# CUSTOMER SEGMENTATION ON CREDIT CARD DATA

Department of Mathematics and Statistics

MTH-209 Project Report

**Group Name- Data Wizards**

Submitted to – Dr. Subhajit Dutta


Submitted by-

- Vinay Kumar Verma
- Laxmi Agarwal
- Riya Mittal
- Anoop Patel
- Mukesh Kumar Suthar

# INTRODUCTION

Customer segmentation is the process of identifying customer segments based on historical purchasing patterns. For example, it can involve identifying repeat/loyal customers, high spending customers, customers that make one time or infrequent purchases and much more. Segments can be created using information like frequency of purchases, transaction amounts, purchase dates and more. All of these characteristics can be used to generate well defined clusters with easy to interpret characteristics.

Not all customers are alike. Consumers usually show a wide variety of behaviours. A lot of times, Segments that are used in businesses are threshold based. With growing number of features and a general theme of personalized products, there is a need for a scientific methodology to group customers together. Clustering based on the behavioural data comes to the rescue. The aim of this analysis is to group credit card holders in appropriate groups to better understand their needs and behaviours and to serve them better with appropriate marketing offers.

We have used k-means algorithm with the K value determined by silhouette score. We have also used PCA for dimension reduction and better visualization to create the appropriate segmentation strategy.

# Motivation for doing this project

We have found several compelling motivations for undertaking this project on customer segmentation using credit card data:

**a)**. **Targeted Marketing**: Customer segmentation allows businesses to tailor their marketing efforts more effectively. By understanding the distinct needs, preferences, and behaviour of different customer segments, companies can create targeted marketing campaigns that resonate with specific groups. This can lead to higher conversion rates and increased customer satisfaction.

**b). Customized Product Offerings:** Through customer segmentation, businesses can identify opportunities to develop and offer customized products or services that cater to the unique needs of different customer segments. This can help enhance customer loyalty and retention by providing solutions that align closely with what each segment values most.

**c). Fraud Detection:** Customer segmentation can also be instrumental in detecting fraudulent activities. By analysing patterns within different customer segments, anomalies indicative of fraudulent behaviour can be identified more readily. This can help financial institutions implement proactive measures to prevent fraud and protect both themselves and their customers.

# About Dataset

This case requires to develop a customer segmentation to define marketing strategy. The sample Dataset summarizes the usage behaviour of 8950 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioural variables.

**Following are the features for credit card dataset: -**

1. CUST_ID: Identification of Credit Card holder (Categorical)

2. BALANCE: Balance amount left in their account to make purchases

3. BALANCE_FREQUENCY: How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)

4. PURCHASES: Amount of purchases made from account

5. ONEOFF_PURCHASES: Maximum purchase amount done in one-go

6.INSTALLMENTS_PURCHASES: Amount of purchase done in instalment

7. CASH_ADVANCE: Cash in advance given by the user

8. PURCHASES_FREQUENCY: How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)

9. ONE_OFF_PURCHASES_FREQUENCY: How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)

10. PURCHASES_INSTALLMENTS_FREQUENCY: How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)

11. CASH_ADVANCE_FREQUENCY: How frequently the cash in advance being paid

12. CASH_ADVANCE_TRX: Number of Transactions made with "Cash in Advanced"

13. PURCHASES_TRX: Number of purchase transactions made

14. CREDIT_LIMIT: Limit of Credit Card for user

15. PAYMENTS: Amount of Payment done by user

16. MINIMUM_PAYMENTS: Minimum amount of payments made by user

17. PRC_FULL_PAYMENT: Percent of full payment paid by user

18. TENURE: Tenure of credit card service for user

# Data Collection Process

We have used the credit card dataset available on Kaggle

https://www.kaggle.com/datasets/arjunbhasin2013/ccdata

# Key Questions of Interest

1. How can we reduce the dimensionality of the dataset using PCA while preserving the variability in the dataset?

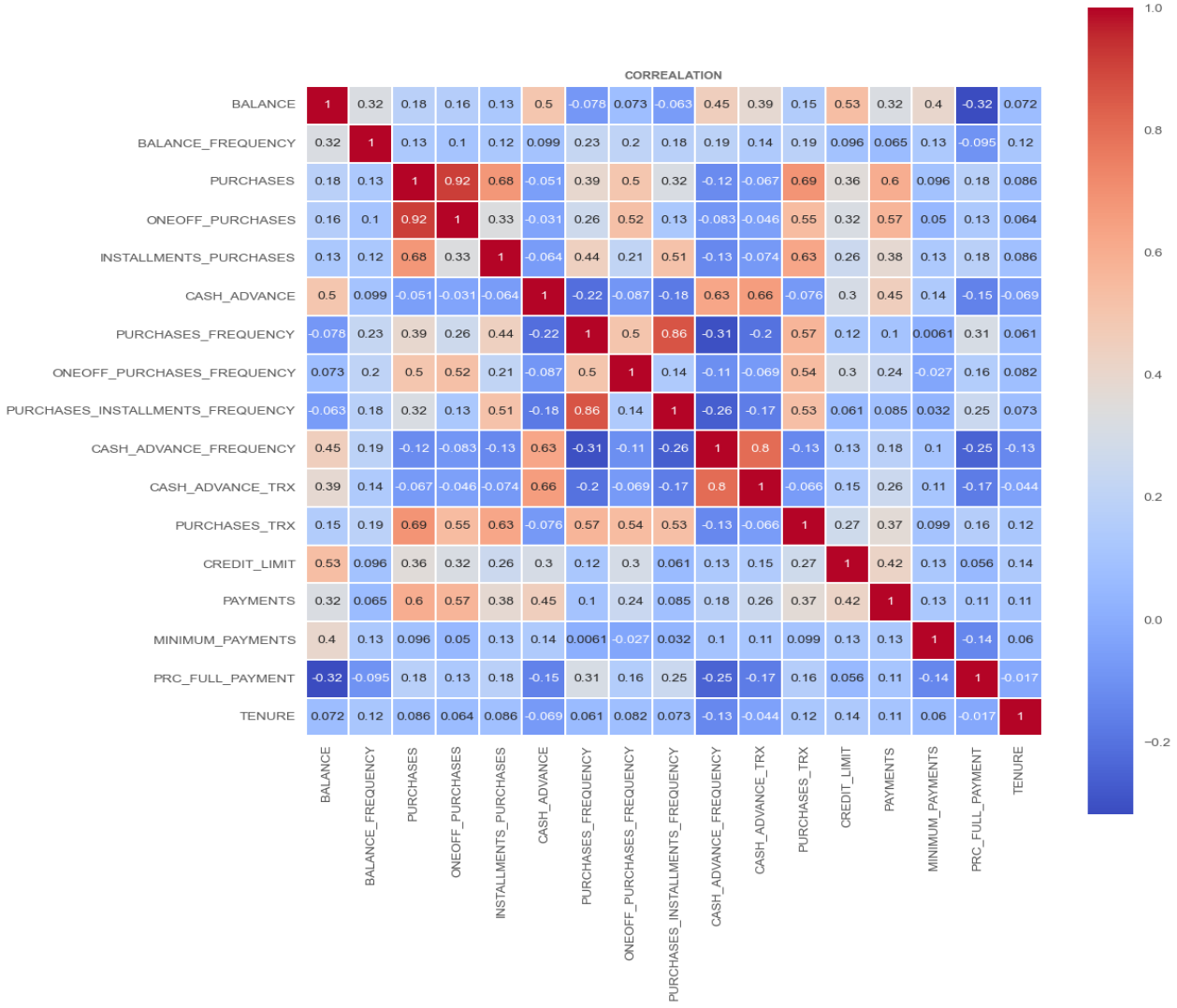
2. How do the identified customer segments differ in terms of spending behaviours, credit card usage, and payment patterns?

3. What recommendations can be made to the credit card company based on the segmentation analysis? How can they tailor their services or marketing strategies to better meet the needs of each segment?

4. How can the results of the segmentation analysis be used to personalize marketing strategies for each customer segment?

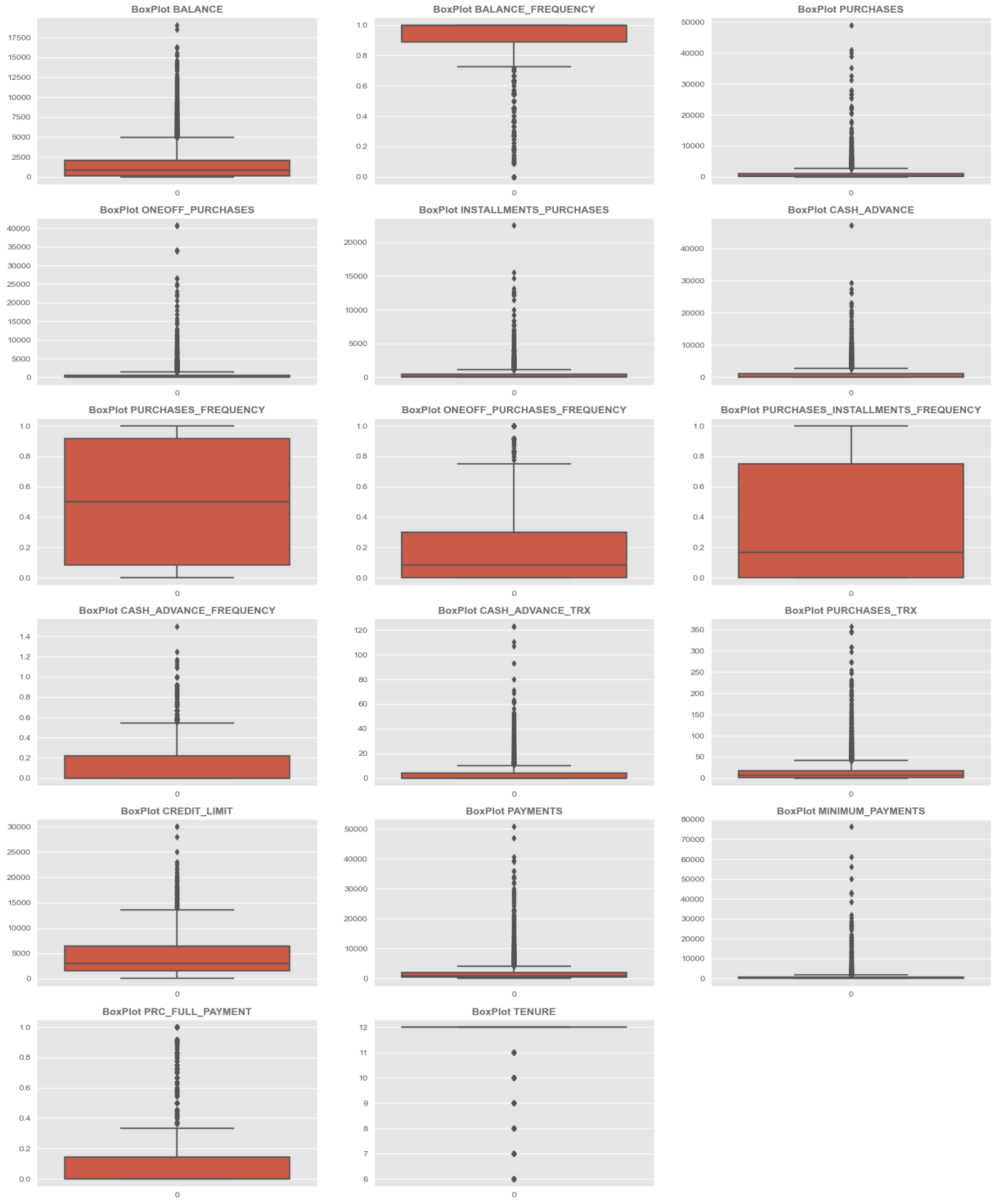
# Important Visualizations

## Visualization Of Correlation matrix of Data



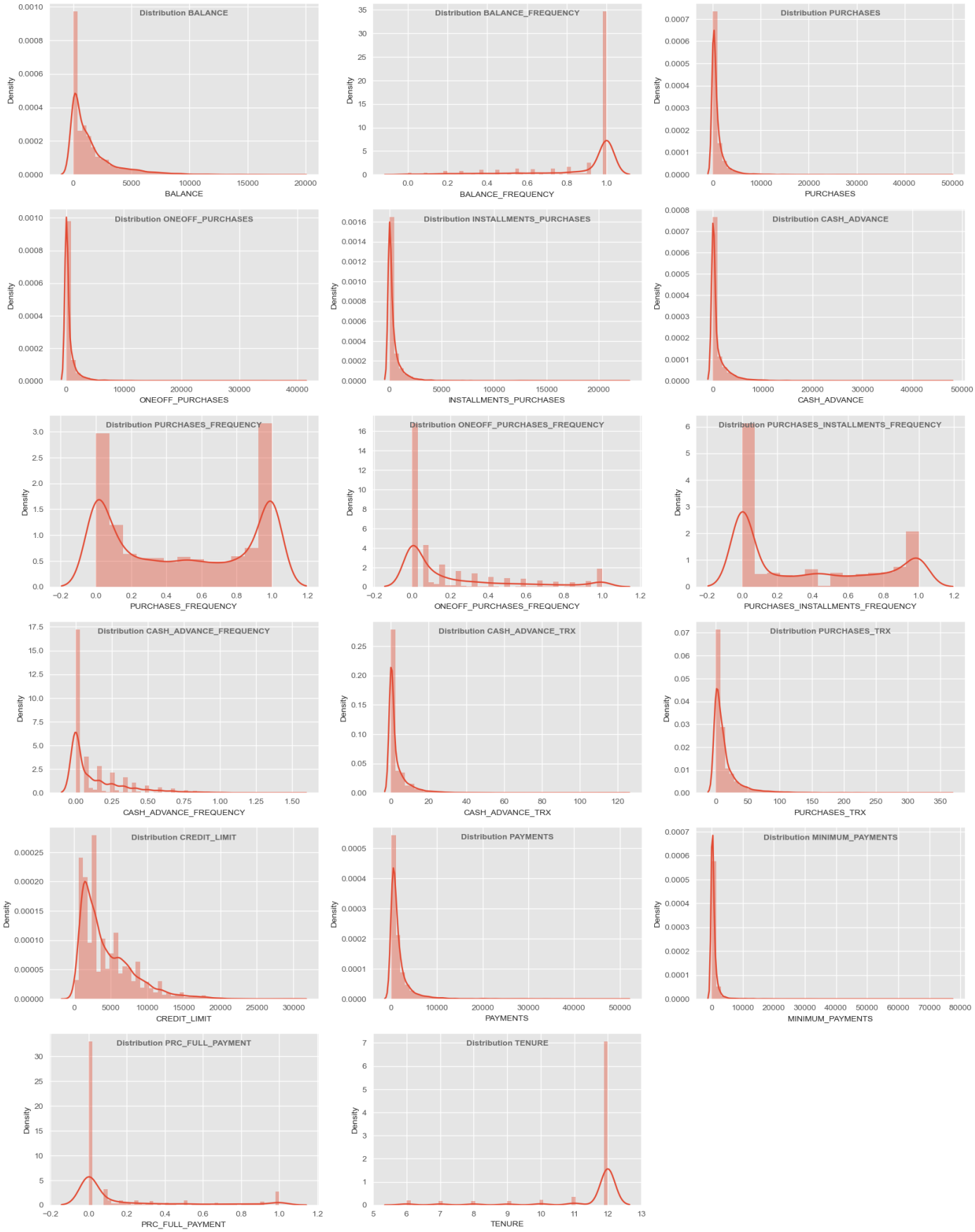

- The features ONEOFF_PURCHASES and PURCHASES have the highest positive correlation among all features i.e. 0.92
- The features PRC_FULL_PAYMENT and BALANCE have the highest negative correlation among all features i.e. – 0.32
- The features MINIMUM_PAYMENTS&PURCHASES_FREQUENCY are highly uncorrelated.

# Box Plots of all the features



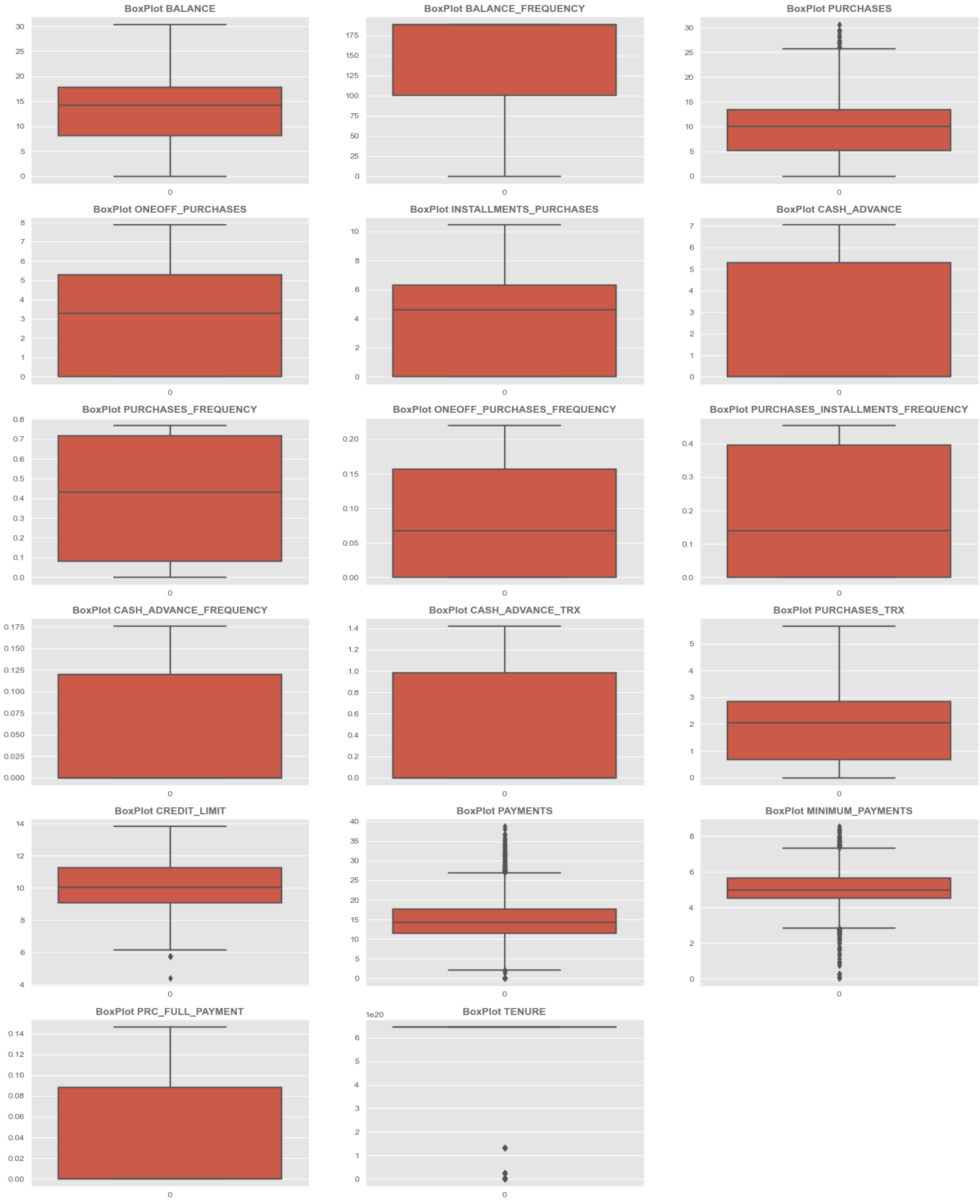

- As we can see that there are many outliers in our data so firstly, we have to deal with these outliers.

# Distribution plots of all the features

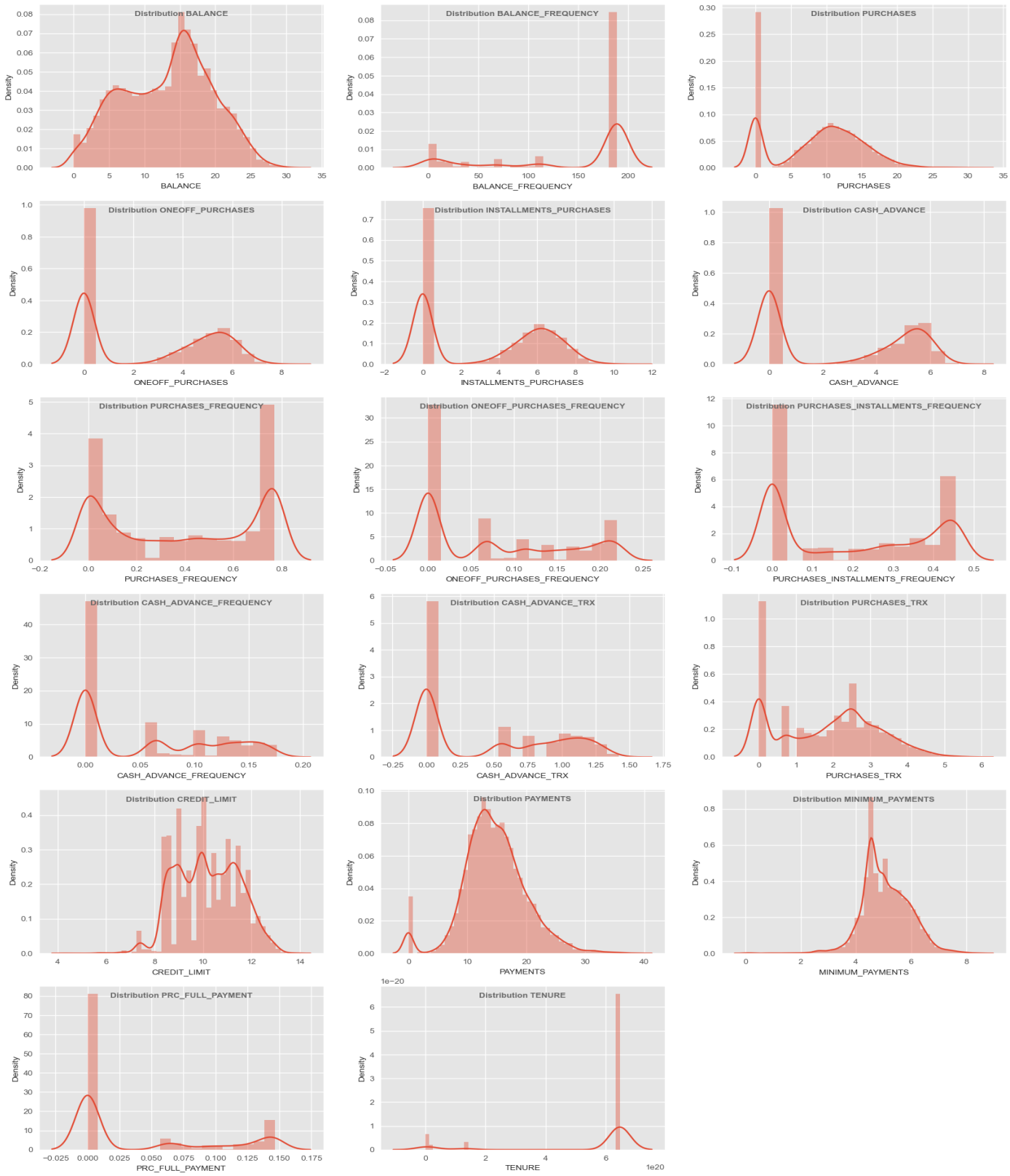



It is clear from the above distribution plots that most of the features are highly skewed.

# After applying the Yeo-Johnson transformations




As we can see the boxplots of transformed data that the most of the outliers has been removed.



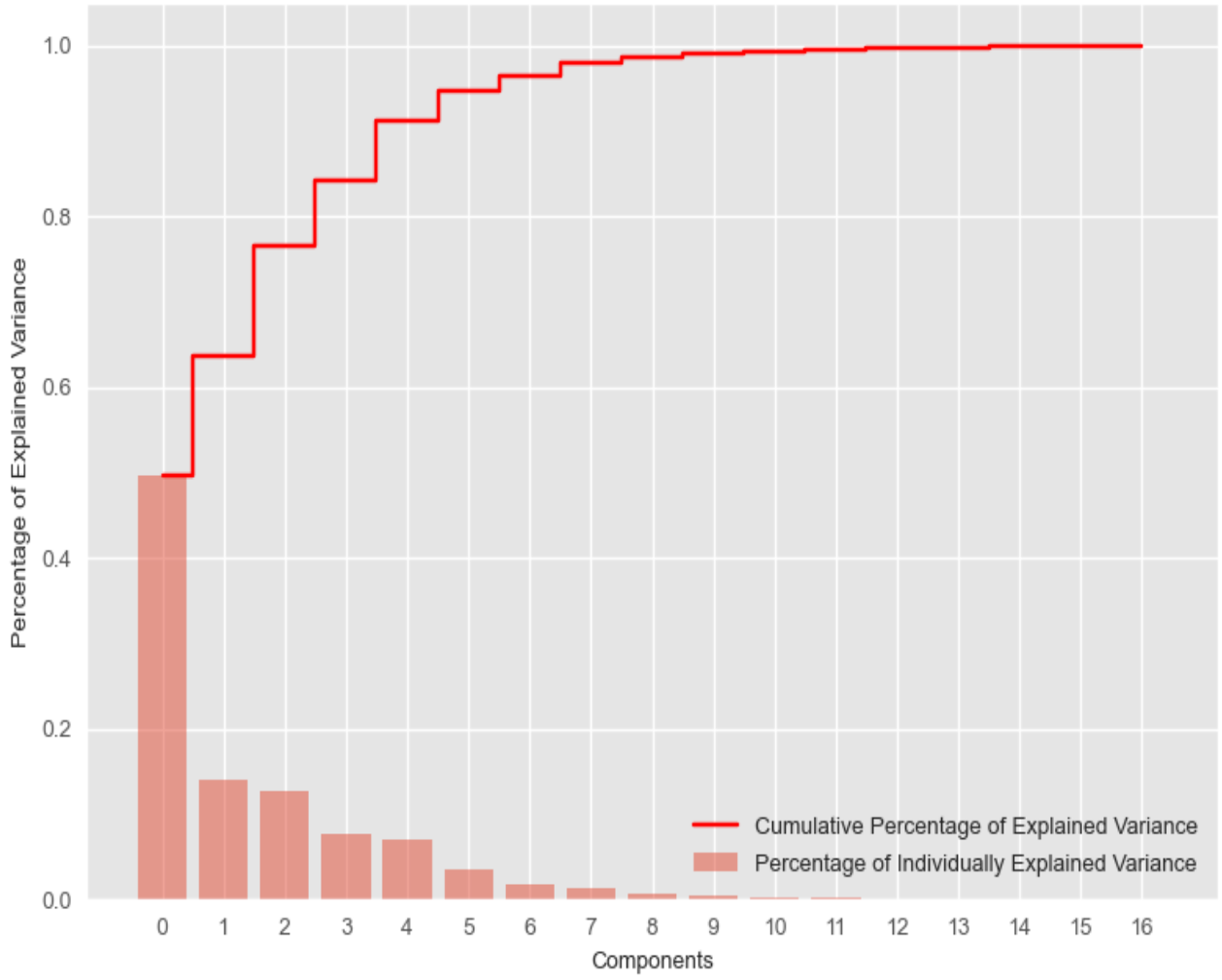

The above distribution plots of transformed data show that now the data is not that much skewed after applying Yeo-Johnson transformations.

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.

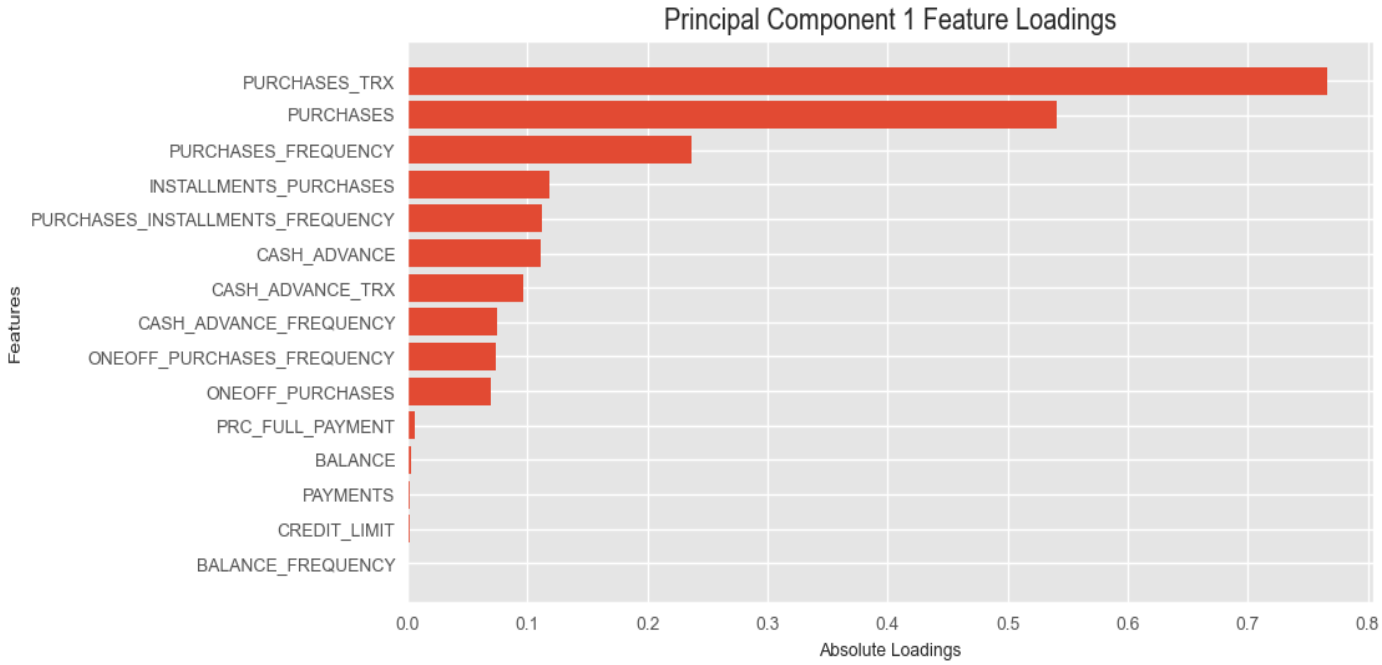
Outlier Handling: Yeo-Johnson transformation can also help in mitigating the impact of outliers. By transforming the data, extreme values can be brought closer to the bulk of the data, reducing their influence on statistical estimates and model performance.

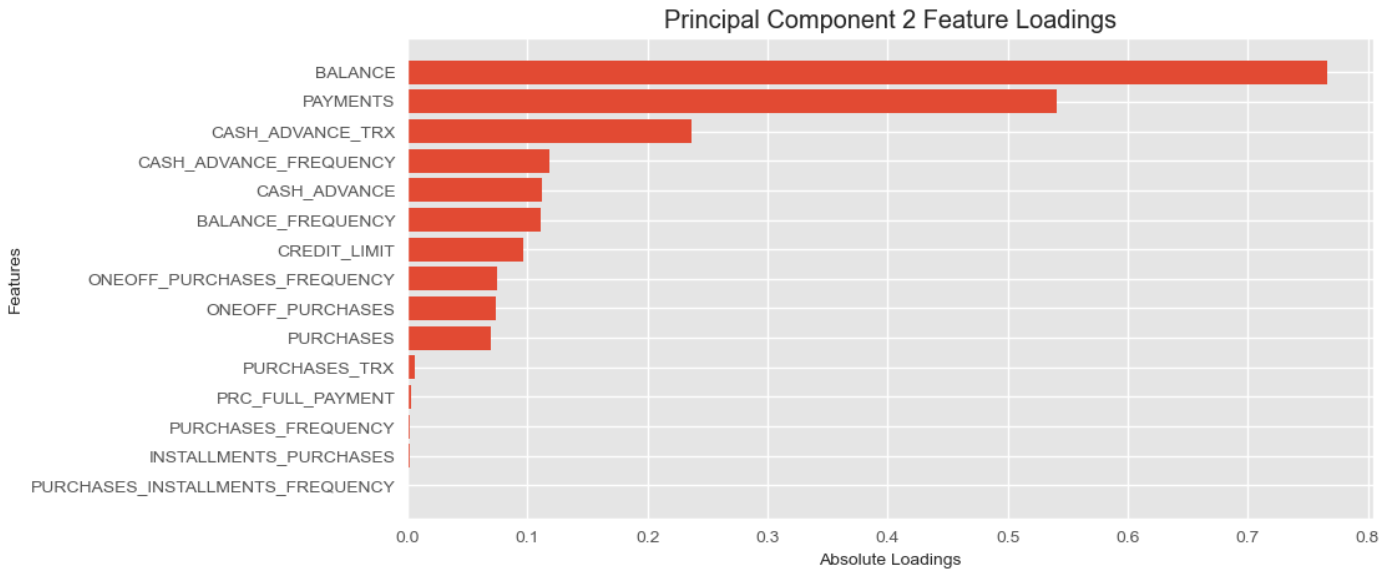
**Visualization of Percentage of Explained Variance by Components**




From the above graph it is clear that first three components explain about 80% variability of the data. So, we have considered three principal components.

## 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

We will use K-means clustering to segment customers based on their purchasing behaviour and other relevant features. These clusters can then be used to recommend similar products or content to users based on their cluster membership. This segmentation can also help tailor marketing strategies and campaigns to specific customer groups.

Now, to choose number of clusters we have used **Silhouette Method**. The silhouette method is a measure used to evaluate the quality of clustering in a dataset. It quantifies how well-separated the clusters are. The silhouette score for each data point is calculated based on the average distance between that point and all other points in the same cluster (a) and the average distance between that point and all points in the nearest neighbouring cluster (b).

The silhouette score (S) for each data point is then calculated as:

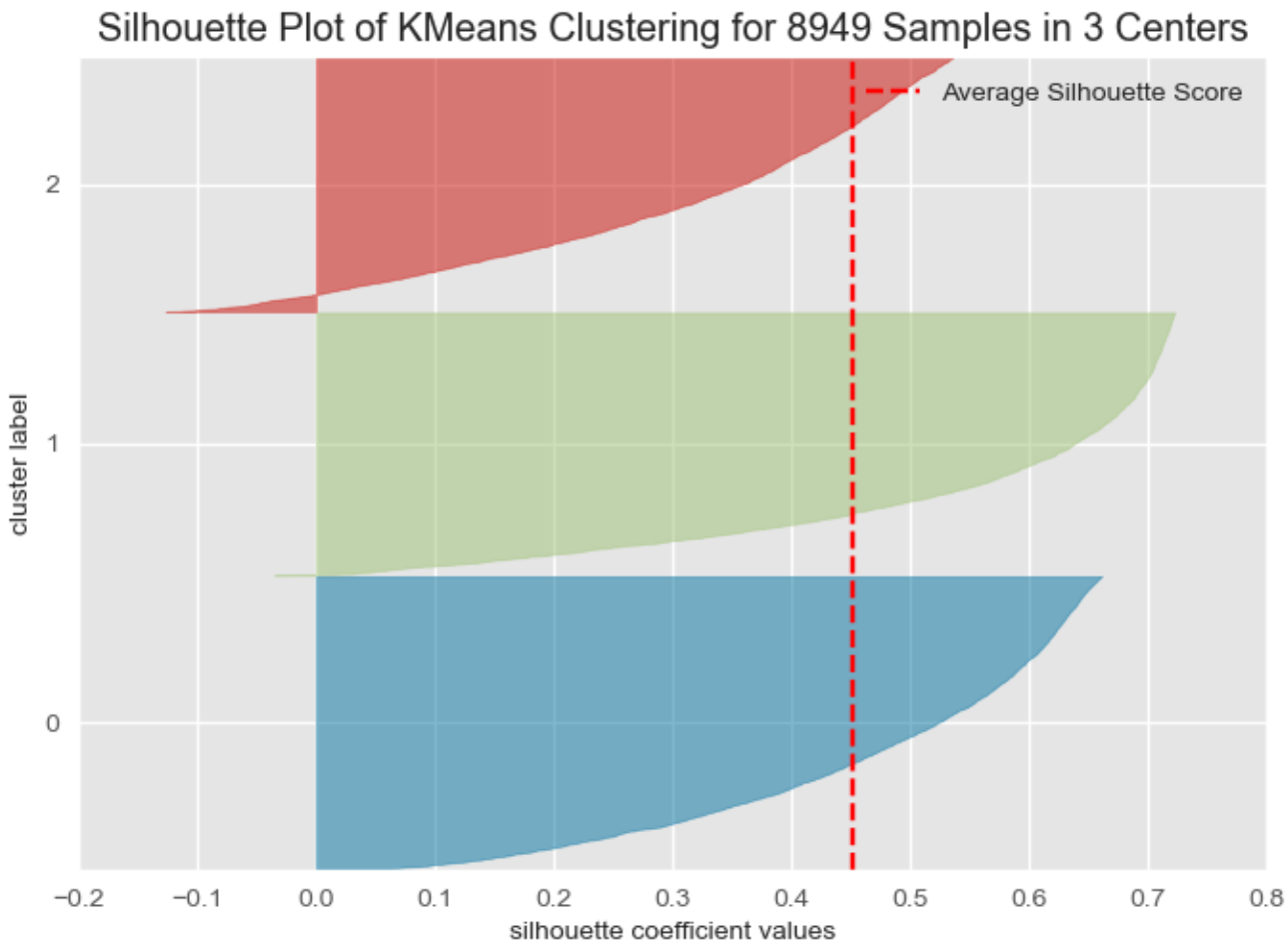
$$\boldsymbol{S = \frac{(b - a)}{\max(a, b)}}$$

The silhouette score ranges from -1 to 1:

- A score close to +1 indicates that the data point is well-clustered and far from neighbouring clusters.
- A score close to 0 indicates that the data point is close to the decision boundary between clusters.
- A score close to -1 indicates that the data point may have been assigned to the wrong cluster.

We have got K = 3 by this method. The Silhouette Plot of K- means clustering for 8949 samples in 3 center is given below.

## Silhouette Plot



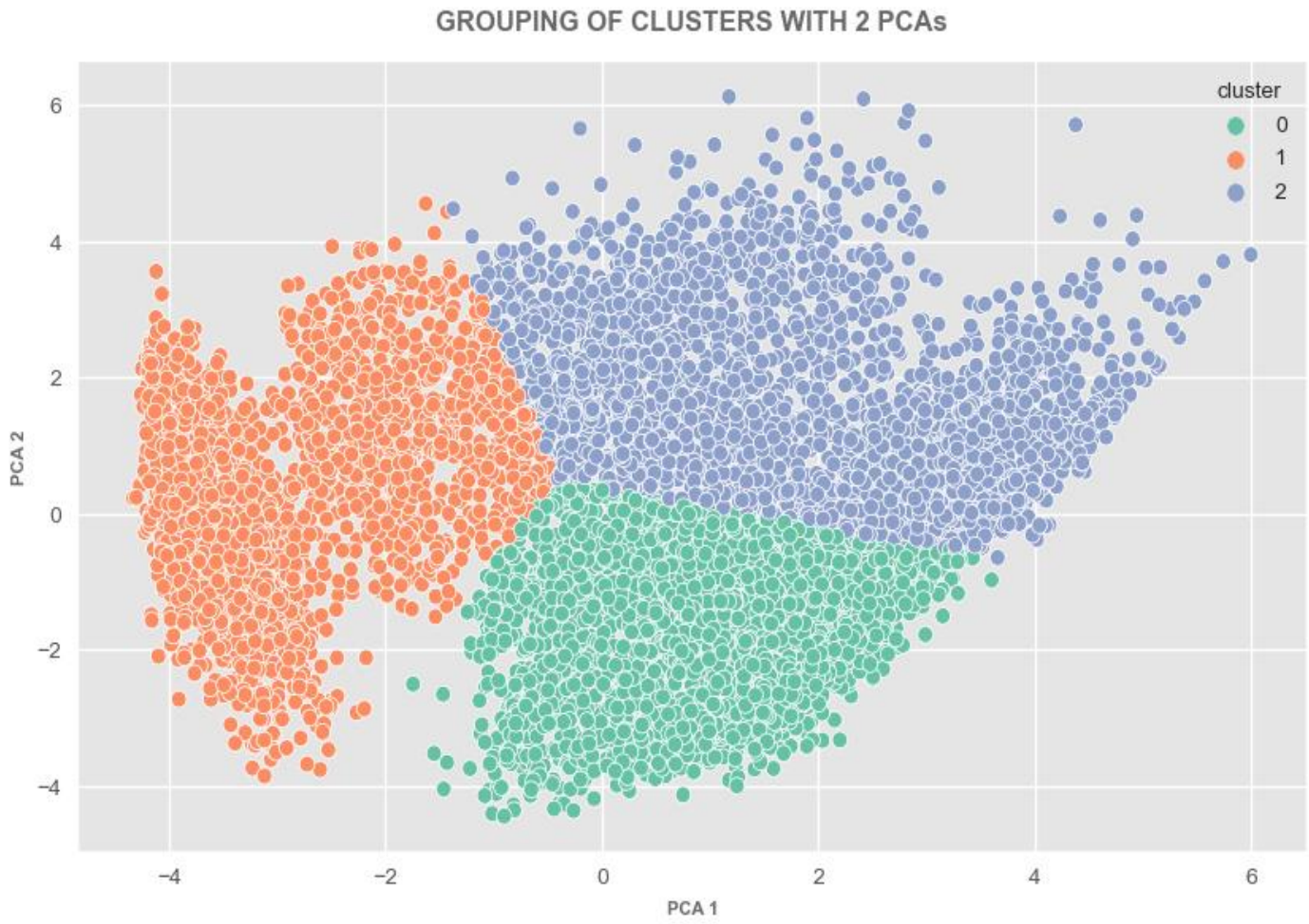

In above plot we can see that there are 3 clusters (Red, Green, Blue). The X-axis represents silhouette coefficient values, ranging from -0.2 to 0.8. and Y-axis represents cluster label.

In the silhouette plot, you may observe clusters with predominantly positive silhouette scores which indicate well-defined clusters. The green cluster has high silhouette coefficients, indicating a well-defined structure. The red cluster also seems fairly well-defined but not as distinct as the green one. The blue cluster might be less distinct or could overlap with other clusters.

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




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

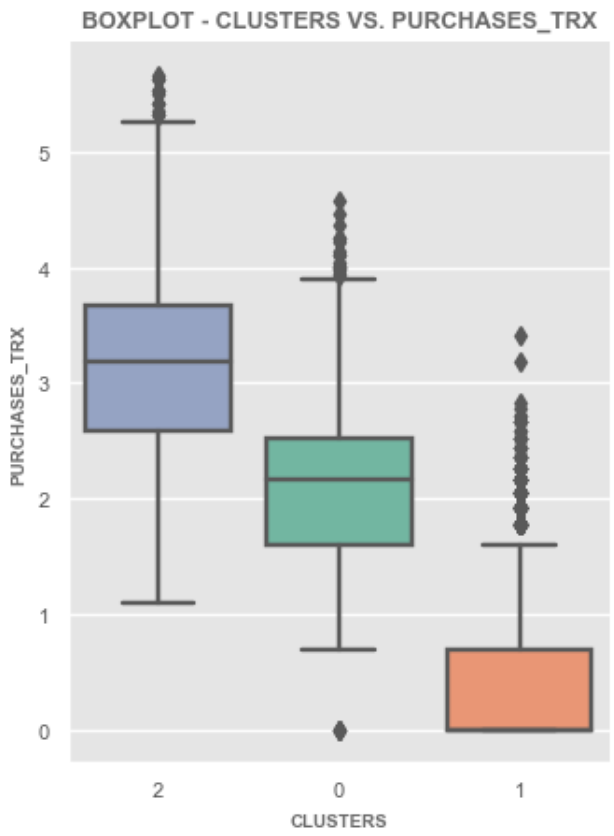
Orange: Concentrated towards the left side, indicating lower values of both PCA 1 and PCA 2.

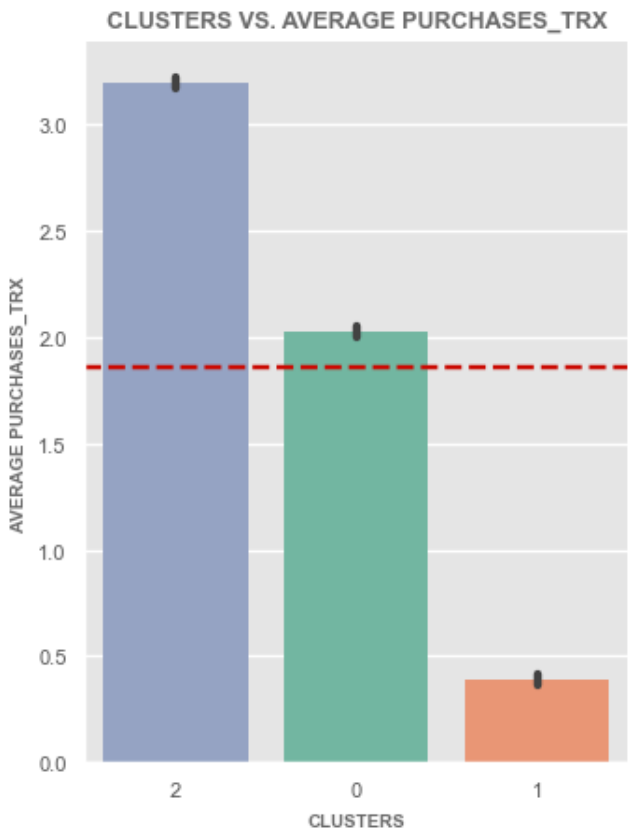
Green: Located in the middle with moderate values of both PCAs.

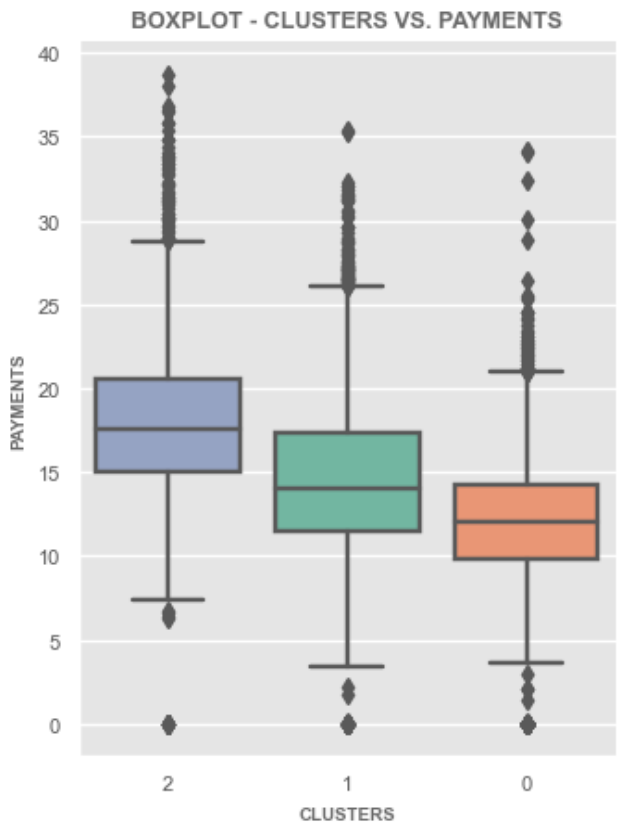
Blue: Spread towards higher values of PCA 1 but similar PCA 2 range as the green cluster.

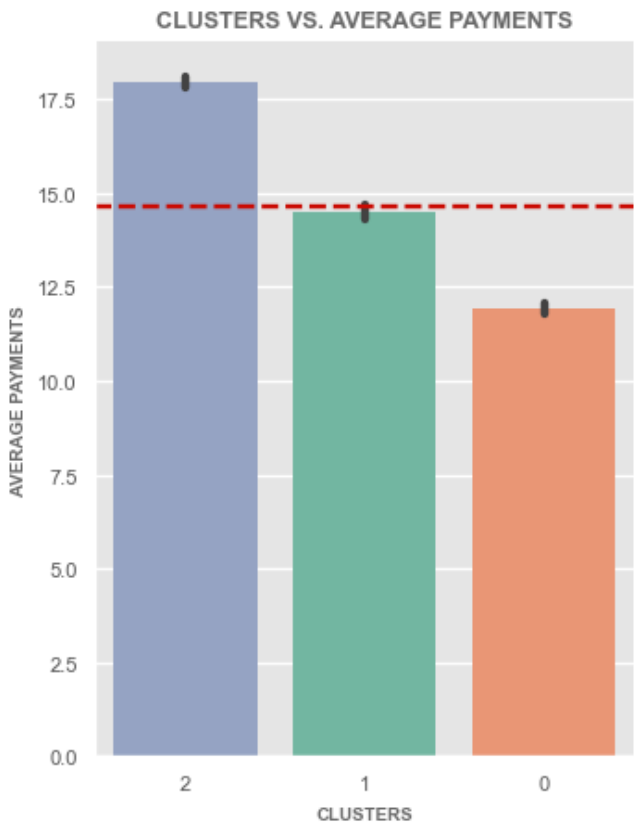
Each colour corresponds to a different cluster, showing how data points are grouped based on similarities.

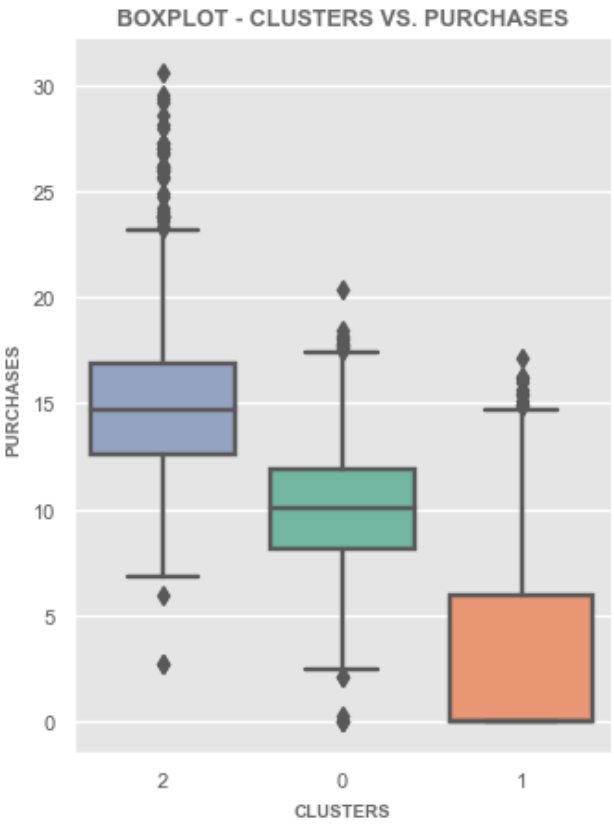
# Comparison of Principal Components among different clusters

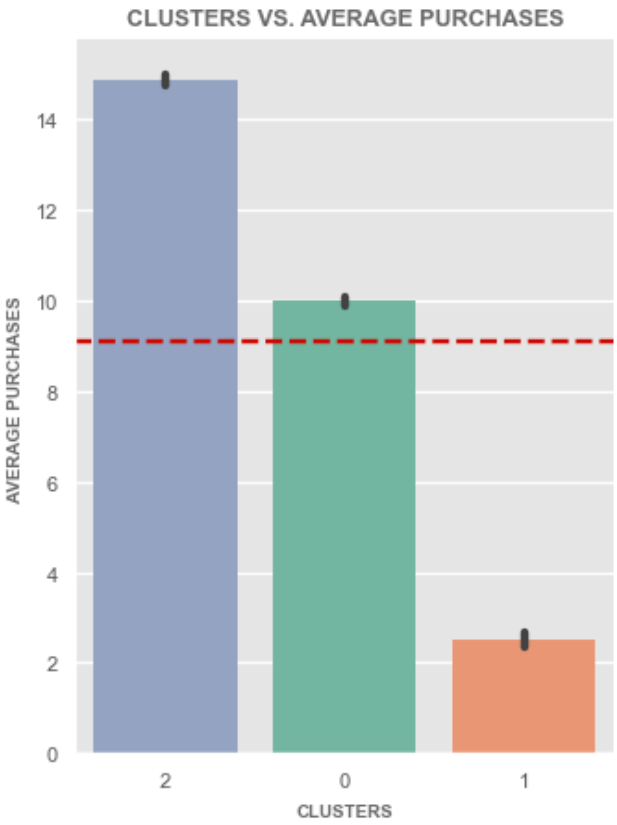


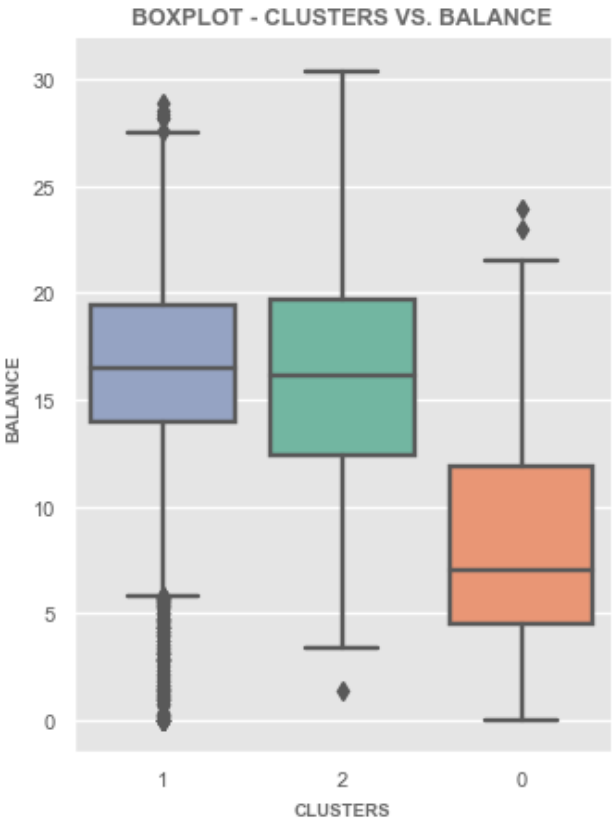



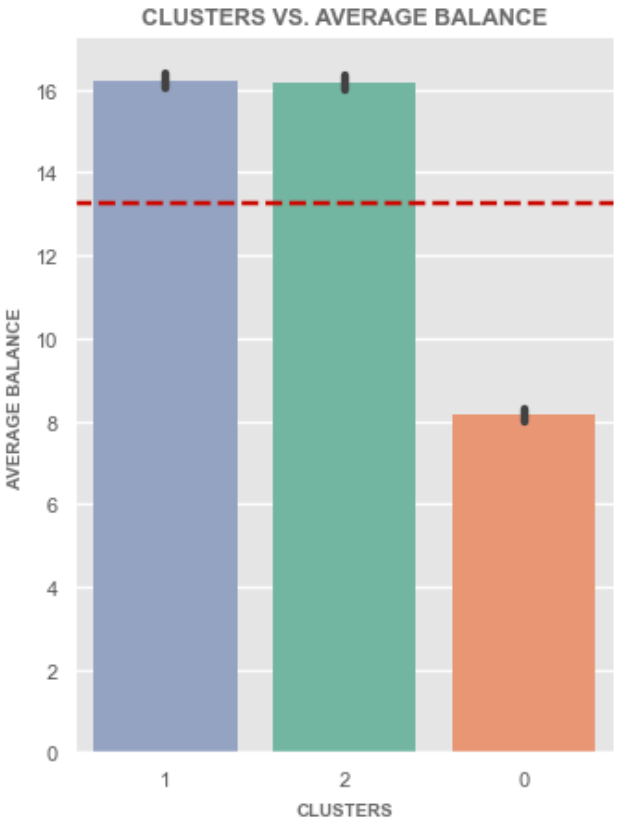




BOXPLOT - CLUSTERS VS. PURCHASES
PURCHASES
0
5
10
15
20
25
30
2
0
1
CLUSTERS


CLUSTERS VS. AVERAGE PURCHASES
AVERAGE PURCHASES
0
2
4
6
8
10
12
14
2
0
1
CLUSTERS


BOXPLOT - CLUSTERS VS. BALANCE
BALANCE
0
5
10
15
20
25
30
1
2
0
CLUSTERS


CLUSTERS VS. AVERAGE BALANCE
AVERAGE BALANCE
0
2
4
6
8
10
12
14
16
1
2
0
CLUSTERS

We have considered four principal components. The above plots compare these PCs across 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

The customer segmentation using PCA and K-Means clustering has provided valuable insights into the credit card data. By reducing dimensionality with PCA, we were able to simplify the dataset while retaining the most significant features. Subsequently, K-Means clustering allowed us to identify distinct groups within the customer base, each characterized by unique purchasing behaviours and credit card usage patterns.

**Recommendations for credit card company based on our segmentation analysis:-**

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

- For the ‘cluster0’ and ‘cluster1’ customers, 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.

- The customers of cluster0 and cluster1 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.

- The customers of cluster1 should be provided valuable scheme to increase the amount of purchases.