CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

SEMINAR-1 REPORT

Submitted by

BHARATHWAJ M - RA2011026020065
HARSHIT V - RA2011026020086
MADHESH B - RA2011026020098

Under the guidance of

Mrs. P. Preethy Jemima, M.E.,

(Assistant Professor, Department of Computer Science and Engineering) ***In***

partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING WITH SPECIALISATION IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM CAMPUS, CHENNAI-600089

NOV 2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that the Seminar-I report titled **"CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING"** is the bonafide work of **BHARATHWAJ M [RA2011026020065], HARSHIT V [RA2011026020086], MADHESH B [RA2011026020098]** submitted for the course 18CSP103L Seminar – I. This report is a record of successful completion of the specified course evaluated based on literature reviews and the supervisor. No part of the Seminar Report has been submitted for any degree, diploma, title, or recognition before.

SIGNATURE

Mrs. P. Preethy Jemima, M.E.,
Assistant Professor
Computer Science & Engineering
SRM Institute of Science and Technology
Ramapuram, Chennai.

SIGNATURE

Dr. K. RAJA, M.E., Ph.D.,
Professor and Head
Computer Science & Engineering
SRM Institute of Science and Technology
Ramapuram, Chennai.

Submitted for the Seminar-1 Viva Voce Examination held on ..... at SRM Institute of Science and Technology, Ramapuram Campus, Chennai-600089.

EXAMINER 1

EXAMINER 2

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, RAMAPURAM, CHENNAI – 89

DECLARATION

We hereby declare that the entire work contained in this project report titled **CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING** has been carried out by **BHARATHWAJ M [RA2011026020065], HARSHIT V [RA2011026020086], MADHESH B [RA2011026020098]** at SRM Institute of Science and Technology, Ramapuram Campus, Chennai- 600089, under the guidance of Mrs. P. Preethy Jemima, Assistant Professor, Department of Computer Science and Engineering.

Place: Chennai Date:

BHARATHWAJ M	HARSHIT V	MADHESH B
RA2011026020065	RA2011026020086	RA2011026020098

ABSTRACT

The objective of this study is to apply business intelligence in identifying potential customers by providing relevant and timely data to business entities in the Retail Industry. The data furnished is based on systematic study and scientific applications in analyzing sales history and purchasing behavior of the consumers. The curated and organized data as an outcome of this scientific study not only enhances business sales and profit, but also equips with intelligent insights in predicting consumer purchasing behavior and related patterns.

In order to execute and apply the scientific approach using K-Means algorithm, the real time transactional and retail dataset are analyzed. Spread over a specific duration of business transactions, the dataset values and parameters provide an organized understanding of the customer buying patterns and behavior across various regions.

This study is based on the RFM (Recency, Frequency and Monetary) model and deploys dataset segmentation principles using K-Means Algorithm. A variety of dataset clusters are validated based on the calculation of Silhouette Coefficient. The results thus obtained with regard to sales transactions are compared with various parameters like Sales Recency, Sales Frequency and Sales Volume.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

In the modern era of innovation, where there is a large competition to be better than everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are a 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. This is where machine learning comes into picture, the various algorithms are applied to identify the hidden patterns in the data for better decision making.

The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. 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.

1.1.1 Problem Statement

In the ever-growing competition and increasing complexity of the business environment, segmentation and its systematic study improves customer loyalty and enhances enterprise-level for long lasting relationships by widening profitable customer databases. The major industries wherein customer segmentation and for data mining can be applied are the Retail Industry, because it requires a vast amount of data on sales, transportation, consumption ratio, redelivery service and many others.

Also, Retail data mining helps in identifying and effectively mapping customer behavior and related patterns during the entire life-cycle of business transactions. This ultimately, leads to improved customer service, effective sales and distribution strategies and many more. This work mainly focuses on tracking the historical purchasing behavior of customers with the aim to find the maximum amount of sale possible in the specific area. Based on the statistical results and indicators, companies in the retail industry can design various sales and marketing strategies like promotional campaigns, extending seasonal discounts or floating sales enabling coupons to increase the sales and improve customer retention.

1.2 Aim of the Project

The objective of this study is to apply business intelligence in identifying potential customers by providing relevant and timely data to business entities in the Retail Industry. Also to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

1.3 Project Domain

The domain of the project is Machine Learning. The progress of machine learning techniques has been challenging when it comes to Data Clustering and RFM Analysis. Machine learning uses K-Means Clustering, which is one of the Unsupervised machine learning algorithms.

1.4 Scope of the Project

In general, the methods used to gather the data for this project can easily be extended into other relevant contexts/analyses. While there is clear value in using the same data to investigate purchasing patterns or to build an item-based collaborative filtering recommender system, neither of these is the focus for this paper. The scope of the paper is limited to the following four intertwined Goals:

- To cluster customers based on common purchasing behaviors for future operations/marketing projects
- To incorporate best mathematical, visual, programming, and business practices into a thoughtful analysis that is understood across a variety of contexts and disciplines
- To investigate how similar data and algorithms could be used in future data mining projects.
- To create an understanding and inspiration of how data science can be used to solve real-world problems. Before delving into the details of the project and its implications, the next chapter discusses what customer segmentation analysis actually is and the reasons for its importance.

1.5 Methodology

This Methodology is based on RFM analysis. When we are provided with raw data extracted from a database, it might be messy and non-informative to look at individual records.

- RFM analysis is applied to present data at aggregate level and is used to segment customers into homogenous groups.
- These three values are important as **F** and **M** indicate value of customers, and **R** indicate customers' engagement and satisfaction.

- The values are easy to obtain from the basic set of information for each purchasing history.
- FM technique is a **cost-efficient marketing strategy** based on customer behavior segmentation.
- With this system the accuracy can be increased by about **20.4%** than the existing system.

Recency	**Frequency**	**Monetary**
When is your last purchase?	How many times have you placed or purchased?	How much have you spent?
Example: Length of duration since last purchase.	**Example:** Number of orders over selected analysis period.	**Example:** Sum of total amount spent over the period.

CHAPTER 2

LITERATURE REVIEW

Paper - I

Title: A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization

Authors: Bhade Kalyani, Vedanti Gulalkari, Nidhi Harwani and Sudhir N Dhage

Methodology:

- K-Means clustering was used for customer segmentation and Singular Value Decomposition was used for providing appropriate recommendations to the customers.

Technical Gap: Drawbacks of the recommender system like sparsity, cold start problem etc and how they can be overcome.

Description: Proposed a systematic approach for targeting customers and providing maximum profit to the organizations. An important initial step was to analyze the data of sales acquired from the purchase history and determine the parameters that have the maximum correlation. Based on respective clusters, proper resources can be assigned towards profitable customers using machine learning algorithms.

Paper - II

Title: Customer Segmentation using K-means Clustering

Authors: Kansal, Tushar, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury

Methodology:

- Performed customer segmentation using K-means clustering.

Technical Gap: By applying clustering, 5 segments of cluster were formed labeled as Careless, Careful, Standard, Target and Sensible customers. However, the authors got two new clusters on applying mean shift clustering labeled as High buyers and frequent visitors and High buyers and occasional visitors.

Description: A python program was developed and the program was trained by applying standard scaler onto a dataset having two features of 200 training samples taken from local retail shops. Both the features are the average of the amount of shopping by customers and average of the customer's visit to the shop annually.

Paper - III

Title: Analysis of Customer Segmentation Based on Broad Learning System

Authors: Wang, Zhenyu, Yi Zuo, Tieshan Li, CL Philip Chen, and Katsutoshi Yada

Methodology:

- Analyzed customer segmentation based on a broad learning system which provides an alternative view of learning in deep structure.

Technical Gap: The customer behavior data used in this paper was collected from a real-world supermarket in Japan. Customer segmentation was considered as a multi-label classification problem based on both POS data and RFID data.

Description: Firstly, in addition to customer purchasing behavior, RFID (Radio Frequency Identification) data was also included, which can accurately represent the consumers' in-store behavior. Secondly, this paper used the Broad Learning System (BLS) to analyze consumer segmentation. BLS is one of the finest machine learning techniques, and quite efficient and effective for classification tasks.

Paper - IV

Title: Research on customer segmentation in retailing based on clustering model

Authors: Li, Zeying

Methodology:

- Proposed a method in which a retail supermarket was taken as a research object, and data mining methods were used to retail enterprise customer segments, and then association rules obtained using Apriori algorithm.

Technical Gap:

First, it has been shown that the worst case running time of the algorithm is super-polynomial in the input size. Second, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering.

Description: Apriori algorithms were used to different groups of customers and get rules about customer characteristics to make customer characteristic analysis efficiently. Finally, the author gave some references to the supermarket's marketing and management work, which helped in understanding it in detail. Data mining was used efficiently to deal with the large amount of historical and current data, from the database to find some potential, useful and valuable information for the retail stores which help us target customers.

Paper - V

Title: K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data

Authors: Kayalvily Tabianan, Shubashini Velu and Vinayakumar Ravi

Methodology:

- Data Mining and K Means Clustering Elbow method used

Technical Gap: There are some limitations such as working with microdata both making comparisons and modeling and clustering, especially if they are business management and sales financial data. The difficulties of working with indicators, indexes, and rates complicate the data mining process and, later, the reading of the results.

Description: This research aims to help researchers and other E commerce stakeholder for a comparison of the structuring and grouping of the E-commerce purchasing pattern in small areas. It also endeavors to show an optimal way of transforming and working on datasets to facilitate the resulting groupings. Therefore, this research allows us to segment a cohort from E-commerce behavior data from multiple category store and purchasing histories. In Addition, it would capture the inequality that can be observed between the high profitable segments and low profitable segments in the category of products better.

Paper - VI

Title: Customer Segmentation using K-means Clustering

Authors: Hemashree Kilari, Sailesh Edara, Guna Ratna Sai Yarra, Dileep Varma Gadhiraju

Methodology:

- K Means Clustering Silhouette method without RFM Analysis.

Technical Gap: You can choose the number of clusters by visually inspecting your data points, but you will soon realize that there is a lot of ambiguity in this process for all except the simplest data sets.This is not always bad, because you are doing unsupervised learning and there's some inherent subjectivity in the labeling process. Here, having previous experience with that particular problem or something similar will help you choose the right value.

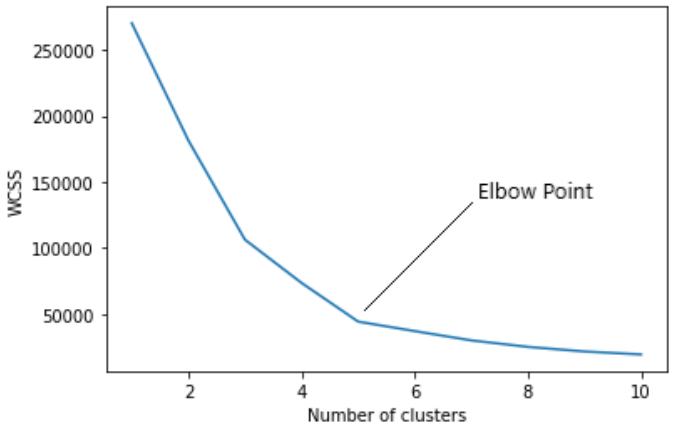
Description: This study demonstrates that client segmentation in shopping malls is achievable despite the fact that this form of machine learning application is highly useful in the market. A manager can concentrate all of his or her attention on each cluster that has been discovered and meet all of their requirements.

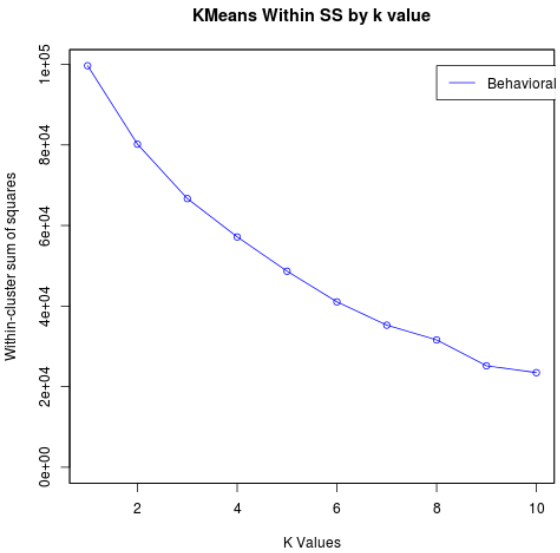
CHAPTER 3

PROJECT DESCRIPTION

3.1 Existing System

In the current system, the K Means Initialization method and Elbow method are used to find the optimal no of clusters (k).

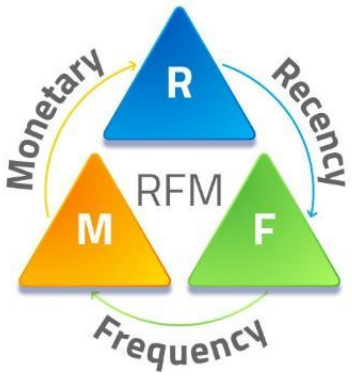


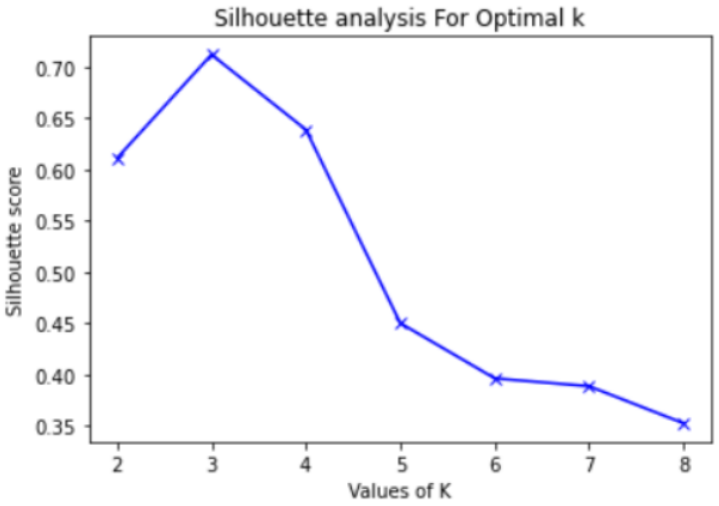


- The elbow method doesn't always work well, especially if the data is not very clustered.
- Because of that it will produce a smooth curve which makes it difficult to find the optimal no of clusters (k).
- Also, RFM technique (which is specifically used in customer segmentation) is not implemented in the current system.

3.2 Proposed System

In our system we are using **K Means silhouette method** to find the optimal clusters (k) and **RFM technique** to improve the accuracy of the clustering based on the customer history, and then it's visualized.





- RFM technique is a cost-efficient marketing strategy based on customer behavior segmentation.
- RFM stands for **Recency, Frequency, Monetary**. Simply, it groups customers based on their purchase history.
    - Recency - How recent was the customer's last purchase?
    - Frequency - How often did this customer make a purchase in a given period?
    - Monetary - How much money did the customer spend in a given period?
- With this system the accuracy can be increased by about **20.4%** than the existing system.

3.3 Feasibility Study

A Feasibility study is carried out to check the viability of the project and to analyze the strengths and weaknesses of the proposed system. The application of usage of masks in crowd areas must be evaluated. The feasibility study is carried out in three forms:

- Economic Feasibility
- Technical Feasibility
- Social Feasibility

3.3.1 Economic Feasibility

The proposed system does not require any high-cost equipment. This project can be developed within the available software.

3.3.2 Technical Feasibility

The proposed system is completely a Machine learning model. The main tools used in this project are Anaconda prompt, Visual studio, Kaggle datasets, Jupyter Notebook And the language used to execute the process in Python. The above-mentioned tools are available for free and technical skills required to use these tools are practicable. From this we can conclude that the project is technically feasible.

3.3.3 Social Feasibility

Social feasibility is a determination of whether a project will be acceptable or not. Our project is Eco-friendly for society and there are no social issues. our project must not be threatened by the system instead must accept it as a necessity. since our project is applicable for every individual in the society to take care about the society and environment. The level of acceptance of the System is very high and it depends on the methods deployed in the system. our system is highly familiar with society.

3.4 System Specification

3.4.1 Hardware Specification

PROCESSOR	Intel i5-8250 @ 3.40GHz
STORAGE	512 GB SSD
RAM	16 GB
GPU	Nvidia GTX 1650 Ti
OPERATING SYSTEM	Windows 11 x64 Bit

3.4.2 Software Specification

Windows 10, 11
Anaconda Jupyter Notebook
Python 3.10
Machine Learning Modules

CHAPTER 4

MODULE DESCRIPTION

4.1 General Architecture

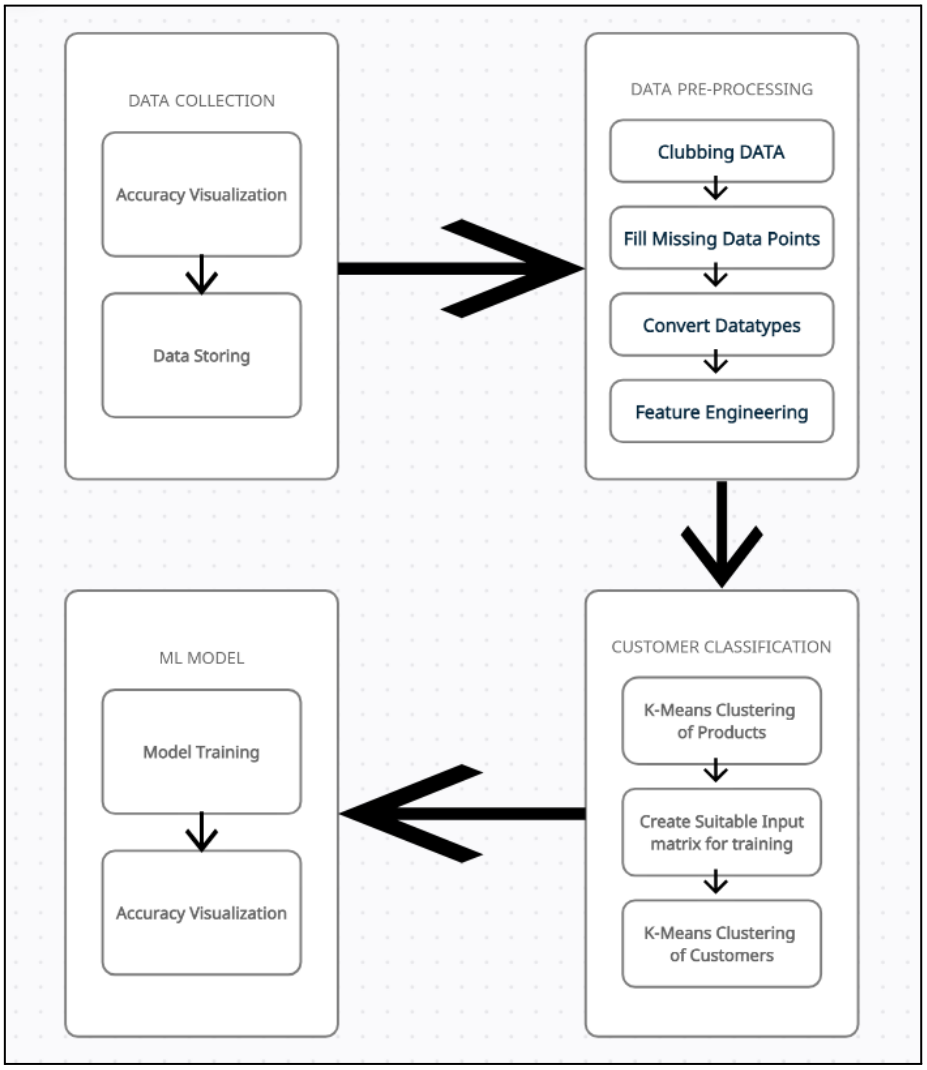


Figure 4.1: **Architecture Diagram**

The data set obtained is passed to the pre-processing stage where duplicates and null rows are dropped. A dataset may contain cancelled orders which are removed as well. The datatype of columns is converted to the required datatype for pre-processing. The refined data set is then used to classify the customers into different categories by un-supervised learning through K-Mean algorithm. The resultant dataset or matrix which contains customer and its categories is used to train and test supervised training models such as Random Forest, K-Nearest Neighbors, and Gradient Boost.

4.2 Design Phase

4.2.1 Data Flow Diagram

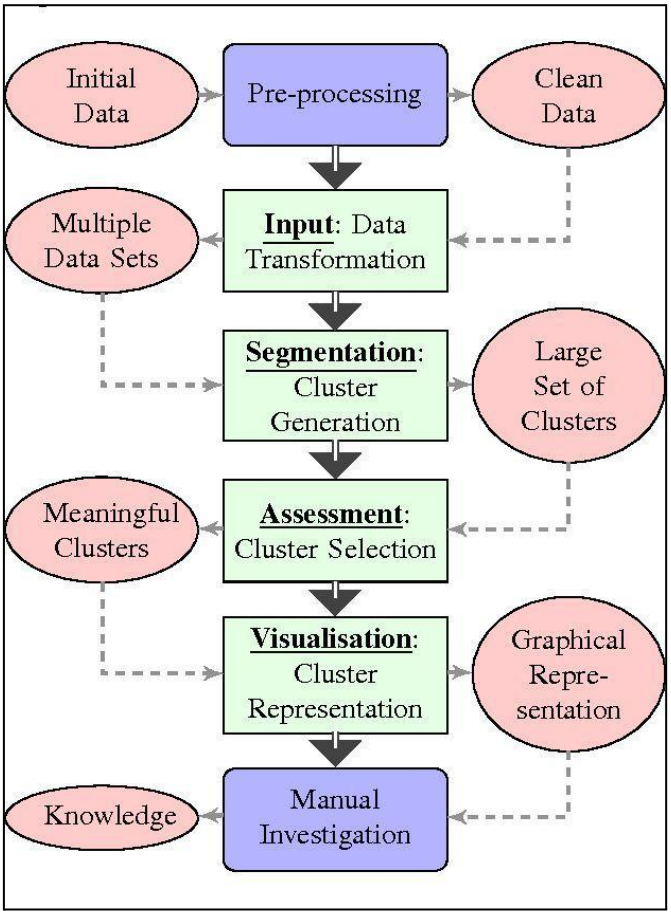


Figure 4.2: **Data Flow Diagram**

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. It shows how data enters and leaves the system, what changes the information, and where data is stored. It can be manual, automated, or a combination of both. At First, the data is preprocessed using pandas py module.

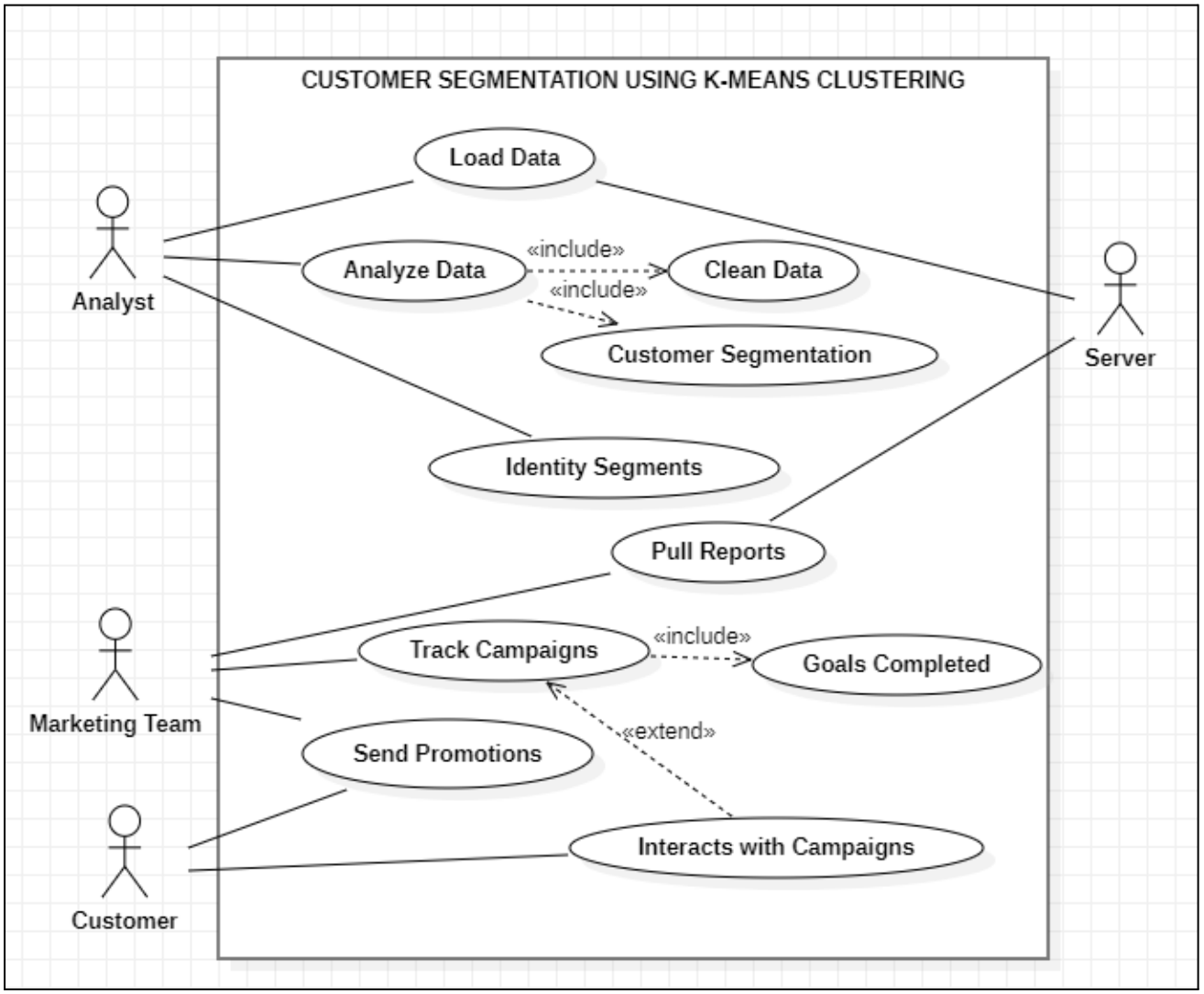
Secondly, the required data is selected from the Pre-processed dataset, and splitted in 80:20 ratio for the train set and test set. Then it is clustered using K-Means clustering algorithm, which is one of the unsupervised machine learning models. Now the result is visualized for manual analysis.

4.2.2 UML Diagram

A UML diagram is a diagram based on the UML (Unified Modeling Language) with the purpose of visually representing a system along with its main actors, roles, actions, artifacts, or classes, to better understand, or document information about the system. For our project we have drawn 2 Diagrams:

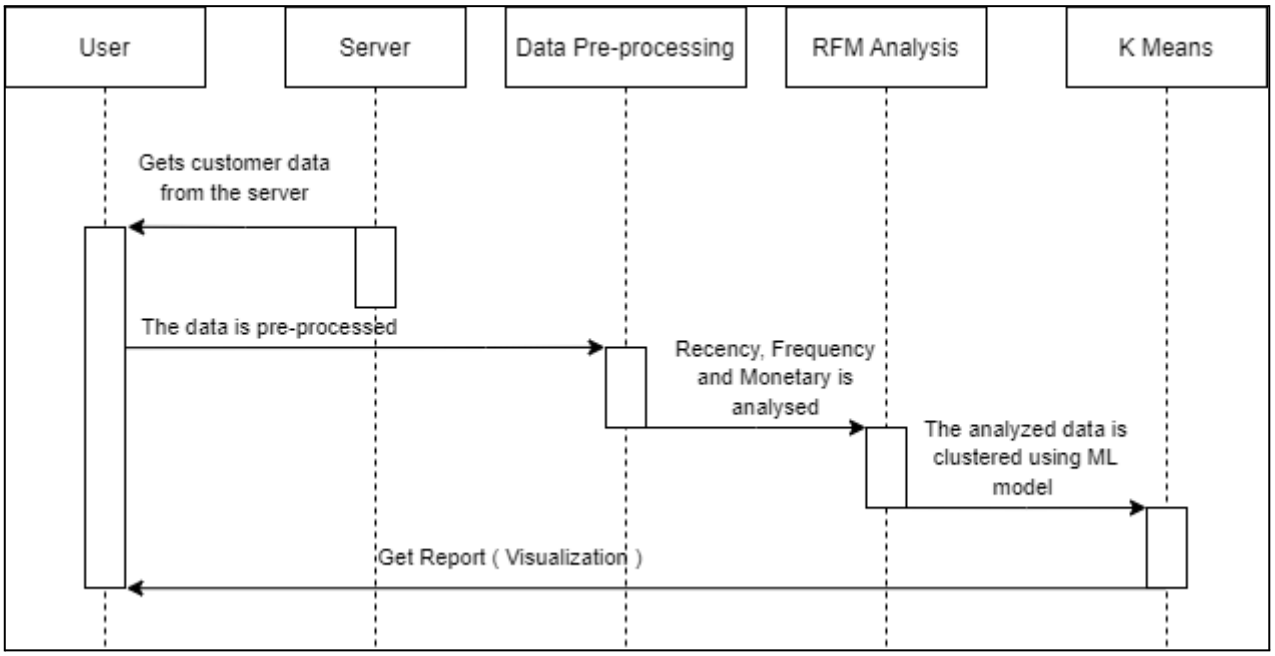
1. Use Case Diagram

A use case diagram is a graphical depiction of a user's possible interactions with a system. A use case diagram shows various use cases and different types of users the system has and will often be accompanied by other types of diagrams as well.



2. Sequence Diagram

Sequence Diagrams are interaction diagrams that detail how operations are carried out. They capture the interaction between objects in the context of a collaboration. Sequence Diagrams are time focused, and they show the order of the interaction visually by using the vertical axis of the diagram to represent time, what messages are sent and when.



4.3 MODULE DESCRIPTION

Our entire project is divided into five modules:

Step 1: Data Collecting

Step 2: Pre-processing of Data

Step 3: RFM Analysis

Step 4: Clustering the analyzed data

Step 5: Visualization

4.3.1 Data Collecting

There are many different ways your online store collects data from consumers. It is vital for eCommerce companies to collect this data responsibly.

There are three basic approaches on collecting data about online store customers:

1. Direct user inquiry (in the case of personal data, informing the user and their consent is mandatory)
2. Indirectly tracking users' movements through the store's tabs
3. By adding customer data from other sources

Depending on the chosen approach, there are several ways and tools to obtain information about individual clients:

- Cookies
- Customer surveys and feedback forms
- Placing orders
- Social media
- Chatbots

4.3.2 Pre-Processing Of Data

Data preparation and filtering steps can take a considerable amount of processing time. Examples of data preprocessing include cleaning, instance selection, normalization, one hot encoding, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set.

4.3.3 RFM Analysis

RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency effects retention, a measure of engagement. Businesses that lack the monetary aspect, like viewership, readership, or surfing-oriented products, could use Engagement parameters instead of Monetary factors. This results in using RFE – a variation of RFM. Furthermore, this Engagement parameter could be defined as a composite value based on metrics such as bounce rate, visit duration, number of pages visited, time spent per page, etc.

4.3.4 Clustering The Analyzed Data

In the context of customer segmentation, customer clustering analysis is the use of a mathematical model to discover groups of similar customers based on finding the smallest variations among customers within each group. The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization.

A common cluster analysis method is a mathematical algorithm known as k-means cluster analysis, sometimes referred to as scientific segmentation. The clusters that result assist in better customer modeling and predictive analytics, and are also used to target customers with offers and incentives personalized to their wants, needs and preferences.

4.3.5 Visualization

Now the clustering is visualized in a scatter plot using matplotlib. The marketing team will now analyze and makes changes to its recommendation system.

4.3 Datasets Sample

DATASET SAMPLE 1

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/2010 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	1/12/2010 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1/12/2010 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/2010 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/2010 8:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	1/12/2010 8:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	1/12/2010 8:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	1/12/2010 8:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	1/12/2010 8:28	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	1/12/2010 8:34	1.69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	1/12/2010 8:34	2.1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	1/12/2010 8:34	2.1	13047	United Kingdom
536367	21754	HOME BUILDING BLOCK WORD	3	1/12/2010 8:34	5.95	13047	United Kingdom
536367	21755	LOVE BUILDING BLOCK WORD	3	1/12/2010 8:34	5.95	13047	United Kingdom
536367	21777	RECIPE BOX WITH METAL HEART	4	1/12/2010 8:34	7.95	13047	United Kingdom
536367	48187	DOORMAT NEW ENGLAND	4	1/12/2010 8:34	7.95	13047	
536368	22960	JAM MAKING SET WITH JARS	6	1/12/2010 8:34	4.25	13047	United Kingdom
536368	22913	RED COAT RACK PARIS FASHION	3	1/12/2010 8:34	4.95	13047	United Kingdom
536368	22912	YELLOW COAT RACK PARIS FASHION	3	1/12/2010 8:34	4.95	13047	United Kingdom
536368	22914	BLUE COAT RACK PARIS FASHION	3	1/12/2010 8:34	4.95	13047	United Kingdom
536369	21756	BATH BUILDING BLOCK WORD	3	1/12/2010 8:35	5.95	13047	United Kingdom
536370	22728	ALARM CLOCK BAKELIKE PINK	24	1/12/2010 8:45	3.75	12583	France
536370	22727	ALARM CLOCK BAKELIKE RED	24	1/12/2010 8:45	3.75	12583	France
536370	22726	ALARM CLOCK BAKELIKE GREEN	12	1/12/2010 8:45	3.75	12583	France
536370	21724	PANDA AND BUNNIES STICKER SHEET	12	1/12/2010 8:45	0.85	12583	France
536370	21883	STARS GIFT TAPE	24	1/12/2010 8:45	0.65	12583	France

DATASET SAMPLE 2

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
552956	21481	FAWN BLUE HOT WATER BOTTLE	6	12/5/2011 12:34	2.95	12431	Australia
552956	48138	DOORMAT UNION FLAG	4	12/5/2011 12:34	7.95	12431	Australia
552956	23284	DOORMAT KEEP CALM AND COME IN	4	12/5/2011 12:34	7.95	12431	Australia
552956	23040	PAPER LANTERN 9 POINT SNOW STAR	3	12/5/2011 12:34	5.75	12431	Australia
552956	23108	SET OF 10 LED DOLLY LIGHTS	2	12/5/2011 12:34	6.25	12431	Australia
552956	23084	RABBIT NIGHT LIGHT	12	12/5/2011 12:34	2.08	12431	Australia
552957	84991	60 TEATIME FAIRY CAKE CASES	240	12/5/2011 12:36	0.55	17404	Sweden
552957	84987	SET OF 36 TEATIME PAPER DOILIES	144	12/5/2011 12:36	1.45	17404	Sweden
552957	21889	WOODEN BOX OF DOMINOES	288	12/5/2011 12:36	1.25	17404	Sweden
552957	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	576	12/5/2011 12:36	0.29	17404	Sweden
552957	22752	SET 7 BABUSHKA NESTING BOXES	2	12/5/2011 12:36	8.5	17404	Sweden
552957	22961	JAM MAKING SET PRINTED	96	12/5/2011 12:36	1.45	17404	Sweden
552957	22721	SET OF 3 CAKE TINS SKETCHBOOK	24	12/5/2011 12:36	4.25	17404	Sweden
552957	23159	SET OF 5 PANCAKE DAY MAGNETS	120	12/5/2011 12:36	2.08	17404	Sweden
552957	23154	SET OF 4 JAM JAR MAGNETS	120	12/5/2011 12:36	2.08	17404	Sweden
552957	22491	PACK OF 12 COLOURED PENCILS	12	12/5/2011 12:36	0.85	17404	Sweden
552957	23191	BUNDLE OF 3 RETRO NOTE BOOKS	12	12/5/2011 12:36	1.65	17404	Sweden
552957	22489	PACK OF 12 TRADITIONAL CRAYONS	432	12/5/2011 12:36	0.42	17404	Sweden
552957	22492	MINI PAINT SET VINTAGE	576	12/5/2011 12:36	0.65	17404	Sweden
552957	22938	CUPCAKE LACE PAPER SET 6	144	12/5/2011 12:36	1.95	17404	Sweden
552957	23080	RED METAL BOX TOP SECRET	4	12/5/2011 12:36	8.25	17404	Sweden
552957	23081	GREEN METAL BOX ARMY SUPPLIES	4	12/5/2011 12:36	8.25	17404	Sweden
552958	23282	FOLDING BUTTERFLY MIRROR IVORY	12	12/5/2011 12:49	0.83	15498	United Kingdom
552958	23284	DOORMAT KEEP CALM AND COME IN	10	12/5/2011 12:49	6.75	15498	United Kingdom
552958	48194	DOORMAT HEARTS	10	12/5/2011 12:49	6.75	15498	United Kingdom
552958	21756	BATH BUILDING BLOCK WORD	6	12/5/2011 12:49	5.95	15498	United Kingdom

Chapter 5

IMPLEMENTATION AND TESTING

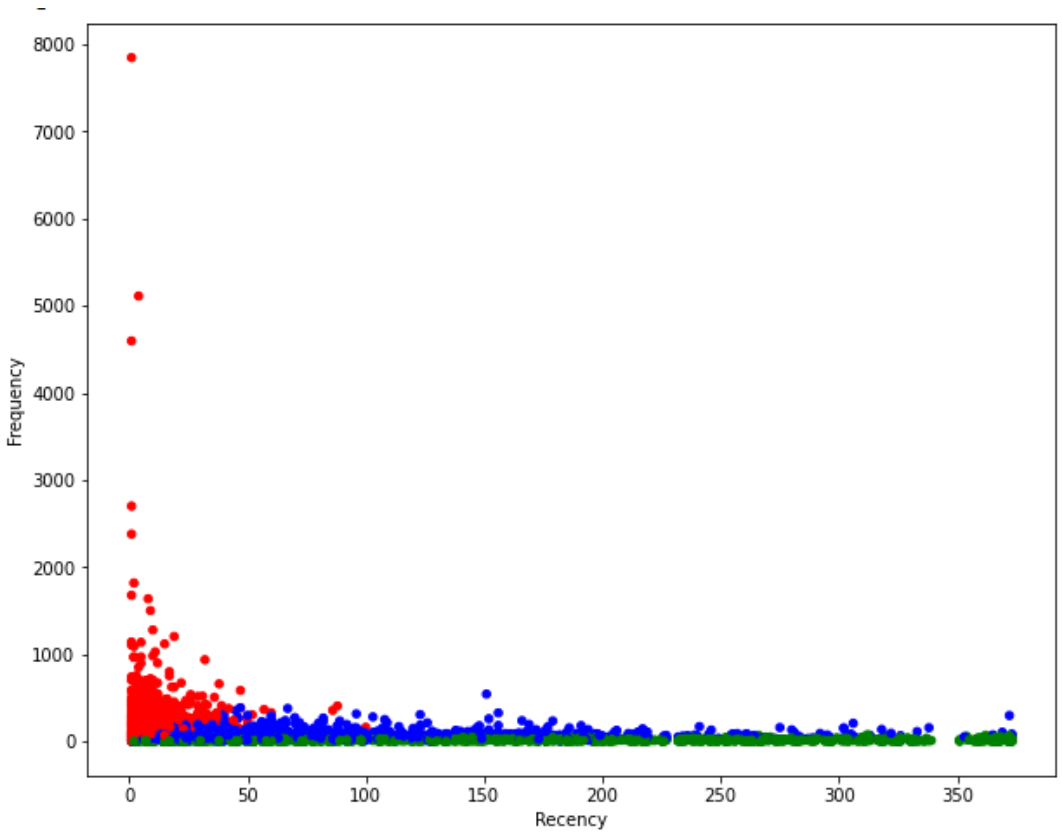
5.1 Input and Output

INPUT:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680.0	France

541909 rows × 8 columns

OUTPUT:



Result

```
[54] RFMScores.head()
```

CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level	Cluster	Color
12346.0	325	1	77183.60	4	4	1	441	9	Silver	2	blue
12747.0	2	103	4196.01	1	1	1	111	3	Platinum	0	red
12748.0	1	4596	33719.73	1	1	1	111	3	Platinum	0	red
12749.0	3	199	4090.88	1	1	1	111	3	Platinum	0	red
12820.0	3	59	942.34	1	2	2	122	5	Platinum	0	red

5.2 Testing

Testing is the process of evaluating a system or its component(s) with the intent to find whether it satisfies the specified requirements or not.

5.2.1 Types of Testing

1. Unit Testing

Unit testing is a beneficial software testing method where the units of source code are tested to check the efficiency and correctness of the program.

2. Integration Testing

In this testing, units or individual components of the software are tested in a group. The focus of the integration testing level is to expose defects at the time of interaction between integrated components or units.

3. Functional Testing

Functional testing is also called as black-box testing, because it focuses on application specification rather than actual code. Testers must test only the program rather than the system. Goal of functional testing.

5.3 Test Result

Result

```
[54] RFMScores.head()
```

CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level	Cluster	Color
12346.0	325	1	77183.60	4	4	1	441	9	Silver	2	blue
12747.0	2	103	4196.01	1	1	1	111	3	Platinum	0	red
12748.0	1	4596	33719.73	1	1	1	111	3	Platinum	0	red
12749.0	3	199	4090.88	1	1	1	111	3	Platinum	0	red
12820.0	3	59	942.34	1	2	2	122	5	Platinum	0	red

CHAPTER 6

RESULTS AND DISCUSSIONS

6.1 Efficiency of Proposed System

The proposed work aims to classify online e-commerce customers into various categories based on their characteristics like spending amount, type of product they buy, and how frequently purchase happens etc. Initially the products were classified into 5 categories using K-means and through this result along with other characteristics like amount spent, frequency etc. online ecommerce customers were classified into 11 categories using unsupervised method i.e. K-means Clustering. To measure the accuracy, silhouette scoring was used. Later different classifiers were trained using the data obtained so far. Our proposed model is 20% more efficient than the existing model.

In order to execute and apply the scientific approach using K-Means algorithm, the real time transactional and retail dataset are analyzed. Spread over a specific duration of business transactions, the dataset values and parameters provide an organized understanding of the customer buying patterns and behavior across various regions.

This study is based on the RFM (Recency, Frequency and Monetary) model and deploys dataset segmentation principles using K-Means Algorithm. A variety of dataset clusters are validated based on the calculation of Silhouette Coefficient. The results thus obtained with regard to sales transactions are compared with various parameters like Sales Recency, Sales Frequency and Sales Volume.

CHAPTER 7

SOURCE CODE & POSTER PRESENTATION

7.1 SOURCE CODE

▾ Import necessary libraries

```
[1] %matplotlib inline
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
```

▾ Import Online Retail Data containing transactions

```
Rtl_data = pd.read_excel('/content/Online Retail.xlsx')
Rtl_data.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Data Pre-processing

```
[ ]
#Keep only United Kingdom data
Rtl_data = Rtl_data.query("Country=='United Kingdom'").reset_index(drop=True)

#Check for missing values in the dataset
Rtl_data.isnull().sum(axis=0)
```

```
[ ]
'''Remove missing values from CustomerID column, can ignore
                    missing values in description column'''
Rtl_data = Rtl_data[pd.notnull(Rtl_data['CustomerID'])]

#Validate if there are any negative values in Quantity column
Rtl_data.Quantity.min()
```

```
-80995
```

```
[ ]
#Validate if there are any negative values in UnitPrice column
Rtl_data.UnitPrice.min()

#Filter out records with negative values
Rtl_data = Rtl_data[(Rtl_data['Quantity']>0)]
```

```
[ ]
#Convert the string date field to datetime
Rtl_data['InvoiceDate'] = pd.to_datetime(Rtl_data['InvoiceDate'])
```

```
[ ]
#Add new column depicting total amount
Rtl_data['TotalAmount'] = Rtl_data['Quantity'] * Rtl_data['UnitPrice']
```

```
[ ]
'''Check the shape (number of columns and rows) in the dataset
                        after data is cleaned'''
Rtl_data.shape
```

```
(354345, 9)
```

RFM Modelling

```
[18] #Recency = Latest Date - Last Inovice Data, Frequency = count of invoice no. of transaction(s),
     #Monetary = Sum of Total
     #Amount for each customer
     import datetime as dt

     #Set Latest date 2011-12-10 as last invoice date was 2011-12-09.
     #This is to calculate the number of days from recent purchase
     Latest_Date = dt.datetime(2011,12,10)

     #Create RFM Modelling scores for each customer
     RFMScores = Rtl_data.groupby('CustomerID').agg({'InvoiceDate': lambda x: (Latest_Date - x.max()).days,
                              'InvoiceNo': lambda x: len(x), 'TotalAmount': lambda x: x.sum()})

     #Convert Invoice Date into type int
     RFMScores['InvoiceDate'] = RFMScores['InvoiceDate'].astype(int)

     #Rename column names to Recency, Frequency and Monetary
     RFMScores.rename(columns={'InvoiceDate': 'Recency',
                               'InvoiceNo': 'Frequency',
                               'TotalAmount': 'Monetary'}, inplace=True)

     RFMScores.reset_index().head()
```

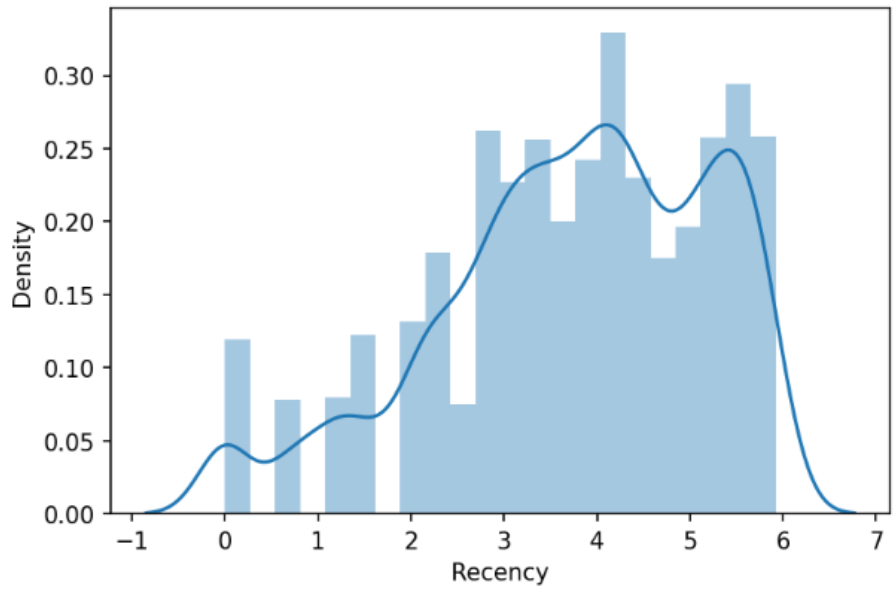
	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12747.0	2	103	4196.01
2	12748.0	0	4596	33719.73
3	12749.0	3	199	4090.88
4	12820.0	3	59	942.34

```
[ ]
#Handle negative and zero values so as 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
RFMScores['Recency'] = [handle_neg_n_zero(x) for x in RFMScores.Recency]
RFMScores['Monetary'] = [handle_neg_n_zero(x) for x in RFMScores.Monetary]

#Perform Log transformation to bring data into normal or near normal distribution
Log_Tfd_Data = RFMScores[['Recency', 'Frequency', 'Monetary']].apply(np.log, axis = 1).round(3)
```

Recency distribution plot (Normalized)

```
[ ]
plt.figure(dpi=100)
Recency_Plot = Log_Tfd_Data['Recency']
ax = sns.distplot(Recency_Plot)
```

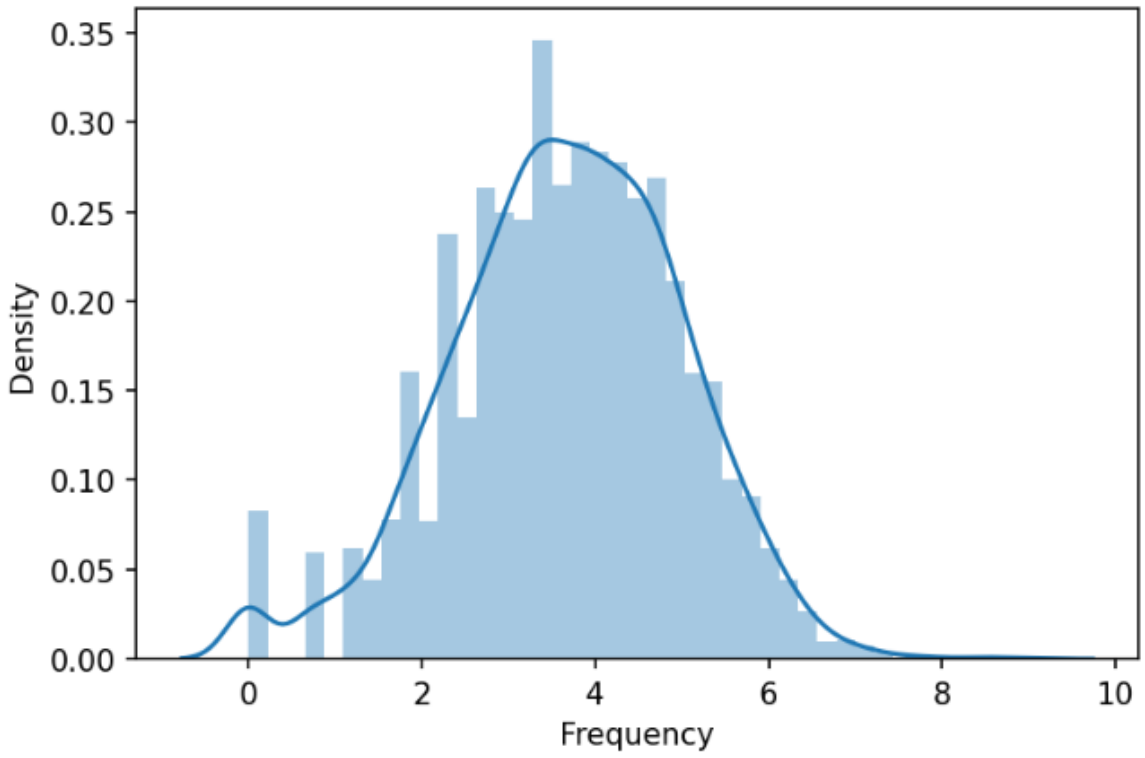
```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level funct
```



Frequency distribution plot (Normalized)

```
[ ]  1  #Data distribution after data normalization for Frequency
     2  plt.figure(dpi=150)
     3  Frequency_Plot = Log_Tfd_Data.query('Frequency < 1000')['Frequency']
     4  ax = sns.distplot(Frequency_Plot)
```

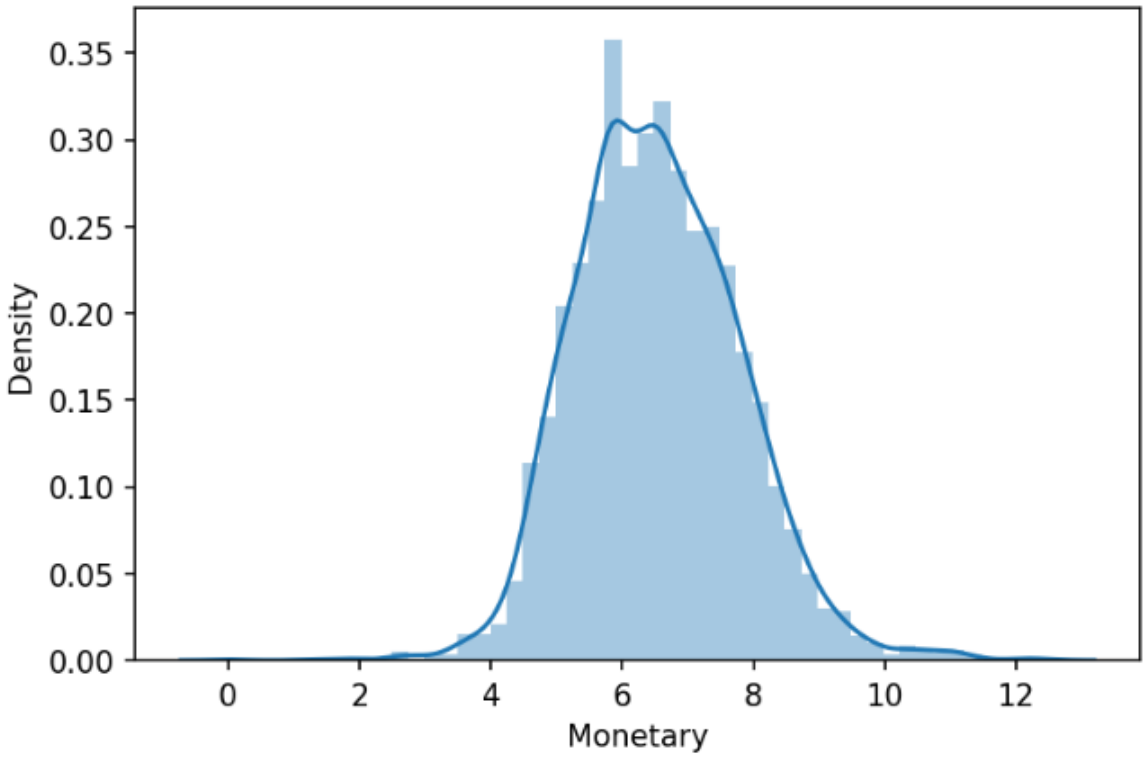
```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use ei
```



Monetary distribution plot (Normalized)

```
[ ]  1  #Data distribution after data normalization for Monetary
     2  plt.figure(dpi=150)
     3  Monetary_Plot = Log_Tfd_Data.query('Monetary < 10000')['Monetary']
     4  ax = sns.distplot(Monetary_Plot)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning:

`distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use ei
```



Split into four segments using quantiles

```
[25] quantiles = RFMScores.quantile(q=[0.25,0.5,0.75])
     quantiles = quantiles.to_dict()
```

```
[26] quantiles
```

```
{'Recency': {0.25: 17.0, 0.5: 50.0, 0.75: 142.0},
 'Frequency': {0.25: 17.0, 0.5: 41.0, 0.75: 99.0},
 'Monetary': {0.25: 300.03999999999996, 0.5: 651.8199999999999, 0.75: 1575.89}}
```

```
[27] #Functions to create R, F and M segments
     def RScoring(x,p,d):
         if x <= d[p][0.25]:
             return 1
         elif x <= d[p][0.50]:
             return 2
         elif x <= d[p][0.75]:
             return 3
         else:
             return 4

     def FnMScoring(x,p,d):
         if x <= d[p][0.25]:
             return 4
         elif x <= d[p][0.50]:
             return 3
         elif x <= d[p][0.75]:
             return 2
         else:
             return 1
```

```
[28] #Calculate Add R, F and M segment value columns in the
     #existing dataset to show R, F and M segment values
     RFMScores['R'] = RFMScores['Recency'].apply(RScoring, args=('Recency',quantiles,))
     RFMScores['F'] = RFMScores['Frequency'].apply(FnMScoring, args=('Frequency',quantiles,))
     RFMScores['M'] = RFMScores['Monetary'].apply(FnMScoring, args=('Monetary',quantiles,))
     RFMScores.head()
```

Finding RFM Scores

```
[30] #Assign Loyalty Level to each customer
     Loyalty_Level = ['Platinum', 'Gold', 'Silver', 'Bronze']
     Score_cuts = pd.qcut(RFMScores.RFMScore, q = 4, labels = Loyalty_Level)
     RFMScores['RFM_Loyalty_Level'] = Score_cuts.values
     RFMScores.reset_index().head()
```

	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver
1	12747.0	2	103	4196.01	1	1	1	111	3	Platinum
2	12748.0	0	4596	33719.73	1	1	1	111	3	Platinum
3	12749.0	3	199	4090.88	1	1	1	111	3	Platinum
4	12820.0	3	59	942.34	1	2	2	122	5	Platinum

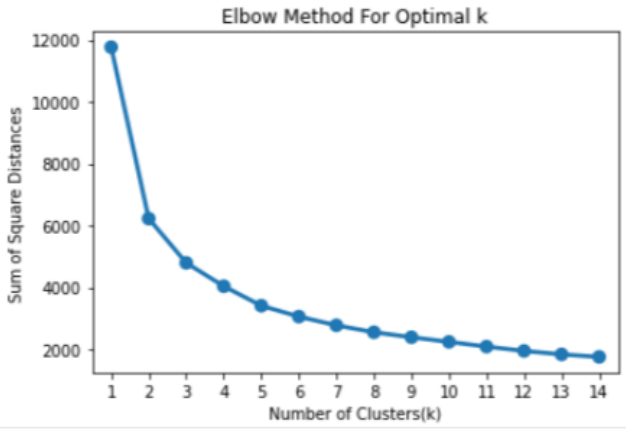
```
[31] #Validate the data for RFMGroup = 111
     RFMScores[RFMScores['RFMGroup']=='111'].sort_values('Monetary', ascending=False).reset_index().head(10)
```

	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level
0	18102.0	0	431	259657.30	1	1	1	111	3	Platinum
1	17450.0	8	337	194550.79	1	1	1	111	3	Platinum
2	17511.0	2	963	91062.38	1	1	1	111	3	Platinum
3	16684.0	4	277	66653.56	1	1	1	111	3	Platinum
4	14096.0	4	5111	65164.79	1	1	1	111	3	Platinum
5	13694.0	3	568	65039.62	1	1	1	111	3	Platinum
6	15311.0	0	2379	60767.90	1	1	1	111	3	Platinum
7	13089.0	2	1818	58825.83	1	1	1	111	3	Platinum
8	15769.0	7	130	56252.72	1	1	1	111	3	Platinum

Finding Optimal Clusters

```
[50] from sklearn.preprocessing import StandardScaler

     #Bring the data on same scale
     scaleobj = StandardScaler()
     Scaled_Data = scaleobj.fit_transform(Log_Tfd_Data)

     #Transform it back to dataframe
     Scaled_Data = pd.DataFrame(Scaled_Data, index = RFMScores.index, columns = Log_Tfd_Data.columns)
```

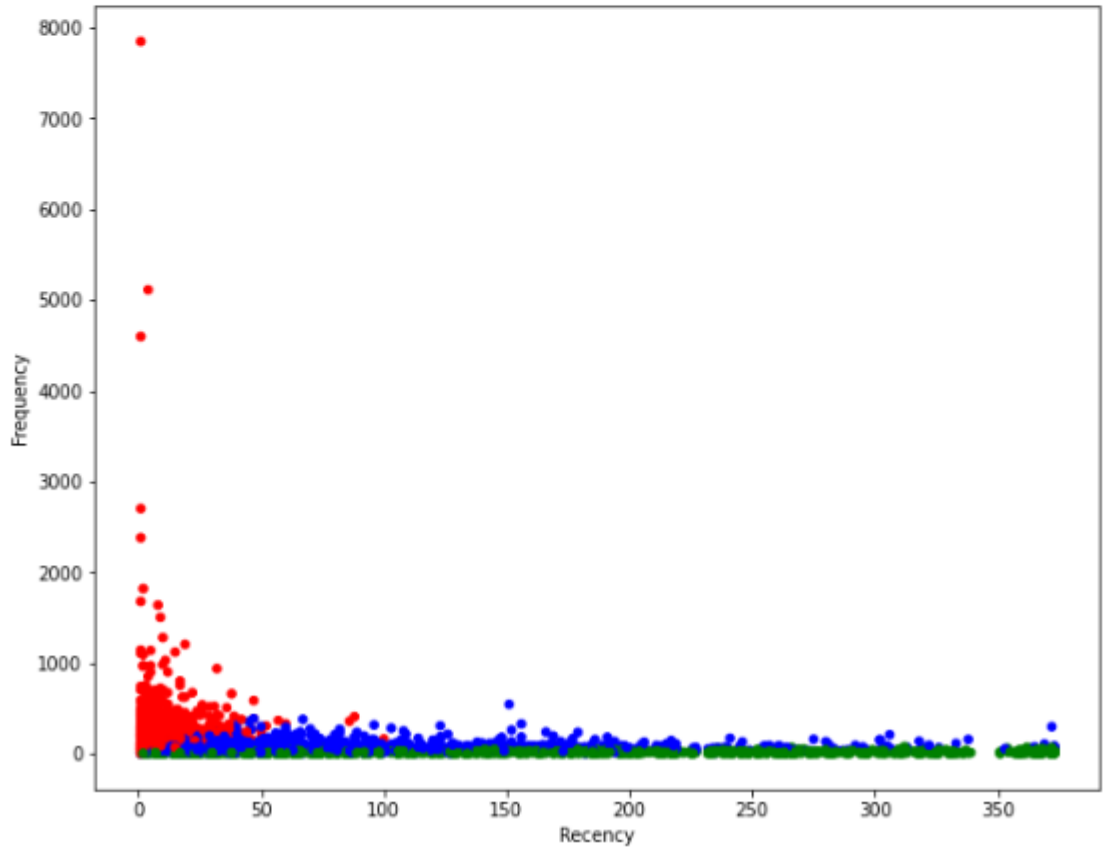
```
[51] from sklearn.cluster import KMeans

     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()
```



Clustering

```
[53] from matplotlib import pyplot as plt
     plt.figure(figsize=(7,7))

     ##Scatter Plot Frequency Vs Recency
     Colors = ["red", "green", "blue"]
     RFMScores['Color'] = RFMScores['Cluster'].map(lambda p: Colors[p])
     ax = RFMScores.plot(
         kind="scatter",
         x="Recency", y="Frequency",
         figsize=(10,8),
         c = RFMScores['Color']
     )
```

```
<Figure size 504x504 with 0 Axes>
```

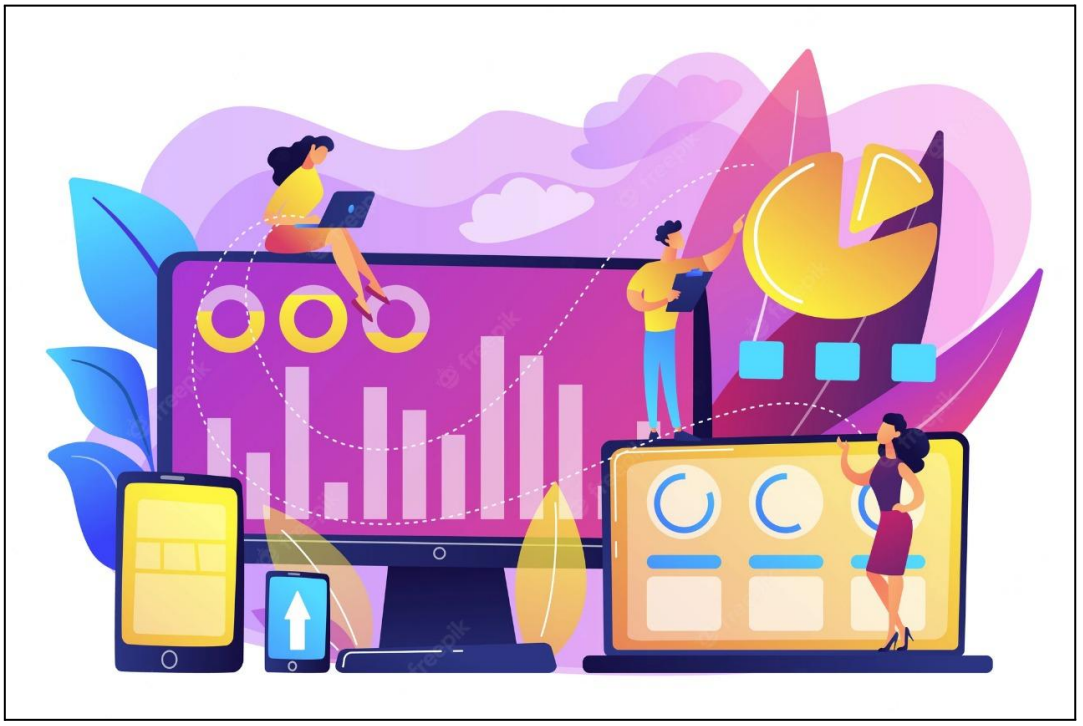


Result

```
[54] RFMScores.head()
```

CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level	Cluster	Color
12346.0	325	1	77183.60	4	4	1	441	9	Silver	2	blue
12747.0	2	103	4196.01	1	1	1	111	3	Platinum	0	red
12748.0	1	4596	33719.73	1	1	1	111	3	Platinum	0	red
12749.0	3	199	4090.88	1	1	1	111	3	Platinum	0	red
12820.0	3	59	942.34	1	2	2	122	5	Platinum	0	red

7.2 POSTER PRESENTATION



CHAPTER 8

CONCLUSION AND FUTURE ENHANCEMENTS

8.1 CONCLUSION

Customer segmentation is a way to improve communication with the customer, to know the wishes of the customer, customer activity so that appropriate communication can be built. Customer Segmentation needed to get potential customers used to increase profits. Potential customer data can be used to provide service the characteristics of customer including ecommerce services as a media buying and selling online. This paper discusses several components to do customer segmentation, which is: Customer segmentation is an activity to divide customers or item into groups that have the same characteristics.

Data that needed for customer segmentation are internal data and external data. The internal data include demographic data and data purchase history, while the external data include cookies and server logs. Internal data can be obtained from a database when customer do registration or transactions and external data can be obtained from web server or other source. Methods of Customer Segmentation can be classified into Simple technique, RFM technique, Target technique, and Unsupervised technique. On Target technique, researcher focus on one variable, it can be product or purchase. Unsupervised technique was used when clustering process researchers have many variables. Process of Customer Segmentation can be simplified into defining business objectives, collecting data, data preparation, and analyzing variables.

8.2 FUTURE ENHANCEMENTS

Going forward, the recommendations created by the system should be evaluated with real users from the marketplace. Feedback from users should give a clear picture of whether or not the clusters are good. it was quite clear that the results from the silhouette scores and elbow method were not optimal. In order to achieve better clusters, a better rating system for brands should be implemented that utilizes views, likes and conversations to construct ratings. A more extensive analysis of how the data should be preprocessed and weighted needs to be made in order to get a better result from the clustering algorithm. Furthermore, K-means was used as the only clustering algorithm which might not be optimal in this case.

The K-means algorithm is a simple yet popular method for clustering analysis. Its performance is determined by initialisation and appropriate distance measure. There are several variants of K-means to overcome its weaknesses:

- **K-Medoids**: resistance to noise and/or outliers
- **K-Modes**: extension to categorical data clustering analysis

- **CLARA**: dealing with large data sets
- **Mixture models (EM algorithm)**: handling uncertainty of clusters

Therefore, other clustering techniques should be tested and compared to the K-means implementation. Another vital part of the success of the algorithm is which users and brands should be filtered out. In order to test different thresholds for minimum user activity and brand views, the visualization tool shown in this report should give the Plick team the possibility to alter the User-Brand matrix analyzed by the clustering algorithm by e.g. setting different thresholds and by adding/removing brands. For each run, evaluation plots should be updated and made visible in order to compare different runs.

REFERENCES

1. Haotian Wu and Bohua Li, Research on Customer Purchase Prediction Based on Improved Gradient Boosting Decision Tree Algorithm, 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), IEEE, 2022

2. V. Arul, Ashutosh Kumar, Aman Agarwal, Segmenting Mall Customers Data to Improve Business into Higher Target using K-Means Clustering, 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), IEEE, 2021

3. Ruiyu Zhao, Chen Li, Research on E-commerce Customer Segmentation Based on RFAC Model, 2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), IEEE, 2021

4. Ravi Sista, Riya Singh, Sunil Kumar Kumawat, Ritesh Dhanare, Techniques used by E-commerce industries for Customer analysis, 2021 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2021

5. Yong Huang, Mingzhen Zhang, Yue He, Research on improved RFM customer segmentation model based on K-Means algorithm, 2020 5th International Conference on Computational Intelligence and Applications (ICCIA), IEEE, 2020

6. Sahar Allegue, Takoua Abdellatif, Khalil Bannour, RFMC: a spending-category segmentation, 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), IEEE, 2020

7. Rahul Pradhan, Customer Segmentation Using Clustering Approach Based on RFM Analysis, 2021 5th International Conference on Information Systems and Computer Networks (ISCON), IEEE, 2021

8. Asmin Alev Aktaş, Okan Tunal, Ahmet Tuğrul Bayrak, Comparative Unsupervised Clustering Approaches for Customer Segmentation, 2021 2nd International Conference on Computing and Data Science (CDS), IEEE, 2021

9. Ritu Punhani, V. P. S Arora, Sai Sabitha, Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce, 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE, 2021

10. V. Chandra Shekhar Rao, Ishwarya Modika, Niranjan Polala, Customer Segmentation with K-Means++ as Initialization Algorithm, 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), IEEE, 2022