**Capstone Project – “Online Retail Store Capstone Project”**

**Table of Contents**

1. Problem Statement

2. Project Objective

3. Data Description

4. Data Pre-processing Steps and Inspiration

5. Choosing the Algorithm for the Project

6. Motivation and Reasons For Choosing the Algorithm and Assumptions

7. Model Evaluation and Techniques

8. Inferences from the Same

9. Future Possibilities of the Project

10. Conclusion

11. References

1. **Problem Statement**

An online retail store is trying to understand the various customer purchase patterns for their firm; you are required to give enough evidence based insights to provide the same.

**2. Project Objective**

1. Using the above data, find useful insights about the customer purchasing history that can be an added advantage for the online retailer.

2. Segment the customers based on their purchasing behavior.

**3. Data Description**

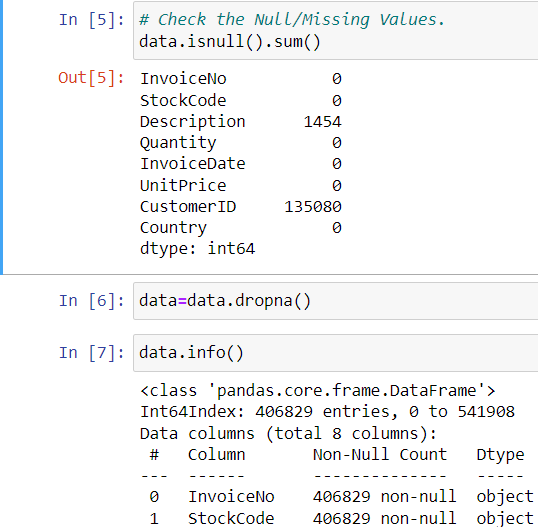
The online\_retail.csv contains 387961 rows and 8 columns.

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Invoice | Invoice number |
| StockCode | Product ID |
| Description | Product Description |
| Quantity | Quantity of the product |
| InvoiceDate | Date of the invoice |
| Price | Price of the product per unit |
| CustomerID | Customer ID |
| Country | Region of Purchase |

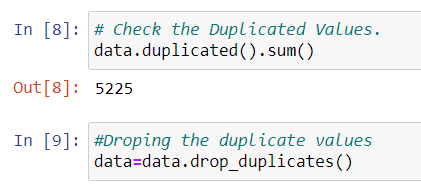
**4. Data Preprocessing Steps and Inspiration**

The preprocessing of the data included the following steps:

* 1. **After loading the dataset, performing the data warling to clear the data for further analysis. Like, check null values and drop the same.**

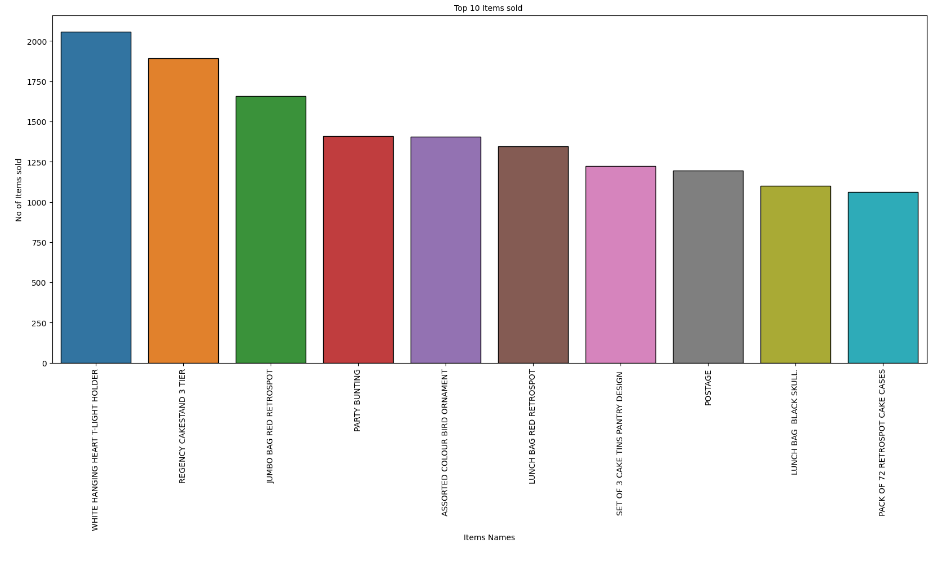


* 1. **Duplicated values and removing the same.**

****

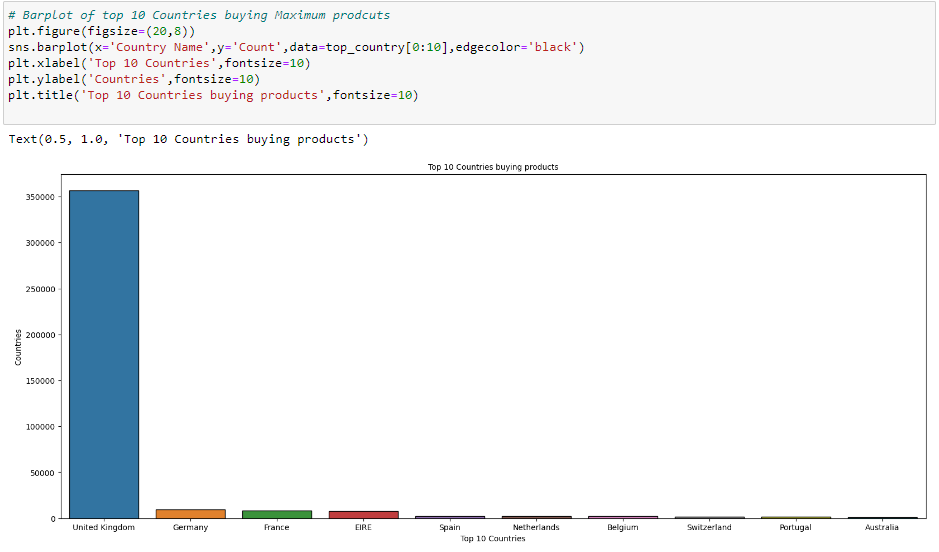
* 1. **Check the top 10 product was sold and plots the graphs.**



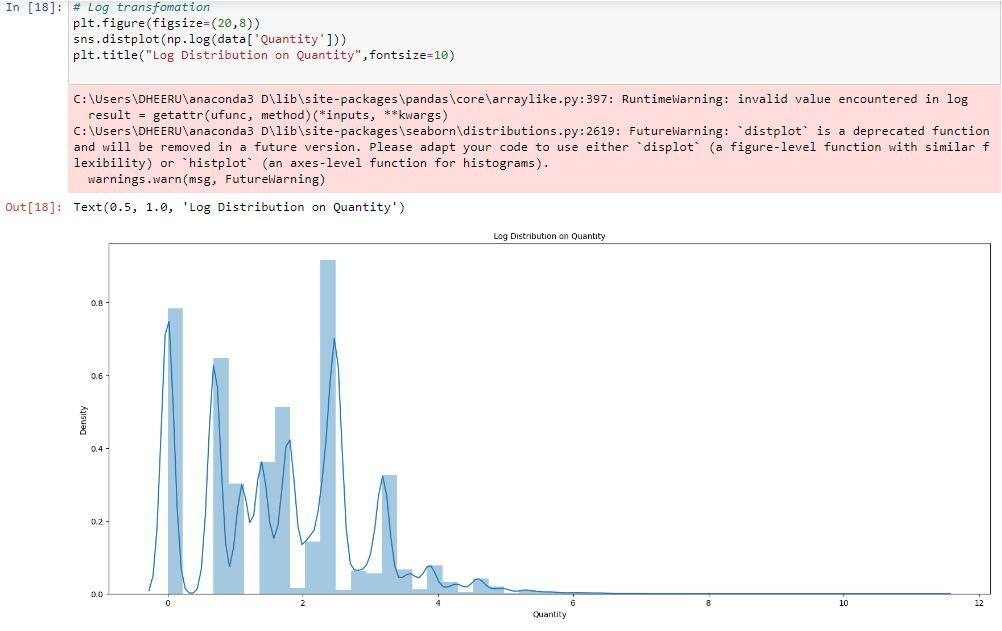
****

* 1. **Check the top 10 Countries buying the maximum products and plot the graphs.**

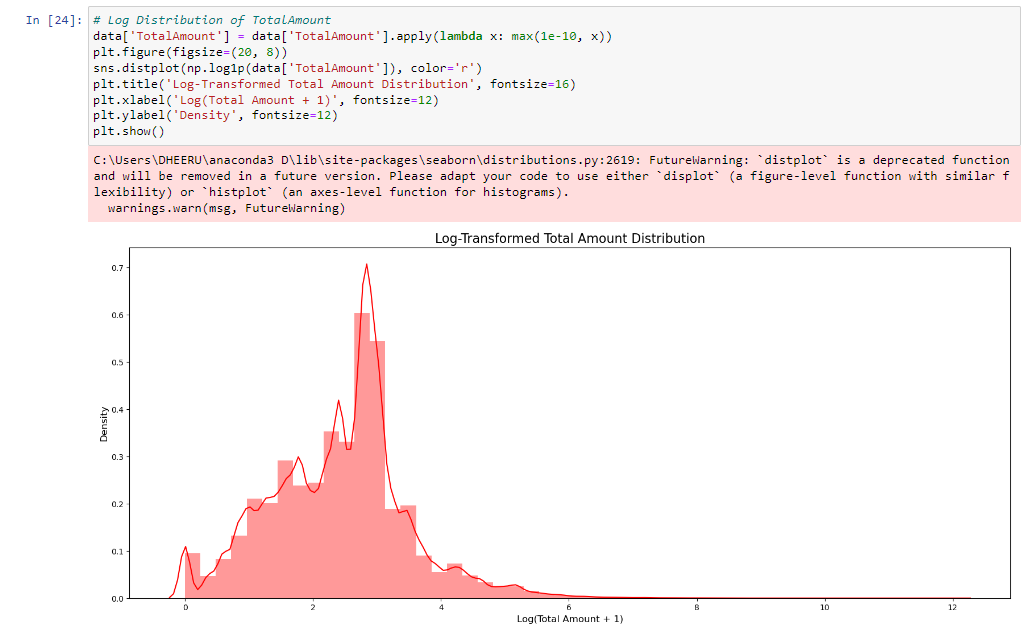
****

****

* 1. **Distribution plot of Quantities by Log Transformation.**

****

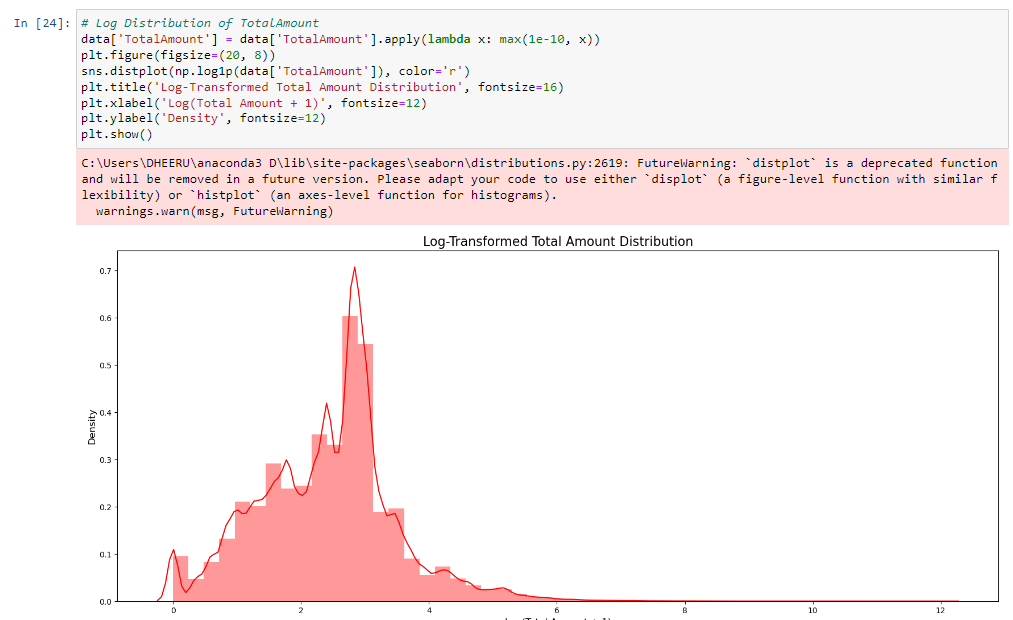
* 1. **Distribution plot of Quantities by Log Transformation**.

****

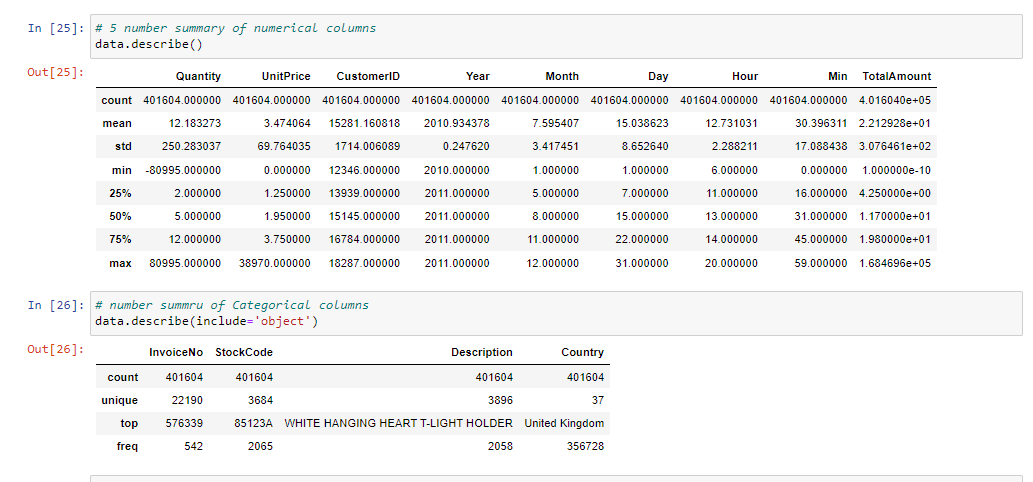
* 1. **Performing the Feature Engineering on InvoiceDate Columns by converting into Day, Month, Year, Hours and Minutes.**



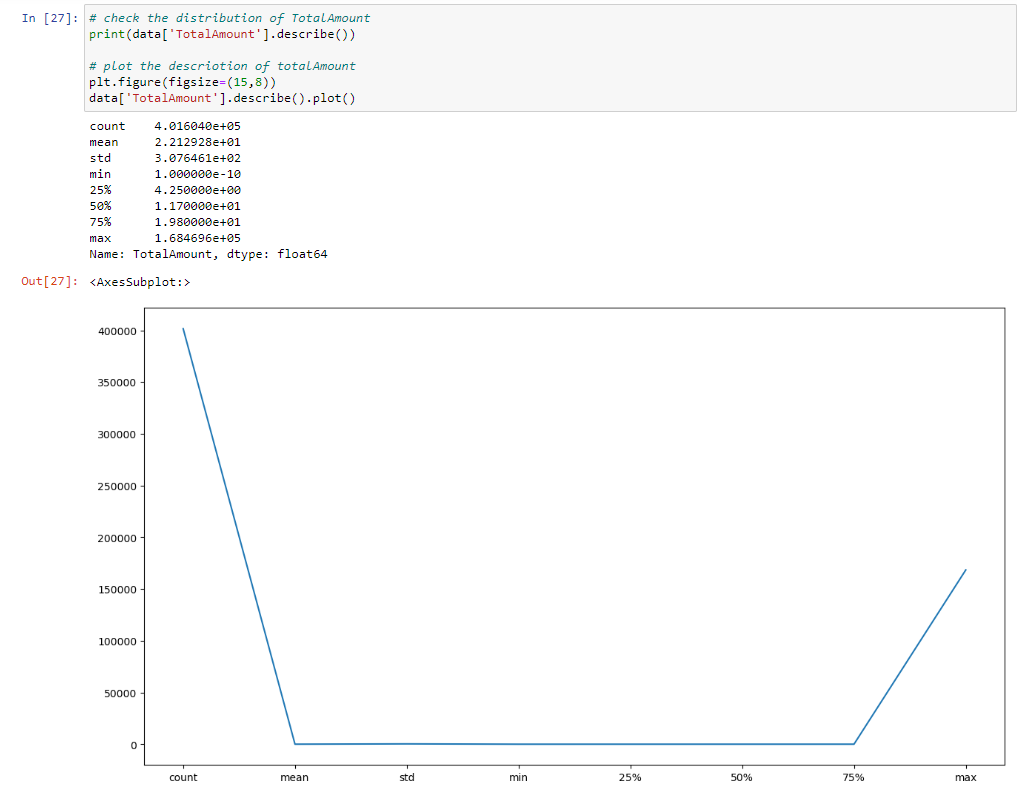
* 1. **Distribution plot of Quantities by Log Transformation**



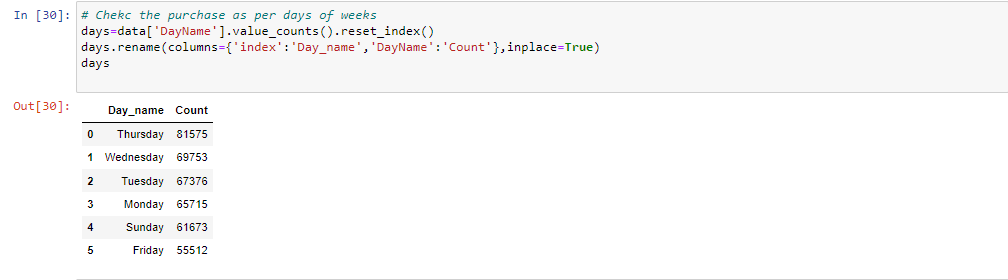
* 1. **Check the 5 number summaries on numerical and categorical features**.

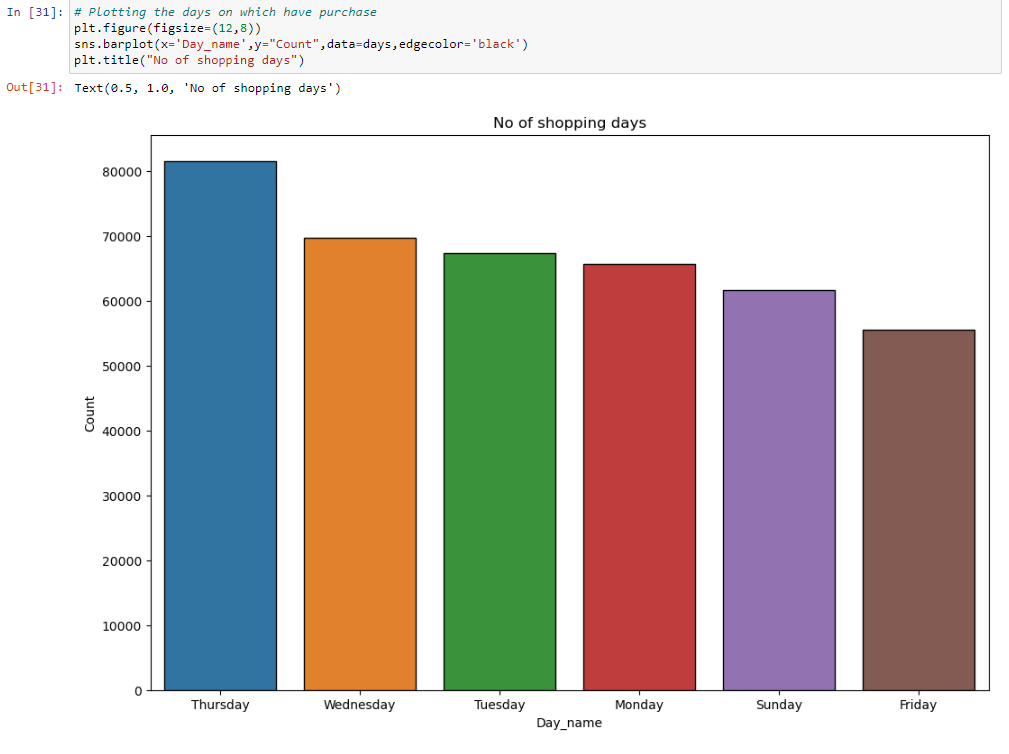
****

**# And distribution of Total Amount.**

****

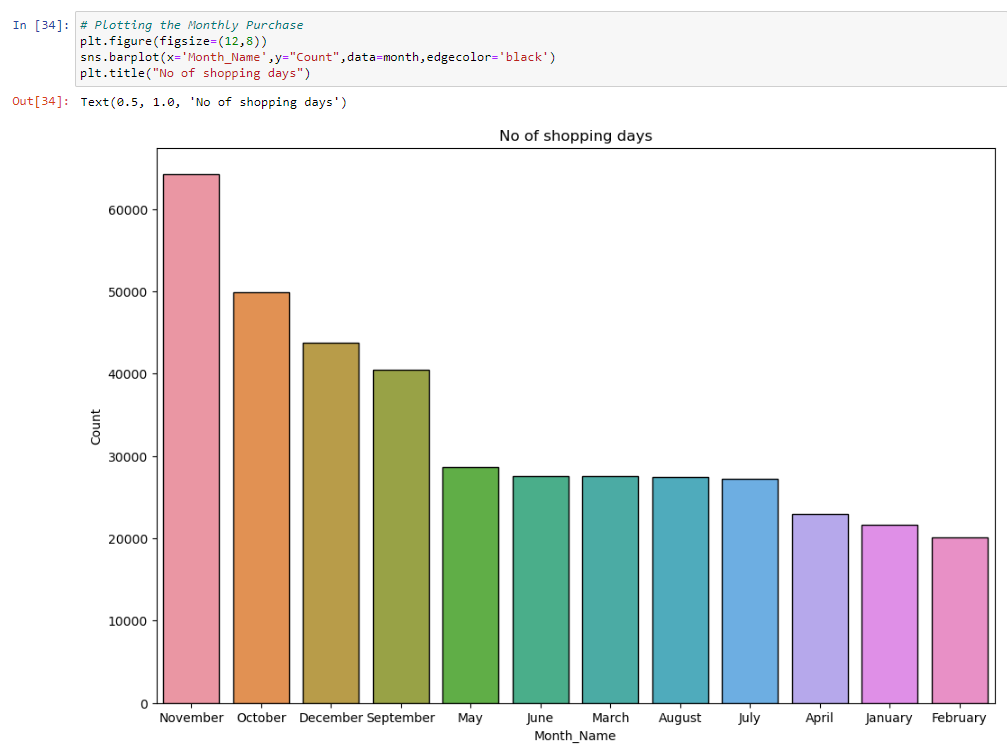
* 1. **Checking the purchase behavior of customer on Week basis.**



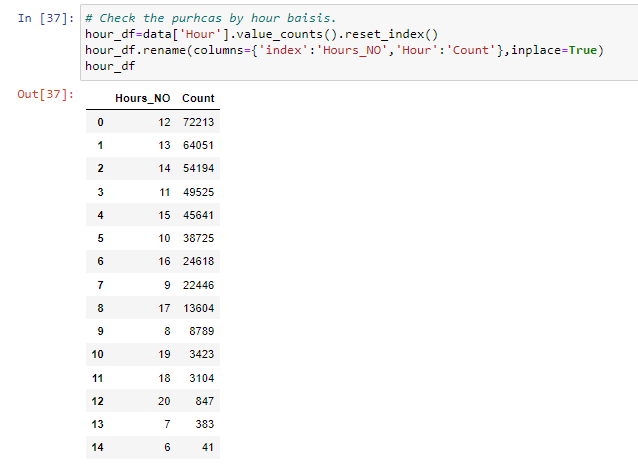


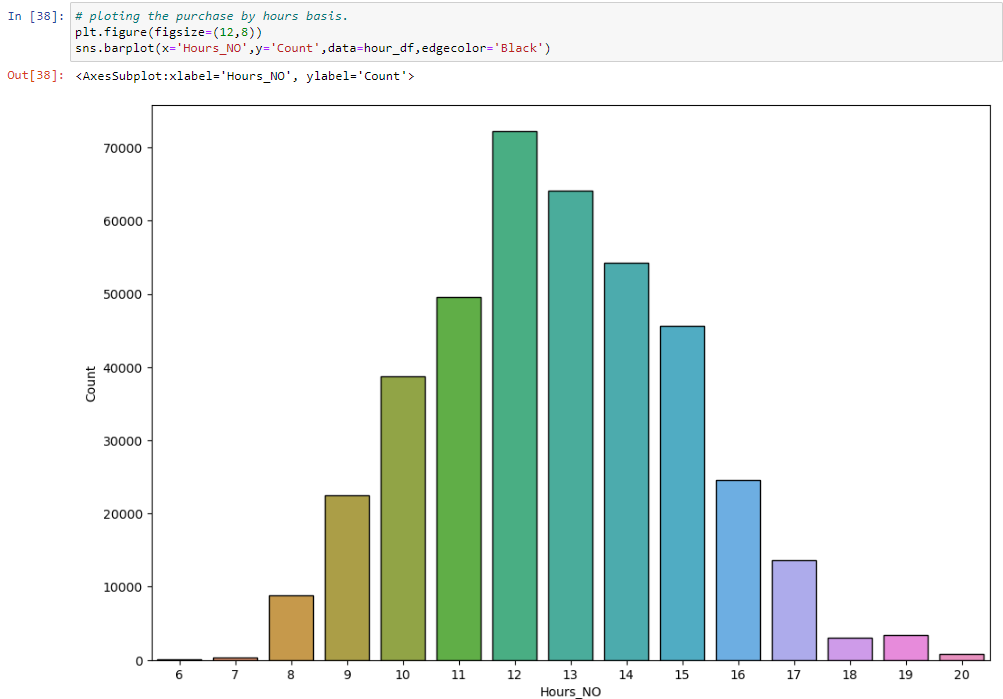
* 1. **Checking the purchase behavior of customer on Monthly basis.**





* 1. **Checking the purchase behavior of customer on Hourly basis.**

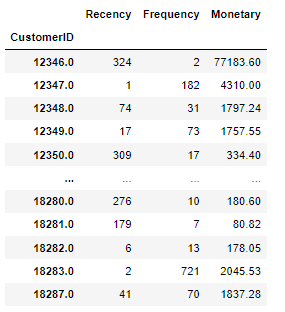
****

****

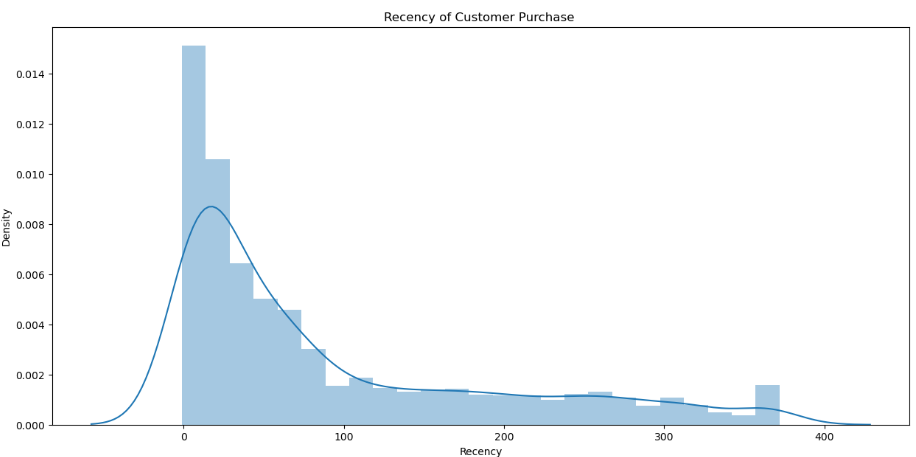
**5. Choosing the Algorithm for the Project**

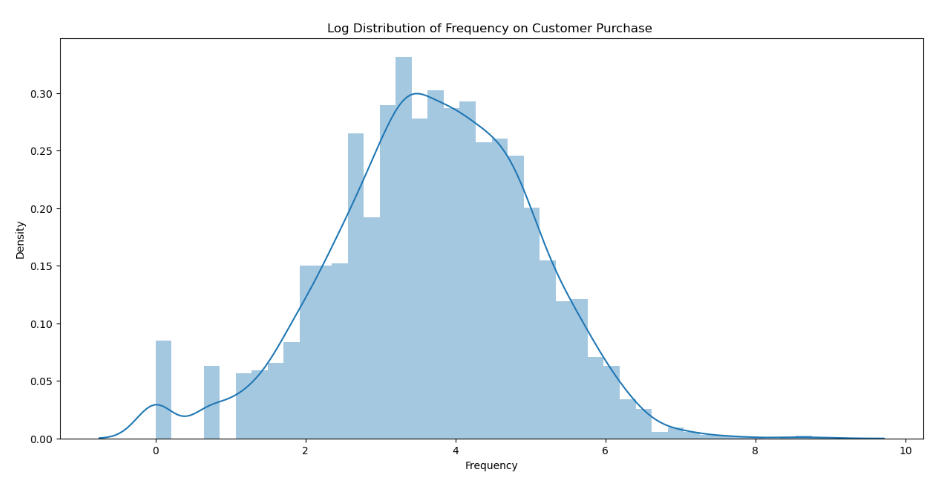
1. **RFM model (Recency, Frequency and Monetary value)**

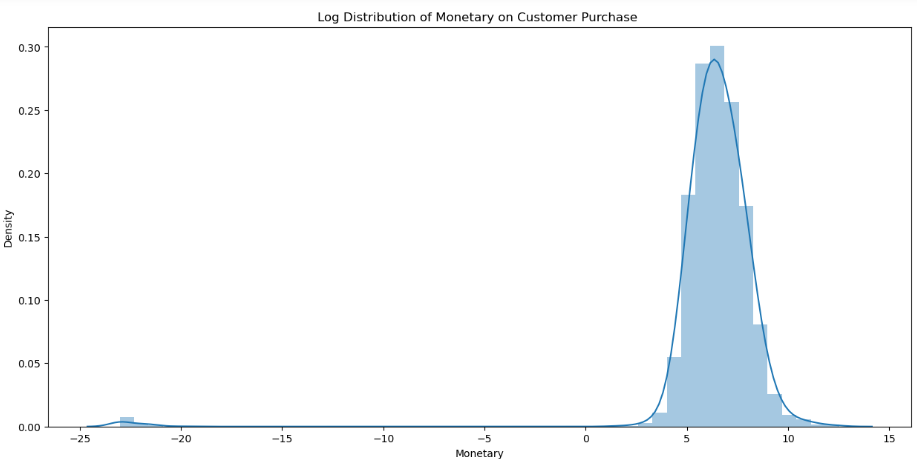
RFM is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors:. Frequency: How often a customer makes a purchase. Monetary Value: How much money a customer spends on.



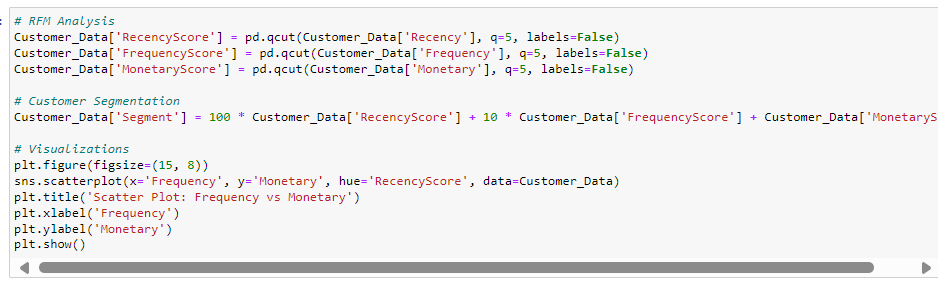
**Distribution plot on Recency, Frequency and Monetary.**

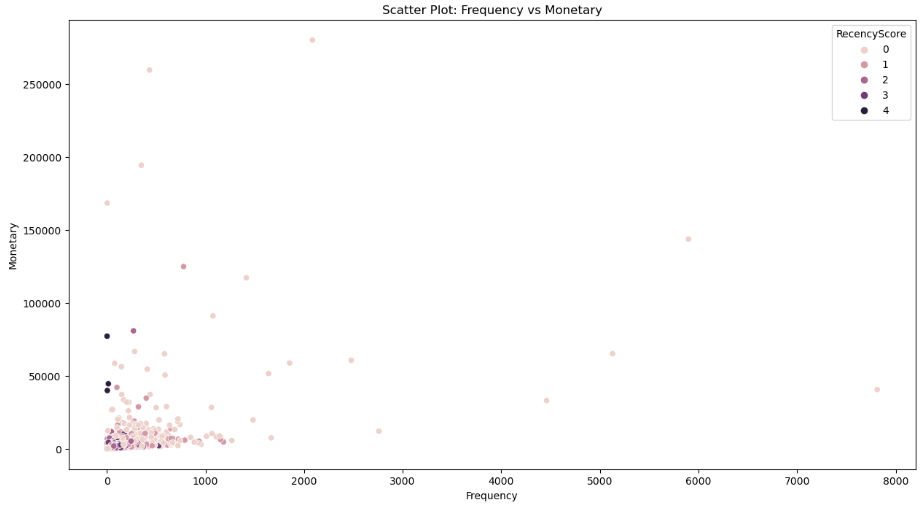
****

****

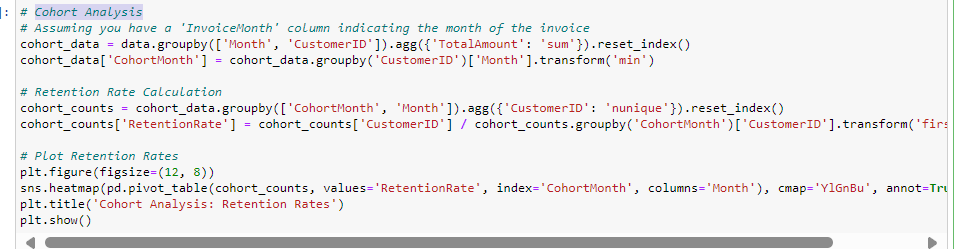
****

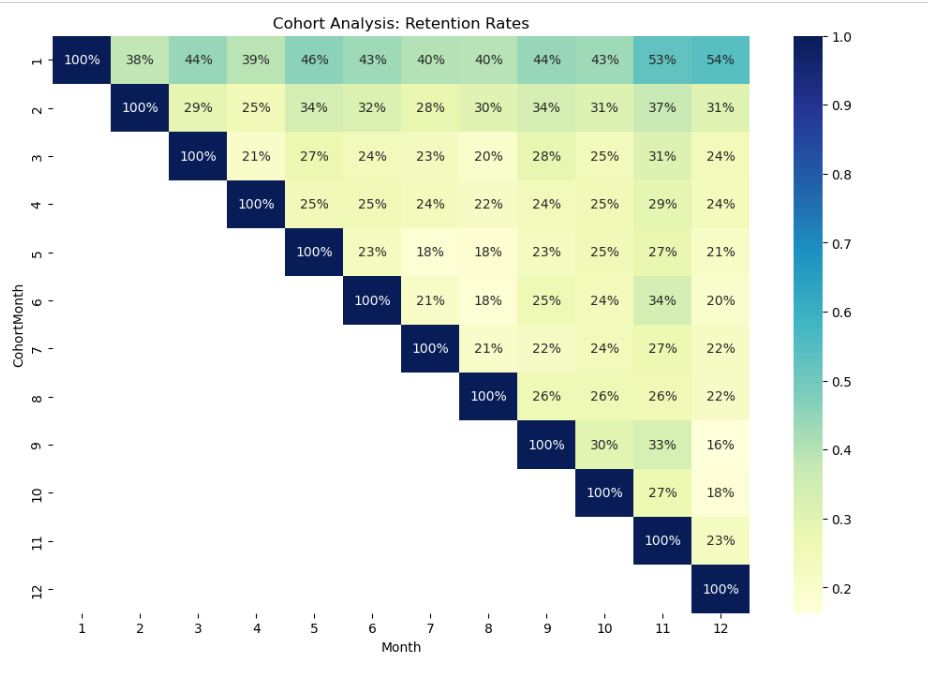
1. **RFM Analysis by showing the scatter plot.**

****

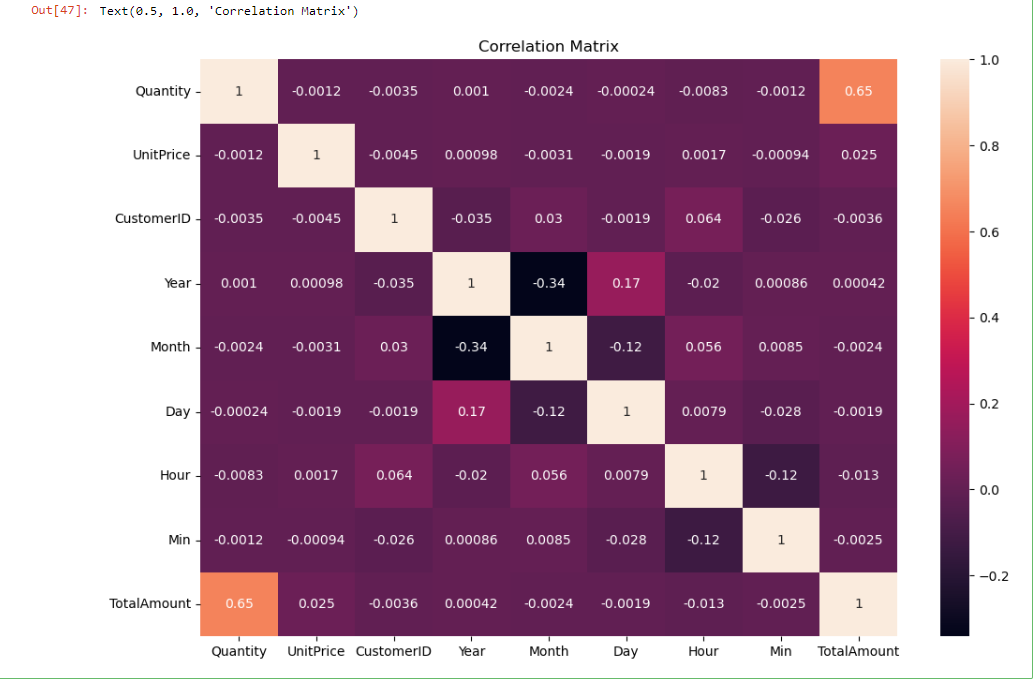
****

1. **Cohort Analysis and heat map.**

****

****

1. **Correlation between various columns.**

****

# 6. Motivation and Reasons for Choosing the Algorithm & Assumptions

RFM stands for Recency, Frequency, and Monetary Value, and it is a method used in marketing and customer relationship management to analyze and categorize customers based on their transaction behavior. Each of the three components represents a key aspect of a customer's purchasing behavior:

Recency (R):

Refers to how recently a customer has made a purchase. Typically measured as the time elapsed since the customer's last purchase. A lower recency value indicates a more recent purchase, which is often associated with higher engagement. Frequency (F):

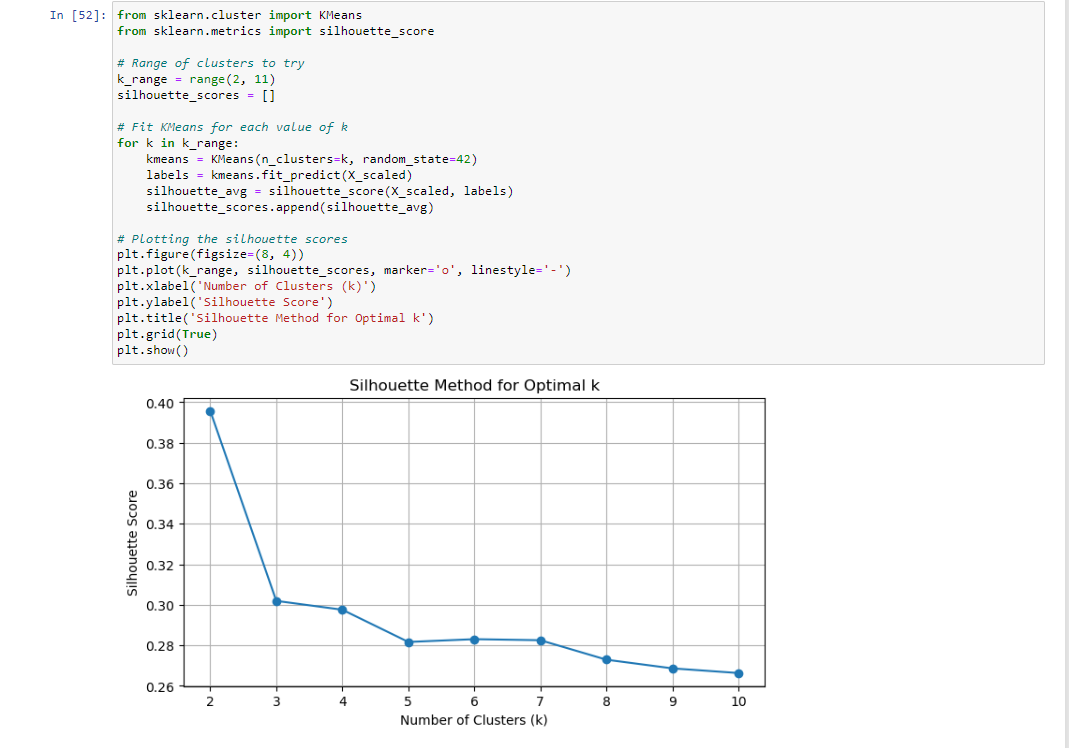
Represents how often a customer makes a purchase. Calculated as the total number of transactions or orders made by the customer. Higher frequency values suggest a more loyal and engaged customer. Monetary Value (M):

Represents the total monetary value of a customer's purchases. Calculated by summing the monetary value of all transactions made by the customer. Indicates the customer's overall spending or contribution to revenue.

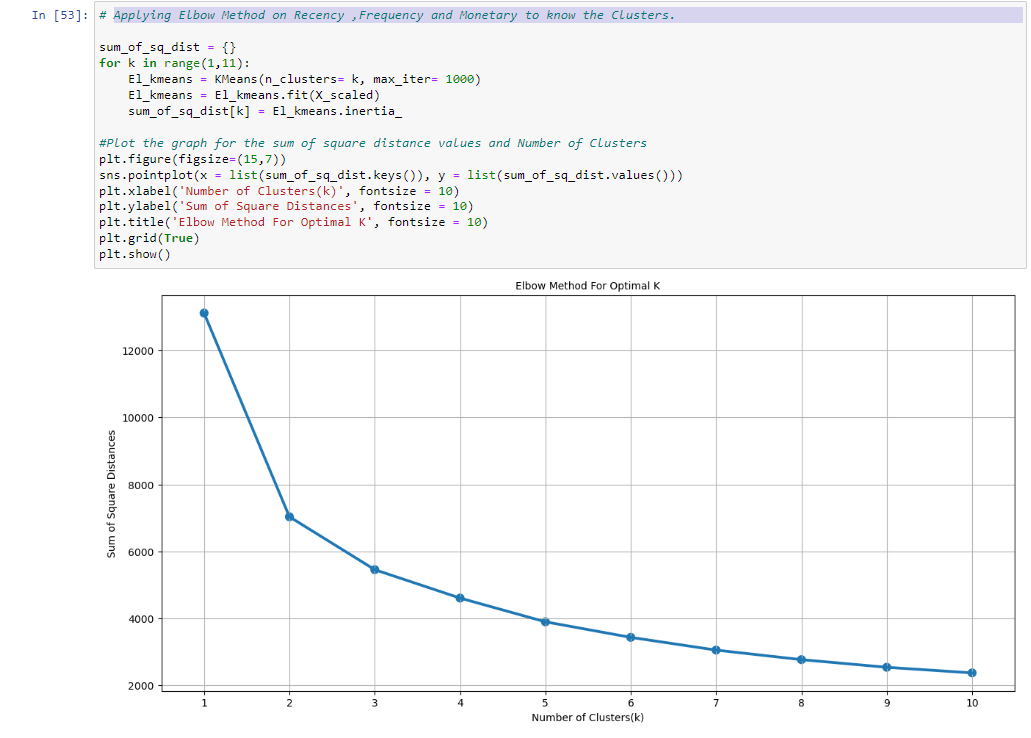
RFM analysis helps businesses understand and target their customers more effectively, allowing for personalized marketing efforts and improved customer retention. It is commonly used in combination with data analytics and machine learning techniques for more sophisticated customer segmentation and prediction.

**7. Model Evaluation and Techniques**

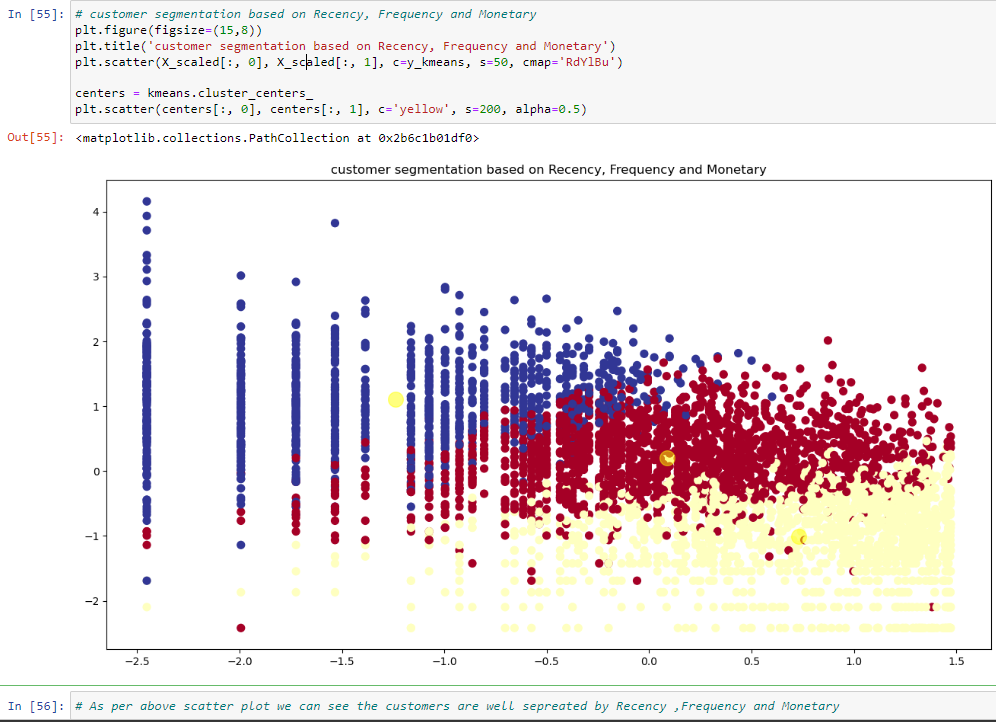
# Applying Silhouette Method on Recency, Frequency and Monetary

****

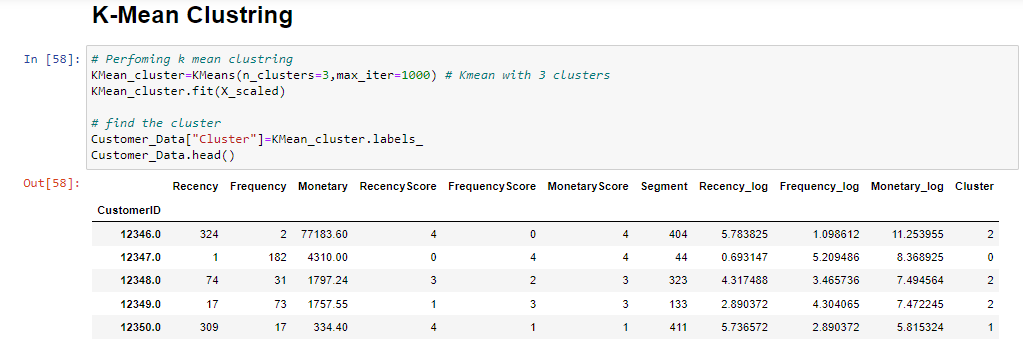
# Applying Elbow Method on Recency, Frequency and Monetary to know the Clusters.

****

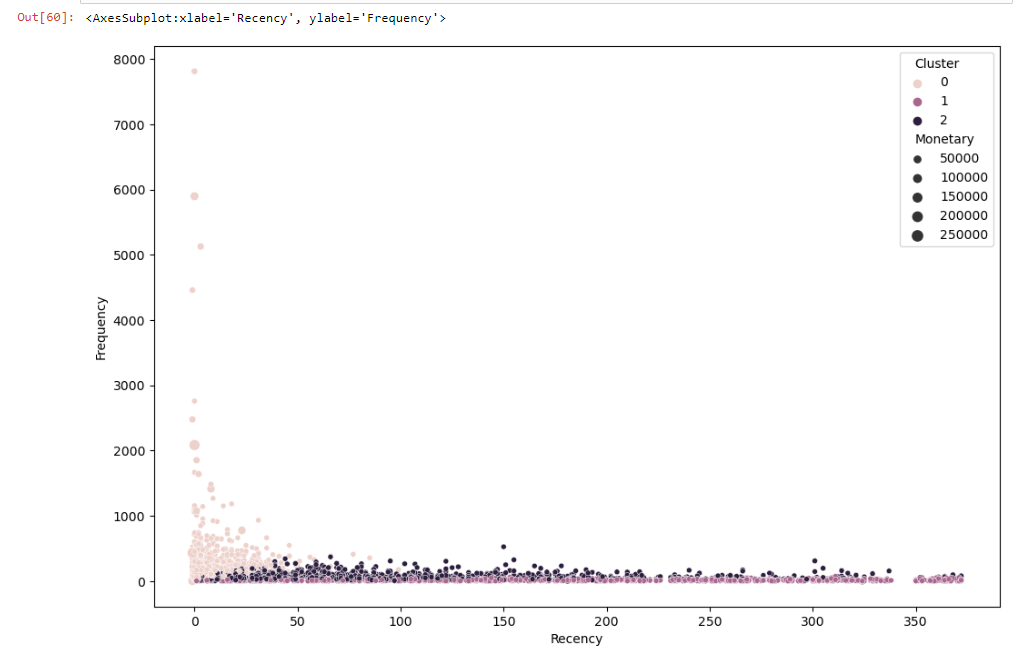
**# Customer segmentation based on Recency, Frequency and Monetary.**

****

**# Applying the K-Mean CLusting and Dendrogram for clustring.**

****

1. **Creating Scatter plot to understand the cluster.**

****

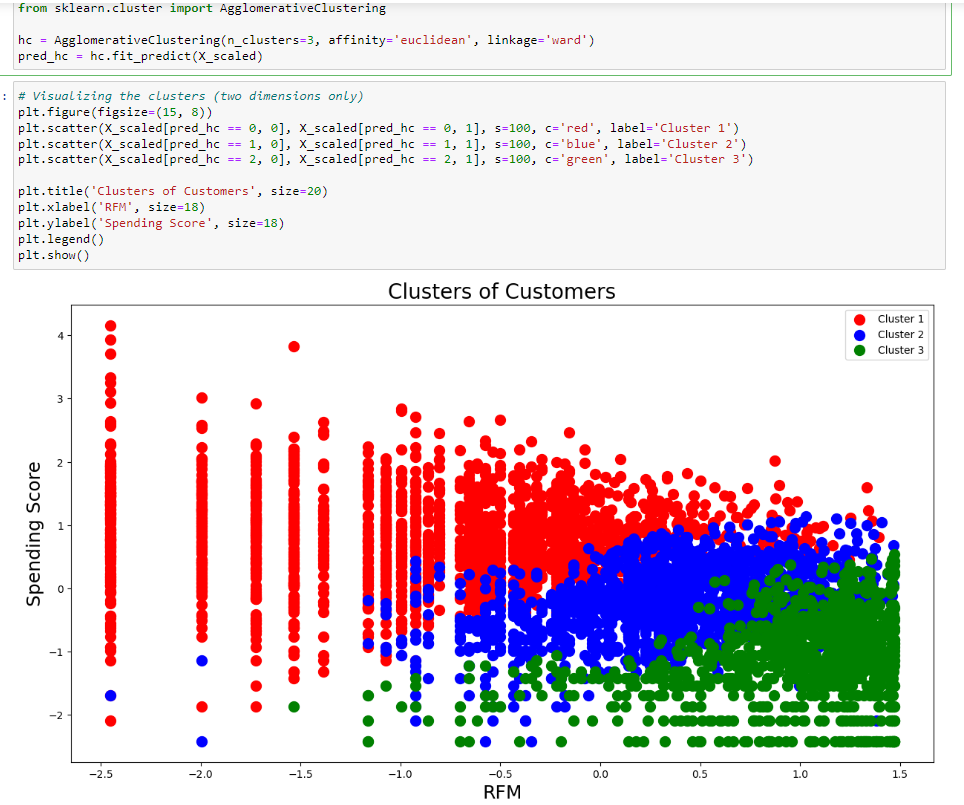
1. **Dendrogram to find the optimal numbers of clusters.**

****

**8. Inferences from the Same**

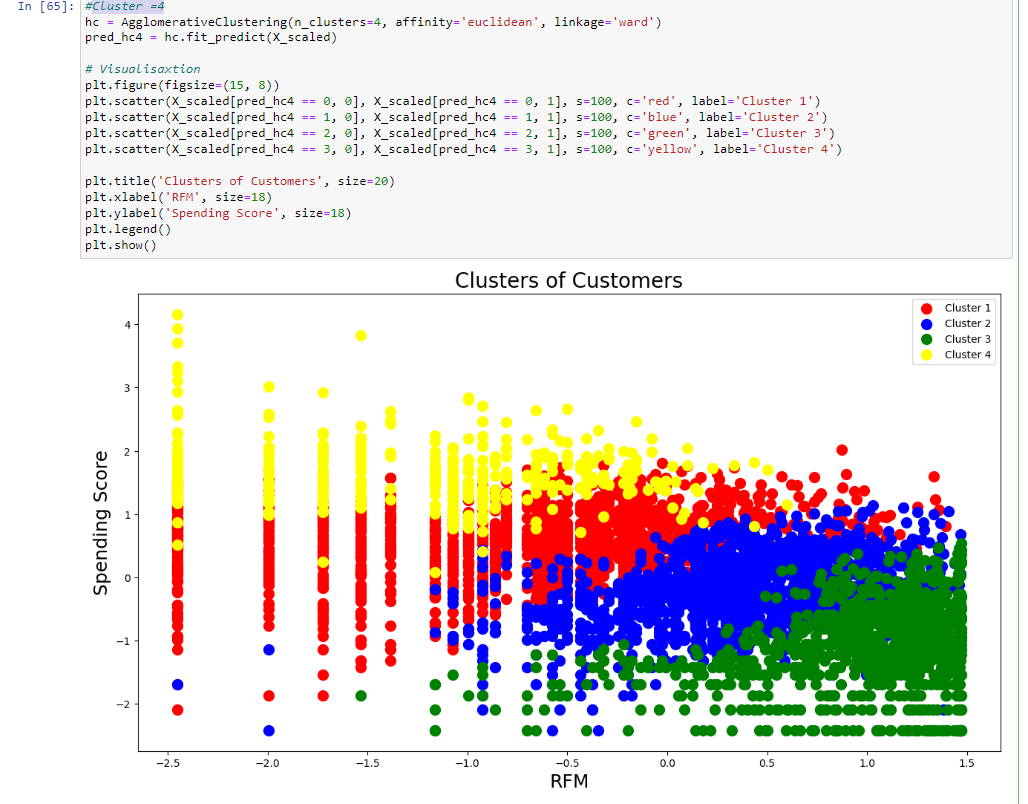
**The no of cluster will numbers of vertical lines which intersecting by the lines using the threshold of 50**

**# fitting the Hierarchical clustering into the datasets and plotting the scatter plot.**

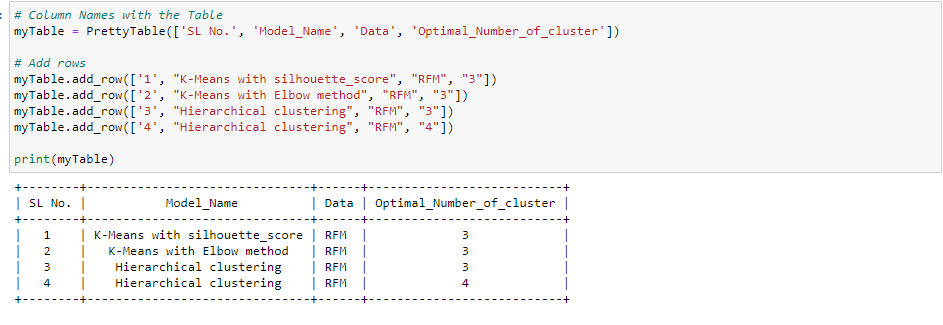
****

#“after applying different clustering algorithm’s to our datasets, we get the 3 optimized clusters, which is overlapped to each other, we can increase the cluster to 4”

**# Applying the Cluster 4 on our dataset.**



**# Creating a Table with the cluster details**.



**9. Future Possibilities of the Project**

Real-time Customer Segmentation:

Implement online clustering algorithms that can adapt to changing customer behavior in real-time. Update customer segments dynamically as new data points are received. Dynamic Personalization:

Provide real-time personalized experiences for users based on their current behavior and preferences. Adapt content, recommendations, and offers dynamically as users interact with the platform. Incremental Learning:

Explore incremental learning techniques to continuously update cluster models with new data without retraining from scratch. Ensure that the model's performance improves over time as more data becomes available. Adaptive Marketing Strategies:

Develop marketing strategies that can adapt on-the-fly based on the latest cluster assignments.

**10. Conclusion**

In this analysis we have performed various steps to perform customer segmentations. we stated with data warling in which we are tried to handle missing values, duplicated, performed feature engineering. After that we have done some EDA and Visualizations to get the insights from the data.

We have also performed the RFM for our entire customer base.

RFM analysis analysis helps us to know various questions with respect to their customers and this help us to make marketing strategies for their customers.

Cluster profiling the average of Recency, frequency and monetary values for each customer segment.

K-means algorithm to segment our customer in various clusters having similar similarity. Also noted more cluster gives better results.

**11. References**

Google and Intellipaat Learning’s