CUSTOMER SEGMENTATION BASED ON PURCHASING BEHAVIOUR

MELVIN BIJU

1631027

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

M.Sc. (Software Systems)

OF ANNA UNIVERSITY

April 2021

DEPARTMENT OF COMPUTING

COIMBATORE INSTITUTE OF TECHNOLOGY

(Autonomous Institution affiliated to Anna University)

COIMBATORE – 641014

COIMBATORE INSTITUTE OF TECHNOLOGY

(Autonomous Institution affiliated to Anna University)

COIMBATORE 641014

(Bonafide Certificate)

Project Work - II
Tenth Semester

CUSTOMER SEGMENTATION BASED ON PURCHASING BEHAVIOUR

Bonafide record of work done by
MELVIN BIJU
(Register No: 1631027)

Submitted in partial fulfilment of the
requirements for the degree of
M.Sc. (Software Systems)
of Anna University

April 2021

_____ _____

Faculty Guide Head of the Department

Submitted for the viva-voce held on _____

_____ _____

Internal Examiner External Examiner

CONTENTS

ACKNOWLEDGEMENT

I sincerely thank **Dr. V. SELLADURAI**, Principal, Coimbatore Institute of Technology, for giving me an opportunity to undertake this full-time industry project.

I would like to express my gratitude towards **Dr. K. SAKTHIMALA,** Head, Department of Software Systems, for her motivation to explore new things and to achieve heights.

I sincerely thank my internal guide**, Dr. C SATHYA,** Assistant Professor, Department of Data Science, for her exhilarating supervision, timely suggestions, and encouragement through all the phase of this work.

I also express my sincere thanks and deep sense of gratitude to my external guide **Mr. Ankur Bharadwaj**, Data Scientist, Ganit Business Solutions Pvt Ltd, Chennai, for helping me in various aspects to complete this project work.

I would like to thank **GANIT BUSINESS SOLUTIONS PVT LTD**, Chennai for providing me an opportunity to work in their concern and others who have helped me in bringing out this project in complete shape.

SYNOPSIS

The project titled "**CUSTOMER SEGMENTATION BASED ON PURCHASING BEHAVIOUR**" aims at segmenting the customers by focusing on the purchasing behaviour in order for companies to derive strategies and insights using historical transaction data. This analysis is done under "**GANIT BUSINESS SOLUTIONS PVT LTD**" for the retail company **- MORE RETAIL LIMITED (MRL)**

In today's fast-moving world of marketing from product-orientation to customer-orientation, the management of customer treatment can be seen as a key to achieve revenue growth and profitability. B2B or business customers are more complex, their buying process is more complicated, and their sales value is greater. Hence no knowledge of customer behaviour can lead to wastage of resources and potential loss of customers and revenue.

For analysis, the MRL retail customer transaction data is used. To find out how value is generated for customers in this segment, a theoretical framework (RFM) consisting of customer-perceived value and its influences on buying behavior has been used. K-means clustering machine learning algorithm is implemented for finding out the clusters in the data. This project makes extensive use of analysis tools like python and R. It also utilizes visualization tools like Tableau. The reports generated after analysis can be used to create segment specific marketing and sales strategies and as a result become customer focused.

PREFACE

Chapter I introduces the organization where the project is done. It continues to give insight into the problem that this project aims to solve and overview of the models used to solve the problem with the summary inferred from the results.

Chapter II describes about the data with its features and exploratory data analysis results which makes the user to understand the data clearly. It continues to explain briefly about the type of analysis, techniques used to solve the problem.

Chapter III provides a detailed note of the learning model design. It continues to give insights about the description of the tools used to analysis process automation. It finally shows the user interface through which the results of the analysis will be displayed to the user.

Chapter IV deals with the analysis process evaluation where the description of the performance measure used to evaluate the analysis process and the results are given.

Chapter V discusses the detailed inferences of the result and visual representations of the analysis report.

Chapter VI has the features of the project and discusses future scope of the completed project.

CHAPTER 1

INTRODUCTION

This section gives a detailed description of the organization for which the system is developed along with the overview of the existing system, problems associated with it, goals, and the scope of the proposed system. It also specifies the system environment used in the development of the proposed system and gives brief introduction about the various technologies used in the development of the system.

1.1 ORGANIZATION PROFILE

Ganit Inc. is a consulting firm specializing in Data Analytics, Big Data, Data Engineering, Machine Learning, AI, and IoT. It was founded in 2017. They provide Decision Making Power (DMP) for companies by providing solutions at the intersection of hypothesis-based analytics, discovery-based AI and IoT.

The solutions are a combination of customized services and functional product suite. The aim of the company is to embed these solutions into the bloodstream of our customers' decision-making process. For this it uses sophisticated tools and techniques to mine Big or small data emerging from transactions, behaviours, macro-economic conditions, social interactions, IOT devices etc. Ganit has capabilities across reporting & dashboarding, inquisitive analytics, predictive analytics, and machine learning. The solutions are easy to consume and implement.

1.2 PROBLEM STATEMENT

Retailers look towards increasing their sales figures and retain customers by segmenting them into different categories based on their purchasing behaviour. Conclusions from the segmentation can be used for deriving business insights and appropriate strategies.

1.2.1 Objective

- To implement a system to find value generated in a segment and how a company can adapt its marketing and sales activities using the metric.
- To find out how value is generated for customers in a particular segment, using a theoretical framework (RFM) consisting of customer-perceived value and its influences on buying behavior.
- To use the understanding for creating segment specific marketing and sales strategies.

1.2.2 Scope

- Isolate the most valuable customers of the company.
- Identify the kinds of customers the company have.
- This can be used for targeted marketing and other marketing strategies.
- Sometimes it can even reveal a potential white space in the marketplace which no company has yet occupied.

1.2.3 Users

Users of this system can include middle and high tier employees like:

- Engagement managers
- Marketing executives
- HR and sales executives

1.2.4 Domain

The RFM metrics are applicable for the retail industry that have huge data dumps. The data dumps are analysed for feature extraction upon which analysis is done and insights are generated.

1.3 DESCRIPTIVE STATISTICAL SUMMARY

1.3.1 Dataset size

For this analysis, the MRL transaction dataset is used. The records involve transaction data for their retail outlets in “Punjab” and between the time period: 1/01/2020 - 31/12/2020.

```
df1 = df.dropna()
df1.shape
```

```
(17075577, 7)
```

Figure 1.1 Dataset Size

1.3.2 Dataset summary

The summary table gives the total statistical description for the chosen data set. For example, the “**bill_qty”** column’s descriptive statistics for centre of data and spread of data are:

1. **Mean:** 102.84
2. **Standard Deviation:** 186.1

	location_code	mobile_no	tran_date	bill_no	item_no	bill_qty	bill_value
count	1.707558e+07	1.707558e+07	17075577	1.707558e+07	1.707558e+07	1.707558e+07	1.707558e+07
unique	NaN	NaN	334	NaN	NaN	NaN	NaN
top	NaN	NaN	2020-03-19	NaN	NaN	NaN	NaN
freq	NaN	NaN	137906	NaN	NaN	NaN	NaN
mean	2.526180e+03	9.222074e+09	NaN	5.496539e+09	1.003812e+08	1.517188e+00	1.028458e+02
std	7.074325e+02	8.846711e+08	NaN	2.235727e+09	5.258117e+05	2.860047e+00	1.860988e+02
min	1.459000e+03	6.000001e+09	NaN	2.083221e+08	1.000000e+08	-1.800000e+02	-2.691000e+04
25%	1.804000e+03	8.749056e+09	NaN	4.316231e+09	1.000559e+08	1.000000e+00	3.430000e+01
50%	3.045000e+03	9.646204e+09	NaN	6.420107e+09	1.001218e+08	1.000000e+00	6.135000e+01
75%	3.096000e+03	9.872207e+09	NaN	7.266137e+09	1.005090e+08	1.144000e+00	1.200000e+02
max	3.582000e+03	1.000000e+10	NaN	9.000003e+09	1.019329e+08	1.000000e+03	1.070816e+05

Figure 1.2 Summary table

1.3.3 Shape and spread of data

On visualization of the dataset (Figure 1.3), it is observed that **bill_qty** is **positively skewed.** A large range of customers shop for a low amount per transaction. Hence suitable modifications are made during analysis to obtain a more accurate model.

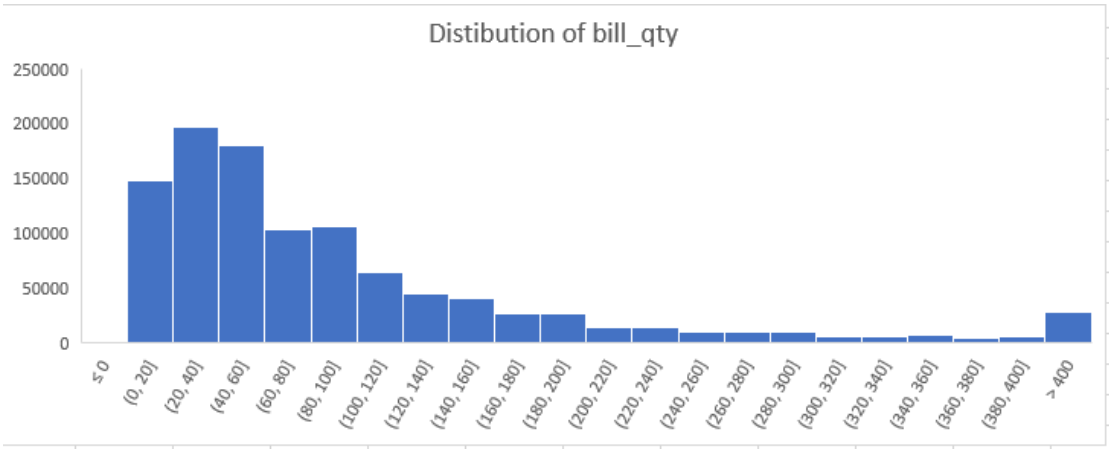


Figure 1.3 bill_qty distribution

Figure 1.3 shows the positively skewed distribution of bill_qty.

1.4 INFERENCES SUMMARY

After exhaustive testing and clustering, seven segments are identified based on customer's purchasing behaviour. They are segmented as:

- Best Customers
- Newbies
- Promising Customers
- Loyal
- High spending Customers
- Slipping Customers
- Lost customers

After segmentation, these customers can be included into some alerting system where SMS and emails can be sent on daily basis regarding the offers and discounts while for other segments an e-mail blast can be set once per week in a month for blast SMSs to notify them about corresponding products.

There can be many marketing, customer retention strategies applied for Customers on these Cluster Analysis. Sample suggested strategy is:

- **Best Customers:** Value added offers along with personalized product recommendations.
- **Newbies:** Welcome emails with introductory offers.
- **Promising Customers:** Product recommendations based on past purchases.
- **Loyal:** Membership benefits.
- **High Spending:** Recommend the best and expensive products.
- **Lost Customers:** "We miss you" e-mails along with information about new products.
- **Slipping Customers:** Retention strategies to increase customer engagement e.g., live events.

A customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities.

When a group of customers is sent personalized messages as part of a marketing mix that is designed around their needs, it is easier for companies to send those customers special offers meant to encourage them to buy more products.

Customer segmentation can also improve customer service and assist in customer loyalty and retention. As a by-product of its personalized nature, marketing materials sent out using customer segmentation tend to be more valued and appreciated by the customer who receives them as opposed to impersonal brand messaging that does not acknowledge purchase history or any kind of customer relationship.

Finally with customer segmentation, Companies will stay a step ahead of competitors in specific sections of the market and identify new products that exist, or potential customers could be interested in or improving products to meet customer expectations.

CHAPTER 2

DATA MODELLING AND EXPLORATION

This chapter depicts the data and its features with exploratory data analysis results which makes the user to understand the data clearly. It continues to give deep insights about the type of analysis, algorithms and methods used to achieve the objective of the project.

2.1 PROBLEM ANALYSIS

2.1.1 Problem Understanding

In a world where large and vast amount of data is collected daily, analysing such data is an important need. In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according to the modern conditions. The business done today runs based on innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers.

2.1.2 Business Understanding

Over the years, the competition amongst businesses is increased and the large historical data that is available has resulted in the widespread use of data mining techniques in extracting the meaningful and strategic information from the database of the organisation. Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. The customer segmentation has the importance as it includes the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with each customer segment and to manage the demand and supply

of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions. The customer segmentation technique allows the business to make better use of their marketing budgets, gain a competitive edge over their rival companies, demonstrating the better knowledge of the needs of the customer. It also helps an organization in, increasing their marketing efficiency, determining new market opportunities, making better brand strategy, identifying customers retention.

2.1.3 Feature Identification

The features identified for this analysis are:

- **mobile_no**
- **trans_date**
- **item_no**
- **bill_value**

The figure below shows that after dropping duplicates, the dataframe has the same number of rows as the original. Thus, it can be confirmed that a combination of these columns gives the total records

```
df = pd.read_csv("punjab_data.csv")
```

```
df1 = df.dropna()
df1.shape
```

```
(17075577, 7)
```

```
df1[['mobile_no','bill_no','item_no','tran_date']].drop_duplicates().shape
```

```
(17075577, 4)
```

Figure 2.1 Feature Identification

2.2 DATA MODEL

2.2.1 Data collection

The data used for analysis is stored in the Amazon Redshift Warehouse. The data is extracted using data pull queries. Table used is **Transaction_data.**

Dataset schema

A dataset schema helps in the interpretation of information in the table. Schemas can be useful because it allows to take shortcuts in interpreting the vast amount of information that is available in our environment.

Table 2.1 Transaction_data

Sno.	***Field Name***	***Data_type***	***Data Description***	***Example***	***Nullable***
1	location_code	varchar	Unique code for each store	1357	YES
2	cust_id	varchar	Unique ID for each customer	LP03666196	YES
3	mobile_no	varchar	Mobile number of customers	986****570	YES
4	tran_type	varchar	Indicates the transaction type: ale/Return	SALE	YES
5	sub_tran_type	varchar	Indicates the mode of Payment such as Online/Offline	OFFLIN	YES
6	tran_datetime	timestamp	Time of transaction	2019-11-01 9:08:33	YES
7	bill_no	varchar	Bill number	5433435616	YES
8	amz_bill_no	varchar	Amazon bill number	407-1897-5321	YES
9	item_no	varchar	Unique code for each SKU	100000001	YES
10	bill_qty	numeric	No. of items bought	1	YES
11	bill_value	numeric	Realised sales value (in INR)	14.25	YES

Table 2.1 illustrates the various fields and its description for analysis

2.2.2 Data transformation

The level of the table is identified and desired feature extraction. The level of the table is identified as:

mobile_no, tran_datetime, bill_no, item_no

These particular columns are chosen because they constitute the attributes of the RFM metrics. **trans_date** is a timestamp from which month and date is extracted for further analysis of the R (Recency) Metric.

2.2.3 Data loading and preparation

The chosen columns are compiled as a table and are loaded for further analysis. The dataset is validated for null values of **mobile_no** and **bill_no.** Null values in these fields are not considered because tracing back to the customer is not possible without a valid mobile number. Also, a null bill_no voids the entire transaction. Hence row not conforming to these are completely excluded.

Table 2.2 Dataset after feature extraction and transformation

	location_code	mobile_no	tran_date	bill_no	item_no	bill_qty	bill_value
0	2001	9.877990e+09	2020-05-06	7766044401	100026048	1.0	95.0
1	1956	9.779075e+09	2020-04-27	4130392627	100005332	1.0	110.0
2	2090	9.501055e+09	2020-12-01	4358571939	100068234	1.0	65.0
3	3315	8.146930e+09	2020-06-03	7670045135	100115606	3.0	30.0
4	3055	7.696762e+09	2020-09-12	7229133922	100142024	1.0	37.0
...	...	...	...	...	...	...	...
18993680	3582	9.041621e+09	2020-12-02	8917001342	100148966	1.0	45.0
18993681	1951	8.950526e+09	2020-03-23	4033415721	100096450	1.0	29.4
18993682	1951	8.146908e+09	2020-03-23	3794453792	100141041	2.0	18.0
18993683	3582	9.417085e+09	2020-12-19	8918001652	100083280	2.0	80.0
18993684	1951	9.852690e+09	2020-03-13	4033414969	100146752	1.0	20.0

2.3 EXPLORATORY DATA ANALYSIS

2.3.1 Handling Outliers

A data point is defined as an outlier if its isolation number is lower than the threshold. The threshold is defined based on the estimated percentage of outliers in the data, which is found through outliner detection algorithms.

2.3.2 Turkey outliner detection

Tukey Fences uses the interquartile range ("IQR") to flag observations that are considered outliers. Tukey Fences defines an outlier based on the below formula:

Outlier = Transaction > Q3 + 1.5 x IQR OR Transaction < Q1–1.5 x IQR

- Q1 & Q3 represent the first and third quartile values
- IQR represents Q3 minus Q1
- 1.5 is used as the multiplier and typically represents the upper and lower ends of a boxplot.

2.3.3 Grouping Operations

The dataset is grouped by **mobile_no since** the objective of this analysis is to target individual customers. Thus, each unique mobile number is grouped, and further analysis is proceeded.

Table 2.3 R metric after grouping based on mobile_no

	mobile_no	bill_value	month
0	6.000001e+09	698.6260	1
1	6.000034e+09	111.0000	1
2	6.000289e+09	0.0270	1
3	6.000432e+09	3148.1300	1
4	6.000470e+09	2390.2000	1

2.3.4 Binning

Binning is a way to group several continuous values into a smaller number of "bins". For the Recency metric (R score) the data is binned into 4 clusters:

- Less than 1 month
- 2-3 months
- 3-6 months
- 6-12 months

Table 2.4 R metric after binning

B	C	D	E
mobile	tran_date	Rvalue	Bins
6E+09	17-07-2020	168	6-12 months
6E+09	16-10-2020	77	< 2 - 3 months
6E+09	26-05-2020	220	6-12 months
6E+09	09-10-2020	84	< 2 - 3 months
6E+09	14-11-2020	48	< 2 - 3 months
6E+09	12-11-2020	50	< 2 - 3 months
6E+09	04-11-2020	58	< 2 - 3 months
6E+09	30-04-2020	246	6-12 months
6E+09	16-10-2020	77	< 2 - 3 months
6E+09	25-12-2020	7	< 1 month
6E+09	23-09-2020	100	3-6 months
6E+09	15-07-2020	170	6-12 months
6E+09	18-12-2020	14	< 1 month
6E+09	17-07-2020	168	6-12 months
6E+09	14-07-2020	171	6-12 months
6E+09	12-10-2020	81	< 2 - 3 months

Table 2.4 shows the distinct customers being binned to the above-mentioned clusters.

2.3.4 Log Transformation

The log transformation is used to transform skewed data to approximately conform to normality. If the original data follows a log-normal distribution or

approximately so, then the log-transformed data follows a normal or near normal distribution. In this case, the log-transformation does remove or reduce skewness. The logarithmic transformation compresses the differences between the upper and lower part of the original scale of data.

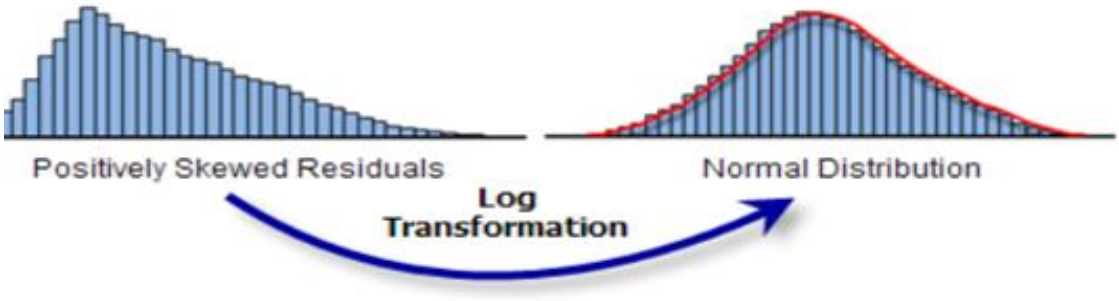


Figure 2.1 Log Transformation

Figure 2.1 shows the structural difference between positively skewed data and a normal distribution.

CHAPTER III

PREDECTIVE ANALYTICS PROCESS

3.1 PREDECTIVE ANALYTICS MODEL

Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyse current and historical facts to make predictions about future or otherwise unknown events. In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision-making for candidate transactions.

3.1.1 Clustering model

Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more like other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals, or statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. It is often necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

3.1.2 RFM Analysis

After data is pre-processed, check for recent transactions, frequency and the amount spent by the customers is done. To create recency variable, the reference date - that is one day prior to the last transaction is decided. RFM analysis is a very popular customer segmentation and identifiable technique in database marketing. It is significant especially in Retail Industry. Each customer under RFM is scored based on three factors.

- ***Recency****:* It refers to the number of days before the reference date when a customer made the last purchase. Lesser the value of recency, higher is the customer visit to a store.

- ***Frequency****:* It is the period between two subsequent purchases of a customer. Higher the value of Frequency, more is the customer visit to the company.

- ***Monetary****:* This refers to the amount of money spent by a customer during a specific period of time. Higher the value, more is the profit generated to the company.

3.1.3 K-means Clustering

K-means clustering is widely used in the field of cluster analysis and customer segmentation. K-means is an algorithm designed to group a set of items into K subgroup or clusters. The algorithm is dependent on a manually set value for K. The K centroids are initialized to random observations in the dataset. K-means is then tasked with iteratively moving these centroids to minimize the cluster variance using two steps:

- for each centroid c identify the subset of items that are closer to c than any other centroid using some similarity measure.
- calculate a new centroid each cluster after every iteration which is equal to the mean vector of all the vectors in the cluster.

This two-step process is repeated until convergence is reached.

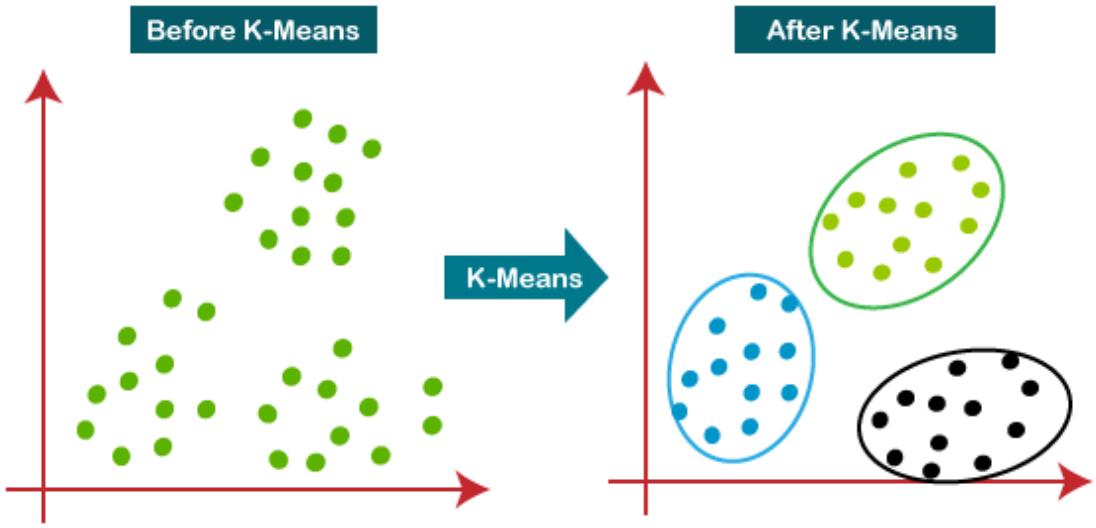


Figure 3.1 K-means Clustering

Figure 3.1 shows the data points being clustered after implementing K-Means algorithm.

Mathematically it is given by,

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

where w_{ik}=1 for data point x_i if it belongs to cluster k; otherwise, w_{ik}=0.

Also, μ_k is the centroid of x_i's cluster.

The standard implementation of K-means uses Euclidian distance measure described in the section above to find the subset of items that corresponds to each cluster. This is done by calculating mean squared error, which in this case is equivalent with the Euclidian distance, of each item's feature vector with the K centroid and

choosing the closest result. However, other distance measures can be used instead of Euclidian distance

3.1.4 The Elbow Method

The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned centre.

When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k. The "arm" can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

Elbow Method uses Within Cluster Sum of Squares (WCSS) against the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formula:

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

where Y_i is centroid for observation X_i.

The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.

3.1.5 Results

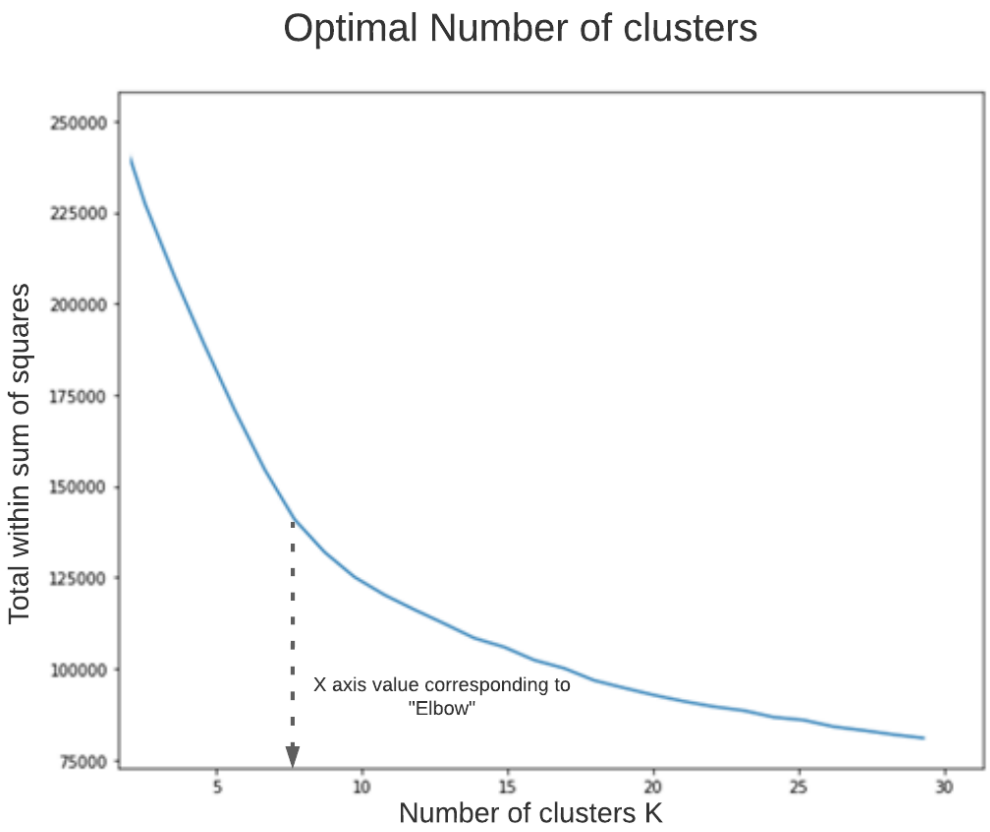


Figure 3.2 Elbow Method

Figure 3.2 shows the graph obtained using elbow method and how optimal k value is selected.

To determine the optimal number of clusters, the value of k at the “elbow” i.e., the point after which the distortion/inertia start decreasing in a linear fashion should be selected. Thus, for the given data, the optimal number of clusters for the data is concluded as **7**.

3.2 TOOLS DESCRIPTION

3.2.1 Software Used

Microsoft Excel VBA

Microsoft Excel has the basic features of all spreadsheets using a grid of *cells* arranged in numbered *rows* and letter-named *columns* to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering, and financial needs. In addition, it can display data as line graphs, histograms, and charts, and with a very limited three-dimensional graphical display.

AWS Redshift

It is an Internet hosting service and data warehouse product which forms part of the larger cloud-computing platform Amazon Web Services. Redshift handles large scale data sets and database migrations.

Tableau

Tableau helps user transform data into actionable insights. Explore with limitless visual analytics. Build dashboards and perform ad hoc analyses in just a few clicks. Share your work with anyone and make an impact on user business. From global enterprises to early-stage start-ups and small businesses, people everywhere use Tableau to see and understand their data.

R Studio

The RStudio IDE is a set of integrated tools designed to help you be more productive with R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing

history, debugging, and managing your workspace. R is an environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

3.2.2 Libraries Used

Pandas

Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

NumPy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The core functionality of NumPy is its "ndarray", for n-dimensional array, data structure. These arrays are strided views on memory. In contrast to Python's built-in list data structure, these arrays are homogeneously typed: all elements of a single array must be of the same type.

Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. This is a standard data science library that helps to generate data visualizations such as two-dimensional diagrams and graphs (histograms, scatterplots, non-Cartesian coordinates graphs). Matplotlib is one of those plotting libraries that provides an object-oriented API for embedding plots into applications.

Seaborn

Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures. It is a dataset-oriented API for examining relationships between multiple variables. It has specialized support for using categorical variables to show observations or aggregate statistics. It contains options for visualizing univariate or bivariate distributions and for comparing them between subsets of data. It employs automatic estimation and plotting of linear regression models for different kinds of dependent variables

Scikit-learn

Scikit-learn is an indispensable tool in data science used for making machine learning models. Scikit-learn is probably the most useful library for machine learning in Python. Scikits is a group of packages in the SciPy Stack that were created for specific functionalities – for example, image processing. Scikit-learn uses the math operations of SciPy to expose a concise interface to the most common machine learning algorithms. Data scientists use it for handling standard machine learning and data mining tasks such as clustering, regression, model selection, dimensionality reduction, and classification.

dplyr

dplyr provides a set of tools for efficiently manipulating datasets in R. dplyr is the next iteration of plyr, focussing on only data frames. dplyr is fast, has a more consistent API and should be easier to use. Dplyr can be used to do manipulations into a local data frame. PostgreSQL, MySQL, SQLite, and Google big query support are built in. Adding a new backend can be done by implementing a handful of S3 methods.

3.3 IMPLEMENTATION USING TOOL

3.3.1 Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import date
import datetime as dt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
```

3.3.2 Pseudocode for outliner detection

```
for x in ['Fvalue']:
    q75,q25 = np.percentile(RFM_Values2.loc[:,x],[90,5])
    intr_qr = q75-q25

    max = q75+(1.5*intr_qr)
    min = q25-(1.5*intr_qr)

    RFM_Values2.loc[RFM_Values2[x] < min,x] = np.nan
    RFM_Values2.loc[RFM_Values2[x] > max,x] = np.nan
```

3.3.3 Pseudocode for K- means clustering

```
#Handle negative and zero values to handle infinite numbers during log transformation

def handle_neg_n_zero(num):

    if num <= 0:

        return 1

    else:

        return num

#Apply handle_neg_n_zero function to Recency and Monetary columns

RFM_Values1['Rvalue'] = [handle_neg_n_zero(x) for x in RFM_Values1.Rvalue]

RFM_Values1['Mvalue'] = [handle_neg_n_zero(x) for x in RFM_Values1.Mvalue]


#Perform Log transformation to bring data into normal or near normal distribution

Log_Tfd_Data = RFM_Values1[['Rvalue','Fvalue', 'Mvalue']].apply(np.log, axis = 1).round(3)


from sklearn.cluster import KMeans

sum_of_sq_dist = {}

for k in range(1,15):

    km = KMeans(n_clusters= k, init= 'k-means++', max_iter= 1000)

    km = km.fit(Scaled_Data)

    sum_of_sq_dist[k] = km.inertia_
```

```
#Plot the graph for the sum of square distance values and Number of Clusters
sns.pointplot(x = list(sum_of_sq_dist.keys()), y = list(sum_of_sq_dist.values()))
plt.xlabel('Number of Clusters(k)')
plt.ylabel('Sum of Square Distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```

3.3.4 Pseudocode for log transformation

```
def handle_neg_n_zero(num):
    if num <= 0:
        return 1
    else:
        return num
#Apply handle_neg_n_zero function to Recency and Monetary columns
RFM_Values1['Rvalue'] = [handle_neg_n_zero(x) for x in RFM_Values1.Rvalue]
RFM_Values1['Mvalue'] = [handle_neg_n_zero(x) for x in RFM_Values1.Mvalue]

#Perform Log transformation to bring data into normal or near normal distribution
Log_Tfd_Data = RFM_Values1[['Rvalue','Fvalue', 'Mvalue']].apply(np.log, axis = 1).round(3)
```

3.3.5 Pseudocode for analysis and visualization

```
plt.figure(figsize=(20,10))

sns.scatterplot(data=RFM_Values2_sample, x="mobile_no", y="Rvalue",
hue="Segment", palette="Set2")

# plt.yticks(np.arange(100, 10000, 1000))

plt.show()

plt.figure(figsize=(20, 20))

RFM_Values1.Segment.value_counts(sort=False).plot.pie(textprops={'fontsize': 27})

plt.show()

ax2 = R_score_group1.plot.count(x='',

                y='Rvalue',

                c='DarkBlue')


plt.figure(figsize=(20, 20))

sns.distplot(R_score_group1.Rvalue, rug = True,bins = 3)


sns.distplot(Rvalue,bins=30,kde=False)

plt.yticks(np.arange(1, 365, 30))

plt.show()

R_score_group = df1[['mobile_no','tran_date']].groupby('mobile_no')

R_score_group = R_score_group.max()

R_score_group['Rvalue'] = today_date - R_score_group['tran_date']
```

```
R_score_group1 = R_score_group.reset_index()

R_score_group1['Rvalue'] = R_score_group1['Rvalue'].dt.days.astype('int16')

R_score_group = df1[['mobile_no','tran_date']].groupby('mobile_no')

R_score_group = R_score_group.max()

R_score_group['Rvalue'] = today_date - R_score_group['tran_date']

R_score_group1 = R_score_group.reset_index()

R_score_group1['Rvalue'] = R_score_group1['Rvalue'].dt.days.astype('int16')

F_score_group = df1.groupby(['mobile_no'],as_index=False).agg({"bill_no": "count",
"month": pd.Series.nunique})


F_score_group['VMA'] = F_score_group['bill_no']/F_score_group['month']

F_score_group

M_score_group['TMV'] = M_score_group['bill_value']/M_score_group['month']

M_score_group.describe()
```

CHAPTER 4

ANALYTICAL MODEL EVALUATION

Model evaluation metrics are used to assess goodness of fit between model and data, to compare different models, in the context of model selection, and to predict how predictions (associated with a specific model and data set) are expected to be accurate.

4.1 HYPOTHESIS TESTING

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. To perform k-means clustering, the dataset must follow a normal distribution. To find if the dataset (after log transformation and removing outliners) is from a normally distributed dataset, the Shapiro wilk's Test is performed.

4.1.1 Shapiro Wilk's Test

The Shapiro-Wilk test is a way to tell if a random sample comes from a normal distribution. The test gives a W value; small values indicate the sample is not normally distributed.

Null Hypothesis: The given sample follows a normal distribution

Alternate Hypothesis: The given sample does not follow a normal distribution

The formula for the W value is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where:

- x_i are the ordered random sample values
- a_i are constants generated from the covariances, variances and means of the sample (size n) from a normally distributed sample.

The test has limitations, most importantly that the test has a bias by sample size. The larger the sample, the more likely you will get a statistically significant result.

The null hypothesis of this test is that the population is normally distributed. Thus, if the p value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the p value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) cannot be rejected (e.g., for an alpha level of .05, a data set with a p value of less than .05 rejects the null hypothesis that the data are from a normally distributed population).

Like most statistical significance tests, if the sample size is sufficiently large this test may detect even trivial departures from the null hypothesis (i.e., although there may be some statistically significant effect, it may be too small to be of any practical significance); thus, additional investigation of the effect size is typically advisable

4.1.2 Results

The dataset is tested for the M value before transformation and is found to be positively skewed and not normally distributed as shown in Figure 4.1. Skewness should be approximately zero for a normal distribution.

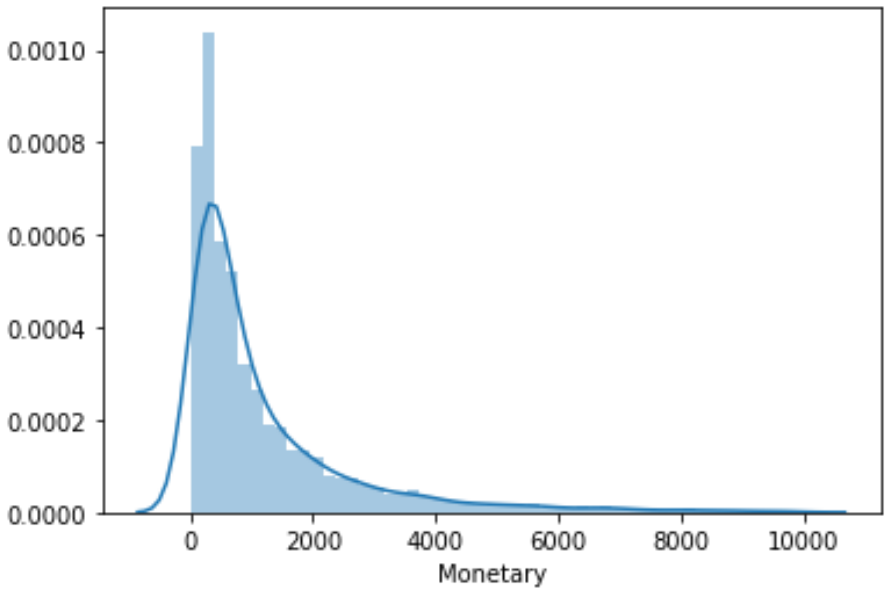


Figure 4.1 Visualization of distribution

After the calculation of the W statistic form the Shapiro Wilk's test, a very low p value is obtained as shown in Figure 4.2, thus the null hypothesis is not accepted, and it is concluded that the sample is not from a normally distributed dataset.

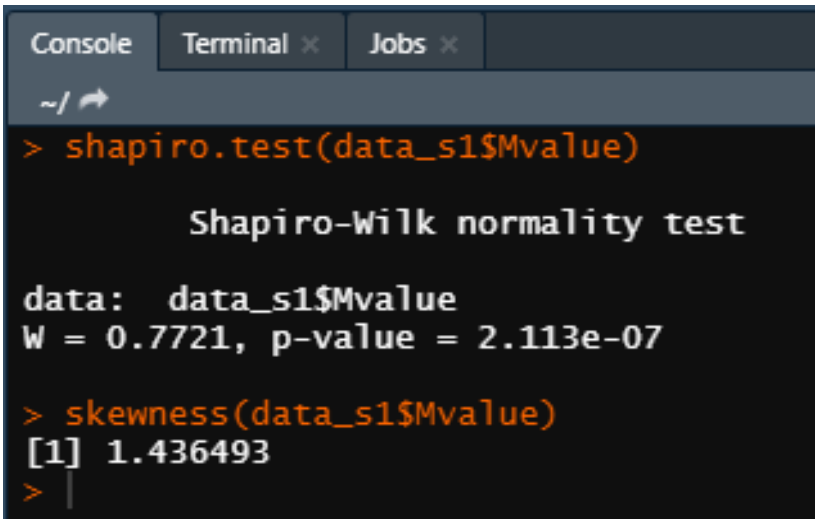


Figure 4.2 Shapiro Wilk's test for skewed data

After scaling and applying log transformation, visualization of the distribution is done again (Figure 4.3). This time a “bell curve” shape is obtained, indicating a normal distribution.

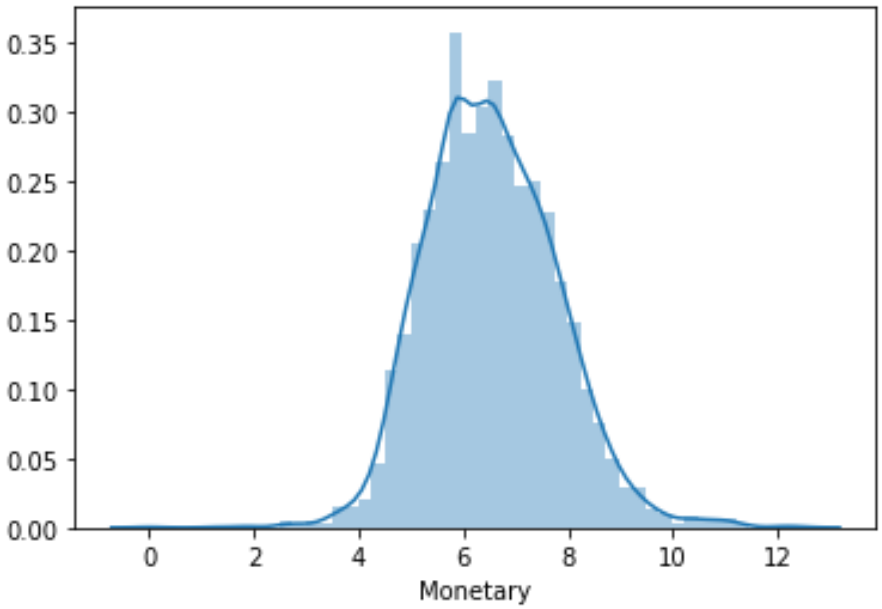


Figure 4.3 Visualization of distribution (After transformation)

Subsequently the Shapiro Wilk’s test is performed again which gives p value = 0.6524 as shown in Figure 4.4. The skewness is also greatly reduced. With this transformed dataset it is possible to get a better outcome after performing k-means clustering.

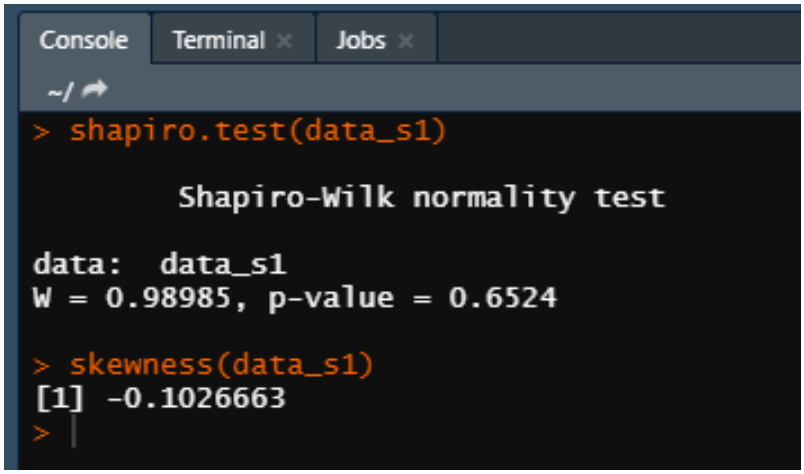


Figure 4.4 Shapiro Wilk’s test (After transformation)

CHAPTER 5

ANALYSIS REPORTS AND INFERENCES

This chapter provides deep insights of the analysis done and it also continues to explain the inferences of the chart results.

5.1 VISUAL FORMATS

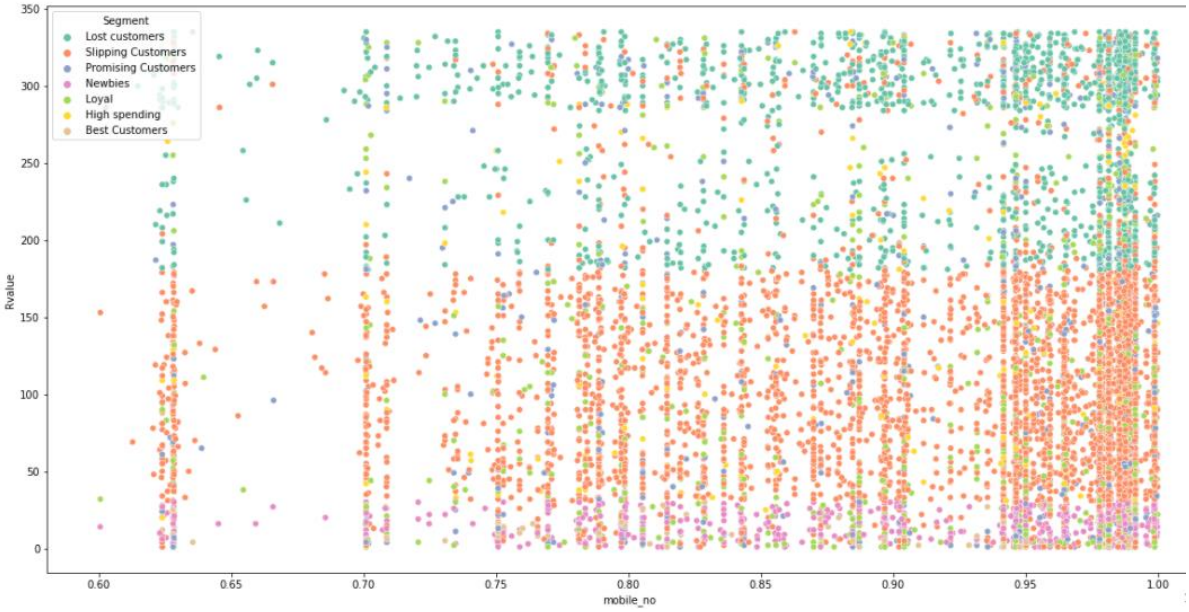
Visual formats include infographics, illustrations, and photographs, which are good for communicating a lot of information in a small space. Some visual formats can also be included in written reports, summaries, blogs, or presentations

The report submitted is in the form of infographic which gives a higher visual look of the data and insights. The visual representation of the charts produces an easy understanding of the analysis data, along with the insights it becomes a good visual report.

The visual report contains many charts that help users to look at the output results. A sample report is shown for a better understanding. The insights are published to the client periodically with the help of infographic for a better understanding.

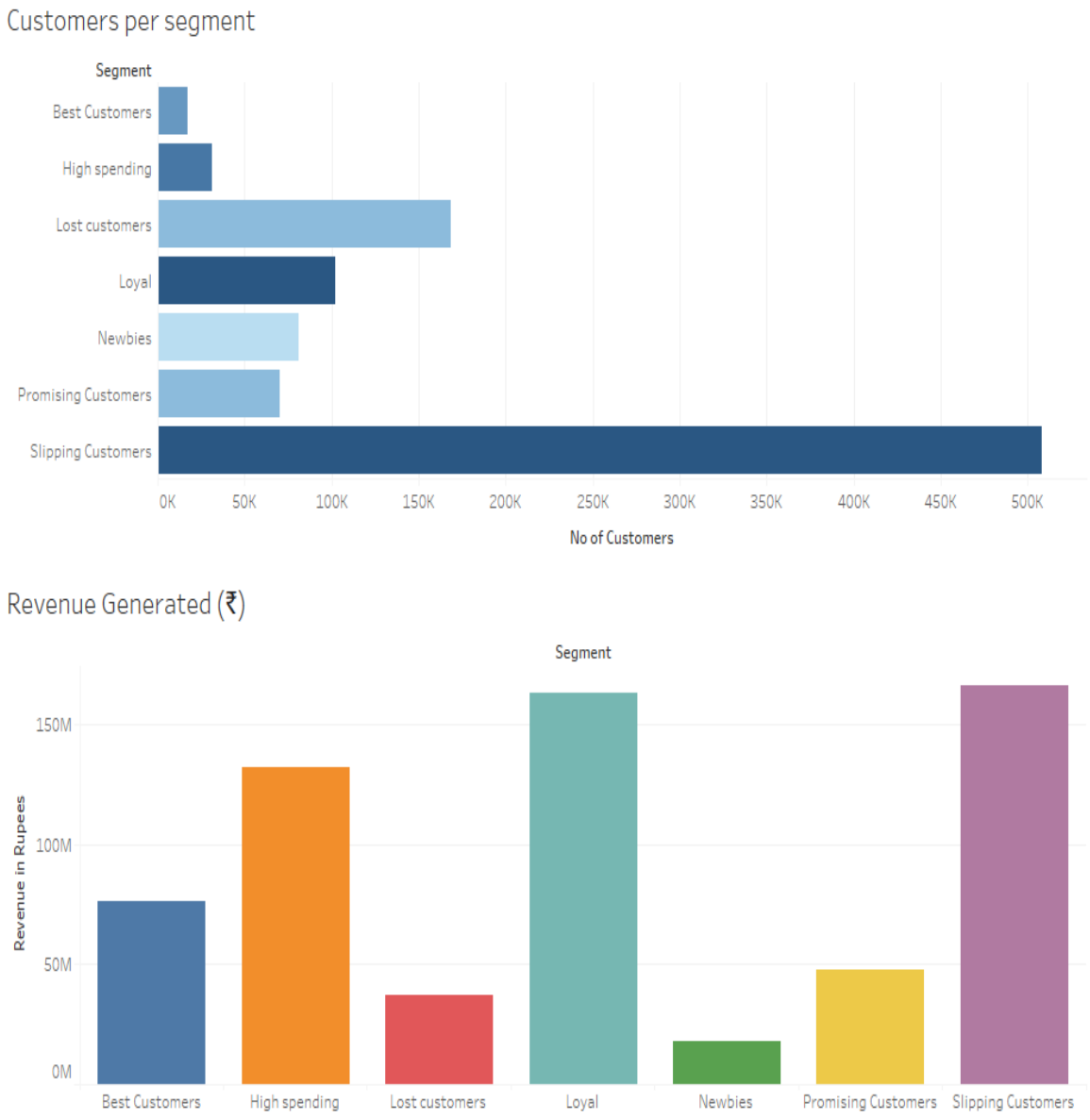
Screen 5.1 Cluster visualization for R value

Screen 5.1 shows the segments of customers based on the recency score. The customers with high recency values are customers that have not visited the store in a very long time.



Screen 5.2 Cluster visualization for M value

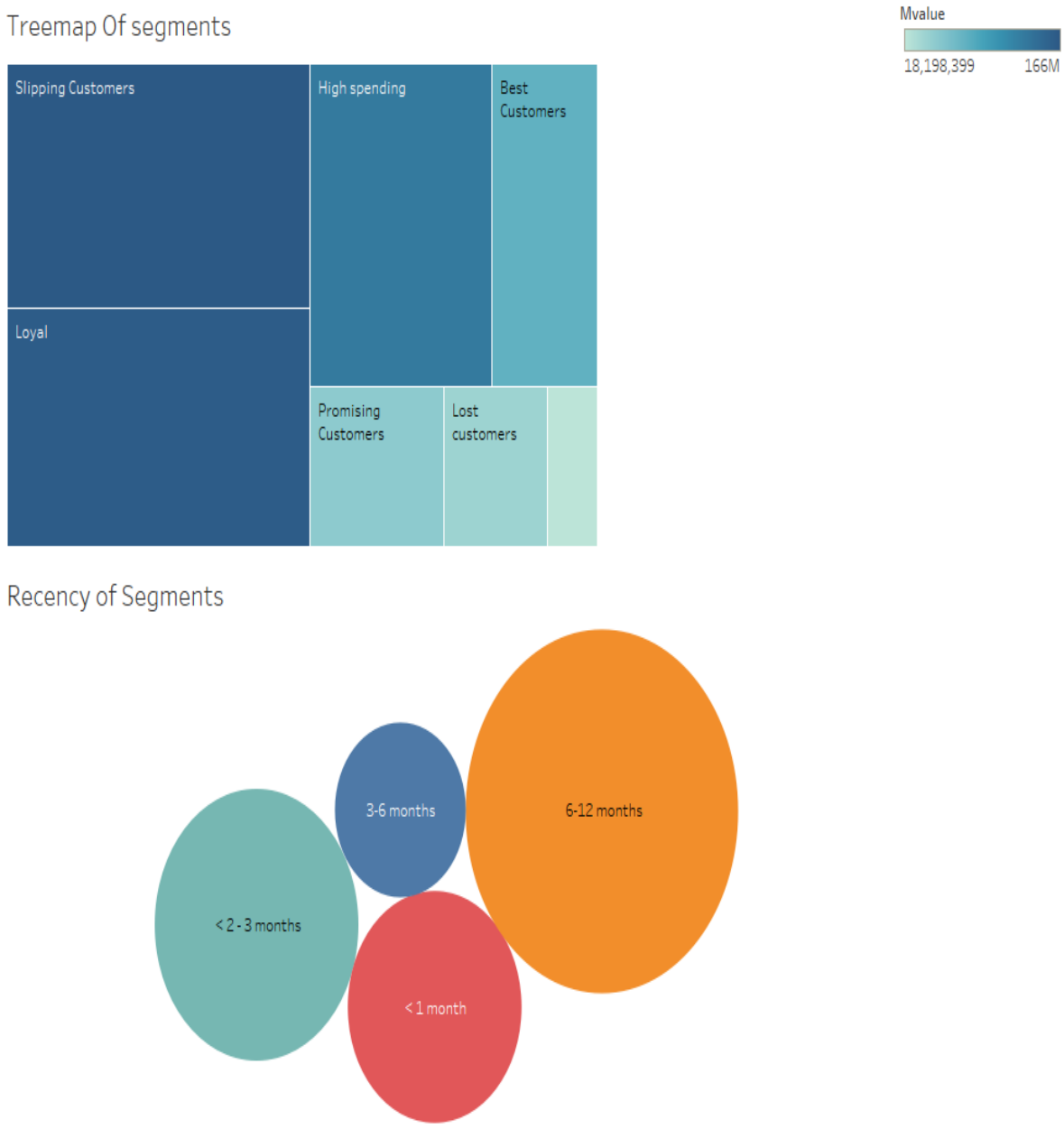
Screen 5.2 shows the segments of customers based on the monetary value. The customers with high monetary values are the most valuable customers and are given appropriate benefits. Promotion and engagement decisions can be taken based on the visualizations after taking other metrics into consideration.



Screen 5.3 Reports for generated revenue based on segments

Screen 5.3 contains a report that consolidates the count of customers based on two metrics

- **Distinct customers per segment**
- **Total revenue generated by each segment**



Screen 5.4 Report for the segment size and shopping frequency

Screen 5.4 contains a report that consolidates the count of customers based on two metrics

- **Customer aggregation based on M value per segment**
- **Customer aggregation based on R value per segment**

CHAPTER 6
CONCLUSION

To conclude, Exploratory Data Analysis (EDA) is done on the dataset to identify seven customer segments based on their shopping behaviour. The segments of customers can be targeted individually based on their respective needs. This will result in a win-win situation for both the customer as well as the company. The company will gain more revenue and customer satisfaction. The customers will be valued more and get personalized services. Specific marketing strategies can be planned by the marketing executives for each segment. A customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities. When a group of customers is sent personalized messages as part of a marketing mix that is designed around their needs, it is easier for companies to send those customers special offers meant to encourage them to buy more products.

6.1 FUTURE ENHANCEMENTS

Further enhancements can be done to this analysis by automating the data pull process. Manual pulling of data can be avoided in this case thereby decreasing the processing time. It will also result in the latest segmentation of customers at any given point of time. This can help the companies to detect a change in trend and take suitable actions earlier.

BIBLIOGRAPHY

1. https://www.statisticshowto.com/shapiro-wilk-test/
2. https://clevertap.com/blog/rfm-analysis/
3. https://www.optimove.com/resources/learning-center/rfm-segmentation
4. https://medium.com/@ODSC/transforming-skewed-data-for-machine-learning-90e6cc364b0
5. https://opendatascience.com/transforming-skewed-data-for-machine-learning/

APPENDIX

Abbreviations

SQL - Structured Query Language

EDA – Exploratory Data Analysis

RFM – Recency Frequency Monetary

IDE - Integrated Development Environment

JSON- JavaScript Object Notation

AI – Artificial Intelligence

IoT – Internet of Things