Capstone Project 4

Unsupervised ML

Online retail customer segmentation

by
Syed Adnan S

Problem statement:

- Our main task here is to identify major customer segments on a transactional 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.
- Customer segmentation is the process by which you divide your customers up based on common characteristics so you can market to those customers more effectively.
- There are alot of benefits of customer segmentation like organised customer bases, targeted communication becomes easy and much more.
- There are many types and ways of segmenting, but I'll be using behavioral segmentation to perform the task at hand.

❖ Understanding the data:

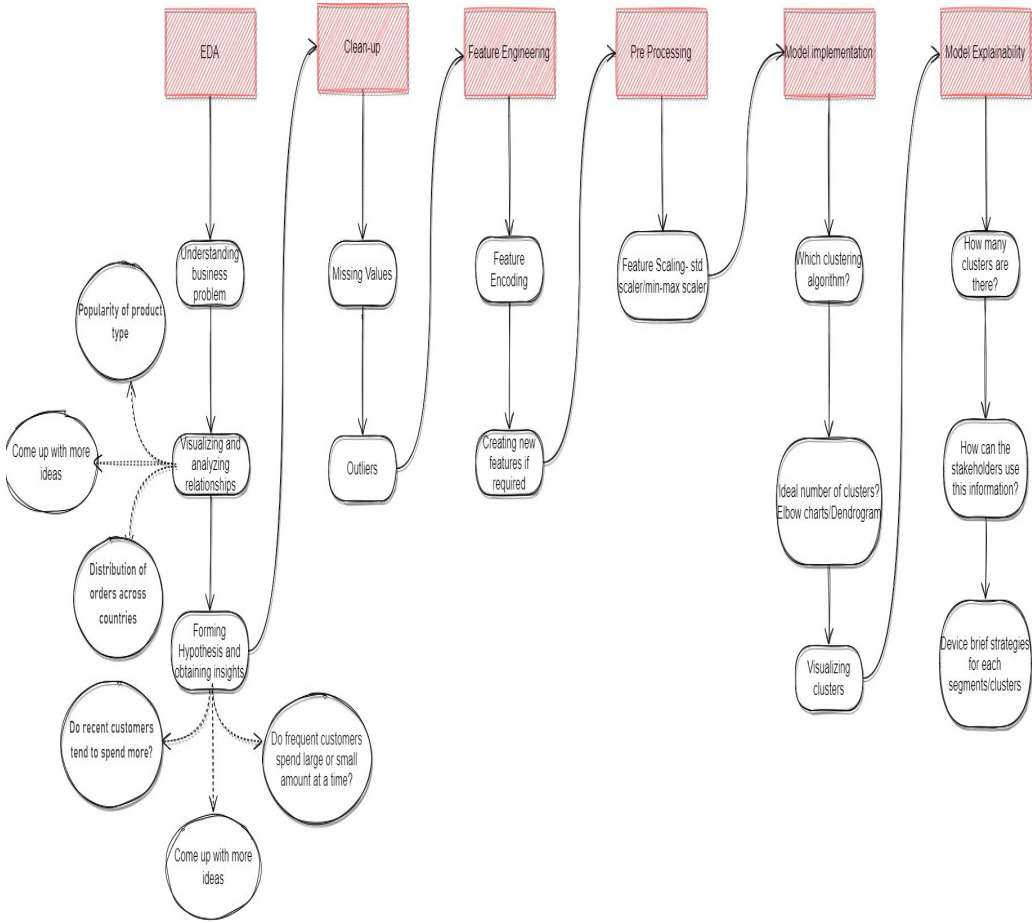
- To increase the efficiency of our analysis we will first have to understand the data and also check if there are some corruptions in the data and if any found we will try to treat it.
- As stated earlier this dataset is from an online retail company. It has 5,41,909 observations and 8 columns.
- Let us now look at what these 8 columns are:
  - 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.
  - StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Understanding the data (Contd):

- 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.
- Unit Price: 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.

❖ Project Flowchart:

1. Initial preparations.
2. Data Cleaning.
3. Feature engineering.
4. EDA.
5. Forming the segmentation criteria.
6. Pre processing of data
7. Model implementation



1. Initial preparation:

- In this section I've loaded in the dependencies, like pandas, seaborn, and many more from the scikit learn library.
- The next step was to mount the drive where the data was stored.
- After mounting the drive I used the pandas.read_excel() function to read the data given to us in excel format.

Note: The data for this project is given to us by the company, AlmaBetter.

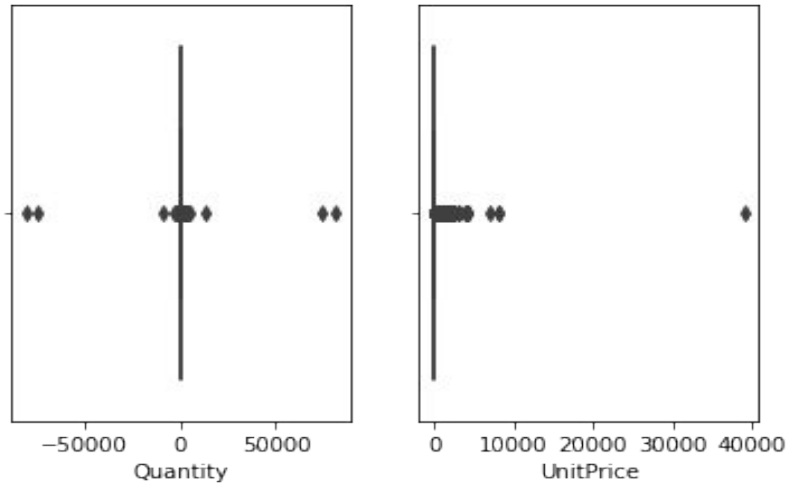
2. Data Cleaning:

- Data exploration: In this step I've just explored the data and its individual properties to get a better idea of how to work on it.
- Handling null values: Null values affect the quality of our ML models and therefore we will have to treat them. There are two ways of dealing with null values, one is directly deleting the null values and the other is to impute the null values with some meaning full values.
- In this project only two columns had null values and they are, Description and CustomerID, I couldn't find any meaningful way to impute these values, therefore I have just deleted them.
- Handling duplicate values: When two features have the same set of values they are known as duplicate values. Duplicate values can cause detrimental effect on our accuracy.

2. Data Cleaning (Contd):

- Duplicate values can ruin the split between train, test and validation set, which ultimately leads to biased performance estimates that disappoint the model in production. The best way of dealing with duplicate values is to delete them and therefore I've just deleted them.

- Removing outliers: Outliers are those points which are significantly different from the other points. I've used boxplot to detect these outliers and deleted them with the help of IQR method



- Removing cancelled orders: There are cancelled orders represented by a 'c' before the Invoice number, I've deleted these transactions as they are not so important to me.

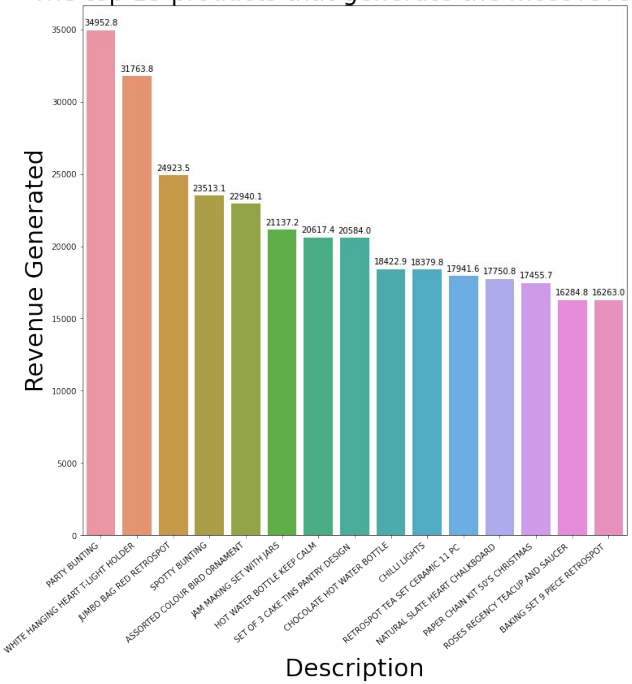
3. Feature engineering:

- **Extracting columns from the Invoice date column:** Here I've used the Invoice date column to form columns like day of the month, day of the week, year, month and hour so that we can have a better understanding of the data for each timespan.
- Creating the total amount column: I've created a column called total_amount using the formula,
  *total amount = Quantity*UnitPrice.*
  This represents the total amount spent by the customer for a particular transaction.

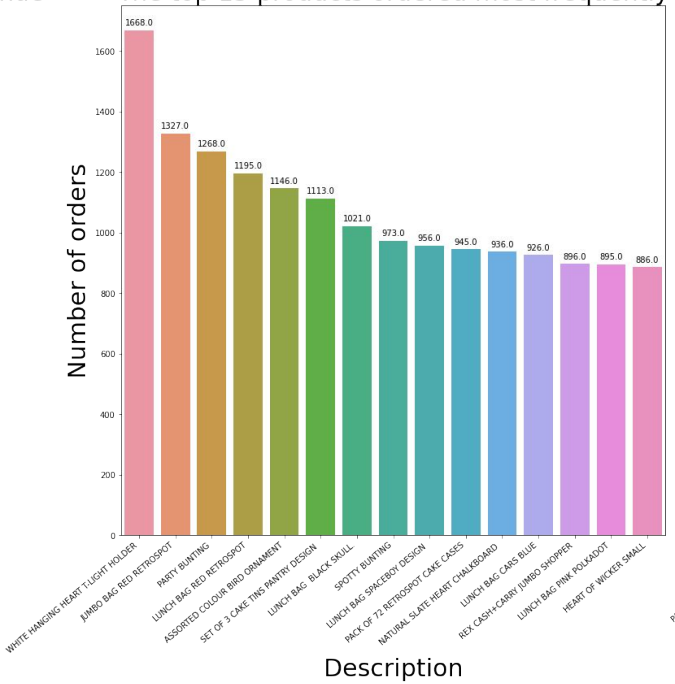
4. EDA:

- **In this step I have done exploratory data analysis on the data to see if I can find some valuable insights that can be directly applied to increasing the success of the business.**

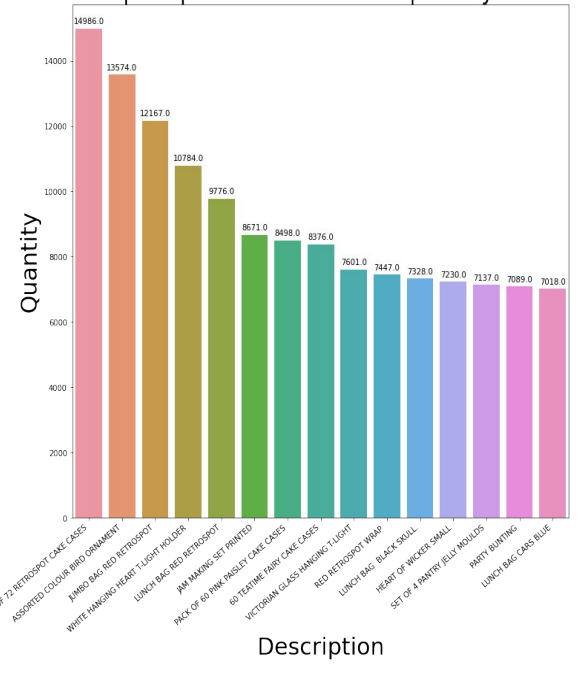
The top 15 products that generate the most revenue



The top 15 products ordered most frequently

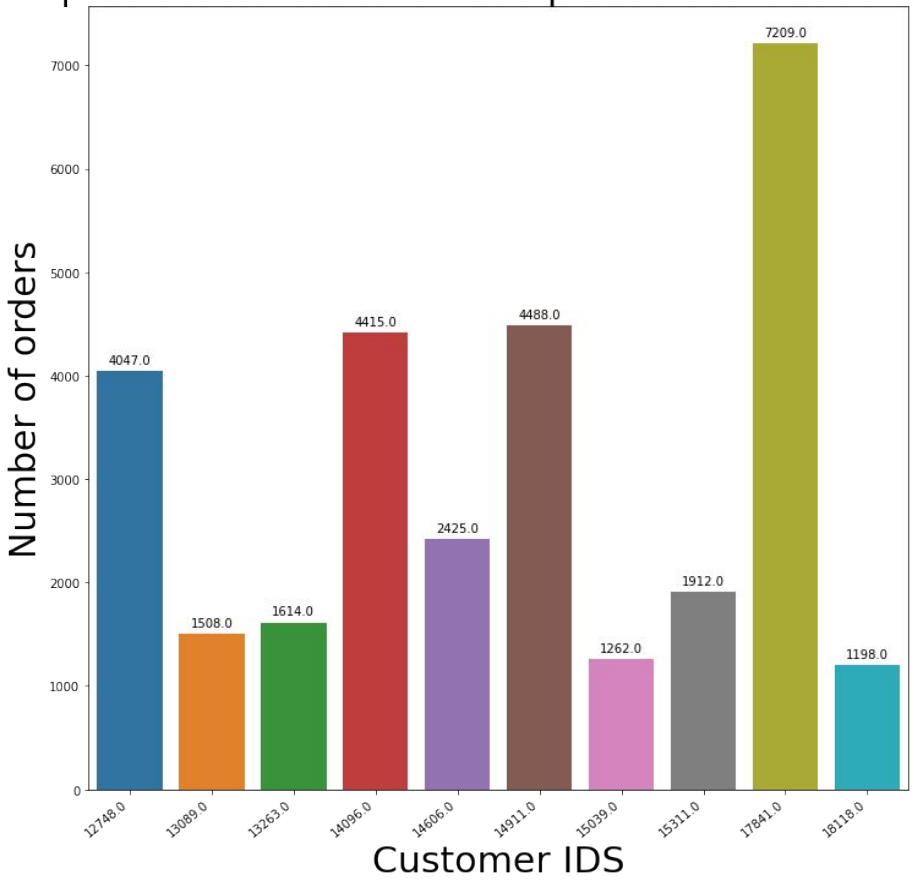


The top 15 products with most quantity ordered

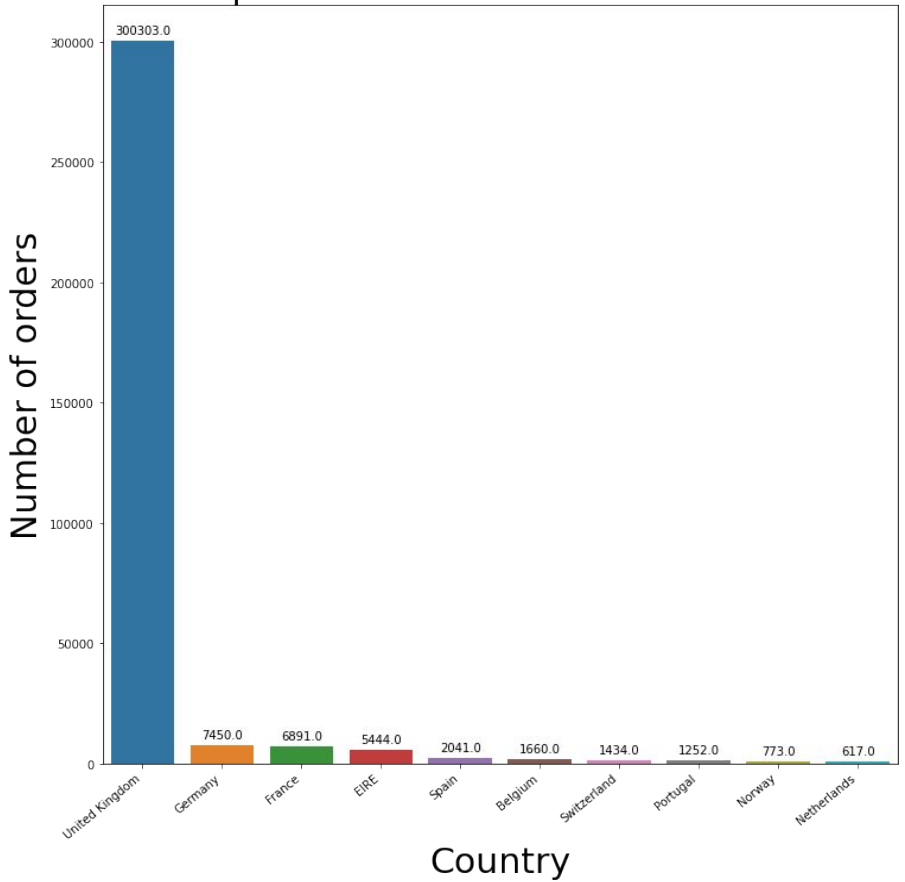


4. EDA (Contd):

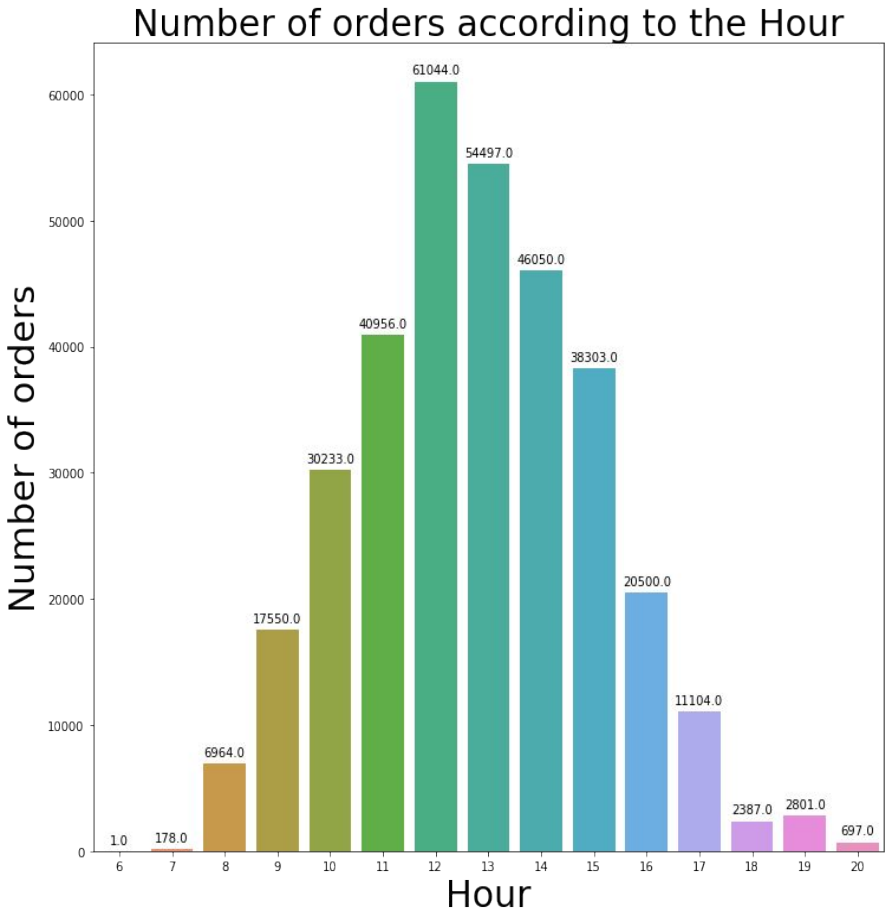
Top 10 customers who have placed the most orders

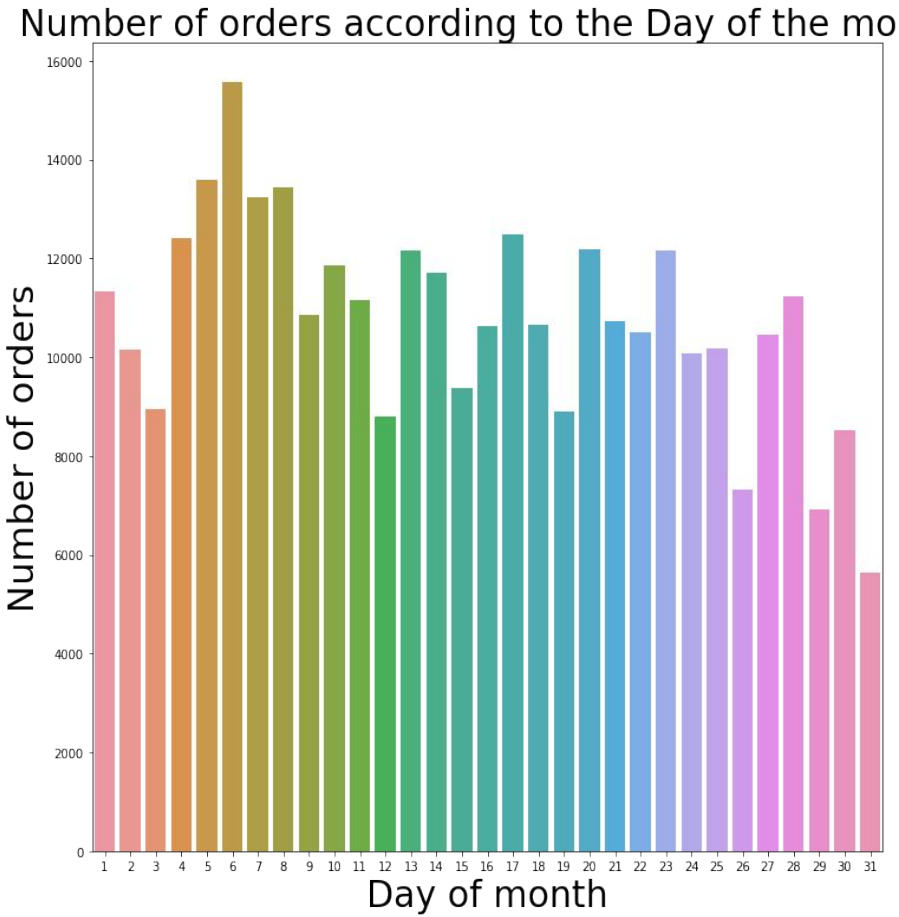


The top 10 countries with the most orders.



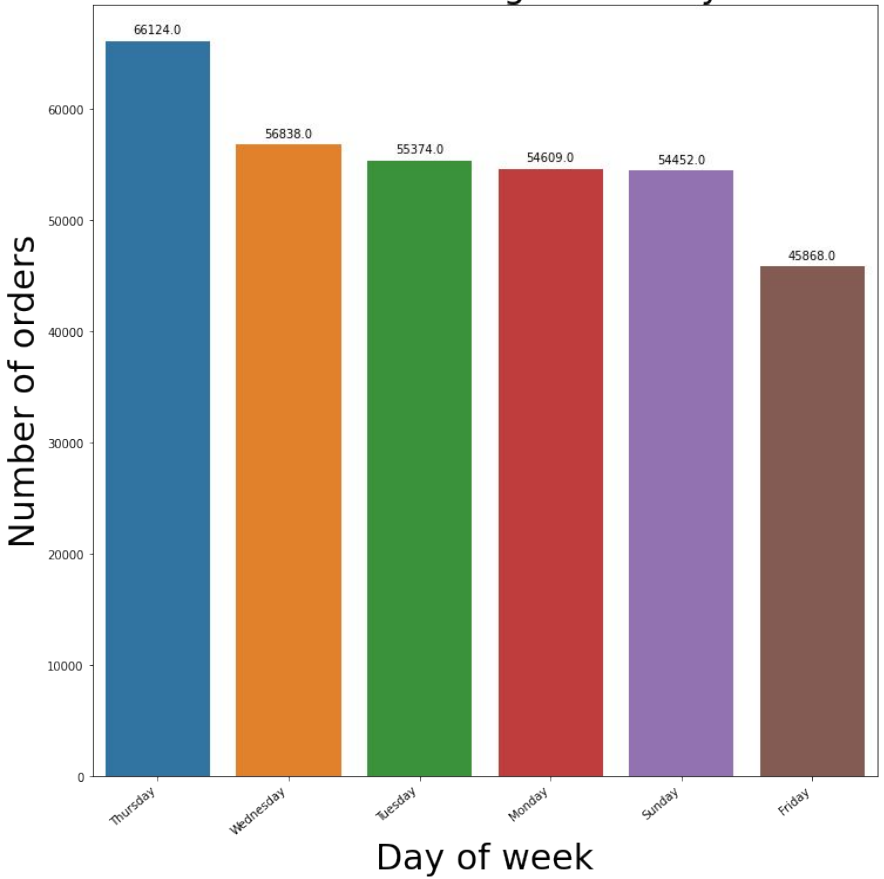
4. EDA (Contd):


Number of orders according to the Hour
Number of orders
Hour
1.0
178.0
6964.0
17550.0
30233.0
40956.0
61044.0
54497.0
46050.0
38303.0
20500.0
11104.0
2387.0
2801.0
697.0

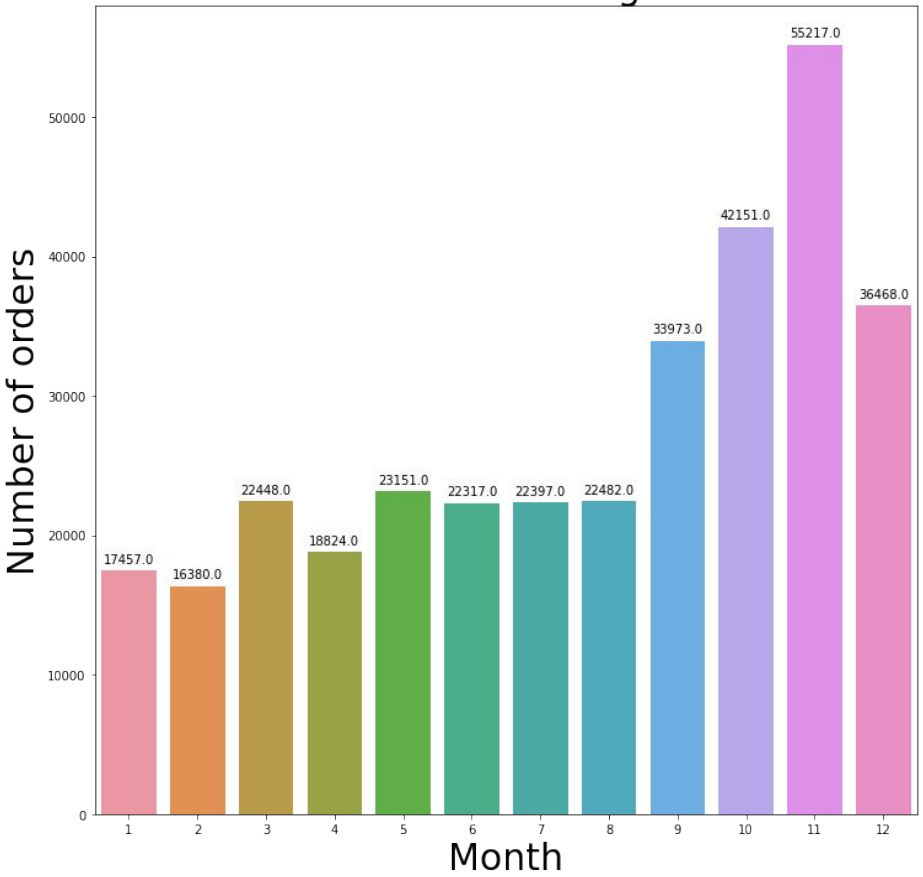

Number of orders according to the Day of the month
Number of orders
Day of month

4. EDA (Contd):

Number of orders according to the Day of the week

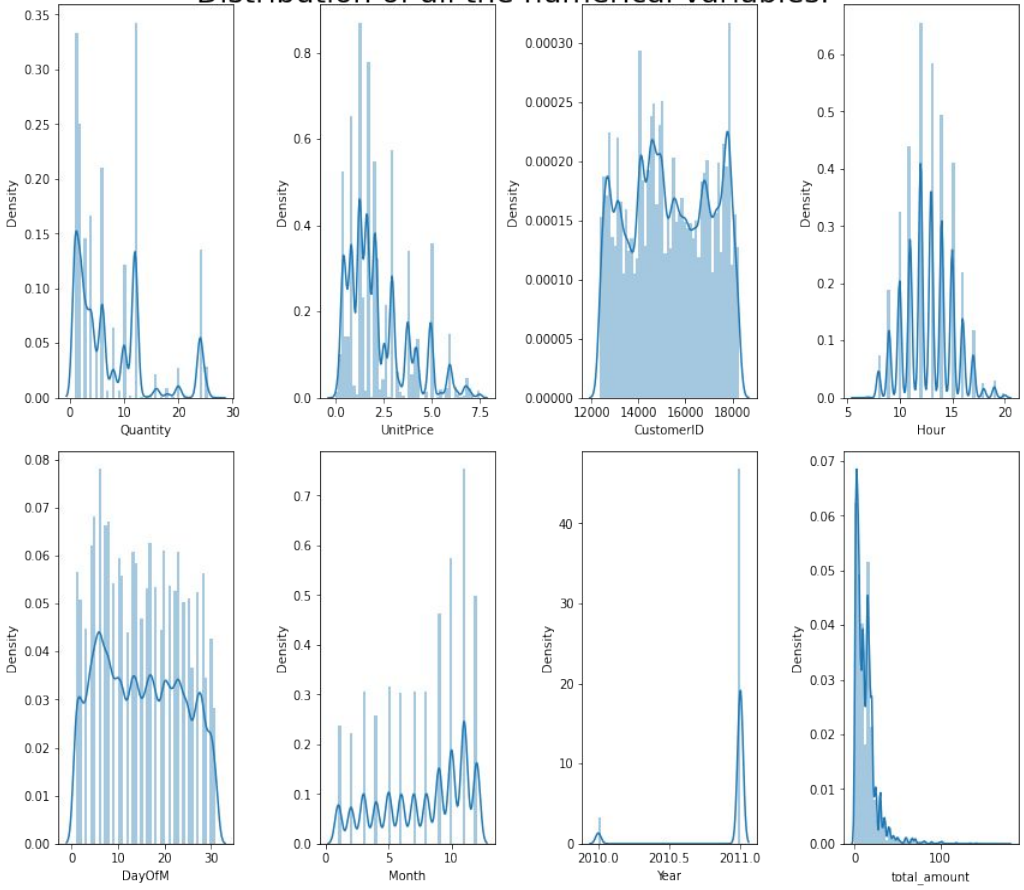


Number of orders according to the Month

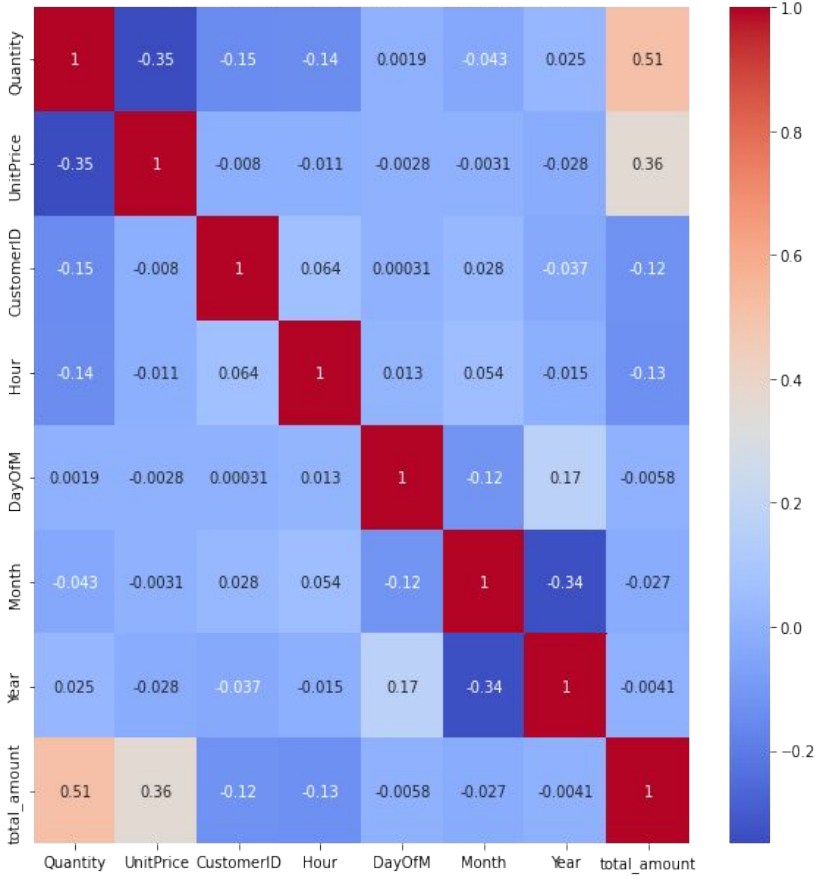


4. EDA (Contd):

Distribution of all the numerical variables.

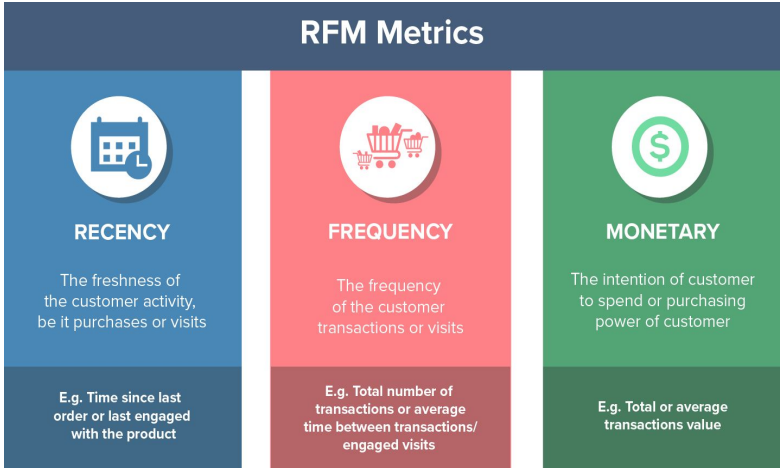


The correlation matrix



5. Forming the segmentation criteria.

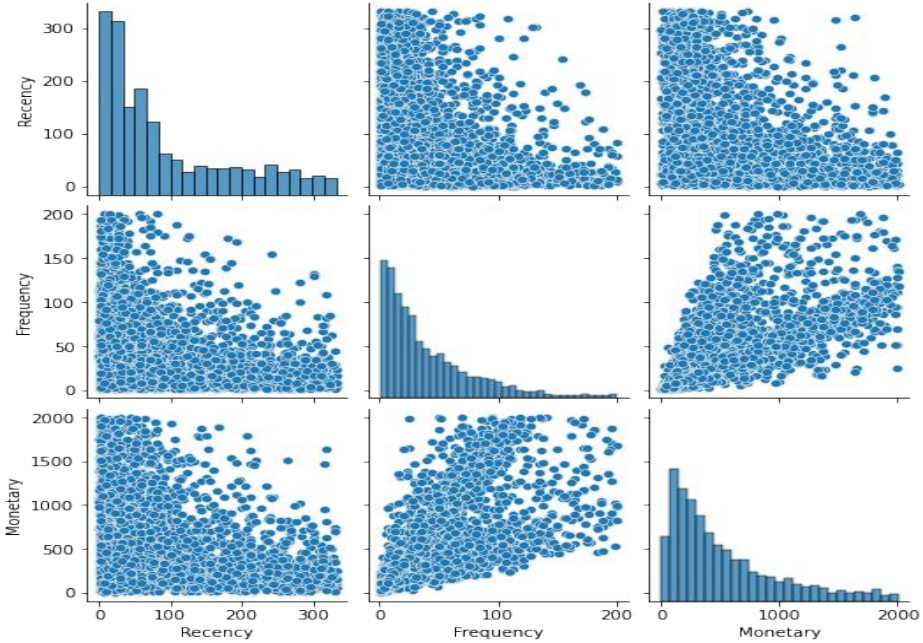
- As stated earlier I'll be using behavioral segmentation to perform the customer segmentation.
- To do this I've used the classical RFM model.

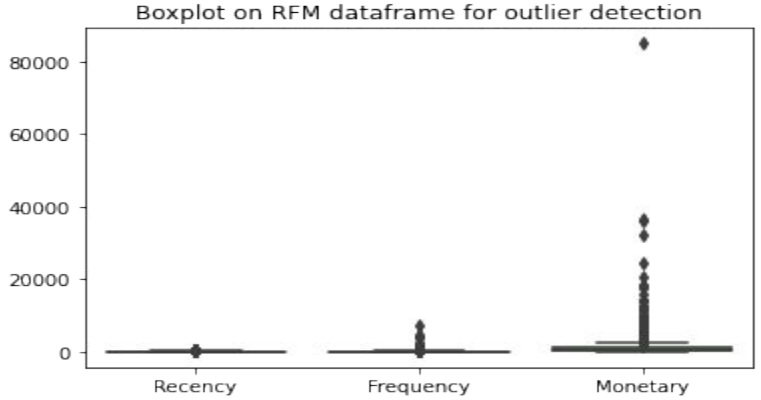


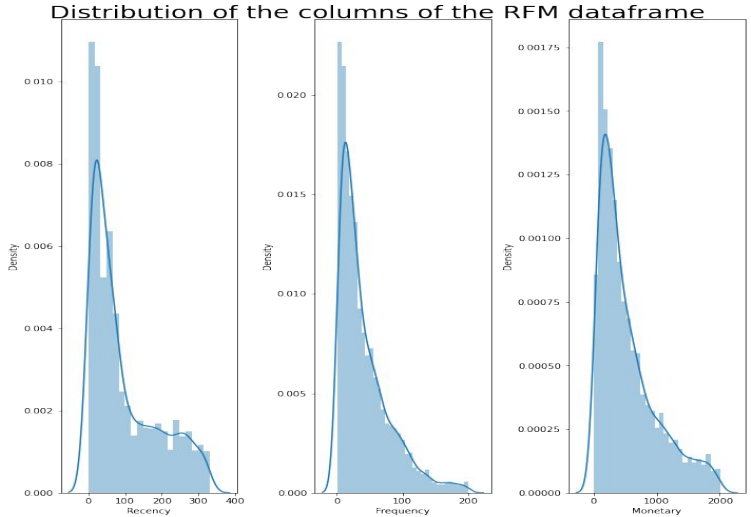
- I've created a new dataset called the RFM data frame which contains the information about the RFM metrics for each customer present in the original dataframe.
- Using these metrics we can easily perform customer segmentation.

5. Forming the segmentation criteria.

- After forming the RFM data frame I've removed the outliers in this data set.







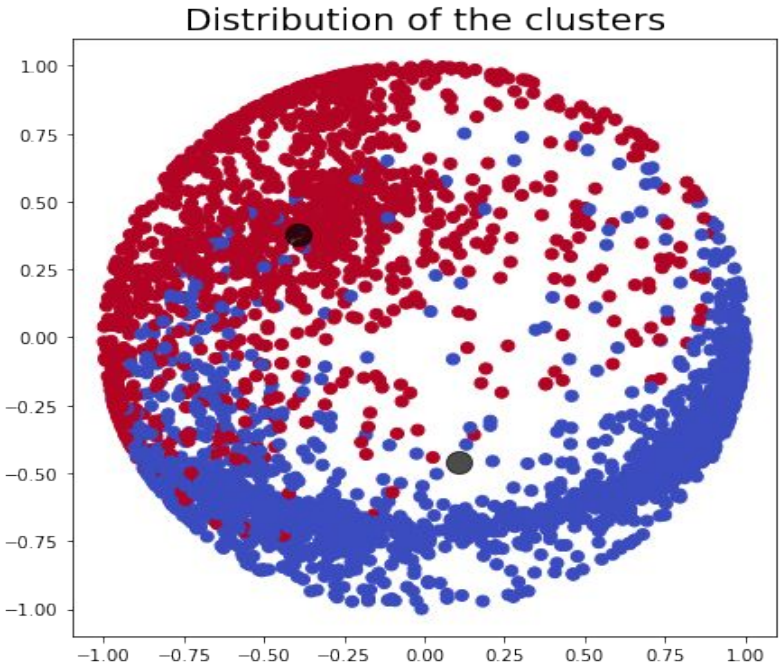
6. Pre processing the data

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values higher. This gives rise to a bias in the ML model. Therefore, we have to scale all the variables in our data.
- In this step I've just scaled and normalized the data in the RFM data frame so that there is no bias given to larger values by the ML models.

7. Model implementation:

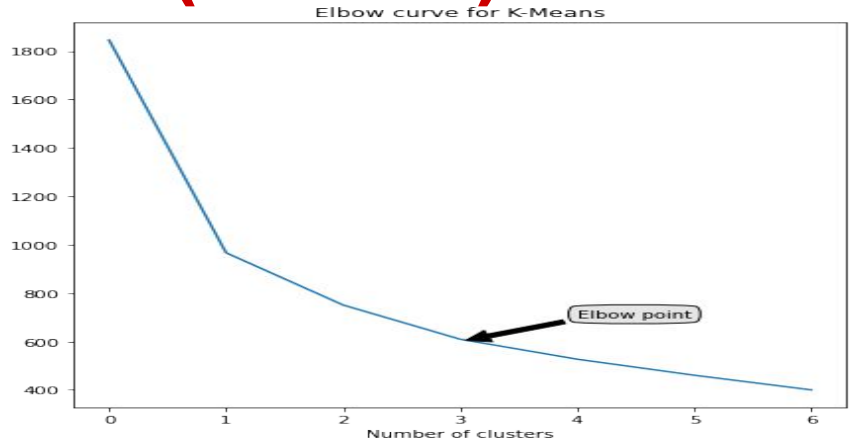
- I've implemented three models on the RFM data frame to form customer segments and they are Simple K Means, K Means with elbow method, and Agglomerative clustering (Hierarchical clustering).

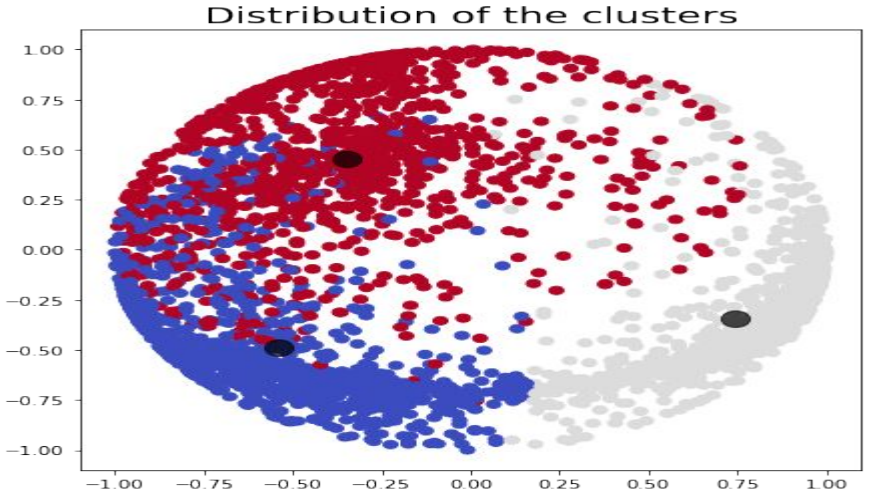
Model : Simple K Means
Number of clusters: 2
Silhouette score: 0.4361
Calinski harabasz index: 2378
Davies Bouldin score: 1.03



7. Model implementation(Contd):

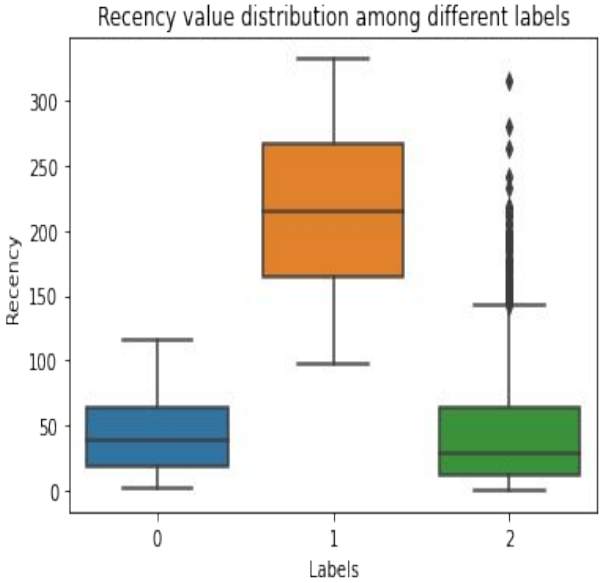
Model : K means with elbow method.
Number of clusters: 3
Silhouette score: 0.5233
Calinski harabasz index: 4171
Davies Bouldin score: 0.688

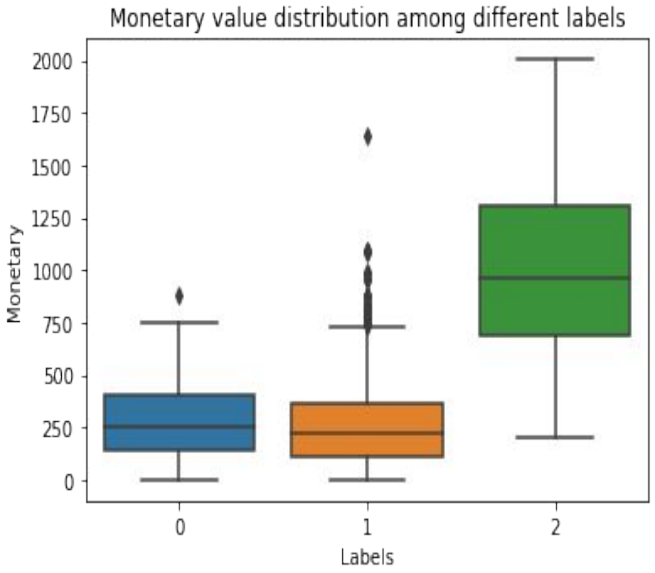


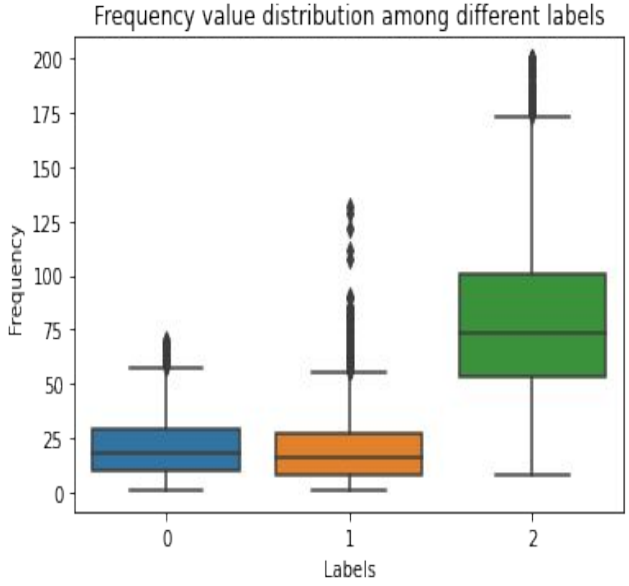


7. Model implementation(Contd):

❖ Cluster Profiling:

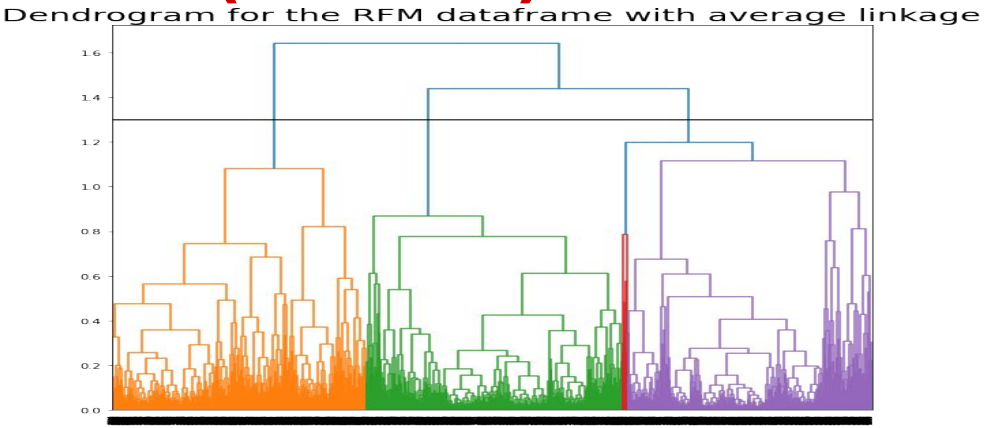


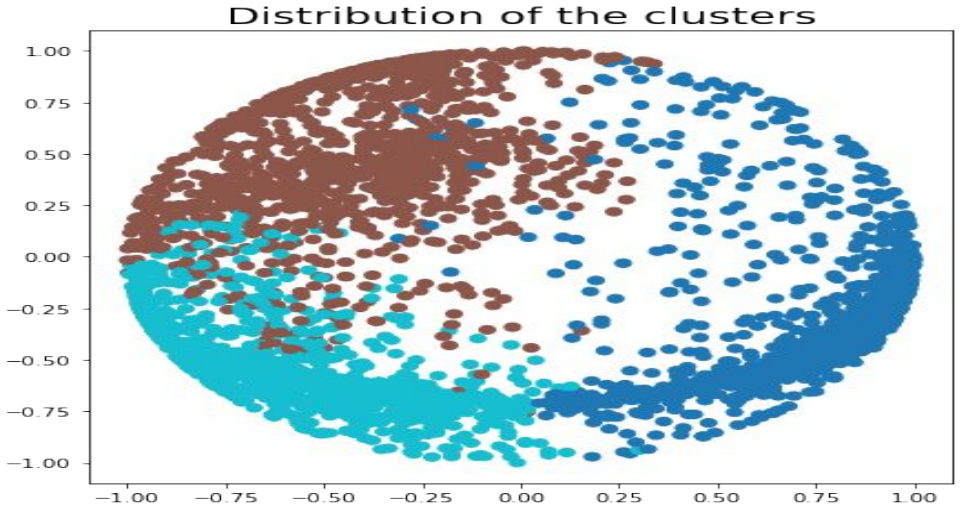




7. Model implementation(Contd):

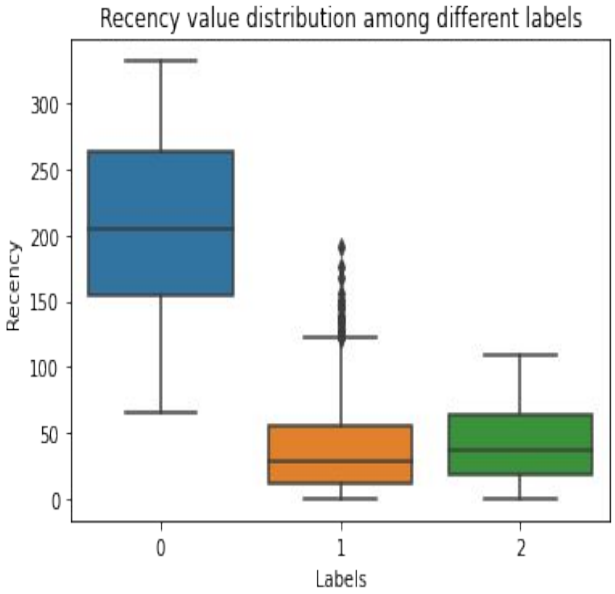
Model : Agglomerative clustering.
Number of clusters: 3
Silhouette score: 0.4958
Calinski harabasz index: 3616
Davies Bouldin score: 0.711

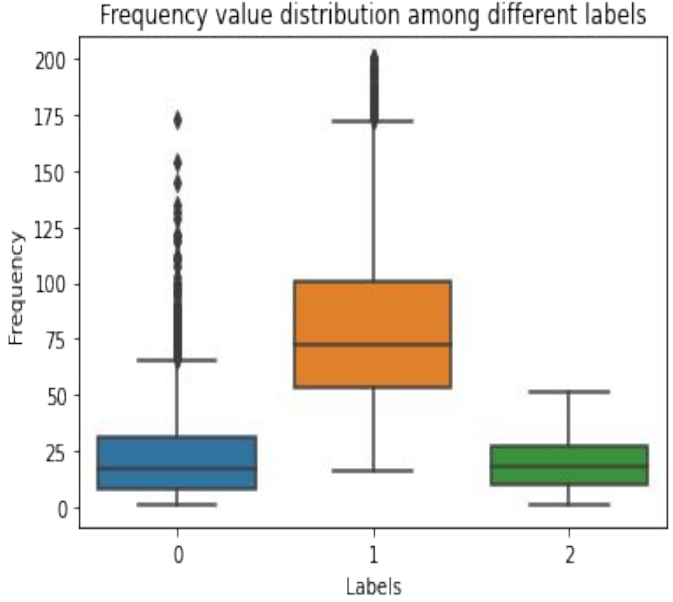


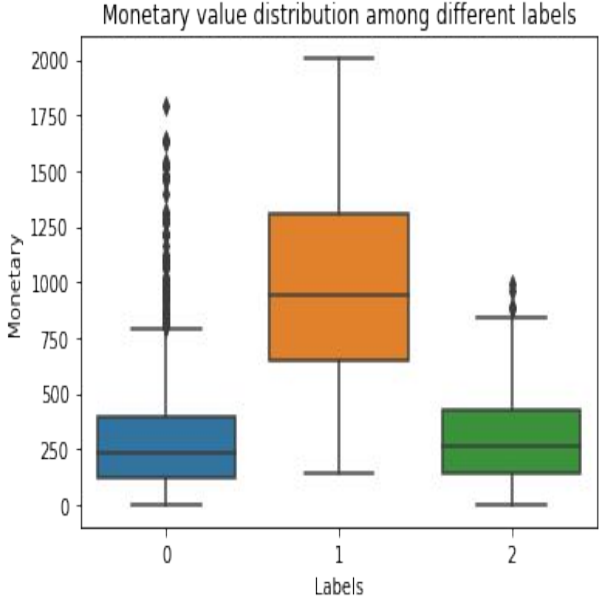


7. Model implementation(Contd):

❖ Cluster Profiling:







Conclusion:

EDA Insights:

- The product 'White hanging heart t-light holder' is the most frequently ordered product, around 1700 times. 'Jumbo bag red retrospot' is the second most ordered product, around 1300 times.
- The product 'Pack of 72 retrospot cake cases' has the most quantity ordered, around 15,000 units. 'Assorted colour bird ornament' is second with around 13,000 units ordered.
- The product "Product Bunting" has made the most money, around 35,000 sterling. "White Hanging heart T-light holder" being the second, which has made around 32,000 sterling.
- The customer with the ID: 17841 has the highest number of orders and the customer with the ID: 18118 has the lowest number of orders.
- United Kingdom has the most orders placed, with around 3 lakh orders. Germany being second, but way less than United Kingdom.
- Most orders are made in the 12th hour, i.e 12pm to 1pm, and the least orders are made in the 6th hour, i.e 6am to 7am.
- The 6th day of the month has the highest number of orders and the 31st day has the lowest.
- Most of the orders are made on Thursday, around 66 thousand, and the least number of orders are made on Friday, around 46 thousand.
- The most number of orders are made in the 11th month, i.e December, and the least in the 2nd month, i.e February.

Conclusion:

Conclusions from Model implementation:

- Simple K Means model has a silhouette score of 0.4361, a Calinski harabasz index of 2378 and a Davies Bouldin score of 1.03.
- K Means model with elbow method has a silhouette score of 0.5233, a Calinski harabasz index of 4171 and a Davies Bouldin score of 0.688.
- Agglomerative clustering has a silhouette score of 0.4958, a Calinski harabasz index of 3616 and a Davies Bouldin score of 0.711.
- K Means model with elbow method is the best performing model.
- Simple K Means model is the worst performing model.
- Actions to take for each cluster:
    i. Perform targeted analysis and targeted advertisement for each cluster.
    ii. Advertise products that can be presented with a discount to the customers in the lesser important clusters, which could convert the customers in these less important groups to customers of more important clusters.

Challenges faced:

- Removing outliers.
- Choosing the right approach for segmentation.
- Choosing the right ML models and evaluation metrics.

THANKYOU!!

I’m very glad that you have tagged along until the end. I hope you enjoyed it and if you have any suggestions about my work, please let me know :)