Capstone Project - 4

Online Retail Customer Segmentation

Submitted by

Kousik Dutta

Contents

Abstract

- 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.

- The groups should be homogeneous within themselves and should also be heterogeneous to each other. The overall aim of this process is to identify high-value customer base i.e. customers that have the highest growth potential or are the most profitable.

Problem Statement

- The main goal is to identify major customer segments like customers that are most profitable and the ones who churned out to prevent further loss of customer by redefining company policies.

Data Summary

Categorical Type	• Invoice No • Stock Code • Description • Country • CustomerID
Numerical Type	• Quantity • Unit Price
Datetime Type	• Invoice Date

Data Summary

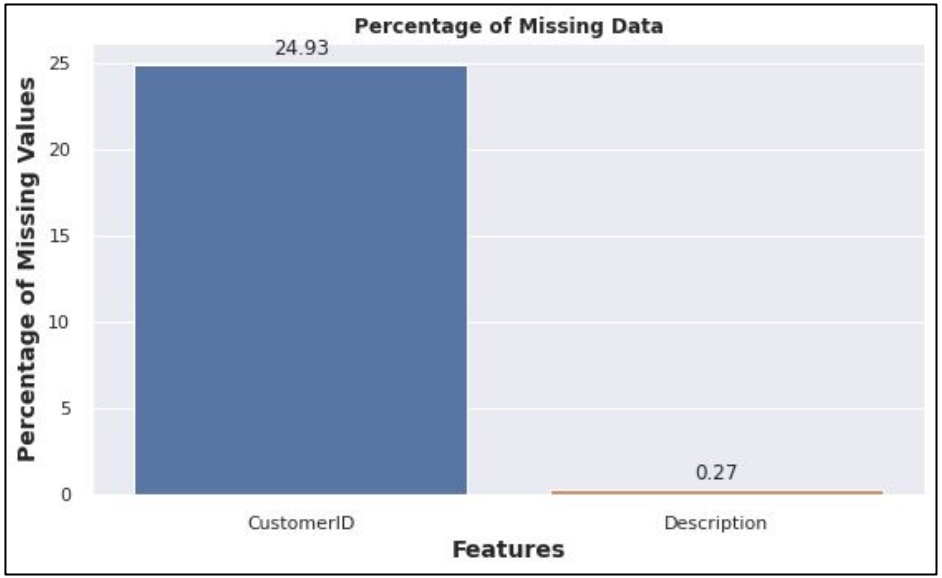
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365.0	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6.0	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365.0	71053.0	WHITE METAL LANTERN	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365.0	84406B	CREAM CUPID HEARTS COAT HANGER	8.0	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365.0	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365.0	84029E	RED WOOLLY HOTTIE WHITE HEART.	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

- This Dataset contains 541909 rows and 8 columns.
- A transactional data set with transactions occurring between 1st December 2010 and 9th December 2011 for a UK-based online retailer.

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  float64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(3), object(4)
memory usage: 33.1+ MB
```

Data Preprocessing

Null Values



After Handling Missing Values

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

Data Preprocessing

Duplicate Rows

- Total Number of Duplicate Rows : 5225

Dropping Invoice No starting with 'C' that represents cancellation.

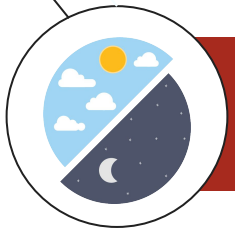
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27	14527	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4	15311	United Kingdom
235	C536391	22556.0	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1	17548	United Kingdom
236	C536391	21984.0	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0	17548	United Kingdom
237	C536391	21983.0	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0	17548	United Kingdom
...	...	...	...	...	...	...	...	...
540449	C581490	23144.0	ZINC T-LIGHT HOLDER STARS SMALL	-11	2011-12-09 09:57:00	0	14397	United Kingdom
541541	C581499	M	Manual	-1	2011-12-09 10:28:00	224	15498	United Kingdom
541715	C581568	21258.0	VICTORIAN SEWING BOX LARGE	-5	2011-12-09 11:57:00	10	15311	United Kingdom
541716	C581569	84978.0	HANGING HEART JAR T-LIGHT HOLDER	-1	2011-12-09 11:58:00	1	17315	United Kingdom
541717	C581569	20979.0	36 PENCILS TUBE RED RETROSPOT	-5	2011-12-09 11:58:00	1	17315	United Kingdom

8872 rows × 8 columns

Feature Engineering

Extracting Day, Month, Year and Hour from Invoice Date column.

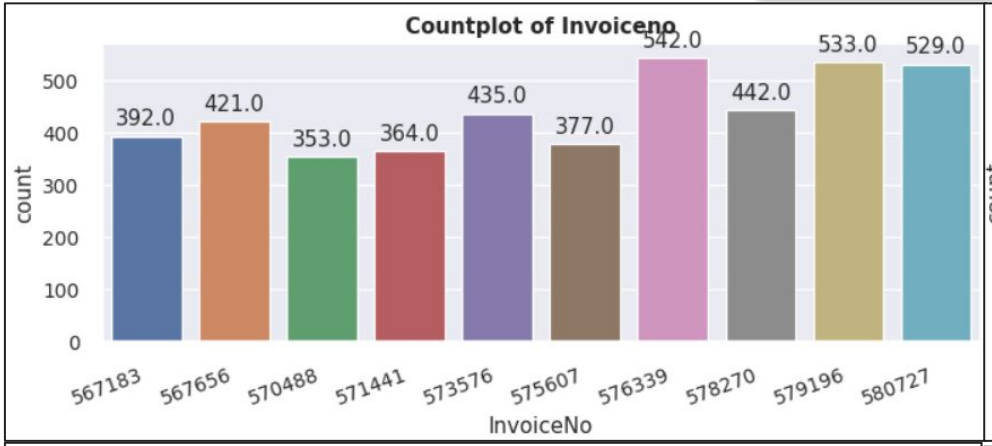
Added Feature 'Total Amount' by multiplying values from the Quantity and Unit Price column.

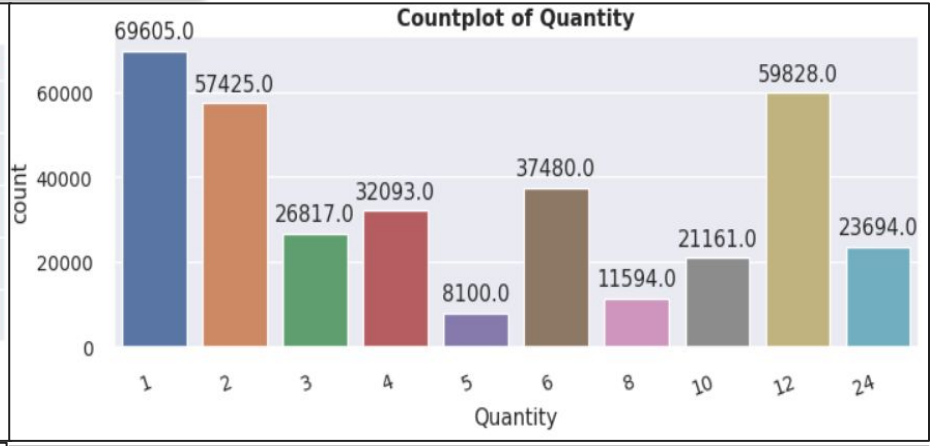


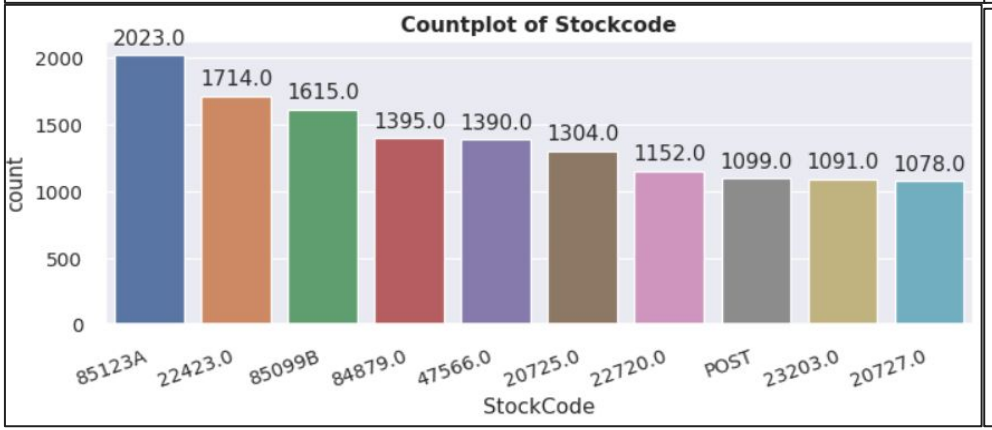
Added feature 'Time_Type' based on hours to define whether its Morning, Afternoon or Evening.

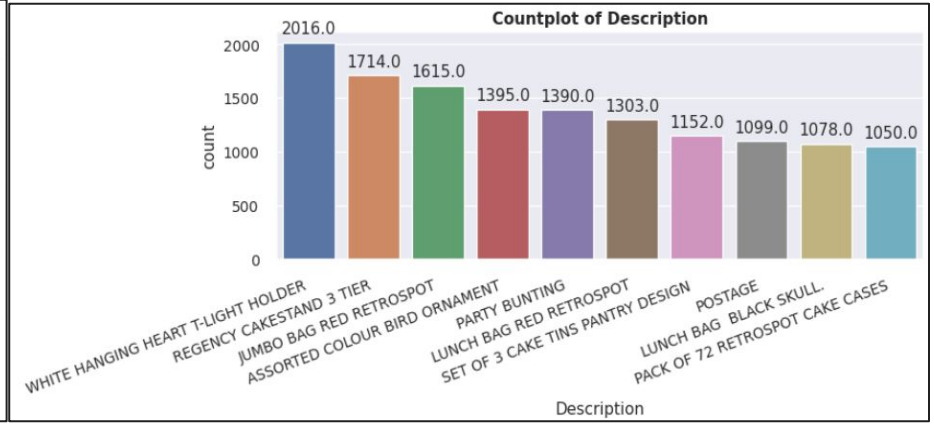
Exploratory Data Analysis

Univariate Analysis



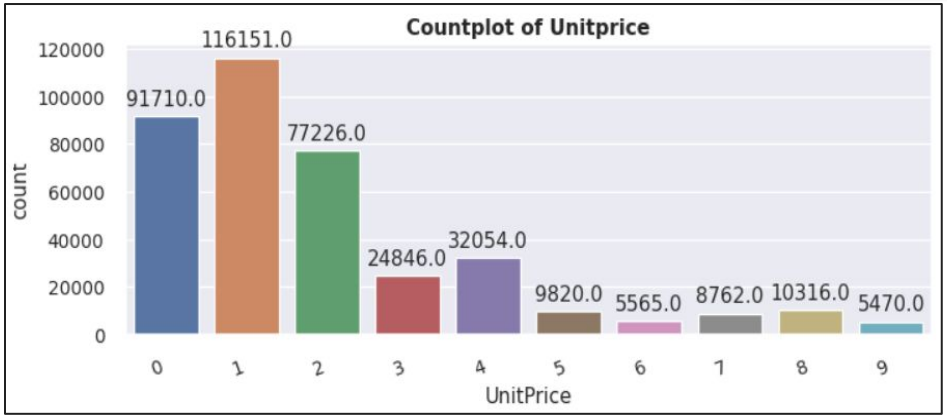


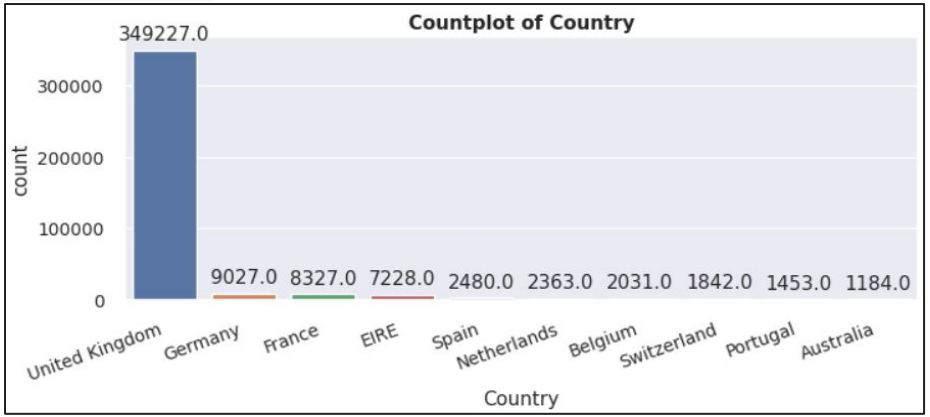


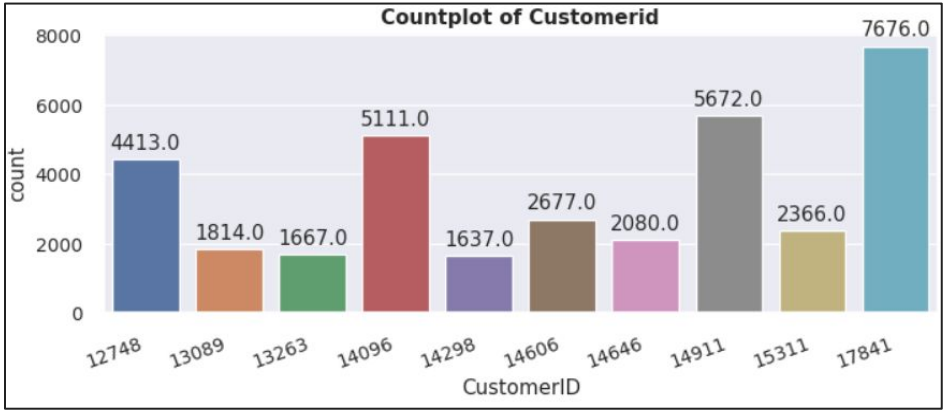


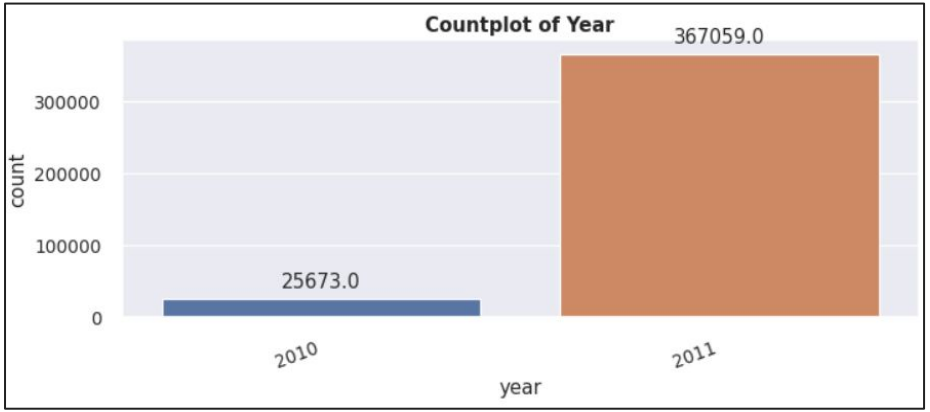
Exploratory Data Analysis

Univariate Analysis



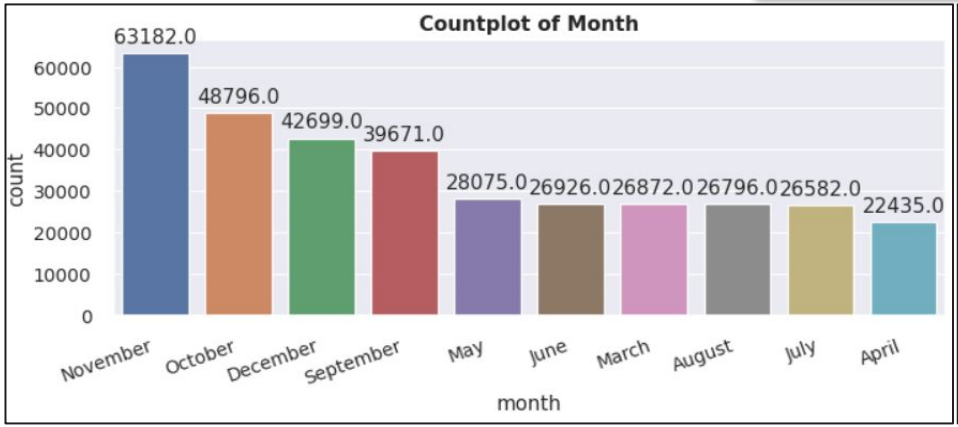


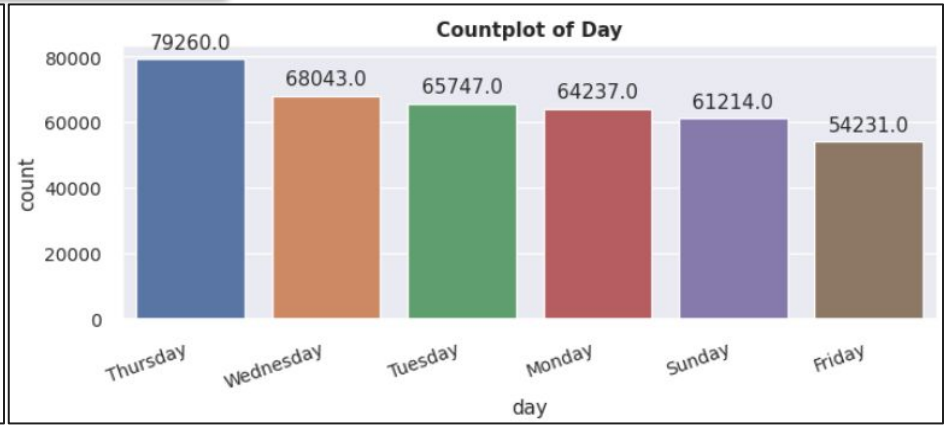


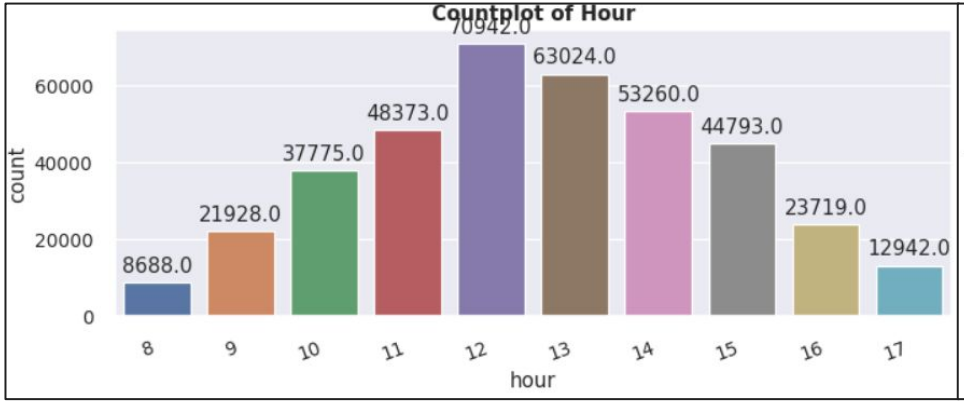


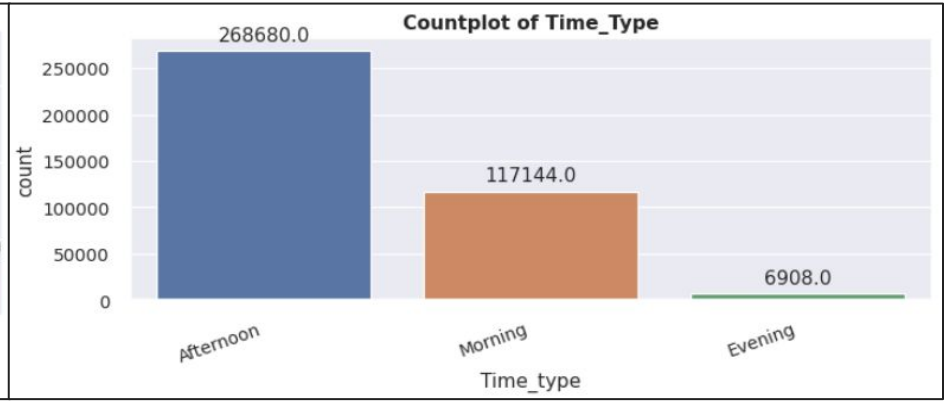
Exploratory Data Analysis

Univariate Analysis









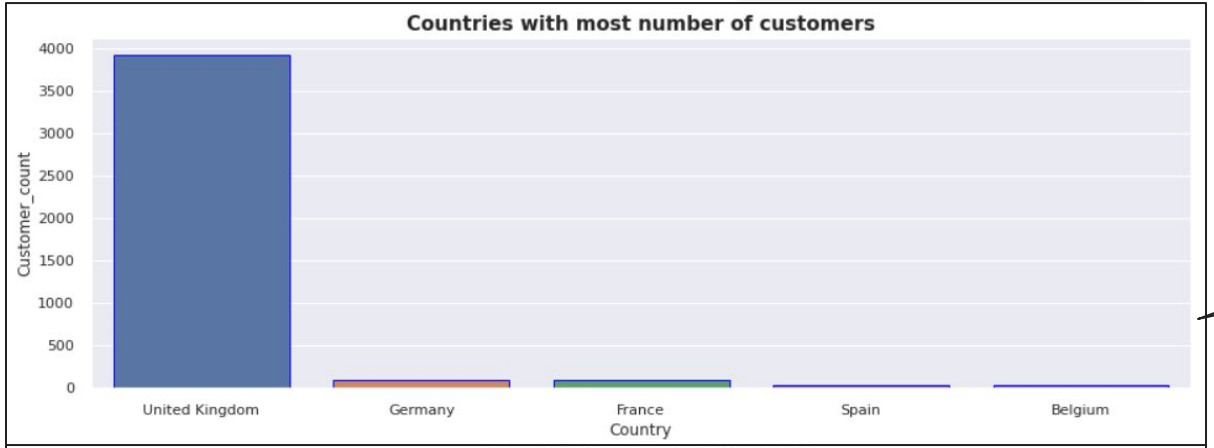
Exploratory Data Analysis

Observations Drawn From The Univariate Analysis :

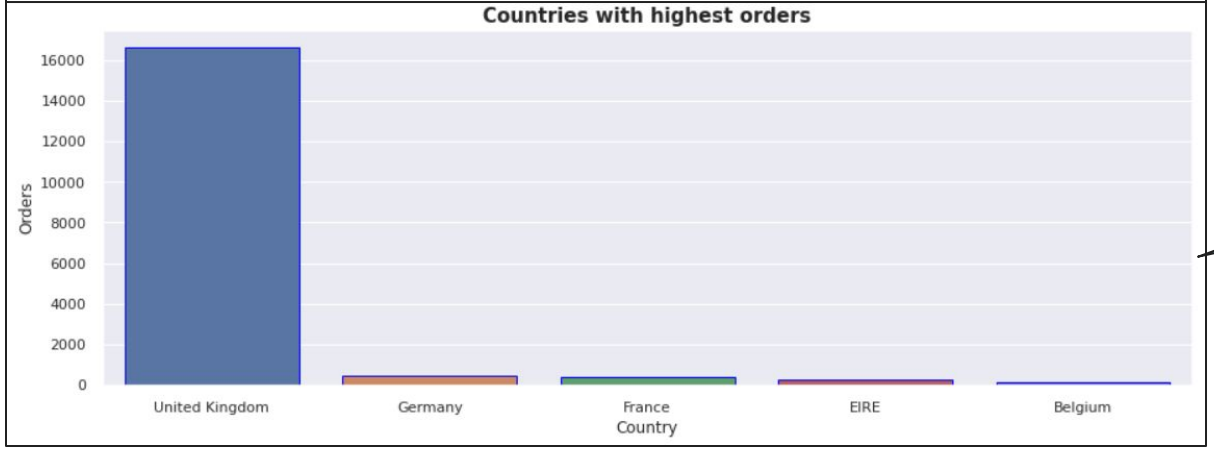
- 'WHITE HANGING HEART T-LIGHT HOLDER' (Stock Code - 85123A), 'REGENCY CAKESTAND 3 TIER' (Stock Code - 22423) are the top 2 most ordered products.
- Most customers are from 'United Kingdom' also considerable number of customers are also from 'Germany' , 'France', 'Eire' and 'Spain'.
- Most of the customer have purchased items in the month of 'October', 'November', 'December' the reason may be most of the festivals are in these months. Less number of customers have purchased the items in the month of 'January', 'February', 'April'.
- There are no orders placed on 'Saturdays', the reason maybe all retail shop stay closed on this day.
- Most of the customers have purchased the items in Afternoon, moderate number of customer have purchased the items in Morning and least number of customers in Evening.

Exploratory Data Analysis

Bivariate Analysis



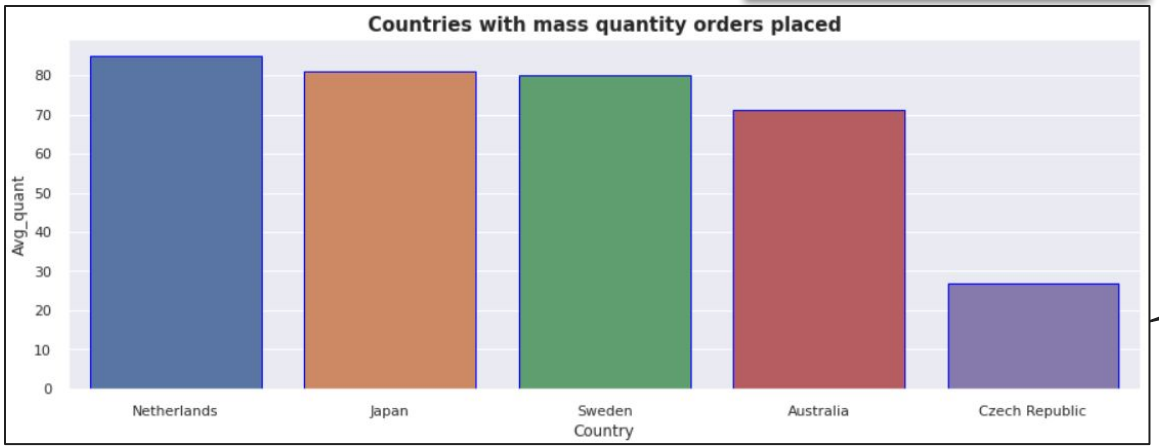
'United Kingdom' has most number of customers than any other countries.



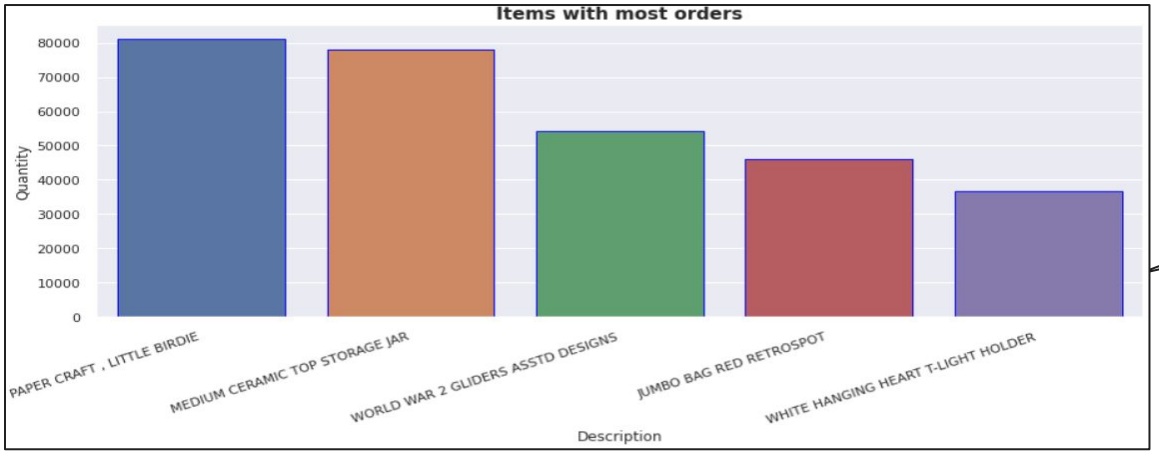
'United kingdom' also topped with most order placed compared to other countries.

Exploratory Data Analysis

Bivariate Analysis



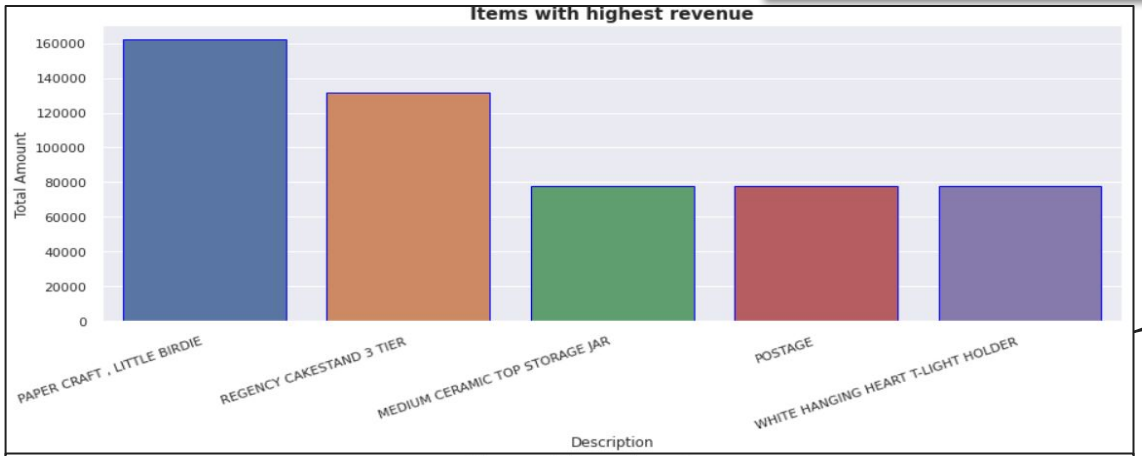
Orders with mass quantity placed by the customer from Netherlands.



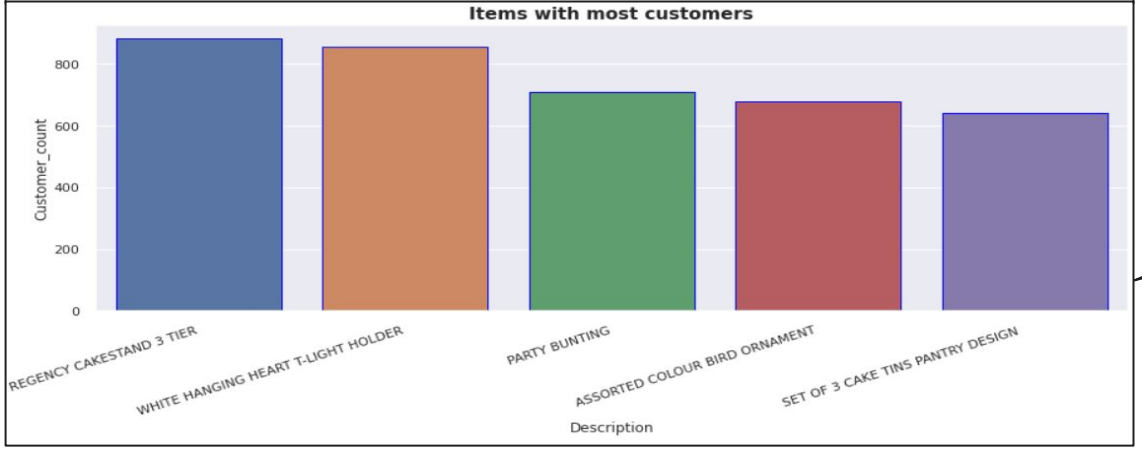
'PAPER CRAFT , LITTLE BIRDIE' , 'MEDIUM CERAMIC TOP STORAGE JAR' these are the top 2 items with most purchased in quantity.

Exploratory Data Analysis

Bivariate Analysis



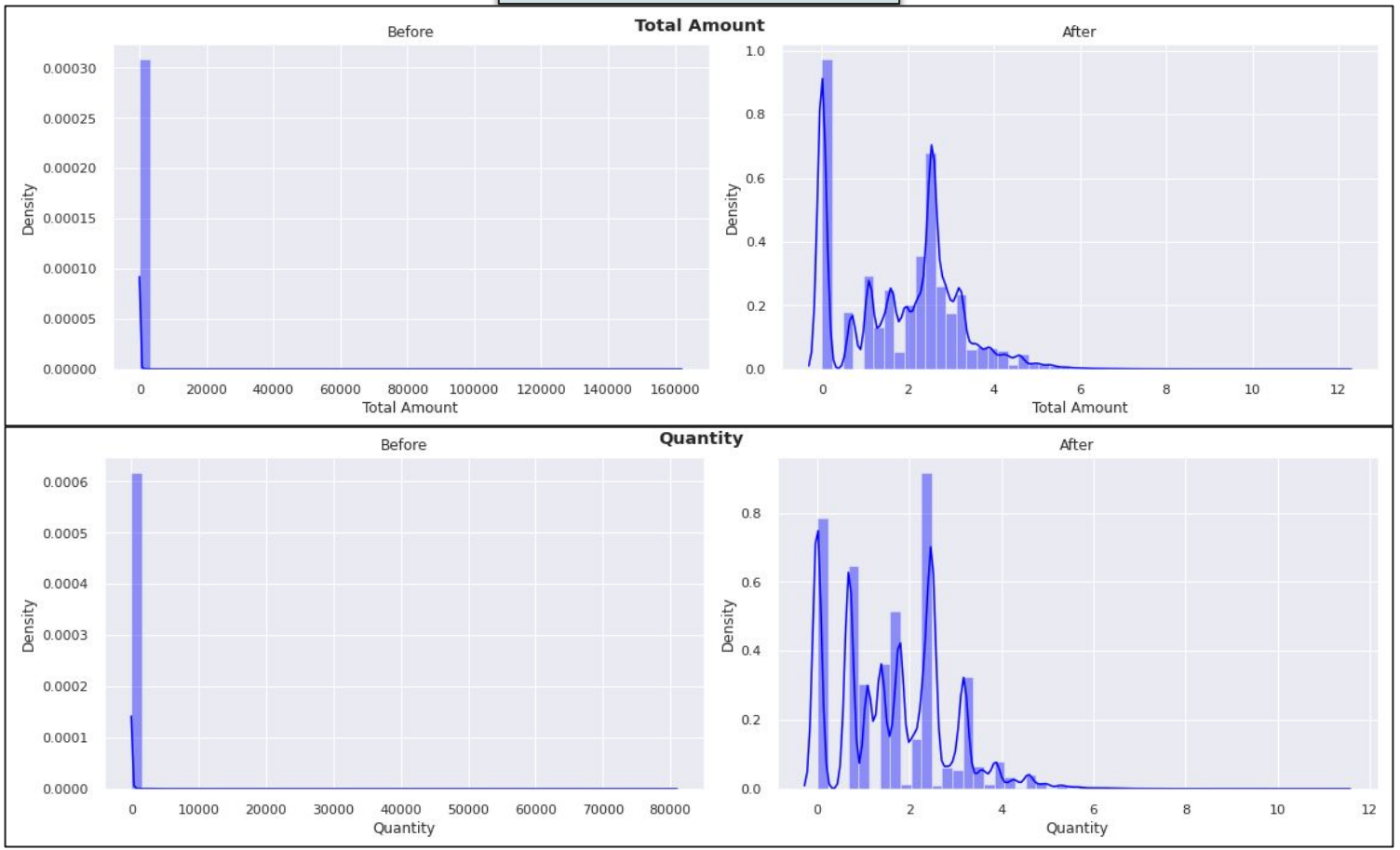
'PAPER CRAFT , LITTLE BIRDIE' product has made highest revenue.



'REGENCY CAKESTAND 3 TIER' is the choice of most of the customer.

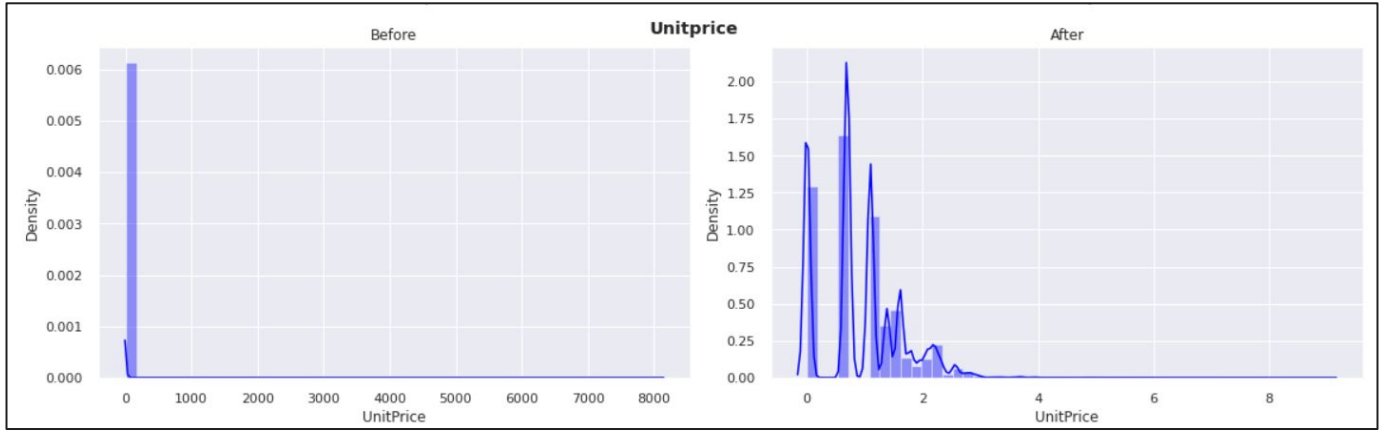
Exploratory Data Analysis

Visualizing Distributions



Exploratory Data Analysis

Visualizing Distributions



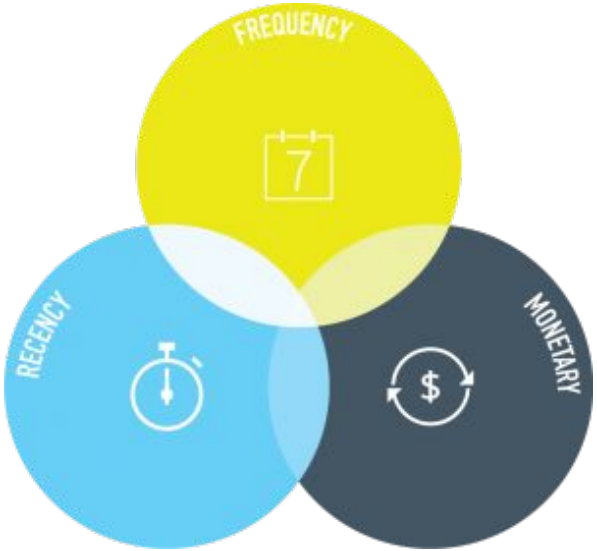
- After applying log transformation now the distribution plot looks comparatively better than being skewed.

Model Building

What Is RFM Analysis ?

RFM Model which stands for Recency, Frequency and Monetary is one of such steps in which we determine the followings.

- Recency : How recently did the customer visit our website or how recently did the customer purchase.
- Frequency : How often do they visit or how often do they purchase.
- Monetary : How much revenue we get from their visit or how much do they spend when they purchase.



What It Is Needed ?

- RFM analysis is a marketing framework that is used to understand and analyze customer behaviour based on the above three factors Recency, Frequency and Monetary.
- The RFM analysis will help the businesses to segment their customer base into different homogenous groups so that they can engage with each group with different targeted marketing strategies.

Model Building

RFM Model Analysis

RFM Dataset

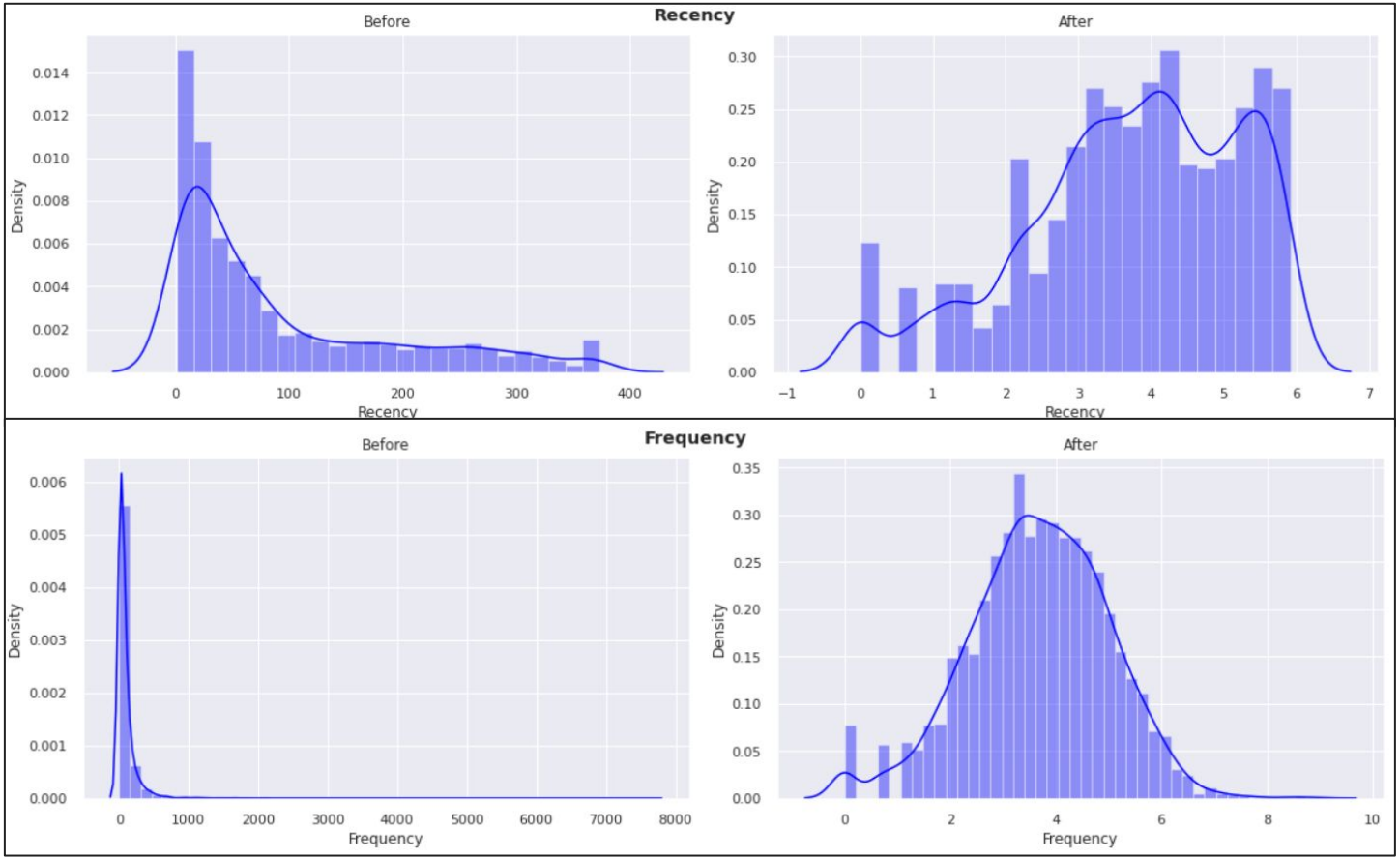
CustomerID	Recency	Frequency	Monetary	R	F	M	RFM_Group	RFM_Score
12346	325	1	74215	4	4	1	441	9
12347	2	182	3012	1	1	1	111	3
12348	75	31	944	3	3	2	332	8
12349	18	73	1404	2	2	1	221	5
12350	310	17	244	4	4	3	443	11

Conditions For Best Customers

- Recency - Less
- Frequency - More
- Monetary - More

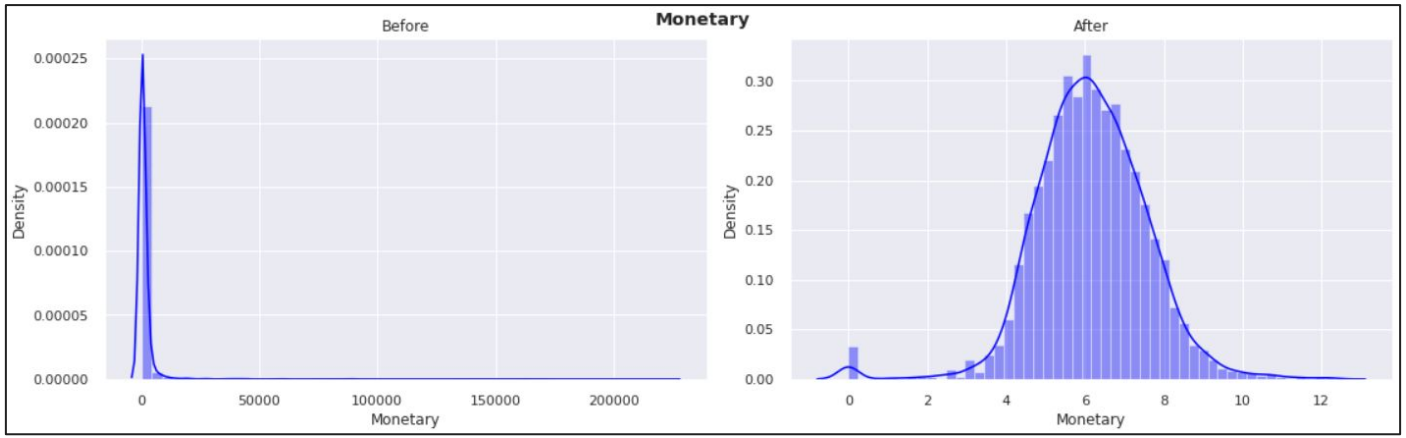
Model Building

RFM Model Analysis



Model Building

RFM Model Analysis



- Earlier the distributions of Recency, Frequency and Monetary columns were positively skewed but after applying log transformation the distributions appear to be symmetrical and normally distributed.

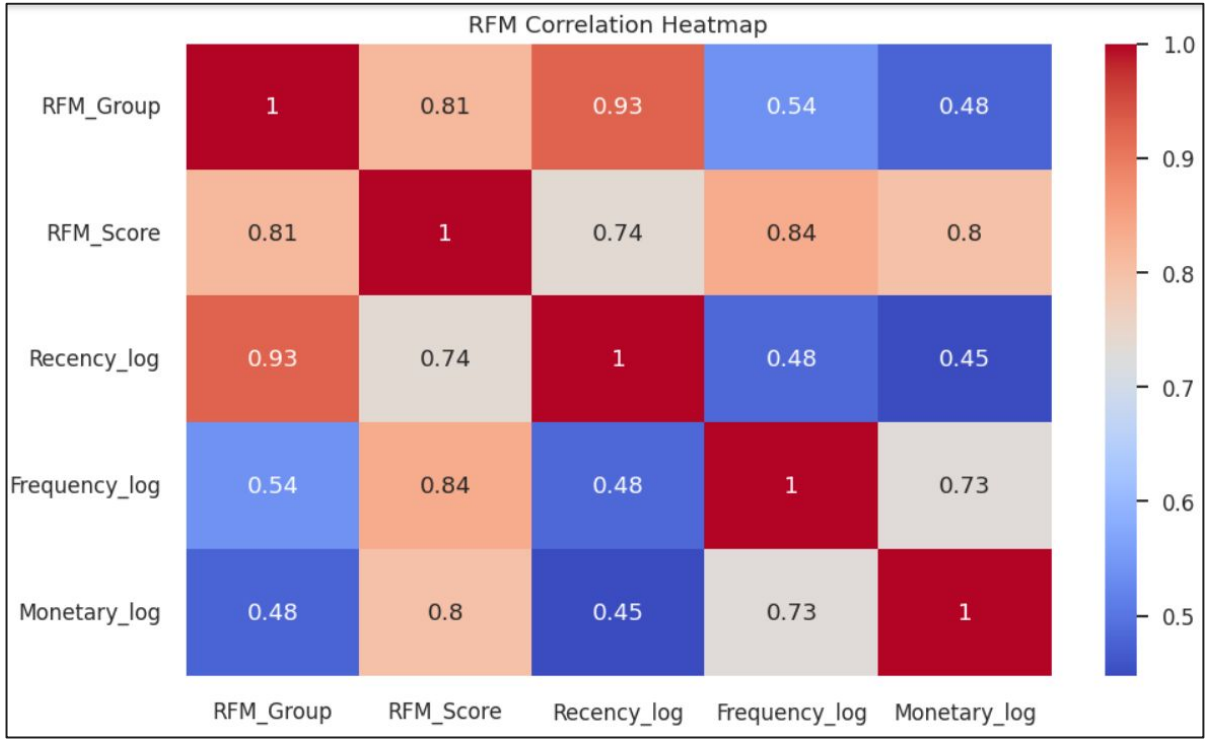
Model Building

RFM Model Analysis

RFM CORRELATION HEATMAP

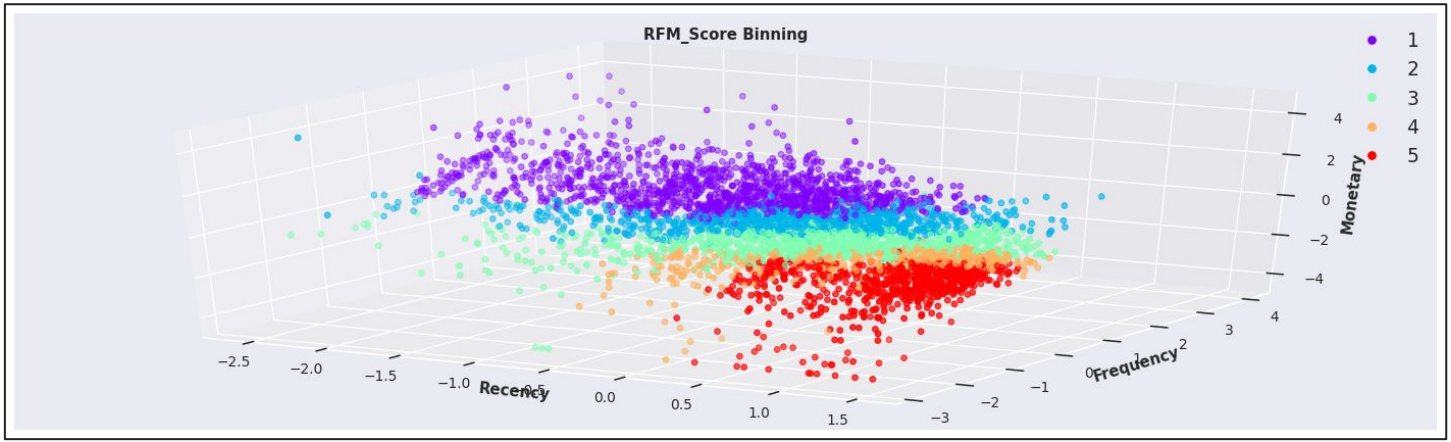
- It is clear from above that 'Recency' is highly correlated with RFM_Group value , whereas Frequency and Monetary is moderately correlated with RFM_Group.

- Also RFM_Score is equally correlated with 'Recency', 'Frequency' and 'Monetary'.



Model Building

RFM Score Binning

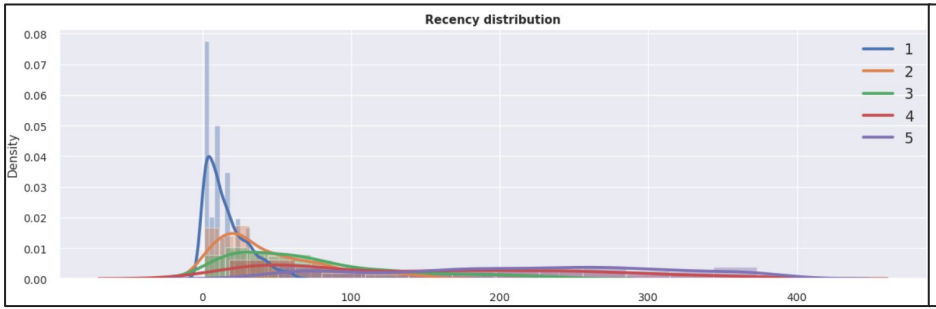


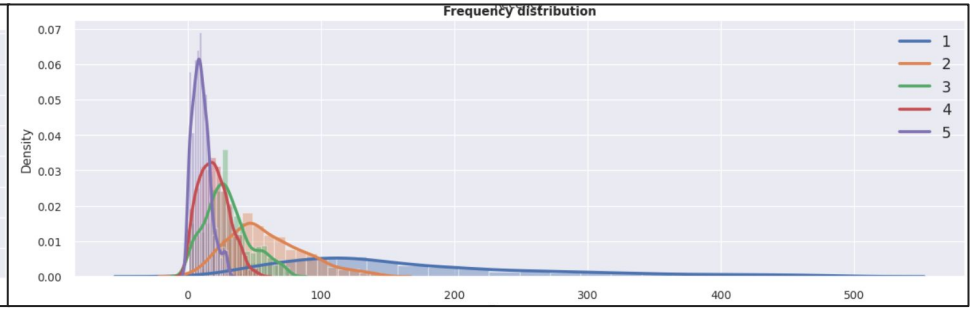
Displaying the mean, median of Recency, Frequency and Monetary for each group of customer

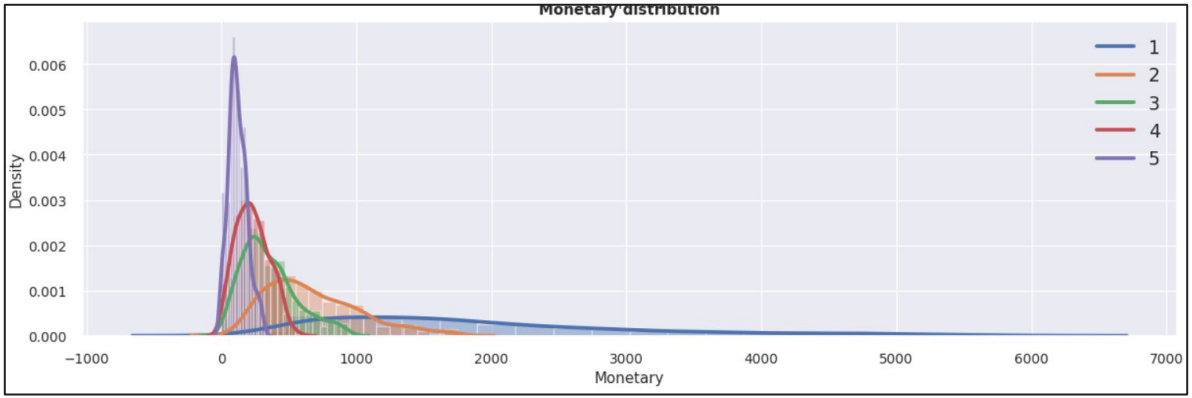
Binning RFM_Score	Recency mean	Recency median	Frequency mean	Frequency median	Monetary mean	Monetary median	count
1	19.497600	12.000000	227.270400	147.000000	3774.648000	1754.500000	1250
2	54.035068	36.000000	66.074661	56.000000	1000.200226	636.000000	884
3	89.537378	64.000000	33.509209	29.000000	583.310943	330.000000	923
4	150.107004	138.000000	21.295720	19.000000	247.933852	224.000000	514
5	218.054688	225.000000	10.884115	10.000000	131.434896	117.000000	768

Model Building

RFM Score Binning







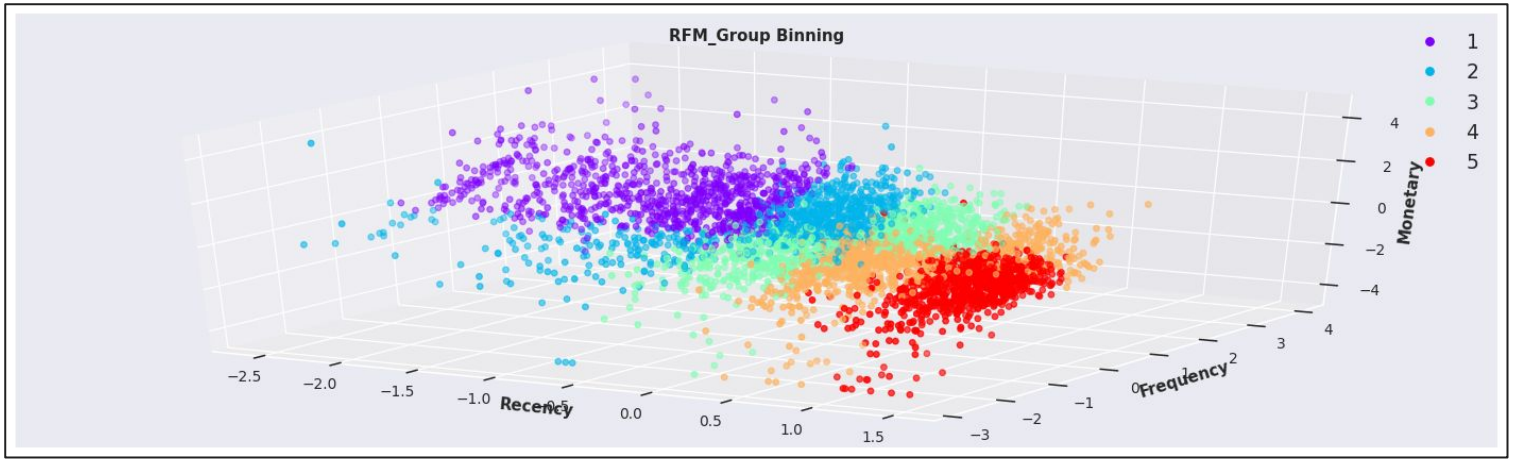
Model Building

RFM Score Binning

	Binning RFM_Score	Last visited	Purchase frequency	Amount spent	Customer_Type
0	1	4 to 27 days ago	Bought 100 to 252 times	Spend around 1078 to 3000 Pound Sterling	Best Customers
1	2	18 to 72 days ago	Bought 40 to 82 times	Spend around 421 to 953 Pound Sterling	Good Customers
2	3	31 to 119 days ago	Bought 19 to 42 times	Spend around 210 to 502 Pound Sterling	Average Customers
3	4	57 to 236 days ago	Bought 11 to 28 times	Spend around 141 to 329 Pound Sterling	Risky to Churn segment customers
4	5	149 to 289 days ago	Bought 5 to 15 times	Spend around 76 to 173 Pound Sterling	Churned Customers

Model Building

RFM Group Binning

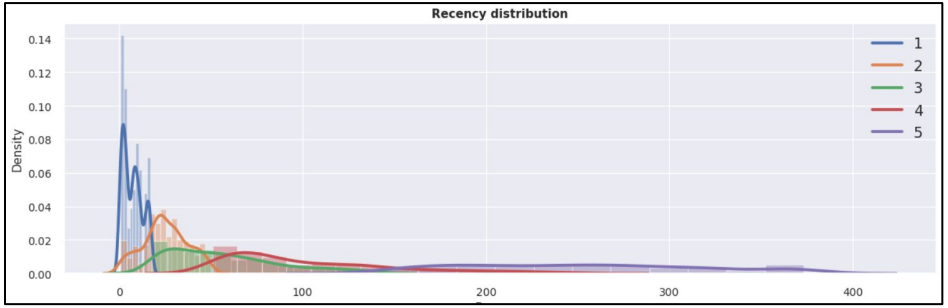


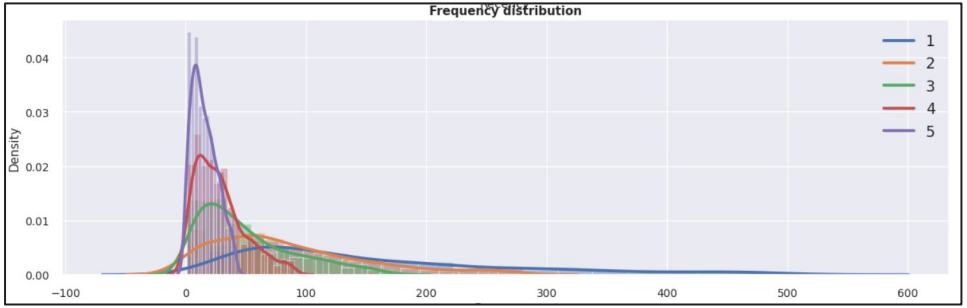
Displaying the mean, median of Recency, Frequency and Monetary for each group of customer

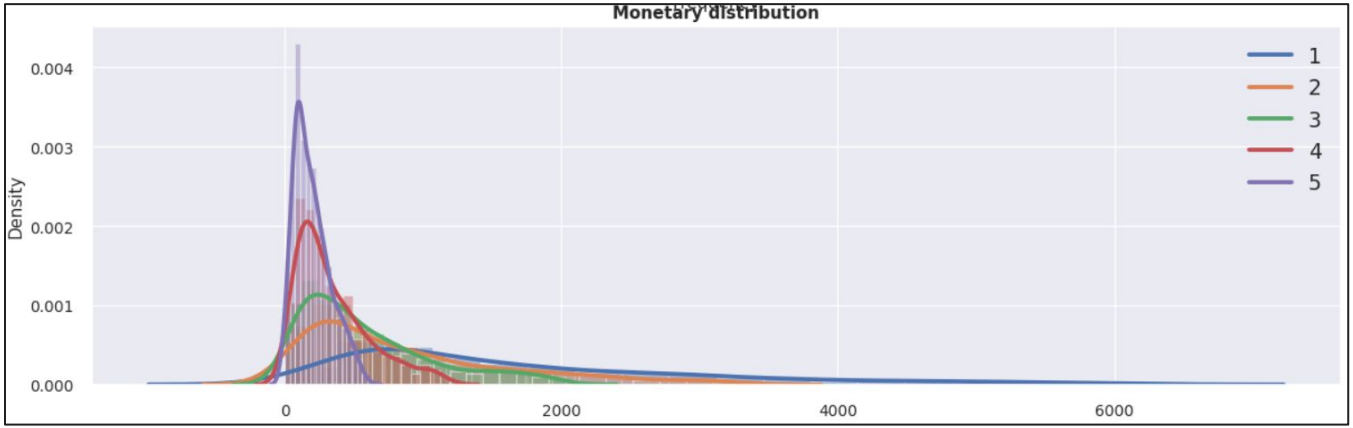
Binning RFM_Group	Recency mean	Recency median	Frequency mean	Frequency median	Monetary mean	Monetary median	count
1	7.493896	7.000000	226.724750	126.000000	3911.526082	1467.000000	901
2	25.327081	25.000000	108.466589	76.000000	1632.607268	703.000000	853
3	56.987074	51.000000	59.983549	42.000000	788.172738	488.000000	851
4	120.543779	93.000000	35.965438	26.500000	495.162442	294.000000	868
5	251.638568	247.500000	15.781755	14.000000	406.267898	179.000000	866

Model Building

RFM Group Binning







Model Building

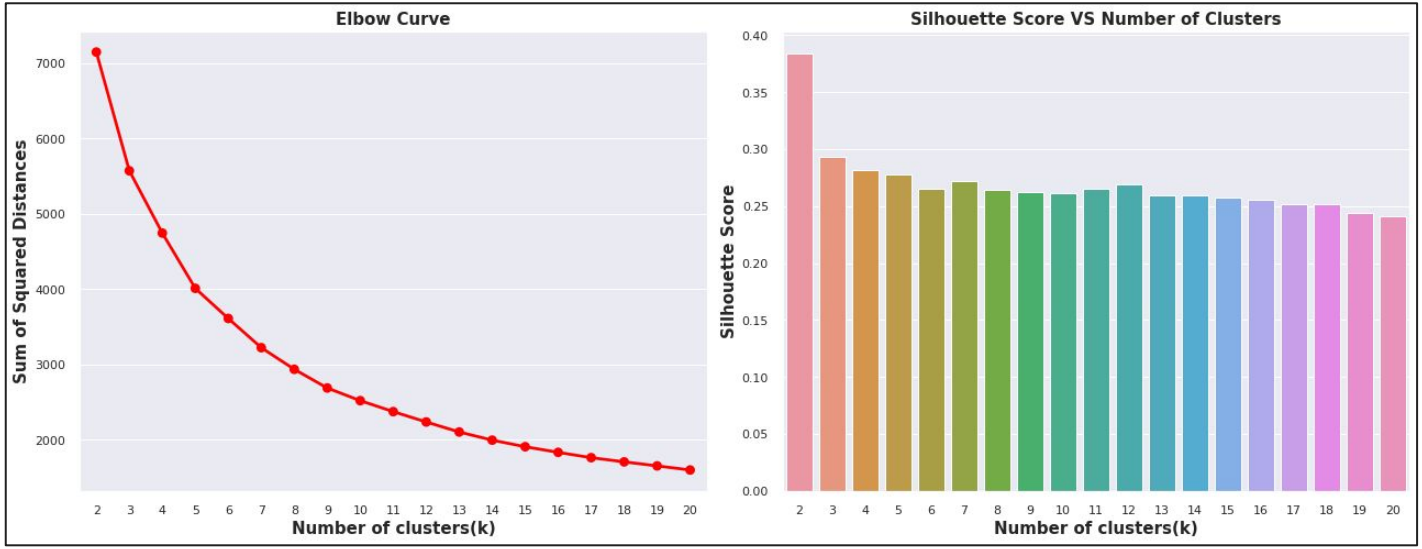
RFM Group Binning

	Binning RFM_Group	Last visited	Purchase frequency	Amount spent	Customer_Type
0	1	3 to 11 days ago	Bought 70 to 250 times	Spend around 756 to 2934 Pound Sterling	Best Customers
1	2	18 to 33 days ago	Bought 42 to 136 times	Spend around 319 to 1519 Pound Sterling	Good Customers
2	3	31 to 74 days ago	Bought 20 to 81 times	Spend around 231 to 959 Pound Sterling	Average Customers
3	4	70 to 143 days ago	Bought 13 to 44 times	Spend around 156 to 573 Pound Sterling	Risky to Churn segment customers
4	5	196 to 302 days ago	Bought 7 to 23 times	Spend around 98 to 299 Pound Sterling	Churned Customers

Model Building

K - Means Clustering

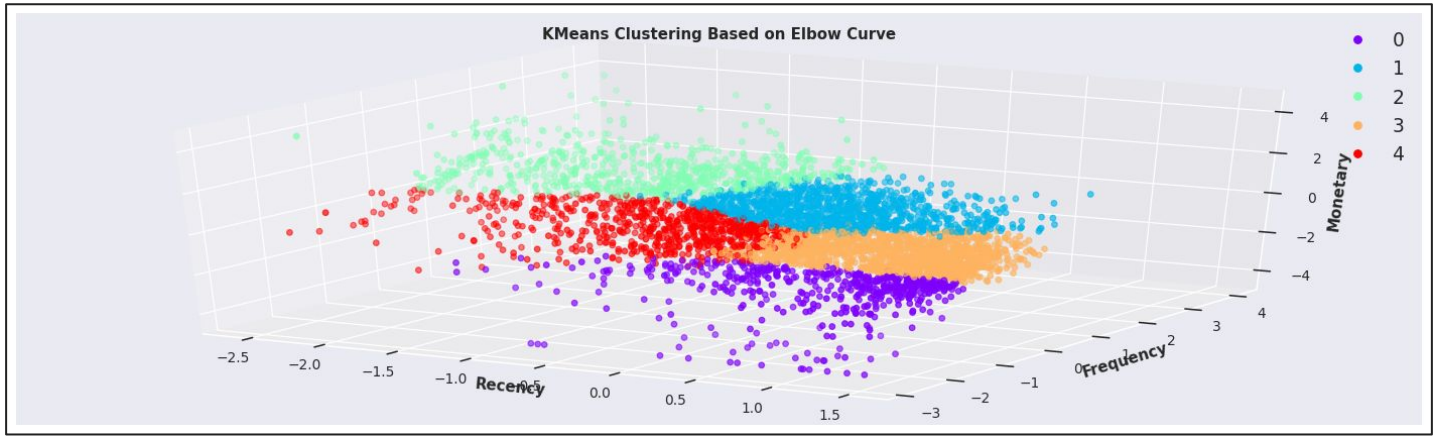
Finding The Optimal ‘k’ : Elbow Curve & Silhouette Score



- From the above Elbow Curve it seems that number of cluster as 5 is the most suitable.
- If we go by maximum silhouette score as the criteria for selecting optimal number of clusters then n_clusters can be chosen as 2.

Model Building

K - Means Clustering : 5 Clusters

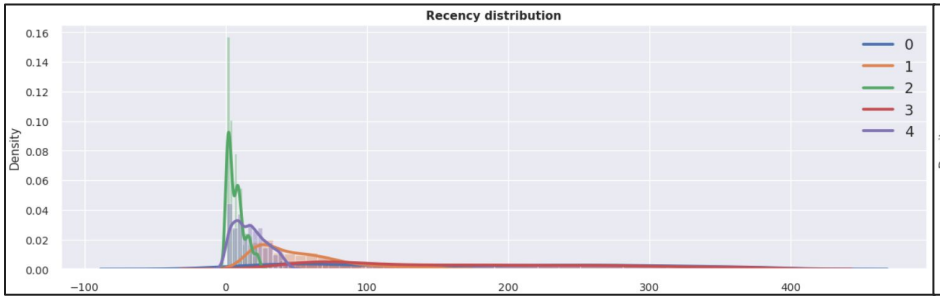


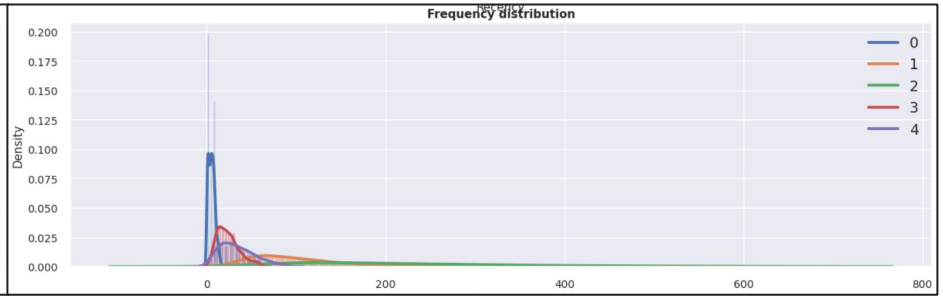
Displaying the mean, median of Recency, Frequency and Monetary for each group of customer

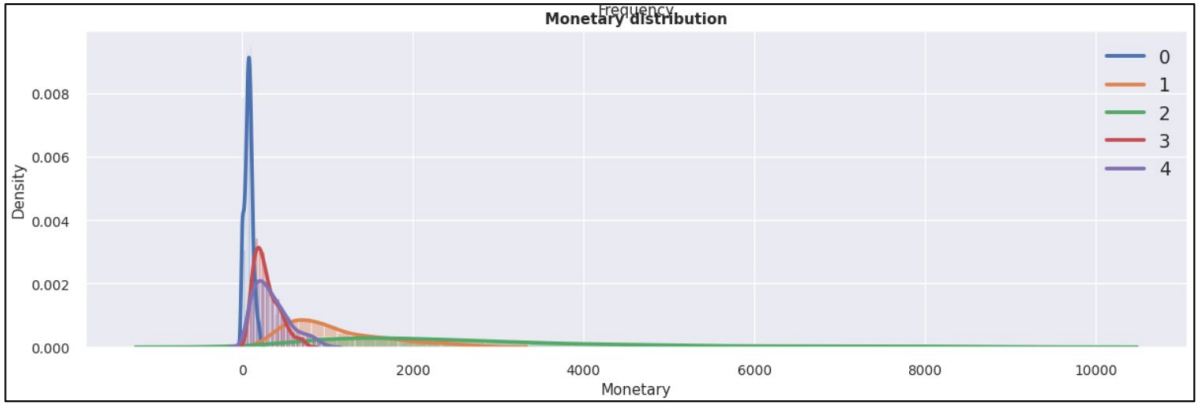
KMeans : 5 Clusters	Recency mean	Recency median	Frequency mean	Frequency median	Monetary mean	Monetary median	count
0	174.890173	173.000000	6.198459	6.000000	98.751445	82.000000	519
1	63.699732	47.000000	106.557641	90.000000	1332.581769	973.000000	1119
2	8.295588	7.000000	308.035294	205.000000	5914.725000	2515.000000	680
3	171.169797	157.000000	26.338028	23.000000	391.728482	260.500000	1278
4	17.441454	16.000000	36.545087	32.000000	409.434724	318.000000	743

Model Building

K - Means Clustering : 5 Clusters







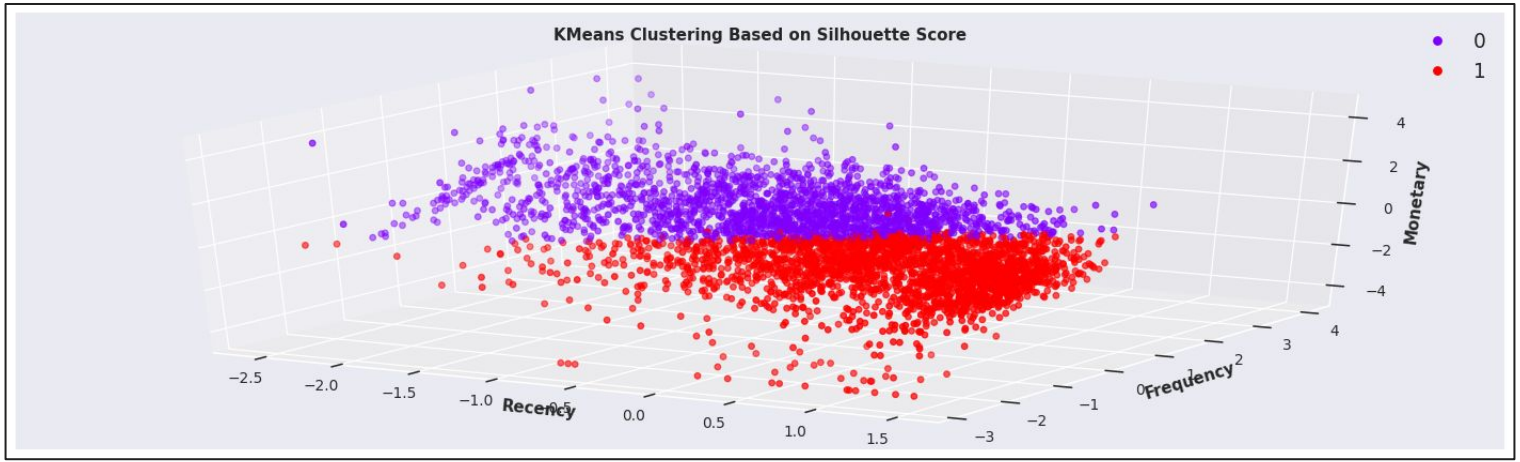
Model Building

K - Means Clustering : 5 Clusters

	KMeans : 5 Clusters	Last visited	Purchase frequency	Amount spent	Customer_Type
0	0	65 to 267 days ago	Bought 3 to 9 times	Spend around 52 to 118 Pound Sterling	Churned Customers
1	1	78 to 248 days ago	Bought 15 to 34 times	Spend around 177 to 416 Pound Sterling	Good Customers
2	2	2 to 11 days ago	Bought 125 to 338 times	Spend around 1467 to 4518 Pound Sterling	Best Customers
3	3	8 to 25 days ago	Bought 19 to 48 times	Spend around 193 to 513 Pound Sterling	Risky to Churn segment customers
4	4	28 to 75 days ago	Bought 63 to 134 times	Spend around 668 to 1588 Pound Sterling	Average Customers

Model Building

K - Means Clustering : 2 Clusters

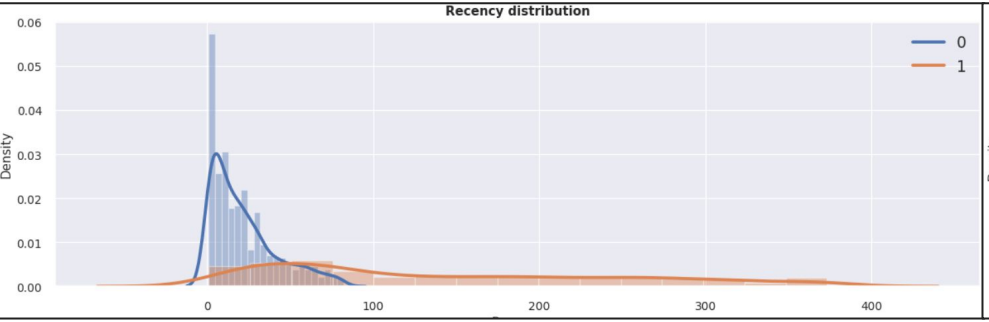


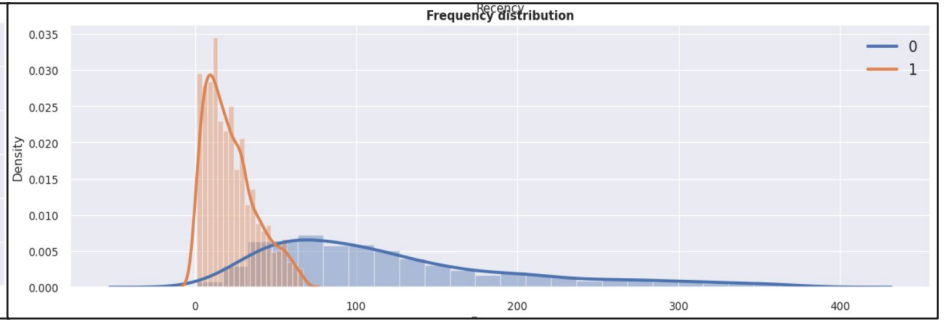
Displaying the mean, median of Recency, Frequency and Monetary for each group of customer

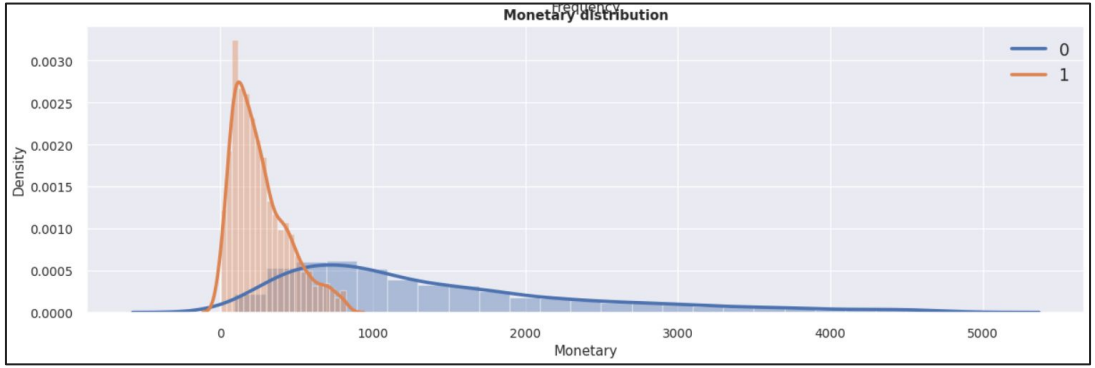
KMeans : 2 Clusters	Recency mean	Recency median	Frequency mean	Frequency median	Monetary mean	Monetary median	count
0	29.526480	17.000000	172.661475	108.000000	2864.098131	1236.500000	1926
1	141.953999	112.000000	24.942395	20.000000	353.506009	226.000000	2413

Model Building

K - Means Clustering : 2 Clusters







Model Building

K - Means Clustering : 2 Clusters

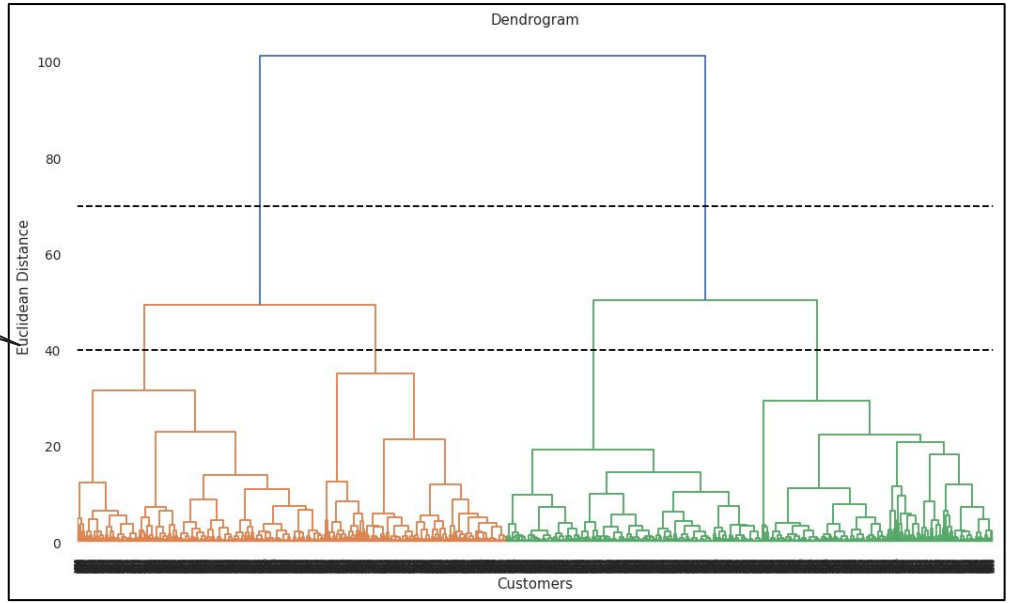
	Hierarchical : 2 Clusters	Last visited	Purchase frequency	Amount spent	Customer_Type
0	0	61 to 235 days ago	Bought 10 to 38 times	Spend around 128 to 480 Pound Sterling	Churned Customers 💔
1	1	7 to 31 days ago	Bought 52 to 181 times	Spend around 469 to 2212 Pound Sterling	Best Customers 🏅

Model Building

Hierarchical Clustering

Finding The Optimal ‘k’ : Dendrogram

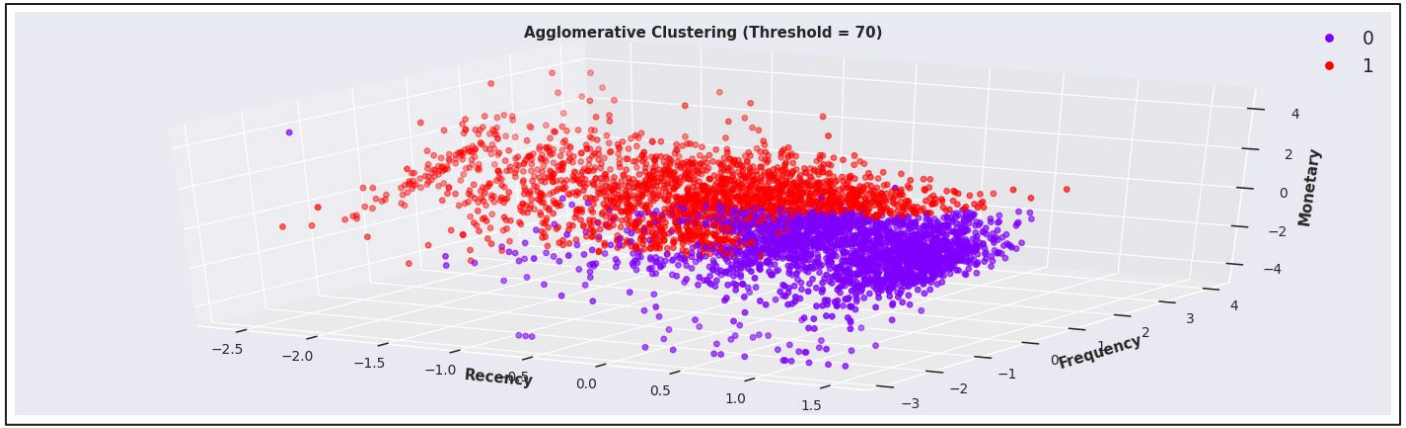
Agglomerative Clustering



- The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold, larger threshold (y = 70) results in 2 clusters, while the smaller threshold (y = 40) results in 4 clusters.

Model Building

Hierarchical Clustering : 2 Clusters

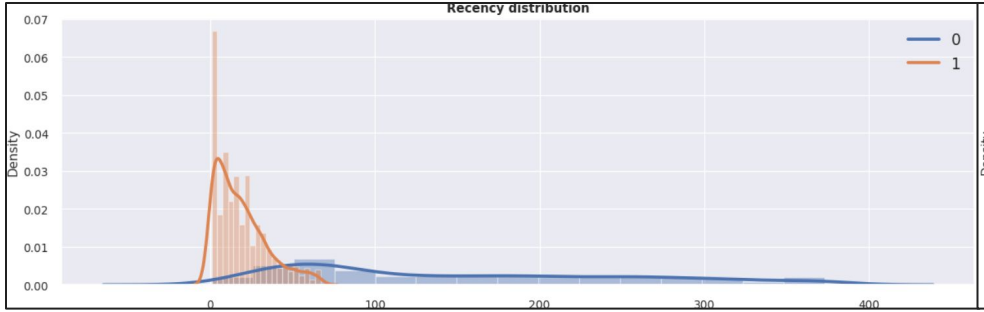


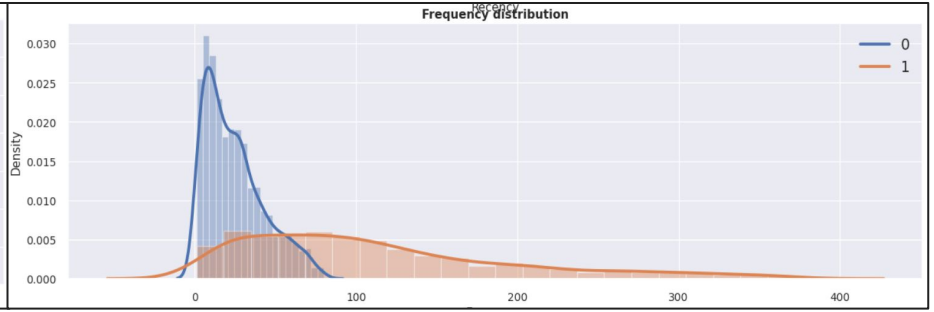
Displaying the mean, median of Recency, Frequency and Monetary for each group of customer

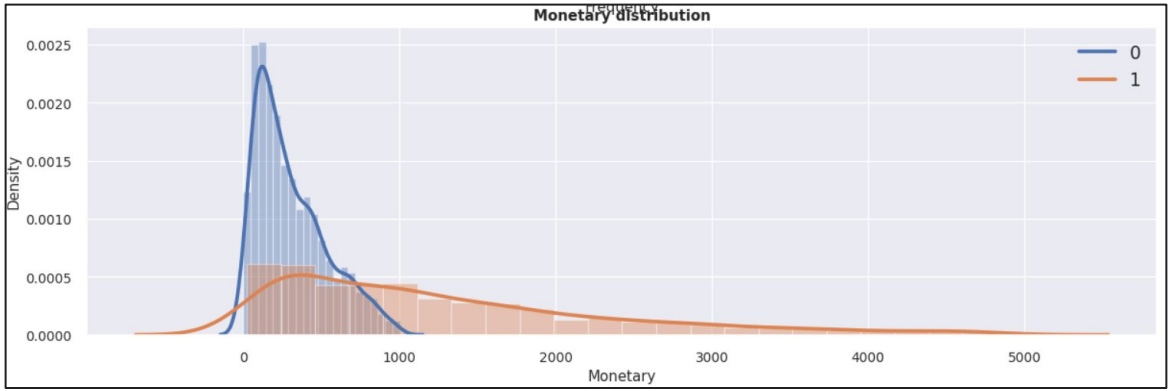
Hierarchical : 2 Clusters	Recency mean	Recency median	Frequency mean	Frequency median	Monetary mean	Monetary median	count
0	150.566218	126.000000	27.959618	22.000000	497.972644	263.000000	2303
1	25.859037	17.000000	161.267682	101.000000	2565.045187	1106.500000	2036

Model Building

Hierarchical Clustering : 2 Clusters







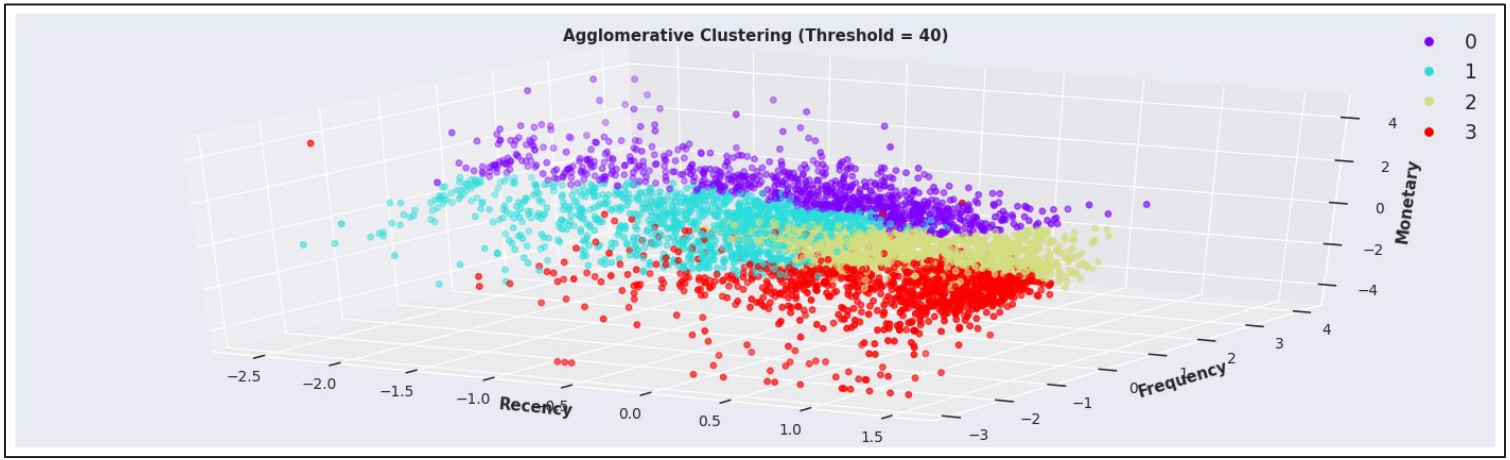
Model Building

Hierarchical Clustering : 2 Clusters

	Hierarchical : 2 Clusters	Last visited	Purchase frequency	Amount spent	Customer_Type
0	0	61 to 235 days ago	Bought 10 to 38 times	Spend around 128 to 480 Pound Sterling	Churned Customers 💔
1	1	7 to 31 days ago	Bought 52 to 181 times	Spend around 469 to 2212 Pound Sterling	Best Customers 🏅

Model Building

Hierarchical Clustering : 4 Clusters

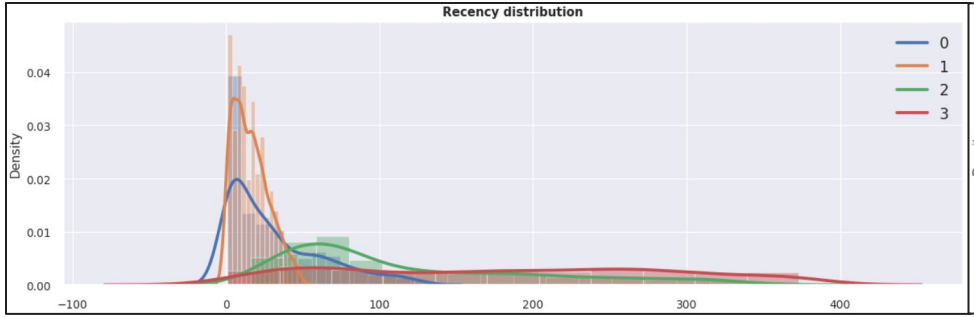


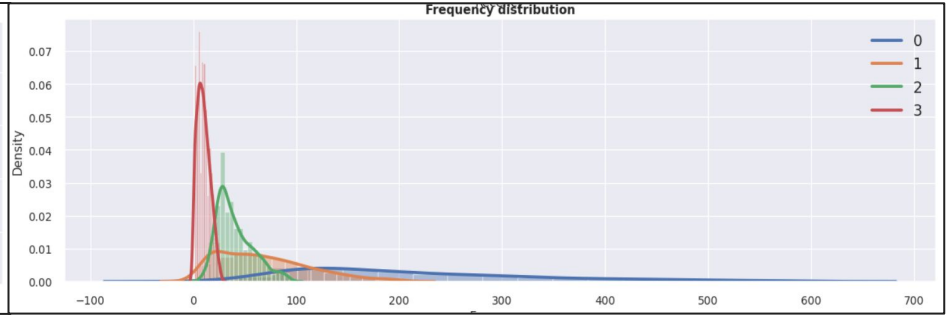
Displaying the mean, median of Recency, Frequency and Monetary for each group of customer

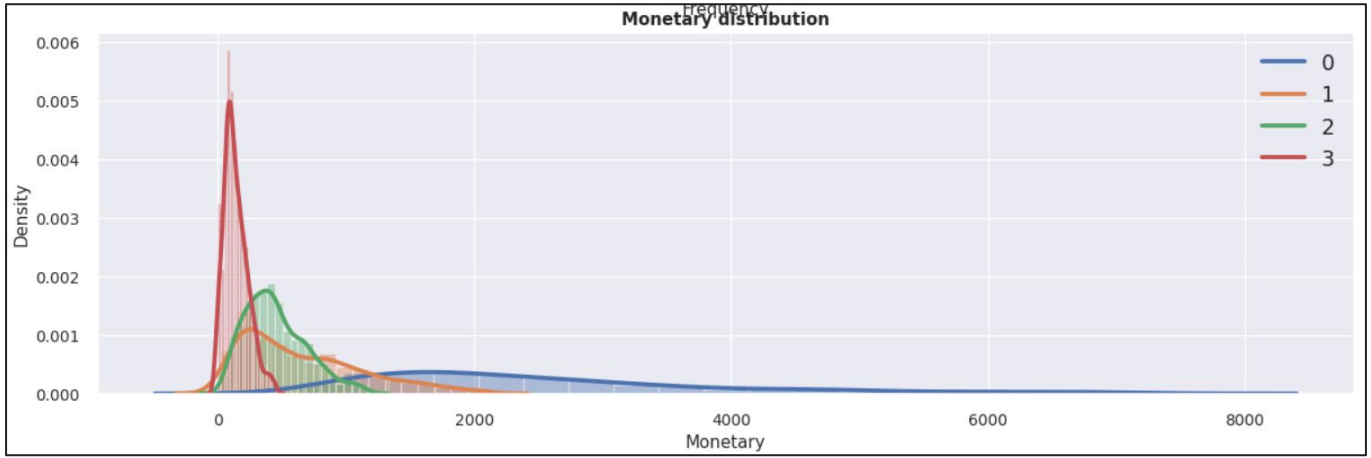
Hierarchical : 4 Clusters	Recency mean	Recency median	Frequency mean	Frequency median	Monetary mean	Monetary median	count
0	37.724739	23.000000	282.619048	194.000000	5053.137050	2365.000000	861
1	17.164255	15.000000	72.345532	63.000000	741.856170	555.000000	1175
2	122.182270	85.000000	44.010771	37.000000	487.097763	422.000000	1207
3	181.824818	184.000000	10.282847	9.000000	509.948905	136.500000	1096

Model Building

Hierarchical Clustering : 4 Clusters







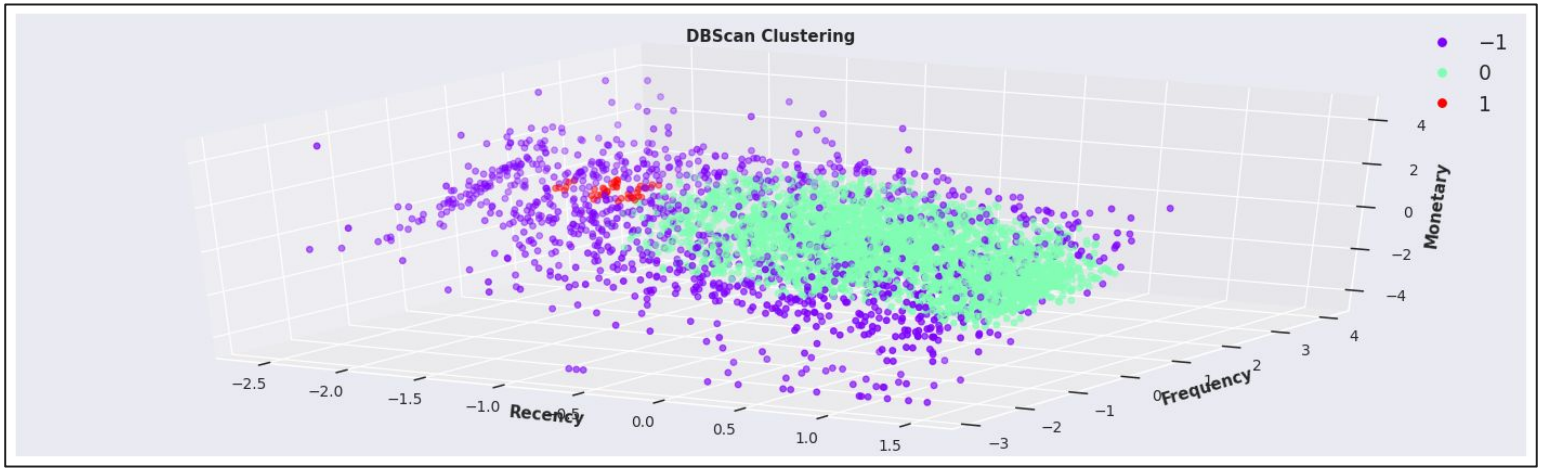
Model Building

Hierarchical Clustering : 4 Clusters

	Hierarchical : 4 Clusters	Last visited	Purchase frequency	Amount spent	Customer_Type
0	0	7 to 56 days ago	Bought 122 to 308 times	Spend around 1588 to 3912 Pound Sterling	Best Customers
1	1	7 to 24 days ago	Bought 31 to 99 times	Spend around 271 to 994 Pound Sterling	Good Customers
2	2	56 to 177 days ago	Bought 28 to 54 times	Spend around 277 to 631 Pound Sterling	Risky to Churn segment customers
3	3	78 to 269 days ago	Bought 5 to 14 times	Spend around 81 to 228 Pound Sterling	Churned Customers

Model Building

DBSCAN Clustering

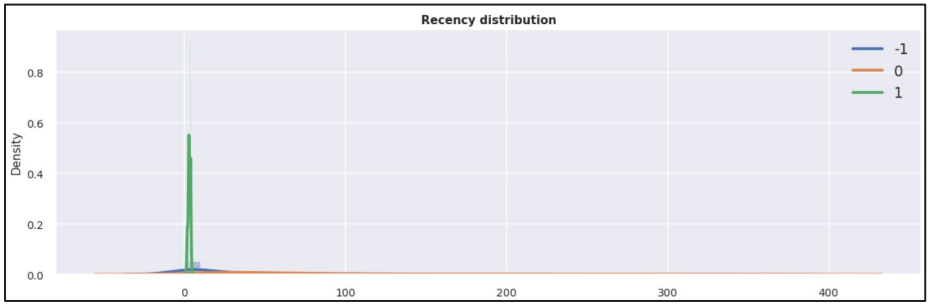


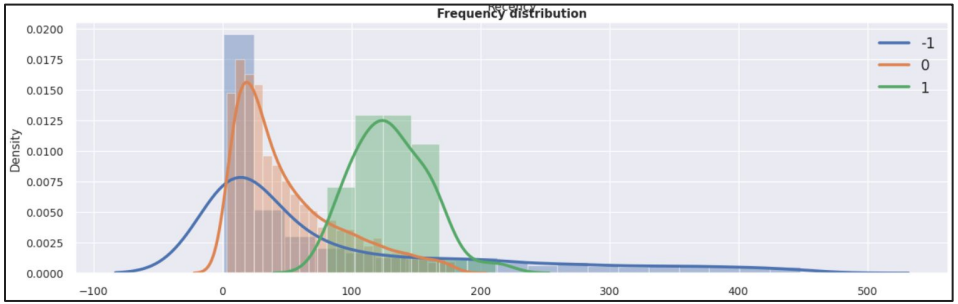
Displaying the mean, median of Recency, Frequency and Monetary for each group of customer

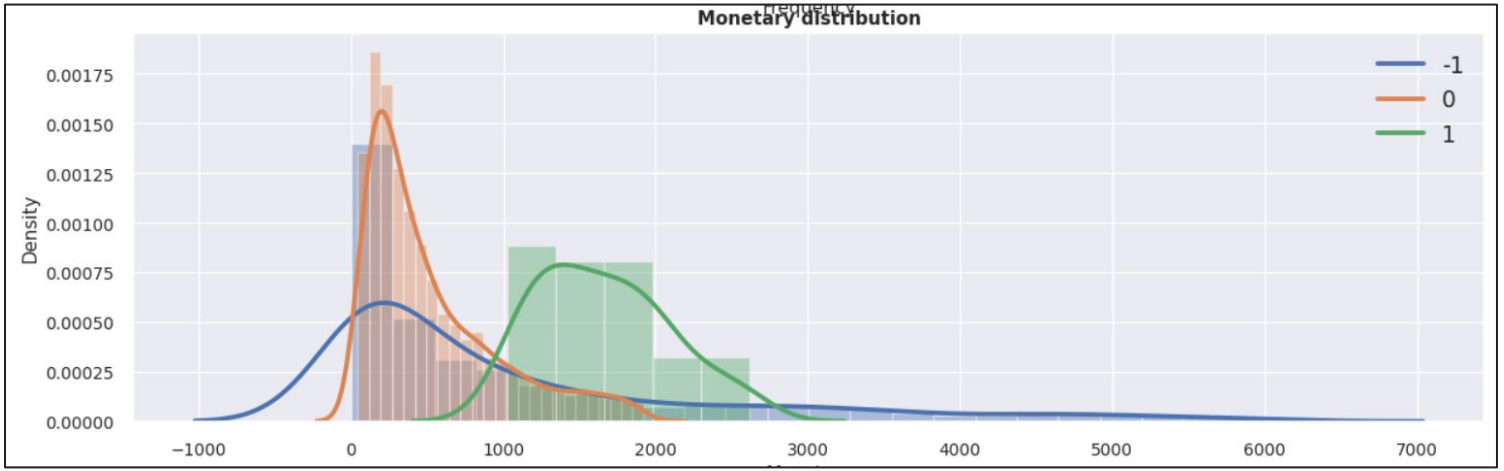
DBScan	Recency mean	Recency median	Frequency mean	Frequency median	Monetary mean	Monetary median	count
-1	63.232082	16.000000	159.396758	40.500000	3456.244027	681.500000	1172
0	103.954284	61.000000	64.203964	40.000000	720.537404	417.500000	3128
1	3.230769	3.000000	130.487179	129.000000	1659.076923	1614.000000	39

Model Building

DBSCAN Clustering







Model Building

DBSCAN Clustering

	DBScan	Last visited	Purchase frequency	Amount spent	Customer_Type
0	-1	3 to 84 days ago	Bought 7 to 186 times	Spend around 124 to 2490 Pound Sterling	Average Customers
1	0	26 to 166 days ago	Bought 20 to 85 times	Spend around 210 to 896 Pound Sterling	Risky to Churn segment customers
2	1	3 to 4 days ago	Bought 109 to 151 times	Spend around 1277 to 1937 Pound Sterling	Good Customers

Model Building

Summary

	Model Name	Category	Optimal Number of Clusters
0	RFM Score	RFM Score Binning	5
1	RFM Group	RFM Group Binning	5
2	K - Means Clustering	Elbow Curve	5
3	K - Means Clustering	Silhouette Score	2
4	Hierarchical Clustering	Agglomerative Clustering (y = 70)	2
5	Hierarchical Clustering	Agglomerative Clustering (y = 40)	4
6	DBScan Clustering	eps = 0.3, min_samples = 20	3

Note : - The number of optimal clusters are approx

Conclusion

Different Customer Segments Obtained

	RFM Score Binning	RFM Group Binning	K-Means Clustering (Elbow Curve)	K-Means Clustering (Silhouette Score)	Hierarchical Clustering (y = 70)	Hierarchical Clustering (y = 40)	DBSCAN Clustering(eps = 0.3, min_samples = 20)
0	Best Customers	Best Customers	Churned Customers	Best Customers	Churned Customers	Best Customers	Average Customers
1	Good Customers	Good Customers	Good Customers	Churned Customers	Best Customers	Good Customers	Risky to Churn segment customers
2	Average Customers	Average Customers	Best Customers	NaN	NaN	Risky to Churn segment customers	Good Customers
3	Risky to Churn segment customers	Risky to Churn segment customers	Risky to Churn segment customers	NaN	NaN	Churned Customers	NaN
4	Churned Customers	Churned Customers	Average Customers	NaN	NaN	NaN	NaN

- However, there can be more modifications on this analysis. One may choose to cluster into more number of 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, segmenting on the basis of time period they visit and much more.

- As machine learning has become more of an ART, there is nothing such as right or wrong. We only try to get the best outcomes that can suit our final objectives. There is, and always will be, a need to improve, going forward.

Challenges Faced

- Handling Missing values in the dataset.
- Creating new features from the existing features for better understanding of the dataset.
- Difficulty in deciding ‘K’ Value before performing K - Means Clustering.

References

- MachineLearningMastery
- GeeksforGeeks
- Analytics Vidhya Blogs
- Towards Data Science Blogs
- Stack Overflow

Thank You!