Capstone Project - IV

ONLINE RETAIL CUSTOMER SEGMENTATION

BY

SHADAB MAHEMUD SHAIKH

Problem Definition

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Building a model to predict the optimal number of customers is essential for a business to understand customer behaviour, plan business strategies, marketing campaigns, etc. to target, incentivise and attract customer base.

Business Understanding

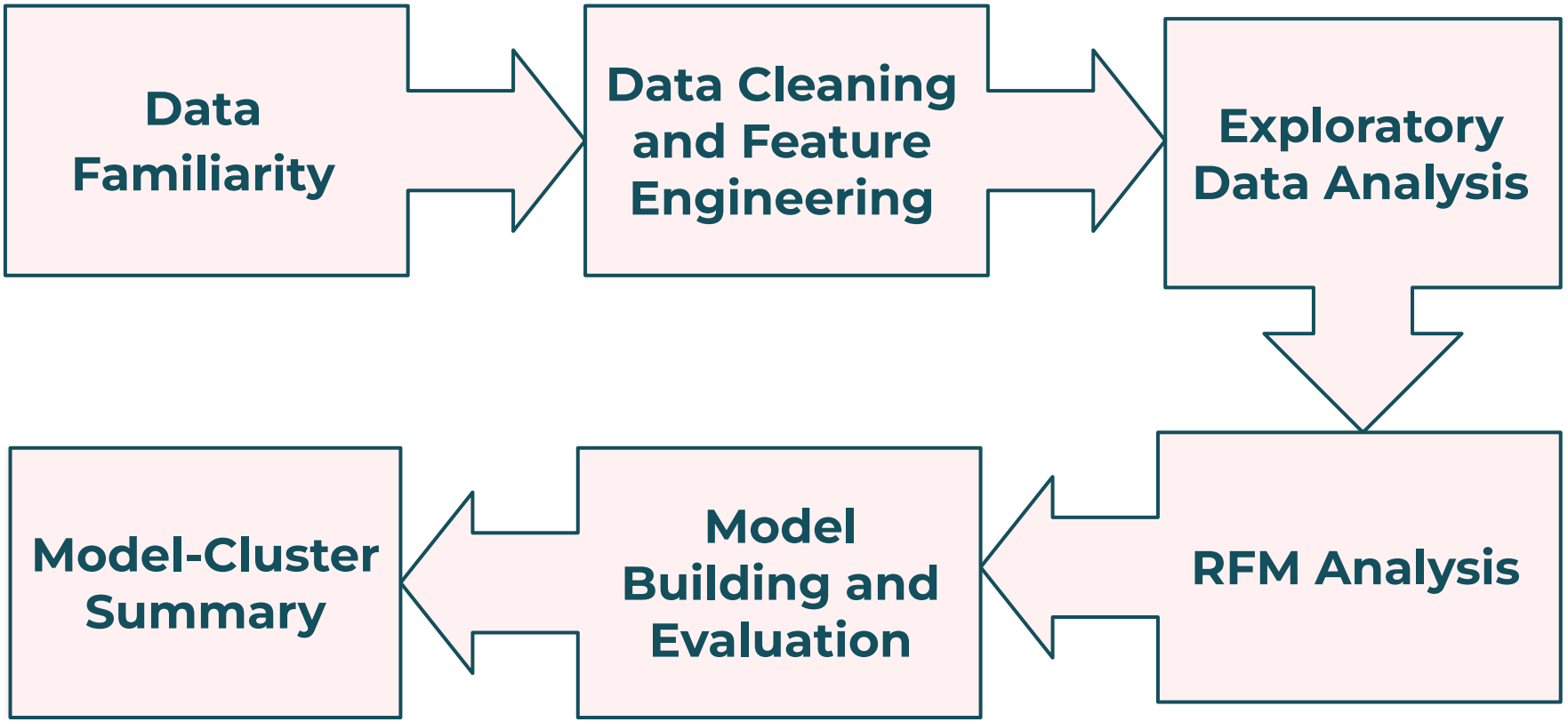
What is Customer Segmentation:

Customer segmentation is the process by which you divide your customers up based on common characteristics – such as demographics or behaviours, so you can market to those customers more effectively.

These customer segmentation groups can also be used to begin discussions of building a marketing persona. This is because customer segmentation is typically used to inform a brand's messaging, positioning and to improve how a business sells – so marketing personas need to be closely aligned to those customer segments in order to be effective.

For the scope of this project, we are dealing with obtaining major customer segments for a UK-based and registered non-store online retail company. This company specialises in gifting needs and most of its customers are B2C wholesalers.

Project Pipeline



Data Familiarity

Let’s understand our dataset...

- **Dataset Name:** TRANSNATIONAL ONLINE TRANSACTION Dataset

- **Size and Shape:** The shape of dataset is (541909 x 8)
  i.e, 541909 observations and 8 features

- **Missing Value Count:**
  - CustomerID - 135080 or 24.93%
  - Description - 1454 or 0.27%

- **Missing Value Count:** 135080

Let’s understand our dataset...

Feature 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:** 1 : Country name. Nominal, the name of the country where each customer resides.

Data Cleaning
and
Feature Engineering

Data Cleaning

- **Missing value treatment:** As explored, our dataset contains almost 25% missing data. We drop these values as they do not contain a valuable information, i.e, CustomerID.

- **Duplicate value treatment:** We also drop these values.

- **Cleaning some more:** The dataset contains record of cancelled order. For the scope of this product we only consider the ordered placed, and drop the cancelled transaction observations.

Feature Engineering

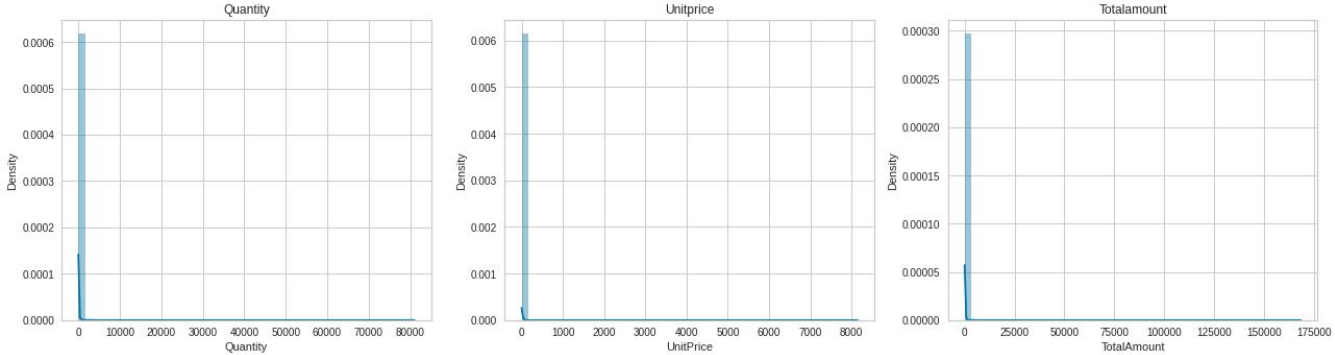
- **Date-Time features:** We construct new features from the InvoiceDate column. We get ‘Year’, ‘Month’, ‘MonthNum’, ‘Day’, ‘DayNum’, ‘Hour’, ‘Minute’

- **TotalAmount:** We construct a new feature, ‘TotalAmount’ from the ‘Quantity’ and ‘UnitPrice’ columns.

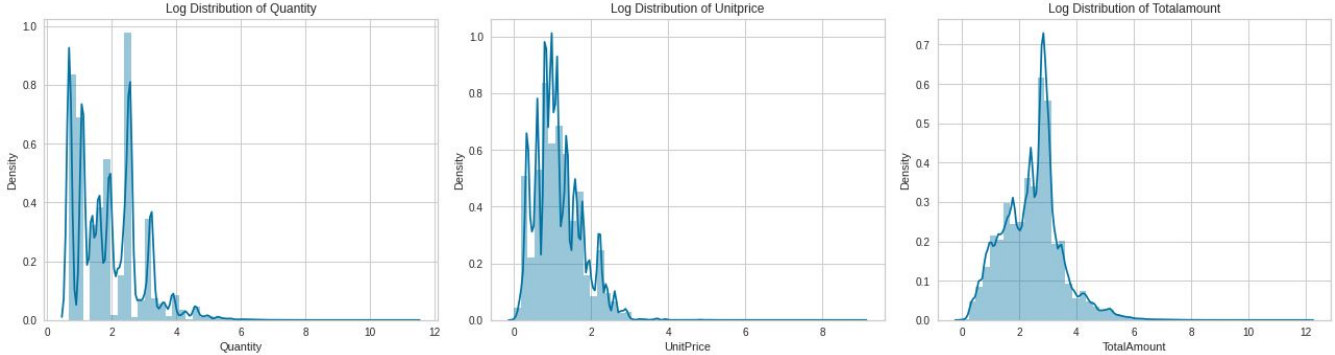
- **DayPart:** 'DayPart' gives us the part of the day(Morning, Afternoon, Evening) based on the hour of that day. We extracted this feature from the newly engineered ‘Hour’ column.

Exploratory Data Analysis

Distribution of Numeric Features

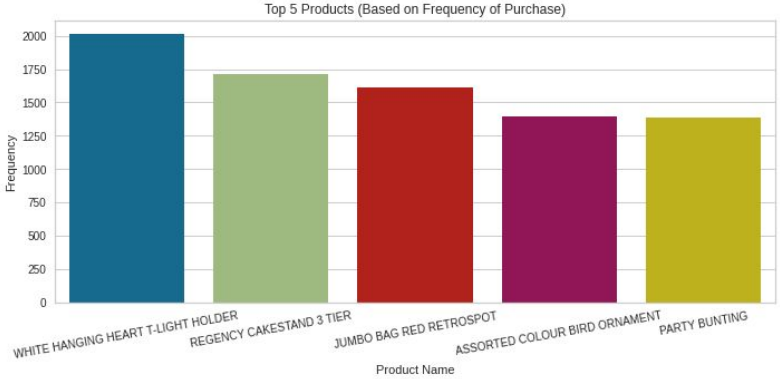


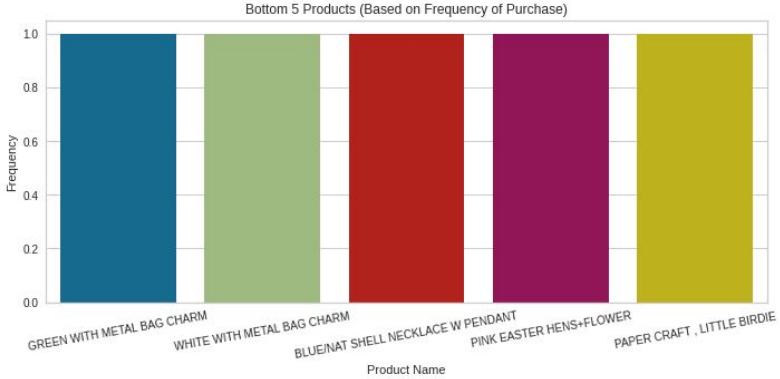
- We can see that the features are all positively skewed.(mean > median > mode)
- Ideally we these features must be symmetric.(mean = median = mode)
- For this we apply Log Transformation



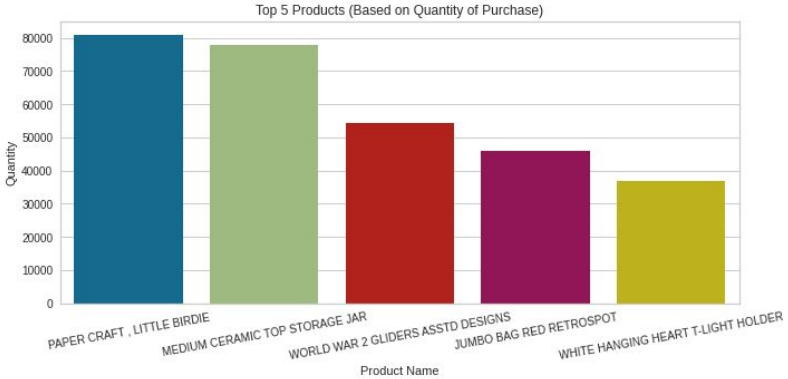
Product Wise Analysis

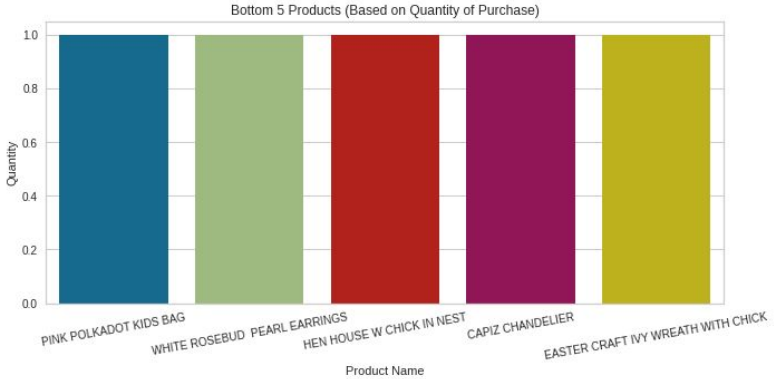
- ## Top and Bottom Products (Based on Frequency of Purchase)





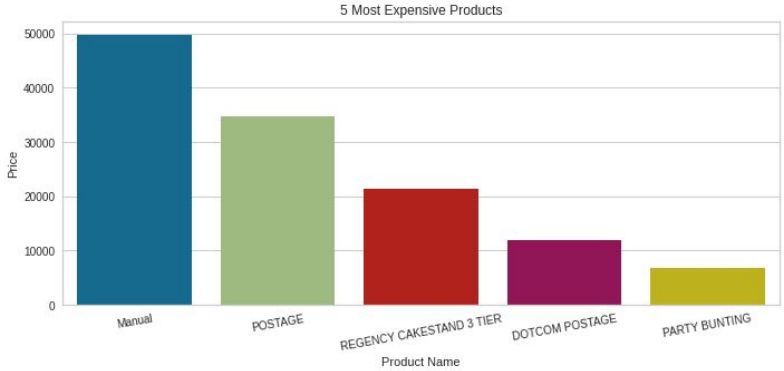
- ## Top and Bottom Products (Based on Quantity of Purchase)

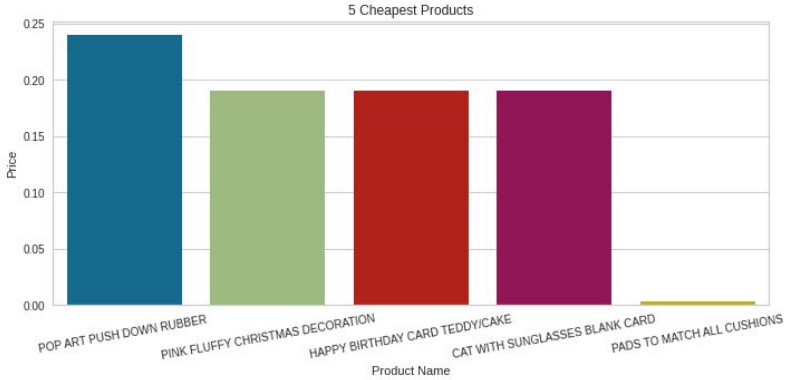




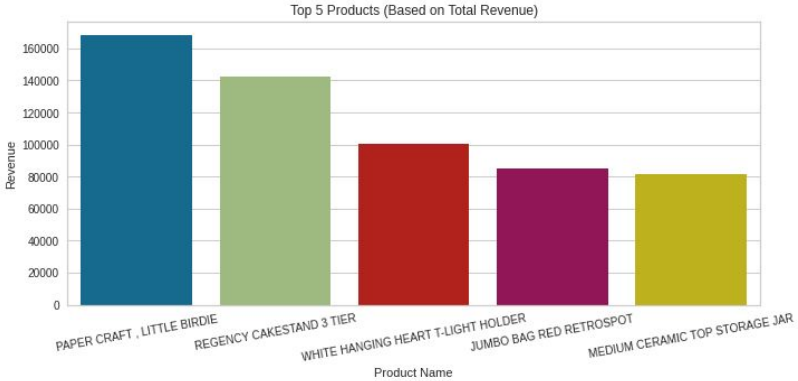
Product Wise Analysis

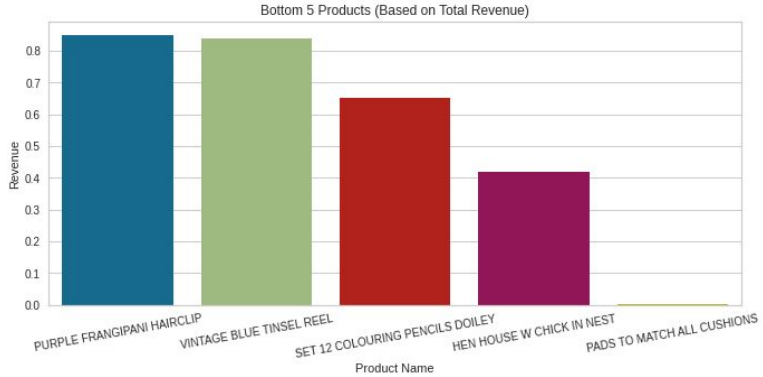
- **Most Expensive and Cheapest Products per Unit**



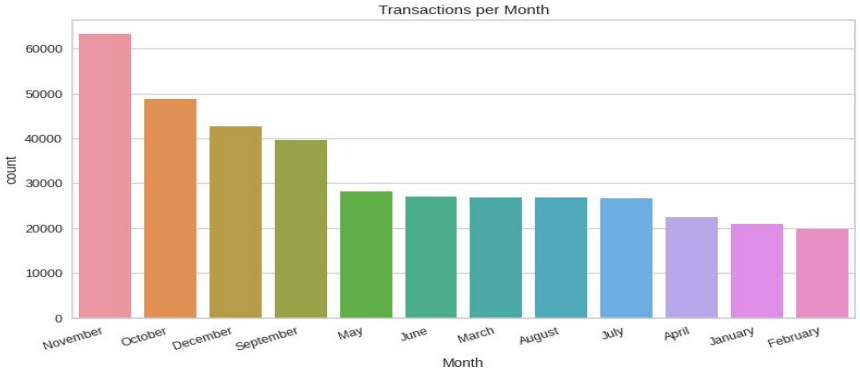


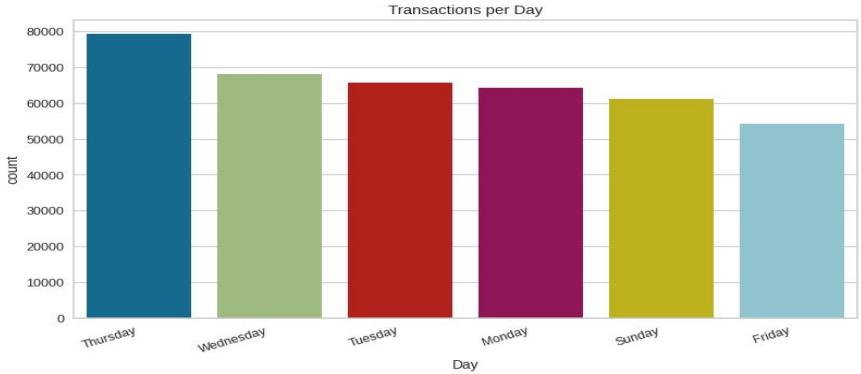
- **Top and Bottom Products (Based on Total Revenue)**

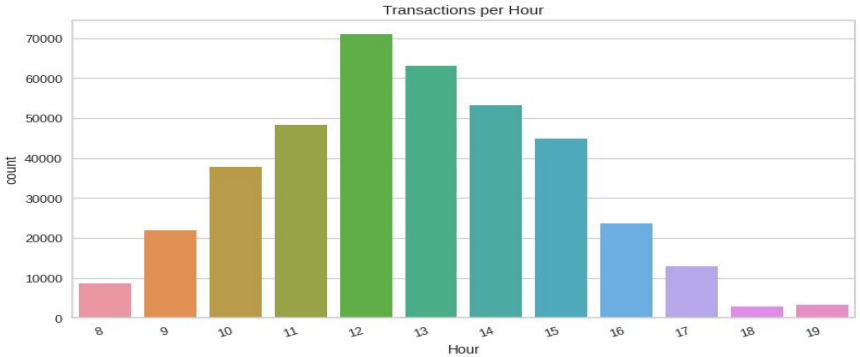


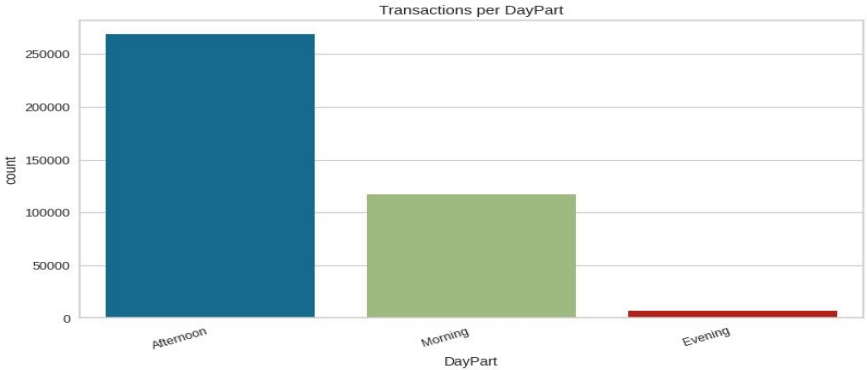


Transactions w.r.t Date and Time





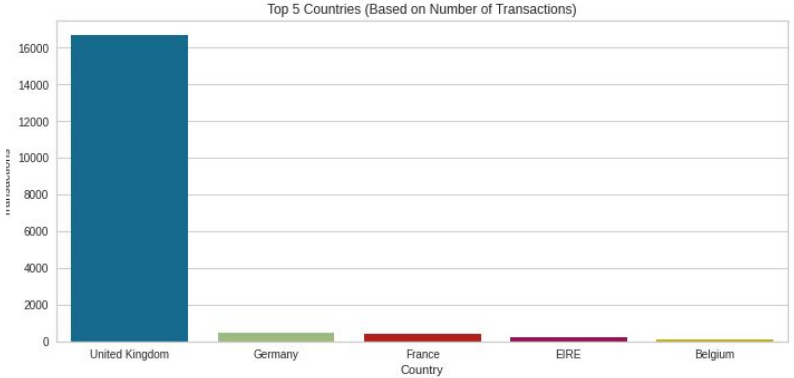


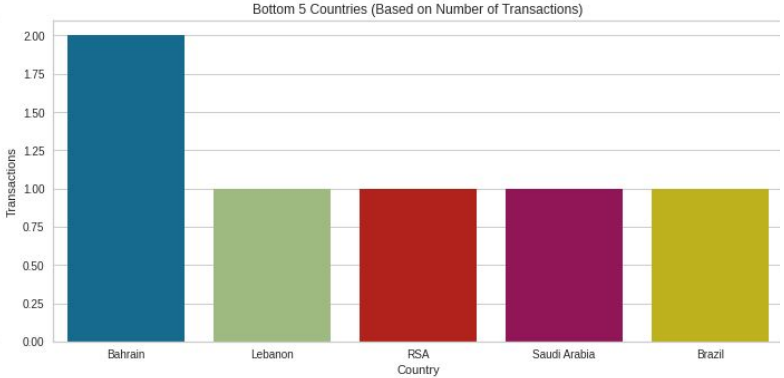


- November is the busiest month, followed by October and December.
- Thursday is the busiest day of the week.
- 12 pm is the peak hour of transactions.
- Most orders are placed in the afternoon.

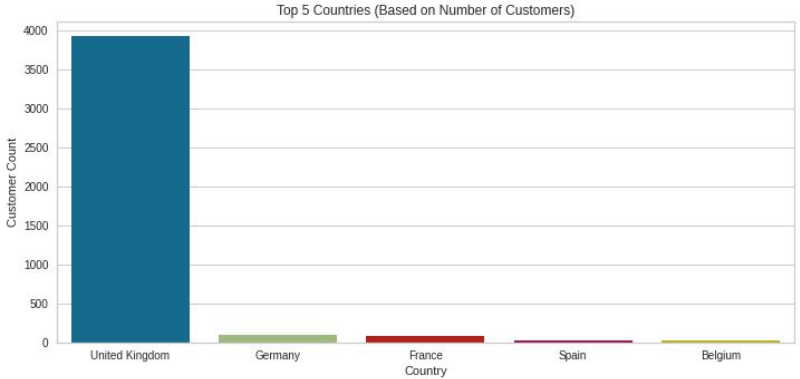
Location Based Analysis

- **Most Expensive and Cheapest Products per Unit**



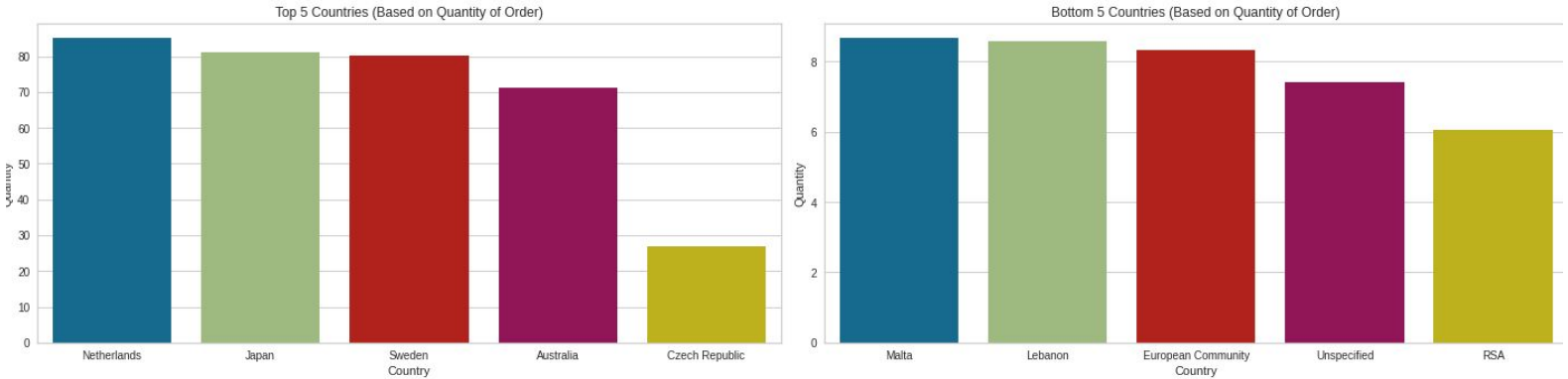


- **Top and Bottom Countries (Based on Number of Customers)**



Location Based Analysis

- **Top and Bottom Countries (Based on Quantities Ordered)**



RFM Analysis

RFM Analysis

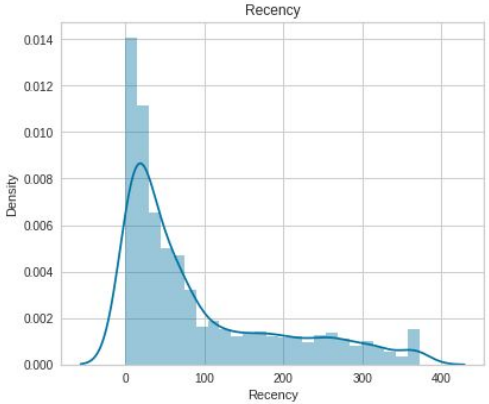
- **Recency**, **Frequency**, **Monetary** value is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors:
  - **Recency**: How recently a customer has made a purchase
  - **Frequency**: How often a customer makes a purchase.
  - **Monetary Value**: How much money a customer spends on

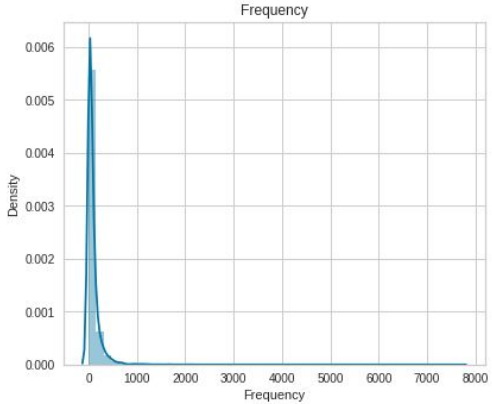
For this project we will calculate the Recency, Frequency, Monetary values as follows:
Recency = Latest Date - Last Invoice Data
Frequency = Count of invoice no. of transaction(s)
Monetary = Sum of total amount for each customer

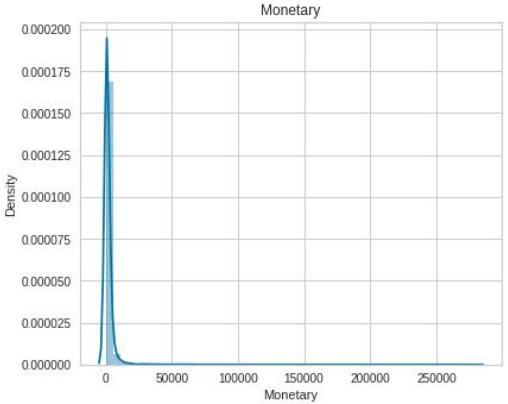
	CustomerID	Recency	Frequency	Monetary	R	F	M	RFM_Group	RFM_Score	Recency_Log	Frequency_Log	Monetary_Log
0	12346.0	325	1	77183.60	1	1	4	114	6	5.783825	0.000000	11.253942
1	12347.0	2	182	4310.00	4	4	4	444	12	0.693147	5.204007	8.368693
2	12348.0	75	31	1797.24	2	2	4	224	8	4.317488	3.433987	7.494007

RFM Analysis

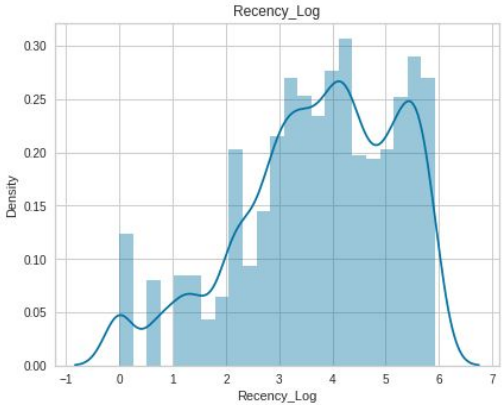
Before Log Transformation

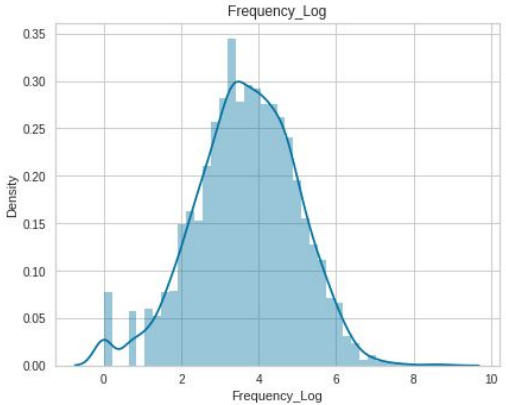


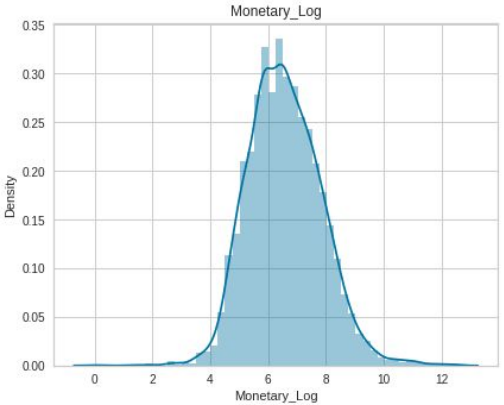




After Log Transformation







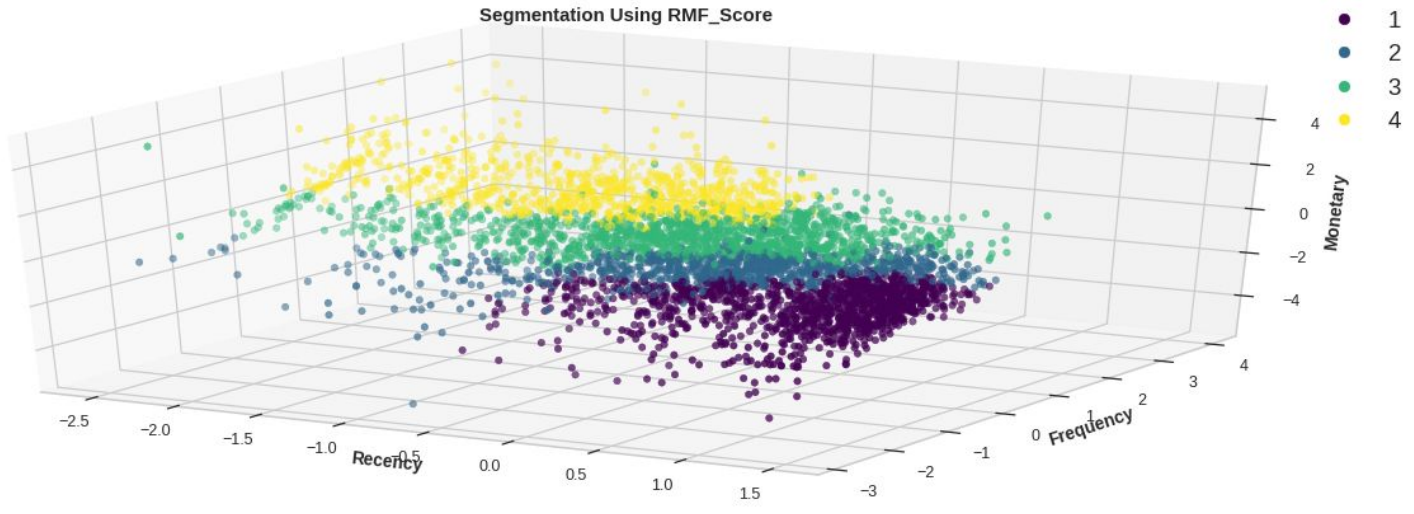
MODEL BUILDING AND EVALUATION

Prerequisites

- **Feature Matrix:**
  - Our feature matrix for model building to come up with clustering was constructed after RFM analysis on the original dataset.
  - The columns are 'Recency_Log', 'Frequency_Log', 'Monetary_Log'

- **Feature Scaling:**
  - We scale our features using StandardScaler.
  - StandardScaler is used to resize the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1.

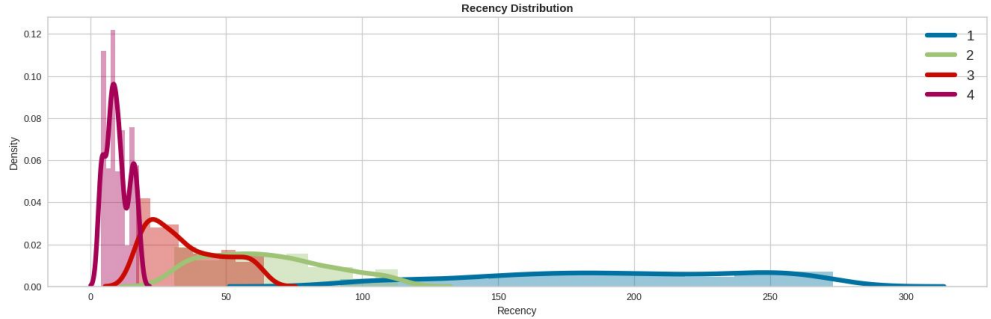
Segmentation Using RMF_Score

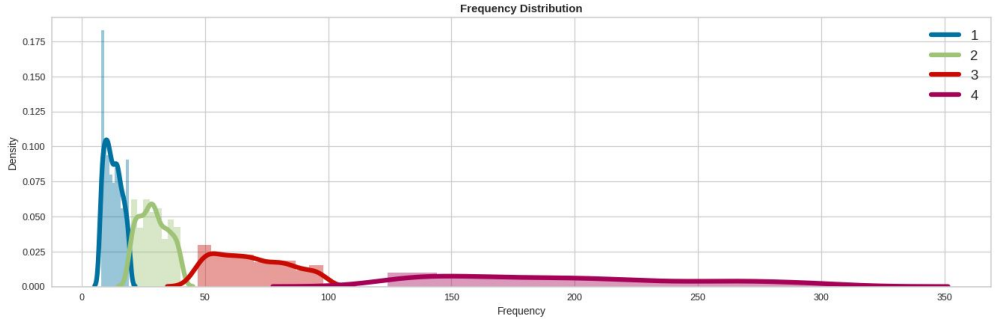


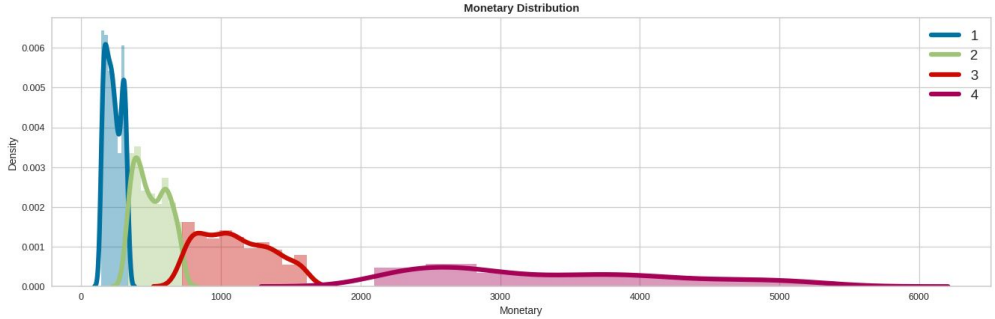
RMF	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
1	191.181255	196.000000	15.040279	12.000000	266.465648	225.850000	1291
2	87.356209	63.000000	33.026144	29.000000	790.243182	491.040000	918
3	47.344641	31.000000	81.417887	67.000000	1601.348359	1079.610000	1297
4	13.192077	9.000000	284.991597	191.000000	6891.729508	3170.980000	833

Segmentation Using RMF_Score

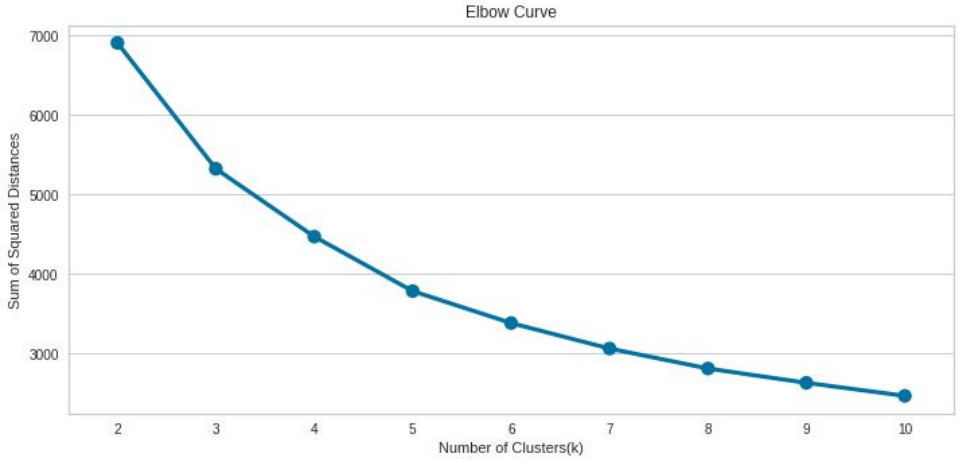
RMF	Last_Transaction	Transaction_frequency	Transaction_Value
1	91 to 273 days ago	7 to 20 times	142 to 335 Sterling
2	30 to 113 days ago	19 to 41 times	328 to 725 Sterling
3	16 to 65 days ago	46 to 99 times	720 to 1618 Sterling
4	3 to 19 days ago	123 to 306 times	2093 to 5411 Sterling

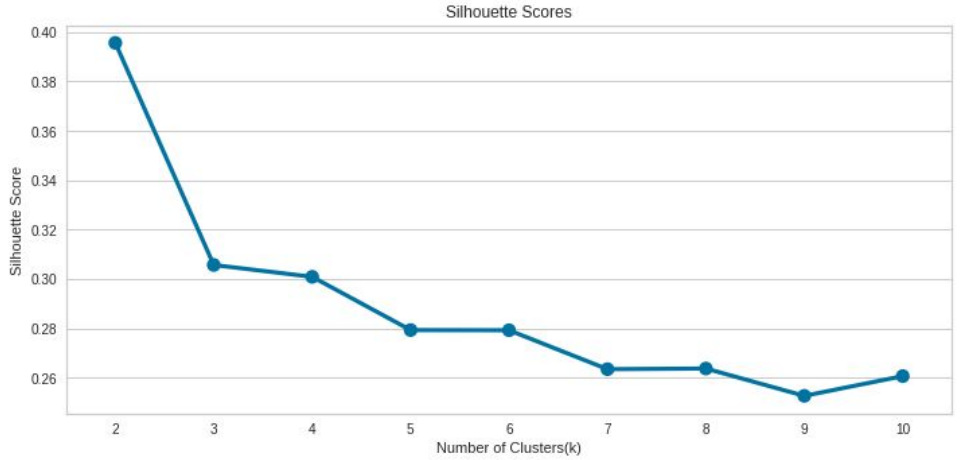






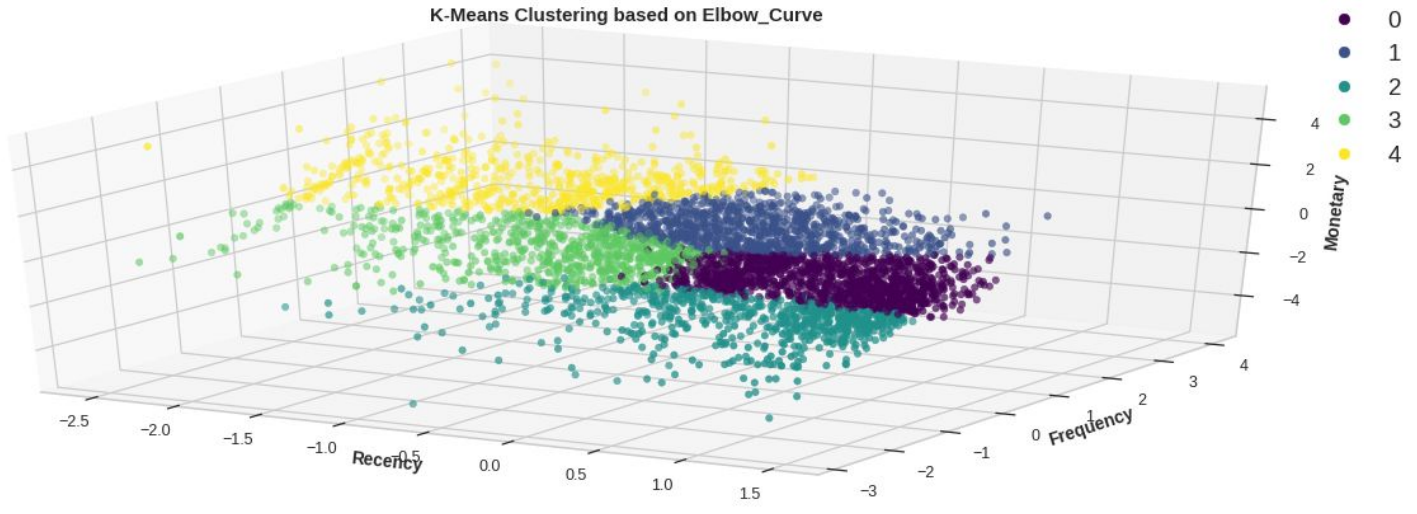
K-Means Clustering





- We have considered the elbow curve and to start with 2 as the minimum number of clusters and calculated their respective silhouette scores.
- 5 appears to be the elbow, looking at the elbow curve.
- Silhouette score for 2 clusters is the highest
- Considering the tradeoff between the two plots we can also consider 4 to be the optimal number of clusters.

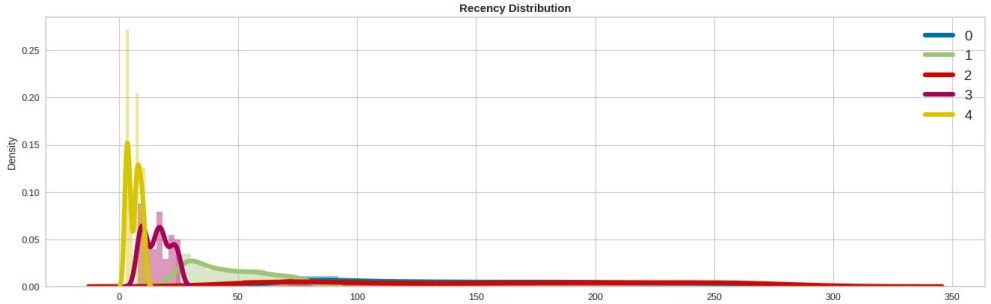
K-Means clustering based on Elbow Curve

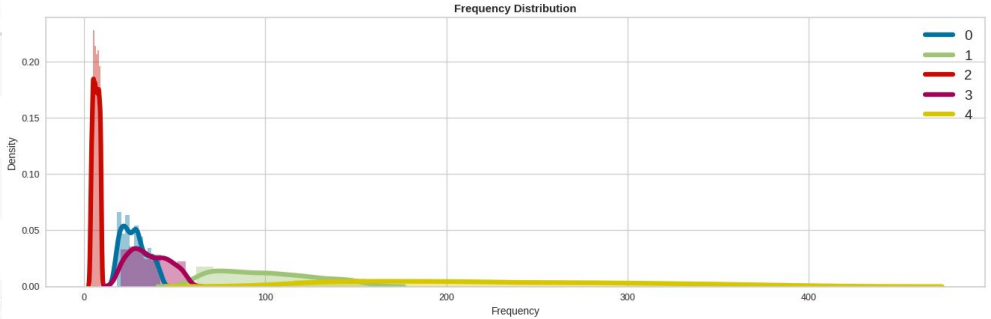


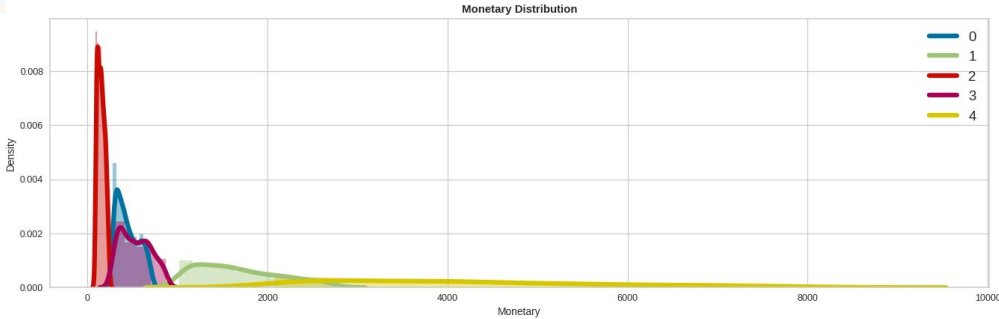
K-Means with n=5	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	164.809135	147.000000	31.154976	27.000000	527.786682	419.085000	1226
1	59.233654	43.000000	114.920192	98.000000	2122.224674	1582.700000	1040
2	168.301565	165.000000	7.092461	7.000000	198.154780	152.650000	703
3	16.706507	16.000000	40.397078	34.000000	632.357252	523.110000	753
4	7.857374	5.000000	323.523501	217.000000	8780.489368	3902.280000	617

K-Means clustering based on Elbow Curve

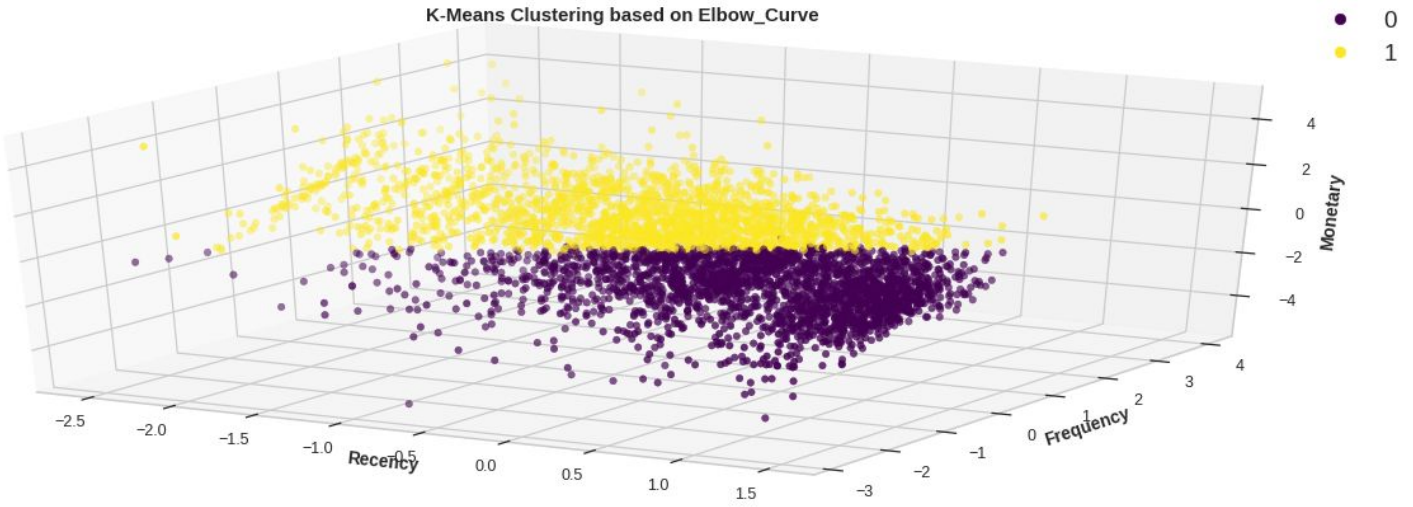
K-Means with n=5	Last_Transaction	Transaction_frequency	Transaction_Value
0	75 to 242 days ago	18 to 39 times	305 to 643 Sterling
1	26 to 72 days ago	67 to 142 times	1100 to 2425 Sterling
2	60 to 264 days ago	4 to 10 times	104 to 216 Sterling
3	8 to 24 days ago	20 to 53 times	324 to 789 Sterling
4	2 to 10 days ago	129 to 349 times	2330 to 6761 Sterling







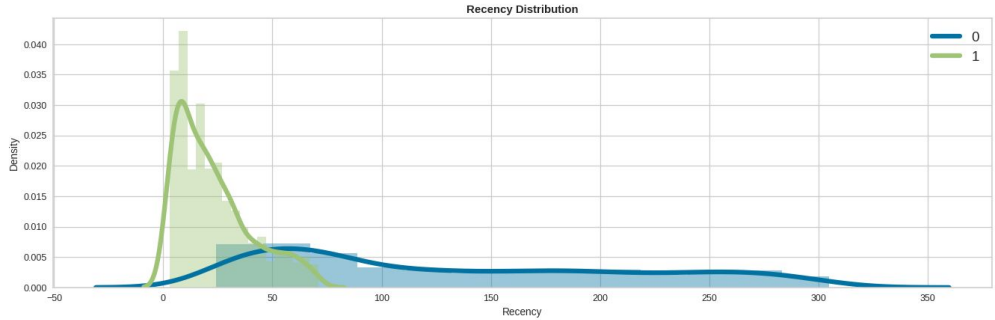
K-Means clustering based on Silhouette Score

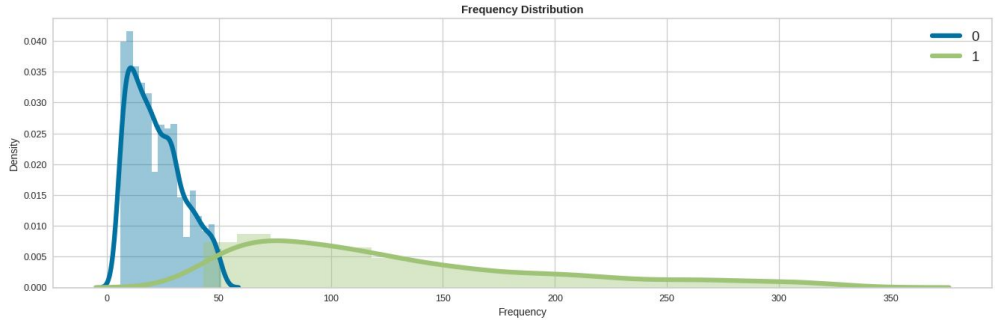


K-Means with n=2	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	140.736777	109.000000	24.831818	20.000000	469.390100	331.210000	2420
1	30.651381	18.000000	173.339760	109.000000	4039.231814	1827.800000	1919

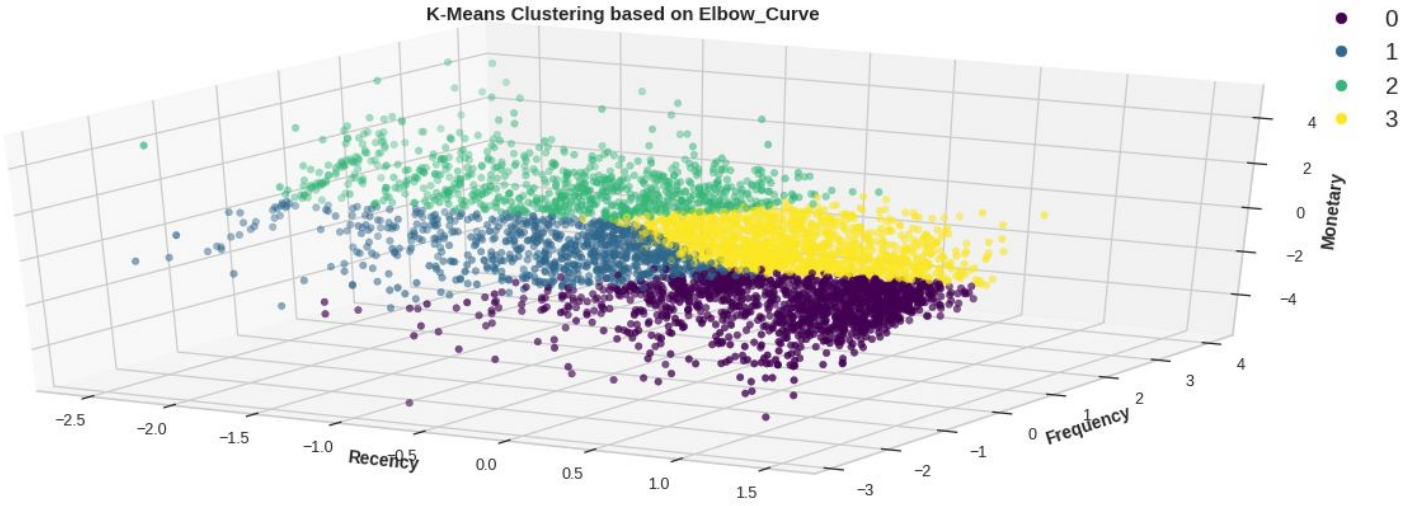
K-Means clustering based on Silhouette Score

K-Means with n=2	Last_Transaction	Transaction_frequency	Transaction_Value
0	50 to 226 days ago	10 to 33 times	189 to 570 Sterling
1	7 to 38 days ago	67 to 192 times	1070 to 3369 Sterling





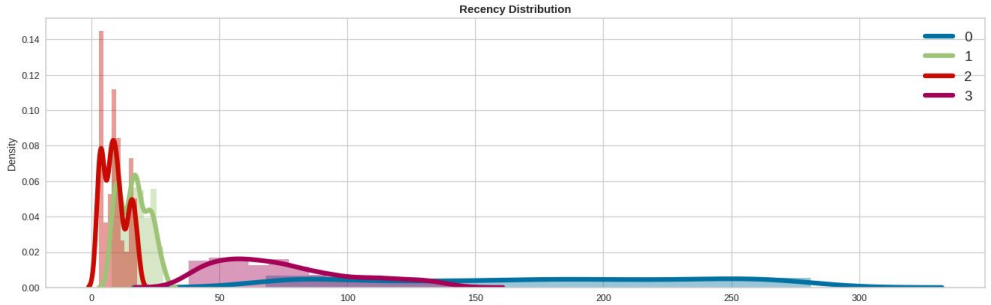
K-Means clustering based on Tradeoff between Elbow Curve and Silhouette Score

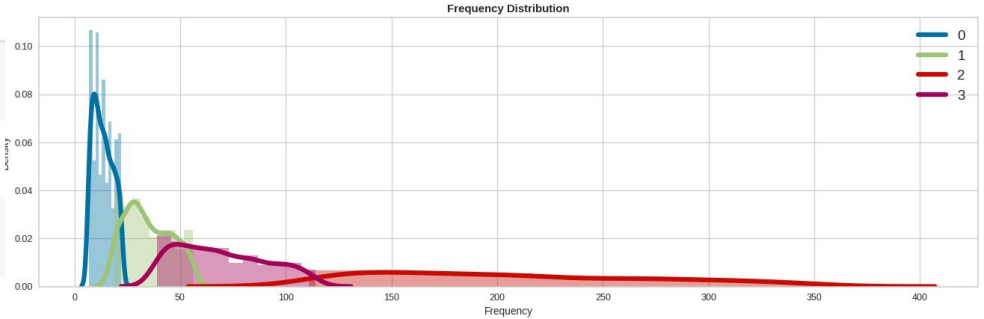


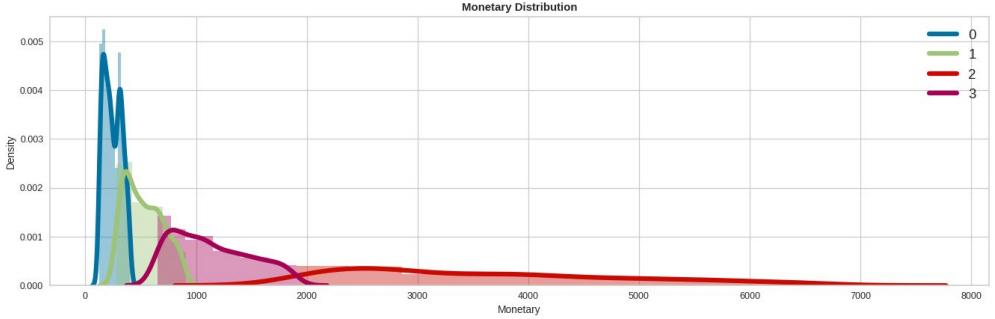
K-Means with n=4	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	180.447612	179.000000	14.779045	12.000000	293.077435	237.360000	1403
1	18.360775	17.000000	39.773608	33.000000	629.136707	503.280000	826
2	11.666667	8.500000	286.562044	195.000000	7299.176837	3408.130000	822
3	94.315994	71.000000	80.427019	67.000000	1518.968923	1079.180000	1288

K-Means clustering based on Tradeoff between Elbow Curve and Silhouette Score

K-Means with n=4	Last_Transaction	Transaction_frequency	Transaction_Value
0	78 to 266 days ago	7 to 21 times	144 to 363 Sterling
1	9 to 25 days ago	20 to 53 times	316 to 800 Sterling
2	3 to 17 days ago	123 to 311 times	2085 to 5655 Sterling
3	44 to 120 days ago	42 to 103 times	701 to 1711 Sterling

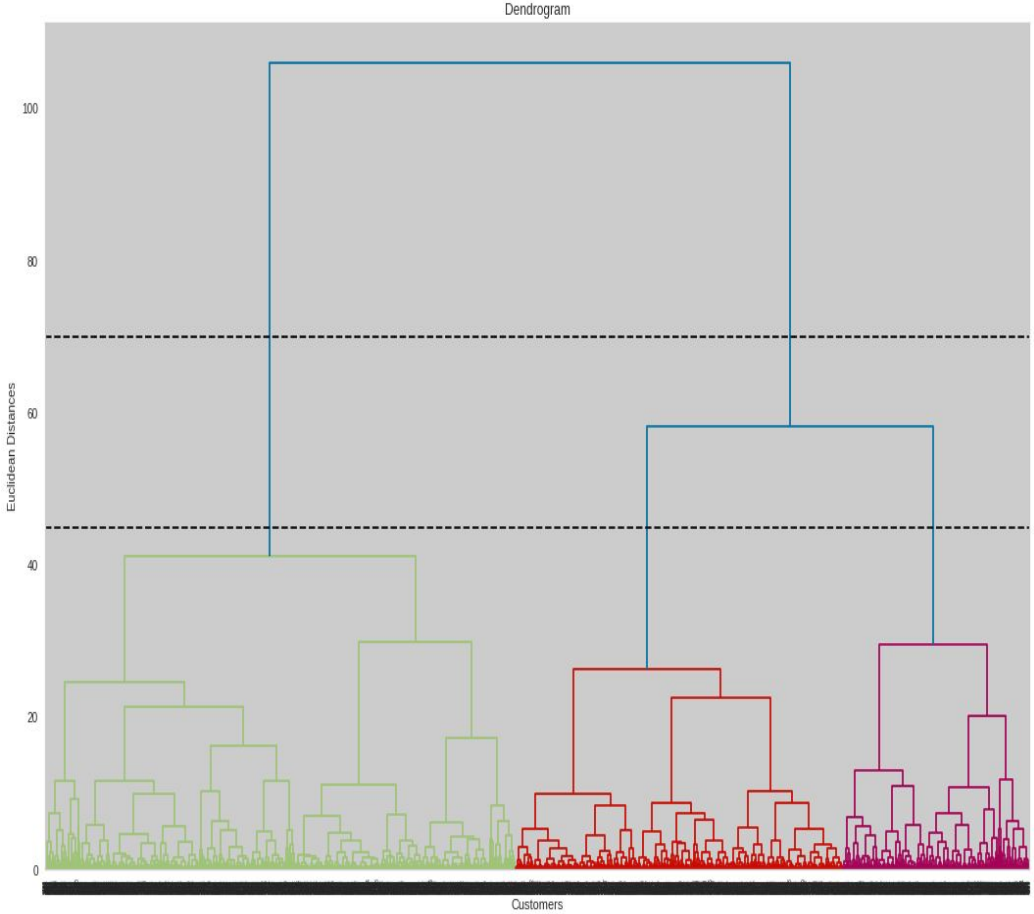




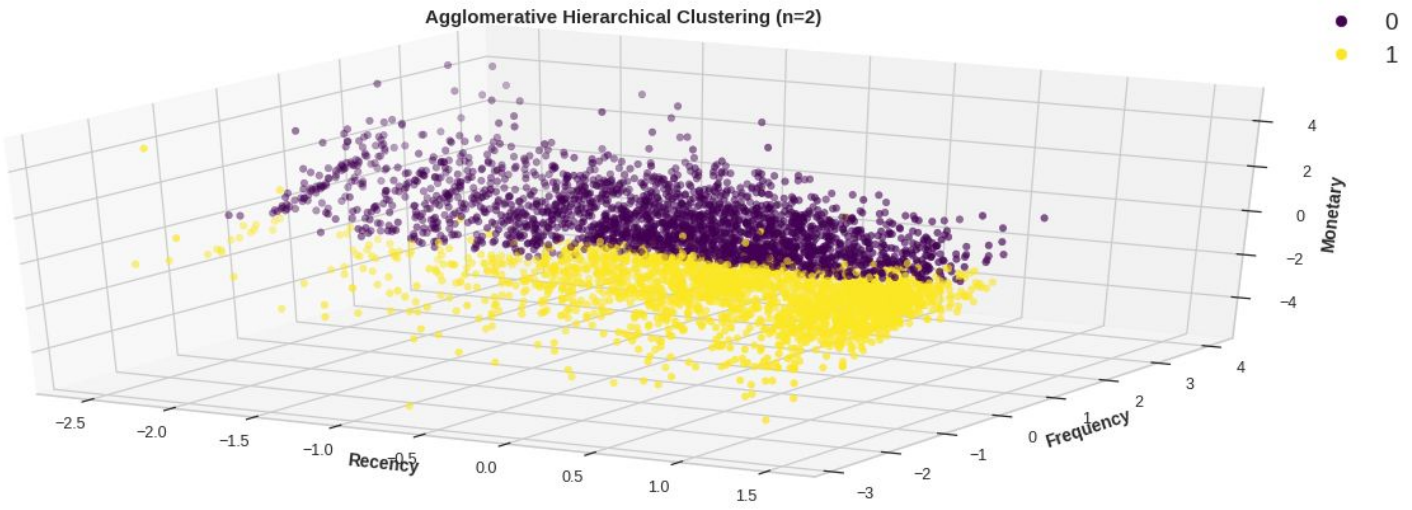


Hierarchical Clustering

- We try to set the threshold in such a way that it cuts the tallest vertical line.
- Here we will try two thresholds to cut the two visibly tallest lines to get the optimal number of clusters. We chose these as y=70 and y=45.
- y=70 that gives us 2 clusters.
- y=45 that gives us 3 clusters.



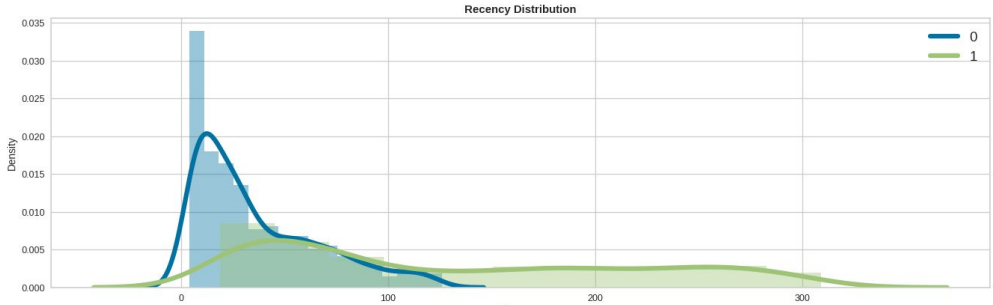
Agglomerative Hierarchical Clustering for Dendrogram threshold y=70

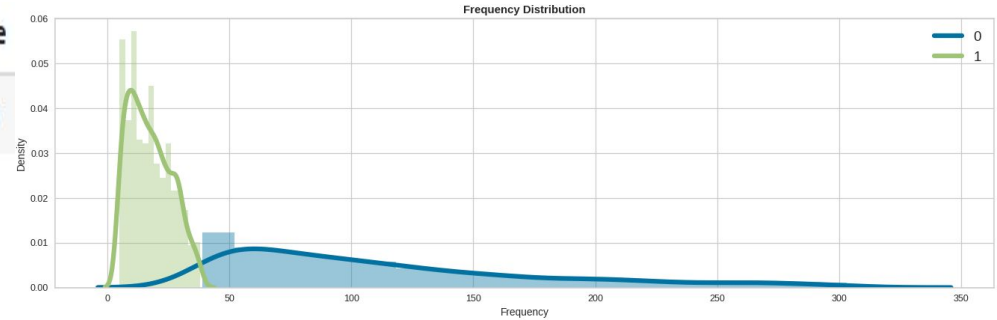


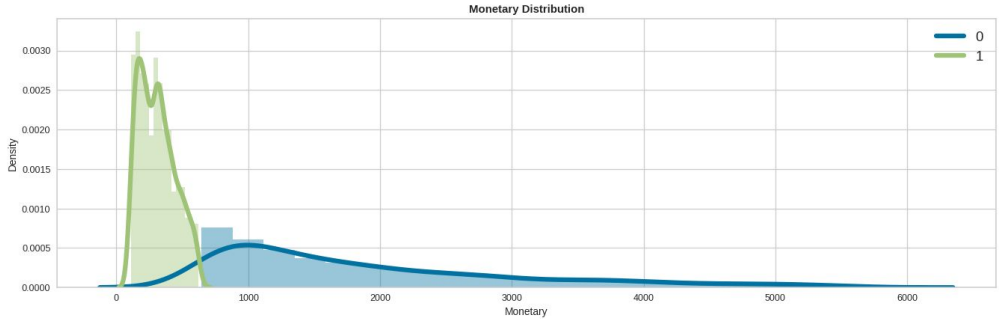
AHC with n=2	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	48.522296	25.000000	155.105960	95.000000	3434.244540	1531.620000	2265
1	139.585342	102.000000	19.969624	16.000000	534.544847	300.390000	2074

Agglomerative Hierarchical Clustering for Dendrogram threshold y=70

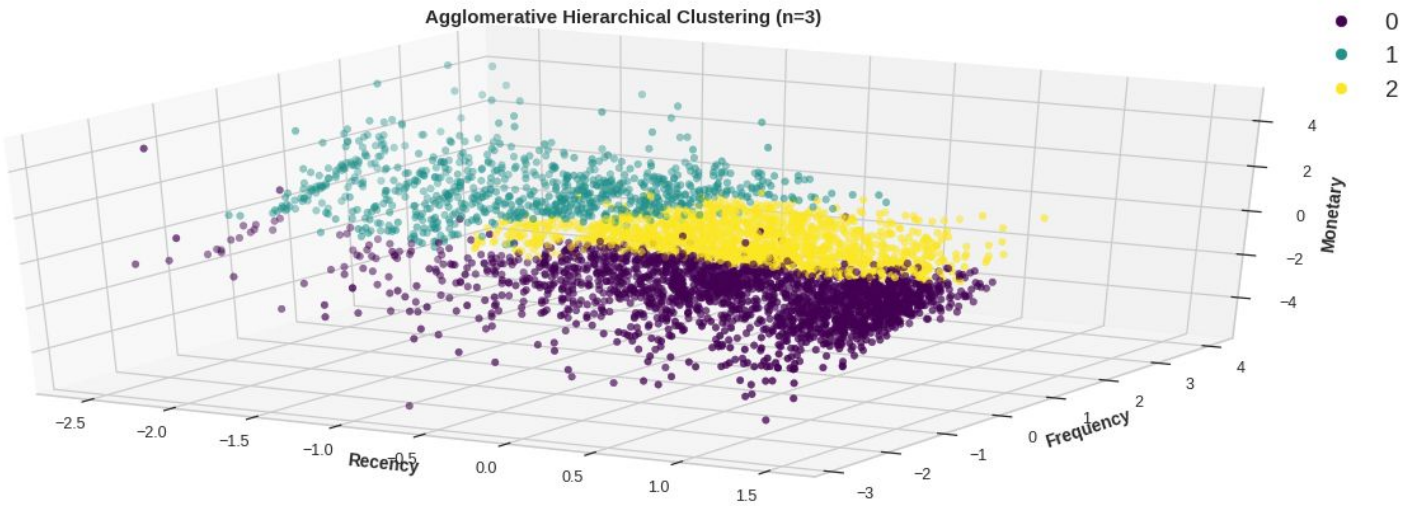
AHC with n=2	Last_Transaction	Transaction_frequency	Transaction_Value
0	9 to 65 days ago	54 to 166 times	891 to 2863 Sterling
1	39 to 237 days ago	9 to 27 times	171 to 439 Sterling







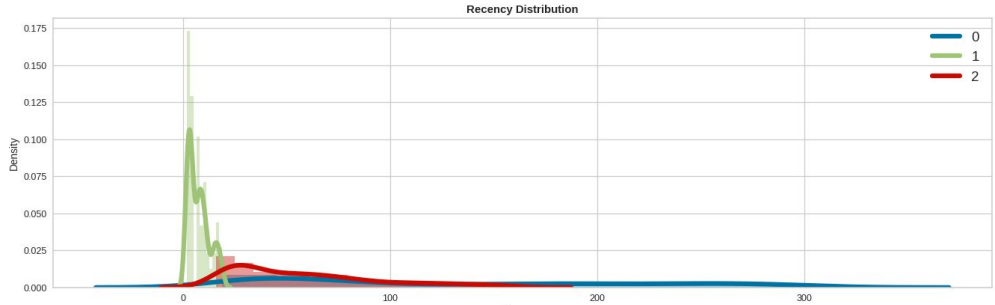
Agglomerative Hierarchical Clustering for Dendrogram threshold y=45

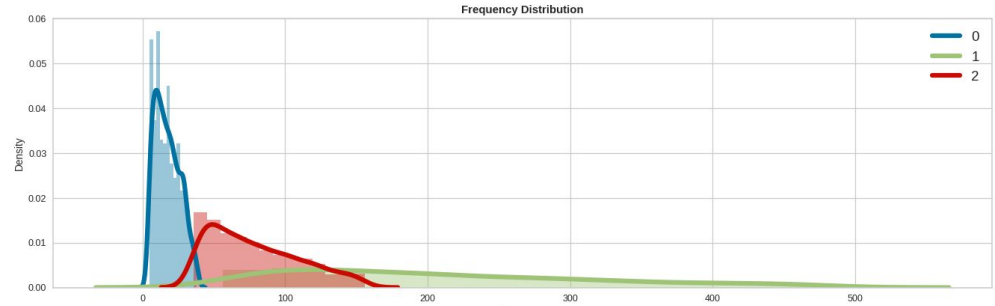


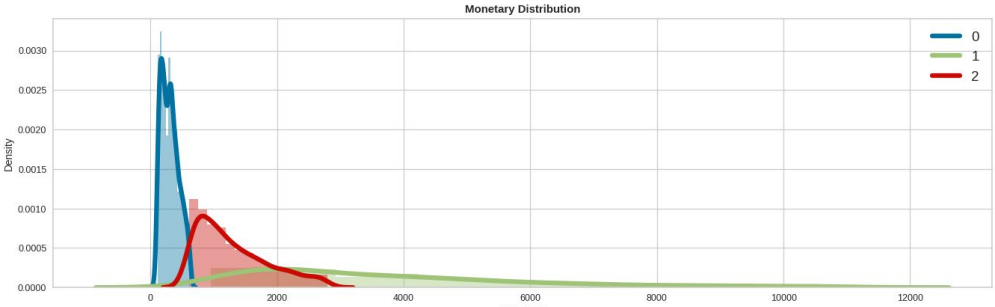
AHC with n=3	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
0	139.585342	102.000000	19.969624	16.000000	534.544847	300.390000	2074
1	9.063183	7.000000	274.784933	186.000000	6791.309465	3023.080000	823
2	71.042996	50.000000	86.800971	71.000000	1518.249786	1135.335000	1442

Agglomerative Hierarchical Clustering for Dendrogram threshold y=45

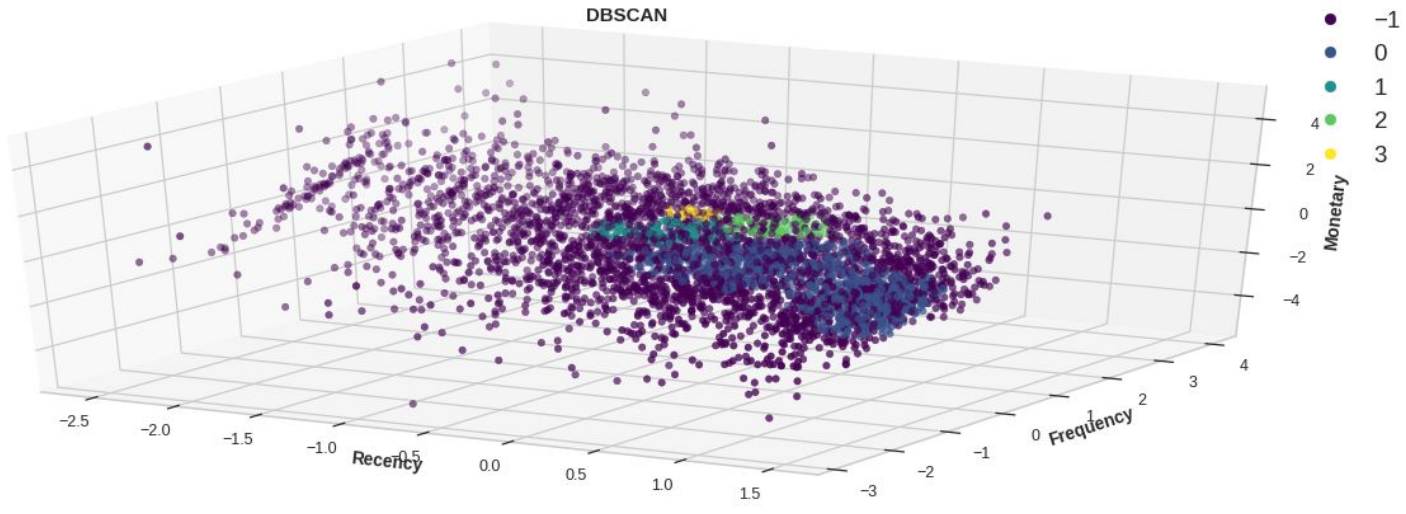
AHC with n=3	Last_Transaction	Transaction_frequency	Transaction_Value
0	39 to 237 days ago	9 to 27 times	171 to 439 Sterling
1	3 to 12 days ago	103 to 310 times	1711 to 5490 Sterling
2	25 to 91 days ago	47 to 110 times	776 to 1823 Sterling







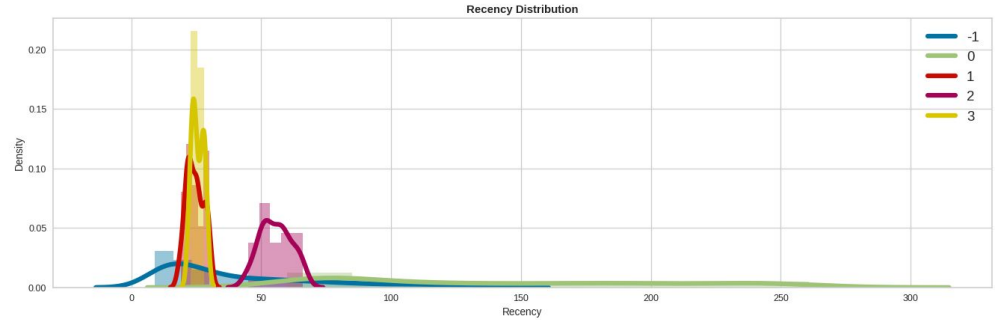
DBSCAN

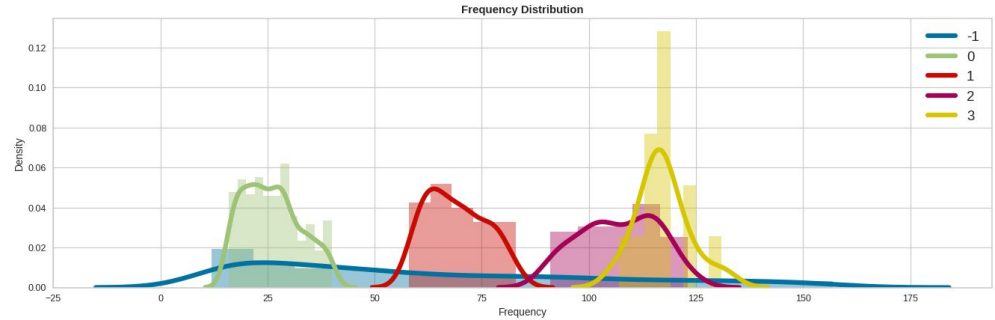


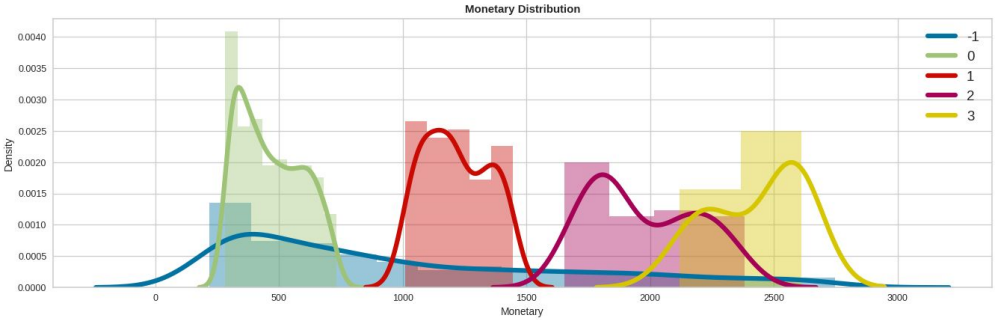
DBSCAN	Recency_mean	Recency_median	Frequency_mean	Frequency_median	Monetary_mean	Monetary_median	Count_
-1	75.556762	31.000000	111.785127	51.000000	2606.060747	778.000000	3039
0	154.735977	125.500000	28.562863	25.000000	516.779916	450.445000	1034
1	25.317241	25.000000	70.765517	68.000000	1234.432690	1198.230000	145
2	56.428571	56.000000	106.979592	105.500000	2024.019490	1887.995000	98
3	25.565217	25.000000	119.043478	117.000000	2421.486957	2527.920000	23

DBSCAN

DBSCAN	Last_Transaction	Transaction_frequency	Transaction_Value
-1	10 to 106 days ago	15 to 131 times	276 to 2198 Sterling
0	65 to 244 days ago	17 to 37 times	308 to 663 Sterling
1	21 to 30 days ago	60 to 79 times	1048 to 1415 Sterling
2	46 to 64 days ago	92 to 117 times	1703 to 2301 Sterling
3	23 to 28 days ago	109 to 127 times	2153 to 2610 Sterling







Model-Cluster Summary and Conclusion

Model-Cluster Summary

Model	Criterion	Clusters
Segmentation Using RMF_Score	RMF_Score	4
K-Means	Elbow Curve	5
K-Means	Silhouette Score	2
K-Means	Elbow Curve & Silhouette Score	4
Hierarchical	Threshold y = 70	2
Hierarchical	Threshold y = 45	3
DBSCAN	eps=0.2, min_samples=25	5

Conclusion

- We cleaned the data, performed feature engineering and EDA to get important business insights.
- We then performed RFM analysis of the transaction data. This helped us gain importance metrics to build models for customer segmentation.
- Model building included **Segmentation Using RFM Scores**(Heuristic Model) which gave us *4 major segments*.
- Further we used the Machine Learning Models giving us the results as follows:
  - **K-Means Clustering**
    - *Elbow Method - 5 Clusters*
    - *Silhouette Score - 2 Clusters*
    - *Elbow Curve & Silhouette Score - 4 Clusters*
  - **Agglomerative Hierarchical Clustering**
    - *Dendrogram Threshold=70 - 2 Clusters*
    - *Dendrogram Threshold=45 - 3 Clusters*
  - **DBSCAN** - *5 Clusters*
- Based on the companies goals, we can use these models and their obtained clusters to build business strategies, marketting campaigns, etc. to target, insentivise and attract customer base.

THANK YOU