Capstone Project 4

Online Retail Customer Segmentation

Kiran Ahire

Points for Discussion

- Business Objective
- Data Summary
- Feature Summary
- Data Preprocessing
- Exploratory Data Analysis
- Corelation between data
- Algorithms implementation
- Conclusion

Business Objective

- Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.
- Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators.
- Given the dataset, our objective is to build clustering model that would perform Customer Segmentation.

Data Summary

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

- There are total 541909 rows and 8 columns in our dataset.
- This is a transnational dataset with transactions occurring between 1st December 2010 and 9th December 2011 for UK based online retailer.
- Many customers of the company are wholesalers.

Feature Summary

We have given the following information in our dataset:

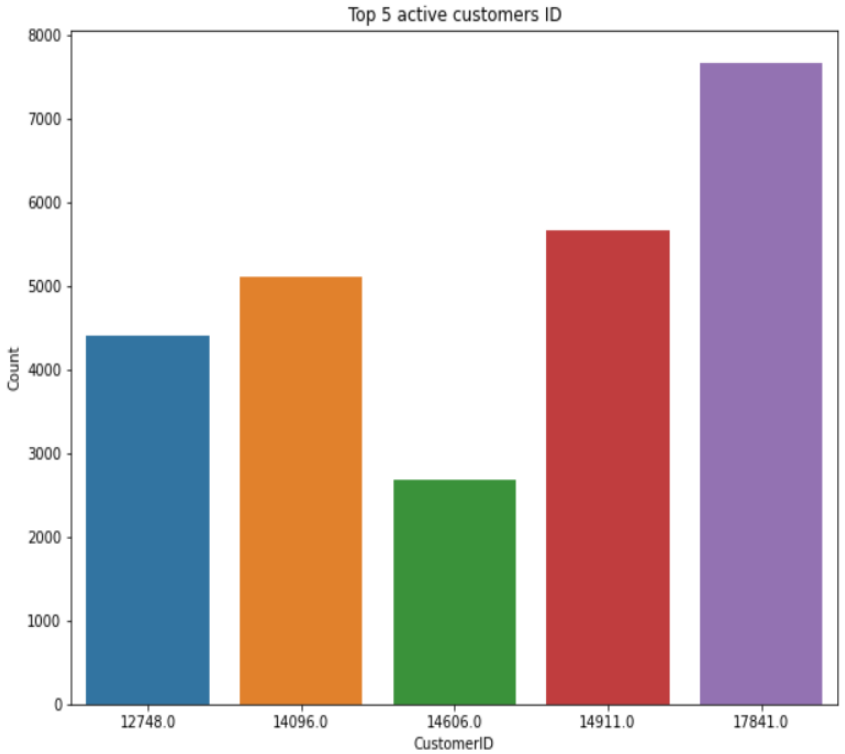
- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product
- Description: Product (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.

Data Preprocessing

- The dataset contains 541909 rows and 8 columns
- There are 4 categorical features namely ‘InvoiceNo’ , ‘StochCode’,’Country’ and ‘Description’
- There are missing values present in ‘Description’ and ‘CustomerID’ columns and removed null values.
- There are duplicate values present so removed them.
- One Datetime feature ‘InvoiceDate’
- Outliers present in ‘Quantity’ and ‘UnitPrice’ column.
- Removed cancelled orders
- Added new features from Datetime columns such as ‘Day’, ’Month’ and ’Hour’
- Added new feature ‘Total Amount’

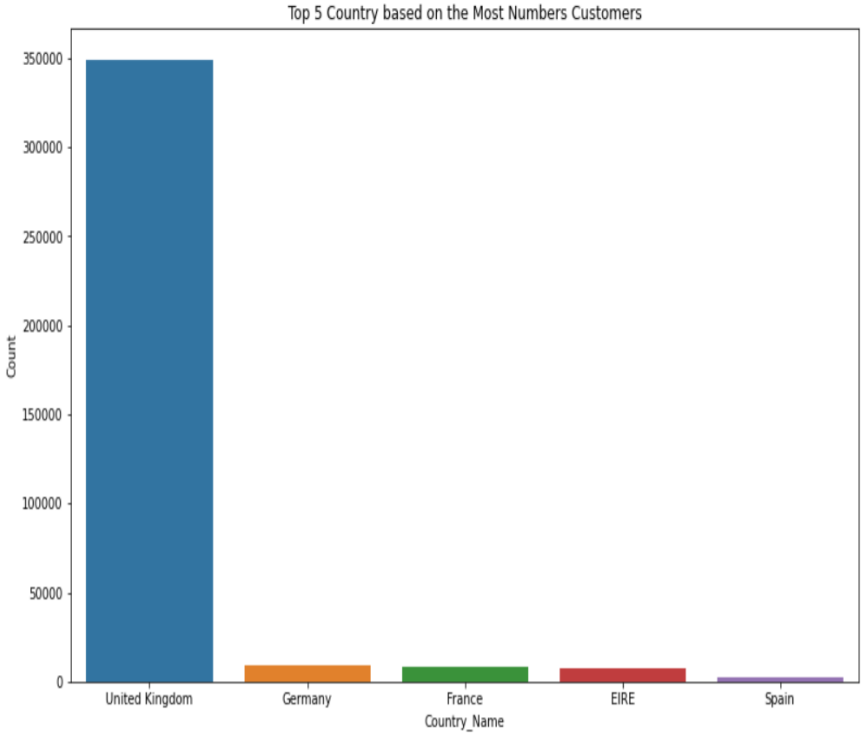
Analysis of CustomerID

- There are 4339 unique Customers id
- Customer with id 17841 is the most Active customer.



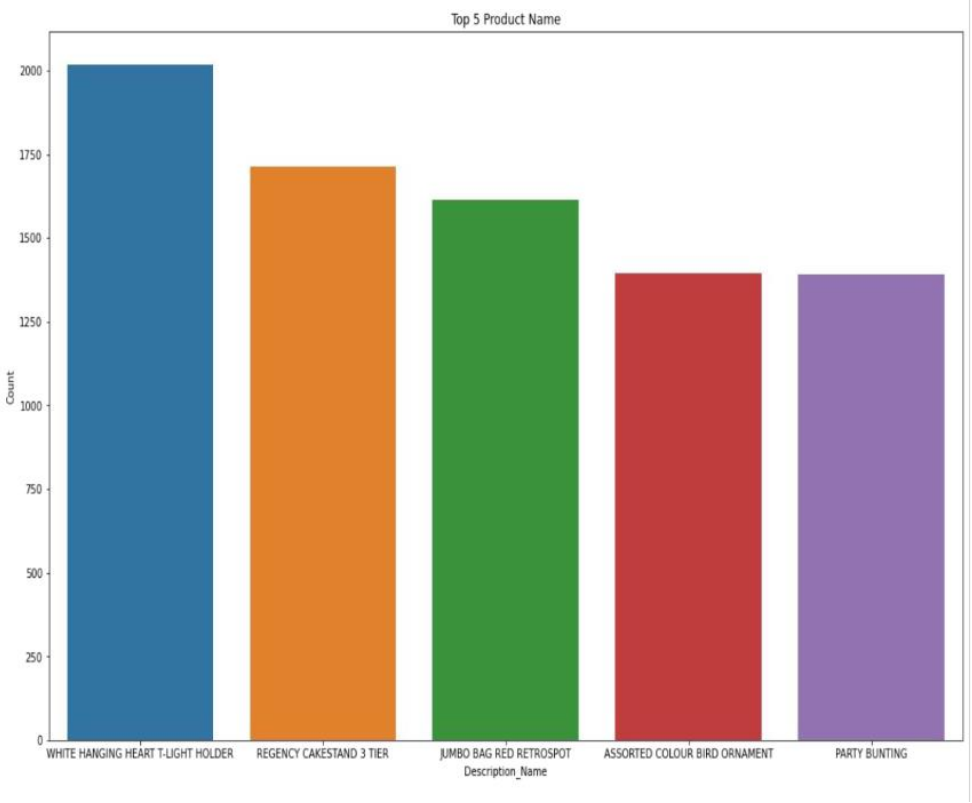
Analysis of Country

- Since the Data belonged to UK company, UK had majority of the customers.
- UK , Germany and France were top countries having most no. of customers.



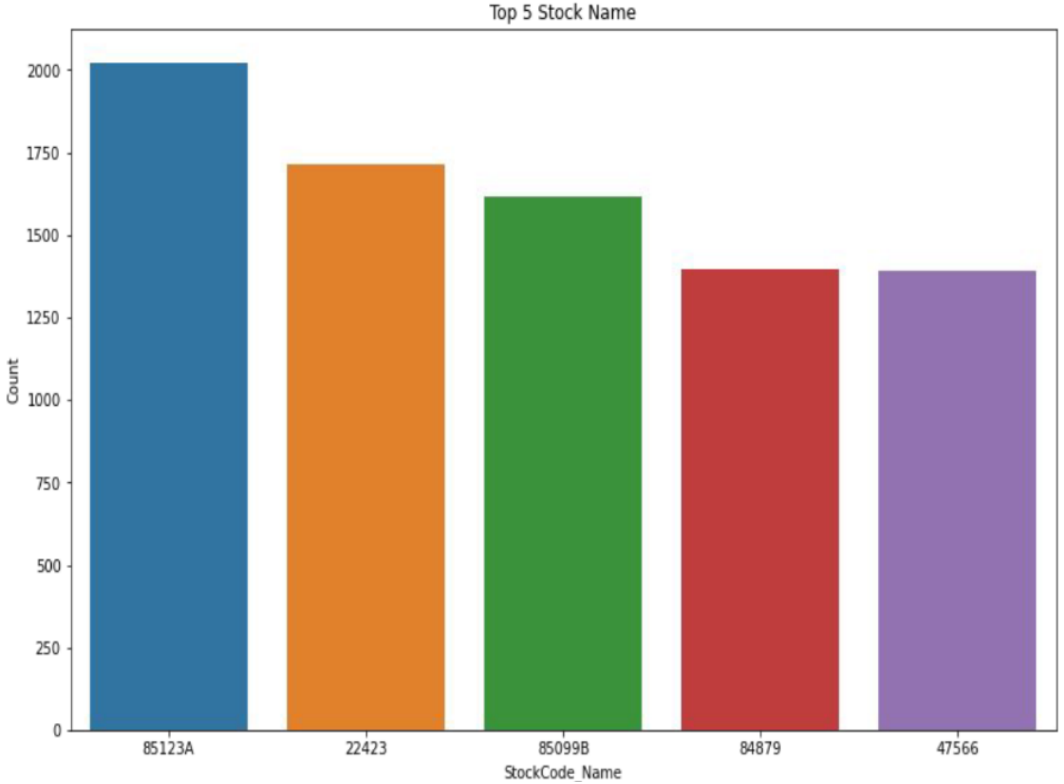
Analysis of Description

	Description_Name	Count
0	WHITE HANGING HEART T-LIGHT HOLDER	2016
1	REGENCY CAKESTAND 3 TIER	1714
2	JUMBO BAG RED RETROSPOT	1615
3	ASSORTED COLOUR BIRD ORNAMENT	1395
4	PARTY BUNTING	1390

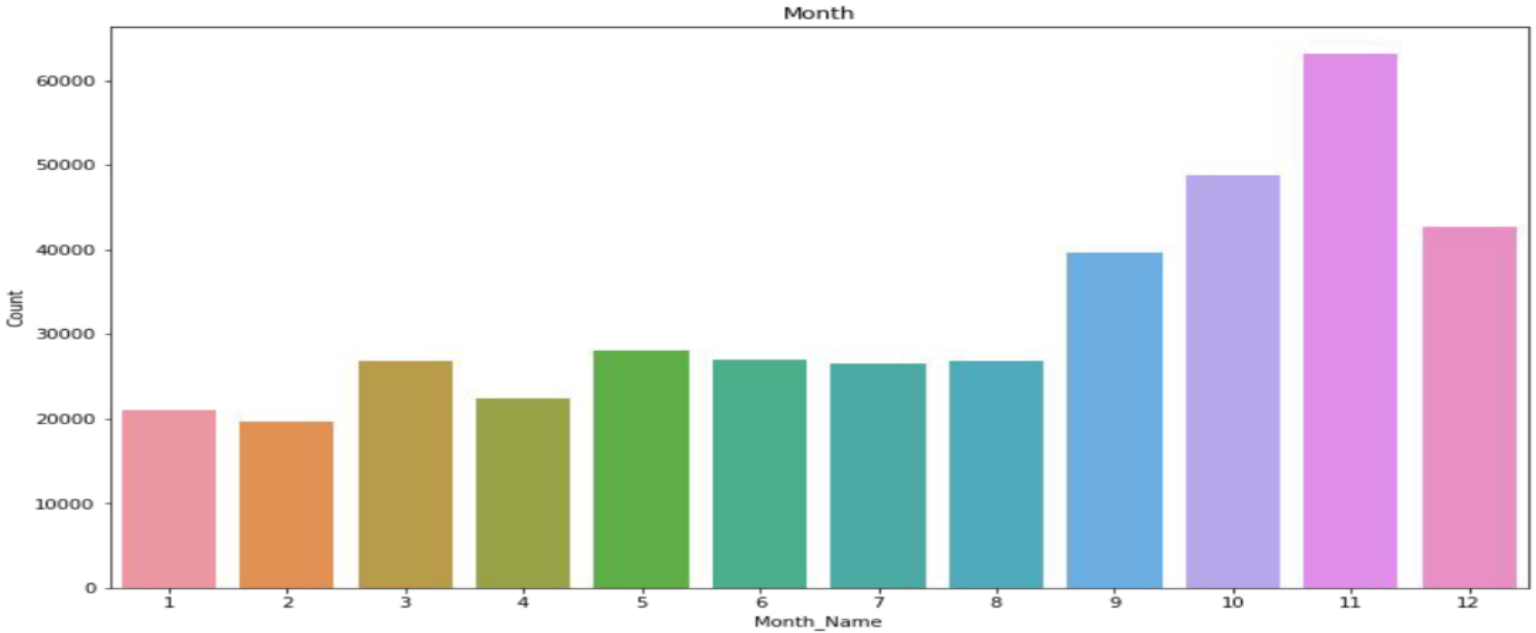


Analysis of StockCode

	StockCode_Name	Count
0	85123A	2023
1	22423	1714
2	85099B	1615
3	84879	1395
4	47566	1390

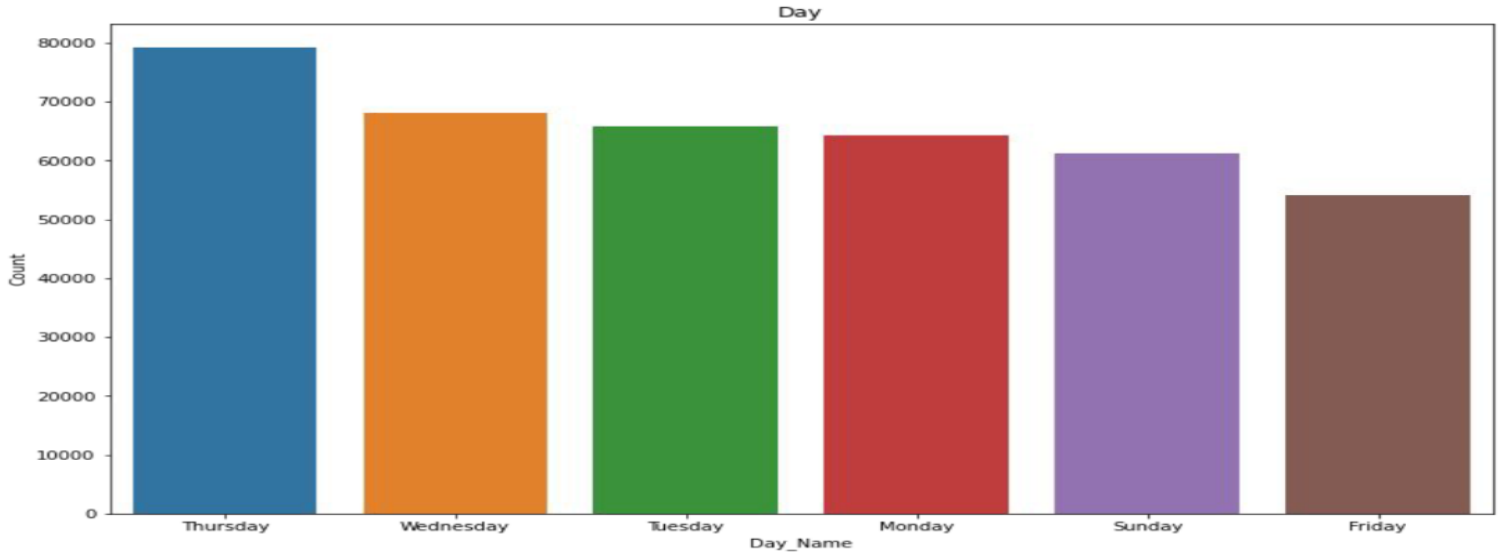


Analysis of Month



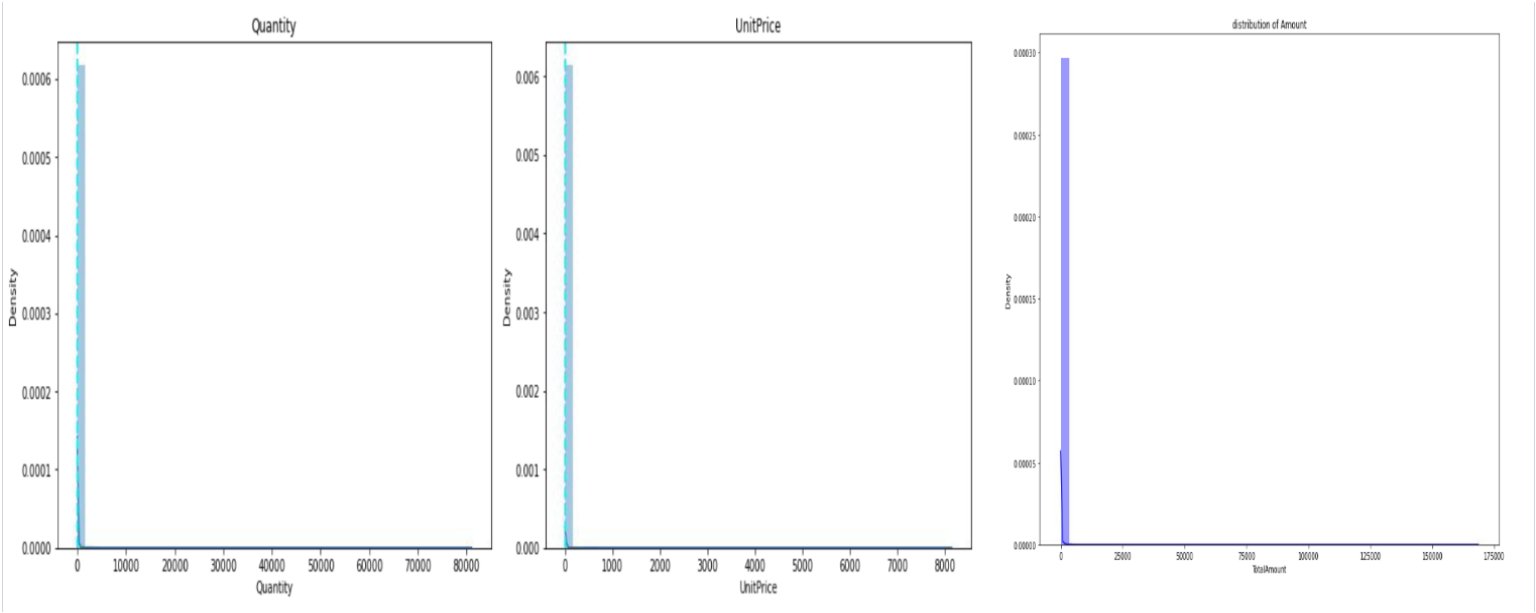
November month has highest sales followed by October and December.

Analysis of Day



- Most of the customers have purchased items on Thursday , Wednesday and Tuesday.

Analysis of Numerical variable



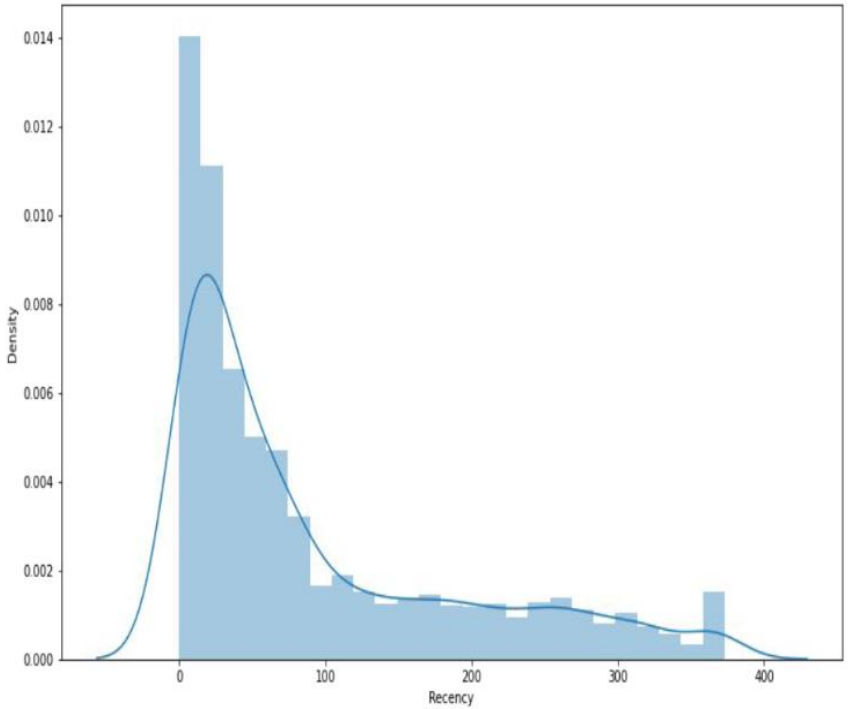
- Highly positively skewed data we need to do log transformation.

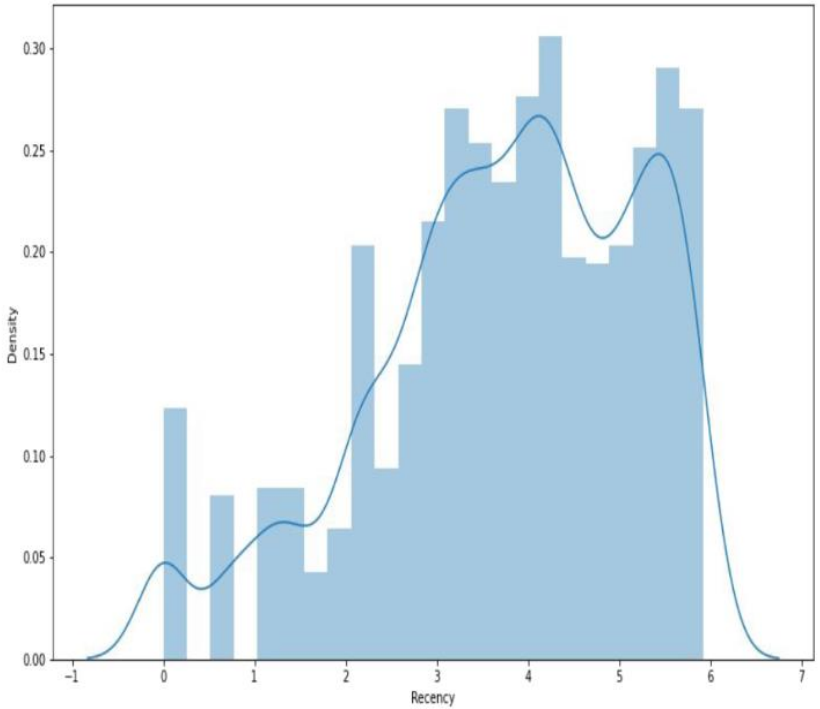
RFM Model

- RFM model which stands for Recency, Frequency, and Monetary is one of such steps in which we determine the recency - days to last visit, frequency - how actively the customer repurchases and monetary - total expenditure of the customer, for each customer
- RFM analysis 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.

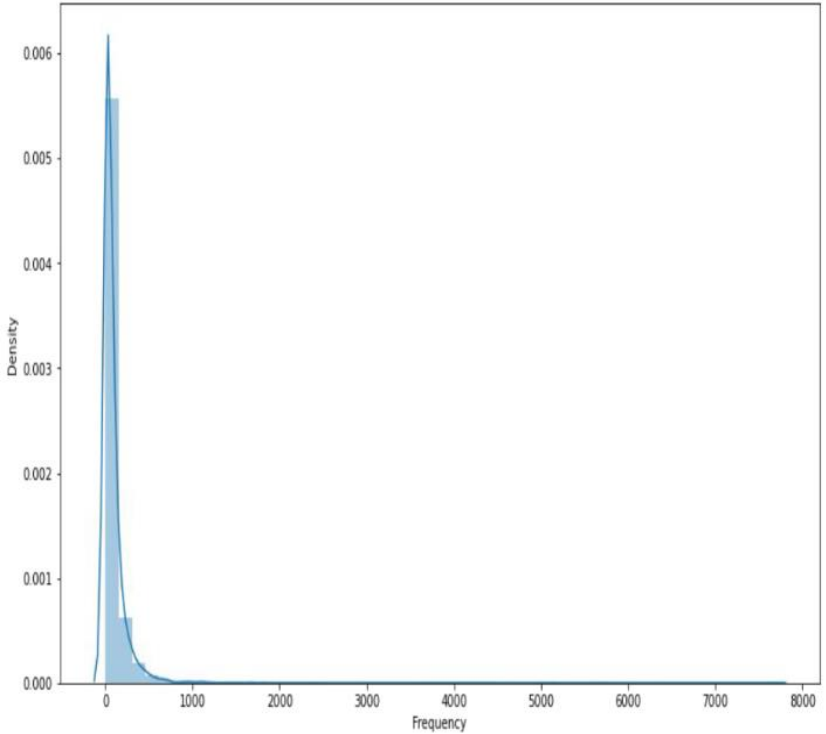
  Recency = Latest Date - Last Inovice Data,
- Frequency = count of invoice no. of transaction(s),
- Monetary = Sum of Total Amount for each customer

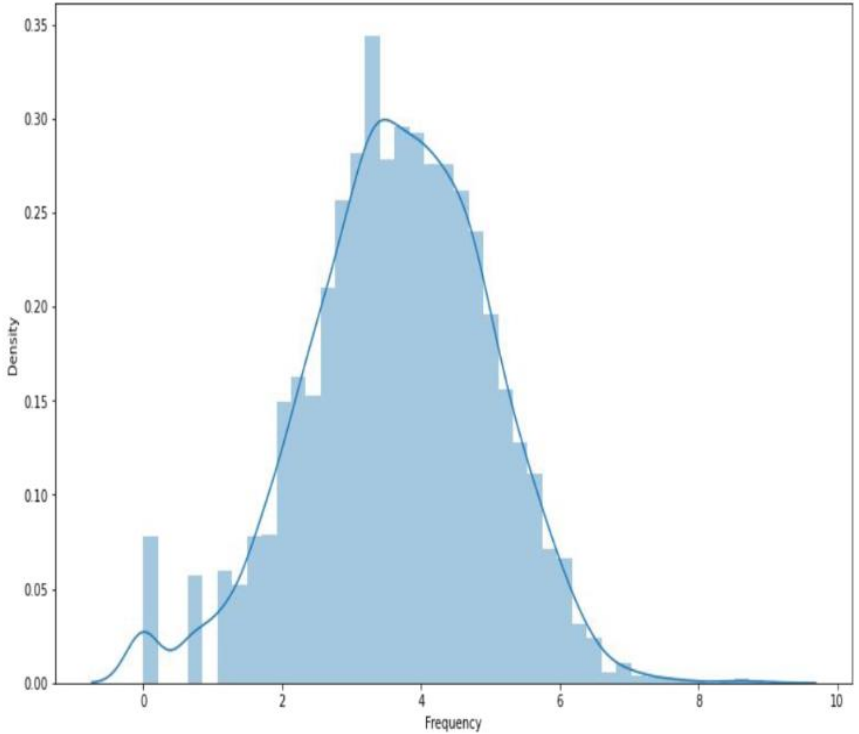
Recency


0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
100
200
300
400
Recency

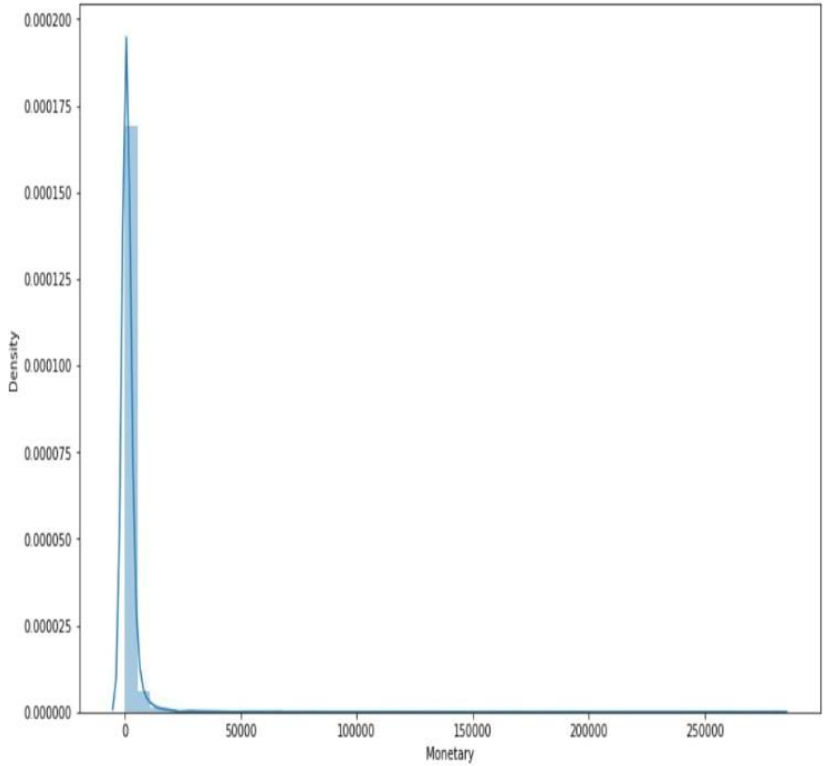

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
−1
0
1
2
3
4
5
6
7
Recency

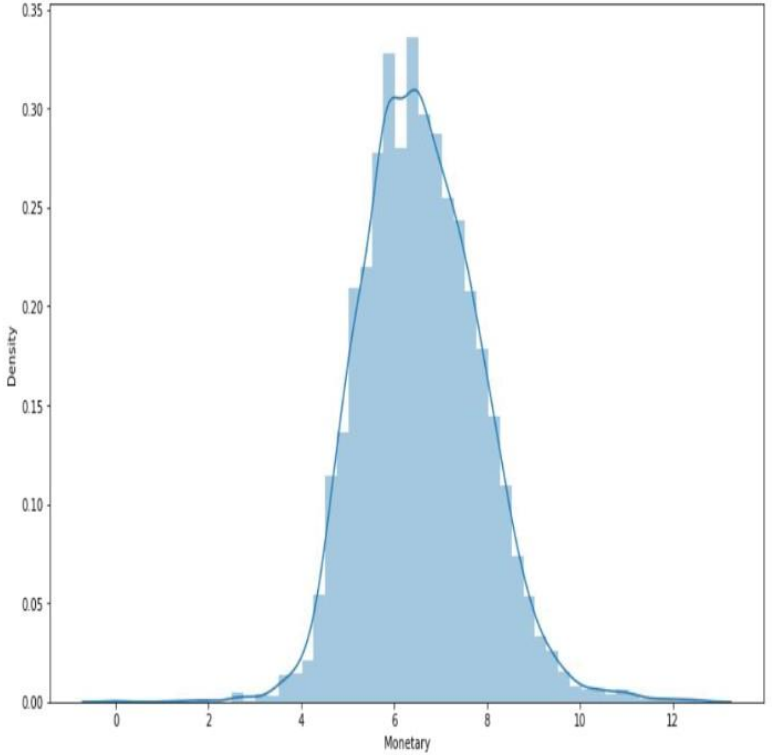
Frequency


Density
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
1000
2000
3000
4000
5000
6000
7000
8000
Frequency

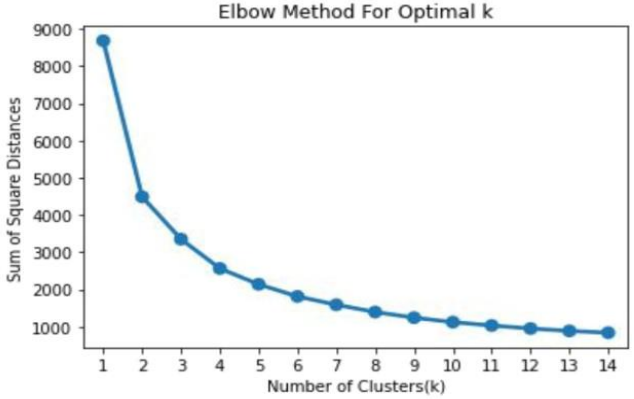

Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
Frequency

Monetary

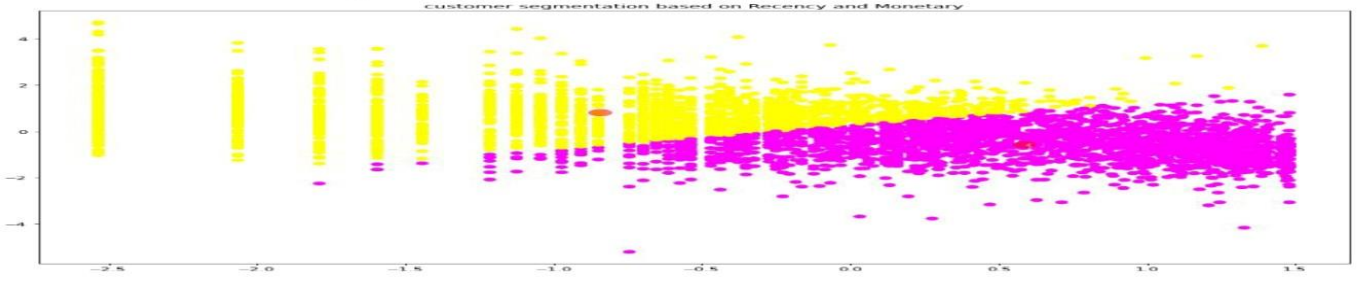

0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
Density
0
50000
100000
150000
200000
250000
Monetary


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
2
4
6
8
10
12
Monetary

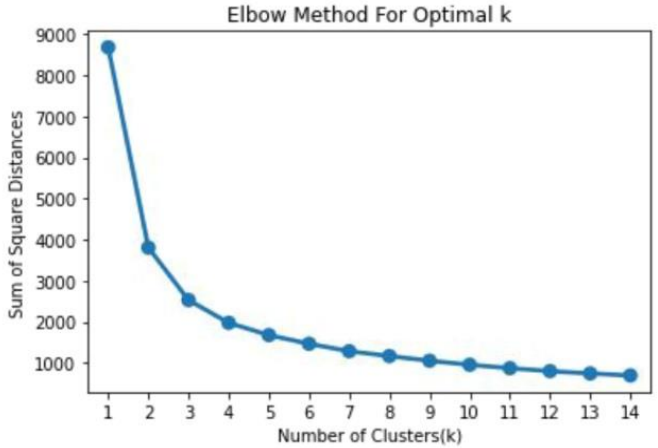
Silhoutte score and Elbow method on R&M



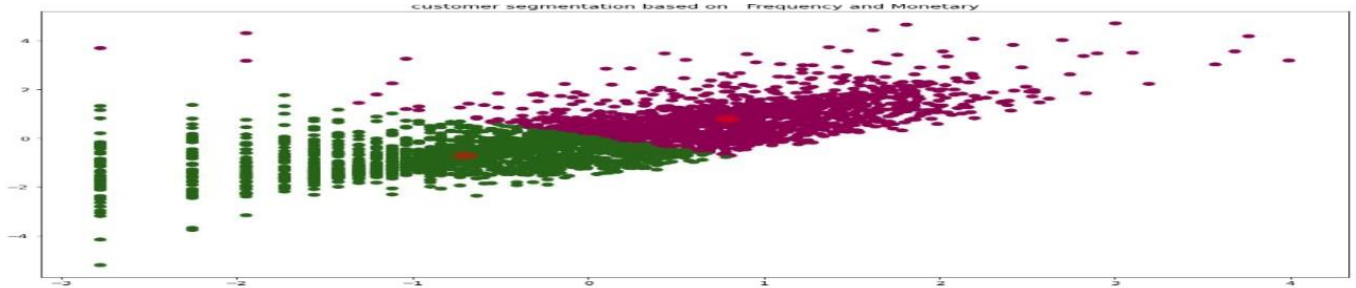

```
For n_clusters = 2, silhouette score is 0.42047430853642515
For n_clusters = 3, silhouette score is 0.34263694998026195
For n_clusters = 4, silhouette score is 0.36471463504091317
For n_clusters = 5, silhouette score is 0.3373938872753767
For n_clusters = 6, silhouette score is 0.34314878344616223
For n_clusters = 7, silhouette score is 0.34497773349086586
For n_clusters = 8, silhouette score is 0.33799261511793943
For n_clusters = 9, silhouette score is 0.34607062536188865
For n_clusters = 10, silhouette score is 0.3475247420875672
For n_clusters = 11, silhouette score is 0.33693656579850056
For n_clusters = 12, silhouette score is 0.3373606876306994
For n_clusters = 13, silhouette score is 0.3389910656356672
For n_clusters = 14, silhouette score is 0.34138143812594274
For n_clusters = 15, silhouette score is 0.33827205107215497
```



Silhoutte score and Elbow method on F&M



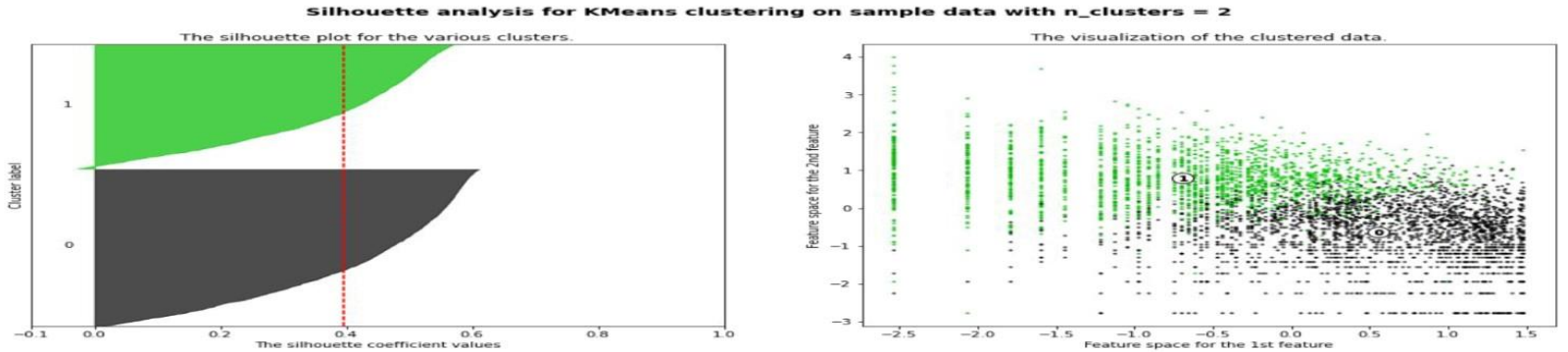

```
For n_clusters = 2, silhouette score is 0.4784099179679686
For n_clusters = 3, silhouette score is 0.40773549715950697
For n_clusters = 4, silhouette score is 0.37231818810915773
For n_clusters = 5, silhouette score is 0.3466695882675493
For n_clusters = 6, silhouette score is 0.36216641752478196
For n_clusters = 7, silhouette score is 0.3388444115986888
For n_clusters = 8, silhouette score is 0.3502006422275672
For n_clusters = 9, silhouette score is 0.3463581243091397
For n_clusters = 10, silhouette score is 0.359712396199174
For n_clusters = 11, silhouette score is 0.3667563841808519
For n_clusters = 12, silhouette score is 0.35450579287148154
For n_clusters = 13, silhouette score is 0.3522388459765374
For n_clusters = 14, silhouette score is 0.3656296509855817
For n_clusters = 15, silhouette score is 0.3387892798152116
```



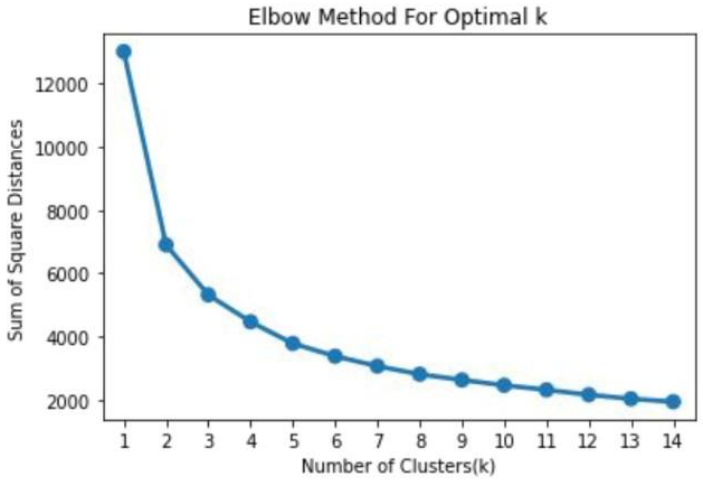
Silhoutte score on R , F & M

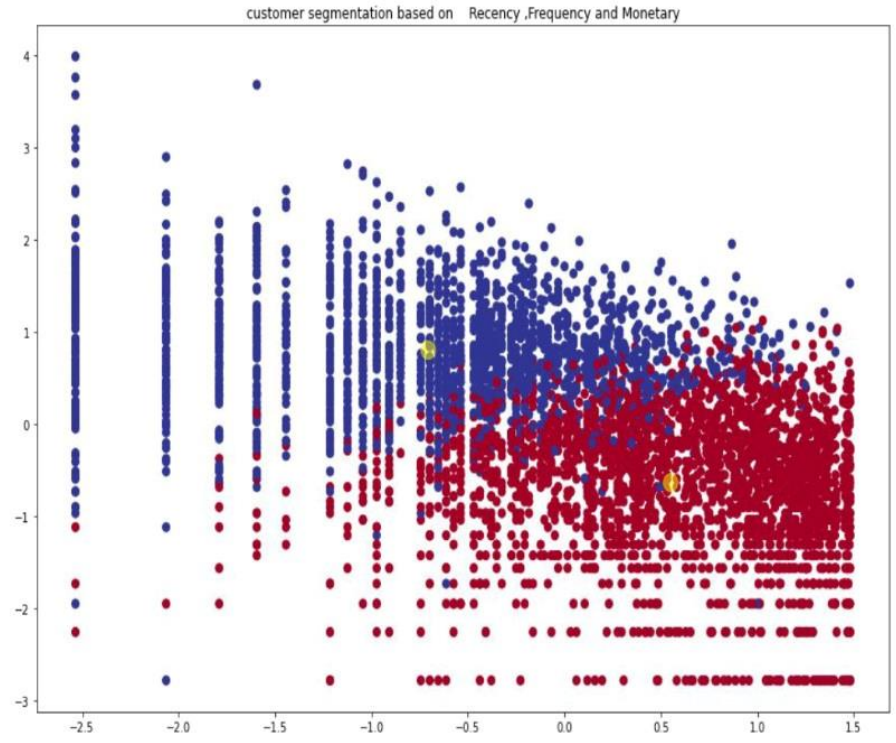
```
For n_clusters = 2 The average silhouette_score is : 0.395588261083924
For n_clusters = 3 The average silhouette_score is : 0.3056346992700891
For n_clusters = 4 The average silhouette_score is : 0.30270317426951315
For n_clusters = 5 The average silhouette_score is : 0.2792460561657022
For n_clusters = 6 The average silhouette_score is : 0.27920490924462127
For n_clusters = 7 The average silhouette_score is : 0.26639589282641335
For n_clusters = 8 The average silhouette_score is : 0.26439181568270587
For n_clusters = 9 The average silhouette_score is : 0.2530950678302339
For n_clusters = 10 The average silhouette_score is : 0.2523587969335804
For n_clusters = 11 The average silhouette_score is : 0.26053571388536995
For n_clusters = 12 The average silhouette_score is : 0.2658660395456048
For n_clusters = 13 The average silhouette_score is : 0.2633746937000211
For n_clusters = 14 The average silhouette_score is : 0.25492771945211434
For n_clusters = 15 The average silhouette_score is : 0.25490689188578847
```

Silhoutte Analysis on R , F & M



Elbow method and cluster chart on RFM

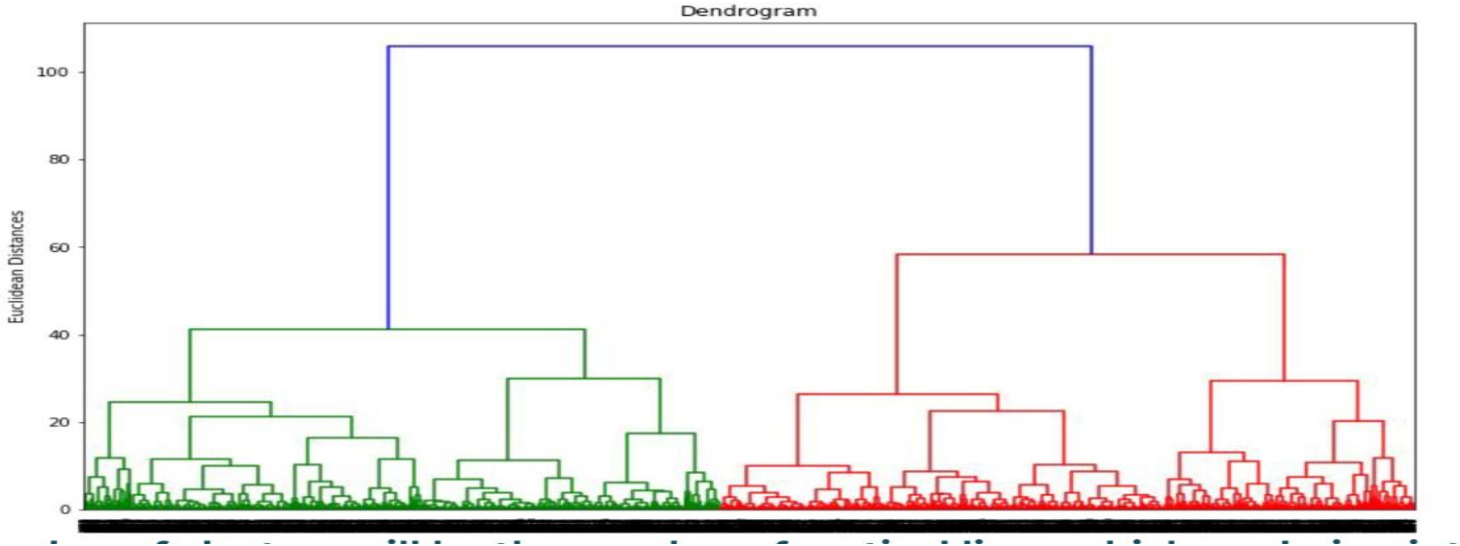




RFM Analysis

CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	Recency_log	Frequency_log	Monetary_log	Cluster
12346.0	325	1	77183.60	4	4	1	441	9	5.783825	0.000000	11.253942	1
12347.0	2	182	4310.00	1	1	1	111	3	0.693147	5.204007	8.368693	0
12348.0	75	31	1797.24	3	3	1	331	7	4.317488	3.433987	7.494007	1
12349.0	18	73	1757.55	2	2	1	221	5	2.890372	4.290459	7.471676	0
12350.0	310	17	334.40	4	4	3	443	11	5.736572	2.833213	5.812338	1
12352.0	36	85	2506.04	2	2	1	221	5	3.583519	4.442651	7.826459	0
12353.0	204	4	89.00	4	4	4	444	12	5.318120	1.386294	4.488636	1
12354.0	232	58	1079.40	4	2	2	422	8	5.446737	4.060443	6.984161	1
12355.0	214	13	459.40	4	4	3	443	11	5.365976	2.564949	6.129921	1
12356.0	22	59	2811.43	2	2	1	221	5	3.091042	4.077537	7.941449	0

Hierarchical Clustering



- The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold=90
- No. of Cluster = 2

Conclusion

- Throughout the analysis we went through various steps to perform customer segmentation. We started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications. Next, we did some exploratory data analysis and tried to draw observations from the features we had in the dataset.
- Then, we formulated some quantitative factors such as recency, frequency and monetary known as rfm model for each of the customers. We implemented KMeans clustering algorithm on these features. We also performed silhouette and elbow method analysis to determine the optimal no. of clusters which was 2. We saw customers having high recency and low frequency and monetary values were part of one cluster and customers having low recency and high frequency, monetary values were part of another cluster.