Capstone Project : 4

Online Retail Customer Segmentation

Team Member

Atul Chouhan

Content

1. ❑ Understanding Business Problem
2. ❑ Dataset Information
3. ❑ Feature Analysis
4. ❑ Exploratory Data Analysis
5. ❑ Data Pre-processing
6. ❑ Model Implementing
7. ❑ Challenges
8. ❑ Conclusions

Understanding Business Problem

→ Topic – “Online Retail Customer Segmentation”

→ 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”.

-> Target is to know our customer, and maximize the profit of the retail company

Dataset Information

→This dataset contains 541909 observations and 8 features that contain the data of between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

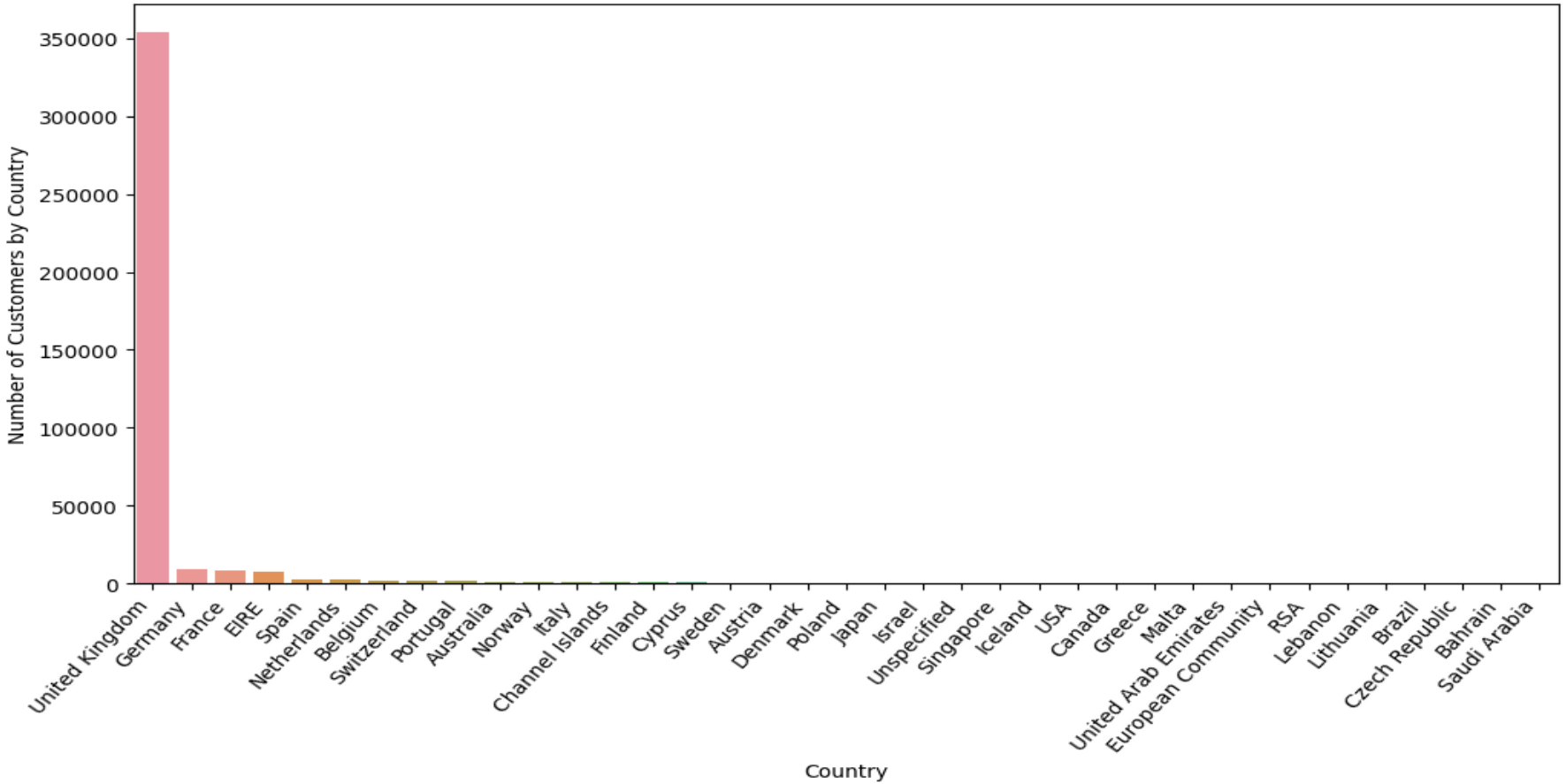
- There are 3 categorical features in our dataset.
- This dataset have null and duplicate values.

	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

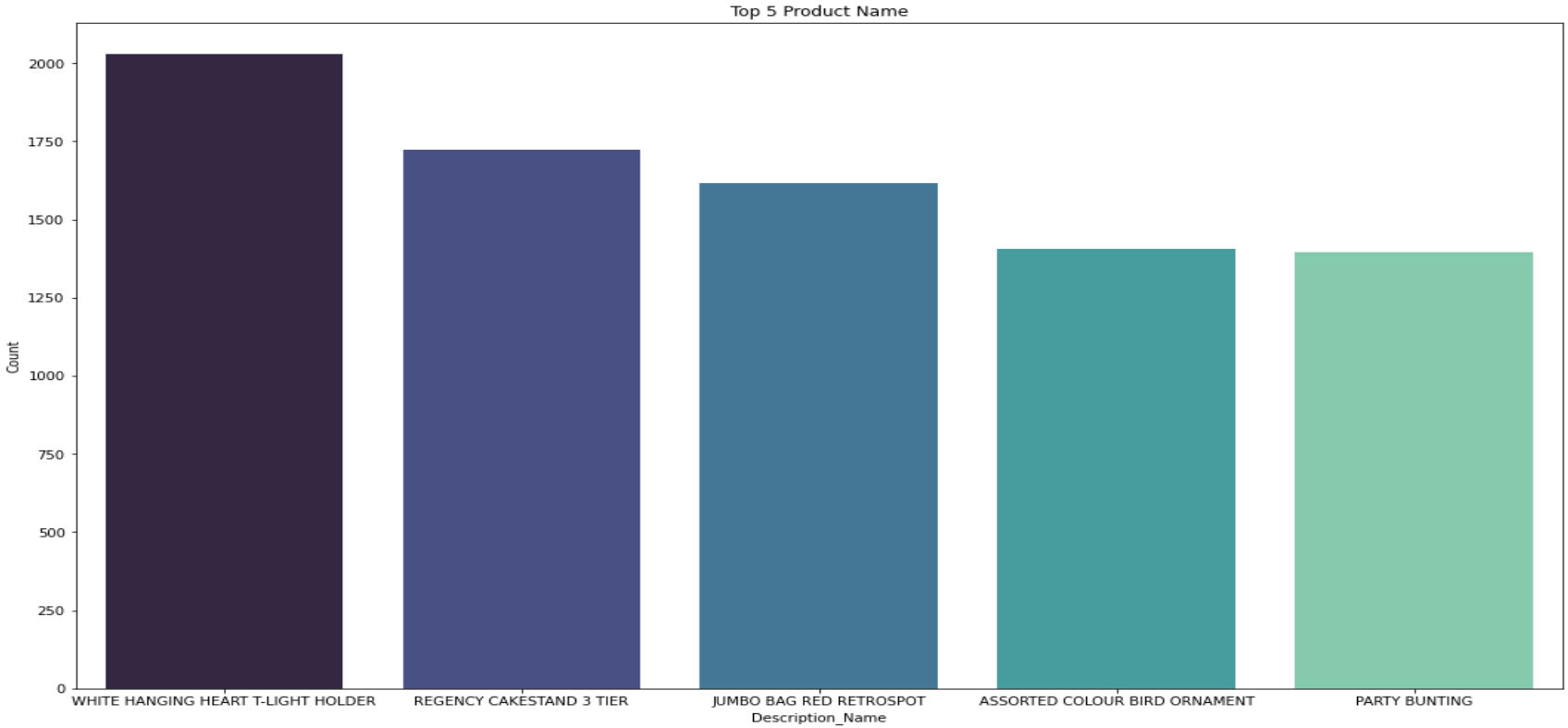
Feature Summary

- **InvoiceNo**: 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**: Invoice 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.

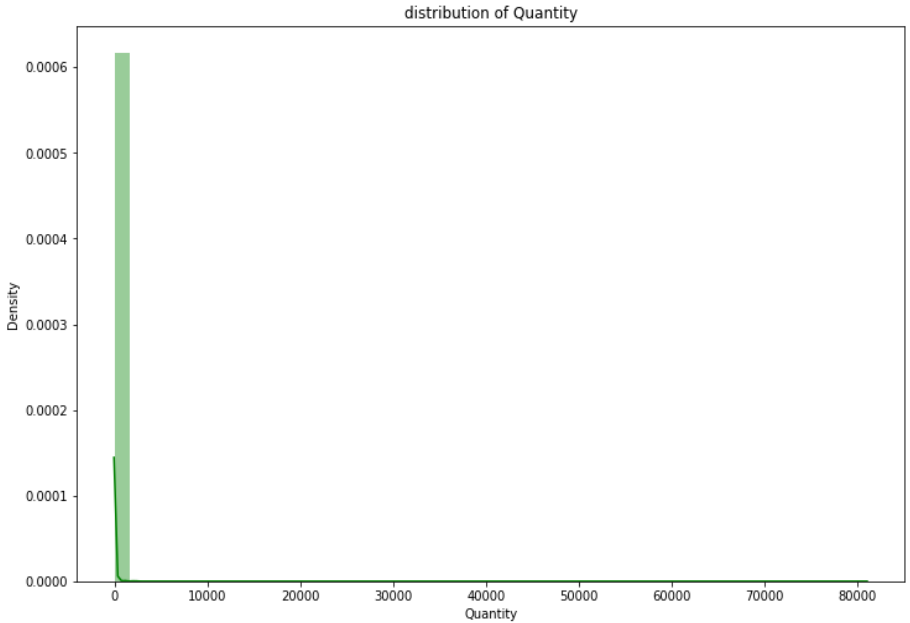
Categorical Feature Analysis Of Country column

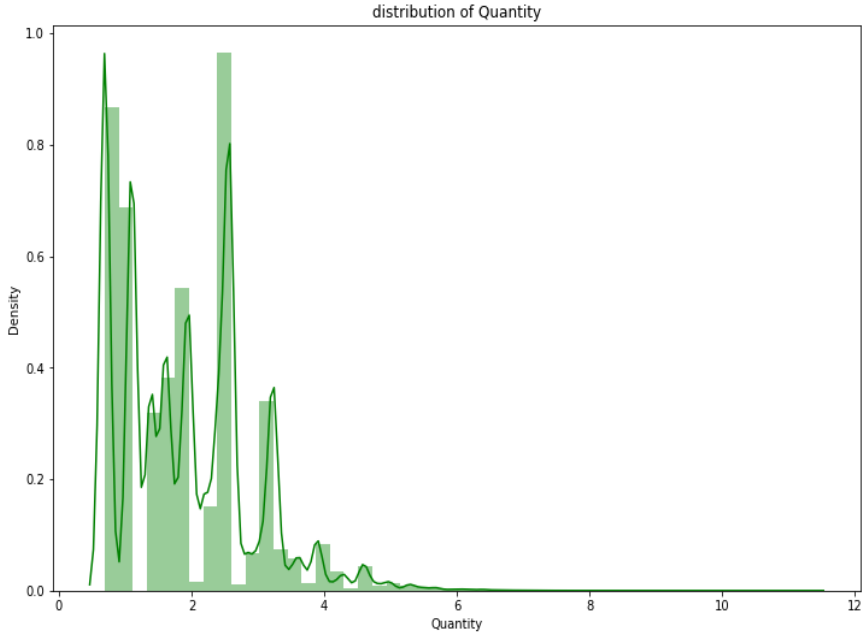


Analysis Of Product column



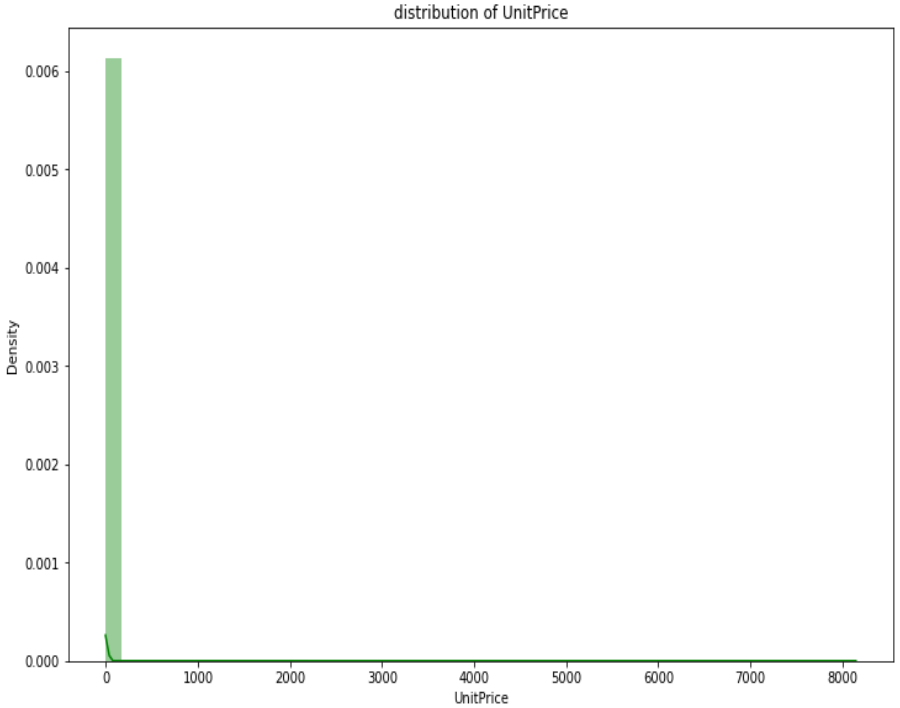
Analysis Of Numerical column

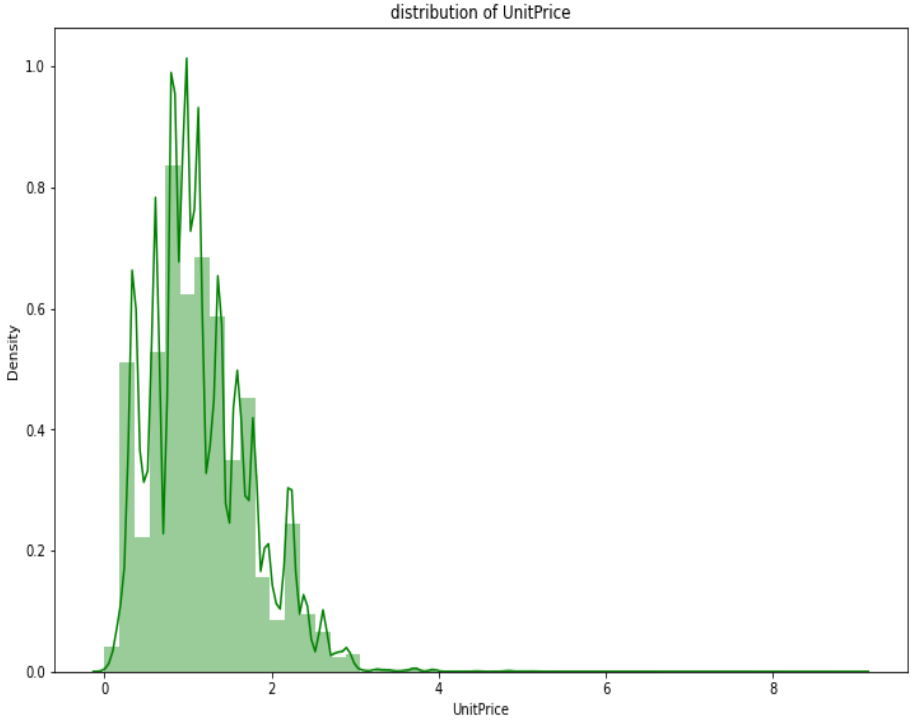




Data transformation

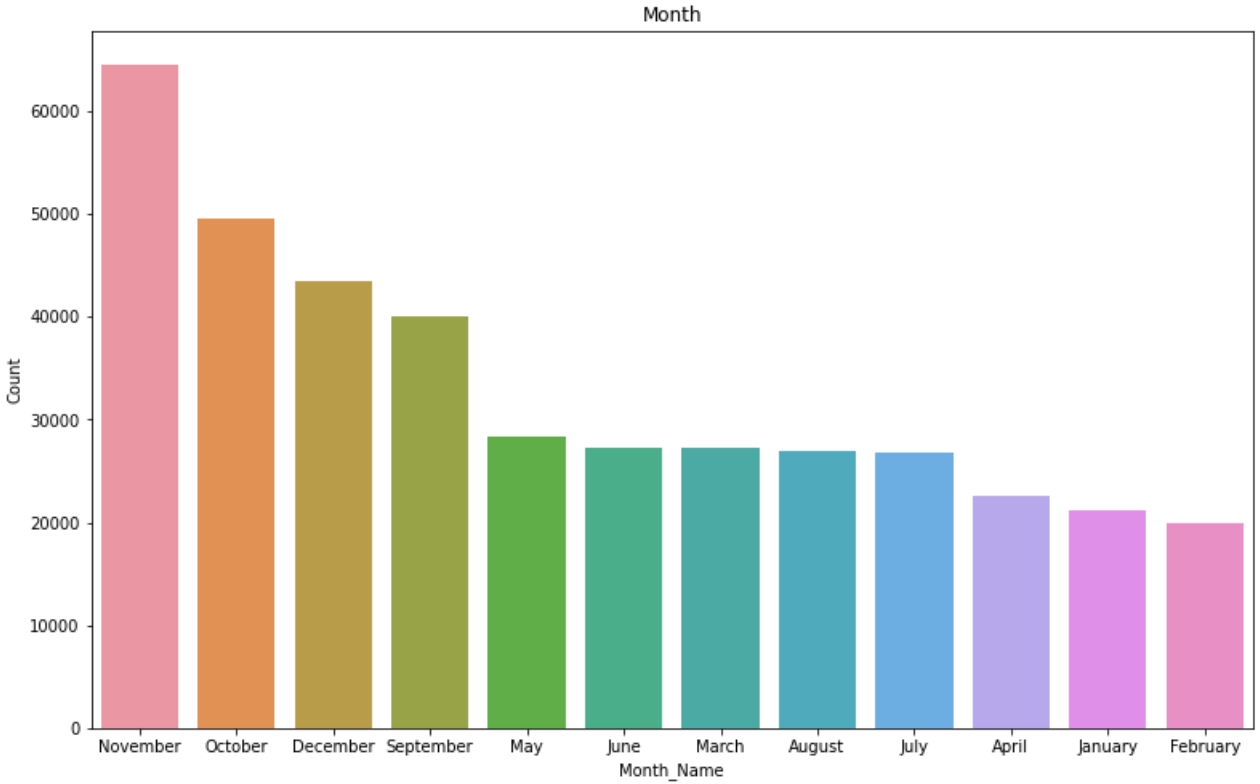
Analysis Of Numerical column



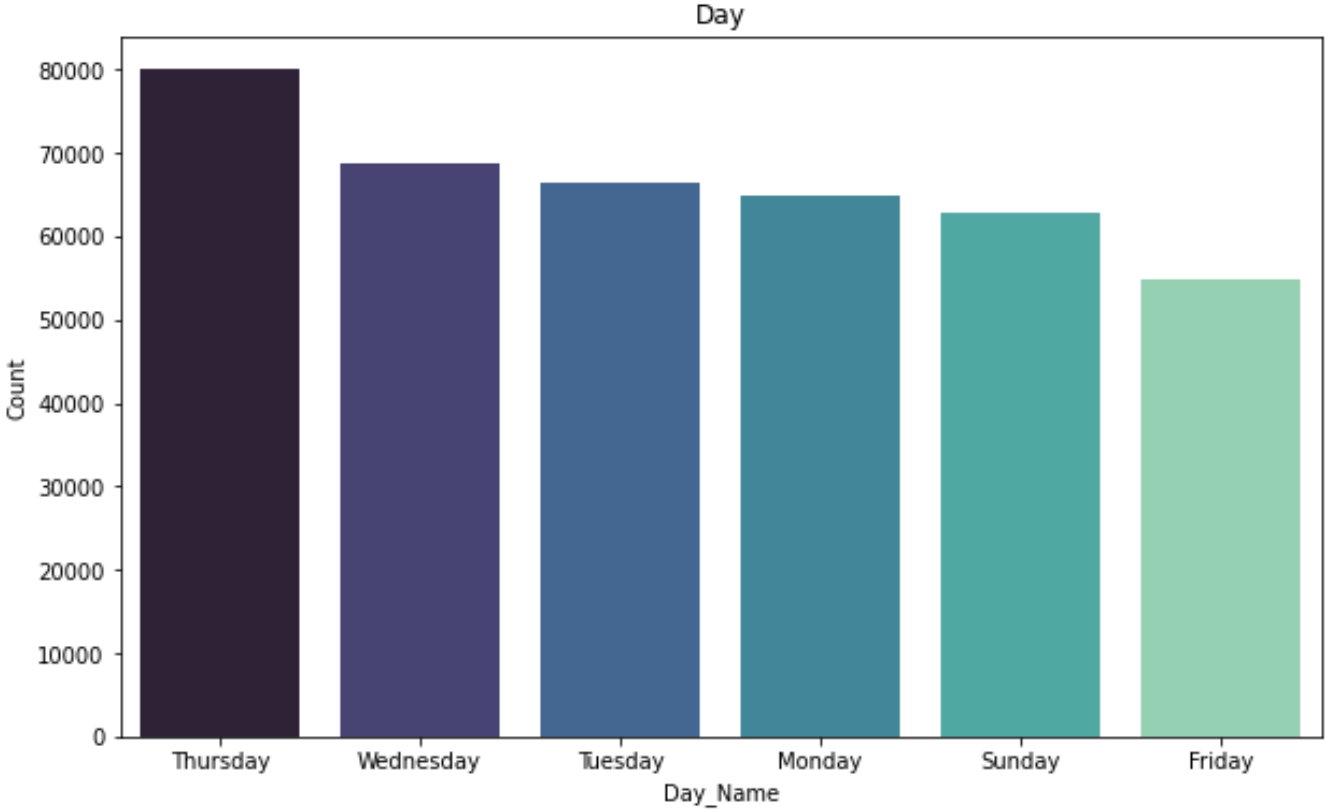


Data transformation

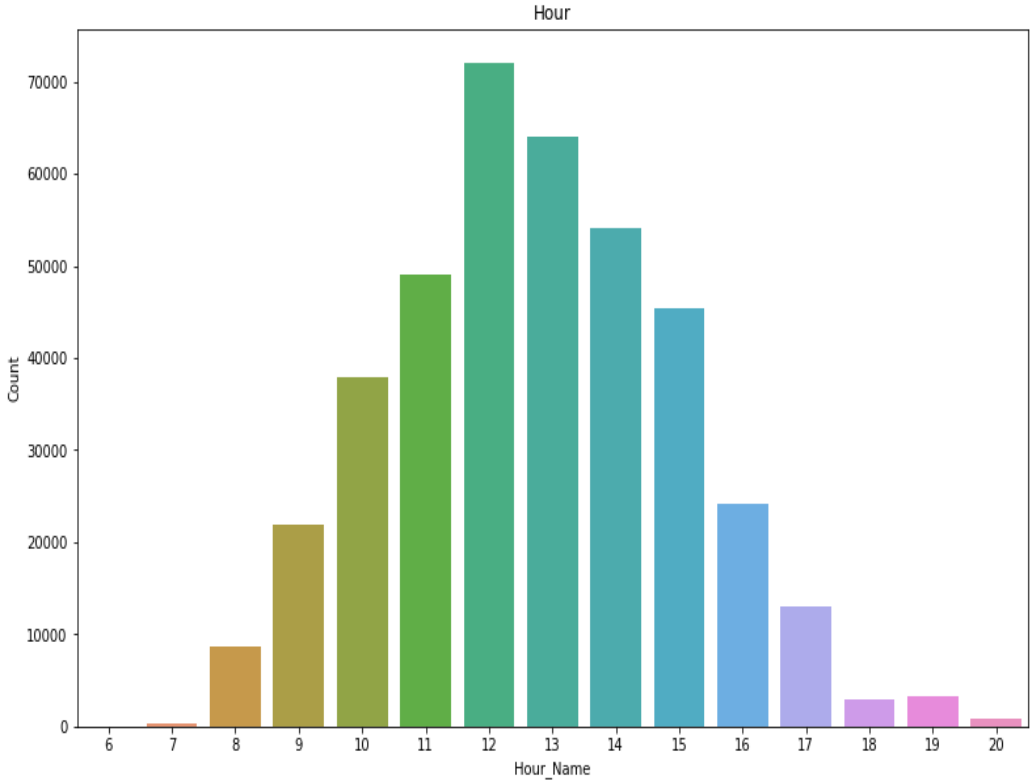
Analysis Of Month Column

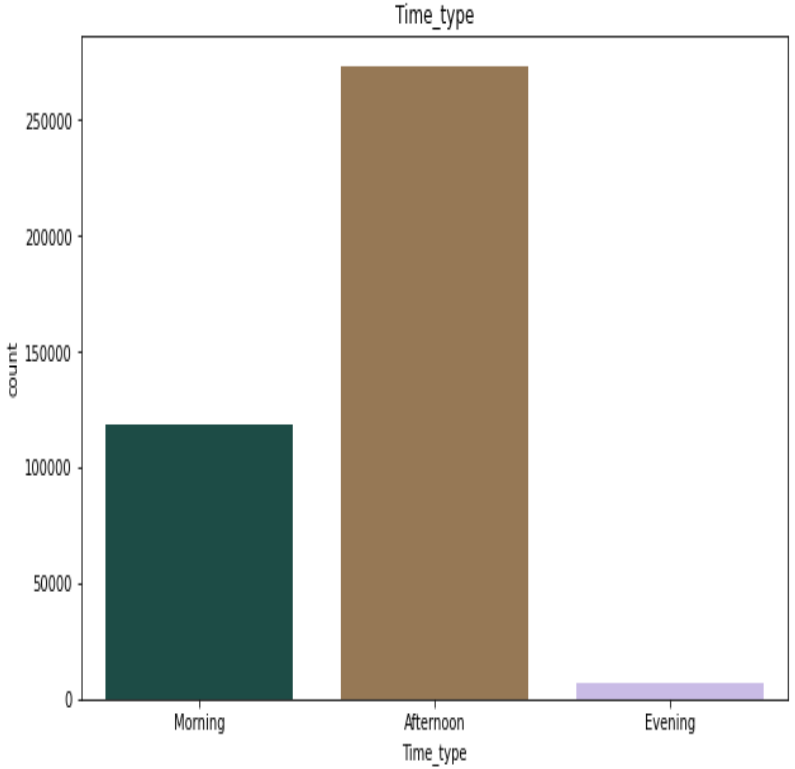


Analysis Of Day Column



Analysis Of Hour Column





RFM Implementation

RFM simply mean → Recency, Frequency, Monetary

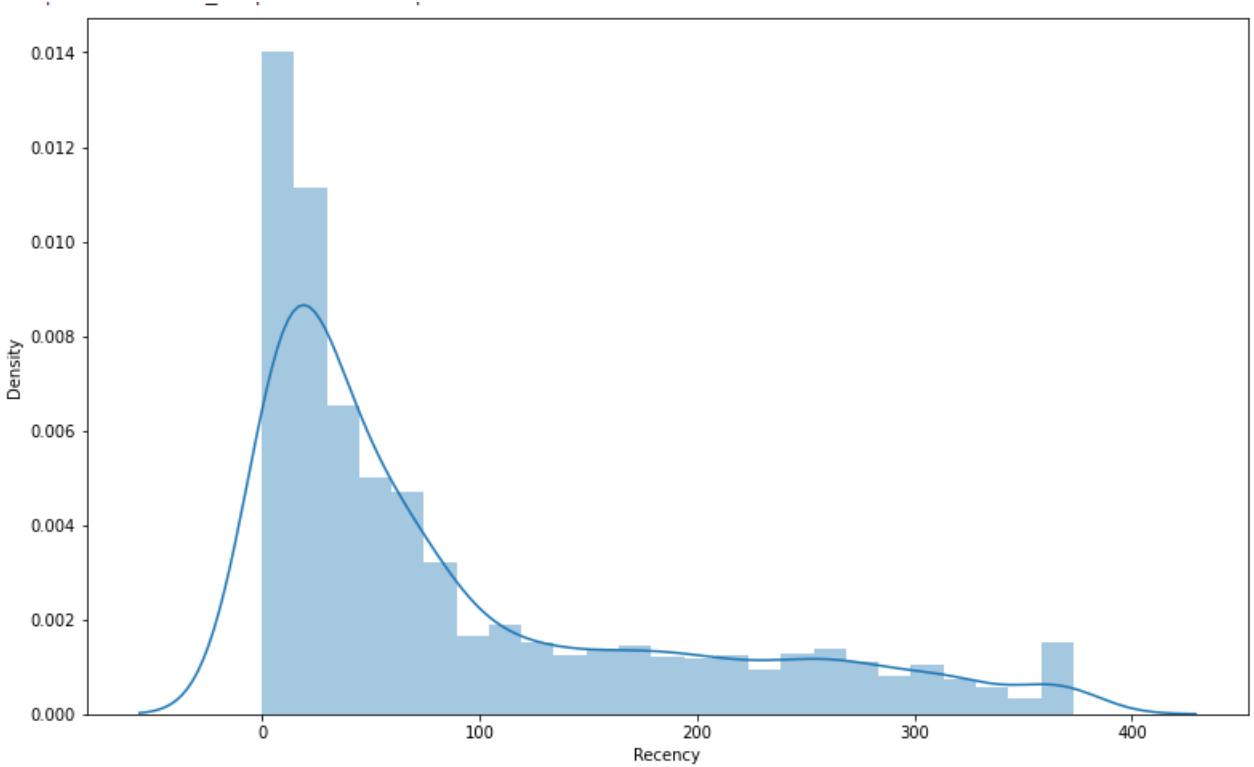
CustomerID	Recency	Frequency	Monetary	R	F	M
12346.0	325	1	77183.60	4	4	1
12347.0	2	182	4310.00	1	1	1
12348.0	75	31	1797.24	3	3	1
12349.0	18	73	1757.55	2	2	1
12350.0	310	17	334.40	4	4	3

→

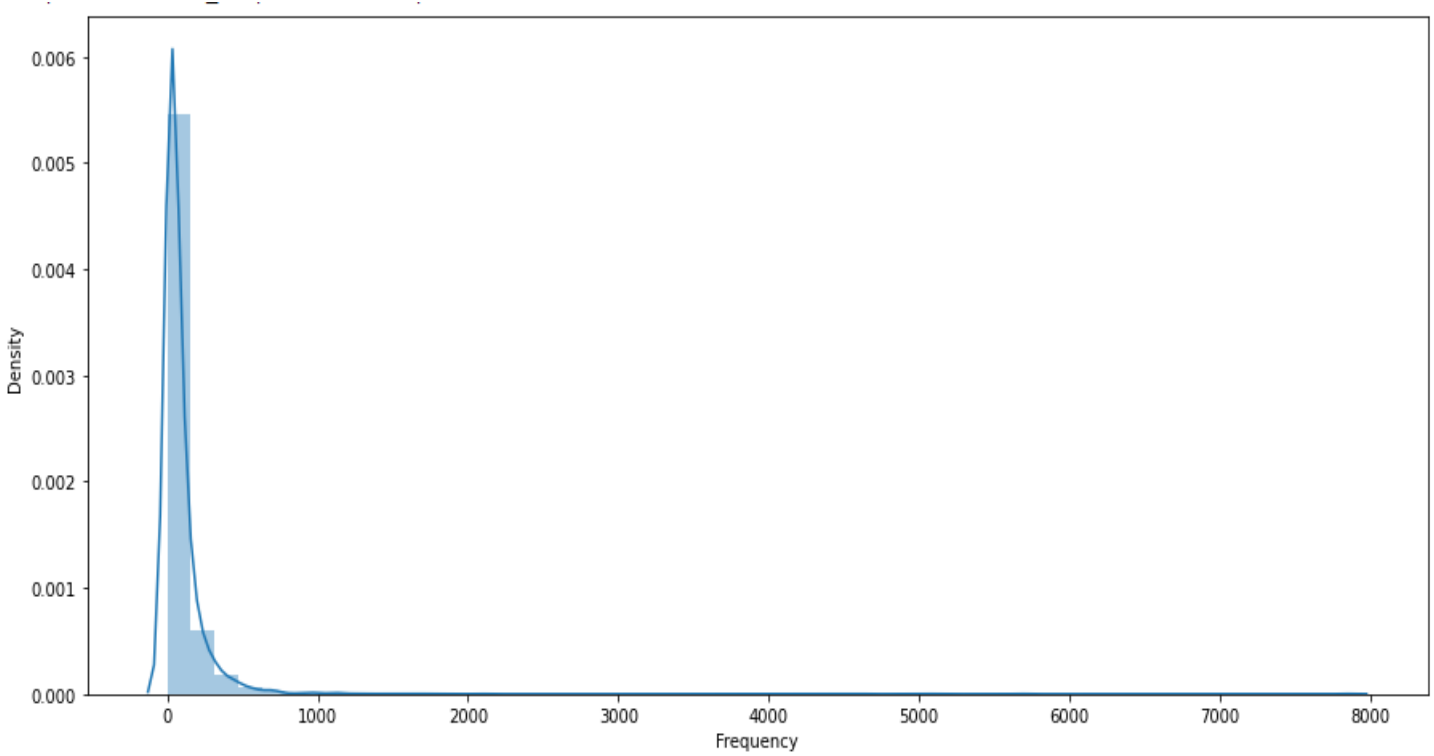
CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore
12346.0	325	1	77183.60	4	4	1	441	9
12347.0	2	182	4310.00	1	1	1	111	3
12348.0	75	31	1797.24	3	3	1	331	7
12349.0	18	73	1757.55	2	2	1	221	5
12350.0	310	17	334.40	4	4	3	443	11

RFM Score

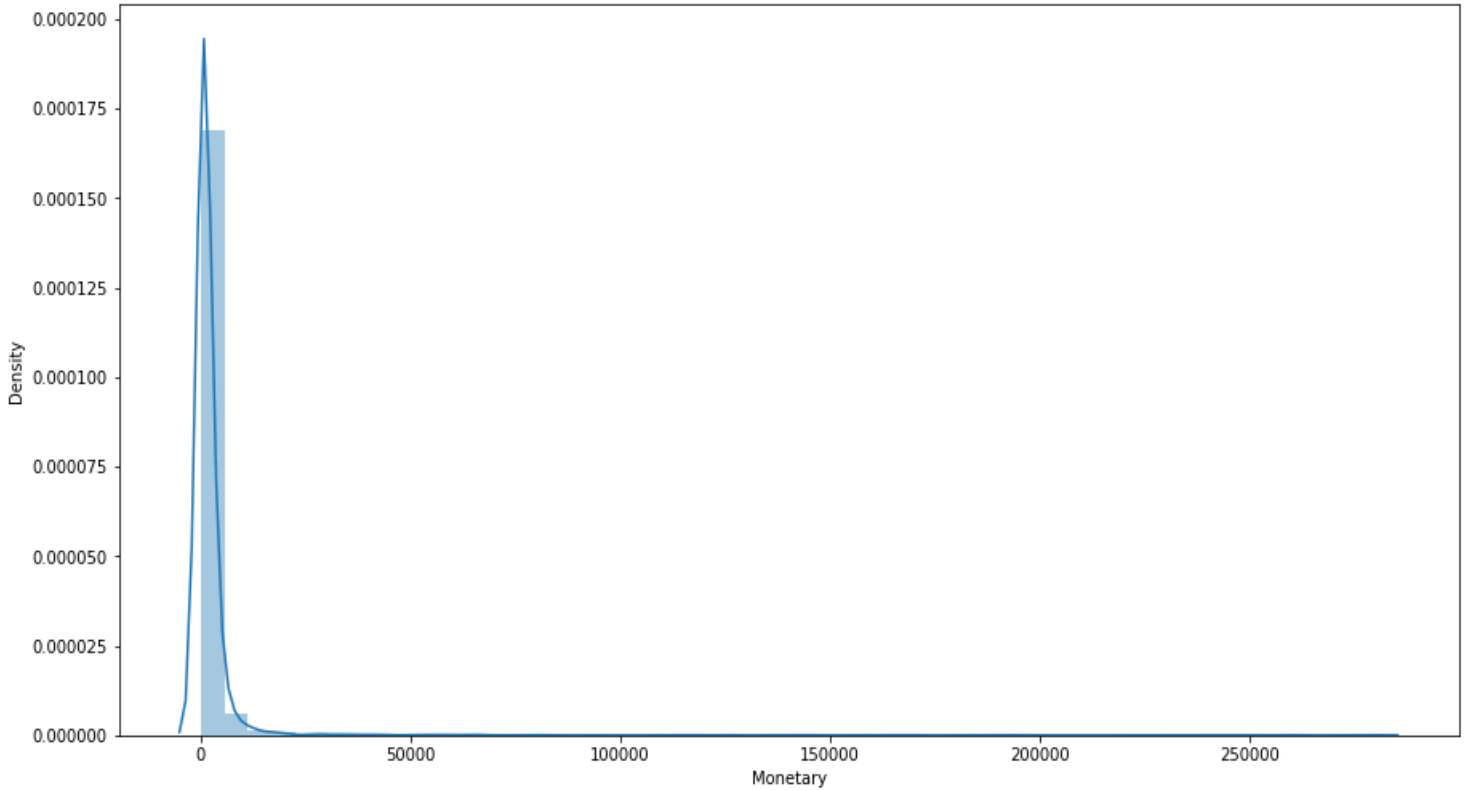
Recency Distribution



Frequency Distribution



Monetary Distribution

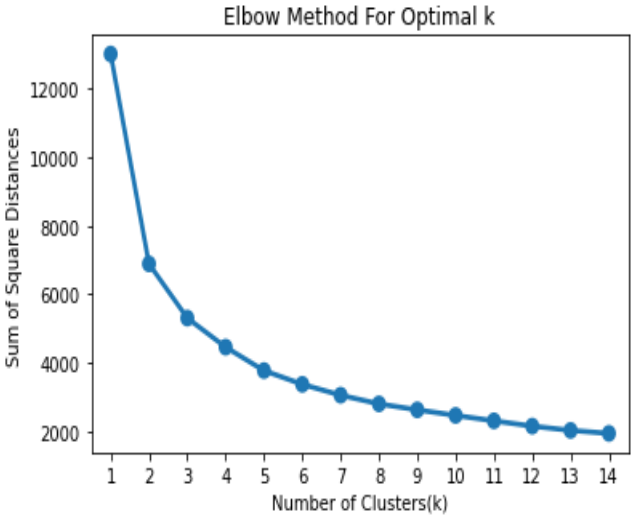


Model Implementation

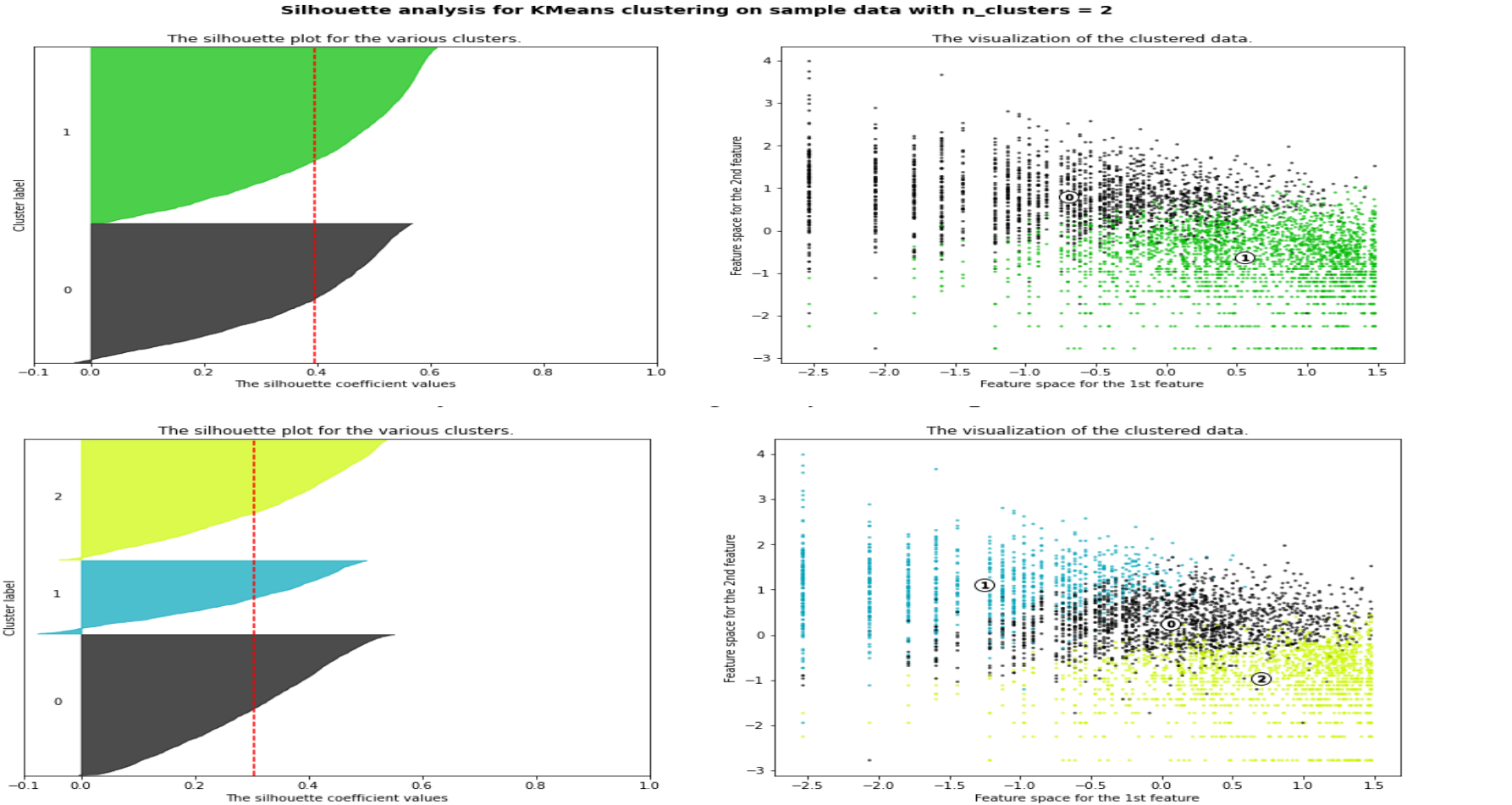
→K-Mean Clustering

Evaluation method

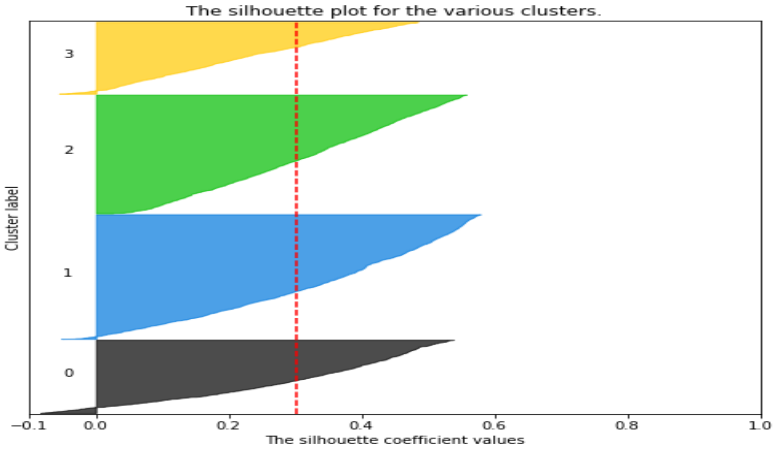
```
For n_clusters = 2 The average silhouette_score is : 0.3956478042246982
For n_clusters = 3 The average silhouette_score is : 0.3049826724447913
For n_clusters = 4 The average silhouette_score is : 0.30279724233096916
For n_clusters = 5 The average silhouette_score is : 0.2785519277480847
For n_clusters = 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
For n_clusters = 15 The average silhouette_score is : 0.2587546253386377
```

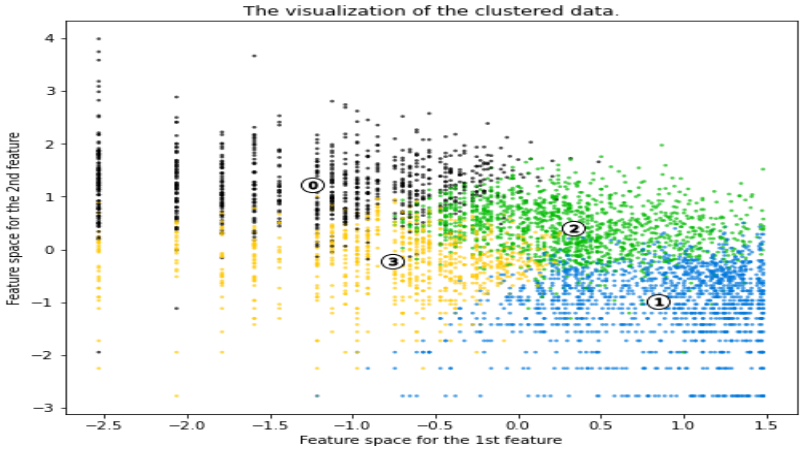


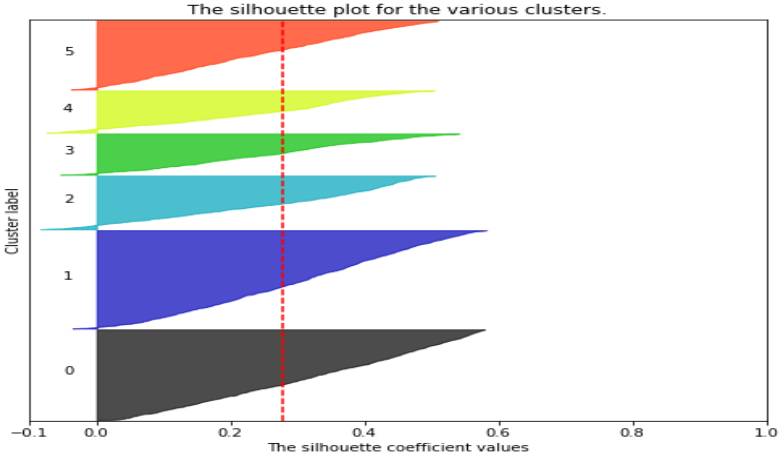
Silhouette Visualization

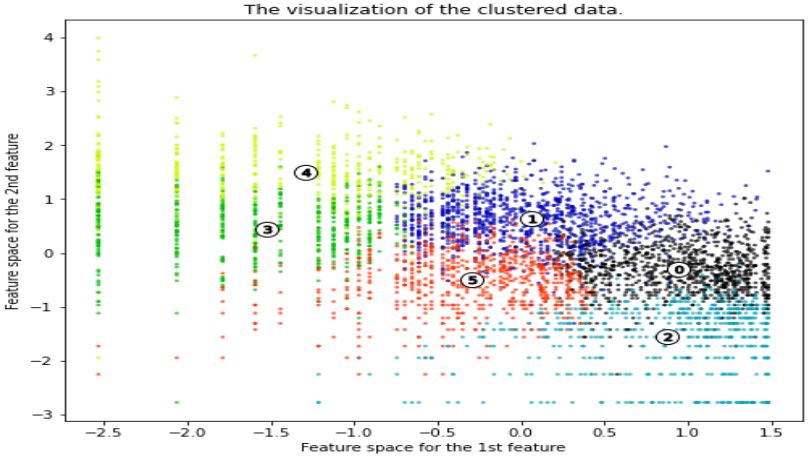


Silhouette Visualization

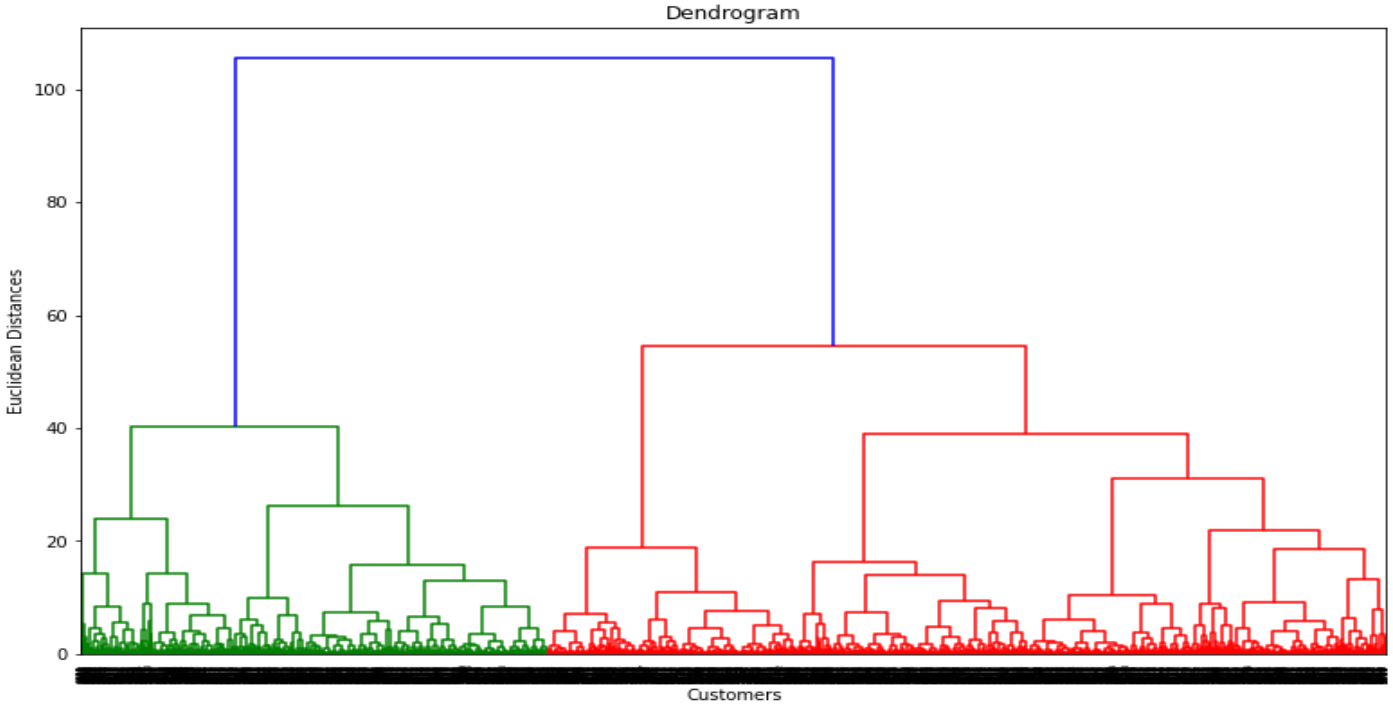






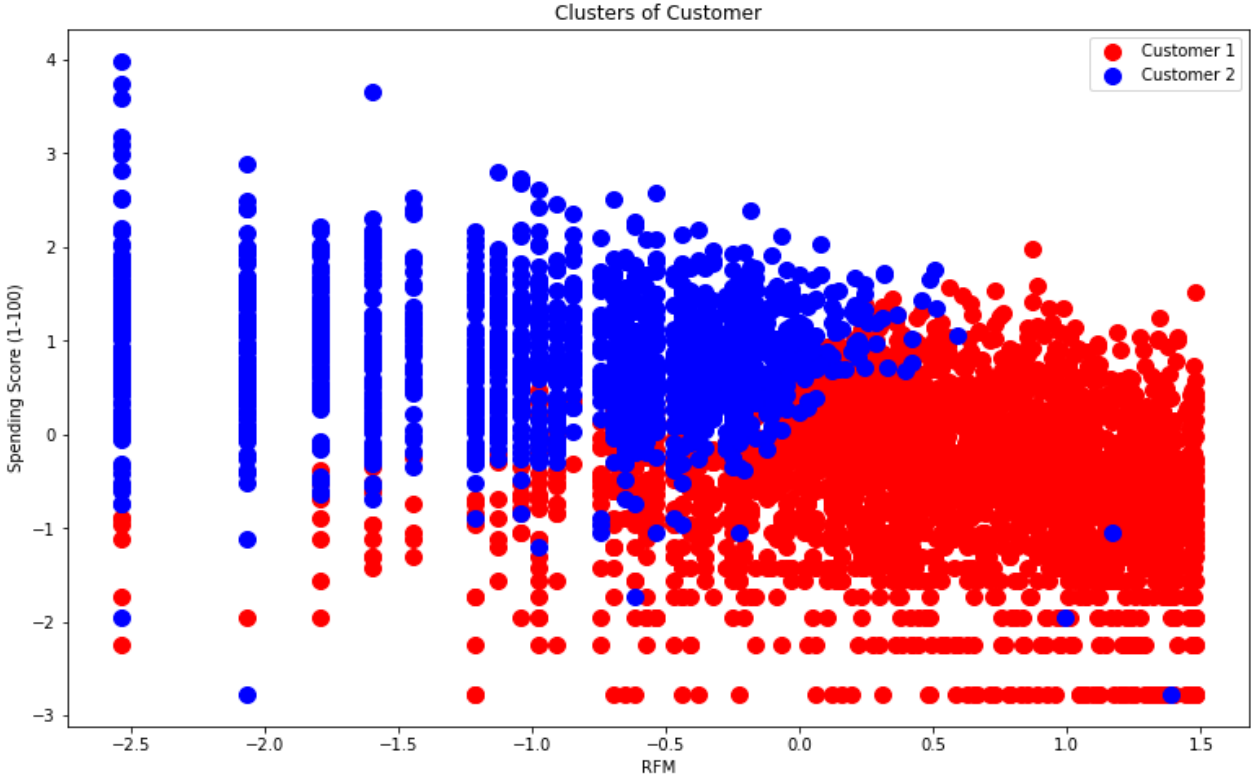


Hierarchical clustering method

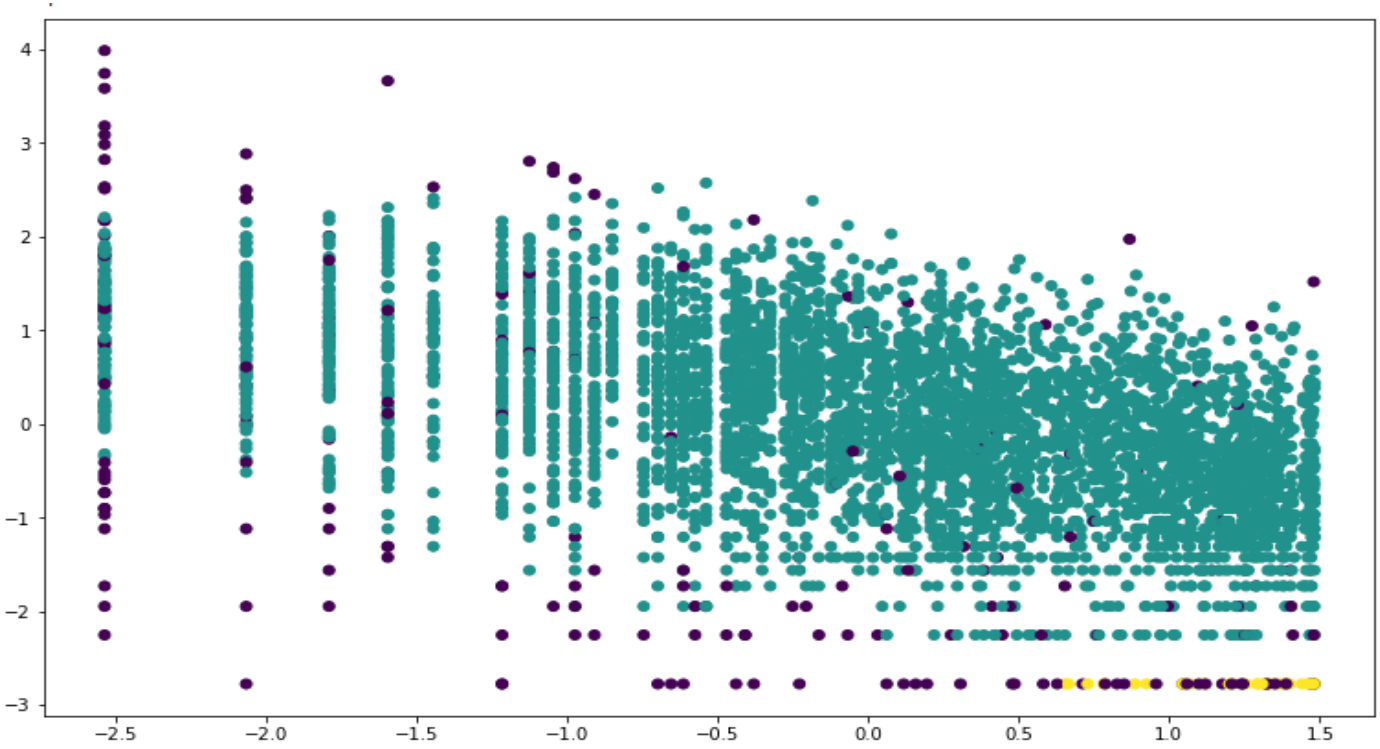


optimum number of cluster are 2

Agglomerative Clustering



DBSCAN Model



Optimum cluster are 3

Summary

	Model_Name	Data	Optimal_Number_of_cluster
0	K-Means with silhouette_score	RFM	2
1	K-Means with Elbow methos	RFM	2
2	Hierarchical clustering	RFM	2
3	DBSCAN	RFM	3

- We can conclude from this that, the optimum number of cluster’s are 2

Challenges

- Large Dataset to handle
- Need to analyze lot of variable
- Null value handling
- Feature engineering
- Selecting Optimum number of cluster
- Deciding the flow of the presentation

Conclusion

Although we didn't obtain two clearly separated clusters, we were able to build a model that can classify new customers into "low value" and "high value" groups. Generally, if a customer only transacted with us a few times, they needed to be at least in the top 50th percentile in monetary spending to be considered a "high value customer".

- Most of the customers are from UK i.e. more than 350000
- Most purchased product is 'WHITE HANGING HEART T-LIGHT HOLDER' (quantity=2028)
- Most people buy in the range of 1 to 10 pound
- Most people buy around 10 units
- Most people buy on Thursday
- Most people buy on November month
- Most people buy the product in the afternoon i.e. around 12'0 clock

Q&A

Thank you