Capstone Project

Customer Segmentation

Chandrashekhar Awate

Problem Statement

- 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.

Data Dicsription

- **InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this 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: Invice Date and time. Numeric, the day and time when each transaction was generated.**
- **UnitPrice: 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.**

Contents

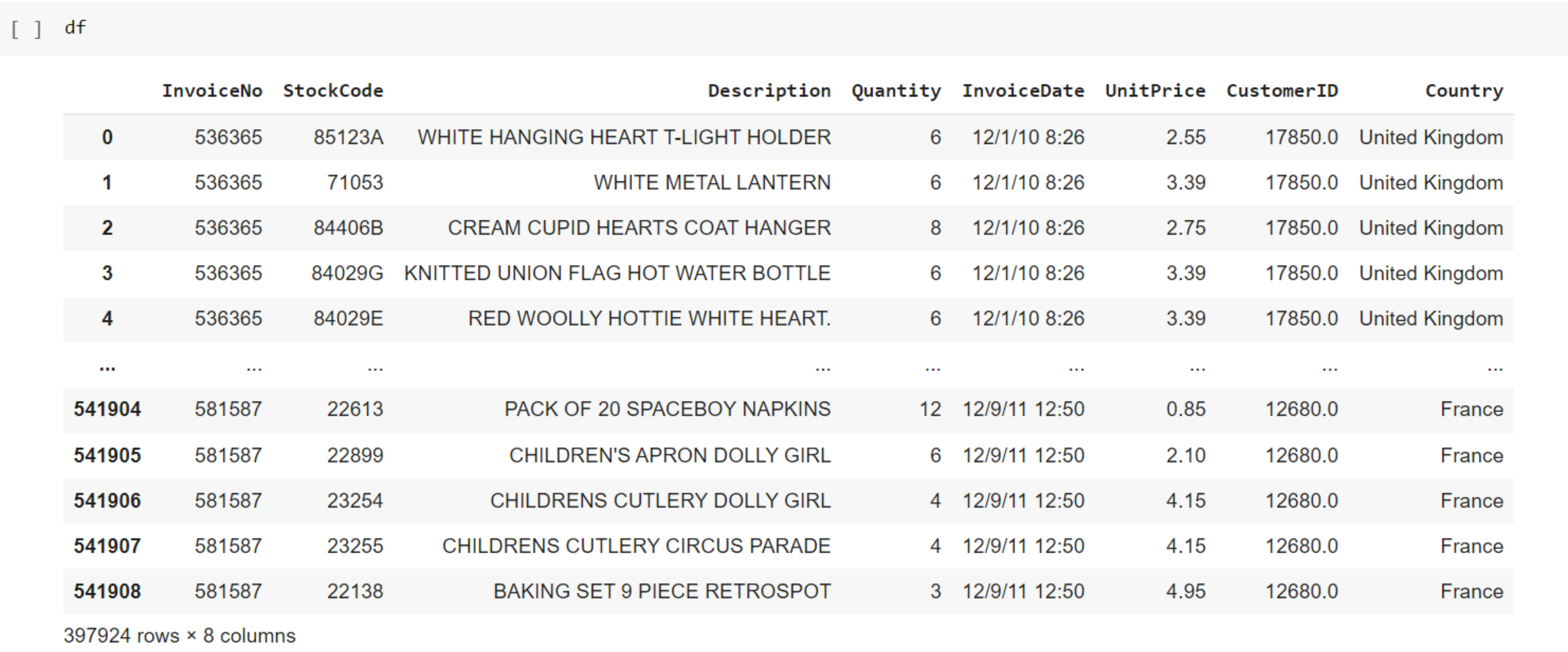
Step 1

1)Exploratory Data Analysis
2)Finding Key Insights
3)Creating RFM Model

Step 2

1)Using Clustering Algorithms on RFM Model
2)Using K means Clustering on various RFM Model
3)Using DBSCAN on various RFM Model
4)Using Various methods to know the optimum number of clusters
5) Silhouette score,Elbow Method,Dendogram
6)Conclusions

Exploratory Data Analysis

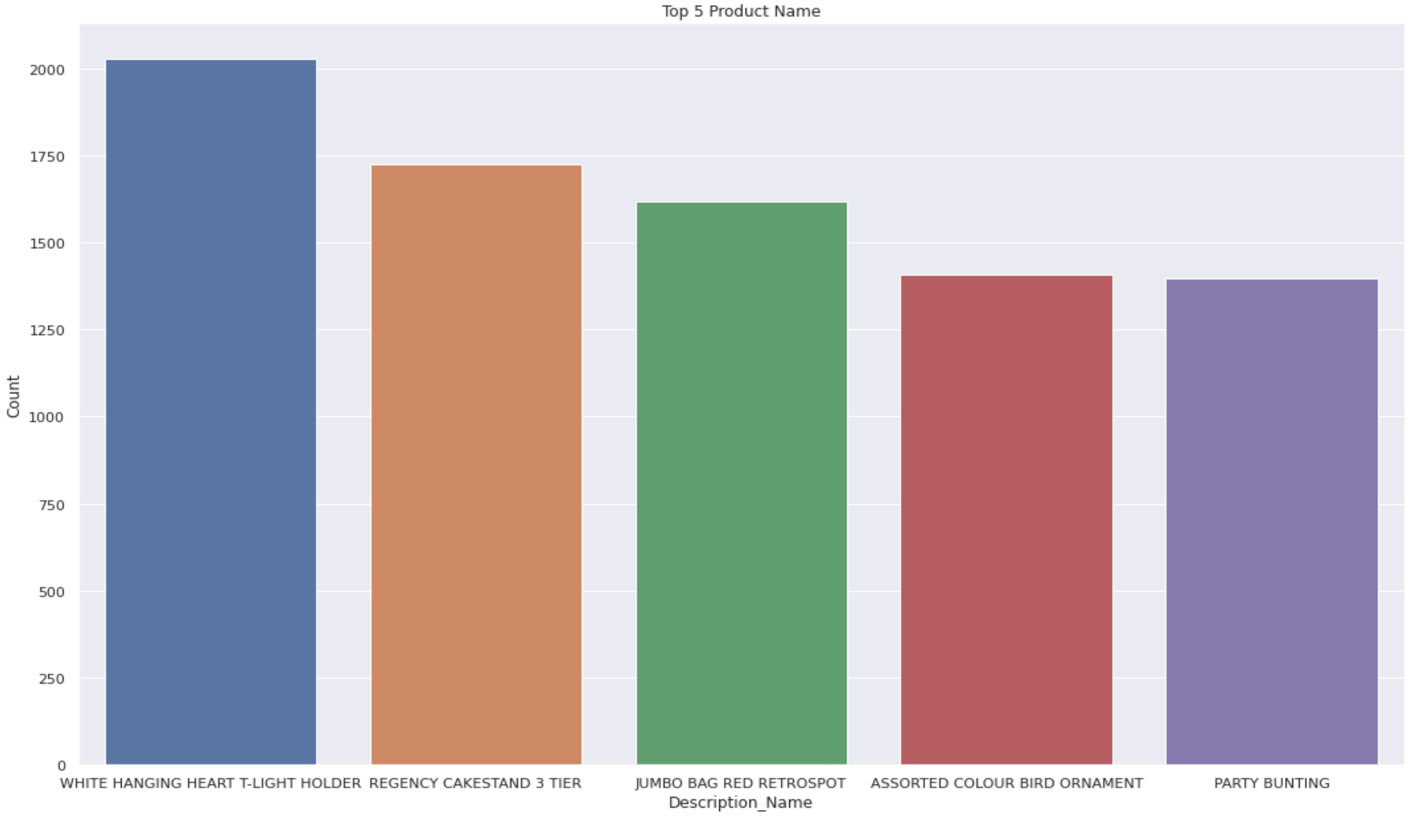


```
df
```

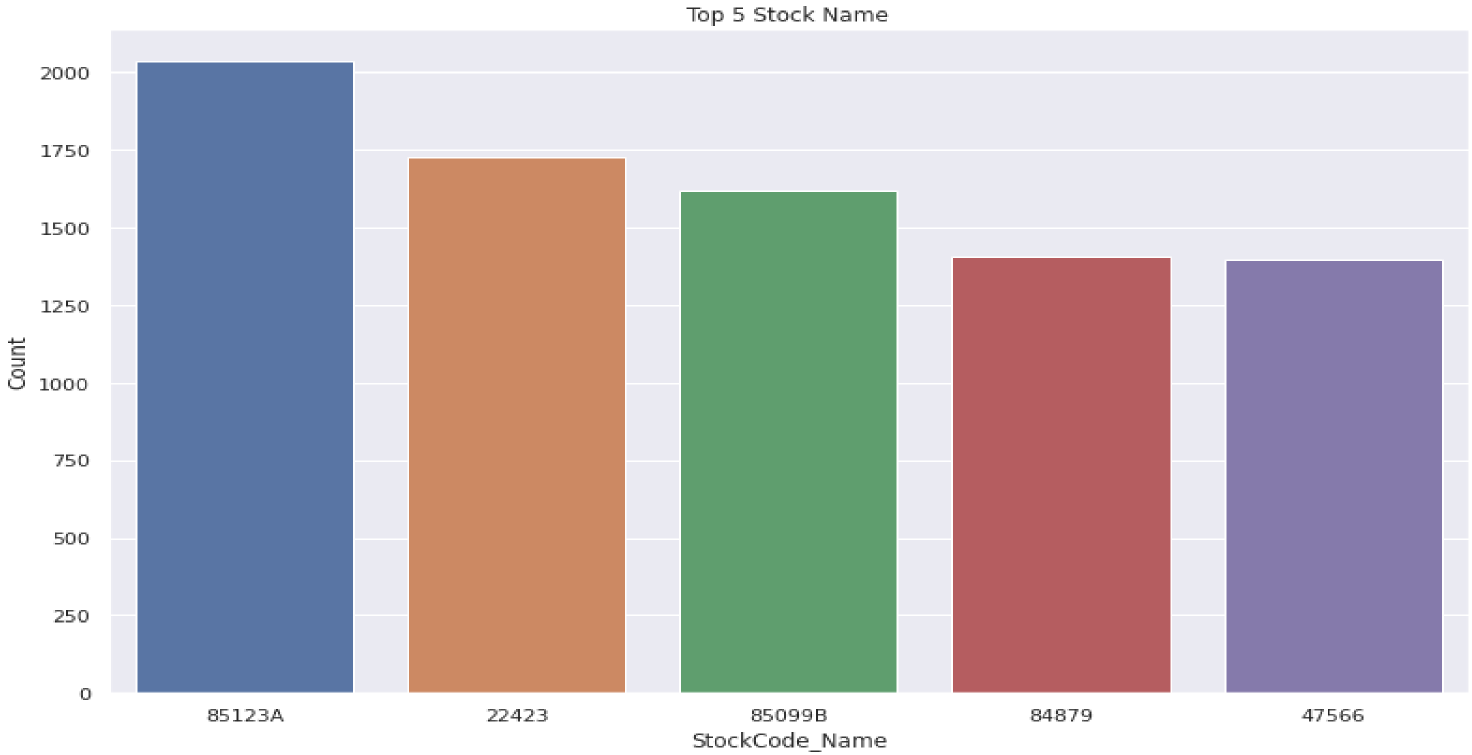
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/10 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/11 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/11 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/11 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/11 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/11 12:50	4.95	12680.0	France

397924 rows × 8 columns

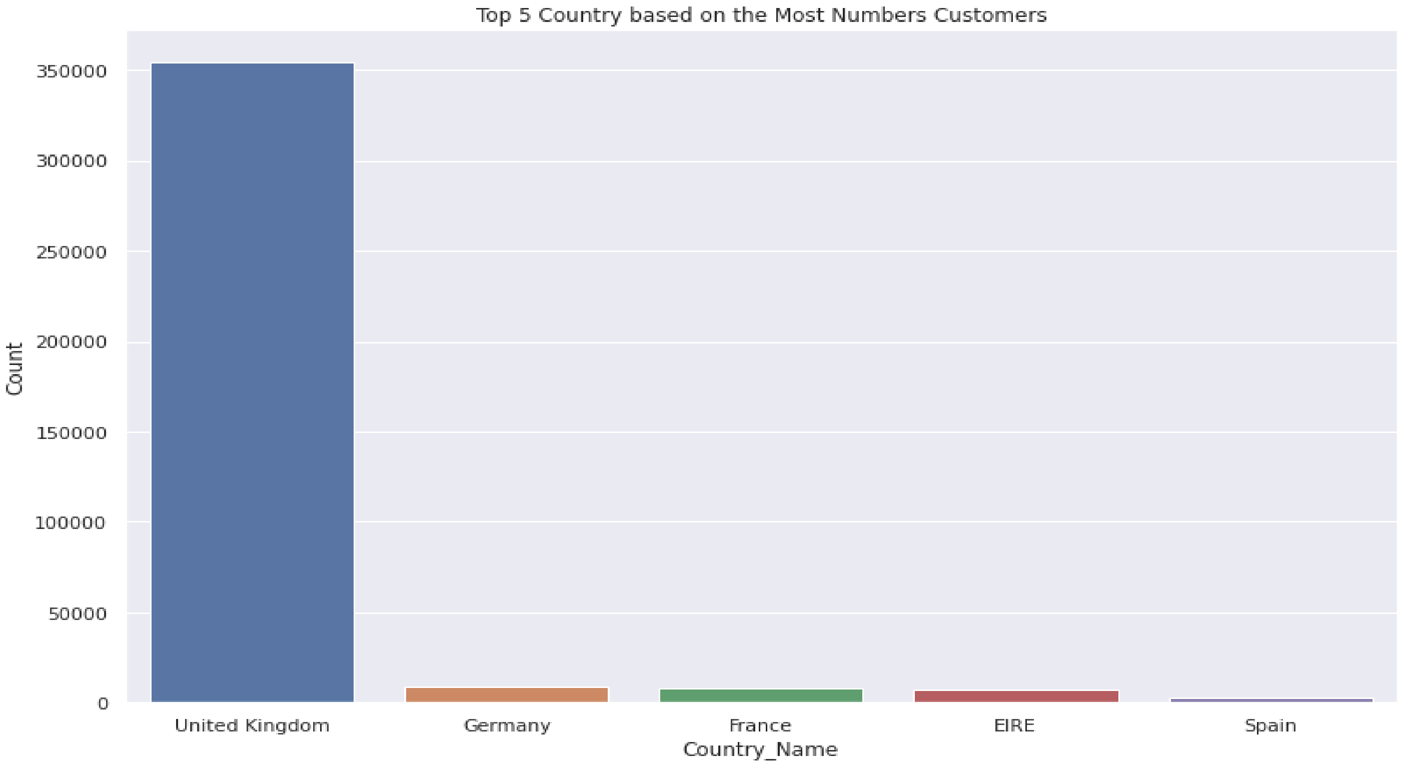
Top Selling Products



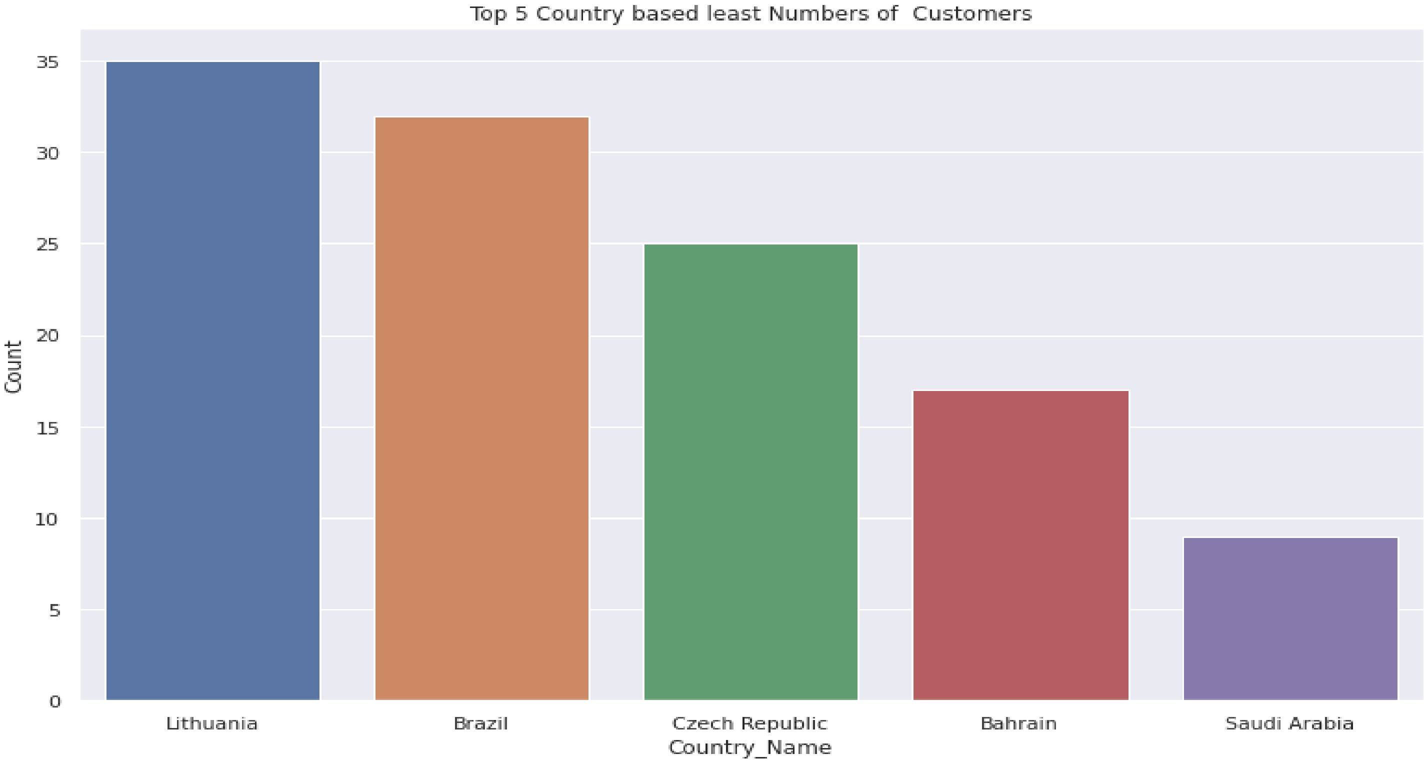
Top 5 StockName



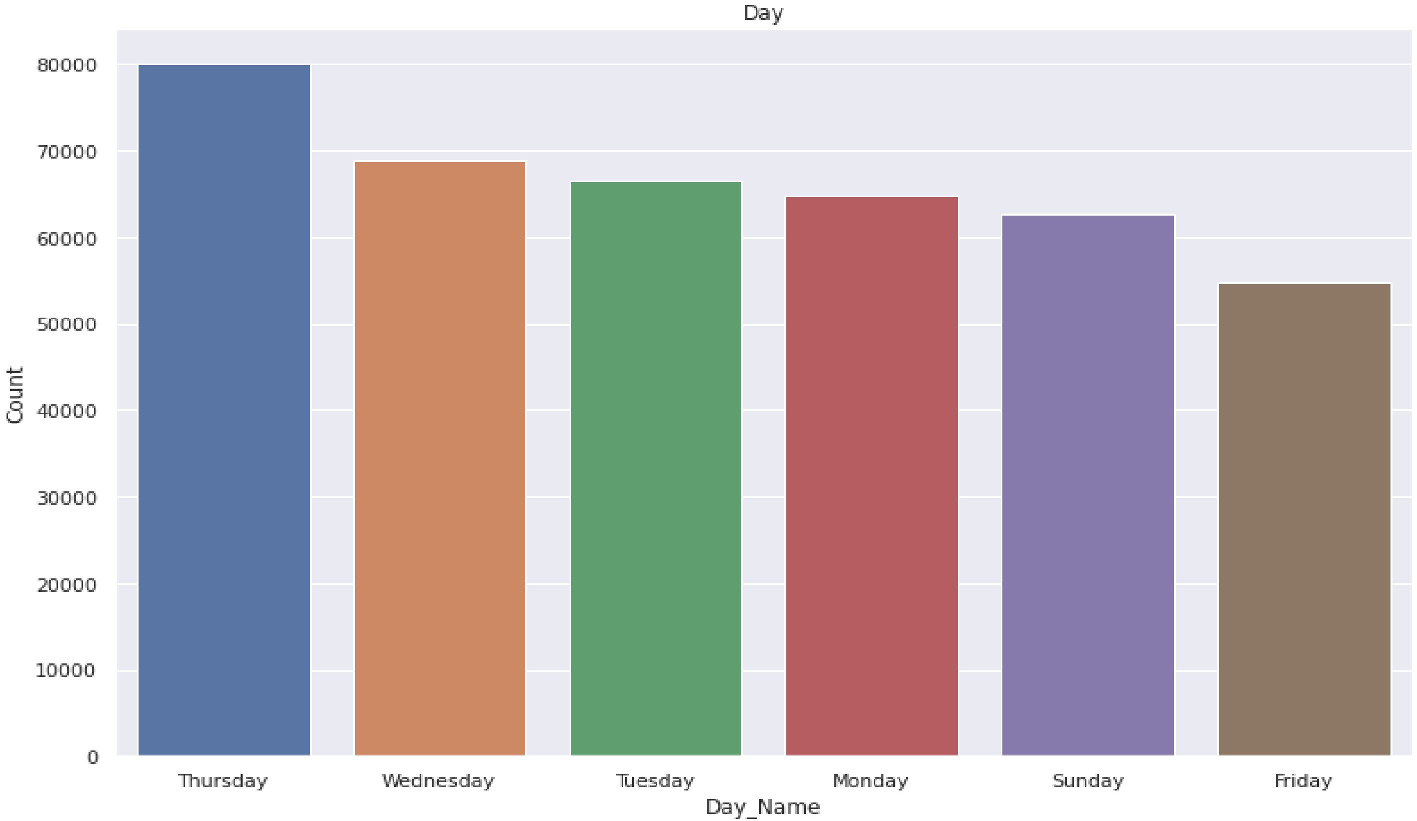
Top 5 Country based on customers



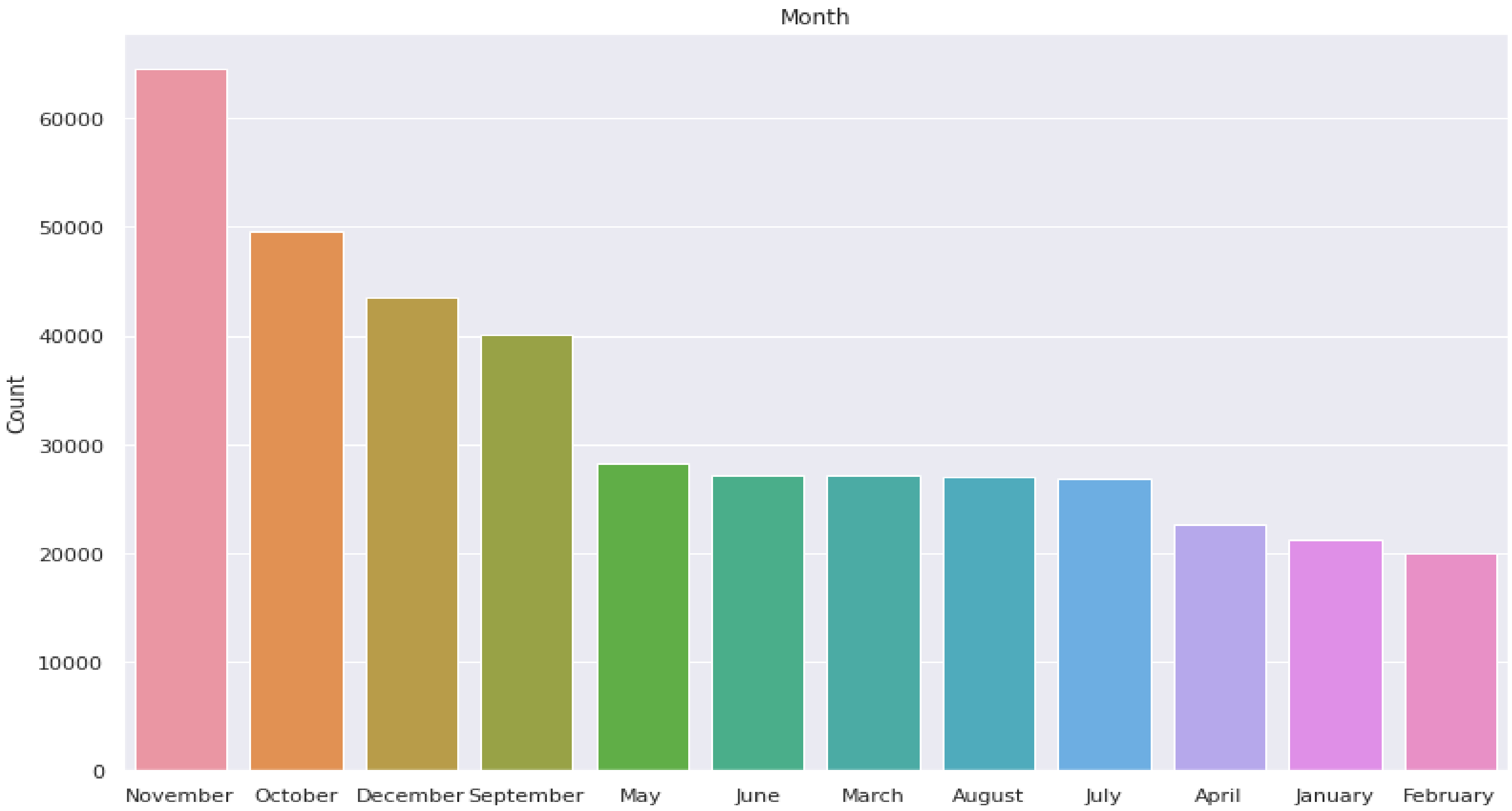
Top 5 Country based least numbers of Customers



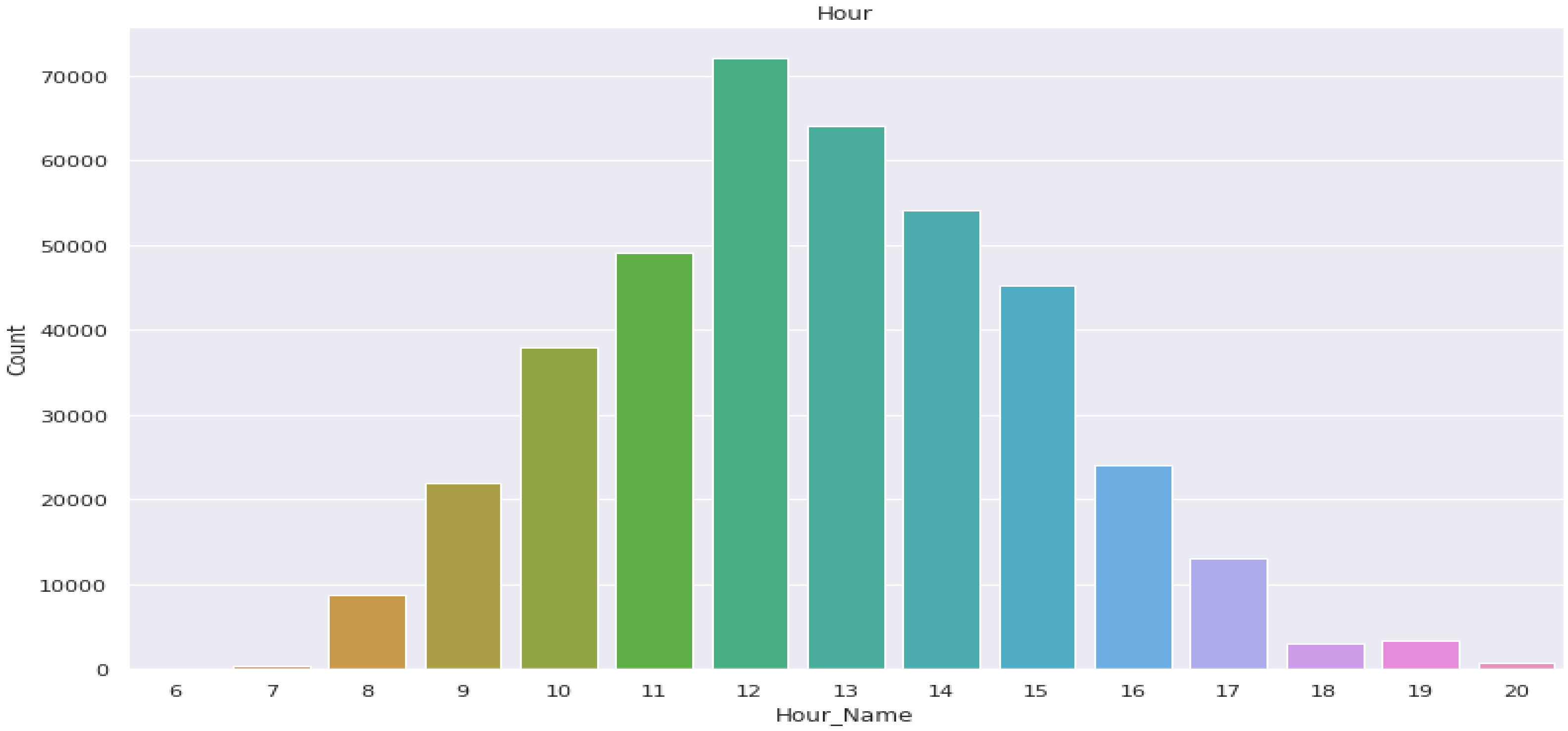
Day wise sales



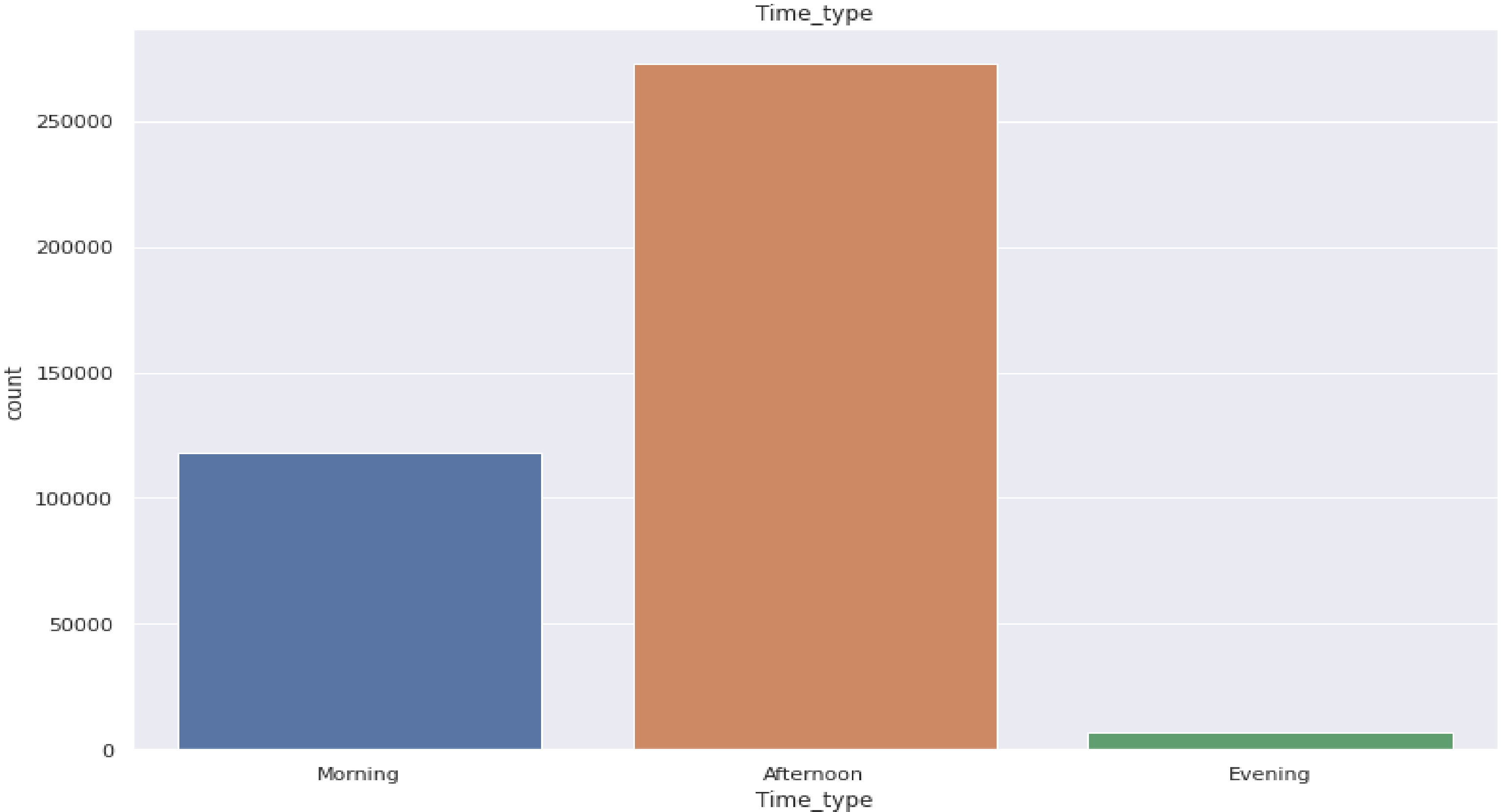
Month Wise Sales



Hour wise Sales



Time Wise Sales



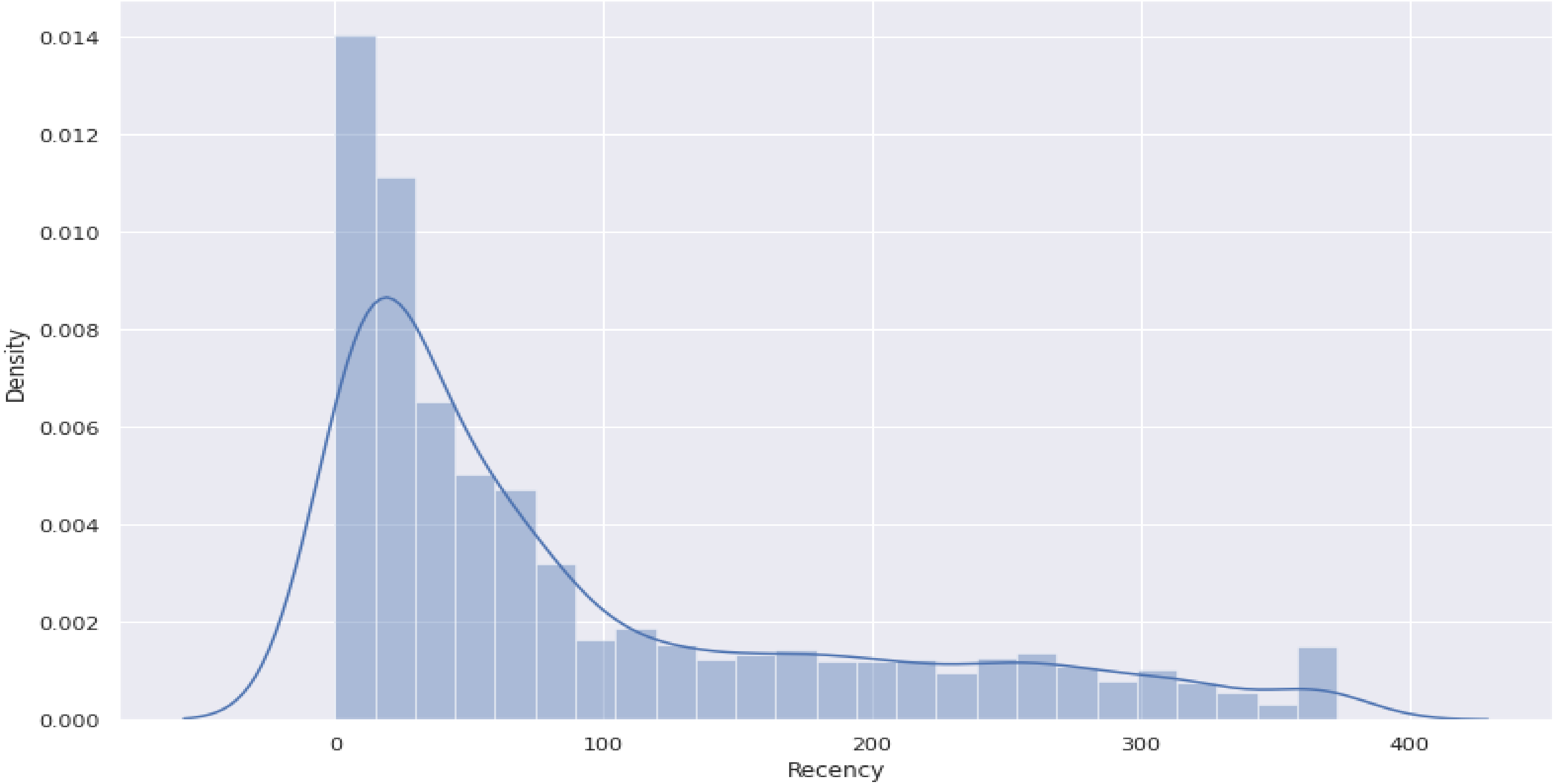
Creating RFM Model

- RFM model (Recency, Frequency,Monetary value)
- Recency, frequency, monetary value is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors: . Frequency: How often a customer makes a purchase. Monetary Value: How much money a customer spends on Performing RFM Segmentation and RFM Analysis, Step by Step 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), using Excel or another tool.

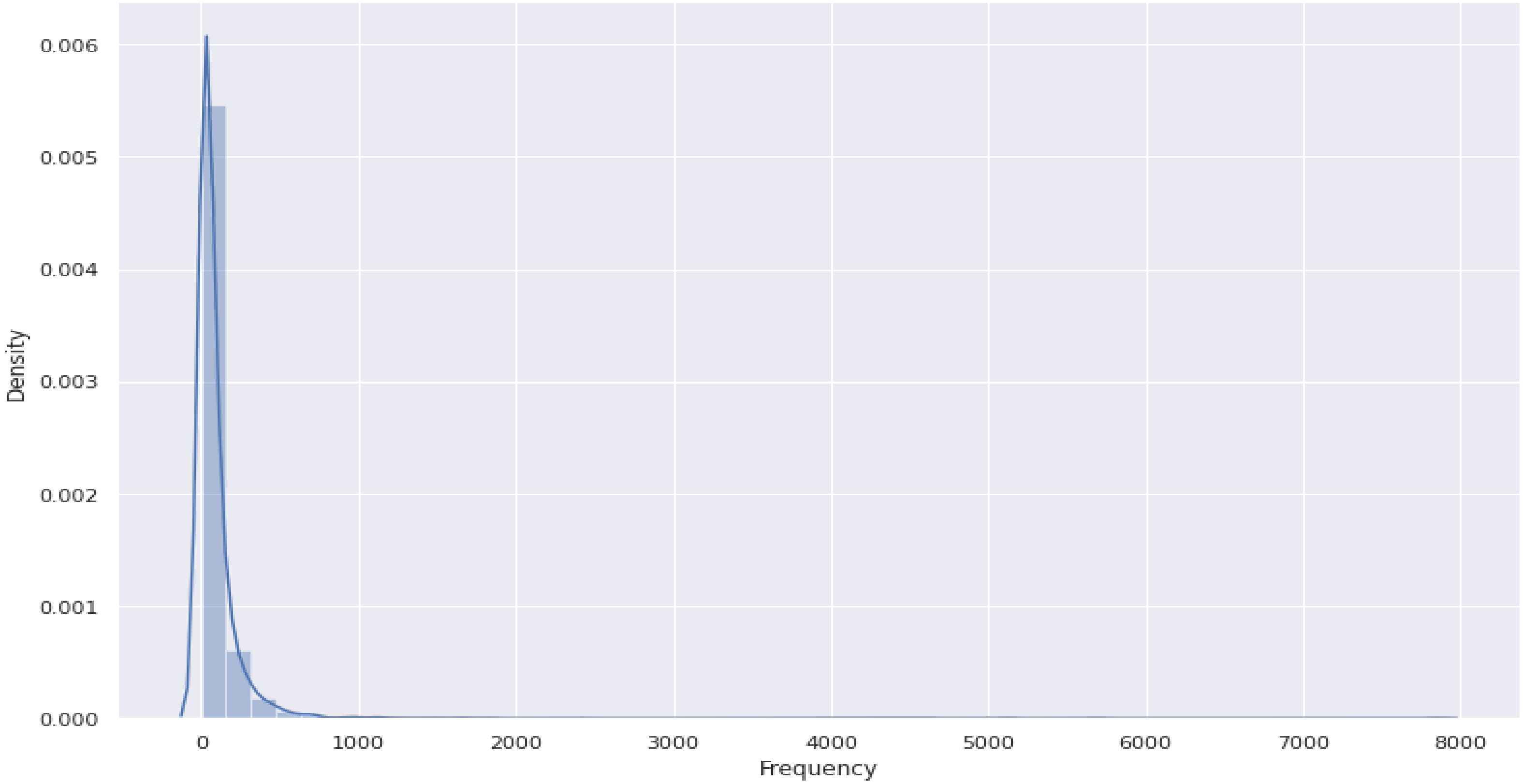
Calculating RFM Score

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.

Recency Distribution



Frequency Distribution



K Means Clustering

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter ‘K’ in K-means. In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster.

DBSCAN

DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers). The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.

Hierarchical Clustering

- In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called **Dendrogram** (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or clusters are broken up (top-down view).
- The basic method to generate hierarchical clustering is
- **1. Agglomerative:** Initially consider every data point as an **individual** Cluster and at every step, **merge** the nearest pairs of the cluster. (It is a bottom-up method). At first, every dataset is considered as an individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.
- The algorithm for Agglomerative Hierarchical Clustering is:
  - Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
  - Consider every data point as an individual cluster
  - Merge the clusters which are highly similar or close to each other.
  - Recalculate the proximity matrix for each cluster
  - Repeat Steps 3 and 4 until only a single cluster remains.

- **2. Divisive:**
- We can say that the Divisive Hierarchical clustering is precisely the **opposite** of the Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren’t comparable. In the end, we are left with N clusters.

Calculation of Silhouette score

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 $S=(b-a)max(a,b)$

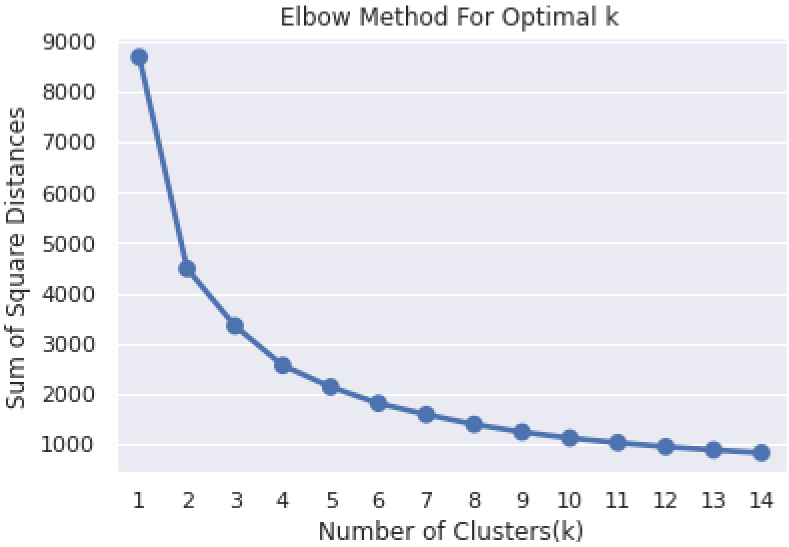
Elbow method

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

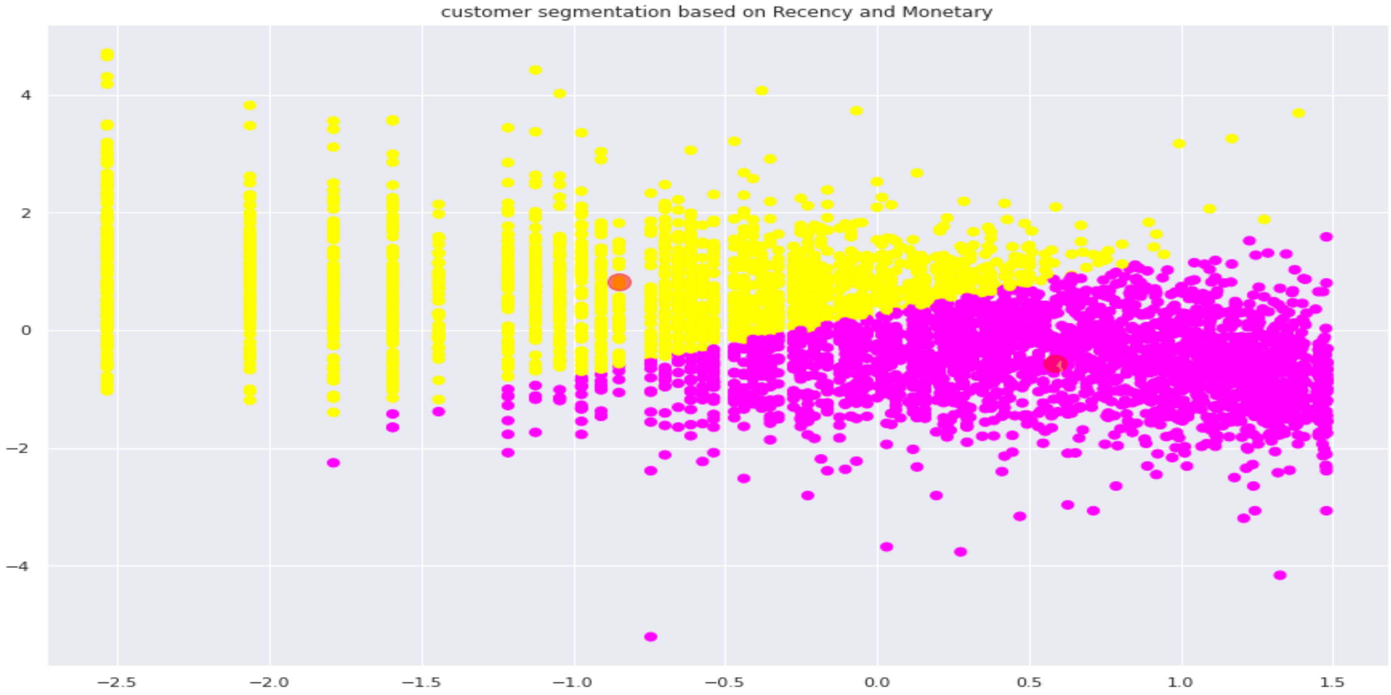
Dendrogram

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

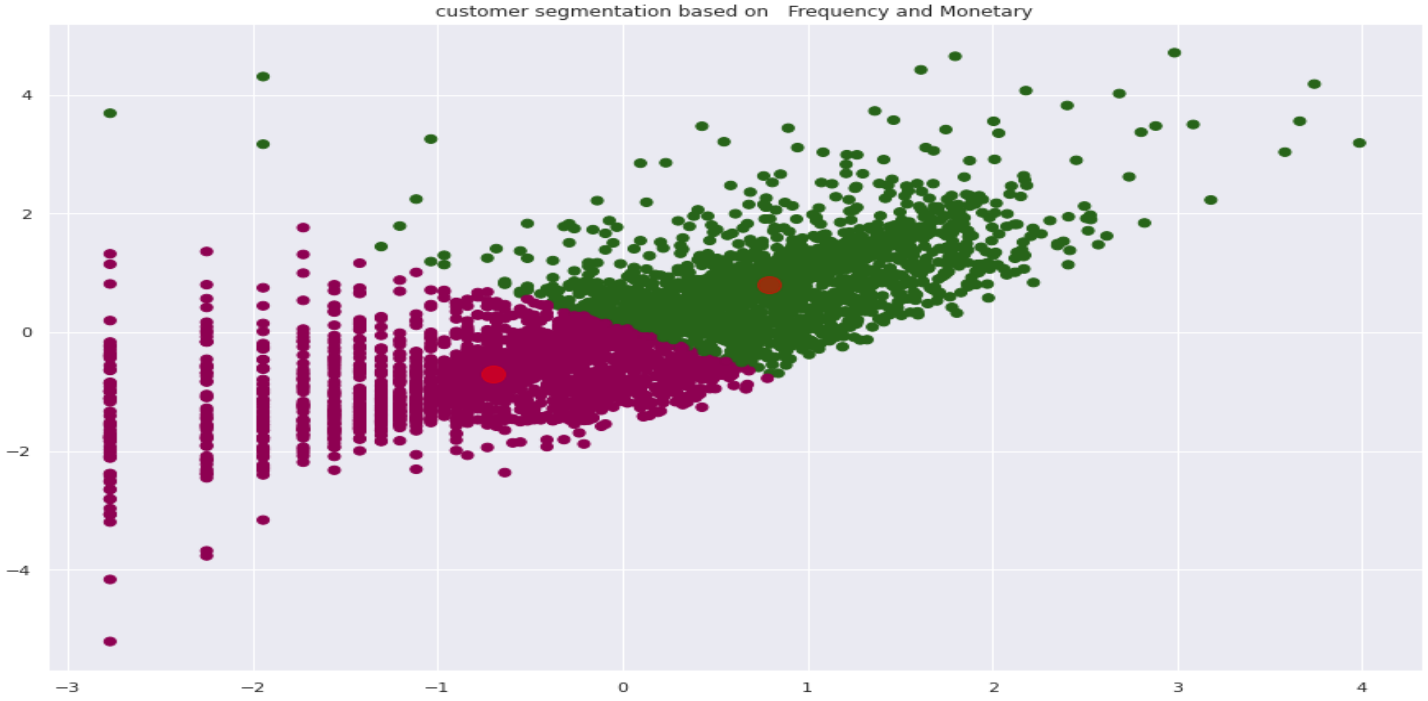
Elbow method for optimum k



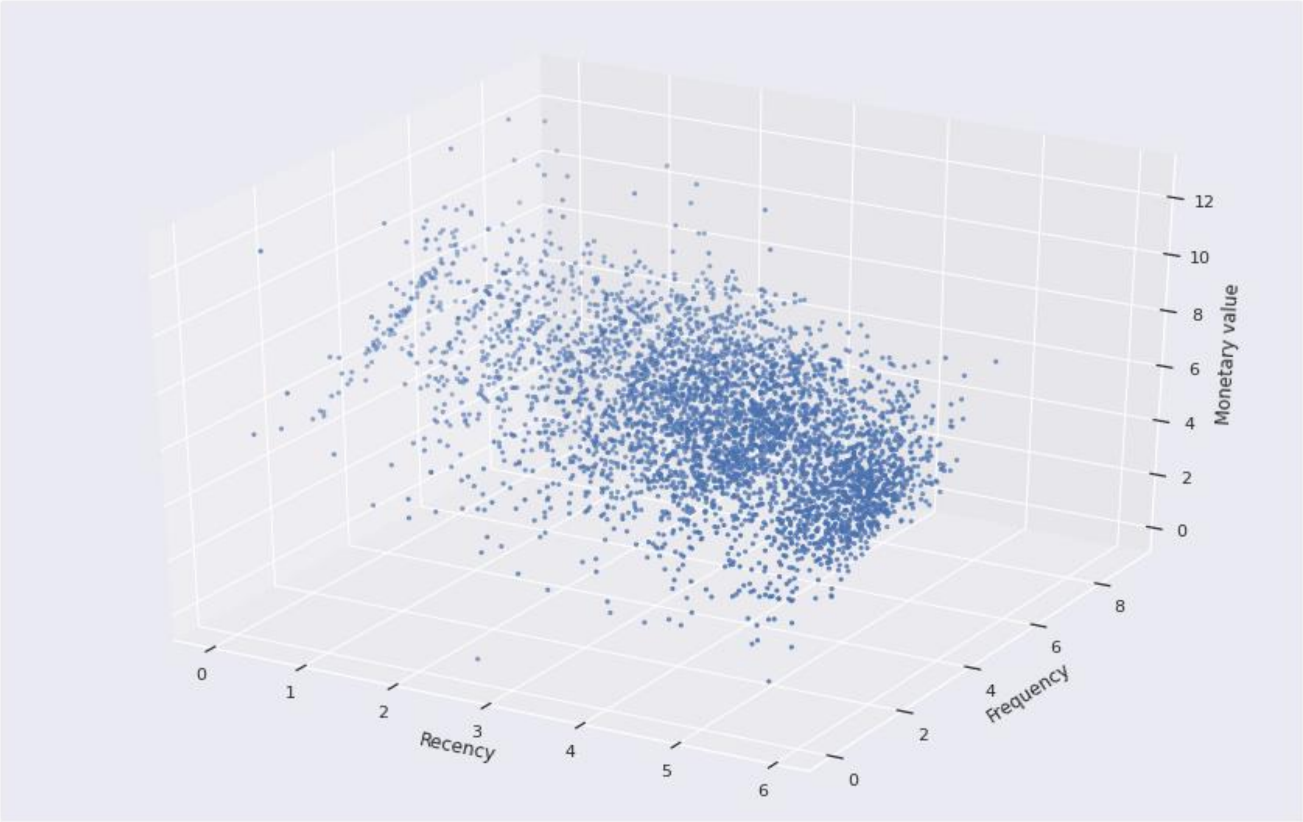
Customer segmentation based on Recency & Monetary



Customer segmentation based on Frequency & Monetary



3D visualization of RFM



Conclusion

From Exploratory Data Analysis

1)Top 5 selling products are WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER, JUMBO BAG RED RETROSPOT, ASSORTED COLOUR BIRD ORNAMENT and PARTY BUNTING.

2) Top 5 stock name based on selling 1. 85123A 2. 22423 3. 85099B 4. 84879 5. 47566

3) Most of the customers are from unitedkingdom,Germany,france,eire,spain

4) Most of the customer buy things on thursday,Wednesday and Tuesday.

5) most numbers of customers have purches the gifts in the month of November ,October and December September

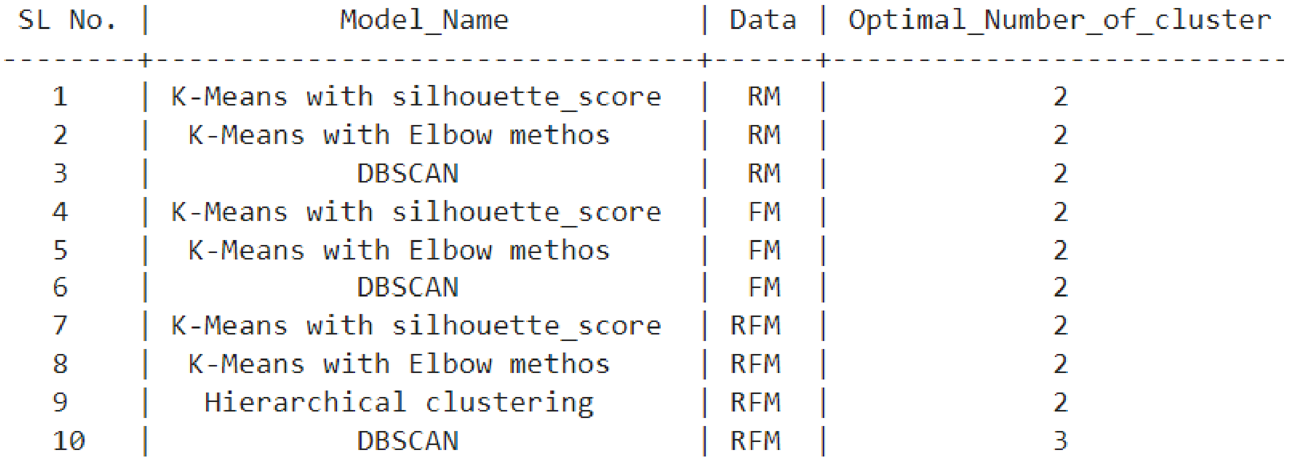
6)Less numbers of customers have purches the gifts in the month of April ,january and February.

7)In AfterNone Time most of the customers have purches the item.

8) Most of the customers have purches the items in Aftrnoon ,moderate numbers of customers have purches the items in Morning and least numbers of customers have purches the items in Evening.

Conclusion

The optimum number of clusters from all algorithms is 2 but in some RFM cases and in DBSCAN it shows 3.But as in lot of cases its 2 so we use two clusters to represent our clustering algorithm in this case.



```
 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
```