Capstone Project-4

Online Retail Customer Segmentation

Team Members

Adityasingh Thakur
Meenakshi
Tushar R. Wagh

Let's Predict!

1. Introduction and Problem statement
2. Objective
3. Data Description
4. Data Cleaning
5. EDA (Exploratory Data Analysis)
6. Data Transformation by using RFM (Recency Frequency Monetary) model
7. Building model
    a) K-Means Clustering with Silhouette analysis
    b) K-Means Clustering with Elbow method
    c) Agglomerative Hierarchical Clustering with dendograms
7. Clustering Profiling
8. Types of linkages in Hierarchical Clustering with dendograms
9. Conclusion

Introduction & Problem Statement

Introduction:-

Business all over the world are growing today. With the help of technology,they have access to a wider market and hence, a large customer base.Customer Segmentation refers to categorizing customers into different groups with similar characteristics .It can help businesses focus on each customer group in different way ,in order to maximize benefits for customers as well as for the business.
This project is mainly deals in segmenting the customers of a online retail stores in UK.

Problem Statement:-

In this project,your task is to identify major customer segments on a transactional dataset which contain all the transactions occuring between 01/12/2010 and 09/12/2011 for a UK based and registered non-store online retail. The company mainly sells unique all-occasion Gifts.Many customers of the company are wholesalers.

Objective

- The main objective of our project is to segment customers into different groups on the basis of their similarities in same group and difference in other group.
- The members of one group is different from other cluster on the basis of their properties and nature.

Data Summary

Understanding the attributes of the dataset better:

This dataset contain transactional data of online retail store. It contain 541908 rows and 8columns.

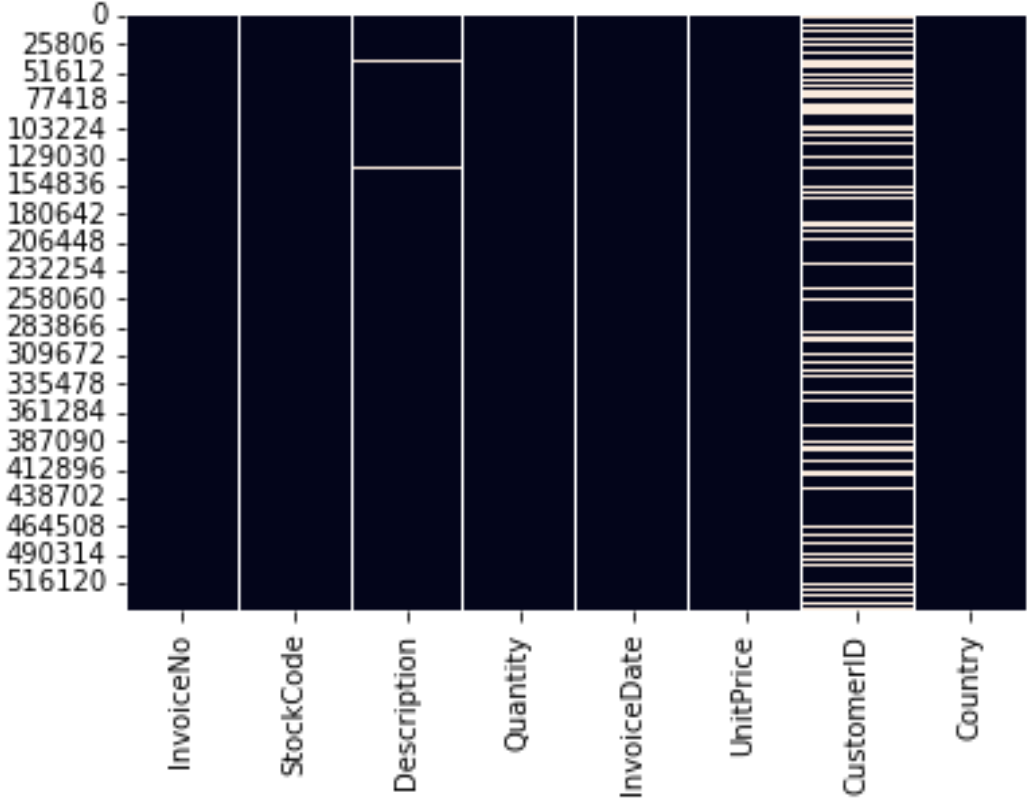
- **InvoiceNo(Invoice Number):** Nominal, a 6-digit integral number uniquely assigned to each transaction. If code starts with letter ‘c’ ,it indicates a cancellation.
- **StockCode:** Product(item) code (Nominal), a 5-digit integral number uniquely assigned to each distinct product.
- **Description:** 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.
- **Unit Price:** 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 country where customer resides.

Data Cleaning

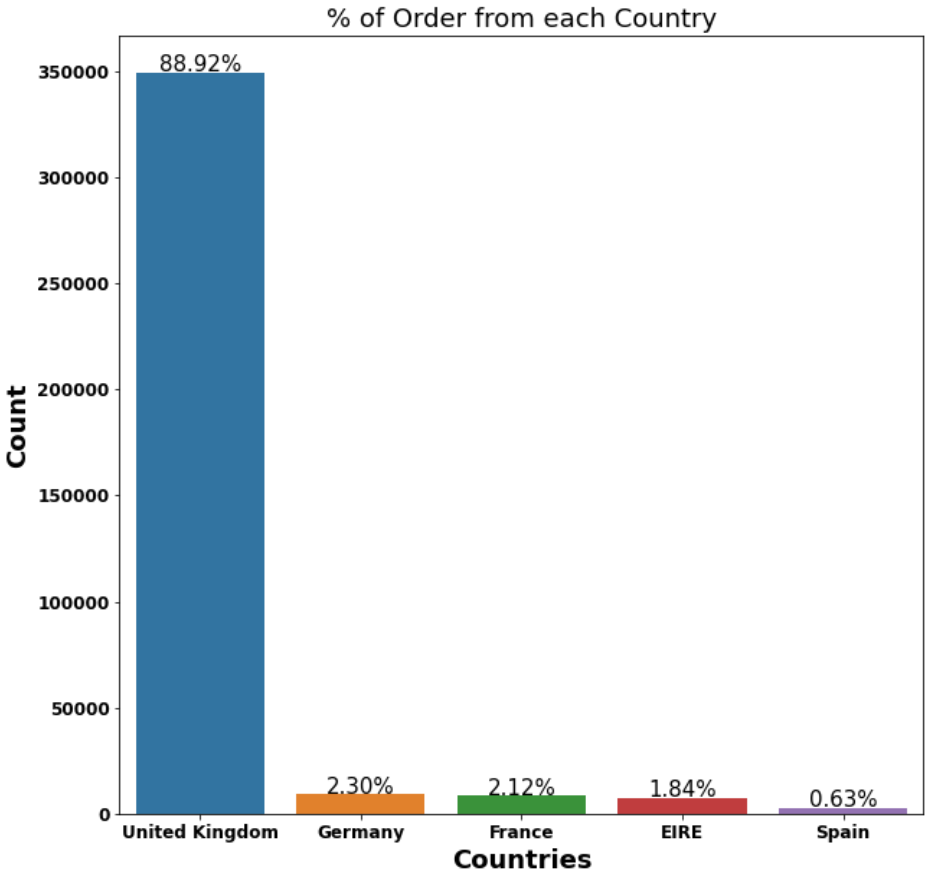
- There are null values present in the ‘CustomerID’ and ‘Description’ column. These have to be dropped as there is no way of filling them strategically.
- We need to remove the cancelled order exist in the dataset.
- Date, Month and Year are extracted from the ‘InvoiceDate’ column. We have created new column for Date, Month and Year.
- Outliers in the ‘Recency’ , ‘Frequency 'and ‘Monetary’ column have been removed.
- We have removed some duplicate values.

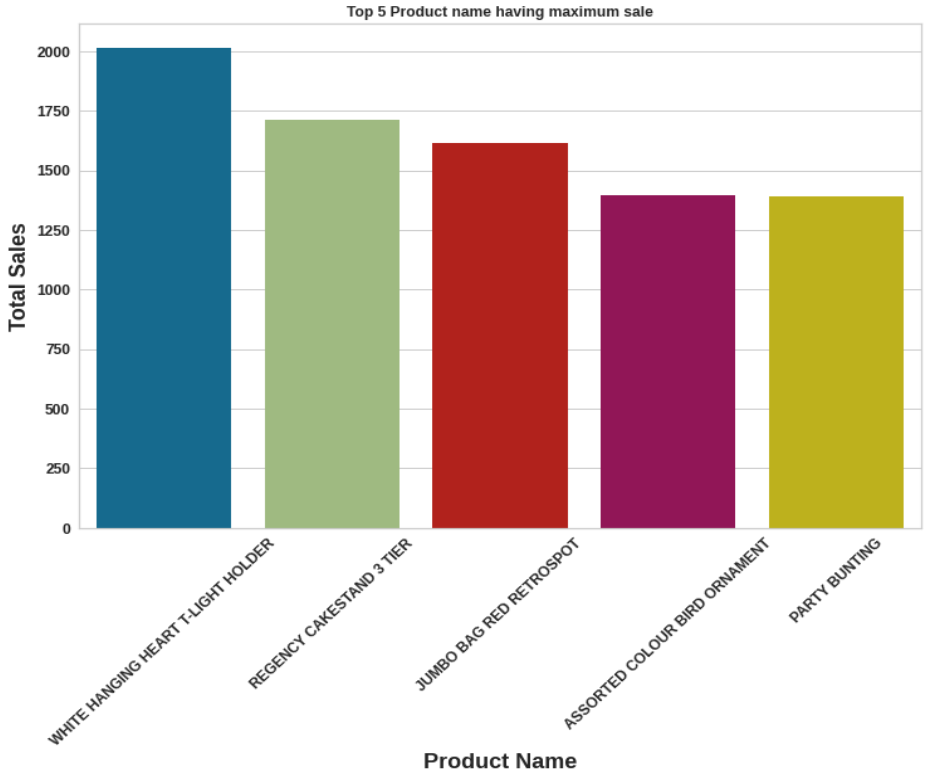
Data Cleaning(Continued…)

- Heat map to show null values in ‘Description’ and ‘CustomerID’ column.



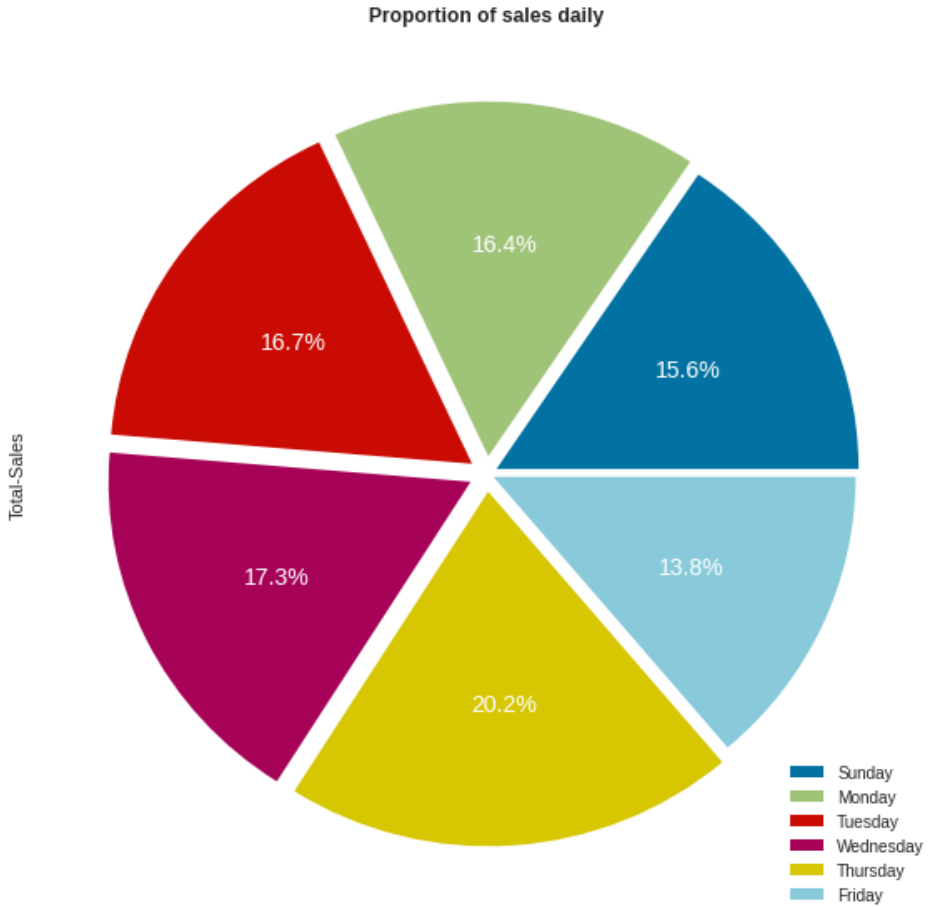
Exploratory Data Analysis(EDA)



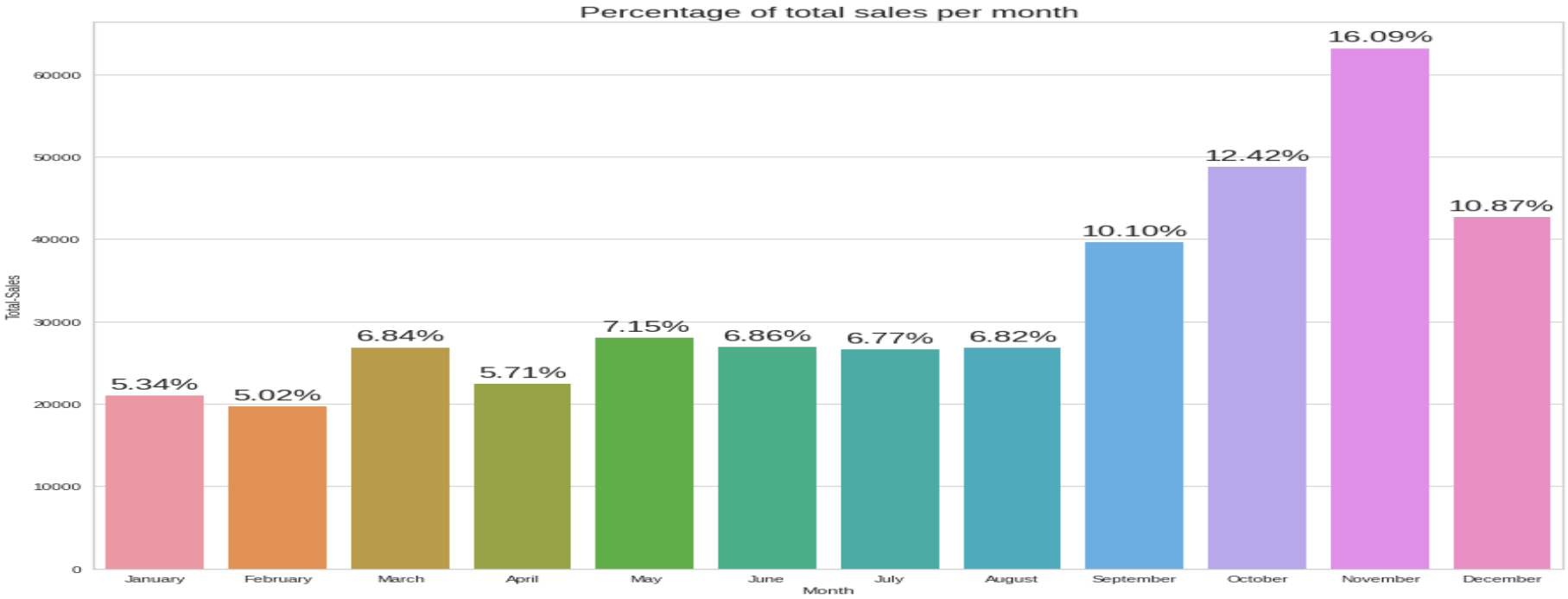


EDA(Continued...)

- There is no sale on Saturday.
- Thursday have highest sale followed by Wednesday and Tuesday.

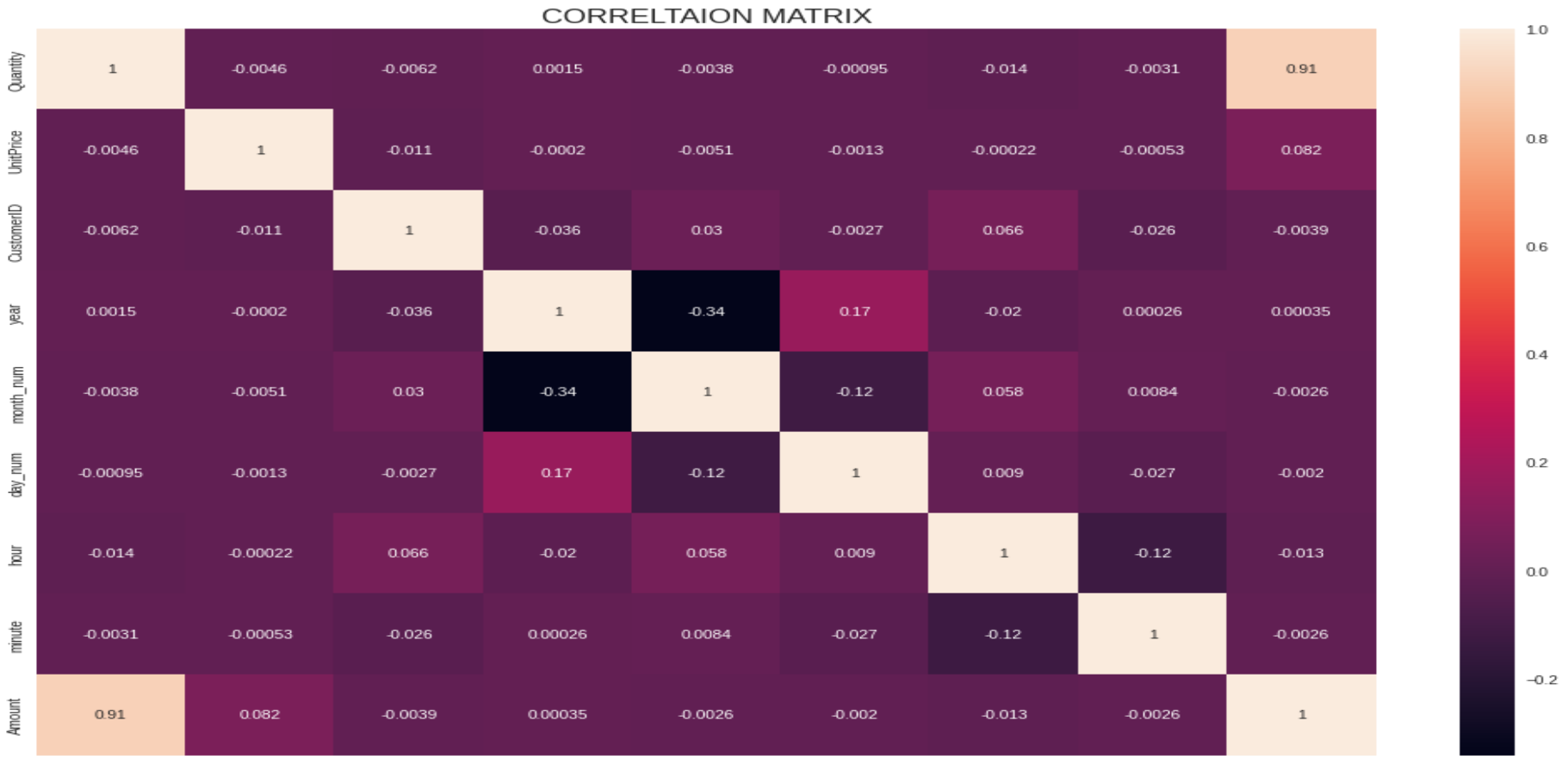


EDA(Continued…)



- November have highest sale followed by October and December.
- February have least sale.

Heat Map for whole dataframe



Data Tranformation

- RFM segmentation allows marketers to target specific clusters of customers with communications that are much more relevant for their particular behavior – and thus generate much higher rates of response, plus increased loyalty and **customer lifetime value.**

- RFM segmentation is a powerful way to identify groups of customers for special treatment. RFM stands for recency, frequency and monetary

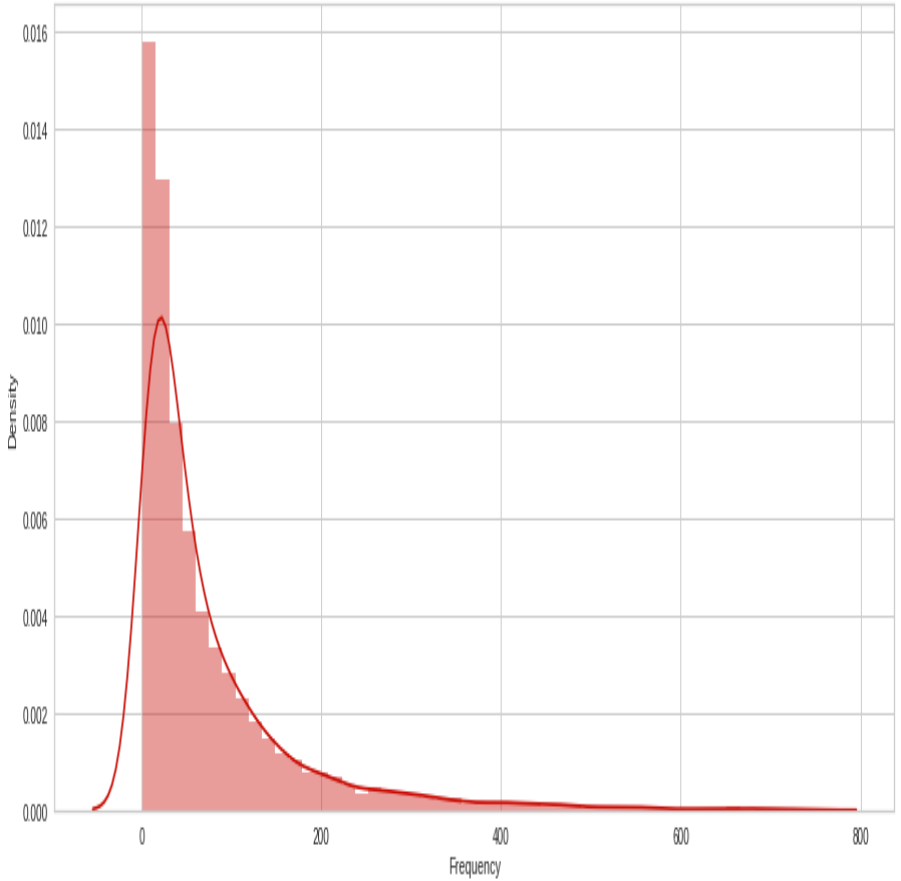
- Recency signifies the days since order , frequency signifies the number of times the customer is been billed and monetary signifies the sales each customer has provided.

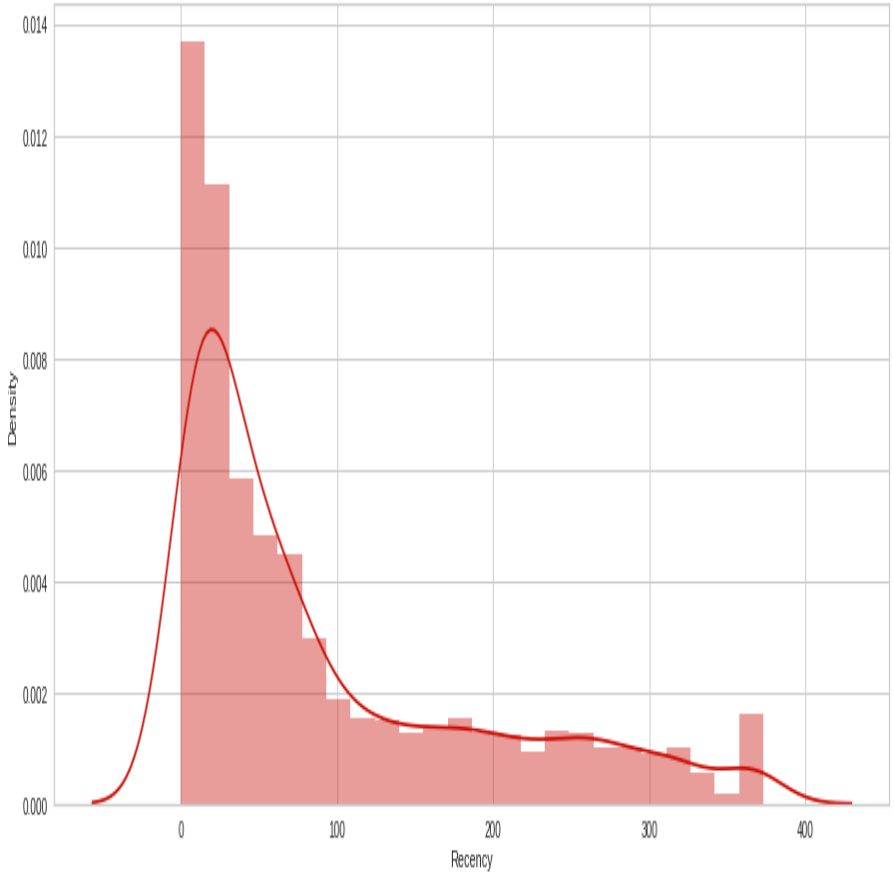
- It can be seen that, frequency and monetary variables have a linear trend and frequency of orders have been high recently.

- The RFM dataframe is grouped on the basis of Customer ID.The data now contains 4192 rows or customers.

- The ranges in the data differ , hence , the data is also scaled using a Standard Scaler.

Distribution plot for Recency , Frequency





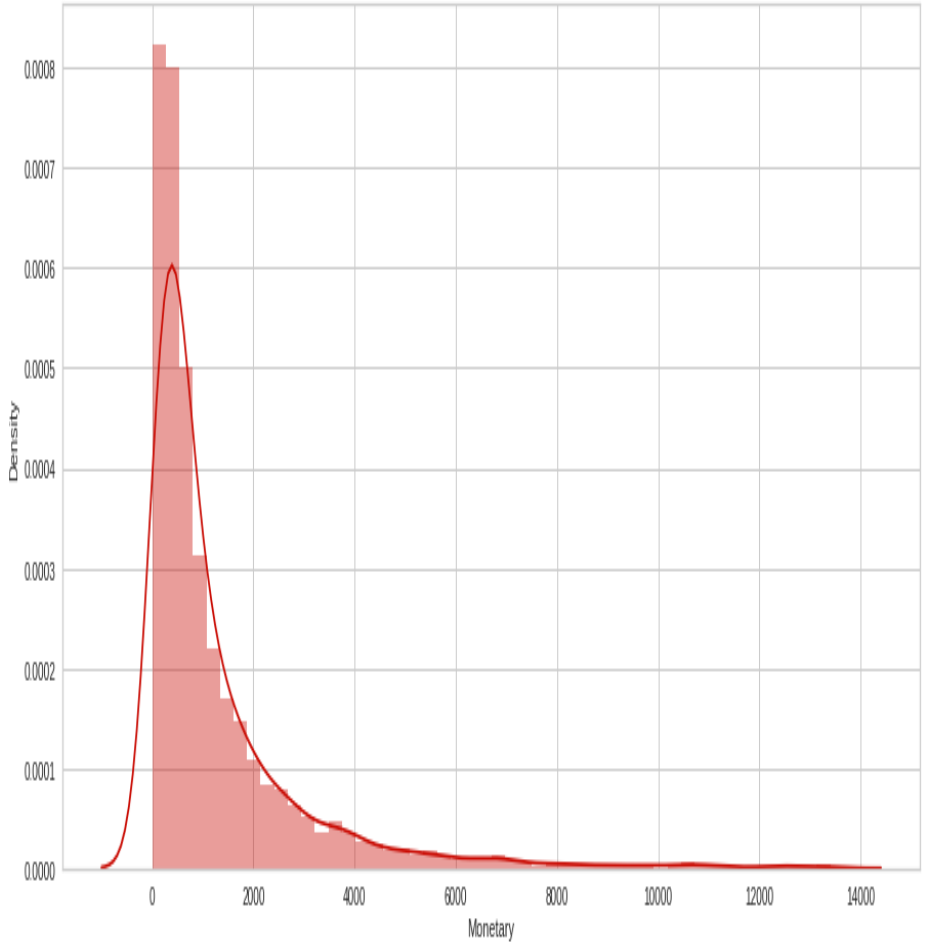
Distribution plot for monetary

From all distribution plot, we conclude that:

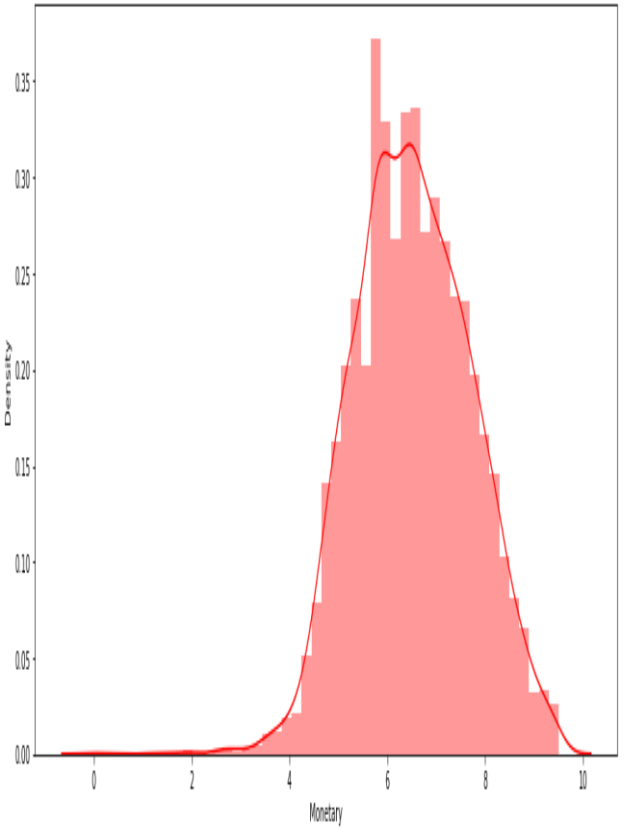
- All the displots are right skewed.

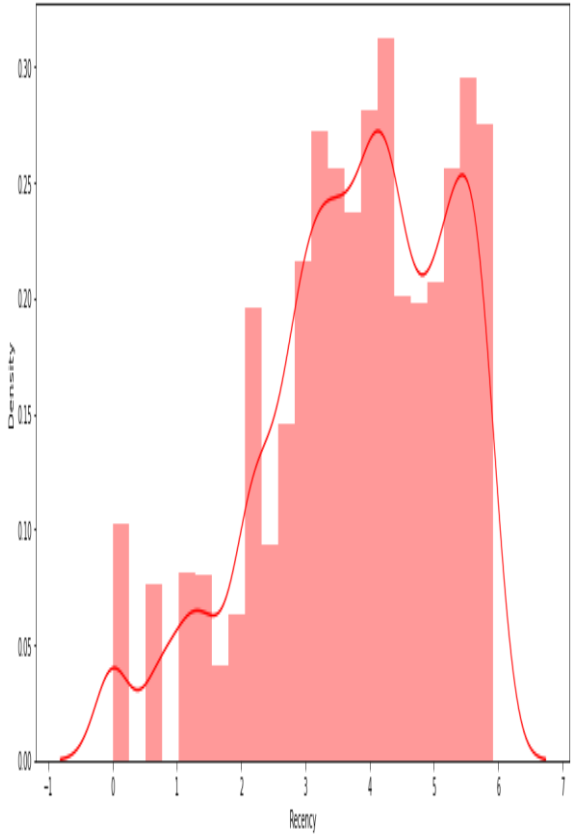
- From recency displot,we can say that there are huge no. of customers who are purchasing the product frequently.

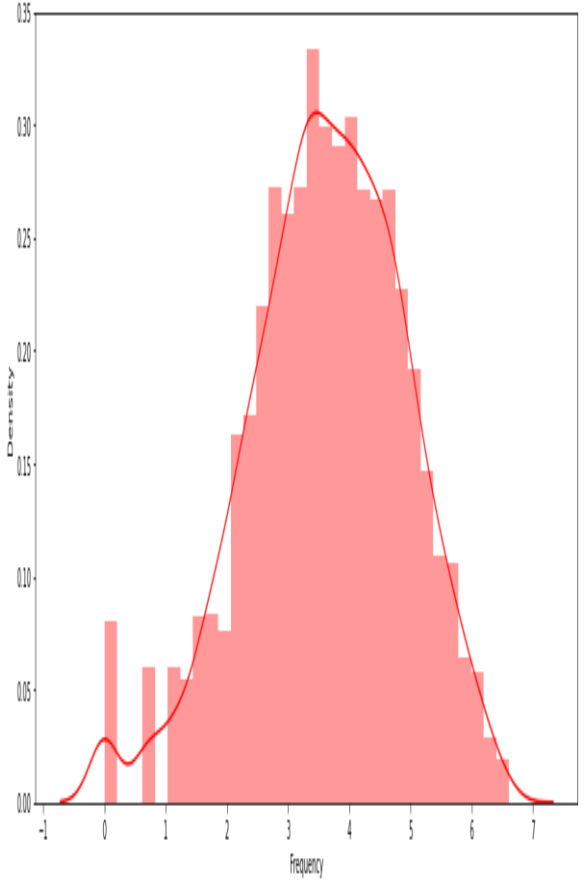
- Displots are not a normal distribution.



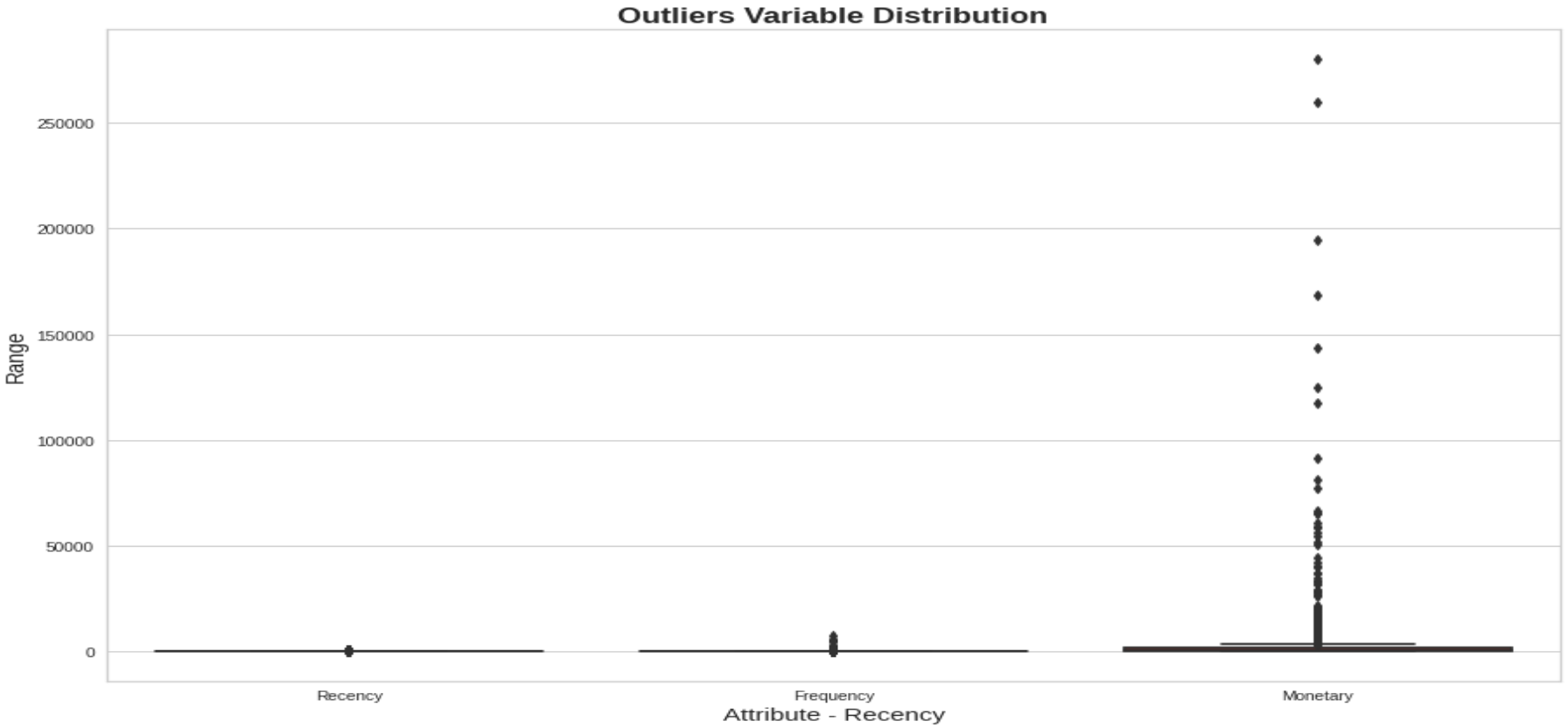
Plot after log transformation



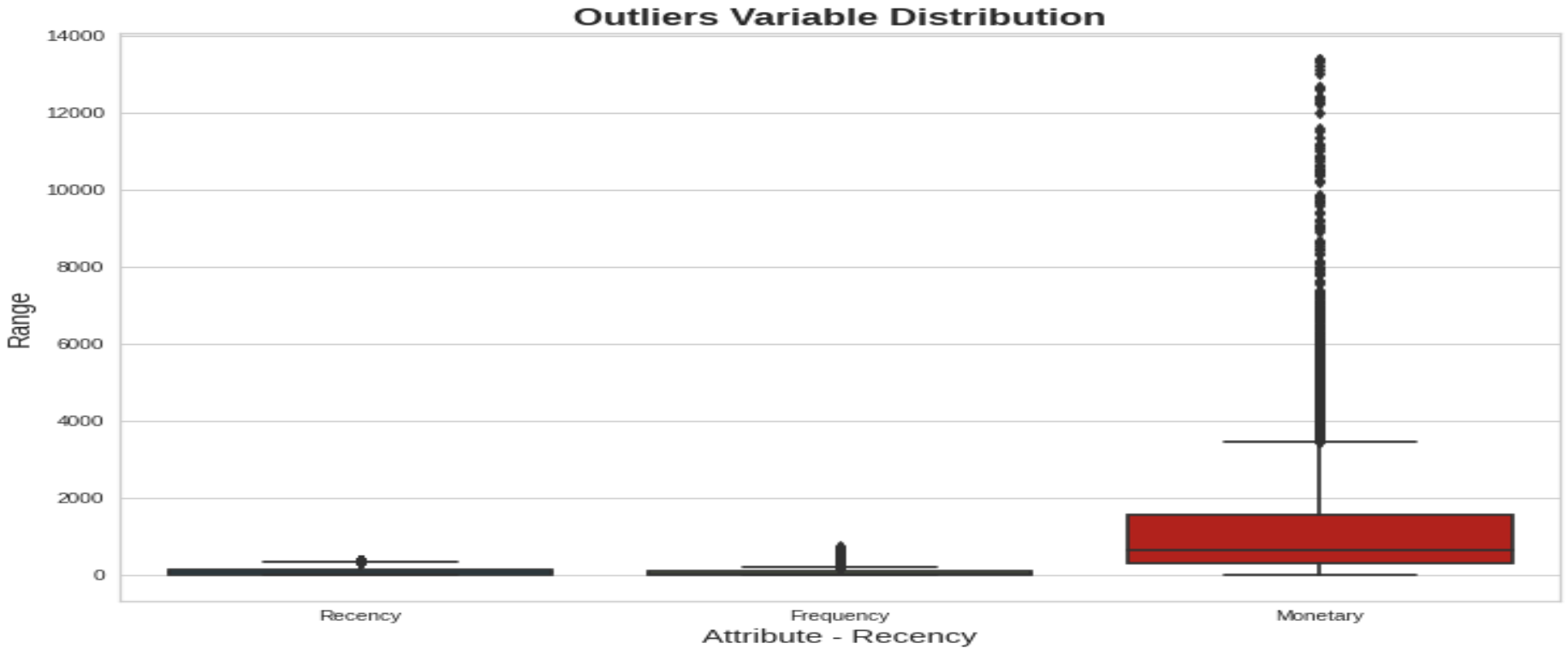




Checking outliers using box plot in recency , frequency and monetary values



After removing outliers using box plot in recency , frequency and monetary values



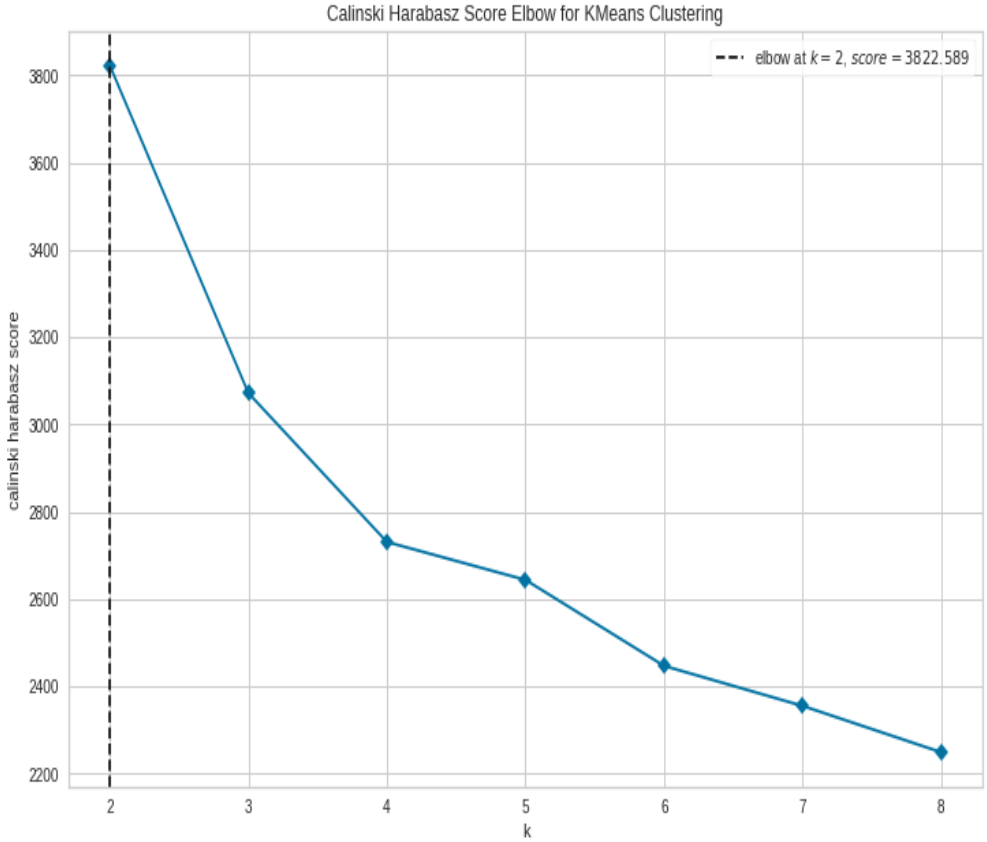
K-Means Clustering Model

- K-Means algorithm is used to cluster the customers into different segments.
- To identify optimal no. of clusters,we have use the Elbow method and Silhouette analysis.
- Find the silhouette score and get optimal clusters
- With both the method 2 is optimal no. of cluster.
- Below is the table to show silhouette score for different clusters.
- We have found that when no. of cluster is 2 and silhouette score is maximum i.e. 0.39408

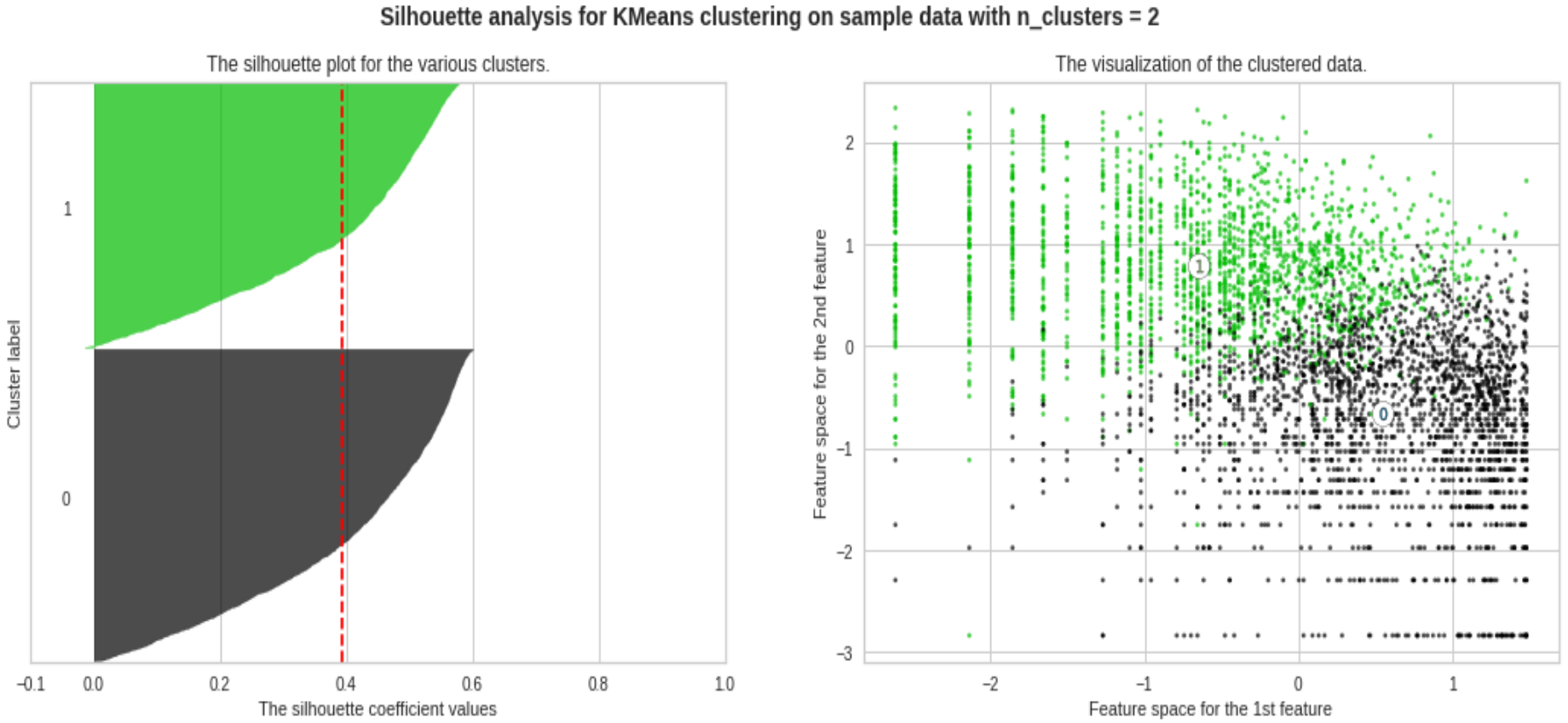
```
For n_clusters = 2 The average silhouette_score is : 0.3940810337905449 3
For n_clusters = 3 The average silhouette_score is : 0.2947593636511543 5
For n_clusters = 4 The average silhouette_score is : 0.2975051811313832
For n_clusters = 5 The average silhouette_score is : 0.2829104892251716 5
```

Elbow method

- This graph is line plot to find the optimal no. of cluster using elbow method.
- The no. of optimal clusters is 2.



Silhouette Analysis(with optimal no. of clusters)



Cluster Profiling

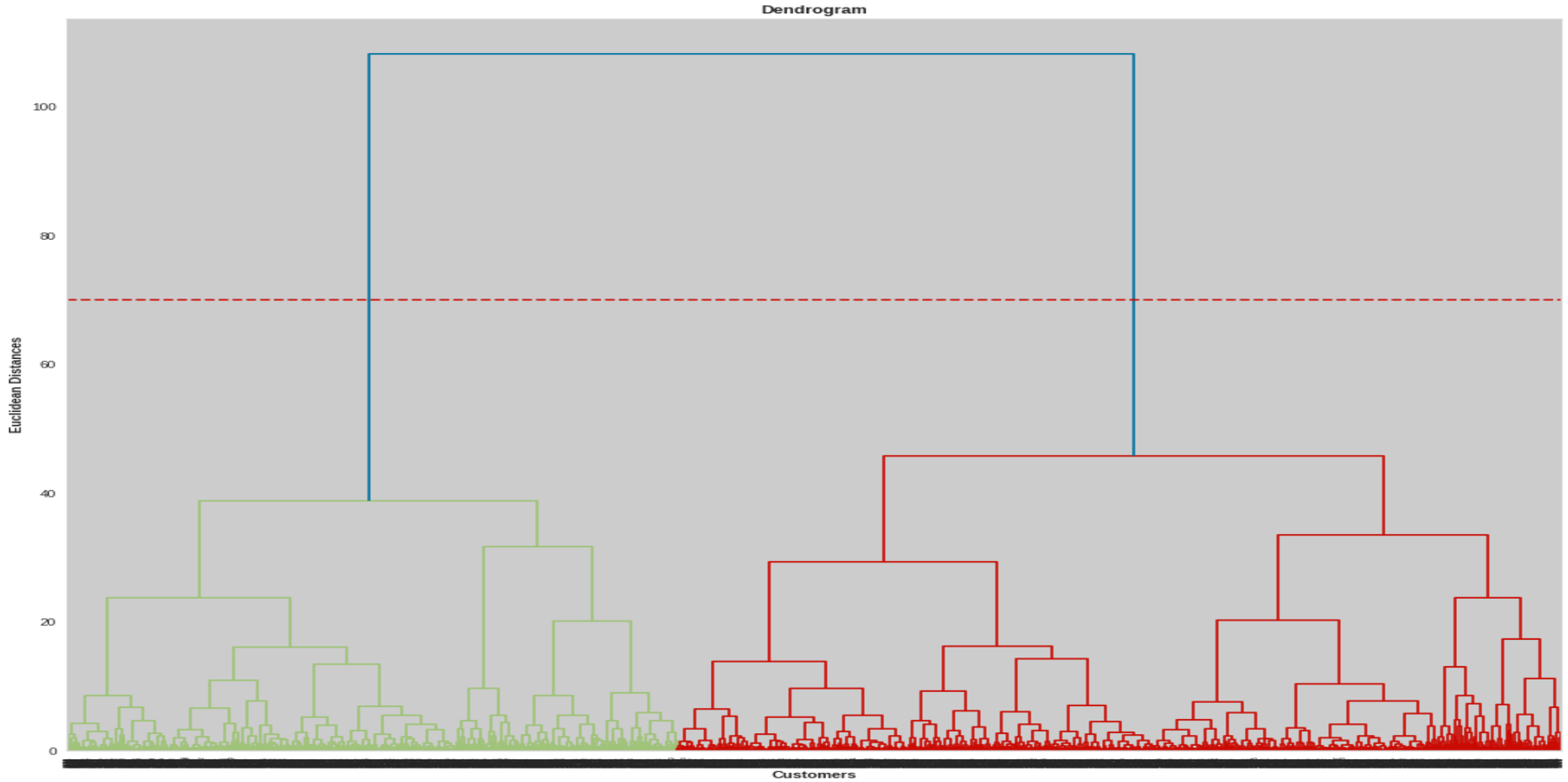
Cluster	Recency	Frequency	Monetary	R	F	M	RFMScore	Recency_log	Frequency_log	Monetary_log
0	143.480694	23.199566	411.099827	3.101518	3.307592	3.278525	9.687636	4.570699	2.775255	5.724485
1	34.482829	137.630446	2365.316854	1.763711	1.574577	1.580215	4.918503	2.844539	4.632407	7.450798

- **Wholesale Customer:'Cluster 1'** is the high value customer segment as the customer in this group place the highest value orders with a very high frequency than other members.They are also the one who have more no. of transactions.These are the wholesale customer of retail store.
- **Average Cutomer:'Cluster 0'** is the average customer segment.These customers order less frequently as compared to wholesale customer and their orders are pretty low valued.

Agglomerative Hierarchical Clustering

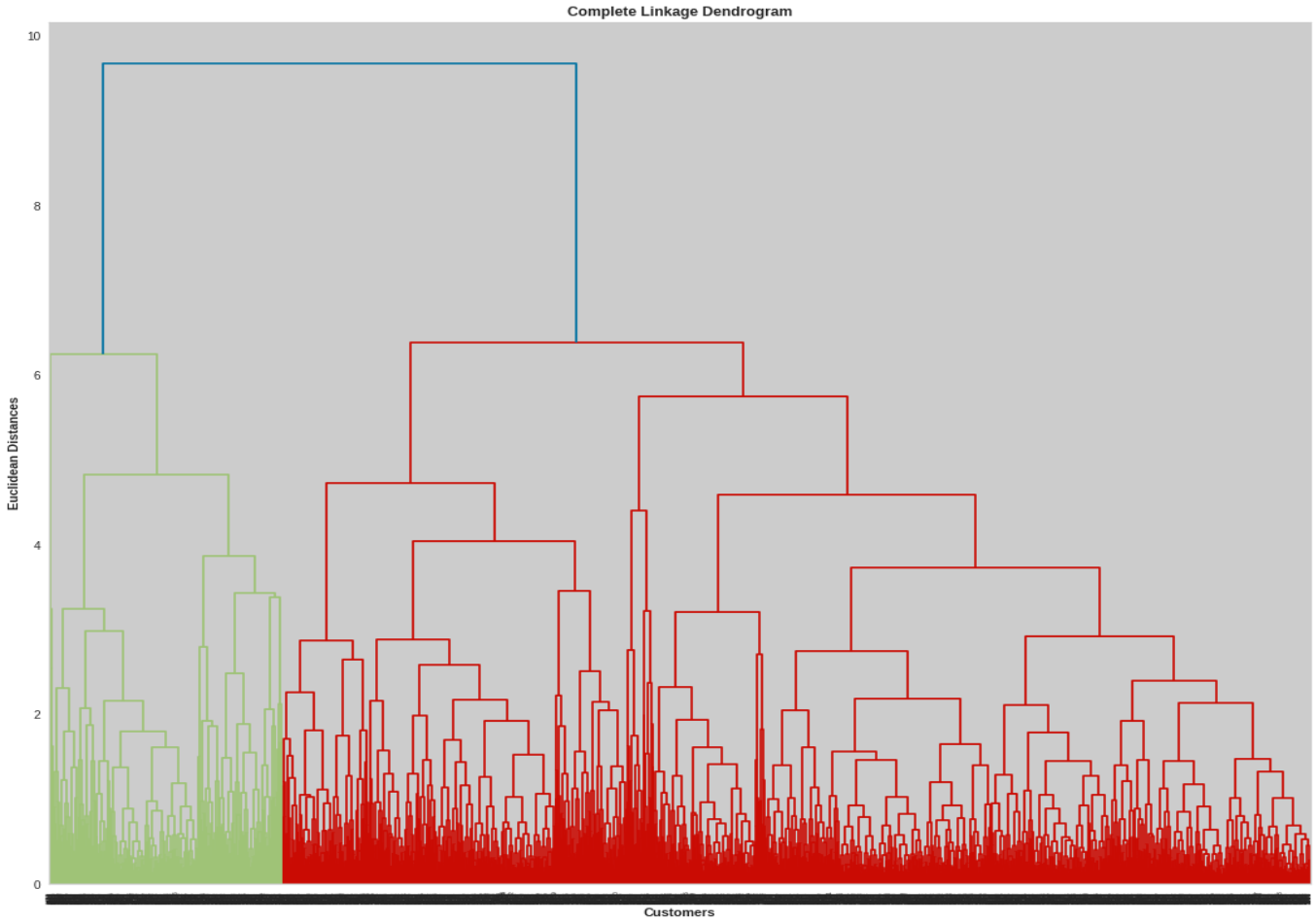
- It a **bottom up** approach.
- In this technique,we assign each point to an individual cluster in the beginning.
- Then,we merge the closest pair.And,finally we got a single cluster.
- This technique is also known as Additive Hierarchical Clustering.
- We have used dendogram to find optimal no. of clusters.
- Dendogram is a tree-like diagram that records the sequences of merge or splits.More the distance of vertical lines in the dedogram,more the distance between those clusters.
- We can set a threshold distance and draw a horizontal line(Generally,we try to set the threshold line in such a way that it cuts the tallest vertical line).
- The no. of clusters willl be the no. of vertical lines which are being intersected by the line drawn using the threshold.

Dendogram with 2 as optimal no. of clusters



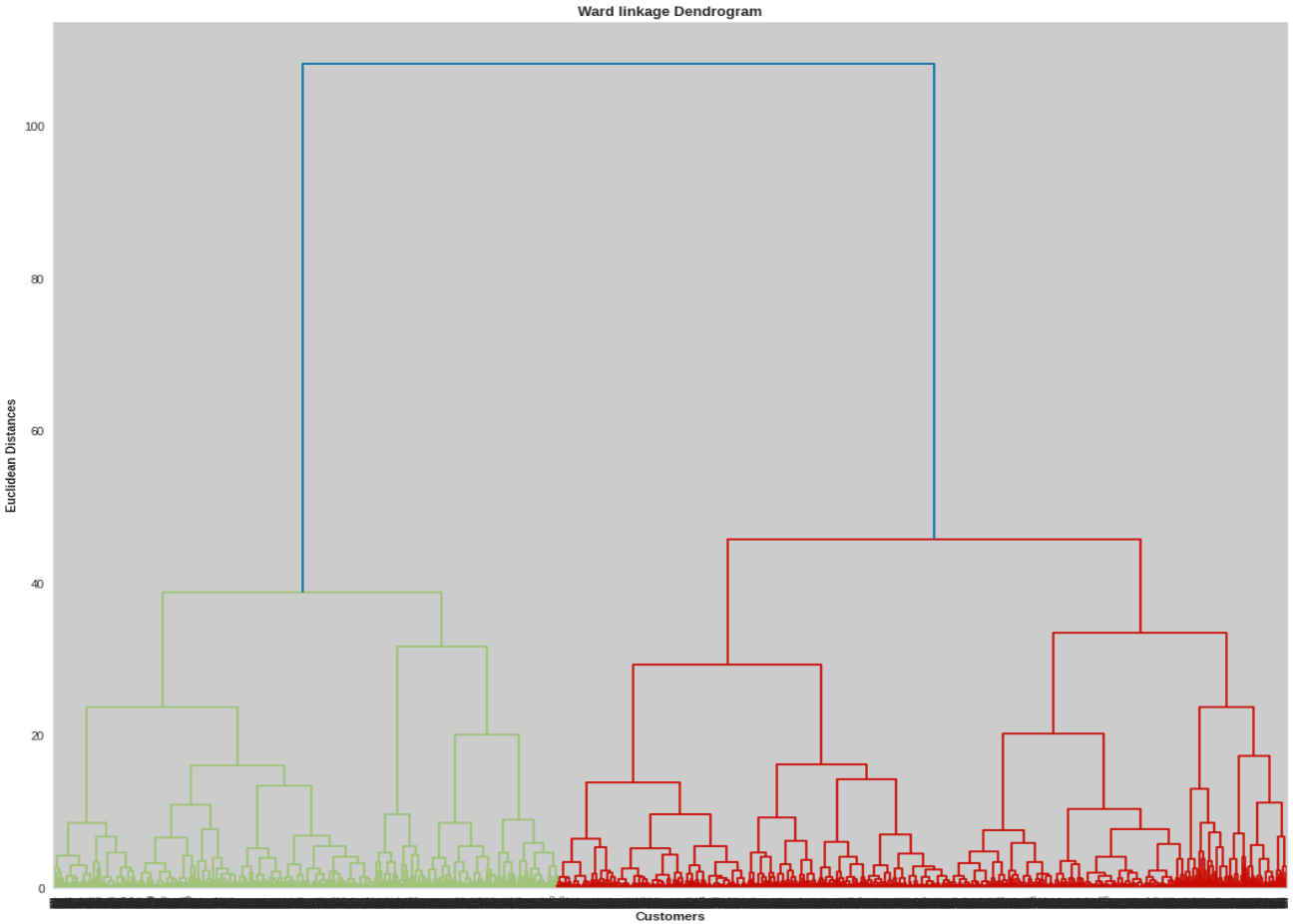
Complete linkage

- In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster.



Ward Linkage

- In ward linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.



Pretty table

Sl No.	Model_Name	Data	Optimal_Number_of_cluster
1	K-Means with silhouette_score	RFM	2
2	K-Means with Elbow method	RFM	2
3	Agglomerative Hierarchical Clustering with threshold value 70	RFM	2

Conclusion

- Throughout the analysis we went through various steps to perform customer segmentation.We start with data wrangling in which we tried to handle null values, duplicates and performed feature modifications.
- We have used dendogram to find optimal no. of clusters.
- Dendogram is a tree-like diagram that records the sequences of merge or splits.More the distance of vertical lines in the dedogram,more the distance between those clusters.
- We casn set a threshold distance and draw a horizontal line(Generally,we try to set the threshold line in such a way that it cuts the tallest vertical line).
- Using cluster profiling the average of recency, frequency and monetary values for each customer segment was identified.

Conclusion

- We have made dendograms for all types of linkages i.e. Single , Average ,Centroid , Complete and ward.
- Ward is most popular among all linkages.
- The optimal no. of clusters in K-Means with Silhouette Analysis and elbow method is 2.
- We have used Agglomerative clustering with different threshold value and see how clusters differ and find optimal no. of clusters.
- The optimal no. of clusters with threshold value 70 in Agglomerative Clustering is 2.

Thank you