Capstone Project

Online retail customer segmentation

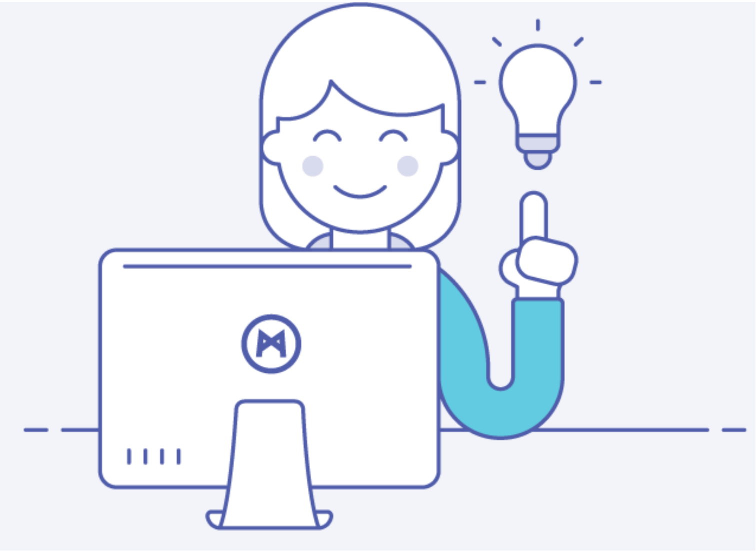
TEAM MEMBER

[NARAYAN SINGH PARMAR , SUMIT GAIKWAD, SAGAR JAIN]

CONTENTS

- PROBLEM DESCRIPTION
- DATA DESCRIPTION
- DATA WRANGLING
- EXPLORATORY DATA ANALYSIS
- K-MEANS CLUSTERING IMPLEMENTATION
- CREATING RECENCY, FREQUENCY AND MONETARY(RFM) MODEL
- CONCLUSION

Problem Description



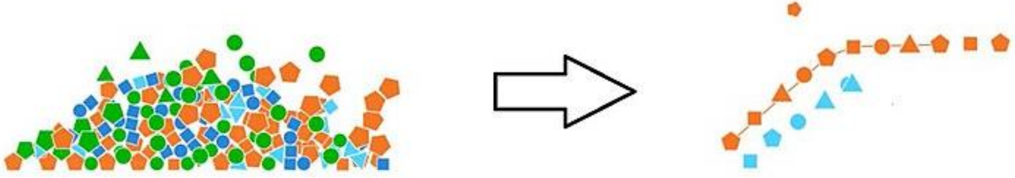
In this project, your task is to identify major customer segments on a transnational dataset which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail .
The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers

Data Description

Attribute Information:

- Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- Stock Code: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- Invoice Date: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- Unit Price: Unit price. Numeric, Product price per unit in sterling.
- Customer ID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

Data Wrangling



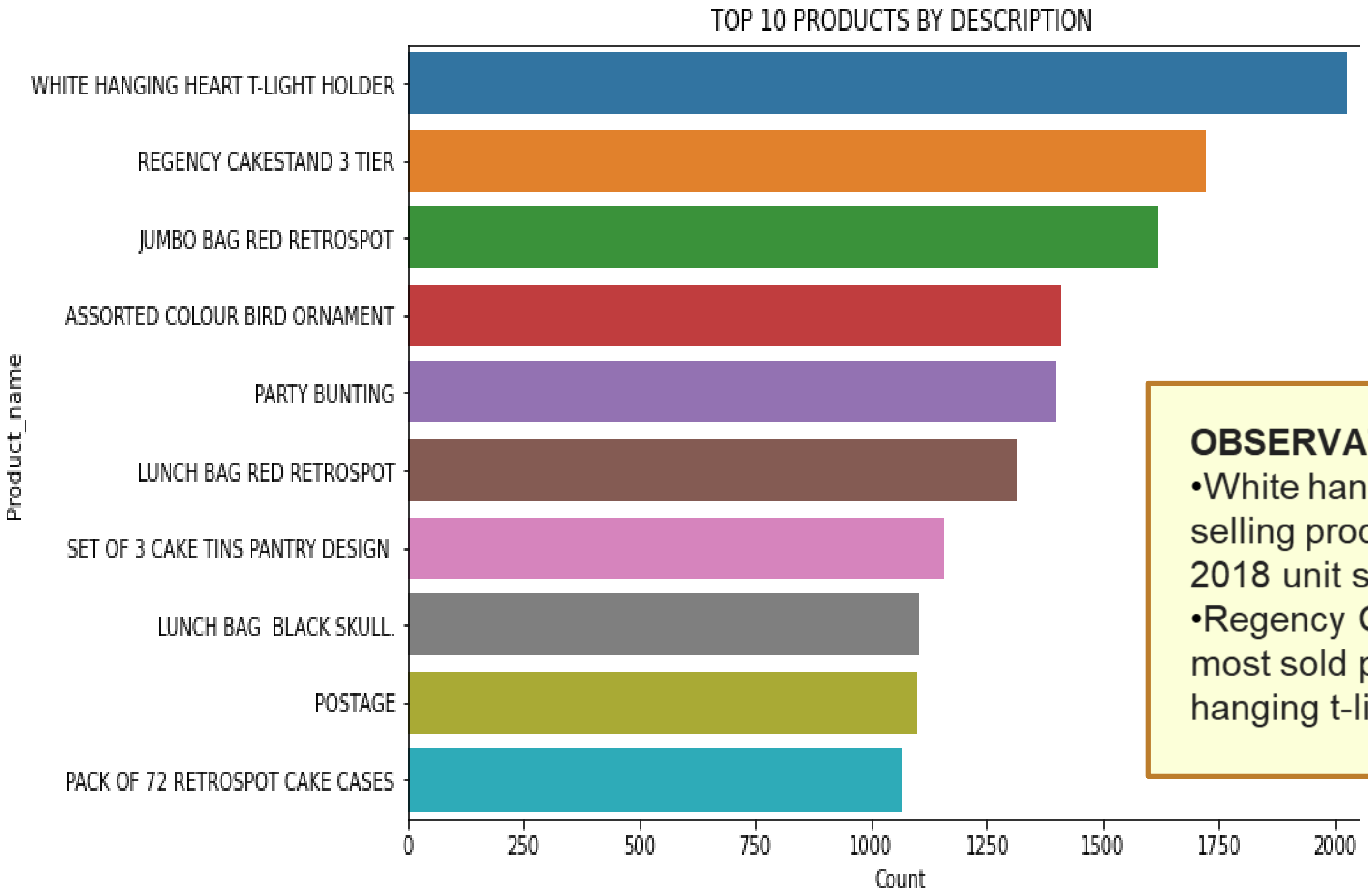
- Dataset is from UK,
- In dataset there are 541909 rows and columns.
- Categorical Features: ‘Invoice No’, ‘Stock Code’, ‘Description’ and ‘Country’.
- There are 1454 null values in “Description’ and 135080 in ‘ Customer ID’.
- There are 5225 duplicates values present in our data.
- One Date time [ns] feature: Invoice Date.
- Outliers present only in ‘Quantity’ and ‘Unit Price’ column.
- Dropped cancelled orders.
- New features from Date time column added such as months, days and hours.

  Total amount is added.
- Data types is converted

EDA

Exploratory Data Analysis

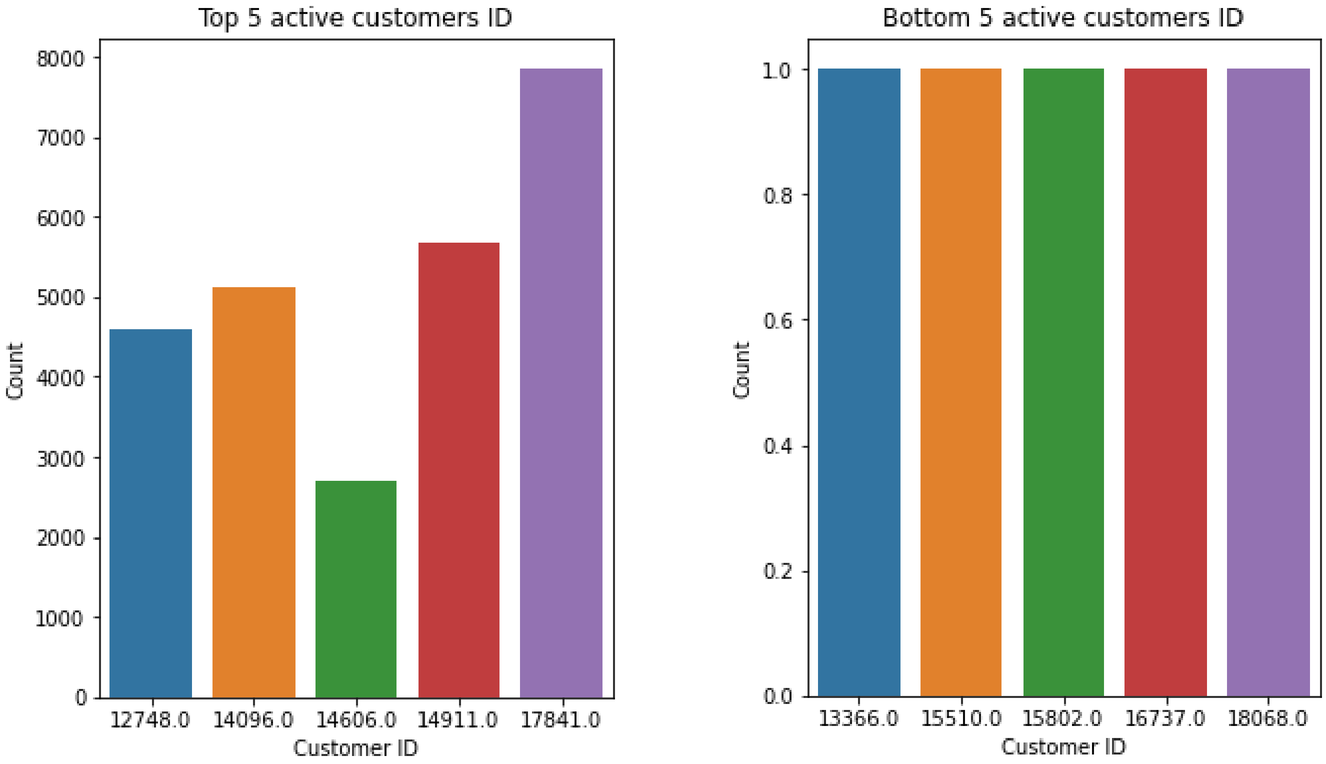
Top 10 products in terms of description



OBSERVATION-

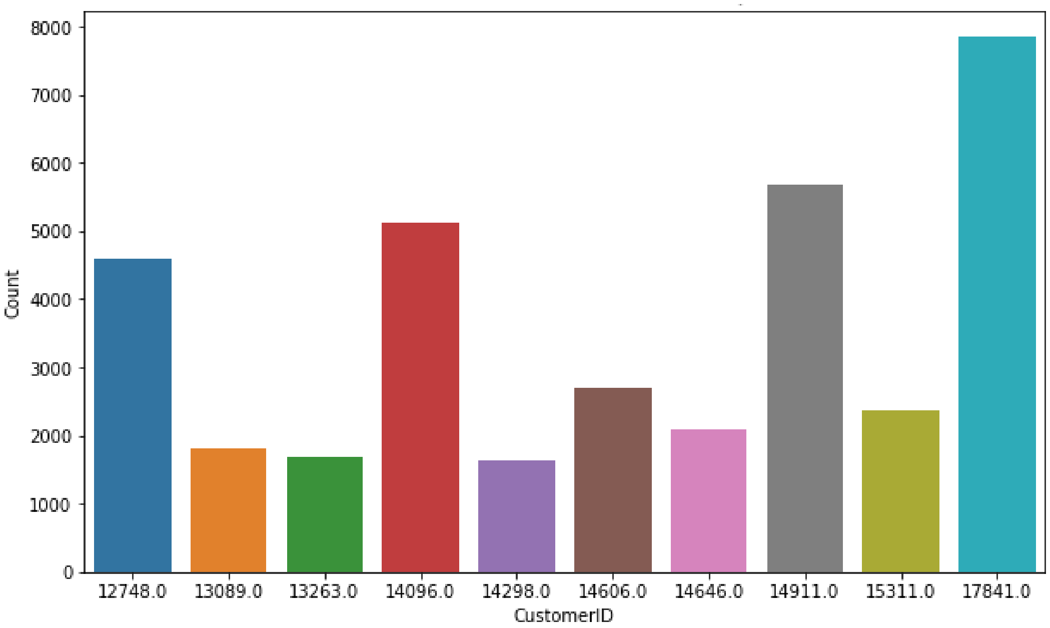
- White hanging T-Light holder is most selling product which has almost 2018 unit sold.
- Regency Cake stand 3 Tier is 2nd most sold product after the white hanging t-light holder.

Top and bottom Customer Id



- ID number 17841 was the most active customer.

Top 10 Customer By Purchasing



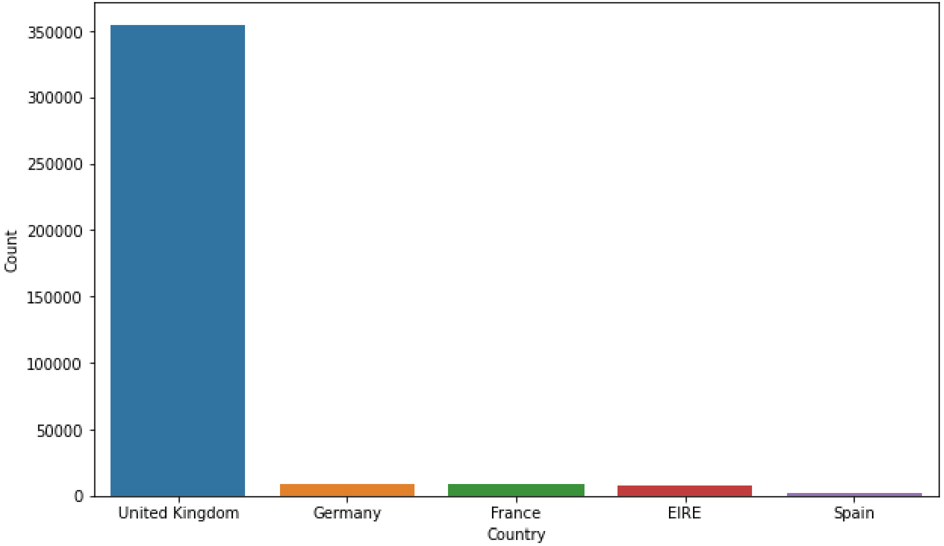
OBSERVATION-

•CustomerID-17841 has bought most products.

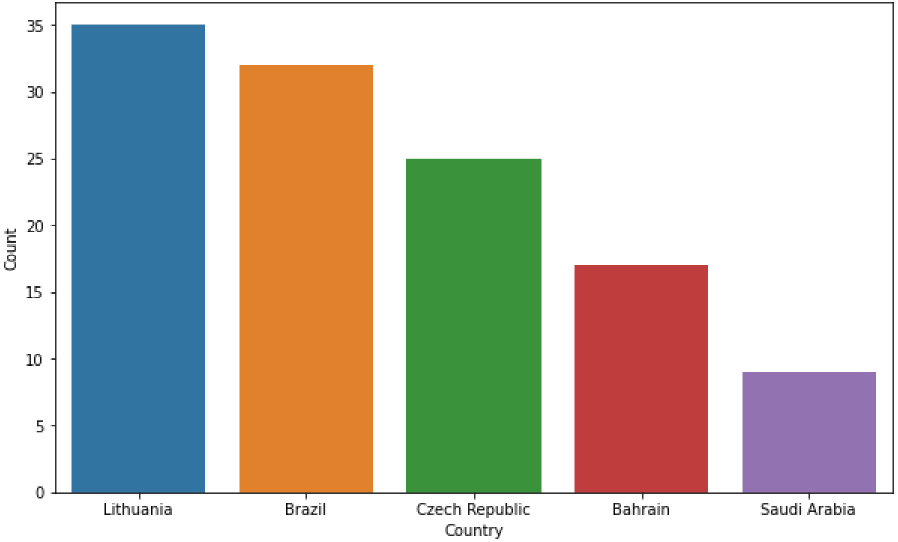
•.CustomerID-14911 has bought 2nd highest products.

Top 5 Highest and least sales Countries

Highest



Least

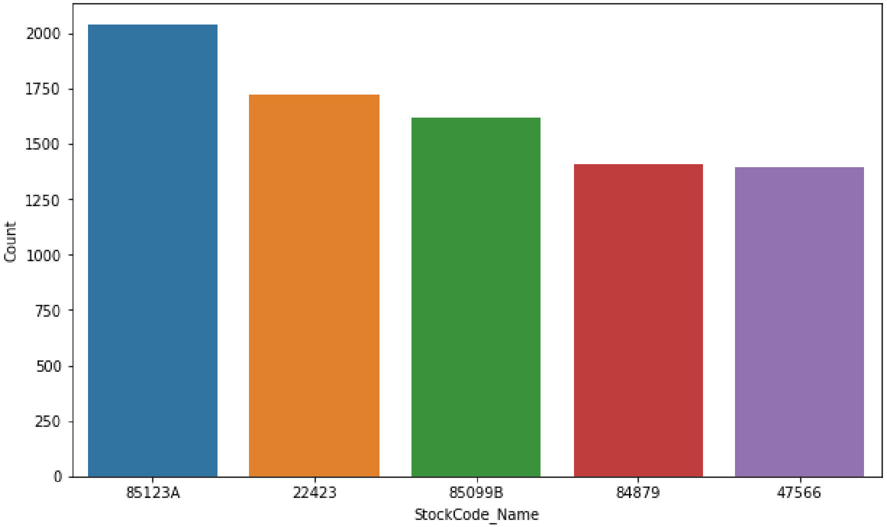


OBSERVATION-

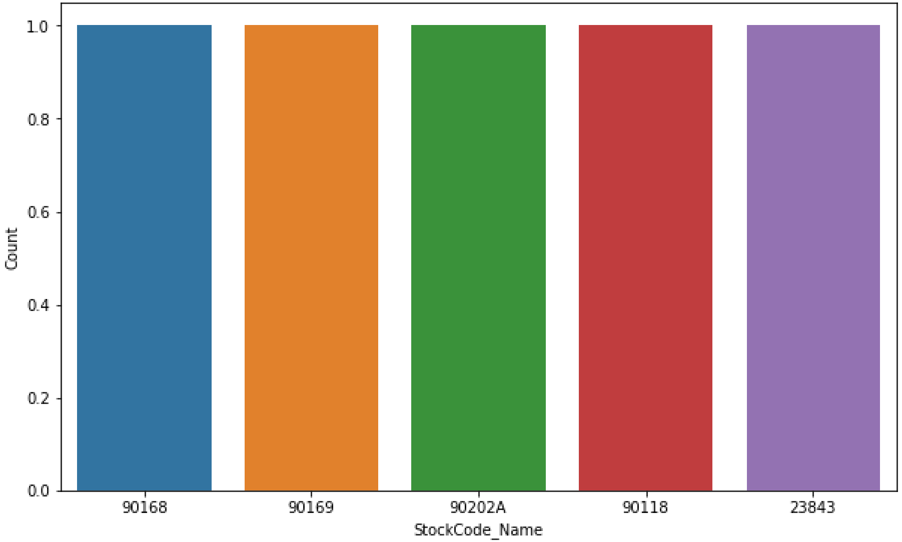
- The country United Kingdom has the highest number of customer, which is preety obvious as per statement out data is mainly UK based data.
- On the other hand Saudi Arabia has least number of customers or low sales.

Top And Bottom 5 Stock code Name

Top



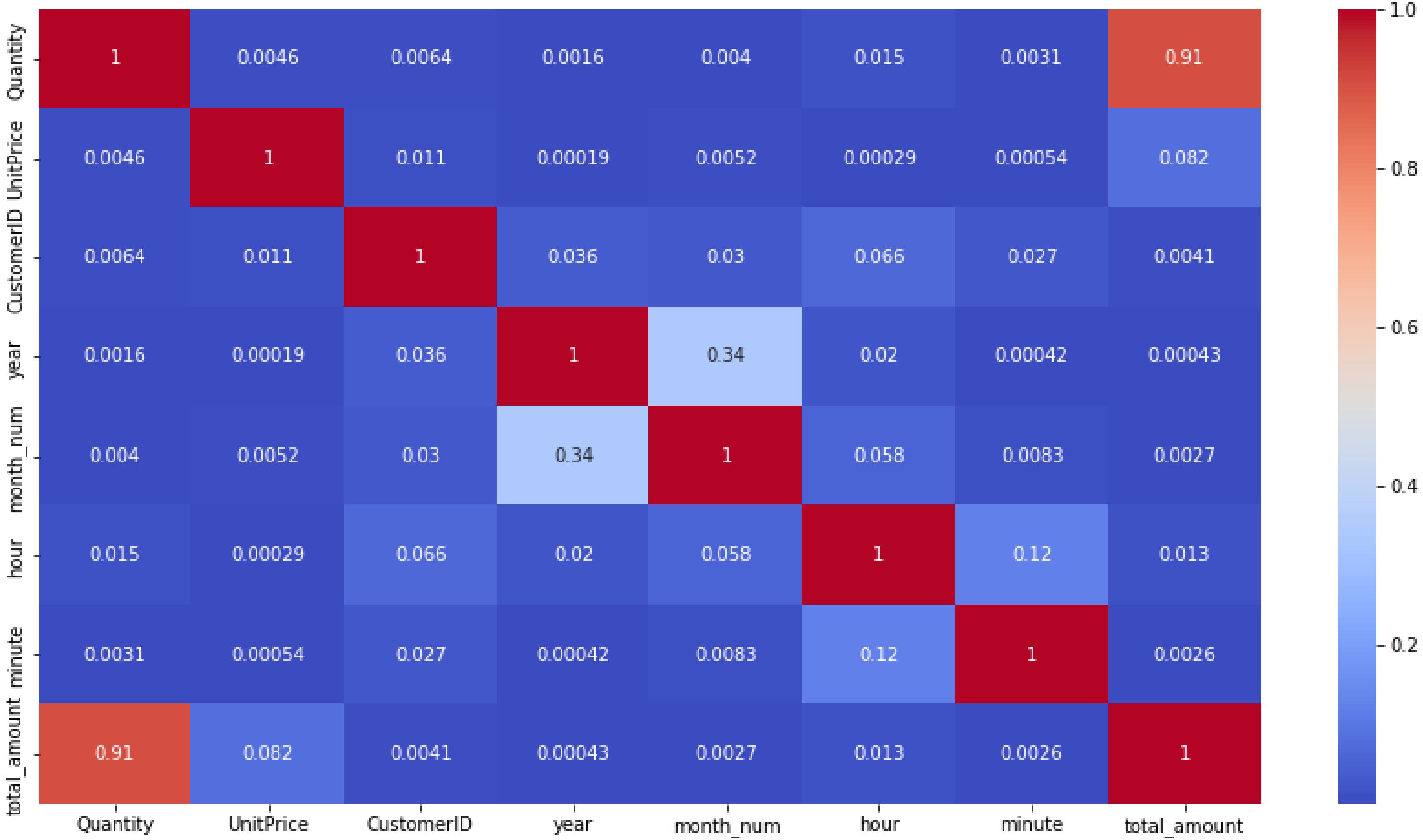
Bottom



The 85123A is top Stock code name

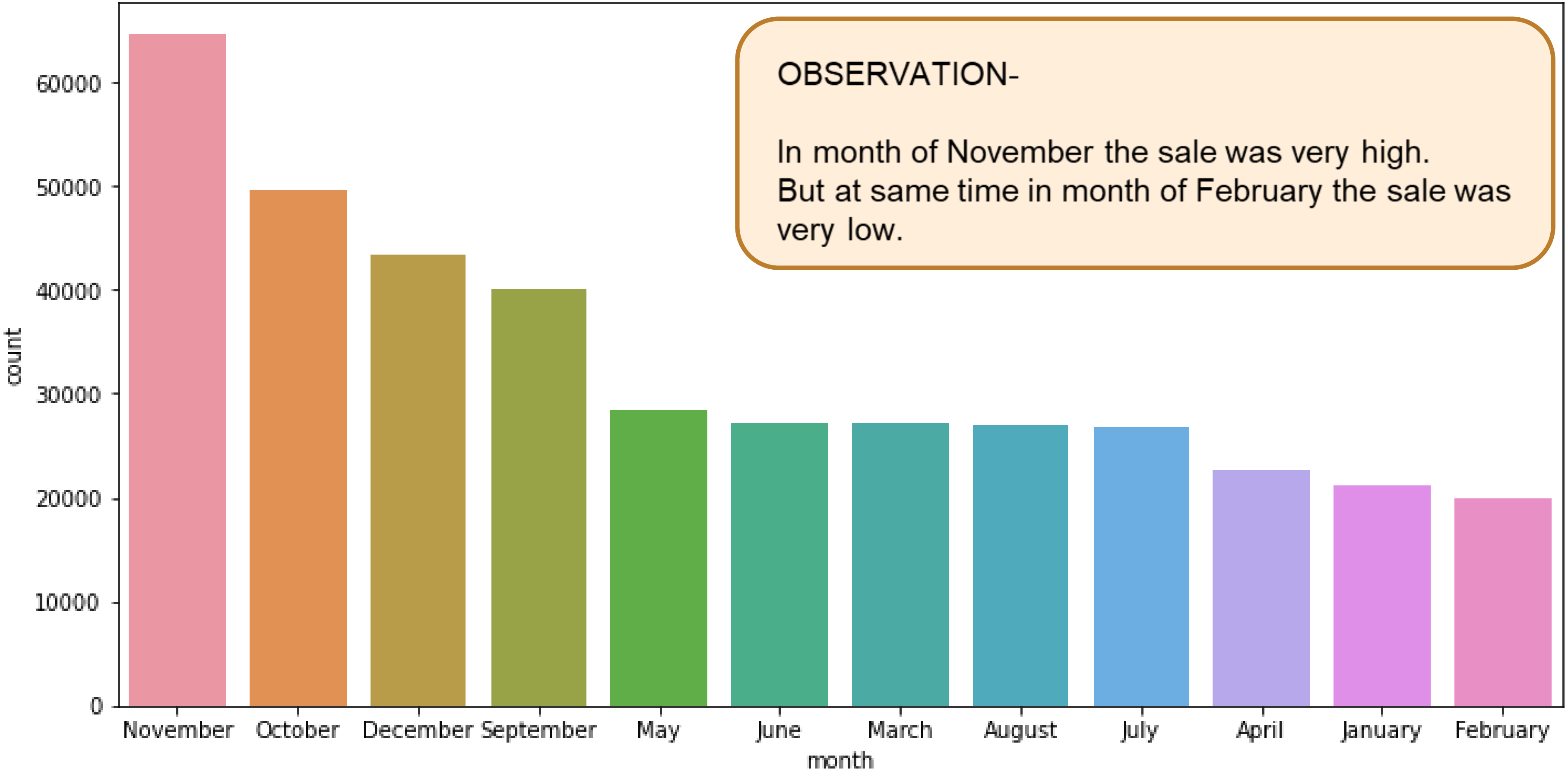
Correlation Matrix

- Heatmap show correlation of different variables



	Quantity	UnitPrice	CustomerID	year	month_num	hour	minute	total_amount
Quantity	1	0.0046	0.0064	0.0016	0.004	0.015	0.0031	0.91
UnitPrice	0.0046	1	0.011	0.00019	0.0052	0.00029	0.00054	0.082
CustomerID	0.0064	0.011	1	0.036	0.03	0.066	0.027	0.0041
year	0.0016	0.00019	0.036	1	0.34	0.02	0.00042	0.00043
month_num	0.004	0.0052	0.03	0.34	1	0.058	0.0083	0.0027
hour	0.015	0.00029	0.066	0.02	0.058	1	0.12	0.013
minute	0.0031	0.00054	0.027	0.00042	0.0083	0.12	1	0.0026
total_amount	0.91	0.082	0.0041	0.00043	0.0027	0.013	0.0026	1

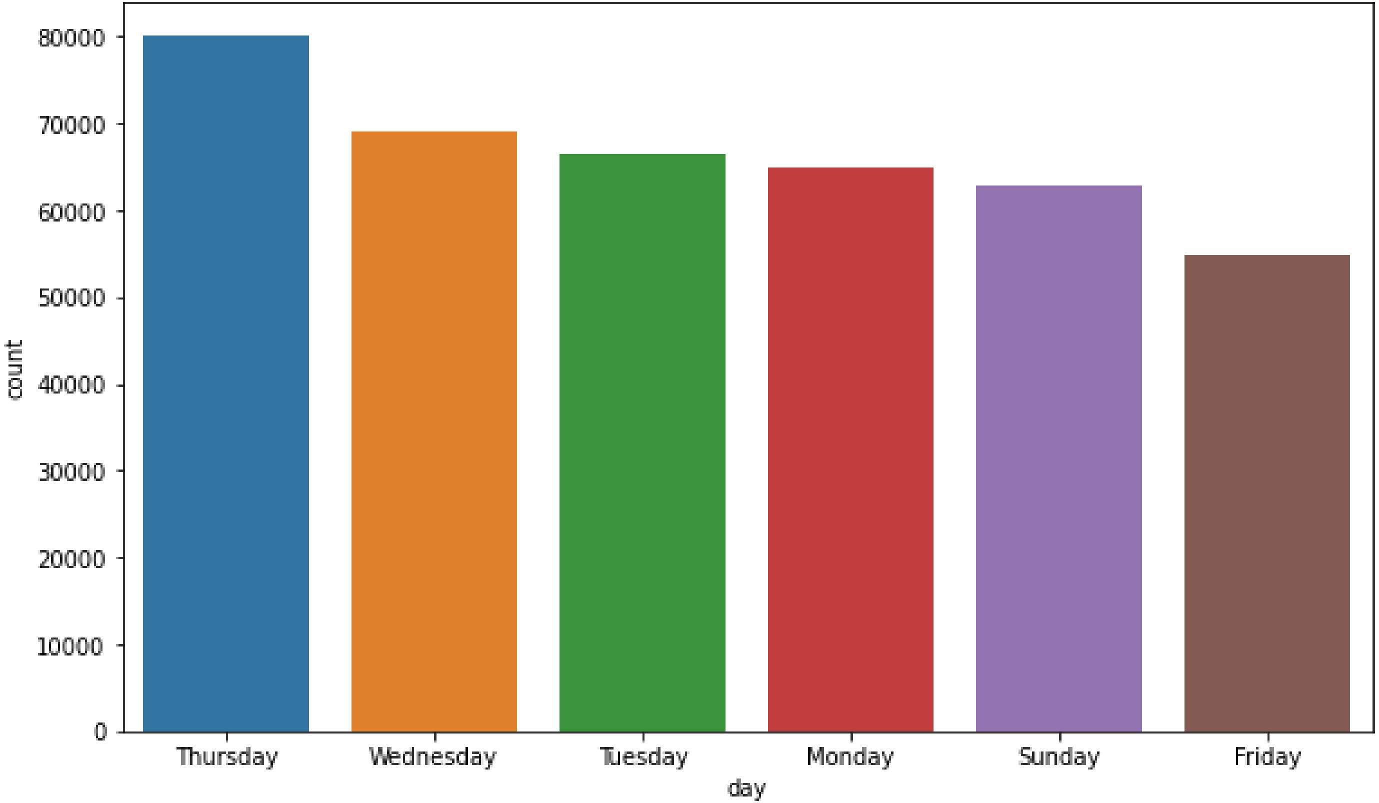
The Sale By Month



OBSERVATION-

In month of November the sale was very high.
But at same time in month of February the sale was very low.

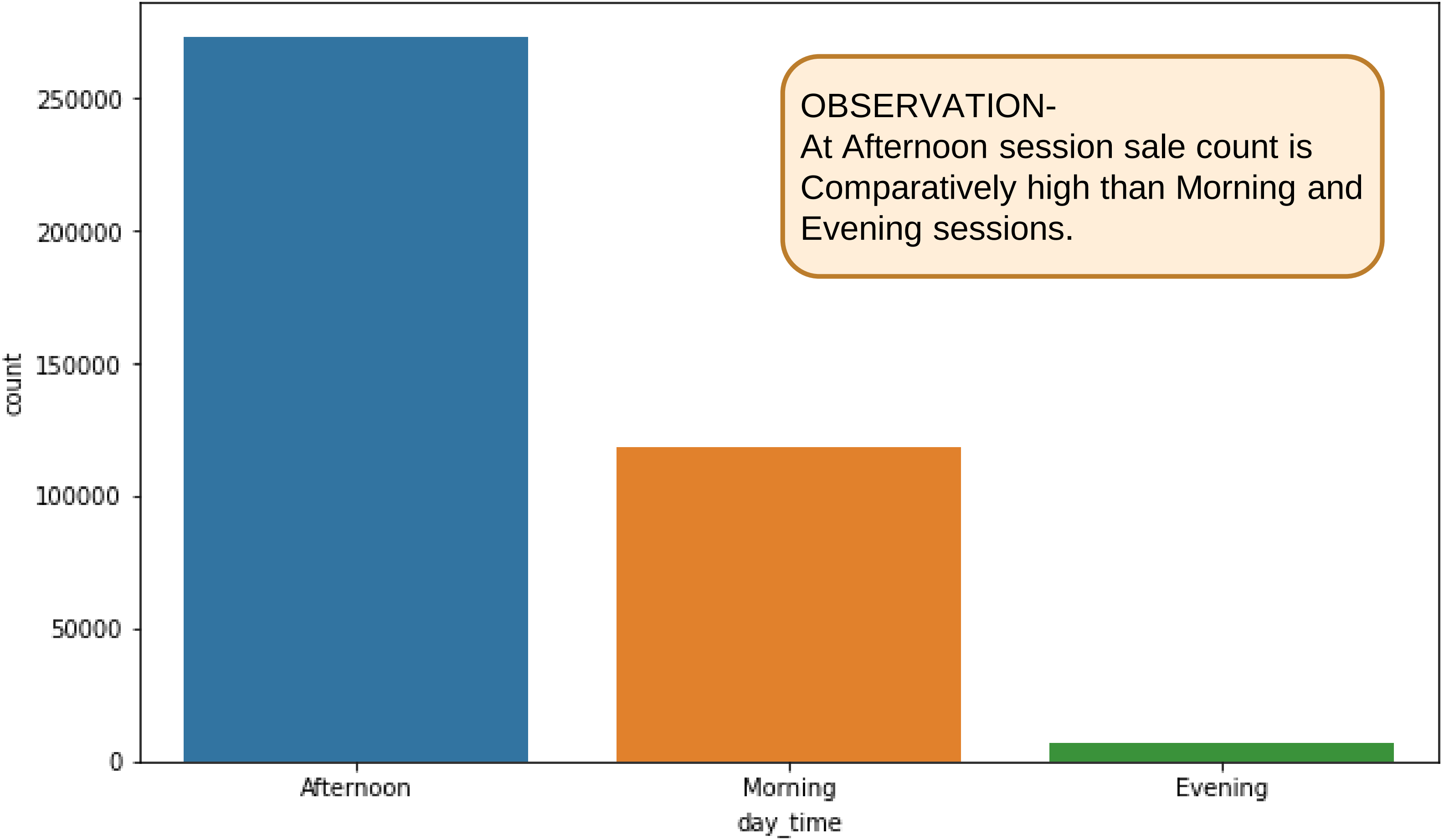
The Sale On Day Basis



OBSERVATION-

•The highest count of sales occurred on the Thursday, and sale at other day are significantly equal amount.

The Count Of Sale On Day Time

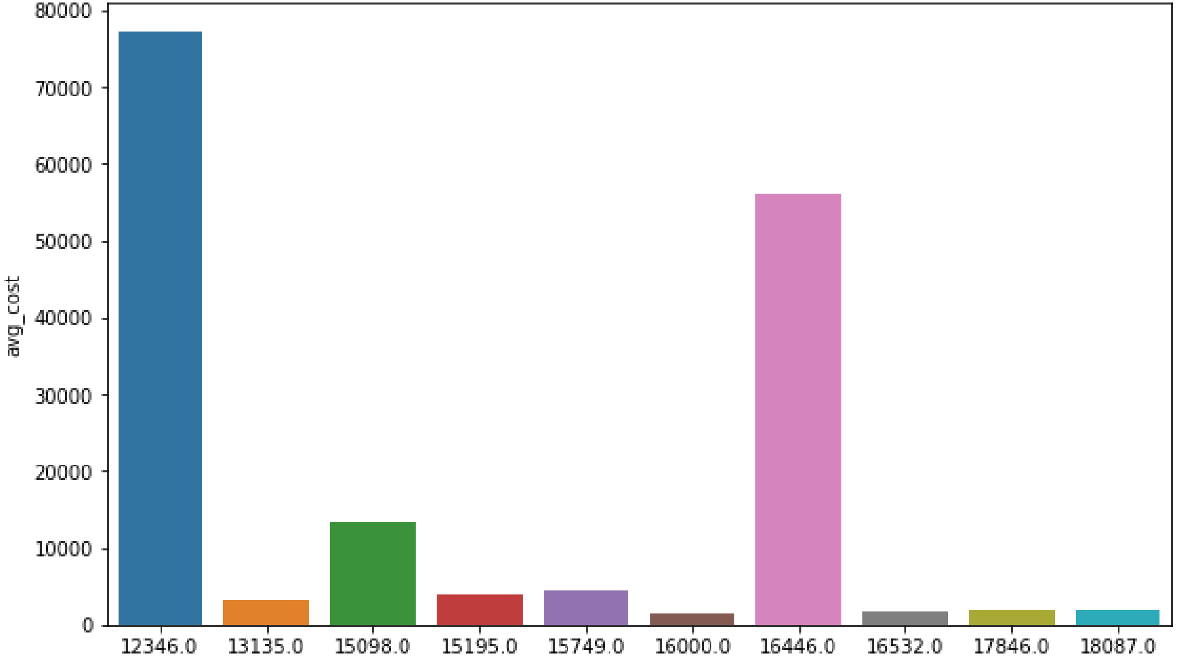


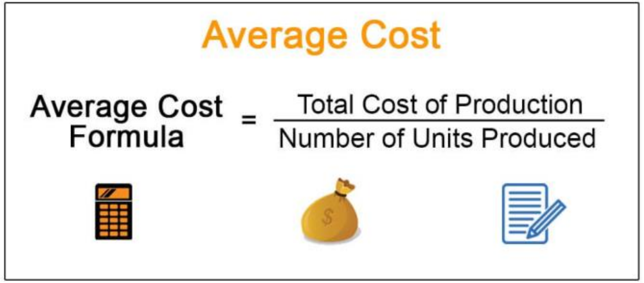
OBSERVATION-
At Afternoon session sale count is Comparatively high than Morning and Evening sessions.

Average Cost Per Customer

OBSERVATION-

- 77183 Is the highest Avg. cost spent by customer 12346.
- 56157 Is the second highest Avg. cost spent by customer 16446.



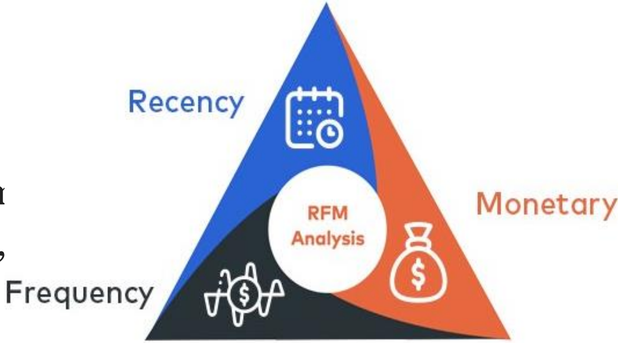


Average Cost

Average Cost Formula = Total Cost of Production / Number of Units Produced

Creating Recency, Frequency and Monetary(RFM) Model

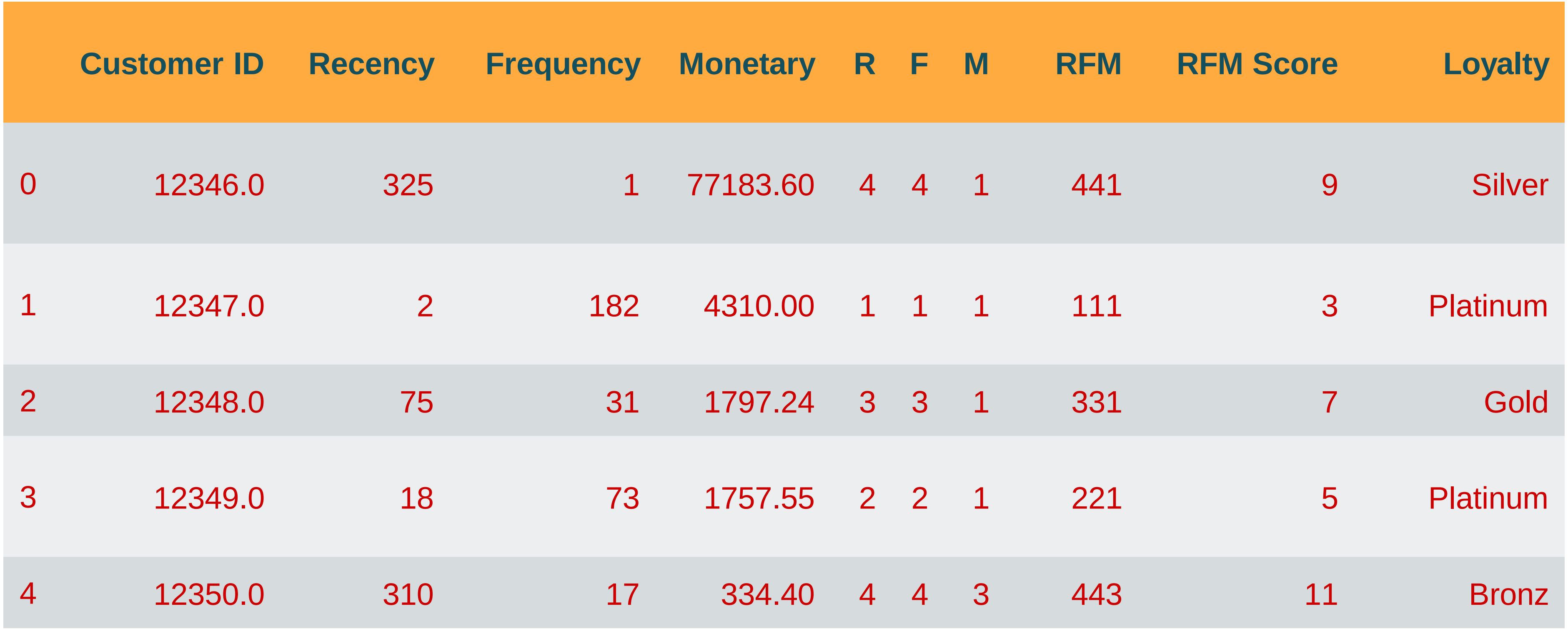
- Before applying any clustering algorithms it is always necessary to determine various quantitative factors on which the algorithm will perforn segmentation. Examples of these would be features such as amount spend, activeness of the customer, their last visit, etc.



- RFM model which stands for Recency, Frequency, and Monetary is one of such steps in which we determine the Recency - days to last visit, frequency - how actively the customer repurchases and monetary - total expenditure of the customer, for each customer. There are other steps too in which we divide each of these features accordingly and calculate a score for each customer. However, this approach does not require machine learning algorithms as segmentation can be done manually. Therefore we will skip the second step and directly use the rfm features and feed it to clustering algorithms.

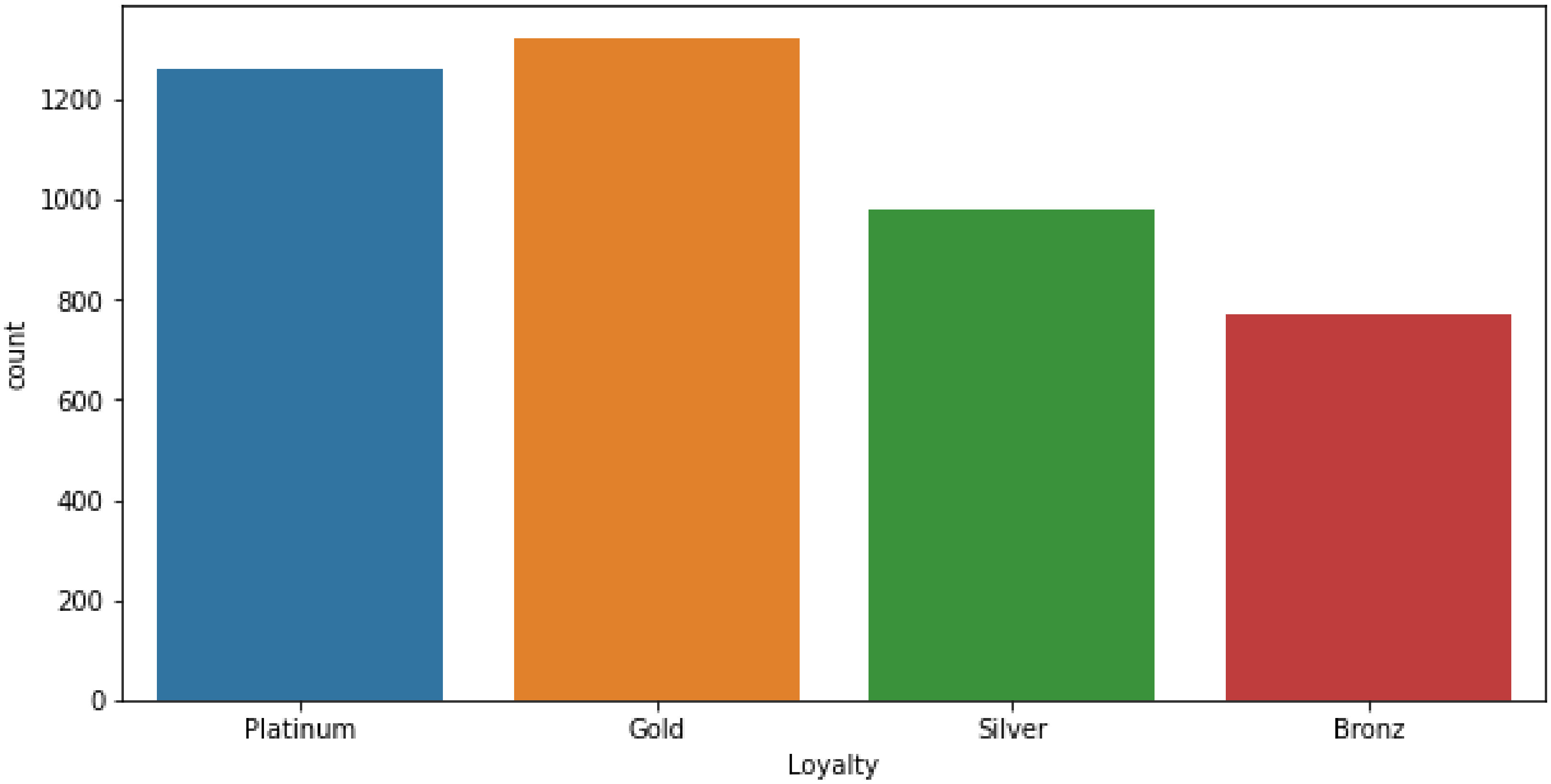
- Recency = Latest Date - Last Inovice Data
- Frequency = count of invoice no. of transaction(s)
- Monetary = Sum of Total Amount for each customer

Loyalty Level To Each Customer

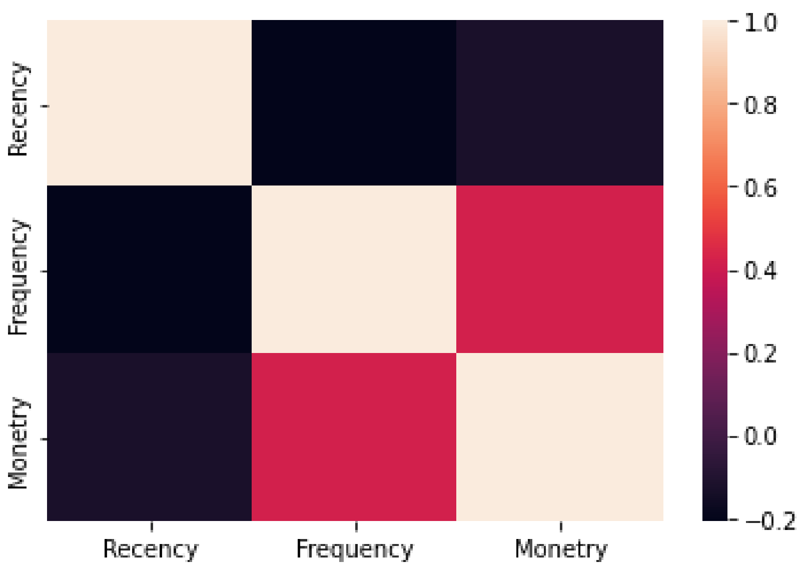


	Customer ID	Recency	Frequency	Monetary	R	F	M	RFM	RFM Score	Loyalty
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver
1	12347.0	2	182	4310.00	1	1	1	111	3	Platinum
2	12348.0	75	31	1797.24	3	3	1	331	7	Gold
3	12349.0	18	73	1757.55	2	2	1	221	5	Platinum
4	12350.0	310	17	334.40	4	4	3	443	11	Bronz

Plot For Loyalty Level Of Customer

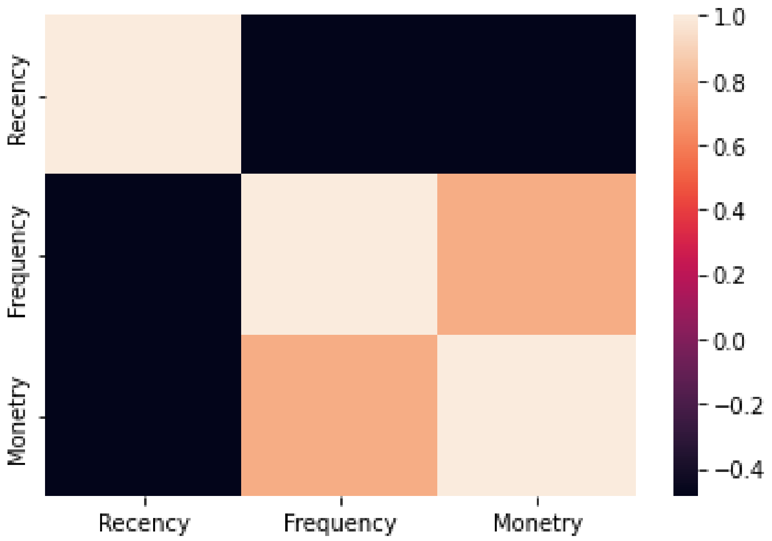


Correlation for RFM And Log RFM



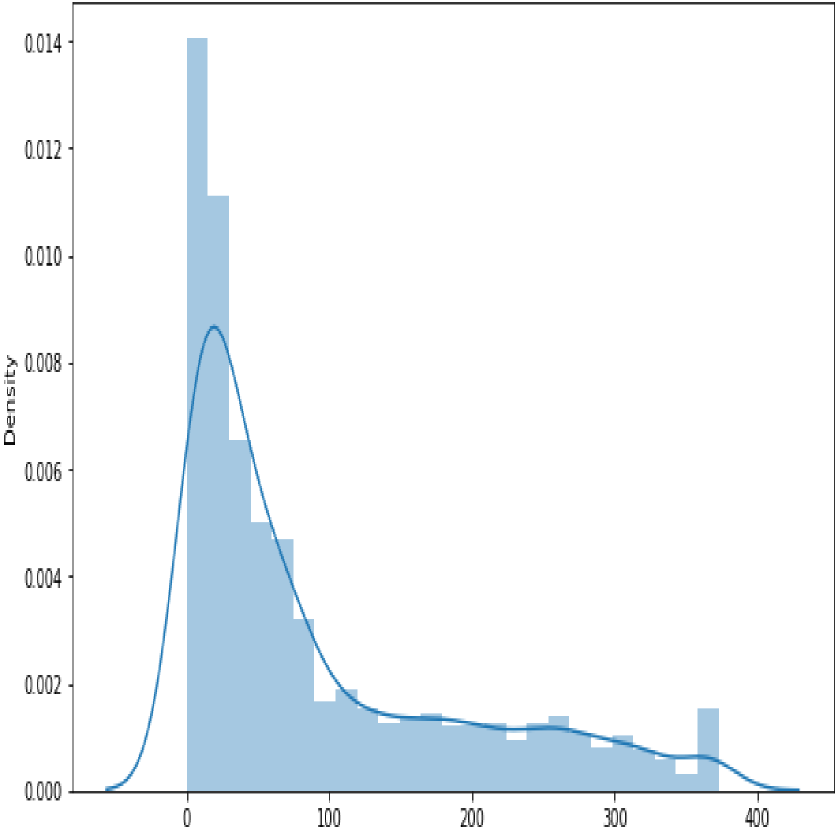
	Recency	Frequency	Monetary
Recency	1.000000	-0.206085	-0.122190
Frequency	-0.206085	1.000000	0.422289
Monetry	-0.122190	0.422289	1.000000

Log Transformation

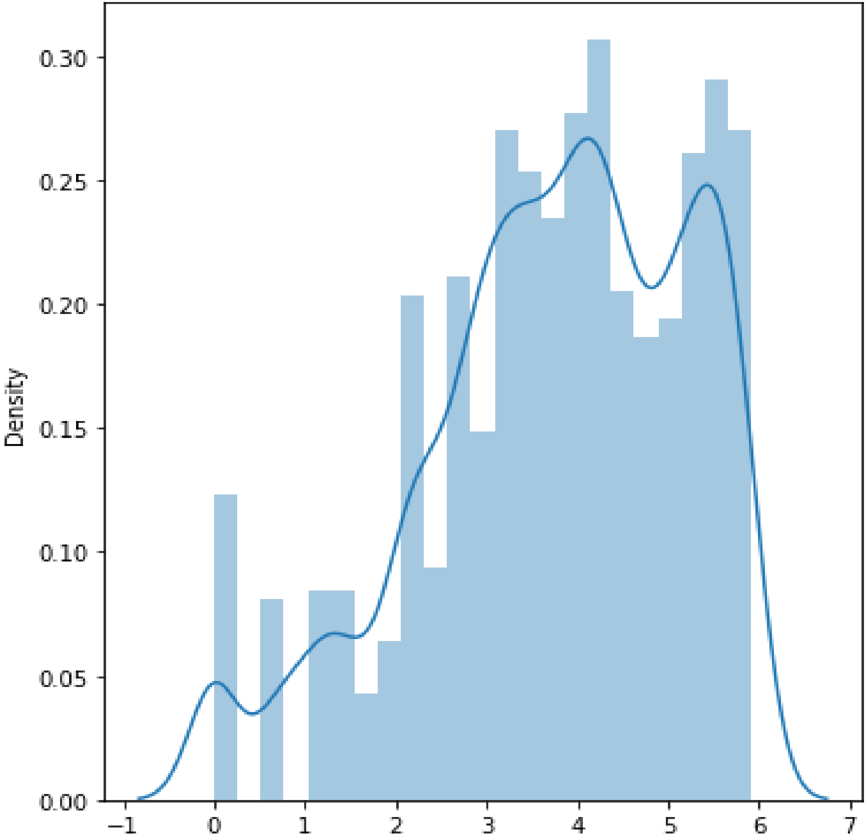


	Recency	Frequency	Monetry
Recency	1.000000	-0.483243	-0.481284
Frequency	-0.483243	1.000000	0.757179
Monetry	-0.481284	0.757179	1.000000

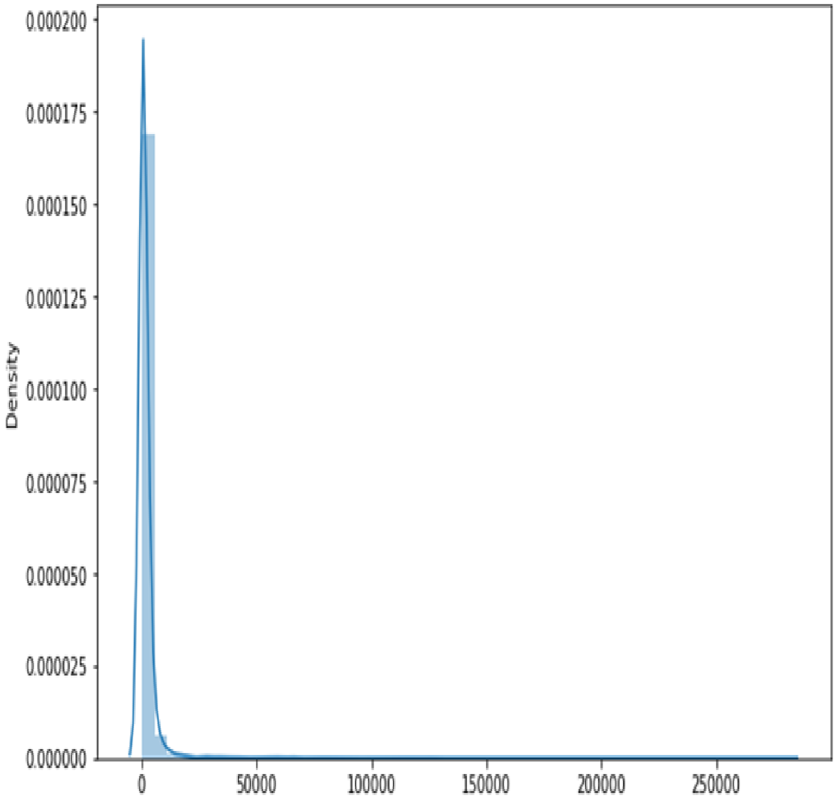
Distribution Plot Recency



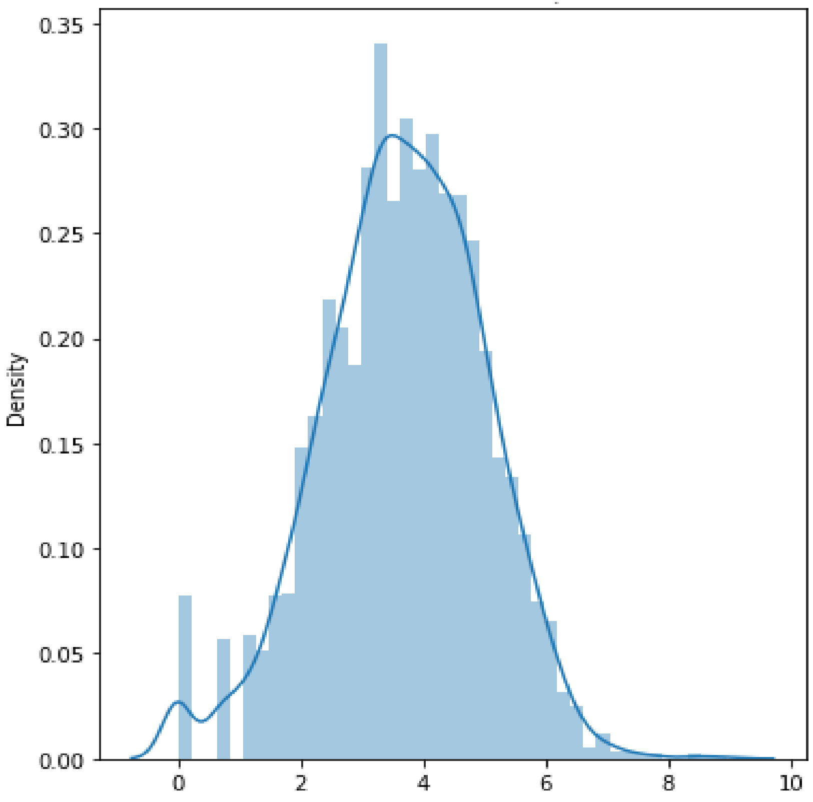
Log Transformation



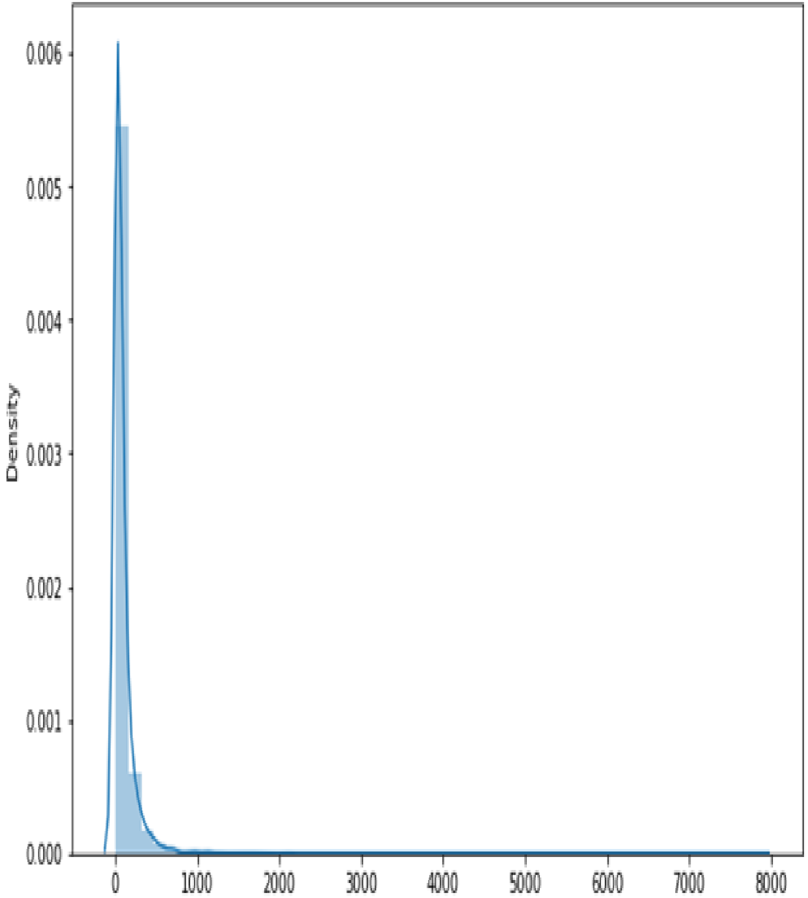
Frequency



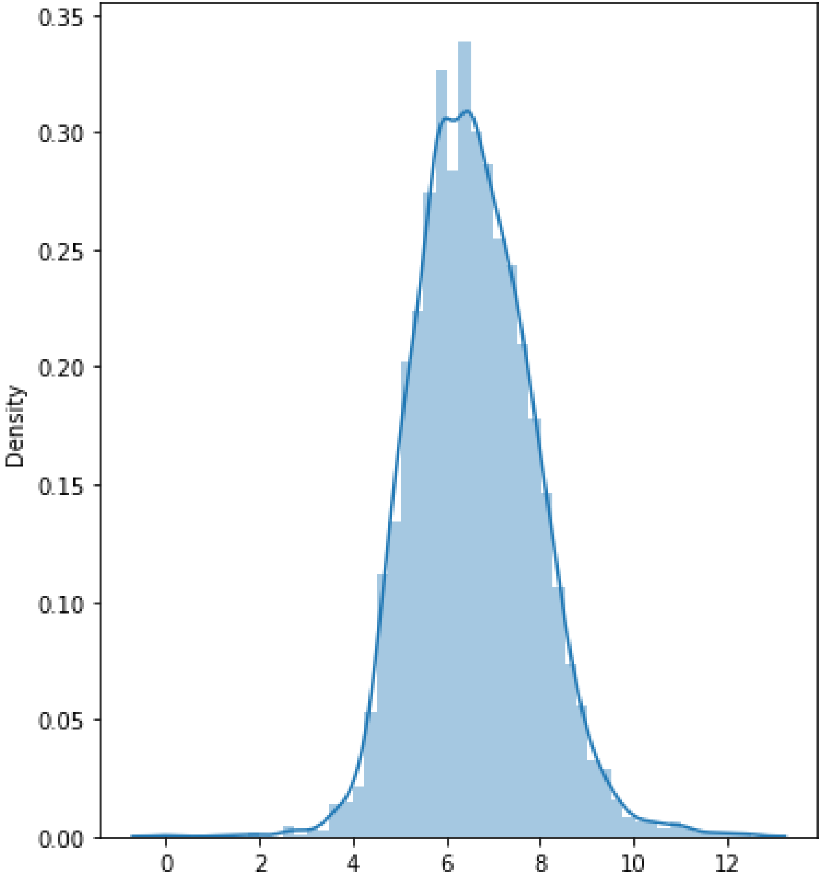
Log Transformation



Monetry



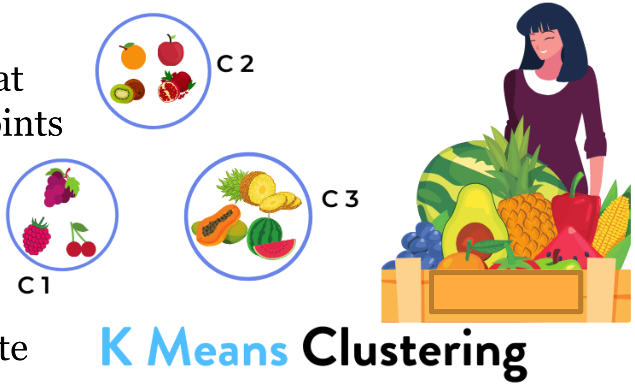
Log Transformation



K-Means Clustering Implementation

- It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. Calculation of Silhouette Score

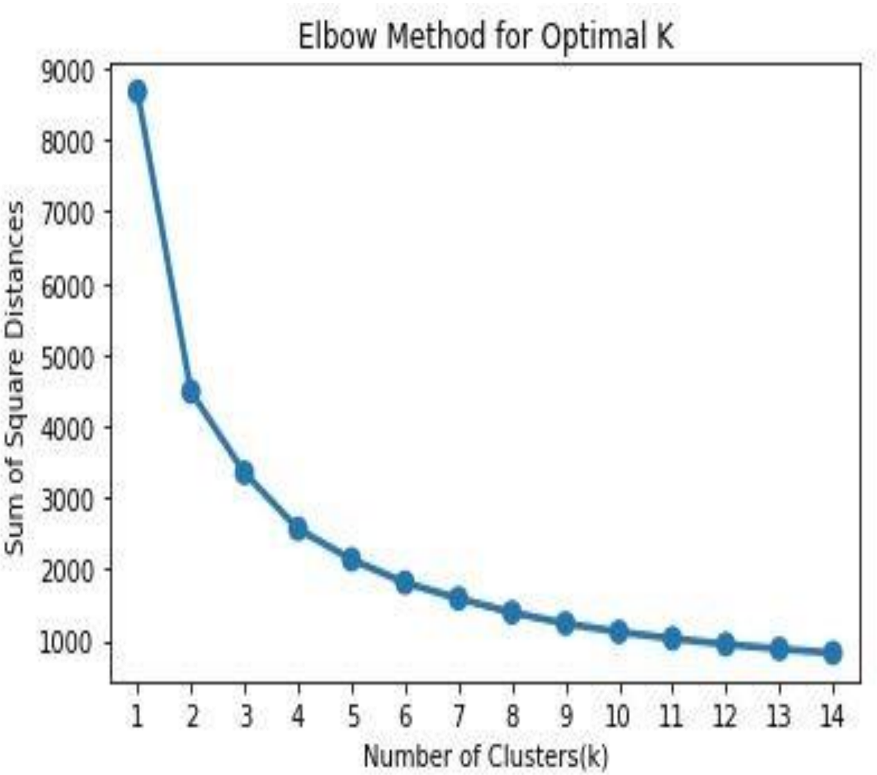
- Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters.

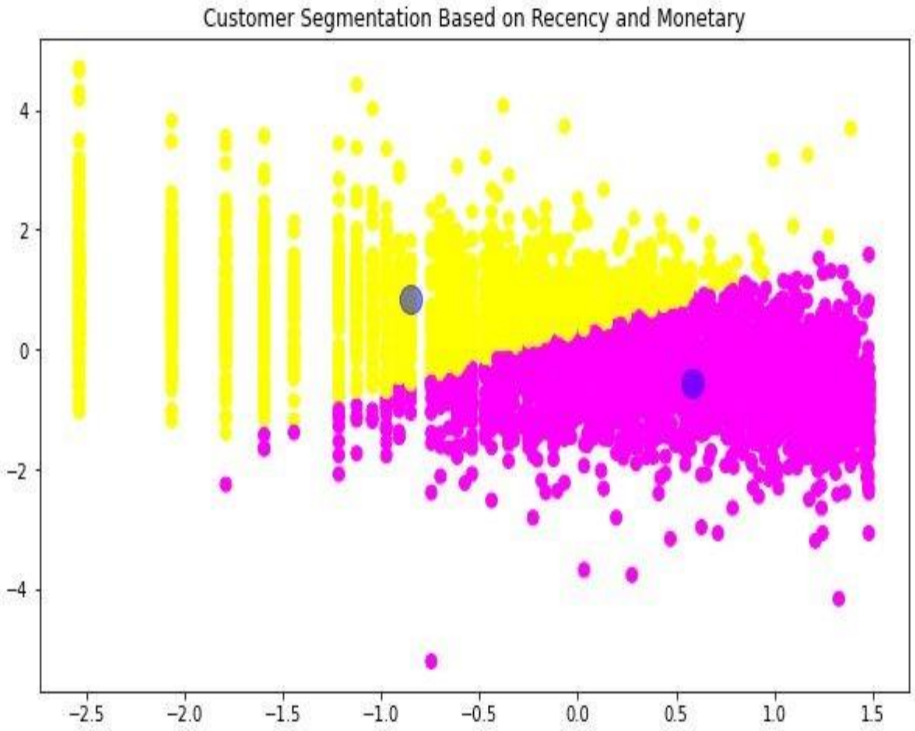


Models Used

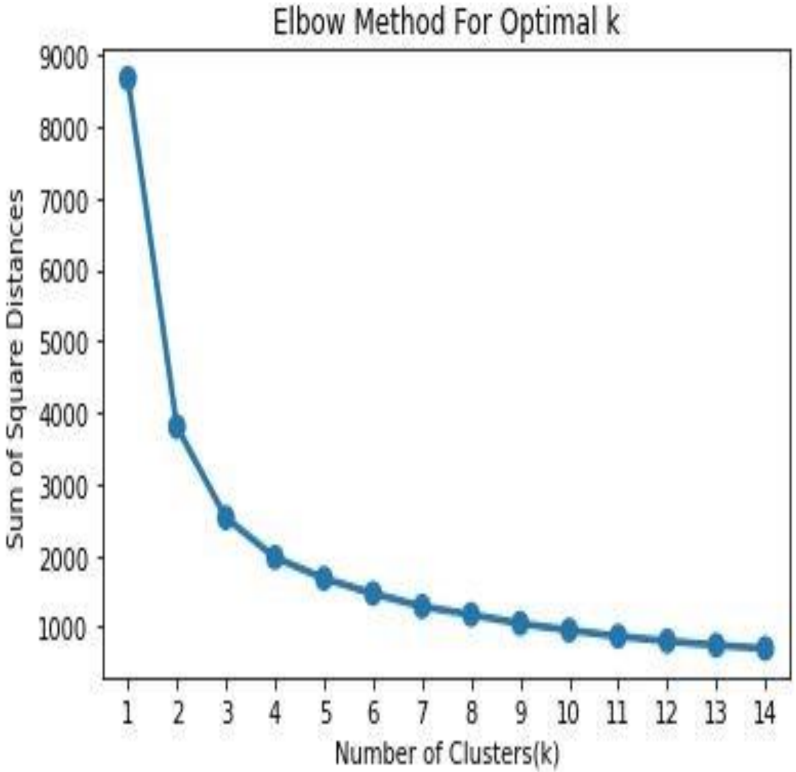
- K-Means Clustering
- DBSCAN
- Hierarchical Clustering (Dendrogram)

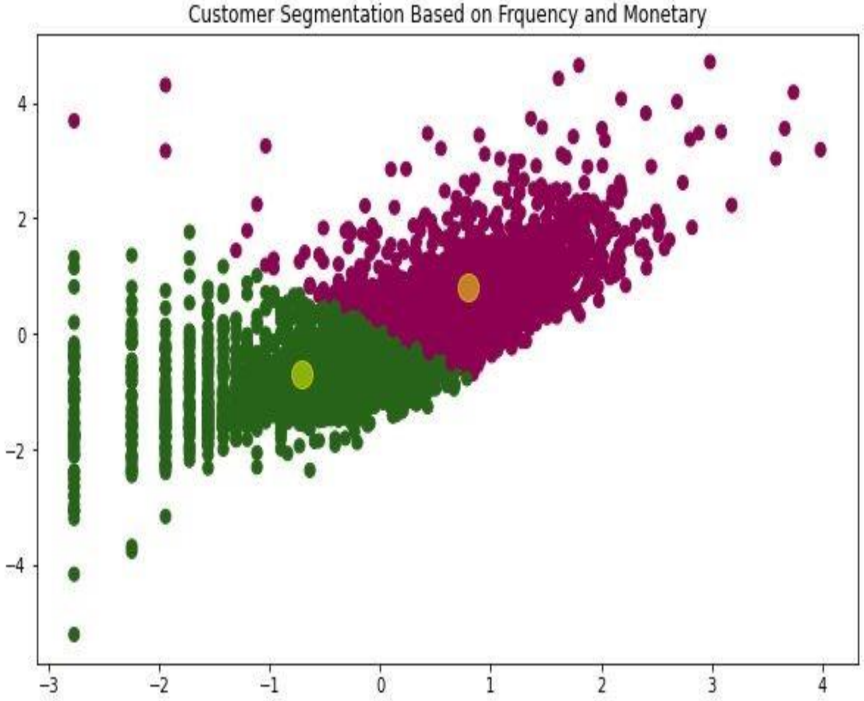
Elbow Method on Recency and Monetary



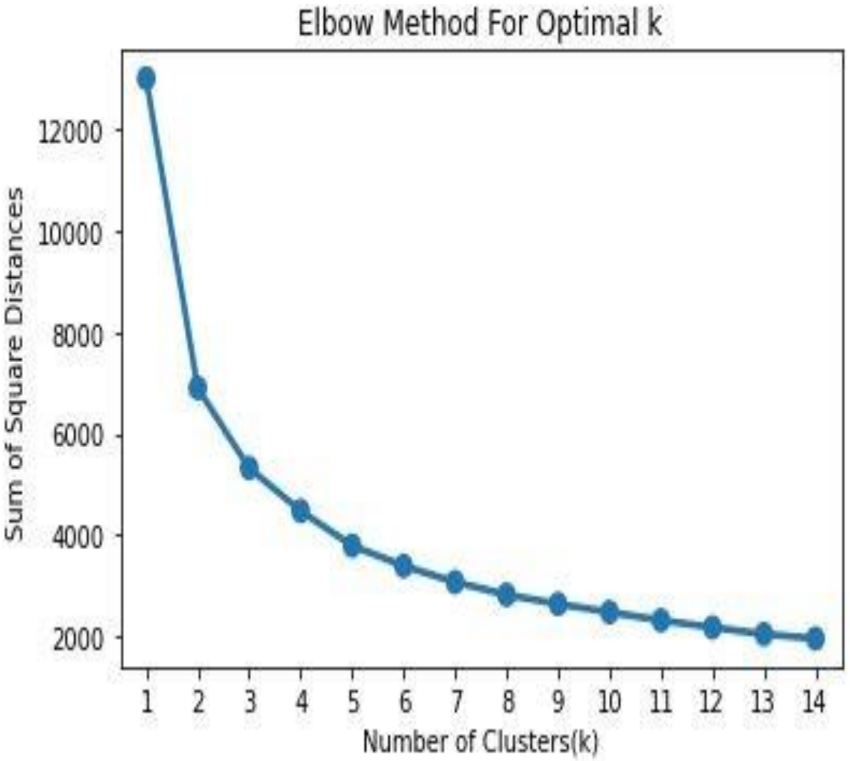


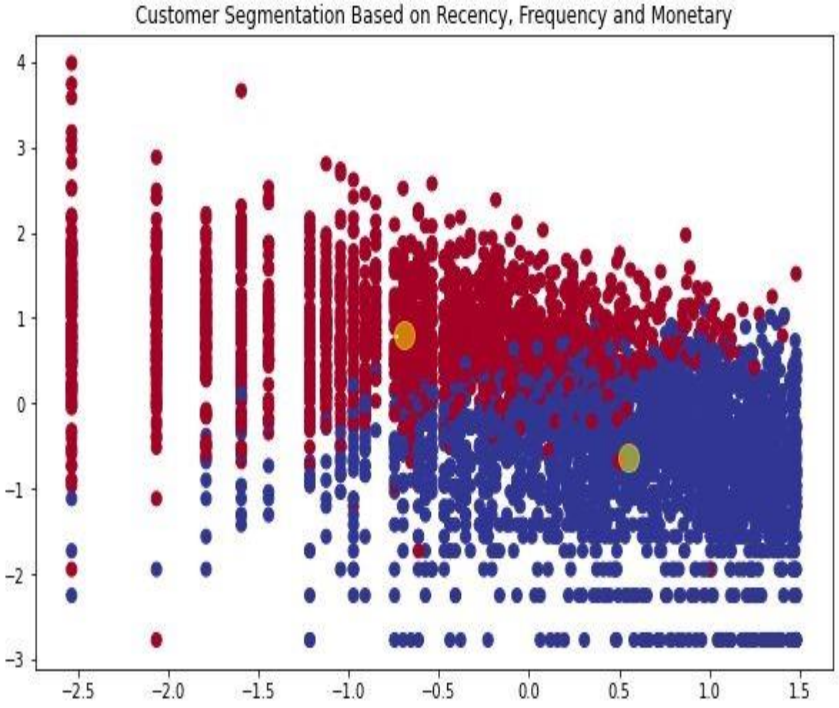
Elbow Method on Frequency and Monetary





Elbow Method on Frequency , Recency and Monetary



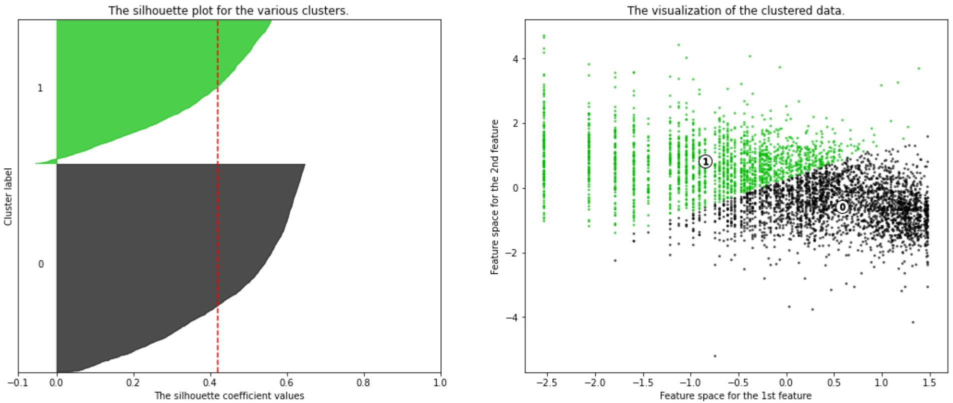


Silhouette Score With N cluster

```
For n_clusters = 2 The average silhouette_score is : 0.4210111657624216
For n_clusters = 3 The average silhouette_score is : 0.34284354535207
For n_clusters = 4 The average silhouette_score is : 0.36431591478500835
For n_clusters = 5 The average silhouette_score is : 0.33674903317539934
For n_clusters = 6 The average silhouette_score is : 0.34421514431762534
For n_clusters = 7 The average silhouette_score is : 0.34738037136141997
For n_clusters = 8 The average silhouette_score is : 0.3373928797084938
For n_clusters = 9 The average silhouette_score is : 0.34551637075041824
For n_clusters = 10 The average silhouette_score is : 0.34826185975167395
```

Silhouette Analysis With N cluster 2 to 10

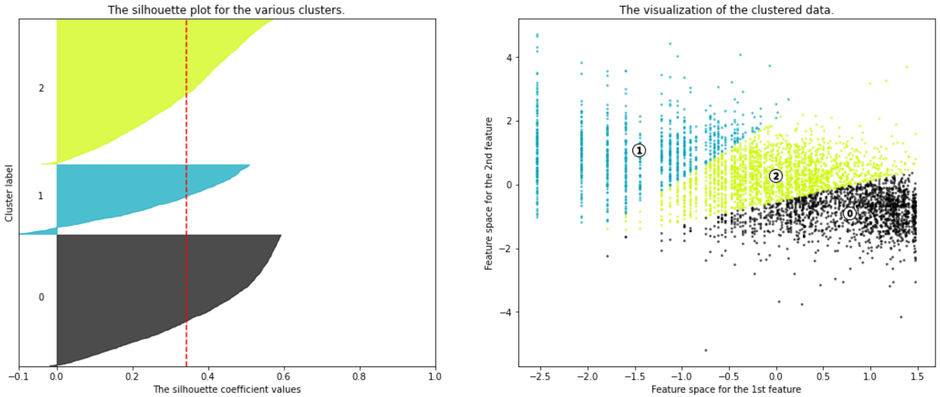
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



The silhouette plot for the various clusters.

The visualization of the clustered data.

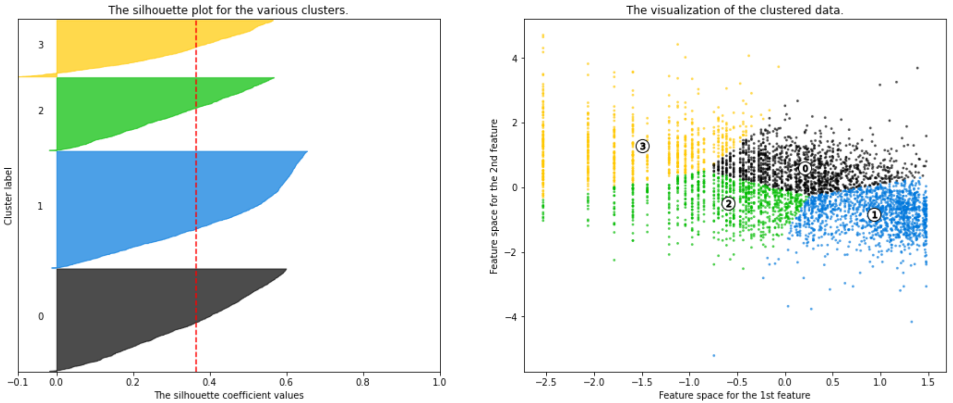
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



The silhouette plot for the various clusters.

The visualization of the clustered data.

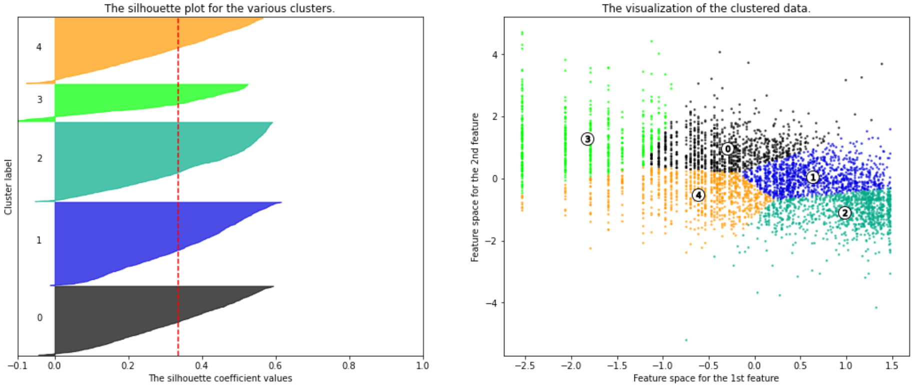
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



The silhouette plot for the various clusters.

The visualization of the clustered data.

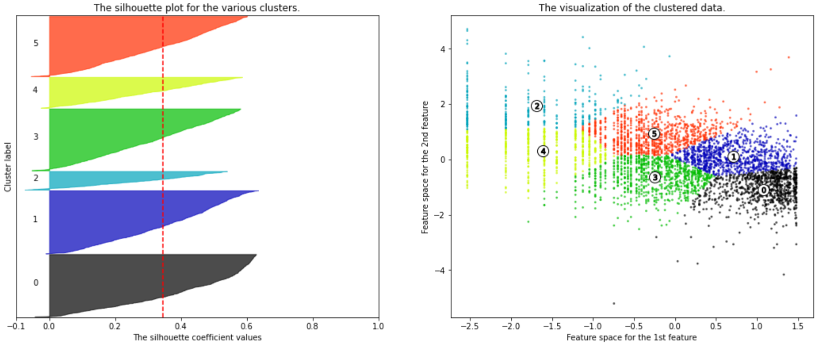
Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



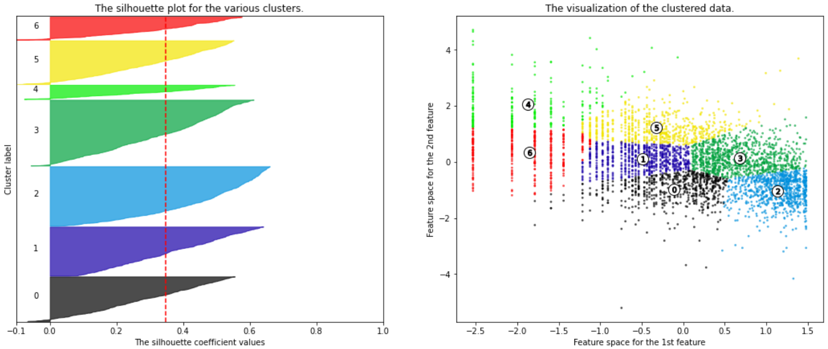
The silhouette plot for the various clusters.

The visualization of the clustered data.

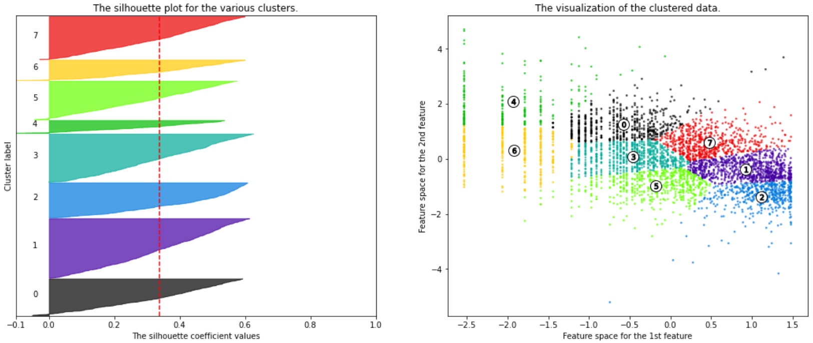
Silhouette analysis for KMeans clustering on sample data with n_clusters = 6



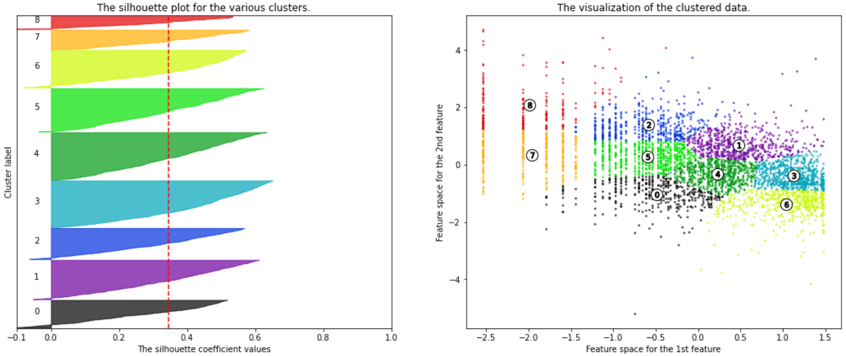
Silhouette analysis for KMeans clustering on sample data with n_clusters = 7



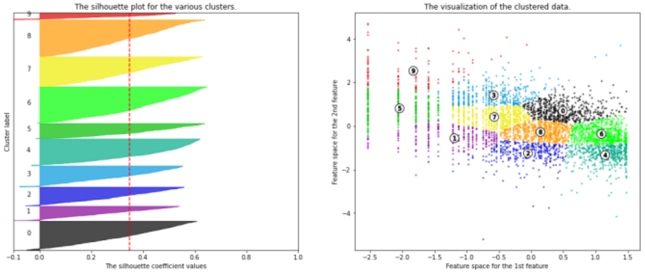
Silhouette analysis for KMeans clustering on sample data with n_clusters = 8



Silhouette analysis for KMeans clustering on sample data with n_clusters = 9



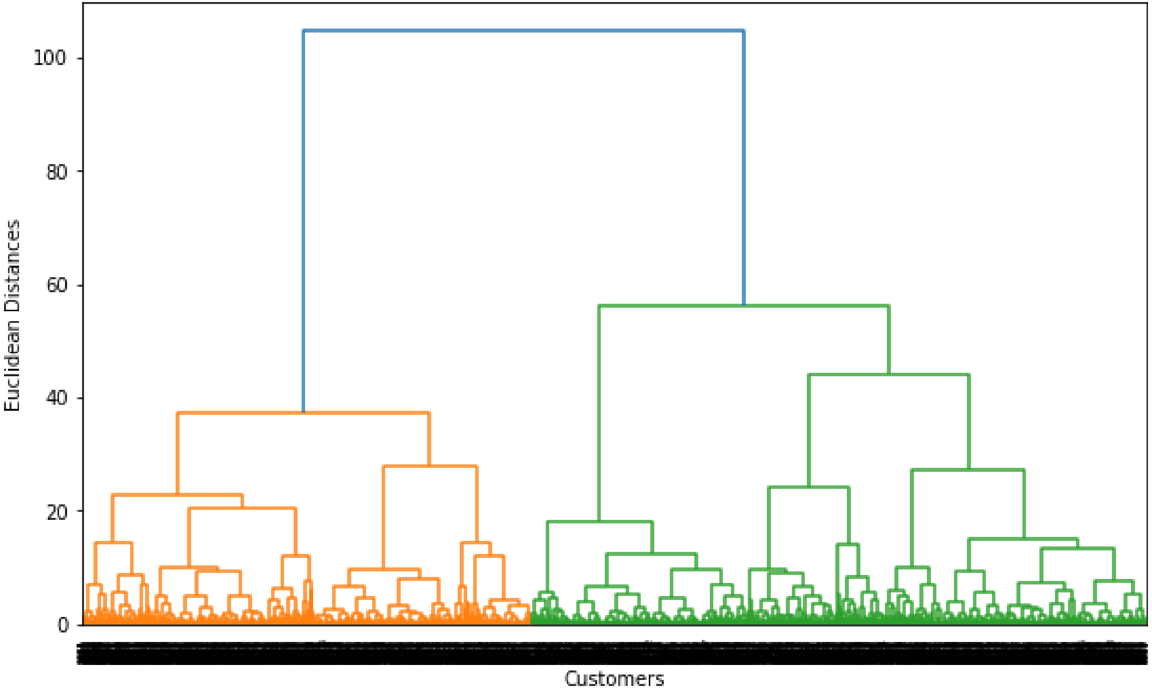
Silhouette analysis for KMeans clustering on sample data with n_clusters = 10



RFM Analysis

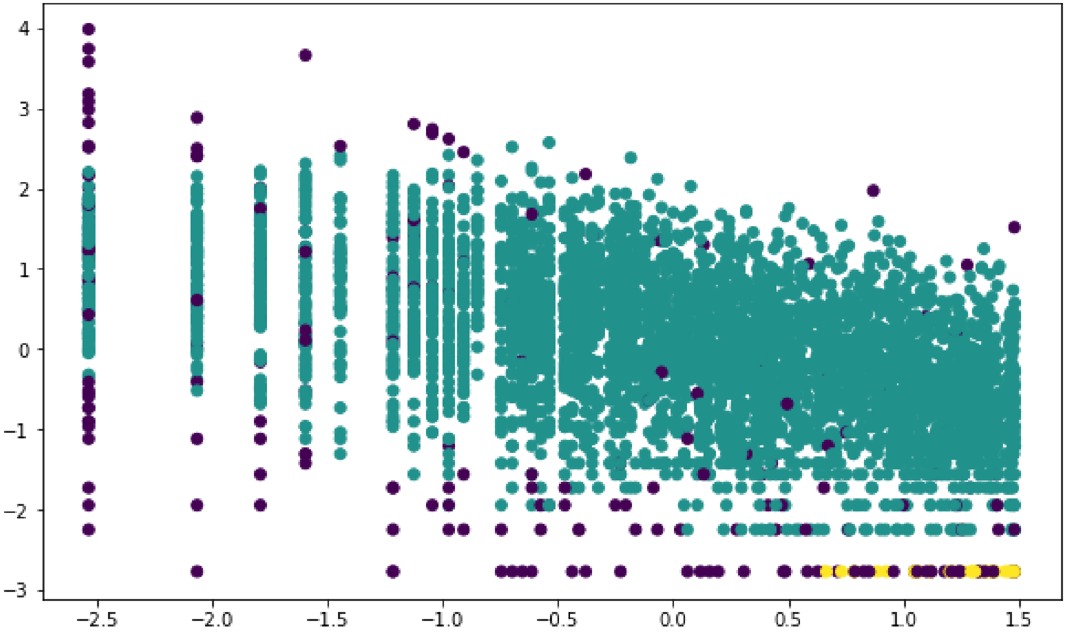
CustomerID	Recency	Frequency	Monetry	R	F	M	RFM	RFM Score	Loyalty	Recency_log	Frequency log	Monetry_log	Cluster
12346.0	325	1	77183.60	4	4	1	441	9	Silver	5.783825	0.000000	11.253942	1
12347.0	2	182	4310.00	1	1	1	111	3	Platinum	0.693147	5.204007	8.368693	0
12348.0	75	31	1797.24	3	3	1	331	7	Gold	4.317488	3.433987	7.494007	1
12349.0	18	73	1757.55	2	2	1	221	5	Platinum	2.890372	4.290459	7.471676	0
12350.0	310	17	334.40	4	4	3	443	11	Bronz	5.736572	2.833213	5.812338	1
12352.0	36	85	2506.04	2	2	1	221	5	Platinum	3.583519	4.442651	7.826459	0
12353.0	204	4	89.00	4	4	4	444	12	Bronz	5.318120	1.386294	4.488636	1
12354.0	232	58	1079.40	4	2	2	422	8	Gold	5.446737	4.060443	6.984161	1
12355.0	214	13	459.40	4	4	3	443	11	Bronz	5.365976	2.564949	6.129921	1
12356.0	22	59	2811.43	2	2	1	221	5	Platinum	3.091042	4.077537	7.941449	0
12357.0	33	131	6207.67	2	1	1	211	4	Platinum	3.496508	4.875197	8.733541	0
12358.0	1	19	1168.06	1	3	2	132	6	Gold	0.000000	2.944439	7.063100	0
12359.0	57	248	6372.58	3	1	1	311	5	Platinum	4.043051	5.513429	8.759760	0
12360.0	52	129	2662.06	3	1	1	311	5	Platinum	3.951244	4.859812	7.886856	0
12361.0	287	10	189.90	4	4	4	444	12	Bronz	5.659482	2.302585	5.246498	1

Dendogram



The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold = 90 and Number of cluster is 2

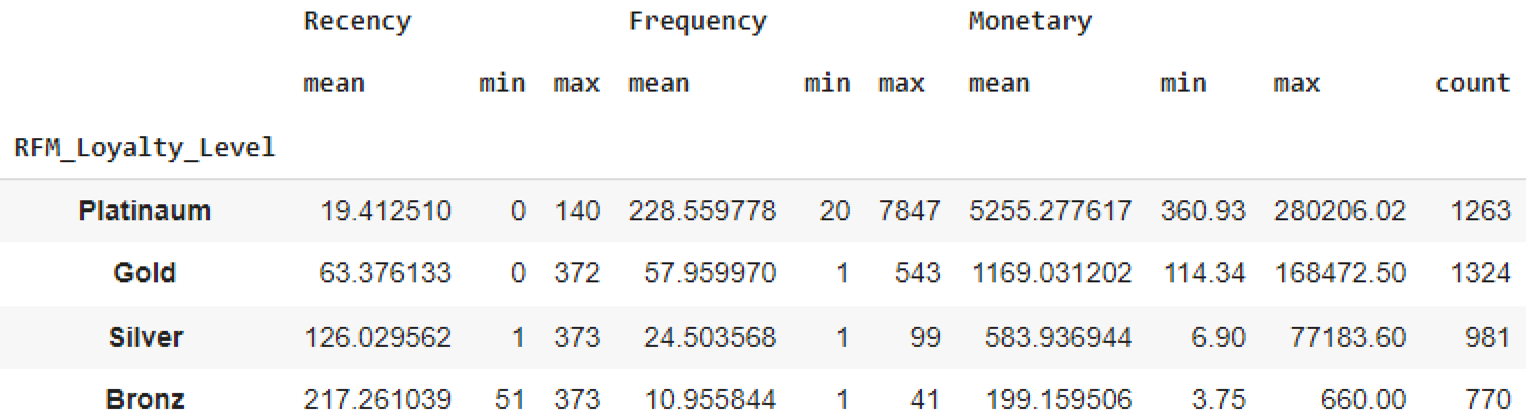
DBSCAN on R, F and M



Conclusion: We can conclude by above plot that customers are well segmented by Recency, Frequency and Monetary. Also, the number of clusters is equal to 3.

Summary And Conclusion of project

Firstly we did clustering based on RFM analysis. We had 4 clusters/Segmentation of customers based on RFM score.



	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
RFM_Loyalty_Level										
Platinaum	19.412510	0	140	228.559778	20	7847	5255.277617	360.93	280206.02	1263
Gold	63.376133	0	372	57.959970	1	543	1169.031202	114.34	168472.50	1324
Silver	126.029562	1	373	24.503568	1	99	583.936944	6.90	77183.60	981
Bronz	217.261039	51	373	10.955844	1	41	199.159506	3.75	660.00	770

- Later we implemented the machine learning algorithms to cluster the customers.

- Above clustering is done with recency , frequency and monetary data(K means Clustering) as all 3 together will provide more information.
- Cluster 0 has high recency rate but very low frequency and monetary. Cluster 0 conatins 2414 customers.
- Cluster 1 has low recency rate but they are frequent buyers and spends very high money than other customers as mean monetary value is very high.Thus generates more revenue to the retail business

Overall Conclusion

Throughout the analysis we went through various steps to perform customer segmentation. We started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications. Next, we did some exploratory data analysis and tried to draw observations from the features we had in the dataset.

Next, we formulated some quantitative factors such as recency, frequency and monetary known as rfm model for each of the customers. We implemented K Means clustering algorithm on these features. We also performed silhouette and elbow method analysis to determine the optimal no. of clusters which was 2. We saw customers having high recency and low frequency and monetary values were part of one cluster and customers having low recency and high frequency, monetary values were part of another cluster.

However, there can be more modifications on this analysis. One may choose to cluster into more no. depending on company objectives and preferences. The labelled feature after clustering can be fed into classification supervised machine learning algorithms that could predict the classes for new set of observations. The clustering can also be performed on new set of features such as type of products each customer prefer to buy often, finding out customer lifetime value (clv), segmenting on the basis of time period they visit and much more.

Thank
You