UNSUPERVISED MACHINE LEARNING (CUSTOMER SEGMENTATION) ONLINE RETAIL

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

Data Description

Attribute	Data Type	Description
Invoice Number	Nominal	6-digit unique number / code starts with letter 'c', it indicates a cancellation
Stock Code	Nominal	a 5-digit unique number assigned to each distinct product.
Description	Nominal	Product (item) name
Quantity	Numeric	Quantities of each product (item) per transaction
Invoice Date	Numeric	Date and time when each transaction was generated
Unit Price	Numeric	Product price per unit in sterling.
CustomerID	Nominal	5-digit unique number for Customer
Country	Nominal	the name of the country where each customer resides.

IMPORTING AND INSPECTING DATASET

Data set Name:- Online Retail

No of Observation:541908 (shape=8x541908)

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

Data Cleaning

Checking Missing data

1. 25 % of items (i.e 135080)purchased are not assigned to Customers
2. Products – 1454 (0.27% Missing Values)

No use of this data it can be dropped

Checking duplicates

5268 data points were duplicated

Dropped duplicates

Total data points left

No of Observation left :401604 (shape=8x 401604)

FEATURE ENGINEERING

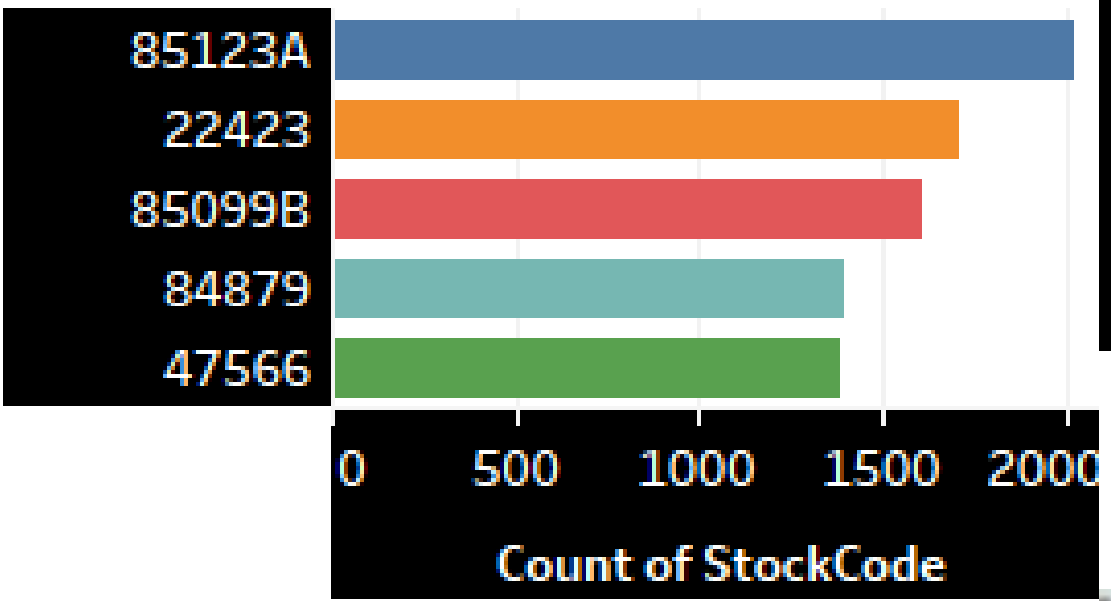
Extracting Year Date and Month from Invoice Date

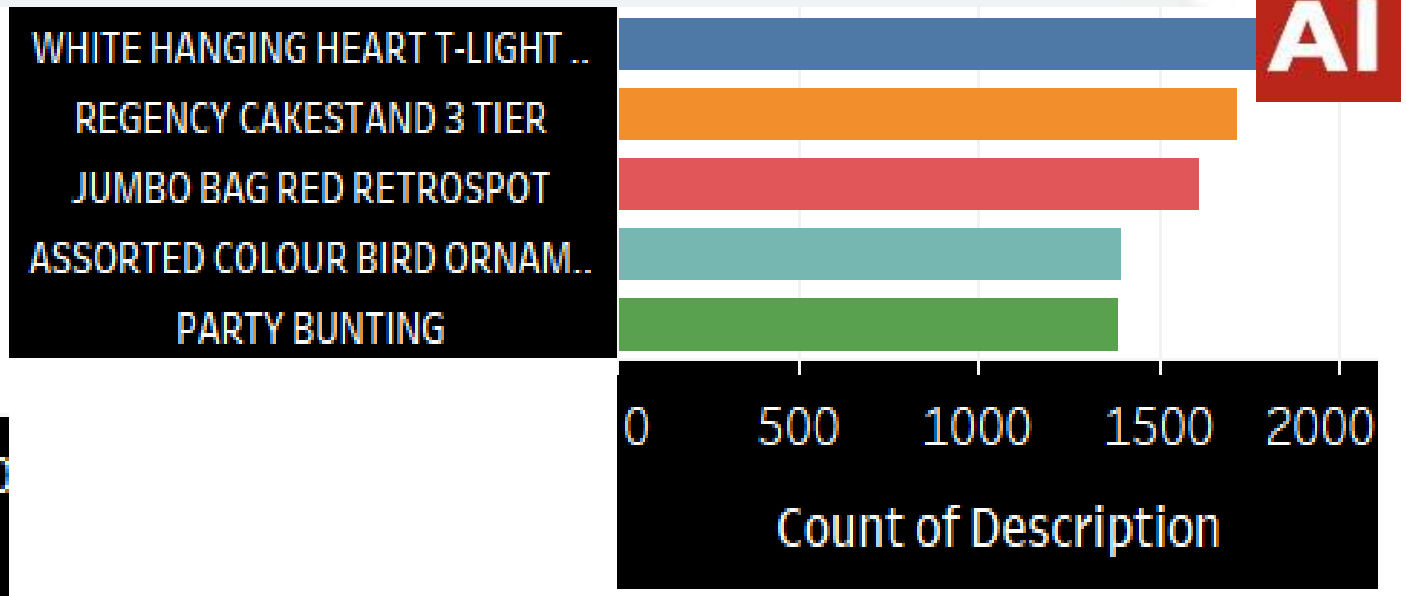
Added Feature 'TotalAmount' by multiplying values from the **Quantity** and **UnitPrice** column.(Sterling)

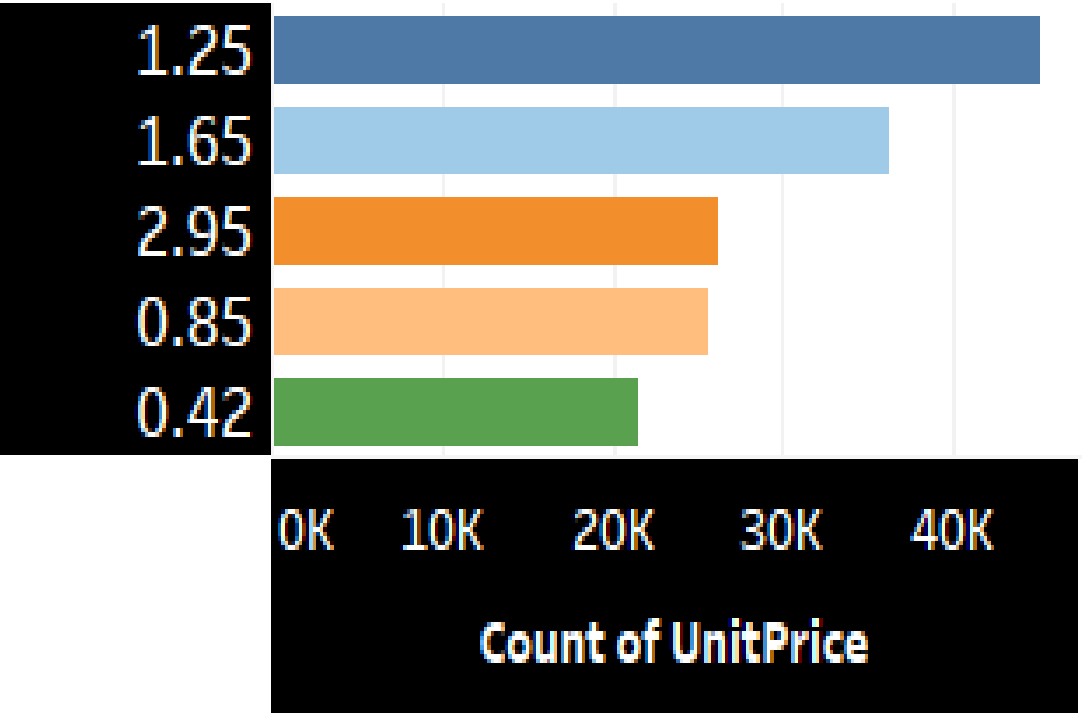
Added feature 'TimeType' based on hours to define whether its Morning, Afternoon, or Evening

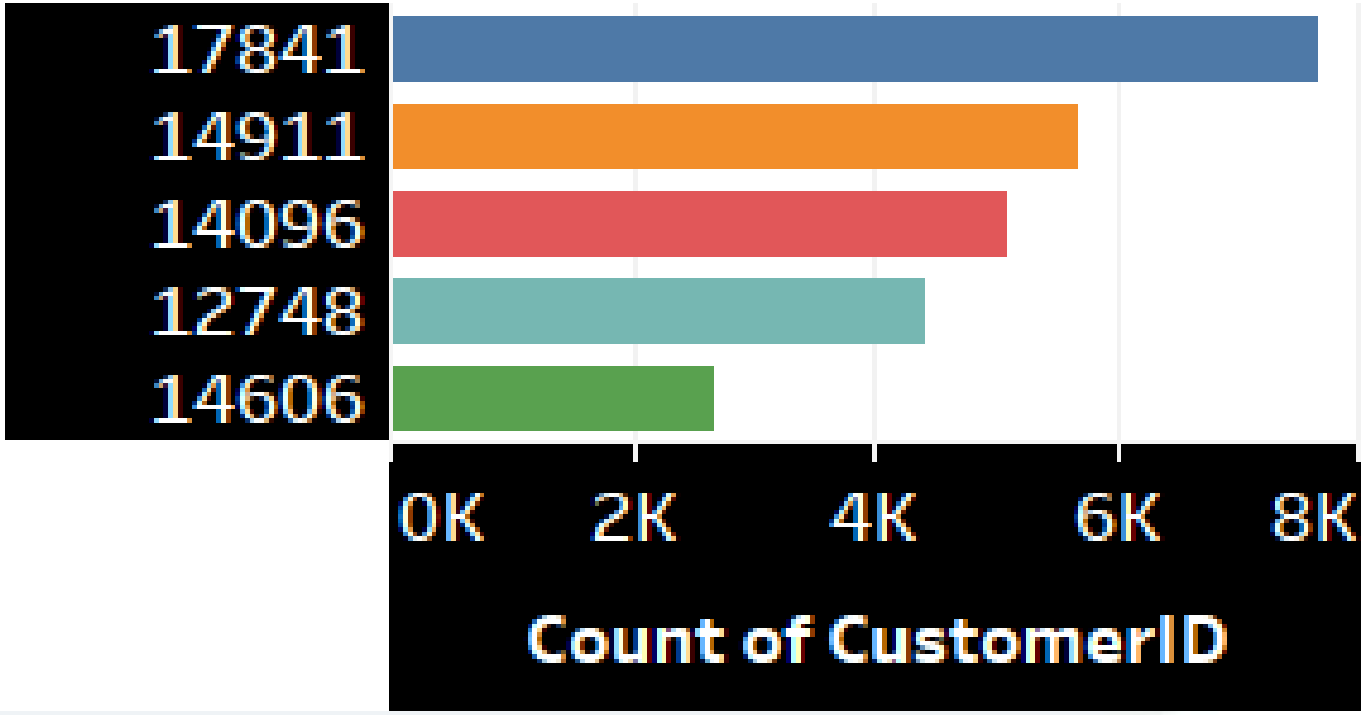
Dropping InvoiceNo starting with 'C' that represents cancellation

TOP FREQUENT VALUES

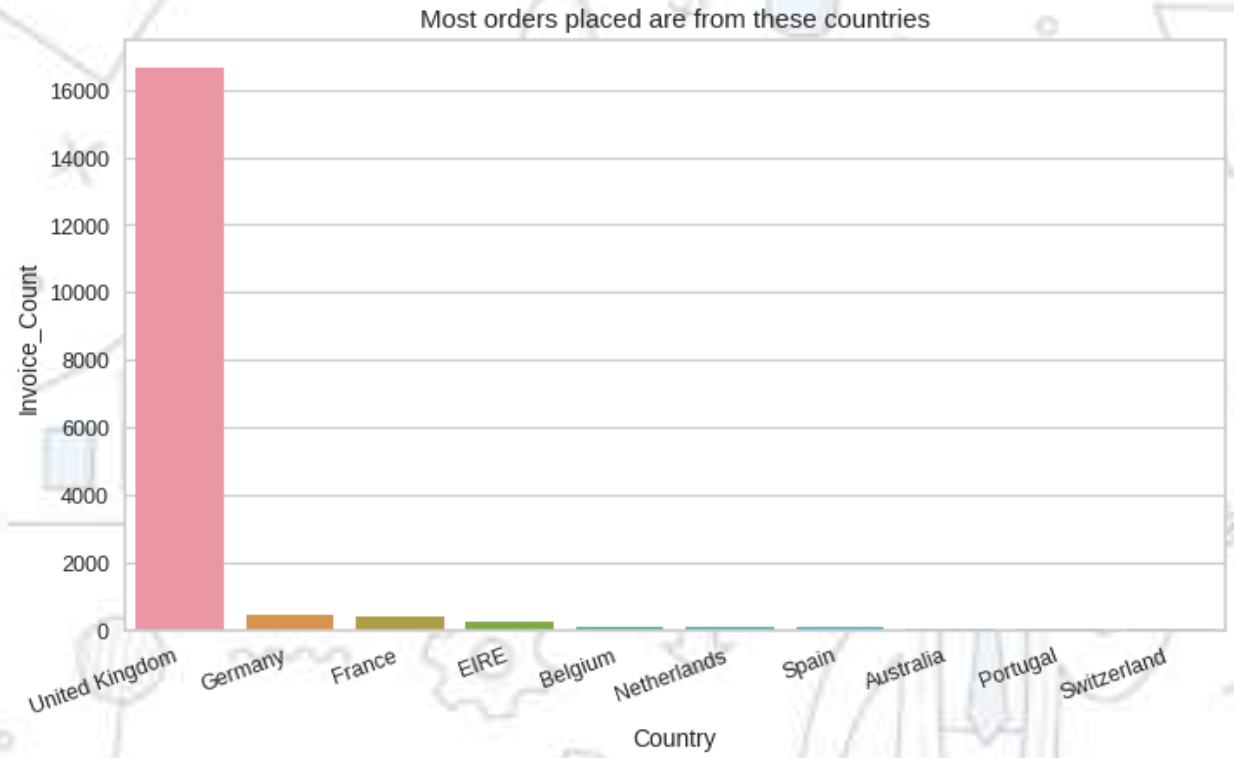


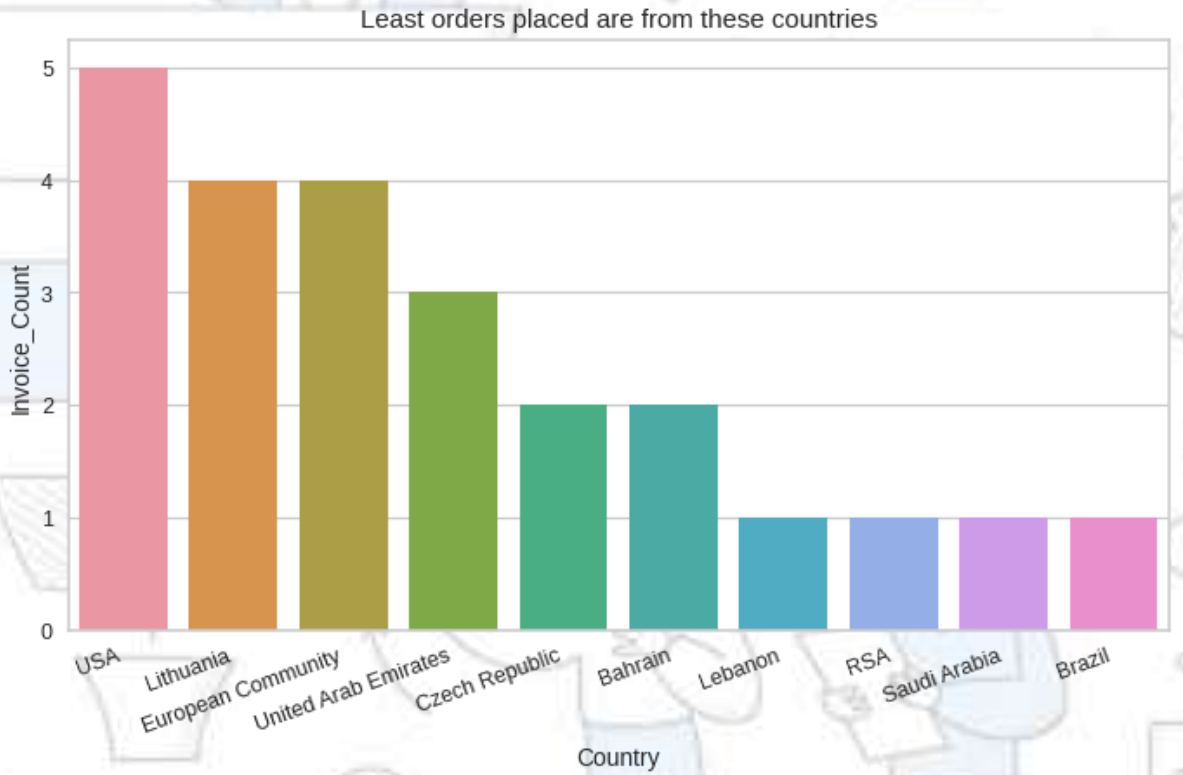




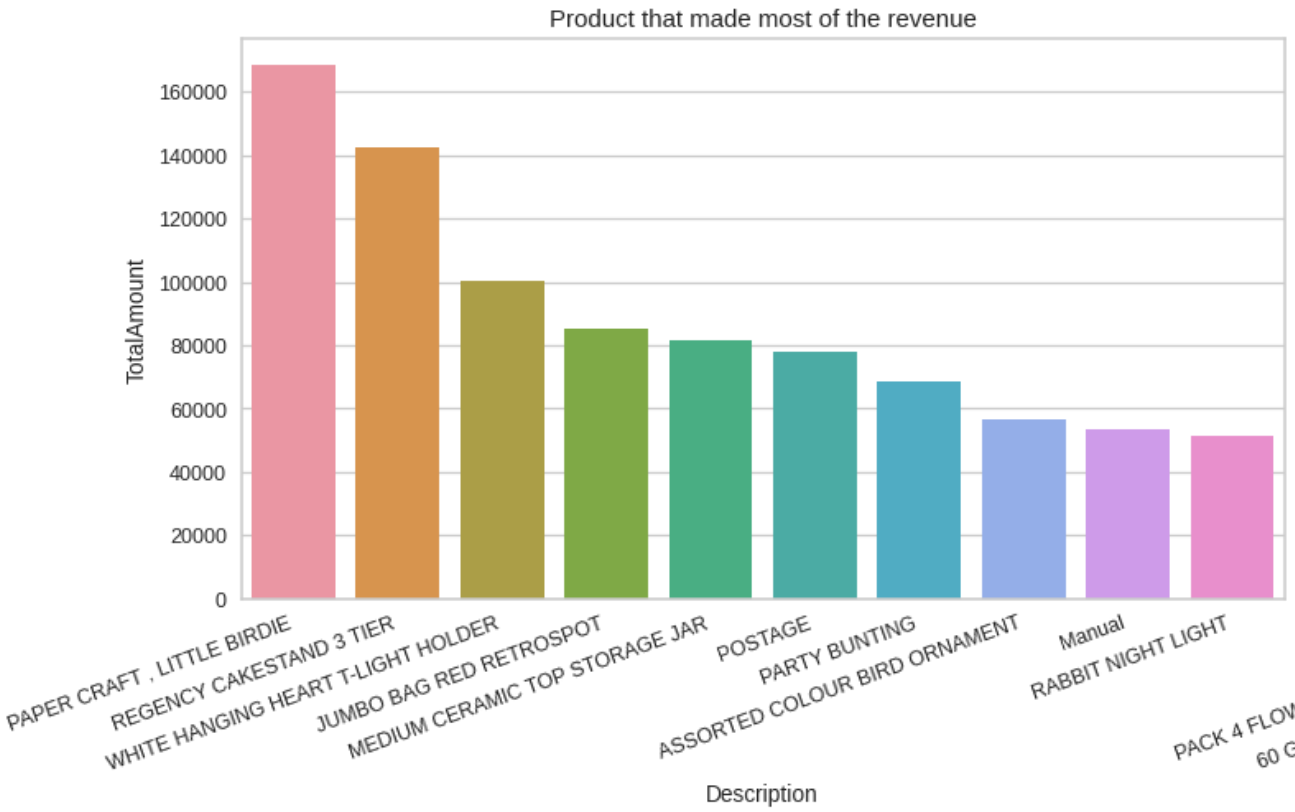


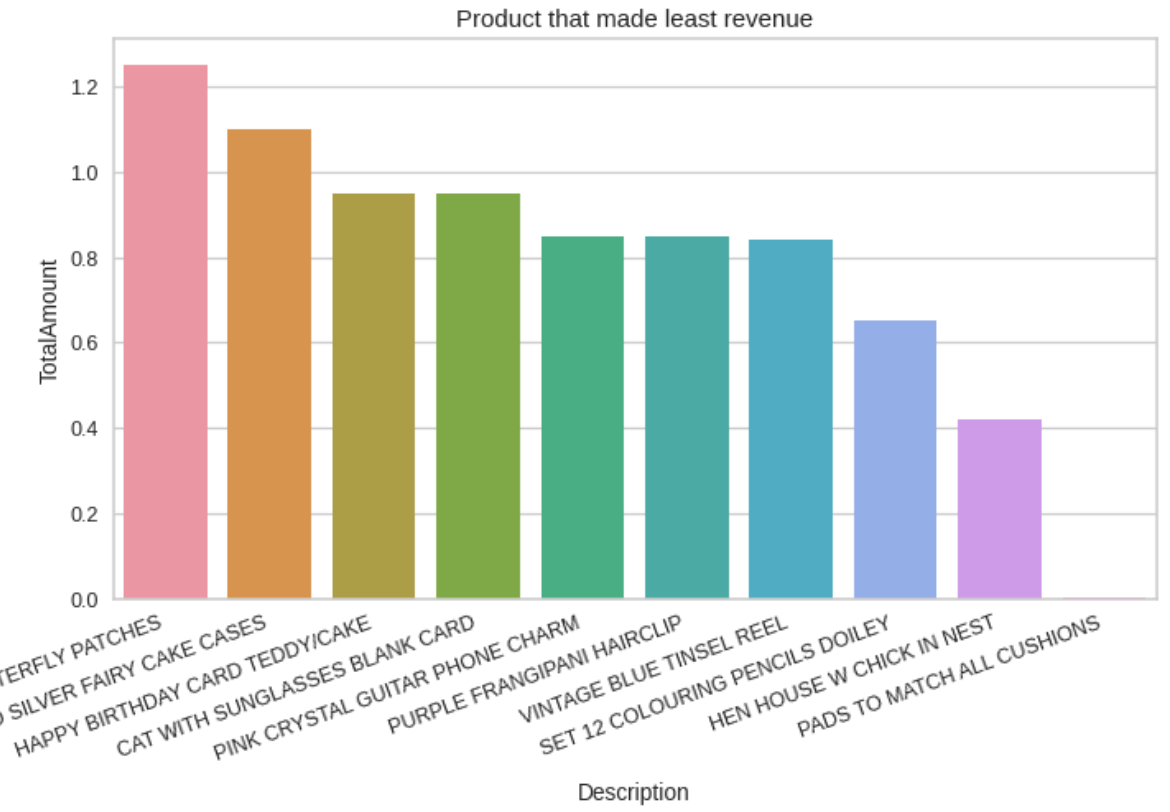
COUNTRY WISE ORDERS



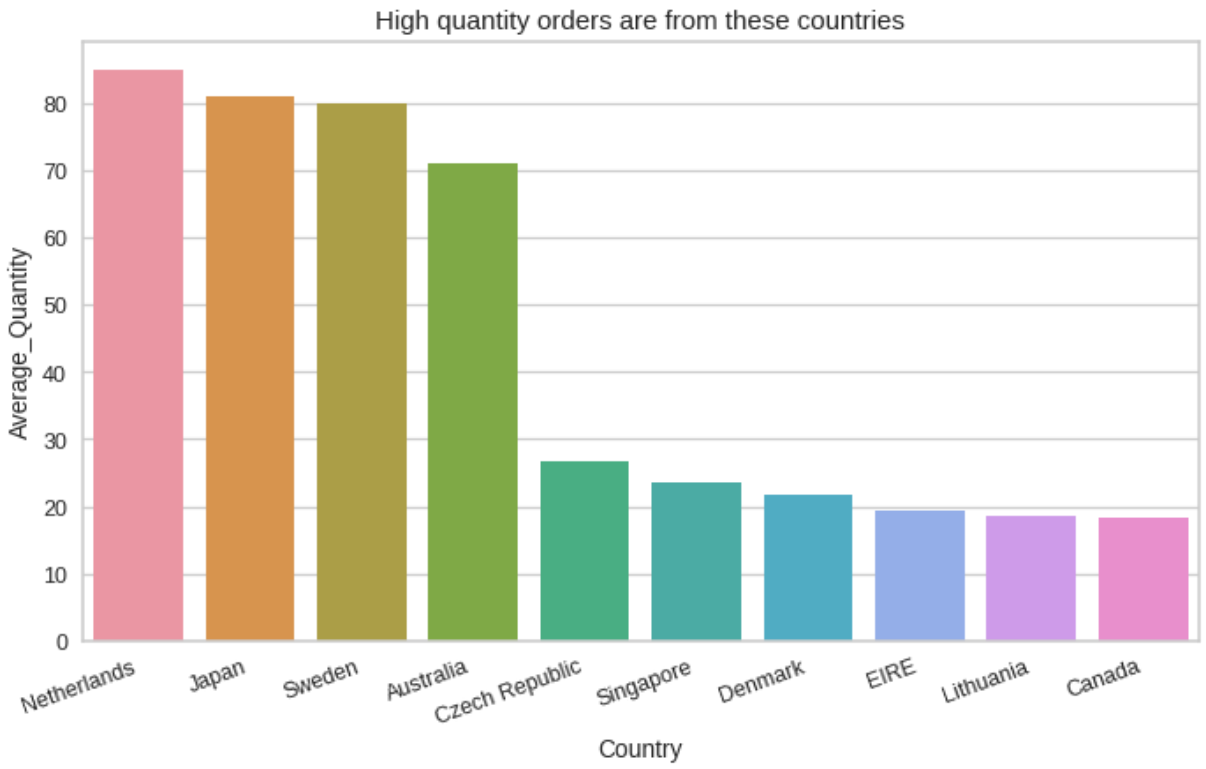


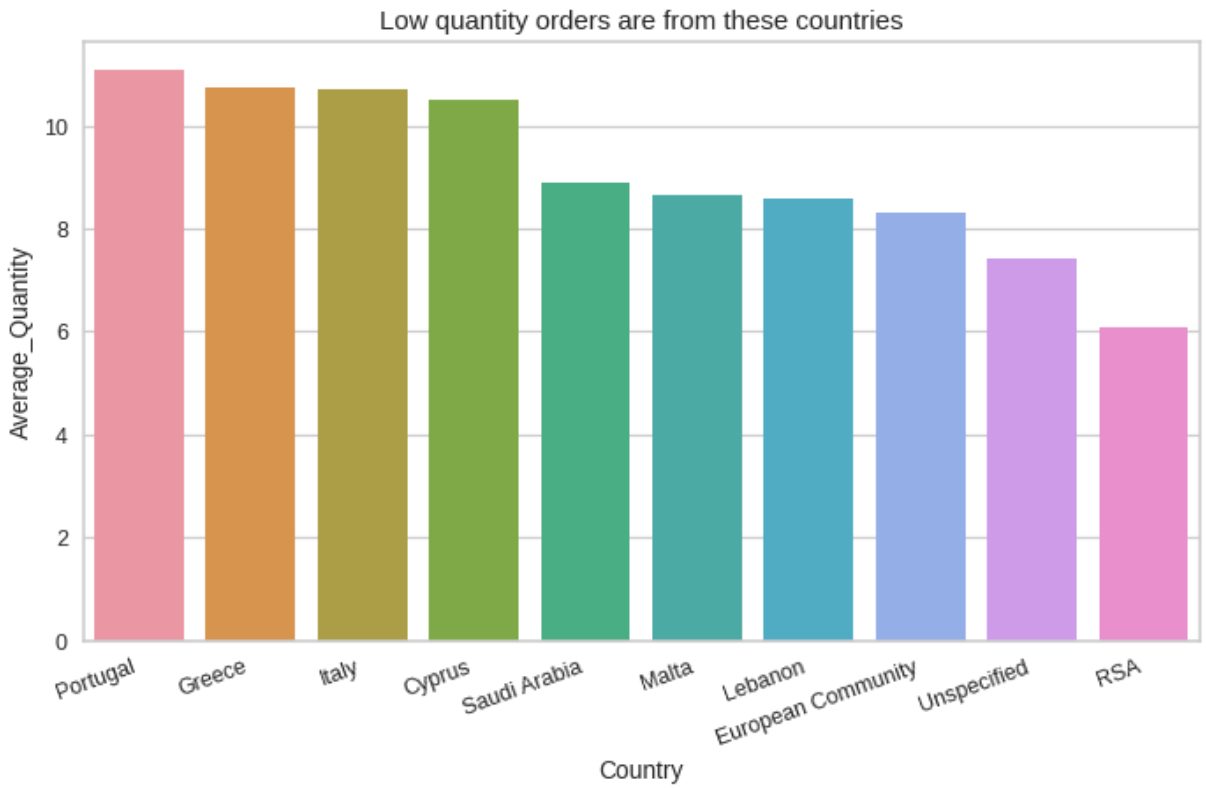
PRODUCT WISE REVENUE



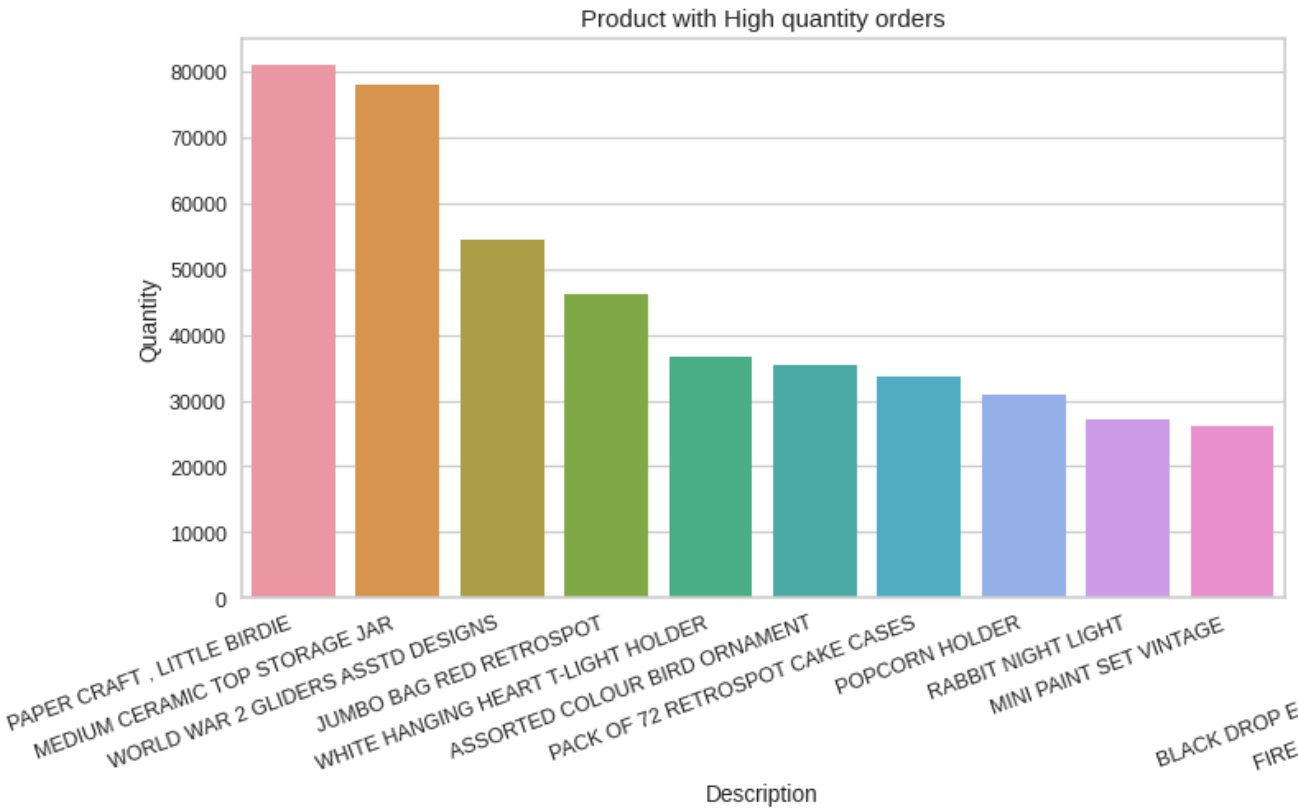


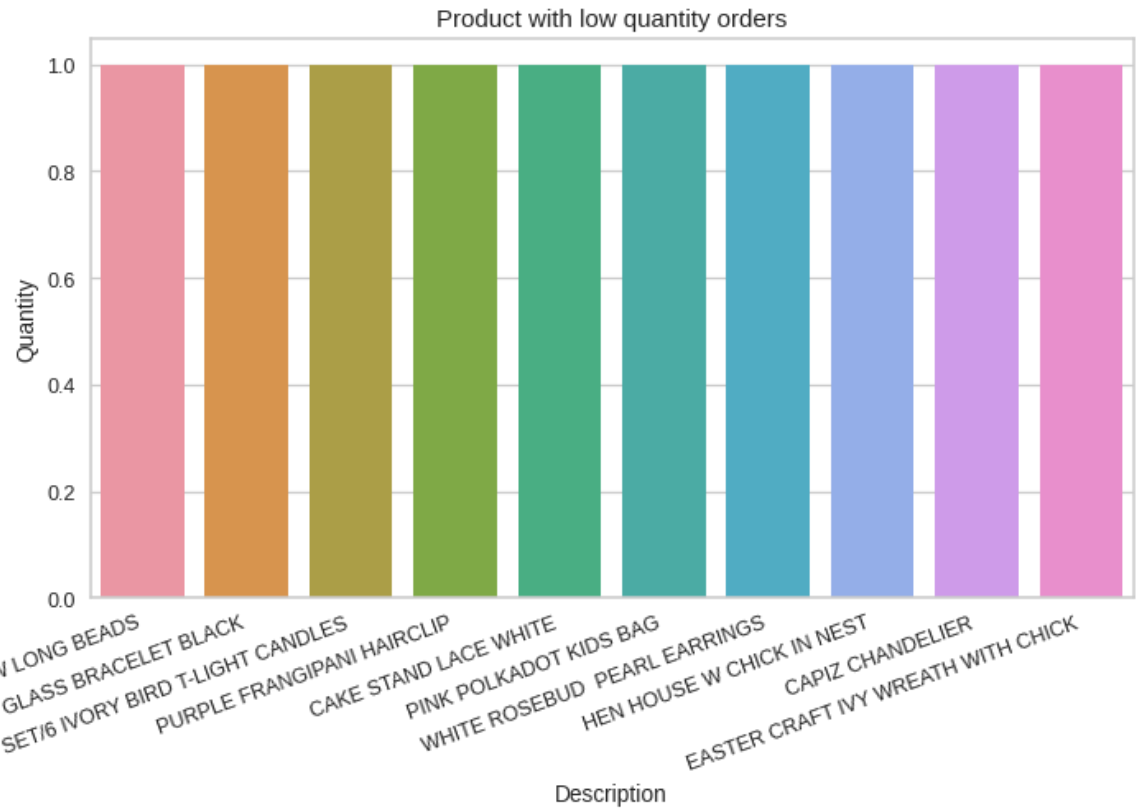
COUNTRY WISE PURCHASE QUANTITY



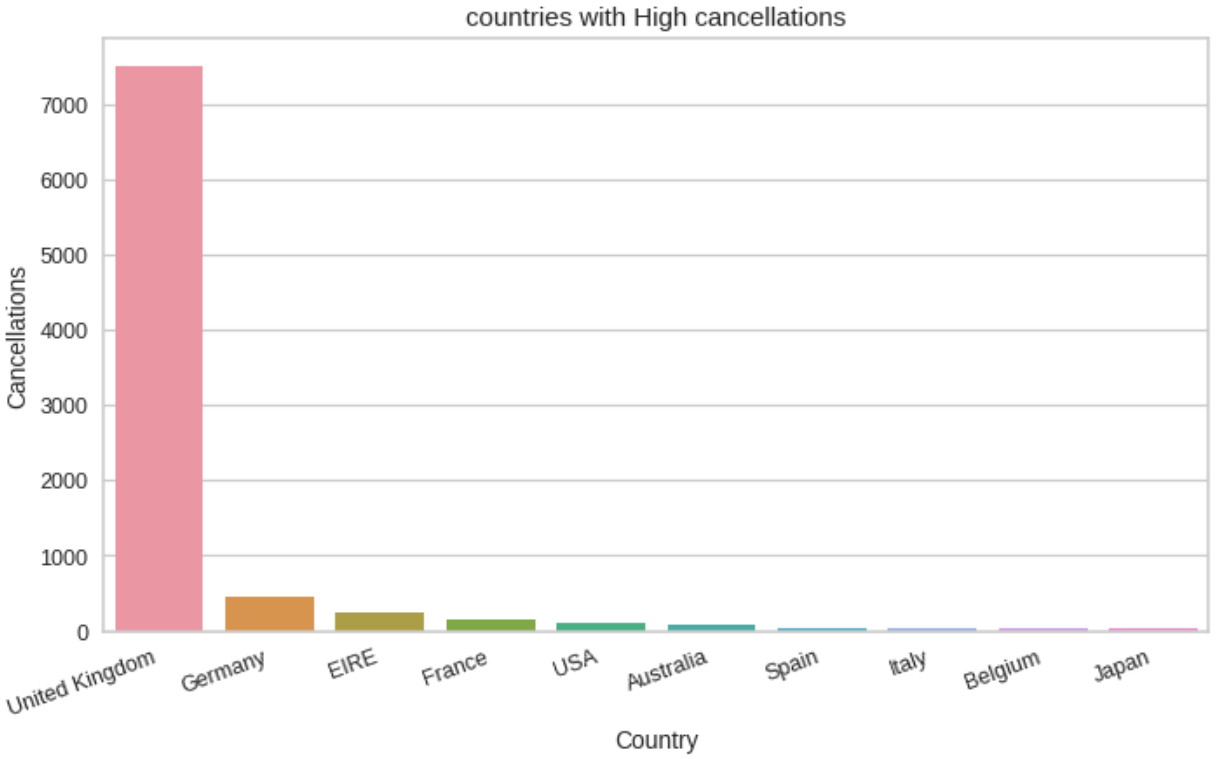


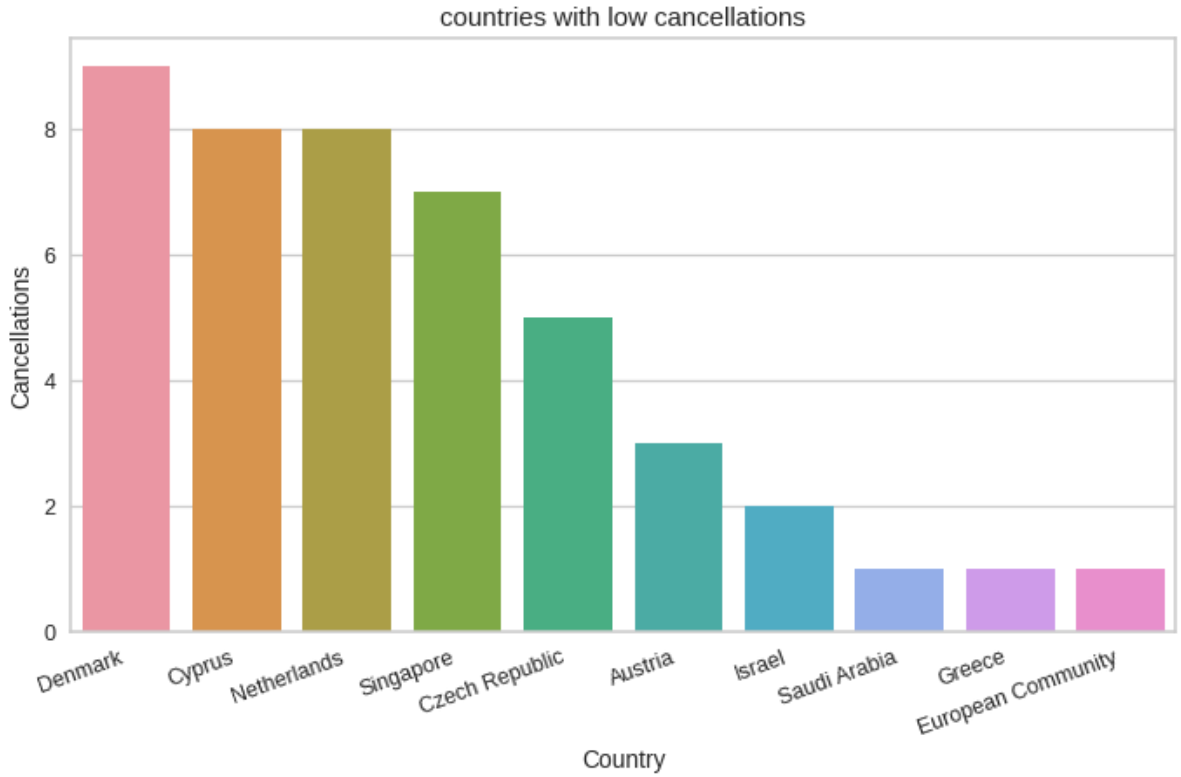
PRODUCT WISE PURCHASE QUANTITY



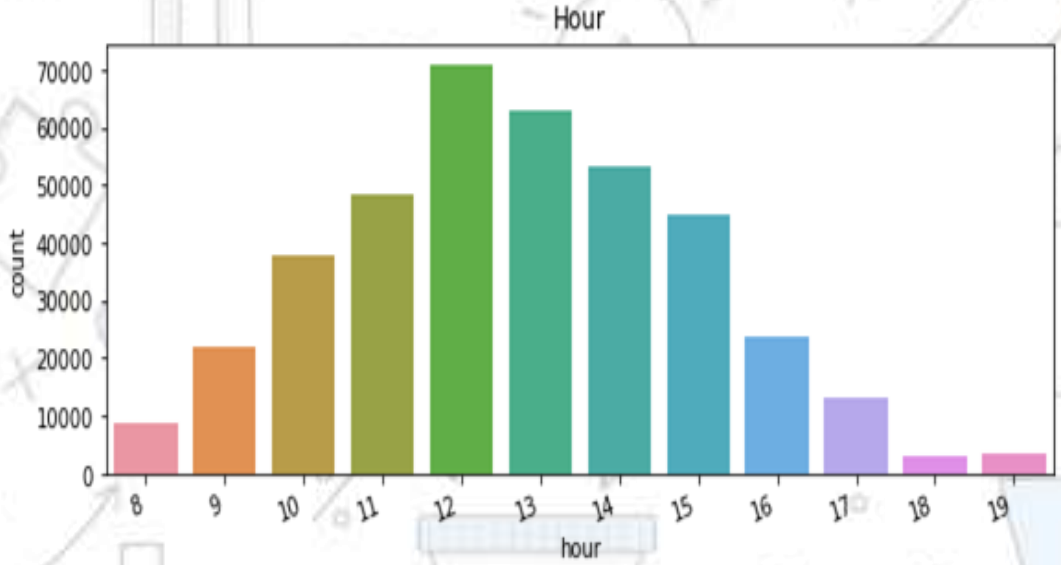


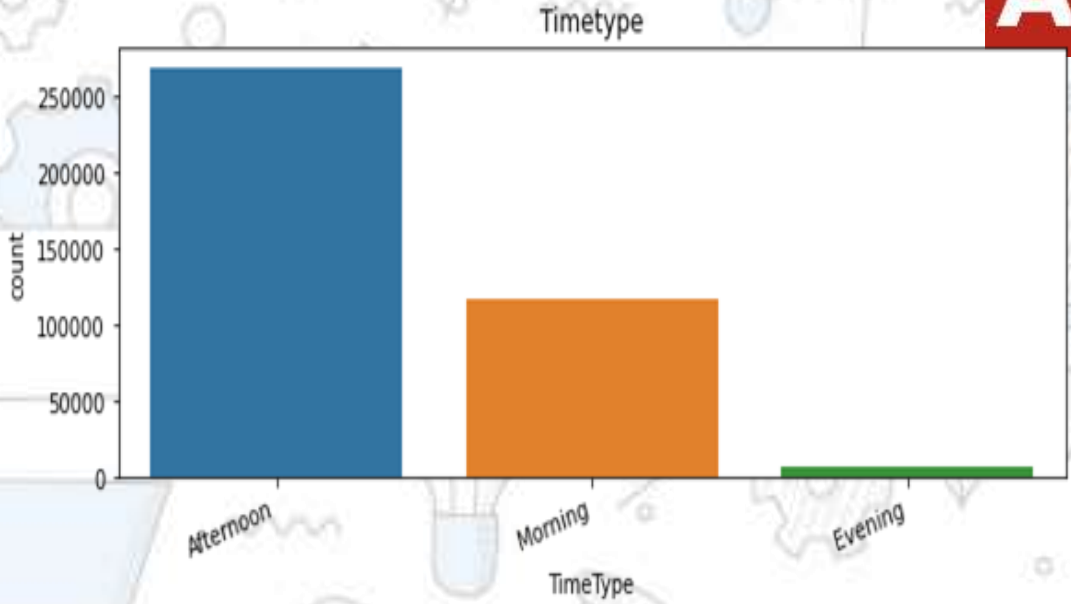
COUNTRY WISE CANCELLATIONS





FREQUENT VALUES





Observations/Hypothesis

1. Most Customers are from the United Kingdom. A considerable number of customers are also from Germany, France, EIRE and Spain. Whereas Saudi Arabia, Bahrain, the Czech Republic, Brazil, and Lithuania has the least number of customers

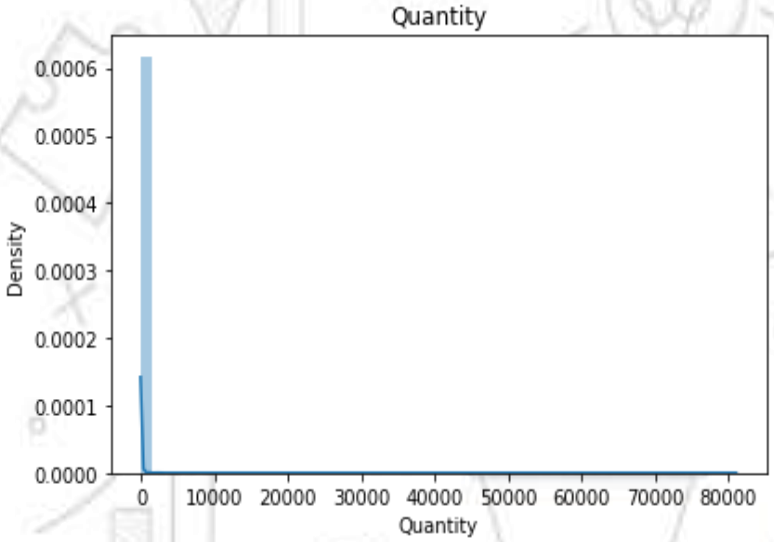
2. PAPER CUT LITTLE DRAFT, REGENCY CAKESTAND 3 TIER are the products making most of the revenue

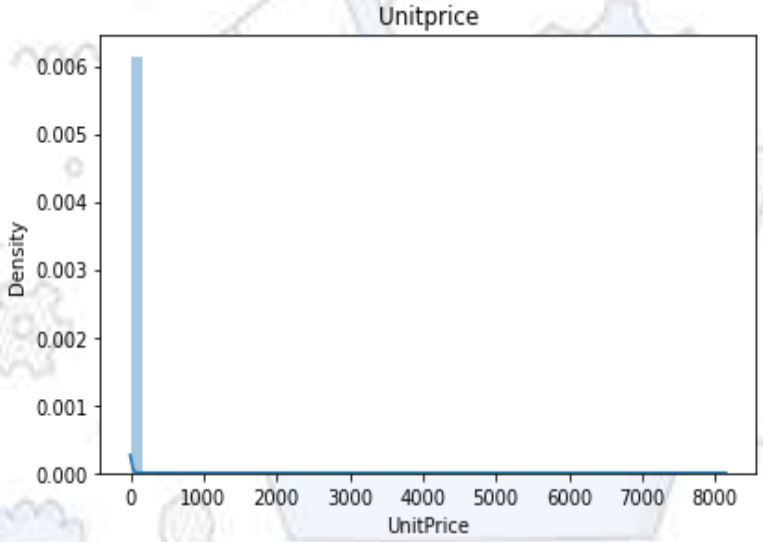
3. Countries like the United Kingdom, Germany, and EIRE have the most number of cancellations, countries like Saudi Arabia, Greece and the European community group have fewer cancellations

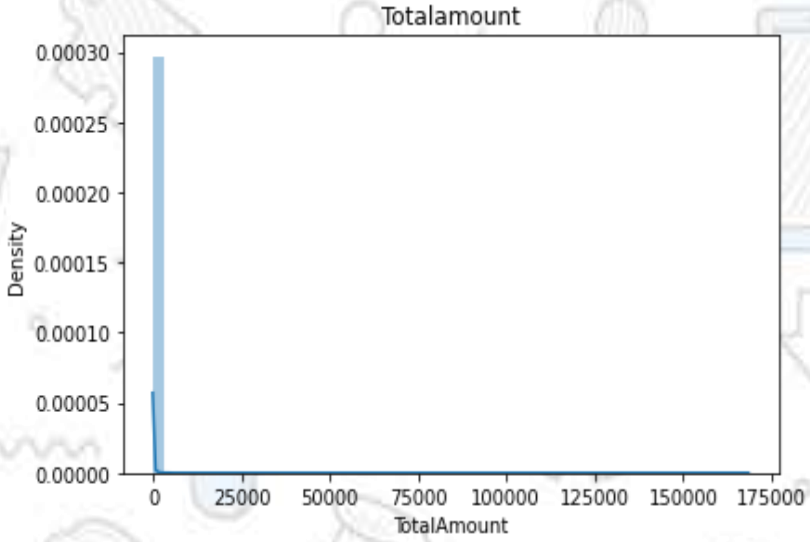
4. Most of the customers have purchased the items in the Afternoon, moderate numbers of customers have purchased the items in Morning and the least in the Evening.

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

VISUALIZING DISTRIBUTIONS





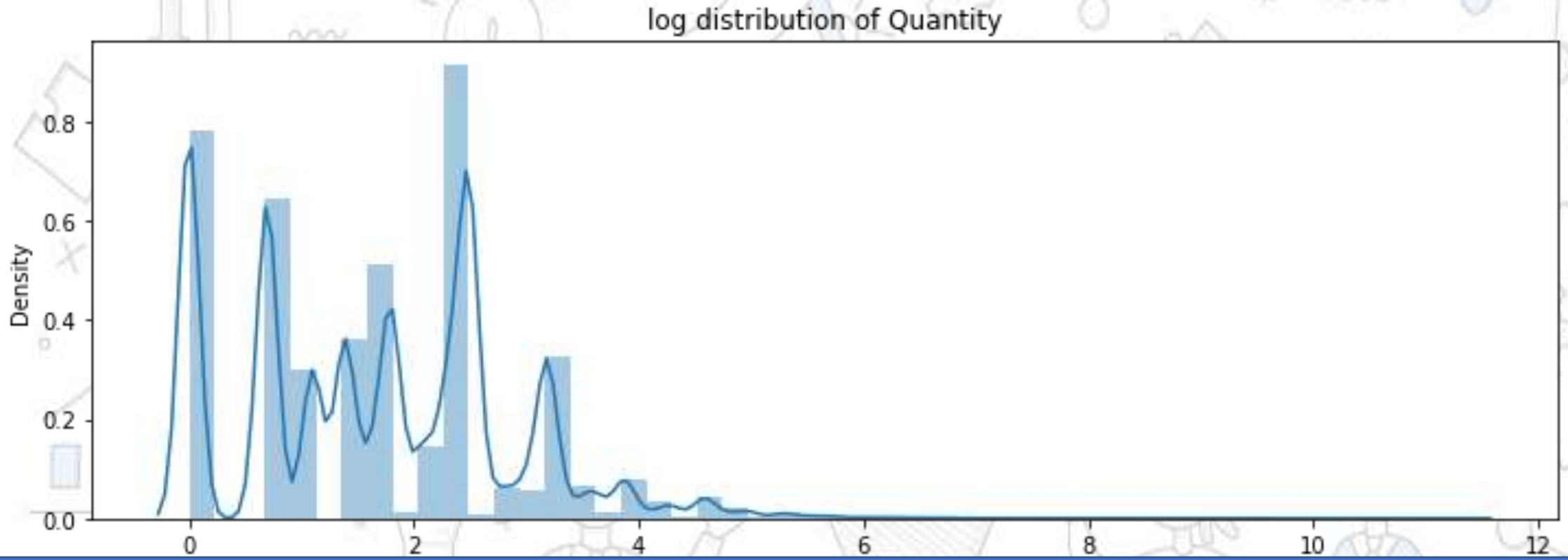


1. Visualizing the distribution of quantity, unitprice and total amount columns

2. 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

3. For symmetric graph mean=median=mode.

LOG TRANSFORMATION



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

2. 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.

RFM ANALYSIS

FREQUENCY
OF PURCHASE

How frequently do customers visit

Money Spent By Customer

Recent visit by the customers

MONETARY
VALUE

Recency
Of Purchase

RFM MODELLING

Customer Name	Recency	Frequency	Monetary
Anthony	326	15	7183
Rahul	2	182	4310
Syed	75	31	1765

RFM TABLE

CONCLUSIONS:

Anthony

Anthony visited 326 days (approx. 1 year) ago and visited 15 times and spent around 7183 Sterlings

Lost Potential Customer

Rahul

Rahul visited 2 days ago and visited 182 times and spent around 4310 Sterlings

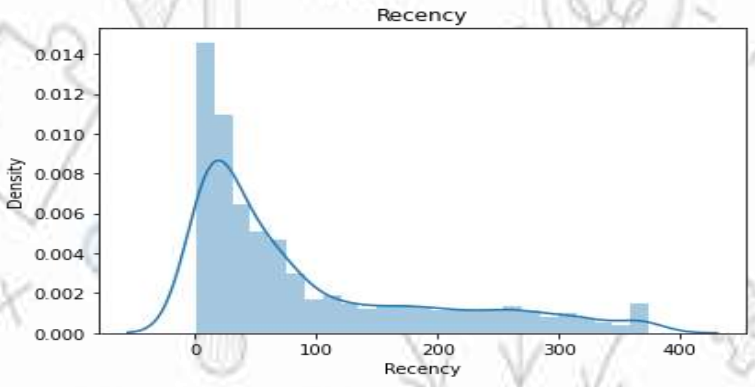
Recently visited Potential Customer

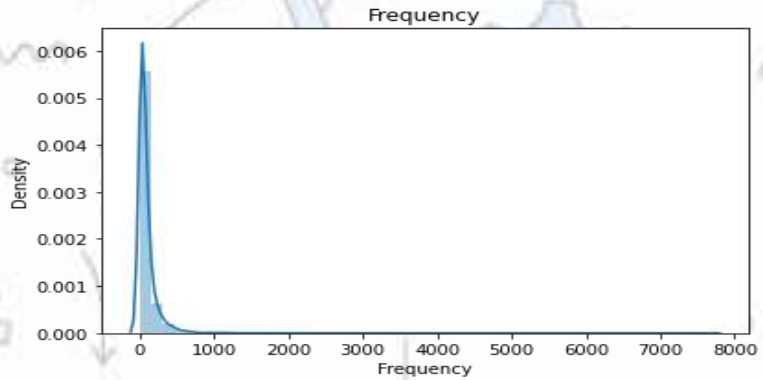
Syed

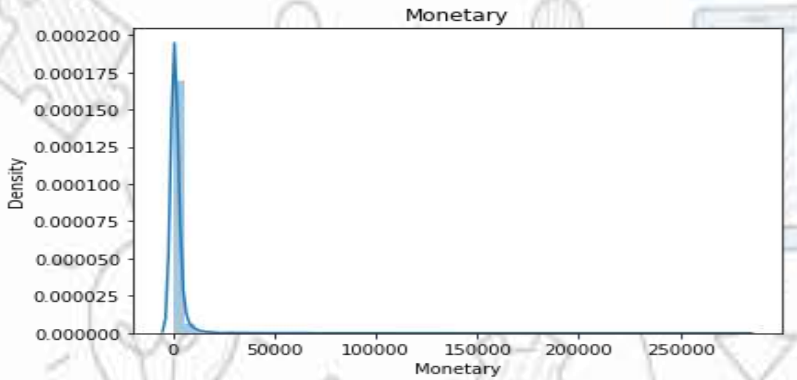
Syed visited 75 days ago (2.5 months)and visited 31 times and spent around 1765 Sterlings

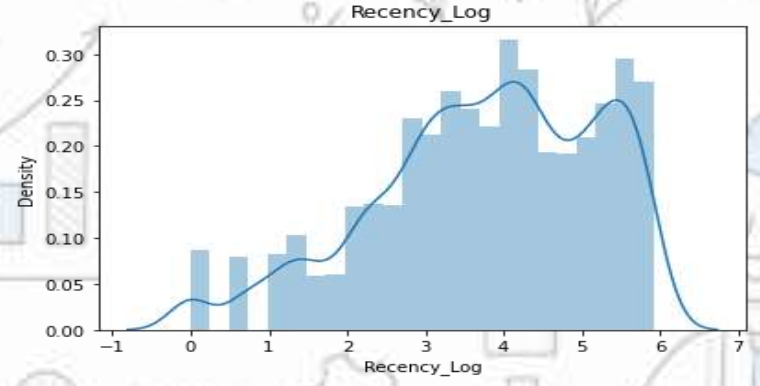
About to Lose Average Customer

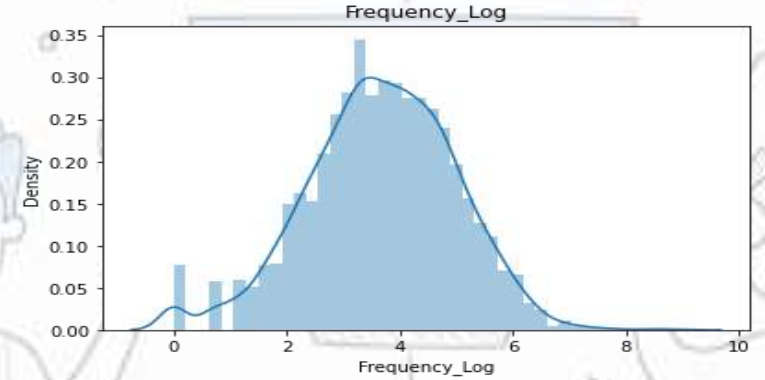
RFM MODELLING

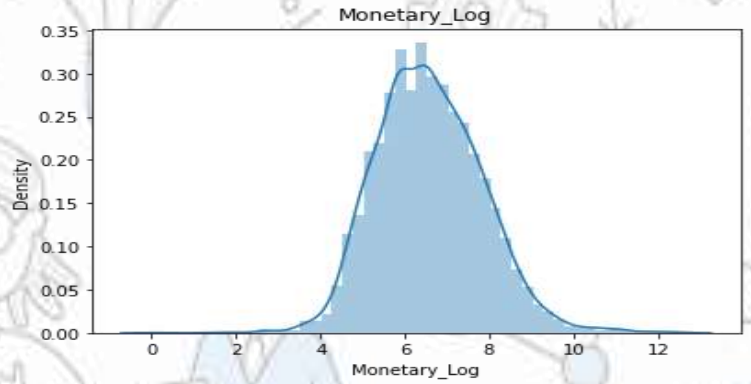












1. Earlier the distributions of Recency, Frequency and Monetary columns were positively skewed but after applying log transformation, the distributions appear to be symmetrical and normally distributed.

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

Pipeline

AI

EXTRACTING DATA → DATA CLEANING → DATA VISUALIZATION → RFM ANALYSIS

Online Retail

Observation:541908
(shape=8x541908)

Checking Missing data

1. 25 % of items (i.e 135080)
2. Products– 1454

Checking duplicates

5268 data points were Duplicated

401604 DATA POINT LEFT

RECENCY: Must be **LESS**

FREQUENCY: Must be **MORE**

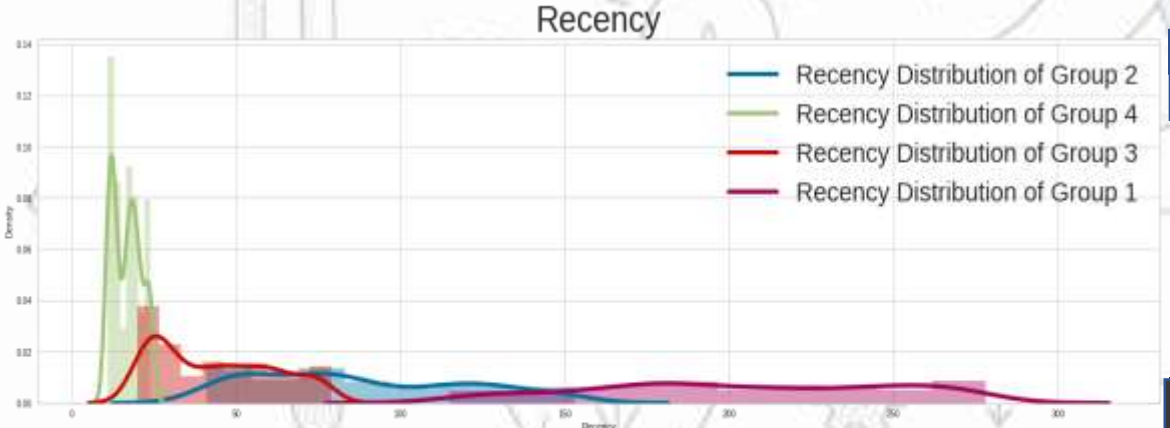
MONETARY: Must be **MORE**

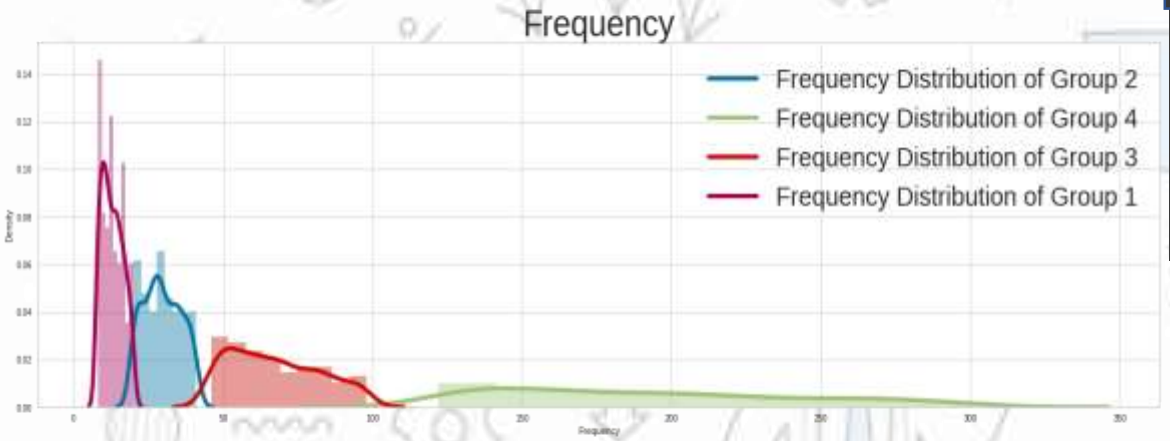
Condition: For Best Customers

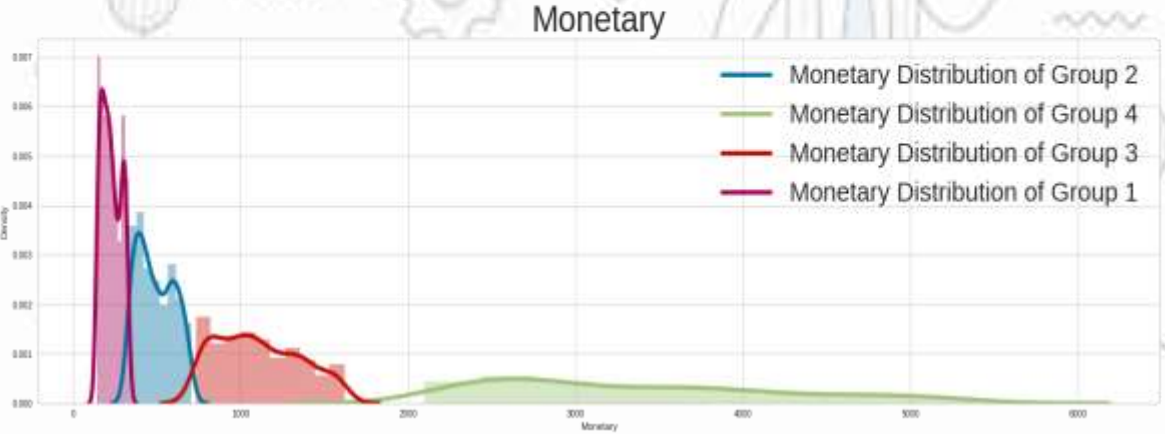
MODELLING → CUSTOMER SEGMENTATION → CONCLUSION

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

QUANTILE BASED CLUSTERING (RFM)







Recency		Frequency		Monetary		
mean	median	mean	median	mean	median	Count
212.503529	202.000000	15.282353	12.000000	263.277420	219.390000	1275
108.752711	81.000000	33.649675	29.000000	922.228450	481.330000	922
56.855403	42.000000	84.457382	65.000000	1497.687619	1079.285000	1314
19.178744	17.000000	279.281401	184.500000	6924.245193	3176.460000	828

Segment	Visited	Brought	Money Spent
1	Visited 114 to 279 days ago	Bought 7 to 21 Times	Spent Around 141 to 330 Sterling
2	Visited 40 to 154 days ago	Bought 19 to 42 Times	Spent Around 329 to 708 Sterling
3	Visited 19 to 79 days ago	Bought 45 to 99 Times	Spent Around 732 to 1623 Sterling
4	Visited 10 to 25 days ago	Bought 121 to 301 Times	Spent Around 2096 to 5396 Sterling

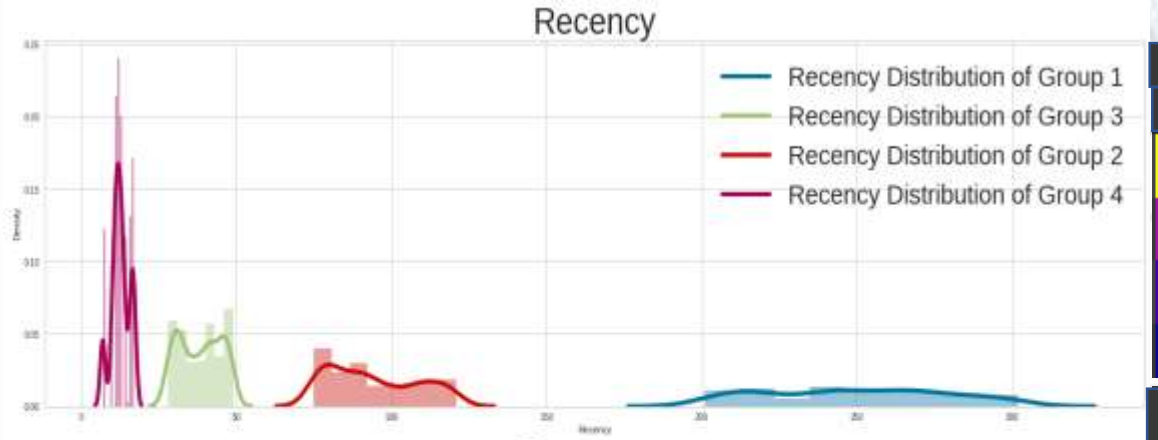
Group 1 → Lost Poor Customers

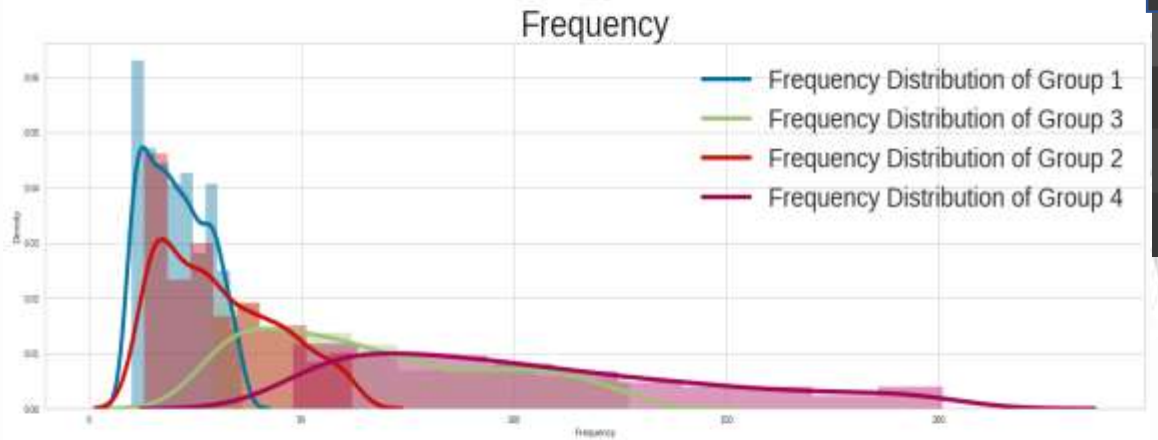
Group 2 → Average Customers

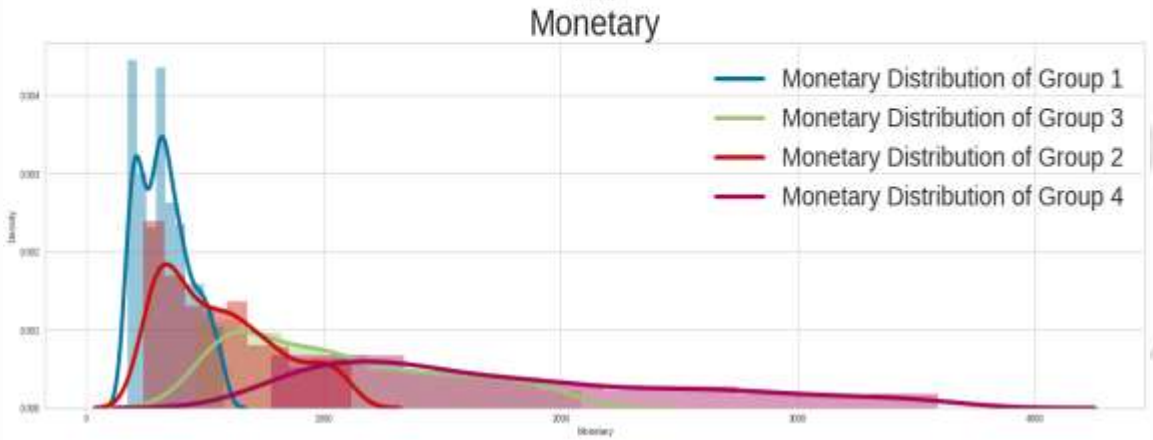
Group 3 → Good Customers

Group 4 → Best Customers

QUANTILE BASED CLUSTERING (RFM SCORE)





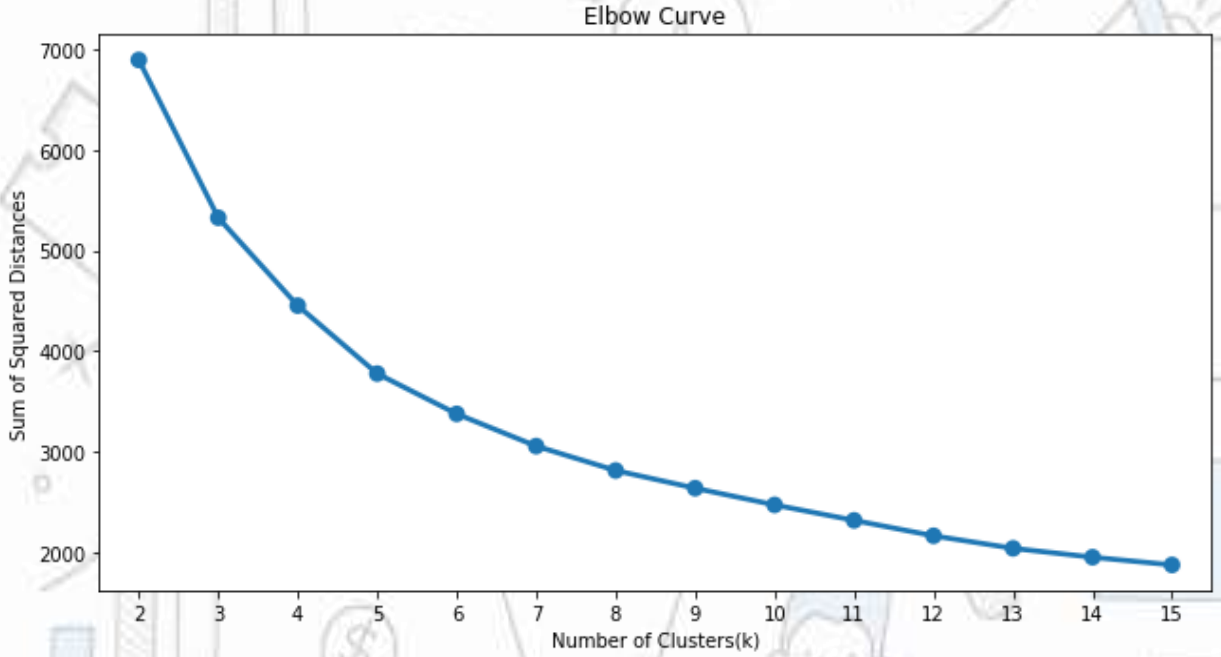


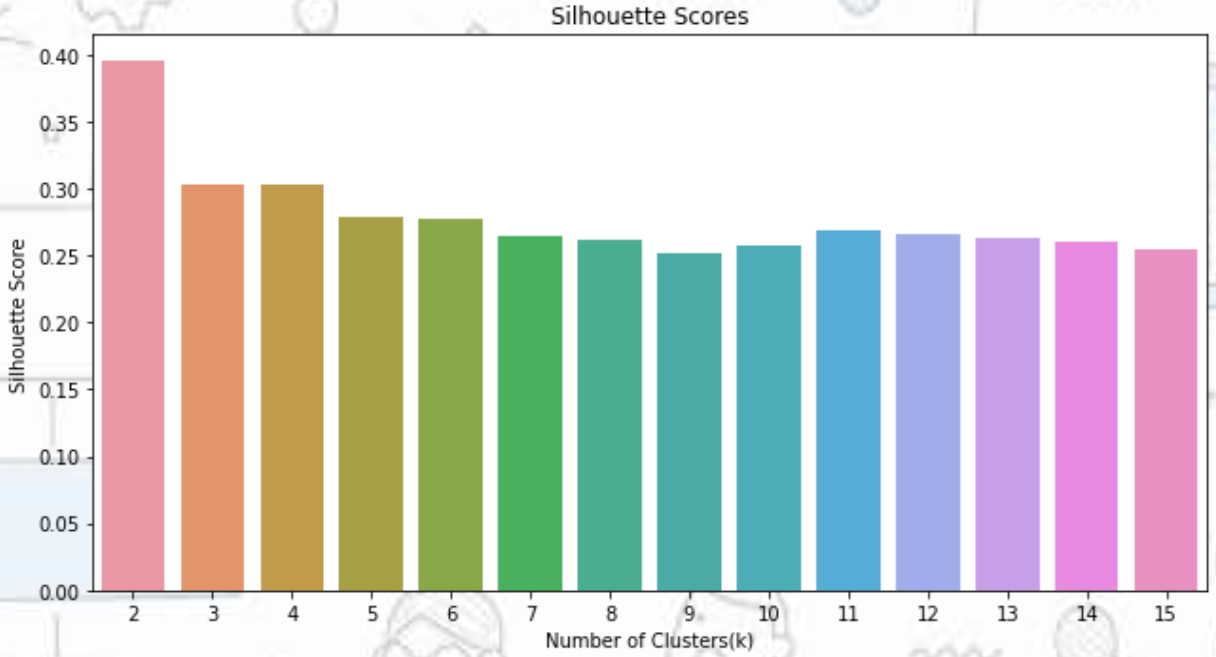
Recency		Frequency		Monetary		
mean	median	mean	median	mean	median	Count
271.032258	250.000000	30.895853	20.000000	575.341595	317.150000	1085
97.374264	91.000000	51.946173	28.000000	1160.407479	520.800000	1189
38.404130	37.000000	99.256637	60.000000	2129.257227	979.720000	1017
12.304389	13.000000	187.501908	97.000000	4501.703406	1586.645000	1048

Segment	Visited	Brought	Money Spent
1	Visited 200 to 303 days ago	Bought 9 to 37 Times	Spent Around 172 to 582 Sterling
2	Visited 74 to 122 days ago	Bought 12 to 63 Times	Spent Around 240 to 1120 Sterling
3	Visited 27 to 51 days ago	Bought 28 to 128 Times	Spent Around 463 to 2093 Sterling
4	Visited 6 to 18 days ago	Bought 47 to 202 Times	Spent Around 778 to 3593 Sterling

- Group 1 → Lost Poor Customers
- Group 2 → Lost Loyal Customers
- Group 3 → Good Customers
- Group 4 → Best Customers

K-MEANS CLUSTERING



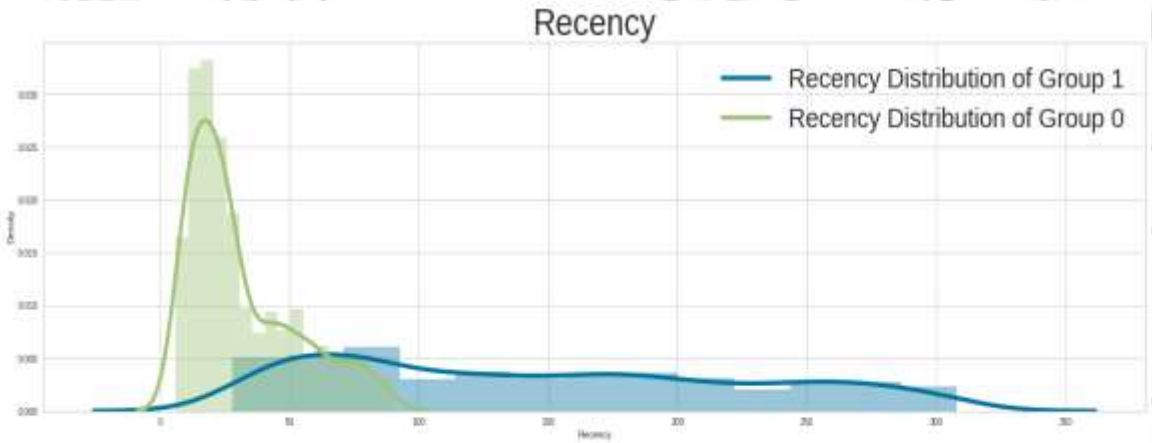


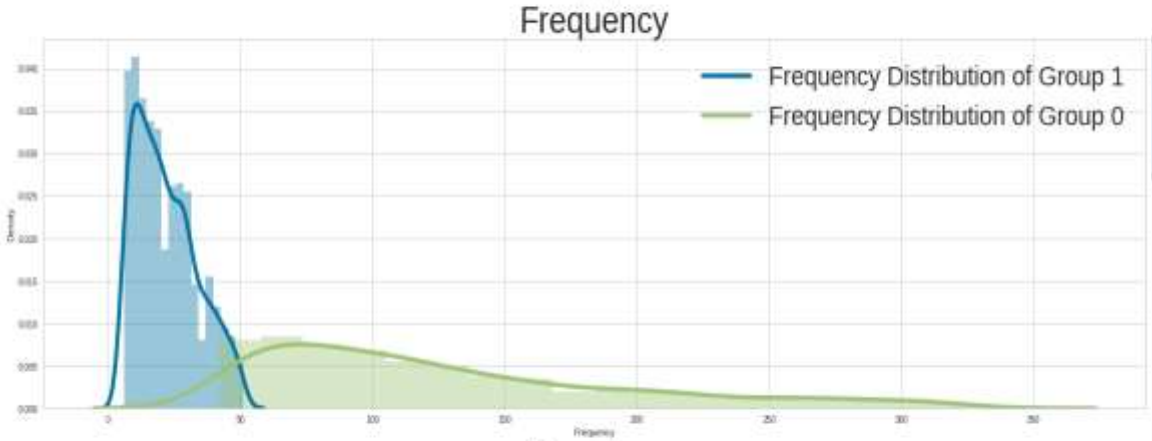
1. From the Elbow curve 5 appears to be at the elbow and hence can be considered as the number of clusters. n_clusters=4 or 6 can also be considered.

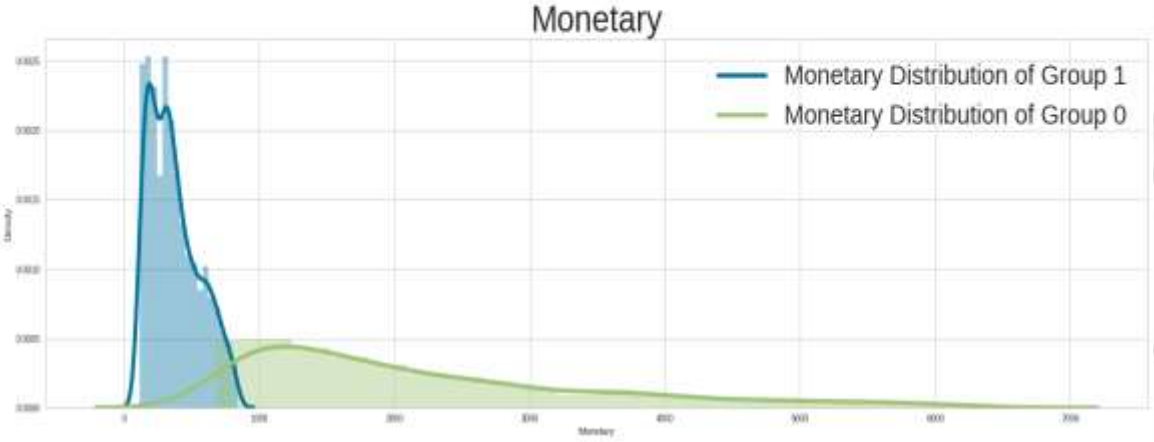
2. If we go by the maximum Silhouette Score as the criteria for selecting an optimal number of clusters, then n_clusters=2 can be chosen. 3 and 4 is also a good choice if we want more segments.

3. 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 CLUSTERING (2-Clusters)







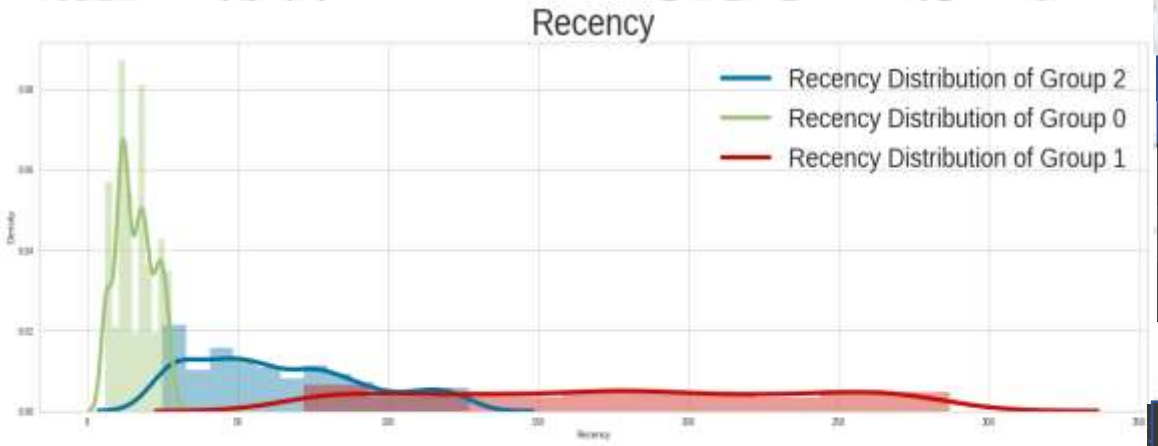
Recency		Frequency		Monetary		
mean	median	mean	median	mean	median	Count
39.313786	24.000000	171.549383	107.500000	4001.372475	1802.025000	1944
160.907724	135.000000	24.734864	19.000000	462.856703	329.600000	2395

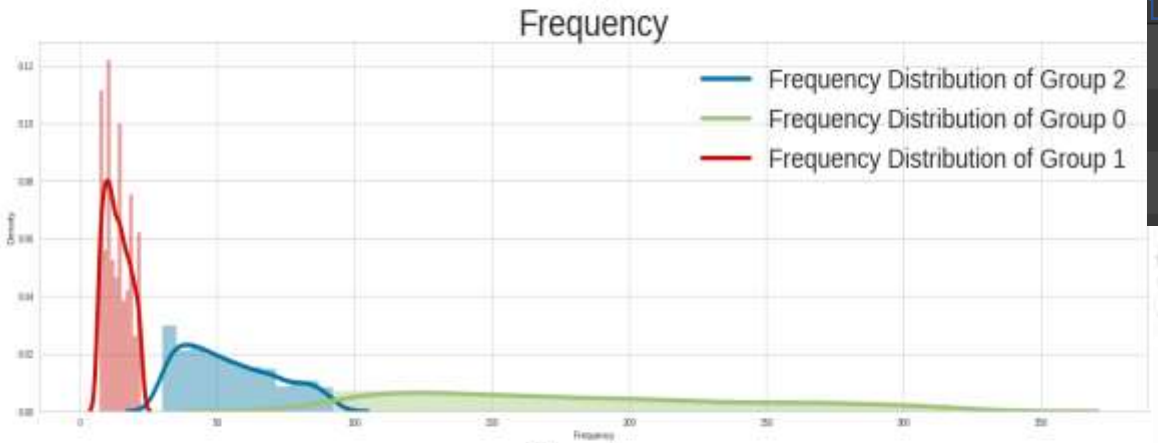
Segment	Visited	Brought	Money Spent
0	Visited 12 to 52 days ago	Bought 65 to 190 Times	Spent Around 1057 to 3340 Sterling
1	Visited 61 to 240 days ago	Bought 10 to 33 Times	Spent Around 187 to 558 Sterling

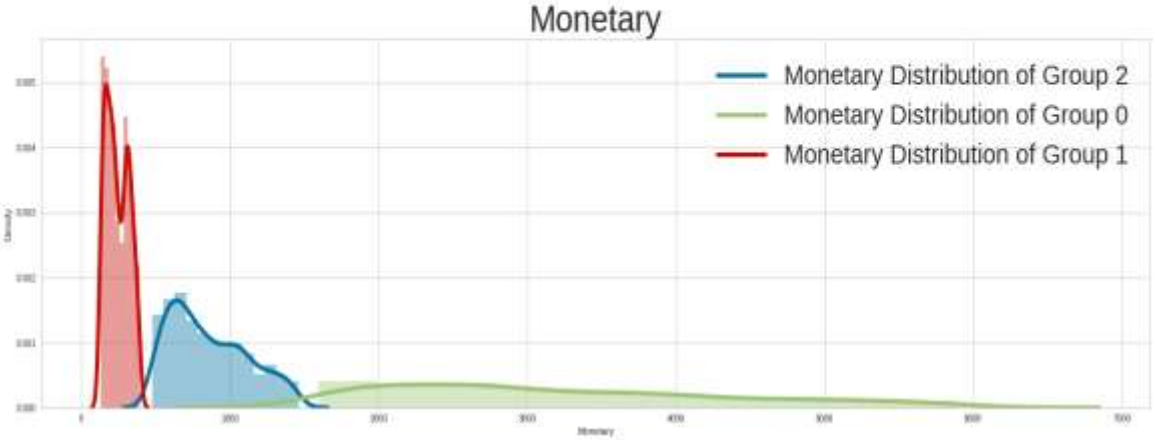
Group 0 → **Best Customers**

Group 1 → **Lost Poor Customers**

K-MEANS CLUSTERING (3-Clusters)







Recency		Frequency		Monetary		
mean	median	mean	median	mean	median	Count
20.309339	16.000000	249.069066	166.000000	6275.850409	2862.595000	1028
192.331997	179.000000	15.004008	13.000000	289.821096	230.250000	1497
84.344542	61.000000	62.970232	51.000000	1103.513502	790.375000	1814

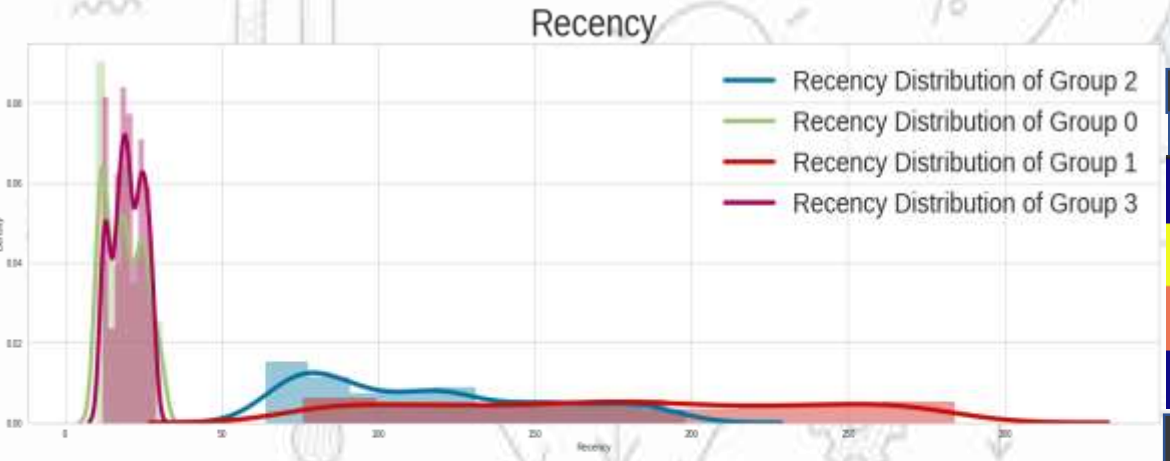
Segment	Visited	Brought	Money Spent
0	Visited 6 to 25 days ago	Bought 100 to 277 Times	Spent Around 1732 to 5040 Sterling
1	Visited 81 to 268 days ago	Bought 7 to 21 Times	Spent Around 145 to 351 Sterling
2	Visited 27 to 117 days ago	Bought 32 to 82 Times	Spent Around 528 to 1281 Sterling

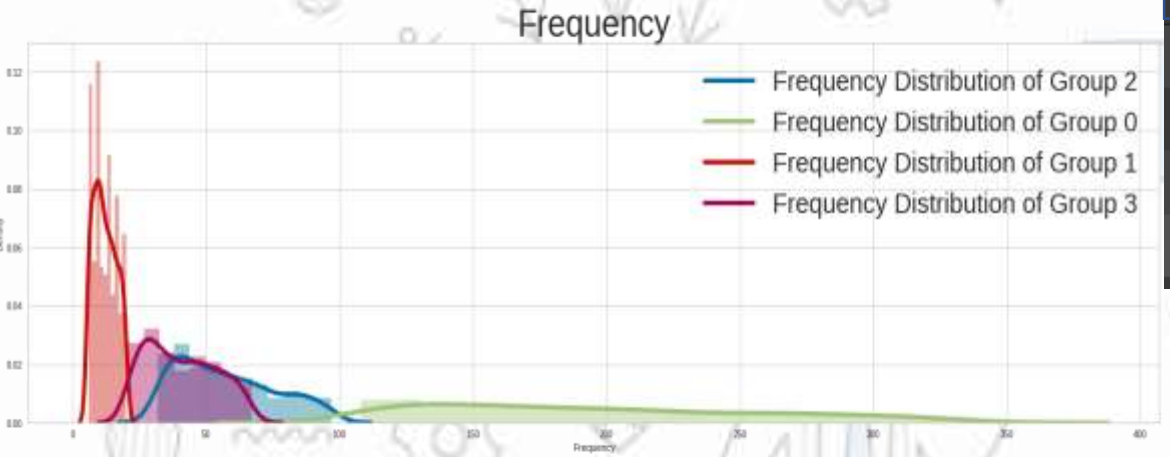
Group 0 → **Best Customers**

Group 1 → **Lost Poor Customers**

Group 2 → **Average Customers**

K-MEANS CLUSTERING (4-Clusters)



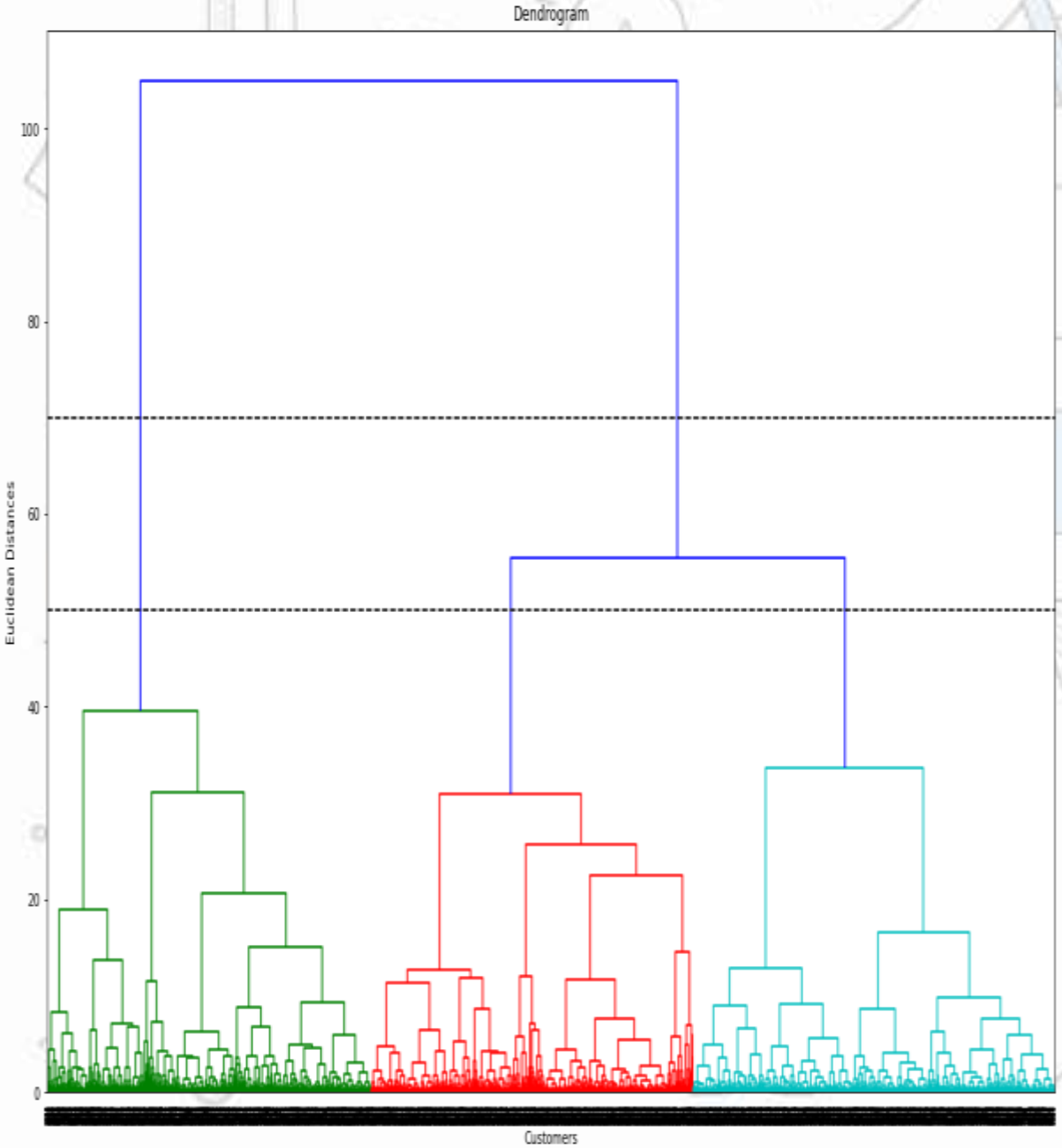


Recency		Frequency		Monetary		
mean	median	mean	median	mean	median	Count
22.438865	18.000000	272.385371	186.000000	6924.051681	3213.215000	916
194.534351	178.000000	13.388550	12.000000	270.886497	217.895000	1310
134.384492	109.000000	68.603517	54.000000	1258.928795	881.070000	1251
21.219258	20.000000	46.247100	39.000000	713.453968	617.620000	862

Segment	Visited	Brought	Money Spent
0	Visited 10 to 28 days ago	Bought 117 to 296 Times	Spent Around 2076 to 5451 Sterling
1	Visited 86 to 266 days ago	Bought 6 to 19 Times	Spent Around 137 to 331 Sterling
2	Visited 70 to 181 days ago	Bought 35 to 88 Times	Spent Around 588 to 1483 Sterling
3	Visited 13 to 28 days ago	Bought 22 to 61 Times	Spent Around 355 to 970 Sterling

- Group 0 → Best Customers
- Group 1 → Lost Poor Customers
- Group 2 → Lost Loyal Customers
- Group 3 → Good Customers

DENDOGRAM



HIERARCHICAL CLUSTERING

AI

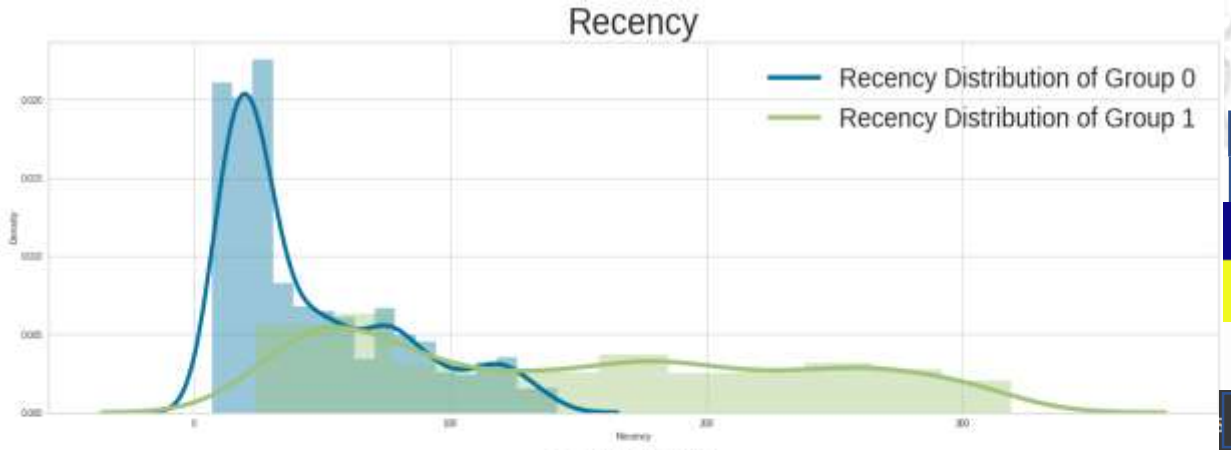
In the K-means clustering there is a challenge to predetermine the number of clusters, and it always tries to create the clusters of the same size. To solve these two challenges, we can opt for the hierarchical clustering algorithm because, in this algorithm, we don't need to have knowledge about the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

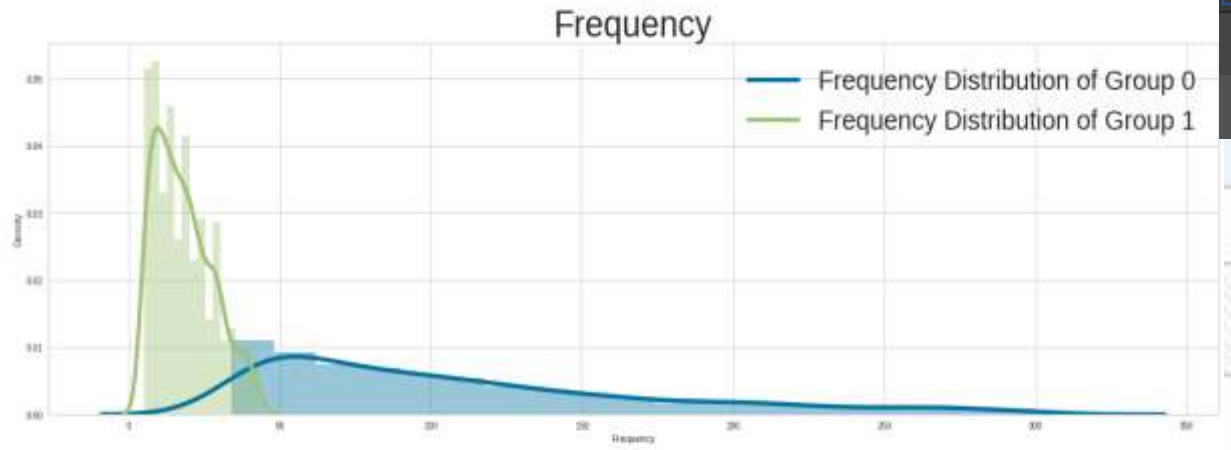
a. Agglomerative: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

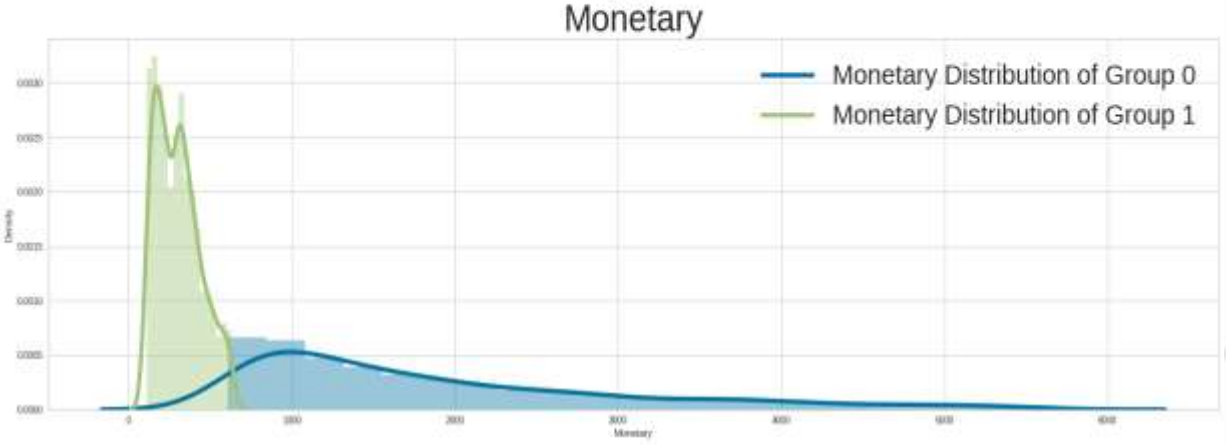
b. Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

1. We can set a threshold distance and draw a horizontal line (Generally, we try to set the threshold in such a way that it cuts the tallest vertical line). We can set the likes threshold as 50 or 70 and draw a horizontal line as shown in the dendrogram above.like
2. The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold. The larger threshold (y=70) results in 2 clusters while the smaller (y=50) results in 3 clusters.

HIERARCHICAL CLUSTERING (2-Clusters)







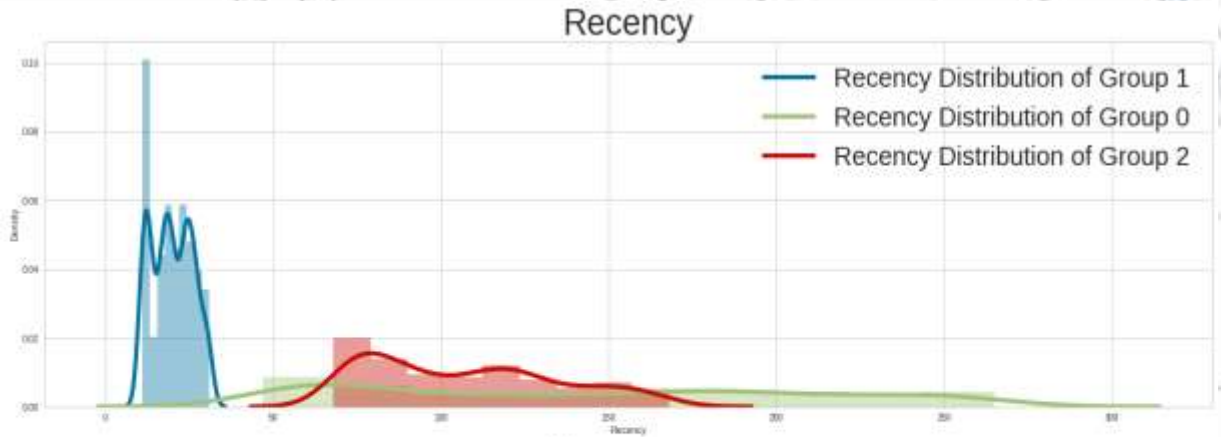
Recency		Frequency		Monetary		
mean	median	mean	median	mean	median	Count
57.281453	30.000000	151.238219	91.000000	3534.730633	1522.750000	2313
162.540967	139.500000	21.183613	16.000000	351.124354	294.620000	2026

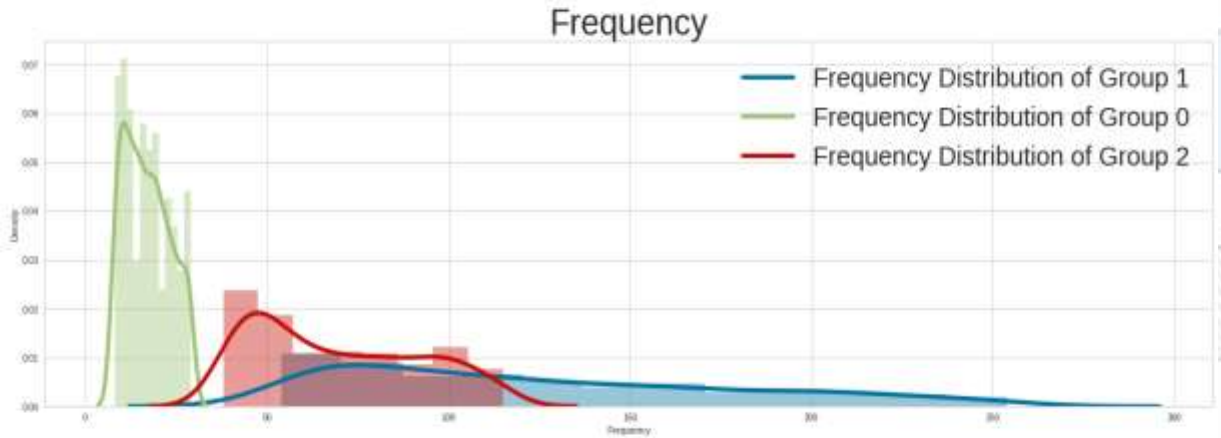
Segment	Visited	Brought	Money Spent
0	Visited 16 to 79 days ago	Bought 50 to 165 Times	Spent Around 887 to 2861 Sterling
1	Visited 54 to 245 days ago	Bought 9 to 28 Times	Spent Around 168 to 427 Sterling

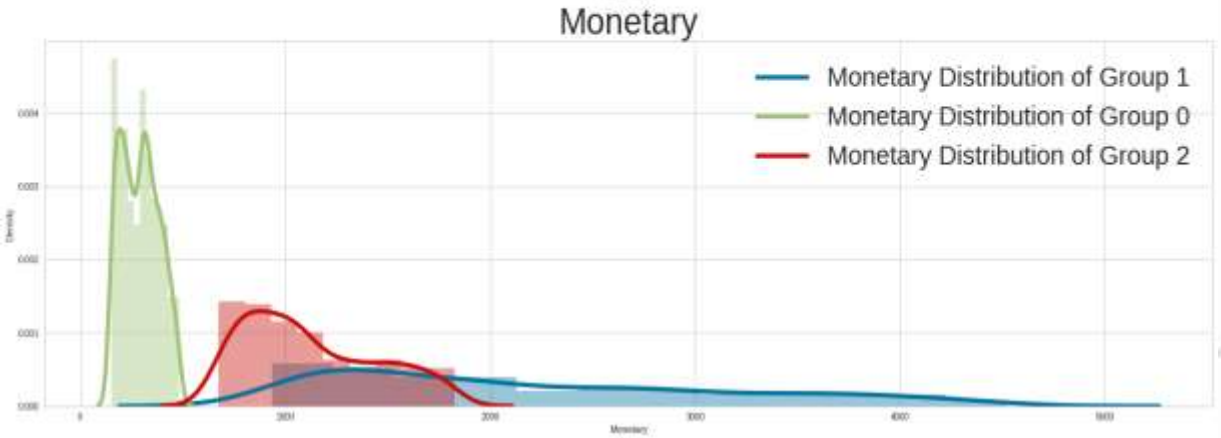
Group 0 → **Best Customers**

Group 1 → **Lost Poor Customers**

HIERARCHICAL CLUSTERING (3-Clusters)







Recency		Frequency		Monetary		
mean	median	mean	median	mean	median	Count
162.540967	139.500000	21.183613	16.000000	351.124354	294.620000	2026
24.338773	20.000000	188.546345	111.000000	4672.209837	1950.845000	1532
121.901408	105.000000	78.055058	63.000000	1303.465407	1046.740000	781

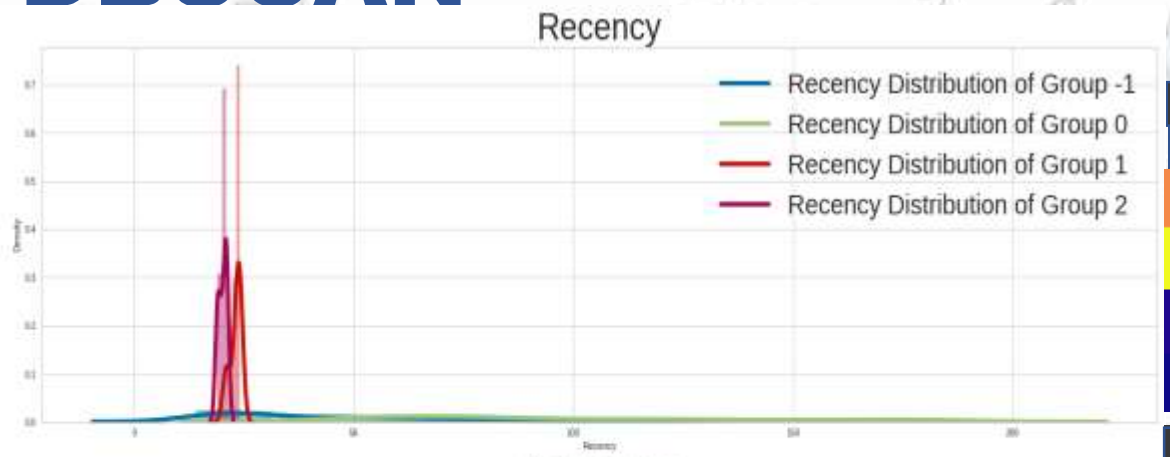
Segment	Visited	Brought	Money Spent
0	Visited 54 to 245 days ago	Bought 9 to 28 Times	Spent Around 168 to 427 Sterling
1	Visited 11 to 30 days ago	Bought 60 to 215 Times	Spent Around 1074 to 3807 Sterling
2	Visited 73 to 153 days ago	Bought 41 to 105 Times	Spent Around 729 to 1661 Sterling

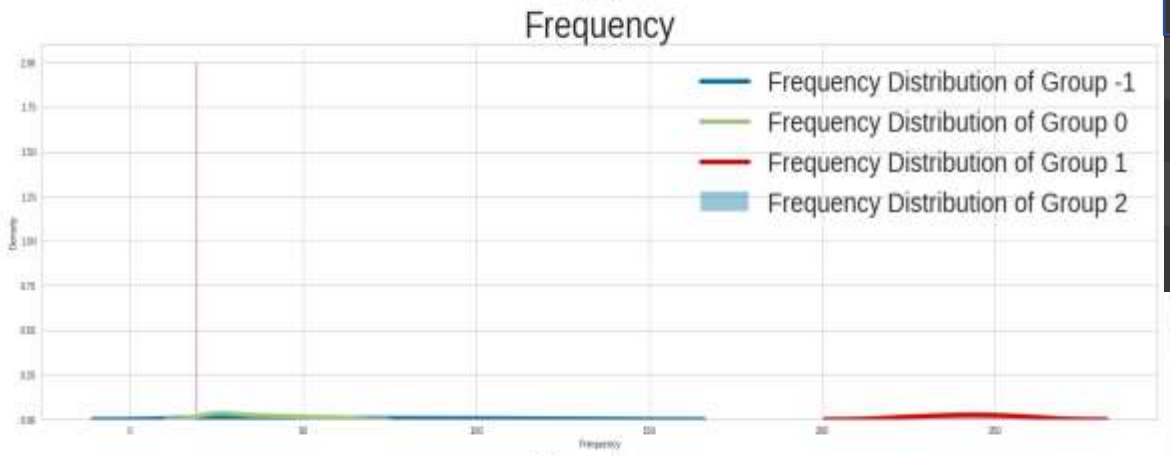
Group 0 → **Lost Poor Customers**

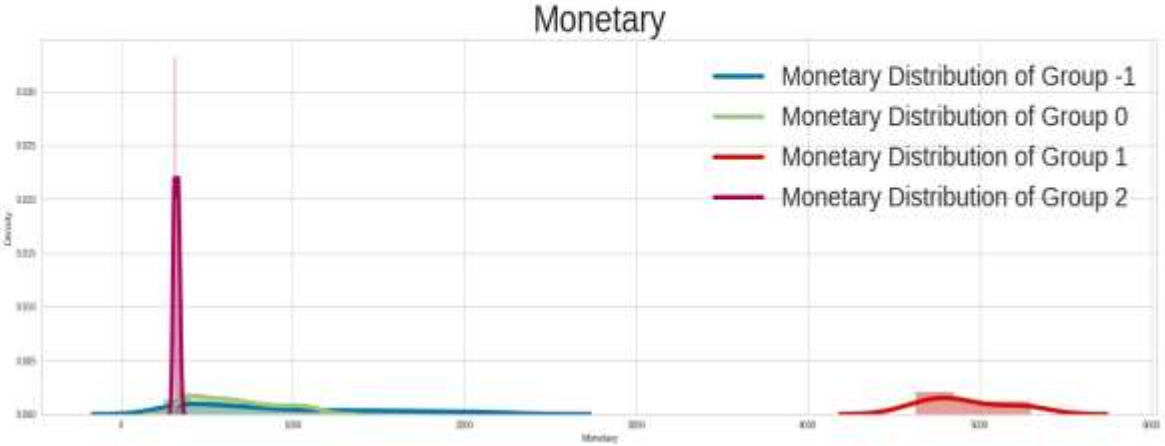
Group 1 → **Best Customers**

Group 2 → **Losing Loyal Customers**

DBSCAN







Recency		Frequency		Monetary		
mean	median	mean	median	mean	median	Count
94.332190	42.000000	124.286755	53.000000	2992.340935	739.200000	2333
123.277806	86.000000	49.270118	35.000000	908.771206	626.650000	1951
23.480000	24.000000	243.200000	245.000000	4932.972400	4867.720000	25
20.700000	20.000000	18.833333	19.000000	324.718667	323.880000	30

Segment	Visited	Brought	Money Spent
-1	Visited 13 to 129 days ago	Bought 14 to 141 Times	Spent Around 247 to 2317 Sterling
0	Visited 42 to 190 days ago	Bought 19 to 66 Times	Spent Around 331 to 1186 Sterling
1	Visited 19 to 25 days ago	Bought 222 to 265 Times	Spent Around 4544 to 5417 Sterling
2	Visited 18 to 24 days ago	Bought 18 to 20 Times	Spent Around 306 to 345 Sterling

Group -1 → **Average Customers**

Group 0 → **Lost Loyal Customers**

Group 1 → **Recently Visited Potential Customers**

Group 2 → Recently Visited Average Customers

FINAL CONCLUSION

Segmented Customer Got from Cluster Analysis

Clusters	LOST POOR CUSTOMERS ❌	AVERAGE CUSTOMERS 🏅	RECENTLY VISITED AVERAGE CUSTOMERS 🤍	GOOD CUSTOMERS 🥈	BEST CUSTOMERS ❤️	LOSING LOYAL CUSTOMERS ❌
Binning_Segment_	Yes	Yes	No	Yes	Yes	No
QuantileCut	Yes	Yes	No	No	Yes	Yes
K-Means|2Cluster	Yes	No	No	Yes	Yes	Yes
K-Means|3Cluster	Yes	No	No	No	Yes	No
K-Means|4Cluster	Yes	Yes	No	No	Yes	No
K-Means|5Cluster	Yes	Yes	Yes	Yes	Yes	No
hierarchical|2Cluster	Yes	No	No	No	Yes	No
hierarchical|3Cluster	Yes	No	No	No	Yes	Yes
DBSCAN	No	Yes	Yes	Yes	Yes	No


Thank You