Capstone Project

Online Retail Customer Segmentation

by
Lovejeet Singh

Contents

Problem Statement

- To identify major customer segments on a transnational data set.
- Data set contains all the transactions occurring between 1st December 2010 and 9th December 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

In this Dataset we have 541909 number of observations and 8 features

Main 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 Wrangling

Null Values

In our dataset we have lots of null values.

```
# Let's check the null values count.
retail_df.isnull().sum().sort_values(ascending=False)
```

```
CustomerID     135080
Description      1454
InvoiceNo           0
StockCode           0
Quantity            0
InvoiceDate         0
UnitPrice           0
Country             0
dtype: int64
```

Null Value Treatment

```
[9] #droping null values
    retail_df.dropna(inplace=True)
```

```
retail_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 406829 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype         
---  ------       --------------   -----         
 0   InvoiceNo    406829 non-null  object        
 1   StockCode    406829 non-null  object        
 2   Description  406829 non-null  object        
 3   Quantity     406829 non-null  int64         
 4   InvoiceDate  406829 non-null  datetime64[ns]
 5   UnitPrice    406829 non-null  float64       
 6   CustomerID   406829 non-null  float64       
 7   Country      406829 non-null  object        
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 27.9+ MB
```

After dropping Null values we have 406829 number of observation and 8 features.

Contd...

```
retail_df.describe()
```

	Quantity	UnitPrice	CustomerID
count	406829.000000	406829.000000	406829.000000
mean	12.061303	3.460471	15287.690570
std	248.693370	69.315162	1713.600303
min	-80995.000000	0.000000	12346.000000
25%	2.000000	1.250000	13953.000000
50%	5.000000	1.950000	15152.000000
75%	12.000000	3.750000	16791.000000
max	80995.000000	38970.000000	18287.000000

- the Quantity feature shows negative value which is not possible.
- UnitPrice has 0 as min value.
- We need to explore these column.

Contd...

```
# dataframe have negative valiues in quantity.
retail_df[retail_df['Quantity']<0]
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
...	...	...	...	...	...	...	...	...
540449	C581490	23144	ZINC T-LIGHT HOLDER STARS SMALL	-11	2011-12-09 09:57:00	0.83	14397.0	United Kingdom
541541	C581499	M	Manual	-1	2011-12-09 10:28:00	224.69	15498.0	United Kingdom
541715	C581568	21258	VICTORIAN SEWING BOX LARGE	-5	2011-12-09 11:57:00	10.95	15311.0	United Kingdom
541716	C581569	84978	HANGING HEART JAR T-LIGHT HOLDER	-1	2011-12-09 11:58:00	1.25	17315.0	United Kingdom
541717	C581569	20979	36 PENCILS TUBE RED RETROSPOT	-5	2011-12-09 11:58:00	1.25	17315.0	United Kingdom

8905 rows × 8 columns

- After exploring these feature we found that some InvoiceNo has ‘C’ which means cancellation order.
- We also have to remove those obvervations which shows unitprice as 0.
- So after removing these observation we have 397884 observation left and 8 features.

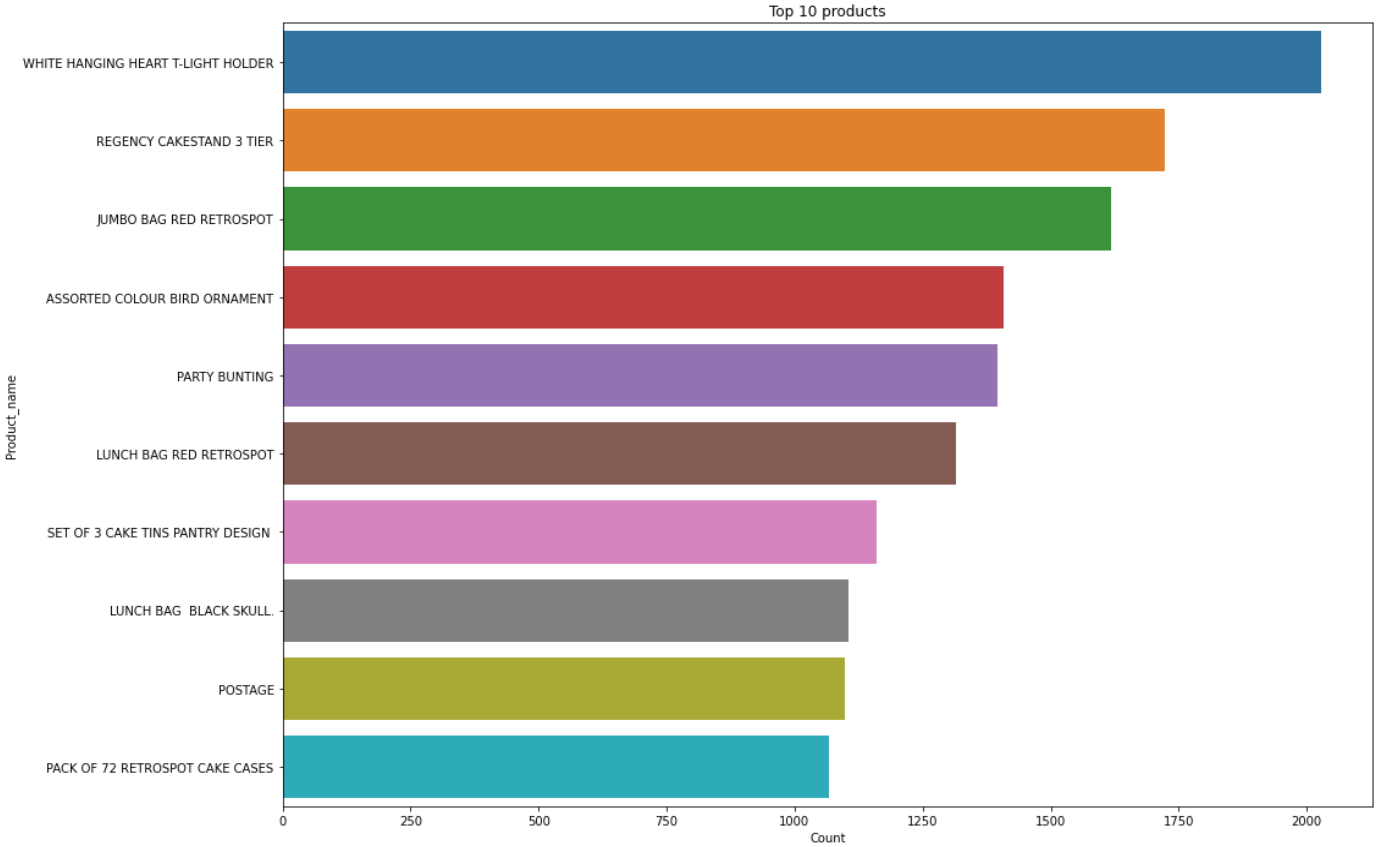
Feature Engineering

Creating New Feature

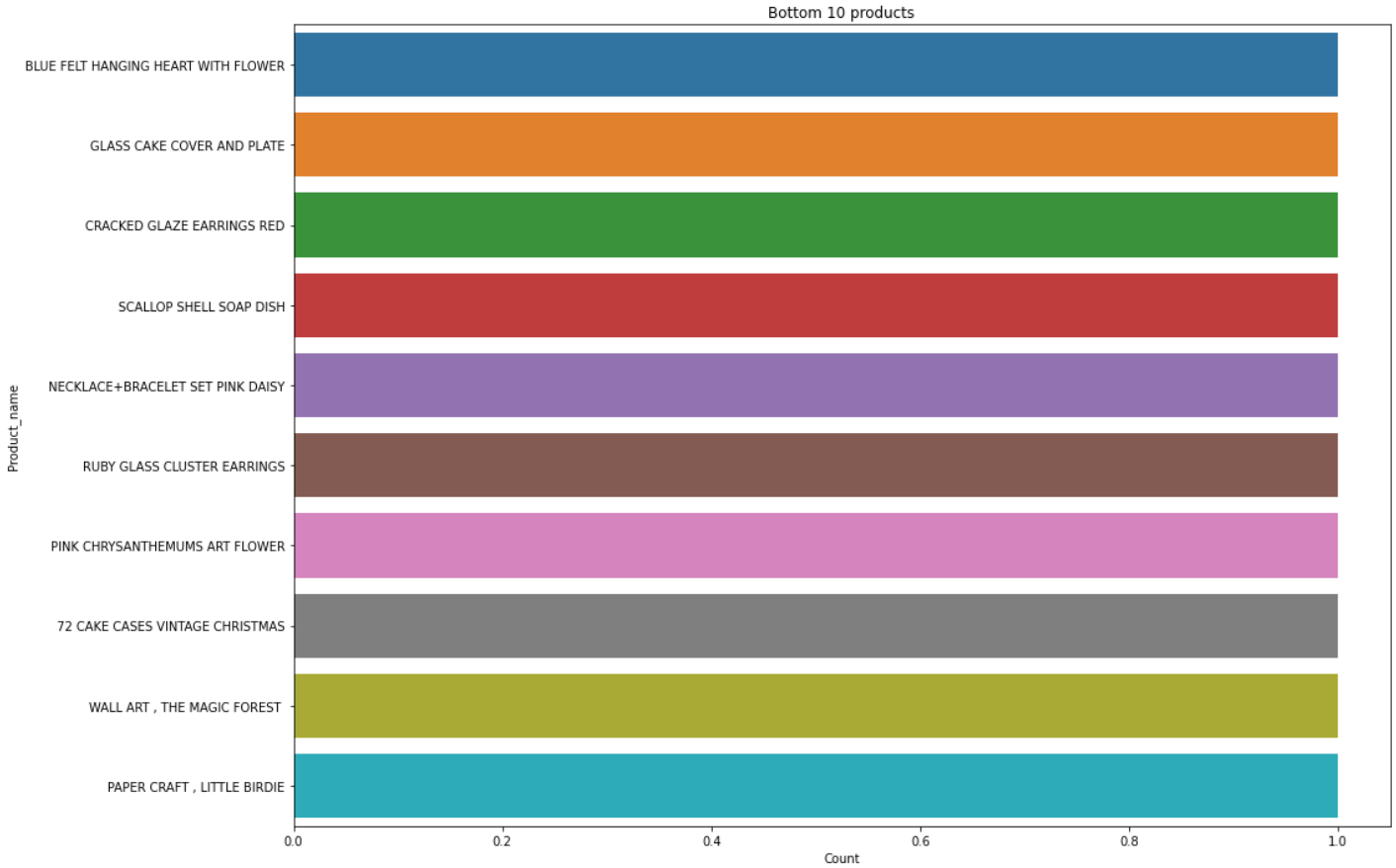
- We can make a new feature by our Datetime Feature which is InvoiceDate to extract Day, Month, Year, Hour and Minute.
- We also create a new feature called TotalAmount by multiplying Unitprice and Quantity.
- After extract Day, Month, Year, Hour and Minute we can make another new feature called Day_time_type, which tells us the three label Morning, Afternoon and Evening according to the hourly time frame.

Exploratory Data Analysis

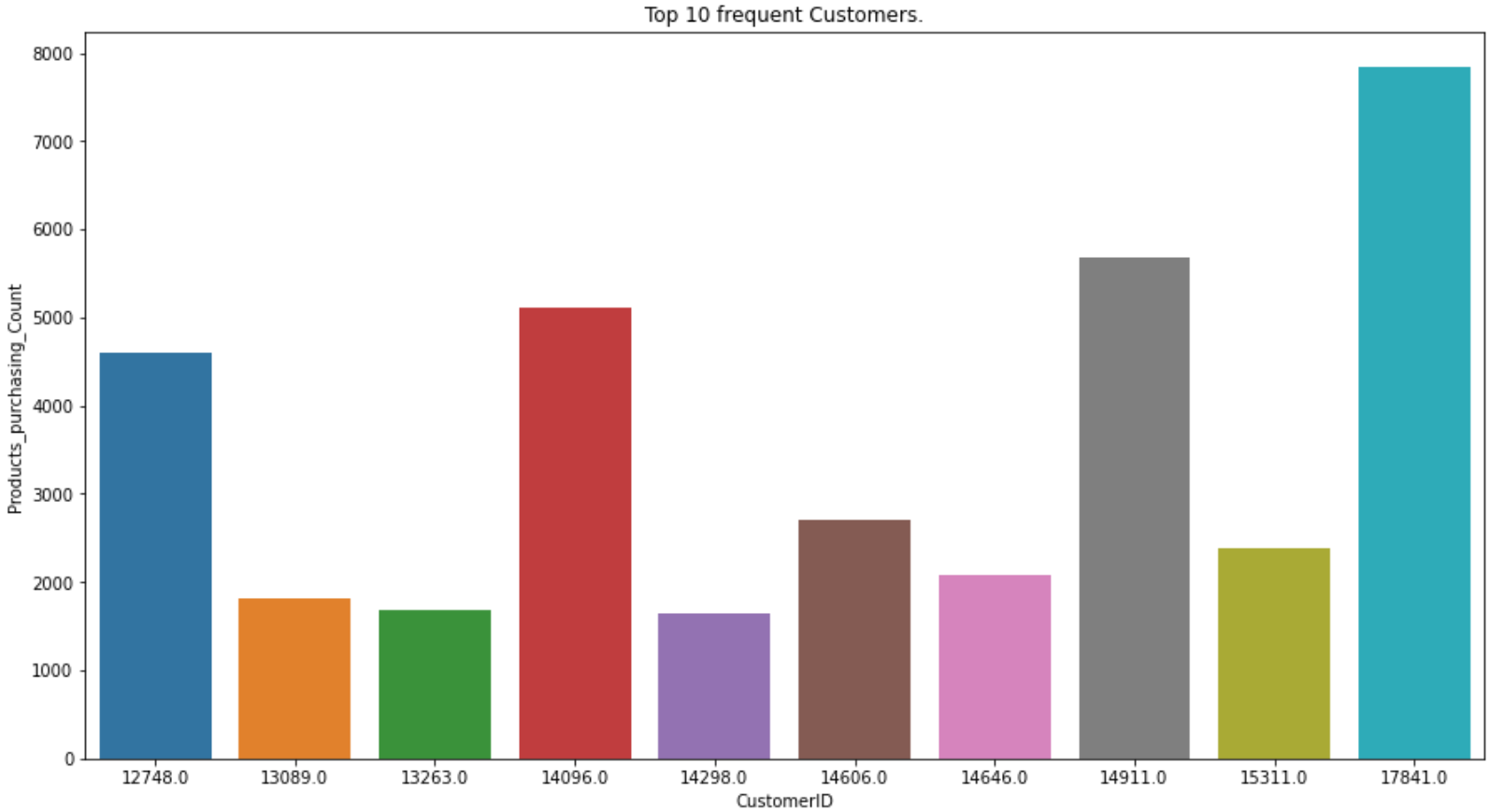
Checking Top 10 Product



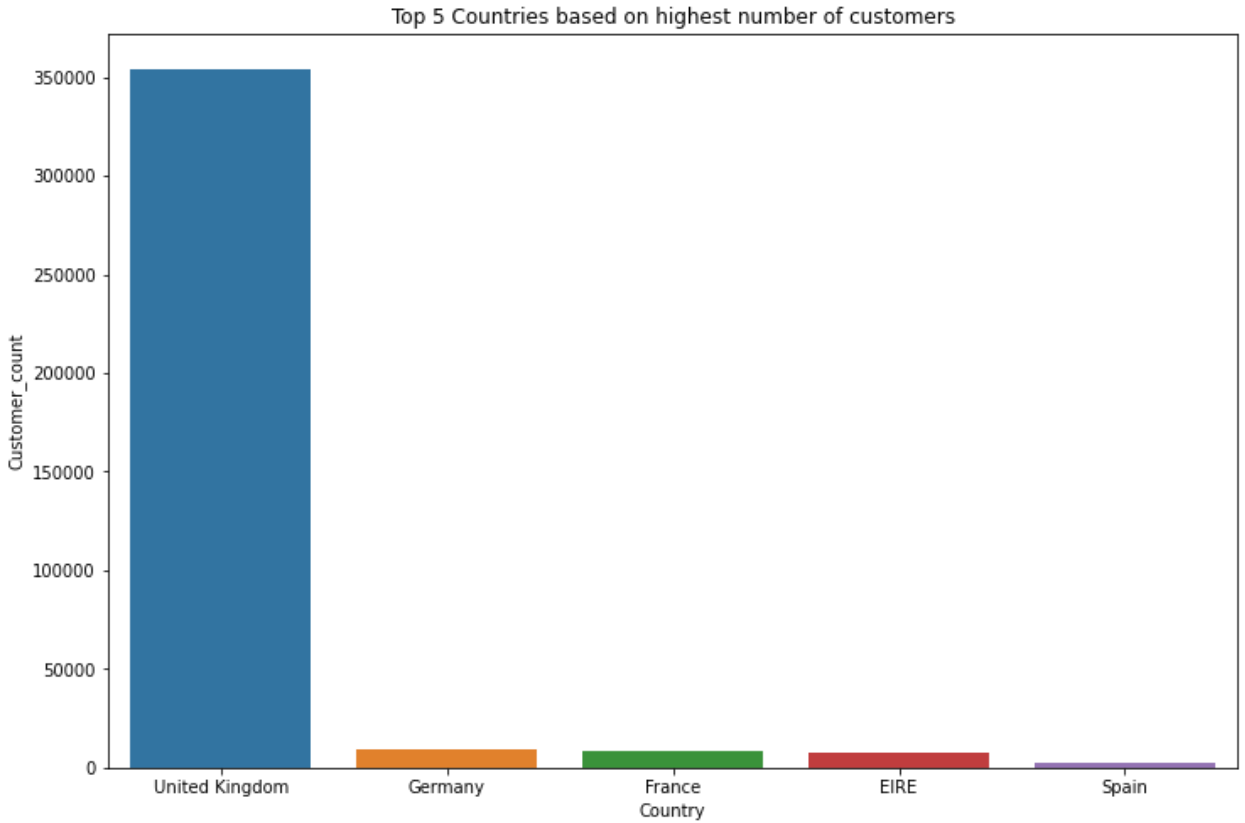
Top 10 Rarely Sold Product



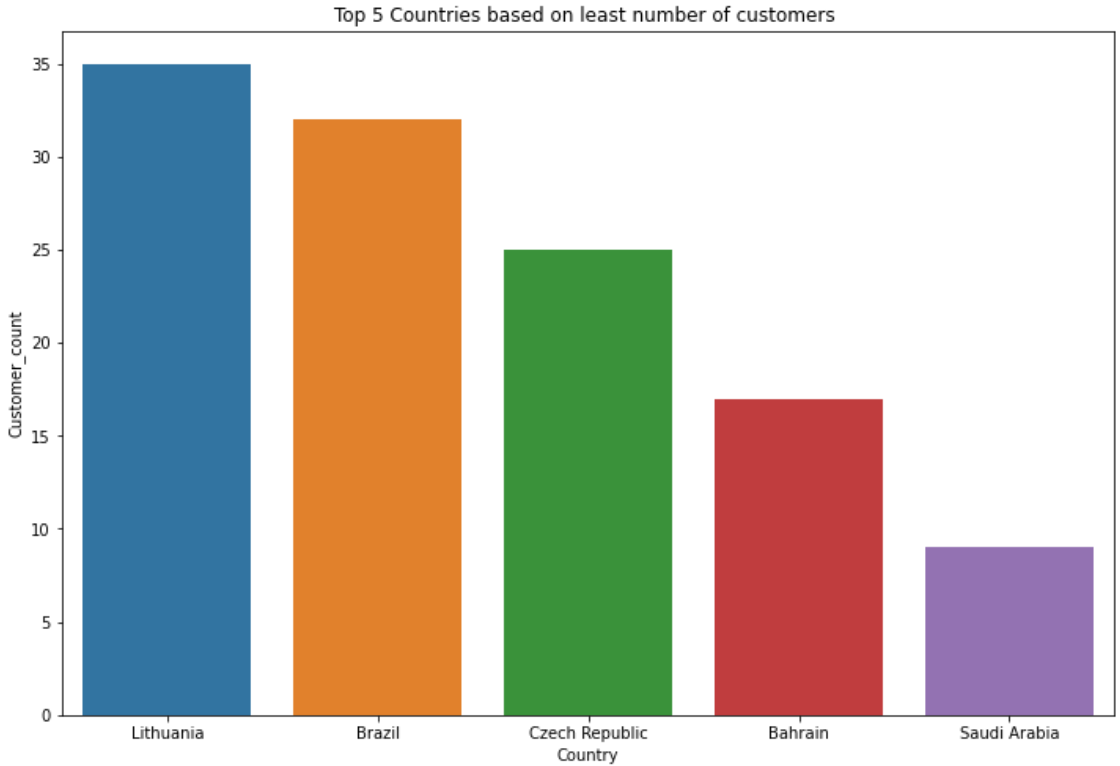
Top 10 Frequent Customers



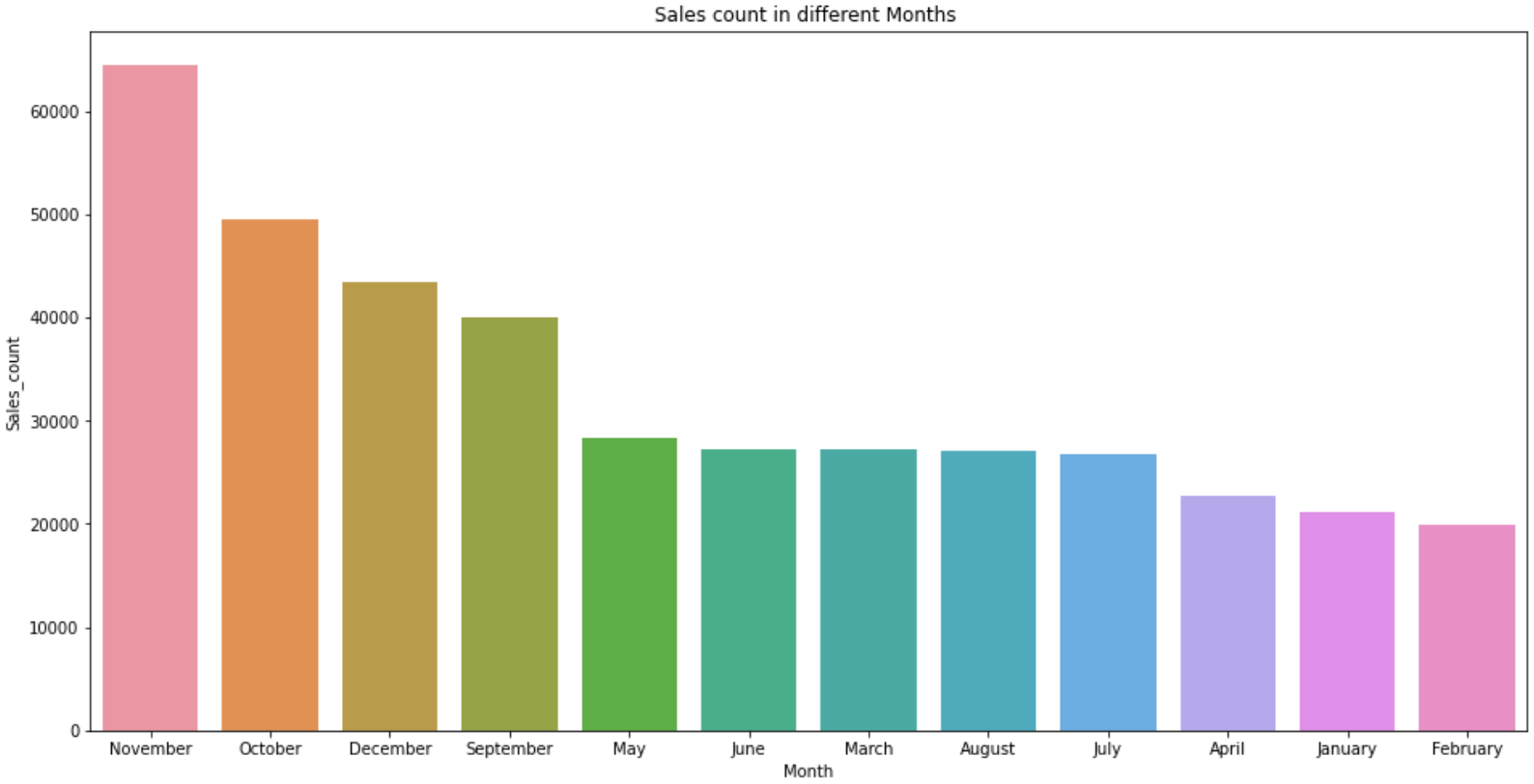
Top 5 Country Based On Number of Customers



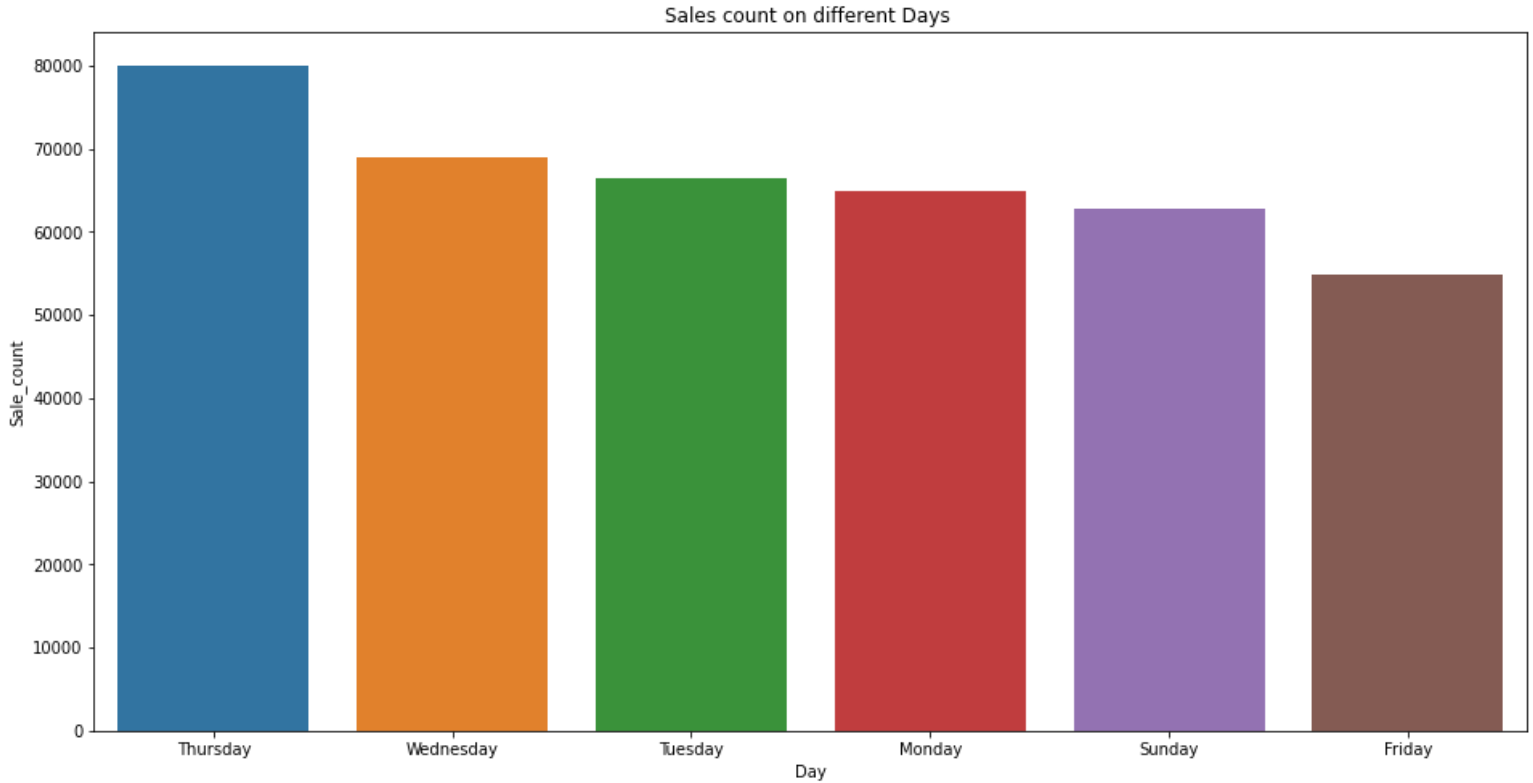
Country Having Least Number Of Customers



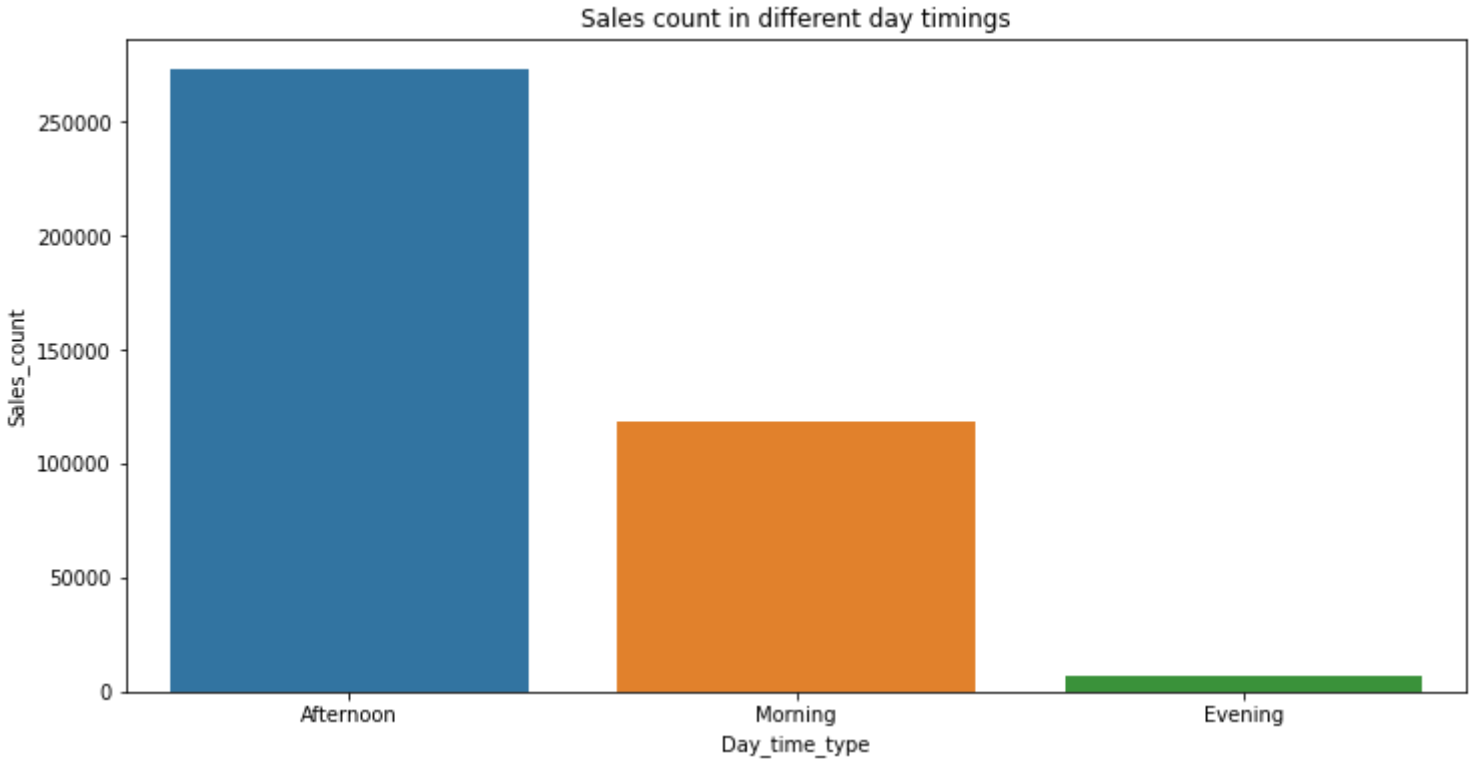
Sales According To Month



Sales On Different Days



Sales On Different Day Timing



Anomaly Detection

Anomaly Detection Using Isolation Forest

- We perform Univariate Anomaly Detection on Quantity and UnitPrice
- After That we perform Multi-variate Anomaly Detection on Quantity and UnitPrice
- So, after combining both Univariate and Multivariate Anomaly Detection we found 38 observations which seems Anomaly. So we have to remove them because of less number of anomaly.
- After removing the Anomaly we have 390222 number of observation in our dataset left.

RFM Model

What is RFM?

- RFM is a method used to analyze customer value. RFM stands for RECENCY, Frequency, and Monetary.
- RECENCY: How recently did the customer visit our website or how recently did a customer purchase?
- Frequency: How often do they visit or how often do they purchase?
- Monetary: How much revenue we get from their visit or how much do they spend when they purchase?

Why it is Needed?

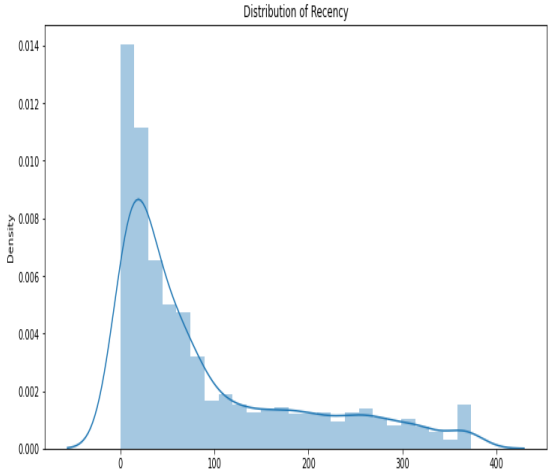
- RFM Analysis is a marketing framework that is used to understand and analyze customer behavior based on the above three factors Recency, Frequency, and Monetary.

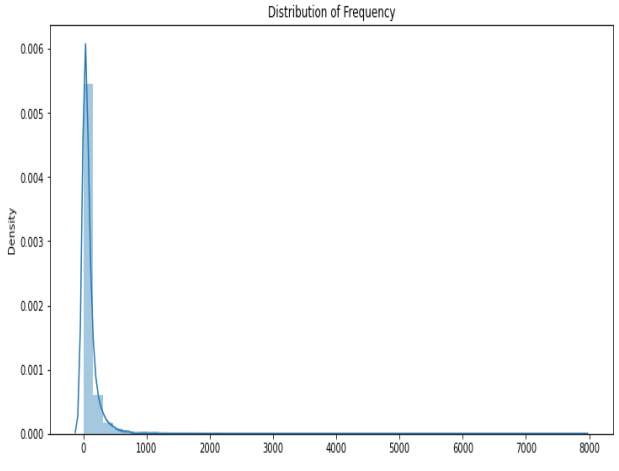
- The RFM Analysis will help the businesses to segment their customer base into different homogenous groups so that they can engage with each group with different targeted marketing strategies.

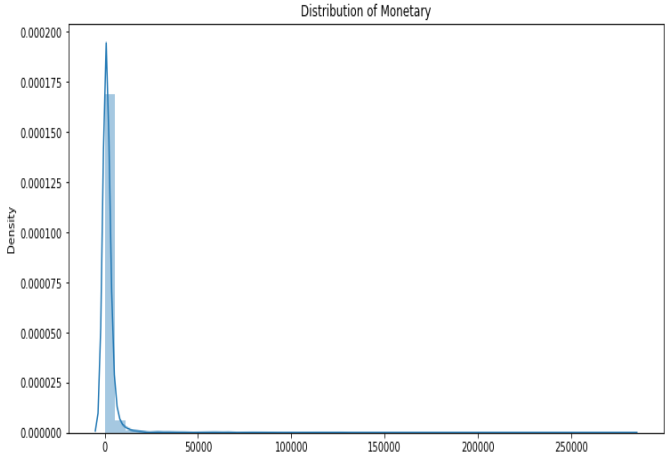
RFM Model Analysis

	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12347.0	2	182	4310.00
2	12348.0	75	31	1797.24
3	12349.0	18	73	1757.55
4	12350.0	310	17	334.40

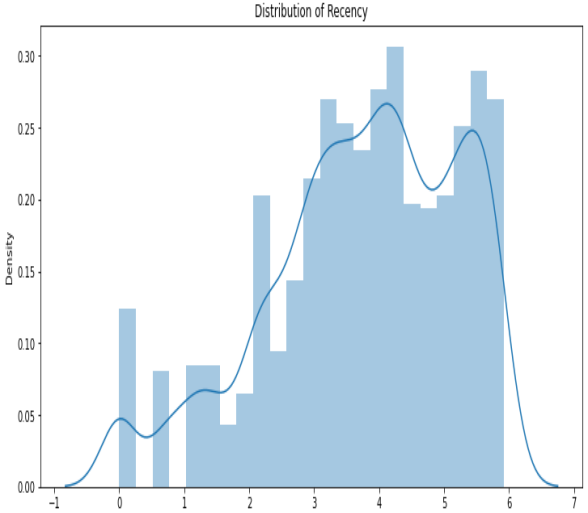
- Recency = Latest Date - Last Invoice Data.
- Frequency = Count of invoice no. of transaction(s).
- Monetary = Sum of Total Amount for each customer.

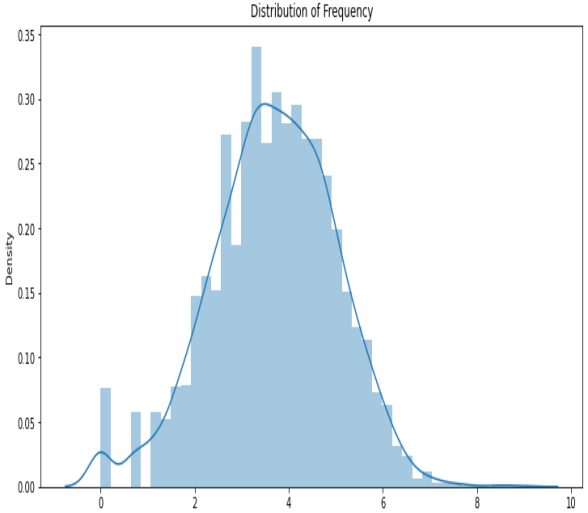


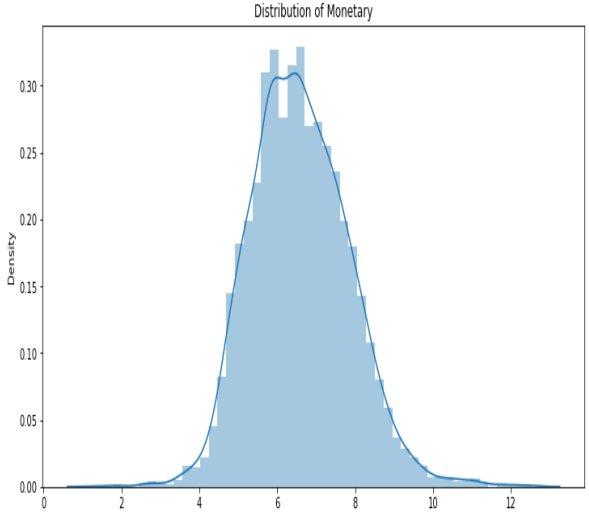




Contd...

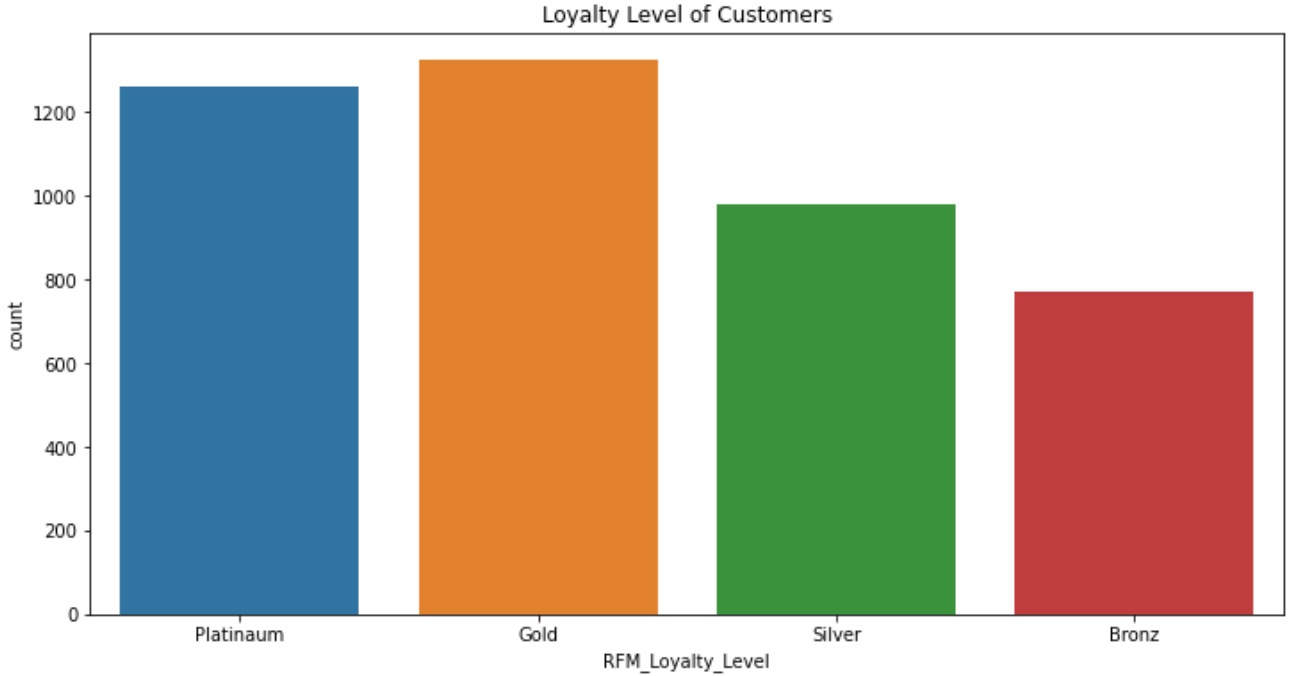






- Apply Log transformation on Frequency, Recency and Monetary. To make distribution more look like normal distribution

Loyalty Level



- So just using RFM Model analysis we created 4 clusters namely Platinum, Gold, Silver and Bronze.

Model Implementation

Algorithm used

Following are the Unsupervised algorithm used.

1. K-Means Clustering
2. DBSCAN
3. Hierarchical Clustering

Following are the Methods to find Optimal Cluster.

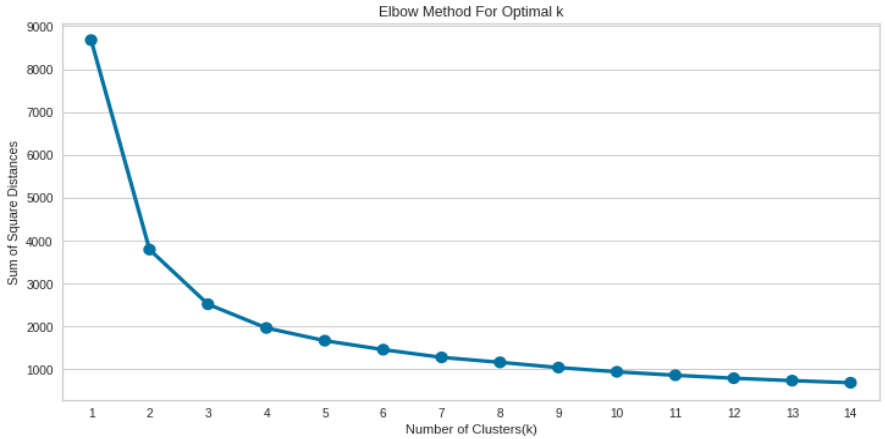
1. Silhouette Score
2. Elbow Method
3. Dendrogram

Comparing All Model with their Optimal Number Of Cluster

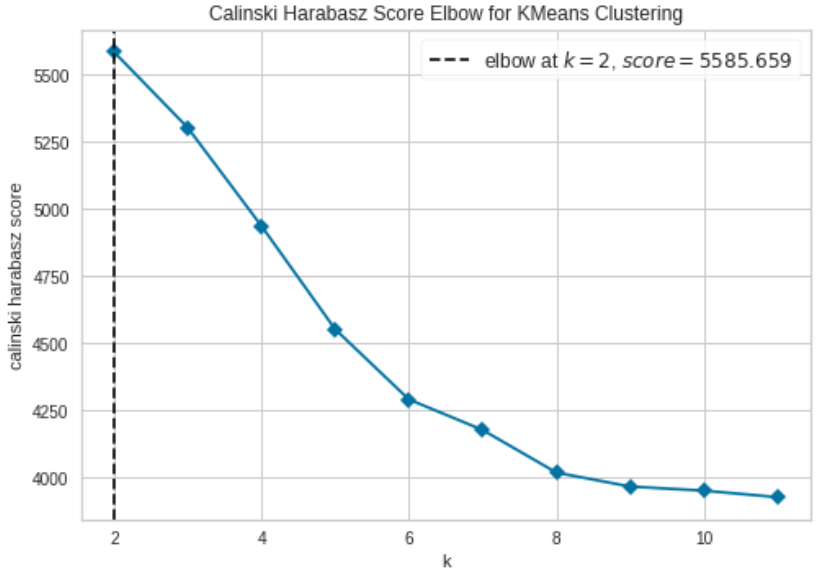
SL.No	Model Name	Data	Optimal Number of Clusters
1	Kmeans with Elbow method(Elbow Visualizer)	Recency and Monetary	2
2	Kmeans withSilhouette Score method	Recency and Monetary	2
3	DBSCAN	Recency and Monetary	2
4	Kmeans with Elbow method(Elbow Visualizer)	Frequency and Monetary	2
5	Kmeans withSilhouette Score method	Frequency and Monetary	2
6	DBSCAN	Frequency and Monetary	2
7	Kmeans with Elbow method(Elbow Visualizer)	Recency ,Frequency and Monetary	2
8	Kmeans withSilhouette Score method	Recency ,Frequency and Monetary	2
9	DBSCAN	Recency ,Frequency and Monetary	2
10	Hierarchical clustering	Recency ,Frequency and Monetary	2

- All model gives same number of optimal cluster which are 2.
- The K-Means Clustering with data Frequency and Monetary gives best Silhouette Score which is 0.47.
- So we are only focusing on that particular model.

K-means Clustering: (Frequency and Monetary)

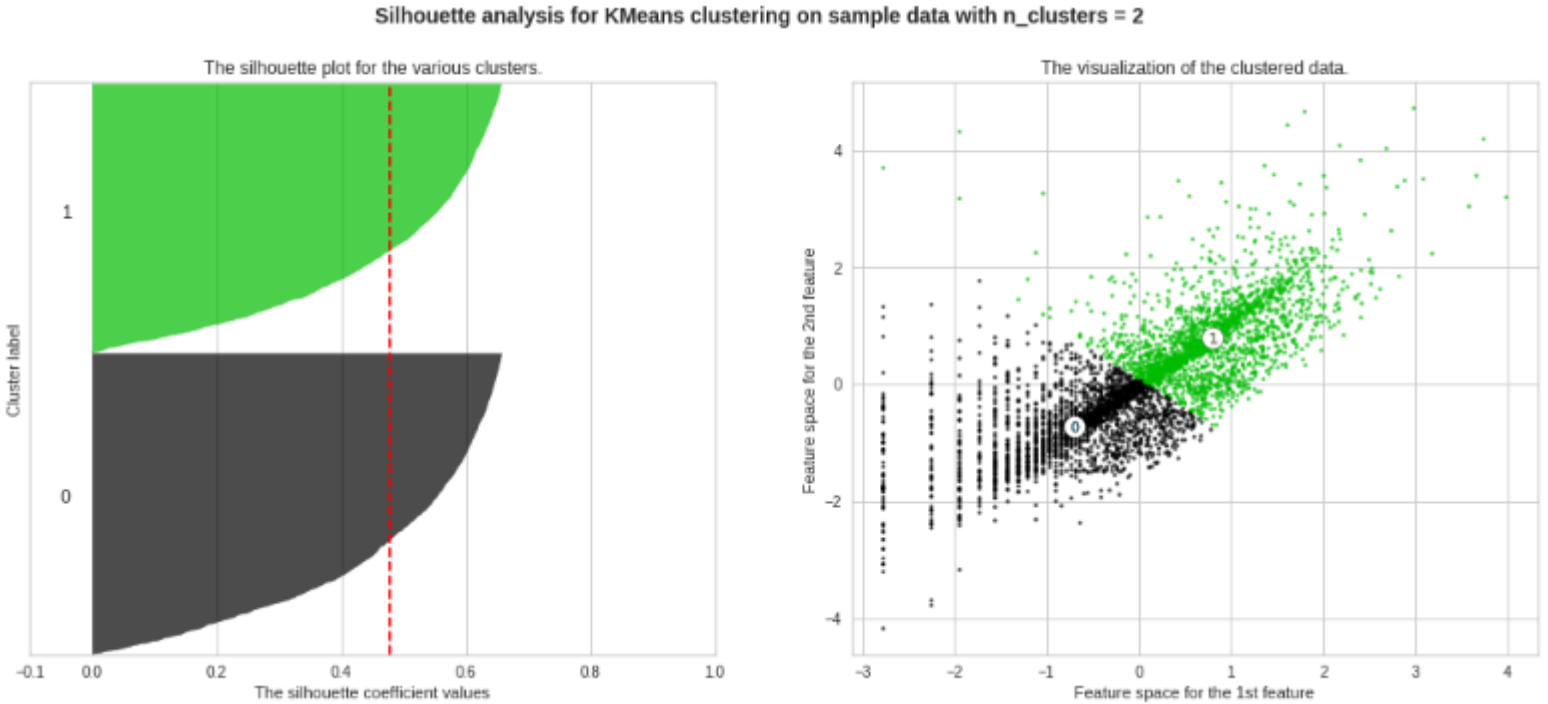



```
For n_clusters = 2, silhouette score is 0.478535709506603
For n_clusters = 3, silhouette score is 0.40764120562174455
For n_clusters = 4, silhouette score is 0.3715810384601166
For n_clusters = 5, silhouette score is 0.3442965607959301
For n_clusters = 6, silhouette score is 0.3586829219947334
For n_clusters = 7, silhouette score is 0.34342098057749704
For n_clusters = 8, silhouette score is 0.3500546906243836
For n_clusters = 9, silhouette score is 0.34419928062567495
For n_clusters = 10, silhouette score is 0.36238664926507114
For n_clusters = 11, silhouette score is 0.3682455762844025
For n_clusters = 12, silhouette score is 0.3534862139672636
For n_clusters = 13, silhouette score is 0.36139542577471895
For n_clusters = 14, silhouette score is 0.3486849890768239
For n_clusters = 15, silhouette score is 0.3628225939841498
```

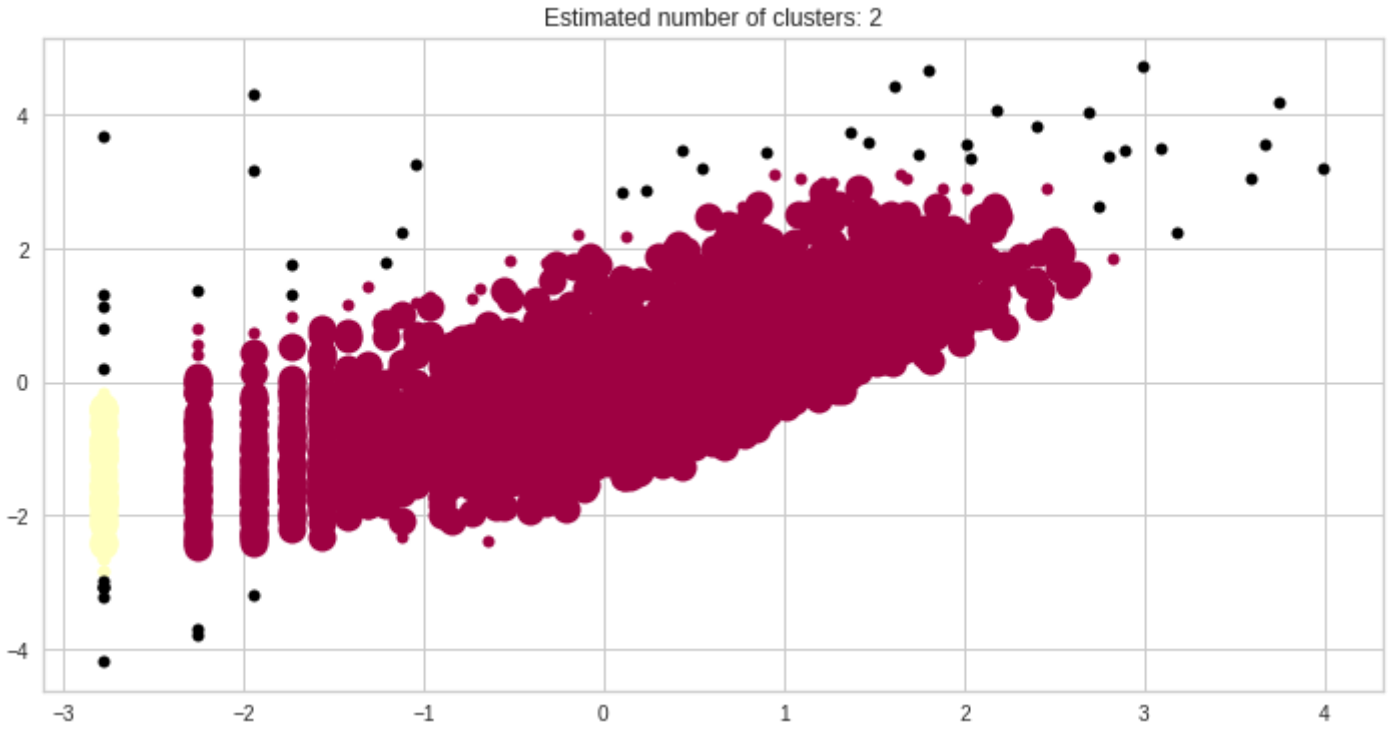


We can see the optimal number of cluster is 2

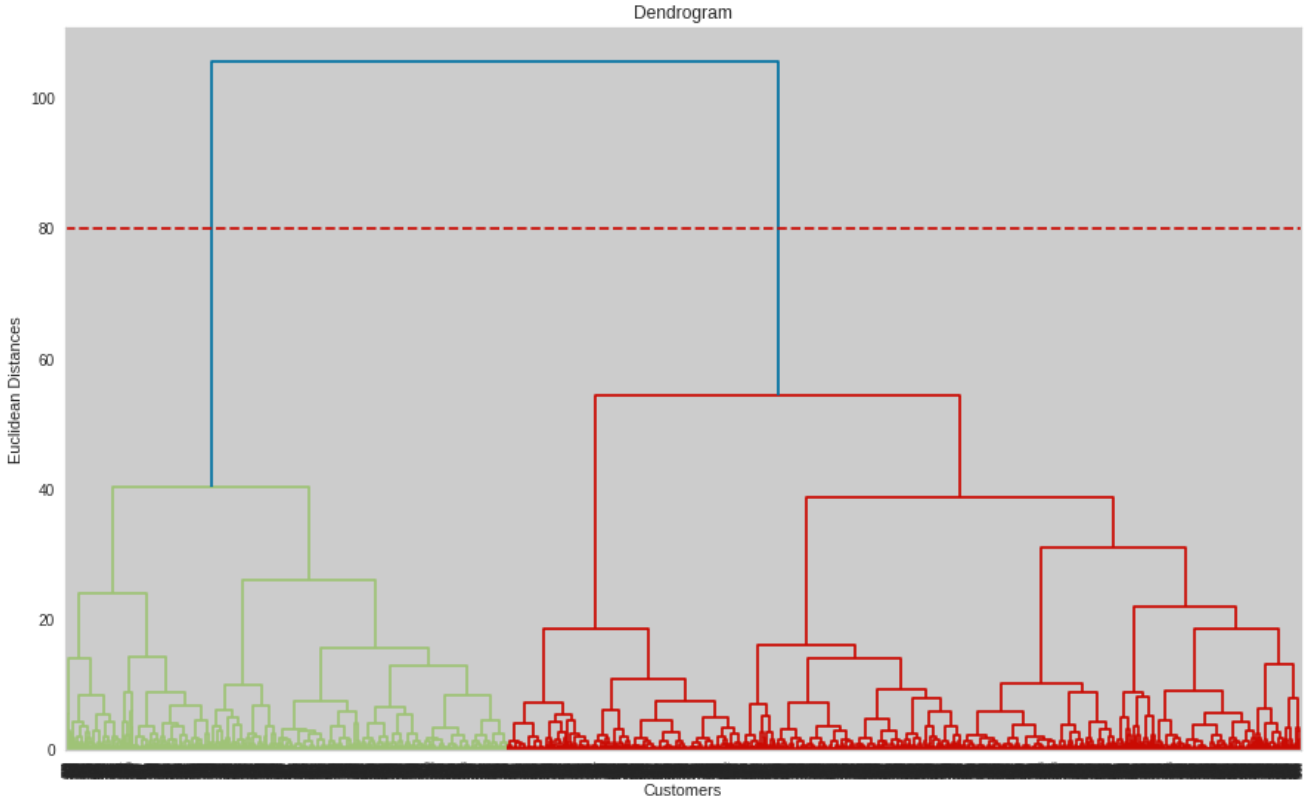
Silhouette Plot (Frequency and Monetary)



DBSCAN Plot



Hierarchical Clustering(Recency, Frequency and Monetary)



- Optimal Number of clusters using Dendrogram.(Optimal Clusters = 2)

Cluster Details

	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
Cluster_based_on_freq_mon_rec										
0	140.975509	1	373	24.833126	1	157	469.732795	3.75	77183.60	2409
1	30.989632	1	372	175.251944	1	7847	4033.085330	161.03	280206.02	1929

- Cluster 0 has high recency rate but very low frequency and monetary. Cluster 0 contains 2414 customers.
- Cluster 1 has low recency rate, but they are frequent buyers and spends very high money than others as mean monetary value is very high.

Conclusions

- The Top 3 products are white hanging T-Light Holder, Regency cake stand 3 tier and Jumbo Bag red retro spot.
- Top 3 rarely sold product are Blue felt heart flower, Glass cake cover and Cracked glaze earrings red.
- The most Frequent Customer Id is 17841.
- United Kingdom has highest number of customers on the other hand Saudi Arabia has the least number of customers.
- In the Month of November, the sales are very high and in the month of February sales are very low as compered to others.
- The sales are very high on Thursday.
- At afternoon, the sales count is very high.
- We Found 38 observation which are Anomaly.
- The Silhouette Score on data Frequency and Monetary we get 0.47 which is higher among the all.
- We get Optimal number of cluster are 2 using Elbow method, Silhouette score and dendrogram for all models.
- We can say that our unsupervised model is ready to deploy to solve business problem.

THANK YOU