Capstone Project - 4

Online Retail Customer Segmentation

Swapnil Patil

Problem Statement

In this project, our task is to identify major customer segments on a transactional data set which contains all the transactions for a UK-based and registered non-store online retail.

The main purpose of this project is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively.

What is Customer Segmentation

Customer segmentation is the process of separating customers into groups on the basis of their shared behavior or other attributes.

The groups should be homogeneous within themselves and should also be heterogeneous to each other.

Customer segmentation can help a company to understand how its customers are alike, what is important to them, and what is not.

The overall aim of this process is to identify high-value customer base i.e. customers that have the highest growth potential or are the most profitable.

Such information can be used to develop personalized relevant content for different customer bases. Many studies have found that customers appreciate such individual attention and are more likely to respond and buy the product.

Data Overview

- The data used in this project is a transactional dataset which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based online retail store.
- The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.
- The transaction dataset of this online retail store has 8 variables as shown below

Attributes Information

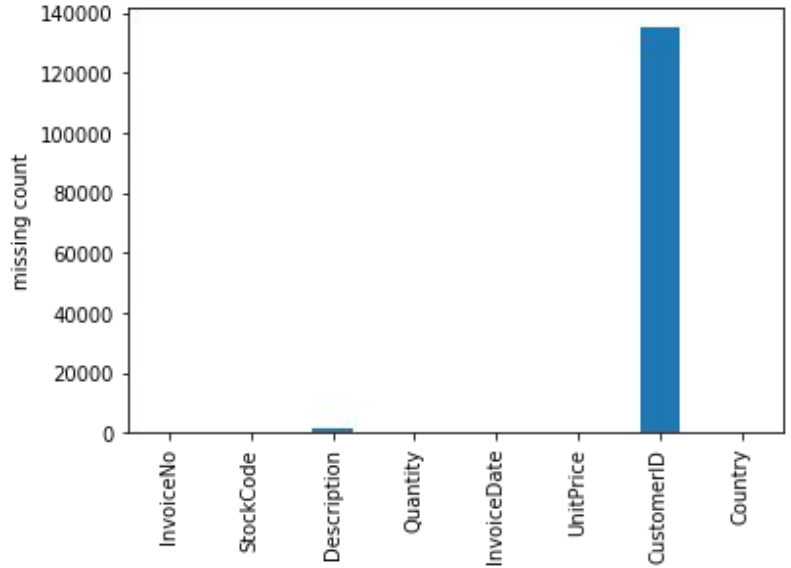
1. **InvoiceNo:** Invoice number.If this code starts with letter ‘c’, it indicates a cancellation.
2. **StockCode:** Product (item) code, a 5-digit integral number uniquely assigned to each distinct product.
3. **Description:** Product (item) name.
4. **Quantity:** The quantities of each product (item) per transaction.
5. **InvoiceDate:** Invoice Date and time, the day and time when each transaction was generated.
6. **UnitPrice:** Unit price, Product price per unit in sterling.
7. **CustomerID:** Customer number, a 5-digit integral number uniquely assigned to each customer.
8. **Country:** Country name.

Data Exploration

- The dataset has 541909 rows (transactions) and 8 columns.
- Dataset has four categorical features - InvoiceNo, StockCode, Description and Country. It also has one Datetime feature.

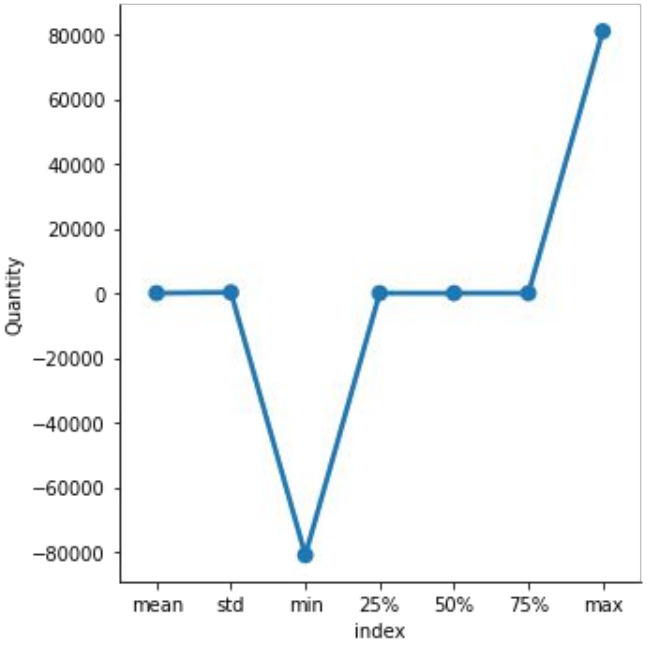
Data Inspection & Cleaning

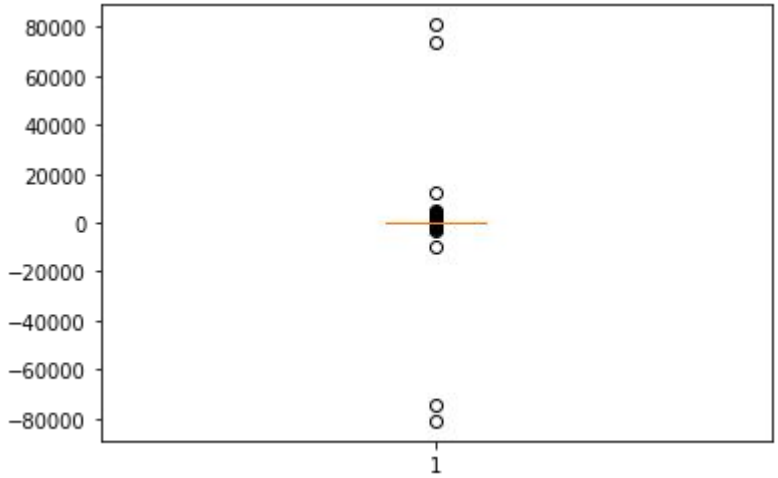
- The data contained 5268 duplicate entries.
- The missing values present in the Description and CustomerID columns.
- Dropped missing CustomerID values which are around 25%.
- Dropped remaining missing values

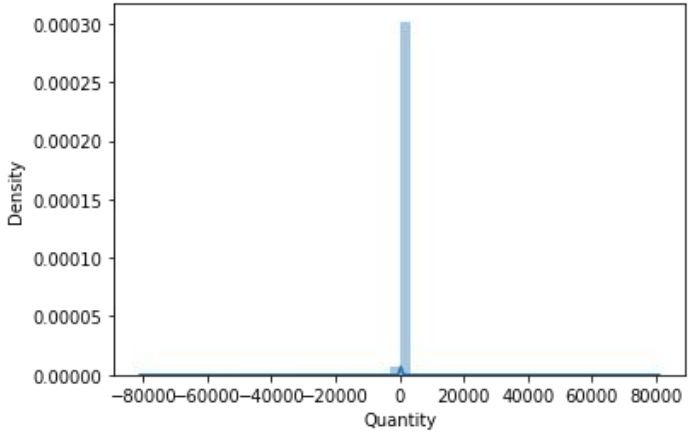


Exploratory Data Analysis - Columns

Quantity Column

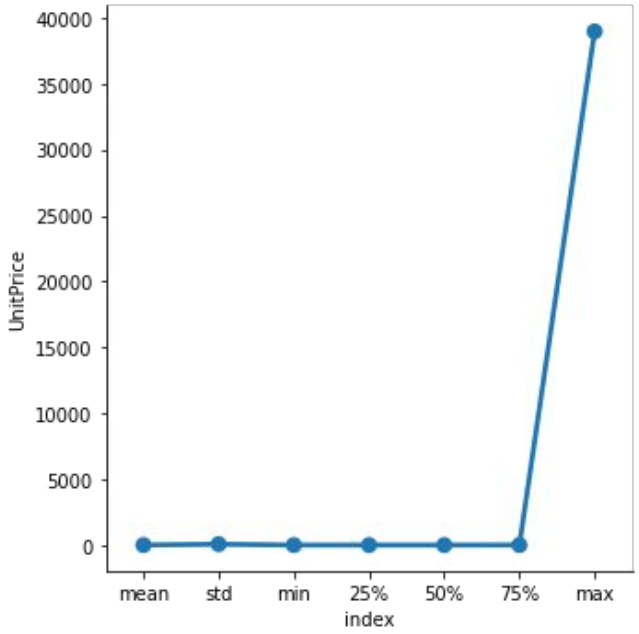


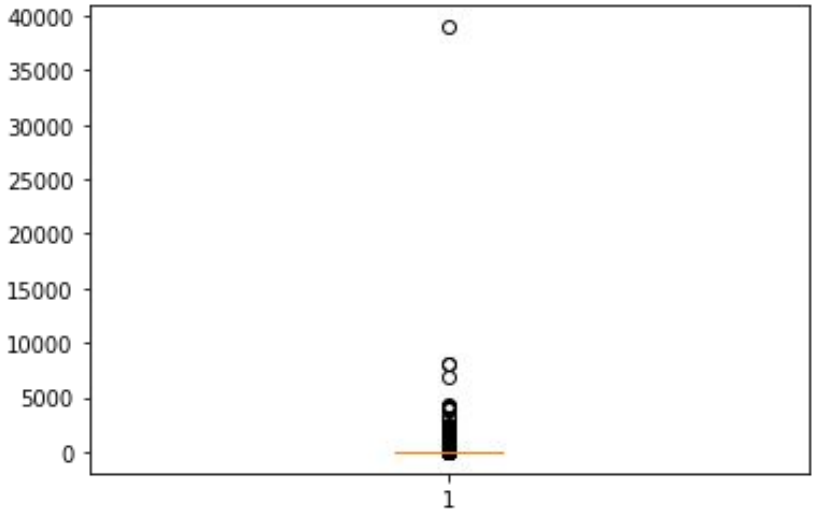


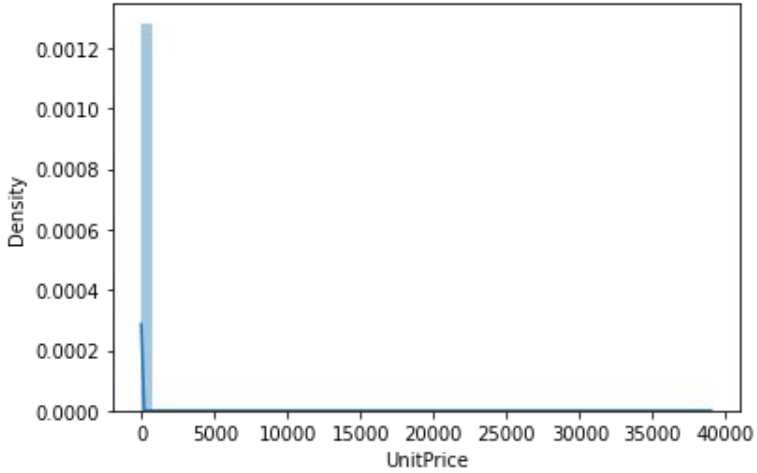


There are 8872 cancelled orders in our dataset

UnitPrice Column

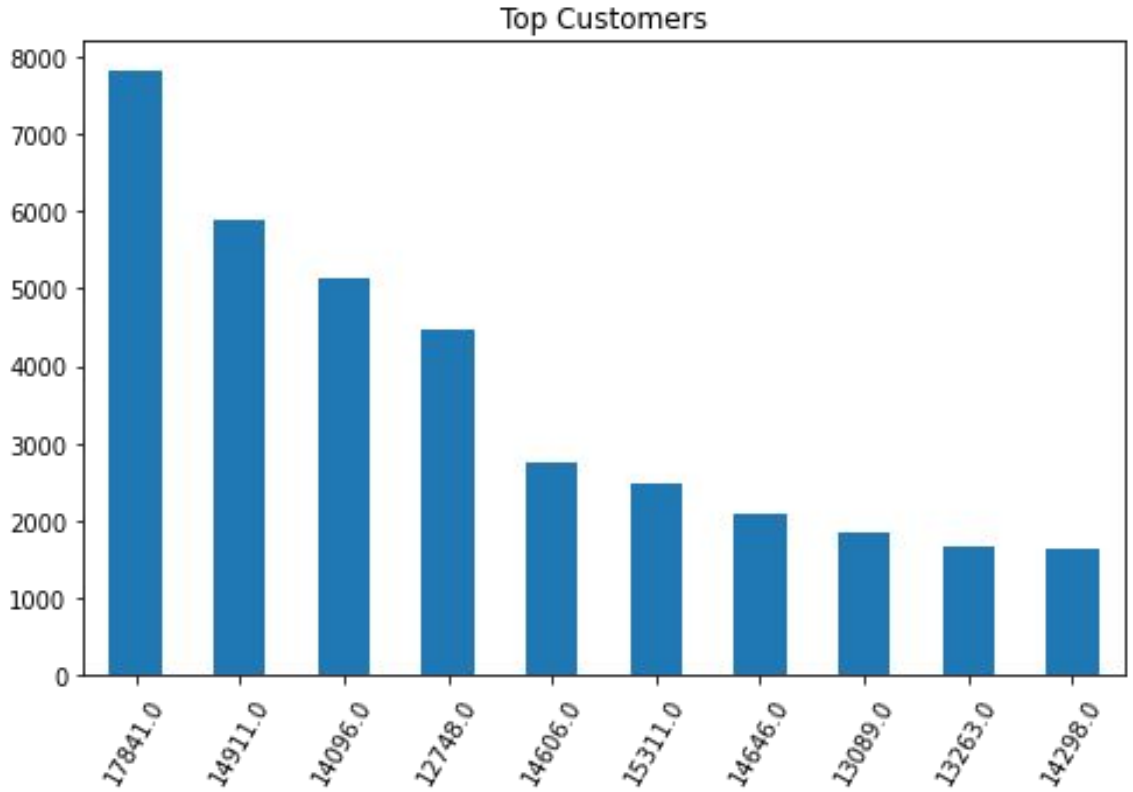






The order with max unitprice is a cancelled order

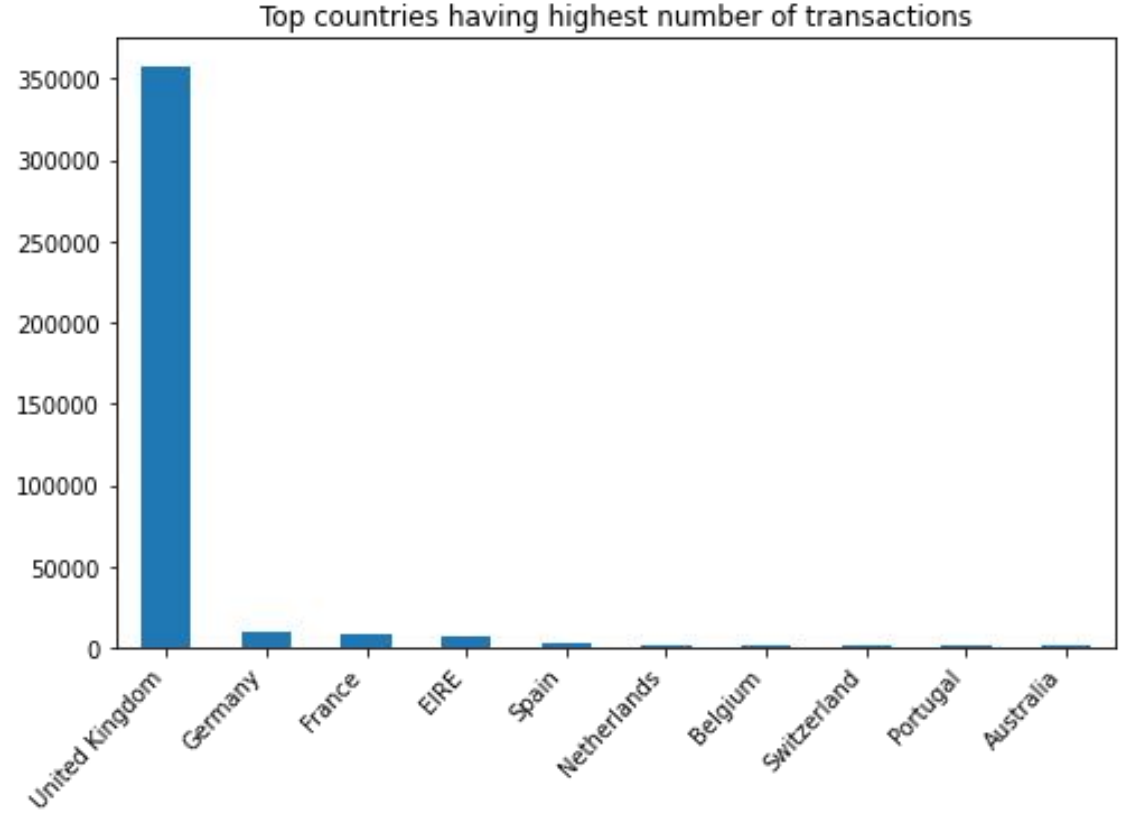
Exploring CustomerID and InvoiceNo

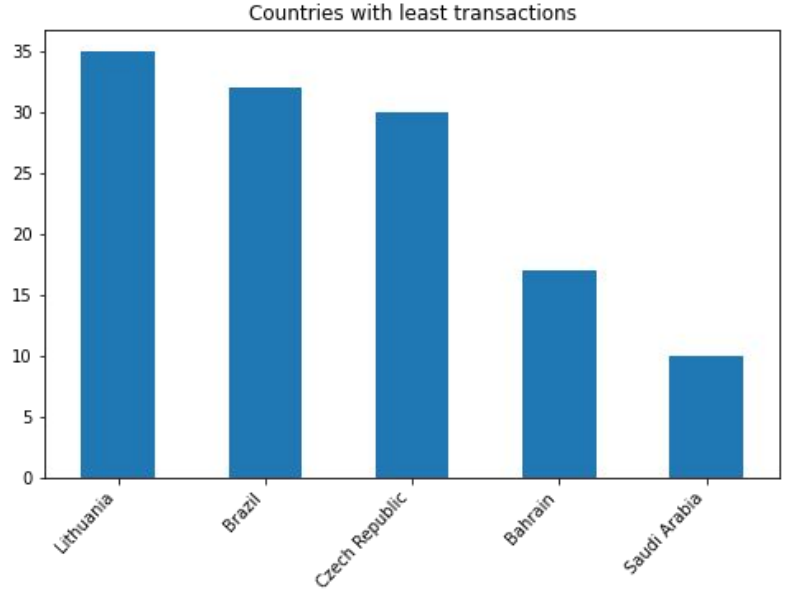


4372 customer records are present in the dataset having 22190 transactions in total.

These are the top customers of the retail store and 17841 CustomerID has most number of transactions 7500+.

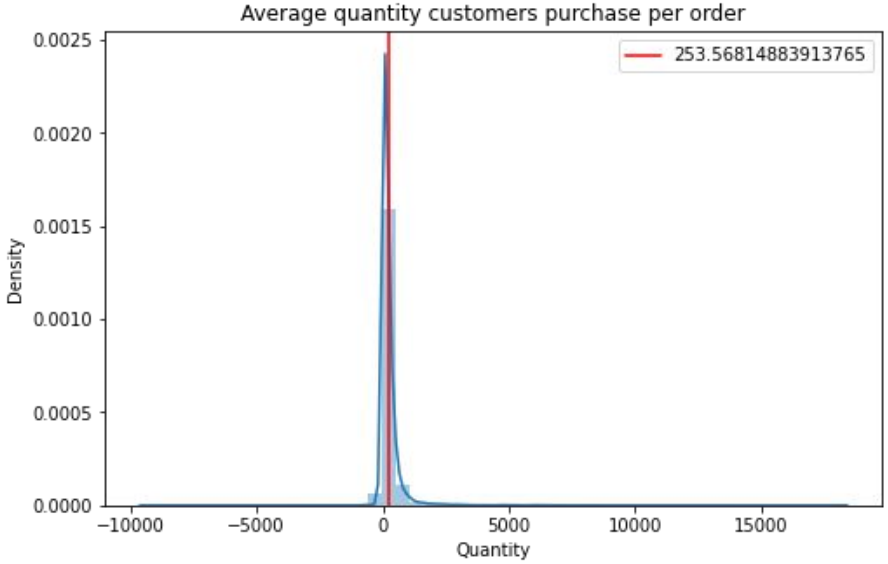
Exploring Country column

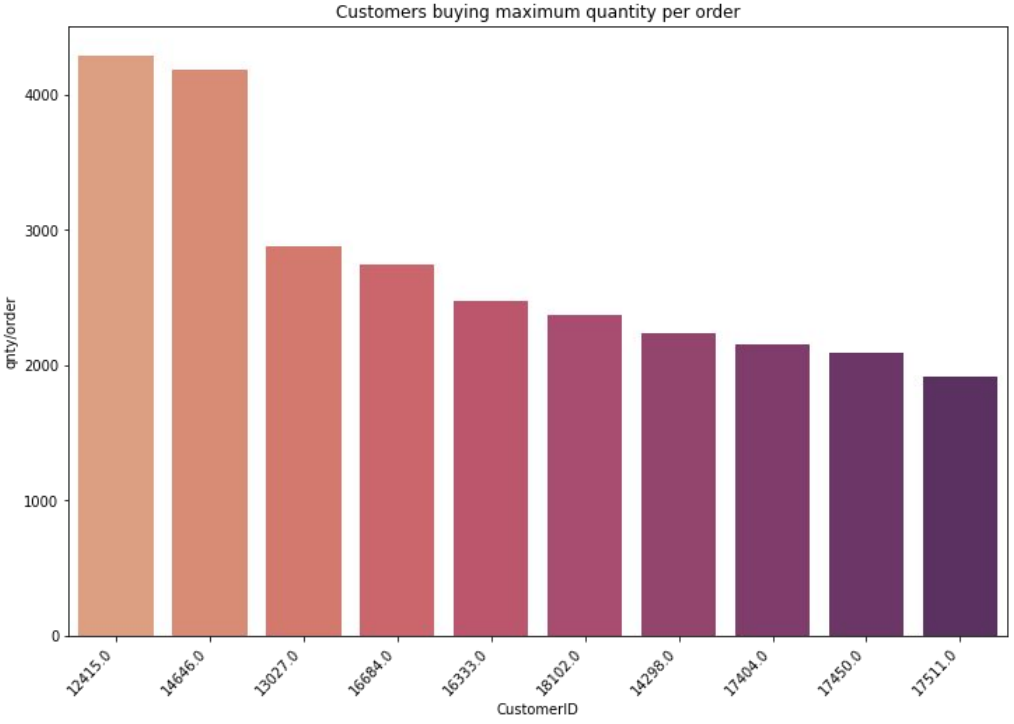




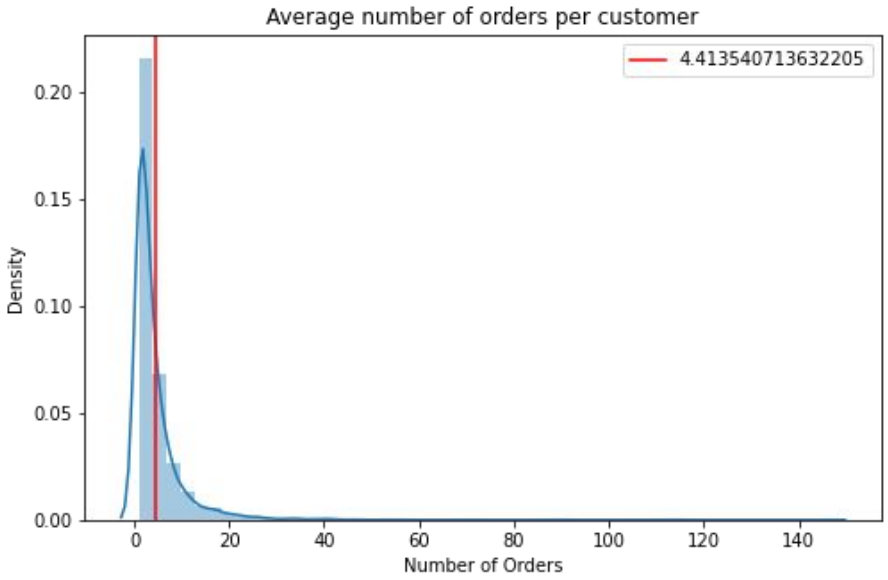
Data Insights

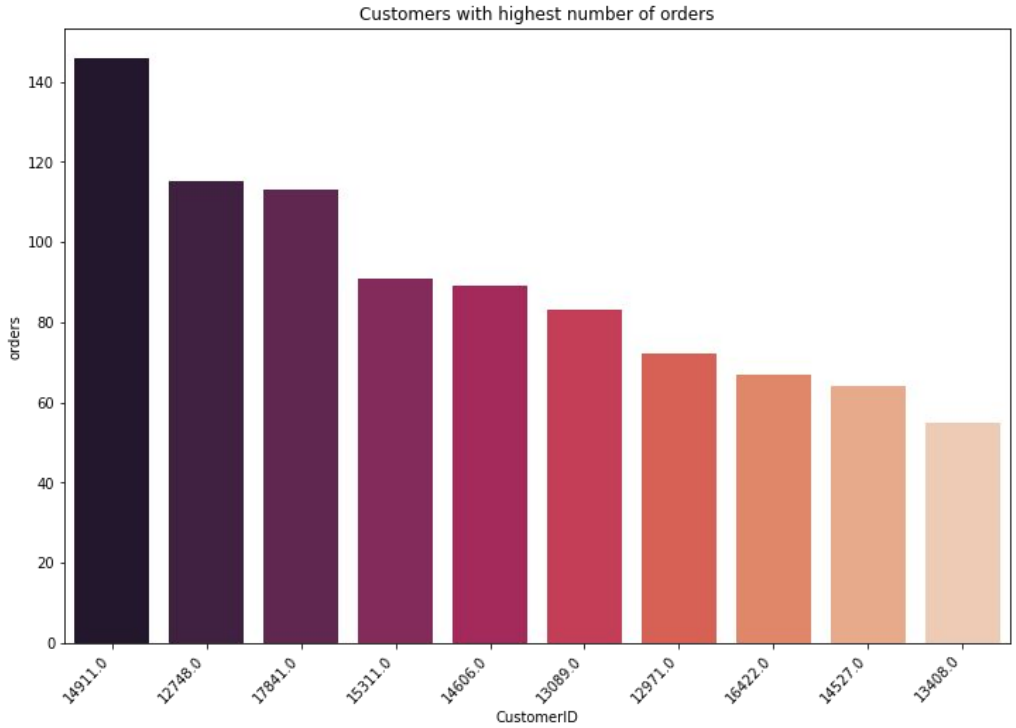
Average quantity customer purchased per order



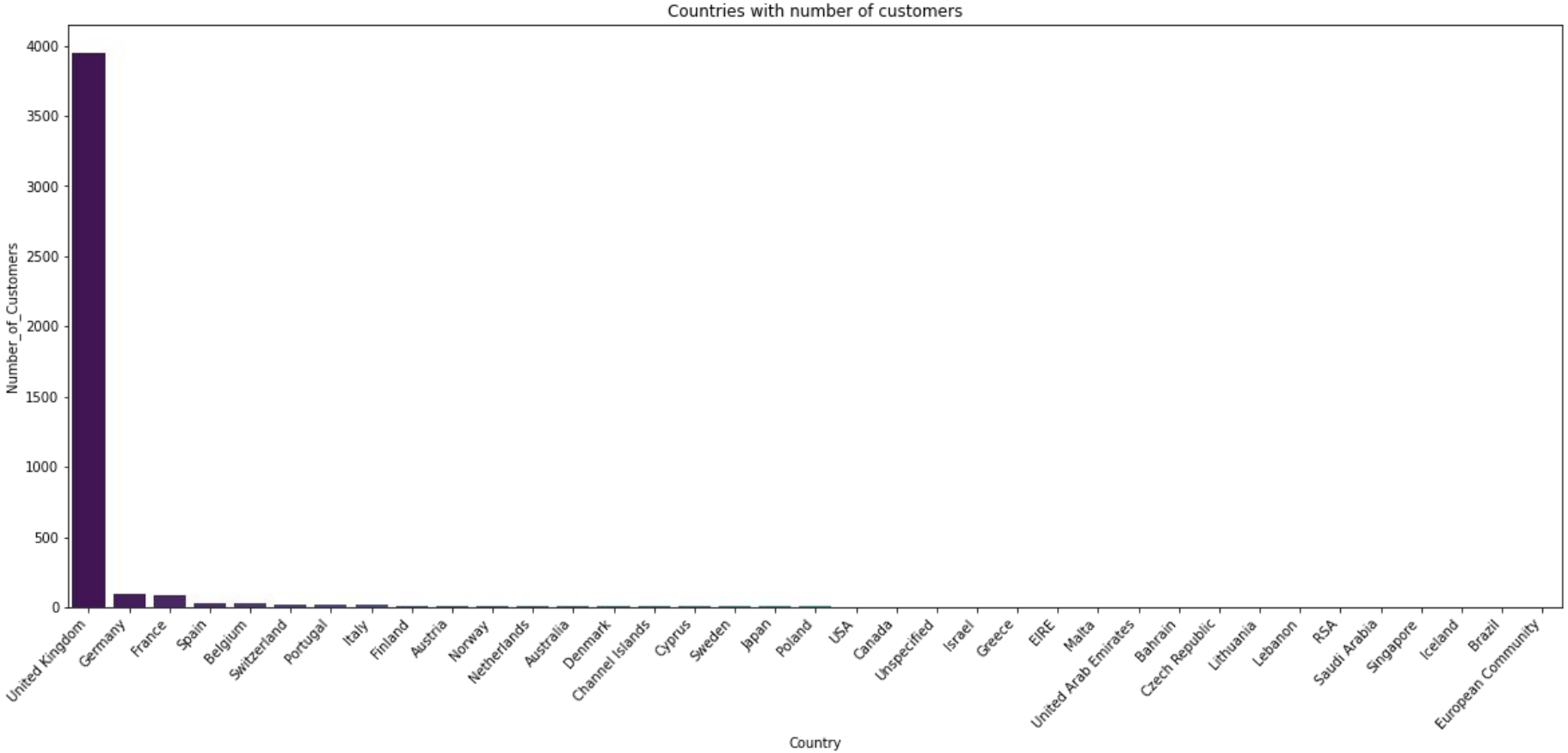


Average number of orders per customer





Which Country has most customers?



Data Preprocessing

- Consider only United Kingdom retail data for maximum impact and not to form clustering on geographical conditions.
- Filtered Cancelled orders also
- Created total cost and date column

```
Data columns (total 10 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 1   InvoiceNo    349227 non-null  object
 2   StockCode    349227 non-null  object
 3   Description  349227 non-null  object
 4   Quantity     349227 non-null  int64
 5   InvoiceDate  349227 non-null  datetime64[ns]
 6   UnitPrice    349227 non-null  float64
 7   CustomerID   349227 non-null  float64
 8   Country      349227 non-null  object
 9   date         349227 non-null  object
 10  total_cost   349227 non-null  float64
```

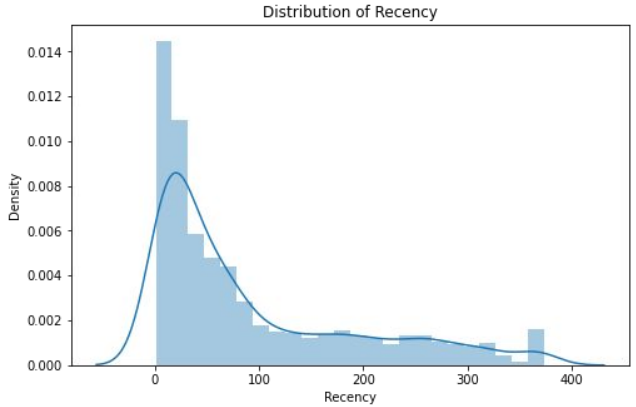
RFM Analysis

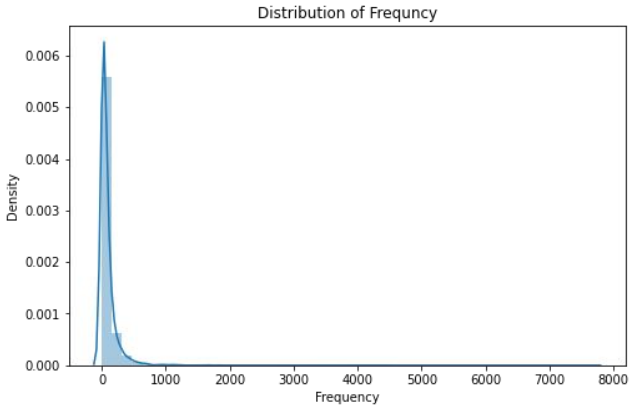
R (Recency) - Number of days since last purchase

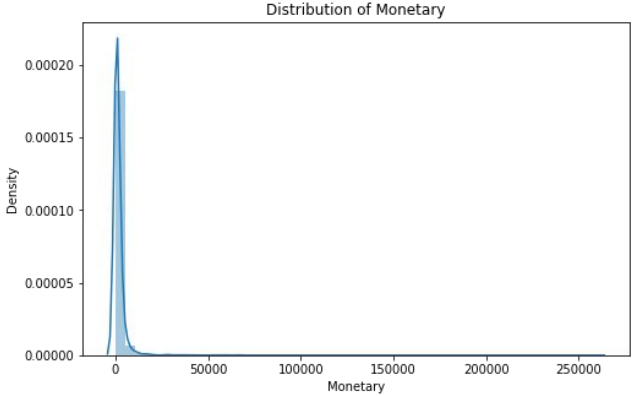
F (Frequency) - Number of times customer has purchased

M (Monetary) - Total Amount of all purchases

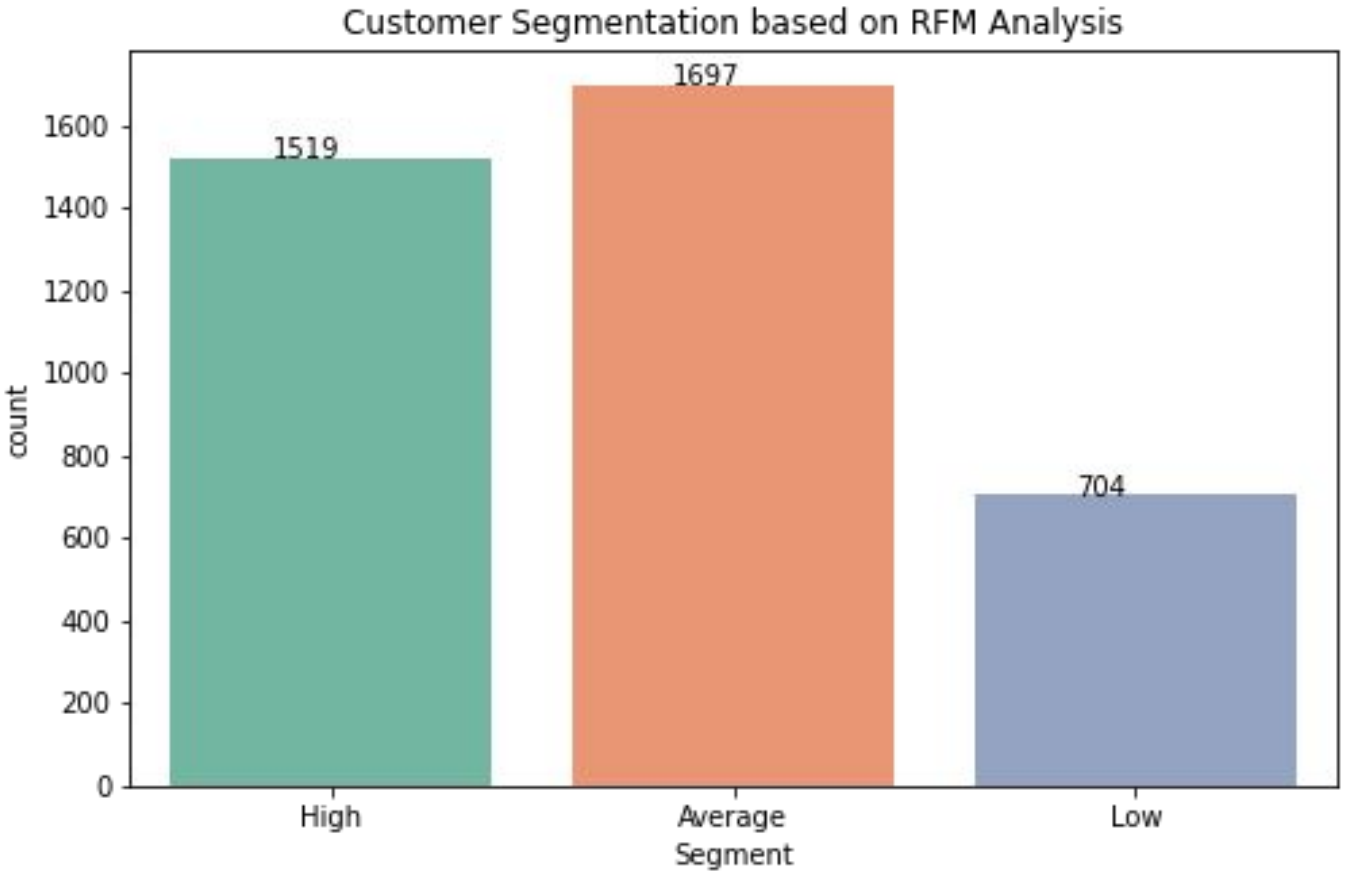
CustomerID	Recency	Frequency	Monetary	R	F	M	RFM_Segment	RFM_score
12346.0	326	1	77183.60	1	1	4	114	6
12747.0	3	103	4196.01	4	4	4	444	12
12748.0	1	4413	33053.19	4	4	4	444	12
12749.0	4	199	4090.88	4	4	4	444	12
12820.0	4	59	942.34	4	3	3	433	10





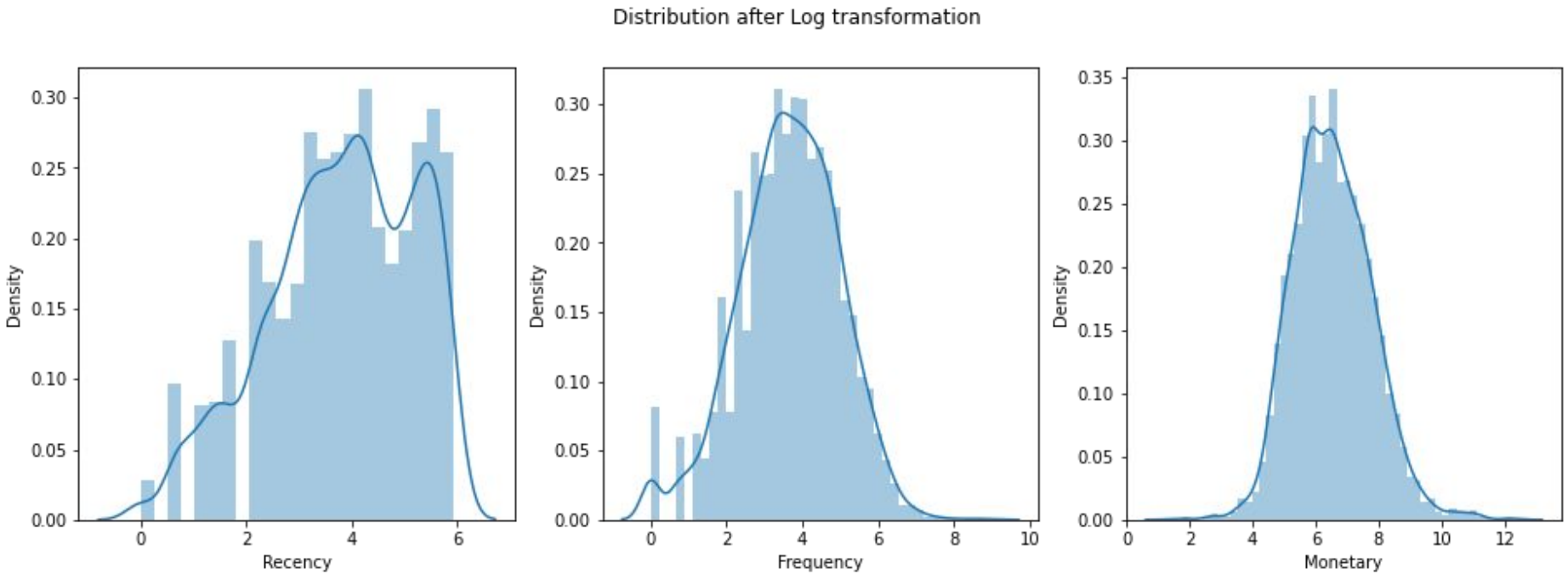


RFM Analysis

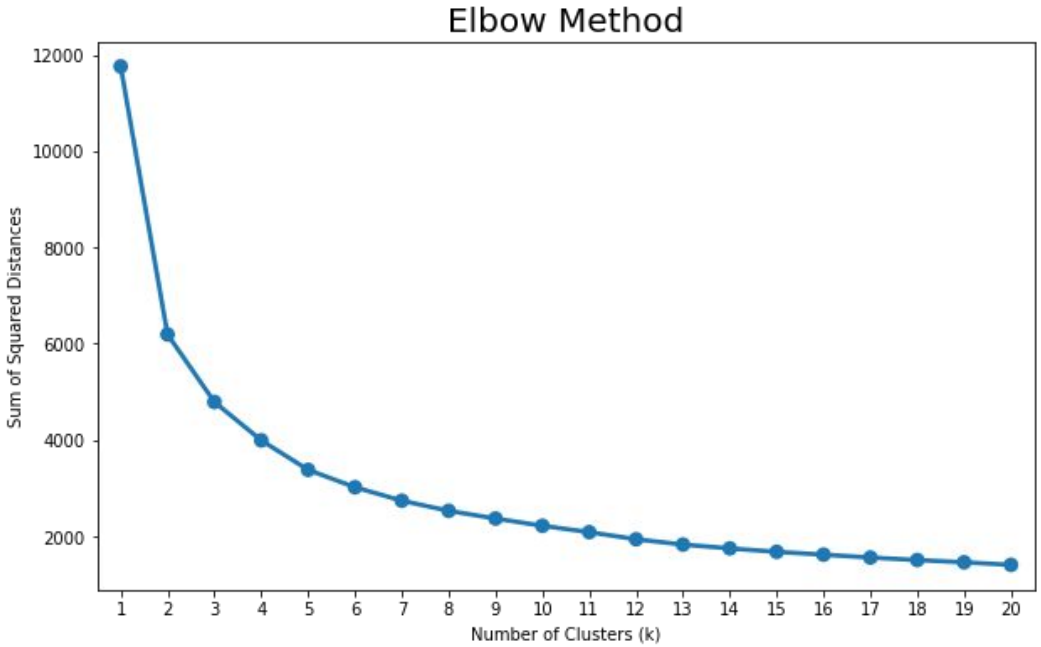


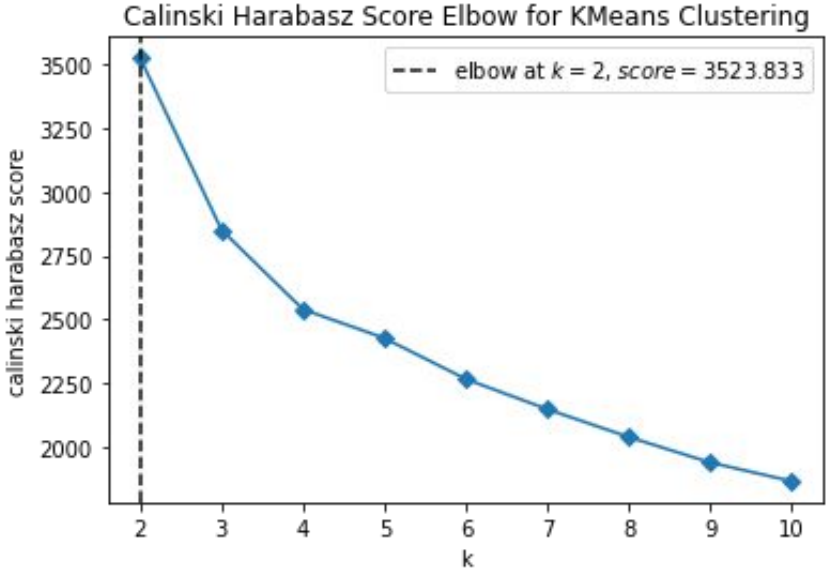
Data Processing

- Clustering algorithms require normally distributed data. We applied log transformation to RFM data to reduce skewness.
- Scaled the data using StandardScaler



K-Means Clustering - Elbow Method

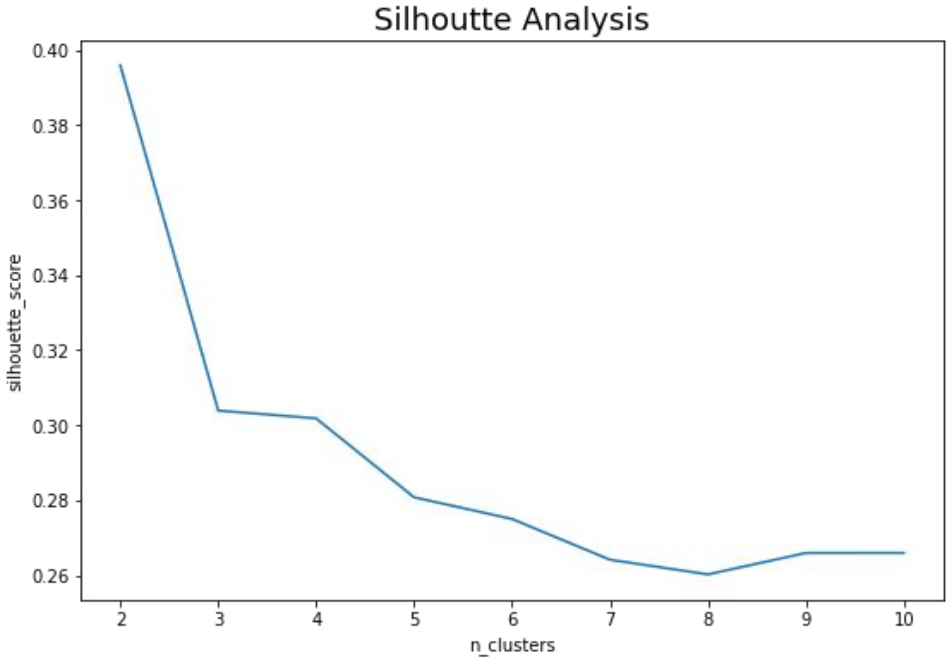




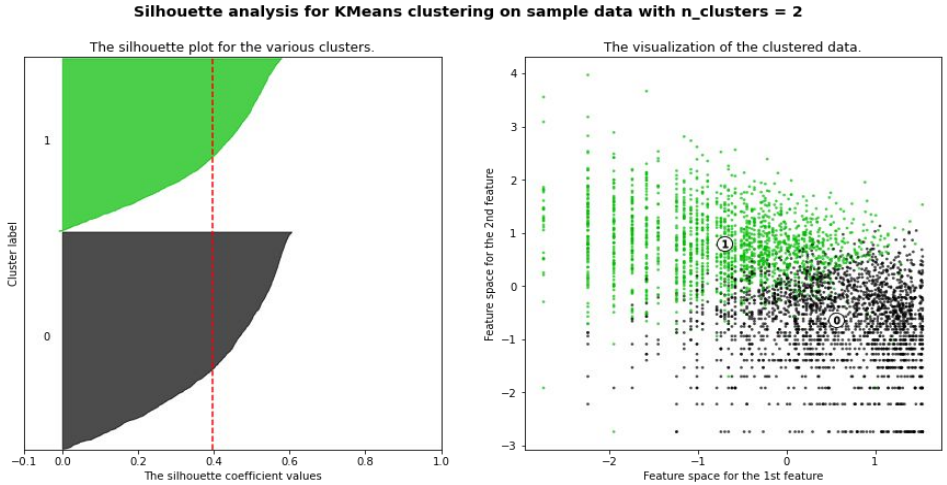
Here, we cannot see a very distinct elbow point. We might infer the optimal value of K to be 2, 3 or 4.

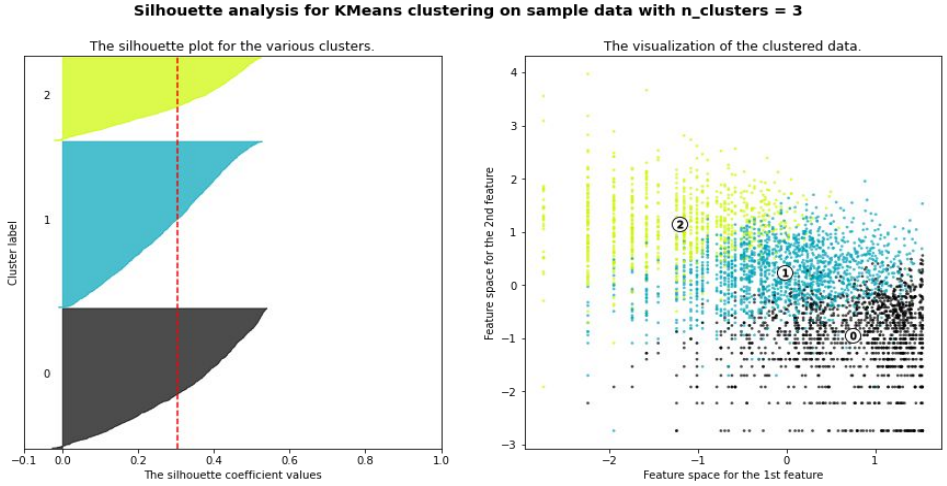
From the above plot, we can see that the optimal number of clusters is 2.

Silhouette Analysis

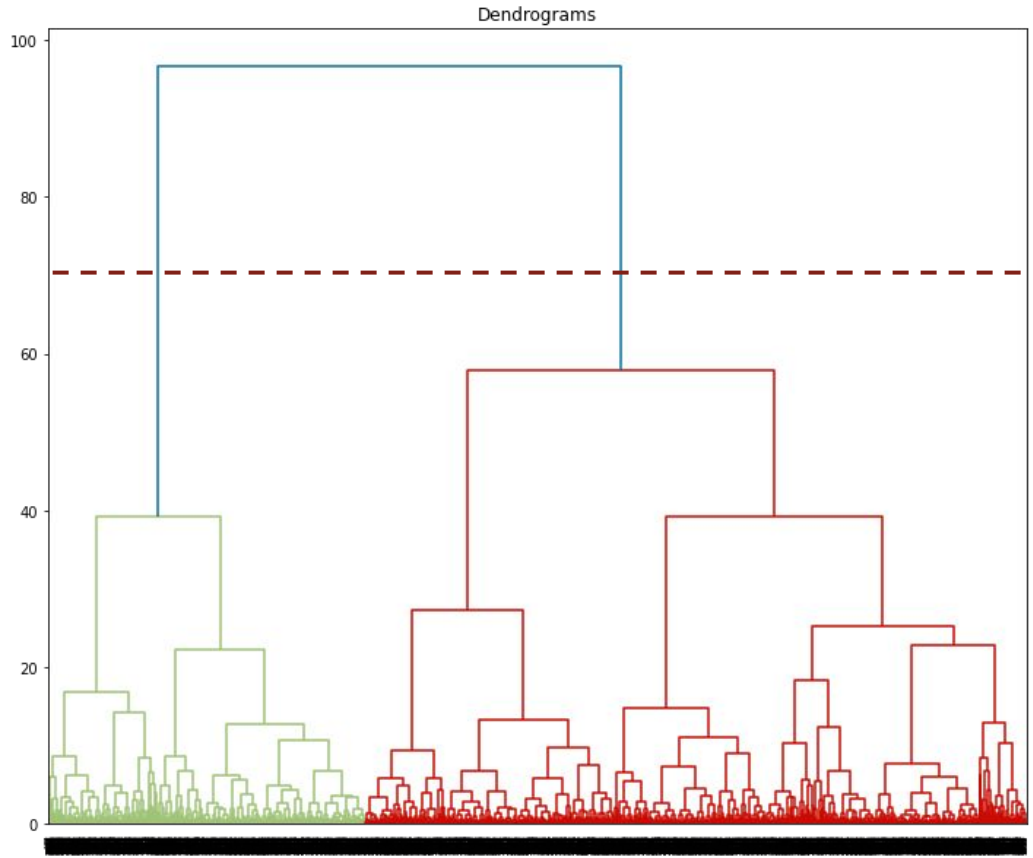


The best silhouette score obtained is when the number of clusters is 2.

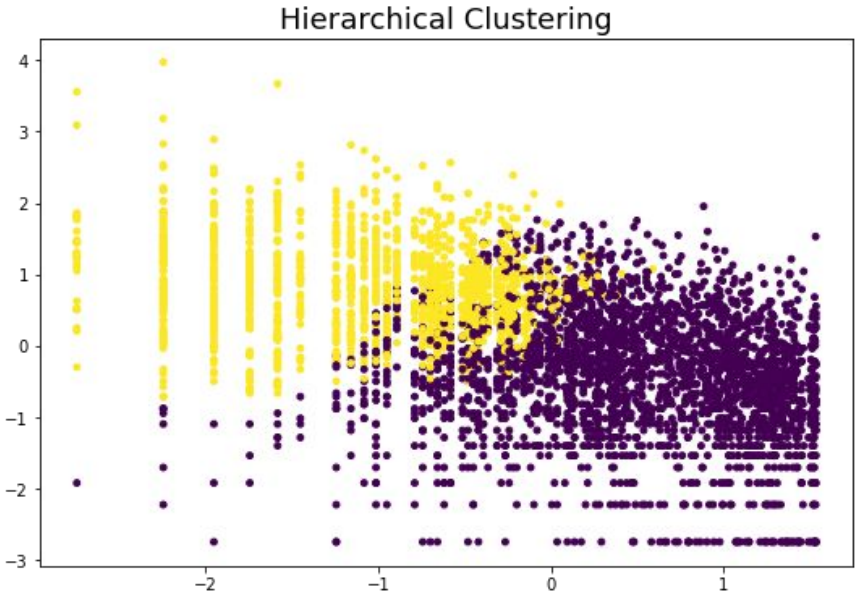




Hierarchical Clustering

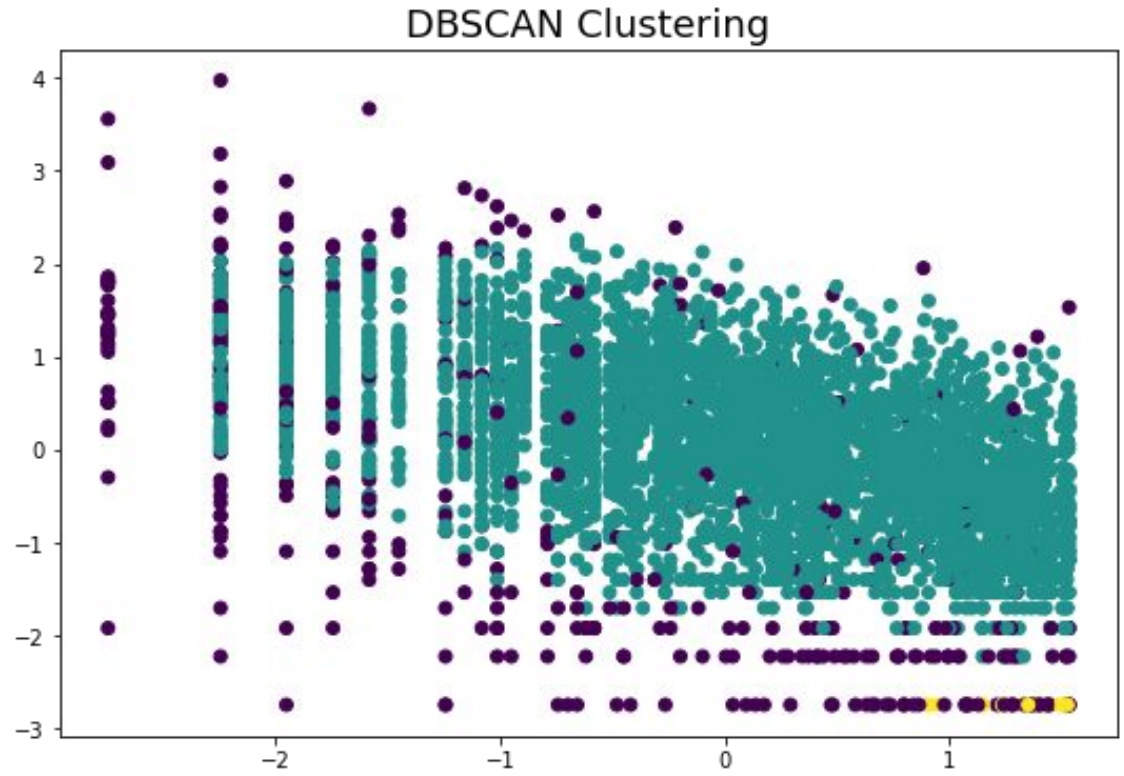


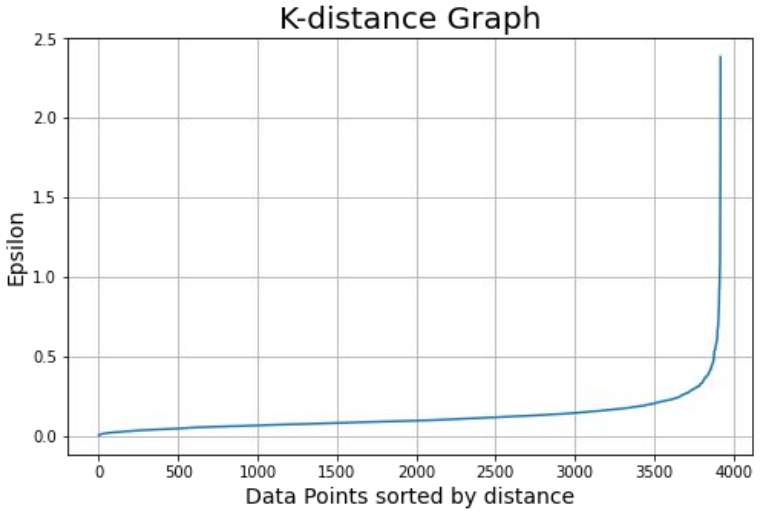

```
Model = AgglomerativeClustering(n_clusters=2,
affinity='euclidean', linkage='ward')
```



Customers are well separate when we use Hierarchical clustering and number of clusters equal to 2

DBSCAN






```
Clusters formed
 0     3503
-1      395 - Noise
 1       22
```

Result

K-means Clusters Summary			
Cluster	**Recency**	**Frequency**	**Monetary**
0	142.0	24.0	451.0
1	31.0	170.0	3625.0

Hierarchical Clusters Summary			
HR_labels	**Recency**	**Frequency**	**Monetary**
0	130.0	41.0	833.0
1	16.0	189.0	3993.0

Customer Segmentation

Wholesale Customers - 'Cluster 1' is the high value customer segment as the customers in this group place the highest value orders with a very high relative frequency than other members. They are also the ones who have transacted the most recently. These are the wholesale customers of the retail store.

Average Customers - 'Cluster 0' is the average customer segment. These customers order less frequently than the wholesale customers and their orders are pretty low valued.

Challenges

- Large dataset to handle
- Lot of duplicate and missing values
- Data analysis was challenging
- Tuning the hyperparameters of models and fitting models

Conclusion

- From RFM analysis, we manually created three clusters on a quartile basis as high, average and low value customers.
- In K-means clustering, using the elbow method and the silhouette analysis we got 2 as optimal number of clusters.
- In Hierarchical clustering the customers were well grouped using 2 clusters.
- Two customer segments have been formed as Wholesale customers and average customers.
- We can conclude that K-means clustering and Hierarchical clustering can be used for this dataset to segment customers based on RFM analysis.

Q & A