Capstone Project - 4

Customer Segmentation

Faraz Ahmad
(Individual)

Problem statement

- To identify major customer segments on a transnational data set 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

- **InvoiceNo:** Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode:** 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.
- **InvoiceDate:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice:** Unit price. Numeric, Product price per unit in sterling.
- **CustomerID:** 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.

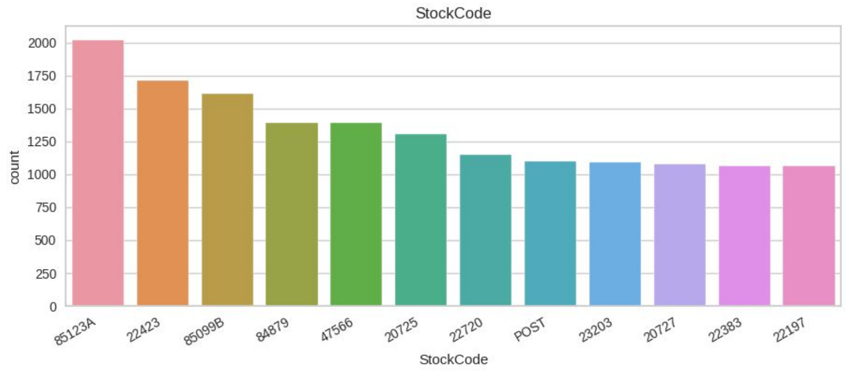
Inspecting dataset

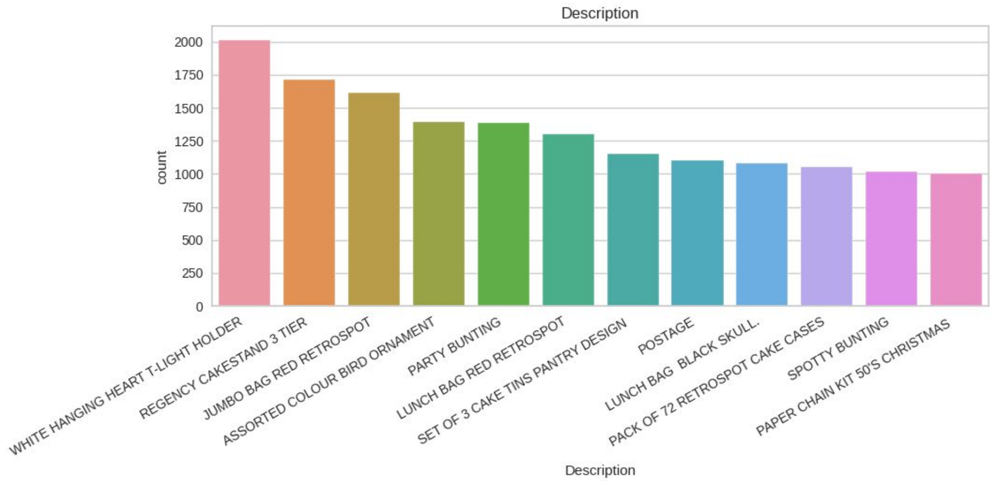
- The dataset contains 541909 rows and 8 columns.
- Missing data count and percentage for each column are as follows:
    - CustomerID 135080 (24.93%)
    - Description 1454 (0.27%)
- There is no use of this data. So, it can be dropped.
- There are 5225 duplicated data points. So, these data points are dropped too.
- Therefore, total number of observations after cleaning the dataset are 401604.

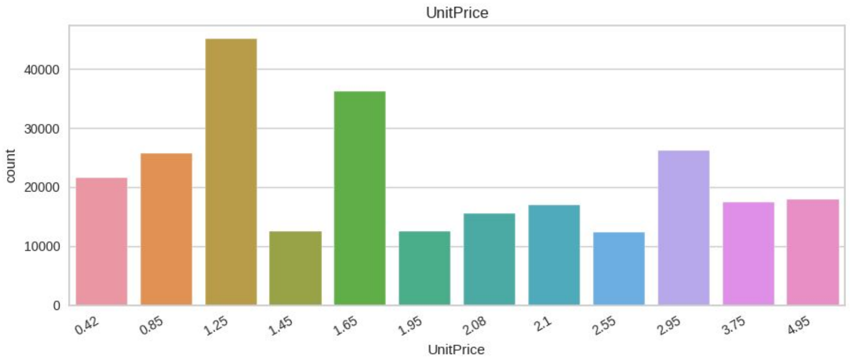
Feature engineering

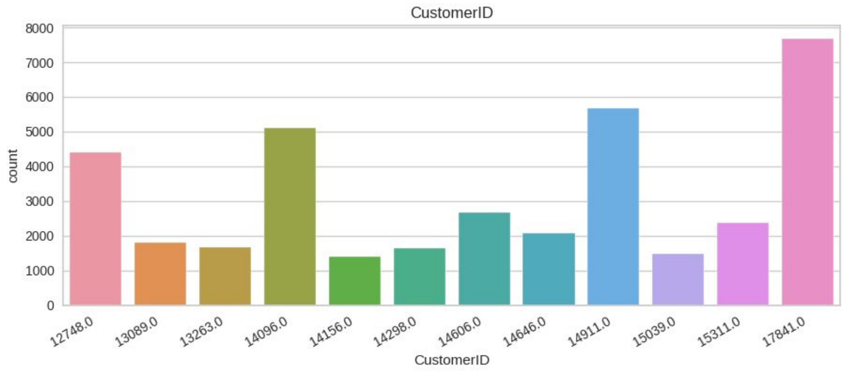
- Extraction year, month, day and hour from Invoice data.
- Adding feature **'TotalAmount'** by multiplying values from the **Quantity** and **UnitPrice** column.
- Adding feature **'TimeType'** which is based on hours to define whether it is Morning, Afternoon or Evening.
- Dropping InvoiceNo starting with **'C'** as it represents cancellation.

Most frequent values

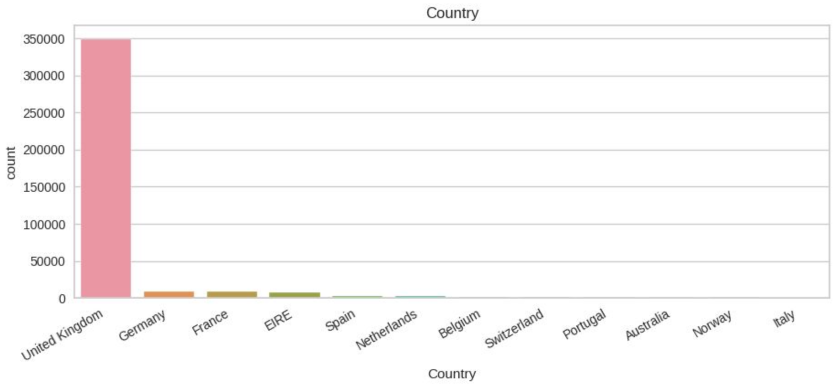


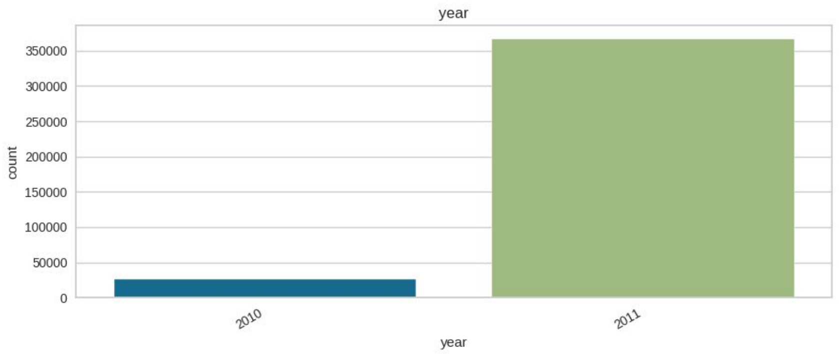


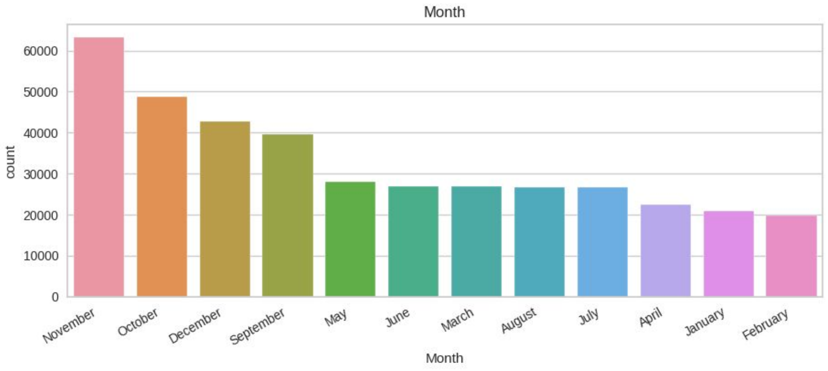


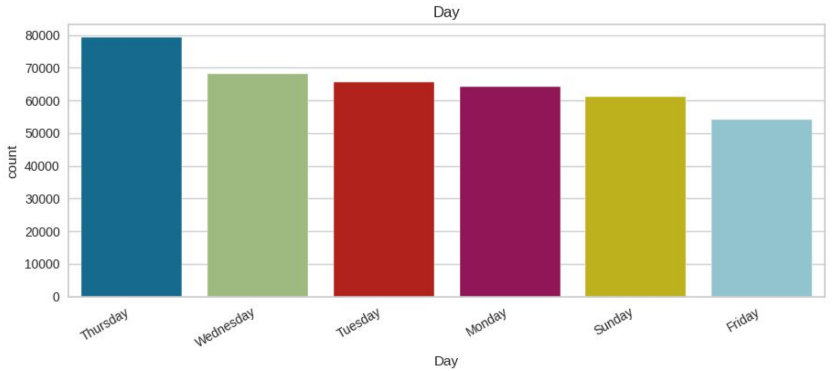


Most frequent values

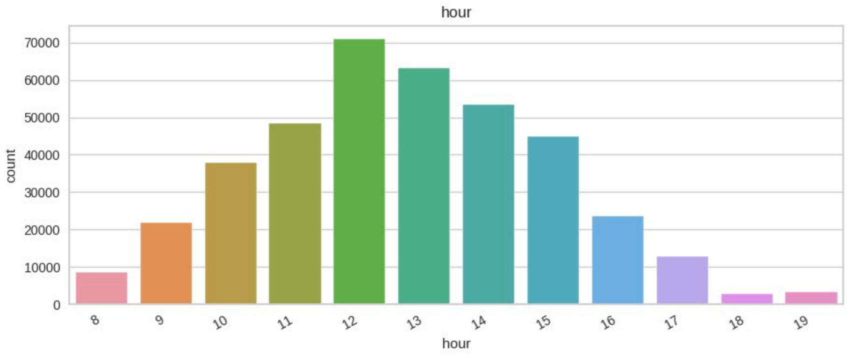


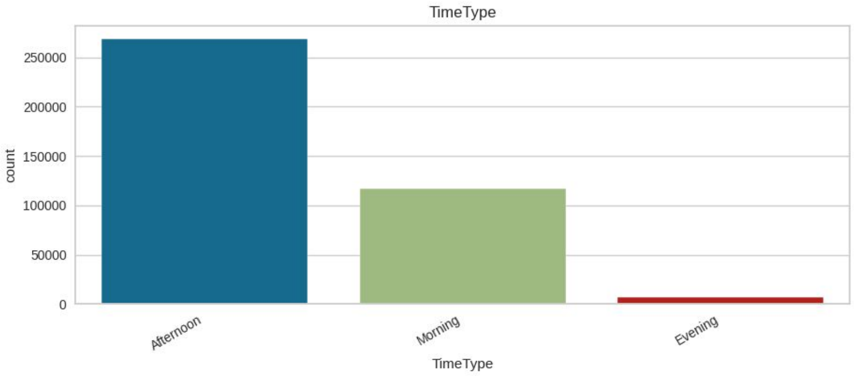






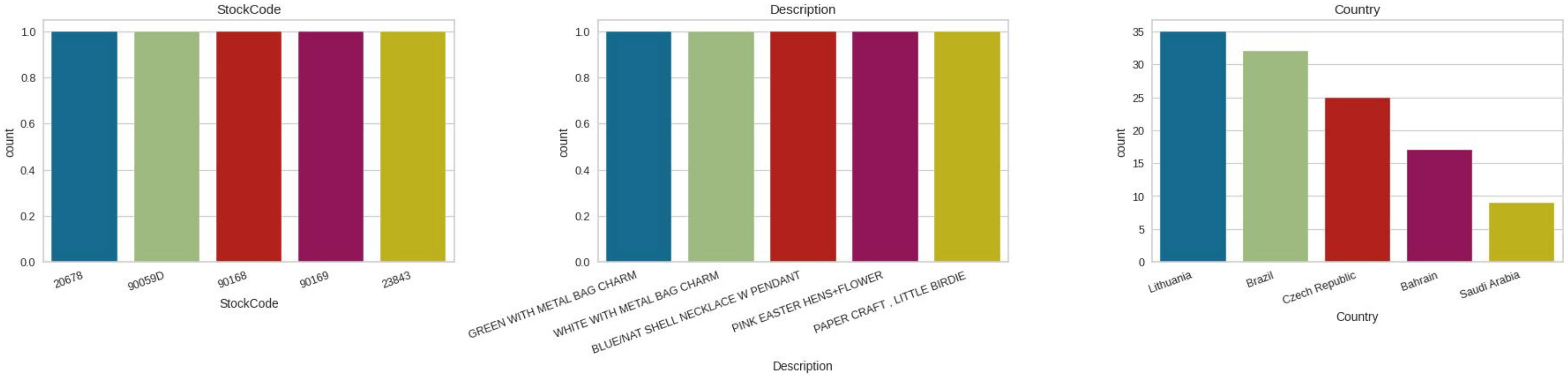
Most frequent values





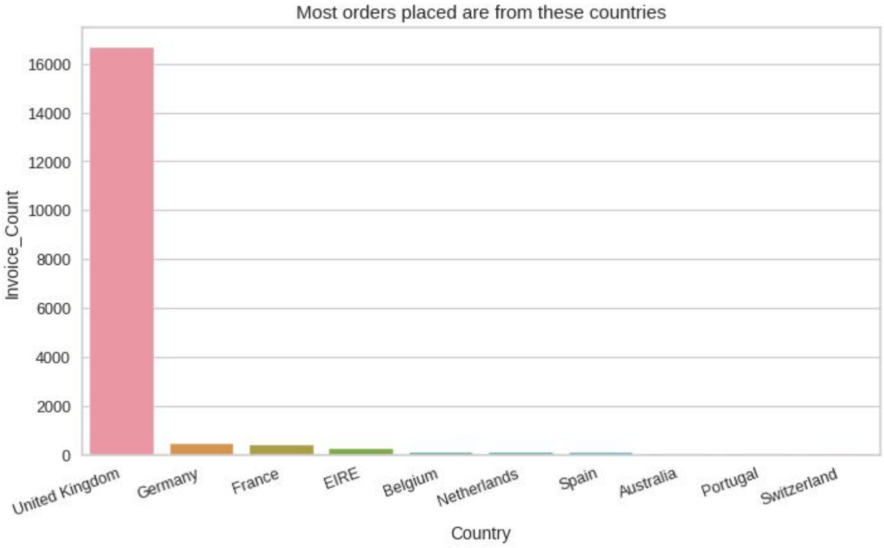
- Most Customers are from United Kingdom. Considerable number of customers are also from Germany, France, EIRE and Spain. Whereas Saudi Arabia, Bahrain, Czech Republic, Brazil and Lithuania has least number of customers.
- There are no orders placed on Saturdays. Looks like it's a non working day for the retailer.
- Most of the customers have purchased the gifts in the month of November, October, December and September. Less number of customers have purchased the gifts in the month of April, January and February.
- Most of the customers have purchased the items in Afternoon, moderate numbers of customers have purchased the items in Morning and the least in Evening.
- WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER, JUMBO BAG RED RETROSPOT are the most ordered products.

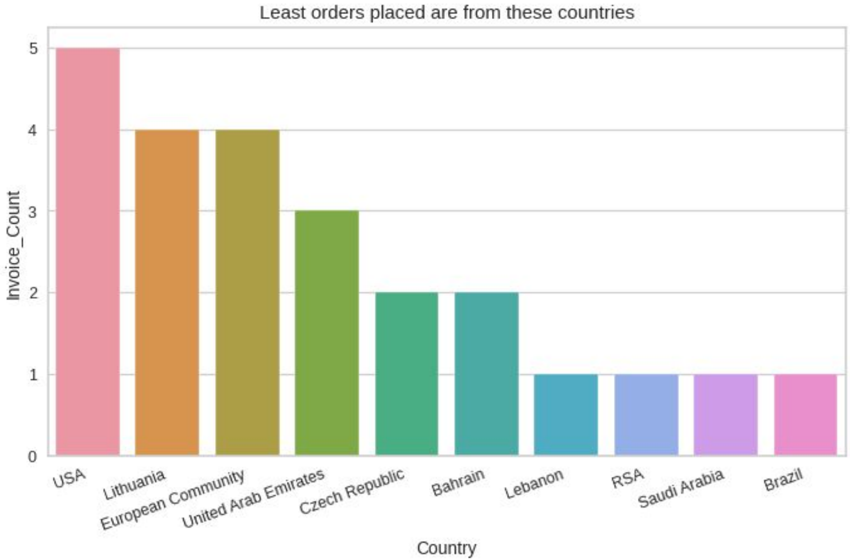
Least frequent values



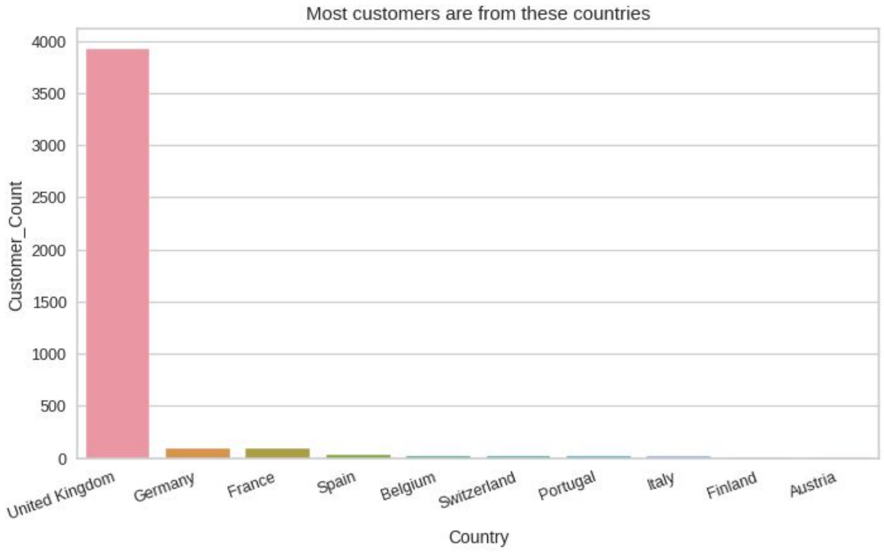
- Saudi Arabia, Bahrain, Czech Republic, Brazil and Lithuania has the least number of customers.
- GREEN WITH METAL BAG CHARM, WHITE WITH METAL BAG CHARM, BLUE/NAT SHELL NECKLACE W PENDANT, PINK EASTER HENS+FLOWER, PAPER CRAFT, LITTLE BRIDE are some of the least sld products.

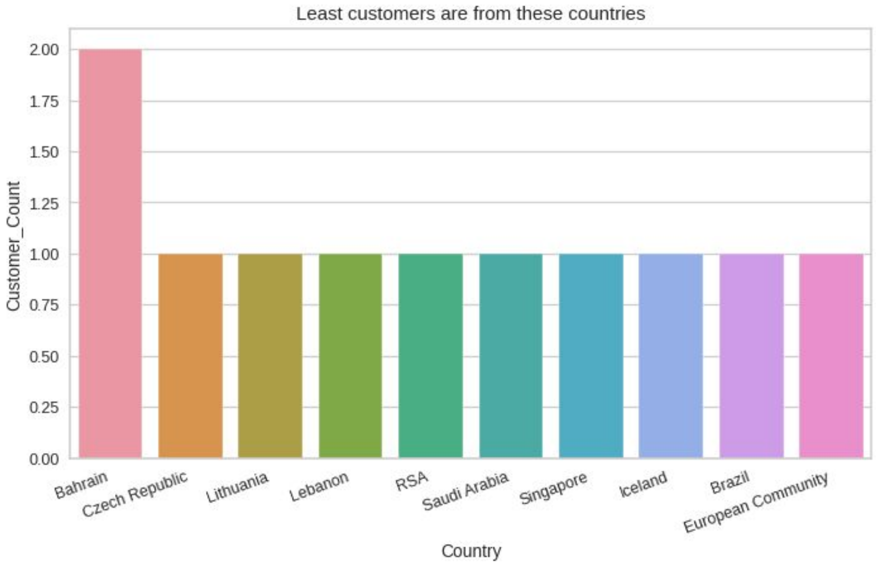
Country wise Orders



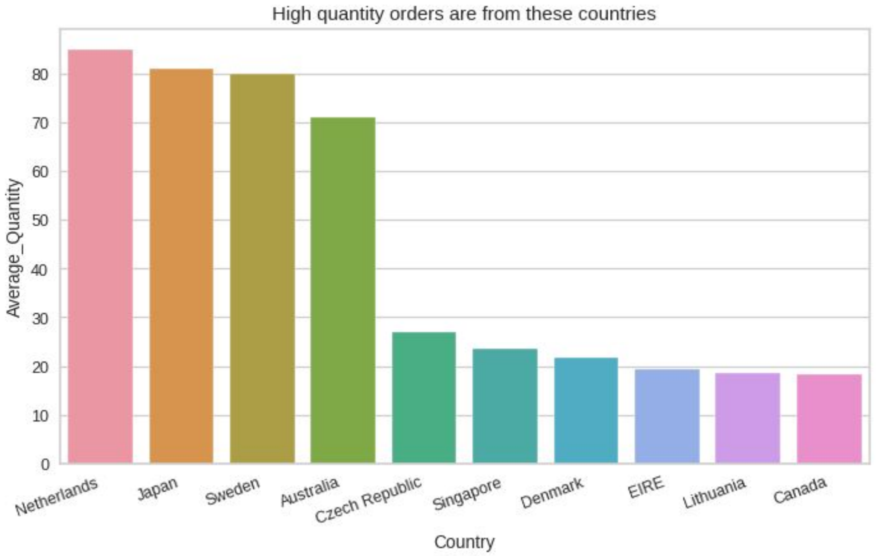


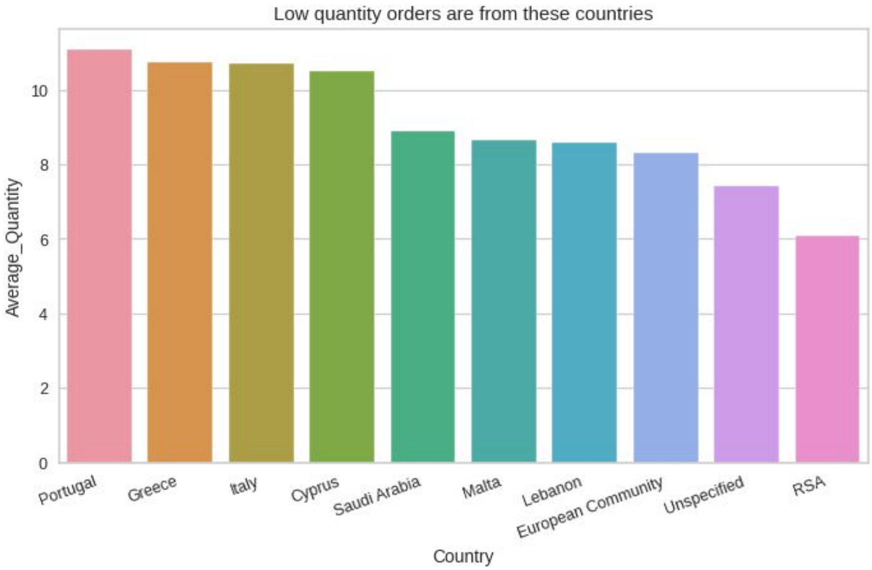
Country wise Customers



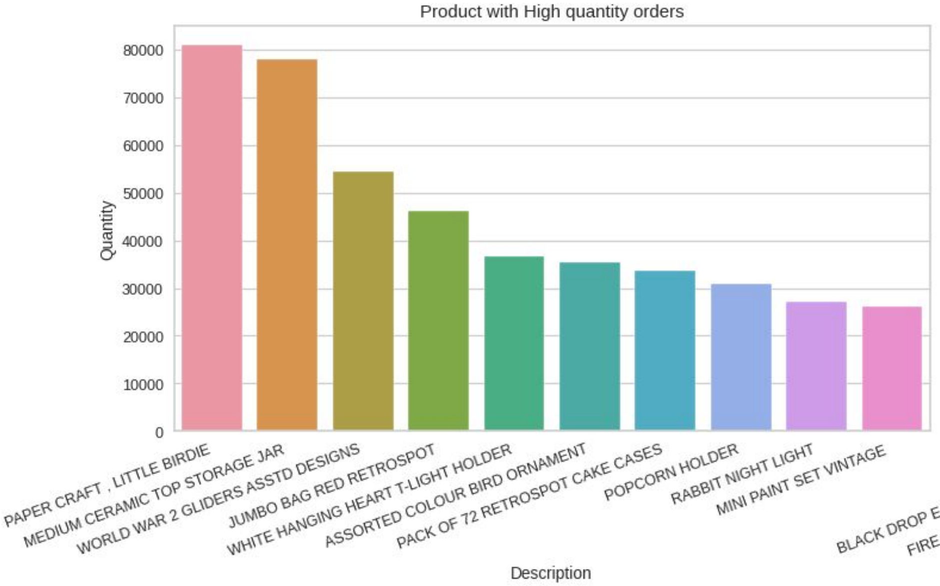


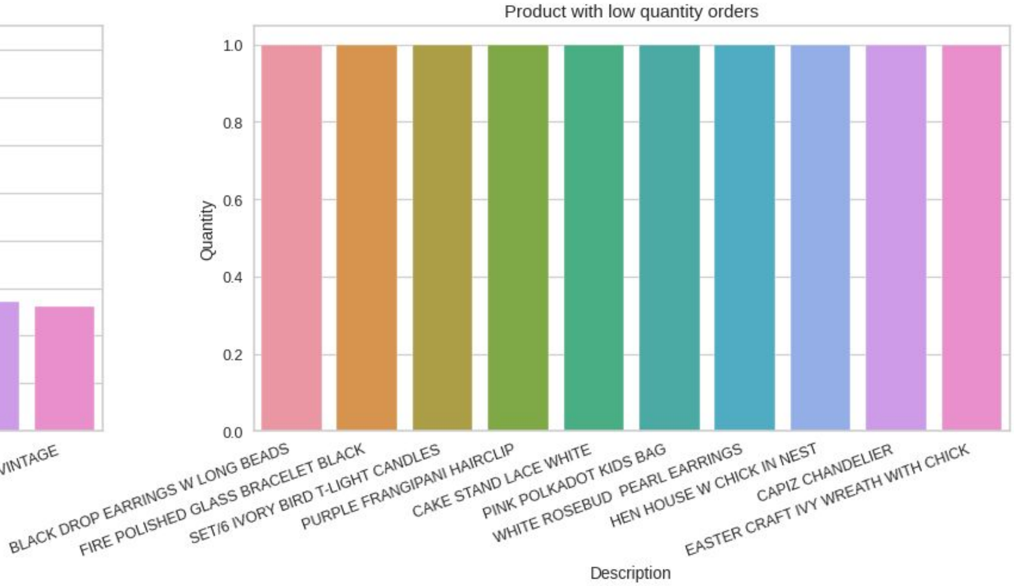
Country wise Purchase Quantity



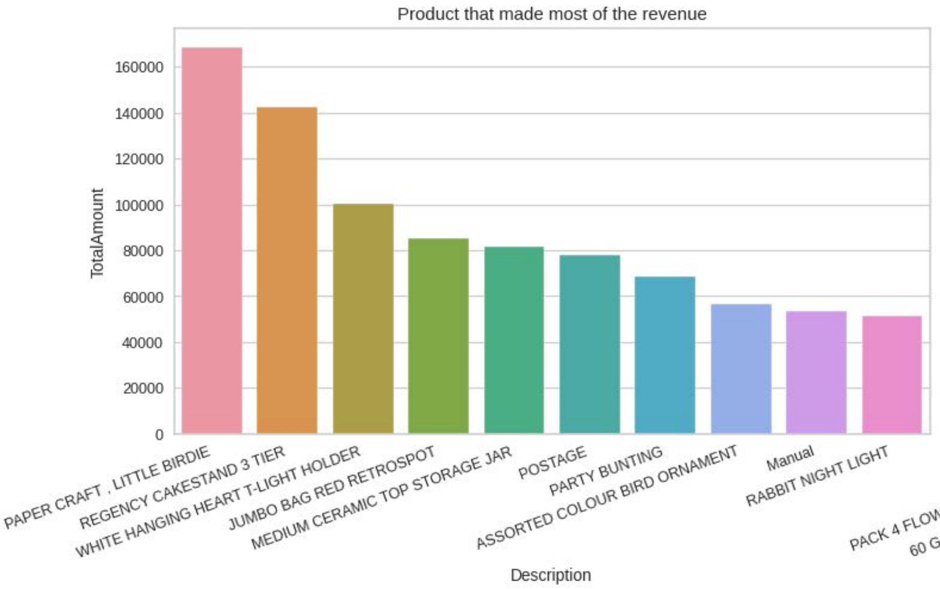


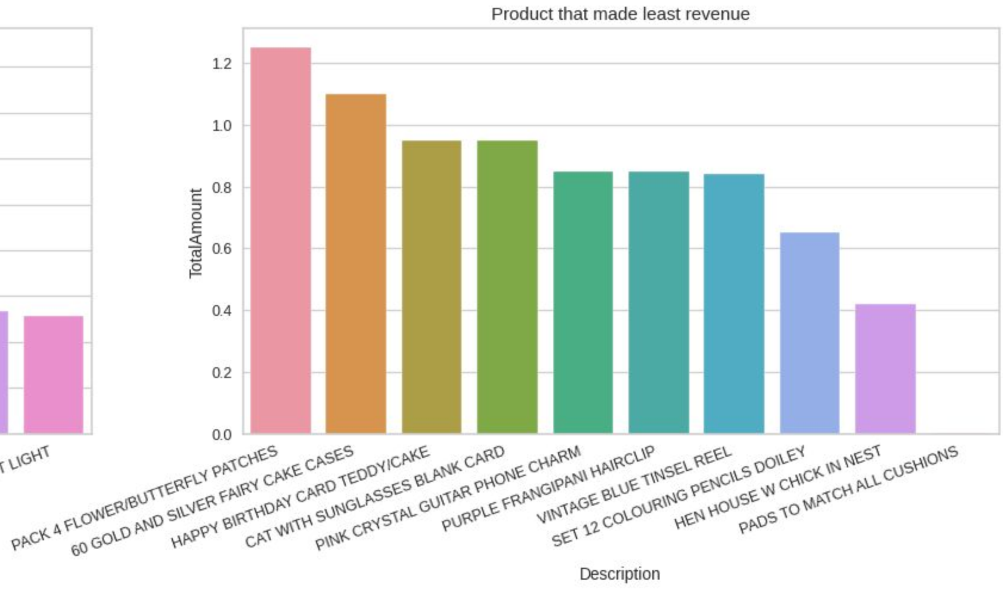
Product wise Purchase Quantity



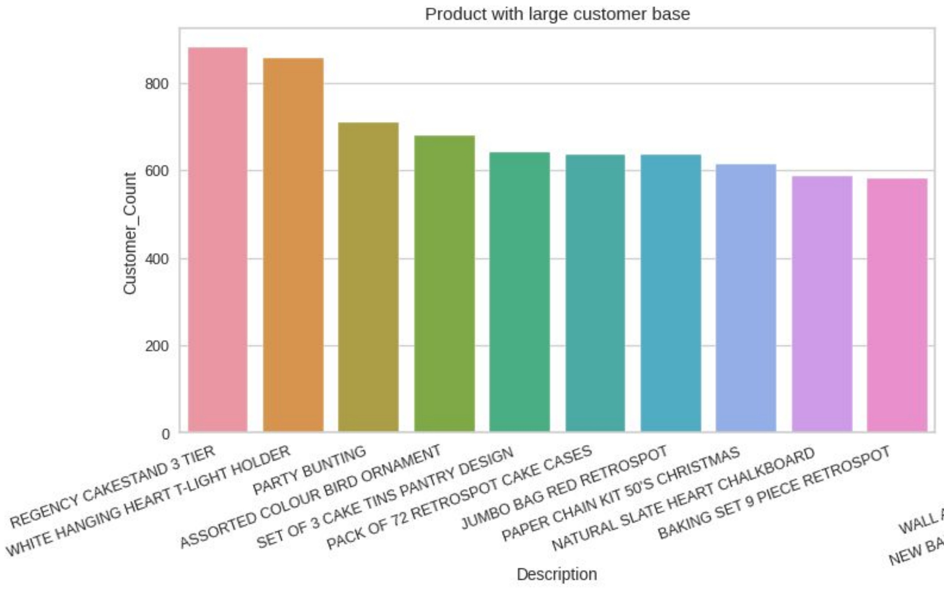


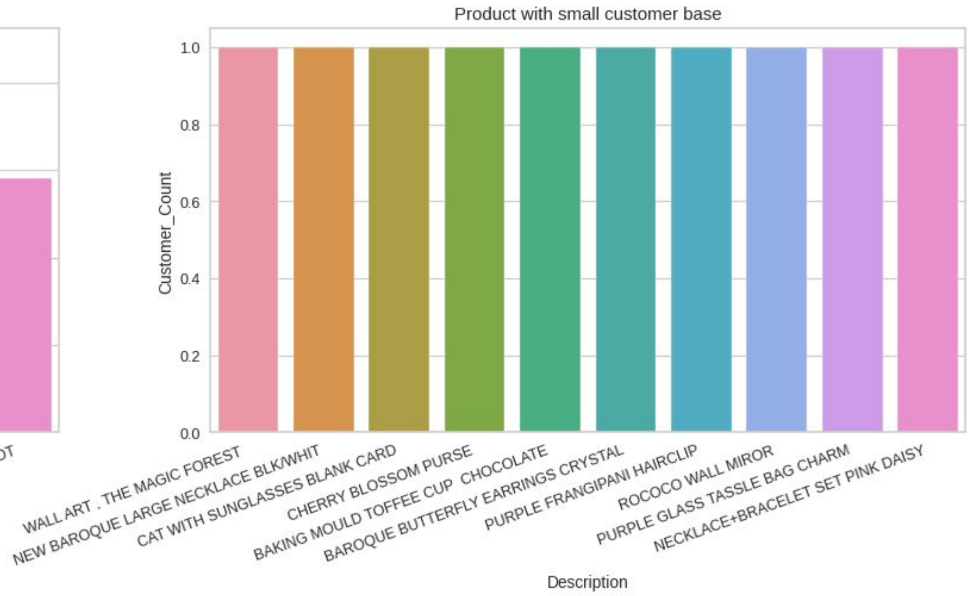
Product wise Revenue



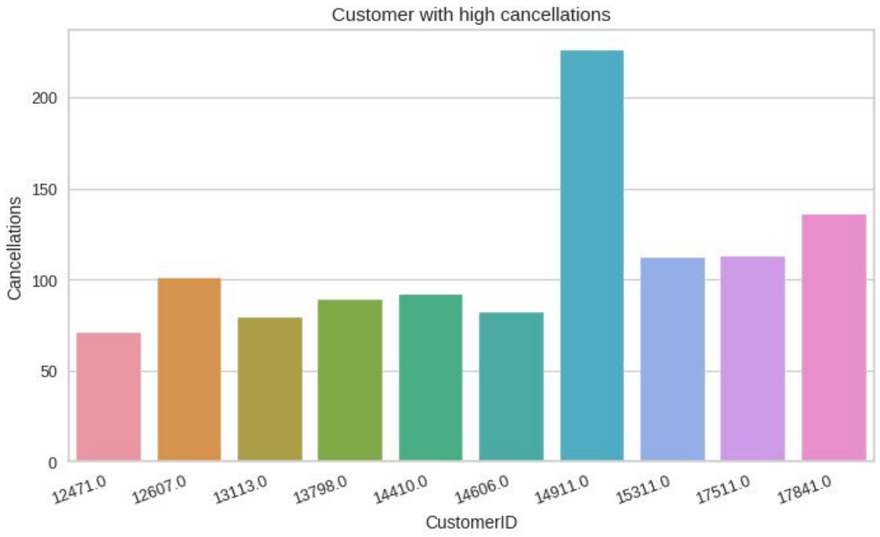


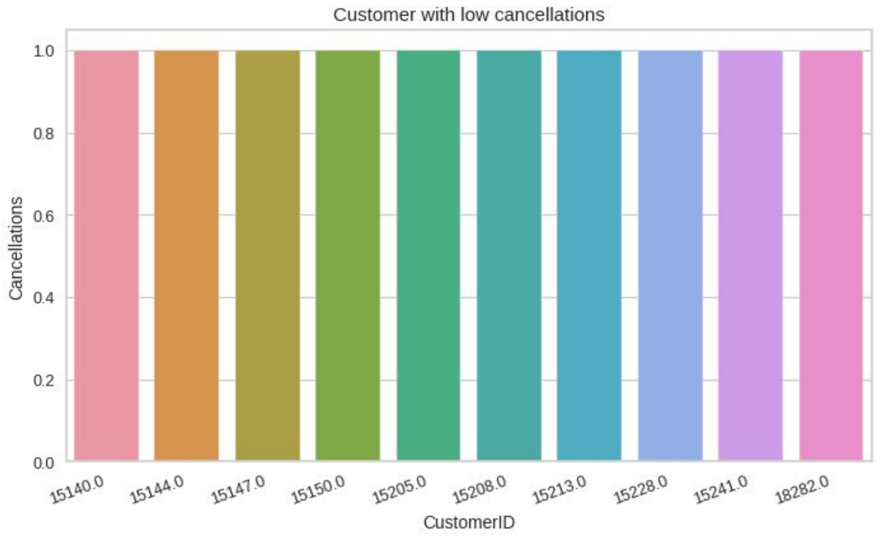
Product wise Customers



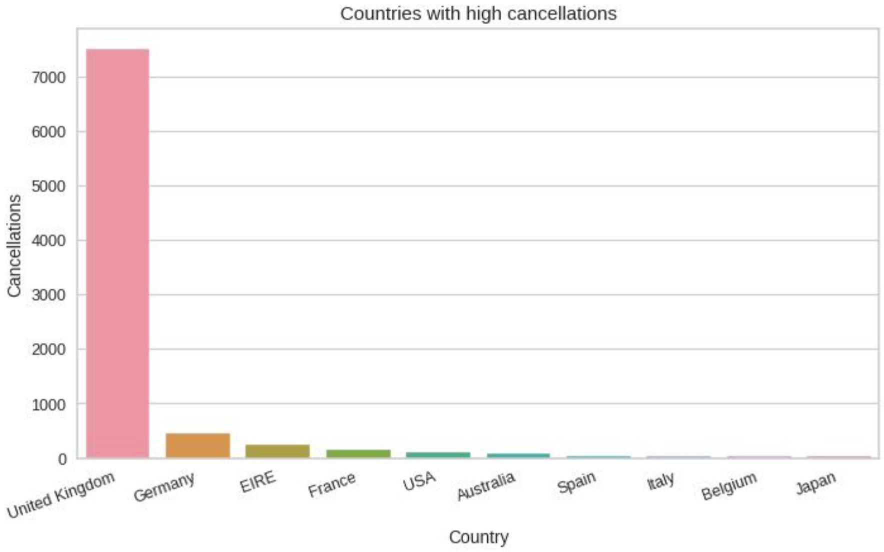


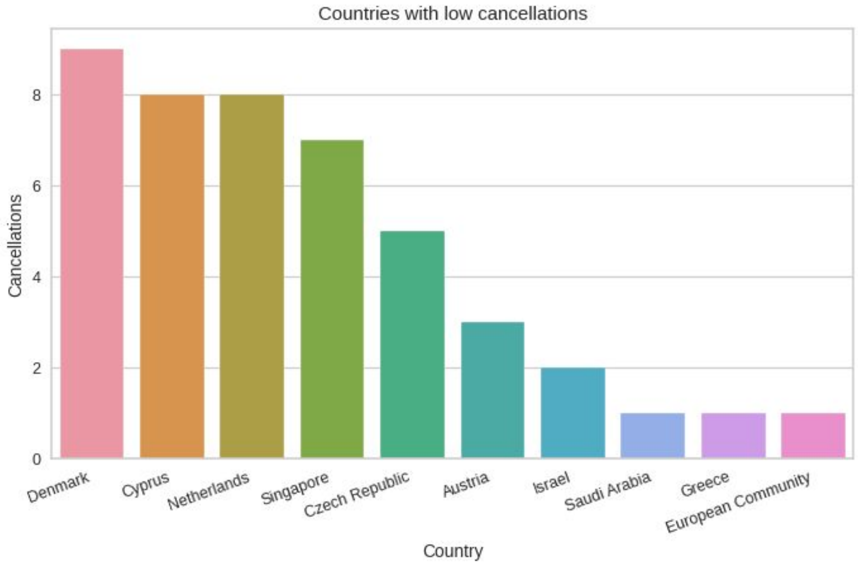
Customer wise Cancellations



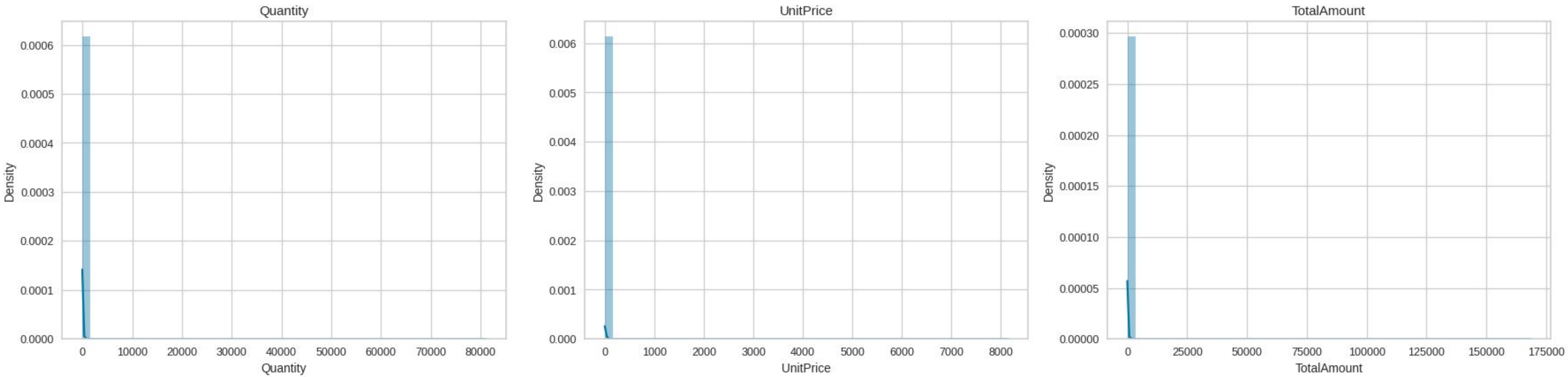


Country wise Cancellations



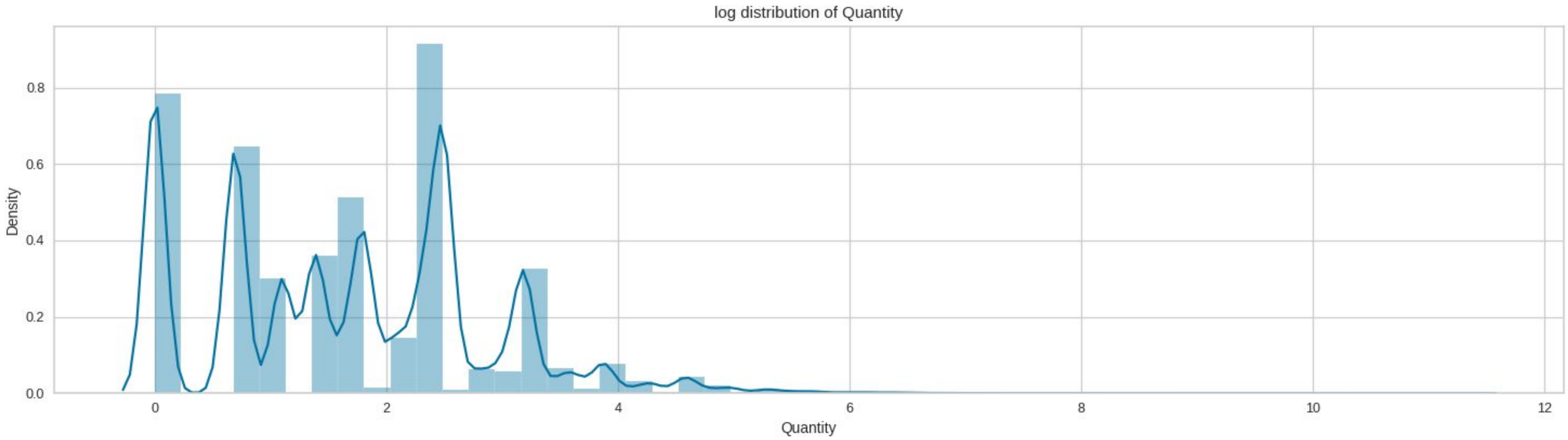


Visualizing distributions



- Visualizing the distribution of Quantity, UnitPrice and TotalAmount columns.
- It shows a positively skewed distribution because most of the values are clustered around the left side of the distribution while the right tail of the distribution is longer, which means mean>median>mode.
- For symmetric graph mean=median=mode.

Log transformation



- After applying log transformation, the distribution plot looks comparatively less skewed.
- We use log transformation when our original continuous data does not follow the bell curve. We use log transformation to make the data as ‘normal’ as possible so that the analysis results from this data become more valid.

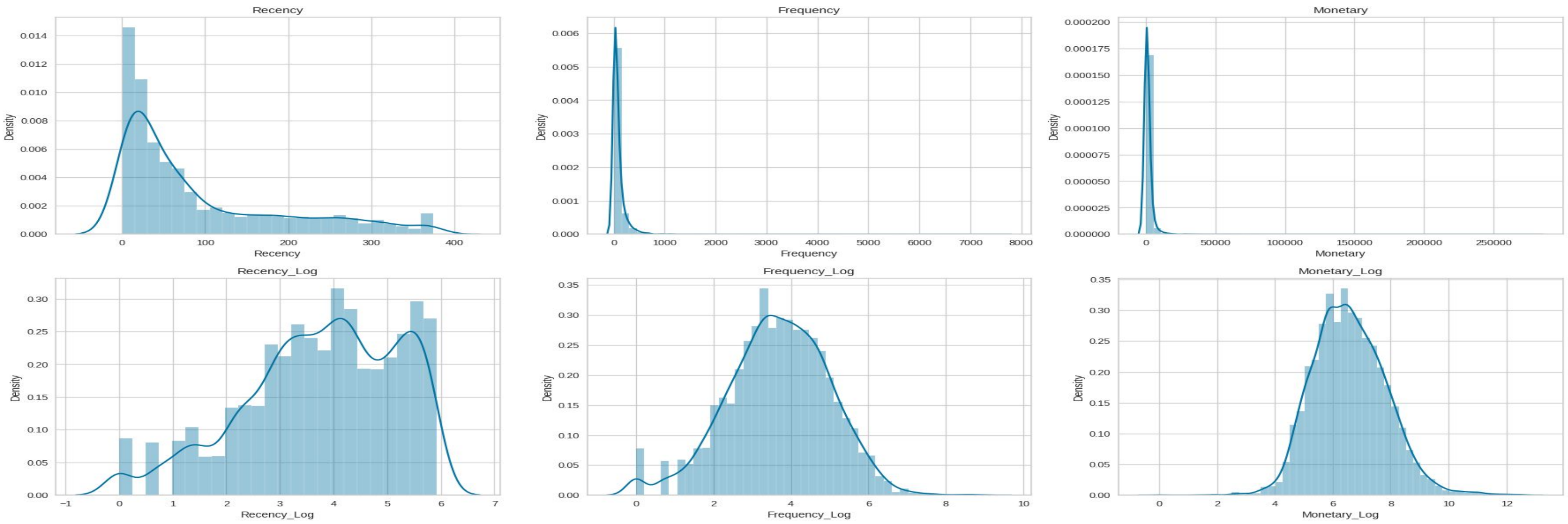
RFM Model Analysis

RFM is a method used to analyze customer value. RFM stands for Recency, Frequency, and Monetary.

- **Recency:** How recently did the customer visit our website or how recently did a customer purchase?
- **Frequency:** How often do they visit or how often do they purchase?
- **Monetary:** How much revenue we get from their visit or how much do they spend when they purchase?

RFM Analysis is a marketing framework that is used to understand and analyze customer behavior based on the above three factors Recency, Frequency, and Monetary.
The RFM Analysis will help the businesses to segment their customer base into different homogenous groups so that they can engage with each group with different targeted marketing strategies.

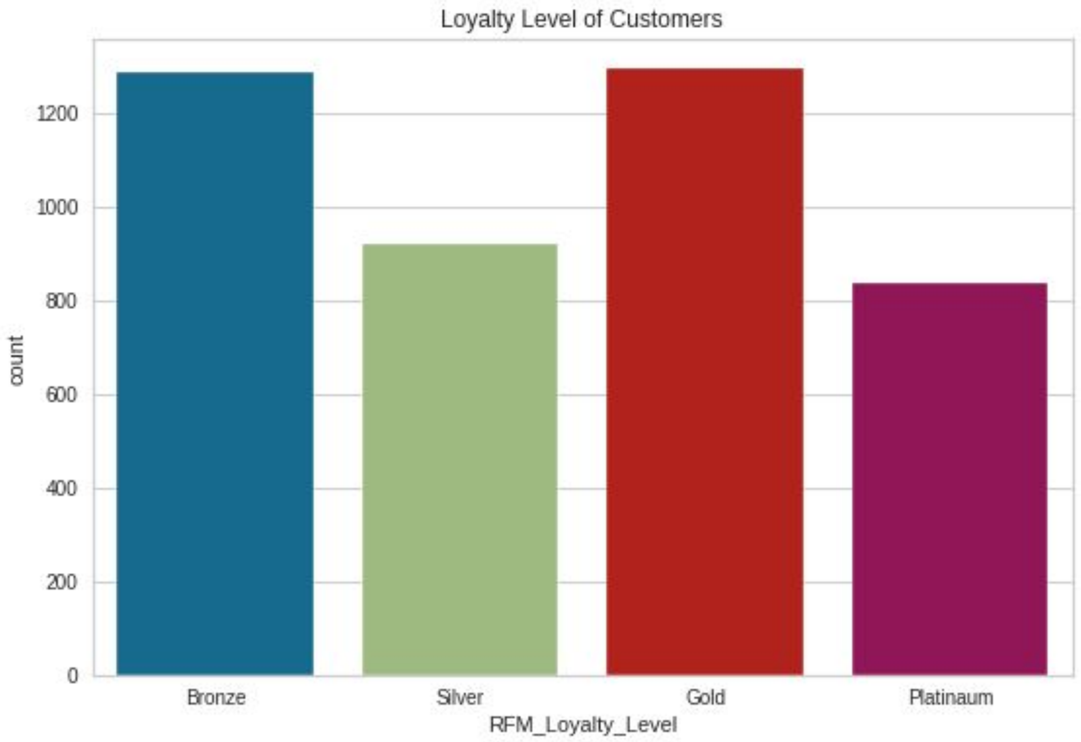
RFM Model Analysis



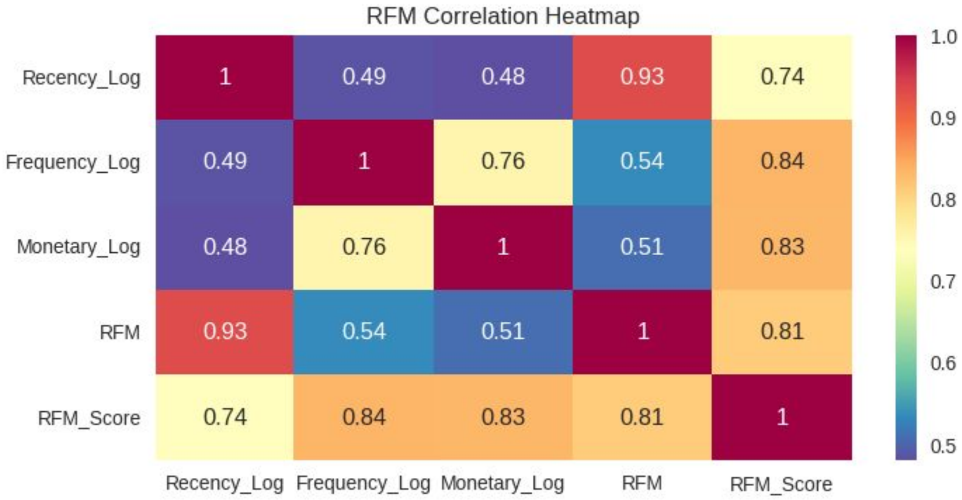
- Earlier the distributions of Recency, Frequency and Monetary columns were positively skewed but after applying log transformation, the distributions appear to be symmetrical and normally distributed.
- It will be more suitable to use the transformed features for better visualization of clusters.

RFM Model Analysis

Using RFM Model analysis, we created 4 clusters namely Bronze, Silver, Gold and Platinum.



RFM Correlation heatmap



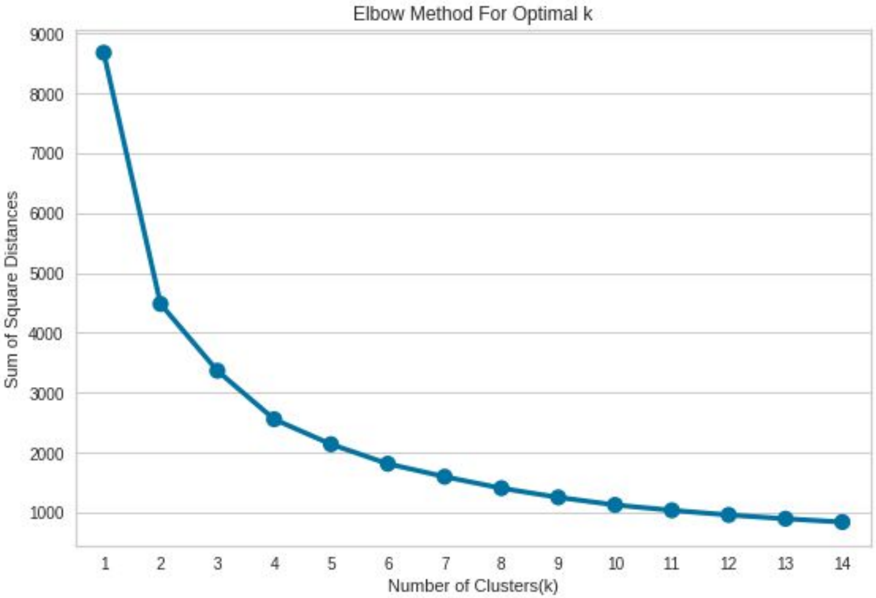
- We can see that Recency is highly correlated with the RFM value.
- Frequency and Monetary are moderately correlated with the RFM.

Scaling for Clustering Analysis

- Firstly, Log Transformation of features like Recency, Frequency and Monetary.
- Then, we applied StandardScaler to scale the features.

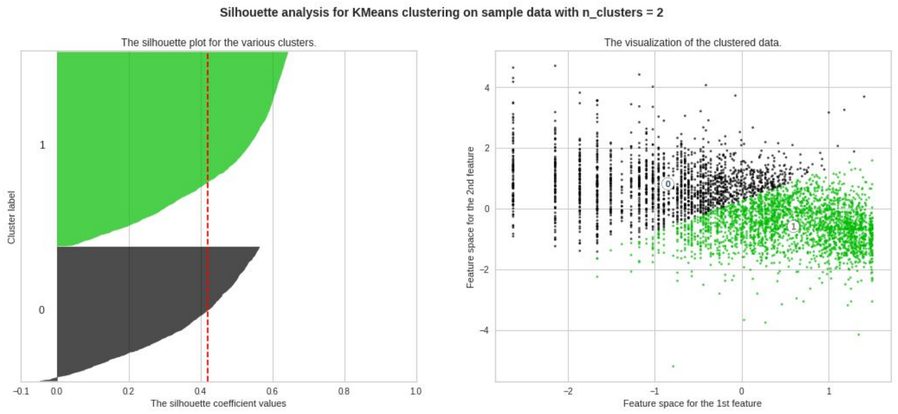
K-means Clustering (Recency and Monetary)

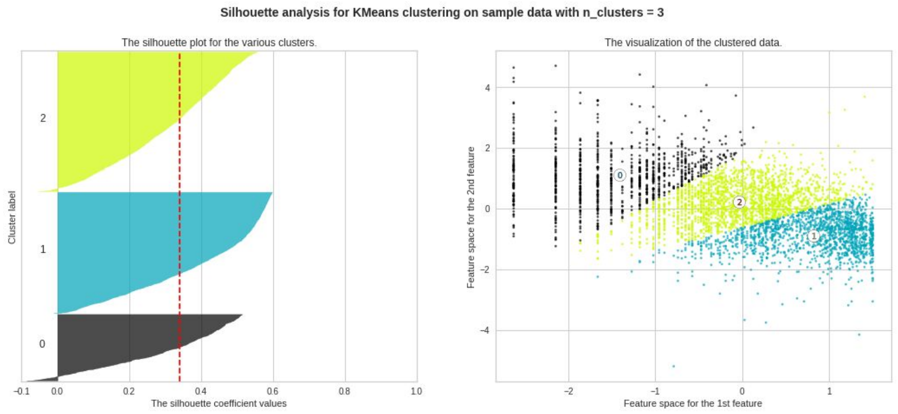
Finding the Optimal value of cluster using Elbow method and Silhouette Score.

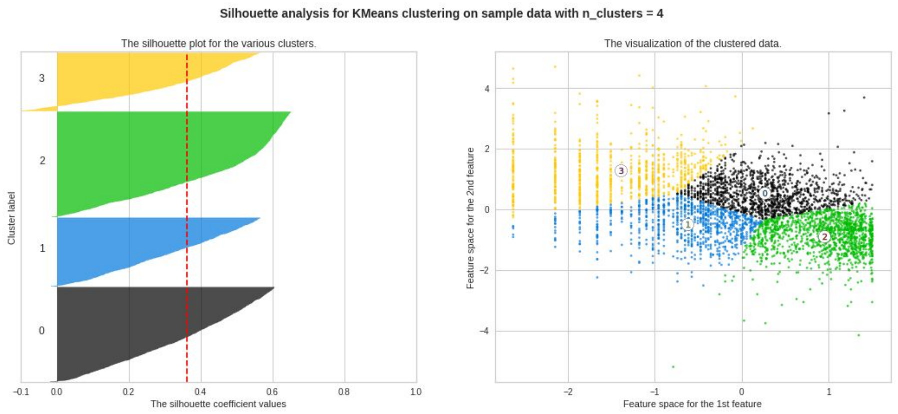


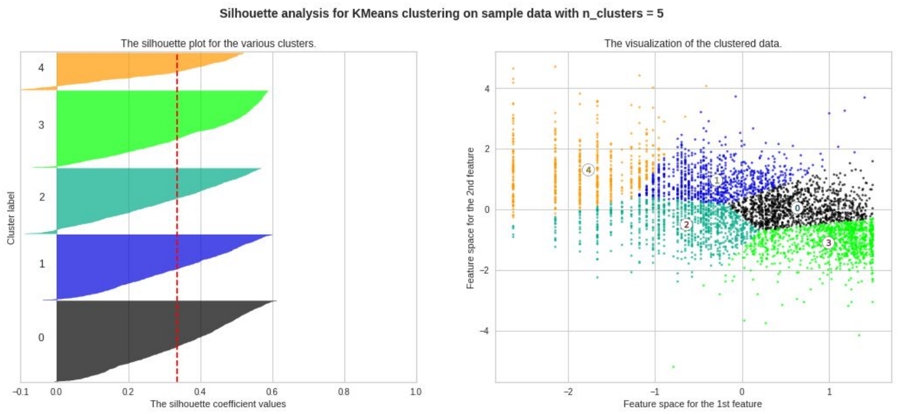
For n_clusters = 2 The average silhouette_score is : 0.42
For n_clusters = 3 The average silhouette_score is : 0.341
For n_clusters = 4 The average silhouette_score is : 0.362
For n_clusters = 5 The average silhouette_score is : 0.336
For n_clusters = 6 The average silhouette_score is : 0.343
For n_clusters = 7 The average silhouette_score is : 0.341
For n_clusters = 8 The average silhouette_score is : 0.337
For n_clusters = 9 The average silhouette_score is : 0.344
For n_clusters = 10 The average silhouette_score is : 0.347
For n_clusters = 11 The average silhouette_score is : 0.337
For n_clusters = 12 The average silhouette_score is : 0.339
For n_clusters = 13 The average silhouette_score is : 0.338
For n_clusters = 14 The average silhouette_score is : 0.34
For n_clusters = 15 The average silhouette_score is : 0.338

K-means Clustering (Recency and Monetary)



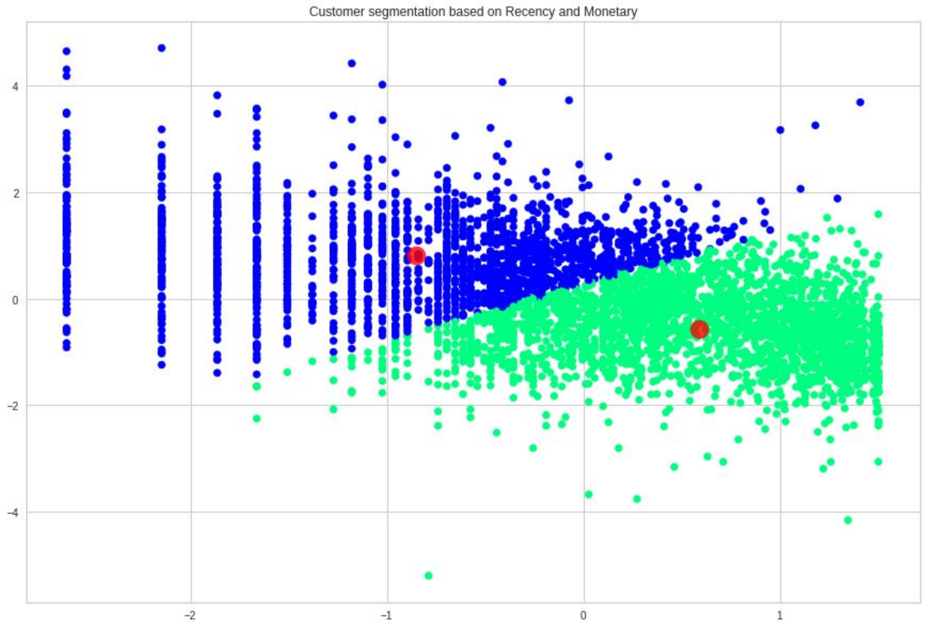




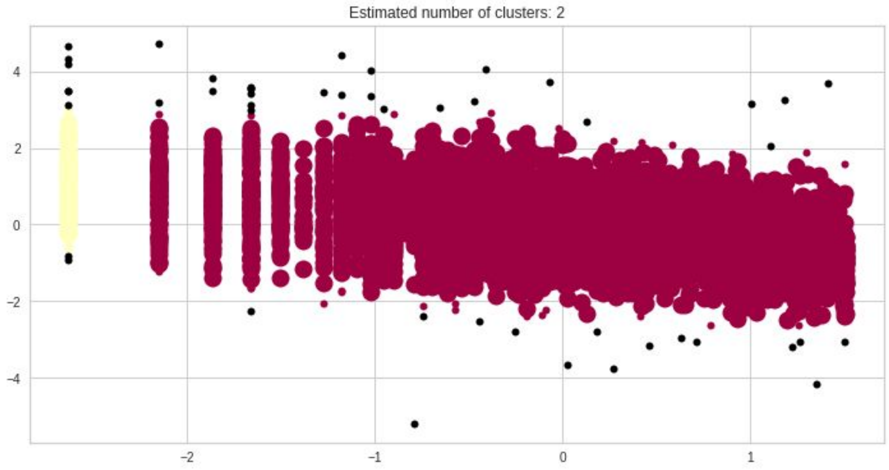


Clustering (Recency and Monetary)

K-means Clustering

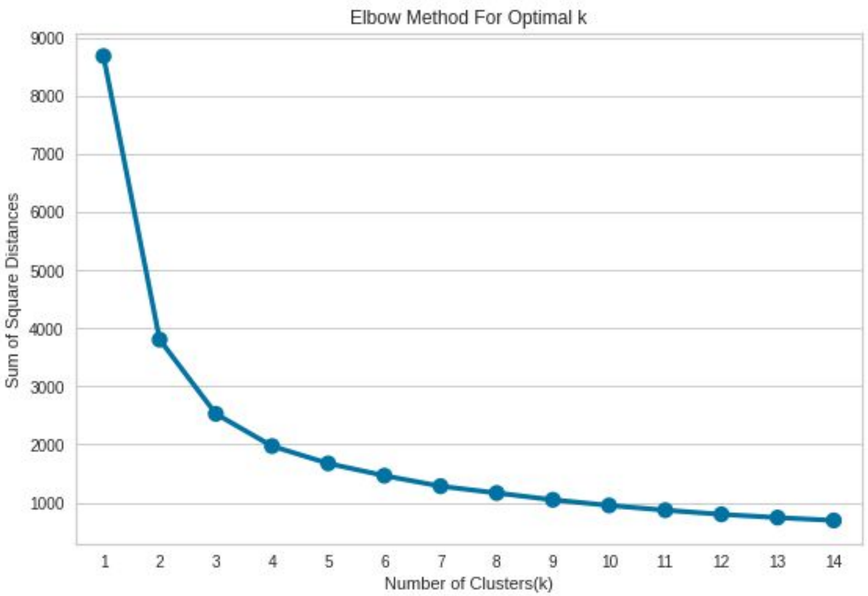


DBSCAN Algorithm



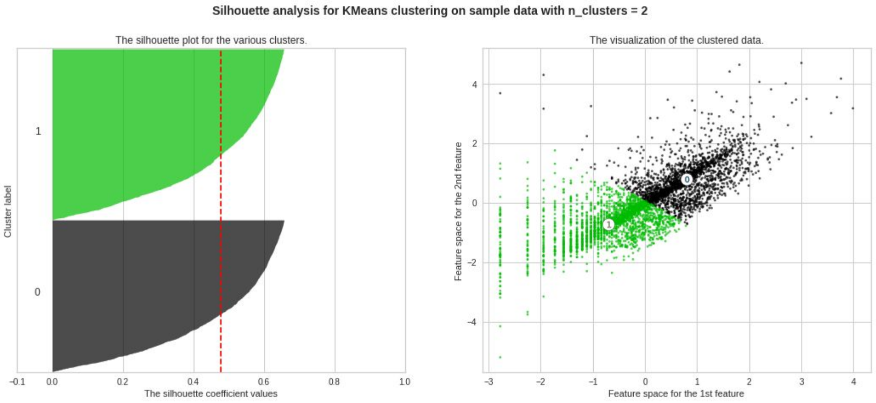
K-means Clustering (Frequency and Monetary)

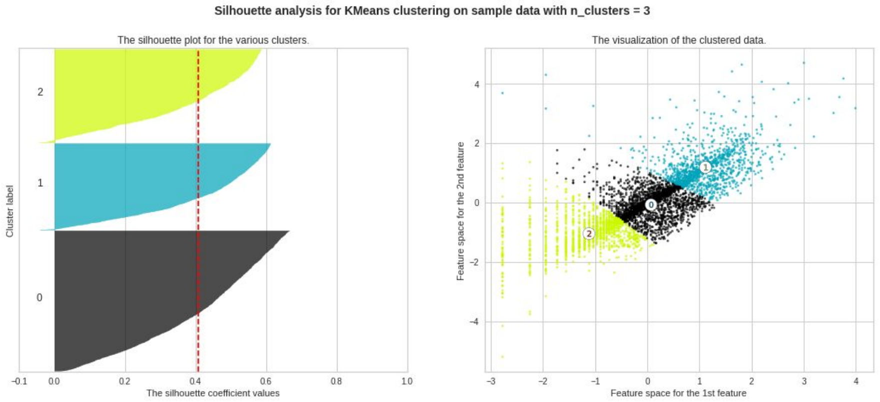
Finding the Optimal value of cluster using Elbow method and Silhouette Score.

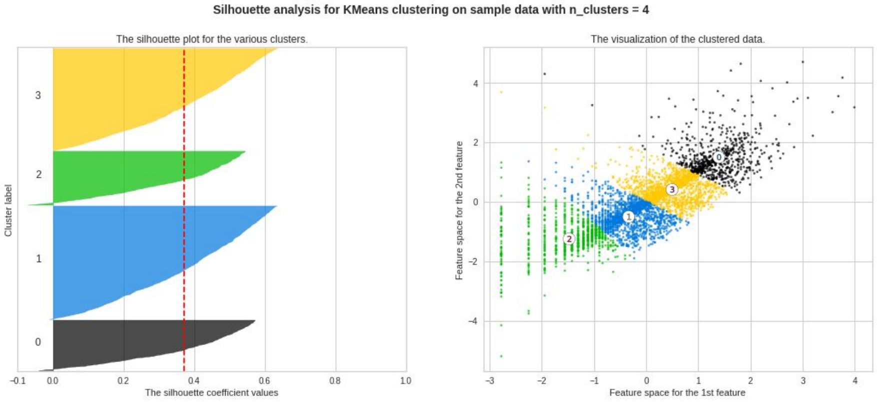


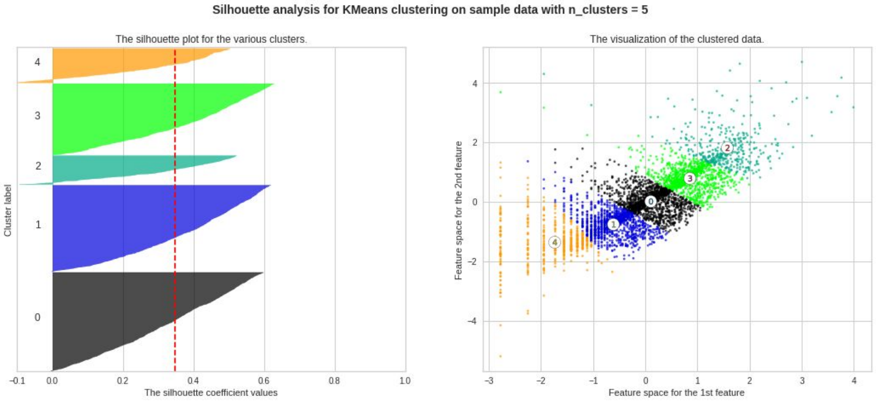
For n_clusters = 2 The average silhouette_score is : 0.478
For n_clusters = 3 The average silhouette_score is : 0.408
For n_clusters = 4 The average silhouette_score is : 0.372
For n_clusters = 5 The average silhouette_score is : 0.347
For n_clusters = 6 The average silhouette_score is : 0.361
For n_clusters = 7 The average silhouette_score is : 0.345
For n_clusters = 8 The average silhouette_score is : 0.354
For n_clusters = 9 The average silhouette_score is : 0.342
For n_clusters = 10 The average silhouette_score is : 0.361
For n_clusters = 11 The average silhouette_score is : 0.368
For n_clusters = 12 The average silhouette_score is : 0.356
For n_clusters = 13 The average silhouette_score is : 0.362
For n_clusters = 14 The average silhouette_score is : 0.359
For n_clusters = 15 The average silhouette_score is : 0.351

K-means Clustering (Frequency and Monetary)



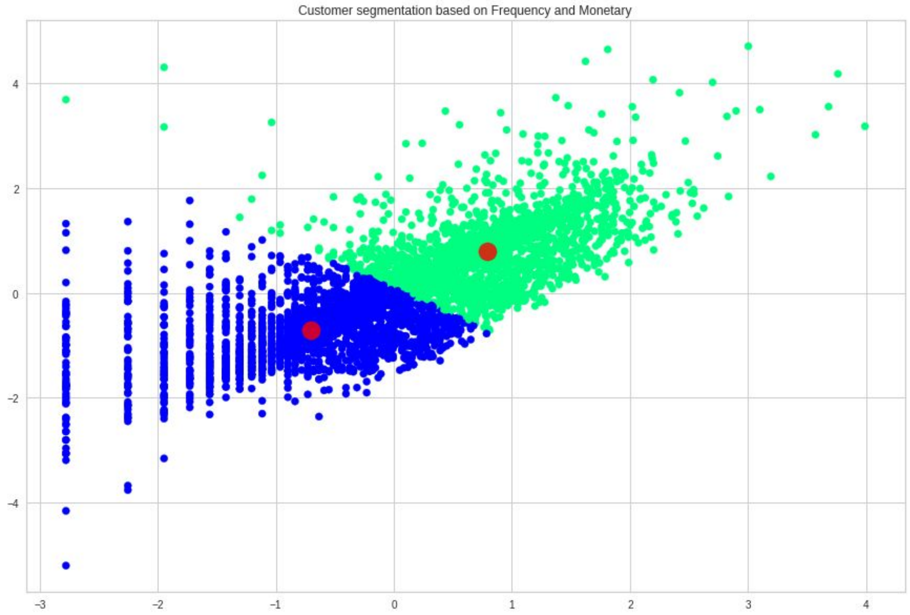




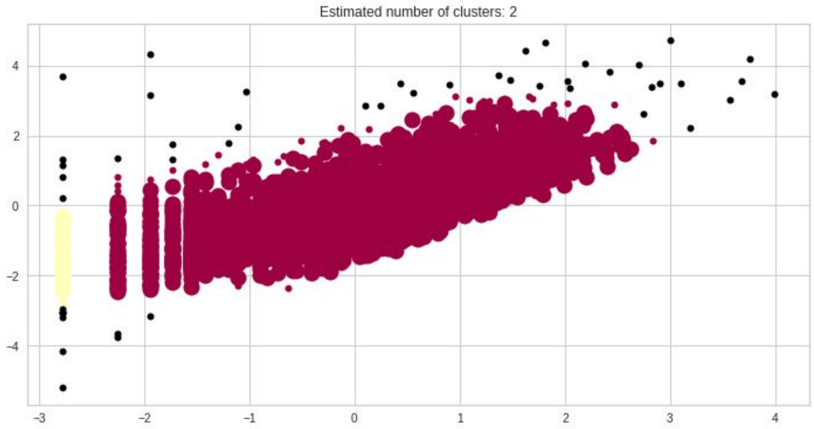


Clustering (Frequency and Monetary)

K-means Clustering

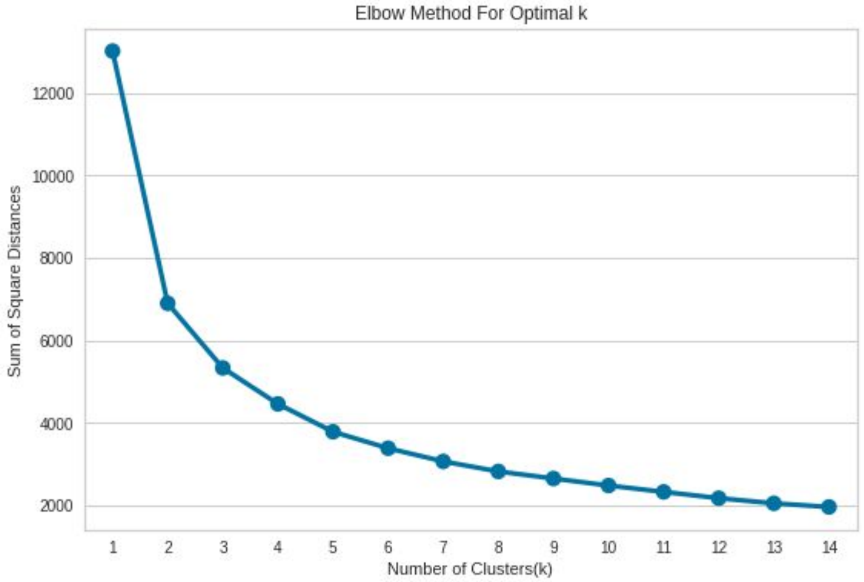


DBSCAN Algorithm



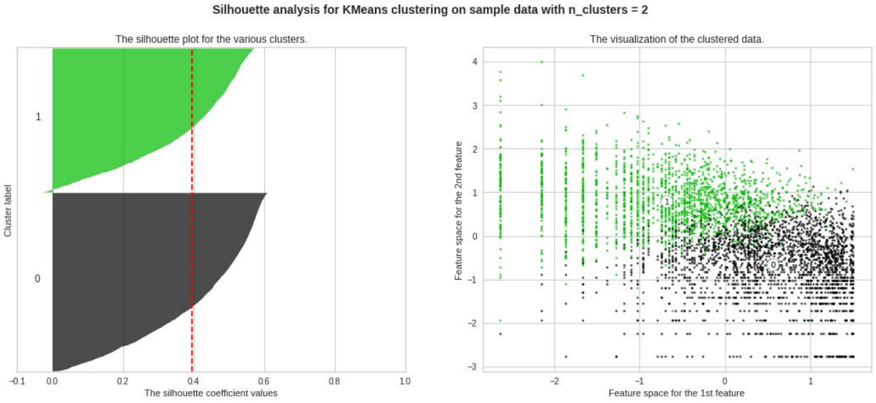
K-means Clustering (Recency , Frequency and Monetary)

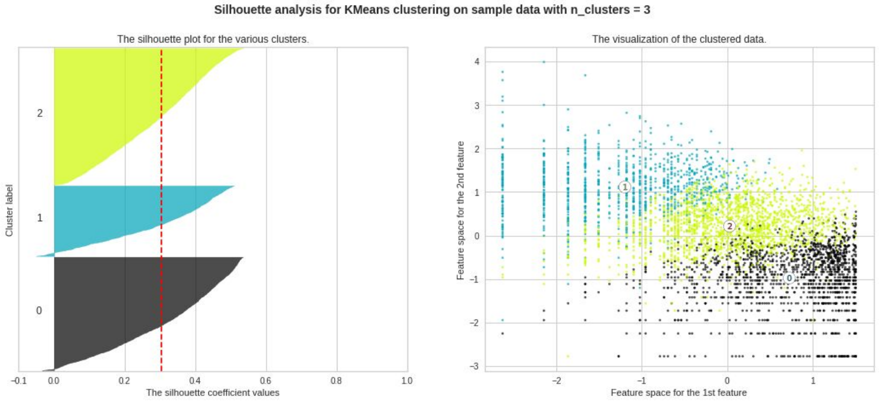
Finding the Optimal value of cluster using Elbow method and Silhouette Score.

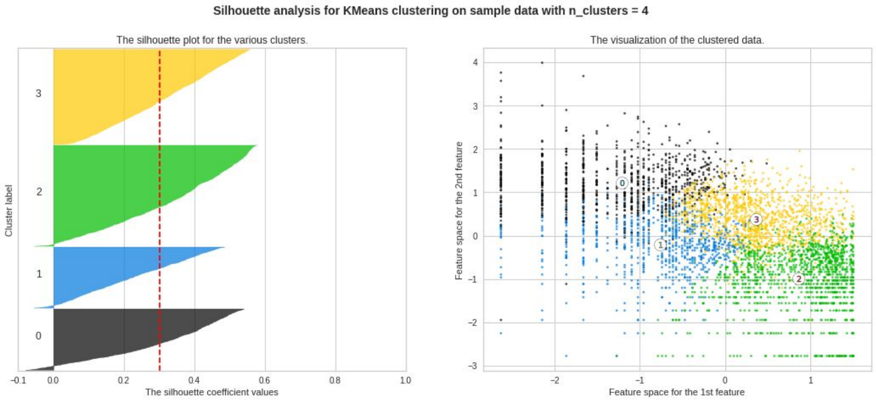


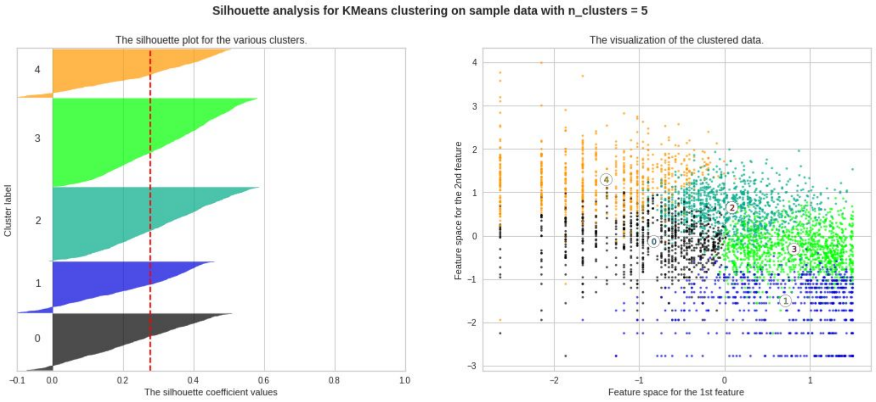
For n_clusters = 2 The average silhouette_score is : 0.395
For n_clusters = 3 The average silhouette_score is : 0.303
For n_clusters = 4 The average silhouette_score is : 0.303
For n_clusters = 5 The average silhouette_score is : 0.278
For n_clusters = 6 The average silhouette_score is : 0.277
For n_clusters = 7 The average silhouette_score is : 0.264
For n_clusters = 8 The average silhouette_score is : 0.261
For n_clusters = 9 The average silhouette_score is : 0.252
For n_clusters = 10 The average silhouette_score is : 0.261
For n_clusters = 11 The average silhouette_score is : 0.262
For n_clusters = 12 The average silhouette_score is : 0.264
For n_clusters = 13 The average silhouette_score is : 0.262
For n_clusters = 14 The average silhouette_score is : 0.261
For n_clusters = 15 The average silhouette_score is : 0.256

Contd...


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.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature

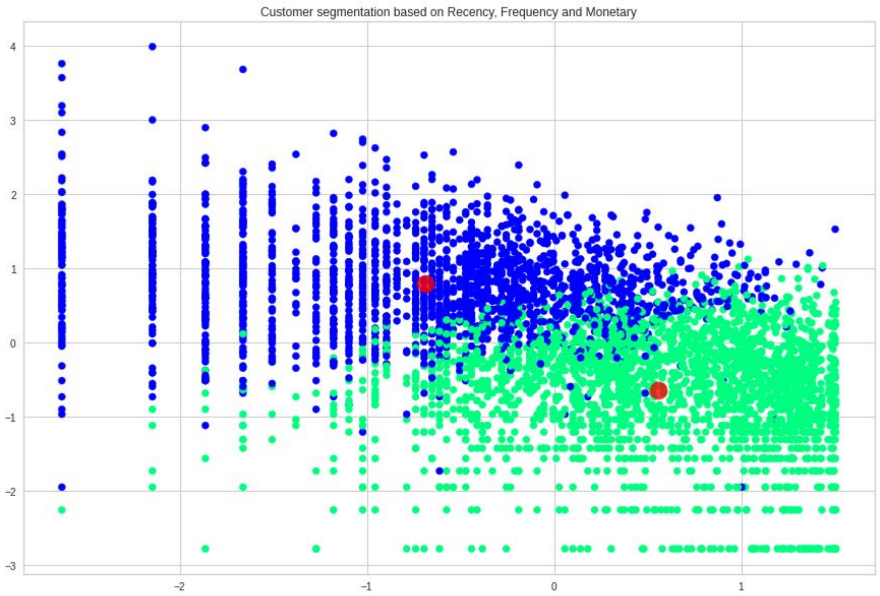

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature


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.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature

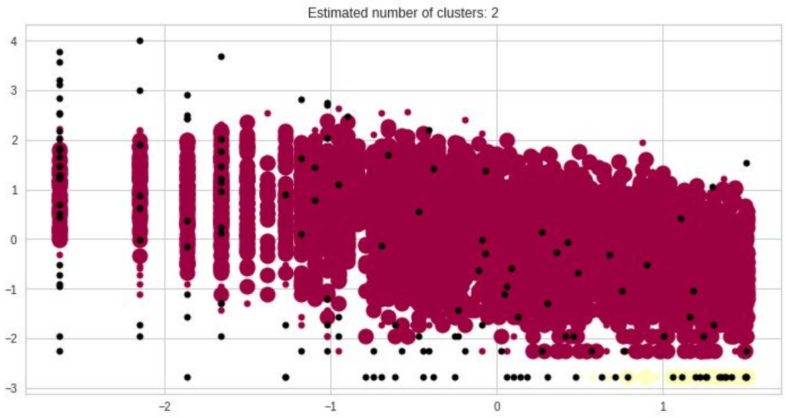

Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature

Clustering (Recency , Frequency and Monetary)

K-means Clustering

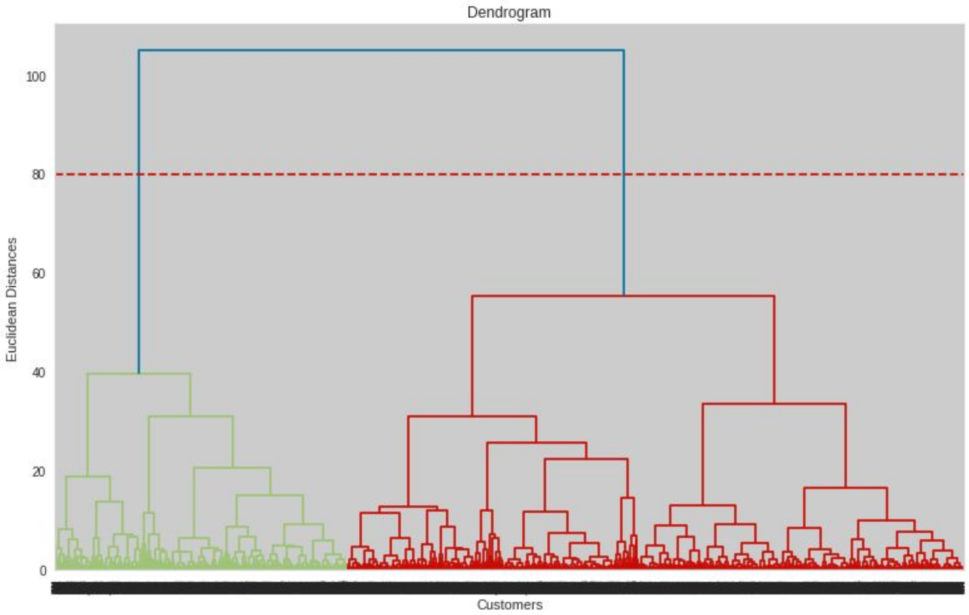


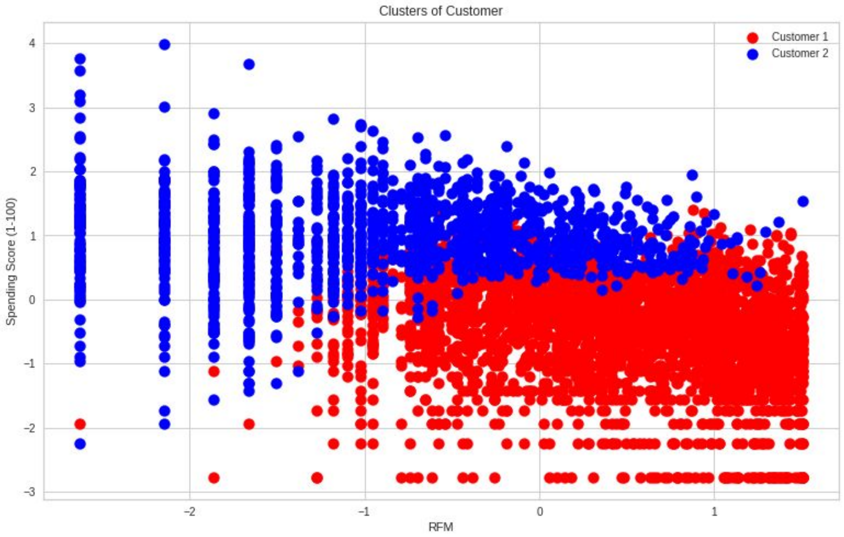
DBSCAN Algorithm



Hierarchical Clustering (Recency , Frequency and Monetary)

Optimal Number of clusters using Dendrogram is 2.





Conclusion

- 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										
Bronze	192.165501	19	374	15.062160	1	84	266.505704	3.75	1542.08	1287
Silver	87.606949	1	374	32.930510	1	123	788.400045	0.00	77183.60	921
Gold	47.848532	1	372	81.241886	1	521	1597.725141	120.03	168472.50	1294
Platinaum	13.761051	1	51	284.218638	43	7676	6870.541553	674.82	280206.02	837

- Bronze customers=1287 (very high recency but very low frequency and spendings).
- Silver customers=921 (high recency, low frequency and low spendings).
- Gold customers=1294 (good recency, frequency and monetary).
- Platinum customers=837 (less recency but high frequency and heavy spendings).

Conclusion

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

SL No.	Model_Name	Data	Optimal_Number_of_cluster
1	K-Means with silhouette_score	RM	2
2	K-Means with Elbow methos	RM	2
3	DBSCAN	RM	2
4	K-Means with silhouette_score	FM	2
5	K-Means with Elbow methos	FM	2
6	DBSCAN	FM	2
7	K-Means with silhouette_score	RFM	2
8	K-Means with Elbow methos	RFM	2
9	DBSCAN	RFM	2
10	Hierarchical clustering	RFM	2

Conclusion

	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
Cluster_based_on_freq_mon_rec										
0	31.119667	1	372	173.174298	3	7676	4032.232935	150.61	280206.02	1922
1	141.342573	1	374	24.779065	1	174	470.524697	1.00	77183.60	2417

- Above clustering is done with recency, frequency and monetary data (K-means Clustering) as all 3 together will provide more information.
- Cluster 0 has a high recency rate but very low frequency and monetary. Cluster 0 contains 1922 customers.
- Cluster 1 has a low recency rate but they are frequent buyers and spends very high money than other customers as mean monetary value is very high. Cluster 1 contains 2417 customers. This generates more revenue to the retail business.