Capstone Project – 4

Customer Segmentation

Team Members

Asim Siddiqui

Sagar Rokad

Suraj Kumar Mishra

Content

- Problem Statement
- Data Summary
- Data Preprocessing
- Exploratory Data Analysis
- Feature Engineering
- RFM
- Model Implementation
- Conclusion

Problem Statement

- **Aim**:- **Customer Segmentation**

- **Problem Statement:** In this project, your 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.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Summary

Dataset contains 541909 entries & total 8 columns with attributes as:

- **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.
- **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.

Data Insights

- Average number of orders per customer is 3.
- The percentage of order cancellation is 35.72% .
- United Kingdom has the highest number of cancelled orders.
- The average number of items per orders is 20.5 .
- The average number of items per customer is 61.2 .

Data Preprocessing

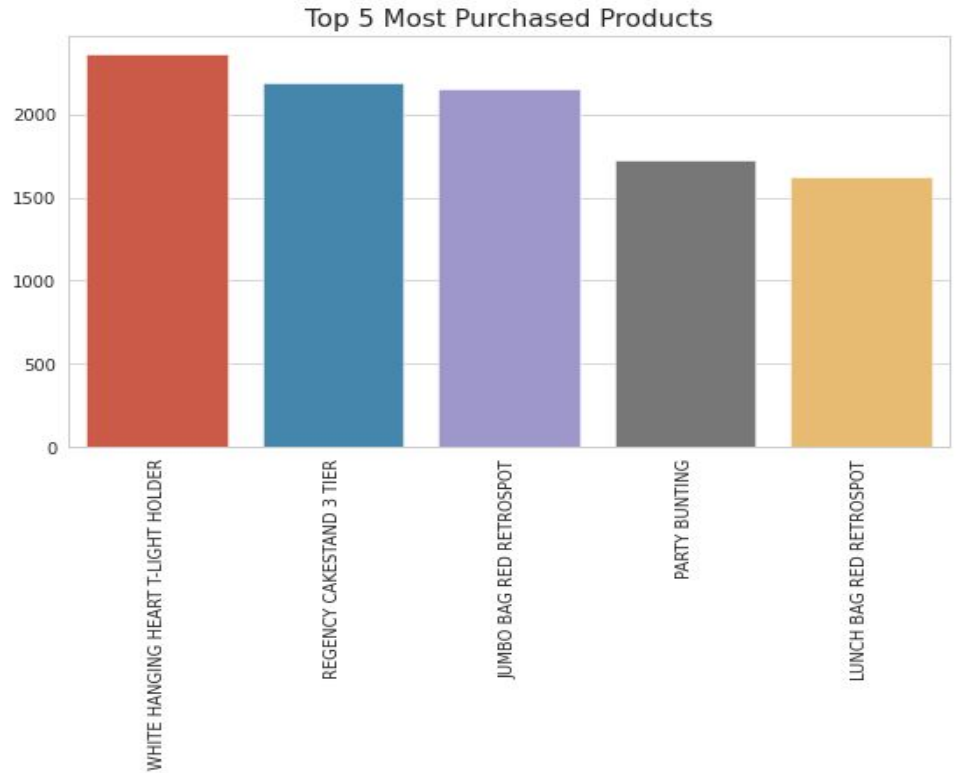
- Removing duplicate entries.
- Dropping Invoice observations ending with ‘c’ (cancelled).
- Removing observations having Quantity & Unit Price less than zero.
- Extracting day, weekday, month, year from Invoiceno.
- Removing all null values in the dataset.

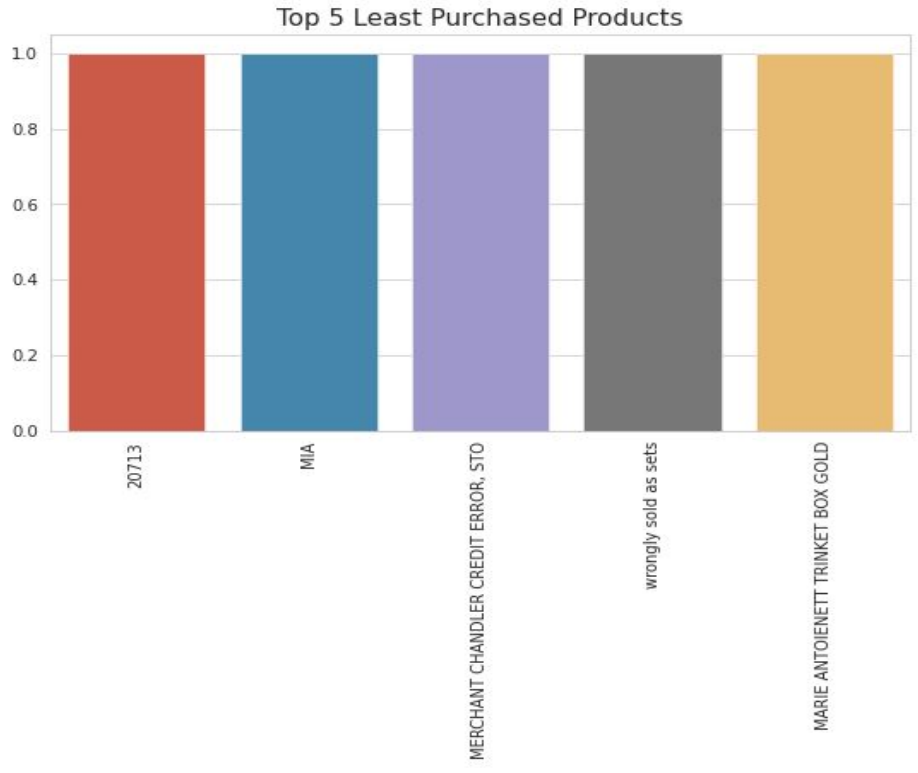
Exploratory Data Analysis

- This online store is serving 4373 customers and operating in 38 countries. The performance of those countries is :-

EDA(continued)

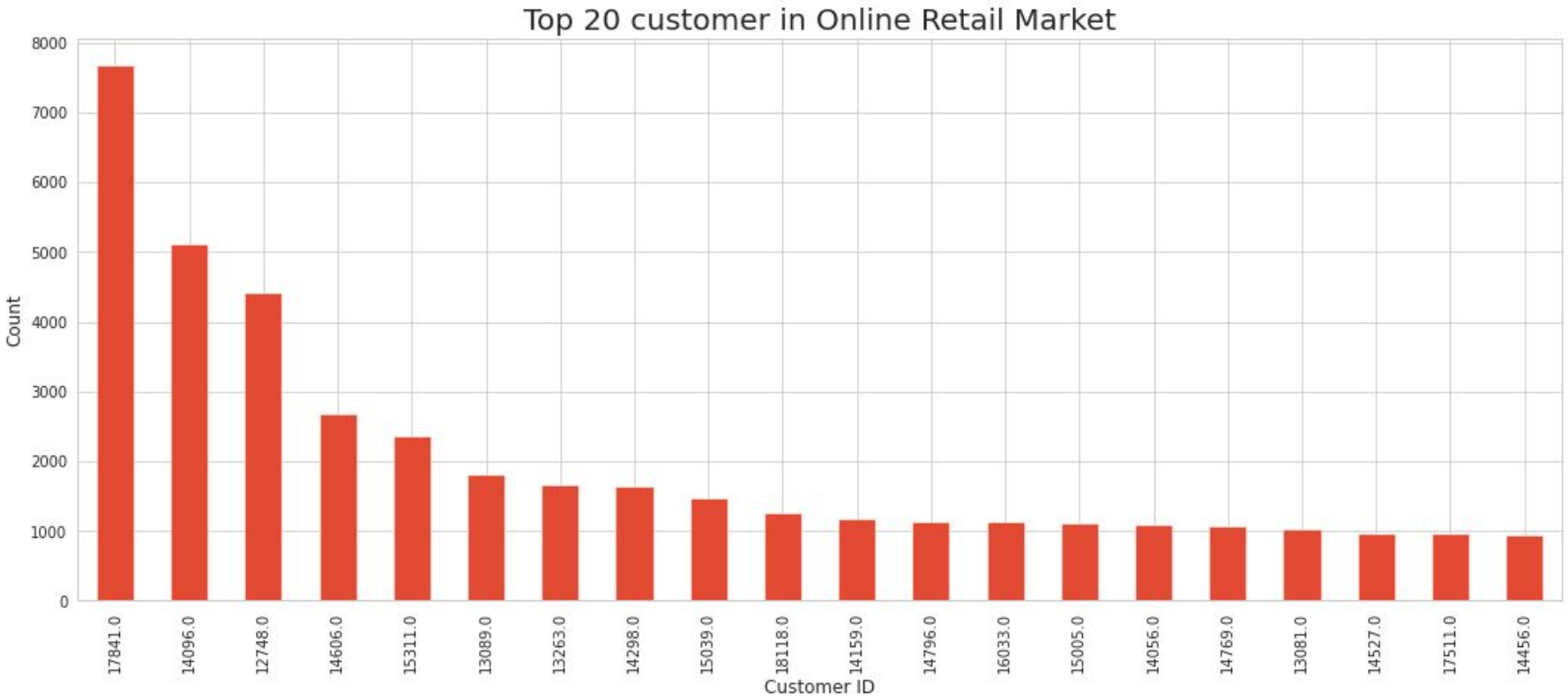
- Analysing Products sold in retail store :-





EDA(continued)

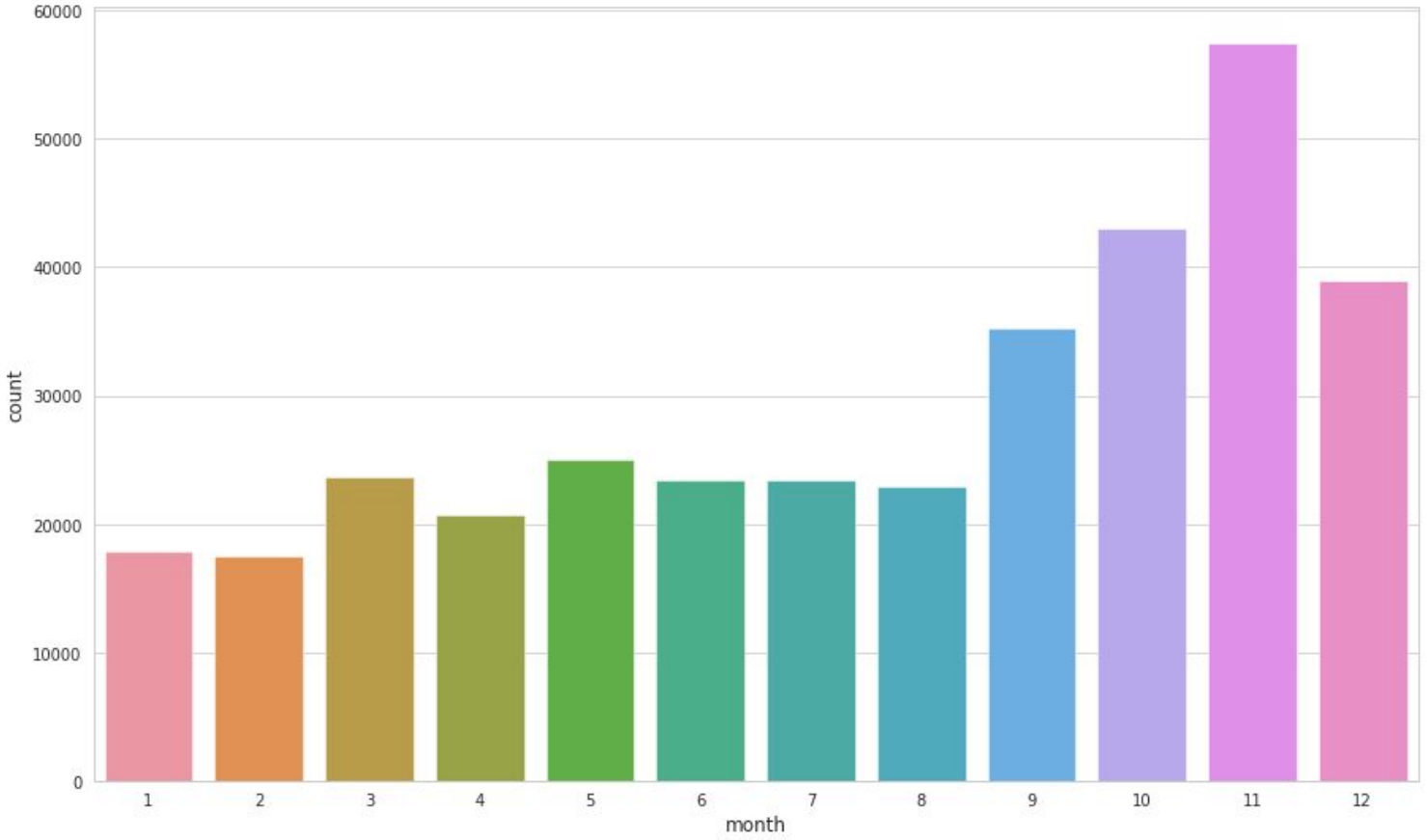
Analysing customers in our dataset



EDA(continued)

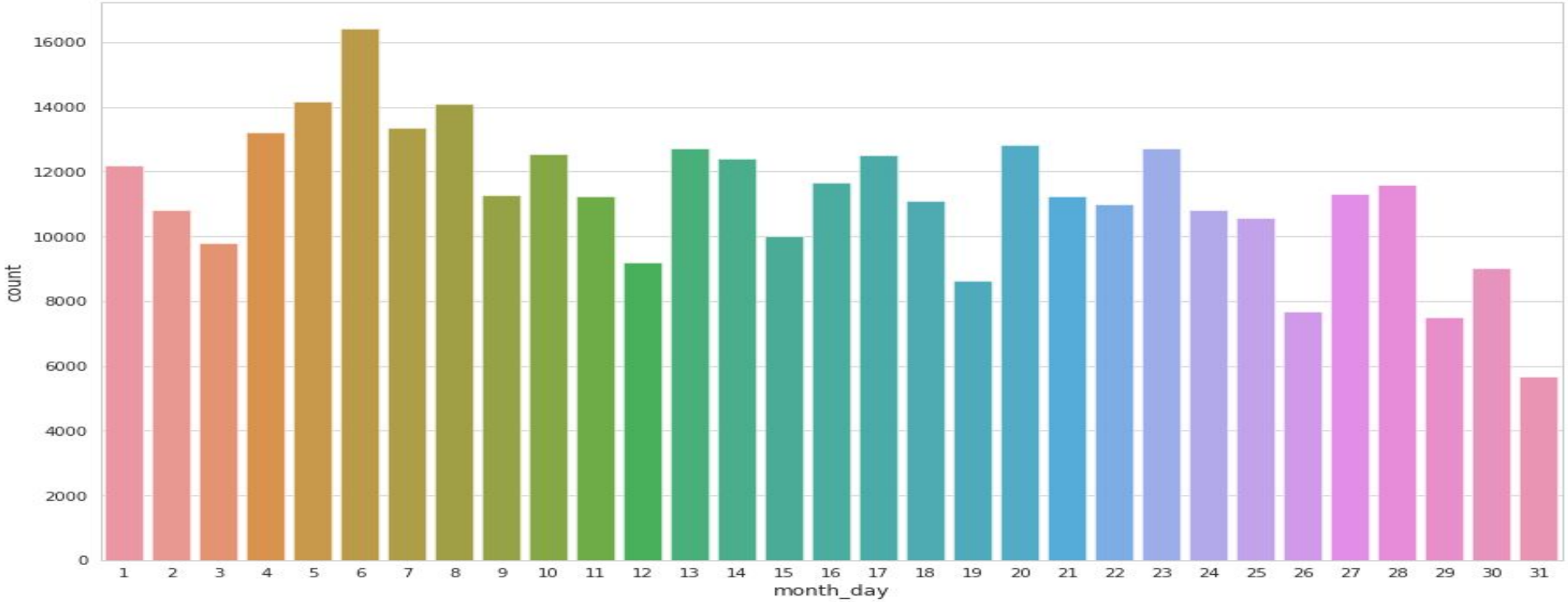
Monthly trend of customers. In last quarter of the year, store is performing good especially in November.

Generating sales count close to 60000.



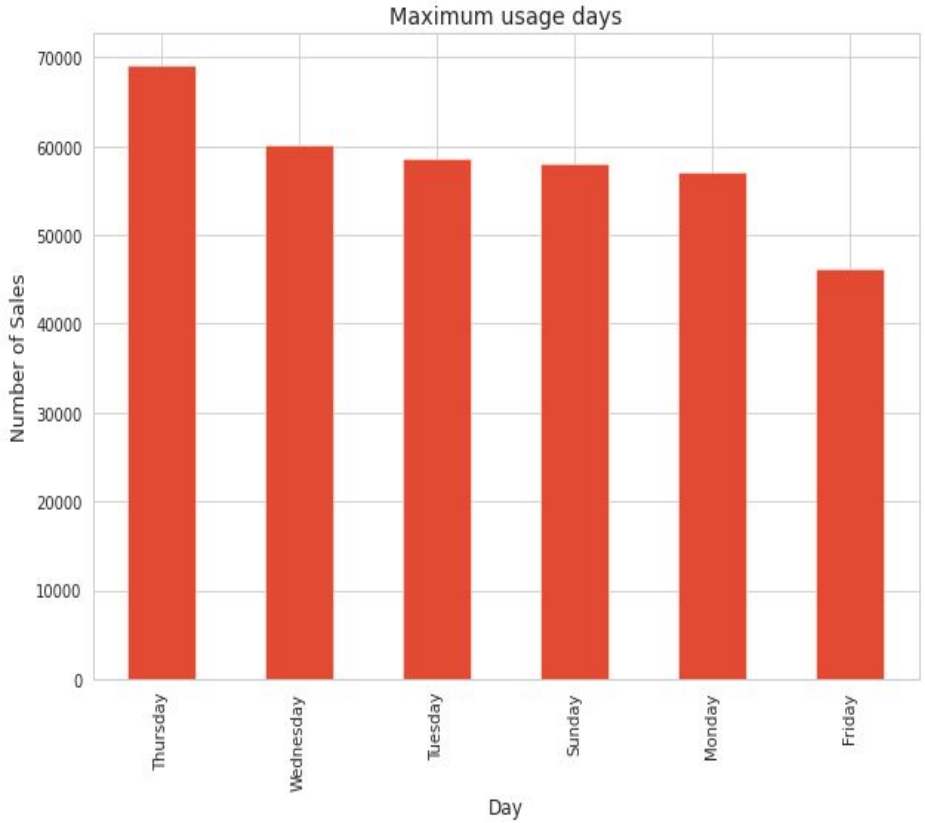
EDA(continued)

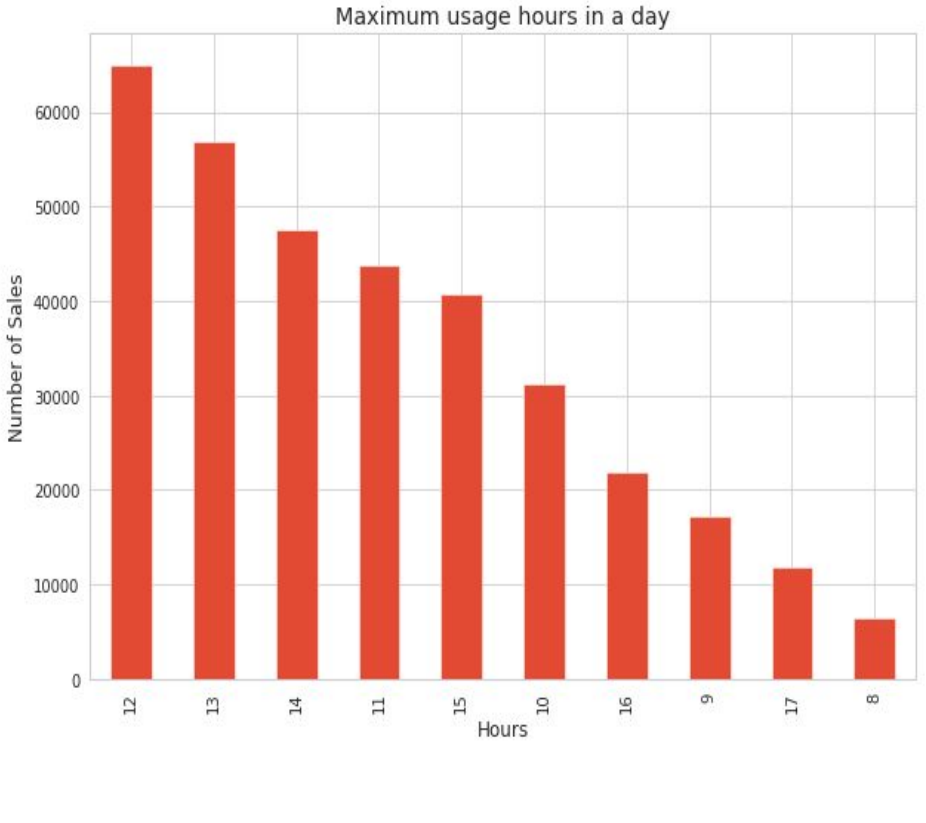
We are experiencing maximum number of customers in first week of the month especially in the first weekend.



EDA(continued)

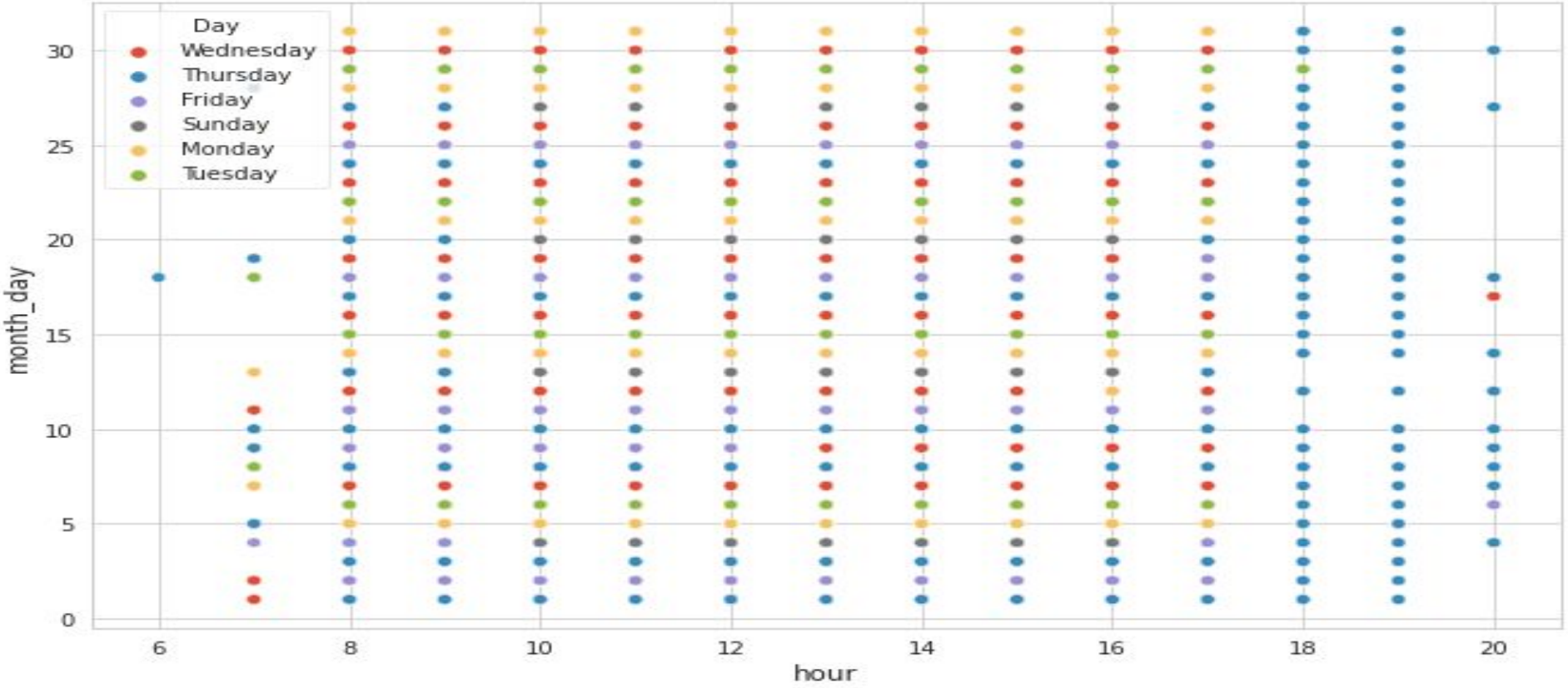
Day and hourly trend of customers





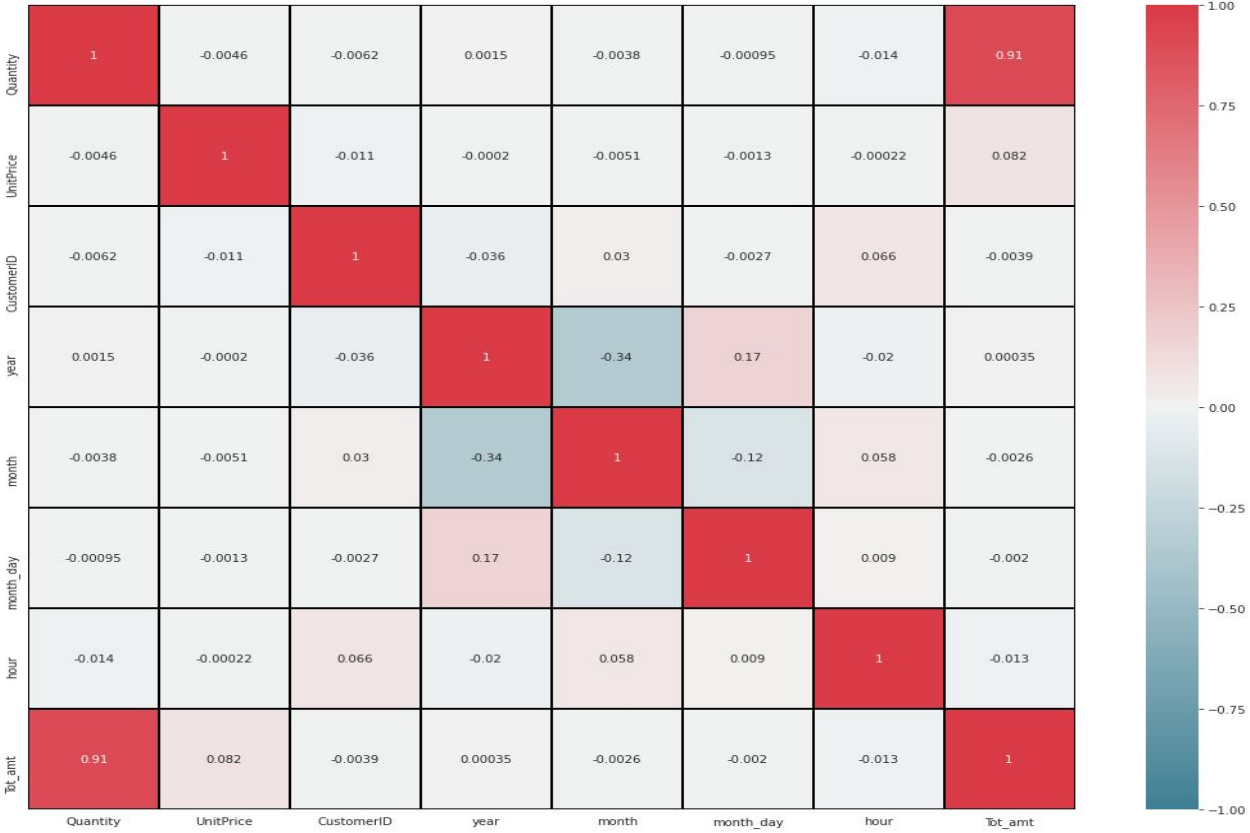
EDA(continued)

Monthly days and hourly trend in sales. On Thursday evening after 5 p.m. the retail is having more number of customers.



EDA(continued)

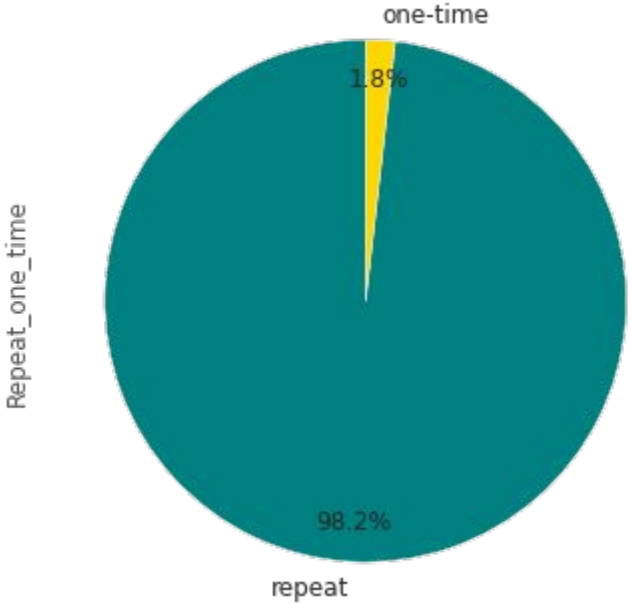
Heatmap of dataset

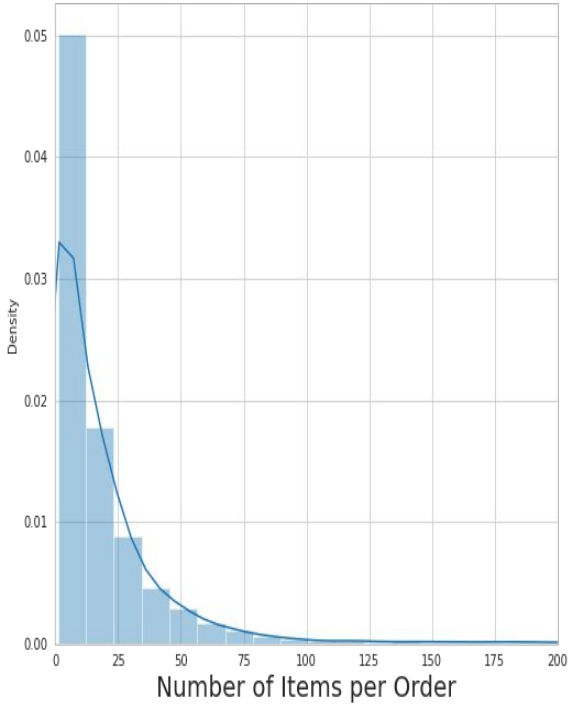


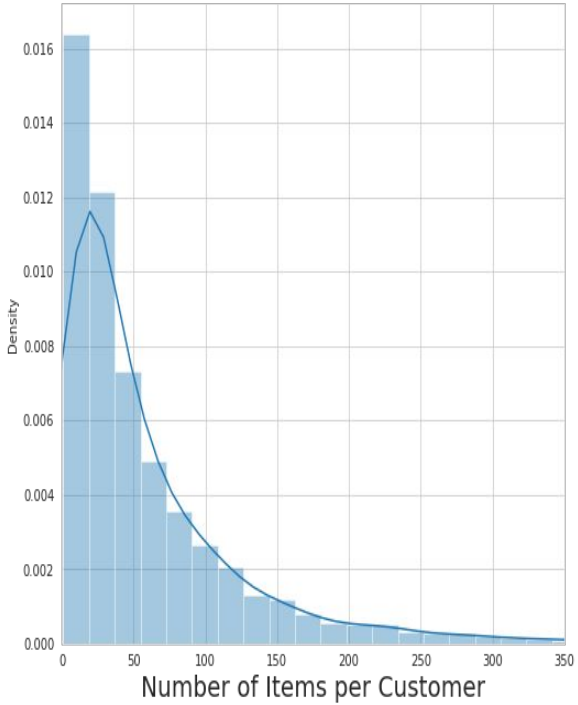
One interesting thing to observe here is that CustomerId and hour also has a correlation that means there is a relation between specific customer and his/her visiting hour in online retail store.

EDA(continued)

Repetition, Orderwise and Items distribution.
We have skewed left distributions for both plots. The average number of items per order is 20.5 & the average number of items per customer is 61.2.







Feature Engineering

Taking United Kingdom as the only country for further analysis because-

Percentage of customers from the UK: 90.35 %

Number of transactions: 23494

Number of products bought: 4065

Number of customers: 3950

Introducing Recency, Frequency & Monetary values with their corresponding quartile.

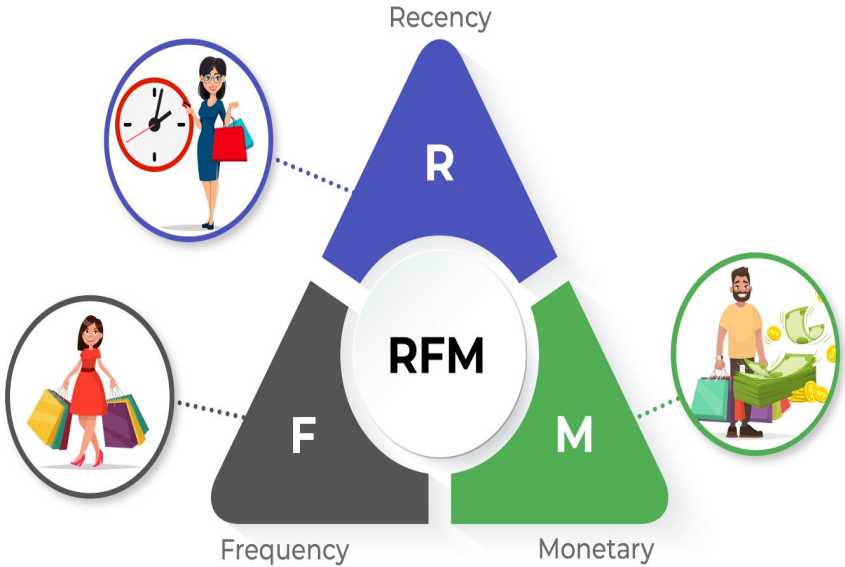
Adding new feature as rfm_rating for evaluation of customer segmentation.

RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

RFM

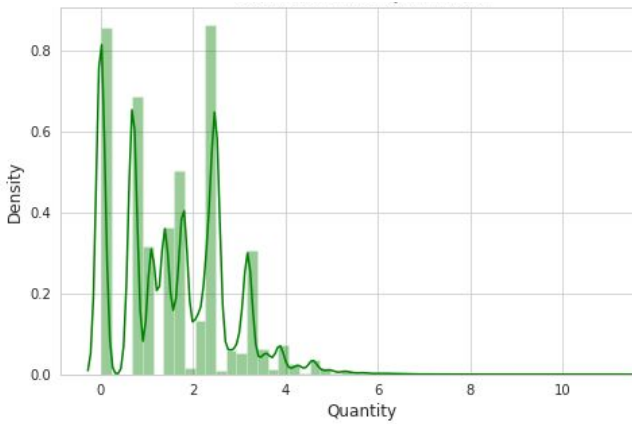
- **Recency = Latest Date - Last Invoice Data**
- **Frequency = count of invoice no. of transaction(s)**
- **Monetary = Sum of Total Amount for each customer**

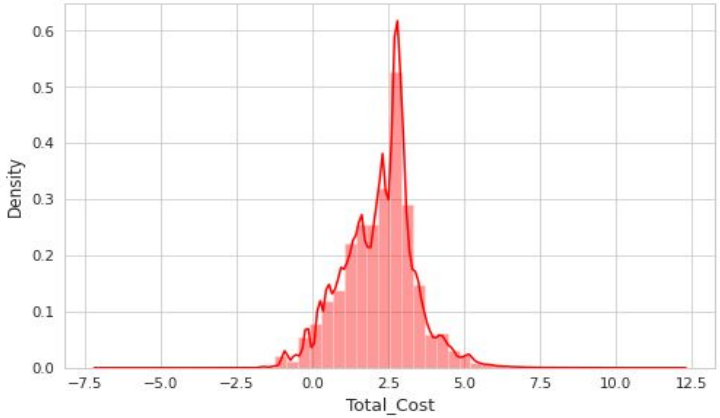
Customers with the lowest recency, highest frequency and monetary amounts considered as top customers.

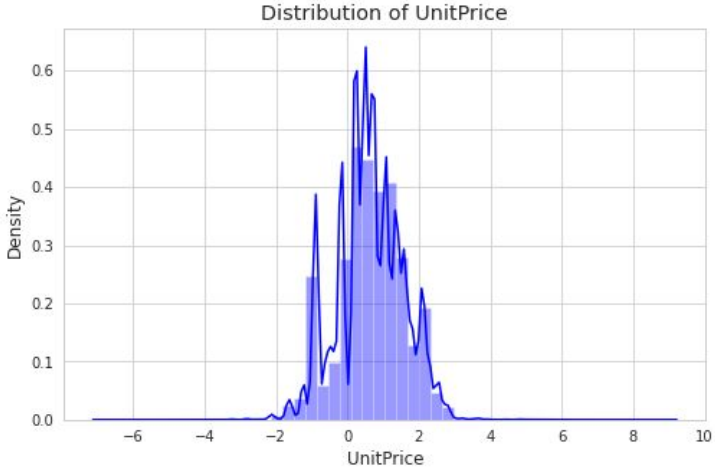


RFM(contd..)

Distribution of various entities in our dataset

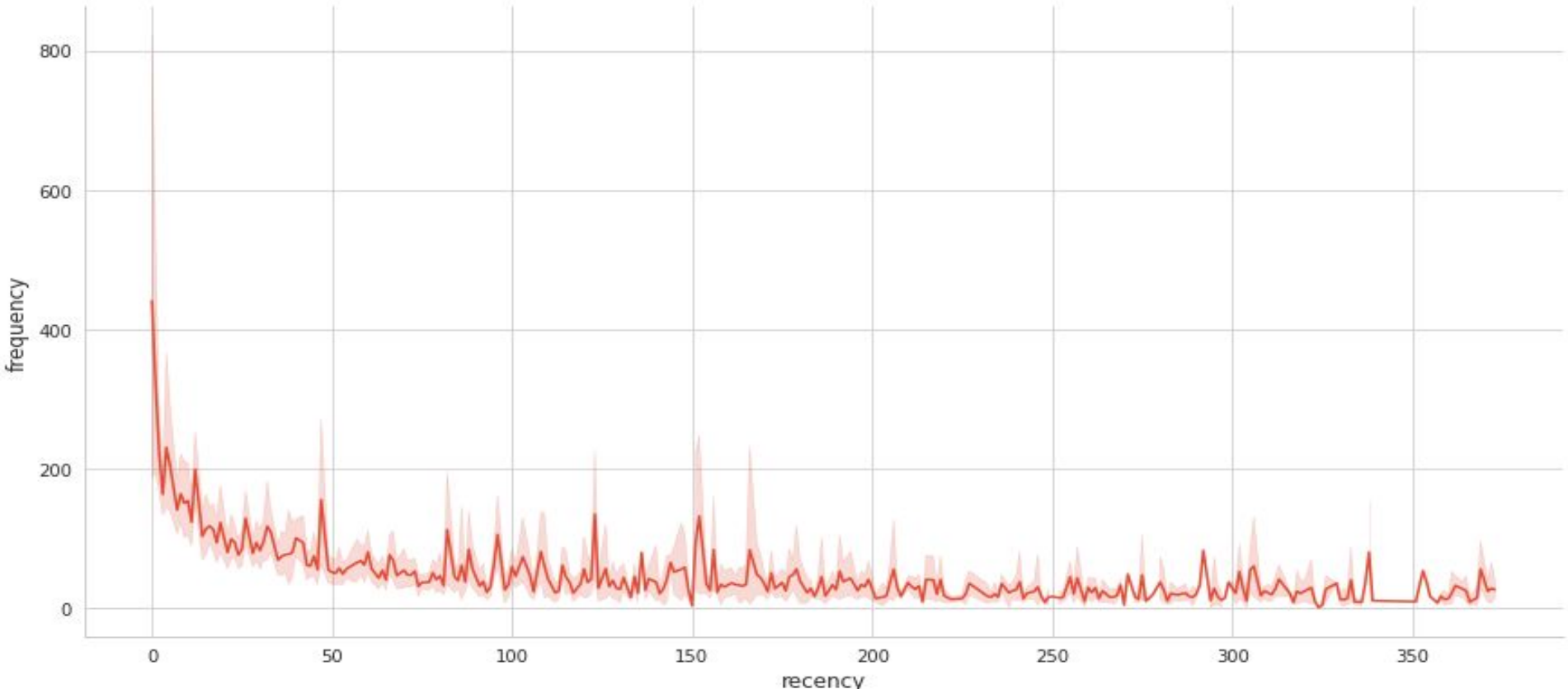






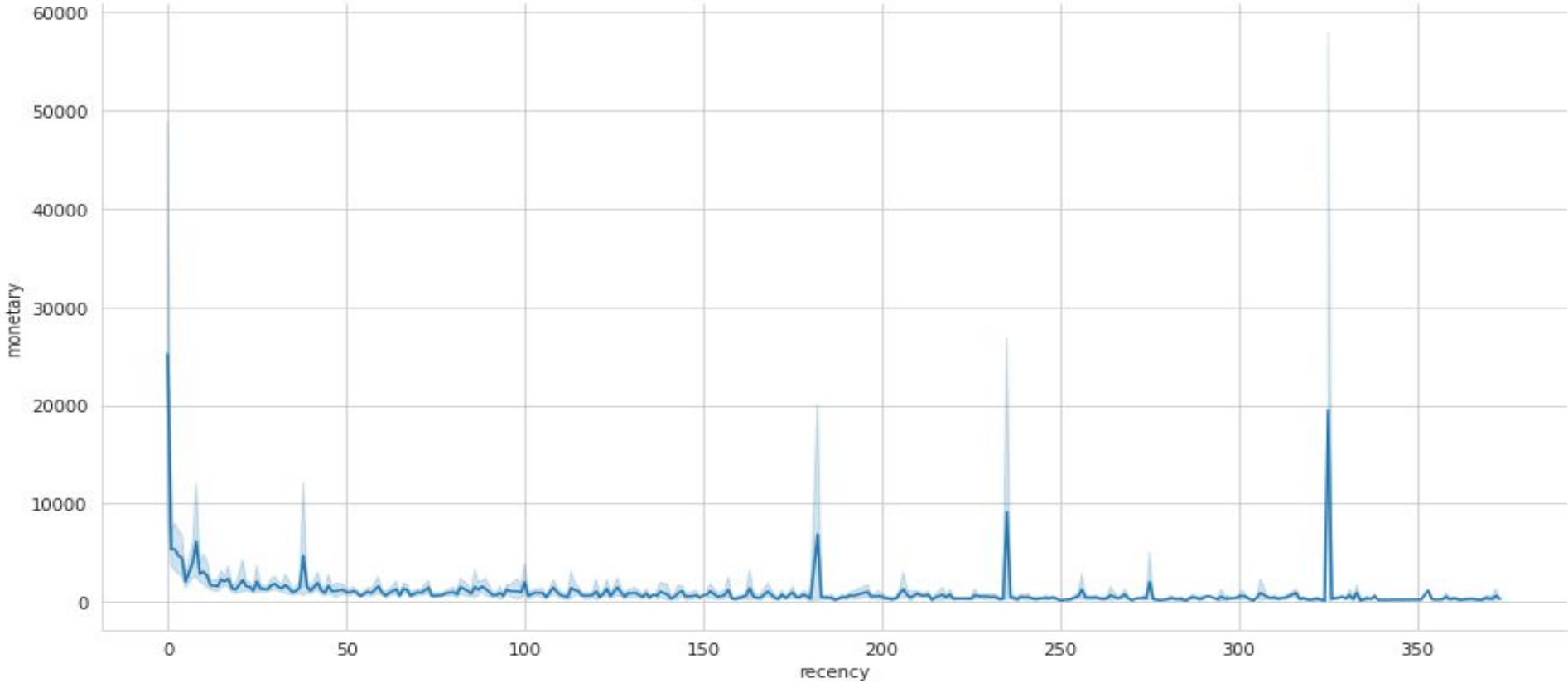
RFM(contd..)

How recency and frequency are related?



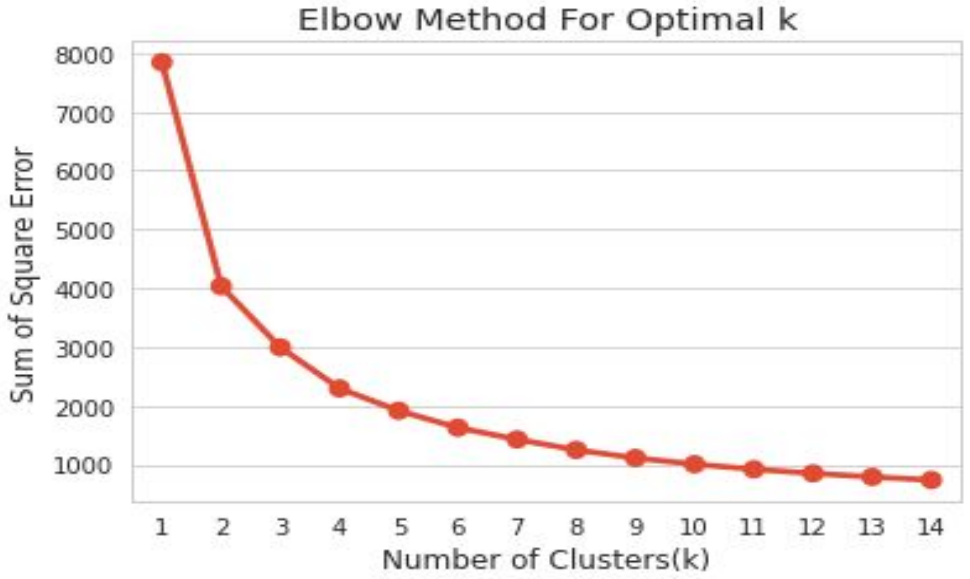
RFM(contd..)

How recency and monetary are related?

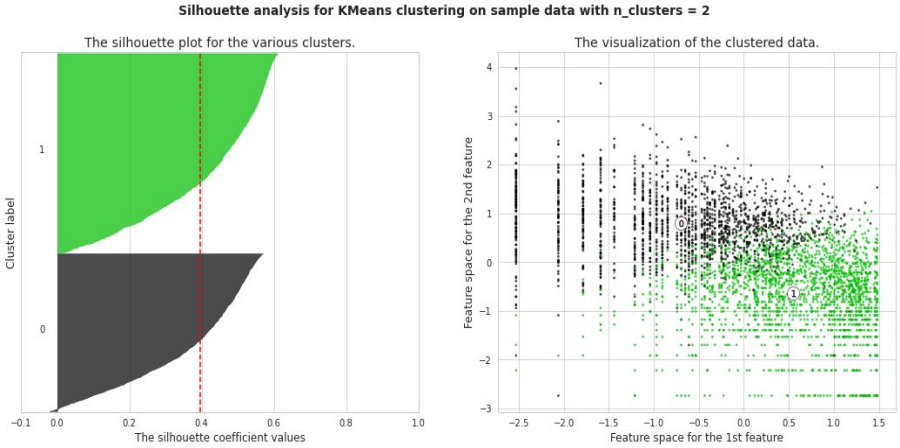


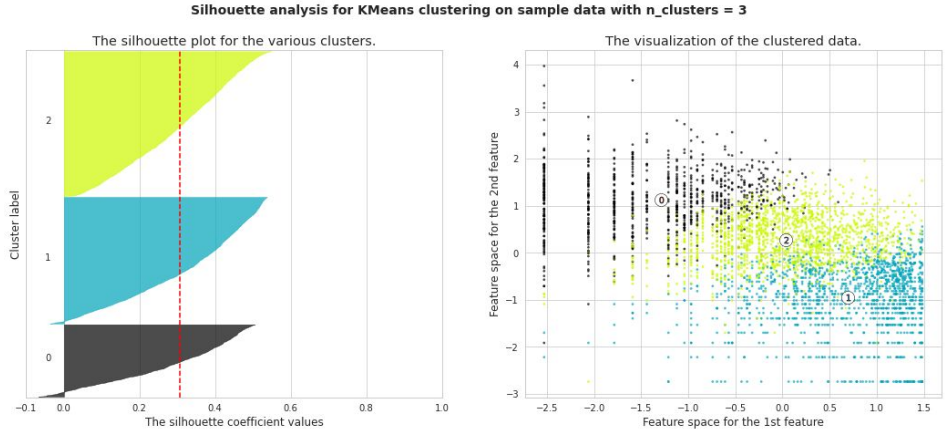
Model implementation

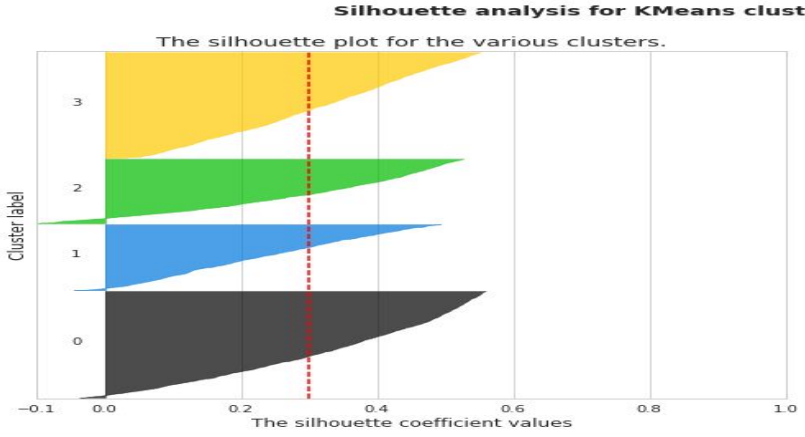
K means algorithm for:-
n = 2, silhouette score is 0.4224
n = 3, silhouette score is 0.3463
n = 4, silhouette score is 0.3661



Model implementation (contd..)

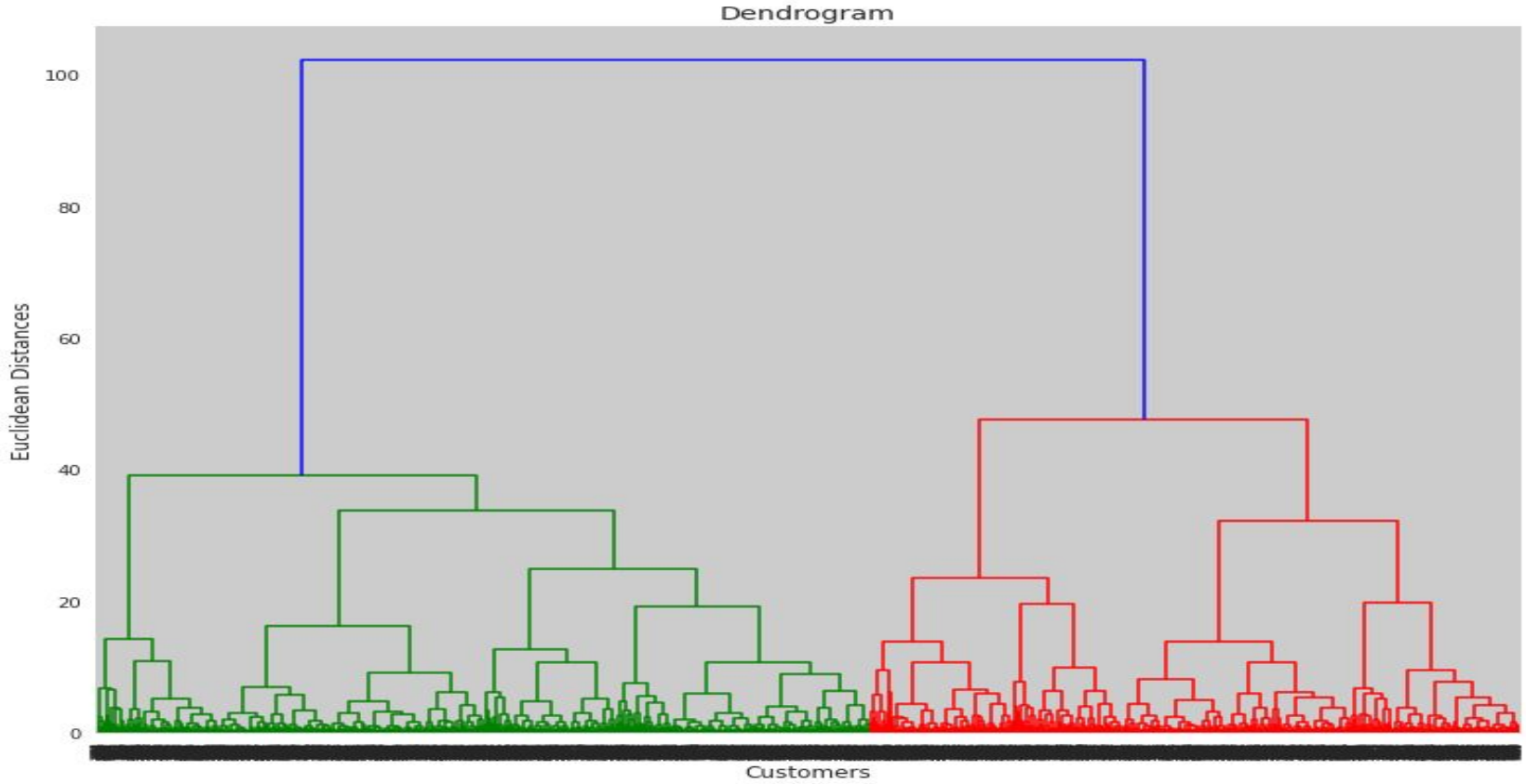




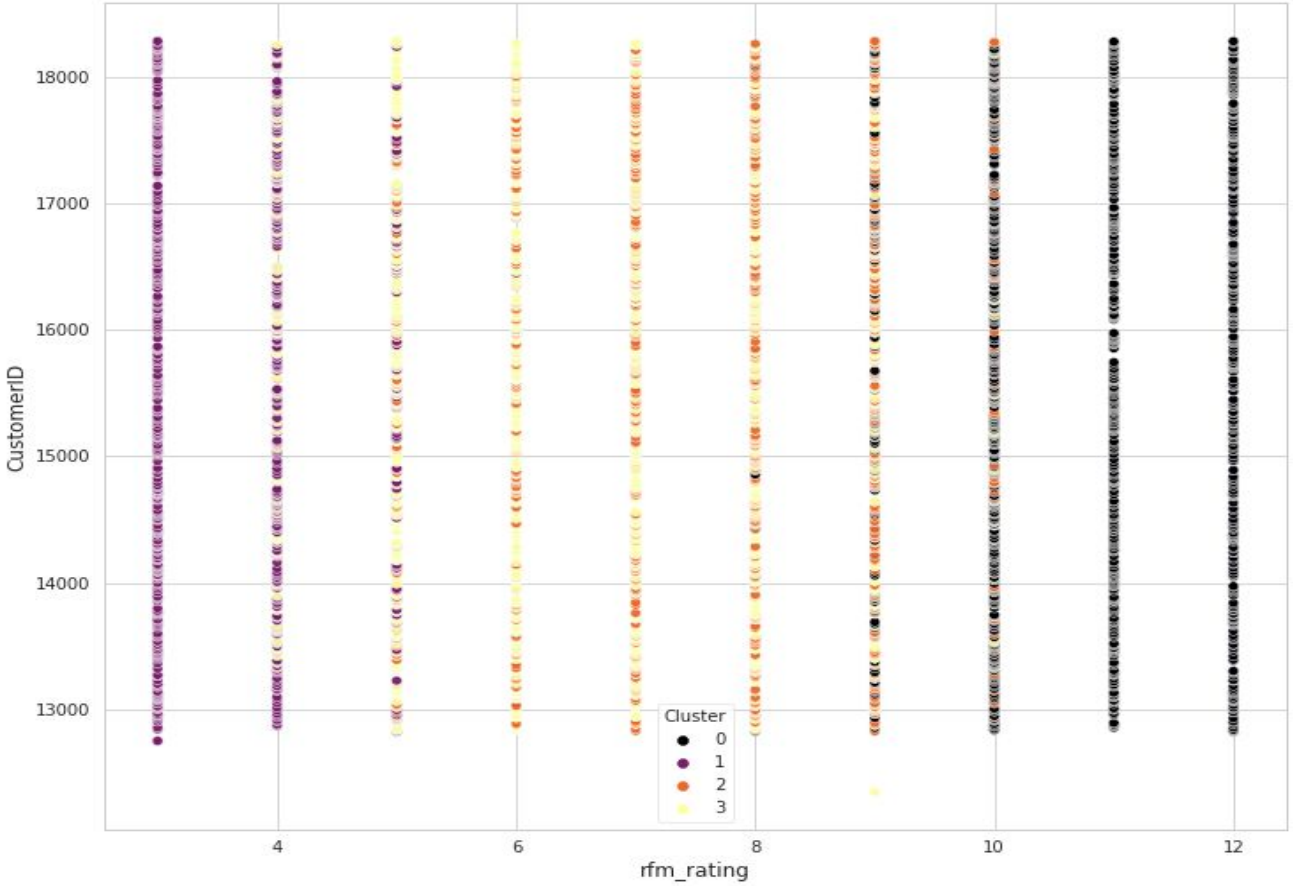


Model implementation(contd...)

Hierarchical clustering:-



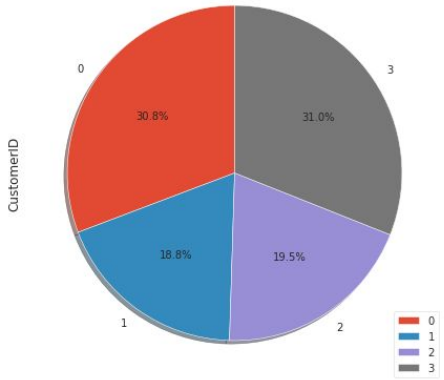
Model Implementation(contd..)

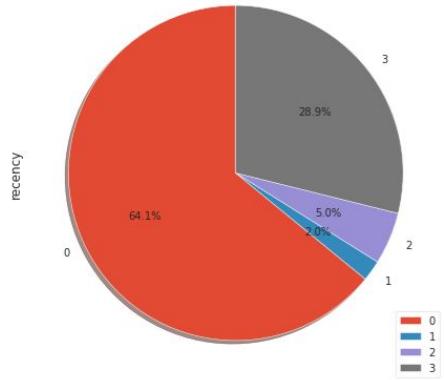


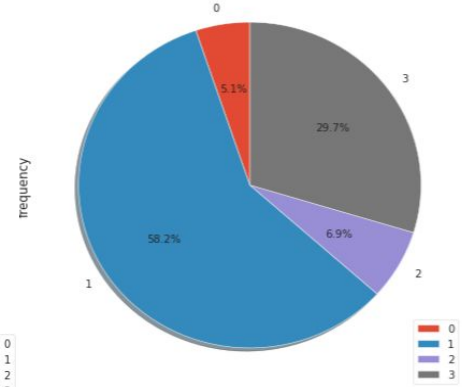
K means algorithm for:-
n = 4,

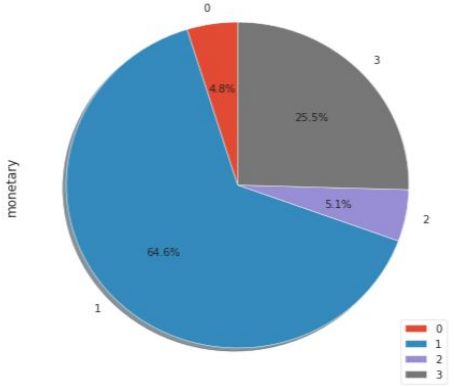
We have cluster 0 as the highest rfm_rating with cluster 1 as lowest rfm_rating.

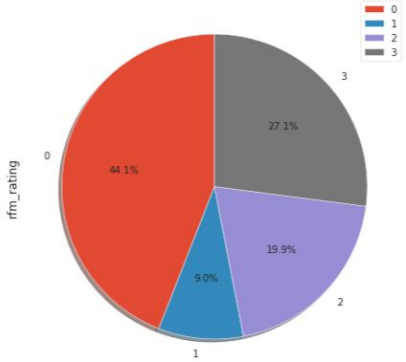
Model Implementation(contd..)











This shows that we have 31% of customers belonging to 0th cluster(first) and they also contribute to our 5% monetary value. Also note that 65% of our monetary value comes from cluster 1(second) having very less rfm rating.

Conclusion

- We can make 2 to 4 cluster according to business requirement.

- Majority of customers are from United Kingdom.

- The month of November has highest sales similarly Thursday have highest sales during the week. Number of customers are more during afternoon session.

- We can do marketing, strategic planning according to our cluster formation to have more effective business growth. We can segment 4 clusters as low, mid, high spenders and customers at risk of churn.

Thank You