Capstone Project-4

Online Retail Customer Segmentation

Team Members

Mohd Danish
Abdul Rahman Talha

Table of Contents:

- Introduction and Problem statement
- Data Description
- Data Cleaning
- Feature Engineering
- Exploratory Data Analysis
- Data Transformation
- RFM Table for Customer ID
- Model Building (Clustering)
- Clustering Profiling
- Conclusion

Introduction

All over the world business is growing in every field. With the help of online platform and new technologies it strengthen its root more deeper, having access to wider market and large customers. Online Retail business is also one of them, Customer segmentation refers to categorizing customers into different groups with similar characteristics. It can help to each customer group in a different way, in order to maximize their business and deliver true services who actually deserve

Problem Statement

- Identify major customer segments on UK based online retail dataset.
- Create RFM table

Data Description

We have been provided with UK-based and registered online retail company which contains transaction between 01/12/2010 and 09/12/2011 with 541909 instances and 8 features.

- **InvoiceNo:** A 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation.
- **StockCode**: It is a 5-digit number assigned to each distinct product.
- **Descroption:** Name of each product.
- **Quantity:** Number of a particular product sold on per transaction.
- **InvoiceDate:** It holds the information of Date and time when transaction was generated.
- **UnitPrice:** Price per unit of a particular product sold
- **CustomerID:** A 5-digit number uniquely assigned to each customer.
- **Country:** The name of the country where customer resides

Data Cleaning

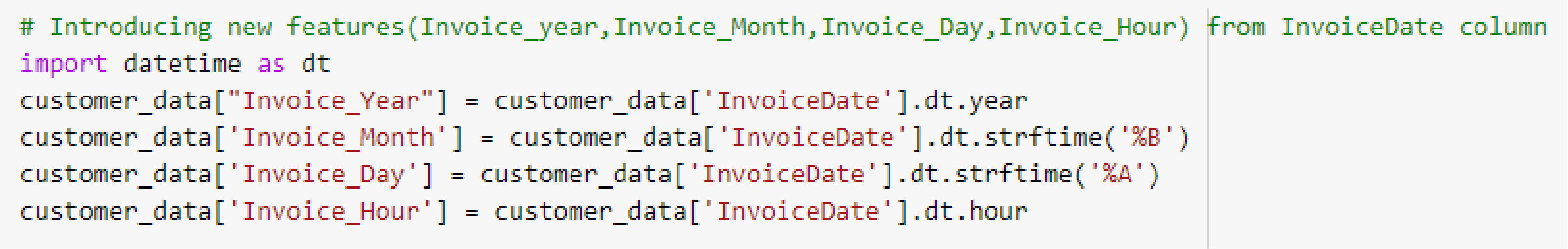
- The dataset has 541909 rows and 8 features(columns).
- Our Dataset possess with high null values in Description and CustomerID columns:

Feature	Percentage of null values
Description	0.2683%
CustomerID	24.92%

We have to drop all null values because each customer IDs are uniquely assigned to a particular customer if it is missing from the dataset means we can impute it with other values it does not make any sense and if we do then we could have ended up with biased results.

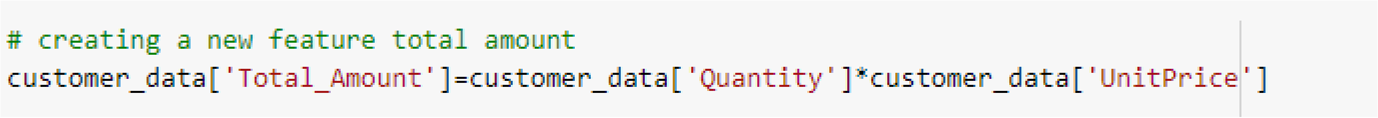
Feature Engineering

Introduced new features: Invoice_Year, Invoice_Month, Invoice_Day, Invoice_Hour from InvoiceDate feature & Total_Amount



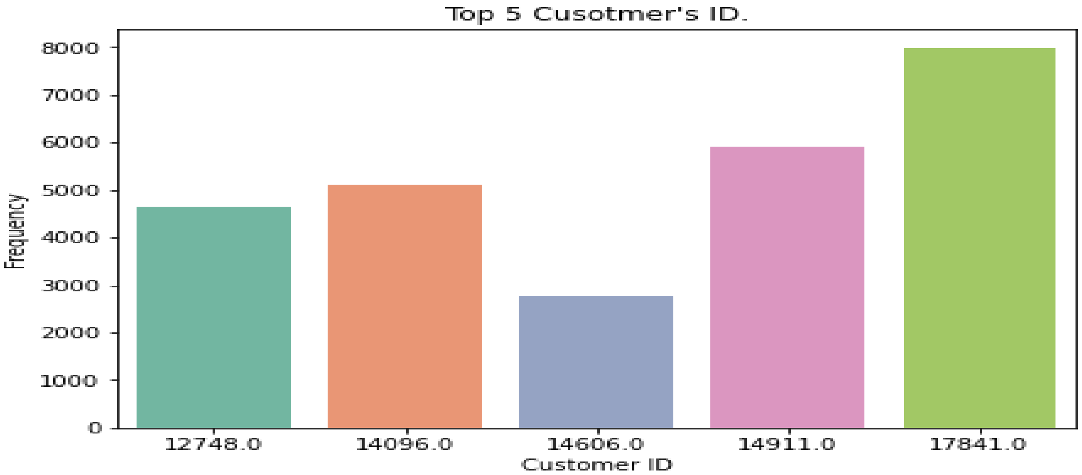
```
# Introducing new features(Invoice_year,Invoice_Month,Invoice_Day,Invoice_Hour) from InvoiceDate column
import datetime as dt
customer_data["Invoice_Year"] = customer_data['InvoiceDate'].dt.year
customer_data['Invoice_Month'] = customer_data['InvoiceDate'].dt.strftime('%B')
customer_data['Invoice_Day'] = customer_data['InvoiceDate'].dt.strftime('%A')
customer_data['Invoice_Hour'] = customer_data['InvoiceDate'].dt.hour
```

$$TotalAmount = Quantity * UnitPrice$$

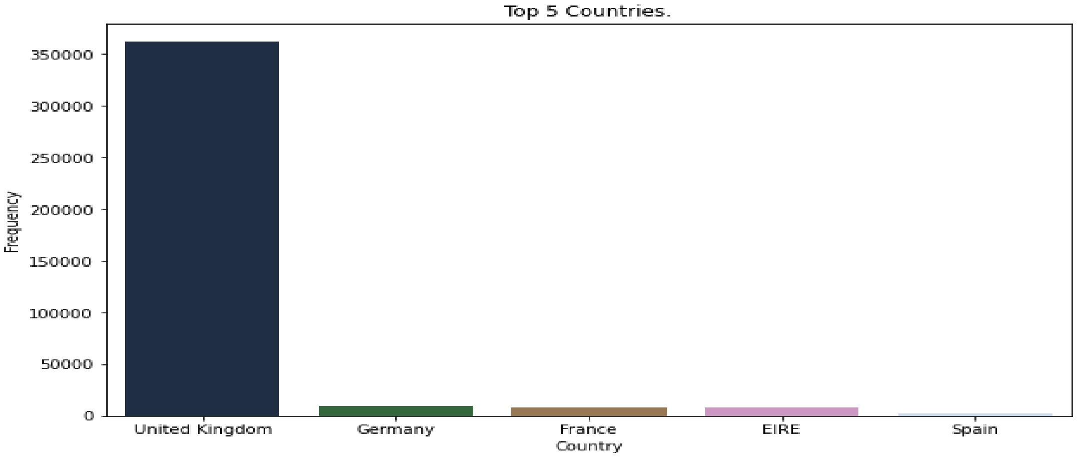


```
# creating a new feature total amount
customer_data['Total_Amount']=customer_data['Quantity']*customer_data['UnitPrice']
```

Exploratory Data Analysis

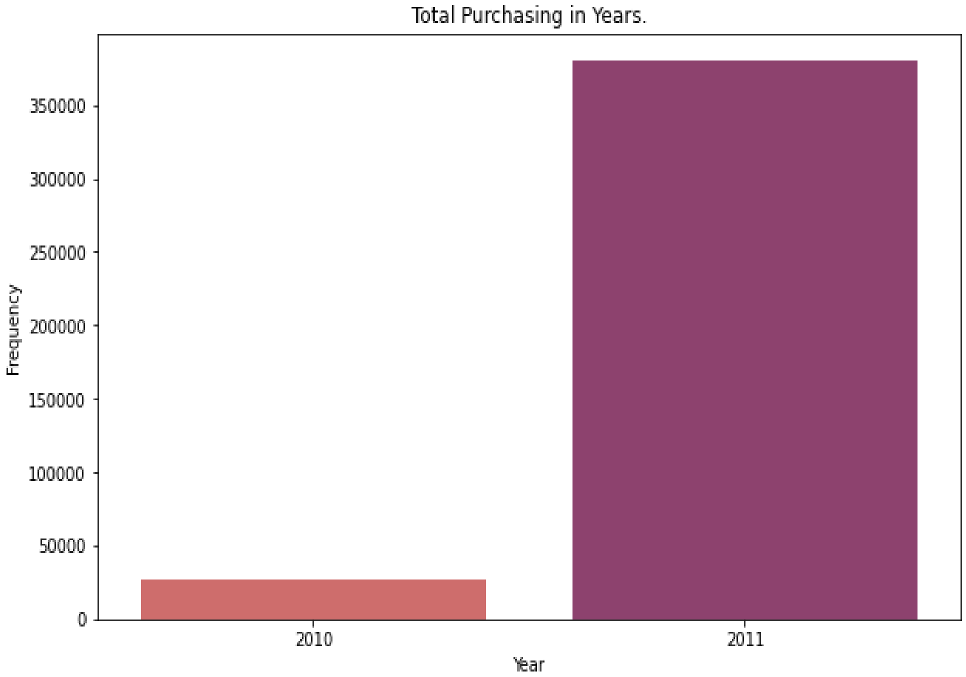


These are the top 5 customer's ID whose frequency of placing order is very high in comparison to others.



Majority of the customers are belong to United Kingdom, Germany, France, Ireland and Spain.
Just because company is established in UK that's why around 90% of orders are placed from UK.

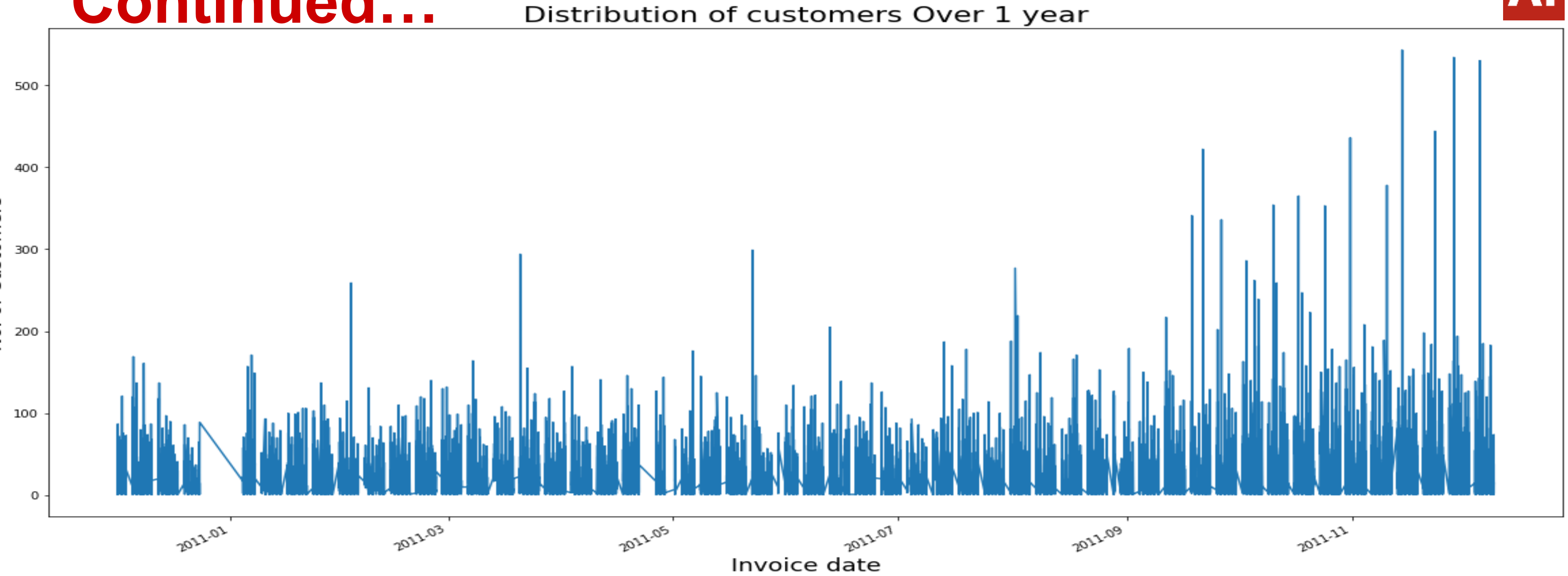
Purchasing in Years



Frequency of purchasing in 2011 is much higher than 2010 because in our dataset we have only December-2010 entries of customers.

Around 3000 orders placed in just one month.

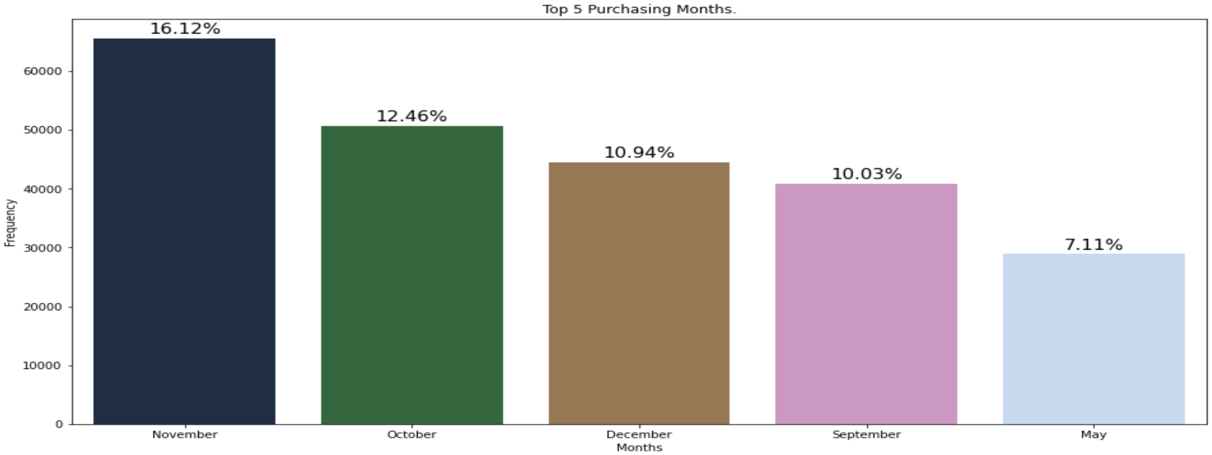
Continued…

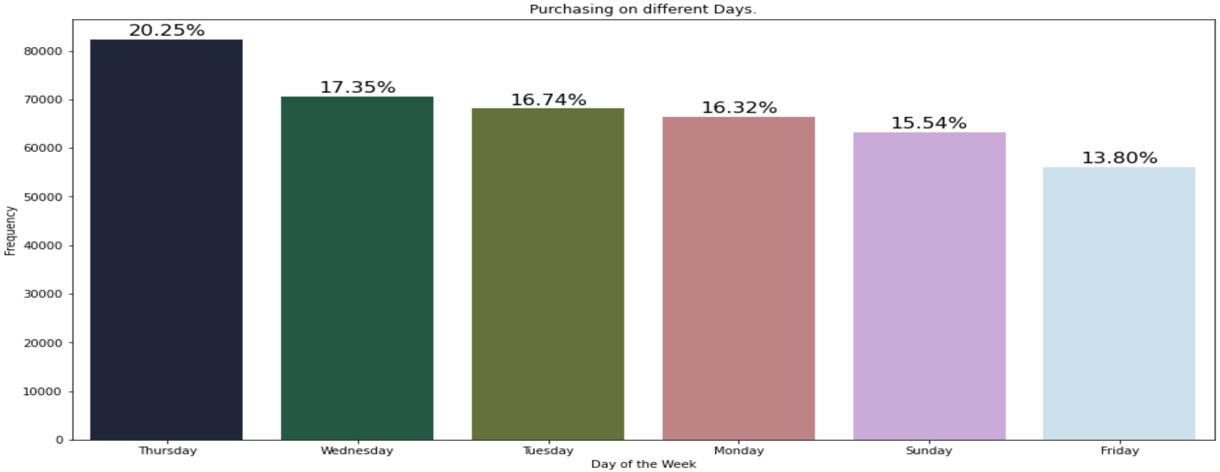


It can be easily concluded from the above graph that number of customers are increasing as we reaching towards the end of the year 2011.

September and November are getting highest purchasing order in comparison to January and March.

Days and Months

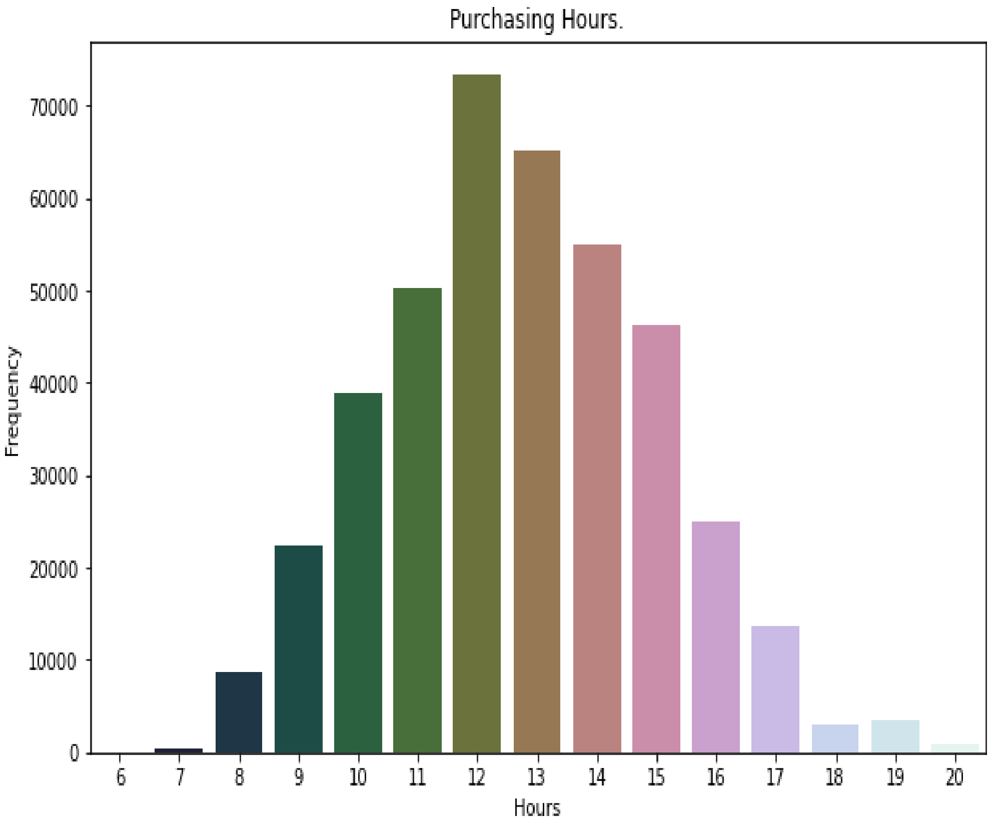




Purchasing by customers are high in Nov, Oct, Dec, Sept and May.
Around 55% of orders are placed in these 5 months.
It may be due to festivals occur in these months.

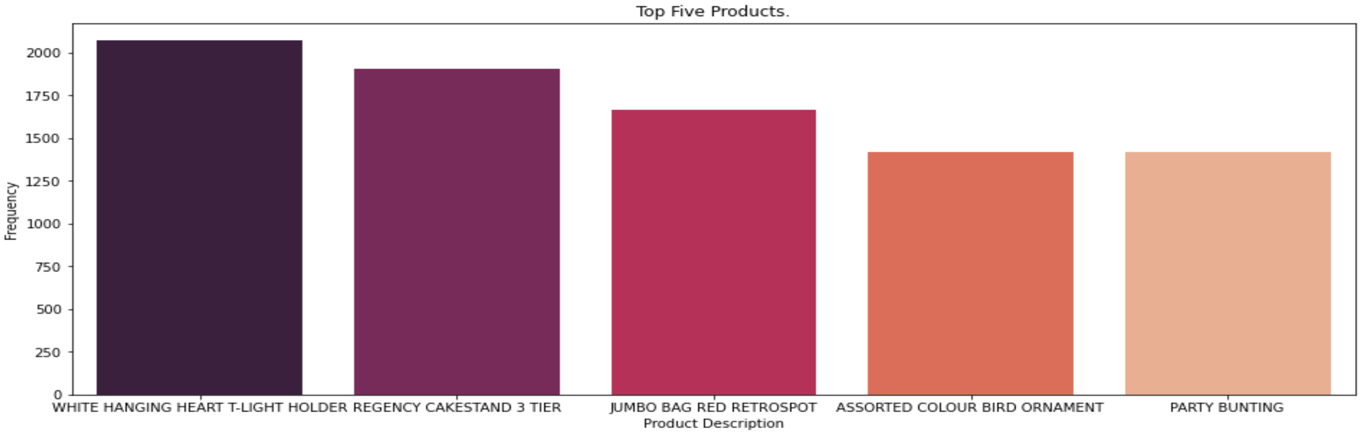
Distribution of purchasing on different days clearly shows that most of the orders were placed on Thursday, Wed, Fri, Tuesday and Monday.

Purchasing at different hours

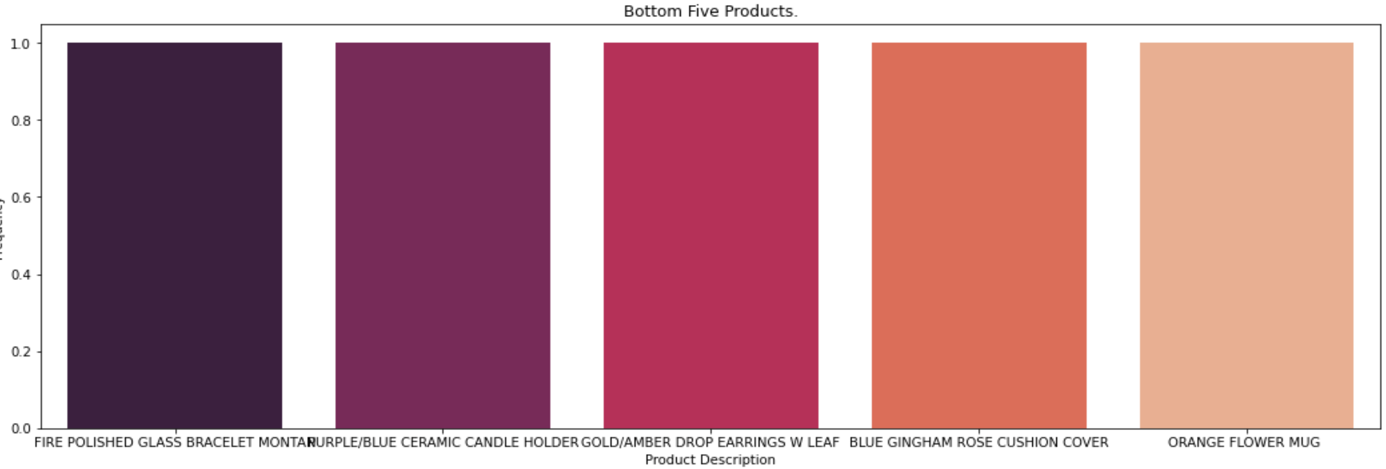


- As it can be seen easily from this countplot customers usually purchase in between 10:00 A.M to 2:00 P.M.
- There are very few purchases in early morning and at midnight.

Top products



Bottom Five Products



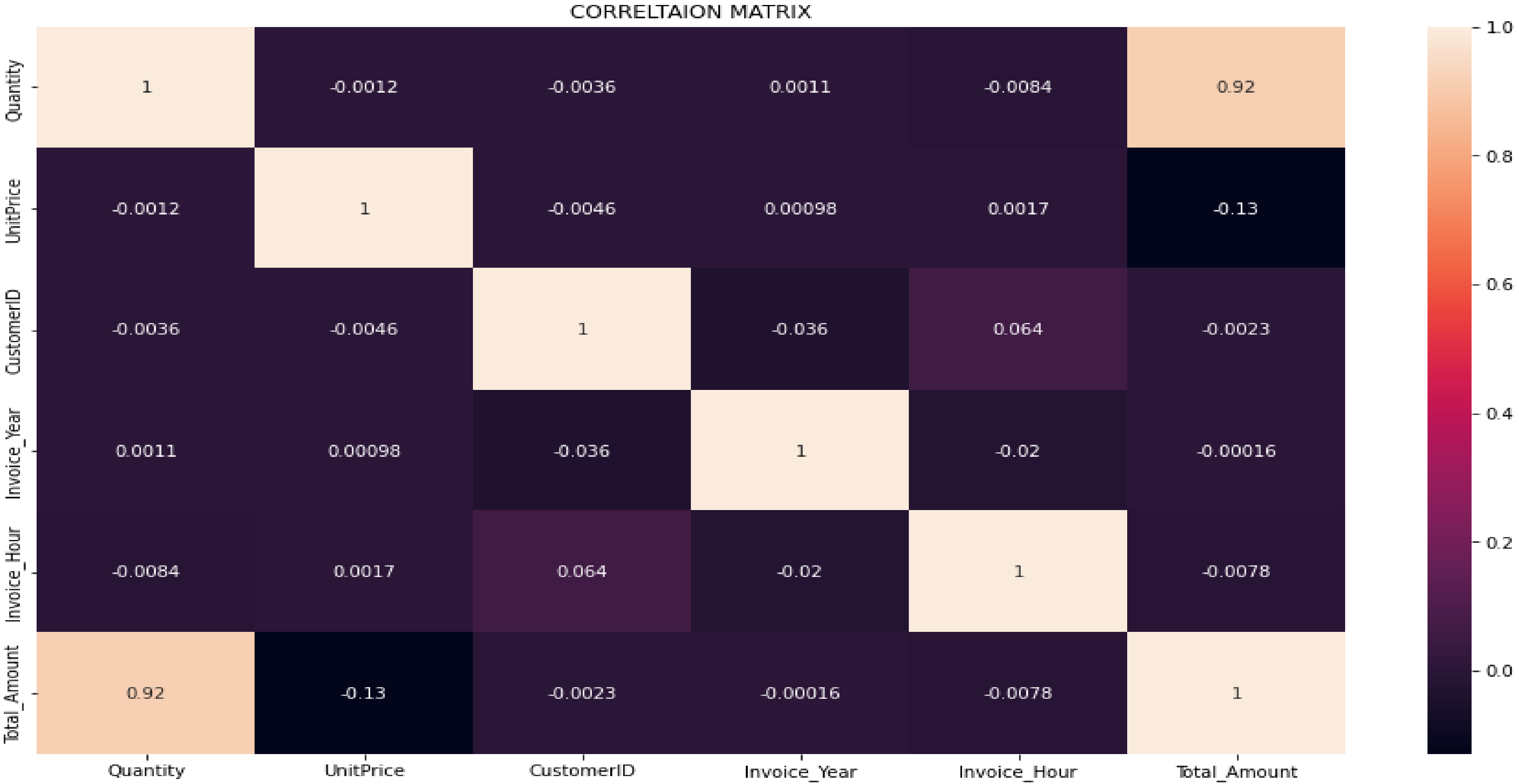
Top sold products

- WHITE HANGING HEART T-LIGHT HOLDER
- REGENCY CAKESTAND 3 TIER
- JUMBO BAG RED RETROSPOT
- ASSORTED COLOUR BIRD ORNAMENT
- PARTY BUNTING

Least sold products

- FIRE POLISHED GLASS BRACELET MONTAN
- PURPLE/BLUE CERAMIC CANDLE HOLDER
- GOLD/AMBER DROP EARRINGS W LEAF
- BLUE GINGHAM ROSE CUSHION COVER
- ORANGE FLOWER MUG

CORRELATION MATRIX



CORRELTAION MATRIX

	Quantity	UnitPrice	CustomerID	Invoice_Year	Invoice_Hour	Total_Amount
Quantity	1	-0.0012	-0.0036	0.0011	-0.0084	0.92
UnitPrice	-0.0012	1	-0.0046	0.00098	0.0017	-0.13
CustomerID	-0.0036	-0.0046	1	-0.036	0.064	-0.0023
Invoice_Year	0.0011	0.00098	-0.036	1	-0.02	-0.00016
Invoice_Hour	-0.0084	0.0017	0.064	-0.02	1	-0.0078
Total_Amount	0.92	-0.13	-0.0023	-0.00016	-0.0078	1

DATA TRANSFORMATION

- In this section, a Recency, Frequency and Monetary (RFM) is created.

R(Recency): Number of days since last purchase

F(Frequency): How frequent customers are(How many times they purchased)

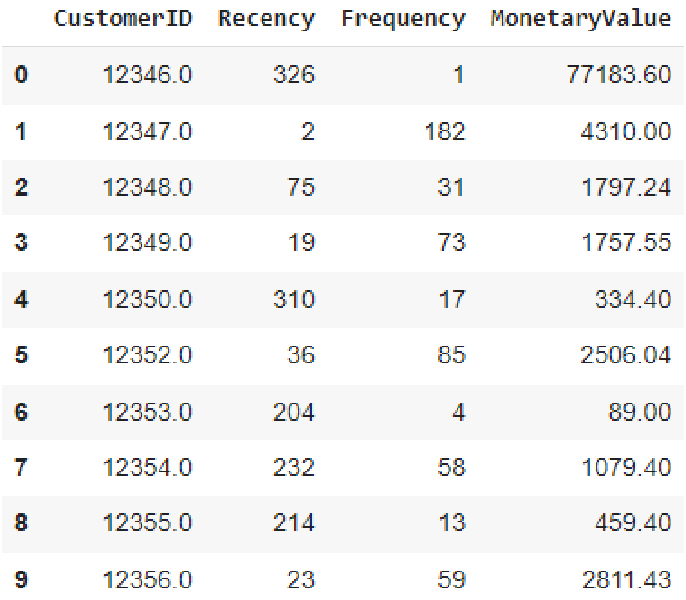
M(Monetary): Total amount of transactions (revenue contributed)

- The RFM DataFrame is grouped on the basis of customer ID. The data now contains 4192 rows or customers.
- There are some outliers in the RFM table, to cure this problem we've applied IQR(Inter Quantile Range) method.

IQR = Q3 – Q1

Q3: Third quartile

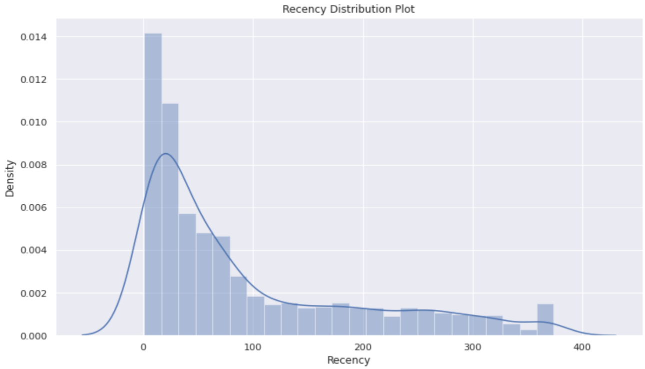
Q1: First quartile

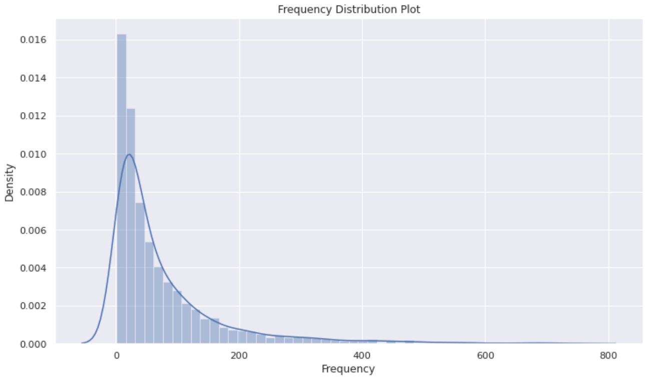


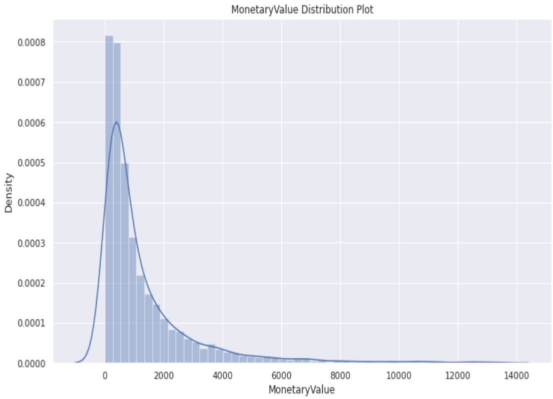
	CustomerID	Recency	Frequency	MonetaryValue
0	12346.0	326	1	77183.60
1	12347.0	2	182	4310.00
2	12348.0	75	31	1797.24
3	12349.0	19	73	1757.55
4	12350.0	310	17	334.40
5	12352.0	36	85	2506.04
6	12353.0	204	4	89.00
7	12354.0	232	58	1079.40
8	12355.0	214	13	459.40
9	12356.0	23	59	2811.43

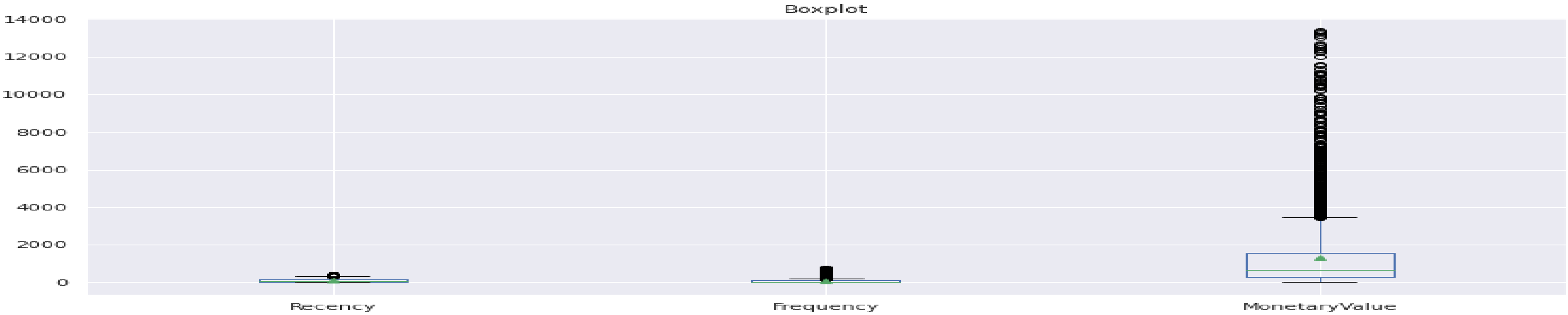
Continued:

After applying IQR on RFM most of the outliers are removed from the dataset

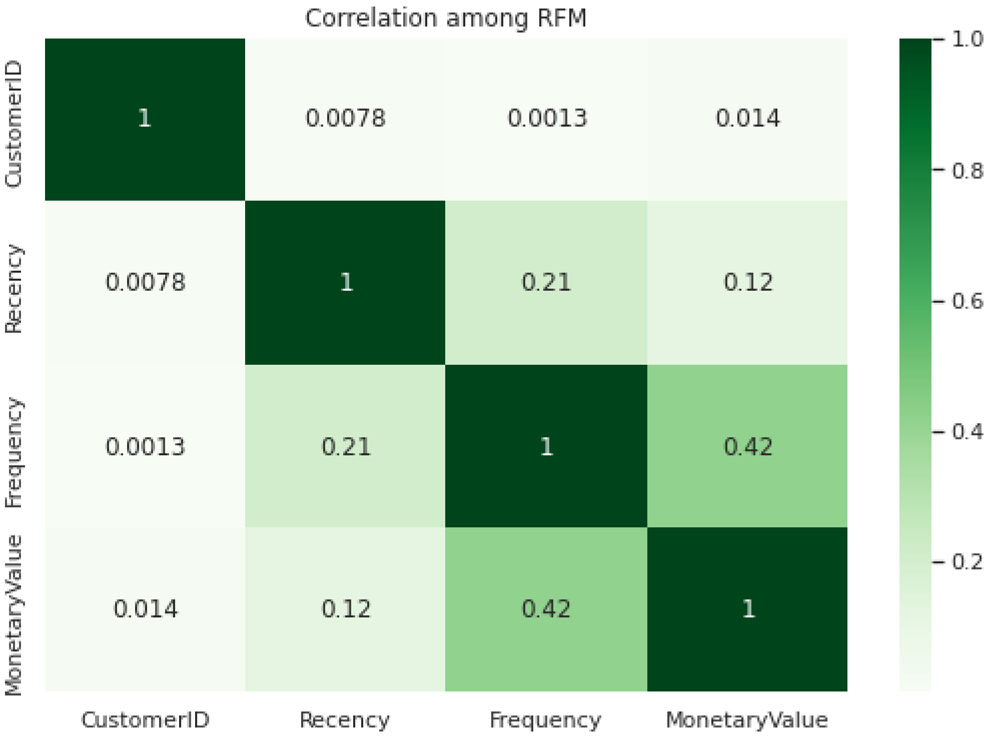








Correlation among RFM



Correlation among RFM

	CustomerID	Recency	Frequency	MonetaryValue
CustomerID	1	0.0078	0.0013	0.014
Recency	0.0078	1	0.21	0.12
Frequency	0.0013	0.21	1	0.42
MonetaryValue	0.014	0.12	0.42	1

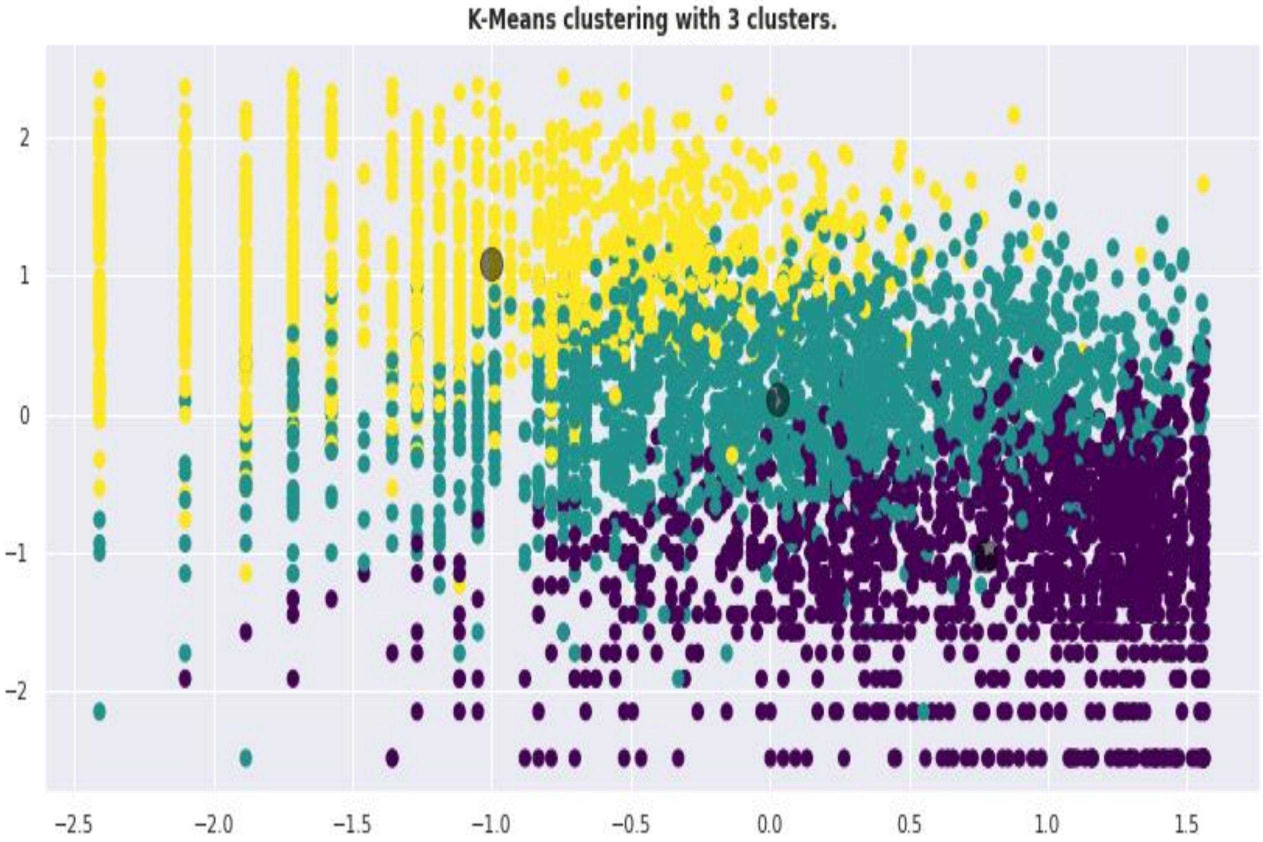
In the correlation matrix of RFM:

- Frequency and Monetary value is positively correlated, somehow frequency of purchasing affects monetary value too.
- Frequency and Recency is also positive correlated but not having very high correlation between them.

Model Building (Clustering)

- In this section, we use K-Means algorithm to cluster the customers into different segments.
- To identify the optimum number of clusters, we use the elbow method and silhouette analysis.
- With both the methods, 3 clusters is optimum in this case.
- A K-Means model with 3 clusters is developed and customers are segmented into different clusters.

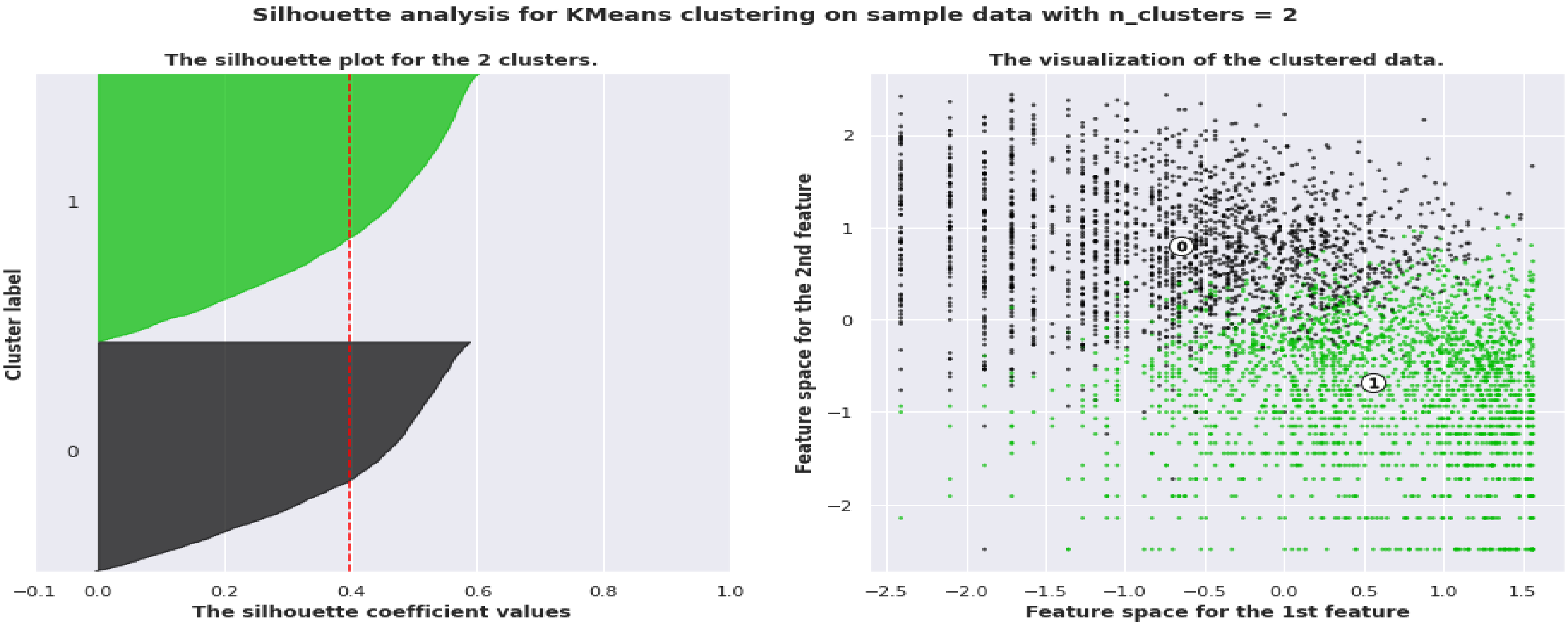
K-Means



The 3 clusters and their centroid are clearly defined but having some overlapping can be seen among these clusters.

To identify it clearly we have to apply Elbow method and Silhouette Score

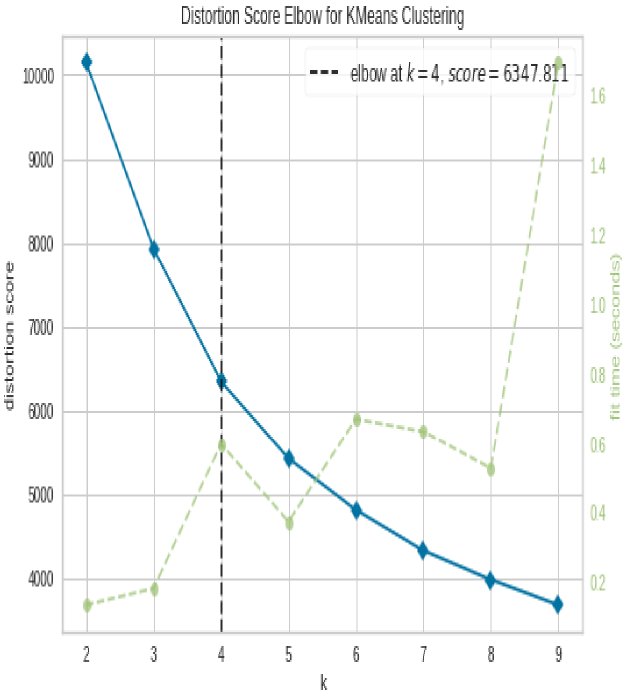
K-means Clustering with Silhouette

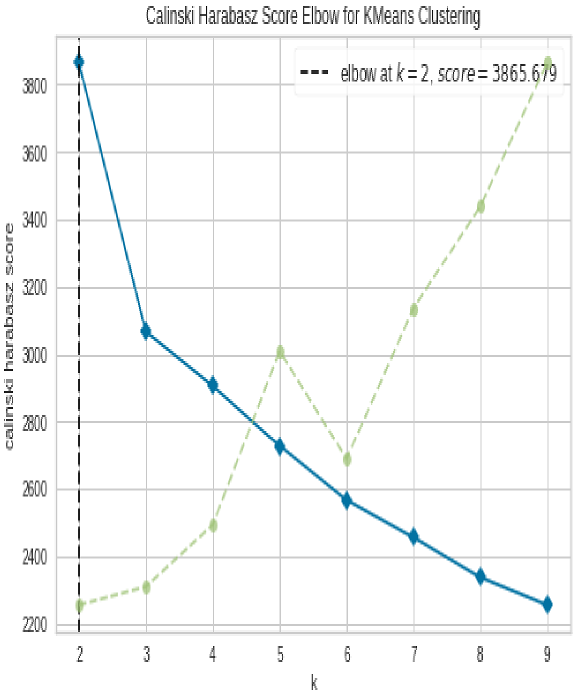


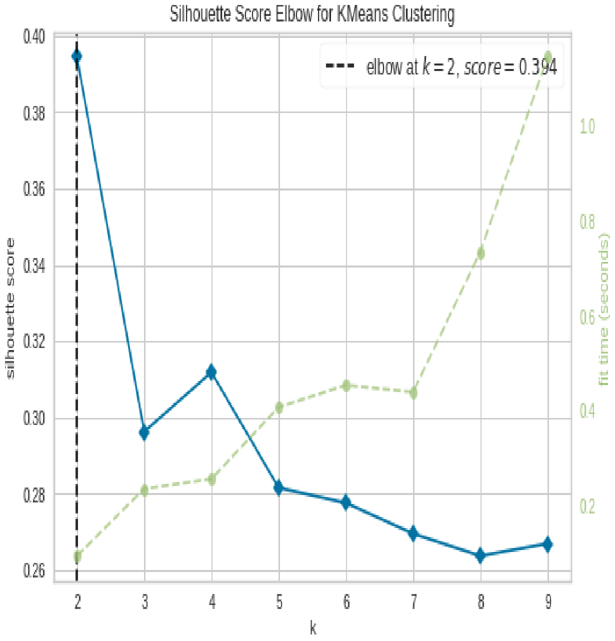
Cont.

- **Silhouette Score is around 0.40 having n_clusters = 2 which means neighboring clusters are away from each other, there are less chance for assigning the customers into wrong clusters.**

K-means with Elbow Method

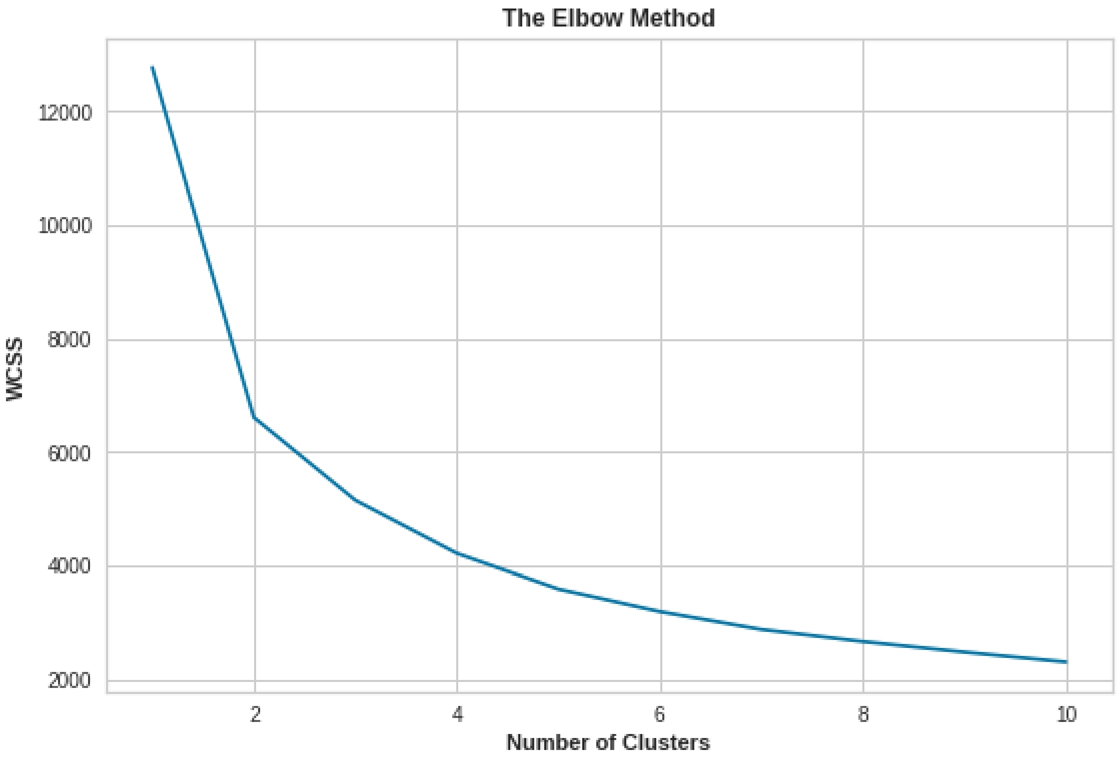






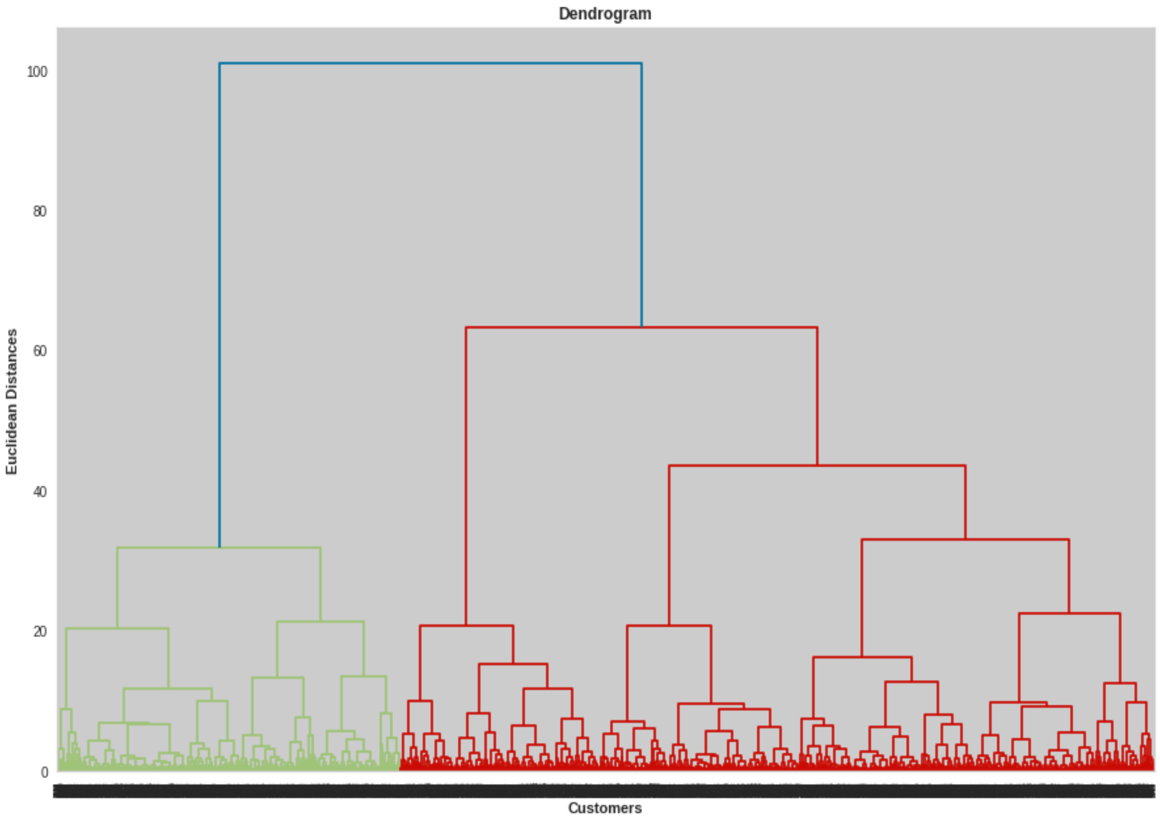
Elbow plot to identify better clusters

- In order to choose a better cluster we need to choose the number of cluster which has minimum WCSS.
- As it can be seen in above Elbow Method 4 seems to be the better cluster which has lower WCSS.
- If we go further then there is very slight downfall in WCSS so, 4 seems to be a good no of cluster.

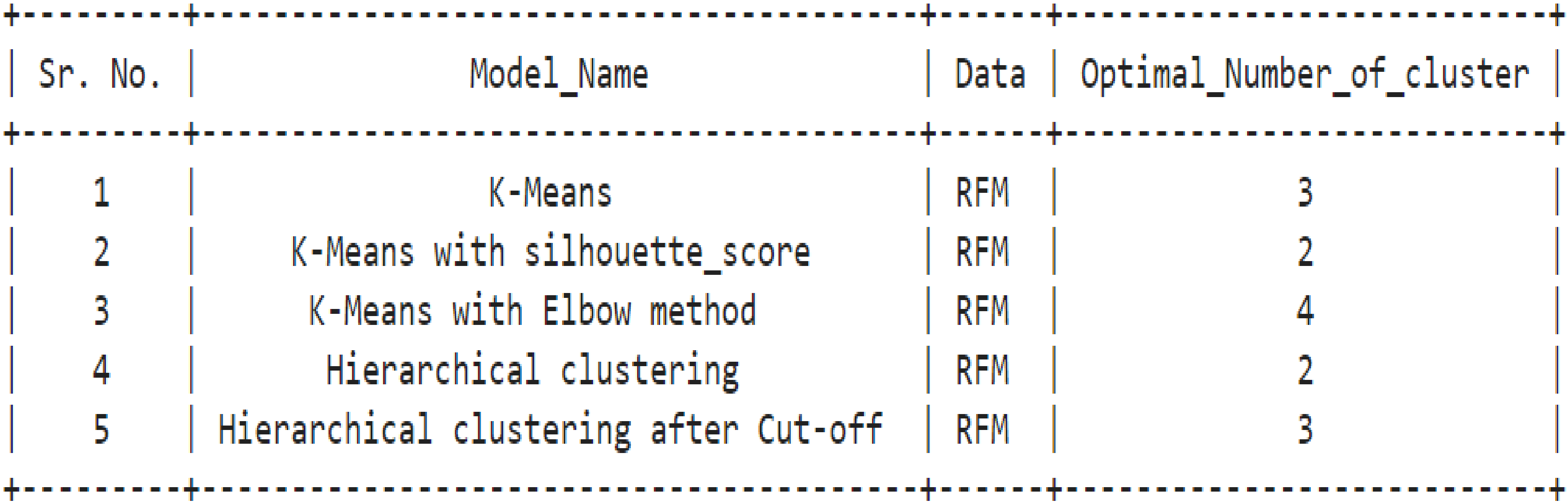


Dendrogram

- The more we climb the tree, the more the classes are grouped together and the less they are homogeneous (less intra-class inertia).
- The number of classes is a compromise between the similarity in the classes and the dissimilarity between the classes.



Pretty Table -The visual representation of the data in tabular forms.



```
+---------+---------------------------------------+------+---------------------------+
| Sr. No. |               Model_Name              | Data | Optimal_Number_of_cluster |
+---------+---------------------------------------+------+---------------------------+
|    1    |                K-Means                | RFM  |             3             |
|    2    |     K-Means with silhouette_score     | RFM  |             2             |
|    3    |       K-Means with Elbow method       | RFM  |             4             |
|    4    |        Hierarchical clustering        | RFM  |             2             |
|    5    | Hierarchical clustering after Cut-off | RFM  |             3             |
+---------+---------------------------------------+------+---------------------------+
```

Conclusion

Top Customer IDs: 17841.0, 14911.0, 14096.0, 12748.0, 14606.0

Top Five Countries: Uniter Kingdom(**88.95%**), Germany(**2.33%**), France(**1.84%**) and Ireland(**1.84%**)

Month which give maximum business: November, October, December, September and May.

Maximum purchasing on different days: Thursday > Wednesday > Tuesday > Monday > Saturday

Most of the customers usually purchase products in between 10:00 A.M to 2:00 P.M.

Top Four products purchasing based on frequency:

1) WHITE HANGING HEART T-LIGHT HOLDER
2) REGENCY CAKESTAND 3 TIER
3) JUMBO BAG RED RETROSPOT
4) ASSORTED COLOUR BIRD ORNAMENT

Conclusion Contd..

- RFM(Recency, Frequency and Monetary) data frame ease our problem to solve in a particular order, it makes easy to recommend and display new launched products to few customers.

- Applied different clustering algorithms:

  - **1) K-Means** = Optimal Clusters(**3**)
  - **2) K-Means with Silhouette** = Optimal_Clusters: (**2**)
  - **3) K-Means with Elbow Method** = Optimal_Clusters: (**4**)
  - **4) Hierarchical Clustering** = Optimal_Clusters: (**2**)
  - **5) Hierarchical Clustering with cut-off** = Optimal_Cluster: (**3**)

THANK YOU!!!

- Q & A