Capstone Project

Online Retail Customer Segmentation

Individual Project

Name : Aakash Ramnani

Content

- Problem Statement
- Data Summary
- Project Summary
- EDA (Exploratory Data Analysis)
- Hypothesis Testing
- Feature Engineering and Data Pre-processing
- Model Implementation
    1. KMeans Clustering
    2. DBScan Clustering
    3. Hierarchical Agglomerative Clustering
- Customer Segmentation
- Conclusion

Problem Statement

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.

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:** Invice 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.

Project Summary

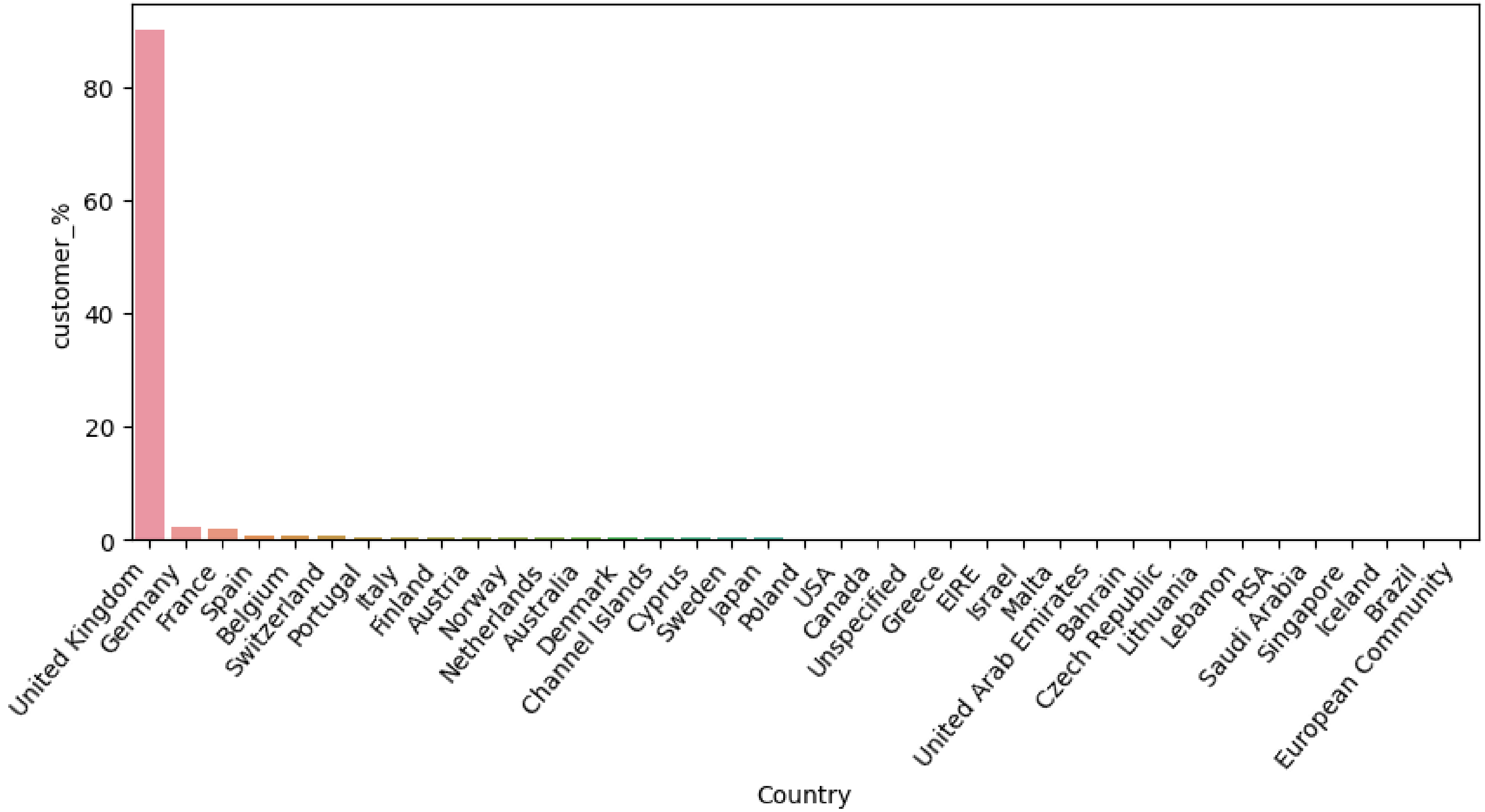
- In this project, the task is to indentify major customer segments on a traditional data set which contains all the transaction occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store retail.
- Project is performed in following steps

1. **Basic EDA on Dataset -** This step involves exploring data set and checking relationship between variables and checking their distributions.
2. **RFM Analysis -** RFM (Recency, Frequency, Monetary) analysis is a customer segmentation technique that uses past purchase behavior to divide customers into groups.
    1. **RECENCY (R):** Days since last purchase
    2. **FREQUENCY (F):** Total number of purchases
    3. **MONETARY VALUE (M):** Total money this customer spent.
3. **Visualization Using different Charts -** This step involves creating various charts and graphs to visualize the data and identify the patterns and relationships among the features. Some of the charts that can be used are bar charts, scatter plots, heat maps, etc.
4. **Hypothesis Testing -** This step involves testing some hypotheses or assumptions about the data using statistical methods.
5. **Feature Engineering for clustering -** This step involves creating new features or transforming existing features to make them suitable for clustering.
6. **Clustering analysis using k-means and agglomerative -** This step involves applying k-means and agglomerative clustering algorithms to group the customers based on their RFM score. This step can help to identify the optimal segments of customers.

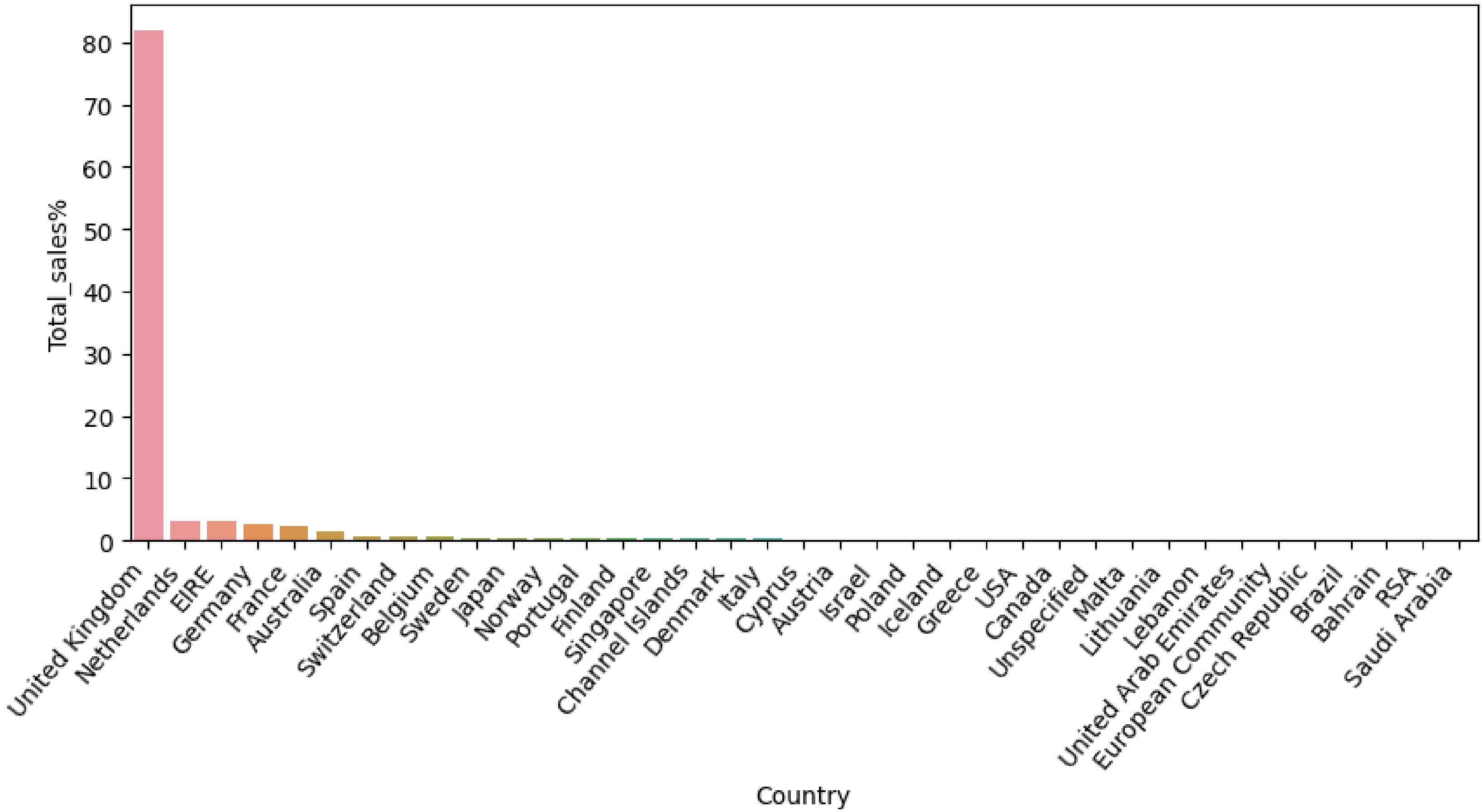
EDA(Exploratory Data Analysis)

- We did a detailed bi-variate and multivariate exploratory data analysis and came up with some insights that might be useful for the business.

EDA W.R.T Country

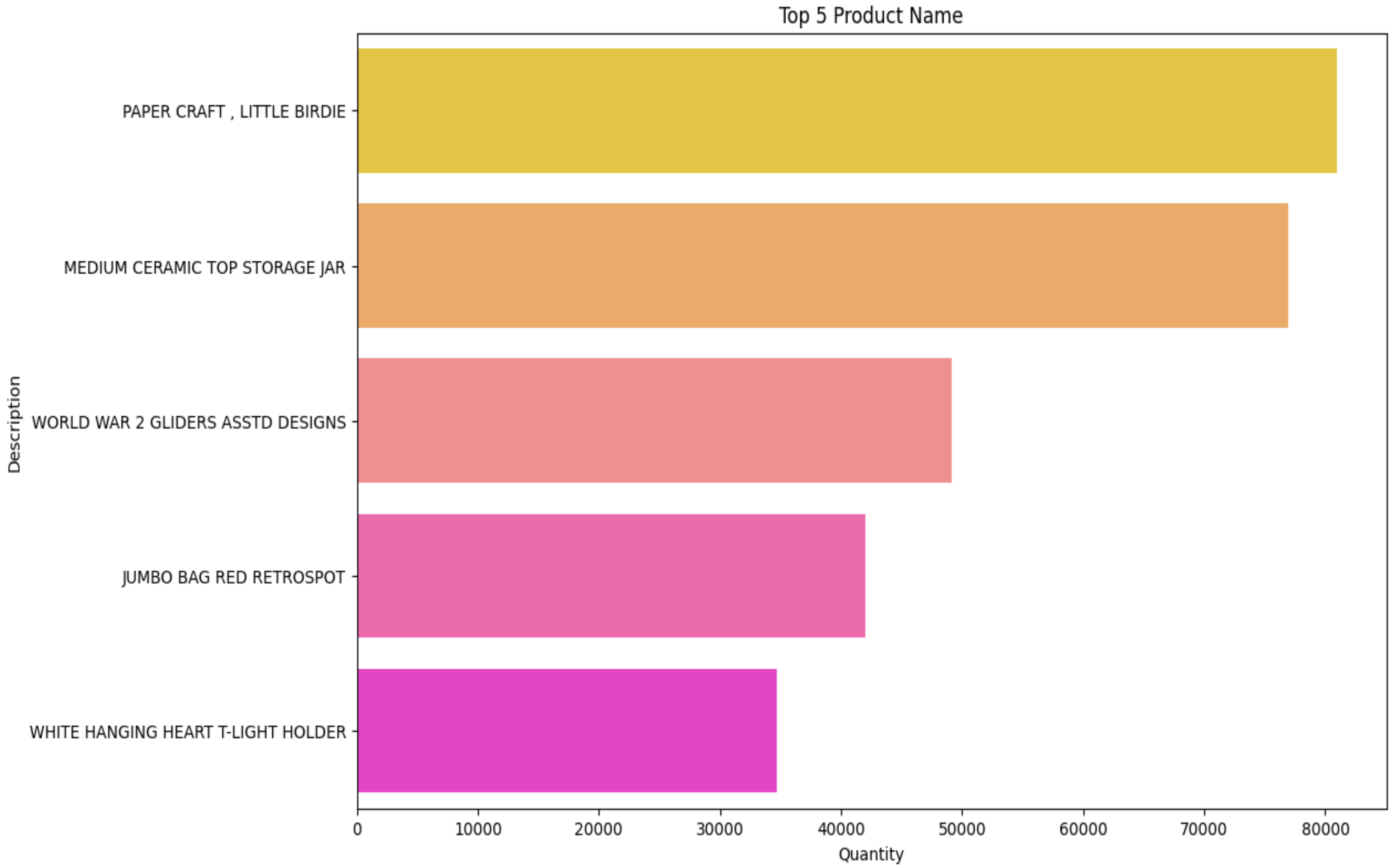


- 90% of total customer are from United-Kingdom.



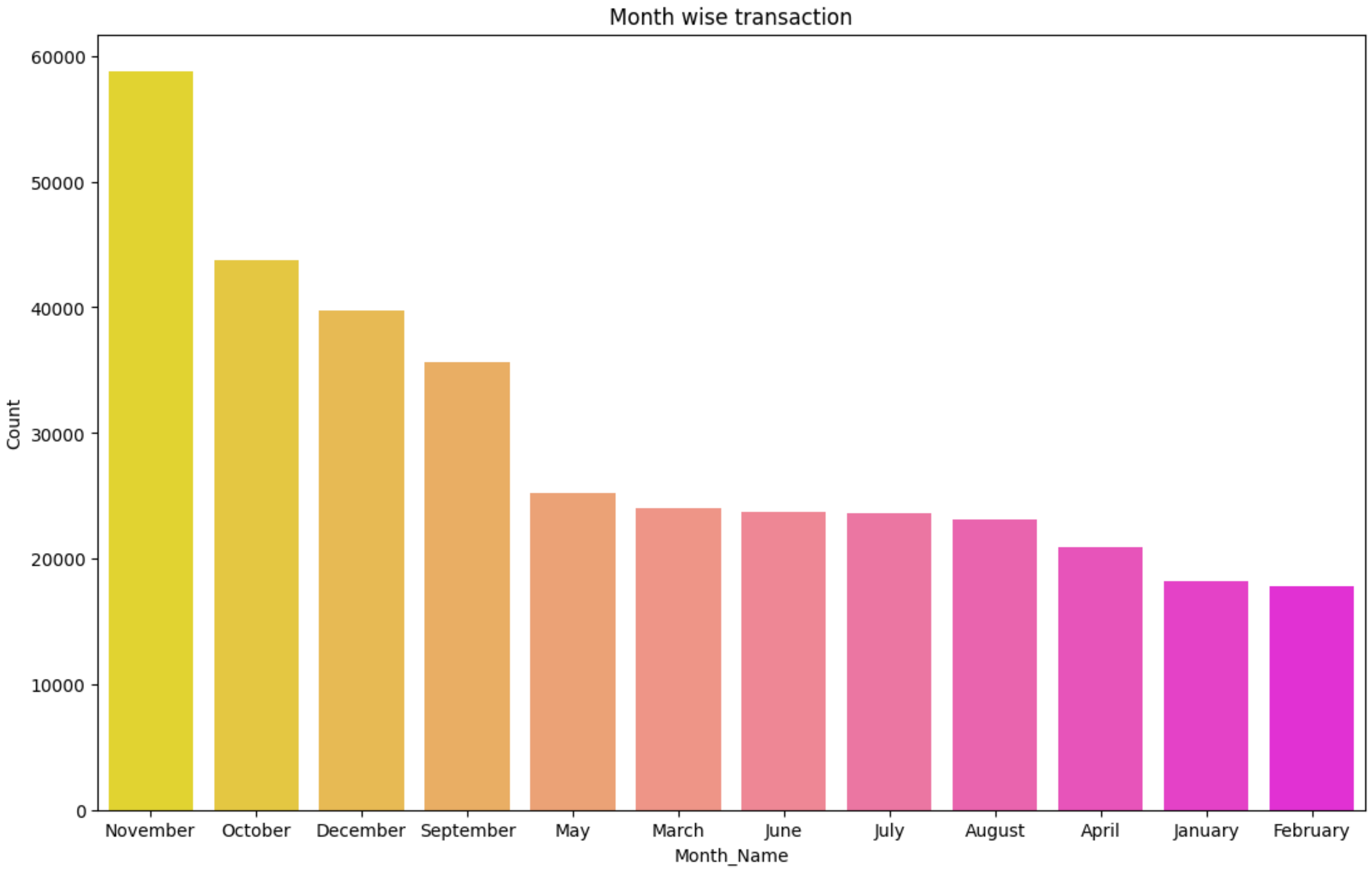
- 82% of total sales revenue is from United Kingdom.

Top 5 Products



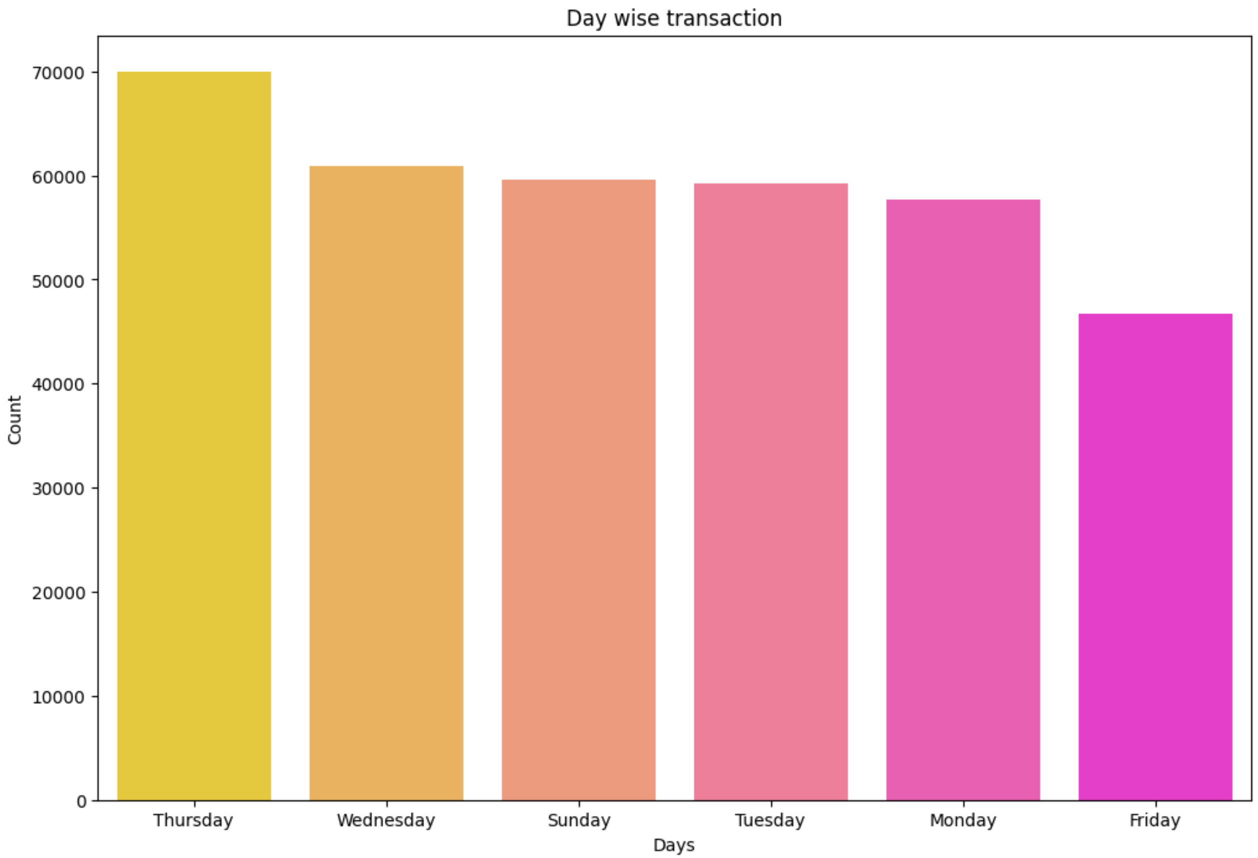
- Paper Craft, little birdie and Medium Ceramic top storage jar are the top 2 products that are ordered the most.
- This insight helps the store to understand the demand of a particular product so that they can keep the stock intact and reduce that their purchase price by buying in quantity as they are sure about the fact that it is one of the most ordered product.

Month-wise Transaction



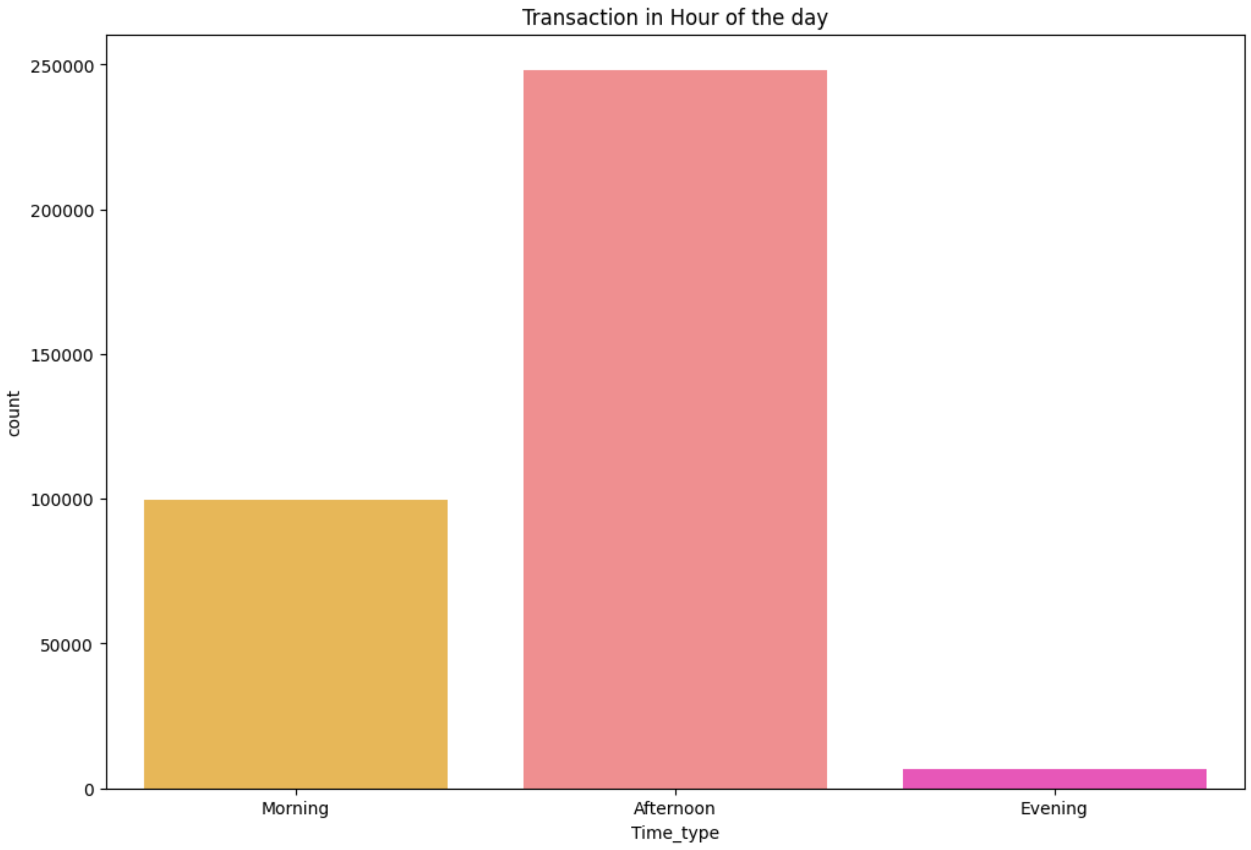
- Most Number of transactions are done in the month of November followed by October and December.
- These behavior can be expected because of festive season in those month. This insight can help store to be prepared for all sales by keeping stock intact and also running promotion accordingly to drive more sales.

Day-wise Transaction



- Most Number of transactions are done on Thursday.
- There are no sales on Saturday store should look into it and figure out the reason behind it.

Transaction in Hour of the Day

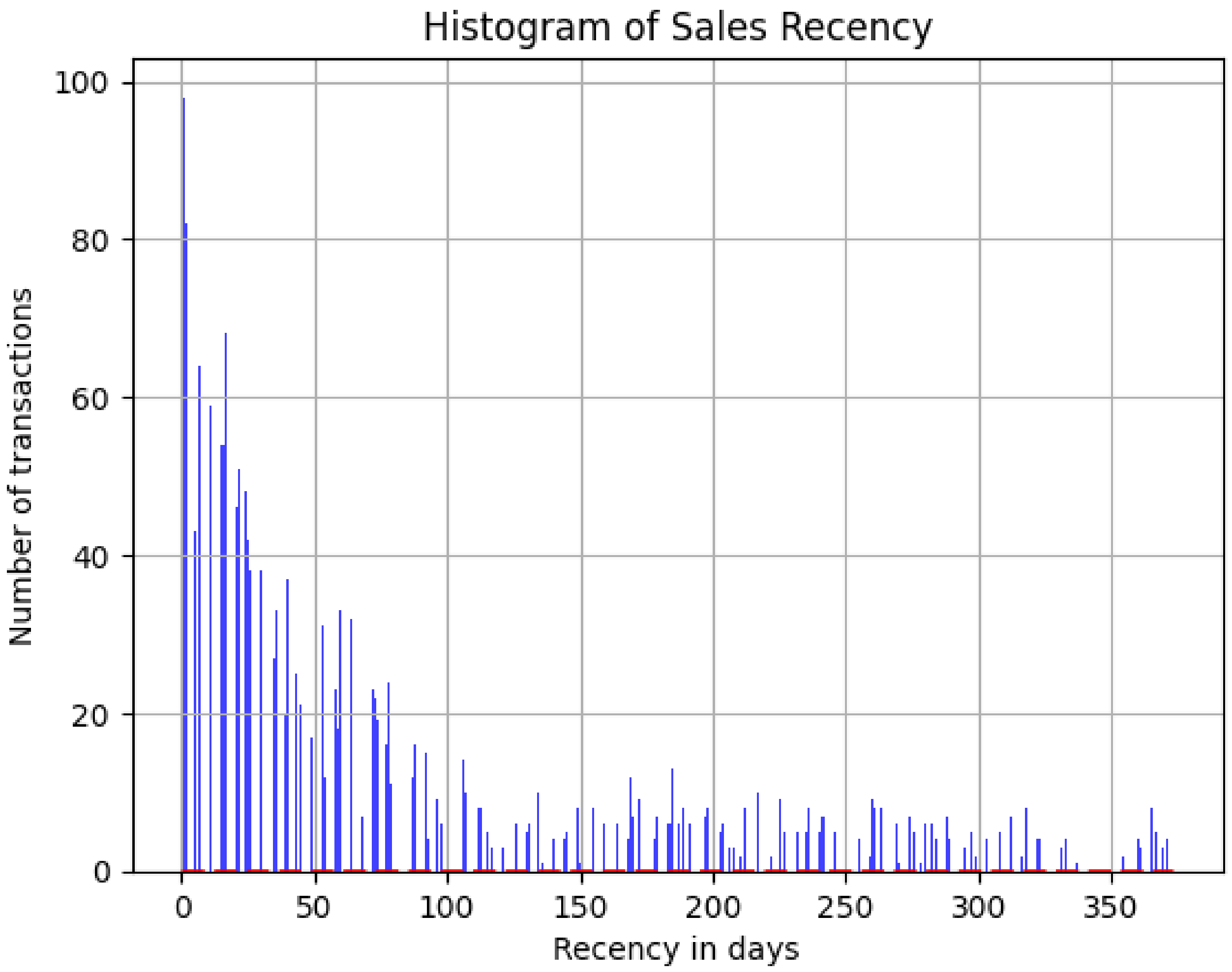


- Most number of transaction are done in afternoon, followed by morning.
- Least number of transaction are done in Evening.
- This can help store to schedule there digital advertisement accordingly for better optimization.

RFM Analysis

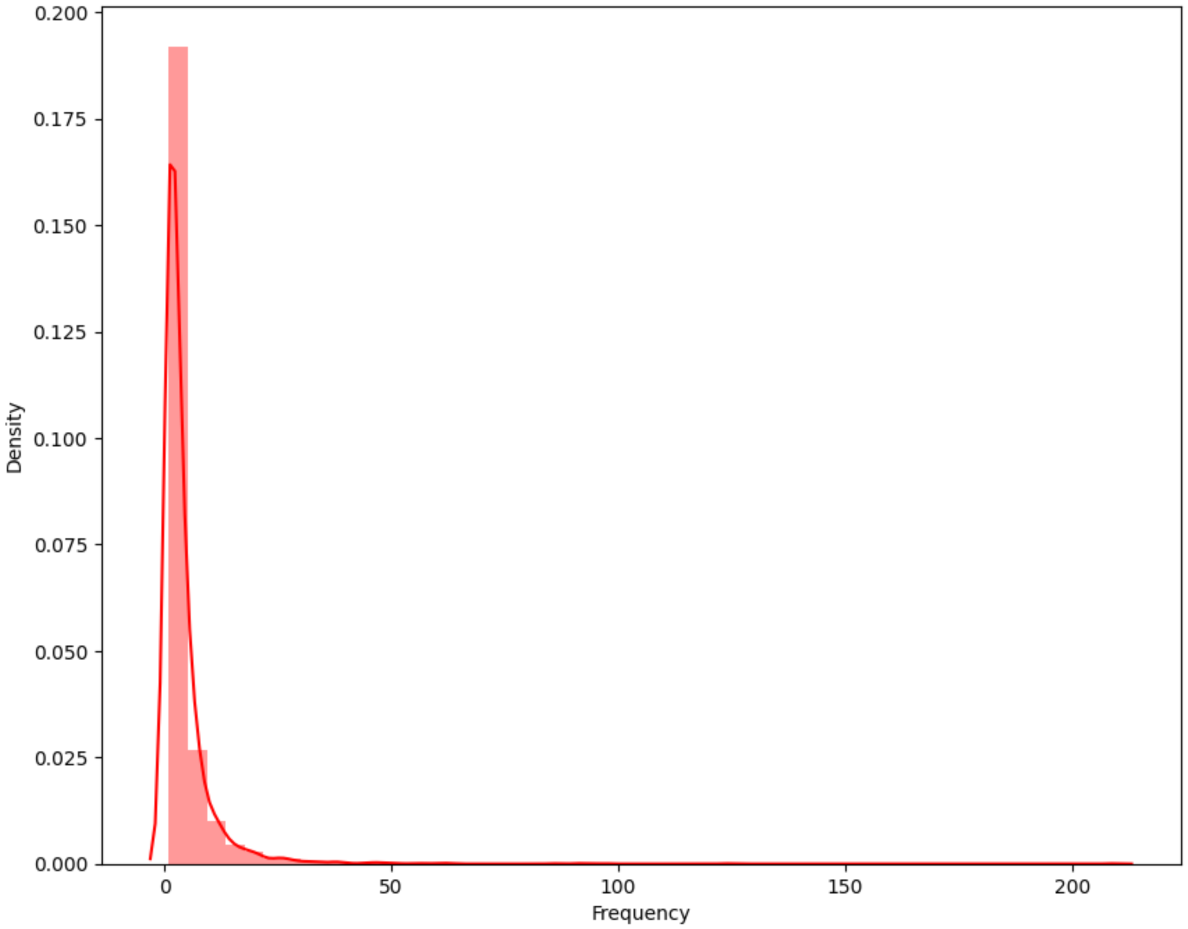
- RFM (Recency, Frequency, Monetary) analysis is a customer segmentation technique that uses past purchase behavior to divide customers into groups.
- RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.
    - RECENCY (R): Days since last purchase
    - FREQUENCY (F): Total number of purchases
    - MONETARY VALUE (M): Total money this customer spent.

Recency



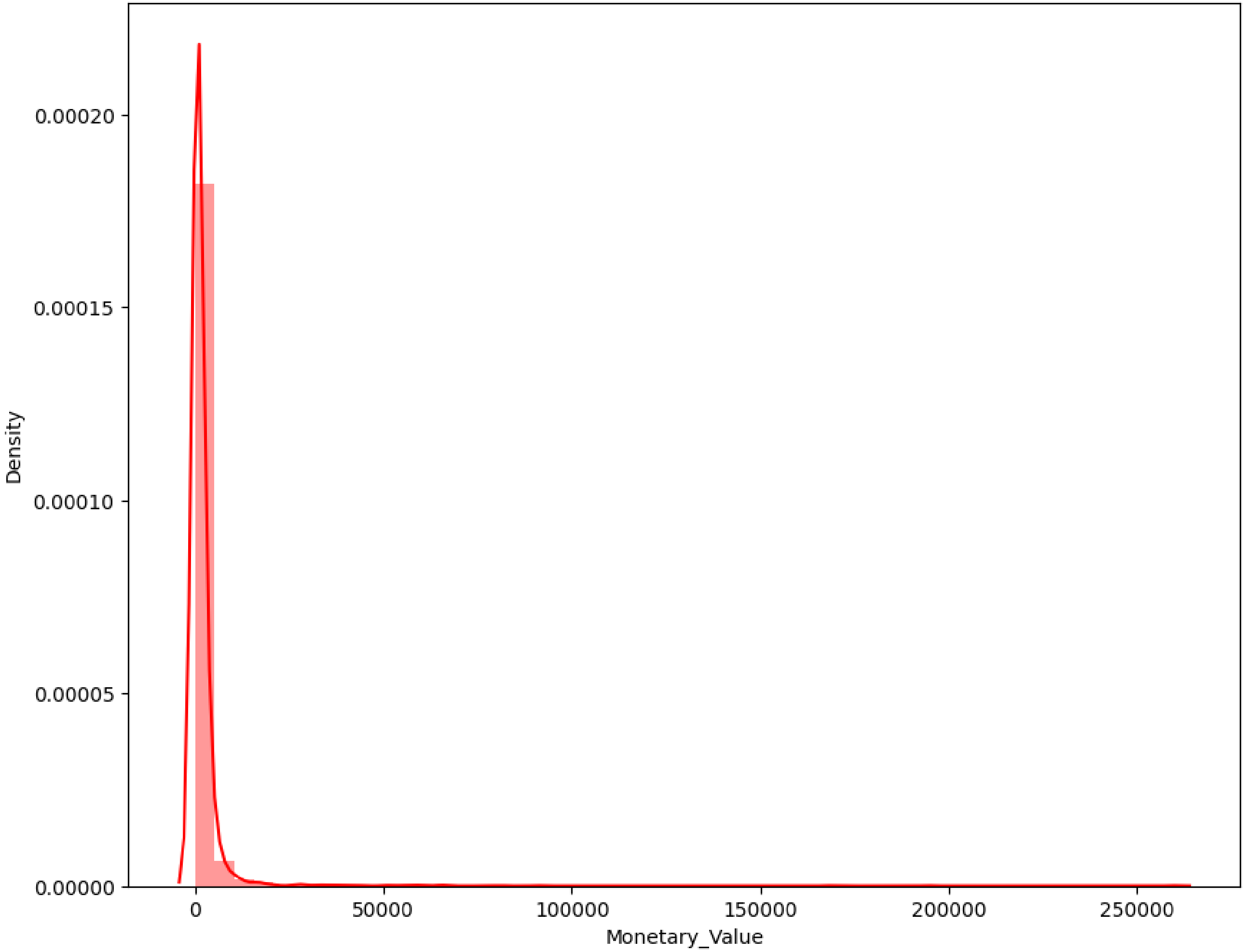
- We have a skewed distribution for recency showing higher transaction in recent days.
- Recency shows days since last purchase this information can help store to see the which customers they are about to lose and which they have lost. And thus come up with a strategies to retain them.

Frequency



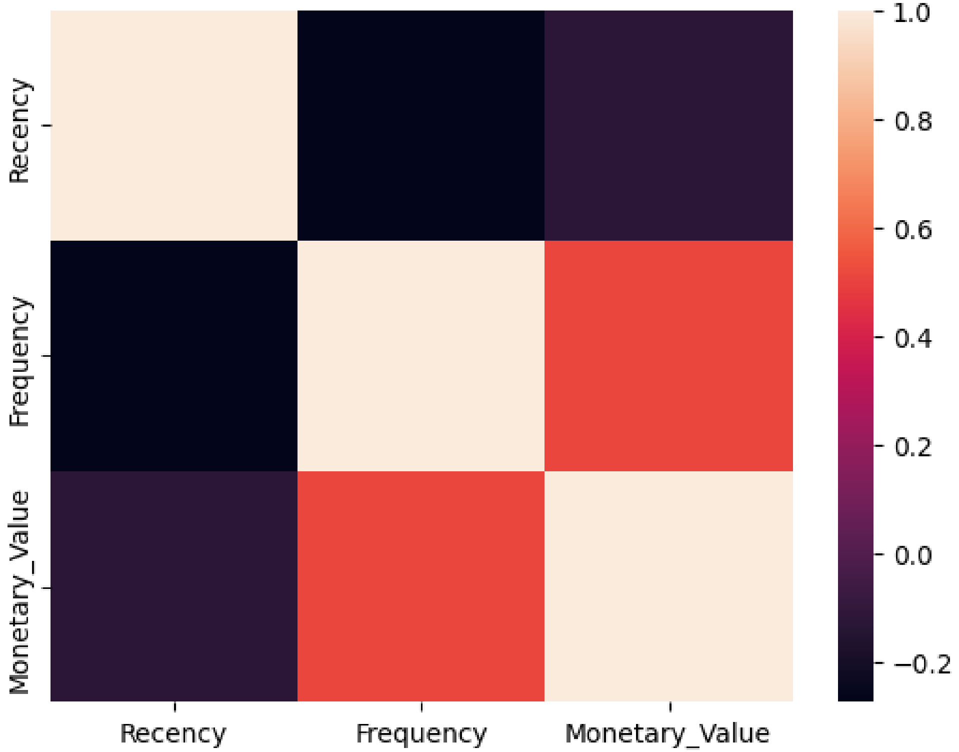
- Frequency helps us to know how many times a customer purchased from us.
- Number of purchases are between 1 to 15. Customers with higher number of purchases are loyal customer and could be included in loyalty program.

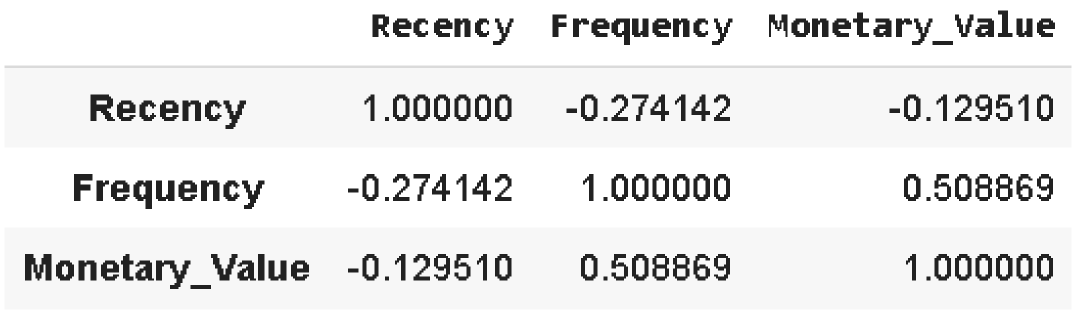
Monetary Value



- Monetary attribute answers the question: How much money did the customer spent over time?
- Value ranges from 0-2000. Customer on higher end of this spectrum are premium customer.

Correlation Matrix for RFM





	Recency	Frequency	Monetary_Value
Recency	1.000000	-0.274142	-0.129510
Frequency	-0.274142	1.000000	0.508869
Monetary_Value	-0.129510	0.508869	1.000000

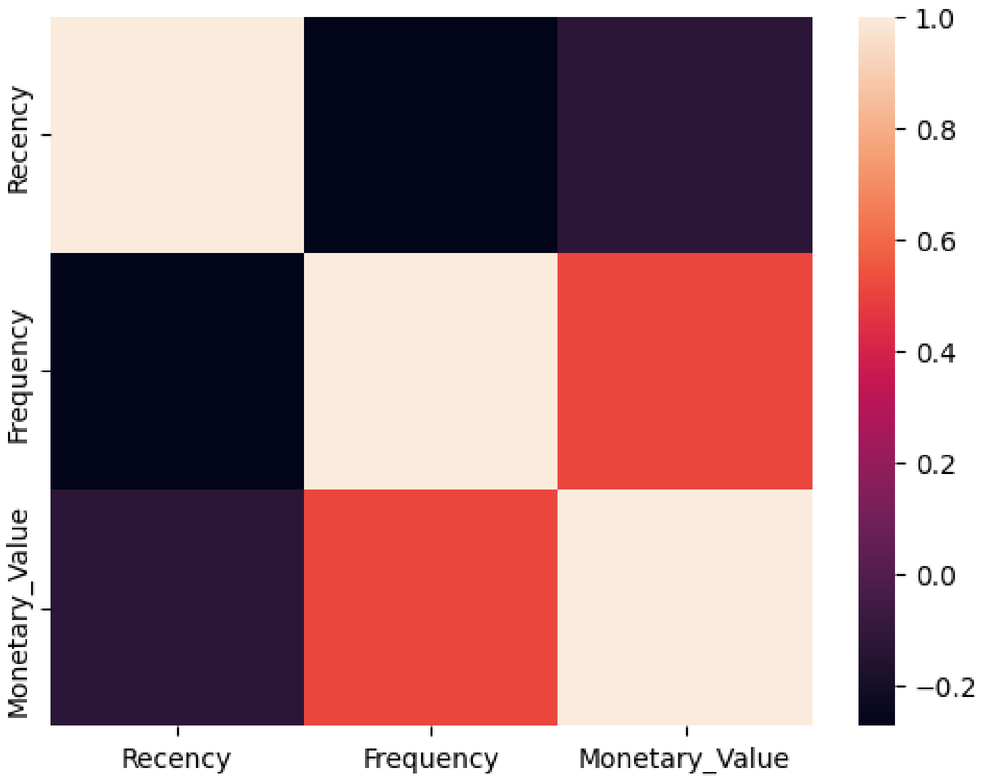
- Recency is negatively correlated to both Frequency and Monetary value.
- Frequency and Monetary Value are positively correlated.

Hypothesis Testing

- Based on our EDA, we have defined these three hypothesis.

1. Recency is negatively correlated to Frequency.
2. Monetary Value and Frequency are positively correlated.
3. Recency is negatively correlated to Monetary value.

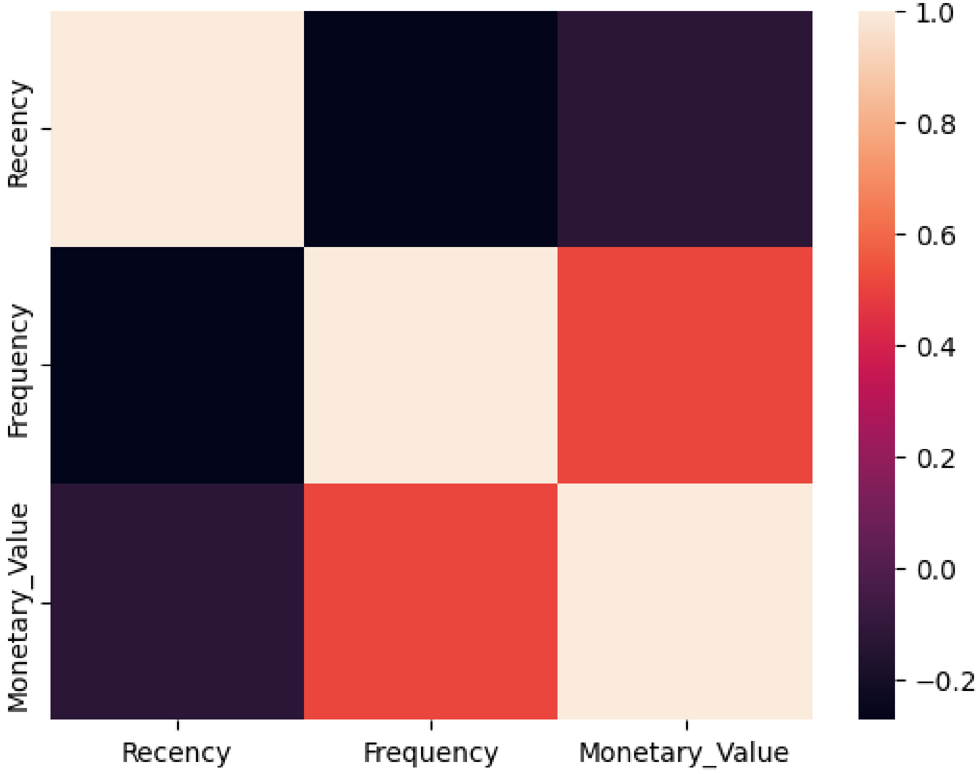
Hypothesis - 1



- **Null Hypothesis(H0) -** There is no correlation between Recency and Frequency.
- **Alternate Hypothesis(HA) -** Recency is negatively correlated to Frequency.
- We Performed 2 sample t-test for this hypothesis testing

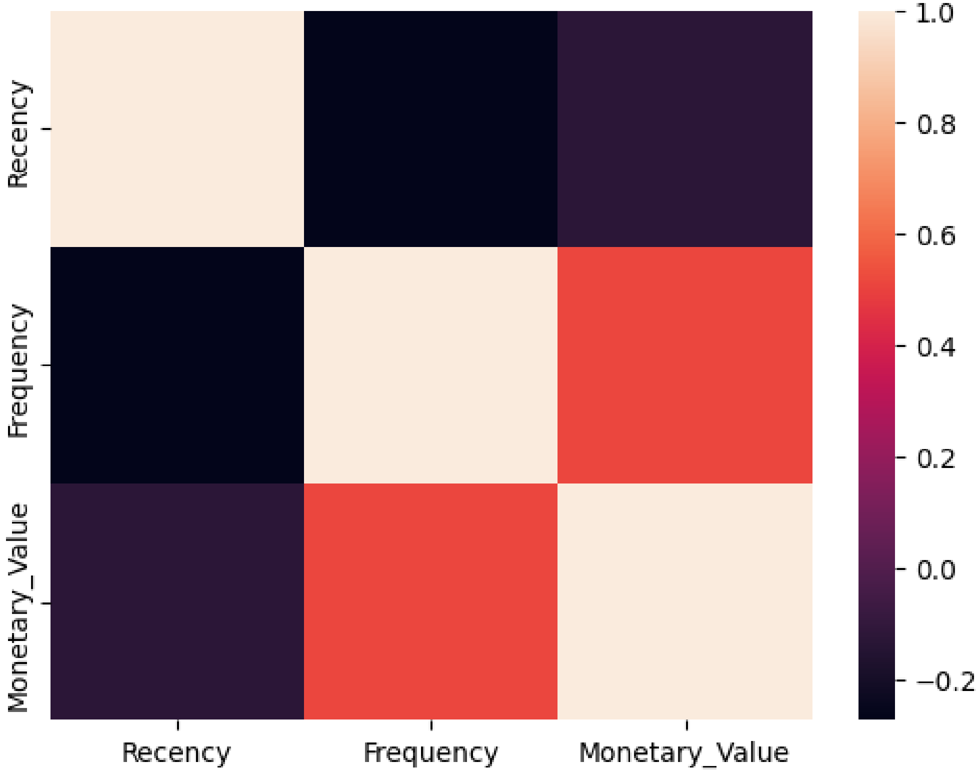
- p-value is 0.00 which is less than significance level 0.05.
- We have enough evidence to reject the null hypothesis.

Hypothesis - 2



- **Null Hypothesis(H0) -** There is no correlation between Monetary value and Frequency.
- **Alternate Hypothesis(HA) -** Monetary Value is positively correlated to Frequency.
- We Performed 2 sample t-test for this hypothesis testing
- p-value is very less than significance level 0.05.
- We have enough evidence to reject the null hypothesis.

Hypothesis - 3



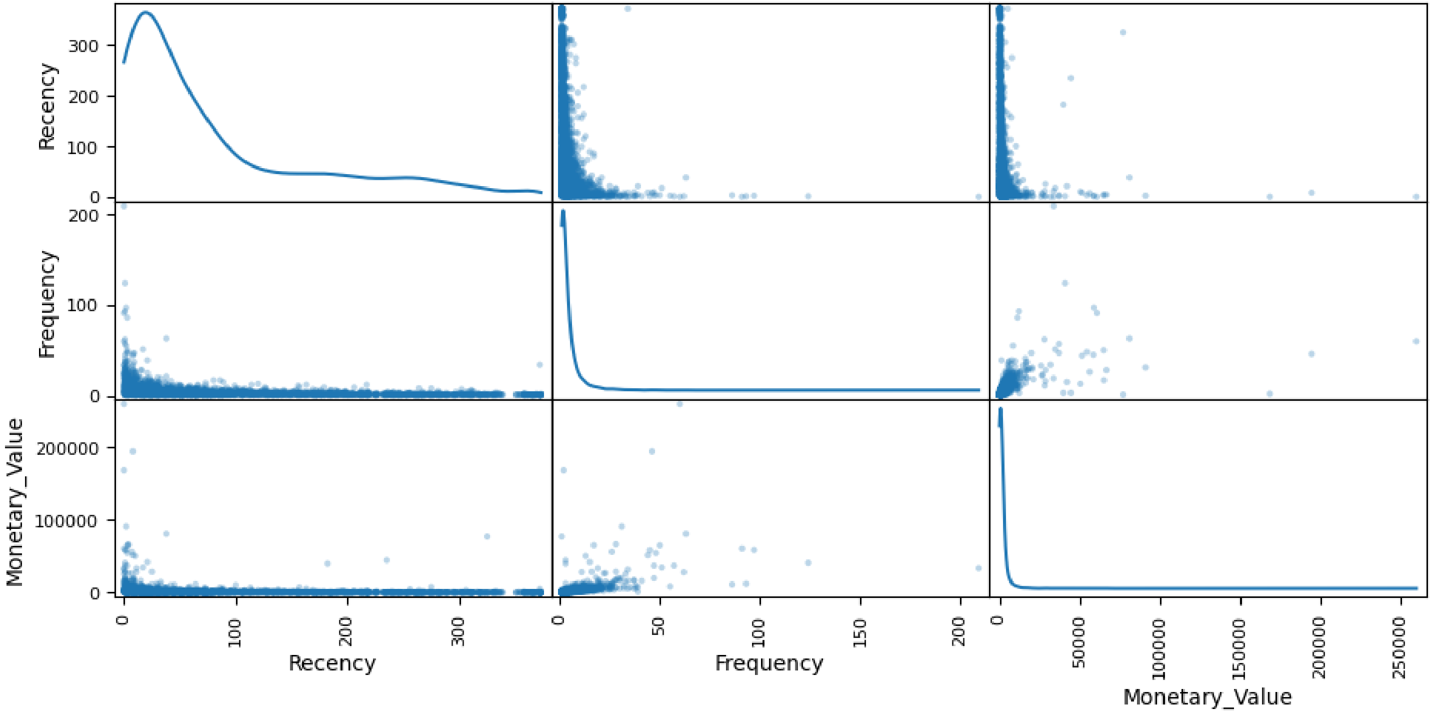
- **Null Hypothesis(H0) -** There is no correlation between Monetary value and Recency.
- **Alternate Hypothesis(HA) -** Recency is Negatively correlated to Monetary Value.
- We Performed 2 sample t-test for this hypothesis testing

- p-value is less than significance level 0.05.
- We have enough evidence to reject the null hypothesis.

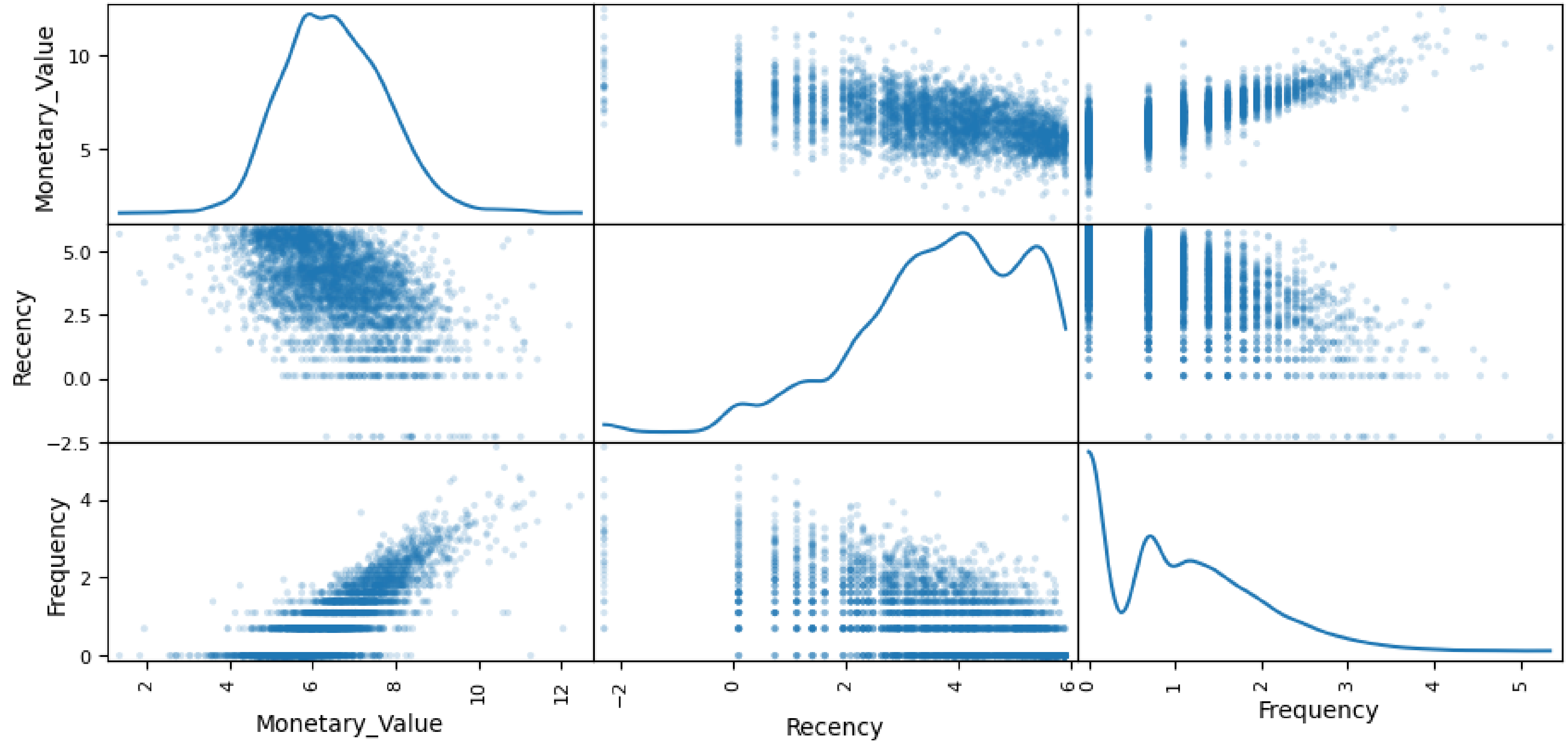
Handling Missing Values

- There were missing values in CustomerID and Description.
- The reason behind this could be that customer buying from the online store was not a registered customer.
- And there is no way we can impute the CustomerID and Description as CustomerID is unique to every customer and Description is unique to every product.
- So we dropped all the information with missing values.

Data Transformation



- Distribution for all three features, Recency, Frequency and Monetary value are right skewed.
- Clustering algorithms require normal distribution.
- We have converted these right skewed distribution to near normal distribution by applying log transformation.

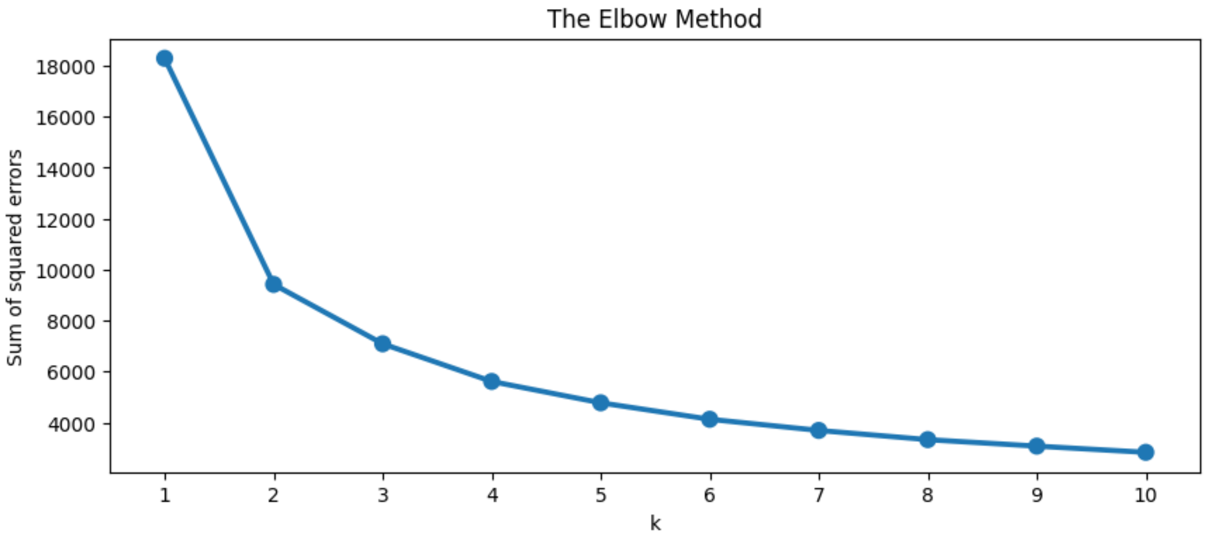


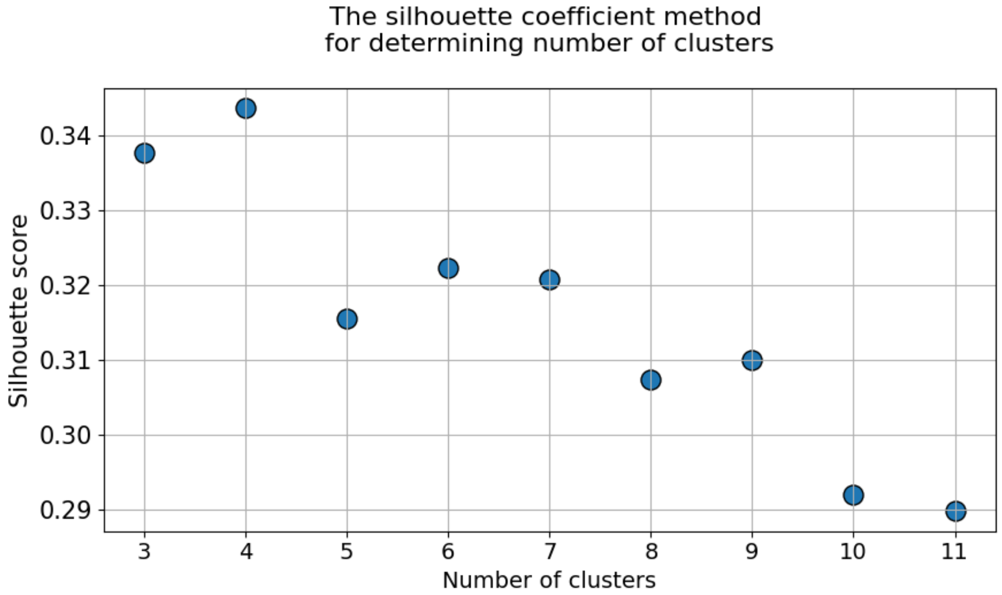
- The Distribution of Monetary value is better, However the distribution of Recency and Frequency have improved but not as much.

Model Implementation

- We have Tried implementing the following models on our dataset :
  - KMeans Clustering
  - DBScan Clustering
  - Hierarchical Agglomerative Clustering

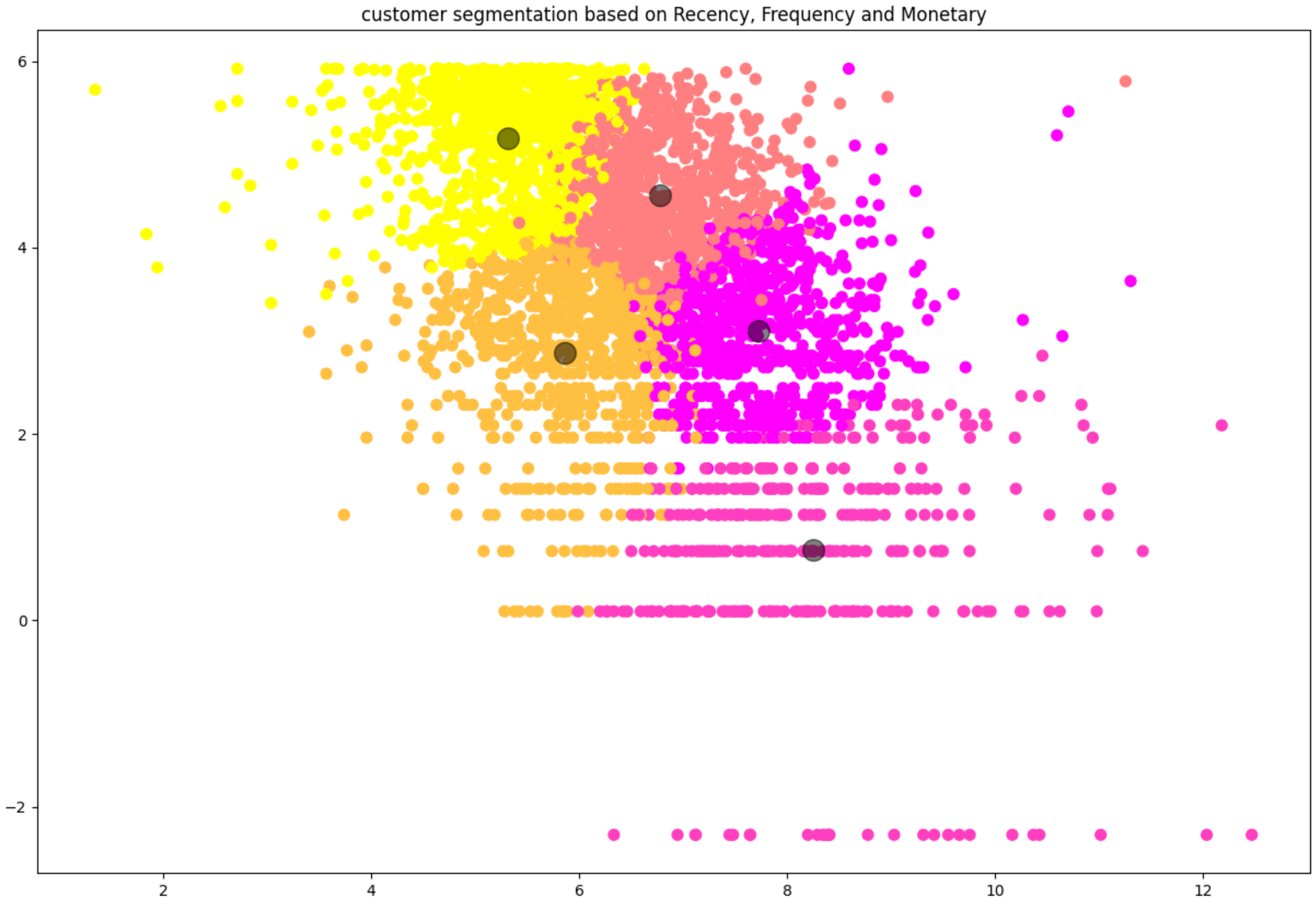
Optimal Cluster



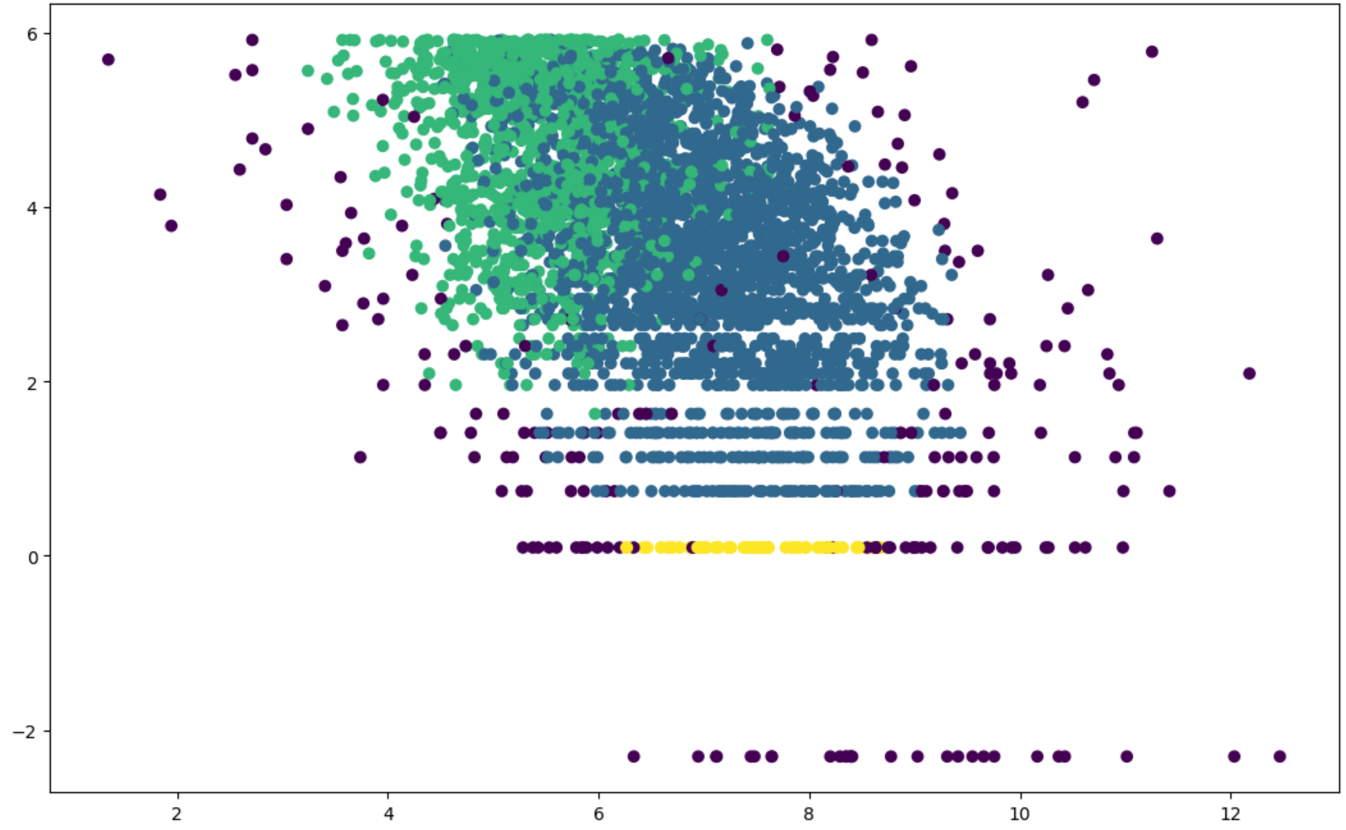


- A common challenge with k-means is that you must tell it how many clusters we expect.
- Figuring out how many clusters we need is not obvious from data, thus we try different clusters numbers and check their Silhouette Coefficient. The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar).
- We have also used Elbow method to determine the number of clusters.
- The optimal number of cluster is 5. Because that is the only point after which the mean cluster distance looks to be plateaued after a steep downfall.
- So we have considered the number of cluster as 5 for our model implementation.

KMeans for 5 Clusters

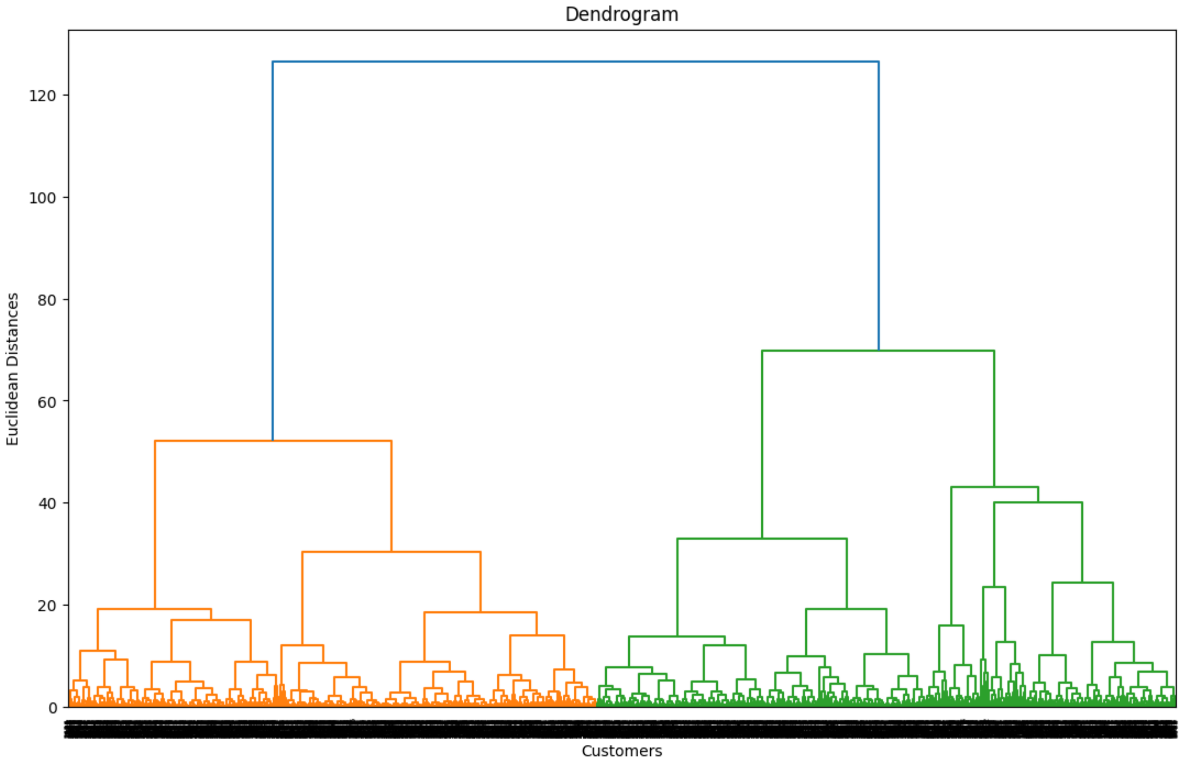


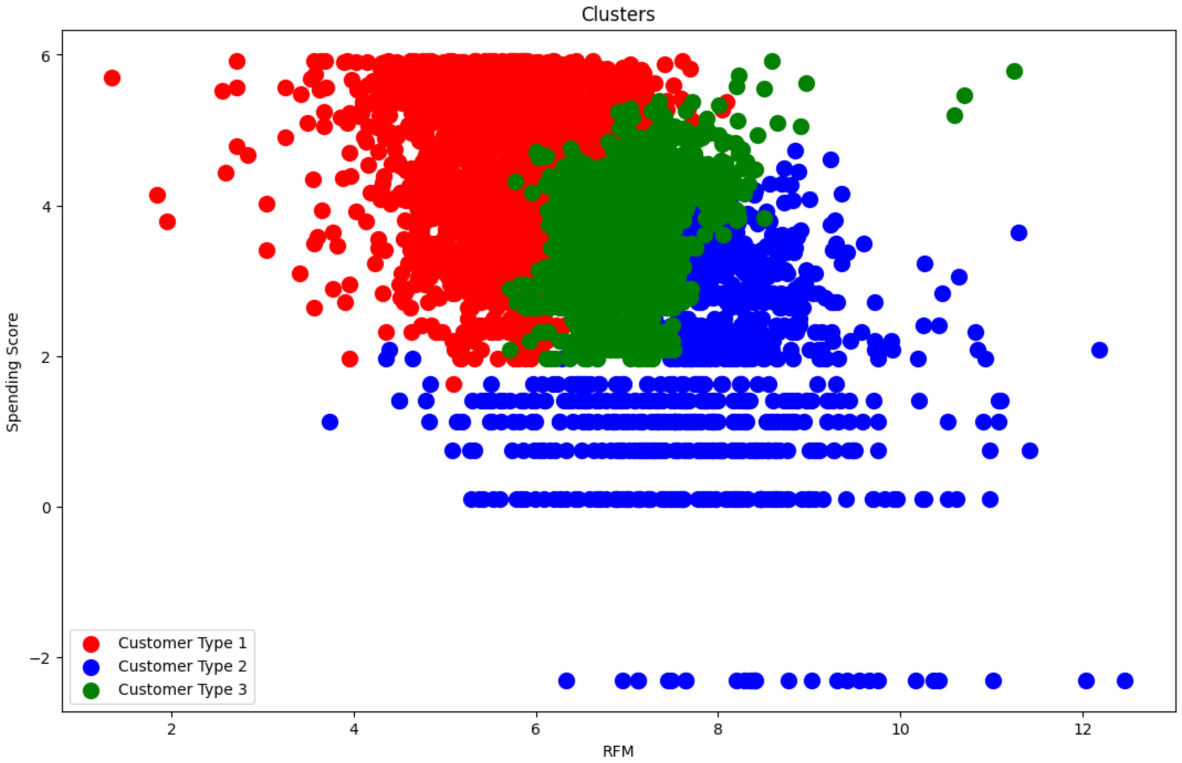
DBScan Clustering



- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is the most well-known density-based clustering algorithm
- Unlike k-means, DBSCAN does not require the number of clusters as a parameter. Rather it infers the number of clusters based on the data, and it can discover clusters of arbitrary shape (for comparison, k-means usually discovers spherical clusters).
- DBSCAN categories the data points into three categories
  - Core Points - (Steel blue points in scatter plot)
  - Border Points - (Green points in scatter plot)
  - Outliers - (Dark Blue points in scatterplot)
- (As we can see from the DBscan Visualization)

Agglomerative Clustering





- For number of cluster = 3 as per dendrogram.

Customer Segmentation

- Platinum customers belong to cluster : [3 4]
- Big Spenders belong to cluster : [2 3 4] High
- Spend new Customers belong to cluster : [1 2]
- Lowest-Spending Active Loyal Customers belong to cluster : [1]
- Recent Customers belong to cluster : [3 1 4]
- Good Customers Almost Lost belong to cluster : [2 4]
- Churned Best Customers belong to cluster : [2]
- Lost Cheap customers belong to cluster : [0 2 1]

Conclusion

- Customer are segmented in 5 different clusters.
- Customer belonging to cluster 0 have lowest RFM value and RFM Score and they belong to Lost Cheap Customer segment.
- Customer belonging to cluster 1 have high recency value and low monetary and frequency value and most of them fall into Lowest-Spending Active Loyal Customer and Recent Customer segments.
- Customer belonging to cluster 2 shows the same characteristics as customer in cluster 0, with low RFM value and RFM score, and they also belong to Lost Cheap Customer and others segments.
- Customer belonging to cluster 3 have very high RFM value and RMF score, most of the customer belong to Platinum Customer segment and some to Big Spenders segment.
- Customer belonging to cluster 4 have good RFM value and RFM score, most of the customer belong to Platinum Customer and Big Spender segments and some also belong to Recent Customer with high Monetary Value.