Capstone Project

UNSUPERVISED MACHINE LEARNING

(CUSTOMER SEGMENTATION)

ONLINE RETAIL

PRAJWAL D U

INTRODUCTION

1 The main goal is to identify customers that are most profitable and the ones who churned out to prevent further loss of customer by redefining company policies.
2 CLUSTER ANALYSIS: Statistically Segment Customers into groups Observation by using the features given below.

IMPORTING AND INSPECTING DATASET

Data set name :- online retail

No of observation :- 541908 (shape=541908 x 8)

dtypes :- datetime=(1), float64=(2), int64=(1), object=(4), 1+2+1+4=8 columns

DATA DESCRIPTION

Attribute Information:

- 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.

DATA CLEANING

Checking Duplicates

5268 data points were duplicated

Dropped Duplicates

Checking Missing Data

1. CustomerID - 135080(25% missing values)
2. Description - 1454(0.27% missing values)

No use of this data it can be dropped

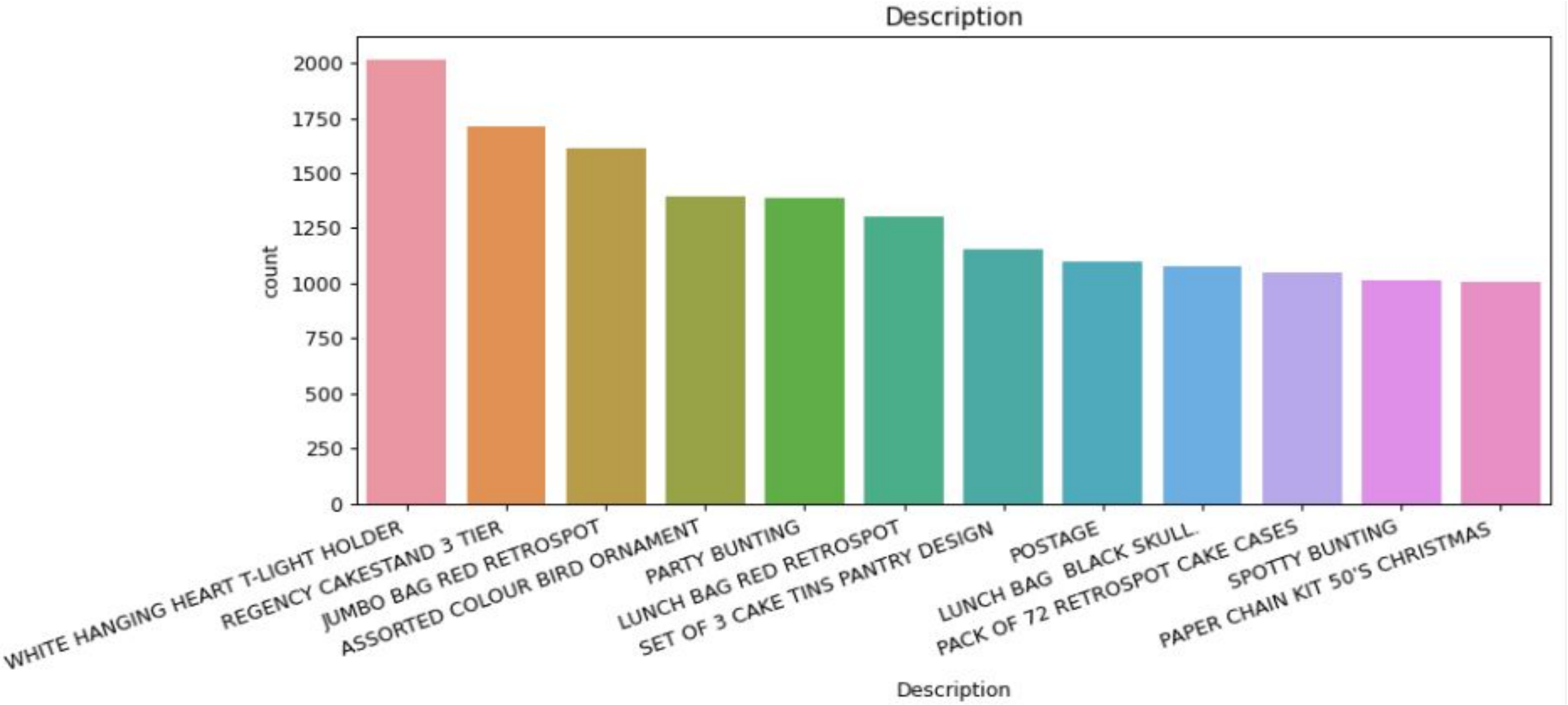
Total data points left

No of observation left - 401604 (shape = 8x401604

FEATURE ENGINEERING

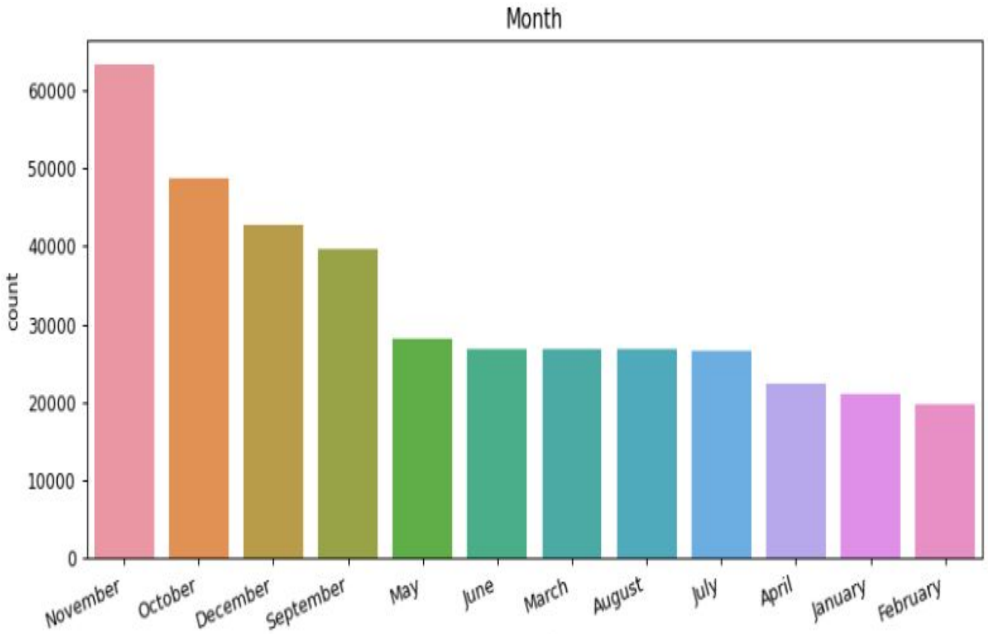
- Extracting year month and day from Invoice Date
- Creating new feature 'TotalAmount' by multiplying values from the Quantity and UnitPrice column.
- Creating new feature 'Timetype' based on hours to define whether its Morning, Afternoon, or evening
- Dropping InvoiceNo starting with 'C' that represents cancellation

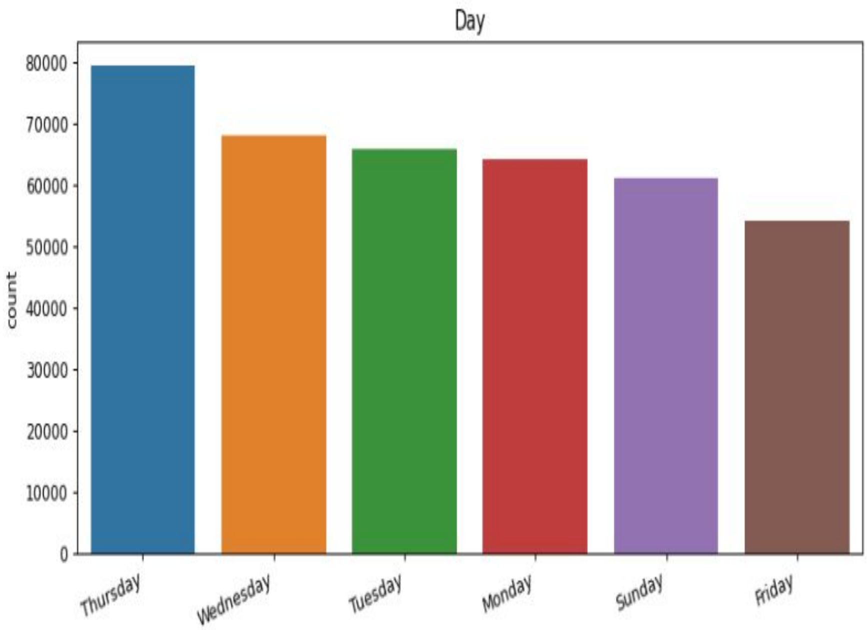
MOST FREQUENT VALUES



WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER, JUMBO BAG RED RETROSPOT are the most ordered products,

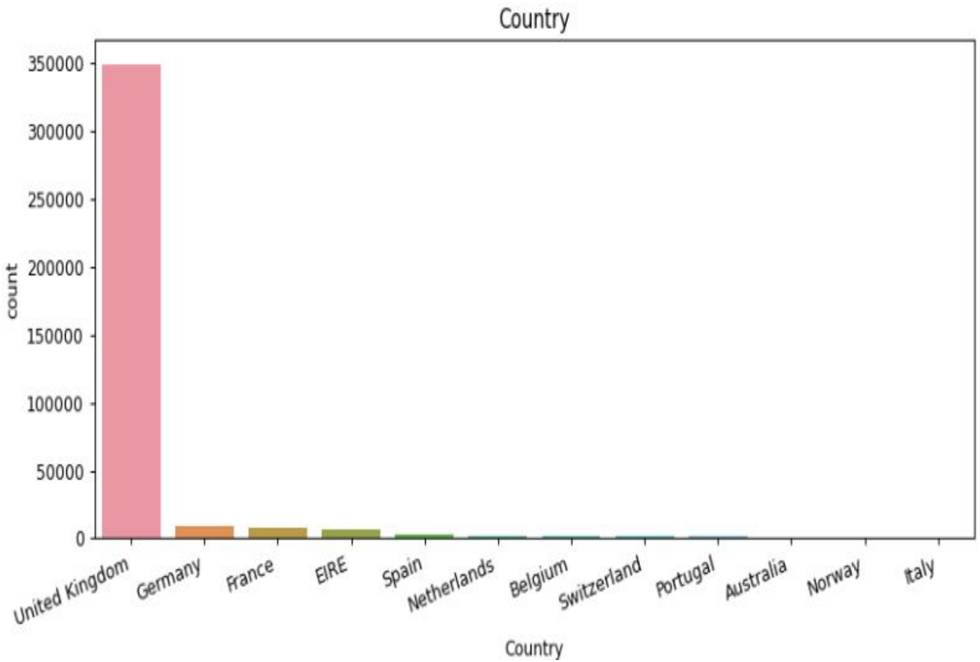
MOST FREQUENT VALUES

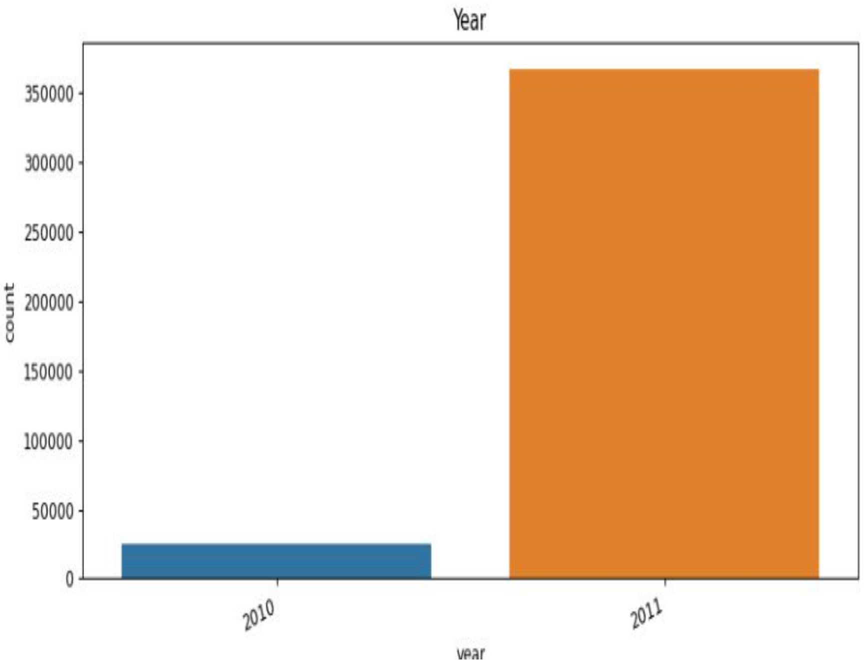




- Most of the customers have purchased the gifts in the month of November, October, December and September, Less number of customers have purchased the gifts in the month of April, January and February.
- Thursday is high selling day according to data and There are no orders placed on Saturdays. Looks like it's a non working day for the retailer.

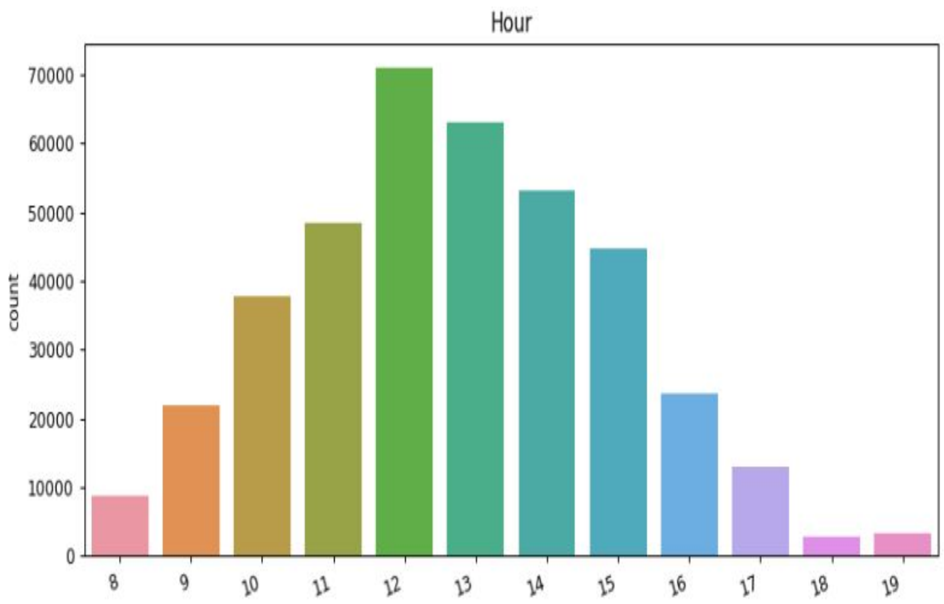
MOST FREQUENT VALUES

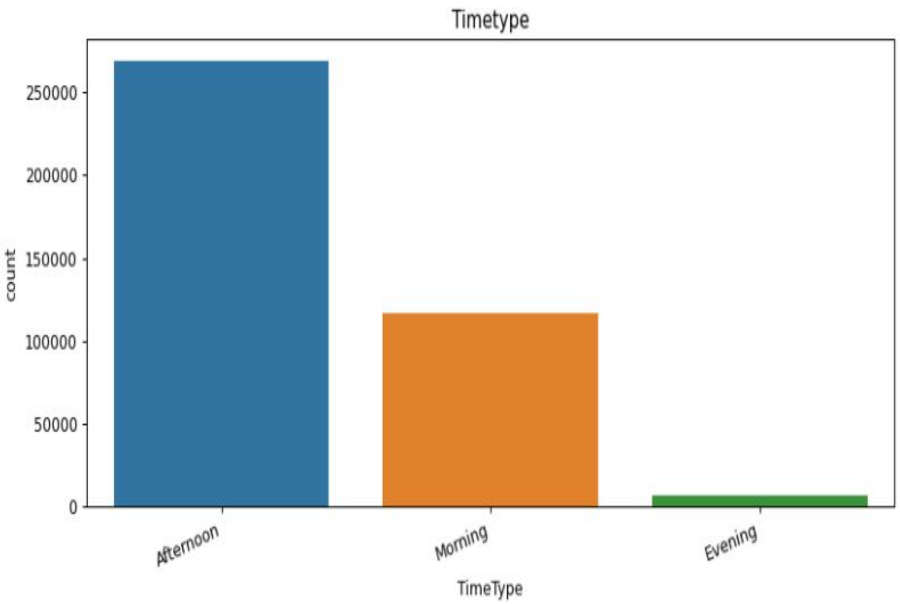




- Most Customers are from United Kingdom. Considerable number of customers are also from Germany, France, EIRE and Spain. Whereas Saudi Arabia, Bahrain, Czech Republic, Brazil and Lithuania has least number of customers.
- 2011 is our high selling year and 2010 is least

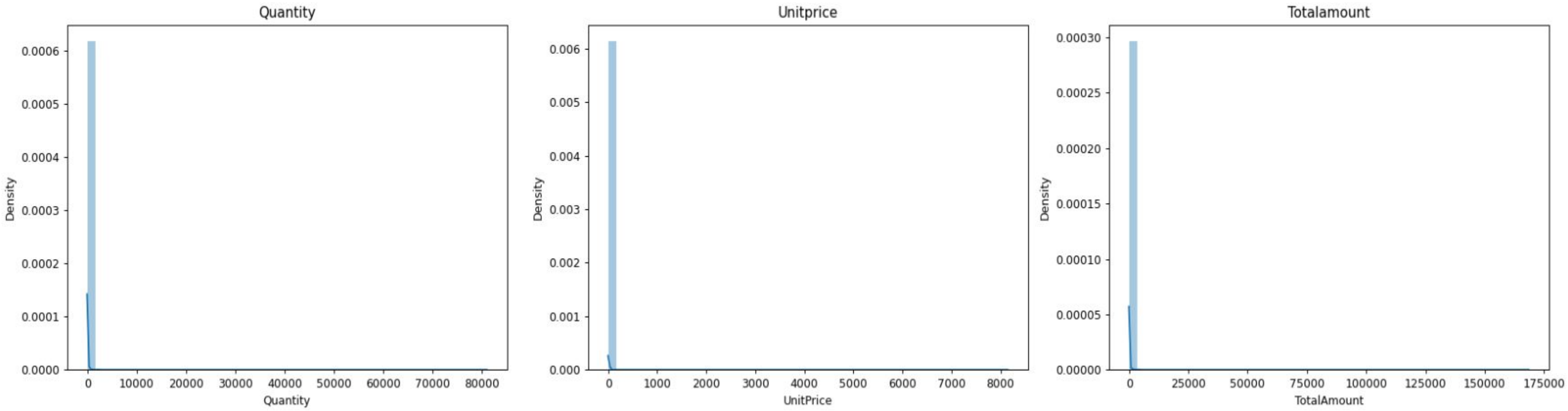
MOST FREQUENT VALUES





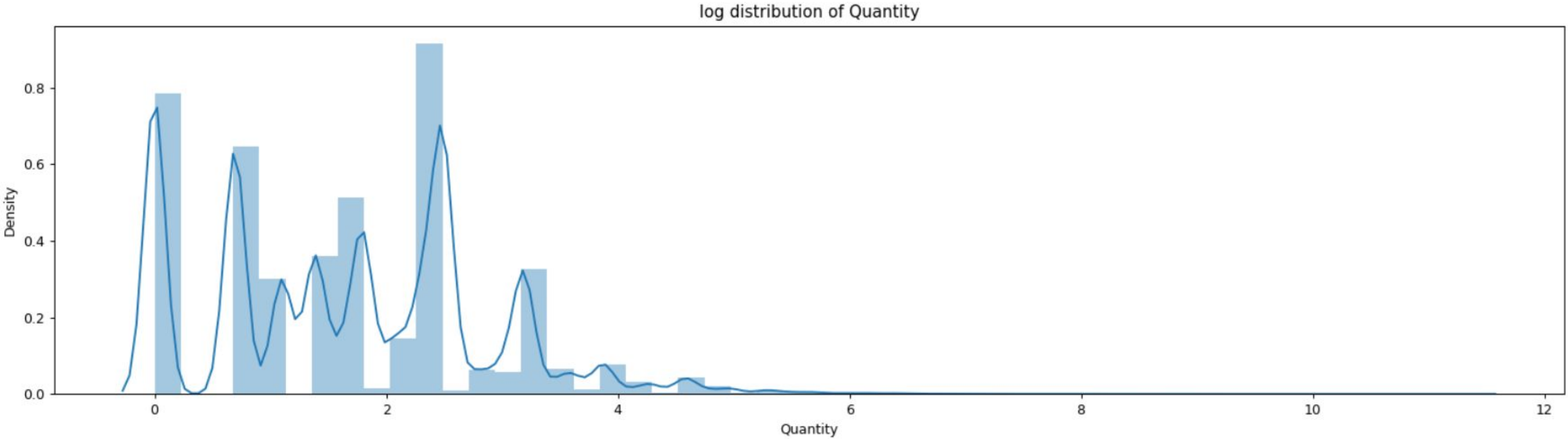
- Most of the customers have purchased the items in Afternoon, moderate numbers of customers have purchased the items in Morning and the least in Evening.

VISUALIZING DISTRIBUTION



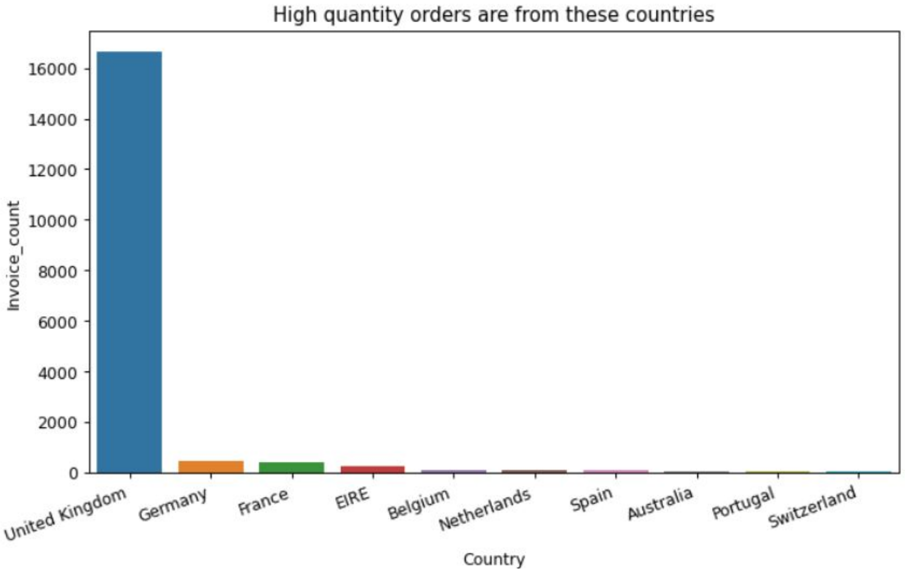
- It shows a positively skewed distribution because most of the values are clustered around the left side of the distribution while the right tail of the distribution is longer, which means mean>median>mode
- For symmetric graph mean=median=mode

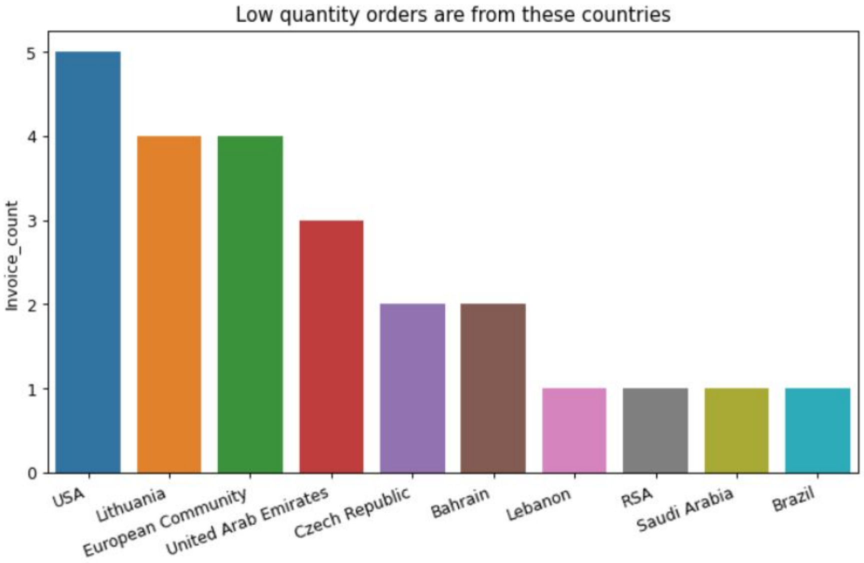
LOG TRANSFORMATION



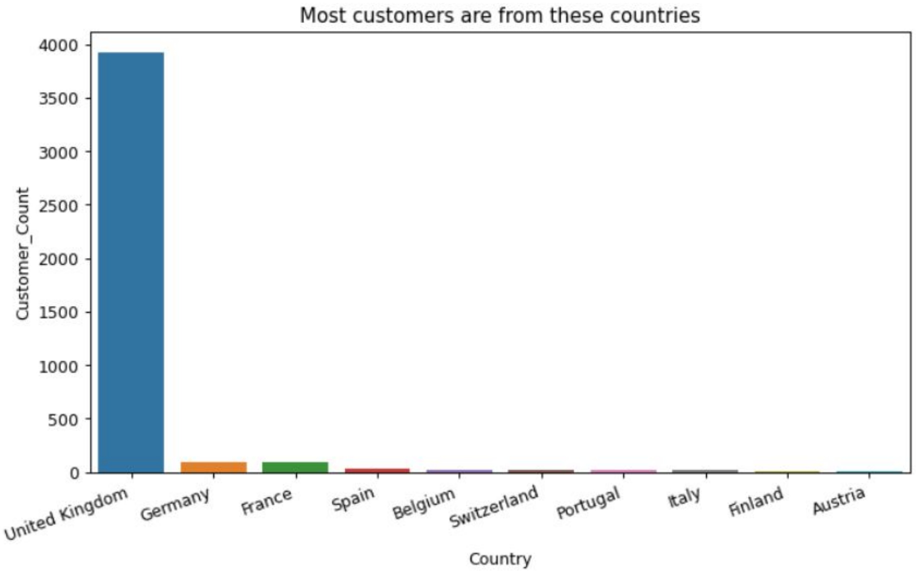
- After applying log transformation now the distribution plot looks comparatively better than being skewed.
- We use log transformation when our original continuous data does not follow the bell curve, we can log transform this data to make it as “normal” as possible so that the analysis results from this data become more valid.

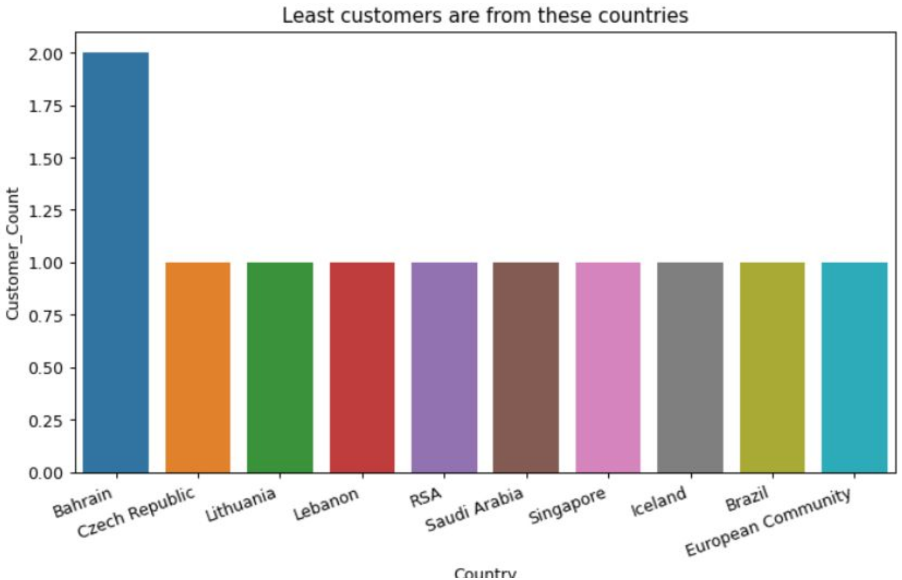
QUANTITY WISE ORDERS



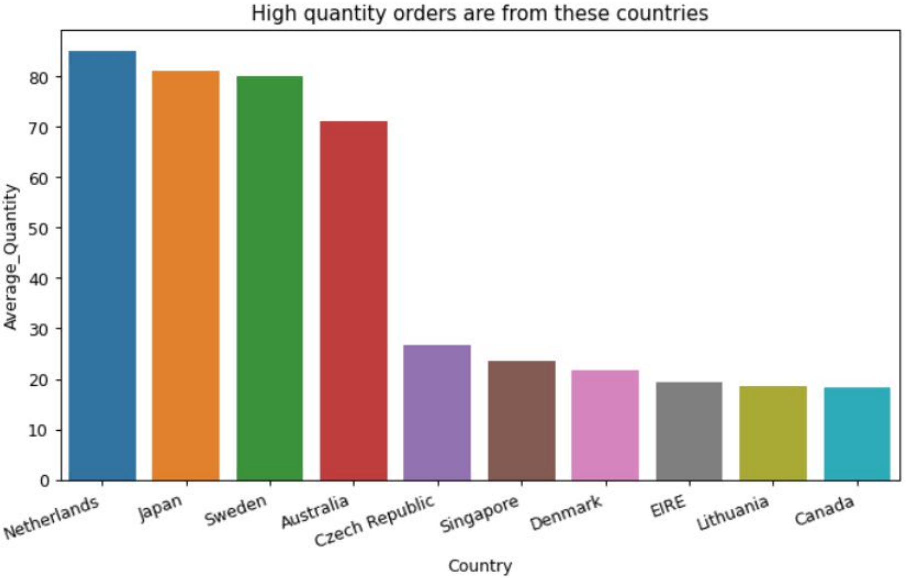


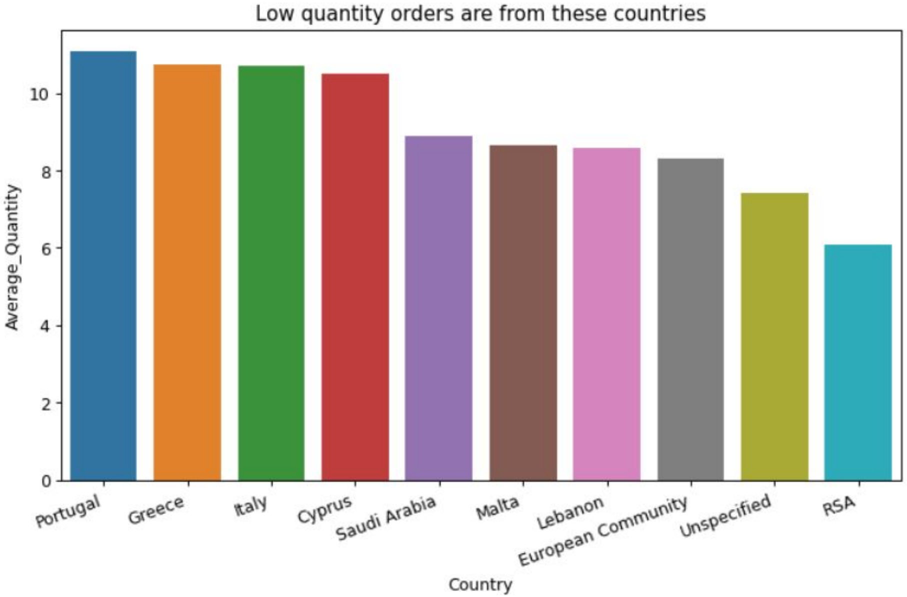
COUNTRY WISE CUSTOMERS



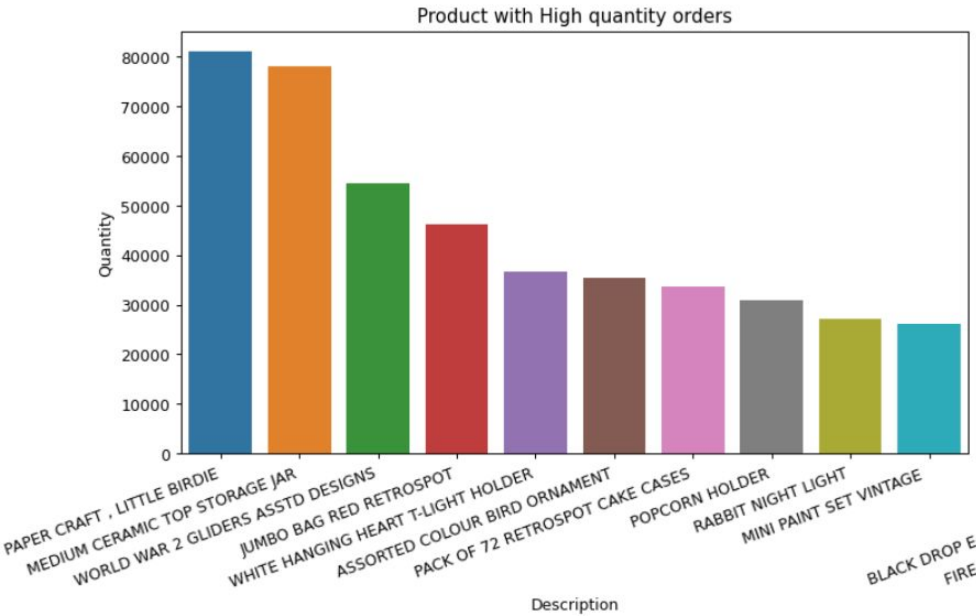


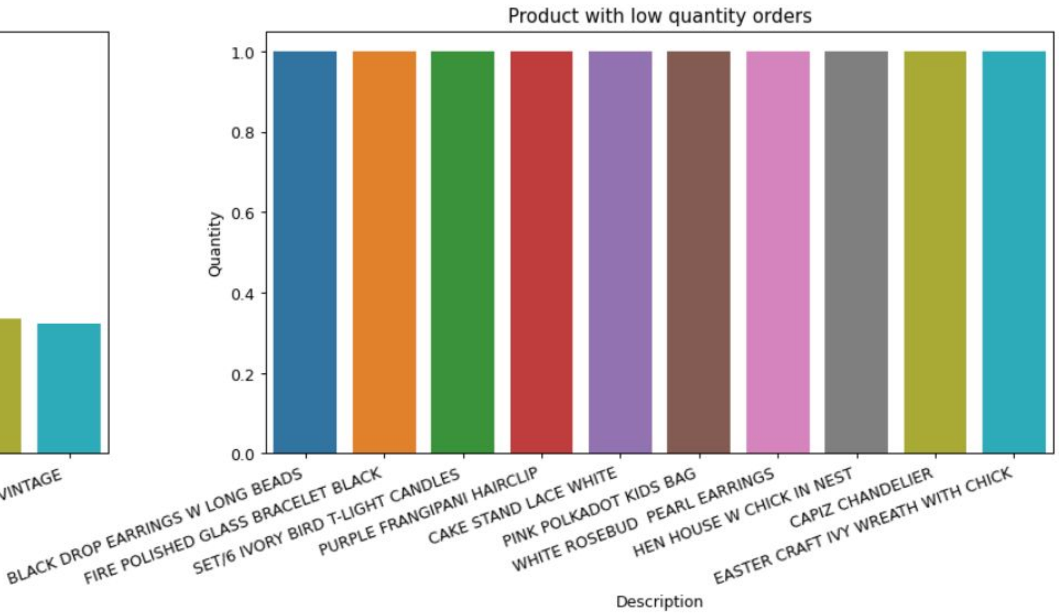
COUNTRY WISE PURCHASE QUANTITY



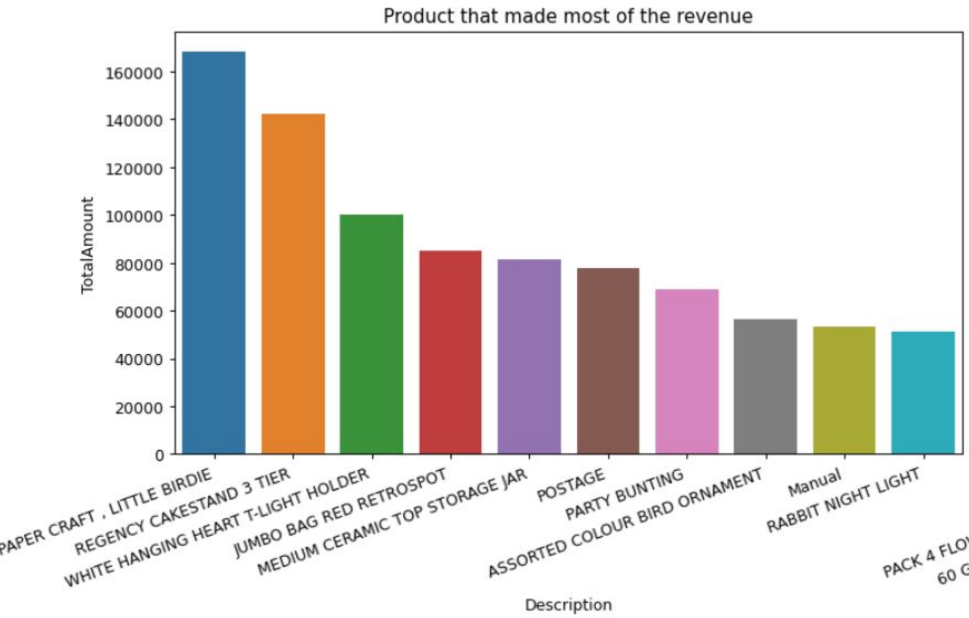


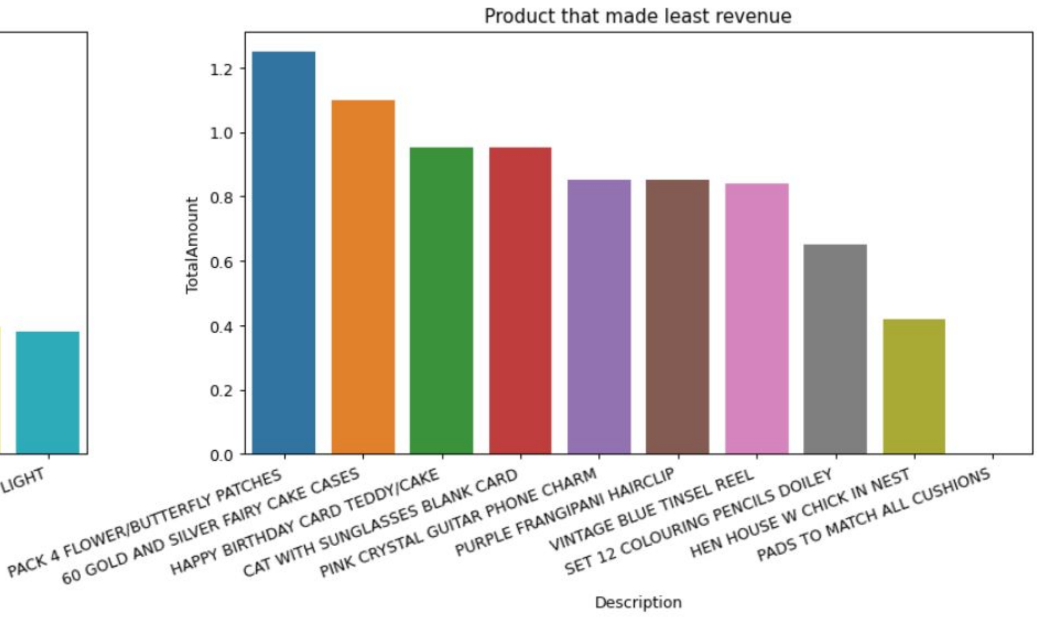
PRODUCT WISE PURCHASE QUANTITY



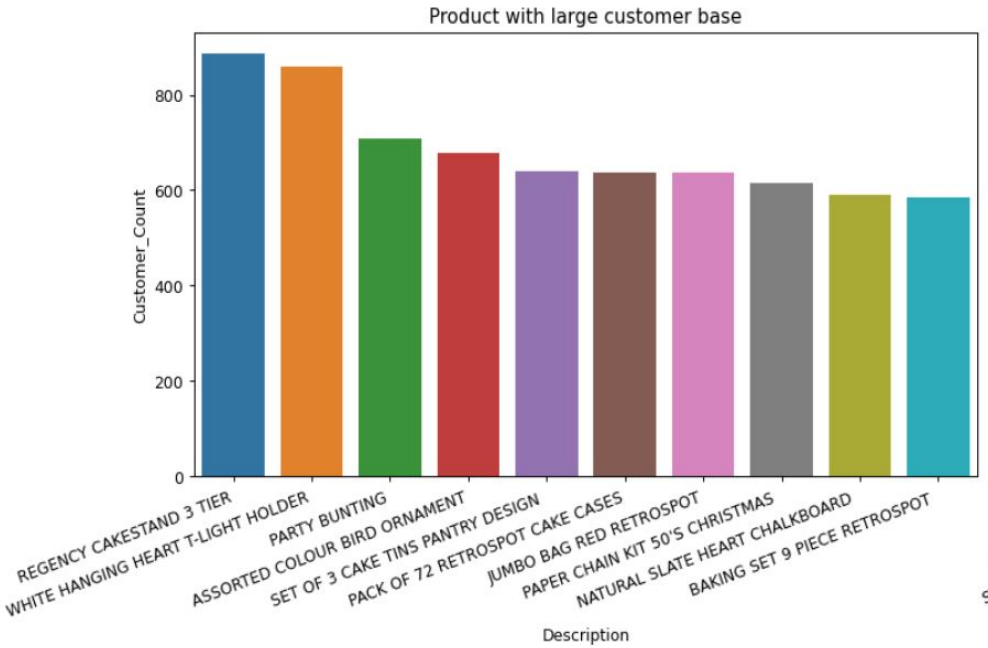


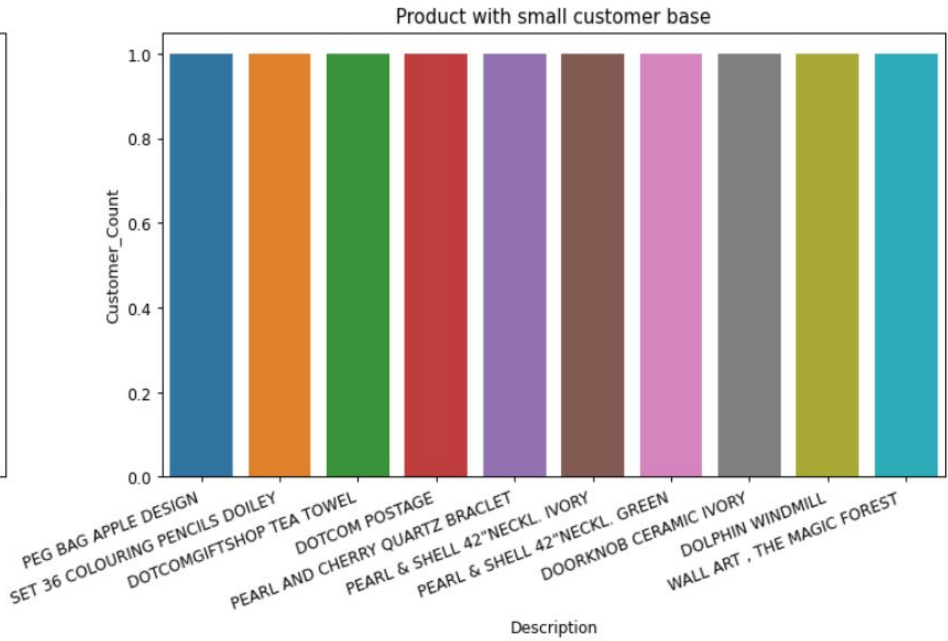
PRODUCT WISE REVENUE



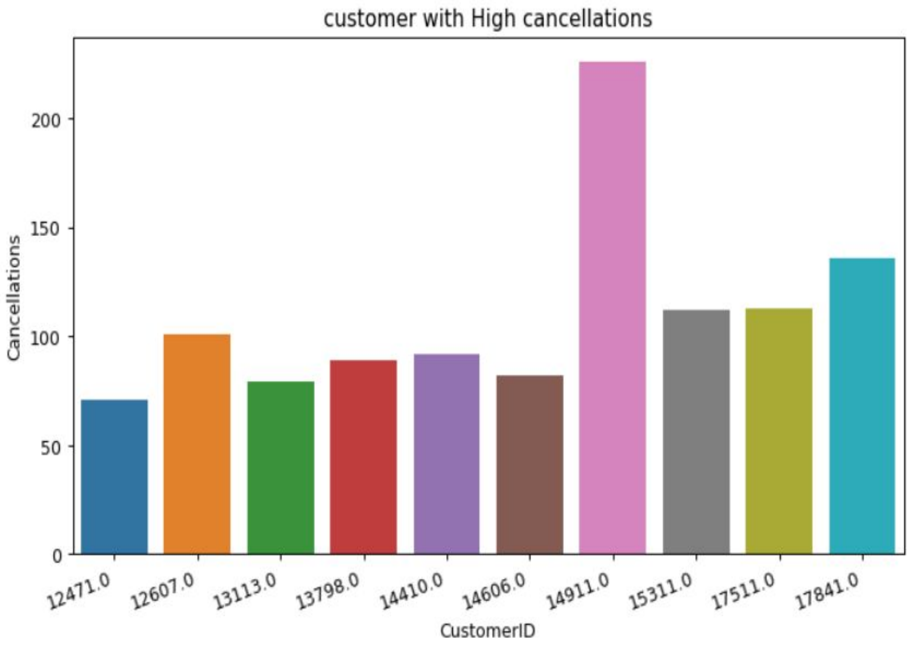


PRODUCT WISE CUSTOMERS

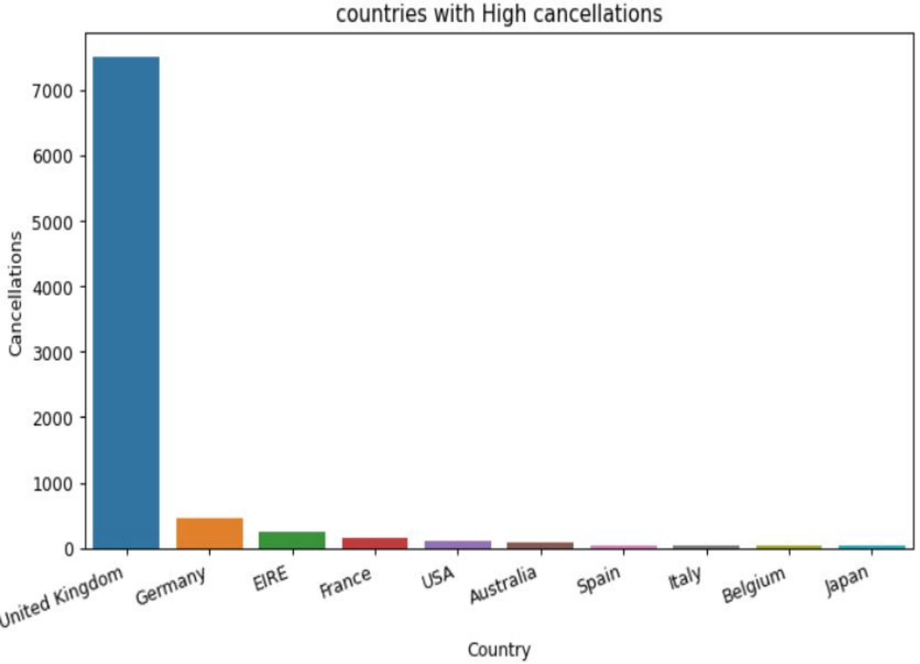


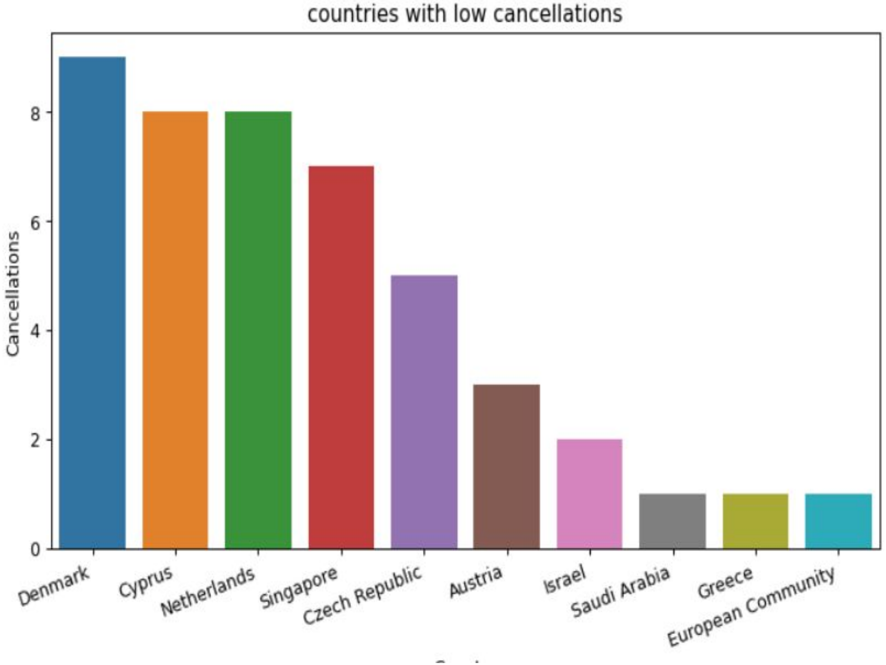


CUSTOMER WISE CANCELLATIONS



COUNTRY WISE CANCELLATIONS





RFM ANALYSIS

	CustomerID	Recency	Frequency	Monetary
0	12346.0	326	1	77183.60
1	12347.0	2	182	4310.00
2	12348.0	75	31	1797.24
3	12349.0	19	73	1757.55
4	12350.0	310	17	334.40

RFM TABLE

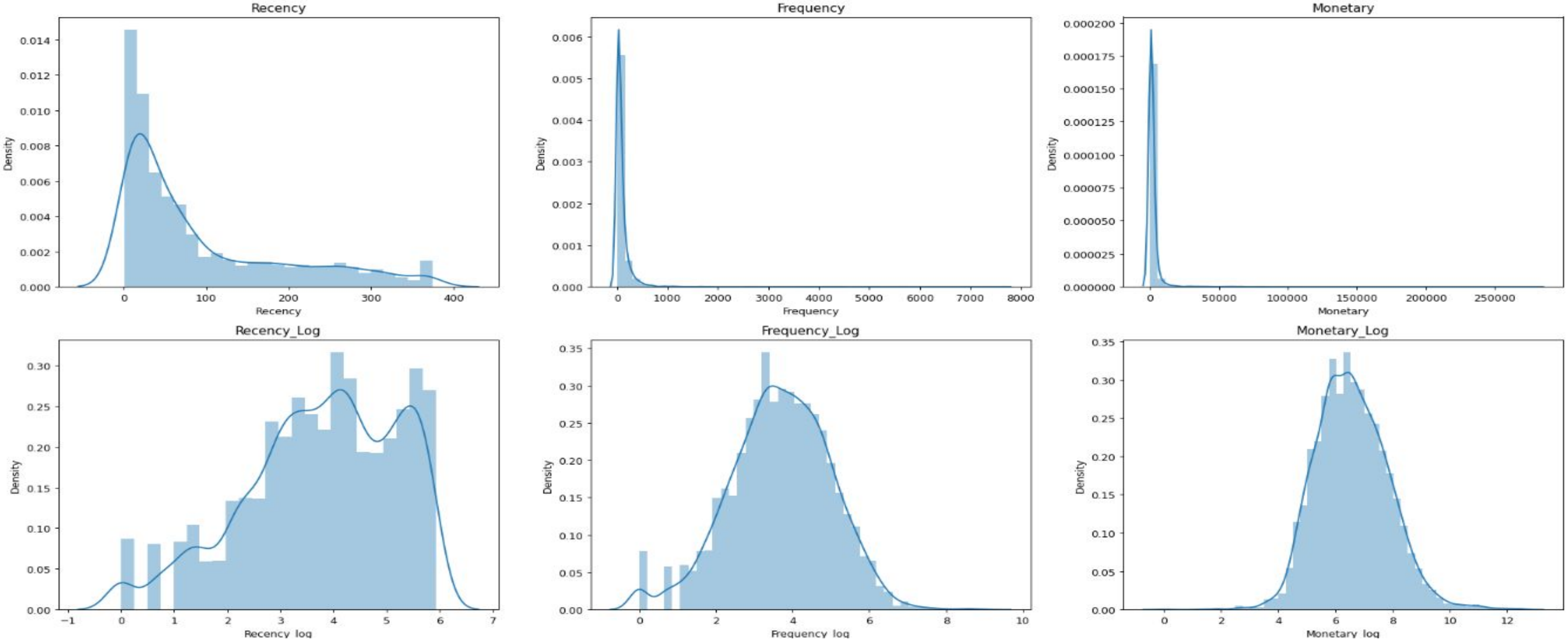
Conclusions :-

1 If the RFM of any customer is 444. His Recency is good, frequency is more and Monetary is more. So, he is the best customer.

2 If the RFM of any customer is 111. His Recency is low, frequency is low and Monetary is low. So, he is the churning customer.

3 If the RFM of any customer is 144. He purchased a long time ago but buys frequently and spends more. And so on.

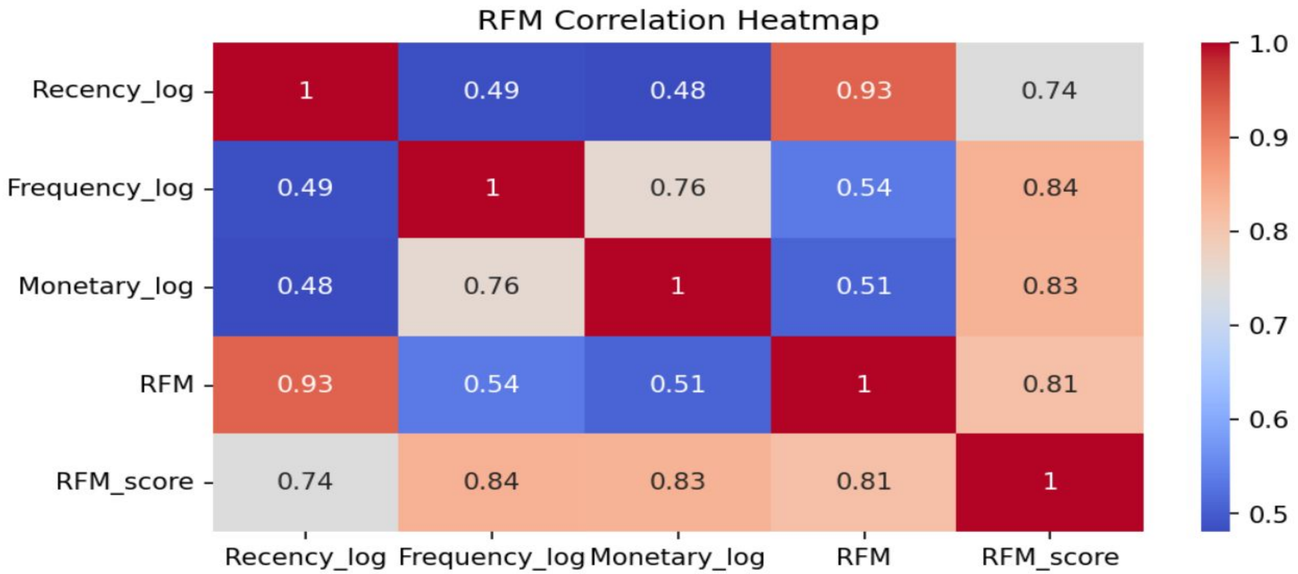
RFM MODELING



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.

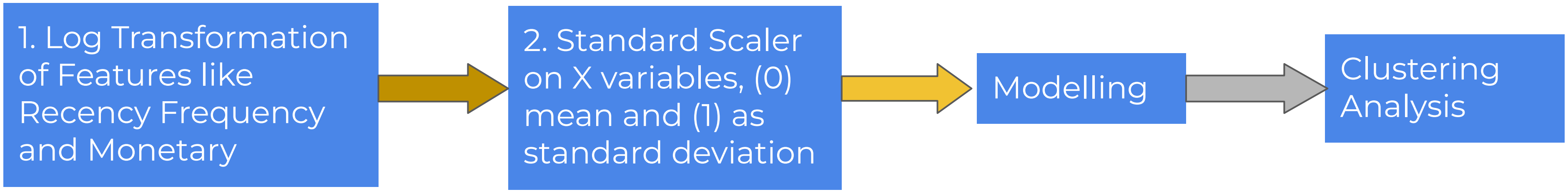
It will be more suitable to use the transformed features for better visualization of clusters

RFM CORRELATION HEATMAP



- We can see that Recency is highly correlated with the RFM value.
- Frequency and Monetary are moderately correlated with the RFM.

Scaling for CLUSTERING Analysis



PIPELINE

EXTRACTING DATA

Online Retail
Observation:
541908
(shape=8x541908)

DATA CLEANING

Checking missing data
1. 25 % of items
(i.e 135080)
2. CustomerID – 1454
Checking duplicates
5268 data points were
Duplicated
401604 DATA POINT LEFT

DATA VISUALIZATION

RFM ANALYSIS

RECENCY: Must be **LESS**

FREQUENCY: Must be **MORE**

MONETARY: Must be **MORE**

Condition: For Best Customers

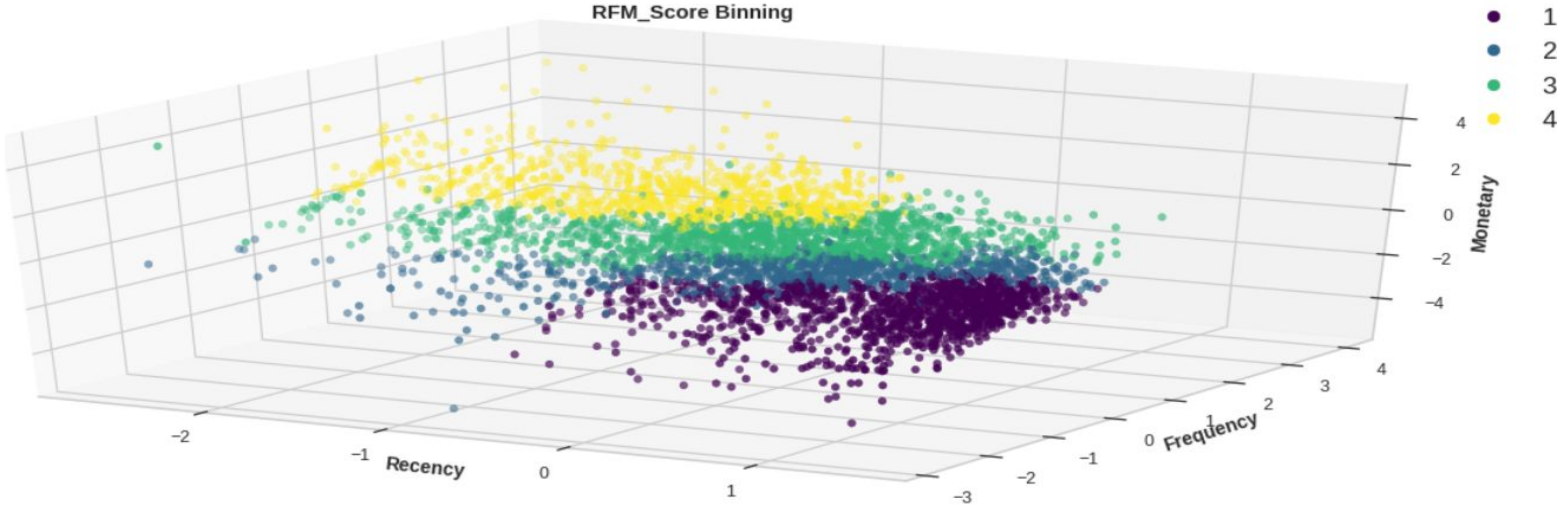
MODELLING

Binning (RFM SCORE)
Binning (RFM combination)
K-Means
Hierarchical
DBSCAN Clustering

CUSTOMER SEGMENTATION

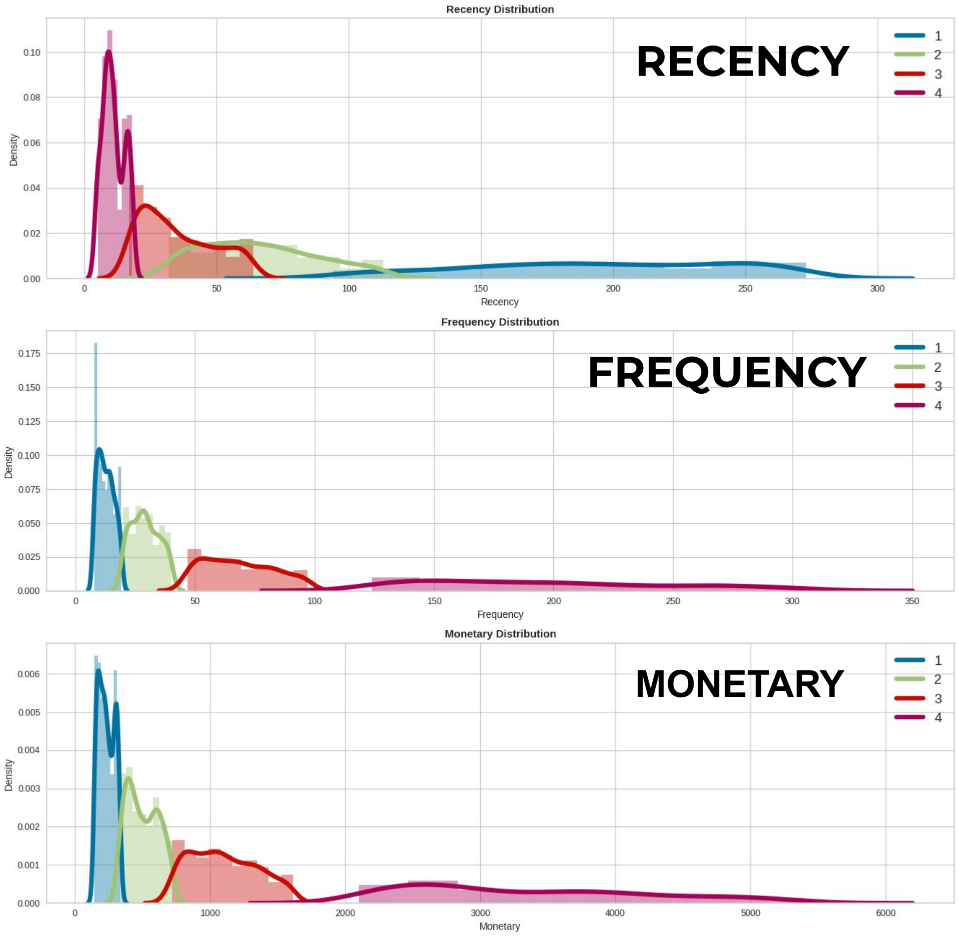
CONCLUSION

BINNING RFM SCORES



Binning	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
1	192.165501	196.000000	15.062160	12.000000	266.505704	225.900000	1287
2	87.606949	64.000000	32.930510	29.000000	788.401130	488.200000	921
3	47.848532	31.000000	81.241886	67.000000	1597.725141	1076.100000	1294
4	13.761051	10.000000	284.218638	190.000000	6870.541553	3158.130000	837
Binning	Last_visited	purchase_frequency			Money_spend		

BINNING RFM SCORES



Binning	Last_visited	purchase_frequency	Money_spend
1	93 to 274 days ago	Bought7 to 20 times	spent around 142 to 335 sterling
2	31 to 114 days ago	Bought19 to 41 times	spent around 327 to 725 sterling
3	16 to 65 days ago	Bought46 to 98 times	spent around 717 to 1613 sterling
4	4 to 19 days ago	Bought123 to 305 times	spent around 2093 to 5398 sterling

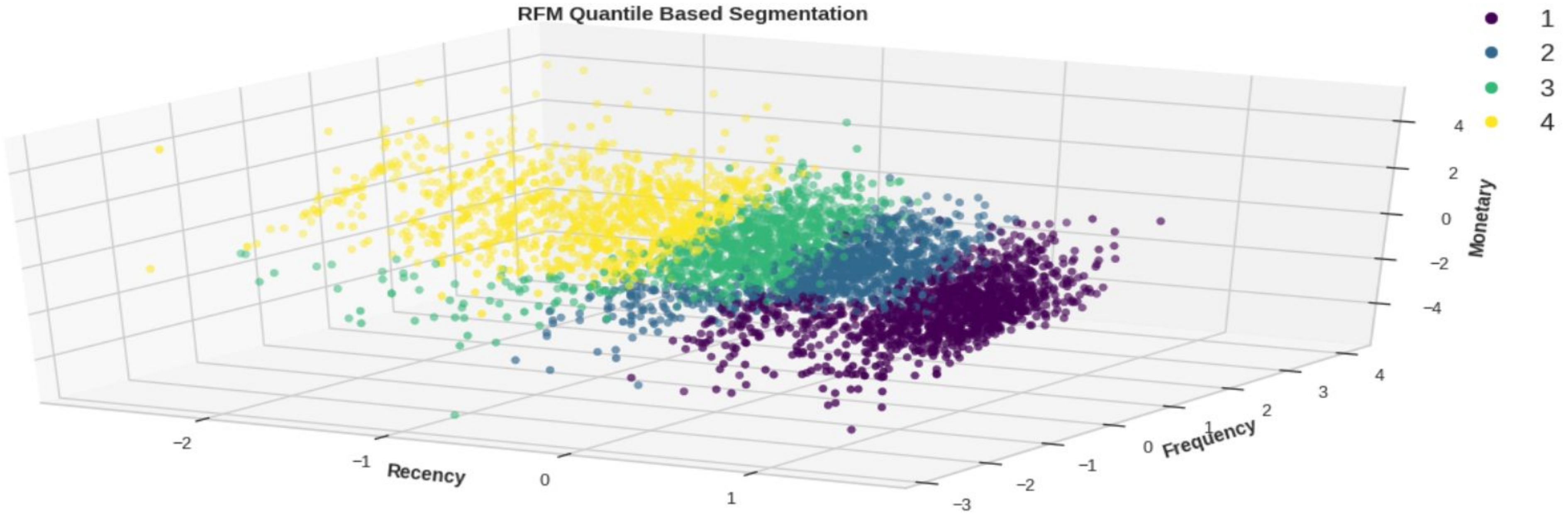
GROUP 1 — LOST POOR CUSTOMERS

GROUP 2 — AVERAGE CUSTOMERS

GROUP 3 — GOOD CUSTOMERS

GROUP 4 — BEST CUSTOMERS

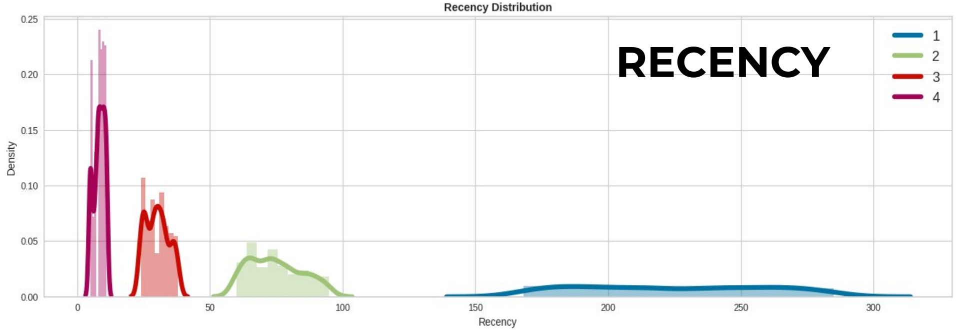
QUANTILE CUT

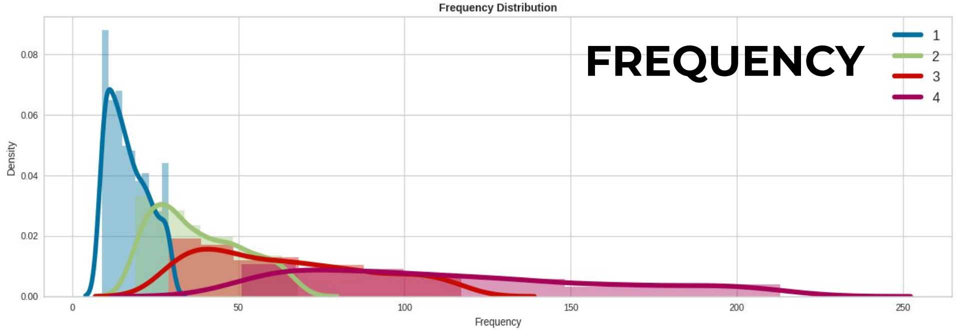


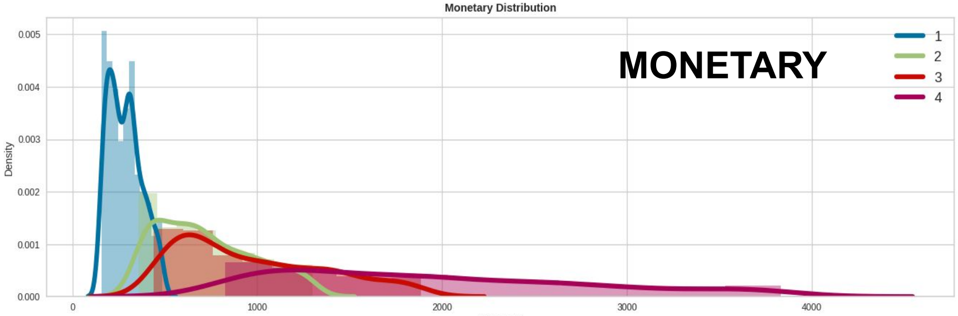
QuantileCut	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
1	224.110055	220.000000	26.190024	15.000000	582.373025	280.550000	1263
2	77.805941	73.000000	54.198020	36.000000	1078.258853	675.645000	1010
3	30.647175	30.000000	94.935580	61.000000	1831.494709	881.290000	1009
4	8.400189	8.000000	197.846736	106.000000	4933.446698	1814.120000	1057

QuantileCut	Last_visited	purchase_frequency	Money_spend

QUANTILE CUT







QuantileCut	Last_visited	purchase_frequency	Money_spend
1	166 to 286 days ago	Bought8 to 30 times	spent around 156 to 486 sterling
2	59 to 96 days ago	Bought18 to 69 times	spent around 355 to 1301 sterling
3	23 to 39 days ago	Bought28 to 118 times	spent around 439 to 1887 sterling
4	4 to 12 days ago	Bought50 to 214 times	spent around 822 to 3849 sterling

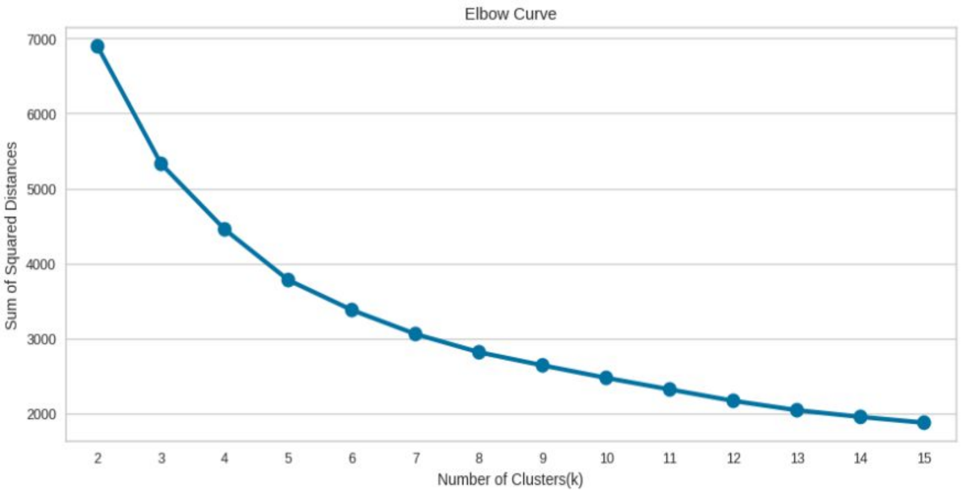
GROUP 1 — LOST POOR CUSTOMERS

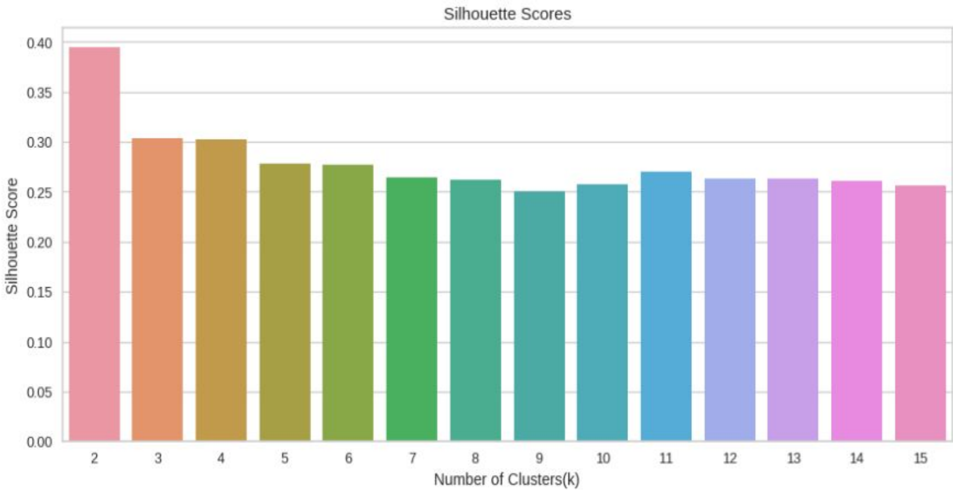
GROUP 2 — LOSING LOYAL CUSTOMERS

GROUP 3 — GOOD CUSTOMERS

GROUP 4 — BEST CUSTOMERS

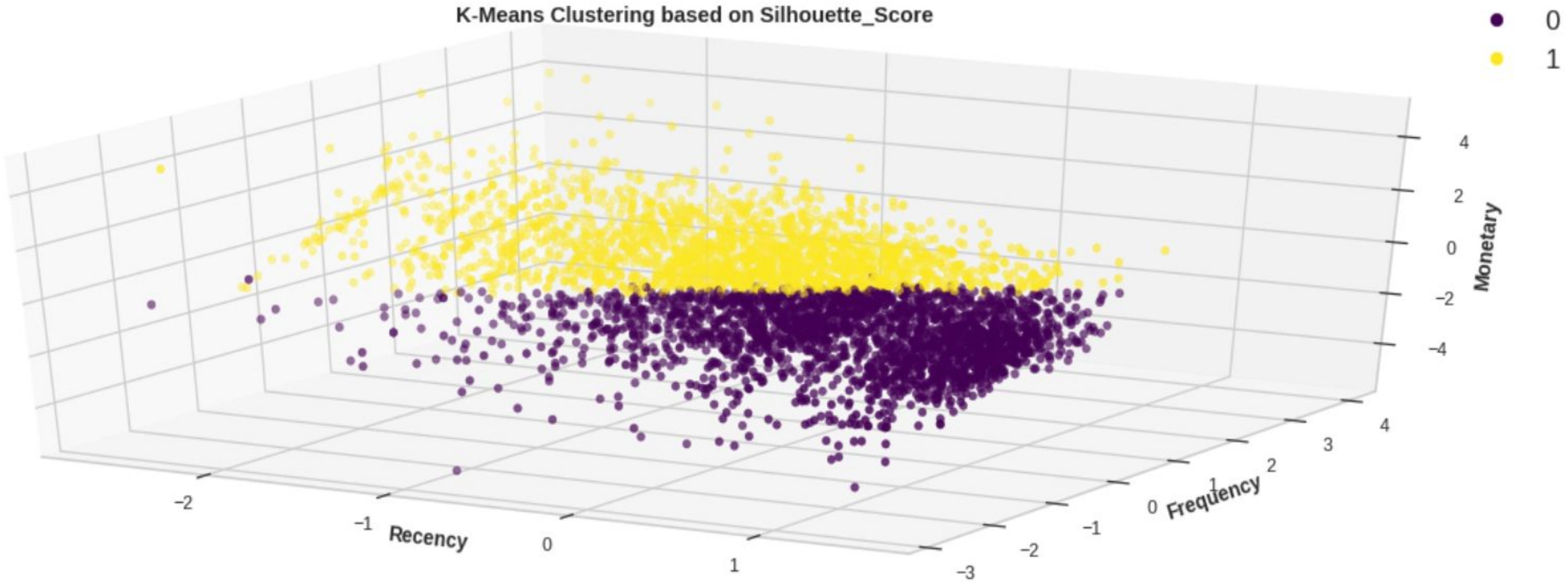
K-MEANS CLUSTERING





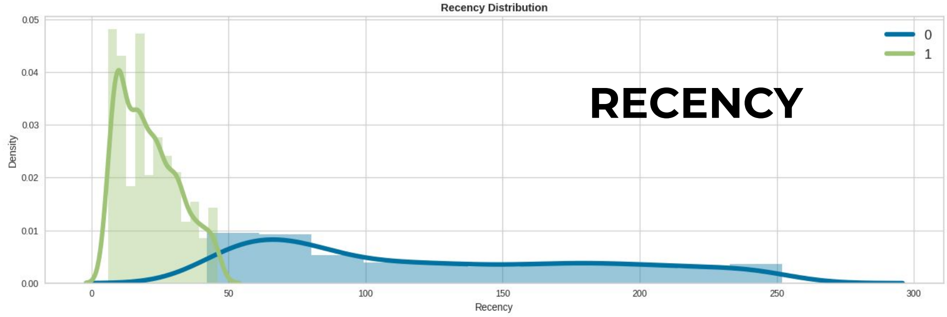
- From the Elbow curve 5 appears to be at the elbow and hence can be considered as the number of clusters. n_clusters=4 or 6 can also be considered.
- If we go by the maximum Silhouette Score as the criteria for selecting an optimal number of clusters, then n_clusters=2 can be chosen.
- If we look at both of the graphs at the same time to decide the optimal number of clusters, So 4 appears to be a good choice, having a decent Silhouette score as well as near the elbow of the elbow curve.

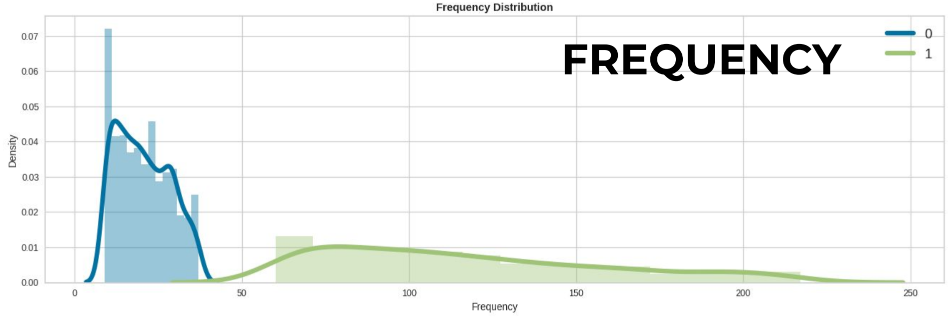
K-MEANS | 2CLUSTER

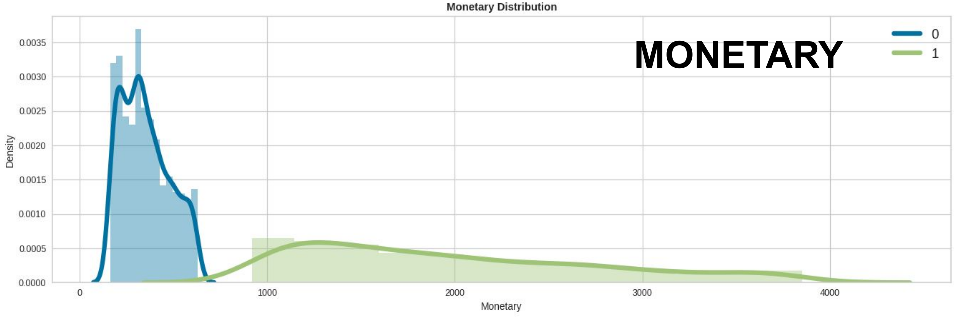


K-means|2cluster	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	31.282620	19.000000	172.146467	108.000000	4003.325535	1804.560000	1939
1	141.991667	110.500000	24.558333	19.000000	468.650701	330.070000	2400

K-MEANS | 2CLUSTER





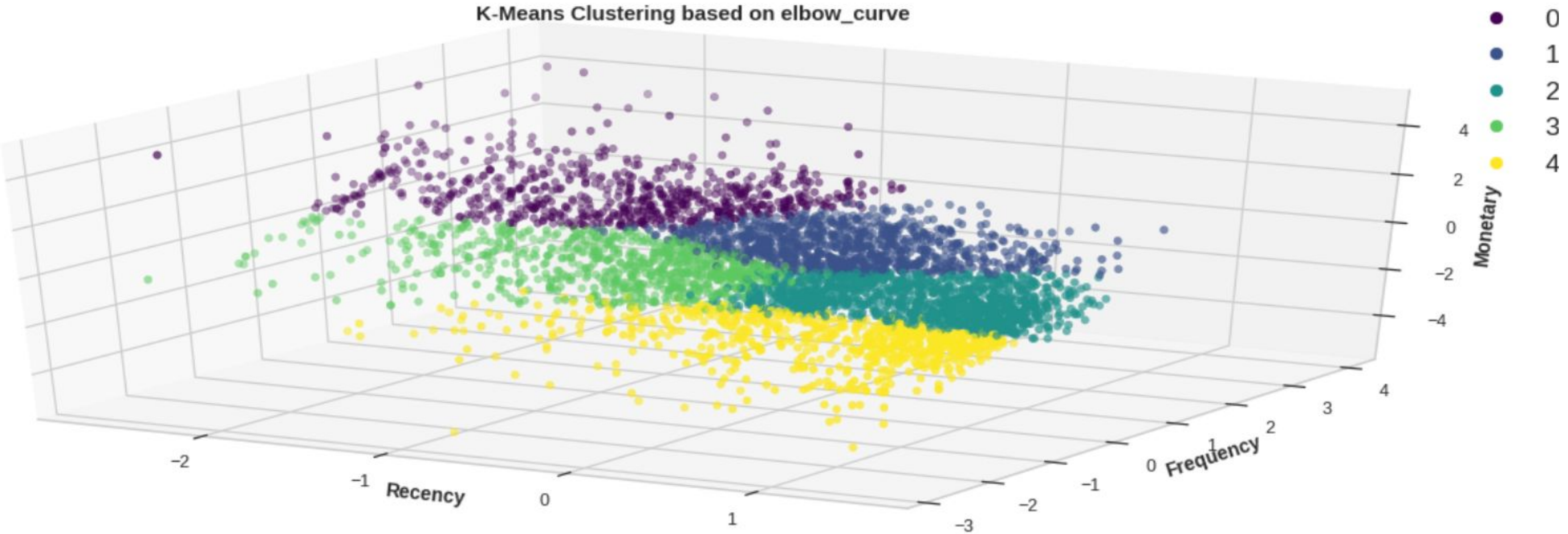


K-means|2cluster	Last_visited	purchase_frequency	Money_spend
0	8 to 39 days ago	Bought66 to 190 times	spent around 1058 to 3342 sterling
1	51 to 227 days ago	Bought10 to 33 times	spent around 187 to 562 sterling

GROUP 0 — BEST CUSTOMERS

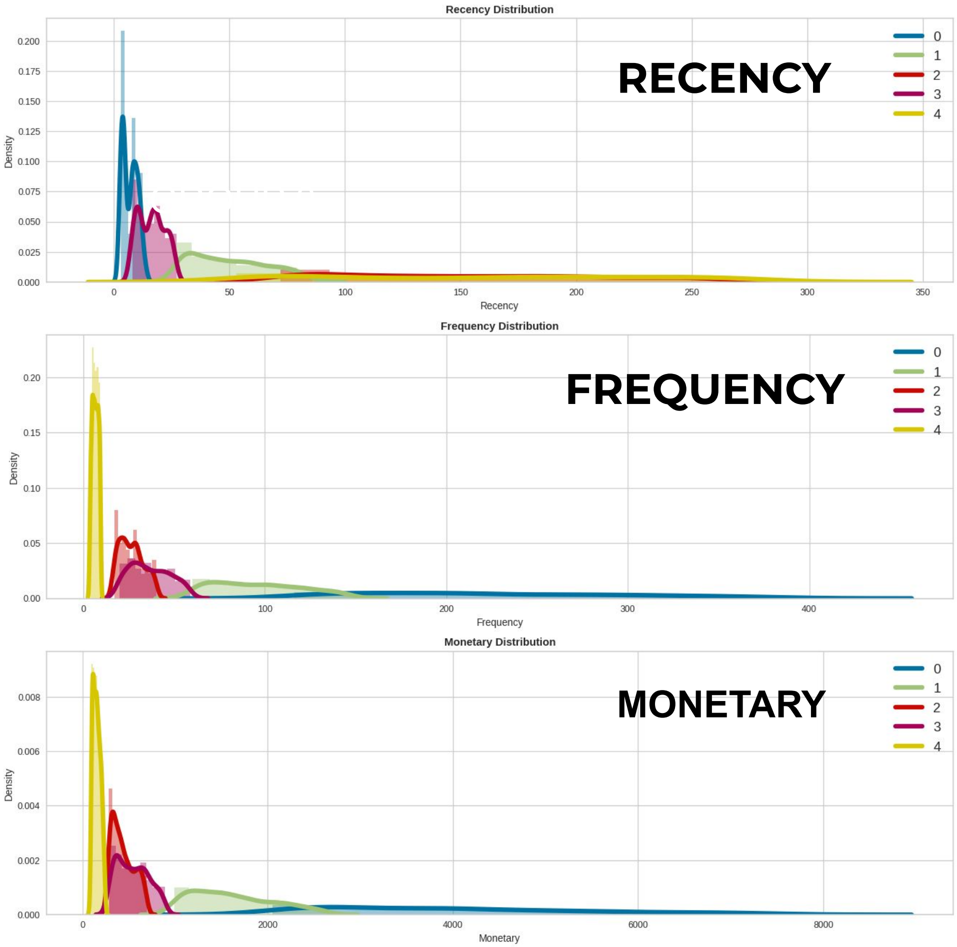
GROUP 1 — LOST POOR CUSTOMERS

K-MEANS | 5CLUSTER



K-means|5cluster	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	17.471895	17.000000	40.967320	35.000000	629.317361	523.110000	765
1	168.150042	152.000000	30.249790	26.000000	512.480094	414.870000	1193
2	62.805882	47.000000	109.121569	95.000000	2052.425148	1539.650000	1020
3	168.642241	165.000000	6.971264	7.000000	199.378966	152.600000	696
4	9.069173	7.000000	314.508271	212.000000	8364.138271	3799.490000	665

K-MEANS | 5CLUSTER



K-means|5cluster	Last_visited	purchase_frequency	Money_spend
0	3 to 12 days ago	Bought130 to 340 times	spent around 2298 to 6483 sterling
1	29 to 76 days ago	Bought63 to 135 times	spent around 1071 to 2320 sterling
2	78 to 246 days ago	Bought18 to 37 times	spent around 302 to 632 sterling
3	8 to 25 days ago	Bought21 to 53 times	spent around 327 to 807 sterling
4	61 to 264 days ago	Bought4 to 10 times	spent around 104 to 215 sterling

GROUP 0 — LOST POOR CUSTOMERS

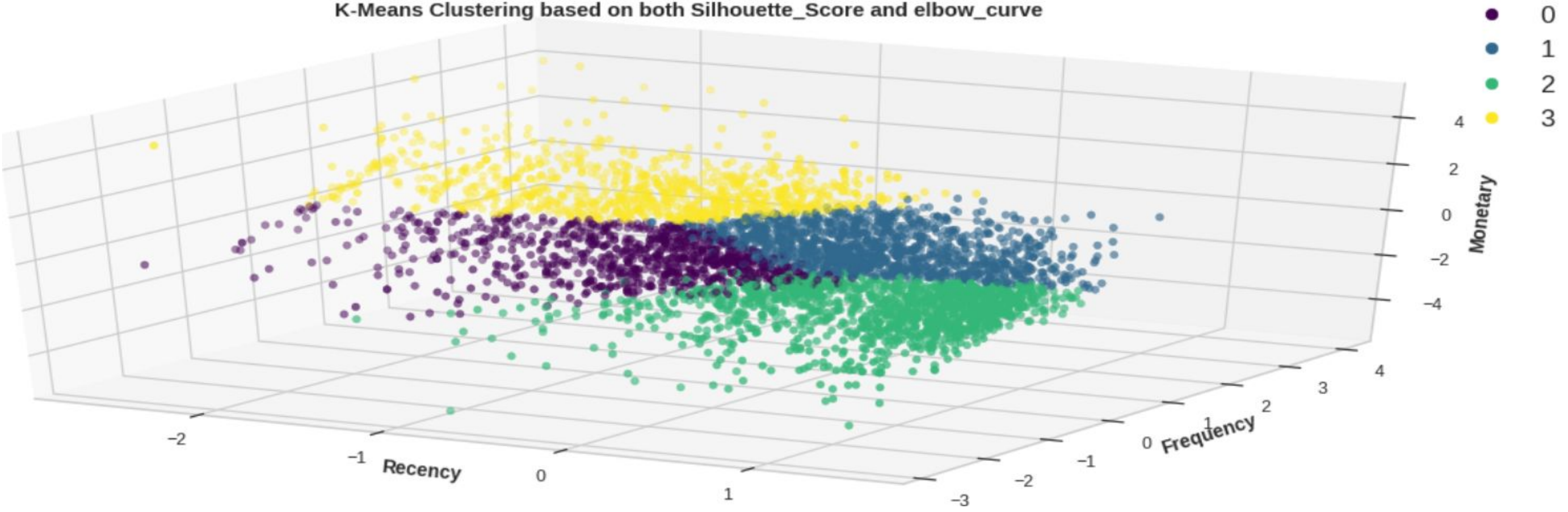
GROUP 1 — BEST CUSTOMERS

GROUP 2 — RECENTLY VISITED AVERAGE CUSTOMERS

GROUP 3 — LOSING LOYAL CUSTOMERS

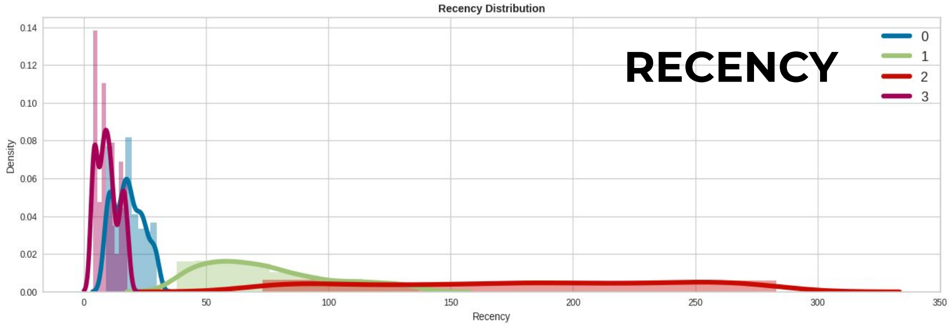
GROUP 4 — AVERAGE CUSTOMERS

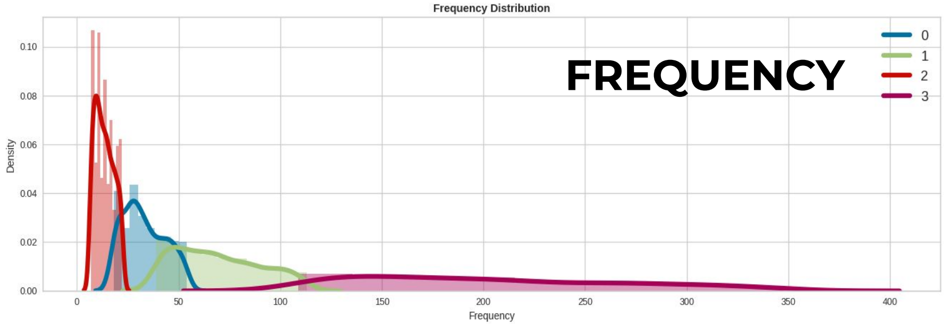
K-MEANS | 4CLUSTER

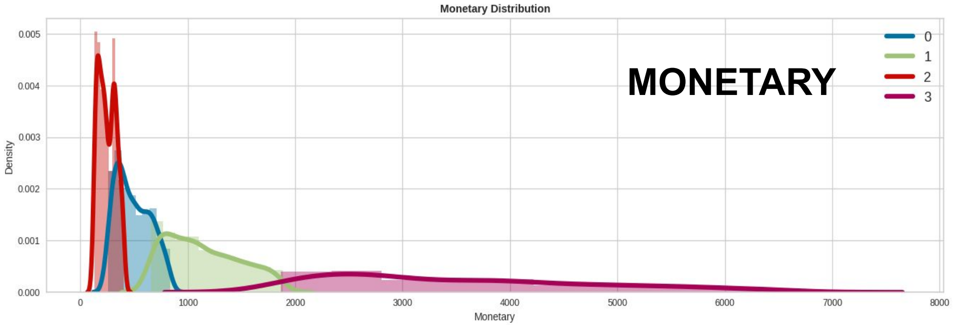


K-means|4cluster	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	184.750364	185.000000	14.724891	12.000000	295.959819	240.275000	1374
1	12.136364	9.000000	283.193780	192.500000	7205.348792	3316.310000	836
2	19.645509	17.000000	38.350898	32.000000	589.801401	470.760000	835
3	93.539413	71.000000	80.159969	66.000000	1518.087591	1083.840000	1294

K-MEANS | 4CLUSTER







K-means|4cluster	Last_visited	purchase_frequency	Money_spend
0	82 to 268 days ago	Bought7 to 21 times	spent around 144 to 368 sterling
1	4 to 17 days ago	Bought120 to 308 times	spent around 2068 to 5601 sterling
2	9 to 28 days ago	Bought20 to 51 times	spent around 291 to 742 sterling
3	43 to 119 days ago	Bought42 to 103 times	spent around 708 to 1706 sterling

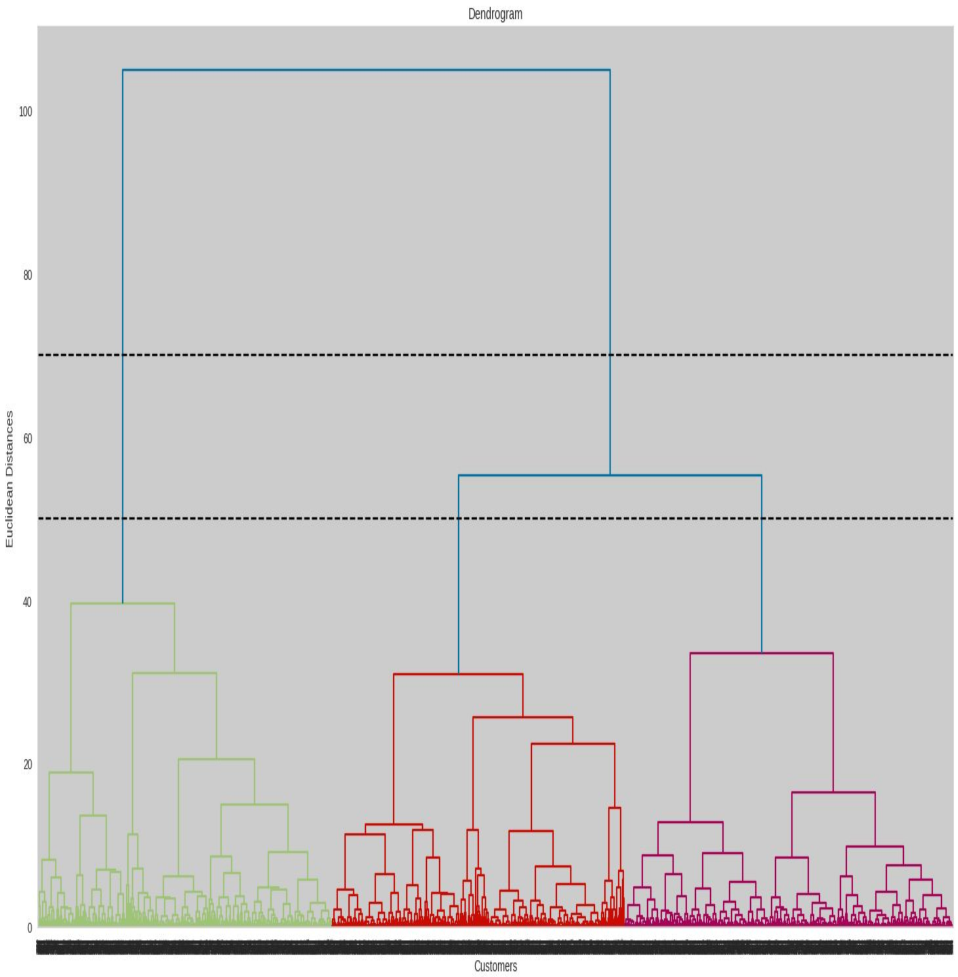
GROUP 0 — LOSING LOYAL CUSTOMERS

GROUP 1 — BEST CUSTOMERS

GROUP 2 — LOST POOR CUSTOMERS

GROUP 3 — RECENTLY VISITED AVERAGE CUSTOMERS

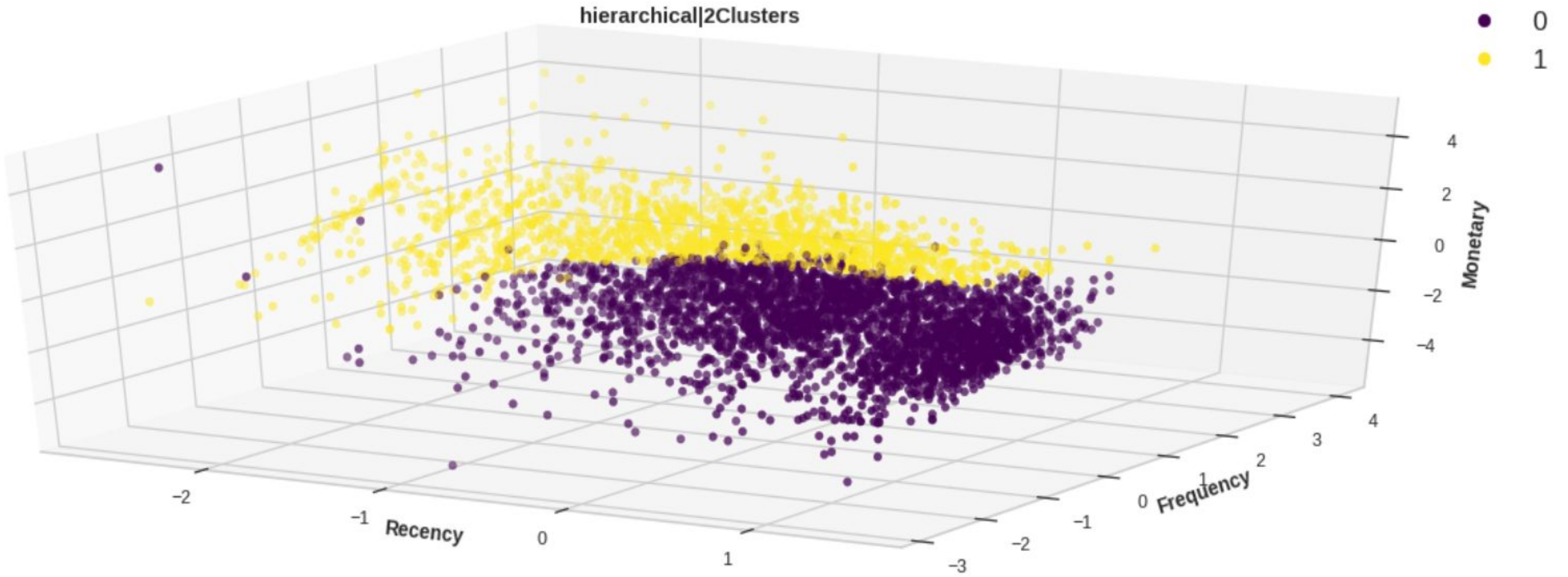
HIERARCHICAL CLUSTERING



In the K-means clustering there is a challenge to predetermine the number of clusters, and it always tries to create the clusters of the same size. To solve these two challenges, we can opt for the hierarchical clustering algorithm because, in this algorithm, we don't need to have knowledge about the predefined number of clusters. Hierarchical clustering is based on two techniques:

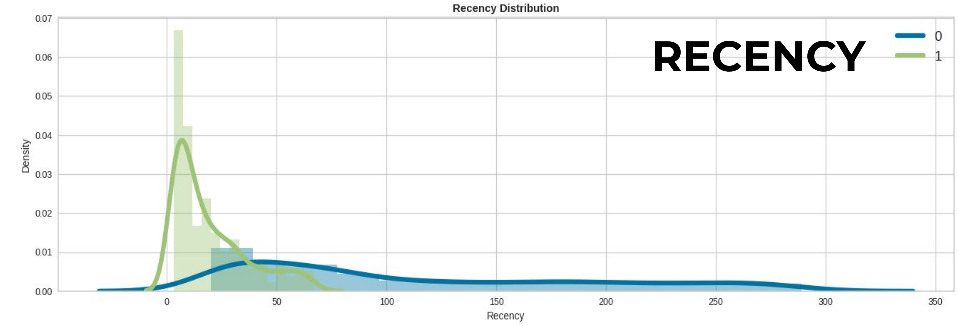
- Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
- Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

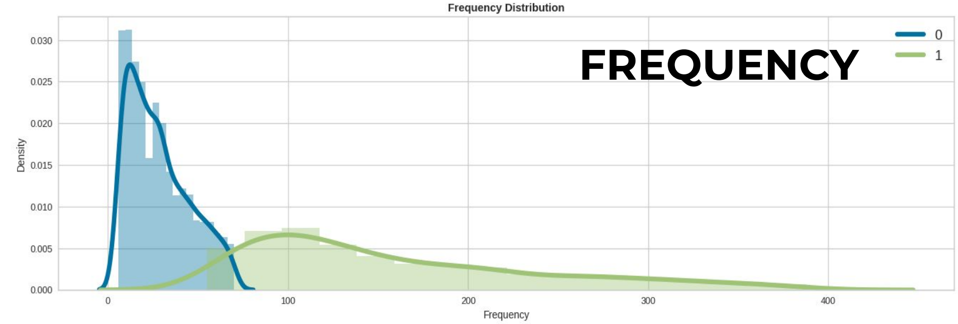
HIERARCHICAL | 2CLUSTER

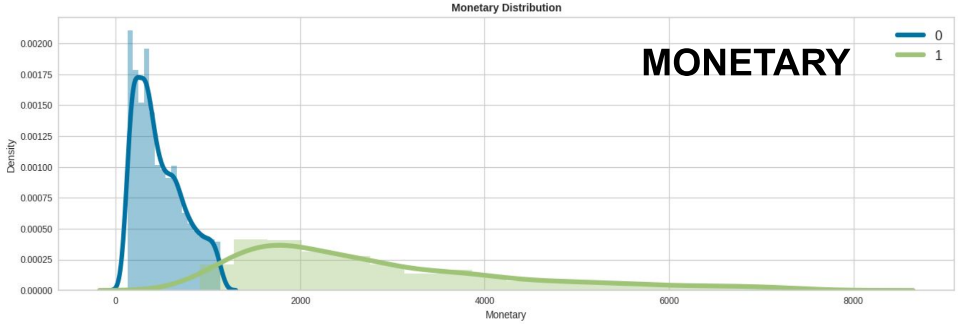


hierarchical|2Clusters	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	123.608696	80.000000	33.610394	25.000000	684.108391	409.685000	2944
1	26.905376	12.000000	210.597133	135.000000	4927.021356	2404.170000	1395

HIERARCHICAL | 2CLUSTER





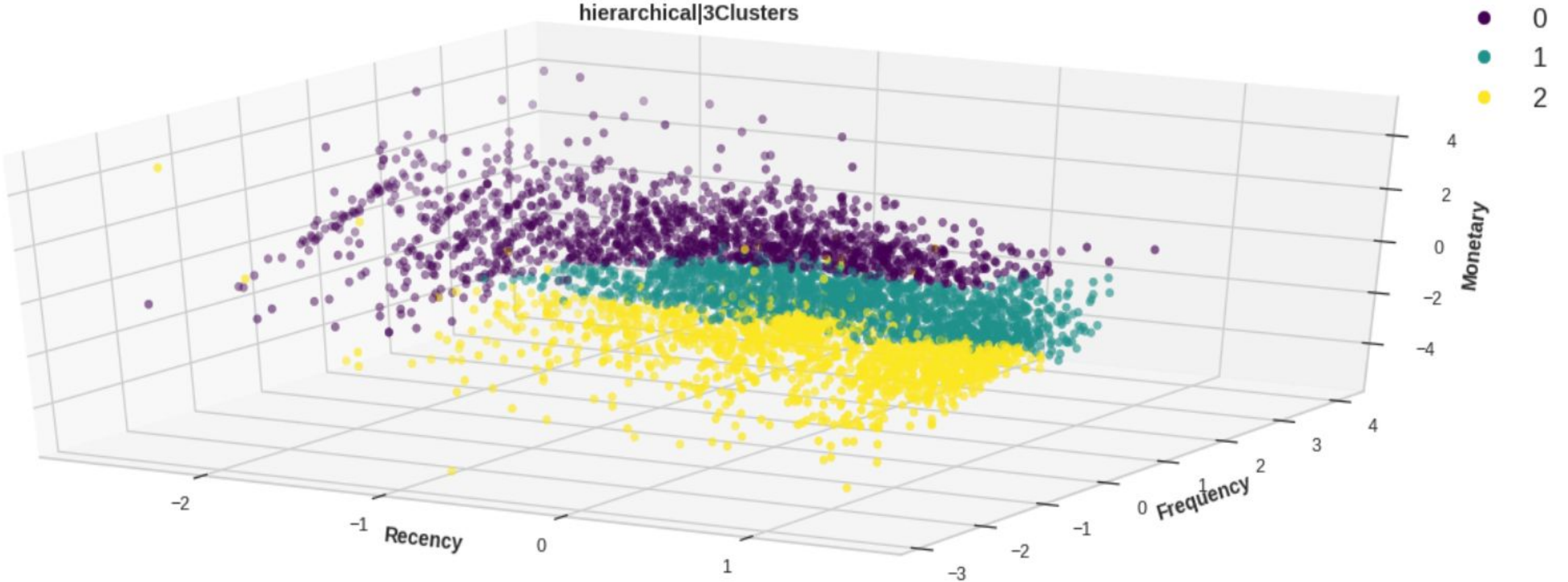


hierarchical|2Clusters	Last_visited	purchase_frequency	Money_spend
0	36 to 200 days ago	Bought12 to 46 times	spent around 219 to 739 sterling
1	4 to 33 days ago	Bought87 to 233 times	spent around 1546 to 4020 sterling

GROUP 0 → AVERAGE CUSTOMERS

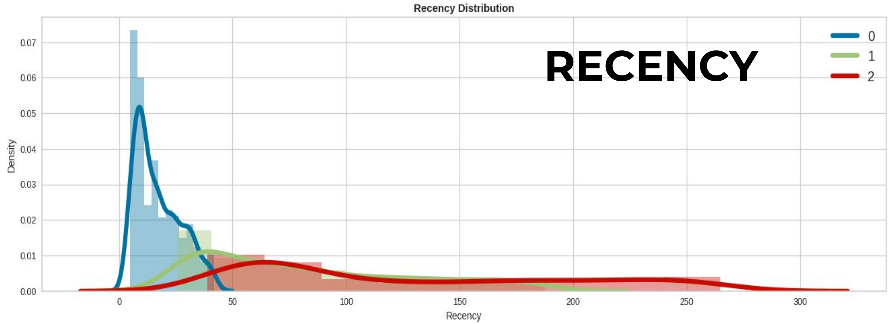
GROUP 1 → BEST CUSTOMERS

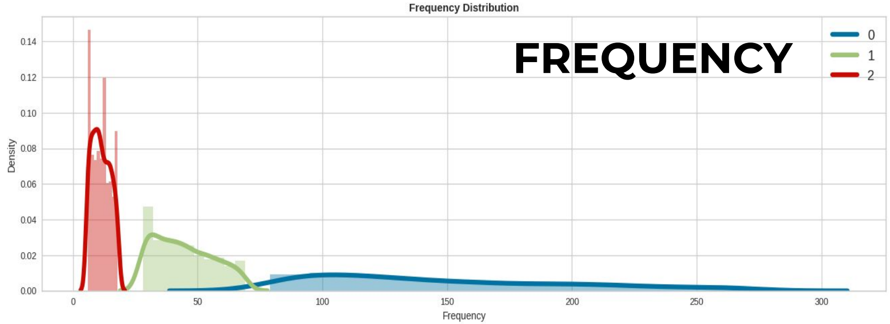
HIERARCHICAL | 3CLUSTER

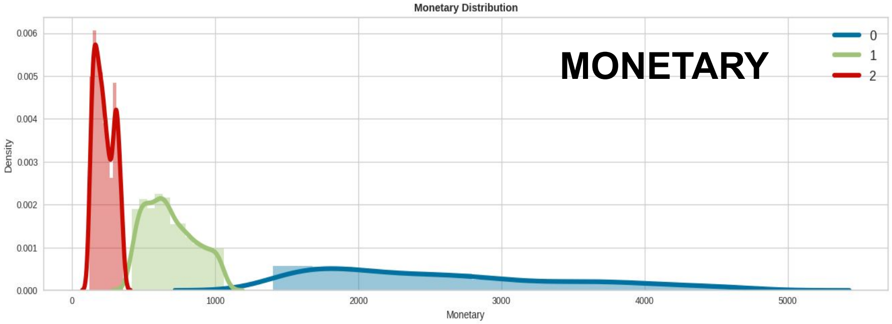


hierarchical|3Clusters	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	26.905376	12.000000	210.597133	135.000000	4927.021356	2404.170000	1395
1	105.246312	71.000000	51.658114	43.000000	756.610450	657.300000	1559
2	144.277978	99.000000	13.295307	11.000000	602.497770	215.480000	1385

HIERARCHICAL | 3CLUSTER







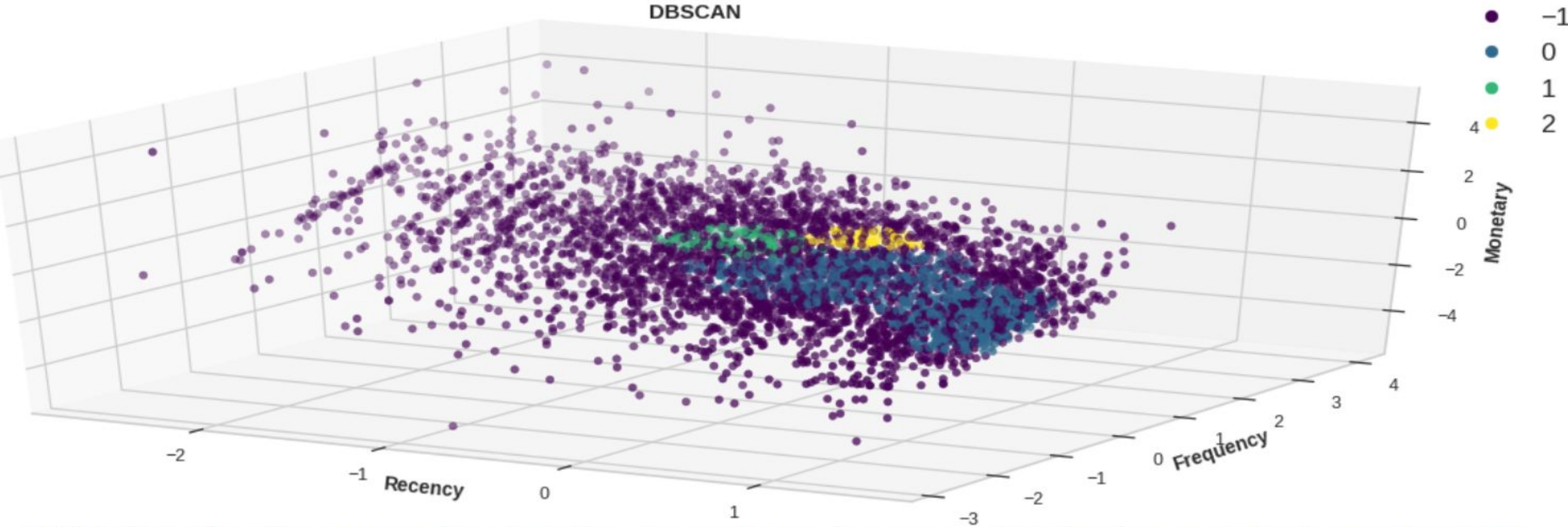
hierarchical|3Clusters	Last_visited	purchase_frequency	Money_spend
0	4 to 33 days ago	Bought87 to 233 times	spent around 1546 to 4020 sterling
1	30 to 163 days ago	Bought29 to 63 times	spent around 463 to 977 sterling
2	49 to 243 days ago	Bought6 to 18 times	spent around 139 to 327 sterling

GROUP 0 — BEST CUSTOMERS

GROUP 1 — LOSING LOYAL CUSTOMERS

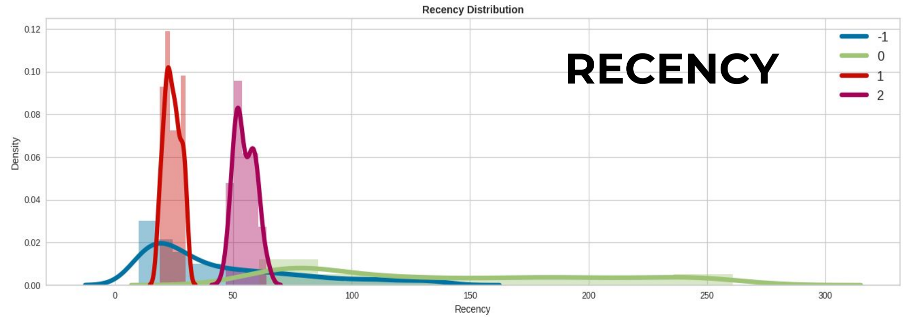
GROUP 2 — LOST POOR CUSTOMERS

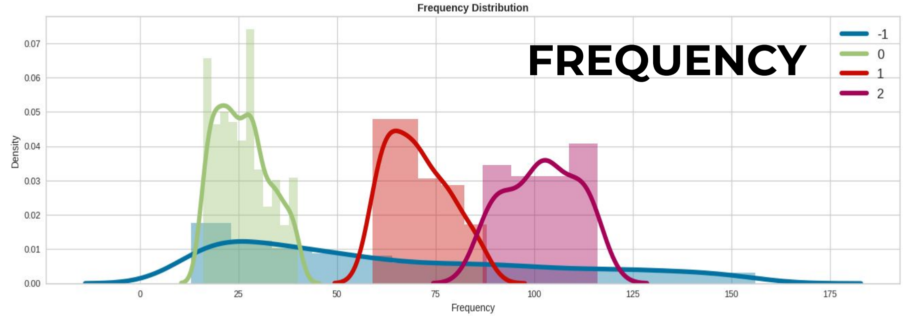
DBSCAN

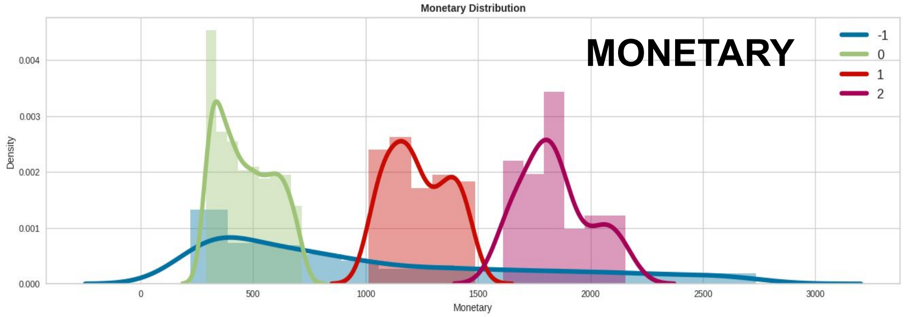


DBSCAN Clustering	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
-1	76.209745	32.000000	111.290416	52.000000	2591.443756	797.960000	3099
0	155.415430	124.000000	28.556874	25.000000	514.588776	451.440000	1011
1	25.253247	24.500000	73.376623	69.500000	1264.077597	1212.910000	154
2	56.653333	54.000000	102.293333	102.000000	1885.446533	1824.230000	75

DBSCAN







DBSCAN Clustering	Last_visited	purchase_frequency	Money_spend
-1	11 to 107 days ago	Bought15 to 130 times	spent around 278 to 2245 sterling
0	66 to 245 days ago	Bought17 to 37 times	spent around 309 to 658 sterling
1	19 to 30 days ago	Bought60 to 84 times	spent around 1052 to 1425 sterling
2	50 to 62 days ago	Bought89 to 114 times	spent around 1637 to 2094 sterling

GROUP 1 — AVERAGE CUSTOMERS

GROUP 2 — LOST POOR CUSTOMERS

GROUP 3 — GOOD CUSTOMERS

GROUP 4 — LOSING LOYAL CUSTOMERS

SUMMARY

clusterer	Binning	Quantile cut	k-means	k-means	k-means	Agglomerative	Agglormerative	DBScan
criterion	RFM Score Binning	RFM quantile Cut	Elbow Curve	silhouette Score	Elbow Curve & Silhouette Score	Dendogram (y=70)	Dendogram (y=50)	eps=0.2, min_samples=0.2
Segments	4	4	5	2	4	3	2	4

1 We started with a simple binning and quantile based simple segmentation model first then moved to more complex models because simple implementation helps having a first glance at the data and know where/how to exploit it better.
2 Then we moved to k-means clustering and visualized the results with different number of clusters. As we know there is no assurance that k-means will lead to the global best solution. We moved forward and tried Hierarchical Clustering and DBSCAN clusterer as well.
3 We created several useful clusters of customers on the basis of different metrics and methods to categorize the customers on the basis of their behavioural attributes to define their volubility, loyalty, profitability etc for the business. Though significantly separated clusters are not visible in the plots, but the clusters obtained is fairly valid and useful as per the algorithms and the statistics extracted from the data.
4 Segments depends on how the business plans to use the results, and the level of granularity they want to see in the clusters. Keeping these points in view we clustered the major segments based on our understanding as per different criteria as shown in the summary dataframe.

FINAL CONCLUSION

CUSTOMER SEGMENTS OBTAINED FROM CLUSTERING ANALYSIS

	LOST POOR CUSTOMERS❌	AVERAGE CUSTOMERS🥉	RECENTLY VISITED AVERAGE CUSTOMERS🖤	GOOD CUSTOMERS🥈	BEST CUSTOMERS❤️	LOSING LOYAL CUSTOMERS❌
Binning	Yes	Yes	No	Yes	Yes	No
QuantileCut	Yes	No	No	Yes	Yes	Yes
K-means|2cluster	Yes	No	No	No	Yes	No
K-means|4cluster	Yes	No	Yes	No	Yes	Yes
K-means|5cluster	Yes	Yes	Yes	No	Yes	Yes
hierarchical|2Clusters	No	Yes	No	No	Yes	No
hierarchical|3Clusters	Yes	No	No	No	Yes	Yes
DBSCAN	Yes	Yes	No	Yes	No	Yes