**ONLINE RETAIL CUSTOMER**

**SEGMENTATION**

**1. Ms.Madhuri Bonela**

**2. Mr.Pratik Ghodke**

DataScience Trainee Almabetter,Bangalore

# INTRODUCTION

Customer segmentation is the process of classifying customers based on their shared behaviour or other attributes. The groups should be homogeneous within them and should also be heterogeneous to each other. The main goal is to identify customers that are most profitable and loyal and the ones who churned out, to prevent further loss of customers by redefining company policies. Having a large number of customers, each with different needs it is difficult to find which customer is most important for business and target them with an appropriate strategy.

Effective decisions and stratergies to be taken are mandatory for a company to grow and result in good revenue outputs. In these days competition is huge and all companies are moving forward with their own different strategies. We should analyise data and take a proper decision. Every

person is different from one another and we don’t know what he/she buys or what their likes and requirements. But, with the help of machinelearning technique one can sort out the data and can find the targetgroup by applying several algorithms to the dataset. Without this, It will bevery difficult and no better techniques are available to find the group of people with similar character and interests in a large dataset.

Here, The customer segmentation using various techniques to know the customer likes and their purposes so that we can apply it on business area to develop the business and it's growth interms of business.

# PROBLEM STATEMENT

Problem Description

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring 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.

Existing System

The existing methods are storing customer information through paperwork and as a digital data we are storing and analysing it to predict customer benifits and needs and to know how much sales were sold and who are the regular customers and which products are having more demand in sales and what is the profit of business so ,all the required data is out from the data for business improvement so ,storing and data analysis is playing vital role to predict the business up and downs .

# INSPECTING DATASET

1. Importing the dataset and checking the head and tail rows to get an overall idea.
2. After that exploration of the dataset, by checking the info which gives us some intuition about null values and data types of columns present in it.
3. And then we checked the descriptive summary of the data frame which gives some quantitative idea about the dataset i.e., average values of the columns, frequency of the values, variability, and dispersion concerns on how to spread out the values.

# DATA CLEANING

Data Cleansing is the process of detecting and changing raw data by identifying incomplete, wrong, repeated,or irrelevant parts of the

data. In our data, there is a feature called Customer-ID which has more than 24% missing values, hence there is no use in having the data with no customer assignment. So, we dropped it. Coming to the duplication values

We have encountered some duplicated observations in the dataset, when we have frequent duplicates in our database, we may inadvertently send multiple clustering messages to the same person CustomerID. As a Consequence, we dropped it. Finally we ended up with a data frame shape of 401604 records with 8 columns.

We also checked the number of cancellations by each customer, where the InvoiceNo starting with 'C' represents cancellation. We dropped the cancellation data records from the main dataset.

# FEATURE ENGINEERING

* + Extracting the new feature named Year, Month, Day, Hour from the datetime column named InvoiceDate.
  + Creating a new feature 'TotalAmount' by multiplying Quantity and UnitPrice.
  + Creating a newfeatur'TimeType' based on hours to define whether it is morning , afternoon or else the evening

# EXPLORATORY DATA ANALYSIS

After feature engineering, we did some data analysis and drew some hypotheses like -

the United Kingdom has the most number of customers and most numbers purchase history as compared to other countries and Orders with mass quantity are placed by the customers from the Netherlands.

# RFM - RECENCY, FREQUENCY, MONETARY

we can use the RFM-based model for finding segments where R is Recency (how recently a purchase happened), F is Frequency (how frequent transactions are made), and M is Monetary value (Value of all transactions). Recency, Frequency, and Monetary score for each customer is calculated. The latest date is assigned as a placeholder to calculate recent purchases. All the transactions are grouped using CustomerID and then aggregate lambda operations are performed. As a result of this operation, numbers will be obtained which depict

depict the recency., frequency, and how much a specific customer spent to date. All these are stored in a new data frame RFM. Earlier the distributions of Recency, Frequency, and Monetary columns were positively skewed but after applying log transformation, the distributions appear to be symmetrical and normally distributed perception

1. If the RFM of any customer is

444. His Recency is good, frequency is more and Monetary is more. So, he is the best customer.

1. If the RFM of any customer is

111. His Recency is low, frequency is low and Monetary is low. So, he is the churning customer.

1. If the RFM of any customer is

144. He purchased a long time ago but buys frequently and spends more. And so on.

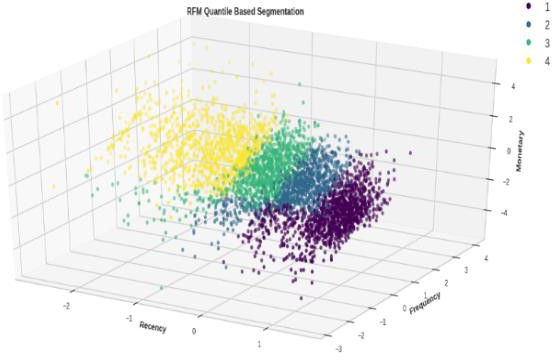
1. Like this we can come up with several segments for all combinations of R, F, and M based on our use case. The higher the RFM score, the more valuable the customer is

# NORMALISATION OF THE DATA

In machine learning, some feature values at times differ from others multiple times. The features with higher values will always dominate the learning process. Before giving our data to clustering algorithms we need to perform the data normalizationtask (i.e StandardScaler which will give equal importance to each variable so that no single variable drives the model performance.

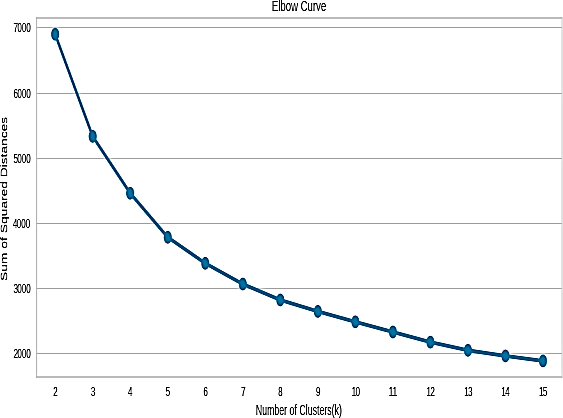
# CLUSTERING

Customer segmentation has been demonstrated to benefit from clustering. Clustering is a sort of unsupervised learning that allows us to locate clusters in unlabelled datasets. Clustering techniques include Binning, Quantile based, K- means, hierarchical clustering, and DBSCAN clustering.



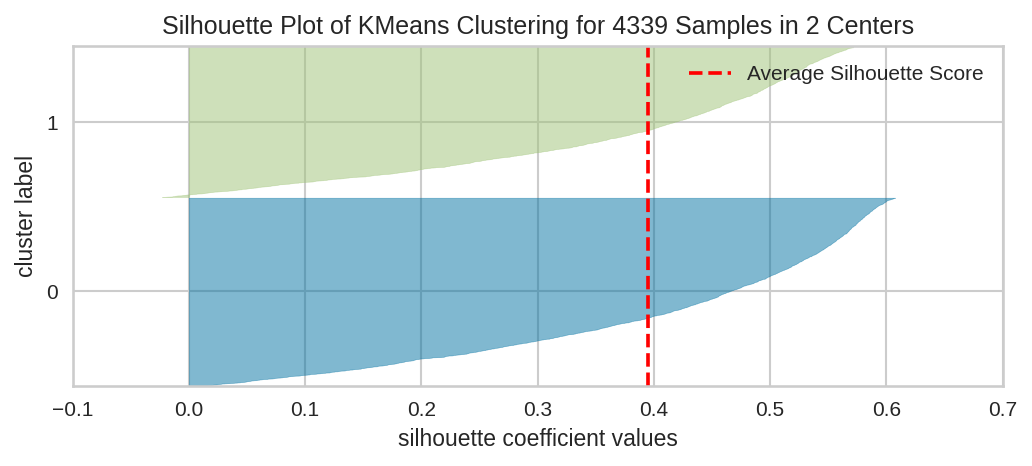
# K-MEANS CLUSTERING

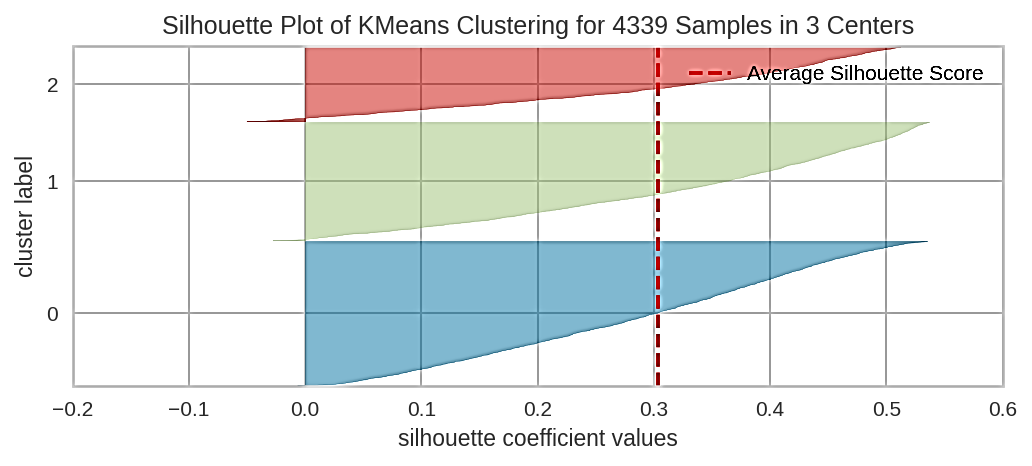
K-Means is an unsupervised learning algorithm used for clustering tasks that work well with complex datasets. It is an iterative algorithm that partitions the dataset into “k” pre-defined non-overlapping subgroups (clusters) where each

data point belongs to only one group.

It is important to determine the optimum number of clusters i.e., “k value”. For this, we used:

**The Elbow method**.

It involves running the algorithm multiple times over a loop with an increasing number of cluster choices and then plotting a score as a function of the number of clusters. When “k” increases, the centroids are closer to cluster centroids. The improvement will decline at some point rapidly creating an elbow-like shape in the graph and that is the whole reason this method is called the elbow. We take the count of the cluster, k-value at the point where this elbow is bending.

**Silhouette scores:** We picked up the range of the k values and drew the silhouette graph by calculating the silhouette coefficient of every point

**SILHOUETTE PLOT OF KMEANS**

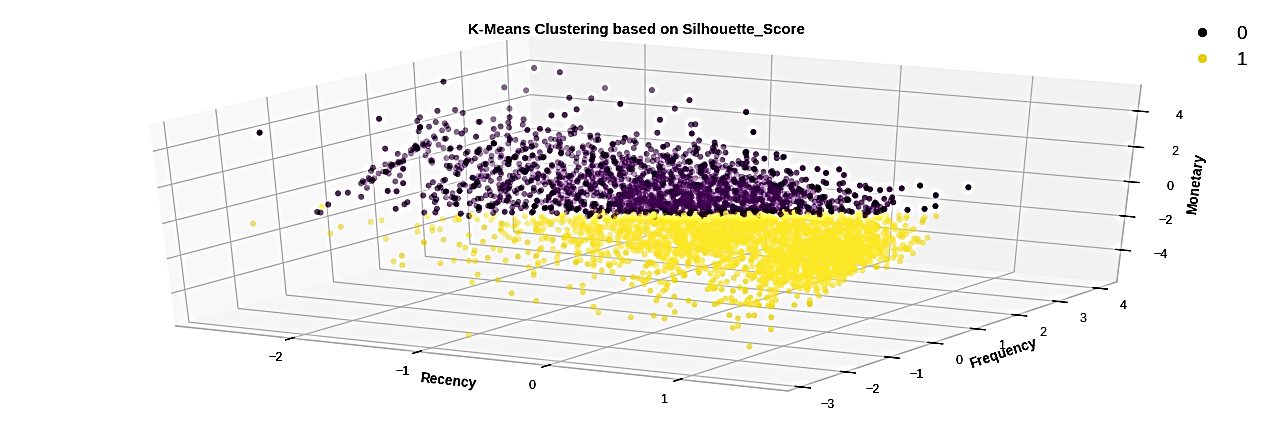
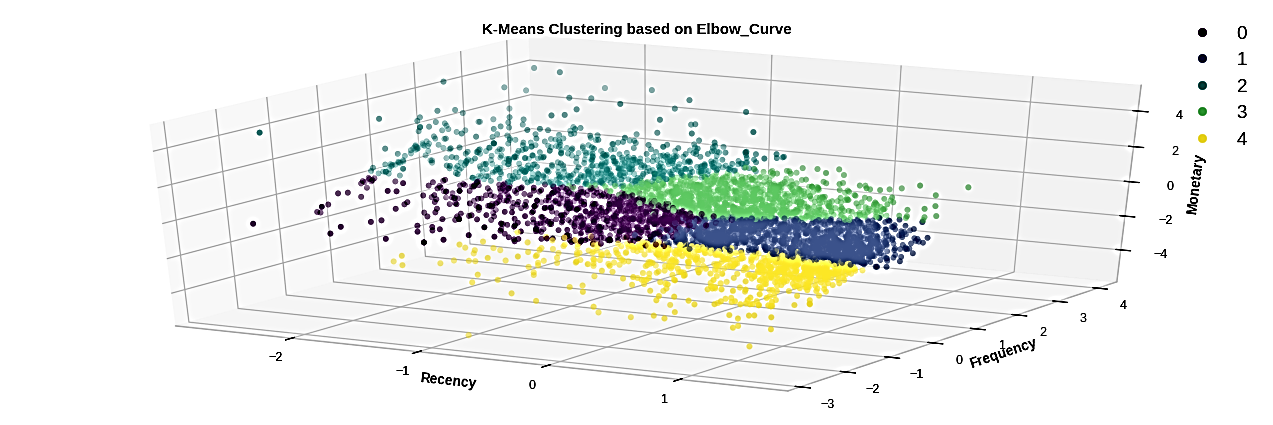
1.The silhouette plot of KMeans

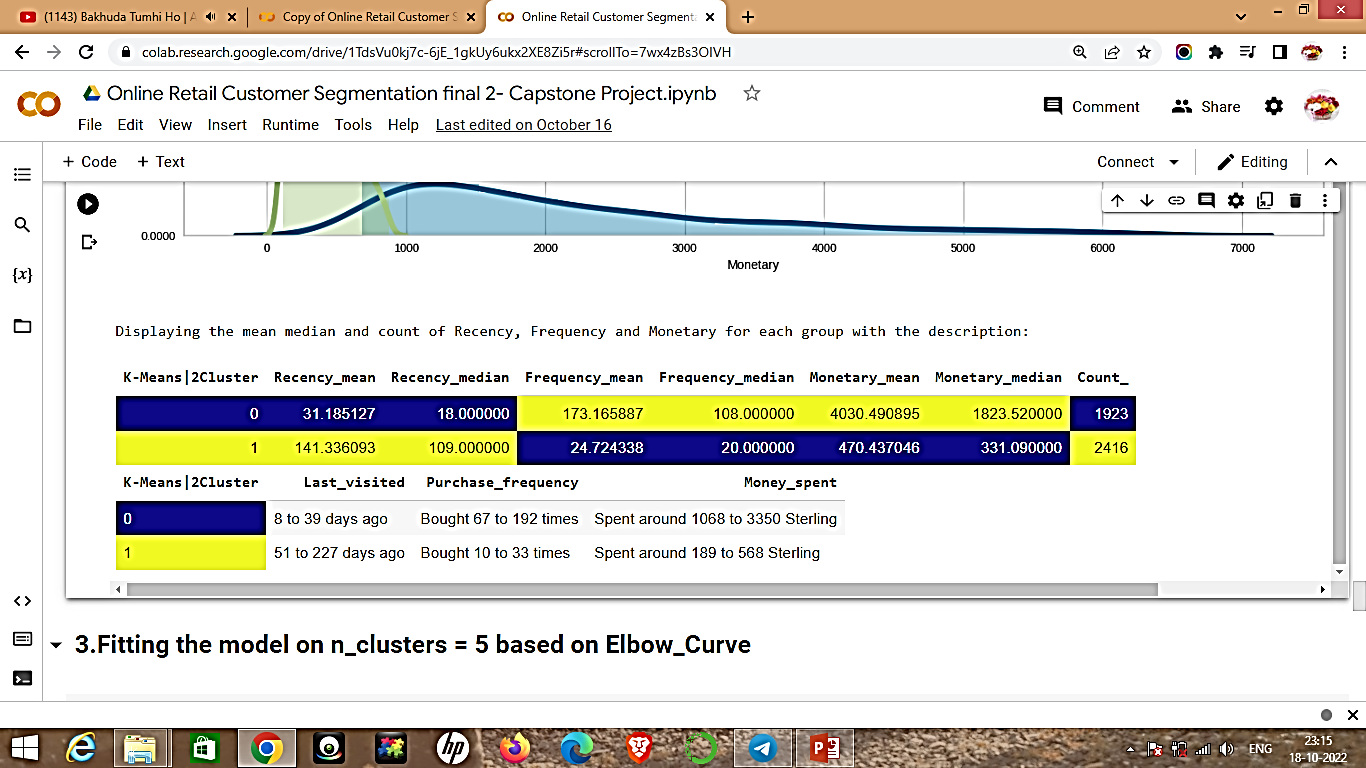
Clustering for 4339 samples is 2 samples and their average score is 0.4

2.The silhouetee plot of kmean Clustering for 4339 samples is 3 samples and their average score is 0.3.

The silhouette plot of KMeans

Clustering for 4339 samples is 2 samples and their average score is 0.4 .

The silhouetee plot of K means clustering for 4339 samples is



This is the K-Means of clustering

Where the number of clusters is 2 and the silhouetee score is used so ,the number of segments were 2.

The RFM values are described in table which is used for further analysis of data.

We have applied the k-means clustering that it is sensitive to outliers means Cluster Centroids can be dragged by outlier.

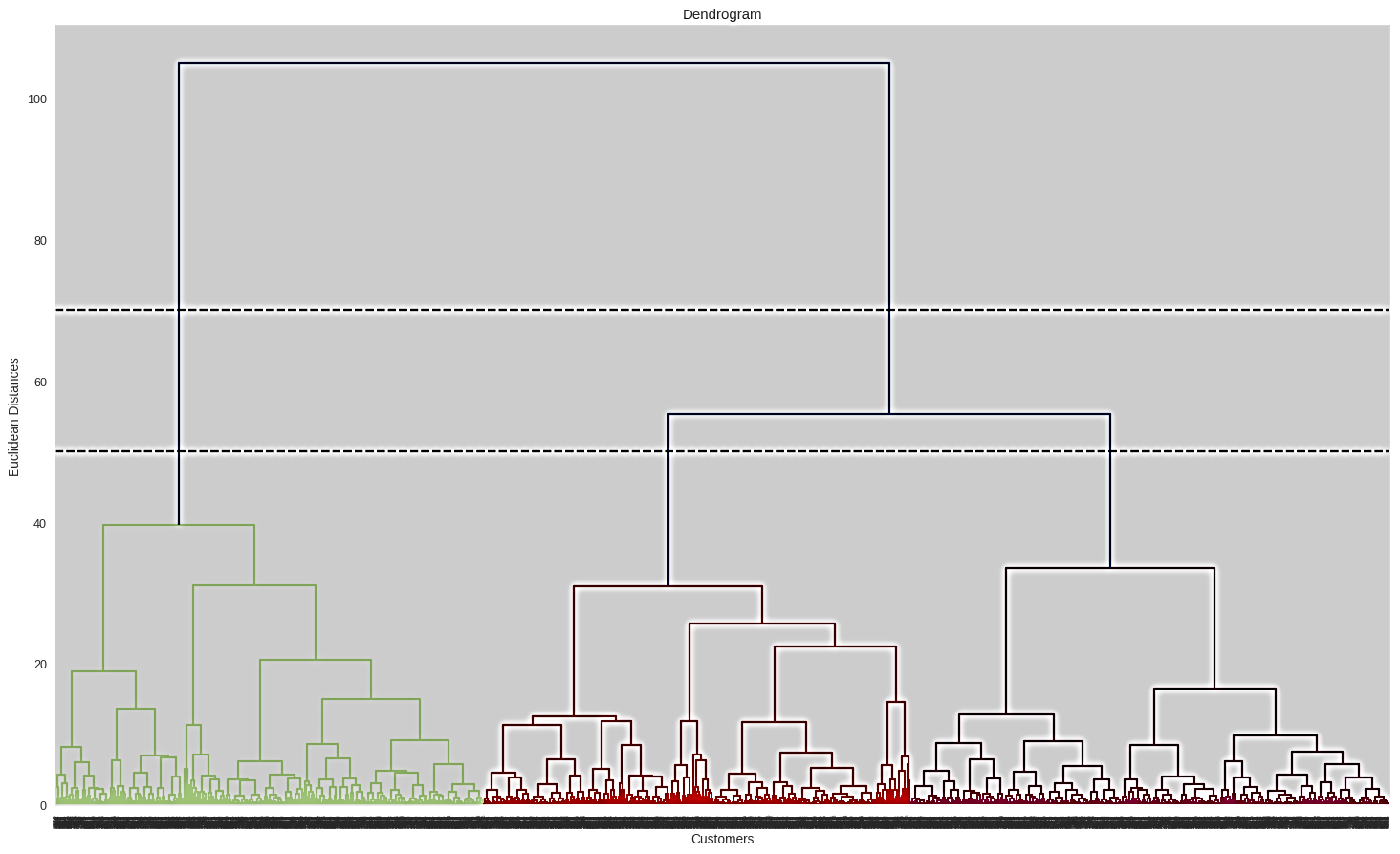
The criterion we used here is Elbow\_Curve .

The number of segments we got here is 5 .

# HIERARCHICAL CLUSTERING

Hierarchical clustering is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from the other cluster, and the objects within each cluster are broadly similar to each other

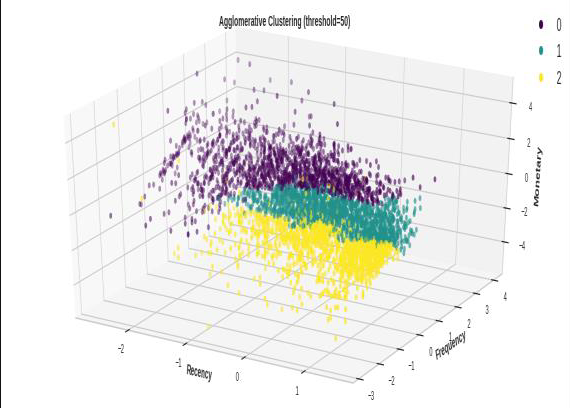
Two techniques are used inalgorithm-Agglomerative and divisive .



Divisive. We have used the agglomerative approach which starts by making n clusters and aggregates the data points until clusters are obtained.

**We can determine the number of clusters on the basis of dendogram as follows:**

1. We can set a threshold distance and draw a horizontal line (Generally, we try to set the threshold in such a way that it cuts the tallest vertical line). We can set this threshold at 70 or 50 and draw a horizontal line.
2. The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold. The larger threshold (y=70) results in 2 clusters while the smaller (y=50) results in 3 clusters.



Hierarchical clustering work for finding the convex clusters. In other words, they are suitable only for compact and well-separated clusters. Moreover, they are also severely affected by the presence of noise and outliers in the data. So we also experimented other algorithms for clustering.

# CONCLUSION

1. We started with a simple binning and quantile-based simple segmentation model first then moved to more complex models.
2. We created several useful clusters of customers based on different metrics and methods to categorize customers based on their behavioural attributes to define their evaluability, loyalty, profitability, etc for the business.
3. Segments depend on how the business plans to use the

results, and the level of granularity they want to see in the clusters.

1. Keeping these points in view we clustered the major segments based on our understanding as per different criteria as shown.
2. **RFM QUANTILE CUT**
3. **Lost poor customers**
4. **Lost loyal customers**
5. **Good customers**
6. **Best customers**

# K-MEANS (2 CLUSTERS)

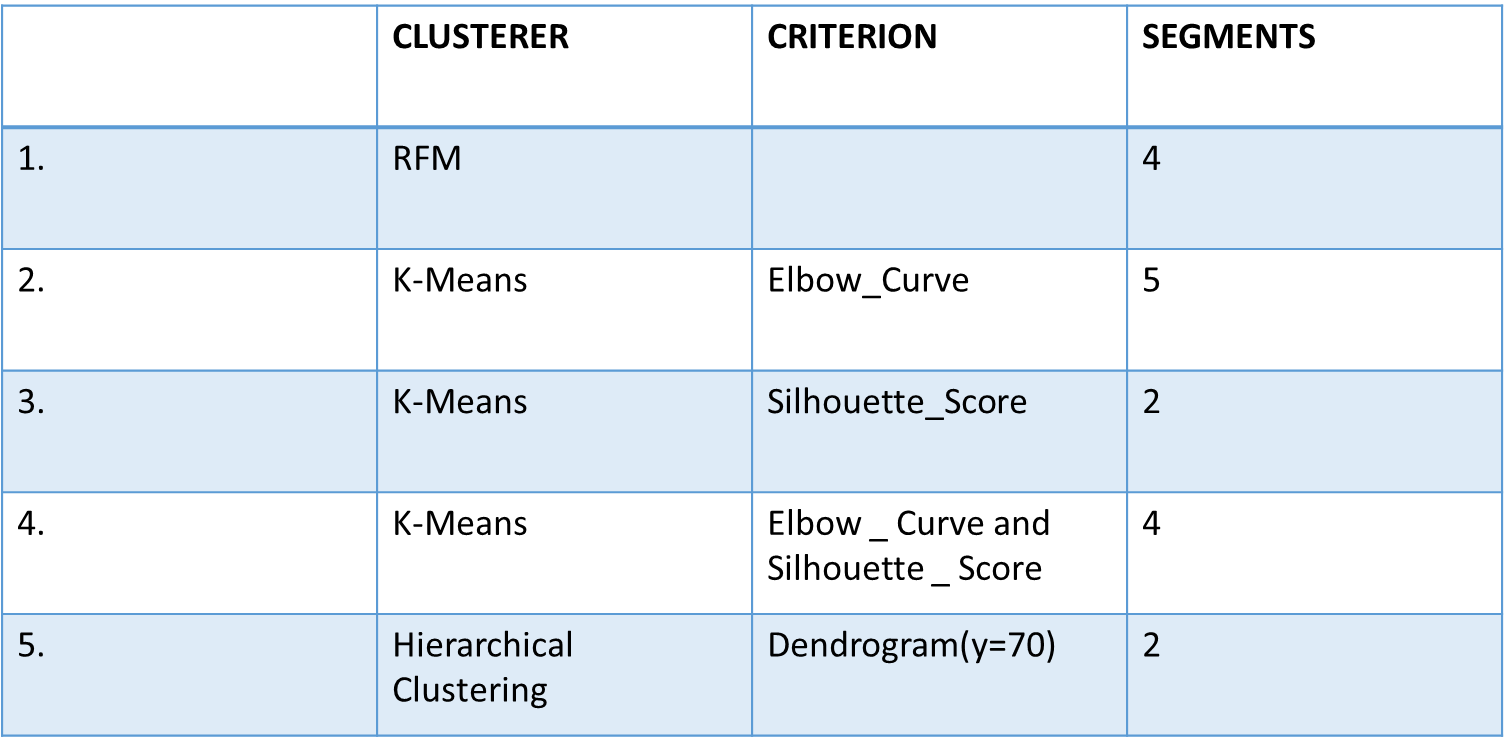
1. **Best customers**
2. **Lost poor customers K-MEANS (4 CLUSTERS)**
3. **Losing loyal customers**
4. **Best customers**
5. **Lost poor customers**
6. **Recently visited average customers**

# K-MEANS (5 CLUSTERS)

1. **Lost poor customer**
2. **Best customers**
3. **Recently visited average customers**
4. **Losing loyal customers**
5. **Average customers**

# AGGLOMERATIVE (2 CLUSTERS)

1. **Average customers**
2. **Best customers**



**EDA Observations:-**

**1.United Kingdom has most number of orders With a count of more than 16000.**

**2.First country with high quantity of orders is from Netherlands with a count of 80 above**

**3.The product with high quantity of order is Paper Craft Little Birdie with a count of 80000.**

**4.Product that made most of the profit in terms of revenue is Paper Craft Little Bride. With more than the count of 160000.**

**5.The customer id with most number of cancellations is 149110.**

**6. Country with highest number of cancellation is United Kingdom with a Count of more than 7000.**