LRFMV$_D$: A Customer Segmentation Model

by

Kawsar Mahmud Sagor
18101638
Masrur Arefin Sadhin
18101626
Ishrat Jahan
18101310
Rezwanul Karim Prottay
18101308

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
May 2023

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing the 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 that 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:

Sagor

Kawsar Mahmud Sagor
18101638

Sadhin

Masrur Arefin Sadhin
18101626

Ishrat Jahan

Ishrat Jahan
18101310

Rezwanul.

Rezwanul Karim Prottay
18101308

Approval

The thesis titled “LRFMV$_D$: A Customer Segmentation Model ” submitted by

1. Kawsar Mahmud Sagor(18101638)

2. Masrur Arefin Sadhin(18101626)

3. Ishrat Jahan(18101310)

4. Rezwanul Karim Prottay(18101308)

Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 25, 2023.

Examining Committee:

Supervisor:

(Member)

Shakila Zaman
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:

(Member)

Jannatun Noor
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

____

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

Ethics Statement

In this thesis, overall results and comparisons made are based on the group's own research and study on this topic. The dataset used here has been properly cited, as have all essential sources. To ensure the analysis's transparency, appropriate precautions have been taken. This thesis has never been submitted to another institution.

Abstract

Customer segmentation is a big part of the superstore industry. Traditionally, the RFM model has been used to segment customers to maximize profit. This work proposes a new customer segmentation named LRFMVD based on RFM and LRFMV models in hopes of providing a more sure-fire way of segmenting customers. The k-means clustering method will be used for the proposed model. The clusters created by K-means are then analyzed using the LRFMV$_D$ model to find a correlation between profit and volume. Many works have been done previously on customer segmentation for maximizing profit, but none of those were able to show a straightforward representation of profit, volume, and discounts on products. Unsupervised learning was used to investigate the correlations between volume, discount, and profit. Customers are then segmented using the Customer Classification Matrix, which looks at the properties of all clusters. The L, R, F, M, V$_D$ parameters' values are compared to the cluster mean values, and based on whether these values are higher or lower than the average, customers are segmented. Comparisons among the three models reveal that the latter provides more profit per head than the other two, and is able to identify customers who cause superstores to lose money or make a loss.

Keywords:volume, silhouette, elbow, RFM analysis, LRFMV and LRFMV$_D$ analysis, K- means

Adherence

The respected departmental faculty and our adored parents, who have supported and encouraged us throughout the thesis and inspired us to strive for excellence in all areas, are honored by receiving this dissertation.

Acknowledgement

Our thesis was successfully finished without any major setbacks, and we are grateful to Allah for that. We would also like to express our gratitude to Mrs. Shakila Zaman, for her kind help and guidance over the course of our work. She was always ready to help us. Finally, it might not have been possible without our parent's ongoing support. Thanks to their tremendous guidance as well as prayers, our graduation is near the end.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Customer segmentation is the procedure of categorizing customers based on some common or shared traits. Using this, superstores can create strategies that can increase the likelihood that a customer will make a purchase. Information like ethnic backgrounds, religious affiliations, income levels, and education of customers, as well as their geographic location, are taken into account when segmenting them. Marketing automation software is responsible for using the information to identify the client segments and to deliver strategies for targeted marketing. This enables organizations to efficiently use marketing resources and maximize profit. This also helps to improve customer service which boosts customer satisfaction. Conventional RFM model segmentation helps in this regard by segmenting customers and by identifying more profitable customers. Recently a new model was developed named LRFMV which introduces a new attribute called volume (V) with a view to further enhance the clustering process that can improve customer segmentation. Another study implemented the LRFM model for this same purpose that uses length (L) as a variable to determine clusters.

For the LRFMV model, length means the duration of a customer's first and last purchase. Volume is a re-measured version of the quantity of products that a client purchases during his entire customer life cycle. The length was introduced to get the most out of recency information. The proposed LRFMV$_D$ model hopes to further expand upon the RFM and LRFMV models by using discount as a factor as discounts directly affect the quantity of products purchased.

LRFMV model builds upon the RFM Model by introducing a new variable volume(V) which uses customers' quantity of purchased products. However, this does not take the products' discount into consideration. But introducing discounts into volume calculation can deal with this issue and in the ideal scenario segment customers more effectively.

The research is attempting to tackle the following questions:

Does introducing discount to volume affect customer segmentation and profit generation?

Does it outperform the RFM and LRFMV models in maximum profit generation?

Can it effectively identify customer segments?
The objectives of this research are:
To show a volume-profit correlation for each customer when the discount is factored in.
Comparing the profits of the conventional RFM analysis, the LRFMV analysis, and the LRFMV$_D$ analysis for maximum profit generation among the clusters.
Calculating the LRFMV$_D$ values and segmenting clients effectively.

1.1 Motivation

A superstore is a self-service establishment that is divided into different sections to ensure a greater range of products[31]. Typically, local people and small companies are the customers of such a store. The superstores serve as a backdrop for shopping and bridging the gap between customers and suppliers.

In Bangladesh, superstores have gained popularity since the start of the 2000s[31].Because of its rising market, different strategies need to be built to retain customers, and Segmenting clients is an effective method among others. This allows businesses to retain clients, reduce expenses, and provide better customer satisfaction.

There are many customer segmentation models of which RFM is the most used. But RFM does not take the number of purchased products of a customer into account. This is a concern because the number of products can be used to predict customer behavior. LRFMV is an extension of the RFM model and uses quantity in its clustering method. But another problem arises when there may be discounts on products. The volume (V) calculated with quantity in the LRFMV model did not take a discount on products into account. Discounts may often affect buying tendencies and directly contribute to the number of purchased products thus affecting volume.

The above-mentioned problems gave incentive to the creation of the proposed model. This research was done with a view to creating a customer segmentation system that will allow superstores to differentiate between customers more effectively so that they can come up with better marketing strategies and identify more profitable customers. This can enable the superstores to devise strategies more catered towards those profitable customers with the intention of generating more profit.

1.2 Research Methodology

The dataset used in this study was created by another party as it is extremely resource intensive to obtain data from primary sources. Only quantitative data from the dataset has been used. Customers' personal information was not used.

1.2.1 RFM Model

Organizations use three marketing measurement approaches recency, frequency, and financial value to pinpoint the most valuable clients according to Segal (2019) [33].

- **Recency:** Refers to how a recent purchase was made by a customer.
- **Frequency:**How many transactions were made customer in his total customer life cycle.

- **Monetary value:** Total spendings of a customer.

1.2.2 LRFMV Model

The LRFMV model is an extension of LRFM model. It introduces a new parameter called volume which takes a customer's quantity of purchased into account [35]. This enables the model to produce more clusters than the RFM and LRFM model.

1.2.3 LRFMV$_D$ Model

This study is investigating a new model that can segment customers more effectively than RFM and LRFMV model and provide a better profit margin for those segments. It builds upon the LRFMV model's volume by considering discounts on purchased products. Volume, when calculated with discount has a favorable impact on the revenue from a customer base.

1.3 Opportunity and Restrictions

This study hopes to provide a comparative analysis of RFM, LRFMV, and the proposed LRFMV$_D$ model by segmenting customers using the K-means algorithm. This data works only for quantitive data. No types of qualitative data were used in this model.

If there is less variance in product numbers, then the change of values of LRFMV and LRFMV$_D$ may be insignificant. Units of amounts have been ignored and PCA has been used to reduce dimensions for each sample pair.

Chapter 2

Literature Review

Data analysis is a huge part of modern-day organizations and various methods have been created for this. As the business ' transaction volume rises, it is getting harder to segment profitable consumers to increase sales business' transaction volume rises. The RFM model can be used in the client's segmentation process to analyze a client's purchasing habits and help grow better inclination and classifying methods [22], [12]. Before using clustering techniques such as normal K-means, [29], [3], [13], fuzzy C-means [19], [25], and repetitive median-based K-Means (RM K-Means) algorithms for clustering [21], the transactional data is first subjected to an RFM analysis. After that, The clusters are then evaluated in order to categorize clients effectively.

Many researchers have used the RFM segmentation model to effectively segment and identify potential customers. Customer response to direct marketing is predicted, for example, by satisfying customer [10], the value of customer's to lifetime [6], churn prediction[4], [22], and CLV measurement [7], [11] may be examined using data mining methods, and clients can be divided into groups based on how profitable they are. Some writers developed a mechanism based on two phases that can be a solution using the RFM model
[24]. RFM is used in various industries in addition to superstore research, such as finance and insurance [5], [7], telecommunications [8], political score generating [32], on-line companies[14], tourist organizations [9], wholesale industry [23], curative field [22], [16], and so on. The Tavakoli.M et al. study [30] highlights the significance of treating consumers with respect based

on their background and categorization, which has changed dramatically in recent years. First, the best cluster numbers are determined using Daoud's. An et al [20] self-organizing maps system (SOM).

Cheng along with Chen used K-means to mine classification rules in rough set theory [10] by fusing the quantitative value of RFM characteristics. The suggested methodology aids trade in creating CRM while also minimizing the downsides of various data learning tools for creating CRM. This clustering is used to divide the output into an odd number of classifications based on subjective evaluation and the chosen class is the one that has the consistency rating. To create decision rules, the authors employed the uneven set LEM2 structure. However, an updated RFM model typically gives satisfactory accuracy and profit margin.

Cho and Moon [17] suggested a personalized suggestion method that uses weighted frequent pattern mining to classify potential customers using the RFM model. They successfully offered clients an appropriate recommendation as a result. The company's capacity to track important clients also increased the net margin. However, they might get a good outcome by including any other variables in place of utilizing traditional RFM models.

In order to enhance data mining approaches Bachtiar. A created a model based on RFM which uses to step mining process [26]. Following segmentation using K-means clustering, the acquired data is initially judged by using the RFM approach. After the clusters are further investigated using association rule learning, customer attributes are specified by IF-THEN rules. The cluster outcomes are evaluated based on the use of silhouette coefficients and connectivity measurements. The two-step technique to customer analysis provides precise insight into customer behavior and purchase tendencies, which can truly aid in greatly improving marketing strategy. As a result, there might not be enough valuable customer research data at this point. Furthermore, by more accurately assessing lucrative clients while taking into consideration some extra aspects, the usage of contemporary RFM models could boost the research process.

Christy, A.J., et al. [27] carried out another outstanding study in which they created a novel method for using transactional data with the RFM model. They contrasted the RFM results with the traditional data mining methods. The repetitive Median centered K-means algorithm, or RM K-means, is the name of the modified algorithm. For this comparative analysis, the execution time, iterations, and cluster compactness of both methods were taken into account. Because it executes faster and requires fewer iterations than the other two algorithms, RM K-means outperforms them, according to the authors. On the other hand, using the conventional RFM method will make it challenging to locate potential clients.

The LRFMV model builds upon the LRFM model by introducing volume to the LRFM model[24][42]. The volume takes the number of purchased products of a customer into account and segments them accordingly. But a drawback of this model is that it does not take the product's discount into account which results in it not being able to identify which customers cause the superstore to make a loss.

Chapter 3

Proposed Model(LRFMV$_D$)

Following some methodologies, the model definition process is concerned with the inclusion and exclusion of variables from the research that is independent.

3.1 Collection of Data

We found it difficult to physically look for a real-life dataset as many superstores were unwilling to provide their customer information to the public. The Tableau Software Company, which is located in the United States, was responsible for shaping and publishing the dataset after it was gathered online. Descriptions of a few extracted attributes are provided below.
Order id: Used to determine the number of purchased products from a single order.
Date of the order: Used to determine the number of orders placed on a single day by a customer.
Customer ID: A customer id is required in order to identify each client's sales.
Sales: The entire purchase and the total price of the product after calculation.
Quantity: The amount of a particular commodity is referred to as quantity.
Discount: A deduction from the usual cost of something.
Profit: Needs further analysis of the dataset. Features were renamed.
Features were renamed to Sales:: The entire purchase and the total price of the product after calculation.
These are some of the extracted attributes used in the study.

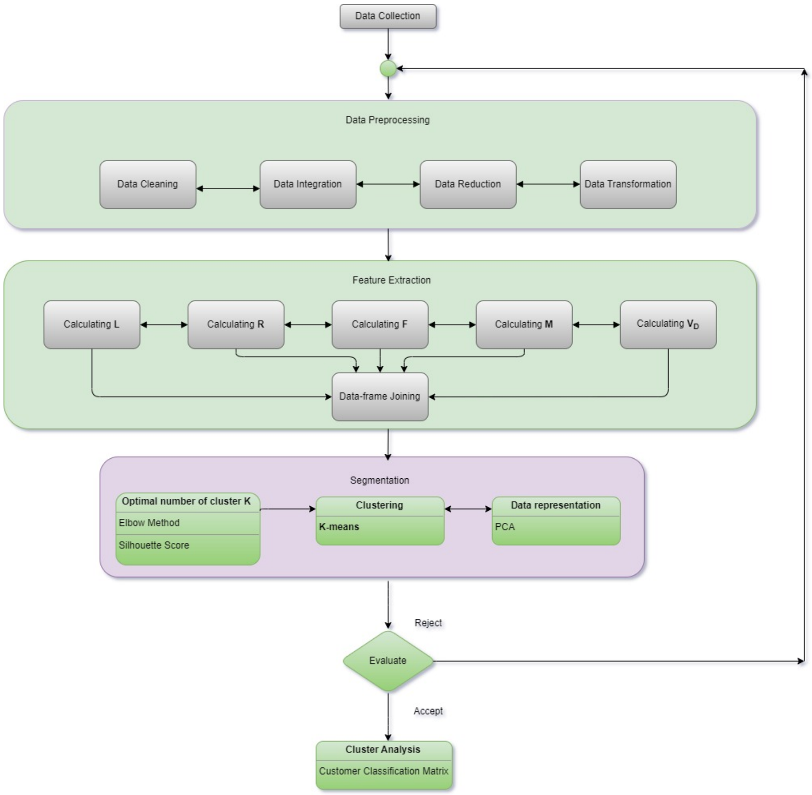
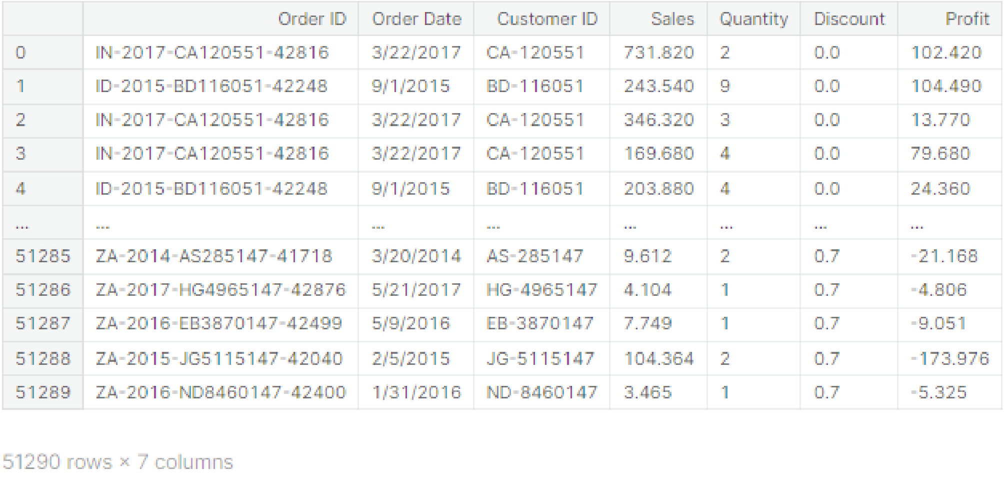


Figure 3.1: Proposed LRFMV$_D$ model



	Order ID	Order Date	Customer ID	Sales	Quantity	Discount	Profit
0	IN-2017-CA120551-42816	3/22/2017	CA-120551	731.820	2	0.0	102.420
1	ID-2015-BD116051-42248	9/1/2015	BD-116051	243.540	9	0.0	104.490
2	IN-2017-CA120551-42816	3/22/2017	CA-120551	346.320	3	0.0	13.770
3	IN-2017-CA120551-42816	3/22/2017	CA-120551	169.680	4	0.0	79.680
4	ID-2015-BD116051-42248	9/1/2015	BD-116051	203.880	4	0.0	24.360
...	...	...	...	...	...	...	...
51285	ZA-2014-AS285147-41718	3/20/2014	AS-285147	9.612	2	0.7	-21.168
51286	ZA-2017-HG4965147-42876	5/21/2017	HG-4965147	4.104	1	0.7	-4.806
51287	ZA-2016-EB3870147-42499	5/9/2016	EB-3870147	7.749	1	0.7	-9.051
51288	ZA-2015-JG5115147-42040	2/5/2015	JG-5115147	104.364	2	0.7	-173.976
51289	ZA-2016-ND8460147-42400	1/31/2016	ND-8460147	3.465	1	0.7	-5.325

51290 rows × 7 columns

Figure 3.2: Selected column of data set

Initially, Tableau Software gathered this dataset (figure 3.2) from a superstore and adjusted it later. This dataset does not contain any personally identifiable data, like client names, income, occupation, or age. Tableau Software conceals this to protect the privacy and confidentiality of its users. A data set called Global Superstore with about 50,000 records in it has been used in this study which contains informational data of customers.

3.2 Preprocessing

This study used a superstore dataset that had been preprocessed to examine a group of potential customers.

3.2.1 Cleaning the Data

In order to fill in missing numbers and repair inaccurate data, data cleaning processes are performed. Depending on the user's preferences, data cleaning can be completed in different ways. Binding, regression, or grouping are methods for removing random data [15]. Data smoothing using the binding approach is done with sorted data. In a variety of ways, the technique is carried out once the data is divided into equal-sized pieces. A segment's average can be used to replace all the data in that segment or the task can be completed using boundary values. Outlier analysis is used to identify and deal with outliers with the help of clustering tools. It is possible to replace the missing value by using the mean or median when employing a central tendency for the characteristics. Furthermore, every class has its own central tendency measurements. It's generally advised against eliminating the training dataset because doing so leads to data loss as attribute values that could add value to the data collection are being eliminated[25]. When utilizing a standard value to replace it, the missing value can be filled in with global constants like 'N/A' or 'Unknown'. Regression and decision-tree techniques, which fill in the missing value with the most probable value, can be used to anticipate and replace missing data. The fundamental tendency of characteristics to fill in the missing value has been exploited to tidy up data [28]. Numerous imaginary and null values were preventing the suggested model from being established. Therefore, for those columns where null values regularly appeared, the mean value was calculated and used.

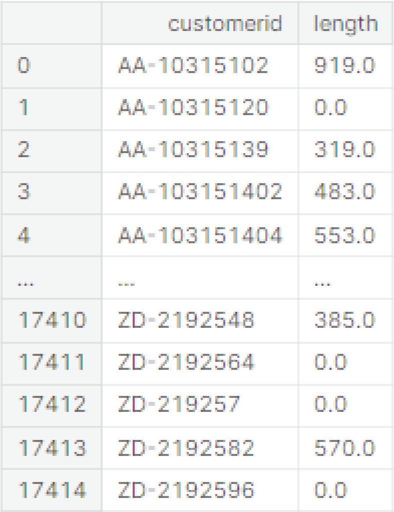
3.3 Extracting Feature

3.3.1 Calculating L

The length in the LRFMV$_D$ model is defined as the number of days separating a customer's first and last purchase. The length(L) can be computed using the following formula:

$$L = p_l - p_f$$

Where p$_l$=final purchased date, p$_f$= first purchased date.



	customerid	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

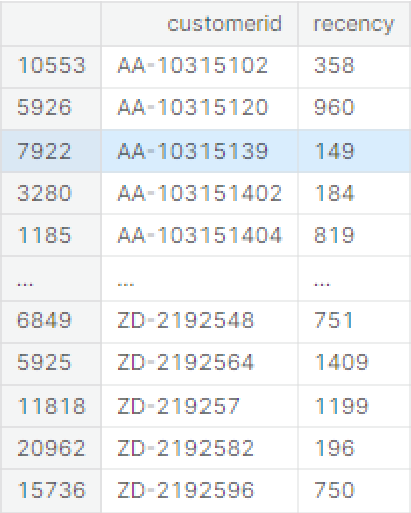
Figure 3.3: Value of Length (L)

3.3.2 Calculating R

The days after a customer's last purchase is recency. Recency (R) can be computed :

$$R = D_r - C_r$$

Where D$_r$= the most present date of the dataset, C$_r$=the most present data of a particular customer.



	customerid	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

Figure 3.4: Value of Recency (R)

Calculated recency is shown in figure 3.4.
Every customer's most recent visit or purchase date was subtracted from the most recent date of the dataset which was then assigned as Recency.

3.3.3 Calculating F

How many times a customer made a purchase in his customer life cycle with a shop is referred to as frequency. If p_f denotes a single purchase of a customer, then Frequency, F is:

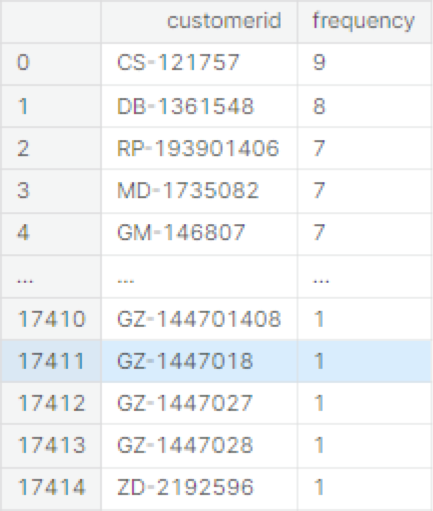
$$F = \text{count}(p_f)$$

The frequency value for each client is shown in figure 3.5.

3.3.4 Calculating M

Monetary was calculated using the number of transactions and total money spent by customers on those transactions. If P_s refers to total spending and x is the sum of all transactions, Monetary, M is,

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



	customerid	frequency
0	CS-121757	9
1	DB-1361548	8
2	RP-193901406	7
3	MD-1735082	7
4	GM-146807	7
...	...	...
17410	GZ-144701408	1
17411	GZ-1447018	1
17412	GZ-1447027	1
17413	GZ-1447028	1
17414	ZD-2192596	1

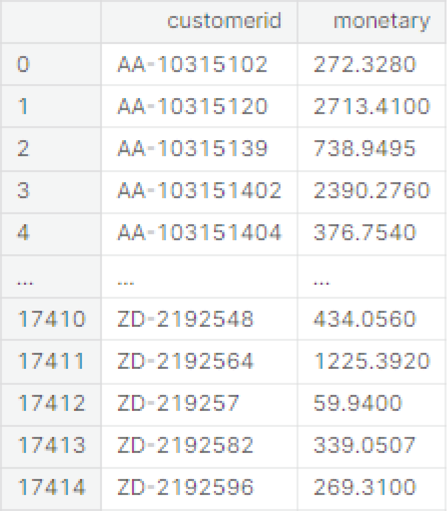
Figure 3.5: Value of Frequency (F)

The monetary value for each customer is shown in figure 3.6.

3.3.5 Calculating V with discount factored in

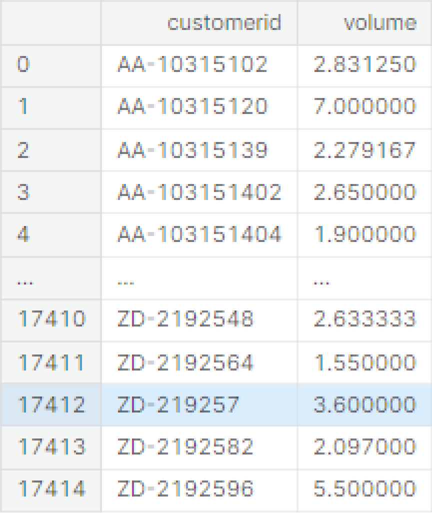
Quantity is how much product a customer has purchased. Volume finds out the number of purchased products over a set period and uses this to identify potentially profitable customer segments. For the proposed model, V$_D$ is a modified version of volume, which has discount factored in its equation. If Q denotes quantity, x denotes the number of purchases of a customer for a specific day, n is the number of days when transactions were made, and the discount is denoted by D, then V$_D$ is

$$V = \frac{\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{x}(Q_j\cdot(1-\text{discount}))}{x}\right)}{n}$$



	customerid	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

Figure 3.6: Value of Monetary (M)



	customerid	volume
0	AA-10315102	2.831250
1	AA-10315120	7.000000
2	AA-10315139	2.279167
3	AA-103151402	2.650000
4	AA-103151404	1.900000
...	...	...
17410	ZD-2192548	2.633333
17411	ZD-2192564	1.550000
17412	ZD-219257	3.600000
17413	ZD-2192582	2.097000
17414	ZD-2192596	5.500000

Figure 3.7: Value of Volume (V$_D$)

3.4 Relationship between volume with other attributes

It is necessary to consider all five LRFMV$_D$ characteristics when choosing profitable clients for superstores. The parameter should not overlap with each other thus establishing their uniqueness. The two additional variables L and V$_D$ have been added to the RFM model. A previous study has shown how L is related to other attributes and why it is crucial to take it into account when aiming to increase market segment profitability. [22] [16] [37] [34].

In figure 3.8 [36] a heatmap is created to show the link between the recently modified characteristic called volume and the other variables L, R, F, and M.

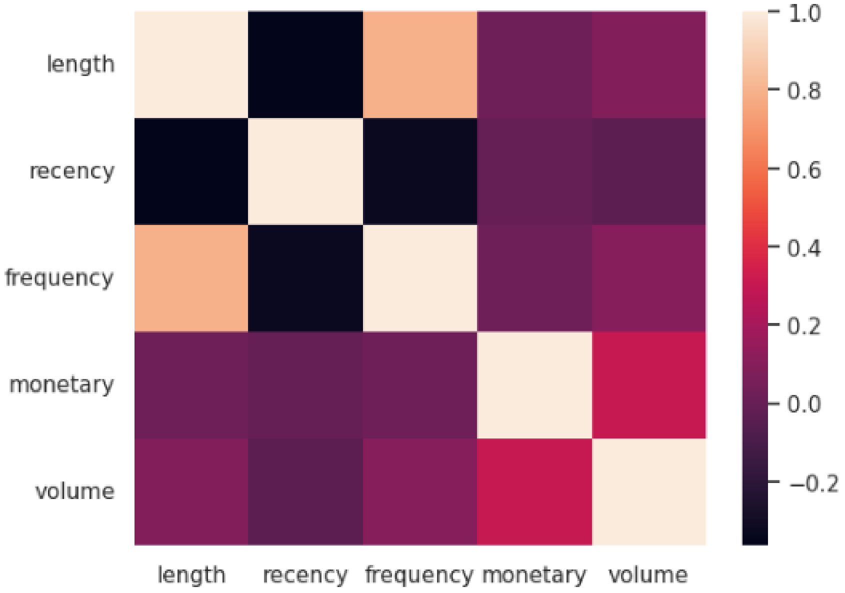


Figure 3.8: Heatmap (LRFMV$_D$)

The graph varies from light to dark or shows the association of firm to less (1.0 to -0.2). In contrast to L (0.2), R (0.0), F (0.2), and M (0.4), the newly added parameter V$_D$ is discovered to have a very weak association with each of them while having a significant link with itself.

3.5 Specification of Model

3.5.1 Elbow Method

An elbow method is an effective approach in order to determine the appropriate number of clusters. Thorndike initially raised the topic in 1953 [16] and introduced this widely used technique. The fundamental main concept of this tactic is choosing the elbow point of the cluster associated graph with error reduction which is conducted by increasing the K value until the gain of K is at a steady pace. This method uses a heuristic approach in order to determine the number of clusters[27].

$$\alpha_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i \tag{3.1}$$

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in S_k} \|x_i - \alpha_k\|^2 \tag{3.2}$$

Here x denotes a point an in the dataset which can be represented like this Let X = x1, x2, x3 ... xn, K refers to a number of clusters. Here K < n and K 1, 2, 3 ... K. k denotes the centroids of the related cluster S$_k$. It is the average position of all the data points of the selected cluster. First, it iterates through the data points that belong to the cluster S$_k$ and cumulates the data points within that cluster. The sum of data points is then divided by the number of elements present within that cluster and the result represents the centroid k of the cluster S$_k$. In order to calculate SSE, first it iterates through each of the clusters of S$_k$ and cumulates the squared distance of each data point xi and its centroid ${}_k$. Then by summing up all of the squared distances for all the clusters of K, we get the total SSE value.
The figure 3.9 [36] represents the graphical representation of the elbow method for the LRFMV$_D$ model. The results of computing SSE for each cluster are displayed on the y-axis together with K, which is displayed on the x-axis. When plotting the SSE, it starts to form a linear graph that begins to descend and starts to form an elbow-like curvature. We see this characteristic in the graph when the number of cluster values moves from 4 to 5 and shows a bend in the linear line. The graph gradually smoothness out as the number of clusters is increased from 5 to a higher value. If the value of k is too

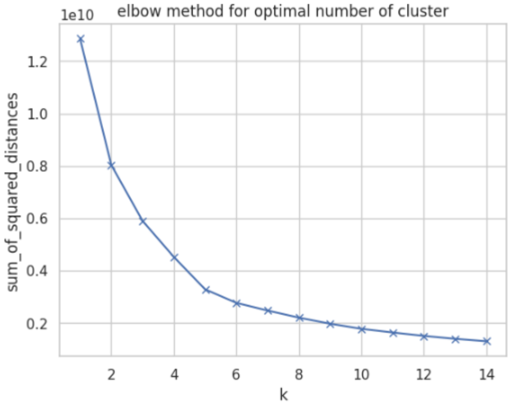


Figure 3.9: Elbow method (LRFMV$_D$ model)

low, numerous crucial clusters might be removed from the data. The elbow method is really simple to implement, it is intuitive, it provides a quantitative measure, and provides straightforward visualization through SSE value. But it has drawbacks as well and one of them is the subjective interpretation by different analysts as different analysts might consider different elbow points.

3.5.2 Silhouette Method

Silhouette Coefficient is a statistical approach in order to drive out the quality of the clustering results. It measures how each data point fits how well inside its cluster and its differences from its neighboring clusters. If the data points are labeled incorrectly then it shows a value close to a negative one, values are close to zero when there is a small space between neighboring clusters and the value is close to a positive one when the clusters are distinct and isolated. The silhouette coefficient is represented mathematically by the following

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{if} \quad x_i > 1 \tag{3.3}$$

where a d(i) denotes the typical distance of each data point inside the cluster which is between object i and other data points. And c(i) describes the average separation between the cluster to which object it stands and all other clusters to which object I do not belong, i.e., the separation between all of the clusters, with x standing for the number of the cluster. Silhouette Coefficients is a quantitative metric for measuring clustering quality that takes into account individual data point fit, nearby cluster distances, and a range from -1 to 1 for simple understanding. They are sensitive to cluster structure, depending on distance measurements, and have trouble with overlapping clusters. Additionally, their effectiveness is mostly limited to partition-based clustering methods like k-means rather than hierarchical or density-based methodologies. It is essential to consider the limitations of Silhouette Coefficients and make use of them in conjunction with other evaluation techniques for a full assessment of clustering quality.

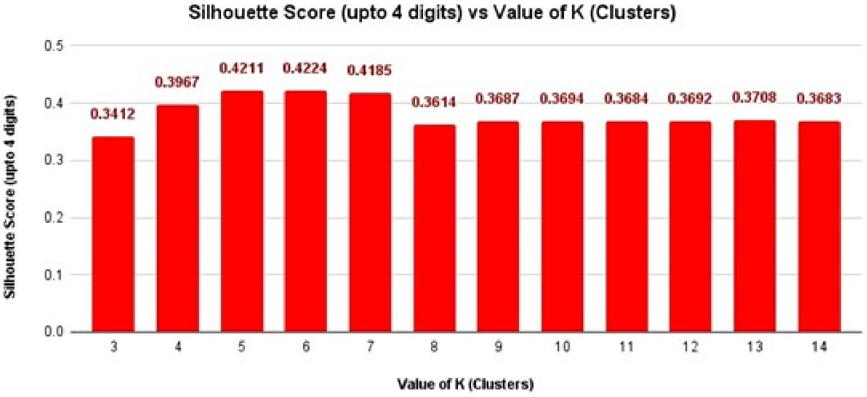


Figure 3.10: Silhouette result

When the Silhouette Coefficient is calculated, it is shown in figure 3.10 [36]that starting with cluster 7, the Coefficient value has begun to decline. Up to cluster 6, it was increasing, reaching 0.4224 for cluster 6 and 0.4185 for cluster 7. It denotes the coefficient value's downward trend, which lasts until cluster 14.

3.5.3 Cumulative Explained Variance Ratio

We must lower the dimensionality of the data set before we can simplify our suggested model. This involves selecting the right dimension for the data set in order to make our model simple to use and quick to execute. Principal Component Analysis (PCA) is a potent tool for this. Before starting PCA, first, we need to assess the value of our primary component. Utilizing the explained variance ratio, we understood the primary component that would be relevant to our model. This demonstrates the percentage of variance in the data that each element contributes to. The cumulative percentage is how the first n components explain variation.

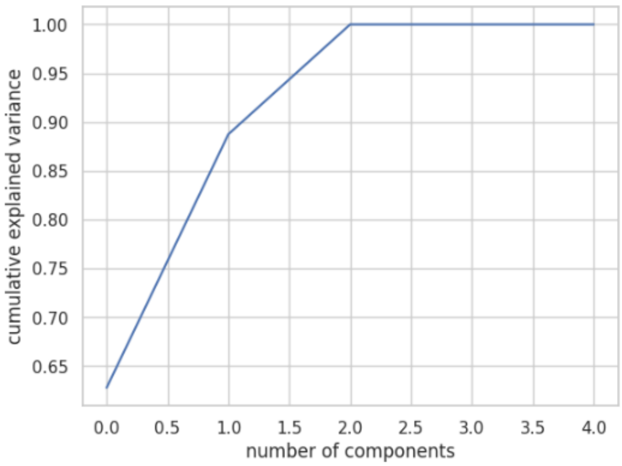


Figure 3.11: Cumulative explained variance for proposed data set

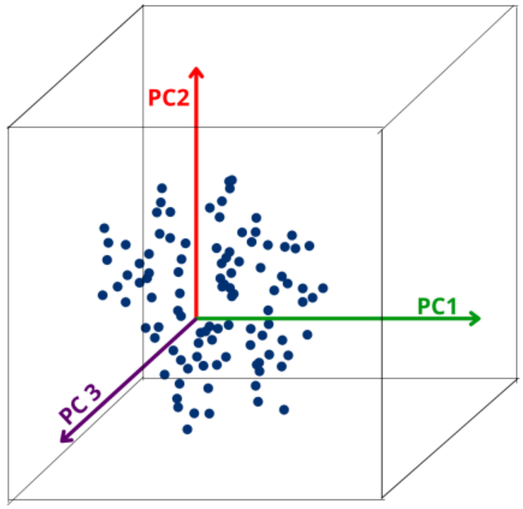
Here in figure 3.11 it seems that a little less than 2 components contain almost over 90 percent of the variance. So, if we can take a round figure like 2 which covers almost 100 percent of the variance.

3.5.4 PCA

PCA is a very useful method that is used in unsupervised learning models. Principal component analysis was first put forth by Pearson [1] in 1901. Then it was further improved by Hotelling [2] and coined the term ”principal components” to refer to its components.

With the aim of preserving the bulk of the pertinent data, it is a dimensional reduction technique that entails compressing a data set. It converts features from a higher dimension to a lower dimension subspace[1],[18].

As a result, dimension reduction removes unwanted variation from the original data set which helps to improve algorithm accuracy. PCA helps to create a linear mapping so that it can maximize the variation of original data set.



]

Figure 3.12: 3D Characteristics Dataset

In figure 3.12, some data points are scattered along the High dimensional 3D region. From figure 3.13 it can be visible that the PCA reduced the original data sets 3D into a 2D space by finding the maximum variance. Since our data sets had a lot of parameters and it was difficult to visualize, it was beneficial to apply PCA in order to reduce the dimension. it has helped to reduce the dimension without losing any valuable data. The original data set had five features L, R, F, M, V$_D$ which was reduced to these two parameters PC1 and PC2.

3.5.5 K-means

K-means is used to cluster data from a dataset, where the clusters contain a specific type of data. Elbow method, Silhouette coefficient approach, gap statistic, etc., can be used to get the best cluster numbers. For the K-means algorithm first, a centroid is created and inserted into an unlabeled dataset which can be random or real data from the dataset provided, All the data points that are closer to the centroid are grouped together after using a predetermined distance metric. The centroid's value will then be determined once more by calculating the average value for each group. This method will keep going unless criterion function Figure on the data set of the proposed model becomes minimum.

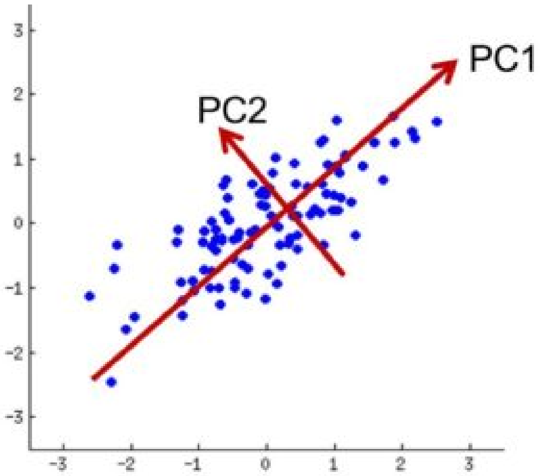


Figure 3.13: Conversion to 2D using PCA

Euclidean distance was used in this study. Euclidean distance can be defined using the following formula:

$$E_{ab} = \sqrt{\sum_{k=1}^{n}(a_k - b_k)^2}$$

Where the provided dataset is, X =x$_1$..........x$_n$ with endpoints and vectors a =a$_1$........a$_n$, b= b$_1$........b$_n$.
The following is a definition of the criterion function:

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

Here, C$_i$ stands for the cluster where α and αi are that cluster's average. The term 'SSE' stands for the total squared errors across all data set data points.

Chapter 4

Comparative Analysis

4.1 Determining Cluster Number for K-Means

To find the optimal cluster numbers for the RFM, LRFMV, and LRFMV$_D$ models elbow and silhouette methods were used. After using the Elbow method the RFM model showed that the value of k kept increasing, and the graph got flatter. The last k value for which the graph showed the most tangent was k=4. Also, the Silhouette score showed that a k = 4 the value was largest giving a score of 0.471724277212437. After that the values kept dropping, showing k = 5 with value. 4674035614222174 and k =6 with score .4332662159270996 and so on. So, RFM's cluster was determined to be 4.Figure 4.1 [36] shows the RFM model's four clusters after using k-means. Clusters were shown with different colors (cluster 0= green, cluster 1= blue, cluster 2= red, and cluster 3= purple).

For the LRFMV model figure 4.2, the ideal cluster number is 6. The silhouette score peaked at cluster 6 with approximately 0.422406. Afterward, the values kept dropping showing values, at k=7 (0.418570), at k=8 (0.361437) and so on. So, the cluster number was determined to be 6.

Because the LRFMV and LRFMV$_D$ (figure 4.3) models use the same number of parameters, the number of optimal clusters is the same for both. But the point to be taken from this is, although they have the same number of clusters, cluster values were different for every cluster of the two models which is true for both customer count and profit amount of those clusters.

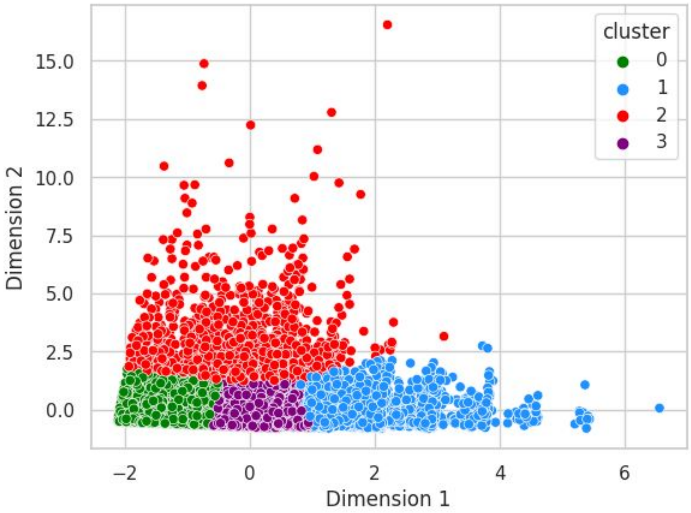


Figure 4.1: K-means (RFM model)

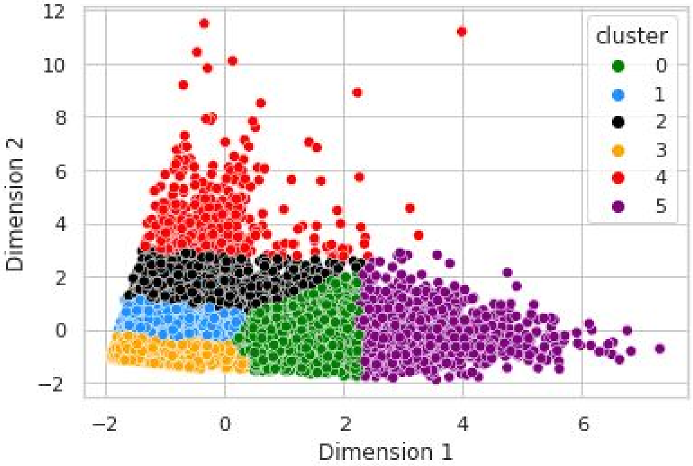


Figure 4.2: K-means (LRFMV model)

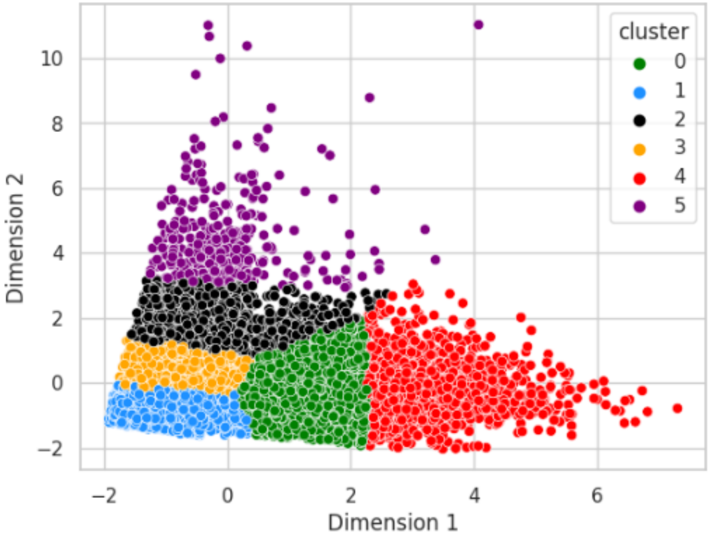
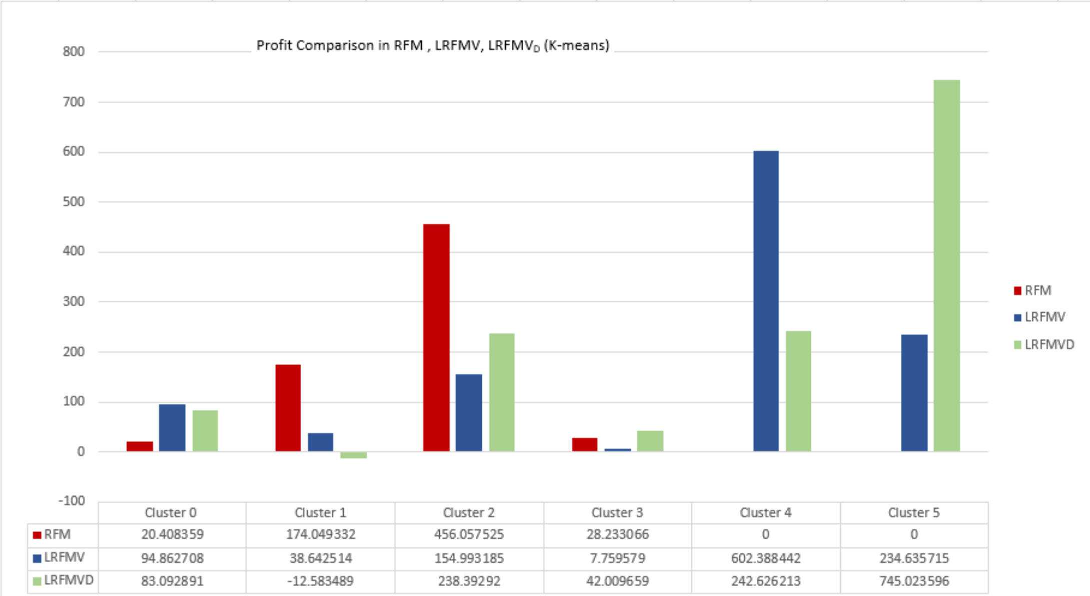


Figure 4.3: K-means (LRFMV$_D$ model)

4.2 Profit Analysis of the Models

For every cluster created after using k-means on RFM, LRFMV, and LRFMV$_D$ models, profit per head has been calculated.
The graph in 4.4 [36] demonstrates the profit per head for RFM, LRFMV along with LRFMV$_D$ models. Different colors have been used to represent the three models. "red", "blue" and "green" have been used to represent RFM, LRFMV, and LRFMV$_D$ models respectively. The red bar denotes the RFM model with four groups, each with its own profit per head. Cluster 2's customers generate the most revenue numbering at 456.0575, and the least revenue is about 20.4083, and cluster 0. The total mean profit for the RFM model is about 678.748.

It is visible that the RFM models' profit from Cluster 2 is greater than LRFMV$_D$ models' Cluster 2. This is because the RFM model has fewer clusters than the LRFMV$_D$ model. RFM model clustered customers together who are represented in different clusters in the LRFMV$_D$ model. Because the RFM model has fewer clusters, it has to cluster more customers together with very little similarity to encompass the entirety of customers. These customers were clustered properly in the LRFMV$_D$ model.



	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
🟥 RFM	20.408359	174.049332	456.057525	28.233066	0	0
🟦 LRFMV	94.862708	38.642514	154.993185	7.759579	602.388442	234.635715
🟩 LRFMVD	83.092891	-12.583489	238.39292	42.009659	242.626213	745.023596

Figure 4.4: Profit analysis for RFM, LRFMV and LRFMV$_D$ model

In the graph, blue bars indicate the 6 clusters of the LRFMV model. The most profit achieved is by cluster 4's customers, which is about 602.3884 and the lowest profit is 7.7596 generated by the customers of cluster 3. Total mean profit for the LRFMV model is equal to 1133.352.

Green denotes the LRFMV$_D$ model in the graph. The maximum profit is achieved by cluster 5 with 745.0236 profits per head. The least profit is about -12.5834 which is achieved from cluster 1. This means that customers from cluster 1 procured a loss for the superstore. This will be explained thoroughly in a later chapter. This also implies that customers from cluster 5 are essential for generating more profit for the superstore. The total mean profit for the LRFMV$_D$ model is about 1338.653.

By analyzing the models, it becomes apparent that the LRFMV and LRFMV$_D$D models have more variance of clusters than the RFM model. Comparing the profit per head of the LRFMV and LRFMV$_D$ models, it becomes clear that the LRFMV$_D$ model generates the highest profit of the two models. It is about 23 percent higher than the LRFMV models' highest profit.

The LRFMV$_D$ model has the highest profit of 745.023596 which is more than the RFM and LRFMV models. It also has the maximum profit when all clusters' profit per head is added together which is about 1338.563. This implies that for RFM the clustering is not up to par with the other models and a primary clustering parameter is not taken into consideration.For the LRFMV model, a key parameter is missing when calculating volume which results in it having the same number of clusters as the LRFMV$_D$ model but not the maximum profit per head. These ultimately result in the RFM and LRFMV model not being able to locate valuable customers and construct marketing strategies correspondingly

Further analysis also reveals that among the three models, the LRFMV$_D$ model was the only one that identified which customer segments caused the superstore to make a loss. This can be found in the cluster 1 of the model. In the case of LRFMV$_D$, cluster 1 customer incurs a loss for the superstores. The phenomena of cluster 1 will be explained in a later chapter.

Chapter 5

Result Analysis

5.1 Statistical Analysis

5.1.1 Correlation between Profit and Volume

The objective of the LRFMV$_D$ model was to improve upon the volume of the LRFMV model to ensure better segmentation, and identification of potential, valuable, and more profitable customers. Volume when calculated by factoring in discount yields better results than calculating volume without discount.

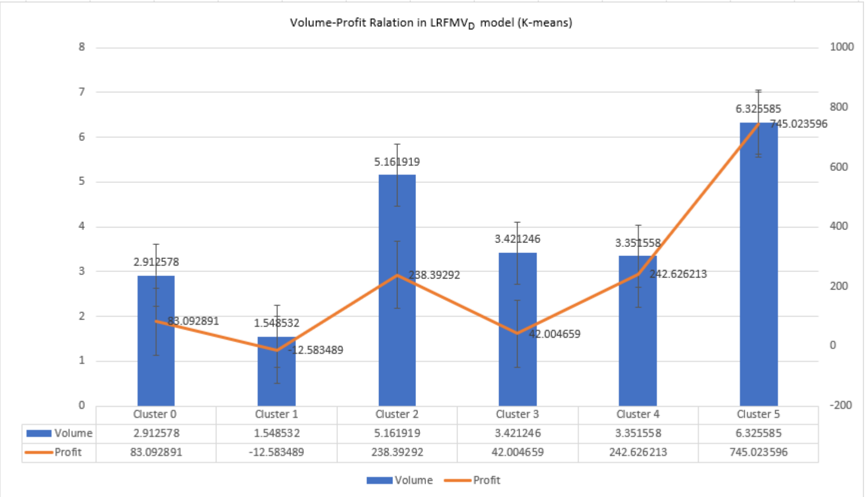


Figure 5.1: Volume-Profit relation of LRFMV$_D$ model

in Figure 5.1 [36], Volume-Profit correlation has been shown and each cluster shows the volume of customers' purchases. Upon analyzing the figure, it is clear that volume positively correlates to profit. Cluster 5 has the highest volume and profit, numbering at about 6.33 and 745.02 respectively. The volume of cluster 4 is about 3.35 and the profit is about 242.63. Cluster 2 also shows the same characteristics as Cluster 5 with high volume and profit numbering at 5.16 and 238.39 respectively. This further reinforces the hypothesis that more volume positively correlates to greater profit. This research can be used to find out how many products businesses need to sell to gain a certain revenue.

5.1.2 Profit Analysis for each Segment

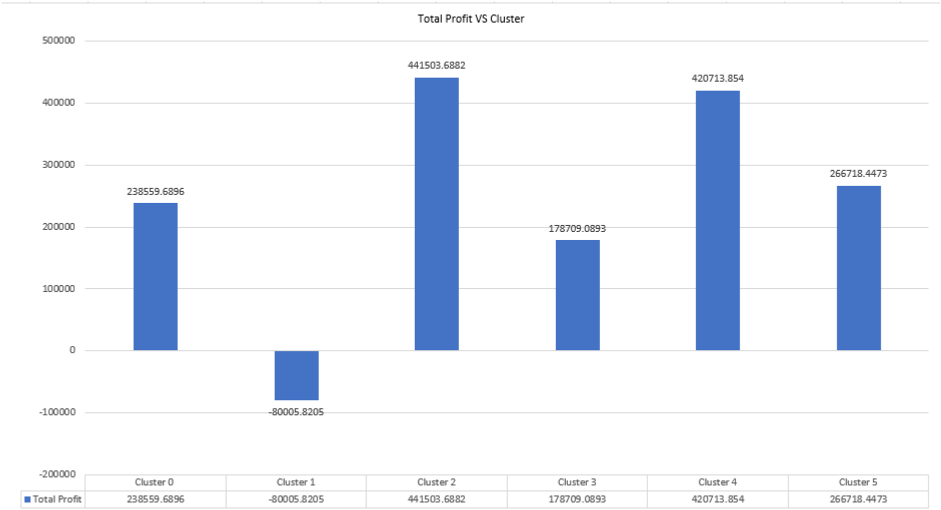


Figure 5.2: Profit of each cluster

The bar chart in figure 5.2 [36] shows the profit for each cluster. Most of the profit has been made by clusters 2, 4, and 5.

Upon analysis, it is visible that Cluster 2 has less profit per head than Cluster 4. But as was previously shown, cluster 4 has less volume than cluster 2. This phenomenon can be explained through discounts. Customers of Cluster 2 purchased more products than that of Cluster 4, but they purchased more discounted products than Cluster 2's customers. That's why purchasing more products did not result in more profit for Cluster 2's customers.

5.1.3 Each Clusters' Customer Number Analysis

Six different clusters have been discovered by using K-means. All of these clusters have different customer counts, and these clusters will be used to segment customers with the help of L, R, M, and V_D parameters.

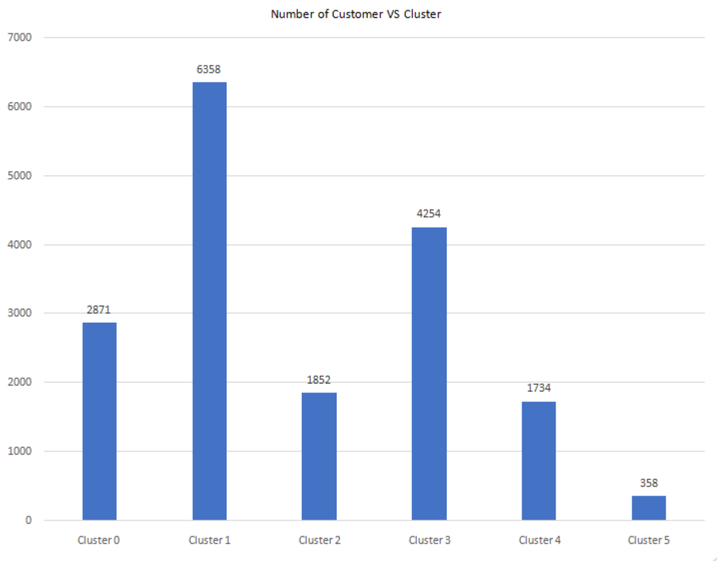


Figure 5.3: Each cluster customer count

Figure 5.3 [36] shows the customer count of individual clusters. Cluster 1 has the most customers, numbering 6358. The smallest one is cluster 5 having 358 clients, and the second smallest is cluster 4 having 1734 clients. Cluster 5 has the lowest customer count with the highest profit margin, and clusters 2 and 4 have comparable customer count with comparable profit margins.

5.2 Client Segmentation with Client Classification Matrix

5.2.1 Profit Generation of Clusters

A classification matrix is necessary to effectively check customer profitability. This matrix helped to group the customers into four different segments. Passive customers are the most valuable ones as they provide huge profits at a low cost. Carriages trade customers also provide big profits but have an expensive serving cost. Some consumers are easy to satisfy but don't contribute to profit generation and are classified as bargain basements. Finally, there are the aggressive customers, who provide very little profit or even a loss.
Aggressive- - Wants the best product with the best service at minimum cost.

Bargain basement – Bargain basement clients are price-conscious. They require a lower service cost than carriage traders.

Passive –They provide the greatest profit to the firm. Passive customers are willing to pay high prices, so they are easy to serve.

Carriage trade - – These customers provide big profits but at a big serving cost. So, lowering these serving costs should be the company's focus.

Table 5.1: Customer classification

Category	Revenue	Expense of Serving
Passive	High	Low
Carriage Trade	High	High
Bargain Basement	Low	Low
Aggressive	Low	High

All customers are not equal when it comes to profit generation. Some are more profitable than others. Some, on the other hand, procure a loss. In most cases, businesses are unable to distinguish between them resulting in them having a loss. This puts an emphasis on customer segmentation and the use of a customer classification matrix. The table 5.1 gives an overview of the customer classification matrix.

Customers of clusters 2 and 5 are classified as Passive customers. Passive customers have huge profit generation and low serving costs. Clusters 2 and 5 show the traits of length, recency, frequency, monetary, and volume when compared to an average of these parameters of the whole dataset: Length goes down, Recency goes up, Frequency goes down, Monetary goes up, and Volume also goes up. Though cluster 5 has the most profit, it has the lowest number of customers.

Cluster 4's customers are defined as Carriage traders. They can generate bigger profits but at the cost of expensive serving costs. Compared to passive customers they show the following traits: Length goes up, Recency goes down, Frequency goes up, Monetary goes up, and Volume goes up. After passive types, they are the most profitable customer types.

Reducing their service cost can turn them into passive types, so this should be the focus of superstores when it comes to carriage traders. Cluster 0 has been classified as a Bargain basement. They show these traits: Length goes up, Recency goes down, Frequency goes up, Monetary goes down, and Volume goes up.

Cluster 1 and Cluster 3 have the greatest number of customers with the lowest amount of profit. They are categorized as Aggressive customers. The customers of cluster 1 caused the superstore to make a loss. Cluster 3 has the second-lowest profit margin. Cluster 1's traits are as follows: The length goes down, Recency goes up, Frequency goes down, Monetary goes down, and Volume goes down. Traits of cluster 3: Length goes down, Recency goes up, Frequency goes down, Monetary goes down, Volume goes up. It is apparent that these two clusters share almost every trait except for volume which is depicted in cluster 3 having a profit and cluster 1 making a loss.

5.2.2 Attribute Recognition

Each component of LRFMV_D was calculated separately beforehand. The average of profit and LRFMV_D variables were calculated for the whole dataset.

The mean of all parameters was also calculated for every cluster.

Mean of LRFMV$_D$ variables for the whole dataset (processed): Length= 181.12, Recency= 507.31, Frequency= 1.47, Monetary= 481.94, Volume= 2.89, Profit= 84.13.

The average value of cluster components where compared with the hold datasets' average components values and with these values some traits were identified. With the help of these traits the following table 5.2 was made.

Table 5.2: L, R, F, M, V$_D$ traits of Customer Segmentation

Customer Type	L	R	F	M	V$_D$
Bargain	up	down	up	down	up
Aggressive	down	up	down	down	up or down
Carriage	up	down	up	up	up
Passive	down	up	down	up	up

Chapter 6

Conclusion

6.1 Overview

This contemporary study compares the RFM model with the proposed LRFMV$_D$ model which uses volume with discount and length. It also compares the LRFMV model with the proposed model. We attempted to prove that the LRFMV$_D$ model produces better outcomes when compared with the RFM model in terms of profit per head. We also tried to demonstrate that calculating volume with discount, and profit per head is better than the LRFMV model. We investigated customer behavior with only sales data and no personal data of customers were used. No predetermined score was assigned as the equations were automated in nature.

The values used for the customer classification matrix were obtained through the K-means algorithm. After that, a Volume-Profit correlation was shown along with individual profit and customer count analysis for each cluster. When the same algorithm was used on RFM and the LRFMV model, the LRFMV$_D$ model showed that it had the maximum profit per head for most of the clusters.

6.2 Contribution

There are several studies have been done on customer segmentation but only a little number of studies can show a relation between customers' product quantity and discounts together. The model agreement came up with shows a strong relation between profit per head as well as purchased products with discounts. This proposed prototype shows an effective cluster while segmenting customers on discount of volume product.

Because of the effective segmentation most profitable customers can now be easily identified and marketing can be targeted catering to those customers.

6.3 Suggestion and Future Work

A problem we faced when conducting the study was the secrecy of data and very few dissimilarities in data points. This hampered our work to use the discount as a clustering parameter and how it affected the other parameters such as L, R, F, and M. It is why only volume could be shown with a discount. Future work can be done on this by implementing discounts with all the clustering parameters. We hope this will improve the clustering quality even more and will allow for more efficient customer segmentation, and from that research profit and discount correlation can also be derived. Consequently, it will be feasible to comprehend consumer behavior better and help organizations in differentiating between customers more effectively.

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] A. Hughes, *Database Marketing.* Probus Publishing Company, 1994.

[4] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. John-son, H. Kaushansky, and A. Software, “Churn reduction in the wireless industry,” in *Advances in Neural Information Processing Systems*, vol. 12, 2000, pp. 935–941.

[5] N. Hsieh, “An integrated data mining and behavioral scoring model for analyzing bank customers,” *Expert Systems with Applications*, vol. 27, pp. 623–633, 2004.

[6] 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,” *Information Systems Frontiers*, vol. 7, pp. 421–434, 2005. DOI: 10.1007/s10796-005-4812-6. [Online]. Available: https://doi.org/10.1007/s10796-005-4812-6.

[7] B. Sohrabi and A. Khanlari, “Customer lifetime value (clv) measurement based on rfm model,” *Iranian Accounting and Auditing Review*, vol. 14, no. 47, pp. 7–20, 2007.

[8] S. Li, L. Shue, and S. Lee, “Business intelligence approach to supporting strategy-making of isp service management,” *Expert Systems with Applications*, vol. 35, pp. 739–754, 2008.

[9] S. Lumsden, S. Beldona, and A. Morison, “Customer value in an all-inclusive travel vacation club: An application of the rfm framework,” *Journal of Hospitality & Leisure Marketing*, vol. 16, no. 3, pp. 270–285, 2008.

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

[11] T. Jiang, A. Tuzhilin, and March, “Improving personalization solutions through,” in *Proceedings of the conference*, 2009.

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

[13] H.-H. Wu, E.-C. Chang, and C.-F. Lo, “Applying rfm model and k-means method in customer value analysis of an outfitter,” in *16th ISPE International Conference on Concurrent Engineering*, 2009, pp. 665–672.

[14] Y. Li, L. CH, and L. CY, “Identifying influential reviewers for word-of-mouth marketing,” *Electronic Commerce Research and Applications*, vol. 9, pp. 294–304, 2010.

[15] 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,” *Journal of Statistics and Management Systems*, vol. 14, no. 2, pp. 267–276, 2011. DOI: 10.1080/09720510.2011.10701555.

[16] 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,” *Expert Systems with Applications*, vol. 39, no. 5, pp. 5529–5533, 2012.

[17] Y. Cho and S. Moon, “Weighted mining frequent pattern based customer’s rfm score for personalized u-commerce recommendation system,” *International Journal of Computer Science and Information Security*, vol. 12, no. 12, pp. 106–111, 2014.

[18] A. Sarveniazi, “An actual survey of dimensionality reduction,” *American Journal of Computational Mathematics*, vol. 4, pp. 55–72, 2014. DOI: 10.4236/ajcm.2014.42006.

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

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

[21] J. Li, S. Song, Y. Zhang, and Z. Zhou, "Robust k-median and k-means clustering algorithms for incomplete data," *Mathematical Problems in Engineering*, vol. ID 4321928, p. 8, 2016. DOI: 10.1155/2016/4321928.

[22] 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," *Procedia Computer Science*, vol. 120, pp. 23–30, 2017.

[23] S. Peker, A. Kocyigit, and P. Eren, "LR-FMP model for customer segmentation in the grocery retail industry: A case study," *Marketing Intelligence & Planning*, vol. 35, no. 4, pp. 544–559, Jun. 2017.

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

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

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

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

[28] S. Monalisa, "Analysis outlier data on RFM and LRFM models to determining customer loyalty with DBSCAN algorithm," in *International Symposium on Advanced Intelligent Informatics (SAIN)*, 2018, pp. 1–5. DOI: 10.1109/SAIN.2018.8673380.

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

[30] 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," in *2018 IEEE 15th International Conference on E-Business Engineering (ICEBE)*, 2018.

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

[32] J. Nagesh, "Generating political scores using RFM model and cluster prediction by XGBoost," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 1333–1336. DOI: 10.1109/CSCI49370.2019.00249.

[33] T. Segal. (Jul. 2019). "Inside recency, frequency, monetary value (rfm)," [Online]. Available: https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp.

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

[35] M. R, I. N, T. M, E. MAF, C. SA, and A. MGR, "Lrfmv: An efficient customer segmentation model for superstores," *PLoS ONE*, vol. 17, no. 12, e0279262, 2022. DOI: 10.1371/journal.pone.0279262. [Online]. Available: https://doi.org/10.1371/journal.pone.0279262.

[36] M. A. Sadhin, K. M. Sagor, I. Jahan, and R. K. Prottay, “Lfrmvd: A customer segmentation model,” Unpublished master’s thesis, M.S. thesis, BRAC University, 2023. [Online]. Available: https://github.com/masrurarefinsadhin/LRFMVd.git.

[37] Soeini and E. Fathalizade, *Customer segmentation based on modified RFM model in the insurance industry.*