Capstone Project

Clustering on Online Retail Customer Segmentation

By - Shourya Chandra Sai

Project Agenda

- Segment 1: Data Description
- Segment 2: Defining the Problem Statement
- Segment 3: Data Preprocessing & Feature Engineering
- Segment 4: Exploratory Data Analysis(EDA)
- Segment 5: RFM Analysis
- Segment 6: Hypothesis on the number of clusters
- Segment 7: Model 1: K-means Clustering using the K-Elbow method
- Segment 8: K-means Clustering using the Silhouette scores method.
- Segment 9: Conclusion

Segment 1: Data Description

Below is the description of all the columns in the dataset.

- **InvoiceNo**: Invoice number. A 6-digit integral number is uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
- **StockCode**: Product (item) code. A 5-digit integral number is uniquely assigned to each distinct product.
- **Description**: Product (item) name.
- **Quantity**: The quantities of each product (item) per transaction.
- **InvoiceDate**: Invoice Date and time. The day and time when each transaction was generated.
- **UnitPrice**: Unit price. Product price per unit in sterling.
- **CustomerID**: Customer number. A 5-digit integral number is uniquely assigned to each customer.
- **Country**: Country name. The name of the country where each customer resides.

Segment 2: Defining the Problem Statement

Segregating the variables into categorical, continuous/discrete, date-time and, textual in nature.

Variables:

- **InvoiceNo** = It is categorical and Nominal in nature.
- **StockCode** = It is categorical and Nominal in nature.
- **Description** = It is textual in nature.
- **Quantity** = It is continuous in nature.
- **InvoiceDate** = It is of Datetime in nature.
- **UnitPrice** = It is continuous in nature.
- **CustomerID** = It is categorical and Nominal in nature.
- **Country** = It is categorical and Nominal in nature.

Segment 3: Data Preprocessing & Feature Engineering

- **Dealing with Null values -** In the dataset, Columns Description and CustomerID had null values. But can't impute them because the values in the columns are unique. So I dropped all the rows that had null values.

- **Dealing with Duplicate values -** I removed the duplicate rows from the dataset.

- **Feature engineering -** I have created the Total Price column from columns Quantity & Price. And used this column to calculate the monetary values of the customers.

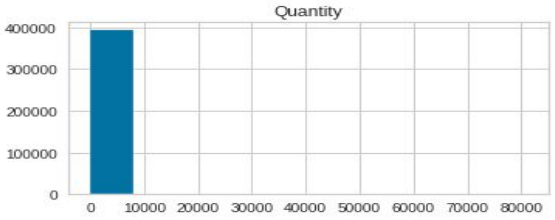
  I have also created the Last Purchase column to calculate the recency values of the customers.

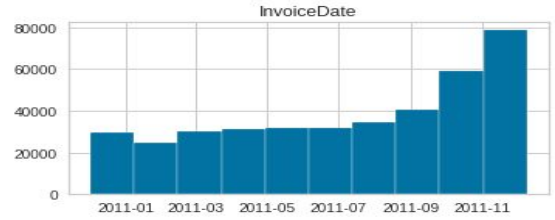
  I have also created various other columns like Year, month, day, and hour for EDA.

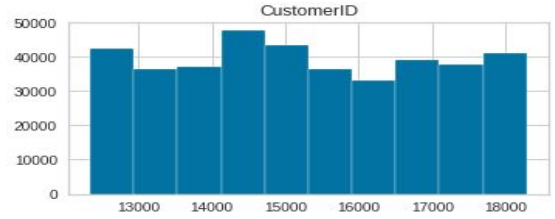
Segment 4: Exploratory Data Analysis

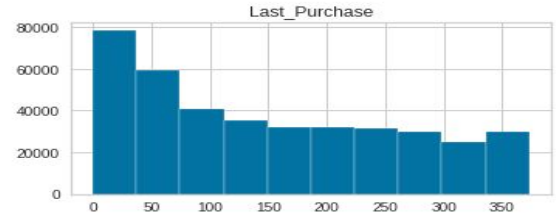
Objective: Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, and patterns using Visualization.

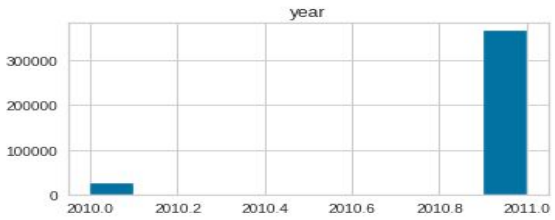
Below are the histograms of the variables to get a better understanding of the data.

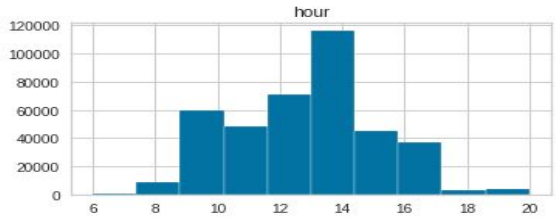






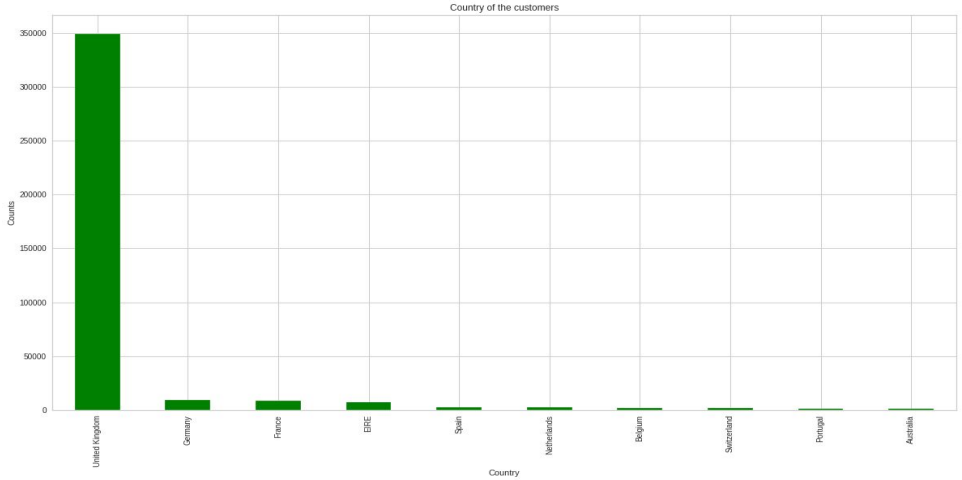


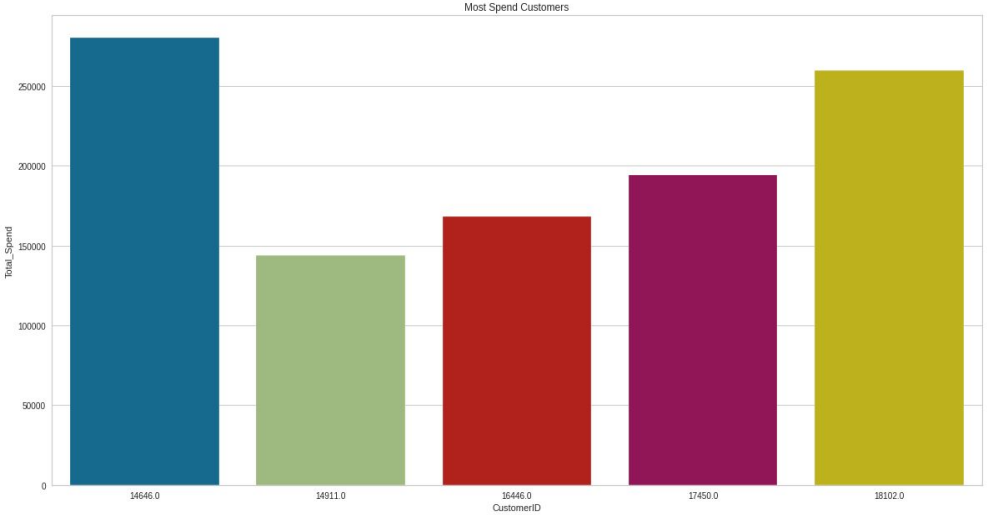




This is a bar plot showing the top 10 countries of the customers who have placed the most orders.

Most of the customers are from UK itself. That makes sense because the store is based out of UK



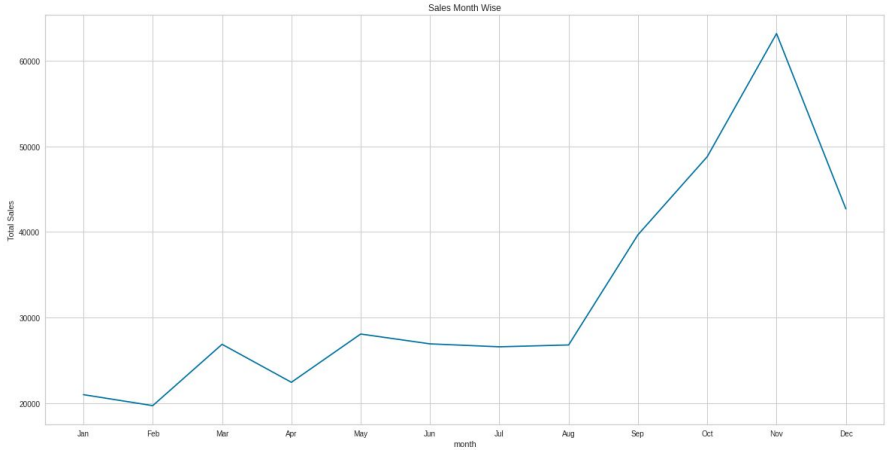


This is also a bar plot showing the customer who has spend the most amount of money with the online e-commerce store.

Customers 14646 & 18102 spent a lot of money buying the e-commerce store products.

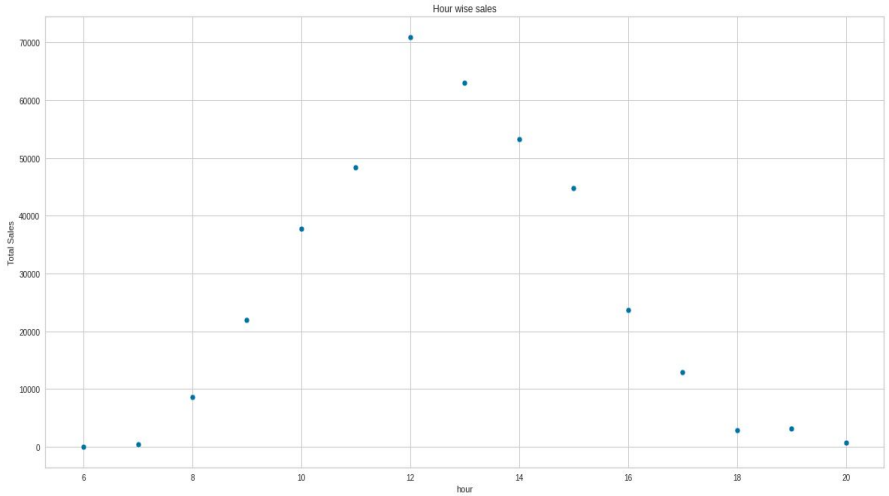
This is a line plot showing the total sales of an online store month wise.

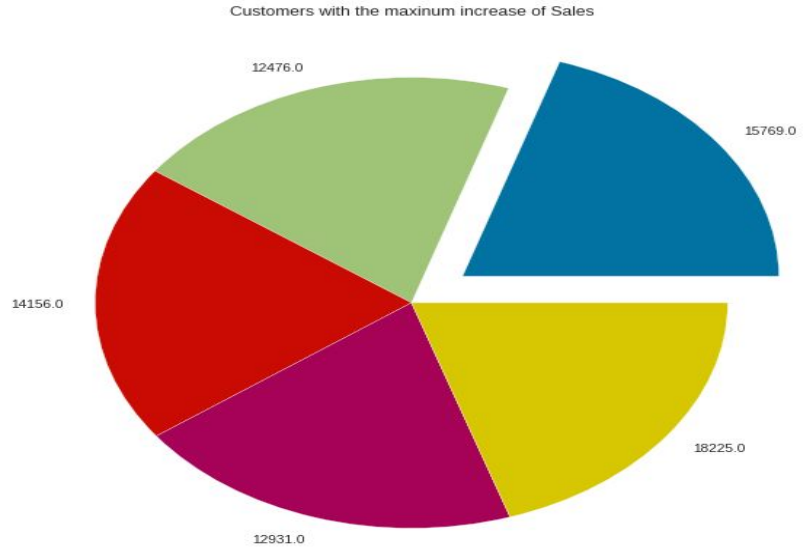
From the line plot we conclude that most of the sales taken place just before the holiday season. The online store might be offering discounts during this period.



This is a scatter plot showing the sales of the products hour wise.

As you can see most of the sales have taken place during the general working hours(9Am to 5Pm)





This is a pie chart showing the top 5 customers with percentage increase in sales from 2010 to 2011

As you can see the percentage increase in sales for all the 5 customers is close with customer 15769 having a marginally better percentage.

Segment 5: RFM Analysis

Recency, frequency, monetary value is a marketing analysis tool used to identify a company's or an organisations best customers by using certain measures. The RFM model is based on three quantitative factors:

- **Recency**: When was the last purchase made.
- **Frequency**: How often a customer makes a purchase.
- **Monetary**: How much money a customer spent between a time interval.

RFM process.

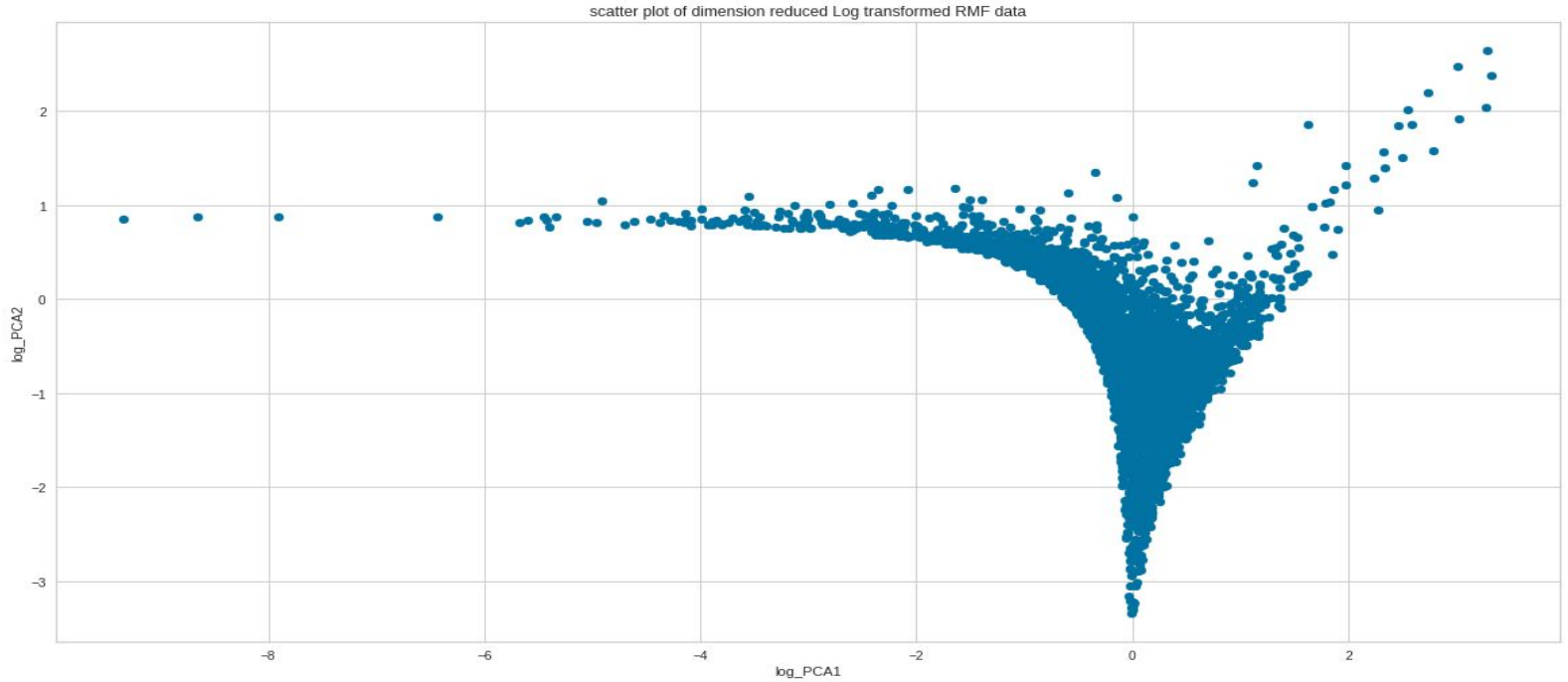
- The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer.
- The second step is to identify a threshold for Recency, Frequency, and Monetary values for encoding(A common way of deciding the threshold is through percentile values).
- The third step is to add the encoded Recency, Frequency, Monetary values to obtain RFM scores.

Under this segment, I have assigned each customer a Recency, Frequency, and Monetary value and then added them up to obtain an RFM score. The Higher the score the better the customer. After analysing the RFM scores I segmented the customers into 3 types:

- Important Customers: All those customers whose RFM score is more than 10(75th percentile).
- Modest Customers: All those customers whose RFM score is between 7(50th percentile) and 10(75th percentile).
- Irrelevant Customers: All those customers whose RFM score is less than 7(50the percentile).

Segment 6: Hypothesis on the number of clusters

To visually identify the clusters I plotted a 3-D scatter plot for Recency, Frequency, and Monetary values but I could not identify any. So I used PCA to reduce the dimensions to two. After reducing the dimensionality and performing log transformation on the PCA data(because the data points were densely packed) I made a visual hypothesis of 3 clusters. I was partially correct with my hypothesis because K-means clustering using the K-Elbow method suggested defining the value of K as 3.



Segment 7: Model 1-K-means Clustering using the K-Elbow method

K-means Clustering: The *k*-means algorithm searches for a predetermined number of clusters within an unlabelled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:

- The "cluster centre" is the arithmetic mean of all the points belonging to the cluster.
- Each point is closer to its own cluster centre than to other cluster centre.

Those two assumptions are the basis of the *k*-means model.

Expectation–maximisation (E–M) is a powerful algorithm that comes up in a variety of contexts within data science. **k*-means is a particularly simple and easy-to-understand application of the algorithm.

In short, the expectation–maximisation approach here consists of the following procedure:

- Guess some cluster centres
- Repeat until converged
  E-Step: assign points to the nearest cluster centres
  M-Step: set the cluster centres to the mean

Here the "E-step" or "Expectation step" is so-named because it involves updating our expectation of which cluster each point belongs to.

The "M-step" or “Maximisation step" is so-named because it involves maximising some fitness function that defines the location of the cluster centres — in this case, that maximisation is accomplished by taking a simple mean of the data in each cluster.

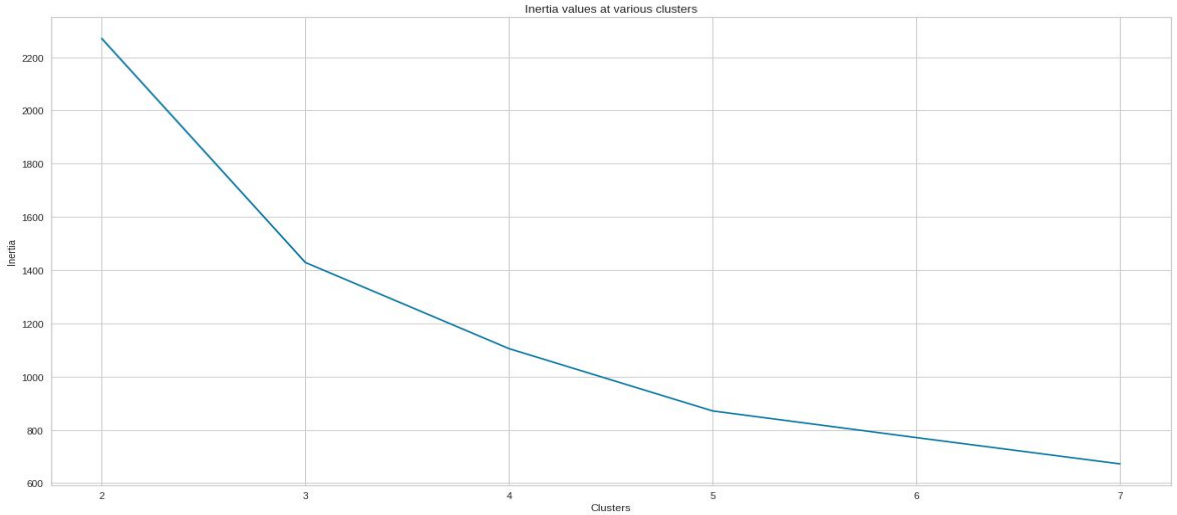
Scaling & Log Transformation

I have rescaled the data using Standard Scaler because distance-based algorithms are always biased towards variables that have a higher range of values. I have also log-transformed the input data as the data points were densely packed.

Identifying the optimal number of clusters

To determine the optimal number of clusters, we have to select the value of k at the “elbow” i.e. the point after which the inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 3.

Inertia Meaning: It is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster.



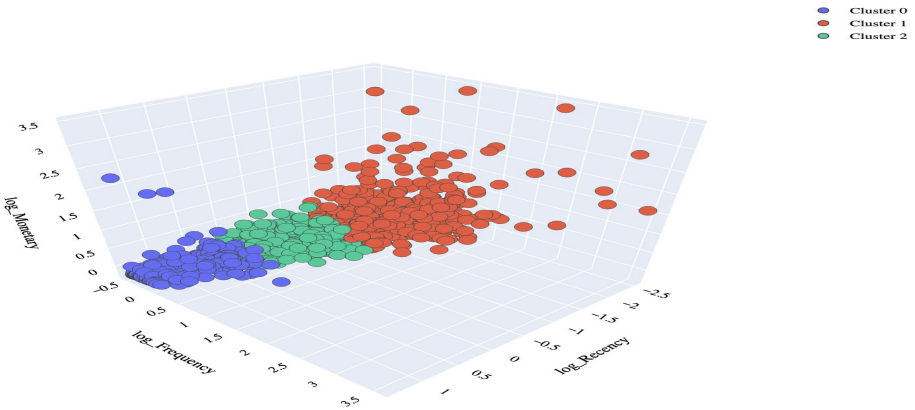
Implementing & Interpretation of the results:

After determining the optimal number of clusters as 3 using the K-Elbow method I have implemented the K-means clustering model on log transformed Recency, Frequency, Monetary values.

After implementation, the model returned us with cluster 0, cluster 1 & cluster 2(3 clusters).

- Cluster 1: Customers of this cluster were considered important because the monetary median and frequency median values were high and the recency median value was low.
- Cluster 2: Customers of this cluster were considered modest because the monetary median and frequency median values were moderate and the recency median value was not too high.
- Cluster 0: Customers of this cluster were considered irrelevant because the monetary median and frequency median values were low and the recency median value was high.

Note: I have used median values of Recency, Frequency, Monetary to identify the clusters since the distributions of recency, frequency, and monetary are all skewed.



Segment 8: Model 2-K-means Clustering using the Silhouette Scores

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observations belonging to all the clusters:

- Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.
- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.

The Silhouette Coefficient for a sample is $S = (b-a) / \max(a,b)$.

The Silhouette Coefficient value ranges from -1 to 1.

1: This Means clusters are well apart from each other and clearly distinguished.

0: This Means clusters are indifferent, or we can say that the distance between clusters is not significant.

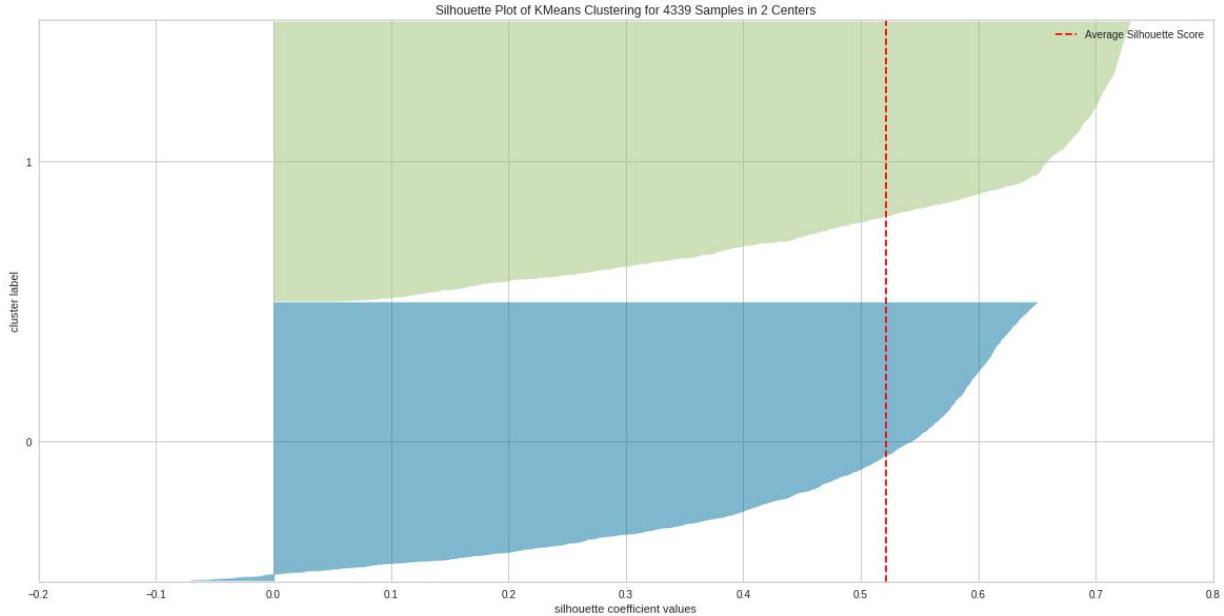
-1: This Means clusters are assigned in the wrong way.

Scaling & Log Transformation

I have rescaled the data using Standard Scaler because distance-based algorithms are always biased towards variables that have a higher range of values. I have also log-transformed the input data as the data points were densely packed.

Identifying the optimal number of clusters

To determine the optimal number of clusters, We have to select the cluster for which the silhouette score is close to 1. Thus for the given data, we conclude that the optimal number of clusters for the data is 2.



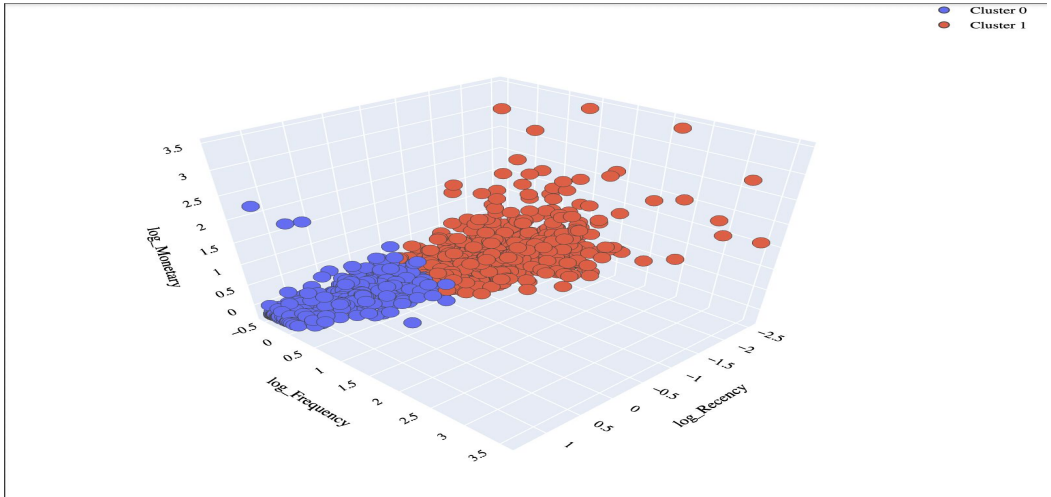
Implementing & Interpretation of the results:

After determining the optimal number of clusters as 2 using the silhouette method I have implemented the K-means clustering model on log transformed Recency, Frequency, Monetary values.

After implementation, the model returned us with cluster 0, cluster 1(2 clusters).

- Cluster 1: Customers of this cluster were considered important because the monetary median and frequency median values were high and the recency median was low.
- Cluster 0: Customers of this cluster were considered irrelevant because the monetary median and frequency median values were low and the recency median value was high.

Note: I have used median values of Recency, Frequency, Monetary to identify the clusters since the distributions of recency, frequency, and monetary are all skewed.



Segment 9: Conclusion.

After implementing K-means clustering using K-Elbow & Silhouette methods, I came to the conclusion that the optimal number of clusters should be 3 because it provides the opportunity to convert modest customers into important customers unlike when K = 2.

Thank you.