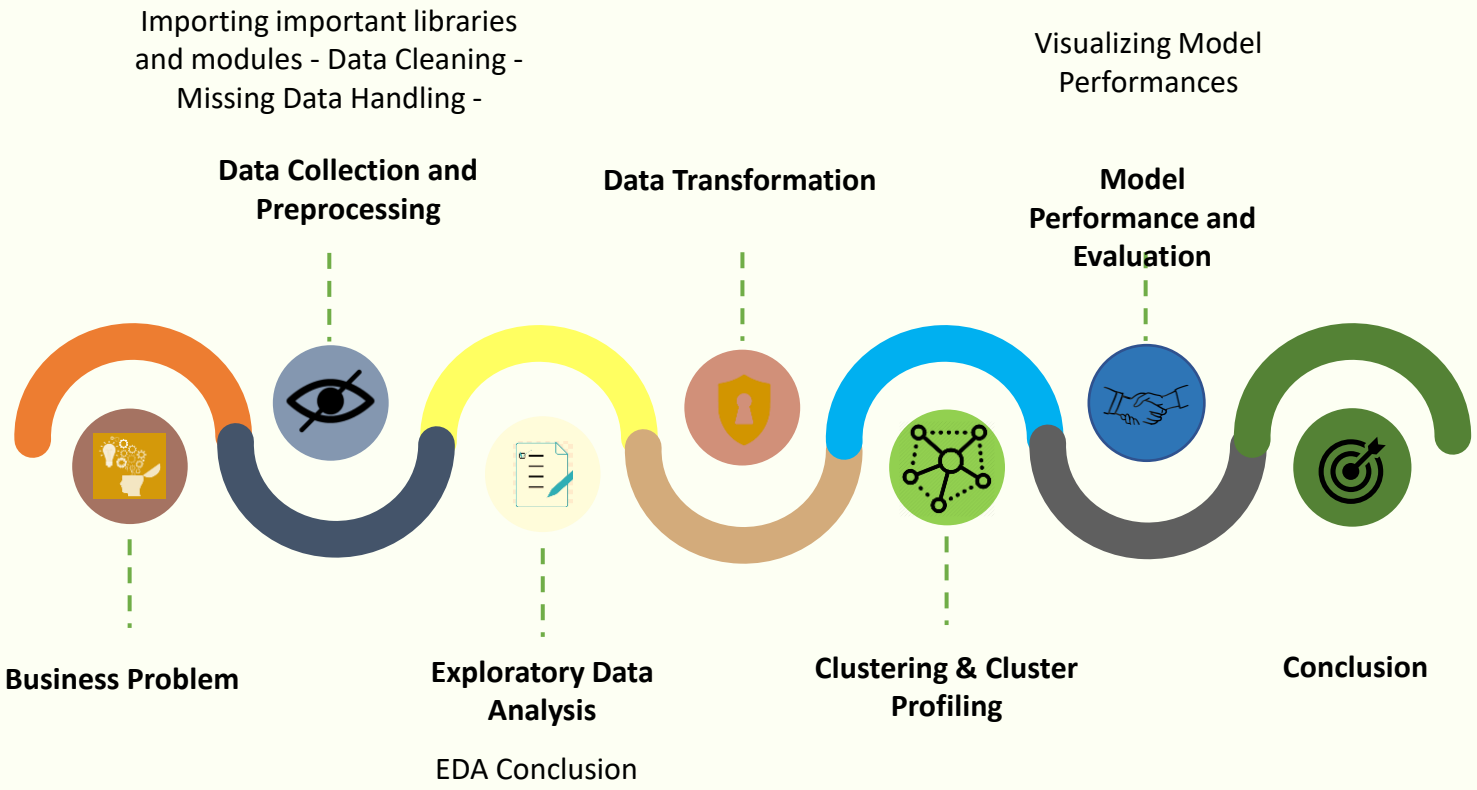
Capstone Project -4

Unsupervised Machine Learning – Clustering

Customer Segmentation

Jouher Lais Khan

Approach


Importing important libraries and modules - Data Cleaning - Missing Data Handling -
Data Collection and Preprocessing
Business Problem
Exploratory Data Analysis
EDA Conclusion
Data Transformation
Clustering & Cluster Profiling
Visualizing Model Performances
Model Performance and Evaluation
Conclusion

Problem Description

1. In this project, our task is 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.
2. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Description

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.Product price per unit in sterling.(Numeric)

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.

InvoiceNo : Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

Data Processing

1. The customer ID contains 133626 null values, as clustering should be done on the basis of customer, so these row will be removed.

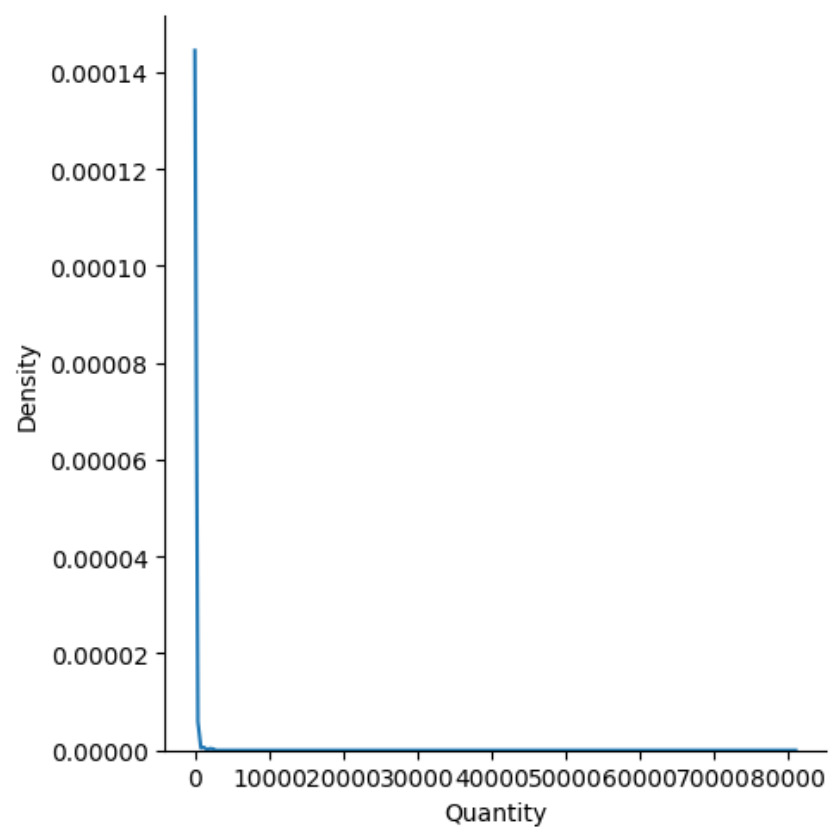
2. The invoice number that end with C is cancelled so we will remove these row and also the item which are cancelled was also bought so this should also be removed.

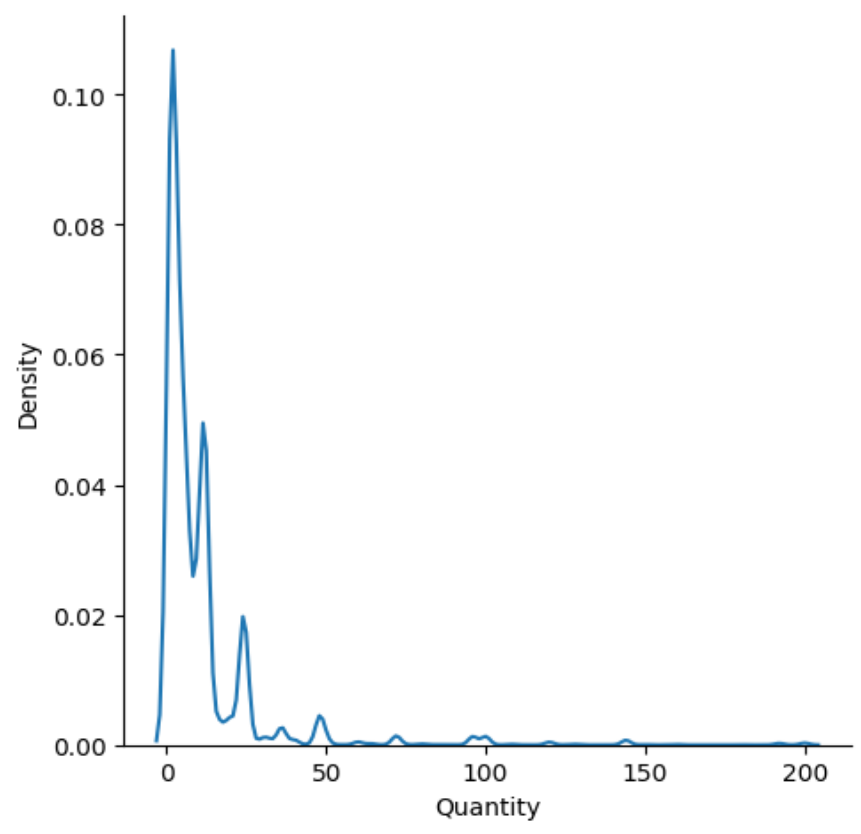
3. Stock code and Item description represent the same thing, but item description describes it more clearly. Hence, we can drop the stock code.

4. Quantity and Unit Price cannot be negative we will drop row where quantity or unit price is negative

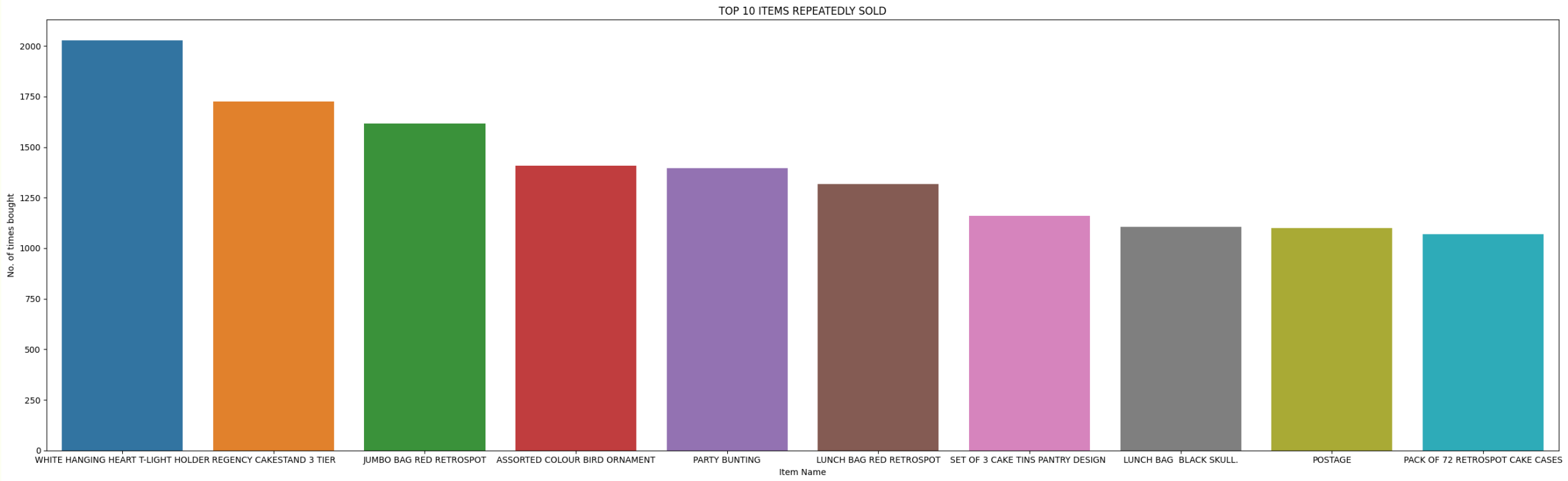
5. Data Contains no duplicates value

Exploratory Data Analysis
Distribution of Quantity

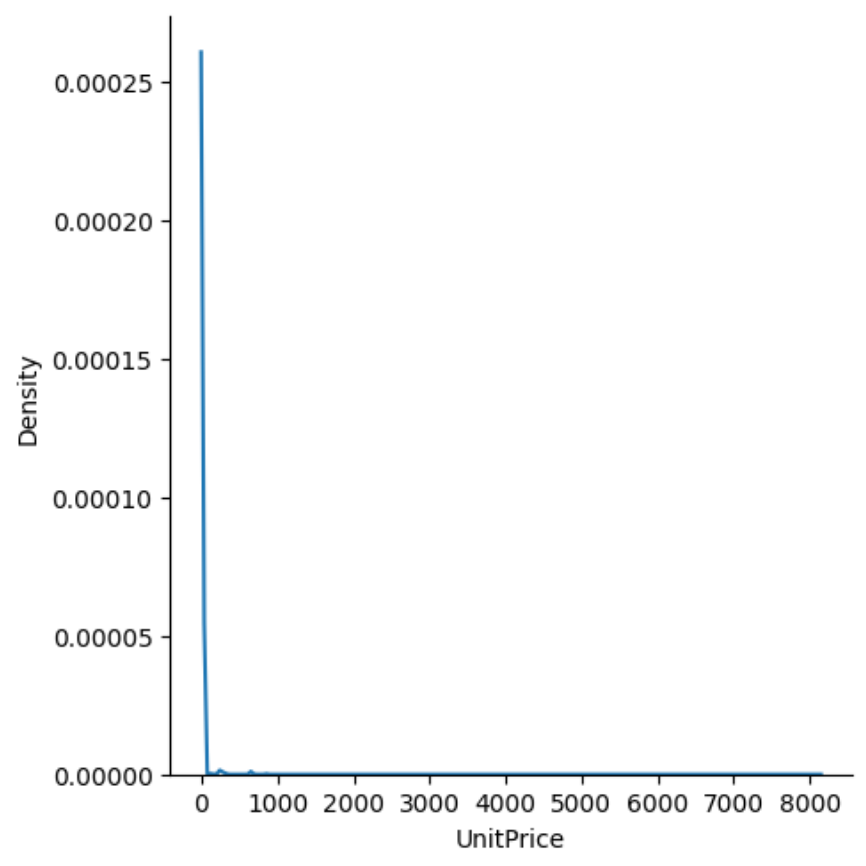


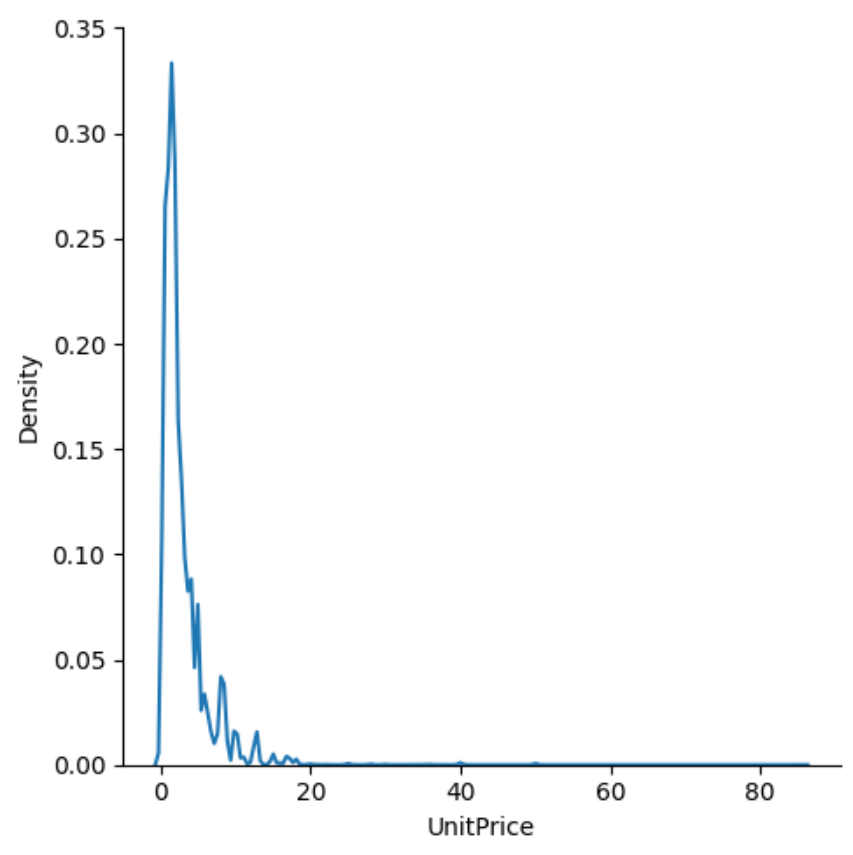


Top 10 most repeatedly sold items

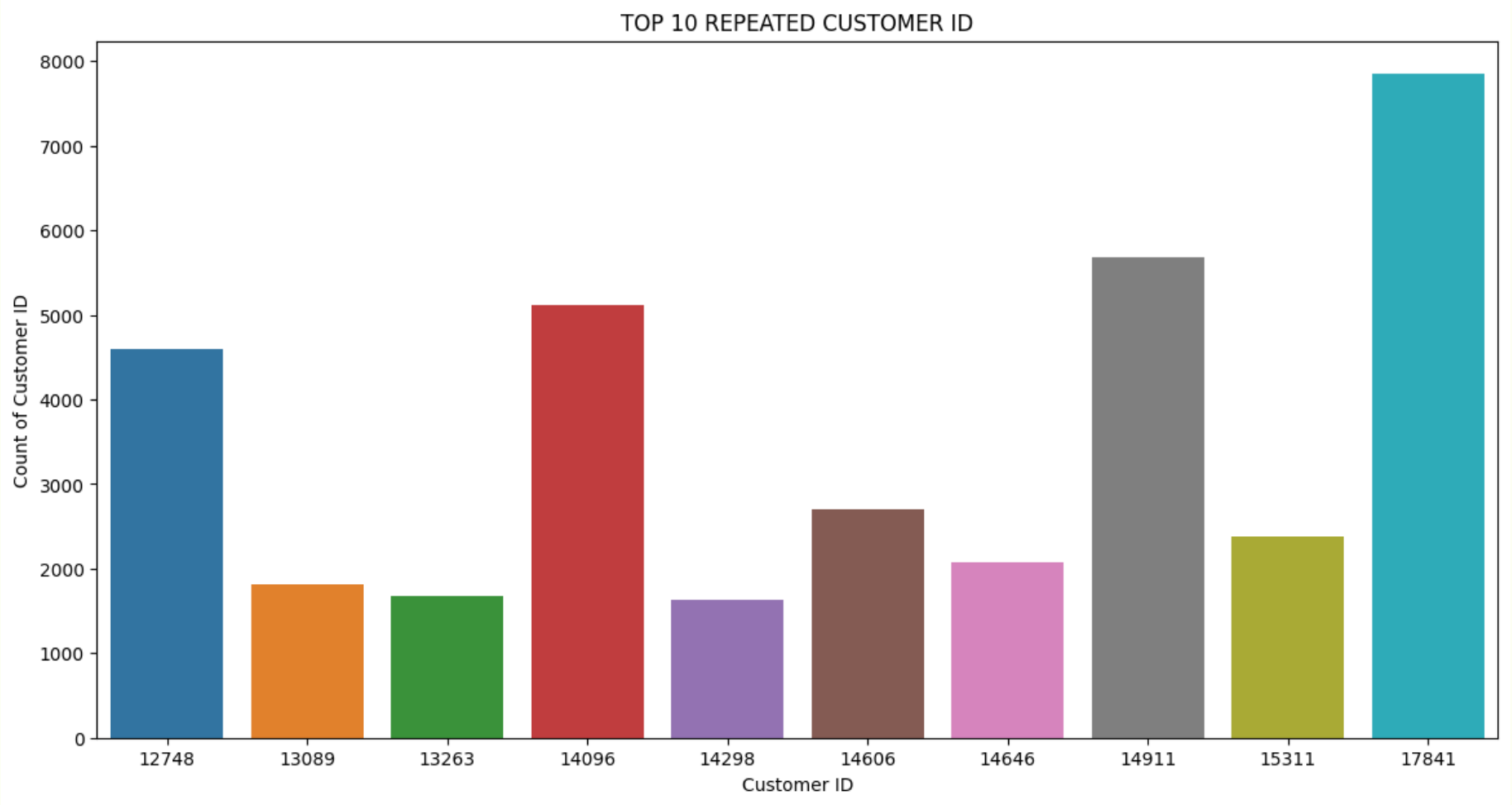


Distribution of the Unit Price

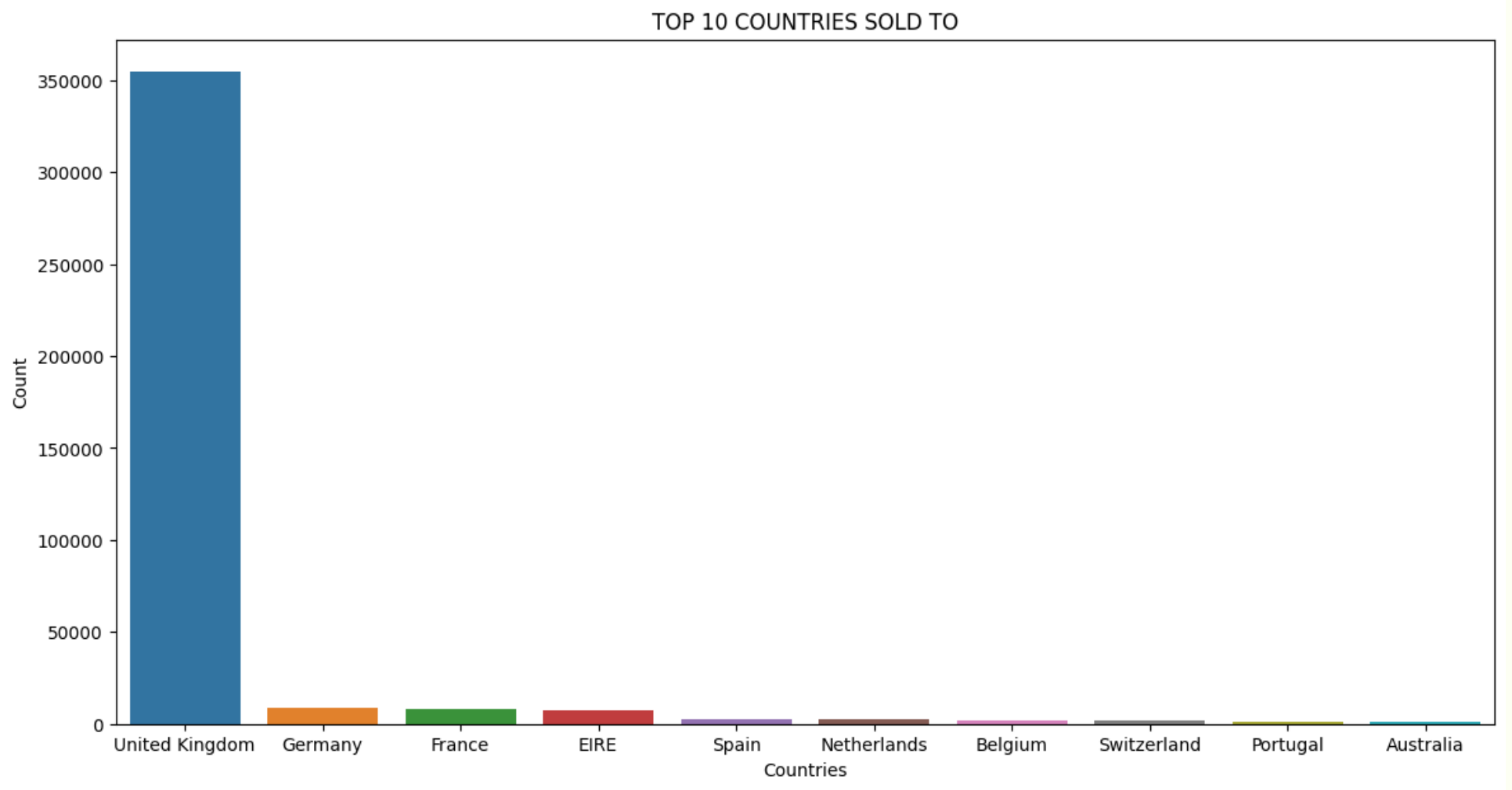




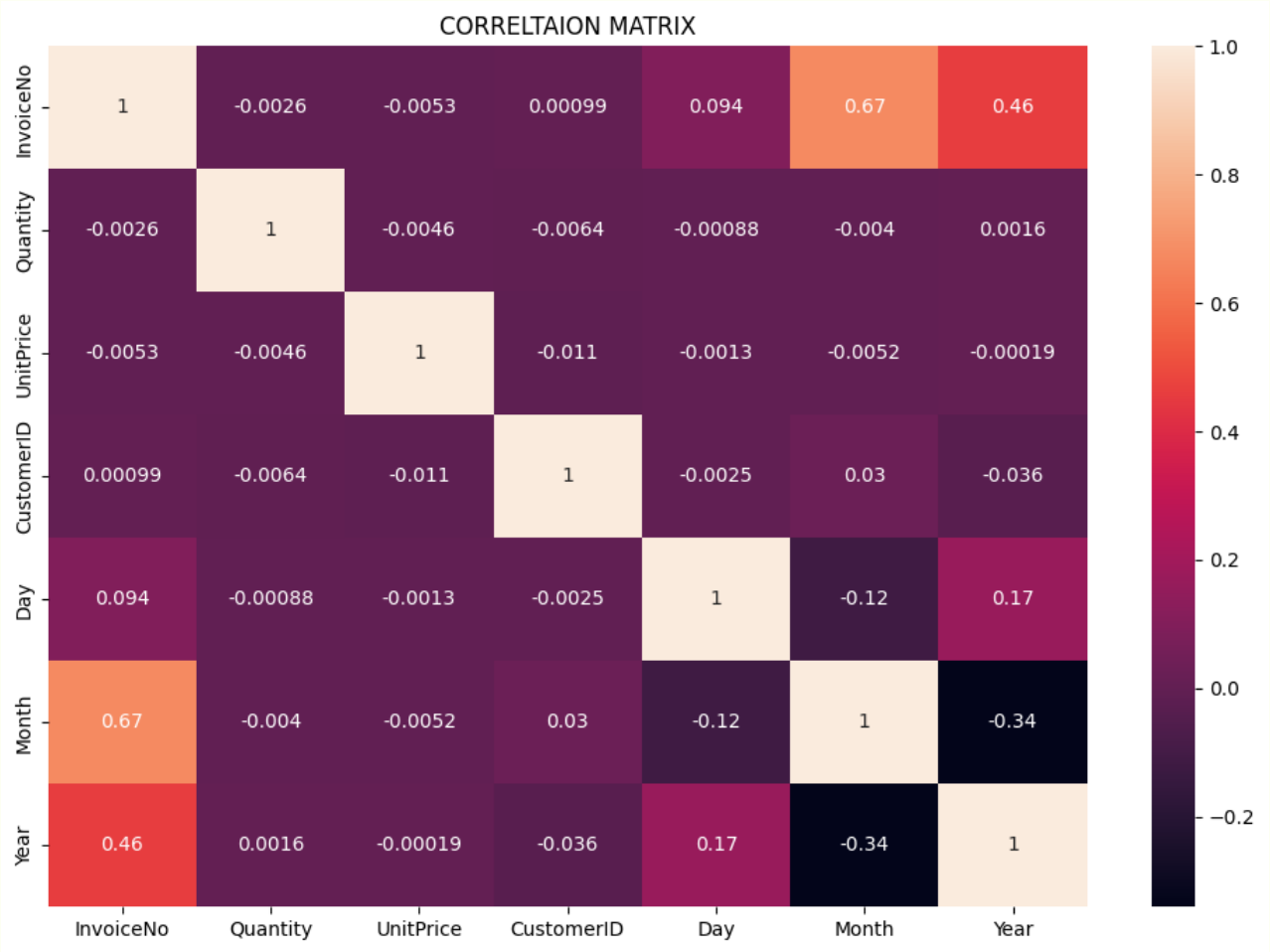
Customer who purchased most number of times



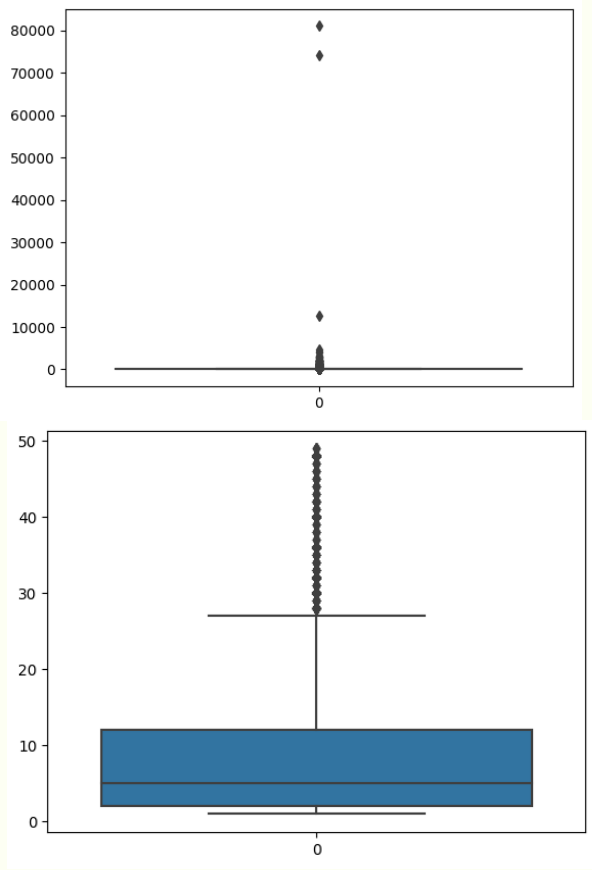
Countries that were sold different items the most



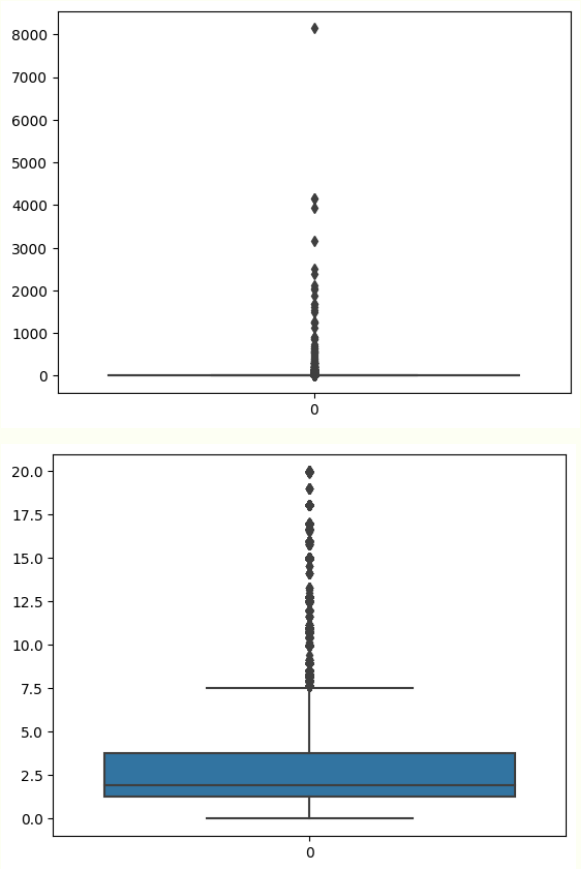
Correlation Map



Outliers Removal



Quantity



Unit Price

Data Transformation

1. For the purpose of this project - Recency, Frequency and Monetory(RFM) analysis shall be conducted.

2. On the basis of these 3 factors, customers can be classified into different groups.

3. They can be catered by the business depending on the cluster they belong to.

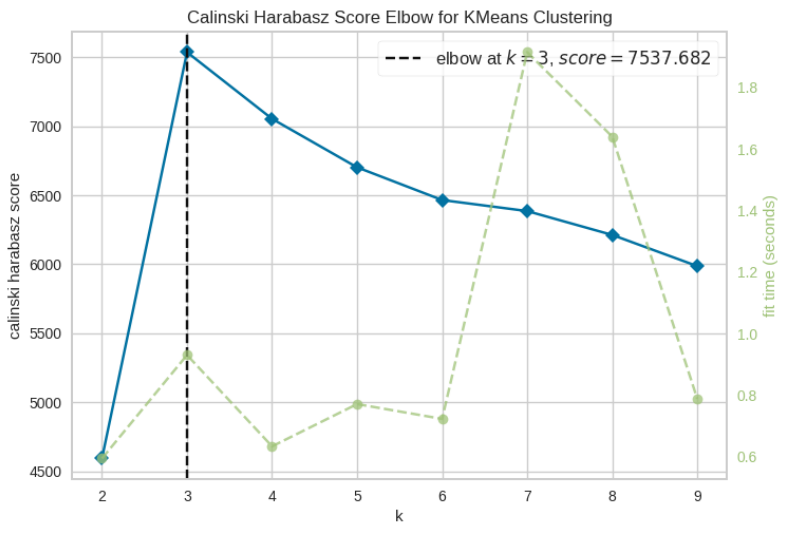
Feature Encoding

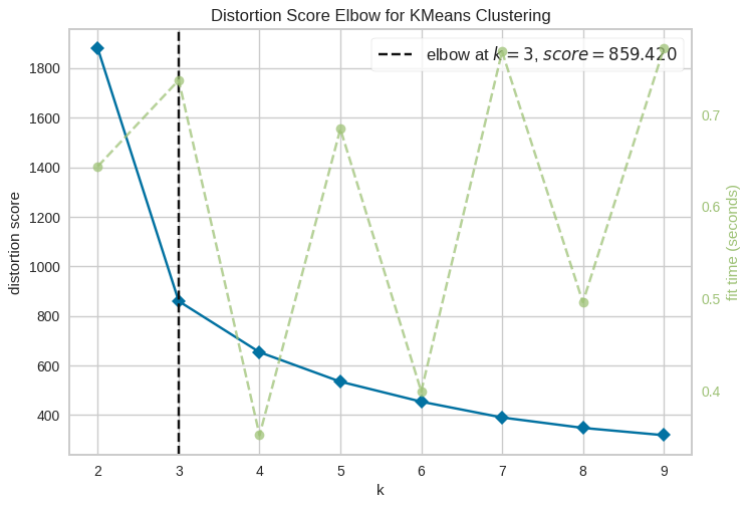
1. Creating copy of Data Frame for Modelling.
2. Creating list of final features which will be used in modelling.
3. Scaling the dataset

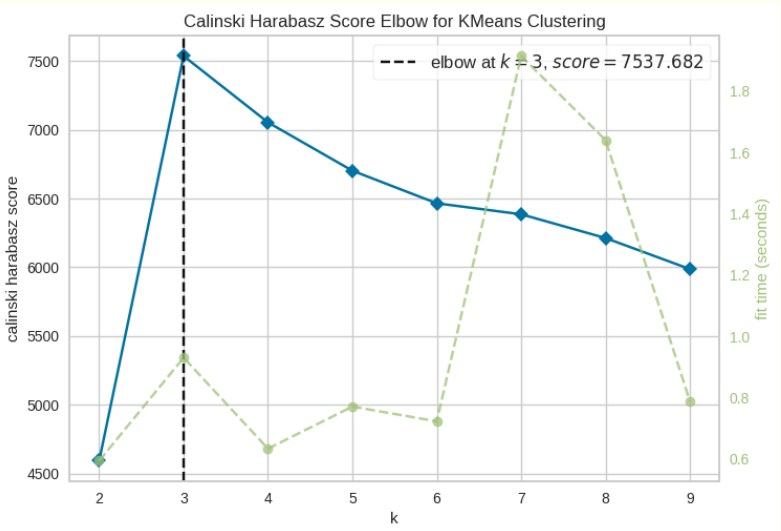
Models Implemented

1. K-Means with silhouette score
2. K-Means with Elbow methos
3. K-Means with distortion methos
4. K-Means with calinski harabasz
5. Hierarchical clustering(ward) with silhouette score
6. Hierarchical clustering(ward) with dendogram and euclidean distance 40

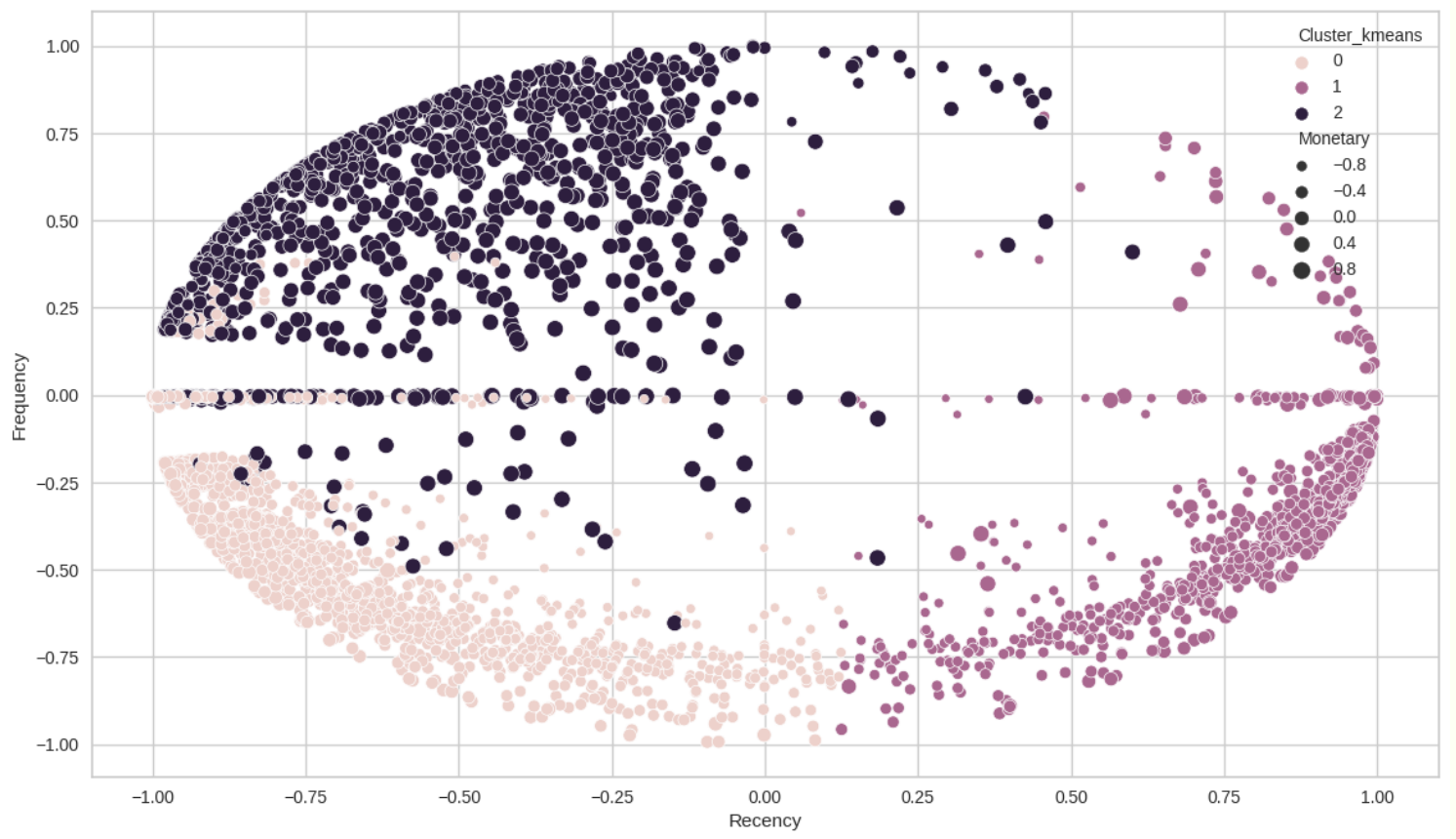
Kmeans Clustering







Kmeans Clusters



Cluster Profiling

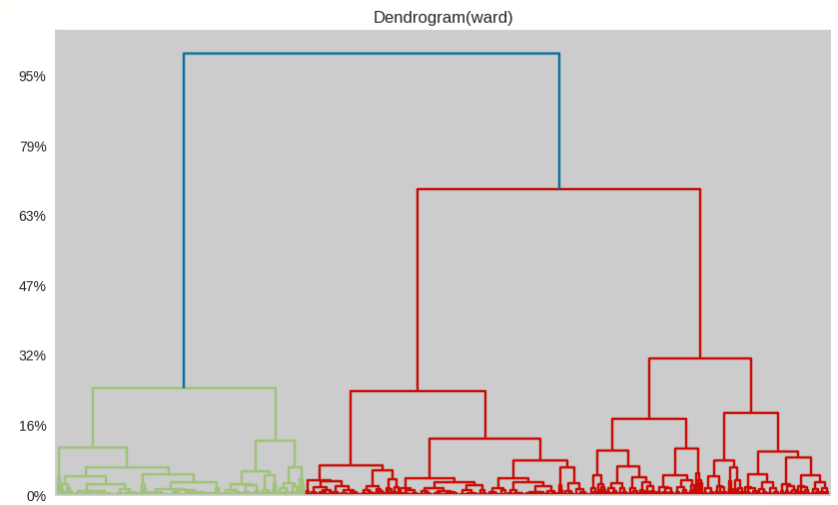
We can associate each clusters by

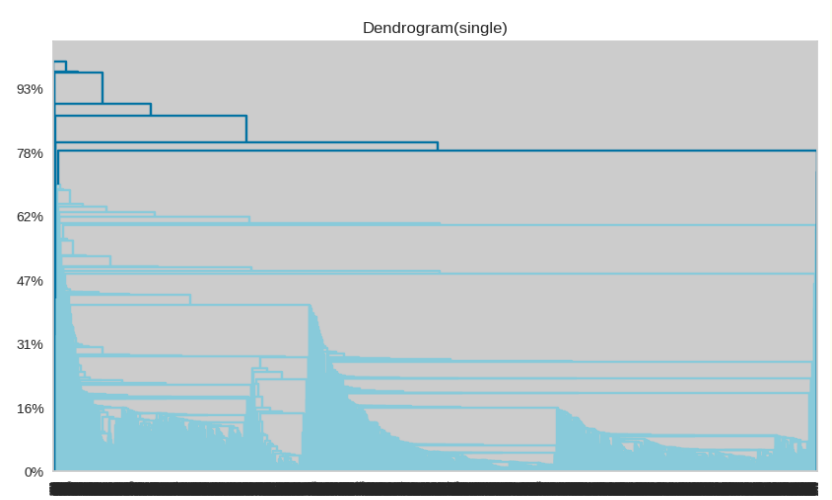
1. Cluster 0 comprises of customers who are moderately recent, frequent and contribute an average amount to sales.

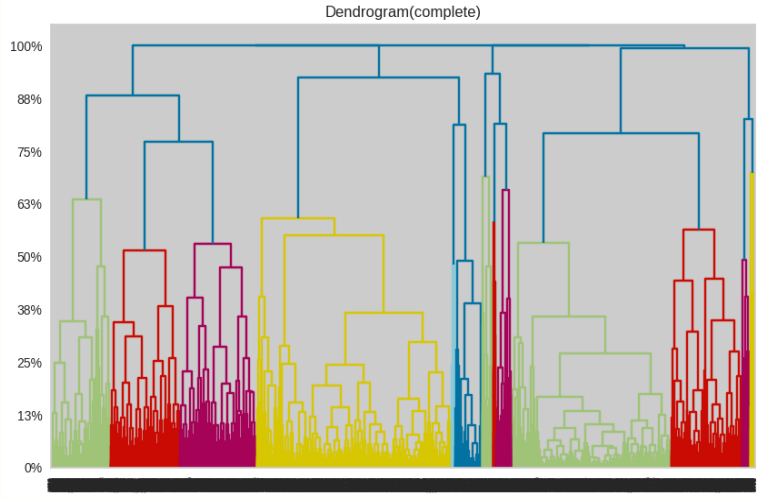
2. Cluster 1 comprises of customers who made purchases a long time ago and purchase infrequently and contribute the least towards the sales of the company.

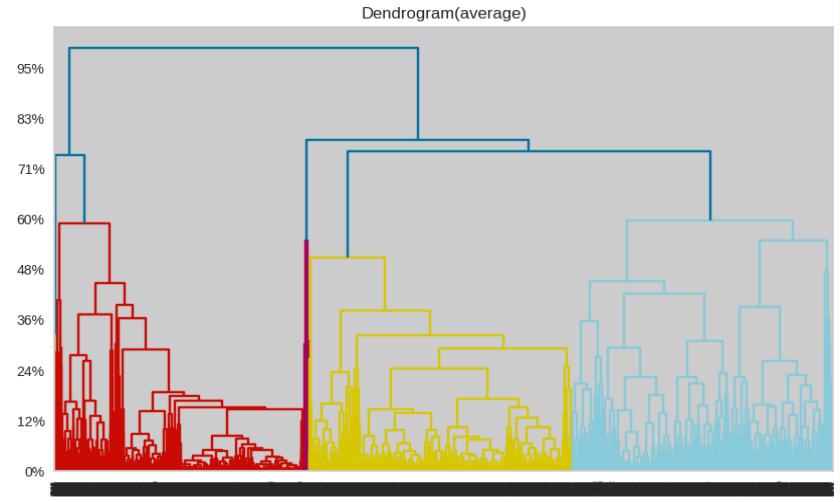
3. Cluster 2 comprises of customers who are very recent, frequent and also contribute largely to the sales.

Hierarchical Clustering

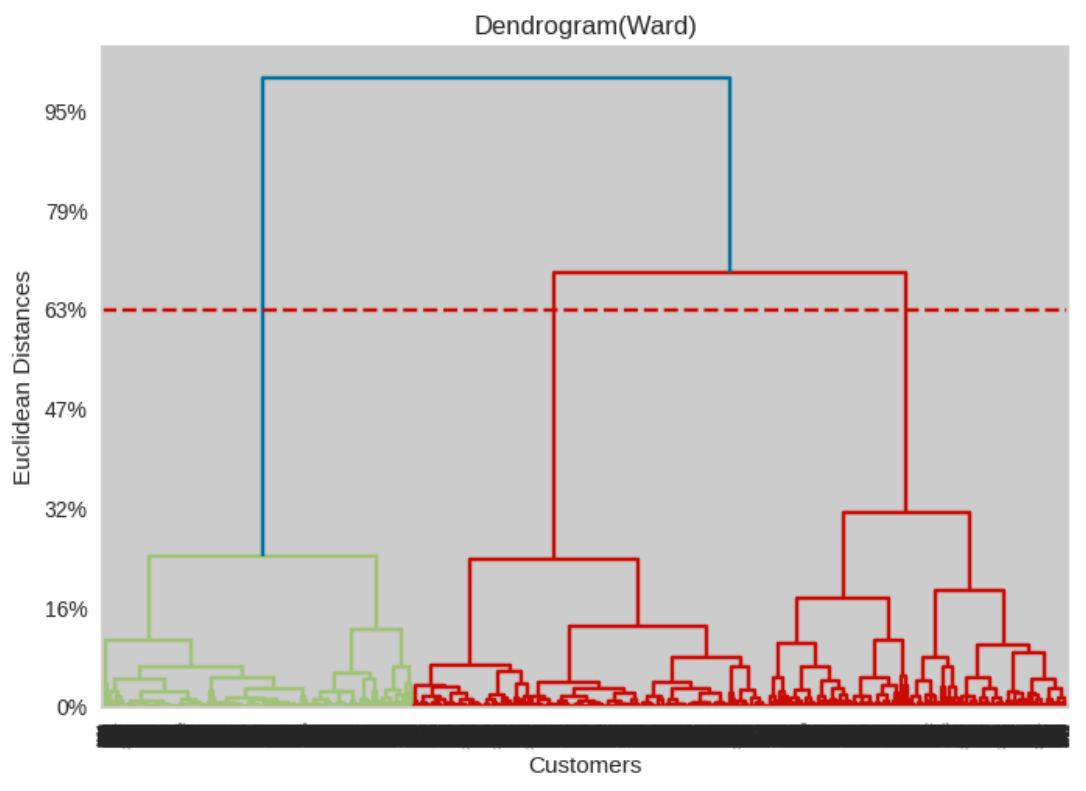






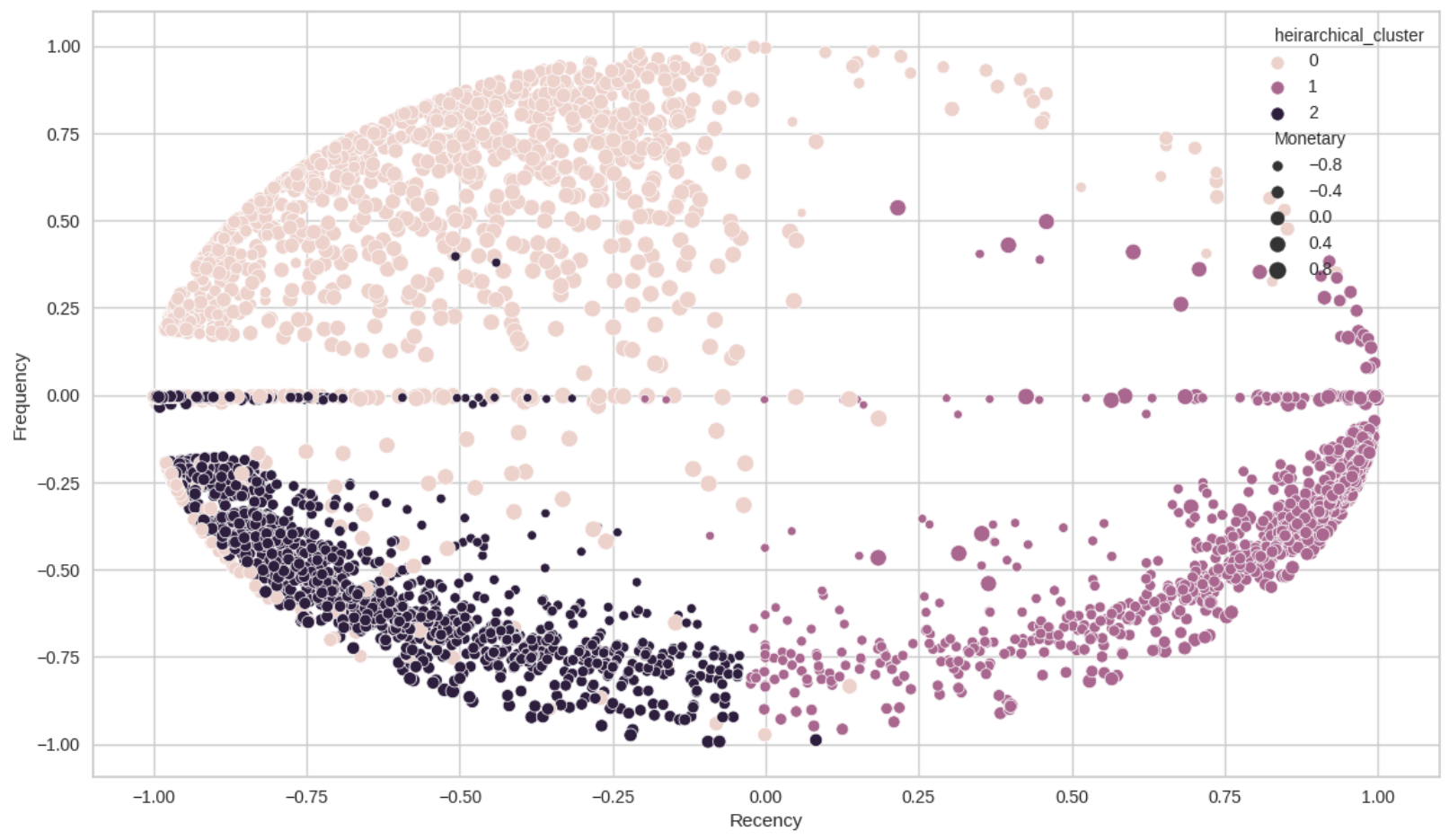


Hierarchical Clustering



Clusters	Silhouette Score	
0	2.0	0.518650
1	3.0	0.549241
2	4.0	0.500661
3	5.0	0.436785
4	6.0	0.399488
5	7.0	0.401470
6	8.0	0.401609
7	9.0	0.367218
8	10.0	0.355250

Hierarchical Clustering



Cluster Profiling

We can associate each clusters by

1. Cluster 0 comprises of customers who are very recent, frequent and also contribute largely to the sales.

2. Cluster 1 comprises of customers who made purchases a long time ago and purchase infrequently and contribute the least towards the sales of the company.

3. Cluster 2 comprises of customers who are moderately recent, frequent and contribute an average amount to sales.

Model Performance

Model_Name	Data	Optimal_Number_of_cluster
K-Means with silhouette_score	RFM	3
K-Means with Elbow methos	RFM	3
K-Means with distortion methos	RFM	3
K-Means with calinski_harabasz	RFM	3
Hierarchical clustering(ward) with silhoutte_score	RFM	3
Hierarchical clustering(ward) with dendograms and eulidean distance 40	RFM	3

Clusters	Silhouette Score	
0	2.0	0.519950
1	3.0	0.580663
2	4.0	0.498186
3	5.0	0.485991
4	6.0	0.479926
5	7.0	0.433104
6	8.0	0.425016
7	9.0	0.419886
8	10.0	0.397676

Kmeans Silhouette Score

Clusters	Silhouette Score	
0	2.0	0.518650
1	3.0	0.549241
2	4.0	0.500661
3	5.0	0.436785
4	6.0	0.399488
5	7.0	0.401470
6	8.0	0.401609
7	9.0	0.367218
8	10.0	0.355250

Hierachical(ward) Silhouette Score

Conclusions of Modelling

1. **Exploratory Data Analysis (EDA):** The company offers medium-to-low quantities of single items at a cheaper unit cost. More orders of various items were placed in the previous quarter, with the UK placing the highest orders overall. The "Paper craft little Berdie" was the best-selling item in terms of amount sold.

2. **Data Transformation:** For each customer ID in this section, Monetary , frequency, and recency analysis was produced. These three elements are essential part of customer segmentation.

3. **Clustering Kmeans:** In this step, the optimal number of clusters was ascertained by using silhouette analysis and the elbow technique. The optimal clusters were found to be three.

4. **Clustering Hierarchical:** Using the silhouette analysis and dendograms distance, the ideal number of clusters is three.

5. **Cluster Profiling:** The three groups clusters were identified as high value and loyal customers, average value and frequent customers, and low value and infrequent customers.

6. **Model Performance:** The best performance is given by kmeans with 3 clusters and in Hierarchical clustering best model is given by ward method with 3 clusters

Thank You