Capstone Project-4

Online Retail Customer Segmentation

ABHISHEK SHUBHAM

Contents

Business Understanding

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers(and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators.

Given the dataset, the objective is to build a clustering model that would perform customer Segmentation.

Data Description

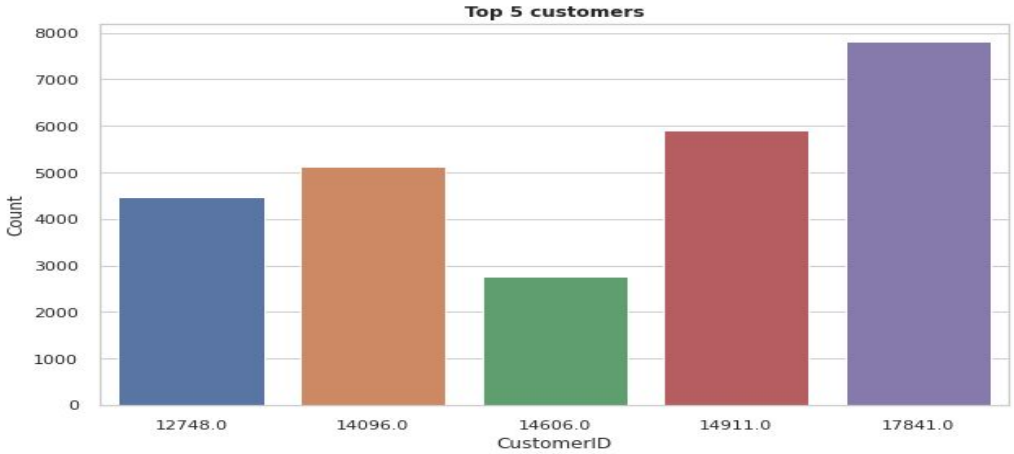
We have been provided with a UK-based and registered online retail company dataset which contains transactions between 01/12/2010 and 09/12/2011 with 541909 instances and 8 features.

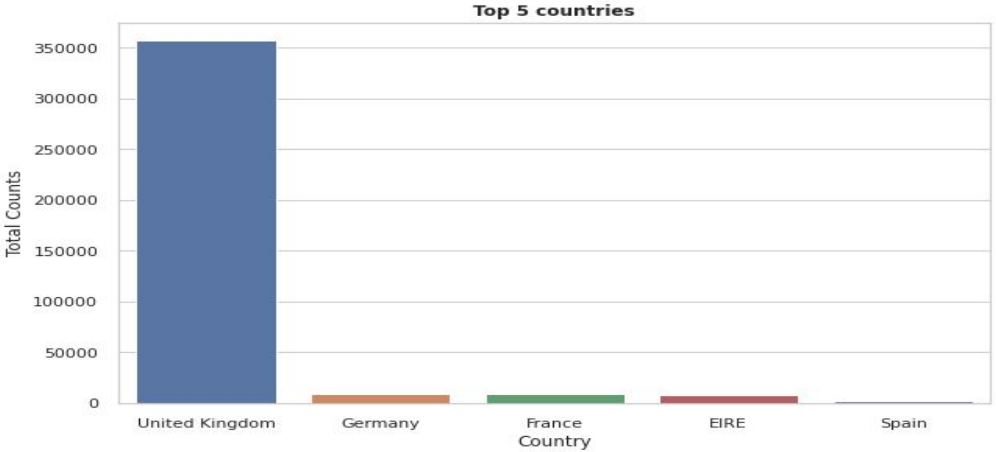
- **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- **Description:** 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 Cleaning

- The dataset has 541909 observations and 8 features/columns.
- Our dataset possess high null values in **CustomerID** and **Description** column. Around 25% missing values are there in **CustomerID** and 0.26% null values present in **Description** column.
- Also the dataset contains a total of 5225 observations which are duplicate.
- We have to drop all null values and duplicate observations because each CustomerID are uniquely assigned to a particular customer which means that we cannot impute it with any other values from the dataset.
- After dealing with null and duplicate values we were left with 401604 instances/rows and 8 features in our dataset.

Exploratory Data Analysis

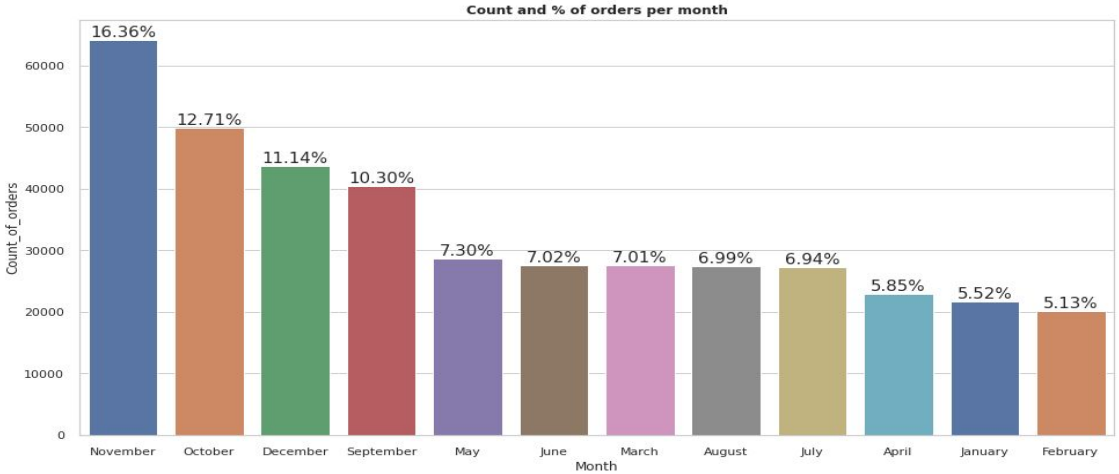


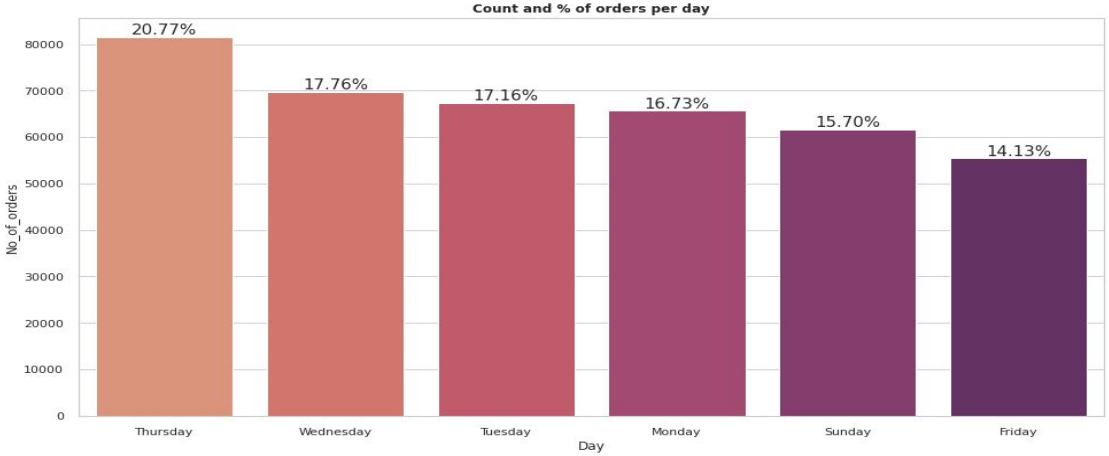


- These are the top 5 customer's ID whose frequency of placing order is very high in comparison to others.
- There are 4339 unique customer for the business.

Majority of the customers belong to the United Kingdom, Germany, France, Ireland and Spain. Since the data belongs to a UK based company, hence 90% of the orders are placed from UK alone.

Days and Months



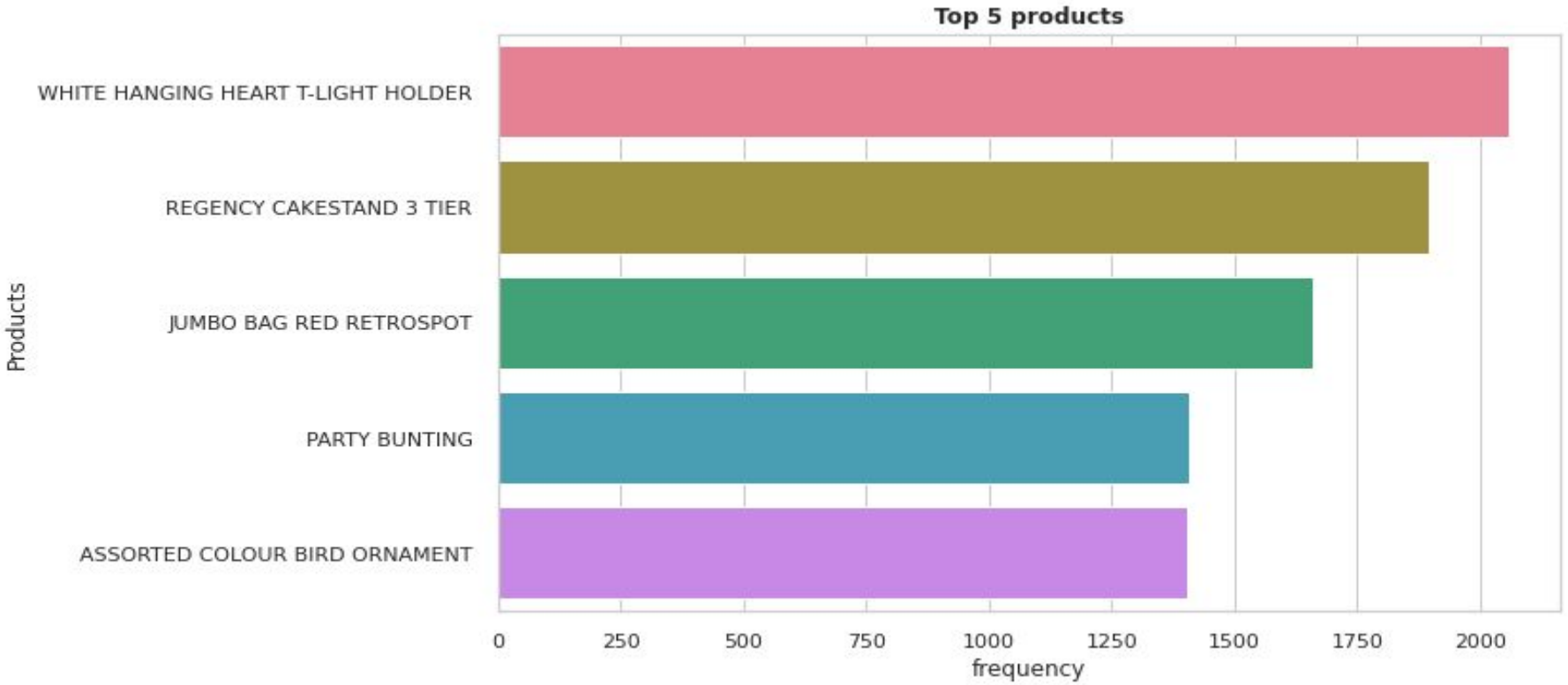


Purchases made by customers are high in the month of Nov, Oct, Dec, Sept and May.

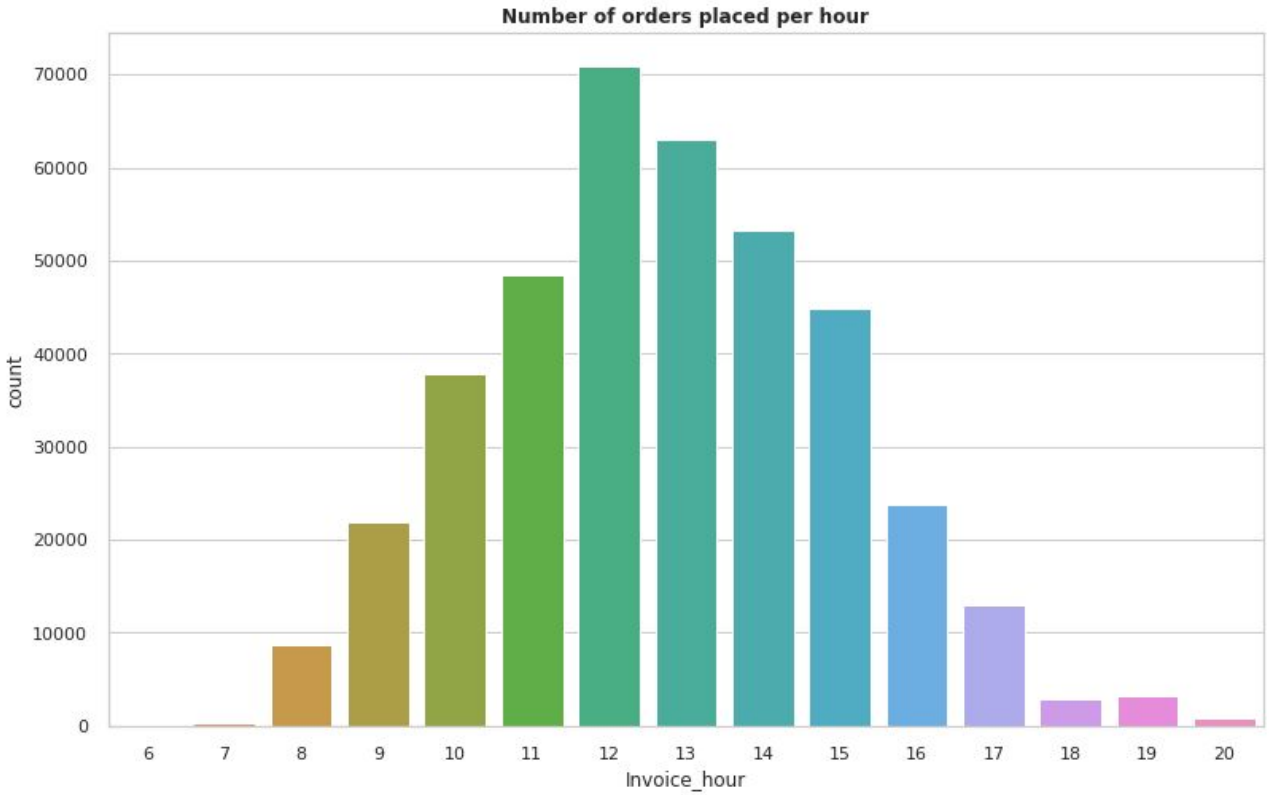
Around 55% of orders are placed in these 5 months. It may be due to festivals or offers during these months.

- It can be seen that **Thursday** is most popular day for our customers to place the orders, followed by **Wednesday**, **Tuesday** etc.
- Also it is noticeable that no orders are placed on **Saturday**. It's possible that the business takes a off on this day of the week.

Top 5 Products

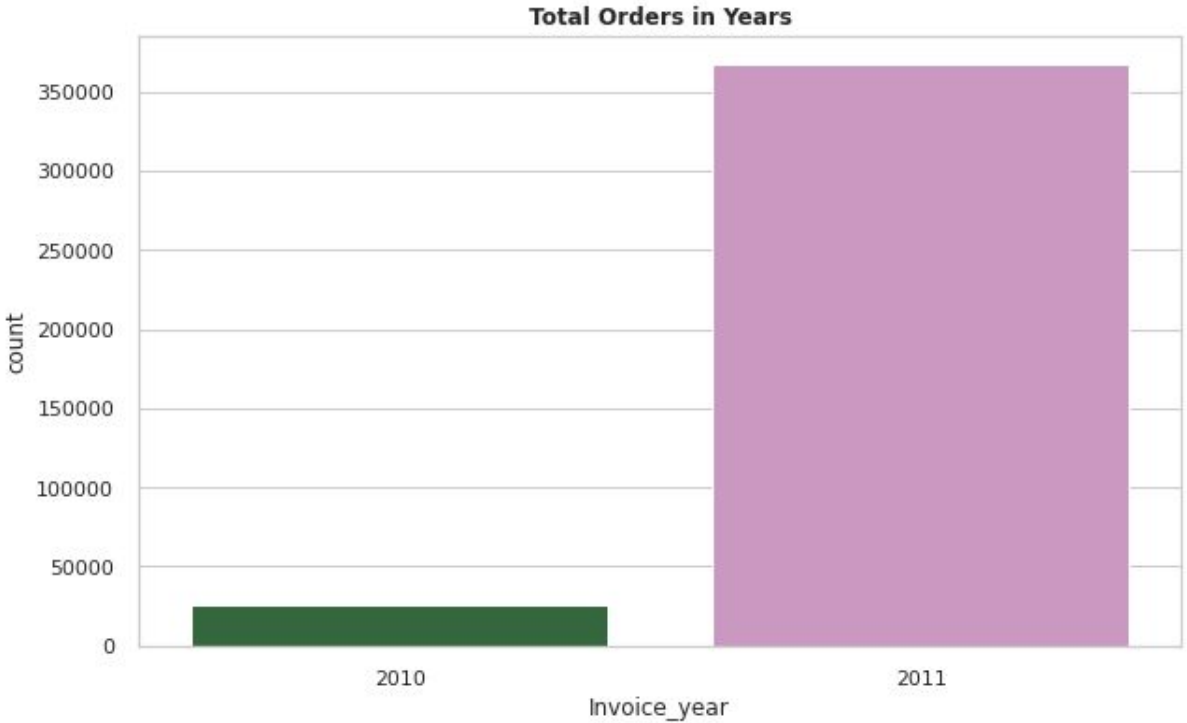


Hourly Order Spikes:



- From this countplot it can be seen that customers usually purchase in between 10:00 AM to 3:00 PM.
- There are very few purchases early morning and at midnight.

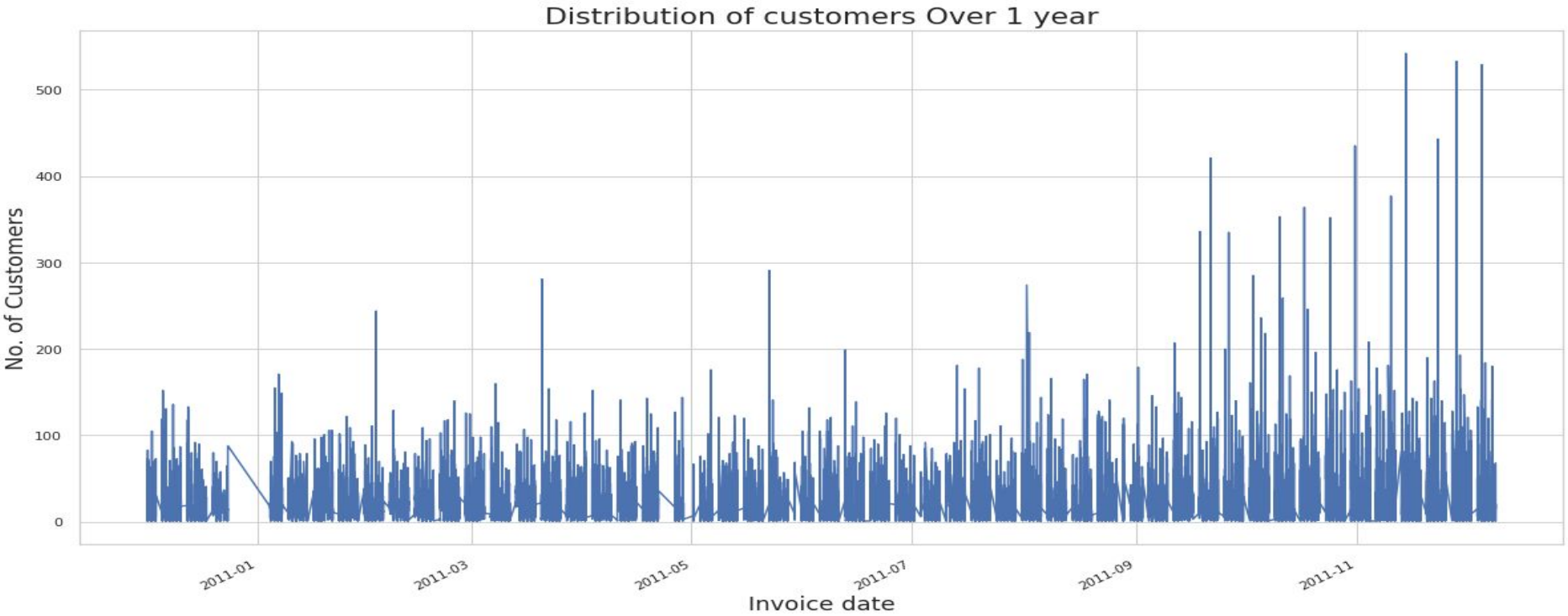
Yearly Purchases



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

Around 3000 orders were placed in just one month.

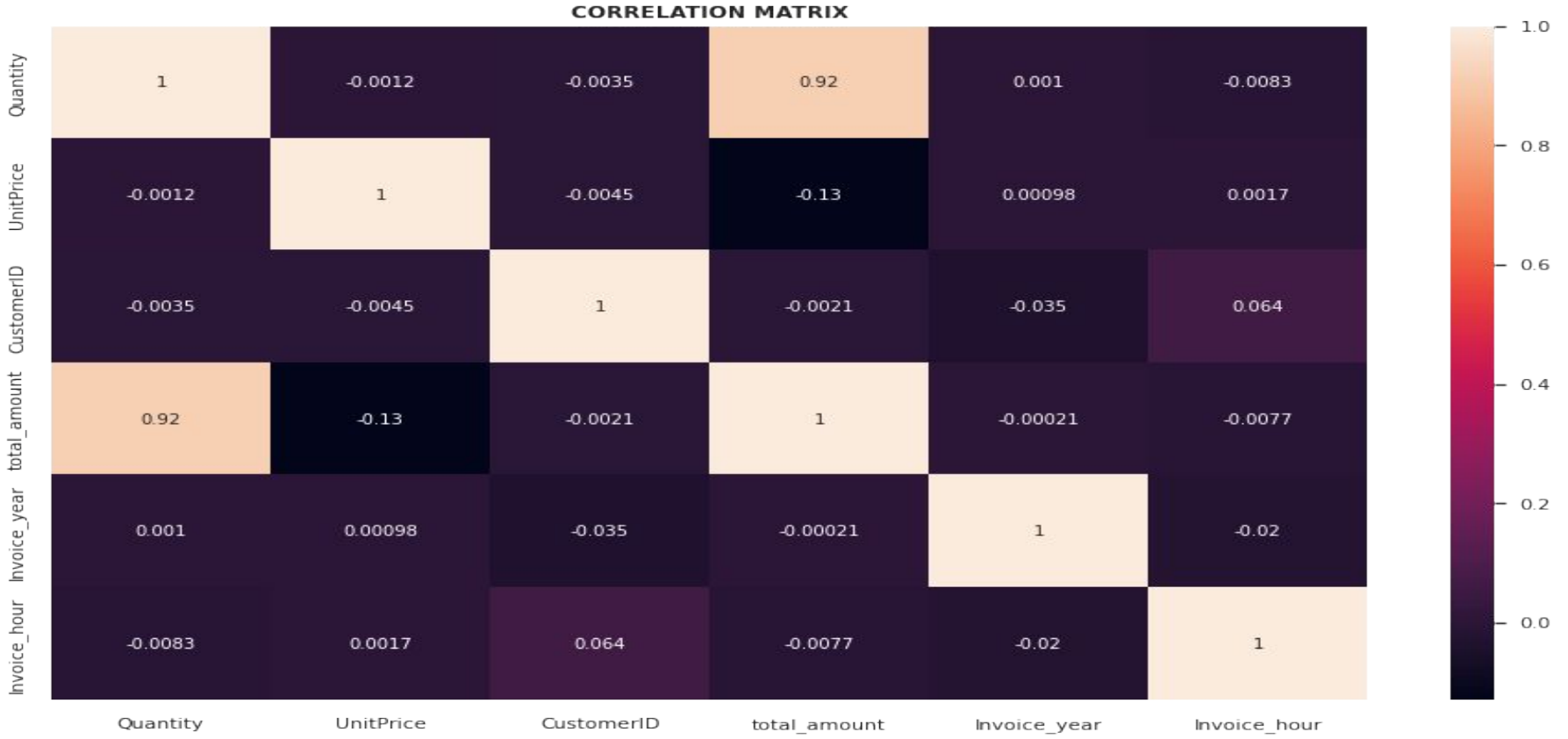
Continued….



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

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

Correlation Matrix



Data Transformation

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

R(Recency): Number of days since last purchase.
F(Frequency): How frequent customers are(Number of times they have purchased)
M(Monetary): Total amount of transactions(revenue contributed)

The RFM DataFrame is grouped on the basis of customer ID.
The data contains 4339 rows or customers.

There are some outliers in the RFM table, to cure this problem we've applied the IQR(InterQuartile Range) method.

IQR= Q3- Q1

Q3: Third quartile

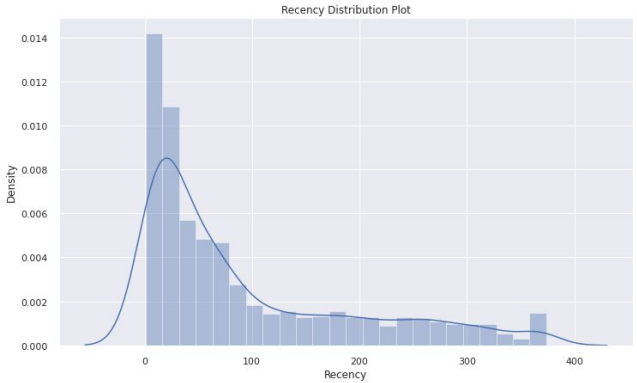
Q1: First quartile

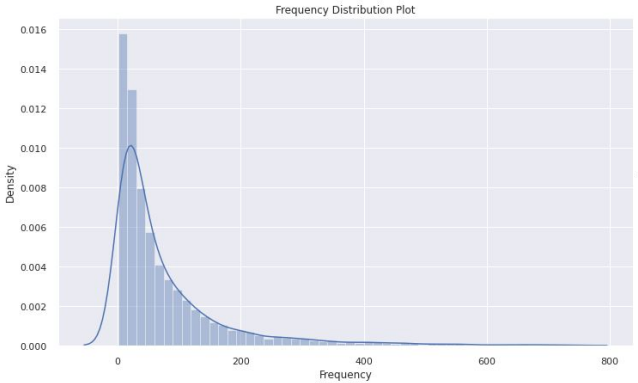
	CustomerID	Recency	Frequency	MonetaryValue
0	12346.0	326	1	77183.60
1	12347.0	2	182	4310.00
2	12348.0	75	31	1797.24
3	12349.0	19	73	1757.55
4	12350.0	310	17	334.40
...	...	...	...	...
4334	18280.0	278	10	180.60
4335	18281.0	181	7	80.82
4336	18282.0	8	12	178.05
4337	18283.0	4	721	2045.53
4338	18287.0	43	70	1837.28

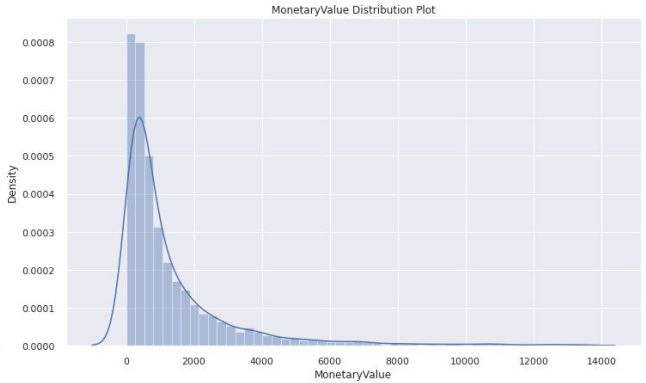
4339 rows × 4 columns

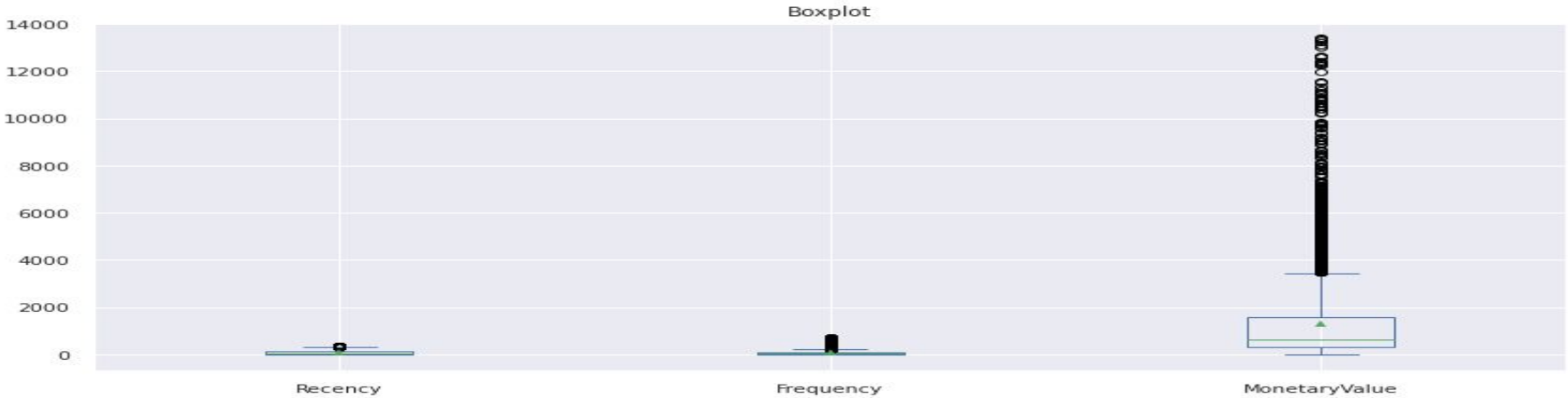
Continued…

After applying IQR on DataFrame, most of the outliers are removed from the dataset.

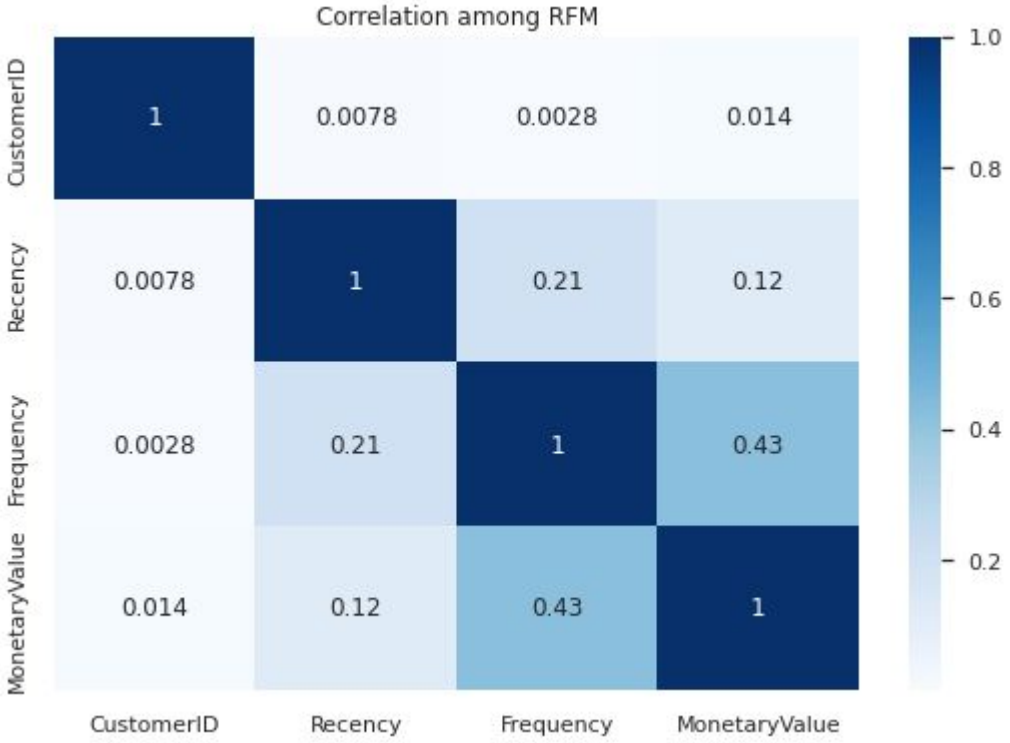








Correlation among RFM



In the Correlation Matrix of RFM:

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

Model Building(Clustering)

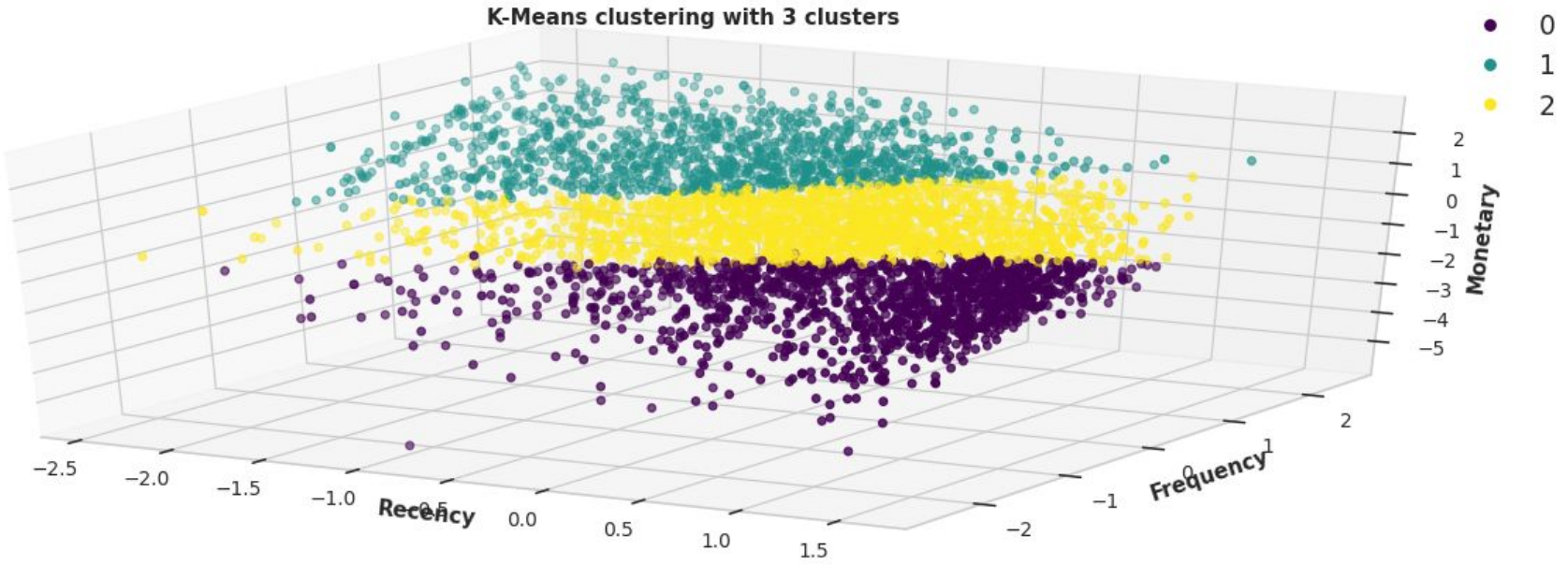
- In this section, we have used the K-Means algorithm to cluster the customers into different segments.

- To identify the optimum numbers of clusters, we have used the elbow method and silhouette analysis.

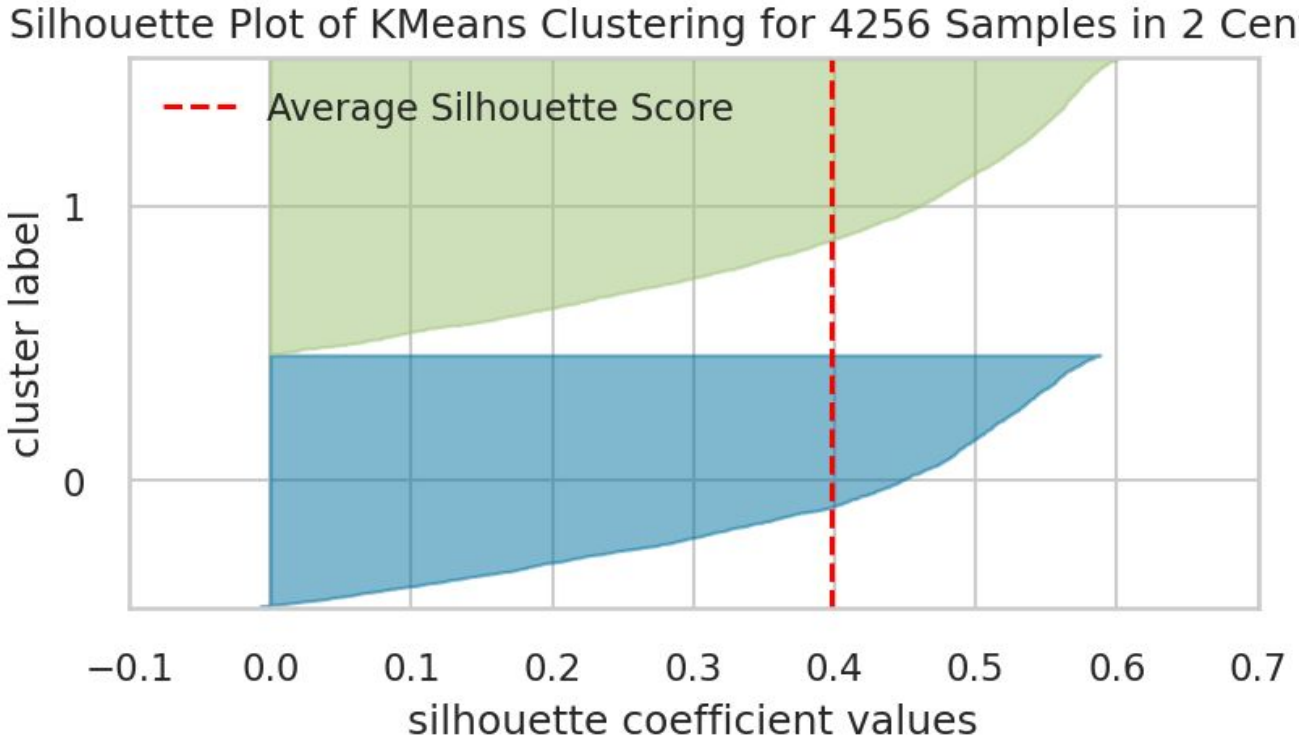
- With both the methods, 3 clusters is optimum in this case.

- Also we used Hierarchical Clustering and plotted a Dendrogram by setting a threshold limit(50) to find the optimal clusters in this case..

K-Means Clustering

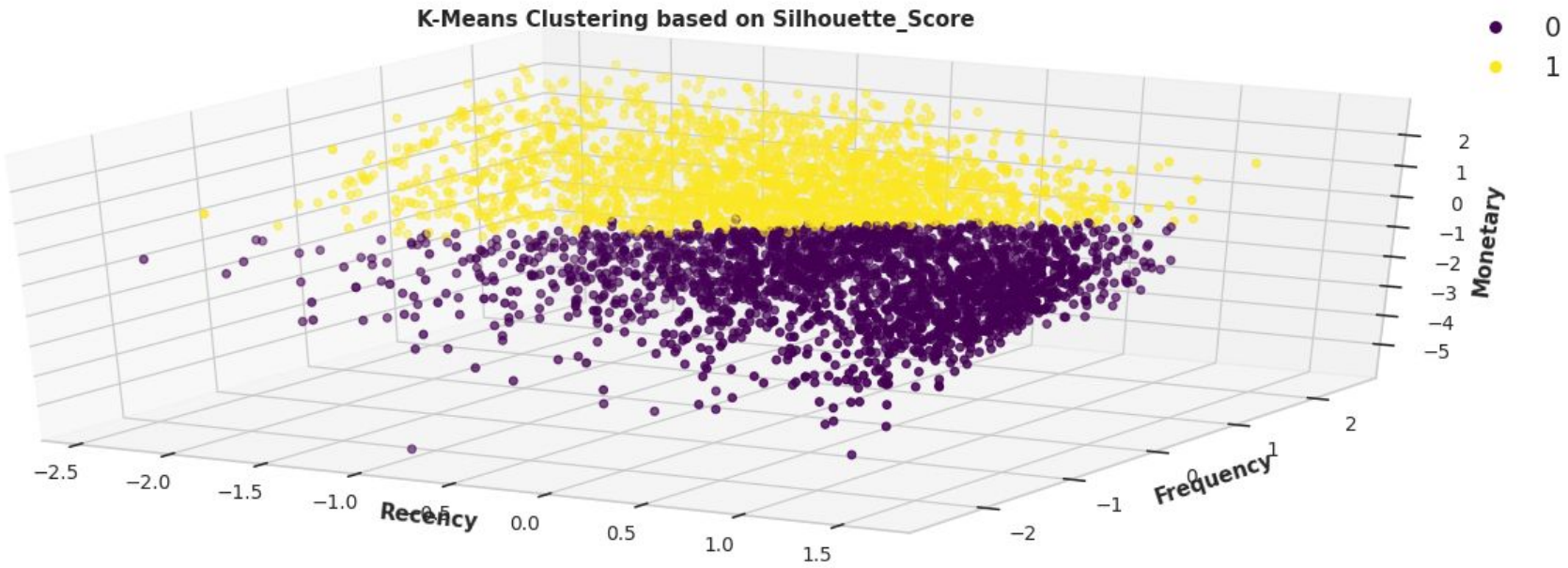


K-Means Clustering with Silhouette

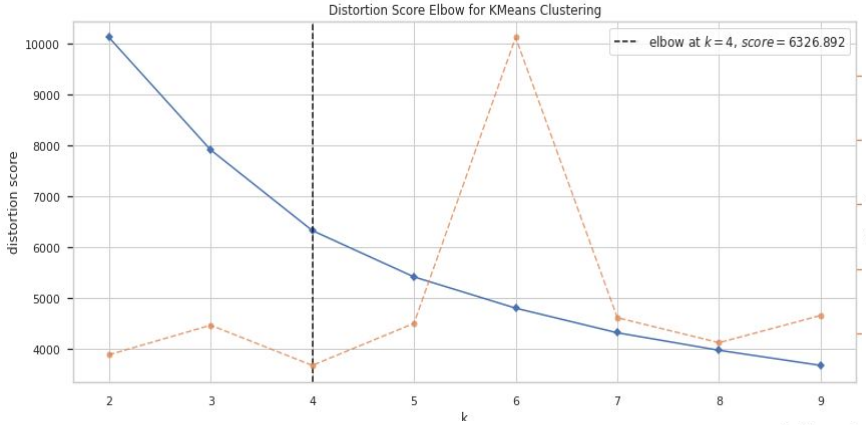


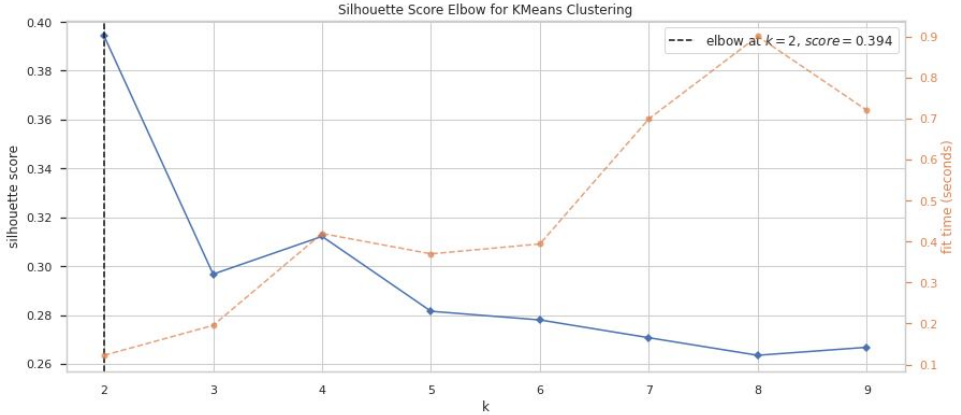
Silhouette Score is around 0.40 having n_clusters = 2 which means neighbouring clusters are away from each other, there are less chance for assigning the customers into wrong clusters.

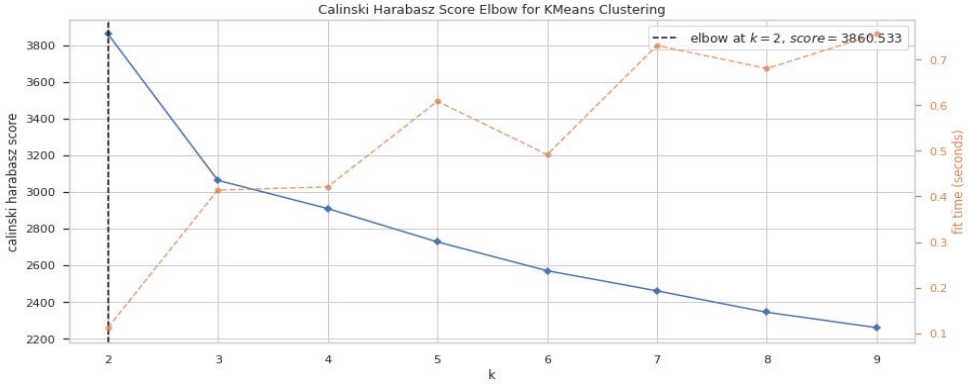
Continued…


K-Means Clustering based on Silhouette_Score
0
1
Recency
Frequency
Monetary
−2.5
−2.0
−1.5
−1.0
0.0
0.5
1.0
1.5
−2
−1
0
1
2
−5
−4
−3
−2
−1
0
1
2

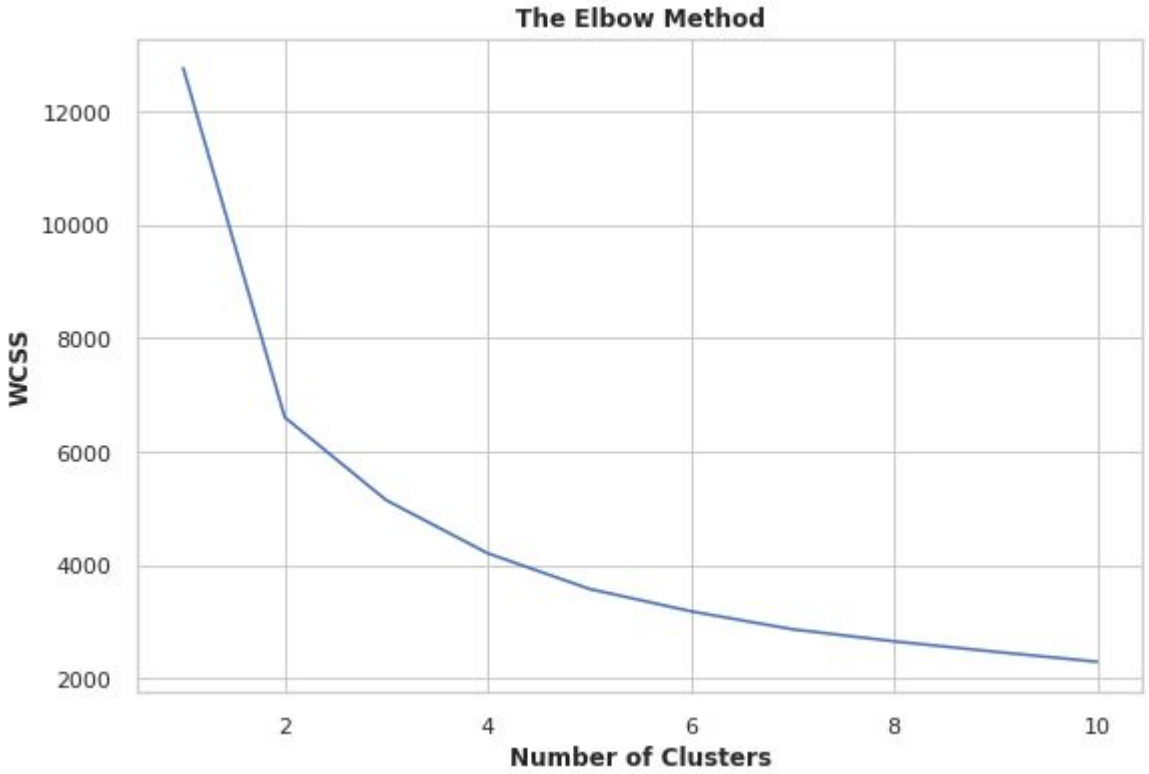
K-Means with Elbow Method







Elbow plot to identify better clusters

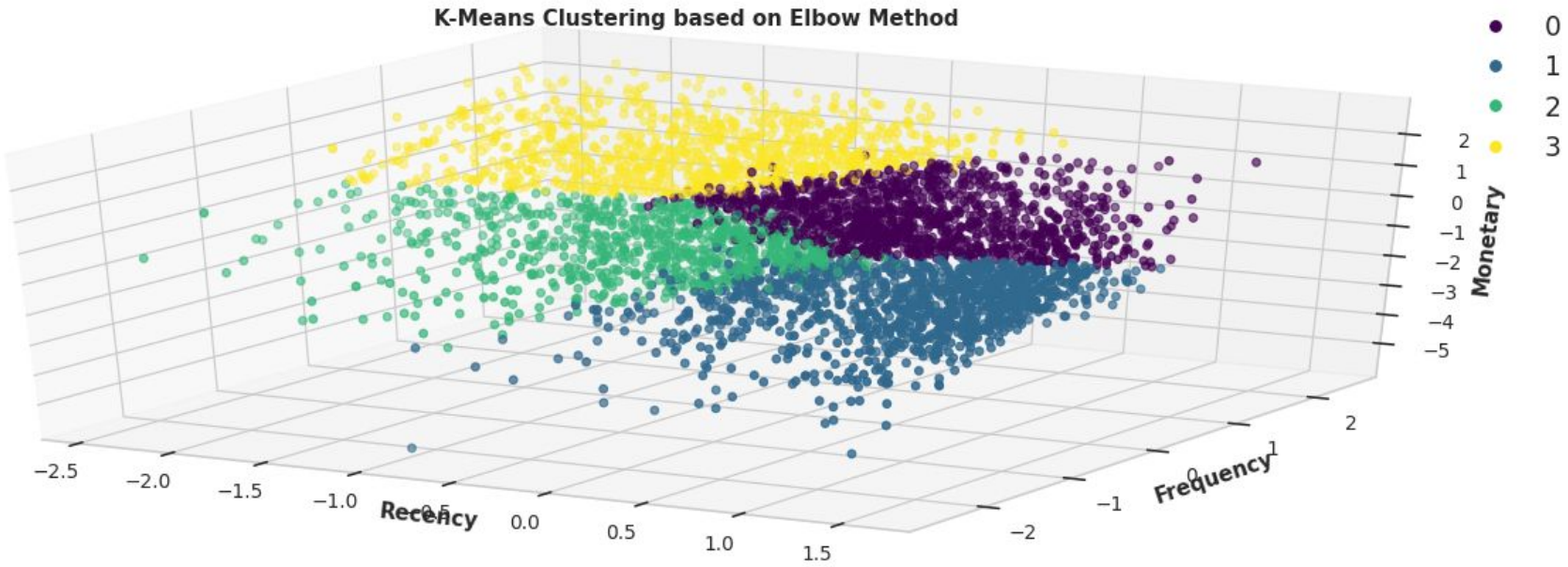


In order to choose a better cluster we need to choose the number of clusters which have minimum WCSS.

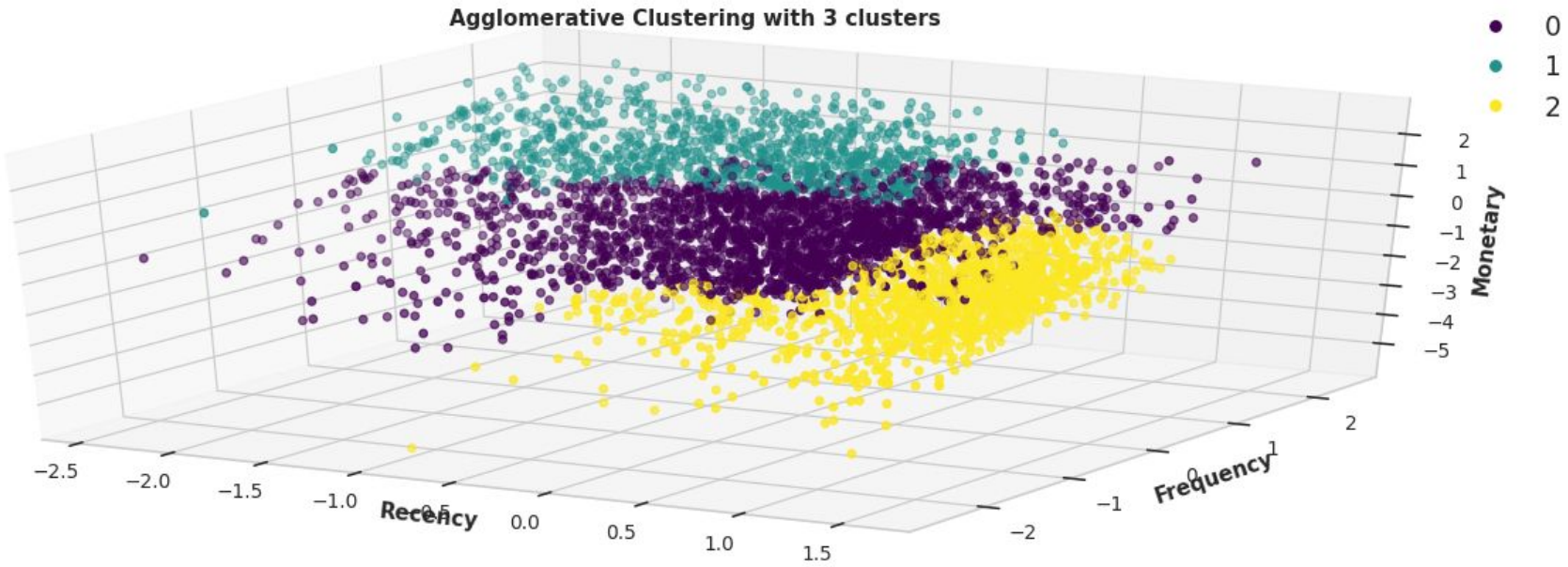
As it can be seen in above Elbow Method 4 seems to be the better cluster which has lower WCSS.

If we go further then there is very slight downfall in WCSS so, 4 seems to be a good no of clusters.

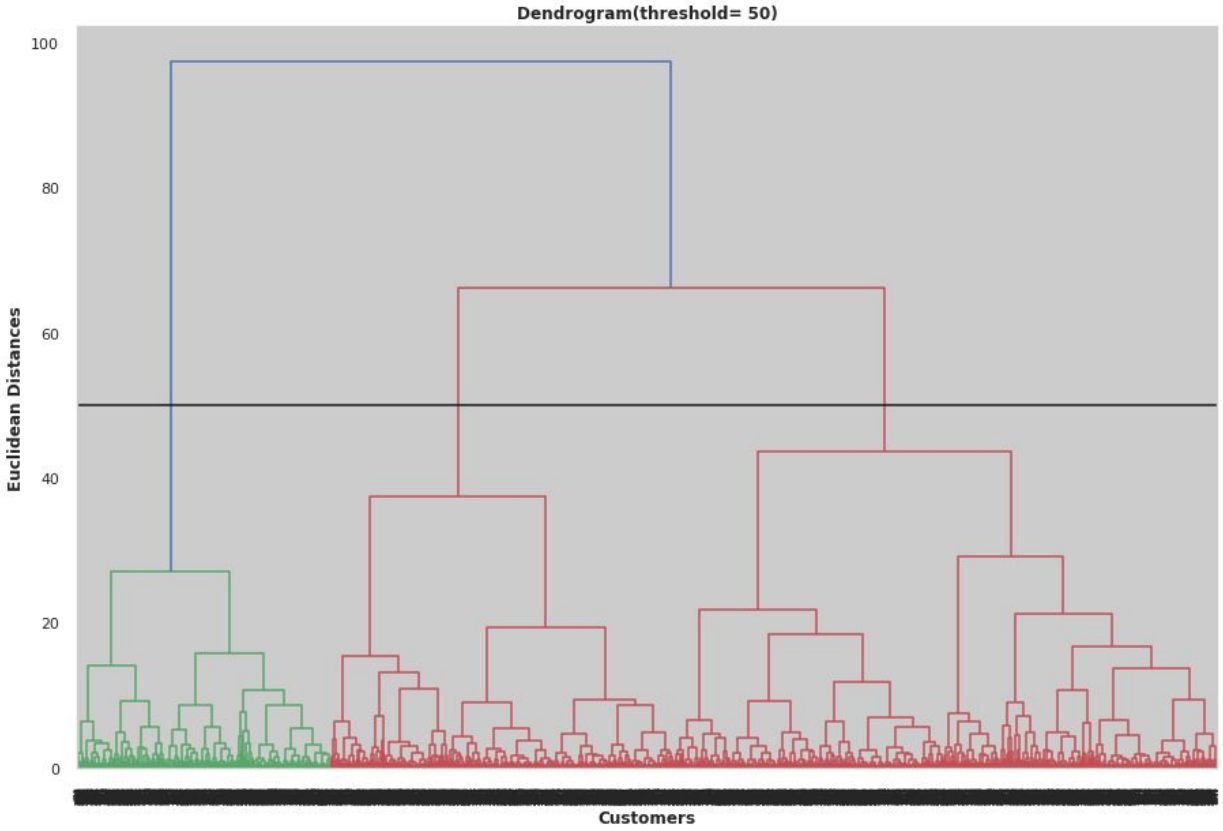
Continued…


K-Means Clustering based on Elbow Method
0
1
2
3
Monetary
2
1
0
−1
−2
−3
−4
−5
Frequency
2
1
0
−1
−2
Recency
−2.5
−2.0
−1.5
−1.0
0.0
0.5
1.0
1.5

Implementing Hierarchical Clustering



Dendrogram



A dendrogram is a branching diagram that represents **the relationships of similarity among a group of entities**.

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

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

Analysis Summary

- **Top Customer IDs:** 17841.0, 14911.0, 14096.0, 12748.0, 14606.0
- **Which Year produced maximum business:** 2011
- **Maximum purchasing on different days:** Thursday > Wednesday > Tuesday > Monday > Sunday > Friday. No Purchases were made on Saturday.
- **Top Five Countries:** United Kingdom(88.83%), Germany(2.36%), France(2.11%), Ireland(1.86%) and Spain(0.63%).
- **Month which give maximum business:** November, October, December, September and May.
- Most of the customers usually purchase products in between 10:00 A.M to 3:00 P.M.

Top Five products purchasing on the basis of frequency(How frequent customers are willing to bought these products):

1. WHITE HANGING HEART T-LIGHT HOLDER
2. REGENCY CAKE STAND 3 TIER
3. JUMBO BAG RED RETROSPOT
4. PARTY BUNTING
5. ASSORTED COLOUR BIRD ORNAMENT

Conclusion

- RFM(Recency, Frequency and Monetary) dataframe 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:
  - **K-Means** = Optimal Clusters(**3**)
  - **K-Means with Silhouette** = Optimal_Clusters: (**2**)
  - **K-Means with Elbow Method** = Optimal_Clusters: (**4**)
  - **Hierarchical Clustering** = Optimal_Clusters: (**2**)
  - **Hierarchical Clustering with cut-off** = Optimal_Cluster: (**3**)

THANK YOU