**PROJECT REPORT**

**E Commerce Analytics**

*Submitted towards the partial fulfillment of the criteria for award of KPMG Data Science Prodegree by Imarticus*

***Submitted By:***

*1. Hrithik Shukla*

*2. Anusha Raut*

*3. Priyanshu Tripathi*

*4. Mahesh Lakshminarayanan*

***Course and Batch****: Data Science Prodegree-DSP 50*



**ACKNOWLEDGEMENTS**

We are using this opportunity to express our gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. We are sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have Pooja Pant as our mentor. She has readily shared her immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the DSP program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

**Table of Contents**

1. **INTRODUCTION**
2. **Title & objective of the study**
3. **Need for the study**
4. **Data sources**
5. **Tools and techniques**
6. **DATA PREPARATION AND UNDERSTANDING**
7. **Feature engineering**
8. **Data dictionary**
9. **Data pre-processing**
10. **Exploratory Data Analysis (EDA)**
11. **RFM ANALYSIS**
12. **Recency**
13. **Frequency**
14. **Monetary**
15. **RFM values and RFM Scores**
16. **RFM segments**
17. **EDA on RFM Segments**
18. **K-MEANS CLUSTERING**
19. **Elbow method**
20. **Silhouette Score**
21. **ACCURACY**
22. **BUSINESS INFERENCE**

**ABSTRACT**

The aim of the experiment is to analyze and segment the customers of an e-commerce company by using the RFM approach. This will enable the e-commerce company to optimize their retention and acquisition strategies.

**INTRODUCTION**

**Title** - E Commerce Analytics

**Objective -** The objective of this project is to build a data model which will

* Identify and segment the customers
* Optimize company’s retention and acquisition strategies.

**Need for the Study -**

* Employee segmentation enables the company to optimize their retention and acquisition strategies.
* E-commerce stores that become success stories did a good job of attracting the right clients. Customer segmentation, or segmenting existing consumers based on frequency of purchases, monetary value, and other factors, was one of the methods they used to do this.
* E-commerce shops that created market strategies based on mass marketing quickly discovered the necessity for client segmentation as a way to save money and time in the digital realm.

**Data Sources –**

* We were provided with a dataset along with the problem statement.
* The dataset contained information about the customers and their e-commerce transactions. This data contains the date-time of sale, customer shipping location, and price of single unit from December 2016 to 2017.

**Tools and Techniques–**

* We are coding in Python using Jupiter notebook
* For Exploratory Data Analysis, we used Tableau for plots an getting insights from them.

**DATA PREPARATION AND UNDERSTANDING**

To get deeper understanding of data or to apply different models to the datasets we need to prepare our data first, so that we can extract the data to read, finding actual predictors and perform visualization part also. Here we performed few steps to get the full knowledge

We imported the essential libraries as below

* import pandas as pd
* import seaborn as sns
* import matplotlib.pyplot as plt
* import numpy as np
* from sklearn import tree
* from sklearn import preprocessing
* import datetime as dt
* from sklearn.clusters import KMeans
* import chart\_studio as cs
* import plotly.offline as po
* import plotly. graph\_objs as gobj

**Setting the working directory:**

* We are setting the working directory to our local environment and later importing the dataset.

**Read the datasets:**

* We read our dataset in the object df.

**Summarizing data:**

The **describe()** function computes a summary of statistics pertaining to the continuous Data Frame columns.

This function gives the **mean, std** and **IQR** values. We passed a hyperparameter include=” all” in order to get insights for categorical variables.

**Feature Engineering:**

Feature engineering has two goals primarily:

* Preparing the proper input dataset, compatible with the machine learning algorithm requirements
* Improving the performance of machine learning models

**List of Feature Engineering Techniques**

We are removing the columns which are not important for our E-Commerce analysis. The columns that we didn’t use for our RFM segmentation followed by K means clustering are as follows:

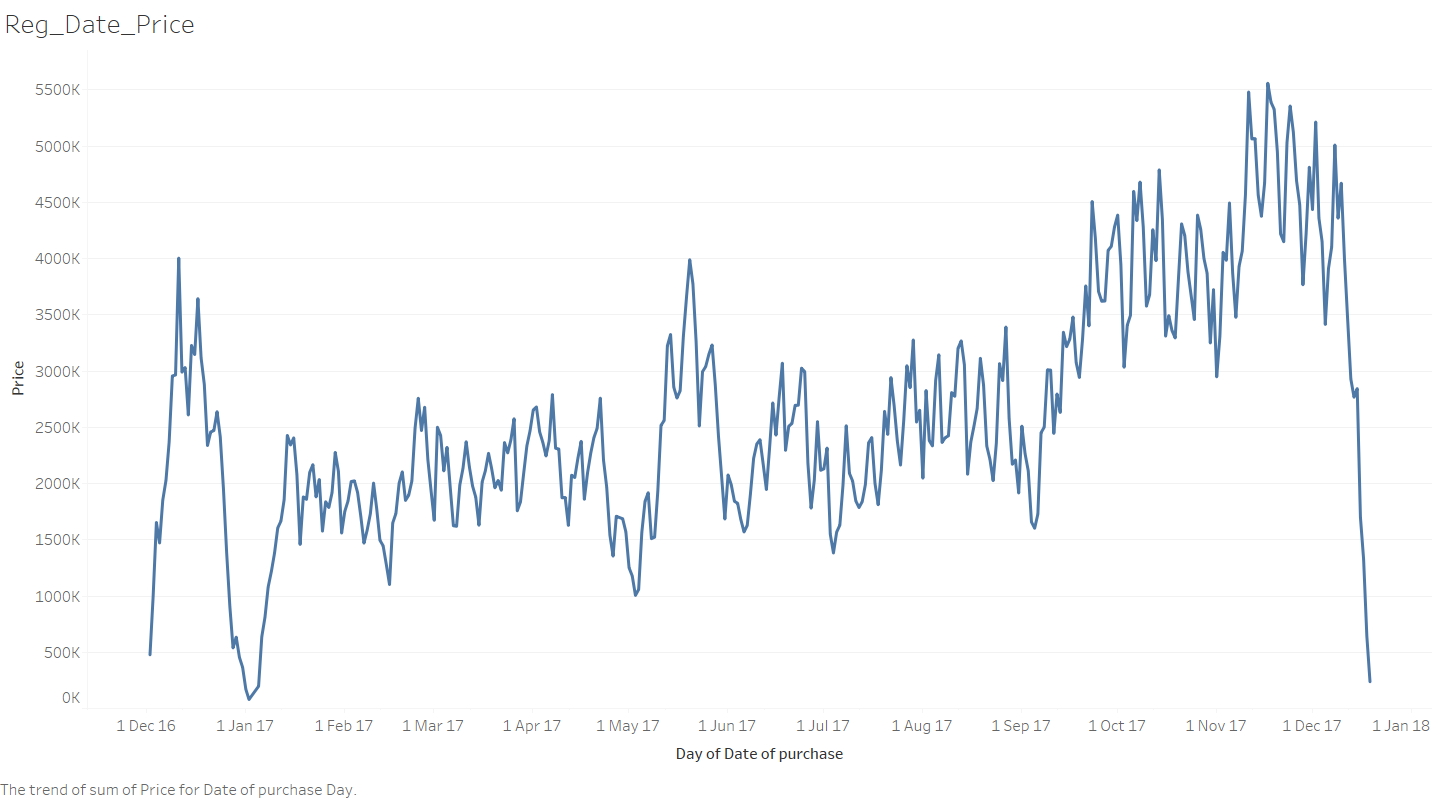
* Item Code
* Reason of return
* Sold as set

**Data dictionary:**

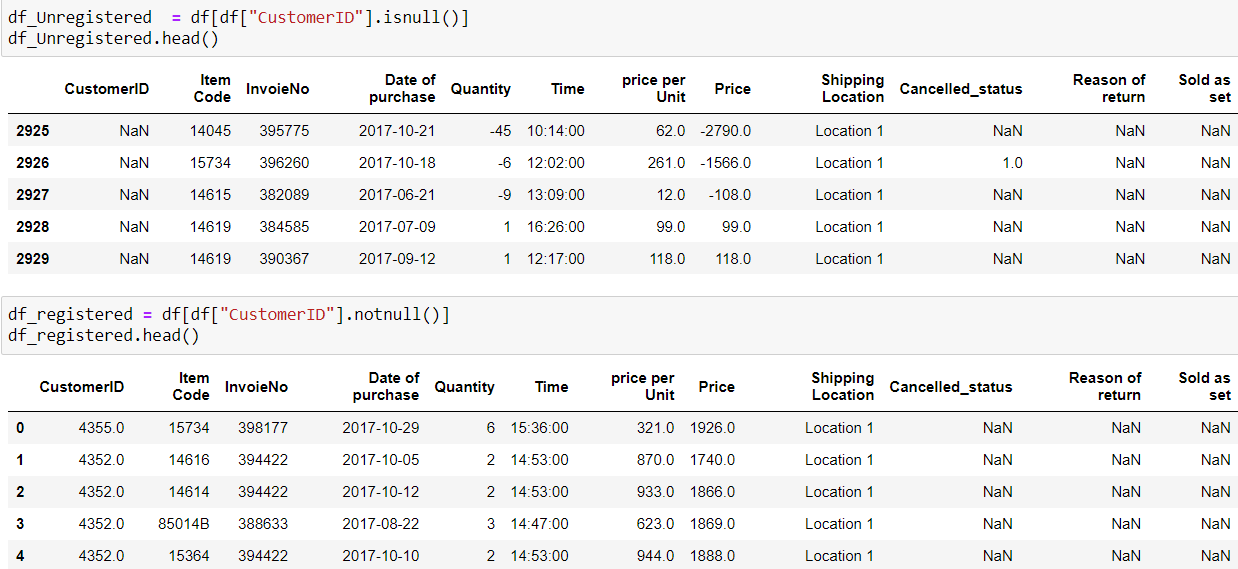
|  |  |
| --- | --- |
| **Column Name** | **Description** |
| **CustomerID** | Unique identifier for each Customer |
| **Item Code** | Unique id for each product |
| **InvoieNo** | Unique id for each purchase |
| **Date of purchase** | Date on which the purchase was made |
| **Quantity** | Number of items bought for each product |
| **Time** | Time at which the purchase was made |
| **price per Unit** | Price of single unit of item purchased |
| **Price** | total purchase price |
| **Shipping Location** | Delivery Location |
| **Cancelled\_status** | Status of Cancellation |
| **Reason of return** | Reason for return of product |
| **Sold as set** | Was the product sold with another product/ Offer |

**Data Pre-processing:**

1] Before starting off with the Exploratory data analysis, we used to tableau to check if the data was complete. We made a line graph using Date of Purchase and Price. The plot depicted that data was from December 2016 to December 2017 without any gaps.

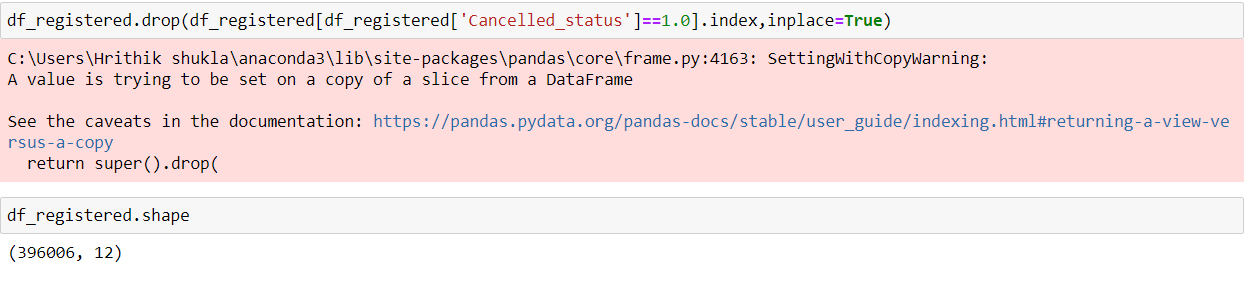


2] Dataset (5,37,979 rows, 12 columns) was mixed as we had missing Customer IDs for some Invoice Numbers. We segregated the dataset on the basis of presence of Customer Id as registered (4,04,189 rows, 12 columns) and non-registered (1,33,790 rows, 12 columns) Data sets.



3] Exploratory Data Analysis [EDA], RFM analysis and K-means clustering is carried out on registered data.

4] After getting some useful insights from the registered data set, we found out negative values for prices. Those observations had cancellation status true, so, we removed those observations and our final registered dataset had 3,96,006 observations. We termed it as Non-Cancelled Registered Data set.

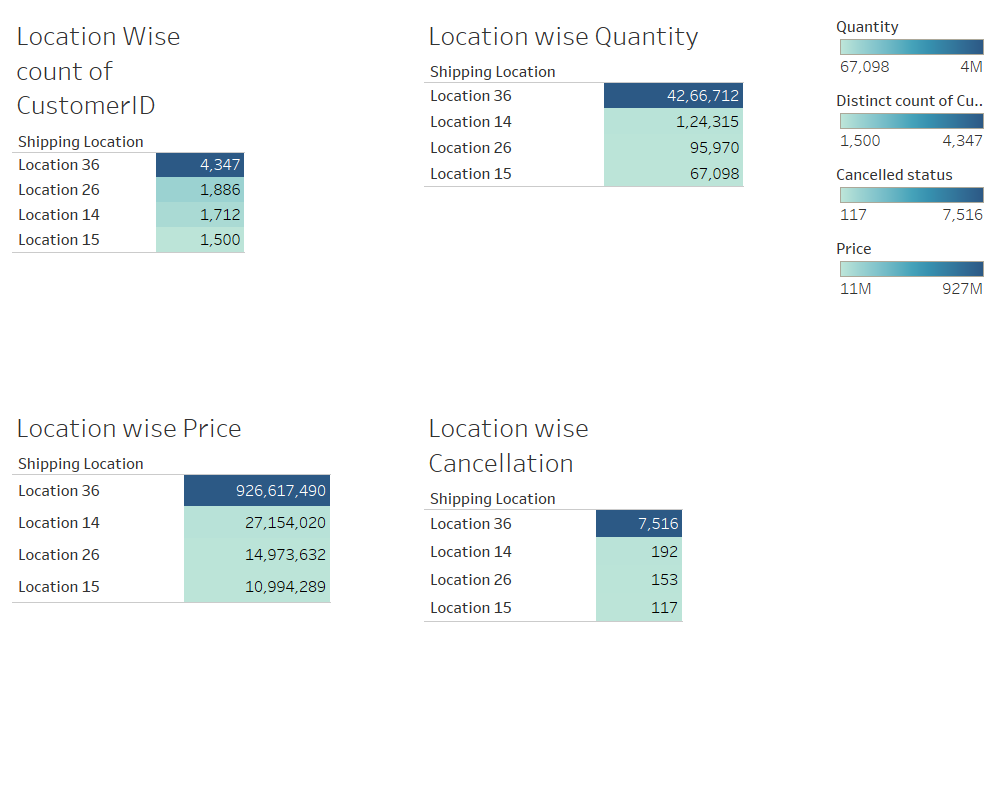


**Exploratory Data Analysis (EDA)**

EDA was carried out with the help of Tableau. We made use of various plots on the basis of variables and its data type to get the possible outcomes. Some of the insights are listed as follows:

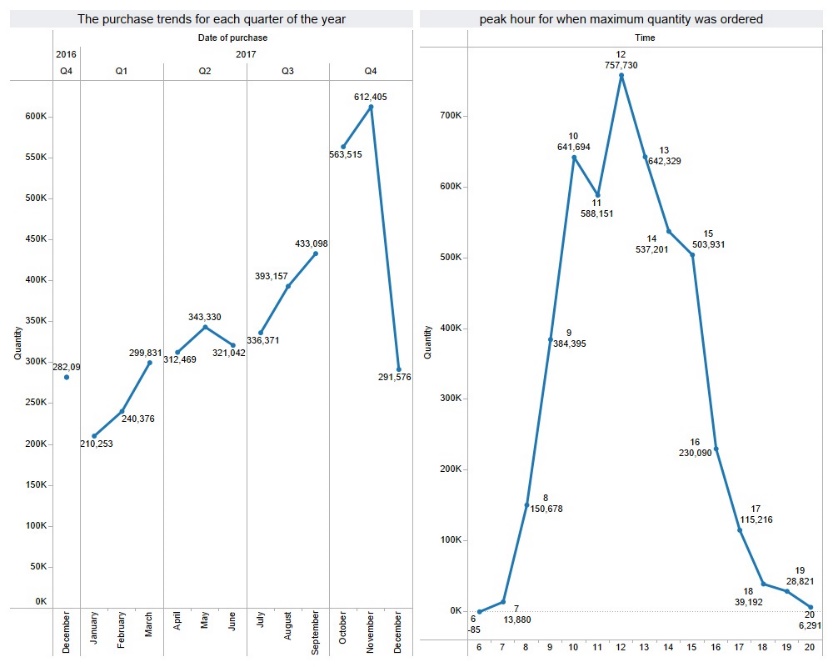
1] Location 36 has been a lime light for the sales of the greatest number of items for each product (42,66,712) along with maximum numbers of registered Customer IDs (4,347). Contrary to this, maximum number of cancellations happed at location 36 (7,516).

2] Other locations which made it to the second third and fourth place for maximum sales were location 14, 26 and 15 respectively. These locations also had a fair share for the number of registered customers.



3] Sales has been constant throughout the year but company was at their peak between October to November 2017.



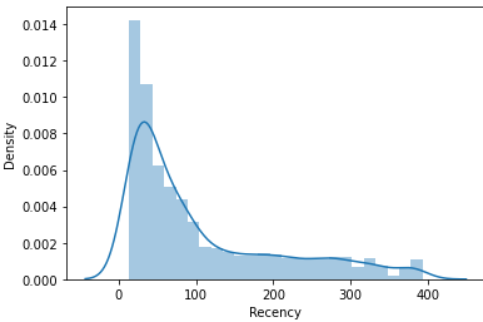


**RFM ANALYSIS**

RFM (Recency, Frequency, Monetary) analysis is a customer segmentation technique that uses past purchase behavior to divide customers into groups. RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

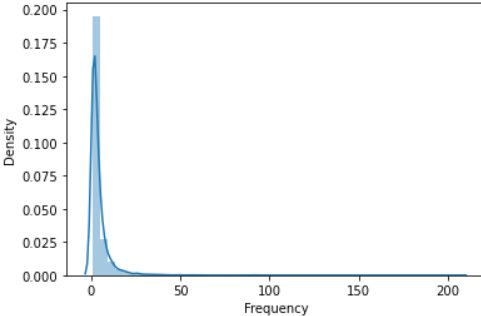
**Recency:**

To calculate recency, we chose a system date as (**01-01-2018**) from which we evaluated customer’s last purchase.



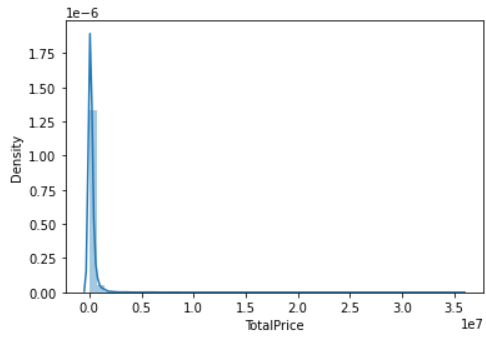
**Frequency:**

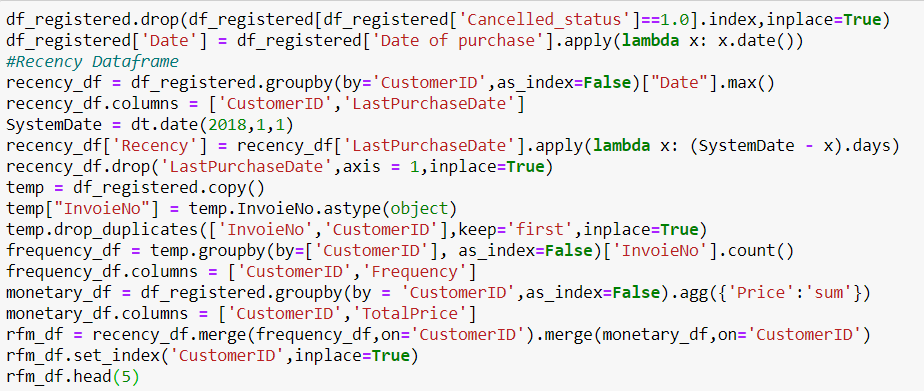
Frequency helps us to know how many times customer purchased from us. To do that we need to check how many invoices are registered from the same Customer ID.

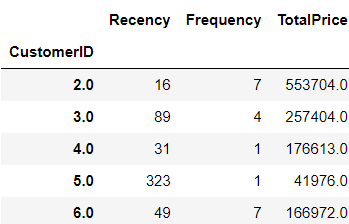


**Monetary Value:**

Monetary value helps us to know how much the customer has spent over time. In this we have calculated by multiplying the quantity with price per unit.

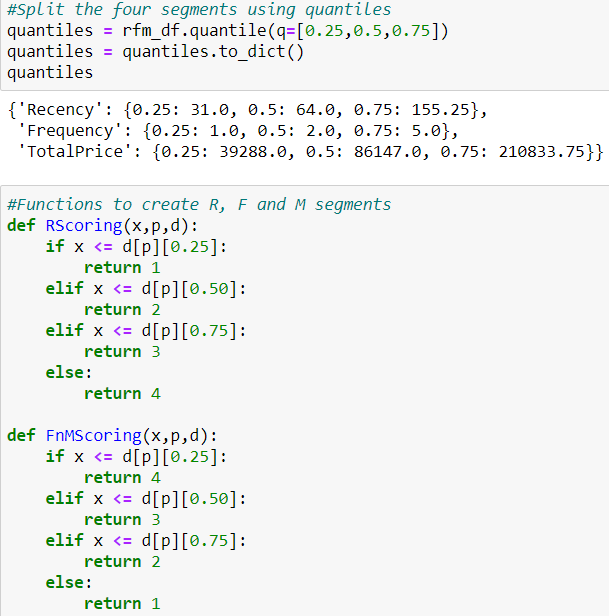


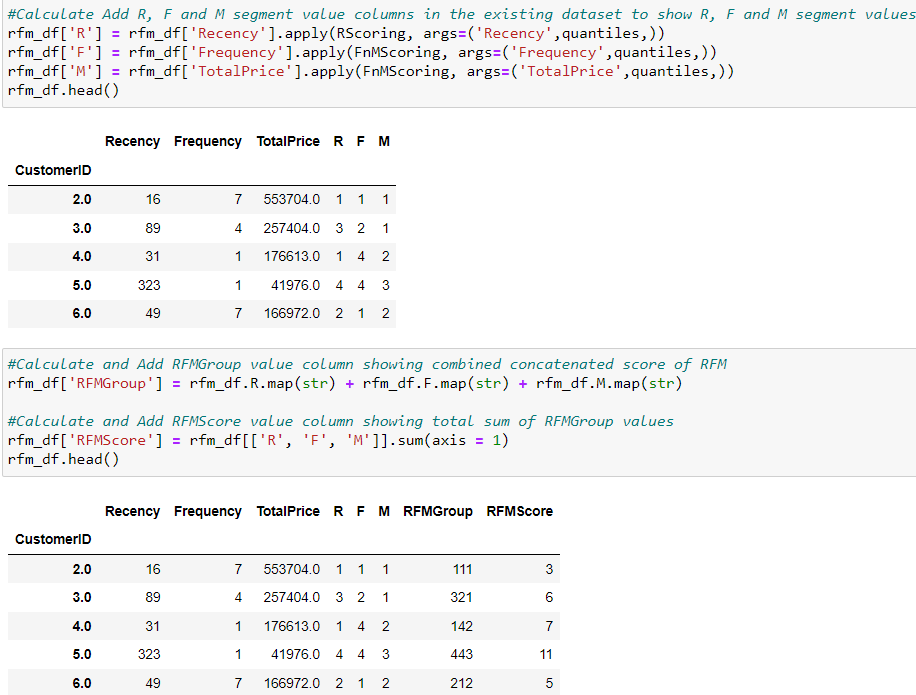




**RFM Group and RFM Scores:**

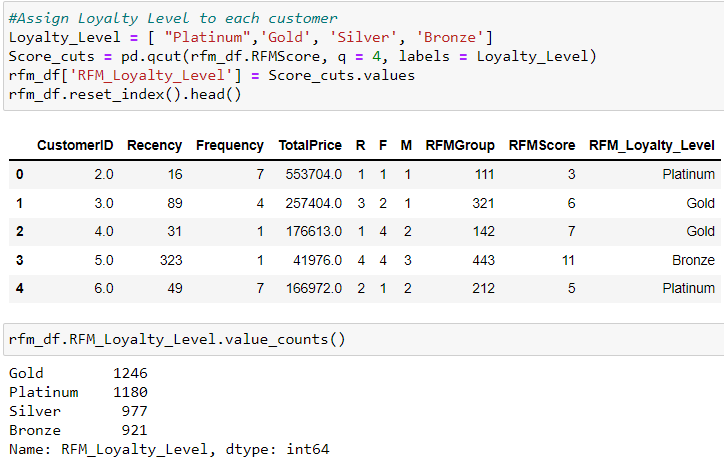
In order to create RFM values, we used the quantiles values of recency, frequency and monetary value. Further we created functions to assign individual R, F, M values. We used the string concatenation functions to get the RFM Group and sum function to get the RFM score.



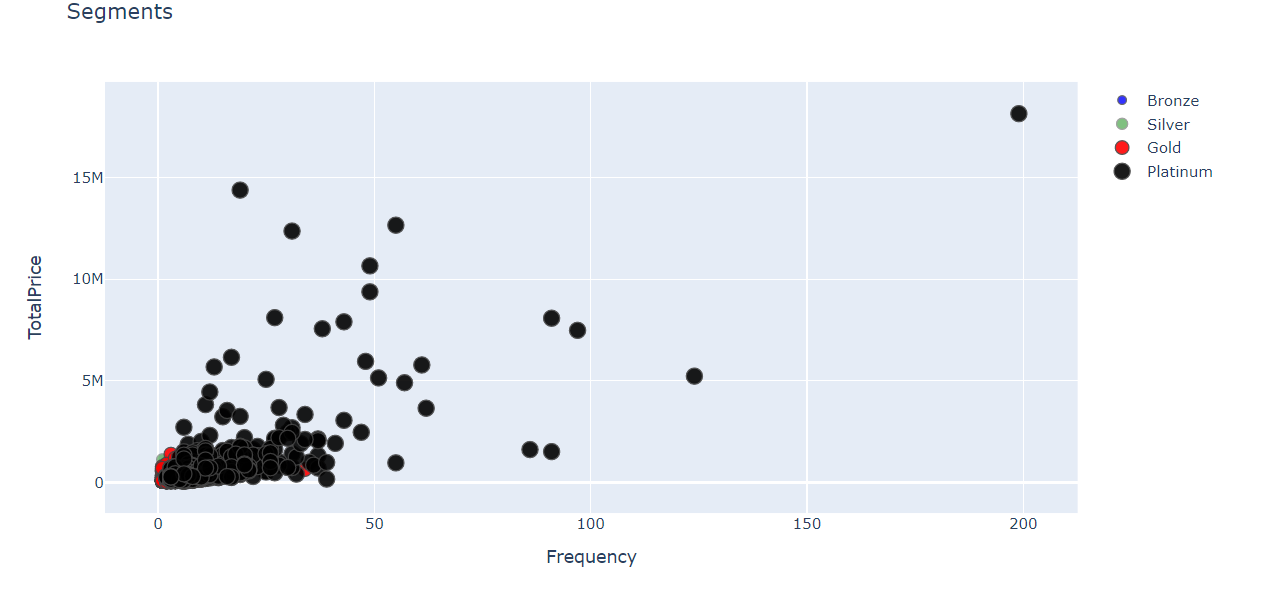


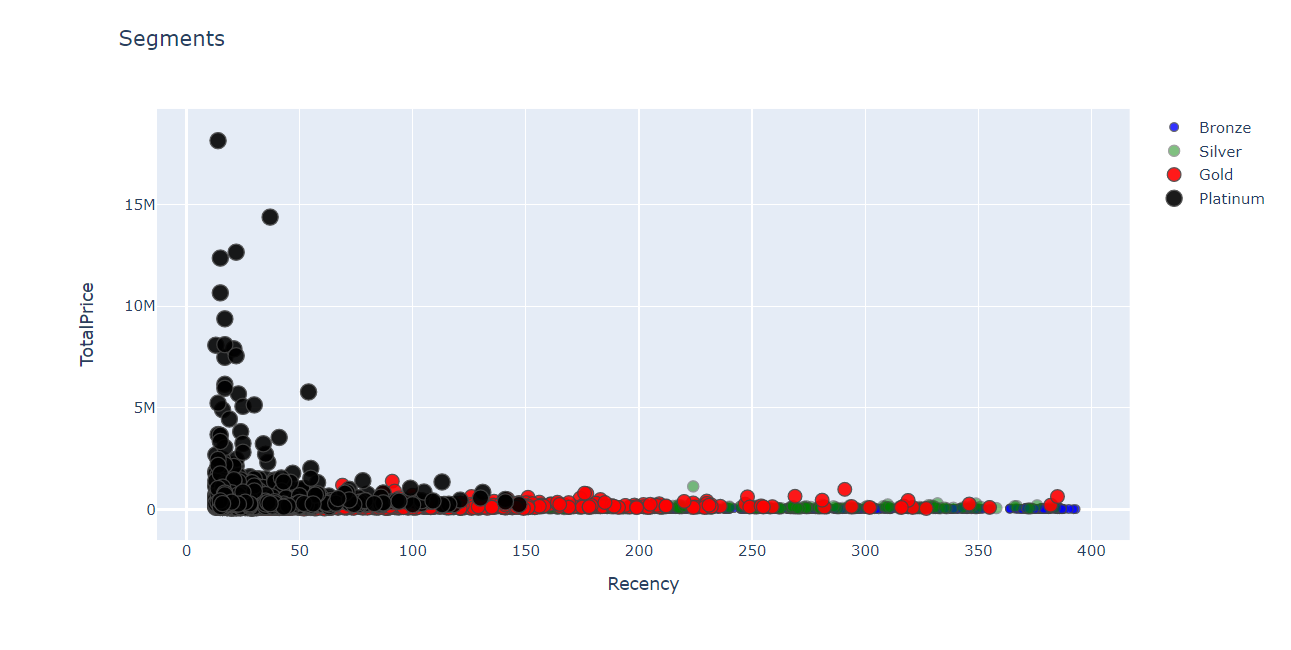
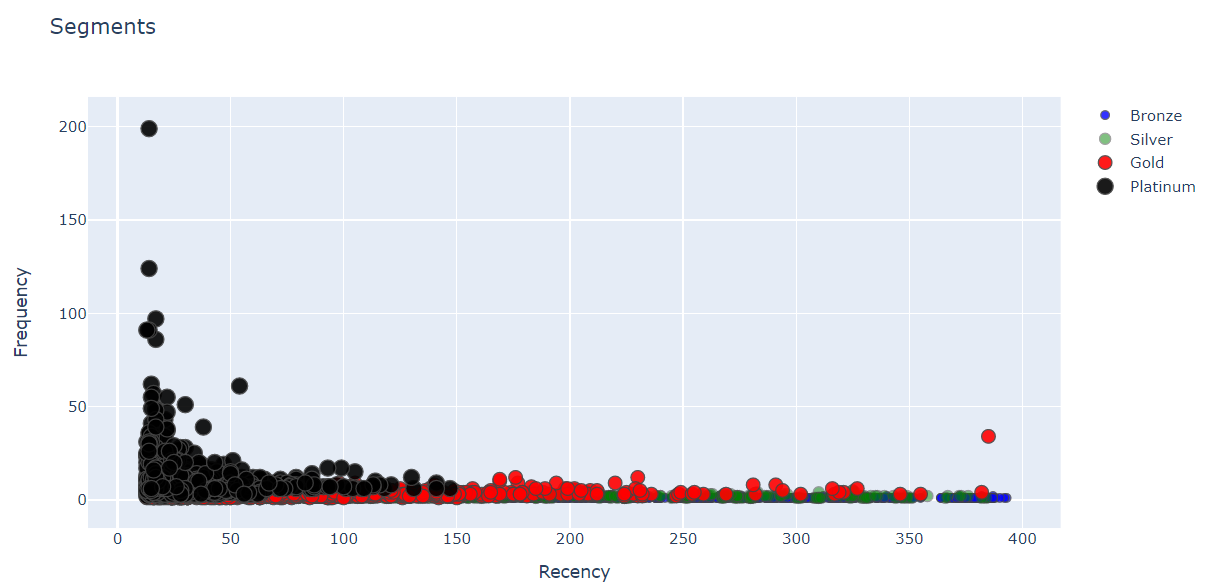
**RFM segments:**

We created segments as Platinum, Gold, Silver and Bronze, where, Platinum standing for the most loyal and bronze for the least loyal. We used the score cut function for segmenting it.

****

**EDA on RFM segments:**

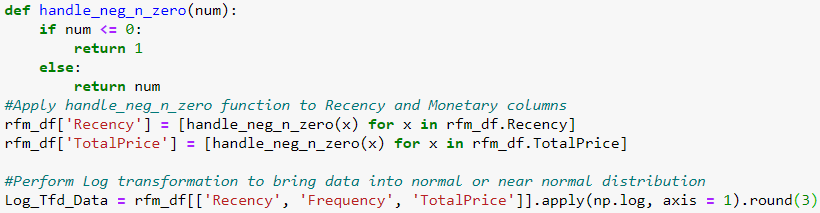
****

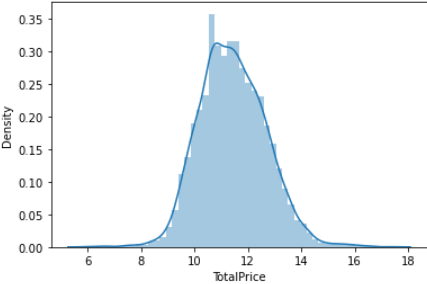
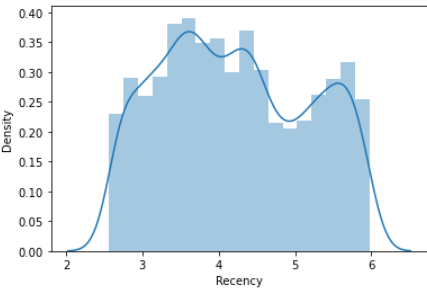
****

**K-MEANS CLUSTERING**

When you have unlabeled data, K-means clustering is a kind of unsupervised learning. The purpose of this technique is to locate groups in the data, with K representing the number of groups. Data points are grouped together based on how comparable their features are.

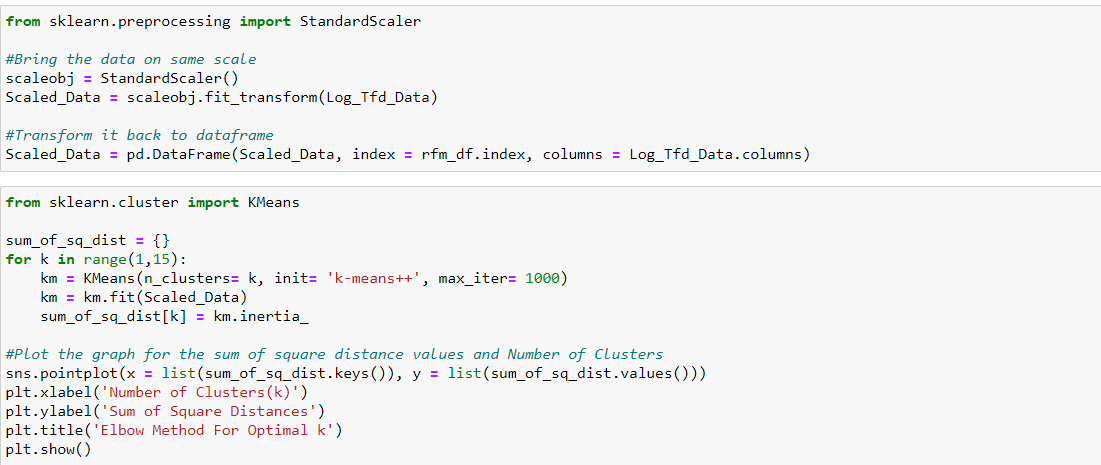
We have a Selected Number of Clusters in our dataset based on Elbow Method and Silhouette Score.

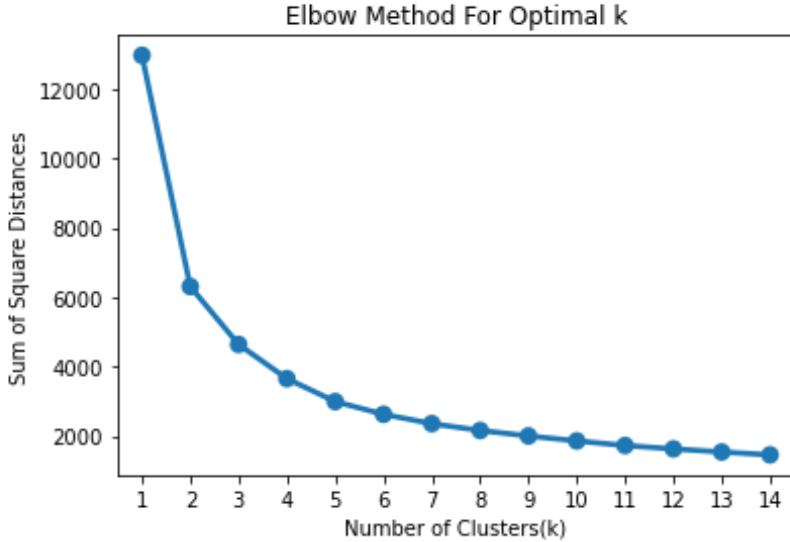




**Elbow Method:**

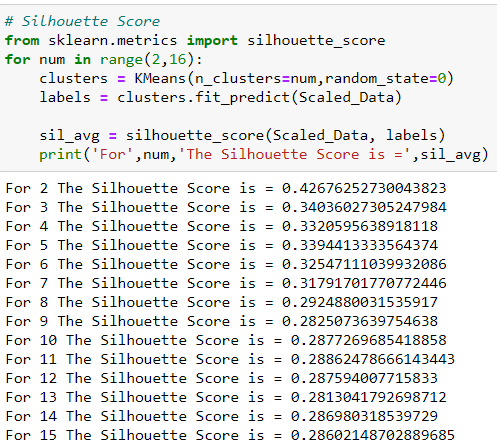
In K-means clustering, the elbow approach is used to estimate the ideal number of clusters. The sum of squared distance caused by various values of K is plotted using the elbow method. We standardized the data using the Standard scaler function from the sklearn.



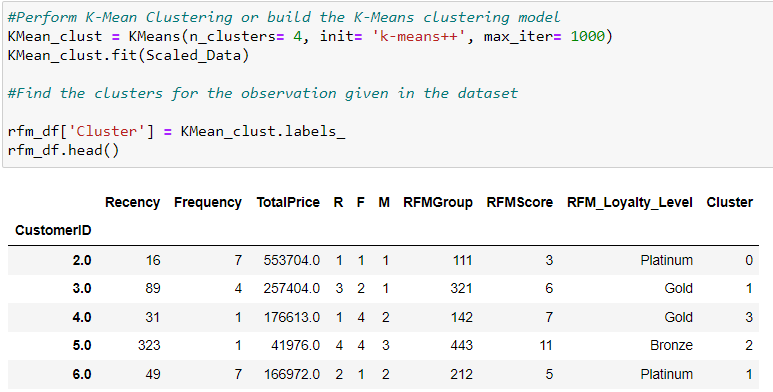


**Silhouette Score:**

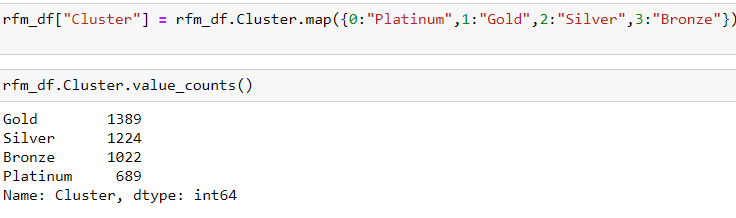
The silhouette **score** is a metric used to calculate the goodness of a clusteringtechnique. Its value ranges from -1 to 1. Values closer to 1 indicate that clusters are well apart from each other and clearly distinguished.

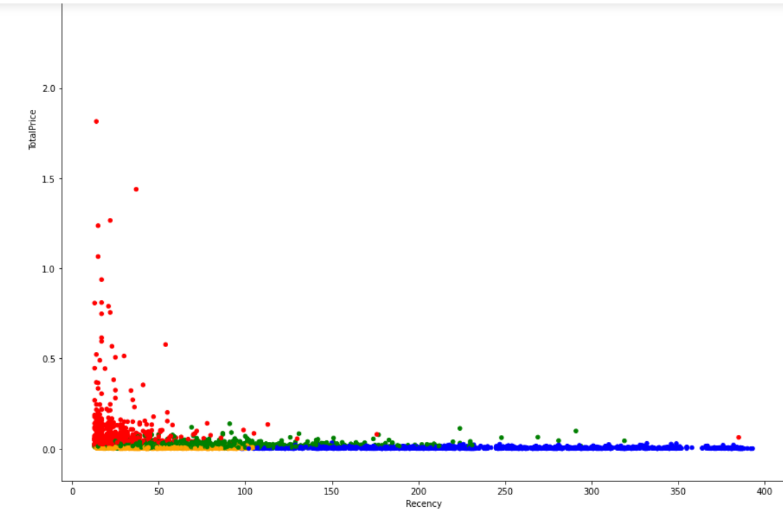


Ideal number of clusters with the help of elbow method and silhouette score was decided as 4. We then applied it to our scaled dataset to find our clusters.



We renamed the clusters using the map function and later calculated the individual frequency of each cluster made.





**ACCURACY**

In order to calculate the accuracy, we compared the RFM segments and K means clusters. We found a total accuracy of 47.82 percent which increased to approximately 80percent when number of clusters was reduced by 1 (3 was the second-best number of clusters on the basis of elbow method and silhouette score).

**BUSINESS INFERENCE**

1] RFM Score 111 are customers that have purchased the most recent, most frequently, and made the greatest revenue. Concentrate on loyalty programs and new product launches. Don't use discount pricing to promote incremental sales because these clients have shown a higher willingness to spend. Instead, concentrate on value-added services such as product suggestions based on previous purchases.

2] RFM score X1X are customers who frequent your store the most. For these frequent visitors, loyalty schemes are successful. X1X methods include advocacy programs and reviews. Finally, consider providing Free Shipping or other similar perks to these clients.

3] RFM Score XX1 are customers that have brought in the most money for your business. These clients have shown a strong willingness to pay. To improve AOV, consider premium deals, subscription tiers, luxury products, or value add cross/up-sells. Discounts are a waste of margin.

4] RFM Score X13 and X14 are customers that come back frequently but do not spend a lot of money. You've already established a loyal following. Increase monetization by making product recommendations based on previous purchases and offering incentives based on spending criteria.

5] RFM Score 14X are first-time buyers. The majority of clients never progress to being loyal. Having clear first-time buyer strategy in place, such as triggered welcome emails, will pay off.

6] RFM Score 44X are great past customers who haven't bought in a while. Customers leave for a variety of reasons. Depending on your situation price deals, new product launches, or other retention strategies.