MATHEMATICS & STATISTICS
IIT KANPUR

MTH209 PROJECT

Customer Segmentation
on
Credit Card Data

SUBMITTED TO
Dr. Subhajit Dutta

SUBMITTED BY
DATA WIZARDS

Introduction

- Customer segmentation identifies groups based on factors like purchase frequency, transaction amounts, and dates (e.g. repeat/loyal customers, high spenders, one-time buyers, etc.
- Clustering by behavior is crucial for personalized products.
- We use k-means with K determined by silhouette score and PCA for dimension reduction to group credit card holders for targeted marketing.

Motivation for doing this project

Source of Data Collection

We have used the credit card dataset available on Kaggle.

Data Preview

1) CUST_ID	7) CASH_ADVANCE	13) PURCHASES_TRX
2)BALANCE	8) PURCHASES_FREQUENCY	14) CREDIT_LIMIT
3) BALANCE_FREQUENCY	9) ONE_OFF_PURCHASES_FREQUENCY	15) PAYMENTS
4) PURCHASES	10) PURCHASES_INSTALLMENTS_FREQUENCY	16) MINIMUM_PAYMENTS
5) ONEOFF_PURCHASES	11) CASH_ADVANCE_FREQUENCY	17) PRC_FULL_PAYMENT
6) INSTALLMENTS_PURCHASES	12) CASH_ADVANCE_TRX	18) TENURE

Outline of Project

STEP 1

- Visualizations of the data collected inorder to see any correlation or important features of the dataset.

STEP 2

- Applied Yeo-Johnson transformation and then again data visualizations

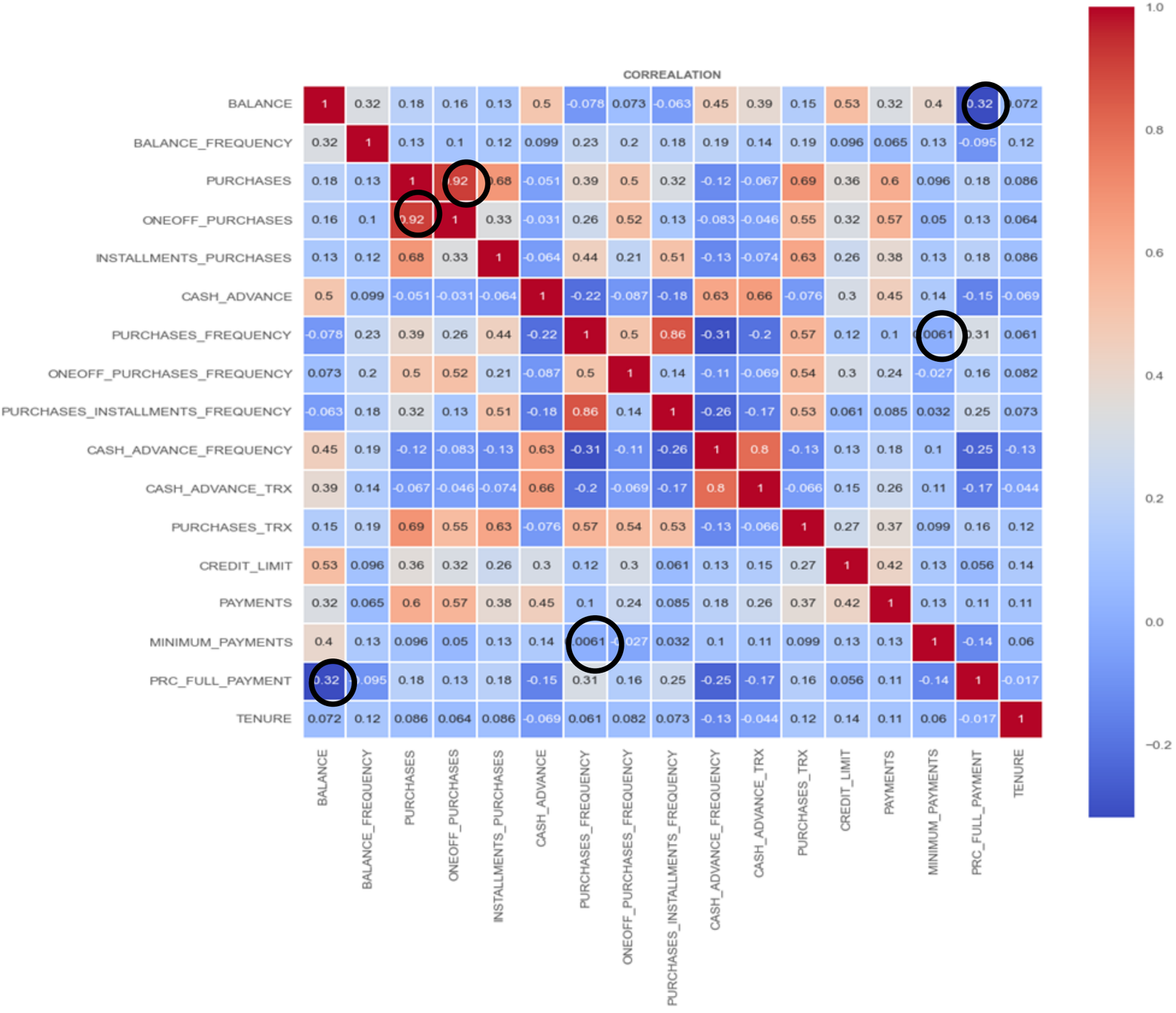
STEP 3

- Applied PCA
- Visualization of percentage of explained variance by components
- Feature loadings graph

STEP 4

- K-Means Clustering using Silhouette Method.
- Silhouette Plot for k = 3
- Comparison of Principal Components among different clusters.

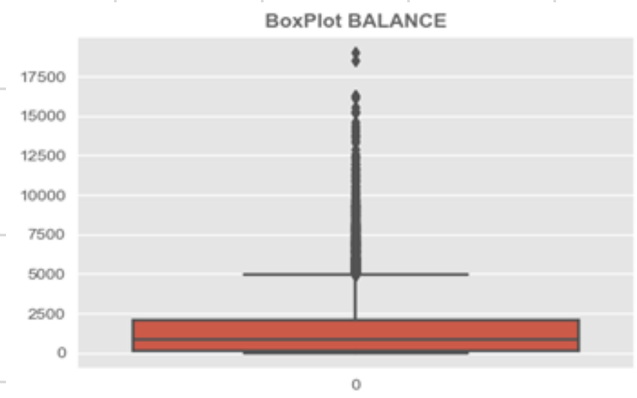
Visualisations of Data Collected

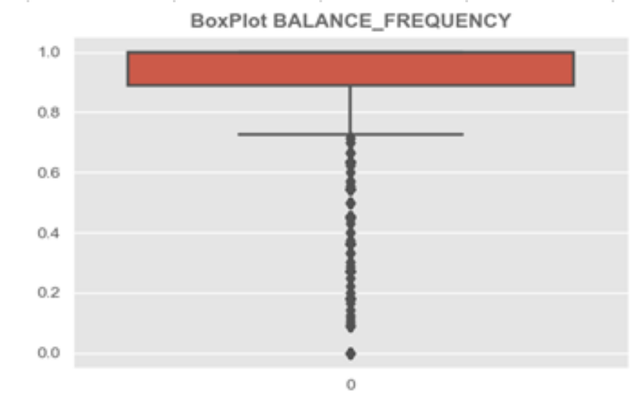


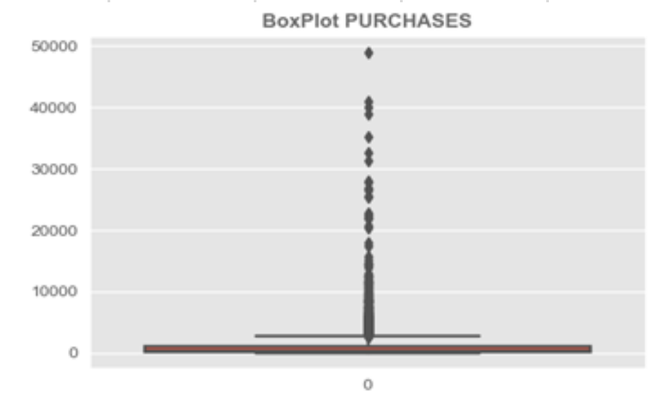
Visualization Of Correlation matrix of Data

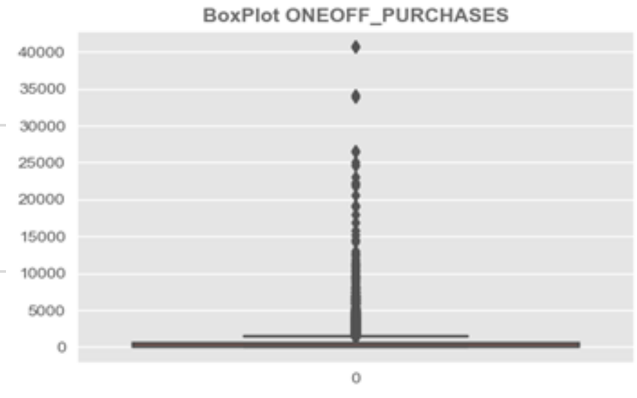
- The features ONEOFF_PURCHASES and PURCHASES have the highest positive correlation among all features i.e. 0.92
- The features MINIMUM_PAYMENTS & PURCHASES_FREQUENCY are highly uncorrelated.
- The features PRC_FULL_PAYMENT and BALANCE have the highest negative correlation among all features i.e. – 0.32

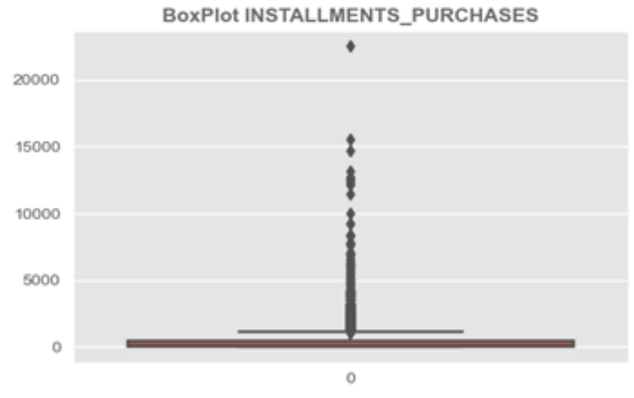
Box Plots of all the features

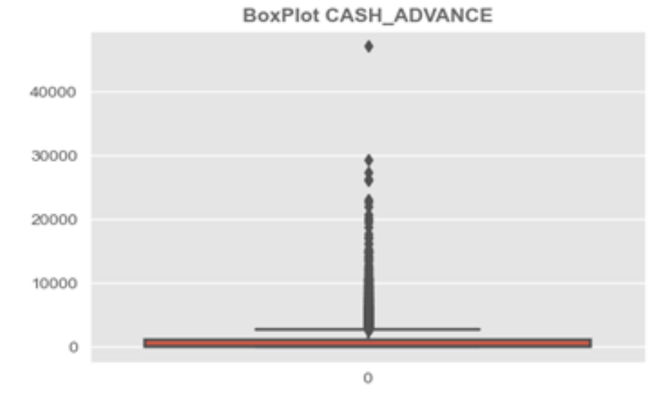


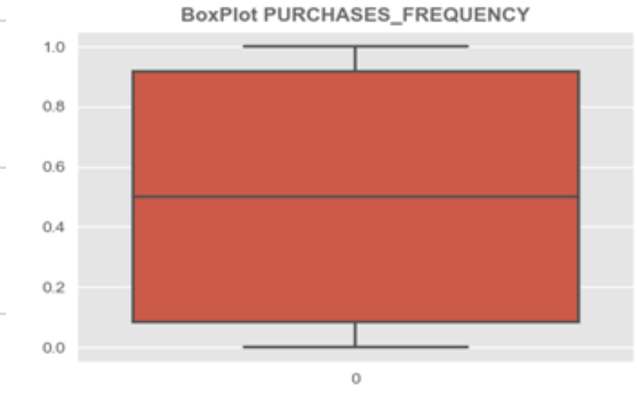


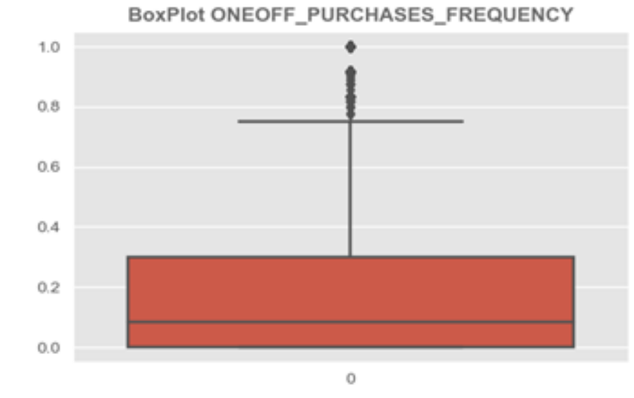


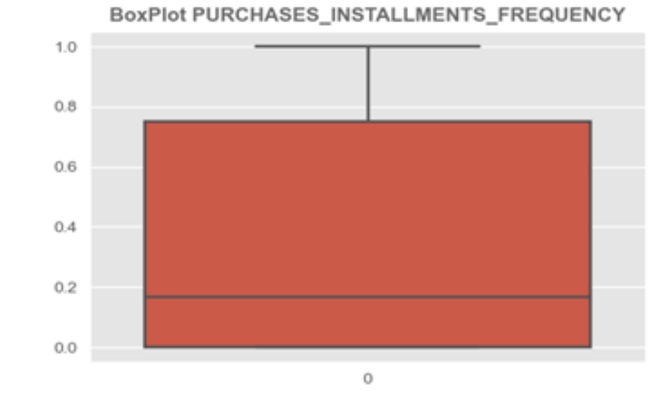




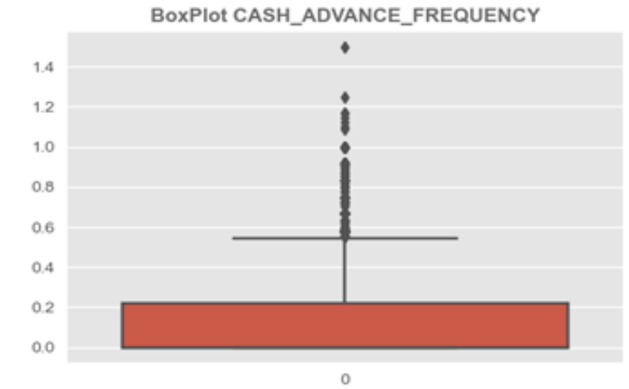


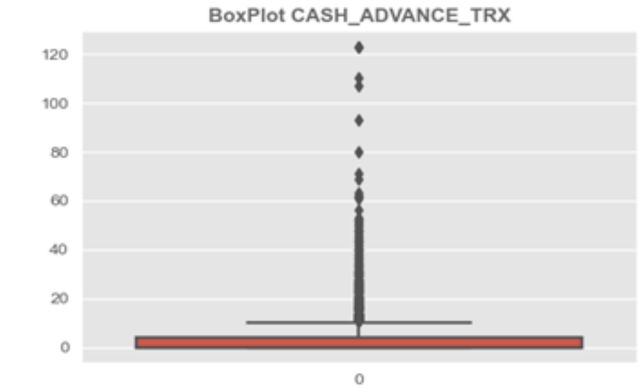


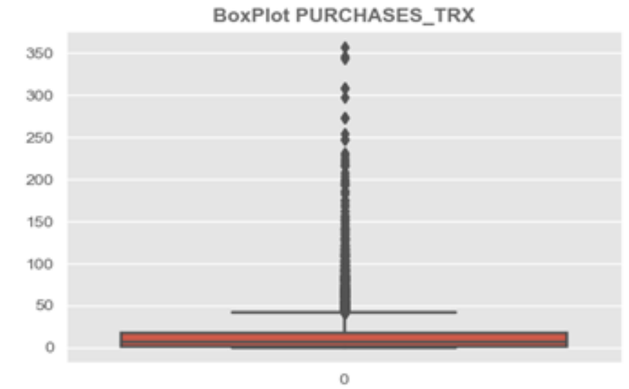


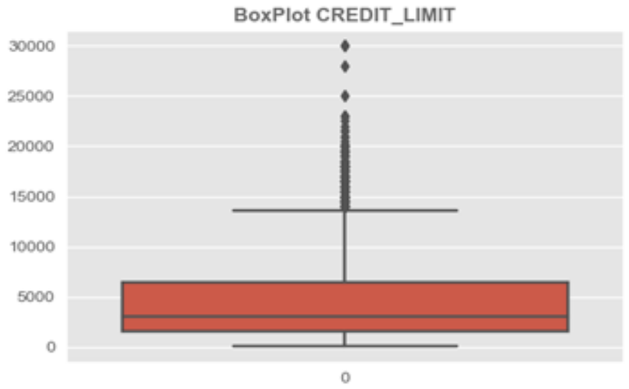


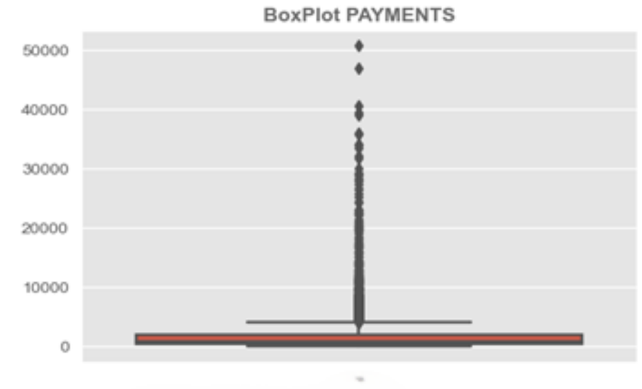
Box Plots of all the features

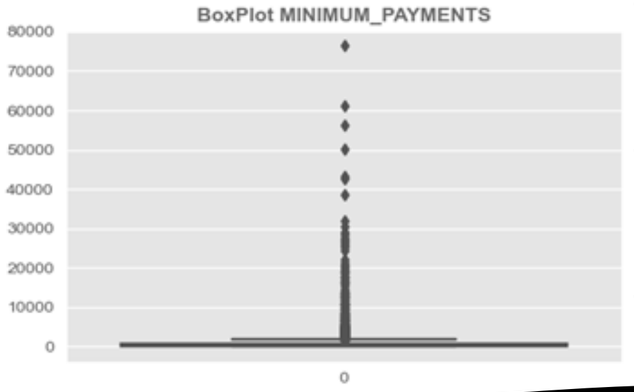


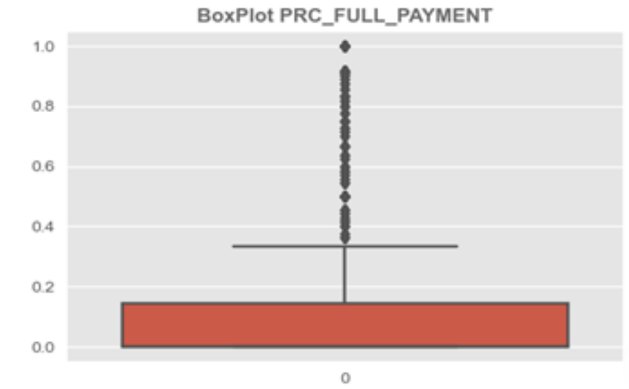




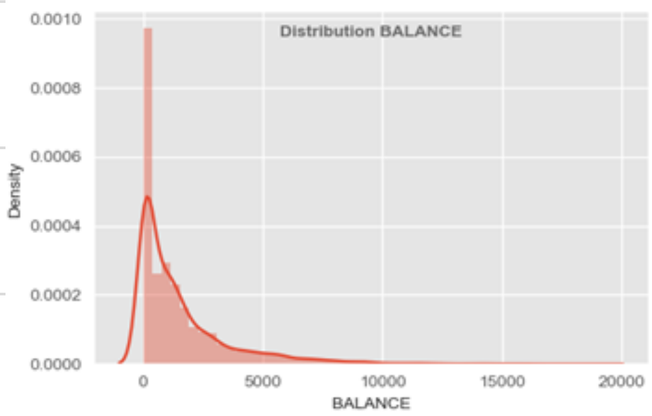


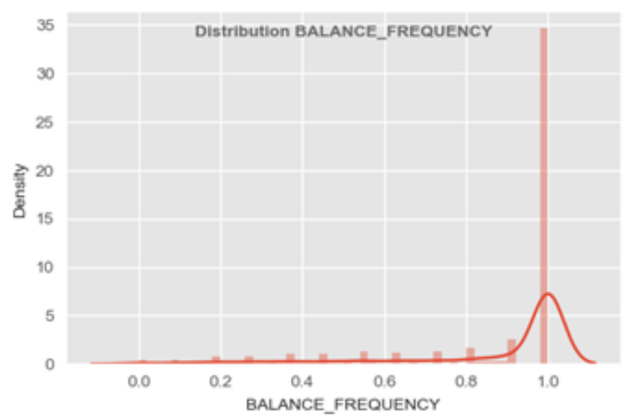


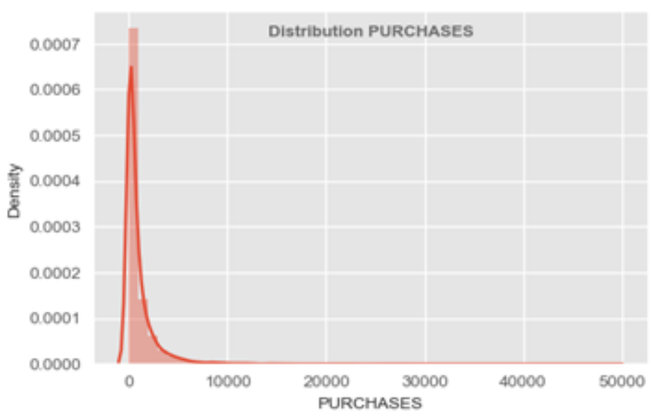


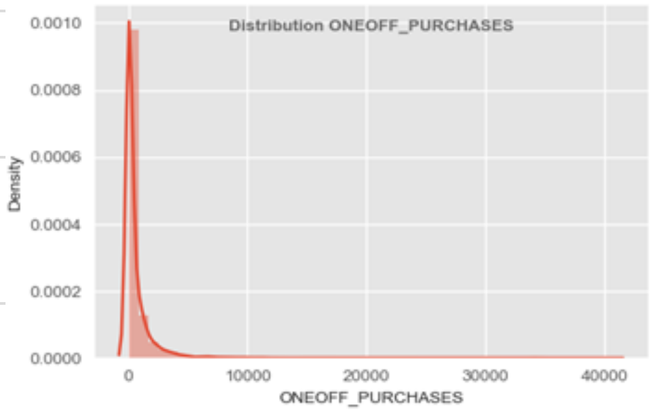


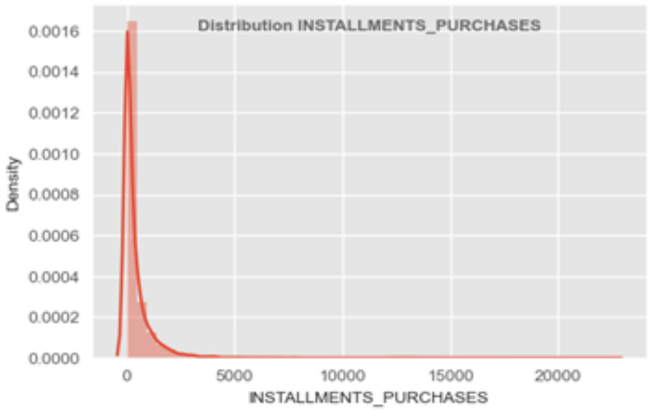
Distribution plots of all the features

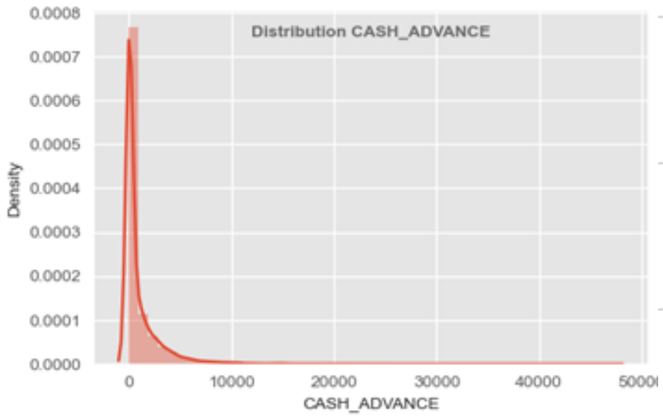


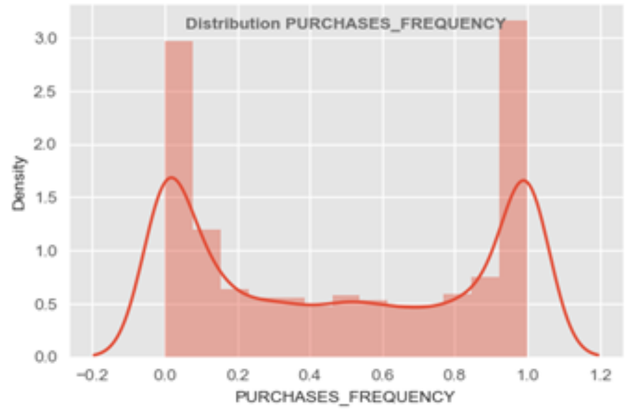


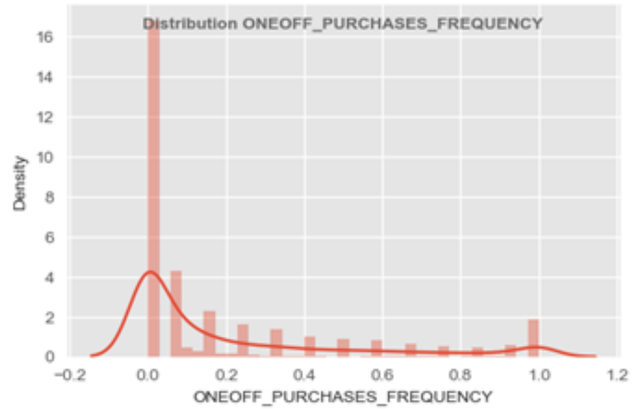


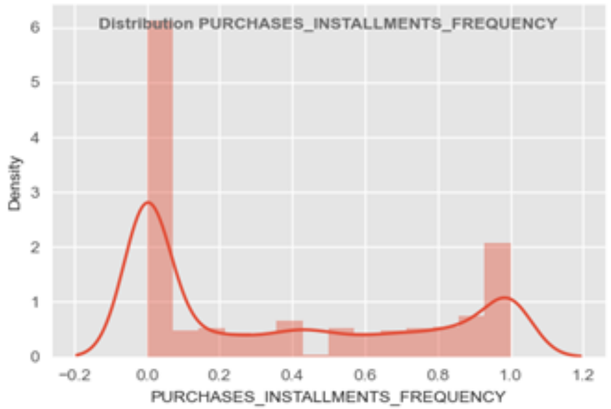




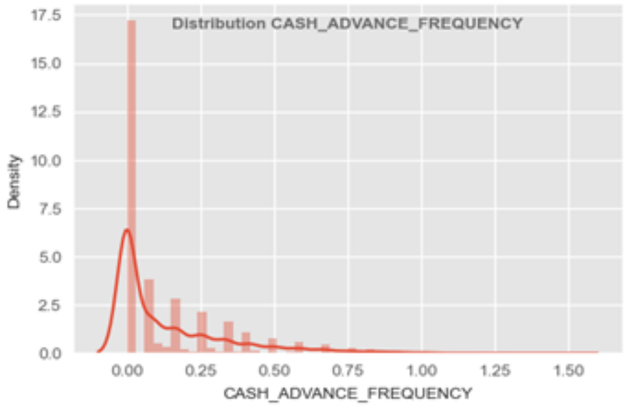


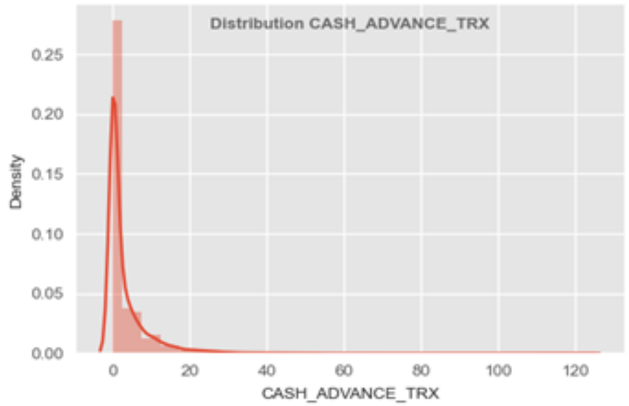


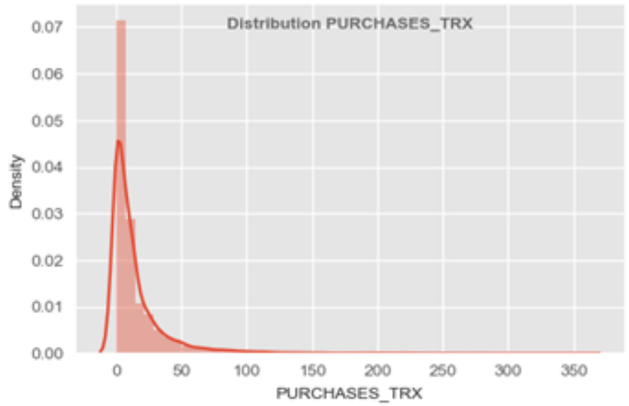


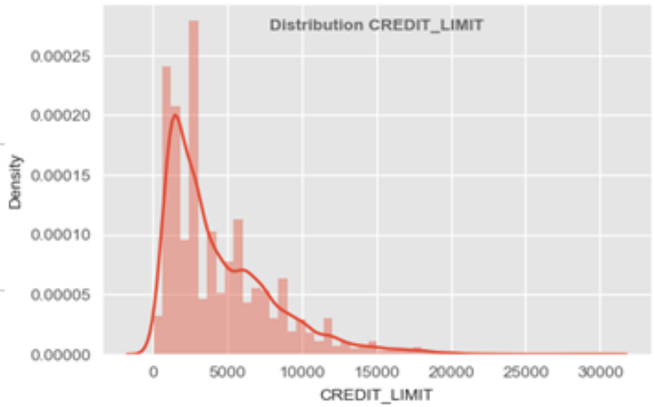


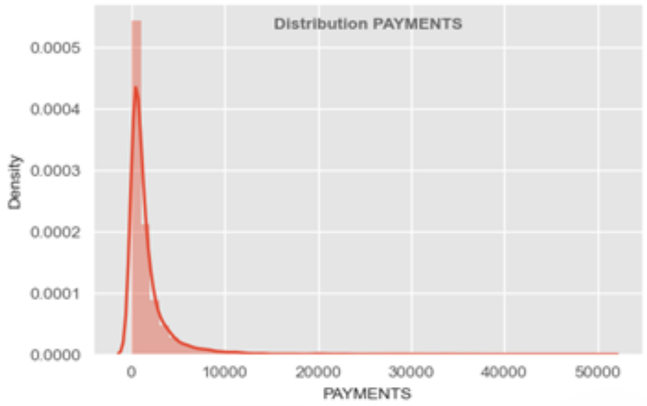
Distribution plots of all the features

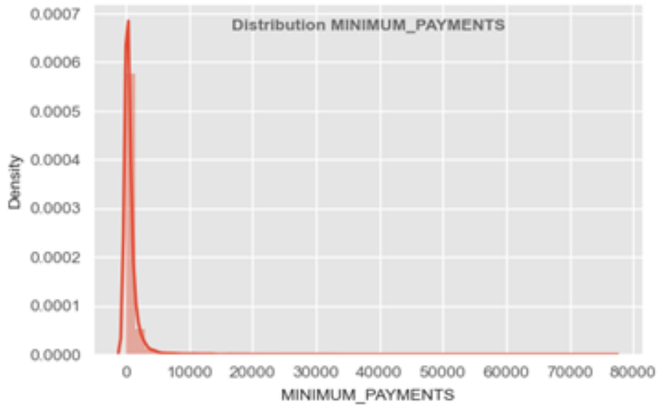


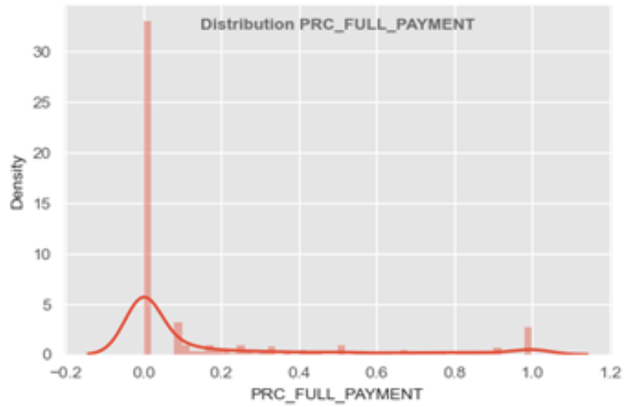




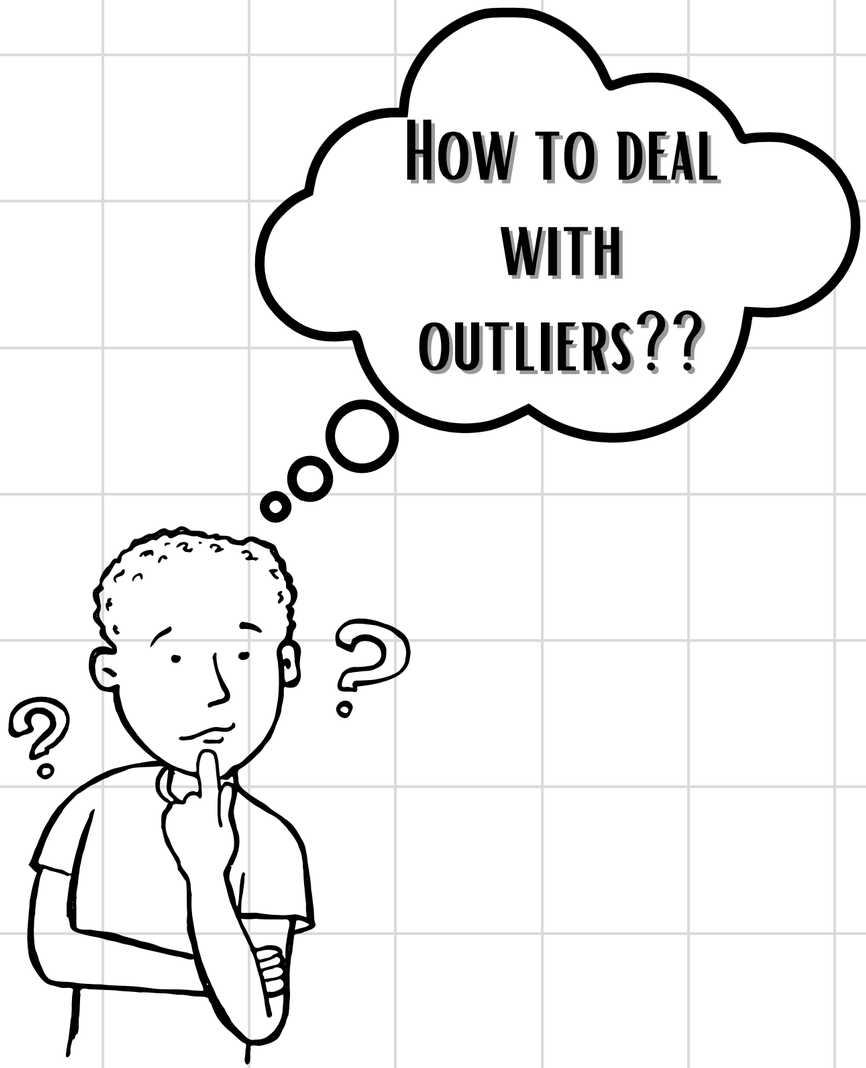








!!Most of the features are highly skewed



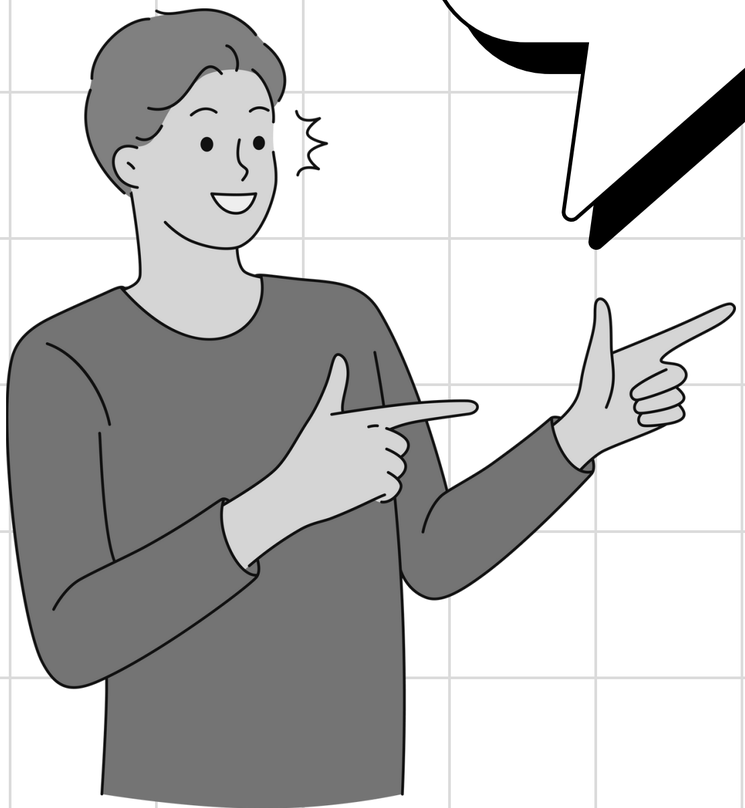
Let's apply the Yeo-Johnson transformations

What is Yeo-Johnson transformations ?

The Yeo-Johnson transformation is a method used to transform non normal data distributions into approximately normal distributions or to stabilize the variance across different groups or samples.
It is a variation of the Box-Cox transformation that handles both positive and negative values.

Outlier Handling: Yeo-Johnson transformation can also help in mitigating the impact of outliers. By transforming the data, extreme values can be brought closer to the bulk of the data, reducing their influence on statistical estimates and model performance.

LET'S LOOK AT THE DATA VISUALIZATIONS AFTER YEO-JOHNSON TRANSFORMATIONS

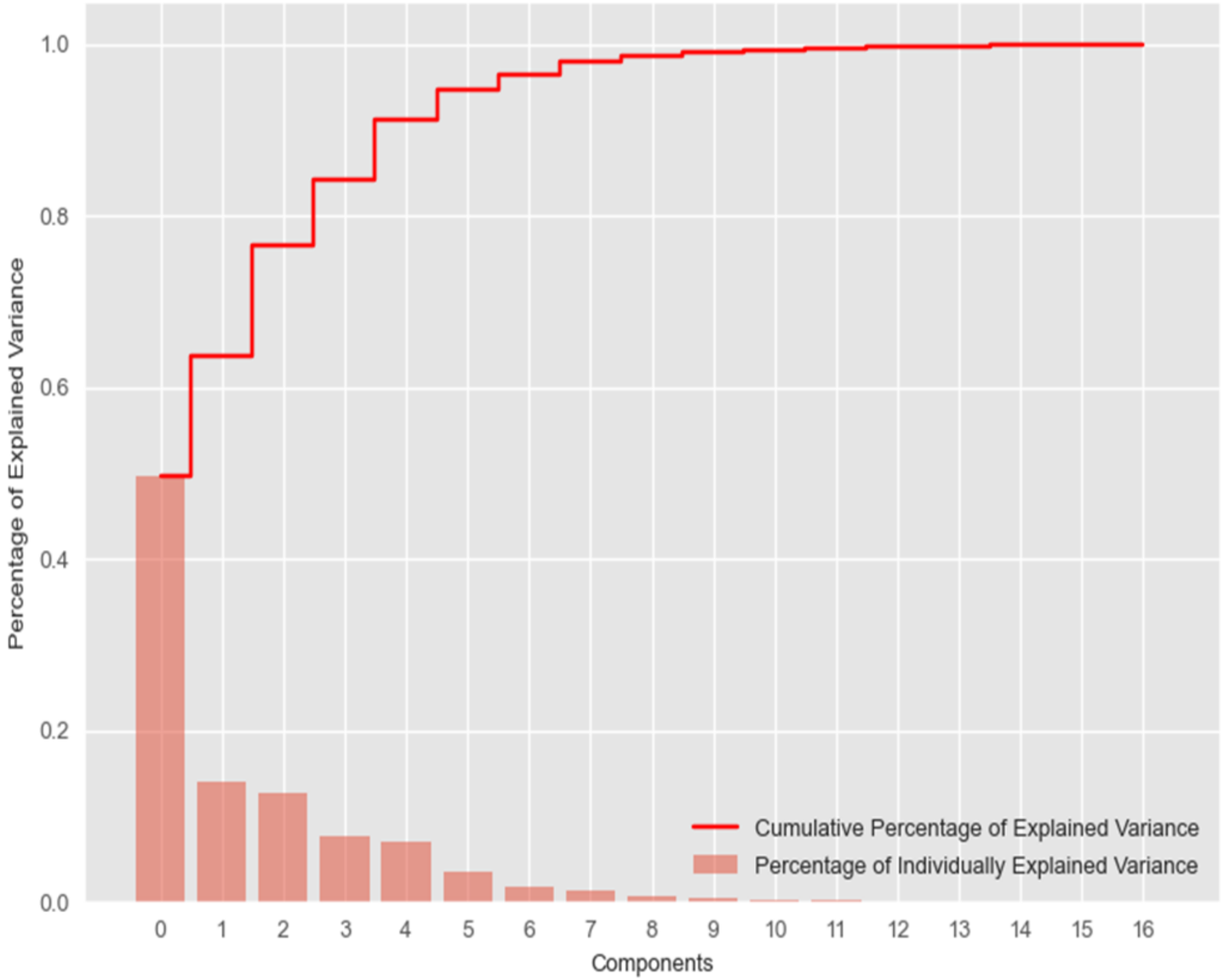


Box Plots of all the features

Distribution plots of all the features

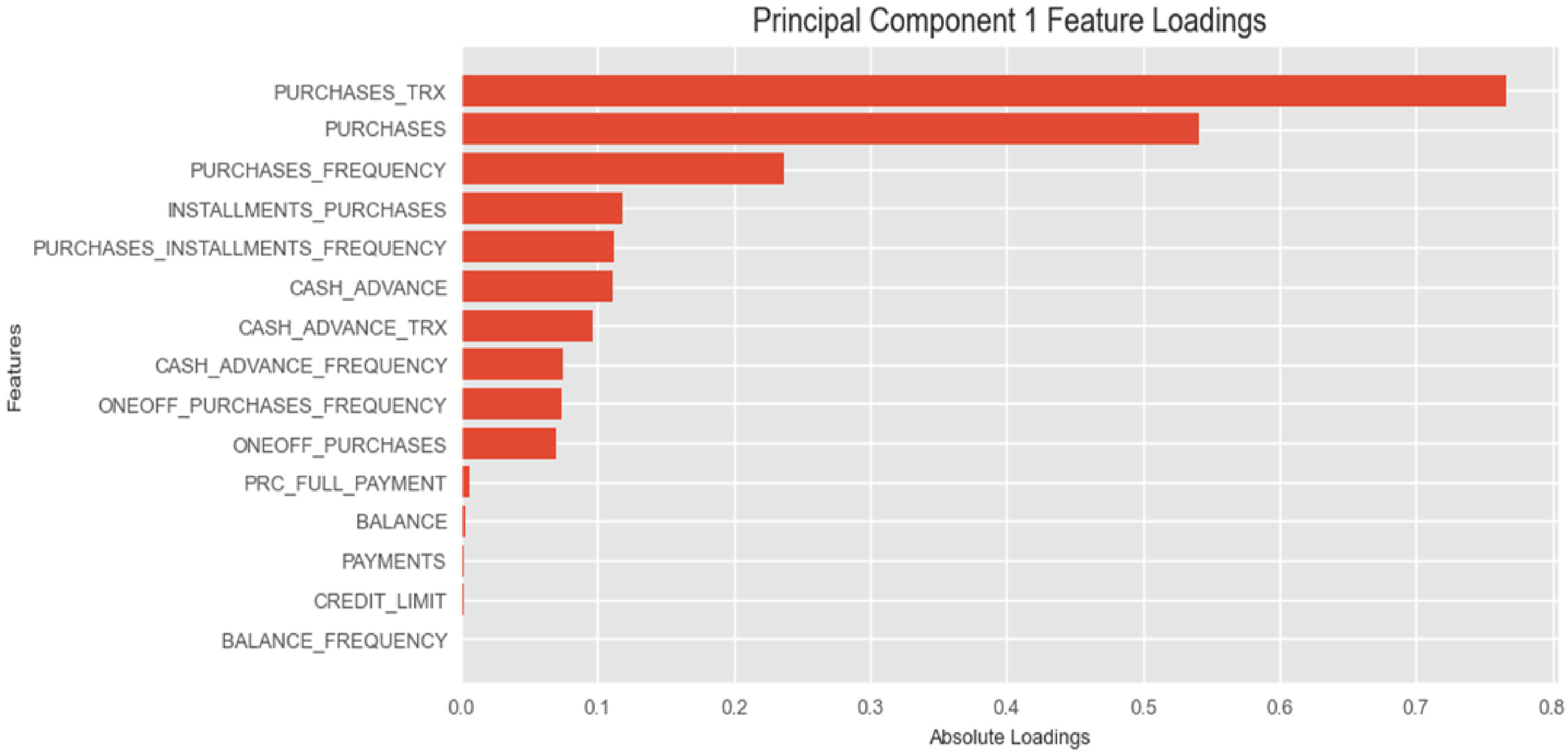
NOW THE DATA IS NOT THAT MUCH SKEWED

Visualization of Percentage of Explained Variance by Components

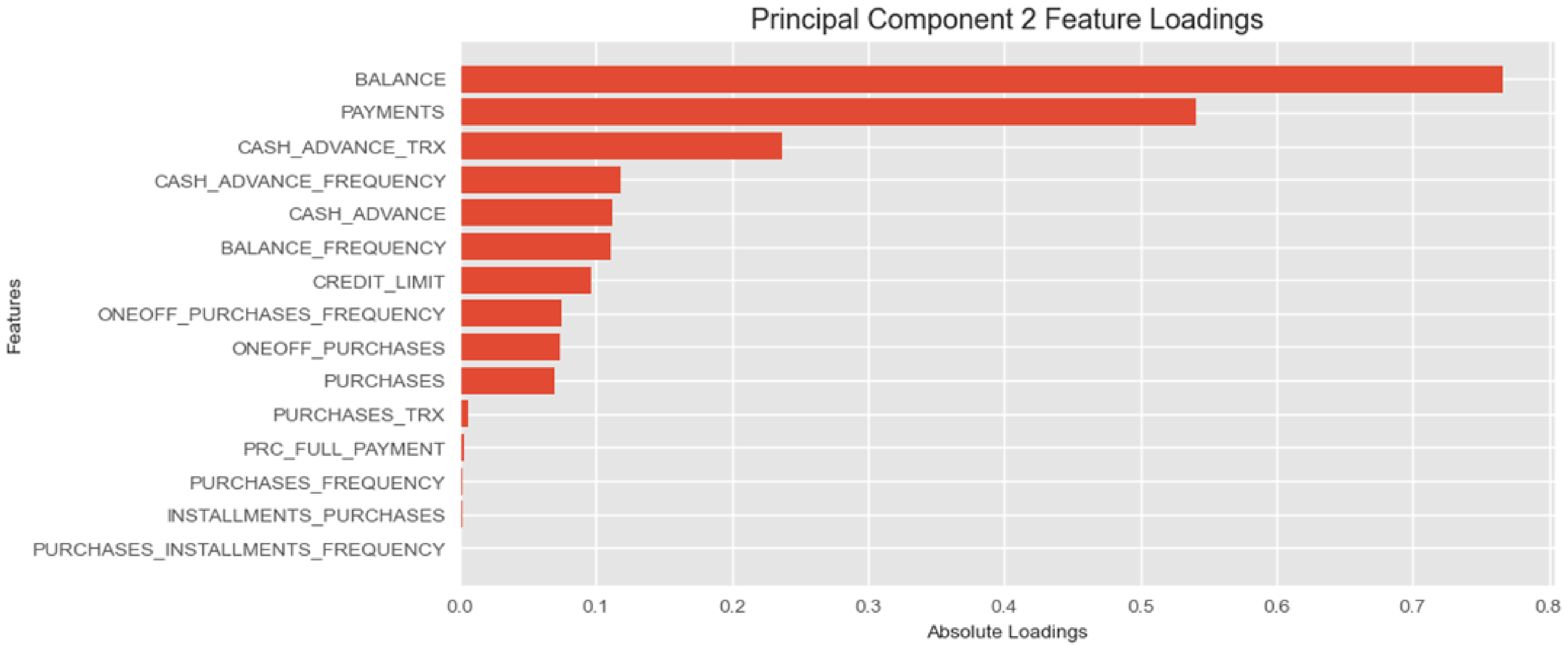


From the above graph it is clear that first three components explain about 62% variability of the data. So, we have considered two principal components.

Feature loadings graph



Feature loadings graph



The length and direction of each bar indicate the strength and direction of the relationship between the original features and that principal component. Features with higher absolute loading values (either positive or negative) contribute more to that component. Therefore, some initial features are contributing more to the component.

K-Means Clustering

To choose optimal number of clusters we used Silhouette Method

The Silhouette Method evaluates how well-defined and separate the clusters are in a dataset. It calculates a silhouette score for each data point based on two factors:

1. The average distance from the point to all other points within the same cluster (a).
2. The average distance from the point to all points in the nearest neighboring cluster (b).

$$S = \frac{(b - a)}{\max(a, b)}$$

The silhouette score ranges from -1 to 1

+1

Indicates that the data point is well-clustered and far from neighbouring clusters

0

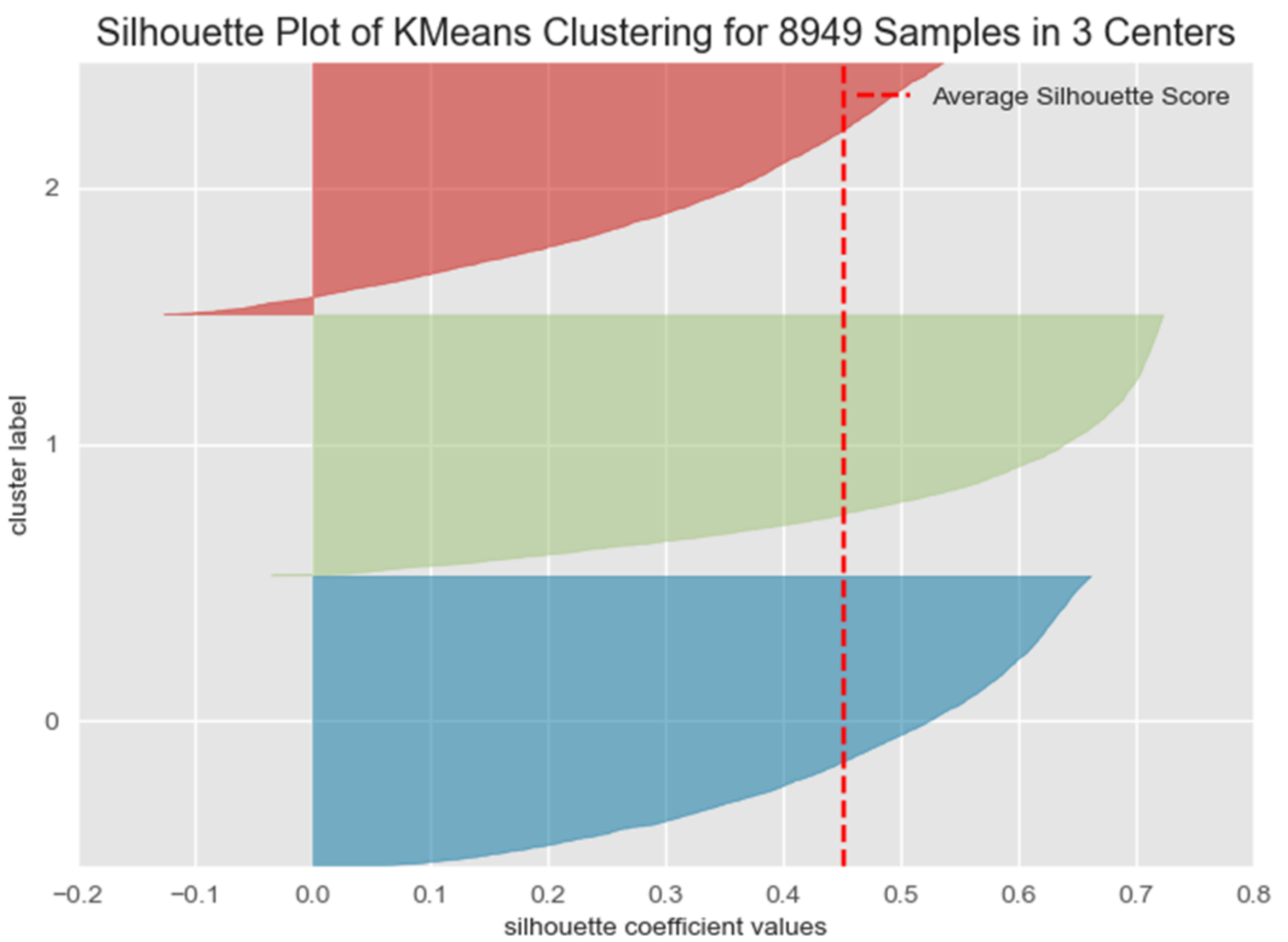
Indicates that the data point is close to the decision boundary between clusters

-1

Indicates that the data point may have been assigned to the wrong cluster

K-Means Clustering

We got K=3

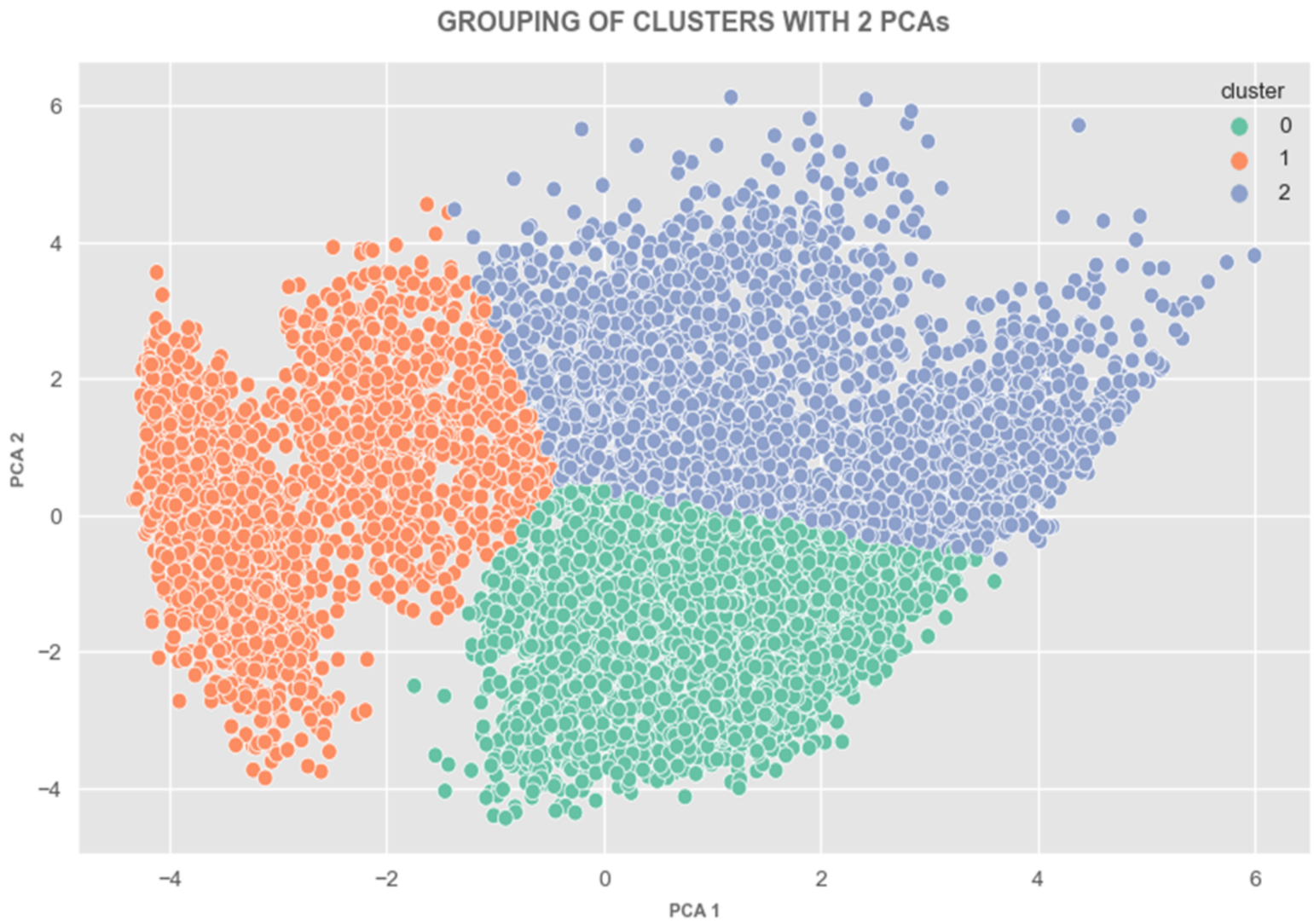


This Silhouette Plot suggests that one cluster (green) is well-separated, while the others (red and blue) may need further investigation.

Average Silhouette score is 0.45

Grouping of clusters

We got K=3



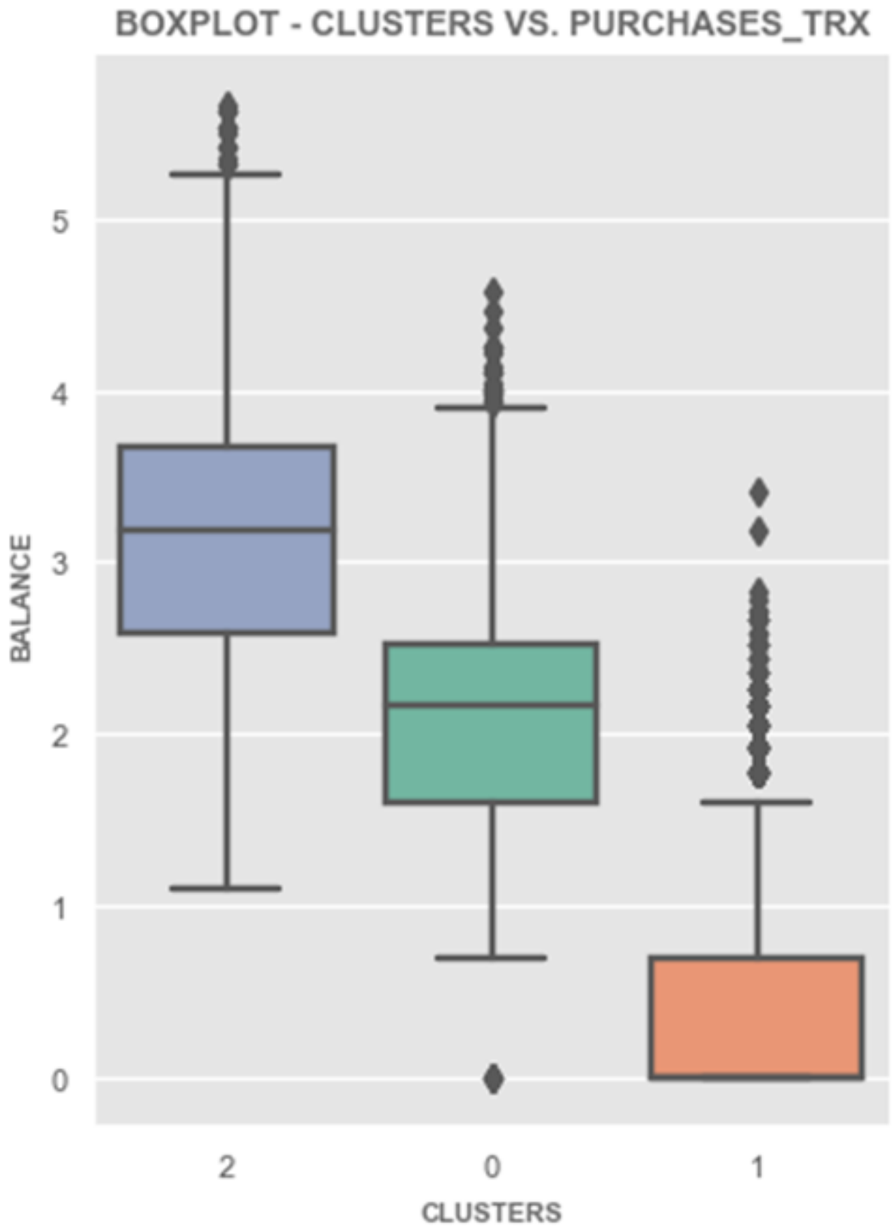
This scatter plot reveals distinct groupings within the dataset when reduced to two principal components. There are three distinct clusters represented by different colours:

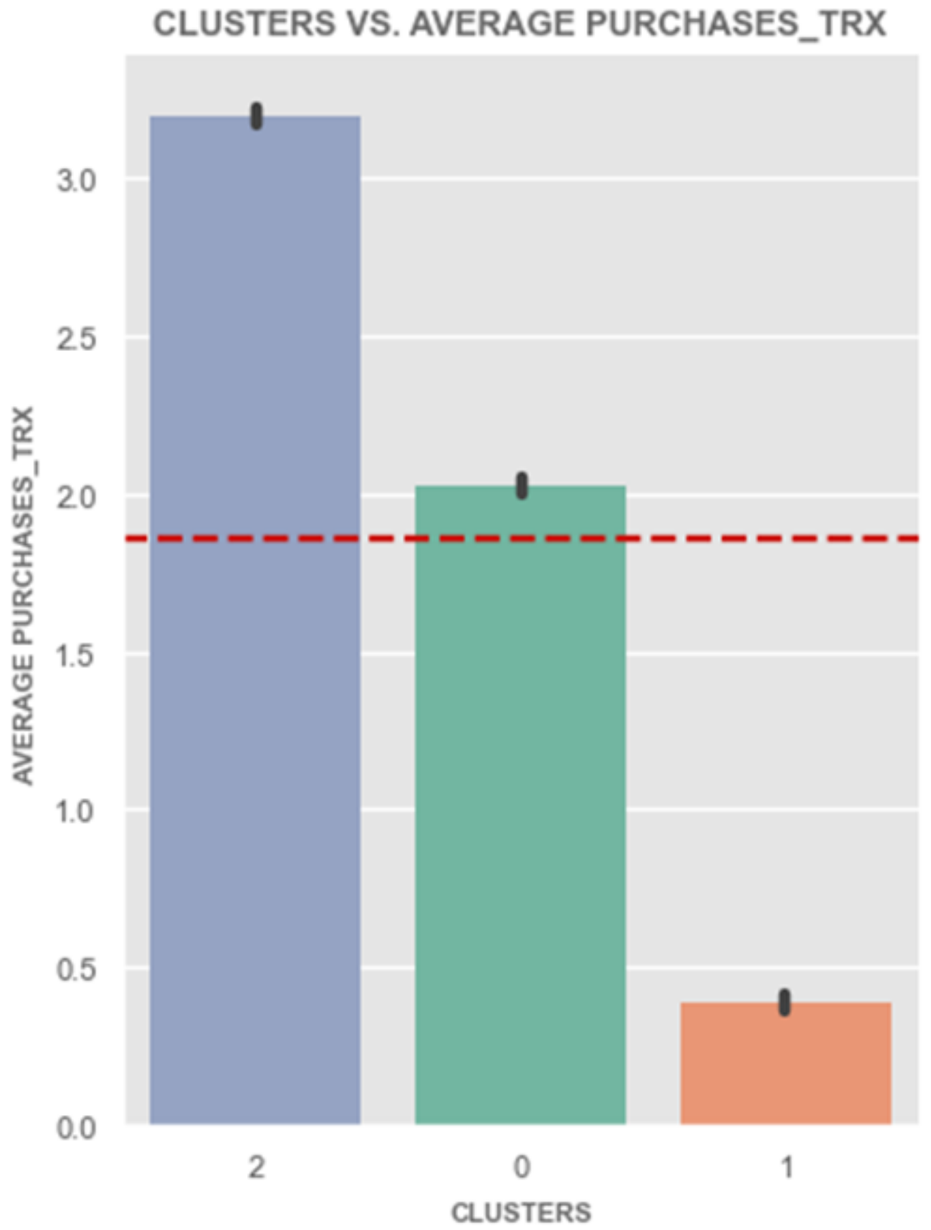
Orange: Concentrated towards the left side, indicating lower values of both PCA 1 and PCA 2.

Green: Located in the middle with moderate values of both PCAs.

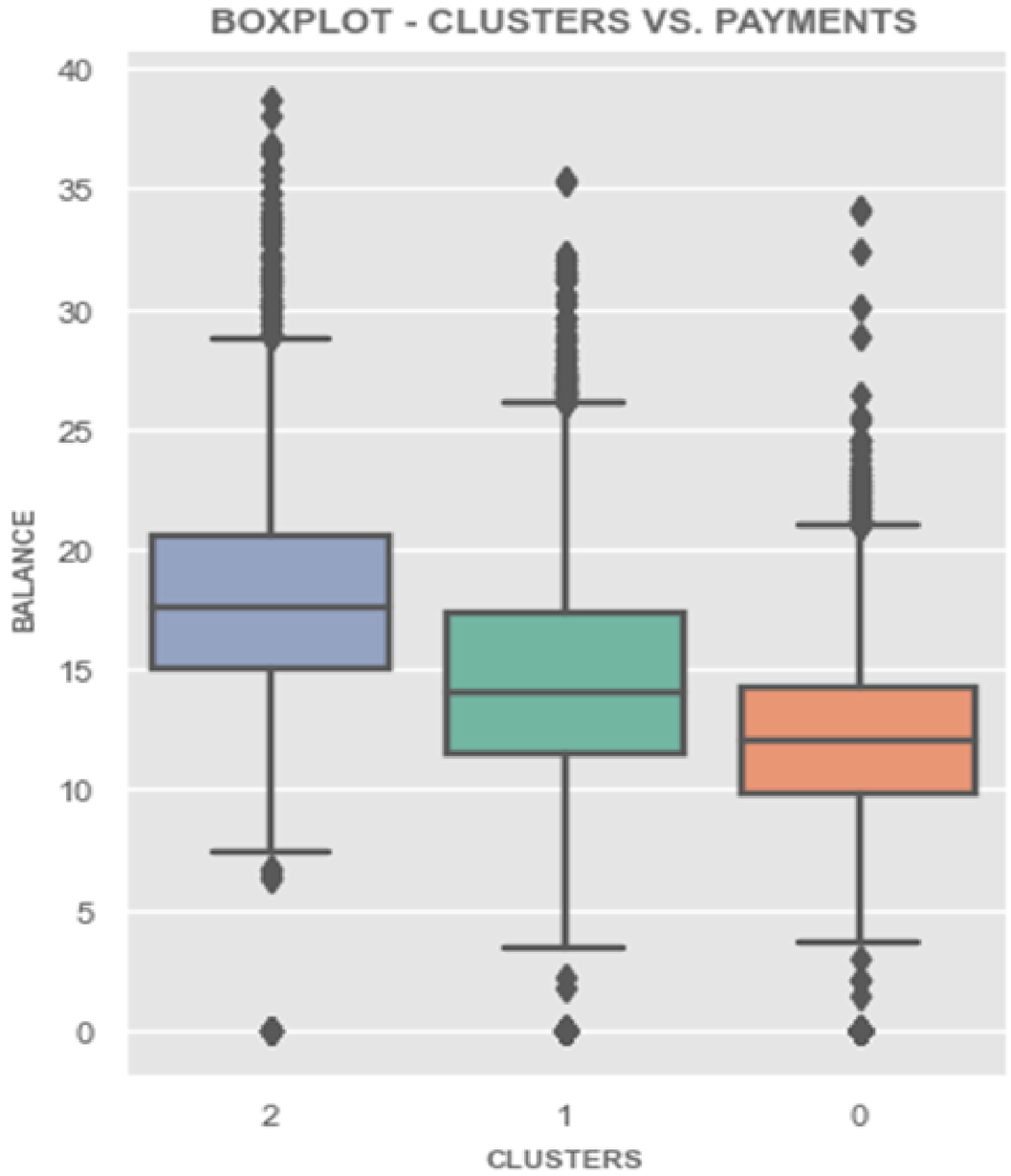
Blue: Spread towards higher values of PCA 1 but similar PCA 2 range as the green cluster.

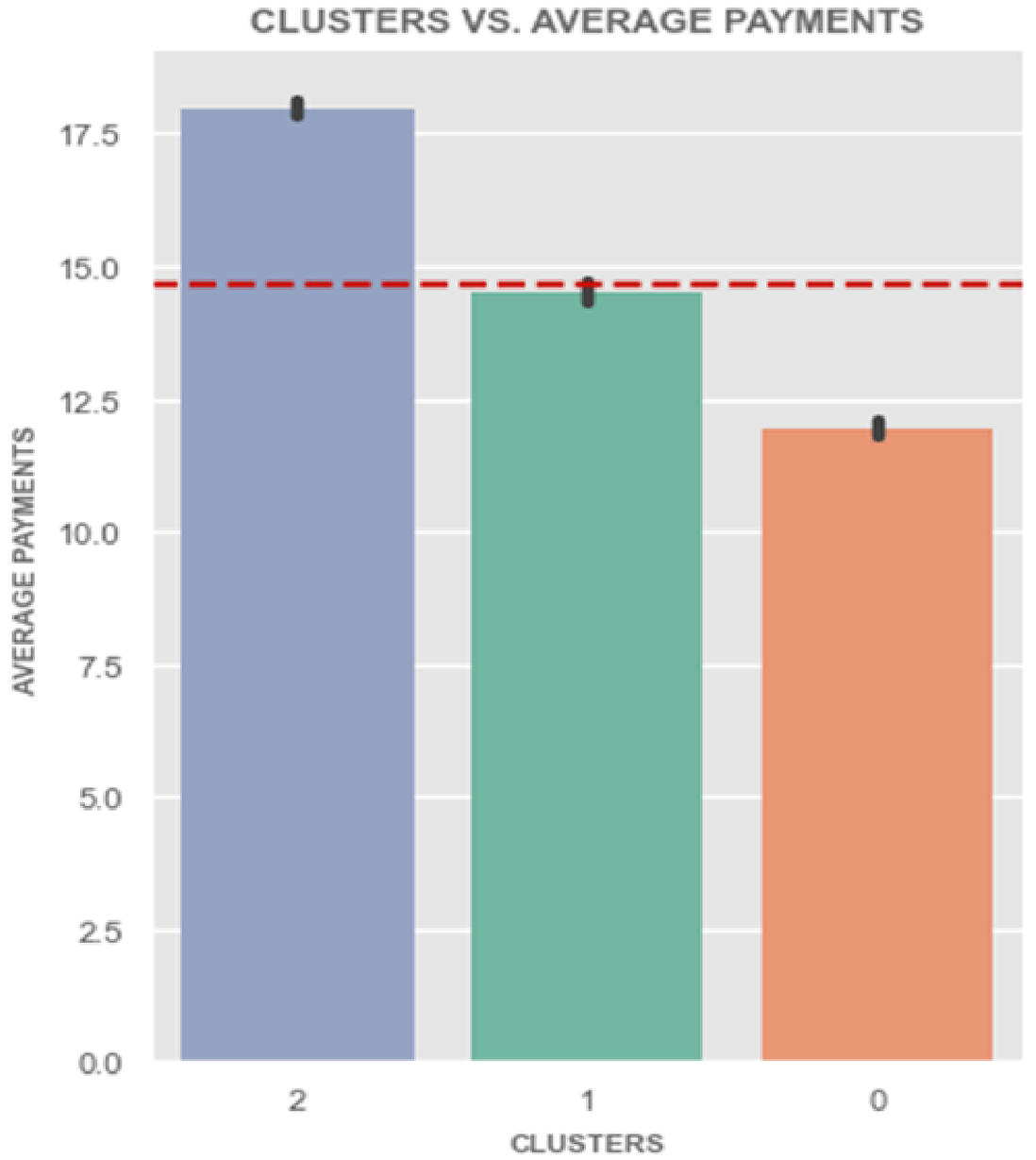
Comparison of Principal Components among different clusters



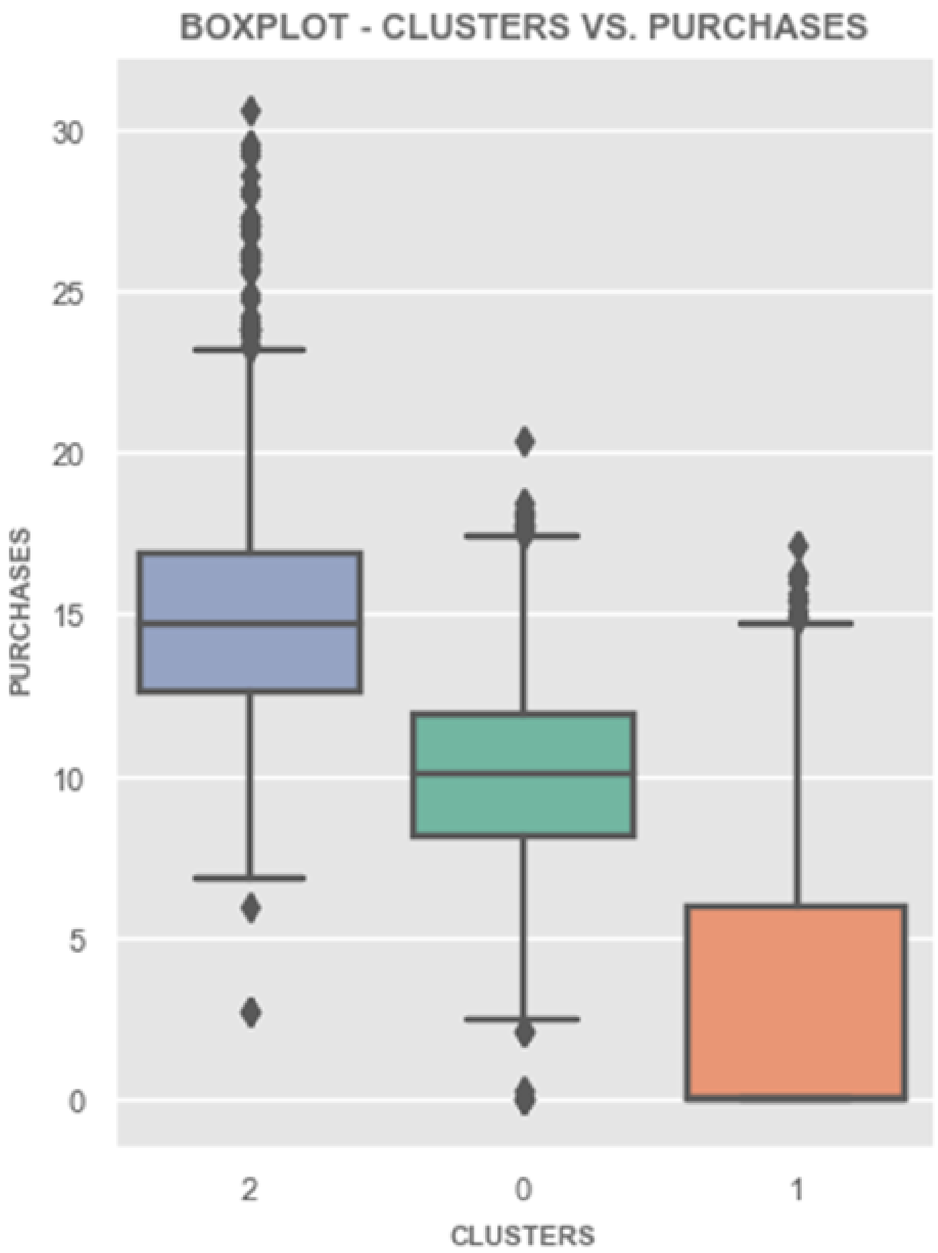


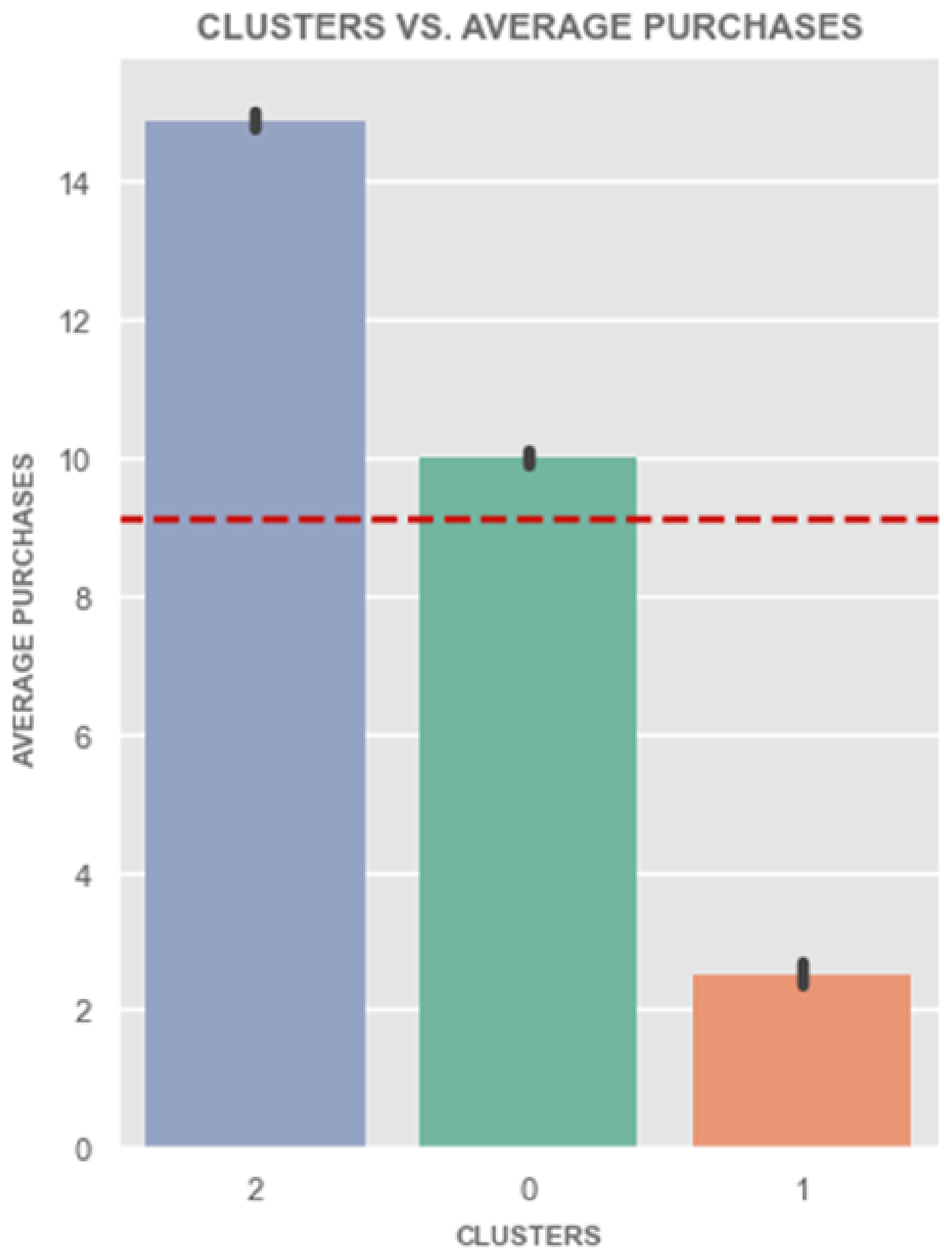
Comparison of Principal Components among different clusters



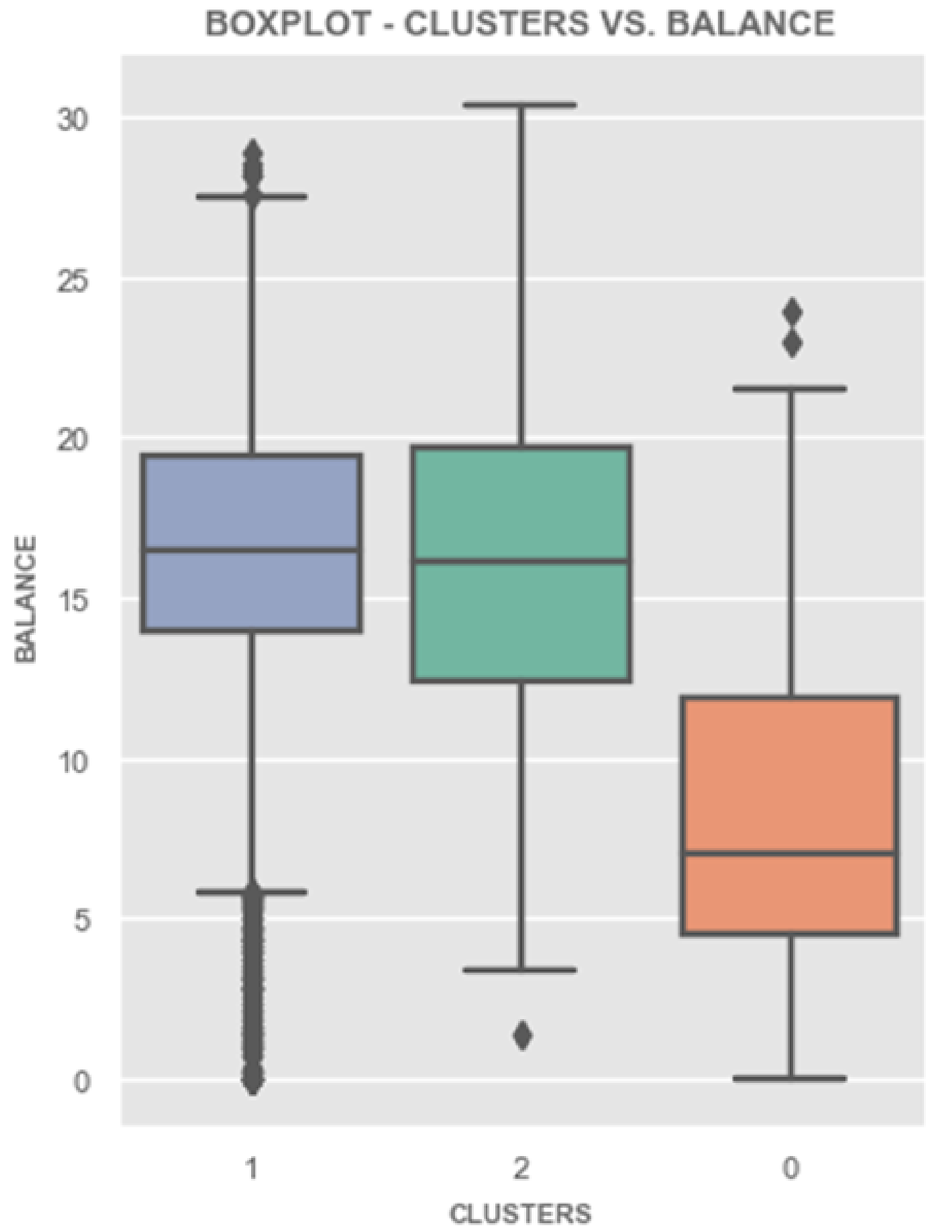


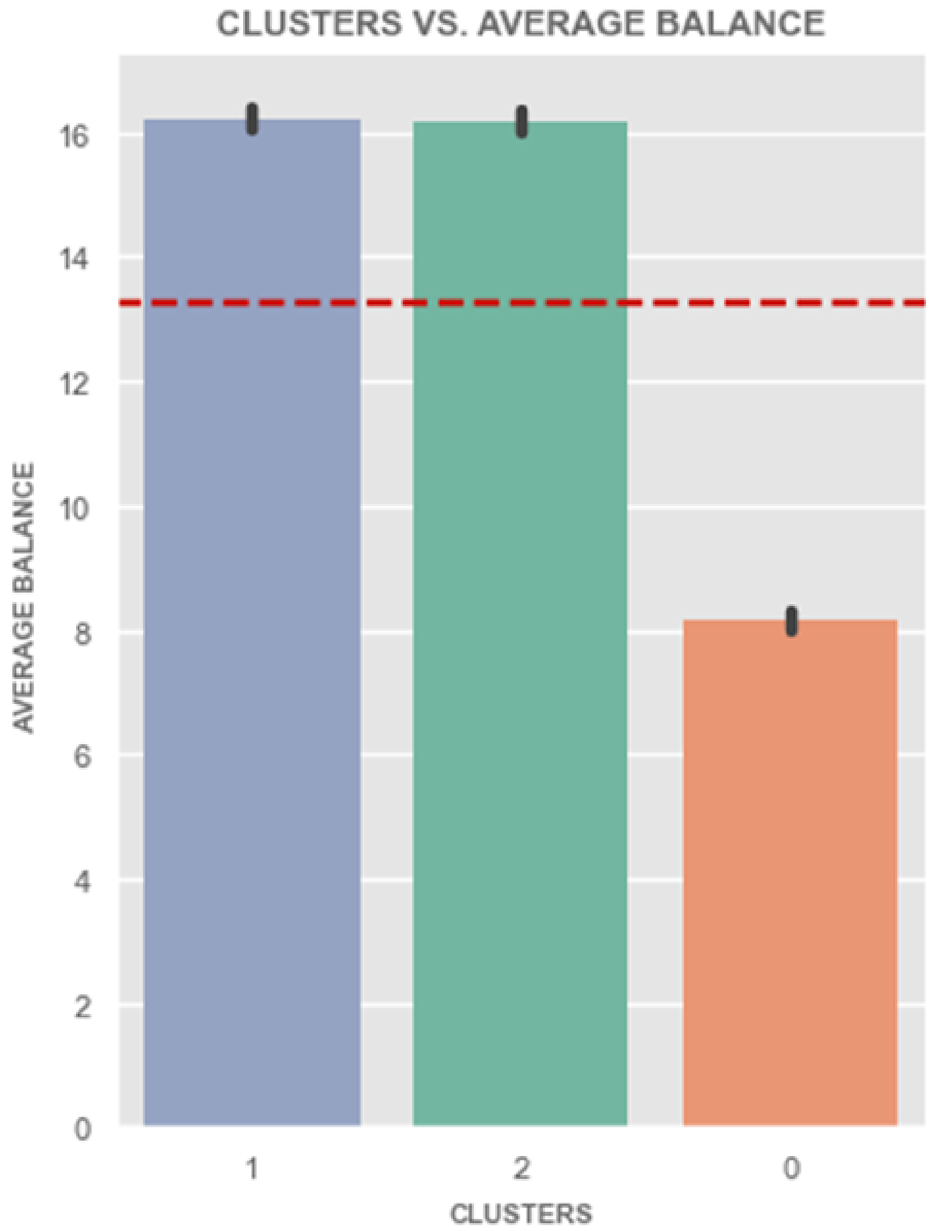
Comparison of Principal Components among different clusters





Comparison of Principal Components among different clusters





Comparison of Principal Components among different clusters

The **Box-plot** shows the distribution of PCs within each cluster. Clusters are ordered by the **median** value of PCs. The boxplot summarizes the central tendency and variability, indicating the consistency of purchasing behaviour within each cluster and highlighting potential outliers.

The **Bar plot** shows the **average** value of PCs for each cluster. Clusters are ordered by the mean value of PCs.

The **red dashed line** represents the **overall mean of PCs**, serving as a benchmark for comparison.

Conclusion

0 Cluster

- Companies should focus on increasing the number of purchase transactions by implementing rewards program that aligns with customer spending habits, offering cashback, travel rewards, or points redeemable for goods and services.
- Customers should be provided flexibility in the billing cycle in which they can now select the bill payment date that aligns with their cash flow, ensuring timely payments

1 Cluster

- Companies should focus on increasing the number of purchase transactions by implementing rewards program that aligns with customer spending habits, offering cashback, travel rewards, or points redeemable for goods and services.
- Customers should be provided flexibility in the billing cycle in which they can now select the bill payment date that aligns with their cash flow, ensuring timely payments
- Customers should be provided valuable scheme to increase the amount of purchases.

2 Cluster

- The users in cluster2 are satisfied with the service providers. So, there are no extra efforts needed from service providers for cluster2 customers.


THANK
YOU