Capstone Project

Customer Segmentation

Content

1. Some EDA

- Dealing with missing Values
- Dealing with cancelled/Returned orders
- Transaction’s Frequency (Count)
  - Year, Month, Day and Time
- Transaction’s Amount (Sum)
  - Year, Month, Day and Time

2. Model Preparation
3. Model
4. Cluster Visualization

Problem Statement

Identifying major customer segments on a transactional data

Data Summary

Data set name – Online Retail

Shape of Dataset- 541909 rows, 8 columns

Columns - 'InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'UnitPrice', 'CustomerID', 'Country'

Some EDA

- **Dealing with missing Values**

- Column “**customerID**” had null values – this was the main thing, because we can not fill these values with any of the number, as these are customers only, so we had to remove them.

- **Dealing with Cancelled/Returned products**

- Column “InvoiceNo”, was the one from which we could see the cancelled order – I have just dropped those rows, because those were not required for clustering

- I have split the “Date” into Month, Year, day, and Hour, and removed the duplicate entries.

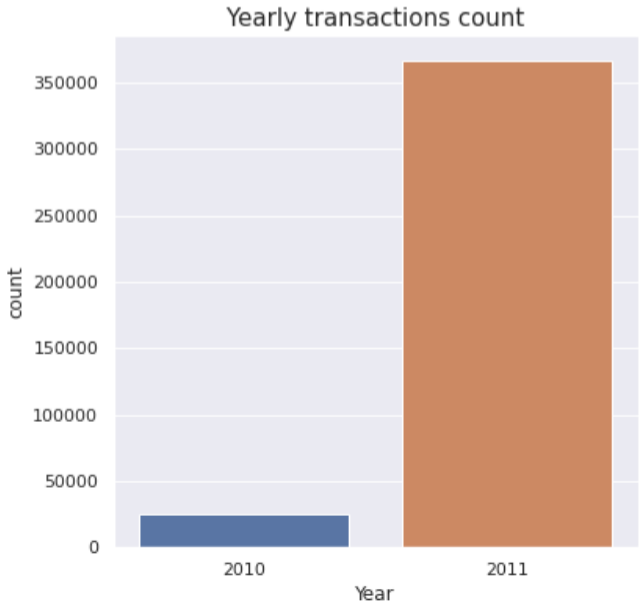
Some EDA

- **Transaction’s Frequency (Count)**
  – Year, Month, Day and Time

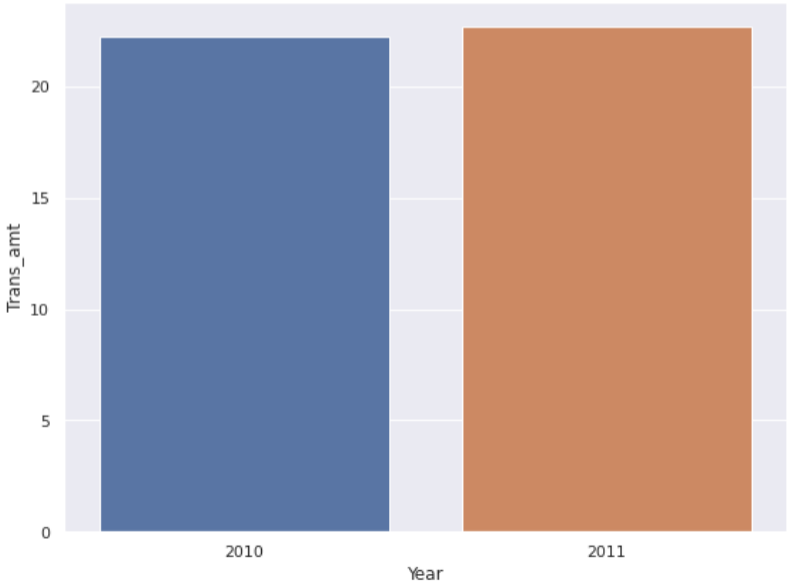
- Here I have plotted some graphs showing the total number of products sold and sum of products sold

Transaction's Frequency (Year)

The highest number of sales was in year 2011

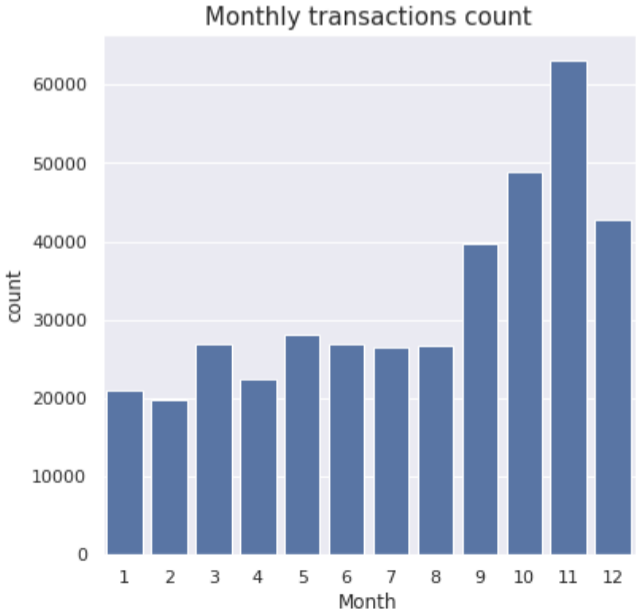


Amount of sales was around same

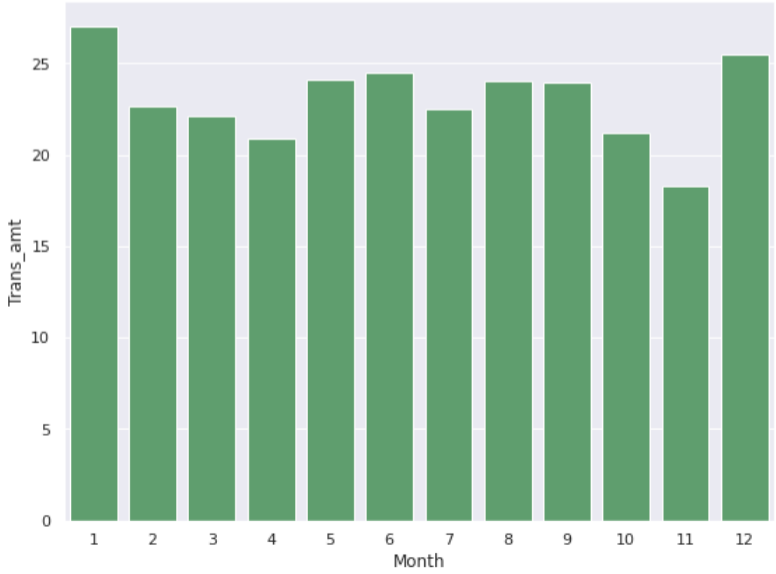


Transaction's Frequency (Month)

Highest number of sales was in November month

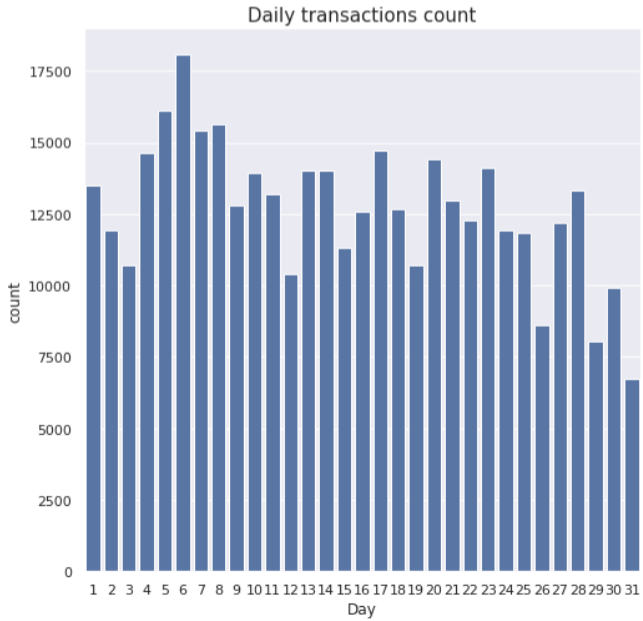


Sum of sales was highest in January

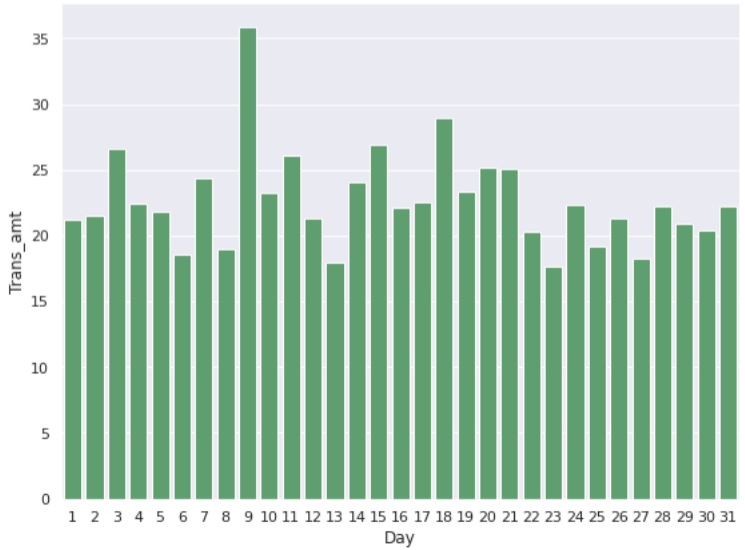


Transaction's Frequency (Day)

The highest number of sales was on 6th day of the month

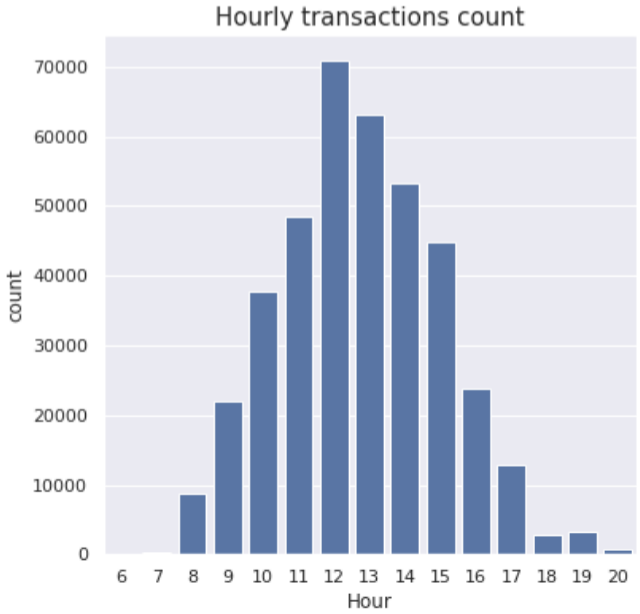


Sum of total sales was highest on the 9th day

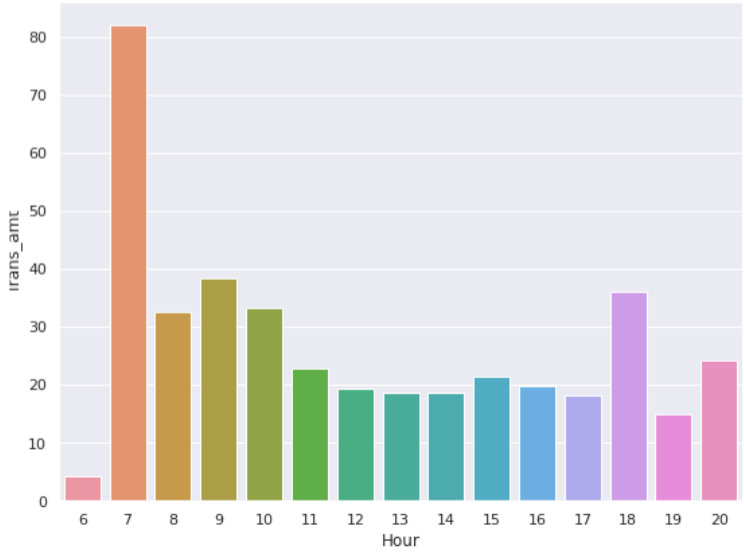


Transaction's Frequency (Hour)

The highest number of sales was around 12:00



The highest Sum of sales was at 7:00



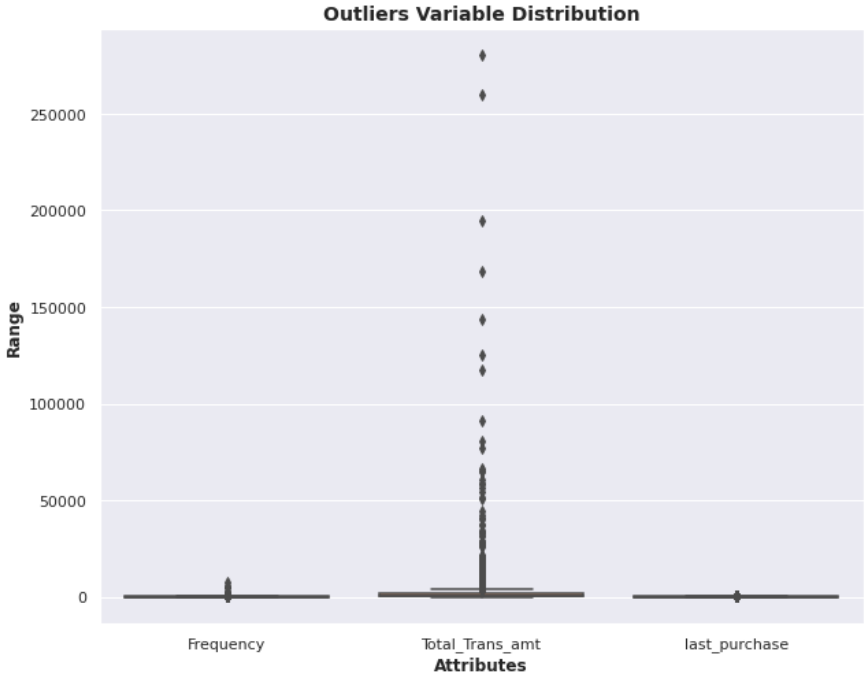
Model Preparation

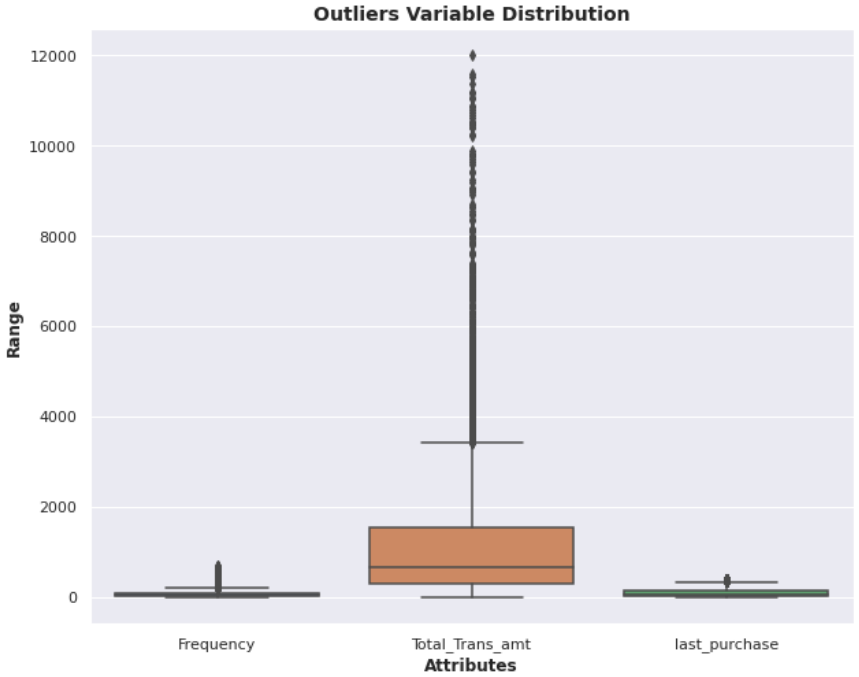
Here I have converted this dataframe into different in which I have put three columns only

- Frequency – Total number of products purchased by a customer
- Total_Trans_amt – Sum of the money a customer has spent on products
- Last_purchase – last transaction of a customer (Days)

Model Preparation

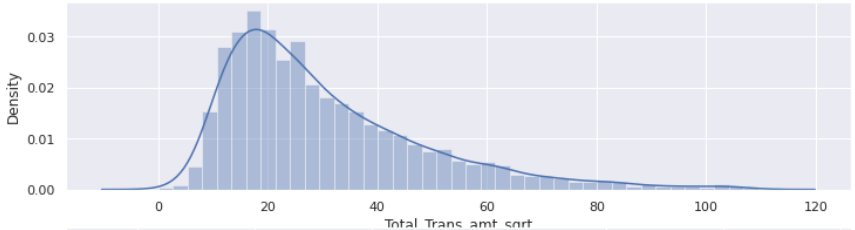
Outliers – Detected and Removed

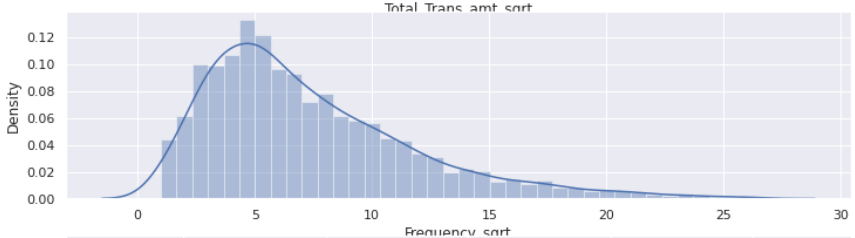


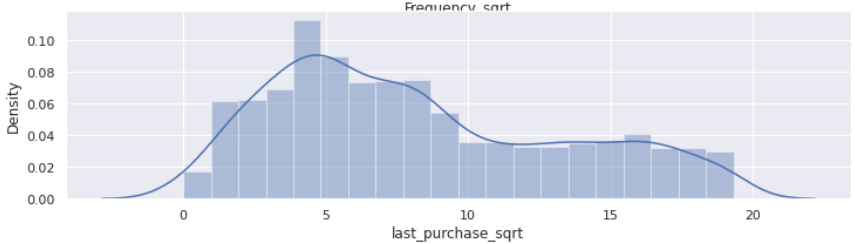


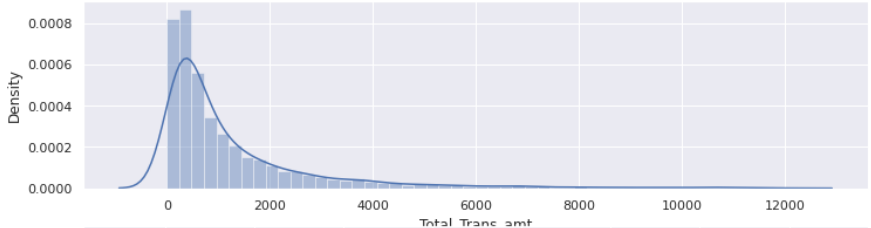
Model Preparation

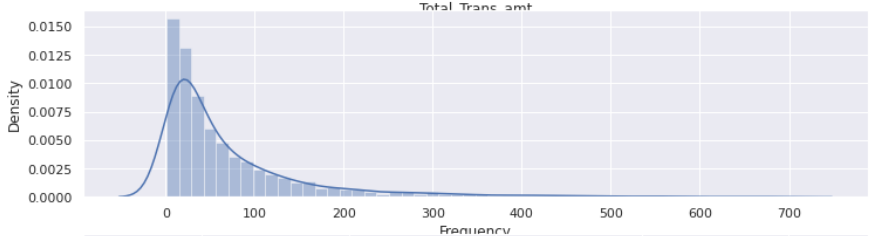
Distribution – Skewed and changed

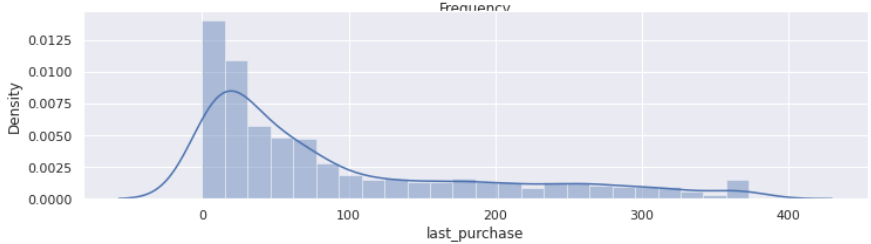






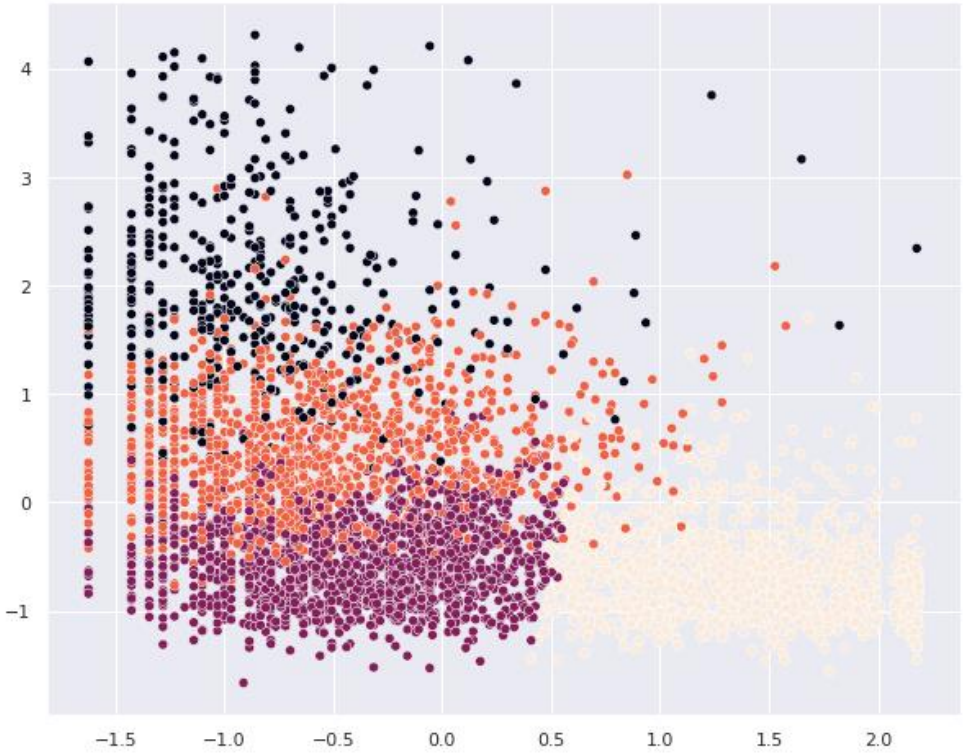






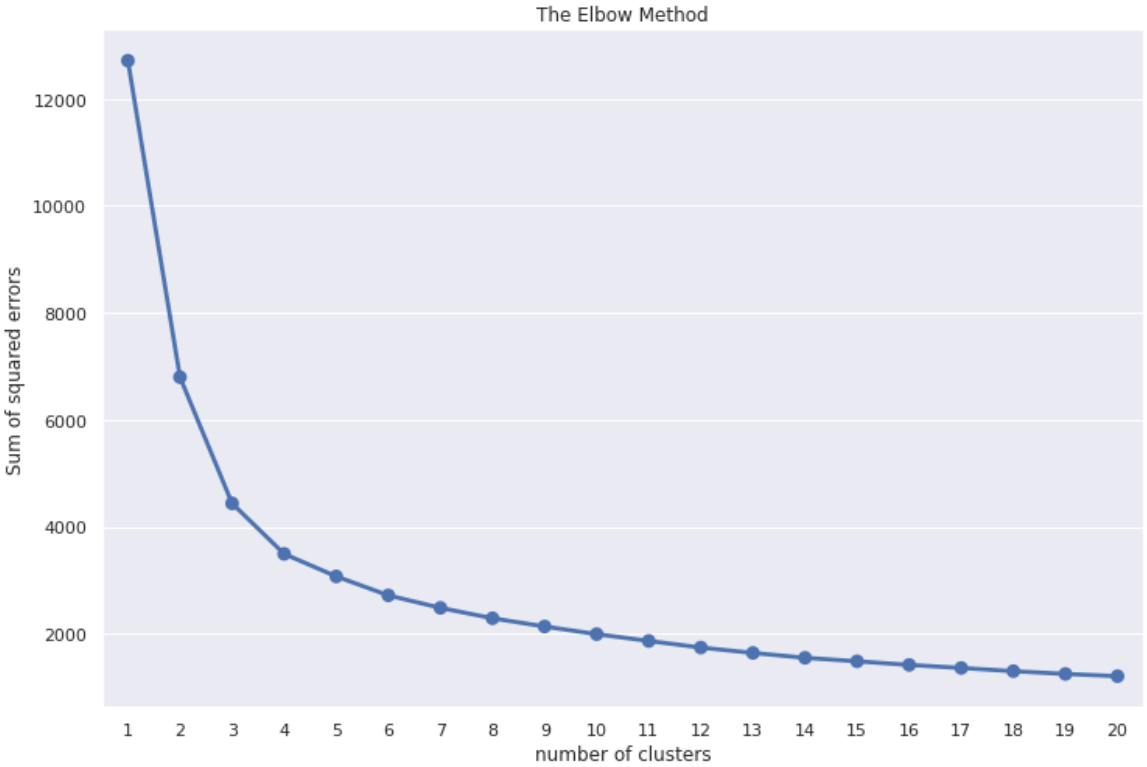
Model

Predicting random with the value k = 4



Model

Checking Optimum number by Elbow Method



Model

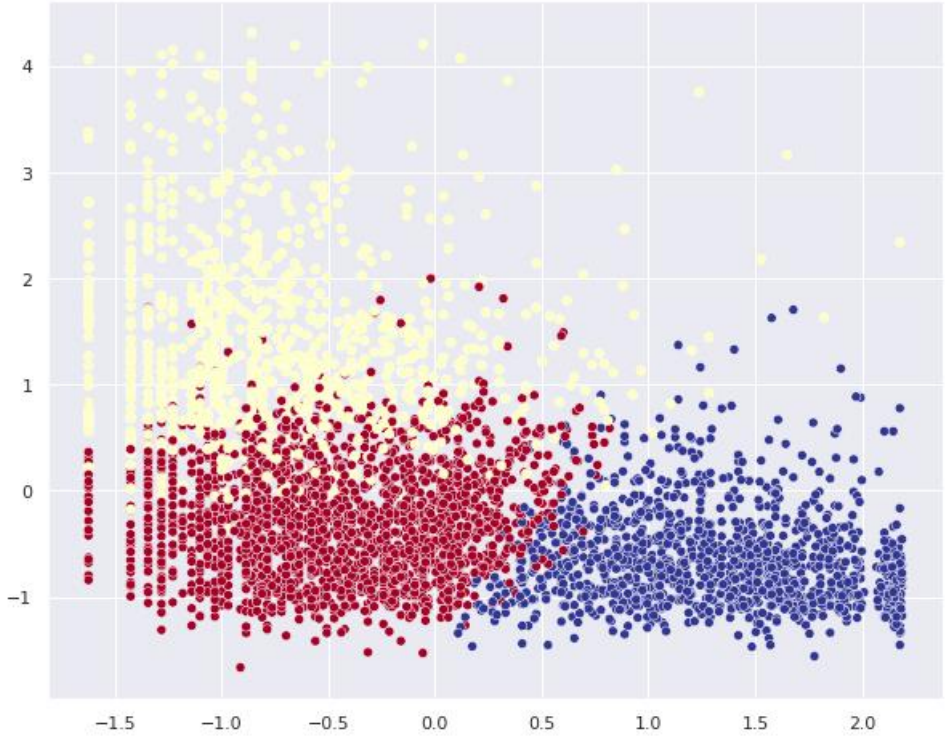
Checking Optimum number by silhouette_score

For n_clusters = 3, silhouette score is 0.3901851051400997
For n_clusters = 4, silhouette score is 0.3600857865526071 5
For n_clusters = 5, silhouette score is 0.345233582260781
For n_clusters = 6, silhouette score is 0.3233378722785445
For n_clusters = 7, silhouette score is 0.3095527732006726 5
For n_clusters = 8, silhouette score is 0.288957487227708 1

So for K = 3, we had the highest score

Model

Cluster for K = 3



Conclusions

- There were 8 features in total, from which we have extracted 3 only
- From those three, we first plotted the cluster with random value 4, but later on we checked it with 2 different techniques
- We have used the Elbow method for having the range of cluster numbers - we had range [3,8]
- We checked this range, by the help of Silhoutte Score, the number 3 was giving the highest score in the given data, so we have chosen k = 3
- At the end we have plotted the cluster visualization with 3 different clusters.