Capstone Project -4

Online Retail Customer Segmentation

Discussion points

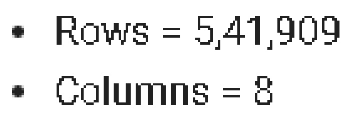
1. Problem statement
2. Data Summary
3. Exploratory Data Analysis
4. Feature Engineering
5. RFM analysis
6. Clustering Analysis
7. Challenges
8. Conclusion

Problem Statement

- **What is customer segmentation?**
- Customer segmentation is the practice of dividing a company’s customers into groups
- that reflect similarity among customers in each group. The goal of segmenting
- customers is to decide how to relate to customers in each segment to maximize the
- value of each customer to the business.

In this project, we have to identify major customer segments on a transnational dataset, 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.

Data Summary

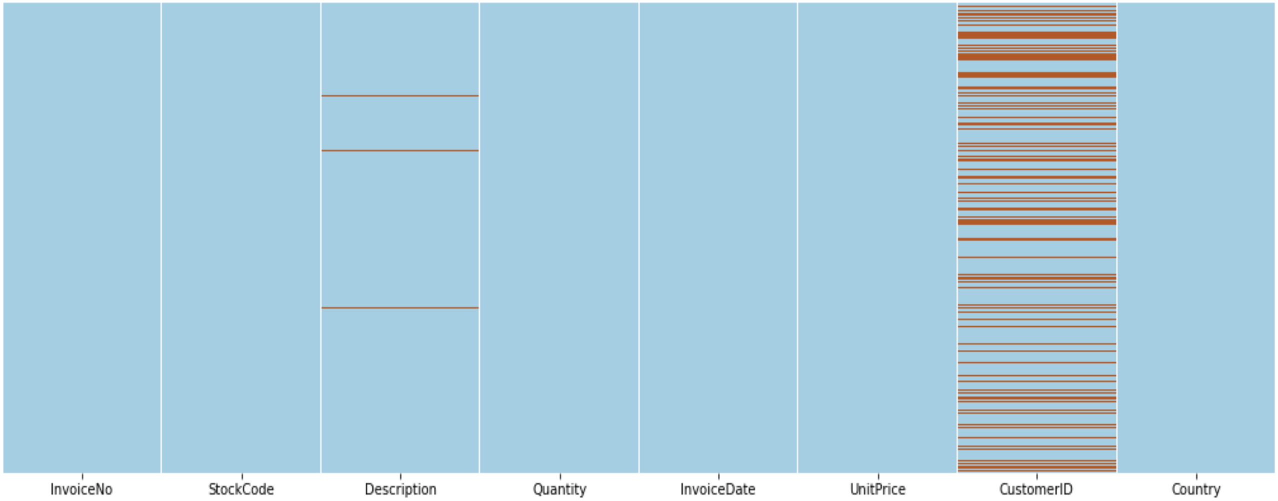


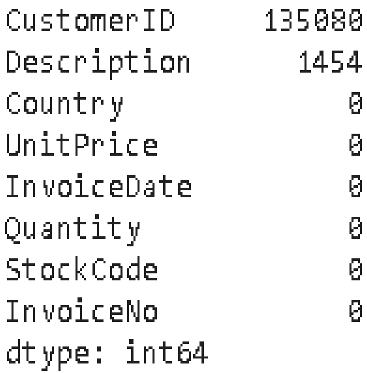
- Rows = 5,41,909
- Columns = 8

- **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- **Description**: Product (item) name. Nominal.
- **Quantity**: The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate**: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice**: Unit price. Numeric, Product price per unit in sterling.
- **CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- **Country**: Country name. Nominal, the name of the country where each customer resides.

Exploratory Data Analysis

Missing values

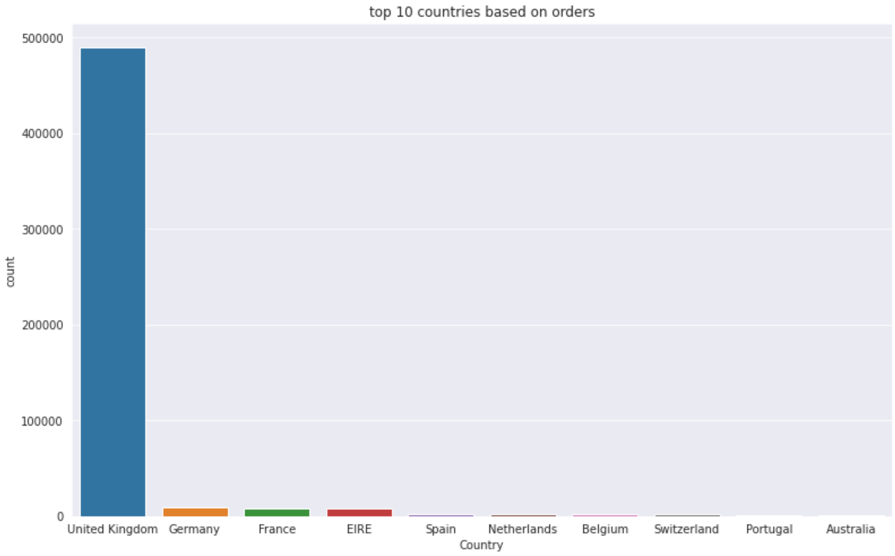


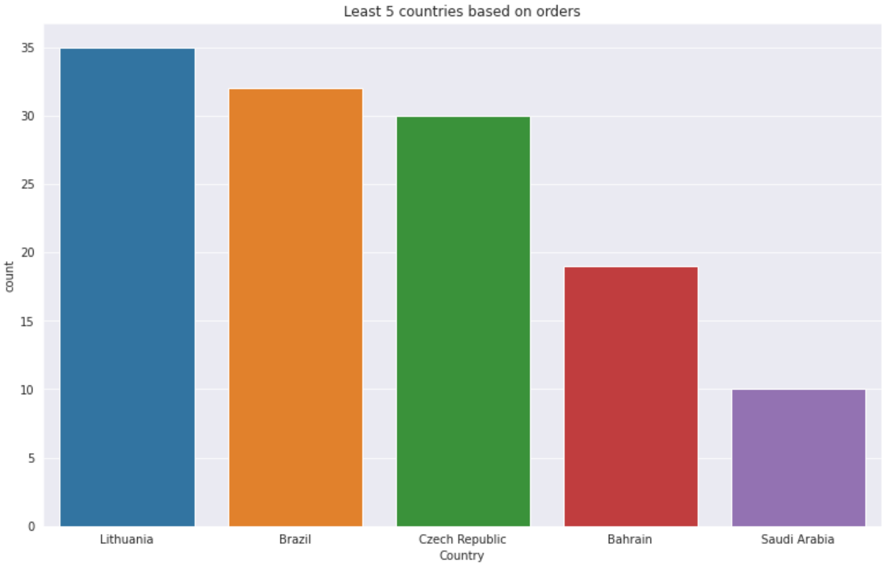


```
CustomerID     135080
Description      1454
Country             0
UnitPrice           0
InvoiceDate         0
Quantity            0
StockCode           0
InvoiceNo           0
dtype: int64
```

- CustomerID and Description have null values

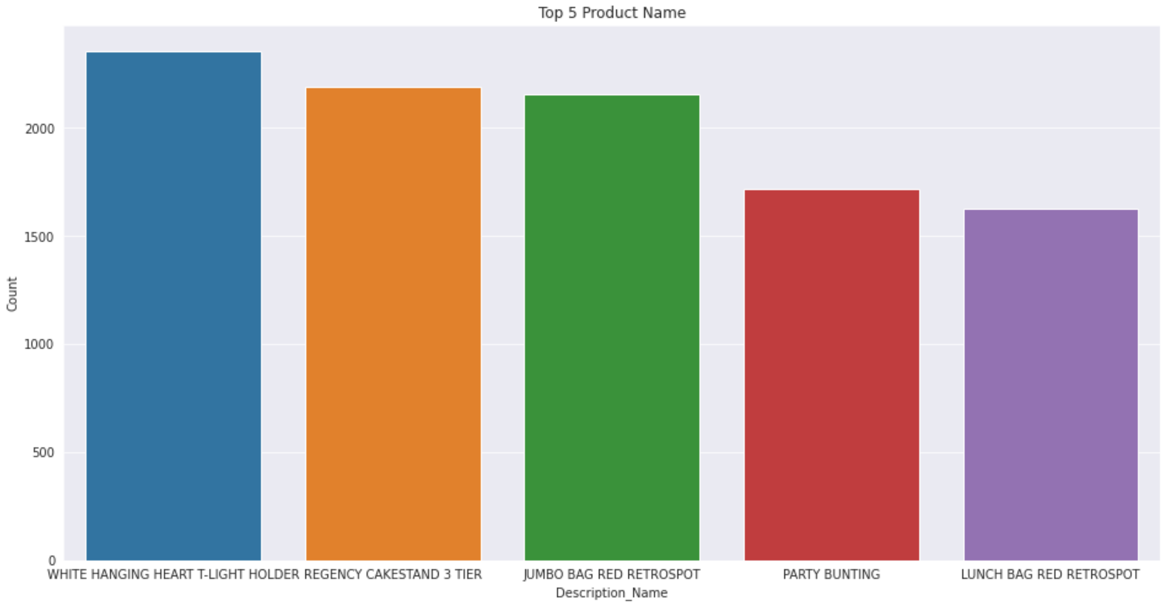
Top and Bottom countries based on orders


top 10 countries based on orders


Least 5 countries based on orders

- Maximum orders are received from United Kingdom
- Least orders are received from Saudi Arabia

Most popular Products

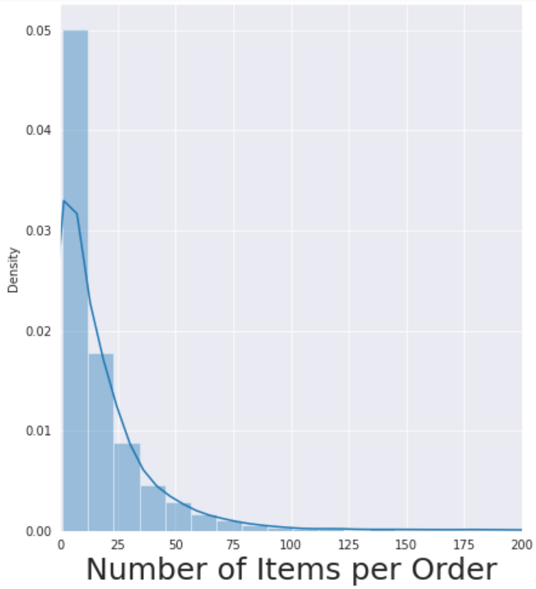


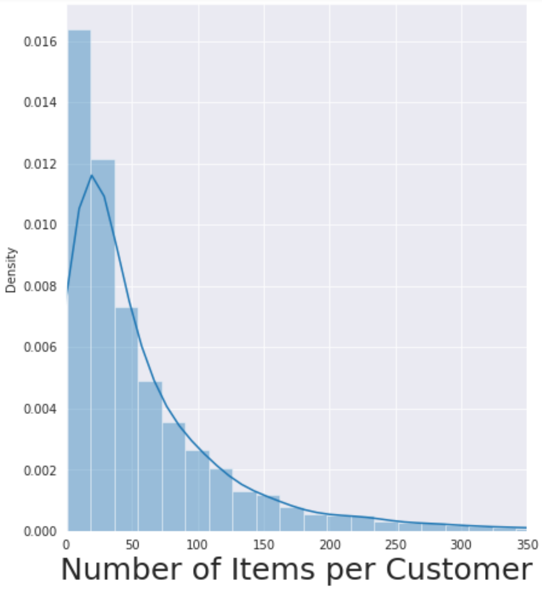
- Most popular item is White hanging Heart T-light

Number of items per Order and Customer

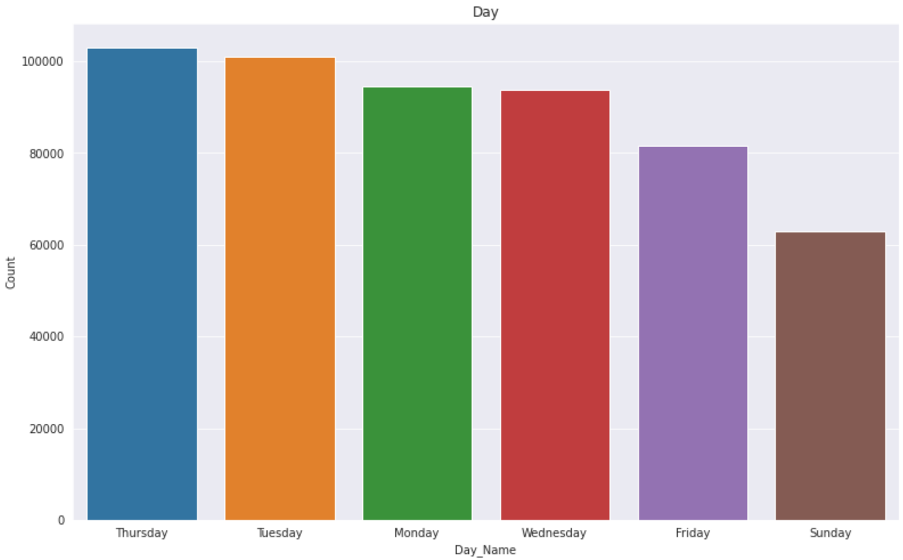
- The average number
- of items per order =20.5

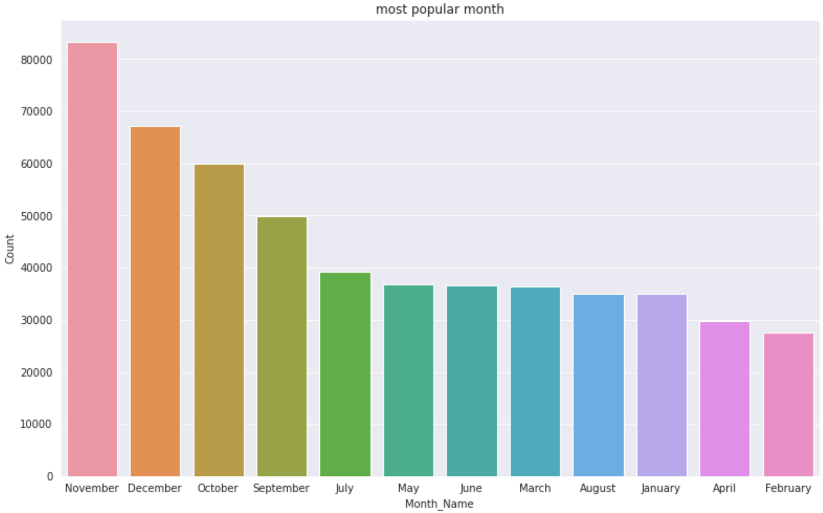
- The average number of
- Items per customer = 60





Daily and Monthly analysis

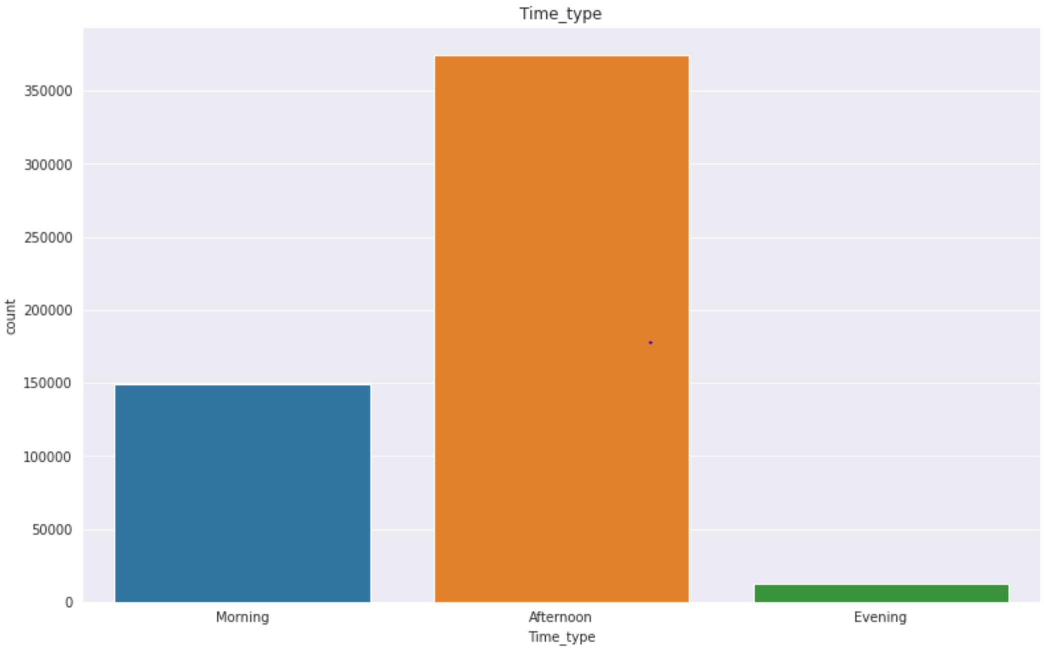




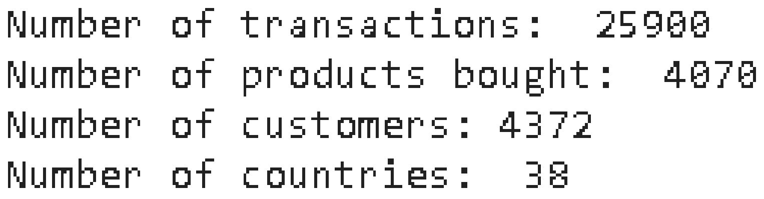
- Most orders are recieved on Thursday followed by Tuesday and Monday
- Most orders are recieved in the month of November followed by December

Hourly Analysis

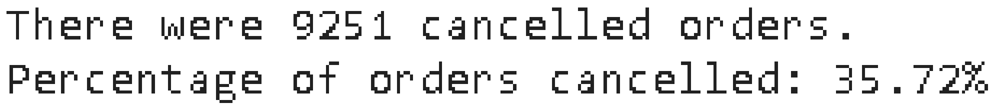
- Most customers buy at Afternoon



General info about orders



```
Number of transactions:  25900
Number of products bought:  4070
Number of customers: 4372
Number of countries:  38
```



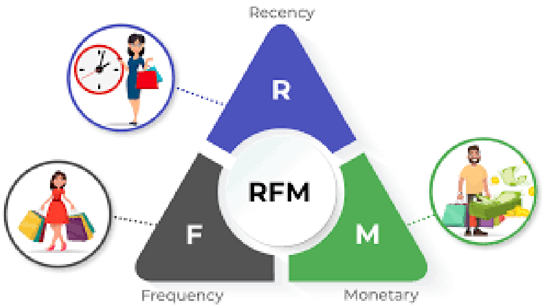
```
There were 9251 cancelled orders.
Percentage of orders cancelled: 35.72%
```

Feature Engineering

1. Null value treatment
   - Applied Most frequent category imputation on Description feature
   - Applied Random Sample Imputation on CustomerID

2. Outlier Treatment
   - Used IQR method to treat outliers in Quantity and UnitPrice features

RFM analysis



What is RFM analysis?

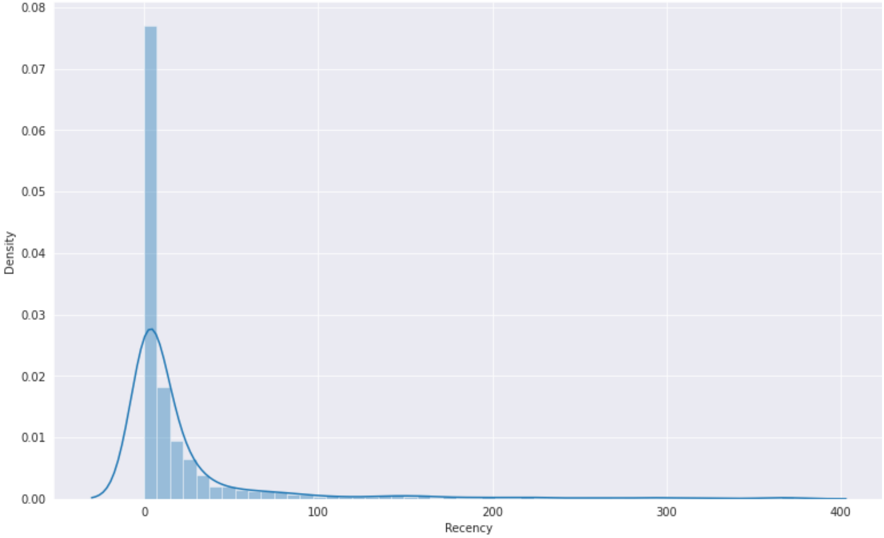
RFM Analysis is a marketing framework that is used to understand and analyze customer behaviour based on the above three factors RECENCY, Frequency, and Monetary.

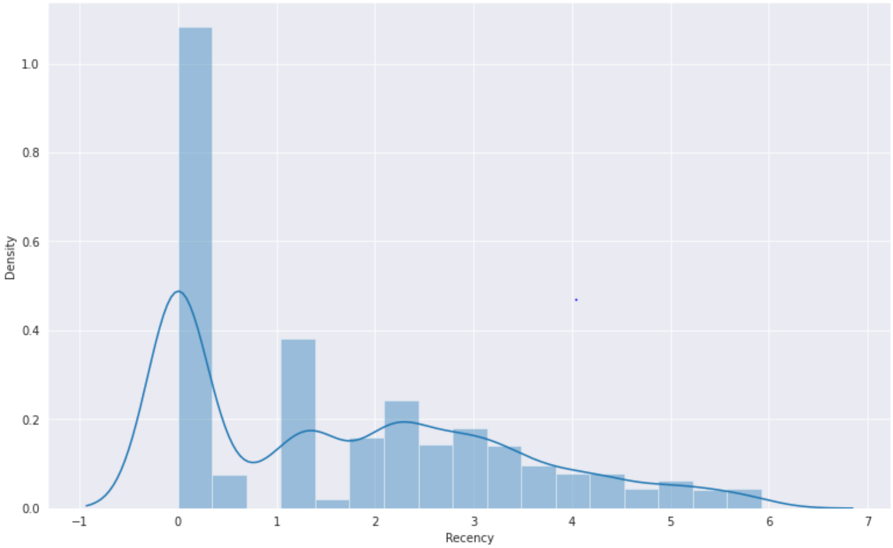
RECENCY: How recently did the customer visit our website or how recently did a customer purchase?
Frequency: How often do they visit or how often do they purchase?
Monetary: How much revenue we get from their visit or how much do they spend when they purchase?

Why is it needed?

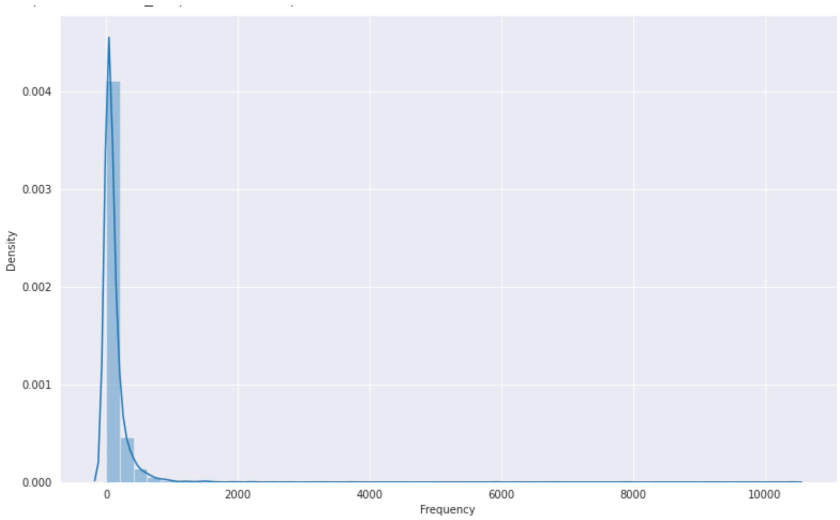
The RFM Analysis will help the businesses to segment their customer base into different homogenous groups so that they can engage with each group with different targeted marketing strategies.

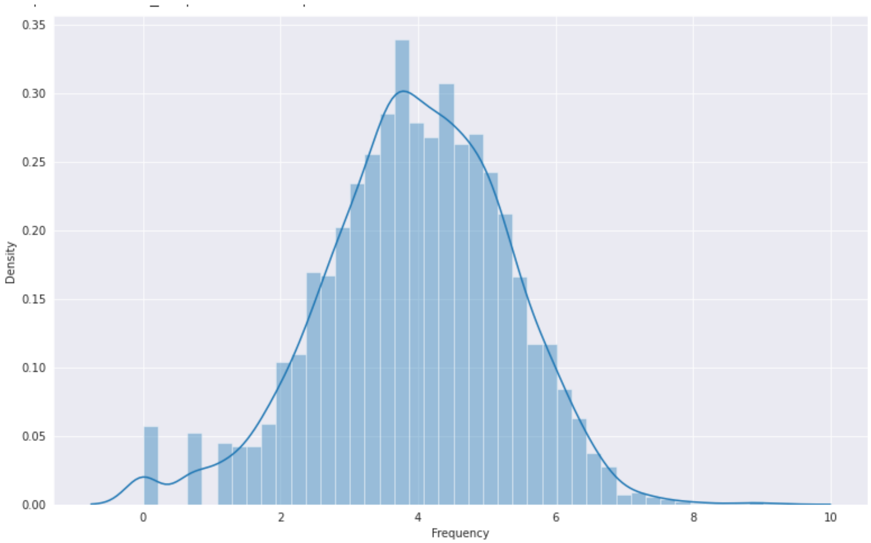
Recency


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
100
200
300
400
Recency

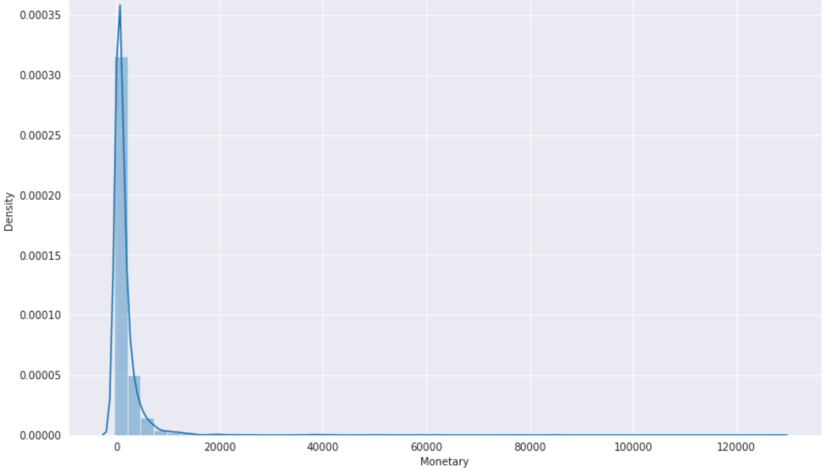

1.0
0.8
0.6
0.4
0.2
0.0
Density
−1
0
1
2
3
4
5
6
7
Recency

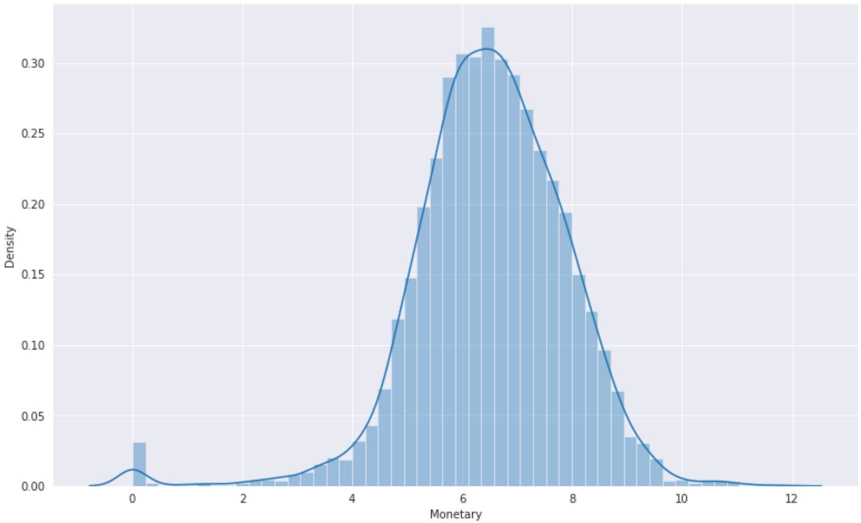
Frequency


0.004
0.003
0.002
0.001
0.000
Density
0
2000
4000
6000
8000
10000
Frequency

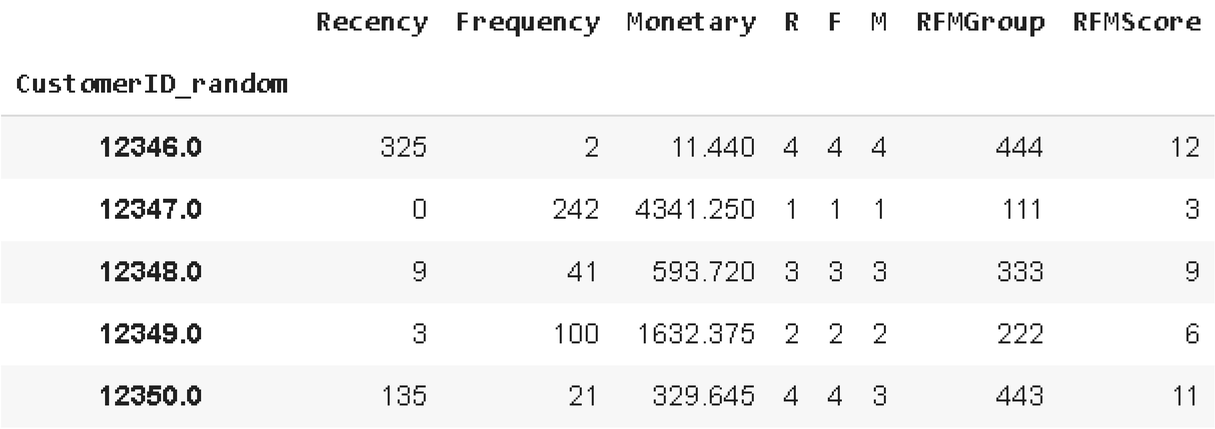

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
2
4
6
8
10
Frequency

Monetary


0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
Density
0
20000
40000
60000
80000
100000
120000
Monetary

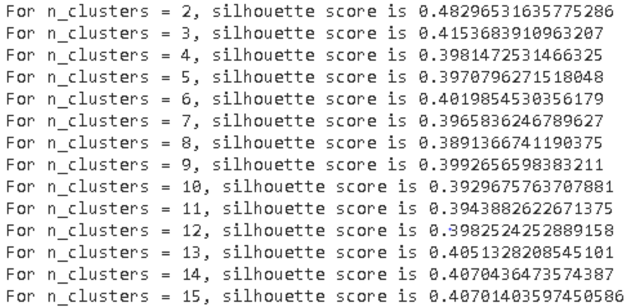

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
2
4
6
8
10
12
Monetary

RFM Scoring

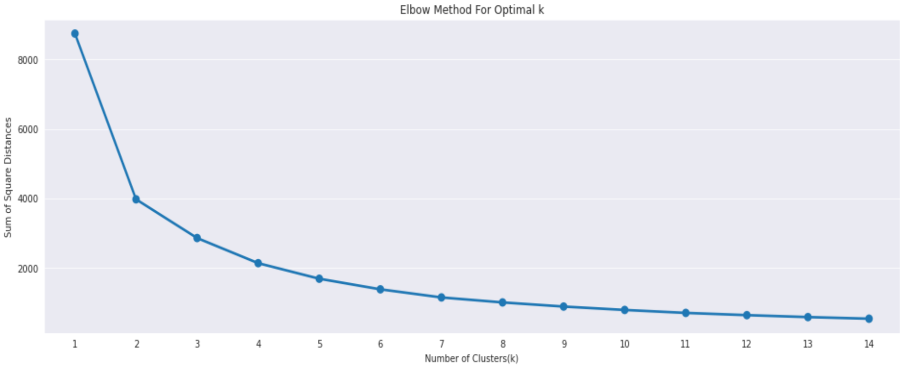


	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore
CustomerID_random								
12346.0	325	2	11.440	4	4	4	444	12
12347.0	0	242	4341.250	1	1	1	111	3
12348.0	9	41	593.720	3	3	3	333	9
12349.0	3	100	1632.375	2	2	2	222	6
12350.0	135	21	329.645	4	4	3	443	11

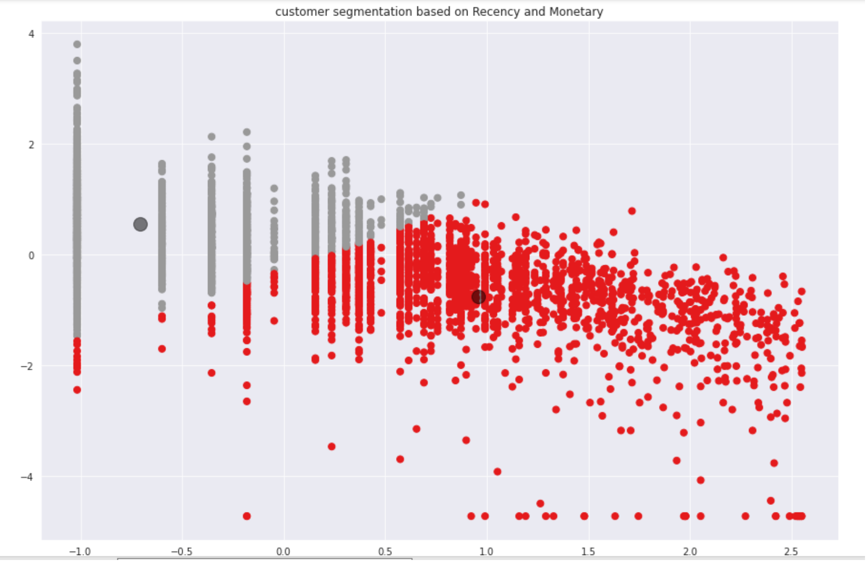
Silhouette score and Elbow method R & M



```
For n_clusters = 2, silhouette score is 0.48296531635775286
For n_clusters = 3, silhouette score is 0.4153683910963207
For n_clusters = 4, silhouette score is 0.3981472531466325
For n_clusters = 5, silhouette score is 0.3970796271518048
For n_clusters = 6, silhouette score is 0.4019854530356179
For n_clusters = 7, silhouette score is 0.3965836246789627
For n_clusters = 8, silhouette score is 0.3891366741190375
For n_clusters = 9, silhouette score is 0.3992656598383211
For n_clusters = 10, silhouette score is 0.3929675763707881
For n_clusters = 11, silhouette score is 0.3943882622671375
For n_clusters = 12, silhouette score is 0.3982524252889158
For n_clusters = 13, silhouette score is 0.4051328208545101
For n_clusters = 14, silhouette score is 0.4070436473574387
For n_clusters = 15, silhouette score is 0.40701403597450586
```

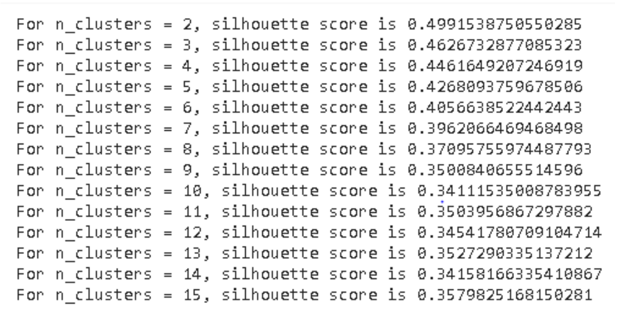


Elbow Method For Optimal k

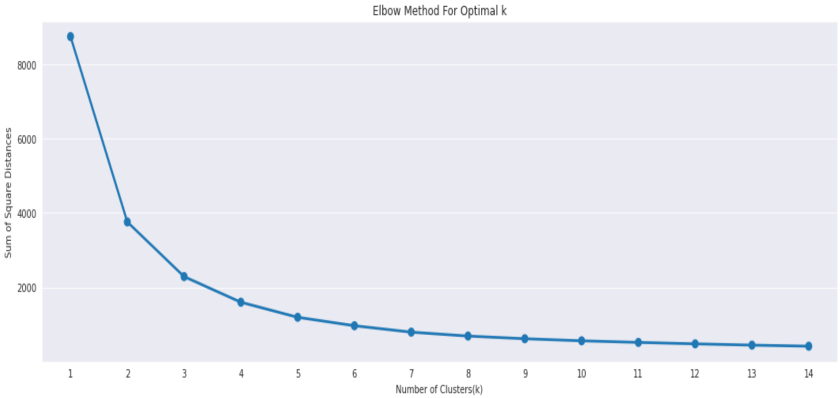


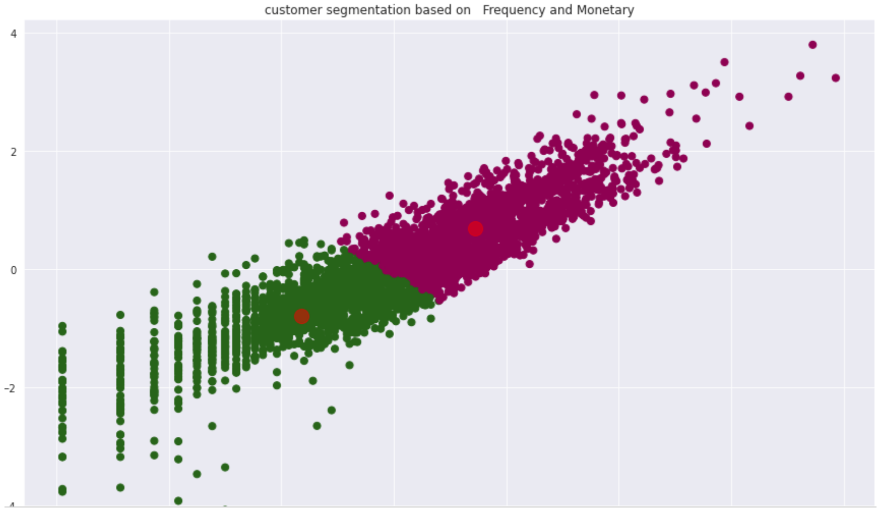
customer segmentation based on Recency and Monetary

Silhouette score and Elbow method F & M

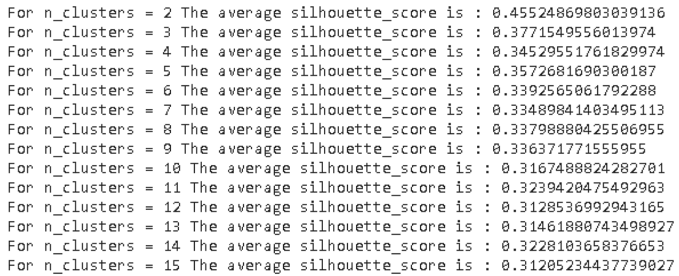


```
For n_clusters = 2, silhouette score is 0.4991538750550285
For n_clusters = 3, silhouette score is 0.4626732877085323
For n_clusters = 4, silhouette score is 0.4461649207246919
For n_clusters = 5, silhouette score is 0.4268093759678506
For n_clusters = 6, silhouette score is 0.4056638522442443
For n_clusters = 7, silhouette score is 0.3962066469468498
For n_clusters = 8, silhouette score is 0.37095755974487793
For n_clusters = 9, silhouette score is 0.3500840655514596
For n_clusters = 10, silhouette score is 0.34111535008783955
For n_clusters = 11, silhouette score is 0.3503956867297882
For n_clusters = 12, silhouette score is 0.34541780709104714
For n_clusters = 13, silhouette score is 0.3527290335137212
For n_clusters = 14, silhouette score is 0.34158166335410867
For n_clusters = 15, silhouette score is 0.3579825168150281
```

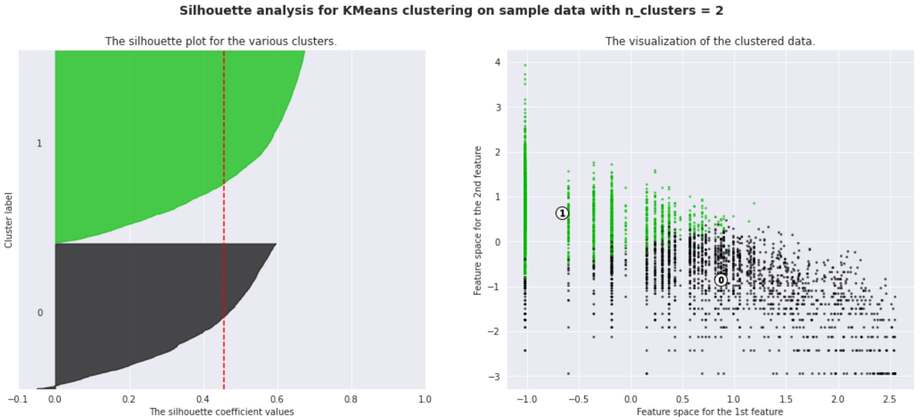




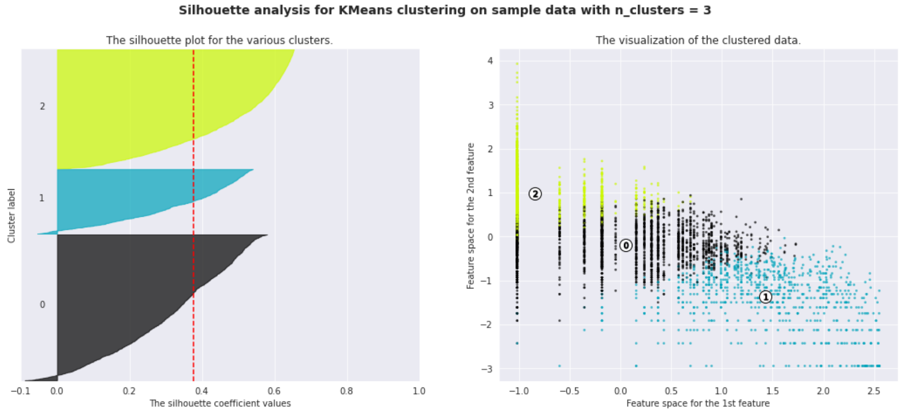
Silhouette analysis on RFM



```
For n_clusters = 2 The average silhouette_score is : 0.45524869803039136
For n_clusters = 3 The average silhouette_score is : 0.3771549556013974
For n_clusters = 4 The average silhouette_score is : 0.34529551761829974
For n_clusters = 5 The average silhouette_score is : 0.3572681690300187
For n_clusters = 6 The average silhouette_score is : 0.3392565061792288
For n_clusters = 7 The average silhouette_score is : 0.33489841403495113
For n_clusters = 8 The average silhouette_score is : 0.33798880425506955
For n_clusters = 9 The average silhouette_score is : 0.336371771555955
For n_clusters = 10 The average silhouette_score is : 0.3167488824282701
For n_clusters = 11 The average silhouette_score is : 0.3239420475492963
For n_clusters = 12 The average silhouette_score is : 0.3128536992943165
For n_clusters = 13 The average silhouette_score is : 0.31461880743498927
For n_clusters = 14 The average silhouette_score is : 0.3228103658376653
For n_clusters = 15 The average silhouette_score is : 0.31205234437739027
```

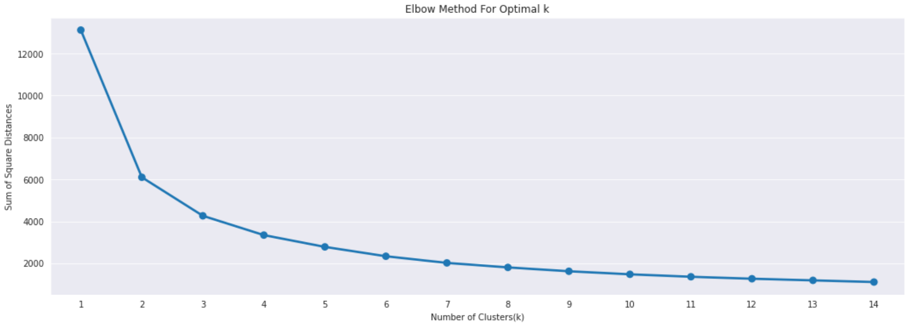


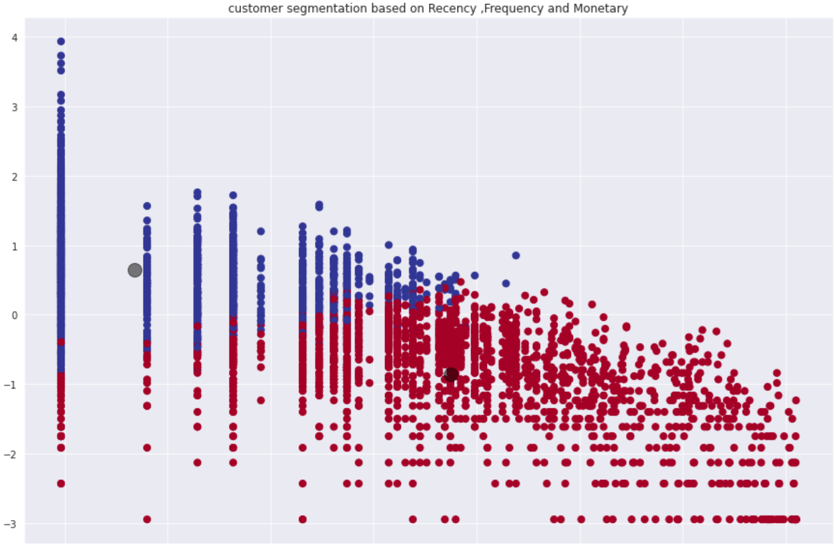
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



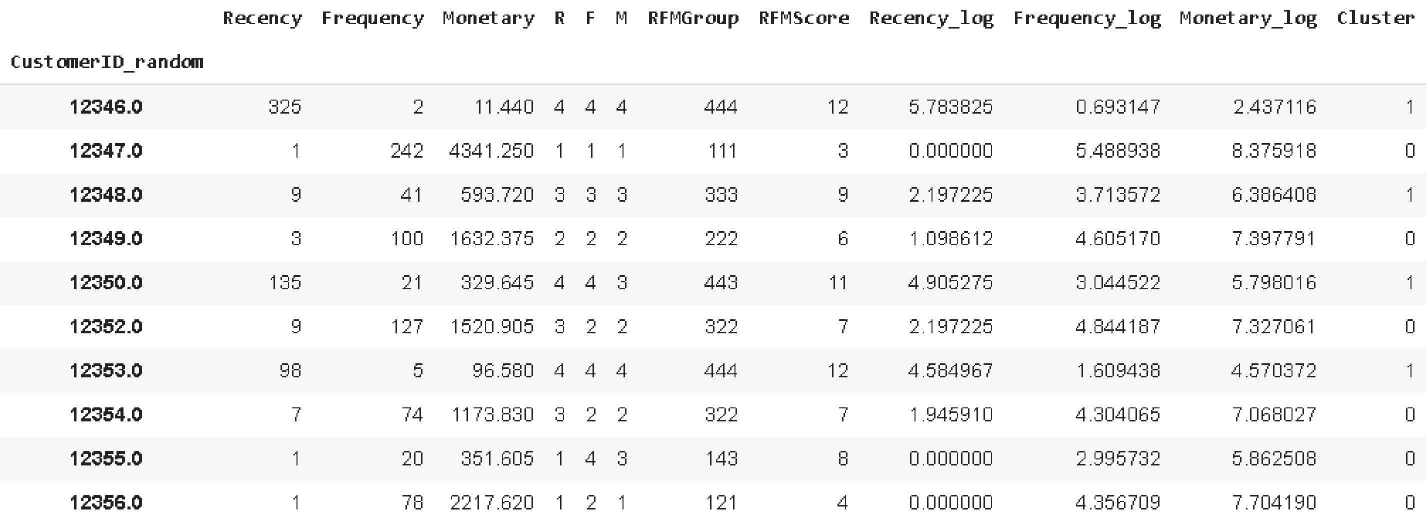
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

Elbow method and Cluster chart of RFM





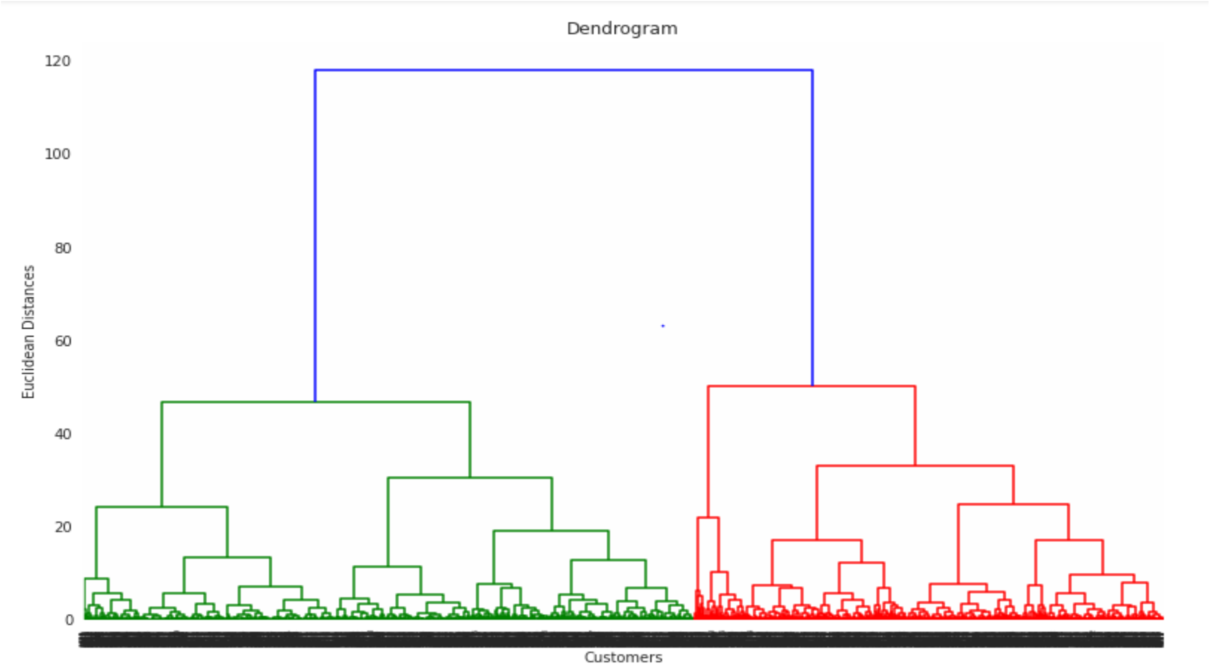
RFM Analysis



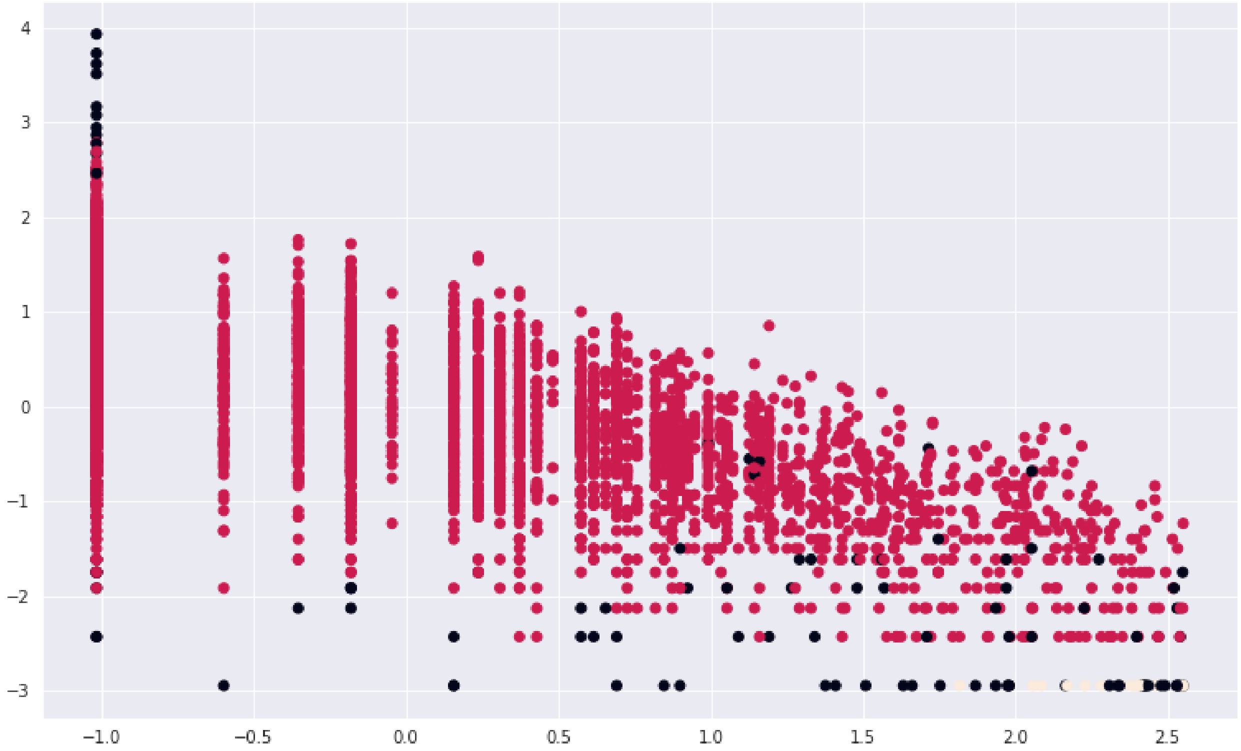
	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	Recency_log	Frequency_log	Monetary_log	Cluster
CustomerID_random												
12346.0	325	2	11.440	4	4	4	444	12	5.783825	0.693147	2.437116	1
12347.0	1	242	4341.250	1	1	1	111	3	0.000000	5.488938	8.375918	0
12348.0	9	41	593.720	3	3	3	333	9	2.197225	3.713572	6.386408	1
12349.0	3	100	1632.375	2	2	2	222	6	1.098612	4.605170	7.397791	0
12350.0	135	21	329.645	4	4	3	443	11	4.905275	3.044522	5.798016	1
12352.0	9	127	1520.905	3	2	2	322	7	2.197225	4.844187	7.327061	0
12353.0	98	5	96.580	4	4	4	444	12	4.584967	1.609438	4.570372	1
12354.0	7	74	1173.830	3	2	2	322	7	1.945910	4.304065	7.068027	0
12355.0	1	20	351.605	1	4	3	143	8	0.000000	2.995732	5.862508	0
12356.0	1	78	2217.620	1	2	1	121	4	0.000000	4.356709	7.704190	0

Dendogram

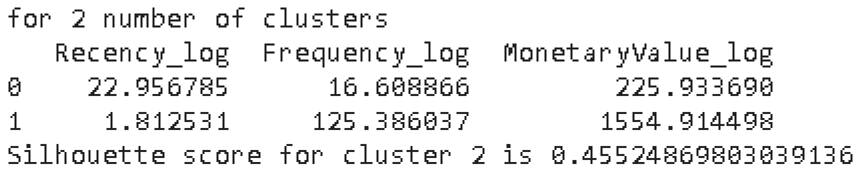
- Hyper Parameter:
- AgglomerativeClustering
- (n_clusters = 2,
- affinity = 'euclidean’,
- linkage = 'ward')



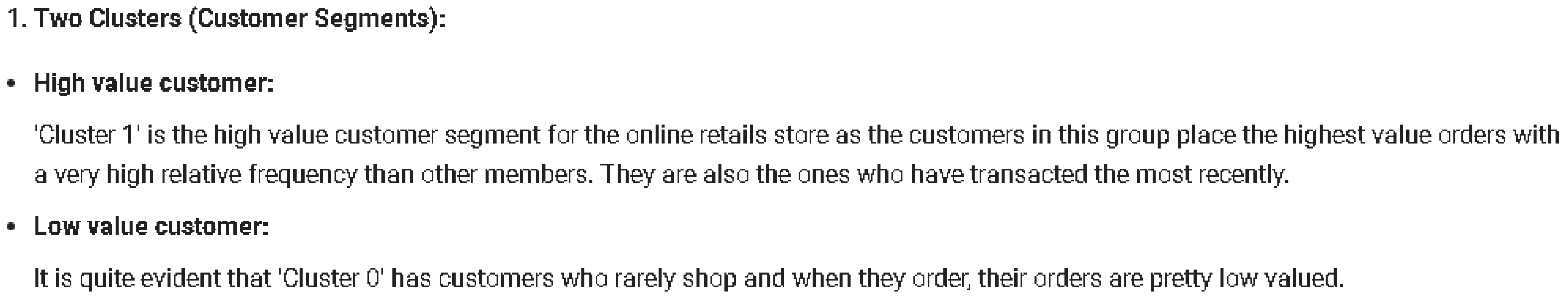
DBSCAN



Cluster Analysis and Business Relevance



```
for 2 number of clusters
   Recency_log  Frequency_log  MonetaryValue_log
0    22.956785      16.608866         225.933690
1     1.812531     125.386037        1554.914498
Silhouette score for cluster 2 is 0.45524869803039136
```



1. Two Clusters (Customer Segments):

- **High value customer:**

  'Cluster 1' is the high value customer segment for the online retails store as the customers in this group place the highest value orders with a very high relative frequency than other members. They are also the ones who have transacted the most recently.

- **Low value customer:**

  It is quite evident that 'Cluster 0' has customers who rarely shop and when they order, their orders are pretty low valued.

Challenges

- Large dataset to handle
- Lot of Missing values in CustomerID
- Choosing right number of ‘k’ clusters

Conclusion

- K means Clustering with silhouette gives highest score with number of clusters = 2

- The customer segments thus deduced can be very useful in targeted marketing, scouting for new customers and ultimately revenue growth. After knowing the types of customers, it depends upon the retailer policy whether to chase the high value customers and offer them better service and discounts or try and encourage low/ medium value customers to shop more frequently or of higher monetary values.

Thank You