform the actual value of the **rfm\_df** data frame into the standardized value and store the standardized value into variable named **rfm\_df\_scaled.** As a result, the value of **rfm\_df** data frame are standardized in the **rfm\_df\_scaled** data frame.

## 5.4 Visualization

### 5.4.1 Boxplot RFM variables

Diagram

Description automatically generated with low confidence

Figure 21: Boxplot with Outliers

The boxplot graph shows that all the three variables contain outliers which can be determined by looking at the graph that it makes the boxplot so tiny, even the box is not even visible. Thus, the outliers must be treated in statistical manners to prevent bias upon the result of k-means clustering.

Chart, box and whisker chart

Description automatically generated

Figure 22: Boxplot without Outliers

As a result of outliers’ removal through the for loop, the boxplots are now can be analysed through the statistic. The statistic for Recency variable of the boxplot shows that the median value is 57, the upper bound is 329 and lower bound is 0. As for Frequency, the median is 34, the upper bound is 160 with the lower bound of 1. Lastly, in terms of Monetary variable, the mid-point fall at 551.04 with upper bound fall at 2352.53 and lower bound at 3.75.

Furthermore, values that belong outside the range of upper bound and lower bound are considered outliers. However, some customers are still considered as outliers although it has been treated. For instance, in Monetary, value outside the maximum whiskers indicates that those are big spender customers. Thus, these outliers are not supposed to be removed.

### 5.4.2 Country

A picture containing timeline

Description automatically generated

Figure 23: Top 10 Country

The visualization above shows the top 10 countries where the transaction took place, and it shows that the top 10 countries are United Kingdom, Germany, France, Eire or Ireland, Spain, Netherlands, Belgium, Switzerland, Portugal, and Australia. Moreover, the visualization also shows that only a single country outside of Europe belong to the top 10 which is Australia. Thus, it shows that most of the customers are based in Europe.

### 5.4.3 CustomerID

Chart, bar chart

Description automatically generated

Figure 24: Top 10 Customer based on Number of Transactions

The figure 24 shows the top 10 customer who had the greatest number of transaction, and customer with CustomerID 17841 is the customer who had the most frequent transaction with the store with more than 7000 frequency count. Moreover, these 10 customers are considered customer who are loyal with store.

### 5.4.4 Stock Code

Chart, bar chart

Description automatically generated

Figure 25: Top 10 Used Stock Code

The figure 25 above shows that most stock codes frequency count lies between 1000 to 1500 frequency counts, and the most stock code found in the historical transaction is stock code 85123A that has more than 1750 frequency counts.

### 5.4.5 Product Sold

Chart

Description automatically generated

Figure 26: Top 10 Product Sold

Out of 4,223 unique products sold in the store, these are the top 10 most product bought by the customers. Product named ‘WHITE HANGING HEART T-LIGHT HOLDER’ is dominating among others product with 1940 count of frequency, followed with ‘JUMBO BAG RED RETROSPOT’ with almost 1500 frequency. Furthermore, among the top 10 product, the least amount of frequency sold for a product is product named ‘HEART OF WICKER SMALL’.

### 5.4.6 Historical Demand

Graphical user interface, application

Description automatically generated

Figure 27: Historical Demand

Figure (Historical Demand)

The graph shows the highest frequency of invoice generated was on 4th quarter, which is on December, with 3246 invoices generated within a day. Moreover, the lowest frequency was on December 2010, with the frequency of invoice generated is just 175 counts, and within the whole range time, the average is 1162 invoice generated every month.

Furthermore, the graph shows that there is pattern that in every 4 months, number of transactions are declining which was on 2011-01, 2011-05, and 2011-09. This could happen due to seasonal change that change the customer behaviour.

## 5.5 Hypothesis

The analyst does not have any hypothesis for the development of the project.

## 5.6 Reports and Dashboard

The developer does not require any Reports and Dashboard for the development of the project.

## 5.7 Modelling

Graphical user interface, text, chat or text message

Description automatically generated

Figure 28: Initialize K-Means Model

As per the objective of the project, which to segment customer into distinct group, the K-Means algorithm from Scikit-Learn library is used to build the K-Means algorithm.

Furthermore, to develop K-Means algorithm, the analyst initialized the model with some parameters included which are n\_cluster to define the number of clusters and centroid to be generated, and max\_iter is the maximum number of iterations of the algorithm in a single run. However, the initialized model is not the final, yet it is to find the right number of clusters through some evaluation metrics, which are Elbow method, and Silhouette Analysis.

Text

Description automatically generated

Figure 29: Elbow Method

The figure 29 shows how the analyst evaluate the appropriate number of clusters through elbow method. The analyst declared a list named **elbow** to store result of the elbow method, and a list containing pre-defined number of clusters which will be used to define the number of clusters in the K-Means.

Furthermore, a for loop created by the analyst to iterate the K-Means algorithm as well as retrieving the score of how well the dataset was clustered by the K-means through inertia\_, an attribute of Scikit-Learn K-Means algorithms that returns the sum of squared distances of samples to their closest cluster center.

Hence, the result of the for loop was then plotted by using Pyplot Application Programming Interface (API) from python matplotlib library.

Chart, line chart

Description automatically generated

Figure 30: Visualization Elbow Method

Upon analysing the result of elbow method in the figure 30, it shows that the appropriate number of clusters can either be 3 or 4 clusters. Hence, it can be considered during the final model development.

However, as a data analyst, it is a must to have more than one evaluation metric to support the decision made by the analyst. Thus, another evaluation metric, silhouette analysis has been conducted to decide the most appropriate number of clusters for the model.

Graphical user interface, text, application

Description automatically generated

Figure 31: Silhouette Analysis

In this step, the analyst used the similar for loop that was created for elbow method, however the difference is instead of retrieving the sum of squared distances of sample to their closest cluster, the analyst used silhouette\_score() function from Scikit-learn library to calculate the mean of the Silhouette Coefficient for each sample in each cluster. Moreover, in each iteration, it will print result of the silhouette score for each number of clusters.

Text

Description automatically generated

Figure 32: Silhouette Score Output

Based on the silhouette analysis score, it shows that the most appropriate number of clusters for the K-Means algorithm are 3 clusters, with the average of silhouette coefficient fall at 0.4457. As a result of analysing both evaluation metrics, the analyst can decide that the most appropriate number of clusters for the model are 3 which has been supported by 2 evaluation metrics. Therefore, the final K-means model will be using n\_clusters = 3 as the total number of customer clusters based on the evaluation metrics.

Graphical user interface, text, chat or text message

Description automatically generated

Figure 33: Declaring Final K-Means Model

Figure (Final Model)

Table

Description automatically generated

Figure 34: Adding Cluster Label to the Data Frame

Upon the completion of final modelling, the analyst must update the **rfm** data frame to add the cluster label that can be used to define each customer belongs to which number of clusters.

## 5.8 Summary

in summary, this section contains the process on how the analyst analyze and developed the program through the dataset of a retail store for sales and marketing department in retail industry through Python programming.

This section also shows on how the analyst performed the initial data exploration about the dataset such as descriptive statistics for each variable, checking the value of each variable if it is containing missing value or not, as well as checking the data type of each variable.

Moreover, this chapter also show on how the analyst prepare the dataset for modelling through cleansing the missing values, and method used to deal with the outlier value. Upon cleansing the dataset, this chapter also shows various techniques implemented by the analyst to transform the dataset into appropriate format for the model development namely from changing data type, creating new variables, as well as reducing the number of observations in the datasets.

Not limited to that, this chapter also show the visualization on what are the top 10 product sold, stock code used, customers transactions frequencies, and historical demand based on the historical transaction.

Lastly, this chapter also contain on how the analyst built the K-Means algorithm model through elbow method and silhouette analysis as the evaluation metrics to determine the most appropriate number of clusters for the customers.

# CHAPTER 6: Results and Discussion

## 6.1 Introduction

In this chapter, the analyst provided the analysis result and suggestions for the retail store based on the characteristic of their customers. Hence, it will help the retail store to leverage their sales and make use of marketing budget effectively and develop customer retention among their customers.

## 6.2 Results and discussion

Table

Description automatically generated

Figure 35: Average of Recency, Frequency, and Monetary in Each Clusters

From the result of two evaluation metrics that shows that the most appropriate number of clusters are 3 clusters, it allows the analyst to retrieve information on what are the characteristics of each cluster based on the mean of the Recency, Frequency, and Monetary.

Cluster with low recency are customers who interacted with the retail recently, and more likely to respond to new marketing campaign. Moreover, cluster with high recency are customers who are at churn-risk. Lastly, cluster with high frequency are considered customers that has high loyalty with the retail.

The higher the recency value for a cluster, the higher the risk of customer churn from the store. Hence, Cluster 0 are customers who has high Recency score which means that this type of customers is at churn risk. This type of customer are customers who has not make any recent transaction for quite long time. Therefore, to retain this type of customers marketing strategy like discount pricing, or exclusive offers can be used for this cluster.

Moreover, customers belong to cluster 1 are considered loyal customer however not generated that much of revenue. Therefore, this type of customer tends to respond towards product recommendation based on their historical purchase. Also, the retail store can introduce a 'customer level' based which will give customer gift based on their spending, each level will get different gift and as the customer level increase, the value of the gift is also higher. Moreover, giving these customers free postage within terms and condition can also be part of the marketing strategy that can be opted to retain this cluster.

Furthermore, customers belong to cluster 2 are customers who are willing to spend high amount of money, this type of customer can be used to offer luxurious product. not limited to that, the retail can also come up with subscription-based customer from this type of customer. Moreover, it is also conceivable to deny this sort of consumer a discount because they are eager to spend money in the store.

Chart

Description automatically generated

Figure 36: Cluster Distribution

Based on the cluster distribution, it shows that most of the customers are belong to cluster 1 with the proportion of 52.85%, followed by customers in cluster 0 with the proportion of 25.48%. Furthermore, the remaining customers are belonging to cluster 2 with the proportion of 21.67%, in which it shows that the analyst able to retrieve insight as per the pareto principle, that tells that 80% of the revenue are generated by the 20% of the customer.

# CHAPTER 7: Conclusions and Reflections

To conclude this paper, the analyst has gain lot of knowledge through the development of the program also through the domain research that needs to be done on earlier stage. Consequently, it allows the analyst to achieve the objective of the project, which are to pre-process and analyze the dataset, to segment customer based on Recency, Frequency, and Monetary value, and to provide suggestions based on the analysis result.

The future enhancement that can be done with regards to this paper are to provide sales predictive-analytics, recommendation engine, and predicting changes in customer behaviour. However, one of the limitations of this paper is the lack of resource paper that focusing on data mining approach for retail industry. Thus, further study in this topic can be done.

# REFERENCES

Agrawal, S. (2021, January 5). *Introduction to Matplotlib and Seaborn*. Analytics Vidhya. <https://medium.com/analytics-vidhya/introduction-to-matplotlib-and-seaborn-e2dd04bfc821#:~:text=Matplotlib%3A%20Matplotlib%20is%20mainly%20deployed>

*An introduction to seaborn — seaborn 0.11.2 documentation*. (2022, March 11). Seaborn.pydata.org. Retrieved March 11, 2022, from <http://seaborn.pydata.org/introduction.html#:~:text=Seaborn%20is%20a%20library%20for>

Analytics Vidhya. (2015, January 5). *Scikit-Learn In Python - Important Machine Learning Tool*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/>

Building and Applying Predictive Models in IBM SPSS Modeler training webinar. (2017). Smart Vision - Europe. <https://www.sv-europe.com/crisp-dm-methodology/>

Carraz, M. (2022). *sate of the developer nation*. Slashdata-website-cms.s3.amazonaws.com. Retrieved 11 March 2022, from <https://slashdata-website-cms.s3.amazonaws.com/sample_reports/ZAamt00SbUZKwB9j.pdf>.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Step-by-step data mining guide. <https://the-modeling-agency.com/crisp-dm.pdf>

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197–208. <https://doi.org/10.1057/dbm.2012.17>

Codeacademy.com. (2021). *Learn the Basics of Machine Learning: Clustering: K-Means Cheatsheet*. Codecademy. https://www.codecademy.com/learn/machine-learning/modules/dspath-clustering/cheatsheet

DATA SCIENTIST 2017 BROUGHT TO YOU BY. (n.d.). <https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf>

Garcke, J., Griebel, M., & Thess, M. (2010). Data Mining for the Category Management in the Retail Market. Production Factor Mathematics, 81–92. <https://doi.org/10.1007/978-3-642-11248-5_5>

Georgian, B. (2021). *Market Basket Analysis in Retail Industry | CPG Analytics*. Applexus.com. Retrieved 11 March 2022, from <https://www.applexus.com/blogs/market-basket-analysis-in-retail-and-cpg-analytics-to-increase-revenue>.

Kaggle.com. (2017). *E-Commerce Data*. Www.kaggle.com. https://www.kaggle.com/datasets/carrie1/ecommerce-data

Koptelov, A. (n.d.). Is ML-Powered Price Optimization Worth it? Total Retail. Retrieved March 11, 2022, from <https://www.mytotalretail.com/article/pricing-optimization-with-machine-learning-is-it-worth-it/>

Li, Y., & Wu, H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, *25*, 1104–1109. <https://doi.org/10.1016/j.phpro.2012.03.206>

‌*Matplotlib — Visualization with Python*. (n.d.). Matplotlib.org. Retrieved March 11, 2022, from <https://matplotlib.org/#:~:text=Matplotlib%20is%20a%20comprehensive%20library>

Mukherjee, S., Mukherjee, T., & Nath, A. (2016). *Sales Prediction Model Using Classification Decision Tree Approach For Small Medium Enterprise Based on Indonesian E-Commerce Data*. researchgate.net. Retrieved 10 March 2022, from <https://www.researchgate.net/publication/304652888_Scope_and_Challenges_in_the_Retail_Industry_Application_of_Data_Science>.

*Package overview — pandas 1.1.5 documentation*. (n.d.). Pandas.pydata.org. <https://pandas.pydata.org/docs/getting_started/overview.html>

(PDF) Sales Prediction Model Using Classification Decision Tree Approach For Small Medium Enterprise Based on Indonesian E-Commerce Data. (n.d.). ResearchGate. Retrieved March 11, 2022, from <https://www.researchgate.net/publication/349787576_Sales_Prediction_Model_Using_Classification_Decision_Tree_Approach_For_Small_Medium_Enterprise_Based_on_Indonesian_E-Commerce_Data>

*Python datetime module*. (2019, November 14). GeeksforGeeks. <https://www.geeksforgeeks.org/python-datetime-module/#:~:text=Python%20Datetime%20module%20supplies%20classes>

*Python vs R vs SAS | Which Data Analysis Tool should I Learn?* (2017, September 11). Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/>

Python, R. (2021). *K-Means Clustering in Python: A Practical Guide – Real Python*. Realpython.com. <https://realpython.com/k-means-clustering-python/>

Python.org. (2019). *What is Python? Executive Summary*. Python.org; Python.org. <https://www.python.org/doc/essays/blurb/>

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack. Journal Of Open Source Software, 3(24), 638. <https://doi.org/10.21105/joss.00638>

Ravindra Savaram. (2018, May 3). *Python VS Sas Vs R*. Mindmajix; Mindmajix. <https://mindmajix.com/python-vs-sas-vs-r>

Russell, C. (2020, April 24). *Worst Ever Decline In U.K. Retail Sales Due To Coronavirus*. Forbes. <https://www.forbes.com/sites/callyrussell/2020/04/24/worst-ever-decline-in-uk-retail-sales-due-to-coronavirus/?sh=6a7d508c1b65>

Sabanoglu, T. (2022, March 31). *Retail sales value monthly growth Great Britain 2021*. Statista. https://www.statista.com/statistics/287171/monthlyl-percentage-change-retail-sales-value-great-britain/

*SAS vs R vs Python - The Battle for Data Science!* (2020, February 24). TechVidvan. <https://techvidvan.com/tutorials/sas-vs-r-vs-python/>

Science, O.-O. D. (2020, July 15). *Why You Should be Using Jupyter Notebooks*. Medium. <https://odsc.medium.com/why-you-should-be-using-jupyter-notebooks-ea2e568c59f2#:~:text=The%20Jupyter%20Notebook%20is%20an>

Scientist, M. L., CEO and Chief Revenue. (n.d.). 8 Of Your Sales Team’s Biggest Challenges And How To Solve Them. Www.square2marketing.com. <https://www.square2marketing.com/blog/8-of-your-sales-teams-biggest-challenges-and-how-to-solve-them>

scikit-learn. (2019). *scikit-learn: machine learning in Python*. Scikit-Learn.org. https://scikit-learn.org/stable/

Tardi, C. (2019). *How to Apply the 80-20 Rule*. Investopedia. <https://www.investopedia.com/terms/1/80-20-rule.asp>

Team, G. L. (2020, November 5). Why using CRISP-DM will make you a better Data Scientist? GreatLearning. <https://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist/>

Techvidvan, T. (2022). https://techvidvan.com/tutorials/sas-vs-r-vs-python/. Retrieved 11 March 2022, from <https://techvidvan.com/tutorials/sas-vs-r-vs-python/>.

Wang, J. (2003). Data Mining. Google Books. Retrieved 11 March 2022, from <https://books.google.com.my/books?hl=en&lr=&id=mm32sNYpmOkC&oi=fnd&pg=PR7&dq=data+mining+approach+literature+review&ots=lHmdF0h2Qx&sig=f9IxihhPn7rbJMIza2Eay2qMPW4&redir_esc=y#v=onepage&q=data%20mining%20approach%20literature%20review&f=false>.

*What is NumPy? — NumPy v1.19 Manual*. (n.d.). Numpy.org. <https://numpy.org/doc/stable/user/whatisnumpy.html>

‌

‌

‌

‌

# APPENDICES

Turnitin Report First Page

Graphical user interface, text, application

Description automatically generated

Figure 37: First Page of Turnitin Report

Second Page of Turnitin Report

Text

Description automatically generated

Figure 38: Second Page of Turnitin Report

Graphical user interface, text

Description automatically generatedFYP poster

Figure 39: FYP Poster

Confidentiality document

Text, letter

Description automatically generated

Figure 40: Confidentiality Document

Library Catalogue

Text, letter

Description automatically generated

Figure 41: Library Cataloguing details

Log sheets

![Text

Description automatically generated with medium confidence]()

Figure 42: Project log 1

![Text, application

Description automatically generated]()

Figure 43: Project Log 2

PPF

![Text, letter

Description automatically generated]()

Figure 44: Page 1 PPF

![Text, letter

Description automatically generated]()

Figure 45: Page 2 PPF

![Text, letter

Description automatically generated]()

Figure 46: Page 3 PPF

![Text, letter

Description automatically generated]()

Figure 47: Page 4 PPF

![Text, letter

Description automatically generated]() PSF

Figure 48: Page 1 PSF

![Text, letter

Description automatically generated]()

Figure 49: Page 2 PSF

![Text, letter

Description automatically generated]()

Figure 50: Page 3 PSF

![Text, letter

Description automatically generated]()

Figure 51: Page 4 PSF

![Text

Description automatically generated]()

Figure 52: Page 5 PSF

![Text

Description automatically generated]()

Figure 53: Page 6 PSF

![Text, letter

Description automatically generated]()

Figure 54: Page 7 PSF

![Text, letter

Description automatically generated]()

Figure 55: Page 8 PSF

![Text, letter

Description automatically generated]()

Figure 56: Page 9 PSF

Ethics form

![Table

Description automatically generated]()

Figure 57: Ethic Form Page 1

![Table

Description automatically generated]()

Figure 58: Page 2 Ethic Form

![Text, letter

Description automatically generated]()

Figure 59: Page 3 Ethic Form

![Text, letter

Description automatically generated]()

Figure 60: Page 4 Ethic Form

![Text, letter

Description automatically generated]()

Figure 61: Page 5 Ethic Form

Gantt Chart for the Whole Final Year Project



Figure 62: Gantt Chart for the Whole Final Year Project