Customer Behaviour Analysis: Identifying High and Low-Value Customers through RFM, Clustering Algorithms, and Market Basket Analysis

॥वसुधैव कुटुम्बकम्॥

THESIS SUBMITTED TO
Symbiosis Institute of Geoinformatics

FOR PARTIAL FULFILLMENT OF THE M. Sc.
DEGREE
By

Niraj S. Gunjal

(Batch 2023-25 / PRN 23070243019)

Symbiosis Institute of Geoinformatics
Symbiosis International (Deemed University)
5th Floor, Atur Centre, Gokhale Cross Road, Model Colony,
Pune - 411016

वसुधैव कुटुम्बकम्

SYMBIOSIS INSTITUTE OF GEOINFORMATICS

Symbiosis International (Deemed University)

(Established under section 3 of the UGC Act, 1956)

Re-accredited by NAAC with 'A++' Grade | Awarded Category - I by UGC

Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

Project Completion Certificate

This is to certify that **Mr.** / Ms. **GUNJAL NIRAJ SHIVAJI** from **M.Sc. Data Science & Spatial Analytics** Programme, Batch **2023-2025** holding PRN **23070243019** has successfully completed Summer Internship entitled as **"Customer Behavior Analysis: Identifying High and Low-Value Customers through RFM, Clustering Algorithms and Market Basket Analysis"** at Symbiosis Institute of Geoinformatics, Symbiosis International (Deemed University) from **01-06-2024** to **31-07-2024.**

Internal Supervisor

Name: **Mr. Sahil Shah.**

Sign.

CERTIFICATE

Certified that this thesis titled **"Customer Behaviour Analysis: Identifying High and Low-Value Customers through RFM, Clustering Algorithms, and Market Basket Analysis"** is a bonafide work done by Mr. Niraj Shivaji Gunjal, at Symbiosis Institute of Geoinformatics, under our supervision.

Supervisor, Internal

Mr. Sahil Shah

Symbiosis Institute of Geoinformatics

Undertaking

The thesis titled **Customer Behaviour Analysis: Identifying High and Low-Value Customers through RFM, Clustering Algorithms, and Market Basket Analysis** is the Bonafede work of **Mr. Niraj S. Gunjal**, University can use and reuse the dissertation work in future.

Date:

Name of the student

Signature

ACKNOWLEDGEMENT

In this regard, I would like to take this opportunity to appreciate all the people who have made their effort in making a successful end of this project. First and foremost my heartfelt and special thanks to my guide Mr. Sahil Shah for this support, constructive feedback and help in the completion of this project. I am grateful for his encouragement and his guideline of where my work should be heading to.

I also feel privileged to have a level of access faculty members and staff of Symbiosis Institute of Geoinformatics for its resource and research support. Their consistent encouragement and information have been greatly benefitted for me.

I extend my heartful thanks to the Internal mentor, who continuously supported me throughout the project.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

LIST OF EQUATION

ABBREVIATION LIST

RFM - Recency, Frequency, Monetary

DBSCAN - Density Based Spatial Clustering of Applications with Noise

MBA – Market Basket Analysis

PREFACE

This project report presents a comprehensive analysis of customer segmentation using clustering techniques and RFM analysis. The main goal of this project is to identify and divide the customer based on purchasing behaviour and characteristics, by tracking their behaviour we can marketing strategies for them and can improve the customer relationship to the store.

In this project the dataset is analysed contains large number of transactions. To extract the valuable insights the dataset undergo through preprocessing like removing null values, handling ouliers and create a parameters for clustering.

In this project we use two clustering algorithm K-Means and DBSCAN for segmentation. K-Means is known for divide the data into well-defined clusters. To form clusters we use silhouette score and elbow method.

The DBSCAN algorithm is a density based algorithm the parameters we need need for dbscan are epsilon and minimum points.

The findings of the project gives the demonstration of the power of clustering algorithm. We also divide the the customer on the basis of the their RFM scoring by using RFM analysis.

The insights are gained from this analysis use to enhanced customer engagement and helps for growing the business through personalized marketing efforts.

Introduction

Currently, any business is surrounded by competition, and that is why defining the needs of various categories of consumers plays a crucial role in achieving the goal of constant development. A major tool that various organizations use to help them identify, differentiate and target different consumers is called customer segmentation (Wedel & Kamakura (2000)) which basically involves the division of a large market which is deemed to consist of a plurality of consumers into smaller sub-markets which are deemed to consist of specific segments of consumers. This paper has sought to establish that it is useful to identify and prioritize the different customers on the basis of their value in order to channel resources to customers that represent the most value and to retain those customers to increase profitability.

This project entails the application of data analysis in an attempt to categorize certain customers based on a given dataset that arises from an online retail setting. The starting point of each of the segments involves data collection and cleaning followed by preparation of the dataset for analysis. This is succeeded by an exploratory data analysis which aims at giving insights of the Richards' frequency, recency and monetary value (FRM) Hughes (1994) to make the segmentation. This is followed by engineering of features that enables the identification of patterns of customers' behaviour to further improve the segmentation model. To determine clusters of customers, Cancelous's Controller uses the Clustering Algorithms: K-Means(Hartigan, J.A. (1979)), DBSCAN cluster analysis Ester et al. (1996). The low risk segments are then further evaluated and substantiated for new and valuable insights that may be useful to the company's strategies. We also used a market basket analysis Agrawal et al. (1993) to find products that are frequently purchased, which product has a more sale this will help in defining business strategies also help in cross-sellng. This will help in improving customer engagement.

Some of the expected outcomes of the project are useful recommendations about the customers to be targeted and how to get the most out of unexpected campaigns. They can also be applied to other high-level tactical decisions that could involve issues of product differentiation, pricing or customer satisfaction that will eventually translate into improved and sustained patrons' loyalty, and greater revenues. Details of the methodology used in the study, along with the conclusions of the segmentation analysis, discussion and application of the findings of the study in business management and student strategy, and recommendations for the future of the study and its use in practice, are outlined in this report.

Objectives:

- Dividing customers into smaller, manageable groups with similar traits.
- Helps businesses understand and target their audience more effectively.
- Traditional methods struggle with large, complex datasets, but machine learning handles them efficiently.
- Machine learning techniques, like clustering, enable personalized marketing strategies that meet customer needs and preferences.
- Enhances customer experience.
- Aids in strategic decision-making, optimizing resource use, driving revenue growth, and gaining a competitive edge.

Literature Review:

The paper explores the use of K-means clustering for segmentation of the customers on the some category or behaviour like spending and income of the customer. The study indicate the importance of customer segmentation for marketing strategies and how to target that group of customers. The methodology contains data collection feature engineering, clustering, hyperparameter tuning to get result. The result give effective clusters or group of customers. (Patankar and Dixit, 2021) [1] This paper is from University of Information Technology Mynmar, the deals with K-Means clustering algorithm for customer segmentation. This study helps business to make group of customers based on their characteristics and apply marketing strategies and improve revenue. The paper use java based tool to apply effective K-means clustering for segmentation. (Maung and New, 2022). [2] The paper use K-means clustering to explore the importance of this algorithm to study customers characteristics and their needs to create marketing strategies. The dataset use in this paper is retail store. The categorise the customer on the basis of purchasing behaviour. The paper shows the importance automated segmentation over traditional. (Banduni and Ilavendhan, 2019) [3] The paper is about customer segmentation by using machine learning techniques like SVM(Support Vector Machine) and Neural Networks. They use the dataset private banking sector. They analyse the dataset and classify on basis of their behaviour. Their model highest accuracy for affluent customers compared to multi-layered neural networks. (Smeureanu et al., 2013) [4] This paper use the RFM analysis they segments the customers on the recency, frequency and monetary analysis. They use dataset of pakistans e-commerce sector. They use algorithms like K-means, Gaussian, hierarchical and DBSCAN. Their aim is to use marketing strategies for targeted customer. (Ullah and Mohmand, 2023) [5] The paper use and explore models like K-means and SAPK for segmentation of customer for making marketing strategies on their purchasing behaviour. (Tabianan et al., 2022) [6] The paper use a RFM analysis that recency, frequency and monetary analysis. They introduce a median based K-means algorithm that good on traditional k-means clustering it reduce the number of iterations. They also use fuzzy means and hierarchical clustering algorithms. There aim is find customer that we have to target using marketing strategies. (Christy et al., 2018) [7] The paper use algorithm that is multi-tier hierarchical super peer to peer (MT-SP2P). It use to enhance the speed and quality of distribution of customers in segment. They also addressed challenges like communication cost and handling large dataset. Their methodology improve the speed over by 90% and reduce the error. The study aim to customers satisfaction and customer service by making marketing strategies based on clusters formed. (Kuruba and Kashef, 2021) [8] The model used in this paper is RFM analysis(Recency, Frequency and monetary) with

FCA(Formal Concept analysis) for customer segmentation and understanding. The model shows the relationship of customers through hierarchical structure. They want to outform the traditional clustering algorithm. The aim is to suggest marketing strategies by implicating of FCA for customer satisfaction and retaintion and profitability of business. (Rungruang et al., 2022) [9] The paper shows the importance of customer segmentation for making marketing strategies and marketing. They group the customers on the based on behaviour or characteristics, geographies. They use K-means clustering algorithm because of its stability. (Thalkar, n.d.) [10] The author used the dataset of sports shop for customer segmentation they use fuzzy C-means, median based K-means algorithm. The use of median based K-Means algorithm effectively work here and reduce time and iteration of clustering of the customer. The aim of this paper is to form a marketing strategies for low values customer by giving them good offers. (Kalaiselvi et al., n.d.) [11] They use mall dataset for customers segmentation. They RFM analysis, K-Means, DBSCAN and hierarchical clustering. The study shows the importance unsupervised machine learning techniques. They form a cluster on the basis of silhouette score. The moto is to improve the customer retention rate and satisfaction (Supraja and Sairamesh, n.d.) [12].

TABLE 1 : Literature Review

Sr. No	**Title**	**Dataset**	**Techniques**	**Results**	**Challenges/ Limitations**
1.	(Patankar and Dixit, 2021)	The dataset using in this research paper is having information about customers including annual income and spending score	K-means clustering algorithm	After analysing the data and categorizing customers based on features like annual income and spending score, they obtained customer clusters	1. Using this method marketing will become more expensive. 2. Due to having less number of customers for segmentation problem the limited production or output will occur

2.	(Maung and New, 2022).	Mall customer segmentation data, source is not mention	K-means clustering unsupervised machine learning technique	The paper likely presents segmented customer groups based on common characteristics like gender, age, interest, and spending habits tohelp marketers effectively target different customer segments	The dataset used in this paper is of very small. Also the methodology does not addressed the challenges associated with high dimensional data
3.	(Banduni and Ilavendhan, 2019)	The dataset used in this paper is between 1/12/10 to 9/12/2011 of unregistered UK broker contains information of transactions and geographic data.	K-means clustering algorithm	The result according to paper that the orange cluster is contains high value customers , green cluster contain low value customers, and the blue and red contains high opportunity customers customers. Additionally, the study highlighted the importance of understanding	The dataset in this paper is unbalanced and it can affect the accuracy results. The paper primarily focus on internal validation on above external validation.

				customer needs and preferences to provide for customers satisfaction, and services	
4.	(Smeureanu et al., 2013)	The dataset used in the paper "Customer segmentation in the private banking sector using machine learning techniques" consists of 2783 represents active carholdersin a commercial bank of romania	Neural networks and SVM (Support Vector Machine)	The Neural Network model with one hidden layer and the logistic function as the activation function demonstrated good performance in customer segmentation.	The less amount of data is used in this research. The model only classify the customers into two groups affluent and mass

5.	(Ullah and Mohmand, 2023)	Pakistani w-commerce dataset by Zeeshan-ul-Hassan, containing data from 1 July 2016, to 28 August 2018.	k-means, agglomerative hierarchy, DBSCAN, and SOM. RFM (Recency, Frequency, and Monetary) Analysis	The study uses RFMT analysis, like K-means, DBSCAN on E-Commerce dataset to identify three distinct factors to enhance the customer relationship, marketing approaches.	The challenges and limitations in the results and methodology of the study include potential bias from the selected clustering algorithms,
6.	(Tabianan et al., 2022)	The dataset used in this paper is malaysias e-commerce dataset	K-means clustering	The paper segments the customer on their behaviour to help business to target their audience and emphasis the importance of data analysis for marketing statergies. The SAPK + K-Means algorithm was	The limitations such as working small data both comparison and modelling and clustering, specially the data is about business management and sales financial data.

				found to have a low error rate in clustering data sets.	
7.	(Christy et al., 2018)	The dataset used in the paper consists of 18,267 instances with eight attributes, including information such as customer ID, product name, price, date and time of purchase, etc. The dataset contains customer purchase information from 1-12-2010 to 09-12-2011	K-means clustering algorithm, Fuzzy C-means clustering, RM K-means clustering	The study first use RFM analysis on the transactional data and then by using this analysis they extend to traditional K-means and fuzzy c-means algorithm	The fuzzy C-means algorithm, RM K-means clustering take more time to analyzed and processed the data.

8.	(Kuruba and Kashef, 2021)	The dataset used in this paper consists of mainly three dataset including insurance, credit and stock dataset.	K-Means (KM) Minibatch K-Means Fuzzy c-Means (FCM) Self- organizing Map (SOM) Density-based spatial clustering (DBSCAN)	The study introduced a multi-tier hierarchical super-peer P2P network architecture for distributed customer segmentation, outperforming centralized methods with a 90% reduction in clustering error and a 90% speed enhancement.	The challenge in methodology making sure the clustering algorithm works Well as the data size grow
9.	(Rungruang et al., 2022)	The dataset used in the research paper is the "online retail II dataset" from the UCI Machine Learning Repository.	RFM Model-Based Clustering Analysis, RFM Model-Based Clustering Analysis	The study integrated the RFM model and Formal Concept Analysis (FCA) for customer segmentation, providing effective clustering and knowledge representation. Results showed improved customer insights, aiding in practical marketing strategies and customer relationship	The study has several limitations, including problems with data quality, algorithm sensitivity to outliers, difficulties interpreting complicated results, and possible lack of generalizability to various business contexts.

				enhancement.	
10.	(Thalkar, n.d.)	The dataset is not explicitly mentioned but the paper shows transactional customer data from shopping vendors,	K-means clustering algorithm	The goal of customer segmentation is to catagorised the customers ontheir buying charactristics and behaviour it will help for increase the sales of the company.	In this research paper only one algorithm is used, using only one algorithm limits comparison between other algorithms that might be more suitable for different types of data.
11.	(Kalaiselvi et al., n.d.)	The dataset of sports store from turkey. The collected is having information of 2 years	K-means clustering Fuzzy C-Means clustering novel Repetitive Median based K-Means method	The use of the K-Means algorithm, Fuzzy C-Means clustering, and a novel Repetitive Median based K-Means method helped in segmenting customers effectively with reduced iteration and time, leading to more targeted marketing strategies and improved customer satisfaction	RFM analysis only relies on transactional data and does not capture other important factors like customer satisfaction, preferences, or demographics. This may limit the depth of customer insights.

12.	(Supraja and Sairamesh, n.d.)	Online Retail Dataset for 2010-2011 and the Mall Customers Dataset	K-means, mini-batch k-means, hierarchical, Density-based spatial clustering of applications with noise (DBSCAN), (Gaussian Mixture Models) GMM, and MeanShift clustering.	Clustering analysis revealed the importance of Recency analysis for efficient clustering and the significant roles of Age and Spending Score in segmentation	They have limited attributes in the dataset and the sample size of dataset is small. The 5 attributes does not provide a comprehensive view of customer behaviour.

Methodology

Dataset Overview:

The dataset is taken from UCI-Machine Learning Repository, The dataset is about a UK-based online non registered retail store. It contain information of transaction between 1/12/2010 to 9/11/2011. The dataset contains 5,41,909 rows and 8 columns.

Attributes/Columns in dataset:

TABLE 2 : Dataset Description

Invoice No.	This is the number is uniquely assigned to each transaction of customer
Stock Code	The number that is uniquely assigned to each product
Description	Product item and name of that item
Quantity	The quantity of each product and item per transaction
Invoice Date	Transaction Date
Unit Price	Product price per unit
Customer ID	The unique number of each customer
Country	The country name each customer that they belong

Flowchart

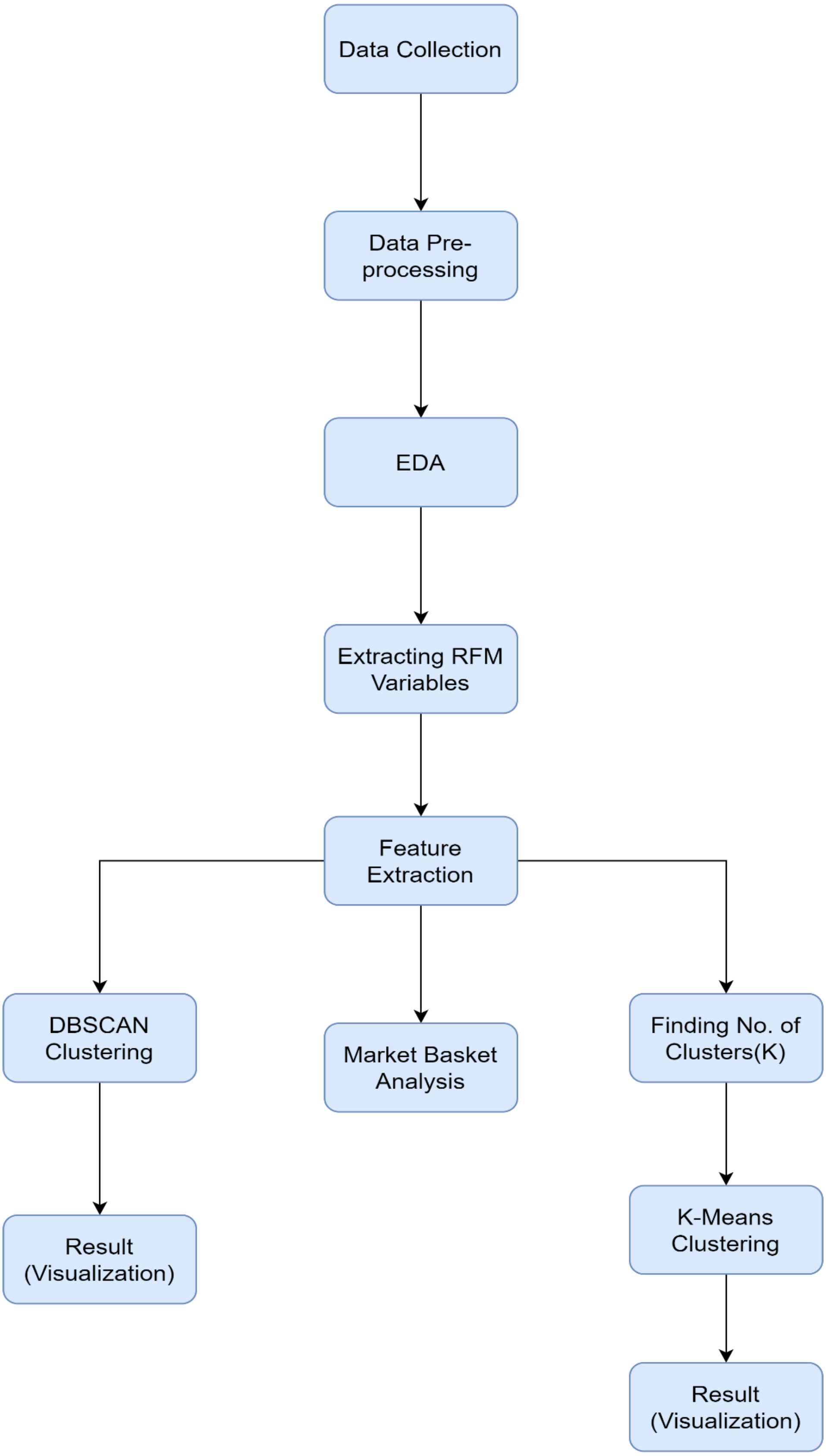


Figure 1: Methodology Flowchart

Importing Necessary Libraries:

TABLE 3 : Libraries

Library Name	Description
Numpy	To perform a numeric operation.
Pandas	It is used to work with dataframes.
Datetime	It provide a classes to work with date and time.
Matplot	It is use to visualize the data.
Seaborn	It is use to plot and visualize the graph.
Sklearn	It is a machine learning library it provide a tools for data analysis, and Algorithms.
Standard Scaler	It is use for scaling the features.
Time	This library is use to work with time formats.
Sklearn.clusters	This module provides a various unsupervised clustering techniques.
Silhoutte_Score	It is a metric to evaluate the number of cluster.
KMeans	It is algorithm it seperates the data on the basis of similarity.
DBSCAN	It is a algorithm to formulate a clusters.
Counter	It is use to count the hashable objects.

Preparation of the data:

1. Data Cleaning:

First, we load the dataset of online retail store. Then we perform the operations on data to check the size and other information of the data. In dataset we can see the columns customer ID and description contains the null values. The null values are about 25% in customer id. Since the we have to check null values we are creating new column and fill the 1 when the ID is null and 0 when ID is not null. And we don’t doing a analysis on the description hence we are not filling the null values. Since the some customer id is null hence this order are not made by the customer that are already having a customer id that’s why we can not assign the orders to them this would alter the insights drawn from the data. We can fill the null values by using invoice numbers and it is a straightforward approach.

After that we can notice that the new id does not have any null values, but the column includes object datatype and we have to convert it into numeric, as mentioned if invoice number contain ‘C’ it means the orders were cancelled.

2. EDA(Exploratory Data Analysis) :

Now the dataset is clean now we can look to a numeric columns specifically quantity and unit price. By using df.describe() we can see that there are some negative values in Quantity and Unit price we can assume that the orders are returned or cancelled. By checking cancelled orders there are 36% orders were cancelled. Here we find that the average number of order per customer is 3.

After exploring the data:

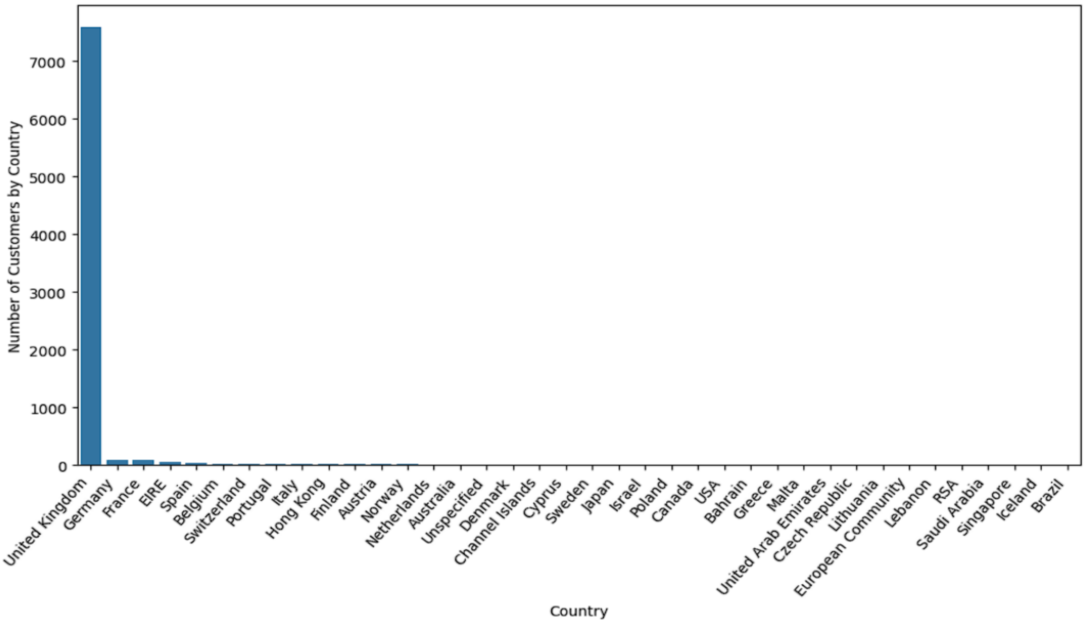


Figure 2: Number of Customer by Country

After we exploring the order and we find that in our dataset the customers from UK are most and produce a more revenue. Exploring the UK market we found that the 93% customers are from there so here we can continue analyzing UK transactions with customer segmentations. In the dataset the total customers of UK are 4,95,478. Now we have to drop the id's that have order quantity have less than 0. We have final count of UK customers are 3,54,345 and the total number for UK transactions are 16,649 and number of customers are 3921. Now our dataset is ready for RFM analysis.

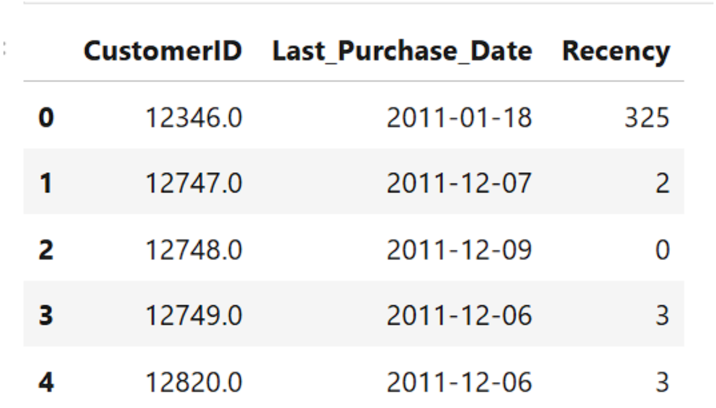
3. RFM Analysis and Feature Engineering:

RFM analysis(Blattberg, R.C(2008)) is a technique that categorizes customers based on three key metrics: Recency, Frequency, and Monetary Value.

- Recency (R): Recency means the time period between the time since a customer's last purchase
- Frequency (F): Frequency represents the number of purchases. How many times customer visits the store
- Monetary Value (M): Monetary Value means the total amount spent by the customer

First, we have to remove fake IDs that trouble our analysis; fake IDs, including the customer, do not make any orders. To calculate recency, we need to choose a date as a point reference we have to check the last purchased date of a customer that how many days before customer ordered and in dataset the most recent date is 12/09/2011 so we will use is as a reference.

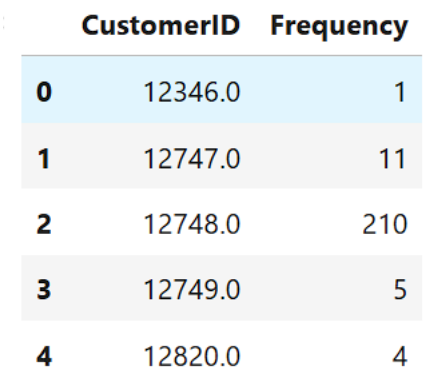
TABLE 4: Recency



	CustomerID	Last_Purchase_Date	Recency
0	12346.0	2011-01-18	325
1	12747.0	2011-12-07	2
2	12748.0	2011-12-09	0
3	12749.0	2011-12-06	3
4	12820.0	2011-12-06	3

For calculating frequency we have to check how many times customer purchase and we have to count how many invoices each customer has.

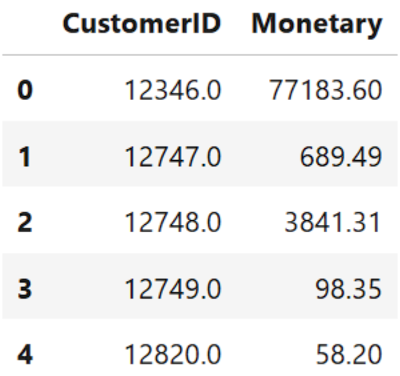
TABLE 5: Frequency



	CustomerID	Frequency
0	12346.0	1
1	12747.0	11
2	12748.0	210
3	12749.0	5
4	12820.0	4

Monetary value is calculated by adding the cost of customer purchases, that is total spending of the customer.

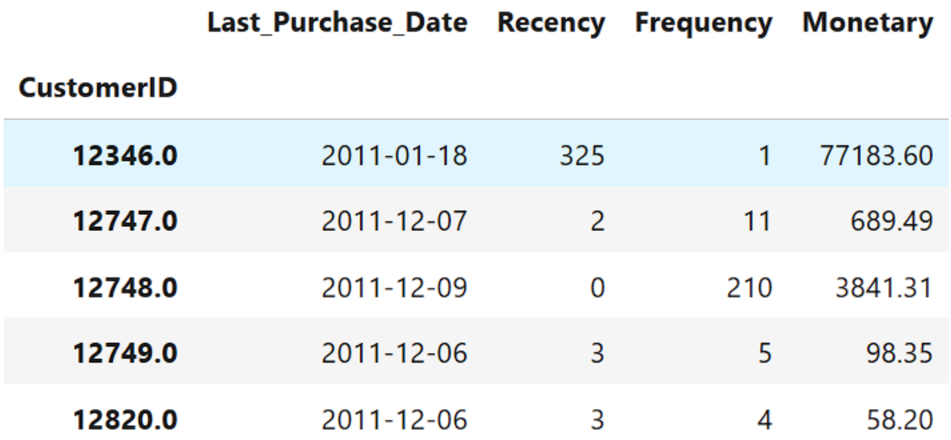
TABLE 6: Monetary



	CustomerID	Monetary
0	12346.0	77183.60
1	12747.0	689.49
2	12748.0	3841.31
3	12749.0	98.35
4	12820.0	58.20

After that we created a RFM table,

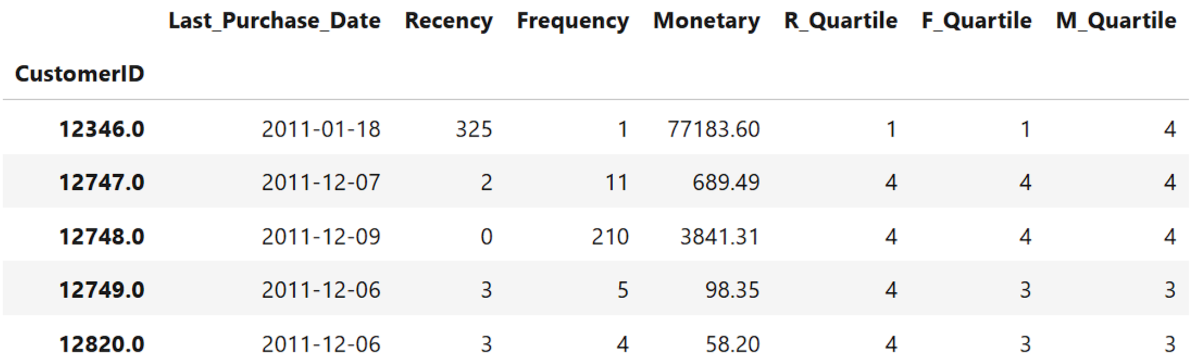
TABLE 7: RFM Table



	Last_Purchase_Date	Recency	Frequency	Monetary
CustomerID				
12346.0	2011-01-18	325	1	77183.60
12747.0	2011-12-07	2	11	689.49
12748.0	2011-12-09	0	210	3841.31
12749.0	2011-12-06	3	5	98.35
12820.0	2011-12-06	3	4	58.20

we merge all three attributes recency, frequency and monetary. On the basis of the table, we create a customer segmentation with the RFM model. We use a RFM scoring, for creating customer segments we use quartile. We will assign a score from 1 to 4 to each category (Recency, Frequency, and Monetary) with 4 being the highest/best value.

TABLE 8: Quartile Table

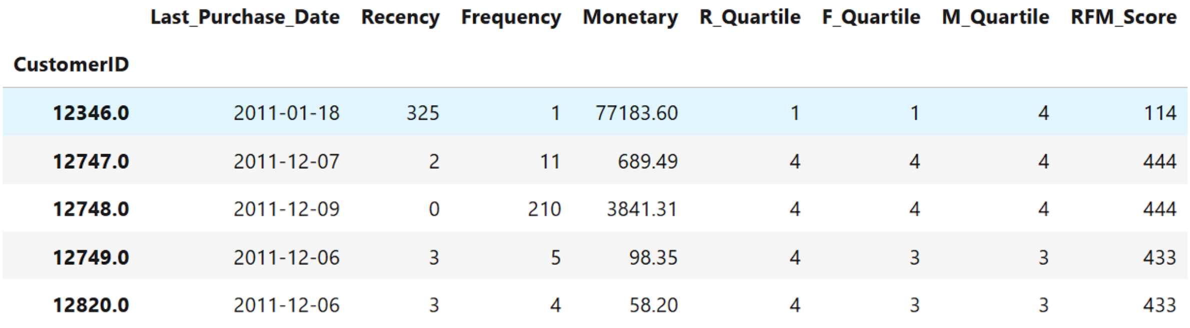


	Last_Purchase_Date	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile
CustomerID							
12346.0	2011-01-18	325	1	77183.60	1	1	4
12747.0	2011-12-07	2	11	689.49	4	4	4
12748.0	2011-12-09	0	210	3841.31	4	4	4
12749.0	2011-12-06	3	5	98.35	4	3	3
12820.0	2011-12-06	3	4	58.20	4	3	3

The final RFM score is calculated by combining the individual RFM values. And we create a two segments since high recency is bad, while high frequency and monetary is value is good. Here we create a two functions for that. After scoring each customer we combine the scores for segmentation. By using we can find the number of customers in each segment like best customer, loyal customer, big spender, almost lost, lost customer, lost cheap customer.

Table: RFM Score based on Quantile:

TABLE 9: RFM Scoring



	Last_Purchase_Date	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFM_Score
CustomerID								
12346.0	2011-01-18	325	1	77183.60	1	1	4	114
12747.0	2011-12-07	2	11	689.49	4	4	4	444
12748.0	2011-12-09	0	210	3841.31	4	4	4	444
12749.0	2011-12-06	3	5	98.35	4	3	3	433
12820.0	2011-12-06	3	4	58.20	4	3	3	433

4. Choosing Predictive Model:

I. **Applying K-means Clustering:** K-means clustering is a unsupervised machine learning algorithm which divides the data points that have same characteristics. HARTIGAN, J.A., WONG, M.A.(1979)

The objective function for K-means clustering:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 ......(1)$$

EQUATION 1: Objective Function

Where,

J: objective function

K: number of cluster

Ci: It represents a points that are assigned to centroid

X: data points

Ui: it represents a centroid of i-th cluster

$\boldsymbol{||x - \mu_i||^2}$**:** It is squared Euclidean distance between points and centroid.

Clustering algorithm K-Means operates on data dividing it into K groups. The detailed procedure of k-means refers to assigning points repeatedly to the closest cluster center while at the same time, revising the cluster center until the algorithm matches.

First, we drop columns of quartiles. Then we find a correlation between the features using heatmap, and here we can see that there is a negative correlation between recency: frequency and recency : monetary, but there is a positive correlation between Frequency : Monetary.

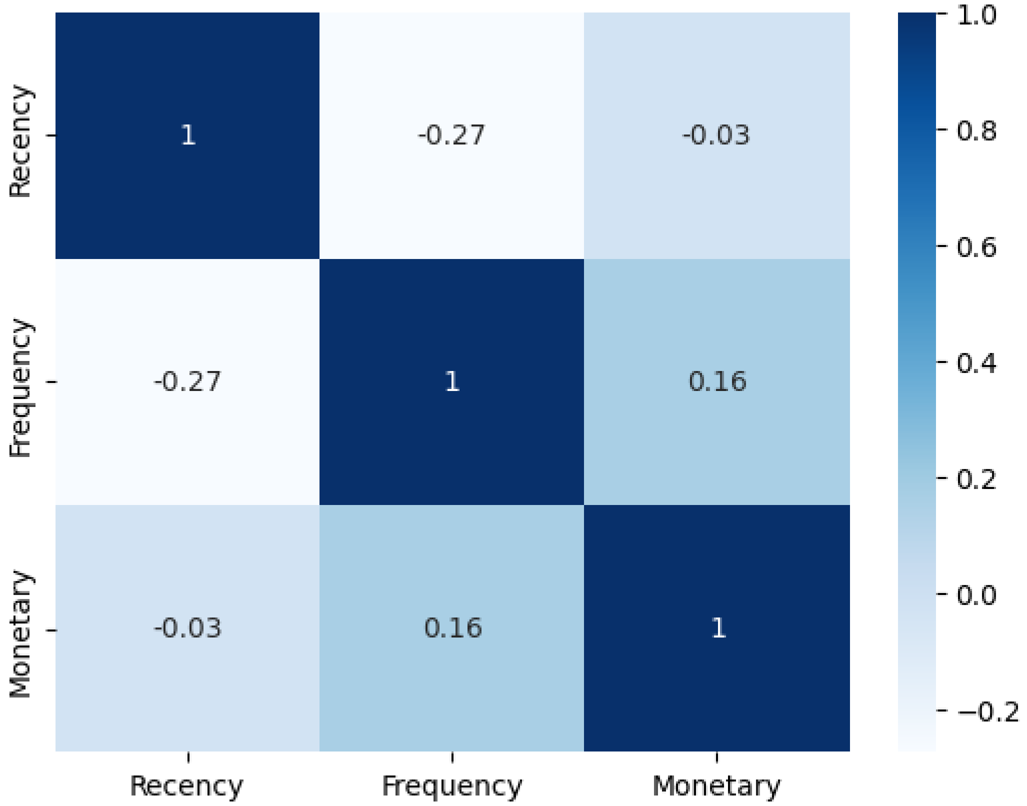
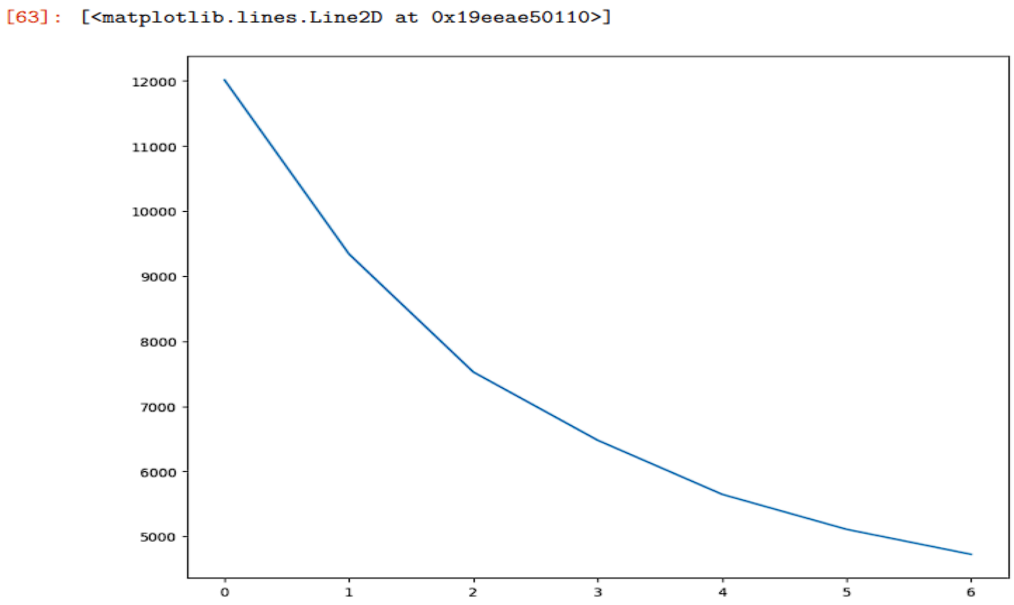


Figure 3: Correlation

And there are outliers present in 3 variables, since clustering algorithm needs a normal distribution, normalization of data is required, logarithmic transformation is used here for normalization, This technique helps to stabilize the variance and make the data more normally distributed. Then we again check the correlation between them and we can see that monetary and frequency are strongly related so we can use those two variables in our K-means model.

For determine the number of clusters we elbow method it shows us number of clusters.

Elbow Method:



```
[63]: [<matplotlib.lines.Line2D at 0x19eeae50110>]
```

Figure 4: Elbow Method

Here we can see that is the curve has taken sharp turn at 2 means the shape of curve is like our elbow so we take the value of K=2.

Silhouette Score:

For K-means algorithm we have to set number of cluster K as our wish, but initially how many required is not specifies. We will attempt various clustering numbers and see their [silhouette coefficient] The silhouette coefficient of a state reflects how close it is to the interacting cluster it has been assigned, ranging from negative one, for dissimilar, to positive one for similar.

To calculate the silhouette score following formula is used:

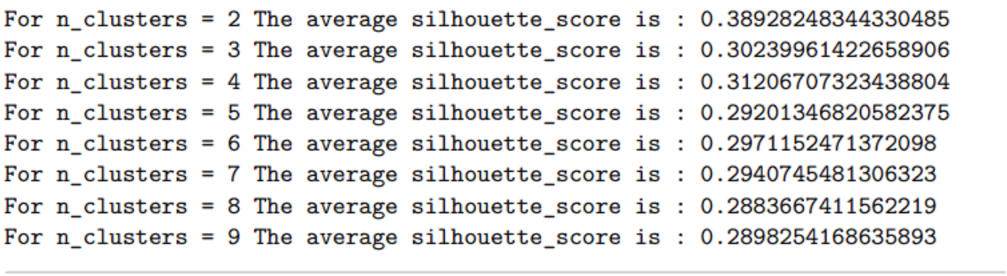
$$s = \frac{b-a}{max(a,b)} \quad \ldots\ldots.(2)$$

EQUATION 2: Silhouette Score

S: It is a silhouette score for a data points

a: It is mean intra cluster distance for a data point

b: It is a mean nearest cluster distance



```
For n_clusters = 2 The average silhouette_score is : 0.38928248344330485
For n_clusters = 3 The average silhouette_score is : 0.30239961422658906
For n_clusters = 4 The average silhouette_score is : 0.31206707323438804
For n_clusters = 5 The average silhouette_score is : 0.29201346820582375
For n_clusters = 6 The average silhouette_score is : 0.2971152471372098
For n_clusters = 7 The average silhouette_score is : 0.2940745481306323
For n_clusters = 8 The average silhouette_score is : 0.2883667411562219
For n_clusters = 9 The average silhouette_score is : 0.2898254168635893
```

Here we can see that when n_cluster = 2 the best silhouette score is obtain. After that we visualize the cluster using scatter plot.

II. Applying DBSCAN:

In this project we use algorithm that is density based spatial clustering of application with Noise algorithm based on that we segment a customers on their transaction behavior on the features that we selected that are frequency and monetory.

DBSCAN is powerfull clustering algorithm that groups the point together that are closely related with each other and the points with low density that are outliers. In DBSCAN we do not have specify the number of clusters in advance, DBSCAN use two parameter that are Epsilon and min_samples.

Objective Function for DBSCAN:

DBSCAN does not come with a criterion which is optimized unlike k-means that aims at minimizing the sum of the squared distances to the centroid. Instead, its objective is to:

Describe regions in the space made of points which form a cluster and are density-reachable. Particularly Differentiate spares areas are noise or outliers. The algorithm uses two parameters to define the nation of density

Ɛ (epsilon): The size of the neighborhood, based on which radius the shows should be selected.

MinPts: The least number of points that need to exist in a region to be dense is known as Core point.

Here we take the value eps = 0.5 and min_sample =15 and plot the cluster.

The algorithm identify different clusters along with their noise points. The cyan color cluster shows the highest proportion of customer that they have similar characteristics according to their monetary and frequency values.

The yellow is the secondary cluster and shows another segment of customers that have the same patterns. The red color cluster has a low value of frequency and monetary, that is why they are low-value customers, and we can see that there are some noise points in the purple color.

III. Market Basket Analysis/Association Rule:

MBA(Market Basket Analysis) is a technique of association and data mining it is use to identify the relationships and patterns between the items that are frequently purchased with other item or purchased together. It analyze the large datasets of customer data and find the relationship between products and helps business to make strategies to get profit it suggests strategies such as where place the product, it helps in promotions and help in marketing and also in cross-selling of product.

In this project, we applied a market basket analysis to determine the which products are buy frequently, which products have most sales.

Here we find the which item is having most sell,

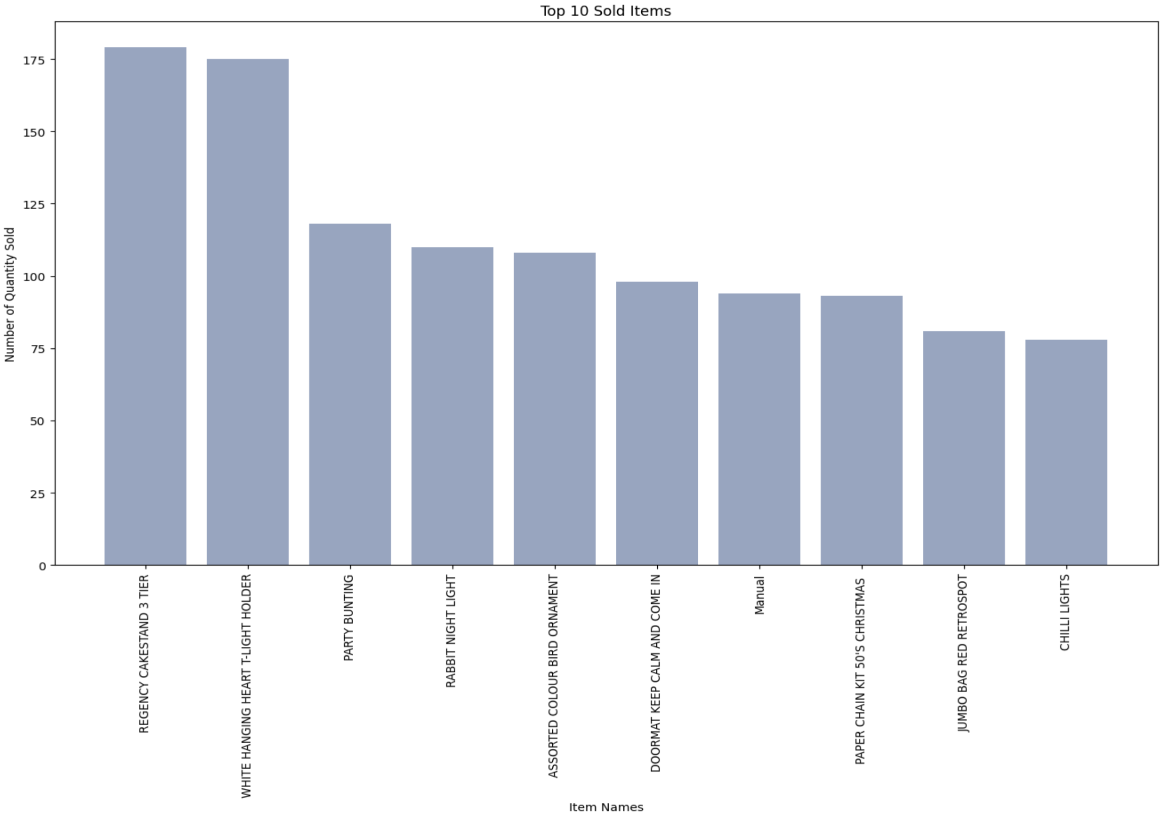


FIGURE 5: Number of Item Sale

Also we find in analysis highest number of items get sold in November,

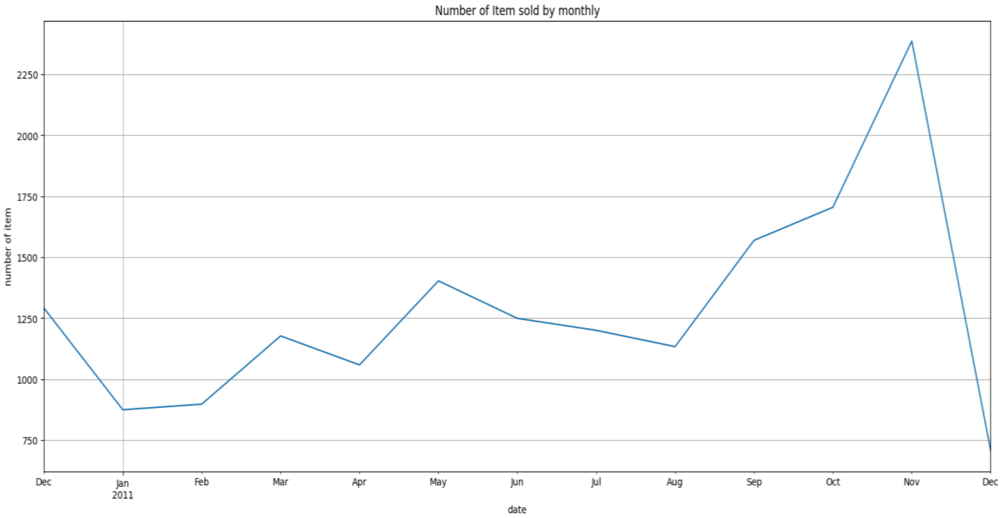


FIGURE 6: Number of Items Sold Monthly

After the analysis we drop the columns that unwanted and sort the values according to the customer id and we only take two columns for basket analysis. We use ‘customer id’ and ‘description’ for analysis and list all the items according to there description.

After that we import apriori algorithm.

Apriori Algorithm:

This is the method that is used in association rule mining. This widely use in market basket analysis to discover the relationship between the items. Agrawal, R., & Srikant, R. (1994). It finds the frequent items in transactional itemset. It will efficient for large dataset.

Following are steps for apriori algorithm:

1) Step 1: In first step we set a minimum support and confidence
2) Step 2: Generate Candidate Itemsets
3) Step 3: Prune Infrequent Itemsets
4) Step 4: Generate Association Rules
5) Step 5: In last step we evaluate rules using Lift

For the apriori algorithm, following parameters, we have defined min_support= 0.02. We taking it low because we have considered the items that are frequently appear. Min_Confidence= 0.05, min_lift= 3 here because lift over 1 shows a positive relation here we gave 3 for finding a strong relation, min_length=2 here we gave 2 because we have to ensure that the association rule involves at least 2 items..

For apriori algorithm following are the metrics:

- **Support:**

Support contain a specific items of a transaction in a proportion, the support tells you the how repitively or frequently the item is appear in data.

Formula: $\text{Support}(A) = \frac{\text{Number of transactions containing item A}}{\text{Total number of transactions}}$………(3)

EQUATION 3: Support

- **Confidence:**

It is a relationship that tells and measures that if item A is purchased then how likelihood that item B is also purchased. It gives a conditional probability of B given A.

Formula: Confidence $(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A)}$…………..(4)

EQUATION 4: Confidence

- ## Lift:

Lift makes it easy to establish the strength of the connection between A and B Lift gives the probability if an item B being bought whenever it is given that an item A had been bought compared to a random probability

Formula: Lift$(A \Rightarrow B) = \frac{\text{Confidence } (A \Rightarrow B)}{\text{Support } (B)}$…………..(5)

EQUATION 5: Lift

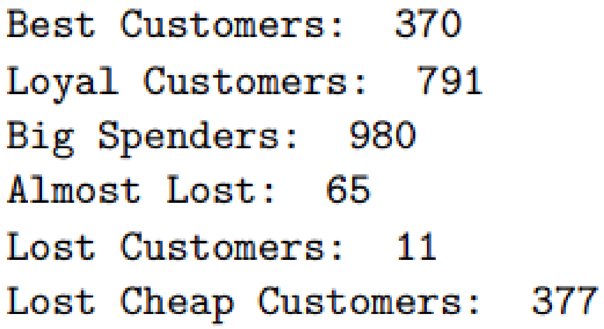
Here, we use min_support = 0.002, min_confidence = 0.05, min_lift = 3, min_length = 2 after that we get results after using this parameters and we get the item set that are most frequent to each other and have probability of that item B is purchased with item A.

Results

This section represents the key findings of the project by analysis of customer data using RFM analysis, K-means and DBSCAN clustering algorithms.

RFM Scoring: In this we give score on the basis of recency, frequency and monetary value. We assign 4 as highest and 1 as lowest. By merging the recency, frequency and monetary score the 444 RFM score is highest and this are the best value customers, and 111 is the lowest this are the low value customers and some customers have mix scores because some of them buy one time but spend more that’s why they have best monetary score and low recency and frequency. And on the basis of RFM scoring we get best, loyal, big spenders, lost, lost customer and lost cheap customer.

After applying RFM scoring analysis we found that:



```
Best Customers:  370
Loyal Customers:  791
Big Spenders:  980
Almost Lost:  65
Lost Customers:  11
Lost Cheap Customers:  377
```

Figure 7. Segment of customers

K-Means Clustering:

In K-Means clustering we find the number of clusters using silhouette score and elbow method and we get best silhouette score at value K = 2. We segment the customer as high value and low value customers according to their centroid value.

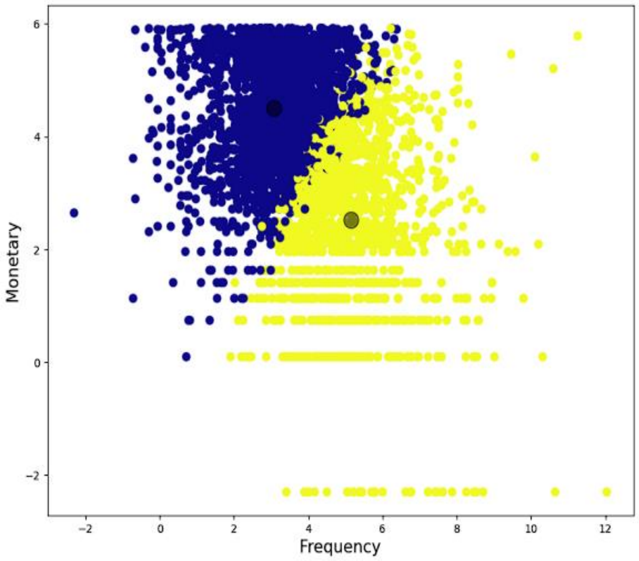


Figure 8: K-means Cluster

The yellow color have the centroid value between (0.5,4) cluster shows the low value customers and the blue color cluster have the centroid value between (2,6) shows the high value customers.

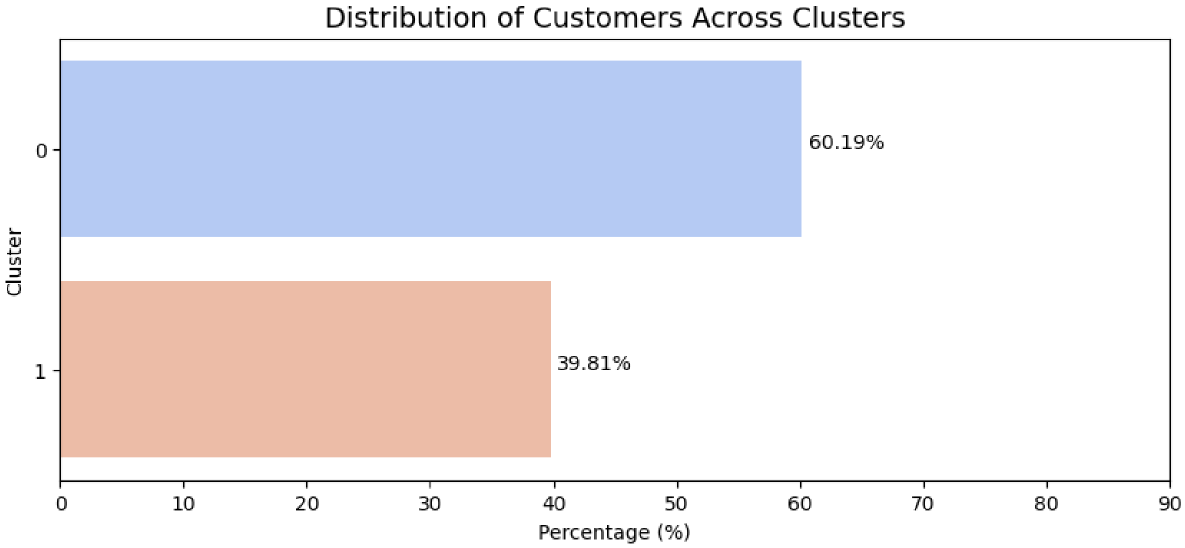
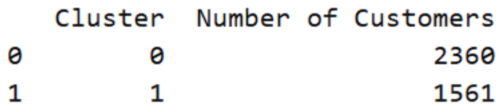


Figure 9: Comparative Analysis

- **Cluster 1 (High-Value Customers):** This cluster consist of high value customers that they are high spenders and have frequent purchases and strong engagement and loyalty.

- **Cluster 2 (Low-Value Customer):** This cluster consist of customer who having low value that is they are low in total spending and they don’t have frequent purchases and less engaged with store.



```
   Cluster  Number of Customers
0        0                 2360
1        1                 1561
```

FIGURE 10: Number of customer

DBSCAN Clustering:

In DBSCAN we provide 2 parameters epsilon and min_sample here we provide epi = 0.5 and min_sample =15 and we get 3 clusters with noise point and outliers. Here we get dense cluster with cyan color that means the points are having same characteristics. And the purple color is the noise point and we get another segment with yellow having high density and the red cluster having a outliers.

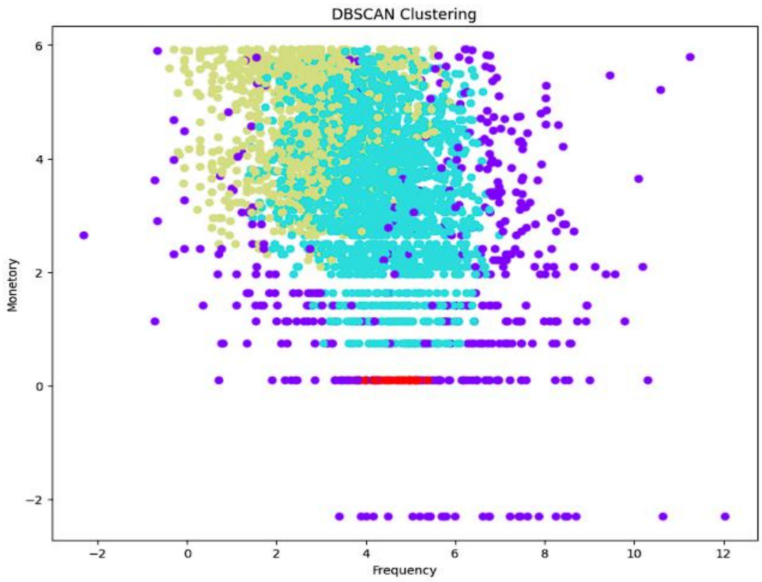
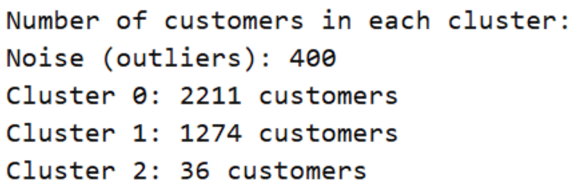


Figure 11: DBSCAN Cluster

- **Cluster 1 (High-Density and High-Value):** This cluster represent a cluster with consistent in purchasing and having similar characteristics.
- **Cluster 2 (Secondary Clusters):** This cluster is consist of diverse purchasing behaviour at have less frequency.
- **Noise Point:** In this we get customers who are irregular to buy from store and don't match other clusters or group.



```
Number of customers in each cluster:
Noise (outliers): 400
Cluster 0: 2211 customers
Cluster 1: 1274 customers
Cluster 2: 36 customers
```

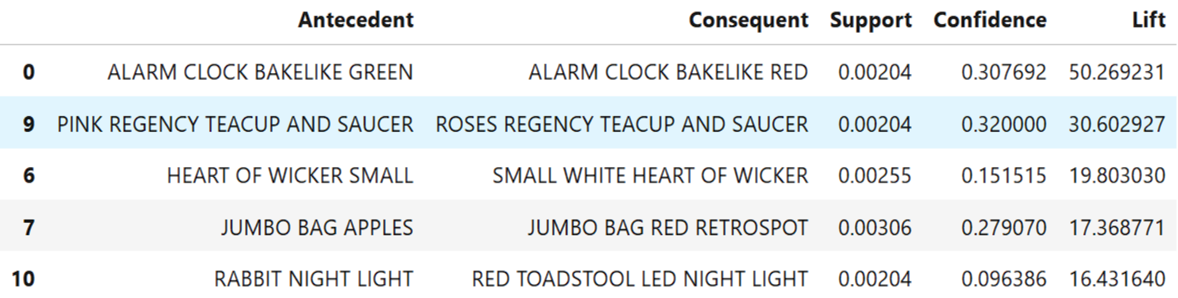
FIGURE 12: Number of Customers

Market Basket Analysis:

In the result we can the association rule with having columns like support, confidence and lift, right hand side and left hand side.

Here is the analysis,

TABLE 10: Market Basket Analysis



	Antecedent	Consequent	Support	Confidence	Lift
0	ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED	0.00204	0.307692	50.269231
9	PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.00204	0.320000	30.602927
6	HEART OF WICKER SMALL	SMALL WHITE HEART OF WICKER	0.00255	0.151515	19.803030
7	JUMBO BAG APPLES	JUMBO BAG RED RETROSPOT	0.00306	0.279070	17.368771
10	RABBIT NIGHT LIGHT	RED TOADSTOOL LED NIGHT LIGHT	0.00204	0.096386	16.431640

1. High Lift and Confidence Rule:

In the table alarm clock bakelike green and red have a highest lift value that is 50.269231 and the confidence is a 0.3076 this suggest that if customer buy alarm clock bakelike green then there is strong chances of buying of alarm clock bakelike red there are 50 times more chances than random.

2. Other Significant Associations:

Pink regency tea cup and saucer and roses regency teacup and saucer having a lift 30.60 and confidence 0.32 this showing the association rule between these two items. Also a heart of wicker small and small white heart of wicker have lift of 19.80 that means these two are bought frequently.

3. Moderate Associations:

The jumbo bag apples and jumbo bag red retrospot have lift of 17.37 and rabbit night light and red toadstool led night light having a low 16.43 and also having a lower confidence respectively they are showing a moderate asscociation between them.

Frequently Purchased Item’s in Market Basket Analysis:

Table 11: Frequently Purchased Items:

Antecedent	Consequent
'ALARM CLOCK BAKELIKE RED'	'ALARM CLOCK BAKELIKE GREEN'
'ANTIQUE SILVER T-LIGHT GLASS'	'VICTORIAN GLASS HANGING T-LIGHT'
'PARTY BUNTING'	'ASSORTED COLOUR BIRD ORNAMENT'
'DOORMAT KEEP CALM AND COME IN'	'DOORMAT UNION FLAG'
'DOORMAT KEEP CALM AND COME IN'	‘WHITE HANGING HEART T-LIGHT H OLDER'
'HEART OF WICKER SMALL'	'REGENCY CAKESTAND 3 TIER'
'SMALL WHITE HEART OF WICKER'	'HEART OF WICKER SMALL'
'JUMBO BAG RED RETROSPOT'	'JUMBO BAG APPLES'
'JUMBO BAG RED RETROSPOT'	'LUNCH BAG RED RETROSPOT'
'ROSES REGENCY TEACUP AND SAUCER'	'PINK REGENCY TEACUP AND SAUCER'
'RABBIT NIGHT LIGHT'	'RED TOADSTOOL LED NIGHT LIGHT'
'RABBIT NIGHT LIGHT'	'REGENCY CAKESTAND 3 TIER'
'SET OF 3 REGENCY CAKE TINS'	'REGENCY CAKESTAND 3 TIER'
'SET OF 3 REGENCY CAKE TINS'	'ROSES REGENCY TEACUP AND SAUCER'
'SET OF 3 CAKE TINS PANTRY DESIGN'	'SET OF 3 REGENCY CAKE TINS'

Discussion

In this project we use a application of RFM scoring, K-Means and DBSCAN analysis for customer segmentation and to understand their behaviour and characteristics. The RFM analysis categorised the customers based on purchasing behaviour, transaction history and gives score according to it and classifies the customer like best, loyal, big spenders, almost lost, cheap customers.

We use K-Means algorithm using this segments based on their RFM scores and forms clusters for targeted marketing strategies. Here we use elbow and silhouette score method for define a number of clusters and it form 2 clusters as low value and high value customers.

DBSCAN algorithm analyze the dense clusters as well as outlies and it for 4 clusters including noise points and outliers. The dbscan gives high value and most spender customers group and outliers and anomalies it provide deep understand of diverse customer base.

The Market Basket Analysis and association rule identifies key products that buy together frequently in transactions, it show good opportunities to improve marketing strategies and cross selling. For example products like ‘Alarm Clock BakeLike Green’ and ‘Alarm Clock BakeLike Red’ showing high lift and confidence that shows customers buying them together. It helps businesses to understand the buying pattern of customer, to make strategies and helps to analyze the customers trend.

Overall analysis provide a valuable insights for customer segmentation for making personalised marketing strategies for customers satisfaction and profitability of business.

Conclusion

The clustering methods were employed and used in this project for the purpose of segmenting customers based on their purchase behaviour. The findings provided further to improve the management of customer relations as well as to design the marketing efforts. Algorithms such as K-Means and DBSCAN were applied in the analysis to explore different strategies of classifying customers.

Key Findings:

RFM analysis and scoring: In this project, we use RFM analysis and scoring to cluster the customers based on their purchases by finding their recency, frequency and monetary and based on that we find best and loyal customers and the customers that we need to approach with some marketing stratergies.

K-means Clustering: The k-means clustering algorithm identifies the 2 distinct customer clusters for segmentation by partitioning them. The algoritms identify the cluster as per the high value customers and low value customer. By analysing and using the output we can make a marketing strategies for the customers.

DBSCAN: The DBSCAN algorithm forms a cluster based on epi(∈) and minimum points and this project it gives 4 clusters. It also gives some noise points ouliers.

We get great performance for K-means clustering as we can separate the clusters and it provide a solid foundation for understanding customers and provide marketing strategies according to it. And because of dataset structure the performance of K-means is good over DBSCAN.

Market Basket Analysis(MBA): On the basis of MBA and association rule we can see that there is relationship between some product pair such as clocks with different color and different teacups and saucers. In this analysis the high lift shows customers buy one item frequently with other item that is associated with it. This helps in to draw some marketing strategies to improve product selling and cross selling and helps to understand the customers behaviour to enhance sales performance.

References

1. Patankar, N., Dixit, S., Bhamare, A., Darpel, A., & Raina, R. (2021). Customer segmentation using machine learning. In Advances in parallel computing. https://doi.org/10.3233/apc210200

2. Maung, Yoon & Aung, Zu & Nwe, Than & Tin, Hlaing Htake Khaung. (2022). Customer Segmentation Analysis.

3. Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. Expert Systems With Applications, 237, 121449. https://doi.org/10.1016/j.eswa.2023.121449

4. Manjunath, Y. S. K., & Kashef, R. F. (2021). Distributed clustering using multi-tier hierarchical overlay super-peer peer-to-peer network architecture for efficient customer segmentation. Electronic Commerce Research and Applications, 47, 101040. https://doi.org/10.1016/j.elerap.2021.101040

5. Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., Mahmoud, H. A., & Huda, S. (2023). Customer analysis using Machine Learning-Based classification algorithms for effective segmentation using recency, frequency, monetary, and time. Sensors, 23(6), 3180. https://doi.org/10.3390/s23063180

6. Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243. https://doi.org/10.3390/su14127243

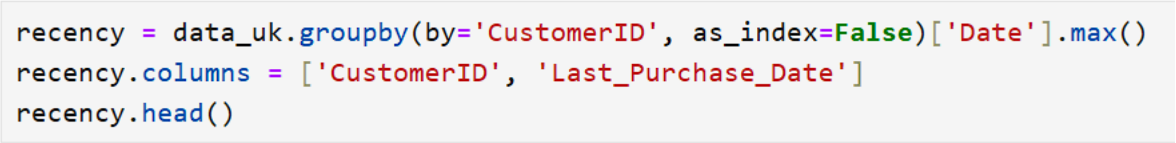
7. Smeureanu, I., Ruxanda, G., & Badea, L. M. (2013). CUSTOMER SEGMENTATION IN PRIVATE BANKING SECTOR USING MACHINE LEARNING TECHNIQUES. Journal of Business Economics and Management, 14(5), 923–939. https://doi.org/10.3846/16111699.2012.749807

8. Thalkar, N. V. R. (2021). Customer segmentation using machine learning. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 207–211. https://doi.org/10.32628/cseit217654

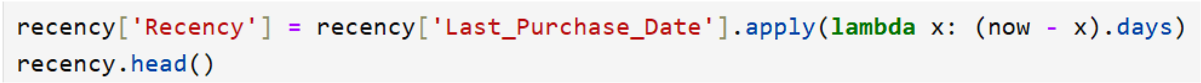
9. Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. Journal of King Saud University - Computer and Information Sciences, 33(10), 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004

10. Nasirian, K., & Taheri, H. (2019). EPRA International Journal of Research & Development (IJRD). EPRA International Journal of Research & Development (IJRD). https://doi.org/10.36713/epra2016

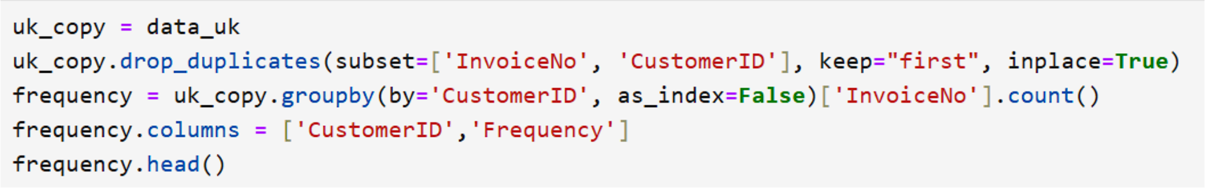
Appendix:



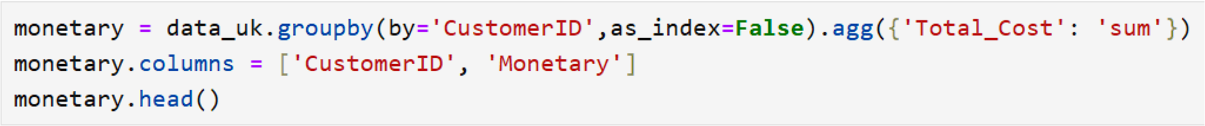
```
recency = data_uk.groupby(by='CustomerID', as_index=False)['Date'].max()
recency.columns = ['CustomerID', 'Last_Purchase_Date']
recency.head()
```



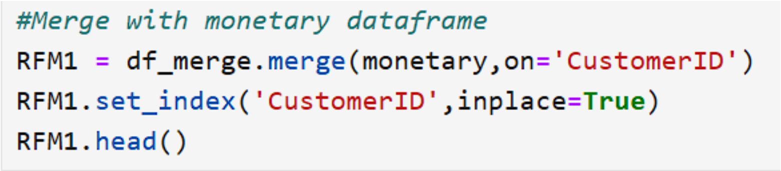
```
recency['Recency'] = recency['Last_Purchase_Date'].apply(lambda x: (now - x).days)
recency.head()
```



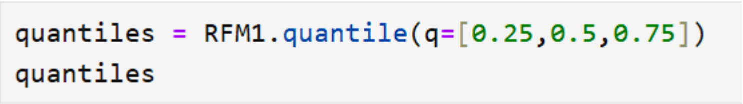
```
uk_copy = data_uk
uk_copy.drop_duplicates(subset=['InvoiceNo', 'CustomerID'], keep="first", inplace=True)
frequency = uk_copy.groupby(by='CustomerID', as_index=False)['InvoiceNo'].count()
frequency.columns = ['CustomerID','Frequency']
frequency.head()
```



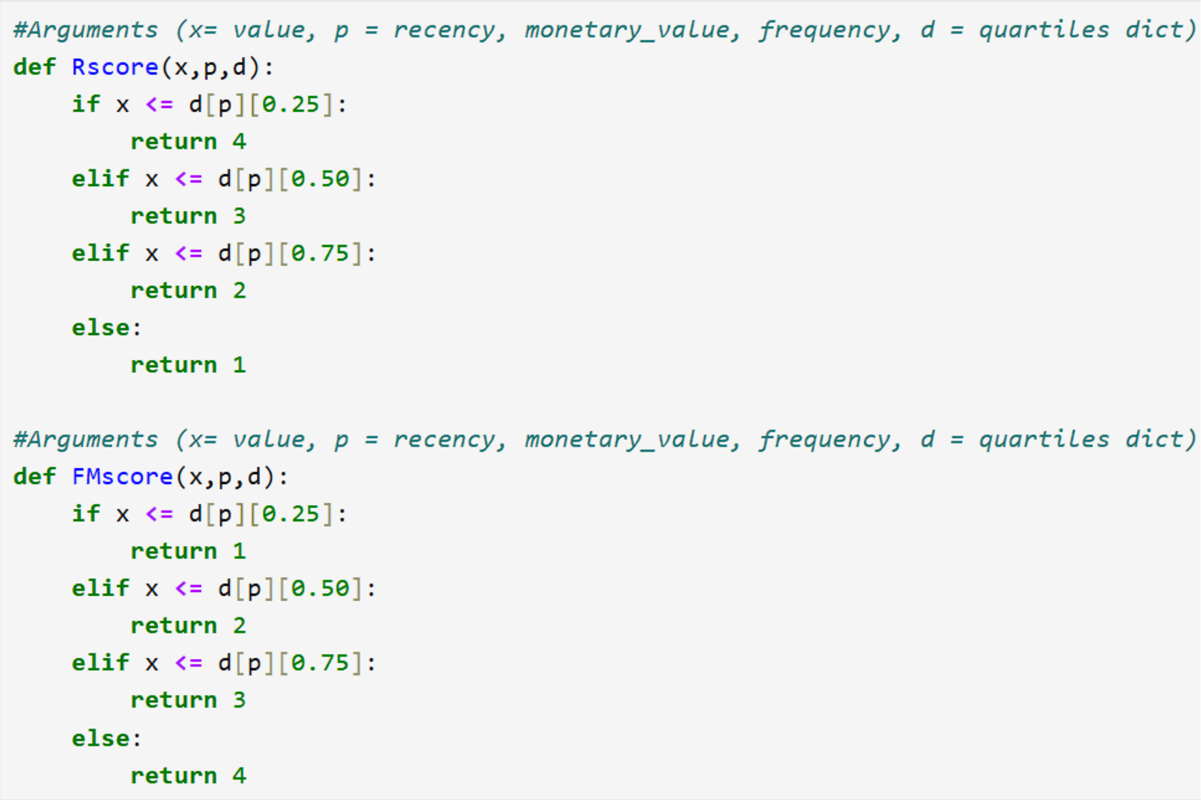
```
monetary = data_uk.groupby(by='CustomerID',as_index=False).agg({'Total_Cost': 'sum'})
monetary.columns = ['CustomerID', 'Monetary']
monetary.head()
```



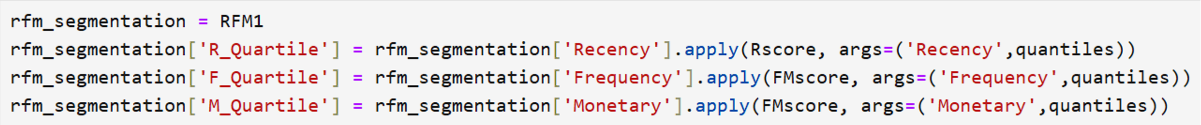
```
#Merge with monetary dataframe
RFM1 = df_merge.merge(monetary,on='CustomerID')
RFM1.set_index('CustomerID',inplace=True)
RFM1.head()
```



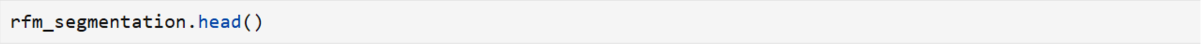
```
quantiles = RFM1.quantile(q=[0.25,0.5,0.75])
quantiles
```



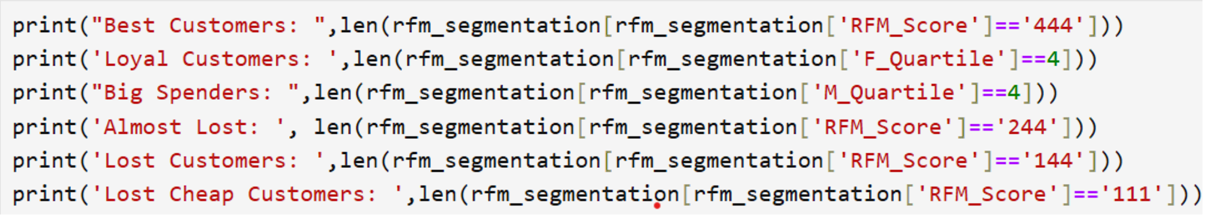
```python
#Arguments (x= value, p = recency, monetary_value, frequency, d = quartiles dict)
def Rscore(x,p,d):
    if x <= d[p][0.25]:
        return 4
    elif x <= d[p][0.50]:
        return 3
    elif x <= d[p][0.75]:
        return 2
    else:
        return 1

#Arguments (x= value, p = recency, monetary_value, frequency, d = quartiles dict)
def FMscore(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
```



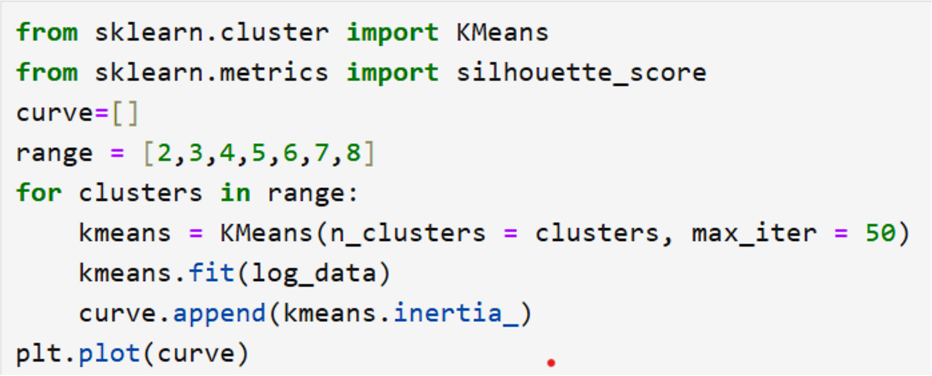
```python
rfm_segmentation = RFM1
rfm_segmentation['R_Quartile'] = rfm_segmentation['Recency'].apply(Rscore, args=('Recency',quantiles))
rfm_segmentation['F_Quartile'] = rfm_segmentation['Frequency'].apply(FMscore, args=('Frequency',quantiles))
rfm_segmentation['M_Quartile'] = rfm_segmentation['Monetary'].apply(FMscore, args=('Monetary',quantiles))
```



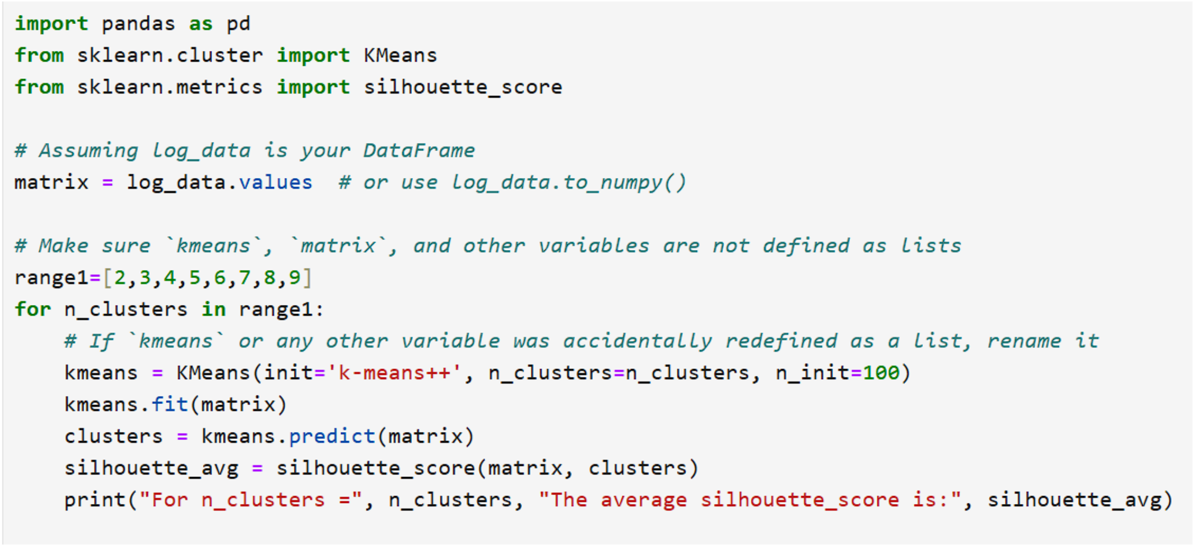
```python
rfm_segmentation.head()
```



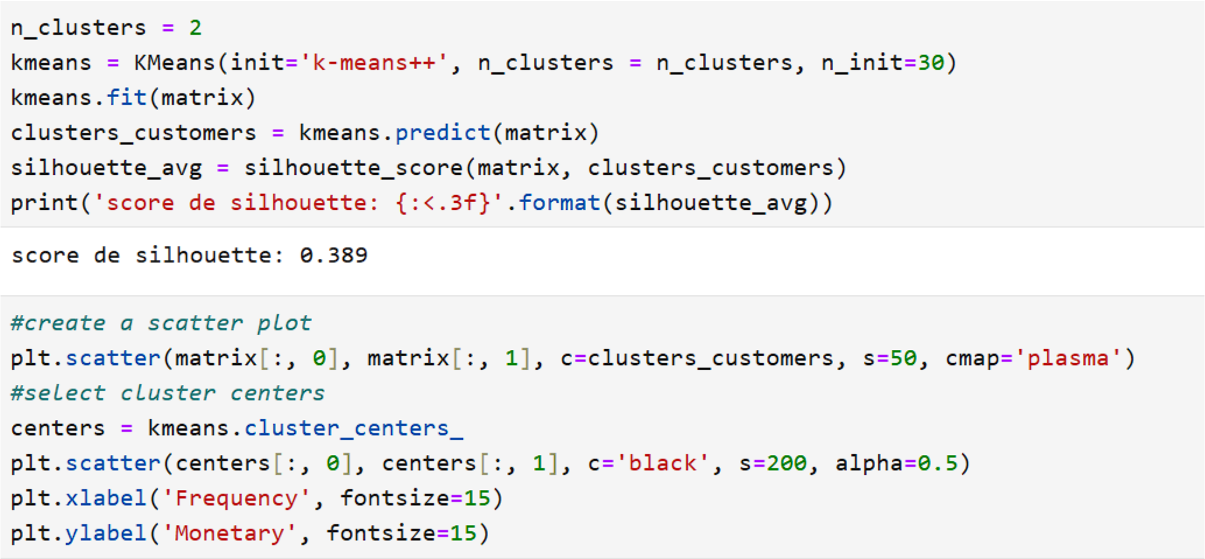
```python
print("Best Customers: ",len(rfm_segmentation[rfm_segmentation['RFM_Score']=='444']))
print('Loyal Customers: ',len(rfm_segmentation[rfm_segmentation['F_Quartile']==4]))
print("Big Spenders: ",len(rfm_segmentation[rfm_segmentation['M_Quartile']==4]))
print('Almost Lost: ', len(rfm_segmentation[rfm_segmentation['RFM_Score']=='244']))
print('Lost Customers: ',len(rfm_segmentation[rfm_segmentation['RFM_Score']=='144']))
print('Lost Cheap Customers: ',len(rfm_segmentation[rfm_segmentation['RFM_Score']=='111']))
```



```python
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
curve=[]
range = [2,3,4,5,6,7,8]
for clusters in range:
    kmeans = KMeans(n_clusters = clusters, max_iter = 50)
    kmeans.fit(log_data)
    curve.append(kmeans.inertia_)
plt.plot(curve)
```



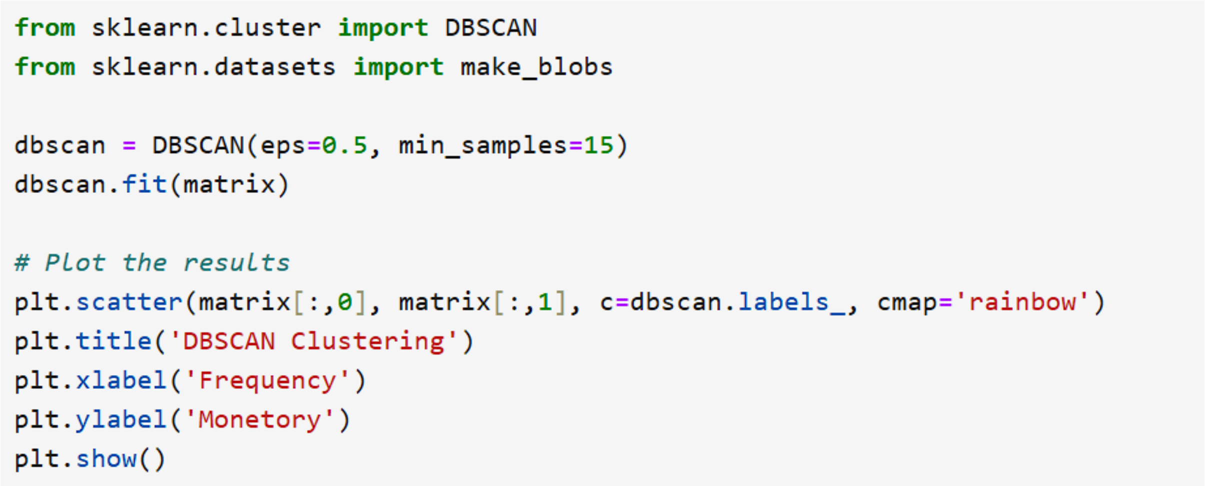
```python
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# Assuming log_data is your DataFrame
matrix = log_data.values  # or use log_data.to_numpy()

# Make sure `kmeans`, `matrix`, and other variables are not defined as lists
range1=[2,3,4,5,6,7,8,9]
for n_clusters in range1:
    # If `kmeans` or any other variable was accidentally redefined as a list, rename it
    kmeans = KMeans(init='k-means++', n_clusters=n_clusters, n_init=100)
    kmeans.fit(matrix)
    clusters = kmeans.predict(matrix)
    silhouette_avg = silhouette_score(matrix, clusters)
    print("For n_clusters =", n_clusters, "The average silhouette_score is:", silhouette_avg)
```



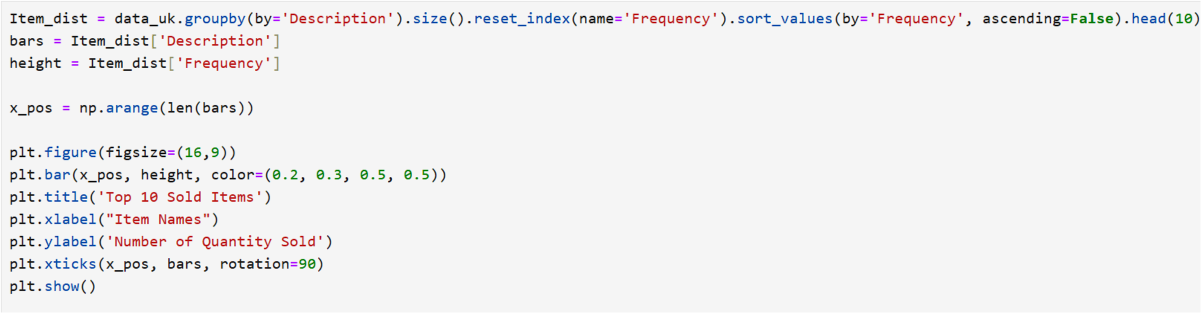
```python
n_clusters = 2
kmeans = KMeans(init='k-means++', n_clusters = n_clusters, n_init=30)
kmeans.fit(matrix)
clusters_customers = kmeans.predict(matrix)
silhouette_avg = silhouette_score(matrix, clusters_customers)
print('score de silhouette: {:<.3f}'.format(silhouette_avg))
```

```
score de silhouette: 0.389
```

```python
#create a scatter plot
plt.scatter(matrix[:, 0], matrix[:, 1], c=clusters_customers, s=50, cmap='plasma')
#select cluster centers
centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200, alpha=0.5)
plt.xlabel('Frequency', fontsize=15)
plt.ylabel('Monetary', fontsize=15)
```

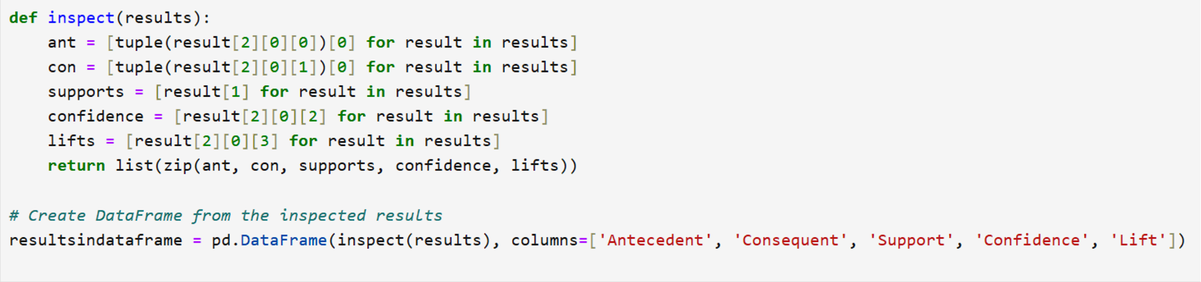


```python
from sklearn.cluster import DBSCAN
from sklearn.datasets import make_blobs

dbscan = DBSCAN(eps=0.5, min_samples=15)
dbscan.fit(matrix)

# Plot the results
plt.scatter(matrix[:,0], matrix[:,1], c=dbscan.labels_, cmap='rainbow')
plt.title('DBSCAN Clustering')
plt.xlabel('Frequency')
plt.ylabel('Monetory')
plt.show()
```

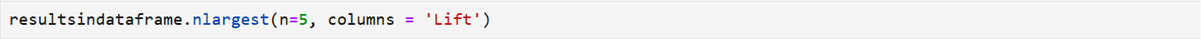


```
Item_dist = data_uk.groupby(by='Description').size().reset_index(name='Frequency').sort_values(by='Frequency', ascending=False).head(10)
bars = Item_dist['Description']
height = Item_dist['Frequency']

x_pos = np.arange(len(bars))

plt.figure(figsize=(16,9))
plt.bar(x_pos, height, color=(0.2, 0.3, 0.5, 0.5))
plt.title('Top 10 Sold Items')
plt.xlabel("Item Names")
plt.ylabel('Number of Quantity Sold')
plt.xticks(x_pos, bars, rotation=90)
plt.show()
```

```
Item_dist = data_uk.groupby(by='Description').size().reset_index(name='Frequency').sort_values(by='Frequency', ascending=False).head(10)
bars = Item_dist['Description']
height = Item_dist['Frequency']

x_pos = np.arange(len(bars))

plt.figure(figsize=(16,9))
plt.bar(x_pos, height, color=(0.2, 0.3, 0.5, 0.5))
plt.title('Top 10 Sold Items')
plt.xlabel("Item Names")
plt.ylabel('Number of Quantity Sold')
plt.xticks(x_pos, bars, rotation=90)
plt.show()
```



```
def inspect(results):
    ant = [tuple(result[2][0][0])[0] for result in results]
    con = [tuple(result[2][0][1])[0] for result in results]
    supports = [result[1] for result in results]
    confidence = [result[2][0][2] for result in results]
    lifts = [result[2][0][3] for result in results]
    return list(zip(ant, con, supports, confidence, lifts))

# Create DataFrame from the inspected results
resultsindataframe = pd.DataFrame(inspect(results), columns=['Antecedent', 'Consequent', 'Support', 'Confidence', 'Lift'])
```



```
resultsindataframe.nlargest(n=5, columns = 'Lift')
```

Proforma
Undertaking from the PG student while submitting his/her final dissertation to his respective institute

Ref. No. 23070243019

I , the following student

Sr. No.	**Sequence of students names on a dissertation**	**Students name**	**Name of the Institute & Place**	**Email & Mobile**
1.	First Author	Niraj S. Gunjal	SIG	Email: 23070243019@sig.ac.in Mobile: 8788511226

Note: Put additional rows in case of more number of students

hereby give an undertaking that the dissertation **Customer Behaviour Analysis: Identifying High and Low-Value Customers through RFM, Clustering Algorithms, and Market Basket Analysis** been checked for its Similarity Index/Plagiarism through Turnitin software tool; and that the document has been prepared by me and it is my original work and free of any plagiarism. It was found that:

1.	The Similarity Index (SI) was: *(Note: SI range: 0 to 10%; if SI is >10%, then authors cannot communicate ms;* ***attachment of SI report is mandatory)***	9%
2.	The ethical clearance for research work conducted obtained from: *(Note: Name the consent obtaining body; if 'not appliable' then write so)*	NA
3.	The source of funding for research was: *(Note: Name the funding agency; or write 'self' if no funding source is involved)*	Self
4.	Conflict of interest: *(Note: Tick √ whichever is applicable)*	No
5.	The material (adopted text, tables, figures, graphs, etc.) as has been obtained from other sources, has been duly acknowledged in the manuscript: *(Note: Tick √ whichever is applicable)*	Yes

In case if any of the above-furnished information is found false at any point in time, then the University authorities can take action as deemed fit against all of us.

Full Name &
Signature of the student

Name &
Signature of SIU Guide/Mentor

Date: 13 September 2024

Endorsement by
Academic Integrity Committee (AIC)

Place: Pune

Note: It is mandatory that the Similarity Index report of plagiarism (only first page) should be appended to the

Document Viewer

Turnitin Originality Report

Processed on: 12-Aug-2024 23:32 IST
ID: 2431081773
Word Count: 4307
Submitted: 2

Project Draft By Niraj Gunjal

Similarity Index	Similarity by Source	
9%	Internet Sources:	4%
	Publications:	6%
	Student Papers:	3%

exclude quoted | exclude bibliography | exclude small matches | mode: quickview (classic) report | print | download

2% match (A. Joy Christy, A. Umamakeswari, L. Priyatharsini, A. Neyaa. "RFM ranking – An effective approach to customer segmentation", Journal of King Saud University - Computer and Information Sciences, 2018)
A. Joy Christy, A. Umamakeswari, L. Priyatharsini, A. Neyaa. "RFM ranking – An effective approach to customer segmentation", Journal of King Saud University - Computer and Information Sciences, 2018

1% match (Internet from 01-Mar-2022)
https://www.igi-global.com/viewtitle.aspx?TitleId=289346&isxn=9781799858959

1% match (student papers from 08-Aug-2023)
Submitted to Coventry University on 2023-08-08

<1% match (student papers from 28-Feb-2020)
Submitted to Coventry University on 2020-02-28

<1% match (student papers from 06-May-2019)
Submitted to University of East London on 2019-05-06

<1% match (student papers from 24-Jul-2023)
Submitted to University of Sunderland on 2023-07-24

<1% match (Nikhil Patankar, Soham Dixit, Akshay Bhamare, Ashutosh Darpel, Ritik Raina. "Customer Segmentation Using Machine Learning", IOS Press, 2021)
Nikhil Patankar, Soham Dixit, Akshay Bhamare, Ashutosh Darpel, Ritik Raina. "Customer Segmentation Using Machine Learning", IOS Press, 2021

<1% match (Internet from 03-Jun-2024)
https://pureadmin.qub.ac.uk/ws/portalfiles/portal/591036236/Enhanced_hydrogen_storage_efficiency.p

<1% match (publications)
Dennis Tay. "Data Analytics for Discourse Analysis with Python - The Case of Therapy Talk", Routledge, 2024

<1% match (student papers from 15-Sep-2022)
Submitted to Fulbright University Vietnam on 2022-09-15

<1% match (K. Manikandan, Niveditha V. R., Sudha K., Magesh S., Radha Rammohan S.. "Design and Development of Customer Relationship Management Recommendations by Clustering and Profiling of Customers Using RFM", International Journal of e-Collaboration, 2021)
K. Manikandan, Niveditha V. R., Sudha K., Magesh S., Radha Rammohan S.. "Design and Development of Customer Relationship Management Recommendations by Clustering and Profiling of Customers Using RFM", International Journal of e-Collaboration, 2021

<1% match ()