Capstone Project
Customer Segmentation

-Rahul Deshmukh

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.

Key Steps:

- Defining the problem statement
- Data Cleaning
- EDA and data visualization
- Data preprocessing
- Feature selection
- Preparing Dataset for model
- Applying model
- Model validation and selection

Why is analytics useful for Customer segmentation ?

- To know Recency of customers.
- To know Frequency of customers.
- To know Monetary Value of customers.

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

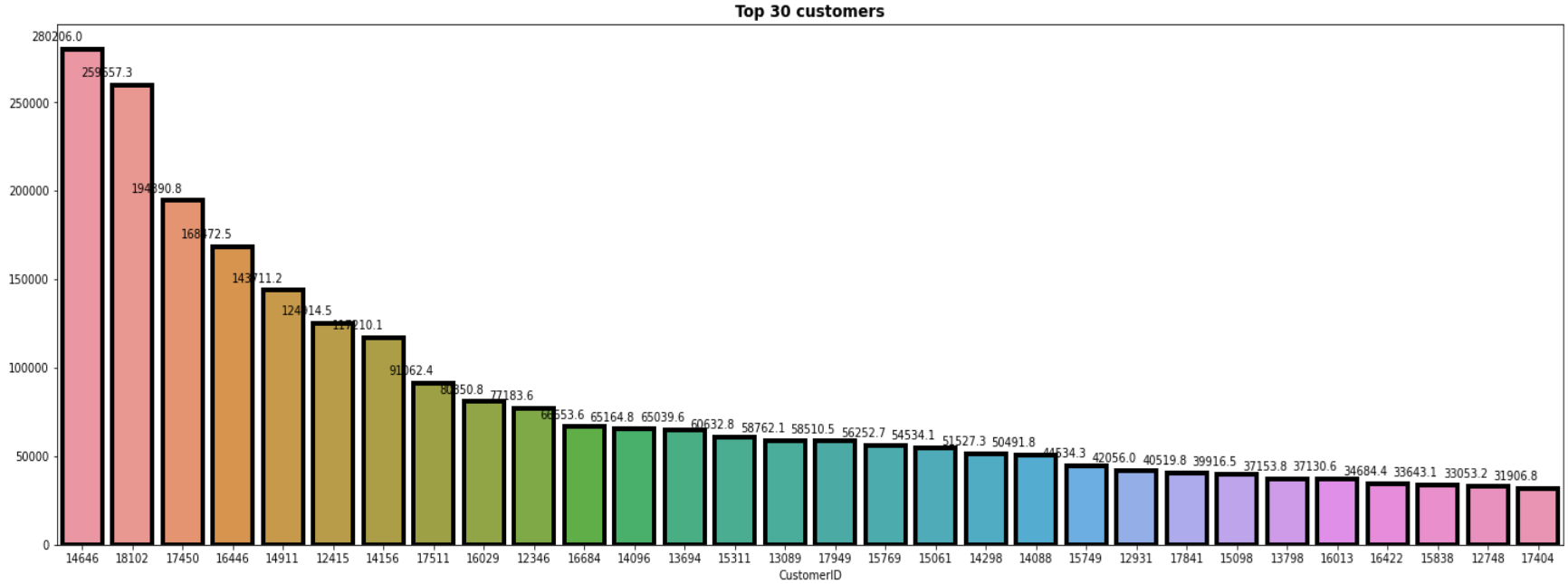
Variable Names:

- InvoiceNo:
- Stock Code:
- Description:
- Quantity:
- InvoiceDate:
- Unit Price:
- CustomerID:
- Country:

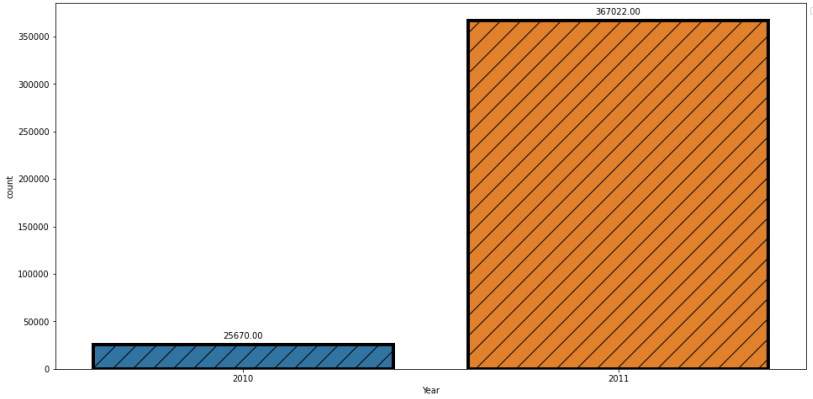
Exploratory Data Analysis:

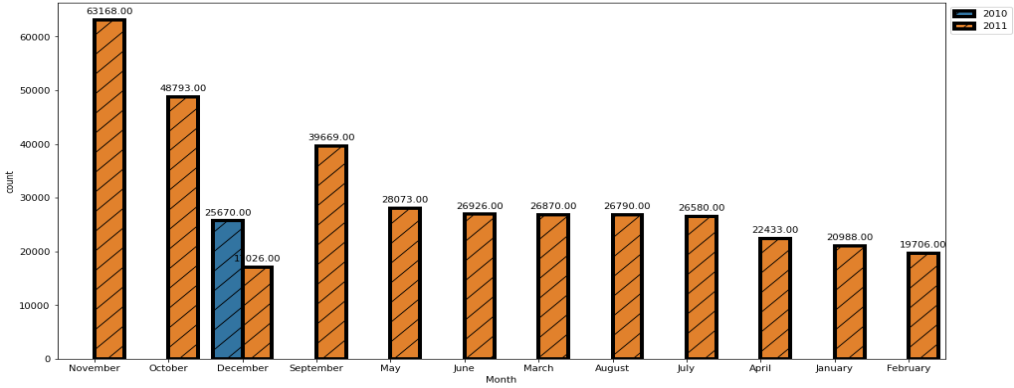
EDA is used for analyzing what the data can tell us before the modeling or by applying any set of instructions/code.

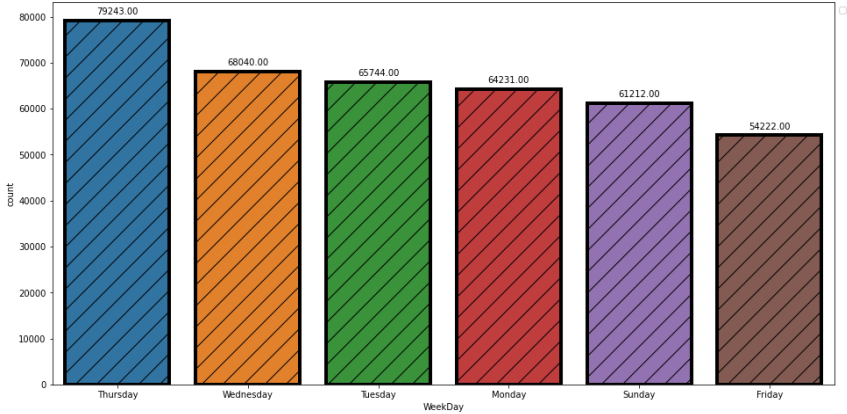
Top 30 customers who visit often and spend well :

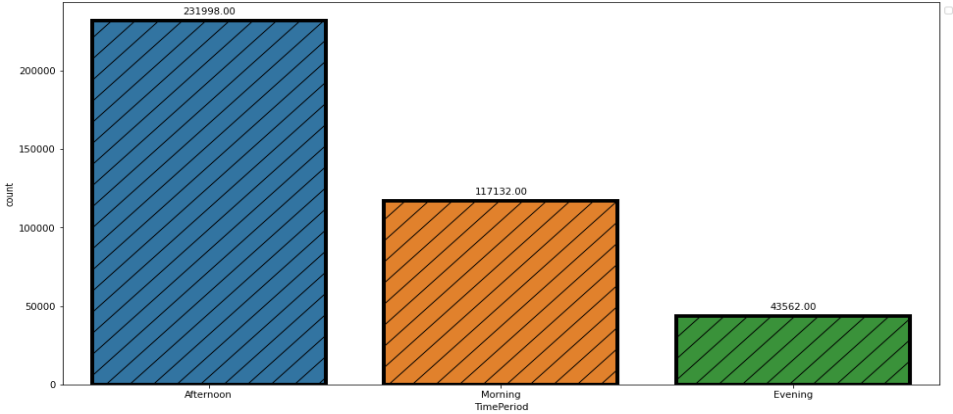


Periodical purchasing stats:

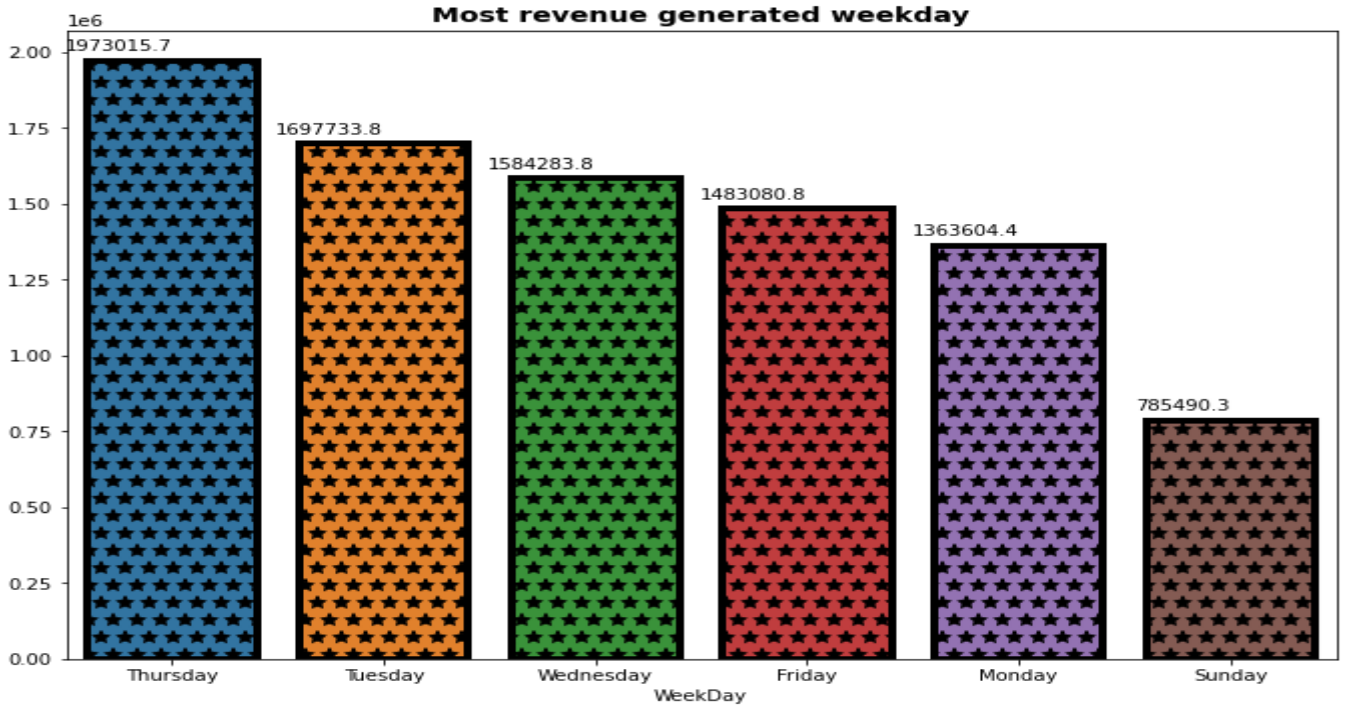




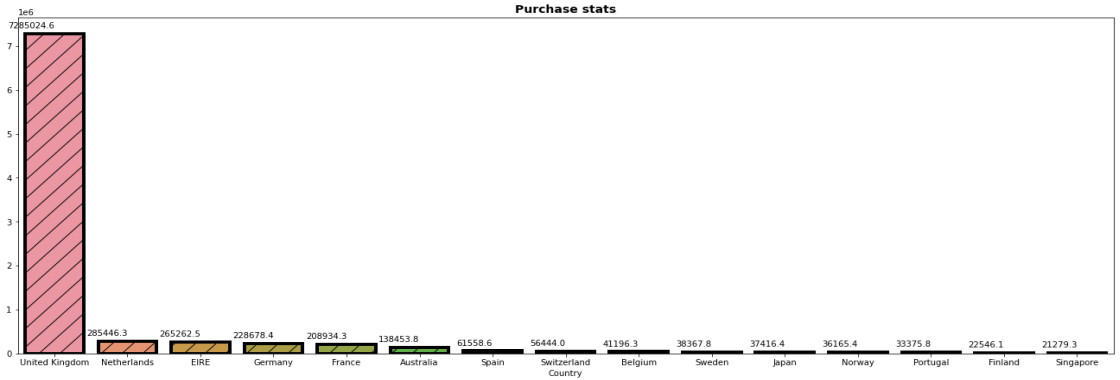


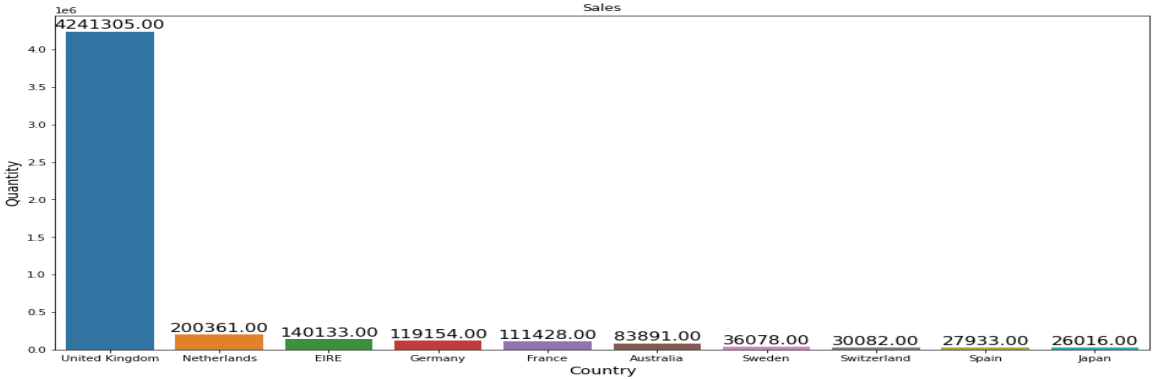


Most revenue generated weekday:



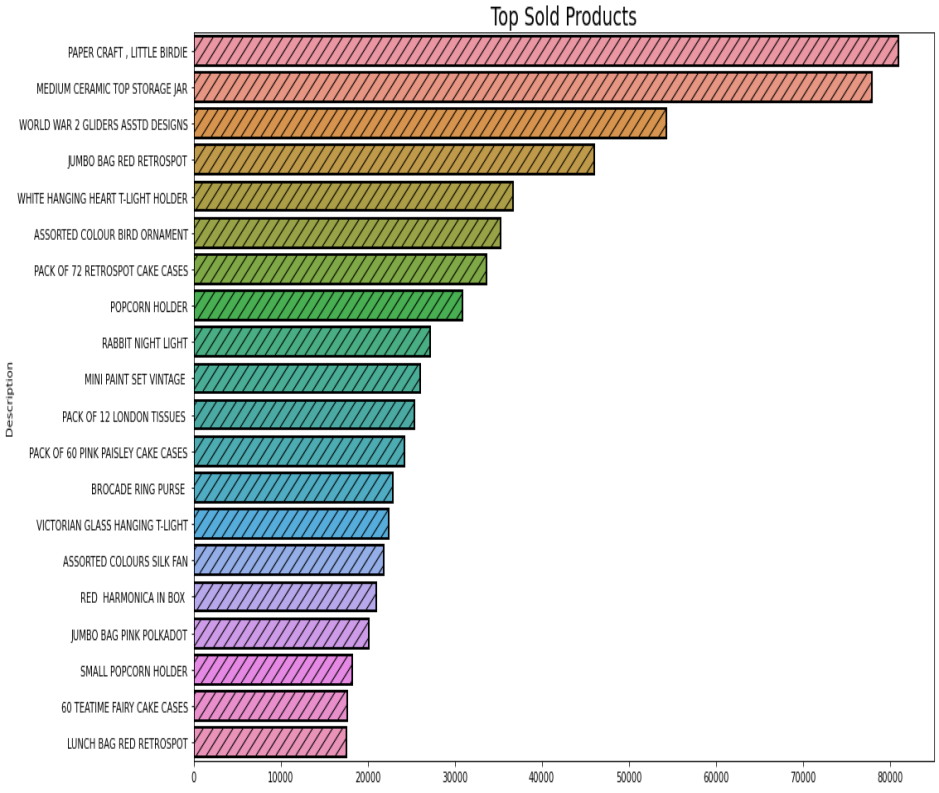
High quantity buying and high purchasing countries stats:

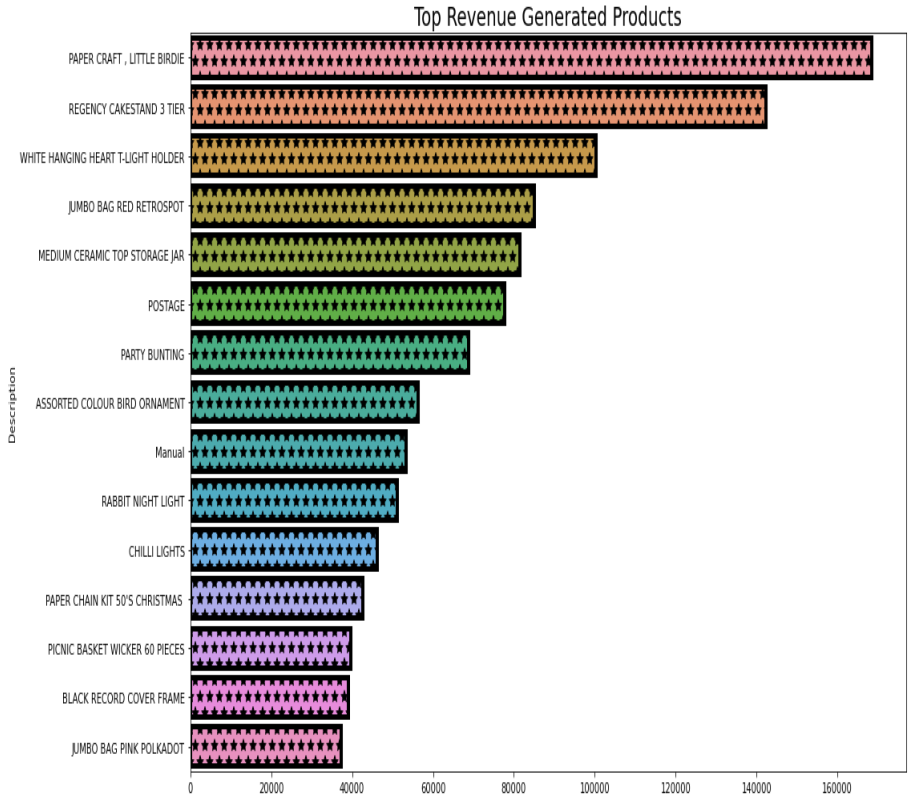




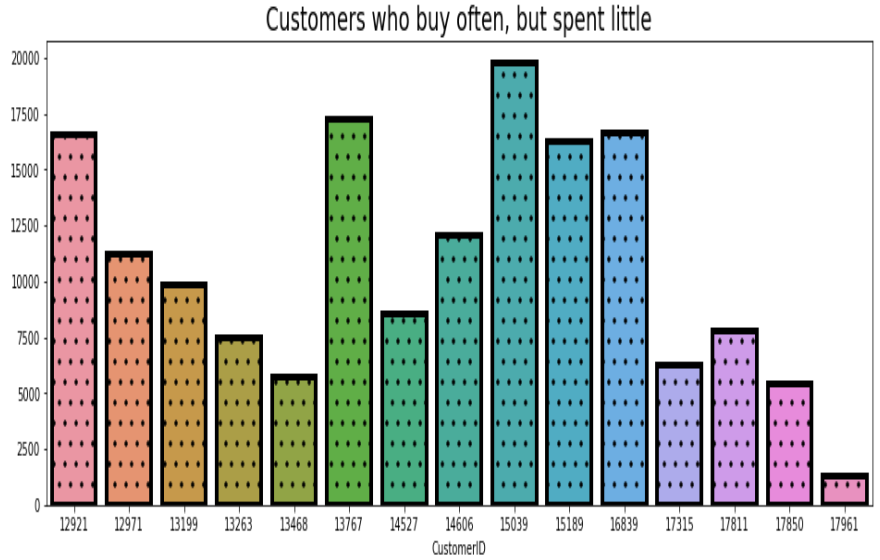
Product Sales Categorization:

Top sold and revenue generated product:

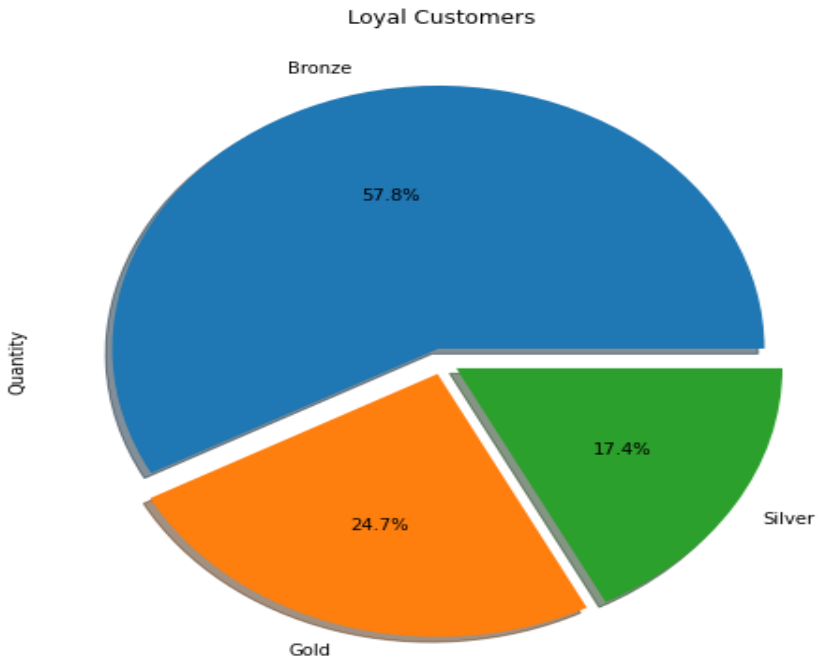


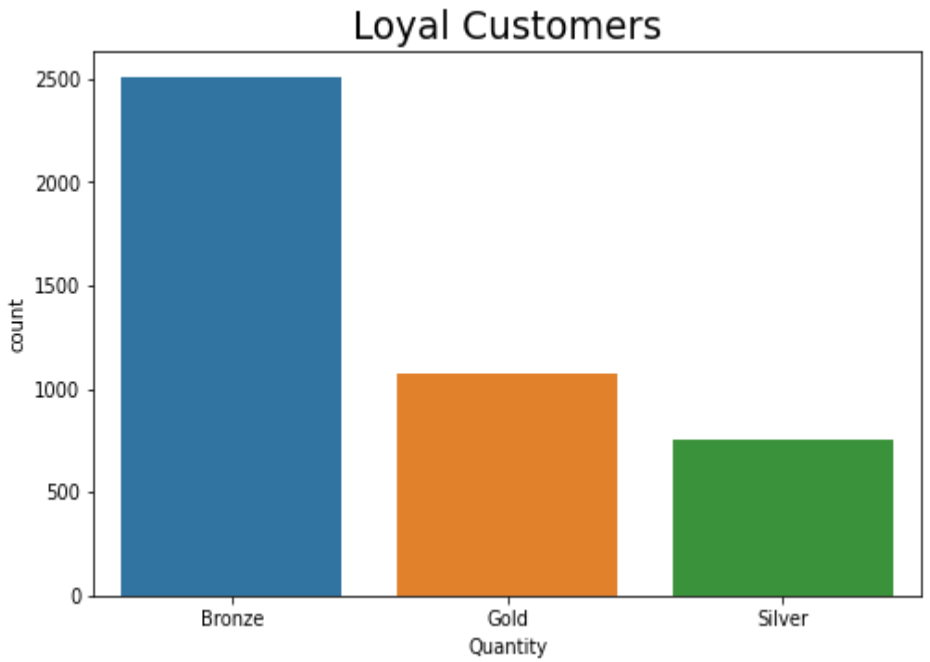


Highest quantity of products purchased by customer and Customers who buy often but spend very little:



Customers category:





Most customers lost in month:

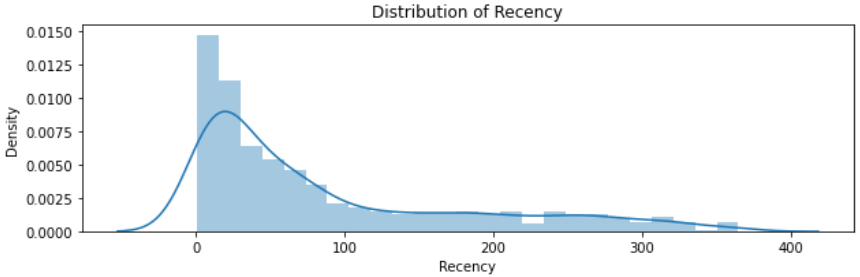
Recency, Frequency and Monetary Value score:

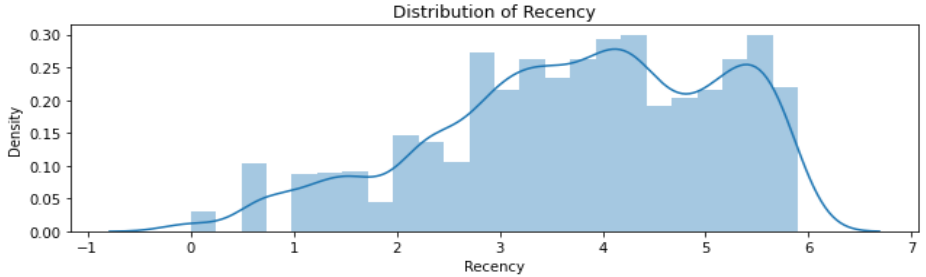
The recency, frequency, monetary value (RFM) model is based on three quantitative factors namely recency, frequency, and monetary value. Each customer is ranked in each of these categories, generally on a scale of 1 to 5 (the higher the number, the better the result). The higher the customer ranking, the more likely it is that they will do business again with a firm. Essentially, the RFM model corroborates the marketing adage that "80% of business comes from 20% of the customers."

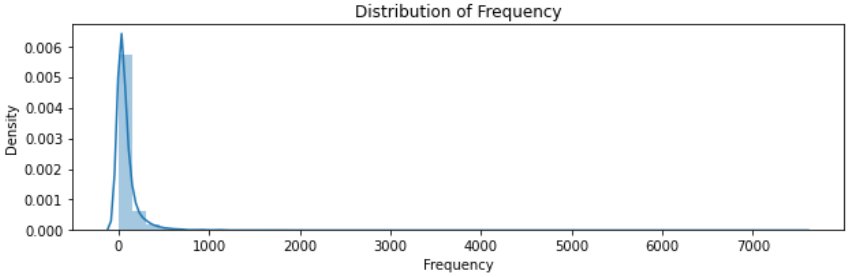
CustomerID	Recency	Frequency	MonetaryValue	Recency_Q	Frequency_Q	MonetaryValue_Q	RFM_Segment	RFM_Score
12346	326	1	77183.60	1	1	4	1.01.04.0	6
12747	3	96	3837.45	4	3	4	4.03.04.0	11
12748	1	4054	31081.74	4	4	4	4.04.04.0	12
12749	4	199	4090.88	4	4	4	4.04.04.0	12
12820	4	59	942.34	4	3	3	4.03.03.0	10

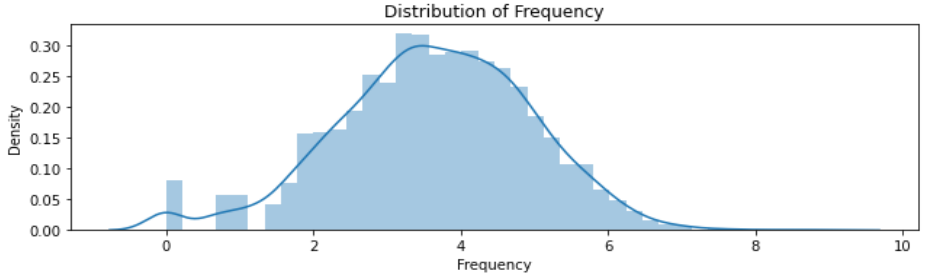
	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
General_Segment				
1.Gold	26.1	182.0	3830.1	1493
2.Silver	95.6	34.0	691.3	1679
3.Bronze	204.8	10.9	188.4	682

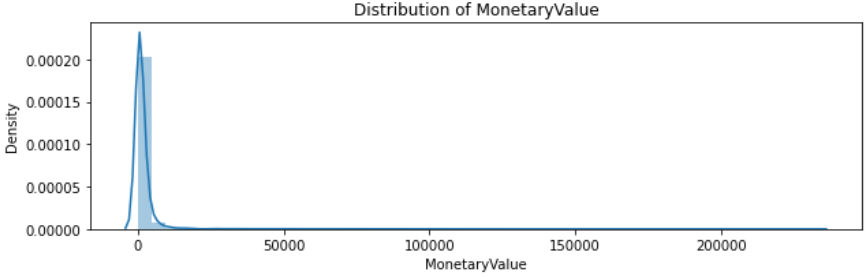
Before and after log transformation:

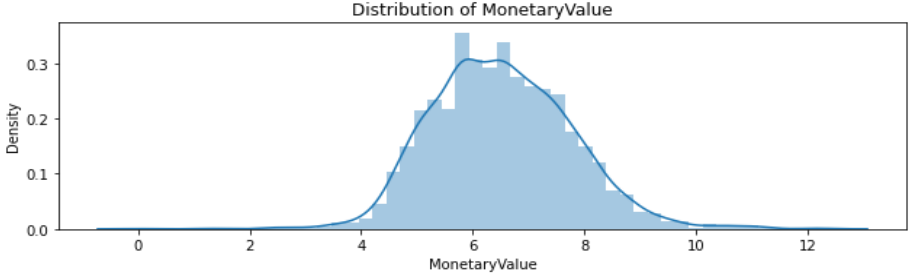




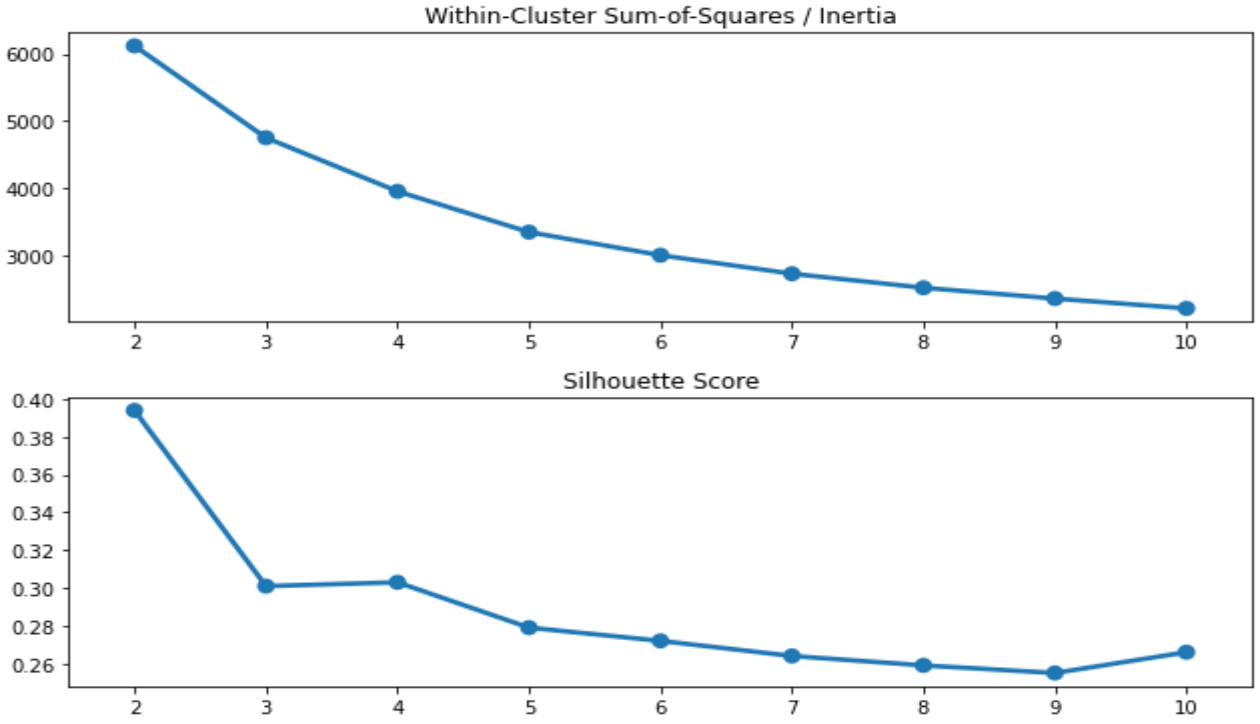






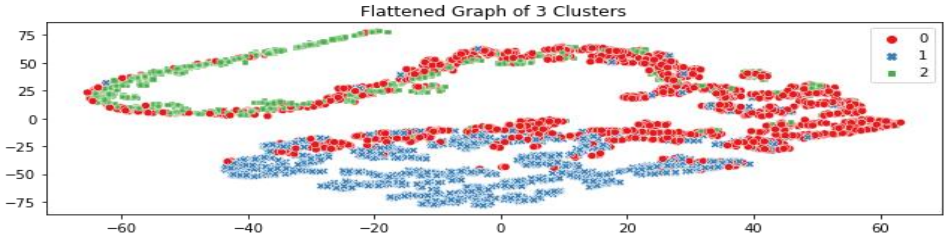


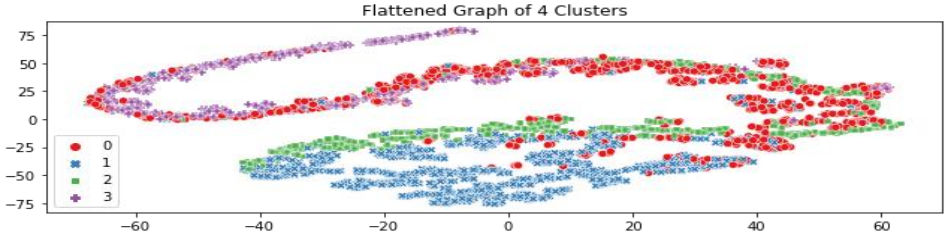
Finding Optimal Cluster:

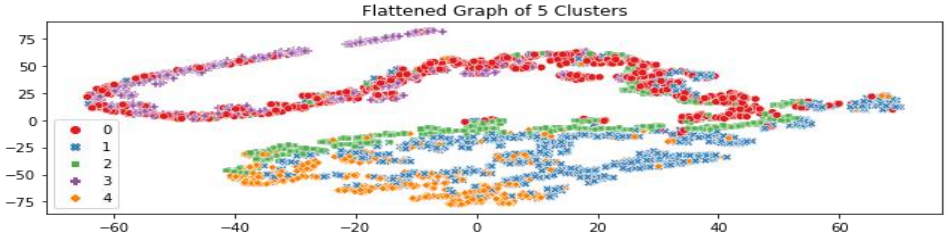


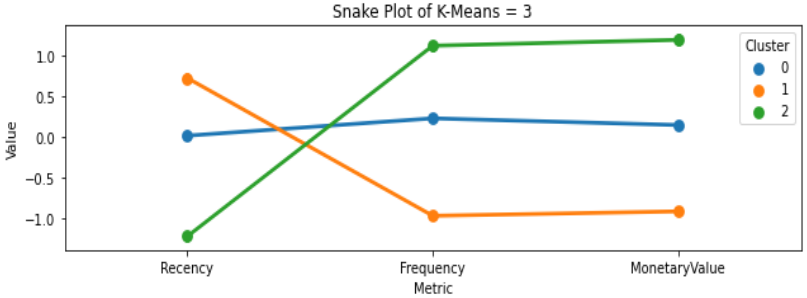
Model used:

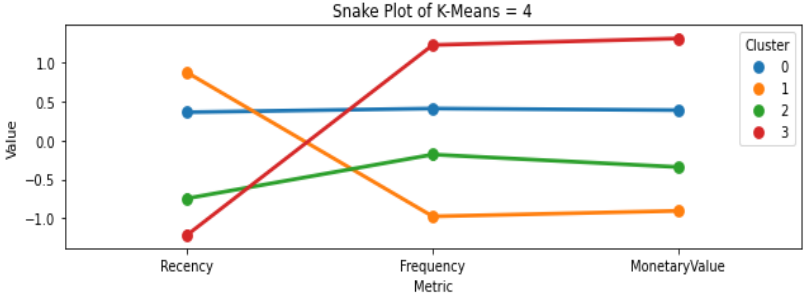
★ KMeans Clustering:

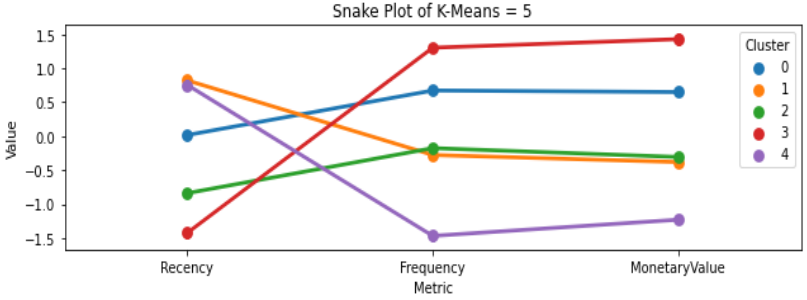












Classification models used for prediction:

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

Observation:

I've used logistic Regression at first but the score was bit less, after that I've used Tree based algorithm i.e. Random forest and XGBoost and with Random forest I got highest score as compared to other two so, random forest is my optimal model which can be used for further.

Evaluation Matrix:

	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

Challenges:

- Loading dataset 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.

Summary of conclusion:

The customer segments thus deduced can be very useful in targeted marketing, scouting for new customers and ultimately revenue growth. After knowing the types of customers, it depends upon the retailer policy whether to chase the high value customers and offer them better service and discounts or try and encourage low/ medium value customers to shop more frequently or of higher monetary values.