Capstone Project - 4

UnSupervised ML - Clustering

Topic - Customer Segmentation

By - Avilash Srivastava

CONTENTS OF THE PRESENTATION

- Problem Statement
- Data Summary
- Data wrangling
- Exploratory data analysis
- RFM model
- Clustering algorithm
- Model Explanation
- Conclusion

Problem Statement

- 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 Summary

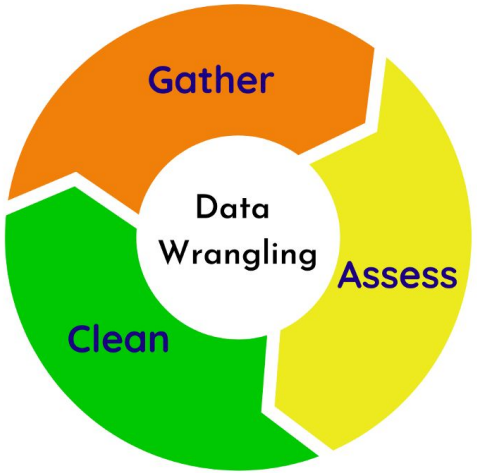
The contents of the data had features such as:

- 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.

Total Rows : 541909

Data Wrangling

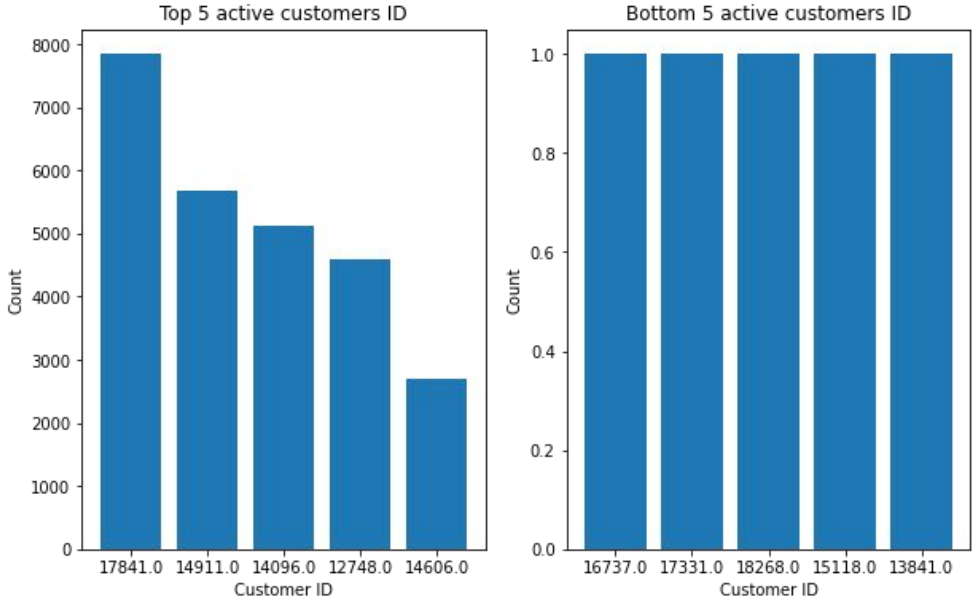
- Removed null values
- Removed duplicates
- Removed cancelled orders.
- Added new features from datetime column such as months, days, hours.
- Added Total Amount
- Converted datatypes



Exploratory Data Analysis

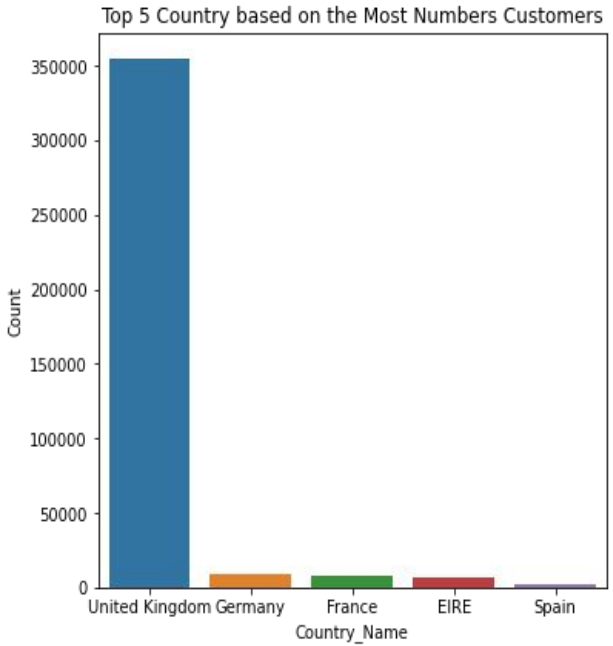
First analysis on customers

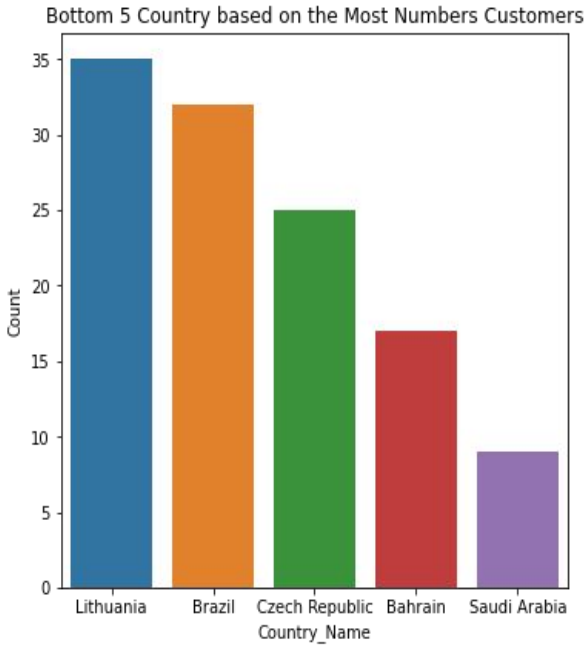
- 4339 unique customer IDs.
- Id 17841 was the most active customer
- Id 13841 was least active



EDA - Country

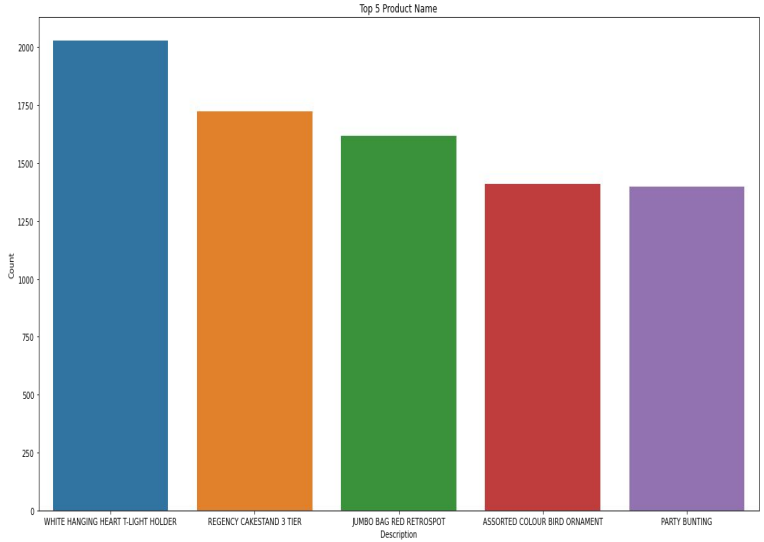
- UK, Germany, France were top countries having more no. of customers.
- Saudi Arabia, Bahrain were least.
- Since data belonged to UK based company, UK had majority of customers

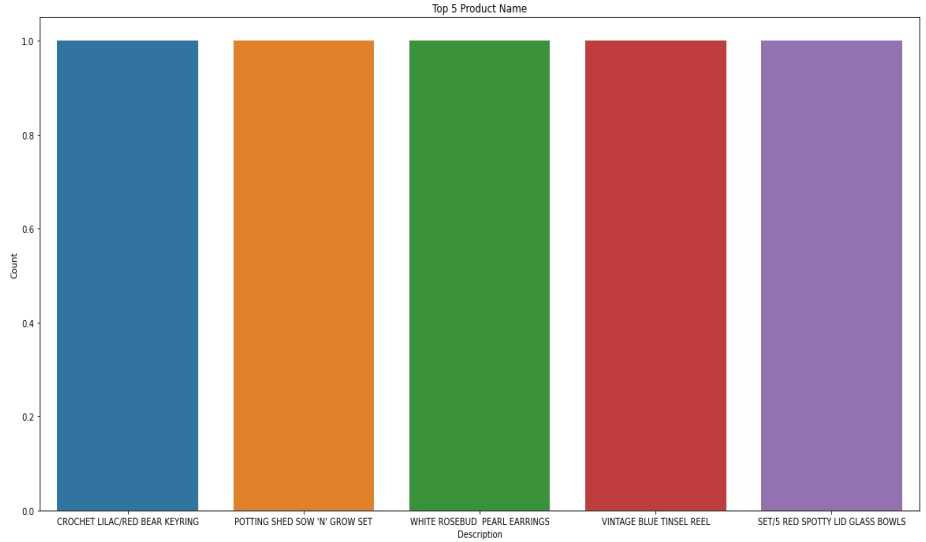




EDA - Products

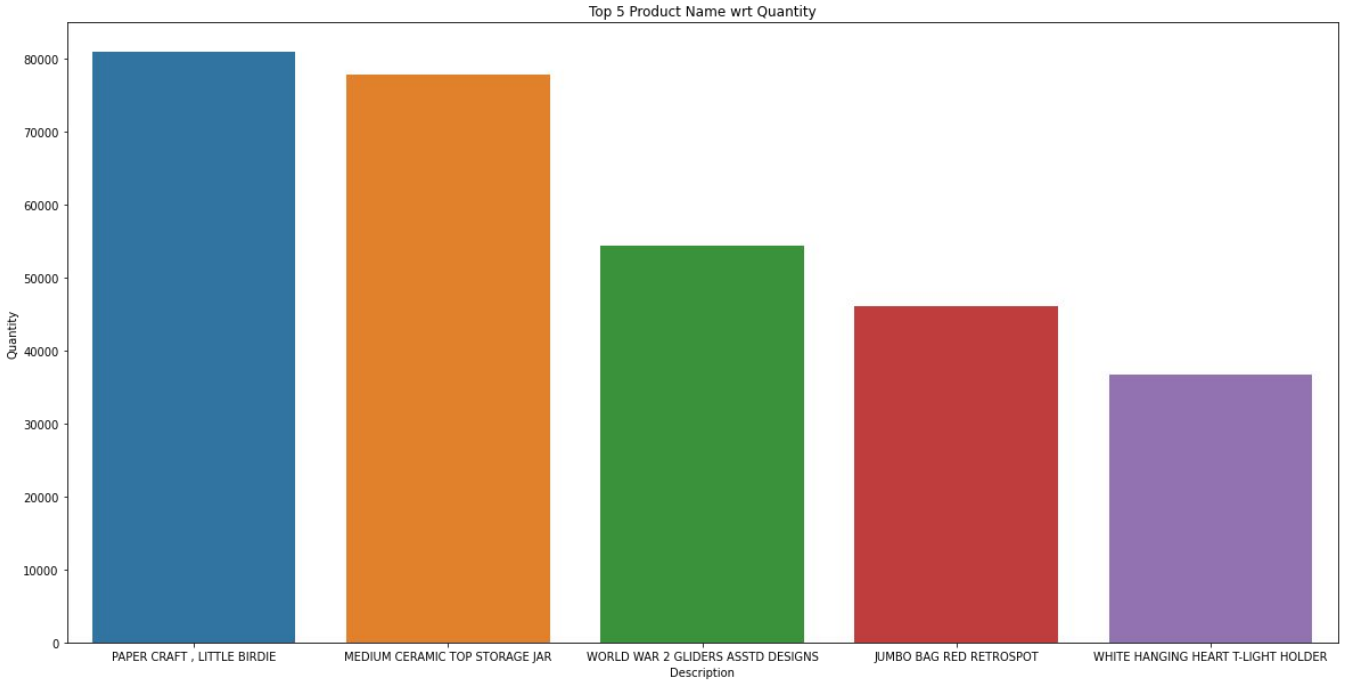
Top and bottom sold products based on count





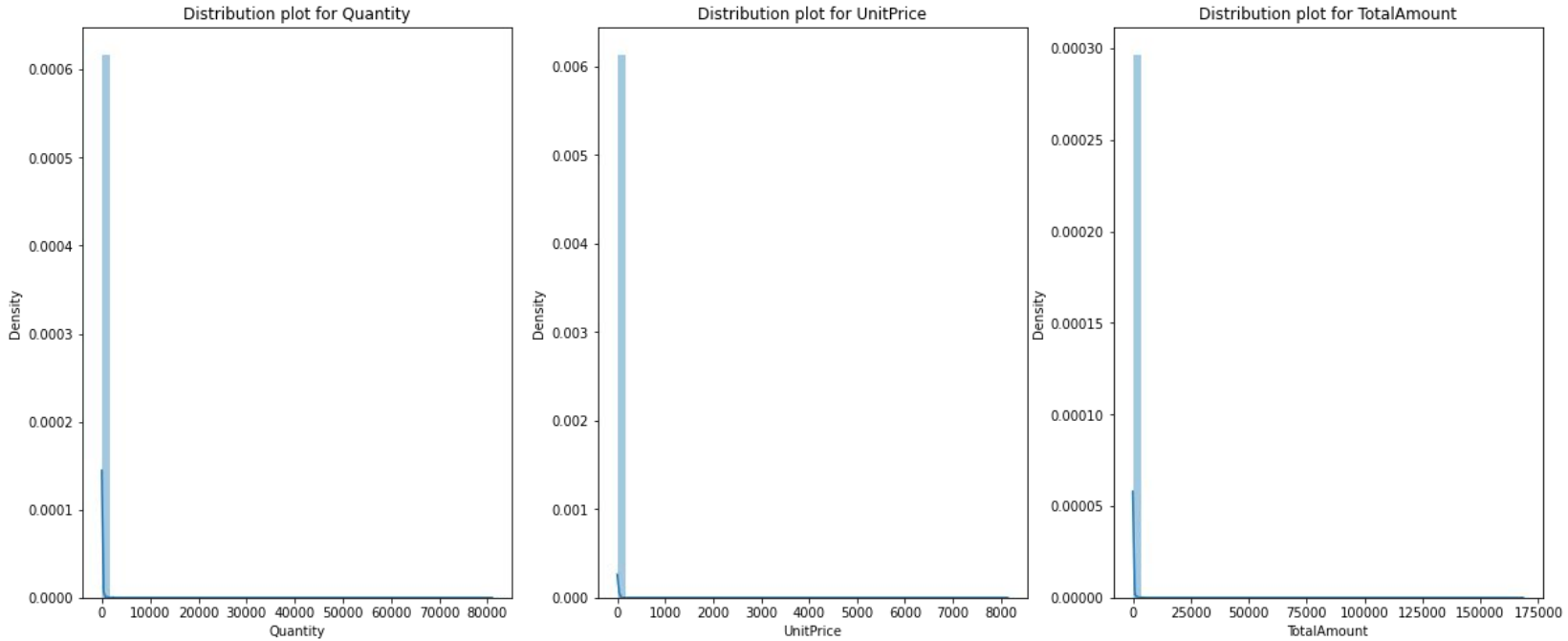
EDA - Continued

Top 5 sold products based on quantity.



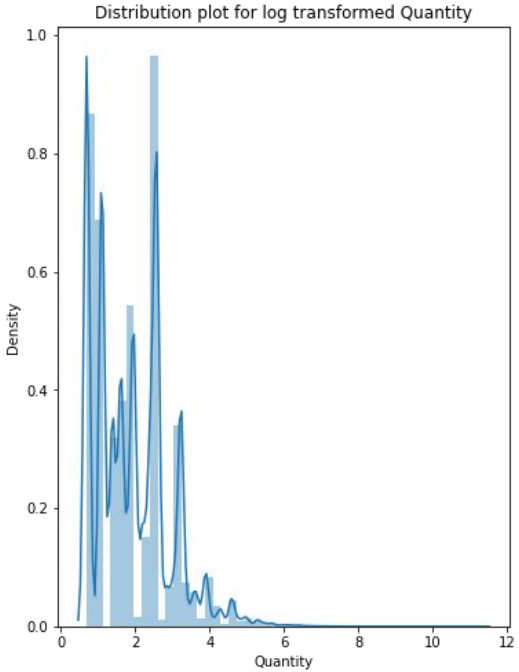
EDA - Numerical Variables

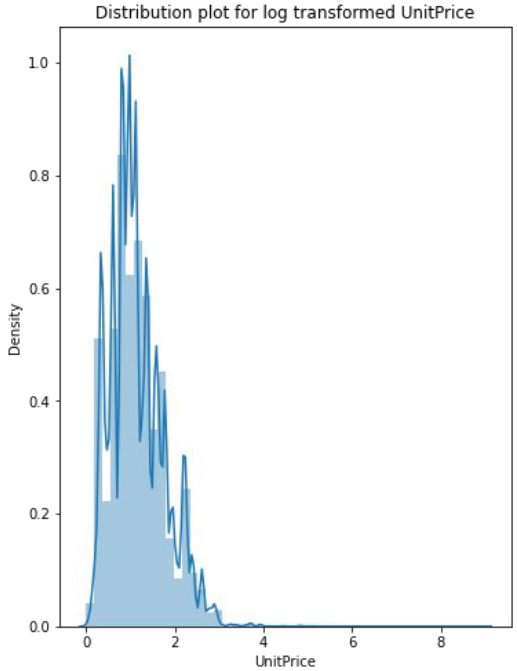
Highly positively skewed

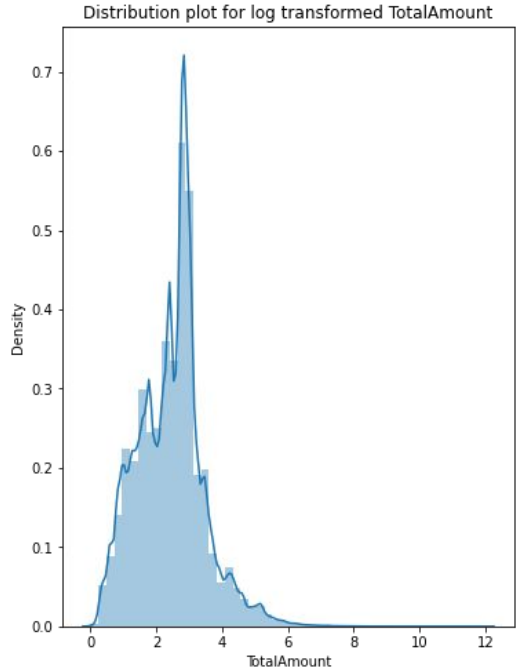


EDA - Continued

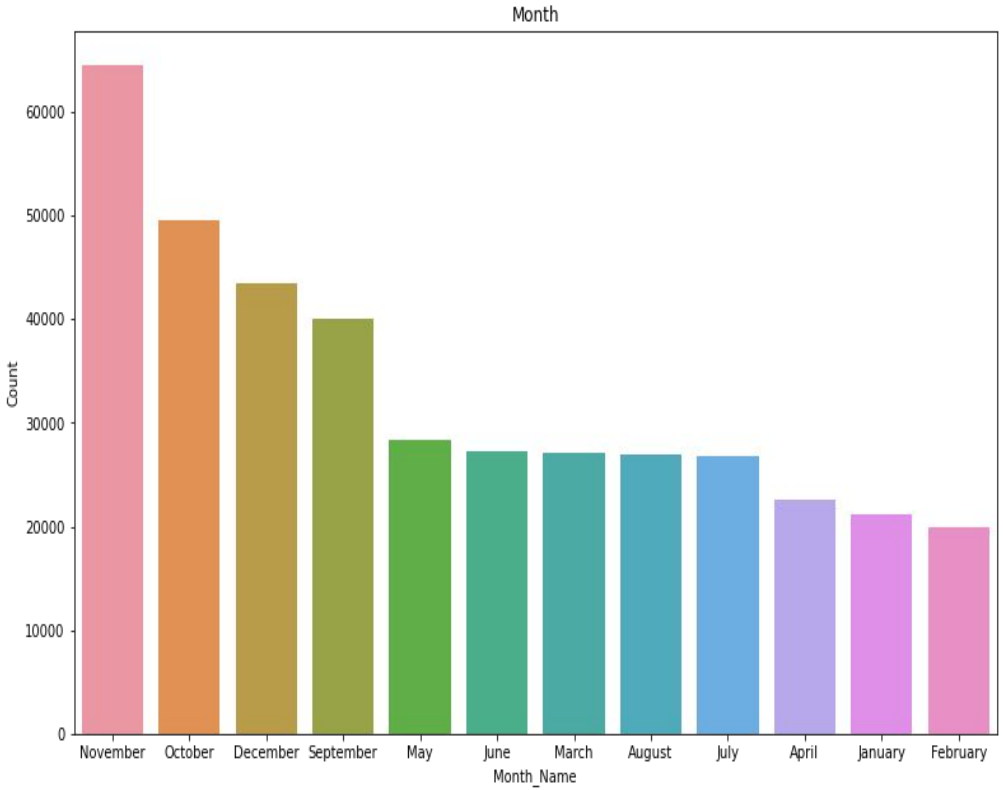
Log transformed

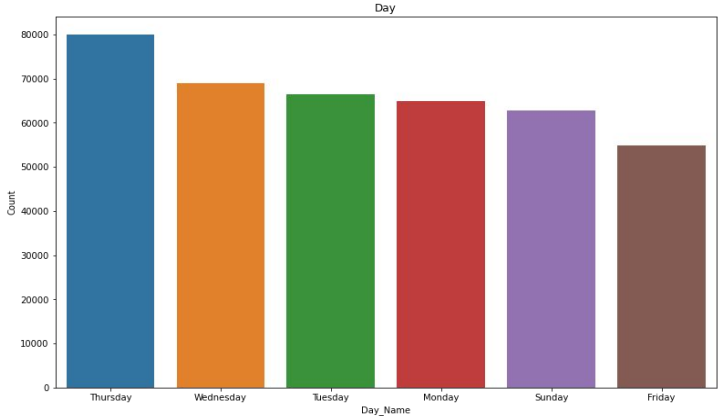


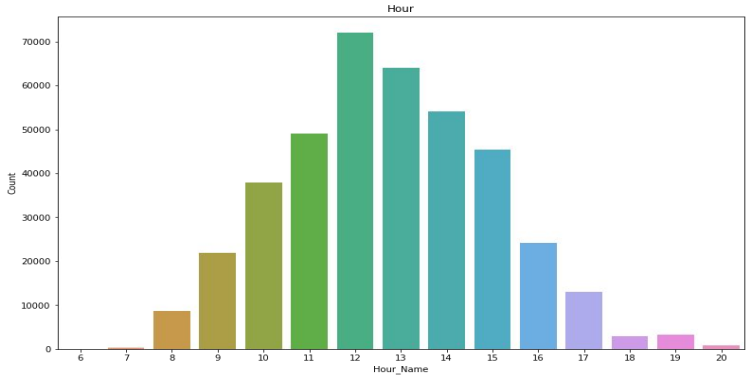




EDA - Months, Days and Hours

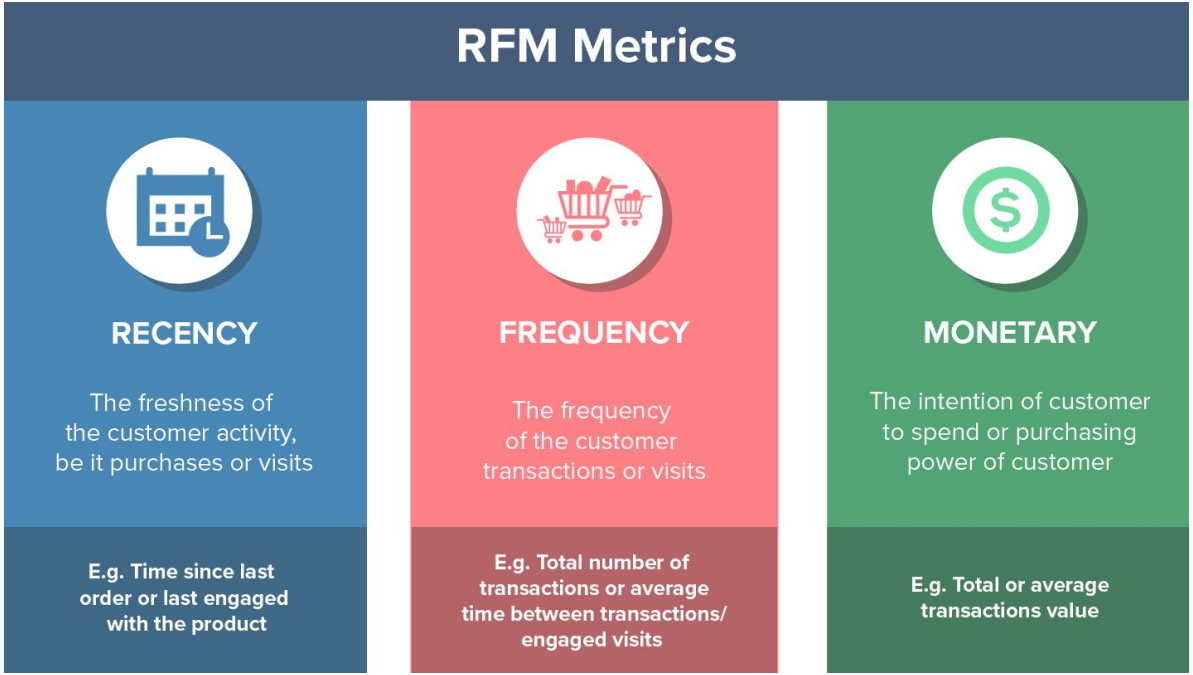






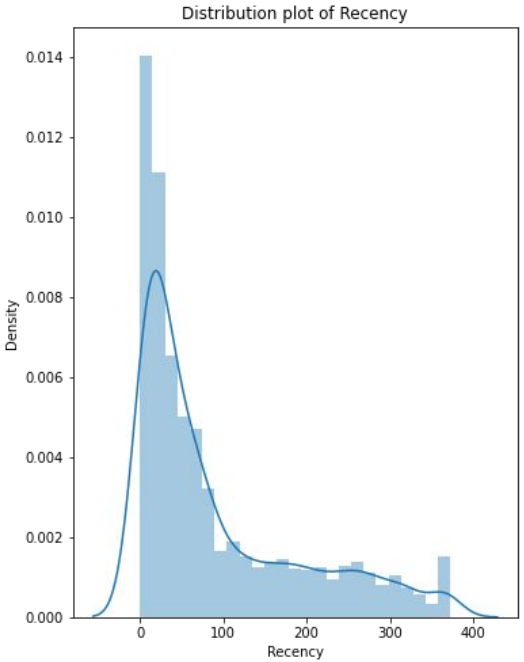
RFM model

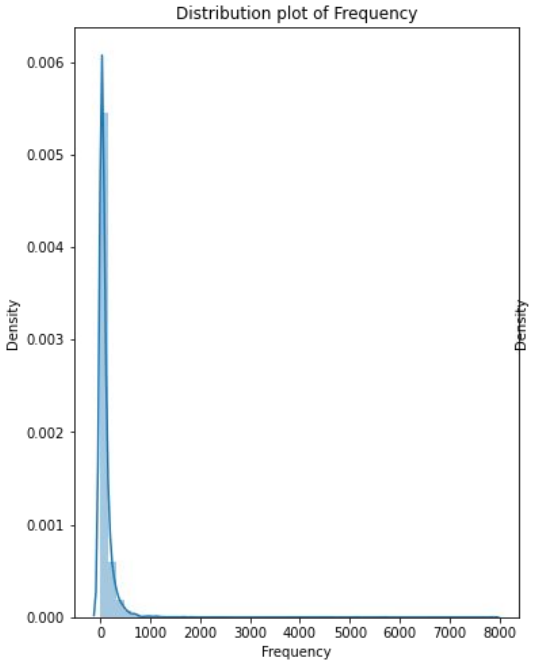
Created features such as recency, frequency and monetary

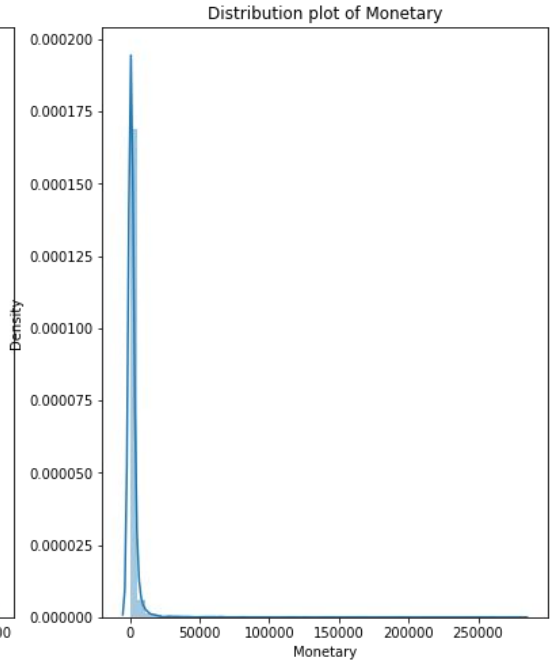


RFM model - Continued

Highly positively skewed

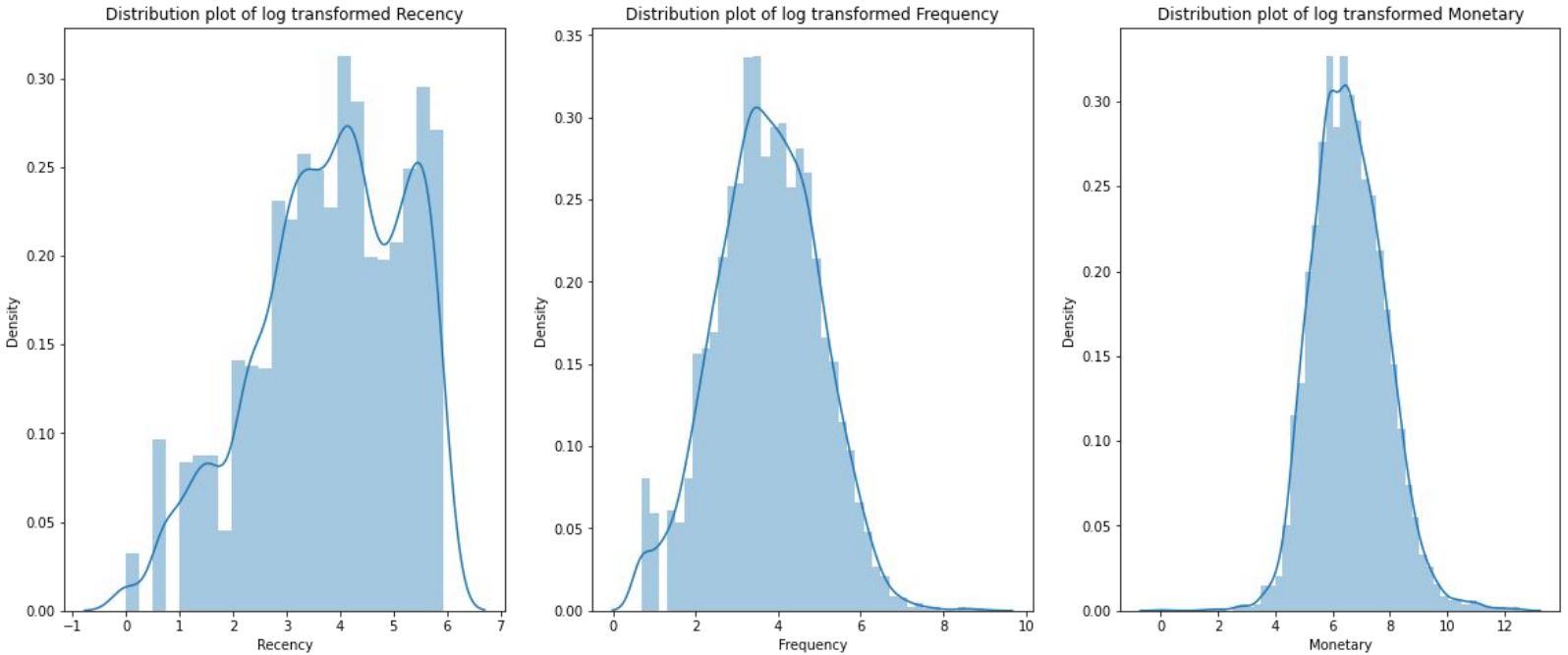






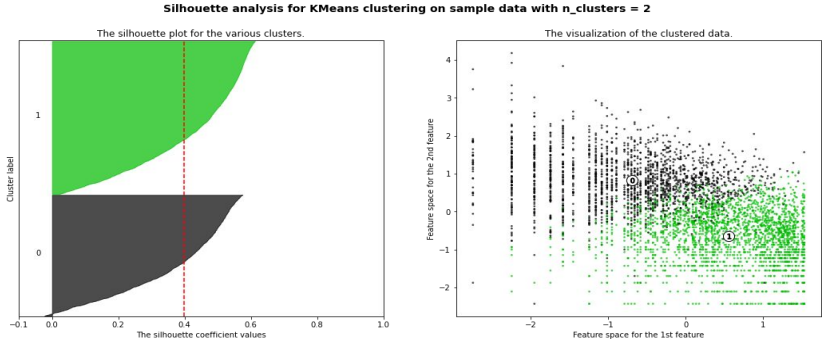
RFM model - Continued

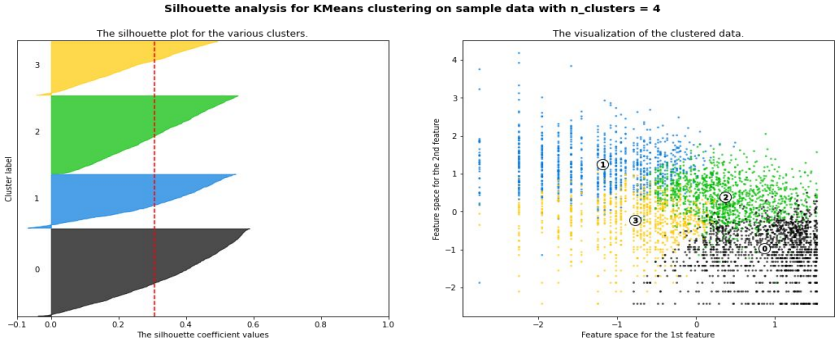
Log transformed

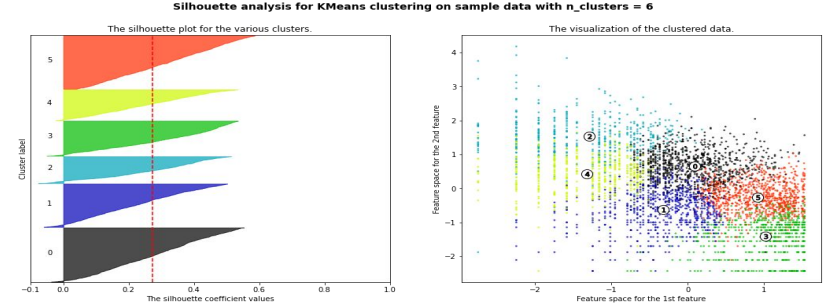


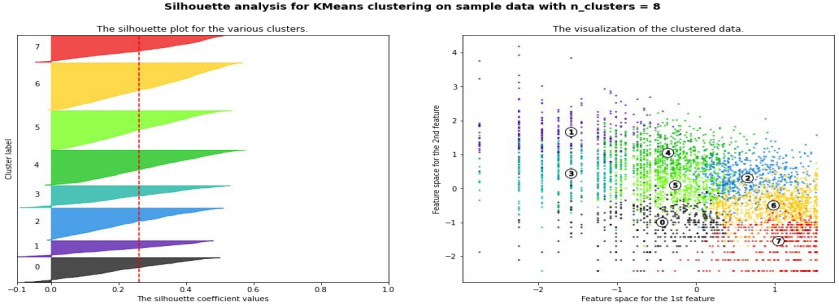
Clustering algorithm - Silhouette score analysis

- KMeans clustering algorithm used on RFM features
- N_cluster = 2 gave the highest silhouette score



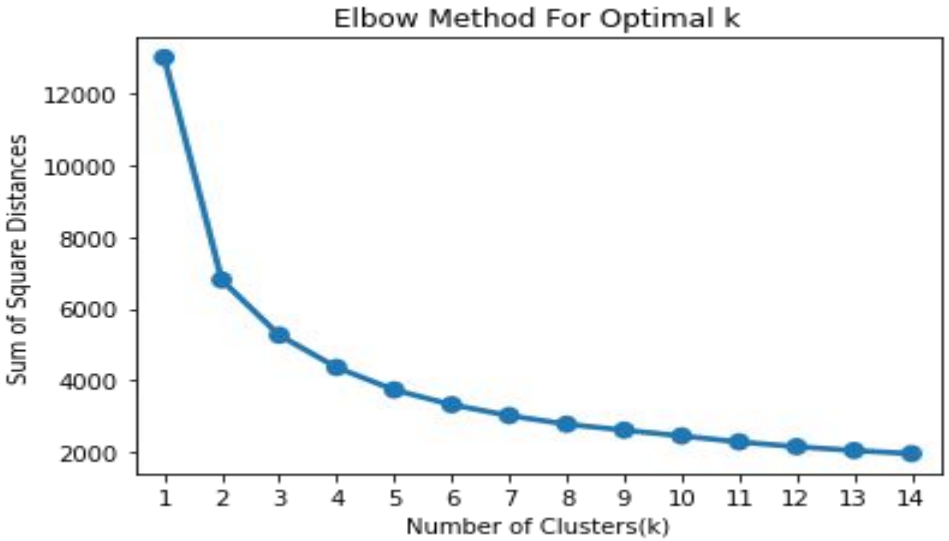




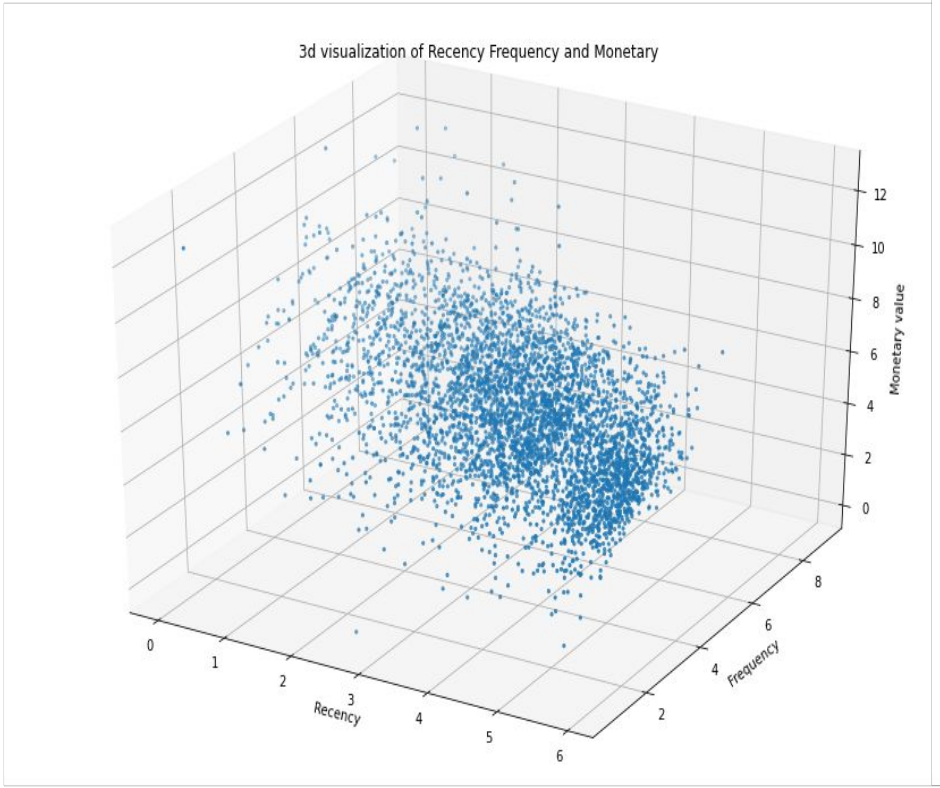


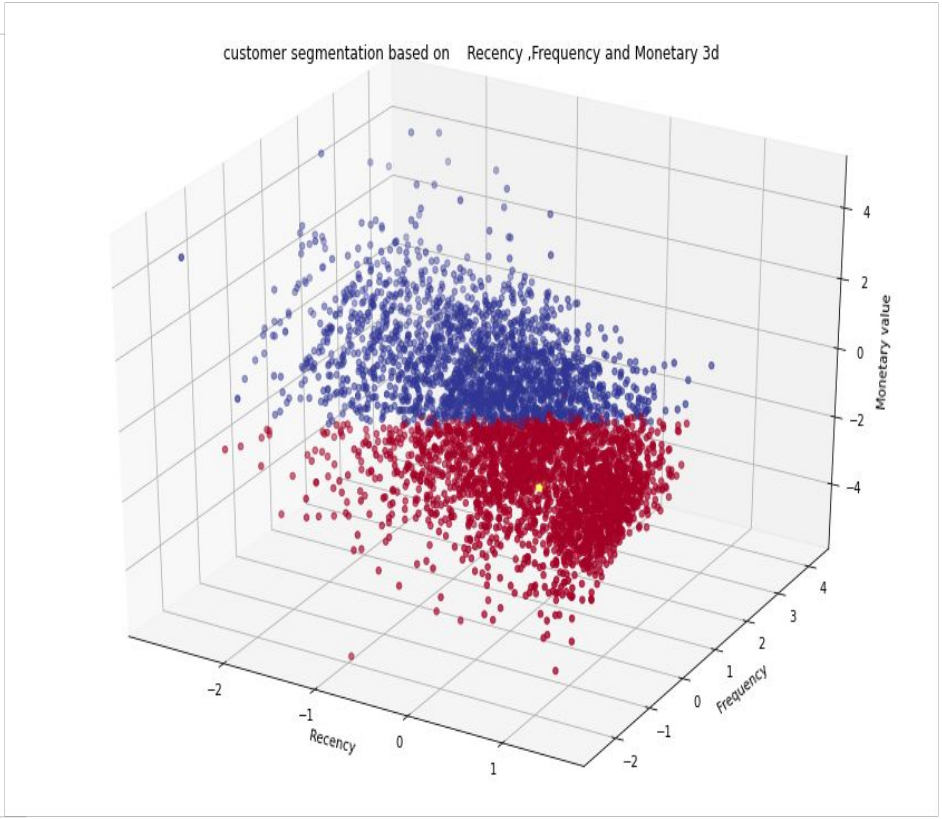
Elbow - method

Line starts decreasing abruptly from 2 and 3 clusters



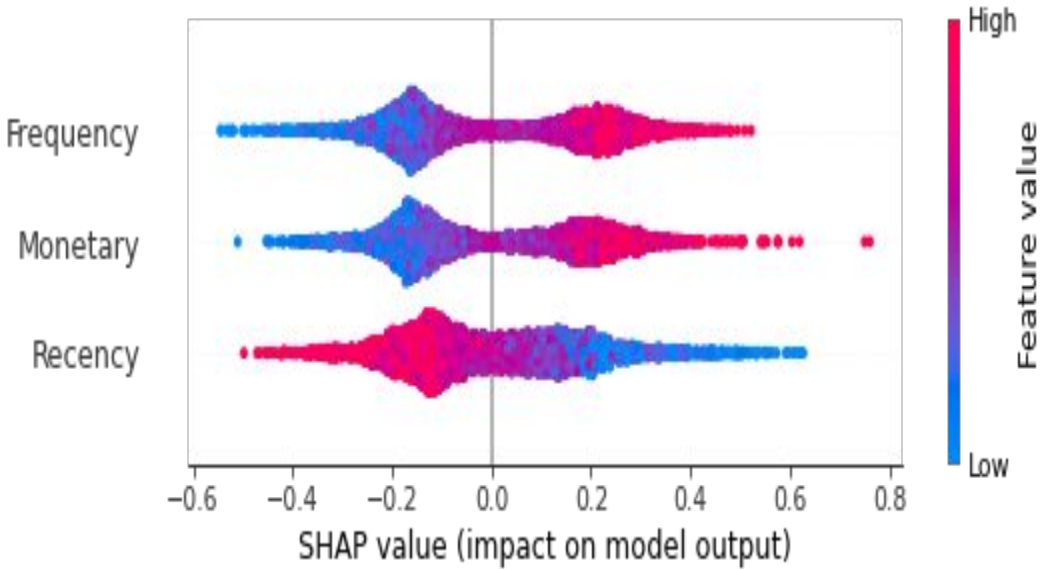
KMeans

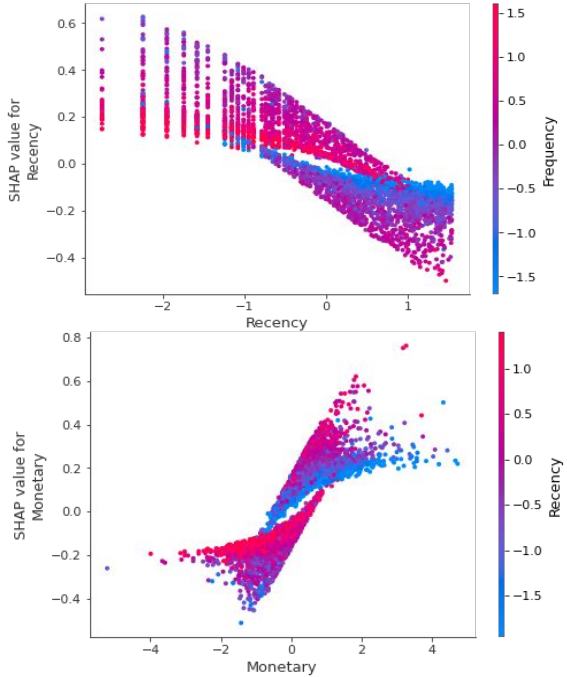

3d visualization of Recency Frequency and Monetary
Recency
Frequency
Monetary value


customer segmentation based on Recency ,Frequency and Monetary 3d
Recency
Frequency
Monetary value

Model Explainability

1. Higher values of frequency, monetary and low values of recency are making the model to predict 1.
2. Low values of frequency, monetary and high values of recency are making the model to predict 0.





Conclusion

- Throughout the analysis we went through various steps to perform customer segmentation. We started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications. Next we did some exploratory data analysis and tried to draw observations from the features we had in the dataset.
- Next we formulated some quantitative factors such as recency, frequency and monetary known as rfm model for each of the customers. We implemented KMeans clustering algorithm on these features. We also performed silhouette and elbow method analysis to determine the optimal no. of clusters which was 2. We saw customers having high recency and low frequency and monetary values were part of one cluster and customers having low recency and high frequency, monetary values were part of another cluster.
- We also implemented shap techniques to understand what is going on inside our model. We saw higher values of frequency, monetary and low values of recency is deciding one class and low values of frequency, monetary and high values of recency is deciding other class.