

# Online Retail Customer Segmentation

Unsupervised Machine  
Learning

**KISHOR SHIVAJI PATIL**

Data Science Trainee,  
AlmaBetter, Bangalore.

# Content

- ❖ **Problem Statement**
- ❖ **Data Summary**
- ❖ **Feature Engineering**
- ❖ **Exploratory Data Analysis**
- ❖ **Analysing Different Types of Models**
- ❖ **Output Summary**
- ❖ **Conclusion**

# Problem Statement

Given a dataset related to a online retailer based out of the UK, we need to analyse and identify major customer segments using K Means algorithm and also using different verification method to confirm the result.

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.

# Data Summary

- **A transnational data set with transactions occurring between 1st December 2010 and 9<sup>th</sup> December 2011 for a UK-based online retailer.**
- **Shape (rows- 541909, columns-8)**
- **The company mainly sells unique all-occasion gifts.**
- **Many customers of the company are wholesalers.**

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.

# INSPECTING DATASET

AI

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

## Data Cleaning

### 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

### Checking duplicates

5268 data points were duplicated



Dropped  
duplicates

### Total data points left

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

# Pipeline

AI

## EXTRACTING DATA

Online Retail  
Observation: 541908  
(shape=8x541908)

## DATA CLEANING

### Checking Missing data

25% of items ie. 135080 were  
Missing  
CustomerID – 1454 Missing

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

## CUSTOMER SEGMENTATION

## CONCLUSION

Binning (RFM SCORE)

Quantile Cut

K-Means

Hierarchical

DBSCAN Clustering

# FEATURE ENGINEERING

AI



Extracting year Date and Month from Invoice Date



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

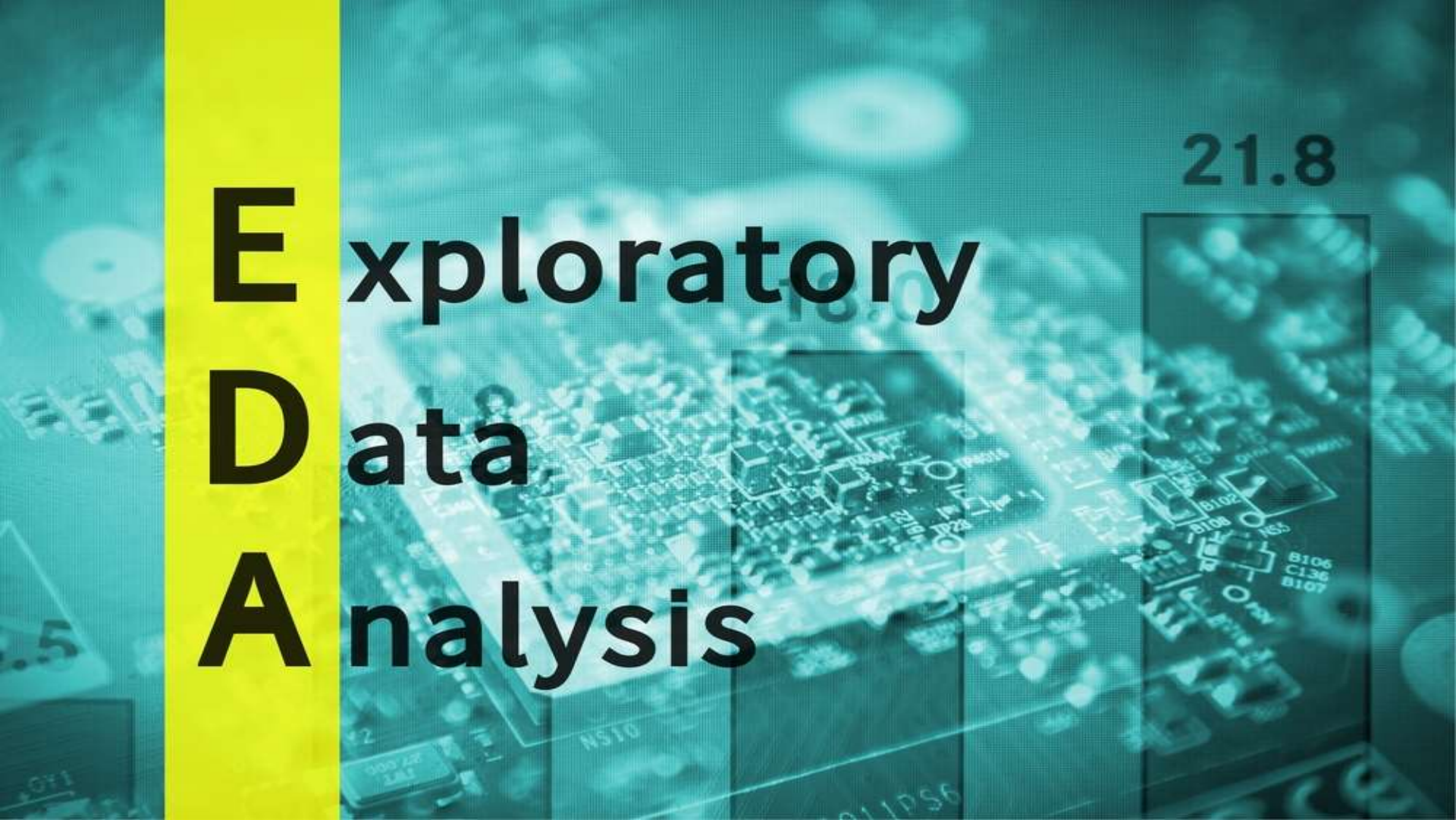


Added feature '**TimeType**' based on hours to define whether its Morning, Afternoon, or Evening



Dropping **InvoiceNo** starting with 'C' that represents cancellation

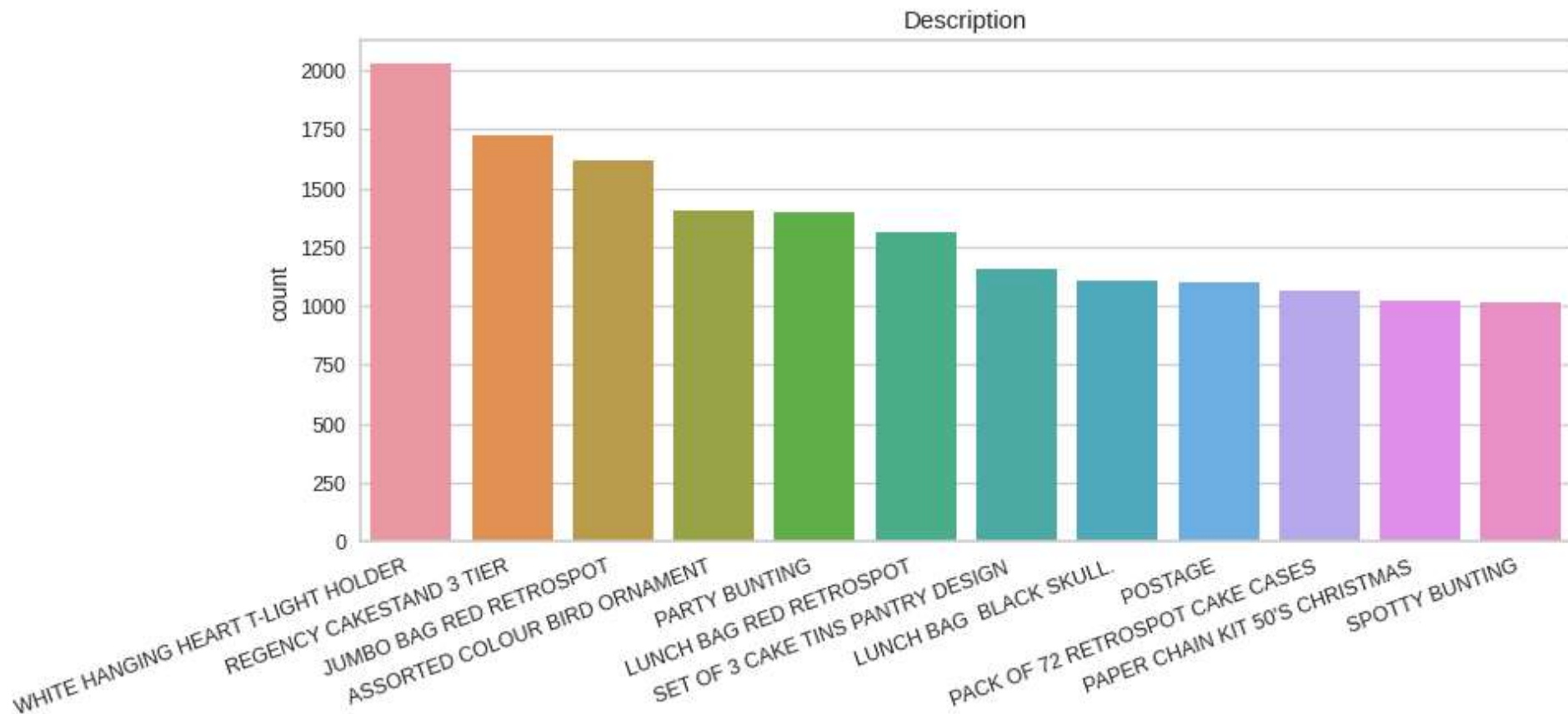




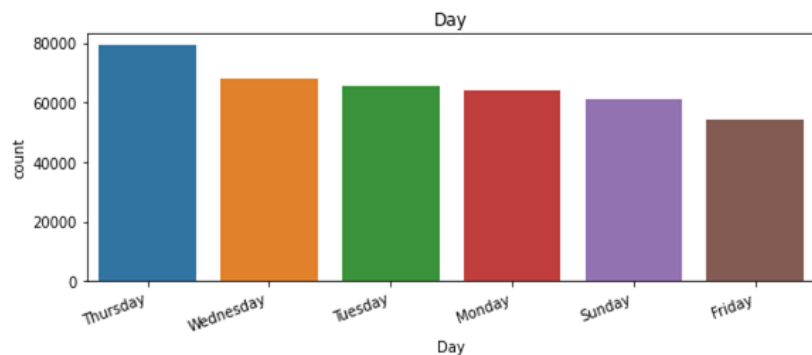
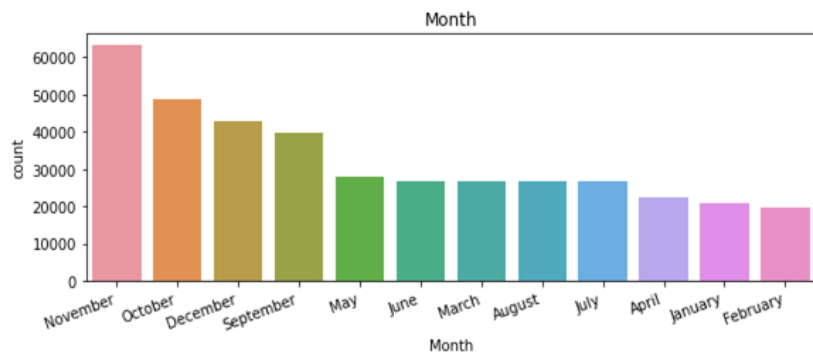
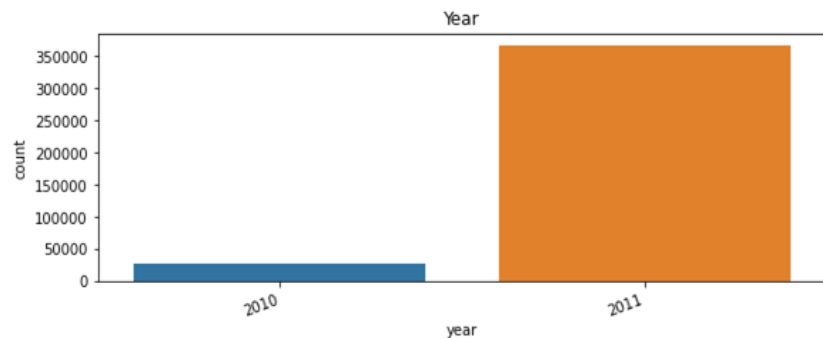
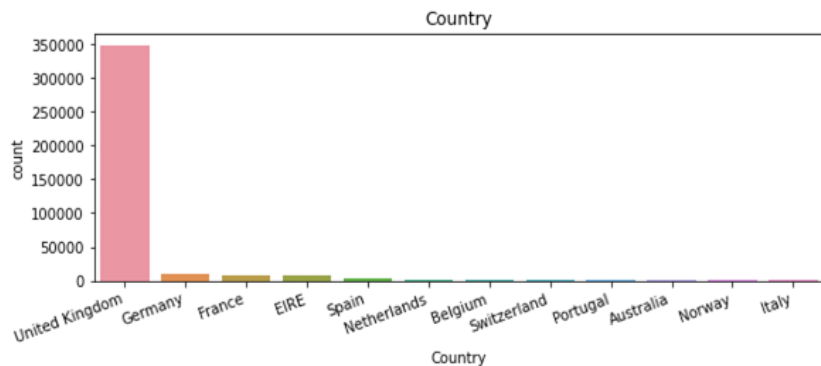
# **E**xploratory **D**ata **A**nalysis

21.8

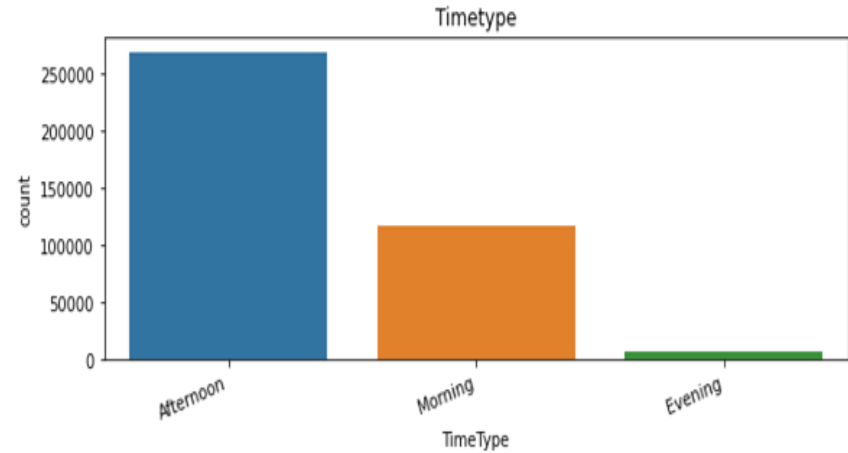
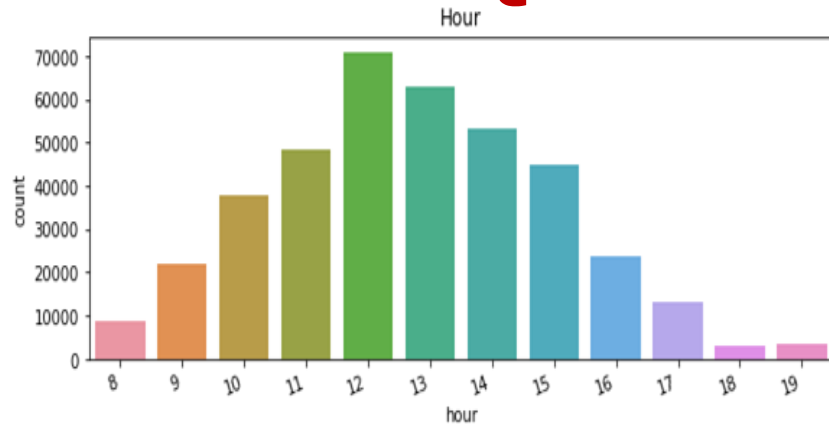
# MOST FREQUENT VALUES



# MOST FREQUENT VALUES



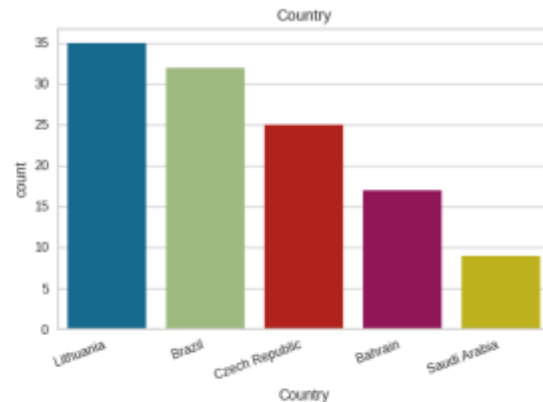
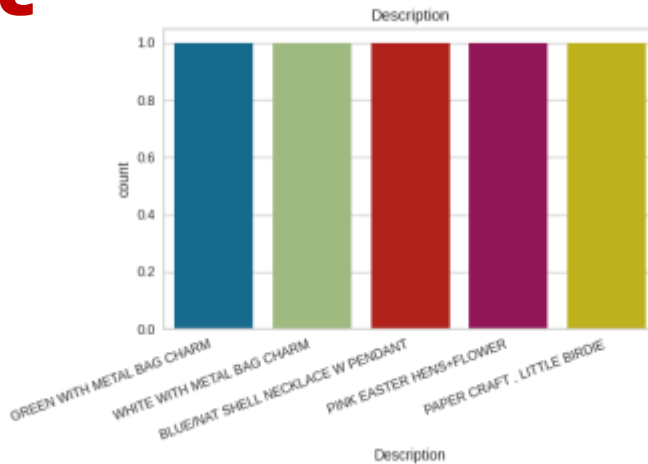
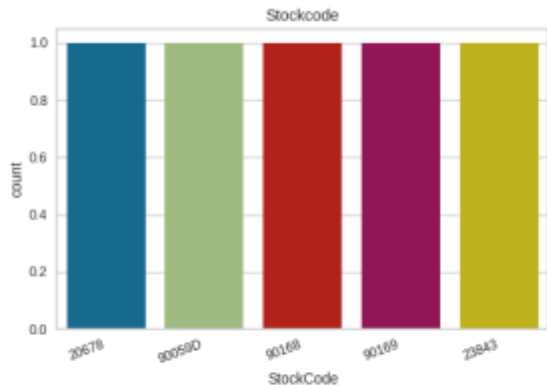
# MOST FREQUENT VALUES



## Observations/Hypothesis :-

1. Most Customers are from United Kingdom. Considerable number of customers are also from Germany, France, EIRR and Spain. Whereas Saudi Arabia, Bahrain, Czech Republic, Brazil and Lithuania has least number of customers.
2. There are no orders placed on Saturdays. Looks like it's a non working day for the retailer.
3. 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.
4. 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.
5. WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER, JUMBO BAG RED RETROSPOT are the most ordered products

# LESS FREQUENT VALUES



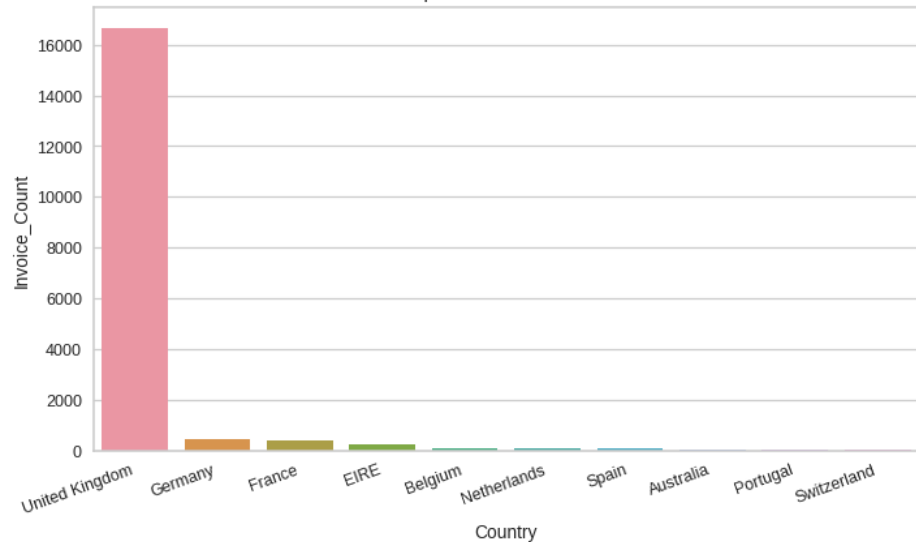
## Observations -

1. Saudi Arabia, Bahrain, the Czech Republic, Brazil, and Lithuania has the least number of customers

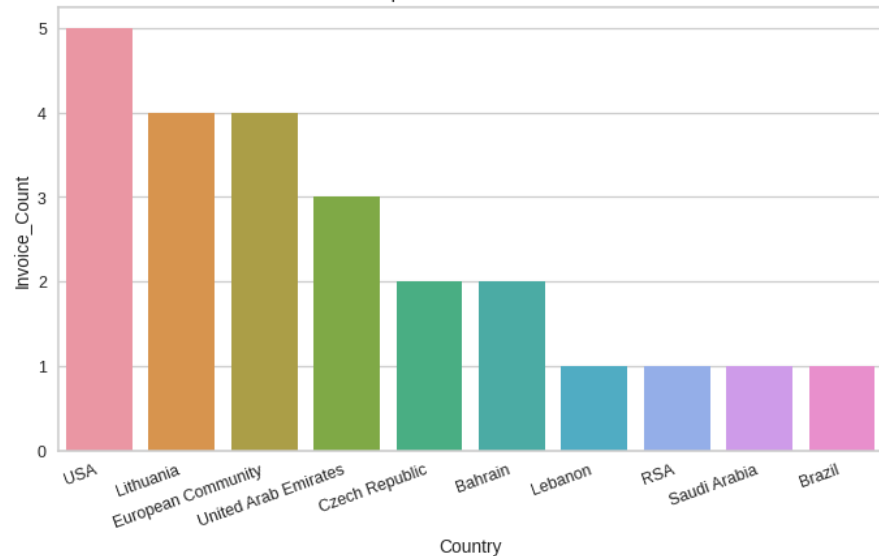
2. GREEN WIT METAL BAG CHARM, WHITE WITH METAL BAG CHARM, BLUE/NAT SELL NECLACE W PENDENT, PINK EASTER ENS FLOWER, PAPER CRAFT LITTLE BIRDIE are some of the least sold products.

# COUNTRY WISE ORDERS

Most orders placed are from these countries

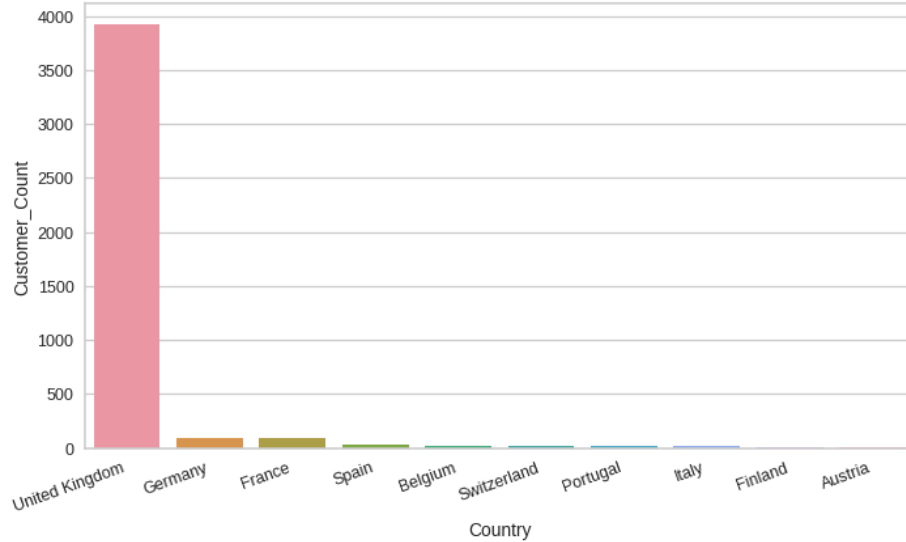


Least orders placed are from these countries

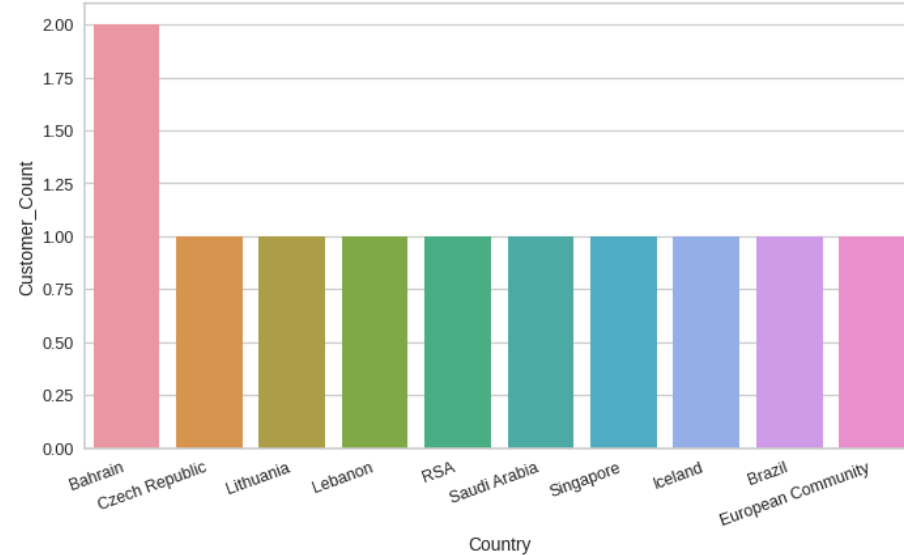


# COUNTRY WISE CUSTOMERS

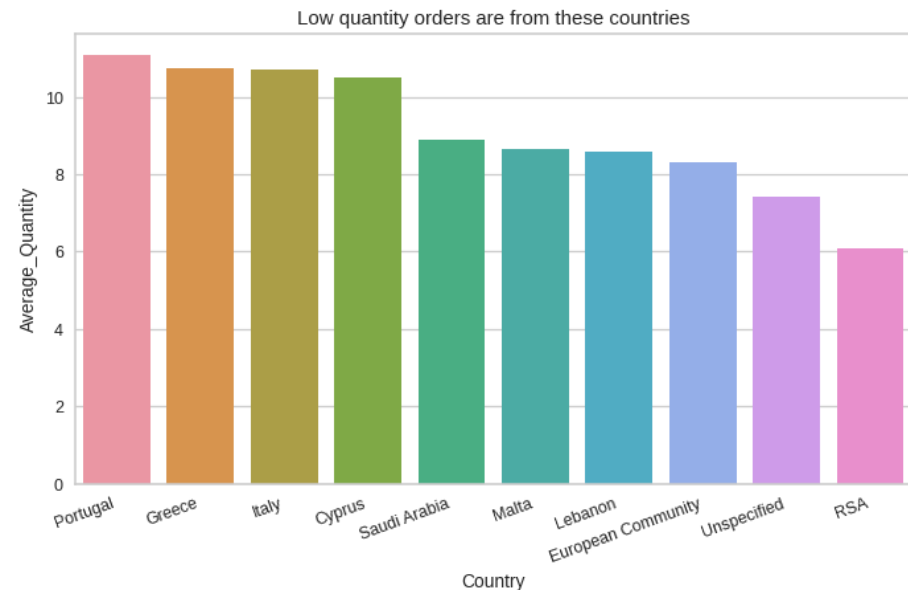
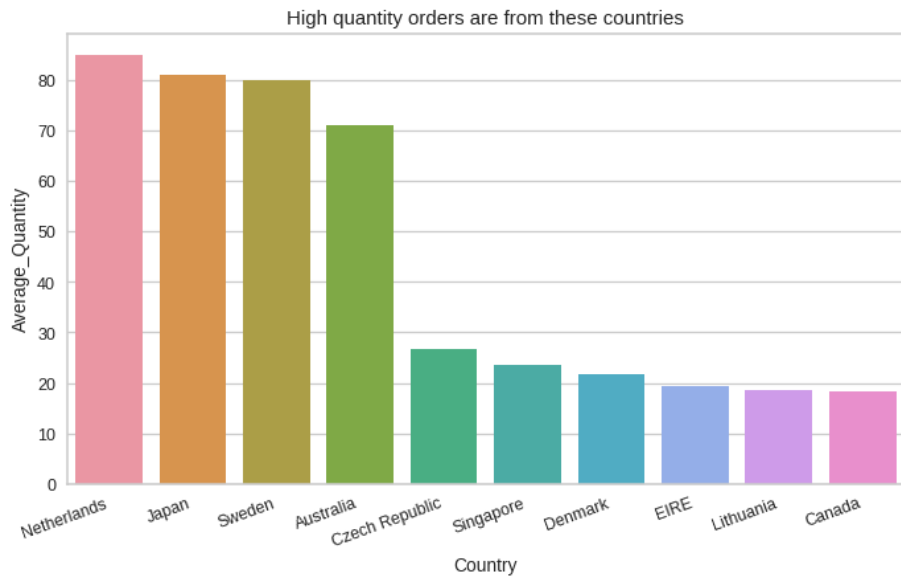
Most customers are from these countries



Least customers are from these countries

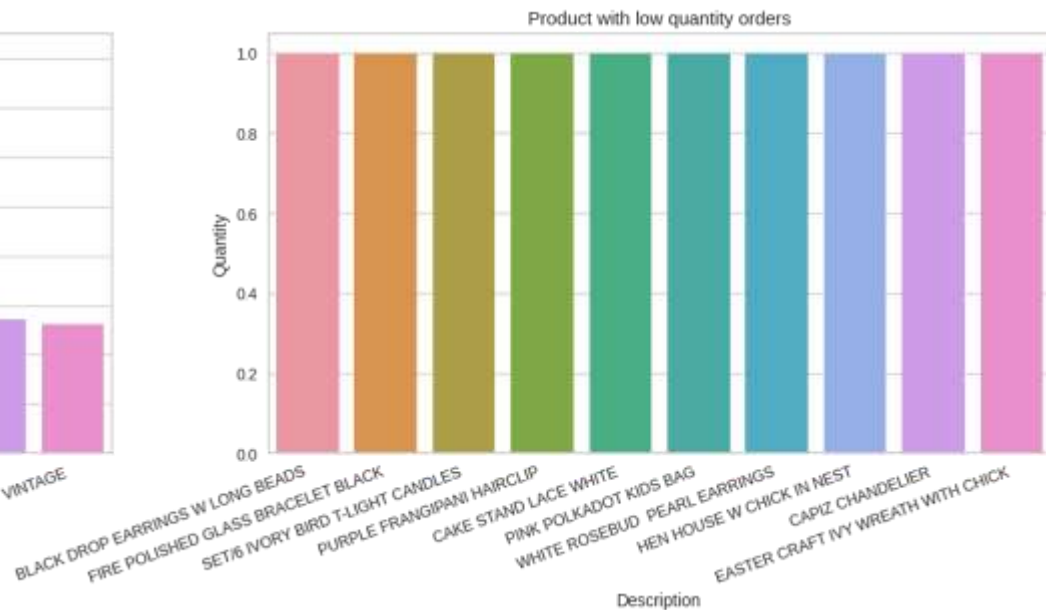
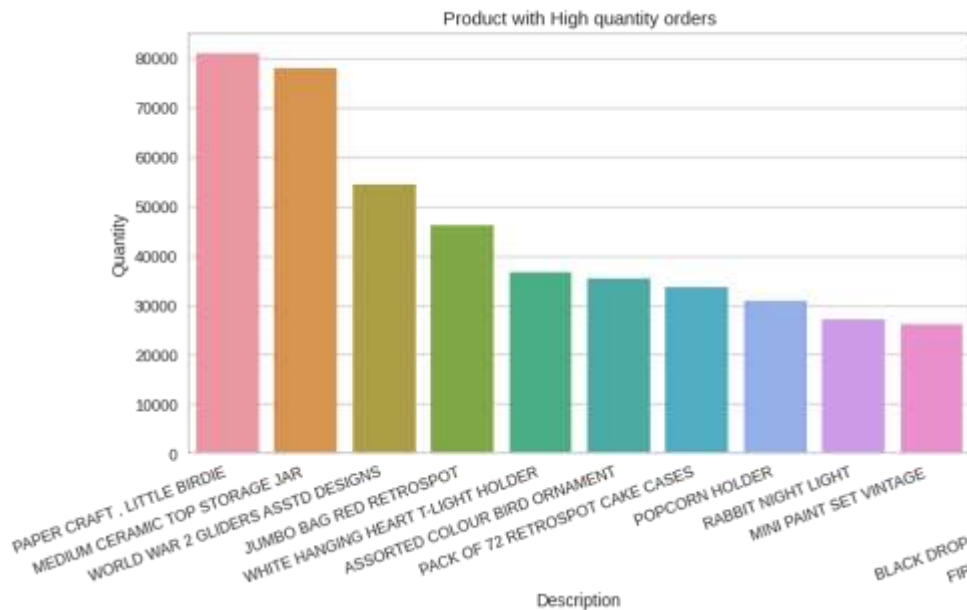


# COUNTRY WISE PURCHASE QUANTITY



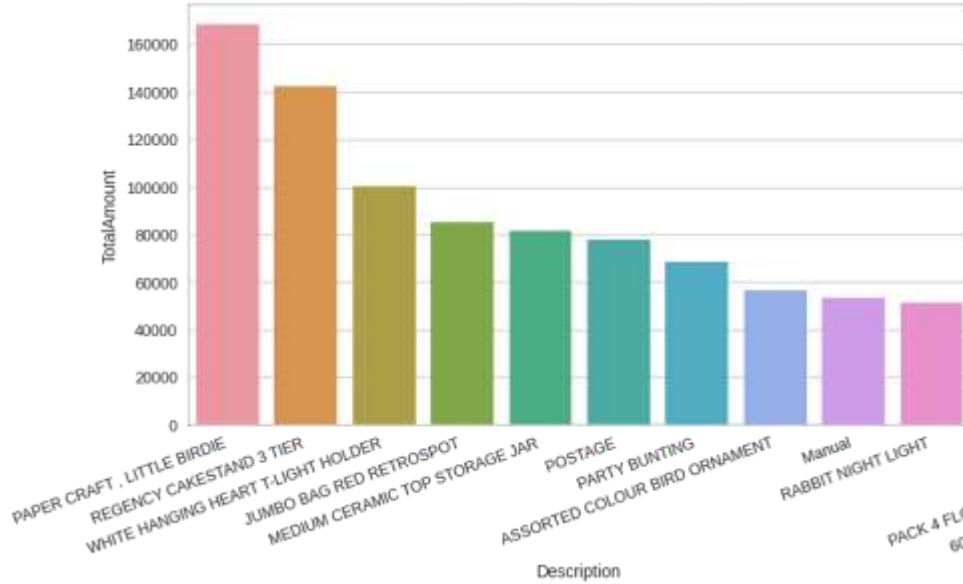


# PRODUCT WISE PURCHASE QUANTITY

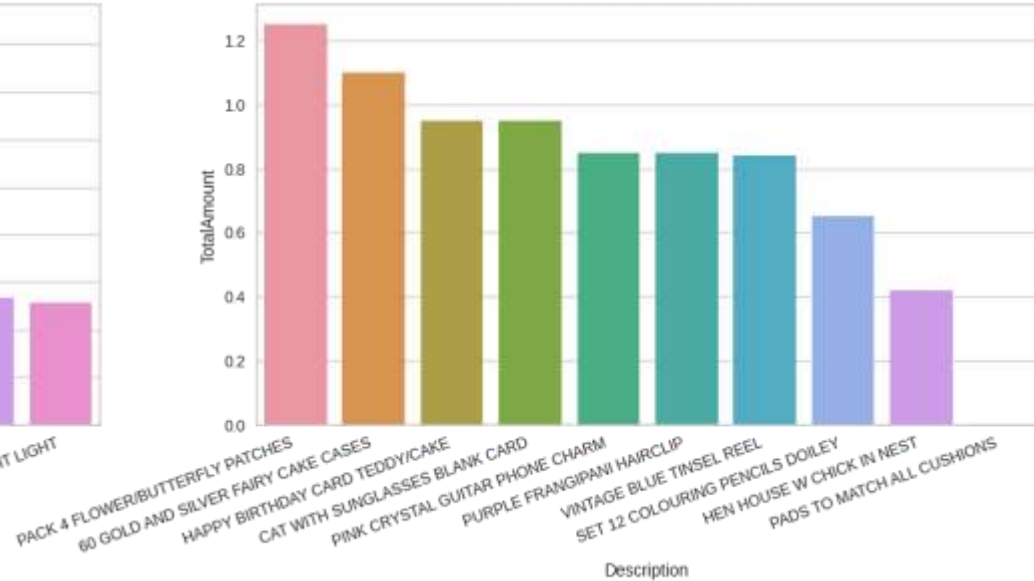


# PRODUCT WISE REVENUE

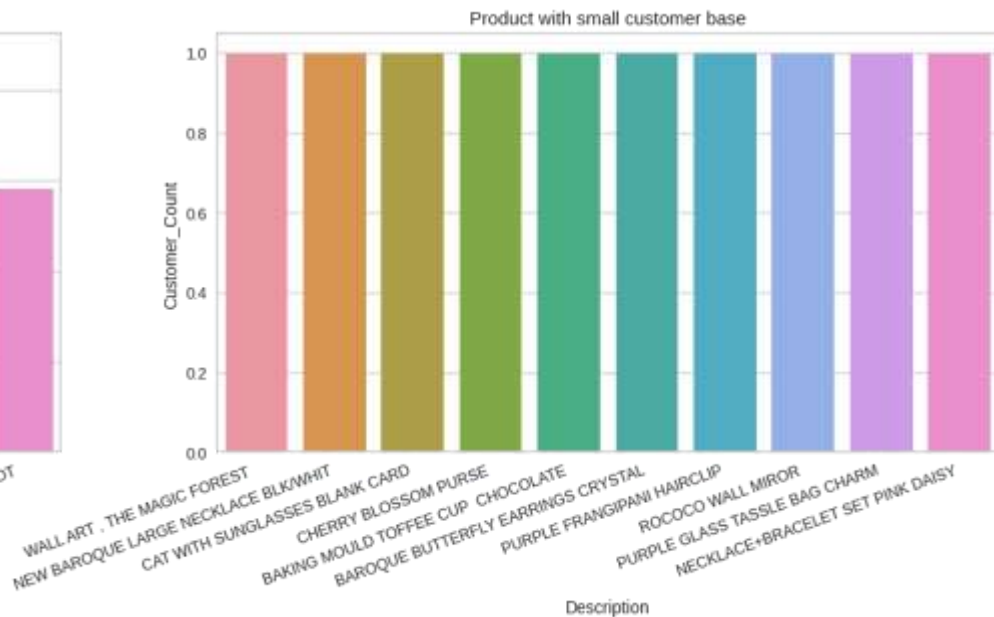
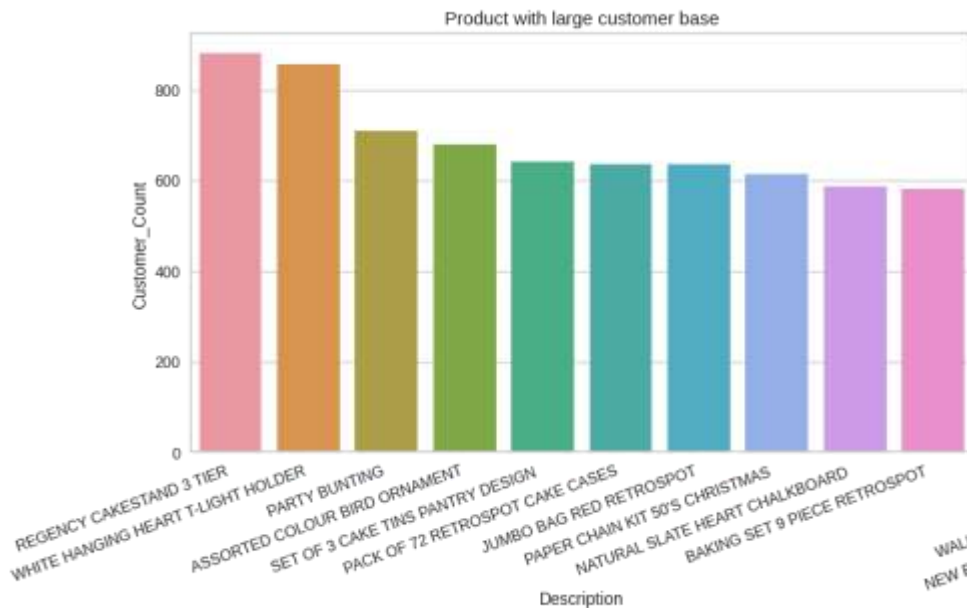
Product that made most of the revenue



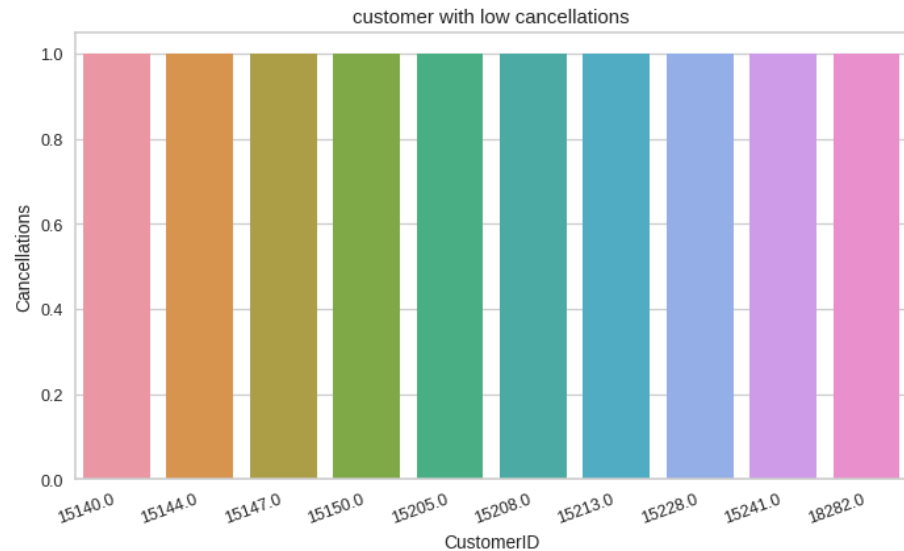
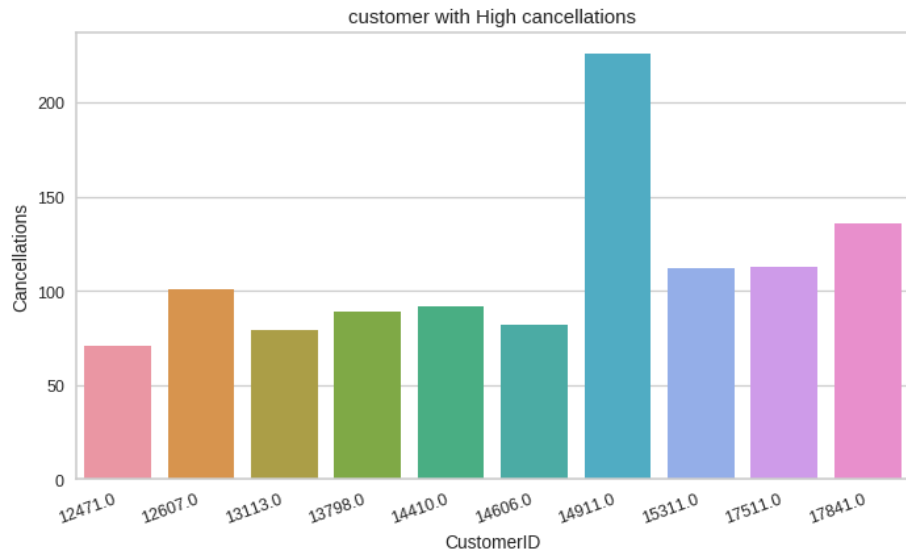
Product that made least revenue



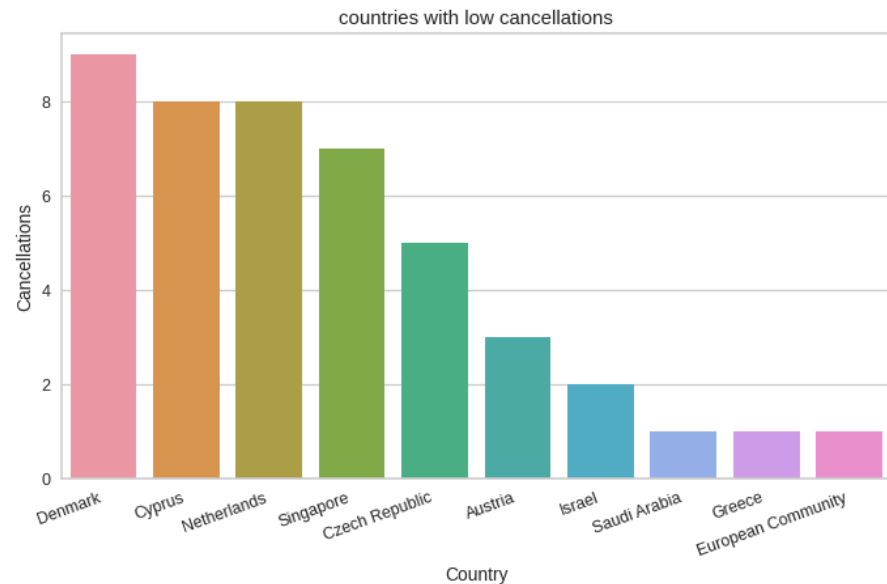
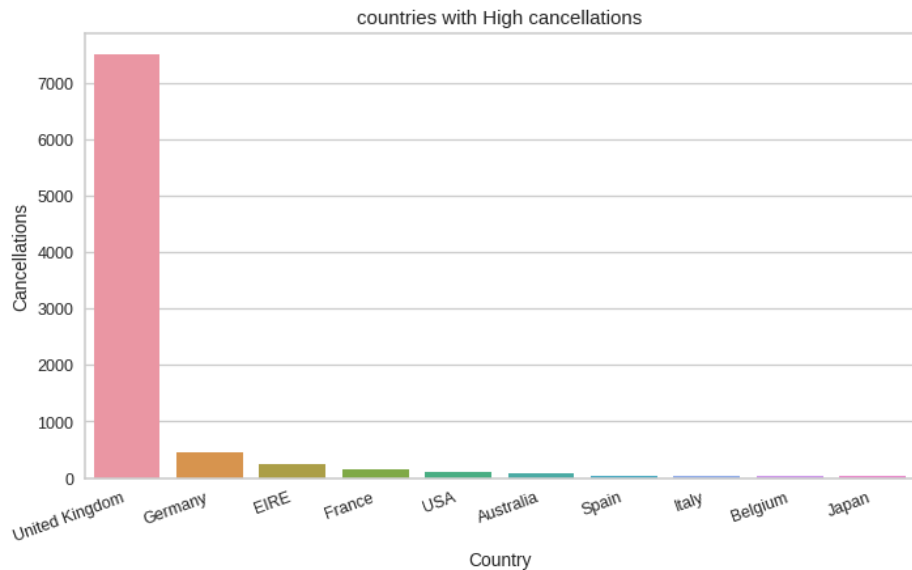
# PRODUCT WISE CUSTOMERS



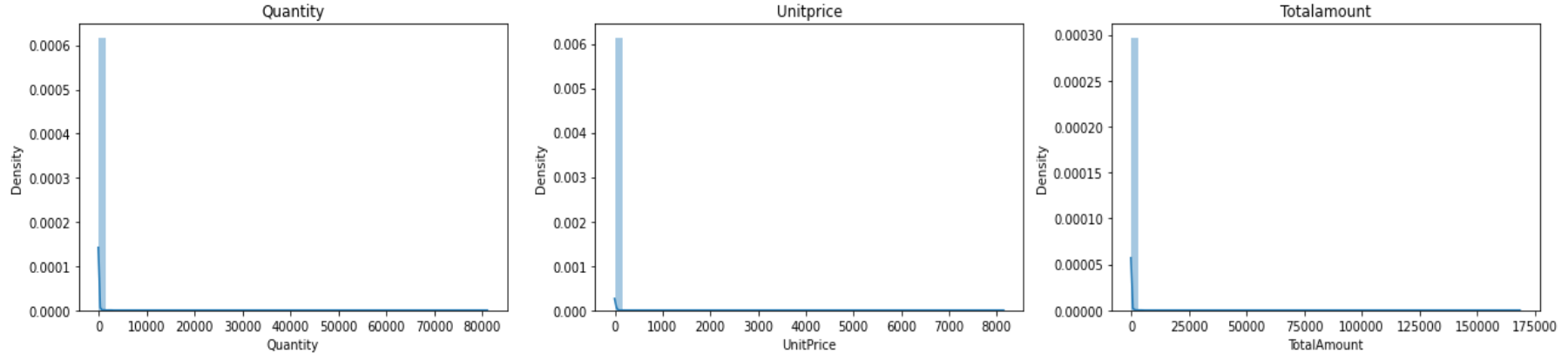
# CUSTOMER WISE CANCELLATIONS



# COUNTRY WISE CANCELLATIONS



# VISUALIZING DISTRIBUTIONS

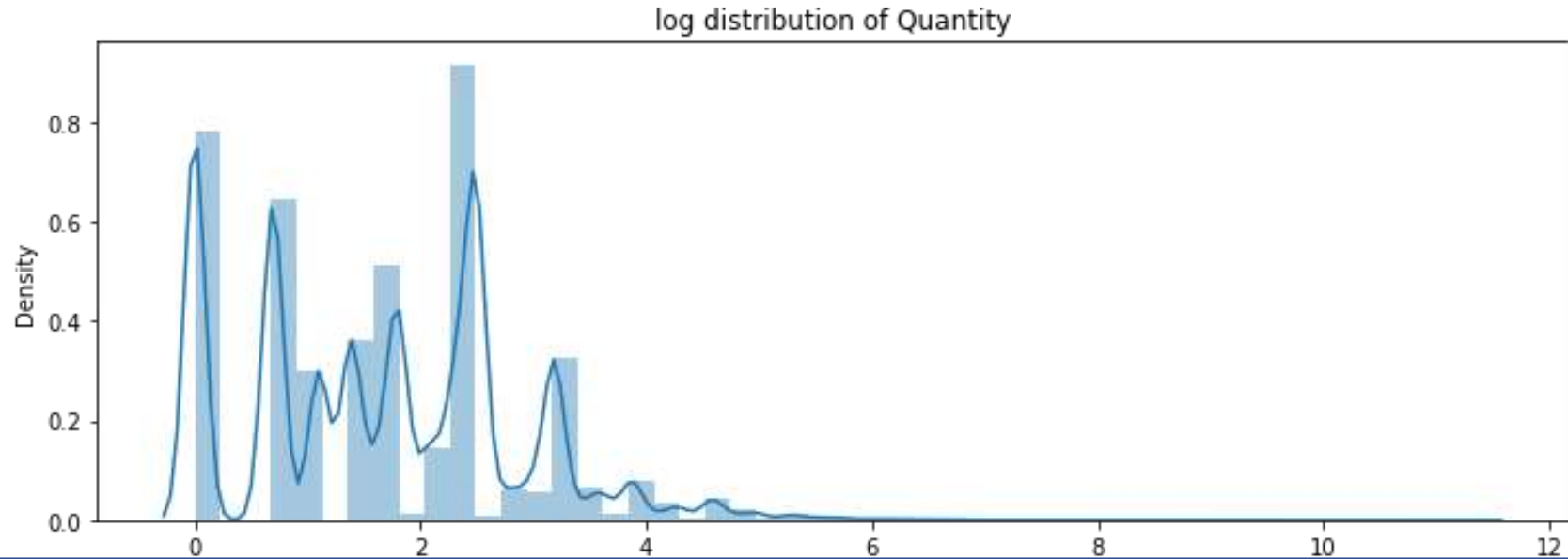


1. Visualizing the distribution of quantity, unitprice and total amount columns

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  $\text{mean} > \text{median} > \text{mode}$

3. For symmetric graph  $\text{mean} = \text{median} = \text{mode}$ .

# 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.

# Recency Frequency Monetary values

## RFM Metrics



### RECENCY

The freshness of the customer activity, be it purchases or visits

E.g. Time since last order or last engaged with the product



### FREQUENCY

The frequency of the customer transactions or visits

E.g. Total number of transactions or average time between transactions/engaged visits



### MONETARY

The intention of customer to spend or purchasing power of customer

E.g. Total or average transactions value



# RFM MODELLING

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

**RFM TABLE**

## CONCLUSION

**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

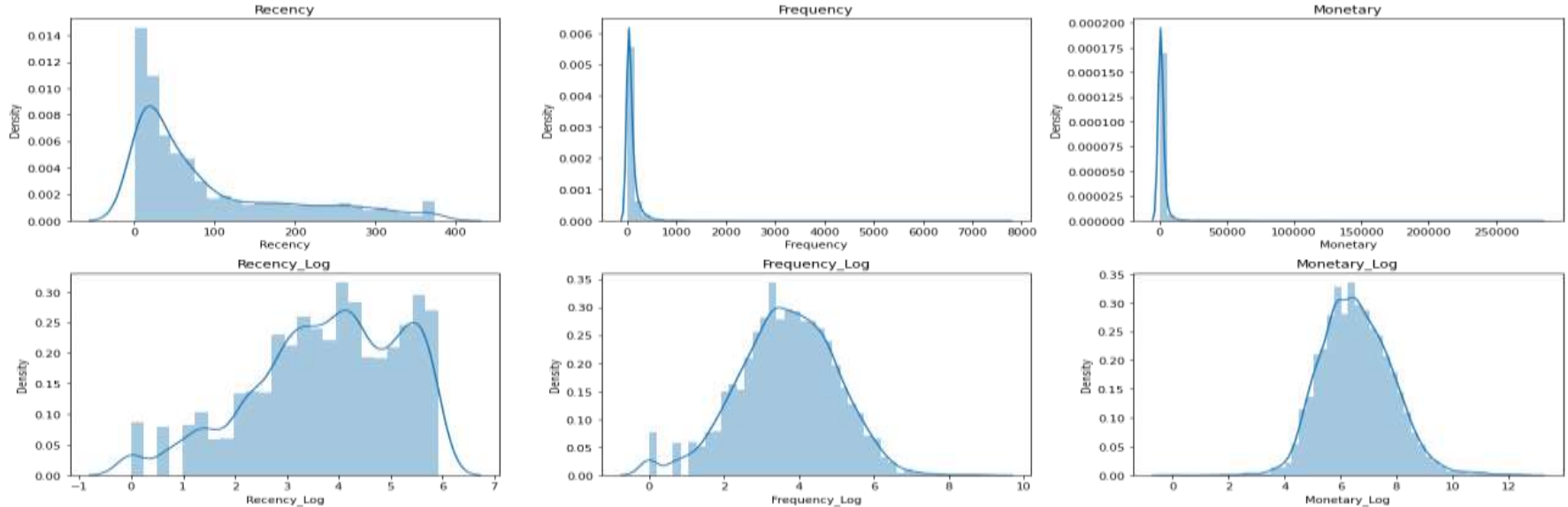
**Recently visited Potential Customer**

**Syed**

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

**About to Lose Average Customer**

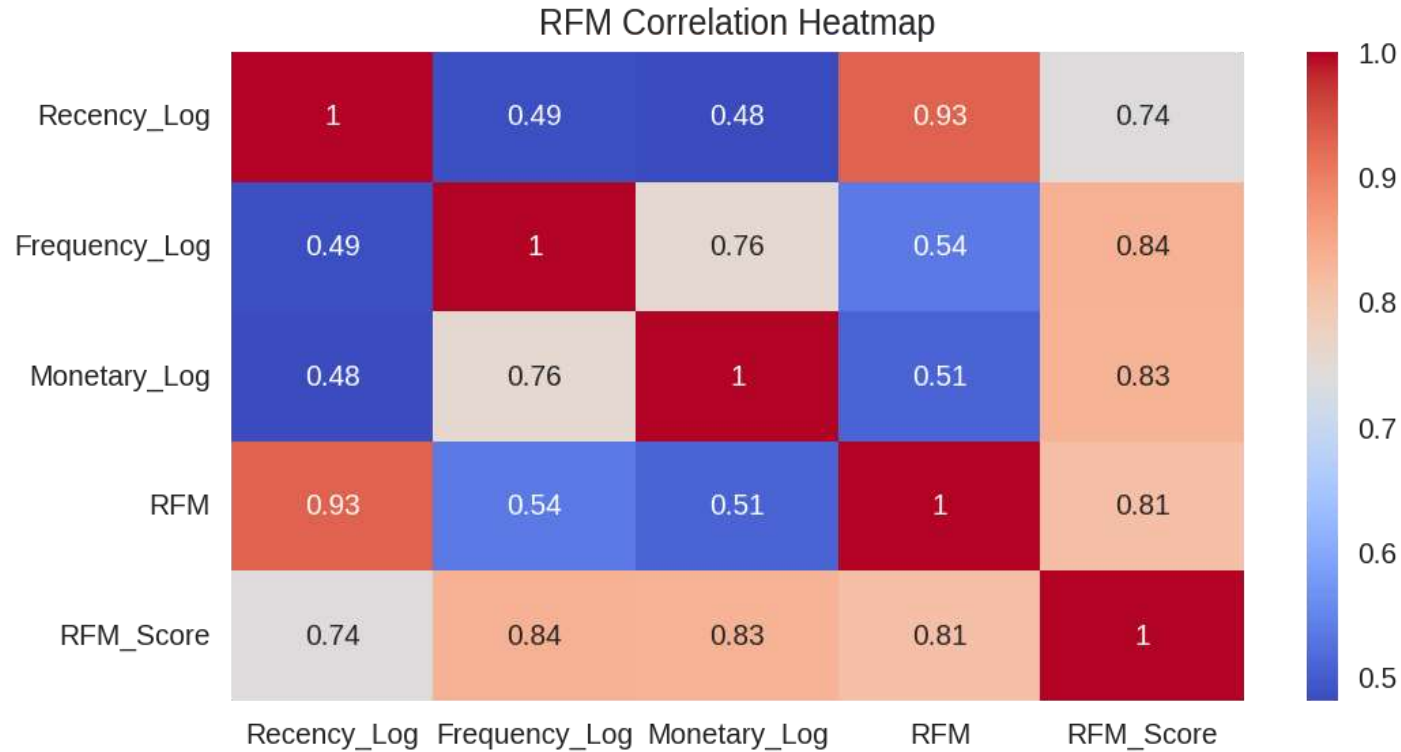
# RFM MODELLING



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.

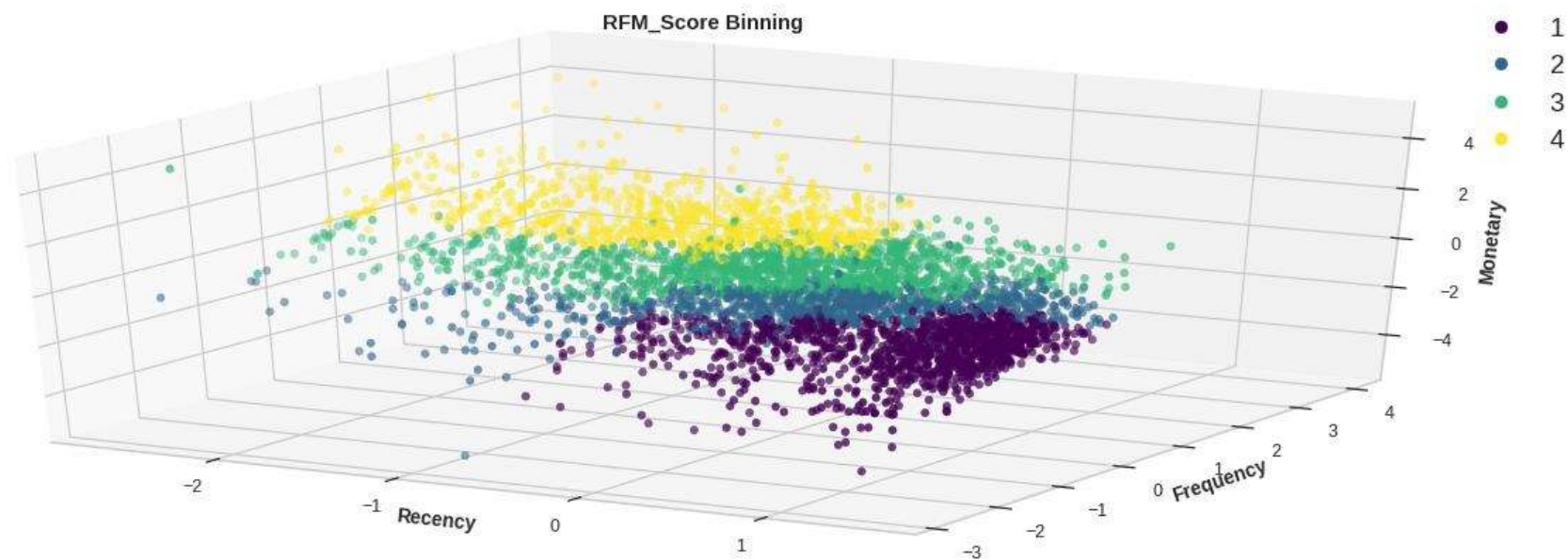
# RFM CORRELATION HEATMAP



1. We can see that Recency is highly correlated with the RFM value.

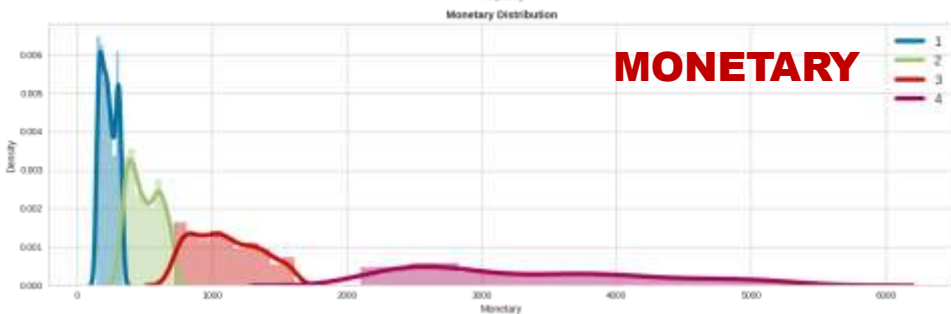
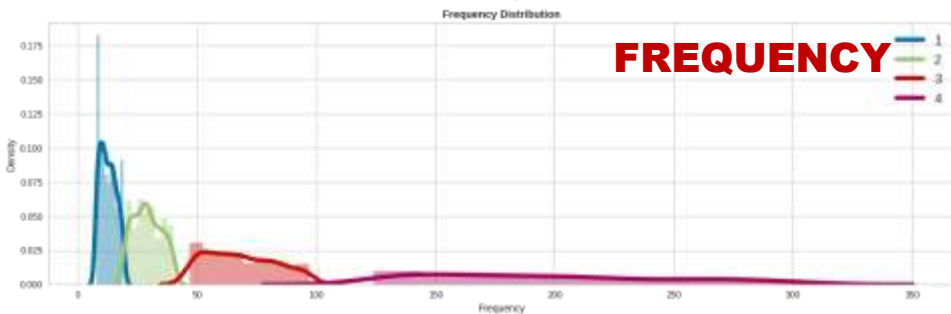
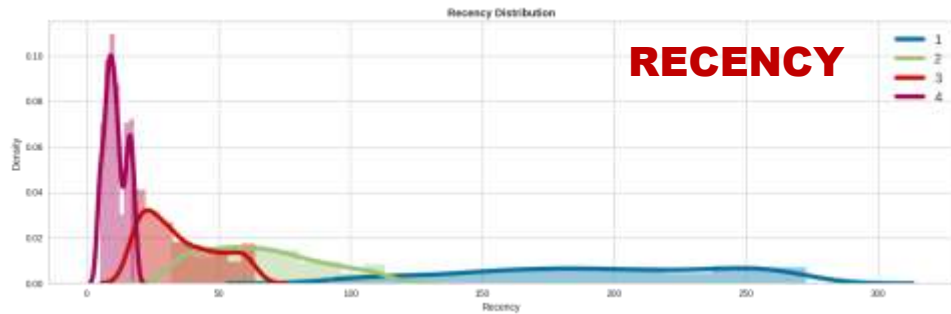
2. Frequency and Monetary are moderately correlated with the RFM.

# 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

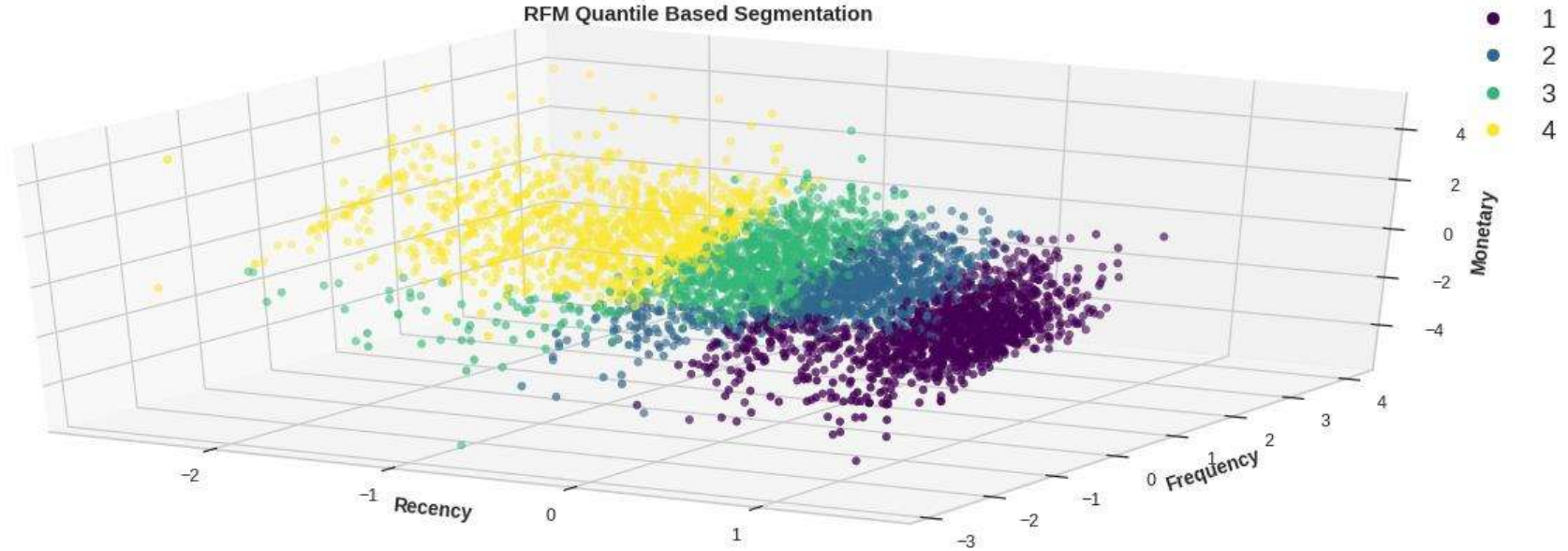
# BINNING RFM SCORES



Binning	Last_visited	Purchase_frequency	Money_spent
1	93 to 274 days ago	Bought 7 to 20 times	Spent around 142 to 335 Sterling
2	31 to 114 days ago	Bought 19 to 41 times	Spent around 327 to 725 Sterling
3	16 to 65 days ago	Bought 46 to 98 times	Spent around 717 to 1613 Sterling
4	4 to 19 days ago	Bought 123 to 305 times	Spent around 2093 to 5398 Sterling

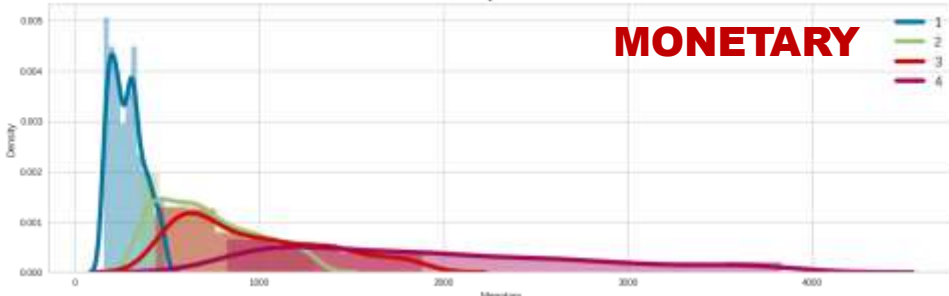
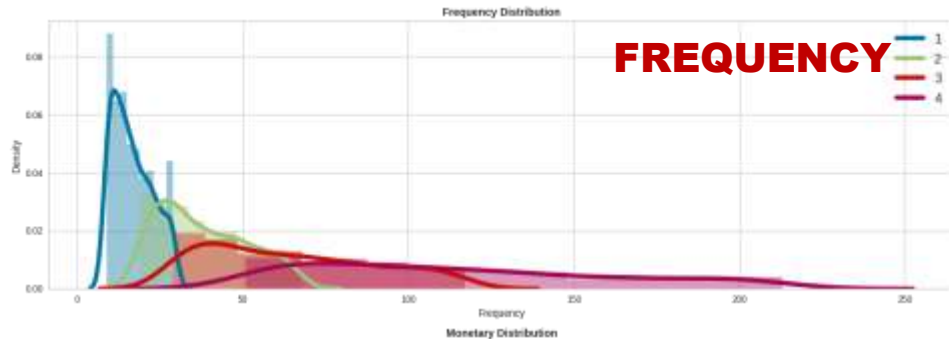
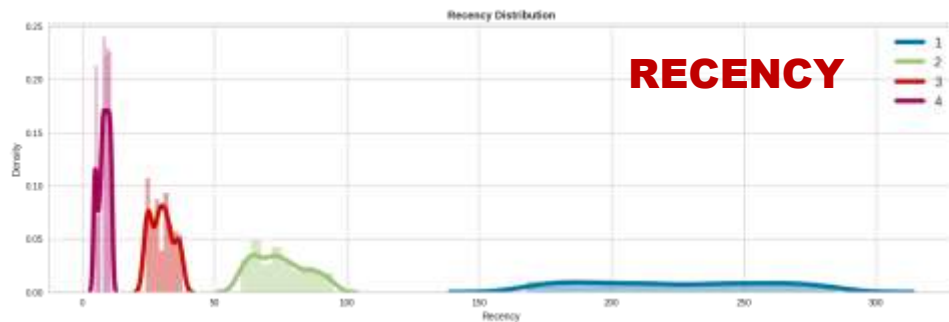


# QUANTILE CUT



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

# QUANTILE CUT



QuantileCut	Last_visited	Purchase_frequency	Money_spent
1	166 to 286 days ago	Bought 8 to 30 times	Spent around 156 to 486 Sterling
2	59 to 96 days ago	Bought 18 to 69 times	Spent around 355 to 1301 Sterling
3	23 to 39 days ago	Bought 28 to 118 times	Spent around 439 to 1887 Sterling
4	4 to 12 days ago	Bought 50 to 214 times	Spent around 822 to 3849 Sterling

GROUP 1

LOST POOR CUSTOMERS

GROUP 2

LOSING LOYAL CUSTOMERS

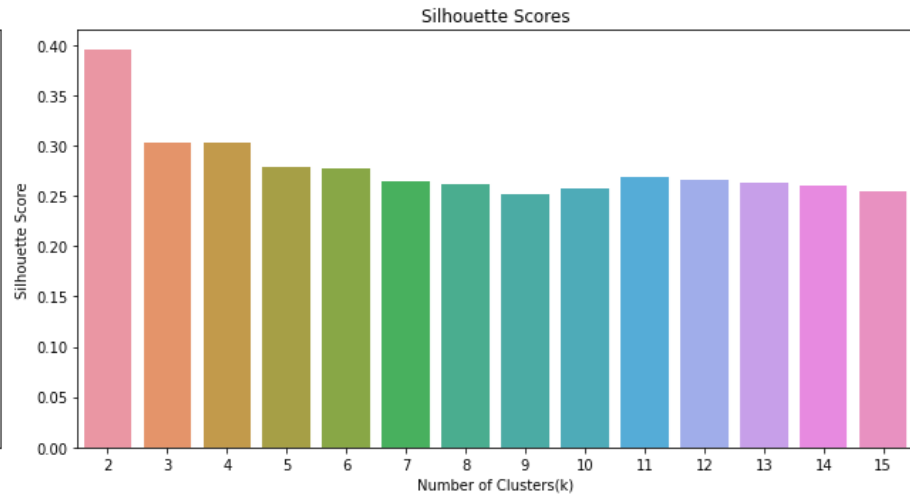
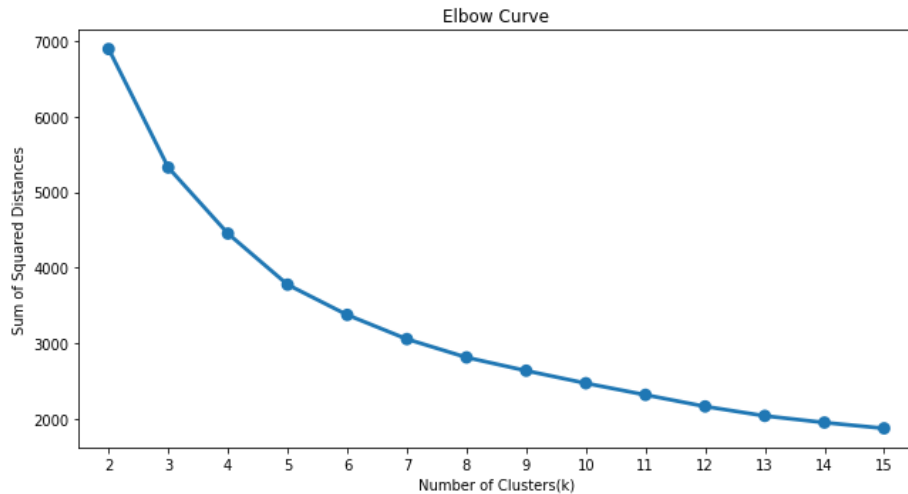
GROUP 3

GOOD CUSTOMERS

GROUP 4

BEST CUSTOMERS

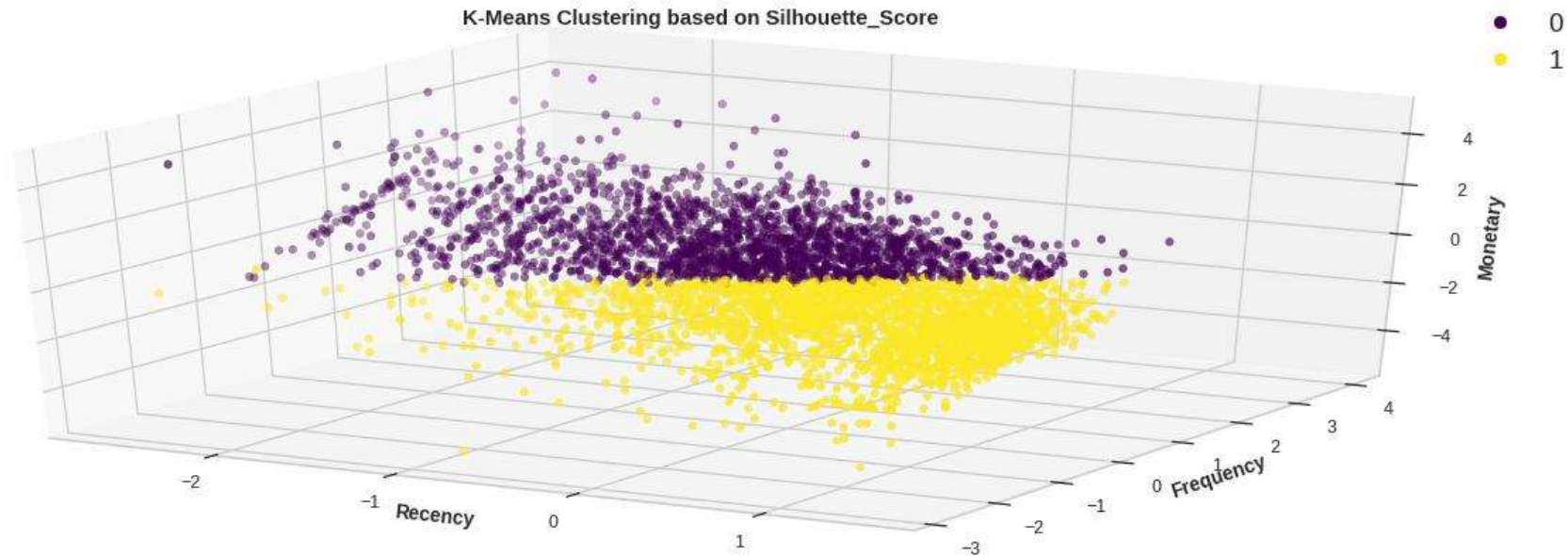
# K-MEANS CLUSTERING



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. 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.

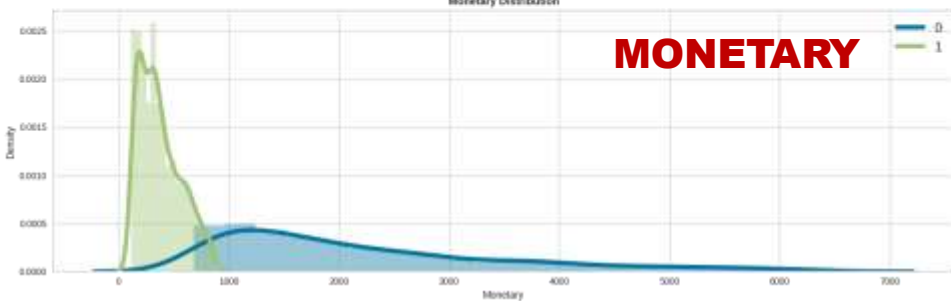
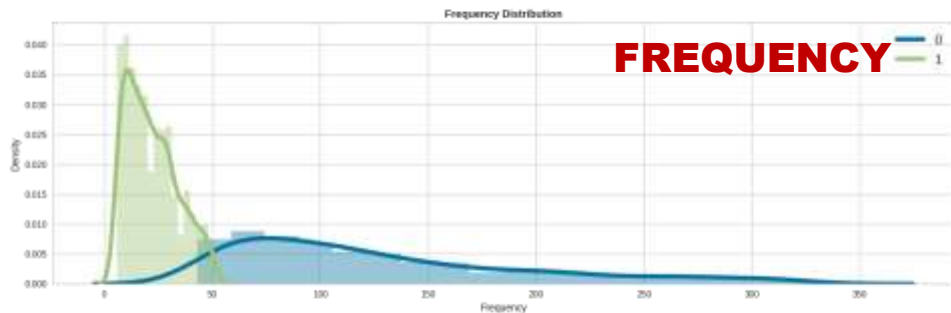
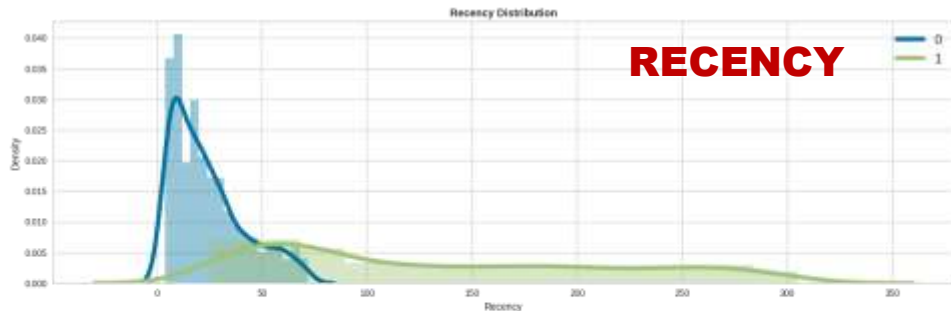


# K-MEANS (2 CLUSTER)



K-Means 2Cluster	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	31.074883	18.000000	173.084763	108.000000	4029.985352	1823.520000	1923
1	141.423841	109.000000	24.788907	20.000000	470.839430	331.210000	2416

# K-MEANS (2 CLUSTER)



K-Means 2Cluster	Last_visited	Purchase_frequency	Money_spent
0	8 to 38 days ago	Bought 67 to 192 times	Spent around 1068 to 3350 Sterling
1	51 to 227 days ago	Bought 10 to 33 times	Spent around 189 to 571 Sterling

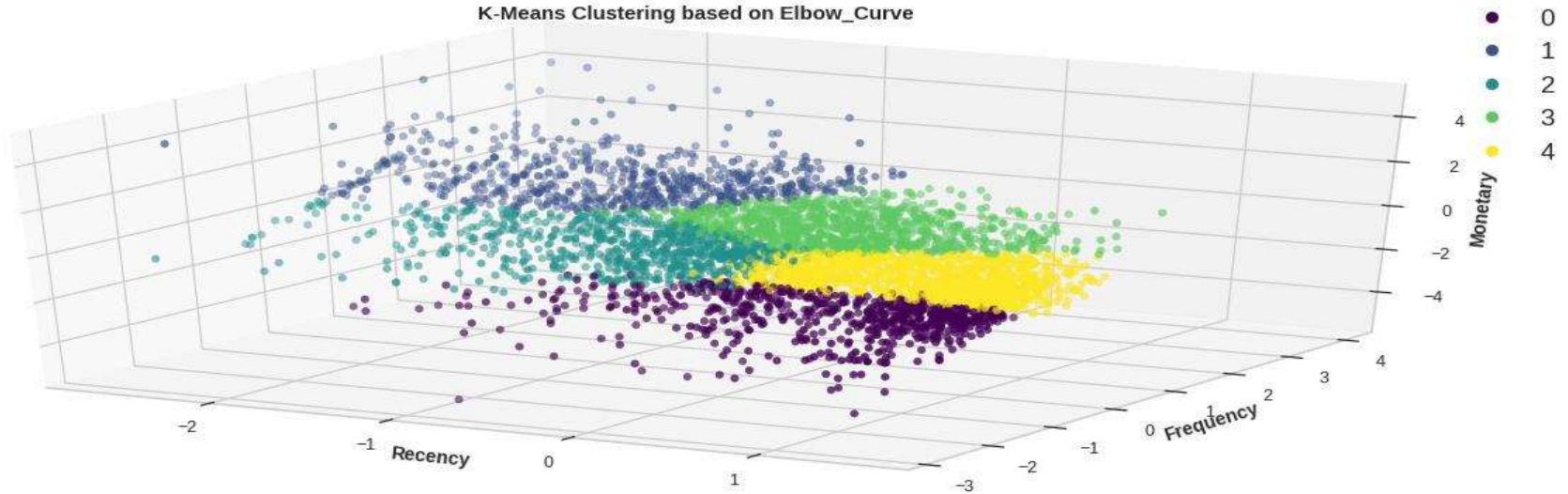
GROUP 0

BEST CUSTOMERS

GROUP 1

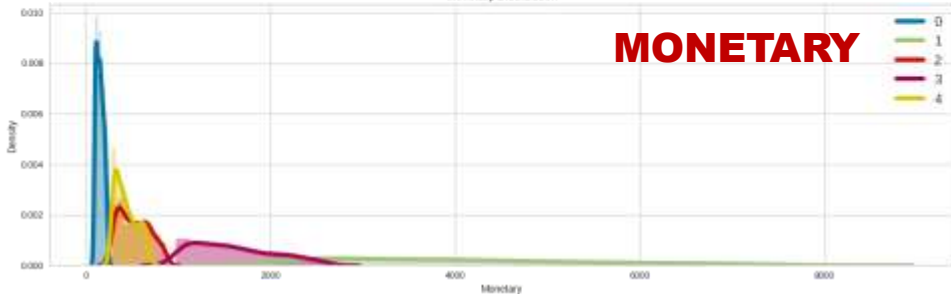
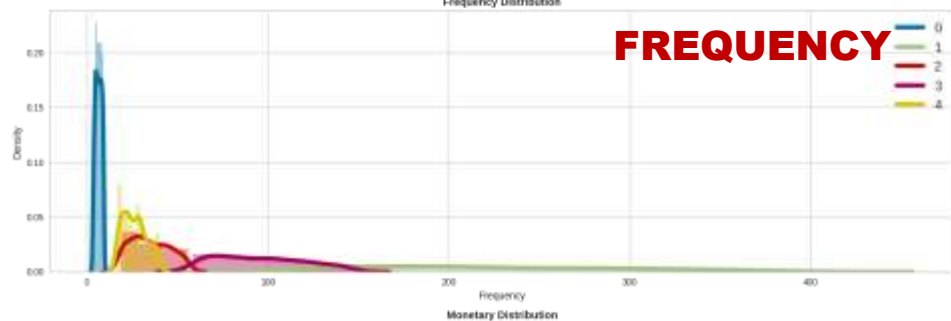
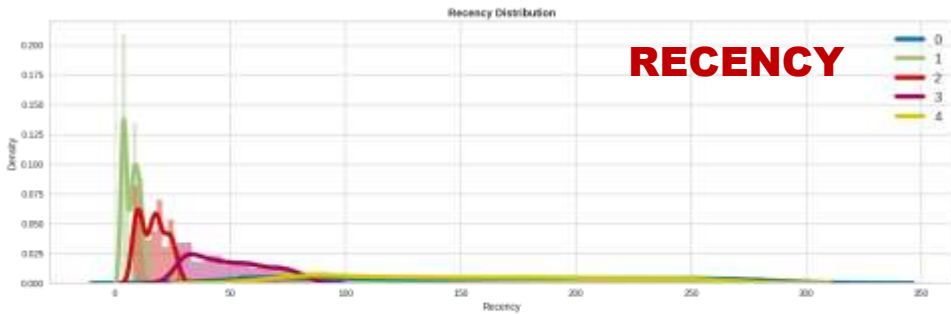
LOST POOR CUSTOMERS

# K-MEANS (5 CLUSTER)



K-Means 5Cluster	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	170.579330	170.000000	6.950509	7.000000	198.009854	152.550000	687
1	9.066265	7.000000	314.746988	212.000000	8374.983886	3803.320000	664
2	17.418848	17.000000	40.079843	34.000000	623.428522	512.180000	764
3	62.401942	46.000000	108.941748	94.000000	2040.923865	1531.530000	1030
4	168.046064	152.000000	30.262982	26.000000	512.342422	414.570000	1194

# K-MEANS (5 CLUSTERS)



K-Means 5Cluster	Last_visited	Purchase_frequency	Money_spent
0	64 to 265 days ago	Bought 4 to 10 times	Spent around 103 to 215 Sterling
1	3 to 12 days ago	Bought 128 to 340 times	Spent around 2289 to 6480 Sterling
2	8 to 25 days ago	Bought 20 to 52 times	Spent around 316 to 780 Sterling
3	29 to 75 days ago	Bought 63 to 135 times	Spent around 1058 to 2297 Sterling
4	78 to 246 days ago	Bought 18 to 37 times	Spent around 302 to 632 Sterling

GROUP 0

LOST POOR CUSTOMERS

GROUP 1

BEST CUSTOMERS

GROUP 2

RECENTLY VISITED AVERAGE  
CUSTOMERS

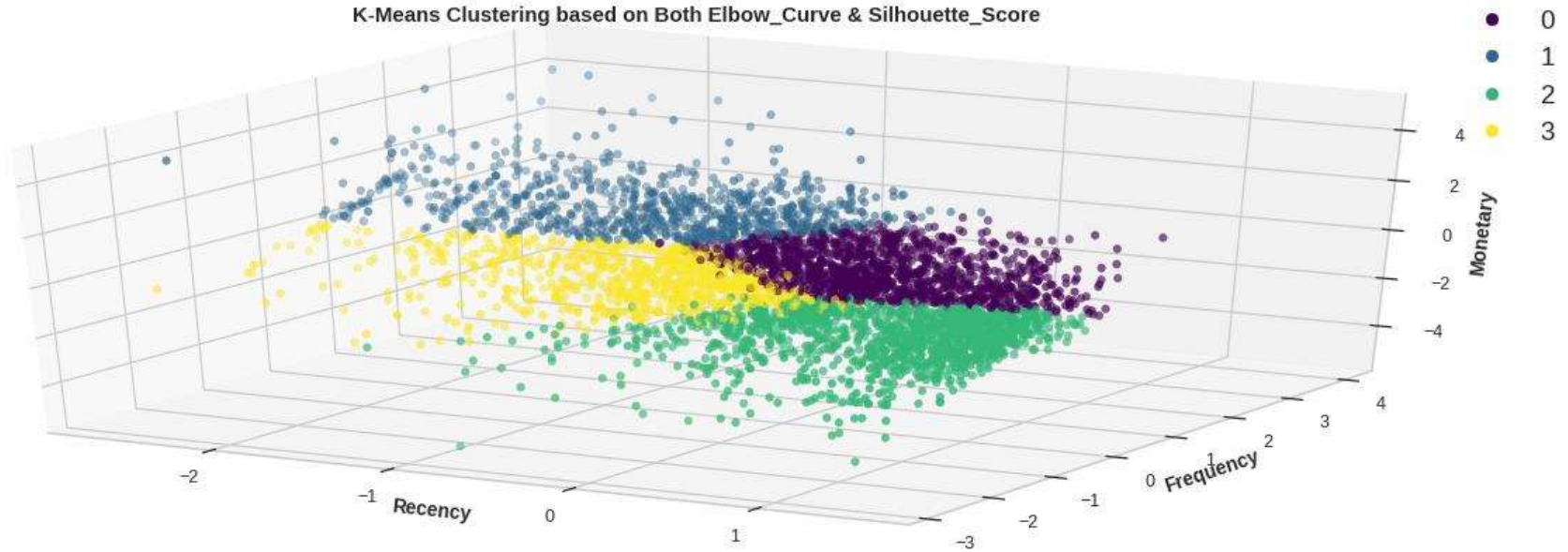
GROUP 3

LOSING LOYAL CUSTOMERS

GROUP 4

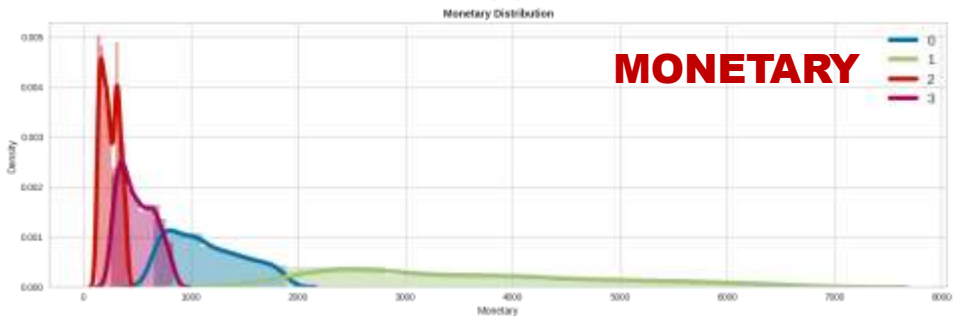
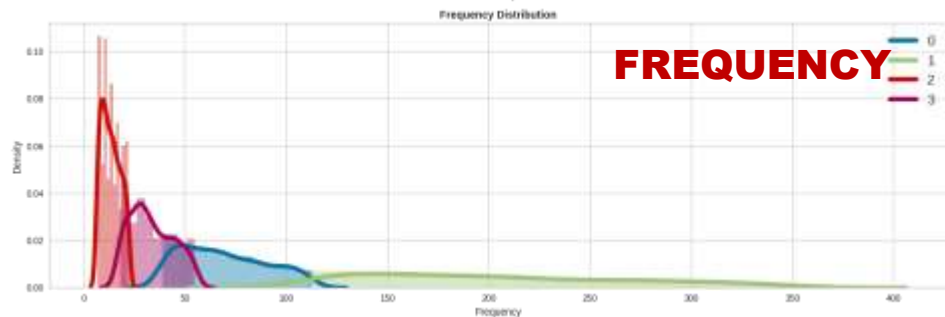
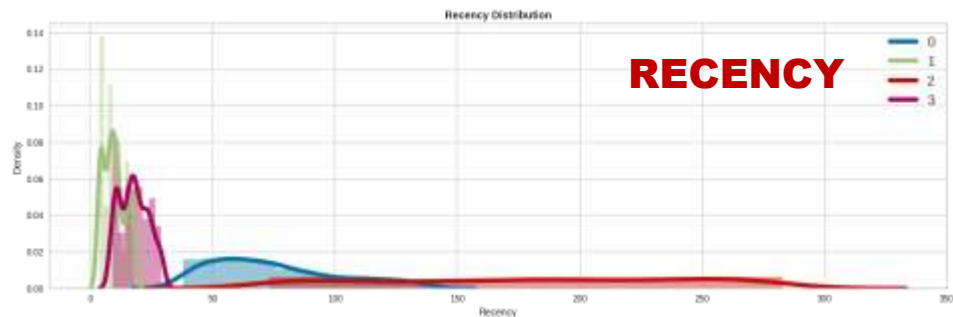
AVERAGE CUSTOMERS

# K-MEANS (4 CLUSTERS)



K-Means 4Cluster	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	93.587007	71.000000	80.149265	66.000000	1518.949902	1086.920000	1293
1	12.146108	9.000000	283.461078	193.000000	7212.437509	3347.310000	835
2	184.356313	184.000000	14.753991	12.500000	295.824551	240.275000	1378
3	19.500600	17.000000	38.509004	32.000000	592.048163	471.400000	833

# K-MEANS (4 CLUSTER)



K-Means 4Cluster	Last_visted	Purchase_frequency	Money_spent
0	43 to 120 days ago	Bought 42 to 103 times	Spent around 709 to 1706 Sterling
1	4 to 17 days ago	Bought 120 to 309 times	Spent around 2071 to 5609 Sterling
2	81 to 268 days ago	Bought 7 to 21 times	Spent around 144 to 368 Sterling
3	9 to 28 days ago	Bought 20 to 52 times	Spent around 293 to 744 Sterling

GROUP 0

LOSING LOYAL CUSTOMERS

GROUP 1

BEST CUSTOMERS

GROUP 2

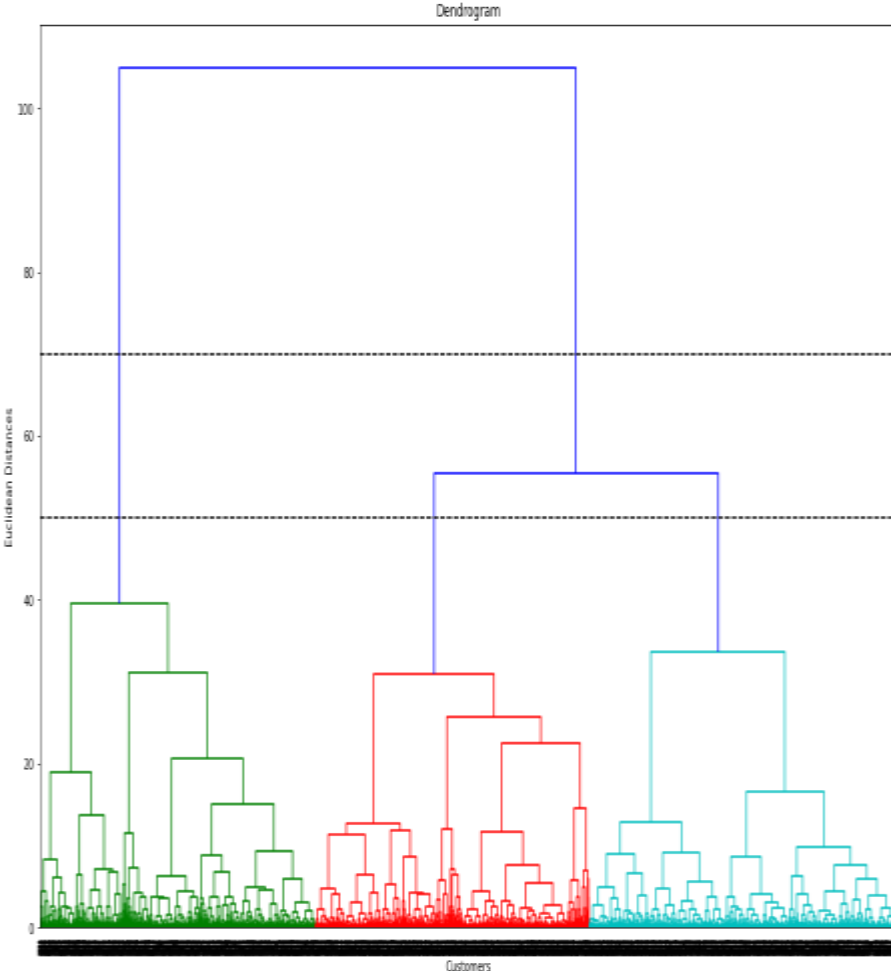
LOST POOR CUSTOMERS

GROUP 3

RECENTLY VISITED AVERAGE CUSTOMERS



# HIERARCHICAL CLUSTERING



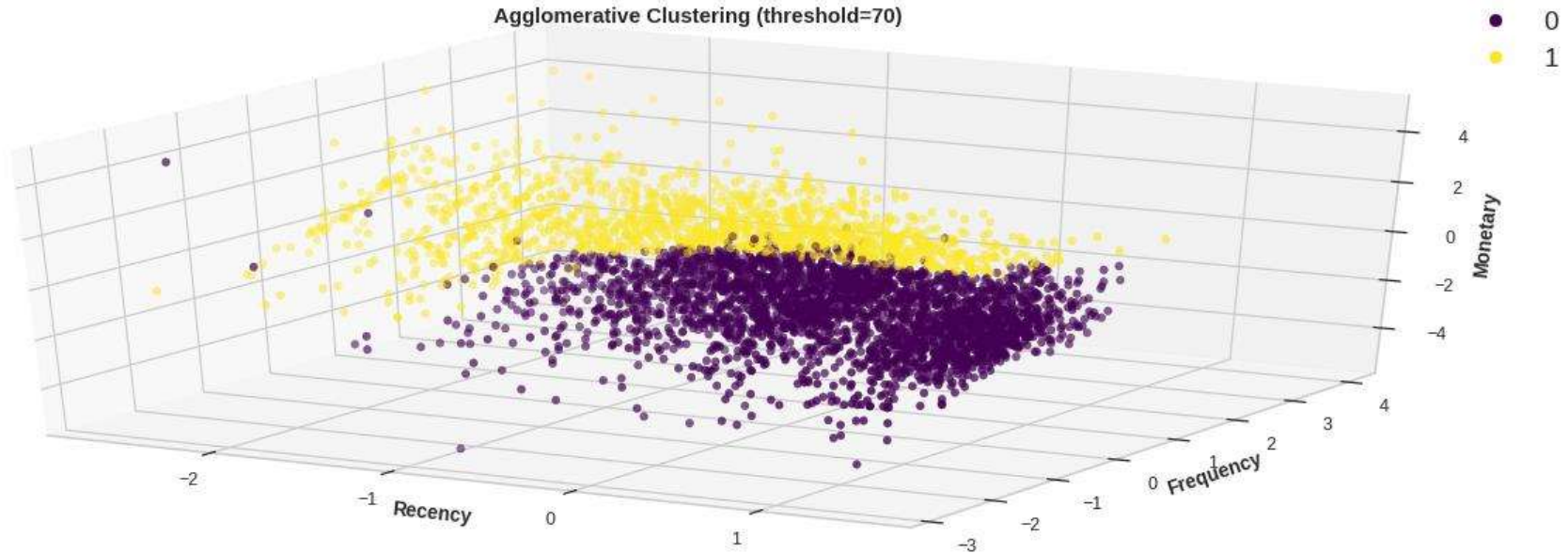
In the K-means clustering there is a challenge to predetermine the number of clusters, and it always tries to create the clusters of the same size. To solve these two challenges, we can opt for the hierarchical clustering algorithm because, in this algorithm, we don't need to have knowledge about the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

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.

← We have defined the optimal number of clusters based on dendrogram as shown here

# HIERARCHICAL (2 CLUSTERS)



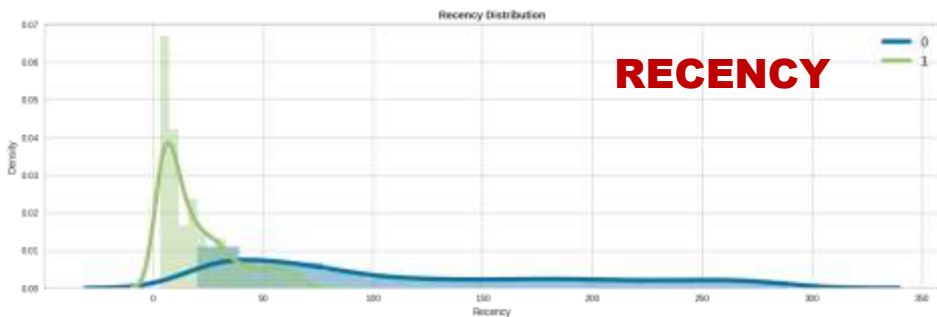
hierarchical 2Cluster	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	123.608696	80.000000	33.610394	25.000000	684.108391	409.685000	2944
1	26.905376	12.000000	210.597133	135.000000	4927.021356	2404.170000	1395



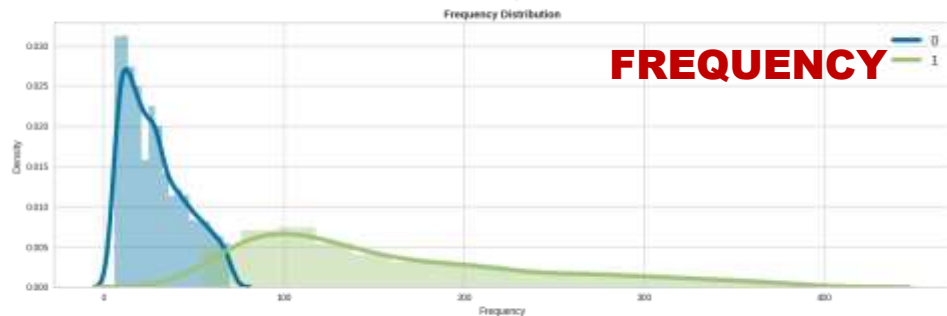
# HIERARCHICAL (2 CLUSTERS)



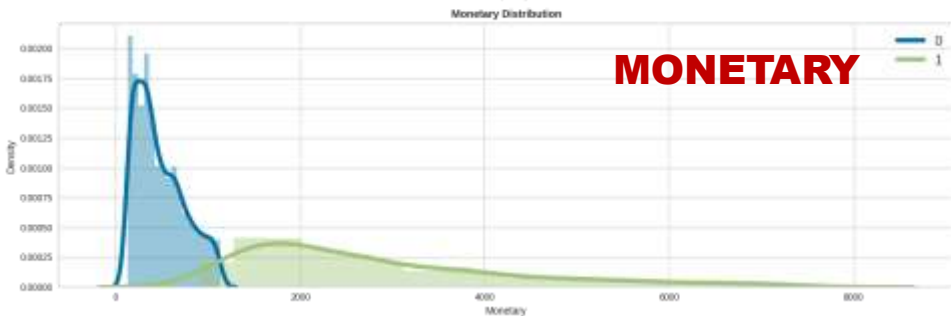
**REGENCY**



**FREQUENCY**



**MONETARY**



hierarchical 2Cluster	Last_visited	Purchase_frequency	Money_spent
0	36 to 200 days ago	Bought 12 to 46 times	Spent around 219 to 739 Sterling
1	4 to 33 days ago	Bought 87 to 233 times	Spent around 1546 to 4020 Sterling

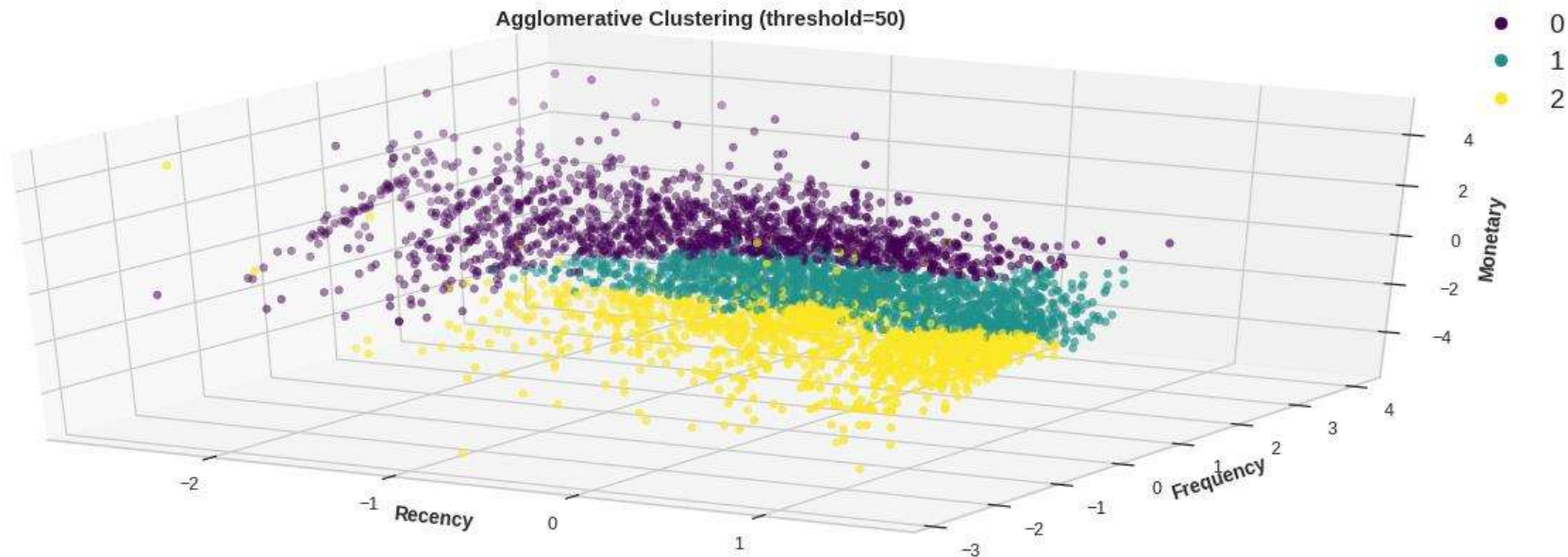
**GROUP 0**

**AVERAGE CUSTOMERS**

**GROUP 1**

**BEST CUSTOMERS**

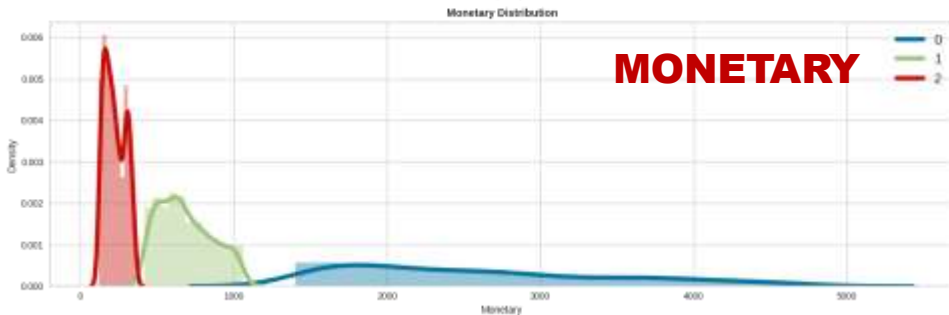
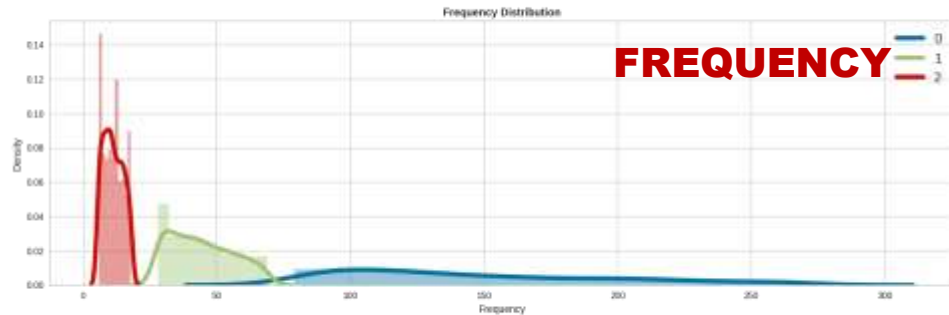
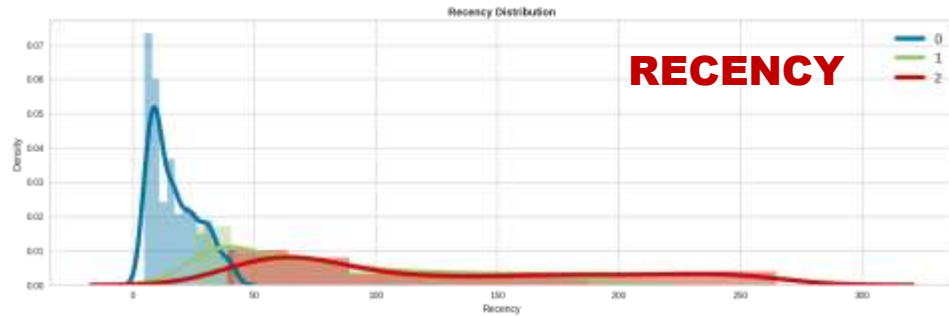
# HIERARCHICAL (3 CLUSTERS)



hierarchical 3Cluster	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 (3 CLUSTERS)

AI



hierarchical 3Cluster	last_visited	Purchase_frequency	Money_spent
0	4 to 33 days ago	Bought 87 to 233 times	Spent around 1546 to 4020 Sterling
1	30 to 163 days ago	Bought 29 to 63 times	Spent around 463 to 977 Sterling
2	49 to 243 days ago	Bought 6 to 18 times	Spent around 139 to 327 Sterling

GROUP 0

BEST CUSTOMERS

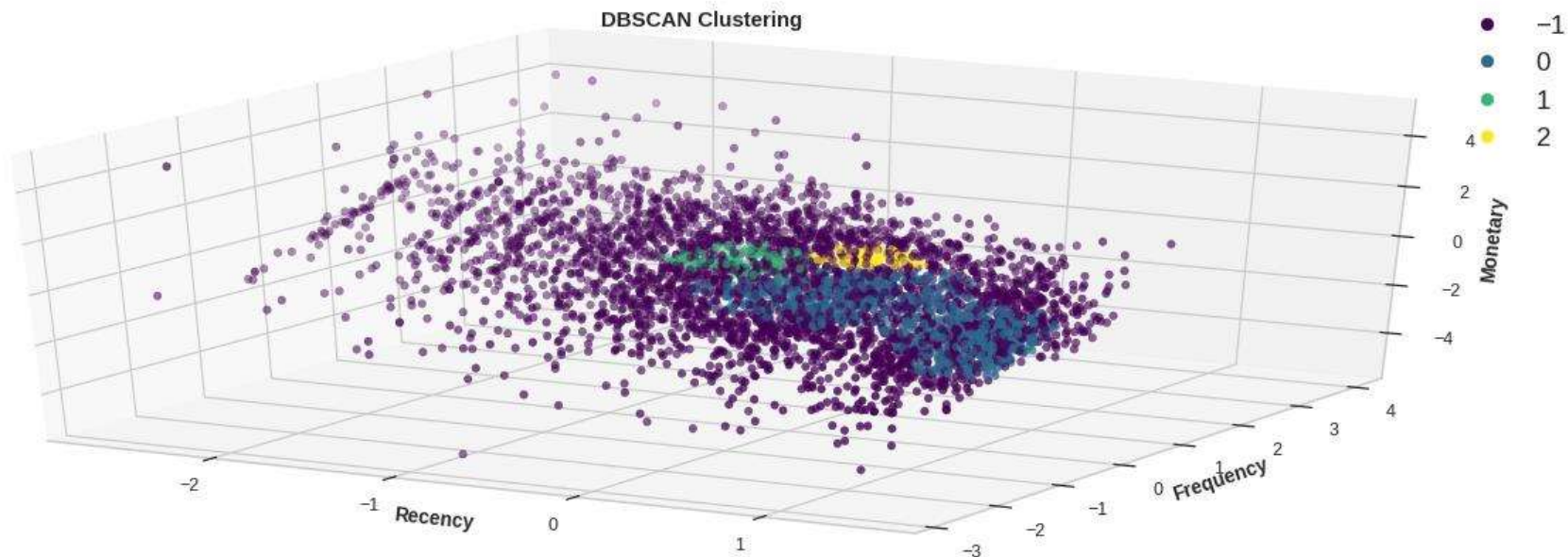
GROUP 1

LOSING LOYAL CUSTOMERS

GROUP 2

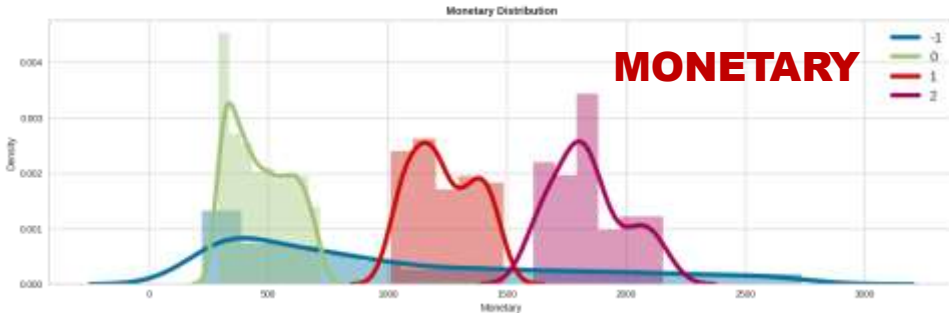
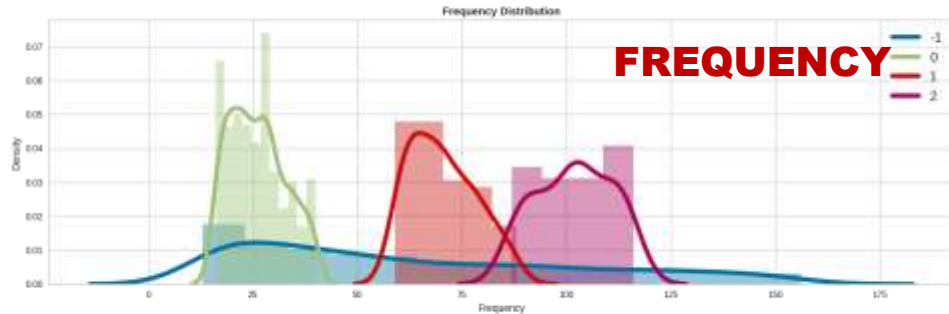
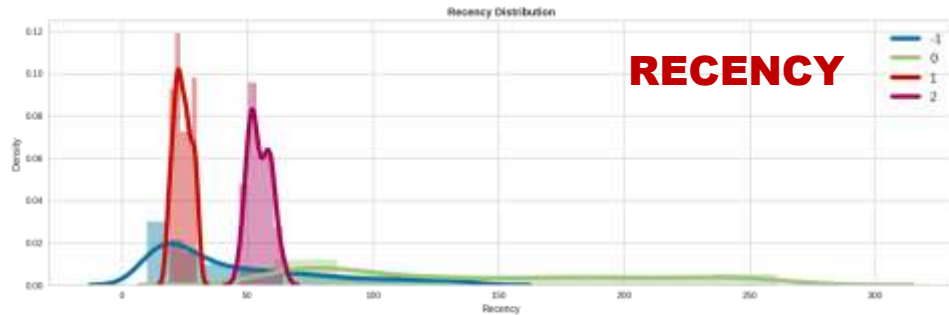
LOST POOR CUSTOMERS

# DBSCAN



DBSCAN	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	Last_visited	Purchase_frequency	Money_spent
-1	11 to 107 days ago	Bought 15 to 130 times	Spent around 278 to 2245 Sterling
0	66 to 245 days ago	Bought 17 to 37 times	Spent around 309 to 658 Sterling
1	19 to 30 days ago	Bought 60 to 84 times	Spent around 1052 to 1425 Sterling
2	50 to 62 days ago	Bought 89 to 114 times	Spent around 1637 to 2094 Sterling

GROUP -1

AVERAGE CUSTOMERS

GROUP 0

LOST POOR CUSTOMERS

GROUP 1

GOOD CUSTOMERS

GROUP 2

LOSING LOYAL CUSTOMERS

# SUMMARY & CONCLUSION

Clusterer	Binning	Quantile Cut	K-Means	K-Means	K-Means	Agglomerative	Agglomerative	DBSCAN	
Criterion	RFM Score Binning	RFM Quantile Cut	Elbow Curve	Silhouette Score	Elbow Curve & Silhouette Score	Dendrogram (y=70)	Dendrogram (y=50)	eps=0.2, min_samples=25	
Segments	4	4	5	2		4	2	3	4

- 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.
- 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.
- We created several useful clusters of customers on the basis of different metrics and methods to categorize the customers on the basis of their behavioral attributes to define their valuability, 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.
- 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.

# CUSTOMER SEGMENTS OBTAINED FROM CLUSTERING

## FINAL CONCLUSION

	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 2Cluster	No	Yes	No	No	Yes	No
hierarchical 3Cluster	Yes	No	No	No	Yes	Yes
DBSCAN	Yes	Yes	No	Yes	No	Yes

**THANK YOU**