Capstone Project-4

Online Retail Customer Segmentation

Individual Member
kota Lakshmana Rao

Content

- BUSINESS UNDERSTANDING
- DATA SUMMARY
- FEATURE ANALYSIS
- EXPLORATORY DATA ANALYSIS
- DATA PREPROCESSING
- IMPLEMENTING ALGORITHMS
- CHALLENGES
- CONCLUSIONS

Introduction

- 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.

Problem Statement

Identify major customer segments on transactional data using cluster analysis

Points for Discussion

Data summary
Feature summary
Insights from our Dataset
Analysis on top products
Analysis on bottom products
Analysis on Stock Code
Analysis on Country Based

Distribution
Analysis day Wise
Analysis Hour Wise
Analysis Numerical variable
RFM MODEL
Recency
Frequency
Monetary

Calculation of Silhouette score

Silhouette score and Elbow method on R , M

Silhouette score and Elbow method on F, M

Silhouette Analysis on R ,F, M

3D visualization of R, F,M

Elbow method and cluster chart on RFM

RFM ANALYSIS

Hierarchical clustering

DBSCAN ON R,F,M

Challenges
Conclusion

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

- **Total Rows : 541909**
- **Total Column: 8**
- **A transactional data set with transactions occurring between 1st December 2010 and 9th December 2011 for a UK-based online retailer.**
- **Many customers of the company are wholesalers.**

Feature Summary

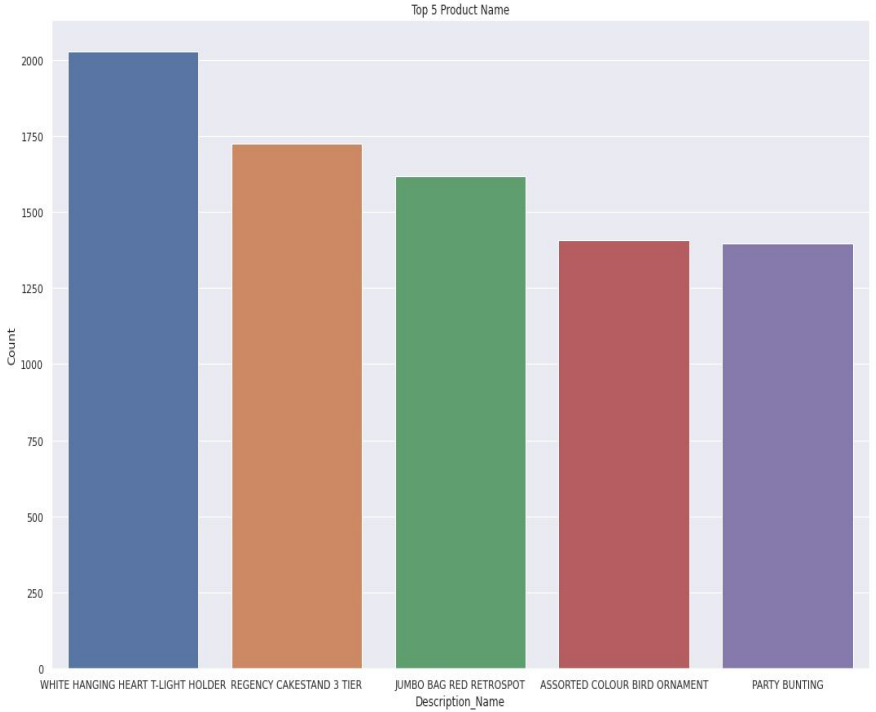
- The contents of the data had features such as:
- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- Stock Code: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- 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 the country where each customer resides.

Insights From our Dataset

- This Dataset is from the UK
- In our data set there are 541909 rows, 8 columns
- Four categorical features ‘InvoiceNo’, ‘Stock Code’, & ‘Description’, ‘ Country’.
- There are Missing Values present on Description & CustomerID columns, Removed null values
- There are Duplicate values present, Removed duplicates
- One Datetime[ns] features ‘InvoiceDate’.
- Outliers present only in “Quantity” & “Unit Price”column.
- Removed cancelled orders.
- Added new features from datetime column such as months, days, hours.
- Added Total Amount
- Converted data types

Analysis On Top Products

1.WHITE HANGING HEART T-LIGHT HOLDER,
2.REGENCY CAKE STAND 3 TIER
3.JUMBO BAG RED RETROSPOT
4.PARTY BUNTING
5.LUNCH BAG RED RETROSHOP



Analysis on Bottom Products

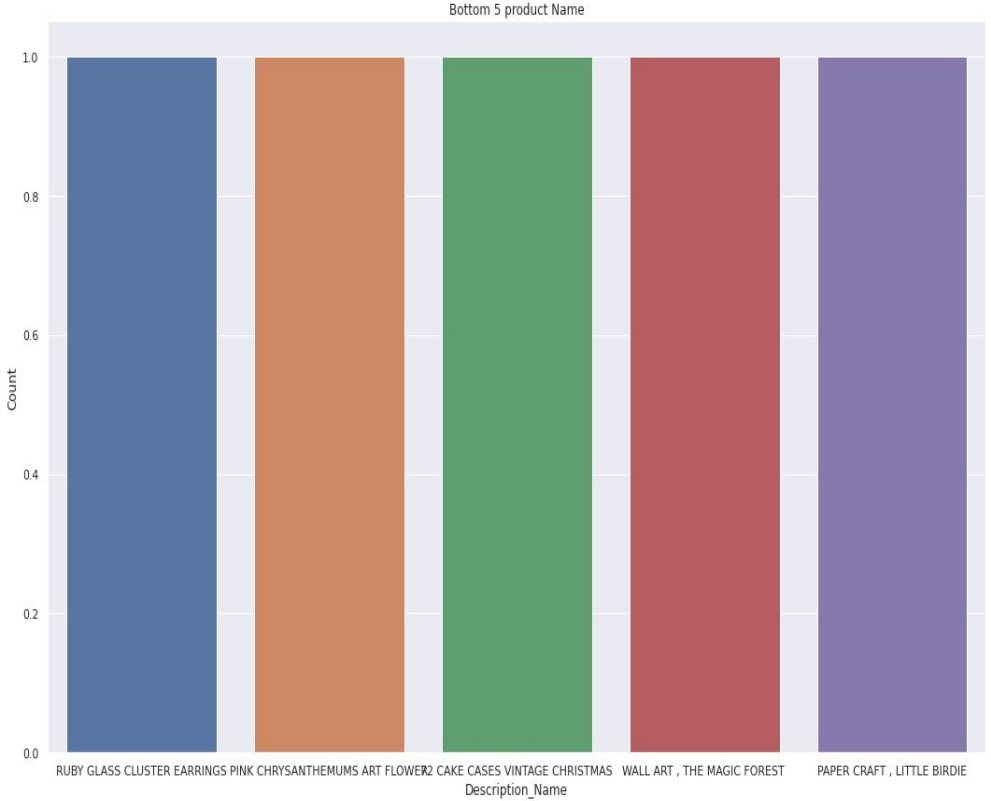
1.LIGHT DECORATION BATTERY OPERATED

2.Water damaged

3.throw away

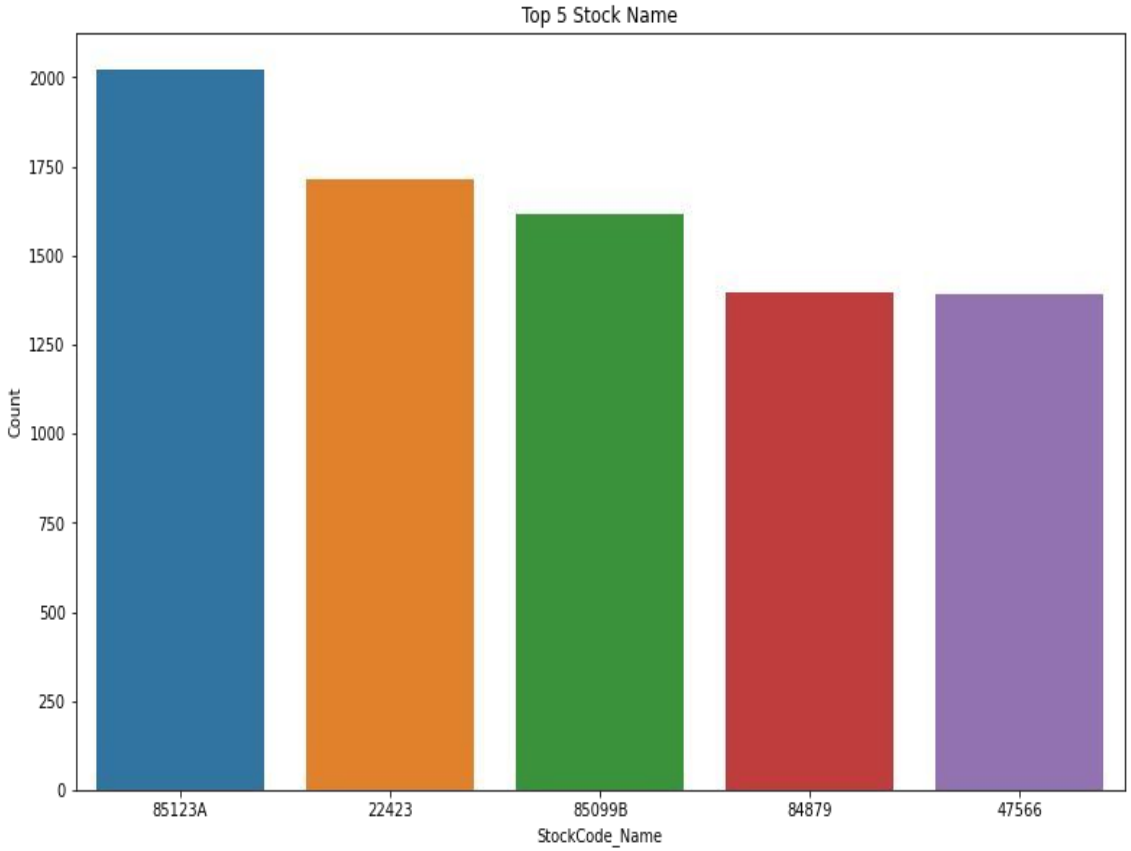
4.re dotcom quick fix.

5.BIRTHDAY BANNER TAPE



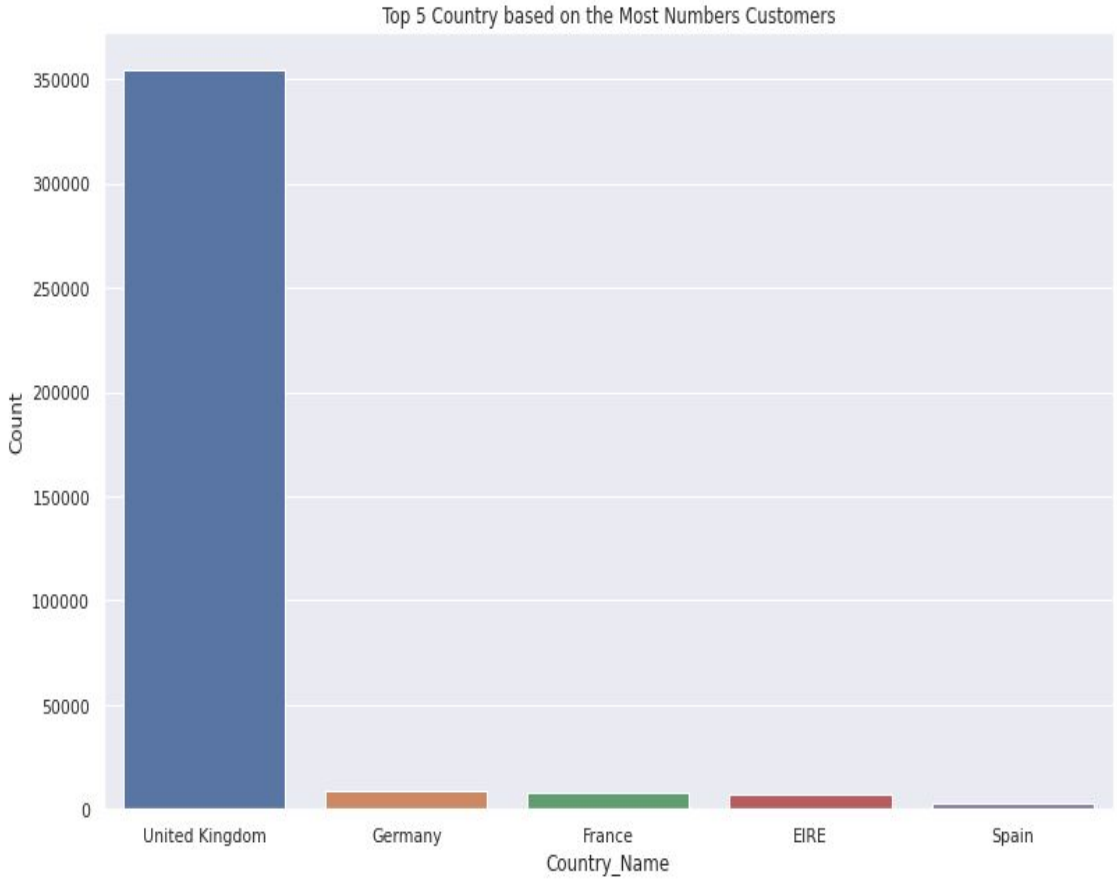
Analysis on top stock code

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



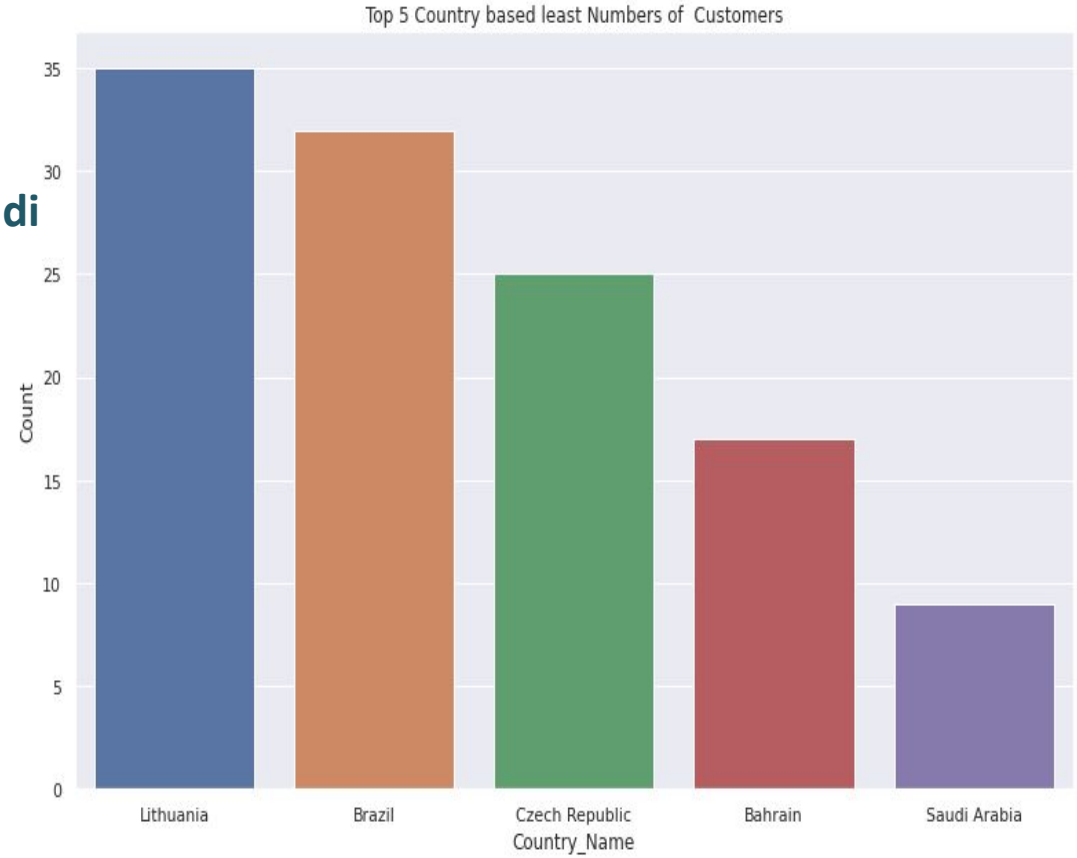
Analysis on top customers country wise

	Country_Name	Count
0	United Kingdom	354345
1	Germany	9042
2	France	8342
3	EIRE	7238
4	Spain	2485

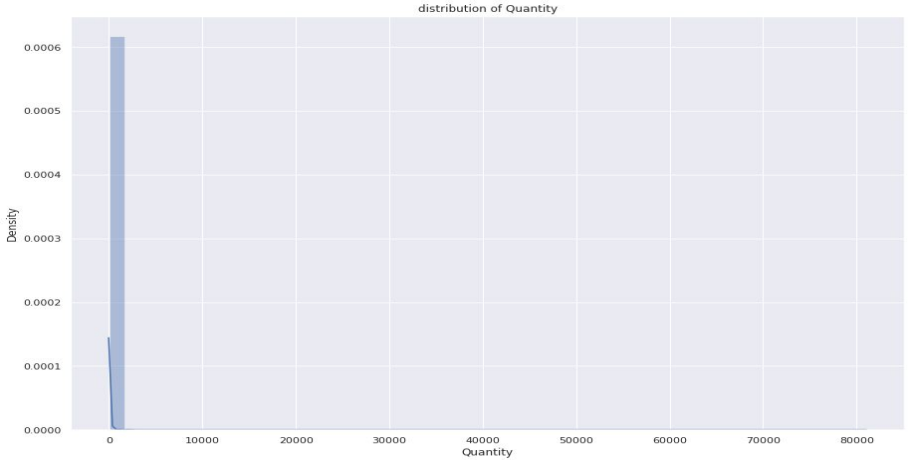


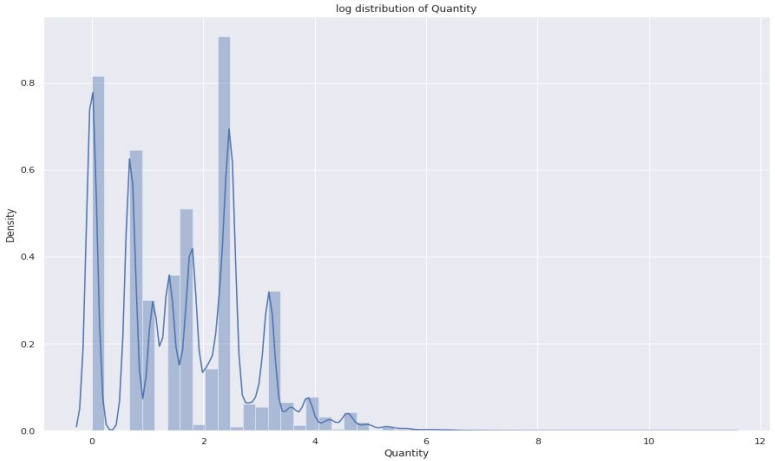
Analysis on least customers country wise

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



DISTRIBUTION



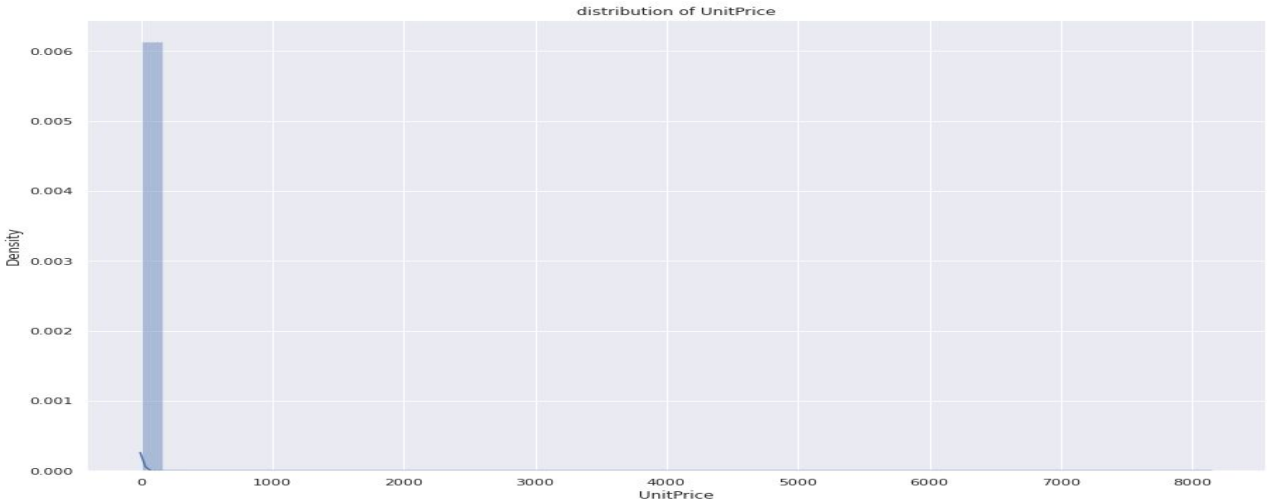


1.Positively skewed (or right-skewed) distribution is a type of distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer.hear mean>median>mode

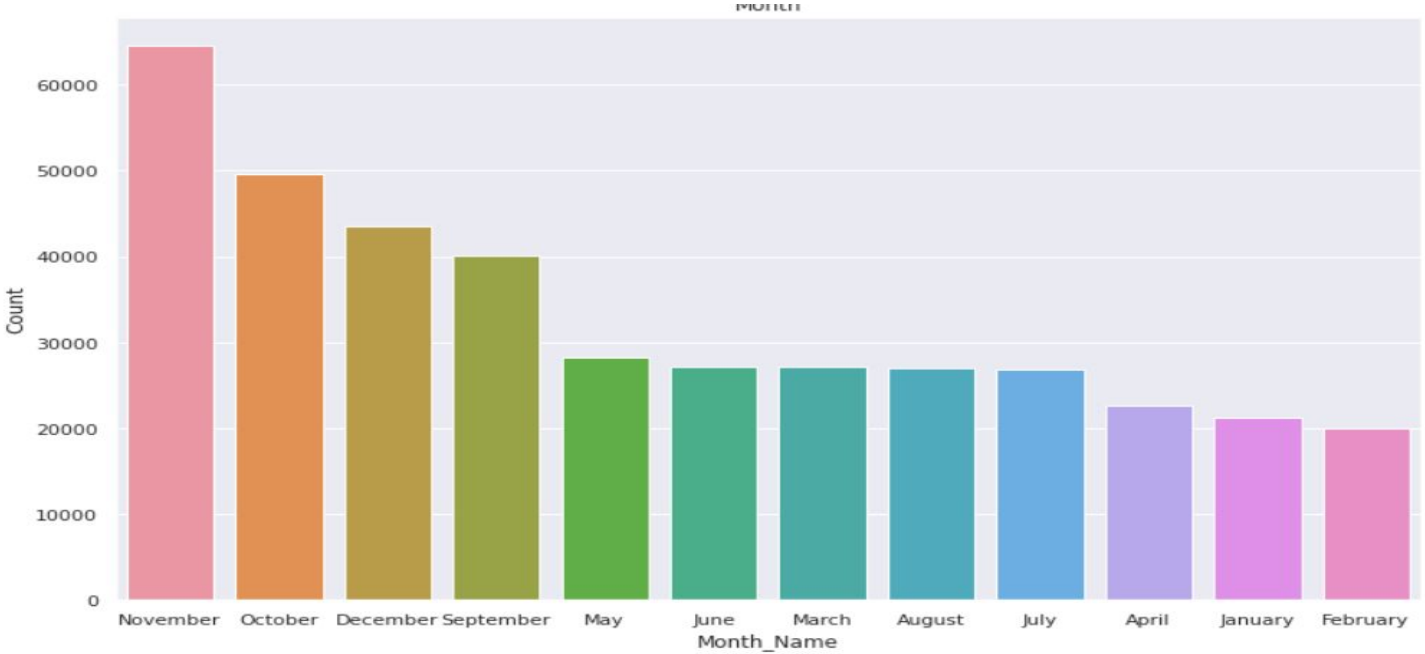
2. Negatively skewed (also known as left-skewed) distribution is a type of distribution in which more values are concentrated on the right side (tail) of the distribution graph while the left tail of the distribution graph is longer.hear mean<median<mode1

DISTRIBUTION

For symmetric graph mean=median=mode

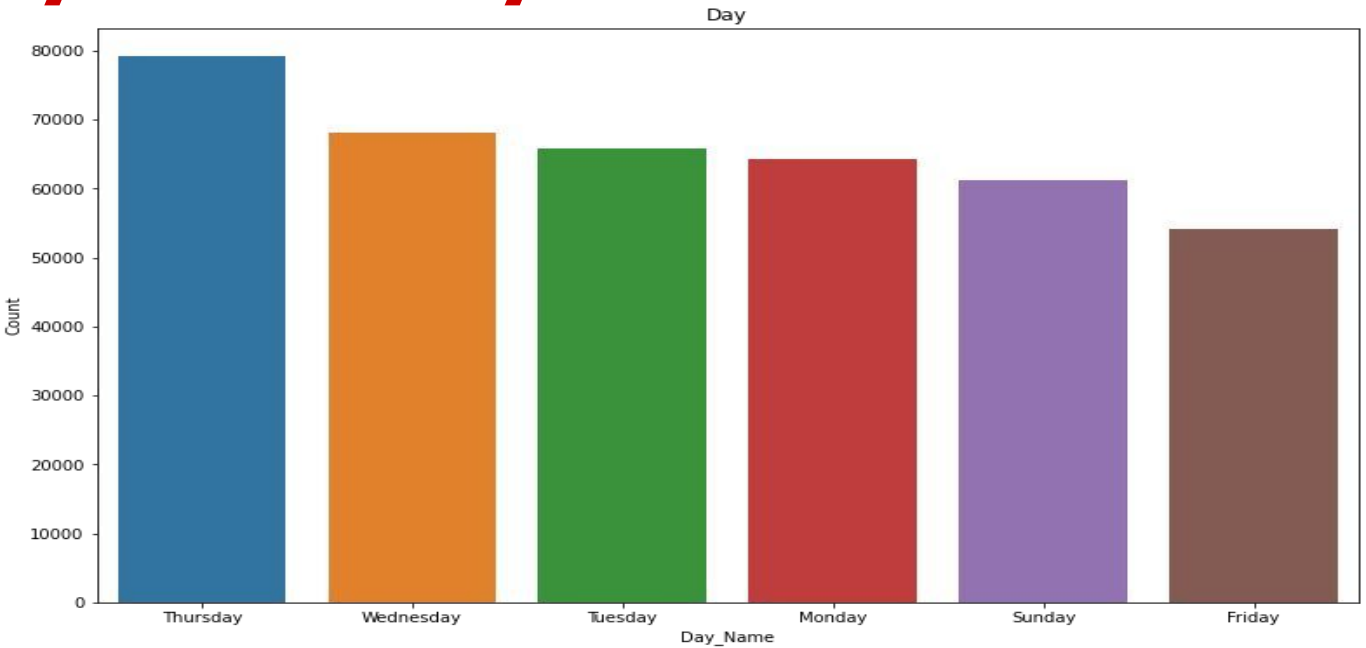


Analysis on Month wise



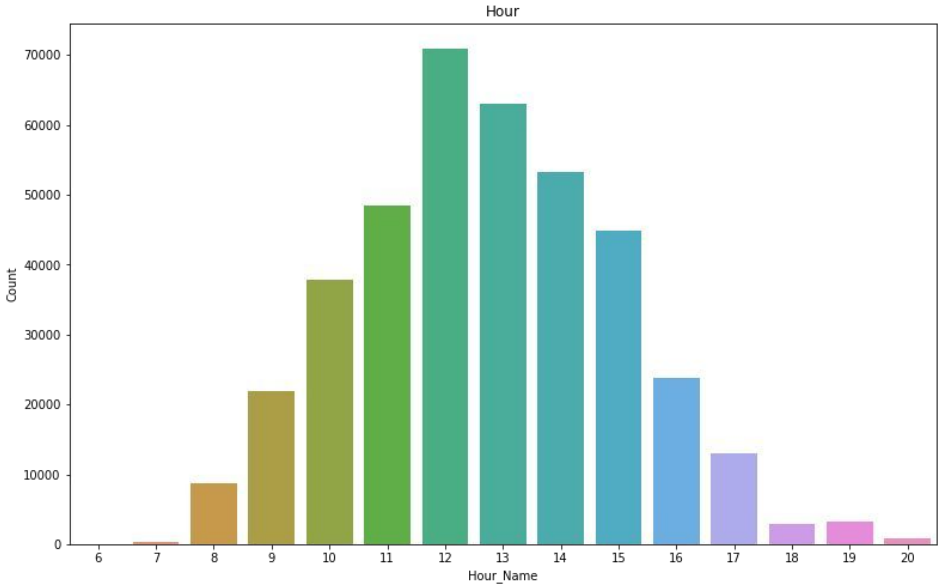
Most of the customers purchased gifts in september,october, november, December

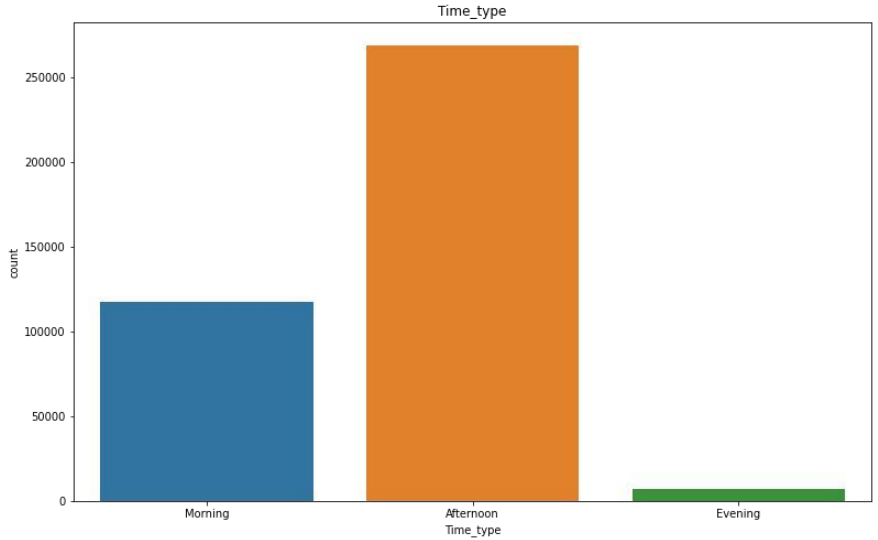
Analysis on Day wise



- **Most of the customers have purchased the items in Thursday, Wednesday and Tuesday**

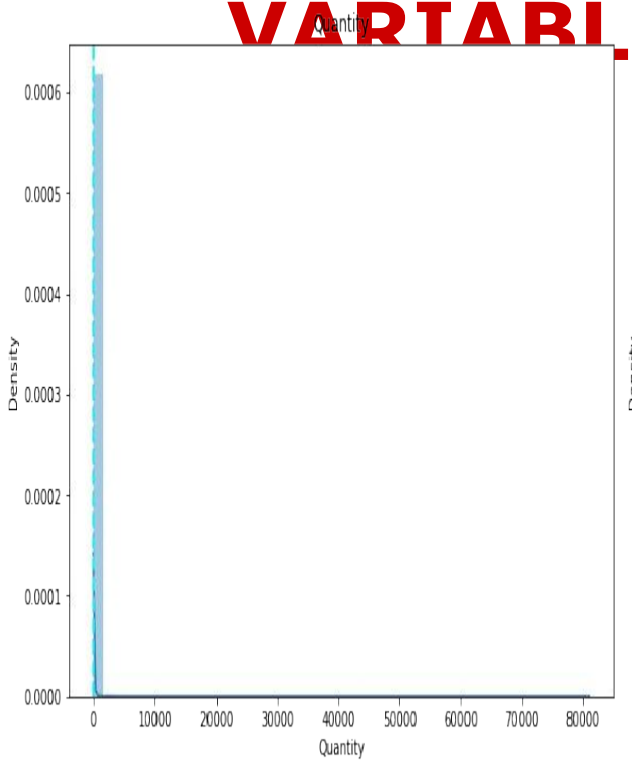
Analysis on Hours wise

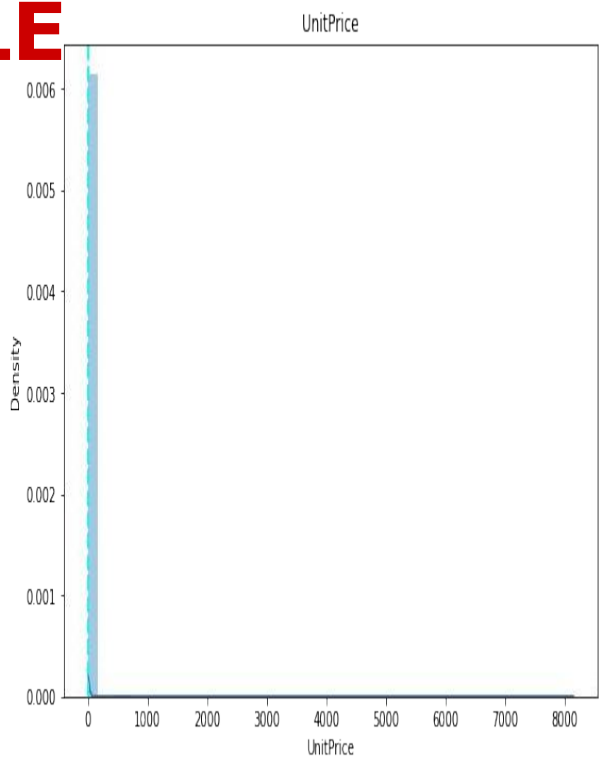


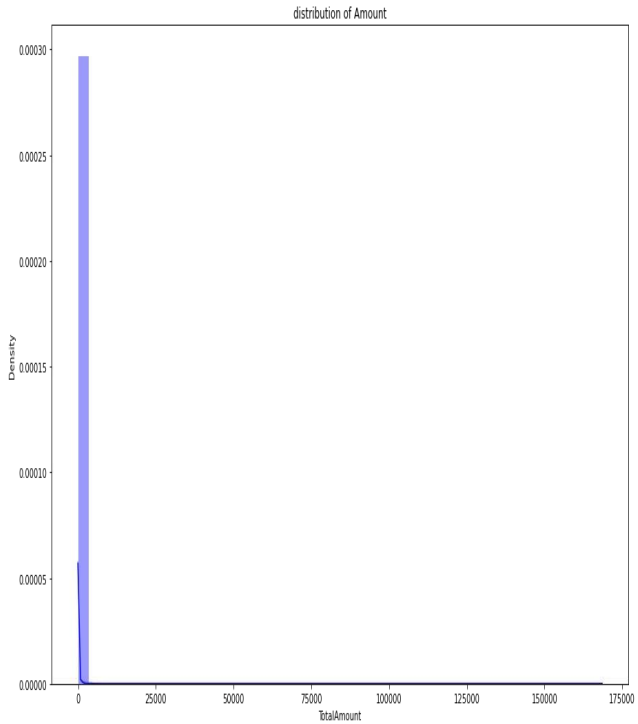


- **Working hours witnessing most of the customer purchased item in afternoon time and moderate in morning and least in evening**

ANALYSIS NUMERICAL VARIABLE



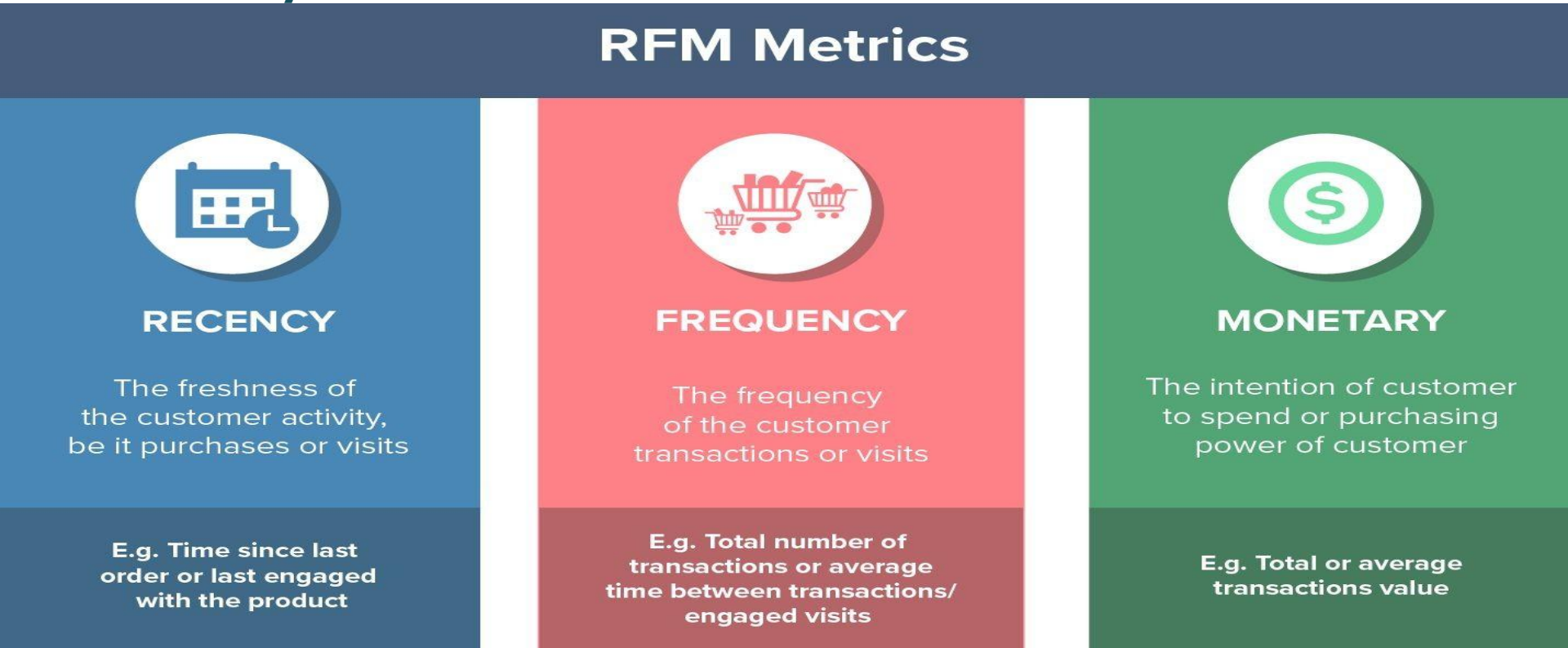




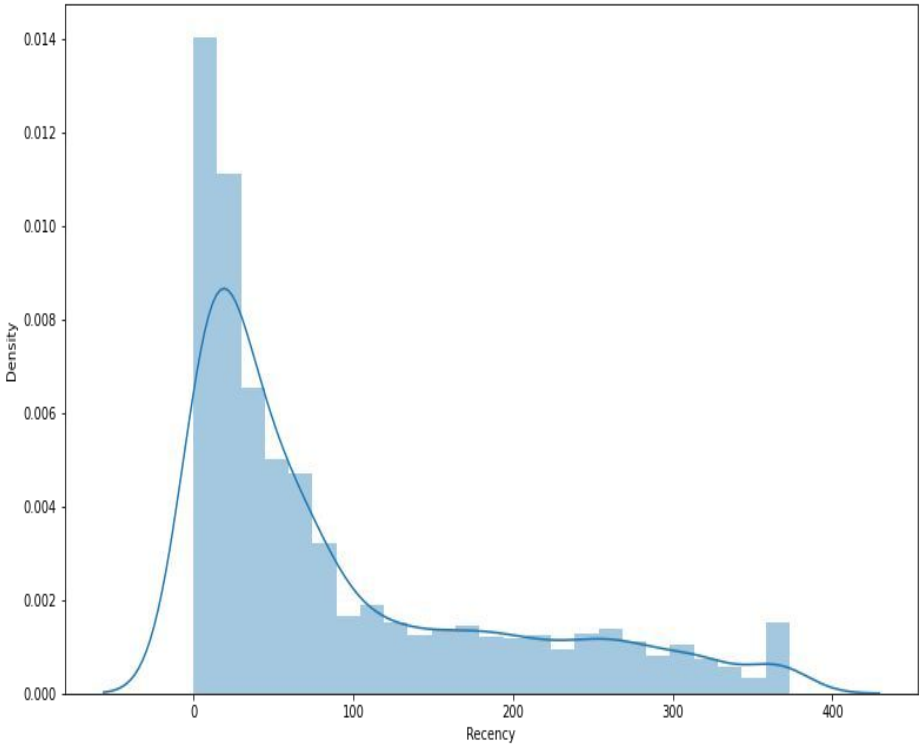
- **Highly positively skewed, need to do log transformation**

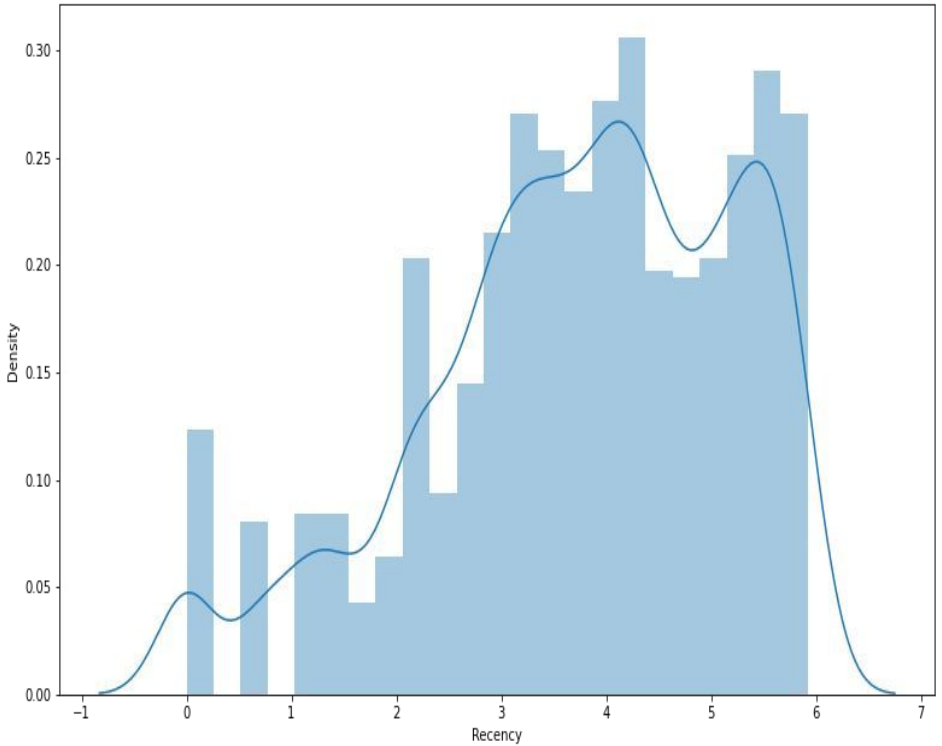
RFM MODEL

- **Created features such as recency, frequency and monetary**



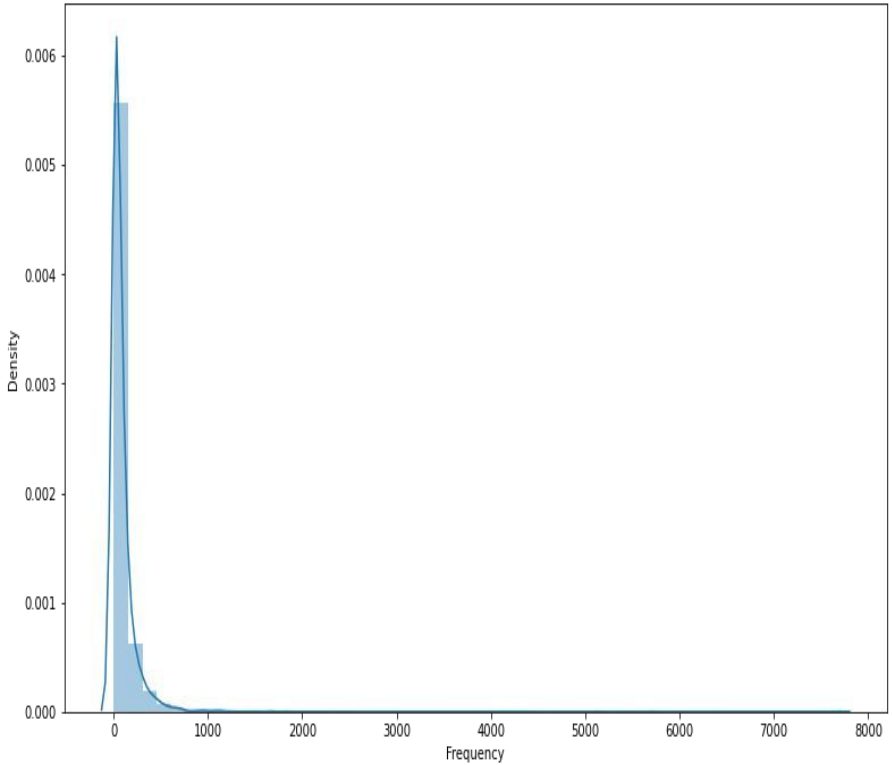
Recency

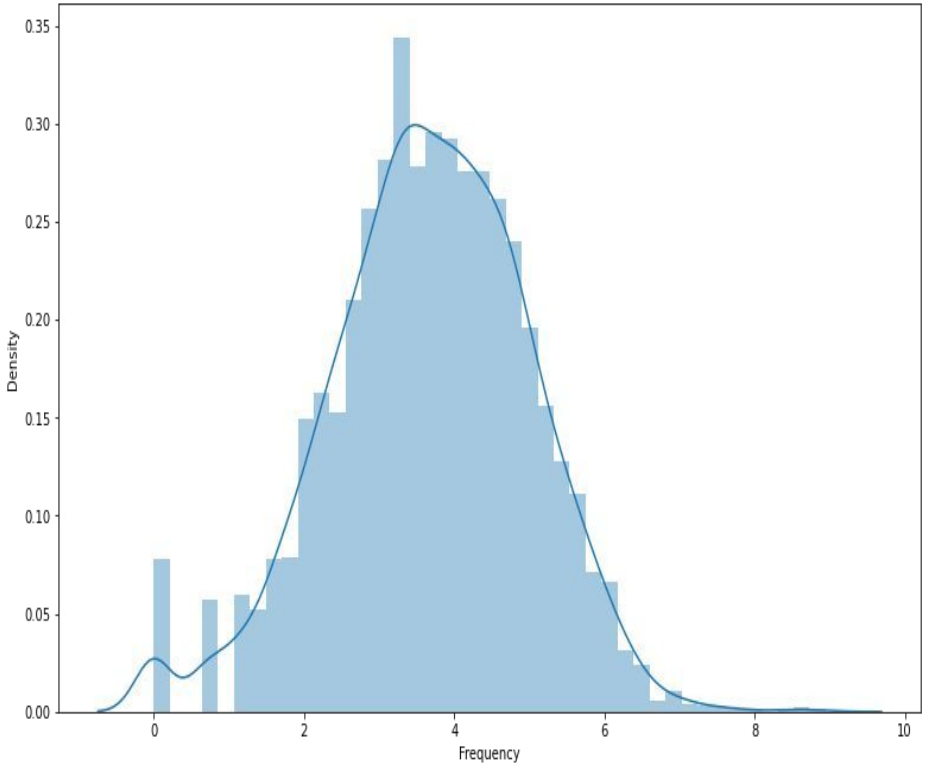




distribution before and after normal distribution with log transformation

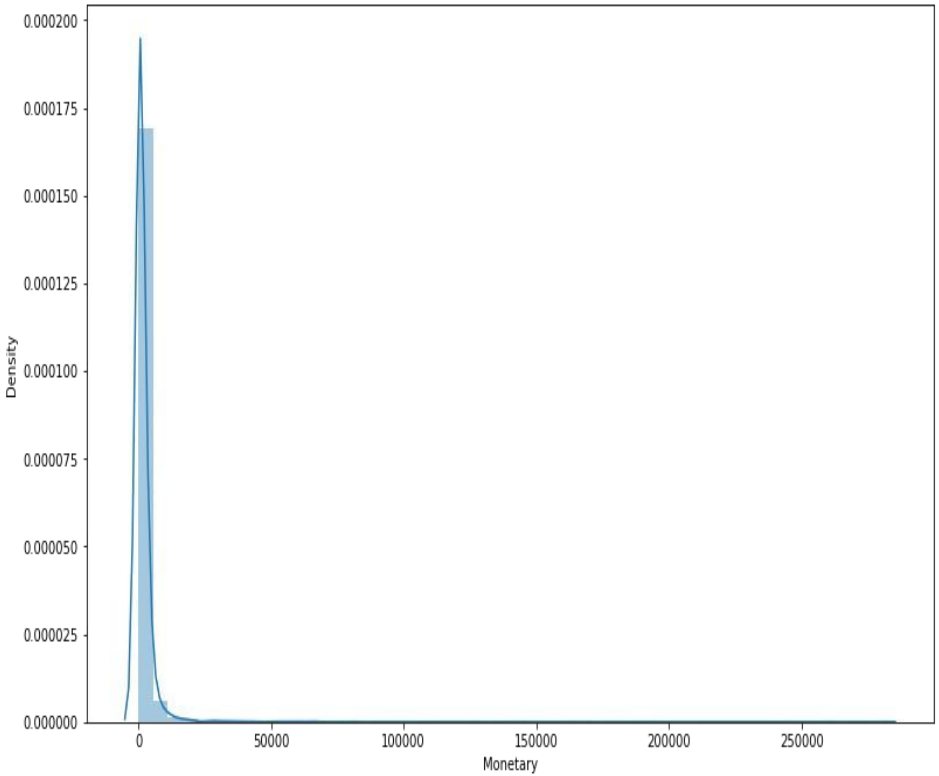
Frequency

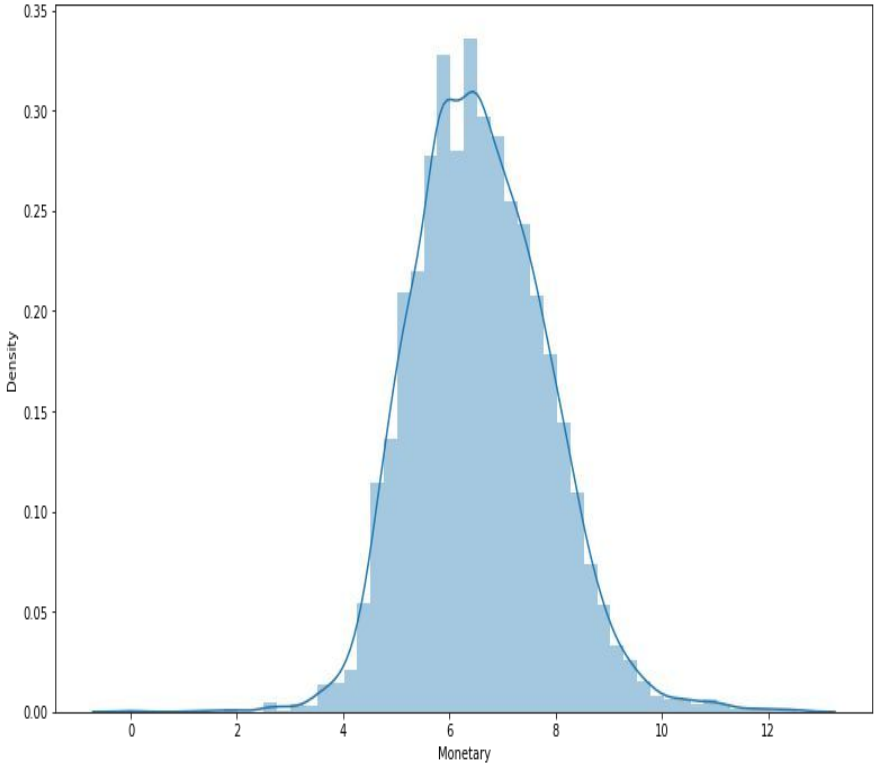




frequency distribution plot and normal distribution after log transformation

Monetary





Data distribution after data normalization for Monetary

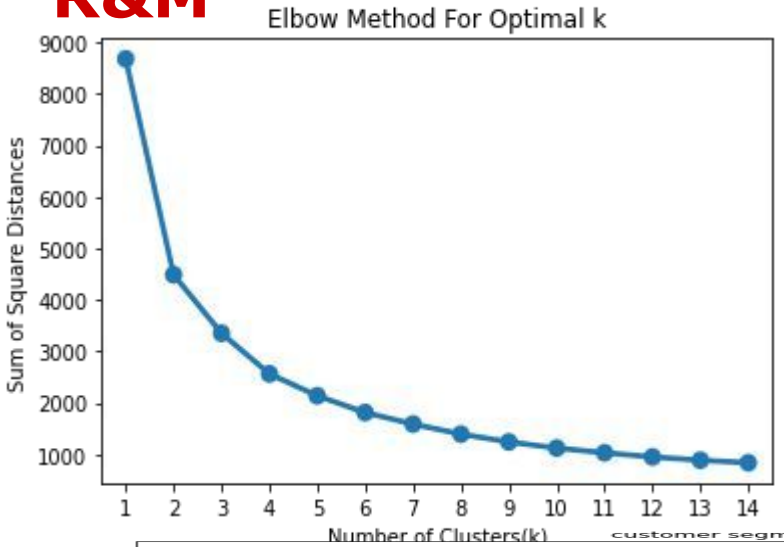
Calculation of Silhouette score

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:

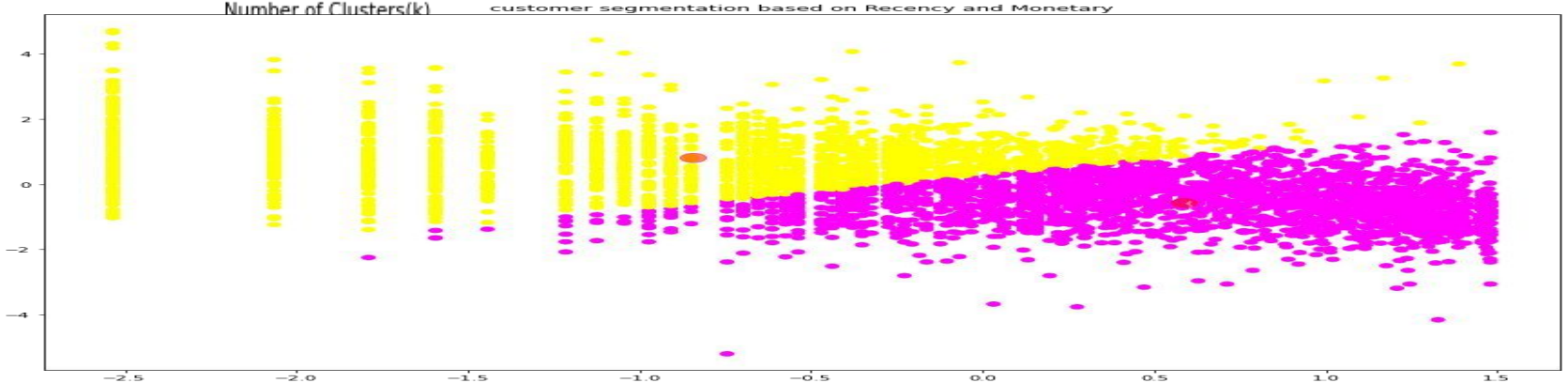
- Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.

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

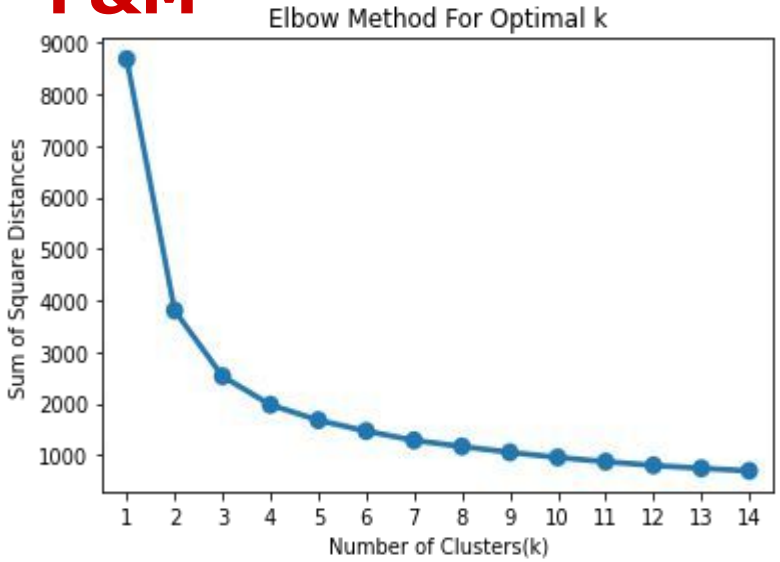
SILHOUETTE SCORE AND ELBOW METHOD ON R&M



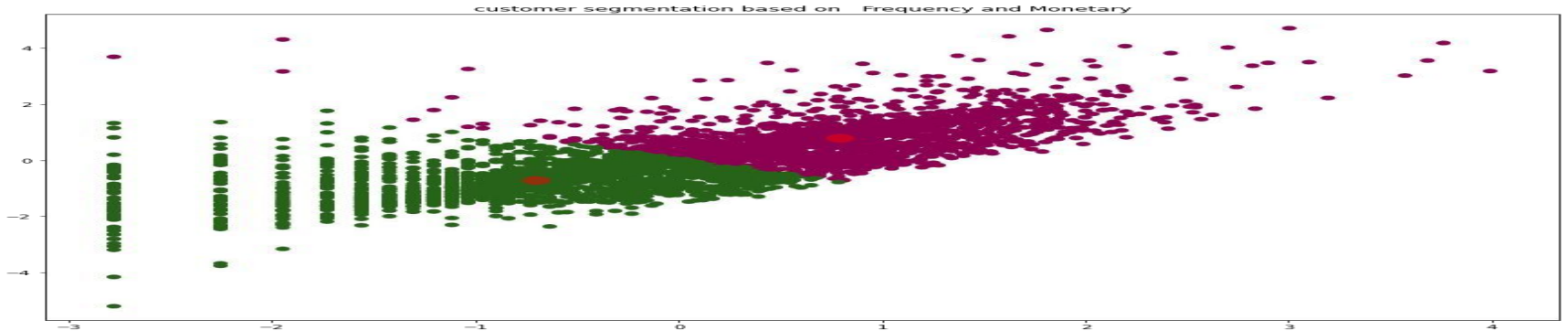

```
For n_clusters = 2, silhouette score is 0.421461308316105
For n_clusters = 3, silhouette score is 0.3434171144833716
For n_clusters = 4, silhouette score is 0.36470899651128513
For n_clusters = 5, silhouette score is 0.3352920215016303
For n_clusters = 6, silhouette score is 0.344125915245459
For n_clusters = 7, silhouette score is 0.34782996809498024
For n_clusters = 8, silhouette score is 0.3387645516221769
For n_clusters = 9, silhouette score is 0.3451188527859384
For n_clusters = 10, silhouette score is 0.3485389294103986
For n_clusters = 11, silhouette score is 0.3380232921546669
For n_clusters = 12, silhouette score is 0.3437933457501528
For n_clusters = 13, silhouette score is 0.34078814349213
For n_clusters = 14, silhouette score is 0.3375910483707973
For n_clusters = 15, silhouette score is 0.3358256706308165
```



SILHOUETTE SCORE AND ELBOW METHOD ON F&M



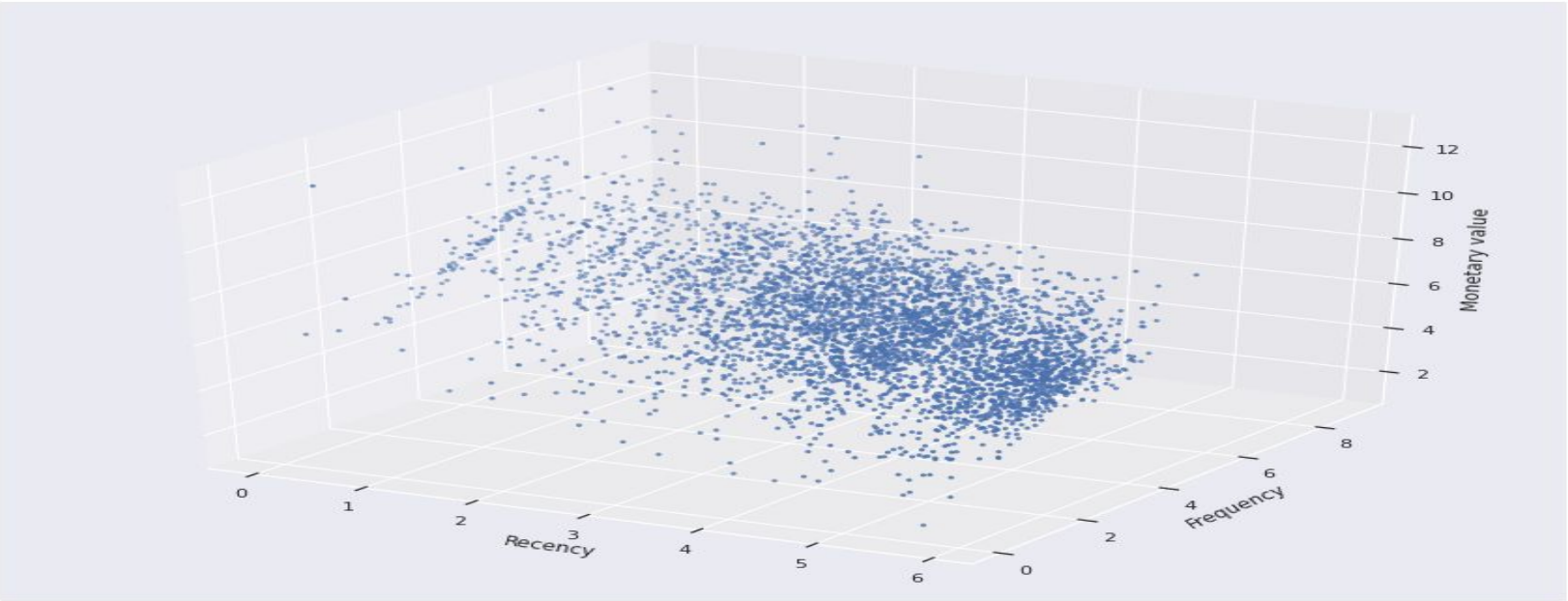

```
For n_clusters = 2, silhouette score is 0.478535709506603
For n_clusters = 3, silhouette score is 0.40764120562174455
For n_clusters = 4, silhouette score is 0.37205487483957167
For n_clusters = 5, silhouette score is 0.34512350681962106
For n_clusters = 6, silhouette score is 0.35915338840993544
For n_clusters = 7, silhouette score is 0.3405727767262927
For n_clusters = 8, silhouette score is 0.350166247976559
For n_clusters = 9, silhouette score is 0.34163531768217203
For n_clusters = 10, silhouette score is 0.3586075218108946
For n_clusters = 11, silhouette score is 0.3427425807832202
For n_clusters = 12, silhouette score is 0.35483075246193607
For n_clusters = 13, silhouette score is 0.36497400916106304
For n_clusters = 14, silhouette score is 0.3463653985250052
For n_clusters = 15, silhouette score is 0.3541193894768307
```



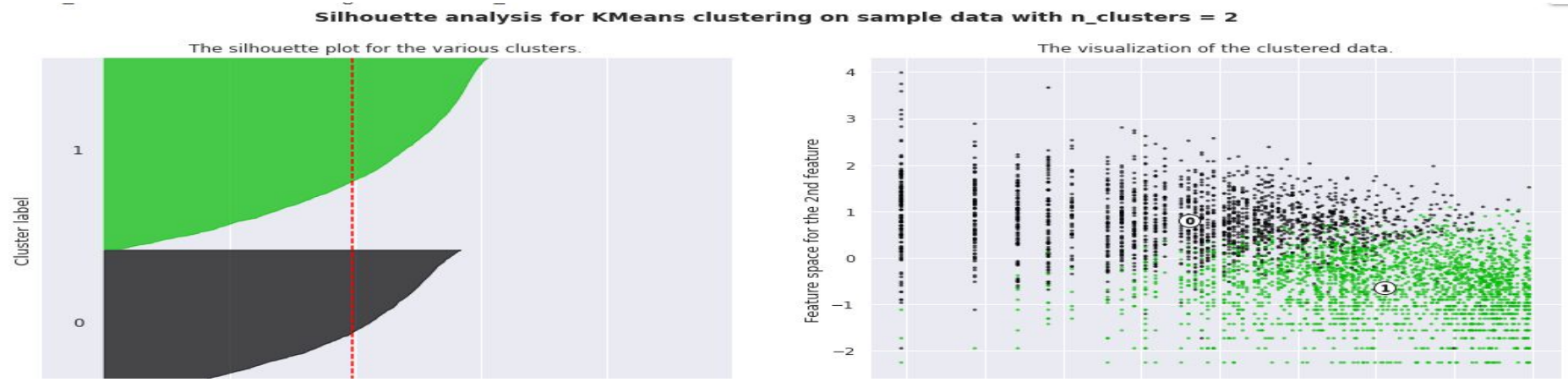
SILHOUETTE ANALYSIS ON R, F AND M

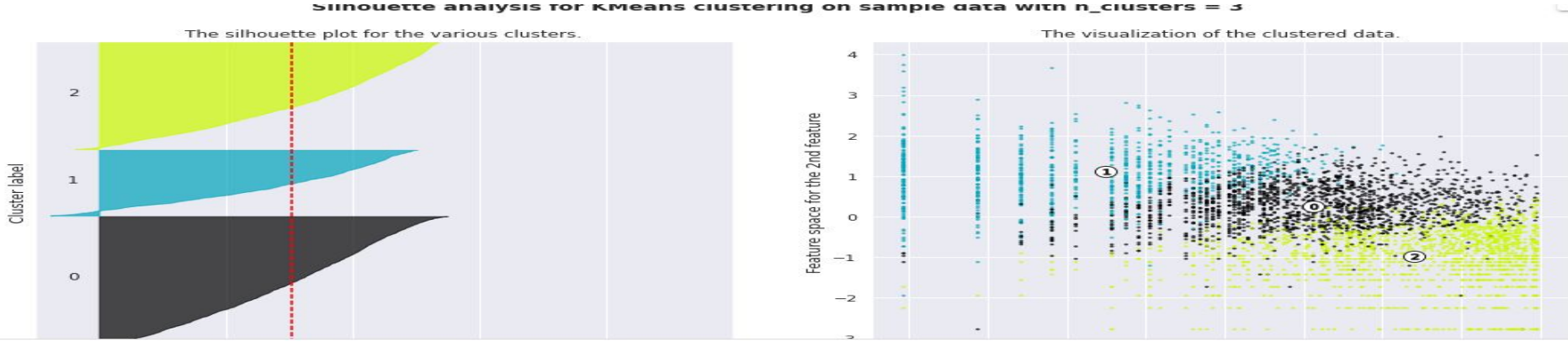
```
For n_clusters = 2 The average silhouette_score is : 0.3956478042246982
For n_clusters = 3 The average silhouette_score is : 0.3049826724447913
For n_clusters = 4 The average silhouette_score is : 0.30279724233096916
For n_clusters = 5 The average silhouette_score is : 0.2785519277480847
For n_clusters = 6 The average silhouette_score is : 0.2789560652501828
For n_clusters = 7 The average silhouette_score is : 0.2613208163968789
For n_clusters = 8 The average silhouette_score is : 0.2640918249728342
For n_clusters = 9 The average silhouette_score is : 0.2585642595481418
For n_clusters = 10 The average silhouette_score is : 0.2644733794304285
For n_clusters = 11 The average silhouette_score is : 0.2592423011915937
For n_clusters = 12 The average silhouette_score is : 0.26503813251658404
For n_clusters = 13 The average silhouette_score is : 0.2621555416679574
For n_clusters = 14 The average silhouette_score is : 0.26140947155997746
```

3D visualization of Recency Frequency and Monetary

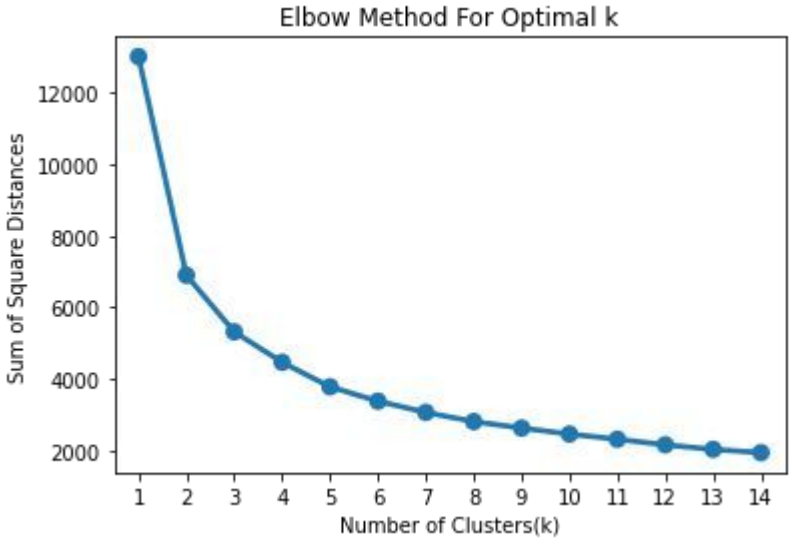


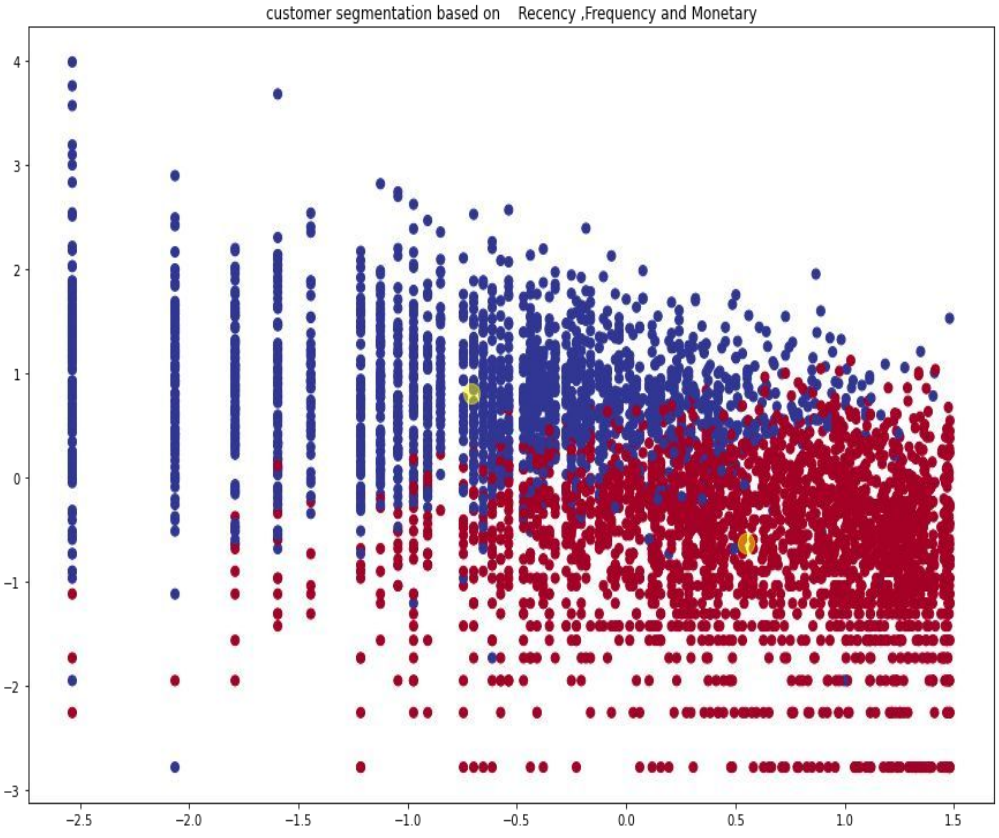
SILHOUETTE ANALYSIS ON R, F , M





ELBOW METHOD AND CLUSTER CHART ON RFM

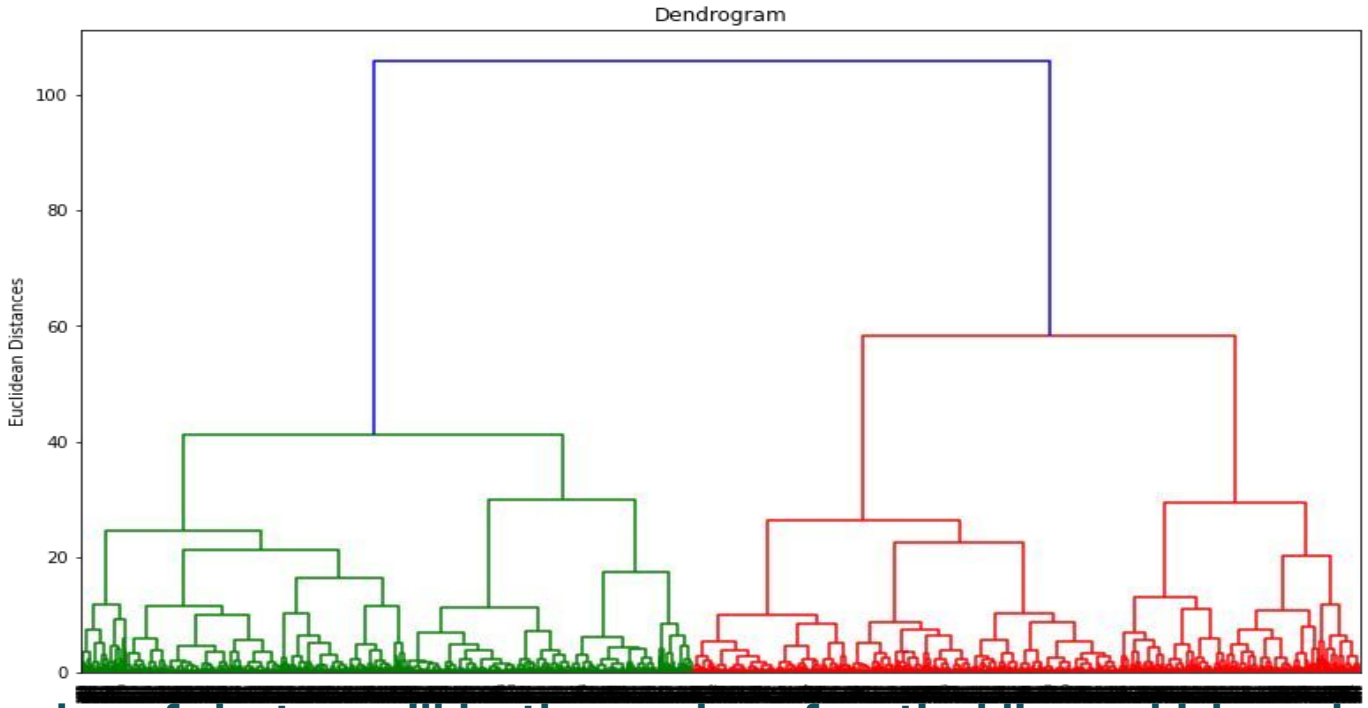




RFM ANALYSIS

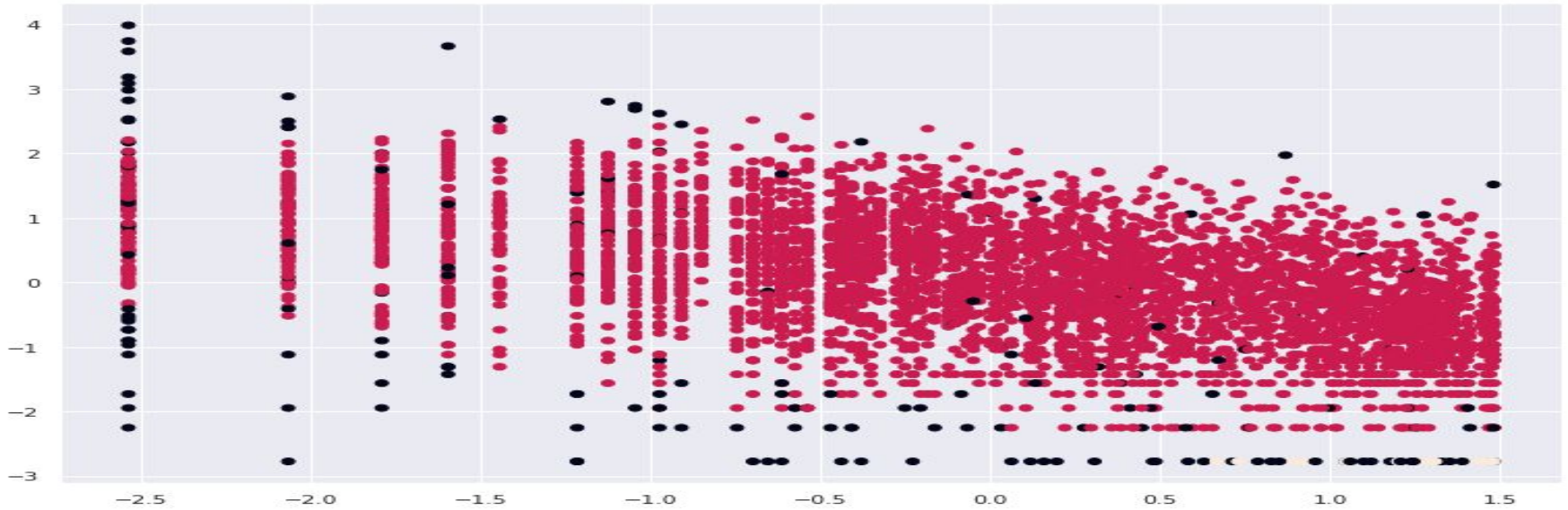
CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	Recency_log	Frequency_log	Monetary_log	Cluster
12346.0	325	1	77183.60	4	4	1	441	9	5.783825	0.000000	11.253942	0
12347.0	2	182	4310.00	1	1	1	111	3	0.693147	5.204007	8.368693	1
12348.0	75	31	1797.24	3	3	1	331	7	4.317488	3.433987	7.494007	0
12349.0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

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

DBSCAN TO RECENCY ,FREQUENCY AND MONETARY



- **Density-based spatial clustering of applications with noise** (DBSCAN)
- **we see that ,Customers are well separate when we cluster them by Recency ,Frequency and Monetary and optimal number of cluster is equal to 3**

CHALLENGES

- Large Dataset to handle.
- Needs to plot lot of Graphs to analyse.
- Lot of NaN values.
- Continuous Runtime and RAM Crash due to large dataset.
- Right number of ‘K’ for clusters

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

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.
- Next we formulated some quantitative factors such as recency, frequency and monetary known as rfm model for each of the customers. We implemented K-Means 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.
- We saw higher values of frequency, monetary and low values of recency is deciding one class and low values of frequency, monetary and high values of recency is deciding other class.

Q&A

THANK YOU