DETAILED REPORT

The OnlineRetail dataset from the UCI Machine Learning Repository is a transactional dataset that contains online retail transactions that occurred between 01/12/2010 and 09/12/2011 for a UK-based online retailer. The dataset includes transactions from customers all over the world, but the majority of transactions are from the UK.

To analyze and understand the data we used many techniques. But we majorly follow four major Machine Learning algorithms

RFM Analysis: This technique involves segmenting customers based on their Recency, Frequency, and Monetary value of their purchases.

Clustering: This technique involves grouping customers into clusters based on their similarities and differences in purchasing behavior and demographics.

Dimensionality reduction:This is a machine learning (ML) or statistical technique of reducing the amount of random variables in a problem by obtaining a set of principal variables.

RandomForestClassifier: Random Forest Classification (RFC) is a machine learning algorithm that can be used in customer segmentation to predict the class or segment to which a customer belongs. It is a type of ensemble learning method that uses a combination of decision trees to make predictions.

Performing EDA on the data:

After importing the “Online Retail” dataset, we found that the columns “CustomerID” and “Description” are having null values. We found a unique number of elements in each column which is used to identify the most common values in the dataset and outliers(if any).

Preprocessing the data: Firstly, we filtered out the rows where the CustomerID column is not null. Then we perform some operations on the data like removing the rows : where the ‘Quantity’ column is negative or zero, where the ‘UnitPrice’ column is zero, where the ‘InvoiceNo’ column have ‘C’ as the starting character and corresponds to cancelled transactions (also assigned them to a new data frame called canc_trans), where the first four characters of the StockCode column are not digits. We then also dropped the duplicate rows. We then prepared a finalized dataset namely final_data and checked the required data. Also, we created the data frame tempo_data by removing the rows with negative Quantity and zero UnitPrice from the canc_trans data frame which is used for our interpretation.

Interpretations:

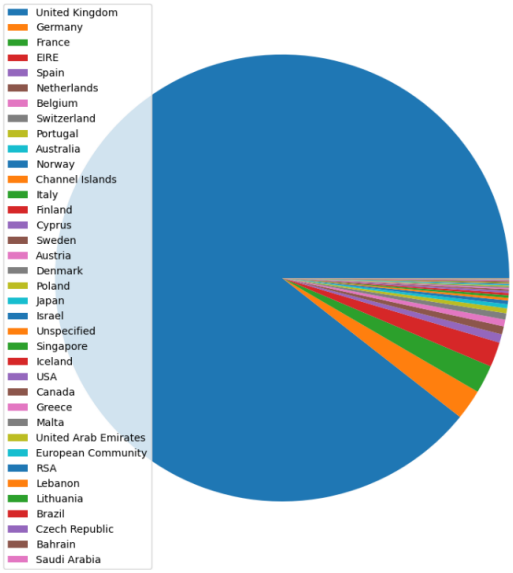
- The percentage of canceled transactions : 1.643%.
- The country with the highest transactions is the United Kingdom with 353985 Transactions.
- The country with highest no of canceled transactions is the United Kingdom with 7533 Transactions.
- The Customer with CustomerID 17841.0 has the highest transactions with the count of 7838 is from the United Kingdom Country.
- The most bought item was WHITE HANGING HEART T-LIGHT HOLDER was from the United Kingdom.
- The most costliest item was bought by the customer with CustomerID 15098.0 of cost 649.5.
- The most costliest item was bought by the customer with CustomerID 15632.0 of cost 649.5.
- The most no of Products were brought by customer with CustomerID 17389.0 of Quantity 80995.

Adding new columns: We have added a few columns namely ‘Month’, ‘Year’, ‘WeekDay’, ‘Month_Year’, ‘Hour’ and ‘Revenue’ for a better interpretation and found the unique values for the newly added columns.
Here Column Revenue is derived by multiplying the UnitPrice with Quantity. And the other functions are derived by using python default

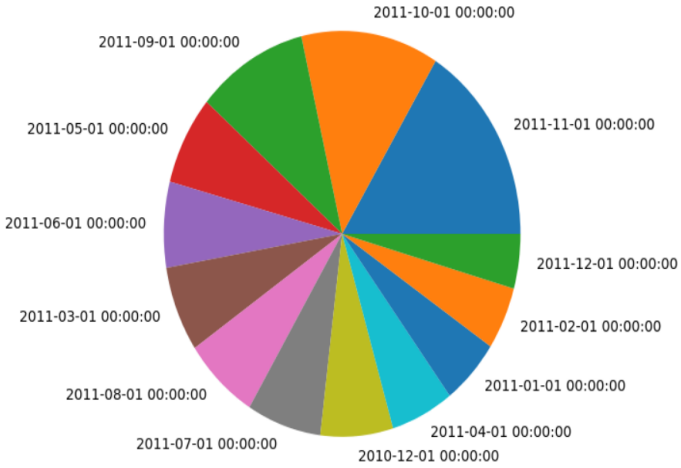
function datetime function. We later convert the names in the columns of Month using the Converter function.

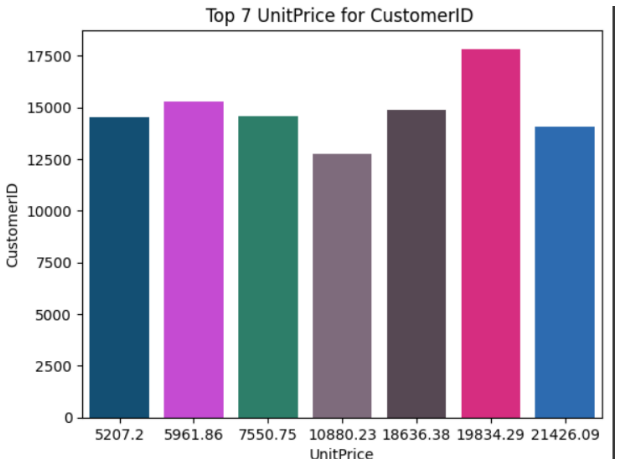
Visualizations:

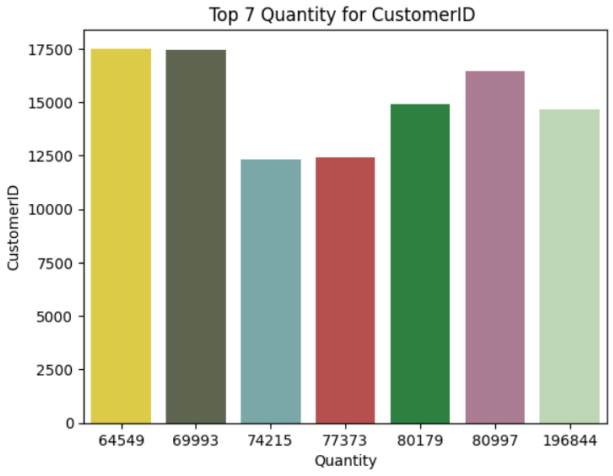
Distribution of the ‘Country’ column:

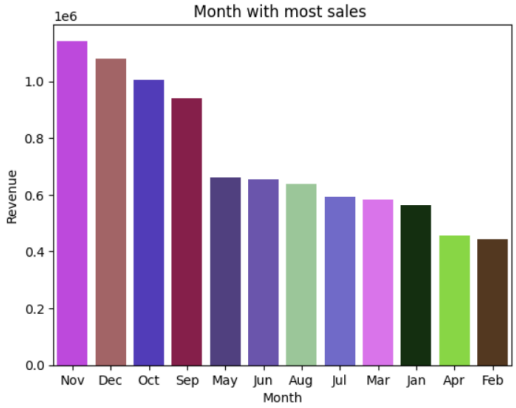


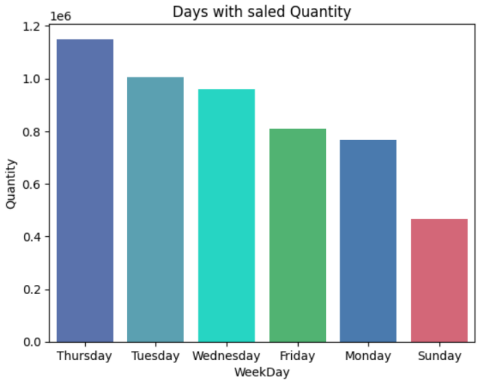
Distribution of the ‘Month_Year’ column:

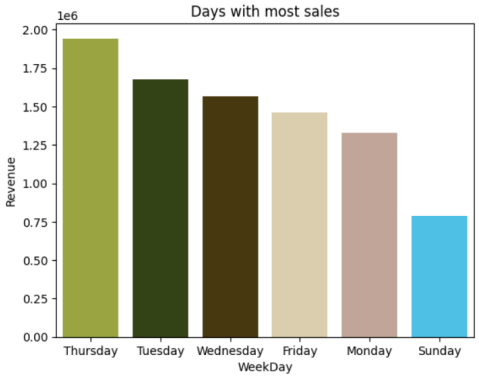


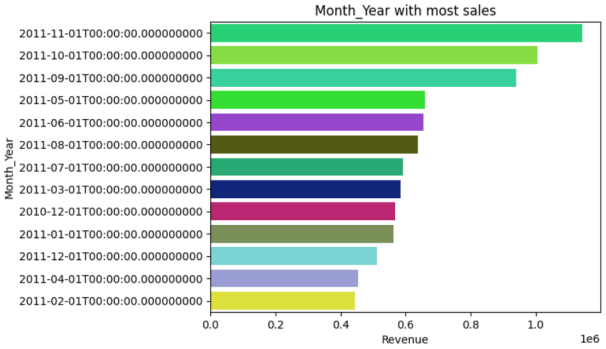


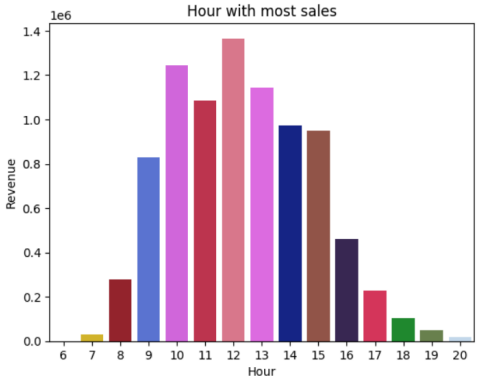


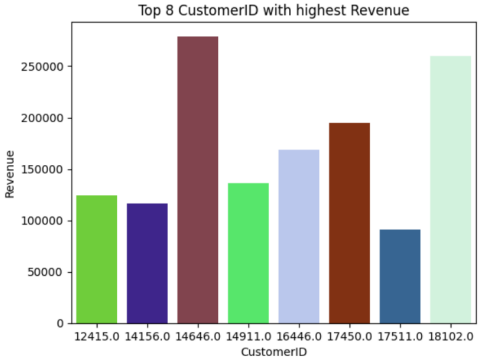




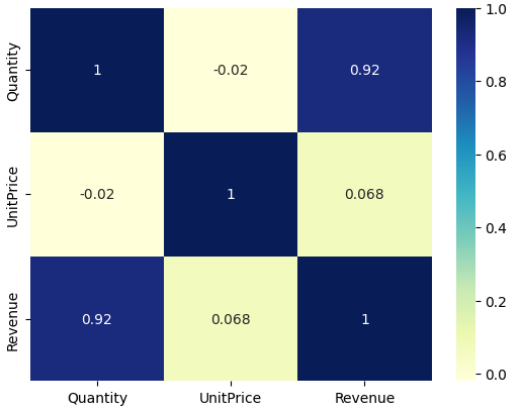








Correlation Matrix:



Customer Segmentation:
Customer segmentation is a process of dividing a customer base into groups or segments based on their similarities or shared characteristics using statistical and machine learning techniques. This allows businesses to understand and target specific groups of customers with more relevant marketing strategies, products, and services.

What is RFM Analysis??
RFM model is a widely used customer segmentation technique in the field of marketing and e-commerce. RFM stands for Recency, Frequency, and Monetary value, which are three key factors used to segment customers based on their purchasing behavior.

- Recency: How recently a customer has made a purchase. Customers who have made a purchase more recently are considered more valuable than those who have not made a purchase in a long time.
- Frequency: How frequently a customer makes purchases. Customers who make purchases more frequently are considered more valuable than those who make purchases less often.
- Monetary Value: How much money a customer has spent on purchases. Customers who have spent more money are considered more valuable than those who have spent less money.

Using RFM analysis, businesses can segment customers into different groups based on their scores for each of the three factors.

Why RFM Analysis??
RFM analysis is better than other methods because of its ease of understandability.It is widely used because of its simplicity and effectiveness in identifying high-value customers, predicting future customer behavior, and improving marketing campaigns.

What is Clustering??
Clustering is a type of customer segmentation technique that involves grouping customers based on their similarities in behavior, demographics, or other characteristics. In clustering, the goal is to

identify groups of customers who share common traits and behaviors, which can help businesses create targeted marketing campaigns and improve customer engagement.

Why Clustering??
Clustering is a popular technique in customer segmentation because it enables businesses to group customers based on similarities in behavior, preferences, or other characteristics. By clustering customers into different groups, businesses can gain insights into customer behavior, identify high-value customers, and develop targeted marketing campaigns to improve customer engagement and satisfaction.

What is Dimensionality Reduction??
Dimensionality reduction is a technique used in customer segmentation to reduce the number of features or variables used to describe customer behavior. It involves transforming high-dimensional data into a lower-dimensional space while retaining the most important information.

Why Dimensionality reduction??
There are two main reasons why dimensionality reduction is used in customer segmentation:

- Simplification: In many cases, customer behavior is described using a large number of variables, making it difficult to interpret and visualize the data. By reducing the number of variables, dimensionality reduction can simplify the data and make it easier to understand.
- Improved Performance: High-dimensional data can be difficult to process and analyze, and can lead to overfitting in machine learning models. By reducing the number of variables, dimensionality reduction can improve model performance and reduce the risk of overfitting.

In this Dimensional Reduction we use PCA technique on our data.

What is Random Forest Classifier??

In customer segmentation, RFC can be used to predict the segment or cluster to which a customer belongs based on their demographic, transactional, and behavioral data. This can be useful for businesses to identify the most important features that contribute to a customer's segment, and to better understand their behavior and preferences.

Why Random Forest Classifier??

There are four main reasons why RFC is used in customer segmentation:

- Accurate predictions: RFC is a machine learning algorithm that can accurately predict the class or segment to which a customer belongs. It uses a combination of decision trees to make predictions, which reduces overfitting and improves accuracy.
- Feature selection: RFC can help businesses identify the most important variables that contribute to customer segmentation. By identifying the most important variables, businesses can focus on the most relevant customer attributes and design targeted marketing strategies to better engage their customers.
- Scalability: RFC can be applied to large datasets, which is particularly useful in customer segmentation where businesses need to analyze vast amounts of data.
- Interpretability: The results of RFC can be easily interpreted, allowing businesses to understand which variables are most important in predicting customer segmentation.

Customer Segmentation:

We begin our customer segmentation by building another dataframe for analysis. We need to built three columns
Recency,Frequency,Monetary
Recency columns can be built by taking the count of the date from a reference data. In our analysis we took the reference date as (2011,11,12). And we fill the recency column by subtracting the days difference between the reference date and first purchase date by grouping the CustomerID column.
Frequency columns can be built by taking the count of the no of Invoice No generated in the given set of days again by grouping CustomerID columns.

Monetary columns can be generated by taking the help of revenue columns in the dataset and grouping them using CustomerID's.
We next generate quartiles using the default qcut function in date and divide our data into 5 parts and naming them from 1 to 5. Based on observing the values in RFM.
We create new columns in the RFM dataset named RecencyScore,frequencyScore,MonetaryScore that store the ranking of the respective columns.
We then create our RFM score column in our dataset using the scores obtained from RFM columns. We convert our scores to strings and add them and store them as strings.
We then create a map for regex conversion of our dataset.

For example:
r'[1-2][1-2]': 'Inactive_Customers',
We are converting the regular expressions that can have 11,12,21,22 values and naming it as Inactive_Customers.

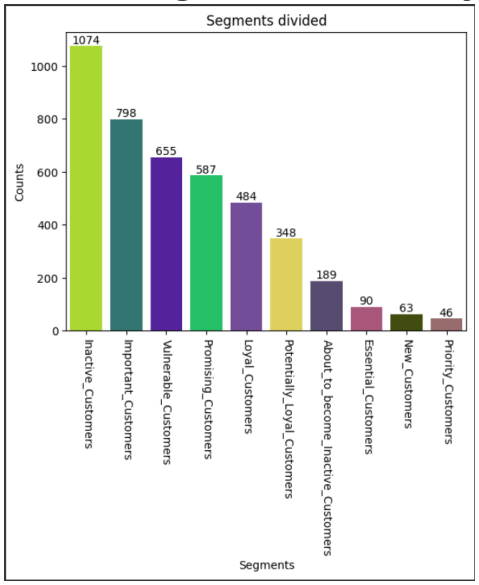
We have used 10 different categories for our CustomerSegmentation
We prioritize only RecencyScore,FrequencyScore because MonetaryScore may be equal to some customers but they may have different behavioral patterns on buying things. So we mostly keep Monetary Score out of our mapping.

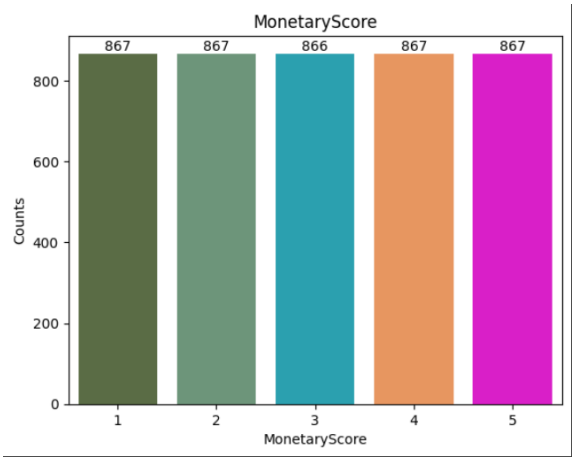
'Inactive_Customers' - This is labeled for customers who are almost not interested in buying Items
'Vulnerable_Customers' - This is labeled for customers who we are at a risk of losing.
'Priority_Customers' - This is labeled for customers who are valuable and cannot be lost at any cost
'About_to_become_Inactive_Customers' - This is labeled to customer whom we are about to become Inactive
'Essential_Customers' - This is labeled for customers who are needed to focused/need attention on them to improve business
'Loyal_Customers' - This is labeled to customers who are frequent and Periodic

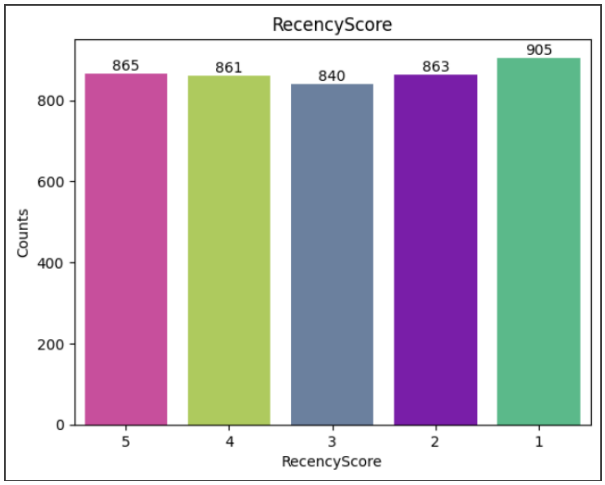
‘Promising_Customers’ - This is labeled to customers who are very promising
'New_Customers' - This is labeled to customers who are new to the market
'Potentially_Loyal_Customers' - This is labeled to customers who are very recent and spent a good amount.
'Important_Customers' - This is labeled to customers who are more regular, more periodic and spend more money.

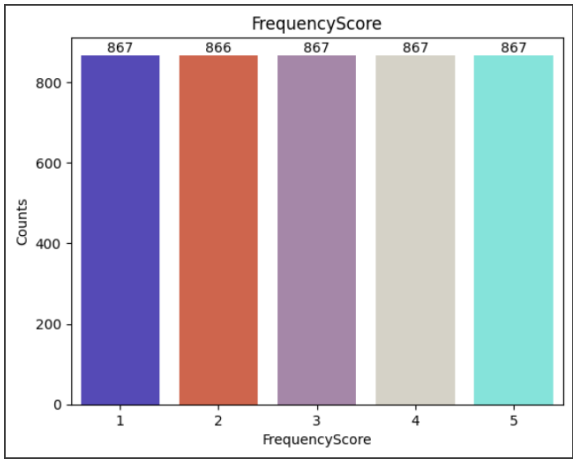
Visualizing the RFM Analysis:

:

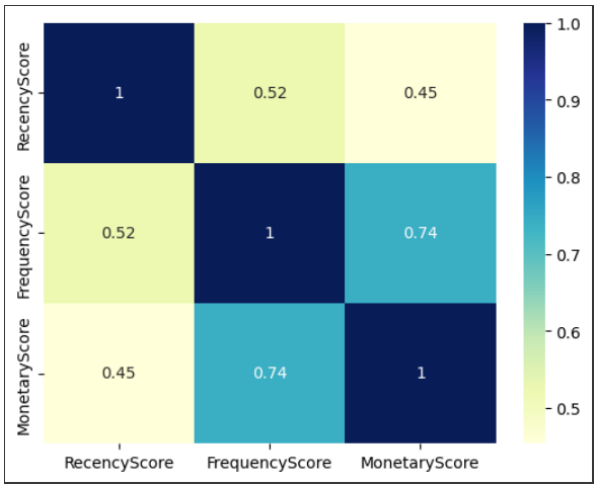


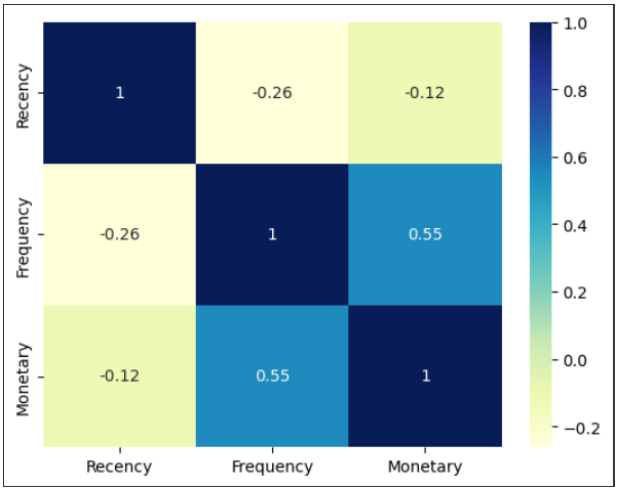


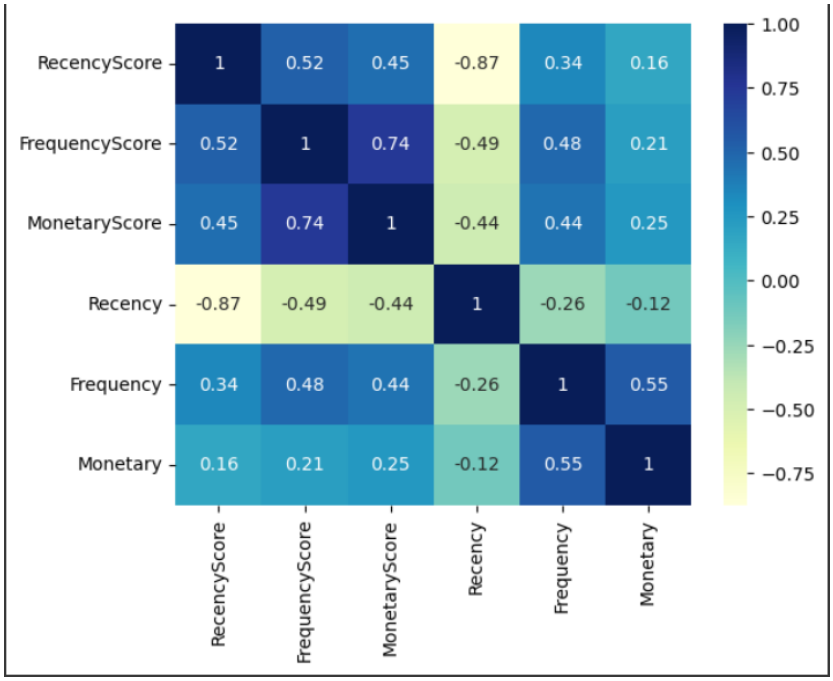




Correlation Matrix for RFM Analysis







Interpretations from RFM Analysis:

- The percentage of customer who are Rank_1 in RecencyScore are : 20.881%
- The percentage of customer who are Rank_2 in RecencyScore are : 19.958 %

- The percentage of customer who are Rank_3 in RecencyScore are : 19.912 %
- The percentage of customer who are Rank_4 in RecencyScore are : 19.866 %
- The percentage of customer who are Rank_5 in RecencyScore are : 19.381 %
- The percentage of customer who are Rank_1 in FrequencyScore are : 20.004 %
- The percentage of customer who are Rank_2 in FrequencyScore are : 20.004%
- The percentage of customer who are Rank_3 in FrequencyScore are : 20.004 %
- The percentage of customer who are Rank_4 in FrequencyScore are : 20.004 %
- The percentage of customer who are Rank_5 in FrequencyScore are : 19.981 %
- The percentage of customer who are Rank_1 in MonetaryScore are : 20.004 %
- The percentage of customer who are Rank_2 in MonetaryScore are : 20.004%
- The percentage of customer who are Rank_3 in MonetaryScore are : 20.004%
- The percentage of customer who are Rank_4 in MonetaryScore are : 20.004%
- The percentage of customer who are Rank_5 in MonetaryScore are : 19.981%
- The percentage of Inactive_Customers are : 24.780%
- The percentage of Important_Customers are : 18.412%
- The percentage of Vulnerable_Customers are : 15.113 %
- The percentage of Promising_Customers are : 13.544%
- The percentage of Loyal_Customers are : 11.1675%
- The percentage of Potentially_Loyal_Customers are : 8.029 %
- The percentage of About_to_become_Inactive_Customers are : 4.360%
- The percentage of Essential_Customers are : 2.076%
- The percentage of New_Customers are : 1.453%
- The percentage of Priority_Customers are : 1.061%

PCA

We have applied a dimensionality reduction technique for the Data using PCA. In PCA we equated the no of components as 2. Hence our data is reduced to 2 dimensional data and then we observe the below interpretations from the data.

From the pca's default attributes we import and find the explained_variance_ratio for our data. We then plot the scree plot for our data and find the reconstruction error. And then find the correlation matrix.

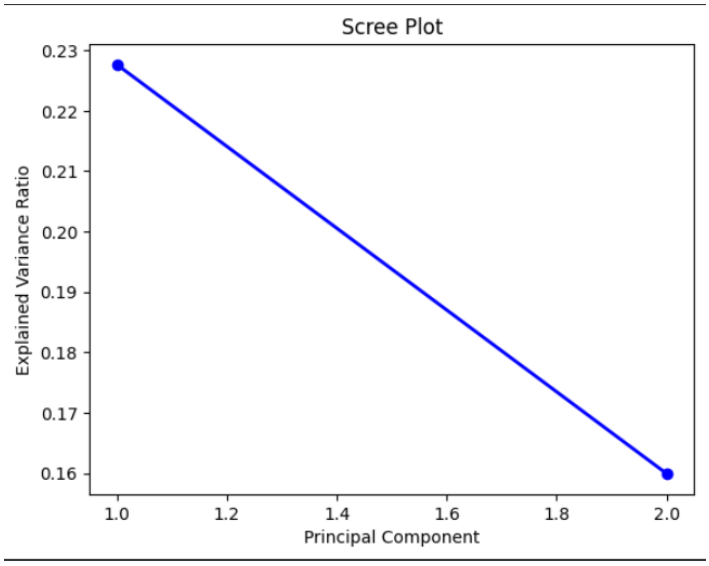
- Explained variance ratio: [0.22764705 0.15992418]
- Reconstruction error: 0.61242877907916

Correlation Matrix

```
Correlation matrix:
    InvoiceNo  StockCode  Description  Quantity  UnitPrice   Country     Month
0  -0.568608  -0.078874    -0.025301  0.006085   0.026074 -0.003890 -0.426074
1   0.004777   0.008234     0.001348  0.705237   0.037539 -0.023476  0.002403

       Year   WeekDay  Month_Year      Hour   Revenue
0 -0.393842  0.025746   -0.575392 -0.025675  0.006523
1  0.008195  0.002719    0.005229 -0.024007  0.707039
```

ScreePlot



KMeansClustering

We have applied KMeansClustering for our dataset. With the required number of clusters equal to 3.

KNearestNeighbors

We have applied KnearestNeighbors for our dataset. With the required numbers of neighbors as 5

- Accuracy score that is achieved by KNN is : 87.662%
- The Precision score is : 0.836
- The Recall score is : 0.876
- The F1 score is : 0.854

Correlation Matrix

```
The Confusion matrix is :
[[    46      0      0 ...      0    223      0]
 [     0      6      0 ...      0     62      0]
 [     0      0      0 ...      0      7      0]
 ...
 [     0      0      0 ...      0     18      0]
 [   134     42      2 ...      2 103076     11]
 [     0      0      0 ...      0     72      4]]
```

Random Forest Classifier

We apply Random Forest Classifier to our data with attributes n_estimators = 100, with cost function/criterion = gini index. We apply the normal data to the RFC and PCA reduced data to RFC and find the accuracies of the models.

- Accuracy that is achieved from RFC before applying PCA is : 99.98%
- Accuracy that is achieved from RFC after applying PCA is : 89.92%

References:

- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- https://www.shopify.com/blog/what-is-customer-segmentation#:~:text=What%20does%20customer%20segmentation%20mean,%2C%20services%2C%20and%20marketing%20stra
- https://clevertap.com/blog/rfm-analysis/