

Customer Segmentation Using K-means

by

Nafis Mahdee

18301035

Ishrak Rahman Shourav

18101664

Tasneem Tabassum

17101219

Eman Nur

17101375

Md Amir Hamza Howlader

17101528

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
May 2022

© 2022. Brac University
All rights reserved.

Declaration

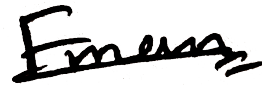
It is hereby declared that

1. The thesis submitted is our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



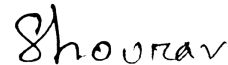
Tasneem Tabassum
17101219



Eman Nur
17101375



Md Amir Hamza Howlader
17101528



Ishrak Rahman Shourav
18101664



Nafis Mahdee
18301035

Approval

The thesis titled “Customer Segmentation Using K-means”

1. Nafis Mahdee (18301035)
2. Ishrak Rahman Shourav (18101664)
3. Tasneem Tabassum (17101219)
4. Eman Nur (17101375)
5. Md Amir Hamza Howlader (17101528)

As of spring, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 29, 2022.

Examining Committee:

Supervisor:
(Member)

Annajiat Alim Rasel
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

The thesis is carried out in complete compliance with research ethics, policies, regulations and codes set by BRAC University. We have used various information from different sources in order to pursue the research. To collect data, we read articles, journals from different websites, etc. The sources we have used here are interpreted in our own terms and are properly mentioned as a reference. We appreciate and give credit to every source that helped us to continue our work. Lastly, we declare that five authors of this paper hold liability if any violation of BRAC University standard is found.

Abstract

Sales Maximization is a critical aspect of operating any business. Our thesis aims to help businesses to probe deep into their market reach as we group customers using the customer segmentation approach. Our dataset is formed based on customer behavior and purchase history. The outcome of this organized study is expected to yield powerful insights in predicting consumer purchasing behavior and related patterns. Using the K-means algorithm, we analyze real-time transactional and retail datasets. The analyzed data forecasts purchasing patterns and behavior of customers. This study uses the RMF (Recency, Frequency Monetary), LRFM (Length, Recency, Frequency, Monetary), and PCA model deploying K-means on a dataset. The results thus obtained concerning sales transactions are compared with multiple parameters like Sales Recency, Sales Frequency, and Sales Volume.

Keywords: Segmentation, Customer segmentation, Clustering, k-means, RFM, LRFM, PCA, Data mining, Machine learning.

Dedication

We would like to dedicate this thesis paper to our respected parents, and all the friends and family members who supported us in the undergraduate run. Specially, we would like to dedicate this paper to our respected supervisor Annajiat Alim Rasel sir and to our respected faculty members who helped us throughout the years inside and outside classroom with their knowledge and how to implement it in real life. Last but not the least, we also want to mention the names of (Late) Md. Abdul Mottalib sir (May he rest in peace) and Mahbubul Alam Majumder sir for showing us and the whole department, the right path to gain knowledge.

Acknowledgement

We would like to acknowledge that this thesis report has been done under the supervision and guidance of our honorable faculty, Annajiat Alim Rasel sir. Above all, we would like to thank the Almighty for providing us with the opportunity to do the research successfully, for giving us the confidence and guidance by Annajiat Alim Rasel sir.

Table of Contents

| | |
|--|------------|
| Declaration | i |
| Approval | ii |
| Ethics Statement | iii |
| Abstract | iv |
| Dedication | v |
| Acknowledgment | vi |
| Table of Contents | vii |
| List of Figures | ix |
| 1 Introduction | 1 |
| 1.1 Overview | 1 |
| 1.2 Problem Statement | 2 |
| 1.3 Research Objectives | 2 |
| 1.4 Customer Segmentation | 2 |
| 1.4.1 Description | 2 |
| 1.4.2 Methods of Customer Segmentation | 3 |
| 2 Literature Review | 4 |
| 2.1 Overview | 4 |
| 2.2 Related Works | 4 |
| 3 Methods and Models | 10 |
| 3.1 Recency Frequency Monetary (RFM) Model | 10 |
| 3.2 Length Recency Frequency Monetary (LRFM) Model | 11 |
| 3.3 Clustering Algorithm | 11 |
| 3.4 K-Means Algorithm | 11 |
| 3.5 K-Means ++ | 12 |
| 3.6 Elbow Method | 13 |
| 3.7 Principal Component Analysis | 13 |
| 3.8 Customer Relationship Management | 14 |
| 3.9 Customer Lifetime Value (CLV) | 15 |
| 3.10 Silhouette Score | 15 |

| | | |
|----------|---|-----------|
| 4 | Methodology | 16 |
| 4.1 | Research Design | 17 |
| 4.2 | Data Collection | 17 |
| 4.3 | Data Exploration | 17 |
| 4.4 | Data Processing | 18 |
| | 4.4.1 Dataset Analysis | 18 |
| | 4.4.2 Data Cleaning and Preparing | 18 |
| 4.5 | Applying Different Algorithms | 19 |
| 5 | Result Analysis | 21 |
| 5.1 | PCA with K-means | 21 |
| 5.2 | PCA with K-means++ | 22 |
| 5.3 | RFM with K-means++ | 24 |
| 5.4 | LRFM with K-means++ | 25 |
| 5.5 | Product-Based Clustering | 26 |
| 5.6 | State-Based Clustering | 27 |
| 5.7 | Data Insights | 29 |
| | 5.7.1 Comparison | 29 |
| | 5.7.2 Equipment Distribution in Industry | 30 |
| | 5.7.3 Order Placement in Different Market | 31 |
| | 5.7.4 Profit and Sales Comparison | 32 |
| 6 | Conclusion | 33 |
| 6.1 | Research overview | 33 |
| 6.2 | Contribution and Impact | 33 |
| 6.3 | Future Work | 34 |
| | Bibliography | 37 |

List of Figures

| | | |
|------|---|----|
| 3.1 | Recency Frequency Monetary (RFM) Model | 10 |
| 3.2 | Clustering data using the K-means algorithm | 12 |
| 3.3 | Determining no. of clusters with the Elbow method | 13 |
| 3.4 | Overview of PCA | 14 |
| 4.1 | Work plan Flow Chart | 16 |
| 4.2 | Distribution of Data Based on Location | 18 |
| 4.3 | Distribution of Data Based on Category | 18 |
| 5.1 | Elbow Method For K-means | 21 |
| 5.2 | K-means Normal Scatter Plot | 22 |
| 5.3 | Elbow Method For K-means ++ | 22 |
| 5.4 | K-means++ Scatter Plot | 23 |
| 5.5 | Elbow Method For RFM with K-means++ | 24 |
| 5.6 | After K-means++ Scatter Plot | 24 |
| 5.7 | Elbow Method For LRFM with K-means++ | 25 |
| 5.8 | After K-means++ Scatter Plot | 25 |
| 5.9 | Product demand for cluster 0 | 26 |
| 5.10 | Product demand for cluster 1 | 26 |
| 5.11 | Product demand for cluster 2 | 26 |
| 5.12 | State-Based demand for cluster 0 | 27 |
| 5.13 | State-Based demand for cluster 1 | 27 |
| 5.14 | State-Based demand for cluster 2 | 28 |
| 5.15 | Dataset of cluster 0 | 29 |
| 5.16 | Dataset of cluster 1 | 29 |
| 5.17 | Dataset of cluster 2 | 29 |
| 5.18 | Equipment Distribution in Industry | 30 |
| 5.19 | Order Placement in Market | 31 |
| 5.20 | Dataset Distribution of Market | 31 |
| 5.21 | Profit Comparison Between Clusters | 32 |
| 5.22 | Sales Comparison Between Clusters | 32 |

Chapter 1

Introduction

1.1 Overview

Initializing, any form of business revolves around its potential customers and sales. In this modern world, the competition among business industries is intense. All industries struggle with lowering the total cost and optimizing customer service to amplify added value (Martin Christopher 2011, 5-6.) [10]. To ensure better service and customer satisfaction, companies need to identify their customers' needs and wants to offer preferable products and services to them. For the past few decades, on-line shopping popularity has been increasing immensely. Companies have huge data collection on customers' purchase history. There has been a good deal of research on utilizing this data for the betterment of the companies' future. By analyzing the available stored data from within the company database, marketers can develop strategies to maintain the customer loyalty of existing customers and attract new customers. And that is expected to increase the aggregate sales of the company. In this regard, researchers took an interest in Customer segmentation. Customer segmentation is a process of grouping companies' customers based on some factors (age, interest, spending history). Companies can target their customers by categorizing customers' details and group their customer base into smaller segments. This process requires a company to assemble certain data about the customers and analyze them to figure out patterns to generate segments. Customer segmentation can provide a better understanding of customer needs. Clustering, an unsupervised learning method that identifies patterns in a dataset, is a prevalent activity. K-means, Mean-Shift Clustering, DBSCAN, OPTICS, Hierarchical Clustering, and others are well-known clustering methods. In this research, we have segmented customers using the k-Means clustering algorithm. The K-means algorithm is the most popular worldwide. Using the K-means clustering technique, groups are extracted from unlabeled data. This algorithm can be used to identify what types of groups exist or to acknowledge unknown groups in complex data sets. When we are given raw data from a database, it can be random and uninformative. RFM, LRFM, and PCA analysis are applied to identify which customers in which market regions have which type of characteristics. Based on these values, the K-means algorithm is executed.

1.2 Problem Statement

While mass marketing tactics can be effective, assuming that everyone wants to buy what people are selling is a time-consuming, inefficient, and costly strategy. Successful segmentation, rather than employing a one-size-fits-all approach, categorizes consumer data based on similar resources or behavioral factors, enabling for more complicated advertising and personalization campaigns for more timely, relevant results. Marketing communications that are both substantive and dependable.

The main purpose of this research is to create a system that will segment the customers of any given organization using raw data from their dataset. So that, according to those segment characteristics and insights, business owners will be able to make better-informed decisions in order to optimize sales.

1.3 Research Objectives

This research aims to develop a segmentation system by which companies can analyze their customers' behavior and find their purchase patterns using either PCA with the K-means or LRFM with the K-means or RFM with the K-means model. Based on those patterns, using the K-means algorithm, they will be able to divide customers into segments. Marketers can come up with particular marketing strategies for particular segments. Companies will be able to capture new targeted customers, focus on the customers who are on the verge of churning out, as well as increase the customer lifetime value (CLV). Businessmen will be able to find out which regions bring in more sales or profit. This whole process is targeted to increase the gross profit of the particular companies.

1.4 Customer Segmentation

1.4.1 Description

Customer Segmentation is a method of dividing a company's customers into groups that reflect common traits among customers in each group [8]. Segmentation creates an opportunity for a particular company to communicate with its customers to determine their needs and wants to intend to increase profits. It allows a company to identify its most valuable customers and why. According to an eCommerce platform's ebook, Magento [18], customer segmentation is the process of dividing consumers into groups with similar characteristics. Customer segmentation provides several advantages: it helps us find out the similar kind of customers and offer similar products; Based on consumer data, it alters the way we communicate with customers; it recognizes the most profitable customers, and it facilitates us to update the products and services to meet customer needs. Baer [13], in his research, Customer Segmentation Intelligence, states that customer Segmentation enhances marketing by offering products or services that accommodate each customer group.

1.4.2 Methods of Customer Segmentation

Customer segmentation may be implemented in a number of different ways. Schneider divided consumer segmentation approaches into four groups: geographic (based on location), demographic (based on age, gender, family size, income, education, and religion), psychographic (based on social class and personality), behavioral (based on customer behavior), and usage-based (based on browser, buyer and shopper). Similar to Schneider, Magento [18] categorizes its customer segmentation tactics as Profit Potential (average order value), Past Buy (buy history), Demographic, Psychographic, and Behavior. Baer [13] classifies customers based on the business rule approach, the quantile membership technique, supervised clustering using the decision tree method, and unsupervised clustering with the k-means algorithm.

Chapter 2

Literature Review

In this chapter of our report, we've discussed some of the previous research papers that we've looked into. Here are a few examples of prior projects:

2.1 Overview

In today's business world, customer relationships are gaining importance with paradigm-shifting. The relationship between customer needs and providing service is a complex task to manage indeed. Because customers' demand differs from person to person based on their demographic, gender, age, geography, and personal choice. If an organization treats all of its customers in the same manner, it will not be very beneficial. However, the marketing world nowadays focuses more on gaining those customers who will be beneficial for the business in the long run. Customer segmentation is a highly appreciated method to get valued customers and provide services according to customers' needs. In each segment, a small group of people have the same interests. They show similar market behavior, so it becomes easy to treat them according to their preferences, which will bring competitive advantages for both parties.

2.2 Related Works

To begin, we have reviewed some papers on customer segmentation by K-means clustering and the RFM method. Firstly, we want to irradiate a paper named "A novel Centre Point Initialization Techniques for K-means Clustering Algorithm" [16] by Dauda Usman and Ismail Bin Mohamad emphasizes on k-means clustering method. On the other hand, it demonstrates the potential answer of adding a new method in k-means clustering in order to increase the algorithm's effectiveness and efficiency. We know that the k-means method loses effectiveness because an initial point is assumed randomly, so they have suggested a new method that can help get the exact initial point.

Sukela Goyan argues in the journal “Business and Management the Basis of Market Segmentation: A Critical Review” [12] that customer segmentation is a business strategy in which tiny groups are segregated homogeneously for a specific goal. On a list of customer lifetime value, few members who are specialists in this field mentioned the Pareto/NBD Counting Customers model provided by Colombo and Morrison (1987). In 68 models, the authors study customer purchase behavior in situations that result in 70 instances of client purchase abandonment. Nonetheless, the model is effective for analysis. However, this scenario cannot account for more than 70 clients’ behavior, therefore, 73 demonstrates greater complexity. On the other hand, 74 researchers have demonstrated that their implementation claims are accurate [12].

After that, we have reviewed another paper on customer segmentation named “The Impact of Big Data Market Segmentation” [23] by Fahed Yoseph and Nurul Hashimah, where they have described big data and its advantages. It has been used for data predictions, enhancing decision-making ability, customer behavior in marketing, and providing services like fraud control, banking, weather forecasting, and health care. Along with advantages, it also has some disadvantages regarding privacy, security methods, easy accessibility, and the inability to keep track. Such as, keeping data in the POS transactions only shows a single consistent view. Again, sophisticated statistical methods are costly for any business to use, which is not expected. Moreover, Big Data Analysis using data mining with clustering methods is in its initial state to use. On the other hand, complexity and ineffectiveness rather than other methods drawbacks work upon it. From this report, we have reviewed that k-means clustering can be a better option to identify customers using the RFM model. According to Ballestar and Grau-Carles Sainz, unsupervised classification involves feature vectors, data items, and observations in different small segments or clusters, which is clearly shown in the journal named “Customer segmentation in e-commerce: Applications to the cashback business model” [20]. This clustering algorithm works based on objects which are more related to nearby objects. A business analysis was made by using these methods. We can see from the journal ‘A review of the application’ by Jo Ting Wei, Shis Yen Len and His hung Wu of RFM model [21]. They used the RFM model to know customer behavior because, in this recent era, they realized retaining customers play a significant role for a business. This RFM technique is widely utilized in a variety of fields, including for-profit and nonprofit organizations. It is used to determine customer profitability as well. With the advantages of this model, they discussed some drawbacks as well. RFM models can’t give significantly better results or give less priority in terms of choosing customers who have less purchase amount and purchased a long time ago. It’s important because, in a firm, it’s believed that only 20 % of customers purchase 80 % of the company’s sales. Then they pointed out some other disadvantages as well. The RFM model used a limited number of variables, which is not better for segmentation. This paper shows the advantages and disadvantages of RFM models and compares them with some other models. Each model has advantages and drawbacks. Though this model has some disadvantages with extended RFM and some minor extensions and modifications, these issues can be solved [21].

In the telecommunication industry, it's essential to segment data as it's vast in amount. They need to segment in a manner so that it becomes most significant in deciding the service which will be offered to a particular group. We came across a paper named "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA" [22], where they segmented customers based on individuals, spatial and social network features. They reduced the optimized unlevelled data using PCA transformation and an autoencoder neural network built with Keras Tensor flow. When the data set is reduced in principal data set and reduced the dimension they performed, K means clustering method on the data. For the final evaluation, they used two different measures for the algorithm. One is the external cluster, and another one is the internal cluster. They compared the result from clustering to an external validator that knows the correct number of clusters in advance. Secondly, they use an internal cluster validation process to evaluate the performance structure to estimate better algorithm methods without any external data. They evaluated performance using internal indices. They got the best data using the mentioned two methods. Though autoencoders and PCA share similarities, the PCA method worked better in this case because it will bring some benefits to autoencoders like vanishing and exploiting gradients and regularization [22].

We looked into the paper "A comparison of k-means and mean shift algorithms". K means algorithm is a special kind of algorithm where we have to determine an initial centroid point. [11]. It requires the determination of cluster numbers. The center of the cluster keeps changing, so the central value is not unstable [9]. Here, center value works as the basis of cluster determination. And the initial centroid cluster shows a significant impact on developing clusters. K-means works-First of all, determine K (number of clustering) and maximum iteration. Then they perform an initialization process that will process clusters. They have to find an observant from the nearest cluster that we will connect using euclidean space measurement. Then relocation on the distance between each data centroid of the cluster. After that, it recalculates the midpoint position. If there's any change in the midpoint iteration number, or as long as it's smaller than the maximum iteration number, the process will go back to three. Otherwise, it will provide the result.

K-means clustering was used in a bank customer segmentation based on their credit payment similarity [17]. A decision tree was used as a classification technique to predict how close the customer is to falling into the targeted group. Another research illustrates the segmentation of customer profile data based on their usage of online banking using k means and k medoid, where k means was more effective than another method in achieving the desired outcome. They built a clustered method based on RFM segmentation. They then divided customers into six categories based on their buying behavior so that banks can apply CRM strategies to win more customers [19].

Furthermore, the RFM method is also highly used in banking, insurance companies, the marketing industry, and other profitable organizations. We have reviewed an-

other work named “Developing a Model for Measuring Customer Loyalty and Value with RFM Techniques and Clustering Algorithm” [14] says that it gives chances to banks to provide customized service for those customers who are loyal to the company by knowing their behavior by RFM model. Moreover, the RFM method is mostly used and mostly adopted. It depends on three measures. Those are recency, frequency, and monetary. In these three terms, recency has more priority. RFM starts working by sorting each dimension and then divides customers into five segments exactly in equal size. From the paper, we get to know the concept of customer segmentation and the demographic behavior of different customers.

In the banking sector, we found another paper [27], “Segmentation of investor customers using machine learning in banking” by Juho Jääskeläinen- who found K-means clustering as a helpful model for exploring new data. According to this study, the K-means model managed to show some key factors for identifying investor customers and other necessary elements. This study emphasized how clustering is a suitable method to get the first insights from data.

Since the pandemic of COVID-19 emerged, the definition of our lifestyle has been changed, including how people do their shopping according to their requirements. As the pandemic has transformed our lives, it has also introduced a lot of new trends, the most prominent of which may be online shopping. The transition to online shopping commenced prior to the epidemic. Recent IBM figures indicate that the COVID-19 has hastened the transformation by five years. In the article “LRFM model for customer purchase behavior using K-Means algorithm” [28], LRFM and K-Means clustering were used to determine that the success value of online shopping appeared to have been sustained even after the pandemic. “ Consumers’ Online Purchase Intentions and Customer Segmentation During the Period of the COVID-19 Pandemic” [25] investigated consumers’ online purchase intentions in the context of a pandemic and then segmented them based on their intentions and purchasing behaviors using k-means. This study identified three unique consumer categories and concluded that worries of pandemic had a key influence in segmenting customers.

Customer segmentation is helpful in any kind of business/ organization. We wanted to explore more different kinds of fields where customer segmentation can be utilized in marketing. Keeping that in mind, we reviewed the paper “Customer Segmentation for Life Insurance in Iran Using K-means Clustering” [29]. The paper describes how the insurance companies in Iran are lagging compared to the developed countries. To capture the market, the companies need to have a good understanding of the needs and behavior of the customers. Therefore, they have tried to modify and customize their products to meet the needs of different people. To achieve this goal, it is vital to have some idea about the customers. Proper marketing is essential for attracting new customers and also retaining loyal ones. On this point, market and customer segmentation provide a high-level overview of customers’ behavior. K-means clustering was carried out in this study, and four different clusters were obtained containing customers with similar characteristics. From these found clus-

ters, marketers will be able to develop ideas for appropriate marketing strategies. From this paper, we obtained a better understanding of the K-means mechanism.

We have discovered another article titled “Consumer Attitude and Online Purchase Intention: A Segmentation Analysis in Malaysian Halal Cosmetic Industry” [30]. In Malaysia, the Halal Cosmetics Industry has slowed due to a lack of client segmentation data. In this work, a hierarchical cluster analysis was done first, followed by a K-means cluster analysis. Utilizing average factor scores, a hierarchical clustering approach with Ward’s linkage was utilized to determine the number of clusters. This study identified four groupings based on demographics, customer attitudes, and willingness to buy online. And finally, by studying these clusters, marketing plans can be developed.

In another study named “Product recommendation for eCommerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society” [26] PCA method has been applied for segmentation. Mainly it focused on segmentation of customers based on some brand size, brand size and customer income with the help of PCA for reducing dataset dimension. They used the elbow method to determine the optimum number of clusters they made based on these data. They followed five steps to implement PCA to the clustering.

1. They standardize the data set for PCA implementation.
2. Then in the second step, they computed the covariance matrix to identify the correlation between dependencies.
3. They calculated the eigenvectors and eigenvalues from the covariance matrix to get the result of the largest to the smallest eigenvalue of eigenvectors.
4. They gave priority most to the largest component gotten from eigenvalue and eigenvectors, thus getting the principal component. The principal value of lesser significant components is removed.
5. Then after rearranging data of the final principal components, they got the reduced data set.

PCA is important to reduce datasets of similar purchases that helps k means clustering not to overlap data of similar segments. Companies need to analyze static customer behavior when they launch new products. After exploring the result, this method can help brands reach the maximum goal [26].

“Incorporating K-means, Hierarchical Clustering, and PCA in Customer Segmentation” [24] used both k-means and Hierarchical clustering. Based on their dataset’s

Davis-Bouldin, Silhouette, and Dunn indices, K-means clustering fared better than Hierarchical clustering. In addition, it has been shown that Principal Component Analysis (PCA), an unsupervised statistical approach, may analyze data to identify patterns and reduce the dimensionality of a dataset with minimum information loss. In addition to cluster analysis, consumer segmentation, and group recognition, PCA may also be utilized for cluster analysis. This study found that, unlike K-means, PCA is not an instant solution for clustering; nevertheless, it can improve the results of K-means clustering by detecting additional clusters or patterns in the data.

In the article “Customer Mobile Behavioral Segmentation and Analysis in Telecom Using Machine Learning” [31], the K-Means method was used to cluster mobile behavioral data from telecom customers into meaningful and labeled segments. During the data preprocessing phase, the PCA methodology was utilized to minimize the dataset, and the Elbow method was utilized to determine the appropriate number of clusters. After clustering the data into four categories using the K-means technique, the results were evaluated and labeled. This study was able to attain a good Silhouette Score for clustering accuracy.

To conclude, we have reviewed these papers to learn some existing methods and the importance of customer segmentation with different methods and techniques. We have realized the importance of this method. So in this project, we have decided to use the RFM, LRFM, PCA and k-means for the clustering algorithm.

Chapter 3

Methods and Models

3.1 Recency Frequency Monetary (RFM) Model

RFM was initially presented by Hughes of the American Database Research Institute in 1994. As a prominent customer value analysis tool, it is extensively used to quantify the value of customers throughout their life cycle, segment consumers, and study their behavior (Cheng Chen, 2009, p. 24) [6]. RFM is a method for determining customer value. RFM is an acronym for Recency, Frequency, and Monetary, which relates to the significance of the most recent purchase, the frequency of purchases, and the monetary worth of the most recent purchase. R (recency) is the interval of time between the date of the customer's most recent purchase and the end date of the statistics. The shorter the interval, the greater the value of R. F (frequency) represents the number of consumer purchases made within the statistical period.

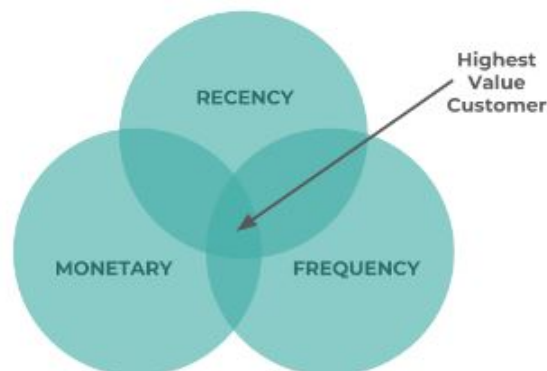


Figure 3.1: Recency Frequency Monetary (RFM) Model

The larger the F-value, the greater client loyalty and propensity to purchase again. M (monetary) reflects the total amount of purchases made by clients over the statistical period. In general, consumer loyalty increases with the overall purchase quantity. It can be utilized as a direct indicator of the manufacturing capability of the sales company. The bigger the value of R or F, the greater the likelihood that the corresponding buyer will transact with the seller again. The bigger the value of

M, the greater the likelihood that the buyer will purchase from the vendor again. Researchers have broadened the application of RFM in customer value analysis to encompass all facets.

3.2 Length Recency Frequency Monetary (LRFM) Model

The LRFM (length, recency, frequency, and monetary) model is an extension of the RFM model. Despite the numerous studies that have been conducted on this subject, some researchers claim that the basic RFM model cannot correctly distinguish between different types of customers based on the length of their relationship (Reinartz Kumar, 2000) [3]. The duration between a customer's first and last purchases is described as the length of the relationship. According to [3], the RFM model cannot detect if a consumer has a long-term or short-term relationship with the organization. Long-term customer relationship management generates customer loyalty based on the relationship between a company and its consumers (Chang and Tsay ,2004) [5]. As a result, Chang and Tsay [5] expanded the RFM model to the LRFM model by factoring in length (L). The LRFM model is considered as a data mining tool in the CRM system (Ngai, Xiu, Chau, 2009) [7]. For the management to identify the essential customers, a combination of the LRFM model and cluster analysis would be beneficial. The administration can then design various marketing tactics in order to optimize consumer value.

3.3 Clustering Algorithm

Clustering is the process of grouping a collection of physical or abstract items into groups of objects with similar characteristics. The clustering method is an unsupervised machine learning approach that detects closely related sets of data points. Clustering is applied to an unlabeled data collection in which the relationship between the variables is unknown. It is a technique for classifying unlabeled data based on similarities. Clustering can be used to segment customers for further study. Standard method for cluster analysis is k-means clustering.

3.4 K-Means Algorithm

The K-Means algorithm is one of a number of well-known clustering algorithms. McQueen initially employed it in 1967 [1]. Numerous industries, including data mining, statistical data analysis, and commercial applications, make extensive use of the K-Means algorithm.

The K-means method is an iterative algorithm that seeks to divide a data set into non overlapping subgroups (clusters). It attempts to make the data points inside the cluster as similar as possible, while making the cluster as distinct (far) as possible. Assign the data points to a group such that the total of the squares of the distance between the data points and the group's center of gravity (the arithmetic mean of all the data points in the group) is equal to The more homogeneous (similar) the data points inside the cluster, the fewer changes that occur within the cluster. Customer segmentation is one of K-Means' primary uses. (Mesforoush A. and Tarokh M. J., 2013) [15].

The K-Means algorithm has a faster calculation speed and minimizes the incidence of data misclassification when compared to other clustering techniques. Therefore, we cluster RFM attributes using the K-Means technique. The algorithm's precision is dependent on the initialization circumstances and cluster count.

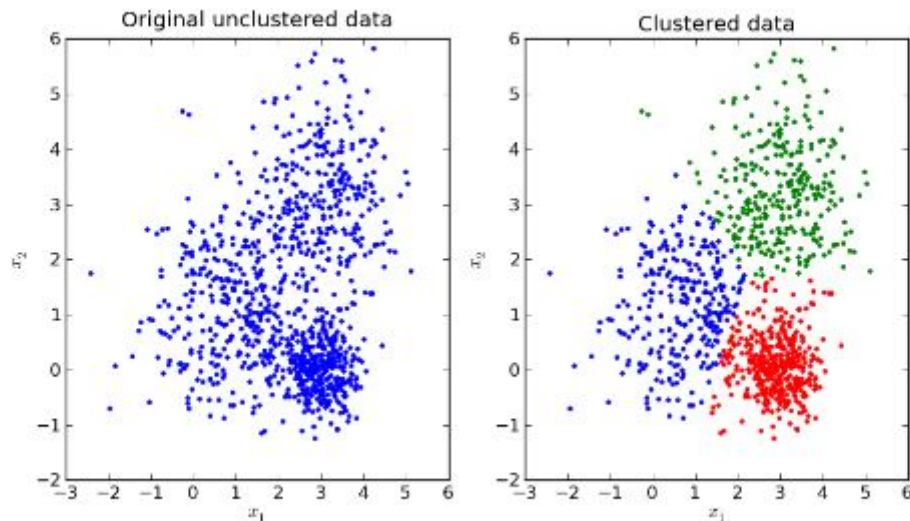


Figure 3.2: Clustering data using the K-means algorithm

3.5 K-Means ++

K-Means algorithm is a centroid-based clustering technique. Each cluster in this clustering technique is represented by a centroid point. After finding out the k number of centroids, The K-means algorithm assigns every data point to the closest cluster. The centroid point is the mean of all the points in the dataset and keeps changing in each step. In standard K means clustering, the centroids are initialized randomly by selecting k random data points from the dataset. This random approach is potentially problematic because if the initialization of clusters is not appropriate, K-Means can result in poor, erratic clusters, and we might get different

clusters every time. In some cases, more than one centroid might be initialized into the same cluster.

To overcome the drawbacks, K-means++ can be used. The K-means++ is the standard K-means algorithm combined with a more appropriate initialization of the centroids. This algorithm optimizes the step where we randomly pick the cluster centroid and improves the clustering quality. Other than the initialization, the rest of the algorithm is the same as the standard K-means algorithm.

3.6 Elbow Method

Fundamental to any unsupervised approach is determining the ideal number of clusters. The Elbow Method is one of the most well-known approaches for determining the ideal number of clusters in k-means clustering. WCSS (Within-Cluster Sum of Square) is the sum of each point's squared distance from the cluster's centroid. When the WCSS is plotted alongside the K value, the resulting graph resembles an Elbow. The WCSS value will decline if the number of clusters grows.

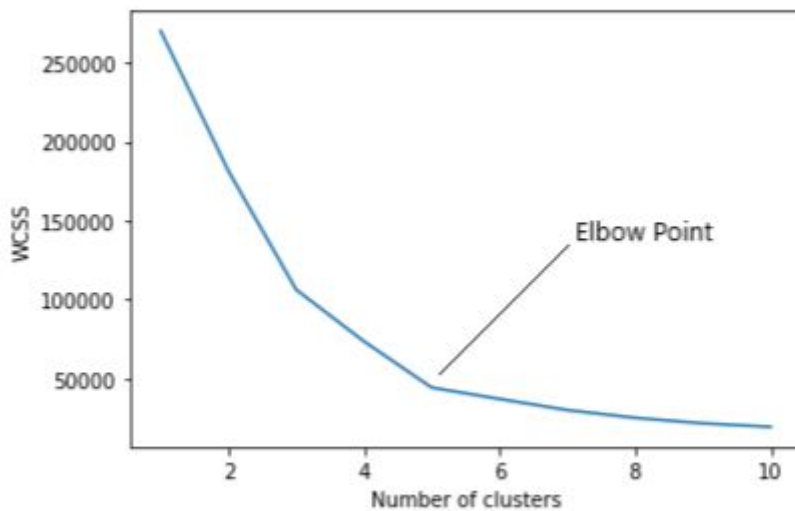


Figure 3.3: Determining no. of clusters with the Elbow method

3.7 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised learning technique. Reducing the number of variables while retaining as much information as feasible, PCA is utilized to identify robust patterns in a huge dataset. It is used to transform enormous data sets into smaller ones for better exploration and analysis. It is a

method for reducing variance and revealing hidden patterns within a dataset. PCA is a method of feature extraction that is modeled. Obviously, reducing the number of variables in a data collection decreases its precision; nevertheless, the secret to dimensionality reduction is to compromise precision for simplicity. Because smaller data sets are more manageable and make data processing considerably simpler and quicker.

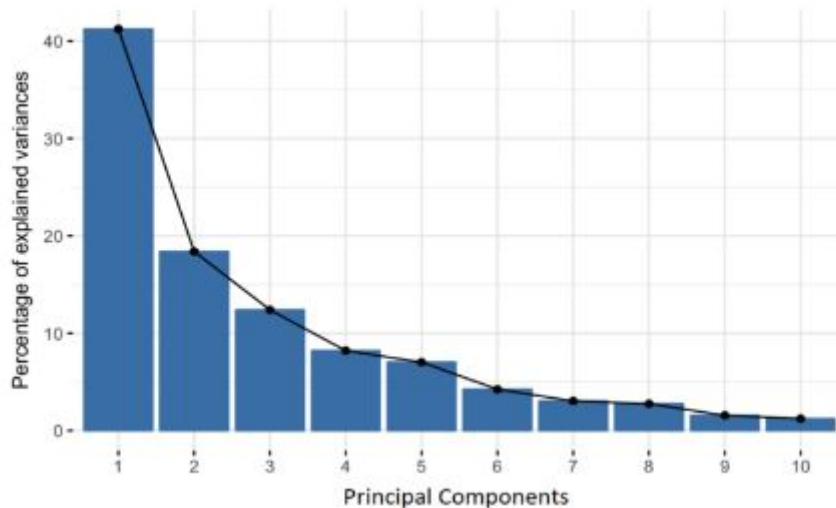


Figure 3.4: Overview of PCA

3.8 Customer Relationship Management

Customer relationship management (CRM) is a software system used by company owners to manage and evaluate customer interactions and related data throughout the lifespan of the customer. The objective is to improve customer service relationships and increase sales. Customer Relationship Management (CRM) software maintains valuable purchase patterns, demographics, and preferences that may be utilized for Customer Segmentation. Several CRM systems give features for segmentation and campaign management to target the appropriate demographics. CRM analyzes sales data and assesses segment-specific patterns in order to provide pertinent insights into customer behavior and attitudes. Connecting CRM software with a customer segmentation strategy can produce even more productive outcomes. Based on a customer's buying history, CRM determines how to target them with product recommendations. CRM aids in the creation of prediction models that identify lucrative clients based on their lifetime value (CLV), a crucial determinant of the net worth of a future engagement with a customer. We may also identify customers with low CLVs and use CRM data to decide how to increase their CLV.

3.9 Customer Lifetime Value (CLV)

Customer lifetime value (CLV) is the total amount of money a customer is predicted to spend with a company over the course of their relationship. Knowing the CLV allows companies to devise strategies for acquiring new consumers and retaining existing ones while preserving profit margins. CLV is a metric for determining how much a company is willing to invest to keep a client relationship intact. For instance, if a customer's CLV is estimated to be \$500, the company will not spend more than that to maintain the relationship. It would simply not be profitable for that company. While calculating CLV, its best to take the total average revenue and average profits made by a customer.

Customer Lifetime Value formula:

$CLV = \text{average purchase value} * \text{average purchase frequency} * \text{average customer lifespan}$

3.10 Silhouette Score

The Rand index [4], corrected Rand index [2], distortion score, and Silhouette index are all ways for evaluating clustering results. The Silhouette index does not require a training set to evaluate the clustering results, unlike most performance evaluation approaches. As a result, it's more suited to a clustering task. The Silhouette index is used in this study to assess clustering performance.

Chapter 4

Methodology

In this chapter, we have provided details of our working process.

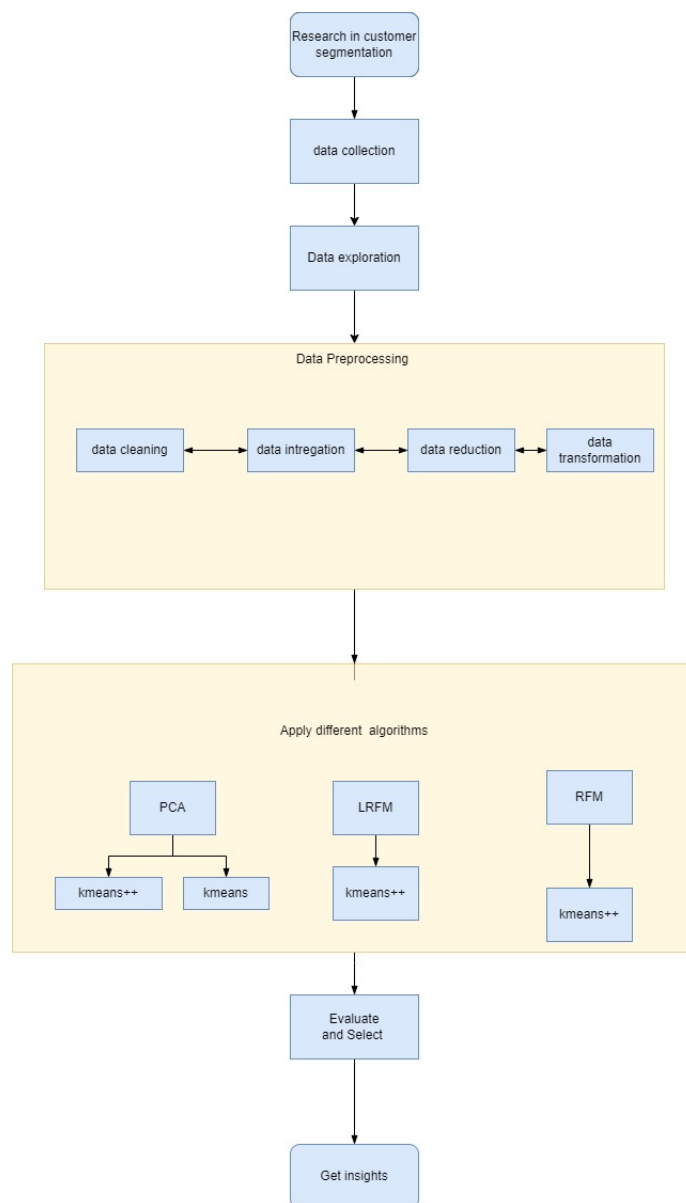


Figure 4.1: Work plan Flow Chart

The proposed system model is inspired by the work of ‘Customer Mobile Behavioral Segmentation and Analysis in Telecom Using Machine Learning’ [31] where PCA and K means are applied , and satisfactory correctness was achieved. Also, we are inspired to use the RFM model from the work of [14] and the extension LRFM model from [28] where LRFM and k means were used to identify the clusters.

4.1 Research Design

A quantitative survey study was conducted, which was based on reviewed literature. The target population in this study is superstore customers.

4.2 Data Collection

Tableau Software, based in the United States, is a developer of interactive data visualization tools that specializes in business intelligence software. The dataset was collected online, then sculpted, and finally published by Tableau Software. They focus primarily on working with the datasets and databases that belong to a variety of different organizations. Tableau facilitates data visualization and comprehension. The way that people utilize data to find solutions to problems is being redefined by their visual analytics technology. Tableau is a tool that helps businesses of all sizes become more data-driven, and it is used by those businesses. Tableau offers a wide variety of services for the development of web portals. It consists of Quick-start features that can be finished in a matter of hours for deployments on a smaller scale. They offer a variety of Tableau products, including desktop software, a server, a reader, and an online service. data.world was utilized in order to obtain the dataset.

4.3 Data Exploration

The dataset has 27 columns and 51290 rows, and 19 out of those 27 columns have been selected to work in this paper. We want our desired clusters to be separated with the description of market2, which is given below.

Market2: This contains the regions where the order was placed. The regions in the dataset are APAC, North America, EMEA, Africa, EU, and LATAM.

4.4 Data Processing

4.4.1 Dataset Analysis

We use charts to explore the distribution of the data based on category and location to see if the data is not too unequally divided.

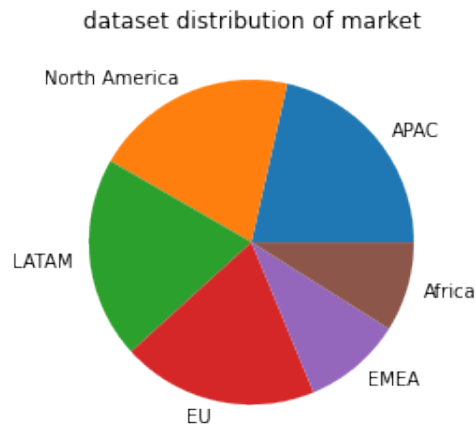


Figure 4.2: Distribution of Data Based on Location

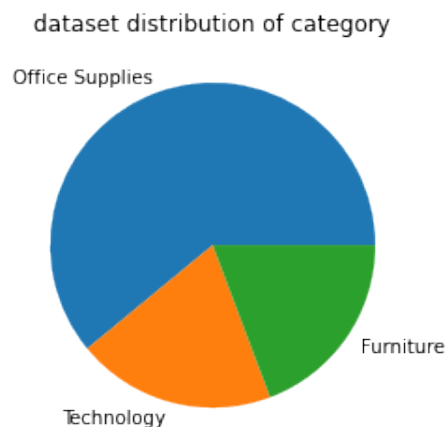


Figure 4.3: Distribution of Data Based on Category

4.4.2 Data Cleaning and Preparing

In the beginning, the functions that are not required will be eliminated. We have put the customer id and market2 data into a separate pandas data frame for use at a later time. Then, utilizing the one hot encoder and dummies package provided by sklearn, we have converted the categorical data to numeric values (one hot encoder is only used for ordinal data.).With the help python date and time package, we derived the shipping date and order date, which are then transformed and added as a new column in our database with the name shipping date.

4.5 Applying Different Algorithms

Step 1: PCA

We used the PCA algorithm to reduce the dimensionality into two dimensions and also to base our clusters on the essential features, which are automatically done by PCA.

Step 2: Elbow Method

We feed in the PCA transformed data into the elbow method to find the optimal number of clusters.

Step 3: Cluster Visualization

We have visualized the clusters with scatter plots for both K-means and K-means++.

Step 4: Choosing Method

We choose the clusters of K-means or K-means++ depending on the scatter plot and group our primary dataset according to those different clusters.

Step 5: Creating LRFM

Creating Length, Recency, Frequency, and Monetary Value columns. We used the panda's group by method to group by customer ID.

L= Length = Last order date – First order date for each customer

R= Recency = The last order date in the dataset-last order date made by the customer

F= Frequency = The number of times a particular customer Id exists in the dataset

M= Monetary = Summing up the total amount of sales for each customer

Step 6: Scaling Data

We used sklearn's standard scaler to scale the LRFM data frame.

Step 7: Elbow method again

We used the elbow method for both LRFM and RFM and notice the number of clusters created in each algorithm.

Step 8: K-means++ visualization for LRFM and RFM

We used a scatter plot and centroid plotting to see how well separated the clusters are.

Step 9: Deciding Best Model

Decide which algorithm we will use with K-means based on the clustering quality which we measured by Silhouette score and choose whether we should use PCA, RFM, or LRFM and use that model to get our Data insights.

Step 10: Data Insights Finding

We used the pandas to describe a method to compare the statistical summary of each cluster group. Then we have checked the Frequency of labels for shipping mode, country, market region, category, and sub-category in each cluster and compare the results.

Chapter 5

Result Analysis

5.1 PCA with K-means

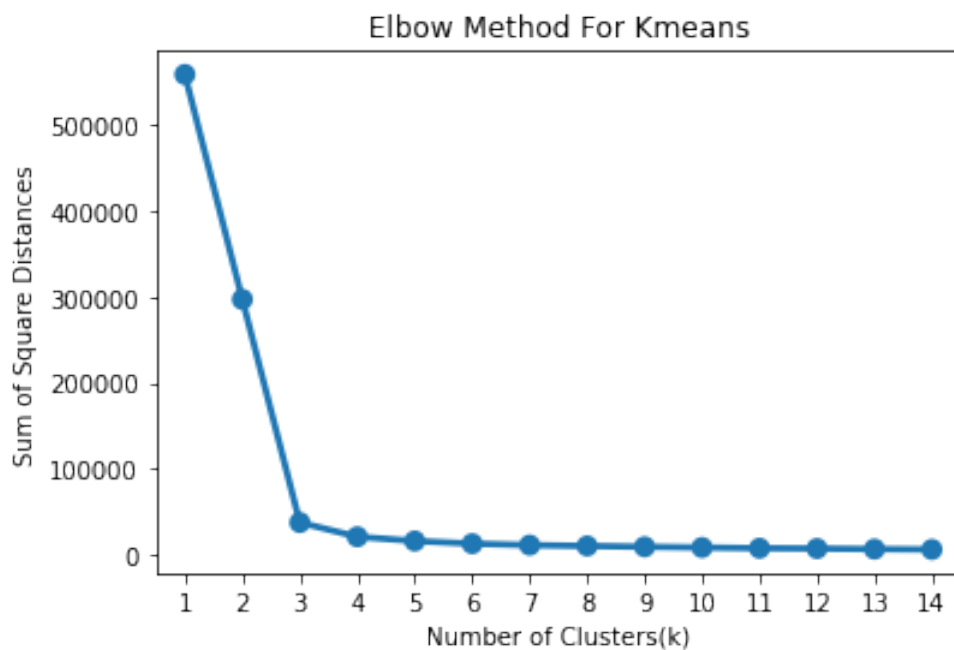


Figure 5.1: Elbow Method For K-means

The Elbow Method is then utilized to calculate the ideal number of clusters. Figure (5.1) depicts the implementation of the Elbow technique. The results of the Elbow approach indicate that the best number of K is three, as illustrated in Figure (5.1), in which abrupt decreasing happens from cluster 1 to 3. Consequently, the optimal number of clusters is 3 ($K = 3$), which is taken into account by the K-means algorithm.

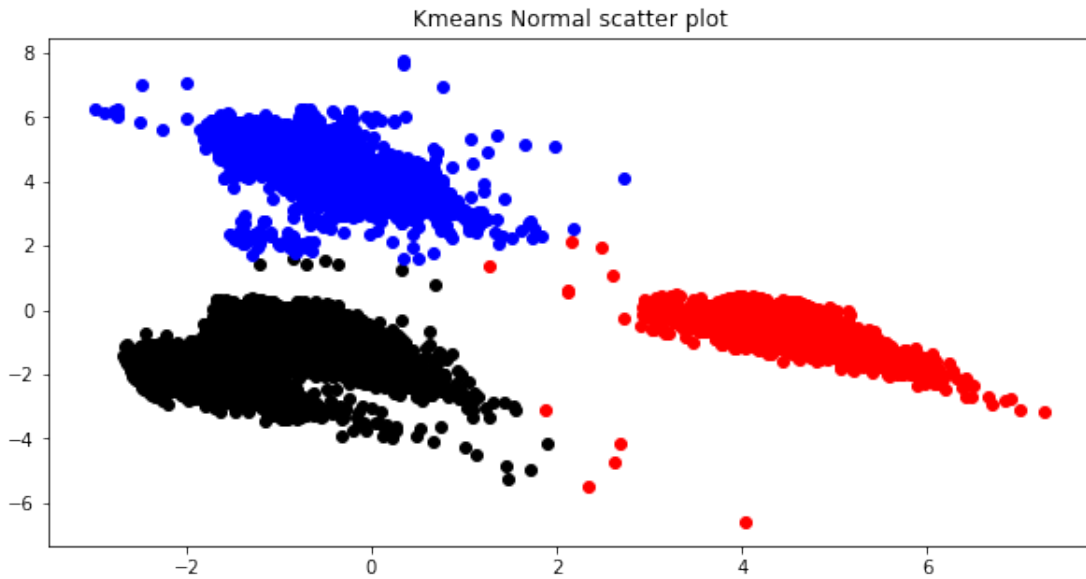


Figure 5.2: K-means Normal Scatter Plot

In figure (5.2), we can see that the scatter plots of K-means++ are significantly far away from each other. From this, we can conclude that the labeling is near to be accurate.

For $k=3$, silhouette score was 0.371.

5.2 PCA with K-means++

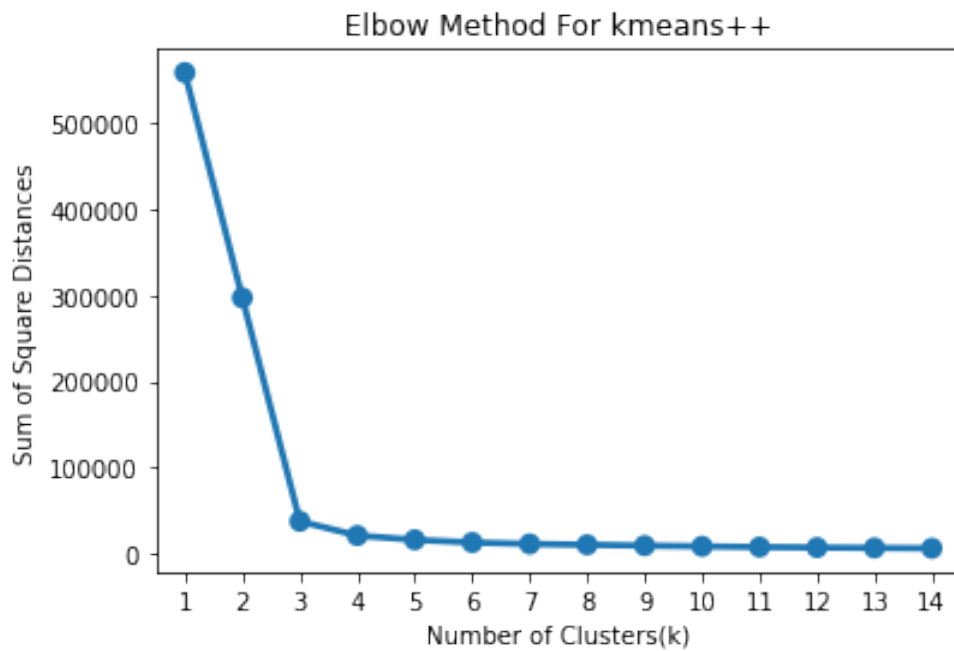


Figure 5.3: Elbow Method For K-means ++

Figure (5.3) depicts the implementation of the Elbow technique. The results of the Elbow approach indicate that the best number of K is three, as illustrated in Figure (5.3), in which abrupt decreasing happens from cluster 1 to 3. Consequently, the optimal number of clusters is 3 ($K = 3$), which is taken into account by the K-means++ algorithm.

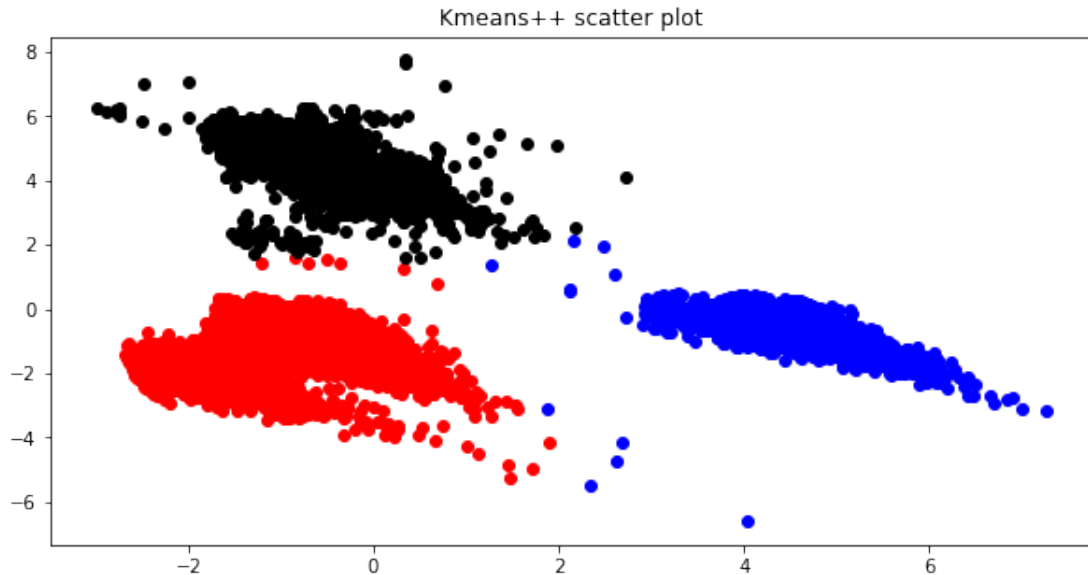


Figure 5.4: K-means++ Scatter Plot

In figure (5.4), we can see that the scatter plots of K-means++ are significantly far away from each other. From this, we can conclude that the labeling is near to be accurate. We can see that both K-means and K-means++ performed the same.

So we will consider just K-means++. The scatter plot shows the clusters have a good distance between each other and each cluster as its data points are mostly packed together.

For $k=3$, silhouette score was 0.817.

Since K-means++ has performed much better, we only decided to test the rest of the experiment with K-means++

5.3 RFM with K-means++

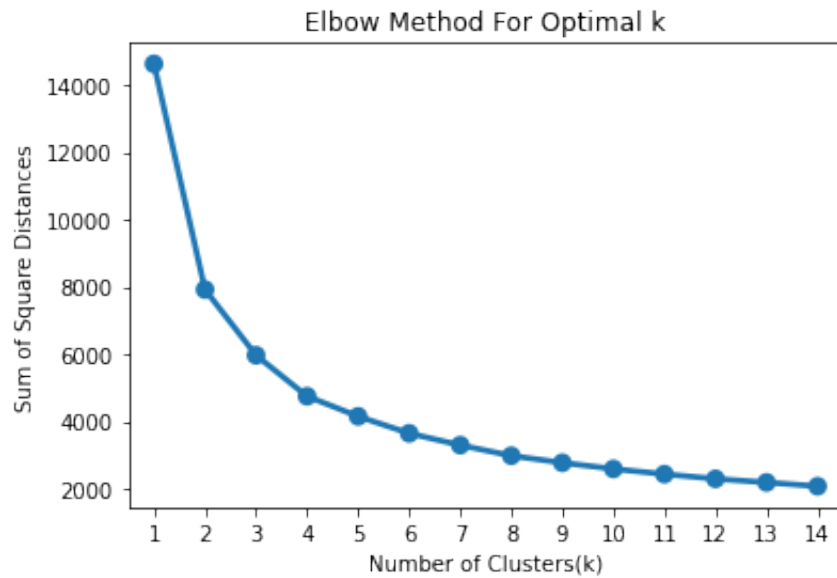


Figure 5.5: Elbow Method For RFM with K-means++

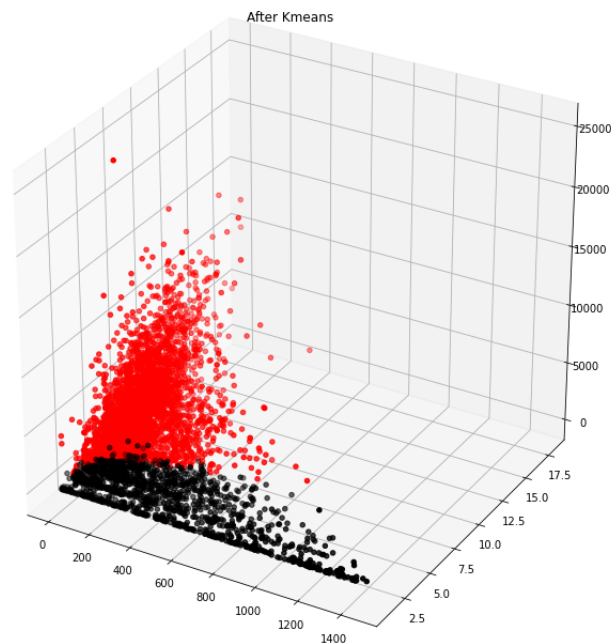


Figure 5.6: After K-means++ Scatter Plot

Here, in figure (5.5), the elbow shows that the number of clusters is 2. The clusters' data points are closed-packed, but the distance between the clusters is almost non-existent, as shown in figure (5.6).

For $k=2$, silhouette score was 0.420.

5.4 LRFM with K-means++

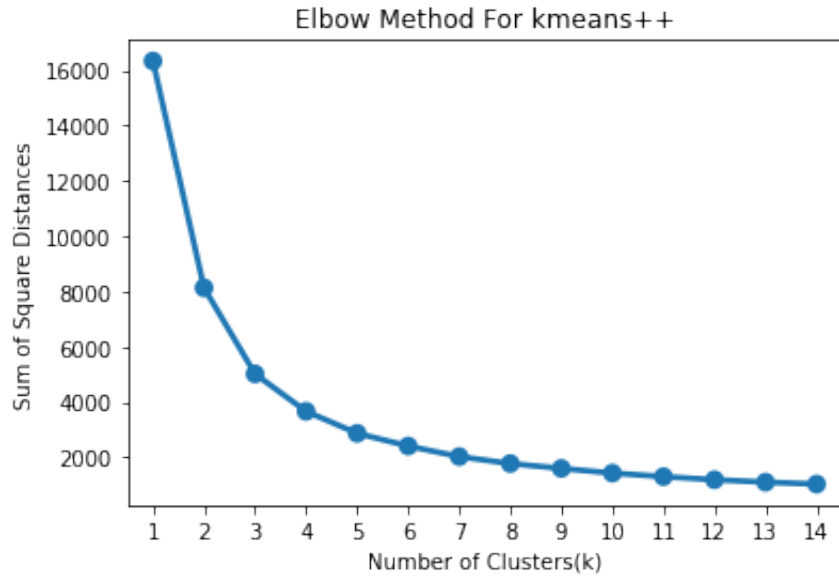


Figure 5.7: Elbow Method For LRFM with K-means++

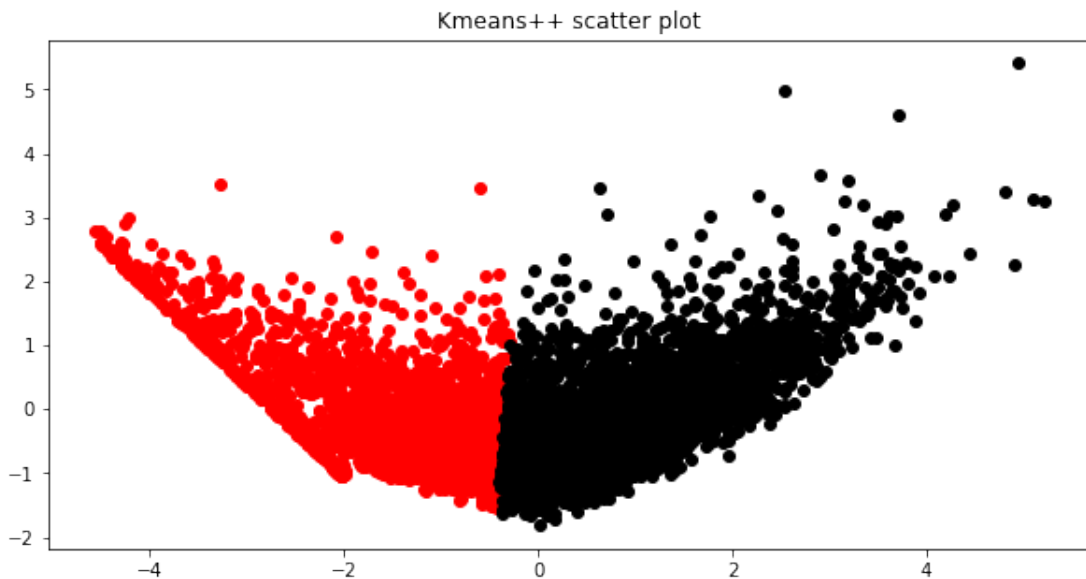


Figure 5.8: After K-means++ Scatter Plot

The elbow here in figure (5.7) also shows us that the number of clusters is 2. The clusters' data points here are also closed-packed like with RFM, but the distance between the clusters is almost non-existent, as is shown in figure (5.8).

For $k=2$, silhouette score was 0.420.0.379.

PCA and K-means++ has a good distance in clusters between them, and the data points in the clusters are mostly packed. Considering the Silhouette score and observing the clustering plots, we choose to go with PCA and K-means++.

5.5 Product-Based Clustering

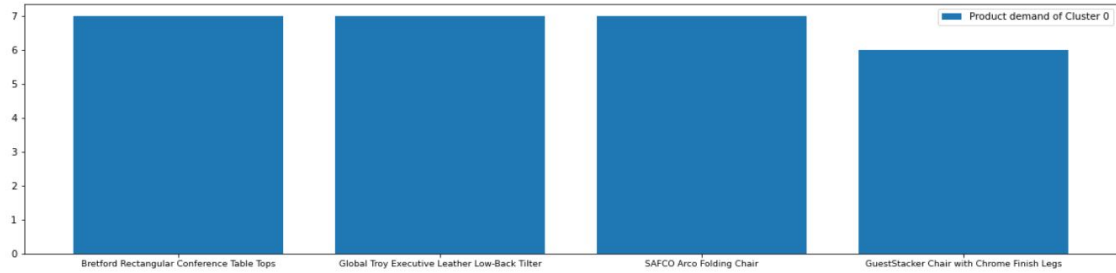


Figure 5.9: Product demand for cluster 0

In figure (5.9), we can see four popular products such as Bretford Rectangular Conference Table, Global Troy executive Leather Low Back Tilter SAFCO Arco Folding Chair, GuestStacker Chair with Chrome Finish Legs which are mostly demandable by cluster 0 on the market. Other products of the market are not so much demandable for cluster 0.

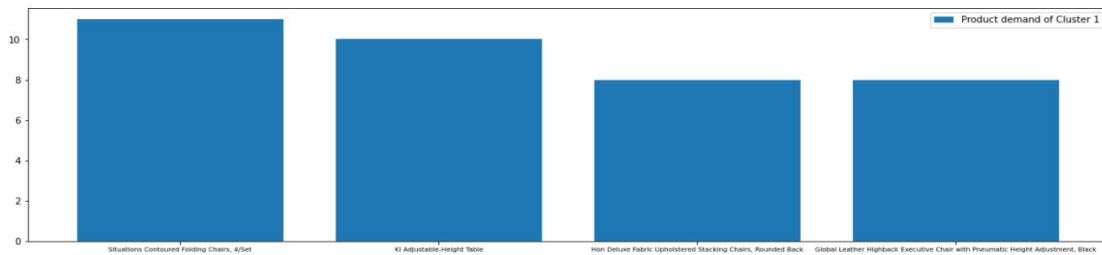


Figure 5.10: Product demand for cluster 1

In figure (5.10), we can see the four products on the market of the second cluster, such as Situations Contoured Folding Chairs, Adjustable Height Table, upholstered stacking chairs, Executive chairs are so much demanded.

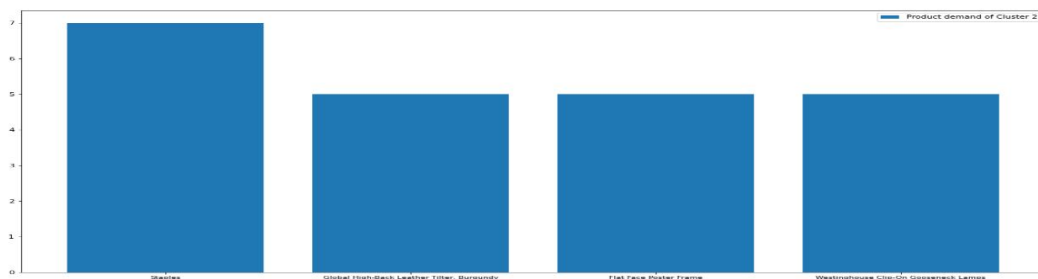


Figure 5.11: Product demand for cluster 2

In figure (5.11), we can see four types of products such as Staples, Global High-Black Leather Tilter, Flat Face Poster Frame, Westinghouse Clip-On Gooseneck Lamps, which are highly demandable by cluster 2. They prefer to buy this kind of product from the market, and other products are not so much preferable.

5.6 State-Based Clustering

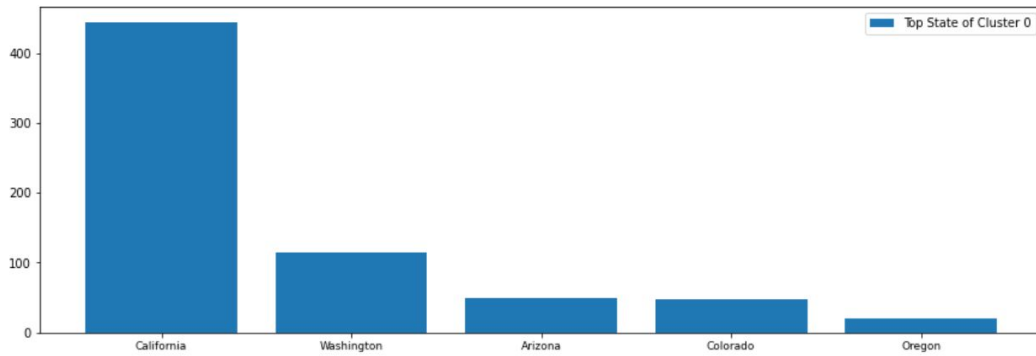


Figure 5.12: State-Based demand for cluster 0

In figure (5.12), there are five states which are California, Washington, Arizona, Colorado, and Oregon. Customers of cluster 0 are highly interested in doing their shopping from these five states. Here we can see that most of the customers are interested in doing their shopping in California.

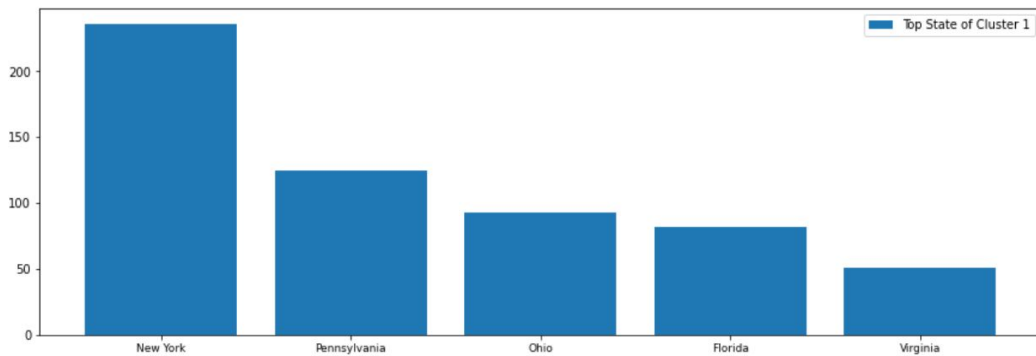


Figure 5.13: State-Based demand for cluster 1

In figure (5.13), there are five states which are New York, Pennsylvania, Ohio, Florida, and Virginia. Customers of cluster 1 are highly interested in doing their shopping from these five states. Here we can see that most of the customers are interested in doing their shopping from New York, then Pennsylvania, then Ohio, and so on.

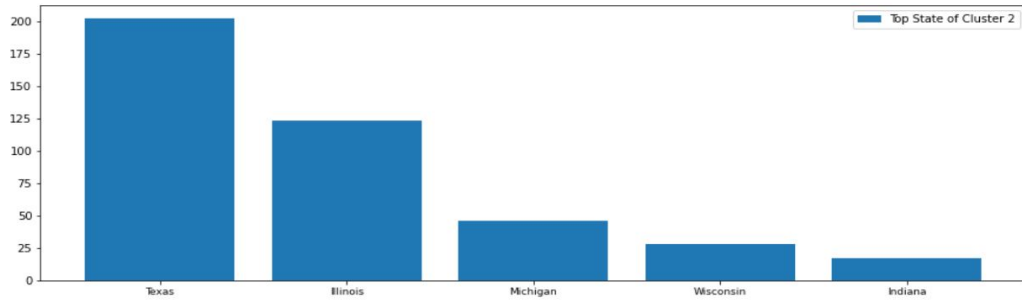


Figure 5.14: State-Based demand for cluster 2

In figure (5.14), there are five states which are Texas, Illinois, Michigan, Wisconsin, and Indiana. Customers of cluster 2 are highly interested in doing their shopping from these five states. Here we can see that most of the customers are more interested in doing their shopping from Texas than in Illinois and Michigan and so on.

5.7 Data Insights

5.7.1 Comparison

```
df_cluster_0.describe()
```

| | discount | profit | quantity | sales | shipping_cost | year | weeknum | clusters |
|--------------|--------------|--------------|--------------|--------------|---------------|--------------|--------------|----------|
| count | 30674.000000 | 30674.000000 | 30674.000000 | 30674.000000 | 30674.000000 | 30674.000000 | 30674.000000 | 30674.0 |
| mean | 0.129990 | 28.664839 | 3.749462 | 242.227391 | 25.686951 | 2012.765208 | 31.884821 | 0.0 |
| std | 0.193356 | 169.799384 | 2.211708 | 478.682300 | 54.703865 | 1.097785 | 14.301845 | 0.0 |
| min | 0.000000 | -6599.978000 | 1.000000 | 0.000000 | 0.002000 | 2011.000000 | 1.000000 | 0.0 |
| 25% | 0.000000 | 0.880000 | 2.000000 | 31.000000 | 2.589000 | 2012.000000 | 21.000000 | 0.0 |
| 50% | 0.000000 | 9.840000 | 3.000000 | 85.000000 | 7.757000 | 2013.000000 | 34.000000 | 0.0 |
| 75% | 0.200000 | 36.433500 | 5.000000 | 252.000000 | 24.534500 | 2014.000000 | 45.000000 | 0.0 |
| max | 0.850000 | 5039.985600 | 14.000000 | 22638.000000 | 933.570000 | 2014.000000 | 53.000000 | 0.0 |

Figure 5.15: Dataset of cluster 0

```
df_cluster_1.describe()
```

| | discount | profit | quantity | sales | shipping_cost | year | weeknum | clusters |
|--------------|-------------|--------------|-------------|-------------|---------------|-------------|-------------|----------|
| count | 9604.000000 | 9604.000000 | 9604.000000 | 9604.000000 | 9604.000000 | 9604.000000 | 9604.000000 | 9604.0 |
| mean | 0.177572 | 12.983994 | 2.290608 | 163.013432 | 18.042296 | 2012.816222 | 30.121408 | 1.0 |
| std | 0.289561 | 146.781816 | 2.121674 | 331.599046 | 43.582826 | 1.096919 | 14.606956 | 0.0 |
| min | 0.000000 | -4088.376000 | 1.000000 | 1.000000 | 0.020000 | 2011.000000 | 1.000000 | 1.0 |
| 25% | 0.000000 | -4.143000 | 1.000000 | 20.000000 | 1.770000 | 2012.000000 | 19.000000 | 1.0 |
| 50% | 0.000000 | 6.180000 | 1.000000 | 52.000000 | 5.100000 | 2013.000000 | 32.000000 | 1.0 |
| 75% | 0.600000 | 26.460000 | 2.000000 | 156.000000 | 15.632500 | 2014.000000 | 43.000000 | 1.0 |
| max | 0.700000 | 1622.880000 | 14.000000 | 5211.000000 | 903.040000 | 2014.000000 | 53.000000 | 1.0 |

Figure 5.16: Dataset of cluster 1

```
df_cluster_2.describe()
```

| | discount | profit | quantity | sales | shipping_cost | year | weeknum | clusters |
|--------------|--------------|--------------|--------------|--------------|---------------|--------------|--------------|----------|
| count | 11012.000000 | 11012.000000 | 11012.000000 | 11012.000000 | 11012.000000 | 11012.000000 | 11012.000000 | 11012.0 |
| mean | 0.148659 | 42.089877 | 3.750636 | 331.206048 | 35.562654 | 2012.776607 | 30.638849 | 2.0 |
| std | 0.176309 | 205.128390 | 2.264084 | 599.833570 | 71.782766 | 1.103180 | 14.539961 | 0.0 |
| min | 0.000000 | -3009.435000 | 1.000000 | 3.000000 | 0.010000 | 2011.000000 | 1.000000 | 2.0 |
| 25% | 0.000000 | -2.983500 | 2.000000 | 47.000000 | 3.960000 | 2012.000000 | 20.000000 | 2.0 |
| 50% | 0.100000 | 10.965000 | 3.000000 | 123.000000 | 11.200000 | 2013.000000 | 33.000000 | 2.0 |
| 75% | 0.270000 | 48.591000 | 5.000000 | 349.000000 | 33.512500 | 2014.000000 | 43.000000 | 2.0 |
| max | 0.800000 | 8399.976000 | 14.000000 | 17500.000000 | 923.630000 | 2014.000000 | 53.000000 | 2.0 |

Figure 5.17: Dataset of cluster 2

Figures 5.15, 5.16, and 5.17 illustrate the sale of products and their discount. In cluster 2, It can be observed that even after having a 20% (approximate) lower mean discount, the sales of the products are doubled compared to cluster 1. Again having a 20% (approximate) higher mean to discount, the sales of the products are increased compared to the clusters 0.

5.7.2 Equipment Distribution in Industry

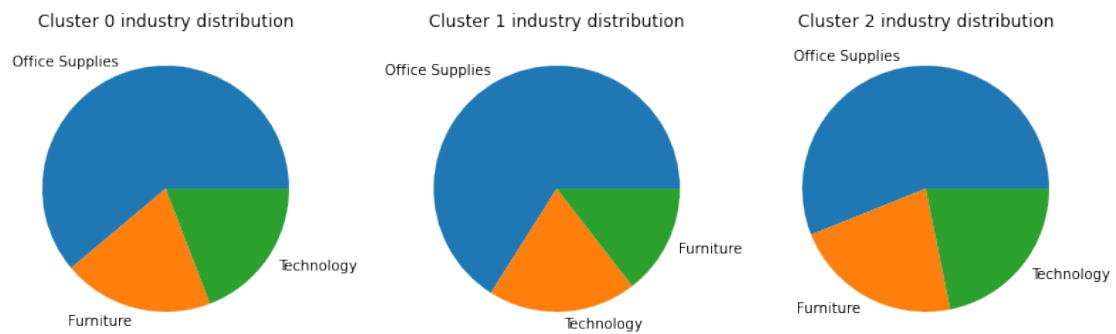


Figure 5.18: Equipment Distribution in Industry

In figure (5.18), it is observed the difference in the industrial distribution of cluster 0, cluster 1, and cluster 2. Where office supplies of cluster 1 are greater than the office supply of cluster 0 and cluster 2. The difference is very low among them. Furniture and technology are more consumed in cluster 2 compared with cluster 0 and cluster 1.

5.7.3 Order Placement in Different Market

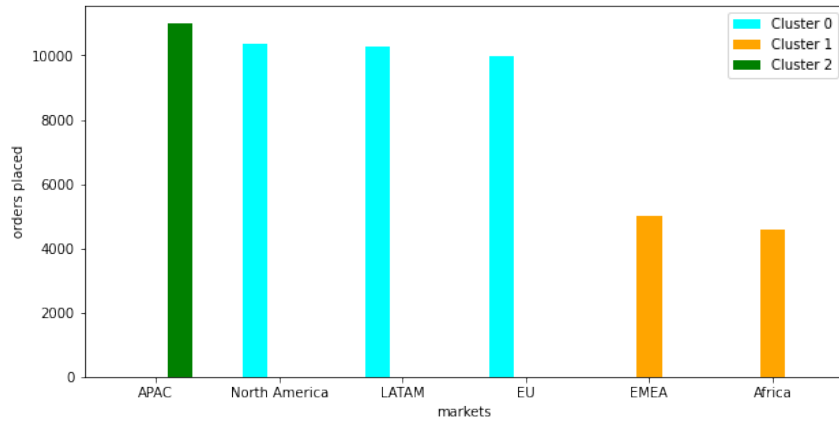


Figure 5.19: Order Placement in Market

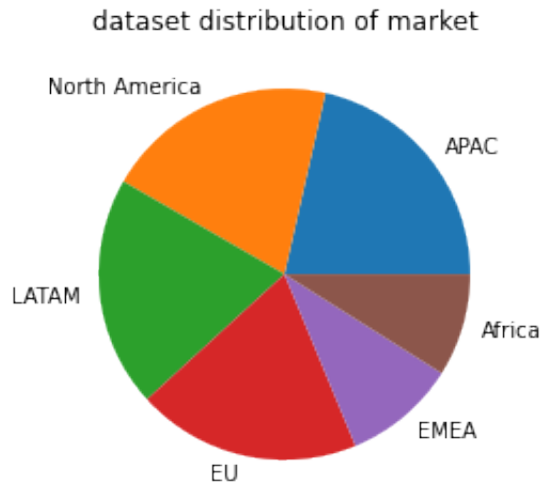


Figure 5.20: Dataset Distribution of Market

Here figure (5.19) shows that orders have been placed in which marketplace. Orders of cluster 2 are mainly placed in APAC, where orders of cluster 0 are placed in North America, LATAM, and EU, and orders of cluster 1 are placed in EMEA and Africa. From figure (5.20), it can be very clear about the dataset distribution of the market where we can see smaller to larger sequences. Sequentially they are APAC, North America, LATAM, EU, EMEA, Africa.

5.7.4 Profit and Sales Comparison

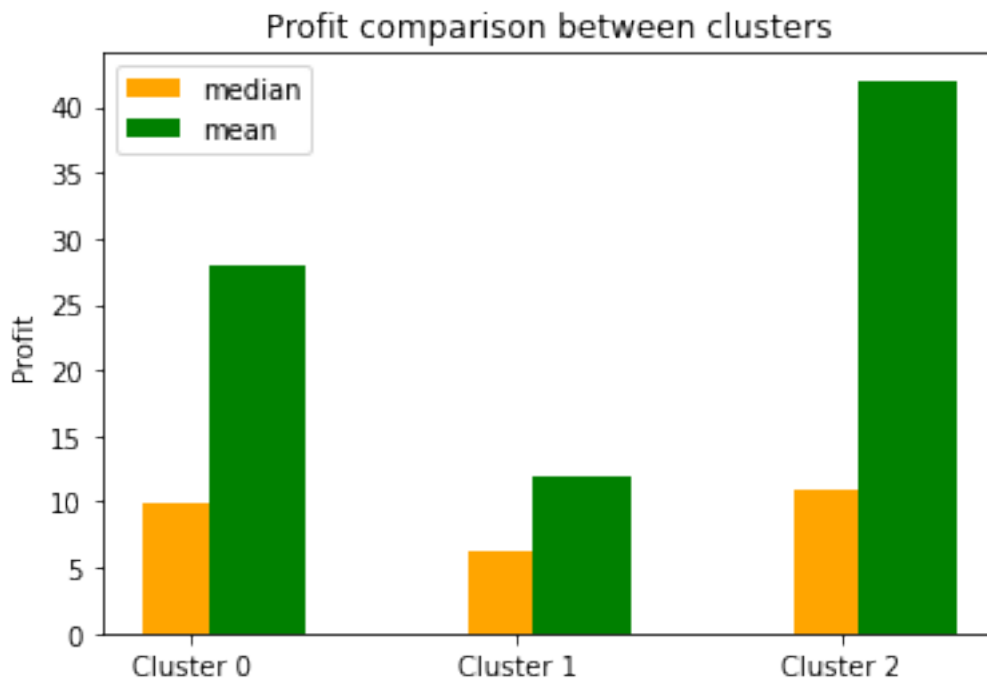


Figure 5.21: Profit Comparison Between Clusters

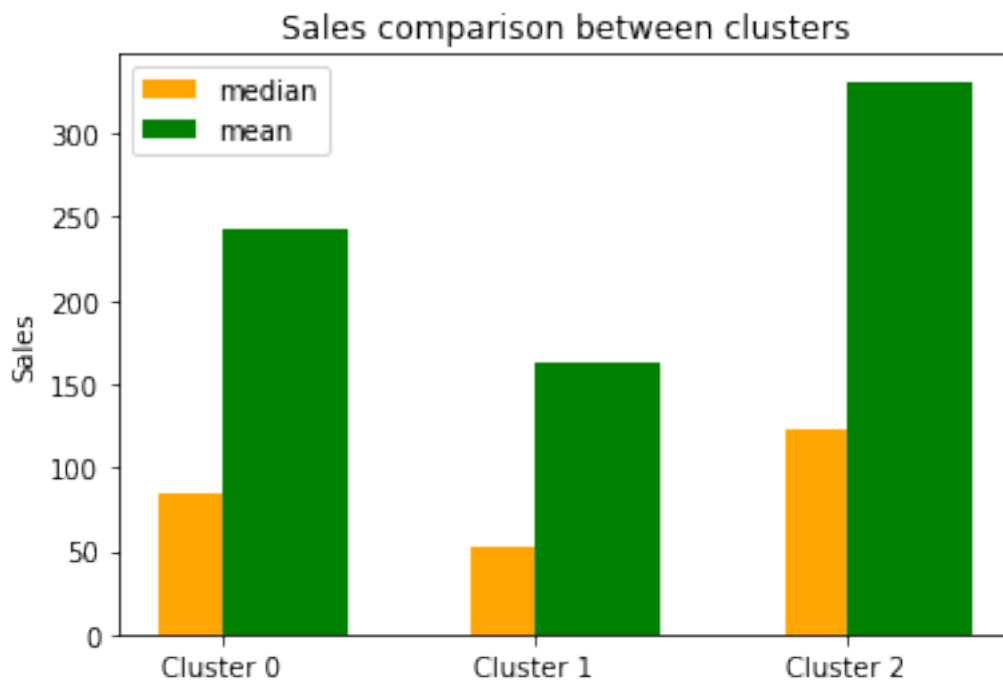


Figure 5.22: Sales Comparison Between Clusters

Profit comparison and sales comparison are shown in Figures 5.21 and 5.22. In both profit and sales figures, cluster two is higher than the other two clusters.

Chapter 6

Conclusion

6.1 Research overview

In this research, we have compared different modified models to see which gave the best results. The four models we have compared are as follows, PCA with K-means, PCA with K-means++, RFM with K-means++, and LRFM with K-means++. We found that PCA with K-means and PCA with k-means++ have the same results. So we choose the combination of PCA and K-means++ as our model as clusters have a good distance between them, and the data points in the clusters, unlike LRFM and RFM, are mostly packed. We attempted to demonstrate that it produces a more accurate outcome than RFM and LRFM research in terms of segmentation.

6.2 Contribution and Impact

Since the inception of RFM and LRFM models, numerous studies on customer segmentation utilizing these models have been conducted. However, only a handful of them can establish a correlation between customers and commodity quantity. The proposed analysis has a significant influence by proving a substantial association between earnings per head and the commodity purchased by each client in a single transaction. Based on the volume of the commodity, the research presents a novel way of segmenting clients into productive clusters.

Similarly, customer segmentation is a technique for enhancing the relationship with customers, learning about their needs and activities to develop organizations' concerns. To increase new customers and maximize profits, customer segmentation is crucial. Prospective customer data may be used to deliver programs based on client type, such as purchasing and selling internet advertising. In addition, the purpose of K-Means is to classify data points into distinct subpopulations that do not overlap. One of the conventional applications of K-means clustering is client segmentation in order to gain a better understanding of them, which can subsequently be used to increase the company's sales. In addition to K-Means, other techniques such as K Medoids and Mini Batch were employed to cross-check the K-Means clusters for this dataset. Here, the K-Means++ method is demonstrated and utilized for consumer

segmentation. After confirming through cluster performance testing we have concluded that the PCA with K-means++ analysis will assist company owners in more efficiently segmenting their clients, which would result in more efficient contact.

6.3 Future Work

PCA with the K-means++ method can be applied to databases with non-discrete details and a smaller number of data point variations. In addition, because k-means clustering is susceptible to outliers, it is advisable to eliminate them beforehand. Numerous businesses may use this model in the future to derive market characteristics from customer research matrices. Similarly, this method can be applied to datasets containing commodities with low variance. In addition to K-Means, other clustering algorithms such as K-medoids and Mini Batch can be used to save time and storage space, respectively. In addition, this model can be applied to the analysis of other facets of advertising, and its reliability can be quantified using specific matrices.

Bibliography

- [1] J. MacQueen *et al.*, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, 1967, pp. 281–297.
- [2] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.
- [3] W. J. Reinartz and V. Kumar, “On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing,” *Journal of marketing*, vol. 64, no. 4, pp. 17–35, 2000.
- [4] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, “Model-based clustering and data transformations for gene expression data,” *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.
- [5] H. Chang and S. Tsay, “Integrating of som and k-mean in data mining clustering: An empirical study of crm and profitability evaluation,” 2004.
- [6] C.-H. Cheng and Y.-S. Chen, “Classifying the segmentation of customer value via rfm model and rs theory,” *Expert systems with applications*, vol. 36, no. 3, pp. 4176–4184, 2009.
- [7] E. W. Ngai, L. Xiu, and D. C. Chau, “Application of data mining techniques in customer relationship management: A literature review and classification,” *Expert systems with applications*, vol. 36, no. 2, pp. 2592–2602, 2009.
- [8] M. J. Zimbardo *et al.*, *Difco & BBL Manual Manual of Microbiological Culture Media*, Mary Jo Zimbardo, BS, MT (ASCP) David A. Power, Ph. D. Sharon M. Miller, BS, MT (ASCP) George E. Wilson, MBA, BS, MT (ASCP) Julie A. Johnson, BA. Becton, Dickinson and Company, 2009.
- [9] L. Aristidis, V. Nikos, and J. Jacob, “The global k-means clustering algorithm ias technical report series,” IAS-UVA-01-02, Tech. Rep., 2011.
- [10] M. Christopher, “Logistics and supply chain management: Forth edition,” UK: Pearson Education Limited, 2011.
- [11] M. F. Eltibi and W. M. Ashour, “Initializing k-means clustering algorithm using statistical information,” *International Journal of Computer Applications*, vol. 29, no. 7, 2011.
- [12] S. Goyat, “The basis of market segmentation: A critical review of literature,” *European Journal of Business and Management*, vol. 3, no. 9, pp. 45–54, 2011.
- [13] D. S. Baer, “Csi: Customer segmentation intelligence for increasing profits,” in *Proceedings of the SAS Global Forum 2012 Conference*, Cary, NC: SAS Institute Inc, Citeseer, 2012.

- [14] R. Qiasi, M. Baqeri-Dehnavi, B. Minaei-Bidgoli, and G. Amooee, “Developing a model for measuring customer’s loyalty and value with rfm technique and clustering algorithms,” *The Journal of Mathematics and Computer Science*, vol. 4, no. 2, pp. 172–181, 2012.
- [15] A. Mesforoush and M. Tarokh, “Customer profitability segmentation for smes case study: Network equipment company,” *International Journal of Research in Industrial Engineering*, vol. 2, no. 1, pp. 30–44, 2013.
- [16] D. Usman and I. B. Mohamad, “A novel center point initialization technique for k-means clustering algorithm,” *Modern Applied Science*, vol. 7, no. 9, p. 10, 2013.
- [17] A. Çaliş, A. Boyaci, and K. Baynal, “Data mining application in banking sector with clustering and classification methods,” in *2015 International Conference on Industrial Engineering and Operations Management (IEOM)*, IEEE, 2015, pp. 1–8.
- [18] J. N. Sari, L. E. Nugroho, R. Ferdiana, and P. I. Santosa, “Review on customer segmentation technique on ecommerce,” *Advanced Science Letters*, vol. 22, no. 10, pp. 3018–3022, 2016.
- [19] M. Aryuni, E. D. Madyatmadja, and E. Miranda, “Customer segmentation in xyz bank using k-means and k-medoids clustering,” in *2018 International Conference on Information Management and Technology (ICIMTech)*, IEEE, 2018, pp. 412–416.
- [20] M. T. Ballestar, P. Grau-Carles, and J. Sainz, “Customer segmentation in e-commerce: Applications to the cashback business model,” *Journal of Business Research*, vol. 88, pp. 407–414, 2018.
- [21] J. T. Wei, S.-Y. Lin, Y.-Z. Yang, and H.-H. Wu, “The application of data mining and rfm model in market segmentation of a veterinary hospital,” *Journal of Statistics and Management Systems*, vol. 22, no. 6, pp. 1049–1065, 2019.
- [22] M. Alkhayrat, M. Aljnidi, and K. Aljoumaa, “A comparative dimensionality reduction study in telecom customer segmentation using deep learning and pca,” *Journal of Big Data*, vol. 7, no. 1, pp. 1–23, 2020.
- [23] F. Yoseph, N. H. Ahamed Hassain Malim, M. Heikkilä, A. Brezulianu, O. Geman, and N. A. Paskhal Rostam, “The impact of big data market segmentation using data mining and clustering techniques,” *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 6159–6173, 2020.
- [24] A. Abdulhafedh, “Incorporating k-means, hierarchical clustering and pca in customer segmentation,” *Journal of City and Development*, vol. 3, no. 1, pp. 12–30, 2021.
- [25] E. Akar, “Customers’ online purchase intentions and customer segmentation during the period of covid-19 pandemic,” *Journal of Internet Commerce*, vol. 20, no. 3, pp. 371–401, 2021.
- [26] S. Bandyopadhyay, S. Thakur, and J. Mandal, “Product recommendation for e-commerce business by applying principal component analysis (pca) and k-means clustering: Benefit for the society,” *Innovations in Systems and Software Engineering*, vol. 17, no. 1, pp. 45–52, 2021.

- [27] J. Jääskeläinen, “Segmentation of investor customers using machine learning in banking,” 2021.
- [28] C. Jamunadevi, S. T. Selvan, M. Govindarajan, C. Saravanan, and B. J. Raman, “Lrfm model for customer purchase behaviour using k-means algorithm,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 1055, 2021, p. 012 111.
- [29] F. Khanizadeh, F. Khamesian, and A. Bahiraie, “Customer segmentation for life insurance in iran using k-means clustering,” *International Journal of Non-linear Analysis and Applications*, vol. 12, no. Special Issue, pp. 633–642, 2021.
- [30] R. N. N. Naseri, F. Rahmiati, and M. M. Esa, “Consumer attitude and online purchase intention: A segmentation analysis in malaysian halal cosmetic industry,” 2021.
- [31] E. H. Sharaf Addin, N. Admodisastro, S. N. S. Mohd Ashri, A. Kamaruddin, and Y. C. Chong, “Customer mobile behavioral segmentation and analysis in telecom using machine learning,” *Applied Artificial Intelligence*, pp. 1–21, 2022.