CUSTOMER SEGMENTATION

K-Means Clustering & RFM Modeling

JULY 31, 2022

KARTHIKEYAN CHELLAMUTHU, SUBHASHINI NATARAJAN

Introduction:

The business problem that we plan to build the model for is, customer segmentation, for the Brazilian e-commerce company, Olist. Customer Segmentation is an important step in marketing for an organization. The process helps in customizing marketing campaigns, prevent customer churn, prioritizing product development or services.

Data Source:

Data for the project is sourced from Kaggle. This is a public dataset of orders made at Olist store. This dataset has several dimensions including, customer, geolocation, order, payment, reviews, product, product category and sellers. For the purposes of customer segmentation, not all dimensions will be considered, but only the following will be used -

- Customer,
- Order
- Customer review

Link to the dataset:

https://www.kaggle.com/code/marianakralco/brazilian-ecommerce-clustering-rfm-and-kmeans/data

Process

The data for the project includes customer data, order, order reviews, order items, product data in separate csv files. The individual csv files are imported in separate python dataframes. The individual data frames are then merged into a single data frame, called olist_df, with required fields.

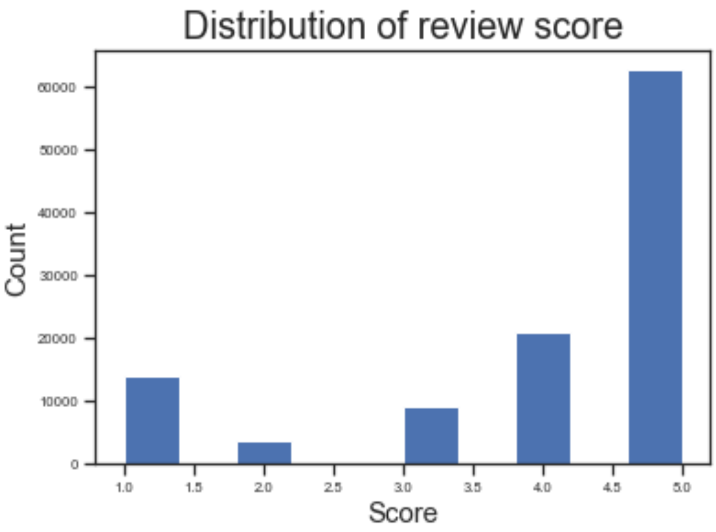
The final data frame is inspected using .info() method and the data frame data types were listed as below –

```
customer_id                 110774 non-null object
customer_unique_id          110774 non-null object
order_id                    110774 non-null object
product_id                  110774 non-null object
review_id                   110774 non-null object
order_purchase_timestamp    110774 non-null object
customer_city               110774 non-null object
product_category_name       110774 non-null object
review_score                110774 non-null int64
order_item_id               110774 non-null int64
review_creation_date        110774 non-null object
price                       110774 non-null float64
freight_value               110774 non-null float64
```

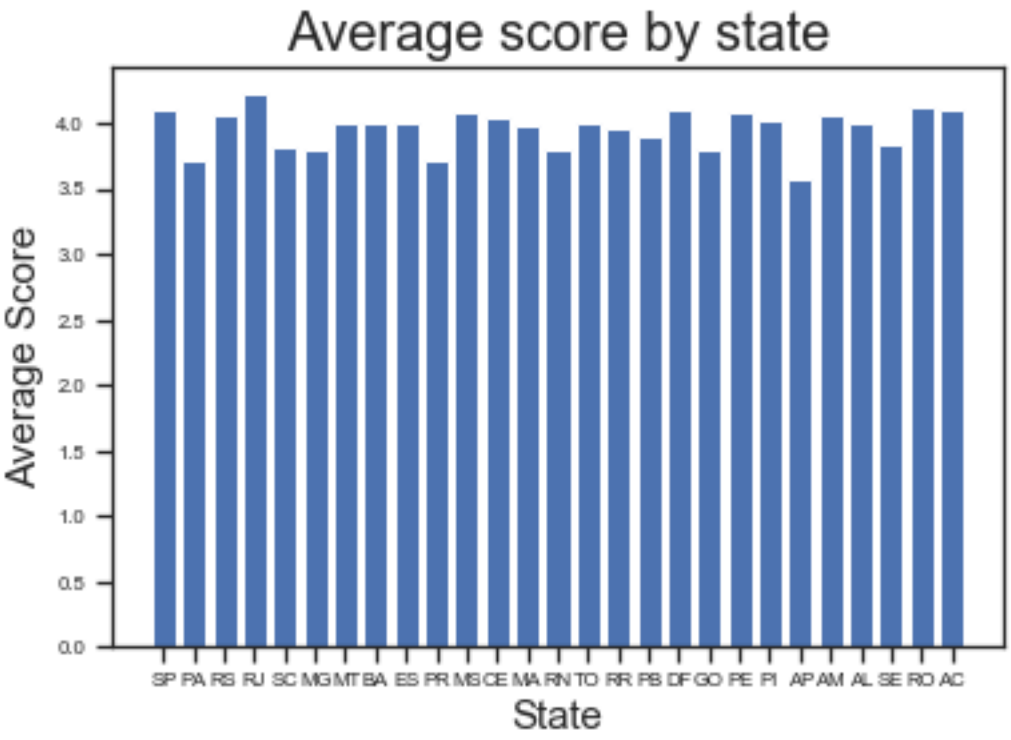
It is observed from the above list that order_purchase_timestamp is populated as an object. In order to obtain trends, the data type of the field has to be converted into date using to_datetime() method.

The following data explorations were performed to understand the data more.

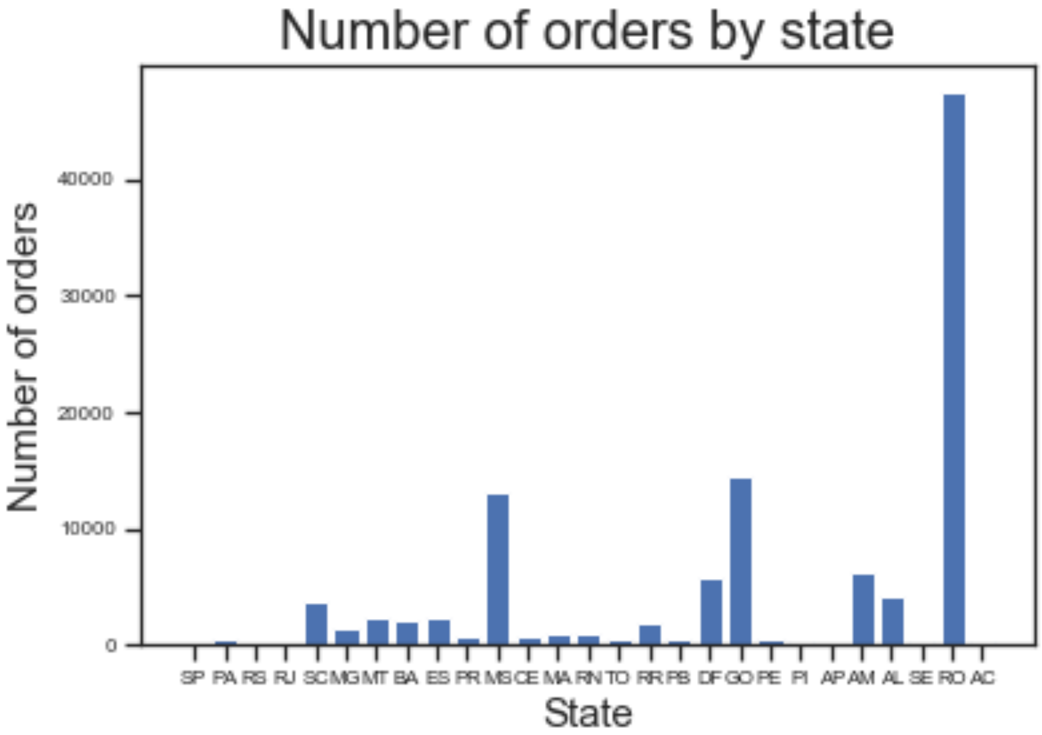
1. Obtain the distribution of review score – The below given distribution shows that number of records with higher ratings are quite high. This implies good quality products and service from Olist.



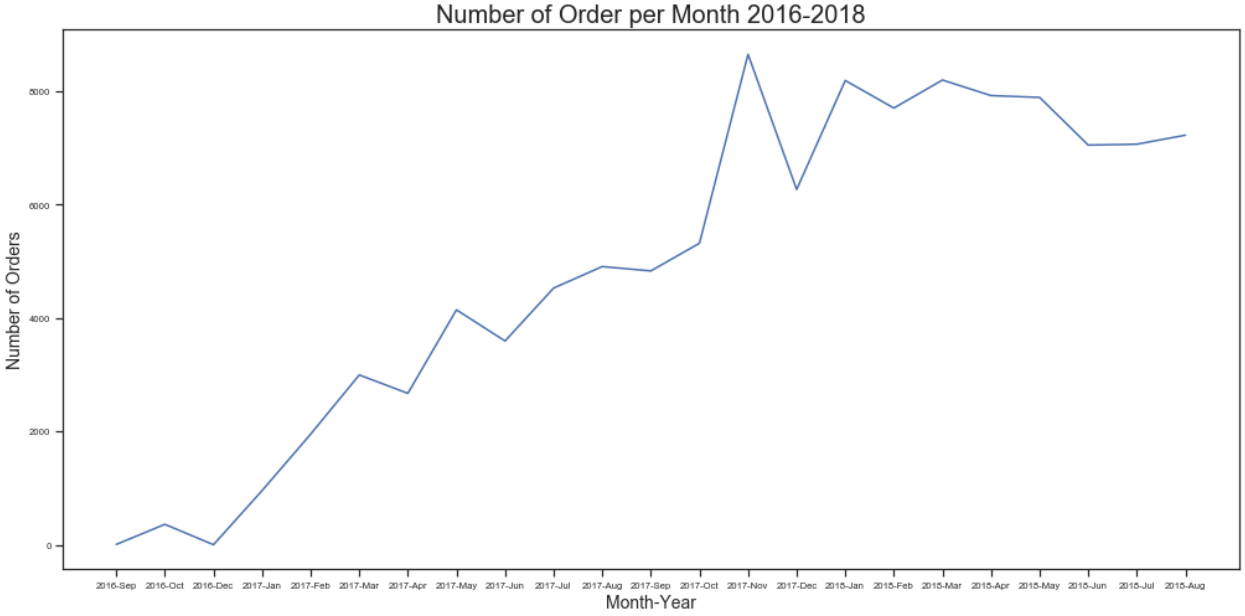
2. To ascertain any regional preferences, did a bar chart of mean review score by state. The plot below shows no strong preferences by state.



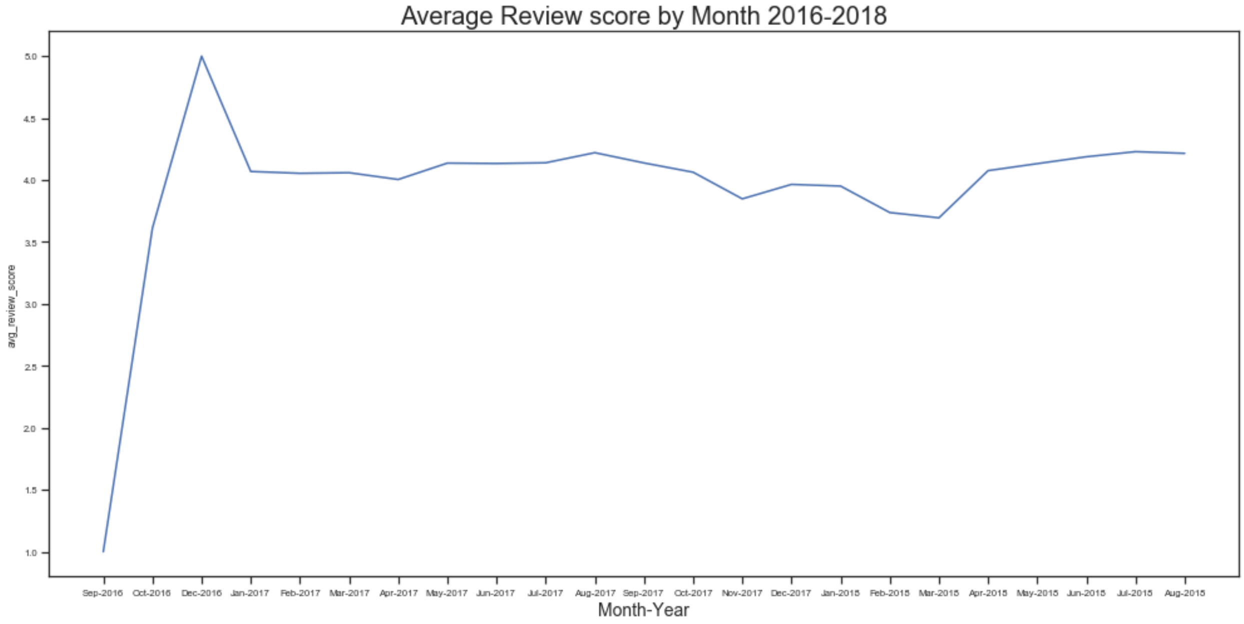
3. To ascertain the number of orders by state, did a bar chart of order count by state. The visualization below indicates that Olist is popular in 5 states and the most popular in Rondonia (RO).



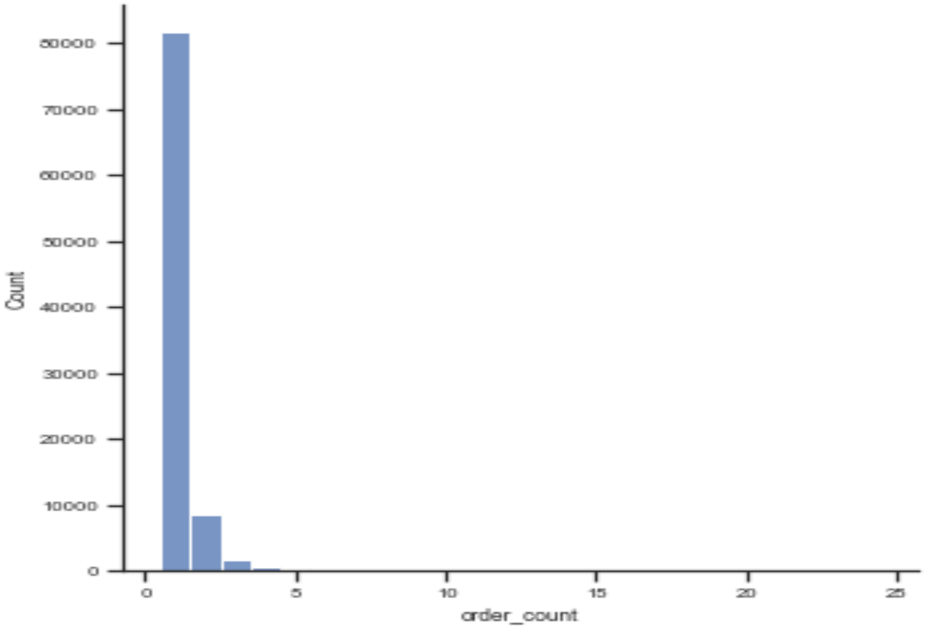
4. In order to plot the trend, grouped the data set based on year – month of order timestamp and plotted the number of orders against the year-month dataset. The trend is shown in the chart below. The number of orders from the plot above can be seen with steady increase apart from few occasional dips and the latest data at the end of 2018 is well above 70000 orders.



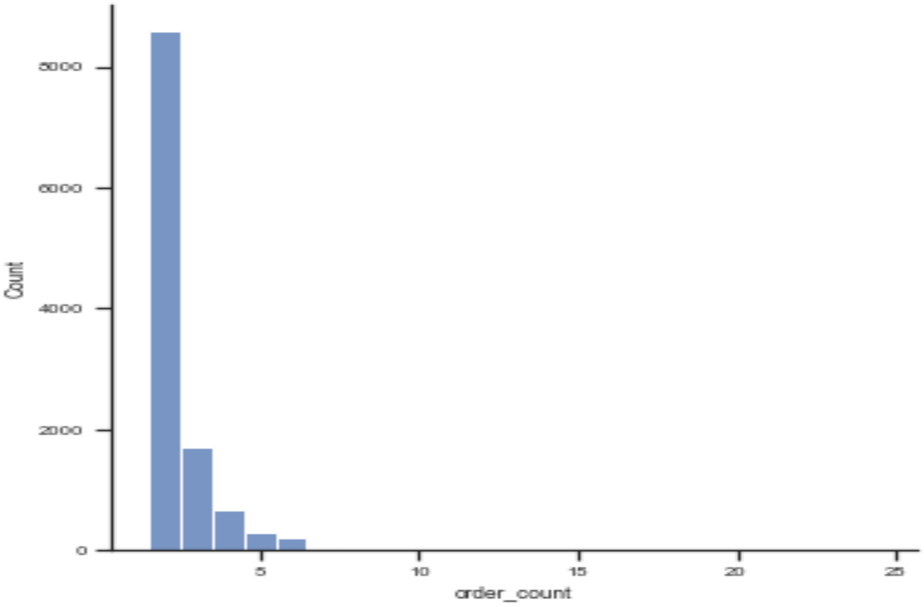
5. Mean review score was also plotted against Month – Year. The average review score for the products from the plot below can be seen as pretty constant and above 4, except for a brief dip between Jan 2018 - April 2018.



6. In order to check the distribution of orders placed by the same customer, performed a distplot for number of orders. The number of one-time orders is the highest, followed by twice and thrice.



7. In order to check the distribution of multiple orders placed by the same customer, performed a distplot for number of orders, where the number of orders is more than 1. Number of three orders and more, by the same customer is lesser than 20,000.



K-Means Clustering & RFM Modeling

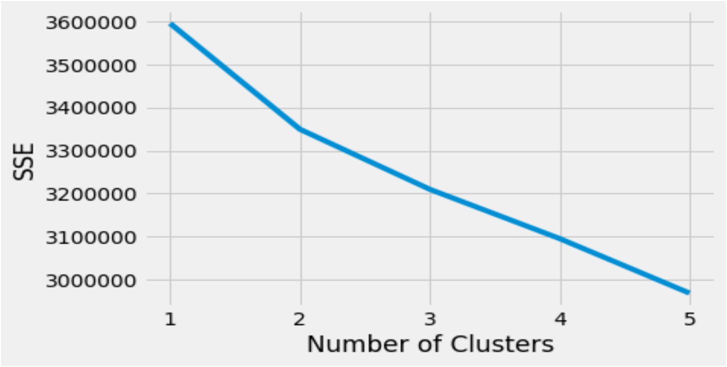
Feature Scaling:

After gaining an understanding of the dataset, through exploratory data analysis, we selected the required features for clustering. The important preprocessing step for unsupervised machine learning algorithms is features scaling. For implementing K-Means algorithm, we chose standardization type of feature scaling using StandardScaler class. This shifts the values for each numerical feature in the dataset so that the features have a mean of 0 and standard deviation of 1.

Choosing appropriate number of clusters:

The optimal number of clusters (k), is a required value for clustering. We performed two methods to determine this value –

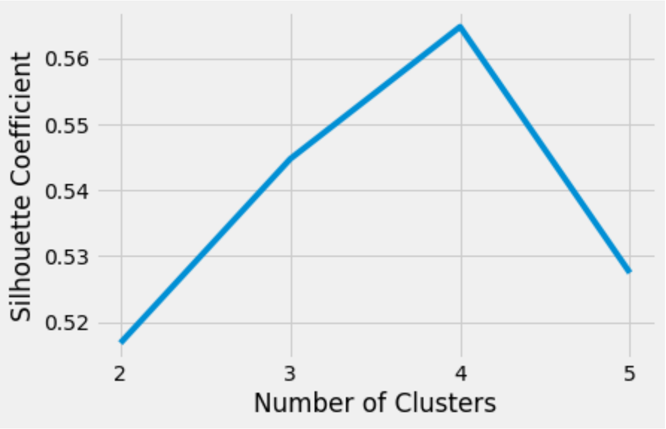
1. Elbow method – In this method, we determined the sum of squared errors and plotted them.



From the plot, the possible elbow point is 2.

2. Silhoutte Coefficient Method –

Silhoutte score is determined for the data and plotted.

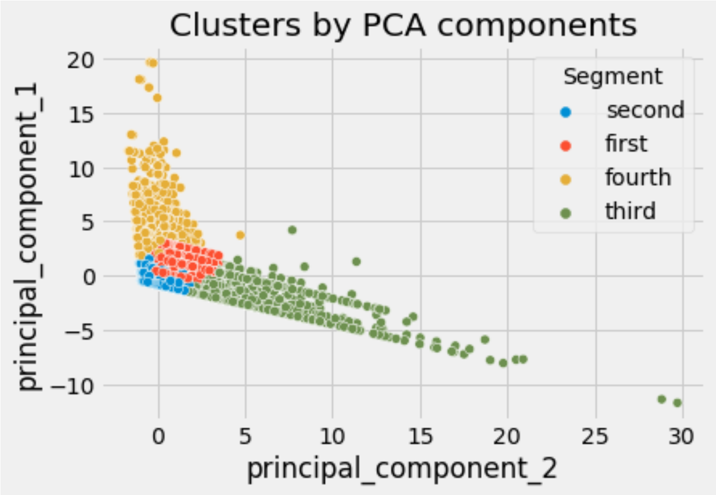


From the plot it is observed that, 4 has the highest Silhoutte score. As we are interested in exploring more than two clusters of customers, we choose 4 as our k-value.

K-Means clustering:

K-Means, clustering is performed with the cluster value of 4. K-Means clustering for better representation requires another transformation step – dimensionality reduction. We performed this using Principal Component Analysis.

We performed PCA and fit the scaled data with it. We stored PCA components into a data frame and named the segments. We plotted the clusters by PCA components.



Customer Segmentation Results:

Upon examining the clusters prepared by the algorithm, we see four segments of customers.

Cluster	Review score	# of orders	Price Range	Description
1	1,2,3	1,2,3	$7-$678	Low to Mid review score , Number of orders < =3 , Low to Mid price range
2	4,5	1,2,3	$9-$584	High review score, Number of orders <= 3, Low to Mid price range
3	1,2,3,4,5	1,2,3,4,5,6,7,8	$570-$6979	All review scores, Number of orders up to 8, Mid to High price range

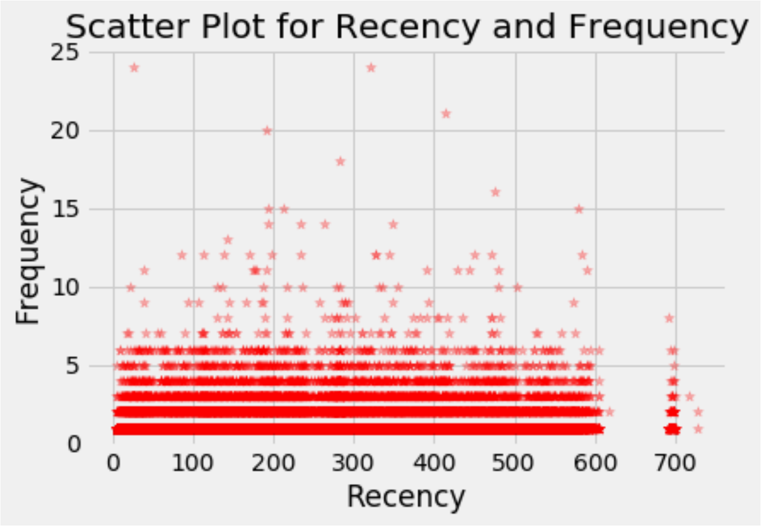
4	1,2,3,4,5	4 to 24	\$6 - \$1000	All review scores, Number of orders > 3, Low to High price range

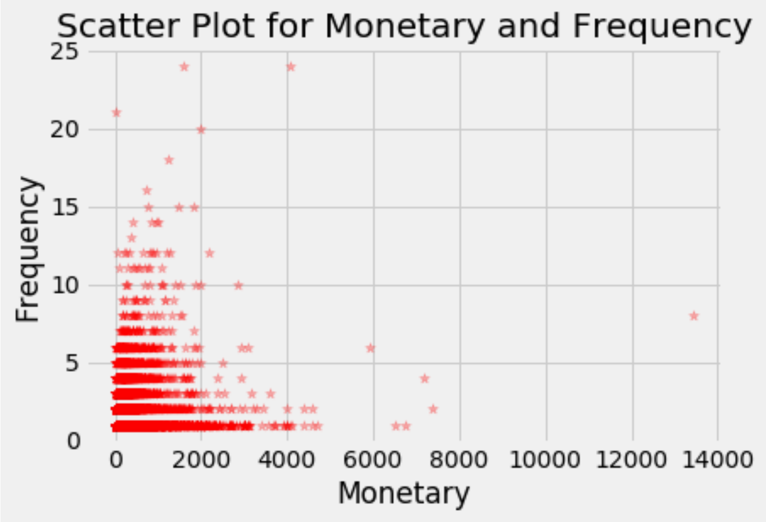
RFM Modeling

RFM modeling is another technique for performing customer segmentation. In this method, we determined, monetary value purchased by every customer, recency of purchase by every customer and frequency of purchase by every customer.

- Monetary Value, is determined by sum of price of products purchased by the customer.
- Frequency is determined by number of orders placed by the customer on different dates.
- Recency is determined by the time period difference between the latest purchase in the dataset and the order date by the customer.

To understand the data, we plotted the data to understand recency, frequency and monetary values.





The next step in the process, is score assignment by bucketing the values.

- R value is bucketed as, 3 - if less than 90 days, 2- if between 90 and 120 and 1 otherwise.
- F value is bucketed as 2, if more than 1 purchase is made and 1 otherwise.
- M is bucketed using qcut method.

Customers are then segmented based on RFM score as below.

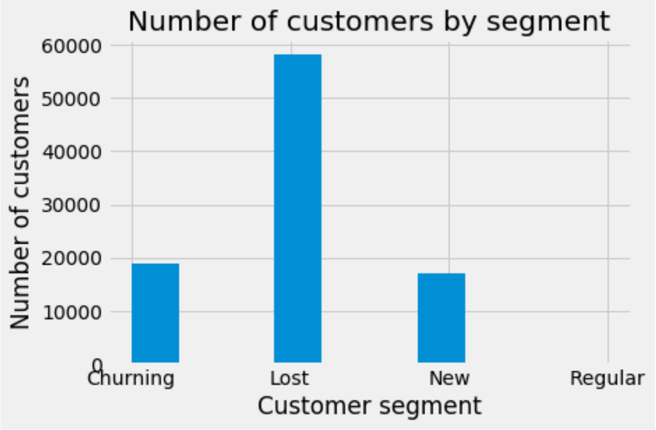
1. If RFM score is in 313, 312, 311 then classified as New.

2. If RFM score is in 323, 321 and 322 then classified as Regular.

3. If RFM score is in 212, 213, 211, 221 and 222 then classified as Churning.

4. Otherwise classified as Lost.

The customer segments are then plotted as below.



RFM results:

For olist e-commerce site, the number of new customers is approximately 18K, the churning customers are approximately 19,000 and lost customers are close to 59,000. This shows that the e-commerce company is not performing as in the beginning and requires significant improvement in products, services and marketing.