Capstone Project : 4

Online Retail Customer Segmentation

By

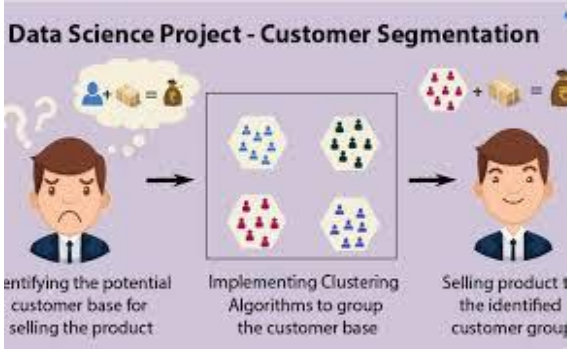
Mayank Mishra

Points for Discussion

- Understanding Problem Statement
- Data Set Information
- Feature Summary
- Data Wrangling
- Approach Overview
- Data Preprocessing
- Exploratory Data Analysis
- Implementing Algorithm
- Challenges
- Conclusion
- Q&A

Understanding Problem Statement

- Topic – “Online Retail Customer Segmentation”
- Problem Statement -- Identify major customer segments on a transnational dataset which contains all the data of transaction of registered non store online retail.

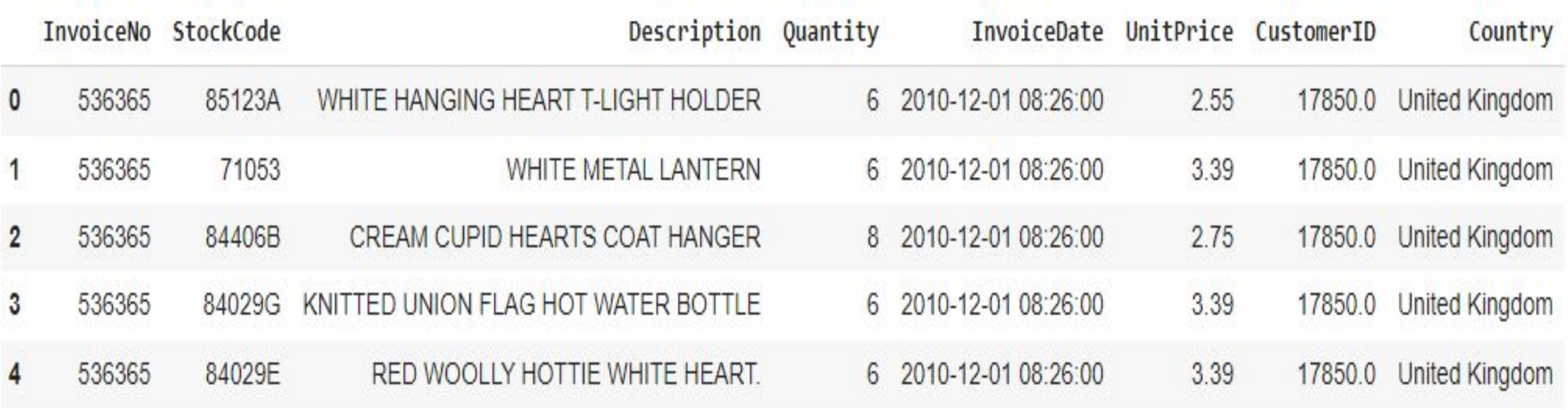

Data Science Project - Customer Segmentation
entifying the potential customer base for selling the product
Implementing Clustering Algorithms to group the customer base
Selling product t the identified customer grou

- Customer segmentation is the practice of dividing a company’s customers into groups that reflect similarity among customers in each group. It allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographics, behavioral and other indicators.

- The goal of this project is to decide how to relate to customers in each segment in order maximize the values of each customers to the business.

Data Set Information

- This dataset contains 541909 observations and 8 features.
- It is a transnational dataset with transactions occurring between 1st December 2010 to 9th December 2011 for a UK based online retailer .
- Many customers are wholesalers.



	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

Feature Summary

- InvoiceNo: It is a nominal 6 digit integral number uniquely assigned to each transaction. If the code starts with letter ‘c’, it indicates cancellation.

- StockCode: Product code. It’s a 5 digit integral number uniquely assigned to each distinct product.

- Description: Product item name.

- Quantity: It’s a numeric value of each item per transaction.

- InvoiceDate: Invoice date and time, it’s a numeric value of the day and time when each transaction was done.

- UnitPrice: It is a numeric value which shows the product price per unit in sterling.

- CustomerId: It’s a 5 digit integral number uniquely assigned to each customer.

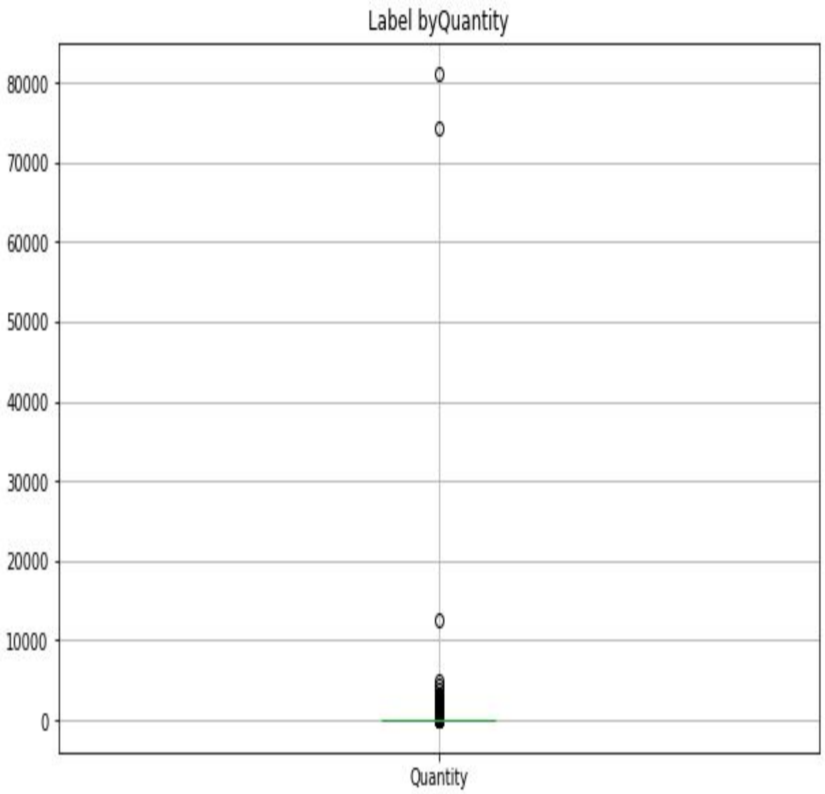
- Country: The name of the country where each customer resides.

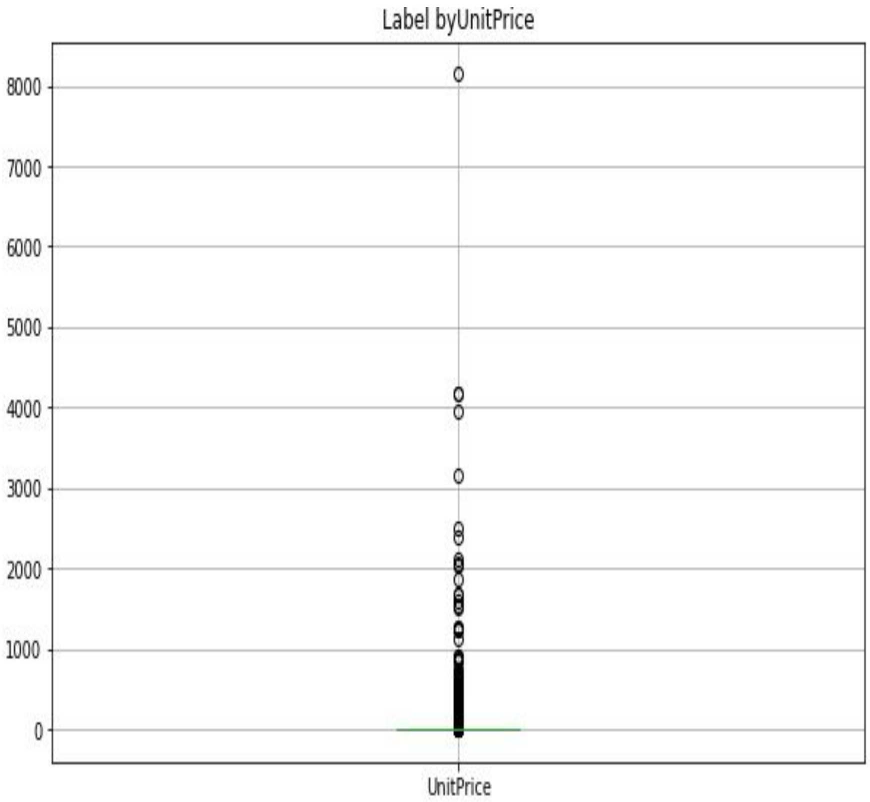
Data Wrangling

- Dataset is from UK,
- In dataset there are 541909 rows and columns.
- Categorical Features: ‘InvoiceNo’, ‘StockCode’, ‘Description’ and ‘Country’.
- There are 1454 null values in “Description’ and 135080 in ‘ CustomerID’.
- There are 5225 duplicates values present in our data.
- One Datetime[ns] feature: InvoiceDate.
- Outliers present only in ‘Quantity’ and ‘UnitPrice’ column.
- Dropped cancelled orders.
- New features from Datetime column added such as months, days and hours.
- Total amount is added.
- Data types is converted

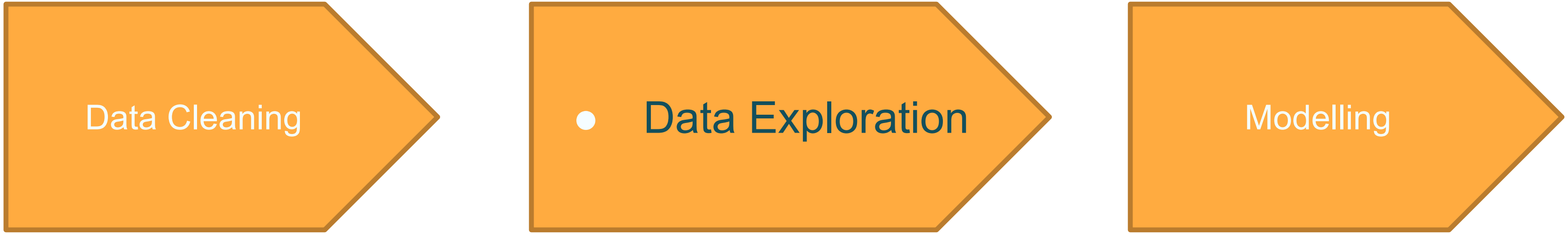
Data Wrangling(Cont.)

- Outliers present in our quantity and UnitPrice column.





Approach Overview



Data Cleaning

Data Exploration

Modelling

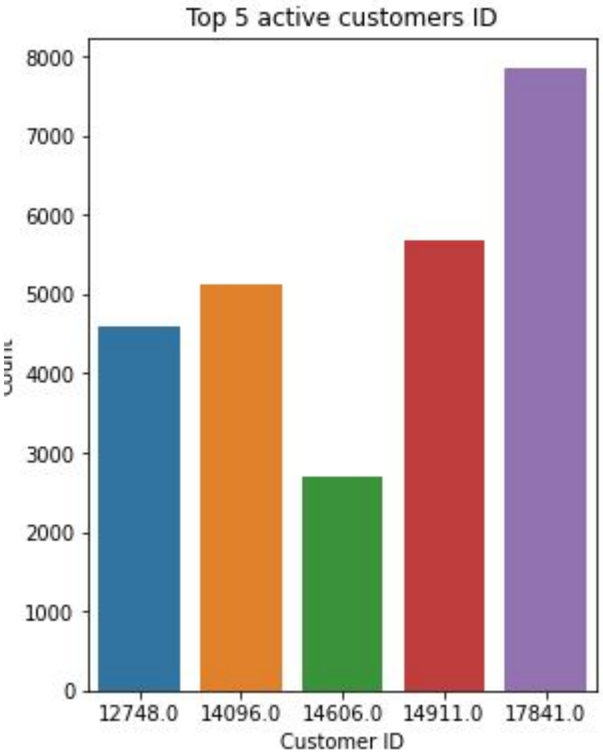
- Find information on documented columns values.
- Clean data for further analysis.

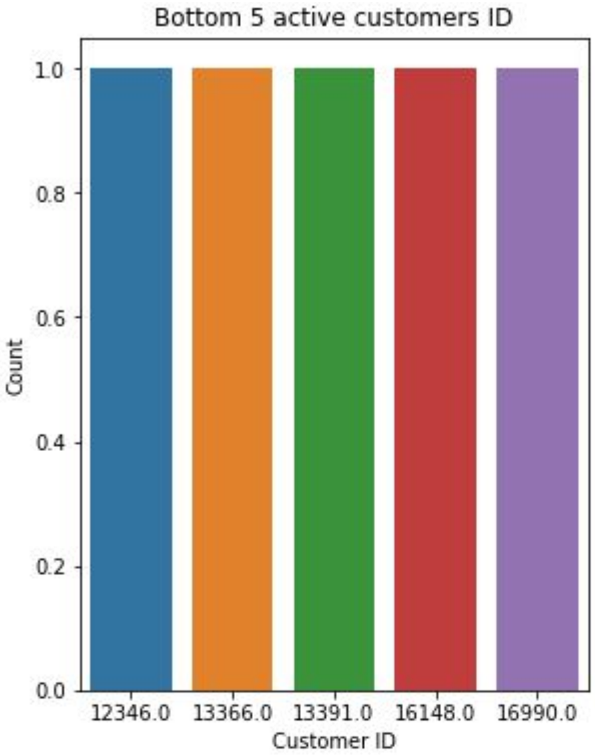
- Analyze the data with EDA

- K Means Clustering
- DBSCAN
- Hierarchical Clustering

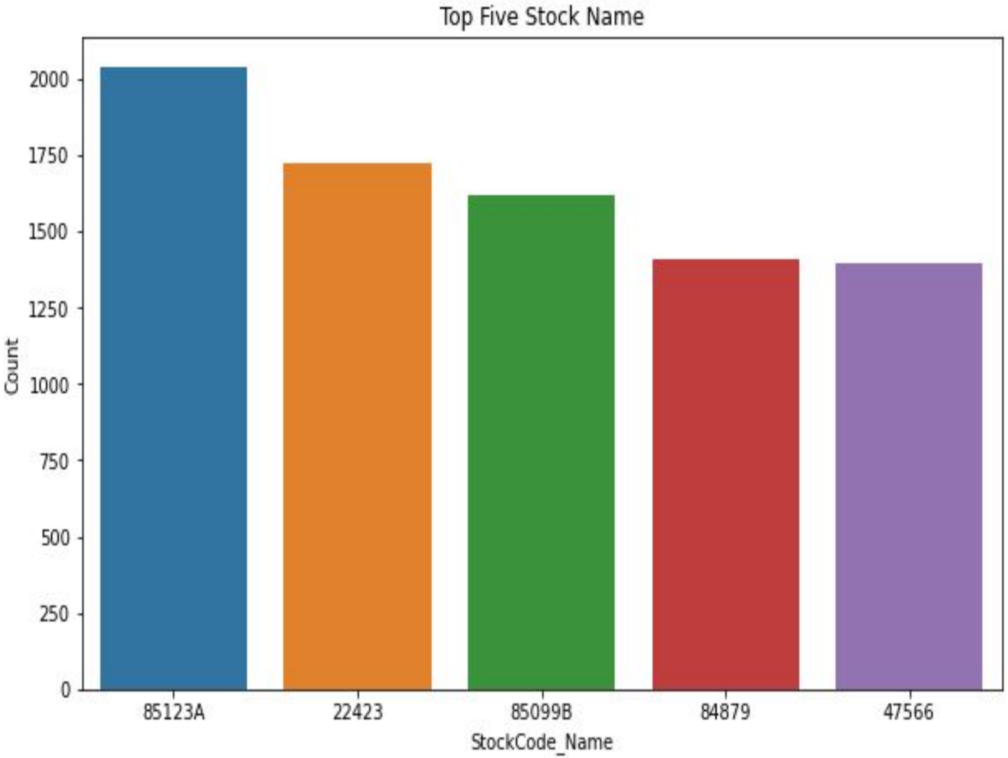
CustomerID Analysis

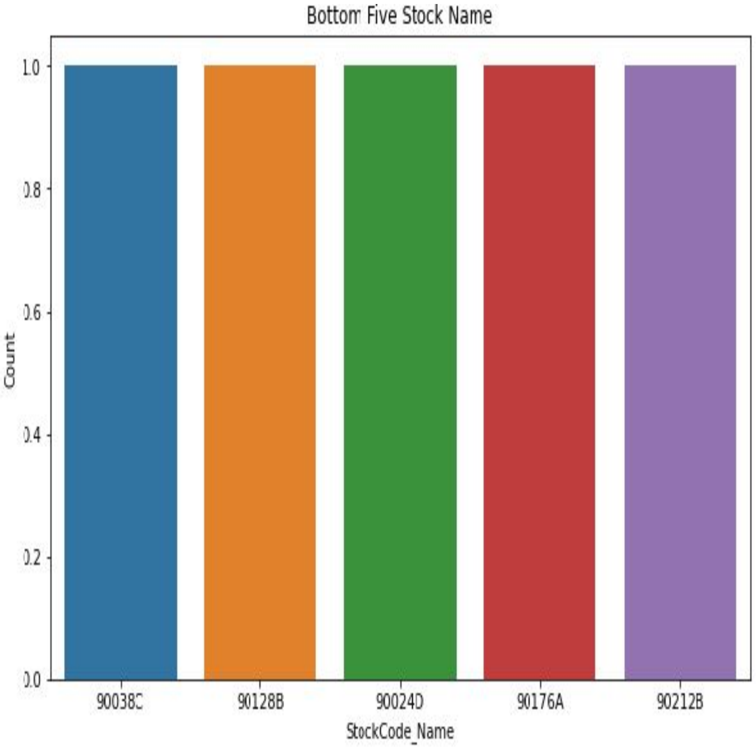
- 4339 unique customer ID.
- ID number 17841 was the most active customer.


Top 5 active customers ID

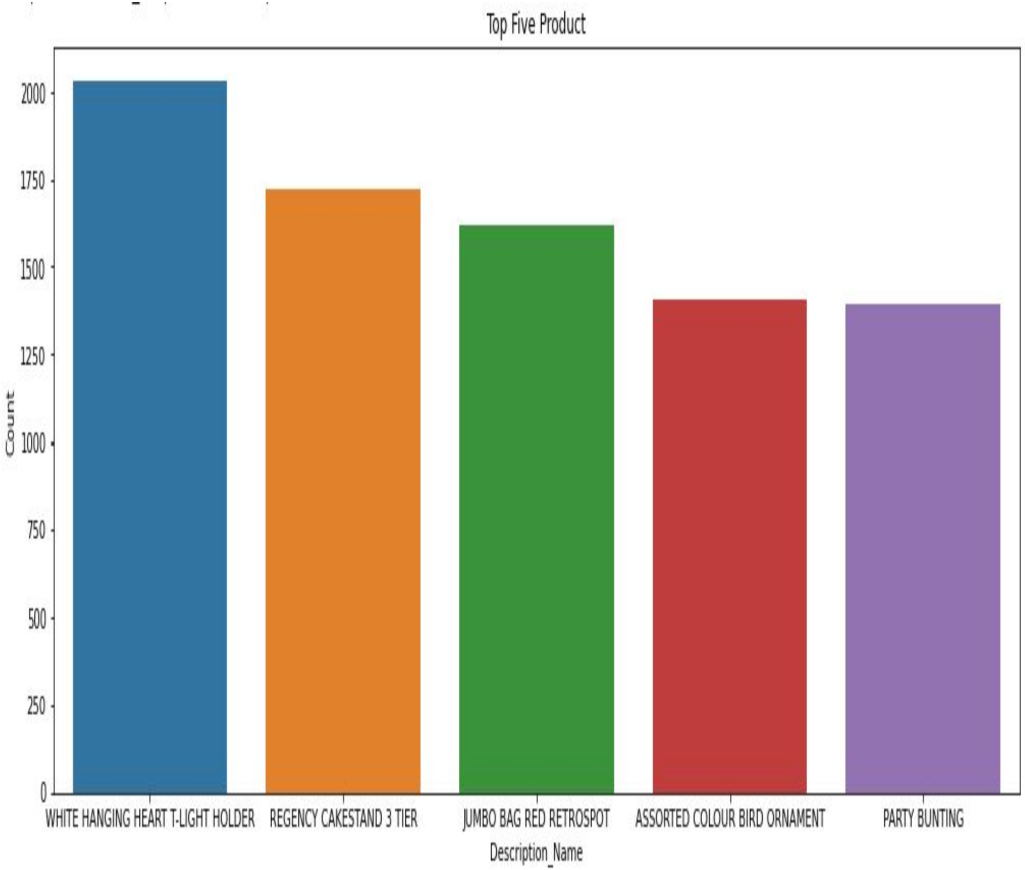

Bottom 5 active customers ID

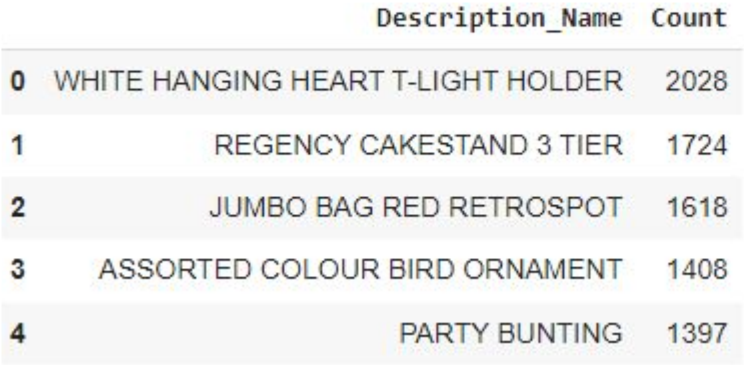
Insights From StockCode





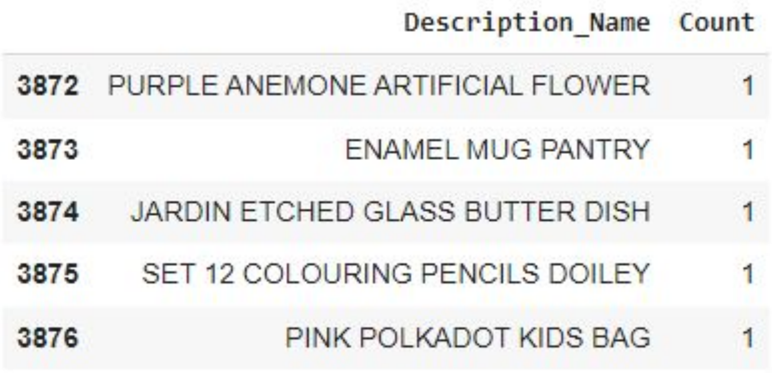
Insights from product Description



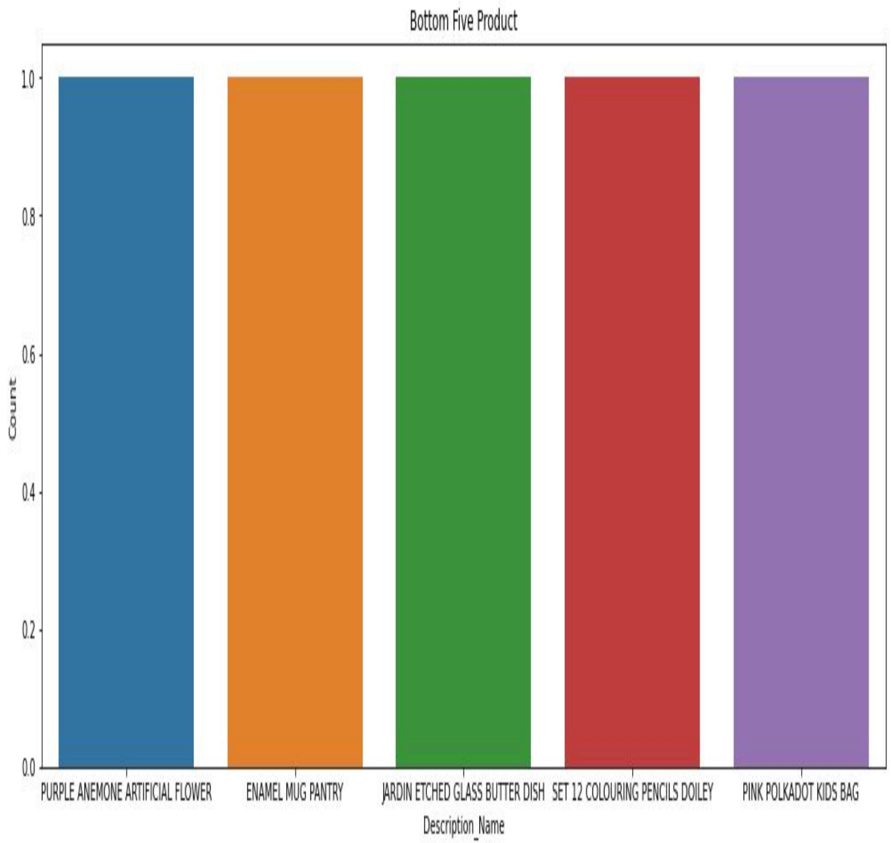


	Description_Name	Count
0	WHITE HANGING HEART T-LIGHT HOLDER	2028
1	REGENCY CAKESTAND 3 TIER	1724
2	JUMBO BAG RED RETROSPOT	1618
3	ASSORTED COLOUR BIRD ORNAMENT	1408
4	PARTY BUNTING	1397

Insights from product Description(Cont.)

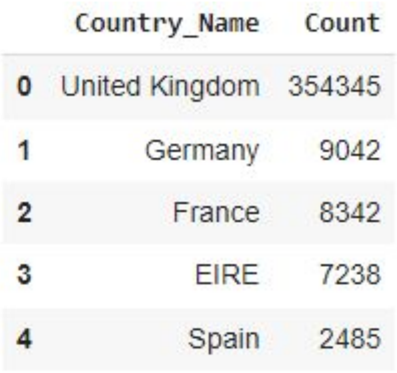


	Description_Name	Count
3872	PURPLE ANEMONE ARTIFICIAL FLOWER	1
3873	ENAMEL MUG PANTRY	1
3874	JARDIN ETCHED GLASS BUTTER DISH	1
3875	SET 12 COLOURING PENCILS DOILEY	1
3876	PINK POLKADOT KIDS BAG	1

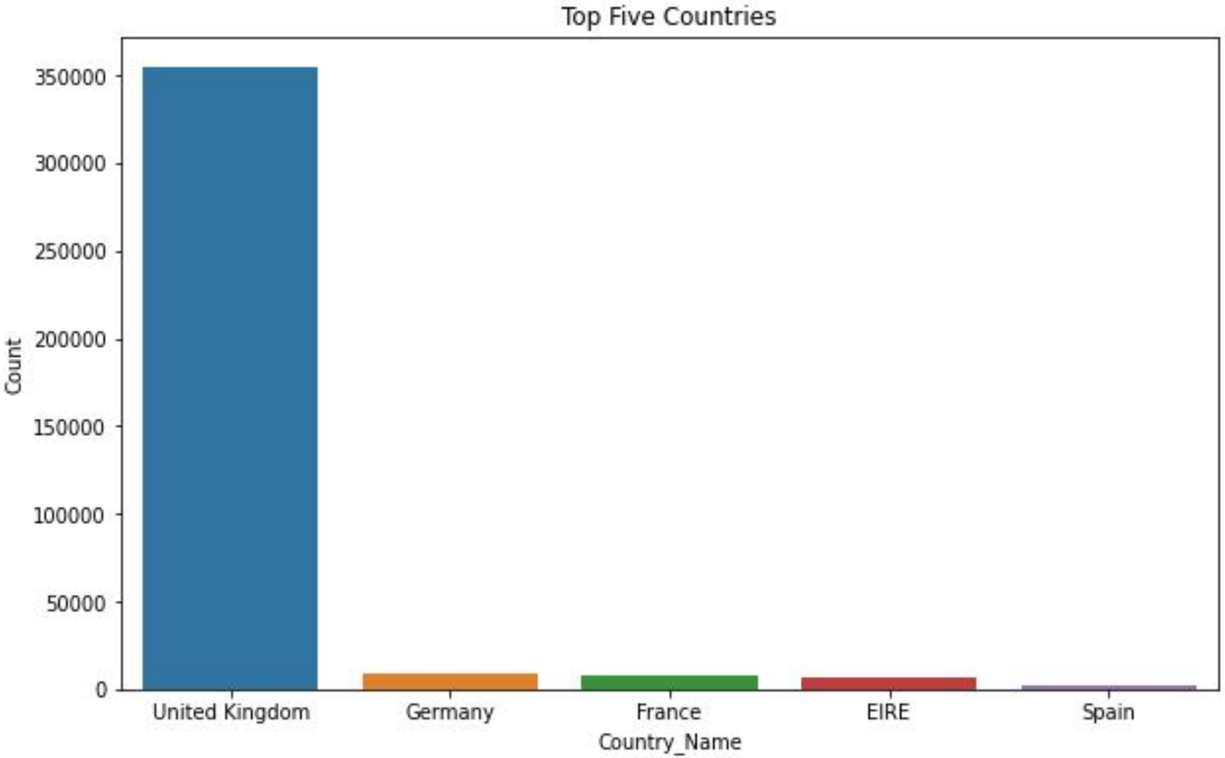


Analysis on Country

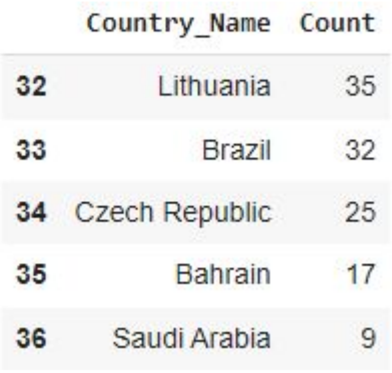
- UK has the majority of customers.



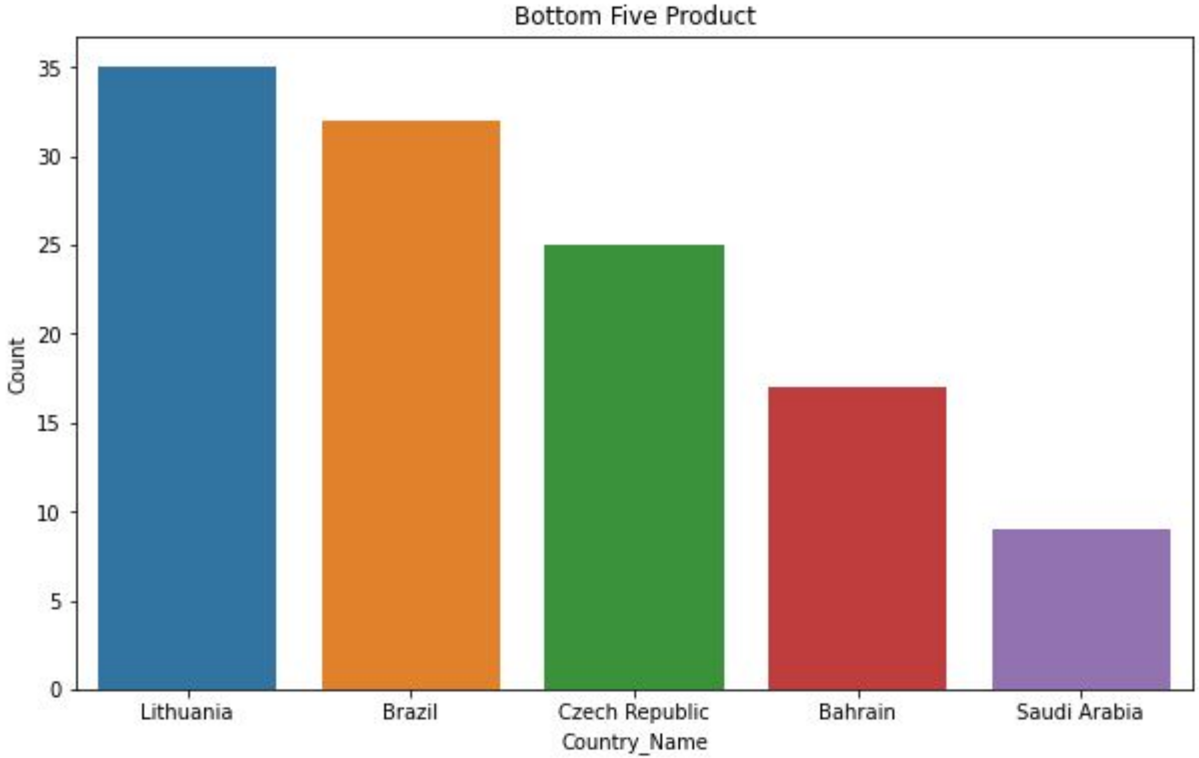
	Country_Name	Count
0	United Kingdom	354345
1	Germany	9042
2	France	8342
3	EIRE	7238
4	Spain	2485



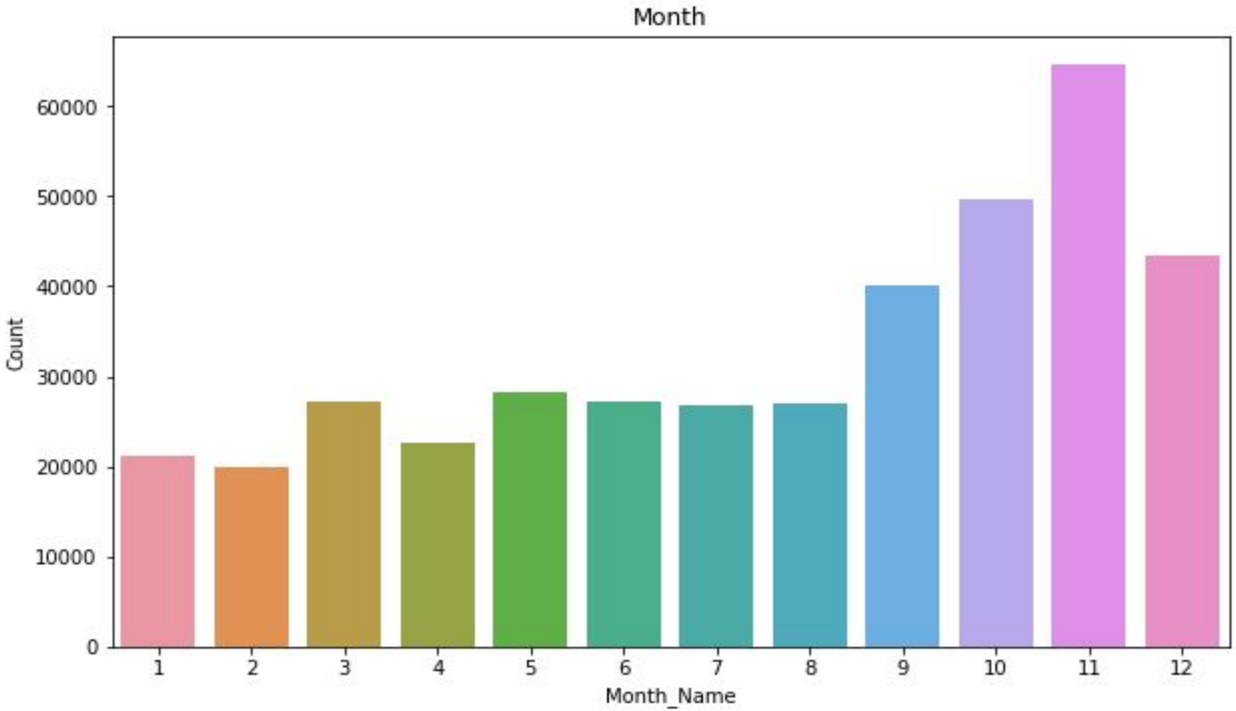
Analysis on Country(Cont.)



	Country_Name	Count
32	Lithuania	35
33	Brazil	32
34	Czech Republic	25
35	Bahrain	17
36	Saudi Arabia	9



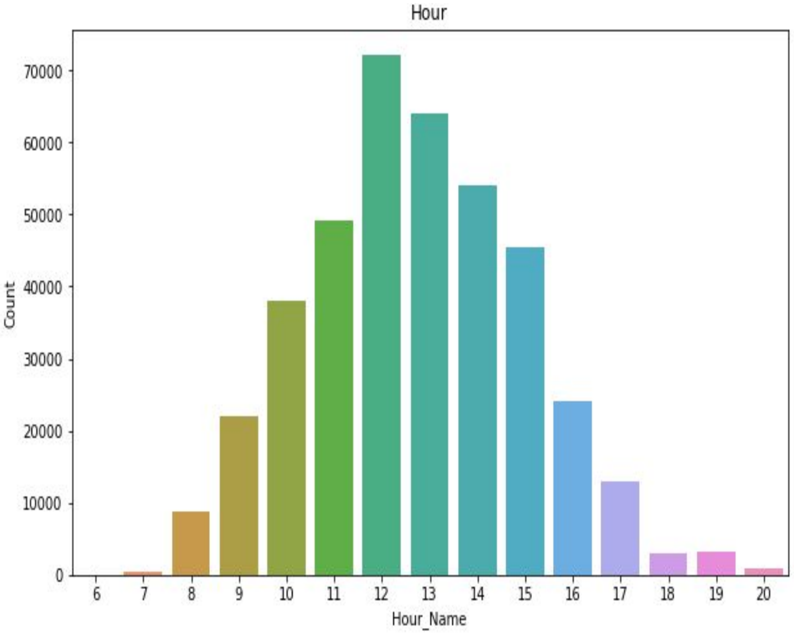
Monthly Purchasing

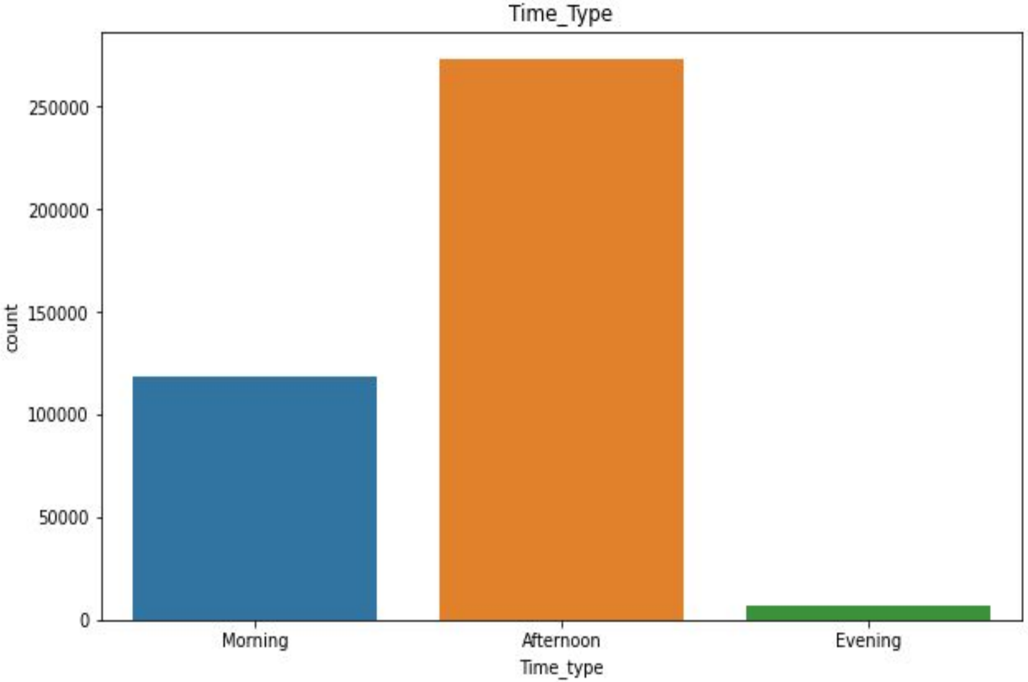


- Most number of purchasing is done in month of November.

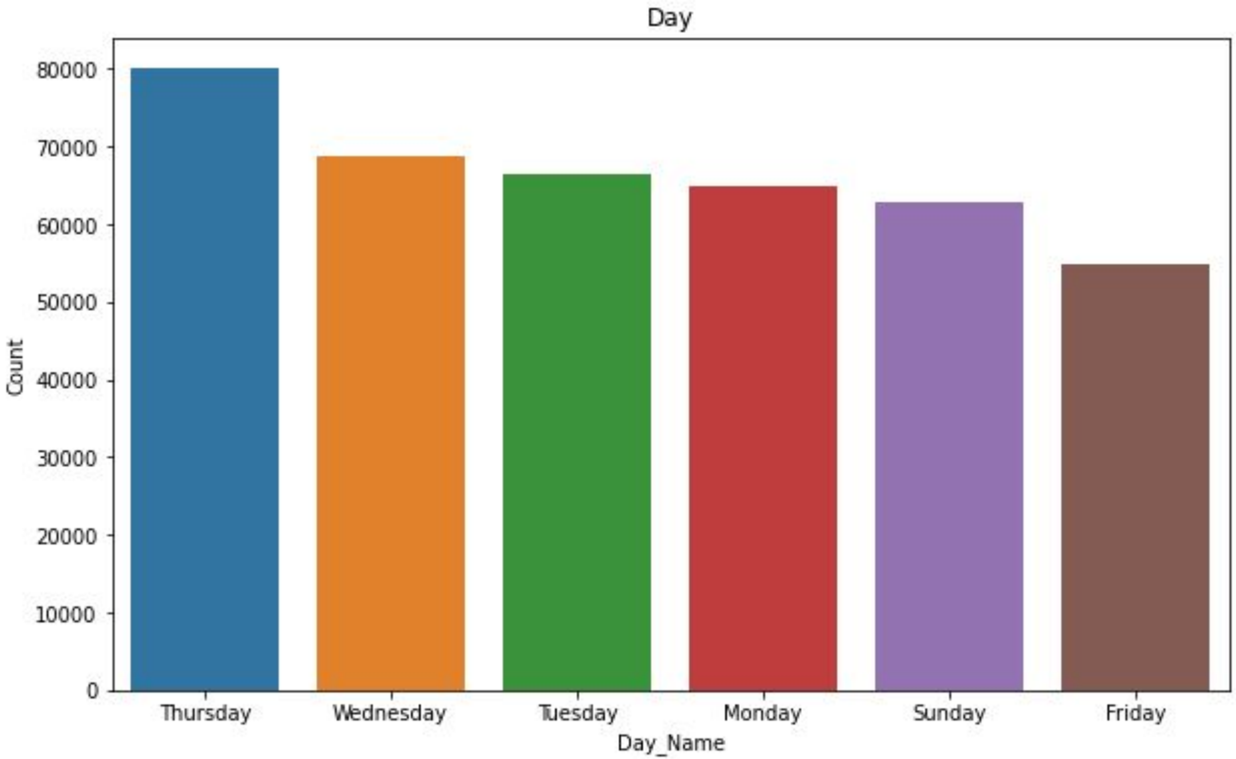
Hourly Purchasing Analysis

- Working has the most sales. This also indicates that the large part of the data is of wholesaler data.



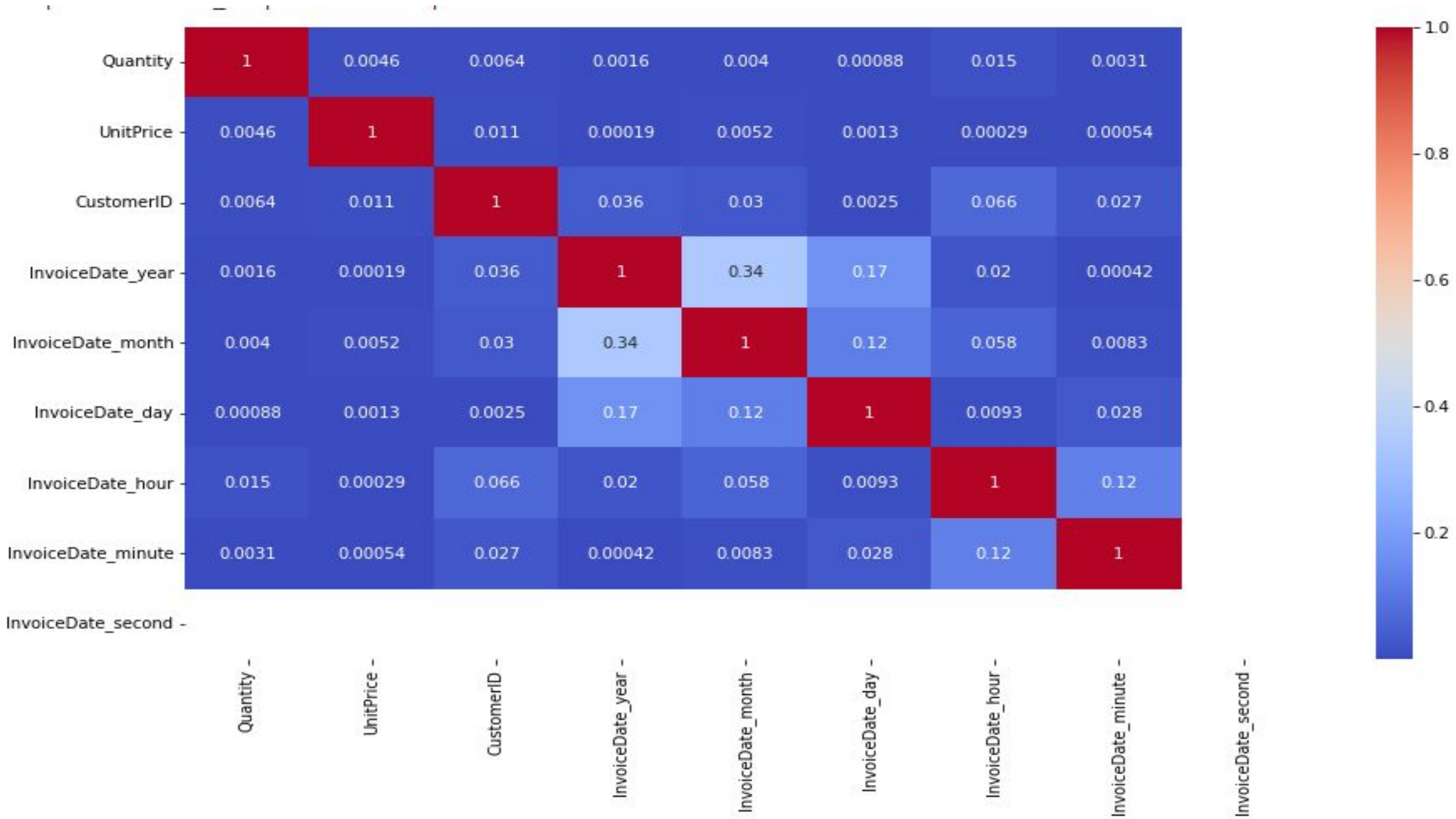


Weekly Purchasing Analysis



- Thursday, Wednesday and Tuesday has highest number of sales.

Correlation Matrix



	Quantity	UnitPrice	CustomerID	InvoiceDate_year	InvoiceDate_month	InvoiceDate_day	InvoiceDate_hour	InvoiceDate_minute	InvoiceDate_second
Quantity	1	0.0046	0.0064	0.0016	0.004	0.00088	0.015	0.0031	
UnitPrice	0.0046	1	0.011	0.00019	0.0052	0.0013	0.00029	0.00054	
CustomerID	0.0064	0.011	1	0.036	0.03	0.0025	0.066	0.027	
InvoiceDate_year	0.0016	0.00019	0.036	1	0.34	0.17	0.02	0.00042	
InvoiceDate_month	0.004	0.0052	0.03	0.34	1	0.12	0.058	0.0083	
InvoiceDate_day	0.00088	0.0013	0.0025	0.17	0.12	1	0.0093	0.028	
InvoiceDate_hour	0.015	0.00029	0.066	0.02	0.058	0.0093	1	0.12	
InvoiceDate_minute	0.0031	0.00054	0.027	0.00042	0.0083	0.028	0.12	1	
InvoiceDate_second									

RFM(Recency, Frequency and Monetary) Model

Recency

The freshness of customer activity

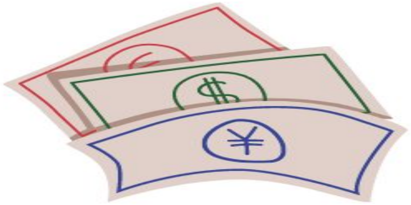
How recently a customer has made a purchase?

Frequency

The frequency of customer visits or transactions

How often a customer makes a purchase?

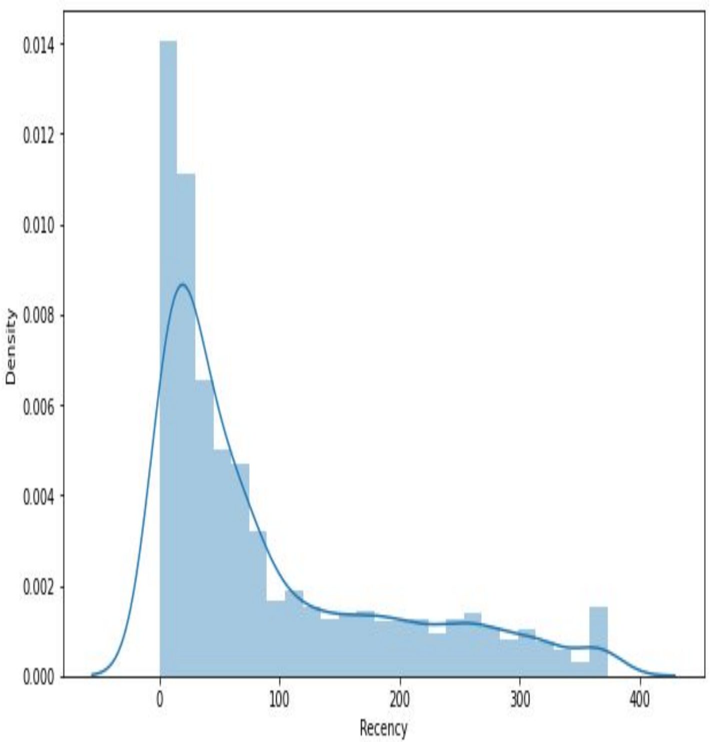
Monetary



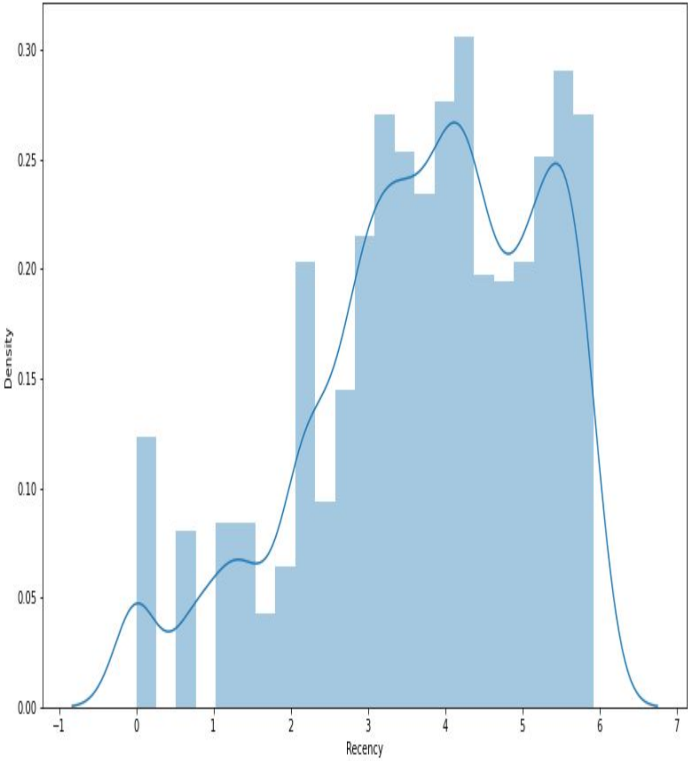
Purchasing power of customer

How much money a customer spends on purchases?

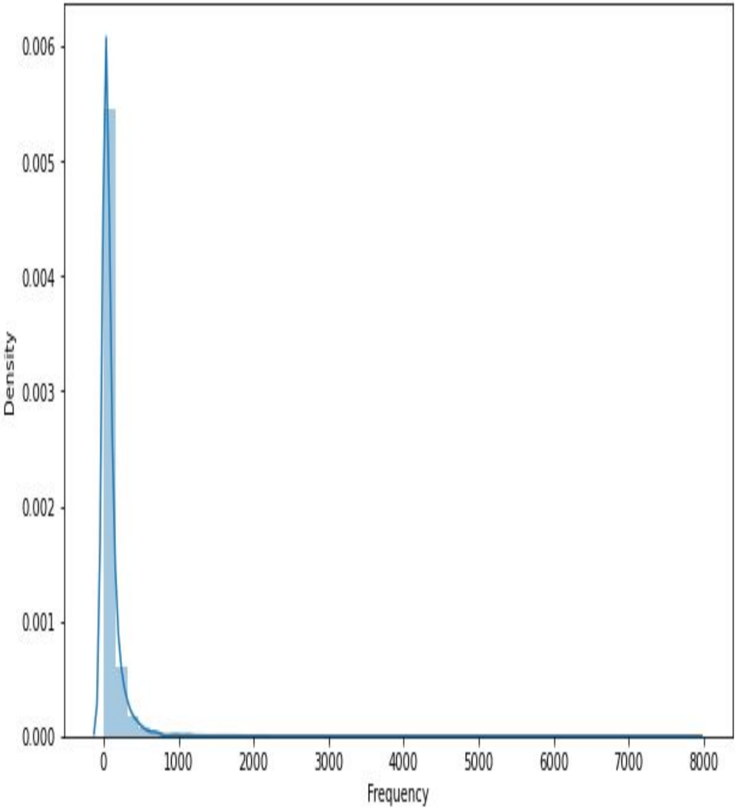
Recency Normalization



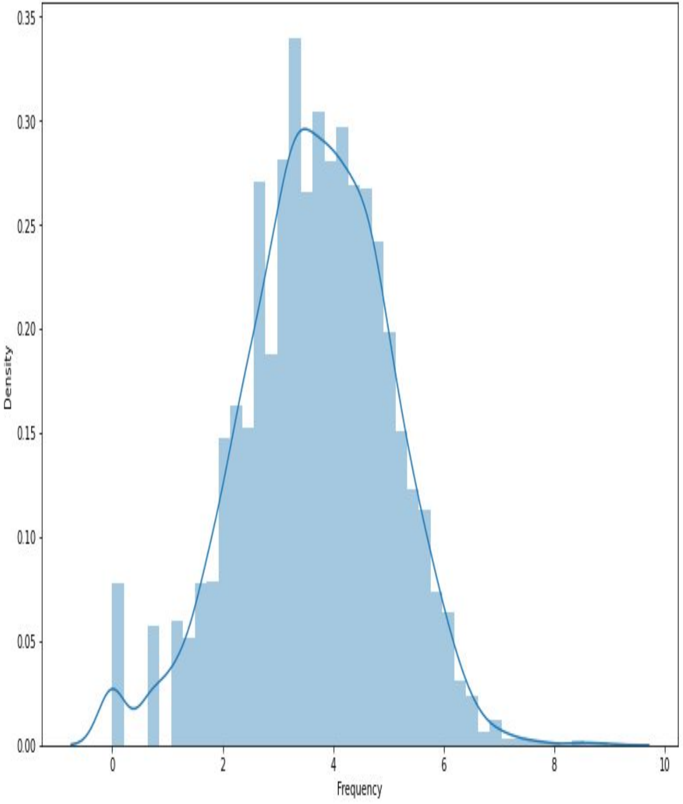
Log Transformation



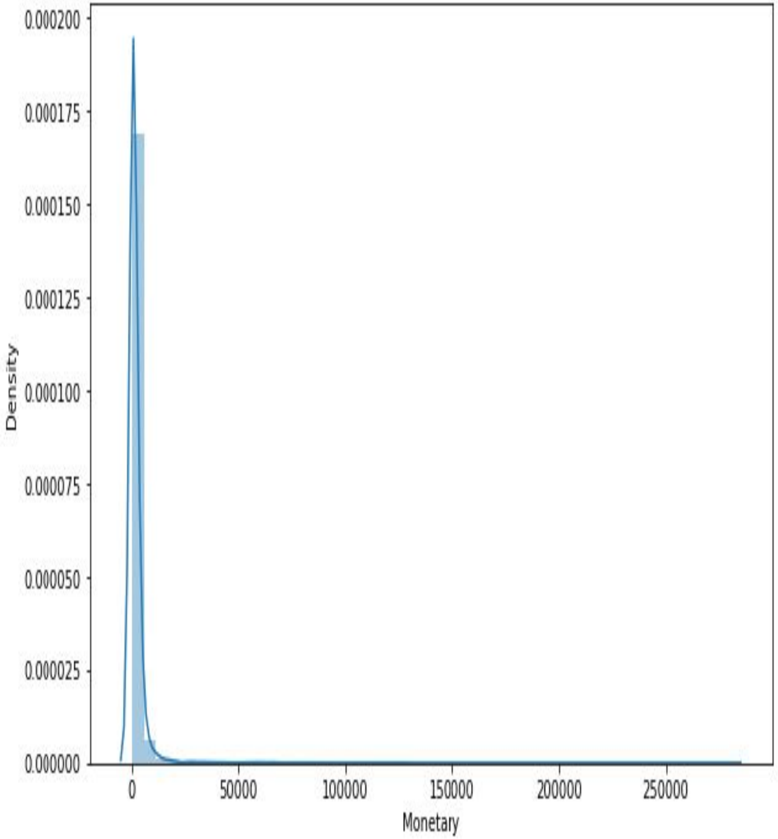
Frequency Normalization



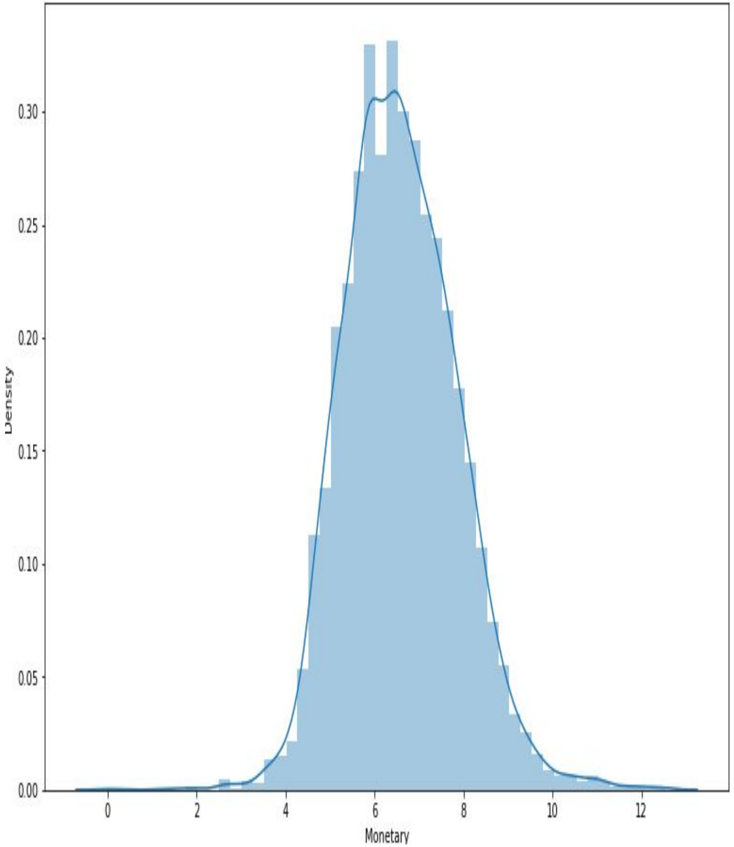
Log Transformation



Monetary Normalization



Log Transformation



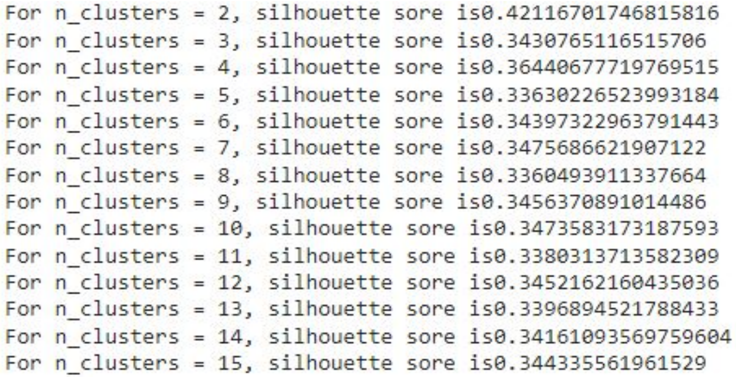
Modeling Overview

- ❑ Unsupervised learning / Clustering
- ❑ Dataset with null values, duplicate values and few outliers

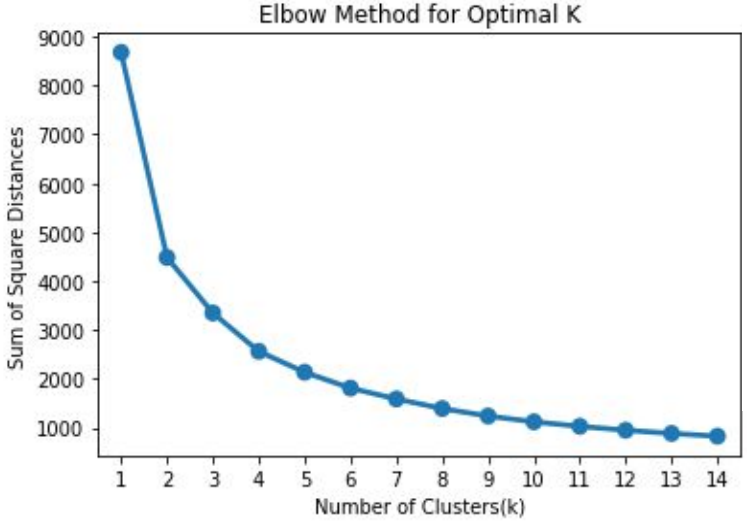
❑ Models Used

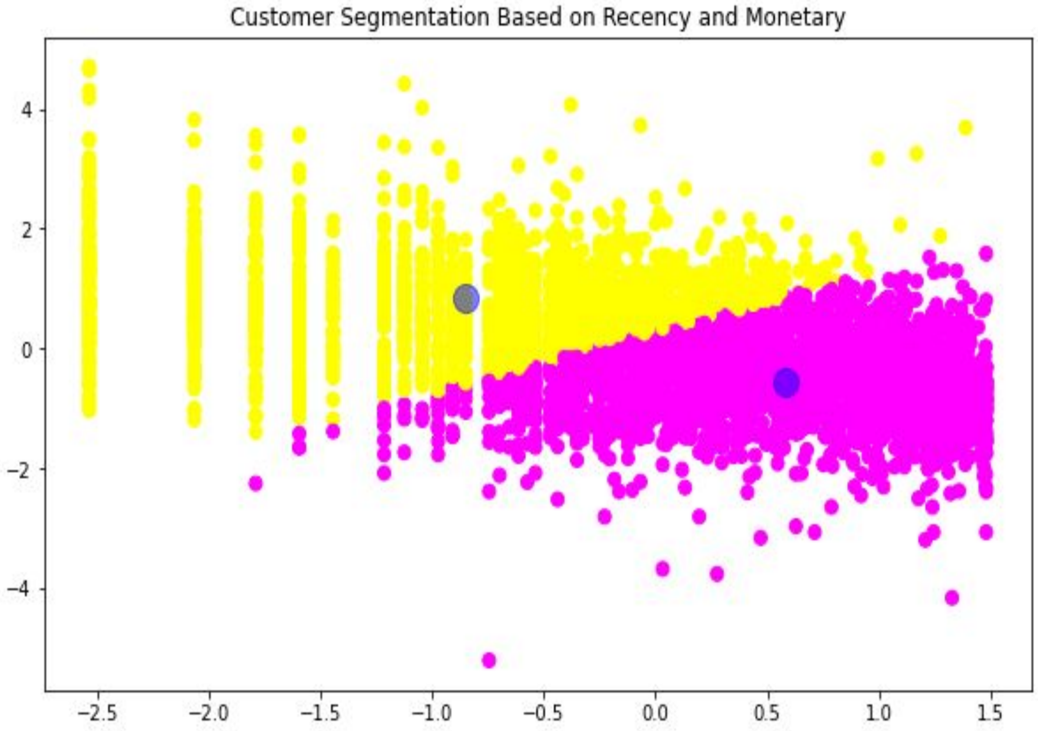
- ❑ K-Means Clustering
- ❑ DBSCAN
- ❑ Hierarchical Clustering (Dendrogram)

Silhouette Score & Elbow Method on R and M

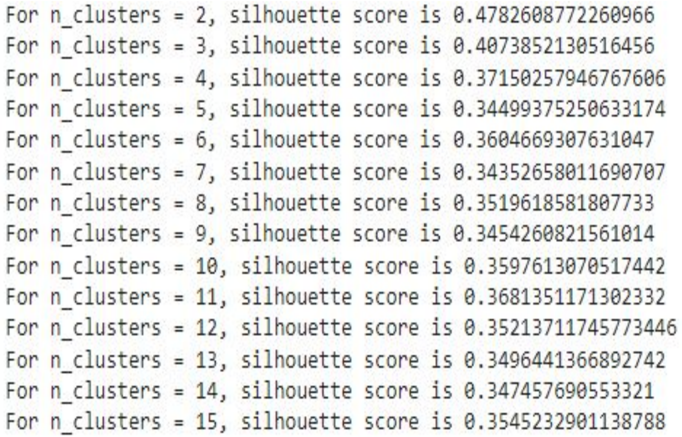


```
For n_clusters = 2, silhouette sore is0.42116701746815816
For n_clusters = 3, silhouette sore is0.3430765116515706
For n_clusters = 4, silhouette sore is0.36440677719769515
For n_clusters = 5, silhouette sore is0.33630226523993184
For n_clusters = 6, silhouette sore is0.34397322963791443
For n_clusters = 7, silhouette sore is0.3475686621907122
For n_clusters = 8, silhouette sore is0.3360493911337664
For n_clusters = 9, silhouette sore is0.3456370891014486
For n_clusters = 10, silhouette sore is0.3473583173187593
For n_clusters = 11, silhouette sore is0.3380313713582309
For n_clusters = 12, silhouette sore is0.3452162160435036
For n_clusters = 13, silhouette sore is0.3396894521788433
For n_clusters = 14, silhouette sore is0.34161093569759604
For n_clusters = 15, silhouette sore is0.344335561961529
```

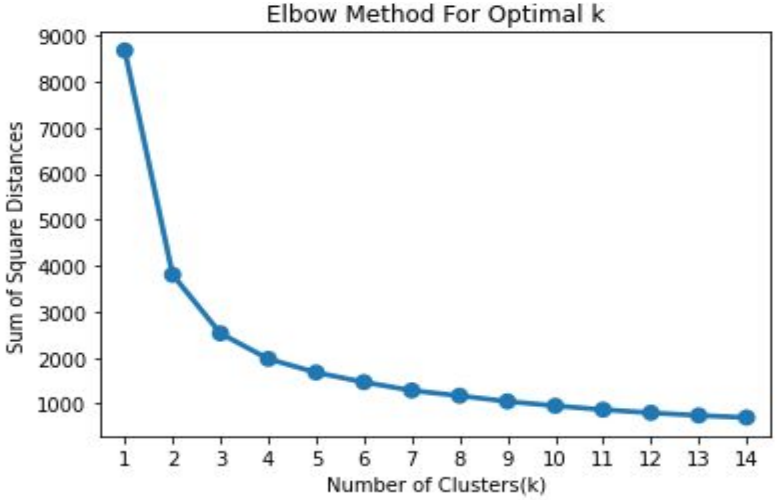




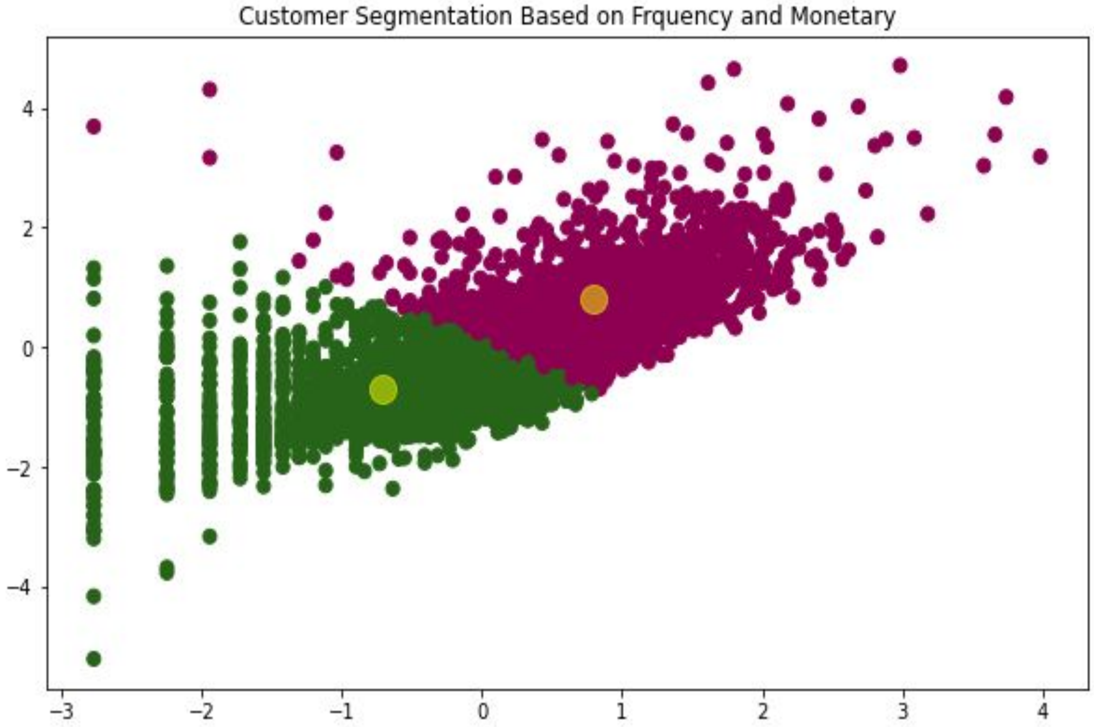
Silhouette Score & Elbow Method on F and M



```
For n_clusters = 2, silhouette score is 0.4782608772260966
For n_clusters = 3, silhouette score is 0.4073852130516456
For n_clusters = 4, silhouette score is 0.37150257946767606
For n_clusters = 5, silhouette score is 0.34499375250633174
For n_clusters = 6, silhouette score is 0.3604669307631047
For n_clusters = 7, silhouette score is 0.34352658011690707
For n_clusters = 8, silhouette score is 0.3519618581807733
For n_clusters = 9, silhouette score is 0.3454260821561014
For n_clusters = 10, silhouette score is 0.3597613070517442
For n_clusters = 11, silhouette score is 0.3681351171302332
For n_clusters = 12, silhouette score is 0.35213711745773446
For n_clusters = 13, silhouette score is 0.3496441366892742
For n_clusters = 14, silhouette score is 0.347457690553321
For n_clusters = 15, silhouette score is 0.3545232901138788
```

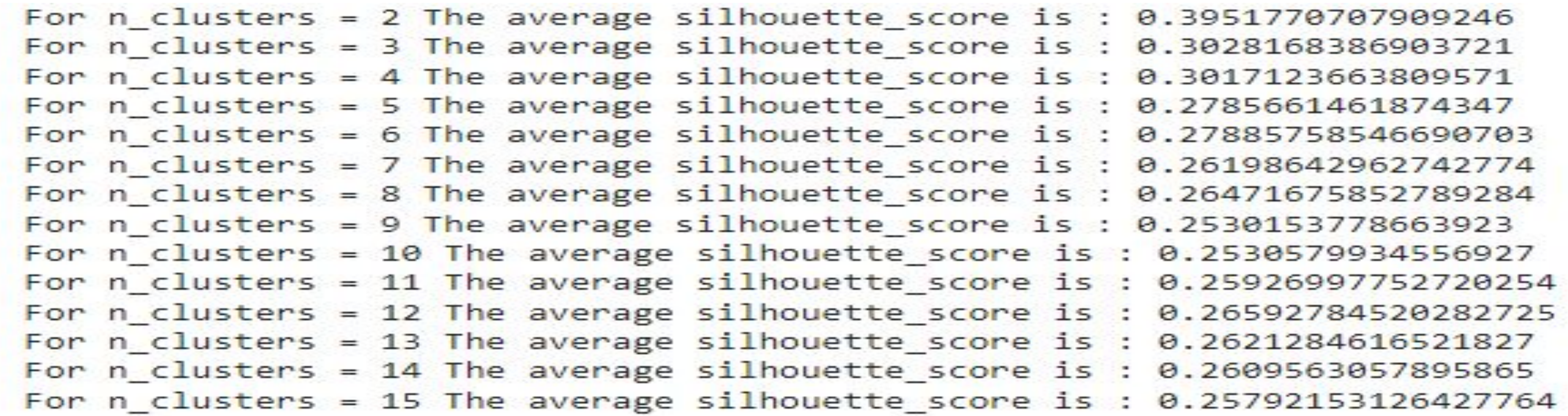


Elbow Method For Optimal k

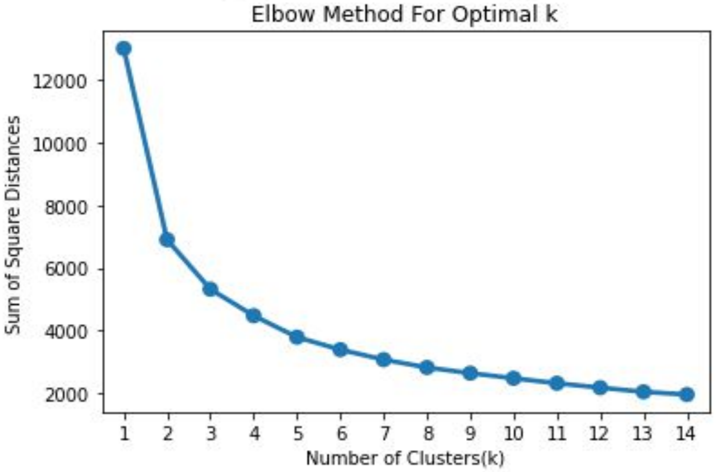


Customer Segmentation Based on Frquency and Monetary

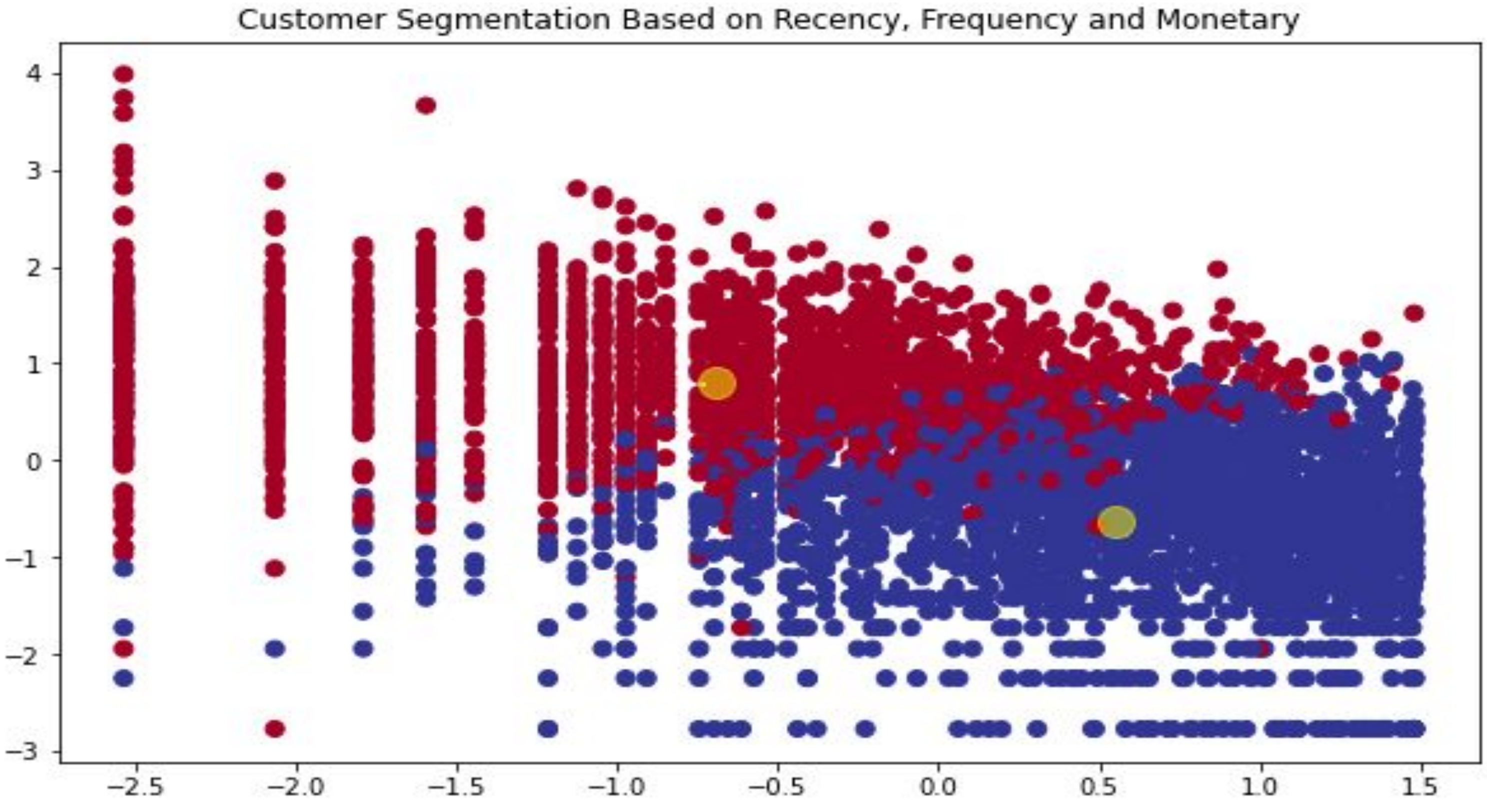
Silhouette Score & Elbow Method on R, F and M



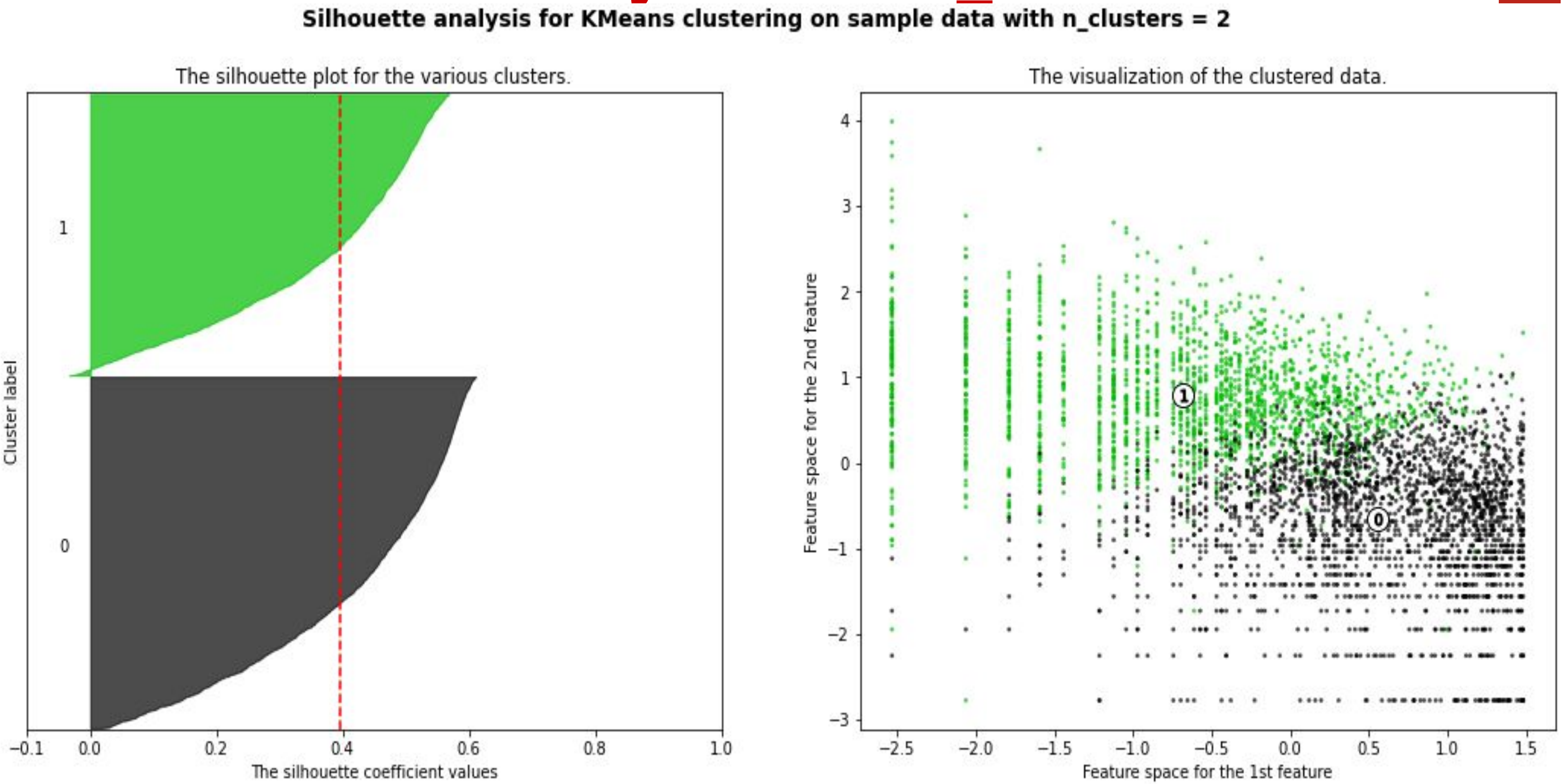
```
For n_clusters = 2 The average silhouette_score is : 0.395177070709246
For n_clusters = 3 The average silhouette_score is : 0.302816838690372l
For n_clusters = 4 The average silhouette_score is : 0.301712366380957l
For n_clusters = 5 The average silhouette_score is : 0.278566146187434 7
For n_clusters = 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 : 0.253057993455692 7
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
```



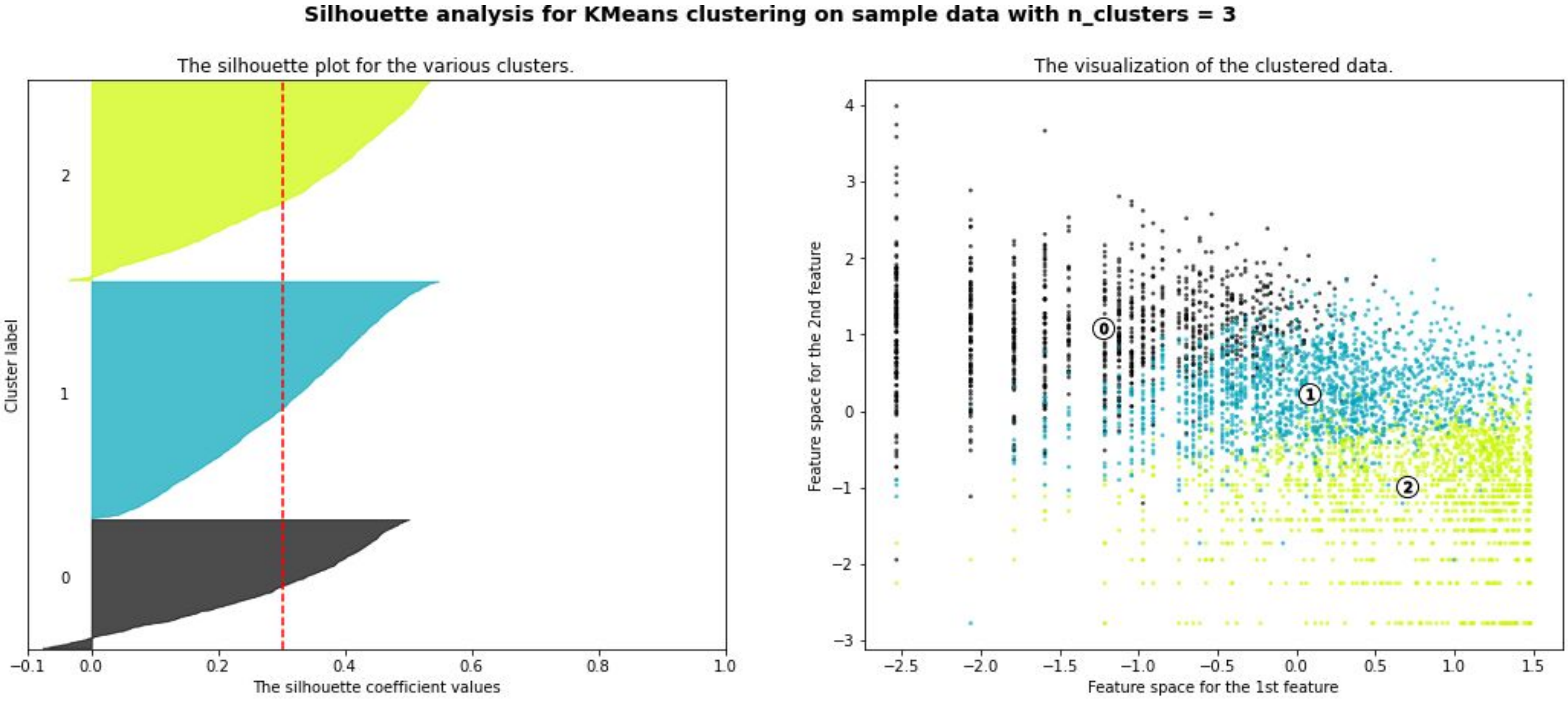
Customer Segmentation based on R, F and M



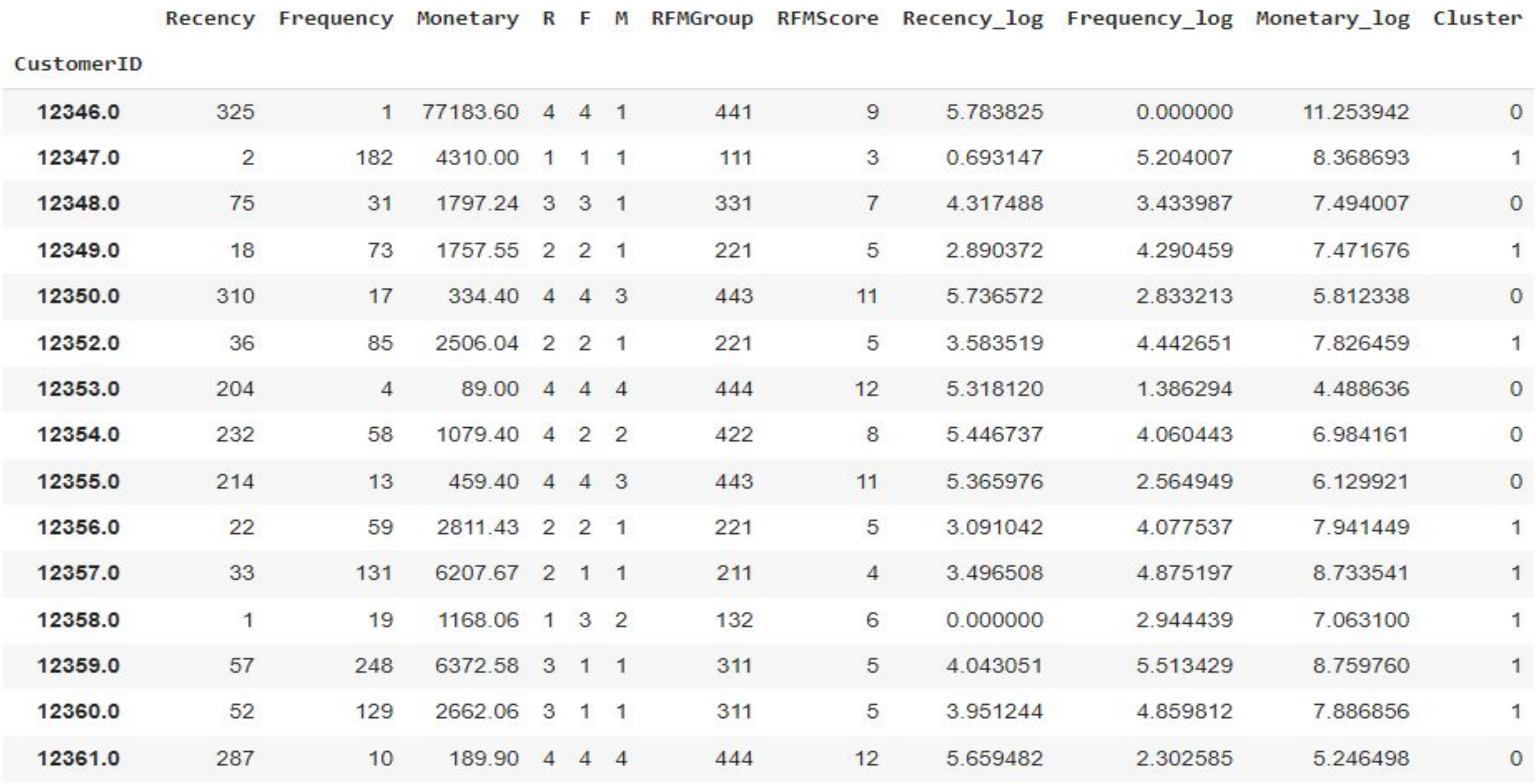
Silhouette Analysis With n_cluster



Silhouette Analysis With n_cluster

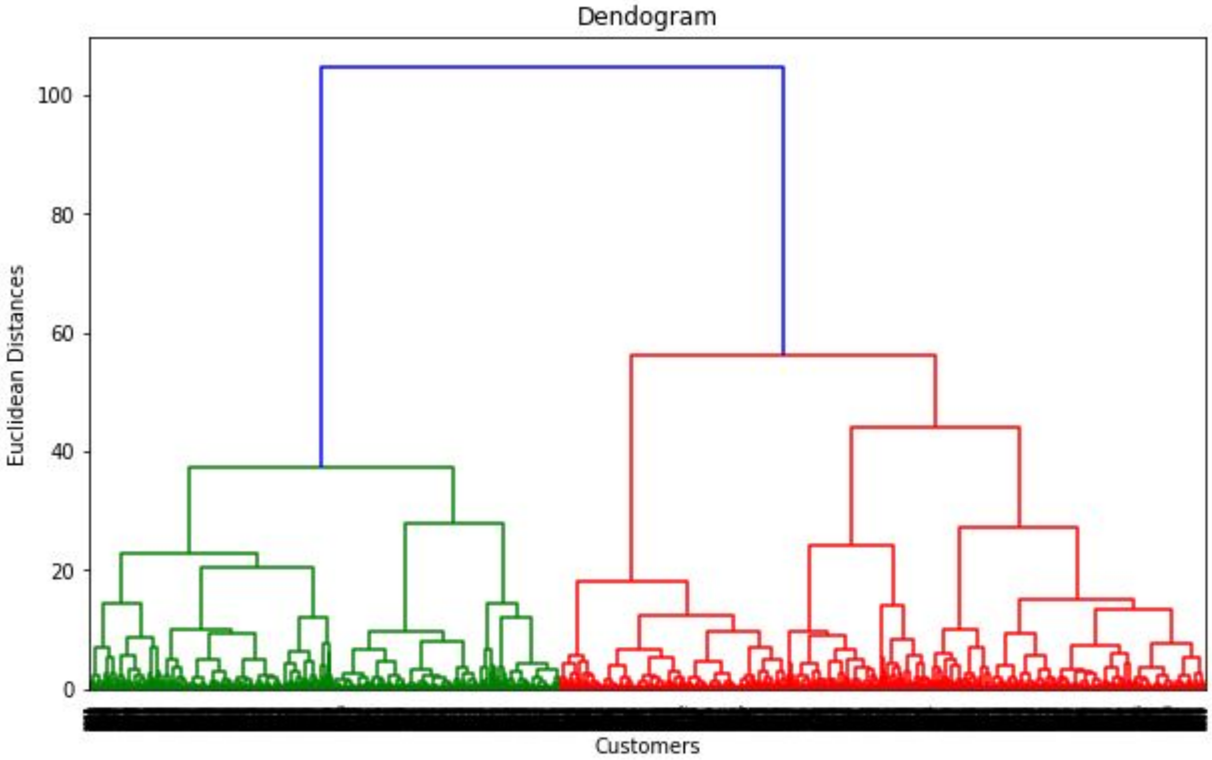


RFM Analysis



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
12357.0	33	131	6207.67	2	1	1	211	4	3.496508	4.875197	8.733541	1
12358.0	1	19	1168.06	1	3	2	132	6	0.000000	2.944439	7.063100	1
12359.0	57	248	6372.58	3	1	1	311	5	4.043051	5.513429	8.759760	1
12360.0	52	129	2662.06	3	1	1	311	5	3.951244	4.859812	7.886856	1
12361.0	287	10	189.90	4	4	4	444	12	5.659482	2.302585	5.246498	0

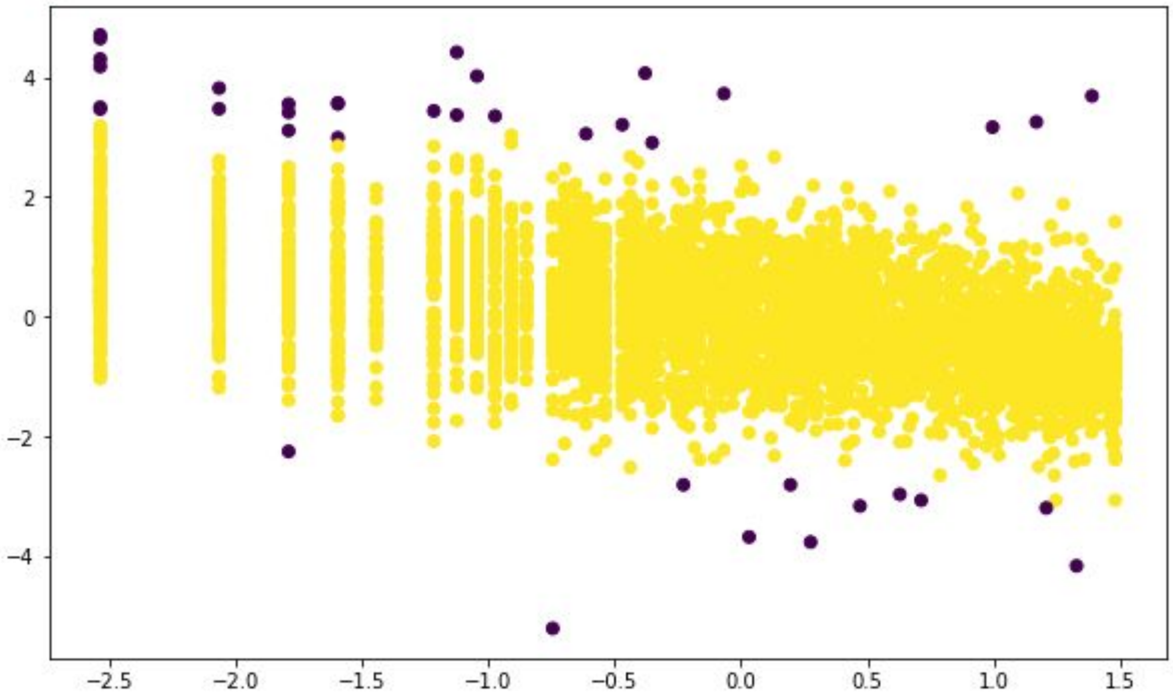
Hierarchical Clustering (Dendogram)



- The number of clusters will be number of vertical lines which are being intersected by the line drawn using the threshold = 90
- Number of cluster is 2

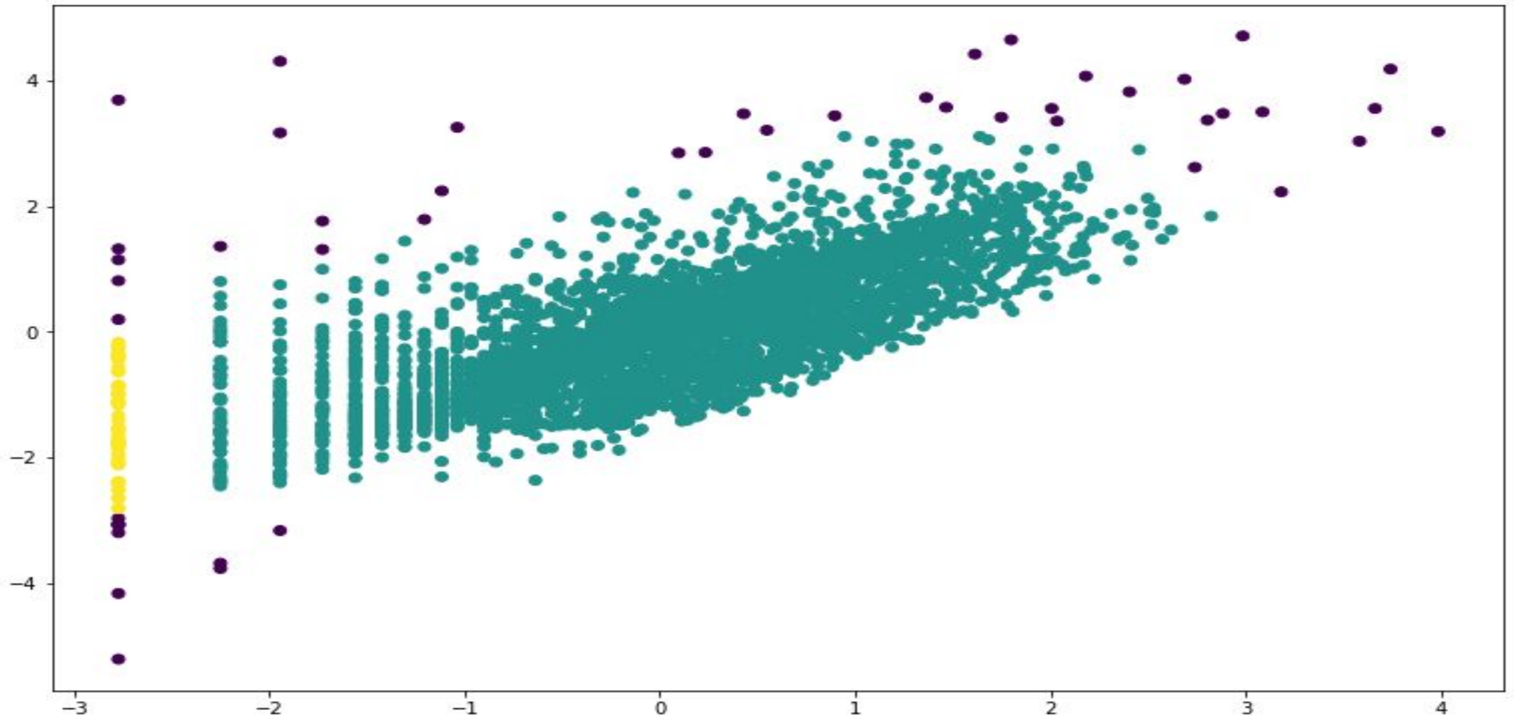
DBSCAN on R and M

- **DBSCAN** is Density Based Spatial Clustering of Applications with Noise. It is distance between nearest points.



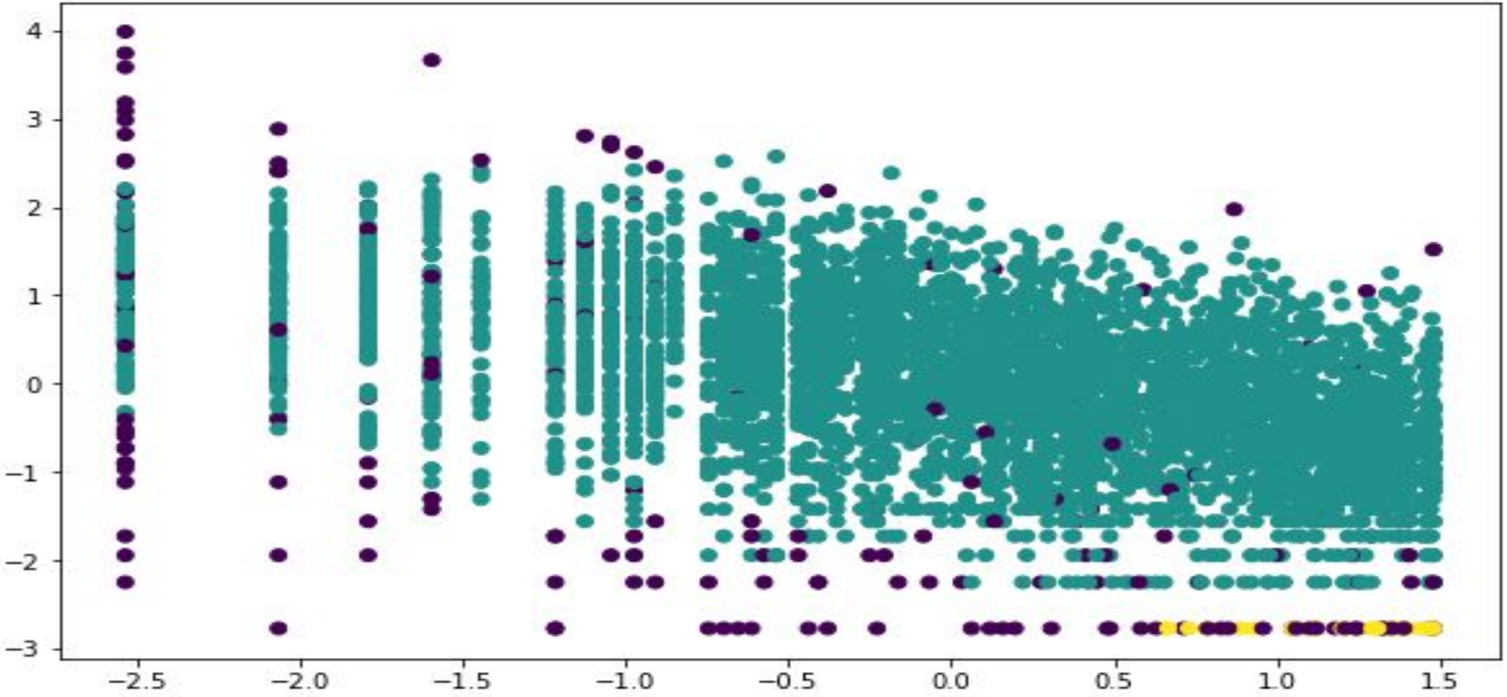
DBSCAN on F and M

- From plot we can see that customers are well segmented on basis of frequency and monetary.



DBSCAN on R, F and M

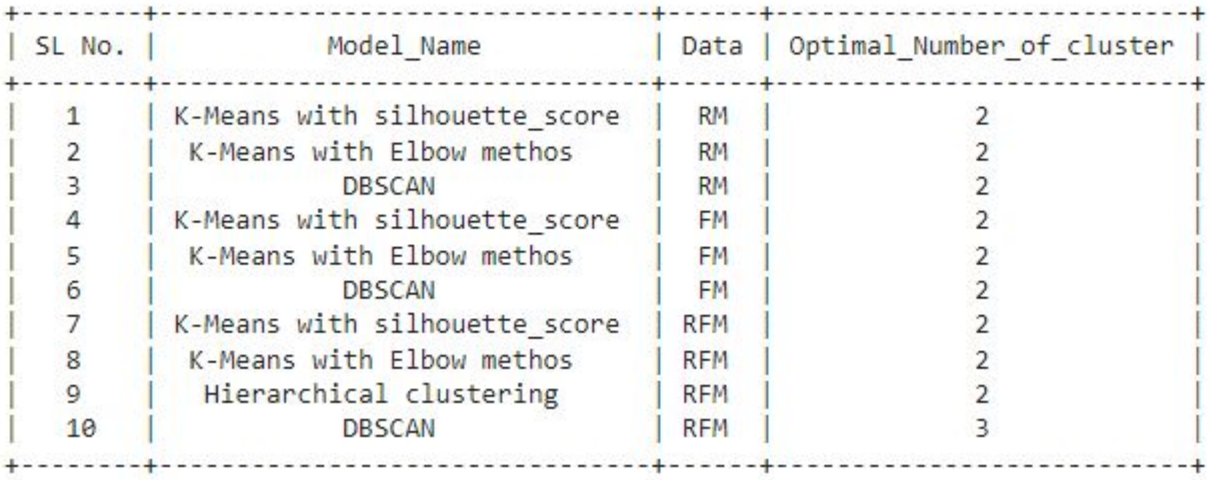
- We can see that customer are well separated on the basis of R, F and M. Here number of cluster are 3.



Challenges

- Data Cleaning
- Data mining
- Lot of null and duplicate values.
- Feature Engineering
- Feature Selection
- Model optimization
- Lot of plots and graphs to analyze
- Normalization
- Deciding the flow of presentation

Overall Conclusion



```
+--------+-------------------------------+------+---------------------------+
| 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             |
+--------+-------------------------------+------+---------------------------+
```

- Throughout the analysis we went through various steps to perform customer segmentation. I started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications. Next, I did some exploratory data analysis and tried to draw observations from the features we had in the dataset.
- I formulated some quantitative factors such as recency, frequency and monetary known as rfm model for each of the customers and implemented K Means clustering algorithm on these features.

Overall Conclusion(Cont.)

- I also performed silhouette and elbow method analysis to determine the optimal no. of clusters which was 2.

- I 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.

- There can be more modifications on this analysis. One may choose to cluster into more no. depending on company objectives and preferences. The labelled feature after clustering can be fed into classification supervised machine learning algorithms that could predict the classes for new set of observations. The clustering can also be performed on new set of features such as type of products each customer prefer to buy often, finding out customer lifetime value (clv), segmenting on the basis of time period they visit and much more.

- Hence segmentation of customer can be concluded by a class which has higher values of frequency and monetary but low values of recency. On the other hand the another class which has low values of frequency and monetary and high values of recency.

Thank You