Capstone Project

Online Retail Customer Segmentation

By- Samiksha Bandbuche

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.

Points to be discussed:

- Introduction
- Data Summary
- Exploratory Data Analysis
- Recency, Frequency and Monetary Value score
- Model Used
- Conclusion

Introduction:

Customer segmentation is the process of separating your customers into groups based on certain traits they share.

Segmentation offers a simple way of organizing and managing your company's relationships with your customers. This process also makes it easy to tailor and personalize your marketing, service, and sales efforts to the needs of specific groups. This helps boost customer loyalty and conversions.The division is based on customers having similar:

- Needs (i.e., so a single whole product can satisfy them)
- Buying characteristics (i.e., responses to messaging, marketing channels, and sales channels, that a single go-to-market approach can be used to sell to them competitively and economically)

Data Summary:

Data Set Name- Online Retail

Dataset -

Rows-541909

Columns-8

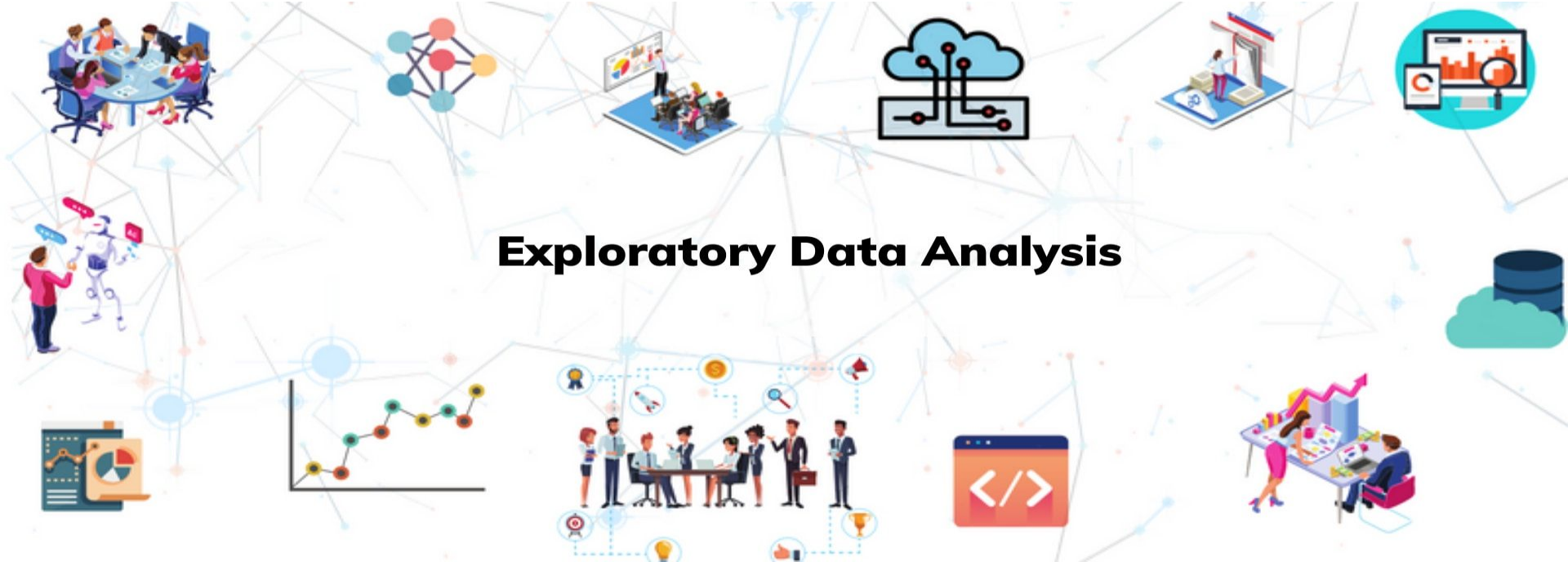
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Variables:

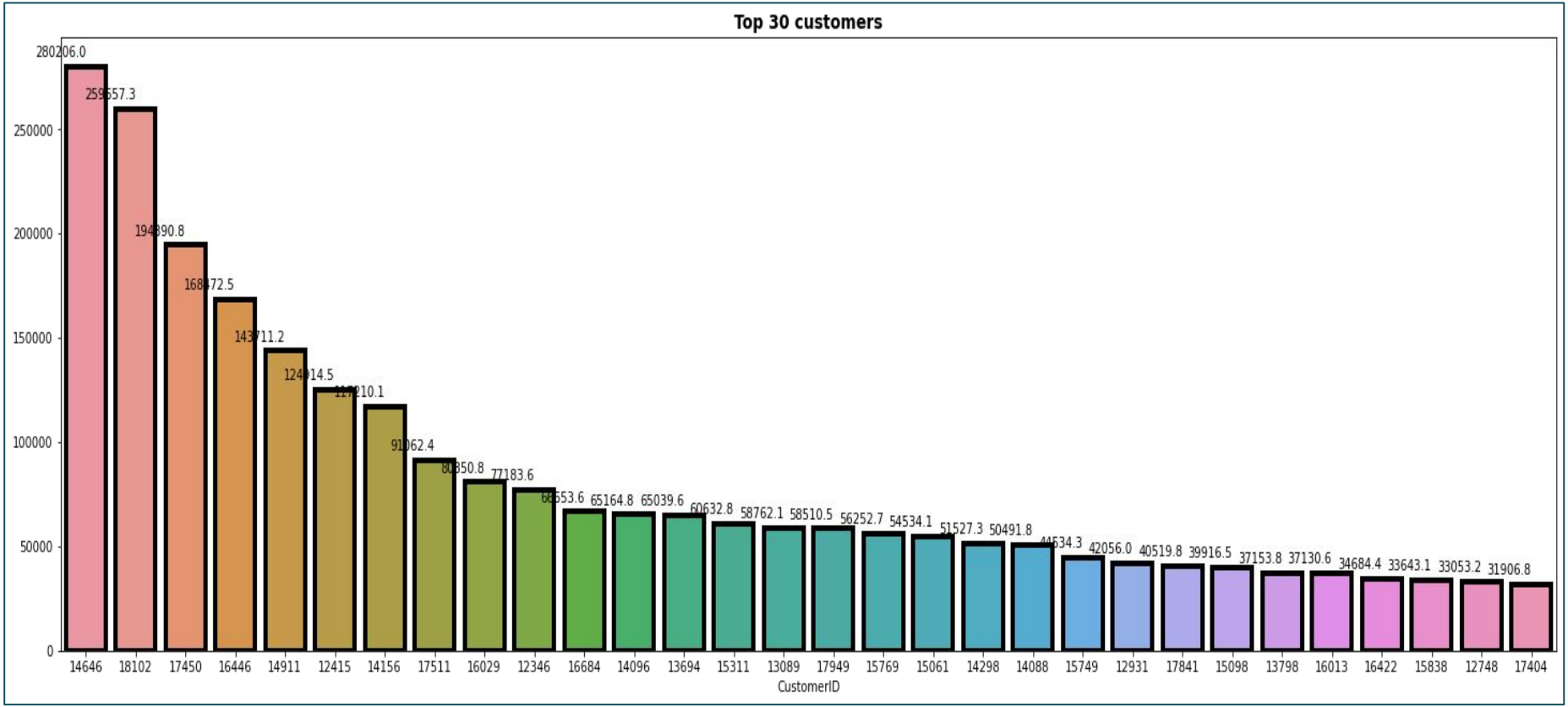
- **InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.**
- **StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.**
- **Description: Product (item) name. Nominal.**
- **Quantity: The quantities of each product (item) per transaction. Numeric.**
- **InvoiceDate: Invice 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.**

Exploratory Data Analysis:

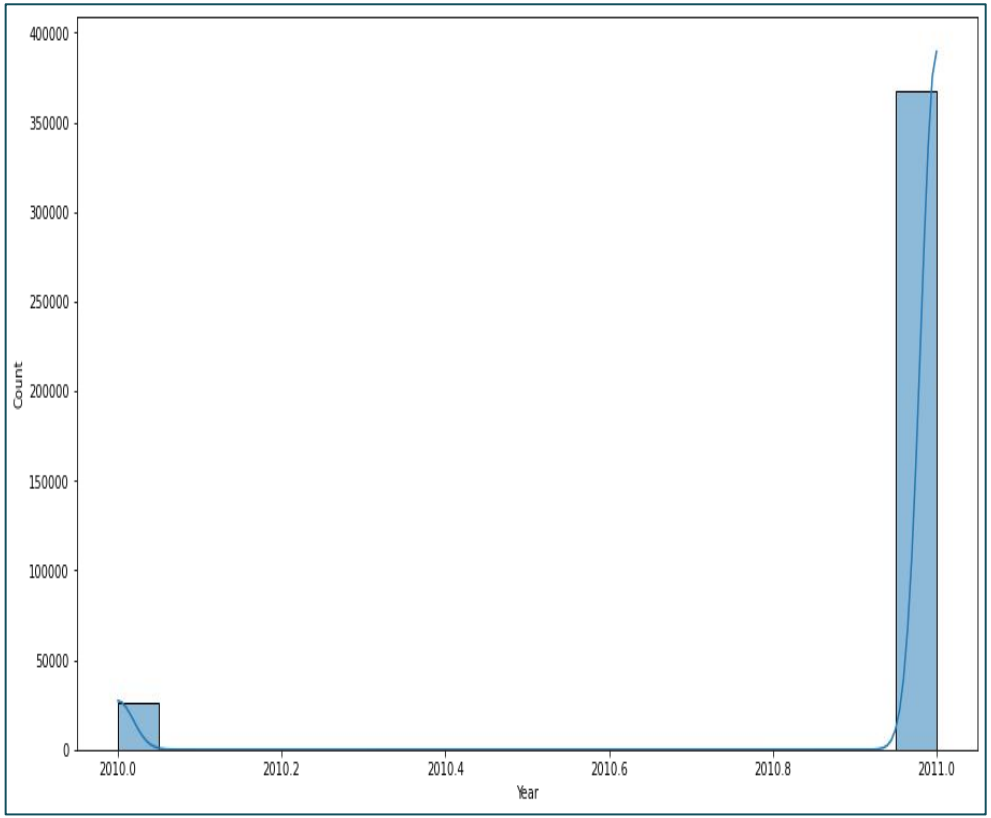
Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods.

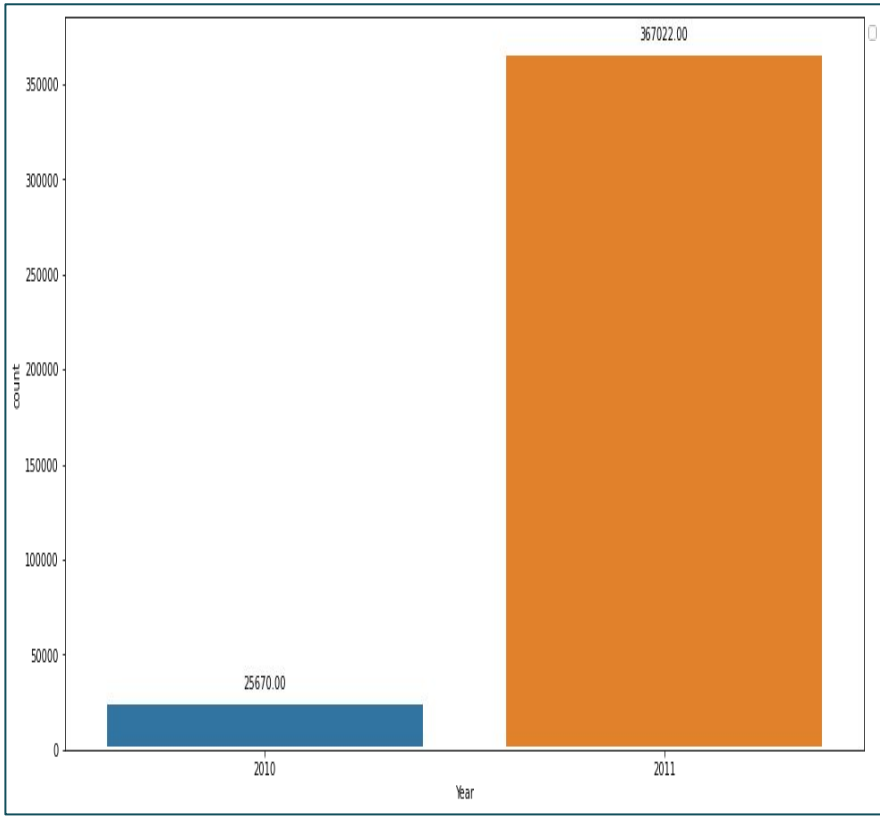


Top 30 Customers of our Retail Chain are:



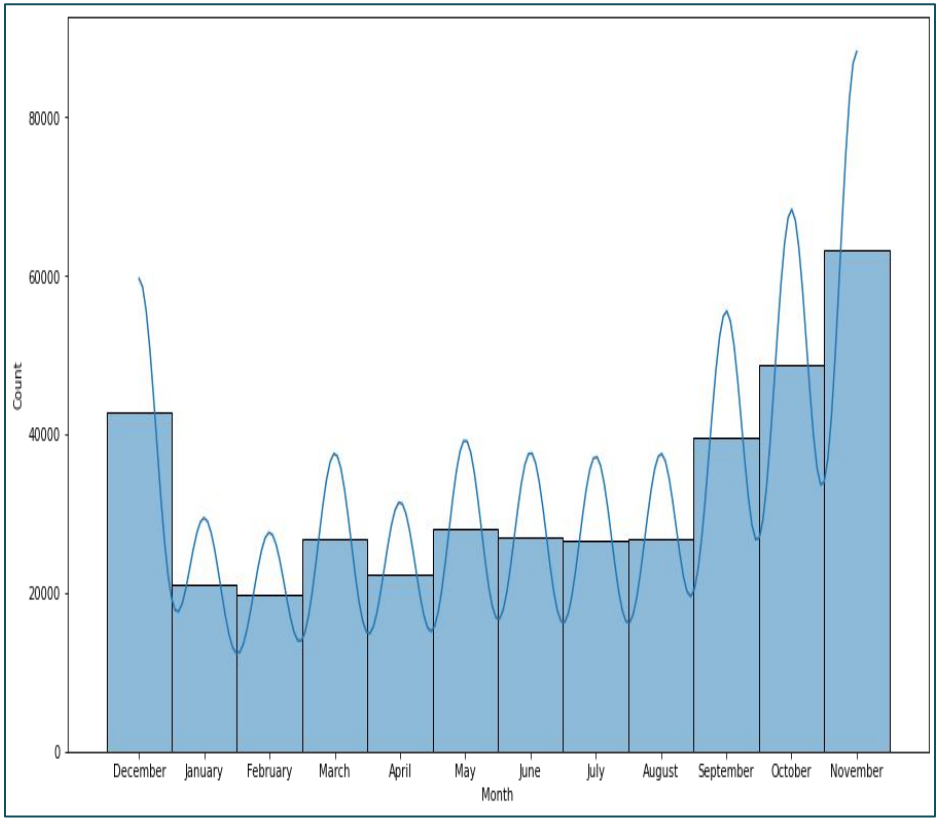
Purchase Per Year:

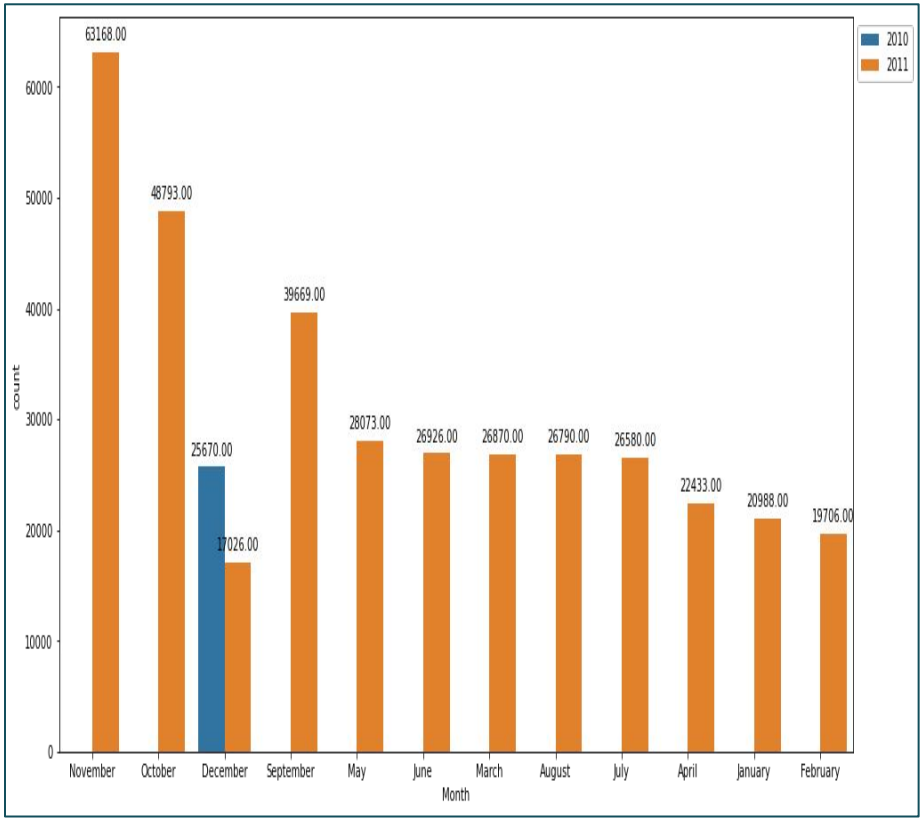




As observed , we have data of only december month of year 2010 thats why we can see that there is huge spike in purchase in year 2011.

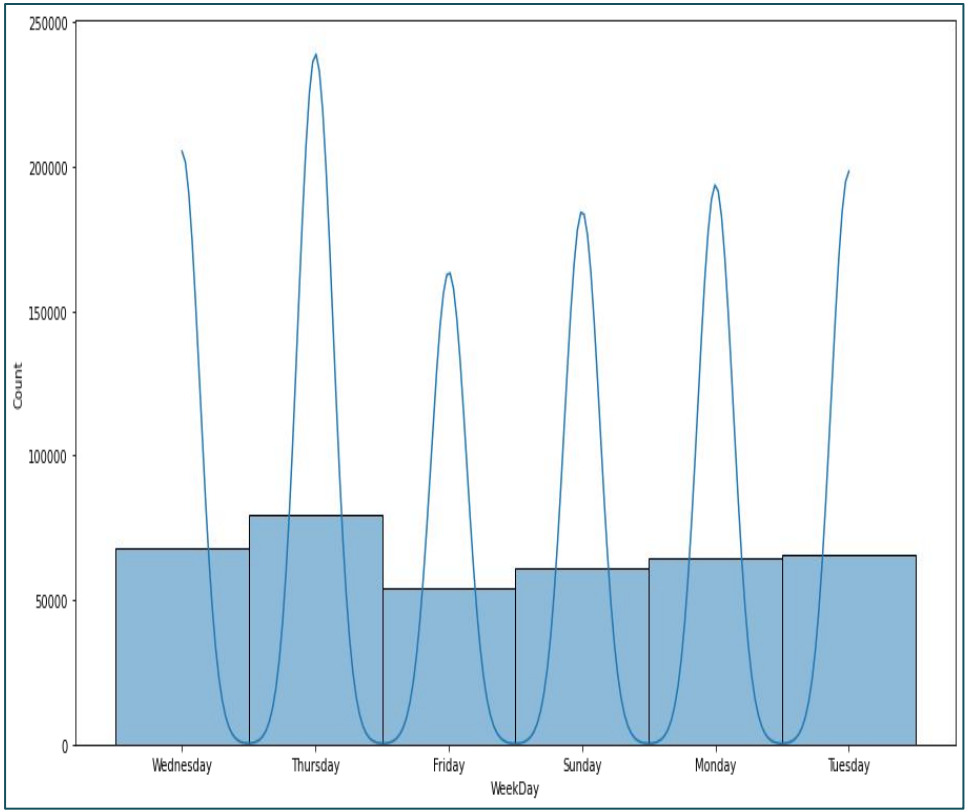
Purchase Per Month:

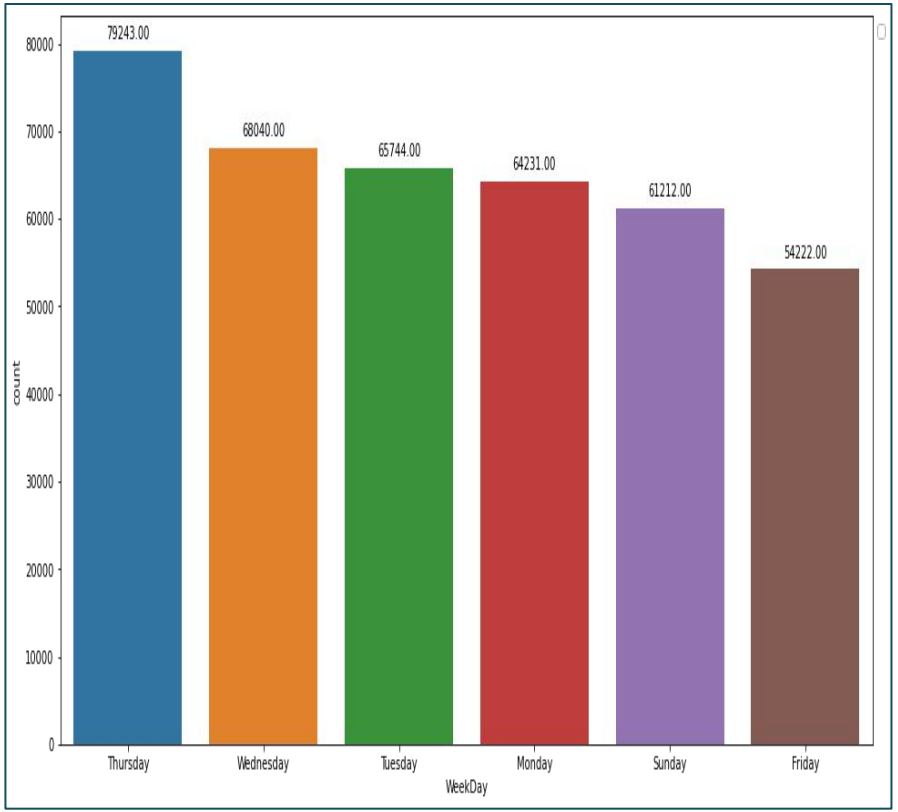




We can see that November and October month have highest number of purchase.

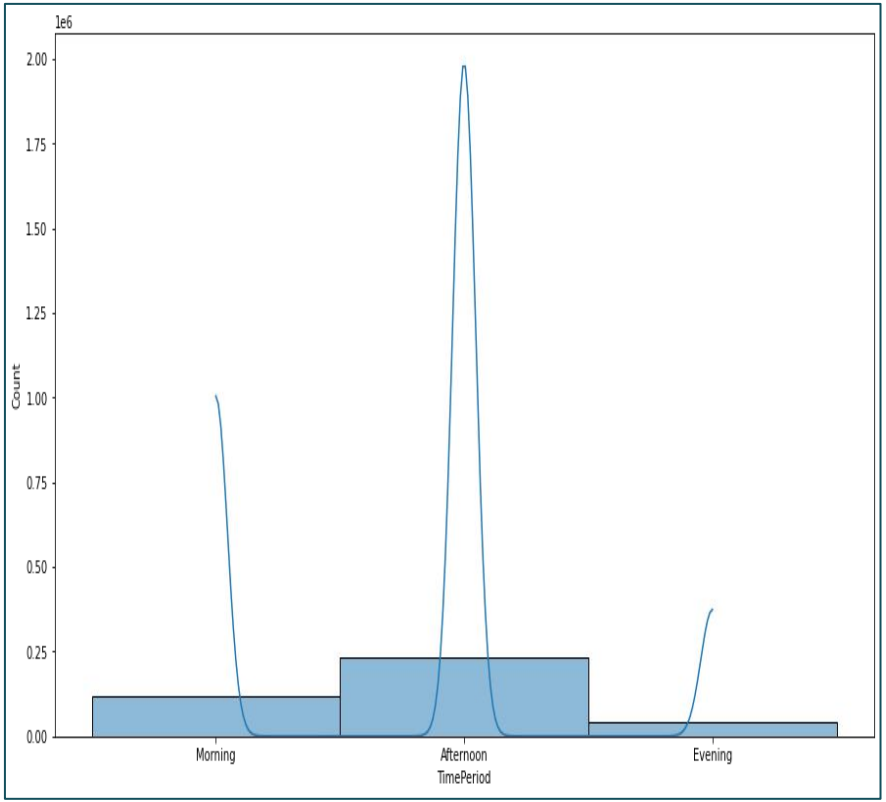
Purchase Per Weekday:

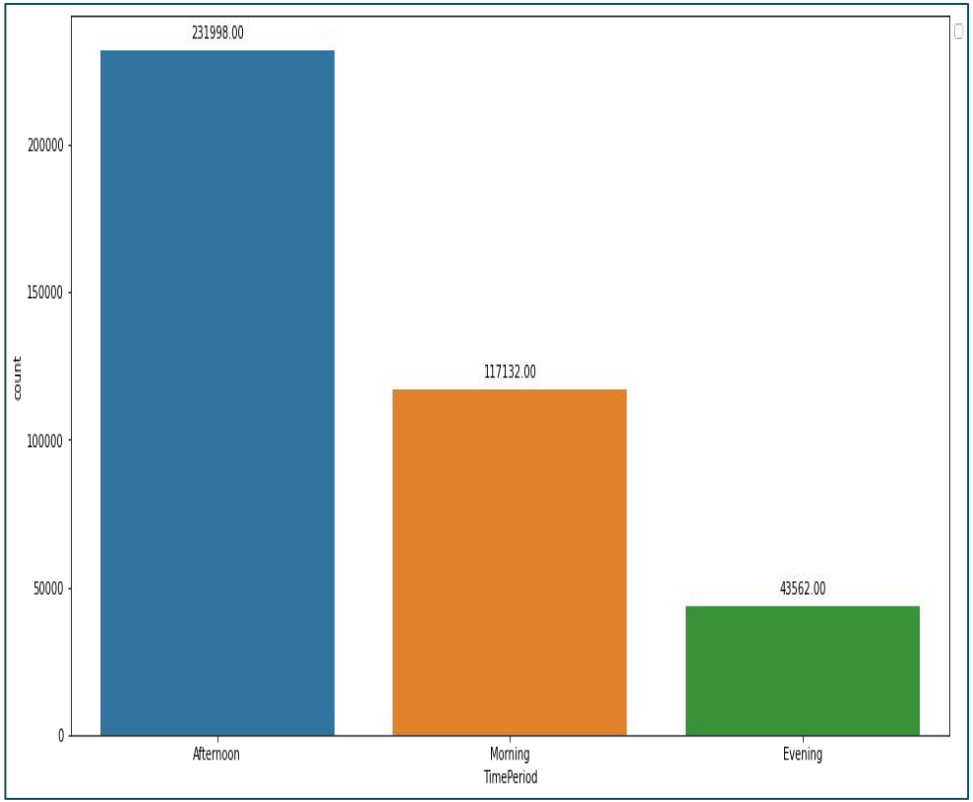




We can see that Thursday and Wednesday have the highest number of purchase.

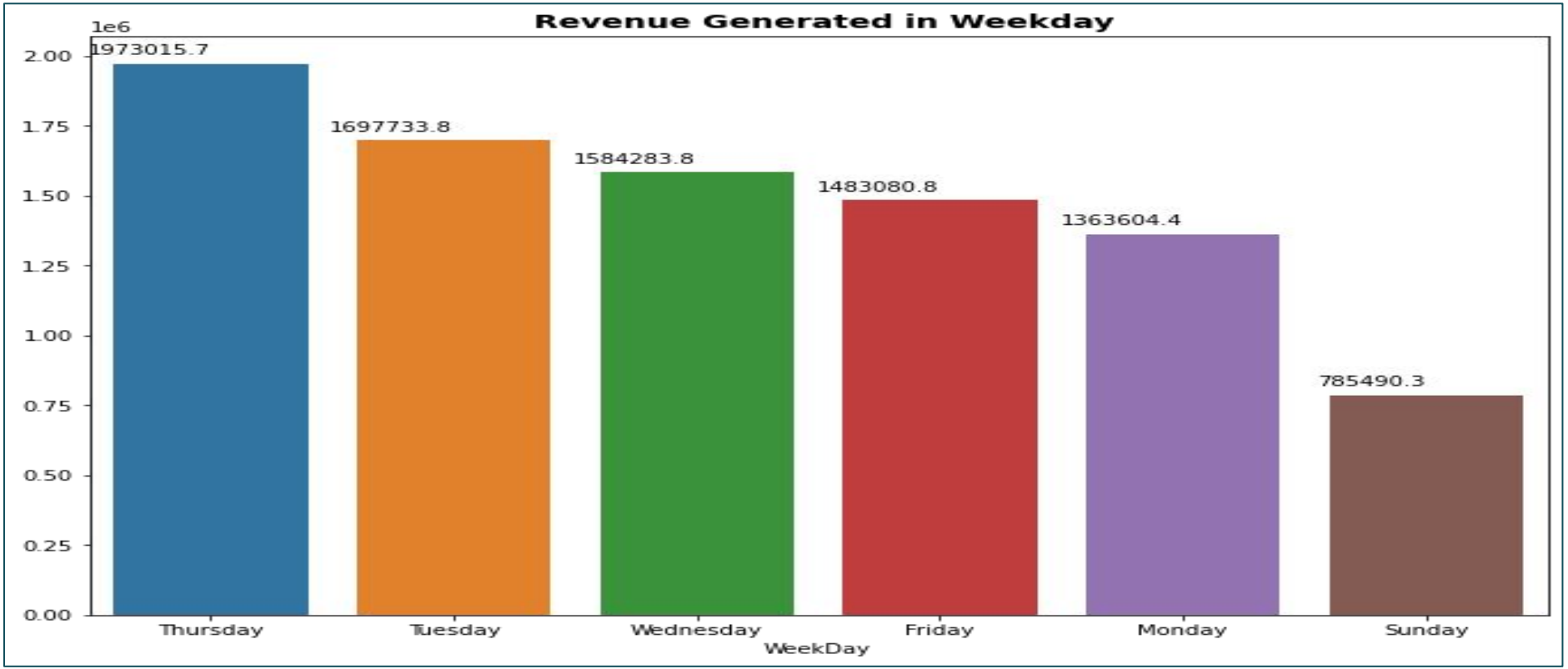
Purchase Per Time Period:





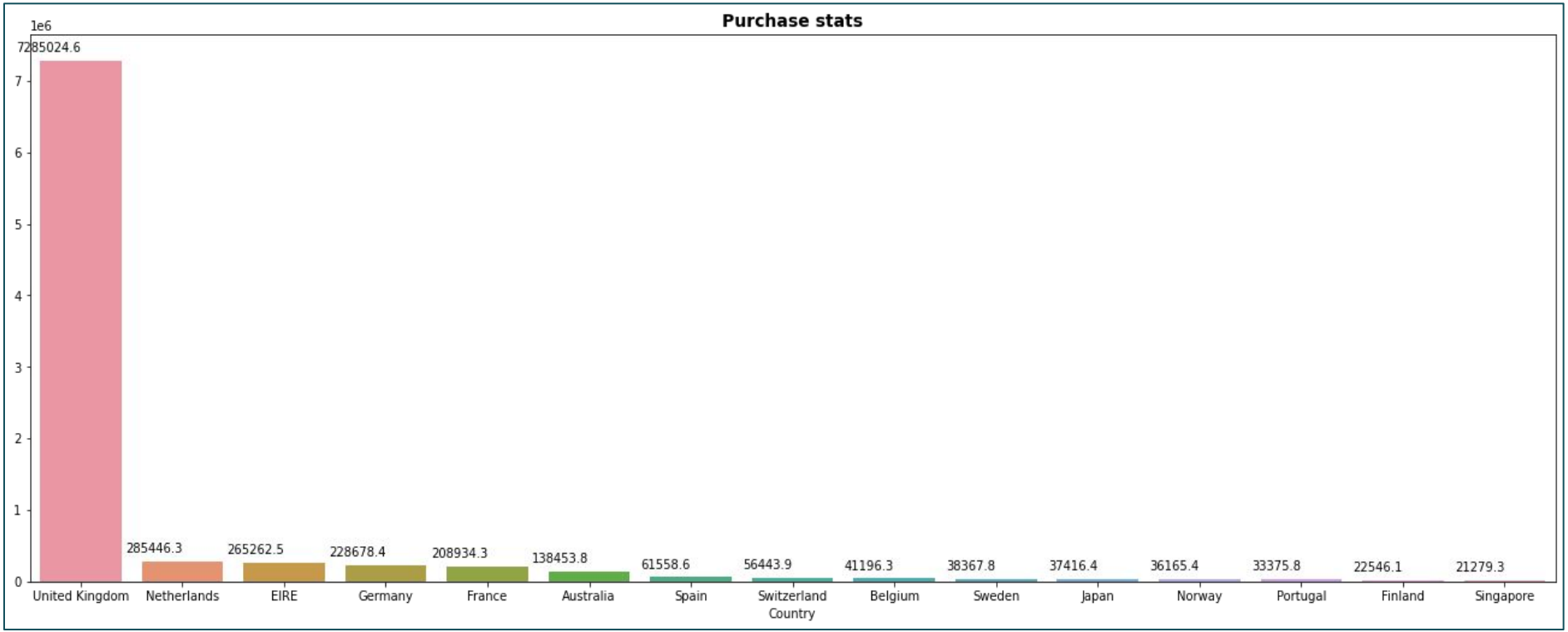
We can see that the most of people like to go for purchasing in afternoon time period.

Revenue Generated in Weekdays:



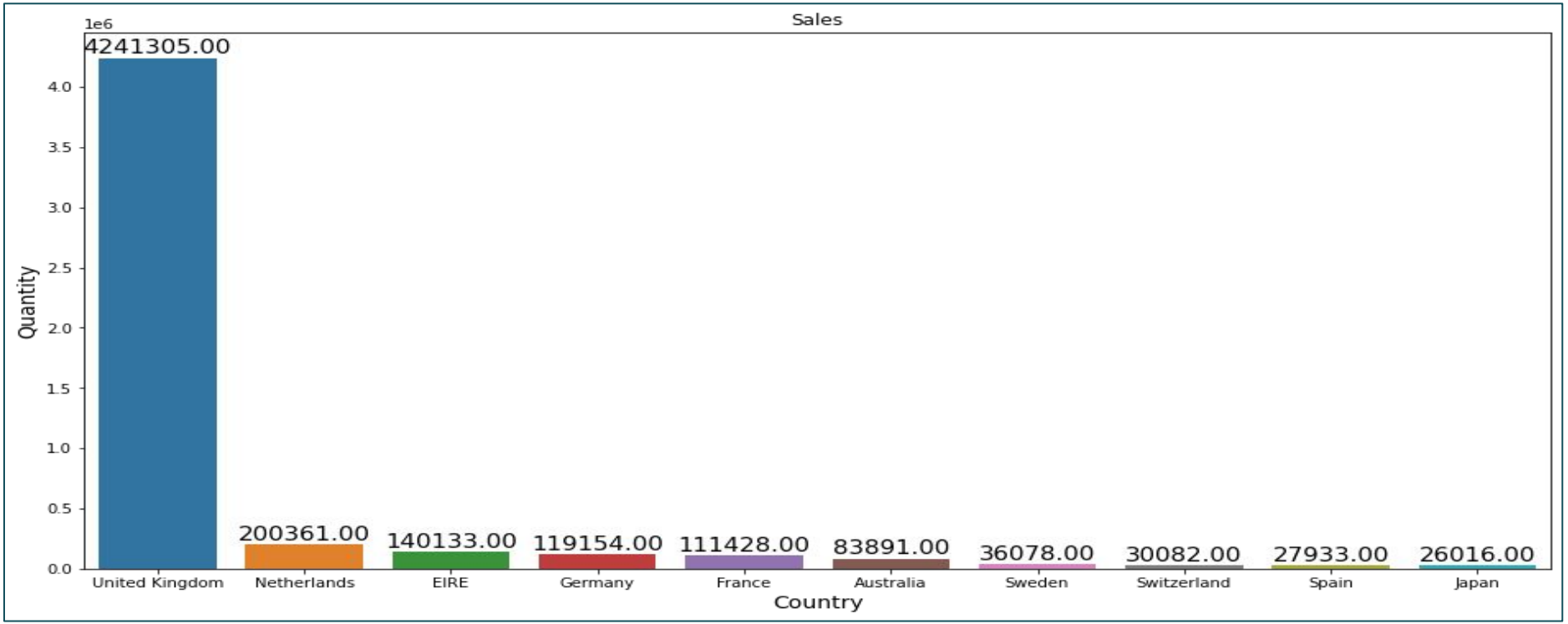
The highest revenue was generated on Thursday and the lowest revenue was generated on Sunday.

Purchase Satistics:



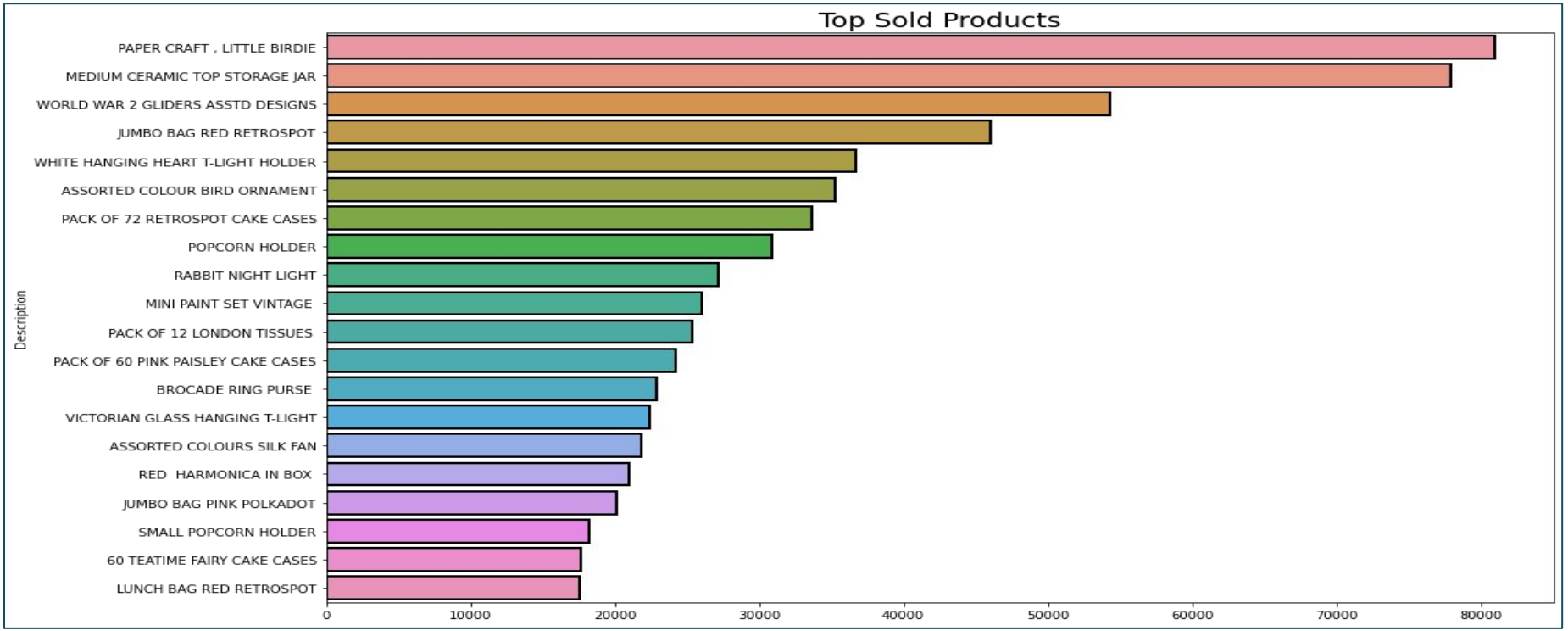
It can be seen that the country which have purchased more number of items as compared to other countries is United Kingdom and the country which have purchased least items is Singapore.

Top 10 Purchasing Countries:



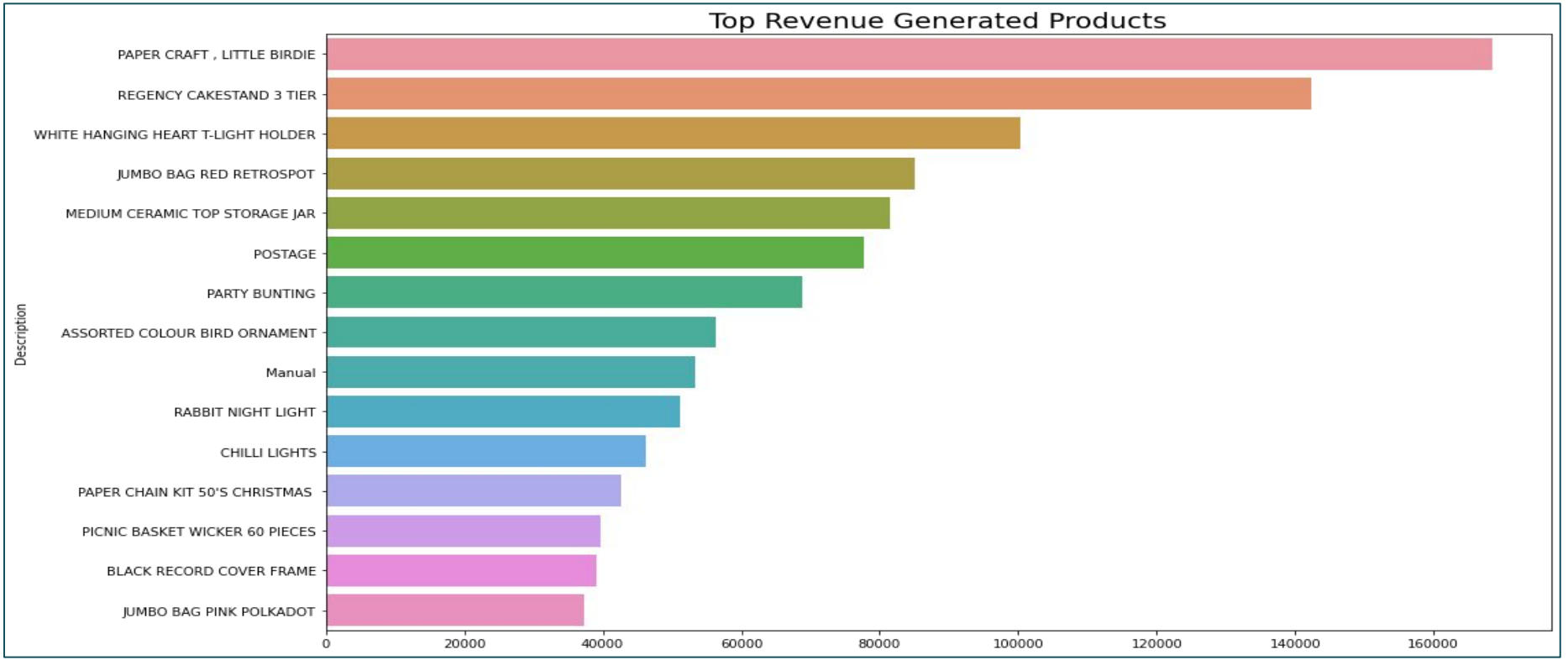
We can see that United Kingdom is at top in the list of top 10 purchasing countries and Japan is at the bottom in the list.

Top Sold Products :



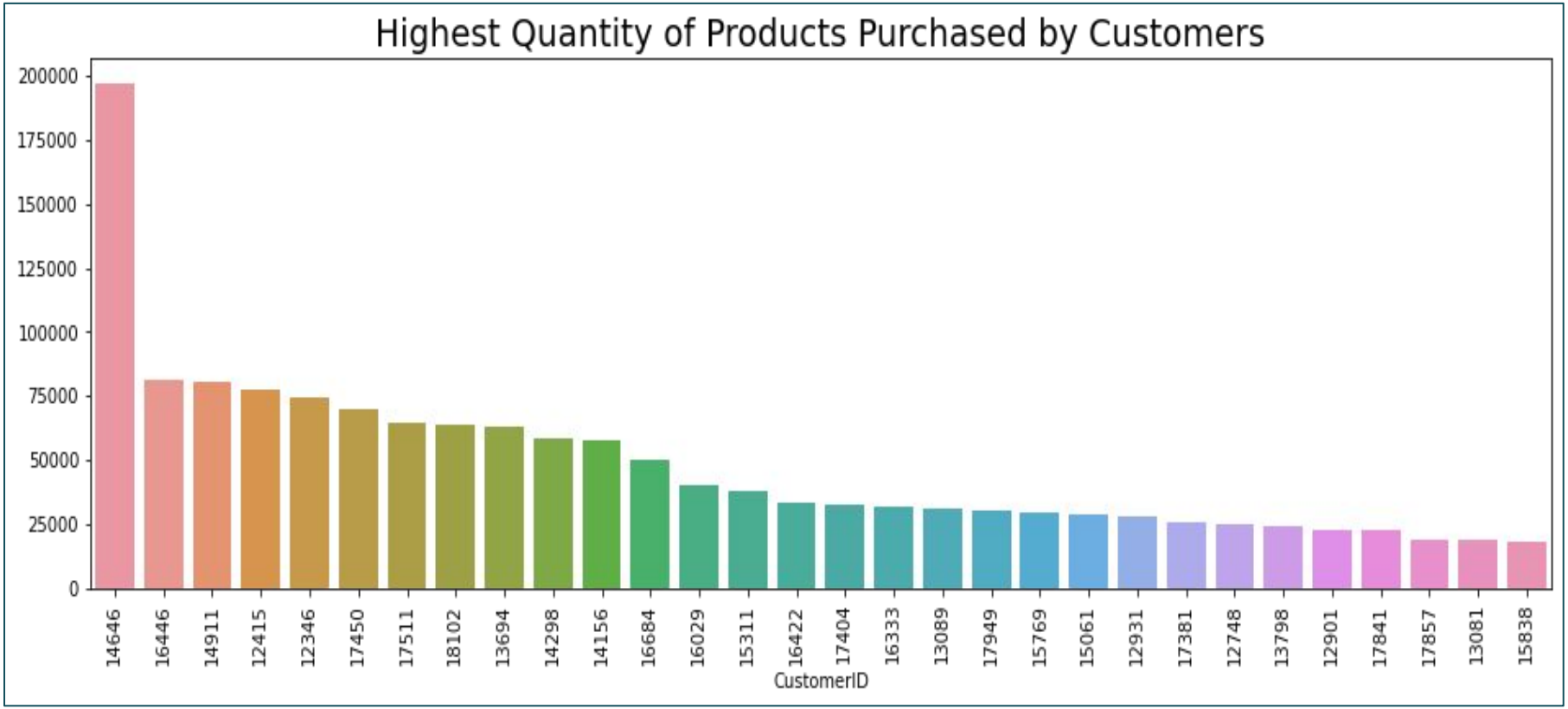
It can be observed that most sold product is "Paper craft, little birdie" and least sold product is "Lunch Bag Red Retrospot".

Top Revenue Generated Products:



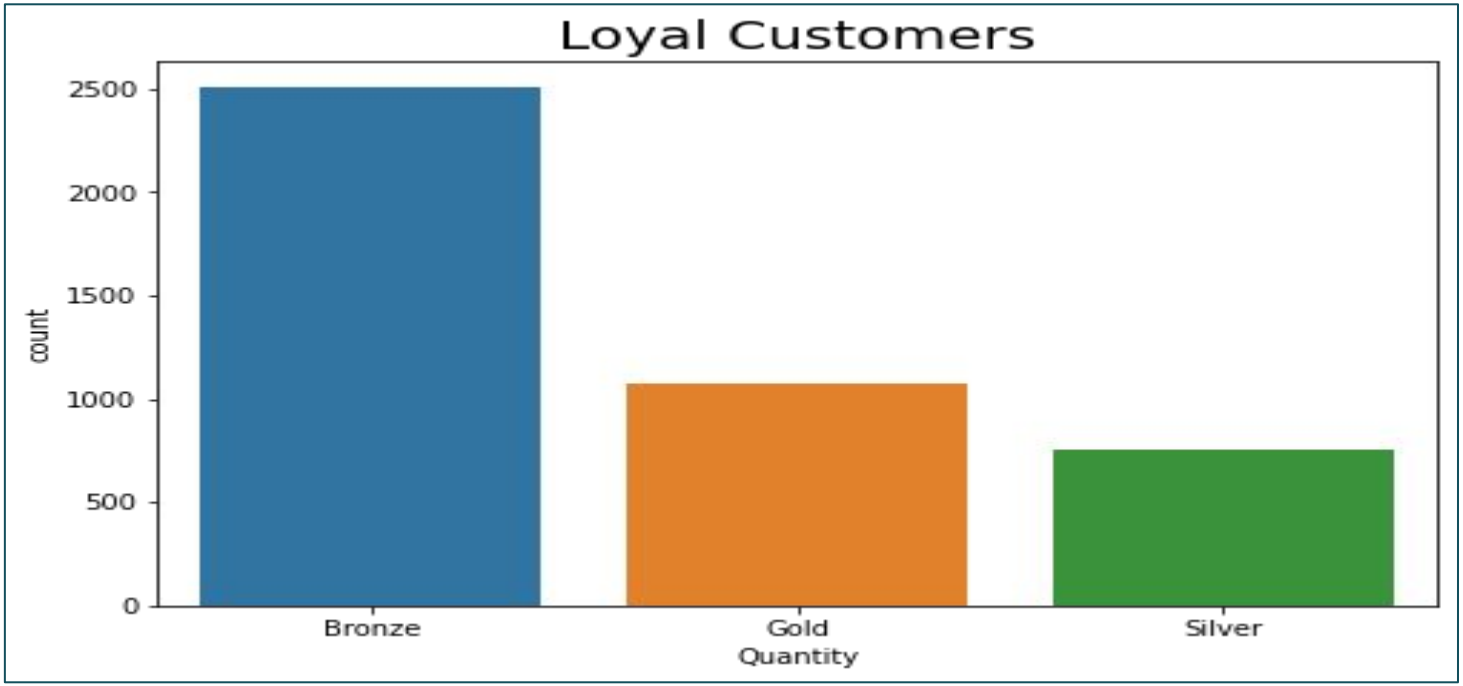
It can be seen that most revenue generated product is "Paper craft, little birdie" and least revenue generated product is "Jumbo Bag Pink Polkadot".

Customer Statistics:



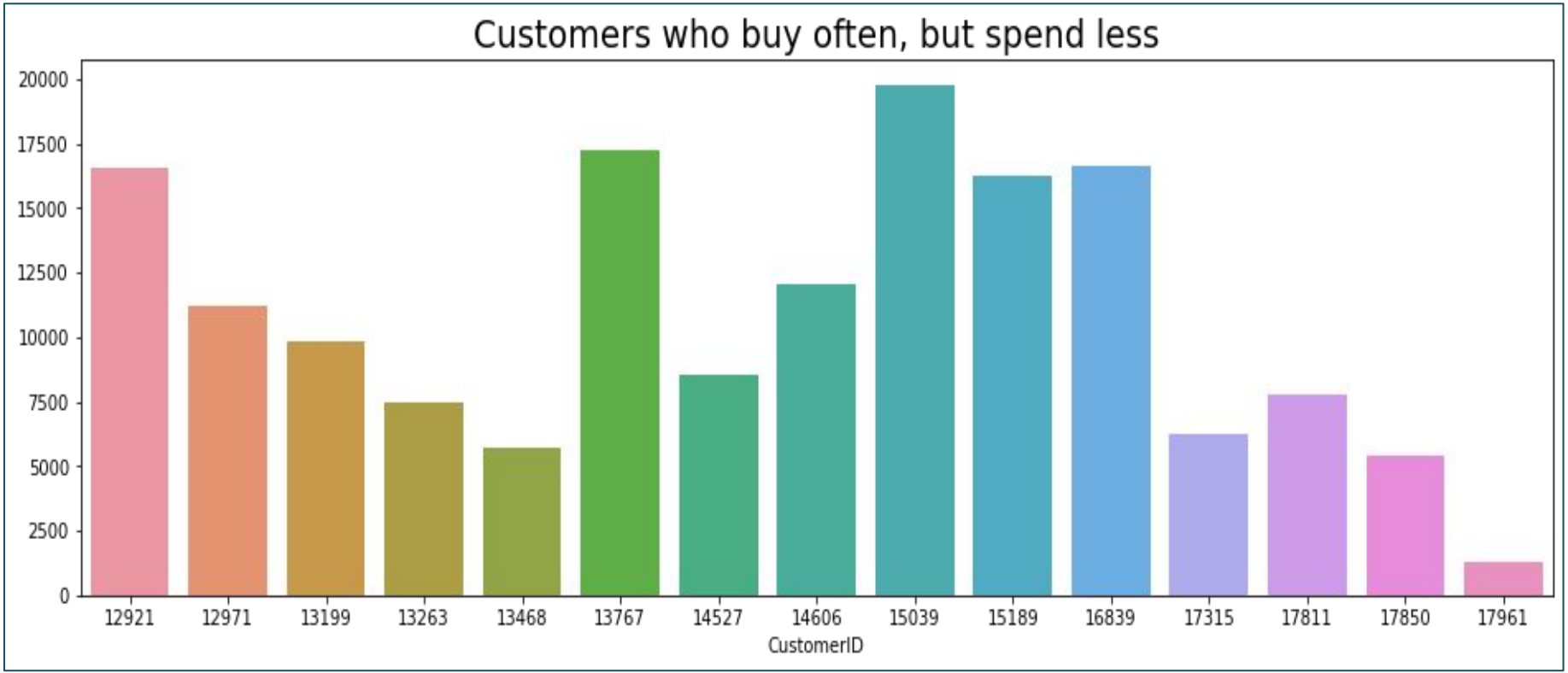
The barplot shows the Customer ID of people who have purchased highest quantity of products

Since we don't have the purchase stats of our customers with the other companies, Let us segregate them into 3 Categories based on their purchases as Bronze, Silver and Gold.



We can see that we have more number of bronze customers.

Which customers who buy often but spend less?

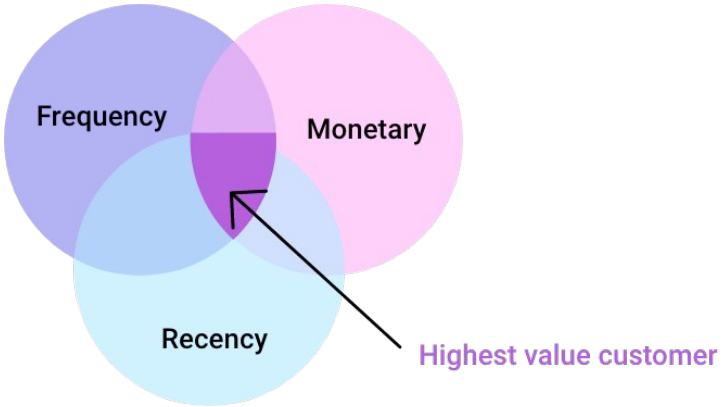


How many customers we have lost?

It can be observed from above that we have lost the maximum customers in the month of November or we can also say that we have lost maximum number of customers in the end of the year.

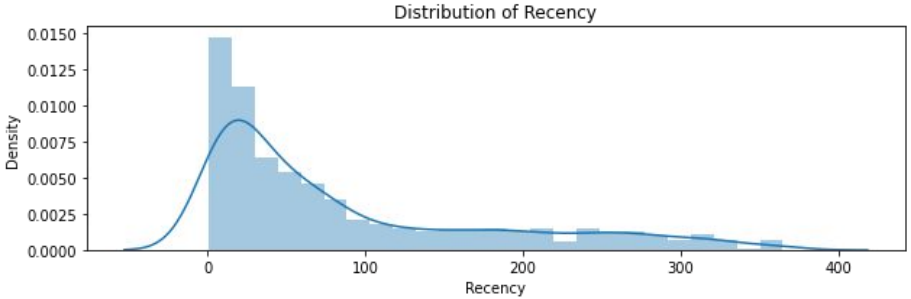
Recency, Frequency and Monetary Value score:

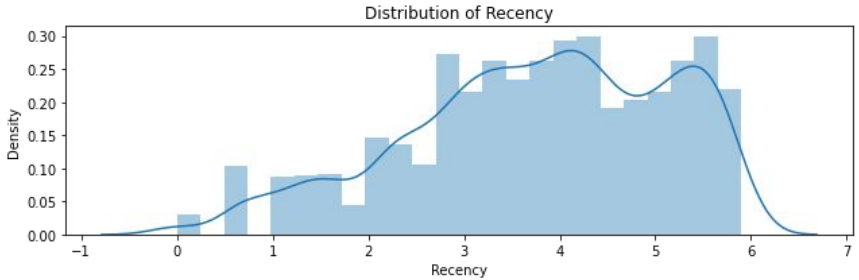
CustomerID	Recency	Frequency	MonetaryValue	Recency_Q	Frequency_Q	MonetaryValue_Q
12346	326	1	77183.60	1	1	4
12747	3	96	3837.45	4	3	4
12748	1	4054	31081.74	4	4	4
12749	4	199	4090.88	4	4	4
12820	4	59	942.34	4	3	3

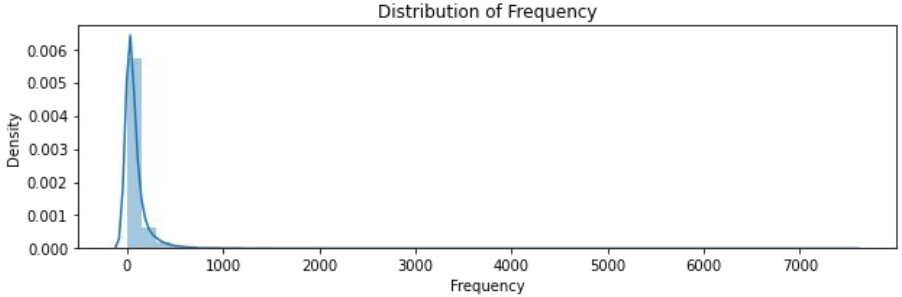


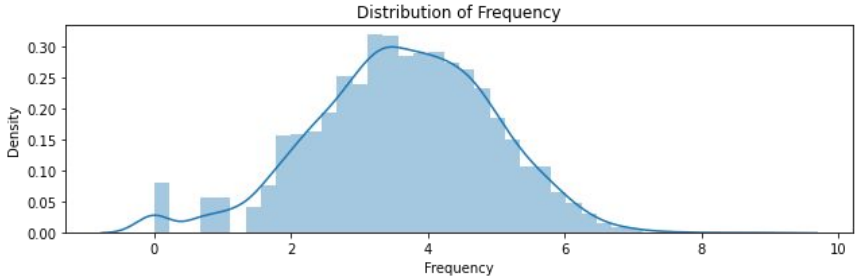
General_Segment	Recency mean	Frequency mean	MonetaryValue mean	MonetaryValue count
1.Gold	26.1	182.0	3830.1	1493
2.Silver	95.6	34.0	691.3	1679
3.Bronze	204.8	10.9	188.4	682

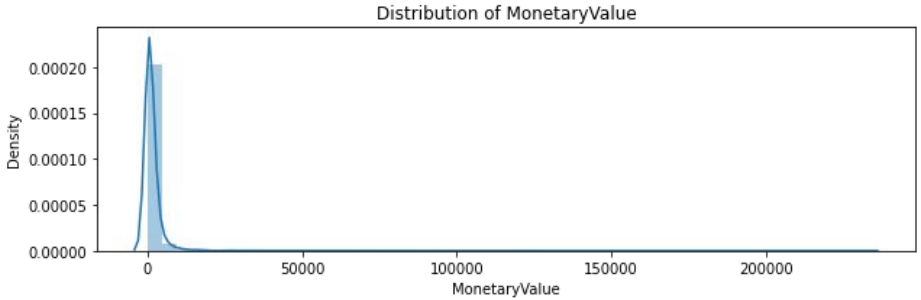
Before and after log transformation:

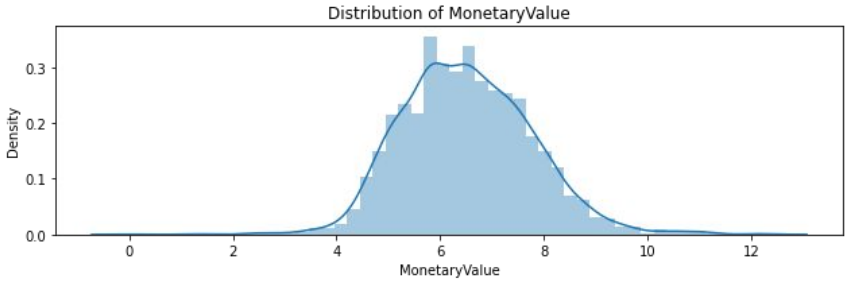




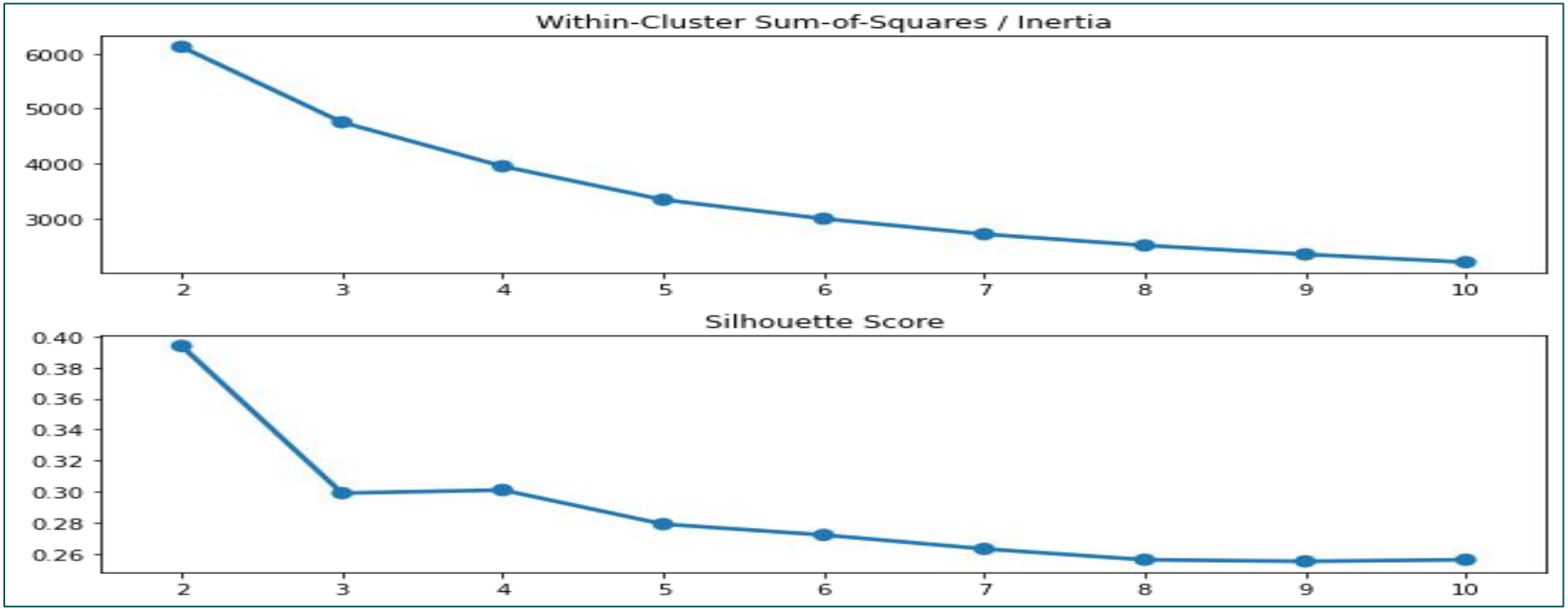






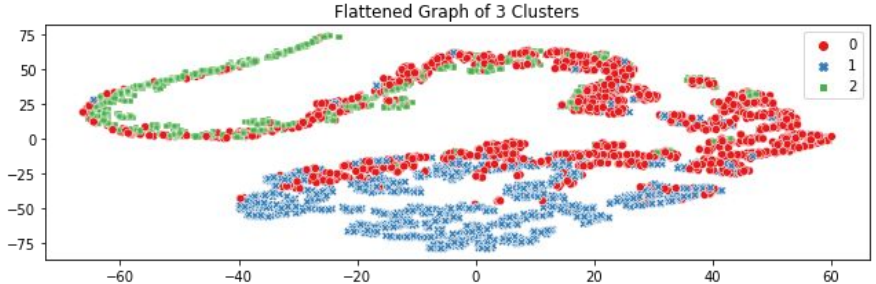


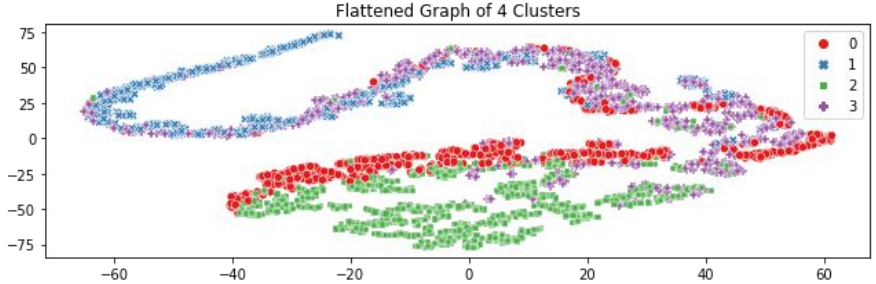
Finding Optimal Cluster:

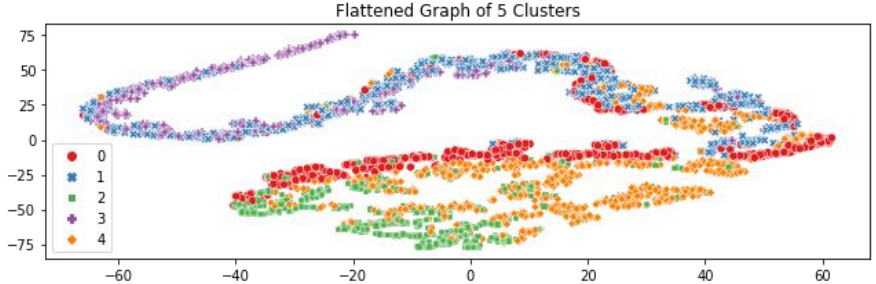


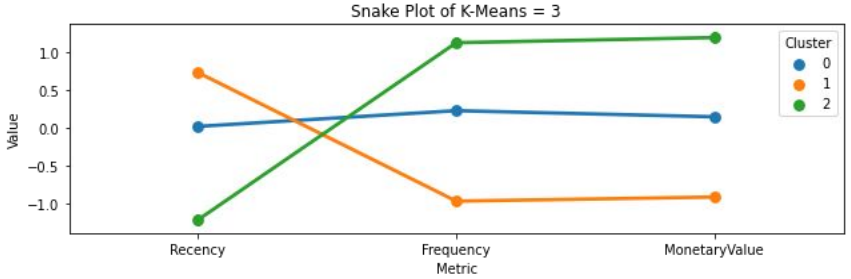
Based on the inertia and silhouette score, the optimal number of cluster is 3. However, during the implementation of K Means, cluster of 3, 4, and 5 will be tested to experiment which cluster makes most business sense.

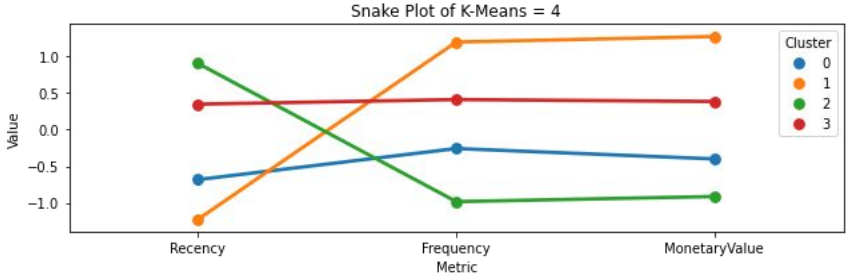
Model used:

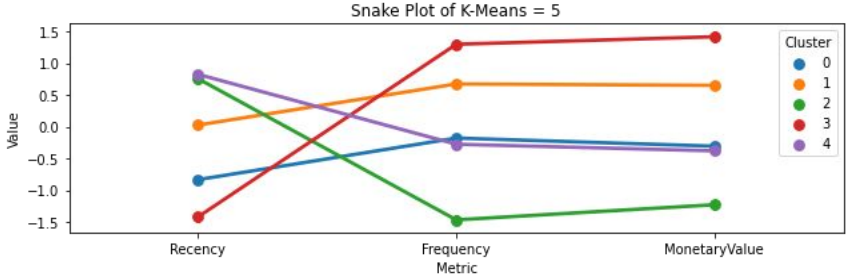

Flattened Graph of 3 Clusters
0
1
2


Flattened Graph of 4 Clusters
0
1
2
3


Flattened Graph of 5 Clusters
0
1
2
3
4


Snake Plot of K-Means = 3
Value
Recency
Frequency
MonetaryValue
Metric
Cluster
0
1
2


Snake Plot of K-Means = 4
Value
Recency
Frequency
MonetaryValue
Metric
Cluster
0
1
2
3


Snake Plot of K-Means = 5
Value
Recency
Frequency
MonetaryValue
Metric
Cluster
0
1
2
3
4

Classification models used for prediction:

1.Logistic Regression:
2. Random Forest:
3. XGBoost:

	Model_Name	Train ROC AUC score	Test ROC AUC score	Train Accuracy score	Test Accuracy score
0	Logistic Regression	0.981703	0.979816	0.92	0.91
1	Random Forest	0.999974	0.998822	1.00	0.98
2	XGBoost	0.999998	0.998982	1.00	0.97

Random Forest have performed really well and got the best scores with Random Forest as compared to other Models, so I conclude Random Forest is my optimal model for use and we can use this model for further in online retail customer segmentation.

Challenges:

- Execution takes time.
- As there were many null values present in data set it took time to clean the dataset.
- Difficulty in selecting the appropriate graph for trend

Conclusion:

- Sales are very high in October, November & December as compared to other months.
- We have sales data of only december month from year 2010.
- The Retail Store is Closed on Saturday as per the observation.
- Maximum number of sales are happening on Tuesday, Wednesday and Thursday in ascending order respectively.
- The highest Revenue was generated on Thursday and the Lowest Revenue was generated on Sunday.
- It can be seen that the country which have purchased more number of items as compared to other countries is United Kingdom and the country which have purchased least items is Singapore.
- We can see that United Kingdom is at top in the list of top 10 purchasing countries and Japan is at the bottom in the list.
- It can be observed that most sold product is "Paper craft, little birdie" and least sold product is "Lunch Bag Red Retrospot".
- It can be seen that most revenue generated product is "Paper craft, little birdie" and least revenue generated product is "Jumbo Bag Pink Polkadot".
- We can see that we have more number of bronze customers.

- It can be observed from above that we have lost the maximum customers in the month of November or we can also say that we have lost maximum number of customers in the end of the year.
- The count of unique stock unit ids and their descriptions should have matched but they do not. This implies some stock units might have more than one descriptions.
- We noticed in the exploratory data analysis phase that majority of the transactions belonged to UK, so it makes sense to consider only this country data for maximum impact.
- Most of the people are visiting the store in the afternoon as compared to morning and evening time period.
- The total number of sales in December of year 2010 is higher than December of year 2011.
- Management needs to concentrate on the decrease of sales in December month.
- With logistic regression we got the accuracy score of 0.91 on train data and 0.92 on test data.
- With Random forest classifier we got the train accuracy of 1.00 and test accuracy of 0.98.
- With XGBoost classifier we got the train accuracy score of 1.00 and test accuracy of 0.97

Thank You!!