Unsupervised Learning

Project : Customer Segmentation

By : Sunil Kumar

Content

- Problem Statement.
- Attribute description.
- Customer segmentation-Project steps.
- Modeling and finding Optimal K.
- Top 5 Countries based on most Numbers of Customers.
- Top 5 Countries based on least numbers of customers.
- Bottom 5 and top 5 products based on selling.
- Month wise analysis for Customers purchases.
- Day wise analysis for Customers purchases.
- Hours and time Zone for Customers Purchases.
- Use of Models.
- Models With distinct approaches.
- Model Summary for Optimal clusters output.
- Conclusions.
- Challenges.

Problem Statement

In this project, your task is to identify major customer segments on a transactional 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.

Attribute description

Data Description

Attribute Information:

Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

Stock Code: 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.

Invoice Date: 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 the country where each customer resides.

Customer segmentation-Project steps

1 **Data-Preprocessing**
Data collection, data understanding and data Loading.
Exploratory Data Analysis and data cleaning.
Data wrangling and transformation.
Feature engineering and feature extraction.

2 Data-Processing-Modeling
RFM Model
K-mean-clustering.
Hierarchical clustering.
DBSCAN.

3 Metrics evaluation
Finding optimal clusters using Elbow method, using Silhouette and DBSCAN.

Modeling and finding Optimal K

Model used

Models

RFM model-Heuristic approach

K-Mean clustering.

Hierarchical clustering.

DBSCAN.

Optimal K-finding

Elbow method.

Silhouette score.

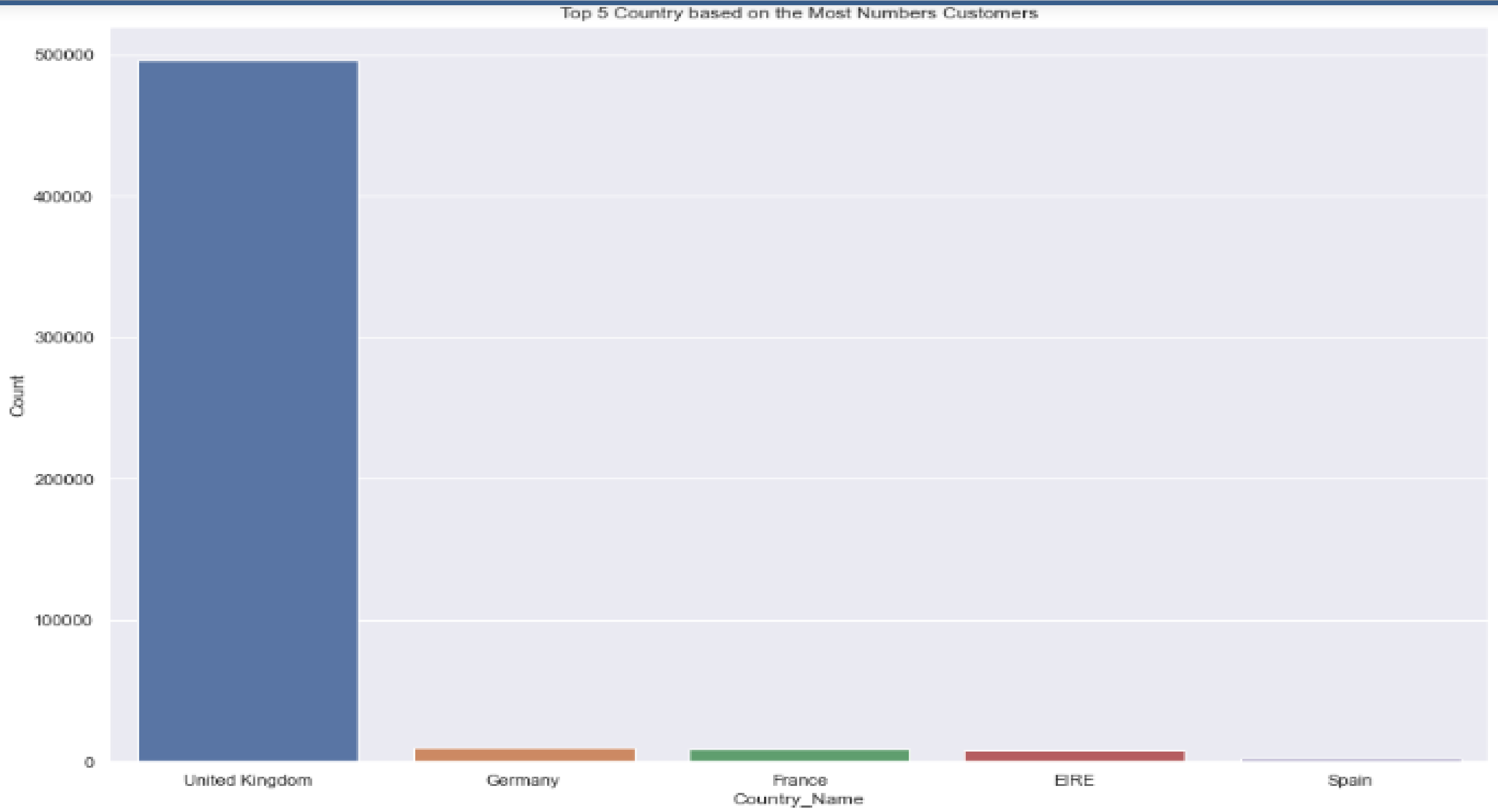
Dendrogram.

Models With distinct approaches

- RFM Model
- K-Means with silhouette score
- K-Means with Elbow method
- DBSCAN
- RM K-Means with silhouette score
- K-Means with Elbow methods DBSCAN
- FM K-Means with silhouette score
- RFM K-Means with Elbow method
- RFM Hierarchical clustering

RFM DBSCAN RFM

Top 5 Countries based on most Numbers of Customers

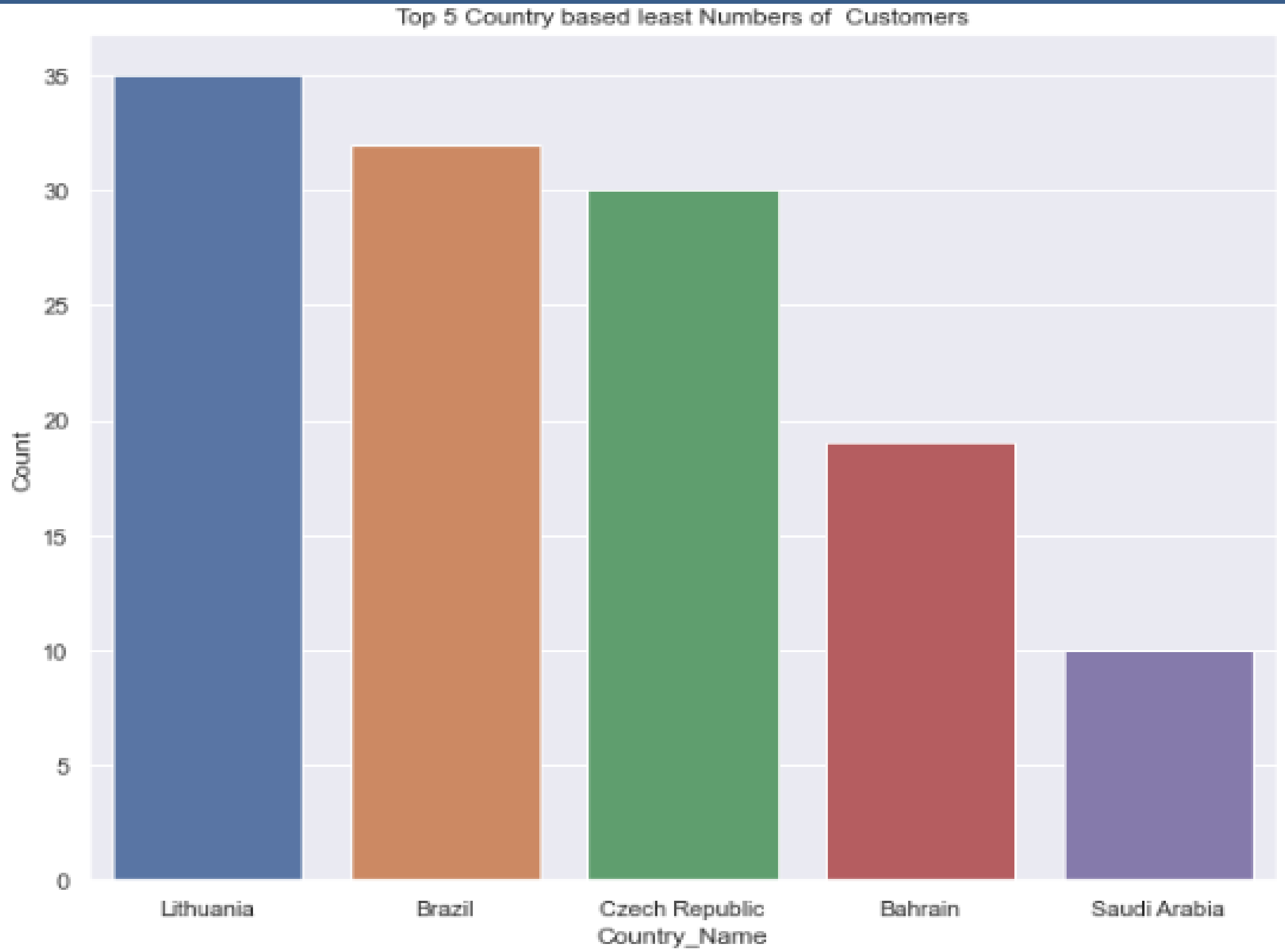


Visualization for country

Findings

- above are top 5 country with most numbers of the customers.
- From this graph we can see that most of the customers are from United Kingdom ,Germany ,France ,EIRE and Spain.**

Top 5 Countries based on least numbers of customers



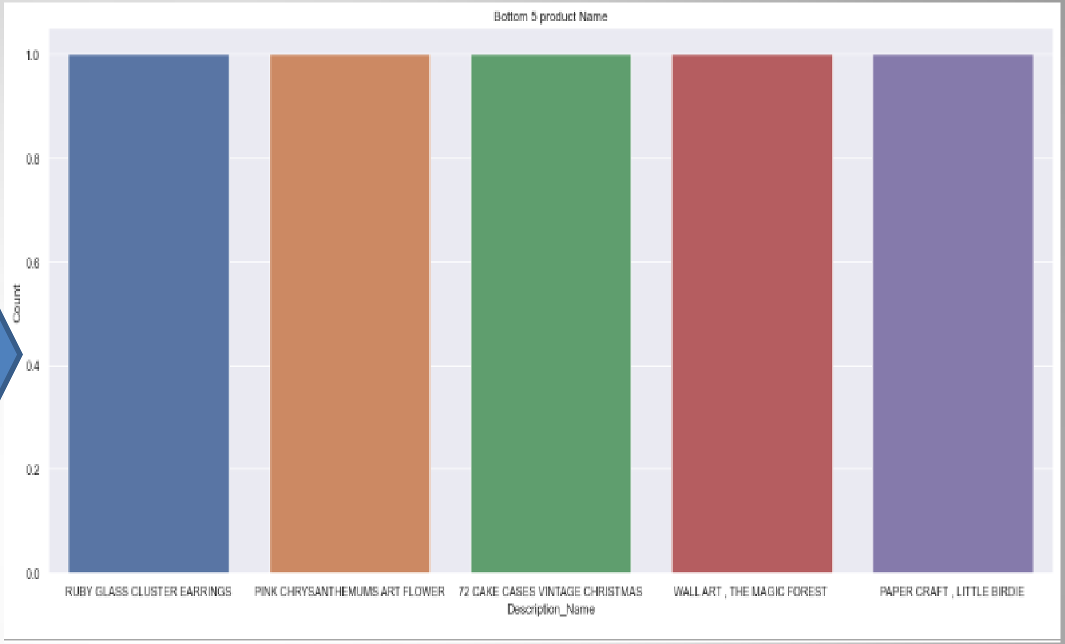
Findings__

- From this graph we can see that least number of customers from Lithuania,Brazil, Czech Republic ,Bahrain and Saudi Arabia

Bottom 5 and top 5 products based on selling

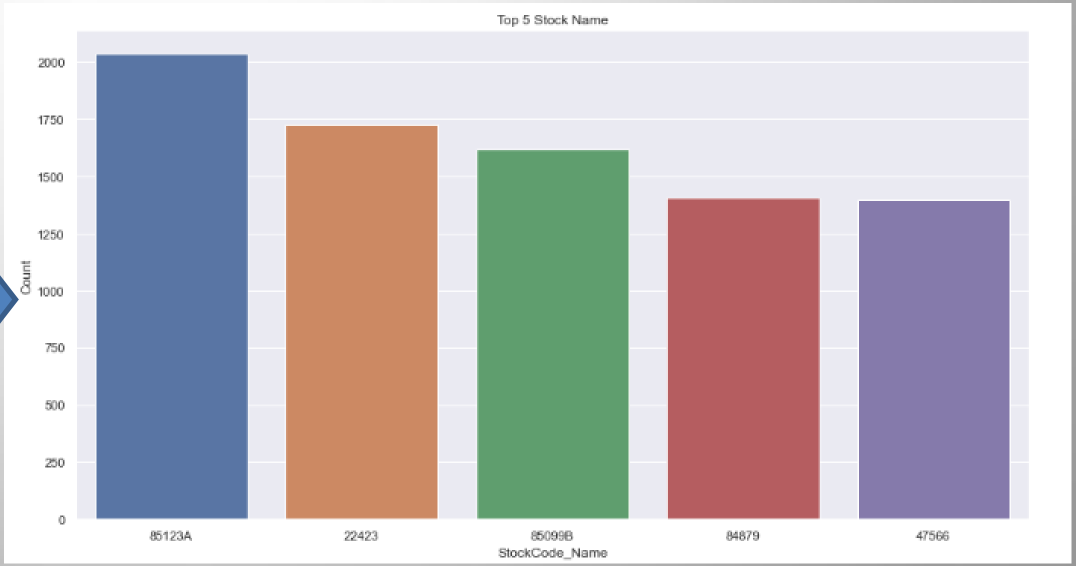
Bottom 5 Products based on the selling are:

1.light decorated battery operated.
2.Water damaged.
3.throw away.
4.re dotcom quick fix.
5.Birthday Banner Tape.

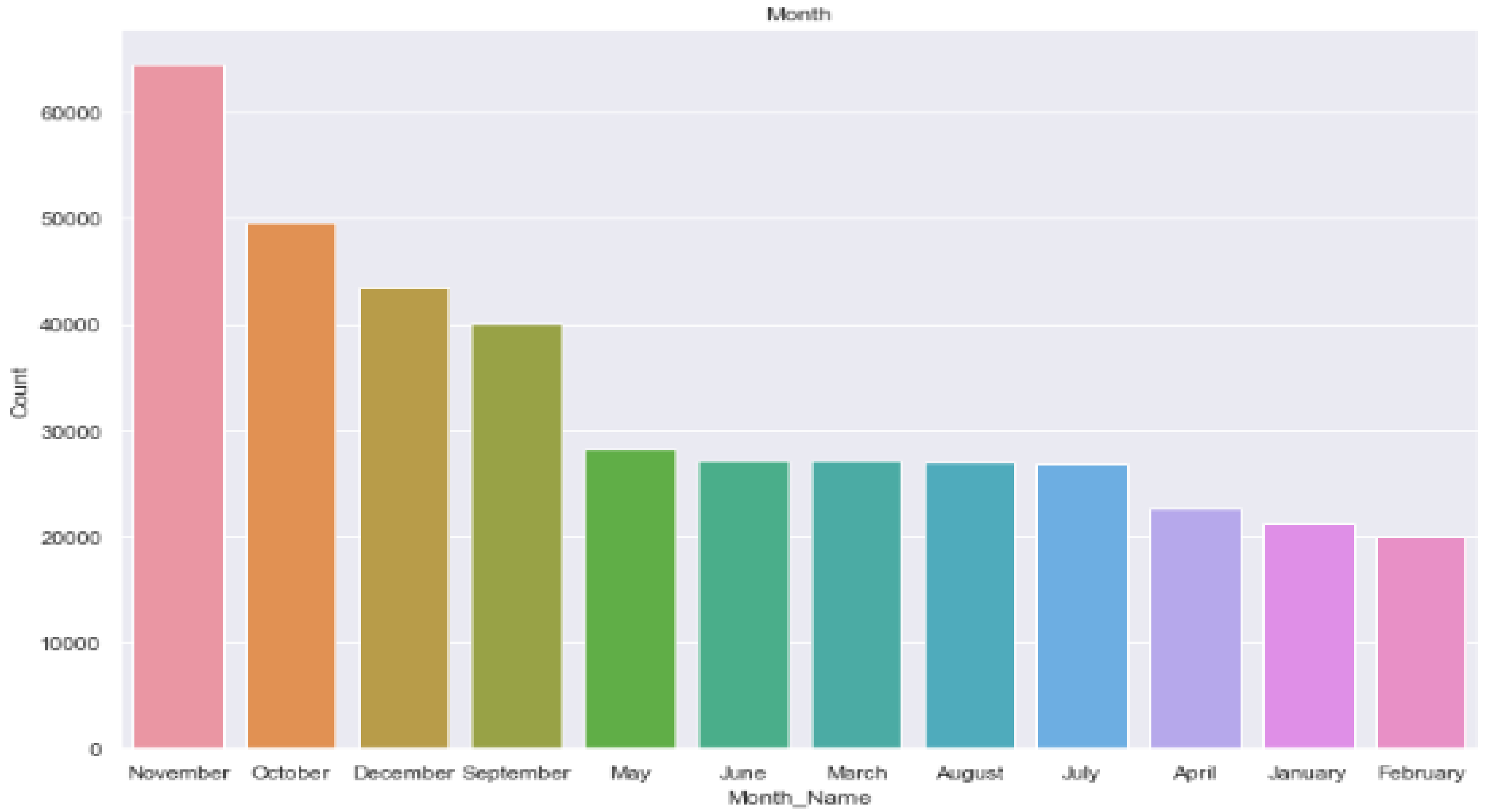


Top 5 Stocks name based on selling are:

85123A
22423
85099B
47566
20725



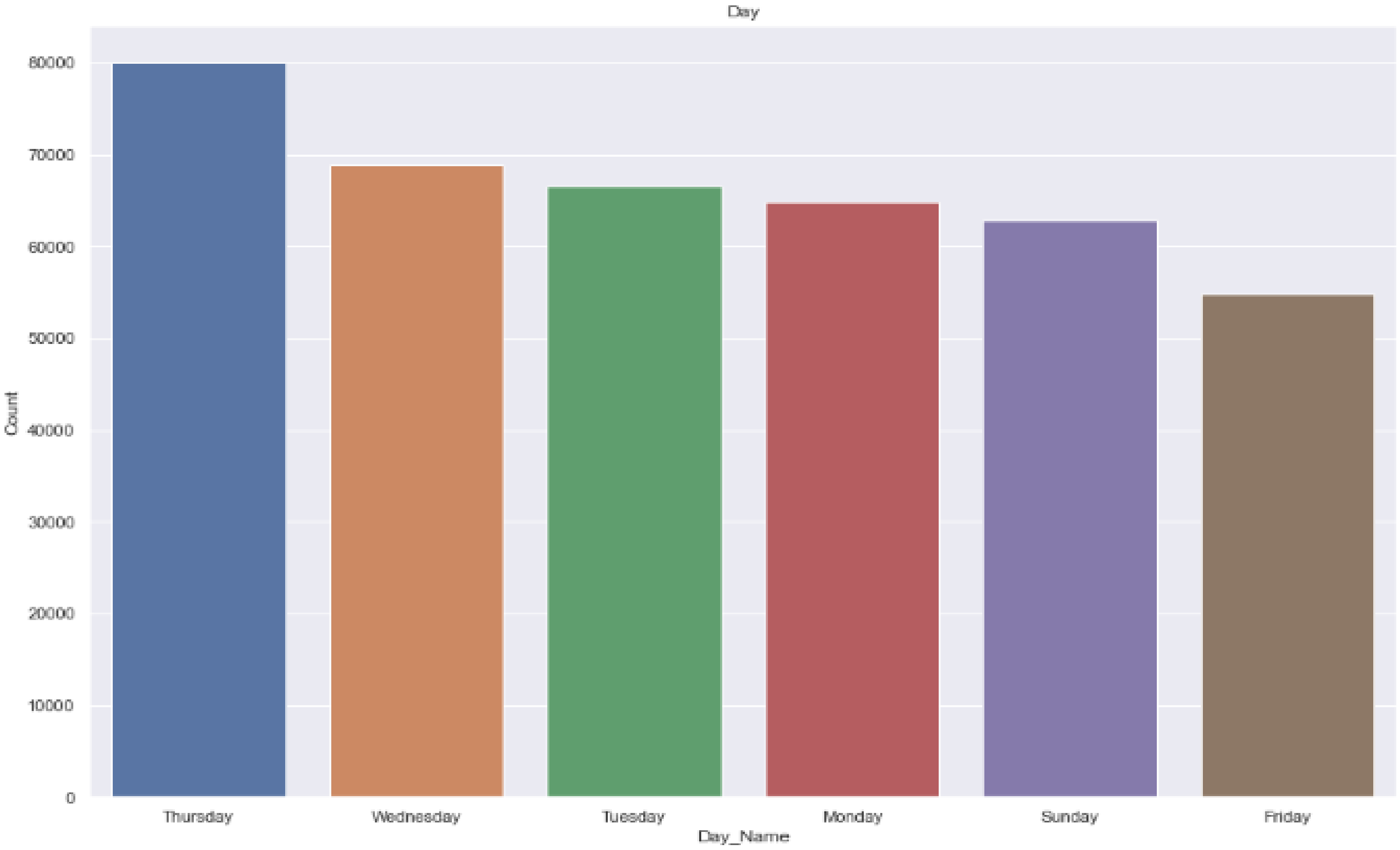
Month wise analysis for Customers purchases



Findings__

- Most numbers of customers have purchased in the months of November ,October, December and September.
- Months of April ,January and February, we have less number of customer buying.

Day wise analysis for Customers purchases.

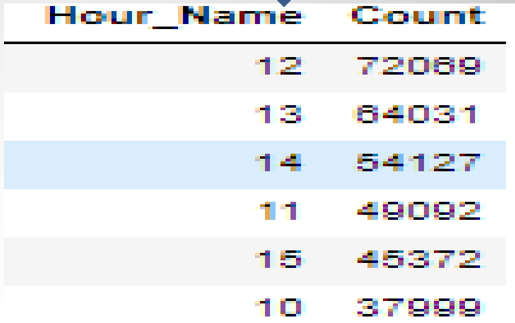


Most of the Customers have purchased Thursday, Wednesday and Tuesday.

Hours and time Zone for Customer Purchases

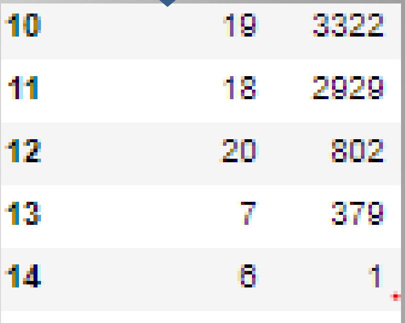
Findings__
We can also say that afternoon timings are popular for the purchasing items. Especially 11-12-13-14-15 gave the more numbers of customer purchasing.

Top 6 Active_hrs



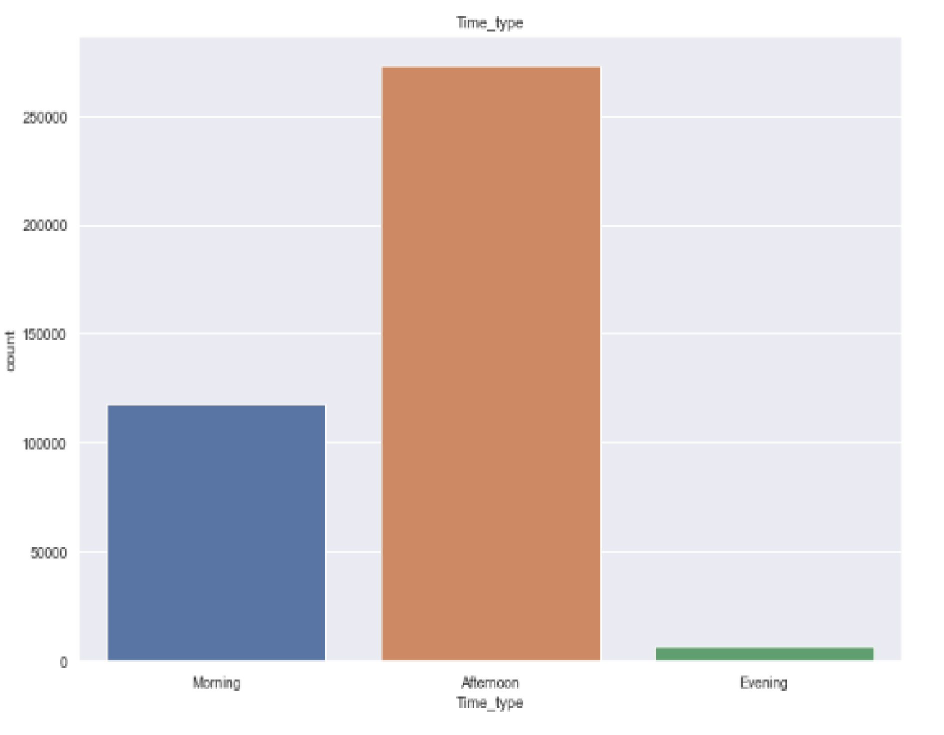
Hour_Name	Count
12	72069
13	64031
14	54127
11	49092
15	45372
10	37999

Least 5 Active_hrs



10	19	3322
11	18	2929
12	20	802
13	7	379
14	6	1

Findings__
We have Afternoon time where we have maximum customer purchasing items/gifts.
Least numbers of customers have purchased the items in Evening.



Use of Models

RFM model (Recency, Frequency, Monetary value)

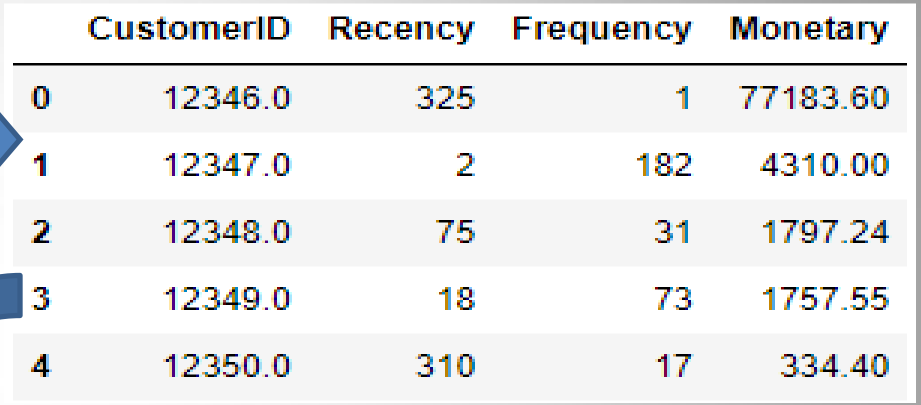
The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer. The second step is to divide the customer list into tiered groups for each of the three dimensions (R, F and M)

Calculating RFM scores.

The number is typically 3 or 5. If you decide to code each RFM attribute into 3 categories, you'll end up with 27 different coding combinations ranging from a high of 333 to a low of 111.

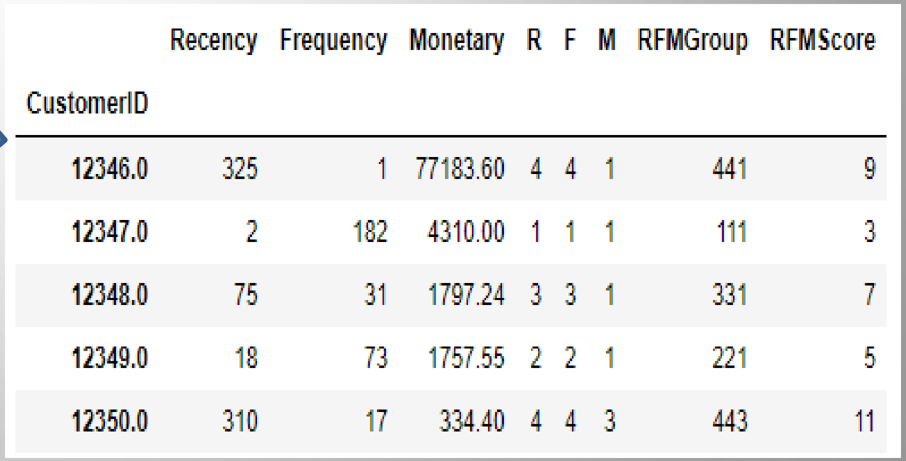
Generally speaking, the higher the RFM score, the more valuable the customer.

Before Quintile Binning



	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12347.0	2	182	4310.00
2	12348.0	75	31	1797.24
3	12349.0	18	73	1757.55
4	12350.0	310	17	334.40

After Quintile Binning



	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore
CustomerID								
12346.0	325	1	77183.60	4	4	1	441	9
12347.0	2	182	4310.00	1	1	1	111	3
12348.0	75	31	1797.24	3	3	1	331	7
12349.0	18	73	1757.55	2	2	1	221	5
12350.0	310	17	334.40	4	4	3	443	11

Use of Models continued

K-mean-clustering(Unsupervised algorithm works by clustering similar groups.)

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters.

Step 1: Choose the number of clusters k.

Step 2: Select k random points from the data as cancroids.

Step 3: Assign all the points to the closest cluster centroid.

Step 4: Recompute the centroids of newly formed clusters.

Step 5: Repeat steps 3 and 4.

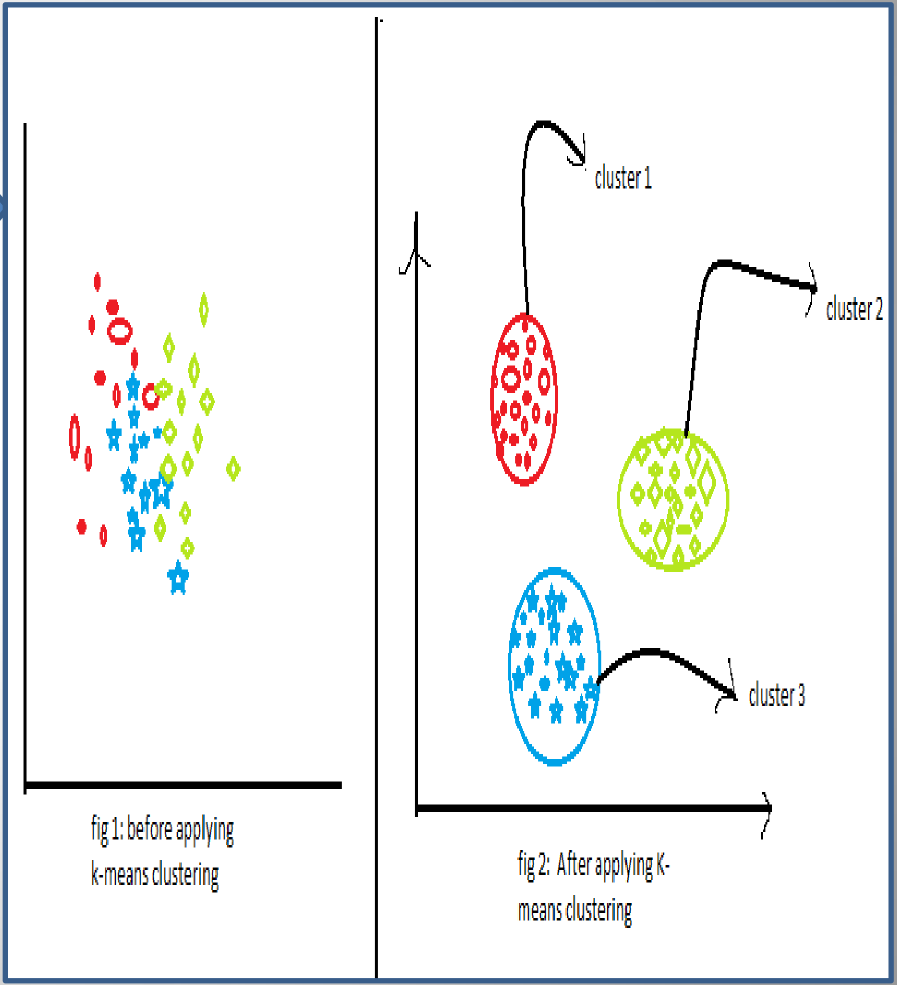
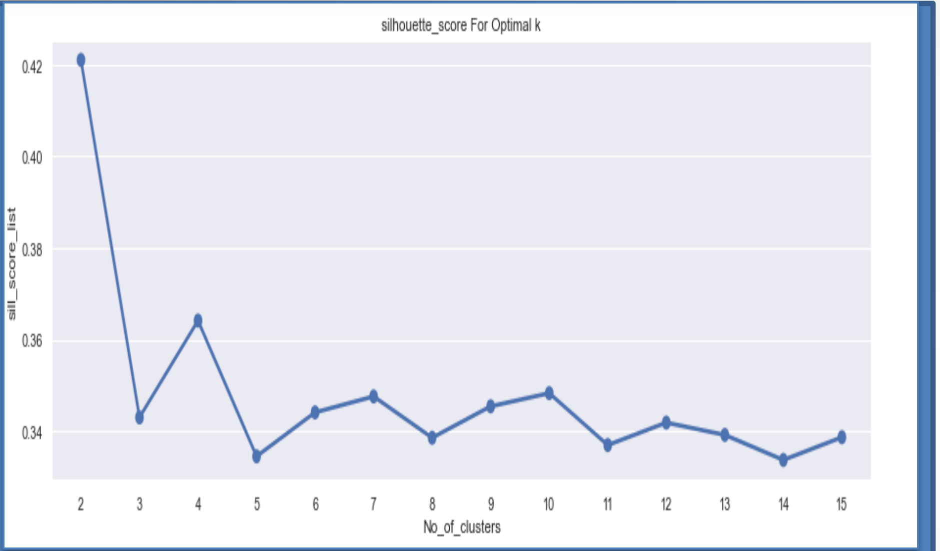


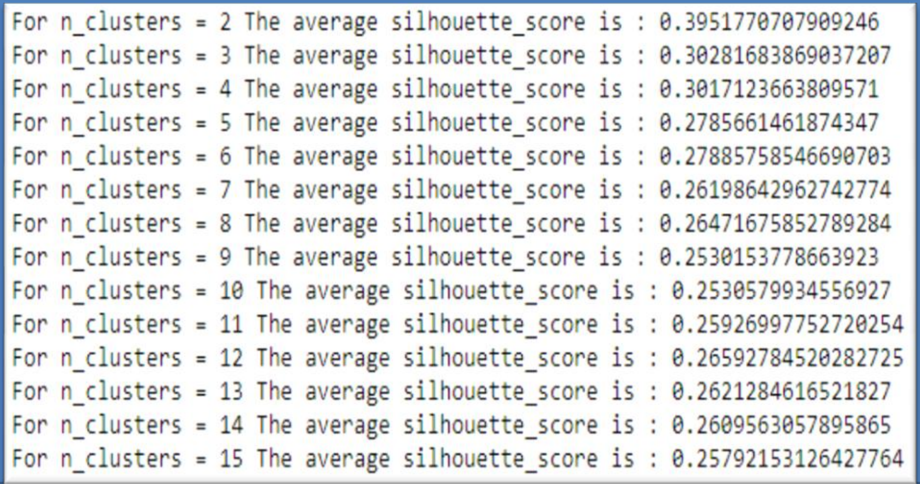
fig 1: before applying k-means clustering

fig 2: After applying K-means clustering

Use of Models continued

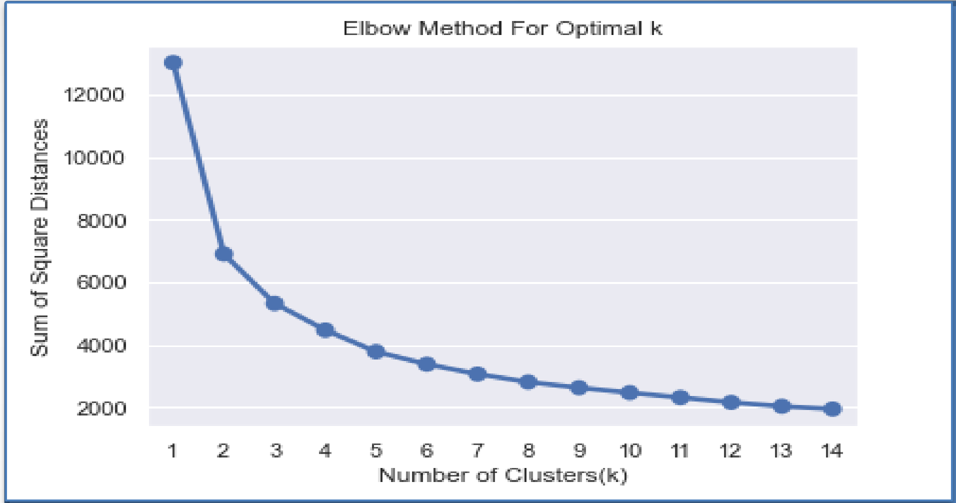
K-mean-Using Silhouette Score method



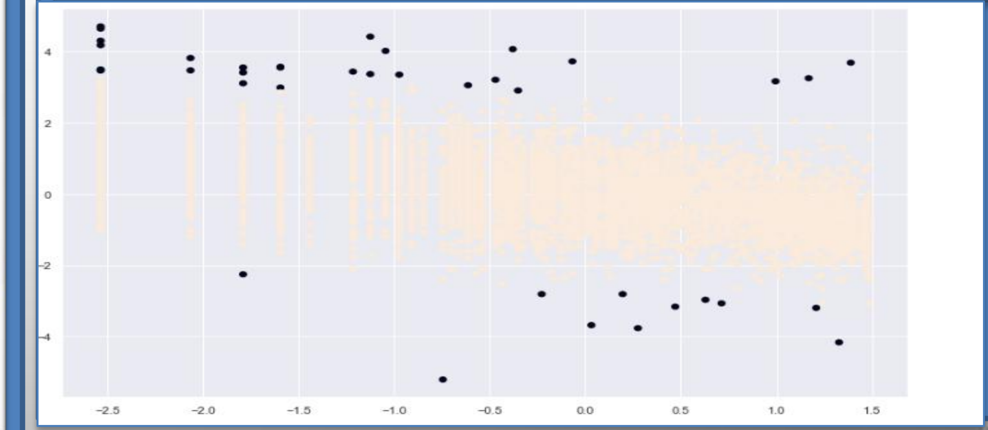


```
For n_clusters = 2 The average silhouette_score is : 0.3951770707909246
For n_clusters = 3 The average silhouette_score is : 0.30281683869037207
For n_clusters = 4 The average silhouette_score is : 0.3017123663809571
For n_clusters = 5 The average silhouette_score is : 0.2785661461874347
For n_clusters = 6 The average silhouette_score is : 0.27885758546690703
For n_clusters = 7 The average silhouette_score is : 0.26198642962742774
For n_clusters = 8 The average silhouette_score is : 0.26471675852789284
For n_clusters = 9 The average silhouette_score is : 0.2530153778663923
For n_clusters = 10 The average silhouette_score is : 0.2530579934556927
For n_clusters = 11 The average silhouette_score is : 0.25926997752720254
For n_clusters = 12 The average silhouette_score is : 0.26592784520282725
For n_clusters = 13 The average silhouette_score is : 0.2621284616521827
For n_clusters = 14 The average silhouette_score is : 0.2609563057895865
For n_clusters = 15 The average silhouette_score is : 0.25792153126427764
```

K-mean using Elbow method



DBSCAN method

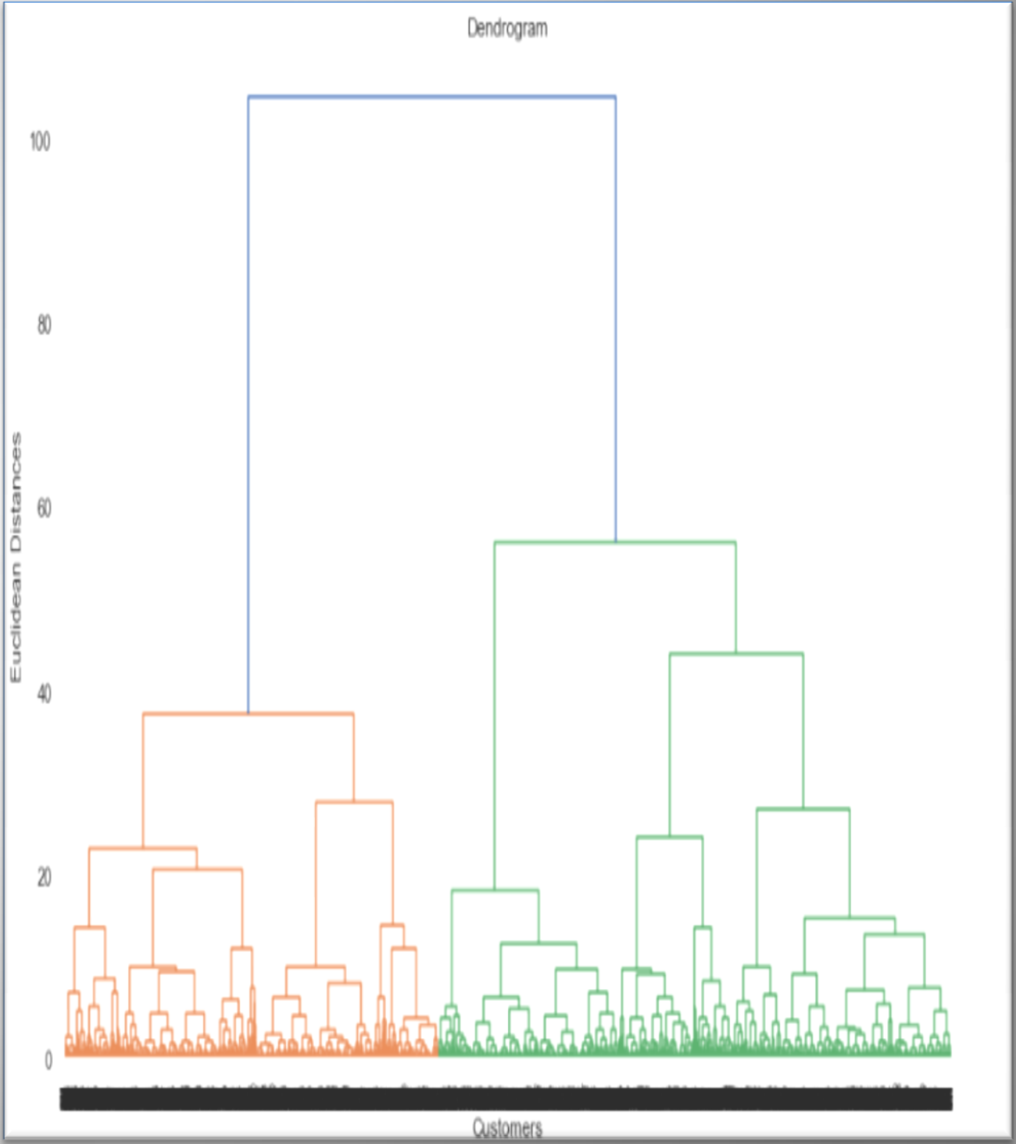


Use of Models continued

Hierarchical-clustering(Optimal clusters Using dendrogram)

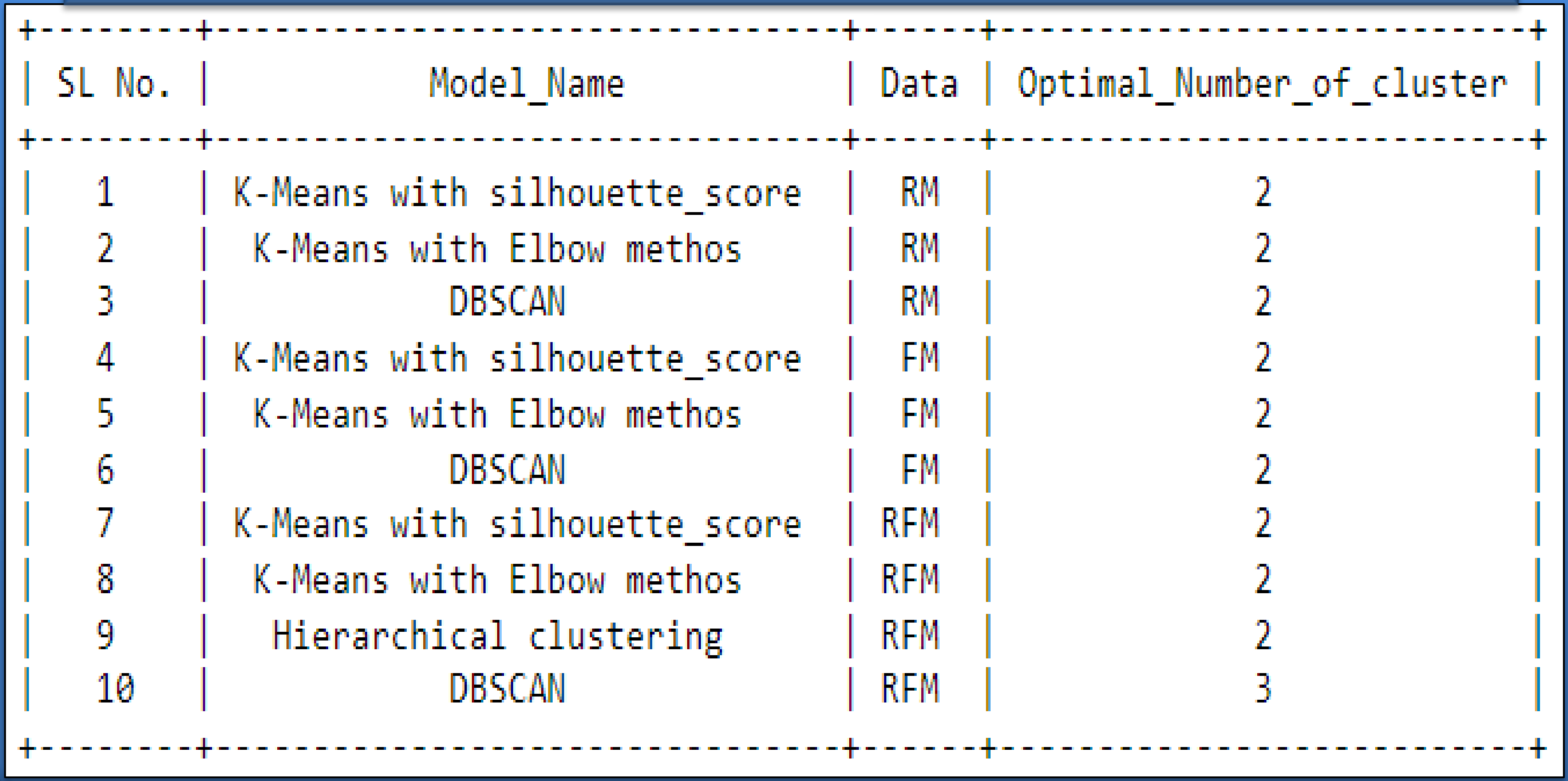
An algorithm that groups similar objects into groups in a hierarchy called hierarchical clustering. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This iterative process continues until all the clusters are merged together.



Model Summary for Optimal clusters output

Finding optimal "K" using different model and features.
Data are the features and Model Name are used methods.



```
+---------+----------------------------------+------+----------------------------+
| SL No.  |            Model_Name            | Data | Optimal_Number_of_cluster  |
+---------+----------------------------------+------+----------------------------+
|    1    | K-Means with silhouette_score    |  RM  |             2              |
|    2    |  K-Means with Elbow methos       |  RM  |             2              |
|    3    |             DBSCAN               |  RM  |             2              |
|    4    | K-Means with silhouette_score    |  FM  |             2              |
|    5    |  K-Means with Elbow methos       |  FM  |             2              |
|    6    |             DBSCAN               |  FM  |             2              |
|    7    | K-Means with silhouette_score    | RFM  |             2              |
|    8    |  K-Means with Elbow methos       | RFM  |             2              |
|    9    |   Hierarchical clustering        | RFM  |             2              |
|    10   |             DBSCAN               | RFM  |             3              |
+---------+----------------------------------+------+----------------------------+
```

Conclusions

I started with the loading, understanding, and exploring the dataset to see the major trends and insights from data regarding the purchases. I have made few observations while performing the project on the given dataset. Which are below:

There were (541909, 8) rows and columns out of which I saw that there were null values in the description and CustomerID columns, which were of float and object data types. I have removed the null values as imputing them with mode would not be meaningful.
After removal of the null values I had (406829, 8) observation and variables respectively.

I did see the overall data distribution and found few points as below:

- In quantity we have values in negative and as well as in Unit Price.
- Found Positively skewed distribution of the dataset.

Conclusions _continued

- Once I started exploring further country wise, monthly basis, day basis and hourly basis and as per time zone my findings were below:
- Countries with top customers are: United Kingdom ,Germany ,France ,EIRE and Spain.
- Most numbers of customers have purchased in the months of November ,October, December and September.
- Most of the customers have purchased the items in Thursday ,Wednesday and Tuesday.
- I have seen that afternoon timings are popular for the purchasing items.
- Especially 11-12-13-14-15 gave the more numbers of customer purchasing.
- Once I have seen the data and for its major minor trends, I then started with modeling techniques which are below:

- Used RFM model for to find out the valuable customers based on Recency, Frequency and Monetary values.

Conclusion _continued

- While using this model I have seen few points that there were customers which were having more Recency and more Monetary, more Recency and less monetary.
- Similarly, like for these combinations I have checked for each customers and found the best set of customers after setting the threshold to 5 and 8 respectively given 1263 customers and 2587 customers with threshold of setting to 8.
- I then started with K-mean clustering to cluster the same set of customers and tried with 2 features and 3 features which were (RFM) Recency, Frequency and Monetary.
- I checked the cluster formation with the help of Silhouette score and elbow method and DBSCAN.
- I found DBSCAN performing good to find out the optimal clusters whereas K-mean clustering is not proven that well with elbow method and silhouette scores. After using all methods I have seen that most of the time optimal numbers of cluster were 2.

Challenges

The major challenges I have faced in this project are mentioned Below:

Null handling for Description column and Customer ID.
Looking for few values which were negative like in Total Price.
Applying Log transformation would be right or not.
Handling skewed dataset.
Binning of quintile for the customers based on 1,2,3,4 score.
Getting Silhouette score for every sample.
Finding optimal “K”.

Thank you!