



Market Outreach for Retail Supermarkets through Customer Segmentation

**A dissertation submitted for the Degree of Master of
Business Analytics**

**M.T.I. PERERA
University of Colombo School of Computing
2021**



Market Outreach for Retail Supermarkets through Customer Segmentation

**M.T.I. Perera
2021**

Declaration


The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: M.T.I. Perera

Registration Number: 2019/BA/027

Index Number: 19880278


Signature:

Date: 22/02/2023


This is to certify that this thesis is based on the work of

~~Mr.~~/Ms. M.T.I. Perera

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. L.N.C. De Silva


Signature:

Date: 22/02/2023

Abstract

This research is built around applying Machine Learning technologies to the supermarket retail sector in Sri Lanka. Two areas were identified for the study: Customer Segmentation for the application of Unsupervised Clustering algorithms and Market Basket Analysis for the application of Association Rule Mining.

The main aim of the research was to identify the different clusters of customers found within the supermarket retail domain of Sri Lanka. To facilitate this, it first required the collecting and analyzing of the POS (Point of Sale) sale data in combination with the customer information. Access to this information was provided by the Keells Supermarket chain, owned, and maintained by Jaykay Marketing Pvt. Ltd which is a Part of the John Keells Group of Companies. They provided limited access to the relevant information and the Nexus Customer Loyalty Program which contained most of their customer data. The data were analyzed in their entirety and various derivative forms yielding diverse results. In the clustering process several clustering algorithms were applied, K-Means, K-Modes, KPrototypes, DBSCAN, and Mean Shift algorithms were some of the successfully tested algorithms. They provided diverse outcomes, some with very clear clusters and others without any coherent meaning. There were also instances where an algorithm could not deliver a clear and coherent outcome with the main dataset but would give a viable result for one of the derived datasets.

The Association Rule Mining (ARM) process considered the Apriori and Frequent Pattern Growth (FP Growth) algorithms are two of the most popular ARM algorithms used today. The outcomes of these algorithms were able to provide consistent association rules between products through tests on different samples of data.

Based on the finding it was successfully concluded that it is indeed possible to apply Clustering to the retail industry in a customer segmentation capacity, albeit the viability of the outcomes may differ based on the requirement and mode of application. Great potential can be found in the application of the findings of both Clustering and ARM in customer attraction and retention. It opens a new frontier for building customer value.

Acknowledgment

Many people's encouragement and inspiration helped to make this research possible, as well as their valuable time, advice, and labor. Before anything else, I would like to express my sincere gratitude to Dr. Lasanthi De Silva my research supervisor from the University of Colombo School of Computing, for her constant direction, inspiration, support, and insightful criticism that inspired me to further my research throughout the research period.

I must also acknowledge Mr. Osanda Warnakulasooriya Head of IT JMSL, Mr. Kasun Hiththathiyage, and Mr. Neomal Perera Chief Operating Officer JKIT for granting me access to classified and sensitive business information for my thesis. I would not have been able to complete the research if not for their trust and generosity.

Finally, I thank my friends, family, and all others who motivated and supported me over the entire period to make the research a success.

Table of Contents

Abstract.....	iii
Acknowledgment.....	iv
List of Figures.....	vii
List of Tables.....	viii
1 Introduction.....	1
1.1 Introduction to the problem	1
1.2 Problem Statement.....	3
1.3 Goals and Objectives	3
1.3.1 Goal	3
1.3.2 Objectives	4
1.4 Scope of the study.....	4
1.4.1 In scope.....	4
1.4.2 Out of Scope	4
1.5 Feasibility Study	5
1.5.1 Technical Feasibility.....	5
1.5.2 Resource Feasibility	5
1.5.3 Legal and Ethical Feasibility	6
1.6 Structure of the Dissertation	6
2 Background & Related Work	7
3 Literature Review	10
3.1 Theories	10
3.1.1 Customer Relationship Management (CRM).....	11
3.1.2 Market Segmentation.....	11
3.1.3 Clustering Algorithms	12
3.2 Critical Review	18
3.3 Gap Analysis.....	19
3.3.1 Shortcomings of Mass Marketing	19
3.3.2 Direct Marketing through customer segmentation	20
4 Design & Methodology	21
4.1 Proposed Approach & Methodology	21
4.2 Research Solution Design.....	24
4.2.1 Solution Components	24
5 Results and Evaluation.....	27
5.1 Data Pre-processing	27
5.1.1 Load Dependencies and Configuration Settings / Load Dataset	27
5.1.2 Exploratory Data Analysis (EDA).....	28

5.2	Hopkins Statistics	31
5.3	RFM Modeling	32
5.3.1	Clustering the Segments	32
5.3.2	Clustering Performance Evaluation metrics	40
5.4	DBSCAN Algorithm.....	41
5.5	Association Rule Mining	43
5.5.1	Apriori Algorithm.....	43
5.5.2	Frequent Pattern Growth Algorithm.....	45
6	Discussion	46
6.1	Challenges.....	46
6.1.1	Data Integrity & Consistency	46
6.1.2	Large Data Volume	47
6.1.3	Limited Computational Power	47
6.2	Application of Algorithms	47
6.3	Assumptions.....	48
6.4	Conclusion	48
7	References.....	49

List of Figures

Figure 4.1: Keells Data Set.....	22
Figure 4.2:Proposed process.....	24
Figure 5.1: No of Lines in the dataset	27
Figure 5.2:Info about the data set	28
Figure 5.3:Summary Statistics.....	28
Figure 5.4: Data Status	29
Figure 5.5:Null Value data	29
Figure 5.6:Sales by Department	30
Figure 5.7: Code for getting Top Products sold in Supermarket.....	30
Figure 5.8:Top Products sold in Supermarket.....	31
Figure 5.9:Hopkins Statistics calculation	31
Figure 5.10:Recency	32
Figure 5.11:Recency Summary Statistics.....	32
Figure 5.12:Plotting the Recency Distribution and QQ-plot.....	33
Figure 5.13:Recency Distribution and QQ-plot	33
Figure 5.14:Recency Summary by Customer.....	34
Figure 5.15: Frequency Distribution and QQ-plot	34
Figure 5.16:Frequency Summary by Customer.....	35
Figure 5.17:Total Value Distribution and QQ-Plot.....	35
Figure 5.18:Customer History Records	36
Figure 5.19:Customer History Summary Statistics	36
Figure 5.20:Scaled Data	36
Figure 5.21:Receny and Frequency Plots	37
Figure 5.22:Code for Cluster selection.....	38
Figure 5.23:Elbow Method.....	39
Figure 5.24:Cluster Center Analysis	39
Figure 5.25:Heat Map for the Clusters	40
Figure 5.26:Silhouette Analysis	40
Figure 5.27:Performance Metrics Calculation	41
Figure 5.28:DBSCAN Execution	42
Figure 5.29:Determining the Epsilon value (eps).....	42
Figure 5.30:DBSCAN Clusters	43
Figure 5.31:Silhouette Score	43
Figure 5.32:Apriori Algorithm Execution.....	44
Figure 5.33:Apriori Algorithm – Rules set I	44
Figure 5.34:Apriori Algorithm – Rules Set II	44
Figure 5.35:Frequent Pattern Growth Algorithm Execution.....	45
Figure 5.36:Frequent Pattern Growth Algorithm - Rules Set I.....	45
Figure 5.37:Frequent Pattern Growth Algorithm - Rules Set II.....	46

List of Tables

Table 1.1 Customer Segmentation	2
Table 1.2 Plan vs. Actual Study	5
Table 5.1:Column Headers of Data Set	27

1 Introduction

1.1 Introduction to the problem

The current retail environment is very challenging and competitive because of the market variety, price change pressure from discounts, increasing price transparency in the industry, and the competition between the companies. The traditional approaches for differentiating strategic pricing and product-related promotions are not effective anymore in the retail industry. Within the competitive nature, the importance of treating customers as the company's main asset increases the organization's value. As a result, many companies invest time and money in developing strategies for better customer maintenance and long-term customer retention. Customer relationship management (CRM) has received more attention because it is an exhaustive process of acquiring and retaining customers, to reach maximum business value. (Ngai et al., 2009)

One of the essential objectives of CRM is customer development through customer insights. Analyzing the customer values for better insights is covered in CRM using the analytical approach by assessing the customer information. Businesses tend to change their business models by engaging in change management along with information technology solutions that help them to acquire new customers and retain the existing customer base and attract new customers by maintaining lifelong customer value. CRM is divided into four dimensions: Customer identification, customer attraction, customer retention, and customer development. Based on four of these categorizations, segmentation is the first step in CRM to identify customers. (Stone et al., 1996)

Customer segmentation is dividing the customer base into groups of similar individuals based on different aspects for marketing purposes. Companies need to obtain a better understanding of their customer's interests and demands and model the segments as in table 1.1 to determine how each category will bring value so that marketing materials can be more precisely tailored to that segment. (HubSpot, n.d.)

The most useful technique in business analytics for customer segmentation is clustering the customers with similarities and behaviors and the customers are grouped into homogeneous clusters. The clustering technique identifies the internally homogeneous and externally

heterogeneous groups. Customers are varied based on their needs, behavior, wants, and characteristics. The goal of clustering is to identify the customer groups and segment the customer base into clusters to align the target marketing processes can be aligned more efficiently.

SEGMENTATION MODEL	HOW TO SEGMENT CUSTOMERS
Demographic Segmentation	Age, gender, income, education, and marital status
Geographic Segmentation	Country, state, city, and town
Psychographic Segmentation	Personality, attitude, values, and interests
Technographic Segmentation	Mobile-use, desktop-use, apps, and software
Behavioral Segmentation	Tendencies and frequent actions, feature or product use, and habits
Needs-Based Segmentation	Product/ service must-haves and needs of specific customer groups
Value-Based Segmentation	The economic value of specific customer groups in the business

Table 1.1 Customer Segmentation

Since the customers play a vital role in profit making in organizations, a detailed understanding of the customer is more important to utilize the personalized experience. The availability of a huge amount of Transactional data on customer purchases allows for retrieving the buying patterns of the customers and personalizing the marketing approaches. Market basket analysis (MBA) is the frequently used technique to find out buying patterns.

Market basket analysis focuses on identifying the associations between products the customers purchase. Understanding the customer intentions during the visit enables retailers to provide satisfactory services that are personalized according to their visit.

1.2 Problem Statement

Due to technological advancements, the evolution of customer food habits, and evolving purchasing expectations, the retail supermarket industry's behavior in Sri Lanka has altered during the last decade.(Mihirani Dissanyake, 2020) Customers are paying more attention to the worth of money as their shopping habits and lifestyles change dramatically. In addition, the emergence of new supermarkets and competitors in the retail supermarket industry has prompted all retail supermarkets to place a greater emphasis on customer happiness and retention.

The supermarkets always use traditional ways such as notice boards, promotion leaflets to reach out the customers, and dashboards to find out the answers to questions *"How much did each store earn during the last 06 months? What are the most selling products? What are products that have the longest shelf life?"*. These can be answered with the existing data that has been collected throughout the supermarket's lifespan. But the dashboards are giving limited in the sense that it misses the data and insights of the transactions. But today retailers want to know *"Who are the customers and what are the defining traits of customers? If they bought product A, will they buy product X or can we predict what customer wants to buy?"*.

In this project, the transaction data collected from the Keells supermarket will be used to identify the consumer segments using data mining algorithms and identify the customers' buying patterns using market basket analysis to recommend products and personalized promotions.

1.3 Goals and Objectives

1.3.1 Goal

The main goal of this project is to study how to increase the basket size of active customers with personalized promotions by increasing the visit frequency based on the buying patterns and reactivating the lapsed customers.

1.3.2 Objectives

The main objectives of this study are to.

- Identify the data attributes that can be used to segment the retail customers
- Identify the suitable algorithms that can be used to segment the retail consumers
- Identify distinct homogeneous groups of customers
- Identify the customer's shopping trends and patterns to introduce personalized marketing which is a marketing technique used for one-to-one marketing or individual marketing with digital technology to deliver individualized messages and product offering to current or prospective customers.

1.4 Scope of the study

1.4.1 In scope

1. Identify the suitable algorithm that can be used to segment the retail consumers and segment the customers of Keells supermarket.

Using several clustering algorithms with the RFM model, the suitable algorithm for Keells customer data segmentation will be selected and segmenting the customers using the selected algorithm.

2. Analyze the customer buying trends and patterns using the Identified distinct homogeneous groups of customers and recommend the products and personalized promotions.

Finding the buying patterns of the customers from the transaction data, using the Market Basket Analysis with Association Rule-Mining, and recommending the products and personalized promotions based on the buying patterns and customer segments.

1.4.2 Out of Scope

Forecasting the inventory requirement based on the customer segments and buying behavior.

1.5 Feasibility Study

1.5.1 Technical Feasibility

The main technologies and tools that are associated with this project are.

- Python Libraries
- Clustering algorithms
- Jupyter Notebook

These technologies are freely available, and the technical skills required are manageable. The ease of segmenting using these technologies is synchronized.

The large volume of data requires an equally large volume of memory and high processing power. Therefore, during the implementation stage of this project, necessary steps and adjustments will be made to accommodate the processing capability of the computer hardware that is available on hand. While it is understood that this may have some adverse effects on the veracity of the outcome utmost care will be taken to minimize the impact of such scaling.

1.5.2 Resource Feasibility

Resources required for this project include;

- Data set – Keells Supermarket data will be used in this project
- Programming device – Laptop
- Programming Languages – Python language with Libraries feely available

The required resources are available for the project.

Area	Plan	Actual
Retail Chain	One of Sri Lanka's Top Supermarket retailers with over 100 active outlets	JayKay Marketing (Pvt) Ltd.
Target Demography	Customers who are a part of the Loyalty program, and whose data is available	Customers who are part of the Nexus loyalty program
Geographical Positioning	Top 5 outlets within the heart of Colombo ranked by footfall	The bulk of outlets regardless of footfall
Data Collection Timeline	24 Months	1 Month

Table 1.2 Plan vs. Actual Study

1.5.3 Legal and Ethical Feasibility

The customer data has been collected based on their consent and the Keells data set will not be disclosed to the public because of company policy.

The software used in the project is free and open source, therefore the risk of software legal issues that can arise is minimum.

1.6 Structure of the Dissertation

The document from this point onward is organized under the following sections.

- Chapter 2 - Background Study

This chapter provides the background knowledge, related terms, and concepts required for an information system professional with basic computer science literacy to understand the context and the project.

- Chapter 3 - Literature Review

From this chapter, a comprehensive overview of the previous studies done which relate to my project will be presented. Furthermore, based on the literature review findings, identified research gaps will be stated in this section.

- Chapter 4 - Design & Methodology

This chapter contains the details of the followed research design and methodology by the researchers for answering the formulated research questions and desired research objectives. Under this section, details about data collection, justifications for the selected data sources, data analysis and implementation details, and the path of the project process will be described.

- Chapter 5 - Results and Evaluation

The evaluation chapter details how the evaluation process of this project was carried out and the findings of the evaluation. The chapter also details the summary of the statistical analysis performed for the obtained results from the final evaluation.

- Chapter 6 – Discussion

This chapter describes the summary of the project findings and the conclusions of the project. Also, it describes what are the limitations identified and possible future works of this study as well as the project ethics which I followed throughout the project period.

2 Background & Related Work

This chapter includes previous studies done in this area. This existing research is divided into the following categories.

- K-means Clustering for customer segmentation
- Data mining techniques for customer relationship management

K-means Clustering for customer segmentation

1. Customer Segmentation using K-means Clustering

In the paper by Kansal et al. (Kansal et al., 2018), they proposed a k-means clustering algorithm for customer segmentation. In this study, Three different clustering algorithms (k-Means, Agglomerative, and Meanshift) were implemented for customer segmentation and k-Means clustering showed a higher result when compared to the results of clusters obtained from the other two algorithms. Also, 5 clusters have been set up labeled as Careless, Careful, Standard, Target, and Sensible customers. However, two new clusters emerged by applying mean shift clustering labeled as High buyers and frequent visitors and High buyers and occasional visitors.

2. An Empirical Study on Customer Segmentation by Purchase Behaviors Using RFM Model and K-Means Algorithm

Jun Wu et al. (Wu et al., 2020) proposed a K-means clustering algorithm to classify customers and make customer segmentation at the enterprise level using a model based on the recency, frequency, and monetary (RFM) factors. The success rate of the proposed K-means clustering algorithm was evaluated by proposing different CRM strategies accordingly. These K-means clustering-based strategies showed a success rate of an increasing number of active customers by 529 and purchase volume by 279%.

3. Customer Segmentation Using Clustering and Data Mining Techniques

The paper presented in 2013 by Kashwan et al. (Kashwan & Velu, 2013) suggests a k-means clustering technique and SPSS Tool to be used to develop a real-time and online system for a particular supermarket to predict sales in various annual seasonal cycles. In this study, they tested, a total of $n = 2138$ customers and divided them into 04 clusters using a K-means clustering algorithm.

4. CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY

The paper presented by Dogan et al. (Dogan, et al., 2016) proposes to segment 700032 customers using two different customer segmentation models. The first segmentation model was developed by using a two-step clustering method and dividing customers into three clusters. The proposed model two was developed by using k-means clustering method optimum value was tested under four different clusters to segment the customers. The clusters obtained by k-means cluster analysis are assigned according to their RFM score.

5. RFM model for customer purchase behavior using K-Means algorithm

This study was conducted by Anitha et al. (Anitha & Patil, 2019) aiming to apply business intelligence to identify potential customers by providing relevant and timely data to business entities in the retail industry. Based on the RFM model they have deployed dataset segmentation using the K-Means algorithm. The obtained clusters were validated using the calculation of the Silhouette Coefficient. RFM log was calculated for when they obtained 03 clusters and 05 clusters. The results show that the silhouette score matrix for $K = 5$ is less optimal compared to $K = 3$ and 03 clusters which are optimal for customer segmentation when using the RFM model along with the K-Means algorithm.

6. Approaches to Clustering in Customer Segmentation

In the paper by Tripathi et al. (Tripathi, et al., 2018) compared the performance of using K-Means clustering and hierarchical clustering for customer segmentation. The result shows K-Means clustering performs better for many observations, while hierarchical clustering is better at handling fewer data points. Also, their study pointed out that the major hindrance to using K-Means clustering is accurately selecting the number of clustering. However, when

considering the performance aspect, they suggest K-Means clustering is suitable for achieving better results.

7. Customer Segmentation in XYZ Bank using K-Means and K-Medoids Clustering

The paper presented by Aryuni al. (Aryuni, et al., 2018) suggests different clustering methods in unsupervised data mining techniques can successfully be used for customer segmentation. This research builds cluster models on customer profile data based on XYZ Bank's use of internet banking. Clustering methods use the K-Means method and the K-Medoids method based on the RFM scores of the customer's online banking transactions. The result shows that the K-Means method outperformed the K-Medoids method based on internal cluster (AWC) distances and the Davies-Bouldin index.

8. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining

The article published by Chen et al. (Chen, et al., 2012) shows customers in the retail industry can be segmented based on the Recency, Frequency, and Monetary (RFM) model using the k-means clustering algorithm and decision tree induction to identify the main characteristics of the consumers in each segment. For the study, they used a data set that includes demand records of UK-based online retailer items of 4070 individual items between 01/12/2010 and 09/12/2011.

9. Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services

Ezenkwu et al. (Ezenkwu, et al., 2015) 2015 used, a MATLAB program of the K-Means algorithm which trained using a Z-score normalized two-feature dataset of 100 training patterns acquired from a retail business for creating efficient customer segmentation. The results show K-Means algorithm can be used for customer segmentation with 95% accuracy with four clusters namely High-Buyers-Regular-Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low-BuyersRegular-Visitors (LBRV) and Low-Buyers-Irregular-Visitors (LBIV).

Data mining techniques for customer relationship management:

1. Study on Application of Customer Segmentation Based on Data Mining Technology

The paper presented by Gong et al. (Gong, et al., 2009) suggests customer segmentation as a key factor for customer relationship management. Furthermore, they suggested customer purchase patterns and other valuable customer knowledge can be identified using data mining techniques, which would be a practical guide for the practice of customer segmentation.

2. Market Basket Analysis Using Apriori and FP- Growth for Analysis Consumer Expenditure Patterns at Berkah Mart in Pekanbaru Riau

The research carried out by Mustakim et al. (Mustakim, et al., 2018) suggests Market Basket Analysis with an FP-Growth algorithm can be used to determine the layout and planning of goods availability of minimarkets in Pekanbaru city relevant to the customer's expenditure patterns. Experimental results show that the FP-Growth algorithm can quickly and efficiently analyze customer shopping patterns and increase market revenue.

3. Using Shopping Baskets to Cluster Supermarket Shoppers

In the paper by Brijs et al. (Brijs, et al., 2004) propose data mining techniques such as the market basket analysis method can be used to identify behavior-based customer segmentation. This study uses loyalty card data as the main segregation factor to segment customers regarding socio-demographic or lifestyle characteristics. The results show several segments that vary significantly from the average purchase rate within a pre-determined set of product categories can be identified.

3 Literature Review

3.1 Theories

Understanding the behavior of consumers has been gaining popularity over recent years. With major online retailers like Amazon, and eBay and even popular streaming platforms like Netflix employ data analytics to improve their revenue through engagement with their customers dramatically. (Karunaratna, 2021)

Many improvements and advancements have been made in the global retail industry in the past decade. Global brand presence, high levels of industry convergence, the doubling of the

number of internet users from 2.14bn in 2011 to 4.88bn in 2021 as well as breakthrough technologies such as IoT have greatly impacted consumer behavior making it incredibly difficult for a retailer to analyze and understand customer behavior (KEMP, 2021). This has placed high importance on building a solid relationship with the customer which will provide sellers with the information they need to provide customers with better products and services thereby increasing customer retention. The process of acquiring new customers as well as retaining both new and existing customers is known as Customer relationship Management or CRM (Khajvand et al., 2011).

3.1.1 Customer Relationship Management (CRM)

The term Customer Relationship Management has been around for almost thirty years. During that time several different definitions have been made for CRM. In a research paper published in 2000 CRM is defined as technology or software that can aid a business to gather data and information about their customers so that they can provide better products and services to their target demographics (“Principles of Managing Customer Experience and Relationships,” 2016). Alternatively, CRM can also be defined as a set of employee training and operating procedures that are geared towards bringing the customers closer to the business and getting to know them better to provide better value to them (Dachyar, M., and Vitasya, L, 2021).

There are three main types of CRM. The first type is Strategic CRM, which is considered the foundation of the CRM strategy. It is built around the goal of acquiring and retaining customers that are valuable to the business. Next, there are Operational CRM, and customer-related departments such as sales, marketing, and after-sales service fall into this category. Finally, there is analytical CRM which uses customer-related data at its core to gain valuable customer insights.

3.1.2 Market Segmentation

Market Segmentation can be defined as “dividing your market into specific groups of customers”. The outcome of this process is the creation of customer groups that are based on common interests, needs, behavior, or demographics. There are four main methods of segmentation.

3.1.2.1 Demographic Segmentation

One of the most common forms of segmentation, demographic segmentation refers to the splitting of a market based on observable demographic characteristics such as Gender, age, occupation, income, etc. For instance, demographic segmentation can be used to target customers based on gender, thereby allowing the advertising team to design ads that are more relevant to the corresponding gender.

3.1.2.2 Psychographic Segmentation

Psychographic segmentation investigates the customer's interests and personality traits. Examples of this would be hobbies, lifestyle, goals, beliefs, etc. However, unlike demographic segmentation, this form is harder to identify. This makes it critically important that proper background research is carried out before the implementation of this form. A classic example would be a promotional campaign targeting cost-conscious shoppers or bargain hunters.

3.1.2.3 Geographic Segmentation

As the name suggests geographic segmentation is focused on the geography of the market. This can be expanded to include country, province, district, and even postal codes. A good example of this would be a promotion or marketing campaign targeting consumers living in a district.

3.1.2.4 Behavioral Segmentation

Understanding customer behavior is paramount, therefore, behavioral market segmentation can be considered the most useful for the supermarket retail industry. This approach can be especially useful when looking at how customer behavior is towards a product or service. The behavioral segmentation method can be used to group customers based on spending habits, buying habits, brand loyalty, and even browsing habits.

3.1.3 Clustering Algorithms

Clustering is the process of dividing a population (in this case a market) into groups based on similarities that are not shared with other groups. In other words, it is to assign elements with the same properties into distinct groups or clusters.

In machine learning, Clustering is considered a form of unsupervised learning. (Witten and Frank, 2006). It is best applied when there are no preexisting groups to categorize the data points into. If the observations/data points are plotted on a feature space, an area of high data point density compared to other areas in the space can be considered a cluster (Murphy, 2021).

These clusters represent a grouping found within the domain the observations/data points are taken from. Clustering can be categorized into two sub-categories.

Hard Clustering

In this category, a data point/observation can either belong entirely to a cluster or not. (Murphy, 2012).

Soft Clustering

In soft clustering, data points/observations are clustered based on the probability of belonging to a particular group (Murphy, 2012)

There are many different clustering algorithms available for machine learning. The following are some of the most used algorithms.

K-Means Algorithm

K-Means can be considered the most popular clustering algorithm due to its speed, ease of use understanding, and availability. It is an unsupervised learning method, which means it has no labeled data. K-means follows an iterative process where the dataset is segmented into K number of predefined clusters. The “K” refers to the number of clusters that need to be created. For example, if there is a need for 5 clusters to be created, then the value of K will be Five.

As popular as it is, this algorithm has a few shortcomings. Notably, since we must specify a value for K and thereby tell the algorithm how many clusters we are looking for, it can be said

that K-means great as a “Partitioning” algorithm than “finding” clusters. This also requires that the users have a good understanding of the data. This means K-Means is not ideal for a scenario that consists of unfamiliar and new data. Another shortcoming of this algorithm is that it requires the entire dataset to be loaded into the system memory for the processing which also consumes large volumes of computational resources slowing down the clustering process. In a scenario where the dataset is extremely large using this algorithm can become impractical (Bergström, S., 2019; Kushwaha, Y. and Prajapati, D., 2018).

Mini-Batch K-Means

This algorithm is a derivation of the original K-Means algorithm. Unlike its predecessor the The mini-Batch K-Means algorithm does not require the entire dataset to be loaded into memory. Instead, it iteratively takes a small random sample and from the dataset, conducts the clustering. As the sample size is small, it does not consume a lot of computing resources and memory to process. In situations where the dataset is extremely large, or the processing of clusters is time-sensitive, Mini-Batch K-Means is the ideal option. However, it must be noted that aside from the high memory consumption and long processing time, the other shortcomings of K-Means algorithms are still present (Béjar Alonso, 2013; Rachman et al., 2021)

K-Modes

Algorithms like K-Means utilize mathematical calculations to measure the distance between clusters of continuous data. However, when the data contains categorical data, it cannot perform its function. K-Modes is an unsupervised machine learning algorithm that specializes in clustering categorical variables. Unlike K-means, K-Modes make use of the differences between data points. The idea is that the lesser the number of differences the greater the similarities are between data points. This algorithm relies on Modes instead of Means (Ayat et al., 2001).

K-Prototypes

K-Means is the go-to algorithm for clustering large volumes of data. However, one of the limitations of this algorithm is that it cannot process data with both categorical and numeric

data. This is mainly due to the use of Euclidean distances to calculate the distances between the clusters. The alternative for this is K-Modes can only process categorical data, which is the inverse of the problem faced by K-means. Enter the K-Prototypes Algorithm. An improvement of the already existing K-Means and K Modes algorithms, the K-Prototypes algorithm was developed to process mixed data types i.e., data with both categorical and numerical variables. The K-Prototypes model relies on partitioning to identify clusters (Ji et al., 2013).

Hierarchical Agglomerative Clustering (HCA)

Hierarchical clustering is an unsupervised learning algorithm that uses a tree-like structure to group data into clusters that are ordered from top to bottom. At the start, HCA considered every data point to be an individual cluster. Next, it iteratively picks the clusters that are the closest to each other and merges them. The outcome of hierarchical clustering is a tree-like structure depicting nested clusters known as a Dendrogram.

There are two main types of HCA, Agglomerative, and Divisive. In the Agglomerative approach also known as the “bottom-up” approach all the data points will have their cluster. The closest of these clusters will merge as we go up the hierarchies. The Divisive approach which is referred to as the “top-down” approach is an inverted form of the agglomerative approach. Here all the data points start in one cluster and data points that are dissimilar to the main cluster are separated forming their cluster (Cibulková, J. and Sulc, Z., 2018).

DBSCAN

Proposed in 1996, Density-based spatial clustering of applications with noise or DBSCAN is an unsupervised clustering algorithm. It is a density-based algorithm that is capable of distinguishing high-density areas from low-density areas. DBSCAN has one major advantage over popular models like K-means in which DBSCAN does not require the number of clusters to be given by the user. This makes it ideal for scenarios where there is little known about the dataset. One disadvantage of this algorithm is that it has difficulties in detecting clusters in data of varying densities (Ram et al., 2010).

Mean Shift

Mean shift is another unsupervised learning algorithm that is popular for clustering. This algorithm is most popular in the image processing domain. It relies on iteratively shifting the data points toward the regional mean. In other words, it assigns the datapoint to the closest cluster centroid in an iterative form. The determining of the centroid is done by identifying the highest concentration of data points nearby is at (Wu and Yang, 2007).

One of the key advantages of this algorithm over traditional K-Means is that does not require the number of clusters to be predefined. However, this also means that the Mean shift algorithm is highly dependent on computational power. Therefore, it is better to adopt this algorithm within an environment with sufficient computing resources.

Spectral Clustering

Spectral clustering is an unsupervised learning algorithm that has its roots in graph theory. Lately, it has become one of the most popular clustering algorithms due to its simplicity, efficiency, and its ability to outperform traditional clustering algorithms such as K-Means (von Luxburg, 2007)

Spectral Clustering functions by identifying groups of nodes, in a graph based on the edges connecting them. It is flexible enough to be able to cluster non-graph data as well. Spectral clustering uses the concepts of Eigenvalues and Eigenvectors which are derived from a special matrix built from the dataset. These Eigenvectors are used to assign data points to clusters.

Affinity Propagation (AP)

Initially published in 2007, Affinity Propagation is another form of hierarchical clustering and is based on Boruvka's MST algorithm. This algorithm considers each data point as a cluster and keeps sending messages between each one, merging the closest ones each iteration (Frey and Dueck, 2007). Unlike K-means or K-Modes, where the user defines the number of clusters, AP does not require a fixed number of clusters to be manually entered. Instead, the algorithm can be stopped once the desired number of clusters has been created.

OPTICS

OPTICS stands for Ordering points to identify the clustering structure. It is a derivative of the DBSCAN algorithm and can address the weaknesses of the DBSCAN algorithm. A unique feature that distinguishes OPTICS from other clustering techniques is that it does not divide data points into clusters. It instead provides a “reachability distant plot” which leaves it up to the researcher to interpret and identify the clusters (Ankerst et al., 1999)

BIRCH Clustering

In environments with limited resources such as memory and processing power, clustering algorithms like K-means find it difficult to perform efficiently. A potential alternative for such scenarios can be found in BIRCH clustering. Balanced Iterative Reducing and Clustering using Hierarchies or BIRCH is an unsupervised data mining algorithm. BIRCH operates by creating a smaller summary of the large original dataset and then clustering the summary in place of the original dataset. BIRCH is combined with other clustering models, where the summary generated by BIRCH is used as the input for the partner clustering algorithm. One of the most significant limitations found in BIRCH is that it can only process numeric attributes – also known as “Metric” attributes. BIRCH is incapable of processing categorical attributes (Fontanini and Abreu, 2018)

RFM/RFV Segmentation

RFM short for Recency, Frequency, and Monetary Value, is becoming a very popular form of segmentation in the retail industry. This is especially due to it being simple to implement with little help needed from data scientists and easy to interpret because of the intuitive nature of its output. The three core factors of RFM can be explained as follows:

Recency: When was the customer’s last purchase/visit?

Frequency: How often do they purchase a product/visit an outlet?

Monetary Value: How much do they spend on a brand/at an outlet?

Its ability to provide outputs such as “How are the biggest spenders?” or “Who visits the store the most?” has made it extremely popular with marketing departments due to its ability to

provide a high-level yet informative view of their customers (Yang, A.X. 2004; Shihab et al.,2019).

Association Rule Mining

Association rule mining is a very popular tool used in retail. It helps identify the relationship between different products within a shopping list/basket. It creates “rules” which will identify buying patterns – If a customer buys item A, then he/she will purchase item B. To this end, many algorithms have been developed. Namely, Apriori-based algorithms, tree-based algorithms, and a handful of other algorithms (Minho Ryu et al., 2021). In a 1996 study, it was found that association rule mining can be effectively used for cross-selling products to customers. Cross-selling is the process of identifying other products that can be sold alongside a targeted product (Nash and Sterna-Karwat, 1996). Two main challenges must be overcome when using association rule mining. First, this method cannot be used when introducing new products into the existing product catalog. This is because this method relies on the mining of historical data to extract the association rules. Without sufficient records in the database, it is impossible to generate rules of any substance. The second challenge is the number of combinations that are generated. This can be extremely difficult if there are thousands of products available on the list (Minho Ryu et al., 2021).

3.2 Critical Review

Market Segmentation benefits both retailers and consumers, some of the biggest advantages for retailers are mentioned below.

- The ability to identify new target customer groups – Market segmentation allows retailers to have a deeper and broader understanding of their customer base. It also helps to identify groups that may have been previously ignored in marketing efforts opening new venues for promotions.
- Improved marketing campaign performance – The output from market segmentation can be used to learn more about the target customer group allowing for the right messages to be sent to the right people at the right time. Thereby drastically increasing the success rate of a marketing campaign.

- Introduction of new products – Consumer feedback can help introduce new products to their stores.

Market segmentation is ideal for direct marketing campaigns. Empirical research has proved that direct marketing yields better results than mass marketing (Thomas and Housden, 2002). An effective direct marketing campaign can reap a host of benefits. It can build a personal relationship with the customer, which in turn builds loyalty and brand image. This has a vicarious effect of word-of-mouth marketing which is the most potent form of advertising which further amplifies sales (B. Stone and Ron Jacobs, 2008). They highlight several key factors that make direct marketing so effective, they are: (Thomas and Housden, 2002)

- **Targets ideal customers** – It is easy to identify the ideal target group of customers. Using customer segmentation. This allowed for tailor-made marketing and promotion campaigns to be launched which have a higher result in a sale.
- **Feedback from the relevant market** - Another benefit of this is that it provides feedback on customer response to the product or service from within the target domain. Which allows for updates and improvements to be made.
- **Cost-effective** – Having the ability to market to a smaller group of customers, frees up resources that can be put forward to designing an effective and relevant marketing campaign yielding better results.
- **Higher customer loyalty** – Direct marketing builds a close relationship with the customer. Personalized promotions, emails, offers, and discounts can create a strong bond with the customer which increases loyalty.
- **Creates new customers** – Having a loyal customer base increases promotion through word of mouth. This drives more consumers who are unfamiliar with the brand to try it out.

3.3 Gap Analysis

3.3.1 Shortcomings of Mass Marketing

The conventional method of advertising has been mass marketing. It was always conceived with the thought, "Here is what we have; come and buy it from us" (Liyanage, U., 2009). Regardless of the consumer's desire in buying the goods or service, it tries to push it to the mass market. Sri Lanka is a country where this behavior is quite prevalent, as seen by the posters covering the walls and the streetside freebies and booklets. Add to that the endless hours of advertising that interfere with our TV and movie watching, as well as the digital billboards that have started to blight some urban areas. The same is true of cyberspace, which features pop-up ads and other sponsored advertisements on websites. (Thomas, A.R. 2007).

Outside of earning the ire of consumers, academic research shows us that mass marketing has several disadvantages.

- High Cost of Implementation – airtime on TV and Radio can be prohibitively expensive
- Difficult to appeal to everyone – a bad first impression could turn away customers for good
- High level of competition – multiple vendors advertising the same type of products can saturate the market making it difficult to gain any market share.
- Little to no adaptability – changes in consumer characteristics are not considered

To counter this market research, in the form of customer segmentation has given rise to the idea of “Direct Marketing”

3.3.2 Direct Marketing through customer segmentation

As the name implies, direct marketing aims to promote goods and services directly to the target market. Retailers must first be able to assess and categorize the types of customers they have to accomplish this. Retailers may gain the advantage they need through market segmentation. This is especially true in Sri Lanka's grocery retail industry, where there is fierce competition, and every customer is given far greater weight.

The study aims to identify the various consumer groups that are present in the Sri Lankan market and give supermarkets the knowledge they require to carry out efficient advertising.

4 Design & Methodology

4.1 Proposed Approach & Methodology

This section explains the proposed implementation process. Overall research methodology can be divided into four main steps as follows.

1. Data preprocessing or data preparation and preprocessing
2. Analyzing the data using the RFM model
3. Customer segmentation using clustering algorithms
4. Identify and describe the customer buying patterns using market basket analysis

The transaction data set has been collected from Keells Supermarket Sri Lanka. The data set consists of the sales data for January in the year 2022.

Age	NexusNumber	Address_Line1	Address_Line2	Outlet_Code	InvoiceNumber	Date	Time	Item_Code	Item_Description	Department	Item_Category	Quantity	TotalValue
47	112105000045290	kottawa	Pannipitiya	S2HK	1298121	1/1/2022	00:00.0	2367	WATAWALA KAHATA TEA POUCH 400G	Beverages	Tea	1	500
47	112105000045290	kottawa	Pannipitiya	S2HK	1298121	1/1/2022	00:00.0	99367	EH GINGER BEER 1.5L	Beverages	Juices & Carbonates	1	220
NULL	1121050001089300	NULL	HOMAGAMA	S2HK	1298122	1/1/2022	00:00.0	10679	RICE SUPURI KEERI SAMBA BULK KG	Grocery	Pulses	1.24	291.4
47	112105000045720	Kottawa	S2HK	1298123	1/1/2022	00:00.0	923048	MANGO - K/C	Fruits	Inorganic		0.588	129.36
51	1121050000521350	Araliya Uyana	Pannipitiya	S2HK	1298126	1/1/2022	00:00.0	13281	NESCAFE ICE COLD COFFEE 180ML	Beverages	RTD Beverages	2	160
51	1121050000521350	Araliya Uyana	Pannipitiya	S2HK	1298126	1/1/2022	00:00.0	83951	RICHLIFE CHOCOLATE FLY.MILK 180ML	Beverages	RTD Beverages	2	140
51	1121050000521350	Araliya Uyana	Pannipitiya	S2HK	1298126	1/1/2022	00:00.0	951015	TOP CRUST BREAD	Production Bakery	Bakery	1	69
46	1121050000029340	Piliyandala	S2HK	1298128	1/1/2022	00:00.0	111875	EVA SANITARY NAPKIN DRYTEX WINGS 20S	Household	Personal Hygiene		1	320
46	1121050000029340	Piliyandala	S2HK	1298128	1/1/2022	00:00.0	112872	VELVET HAND WASH REFILL ROSE 200ML	Household	Hand & Body Care		1	175
46	1121050000029340	Piliyandala	S2HK	1298128	1/1/2022	00:00.0	121606	BABY CHERAMY SOAP POKURUWA/SADUN 5IN1	Household	Baby Needs		2	530
46	1121050000029340	Piliyandala	S2HK	1298128	1/1/2022	00:00.0	121740	MILO PACKET 600G	Beverages	Malt Drink		1	645
46	1121050000029340	Piliyandala	S2HK	1298128	1/1/2022	00:00.0	4191	VIM DISHWASH LIQUID LIME 500ML	Household	Cleaning Consumables		1	250
46	1121050000029340	Piliyandala	S2HK	1298128	1/1/2022	00:00.0	954	CLOGARD TOOTHPASTE 120G	Household	Oral Care		2	248
46	1121050000029340	Piliyandala	S2HK	1298128	1/1/2022	00:00.0	99367	EH GINGER BEER 1.5L	Beverages	Juices & Carbonates		1	220
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	100334	CLOGARD TOOTH PASTE 200G	Household	Oral Care	1	195
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	10406	SUNQUICK B/CURRENT STD 330ML	Beverages	Concentrated Fruit Drinks	1	490
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	1075	MYSOORE DHAL BULK KG	Grocery	Pulses	1.024	265.216
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	122789	NEW RATHNA ROSE RAW RICE 1KG (5U)	Grocery	Pulses	5	750
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	25029	CIC BASMATHI RICE BULK	Grocery	Pulses	2.024	698.28
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	4681	RICE SAMBA BULK KG - LOCAL	Grocery	Pulses	5.066	835.89
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	5228	HIGHLAND I/C VANILLA 1L	Frozen Food	Desserts	1	320
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	94007	KNORR SEASONING CUBE 15*20G	Grocery	Seasoning & Coconut Cream	1	50
33	1121050000313540	NULL	no	S2HK	1298129	1/1/2022	00:00.0	97445	KREST CHICKEN SAUSAGES S/LESS 250G	Frozen Food	Processed/Preserved Meat	1	375
56	1121050000214920	Mattegoda	S2HK	1298130	1/1/2022	00:00.0	105345	IMORICH CHOC CHOC CHIP 1L	Frozen Food	Desserts		1	775
56	1121050000214920	Mattegoda	S2HK	1298130	1/1/2022	00:00.0	117560	AMBEWELA FLAVOURED MILK VANILLA 1000ML	Beverages	Juices & Carbonates		1	320
56	1121050000214920	Mattegoda	S2HK	1298130	1/1/2022	00:00.0	120473	NIPUNA RICE SUPURI KEERI SAMBA 1KG (5U)	Grocery	Pulses		5	1225
56	1121050000214920	Mattegoda	S2HK	1298130	1/1/2022	00:00.0	121018	RITZBURY BLUEBERRY CHOCOLATE 45G	Grocery	Confectionery		1	120

Figure 4.1: Keells Data Set

The original dataset for the study will initially be chosen based on the RFM model's parameters. A new dataset will then be constructed after cleaning the previous dataset to remove outliers and erroneous numbers. The data are then changed into a more straightforward and effective to handle customer value analysis by removing superfluous attributes.

The RFM model will be a basic model for identifying customer groups because the collected dataset was restricted to sales records and did not include any other information about the customers. Three crucial informational qualities about each client will be determined by the RFM model using the transactions of a customer:

- Recency: The value of how recently a customer purchased at the supermarket
- Frequency: How frequent the customer's transactions are at the supermarket
- Monetary value: The value of all the transactions that the customer made at the supermarket

The consumers will then be divided up using clustering techniques, and the number of segments will be decided using a distance computation approach.

After determining the consumer categories, to offer targeted marketing that can be used to enhance sales by better understanding customer buying patterns, the market basket analysis, which incorporates association rule mining, will be used to evaluate transaction data, and determine the consumers' purchasing habits. It is a fundamental filtering technique that aids in anticipating and displaying the products that a user would like to buy in addition to the ones that they have previously bought. The transaction data, which includes the purchase history, will be analyzed using the Apriori algorithm and the Frequent Pattern Growth Algorithm to identify product groups and those that are most likely to be bought together. Searching for combinations of goods that regularly occur together in transactions is how algorithms operate.

There are two different kinds of market basket analyses:

I. **Predictive market basket analysis:** Purchased goods are considered to identify cross-sell which will be used in this study.

II. **Differential market basket analysis:** Considers information from various retailers as well as purchases made by various client groups at various times of the day, week, or year which will not consider in this study.

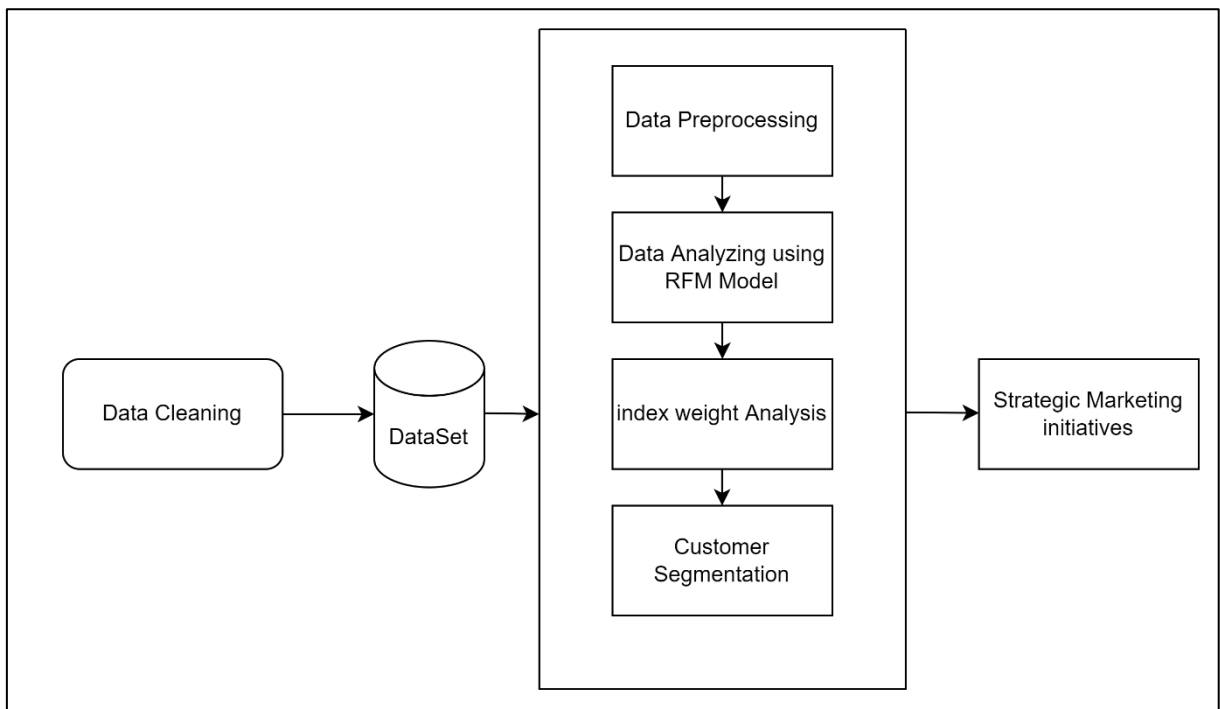


Figure 4.2:Proposed process

4.2 Research Solution Design

4.2.1 Solution Components

Data

With the business's consent, the transactional data was collected from the Keells supermarket. Exploratory data analysis will be used to clean up and aggregate the data collection.

Algorithms

Clustering algorithms will be used to segment the customers.

Primary Clustering Algorithms

- K-Means
- K-Modes
- K-Prototypes
- Hierarchical Agglomerative
- DBSCAN
- Mean Shift
- Spectral Clustering

Secondary Clustering Algorithms

- Affinity Propagation
- WARD
- OPTICS
- BIRD
- Gaussian

Determining the Number of Clusters

The scientific literature suggests several methods to determine the number of clusters which vary in complexity and application (Xu et al., 2016). They are:

- By rule of thumb
- Elbow method
- Information Criterion Approach
- An Information Theoretic Approach
- Choosing k Using the Silhouette
- Cross-validation

To ensure consistency with the number of clusters found for this research, I have chosen the tried-and-true Elbow approach and the Cross-Validation method. (Xu et al., 2016)

Evaluating Clustering Validity

The validity of the clusters produced by the clustering algorithms can be assessed using one of two main techniques. The Silhouette Measure and the Sum of Squared Errors are the names of these two techniques. (Aranganayagi and Thangavel, 2007).

The degree of cohesion and separation inside the clusters serves as the foundation for the Silhouette Measure. The value of the Silhouette Coefficient/Score varies from -1 to 1. The clusters have significant cohesiveness and separation when the value is closer to 1. Conversely, a score closer to -1 indicates that the clusters are inaccurate and have poor separation and cohesion.

Sum of Squared Errors (SSE) is the term used to describe the sum of squared errors between each observation inside a cluster and the cluster centroid.

Association Rule

The most common association rule mining algorithms are as follows.(Ghosh, Samarendra, 2014). This section aims to offer the outlet manager insights into what sort of products are bought frequently together. This will allow for better, more targeted promotions and advertising.

Rule Mining Algorithms

- Apriori
- Frequent Pattern Growth

Association Rule Evaluation Metrics

- Support - It is the frequency of occurrence of an itemset X and Y
- Confidence - This is a ratio that represents the total number of transactions of all of the items in {X} and {Y} to the number of transactions of the items in {X}.
- Lift -This is the ratio between the confidence of the rule and the expected confidence. In this scenario, it is believed that {X} and {Y} are independent of each other. The expected confidence can be calculated as the ratio between the confidence and the frequency of {Y}.

5 Results and Evaluation

5.1 Data Pre-processing

Data Set

The actual point-of-sale data that was used to build the dataset for this study was available in the JMSL information system. Due to the sensitive nature of the information, some information that could be used to identify a specific consumer has been removed.

Field	Description
Age	Age of the Customer
NexusNumber	The unique identification number is assigned to each customer
Address_Line1	Address Line 1
Address_Line2	Address Line 2
Outlet_Code	4 Digit alpha-numeric code used to identify supermarket outlet
InvoiceNumber	Numeric invoice number
Date	Date of invoice
Time	Time of the Transactions
Item_Code	Numeric code used to uniquely identify an item
Item_Description	Description of the item being sold
Department	The primary department the item belongs to
Item_Category	Subdepartment the item belongs to
Quantity	Quantity of the item purchased (Each, Kilograms, Liters, etc)
TotalValue	The cumulative monetary value of the item in the invoice

Table 5.1: Column Headers of Data Set

Total number of datapoints/records available for analysis: **6,895,444**

5.1.1 Load Dependencies and Configuration Settings / Load Dataset

Started with importing the python libraries and data set for the analysis as in Figure 5.1.

```
In [2]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

In [3]: df = pd.read_csv('C:/Users/thilinipe/Desktop/BA-PP/salesdata_all.txt', delimiter = "\t")
C:\Users\thilinipe\AppData\Local\Temp\ipykernel_18956\3134075197.py:1: DtypeWarning: Columns (5) have mixed types. Specify dtype option on import or set low_memory=False.
df = pd.read_csv('C:/Users/thilinipe/Desktop/BA-PP/salesdata_all.txt', delimiter = "\t")

In [6]: len(df)

Out[6]: 6895444
```

Figure 5.1: No of Lines in the dataset

5.1.2 Exploratory Data Analysis (EDA)

Performed exploratory data analysis to discover the information about the data set before starting the customer segmentation analysis.

```
In [7]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6895444 entries, 0 to 6895443
Data columns (total 14 columns):
#   Column          Dtype
---  ---
0   Age             float64
1   NexusNumber     int64
2   Address_Line1   object
3   Address_Line2   object
4   Outlet_Code     object
5   InvoiceNumber   object
6   Date            object
7   Time            object
8   Item_Code       object
9   Item_Description object
10  Department       object
11  Item_Category   object
12  Quantity        float64
13  TotalValue      float64
dtypes: float64(3), int64(1), object(10)
memory usage: 736.5+ MB
```

Figure 5.2: Info about the data set

The data consist of float, integer, and object values.

```
In [5]: df.describe()

Out[5]:
```

	Age	NexusNumber	Quantity	TotalValue
count	6.383474e+06	6.895444e+06	6.895360e+06	6.895360e+06
mean	4.567472e+01	1.121047e+15	1.254501e+00	2.904884e+02
std	1.303088e+01	1.630129e+12	2.276087e+01	4.017343e+02
min	1.600000e+01	1.121050e+14	-9.298000e+03	-2.350000e+04
25%	3.600000e+01	1.121050e+15	1.000000e+00	1.073600e+02
50%	4.400000e+01	1.121050e+15	1.000000e+00	1.947000e+02
75%	5.400000e+01	1.121050e+15	1.000000e+00	3.350000e+02
max	1.000000e+02	1.121058e+15	9.298000e+03	1.887500e+05

Figure 5.3: Summary Statistics

The fact that the Total Value and Quantity figures in the output in figure 5.3 are negative suggests that our data contains return transactions, which may be exaggerated. We will first look to see if there are any records where both are negative or if one is negative and the other is zero. This is because our focus is on market basket analysis and customer segmentation.

```

In [11]: print('Check if we had negative quantity and prices at same register:',
          'No' if df[(df.Quantity<=0) & (df.TotalValue<=0)].shape[0] == 0 else 'Yes', '\n')
print('Check how many register we have where quantity is negative',
      'and prices is 0 or vice-versa:',
      df[(df.Quantity<=0) & (df.TotalValue<=0)].shape[0])
print('\nWhat is the customer ID of the registers above:',
      df.loc[(df.Quantity<=0) & (df.TotalValue<=0),
             ['NexusNumber']].NexusNumber.unique())
print('\n% Negative Quantity: {:.2%}'.format(df[(df.Quantity<=0)].shape[0]/df.shape[0]))

Check if we had negative quantity and prices at same register: Yes

Check how many register we have where quantity is negative and prices is 0 or vice-versa: 31706

What is the customer ID of the registers above: [1121050001493650 1121050000369130 1121050000168340 ... 1121050000833030
1121050002058930 1121050001543650]

% Negative Quantity: 0.46%

```

Figure 5.4: Data Status

There are 31706 records where quantity and Total Value are negative, so those data are removed from the data set.

```

In [12]: print(df.isnull().sum())

Age                511970
NexusNumber        0
Address_Line1     5015175
Address_Line2     130426
Outlet_Code       84
InvoiceNumber     84
Date              84
Time              84
Item_Code         84
Item_Description  84
Department        84
Item_Category     84
Quantity          84
TotalValue        84
dtype: int64

```

Figure 5.5: Null Value data

All the null values have been removed from the final data set taken for the analysis.

Total number of datapoints/records taken for final analysis: 6,255,697

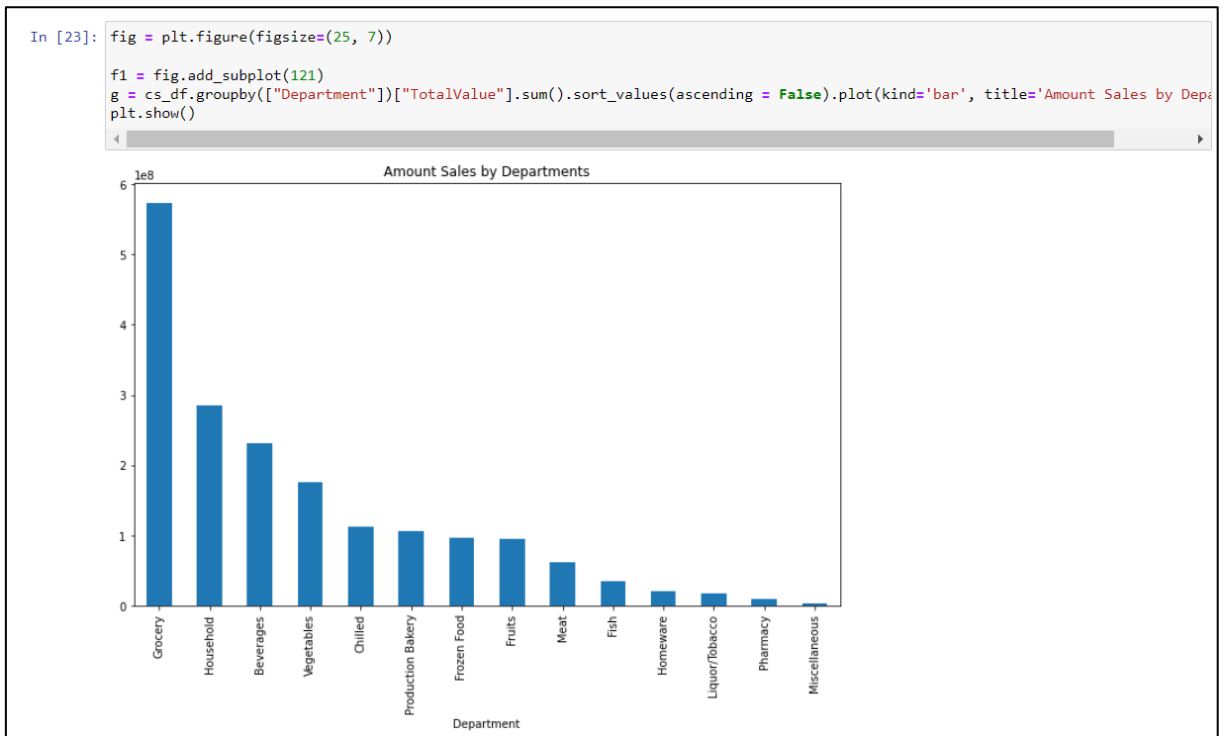


Figure 5.6: Sales by Department

```
In [25]: AmoutSum = cs_df.groupby(["Item_Description"]).TotalValue.sum().sort_values(ascending = False)
inv = cs_df[["Item_Description", "InvoiceNumber"]].groupby(["Item_Description"]).InvoiceNumber.unique().\
agg(np.size).sort_values(ascending = False)

fig = plt.figure(figsize=(25, 7))
f1 = fig.add_subplot(121)
Top10 = list(AmoutSum[:10].index)
PercentSales = np.round((AmoutSum[Top10].sum()/AmoutSum.sum()) * 100, 2)
PercentEvents = np.round((inv[Top10].sum()/inv.sum()) * 100, 2)
g = AmoutSum[Top10].\
plot(kind='bar', title='Top 10 Products in Sales Amount: {:.2f}% of Amount and {:.2f}% of Events'.\
format(PercentSales, PercentEvents))

f1 = fig.add_subplot(122)
Top10Ev = list(inv[:10].index)
PercentSales = np.round((AmoutSum[Top10Ev].sum()/AmoutSum.sum()) * 100, 2)
PercentEvents = np.round((inv[Top10Ev].sum()/inv.sum()) * 100, 2)
g = inv[Top10Ev].\
plot(kind='bar', title='Events of top 10 most sold products: {:.2f}% of Amount and {:.2f}% of Events'.\
format(PercentSales, PercentEvents))
```

Figure 5.7: Code for getting Top Products sold in Supermarket

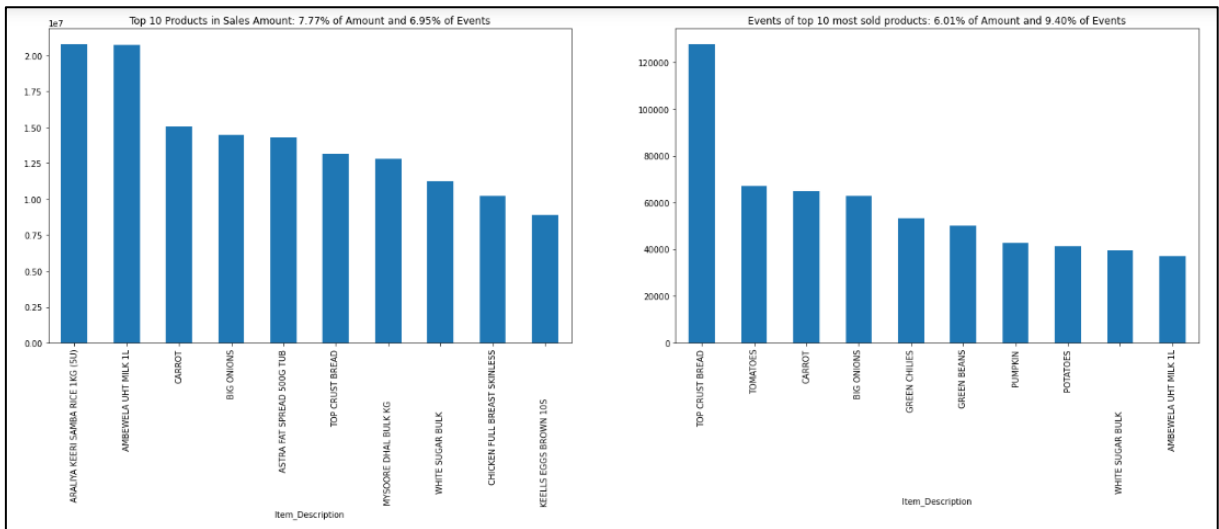


Figure 5.8: Top Products sold in Supermarket

5.2 Hopkins Statistics

The Hopkins statistic was used to test whether the dataset can be clustered. This statistic analyzes the geographic randomization of the data and displays the clustering tendency, or how well the data can be clustered. It also establishes the probability that a set of data was generated by a uniform distribution.

A statistically significant cluster is improbable if the value is 0.5 or less since the data are uniformly dispersed. Clusters have a high possibility of being statistically significant when the value is between 0.7 and 0.99.

```
In [43]: Num_features =dfan.select_dtypes(include=[np.number]).columns
hopkins(dfan[Num_features])

Out[43]: 0.9995120824414789
```

Figure 5.9: Hopkins Statistics calculation

The Hopkins score for the Keells data set is 0.99 which indicates that data has a high tendency to cluster.

5.3 RFM Modeling

The RFM model will take the transactions of a customer and calculate three important informational attributes about each customer:

- **Recency:** The value of how recently a customer purchased at the establishment
- **Frequency:** How frequent the customer's transactions are at the establishment
- **Monetary value:** The dollar (or pounds in our case) value of all the transactions that the customer made at the establishment

5.3.1 Clustering the Segments

Recency

The reference date for our analysis must be chosen to build the recency feature variable. In most cases, we utilize the date of the last transaction plus one day. The number of days before the reference date that a client last made a purchase will then be used to generate the recency variable.

```
In [60]: reference_date = cs_df.Date.max() + datetime.timedelta(days = 1)
print('Reference Date:', reference_date)
cs_df['days_since_last_purchase'] = (reference_date - cs_df.Date).astype('timedelta64[D]')
customer_history_df = cs_df[['NexusNumber', 'days_since_last_purchase']].groupby("NexusNumber").min().reset_index()
customer_history_df.rename(columns={'days_since_last_purchase': 'recency'}, inplace=True)
customer_history_df.describe().transpose()

Reference Date: 2022-01-15 00:00:00
```

Figure 5.10: Recency

```
Out[60]:
```

	count	mean	std	min	25%	50%	75%	max
NexusNumber	95729.0	1.121039e+15	3.260962e+12	1.121050e+14	1.121050e+15	1.121050e+15	1.121050e+15	1.121058e+15
recency	95729.0	5.297517e+00	3.186412e+00	1.000000e+00	2.000000e+00	6.000000e+00	7.000000e+00	1.400000e+01

Figure 5.11: Recency Summary Statistics

```

In [53]: def QQ_plot(data, measure):
          fig = plt.figure(figsize=(20,7))

          (mu, sigma) = norm.fit(data)

          fig1 = fig.add_subplot(121)
          sns.distplot(data, fit=norm)
          fig1.set_title(measure + ' Distribution ( mu = {:.2f} and sigma = {:.2f} )'.format(mu, sigma), loc='center')
          fig1.set_xlabel(measure)
          fig1.set_ylabel('Frequency')

          fig2 = fig.add_subplot(122)
          res = probplot(data, plot=fig2)
          fig2.set_title(measure + ' Probability Plot (skewness: {:.6f} and kurtosis: {:.6f} )'.format(data.skew(), data.kurt()), loc='center')

          plt.tight_layout()
          plt.show()

          QQ_plot(customer_history_df.recency, 'Recency')

```

Figure 5.12: Plotting the Recency Distribution and QQ-plot

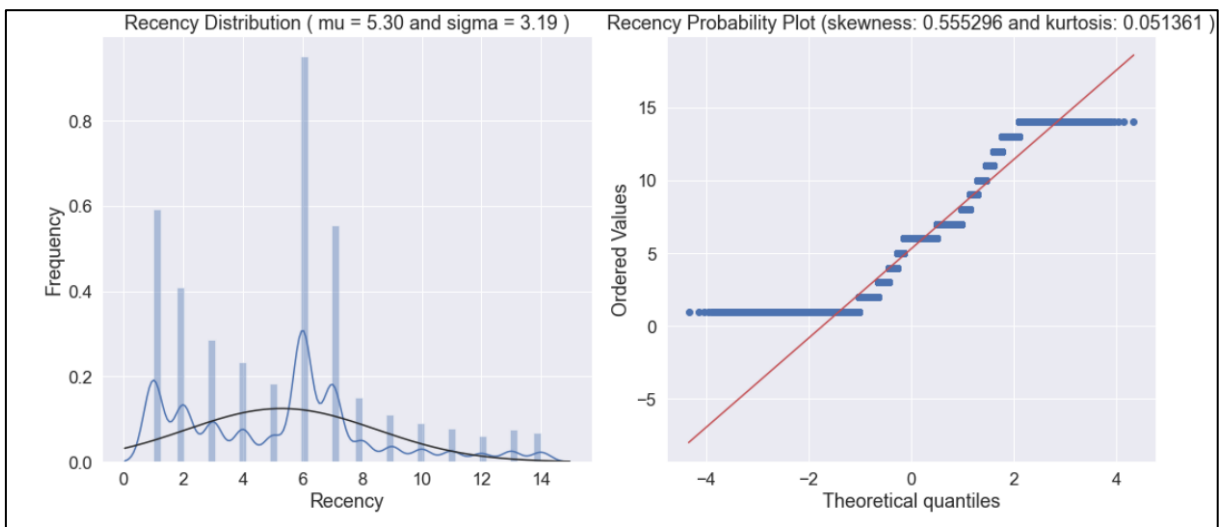


Figure 5.13: Recency Distribution and QQ-plot

We can observe that the sales recency distribution is skewed from the first graph up top. Positive bias and deviation from normal distribution characterize it.

We can observe from the Probability Plot that the diagonal red line, which represents the normal distribution, does not line up with the most recent sales data. Its distribution's shape demonstrates that it is skewed.

With a positive skewness of 0.56, we can validate the lack of symmetry and show that recent sales are skewed to the right. As we can also see from the Sales Distribution map, the skewed right denotes that the right tail is longer than the left tail. Any symmetric data should have a skewness that is close to zero since the skewness for a normal distribution is zero. If a distribution or data set appears the same to the left and right of the center point, it is said to be symmetrical.

Kurtosis is a metric that indicates how heavy-tailed or light-tailed the data are in comparison to a normal distribution. To put it another way, positive kurtosis denotes a heavy-tailed distribution, whereas negative kurtosis denotes a light-tailed distribution, thus data sets with high kurtosis tend to have heavy tails or outliers. Therefore, sales recency has tails and has some outliers with 0.05 of positive kurtosis.

```
In [58]: customer_history_df
Out[58]:
```

	NexusNumber	recency
0	112105000065701	6.0
1	112105000000000	2.0
2	112105000000010	2.0
3	112105000000020	3.0
4	112105000000030	2.0
...
106211	1121058000015710	13.0
106212	1121058000015800	10.0
106213	1121058000015860	7.0
106214	1121058000015900	4.0
106215	1121058000015920	4.0

95729 rows x 2 columns

Figure 5.14: Recency Summary by Customer

Frequency

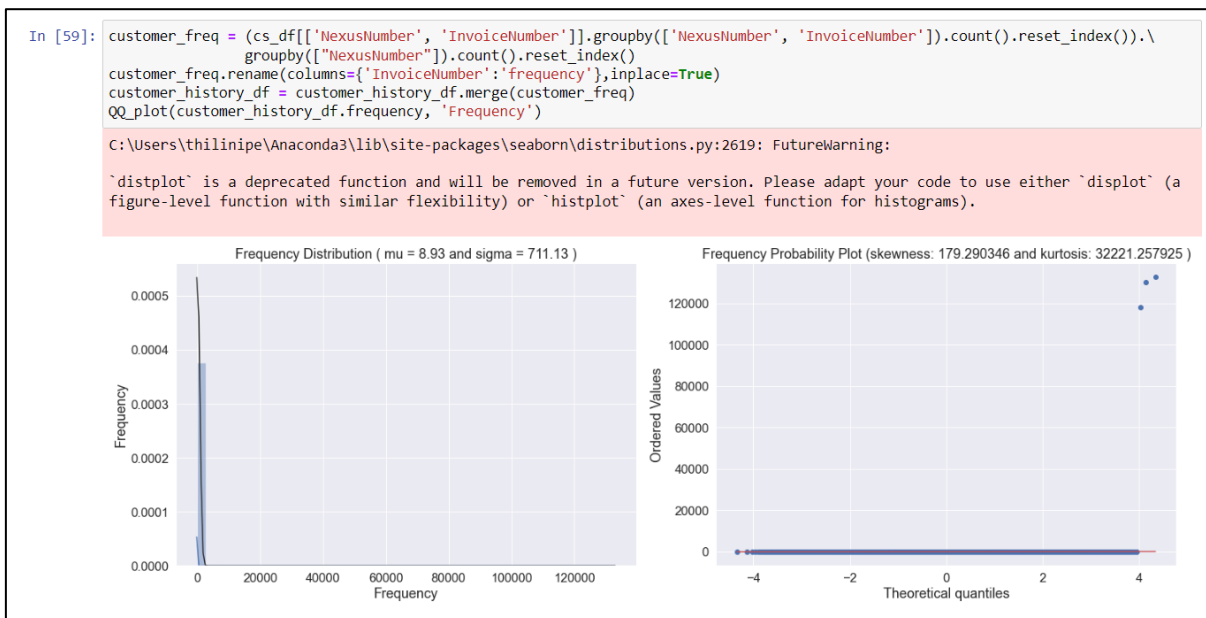


Figure 5.15: Frequency Distribution and QQ-plot

The first graph demonstrates no dispersion in the sales frequency distribution.

The Probability Plot demonstrates that the diagonal and sales frequency are aligned.

We can confirm the considerable absence of symmetry with skewness positive of 179.29, and 32221.26 A distribution with long tails and outliers is indicated by kurtosis.

```
In [60]: customer_freq
Out[60]:
```

	NexusNumber	frequency
0	112105000065701	3
1	1121050000000000	4
2	1121050000000010	14
3	1121050000000020	26
4	1121050000000030	22
...
106211	1121058000015710	1
106212	1121058000015800	1
106213	1121058000015860	1
106214	1121058000015900	1
106215	1121058000015920	2

106216 rows x 2 columns

Figure 5.16: Frequency Summary by Customer

Monetary

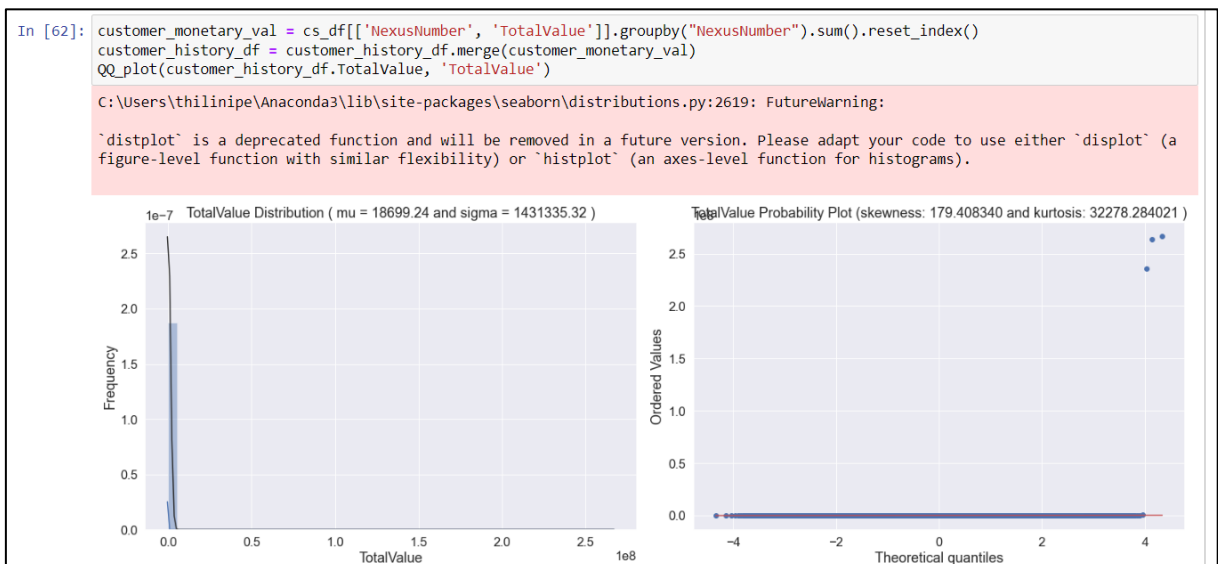


Figure 5.17: Total Value Distribution and QQ-Plot

From the first graph above we can see that the sales value distribution does not show a significant distribution.

From the Probability Plot, we could see that the sales amount aligns with the diagonal, and have some outliers.

With a skewness positive of 179.41, we confirm the high lack of symmetry and with 32278.28 Kurtosis.

```
In [66]: customer_history_df
```

```
Out[66]:
```

	NexusNumber	recency	frequency	TotalValue
0	112105000065701	6.0	3	2227.960
1	1121050000000000	2.0	4	56005.900
2	1121050000000010	2.0	14	49238.970
3	1121050000000020	3.0	26	97580.848
4	1121050000000030	2.0	22	53402.580
...
95724	1121058000015710	13.0	1	12500.000
95725	1121058000015800	10.0	1	3057.000
95726	1121058000015860	7.0	1	2044.000
95727	1121058000015900	4.0	1	8380.300
95728	1121058000015920	4.0	2	1944.820

95729 rows x 4 columns

Figure 5.18: Customer History Records

```
In [67]: customer_history_df.describe()
```

```
Out[67]:
```

	NexusNumber	recency	frequency	TotalValue
count	9.572900e+04	95729.000000	95729.000000	9.572900e+04
mean	1.121039e+15	5.297527	8.931275	1.869924e+04
std	3.260962e+12	3.186403	711.136618	1.431343e+06
min	1.121050e+14	1.000000	1.000000	1.000000e-02
25%	1.121050e+15	2.000000	2.000000	2.526340e+03
50%	1.121050e+15	6.000000	4.000000	6.545000e+03
75%	1.121050e+15	7.000000	7.000000	1.418373e+04
max	1.121058e+15	14.000000	132764.000000	2.667222e+08

Figure 5.19: Customer History Summary Statistics

K-Means Clustering

After creating the customer data set, the K means is used to cluster the data. Before using the algorithm, the dataset values have been standardized from 0 to 1 value.

```
In [68]: customer_history_df['recency_log'] = customer_history_df['recency'].apply(math.log)
customer_history_df['frequency_log'] = customer_history_df['frequency'].apply(math.log)
customer_history_df['value_log'] = customer_history_df['TotalValue'].apply(math.log)
feature_vector = ['value_log', 'recency_log', 'frequency_log']
X_subset = customer_history_df[feature_vector].as_matrix()
scaler = preprocessing.StandardScaler().fit(X_subset)
X_scaled = scaler.transform(X_subset)
pd.DataFrame(X_scaled, columns=X_subset.columns).describe().T
```

```
Out[68]:
```

	count	mean	std	min	25%	50%	75%	max
value_log	95729.0	1.199455e-15	1.000005	-10.268856	-0.613268	0.125612	0.725914	8.365089
recency_log	95729.0	-1.348245e-14	1.000005	-1.852538	-0.952038	0.475220	0.675484	1.575984
frequency_log	95729.0	-1.496969e-14	1.000005	-1.490158	-0.667862	0.154434	0.818319	12.504089

Figure 5.20: Scaled Data

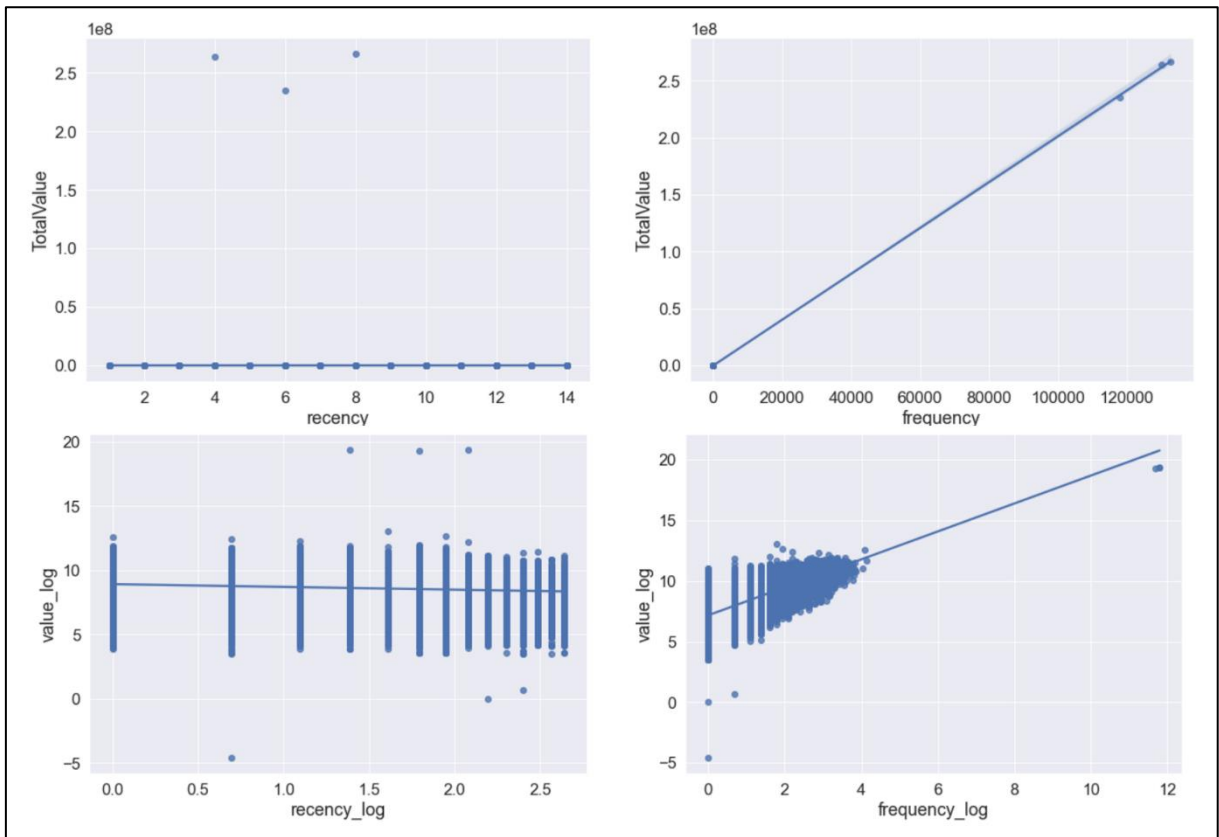


Figure 5.21: Receny and Frequency Plots

According to the increasing trend in Total Value and the accompanying increasing and falling trends for Frequency and Recency, respectively, customers who purchase more frequently and recently tend to spend more. This is evident from the plots above.

The Elbow Method

The elbow approach is used to determine the ideal number of clusters. The elbow approach seeks to locate the value of k at which the distortion increases most quickly. Since the samples will be nearer their assigned centroids as k rises, the distortion will diminish.

The Elbow method suggested clusters number is 6.

```

In [70]: c1 = 50
corte = 0.1

anterior = 1000000000000000
cost = []
K_best = c1

for k in range(1, c1+1):

    model = KMeans(
        n_clusters=k,
        init='k-means++',
        n_init=10,
        max_iter=300,
        tol=1e-04,
        random_state=101)

    model = model.fit(X_scaled)

    labels = model.labels_

    interia = model.inertia_
    if (K_best == c1) and (((anterior - interia)/anterior) < corte): K_best = k - 1
    cost.append(interia)
    anterior = interia

plt.figure(figsize=(8, 6))
plt.scatter(range(1, c1+1), cost, c='red')
plt.show()

print('The best K suggest: ',K_best)
model = KMeans(n_clusters=K_best, init='k-means++', n_init=10,max_iter=300, tol=1e-04, random_state=101)

model = model.fit(X_scaled)

labels = model.labels_

fig = plt.figure(figsize=(20,5))
ax = fig.add_subplot(121)
plt.scatter(X = X_scaled[:,1], y = X_scaled[:,0], c=model.labels_.astype(float))
ax.set_xlabel(feature_vector[1])
ax.set_ylabel(feature_vector[0])
ax = fig.add_subplot(122)
plt.scatter(x = X_scaled[:,2], y = X_scaled[:,0], c=model.labels_.astype(float))
ax.set_xlabel(feature_vector[2])
ax.set_ylabel(feature_vector[0])

plt.show()

```

Figure 5.22:Code for Cluster selection

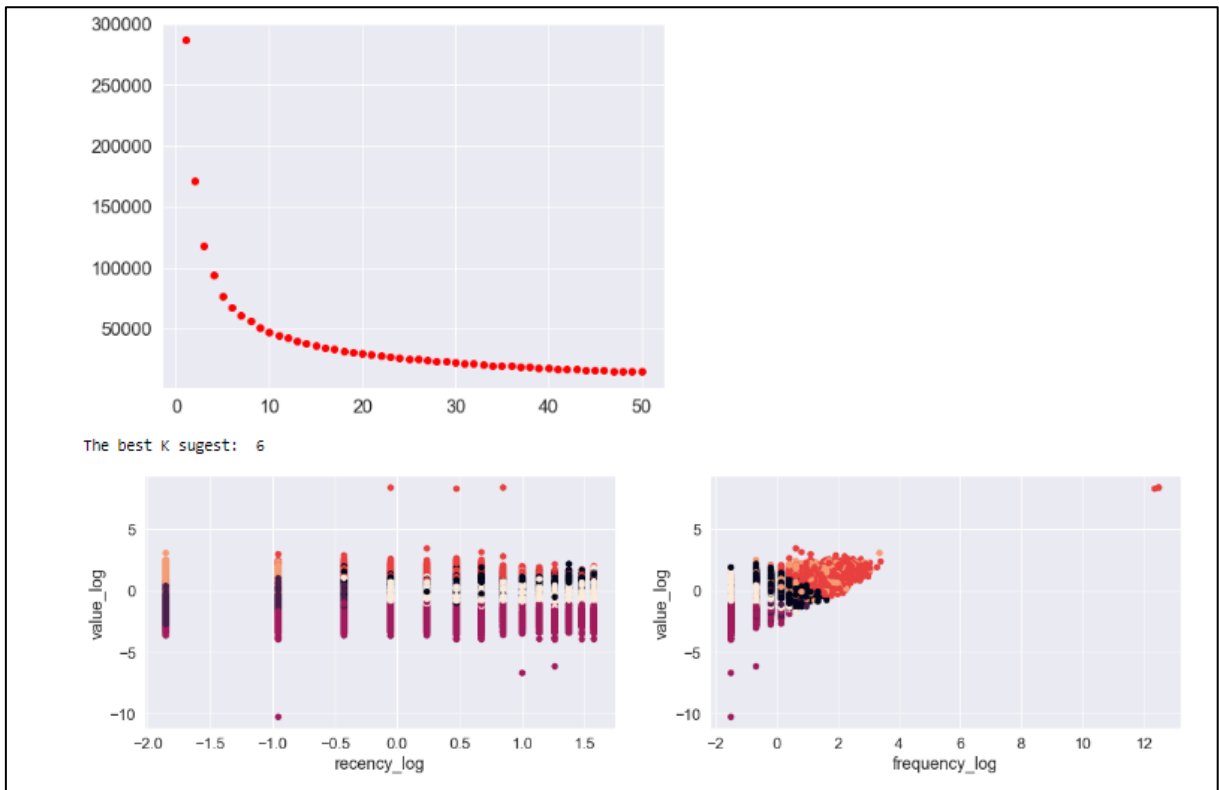


Figure 5.23: Elbow Method

Cluster Analysis

Cluster label 3 indicates good frequency and good spending, which characterizes frequent and heavy shoppers. Cluster label 4 displays a decent frequency and second-best buy. Cluster 0 presents the third-best purchase and fair frequency. This group should be attentive to specials and activations, so they do not get confused and make their next purchase.

```
In [72]: features = ['TotalValue', 'recency', 'frequency']
for i in range(3,K_best+1,1):
    print("for {} clusters the silhouette score is {:.2f}".format(i, cluster_centers[i]['silhouette_score']))
    print("Centers of each cluster:")
    cent_transformed = scaler.inverse_transform(cluster_centers[i]['cluster_center'])
    print(pd.DataFrame(np.exp(cent_transformed),columns=features))
    print('-'*50)

-----
for 6 clusters the silhouette score is 0.30
Centers of each cluster:
   TotalValue  recency  frequency
0  8182.038200  6.172218  4.298144
1  2814.003647  1.586744  2.152265
2   542.774732  6.211264  1.133602
3 23045.682585  5.604848 10.436789
4 14698.029513  1.272630  7.136589
5  3190.258527  7.290481  1.745115
-----
```

Figure 5.24: Cluster Center Analysis

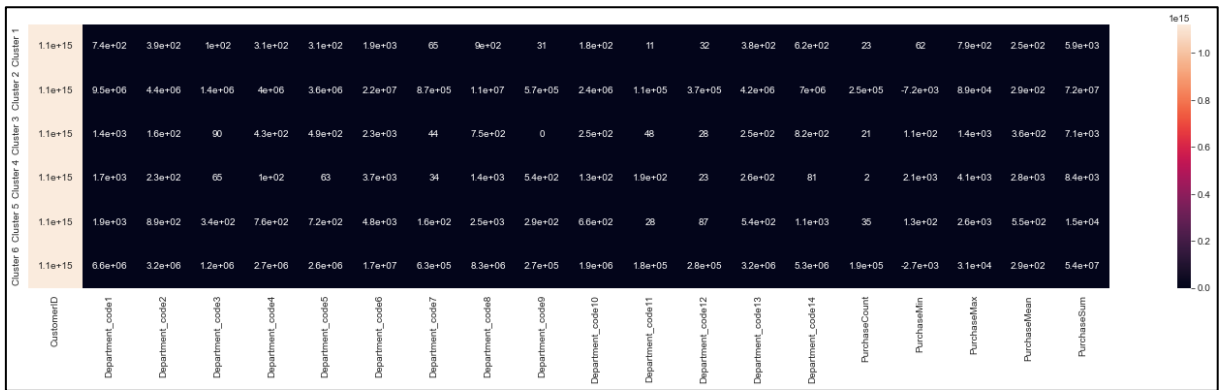


Figure 5.25: Heat Map for the Clusters

5.3.2 Clustering Performance Evaluation metrics

1. Silhouette Score

The separation between clusters is calculated using the Silhouette Score and Silhouette Plot. It shows the distance between each point in a cluster and points in other clusters. This metric, which has a range of [-1, 1], is excellent for visually examining cluster similarities and differences. (Eugenio Zuccarelli, 2021)

Silhouette score for the data set = 0.30

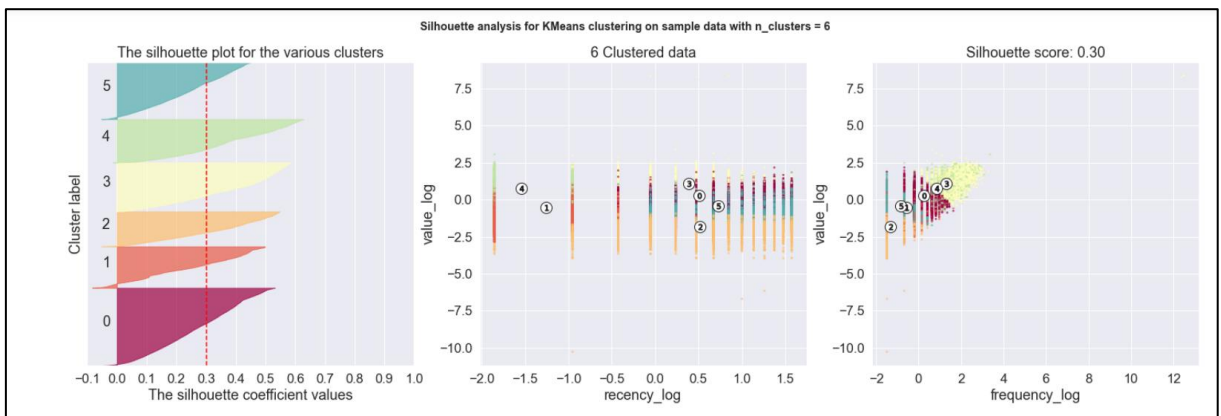


Figure 5.26: Silhouette Analysis

2. Calinski-Harabasz Index

The Variance Ratio Criterion is also known as the Calinski-Harabasz Index. The ratio of within-cluster to between-cluster dispersion is used to define the score. Since it does not require knowledge of ground truth labels, the C-H Index is a great approach to assess the effectiveness of a clustering algorithm. The performance improves as the Index rises. (Eugenio Zuccarelli, 2021)

Calinski-Harabasz Index for the data set = 62178.52

3. Davies-Bouldin Index

The average assessment of each cluster's resemblance to its most comparable cluster is known as the Davies-Bouldin Index. The ratio of within-cluster to between-cluster distances represents similarity. This will result in higher scores for clusters that are further apart and less scattered. The minimum score is zero, and unlike other performance indicators, the greater the clustering performance, the lower the value. (Eugenio Zuccarelli, 2021)

Davies-Bouldin Index for the data set = 1.025

```
In [83]: from sklearn import datasets
        from sklearn.cluster import KMeans
        from sklearn import metrics
        Sil = metrics.silhouette_score(X_scaled, labels)
        CH = metrics.calinski_harabasz_score(X_scaled, labels)
        DB = metrics.davies_bouldin_score(X_scaled, labels)

In [86]: print("silhouette_score =", Sil)
        silhouette_score = 0.30119726510239175

In [87]: print("calinski_harabasz_score=", CH)
        calinski_harabasz_score= 62178.517185990815

In [88]: print("davies_bouldin_score=",DB)
        davies_bouldin_score= 1.024795533639111
```

Figure 5.27: Performance Metrics Calculation

5.4 DBSCAN Algorithm

A density-based unsupervised clustering approach is known as DBSCAN (short for Density-Based Spatial Clustering of Applications with Noise). In DBSCAN, clusters are created from dense areas and are divided into sparsely populated areas. In contrast to k-means clustering, which commonly produces spherical-shaped clusters, DBSCAN computes nearest-neighbor graphs and produces arbitrary-shaped clusters in datasets (which may contain noise or outliers). (Reneshbe, 2022)

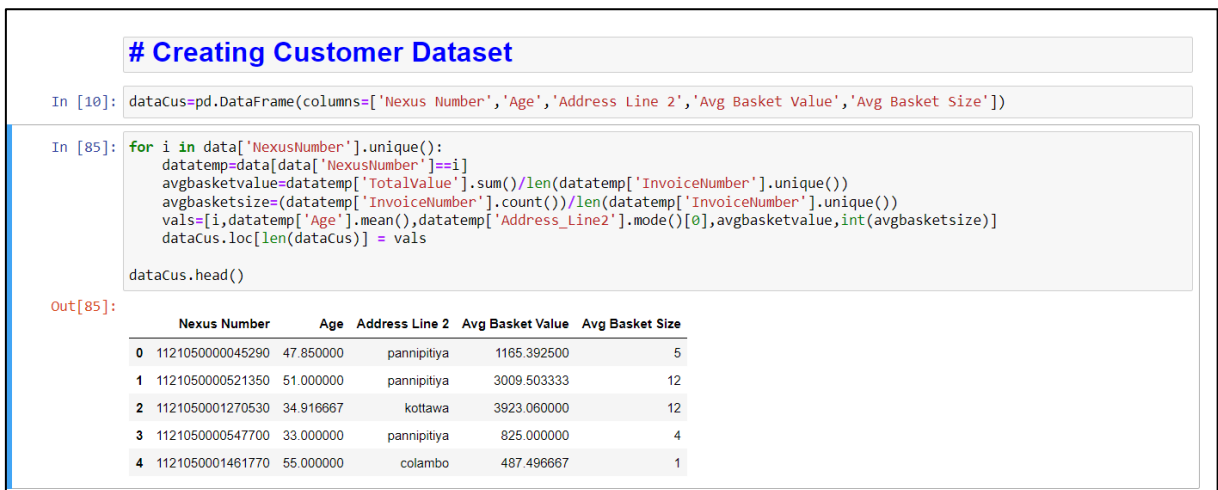


Figure 5.28:DBSCAN Execution

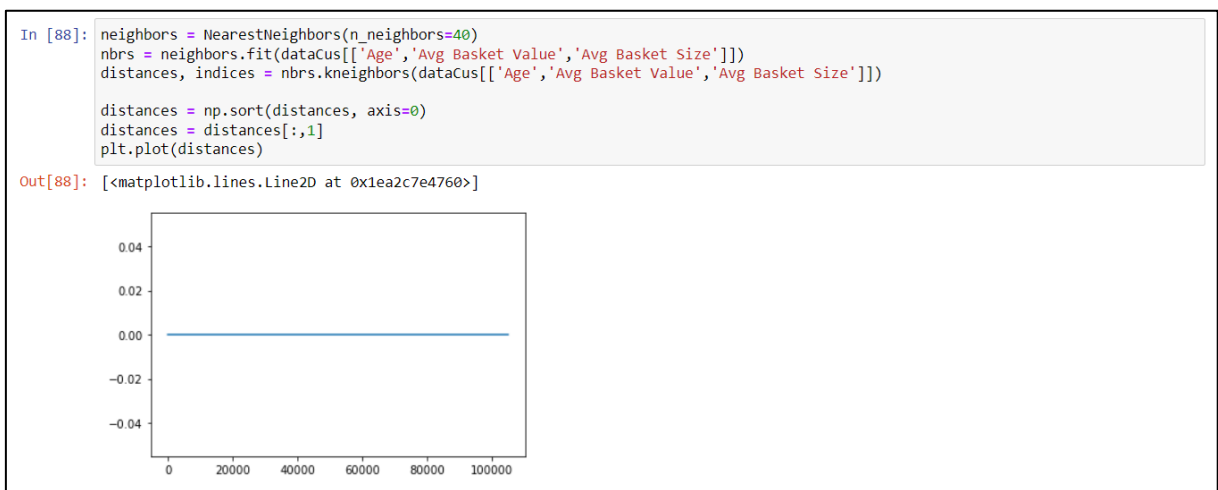


Figure 5.29:Determining the Epsilon value (eps)

Determining the number of clusters

DBSCAN doesn't need a specific number of clusters to function. Epsilon values and Minimum sample values are necessary, though. The epsilon value was set at 0.3 and the value of the minimum sample was set at 7.

Analyzing Clusters

The following Numerical data variables were passed to the DBSCAN algorithm

- Age
- Avg Basket Value
- Avg Basket Size

```

In [91]: db = DBSCAN(eps=0.3, min_samples=7).fit(dataCus[['Age', 'Avg Basket Value', 'Avg Basket Size']])
clusteredDataDB=dataCus.copy()
clusteredDataDB['cluster']=db.labels_
clusteredDataDB.head()

clusteredDataDB['cluster'].unique()

Out[91]: array([-1,  0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11,
 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,
 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37,
 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,
 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63,
 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76,
 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89,
 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102,
103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115,
116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128,
129, 130, 131, 132, 133, 134, 135], dtype=int64)

```

Figure 5.30: DBSCAN Clusters

```

In [93]: silhoutdbcust=silhouette_score(dataCus[['Age', 'Avg Basket Value', 'Avg Basket Size']],clusteredDataDB['cluster'])
silhoutdbcust

Out[93]: -0.8550054640031985

```

Figure 5.31: Silhouette Score

Model Evaluation

Silhouette Score: -0.855

Conclusion

The results here are inconclusive. Therefore, for this dataset, the algorithm is not suitable.

5.5 Association Rule Mining

The association rules discovered by the FPG and Apriori algorithms appear to be related to one another. While the enormous amounts of data at hand can be used to construct a comprehensive list of Rules, processing such data without industrial-level computing capacity is not viable. This led to the adoption of a scaled-down version of the dataset. The following rules were discovered for each week of transactions using this dataset.

5.5.1 Apriori Algorithm

```

In [32]: from mlxtend.frequent_patterns import apriori, association_rules
         from mlxtend.frequent_patterns import fpgrowth

In [33]: #Building the model
         frq_items = apriori(itemsetencoded, min_support = 0.01, use_colnames = True)

         #Collecting the inferred rules in a dataframe
         rules = association_rules(frq_items, metric = "lift", min_threshold = 1)
         rules = rules.sort_values(['confidence', 'lift'], ascending = [False, False])

C:\Users\thilnipe\Anaconda3\lib\site-packages\mlxtend\frequent_patterns\fpcommon.py:111: DeprecationWarning: DataFrames with non-bool types result in worse computational performance and their support might be discontinued in the future. Please use a DataFrame with bool type
  warnings.warn(

In [34]: print(rules[rules['confidence']>0.65])

```

Figure 5.32: Apriori Algorithm Execution

	antecedents	consequents	antecedent support	\
162	(GARLIC, POTATOES)	(BIG ONIONS)	0.013704	
203	(GREEN BEANS, LEEKS)	(CARROT)	0.016056	
222	(TOMATOES, LEEKS)	(CARROT)	0.017188	
167	(TOMATOES, GARLIC)	(BIG ONIONS)	0.015417	
179	(GREEN CHILIES, POTATOES)	(BIG ONIONS)	0.015649	
191	(TOMATOES, POTATOES)	(BIG ONIONS)	0.017827	

	consequent support	support	confidence	lift	leverage	conviction
162	0.078683	0.010394	0.758475	9.639624	0.009316	3.814576
203	0.075547	0.011701	0.728752	9.646305	0.010488	3.408149
222	0.075547	0.011788	0.685811	9.077900	0.010489	2.942344
167	0.078683	0.010568	0.685499	8.712162	0.009355	2.929457
179	0.078683	0.010598	0.677180	8.606433	0.009366	2.853965
191	0.078683	0.012049	0.675896	8.590111	0.010647	2.842656

Figure 5.33: Apriori Algorithm – Rules set I

	antecedents	consequents	antecedent support	\
247	(GREEN BEANS, LEEKS)	(CARROT)	0.015183	
223	(GREEN BEANS, CABBAGE)	(CARROT)	0.018253	
259	(GREEN BEANS, POTATOES)	(CARROT)	0.014879	
284	(TOMATOES, LEEKS)	(CARROT)	0.015985	
241	(GREEN BEANS, GREEN CHILIES)	(CARROT)	0.020576	
181	(TOMATOES, GARLIC)	(BIG ONIONS)	0.016345	
253	(GREEN BEANS, LIME)	(CARROT)	0.015571	
205	(TOMATOES, POTATOES)	(BIG ONIONS)	0.018834	
307	(GREEN BEANS, GREEN CHILIES)	(TOMATOES)	0.020576	
271	(GREEN BEANS, TOMATOES)	(CARROT)	0.028182	

	consequent support	support	confidence	lift	leverage	conviction
247	0.079429	0.010869	0.715847	9.012394	0.009663	3.239701
223	0.079429	0.012916	0.707576	8.908261	0.011466	3.148066
259	0.079429	0.010399	0.698885	8.798842	0.009217	3.057204
284	0.079429	0.011146	0.697232	8.778032	0.009876	3.040514
241	0.079429	0.013939	0.677419	8.528596	0.012304	2.853770
181	0.078848	0.011035	0.675127	8.562343	0.009746	2.835420
253	0.079429	0.010482	0.673179	8.475216	0.009245	2.816747
205	0.078848	0.012667	0.672540	8.529539	0.011182	2.813024
307	0.085956	0.013413	0.651882	7.583893	0.011645	2.625671
271	0.079429	0.018336	0.650638	8.191422	0.016098	2.635005

Figure 5.34: Apriori Algorithm – Rules Set II

5.5.2 Frequent Pattern Growth Algorithm

```
In [38]: #Building the model
frq_itemsfpg = fpgrowth(itemsetencodedfpg, min_support=0.01,use_colnames=True)

#Collecting the inferred rules in a dataframe
rulesfpg = association_rules(frq_itemsfpg, metric="lift", min_threshold = 1)
rulesfpg = rulesfpg.sort_values(['confidence', 'lift'], ascending =[False, False])

In [39]: print(rulesfpg)

rulesfpg[rulesfpg['confidence']>0.50].to_csv('FPG_ARM_Rules.csv')
```

Figure 5.35: Frequent Pattern Growth Algorithm Execution

```
In [39]: print(rulesfpg)

rulesfpg[rulesfpg['confidence']>0.50].to_csv('FPG_ARM_Rules.csv')
```

	antecedents	consequents	antecedent support	\
221	(GREEN BEANS, LEEKS)	(CARROT)	0.015183	
147	(GREEN BEANS, CABBAGE)	(CARROT)	0.018253	
119	(GREEN BEANS, POTATOES)	(CARROT)	0.014879	
215	(TOMATOES, LEEKS)	(CARROT)	0.015985	
44	(GREEN CHILIES, GREEN BEANS)	(CARROT)	0.020576	
..	
162	(TOMATOES)	(GREEN BEANS, CABBAGE)	0.085956	
252	(TOMATOES)	(PUMPKIN, GREEN BEANS)	0.085956	
1	(TOP CRUST BREAD)	(TOMATOES)	0.109049	
7	(TOP CRUST BREAD)	(BIG ONIONS)	0.109049	
17	(TOP CRUST BREAD)	(CARROT)	0.109049	

	consequent support	support	confidence	lift	leverage	conviction
221	0.079429	0.010869	0.715847	9.012394	0.009663	3.239701
147	0.079429	0.012916	0.707576	8.908261	0.011466	3.148066
119	0.079429	0.010399	0.698885	8.798842	0.009217	3.057204
215	0.079429	0.011146	0.697232	8.778032	0.009876	3.040514
44	0.079429	0.013939	0.677419	8.528596	0.012304	2.853770
..
162	0.018253	0.010122	0.117761	6.451498	0.008553	1.112790
252	0.017424	0.010012	0.116474	6.684846	0.008514	1.112108
1	0.085956	0.012058	0.110576	1.286421	0.002685	1.027680
7	0.078848	0.011754	0.107786	1.367003	0.003156	1.032433
17	0.079429	0.010426	0.095612	1.203745	0.001765	1.017894

[324 rows x 9 columns]

Figure 5.36: Frequent Pattern Growth Algorithm - Rules Set I

```
In [42]: print(rulesfpga)
rulesfpga[rulesfpga['confidence']>0.50].to_csv('FPG_ARM_RulesII.csv')
```

	antecedents	consequents	antecedent support	\
221	(GREEN BEANS, LEEKS)	(CARROT)	0.015183	
147	(GREEN BEANS, CABBAGE)	(CARROT)	0.018253	
119	(GREEN BEANS, POTATOES)	(CARROT)	0.014879	
215	(TOMATOES, LEEKS)	(CARROT)	0.015985	
44	(GREEN CHILIES, GREEN BEANS)	(CARROT)	0.020576	
..	
162	(TOMATOES)	(GREEN BEANS, CABBAGE)	0.085956	
252	(TOMATOES)	(PUMPKIN, GREEN BEANS)	0.085956	
1	(TOP CRUST BREAD)	(TOMATOES)	0.109049	
7	(TOP CRUST BREAD)	(BIG ONIONS)	0.109049	
17	(TOP CRUST BREAD)	(CARROT)	0.109049	

	consequent	support	support	confidence	lift	leverage	conviction
221		0.079429	0.010869	0.715847	9.012394	0.009663	3.239701
147		0.079429	0.012916	0.707576	8.908261	0.011466	3.148066
119		0.079429	0.010399	0.698885	8.798842	0.009217	3.057204
215		0.079429	0.011146	0.697232	8.778032	0.009876	3.040514
44		0.079429	0.013939	0.677419	8.528596	0.012304	2.853770
..	
162		0.018253	0.010122	0.117761	6.451498	0.008553	1.112790
252		0.017424	0.010012	0.116474	6.684846	0.008514	1.112108
1		0.085956	0.012058	0.110576	1.286421	0.002685	1.027680
7		0.078848	0.011754	0.107786	1.367003	0.003156	1.032433
17		0.079429	0.010426	0.095612	1.203745	0.001765	1.017894

[324 rows x 9 columns]

Figure 5.37: Frequent Pattern Growth Algorithm - Rules Set II

It should be mentioned that the outcomes of the Frequent Pattern Growth Algorithm are comparable to those of the Apriori algorithm. Several recurring rules are discernible based on the outcomes from both algorithms and all datasets. These will be the best rules to consider when analyzing. Multiple data files won't be required in real-world situations, which will drastically minimize the required work.

6 Discussion

6.1 Challenges

During the research project, several challenges were faced in different domain areas. They have been described in detail below.

6.1.1 Data Integrity & Consistency

A large number of variables that were missing and values that were misspelled in the dataset made data pre-processing extremely difficult. The primary causes of such inaccuracies can be identified as human error and a lack of care during the data collecting and signup procedure. While special characters and null values were easily removed by ordinary data preprocessing,

spelling problems required a more sophisticated approach because there were endless combinations for each city or town name. There are still a variety of misspelled terms, though, which require personal intervention to be removed.

6.1.2 Large Data Volume

For a data mining project, a high data volume is ideal. However, the sheer volume of records in this study made processing and computation difficult. The dataset had to be shrunk to a few hundred thousand records to be mitigated. As a result, a dataset with roughly 3.5 million records would be reduced to between 200,000 and 250,000 records. This stage had the potential to have an impact on the number of clusters and association rules extracted.

6.1.3 Limited Computational Power

The coding scripts were executed both on an Asus Expert book with an Intel 11th Gen Core i5 with 16GB of RAM running on the x64 Architecture. Yet, certain algorithms were still not functioning or would simply hang upon execution, which unfortunately severely limited the different clustering approaches available for application.

6.2 Application of Algorithms

Application of the algorithms can be very straightforward, but depending on how the data is presented, they behave very differently. The original data file was subjected to the algorithms in its original form during this investigation (post cleansing & text pre-processing). Clustering algorithms like DBSCAN and mean shift would occasionally not work. The dataset was shrunk in size to get around this, but the outcomes held. Surprisingly, the "Customer" datasets, which are derived datasets, would work with these identical techniques. It was determined that this was caused by the vast number of variables and the noise that was produced by outliers, misspelled addresses, etc.

When certain algorithms were applied, the findings were ambiguous. The K-Modes method is an effective illustration of this. Given that K-Modes only accept categorical variables, the inaccuracies could be caused by issues with the data itself. In addition to machine learning algorithms, there are various segmentation techniques. Since RFM analysis did not utilize a complex mathematical formula to divide customers into segments, it was one method for

customer segmentation. With this technique, clients are categorized depending on their purchase habits. While straightforward, it lacks an exploratory clustering component.

6.3 Assumptions

Several presumptions had to be true for these clusters to be accurate.

- Customers who are not Nexus Program members are not included in the study.
- Purchases directly reflect the needs and patterns of the buyer.
- Buyers will stick to the outlet closest to where they live.
- All the outlets in question were equal in every way.
- Multiple people could not make purchases through a single invoice.
- Nexus IDs would not be shared during checkout.
- External elements impacting customers, such as outlet size, parking availability, and product portfolios are disregarded.
- "Item Count" refers to the number of goods, not the number of SKUs.
- External influences like geopolitical unrest and financial crises are disregarded.

6.4 Conclusion

In conclusion, the cluster analysis of the chosen set of transactions revealed details on possible demographic groups that might have existed among the target clientele. The RFM model was used to calculate the number of clusters, and then the non-hierarchical k-means clustering method was used.

Because the dataset in this study was unlabeled, internal clustering validation was chosen over external clustering validation, which depends on some external data like labels. Internal cluster validation can be used to choose the clustering algorithm that best fits the dataset and can correctly cluster data into its opposite cluster. The association between the fresh items in the store was ascertained using the market basket analysis with the Apriori and FPG algorithms.

7 References

- Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Rec.* 28, 49–60. <https://doi.org/10.1145/304181.304187>
- Aranganayagi, S., Thangavel, K., 2007. Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure, in *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*. Presented at the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), IEEE, Sivakasi, Tamil Nadu, India, pp. 13–17. <https://doi.org/10.1109/ICCIMA.2007.328>
- Ayat, N.E., Cheriet, M., Remaki, L., Suen, C.Y., 2001. KMOD - a new support vector machine kernel with moderate decreasing for pattern recognition. Application to digit image recognition, in *Proceedings of Sixth International Conference on Document Analysis and Recognition*. Presented at the Sixth International Conference on Document Analysis and Recognition, IEEE Comput. Soc, Seattle, WA, USA, pp. 1215–1219. <https://doi.org/10.1109/ICDAR.2001.953976>
- B. Stone, Ron Jacobs, 2008. *Successful direct marketing methods : interactive, database, and customer-based marketing for the digital age*.
- Béjar Alonso, J., 2013. K-means vs Mini Batch K-means: a comparison.
- Dachyar, M., Vitasya, L., 2021. Customer Loyalty Analysis Using Customer Lifetime Value (A Case Study of Baby Equipment SMEs). *IEOM Soc. Fagnoli M., Lombardi M., Tronci M., Dallasega P., Savino M.M., Costantino F., Di Gravio G., Patriarca R., 2021, Pages 2094-2104*.
- Eugenio Zuccarelli, 2021. *Performance Metrics in Machine Learning — Part 3: Clustering*.
- Fontanini, A.D., Abreu, J., 2018. A Data-Driven BIRCH Clustering Method for Extracting Typical Load Profiles for Big Data, in *2018 IEEE Power & Energy Society General Meeting (PESGM)*. Presented at the 2018 IEEE Power & Energy Society General Meeting (PESGM), IEEE, Portland, OR, USA, pp. 1–5. <https://doi.org/10.1109/PESGM.2018.8586542>
- Frey, B.J., Dueck, D., 2007. Clustering by Passing Messages Between Data Points. *Science* 315, 972–976. <https://doi.org/10.1126/science.1136800>
- Ghosh, Samarendra, 2014. *Identification of Best Algorithm in Association Rule Mining Based on Performance*.
- HubSpot, n.d. *How to Organize Your Customers to Grow Better*. URL <https://blog.hubspot.com/service/customer-segmentation>
- Ji, J., Bai, T., Zhou, C., Ma, C., Wang, Z., 2013. An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing* 120, 590–596. <https://doi.org/10.1016/j.neucom.2013.04.011>
- Kansal, T., Bahuguna, S., Singh, V., Choudhury, T., 2018. Customer Segmentation using K-means Clustering, in the *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. Presented at the 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), IEEE, Belgaum, India, pp. 135–139. <https://doi.org/10.1109/CTEMS.2018.8769171>
- Karunaratna, A.C., 2021. Motives of Customer Loyalty in Supermarket Patronage in Sri Lanka. *Vidyodaya J. Manag.* 7. <https://doi.org/10.31357/vjm.v7i1.4912>
- KEMP, S., 2021. *DIGITAL 2021: SRI LANKA*. URL <https://datareportal.com/reports/digital-2021-sri-lanka>

- Khajvand, M., Zolfaghar, K., Ashoori, S., Alizadeh, S., 2011. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: A case study. *Procedia Comput. Sci.* 3, 57–63. <https://doi.org/10.1016/j.procs.2010.12.011>
- Mihirani Dissanyake, 2020. 2020 AND BEYOND: TRENDS SHAPING SUPERMARKETS FOR THE FUTURE. URL <https://www.dailynews.lk/2020/07/08/finance/222626/2020-and-beyond-trends-shaping-supermarkets-future-retrospective-view-past>
- Minho Ryu, Kwang-II Ahn, Kichun Lee, 2021. Finding Effective Item Assignment Plans with Weighted Item Associations Using A Hybrid Genetic Algorit.
- Murphy, K.P., 2012. *Machine learning: a probabilistic perspective*, Adaptive computation, and machine learning series. MIT Press, Cambridge, MA.
- Nash, D., Sterna-Karwat, A., 1996. An application of DEA to measure branch cross-selling efficiency. *Comput. Oper. Res.* 23, 385–392. [https://doi.org/10.1016/0305-0548\(95\)00046-1](https://doi.org/10.1016/0305-0548(95)00046-1)
- Ngai, E.W.T., Xiu, L., Chau, D.C.K., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Syst. Appl.* 36, 2592–2602. <https://doi.org/10.1016/j.eswa.2008.02.021>
- Principles of Managing Customer Experience and Relationships, 2016. , in: *Managing Customer Relationships*. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 1–1. <https://doi.org/10.1002/9781119239833.part1>
- Rachman, F.P., Santoso, H., Djajadi, A., 2021. Machine Learning Mini Batch K-means and Business Intelligence Utilization for Credit Card Customer Segmentation. *Int. J. Adv. Comput. Sci. Appl.* 12. <https://doi.org/10.14569/IJACSA.2021.0121024>
- Ram, A., Jalal, S., Jalal, A.S., Kumar, M., 2010. A Density-Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases. *Int. J. Comput. Appl.* 3, 1–4. <https://doi.org/10.5120/739-1038>
- Reneshbe, 2022. DBSCAN in Python. URL <https://www.reneshbedre.com/blog/dbscan-python.html>
- Stone, M., Woodcock, N., Wilson, M., 1996. Managing the change from marketing planning to customer relationship management. *Long Range Plann.* 29, 675–683. [https://doi.org/10.1016/0024-6301\(96\)00061-1](https://doi.org/10.1016/0024-6301(96)00061-1)
- Thomas, B., Housden, M., 2002. *Direct marketing in practice*, Chartered Institute of Marketing/Butterworth-Heinemann marketing series. Elsevier, Amsterdam.
- von Luxburg, U., 2007. A tutorial on spectral clustering. *Stat. Comput.* 17, 395–416. <https://doi.org/10.1007/s11222-007-9033-z>
- Witten and Frank, I.W. and E.F., 2006. *Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques 2nd edition*: San Francisco: Morgan Kaufmann Publishers; 2005:560. ISBN 0-12-088407-0, £34.99. *Biomed. Eng. OnLine* 5, 51, 1475-925X-5–51. <https://doi.org/10.1186/1475-925X-5-51>
- Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., Xu, G., 2020. An Empirical Study on Customer Segmentation by Purchase Behaviors Using an RFM Model and K-Means Algorithm. *Math. Probl. Eng.* 2020, 1–7. <https://doi.org/10.1155/2020/8884227>
- Wu, K.-L., Yang, M.-S., 2007. Mean shift-based clustering. *Pattern Recognit.* 40, 3035–3052. <https://doi.org/10.1016/j.patcog.2007.02.006>
- Xu, S., Qiao, X., Zhu, L., Zhang, Y., Xue, C., Li, L., 2016. Reviews on Determining the Number of Clusters. *Appl. Math. Inf. Sci.* 10, 1493–1512. <https://doi.org/10.18576/amis/100428>
- Algorithm. *Journal of King Saud University - Computer and Information Sciences*.
- Aryuni, M., Madyatmadja, E. & Miranda, E., 2018. Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering. *2018 International Conference on Information Management and Technology (ICIMTech)*, pp. 412-416.

- Brijs, T., Vanhoof, K. & Wets, G., 2004. Using Shopping Baskets to Cluster Supermarket Shoppers.
- Chen, D., Sain, S. & 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*, Volume 19.
- Dogan, O., Ayçin, E. & Bulut, Z., 2016. CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY. *International Journal of Contemporary Economics and Administrative Sciences*, Volume 8, pp. 1-19.
- Ezenkwu, C. P., Ozuomba, S. & Kalu, C., 2015. Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services. *International Journal of Advanced Research in Artificial Intelligence*, Volume 4.
- Gong, Xia, H. a. & Qiong, 2009. Study on Application of Customer Segmentation Based on Data Mining Technology. pp. 167 - 170.
- Jun Wu, Li Shi, Wen-Pin Lin, Sang-Bing Tsai, Yuanyuan Li, Liping Yang, and Guangshu Xu, 2020, pp. 1024-123X, doi: 10.1155/2020/8884227. . *An Empirical Study on Customer Segmentation by Purchase*. s.l., s.n.
- Kashwan, K. R. & Velu, C. M., 2013. Customer Segmentation Using Clustering and Data. *International Journal of Computer Theory and Engineering*, 5(6), pp. 856-861.
- Mustakim, et al., 2018. Market Basket Analysis Using Apriori and FP-Growth for Analysis Consumer Expenditure Patterns at Berkah Mart in Pekanbaru Riau. *Journal of Physics: Conference Series*, Volume 1114.
- T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171. *Customer Segmentation using K-means Clustering*. s.l., s.n.
- Tripathi, S., Bhardwaj, A. & Eswaran, P., 2018. Approaches to Clustering in Customer Segmentation. *International Journal of Engineering & Technology*, Volume 7, p. 802.