LRFMV: An Efficient Customer Segmentation Model for Superstores

by

Md.Toyeb
17101399
Rezwana Mahfuza
17301016
Nafisa Islam
17101448
Md Asaduzzaman Faisal Emon
17301188

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

Department of Computer Science and Engineering
BRAC University
June 2021

© 2021. 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:

Md. Toyeb

Md. Toyeb
17101399

REZWANA MAHFUZA

Rezwana Mahfuza
17301016

Nafisa Islam

Nafisa Islam
17101448

Md Asaduzzaman Faisal Emon
17301188

Approval

The thesis titled "LRFMV: An Efficient Customer Segmentation Model for Superstores" submitted by

1. Md. Toyeb (17101399)
2. Rezwana Mahfuza (17301016)
3. Nafisa Islam (17101448)
4. Md Asaduzzaman Faisal Emon (17301188)

Of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 2, 2021.

Examining Committee:

Supervisor:
(Member)

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

Head of Department:
(Chair)

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

Ethics Statement

It is therefore declared that all comparisons made and conclusions formed in this thesis are based on the results of the group's own study on the topic. All essential materials, as well as the dataset provided, have been properly cited. Appropriate precautions have been made to assure the analysis's transparency. This thesis has never been submitted to another institute in any format.

Abstract

In superstore business, the recency, frequency, and monetary (RFM) based on customers' purchase results is preferred to categorize valuable customers in order to increase profit margins. This paper develops an enhanced RFM (recency, frequency, monetary) and LRFM (length, recency, frequency, monetary) model, namely LRFMV (length, recency, frequency, monetary, and volume), and then clusters the data using the standard K-means, K-medoids and Mini Batch K-means algorithms. The results obtained from the three algorithms are compared and the K-means algorithm is chosen for the superstore dataset of the proposed LRFMV model. All clusters created using these three algorithms are evaluated in the LRFMV model and a close relationship between profit and volume is observed. A clear profit-quantity relationship of items has yet not been seen in any prior study on the RFM and LRFM models. Grouping customers aiming at the profit maximization existed previously but there were no clear and direct depiction of profit and quantity of sold items. To establish a relationship between volume and profit, this study applied unsupervised machine learning to investigate the patterns, trends, and correlations between these two variables. The traits of all the clusters are analyzed by the Customer-Classification Matrix. The values of LRFMV variables that are larger or less than the overall average for each cluster are identified and utilised as their traits. The RFM model, the LRFM model and the suggested LRFMV model are compared, and the outcome indicates that the LRFMV model may create more segments with the same number of customers while maintaining a greater profit per head.

Keywords: Customer segmentation; RFM analysis; LRFMV analysis; K-means; K-Medoids; Mini Batch K-means; volume; silhouette; elbow; traits

Dedication

This thesis is dedicated to our loving parents and our department's respectable faculties, who have encouraged and supported us during the entire thesis and motivate us to achieve excellence in every aspects.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor Dr. Md. Golam Rabiul Alam sir for his kind support and advice in our work. He helped us whenever we needed help.
And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Declaration **i**

Approval **ii**

Ethics Statement **iii**

Abstract **iv**

Dedication **v**

Acknowledgment **vi**

Table of Contents **vii**

List of Figures **xi**

List of Tables **xiii**

Nomenclature **xiii**

1 Introduction **1**

1.1 Motivation 2

1.2 Problem Statement 3

1.3 Research Objective 4

1.4 Research Methodology 4

1.4.1 RFM Model 5

1.4.2 LRFM Model 6

1.4.3 LRFMV Model 6

1.5 Scope and Limitation 6

1.6 Thesis Outline 7

2 Literature Review 8

3 Proposed Customer Segmentation Model (LRFMV) 12

3.1 Proposed System Model 12

3.2 Data Collection 14

3.3 Data Preprocessing 16

3.3.1 Data inspection before preprocessing 16

3.3.2 Data Cleaning 19

3.3.3 Data Integration 20

3.3.4 Data dimensionality reduction and Feature selection 21

3.3.5 Data Transformation 21

3.4 Feature Extraction 22

3.4.1 Calculation of L 22

3.4.2 Calculation of R 23

3.4.3 Calculation of F 24

3.4.4 Calculation of M 25

3.4.5 Calculation of V 26

3.5 Correlation of volume with other features 27

3.6 Model Specification 28

3.6.1 Elbow Method 28

3.6.2 Silhouette Coefficients 31

3.6.3 Cumulative Explained Variance Ratio 32

3.6.4 PCA 34

3.6.5 K-means Algorithm 38

3.6.6 K-Medoids Algorithm 43

3.6.7 Mini Batch K-means Algorithm 45

4 Comparative Analysis 48

4.1 Number of clusters 48

4.1.1 Number of cluster determination for KMeans algorithm 48

4.1.2 Number of cluster determination for K-Medoids algorithm . . 51

4.1.3 Number of cluster determination for Mini Batch K-means Algorithm 52

4.2 Clustering Quality 54

4.2.1 Profit analysis for K-Means Algorithm 54

4.2.2 Profit analysis for Mini Batch K-Means Algorithm 56

4.2.3 Profit analysis using K-Medoids algorithm 58

5 Result Analysis 60

5.1 Statistical Analysis 60

5.1.1 Volume-Profit Relationship for K-Medoids Algorithm 60

5.1.2 Volume-Profit correlation for Mini Batch K-Means Algorithm 61

5.1.3 Volume-Profit correlation for K-Means Algorithm 62

5.1.4 Profit Analysis of each Cluster 63

5.1.5 Customer Number Analysis for each cluster 63

5.2 Cluster Analysis with Customer-Classification Matrix 64

5.2.1 Revenue Generation for each cluster and their cost to serve . . 64

5.2.2 Identification of Traits 67

6 Conclusion 69

6.1 Research Overview 69

6.2 Contribution and Impact 69

6.3 Recommendation & Future Work 70

Bibliography **75**

List of Figures

3.1 Work Process of the proposed LRFMV model 13

3.2 Raw Dataset 14

3.3 Raw Dataset 15

3.4 Dataset with selected columns 16

3.5 Profit made from different sub-categories of products 17

3.6 Sales per category 18

3.7 Barchart of sales by category based on order-dates 19

3.8 Heatmap for LRFMV model 28

3.9 Elbow Method 29

3.10 Elbow Method for the LRFMV model 30

3.11 A visual representation of objects involved in the calculation of S(i) . 31

3.12 Silhouette Score 32

3.13 Cumulative Explained Variance for random 200 points 33

3.14 Cumulative Explained Variance for our proposed dataset 34

3.15 3D featured dataset 35

3.16 Dimensionality reduction to 2D using PCA 36

3.17 3D representation of PCA on the dataset of the proposed model . . . 38

3.18 Unlabeled data before applying the K-means algorithm 39

3.19 Labeled data after applying the K-means algorithm 40

3.20 Graphical illustration of the K-means clustering process 41

3.21 K-means clusters for LRFMV model 42
3.22 3D representation of K-means clusters for LRFMV model 43
3.23 Graphical illustration of K-Medoids algorithm 44
3.24 K-Medoids for LRFMV model 44
3.25 3D representation of K-Medoids clusters for LRFMV model 45
3.26 Implementation of Mini Batch K-means algorithm 46
3.27 Applying Mini Batch K-means Algorithm on LRFMV model 46
3.28 3D representation of K-means clusters for LRFMV model 47

4.1 Applying K-means Algorithm on RFM model 49
4.2 Applying K-means Algorithm on LRFM model 50
4.3 Applying K-Medoids algorithm on RFM model 51
4.4 Applying K-Medoids algorithm on LRFM model 52
4.5 Applying Mini Batch K-means algorithm on RFM model 53
4.6 Applying Mini Batch K-means algorithm on LRFM model 54
4.7 Profit analysis for RFM, LRFM and LRFMV model using standard K-means algorithm 55
4.8 Profit analysis for RFM, LRFM and LRFMV model using Mini Batch K-means algorithm 57
4.9 Profit analysis for RFM, LRFM and LRFMV model using K-Medoids algorithm 58

5.1 Volume-Profit Relationship of LRFMV model for K-Medoids Algorithm 60
5.2 Volume-Profit Relationship of LRFMV model for Mini Batch K-Means Algorithm 61
5.3 Volume-Profit Relationship of LRFMV model for K-Means Algorithm 62
5.4 Profit of each cluster in LRFMV model 63
5.5 Number of customers in each cluster 64

List of Tables

3.1 Calculated Length (L) 22

3.2 Calculated Recency (R) 24

3.3 Calculated Frequency (F) 25

3.4 Calculated Monetary (M) 26

3.5 Calculated Volume (V) 27

3.6 PCA on the dataset of the proposed model 37

5.1 Revenue Generation for each cluster and their cost to serve 65

5.2 Comparison between Average of L, R, F, M, V of the whole dataset with each cluster 68

Chapter 1

Introduction

Customer segmentation is the process of classifying or segmenting a customer base according to common factors such as age, gender, interests, and buying habits. This enables marketing departments to engage particular smaller markets with relevant messages that increase the likelihood of consumers purchasing something. Customer segmentation is based on a number of considerations. This involves demographic information about consumers such as ethnicity, religion, wages, and schooling, as well as information about their place, lifestyle, and buying habits. Marketing automation software enables the collection of necessary data, the definition of these customer segments, and the execution of marketing strategies. Businesses may use customer segmentation models to get a deeper understanding of their consumers' tastes and desires, as well as to customize marketing materials to be more customized. Customer segmentation essentially enables companies to optimize sales and deploy marketing capital efficiently. Additionally, it will enhance customer support and increase customer satisfaction and segmentation by RFM analysis can ensure the right customer support for the right customer. RFM segmentation enables marketers to communicate with individual groups of consumers in ways that are far more important to their unique actions – and therefore produce significantly better engagement rates and improved loyalty. As with other segmentation techniques, RFM segmentation is an effective technique for identifying classes of consumers that need special care but it lacks the ability to determine customer loyalty. In this case, LRFM model was developed as an extension of RFM model. Here, ”L” denotes length which indicates the length of a customer's first and last transaction within a certain time period. This must be measured in order to increase the value of recency produced by customers based on the length of the most recent transaction with a certain time period.

In the research, an in-depth comparison of RFM analysis, LRFM analysis and LRFMV analysis is shown. RFM analysis is a form of measured mathematical analysis that enables a company to group all of its users according to their behavior. It is an acronym for Recency, Frequency, and Monetary. Numerous experiments have shown that by adding a component to the initial RFM Model, the predictability of customer behaviour may be increased. By simply increasing the duration of the customer connection (L) and the volume of the customer (V) in the RFM Model, the efficiency and accuracy of customer segments are increased. By using the conventional RFM Model, business entities are unable to differentiate between long-term and short-term clients, which the LRFMV Model enables. After determining RFM, LRFM and LRFMV values respectively, K-means algorithm, K-medoids algorithm and mini batch K- means have been applied. After evaluating customer values using the LRFMV model, customers have been profiled using the k-means clustering algorithm.

1.1 Motivation

The term ”superstore” refers to a store that is self-service and physically organised with multiple divisions to ensure a diverse variety of food and grocery goods[54]. This retail type is often smaller than a hypermarket but larger than a standard grocery store.

A superstore’s customers are generally local individuals and small enterprises that need home goods replenishment on a recurrent basis. The suppliers to a superstore are typically manufacturers of household items located in regions far from the ultimate customers. In fact, the superstore serves as a shopping environment, bridging the divide between distant suppliers and local buyers. The superstore’s ”product” under this setup is its supply chain[30] and the main essence of the business model is to get as many individuals from all socioeconomic backgrounds through the door as possible and to keep them.

Walmart, one of the world’s biggest companies, operates about 4,700 shops in the United States, including almost 600 Sam’s Club locations[12]. It is the biggest private employer in the United States. It has the ability to offer its product at a discount to competitors in the marketplaces in which it operates. This may have an influence on markets other than retail, such as manufacturing and production.

After the industry started in the early 2000s in Bangladesh[54], superstore chains have risen in prominence with the general public. According to the BSOA (Business superstores Owners’ Association), over 121 superstores have opened in Bangladesh over the last 16 years[36]. Many of these retailers often offer online shopping. Given population size and rapid urbanization of Bangladesh, observers and retail owners alike agree that Bangladesh’s superstore network is insufficient. superstore chains have evolved in stages around the globe, not necessarily in first-world economies, as market consumption patterns changed and real incomes increased. To extend these shops, it is essential to monitor customers’ shopping habits. Recently, superstores such as Meena Bazar, Agora, and Shopno have become more common in Bangladesh for regular household supplies than department stores, owing to the wide variety of products available. It can be considered as a thriving industry in Bangladesh, and identifying their target audiences has become a necessary phase in developing marketing campaigns for various customer segments.

According to the facts shown above, it is critical for service businesses to retain customers over time since this is the only way to reduce expenses and increase customer satisfaction via a greater knowledge of their requirements and expectations.

Though customer profiling has been found in other technical contexts, new functionality of the RFM model has been added to improve clustering, and some studies have attempted to profile superstore customers.

The research aims to bridge this divide by creating a customer profiling system focused on real-world retail customers. It suggests, in particular, a combined solution focused on clustering for customer profiling in superstore operations. This feature evaluates customer values using the LRFMV model, followed by customer profiling using the k-means clustering algorithm which is an unsupervised learning algorithm. Unsupervised learning is a process in which the computer takes inputs $x_1, x_2, ...$ but does not gain supervised target outputs or incentives from its environment[8].

Then, in order to elicit the association between a store's expense and revenue, the purchasing habits of shoppers and the quantity of products bought are gathered.

1.2 Problem Statement

Although mass marketing strategies may be successful, believing that everyone is interested in buying what you're selling is a time-consuming, inefficient, and expensive strategy. Rather than using a one-size-fits-all approach, successful segmentation categorizes customer data according to similar resources or behavioral characteristics, allowing complex advertising and personalization campaigns for more timely, substantive, and reliable marketing communications.

According to Desouza et al. (2008)[15], organizations may use segmentation to distinguish and categorize consumers depending on those characteristics, allowing them to define target audiences. These characteristics, when handled properly, can enhance both the service and the product.

Numerous segmentation components are available for segmenting customers[58]. Customer segments may be described using demographic data (age, wages, business, etc.) as well as transaction records (sales and purchasing data). The RFM-analysis is mostly concerned with the latter. The descriptive method is characterized by its simplicity of execution, straightforward treatment, and friendly versatility (Jasmin, 2020)[58].

In the 1960s, catalog distribution firms used RFM analysis to conduct targeted marketing (Blattberg et al., 2009)[18]. RFM is a classification method that includes three factors into consideration: the recency level and monetary value of each consumer. According to Wu et al. (2014)[37], The following meanings apply: recency is the time interval after the most recent purchase; frequency is the number of sales made within the same time span, and cumulative is the overall amount of money invested on all transactions made within the same time period.

One of the shortcomings of the standard RFM paradigm is its inability to distinguish between long-term and short-term buyers. Additionally, it lacks the ability to intelligently associate customers with products. As discussed previously, the LRFM model is capable of quickly identifying a consumer's loyalty, but there is always a disconnect between the product and the customer.

Thus, the study is attempting to address the following questions:

- ***How much influence does the LRFMV model have on superstores?***
- ***How does it outperform the traditional RFM & LRFM model?***
- ***How can LRFMV model be extemporized and compared with the RFM and LRFM model ?***

1.3 Research Objective

The key goals of the proposed model are:

- To present the correlation between the volume of products and the profit earned against each customer.

- Calculate the LRFMV value and perform customer segmentation with the appropriate number of clusters by the standard K-means, K-Medoids and Mini Batch K-means algorithms

- Comparing the findings from the standard K-means, K-Medoids and Mini Batch K-means algorithms

- Comparing the result of LRFMV analysis with the traditional RFM analysis and LRFM analysis

1.4 Research Methodology

This thesis is aimed to propose an extension model of traditional RFM analysis on a given dataset. Just like any other research, Data collection is needed to create an ideal dataset. Since collecting data from primary sources can be time-consuming and costly,the dataset used is created by others to conduct the study. The data used in this dataset are quantitative which is used to calculate the metrics of Length, Recency, Frequency, Monetary value, and Volume. The variables used to calculate these above-mentioned metrics consist mostly of sales data and the purchase history of customers. No personal information of customers is used in this research . The dataset used is a client/customer-driven informational dataset that contains the relative details of orders placed with different vendors and markets between 2011 and 2018. 24 attributes were included in the dataset, including Ship Date, City, Category, and Order Priority.

This massive dataset was generated by Tableau Software, Inc. They are one of the biggest data-related organizations in the world. They maintain a current version of this dataset.

1.4.1 RFM Model

According to Segal (2019)[56], Recency, frequency, and monetary value is a marketing measurement technique utilized to determine a business's or organization's best clients by the application of specific metrics. Three quantitative considerations serve as the foundation for the RFM model.

- Recency: How recently a consumer purchased something.

- Frequency: The frequency at which a consumer makes a transaction

- Monetary value: The amount of money spent by a consumer on sales.

The term recency, frequency, and monetary value (RFM) is thought to have originated with an article titled "Optimal Selection for Direct Mail" in a 1995 edition of Marketing Science by Jan Roelf Bult and Tom Wansbeek.[56].RFM analysis often confirms the advertising cliché that "80% of revenue comes from 20% of customers."[56]." RFM research significantly aids the business owners in determining what they need to do about the customer.

RFM values are determined by integrating their Recency, Frequency, and Monetary values. For instance, customer 111 recently put a low-value order. Customer 333, on the other hand, always places large-value orders and recently made a purchase. Customers that earn three stars in each ranking category are the most valuable.

The RFM model is based on user-business transactions in order to establish a solid data-driven approach that is grounded in real-world facts. This client data is graded, reviewed in more detail, and then segmented in order to target certain client segments. This strategy enables businesses to analyse each customer's historical purchasing behaviour in order to anticipate and affect future consumer interactions.

Segmentation techniques used by market research firms prior to the invention of data analytics segment consumers based on demographic and psychographic factors. Researchers always predict population behaviour using sample audiences, which limits market researchers' ability to forecast user behaviour for niche consumer groups and specific customers.

1.4.2 LRFM Model

According to Wu et al (2014)[37], The LRFM model was built on the basis of the RFM model, a well-known technique for analysing customer values for market research. Reinartz and Kumar (2000)[7] discussed the RFM model's inability to differentiate between long- and short-term partnerships with consumers. Consumer satisfaction is determined by the partnership between a business and its consumers and is built over time by effective customer relationship management.Additionally, Chow and Holden (1997)[6] imply that customer loyalty and profitability are contingent upon the connection between a business and its consumers. In this respect, it is vital to examine the customer's relationship duration in order to identify the most loyal consumers. Customer loyalty is determined by the connection between a business and the customers and is built over time via effective customer relationship management.

As a result, Chang and Tsay (2004)[9] expanded the RFM model to the LRFM model by factoring in duration (L), which is defined as the time span (in days) between the database's first and latest purchase.

1.4.3 LRFMV Model

The study included volume (V) which denotes the quantity of products in a single transaction completed by a customer to segment the consumer base more effectively and to relate between customer and products. Volume may influence the profit generated from certain customer base in a positive way. The values of LRFMV variables that are larger or less than the general average for each cluster are discovered and utilised as their features.

1.5 Scope and Limitation

This thesis is set to provide an evaluation of RFM and the extended version of RFM model, LRFMV model by segmentation of customers using K-means, K-medoids and mini batch K-means algorithms. Since this evaluation is based on numerical values, it is not applicable to qualitative data of textual values. Moreover, as it only deals with the sales data, it is unable to analyze the customer behavior based on personal data like age, demography, gender, race, etc.

Moreover, If there is less variance of quantities of product, then the change of values of LRFM and LRFMV may be insignificant. The dataset used in this research does not distinguish between the units of the quantities (for example, litre, kg etc.) . Additionally, as the number of dimensions rises, a distance-based similarity measure tends to converge to a fixed value for each given pair of samples. So to reduce dimensionality PCA is applied.

1.6 Thesis Outline

The rest of the thesis is designed as follows:

Chapter 2 contains a Literature review and an overview of the related works

Chapter 3 contains a detailed discussion about the proposed model. It is consist of data collection, data preprocessing, proposed workflow, feature extraction, and model specification

Chapter 4 describes the comparison amongst RFM, LRFM and LRFMV model after applying k-means, k-medoids and mini batch k-means algorithms.

Chapter 5 contains a discussion about analyzing the results. It describes statistical result analysis of each clustering algorithm applied in this study.

Chapter 6 is the chapter of the conclusion. It contains a research overview, contribution, impacts, and future works.

Chapter 2

Literature Review

Over the years, many approaches for data analysis have been developed widely applied in different fields. As the volume of transactions in the business grows, it has become more difficult to segment profitable customers in order to enhance sales. In the customer segmentation process, the RFM model can be applied on the purchase experience of a customer, the development of improved prediction and classification techniques [42], [22]. To optimize marketing results, both customer segmentation and customer targeting are needed [21]. The transactional data is first subjected to an RFM analysis,[49], [34] after which clustering techniques like standard K-means,[51], [4], [23] Fuzzy C-means [38] ,[46], and Repetitive Median based K-Means (RM KMeans) algorithms for clustering [40] are used. Following that, the clusters are further evaluated in order to segment customers appropriately.

For better forecasts and identifications, numerous researchers used RFM analysis. For instance, customer satisfaction [20], customer lifetime value [11], churn prediction [59], [42], CLV measurement [13] forecast a customer's reaction to direct marketing [21] can be analyzed using data mining models and customers can be classified according to their profitability. Some authors used the RFM model to create a two-phased model mechanism that can be thought of as a new segmented solution [44]. In addition to superstore research, RFM is also used in a variety of sectors, including banking and insurance [10],[13], telecommunications [16], political score generation [55], on-line industries [25], travel agencies [17], retail industry [43], medical field [42], [31], and so on. The study conducted by Tavakoli.M et al [53] includes the importance of behaving with customers according to their background and category which has evolved significantly in recent years. The optimal cluster numbers is first ascertained using the self-organizing maps system (SOM) by Daoud.A et al [39]. After ascertaining the best cluster numbers, the K-means algorithm is utilized for clustering customer data after performing RFM analysis for each customer of an online selling company in Morocco. Customer segmentation has been used as the basis for customer understanding and classification, and as segmenting customers is the RFM model, the most successful approach has been considered.

Cheng and Chen mined classification rules by combining the quantitative value of

RFM characteristics and the K-means method in the rough set theory RS theory [20]. The proposed model helps companies to develop CRM and also minimize the drawbacks of different data learning tools for establishing CRM. Based on subjective judgement, the obtained output is classified into three, five, and seven classifications with the help of K-means clustering and determines the class which has the highest consistency rating. The authors used the rough set LEM2 framework to generate decision rules. However, an improved RFM model offers satisfying accuracy and profit margin in most situations.

Using weighted frequent pattern mining, Cho and Moon [33] proposed a personalized recommendation scheme in which the RFM model is used to categorize prospective clients. A good customer profiling was generated by applying the RFM model from the client's details. As a consequence, they were able to provide the clients with an effective recommendation. The firm's net margin was also boosted as a result of its ability to track valuable clients. However, instead of using standard RFM models, they could get a decent result by factoring in any other variables.

To improve data mining approaches for customer analysis, Bachtiar.A suggested a two-step mining approach based on the RFM model [48]. The obtained data is first analyzed using the RFM technique and later on segmentation using K-means clustering. Customer characteristics are described by IF-THEN rules after the clusters are further examined using association rule learning. Silhouette coefficients and Connectivity measures are utilized for assessing the cluster results and evaluate accordingly. Because of using the two steps approach for customer analysis, accurate insight on customer behavior as well as purchasing tendency are gained which can actually help to improve marketing strategy significantly. In the association study, only frequent mining patterns have been used, as infrequent patterns are not obtained from the clusters. As a consequence, useful customer research data could be lacking at this stage. Furthermore, the use of modern RFM models could strengthen the research process by analyzing profitable customers more precisely as some additional variables are taken into account.

Many researchers started to develop modern RFM models in order to see how they outperform existing RFM models by introducing new variables. For instance, Yeh et al. proposed the RFMTC model where T indicates Time after first purchase and C demonstrates Churn probability and it can quantify the likelihood of the buyer repeating the acquisition as well as the estimated gain of the cumulative amount of possible sales [14]. Wei.J et al [31] concentrated on the LRFM, which is an expanded RFM that contains an additional parameter named Length. It refers to the summation of days between a client's initial and final appointment at the clinic. The average LRF values for each cluster and all patients are estimated, and the values that are higher than the average are analyzed, indicating core patients. They also formulated their marketing techniques and created twelve clusters for a total of 2258 dental patients using the SOM (self-organizing maps) methodology. The result is definitely better than the traditional RFM model but there are also some limitations. It is clearly observable that the above mentioned papers have not stated about the interconnection between profit and their proposed model as increasing profit is the main goal in most of the business.

Another great research has been conducted by Christy, A.J. et al [49] where they designed a new approach using the RFM model on the transactional data. They compared and contrasted the findings of RFM with the regular K-means and Fuzzy C-means data mining algorithms. In the mentioned paper, a new approach for selecting centroids at the start for the K-means algorithm is suggested. The proposed modified algorithm is named as Repetitive Median centered K-means algorithm or RM K-means. Here for comparative analysis, they considered execution of time, iterations, and cluster compactness for both of the algorithms. The authors observed that RM K-means performs better than the other two algorithms because it takes less time to execute and has fewer iterations. Traditional RFM models, on the other hand, will make it difficult to identify prospective customers. In certain cases, combining parameters with conventional RFM yields the best results and provides the best solution.

In specialized clinics, Mohammadzadeh.M et al [42] discovered opportunities for repeat clients in order to maximize profit and minimize patient failure costs. They used the RFML model to forecast new client turnover and conduct behavioral research on specific existing patients. Following this, the authors applied the K-means algorithm to group clients and compare different groups of three clinics. A decision tree classifier was also used as a churn predictor based on which the number of faithful and turnover patients were identified. Conversely, due to demographic and regional factors, the suggested approach may not be effective in all cases.

Sarvari et al.[41] used a variety of data mining techniques, as well as the K-means algorithm, neural networks and Apriori association rule mining to develop WRFM (Weighted RFM), a revised RFM process [41]. They analyzed the customer data in Turkey of a multinational pizza restaurant network by considering demographic data with RFM variables. The cluster is analyzed using both an unweighted conventional RFM value and a weighted WRFM value, demonstrating that adding demographic variables into account yields an excellent outcome with positive associations. Both the Kohonen algorithm (SOM) and the K-means algorithm have been studied and compared for customer segmentation. Between the two clustering algorithms used by the author, K-means produced more consistent results in terms of cluster consistency and runtime. The proposed model shows how WRFM combined with clustering strategies outperforms conventional RFM and strengthens marketing strategy, resulting in increased profits for the business. More demographics characteristics should be considered and compared to standard RFM to see how they work. Aside from that, other variables may be considered in order to observe customers' purchasing patterns more systematically.

The RFM model was extended by Soeini.AR et al.[60] by adding two more variables: duration and cost. Here, L denotes the period after the first transaction or the duration of the customer relationship, and C denotes the customer's expense. In the paper, SPSS software and the K-means algorithm is utilized for clustering the clients of an insurance company. The authors contrasted the two models using association rules and came to the conclusion that LRFMC outperforms the standard RFM model. The fundamental disadvantage of the suggested model is it is only applicable to a single group, whereas a model should be more universal. A methodology that is suited for all marketing categories should be used to find profitable customers.

Tanaka.T et al.[45] proposed a solution for long-term revenue declines for Japan's superstore stores which is a result of demographic shifts. They introduced a new model for identifying loyal customers by extending the traditional RFM method. Along with the RFM model they combined IF and ISF scores and observed that it gives better performance in respect to the classification accuracy by accuracy, precision, recall, and F value indicators. Good customers are also evaluated successfully by Criterion Variables and Coefficients in their proposed model. By this, a better understanding of features that contribute to the retention of excellent clients in Japan's superstores was accomplished. Rather than using the RFM model, they have developed a new expanded model that outperforms the RFM model.

A composite technique integrating clustering and association rule mining is proposed by Guney.S et al, 2020 [57]. For the segmentation process, the authors used modified RFM that is LRFMP alongside apriori algorithms and K-means algorithm for the association rule mining. With the traditional RFM model, L (Length) and P (Periodicity) are introduced by which customer value can be measured successfully. The aim of the suggested LRFMP model is to assist defining the qualities and consistency of customers with various buying patterns in the grocery and retail sectors [43]. The authors [57] used this model to obtain the values of subscribers with the help of K-means algorithm and determined suitable subscribers on the basis of an actual customer data from IPTV suppliers of services. Customers that are subscribers are divided into groups consists of four after the report determines the most appropriate cluster number with the help of the K-means algorithm. Although using modified RFM in their research purposes identified potential customers, VoD transaction records are obtained by the STB devices only. As a result, transactional data using other devices are not considered here. Besides, the approach is limited to Turkish customers and so it can not be said that the approach will be applicable in all geographic areas.

An innovatory technique in order to obtain all RFM sequential patterns from customer purchase information of a Taiwanese retail chain has been suggested by Chen. Y et al. [19]. With the use of the RFM pattern for Sequential Pattern Mining (SPM), the authors proposed a segmentation approach for analyzing client buying behavior. RFM-Apriori, R^AFM-Apriori and GSP has been executed here with the Java language. The difference between R^AFM-Apriori and traditional RFM-Apriori is in the measurement for recency structure by averaging the recency scores of the data items within the series. The time of execution and total patterns for RFM-Apriori and R^AFM Apriori is analyzed and compared. The GSP and RFM-Apriori are compared considering the execution of time, recency, monetary value and total amount of patterns for both of them. The traditional Apriori algorithm (GSP) [5] is modified and a new algorithm is proposed named RFM-Apriori for generating RFM-patterns by the authors [19]. The dataset collected here is from an offline retailer where there are major differences between online and offline retailing data with respect to the scale of visiting frequency, product purchase rate, customer visits, etc. That's why the proposed model may not be suitable for online retail shops where the model should be more generalized and flexible.

Chapter 3

Proposed Customer Segmentation Model (LRFMV)

3.1 Proposed System Model

Finding a proper and appropriate data set for the intended outcome was the first step in this project. Tableau Software has published the dataset online which has been used. This firm specializes in datasets and databases. The dataset was not preprocessed because it had a large number of null and arbitrary values that could stymie the research. Then, through a series of stages, this dataset has been preprocessed.

Preprocessing consists of four steps which are known as Data cleaning,data integration,data dimensionality reduction and feature selection and data transformation though the proposed dataset did not need any data integration. After preprocessing ,feature extraction had been done of that dataset by calculat- ing the desired L, R, F, M, V components for establishing the LRFMV model. Data frame joining process has been materialized in order to combine L,R,F,M,V together. The optimal and efficient number of clusters has been found for the K- means algorithm by using the Elbow and Silhouette method. The feature reduction step was for reducing the 6 cluster representation to 2 dimensions for complexity. Cumulative Explained Variance and PCA have been used for feature reduction and representing the dataset in two dimensions. The clusters of customers have been generated by applying the K-means algorithm and evaluated by comparing them with the RFM and LRFM model. Evaluation has taken place with the compari- son of LRFMV,LRFM and RFM models. 6 efficient clusters for the dataset for K-Medoids, Mini Batch K-means and K-Means have been demonstrated and afterward, those clusters have been apprised by profit-volume correlation, calculating profit for different clusters, calculating the number of customers in each cluster, etc. Customer segmentation is being assessed by using a customer classification matrix and labeled the customers as aggressive, passive, bargain-basement and carriage trade. For failing into cluster evaluation it was started from the data preprocessing step again.

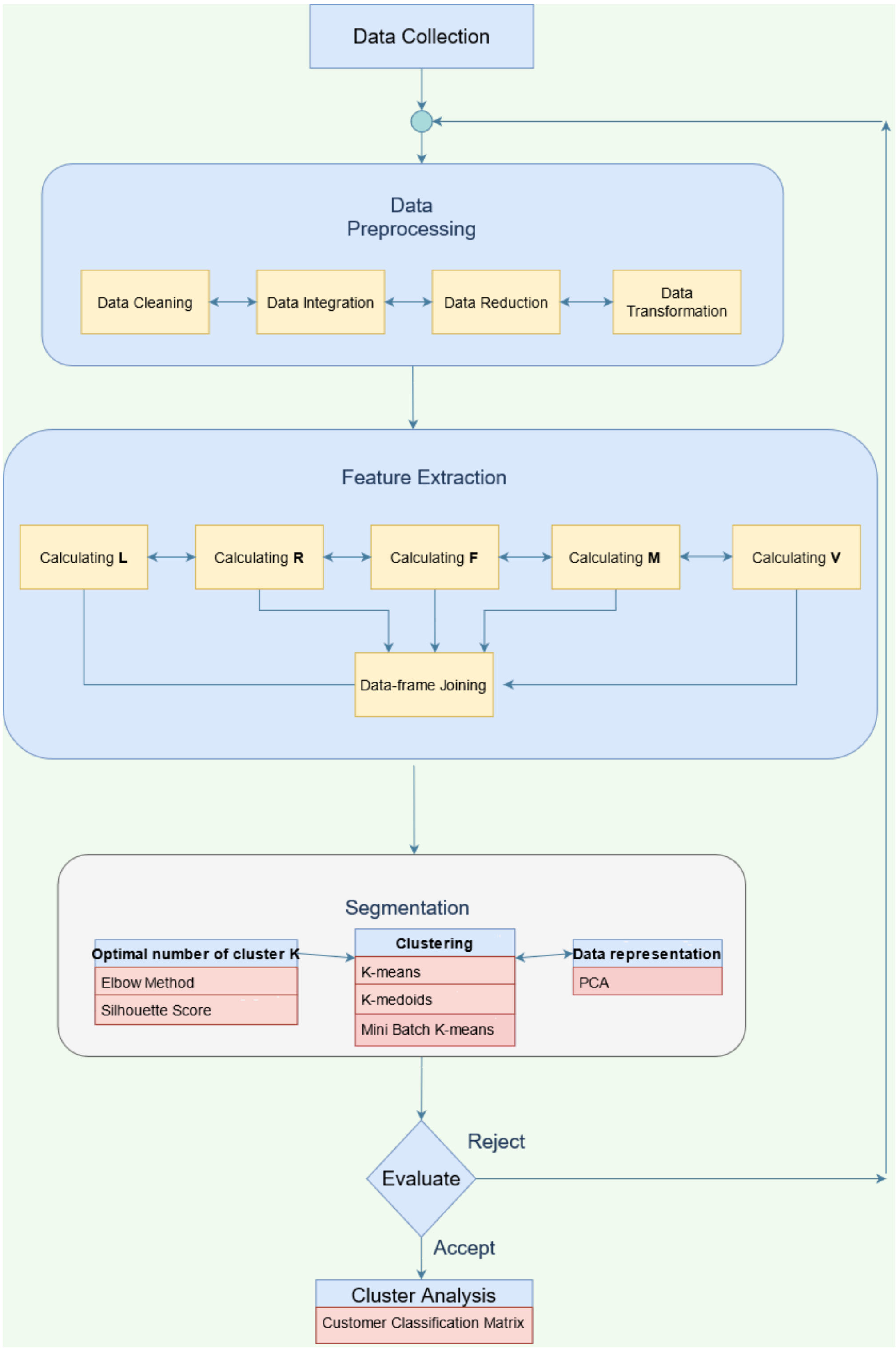
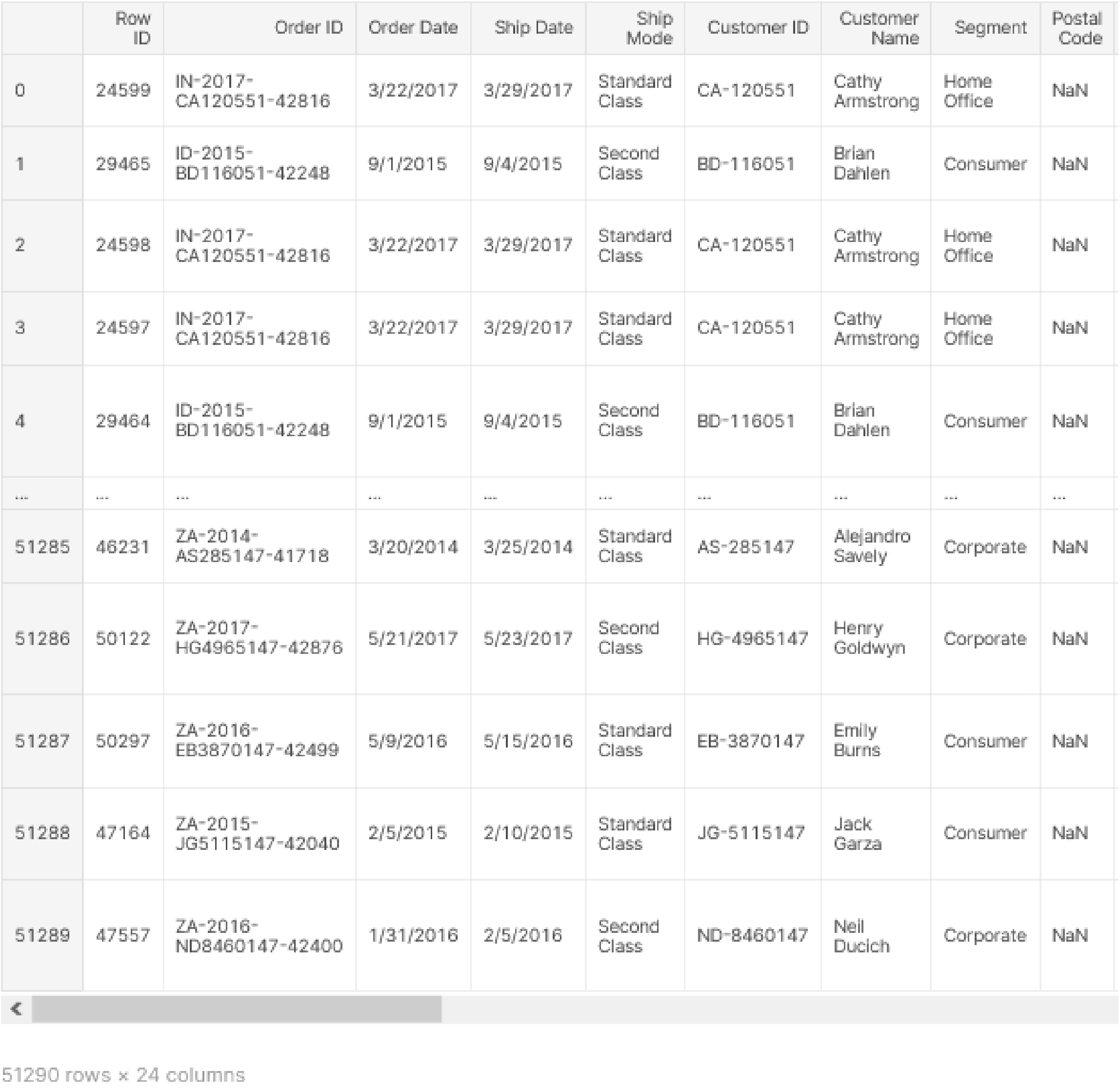


Figure 3.1: Work Process of the proposed LRFMV model

3.2 Data Collection

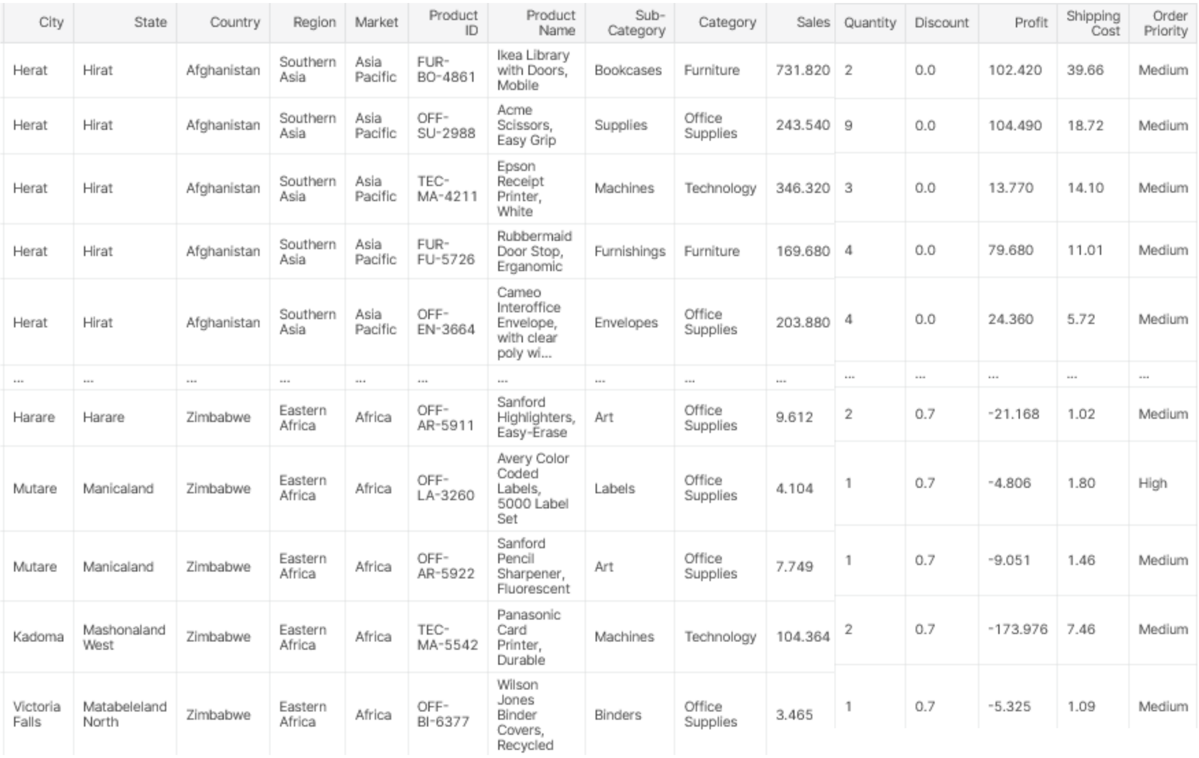
In this global pandemic of Covid-19, it was really impossible for us to search for a real-life dataset in person. The dataset was collected online which was shaped and published by Tableau Software which is a business intelligence software company based in the United States, creating interactive data visualization software. They basically work with databases and datasets of different organizations. Tableau aids in the visualization and comprehension of data. Their visual analytics platform is revolutionizing the way people solve problems with data. Businesses of all sizes rely on Tableau to help them become more data-driven. Tableau has a variety of web portal development services available. For small-scale deployments, it includes quick-start capabilities that can be finished in a matter of hours. They have a Tableau desktop, server, reader, and online service, among other things.



	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Postal Code
0	24599	IN-2017-CA120551-42816	3/22/2017	3/29/2017	Standard Class	CA-120551	Cathy Armstrong	Home Office	NaN
1	29465	ID-2015-BD116051-42248	9/1/2015	9/4/2015	Second Class	BD-116051	Brian Dahlen	Consumer	NaN
2	24598	IN-2017-CA120551-42816	3/22/2017	3/29/2017	Standard Class	CA-120551	Cathy Armstrong	Home Office	NaN
3	24597	IN-2017-CA120551-42816	3/22/2017	3/29/2017	Standard Class	CA-120551	Cathy Armstrong	Home Office	NaN
4	29464	ID-2015-BD116051-42248	9/1/2015	9/4/2015	Second Class	BD-116051	Brian Dahlen	Consumer	NaN
...	...	...	...	...	...	...	...	...	...
51285	46231	ZA-2014-AS285147-41718	3/20/2014	3/25/2014	Standard Class	AS-285147	Alejandro Savely	Corporate	NaN
51286	50122	ZA-2017-HG4965147-42876	5/21/2017	5/23/2017	Second Class	HG-4965147	Henry Goldwyn	Corporate	NaN
51287	50297	ZA-2016-EB3870147-42499	5/9/2016	5/15/2016	Standard Class	EB-3870147	Emily Burns	Consumer	NaN
51288	47164	ZA-2015-JG5115147-42040	2/5/2015	2/10/2015	Standard Class	JG-5115147	Jack Garza	Consumer	NaN
51289	47557	ZA-2016-ND8460147-42400	1/31/2016	2/5/2016	Second Class	ND-8460147	Neil Ducich	Corporate	NaN

51290 rows × 24 columns

Figure 3.2: Raw Dataset

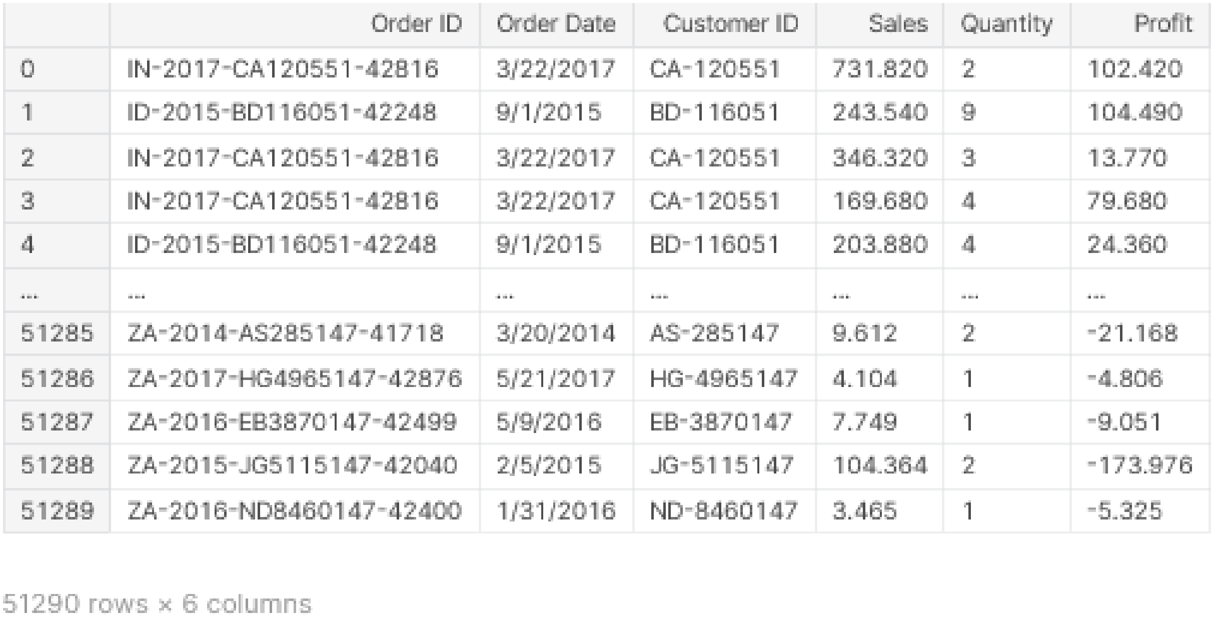


City	State	Country	Region	Market	Product ID	Product Name	Sub-Category	Category	Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority
Herat	Hirat	Afghanistan	Southern Asia	Asia Pacific	FUR-BO-4861	Ikea Library with Doors, Mobile	Bookcases	Furniture	731.820	2	0.0	102.420	39.66	Medium
Herat	Hirat	Afghanistan	Southern Asia	Asia Pacific	OFF-SU-2988	Acme Scissors, Easy Grip	Supplies	Office Supplies	243.540	9	0.0	104.490	18.72	Medium
Herat	Hirat	Afghanistan	Southern Asia	Asia Pacific	TEC-MA-4211	Epson Receipt Printer, White	Machines	Technology	346.320	3	0.0	13.770	14.10	Medium
Herat	Hirat	Afghanistan	Southern Asia	Asia Pacific	FUR-FU-5726	Rubbermaid Door Stop, Erganomic	Furnishings	Furniture	169.680	4	0.0	79.680	11.01	Medium
Herat	Hirat	Afghanistan	Southern Asia	Asia Pacific	OFF-EN-3664	Cameo Interoffice Envelope, with clear poly wi...	Envelopes	Office Supplies	203.880	4	0.0	24.360	5.72	Medium
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Harare	Harare	Zimbabwe	Eastern Africa	Africa	OFF-AR-5911	Sanford Highlighters, Easy-Erase	Art	Office Supplies	9.612	2	0.7	-21.168	1.02	Medium
Mutare	Manicaland	Zimbabwe	Eastern Africa	Africa	OFF-LA-3260	Avery Color Coded Labels, 5000 Label Set	Labels	Office Supplies	4.104	1	0.7	-4.806	1.80	High
Mutare	Manicaland	Zimbabwe	Eastern Africa	Africa	OFF-AR-5922	Sanford Pencil Sharpener, Fluorescent	Art	Office Supplies	7.749	1	0.7	-9.051	1.46	Medium
Kadoma	Mashonaland West	Zimbabwe	Eastern Africa	Africa	TEC-MA-5542	Panasonic Card Printer, Durable	Machines	Technology	104.364	2	0.7	-173.976	7.46	Medium
Victoria Falls	Matabeleland North	Zimbabwe	Eastern Africa	Africa	OFF-BI-6377	Wilson Jones Binder Covers, Recycled	Binders	Office Supplies	3.465	1	0.7	-5.325	1.09	Medium

Figure 3.3: Raw Dataset

Some attributes' description which has been extracted from the main dataset have been given below. These are the desired attributes for our proposed model.

Order id: It has been used to extract how many products were bought in a single order.
Order date: It needs to know how many orders were made by a single customer in a specific day.
Customer ID: In order to identify the sales for every customer,customer id is needed.
Sales: Sales is required to know the amount of transaction to calculate monetary.
Quantity: Quantity refers to the amount of a specific good.
Profit: It requires analyzing the data in the future.
Renamed the features to Sales: It refers to total purchase and calculating the total price on the product.



	Order ID	Order Date	Customer ID	Sales	Quantity	Profit
0	IN-2017-CA120551-42816	3/22/2017	CA-120551	731.820	2	102.420
1	ID-2015-BD116051-42248	9/1/2015	BD-116051	243.540	9	104.490
2	IN-2017-CA120551-42816	3/22/2017	CA-120551	346.320	3	13.770
3	IN-2017-CA120551-42816	3/22/2017	CA-120551	169.680	4	79.680
4	ID-2015-BD116051-42248	9/1/2015	BD-116051	203.880	4	24.360
...	...	...	...	...	...	...
51285	ZA-2014-AS285147-41718	3/20/2014	AS-285147	9.612	2	-21.168
51286	ZA-2017-HG4965147-42876	5/21/2017	HG-4965147	4.104	1	-4.806
51287	ZA-2016-EB3870147-42499	5/9/2016	EB-3870147	7.749	1	-9.051
51288	ZA-2015-JG5115147-42040	2/5/2015	JG-5115147	104.364	2	-173.976
51289	ZA-2016-ND8460147-42400	1/31/2016	ND-8460147	3.465	1	-5.325

51290 rows × 6 columns

Figure 3.4: Dataset with selected columns

The dataset was originally collected from a superstore by Tableau Software, which later modified it. There is no personal information in this dataset, such as age, occupation, income, or name of any customer. These are hidden by Tableau Software to ensure customer confidentiality and privacy. They also filled many items of any attribute with arbitrary or imaginary values in the dataset. Global Superstore is a data set that contains around 50000 records. It's a client/customer-driven informational dataset that contains relative data of orders placed through various sellers and markets from 2011 to 2018. Ship Date, City, Category, and Order Priority were among the 24 attributes included in the dataset.

3.3 Data Preprocessing

3.3.1 Data inspection before preprocessing

Superstore business has been an emerging business for the last few years in the South Asian region and the process of finding potential and profitable customers will be really beneficial for this sector. This research aimed to scrutinize the cluster of potential buyers by preprocessing a dataset of a superstore.Before heading towards preprocessing,some graphical representations have been plotted in order to have a short glimpse of the raw dataset of a superstore.

The following graph is giving a visual representation of profit earned from different products under some sub-category being sold in the supershop. It is visible that copiers make the largest profit in that supershop where tables category is making loss

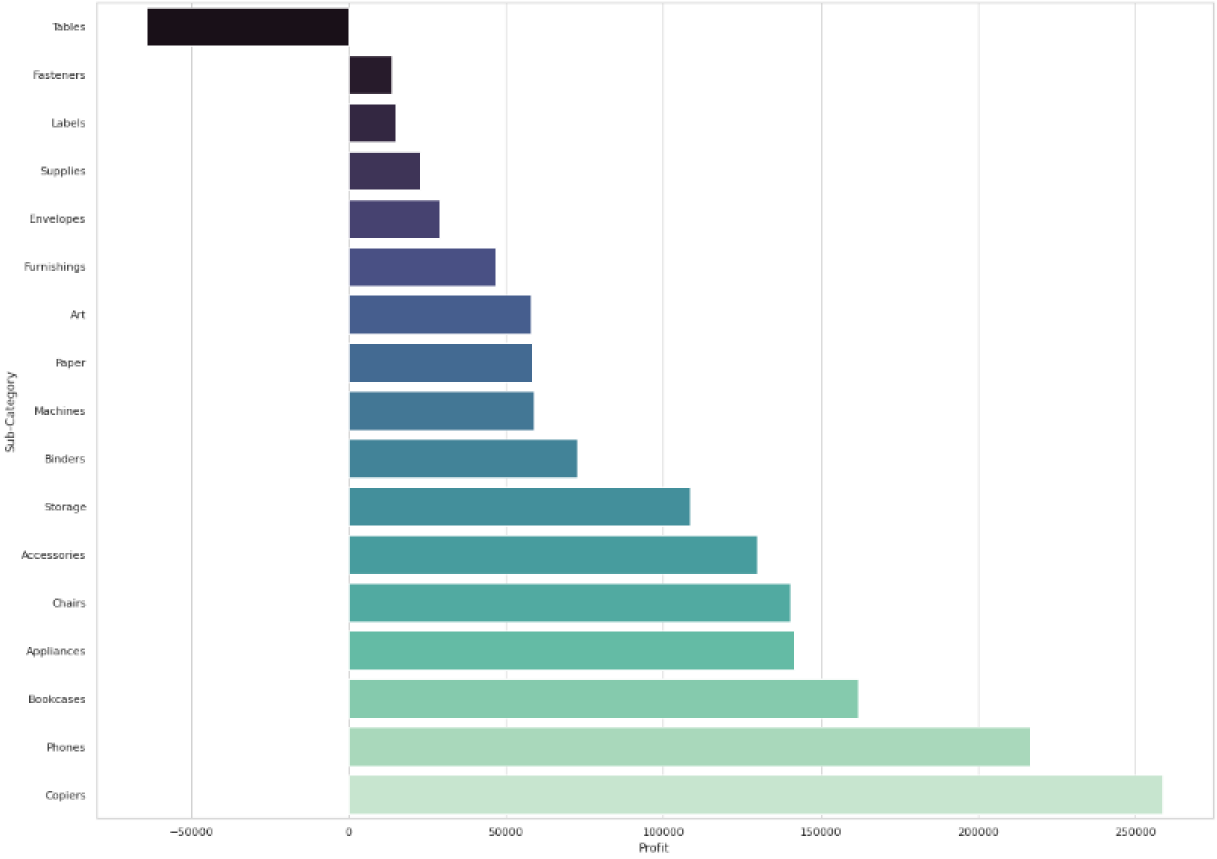


Figure 3.5: Profit made from different sub-categories of products

The following pie chart is showing the sales on the basis of some main categories of products sold in that superstore. It is visible in the pie chart that the technology category is the most sold item in the superstore which means that the superstore earns the most revenue from this item.

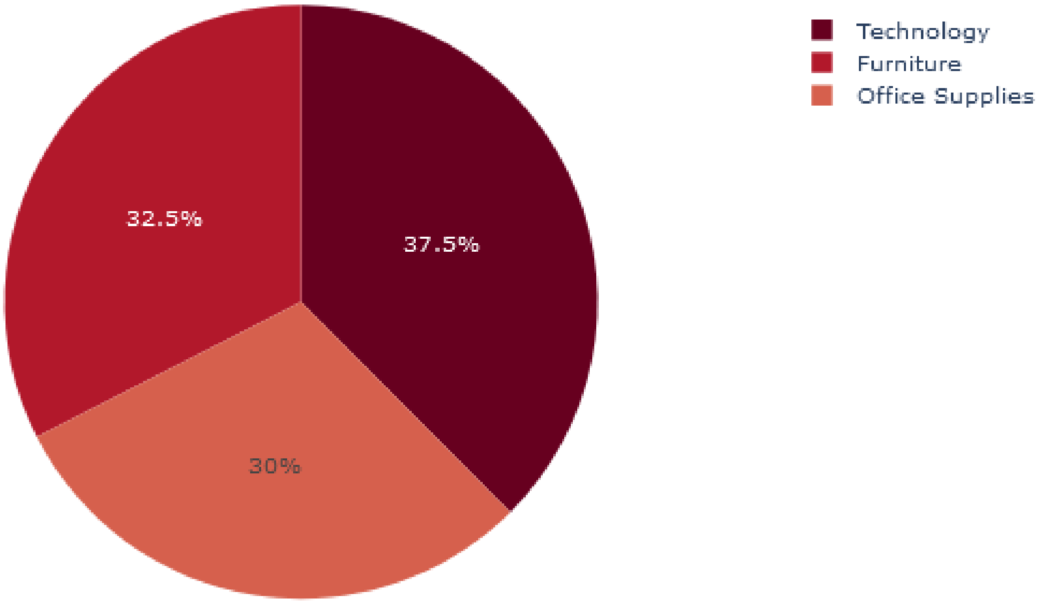


Figure 3.6: Sales per category

The below mentioned bar chart is similar to the figure 3.4 but it has been plotted with respect to date and more sub-categories have been added.Sales of different types of products on different dates is clearly noticeable to understand the earned revenue and it helps to analyze profit margin.

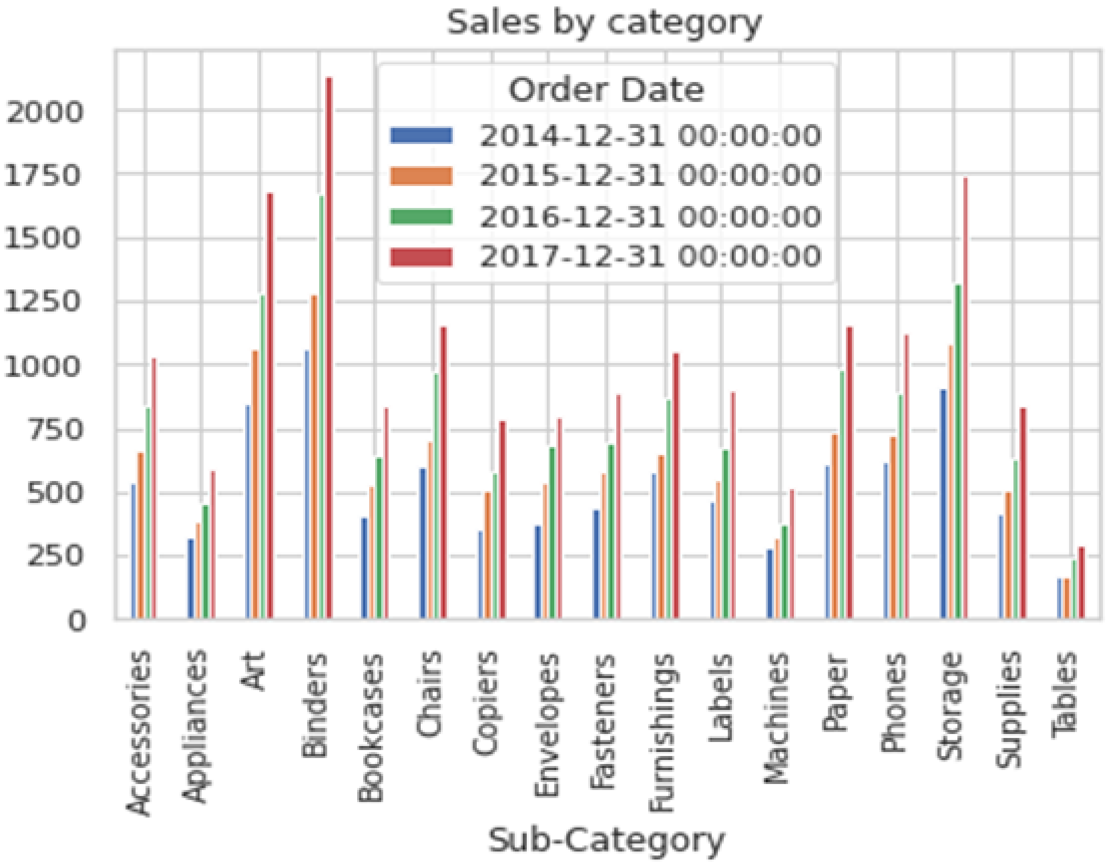


Figure 3.7: Barchart of sales by category based on order-dates

3.3.2 Data Cleaning

Data cleaning procedures are used to substitute missing values and fix incorrect data. Each of these activities can be carried out in a variety of ways, depending on the user's preferences or problem set. Each task is discussed in detail below, along with the methods used to execute it.

The phrase ”noise” refers to a large volume of irrelevant information. Duplicates or semi-duplicates of data records, data segments with no value for specific study, and unnecessary information fields for each variable are examples. Random data can be removed via binning, regression, or clustering. [28].The binding strategy is used to smooth out sorted data. The data is separated into equal-sized chunks, and the procedure is carried out in a number of ways. Each segment is handled separately. All data in a segment can be replaced by its mean, or boundary values can be used to finish the work. Linear regression and multiple linear regression can be used to smooth the data when the values are conformed to a function. The process of discovering and dealing with outliers using techniques such as data clustering is known as outlier analysis. Multiple ways can be employed to cope with missing data. If the output label for the classification problem is absent, the training example can be ignored.

The mean or median can be used to fill in the missing value using central tendency for characteristics to replace the missing value. And each class has its own central tendency metrics.

Eliminating the training is usually discouraged since it results in data loss because attribute values are being removed that can bring value to the data collection[29].

Filling in missing values is lengthy and is not advised for comprehensive data sets.

Global constants like 'N/A' or 'Unknown' can be used to fill in the missing value when using a standard value to replace it. Although this is a simple method, it is not without problems. Missing values can be anticipated and replaced using methods like regression and decision trees, which use the most likely value to fill in the missing value.

For data cleaning, the central tendency for attributes to replace the missing value has been used[50]. In the dataset, there were many arbitrary and null values which were creating a barrier in establishing the proposed model. So, the mean or average value has been found for those columns where null values occurred frequently.

3.3.3 Data Integration

Data integration has turned into a critical aspect of pre-process because data is acquired from numerous sources. This could result in duplicate and inconsistent data, resulting in poor data model accuracy and speed.

To deal with these issues and ensure data integrity, techniques such as tuple repetition detection and data conflict detection are being investigated. There are various techniques for accomplishing this integration, which are listed below.

The term data consolidation refers to the process of physically combining data into a single data repository. Data Warehousing is generally included.

Data propagation is a process that involves the transmission of information[47]. Data propagation is the process of copying data from one area to another utilizing programs. It is event-driven and can be synchronous or asynchronous.In data virtualization, an interface is used to deliver a real-time and uniform view of data from multiple sources. In data virtualization, an interface is used to deliver a real-time and uniform view of data from multiple sources. All of the data is accessible from a single location.

The dataset did not need any data integration as it was integrated enough to do the calculation and plotting.

3.3.4 Data dimensionality reduction and Feature selection

Analysis gets more challenging when dealing with vast volumes of data, as data min- ing is a methodology for dealing with big amounts of data. To prevent this, a data dimensionality reduction technique was used. Its purpose is to increase storage effi- ciency while cutting costs for data storage and processing. The main aim of this research paper is to add a new component volume to the existing LRFM model.To reach the goal, many attributes need to be dropped from the raw dataset and unsupervised learning allows us to drop the inconsequential attributes and select the important features.

3.3.5 Data Transformation

The final stage of data preprocessing is to convert the data into a format suitable for Data Modeling. This stage transforms the data into a format that can be used in the data mining process. This can be done in a variety of ways:

Construction of new attributes/features: To add new attributes, a group of attributes is used. Additional qualities are built from the supplied collection of attributes in this methodology to enhance the mining process..
Aggregation: Summary and aggregate procedures are performed on a group of attributes to create new attributes.
Normalization: The data for each attribute is scaled to a smaller range. This is used to scale data values inside a specific range (-2.0 to -1.0 or 1.0 to 2.0)
Discretization:Discrete or conceptual intervals are used to replace raw numeric attribute values, which can subsequently be structured further into higher-level intervals. The raw values of the numeric attribute are replaced with interval or conceptual levels in this way[50].
Nominal data concept hierarchy generation: Higher-order ideas are generalized from nominal data values. Attributes are changed from one level of the hierarchy to the next in this example.

Transforming the data into a format suitable for Data Modeling is the final phase of data preprocessing. This step is conducted to change the data into a format that can be used in the mining process. In the data transformation phase, a set of attributes (Length, Recency, Frequency, Monetary, Volume, K-means, K-Medoids, Mini Batch K-means etc.) is used to create new attributes. To aid the mining process, additional attributes are built from the supplied collection of attributes in this technique. To create new attributes, summary, and aggregation operations are conducted to a set of attributes. The extracted L, R, F, M, V is rescaled with data standardization technique in such a way that the mean value of the attribute is 0 and standard deviation for the resultant distribution is 1.

3.4 Feature Extraction

The LRFMV model is used to define key and potentially important customers in this study of a superstore's consumers. The K-means strategy is used to partition a total of 51290 records into six clusters when the segmenting variables are L, R, F, M, and V.K-Medoids and Mini Batch clustering approach have also been used but K-Means is considered as the most suitable one.As a result, the superstore will be able to distribute and leverage capital more effectively and efficiently, allowing it to meet a wider range of customer needs while still increasing profits[47]. The customer values in the clusters can be generated and examined to provide useful decision-making information for the superstore's client retention and profitability based on different customer classes.

The LRFMV (length, recency, frequency, monetary, and volume) model is a straightforward but effective market segmentation method. According to this article, LRFMV analysis will segment the client base and maximize the purchase reaction rates of marketing efforts. LRFMV research enhances market segmentation by looking at how long (length), when (recency), how much (frequency), how much money (monetary), and how much money a customer spends (volume). Customers who had recently invested a lot of money and purchased a lot of items were much more likely to react to potential promotions, according to the study. As a result, the scope of LRFMV research has been broad[47]. As a result, direct marketers may be able to benefit from LRFMV research. Customers are categorized into four categories based on frequency and monetary value: best, spender, uncertain, and frequent. Through length and recency, the customer relationship matrix assists management in identifying the characteristics of four different types of customer relationships[28]. The volume highlights the customers who provide more profit to the organization as their buying habit is larger than any other customer segment.

3.4.1 Calculation of L

Difference of days between a customer's first and last visit is referred to as length in the LRFMV model. It denotes the distance between two specific visits, or more specifically, it refers to the purchases made on these two dates. Mathematically, if the last purchase date and the first purchase date for a particular customer is denoted by p_l and p_f respectively then Length, L can be calculated as,

$$L = p_l - p_f$$

Table 3.1: Calculated Length (L)

In the above table, it can be observed that Length (L) has been calculated with respect to a customer id. After the customer has completed their first and last purchase, the length of these two purchases has been calculated. It is, in essence, the time interval between two consecutive sales. So, if all the individual's purchases have been added between these two times,the Length(L) can be found.

	Customer ID	**Length**
0	AA-10315102	919.0
1	AA-10315120	0.0
2	AA-10315139	319.0
3	AA-103151402	483.0
4	AA-103151404	553.0
...	...	...
17410	ZD-2192548	385.0
17411	ZD-2192564	0.0
17412	ZD-219257	0.0
17413	ZD-2192582	570.0
17414	ZD-2192596	0.0

3.4.2 Calculation of R

Recency refers to the days after the last visit of any particular customer. It indicates the days which exist after any valuable customer's last purchase to find the irregularity after that visit. The Recency (R) value was calculated. Mathematically, If the most recent date of the dataset is denoted by D_r and the last purchase date of a particular customer is C_r then Recency, R can be calculated as,

$$R = D_r - C_r$$

In the following table, the above formula has been used for calculating the recency(R) for each customer id.

Table 3.2: Calculated Recency (R)

	Customer ID	**Recency**
10553	AA-10315102	358
5926	AA-10315120	960
7922	AA-10315139	149
3280	AA-103151402	184
1185	AA-103151404	819
...	...	...
6849	ZD-2192548	751
5925	ZD-2192564	1409
11818	ZD-219257	1199
20962	ZD-2192582	196
15736	ZD-2192596	750

To do so, firstly the most recent date in the dataset had been found and saved as a variable, then stored only the most recent dates of each customer in a data frame. Then each customer's most recent visit/purchase date had been subtracted from the dataset's most recent date. Recency was assigned to this data.

3.4.3 Calculation of F

The number of purchases made by a customer in a customer life cycle of time is referred to as frequency. Counting the total number of times a customer purchased any service from the superstore yielded the Frequency (F) value. Mathematically, If the purchase for a customer is denoted by p_f then Frequency, F for that particular customer will be,

$$F = count\,(p_f)$$

Here, by the count method, the total number of purchases for a particular customer is calculated. We will calculate the frequency per customer id, using this formula in the below-stated table.

Table 3.3: Calculated Frequency (F)

	Customer ID	**Frequency**
0	CS-121757	9
1	DB-1361548	8
2	SH-203951406	7
3	RP-193901406	7
4	AI-108551404	7
...	...	...
17410	BS-1755134	1
17411	RB-1946548	1
17412	VP-2173026	1
17413	MC-17425130	1
17414	JF-5415137	1

3.4.4 Calculation of M

Total number of transactions and total expenditure for all those transactions were needed to calculate the monetary.In a customer's whole life cycle,total amount of his comprehensive transactions has been divided with the number of those transactions in order to find the value of monetary.Mathematically, if the total spending on purchasing of a customer is p_s, x is the total number of transactions denoted here. Monetary, M can be calculated as,

$$M = \frac{\sum_{n=1}^{x} p_s}{x}$$

Using this formula, we will calculate the money for each customer id.

Table 3.4: Calculated Monetary (M)

	Customer ID	Monetary
0	AA-10315102	272.3280
1	AA-10315120	2713.4100
2	AA-10315139	738.9495
3	AA-103151402	2390.2760
4	AA-103151404	376.7540
...	...	...
17410	ZD-2192548	434.0560
17411	ZD-2192564	1225.3920
17412	ZD-219257	59.9400
17413	ZD-2192582	339.0507
17414	ZD-2192596	269.3100

As a result, we decided to use the mean value per sale as our monetary value (M) using the above table. Monetary is the com payment fee of any purchase of a customer.

3.4.5 Calculation of V

Volume is a rescaled version of the quantity of goods purchased by a potential customer.It identifies a group of valuable customers for any company by highlighting the amount of their purchased product over a set period of time or a set number of visits.It is the proposed term by us in this research paper which adds value to the LRFM model by figuring out the customer clusters that give more profit to an organization. It shows that if a customer buys a large amount of product in his certain visits regardless of the spent money he will contribute more to the profit of that organization.

Table 3.5: Calculated Volume (V)

	Customer ID	**Volume**
0	AA-10315102	4.875000
1	AA-10315120	7.000000
2	AA-10315139	2.479167
3	AA-103151402	3.000000
4	AA-103151404	2.000000
...	...	...
17410	ZD-2192548	2.833333
17411	ZD-2192564	2.250000
17412	ZD-219257	4.000000
17413	ZD-2192582	2.500000
17414	ZD-2192596	5.500000

In the above-mentioned figure,the volume had been calculated for each customer id using the volume formula. The detailed formula is like this- Mathematically, if the quantity of purchased products for a particular customer is Q, x is the number of transactions made by a specific customer on a specific day while calculating the average of different attributes, n is the number of days while transacting in his customer life cycle.

$$V = \frac{\sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{x} Q_j}{x} \right)_i}{n}$$

In this paper, the mean quantity is being calculated by grouping the dataset with both the customer id and the sale date. It will provide us with the average quantity of products purchased by each customer on specific days.The quantity's mean will be calculated once more. This time, however, it will be grouped solely by customer id. It will be referred to as Volume (V).

3.5 Correlation of volume with other features

To select profitable customers for superstores, all six LRFMV parameters must be considered. Each parameter should have distinct characteristics that do not overlap with other attributes.The RFM model has been by including two new variables, L and V. Previous research demonstrated how L is associated with other attributes and why it is important to consider in order to enhance profitability in market segments. [42] [31] [60] [57] A heatmap is generated here to visualize the relationship between the newly added feature called volume and other parameters L, R, F, and M.

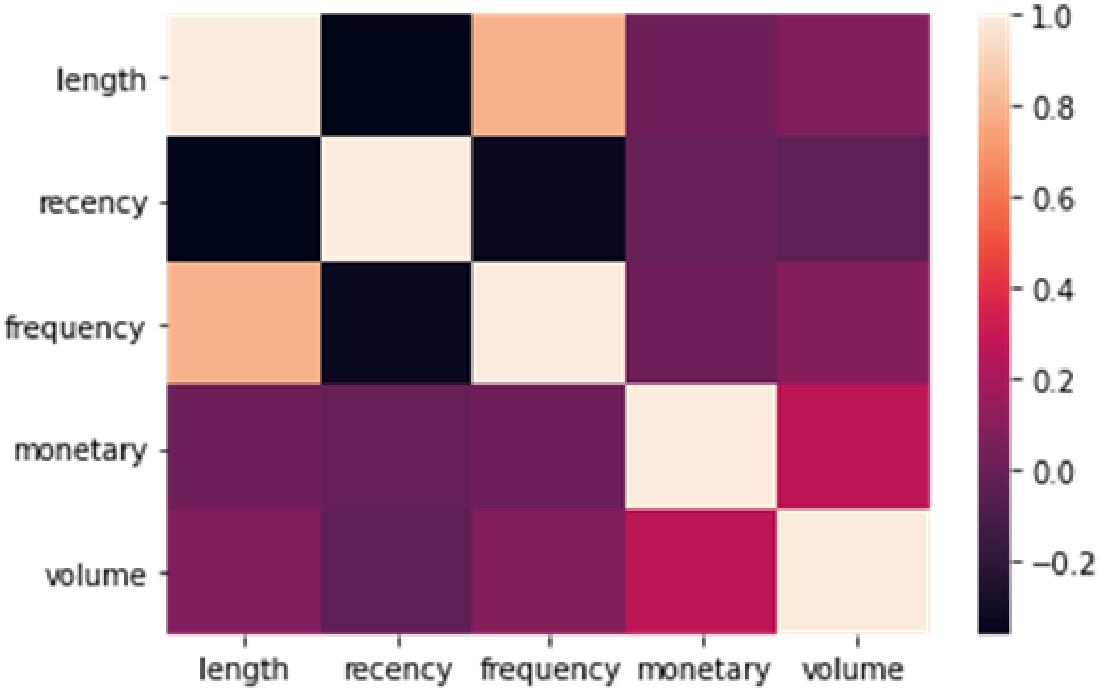


Figure 3.8: Heatmap for LRFMV model

The graph above represents the relationship between each parameter. The indication ranges from light to dark, or from strongly associated to less connected (1.0 to -0.2). It is observed that the newly introduced parameter V has a very weak relationship with each of the other parameters L (0.2), R (0.0), F (0.2), and M (0.4), and is quite strong with itself. It signifies that V is a unique feature and can not be replaced by other features. Therefore, to validate the statement, the LRFMV model can be employed as a unique model and see if it outperforms the traditional RFM and LRFM model.

3.6 Model Specification

Focusing on the process of determining some independent variables need to include and exclude from the research which is referred to as model specification follows some methods.Elbow method,Silhouette coefficient,Cumulative Explained Variance Ratio and PCA have been used as the paramount process as the model specification for this research.

3.6.1 Elbow Method

The elbow approach is a very effective method for identifying the optimal cluster numbers in clustering procedures. The approach first came into the discussion by Thorndike in 1953 [31]. The primary idea behind this strategy is to choose the elbow of the curve of the cluster association graph with error reduction and then increase K values until the gain of K is stable.

The maximum number of clusters that should be employed for a given dataset is determined by the elbow curve that is chosen by the heuristic approach. The elbow method based on SSE values generates the best number of clusters for larger datasets[52]. In this research, K-values were measured to select the best clusters based on SSE as a performance indicator of elbow technique. For a definite clustering methodology, the average distance between the points inside a cluster represents the average internal sum of squares.

In terms of mathematics,

$$SSE = \sum_{k=1}^{k} \sum_{x_i = S_k} \|x_i - \alpha_k\|_2^2$$

For any given dataset let X $= \{x_1, x_2, x_3 \ldots x_n\}$, x having n points, K is the cluster numbers where $K < n$ and $k \in \{1, 2, 3, 4, \ldots K\}$. In the above formula α_k denotes the centroids of the associated cluster S_k ,where

$$\alpha_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} X$$

After computing SSE for each cluster, the result is put onto a graph. The ideal number of clusters is found by identifying the bend in the curve.

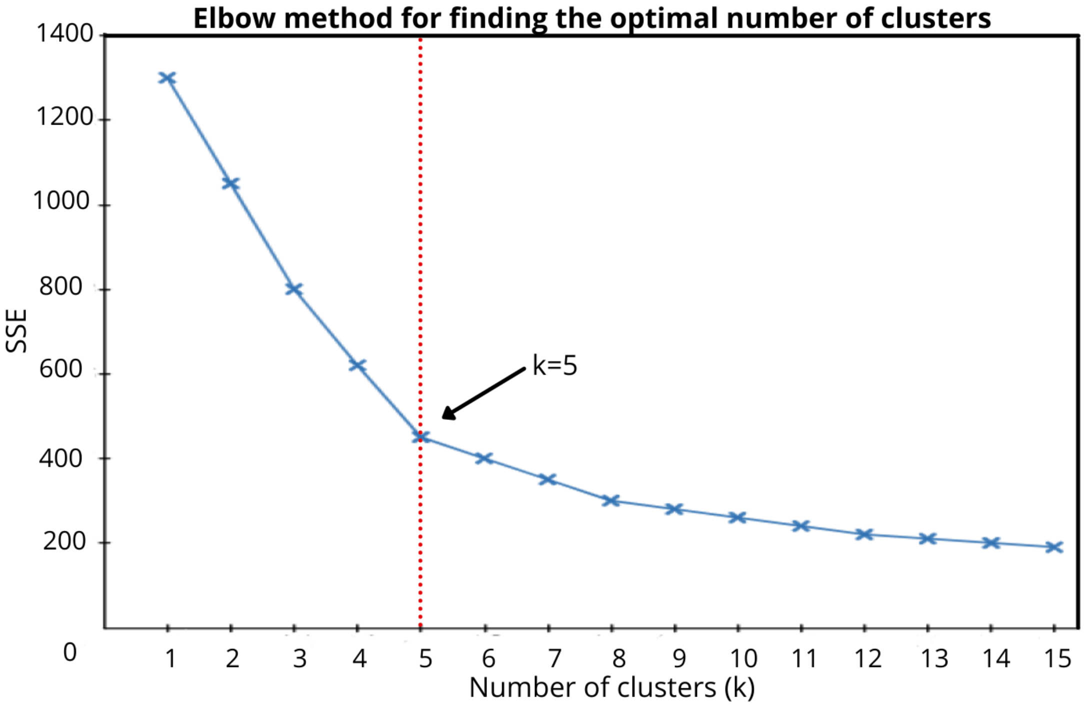


Figure 3.9: Elbow Method

In the above figure, the number of clusters, $k = 1, 2, 3, 4, ...K$ is shown on the x-axis and the corresponding SSE values for each cluster are shown in the y-axis. When the value of k is raised, the graph begins to flatten significantly and the SSE value becomes insignificant. Here a bend is seen on the graph for the value of $k = 5$. The mentioned thing actually demonstrates the ideal amount of clusters which is needed for the segmentation for the particular dataset X. When the value of k is increased from 5 the graph gradually smoothes out. If the k value is excessively high, the bend actually prevents the clusters from overfitting the data. Many essential clusters may be eliminated from the data if the value of k is excessively low.

As a result, selecting the ideal amount of clusters is actually required to achieve effective data segmentation. The Elbow method consists of less complexity and is really easy to use. Additionally, as a validity indicator, SSE works well in terms of determining how effectively an object functions. When there are several bends or the overall curve is flattened, the elbow approach fails. Other approaches for determining the optimal number of clusters is also utilized in certain circumstances. Furthermore, when dealing with huge data sets, the elbow approach requires a lot of time to execute since the SSE value must be computed for each cluster. In most circumstances, however, the elbow technique is quite beneficial for calculating the ideal number of clusters.

This method can be a potential one for determining an optimal clusters for the K-means method. This K-means process can give a variety of clusters depending on the dataset and user's demand. Data variance, dimensions, etc are the reactors for the total amount of clusters in case of particular datasets. But an efficient number of clusters is really important for creating any kind of model to avoid data traffic and a clear understanding of a paper. For ensuring the effective cluster numbers for the dataset, the elbow method is used.

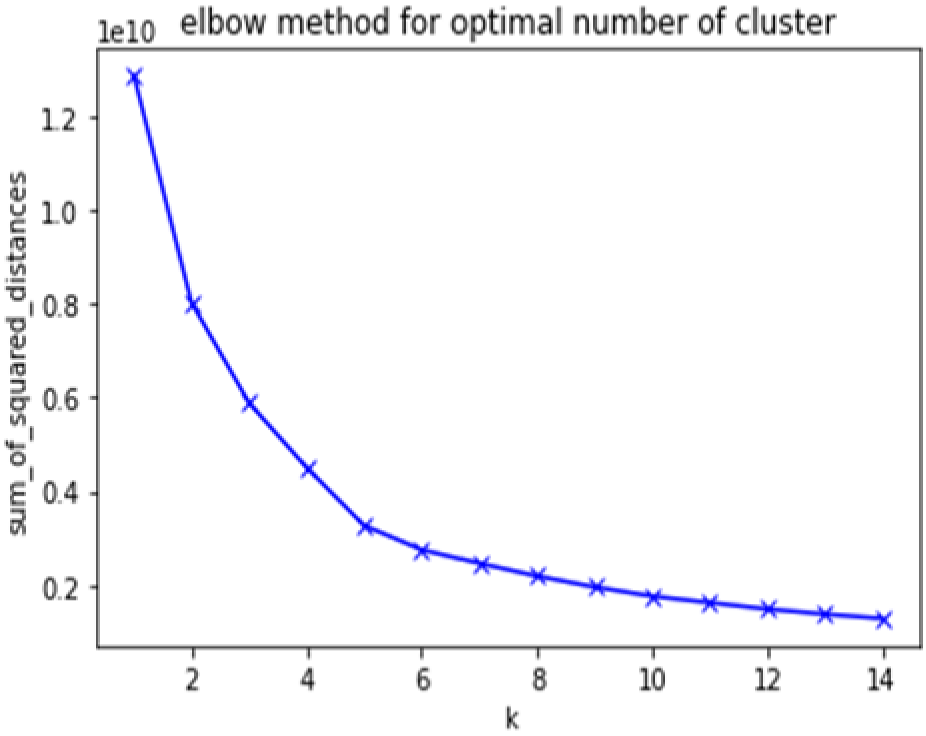


Figure 3.10: Elbow Method for the LRFMV model

In the graph, it is noticed that an optimal cluster numbers is on x-axis, and the sum of squared distances is on y-axis. It can be observed that the most curved line lies between 4 and 6. If the ceiling margin is being considered, 6 will be considered as the number of efficient clusters.

3.6.2 Silhouette Coefficients

The silhouette coefficient is a statistic used to determine the quality of a clustering process. Its value is between -1 and 1.
1: Specifies that clusters are clearly distinct and isolated from one another.
0: Implies that clusters are indifferent to one another or that the spacing between clusters is quite small.
-1: Indicates that clusters have been allocated incorrectly.
According to Rousseeuw (1987)[3], the silhouettes are beneficial when proximities are measured in ratios (as with Euclidean distances) and when dense and clearly differentiated clusters are desired. Indeed, the definition makes use of average proximities in the same way as demonstrating a positive association linkage does, which is known to perform best when clusters are approximately spherical.
Mathematically, silhouette coefficient can be written as,

$$S(i) = \frac{c(i) - d(i)}{\max\{c(i), d(i)\}}, if\ |X_i| > 1$$

where an d(i) represents the average intra-cluster distance between object i and other data points, that is, the average distance between each point within a cluster. And c(i) is the average distance between the cluster where the object i belongs to all other clusters the object i does not belong, i.e. the distance between all clusters and X denotes the cluster.
When a cluster contains a single object, it is impossible to clearly define the value of d. As a result, the value of ”d” should be zero, which appears to be the most neutral value.

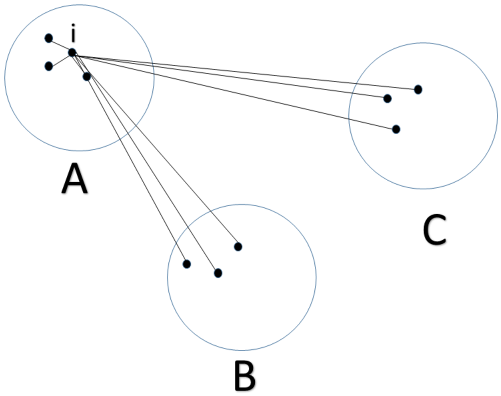


Figure 3.11: A visual representation of objects involved in the calculation of S(i)

As it is known that this method is used for confirming the methodical number of clusters like the elbow method,this method is being used for confirming the cluster number. For applying the standard K-means, K-Medoids and Mini Batch K-means algorithms it is compulsory to know how many clusters are appropriate for the dataset based on the variety of information.The Silhouette Coefficients were used to determine the number of clusters. The Silhouette score yielded the same result of 6 points for the three algorithms.

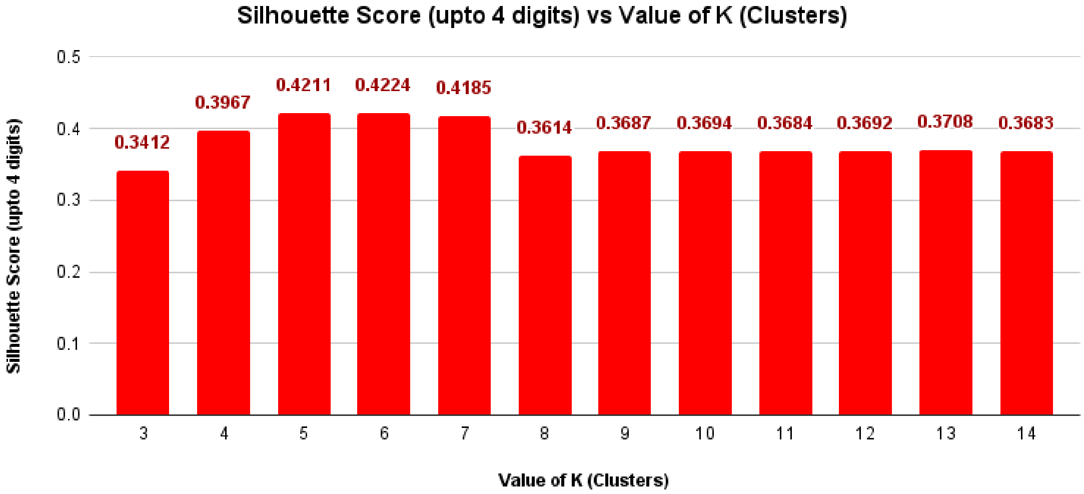


Figure 3.12: Silhouette Score

In the Silhouette Coefficient calculation, it is spotted that the value of the Silhouette Coefficient for the standard K-means has started decreasing from cluster 7. Till cluster 6 it was rising and the value was 0.4224 for cluster 6, and for cluster 7, it was 0.4185. It indicates the downfall of coefficient value which continues till cluster 14.

3.6.3 Cumulative Explained Variance Ratio

To lower the complexity of our suggested model, we must first lower the dataset's dimensionality. To develop an easy-to-understand model that is also fast to execute, we must first determine the suitable dimension of the dataset, and principal component analysis (PCA) is a powerful technique for reducing the dimension of the data. However, before we go into PCA, we must ascertain the utility of our primary components. We utilize the Explained Variance Ratio to determine the utility of your primary components and the number of components to include in our model. The explained variance ratio indicates the proportion of variation that each of the specified components contributes. The cumulative percentage indicates how much variation is explained by the first n components. For instance, the cumulative Rate for the second component is equal to the total of the first and second components' variance percentages.

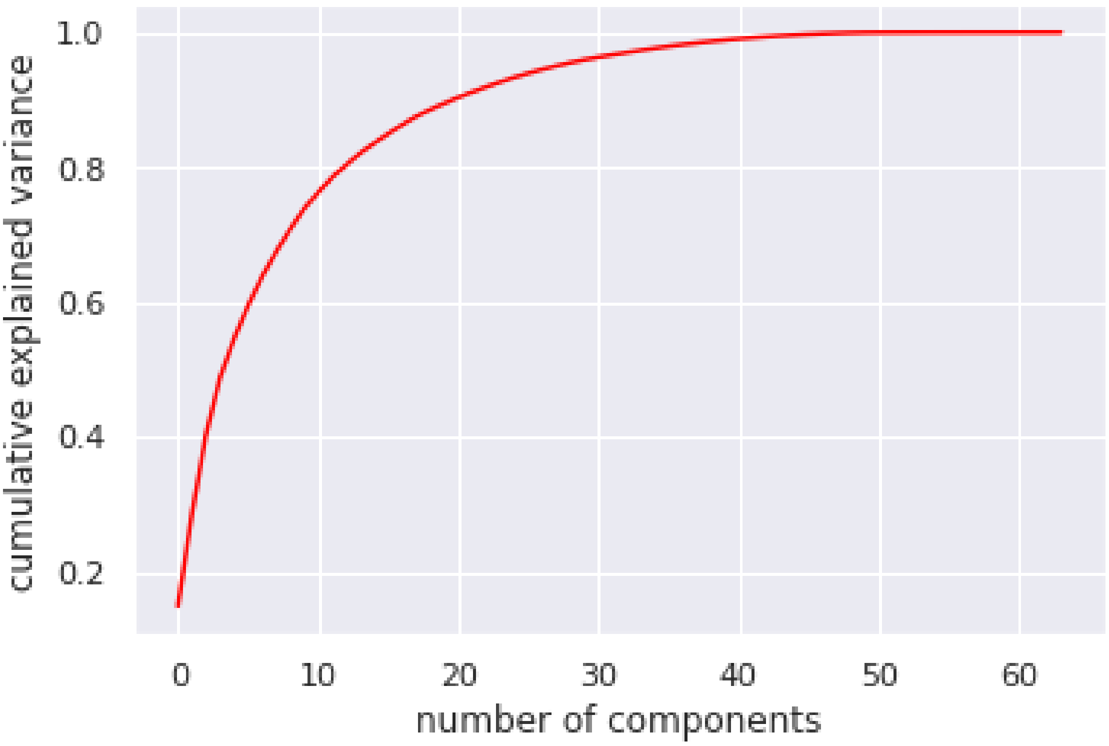


Figure 3.13: Cumulative Explained Variance for random 200 points

Here,it is seen that the first 10 components contain 75% of the variance, on the other hand, approximately 50 components are needed to describe nearly 100% of the variance.
Before using k-means,it is required that how many dimensions are appropriate for representing this dataset or observing trends, jumps, clusters, and outliers.The Cumulative Explained Variance Ratio method is used to do so.

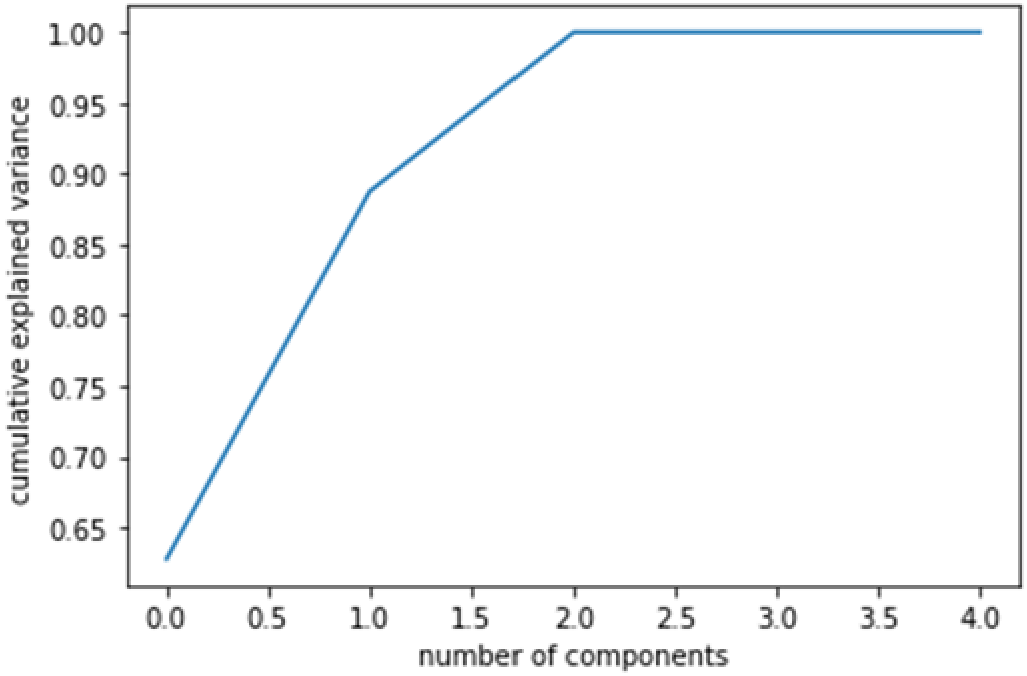


Figure 3.14: Cumulative Explained Variance for our proposed dataset

The number of components vs. variance graph has been plotted using this method.Any number of components can be combined to achieve a ratio of 0.90 to 1.00. (90 to 100 percent). The more the number of components the more will be accurate. So,a round figure has been chosen which is 2 and it shows that 2 or higher numbers as the number of components as the optimal result. So, 2 has been chosen.

3.6.4 PCA

For different fields of unsupervised machine learning, PCA is a remarkably successful approach. Pearson first [1] proposed principal component analysis (PCA) in 1901, and Hotelling [2] later on improved it in 1933 and named it as ”principal components”. It’s a dimensionality reduction approach that involves compressing a dataset of higher dimensional feature space into a lower dimensional feature subspace with the goal of maintaining the majority of the relevant data [1] , [35]. Maintaining a big number of dimensions in the feature space might result in a big volume of space, which can affect the effectiveness of data mining algorithms for a given dataset. As a result, dimension reduction actually aids in increasing the algorithm’s accuracy by extracting as much variation from the original data as necessary. Mathematically PCA aims to create a linear mapping L in order to maximize the variations from the original data D.

In other words, PCA solves the equivalent eigenvalue problem of the covariance matrix by maximizing the cost of the function $L^T ML$ with respect to L[35] Eigenvalue and Eigenvectors always come in pairs and their numbers are equal to the dimension of the data Principal components are actually the direction of the Eigenvectors where Eigenvalues provide the variance from each principal component. For any given data D, if d_{xy} demonstrates the pairwise Euclidean matrix for a high dimensional data and $\|\gamma_x - \gamma_y\|$ is the Euclidean distance between low dimensional data points γ_x and γ_y then PCA will try to find the linear mapping L for maximizing the cost of function:

$$P(Y) = \sum_{x,y} \left(d_{xy}^2 - \|\gamma_x - \gamma_y\|^2\right)$$

By maximizing the variances from the original data, the higher dimensional feature space is reduced to lowerer dimensional feature subspace after computing the Eigenvalues from the Eigenvectors.

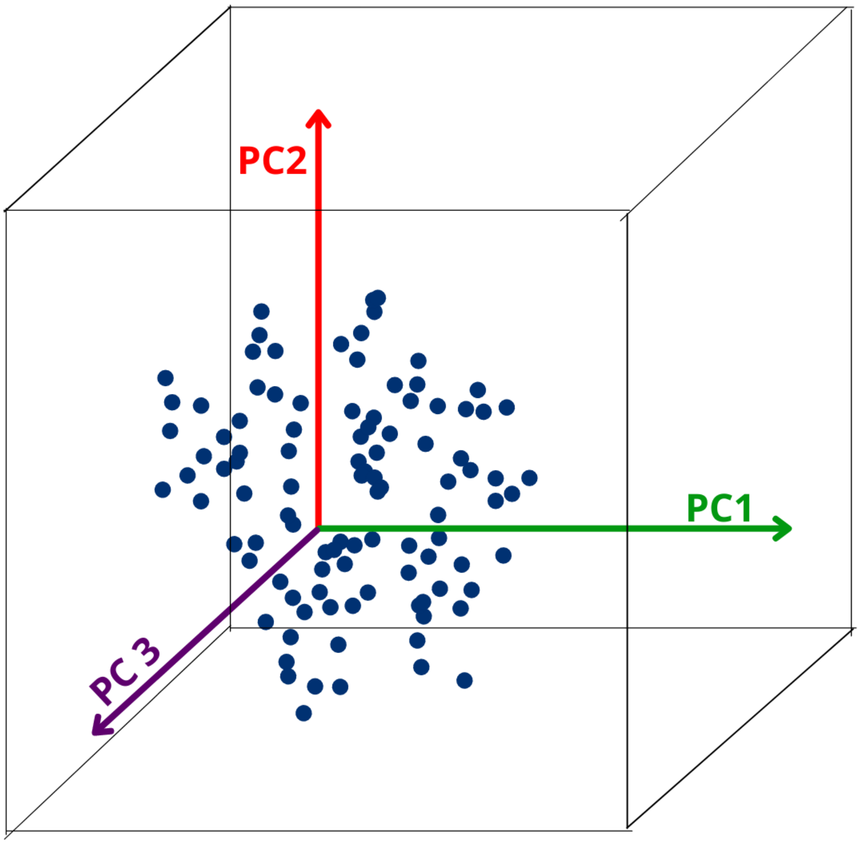


Figure 3.15: 3D featured dataset

Let X, Y, Z be three-dimensional axes where PC1, PC2, and PC3 are 3 principal components of the original space. We observe some scattered data points in the high dimensional 3D region of the original data.

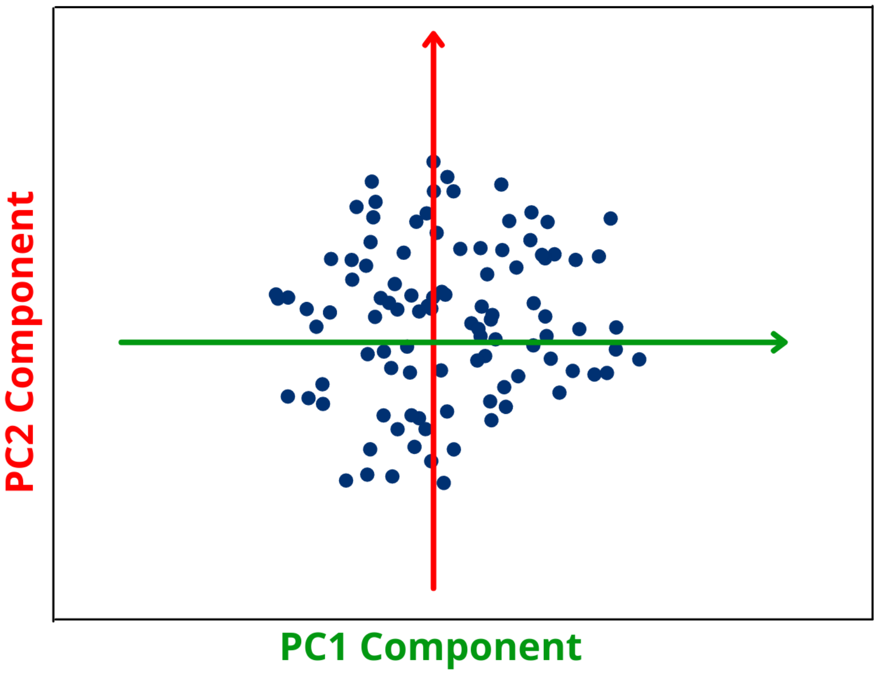


Figure 3.16: Dimensionality reduction to 2D using PCA

After applying PCA in the original data the dimension is reduced from 3D to 2D by finding the maximum variance by computing and reorienting the coefficients of Eigenvalues from the Eigenvectors of the covariance matrix. When input characteristics have a lot of dimensions and data is difficult to visualize, PCA is applied. By minimizing the space by increasing the variations, PCA also ensures that valuable information is present. It also gives the variables a synchronized low-dimensional form of the original space. It also has certain limitations, such as the size of the covariance matrix being identical to the data dimension, and Eigenvalues and Eigenvectors being equal to the dimension of the data. As a result, computing eigenvalues accurately for bigger datasets becomes more complex. Another issue is that small variance components aren't as significant when using PCA, consequently, prediction accuracy might deteriorate. In the majority of instances, however, PCA improves productivity in unsupervised machine learning.
In the study, six features L, R, F, M, V have been extracted respectively. It will be really complex to analyze and portray a dataset in six dimensions.The Cumulative Explained Variance Ratio is being used to know how many dimensions are suitable for the dataset to represent and analyze the graphs.

Table 3.6: PCA on the dataset of the proposed model

	Dimension 1	**Dimension 2**
0	2.026663	0.113448
1	-0.688949	3.941569
2	2.452061	-0.404806
3	1.44086	1.647351
4	0.648339	-0.694921
...	...	...
17422	1.207285	-0.365144
17423	-1.652813	0.649655
17424	-1.412286	0.062498
17425	1.365681	-0.705877
17426	-0.811169	0.749766

PCA transformed the data frame from six to two dimensions. It increased the interpretability of the dataset along with reducing the number of dimensions without any information loss.

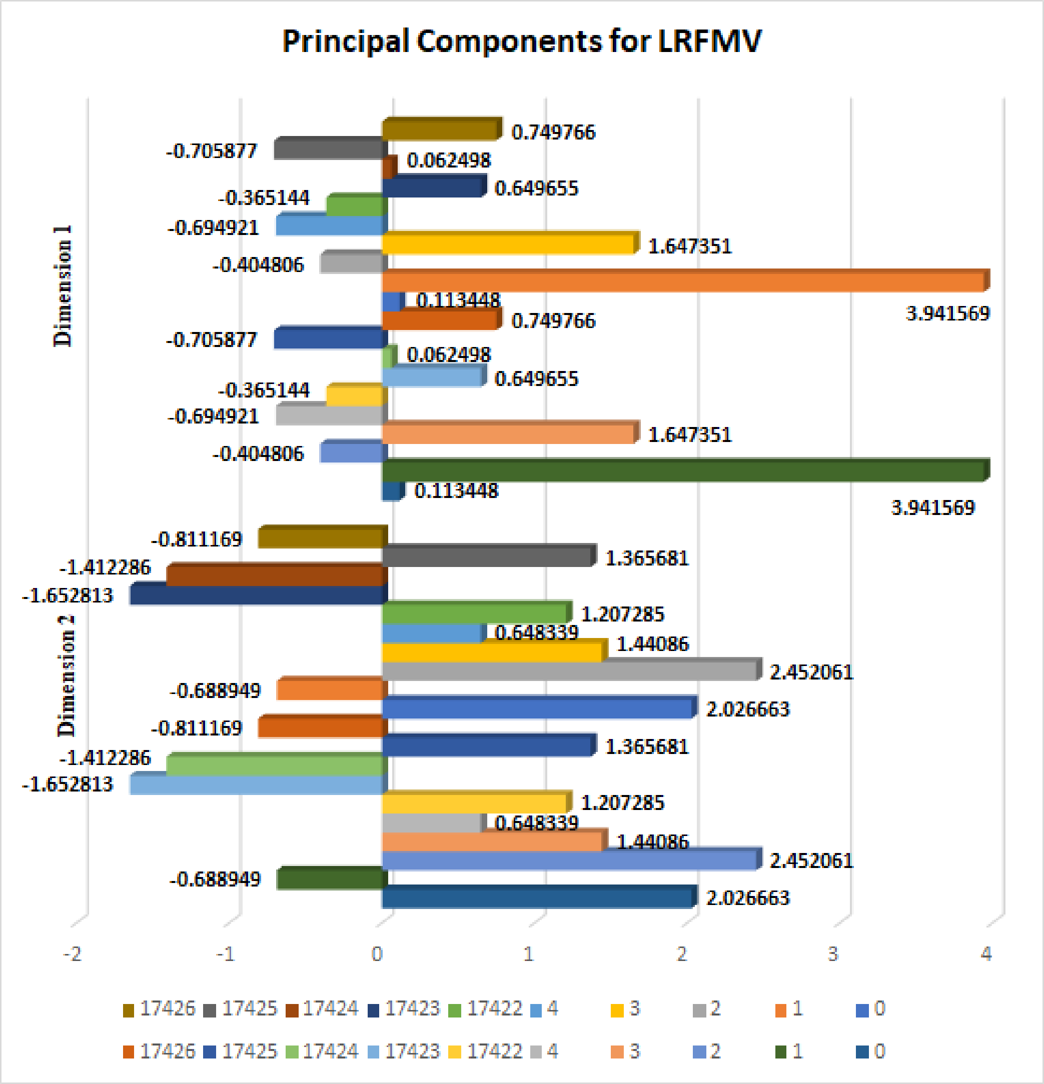


Figure 3.17: 3D representation of PCA on the dataset of the proposed model

The above mentioned figure is the 3D representation of PCA for the Table 3.6 on the dataset of the proposed LRFMV model. Here for both dimensions different colors has been used for different numbers in case of better visualization.

3.6.5 K-means Algorithm

For unsupervised machine learning, the k-means method is known as a popular clustering strategy for segmenting dividing a bunch of unprocessed sample points into distinct groups. To address the clustering issue, MacQueen proposed the k-means method in 1967 which was one of the simplest unsupervised learning algorithms at the time [57]. The algorithm takes a set of unlabeled data points and assigns them to one of k groups, with each cluster containing data of a similar sort. The optimal cluster numbers can be identified with the help of the Elbow method, Silhouette Coefficient method, gap statistic, etc. The process of clustering begins by determining the centroids and placing them in the unlabeled data which can be entirely random or actual data points from the real data set, X. After applying a defined distance metric to compare the distances, the data points with the shortest distance will be assigned together around the centroid and the centroid's value will be computed again by computing the mean value of each group's data points. The process will continue until the criterion function becomes minimum after numerous calculations. At the end of the procedure, each centroid will be placed in the center of its circular decision boundary known as clusters.

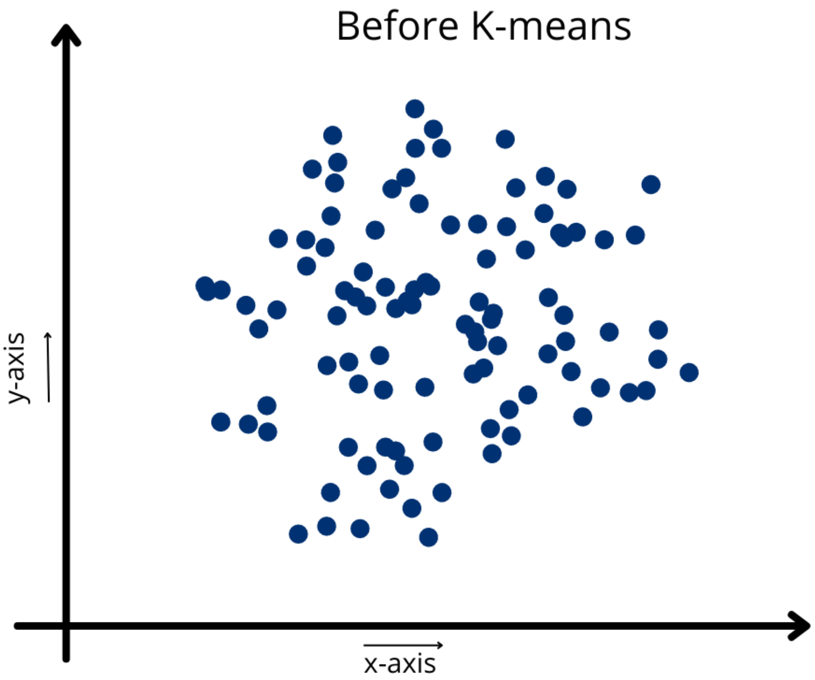


Figure 3.18: Unlabeled data before applying the K-means algorithm

Figure 3.18 indicates a the total number of points together with the x and y-axes for any dataset X. The data points labeled as blue have not yet been clustered and are spread throughout the graph.

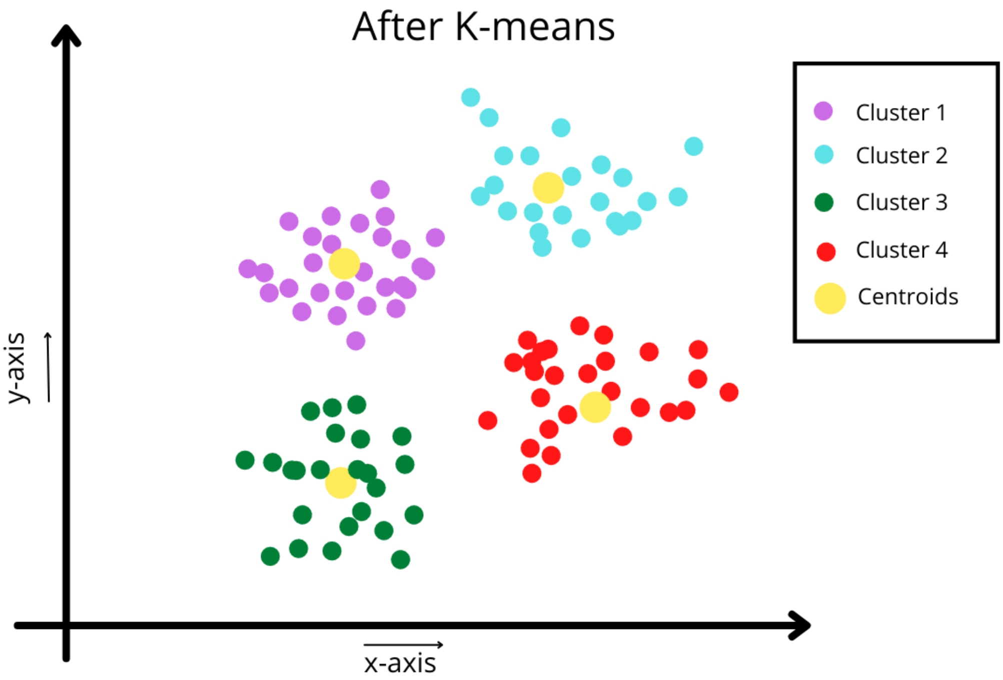


Figure 3.19: Labeled data after applying the K-means algorithm

In figure 3.18 , it is noticed that the unlabeled data points 'n' from figure 3.18 are labeled into 4 clusters with the help of K-means. To measure the distance between the taken points and each centroid several distance metrics, such as Chebyshev distance, Euclidean distance, Manhattan distance, Minkowski distance and so on, can be used to determine the minimal distance between them.Euclidean distance will be used as a distance metric here. Mathematically, for any given dataset: let X = $\{x_1, x_2, x_3 \ldots x_n \ \}$ x having n points, the Euclidean distance between the vectors $i = \{i_1, i_2, i_3, \ldots, i_n\}$ and $j = \{j_1, j_2, j_3, \ldots, j_n\}$ can be written as,

$$E_{ij} = \sqrt{\sum_{k=1}^{n} (i_k - j_k)^2}$$

The criterion function can be defined as below-

$$SSE = \sum_{i=1}^{k} \sum_{\alpha \in C_i} \|a - \alpha_i\|^2$$

Here, C_i denotes the cluster where α and α_i is the average of that cluster. SSE denotes the sum of squared errors for all points of data in the dataset. The distance measured for the mentioned criterion function is Euchdean distance [41]. Using the Euclidean Distance formula each data point's distance from the centroid is computed and utilizing K means the objects that are near to the centroids are placed next to each other The process continues until the next new centroid is found.

The process of creating new centroids is visible for each step until the criteria function becomes minimal in the above figure for each given dataset X = $\{x_1, x_2, x_3 \ldots x_n\}$ with n number of points. The distance between the n points and

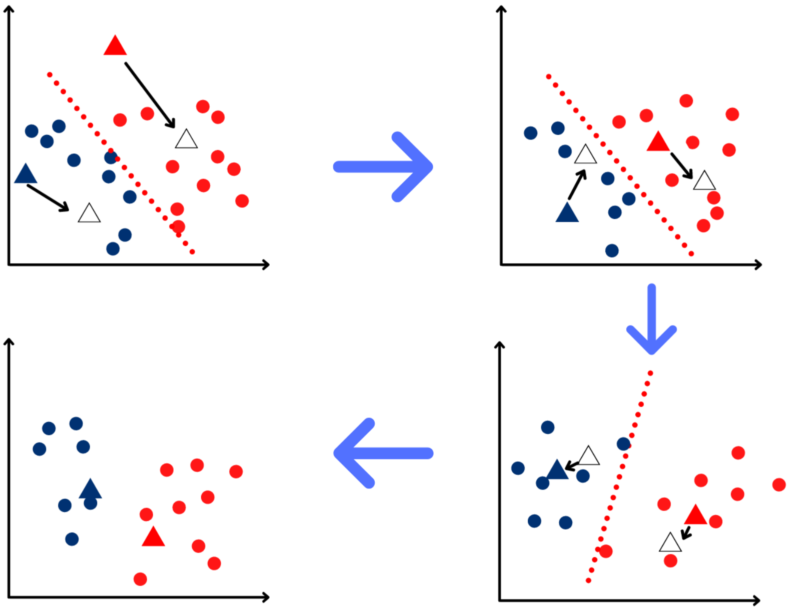


Figure 3.20: Graphical illustration of the K-means clustering process

the current centroids is calculated and the points that are closest to the centroids are gradually grouped together and create clusters. The K -means algorithm is simpler to implement and works well with large data sets. For unsupervised learning, converged data can be easily visualized within a short time using the algorithm. The k-means can be difficult to apply when there are a large number of variations in dense data sets. On the featured data, PCA can be used to reduce dimensions. Outliers affect the grouping of the data when using the K-means algorithm. Nonetheless, the K-means algorithm is an effective technique operates in commercial sectors to segment customer and product data and aid in the development of business models. In case of applying the LRFMV model, we employed the K-means clustering technique to generate efficient clusters and find the potential customers to earn more profit. Cluster number was derived by the Elbow and Silhouette method and clusters have been created by the K-means algorithm.

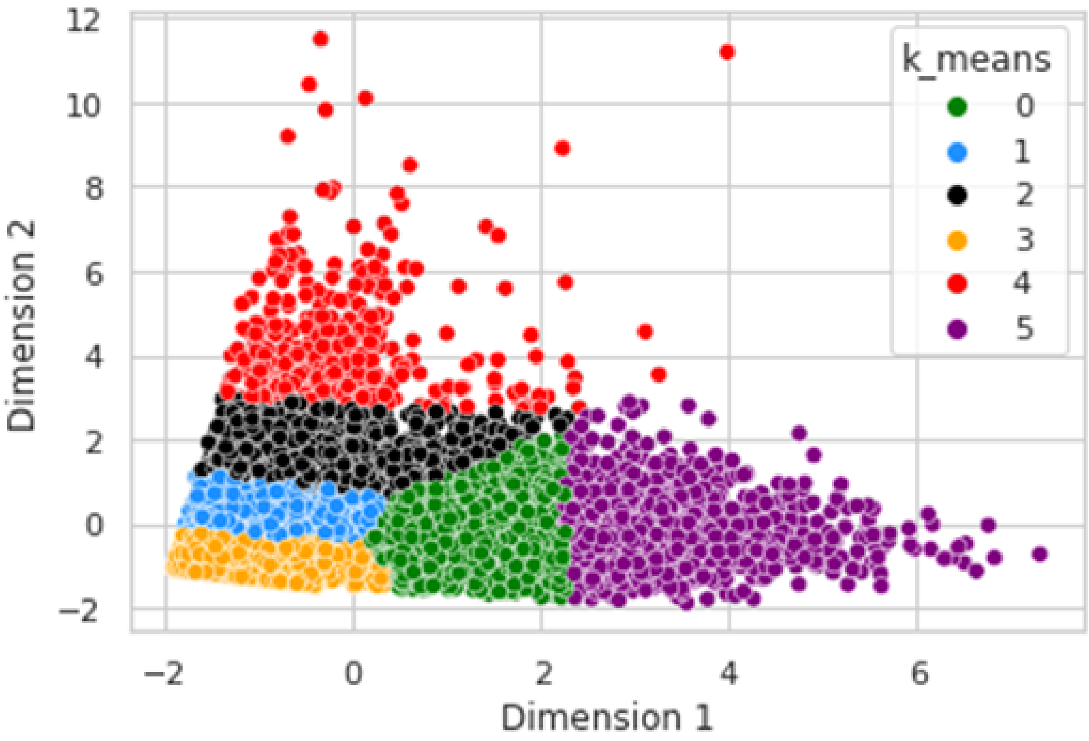


Figure 3.21: K-means clusters for LRFMV model

Following that, the k-means algorithm is applied to get clusters for the proposed LRFMV model and the above graph is projecting the clusters. In the graph,it can be observed that six different clusters have been shown with six different shades of colours in the graph where green is denoted for cluster 0 and purple is assigned for cluster 5. K-Means algorithm has been applied on the RFM and LRFM model also and further explanation and comparison prove that it is the best clustering technique for the proposed LRFMV model .

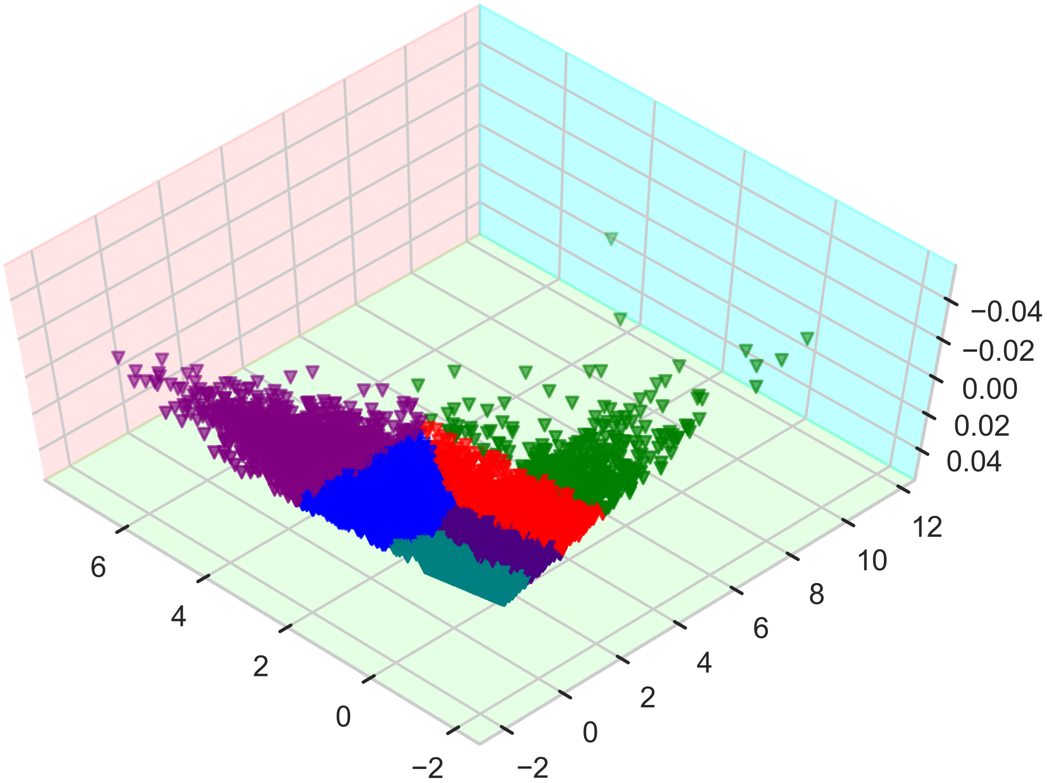


Figure 3.22: 3D representation of K-means clusters for LRFMV model

In the above figure, a three dimensional representation of K-means clusters for LRFMV model has been illustrated. The clusters are labeled with six different colors (teal, indigo, blue, green, purple and red) in order to visualize competently.

3.6.6 K-Medoids Algorithm

K-Medoids is a conventional clustering method that splits an n-object data collection into k a priori determined clusters. The silhouette is a useful tool for figuring out k. When compared to k-means, it is more noisy and outlier sensitive[24]. Actual objects can be used to represent clusters in the K-Medoids methodology, with one representative item per cluster, and a medoid is an object in a cluster with the minimum average dissimilarity to all other objects in the clusterThe other objects are grouped along with the most similar representative object. The partitioning method is based on the idea of reducing the sum of the differences between each object and the reference point with which it is connected.[24].

It chooses 'k' points at random from the data ('k' is the number of clusters to form). The validity of the choice of k's value has been tested using techniques such as the silhouette technique and the K-Means method. Each data point is given to the cluster that contains the closest medoid.The distance between each data point in clusters namely C1,C2,C3 and all other data points is computed and added.

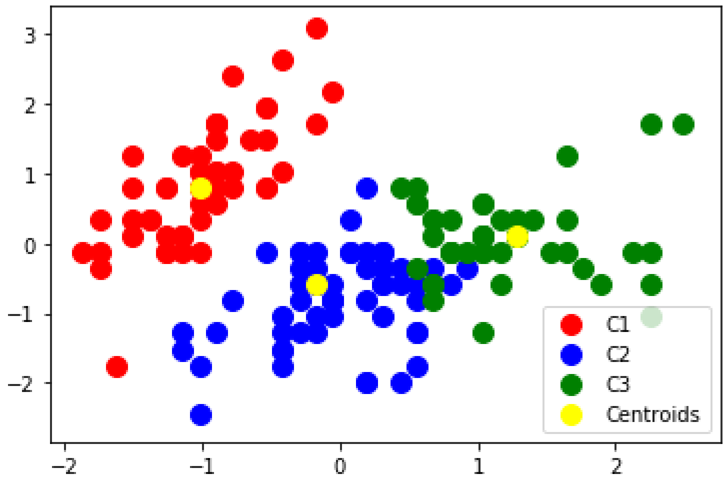


Figure 3.23: Graphical illustration of K-Medoids algorithm

K-Medoids is a clustering approach that is similar to the K-Means methodology. It belongs to the unsupervised machine learning category. In terms of how it determines the clusters' centers, it differs significantly from the K-Means algorithm[26].The former chooses the average of a cluster's points as its center (which may or may not be one of the data points), whereas the latter always chooses the clusters' actual data points as their centers (also known as medoids).

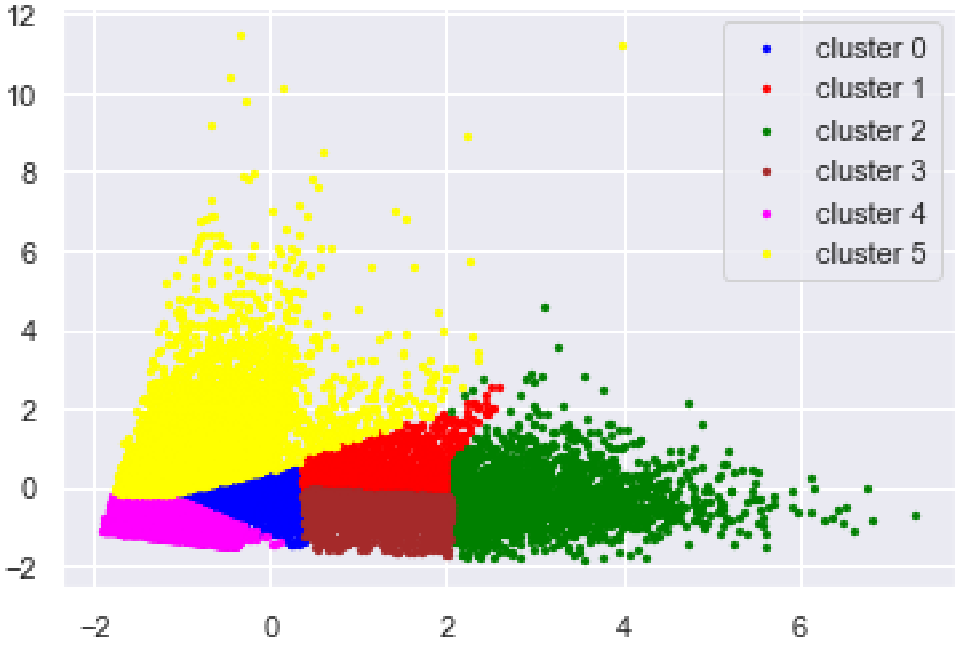


Figure 3.24: K-Medoids for LRFMV model

In figure: 3.24, it is noticeable that 6 clusters have been identified for the pro- posed the LRFMV model after applying the K-Medoids algorithm. Different colors like blue,green,red etc. have been assigned to highlight the clusters and K- Medoids approach has given more number of clusters than the standard K-Means algorithm. Moreover,K-Medoids has also been applied for RFM and LRFM model for further comparison analysis.

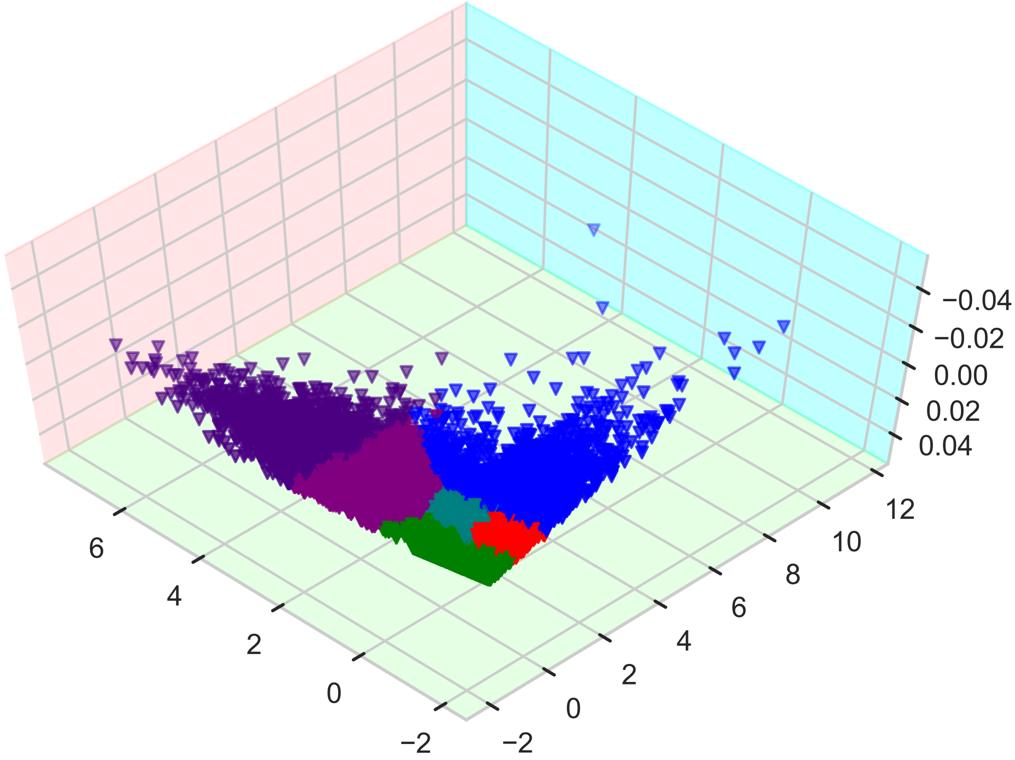


Figure 3.25: 3D representation of K-Medoids clusters for LRFMV model

A three-dimensional depiction of K-Medoids clusters for the LRFMV model is shown in the figure above. To facilitate visualization, the clusters are labeled with six distinct hues (teal, indigo, blue, green, purple, and red).

3.6.7 Mini Batch K-means Algorithm

For clustering huge datasets, Mini Batch K-means has been proposed as a replacement for the K-means algorithm[27]. By sampling a fixed-size subsample of the dataset rather than the entire dataset, this strategy saves time and money. The fundamental concept is to use tiny stochastic groups of adjusted data that may be kept in storage. Each repetition, a fresh random selection of the information is picked and used to refresh the clusters, and the process is repeated once convergence is reached [32]. As the number of iterations grows, each mini batch refreshes the clusters with a convex mix of prototype and example values and a falling learning rate. The number of examples assigned to a cluster is inversely proportional to the speed with which the procedure is completed.The influence of additional instances lessens as the number of iterations increases, and convergence can be detected when no changes in the clusters occur over a long period of time.

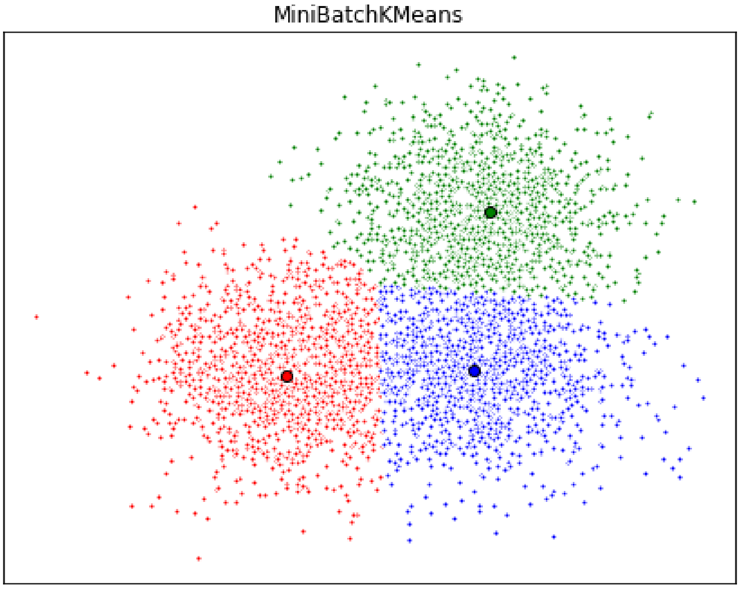


Figure 3.26: Implementation of Mini Batch K-means algorithm

In the above-mentioned figure, the Mini Batch K-means algorithm has been applied on 4000 random data points and the batch size was 200. It can be seen that 3 uniform clusters have been generated after a certain number of iterations.

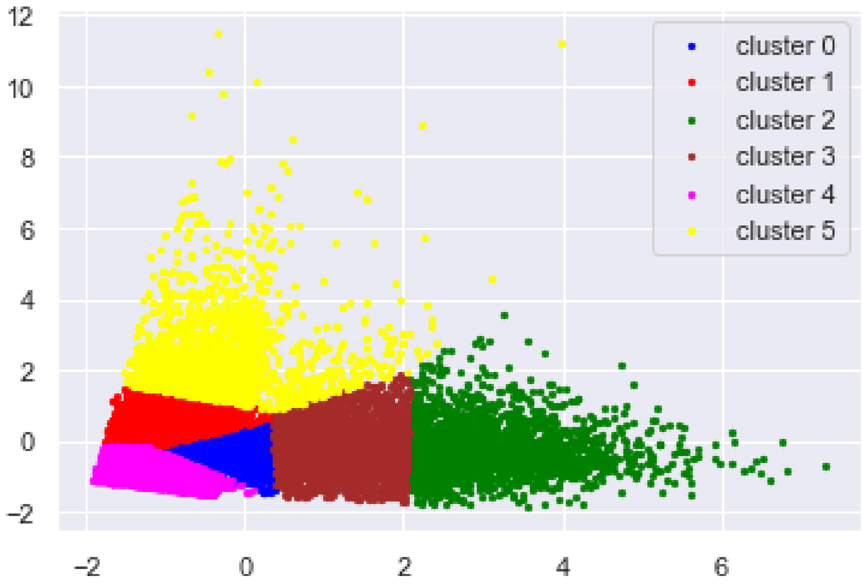


Figure 3.27: Applying Mini Batch K-means Algorithm on LRFMV model

In the figure 3.27, it can be seen that for the proposed LRFMV model 6 clusters have been formed after applying mini batch K-means which is similar to K-means algorithm where the density of clusters 0.1,3,4 is higher than that of clusters 2,5.Mini Batch has also been used for clustering in the RFM and LRFM model in order to find out the efficiency of the proposed LRFMV model.

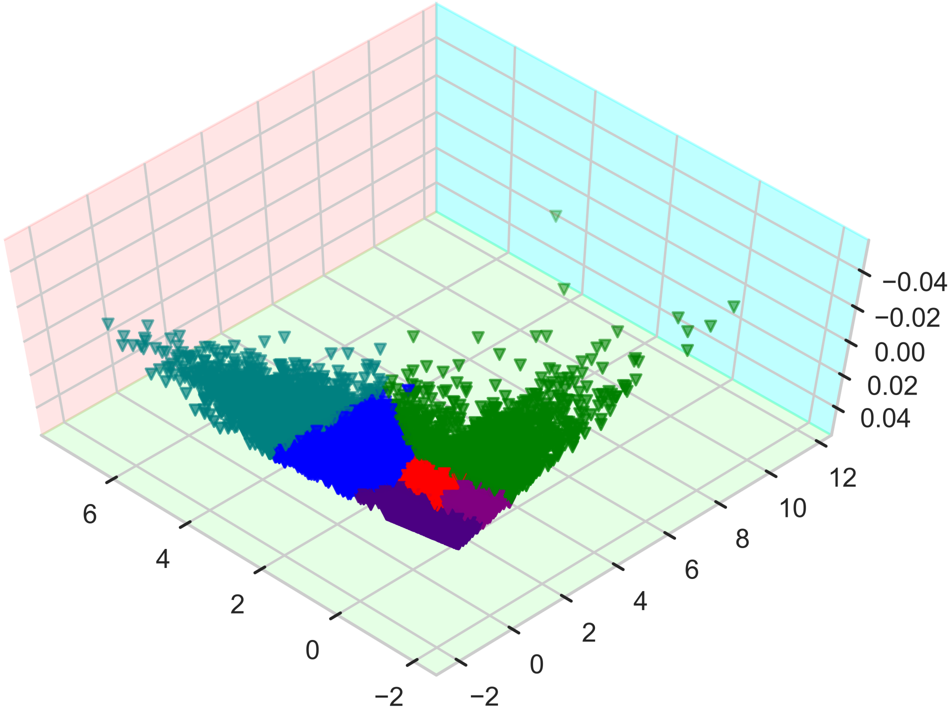


Figure 3.28: 3D representation of K-means clusters for LRFMV model

The graphic above exhibits a three-dimensional representation of Mini Batch K-means clusters for the LRFMV model. The clusters are labeled with six various colors to aid with viewing (teal, indigo, blue, green, purple, and red).

Chapter 4

Comparative Analysis

For any unsupervised learning, analysing the performance from the result by clustering techniques is an important issue. We considered two methodologies to compare the performance of RFM, LRFM and LRFMV. One of them is the number of clusters generated by the three models RFM, LRFM and LRFMV and another one is the clustering quality for each of them.

4.1 Number of clusters

4.1.1 Number of cluster determination for KMeans algorithm

To compute the optimal number of clusters for the RFM, LRFM and LRFMV model the Elbow Method and the Silhouette Coefficient is applied. It's important to know the total number of clusters before applying K-means so that the data doesn't overfit in case of high value for K and important features are not excluded in case of lower value for k.

In case of RFM model with increasing the value of k the graph gradually becomes flatten When k exceeds 4, the graph flattens out and the SSE values become insignificant. As a result, the bend in the above Elbow curve can be calculated as $k = 4$, which is also the RFM model's optimal number of clusters for our dataset. Using the Elbow method,The four clusters have been discovered for the RFM model for the data set.

Now, the cluster numbers for the RFM model for the dataset using the Silhouette score is calculated. Here, the Silhouette Value for k gradually increases till the value of $k = 4$. When the value of k exceeds 4, the Silhouette drastically becomes less (0.4674035614222174 when $k = 5$, 0.43326621579270996 when $k = 6$, 0.3981119827887891 when $k = 7$, ...and so on). Therefore, $k = 4$ is the cluster numbers when the corresponding Silhouette score is 0.471724277212437.

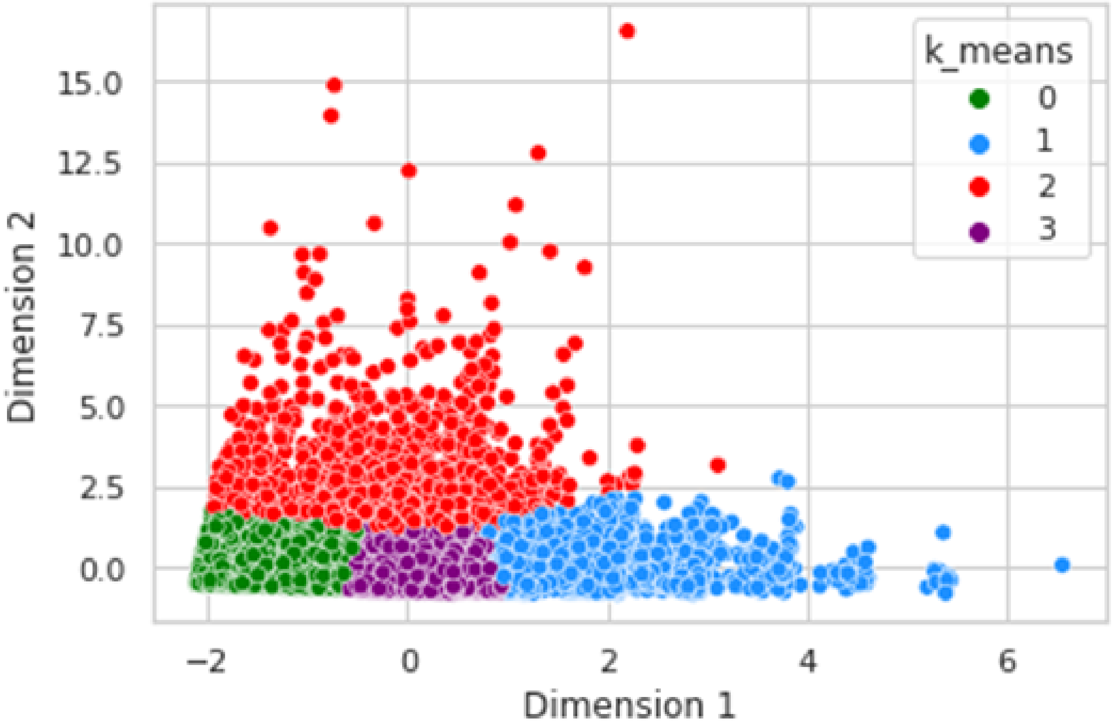


Figure 4.1: Applying K-means Algorithm on RFM model

Figure 4.3 demonstrates the usage of the k-means algorithm to construct four clusters for the RFM model. For the k-means clusters of RFM model four different shades have been used to represent the clusters (cluster 0: green, cluster 1: blue, cluster 2: red and cluster 3: purple).

In case of LRFM model, the ideal cluster numbers are determined using both the Elbow Method and the Silhoutte Method for K-means algorithm. As the value of k is increased, the graph flattens out. When k is greater than 5, the graph flattens and the SSE values become unimportant. As a consequence, the bend in the above Elbow curve can be computed as $k = 5$, which is exactly the ideal number of clusters for our dataset according to the LRFM model.

Besides Elbow method, the cluster numbers for the LRFM model for the dataset are also derived using the Silhouette score. In this case, the Silhouette Value for k steadily grows until it reaches $k = 5$ and the Silhouette Value is 0.4210816270834872. When the value of k surpasses 5, the Silhouette gets significantly smaller (0.41775957732139857 when $k = 7$, 0.402157582846443 when $k = 8$, 0.35841125721368255 when $k = 9$ and so on). As a result, when the matching Silhouette score for LRFM model is 0.4210816270834872, the cluster numbers are $k = 5$.

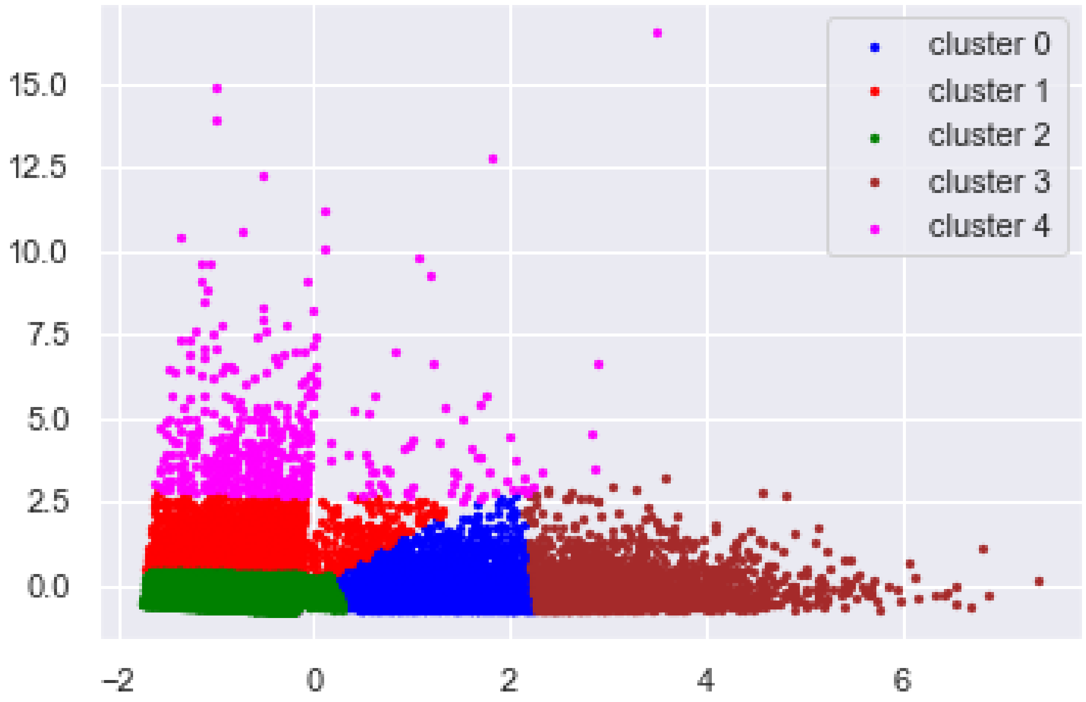


Figure 4.2: Applying K-means Algorithm on LRFM model

From the above figure it is visible that 5 clusters are formed from LRFM model by applying K-means algorithm. The clusters are labeled with five different colours in order to visualize properly (cluster 0: blue, cluster 1: red, cluster 2: green, cluster 3: brown, cluster 4: magenta).

From figure 3.15 it is observed that the clusters are indicated on the horizontal axis, and the SSE values for the clusters are indicated on the vertical axis. Here after exceeding a certain value for k the curve becomes smooth. In this scenario, the value for k is 6. When the value of k exceeds 6 the change of SSE corresponding to k gets smaller. As a result, it creates an elbow point at k = 6 and that is the ideal number of clusters found for the dataset.

To measure the ideal number of the clusters more precisely the Silhouette score is used to determine the cluster number. It is seen from figure 3.17 that the Silhouette Value for k steadily rises until it reaches k = 6. When the value of k crosses 6, it significantly decreases (0.4185709369336526 when k = 7, 0.3614374374627813 when k = 8, 0.36872579425202234 when k = 9, ...and so on). When the value k = 6 , the Silhouette score is 0.4224064188058843, so the number of clusters that is detected is 6. As a consequence, four clusters are formed in the RFM model, five clusters are formed in LRFM model and six clusters in the LRFMV model using both the Elbow technique and the Silhouette coefficient for K-means algorithm.

In case of k-means clusters of LRFMV model six different shades have been used in the figure 3.26 to represent the clusters (cluster 0: green, cluster 1: blue, cluster 2: black and cluster 3: yellow, cluster 4: red and cluster 5: purple). That implies the LRFMV model develops more segmentation for the purchase behavior of the customers, which is highly essential for identifying potential customers. The correct choice to develop the superstore business can be determined by analyzing the 6 clusters of the LRFMV model.

4.1.2 Number of cluster determination for K-Medoids algorithm

In the K-Medoids algorithm, a medoid is the cluster's most central element, with the shortest distance between it and other points. Because medoids are unaffected by extremities, The K-Medoids algorithm is less susceptible to outliers and noise than the K-Means method. The mean of the data points is a metric that is heavily influenced by extreme points. If the data contains outliers, the centroid in the K-Means algorithm may be pushed to an inaccurate position, resulting in erroneous grouping. Both Elbow method and silhouette score is used in case of K-Medoids algorithm for RFM model to determine the optimal cluster numbers. In both cases just like K-means algorithm the four clusters are produced for RFM model.

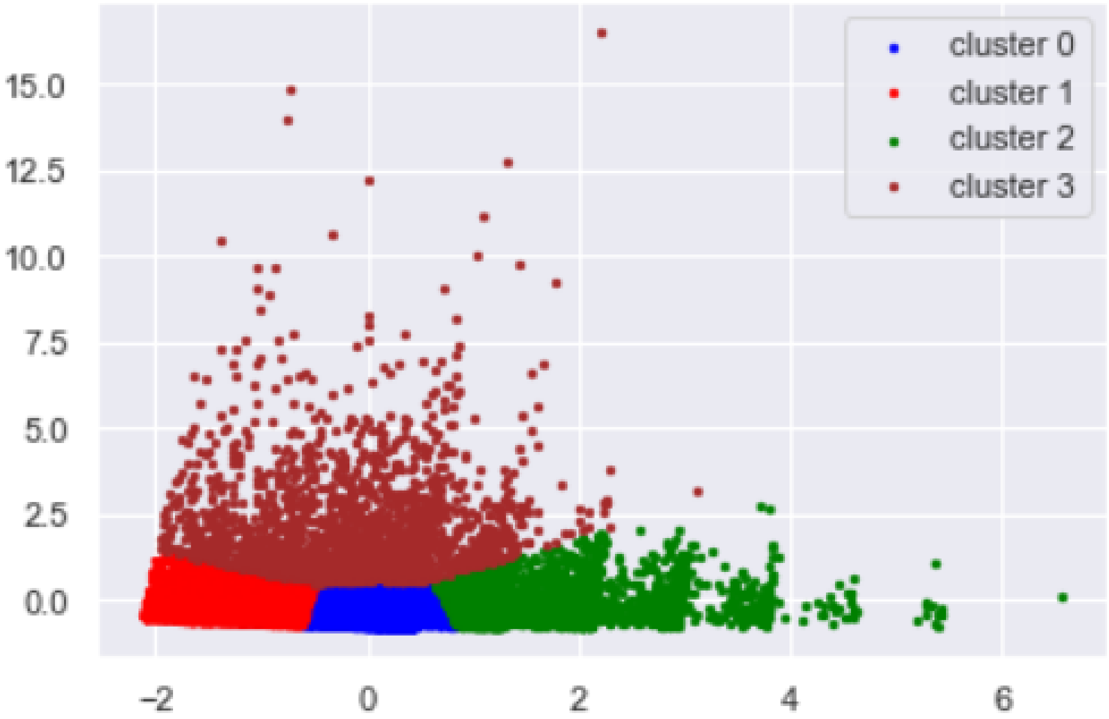


Figure 4.3: Applying K-Medoids algorithm on RFM model

In the above mentioned figure, K-Medoids approach has been applied on RFM model. Four clusters have been identified here and four different colors are assigned to mark them individually (cluster 0: blue, cluster 1: red, cluster 2: green and cluster 3: brown).

For LRFM model, applying K-Medoids algorithm optimal cluster numbers are determined by using both elbow method and silhoutte coefficients just like for K-means algorithm. Hence, the optimal cluster number for LRFM model using K-medoids can be determined as five.

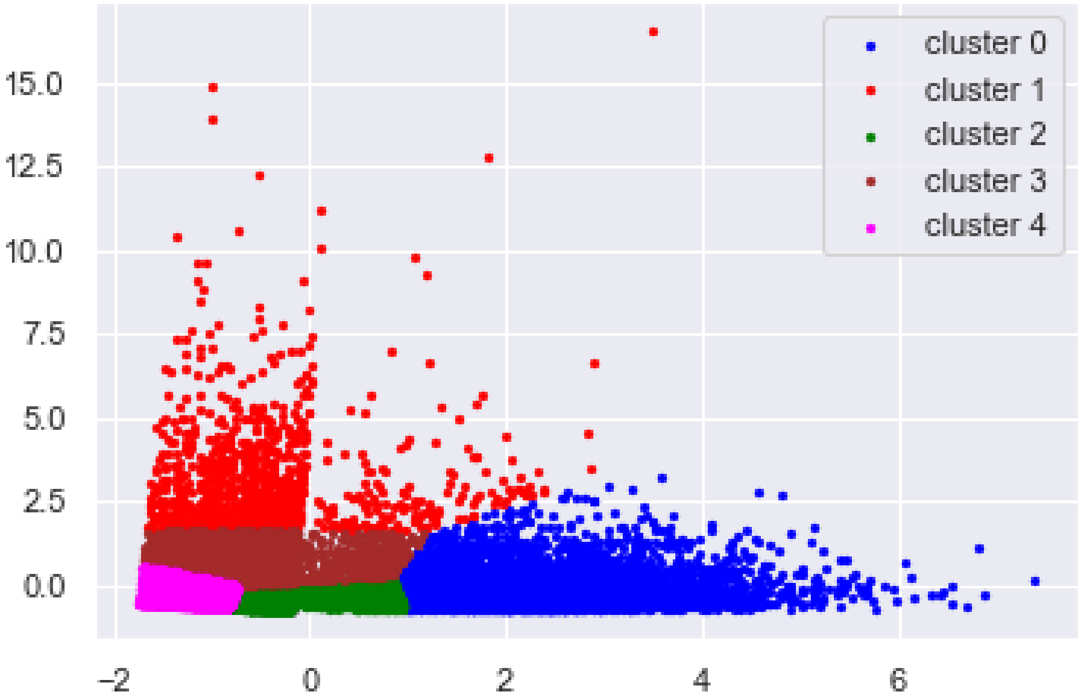


Figure 4.4: Applying K-Medoids algorithm on LRFM model

In the above figure 5 clusters are formed by LRFM model applying K-Medoids algorithm. Each clusters are labeled with five different colors to visualize in a better way (cluster 0: blue, cluster 1: red, cluster 2: green, cluster 3: brown and cluster 4: magenta). Here cluster 1 and 0 is a bit scattered where cluster 2, 3 and 4 are compact.

It is clearly visible that cluster number has been reduced for both RFM and LRFM models compared to the fig 3.23. From LRFMV model, six clusters are formed using K-Mediods algorithm. Six different shades have been used to represent the clusters (cluster 0: blue, cluster 1: red, cluster 2: green, cluster 3: brown, cluster 4: magenta and cluster 5: yellow). Hence, it can be said that the proposed LRFMV model generates clusters for K-Mediods algorithm more efficiently.

4.1.3 Number of cluster determination for Mini Batch K-means Algorithm

K-means is one of the most often used clustering algorithms due to its high efficiency. Due to the fact that K-means needs the complete dataset to be kept in main memory, its calculation time scales with the size of the datasets under examination. As a result, numerous strategies for reducing the algorithm's time and spatial cost have been presented. So, MiniBatch has been proposed to generate good quality clusters for segmentation. For RFM model, 4 clusters have been found using both elbow method and silhoutte coefficients while applying Mini Batch K-means algorithm.

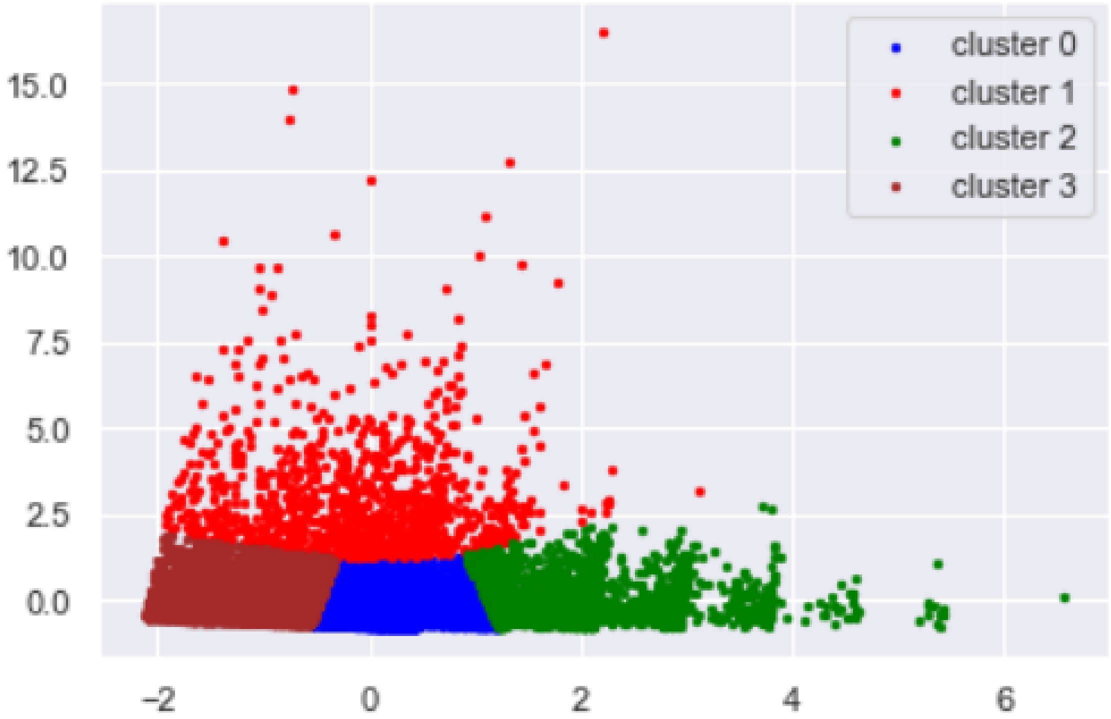


Figure 4.5: Applying Mini Batch K-means algorithm on RFM model

While applying the Mini Batch K-means algorithm in the RFM model, 4 clusters are found where cluster 0 and cluster 3 are compact and cluster 1 and 2 are a bit scattered. Four colors are used to label the clusters for better visualizing (cluster 0: blue, cluster 1: red, cluster 2: green, cluster 3: brown).

In case of LRFM model, five clusters have been determined using both elbow method and silhoutte coefficients by applying Mini Batch K-means algorithm. It is observable that Mini Batch K-means algorithm is a modified algorithm of standard k-means algorithm and produces same number of clusters in most of the cases.

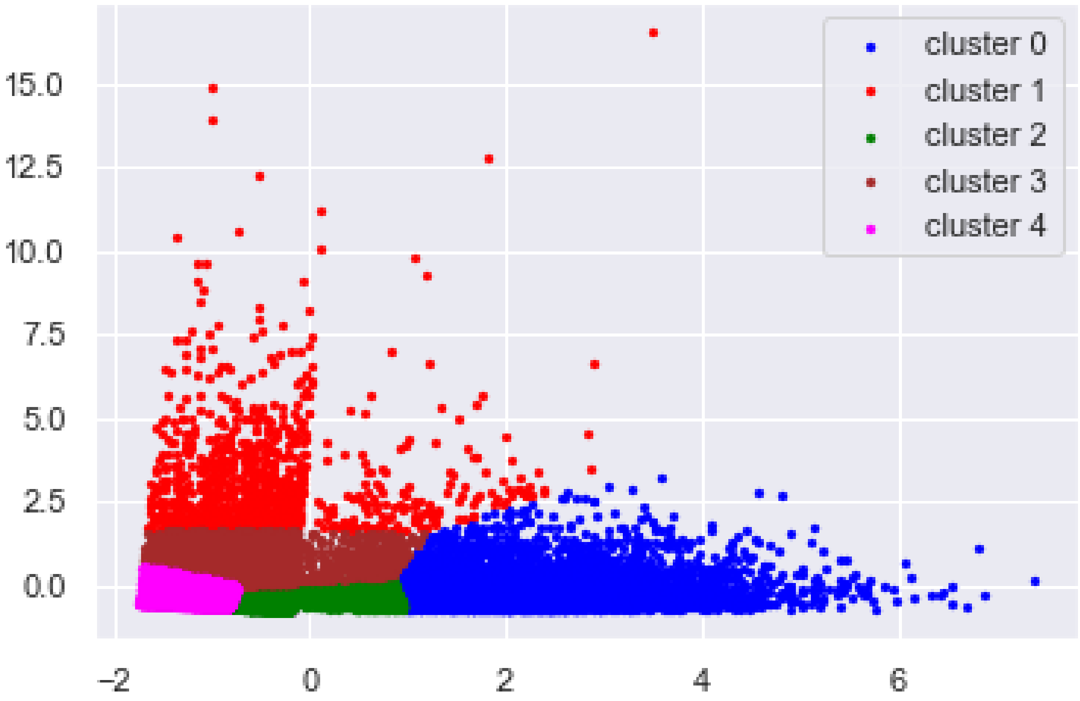


Figure 4.6: Applying Mini Batch K-means algorithm on LRFM model

In the above figure, Mini Batch K-means algorithm on LRFM model has been applied and five different clusters are formed. Here cluster 1 and cluster 0 are a bit scattered where other clusters are compact. Five different colors has been used here to visualize the clusters in an efficient manner (cluster 0: blue, cluster 1: red, cluster 2: green, cluster 3: brown and cluster 4: magenta).

On the other hand, in the LRFMV model by applying Mini Batch K-means algorithm, six clusters have been found which are visible in figure 3.25. Here six different colors have been used to label the clusters for better visualization (cluster 0: blue, cluster 1: red, cluster 2: green, cluster 3: brown, cluster 4: magenta and cluster 5: yellow).

4.2 Clustering Quality

4.2.1 Profit analysis for K-Means Algorithm

The quality of clustering refers to how closely the clusters fit the original data. The cluster fitness is vital because valued customers are not identified effectively by clustering if the resulting clusters do not extract the relevant attributes even after segmentation. Profit per head has been evaluated for clusters formed by the RFM, LRFM and LRFMV models to see how well they worked.

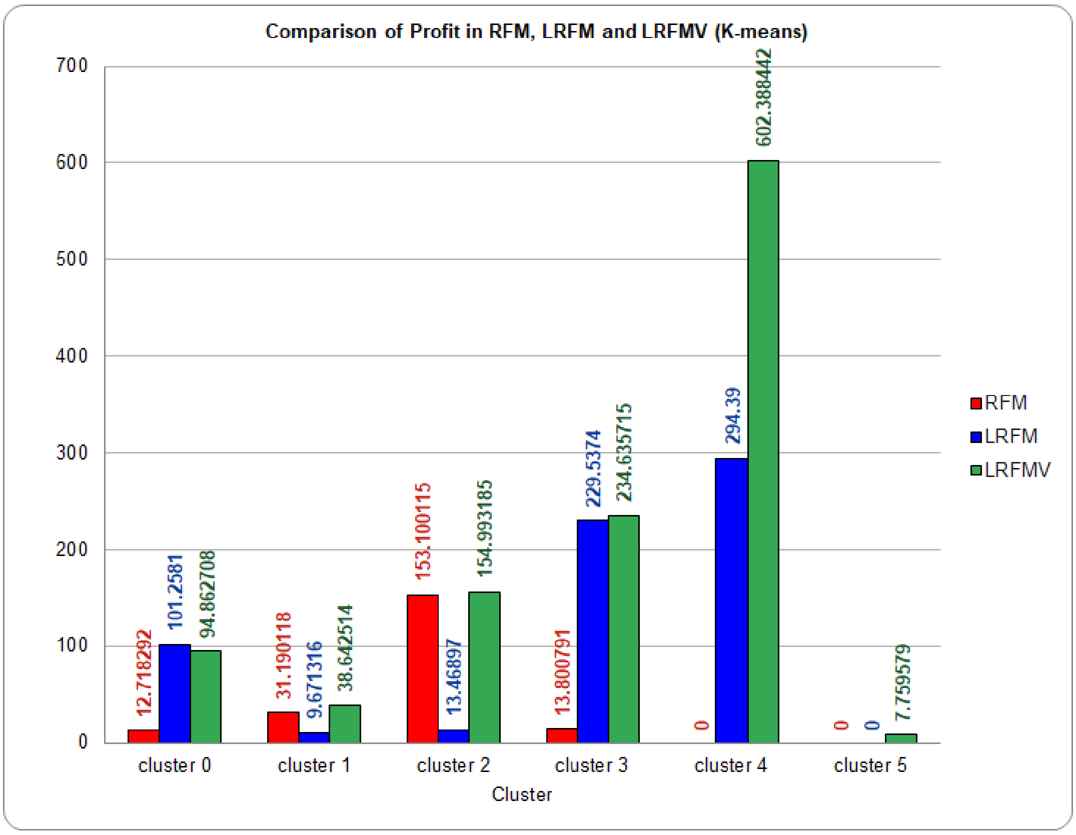


Figure 4.7: Profit analysis for RFM, LRFM and LRFMV model using standard K-means algorithm

The above graphs demonstrates different profit gains for RFM, LRFM and LRFMV models by employing K-means algorithm. Three colors are used for representing the profits for three models. The "red" colored bar indicates RFM model, "green" colored bar demonstrates LRFM model, "blue" colored bar is designated as the proposed LRFMV model. Now, in the red colored bar for the RFM model, it shows that there are four groups, each with its own per-head total revenue to the superstore. The highest profit is achieved from cluster 3, which is about 229.5734, while the lowest profit is achieved from cluster 0, which is about 12.718292. It implies that customers from cluster 3 are more essential in terms of creating more profit for the superstore.

Following that, the LRFM model, which is represented by the 5 blue colored bars denoting the 5 clusters, is perceived applying the K-means algorithm. The maximum profit is obtained from cluster 4, which is around 294.39, and the lowest profit is obtained from cluster 1, which is around 9.671316. That seems to be, if the LRFM model employs the K-means algorithm, the customers from cluster 4 generate higher profit for superstores.

After that, in the graph of the LRFMV model, it is clearly visible that the LRFMV model can develop 6 segments with the same number of customers while providing a larger profit per head. The highest profit is achieved from cluster 4 with 602.38842 profit per head. The lowest profit is achieved from cluster 5 that is 7.759579 per head. It implies that customers from cluster 4 are more essential in terms of creating more profit for the superstore.

By comparing three models with respect to profit per head from the clusters, it is observed that there is more variance of clusters in the LRFMV model than both of the RFM model and LRFM model. The same number of customers can be analyzed more efficiently with the LRFMV model with more variations. Once the three models are compared in terms of profit per head from clusters, it is visible that the LRFMV model has a greater variation of clusters than the RFM model and LRFM model. With additional variations, the LRFMV model can evaluate the same number of customers more efficiently. The LRFMV model produces a cluster with a profit of 602.388442, which is more than the RFM model's maximum profit of 153.100115 and LRFM model's maximum profit of 294.39. As a result, in the case of RFM model and LRFM model, a key cluster is overlooked that provides the greatest results for locating valuable customers and constructing marketing strategies correspondingly.

Furthermore, a detailed examination of both graphs indicates that the profit for the LRFMV model goes from high to moderate to low. For the LRFMV model, cluster 4 has a high profit per head, clusters 2 and 3 have a moderate profit per head, cluster 0 has a lower medium profit per head, cluster 1 has a lower profit per head and cluster 5 has a very low profit per head. However, under the RFM model, profit becomes more polarized, that is, it becomes predominantly either high or poor. Cluster 2 gives a larger profit per head for the RFM model, whereas Cluster 1 gives a medium profit per head and the other clusters give a lower profit per head. Furthermore, the RFM model in superstores does not accommodate all sorts of customers. Now, in the case of the LRFM model, the variances are better than in the RFM model, but it still has certain shortcomings. Cluster 1 and 2 give low profit for LRFM model where cluster 0 gives lower medium profit and cluster 3 and 4 gives larger profits. As a result, in many ways, the clustering quality of the LRFMV model is preferable to that of the standard RFM model and modified LRFM model.

4.2.2 Profit analysis for Mini Batch K-Means Algorithm

To decide the optimal number of clusters and analyze the clustering quality is very influential for segmenting customers. Silhouette coefficient method has been used to decide the cluster number for Mini Batch K-Means algorithm in the RFM and LRFMV model. However, Elbow method has been used to determine the cluster numbers for the LRFM model. Mini Batch K-Means process has been implied on all three models known as RFM,LRFM and LRFMV model. Profits of these three models for different clusters can be analyzed from below mentioned bar chart.

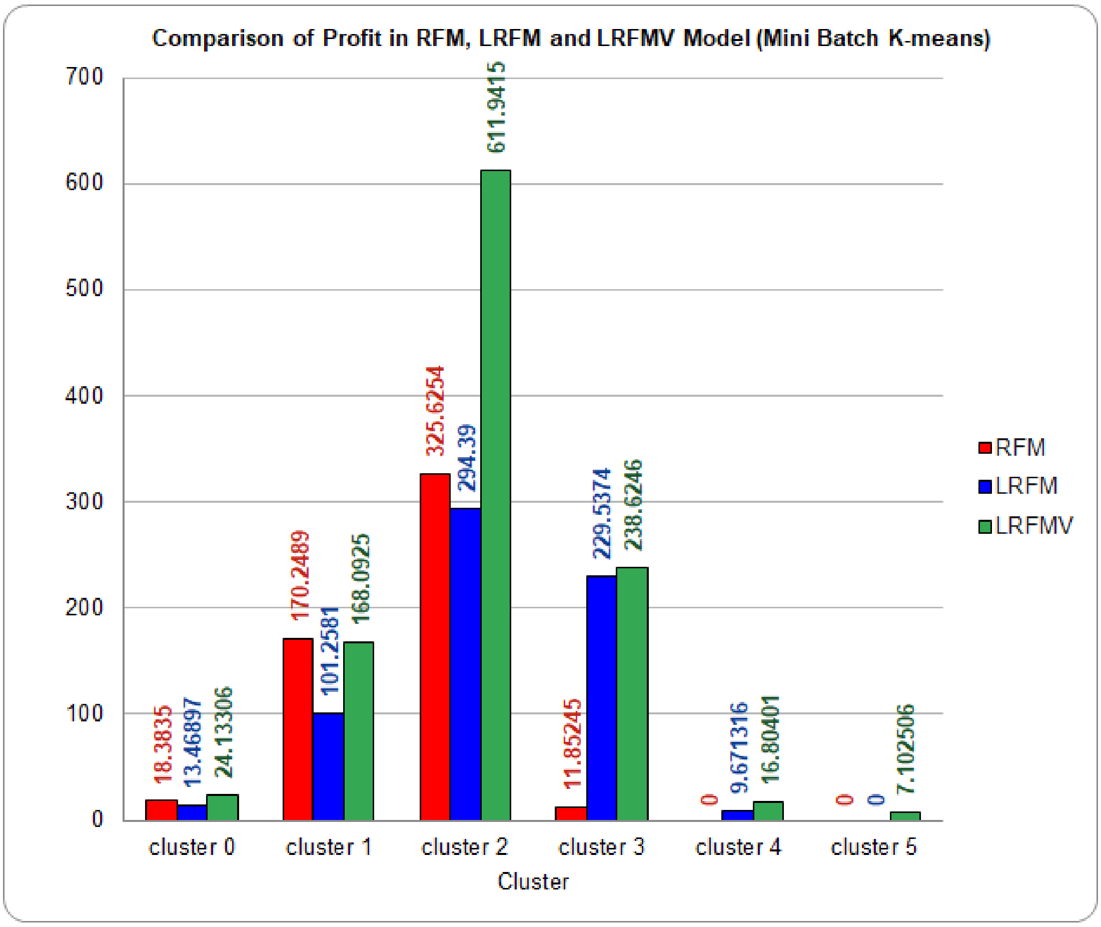


Figure 4.8: Profit analysis for RFM, LRFM and LRFMV model using Mini Batch K-means algorithm

Using the Mini Batch K-means algorithm, the above graphs show distinct profit gains for RFM, LRFM, and LFM models. Three different colors are utilized to symbolize the profits for three different models. The "red" colored bar represents the RFM model, the "green" colored bar represents the LRFM model, and the "blue" colored bar represents the proposed LRFMV model.For the RFM model, 4 clusters have been decided to plot by Silhouette coefficient. Cluster 2 is the largest cluster and gained profit is 325.6254 which is larger than any other profit for RFM model.The smallest one is cluster 3 and earned profit is 11.85245. It is discernible that the difference between cluster size and earned profit is huge and polarity is less in this regard.

For the LRFM model, cluster 3 is the most profitable cluster which is giving a profit of 238.6246 and cluster 4 is the lowest profit making cluster here. Though LRFM model is giving the highest profit in cluster three but this model is giving only 5 clusters and overall profit of this model is less than the LRFMV model.

6 clusters have been found for the LRFMV model and the difference between each cluster is not that large like the RFM model. As the cluster number is larger here,clustering quality and behaviour of particles in each cluster is more similar. Clustered between the largest and smallest one can be labeled easily here. The largest cluster is cluster 2 and earned profit is 611.9415 and the smallest one is

cluster 5 where the profit is 7.102506. The polarity is greater here for less difference among clusters. Comparing the profits of three different models in the above bar chart, it can be concluded that the clusters of LRFMV model are making the highest profit which are denoted with green bars.

4.2.3 Profit analysis using K-Medoids algorithm

The profit for each cluster in the RFM, LRFM and LRFMV models is calculated using the K-Medoids method. The RFM, LRFM and LRFMV models based on their findings are compared by plotting a barchart.

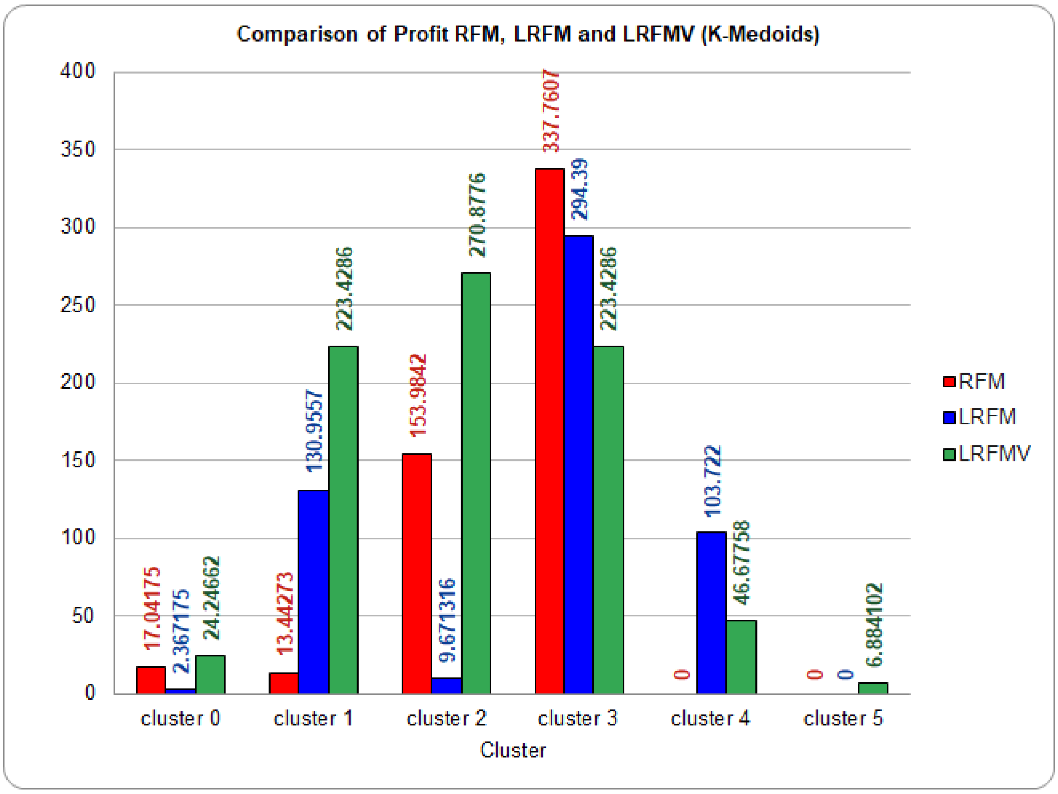


Figure 4.9: Profit analysis for RFM, LRFM and LRFMV model using K-Medoids algorithm

The graphs above show distinct profit gains for RFM, LRFM, and LRFMV models using the K-means algorithm. Three colors are utilized to symbolize the profits for three different models. The ”red” colored bar indicates the RFM model, the ”green” colored bar represents the LRFM model, and the ”blue” colored bar reflects the proposed LRFMV model. It is evident that for RFM model, four clusters have formed, each with their own per-head total income for the superstore. The highest profit is achieved from cluster 3 which is 337.760694 and the lowest profit is gained from cluster 1 that is 13.44273. It means that cluster 3 is the most important group for the superstore data to increase profitability.

In case of LRFM model indicating ”blue” bar, the highest profit is achieved for cluster 3 that is 294.39 and the lowest profit is gained from cluster 0 that is 2.367175. That implies that customers from cluster 3 appear to earn larger profits for superstores if the LRFM model applies the K-means algorithm.

On the other hand, the above figure illustrates the profit per head for the LRFMV model for each cluster adopting K-Medoids. Here cluster 2 generates the highest profit that is 270.877579 and the lowest profit is achieved from cluster 5 that is 6.884102. This suggests that cluster 5 is the most critical group for the superstore data to boost profitability. By comparing the RFM, LRFM and LRFMV models, we see that the highest number of profit can be gained using the RFM model, which is 337.760694 and the highest profit from LRFM model is 294.39. Inspite of giving the highest profit for the RFM model, the second highest profit that is cluster 3 of LRFMV model gives much higher profit than the second highest profit for RFM model and LRFM model. Besides, cluster 0,1 and 2 of LRFMV model yields better profit than cluster 0 and 1 of RFM model and cluster 0. By analyzing each cluster's profit a polarity is seen on the RFM model and LRFM model. For the RFM model cluster 1 and 0 are extremely lower than the other two clusters, for the LRFM model cluster 0 and cluster 2 are extremely lower but for LRFMV model more variations can be generated. For the LRFMV model, Clusters 1, 2 and 3 yield high profit, cluster 4 yields medium profit, cluster 0 yields lower medium profit and cluster 5 yields very low profit. As a result, the cluster quality for the LRFMV model employing the K-Medoids method is undeniably better than that of the RFM model and LRFM model.

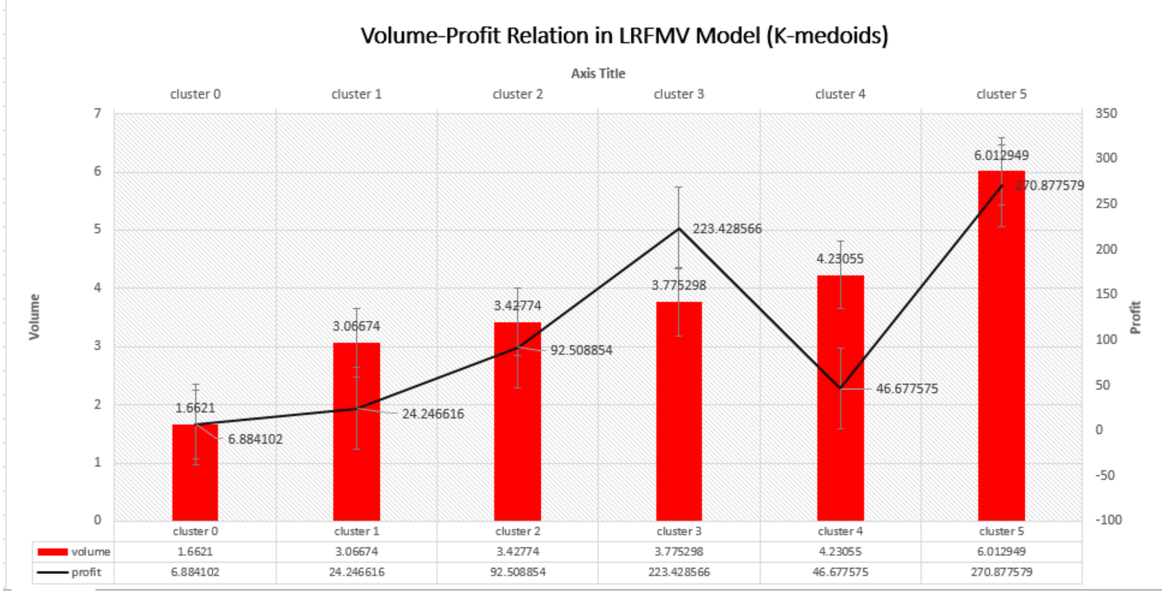
Chapter 5

Result Analysis

5.1 Statistical Analysis

5.1.1 Volume-Profit Relationship for K-Medoids Algorithm

Volume profit analysis calculates the impact of changes in a company's sales volume and price on profit. This is a very effective approach in managerial finance for gaining a better knowledge of profit growth. It is one of the most frequently used management accounting tools to help administrators to make futuristic decisions. In order to understand the volume-profit relationship for K-Medoids algorithm on the proposed LRFMV model,a barchart has been plotted in this section.It is observable that cluster 0 is giving the lowest profit which is 6.8841 where the volume of sold product is 1.6621. On the other hand,cluster 5 is giving the highest amount of profit which is 270.8775 percent and the volume of sold product is 6.0129. A proportional relationship between volume and profit is noticeable for this approach as well.



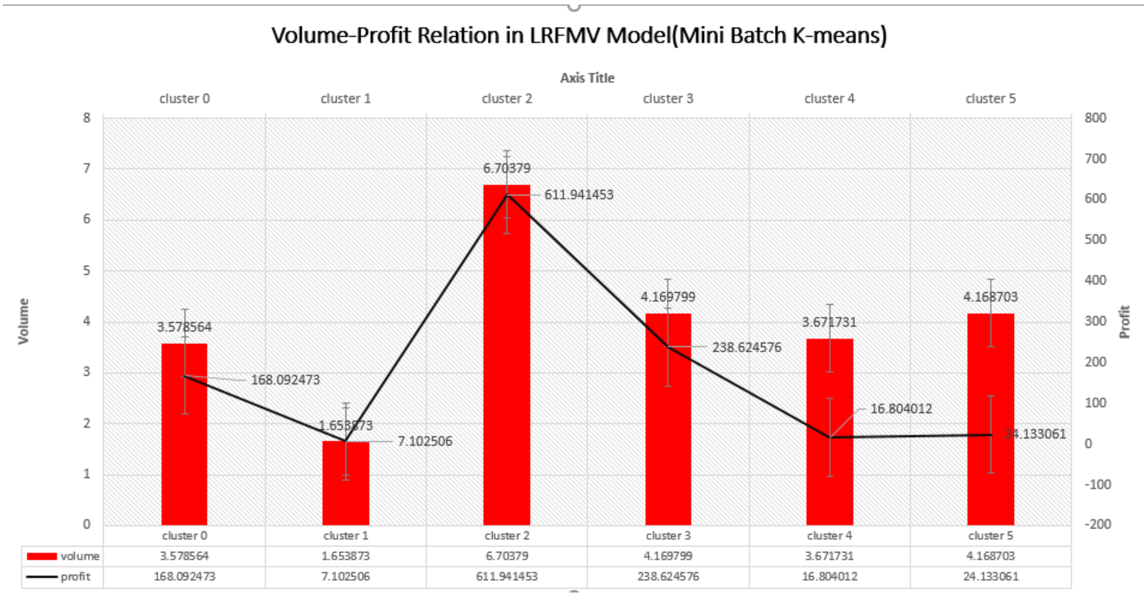
	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
🟥 volume	1.6621	3.06674	3.42774	3.775298	4.23055	6.012949
— profit	6.884102	24.246616	92.508854	223.428566	46.677575	270.877579

Figure 5.1: Volume-Profit Relationship of LRFMV model for K-Medoids Algorithm

Along with the proportionality, it is also visible that the relationship between volume and profit is atypical for cluster 3.Because the profit in cluster 3 is higher than the volume, the K-Means technique has been picked for the provided dataset in this research study. Moreover, K-Medoids is more appropriate for a dynamic dataset which has a prominent possibility to change the values frequently. The proposed dataset is lethargic and K-Means is more suitable for this research.

5.1.2 Volume-Profit correlation for Mini Batch K-Means Algorithm

Many corporate management decisions require a volume-profit analysis. This analysis entails developing a model of the relationships between product prices, volume or degree of activity, and sales mix. This idea is used to forecast the impact of changes in such characteristics on earnings. It is evident that a proportional relationship is also present in the below mentioned bar chart for Mini Batch K-Means approach. Highest profit has been earned from cluster 2 which is 611.9414 and the volume is 6.7037. The lowest profit has been drawn from cluster 1. The profit of cluster 1 is 7.1025 where the volume is 1.6538.



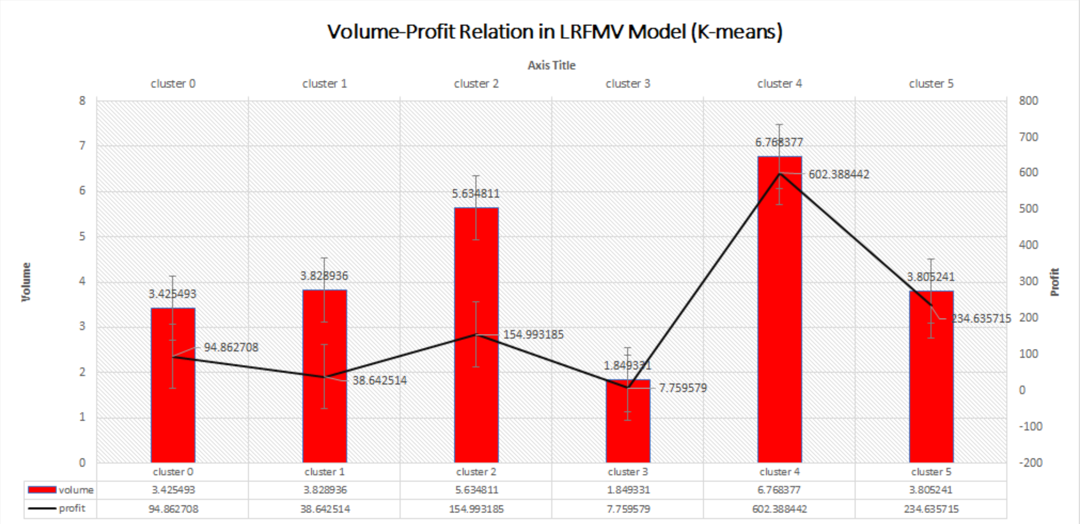
	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
🟥 volume	3.578564	1.653873	6.70379	4.169799	3.671731	4.168703
— profit	168.092473	7.102506	611.941453	238.624576	16.804012	24.133061

Figure 5.2: Volume-Profit Relationship of LRFMV model for Mini Batch K-Means Algorithm

Mini Batch is more like a subsection of K-Means Algorithm. For the proposed dataset this algorithm also gives a similar kind of result as the K-Means algorithm. Though it gurrentied time and space efficiency, it is not widely used like K-Means.The K-Means algorithm has been used in this proposed research study because of its compelling nature.

5.1.3 Volume-Profit correlation for K-Means Algorithm

The goal of this proposed model was to initiate a new term volume (V) with the existing LRFM model to ensure better clustering and finding out potential, valuable and profitable customers. The volume and profit have a proportional relationship with each other. As it was stated that volume is the amount of purchased products by any customer. It will be typified that the customer cluster with a large volume ensures more profit than others.
Regardless of the expenditure, customers with more buying habits have a direct influence on the profit of any firm. Later, it will be proven by different charts and graphs where it is noticed that the more customers buy products the larger profit the superstore can make.



	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
🟥 volume	3.425493	3.828936	5.634811	1.849331	6.768377	3.805241
— profit	94.862708	38.642514	154.993185	7.759579	602.388442	234.635715

Figure 5.3: Volume-Profit Relationship of LRFMV model for K-Means Algorithm

Here in Figure 5.3, a relation between Volume-Profit is shown, each cluster represents customers' volume of purchasing. Volume refers to the amount of purchased product by different types of customer segments. The highest amount of profit was generated through the highest volume of customers. Here cluster 4 has generated the highest profit compared to other clusters. Customers of cluster 4 have purchased the highest volume of products and it is about 6.76 and the profit value is 602.78. Cluster 5 has generated the second-highest profit. The volume of cluster 5 is 3.80 and the amount of profit shows 284.63 The addition of volume has been able to showcase the amount of profit which clearly contributes widely to the segmentation of the market. Cluster 2 is showing the volume is 5.63 and the profit is 154.99. These numerical values clearly indicate that the volume and profit of that super shop are proportionally dependent on each other as profit is higher when the volume is comparatively greater. This volume-profit analysis is used to determine how variations in the amount purchased influence a firm's profit. Businesses can use this research to figure out how many units they need to sell to break even (cover all costs) or achieve a certain profit margin. Despite the fact that all three algorithms were implemented, the K-Means approach was found to be the best fit for the proposed dataset and customer classification matrix.

5.1.4 Profit Analysis of each Cluster

Two different bar charts have been used to see the correlation between the number of clusters and the profit made by each cluster.

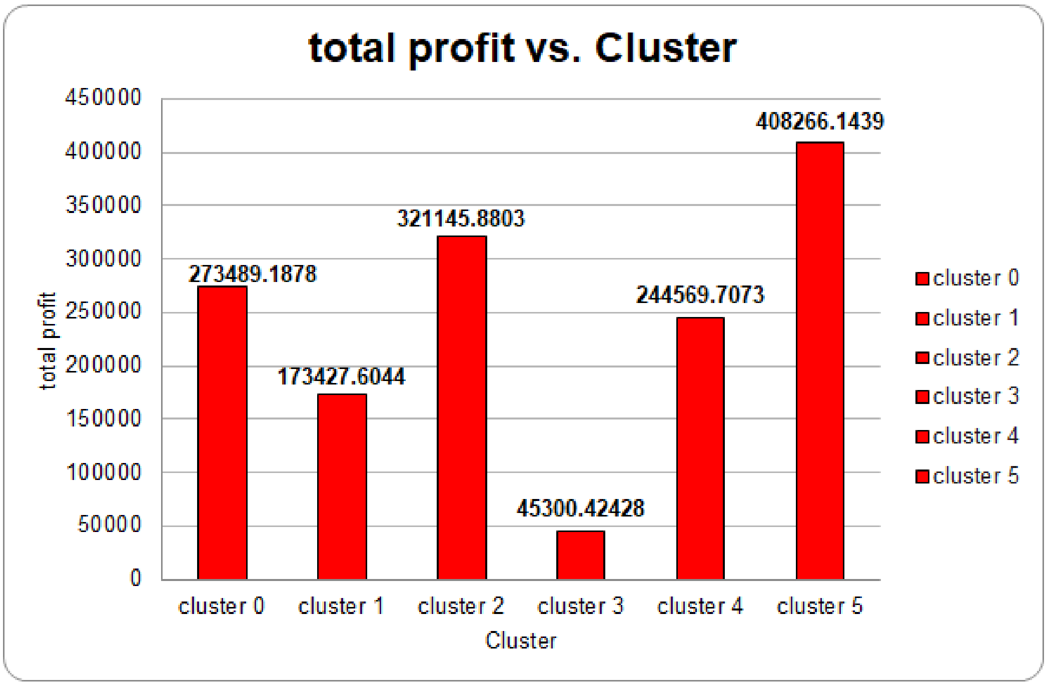


Figure 5.4: Profit of each cluster in LRFMV model

The above-mentioned bar chart is a clear visual representation of the profit made by each individual cluster. Each cluster's profit in numerical figures has been written outside as well.In this bar chart, Cluster 5 is the most profitable cluster, while Cluster 3 is the least profitable. But it is visible that most of the profit has been made by cluster 4 and cluster 5. It is covering approximately half of the profit from the total profit earned by different clusters of customers. The other 4 clusters are covering the rest half of profit in a year.

5.1.5 Customer Number Analysis for each cluster

K-means has been used here to discover 6 efficient clusters of customers. These 6 clusters contain different numbers of customers and these clusters denote a variety of parameters to calculate and understand the most potential buyer group in any organization.

This bar chart is created to show the number of customers in each cluster. The numerical values are also stated outside of the different segments of the bar chart. For instance, the largest cluster of customers is denoted by yellow which contains 5838 customers and it is cluster 3. On the contrary, the smallest one is cluster 4

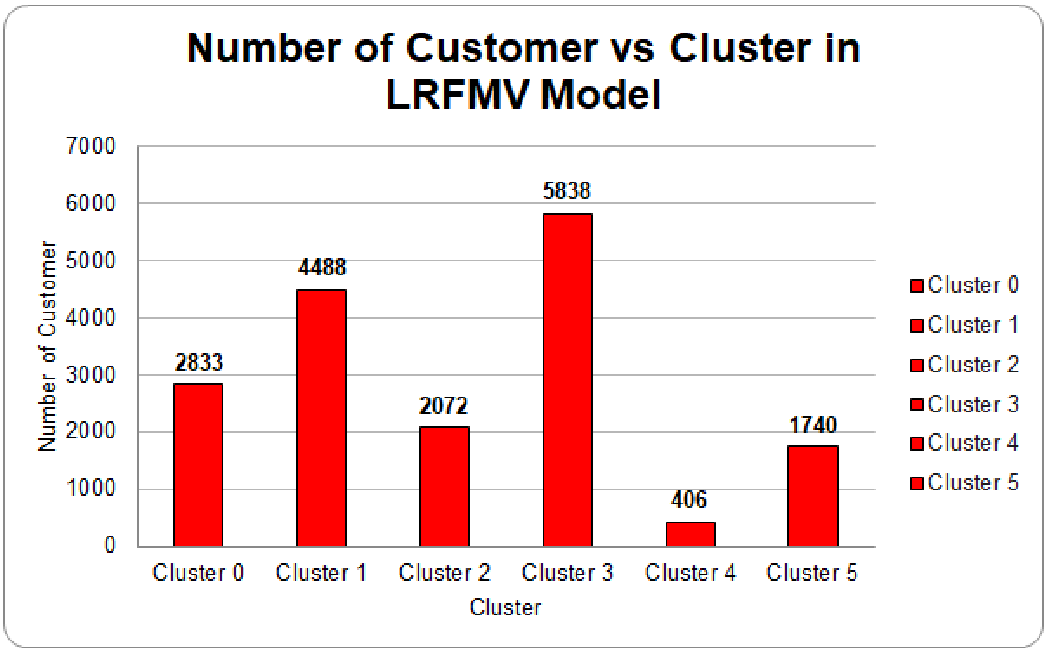


Figure 5.5: Number of customers in each cluster

which has a customer number of 406. It is remarkable that cluster 4 and cluster 5 are really small and cluster 5 has only 1740 customers and if clusters 4 and 5 have been combined, the size will be 1.5 times less than the largest cluster 3.
As can be seen, clusters 4 and 5 have just 12% clients but 45 percent profit. These are the two main clusters that made up the Passive Customer type in the Customer Classification Matrix.

5.2 Cluster Analysis with Customer-Classification Matrix

5.2.1 Revenue Generation for each cluster and their cost to serve

The information processing regarding the income, cost, and revenue are essential to assess the clusters from the LRFMV model, and we must apply the Customer Classification Matrix in our output for a better customer profitability check. Based on the cost to serve per revenue, this matrix depicts four different sorts of clients.
Several literature positions advised using the Customer Classification Matrix (matrix of customer revenue and cost to serve). This classification demonstrates that businesses can service lucrative clients in a variety of ways. The most valuable clients are those who are passive, providing great revenue at low cost. These are the most profitable consumers, and the corporation should pay them special attention. Some customers that generate a lot of money can also be expensive (carriage trade quadrant) - they can be lucrative if the revenue surpasses the expense of serving them. There may be consumers who are easy to satisfy yet don't bring in a lot of money (bargain basement quadrant). Finally, clients with high costs and poor revenue are included in the last quadrant (aggressive).

Aggressive- Aggressive clients seek (and frequently receive) the greatest possible product quality, the greatest possible service, and the lowest possible costs. Procter Gamble has a reputation among its suppliers for paying the least and getting the most, thanks to an effective procurement unit. Aggressive buyers are frequently powerful; their habit of buying in large quantities gives them negotiating strength with suppliers in order to get better prices and better service. The national accounts detailed in the second case at the start of this piece pushed the capital equipment supplier to make tough concessions.

Bargain basement - the corporation may concentrate on boosting revenue from these clients, beginning with research into whether they demand various services, their price sensitivity, and so on. Customers that are price conscious but somewhat unconcerned about service and quality are at the other end of the transport trade spectrum. They can be served at a lower cost than carriages.

Passive - Because passive customers provide the majority of the firm's earnings, the corporation should consider putting more resources into better serving these consumers, hence boosting their satisfaction and loyalty.

Passive clients are also less expensive to serve, but they are willing to pay high prices. These accounts earn a lot of money for you. Their attitude is due to a variety of factors. In some circumstances, the product is insignificant enough that a hard bargaining posture on pricing is unnecessary. Others are unconcerned about the price because the product is critical to their business. Others stick with their present supplier regardless of pricing since switching is too expensive. In some circumstances, vendor capacity and buyer needs are so closely aligned that the cost to serve is inexpensive, even while the client receives (and pays for) excellent service and quality.

Carriage trade - the company should concentrate on lowering the cost of serving customers, evaluating cost causes, and identifying ways to optimize internal procedures. Serving the carriage trade is expensive, but the carriage trade is willing to pay top money.

The horizontal axis is the cost to service, and the vertical axis is net pricing, from low to high. Any marketer will benefit from this classification. Any client connection that begins above the horizontal mid-point, on the other hand, is usually marked by a willingness and/or ability to pay more for services. Any relationship that lies to the right of the vertical axis has a higher cost to serve in general. Relationships that begin below the horizontal axis, on the other hand, have a lower willingness or capability to pay for your services. Low fees and a high cost to serve are found below the midpoint and to the right of the vertical axis, which is an unfavorable combination.

Table 5.1: Revenue Generation for each cluster and their cost to serve

Although all persons are created equal, the same cannot be true of customers. Some customers are more profitable than others, as it is known. Some, on the other hand, are outright unprofitable. The crucial thing is determining which is which. Despite huge disparities in profitability, many businesses maintain unprofitable customer connections, typically providing them with the same pricing and service levels as the most profitable ones. What is the reason for this? The majority of the time, businesses have no idea who their unprofitable clients are. As a result, they are unable to design marketing plans or effectively manage costs. A customer classification

Passive (Revenue high, the expense of serving low)	**Carriage Trade** (Revenue high, the expense of serving high)
Bargain Basement (Revenue low, the expense of serving low)	**Aggressive** (Revenue low, the expense of serving high)

matrix can be an effective and easily understandable way to find out the different epitomes of customers.

The cluster 4 and cluster 5 are found as Passive customers as the main serving guests. These accounts have a strong propensity to pay and low costs, making them very profitable. To improve loyalty and happiness, businesses can aim to focus more resources on them. A higher rise is found in the revenue with less cost for these two groups. They are contributing more to the profit gaining process. Cluster 4 shows the traits if it is compared with its features with the average values of Length, Recency, Frequency, Volume: L↓, R↑, F↓, M↑, V↑. It means that length and volume are rising while the others are decreasing. Cluster 5 also has these traits but with different types of change.If it is written in precise form it can be L↑, R↓, F↑, M↑, V↑.It is visible that if recency is falling other traits are increasing at the same time. Though cluster 5 doesn't have a similar kind of impact as cluster 4 in terms of revenue generated, it has the lowest cost to serve among all the 6 clusters and also has the second-largest profitability.

Secondly, Another most important segment is cluster 2 with the Carriage trade customers. These customers have an important impact on revenue with a small drawback. Despite the fact that they are the largest revenue generator and control the majority of revenue, the difficulty is that they generate high revenue at a significant cost. They have the second-highest cost among the 6 segments. Most stores make the most money by catering to the masses: ”load ’em high and sell ’em cheap.” Some, on the other hand, are able to make more money by focusing on the carriage trade, selling fewer products but at a substantially larger profit per item. The carriage trade is a specialized market, but there are other niche markets as well, therefore it wouldn't work well as a stand-alone term. Their traits are L↓, R↑, F↓, M↑, V↑. It indicates that recency and volume are increasing here and the others are decreasing meanwhile. They are the second important customer type in customer groups and close to the Passive Customer Type. They can be as profitable as the Passive customer type if the superstore does a background investigation to

reduce the cost. Reducing cost should be the goal for superstores to convert these customers to passive ones. These clients have a high cost of service yet are prepared to pay more. The superstore should look into cost drivers and streamline internal processes in order to cut costs.
Thirdly, cluster 0 has been considered as a Bargain customer group. They are the most flexible group and they give more priority to the price than the quality of the product. They are very easy to serve with one of the lowest costs. superstores can increase their profit from this group by conducting surveys and research to know this segment better so that they get more choices in terms of service. Sometimes reducing cost is the most convenient way to convince them. These clients are price-conscious and don't expect high levels of service or quality. To increase income, businesses should aim to turn clients into passive customers. The superstore should focus on understanding their demand to attract this group toward their services. Their traits are: L↑, R↓, F↑, M↓, V↑ where volume, frequency, and length are on a rise and others are falling.
Lastly, cluster 1 and cluster 3 are the most highly populated groups with the lowest impact on revenue. They are considered as the most unprofitable groups with the tag of Aggressive groups in the business. Cluster 3 has the highest number of members with the highest number of times when the superstore had to face loss to serve them. Cluster 1 has the second-highest number of customers with the highest amount of loss in revenue. superstores should renegotiate with these customers in terms of pricing in services. They should not lose these customers rather they should focus to avoid loss while serving these segments. Traits of cluster 1: L↓, R↑, F↓, M↓, V↑. and traits of cluster 3: L↓, R↑, F↓, M↑, V↓. These clients expect the best service but are unwilling to pay a premium price for it. They are the ones with the most power. Firms should endeavor to negotiate pricing, service, and delivery arrangements. Firms should try to raise the price of the service they provide to these customers.

5.2.2 Identification of Traits

Average of L, R, F, M, V for the whole Dataset:

For LRFMV modeling, a pre-processed data set is used, and each component of LRFMV(like length,recency,volume etc.) was calculated separately beforehand. The average of LRFMV components for the whole dataset is calculated, as well as the profit, at this point. The averages that were calculated are listed below. The values of L, R, F, M, V for each of the six clusters are also calculated separately. Giving numerical values in order to simplify the representation is avoided here and a table is created for easier understanding.
Average of LRFMV components for the processed dataset:
L= 181.129225 R= 507.312102 F= 1.478912 M= 481.944309 V= 3.379857 Profit= 84.133755

Average of L,R,F,M,V for each cluster:

The average has been calculated of L, R, F, M, V for each cluster in order to identify customer types in order to stabilize the marketing approach for superstores and label them as up or down. Profit and customer type are also found based on the comparison between average of L, R, F, M, V for each cluster and the whole dataset's LRFMV components.

Table 5.2: Comparison between Average of L, R, F, M, V of the whole dataset with each cluster

Clusters	**L**	**R**	**F**	**M**	**V**	**Profit**	**Customer type**
Cluster 0	up	down	up	down	up	Slightly high	Bargain
Cluster 1	down	up	down	down	up	low	Aggressive
Cluster 2	down	up	down	up	up	Slightly high	Carriage
Cluster 3	down	up	down	up	down	Incredibly low	Aggressive
Cluster 4	down	up	down	up	up	Incredibly high	Passive
Cluster 5	up	down	up	up	up	high	Passive

Every component(L, R, F, M, V) of LRFMV has been calculated and compared with the average of LRFMV components throughout the entire data set in the above table. After comparing, if a higher value has been obtained than the average it was written up and if a lower value is obtained for each component (L, R, F, M, V) then it was denoted as down.The profit of different clusters was also compared and stated the comparison with the up/down method.The last column is describing the customer type on the basis of L, R, F, M, V indications. It helps to get a better idea of different types of customers for divergent clusters.

Chapter 6

Conclusion

6.1 Research Overview

Although the LRFM model gives useful insight into customers based on their purchasing characteristics, the current study compared the standard RFM model to its expanded form, which incorporates a different dimension called volume in addition to the length. The term ”volume” refers to the number of goods delivered to a particular buyer during a single transaction. We attempted to demonstrate that it produces a more accurate outcome than RFM and LRFM researches in terms of segmentation. We used volume to establish a relationship between commodity and customer that previous models lacked. We did not examine customer behavior based on personal characteristics such as age, demography, gender, or race since we were just dealing with sales results. Additionally, a goal score was previously established to calculate the values of L, R, F, M, and V. Due to the automated nature of the equations, no predetermined score was assigned.
All six segments were evaluated obtained through the standard K-means, K-medoids and Mini Batch K-means algorithms and the near relationship between profit and volume for all three algorithms in the proposed LRFMV model is examined.K-Means algorithm was chosen for further analysis and customer labeling according to classification matrix for its widespread and ascertained use. When the same algorithm has been applied on the RFM model, LRFM model and proposed LRFMV model, it is apparent that the LRFMV model will produce additional segments with the same number of customers and a higher profit per head.

6.2 Contribution and Impact

Numerous studies on customer segmentation using RFM and LRFM models have been performed since the invention of these concepts, but only a few of them can create a connection between customers and commodity quantity. The proposed analysis makes a significant impact by establishing a strong correlation between earnings per head and the commodity purchased by each customer in a single transaction. The research demonstrates a novel method for segmenting customers into productive clusters based on the volume of the commodity. The model presented in LRFMV research is capable of resolving a variety of issues related to determining the optimal customer for the optimal product.
Similarly, customer segmentation is a technique for improving the contact with cus-

tomers, for learning about their desires and activities such that companies' issues can be developed. Customer segmentation is essential to acclimate new customers and maximize earnings. Prospective customer data may be used to deliver programs depending on the type of customers, such as internet advertising purchasing and sale. Additionally, the objective of K-Means implies is to classify data points into separate, non-overlapping subpopulations. One of the more popular uses of K-means clustering is segmenting customers in order to get a greater view of them, which can then be used to maximize the company's sales. Along with K-Means,other algorithms like K Medoids, Mini Batch have also been used for cross checking the clusters of K-Means for this dataset. Here, K-Means algorithm has been proved and used for segmenting the customers. The LRFMV analysis will assist company owners in more efficiently segmenting their clients, which would result in more efficient contact.

6.3 Recommendation & Future Work

The LRFMV approach can be applied to databases of non-discrete details and a smaller variation in the number of data points. Additionally, since k-means clustering is prone to outliers, it is preferable to exclude them first. Numerous companies could potentially use this model in the future to derive market characteristics from matrices of customer research. Similarly, this technique can be used for datasets that have a low variance in terms of commodities.Along with K-Means, some other algorithms were used here as well but certain clustering algorithms, such as K-medoids can be used for dynamic dataset, Mini Batch can be used to save time and space etc. Additionally, this model can be used to analyze other aspects of advertising, and its reliability can be quantified using certain matrices.

Bibliography

[1] K. Pearson, "On lines of closes fit to system of points in space, london, e dinb," *Dublin Philos. Mag. J. Sci*, vol. 2, pp. 559–572, 1901.

[2] H. Hotelling, "Analysis of a complex of statistical variables into principal components, j.educ," es, *Psych*, vol. 24, pp. 417–441, 498–520, 1933.

[3] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," en, *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. DOI: 10.1016/0377-0427(87)90125-7. [Online]. Available: https://doi.org/10.1016/0377-0427(87)90125-7.

[4] A. Hughes, en, in *database marketing*, Chicago: Probus Publishing Company, 1994.

[5] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," en, in *Proceedings of the Fifth International Conference on Extending Database Technology*, 1996, pp. 3–17.

[6] S. Chow and R. Holden, "Toward an understanding of loyalty: The moderating role of trust," en, *Journal of Managerial Issues*, vol. 9, no. 3, pp. 275–298, 1997, Retrieved May 27, 2021, from. [Online]. Available: http://www.jstor.org/stable/40604148.

[7] W. Reinartz and V. Kumar, "On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing," en, *Journal of Marketing*, vol. 64, no. 4, pp. 17–35, 2000. DOI: 10.1509/jmkg.64.4.17.18077. [Online]. Available: https://doi.org/10.1509/jmkg.64.4.17.18077.

[8] Z. Ghahramani, "Unsupervised learning," in *Summer School on Machine Learning*, Springer, 2003, pp. 72–112.

[9] H. Chang and S. Tsay, "Integrating of som and k-mean in data mining clustering: An empirical study of crm and profitability evaluation," en, *Journal of Information Management*, vol. 11, pp. 161–203, 2004.

[10] N. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers," it, *Expert Syt Appl*, vol. 27, pp. 623–633, 2004.

[11] O. Etzion, A. Fisher, and S.-C. Wasserkrug, "A modeling approach for customer lifetime evaluation in e-commerce domains, with an application and case study for online auction," en, *Inf Syst Front*, vol. 7, pp. 421–434, 2005. DOI: 10.1007/s10796-005-4812-6. [Online]. Available: https://doi.org/10.1007/s10796-005-4812-6.

[12] C. Fishman, *The Wal-Mart Effect: How the World's Most Powerful Company Really Works–and HowIt's Transforming the American Economy*, en, Reprint. Penguin Books, 2006.

[13] B. Sohrabi and A. Khanlari, "Customer lifetime value (clv) measurement based on rfm model," it, *Iranian Acc. Aud. Rev*, vol. 14, no. 47, pp. 7–20, 2007.

[14] Y. I. C., Y. K. J., and T. T. M, "Knowledge discovery on rfm model using bernoulli sequence," en, *Expert System with Applications*, vol. 36, pp. 5866–5871, 2008.

[15] K. Desouza, Y. Awazu, S. Jha, C. Dombrowski, S. Papagari, P. Baloh, and J. Kim, "Customer-driven innovation," en, *Research Technology Management*, vol. 51, no. 3, pp. 35–44, 2008, Retrieved May 27, 2021, from. [Online]. Available: http://www.jstor.org/stable/24135952.

[16] S. Li, L. Shue, and S. Lee, "Business intelligence approach to supporting strategy-making of isp service management," fr, *Expert Syst. Appl*, vol. 35, pp. 739–754, 2008.

[17] S. Lumsden, S. Beldona, and A. Morison, "Customer value in an all-inclusive travel vacation club: An application of the rfm framework," en, *J. Hosp. Leisure Mark*, vol. 16, no. 3, pp. 270–285, 2008.

[18] R. Blattberg, E. Malthouse, and S. Neslin, "Customer lifetime value: Empirical generalizations and some conceptual questions," en, *Journal of Interactive Marketing*, vol. 23, pp. 157–168, 2009. DOI: 10.1016/j.intmar.2009.02.005..

[19] Y.-L. Chen, M.-H. Kuo, S.-Y. Wu, and K. Tang, "Discovering recency, frequency, and monetary (rfm) sequential patterns from customers' purchasing data," en, *Electronic Commerce Research and Applications*, vol. 8, no. ue 5, pp. 241–251, 2009, ISSN 1567-4223, DOI: 10.1016/j.elerap.2009.03.002.. [Online]. Available: https://doi.org/10.1016/j.elerap.2009.03.002..

[20] C. Cheng and Y. Chen, "Classifying the segmentation of customer value via rfm model and rs theory," en, *Expert Systems with Applications*, vol. 36, pp. 4176–4184, 2009.

[21] T. Jiang, A. Tuzhilin, and March, *Improving personalization solutions through*, en, 2009.

[22] C. Wang, "Outlier identification and market segmentation using kernel-based clustering techniques'," en, *Expert Systems with Applications*, vol. 36, no. 2, pp. 3744–3750, 2009.

[23] H.-H. Wu, E.-C. Chang, and C.-F. Lo, "Applying rfm model and k-means method in customer value analysis of an outfitter," en, *16th ISPE International Conference on Concurrent Engineering*, no. 2, pp. 665–672, 2009.

[24] D. Cao and B. Yang, "An improved k-medoids clustering algorithm," in *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, vol. 3, 2010, pp. 132–135. DOI: 10.1109/ICCAE.2010.5452085.

[25] Y. Li, L. CH, and L. CY, "Identifying influential reviewers for word-of-mouth marketing," en, *Elect. Com. Res Appl*, vol. 9, pp. 294–304, 2010.

[26] B. Pardeshi and D. Toshniwal, "Improved k-medoids clustering based on cluster validity index and object density," in *2010 IEEE 2nd International Advance Computing Conference (IACC)*, 2010, pp. 379–384. DOI: 10.1109/IADCC.2010.5422924.

[27] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 1177–1178.

[28] Y.-T. Kao, H.-H. Wu, H.-K. Chen, and E.-C. Chang, "A case study of applying lrfm model and clustering techniques to evaluate customer values," fr, *Journal of Statistics and Management Systems*, vol. 14, no. 2, pp. 267–276, 2011. DOI: 10.1080/09720510.2011.10701555.

[29] A. Parvaneh, H. Abbasimehr, and M. Tarokh, "Data mining application in retailer segmentation based on lrfm variables: Case study," fr, *Global Journal on Technology*, vol. 1, 2012.

[30] F. Steeneken and D. Ackley, "A complete model of the supermarket business," en, *BPTrends*, 2012, Retrieved from: [Online]. Available: https://www.bptrends.com/publicationfiles/01-03-2012-ART-Supermarket%20Article-steeneken-Ackley%20111226.pdf.

[31] J.-T. Wei, S.-Y. Lin, C.-C. Weng, and H.-H. Wu, "A case study of applying lrfm model in market segmentation of a children's dental clinic," en, *Expert Systems with Applications*, vol. 39, no. 5, pp. 5529–5533, 2012.

[32] J. Béjar Alonso, "K-means vs mini batch k-means: A comparison," 2013.

[33] Y. Cho and S. Moon, *Weighted mining frequent pattern based customer ' s rfm score for personalized u-commerce recommendation system*, en, 2014.

[34] K. Coussement, F. Van Den Bossche, and K. De Bock, "Data accuracy's impact on segmentation performance: Benchmarking rfm analysis, logistic regression, and decision trees"," en, *Journal of Business Research*, vol. 67, pp. 2751–2758, 2014.

[35] A. Sarveniazi, "An actual survey of dimensionality reduction," en, *American Journal of Computational Mathematics*, vol. 04, pp. 55–72, 2014. DOI: 10.4236/ajcm.2014.42006..

[36] R. Shamsher, "Growth of super stores in bangladesh: A theoretical framework," en, *International Journal of Agriculture and Rural Economic Research. International Journal of Agriculture and Rural Economic Research*, vol. 2, pp. 5–13, 2014.

[37] H.-H. Wu, S.-Y. Lin, and C.-W. Liu, "Analyzing patients' values by applying cluster analysis and lrfm model in a pediatric dental clinic in taiwan," en, *TheScientificWorldJournal*, p. 685 495, 2014. DOI: 10.1155/2014/685495.

[38] M. Casabayó, N. Agell, and G. Sánchez-Hernández, "Improved market segmentation by fuzzifying crisp clusters: A case study of the energy market in spain," en, *Expert Systems with Applications*, vol. 42, no. 3, pp. 1637–1643, 2015.

[39] R. Daoud, A. Amine, B. Bouikhalene, and R. Lbibb, “Combining rfm model and clustering techniques for customer value analysis of a company selling online,” en, in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA*, 2015, pp. 1–6. DOI: 10.1109/AICCSA.2015.7507238..

[40] J. Li, S. Song, Y. Zhang, and Z. Zhou, “Robust k-median and k-means clustering algorithms for incomplete data”,” en, *Mathematical Problems in Engineering*, vol. ID 4321928, 8 pages, 2016. DOI: 10.1155/2016/4321928. [Online]. Available: https://doi.org/10.1155/2016/4321928.

[41] P. Sarvari, A. Ustundag, and H. Takci, “Performance evaluation of different customer segmentation approaches based on rfm and demographics analysis,” en, *Kybernetes*, vol. 45, pp. 1129–1157, 2016. DOI: 10.1108/K-07-2015-0180.

[42] M. Mohammadzadeh, Z. Hoseini, and H. Derafshi, “A data mining approach for modeling churn behavior via rfm model in specialized clinics case study: A public sector hospital in tehran,” es, *Procedia Computer Science*, vol. 120, pp. 23–30, 2017.

[43] S. Peker, A. Kocyigit, and P. Eren, “
lrfmp model for customer segmentation in the grocery retail industry: A case study,’” en, *Marketing Intell. Planning*, vol. 35, no. 4, pp. 544–559, Jun. 2017.

[44] A. Sheshasaayee and L. Logeshwari, “An efficiency analysis on the tpa clustering methods for intelligent customer segmentation,” en, in *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore*, 2017, pp. 784–788.

[45] T. Tanaka, T. Hamaguchi, T. Saigo, and K. Tsuda, “Classifying and understanding prospective customers via heterogeneity of superstore stores,” en, *Procedia Computer Science*, vol. 112, pp. 956–964, 2017, ISSN 1877-0509, DOI: 10.1016/j.procs.2017.08.133.. [Online]. Available: https://doi.org/10.1016/j.procs.2017.08.133..

[46] L. Zahrotun, “Implementation of data mining technique for customer relationship management (crm) on online shop tokodiapers.com with fuzzy c-means clustering,” en, in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta*, 2017, pp. 299–303.

[47] A. Alizadeh-Zoeram and A. Karimi, “A new approach for customer clustering by integrating the lrfm model and fuzzy inference system,” en, *Iranian Journal of Management Studies*, vol. 11, pp. 351–378, 2018. DOI: 10.22059/ijms.2018.242528.672839.

[48] F. Bachtiar, “Customer segmentation using two-step mining method based on rfm model,” en, in *International Conference on Sustainable Information Engineering and Technology (SIET*, 2018, pp. 10–15. DOI: 10.1109/SIET.2018.8693173.

[49] A. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, “Rfm ranking – an effective approach to customer segmentation,” en, *Journal of King Saud University - Computer and Information Sciences*, 2018. DOI: 10.1016/j.jksuci.2018.09.004.

[50] S. Monalisa, “Analysis outlier data on rfm and lrfm models to determining customer loyalty with dbscan algorithm,” en, in *International Symposium on Advanced Intelligent Informatics (SAIN*, 2018, pp. 1–5. DOI: 10.1109/SAIN.2018.8673380..

[51] M. Pakyürek, M. Sezgin, S. Kestepe, B. Bora, R. Düzağaç, and O. Yıldız, “Customer clustering using rfm analysis,” en, in *26th Signal Processing and Communications Applications Conference (SIU*, 2018, pp. 1–4. DOI: 10.1109/SIU.2018.8404680.

[52] M. Syakur, B. Khotimah, E. Rohman, D. Satoto, and Budi, “Integration k-means clustering method and elbow method for identification of the best customer profile cluster,” en, *IOP Conference Series: Materials Science and Engineering*, vol. 336, pp. 012017 10 1088 1757–899 336 1 012017, 2018.

[53] M. Tavakoli, M. Molavi, V. Masoumi, M. Mobini, S. Etemad, and R. Rahmani, “Customer segmentation and strategy development based on user behavior analysis, rfm model and data mining techniques: A case study,” en, in *2018 IEEE 15th International Conference on E-Business Engineering (ICEBE). IEE*, 2018.

[54] M. Alam and N. Noor, *Superstore retailing in bangladesh: A comprehensive literature review from consumer perspective*, en, 2019.

[55] J. Nagesh, “Generating political scores using rfm model and cluster prediction by xgboost,” en, in *2019 International Conference on Computational Science and Computational Intelligence (CSCI*, Las Vegas, NV, USA, 2019, pp. 1333–1336. DOI: 10.1109/CSCI49370.2019.00249.

[56] T. Segal, *Inside recency, frequency, monetary value (rfm*, es, Jul. 5, 2019. [Online]. Available: https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp.

[57] S. Guney, S. Peker, and C. Turhan, “A combined approach for customer profiling in video on demand services using clustering and association rule mining,” en, in *IEEE Access*, vol. 8, 2020, pp. 84 326–84 335. DOI: 10.1109/ACCESS.2020.2992064..

[58] Jasmin, *Machine learning in customer segmentation with rfm-analysis*, en, Retrieved from: Nov. 12, 2020. [Online]. Available: https://www.nextlytics.com/blog/machine-learning-in-customer-segmentation-with-rfm-analysis.

[59] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. John-son, H. Kaushansky, and A. Software, “Churn reduction in the wireless industry,” en, in *Advances in Neural Information Processing Systems*, vol. 12, PDF) Combining Sequential and Aggregated Data for Churn Prediction in Casual Freemium Games., pp. 935–941, 2000 1.

[60] R. Soeini and E. Fathalizade, en, Customer Segmentation based on Modified RFM Model in the Insurance Industry.,