Capstone Project - 4

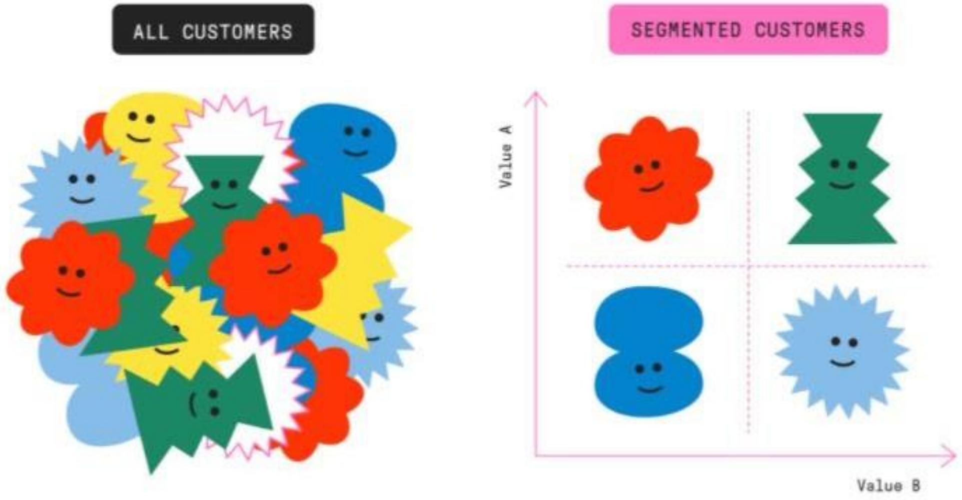
Online Retail Customer Segmentation

Unsupervised ML Model

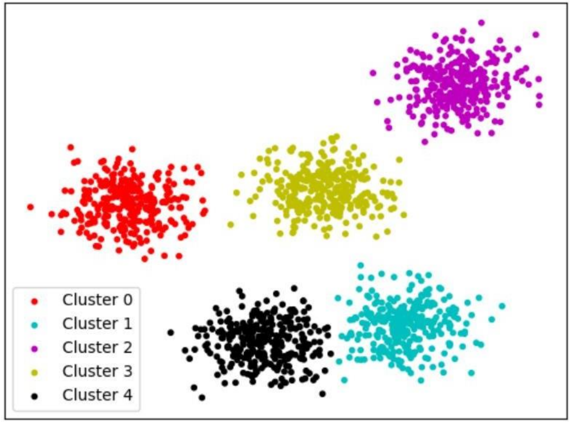
Vineeta Singh

Why Customer Segmentation?

“We needed a new way to understand our customers in a structured manner”



Introduction to Clustering



Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabelled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".

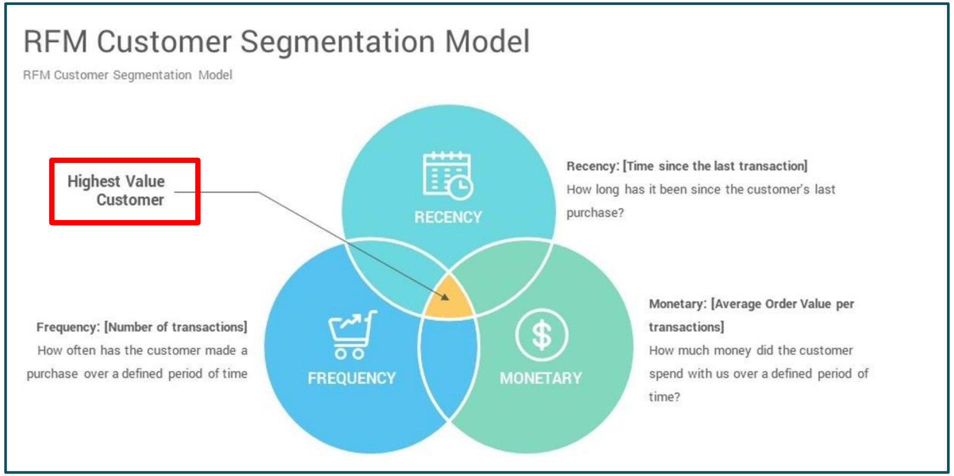
A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Content

- Problem Statement
- Data Summary
- EDA / Feature analysis
- Data preparation
- Create RFM model
- Implementing various clustering Models
- Challenges
- Conclusion

Problem statement

- This project aims to identify major customer segments on a transnational data set for a UK-based online retail.
- Create RFM table
- We need to analyse and identify major customer segmentation using k means algorithm and also different algorithms to confirm our result.



Data Summary

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

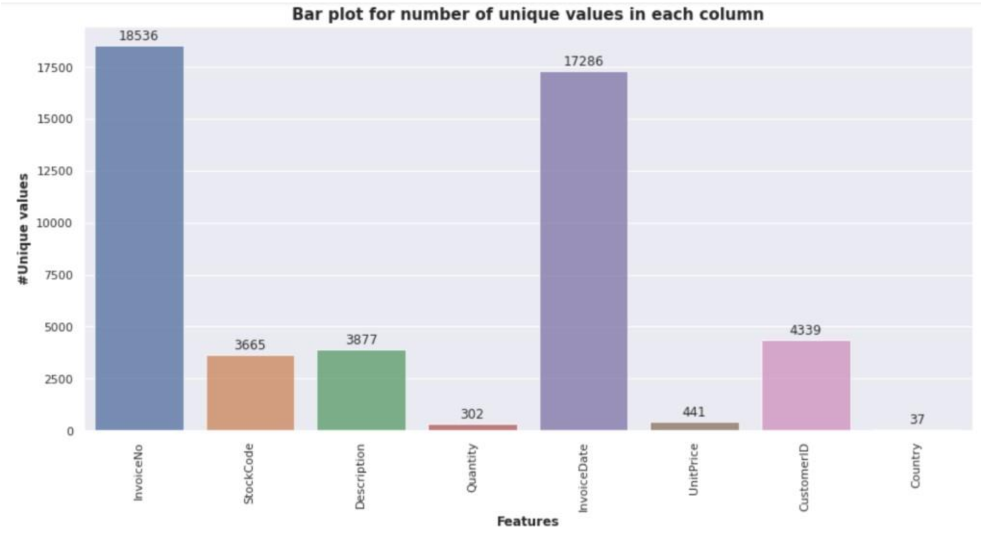
Basic Data Exploration

- A transnational data set with transactions occurring between 1st December 2010 and 9th December 2011 for a UK-based online retailer.
- Dataset has rows- 541909 & columns-8.
- The company mainly sells unique all-occasion gifts.
- Many customers of the company are wholesalers

```
# First look
df.head()
```

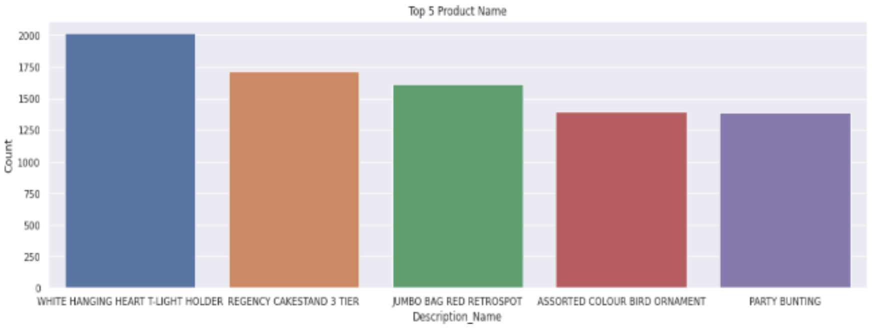
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS 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

EDA - Which feature has the highest number of unique values?



The invoice number is unique for every transaction. Invoice Date has second highest count.

EDA – Finding Top product based on maximum selling

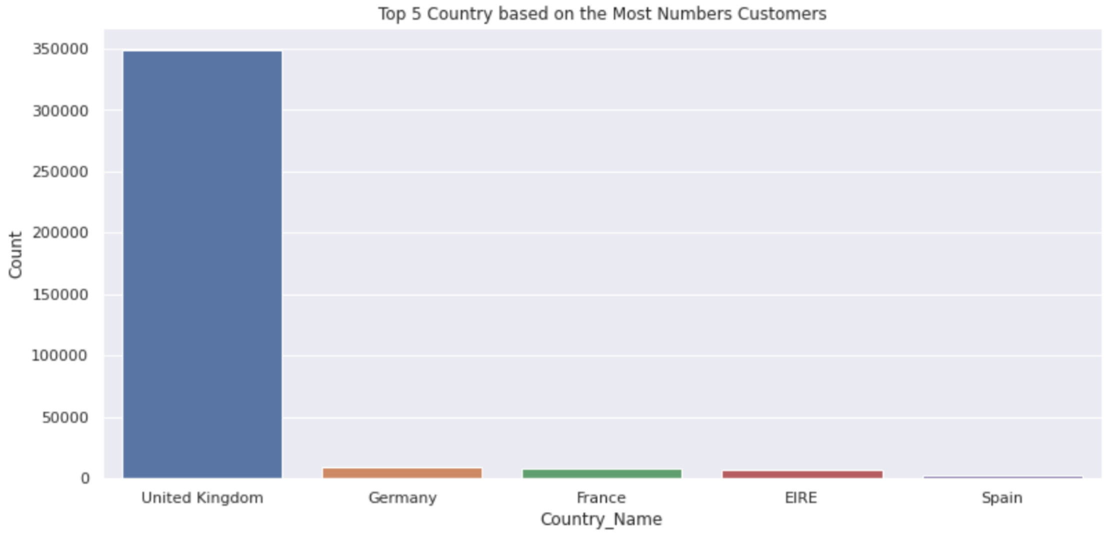


Top product based on maximum selling are :

```
1.WHITE HANGING HEART T-LIGHT HOLDER,
2.REGENCY CAKESTAND 3 TIER
3.JUMBO BAG RED RETROSPOT
4.PARTY BUNTING
5.LUNCH BAG RED RETROSPOT
```

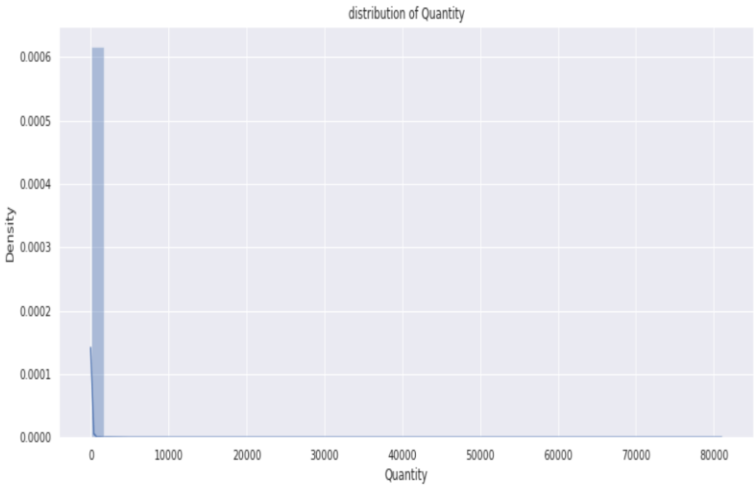
White Hanging Heart T- Light Holder is the top product.

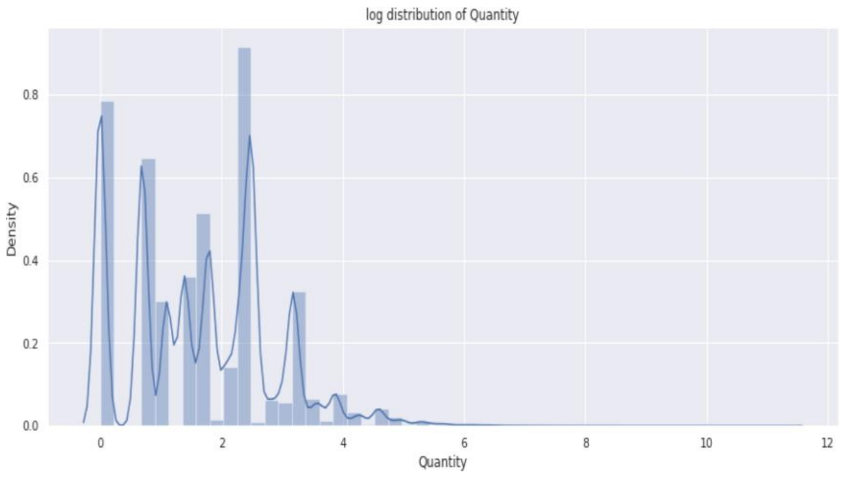
EDA - Top 5 Country based on the Most Numbers Customers?



In This graph, we can observe that most purchases are from the United Kingdom. It is justifiable also, as this is UK's company.

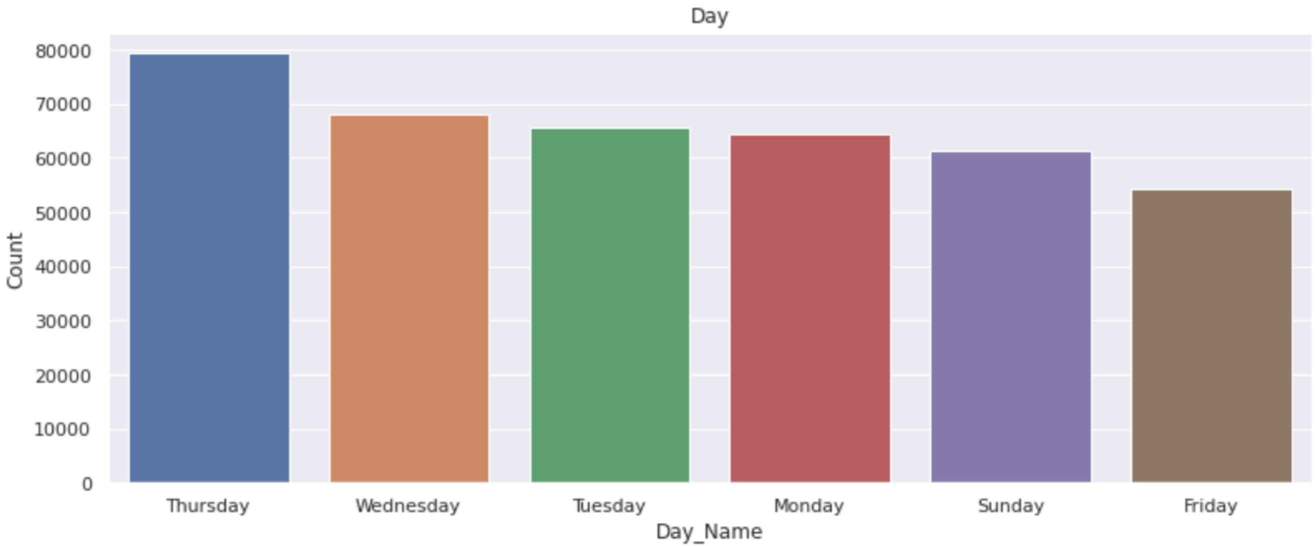
Log transformation of quantity





For better accuracy, applied log distribution on quantity.

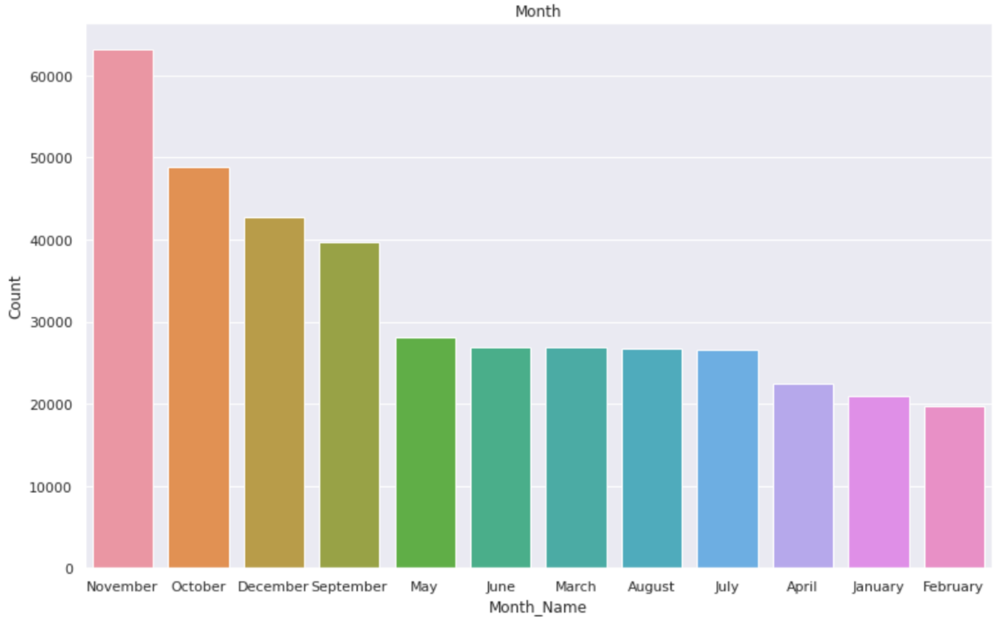
EDA- which day has highest count?



	Day_Name	Count
0	Thursday	79243
1	Wednesday	68040
2	Tuesday	65744
3	Monday	64231
4	Sunday	61212
5	Friday	54222

Most of the customers have purchased items on Thursday, Wednesday, Tuesday.

Which month has the highest count ?



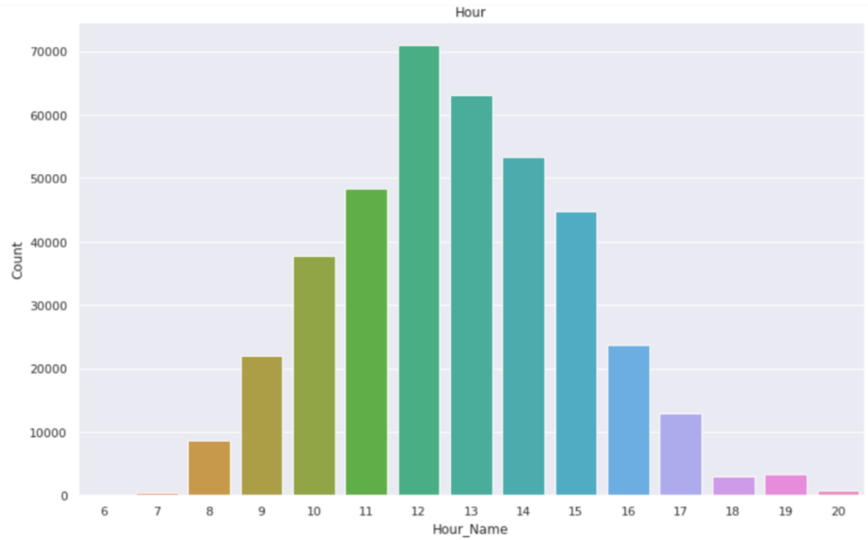
	Month_Name	Count
0	November	63168
1	October	48793
2	December	42696
3	September	39669
4	May	28073
5	June	26926
6	March	26870
7	August	26790
8	July	26580
9	April	22433
10	January	20988
11	February	19706

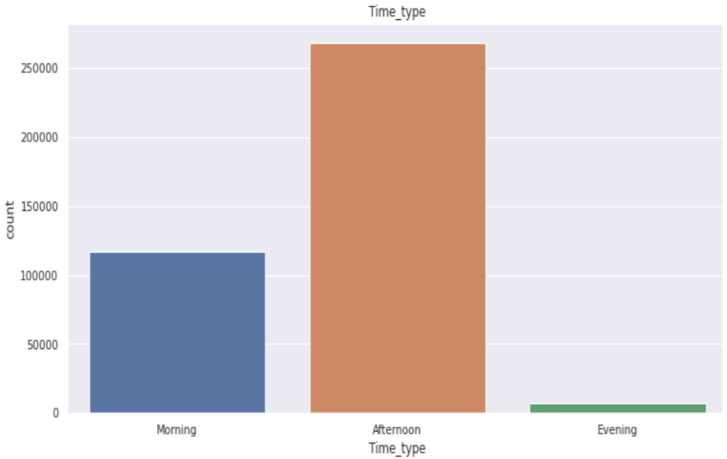
Most of the customers have purchased items in November, October, December, and the least number of purchases in April, January, February.

Hour wise Analysis

We have divided hours of the day into 3-time types.

1. Morning
2. Afternoon
3. Evening



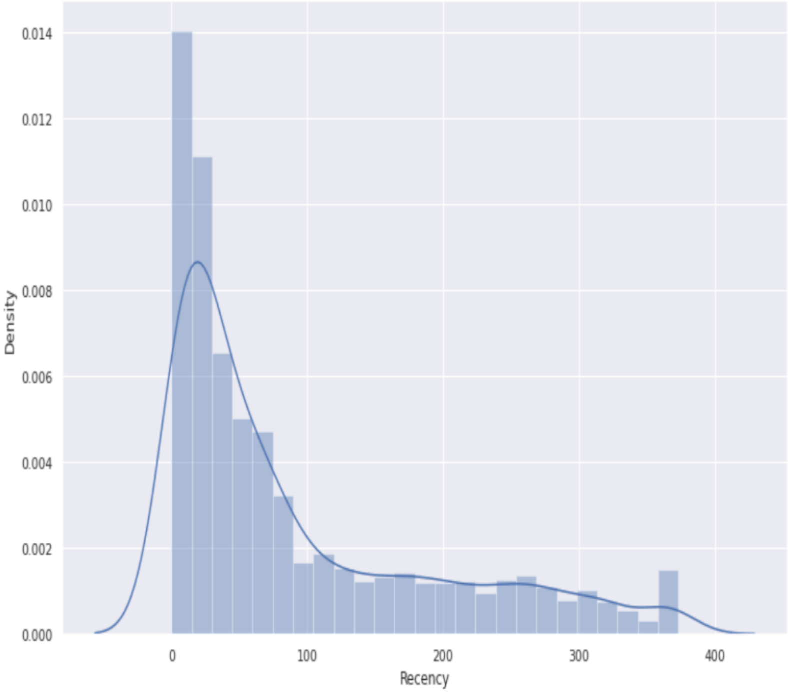


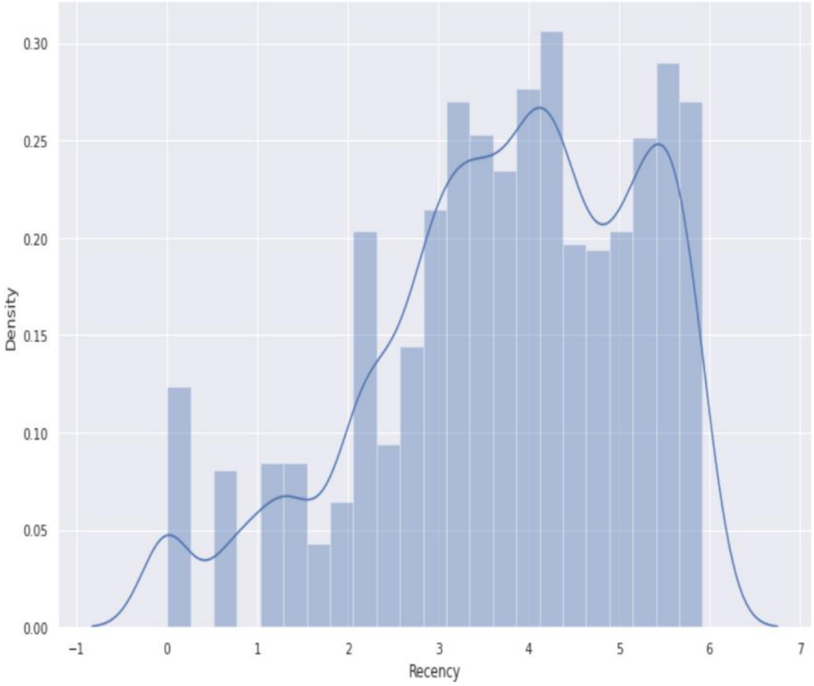
Most of the customers purchase in the afternoon time. The 12th hour of the day is a peak for purchasing items. Moderate numbers of customers have purchased the items in the Morning and the least numbers of customers have purchased the items in the Evening.

Create the RFM model (Recency, Frequency, Monetary value)

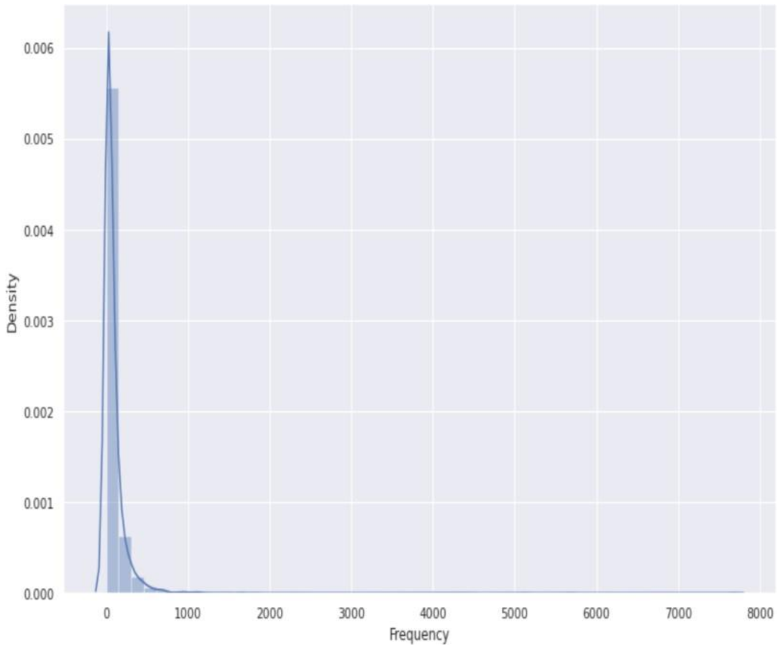
CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore
12346.0	325	1	77183.60	4	4	1	441	9
12347.0	2	182	4310.00	1	1	1	111	3
12348.0	75	31	1797.24	3	3	1	331	7
12349.0	18	73	1757.55	2	2	1	221	5
12350.0	310	17	334.40	4	4	3	443	11

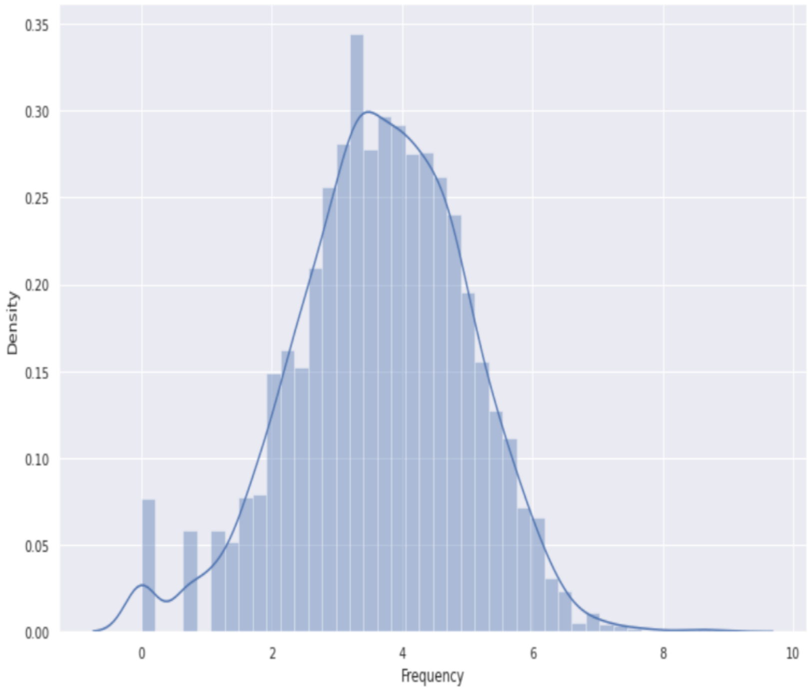
Log Transformation of Recency



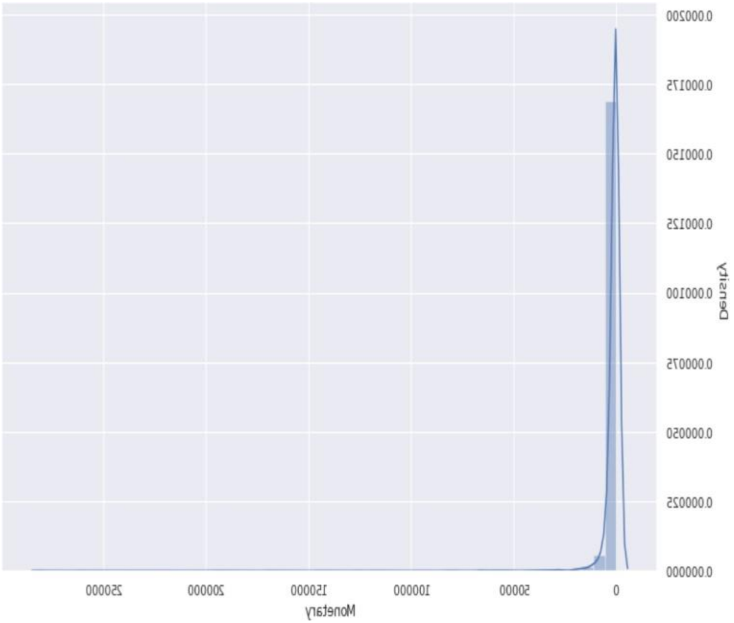


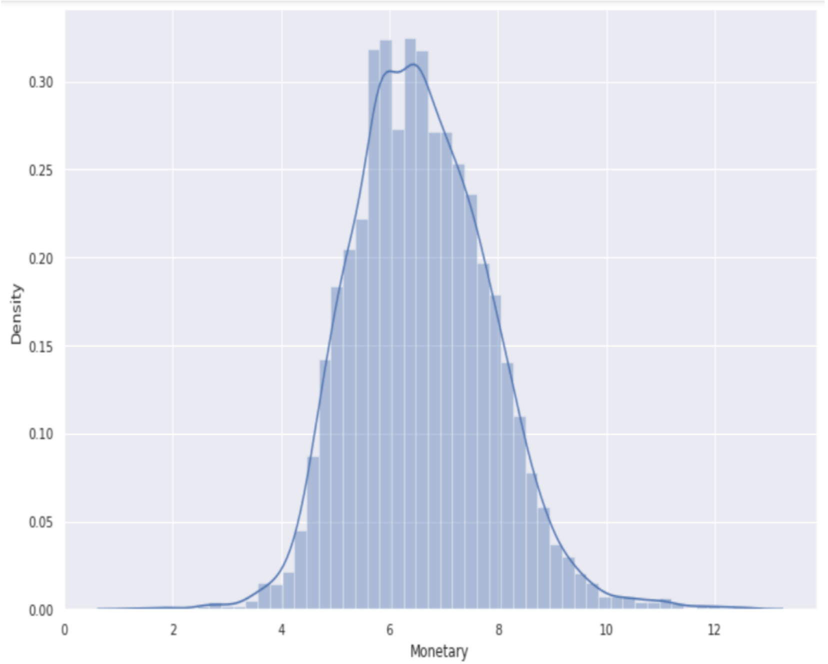
Log Transformation of Frequency





Log Transformation of Monetary





Model Overview

Let's get some insight about Clustering models:

Silhouette score method : It is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. It ranges from -1 to1 , where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

Elbow method : a point from where the value of clusters starts decreasing suddenly, indicates the optimal number of clusters.

DBSCAN (Density Based Spatial Clustering of Application with Noise) :
Finds core samples of high density and expands clusters from them.

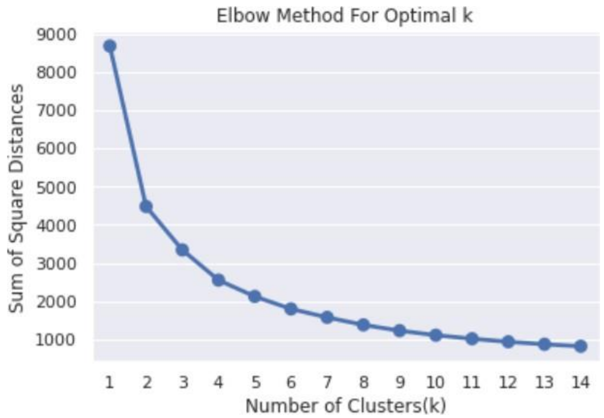
Dendrogram : It is representation of hierarchical clustering.

Model Overview

- K-Means with silhouette score for RM
- K-Means with Elbow method FOR RM
- DBSCAN for RM
- K-Means with silhouette score for FM
- K-Means with Elbow method for FM
- DBSCAN for FM
- K-Means with silhouette score for RFM
- K-Means with Elbow method for RFM
- Hierarchical clustering for RFM
- DBSCAN for RFM

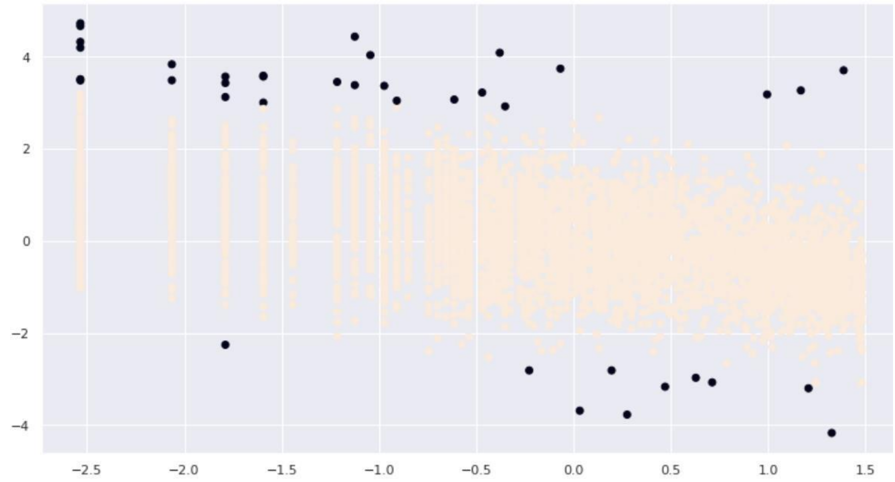
Applying Silhouette Score and Elbow Method on Recency and Monetary

```
For n_clusters = 2, silhouette score is 0.42071509151962466
For n_clusters = 3, silhouette score is 0.3431112622037282З
For n_clusters = 4, silhouette score is 0.3650830552007133
For n_clusters = 5, silhouette score is 0.3348798355166678
For n_clusters = 6, silhouette score is 0.3446603914792049
For n_clusters = 7, silhouette score is 0.3478025808437424
For n_clusters = 8, silhouette score is 0.33801799623366263
For n_clusters = 9, silhouette score is 0.3457774819953329
For n_clusters = 10, silhouette score is 0.3476181905063959
For n_clusters = 11, silhouette score is 0.338004977551094
For n_clusters = 12, silhouette score is 0.34239046089095004
For n_clusters = 13, silhouette score is 0.3421565330406368
For n_clusters = 14, silhouette score is 0.3362226361846414
For n_clusters = 15, silhouette score is 0.33678077415427365
```



We can see that, Customers are well separated when we cluster them by Recency and Monetary.

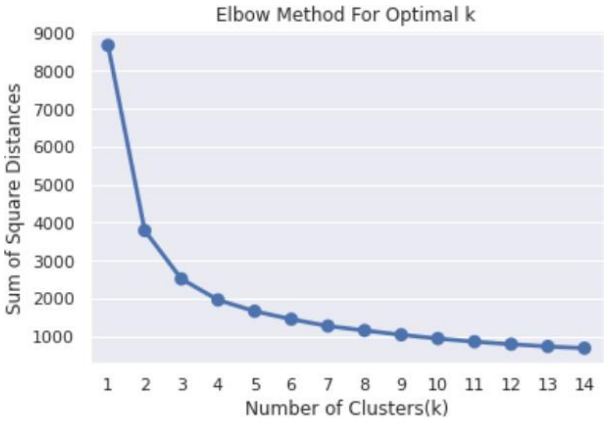
Applying DBSCAN on Recency and Monetary



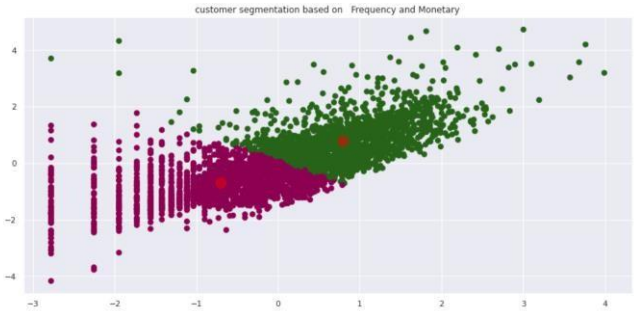
From above plot, we can observe that Customers are well separate when we cluster them by Recency and Monetary. We got 2 as optimal number of clusters.

Applying silhouette Score and Elbow Method on Frequency and Monetary

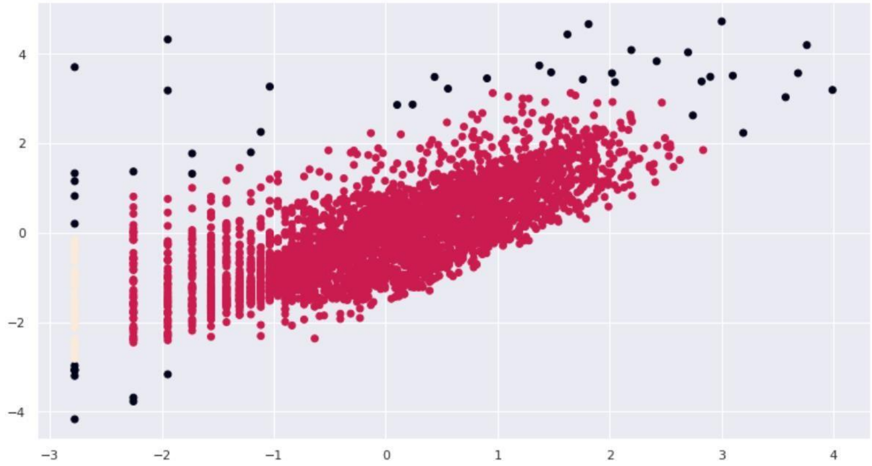
```
For n_clusters = 2, silhouette score is 0.47868810164073394
For n_clusters = 3, silhouette score is 0.40765717862922224
For n_clusters = 4, silhouette score is 0.37225458014072377
For n_clusters = 5, silhouette score is 0.3467764608315663
For n_clusters = 6, silhouette score is 0.36238573376083216
For n_clusters = 7, silhouette score is 0.3446585674869972
For n_clusters = 8, silhouette score is 0.35229572577506524
For n_clusters = 9, silhouette score is 0.3448452433500612
For n_clusters = 10, silhouette score is 0.3589804874619343
For n_clusters = 11, silhouette score is 0.3684087132646038
For n_clusters = 12, silhouette score is 0.35405880230027
For n_clusters = 13, silhouette score is 0.36351442085428287
For n_clusters = 14, silhouette score is 0.35758398562740784
For n_clusters = 15, silhouette score is 0.3444696702423664
```



From this plot, We found Customers are well separated when we cluster them by Frequency and Monetary.



Applying DBSCAN on Frequency and Monetary



We can see that Customers are well separated when we cluster them by Frequency and Monetary. We got 2 as optimal number of clusters.

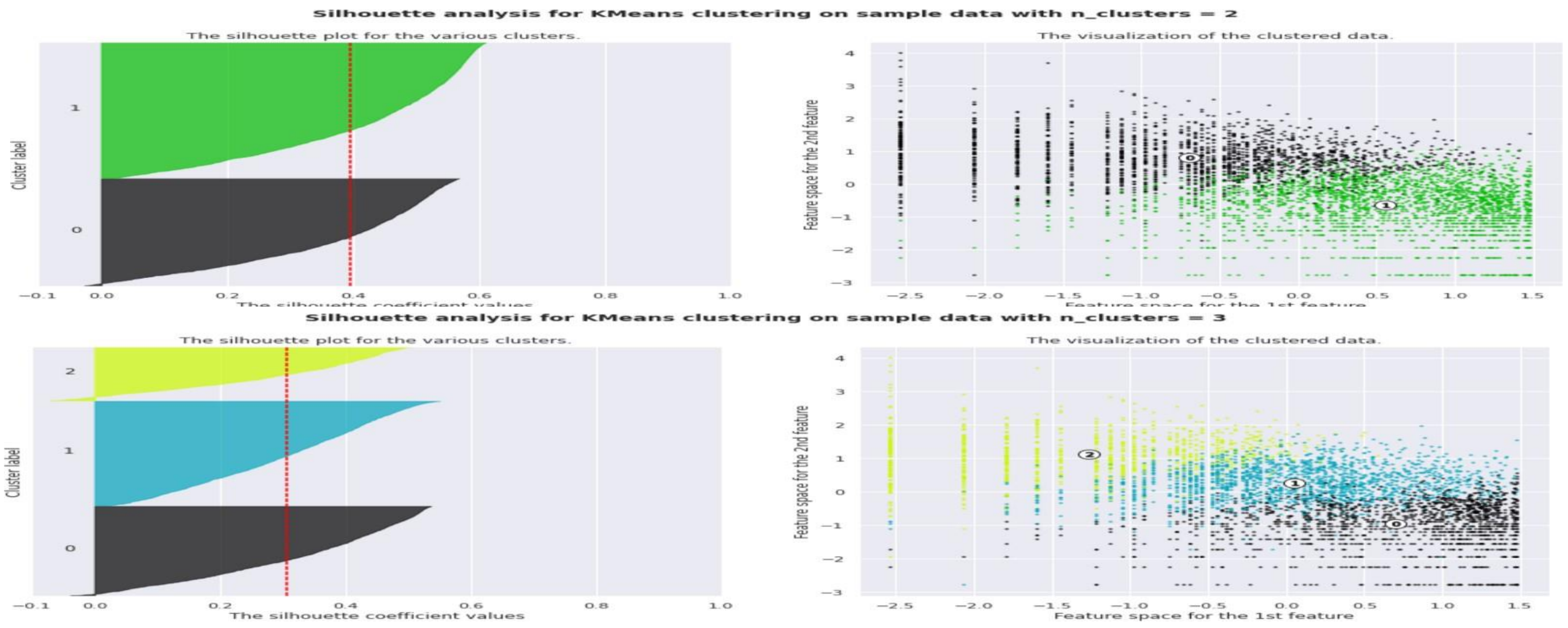
Applying Silhouette Method on Recency, Frequency and Monetary

```
For n_clusters = 2 The average silhouette_score is : 0.3955943249451756
For n_clusters = 3 The average silhouette_score is : 0.3058190516961747
For n_clusters = 4 The average silhouette_score is : 0.3005812873803695
For n_clusters = 5 The average silhouette_score is : 0.279264977284325
For n_clusters = 6 The average silhouette_score is : 0.2791466583409964
For n_clusters = 7 The average silhouette_score is : 0.268196906247297
For n_clusters = 8 The average silhouette_score is : 0.263748148701171
For n_clusters = 9 The average silhouette_score is : 0.2601971253281270
For n_clusters = 10 The average silhouette_score is : 0.2591702086507785
For n_clusters = 11 The average silhouette_score is : 0.2560536365948069
For n_clusters = 12 The average silhouette_score is : 0.261890522577597
For n_clusters = 13 The average silhouette_score is : 0.2629370975986603
For n_clusters = 14 The average silhouette_score is : 0.262235684437257
For n_clusters = 15 The average silhouette_score is : 0.2582846446990525
```

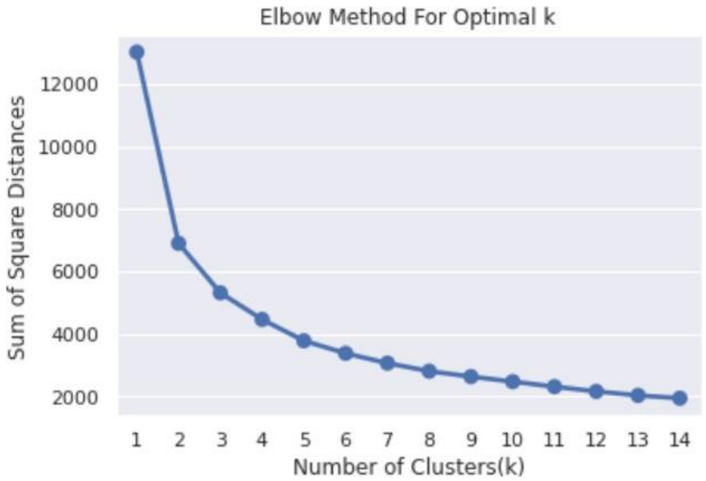
Applying Silhouette Method on Recency, Frequency and Monetary

No. of cluster = 2

No. of cluster = 3



Applying Elbow Method on Recency, Frequency and Monetary

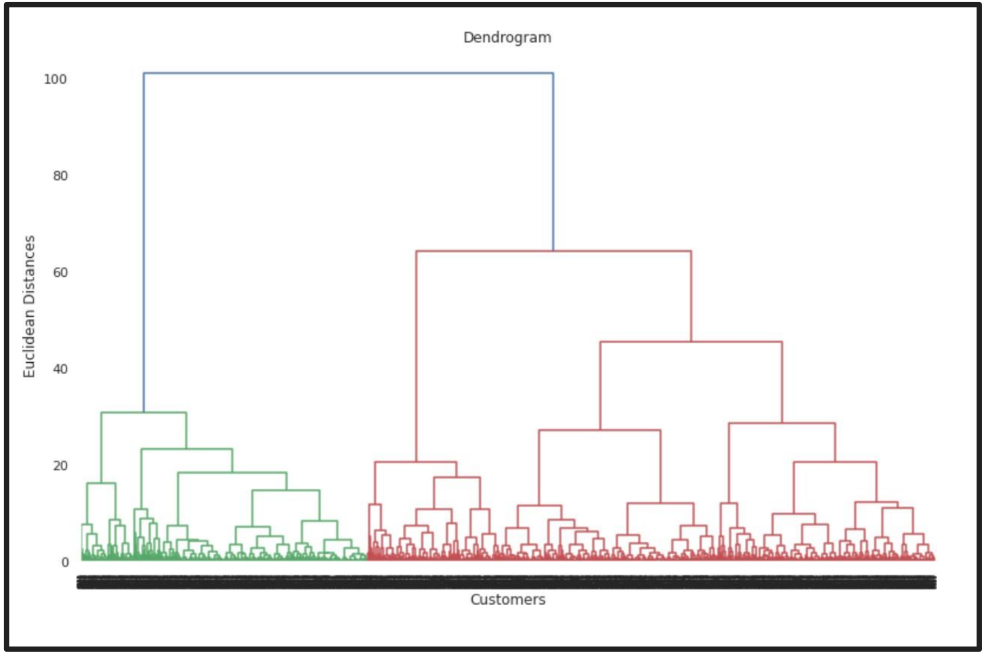


This method also gives information about the optimal number of clusters. According to it, 2 is the optimal number of clusters.

Using the Dendrogram to find the optimal number of clusters

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

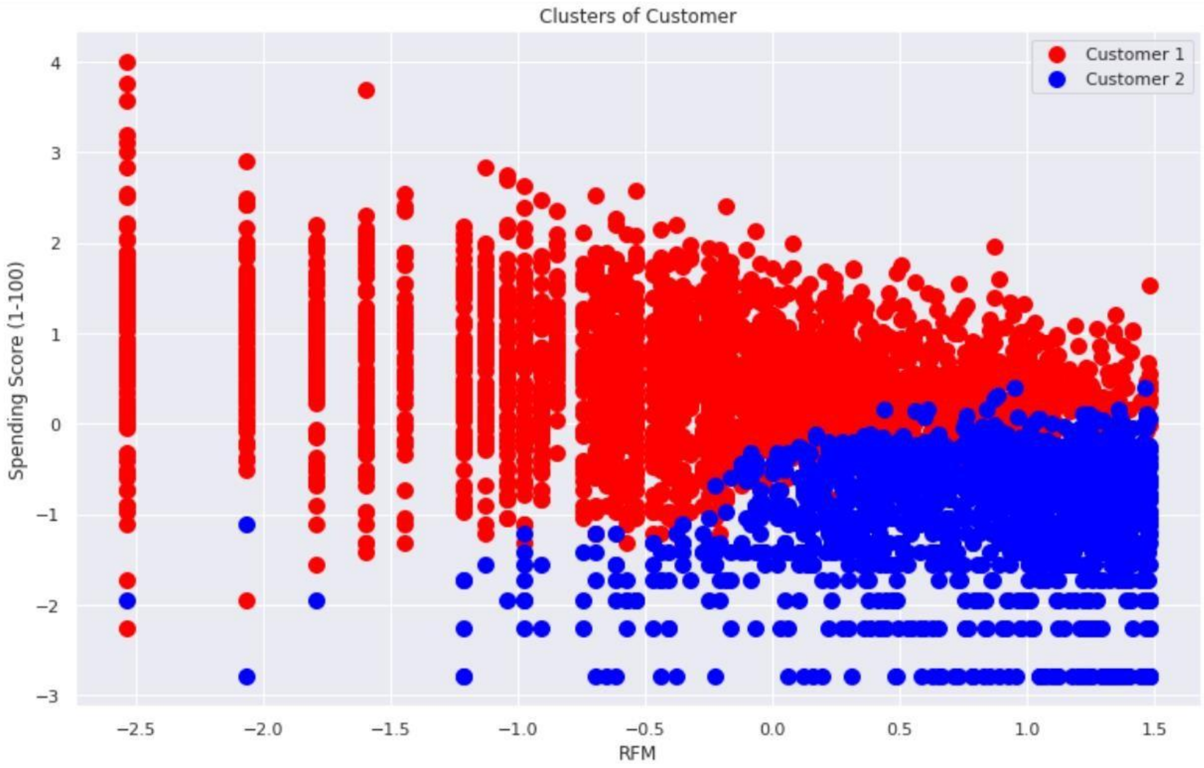
No. of Cluster = 2



Find the clusters on the basis on RFM table

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	0
12347.0	2	182	4310.00	1	1	1	111	3	0.693147	5.204007	8.368693	1
12348.0	75	31	1797.24	3	3	1	331	7	4.317488	3.433987	7.494007	0
12349.0	18	73	1757.55	2	2	1	221	5	2.890372	4.290459	7.471676	1
12350.0	310	17	334.40	4	4	3	443	11	5.736572	2.833213	5.812338	0
12352.0	36	85	2506.04	2	2	1	221	5	3.583519	4.442651	7.826459	1
12353.0	204	4	89.00	4	4	4	444	12	5.318120	1.386294	4.488636	0
12354.0	232	58	1079.40	4	2	2	422	8	5.446737	4.060443	6.984161	0
12355.0	214	13	459.40	4	4	3	443	11	5.365976	2.564949	6.129921	0
12356.0	22	59	2811.43	2	2	1	221	5	3.091042	4.077537	7.941449	1

Visualizing the clusters (two dimensions only)



Challenges

- Understanding the problem statement.
- Figuring Out right Approach
- Dealing with Null And duplicate values
- Treatment of cancelled orders
- Extracting Datetime Column Properly and creating RFM variables.
- Designing multiple visualizations to summarize the Data points in the dataset and effectively communicating the results and insights to the reader.
- Finding optimal number of clusters

Conclusion

Descriptive Analytics

In conclusion, the data exploration of Online customer segmentation dataset shows :

- Missing and duplicate values were found.
- Most of the purchases are from the United Kingdom.
- Most of the customers have purchased items on Thursday, Wednesday, Tuesday.
- Most of the customers have purchased items in November, October, December, and the least number of purchases in April, January, February.
- Most of the customers purchase in the afternoon time. The 12th hour of the day is a peak for purchasing items.

Conclusion

SL No.	Model_Name	Data	Optimal_Number_of_cluster
1	K-Means with silhouette_score	RM	2
2	K-Means with Elbow method	RM	2
3	DBSCAN	RM	2
4	K-Means with silhouette_score	FM	2
5	K-Means with Elbow method	FM	2
6	DBSCAN	FM	2
7	K-Means with silhouette_score	RFM	2
8	K-Means with Elbow method	RFM	2
9	Hierarchical clustering	RFM	2
10	DBSCAN	RFM	3

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

Final Thought

Customer segmentation is an important marketing approach that businesses should employ in order to gain a better understanding of the market and make more informed decisions in order to increase sales.

K-Means clustering is a basic but effective machine learning algorithm that businesses can use. Finally, in order to optimise our marketing success, we must keep the RFM client segmentation up to date.

Further analysis

- New variables have been added, such as tenure, which is the number of days since each customer's first transaction. This will reveal how long each customer has been a member of the system.

- Customers are being segmented more deeply based on their physical location, as well as demographic and psychographic factors.

- Incorporating data from the company's Google Analytics account. Google Analytics is an excellent tool for tracking a variety of essential business data, including Customer Lifetime Value, Traffic Source/Medium, Pageviews per Visit, and Bounce Rate of a company's website, among others.

Q & A