Capstone Project-4

CUSTOMER SEGMENTATION

Team Members:

Soni Rani

Vivek Kumar

Suraj Singh

Content

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.

Introduction

- Businesses all over the world are growing every day. With the help of technology, they have access to a wider market and hence, a large customer base.

- Customer segmentation refers to categorizing customers into different groups with similar characteristics.

- Customer segmentation can help businesses focus on each customer group in a different way, in order to maximize benefits for customers as well as the business.

- This project mainly deals in segmenting customers of an online business store in the UK.

Data Summary

- A transnational data set with transactions occurring between 1st December 2010 and 9th 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.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Feature Summary

- The contents of the data had features such as:
- 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.
- Unit Price: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer. Country: Country name. Nominal, the name of the country where each customer resides.

Data Cleaning

- In this dataset , we have null values present in the ‘CustomerID’ and ‘Description’ column. These have to be dropped as there is no way of filling them strategically.

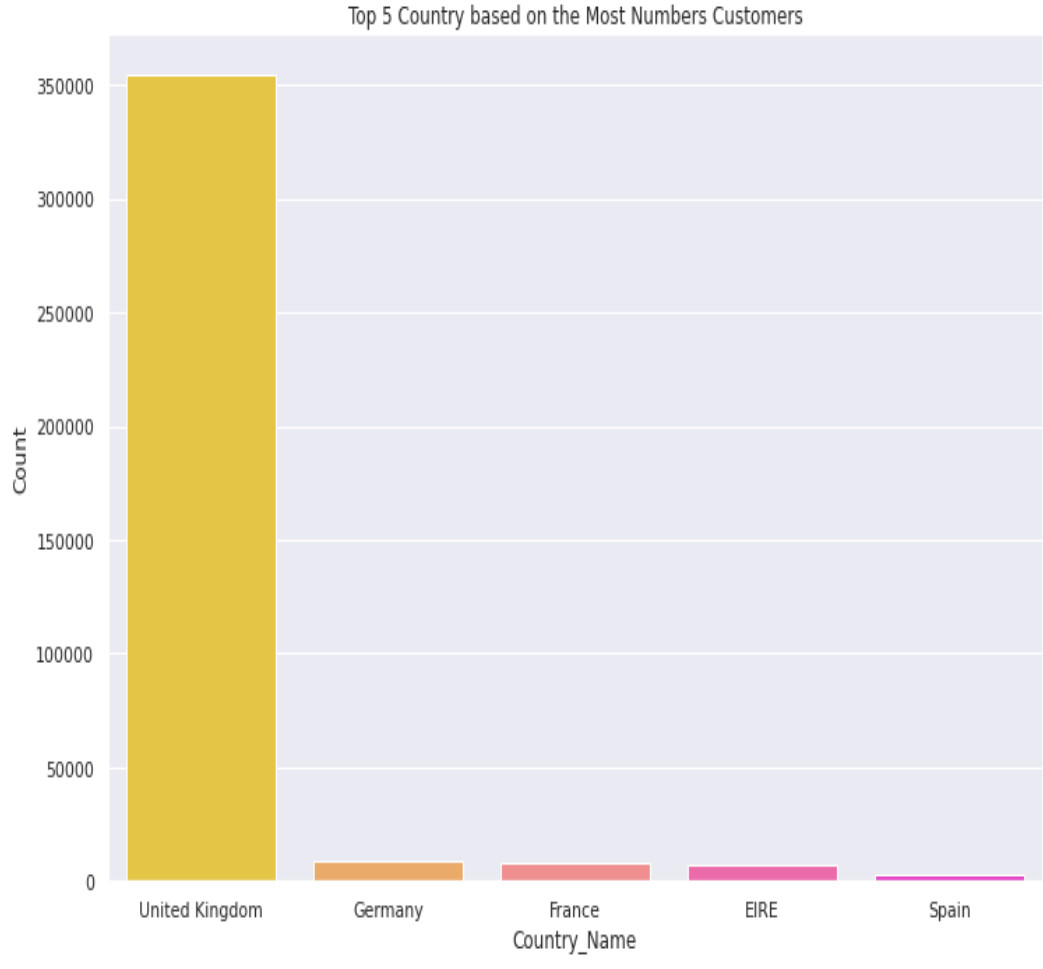
- Cancelled orders exist in the data, these too have been removed.

- Date, month and year were extracted from the ‘InvoiceDate’ column.

Exploratory Data Analysis

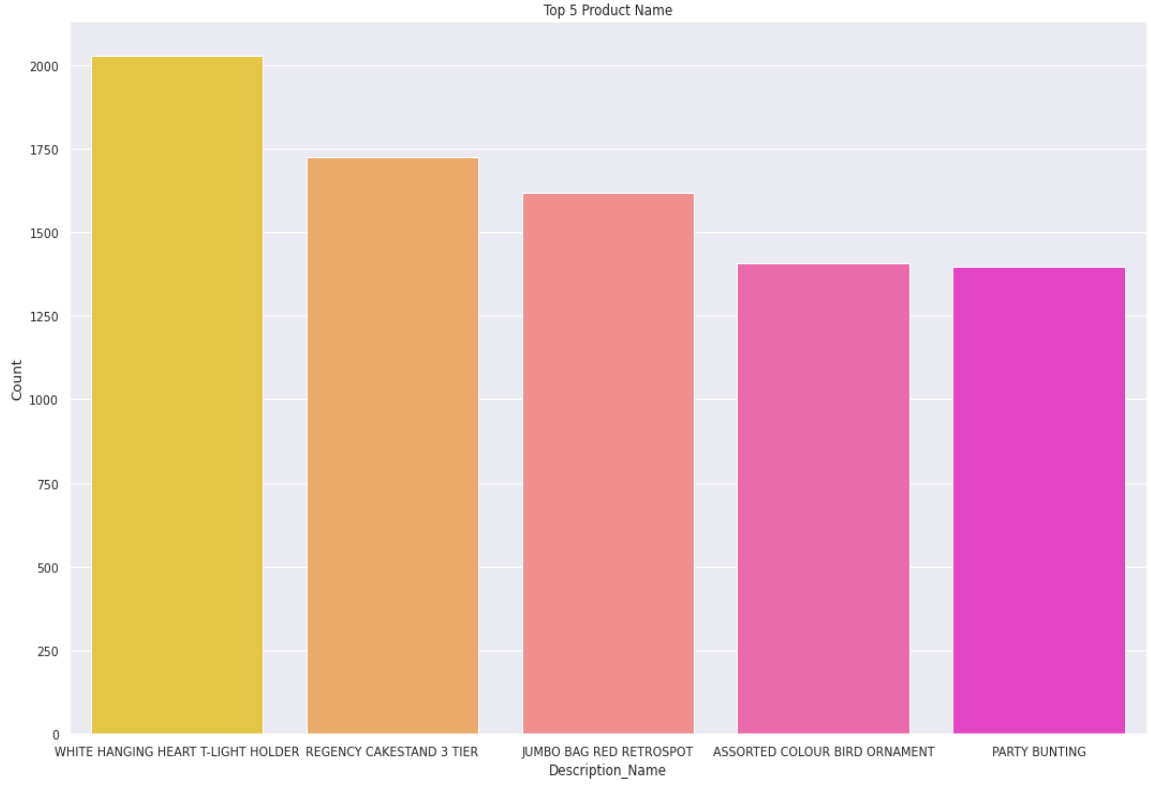
- UK, Germany, France were top countries having more no. of customers.

- Since data belonged to UK based company, UK had majority of customers.



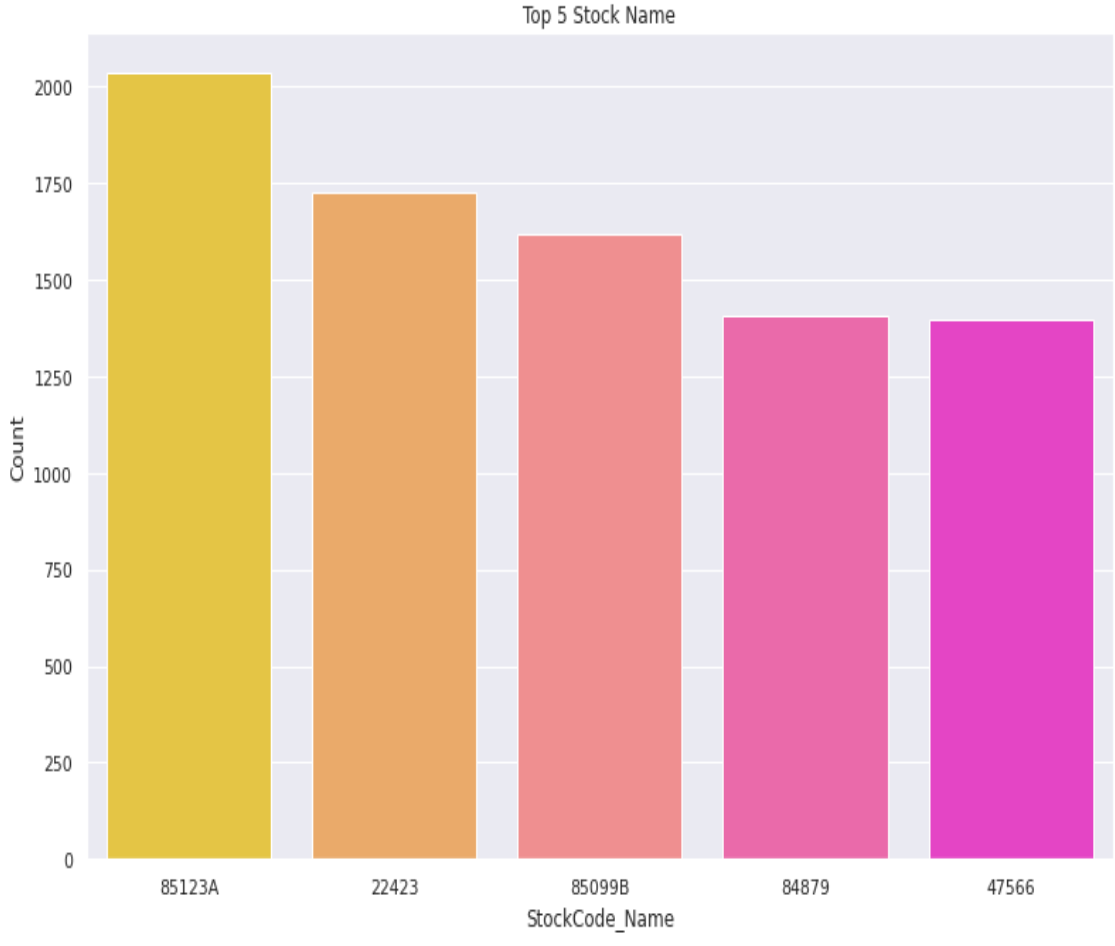
EDA (Continued)

	Description_Name	Count
0	WHITE HANGING HEART T-LIGHT HOLDER	2028
1	REGENCY CAKESTAND 3 TIER	1724
2	JUMBO BAG RED RETROSPOT	1618
3	ASSORTED COLOUR BIRD ORNAMENT	1408
4	PARTY BUNTING	1397

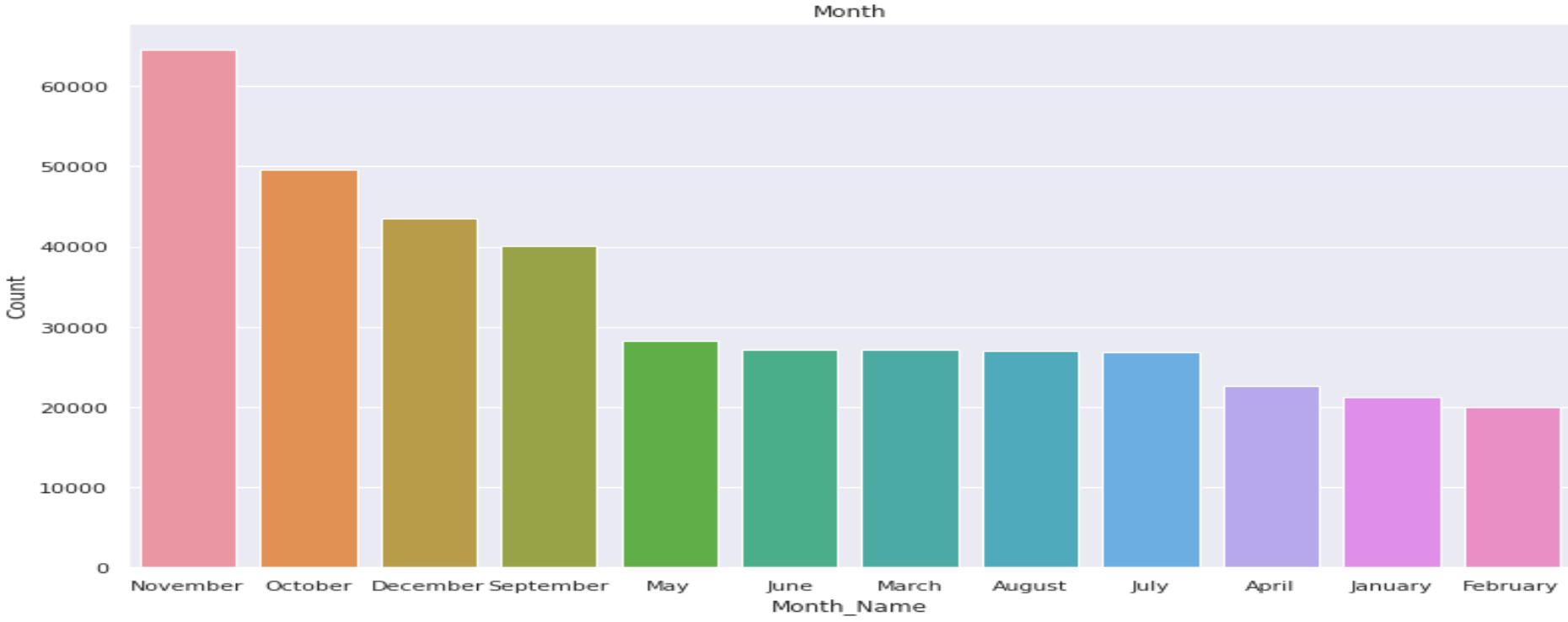


EDA (Continued)

	StockCode_Name	Count
0	85123A	2035
1	22423	1724
2	85099B	1618
3	84879	1408
4	47566	1397

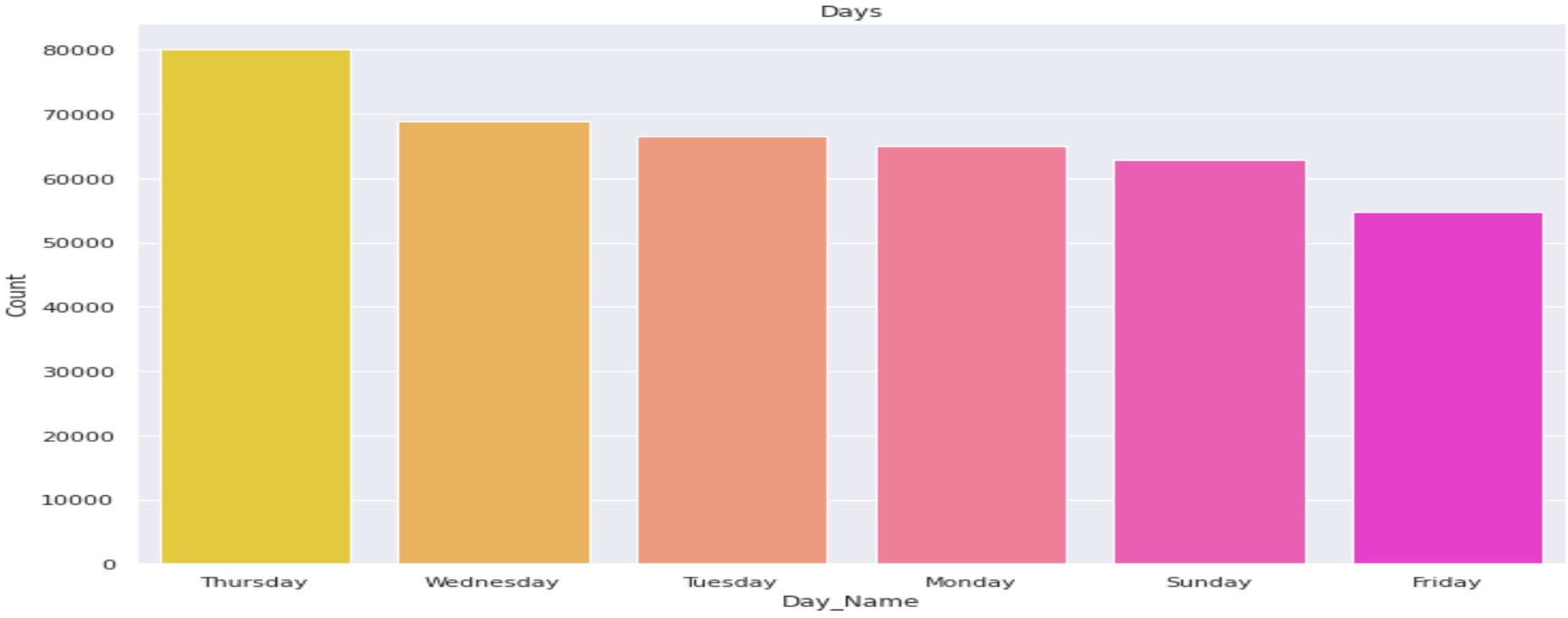


EDA (Continued)



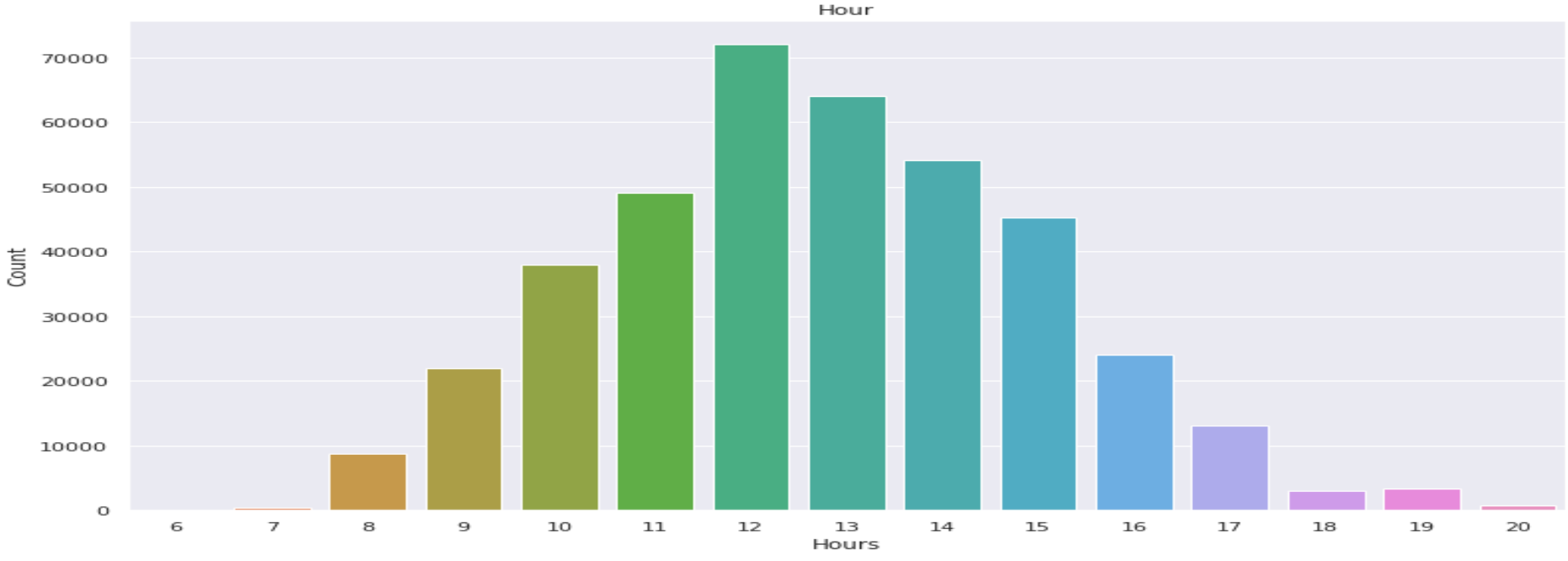
- Most numbers of customers have purchased the gifts in the month of November, October and December.
- Least numbers of purchasing are in the month of April and February.

EDA (Continued)



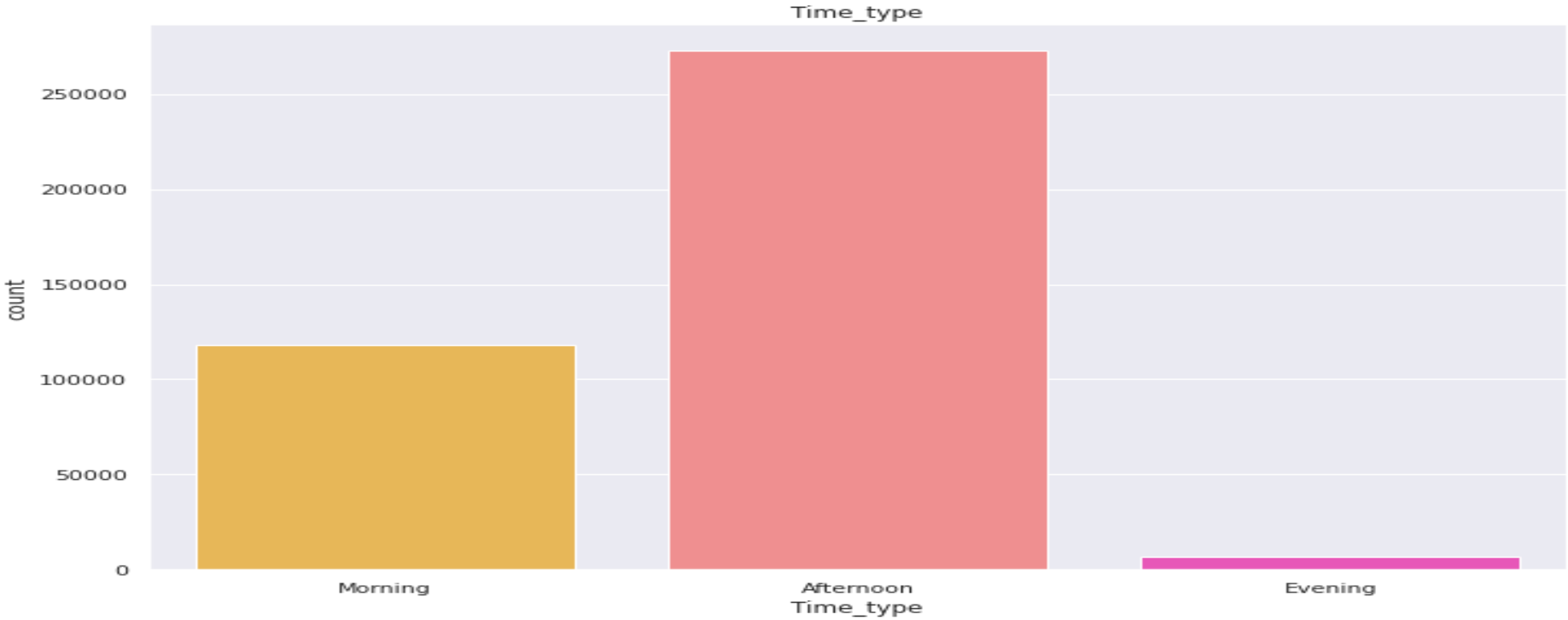
- Most of the customers have purchased the items in Thursday ,Wednesday and Tuesday.

EDA (Continued)



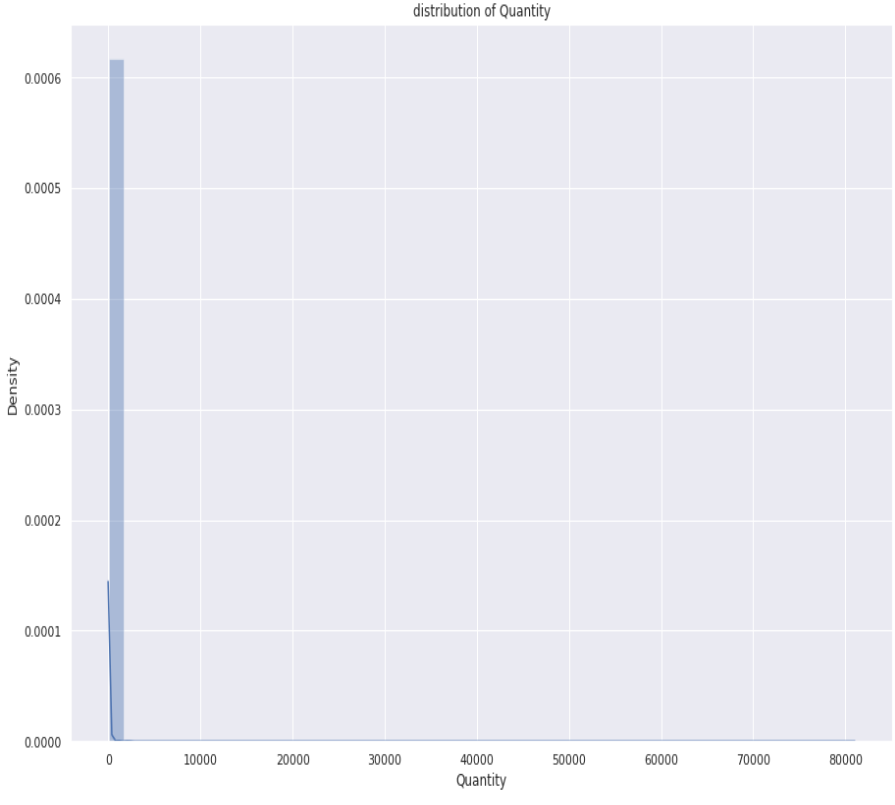
- Most numbers of purchasing is done between 12pm to 3pm.

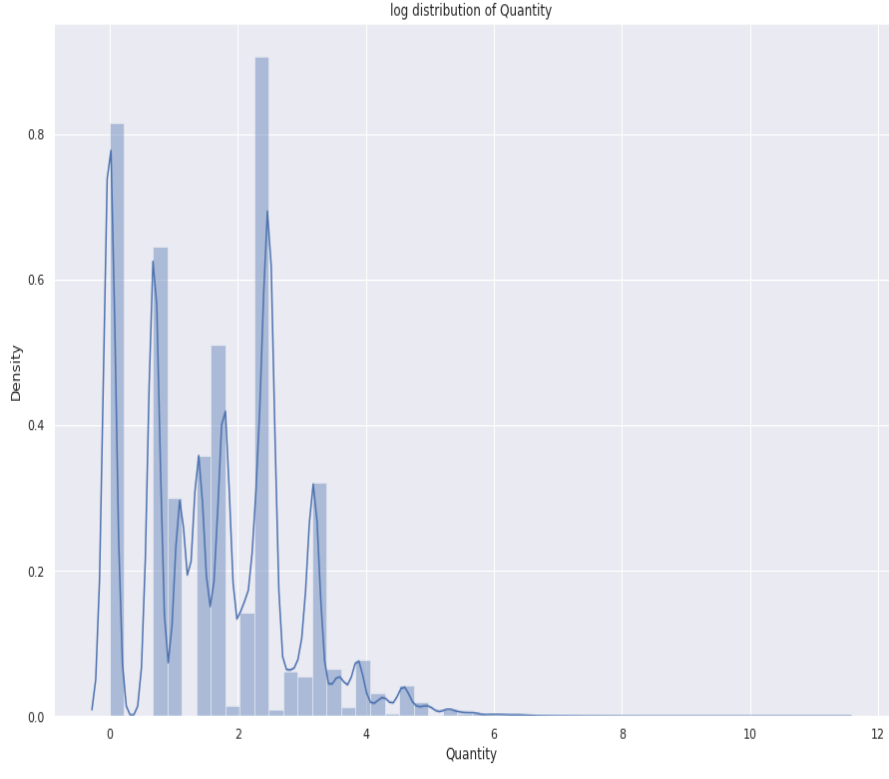
EDA (Continued)



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

EDA (Continued)

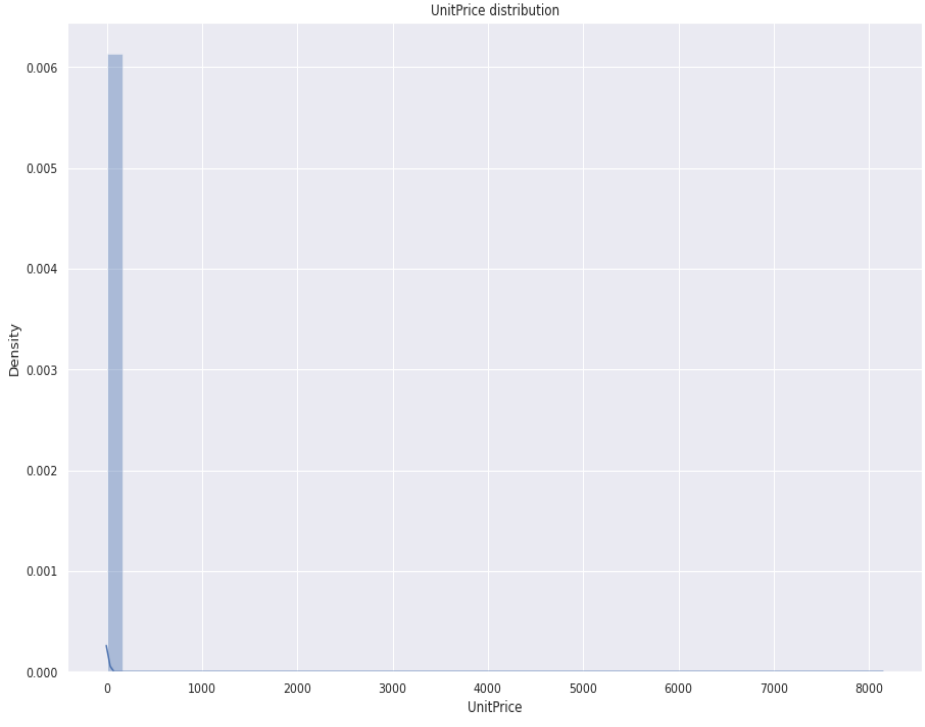


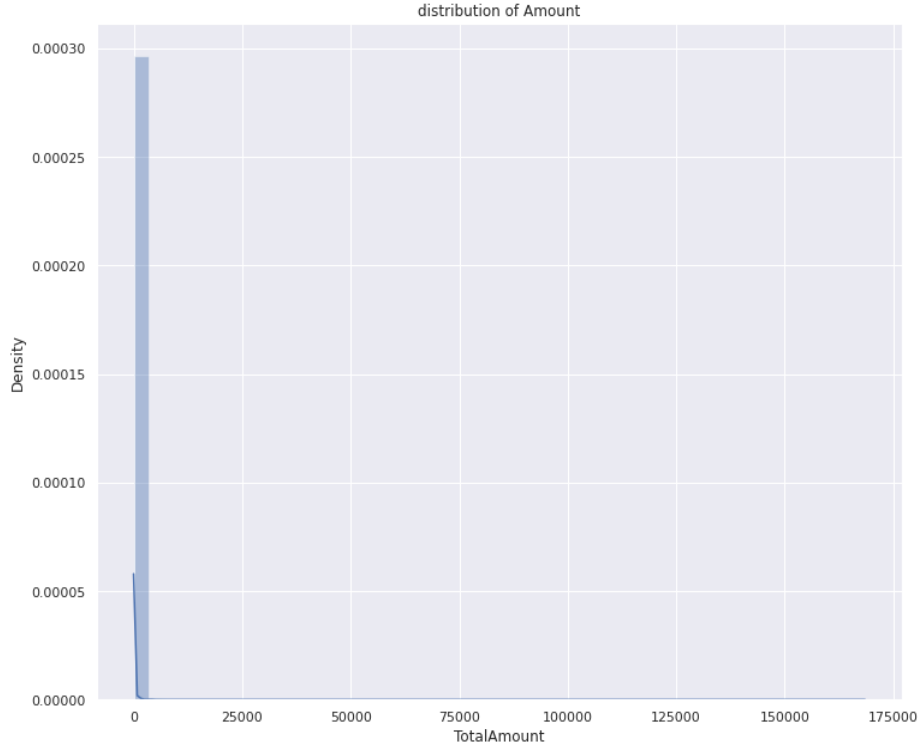


- Distribution of Quantity is highly positively skewed.

- Distribution of Quantity After log transformation.

EDA (Continued)





- Distribution of unit and total amount is highly positively skewed.

Recency, Frequency and 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

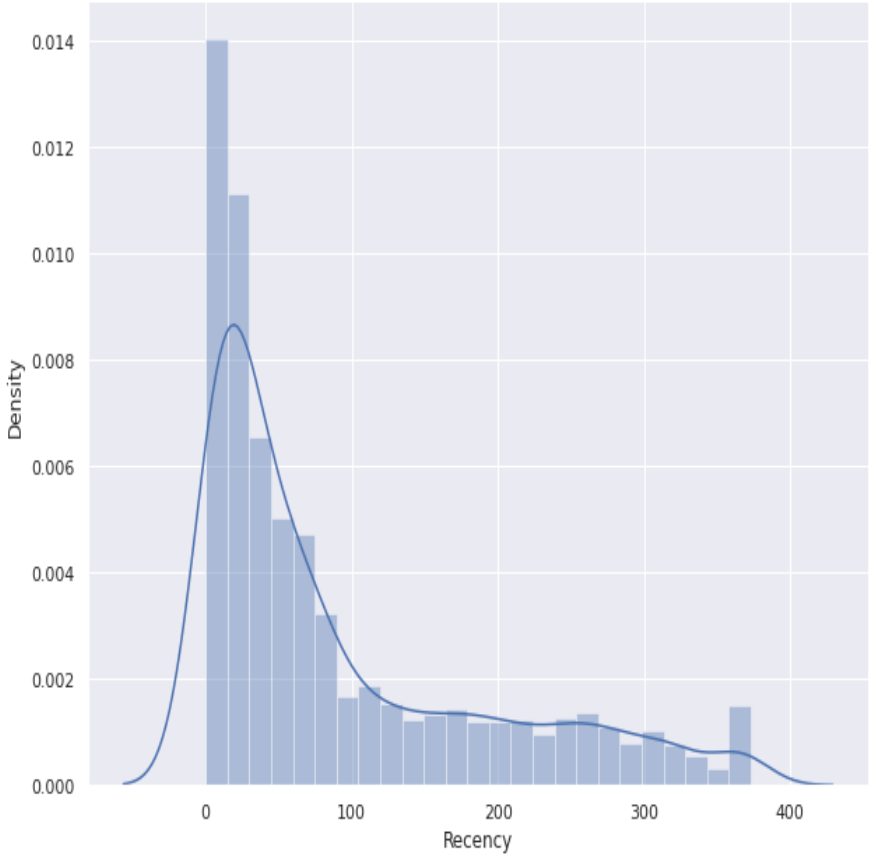
E.g. Total number of transactions or average time between transactions/ engaged visits

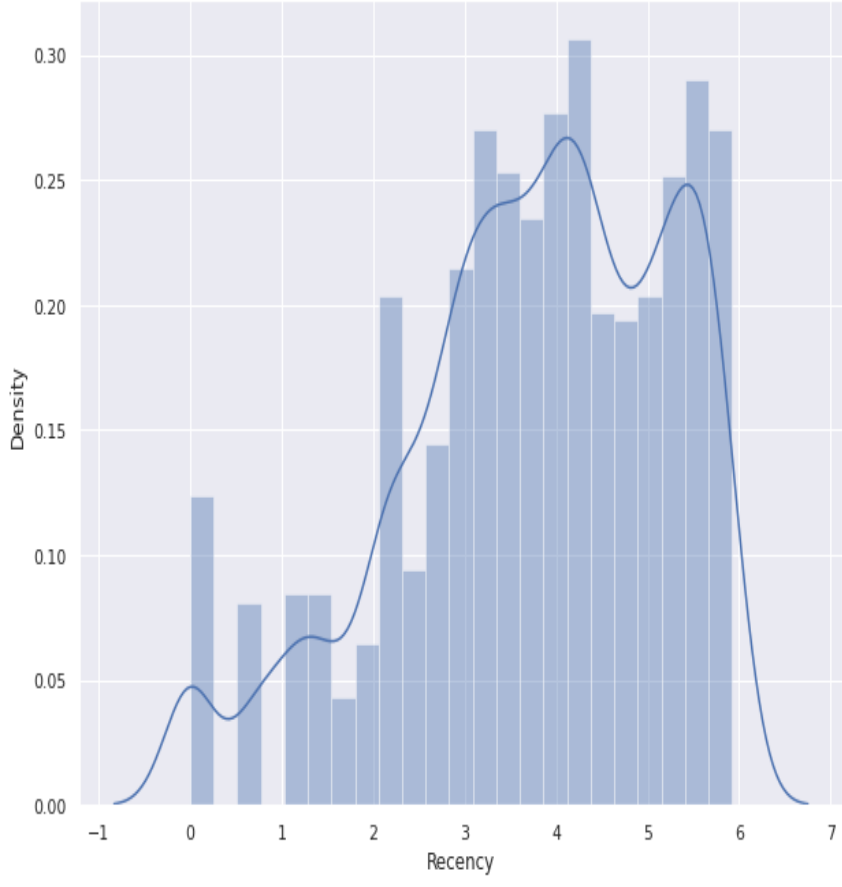
MONETARY

The intention of customer to spend or purchasing power of customer

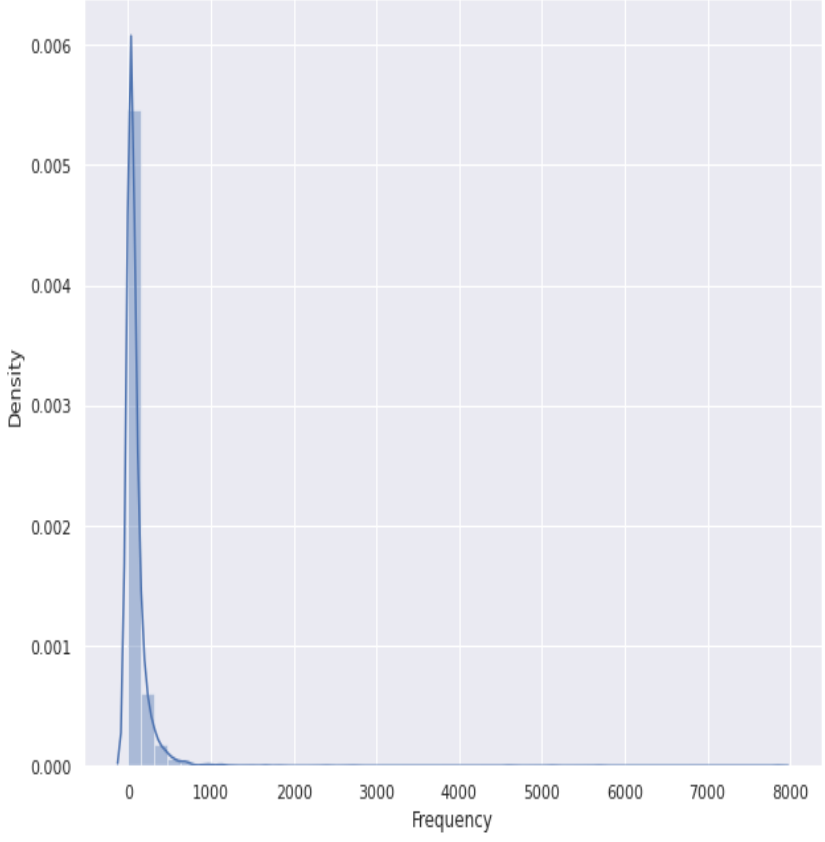
E.g. Total or average transactions value

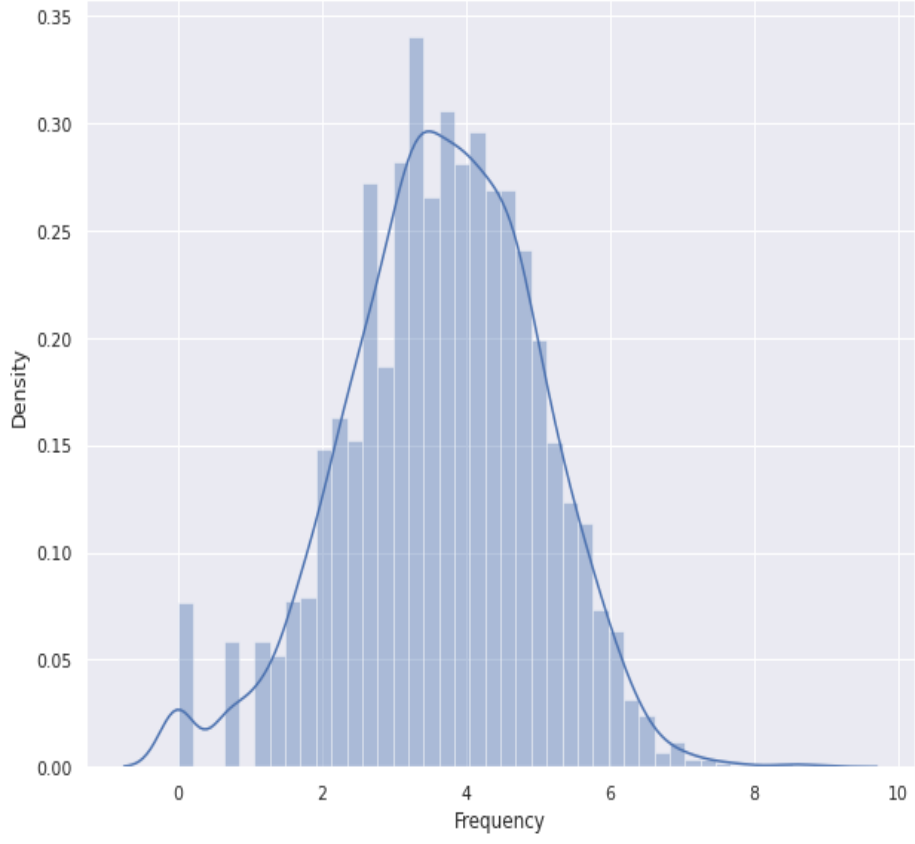
Recency


0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
100
200
300
400
Recency

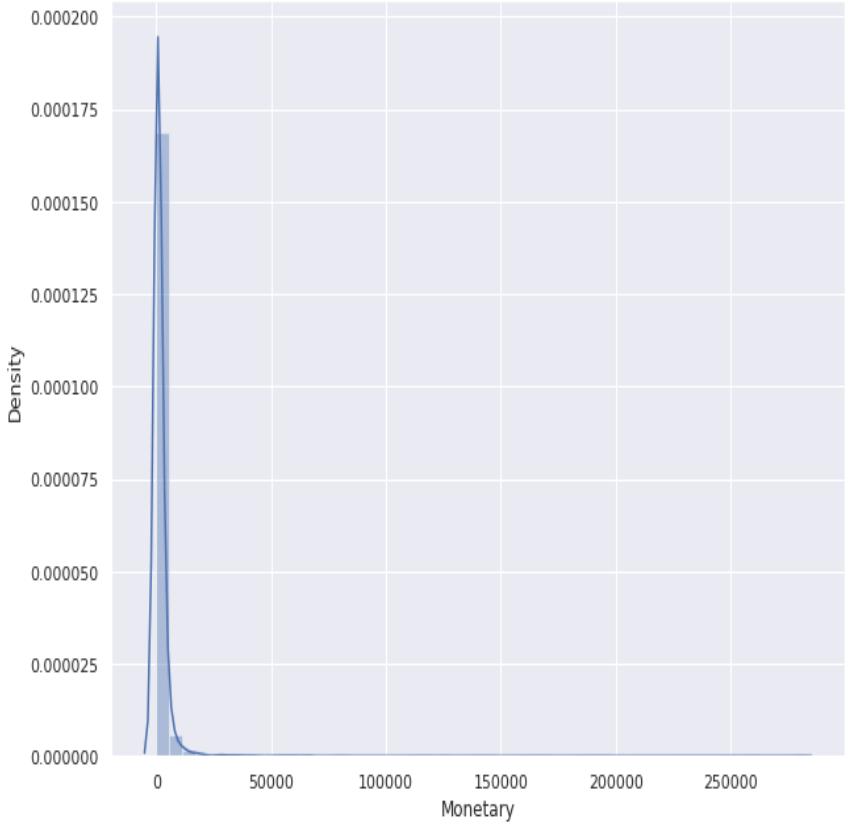

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
−1
0
1
2
3
4
5
6
7
Recency

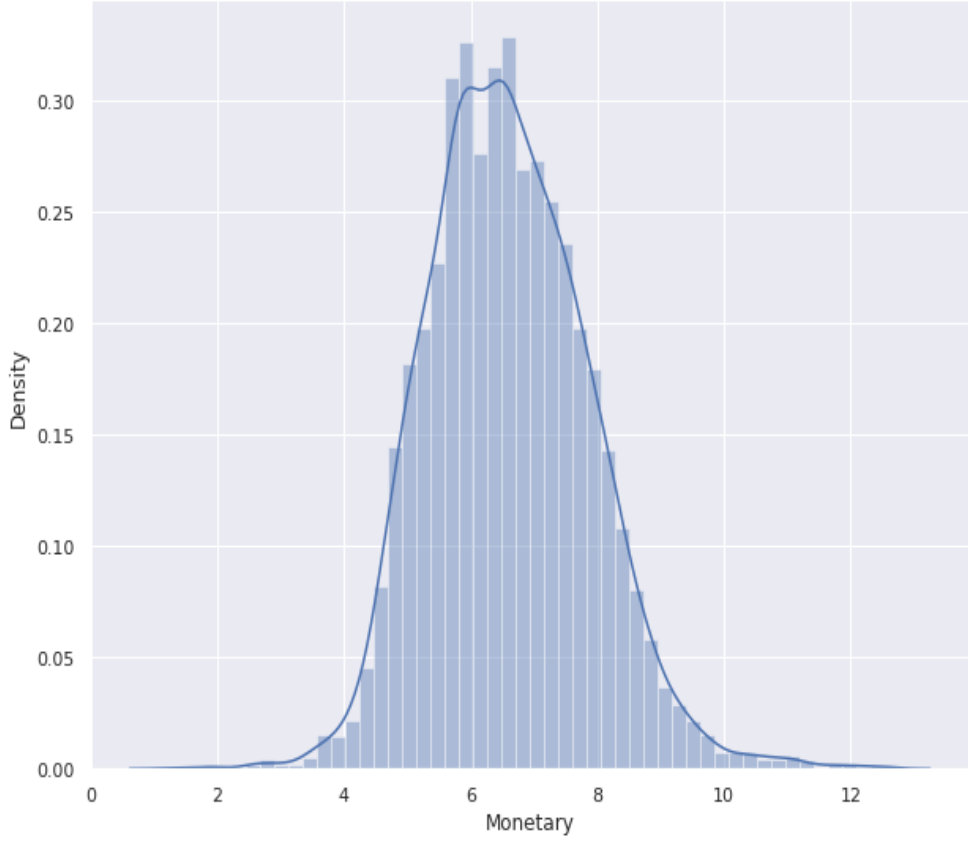
Frequency


0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
0
1000
2000
3000
4000
5000
6000
7000
8000
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

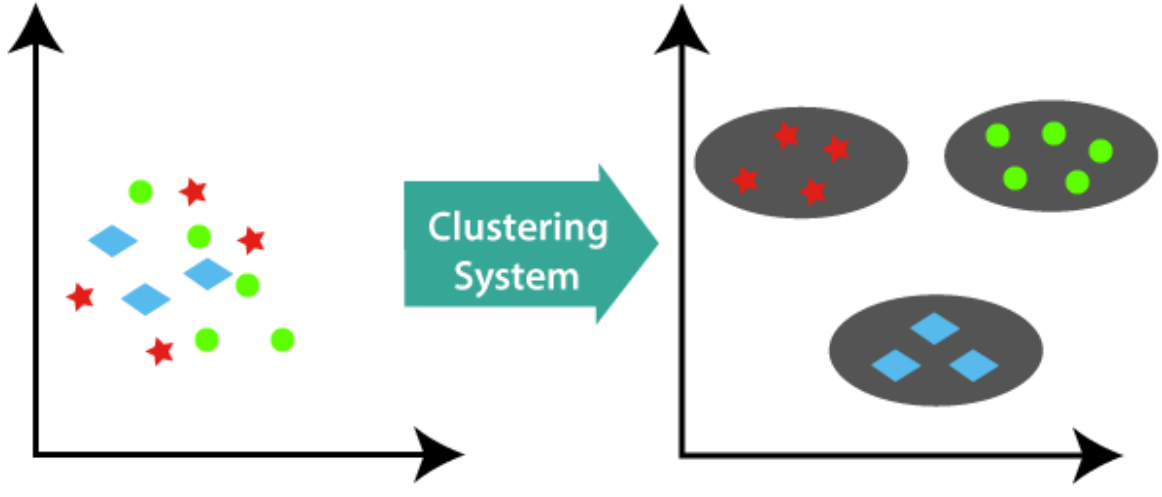

0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
Density
0
50000
100000
150000
200000
250000
Monetary


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

Clustering

- Clustering is an unsupervised classification technique to understand the groups of classes in the data.

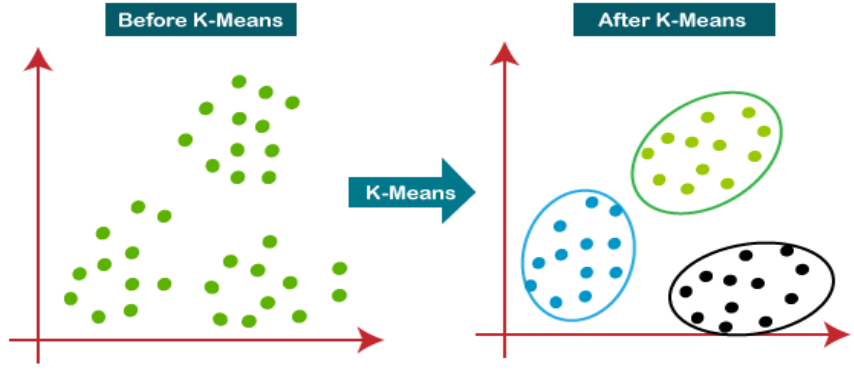
- In this section, we use KMeans algorithm and hierarchical clustering to cluster the customers into different segments.



K-Means Clustering

- K-means algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

- K-Means requires the number of clusters to be specified during the model building process. To know the right number of clusters, methods such as silhouette analysis and elbow method can be used. These methods will help in selection of the optimum number of clusters.



Methods to find optimal clusters

Silhouette score : Silhouette score is used to evaluate the quality of clusters that ranges from -1 to1 , where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

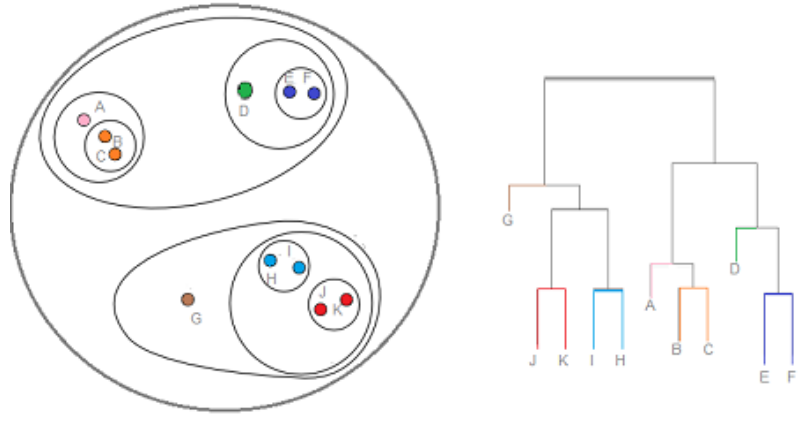
Elbow method : a point from where the value of clusters starts decreasing suddenly. It calculates the sum of the square of the points and calculates the average distance.

DBSCAN : DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It is basically a clustering algorithm based on density.

Hierarchical clustering

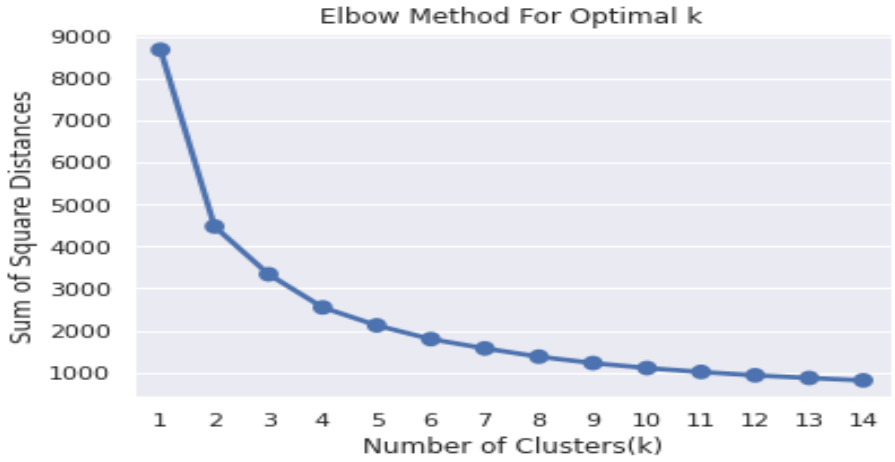
- Hierarchical clustering is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom. To get the number of clusters for hierarchical clustering, we make use of an awesome concept called a Dendogram.

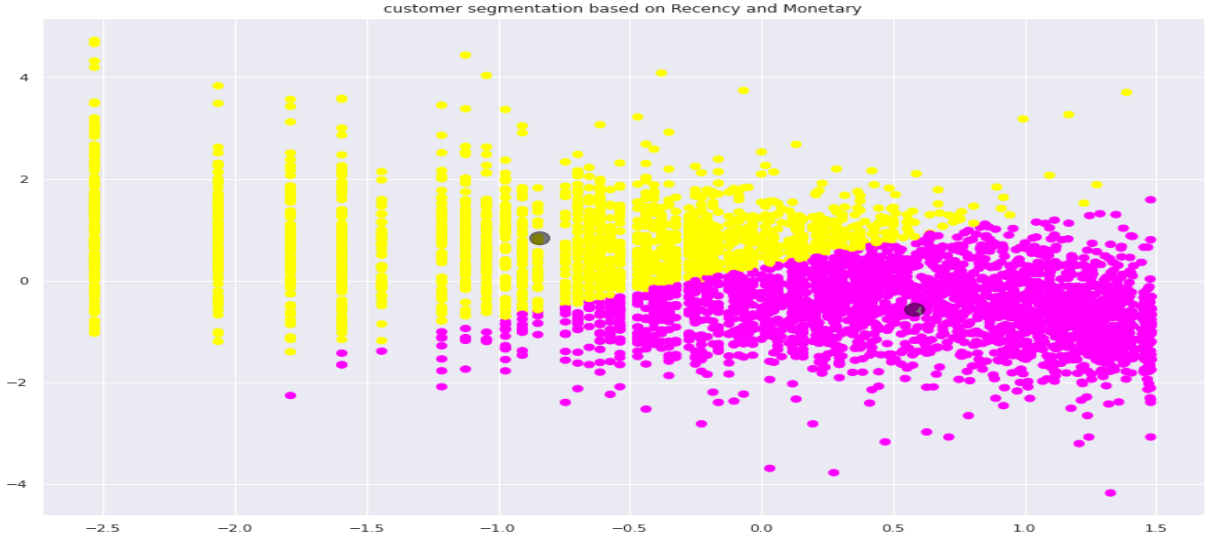
Dendogram : A Dendogram is a type of tree diagram showing hierarchical relationships between different sets of data.



SILHOUETTE SCORE AND ELBOW METHOD ON R&M

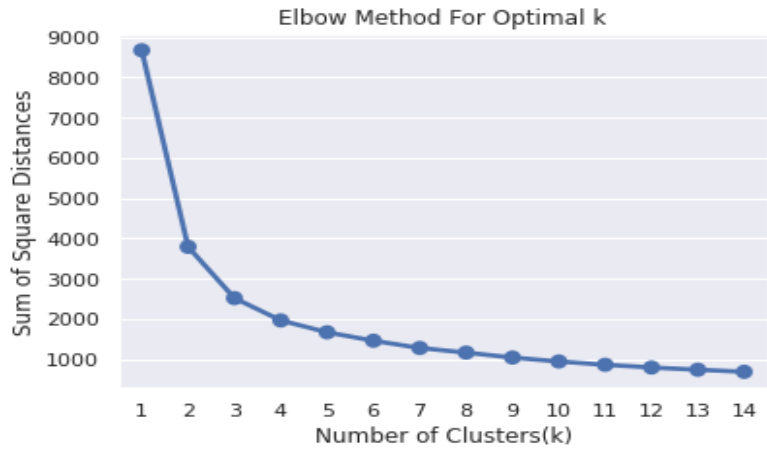
```
For n_clusters = 2, silhouette score is 0.42138910122697165
For n_clusters = 3, silhouette score is 0.34340822529173637
For n_clusters = 4, silhouette score is 0.36489854656957116
For n_clusters = 5, silhouette score is 0.3342349143457906
For n_clusters = 6, silhouette score is 0.34451670273920804
For n_clusters = 7, silhouette score is 0.3447813648137824
For n_clusters = 8, silhouette score is 0.3374707992583374
For n_clusters = 9, silhouette score is 0.3455560185153941
For n_clusters = 10, silhouette score is 0.34777103198333353
For n_clusters = 11, silhouette score is 0.33740543922116056
For n_clusters = 12, silhouette score is 0.3453686104458895
For n_clusters = 13, silhouette score is 0.3418869786215514
For n_clusters = 14, silhouette score is 0.34359568600215823
For n_clusters = 15, silhouette score is 0.33877193803993644
```

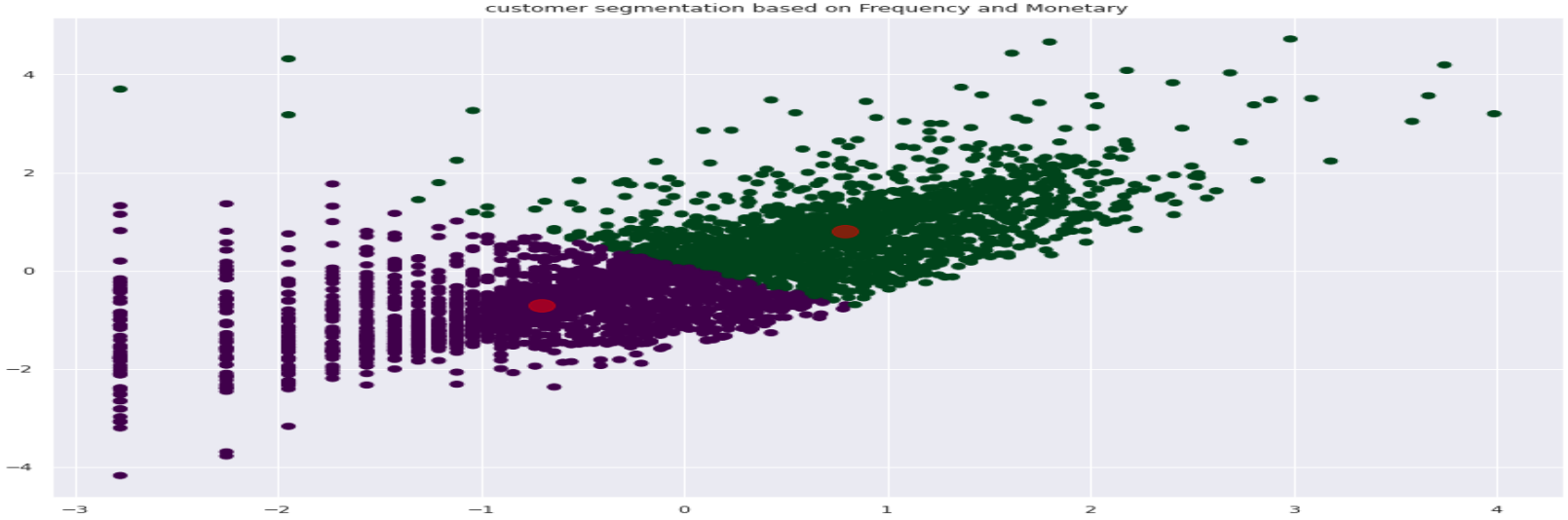




SILHOUETTE SCORE AND ELBOW METHOD ON F&M

```
For n_clusters = 2, silhouette score is 0.478535709506603
For n_clusters = 3, silhouette score is 0.40764120562174455
For n_clusters = 4, silhouette score is 0.3717868483347582
For n_clusters = 5, silhouette score is 0.3442982956807253
For n_clusters = 6, silhouette score is 0.35876031068888614
For n_clusters = 7, silhouette score is 0.3433784005206166
For n_clusters = 8, silhouette score is 0.35023779115767656
For n_clusters = 9, silhouette score is 0.3437281344168227
For n_clusters = 10, silhouette score is 0.3584593363276654
For n_clusters = 11, silhouette score is 0.3681864615017474
For n_clusters = 12, silhouette score is 0.35357508977280694
For n_clusters = 13, silhouette score is 0.36450403893982863
For n_clusters = 14, silhouette score is 0.3647014893236436
For n_clusters = 15, silhouette score is 0.3584819922771737
```

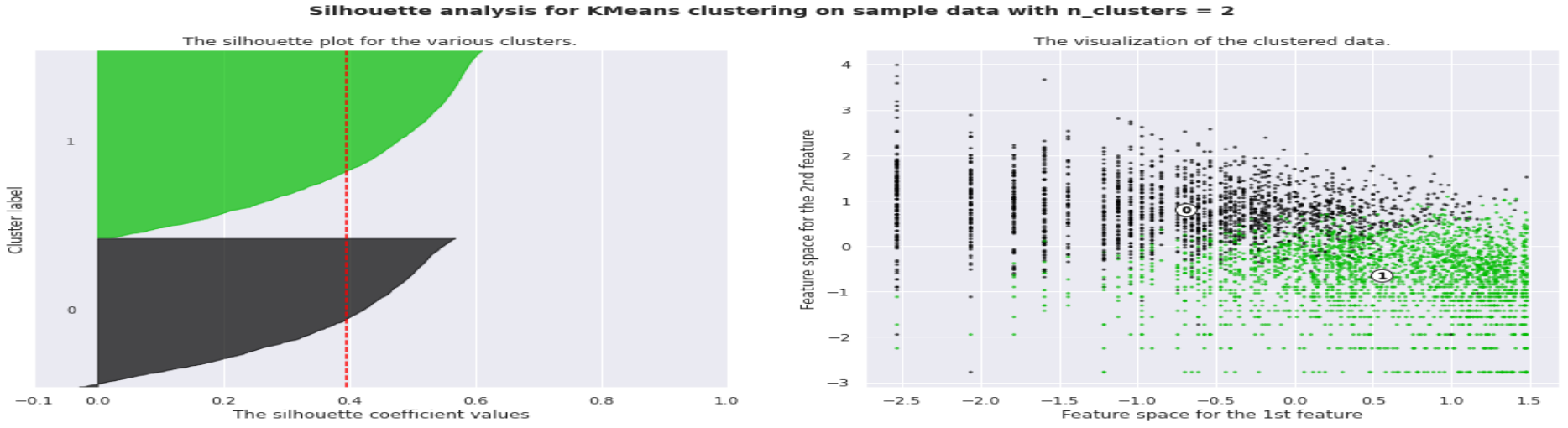




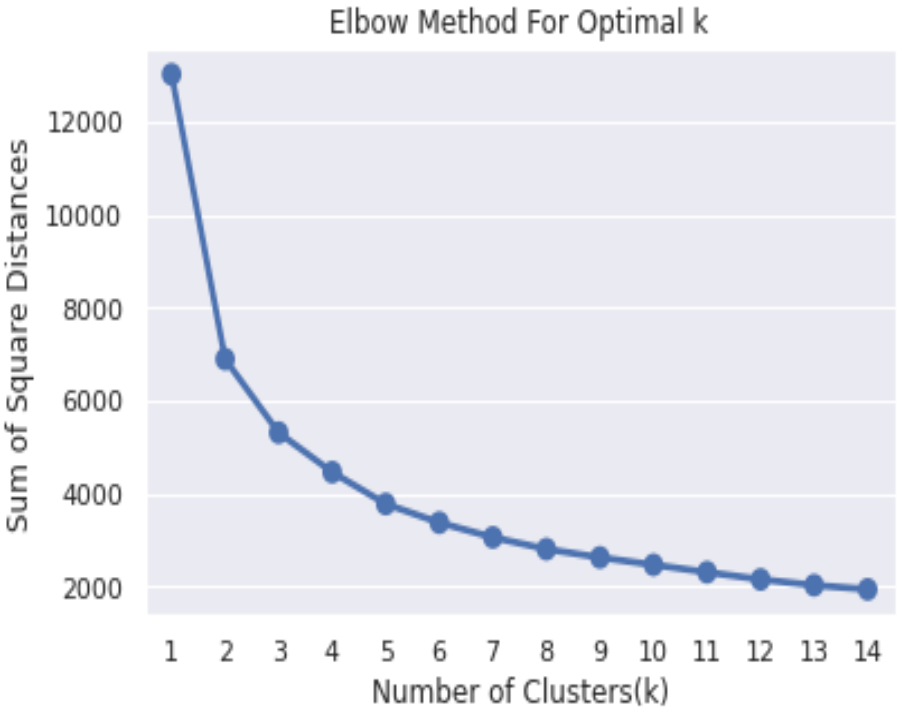
Silhouette analysis on R, F and M

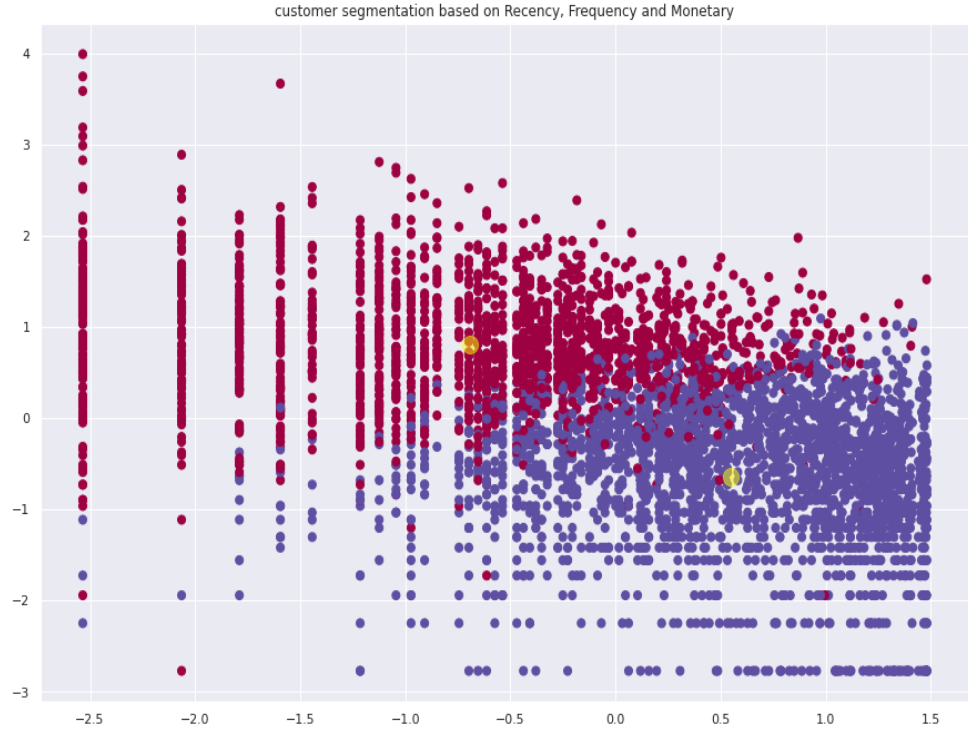
```
For n_clusters = 2 The average silhouette_score is : 0.3956478042246982
For n_clusters = 3 The average silhouette_score is : 0.3049826724447913
For n_clusters = 4 The average silhouette_score is : 0.30279724233096916
For n_clusters = 5 The average silhouette_score is : 0.2785519277480847
For n_clusters = 6 The average silhouette_score is : 0.2789560652501828
For n_clusters = 7 The average silhouette_score is : 0.2613208163968789
For n_clusters = 8 The average silhouette_score is : 0.2640918249728342
For n_clusters = 9 The average silhouette_score is : 0.2585642595481418
For n_clusters = 10 The average silhouette_score is : 0.2644733794304285
For n_clusters = 11 The average silhouette_score is : 0.2592423011915937
For n_clusters = 12 The average silhouette_score is : 0.26503813251658404
For n_clusters = 13 The average silhouette_score is : 0.2621555416679574
For n_clusters = 14 The average silhouette_score is : 0.26140947155997746
For n_clusters = 15 The average silhouette_score is : 0.258754625338637
```

Silhouette analysis on R, F and M

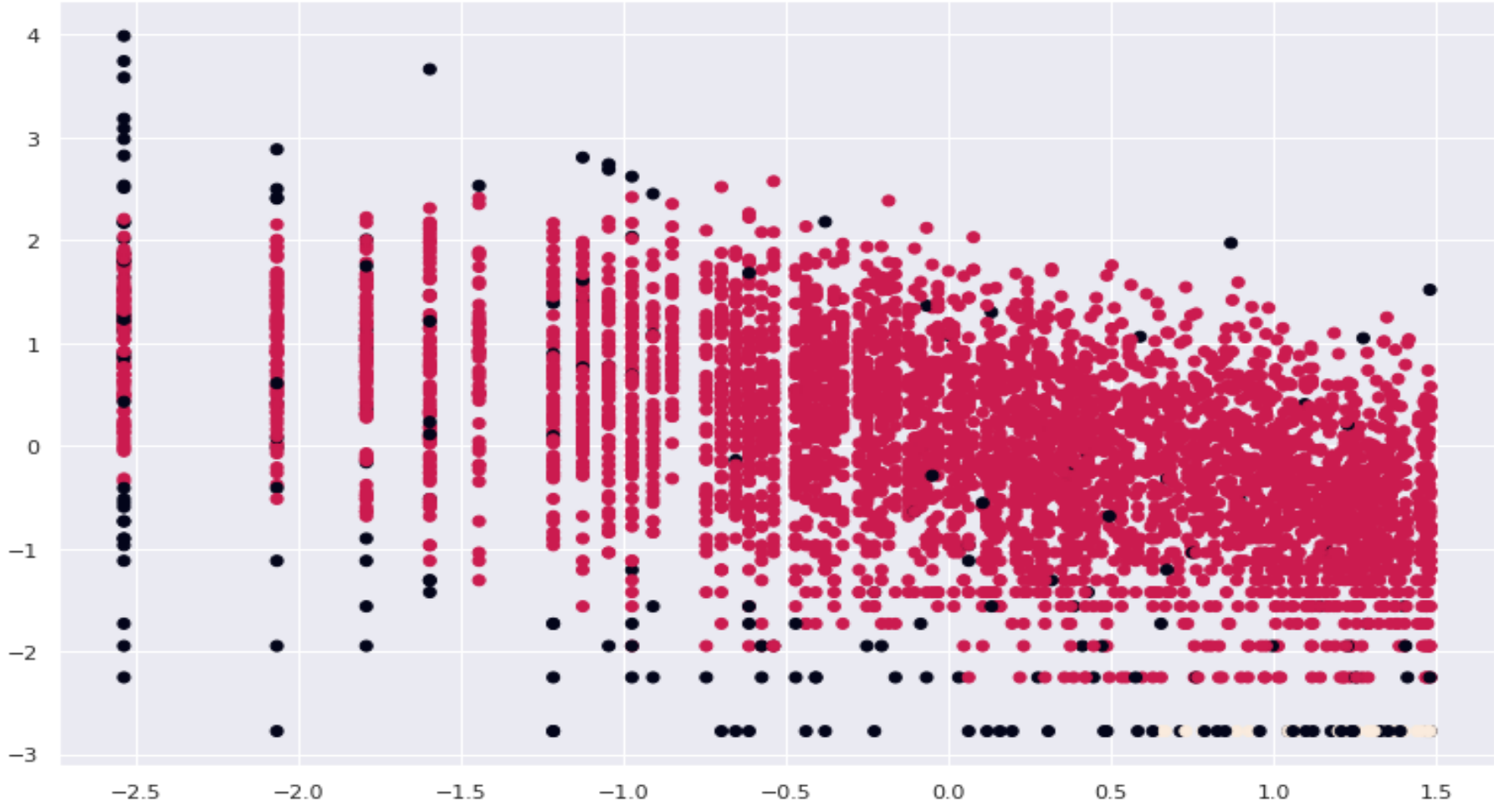


Elbow method and Cluster chart on RFM





DBSCAN TO RECENCY, FREQUENCY AND MONETARY



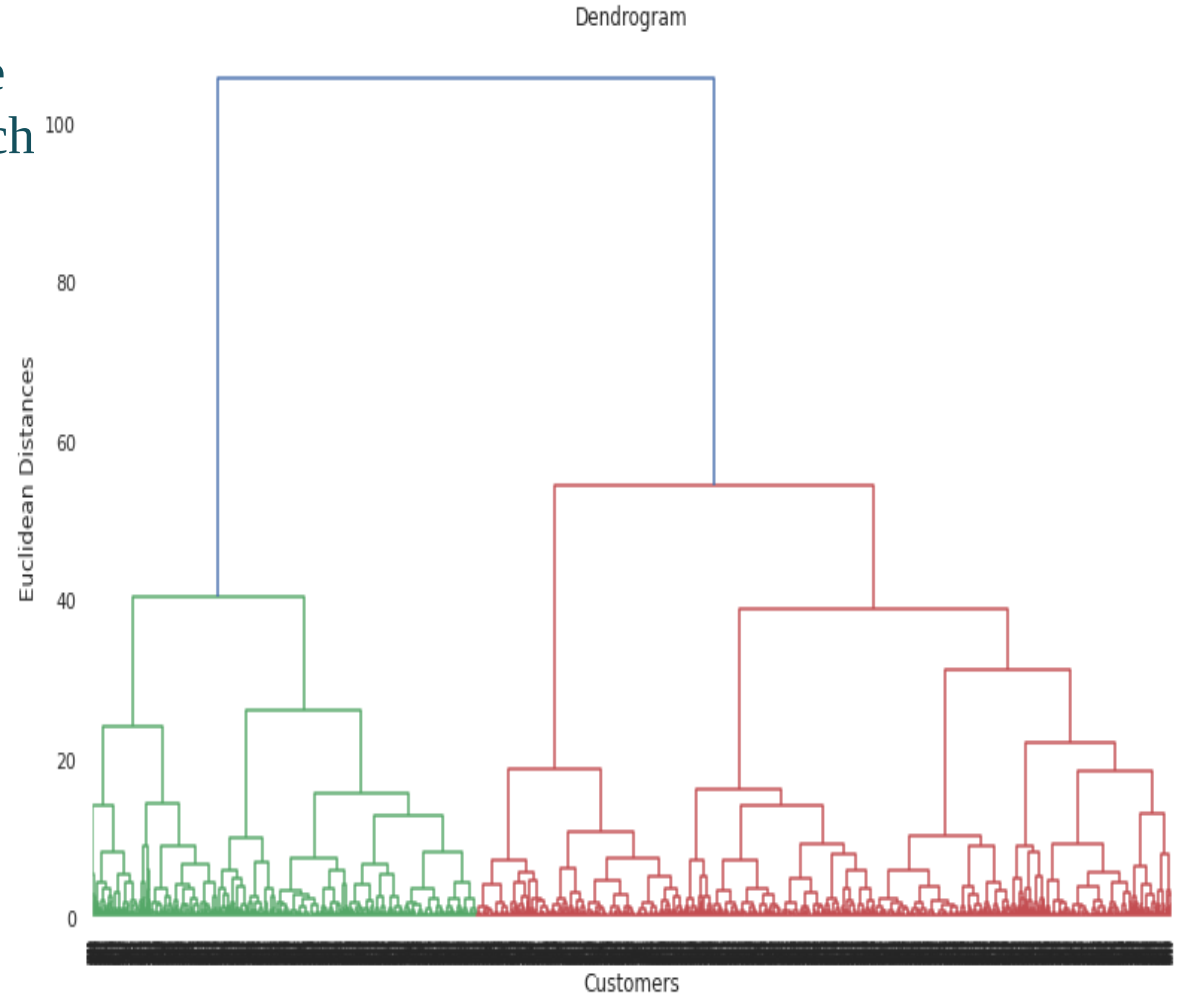
RFM Analysis

CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	Recency_log	Frequency_log	Monetary_log	Cluster
12346.0	325	1	77183.60	4	4	1	441	9	5.783825	0.000000	11.253942	1
12347.0	2	182	4310.00	1	1	1	111	3	0.693147	5.204007	8.368693	0
12348.0	75	31	1797.24	3	3	1	331	7	4.317488	3.433987	7.494007	1
12349.0	18	73	1757.55	2	2	1	221	5	2.890372	4.290459	7.471676	0
12350.0	310	17	334.40	4	4	3	443	11	5.736572	2.833213	5.812338	1
12352.0	36	85	2506.04	2	2	1	221	5	3.583519	4.442651	7.826459	0
12353.0	204	4	89.00	4	4	4	444	12	5.318120	1.386294	4.488636	1
12354.0	232	58	1079.40	4	2	2	422	8	5.446737	4.060443	6.984161	1
12355.0	214	13	459.40	4	4	3	443	11	5.365976	2.564949	6.129921	1
12356.0	22	59	2811.43	2	2	1	221	5	3.091042	4.077537	7.941449	0

HIERARCHICAL CLUSTERING

- The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold=90 degree.

- No. of Clusters = 2



Challenges

- Huge dataset
- Null values Treatment
- Treatment of cancelled orders
- Right number of ‘k’ for clusters

Conclusion

- This project mainly focused on developing customer segments for a UK based online store, selling unique all occasion gifts.

- This project is worked through 5 major steps, starting from data cleaning till cluster analysis.

- Using a recency, frequency and monetary (RFM) analysis, the customers have been segmented into various clusters.

- By applying different clustering algorithm to our dataset , we get the optimal number of cluster is equal to 2.

- The business can focus on these different clusters and provide to customers of each sector in a different way, which would not only benefit the customer but also the business at large.