Capstone Project

Online Retail Customer Segmentation

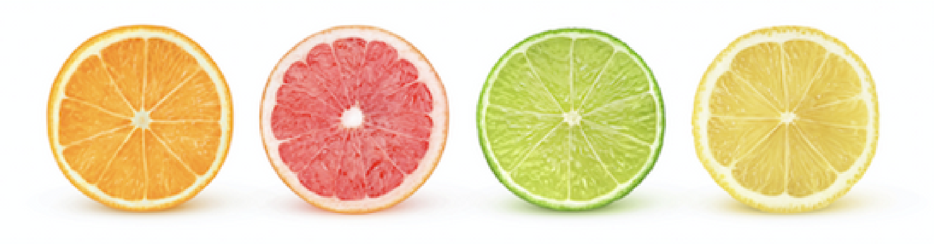
Individual contributor
Mohammed Javeed

Data Summary

- 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.

Why do customers need to be segmented?

In the context of marketing, the process of dividing customers into groups of similar individuals based on certain characteristics, such as age, gender, interests, and spending patterns.



CUSTOMER SEGMENTATION

It helps us focus on our customers more efficiently and improve their customer experience by providing a better understanding of their needs.

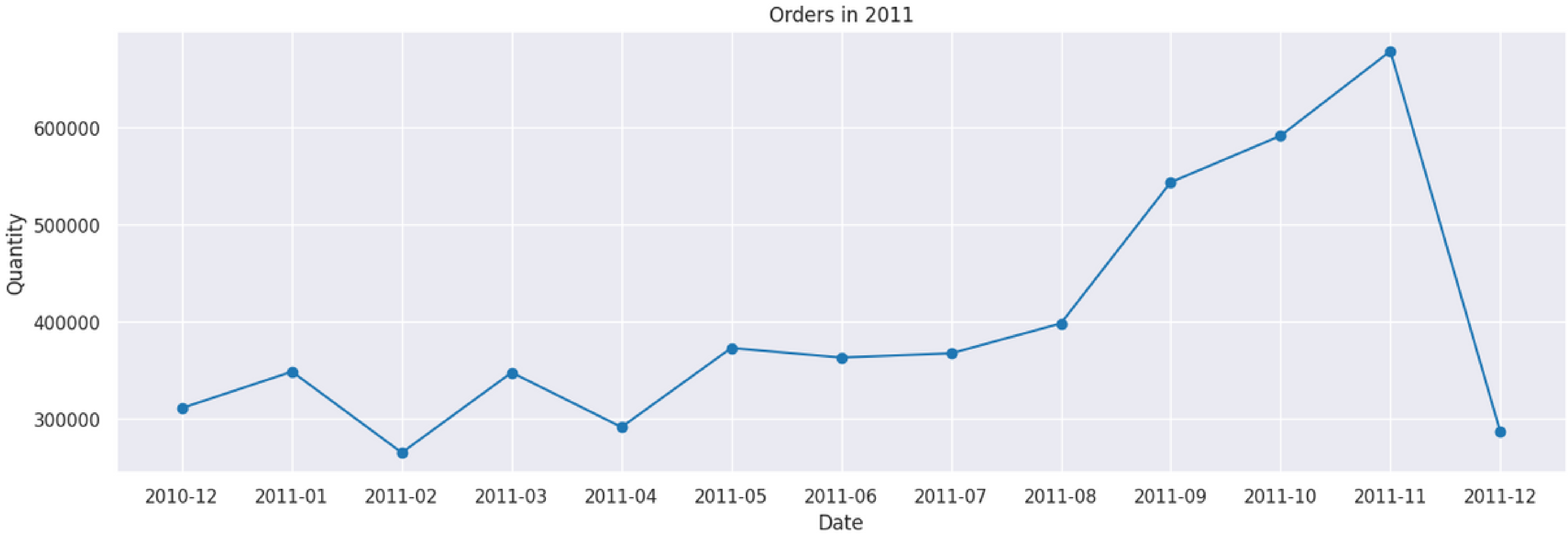
Problem Statement

The goal of this study is to analyze data collected from an online retailer based in the UK.

We will use the K Means algorithm to analyze the data and identify major customer segments. We will also use different verification methods to verify the results that are obtained through this analysis.

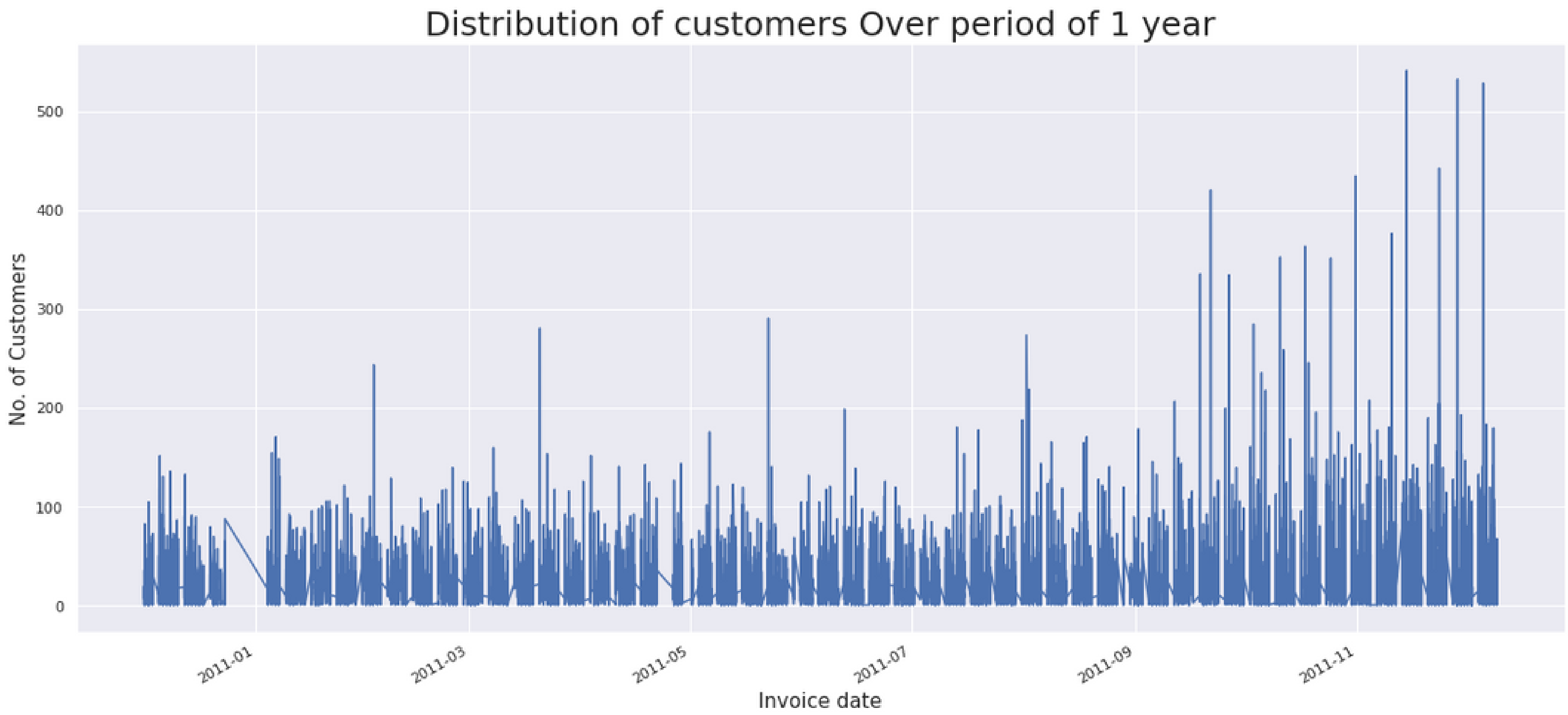
Exploratory Data Analysis .

How many product sold every month?



Over the past year, November has sold the most products, accounting for 13,41% of total sales. As a result, the business team can increase sales this month by promoting the updated products to the target market.

Annual sales of products



It can be easily concluded from the above graph that number of customers are increasing as we reaching towards the end of the year 2011.

September and November are getting highest purchasing order in comparison to January and March.

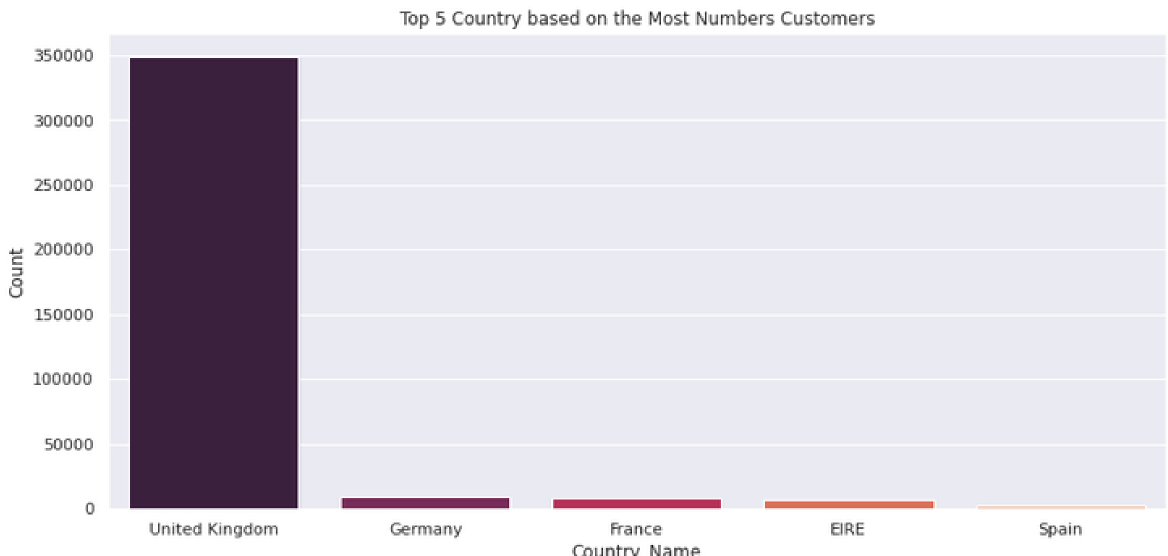
Finding the most Purchased Products

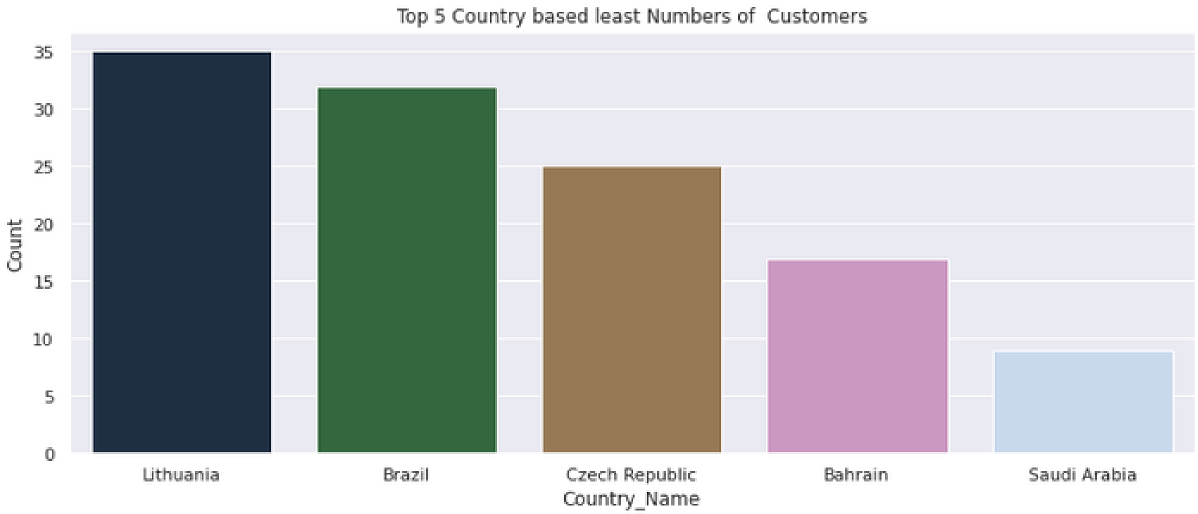
	Description_Name	Count
0	WHITE HANGING HEART T-LIGHT HOLDER	2016
1	REGENCY CAKESTAND 3 TIER	1714
2	JUMBO BAG RED RETROSPOT	1615
3	ASSORTED COLOUR BIRD ORNAMENT	1395
4	PARTY BUNTING	1390

Top 5 vs Bottom 5 countries

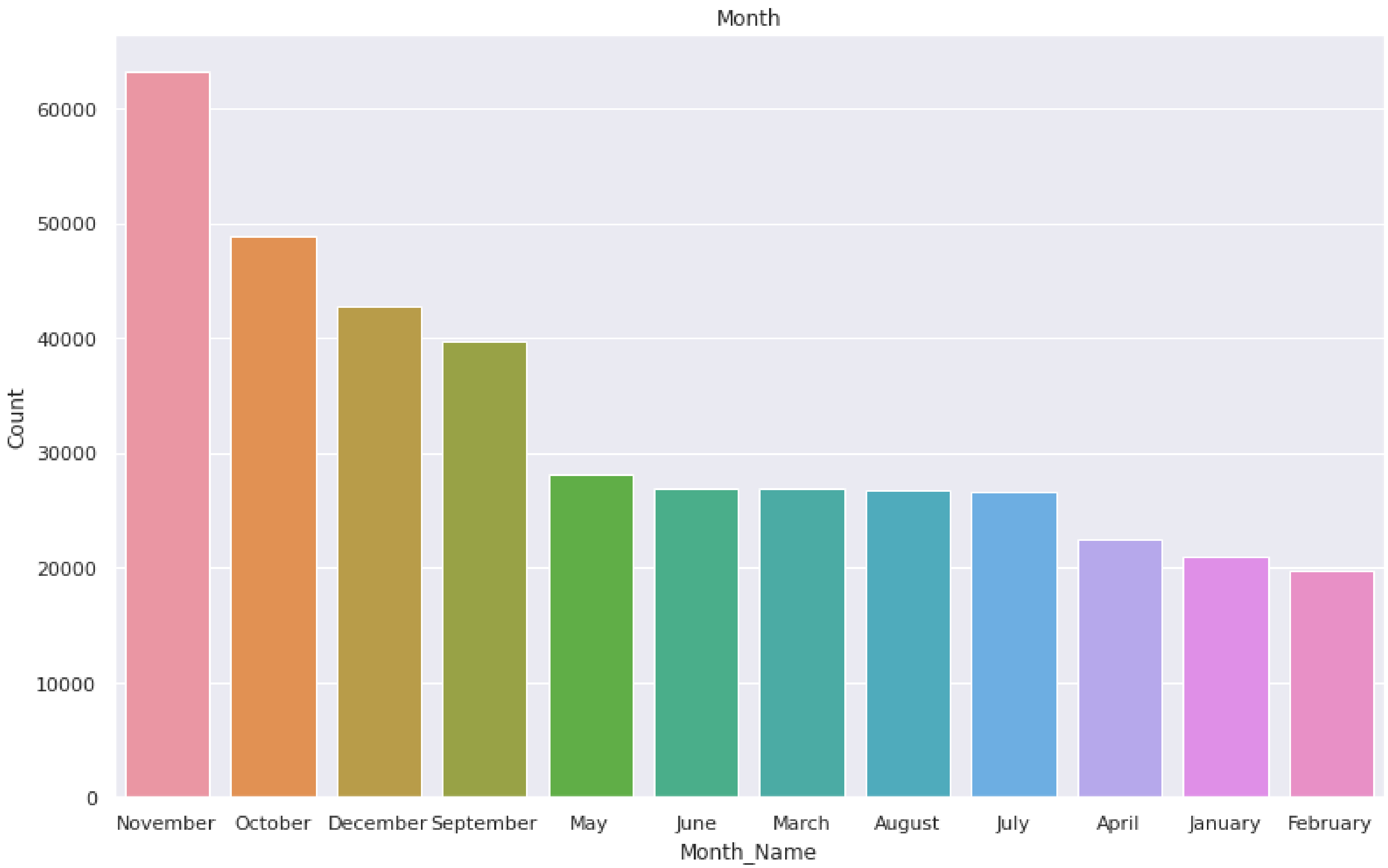
	Country_Name	Count
0	United Kingdom	349227
1	Germany	9027
2	France	8327
3	EIRE	7228
4	Spain	2480

	Country_Name	Count
32	Lithuania	35
33	Brazil	32
34	Czech Republic	25
35	Bahrain	17
36	Saudi Arabia	9



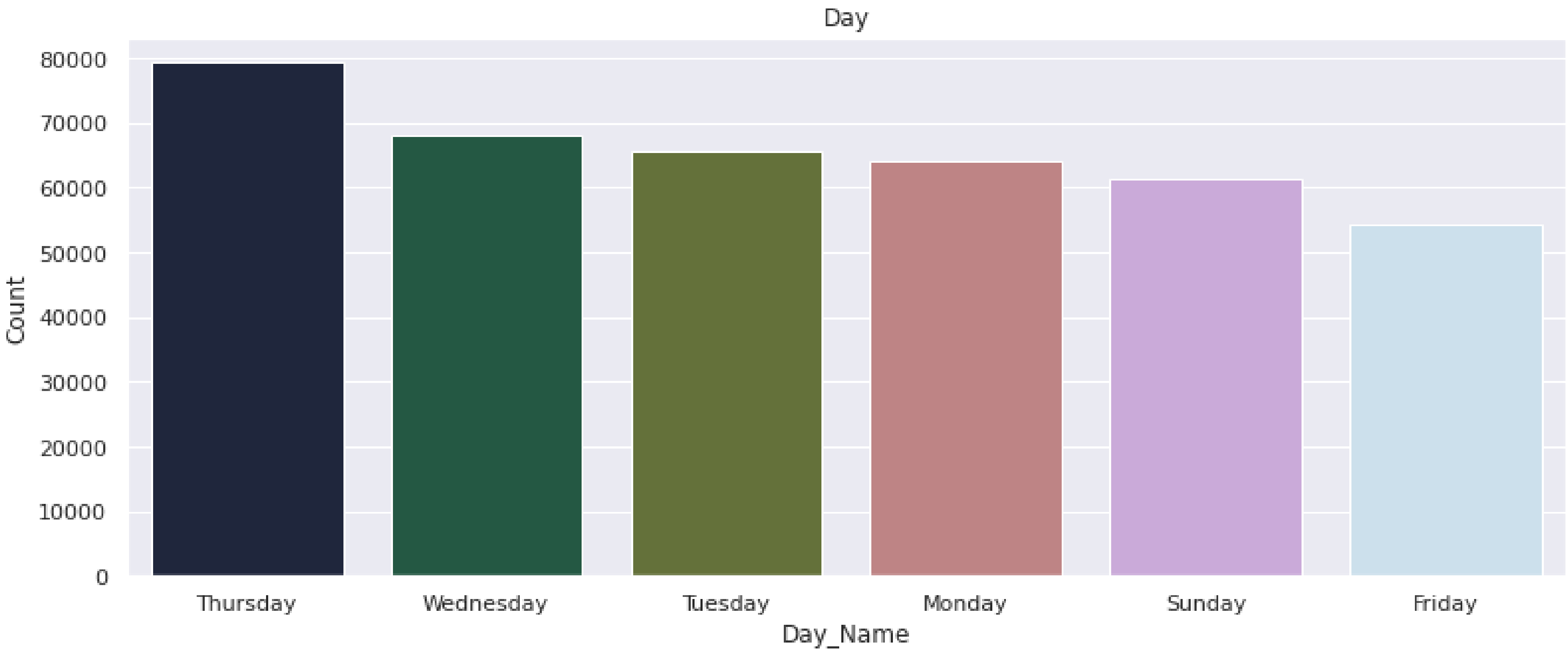


An analysis of each month



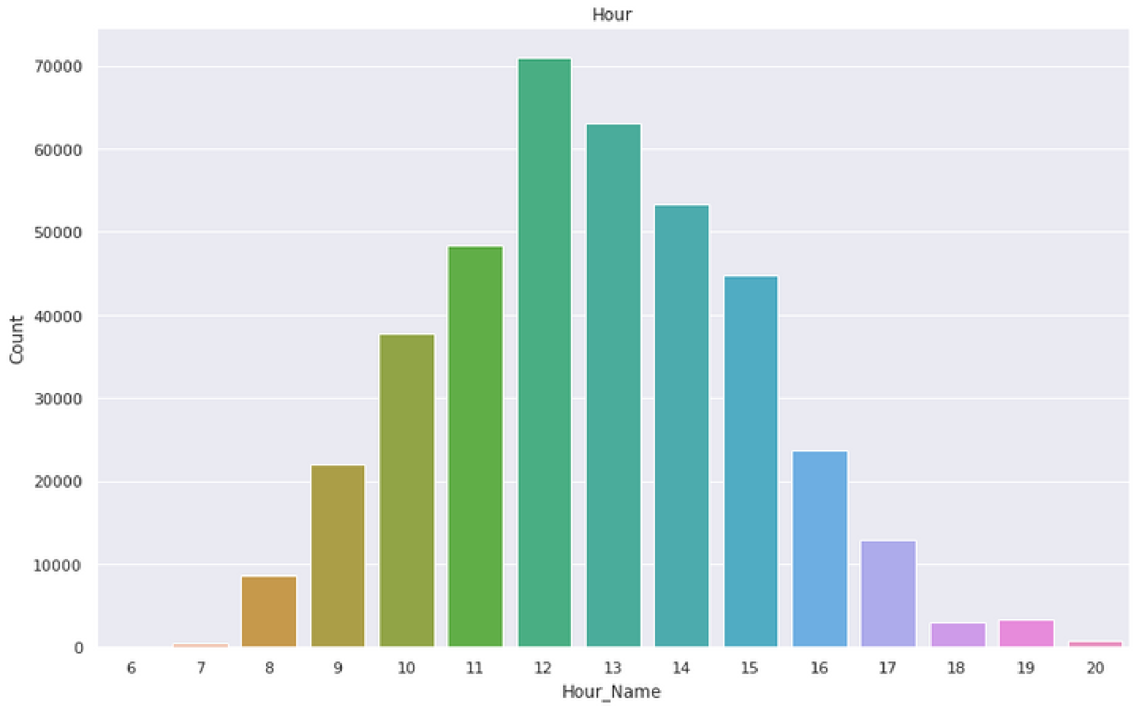
November and December could be the months with highest sales in anticipation of Christmas

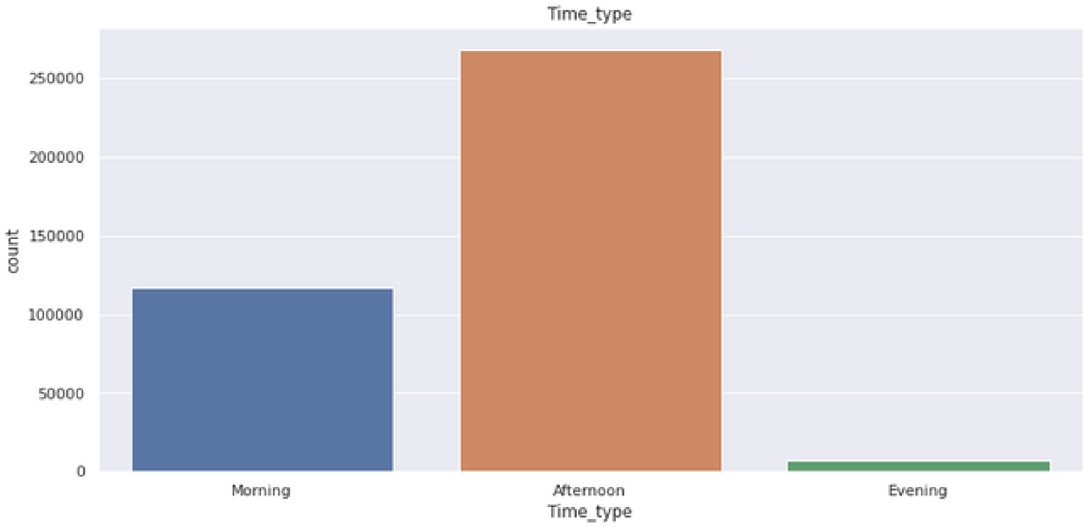
Daywise analysis



In terms of buying activity, Thursday is the busiest day, followed by Wednesday and Tuesday

Hourly analysis



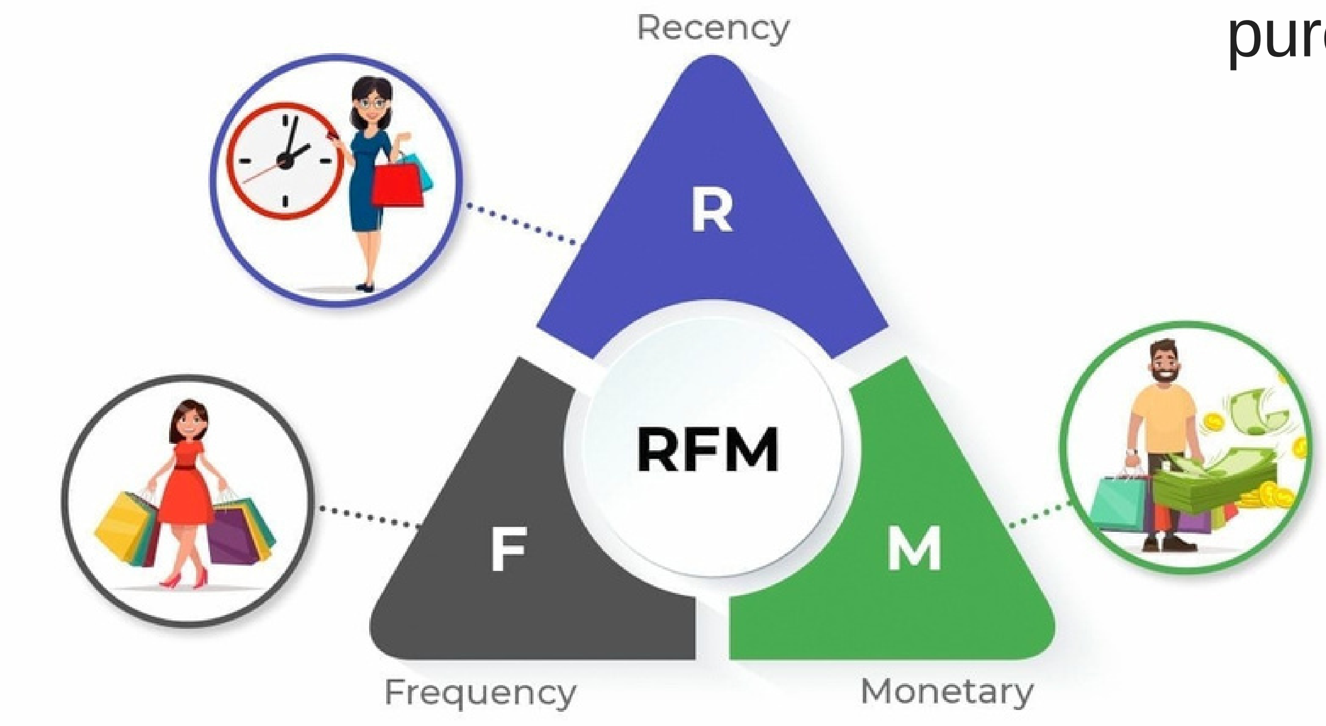


A large portion of the dataset contains wholesalers' data, which could explain the high sales during working hours

Modeling Data: RFM Quantiles

Recency Frequency Monetary(RFM)

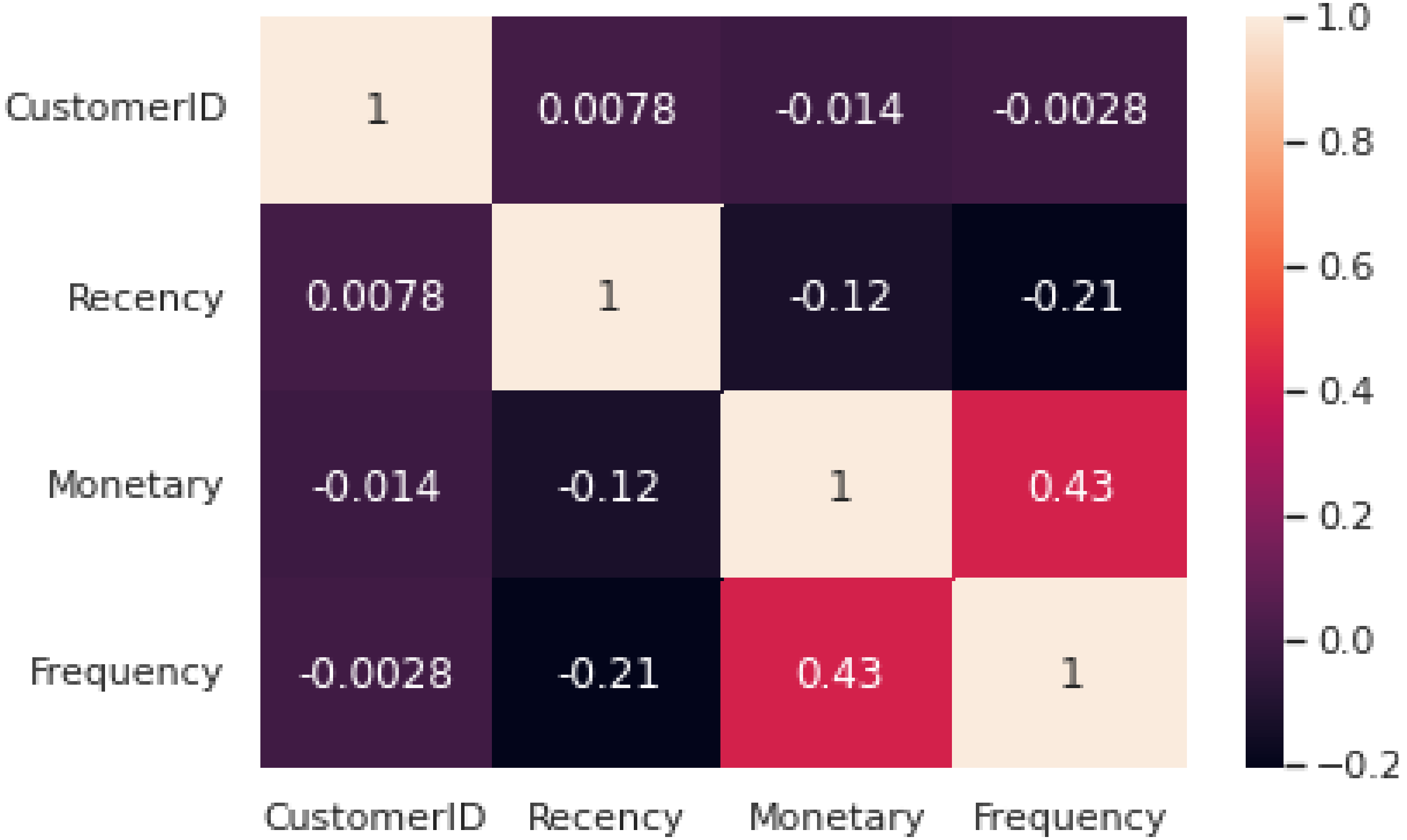
RFM analysis allows you to segment customers by the frequency and value of purchases and identify those customers who spend the most money.



- Recency — how long it's been since a customer bought somethingfromus.
- Frequency—how often a customer buys from us.
- Monetary value—the total value of purchases a customer has made.

	CustomerID	Recency	Monetary	Frequency
0	12346.0	326.0	77183.60	1
1	12347.0	2.0	4310.00	182
2	12348.0	75.0	1797.24	31
3	12349.0	19.0	1757.55	73
4	12350.0	310.0	334.40	17

Correlation among RFM



In the correlation matrix of RFM:

•Frequency and Monetary value is positvely correlated, somehow frequency purchasing affects monetary value too.

•Frequency and Recency is also positive correlated but not having very high correlation between them.

Model Building (Clustering)

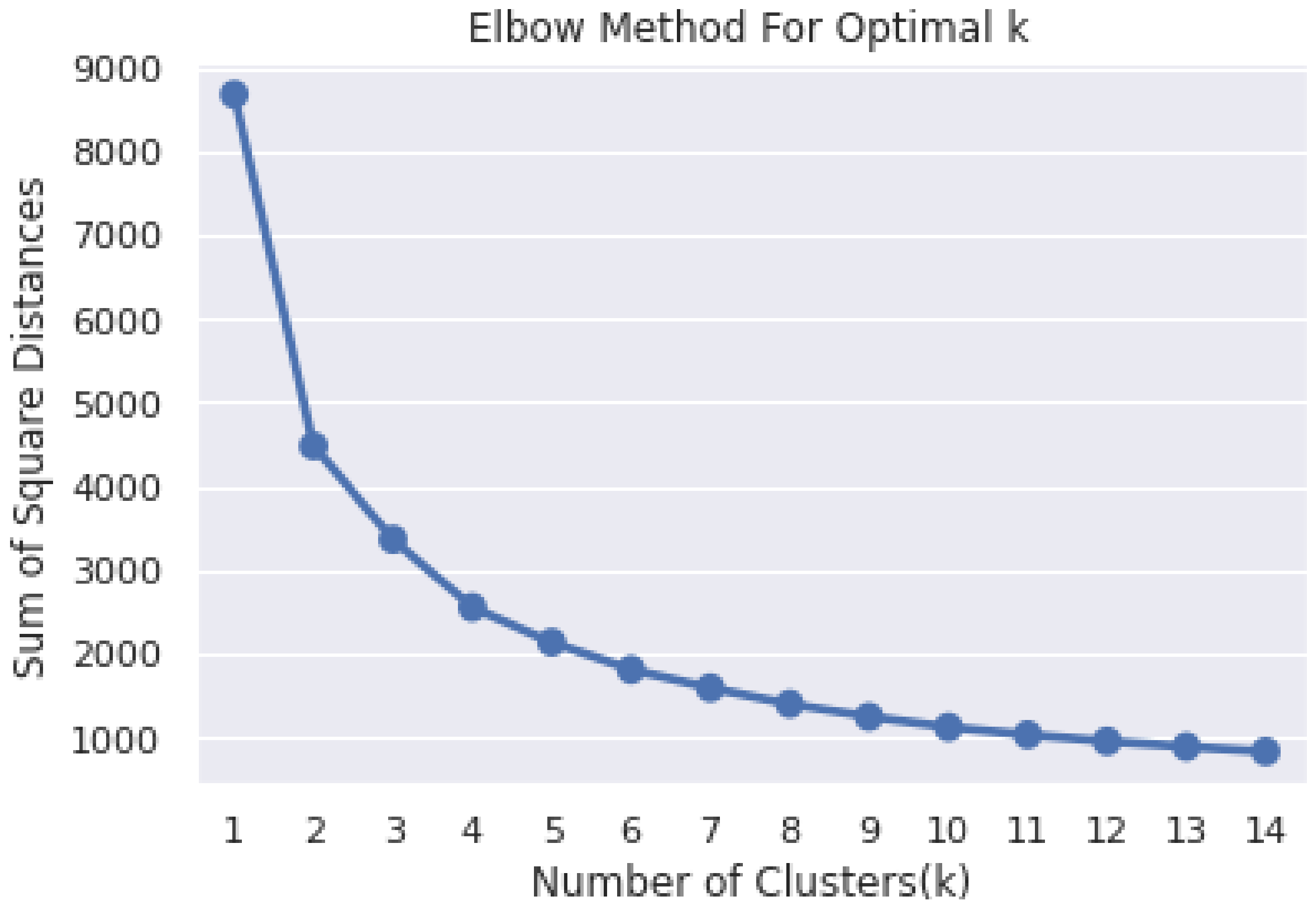
In this section, we use K-Means algorithm to cluster the customers into different segments.

To identify the optimum number of clusters, we use the elbow method and silhouette analysis.

With both the methods,3 clusters is optimum in this case.

A K-Means model with 3 clusters is developed and customers are segmented into different clusters.

Modeling Data: K-Means Clustering .



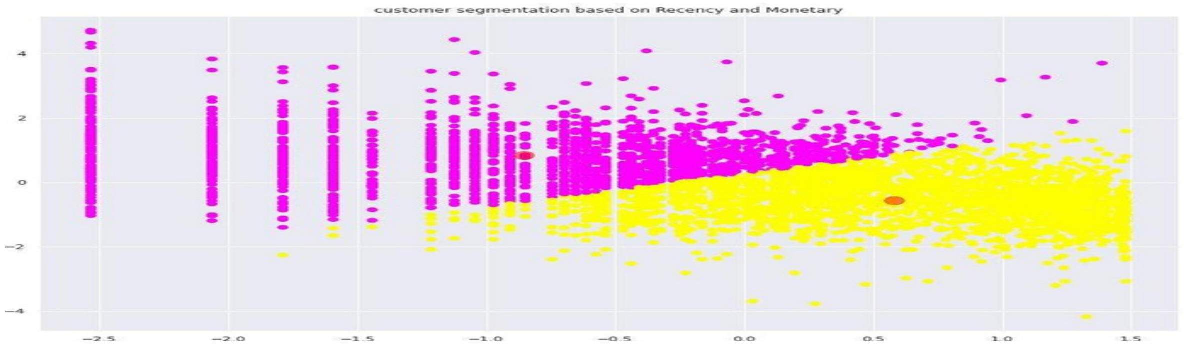
In order to choose a better cluster we need to choose the number of cluster which has minimum WCSS.

As it can be seen in above Elbow Method 4 seems to be the better cluster which has lower WCSS.

If we go further then there is very slight downfall in WCSS so, 4 seems to be a good no of cluster.

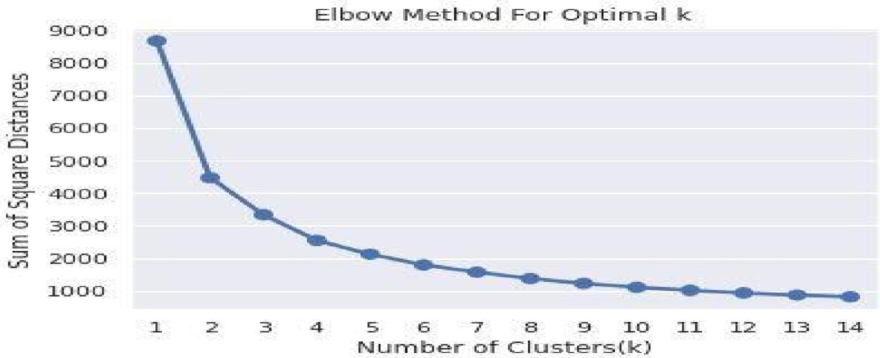
Silhouette score

```
For n_clusters = 2, silhouette score is 0.4216081125935063
For n_clusters = 3, silhouette score is 0.3432957775914936
For n_clusters = 4, silhouette score is 0.36494104664274657
For n_clusters = 5, silhouette score is 0.33668503688485785
For n_clusters = 6, silhouette score is 0.34397809419193187
For n_clusters = 7, silhouette score is 0.3458567202377316
For n_clusters = 8, silhouette score is 0.33919727934627264
For n_clusters = 9, silhouette score is 0.3458423886312394
For n_clusters = 10, silhouette score is 0.34850666375861195
For n_clusters = 11, silhouette score is 0.3385166366909024
For n_clusters = 12, silhouette score is 0.3427649471441594
For n_clusters = 13, silhouette score is 0.34083950250492523
For n_clusters = 14, silhouette score is 0.3406096956008792
For n_clusters = 15, silhouette score is 0.34223526314989594
```

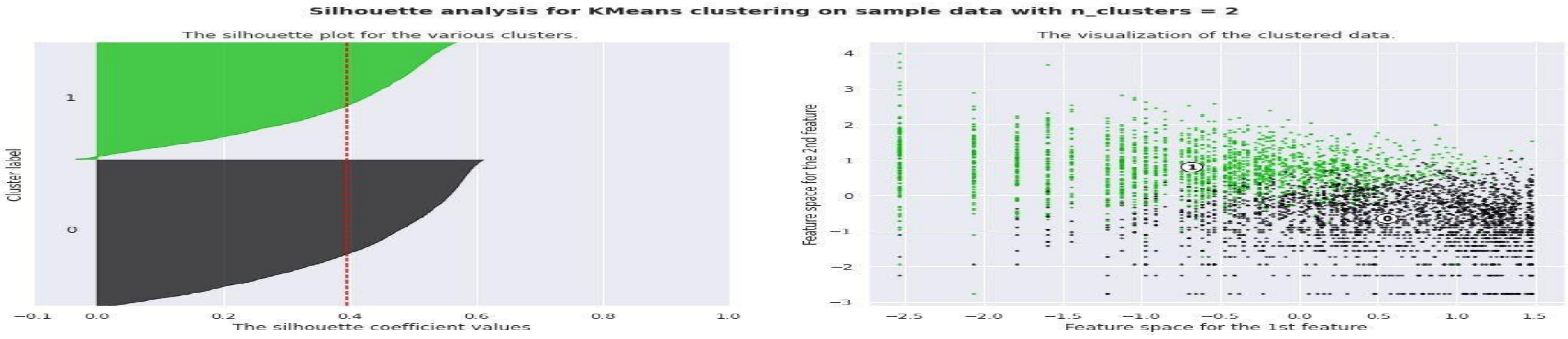


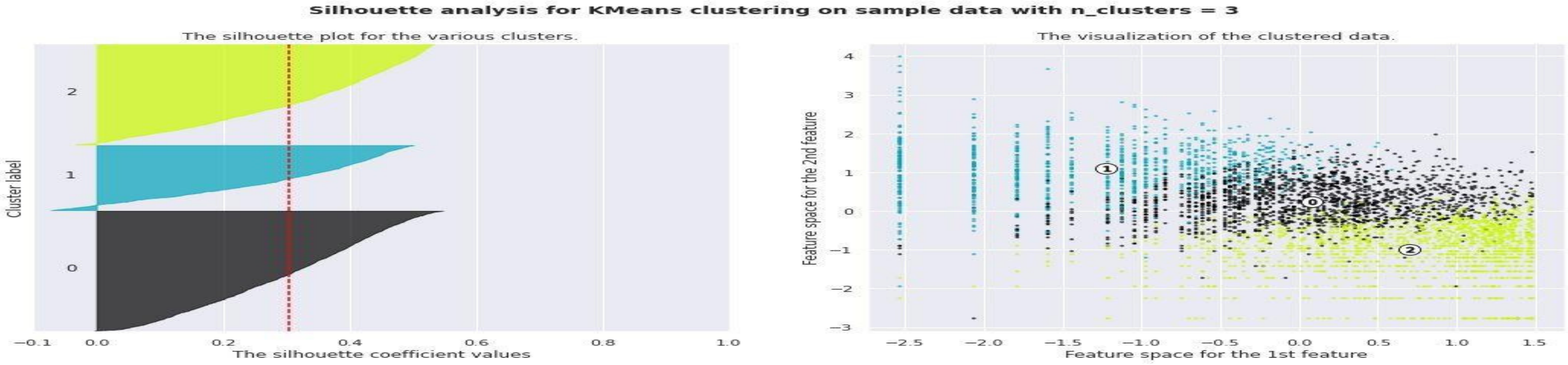
Silhouette score and Elbow method on F&M

```
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 0.371378259651020З
For n_clusters = 5, silhouette score is 0.34479733808079405
For n_clusters = 6, silhouette score is 0.35974563779013946
For n_clusters = 7, silhouette score is 0.33835032540639154
For n_clusters = 8, silhouette score is 0.351989209180013З
For n_clusters = 9, silhouette score is 0.3460160650521864
For n_clusters = 10, silhouette score is 0.3619887930235607
For n_clusters = 11, silhouette score is 0.36822618560766546
For n_clusters = 12, silhouette score is 0.35460489785135785
For n_clusters = 13, silhouette score is 0.3624674157300161
For n_clusters = 14, silhouette score is 0.36520616987776316
For n_clusters = 15, silhouette score is 0.36101570873847355
```

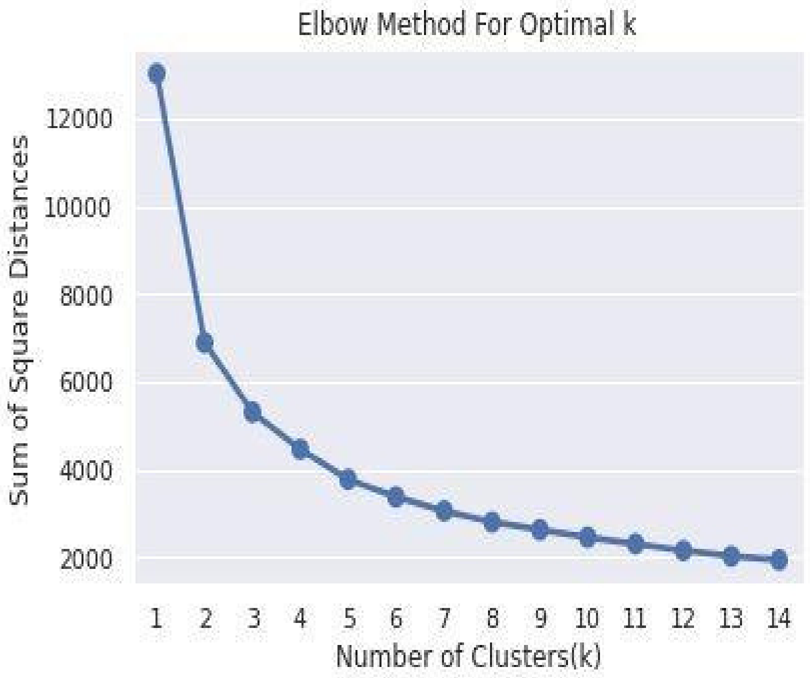


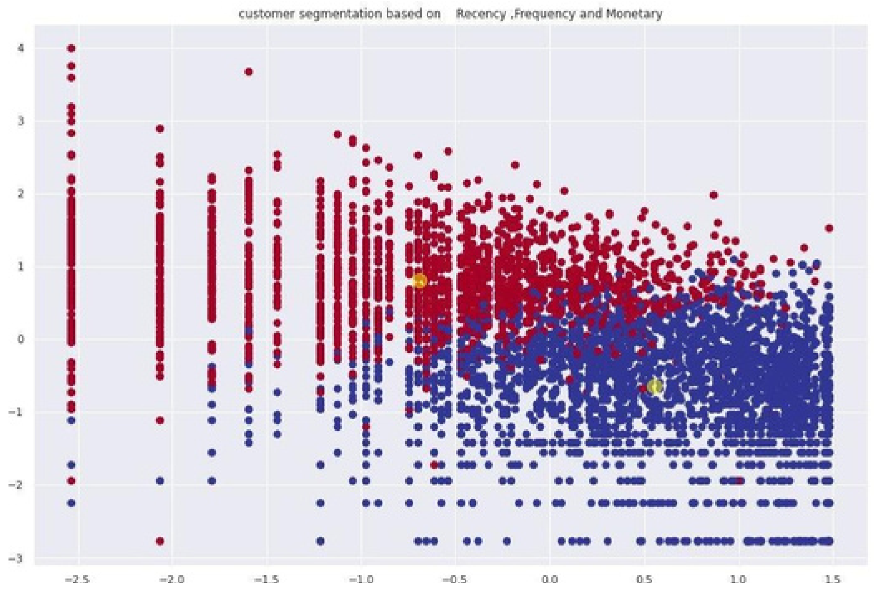
Silhouette analysis on RFM



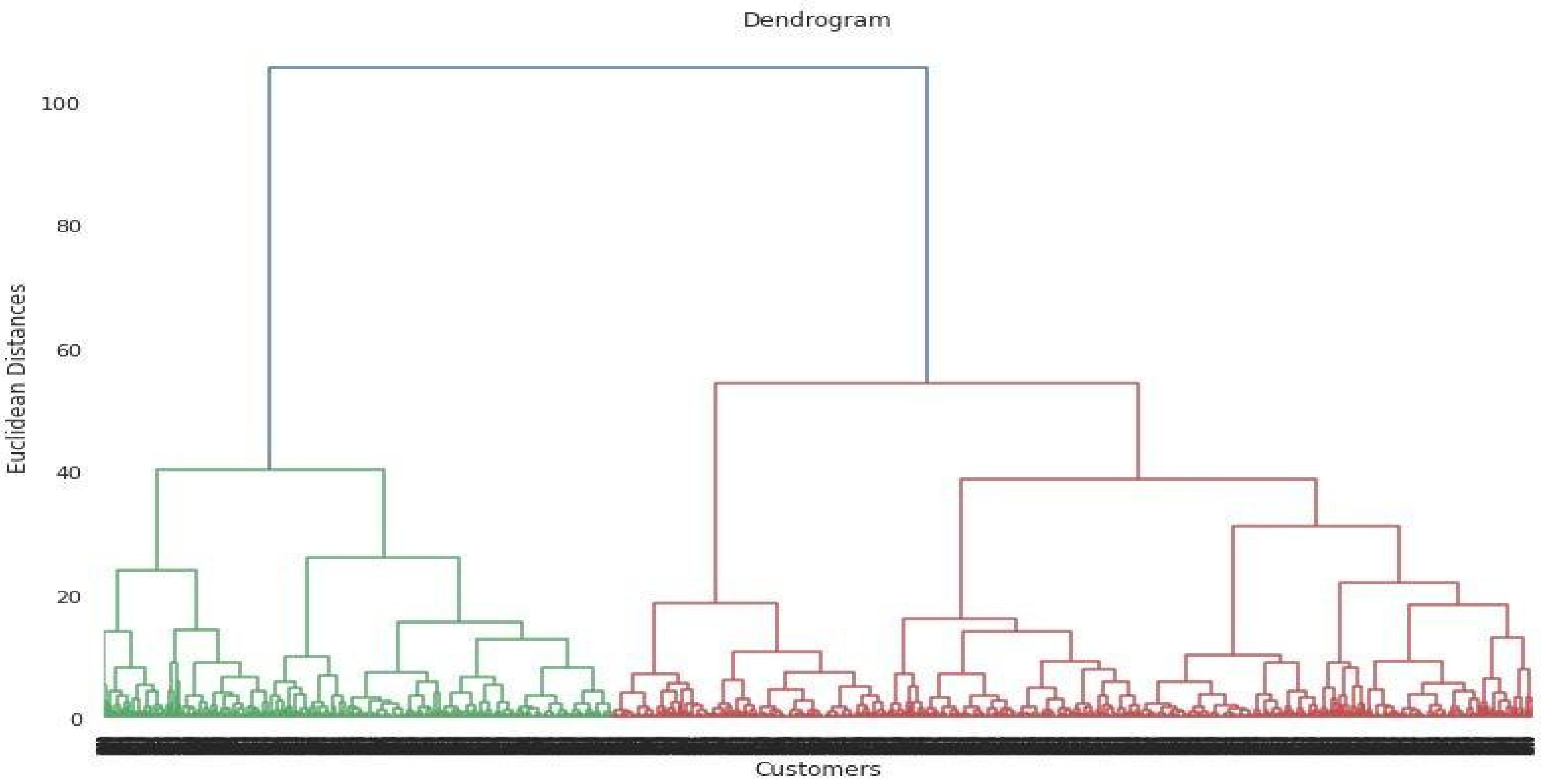


Elbow method and Cluster chart on RFM





Dendrogram



The more we climb the tree, the more the classes are grouped together and the less they are homogeneous (less intra- class inertia).

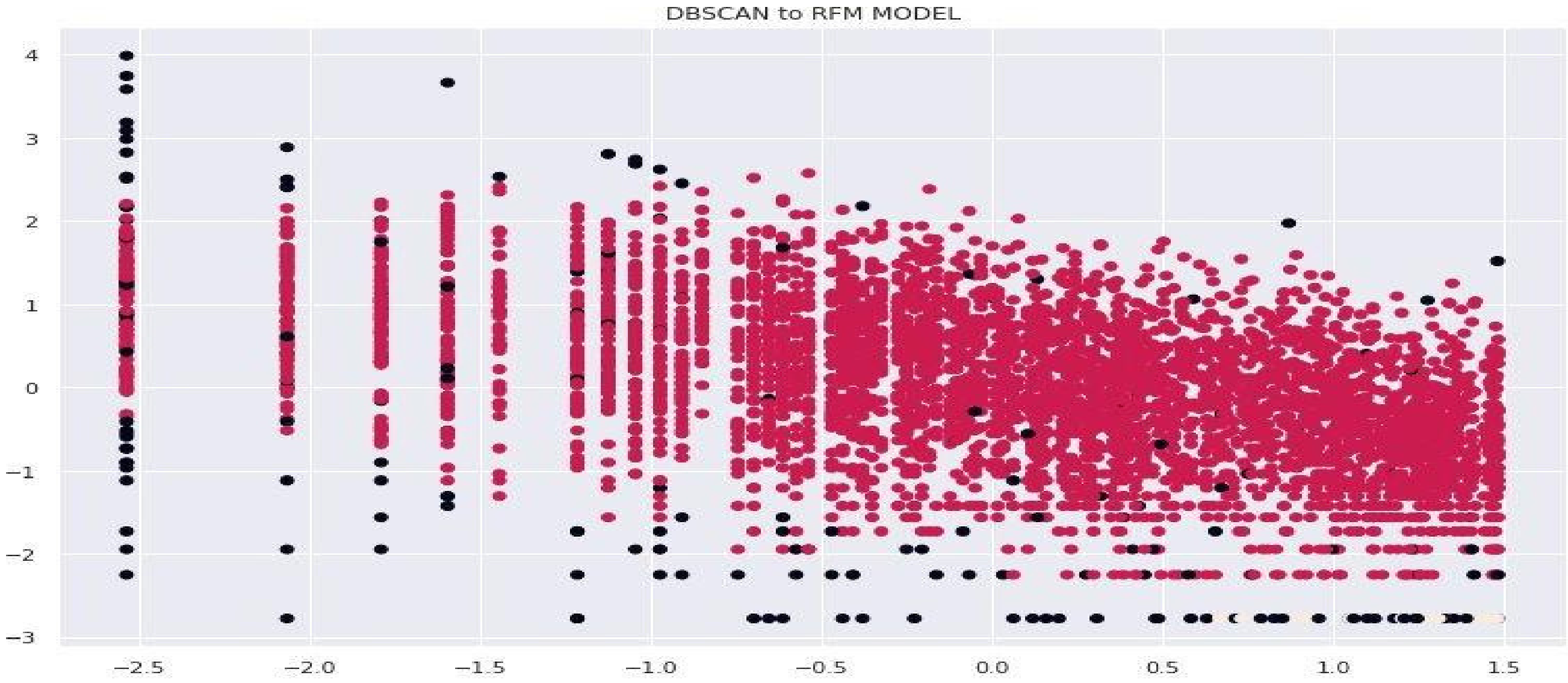
The number of classes is a compromise between the similarity in the classes and the dissimilarity between the classes

Table: -

The visual representation of the data in tabular forms.

Sr. No.	Model_Name	Data	Optimal_Number_of_cluster
1	K-Means	RFM	3
2	K-Means with silhouette_score	RFM	2
3	K-Means with Elbow method	RFM	4
4	Hierarchical clustering	RFM	2
5	Hierarchical clustering after Cut-off	RFM	3

DBSCAN



Challenges

After grouping, the data consists mostly of duplicated data. There were originally over five lakh records.

According to Customer ID, there were 3.5 thousand customers left.

There were many negative values when grouping them according to certain assumptions.

Conclusion

In this instance, we compared RFM Analysis with K Mean Clustering.

What is the most effective clustering algorithm in Kmeans modeling? The first advantage of this dataset is that the data can be scaled and centered better, Since this dataset contains a large number of outliers, a robust scaler should be used.

As a result, we have determined the appropriate cluster size for this data. This will enable us to figure out how much segmentation should be given and what silhouette score each n_cluster should have.

As a final step, we conducted PCA (principal component analysis) to identify the most suitable components.

Thank you