

**PS Test : Finalized Version**

Customer segmentation

Customer segmentation is the process of breaking down the customer base into various groups of people that are similar in many ways that are important to marketing, such as gender, age, interests, and various spending habits.

**Table of content**

[**Data processing and visualization** 3](#_Toc162633760)

[1) What is the data shape of the dataset? (number of columns /rows) 3](#_Toc162633761)

[2) Dress the dataset summary? Is there any NaN values? What does it mean? 3](#_Toc162633762)

[3) Is there any duplicate values or missing values? 3](#_Toc162633763)

[4) Perform a cleaning process on the features Minimum-payments and payments. Explain in detail the adopted reasoning behind it. 3](#_Toc162633764)

[5) Is it necessary to keep the customor\_ID? Justify. 4](#_Toc162633765)

[6) What is the new data shape? 4](#_Toc162633766)

[7) Study the correlations? What are your observations? 4](#_Toc162633767)

[8) Plot the density according to each feature. 4](#_Toc162633768)

[**Kmeans clustering** 5](#_Toc162633769)

[1) Normalize/standardize the data 5](#_Toc162633770)

[2) Find the number of clusters using the elbow method. Explain the method and comment the results. 5](#_Toc162633771)

[3) Find the number of clusters using the silhouette method. Explain the method and comment the results. 5](#_Toc162633772)

[**Principal component analysis** 9](#_Toc162633773)

[1) Perform a PCA on the dataset. 9](#_Toc162633774)

[2) Find the number of components to select. 9](#_Toc162633775)

[3) Perform now the kmeans clustering on the reduced data. 9](#_Toc162633776)

[**Conclusion** 11](#_Toc162633777)

## 

# **Data processing and visualization**

##### **1) What is the data shape of the dataset? (number of columns /rows)**

We first import the dataset.

install.packages("ggplot2") library(ggplot2) customer\_data <- read.csv("C:/Users/fefe1/Downloads/Customer\_Data.csv") str(customer\_data)

dim(customer\_data)

The shape of the dataset is 8950 rows and 18 columns.

##### **2) Dress the dataset summary? Is there any NaN values? What does it mean?**

summary(customer\_data)

The summary indicates that there are 1 NA values in the CREDIT\_LIMIT column. To verify this, we take a look at the initial dataset.

A screenshot of a computer

Description automatically generated

We found multiple NA values in the minimum Payments which means there are missing or undefined numerical values present in the data. This could occur due to various reasons such as incomplete data collection, data entry errors, or data processing issues.

##### **3) Is there any duplicate values or missing values?**

We check for the duplicates with sum(duplicated(customer\_data)) that return 0 duplicates finded.

Same for NA, sum(anyNA(customer\_data)) returns 1.

##### **4) Perform a cleaning process on the features Minimum-payments and payments. Explain in detail the adopted reasoning behind it.**

We are going to clean the features Minimum-payments and payments in the dataset using R. Now that we identified missing values we fill missing values using an appropriate imputation methods:

customer\_data$PAYMENTS[[is.na](http://is.na/)(customer\_data$PAYMENTS)] <- median(customer\_data$PAYMENTS, na.rm = TRUE) customer\_data$MINIMUM\_PAYMENTS[[is.na](http://is.na/)(customer\_data$MINIMUM\_PAYMENTS)] <- mean(customer\_data$MINIMUM\_PAYMENTS, na.rm = TRUE)

rows\_with\_na <- which([is.na](http://is.na/)(customer\_data$PAYMENTS) | [is.na](http://is.na/)(customer\_data$MINIMUM\_PAYMENTS)) print(customer\_data[rows\_with\_na, ])

The missing has not been replaced correctly. So we will to remove rows with any missing values from the dataset:

cleaned\_data <- na.omit(customer\_data)

##### **5) Is it necessary to keep the customor\_ID? Justify.**

In the context of customer segmentation for a credit card company, removing the CUST\_ID variable can simplify analysis, reduce bias, and focus attention on meaningful behavioral patterns rather than individual identities.

customer\_data <- customer\_data[, -which(names(customer\_data) == "CUST\_ID")]

##### **6) What is the new data shape?**

dim(customer\_data)

The shape of the dataset is now 8950 rows and 17 columns after removing the CUST\_ID column.

##### **7) Study the correlations? What are your observations?**

correlation\_matrix <- cor(cleaned\_data) correlation\_matrix

Positive Correlations:

* PURCHASES correlates moderately positively with ONE\_OFF\_PURCHASES (0.92) and INSTALLMENTS\_PURCHASES (0.68).
* CASH\_ADVANCE has a moderate positive correlation with BALANCE (0.50) and CASH\_ADVANCE\_TRX (0.66).
* PURCHASES\_FREQUENCY strongly correlates with PURCHASES\_INSTALLMENTS\_FREQUENCY (0.86).

Negative Correlations:

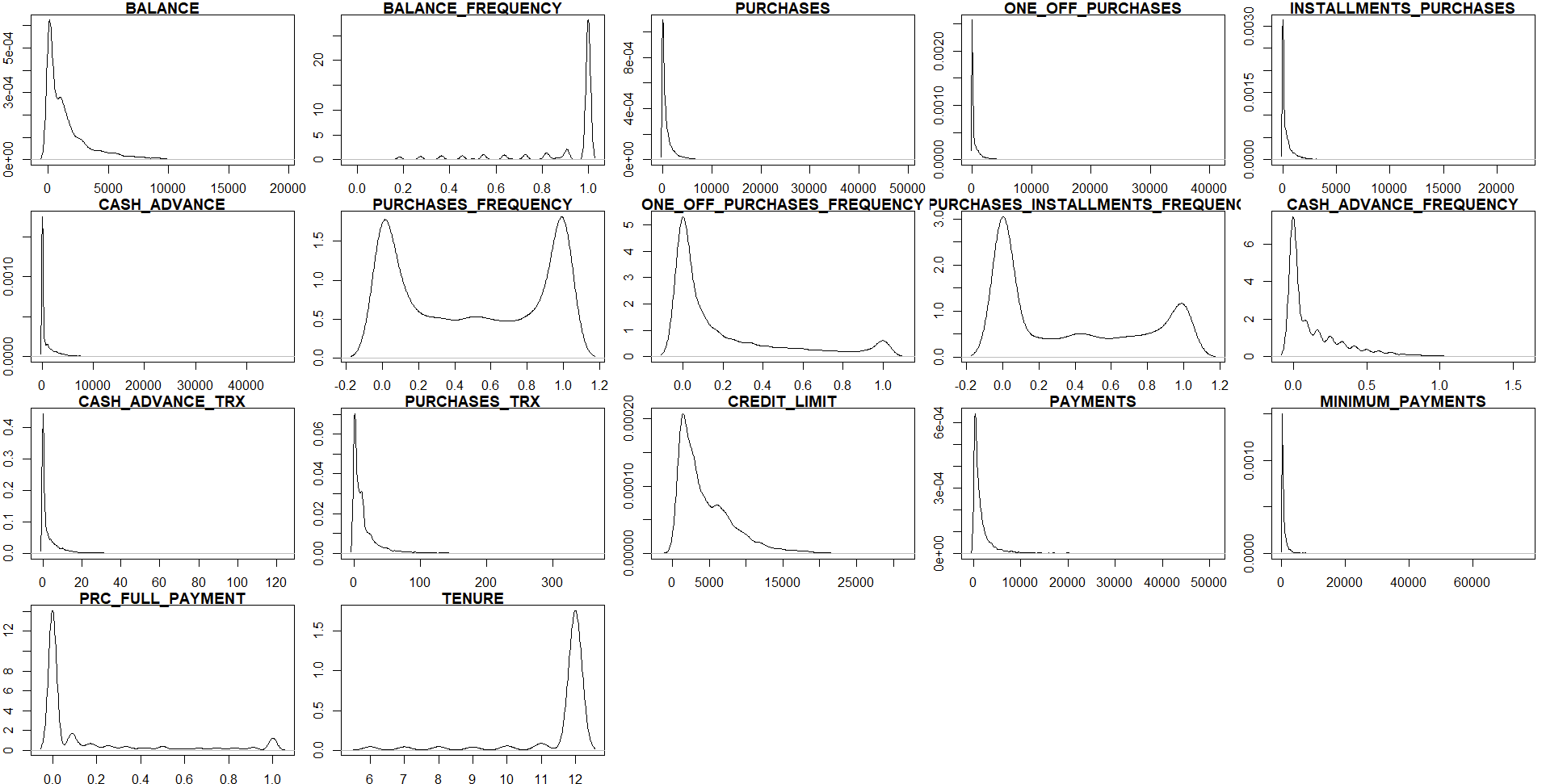
* CASH\_ADVANCE\_FREQUENCY is moderately negatively correlated with PURCHASES\_FREQUENCY (-0.31).

##### **8) Plot the density according to each feature.**

par(mfrow=c(4, 5), mar=c(2, 2, 1, 1)) # Adjust margins as needed

for (i in 1:ncol(cleaned\_data)) {if (!is.numeric(cleaned\_data[, i])) next

density\_values <- density(cleaned\_data[, i], na.rm = TRUE) plot(density\_values, main = colnames(cleaned\_data)[i], xlab = "", ylab = "") }



# **Kmeans clustering**

##### **1) Normalize/standardize the data**

if ("CUST\_ID" %in% colnames(customer\_data)) { customer\_data <- customer\_data[, -which(names(customer\_data) == "CUST\_ID")] }

normalized\_data <- scale(customer\_data)

head(normalized\_data)

Data appears to be well normalized/standardized. Each column has a mean of approximately 0 and a standard deviation of approximately 1, which indicates that the data has been successfully standardized.

##### **2) Find the number of clusters using the elbow method. Explain the method and comment the results.**

cleaned\_data <- normalized\_data[complete.cases(normalized\_data), ]

library(FactoMineR)

library(factoextra)

pca\_result <- PCA(cleaned\_data, graph = FALSE)

reduced\_data <- as.data.frame(pca\_result$ind$coord[, 1:5]) # Adjust the number of components as needed

elbow\_nb <- fviz\_nbclust(reduced\_data, kmeans, method = "wss")

print(elbow\_nb)

A graph with a line

Description automatically generated

So we take the number of cluster k=3 or k=4 are there are at the postion of an elbow.

##### **3) Find the number of clusters using the silhouette method. Explain the method and comment the results.**

cleaned\_data <- na.omit(customer\_data)

cleaned\_data <- cleaned\_data[, -which(names(cleaned\_data) == "CUST\_ID")]

scaled\_data <- scale(cleaned\_data)

library(cluster)

library(FactoMineR)

library(factoextra)

pca\_result <- PCA(scaled\_data, graph = FALSE)

reduced\_data <- as.data.frame(pca\_result$ind$coord[, 1:5])

silhouette\_nb <- fviz\_nbclust(reduced\_data, kmeans, method = "silhouette")

print(silhouette\_nb)

A graph with a line and a line

Description automatically generated

Using the silhouette method we can guess the better number of cluster is the first one

**4) Perform the kmeans with k=3. Plot the obtained clusters with the features of customers.**

* 1. Without the PCA

library(cluster)

library(factoextra)

kmeans\_res <- kmeans(cleaned\_data, centers = 3, nstart = 25)

fviz\_cluster(kmeans\_res, data = cleaned\_data, geom = "point", stand = FALSE,

ellipse.type = "convex", ellipse.level = 0.68,

main = "Clusters of Customers")

A graph showing a green and blue triangle

Description automatically generated with medium confidence

* 1. With the PCA

library(cluster)

library(FactoMineR)

library(factoextra)

pca\_result <- PCA(cleaned\_data, graph = FALSE)

reduced\_data <- as.data.frame(pca\_result$ind$coord[, 1:5]) # Adjust the number of components as needed

kmeans\_res <- kmeans(reduced\_data, centers = 3, nstart = 25)

fviz\_cluster(kmeans\_res, data = reduced\_data, geom = "point", stand = FALSE,

ellipse.type = "convex", ellipse.level = 0.68,

main = "Clusters of Customers")

A graph showing a number of customers

Description automatically generated

**5) Which distance you used in the kmeans clustering?**

In the previous code, the default distance measure used in the kmeans clustering function is the Euclidean distance.

**6) Define the features of each cluster**

cluster\_means <- tapply(cleaned\_data, km\_res$cluster, FUN = colMeans)

cluster\_means\_df <- as.data.frame(cluster\_means)

cluster\_means\_df

**cluster\_means**

Cluster 1: 3.955850e+03, 9.609731e-01, 3.634447e+02, 2.344233e+02, 1.290840e+02, 3.780005e+03, 2.231949e-01, 1.052709e-01, 1.373056e-01, 4.380039e-01, 1.212631e+01, 5.264325e+00, 6.597672e+03, 2.951292e+03, 1.830302e+03, 3.075607e-02, 1.137461e+01

Cluster 2: 2297.4859546, 0.9821114, 4439.2747672, 2800.6530517, 1639.1579224, 515.2567841, 0.9519292, 0.6542323, 0.7715694, 0.0685312, 1.7353448, 58.4155172, 7814.5258621, 4375.3236022, 1294.5198693, 0.3007700, 11.9189655

Cluster 3: 8.103108e+02, 8.594934e-01, 5.324137e+02, 2.724590e+02, 2.602901e+02, 3.165991e+02, 4.812870e-01, 1.449622e-01, 3.531975e-01, 6.799476e-02, 1.183154e+00, 9.144200e+00, 3.294021e+03, 9.474510e+02, 5.111761e+02, 1.669120e-01, 1.150248e+01

*\*These results are based on the k-means with PCA*

**Cluster 1:** High values in spending (Feature 1: 3955.85) and engagement (Feature 6: 3780.01). Possibly represents highly engaged and active customers.

**Cluster 2:** High values in spending (Feature 3: 4439.27) and engagement (Feature 13: 7814.53). May indicate customers with distinct spending patterns compared to Cluster 1.

**Cluster 3:** Moderate to high values in spending (Feature 13: 3294.02) and engagement (Feature 15: 511.18). Represents customers with a moderate level of engagement and spending compared to other clusters.

# **Principal component analysis**

##### **1) Perform a PCA on the dataset.**

pca\_result <- prcomp(cleaned\_data, scale. = TRUE)

summary(pca\_result)

##### **2) Find the number of components to select.**

To find the number of principal components to select, we can use scree plot in this case :

plot(summary(pca\_result)$importance[2,], type = "b", xlab = "Principal Component", ylab = "Eigenvalue", main = "Scree Plot")

A graph with lines and dots

Description automatically generated with medium confidence

After examining the scree plot we identify that the point where the eigenvalues start to level off is around 3. So we will select 3 components

##### **3) Perform now the kmeans clustering on the reduced data.**

- Find the number of clusters using the silhouette method. Is it 3 clusters?

library(cluster)

library(FactoMineR)

library(factoextra)

pca\_result <- PCA(cleaned\_data, scale.unit = TRUE, ncp = 3, graph = FALSE)

reduced\_data <- as.data.frame(pca\_result$ind$coord)

silhouette\_nb <- fviz\_nbclust(reduced\_data, kmeans, method = "silhouette")

print(silhouette\_nb)

A graph with a line and a number of clusters

Description automatically generated

Yes the number of clusters using the silhouette method is 3.

- Plot the obtained clusters with the features of customers.

kmeans\_res <- kmeans(reduced\_data, centers = 3, nstart = 25)

fviz\_cluster(kmeans\_res, data = reduced\_data, geom = "point",

ellipse.type = "convex", ellipse.level = 0.68,

main = "Clusters of Customers")

A graph showing a green triangle with red dots

Description automatically generated

- Define the features of each cluster.

cluster\_centroids <- kmeans\_res$centers

cluster\_centroids\_df <- as.data.frame(cluster\_centroids)

colnames(cluster\_centroids\_df) <- colnames(reduced\_data)

print(cluster\_centroids\_df)

A number with black text

Description automatically generated with medium confidence

For cluster 1 the feature is in majority Dim 2, then cluster 2 is defined by Dim 1 and cluster 3 mostly by Dim 2.

# **Conclusion**

In conclusion, customer segmentation using k-means clustering and principal component analysis revealed distinct groups with varying spending patterns and engagement levels. The analysis identified key features driving customer behavior, enabling targeted marketing strategies for each segment. Overall, these insights provide valuable guidance for businesses seeking to optimize customer relationships and enhance marketing effectiveness.