Customer Segmentation Based on RFM Analysis & Its Impact on Business Strategy Analysis

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering of the University of Asia Pacific

Presented by

Md. Efti Khirul Alam
Roll: 18101014
Tanveer Ahamed Rabby
Roll: 18101025
Sharmin Akter
Roll: 18101065

-

Supervised By

Shammi Akhtar
Assistant Professor
Department of CSE
University of Asia Pacific

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY OF ASIA PACIFIC

June 2022

Certificate of Approval

We hereby recommend that the thesis prepared by Md. Efti Khirul Alam, Tanveer Ahamed Rabby, Sharmin Akter entitled “Customer Segmentation and Its Impact on Business Strategy Analysis” is accepted as fulfilling the requirements for a degree of Bachelor of Science in Computer Science and Engineering.

Shammi Akhtar
Assistant Professor
Department of Computer Science and Engineering
University of Asia Pacific (UAP)

Chairman of the Committee
(Supervisor)

Molla Rashied Hussein
Assistant Professor
Department of Computer Science and Engineering
University of Asia Pacific (UAP)

Member of the Committee
(External)

Dr. Md. Rajibul Islam
Assistant Professor
Department of Computer Science and Engineering
University of Asia Pacific (UAP)

Head of the Department

DECLARATION

We, hereby, declare that the work presented in this Thesis is the outcome of the investigation performed by us under the supervision of **Shammi Akhtar, Assistant Professor, Department of Computer Science and Engineering, University of Asia Pacific**. We also declare that no part of this Thesis and thereof has been or is being submitted elsewhere for the award of any degree or Diploma.

Countersigned
(**Supervisor**)

(Shammi Akhtar)
Assistant Professor
Department of CSE
University of Asia Pacific

Signature
(**Candidates**)

(Md. Efti Khirul Alam)
ID: 18101014
Department of CSE
University of Asia Pacific

(Tanveer Ahamed Rabby)
ID: 18101025
Department of CSE
University of Asia Pacific

(Sharmin Akter)
ID: 18101065
Department of CSE
University of Asia Pacific

ACKNOWLEDGEMENTS

To begin with we would really like to thank the almighty ALLAH. These days we're successful in finishing our work with such ease because He gave us the potential, danger, and co-operating supervisor.

We would really like to take the opportunity to express our gratitude to **Shammi Akhtar, Assistant Professor, (CSE, University of Asia Pacific)** our respected supervisor. She gave us proper guidance and valuable advice. Her limitless persistence, scholarly guidance, persistent encouragement, regular and lively supervision, positive criticism, precious advice, studying many inferior drafts and correcting them in any respect have made it possible to finish this thesis in addition to feedback and guidance helped us in getting our thesis document.

We would like to thank **Molla Rashied Hussein, Assistant Professor, (CSE, University of Asia Pacific)** our respected teacher, who stimulated us in every step and helped us in some ways by imparting numerous resources and moral support. We are also thankful to our instructors who helped us in some ways by means of presenting various assets and ethical help.

Remaining of all we are thankful to our families; who're constantly with us in each step of our lives.

ABSTRACT

At present each business platform has grown to be digitalised. by implementing diverse regulations and methods a person could make their business coverage smarter. Consumer segmentation is the system which separates clients into businesses primarily based on a certain attribute (e.g. character, interests, habit) and elements (e.g. demography, industry, income). nowadays corporations or marketers may focus on a way to attract customers and maintain them responsible. The most crucial aspects of a powerful enterprise are innovation and understanding what clients need. Based on the RFM (Recency, Frequency, and financial) values of the purchasers, an organization's clients are effectively segmented into businesses with comparable behavior. An enterprise's transactional statistics is evaluated over a time period. Segmentation affords a clear photo of the client's wishes and aids in the identity of the organization's capacity customers. The revenue of the company is also elevated by way of segmenting the purchasers.it is the concept that keeping existing customers is more vital than finding new ones. To retain clients, as an instance, the company might use advertising and marketing techniques tailor-made to a positive section. The transactional records are first subjected to an RFM evaluation, and then the statistics is clustered using well known k-means in particular. In this analysis we successfully segmented the customer where we analyze customer data & information for an organization's dataset and then segment consumers regarding to RFM score, including generate business decision what will be the approach over this segmented customer, with this find out the most selling item from this dataset which will further needed for recommendation or promotional purposes in order to benefit both the customer and the business and also evaluate k-means and hierarchical clustering model regarding to Silhouette coefficient & Davies-Bouldin index method.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Introduction

1.1 Introduction

In the present business environment has moved for customer equity. To generate sales, every company is focusing on the concept of customer, how loyal the customers are and how they increase their profit. Every individual has their own variety of needs and wants. Nowadays companies apply many segmentation techniques to understand better customer behavior. segmentation is important to promote necessary and underselling products or services to higher valued customers. If we find a large customer's dataset there is a possibility to generate some patterns where we identify the customers' purchase behavior for a particular company. The real dataset is too big and complex so it is not easy to make a decision about what are the customers' necessities and how to increase the satisfactory level at a first glance. This is where customer segmentation is much needed to determine the targeted customer. The details of future buying patterns of the customer are determined with the techniques of analyzing customer value and their past purchasing records. RFM Analysis is a technique to know customer values. In RFM matrices include three parameters which are recency, frequency and monetary. These parameters are used for determining the value of each customer.

1.2 Motivation

As the buying pattern or searching pattern for a product differs from person to person, it is not an easy task to treat all the customers in the same manner. Nowadays Customers have so many options to fulfill their demands**.** The motivation here is to segments customers and generate a business solution from a particular organizations dataset. At present there is a large competition among all the organizations so they need to make proper arrangements for attracting customers and making profit. According to the Pareto principle (Srivastava, 2016), 20% more revenue of the company is acquired from the contribution of retaining customers than funneling in new customers. Many segmentation methods are established but the purpose of calculating RFM scores on our online retail dataset is to divide the customer into different segment and make decisions for each segments particularly that what should the organization do for each of them for generate their profit and also fulfill their customer satisfaction. From the segmentation it provides an opportunity build a plan which will help to take actions such as whether or not to deliver a promotional offer to the customer or to push the customer with new marketing.

1.3 Customer Segmentation

1.3.1 Basics

Client segmentation is the practice of partitioning an organization's clients into companies that replicate similarity amongst customers in each organization. The goal of segmenting clients is to decide how to narrate to clients in each segment so as to maximize the price of each purchaser to the enterprise. Segmentation gives a simple way of organizing and dealing with your employer's relationships along with your clients.

1.3.2 Types of Customer Segmentation

We are focusing mainly four types of customer segmentation, which are

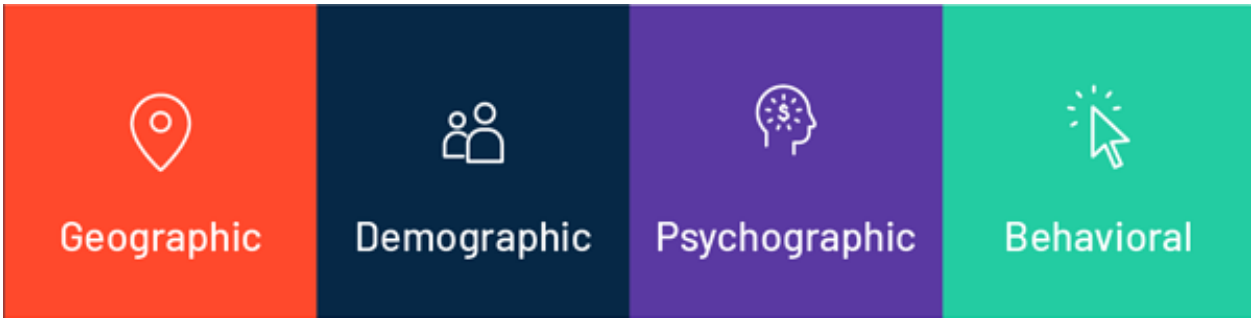

Fig 1.1: Customer segmentation types

- **Geographic Segmentation**
  Based on the geographic limitations, this consists of developing one of a kind agencies of clients. The desires and pastimes of capability clients vary in line with their geographic vicinity, climate and vicinity, and know-how this allows you to decide in which to sell and put it on the market an emblem, in addition to in which to amplify a commercial enterprise.
- **Demographic Segmentation**
  It includes dividing the market via one-of-a-kind variables along with age, gender, nationality, schooling degree, circle of relative's size, career, income, and many others. That is one of the most extensively used sorts of marketplace segmentation, when you consider that it's far based on knowing how customers utilize your services and products and what kind they are willing to pay for them.
- **Psychographic Segmentation**
  Which consists of assembling the target audience based totally on their behavior, lifestyle, attitudes and curiosity. To understand the clients, market analysis strategies which include focus companies, surveys, meetings and case research may be a success in compiling this kind of end.

- **Behavioral Segmentation**
  It specializes in precise reactions, like the client behaviors, styles and the way customers go through their decision and buying techniques. The point of view the public has towards your company, the manner they use it and their focus are examples of behavioral segmentation. gathering this form of information is just like the way you would locate psychographic facts. This lets entrepreneurs increase a more focused technique.

1.3.3 Benefits of Customer Segmentation

Customer segmentation brings it easier for marketing teams to develop highly targeted and effective marketing campaigns and plans. Beneath we've mentioned several benefits which exist with knowledge and defining client segments.

- **Greater organization recognition**
  While an organization has recognized particular marketplace segments, it enables them to be aware of what segments they need to target with precise merchandise/ services/ content material/ blogs and campaigns. While an agency has a focal point on particular segments, they ensure they are concentrated on the right segment with the right product that allows you to see the finest ROI.
- **Better serve a client's want and desires**
  Having described segments enables groups to fulfill a variety of purchaser needs by means of supplying special bundles and incentives. exceptional forms and promotional activities might be used for distinct segments primarily based on that phase's desires/ needs and characteristics.
- **Market competitiveness**
  whilst an enterprise is specializing in a selected part, their market competitiveness will increase. Which in flip will cause an advanced return On investment(ROI). The employer is targeting precise components and learns the whole lot they need to understand about that element, to market their merchandise to them.
- **Marketplace Enlargement**
  With geographic segmentation as mooted ahead, request growth is viable immediately. whilst a business enterprise is familiar with their parts and how to sell to an element in a specific role, they can enlarge incontinently into some other close by location. If segmentation is grounded on demographics, also once the organization knows their demographic part they are able to enlarge in that part with comparable products.
- **Targeted communication**
  Even if product functions and benefits are the same, it's far crucial for corporations to goal segments with unique conversation. As an example, if your phase becomes senior engineers, they may reply higher to technical facts about a product inside the shape of white papers or infographics, however a challenge manager may reply higher to information

concerning fee financial savings, efficiencies and many others in the shape of a blog, case examine or video. Messaging may be distinct for distinctive segments. platforms which can be used to goal exceptional segments might be specific additionally. The secrets to recognize your segments and target conversation applicable to them at the applicable systems.

1.3.4 Customer Segmentation Model

Here we mention some models which are popular for segmenting customers.

- **Demographic**
  At a bare minimum, many groups perceive gender to create and deliver content material primarily based on that consumer phase. Further, parental fame is another important segment and may be derived from buying details, asking extra records from clients, or obtaining the facts from a third party.

- **Recency, frequency, monetary (RFM)**
  RFM is a technique used frequently inside the junk mail segmentation space wherein you perceive clients based on the recency of their closing purchase, the overall wide variety of purchases they have made (frequency) and the quantity they've spent (financial).

- **High-value customer (HVCs)**
  Based totally on an RFM segmentation, any commercial enterprise, regardless of sector or industry, will want to recognize more approximately in which HVCs come from and what traits they proportion so you can accumulate greater of them.

- **Customer status**
  At a minimum, most organizations will bucket customers into active and former, which shows while the remaining time a purchaser made a buy or engaged with you.

1.3.5 RFM Segmentation Model

RFM analysis is a proven marketing model for behavior based customer segmentation, which we are currently choosing for our customer data analysis for segment customers and finding targeted customers. RFM segmentation model consists of three elements which are shown in the following Figure.

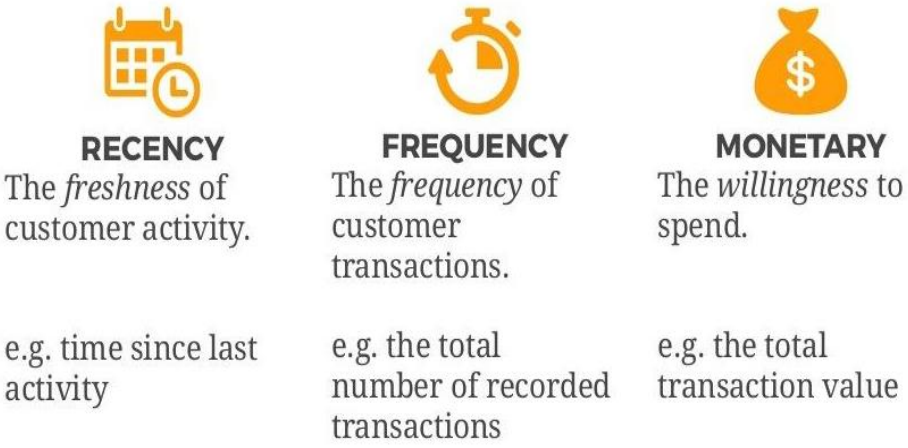

Fig 1.2: RFM Metrics elements

R→ Recency	How long has it been since the customer's last purchase? Activity is usually a buy, despite the fact that versions are every now and then used, assume the closing goes to an internet site or use of a cellular app.
F→ Frequency	How often has the customer made a purchase over a defined period of time? Sincerely, customers with common activities are more engaged, and in all likelihood greater loyal, than customers who do not often do so. And one-time-handiest customers are in a category on their own.
M→ Monetary	How much money did the customers spend with us over a defined period of time? Additionally, known as a "Monterey price," this element reflects how much a consumer has spent with the emblem throughout a selected time frame. large spenders must normally be handled in another way than customers who spend little.

Table 1.1: RFM Analysis

The total shape of RFM states that --Recency, frequency, financial. RFM price is an advertising analysis tool used to discover an organization's or a corporation's first-rate customers through measuring and analyzing their spending behavior. RFM evaluation is a way used to categorize customers consistent with their shopping behavior. An RFM evaluation evaluates clients and clients by way of scoring them in three classes: How these days they've made a purchase, how

regularly they purchase, and the scale in their purchases. RFM is a more efficient approach instead of conventional approach. Here we point out a few variations among RFM and traditional segmentation

RFM Model of Customer Value	Traditional Segmentation Methods
Built on historical transactions between user and the business	Built on consumer understanding and research studies commissioned by business
Uses R,F and M variables of customer data	Uses demographic and psychographic variables
Analyses the entire population	Analyses representative sample sets
No need to create curated sample sets	Requires careful selection of representative sample
Dependent on efficient and accurate data	Dependent on skill researchers
No scope for human error	Scope for human error

Table 1.2: RFM Segmentation Vs Traditional Segmentation

1.4 Purpose of the Study

There are numerous varieties of requirements and wants for a consumer. Companies are using a lot of different segmentation techniques to select the groups of customers and offering them better services also satisfy them with fulfill their necessities. There is also possibility to make the non-profitable groups into profitable customers. Without proper segmentation all of our promotional and advertising work will be worthless if we don't have proper segmentation guidelines. The purpose of the studies is to segment the customer in the proper way using RFM score.

1.5 Objective

This research is machine learning based on certain specifications of segmenting customers. It will try to guess the most loyal customer using various algorithms and models.

- To identify different groups within your target audience so that you can deliver more targeted and valuable messaging for them
- Collect the consumer facts
- To assist advertising and marketing people in meaningful customer segmentation
- Clustering customers into different groups helps decision-makers identify market

segments more clearly and thus develop more effective marketing and sale strategies for customer retention

- Clustered clients into segments consistent with RFM and extended RFM parameters the use k-means set of rules
- Technique that the perfect clients with appropriate advertisement, promotions and offers, centered clients are located with an in-intensity analysis at the clusters.

1.6 Expected Output

Outcome of the analysis in this work is to help advertising humans in meaningful consumer segmentation. Thought out the study we focus on analyzing the customer data, clustering model evaluation, calculating the RFM score for dividing the customer into different groups. Business organizations identify market segments more clearly with the help of clustering customers into different groups and create a better marketing approach for retaining customers.in this analysis we cluster customer's segments with RFM score then view the clustering points with the help of K-means algorithm. From the segmentation result we figure out some decisions which will be applied over the customers with proper promotions and offers.

Chapter 2
Literature Review

2.1 Terminologies

This section contains the basic terminologies which we use throughout the analysis. It includes customer segmentation, RFM analysis, RFM score, K means & hierarchical clustering method and also silhouette, Davies-Bouldin score analysis for model evaluation. So at first we need to understand the customers and organizations relationship. For increase the profit and influence the desired customer a company need to predict the customer purchase behavior with help of customer past data or information, where the term customer segmentation helps to divide customer into some particular groups and which will help a company what approach they will use to make customer happy also make more profit for the company. The term RFM means Recency, frequency and monetary value. Recency means a customer's buying interval time between two purchase dates. If the recency time is short that means the customers visit the particular company on a regular basis. Frequency indicates the number of purchases in a sudden time for a particular company, where the loyal customers are more frequent that means they are the top valued customers for the company. Then, the term monetary describes the total amount a customer spends in a certain time period. If a customer spends more for a particular company that indicates the company gains more profit from this customer. RFM score is a term where we put value in the R, F, M depending on the characteristics of the company. Different values or weights assigned to RFM metrics (from 1 to 5) are called RFM scores. In this analysis we try to develop machine learning techniques to analyze the customer information and which will solve a difficult problem. Lots of machine learning techniques but here an unsupervised machine learning is applied to identify different groups which is known as clustering. The main purpose of clustering is to create some groups of customers to identify some patterns which are needed to develop a business solution. There are many methods for clustering like partitioning, hierarchical, density, and grid, etc. Here K-means clustering algorithm is used, also compare it to the hierarchical clustering model with respect to silhouette and davies-bouldin sores. For getting higher efficiency and better output k-means algorithm is used, on the other hand hierarchical clustering constructs a hierarchy of clusters by either repeatedly merging two smaller clusters into a larger one or splitting a larger cluster into smaller ones.

2.2 Related Works

An awesome wide variety of articles have been written about diverse methods for separating the clients in segments. The RFM is a model which analyzes the value of a customer. Many researchers have applied various methods for segmenting customer dataset. In previous a lot of prediction & classification models were developed by researchers for RFM model analysis. Some works are

included here, V. Vijilesh, A.Harini, M.Hari Dharshini, R.Priyadharshini[1] the use of k-means clustering for client segmentation ,then writer A.joy Christy, A.Umamaheswari, L.Priyadarshini, A.Neyaa [2] named paper 'RFM ranking – An effective method to client segmentation' segmenting patron using three kinds of algorithm (k means, fuzzy c means,rm-k means).Etzion et al. [4] grouped profitable clients and calculated the price for their lifetime with the company. Cui et al. [5] designed a version for predicting reaction use of variables from RFM as well as extra literature including RFM segmentation incorporated with clustering algorithms, sincerely associated with the model to be able to be used in this paper. Cho and Moon [7] implemented a weighted common sample mining in a custom designed recommendation system. Sheshasaayee and Logeshwari [8] evolved a clean technique of segmenting with the RFM and finding lifetime cost of customers. Jiang and Tuzhilin [12] proposed in order to increase performance in marketing, both segmentation of customers and targeting of buyers are necessary. Multiple methods for segmentation of buyers has been utilized by a number of authors after this. In order to find the demand and expectations of customers for providing a good service, the assistance of segmentation is required.

2.3 Challenges

This part contains some challenges for developing a model for customer segmentation in the related work we previously mentioned.V.Vijilesh,A.Harini, M.Hari Dharshini, R.Priyadharshini[1] the use of k-means clustering for customer segmentation but fails to provide any optimal solution ,then writer A.joy Christy, A.Umamaheswari, L.Priyadarshini, A.Neyaa [2] named paper 'RFM ranking – An powerful approach to consumer segmentation' segmenting clients the usage of three forms of set of rules (k means, fuzzy c means,rm-k means). In the work of Jiang and Tuzhilin [12] proposed a better understanding of marketing performances, both segmentation and targeting based on customers and buyers are necessary respectively, but the problem was in terms of optimization, where they developed a k-means segmentation algorithm to solve the issue. Then in the work from,Cho and Moon [5] where they developed a pattern for customized recommendation system, but the challenge is here to apply unique weights to each value of transaction. In the customer management approach fuzzy c means is developed for performing segmentation based on the customers but the issue is time delay regarding multiple iteration. The technique designed by Sheshasaayee and Logeshwari [7] is also not optimized sufficiently. The churn prediction is an experimental implementation with no practical technique.

Chapter 3
Proposed Methodology

3.1 Proposed Workflow Diagram

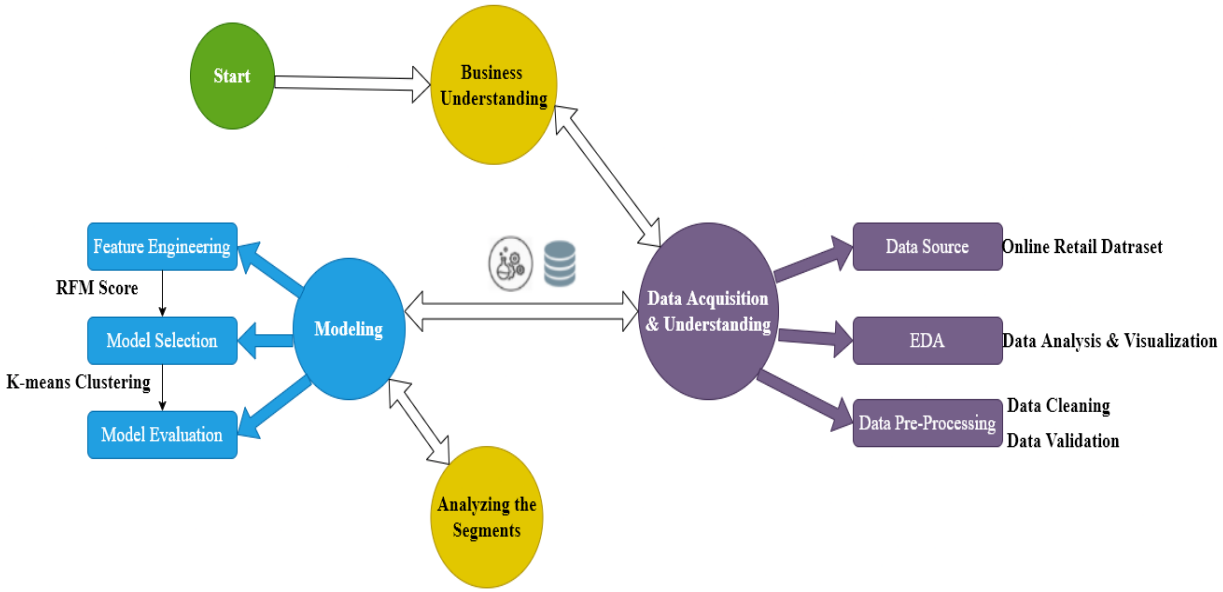

Fig 3.1: Workflow

3.2 Our Contribution

In this research we introduced k-means clustering as well as hierarchical clustering and comparing the output through some factors (i.e. Silhouette Coefficient, Davies-Bouldin Index). Silhouette & Davies-Bouldin Index is used to evaluate clustering algorithms where Silhouette Coefficient score is a metric used to compute the goodness of a clustering technique. Its score varies from -1 to 1. 1 and Davies-Bouldin index, which is a validation metric that is sometimes used in order to evaluate the minimum number of clusters to use. Using these two coefficients we compare two clustering algorithms and find that K means perform better than hierarchical for some points, which will be mentioned in **section 3.9.5**. Then we will analyze the bestselling/ top most selling product according to the dataset which will need to recommend a product for targeted customers.

3.3 Requirements

The requirements that had been applied in conducting the outcome of this research are listed below

Hardware Requirements

- OS type -- Windows 10
- RAM -- 8 GB
- Display -- 22 inch
- Processor -- intel core i5 7th generation
- Hard Disk -- 1TB HDD SATA
- Floppy Drive -- Not necessary
- CD Drive -- Not necessary
- Mouse, Keyboard

Software Requirements

- Programming language: Python version == 3.8.5
- Notebook: Jupyter , Google Colab
- Web Browser: Brave , Chrome
- Microsoft Excel , WPS office

3.4 Business Problem Understanding

An e-trade organization wants to segment its clients and decide advertising techniques in keeping with those segments. For instance, it is desirable to organize different campaigns to retain customers who are very profitable for the company, and different campaigns for new customers. The objective of the analysis is to categorize customers into various clusters (based on past data) to become aware of clients who are in all likelihood to respond to promotions and additionally for future personalization offerings that would help in increased sales and reduced customer churn rate using RFM Analysis that helps to retain customers and increase user engagement.

3.5 Dataset Collection Procedure

For this analysis we use an online Retail dataset which incorporates all of the transactions taking place for a United Kingdom-based totally and registered, non-shop online retail between 01/12/2009 and 09/12/2011.The enterprise particularly sells particular all-event gift-ware. Many customers of the agency are wholesalers. The dataset is accrued and applied from open-supply database, the link indexed underneath

→ **https://archive.ics.uci.edu/ml/datasets/Online+Retail+II**

Attribute Information:

InvoiceNo → No of Invoice.A 6-digit integral number uniquely assigned to each transaction.
However, it indicates a cancellation, If this code starts with the letter' c'.
StockCode → Code No of Product. A Five-digit integral number uniquely assigned to each distinct product.
Description → Name of Product.
Quantity → The product quantities with per transaction.
InvoiceDate → Date and time of an invoice.The day and time when a sale was generated.
UnitPrice → Price of a Unit. Product price per unit in pounds.
CustomerID →Client number. A Five-digit integral number uniquely assigned to each customer.
Country → Name of country.The name of the country where a client resides.

3.6 Required Libraries

Customer segmentation analysis needs to import necessary libraries. The desired libraries for such technique system,
Imported libraries for analysis

- Manipulation library for Date and Time - **Datetime**
- Pandas Data Manipulation Evaluation Library for data analysis - **Pandas**
- Mathematica Plotting and visualization Library for data analysis - **Matplotlib**
- Future statement Library for data analysis - **Division**
- Clustering Library for data analysis - **k means**
- Visualization Library for data plotting - **Plotly**
- Numerical Python Function Library for data analysis - **Numpy**
- Drawing Visualization Library for data analysis - **Seaborn**

3.7 Data acquisition and understanding

Data Exploration

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
```

Table 3.1: Explore Data

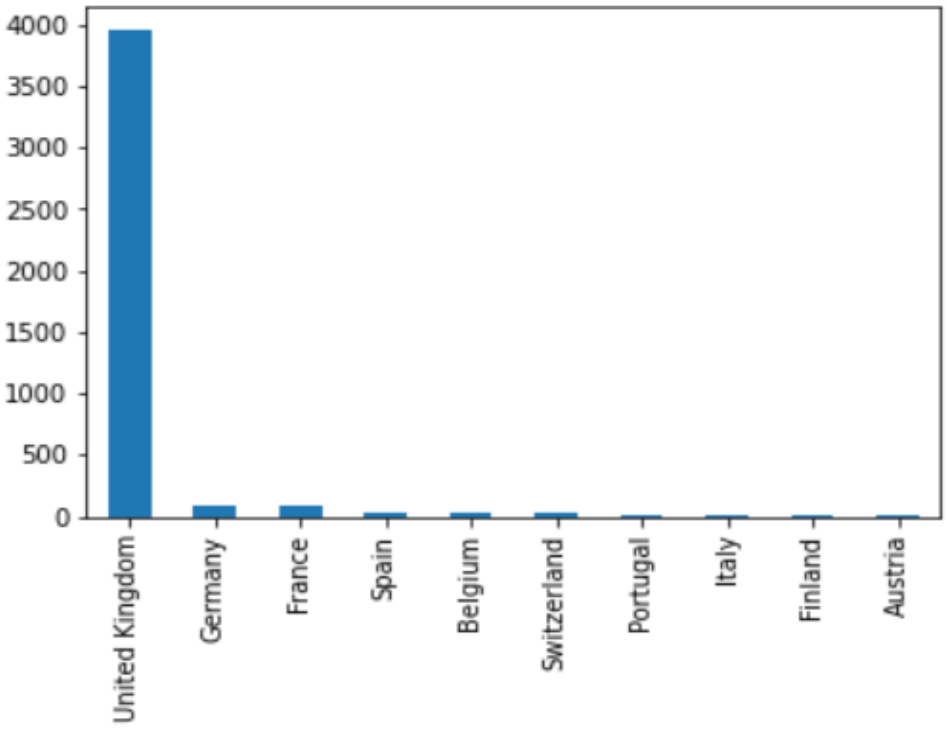

Fig 3.2: Visualize Country Level

Fig 3.2 represents that the UK is on the top most level in terms of selling products and services.

3.8 Data Preprocessing

For data preprocessing we need to solve the following issue,

- Remove null values
- Remove Duplicate values

There are missing values in the Customer ID and Description columns, so Take all the rows where customer ID is not equal to blank. So we need to remove null values from this dataset, after removing null values we see there is 406829 not-null value present in the dataset in Table 3.3

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    406829 non-null  object
 1   StockCode    406829 non-null  object
 2   Description  406829 non-null  object
 3   Quantity     406829 non-null  int64
 4   InvoiceDate  406829 non-null  object
 5   UnitPrice    406829 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      406829 non-null  object
dtypes: float64(2), int64(1), object(5)
```

Table 3.3: Remove Null Value

Then we need to remove duplicate values from the dataset. In table 3.3 we figured out that there are 4380 duplicate values which need to be removed.

	Country	CustomerID
0	United Kingdom	17850.0
9	United Kingdom	13047.0
26	France	12583.0
46	United Kingdom	13748.0
65	United Kingdom	15100.0

```
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Country     4380 non-null   object
 1   CustomerID  4380 non-null   float64
dtypes: float64(1), object(1)
```

Table 3.4: Remove Duplicate Values

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Year
178	C489449	22087	PAPER BUNTING WHITE LACE	-12	2009-12-01 10:33:00	2.95	16321.0	Australia	2009-2010
179	C489449	85206A	CREAM FELT EASTER EGG BASKET	-6	2009-12-01 10:33:00	1.65	16321.0	Australia	2009-2010
180	C489449	21895	POTTING SHED SOW 'N' GROW SET	-4	2009-12-01 10:33:00	4.25	16321.0	Australia	2009-2010
181	C489449	21896	POTTING SHED TWINE	-6	2009-12-01 10:33:00	2.10	16321.0	Australia	2009-2010
182	C489449	22083	PAPER CHAIN KIT RETRO SPOT	-12	2009-12-01 10:33:00	2.95	16321.0	Australia	2009-2010
...	...	...	...	...	...	...	...	...	...
540449	C581490	23144	ZINC T-LIGHT HOLDER STARS SMALL	-11	2011-12-09 09:57:00	0.83	14397.0	United Kingdom	2010-2011
541541	C581499	M	Manual	-1	2011-12-09 10:28:00	224.69	15498.0	United Kingdom	2010-2011
541715	C581568	21258	VICTORIAN SEWING BOX LARGE	-5	2011-12-09 11:57:00	10.95	15311.0	United Kingdom	2010-2011
541716	C581569	84978	HANGING HEART JAR T-LIGHT HOLDER	-1	2011-12-09 11:58:00	1.25	17315.0	United Kingdom	2010-2011
541717	C581569	20979	36 PENCILS TUBE RED RETROSPOT	-5	2011-12-09 11:58:00	1.25	17315.0	United Kingdom	2010-2011

Table 3.5: Remove Negative Quantities

Table 3.5 belongs to negative quantities appearing with canceled orders which we need to remove.

	Quantity	UnitPrice	CustomerID
count	361878.000000	361878.000000	361878.000000
mean	11.077029	3.256007	15547.871368
std	263.129266	70.654731	1594.402590
min	-80995.000000	0.000000	12346.000000
25%	2.000000	1.250000	14194.000000
50%	4.000000	1.950000	15514.000000
75%	12.000000	3.750000	16931.000000
max	80995.000000	38970.000000	18287.000000

Table 3.6: Data description

In Table 3.6 we observe that Field 'Quantity' requires special attention because the minimum

amount is negative and that can't be negative. CustomerID has to be integer but here it is float.

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    354345 non-null  object
 1   StockCode    354345 non-null  object
 2   Description  354345 non-null  object
 3   Quantity     354345 non-null  int64
 4   InvoiceDate  354345 non-null  object
 5   UnitPrice    354345 non-null  float64
 6   CustomerID   354345 non-null  float64
 7   Country      354345 non-null  object
dtypes: float64(2), int64(1), object(5)
```

Table 3.7: Remove Float Type CustomerID

CustomerID has to be integer but there is a float type of value existing in the dataset. In Table 3.6 where is the result after removing those float type CustomerID data.

3.9 Feature Engineering

3.9.1 Perform RFM Analysis

- Calculate the quantity of days between present date and date of last buy for each client to calculate recency.
- Calculate the wide variety of orders for each customer for getting frequency.
- Calculate sum of purchase fee or for every client for getting monetary.

576339	542
579196	533
580727	529
578270	442
573576	435

Table 3.8: Sell Per Invoice

Table 3.8 represents the product with invoice number 576339 has the maximum amount of being sold. There are so many orders for each invoice number which represents frequency of purchases. Now we need to calculate recency. So at first store a date as a reference point for recency calculations then Convert date to datetime format and Check Min and Max dates

3.9.2 Visuals of RFM Data

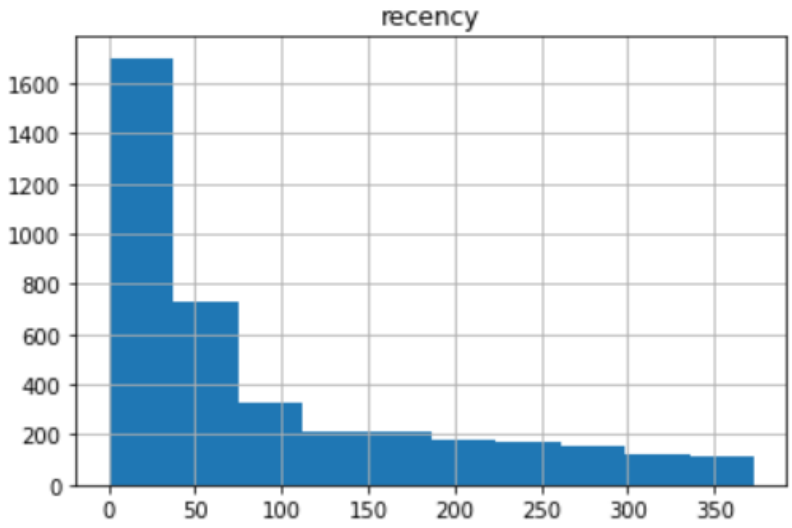

Fig 3.3: Histogram for Recency Data

In the above, Fig 3.3 histogram gives us the information of customers and their purchase time between Present and last purchase. Most of the customers fall under 50 days of recency. Which means most of the customers are regular buyers. As the recency period increased customers decreased which is a good thing.

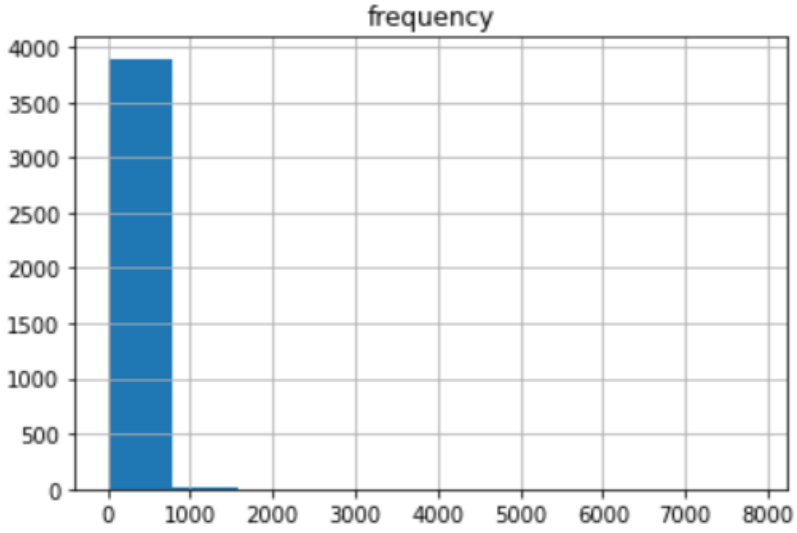

Fig 3.4: Histogram for Frequency Data

Fig 3.4 is histogram of frequency data; this shows that the buying frequency of the customers lies maximum between 0 to 700. Moreover, the histogram is highly skewed because of outliers.

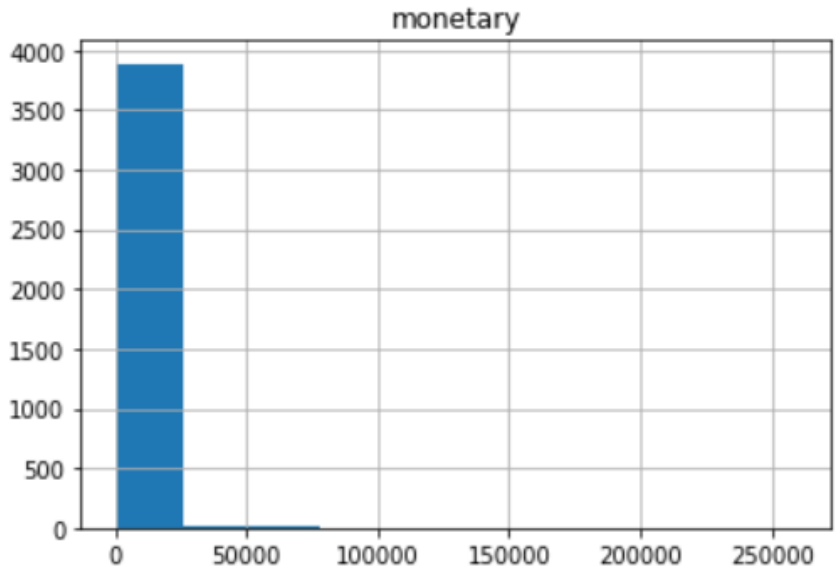

Fig 3.5: Histogram for Monetary Data

The above Fig 3.5 histogram shows the total amount of money spent per customer. Most of the customers fall under the monetary value 30000. Again this histogram is highly skewed because of few outliers.

3.9.3 RFM Score Calculation

We need to calculate the RFM score to segment the customer. Here calculating the RFM score uses quintiles to make 5 equal parts based on the available values. Each quintile contains 20% of the population. Then define functions to allocate ranks from 1 to 5. A lower recency data is better and higher frequency and monetary data are better. There are two separate functions and we need to Calculate RFM score for each customer then Combine the scores

CustomerID	recency	frequency	monetary	R	F	M	RFM Score
12346.0	325	1	77183	1	1	5	115
12747.0	2	103	4196	5	4	5	545
12748.0	0	4596	33719	5	5	5	555
12749.0	3	199	4090	5	5	5	555
12820.0	3	59	942	5	4	4	544

Table 3.9: RFM Score

CustomerID	recency	frequency	monetary	R	F	M	RFM Score	Segment
12346.0	325	1	77183	1	1	5	115	at risk
12747.0	2	103	4196	5	4	5	545	champions
12748.0	0	4596	33719	5	5	5	555	champions
12749.0	3	199	4090	5	5	5	555	champions
12820.0	3	59	942	5	4	4	544	champions

Table 3.10: Segment with RFM Score

In the table 3.9 represents RFM score then in the table 3.10 we Create 6 segments based on R and F scores where the segments are declared as ‘[1-2] [1-4]': 'at risk', '[1-2]5': 'can\'t lose','3[1-3]': 'needs attention', '[3-4] [4-5]': 'loyal customers','[4-5]1': 'new customers','[4-5] [2-5]': 'champions'. This segment indicates customers' value for an organization.

3.9.4 RFM Score Calculation for Clean Data

Here we create box plots which is mentioned in Fig 3.6 to check for outliers

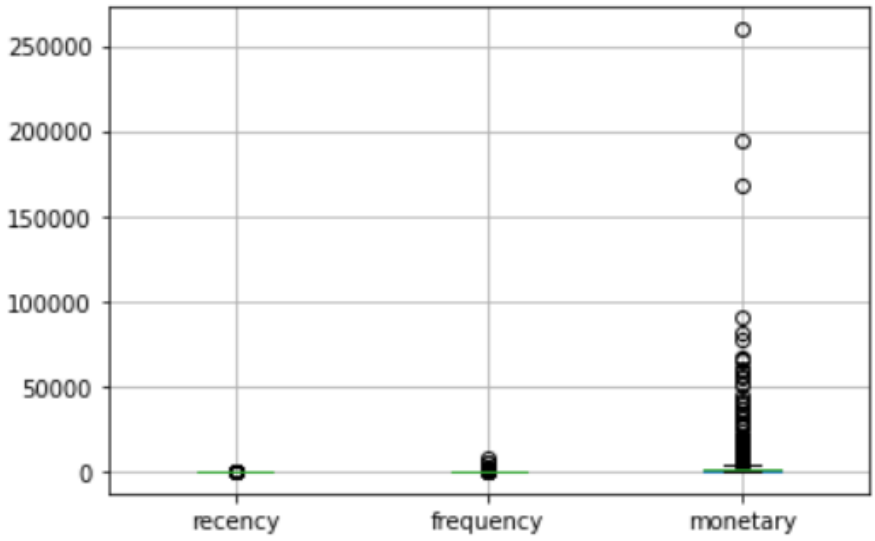

Fig 3.6: RFM Box Plot

Calculate Z scores to normalize the data. The purpose of the above code
First we have plotted the box plot and identified the outliers. Then we have calculated the z scores for recency, frequency, and monetary. Using the above code, we are trying to eliminate the outliers in the dataset.

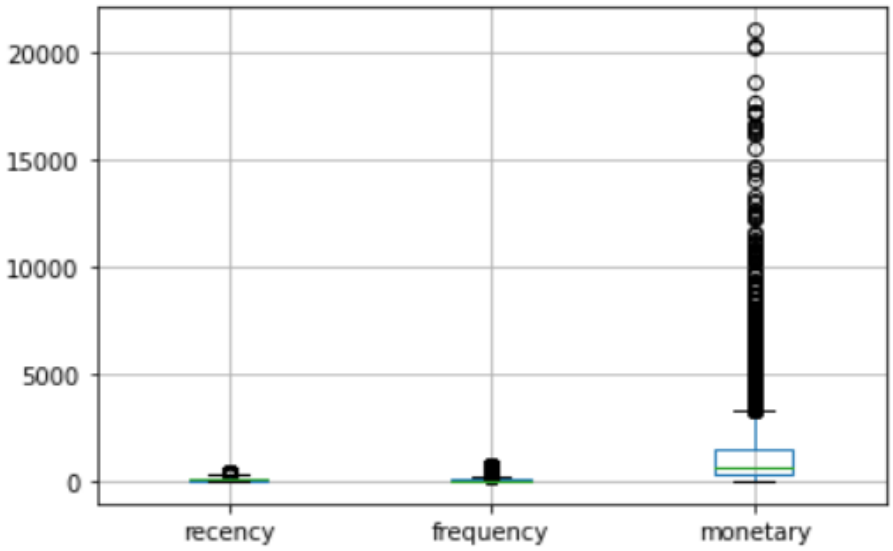

Fig 3.7: RFM Box Plot for Clean Data

Now create a new box plot in Fig 3.7. to check for outliers with the cleaned records and comment on it Now if we examine it with the preceding field plot we see that using z score we have removed the outliers whose values are less than 3. There is a small from the previous box plot to this. Even in this box plot there are outliers but we have removed potential outliers which is creating the skewness in the distribution. Then we again create the Histogram for R, F, and M for the cleaned statistics. Examine it with the preceding field plot

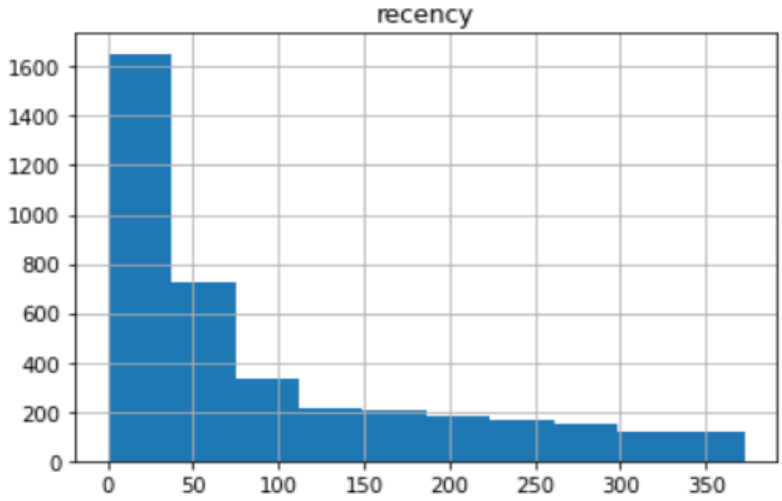

Fig 3.8: Histogram for Cleaned Recency Data

The above histogram in Fig 3.8 shows that most of the customers fall within the recency of 0 to 50 days. As the recency increased, the number of customers who fall in that region. Recency histogram of original rfm and the histogram of rfm clean data looks similar. This shows that there are no outliers in the recency.

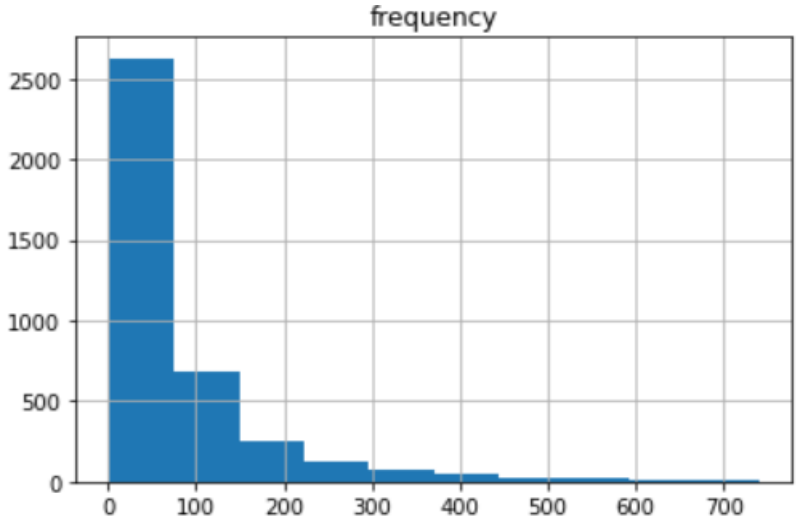

Fig 3.9: Histogram for Cleaned Frequency Data

This histogram at Fig 3.9 shows how customers are spread in terms of frequency. Most of the

customers fall under 100 purchases of frequency. And the spread is till 700.If we compare with the previous histogram which is highly skewed because of outliers, even after removing most of the outliers we can still see it is skewed but it is more inclined to normal distribution.

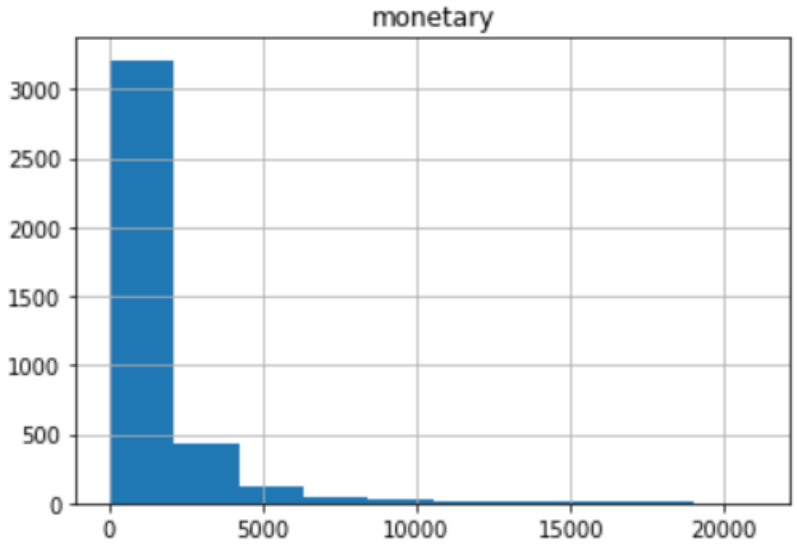

Fig 3.10: Histogram for Cleaned Monetary Data

The above histogram at Fig 3.10 shows the number of customers and their monetary value. After removing outliers all customers fall under 20000 monetary value and most of the customers fall under 3000.When we compare this histogram with the previous one which is highly skewed because of outliers, is now more normally distributed

3.9.5 Model Evaluation

Hierarchical Clustering

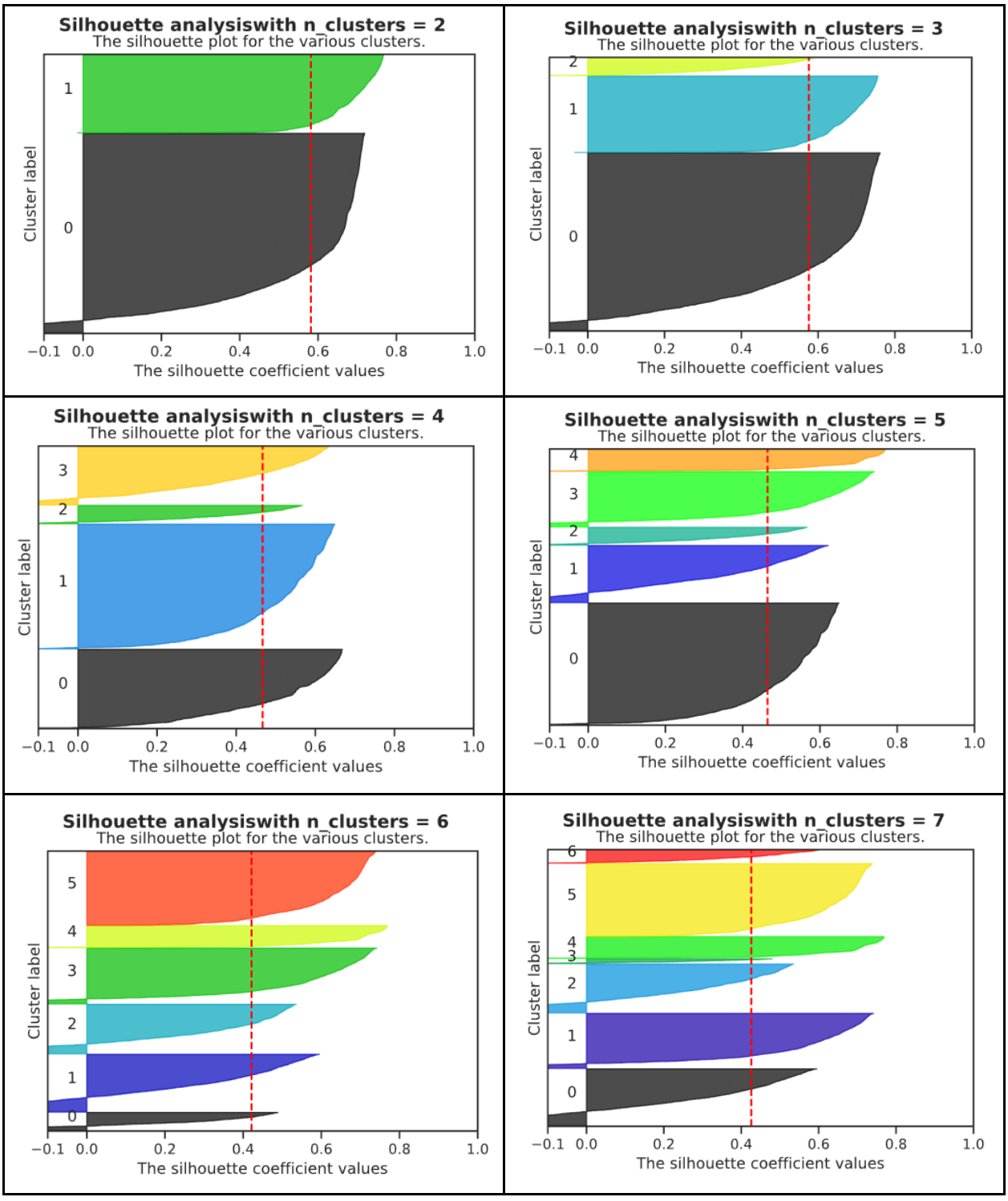

Fig 3.11: Silhouette Analysis for Hierarchical Clustering

K means Clustering

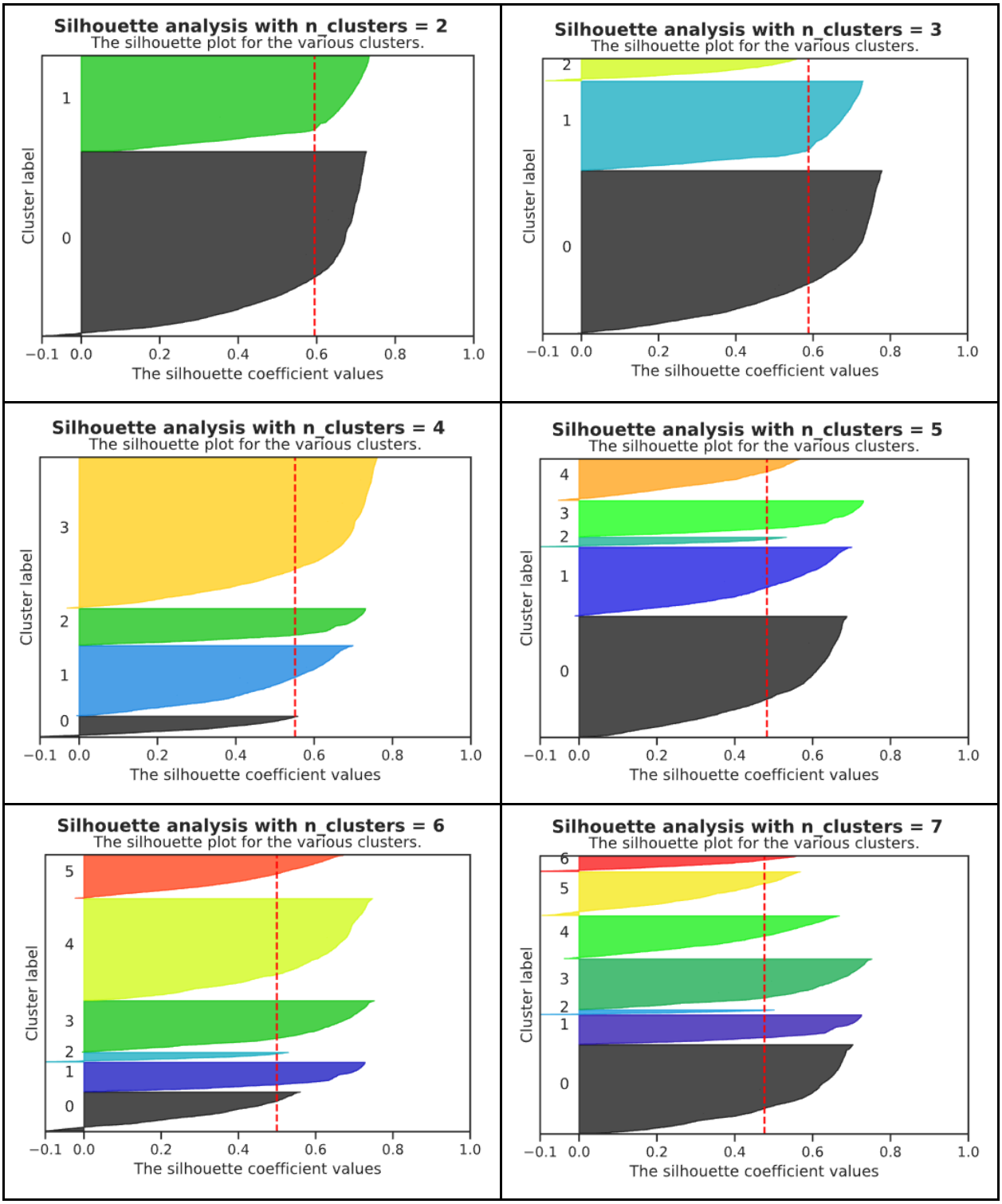

Fig 3.12 Silhouette Analysis for K-means Clustering

Silhouette & Davies-Bouldin Index is used to evaluate clustering algorithms where Silhouette Coefficient score is a metric used to compute the goodness of a clustering technique. Its score

varies from -1 to 1. 1 and Davies-Bouldin index, which is a validation metric that is sometimes used in order to evaluate the minimum number of clusters to use.

	model	n_clusters	s_score	db_score
0	KMeans	2	0.594541	0.582639
0	Hier	2	0.581698	0.559888
1	KMeans	3	0.588156	0.649014
1	Hier	3	0.575753	0.636482
2	KMeans	4	0.551103	0.642604
2	Hier	4	0.466708	0.738229
3	KMeans	5	0.482530	0.728177
3	Hier	5	0.464246	0.690028
4	KMeans	6	0.499492	0.669170
4	Hier	6	0.421601	0.707550
5	KMeans	7	0.476276	0.733422
5	Hier	7	0.425663	0.752059

Table 3.11: Evaluation Between K-means and Hierarchical Clustering

Figure 3.11 & Fig 3.12 shows that K-means and Hierarchical models are very similar in terms of their silhouette scores. We noticed from the table 3.9 For both models, from n_clusters = 6 to 7, K-means outperforms at n_clusters = 6.

K-MEANS gives disjoint sets that want each customer to belong to one and only one segment which is mainly required for a big number of dataset. With a large number of variables, K-means compute faster. Here our datasets contain around 541,000 customers. Therefore, time complexity could be an issue. K-means has a linear time complexity O(n) as opposed to hierarchical which has a quadratic complexity - O(n^2). So getting faster computation we think K means algorithm is suitable for our further analysis.

3.9.6 Cluster Analysis Using K-means Clustering

Now we need to perform clustering for the dataset using k-means clustering for this here use Elbow method, which means it runs k-means algorithm for clustering on the dataset for various values for could be from 1 to 10 and then for every cost of k calculate an average rating for all partitions. by

using default, the distortion score is calculating, the sum of rectangular distances(WCSS) from every point to its assigned center

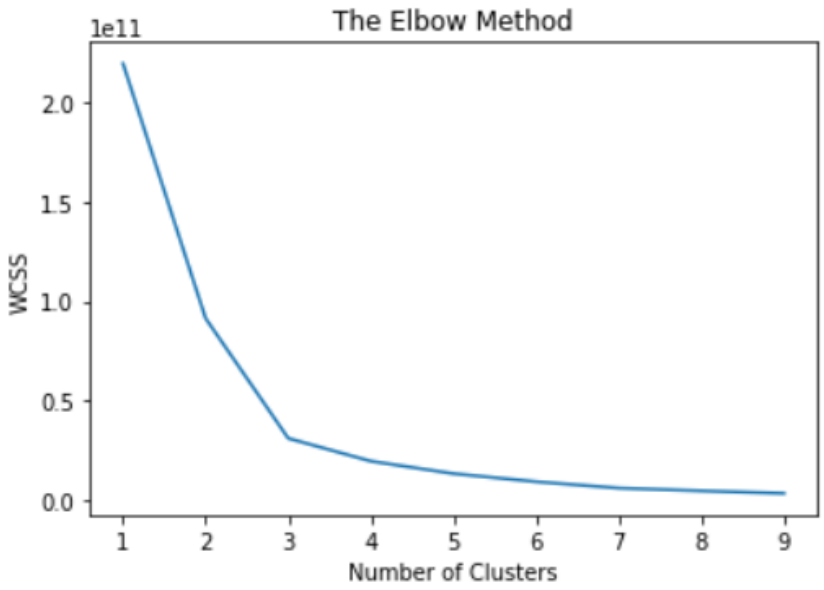

Fig 3.13: Elbow Method with Outliers

In the following figure we fit K means to the dataset using K=3, that means the dataset is segmenting by 3 cluster

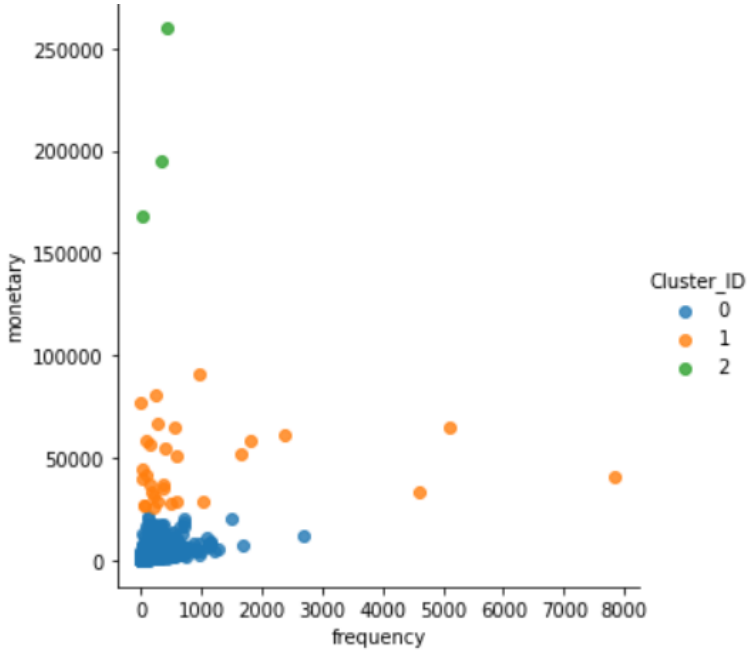

Fig 3.14: Frequency vs Monetary, K= 3

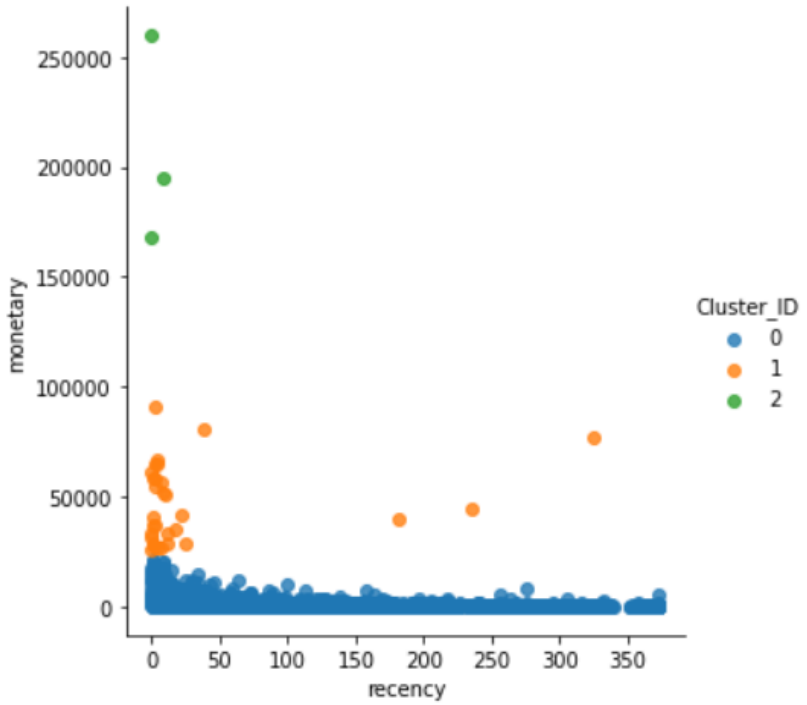

Fig 3.15: Recency vs Monetary, K= 3

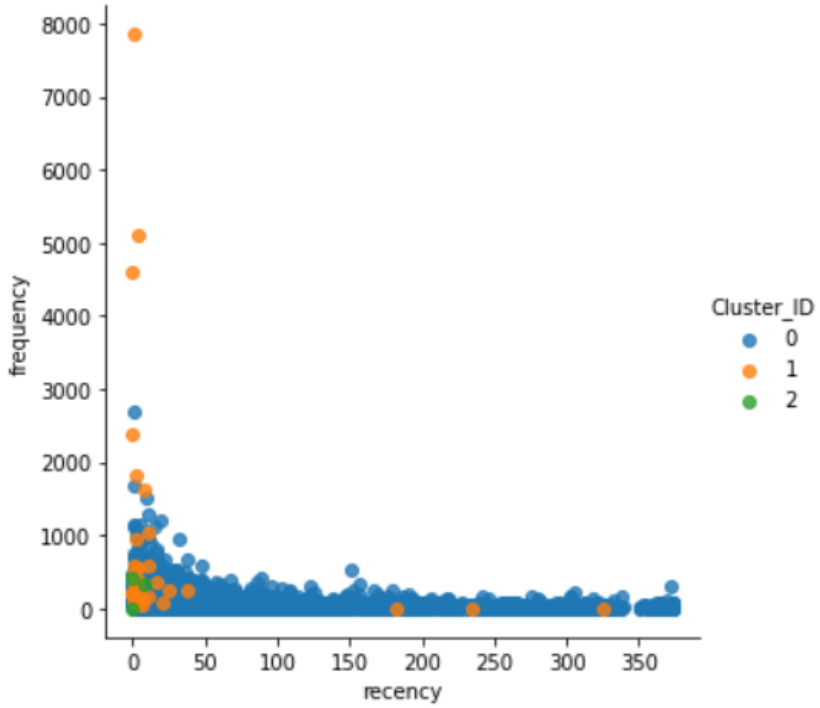

Fig 3.16: Recency vs Frequency, K= 3

We have used the elbow method to determine 3 clusters to be formed. We have implemented and

plotted various scatter plots.

When it comes to frequency vs monetary, the clusters are clear. It represents customers' purchase frequency and their monetary value. Clusters green represents customers with low frequency but high monetary value purchase. Orange cluster shows those customers with high frequency and moderate monetary value. Blue cluster shows those customers whose frequency is low as well as monetary.

Next plot is recency vs monetary, the clusters are relatively clear. It represents days elapsed between present and last purchase by customer and their monetary value. As the recency increased, money decreased. Green cluster represents those customers whose recency is low (which is good) and have high monetary value. Orange cluster shows the customers whose recency is low and monetary is moderate. Blue cluster shows those customers whose recency is low and monetary is also low.

Final plot is recency vs frequency, the clusters are not clear. It represents days elapsed between present and last purchase by customer and purchase frequency. It is very hard to distinguish the clusters clearly.

In all the graphs most of the customers fall in the blue cluster, next orange and finally green clusters.

Now, Perform Cluster Analysis with k=4

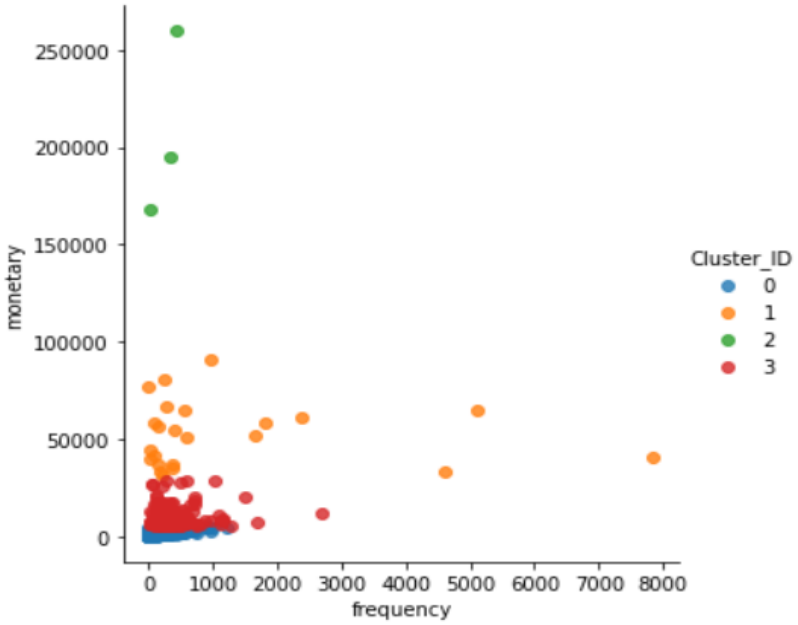

Fig 3.17: Frequency vs Monetary, K= 4

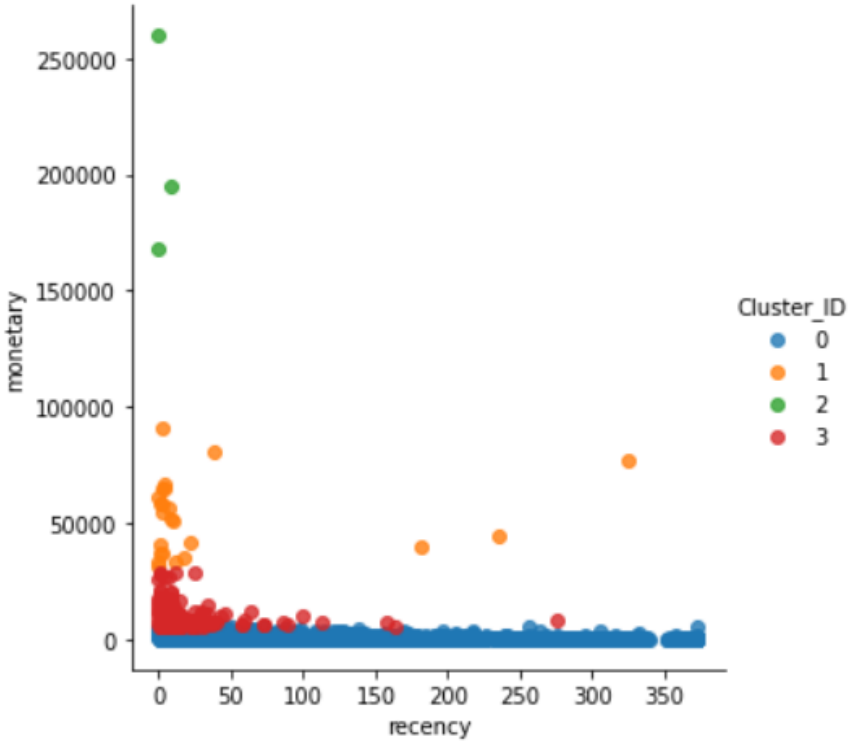

Fig 3.18: Recency vs Monetary, K= 4

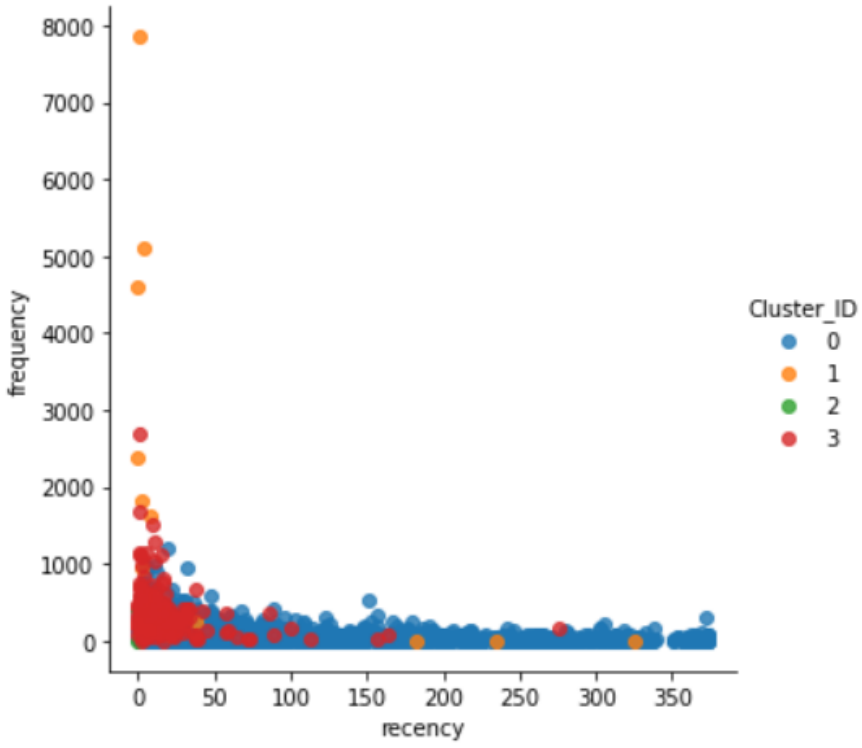

Fig 3.19: Recency vs Frequency, K= 4

Now we have just randomly assigned the k value as 4. Earlier only recency vs frequency plot was not able to be understood. Now even frequency vs recency is hard to distinguish the clusters. We can clearly distinguish the clusters in recency vs monetary plot.

It represents days elapsed between present and last purchase by customer and their monetary value. As the recency increased, money decreased. Green cluster represents those customers whose recency is low (which is good) and have high monetary value. Orange cluster shows the customers whose recency is low and monetary is moderate. Red and Blue clusters show those customers whose recency is low and monetary is also low.

Cluster analysis with the outliers removed rfm dataframe.

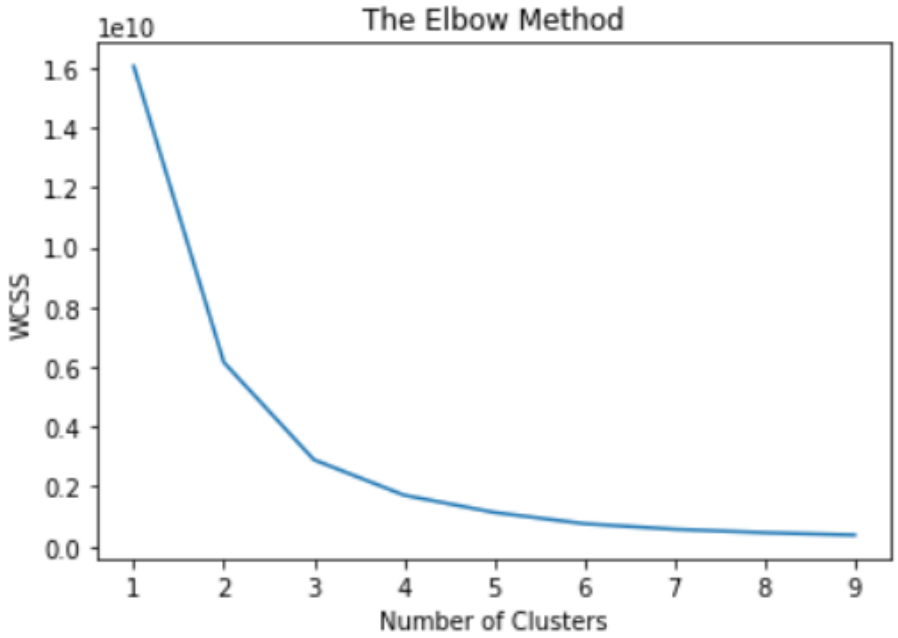

Fig 3.20: Elbow Method Without Outliers

In the following figure 3.21,3.22,3.23 we fit K means to the dataset using K=3, that means the dataset is segmenting by 3 cluster

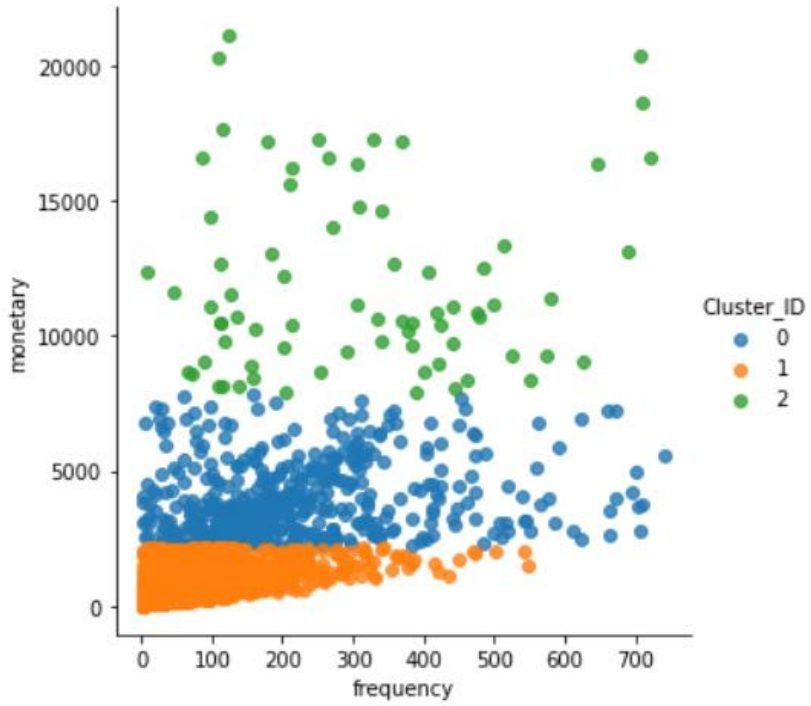

Fig 3.21: Frequency vs Monetary (Without Outlier), K= 3

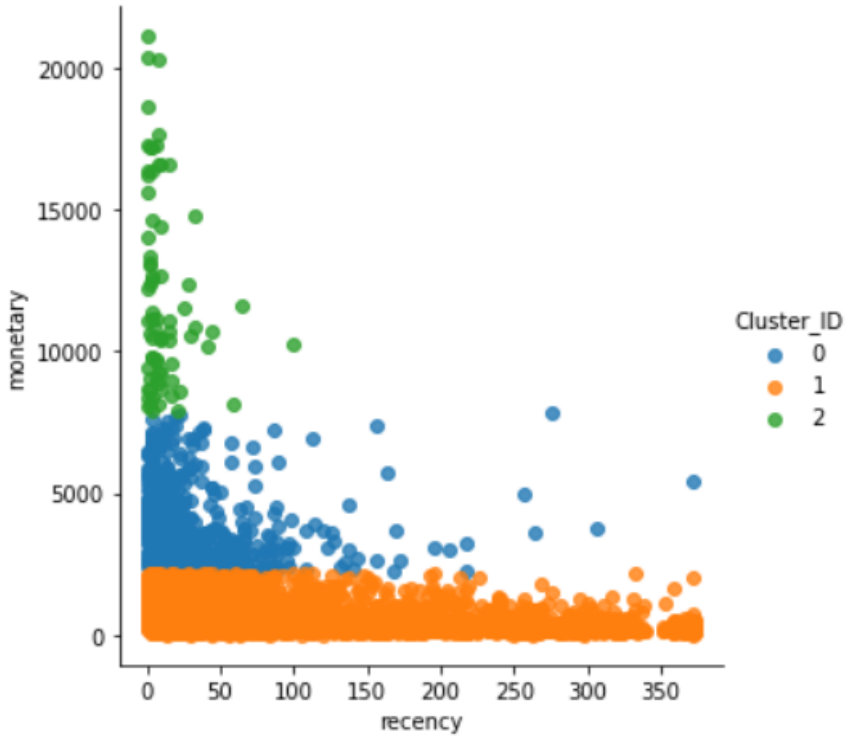

Fig 3.22: Recency vs Monetary (Without Outlier), K= 3

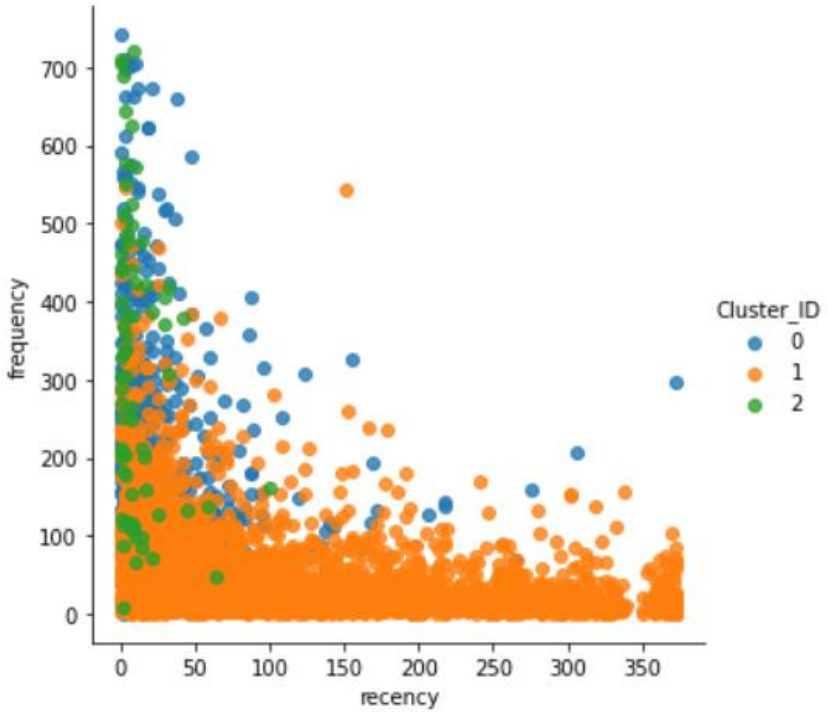

Fig 3.23: Recency vs Frequency (Without Outlier), K= 3

We have used the elbow method to determine 3 clusters to be formed for a cleaned dataset. We have implemented and plotted various scatter plots.

When it comes to frequency vs monetary, the clusters are clear. It represents customers' purchase frequency and their monetary value. Clusters green represents customers with high frequency but high monetary value purchase. Orange cluster shows those customers with high frequency and moderate monetary value. Blue cluster shows those customers whose frequency is low as well as monetary. If we compare with earlier scatter plots, due to outliers only few customers fell into the green cluster.

Next plot Fig 3.22 is recency vs monetary, the clusters are relatively clear. It represents days elapsed between present and last purchase by customer and their monetary value. As the recency increased, money decreased. Green cluster represents those customers whose recency is low (which is good) and have high monetary value. Orange cluster shows the customers whose recency is low and monetary is moderate. Blue cluster shows those customers whose recency is low and monetary is also low. Even for this plot as we removed outliers the formation of clusters is clearer.

Final plot Fig 3.23 is recency vs frequency, the clusters are not clear. It represents days elapsed between present and last purchase by customer and purchase frequency. It is very hard to distinguish the clusters clearly. Even though we have removed the outliers, the cluster division is not clear to distinguish.

In all the graphs most of the customers fall in the blue cluster, next orange and finally green clusters.

3.9.7 Cluster Analysis Using Hierarchical Clustering

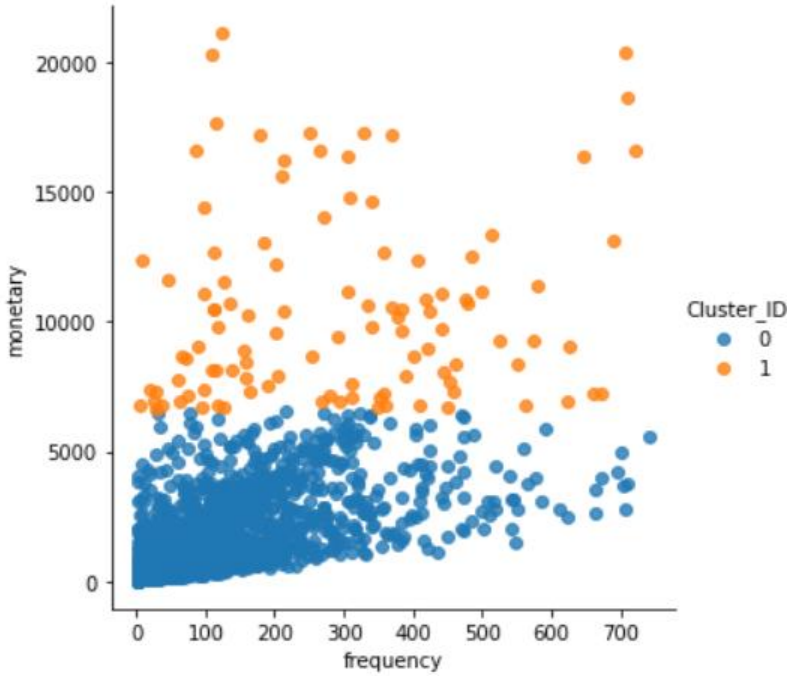

Fig 3.24: Frequency vs Monetary(Hierarchical), Cluster=2

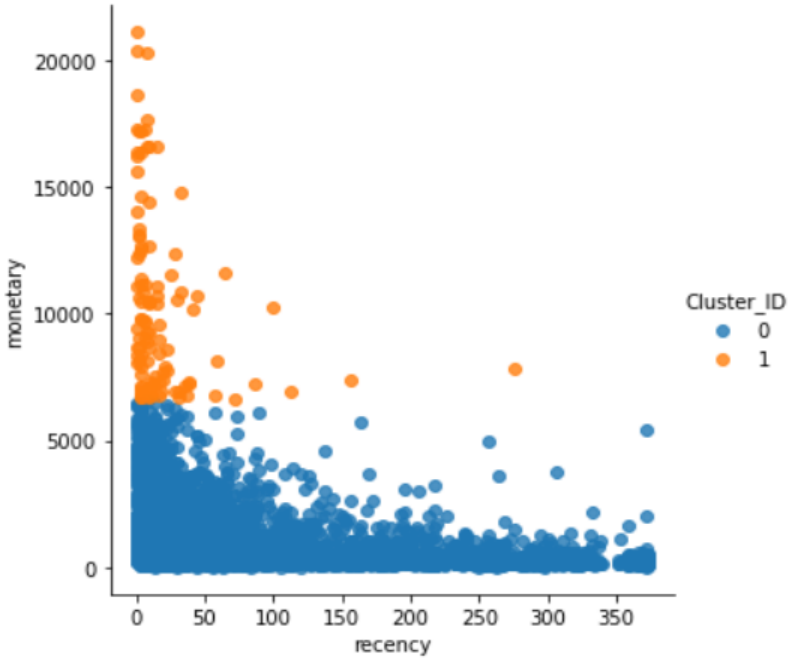

Fig 3.25: Recency vs Monetary(Hierarchical), Cluster=2

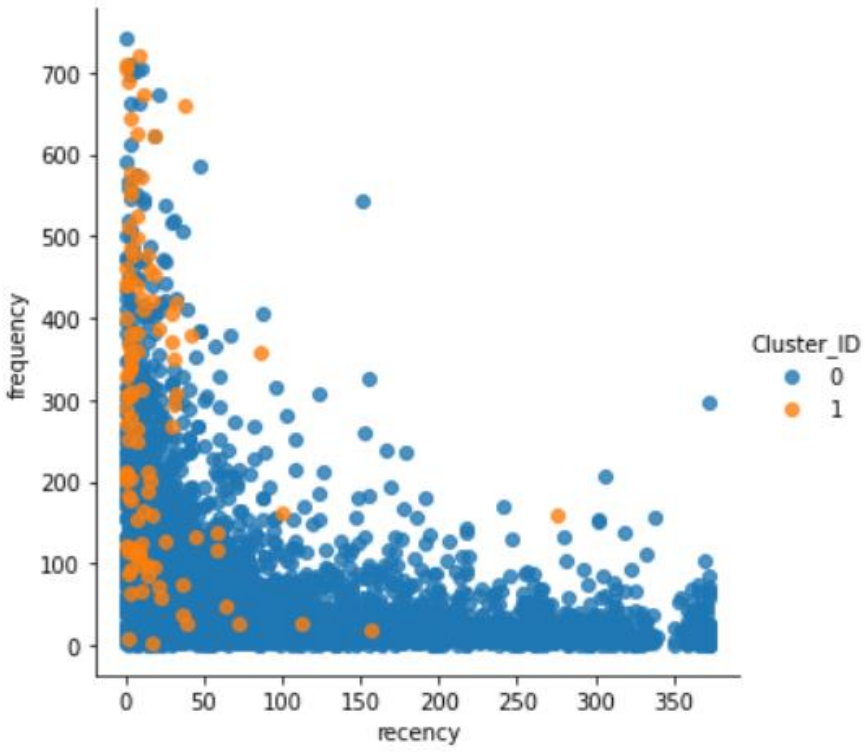

Fig 3.26: Recency vs Frequency(Hierarchical), Cluster=2

Here we have used a hierarchical clustering method to form the clusters. Using dendrogram we have determined to use 2 clusters and form it using agglomerative clustering. When we compare this method of clustering with k means, the clusters are much clearer and easy to distinguish.

When it comes to frequency vs monetary Fig 3.24, the clusters are clear. It represents customers' purchase frequency and their monetary value. Clusters orange represents customers with high frequency but high monetary value purchase. Blue cluster shows those customers with relatively low frequency and moderate monetary value.

Next plot Fig 3.25 is recency vs monetary, the clusters are relatively clear. It represents days elapsed between present and last purchase by customer and their monetary value. As the recency increased, money decreased. Orange cluster represents those customers whose recency is low (which is good) and have high monetary value. Blue cluster shows the customers whose recency is low and monetary is comparatively low.

Final plot Fig 3.26 is recency vs frequency, the clusters are not that clear. It represents days elapsed between present and last purchase by customer and purchase frequency. It is very hard to distinguish the clusters clearly. Even though we have used hierarchical methods to form the clusters, most of the points in this plot are in both cluster spaces.

Chapter 4
Result and Analysis

4.1 Experimental Result and Analysis

4.1.1 Algorithm Comparison

Silhouette & Davies-Bouldin Index is used to evaluate clustering algorithms where Silhouette Coefficient score is a metric used to compute the goodness of a clustering technique. Its score varies from -1 to 1. 1 and Davies-Bouldin index, which is a validation metric that is sometimes used in order to evaluate the minimum number of clusters to use.

	model	n_clusters	s_score	db_score
0	KMeans	2	0.594541	0.582639
0	Hier	2	0.581698	0.559888
1	KMeans	3	0.588156	0.649014
1	Hier	3	0.575753	0.636482
2	KMeans	4	0.551103	0.642604
2	Hier	4	0.466708	0.738229
3	KMeans	5	0.482530	0.728177
3	Hier	5	0.464246	0.690028
4	KMeans	6	0.499492	0.669170
4	Hier	6	0.421601	0.707550
5	KMeans	7	0.476276	0.733422
5	Hier	7	0.425663	0.752059

Table 4.1: Comparison Between K-means and Hierarchical Clustering

Figure 3.11 and Fig 3.12 shows that K-means and Hierarchical models are very similar in terms of their silhouette scores. We noticed from the table 3.9 For both models, from n_clusters = 6 to 7, K-means outperforms at n_clusters = 6.

K-MEANS gives disjoint sets that want each customer to belong to one and only one segment which is mainly required for a large number of datasets. With a large number of variables, K-

means compute faster. Here our datasets contain around 541,000 customers. Therefore, time complexity could be an issue. K-means has a linear time complexity O(n) as opposed to hierarchical which has a quadratic complexity - O(n^2). So getting faster computation we think K means algorithm is suitable for our further analysis.

4.1.2 Customer Segmentation Analysis

We Create 6 segments based on R and F scores where the segments are declared as ‘[1-2] [1-4]': 'at risk', '[1-2]5': 'can\'t lose','3[1-3]': 'needs attention', '[3-4] [4-5]': 'loyal customers','[4-5]1': 'new customers','[4-5] [2-5]': 'champions'. This segment indicates customers' value for an organization.

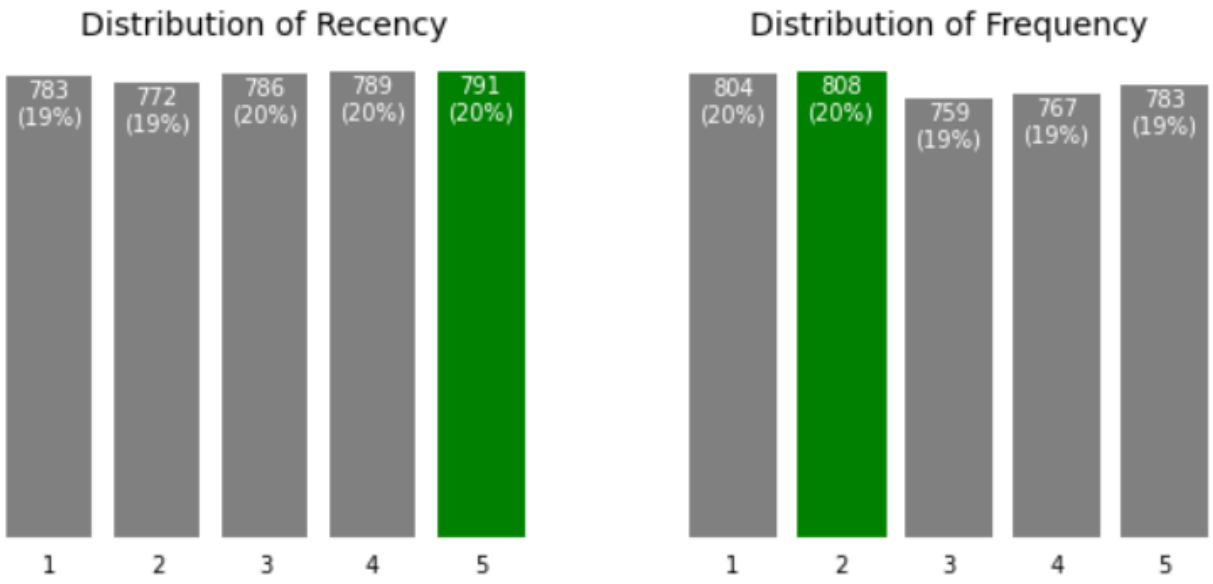

Fig 4.1: Distribution of R, F (For 6 Segment)

Following Fig 4.1 we observe that in the distribution of recency the max value is 791, whereas in the distribution of frequency the max value is 808.

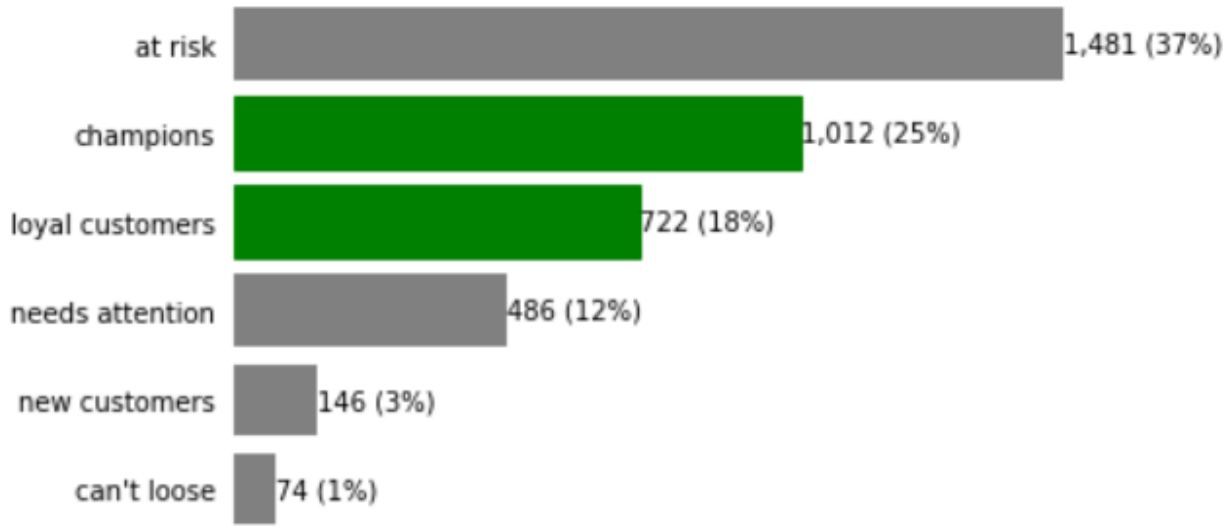

Fig 4.2: Customer Segmentation Category

In the Fig 4.2, highest number of customers are in the at risk category and least are in the can't lose category these two categories are important, we need to develop a special strategy to keep them with the company, champions category is in the second place and the loyal customers category is in the third place, needs attention category is in the 4 th place and new customers is in the 5th place.

Based on the category and the number of people in that category we should have a specified strategy for them.

Best Customer

Now if we try to figure out the best customer regarding to RFM score, the best customers are the ones with the more rfm value here the best customers are the champions category and the loyal customer's category which is in the following Table 4.2

CustomerID	recency	frequency	monetary	R	F	M	RFM Score	Segment
18102.0	0	431	259657	5	5	5	555	champions
17450.0	8	337	194550	5	5	5	555	champions
17511.0	2	963	91062	5	5	5	555	champions
16684.0	4	277	66653	5	5	5	555	champions
14096.0	4	5111	65164	5	5	5	555	champions

Table 4.2: Best Customer Segment (6 Segment)

Churn Customer

Then we need to find Customers who are likely to churn, meaning those customers who have not bought any product for a long time. This is shown in terms of recency, where the value is less than 1 for the customer. These are customers who are likely to churn. Customers whose currency value is low.

CustomerID	recency	frequency	monetary	R	F	M	RFM Score	Segment
12346.0	325	1	77183	1	1	5	115	at risk
15749.0	235	10	44534	1	1	5	115	at risk
15098.0	182	3	39916	1	1	5	115	at risk
13093.0	275	159	7832	1	5	5	155	can't loose
17850.0	372	297	5391	1	5	5	155	can't loose

Table 4.3: Churn Customer Segment (6 Segment)

Lose Customer

Customers with low RFM scores are the customers we are going to lose. Which means customers who have not bought any product for a long time, even if the customer bought a product the frequency and monetary is also low. Following table shows the customers we are going to lose. Customers who's recency, frequency and monetary values are low.

CustomerID	recency	frequency	monetary	R	F	M	RFM Score	Segment
13747.0	373	1	79	1	1	1	111	at risk
14237.0	373	9	161	1	1	1	111	at risk
17643.0	373	8	101	1	1	1	111	at risk
15350.0	373	5	115	1	1	1	111	at risk
13011.0	372	3	50	1	1	1	111	at risk

Table 4.4: Lose Customer Segment (6 Segment)

Loyal Customer

Now we find the loyal customer using RFM score, Customers whose recency is low and high frequency & high monetary value are loyal customers. We have calculated those customers with a frequency more than 3. Following table shows the loyal customers. Customers with high frequency value.

CustomerID	recency	frequency	monetary	R	F	M	RFM Score	Segment
18102.0	0	431	259657	5	5	5	555	champions
17450.0	8	337	194550	5	5	5	555	champions
17511.0	2	963	91062	5	5	5	555	champions
16029.0	38	242	81024	3	5	5	355	loyal customers
16684.0	4	277	66653	5	5	5	555	champions

Table 4.5: Loyal Customer Segment (6 Segment)

Now we calculate RFM scores using 4 quintiles instead of 6, We Create 4 segments based on R and F scores where the segments are declared as ‘[1-2] [1-4]': 'at risk', '3[3-4]': 'loyal customers’, [3-4] [1-2]': 'new customers','4[3-4]': 'Best Customer’'. This segment indicates customers' value for an organization.

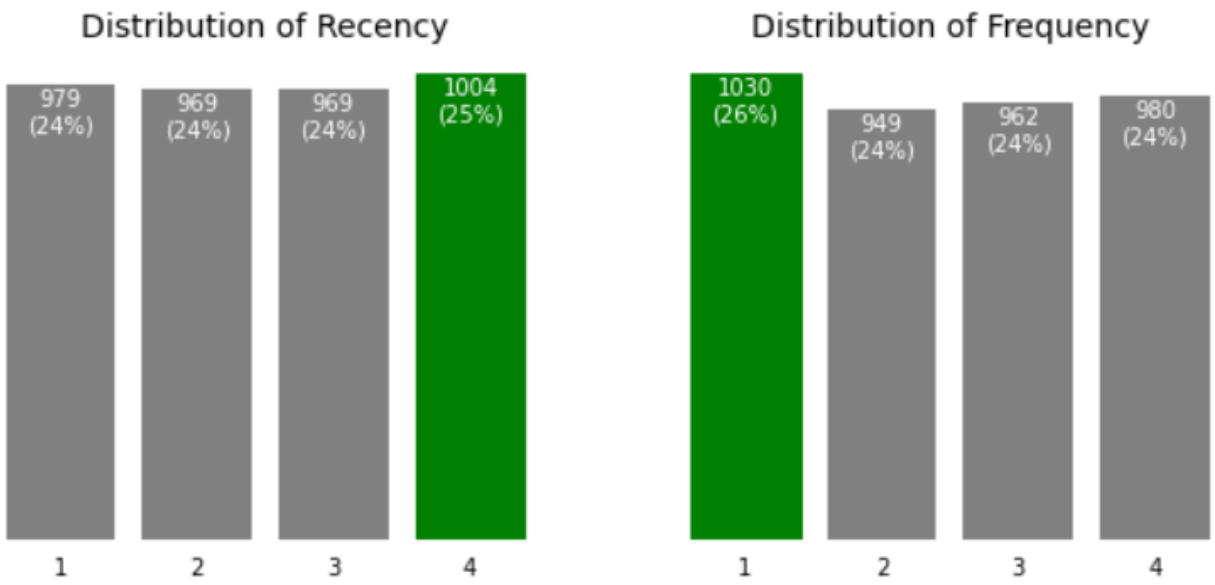

Fig 4.3: Distribution of R, F (For 4 Segment)

We can observe that after dividing the overall data into 4 quintiles the recency is high in the fourth quintile and the frequency is more in the first quintile, from this we can say that the recent purchase customers are more in the fourth quintile set of data and the frequent customers are more in the first quintile set of data.

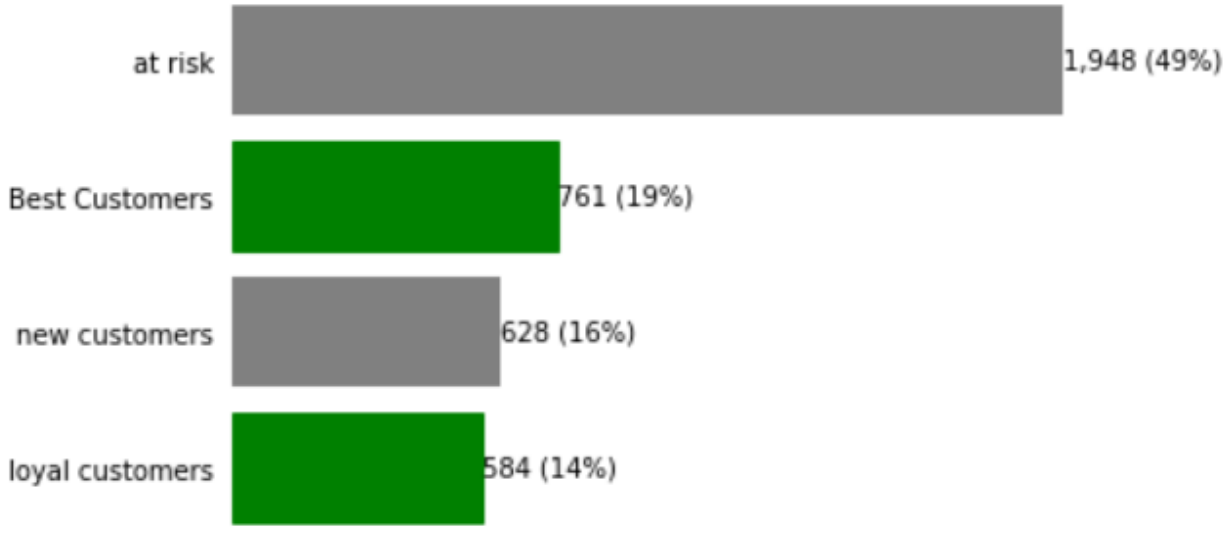

Fig 4.4: Customer Segmentation Category (For 4 Segment)

Best Customer

From the above chart Fig 4.4 we can say that nearly half of the customers fall under the "at risk" category, followed by the "Best customers" category and the "new customers" category the "loyal customers" category is in last place.

Table 4.6 finds out Customers whose rank of RFM is 444 are the best customers, because the recency is low, they are frequent buyers and generate high monetary value.

CustomerID	recency	frequency	monetary	RFM Score	Segment	R_1	F_1	M_1
18102.0	0	431	259657	444	Best Customers	4	4	4
17450.0	8	337	194550	444	Best Customers	4	4	4
17511.0	2	963	91062	444	Best Customers	4	4	4
16684.0	4	277	66653	444	Best Customers	4	4	4
14096.0	4	5111	65164	444	Best Customers	4	4	4

Table 4.6: Best Customer Segmentation (4 Segment)

Churn Customer

Now to find out the churn customers Table 4.7 with more recency value, which means the customers who have not bought any product for a long time. Here are the customers with recency rank 1, who are likely to churn.

CustomerID	recency	frequency	monetary	RFM Score	Segment	R_1	F_1	M_1
12346.0	325	1	77183	114	at risk	1	1	4
15749.0	235	10	44534	114	at risk	1	1	4
15098.0	182	3	39916	114	at risk	1	1	4
13093.0	275	159	7832	144	at risk	1	4	4
12980.0	157	20	7374	124	at risk	1	2	4

Table 4.7: Churn Customer Segmentation (4 Segment)

Lose Customer

Customers whose recency value is high, frequency is low and monetary is low. Which means the customers who have not bought any product for a long time, even if they buy the frequency and money is less. According to our ranking customers whose rfm score is 111 are customers who we are going to lose.

CustomerID	recency	frequency	monetary	RFM Score	Segment	R_1	F_1	M_1
13065.0	373	14	205	111	at risk	1	1	1
16583.0	373	14	233	111	at risk	1	1	1
16048.0	373	8	256	111	at risk	1	1	1
13747.0	373	1	79	111	at risk	1	1	1
17643.0	373	8	101	111	at risk	1	1	1

Table 4.8: Lose Customer Segmentation (4 Segment)

Loyal Customer

Loyal Customers are those customers whose frequency is high. Which means though their recency score is relatively less, their frequency is high.

CustomerID	recency	frequency	monetary	RFM Score	Segment	R_1	F_1	M_1
18102.0	0	431	259657	444	Best Customers	4	4	4
17450.0	8	337	194550	444	Best Customers	4	4	4
17511.0	2	963	91062	444	Best Customers	4	4	4
16029.0	38	242	81024	344	loyal customers	3	4	4
16684.0	4	277	66653	444	Best Customers	4	4	4

Table 4.9: Loyal Customer Segmentation (4 Segment)

4.1.3 Product Analysis

In the product analysis part, we find out the most selling item from the dataset. Where in Table 4.10 it shows us the higher selling product with priority.

WHITE HANGING HEART T-LIGHT HOLDER	2369
REGENCY CAKESTAND 3 TIER	2200
JUMBO BAG RED RETROSPOT	2159
PARTY BUNTING	1727
LUNCH BAG RED RETROSPOT	1638
	...
damages/showroom etc	1
PINK POLKADOT KIDS BAG	1
Sale error	1
CAMOUFLAGE DOG COLLAR	1
72 CAKE CASES VINTAGE CHRISTMAS	1

Table 4.10: Most Selling Product

4.2 Recommendation for Taking Business Decision

Customer segmentation using RFM segmentation analysis we can evaluate some groups of customers which will help us to find out what should be our approach towards each particular segment. Generating some messaging that is tailored for every consumer group. With the aid of focusing on the behavioral sample of precise corporations, (RFM) technique allows entrepreneurs to speak with their clients in a mile's greater effective approach.
once more, right here is just a few instance of enterprise decision, using the groups we named:

- **Best customers**
  While communications with this group have to cause them to experience value and preference. These clients in all likelihood generate a disproportionately high percent of normal sales and for that reason specializing in preserving their happiness need to be a top priority, also reading their individual choices and affinities will offer extra benefits for even greater personalized messaging.
- **High-spending New Customers**
  It is continually an excellent concept to cautiously "incubate" all new customers, however because those new clients spent plenty on their first buy, it's even more vital. Like with the best clients group, it's essential to lead them to experience valued and liked – and to present them incredible incentives to maintain interacting with the brand.
- **Lowest-Spending Active Loyal Customers**
  This segment, repeat customers are energetic and constant, also they may be low investors. Organizations have to create campaigns for this group that lead them to experience value, and incentivize them to boom them spend stages. As loyal customers, it frequently also pays to praise them with unique offers if they spread the word about the brand to their buddies, like through Facebook, twitter, mail.
- **Churned Best Customers**

Where belongs precious customers who don't continue transacting for a long term in the past. While it’s often challenging to re-engage churned clients, the excessive fee of those customers makes it profitable. Like with the best clients group, it’s needed to communicate with them on the idea in their specific choices, as acknowledged from earlier transaction data.

RFM evaluation can assist shops phase the clients and design offers and promotions primarily based on their profile. below are some examples:

- Clients with a usual high RFM score constitute the first-rate clients.
- Customers who have an excessive overall RFM score but a frequency rating of one are new clients. The employer can offer special gifts for these clients that allow you to boom their visits.
- Customers who've a high frequency score but a low recency rating are the ones customers that used to go pretty regularly however have no longer been visiting recently. For these clients, the organization desires to offer promotions to deliver them lower back to the store, or run surveys to discover why they deserted the store.
- RFM scores can be analyzed collectively with the results of the campaigns to get rid of unresponsive clients and similarly improve the campaigns.
- RFM score may be analyzed collectively with the products they buy to design especially centered gives for each consumer section.
- RFM rating may be analyzed together with different records about the customers together with their earnings ranges, gender, whether or not they own a car or not, and so forth. to segment the customers

Of course, figuring out which corporations of customers to target and how to best speak with them is wherein the art of advertising is available!

In the product analysis part, we find out the most selling item from the dataset. Where in Table 4.10 it shows us the higher selling product with priority. If we focus on the most selling product we can easily generate some promotional advertisement above this product, so that this can manipulate customer demand more.

Chapter 5
Effect on Environment and Society

5.1 Effect on Environment

Entrepreneurs can be more efficient in phrases of time, cash, and different assets by segmenting their records. Market segmentation permits organizations to have a higher information of their customers. They gather a greater draw close to the requirements and desires of clients, letting them customize campaigns to the customer corporations most probable to shop for objects.

5.2 Effect on Society

This observation does not now affect the physical surroundings instead of the social surroundings. There are possibilities to increase higher social know-how and grouping customers in day after day life based totally on this take a look at.

5.3 Ethical Aspects

This take a look at is morally secure due to the reality, in this observation does not degrade any kind of privacy topics. Also, this set of data is gathered from an open supply to keep away from privacy laws.

5.4 Critical Challenges

Data collection is one of the most difficult parts of this project. The scarcity of dataset on the internet in this regard made the task even more difficult. As we are working with machine learning models and data analysis, the programming part will be very difficult. Accurately segmenting the customer is a daunting task. The customers are just not concerned about the products, they are motivated by services, discounts, staff behavior and other occasional features of a particular shop or any business hub. So there are various other factors that can attract a customer being loyal even before. It can be extremely difficult to figure out the satisfaction level of a customer.

5.5 Conflict Requirements

Developing a practical machine learning model with proper regularization with low variance while limited given data will be used. If the learning process is unsupervised or reinforcement rather than supervised learning, then predict the output will be commuted. In real life a company, business owner may not want to lose any customers but here the segmentation focuses only on the loyal customers.

Chapter 6
Conclusion and Future Work

6.1 Concise of the study

The observation throughout this analysis here offers incorporating RFM analysis into purchaser segmentation strategies to offer market intelligence. The purpose is to analyze a large customer dataset, then segmenting them based on RFM score using RFM method and develop a better business approach for a company on how they are going to target more customers and gain more profit with necessary approaches.

6.2 Conclusion

There may be an essential role in consumer partition for retail corporations. Better segmentation of clients is vital in obtaining a corporation's sales target. Organizations get a higher know-how of the target marketplace if the customers which have equivalent demands, necessities and conduct are grouped together. As a consequence, companies may want to reevaluate the modern course of action and broaden a new technique for reaching better sales, along with; update marketing, price control, promotions, building greater customer touchpoints, and many others.

6.3 Future Work

In this research our main focus is to help any organization in their business plan and achieve goals. So, there can be a golden opportunity for product segmentation besides this customer segmentation which will help an organization one step further. Market basket analysis will show us which product is better, most selling and which product an organization needs to focus more to increase its market share. Product segmentation and Market basket analysis can bring an organization one step forward.

References

[1] V. Vijilesh, A.Harini, M.Hari Dharshini, R.Priyadharshini - CUSTOMER SEGMENTATION USING MACHINE LEARNING,International Research Journal of Engineering and Technology (IRJET),e-ISSN: 2395-0056

[2] A.Joy Christy, A.Umamaheswari, L.Priyadarshini, A.Neyaa,RFM ranking - An effective approach to customer segmentation,Journal of King Saud University –Computer and Information Sciences

[3] AMAN BANDUNI*1, Prof ILAVENDHAN A.2 - CUSTOMER SEGMENTATION USING MACHINE LEARNING, 1,2School of Computing Science & Engineering,Galgotias University, Greater Noida, U.P.

[4] Etzion, O., Fisher, A., & Wasserkrug, S. (2004, Walk). e-CLV: a modeling approach for client lifetime assessment in e-commerce spaces, with an application and case considered for online barters. In innovation, e- Commerce, and e-Service, 2004. EEE'04. 2004 IEEE Worldwide Conference on (pp. 149-156). IEEE.

[5] Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. Management Science, 52(4), 597-612.Cui, G., Wong, M. L. & Lui, H. K. (2006).

[6] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu - Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer ServicesIJARAI) International Journal of Advanced Research in Artificial Intelligence,Vol. 4, No.10, 2015

[7] Cho, Young, Moon, S.C., 2013. Weighted mining frequent pattern-based customer's RFM score for personalized u-commerce recommendation system. J. Converg. 4, 36–40.

[8] Efficiency analysis of the TPA clustering methods for intelligent customer segmentation. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, pp. 784–788.Seshasayee, A., Logeshwari, L., 2017.

[9] Serrano, Stephan. "[Guide]: RFM Analysis w/ predictive segmentation examples." Barilliance,January 14,2022,https://www.barilliance.com/rfm-analysis/

[10] Vardon, Eric. "How to Use Machine Learning for Customer Segmentation". Morphio,May 22,2022,https://morphio.ai/blog/machine-learning-for-customer-segmentation/

[11] Ritchie.Ng. “Identifying Customer Segments (Unsupervised Learning), Unsupervised learning application by identifying customer segments”.(https://www.ritchieng.com/machine-learning-project-customer-segments/)

[12] Jiang, T., Tuzhilin, A., March 2009. Improving personalization solutions through optimal segmentation of customer bases. IEEE Trans. Knowledge Data Eng. 21 (3), 305–320. https:// doi.org/10.1109/ TKDE. 2008.163N.

Appendix A

CEP Mapping

Ks are addressed through the project and mapping among Ks, COS and POS

Ks	Attribute	How ks are addressed through the project	CO	PO
K2	Mathematics	Details and basic of Knowledge Statistics	**CO-1**	**PO-a**
K3	Engineering Fundamental	Knowledge about Machine Learning, Different types of Learning, Machine Learning Algorithms, Data Analysis Tool, and Python Language, and different types of Framework.	**CO-1** **CO-2**	**PO-a** **PO-b** **PO-c**
K4	Specialist Knowledge	Developing Dataset, Data Cleaning, Data Processing, Data Normalization, Feature Extraction, Design, Train & Test.	**CO-1** **CO-2**	**PO-a** **PO-b** **PO-c** **PO-e**
K6	Engineering Practice	Knowledge of programming language python, knowledge of using library, Machine learning based model, Understanding data, idea of large data set, classes of machine learning problem	**CO-1** **CO-2**	**PO-a** **PO-c** **PO-e**
K8	Research Literature	The research requires a detailed study of the related research field and other sources and documentation	**CO-1** **CO-5**	**PO-a** **PO-d** **PO-h**

Ps are addressed through the project and mapping among Ps, COs, and POs

Ps	Attribute	How Ps are addressed through the project	CO	PO
P1	Depth of Knowledge Requirement	Basic & advance statistics knowledge (**K2**) Project requires study of research on Data Science, Data Analysis & Machine Learning Algorithms (**K8**) Data collection from Online shop, super shop, e-commerce site (Facebook page) **(K3, K4)** knowledge of using Library, Machine learning Based Model, Understanding Data, Idea of Large Data Set, Classes of Machine Learning Problem (**K6**)	**CO-1** **CO-2** **CO-8**	**PO-a** **PO-b** **PO-c** **PO-d** **PO-j** **PO-l**
P2	Range of Conflicting Requirements	Developing a practical machine learning model with proper regularization with low variance while limited given data will be used. If the learning process is unsupervised or reinforcement rather than supervised learning, then accurate segmentation will be commuted	**CO-2** **CO-4**	**PO-a** **PO-c** **PO-g**
P3	Depth of Analysis Required	A huge algorithm can be adopted but choice of the selected algorithm requires in-detail and depth analysis	**CO-1** **CO-2**	**PO-a** **PO-b** **PO-d** **PO-l**
P4	Familiarity of Issues	CSE graduates are not typically familiar with customer management, business analytics and business policy.	**CO-5**	**PO-f**
P5	Extent of Applicable Codes	We maintained user privacy carefully as well as took other ethical approaches and used open-licensed tools to develop the system.	**CO-5** **CO-6**	**PO-f** **PO-h** **PO-i**
P6	Diverse Groups	People of all ages and classes are involved specially the more loyal people	**CO-6**	**PO-i**
P7	Interdependence	Research involves a number of sub-system like Data Collection, Training Dataset, Machine Learning Algorithms, Data Analysis, Data Processing.	**CO-3** **CO-6** **CO-8**	**PO-c** **PO-i** **PO-j** **PO-k**

Addressing Complex Activities (As) through the project

As	Attribute	How As are addressed through the project
A1	Range of Resources	The project requires the use of diverse resources including different types of **materials**, **Information's**: dataset (test & training), **Dataset** (https://archive.ics.uci.edu/ml/datasets/Online+Retail+II) **people**: (Members: Tanveer Ahamed Rabby,Md.Efti Khirul Alam, Sharmin Akter)
A2	Level of interaction	The level of interaction between the group members has been varied when it comes to making the dataset in our model. By using data analysis & Machine learning algorithms to segment customers from a large dataset for a particular company.
A4	Consequences for society and the environment	By segmenting customers, it can be easy to understand the divergence between loyal customers and so-called customers. If we can segment the customer which contains some features, it can be easier for a businessman or a company to realize the loyal and targeted customer which will be a more efficient way for business.
A5	Familiarity	The project deals with data analysis and machine learning algorithms for segmenting customer and market basket analysis for learners.

Course Outcomes (CO) with PO mapping for this project

CO No.	CO Statements	Corresponding POs
CO1	We identified, formulated, and analyzed the real-world problem of Super shop/E-commerce by segmenting their customer analyzing dataset.	**a** **b** **d** **l**
CO2	We proposed a solution using data science and machine learning algorithms to segment the customer for any kind of enterprise as well as find the most selling item of a company.	**a** **c** **e**
CO3	Expected completion time of 6 months, initially for analysis no budget required.	**k**
CO4	Marketers can be extra efficient in phrases of time, cash, and other sources through segmenting their facts. Marketplace segmentation permits organizations to have a higher understanding in their clients. They gather a greater draw close to the requirements and desires of customers, allowing them to personalize campaigns to the consumer groups most likely to buy objects.	**g**
CO5	Understood the concept of professional ethics, confidentiality, industrial standards, risk-benefit analysis and explained the impact of engineering solutions in social safety, data safety, and welfare from the Code of Ethics (https://www.acm.org/code-of-ethics)	**f** **h**
CO6	Function effectively in a multidisciplinary team	**i**
CO8	The present design,analysis,analysis output, documentation through oral presentations	**c** **j**

Washington Accord Program Outcomes (PO) for engineering programs

No.	PO	Differentiating Characteristic
a	Engineering Knowledge	Breadth and depth of education and type of knowledge, both theoretical and practical
b	Problem Analysis	Complexity of analysis
c	Design/ development of solutions	Breadth and uniqueness of engineering problems such as the extent to which problems are original and to which solutions have previously been identified or codified
d	Investigation	Breadth and depth of investigation and experimentation
e	Modern Tool Usage	Level of understanding of the appropriateness of the tool
f	The Engineer and Society	Level of knowledge and responsibility
g	Environment and Sustainability	Type of solutions.
h	Ethics	Understanding and level of practice
i	Individual and Team work	Role in and diversity of team
j	Communication	Level of communication according to type of activities performed
k	Project Management and Finance	Level of management required for differing types of activity
l	Lifelong learning	Preparation for and depth of Continuing learning.