VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

J Component Report

Programme: B. Tech
Course Title: Foundations of Data Analytics
Course Code: CSE3505
Slot: F1/F2

Title: Customer Segmentation

Team Members:

Krishna Sarawagi	20BRS1115
Deepshika V	20BCE1615
Rishabh Jain Patni	20BRS1065
Aakash Hariharan	20BRS1225

Faculty: Dr. S. Brindha

Sign:

Date: 14-11-2022

1. Abstract

- Due to emergence of digitization and many competitors in the market it has become a competing problem among businesses to find new buyers, while keeping the old ones.
- As a result of this, a need for an exceptional customer service becomes necessary for businesses regardless of their size.
  This understanding is possible through structured customer service.
- The ideas of Data Analytics can be used in this context of market analytics where we can use techniques such as clustering algorithm for this purpose.

2. Scope

- Consumer segmentation enables us to uncover valuable audience insights that can help to dictate the content, targeting and media buying strategy of your digital campaigns. It allows us to understand which audiences exist, who to target and why, and the most effective way to reach them.
- Customer segmentation helps formulate a more effective marketing strategy, optimizes the journey of the customer, predicts customer behavior, and personalizes the customer experience.
- It improves customer loyalty and retention, conversion metrics and supports product development.
- Customer segmentation benefits in wise and efficient use of resources, cost efficiency, helps know the customers better and retain the customers, it gives higher rate of success and customer satisfaction and targets each customer based on their interests.

3. Problem Statement

- In this project, we are going to develop a Customer Segmentation Strategy to define the Marketing Strategy for a Credit Card Company.
- To achieve the objective, an unsupervised learning method, K-Means Clustering Algorithm will be used.
- The process will include Data Preparation, Exploratory Data Analysis, Data Pre-Processing, K-Means Clustering, Principal Component Analysis (PCA), and finally the Conclusion including suggestions based on the analysis.

4. Introduction to Customer Segmentation

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group.

Types of customer segmentation:

i. Demographic & Socioeconomic Segmentation

- Demographic segmentation is one of the most widely used and simple-to-understand types of consumer segmentation.

- At its most basic level, demographic segmentation looks at age and gender, but socioeconomic factors such as household income, marital status, occupation or industry and education (e.g., are they educated to degree level) also fall into this segment.

ii. Geographic Segmentation

- Another widely used type of consumer segmentation is geographic segmentation, which clusters audiences by location, usually at country and city level.
- The needs and desires of consumers can vary wildly from city to city, especially in larger countries with particularly topographically or economically diverse regions, e.g., the US. Geographic segmentation is extremely valuable for international businesses to help them to define the nuances in consumer demands in different regions and identify any cultural characteristics which can form part of their targeting strategy.

iii. Behavioral Segmentation

- Behavioural segmentation, as the name suggests, segments audiences by behaviours and habits.
- This includes buying habits and online behaviour, including platform and technology usage and most active hours online.
- This type of consumer segmentation can help your business to tailor your marketing efforts to garner the most positive results, e.g., scheduling organic social posts or email campaigns for when your audience is most active online.

iv. Psychographic Segmentation

- This type of consumer segmentation is like Behavioral Segmentation in that it seeks to understand an audience at a deeper level than demographics.
- The difference is that it is focused on their personality, beliefs, and interests, rather than a measure of their past actions.

v. Social Media Segmentation

- Social Media segmentation groups your different audiences by platform, so you can understand where they are most active.
- This is a useful insight as it can help you to determine where to place your ads or amplify your content for the best results.

5. Literature review

- The commercial world has become more and more competitive over the years, as every organization has desires to meet the need of their customers.
- Since, customers can vary according to their needs, wants, demographics, size, taste, etc. it becomes very difficult for the businesses to meet everyone's demand.
- As the era of digitalization and globalization boomed, companies got access to a very precious tool which was earlier not available to them in great abundance, that tool is "DATA".
- Companies have billions of data about their customers, suppliers, and operations and millions of internally connected data.

- Improved forecasting, saving money, increased efficiency and a lot of other improvements can be achieved with the help of data.
- Now, clustering the data is a process of grouping the information into a dataset based on some commonalities.
- There are many algorithms which can be used for this task based on the conditions given but there does not exist any universal one. So, we need to choose the appropriate clustering techniques as per our requirement.
- K-Means is one of widely used classification algorithms. It relies in the Centro, where each data point is placed in one of the overlapping ones, which is pre-sorted in K-algorithm.
- Clusters are created that correspond to the hidden patterns in the data that provide the necessary information to help decide the execution process.

6. Limitations of Algorithms

- Needs prior specification for the number of cluster centers
- If there are two highly overlapping data, then it cannot be distinguished and cannot tell that there are two clusters
- With the different representations of the data, the results achieved are also different
- Euclidean distance can unequally weigh the factors
- It gives the local optima of the squared error function
- Sometimes choosing the centroids randomly cannot give fruitful results

- It can be used only if the meaning is defined
- Cannot handle outliers and noisy data
- Do not work for the non-linear data set
- Lacks consistency
- Sensitive to scale
- If very large data sets are encountered, then the computer may crash
- Prediction issues

7. Algorithm

- Clustering Algorithm is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses.
- Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.
- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.
- Clustering algorithms can be of various types like Density-based, Distribution-based, Centroid-based or Hierarchical-based. Some of the clustering algorithms are: K-means clustering algorithm, DBSCAN clustering algorithm, Gaussian Mixture Model algorithm, etc.

8. K-Mean Clustering

This algorithm is an iterative algorithm that partitions the dataset according to their features into K number of predefined non-overlapping distinct clusters or subgroups. It makes the data points of inter clusters as similar as possible and also tries to keep the clusters as far as possible. It allocates the data points to a cluster if the sum of the squared distance between the cluster's centroid and the data points is at a minimum, where the cluster's centroid is the arithmetic mean of the data points that are in the cluster. A less variation in the cluster results in similar or homogeneous data points within the cluster.

The algorithm will categorize the items into k groups or clusters of similarity. To calculate that similarity, we will use the Euclidean distance as measurement.

The algorithm works as follows:

1. First, we initialize k points, called means or cluster centroids, randomly.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that cluster so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters.

The "points" mentioned above are called means because they are the mean values of the items categorized in them.

9. Dataset Description

The Dataset that we will be using is an open-source dataset about the Credit Card usage of the Customers. It summarizes the usage behaviour of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioural variables.

Following are the Variables in our Dataset: -

- CUSTID:
  Identification of Credit Card holder (Categorical)
- BALANCE:
  Balance amount left in their account to make purchases
- BALANCEFREQUENCY:
  How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- PURCHASES:
  Number of purchases made from account
- ONEOFFPURCHASES:
  Maximum purchase amount done in one-go
- INSTALLMENTSPURCHASES:
  Amount of purchase done in instalment
- CASHADVANCE:
  Cash in advance given by the user
- PURCHASESFREQUENCY:
  How frequently the Purchases are being made, score between 0 and 1 (1 = frequency, 0 = not frequently purchased)

- ONEOFFPURCHASESFREQUENCY:
  How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
- PURCHASESINSTALLMENTSFREQUENCY:
  How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
- CASHADVANCEFREQUENCY:
  How frequently the cash in advance being paid
- CASHADVANCETRX:
  Number of transactions made with “Cash in Advanced”
- PURCHASESTRX:
  Number of purchase transactions made
- CREDITLIMIT:
  Limit of Credit Card for user
- PAYMENTS:
  Amount of Payment done by user.
- MINIMUM_PAYMENTS:
  Minimum number of payments made by user.
- PRCFULLPAYMENT:
  Percent of full payment paid by user.
- TENURE:
  Tenure (repayment period) of credit card service for user.

10. Work Flow

a. Data Preparation:

- In this phase, we load the dataset and the required libraries. Prepare it. Check for any missing values and if there are any missing values in any observational row then we will omit that row itself.

b. Exploratory Data Analysis

- In this phase, we will explore the data variables, and find out any pattern that can indicate any kind of correlation between the variables.

c. Data Pre-Processing

- Here, we will scale the data set on the same scale using z-scaling, since the data set used is not on the same scale.

d. K-Means Clustering

- Now, we will do the clustering but first we will try to find the optimal number of clusters for our clustering problem using Elbow Method and Silhouette Method. After getting the value for optimal number of clusters we will apply the K-Means Clustering.

e. Goodness of Fit

- The evaluation of the clustering results is very important and for that it can be seen using 3 values i.e., Within Sum of Squares, Between Sum of Squares and Total Sum of Squares.
- For a good cluster, a low value of 'Within Sum of Squares' and 'Total Sum of Squares' near to 1 is needed.

f. Cluster Profiling

- After getting the information about the cluster of each observation, we are going to combine the cluster column into the dataset to interpret each characteristic of the cluster.
- Also, we may add the CUST_ID that we had initially dropped to find out which cluster each customer belongs to in the end.

g. PCA

- PCA is used for the purpose of Dimensionality Reduction. The main purpose is to reduce the no. of variables in the data while retaining as much relevant information as possible.
- Dimensionality Reduction is useful in solving the problem of high-dimensional data. Then, based on the results of PCA, Individual Factor Map and Variables Factor Map will be used to outliers and variable contribution respectively.

11. Results

Based on the analysis of the data set that we have done using K-Means Clustering, we have results as:

1. Cluster 1:
    - These are the customers with lowest amount of all the purchases, not much withdrawals, indicates not many transactions of the credit card as compared to the other clusters.

2. Cluster 2:
    - These are the customers with lowest amount of withdrawal and frequency; however, they have the highest amount of all purchases. They are people with longest tenure and highest percent of full payments paid, indicating that they are aware of their credits.

3. Cluster 3:
    - These are people with high amount of balance, high cash advance and high credit limit. Their balance also seemed to be updated frequently, indicates many transactions of the credit card.
    - The customers of this cluster also have high amount of minimum payments, however, lowest percent of full payments paid, indicating higher loans amount and often they like to withdraw a lot of money from the credit card.

4. Cluster 4:
    - Now, these customers are the one with lowest of balance and lowest credit limit. The customers of this cluster also have the lowest of minimum payments, payments and tenure which indicates that transactions made by these people are small transactions.

12. Implementation

- Code-

```
install.packages('tidyverse')
install.packages('DT')
install.packages('GGally')
install.packages('RColorBrewer')
install.packages('ggplot2')
install.packages('ggforce')
install.packages('concaveman')
install.packages('factoextra')
install.packages('FactoMineR')
```

```
# Libraries Required
library(tidyverse)
library(DT)
library(GGally)
library(RColorBrewer)
library(ggplot2)
library(ggforce)
library(concaveman)
library(factoextra)
library(FactoMineR)

# Data Input
data <- read.csv("Dataset.csv")

datatable(data, options = list(scrollX = TRUE))



# Checking Data Types
str(data)

# Calculating the percentage of missing values
colSums(is.na(data)/nrow(data))

# Checking Missing Values
colSums(is.na(data))

# Drop NA
data_na <- data %>%
  drop_na(CREDIT_LIMIT, MINIMUM_PAYMENTS)

data_na

# Checking the Number of missing values after dropping the
rows with missing values
colSums(is.na(data_na))
```

```
# Checking new dimensions of the data set
dim(data_na)

# dropping CUST_ID columns
data_clean <- data_na %>%
  select(-CUST_ID)

data_clean

#checking the dimension of cleaned data
dim(data_clean)

# Data Summary
summary(data_clean)

# Correlation of each variable
ggcorr(data_clean, hjust=1, layout.exp = 2, label = T,
       label_size = 4, low = "green", mid = "white",
       high = "red")

# Data Scaling
data_z <- scale(data_clean)
data_z

summary(data_z)


# 1. Elbow Method:

fviz_nbclust(data_clean, FUNcluster = kmeans, method =
"wss",
             k.max = 7, print.summary = TRUE) + labs(subtitle =
"Elbow method")
```

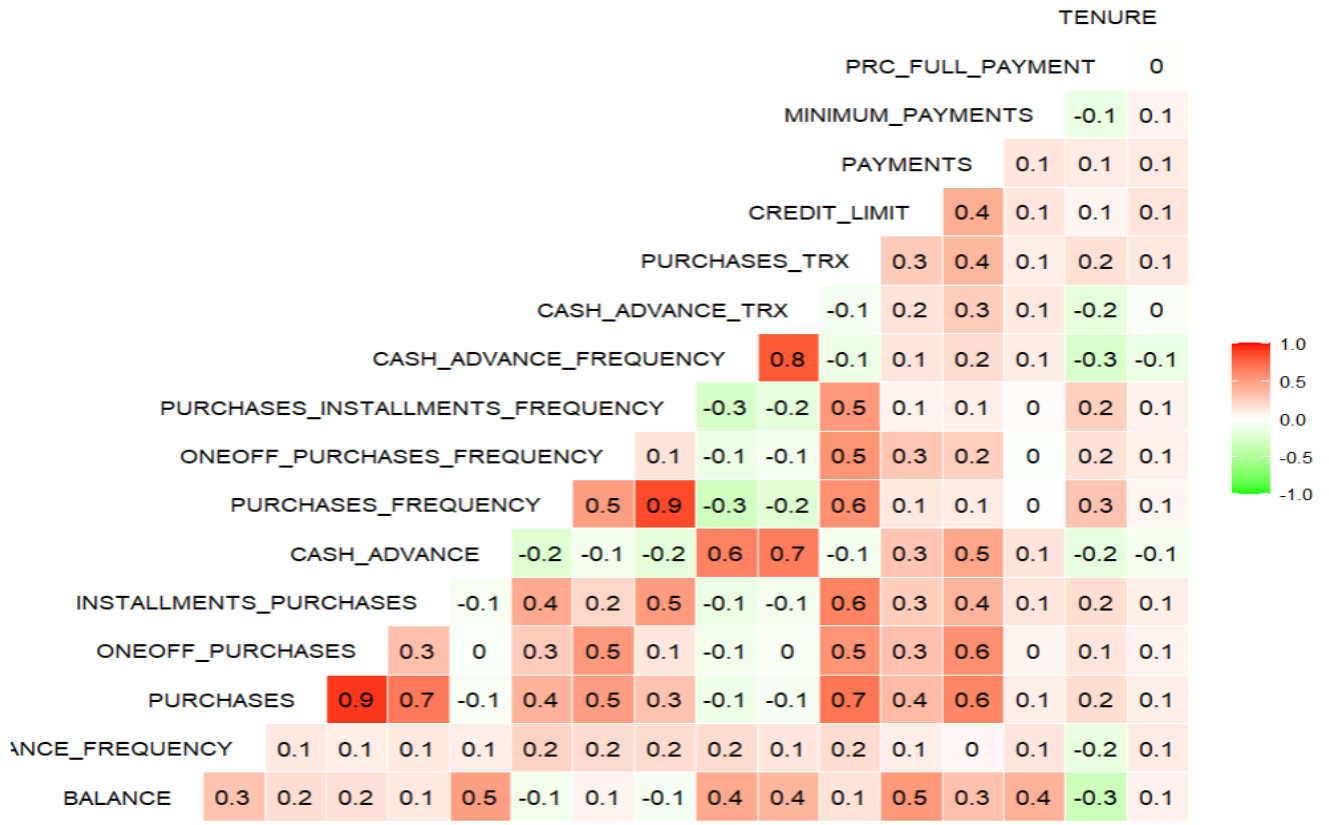
```
# 2. Silhouette Method

fviz_nbclust(data_clean, FUNcluster = kmeans, method =
"silhouette",
             k.max = 10, print.summary = TRUE) + labs(subtitle =
"Silhouette Method")



# Clustering

# K-Means with k = 2

set.seed(123)
data_KM2 <- kmeans(x = data_z, centers = 2)

# Number of observations in each cluster
data_KM2$size

# Location of the center of the cluster/centroid, will be used for
cluster profiling
data_KM2$centers

# Cluster Visualization
fviz_cluster(object = data_KM2, data = data_z, geom =
"point") +
  ggtitle("K-Means Clustering Plot") +
  scale_color_brewer(palette = "Accent") + theme_minimal() +
  theme(legend.position = "bottom")



# K-Means with k = 4

set.seed(123)
data_KM4 <- kmeans(x = data_z, centers = 4)
```

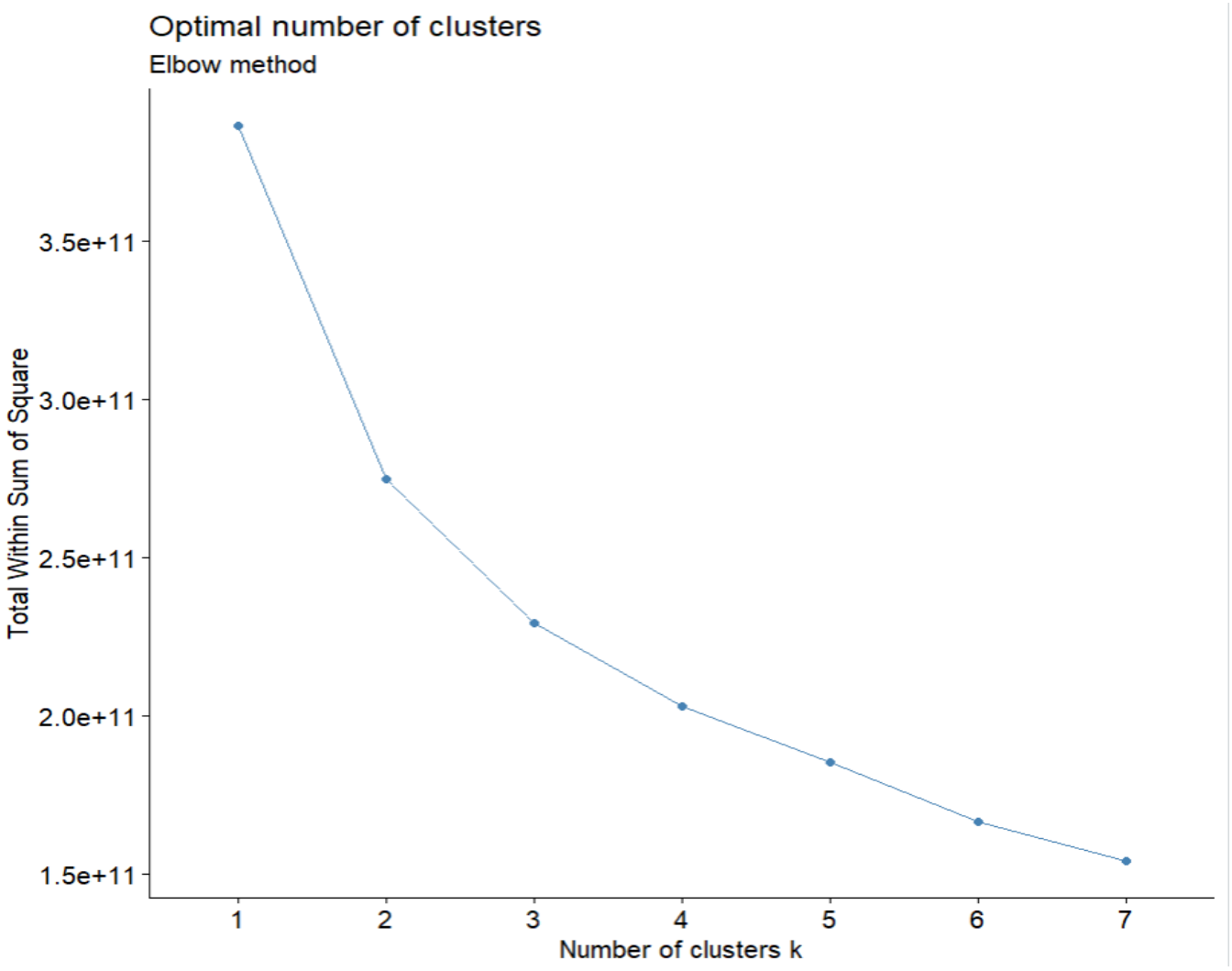
```
# Number of observations in each cluster
data_KM4$size
data_KM4$centers


# Cluster Visualization:
fviz_cluster(object = data_KM4, data = data_z,
             geom = "point") +
  ggtitle("K-Means Clustering Plot") +
  scale_color_brewer(palette = "Accent") +
  theme_minimal() +
  theme(legend.position = "bottom")


# Goodness of Fit
# Goodness of Fit of K-Means with k = 2.

# Within Sum of Squares
data_KM2$withinss

# Total Sum of Squares
data_KM2$betweenss/data_KM2$totss

# Goodness of Fit of K-Means with k = 4.

# Within Sum of Squares
data_KM4$withinss

# Total Sum of Sqaures
data_KM4$betweenss/data_KM4$totss

# Cluster Profiling

# Combining the cluster label into the data set
data_clean$CLUSTER <- as.factor(data_KM4$cluster)
```

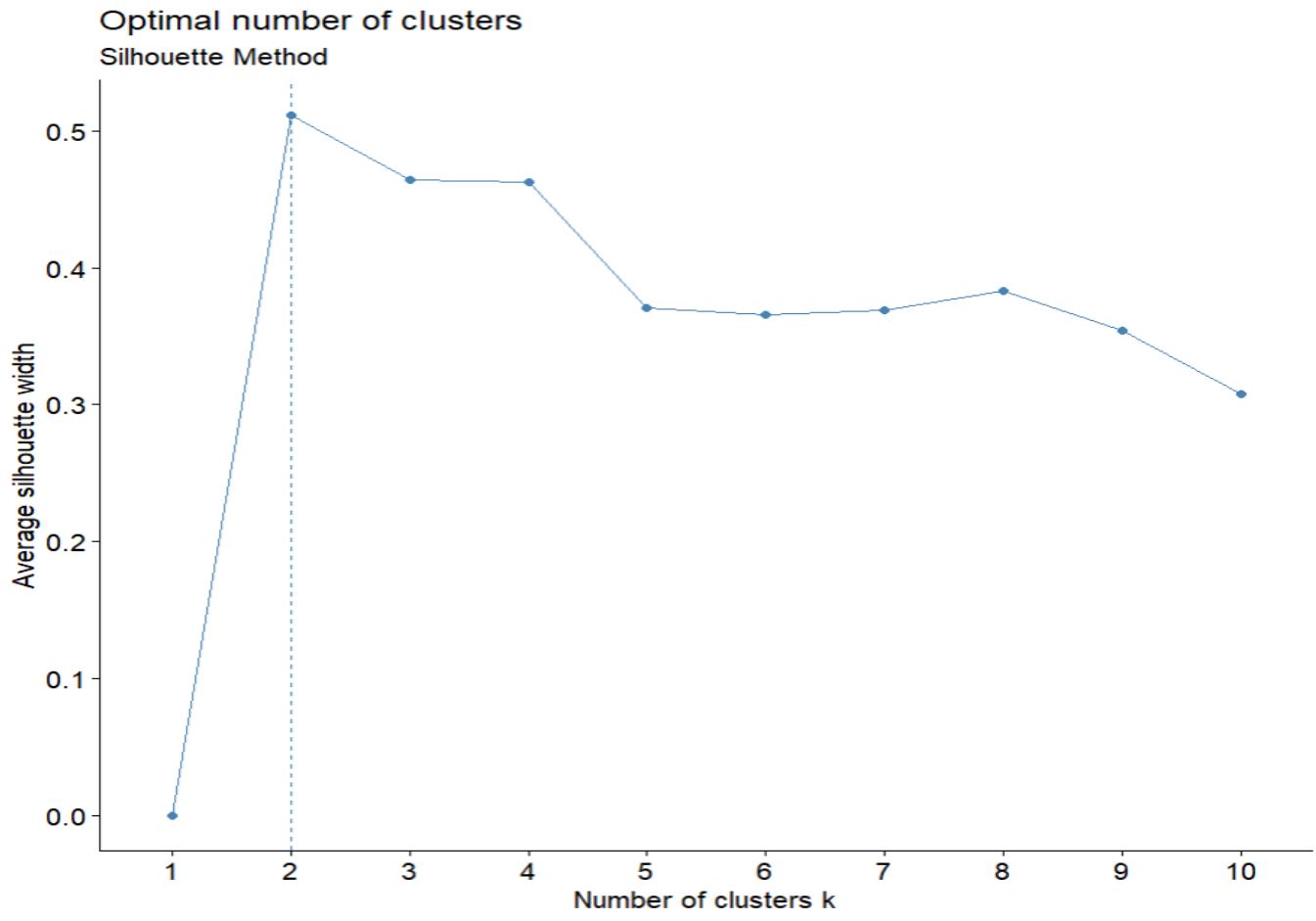
```
# Profiling with aggregation table
data_clean %>%
  group_by(CLUSTER) %>%
  summarise_all(mean)

# Profiling with aggregation table
data_clean %>%
  group_by(CLUSTER) %>%
  summarise_all(mean) %>%
  tidyr::pivot_longer(-CLUSTER) %>%
  group_by(name) %>%
  summarize(cluster_min_val = which.min(value),
            cluster_max_val = which.max(value))

ggplot(data_clean, aes(x = factor(CLUSTER), y =
PURCHASES, fill = CLUSTER, colour = CLUSTER)) +
  geom_bar(stat = "identity", position = "dodge")

ggplot(data_clean, aes(x = factor(CLUSTER), y =
PAYMENTS, fill = CLUSTER, colour = CLUSTER)) +
  geom_bar(stat = "identity", position = "dodge")


# Combining the 'CUST_ID' column into the data set
data_ID <- data_clean %>%
  mutate(CUST_ID = data_na$CUST_ID)

datatable(data_ID, options = list(scrollx = TRUE))

# Principal Component Analysis(PCA)

# PCA using FactoMineR
data_pca <- PCA(X = data_clean, quali.sup = 18,
                scale.unit = T, ncp = 17, graph = F)
data_pca$eig
```

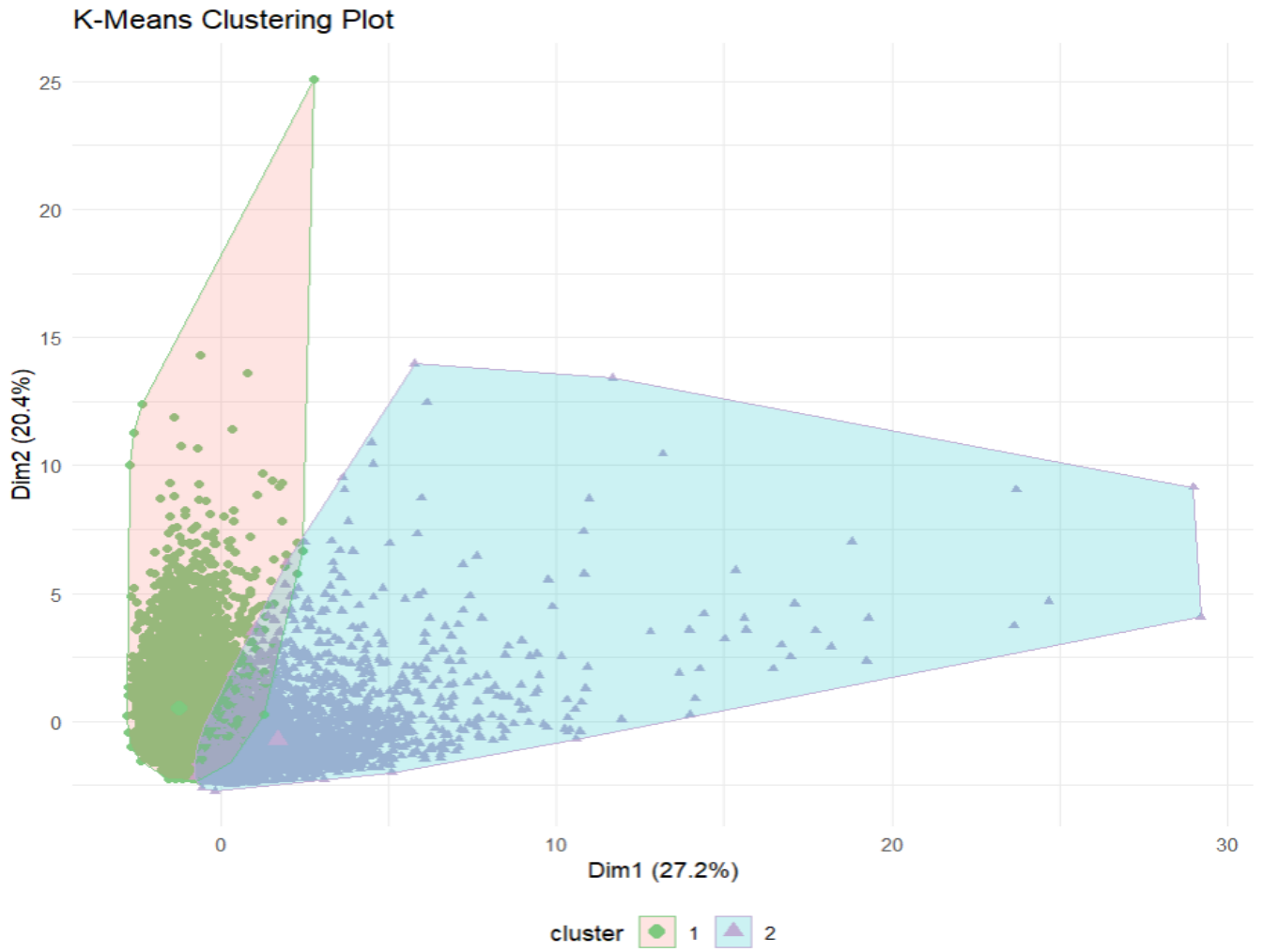
```
# Variance explained by each dimensions
fviz_eig(data_pca, ncp = 17, addlabels = T,
         main = "Variance explained by each dimensions")


# Variable Contribution of PC1
fviz_contrib(X = data_pca, choice = "var", axes = 1)

# Variable contribution untuk PC2
fviz_contrib(X = data_pca, choice = "var", axes = 2)


# PCA Visualization

# Individual Factor Map
plot.PCA(x = data_pca, choix = "ind", invisible = "quali",
         select = "contrib 8", habillage = "CLUSTER") +
  scale_color_brewer(palette = "Accent") +
  theme(legend.position = "bottom")

# Variables Factor Map
fviz_pca_var(data_pca, select.var = list(contrib = 17), col.var =
"contrib",
             gradient.cols = c("red", "white", "blue"), repel =
TRUE)

# Cluster Visualization with PCA

# Visualisasing PCA + K-Means Clustering
fviz_pca_biplot(data_pca, habillage = 18, addEllipses = T,
                geom.ind = "point") +
  theme_minimal() +
  theme(legend.position = "bottom") +
  scale_color_brewer(palette = "Accent")
```

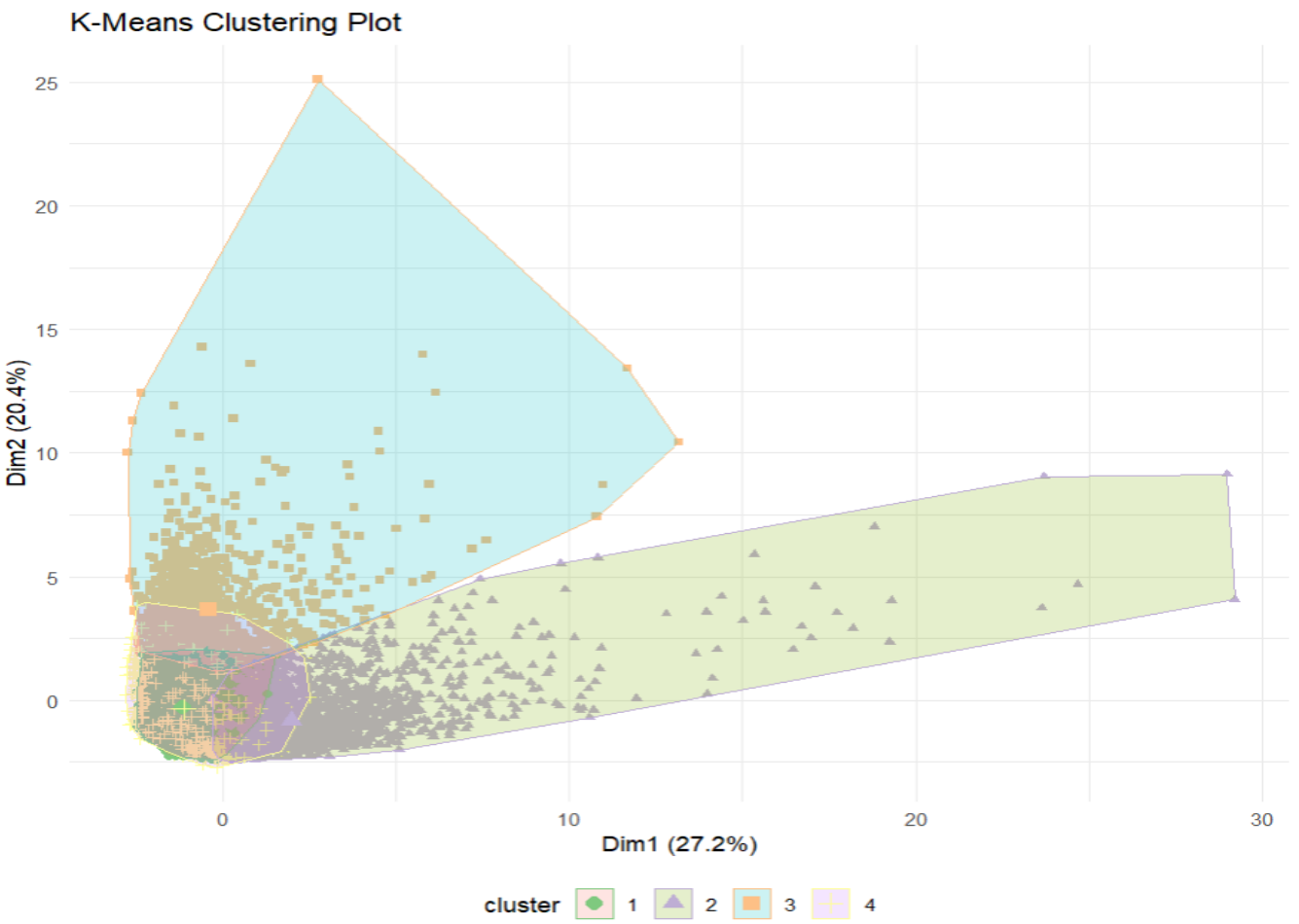
- Output-

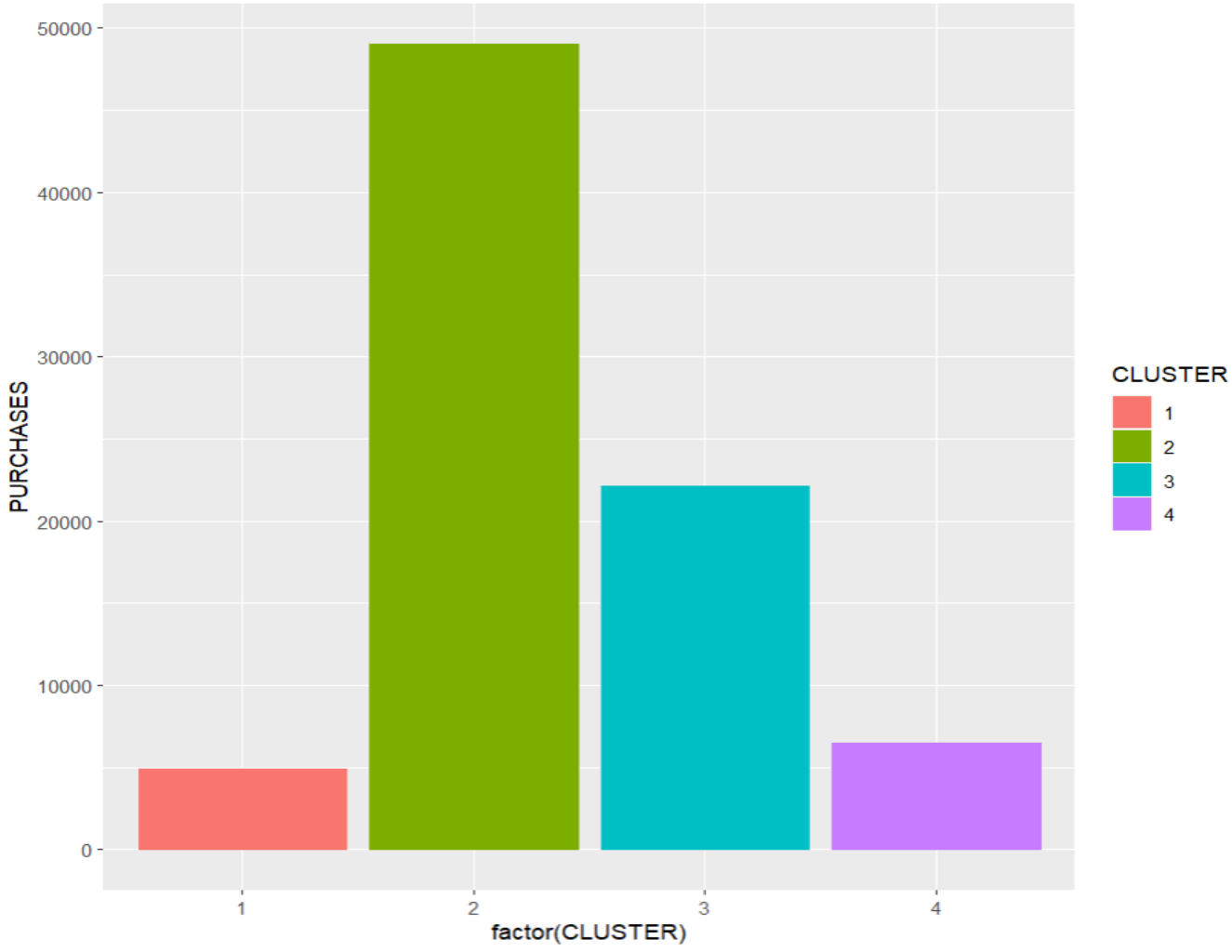

TENURE
PRC_FULL_PAYMENT 0
MINIMUM_PAYMENTS -0.1 0.1
PAYMENTS 0.1 0.1 0.1
CREDIT_LIMIT 0.4 0.1 0.1 0.1
PURCHASES_TRX 0.3 0.4 0.1 0.2 0.1
CASH_ADVANCE_TRX -0.1 0.2 0.3 0.1 -0.2 0
CASH_ADVANCE_FREQUENCY 0.8 -0.1 0.1 0.2 0.1 -0.3 -0.1
PURCHASES_INSTALLMENTS_FREQUENCY -0.3 -0.2 0.5 0.1 0.1 0 0.2 0.1
ONEOFF_PURCHASES_FREQUENCY 0.1 -0.1 -0.1 0.5 0.3 0.2 0 0.2 0.1
PURCHASES_FREQUENCY 0.5 0.9 -0.3 -0.2 0.6 0.1 0.1 0 0.3 0.1
CASH_ADVANCE -0.2 -0.1 -0.2 0.6 0.7 -0.1 0.3 0.5 0.1 -0.2 -0.1
INSTALLMENTS_PURCHASES -0.1 0.4 0.2 0.5 -0.1 -0.1 0.6 0.3 0.4 0.1 0.2 0.1
ONEOFF_PURCHASES 0.3 0 0.3 0.5 0.1 -0.1 0 0.5 0.3 0.6 0 0.1 0.1
PURCHASES 0.9 0.7 -0.1 0.4 0.5 0.3 -0.1 -0.1 0.7 0.4 0.6 0.1 0.2 0.1
ANCE_FREQUENCY 0.1 0.1 0.1 0.1 0.2 0.2 0.2 0.2 0.1 0.2 0.1 0 0.1 -0.2 0.1
BALANCE 0.3 0.2 0.2 0.1 0.5 -0.1 0.1 -0.1 0.4 0.4 0.1 0.5 0.3 0.4 -0.3 0.1
1.0
0.5
0.0
-0.5
-1.0

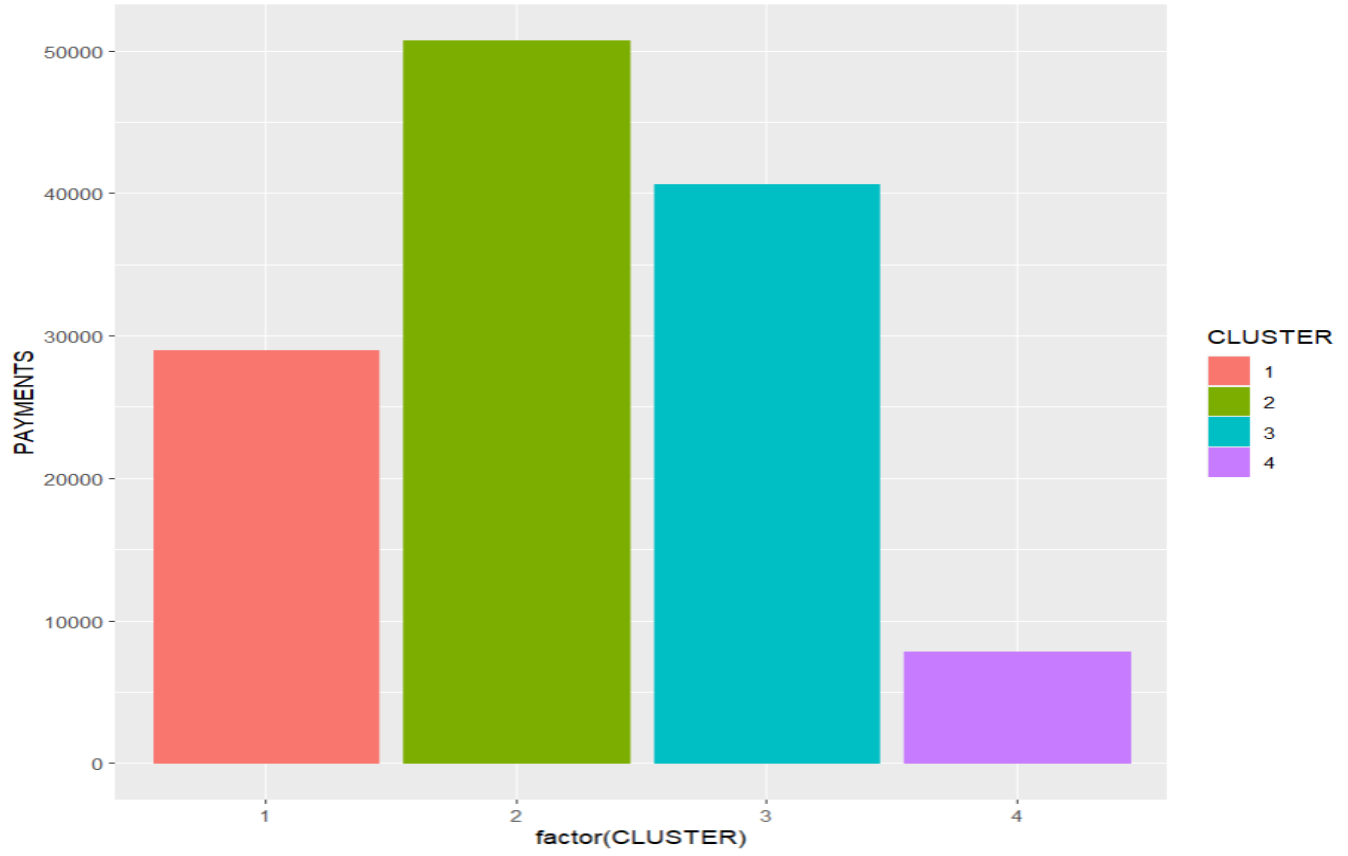

Optimal number of clusters
Elbow method
Total Within Sum of Square
3.5e+11
3.0e+11
2.5e+11
2.0e+11
1.5e+11
1
2
3
4
5
6
7
Number of clusters k

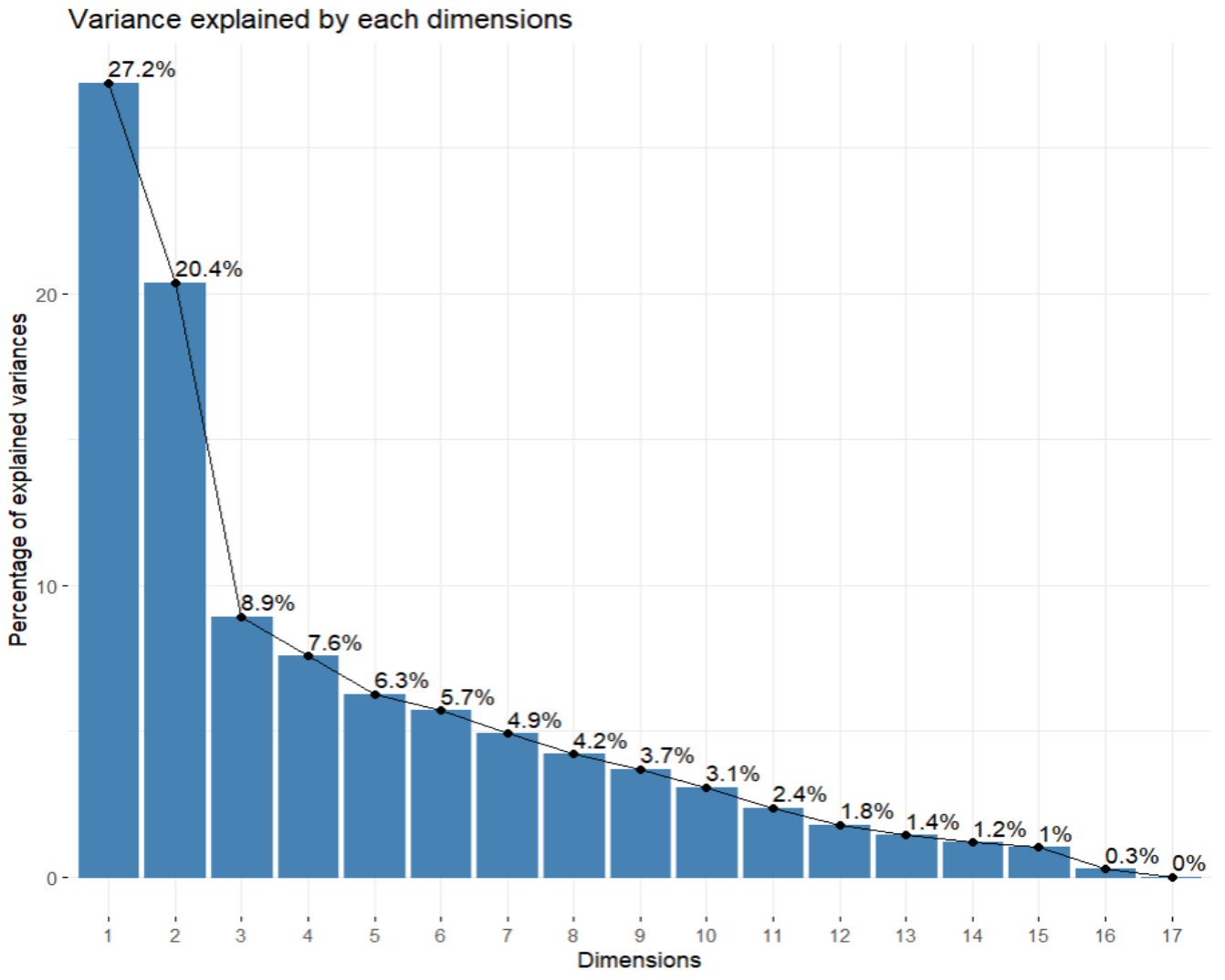

Optimal number of clusters
Silhouette Method
Average silhouette width
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
8
9
10
Number of clusters k

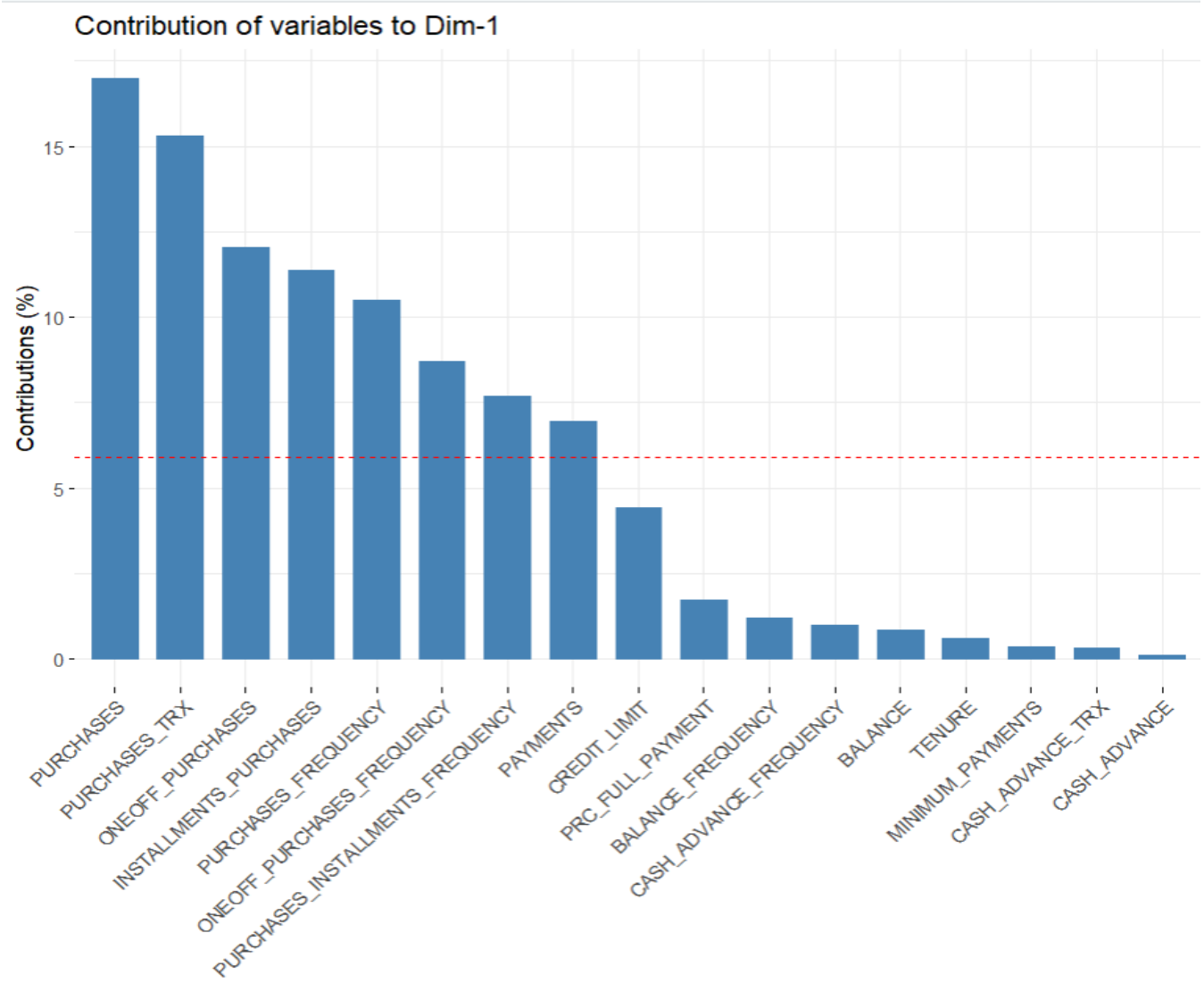

K-Means Clustering Plot
Dim2 (20.4%)
25
20
15
10
5
0
0
10
20
30
Dim1 (27.2%)
cluster
1
2

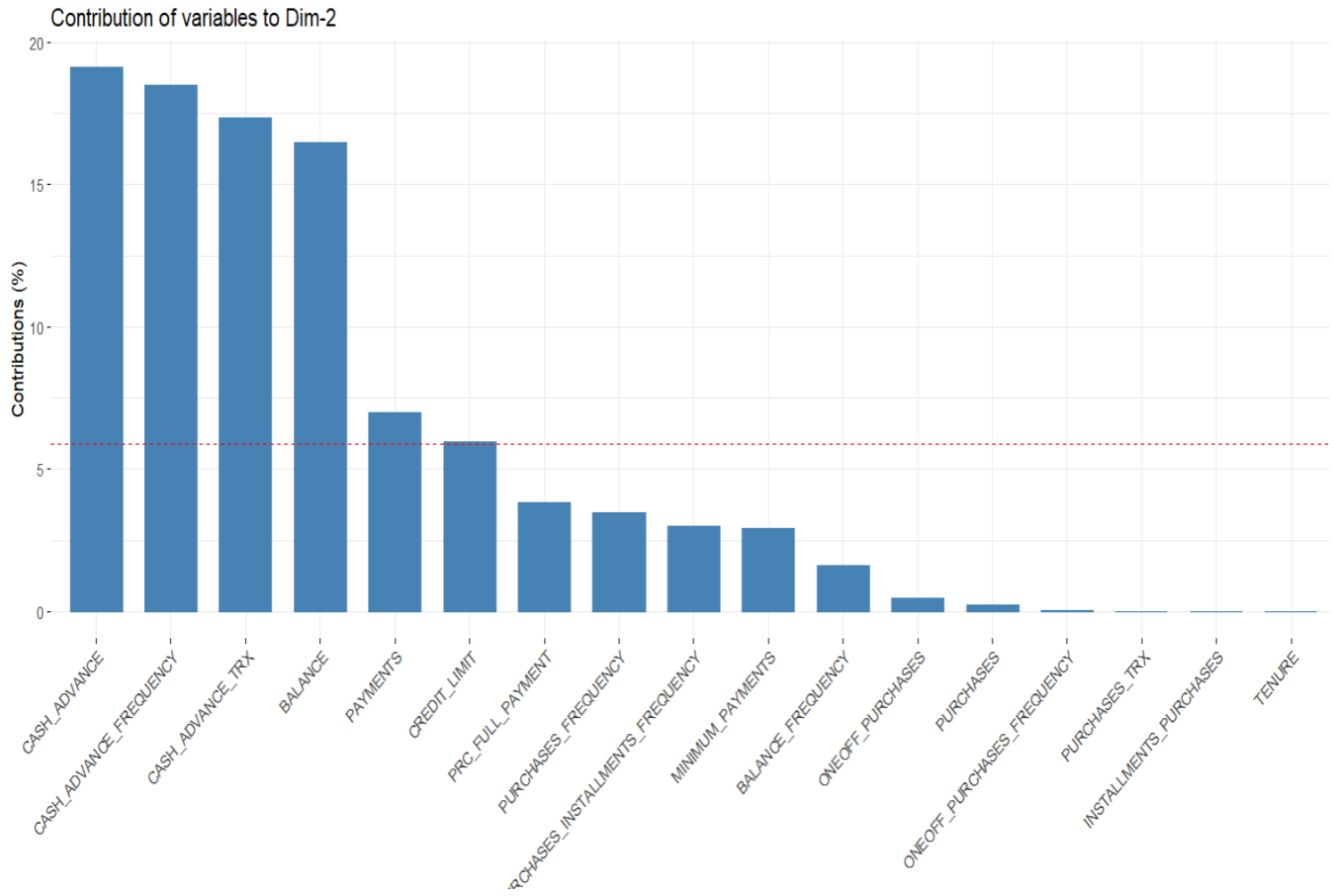

K-Means Clustering Plot
Dim2 (20.4%)
Dim1 (27.2%)
25
20
15
10
5
0
0
10
20
30
cluster
1
2
3
4

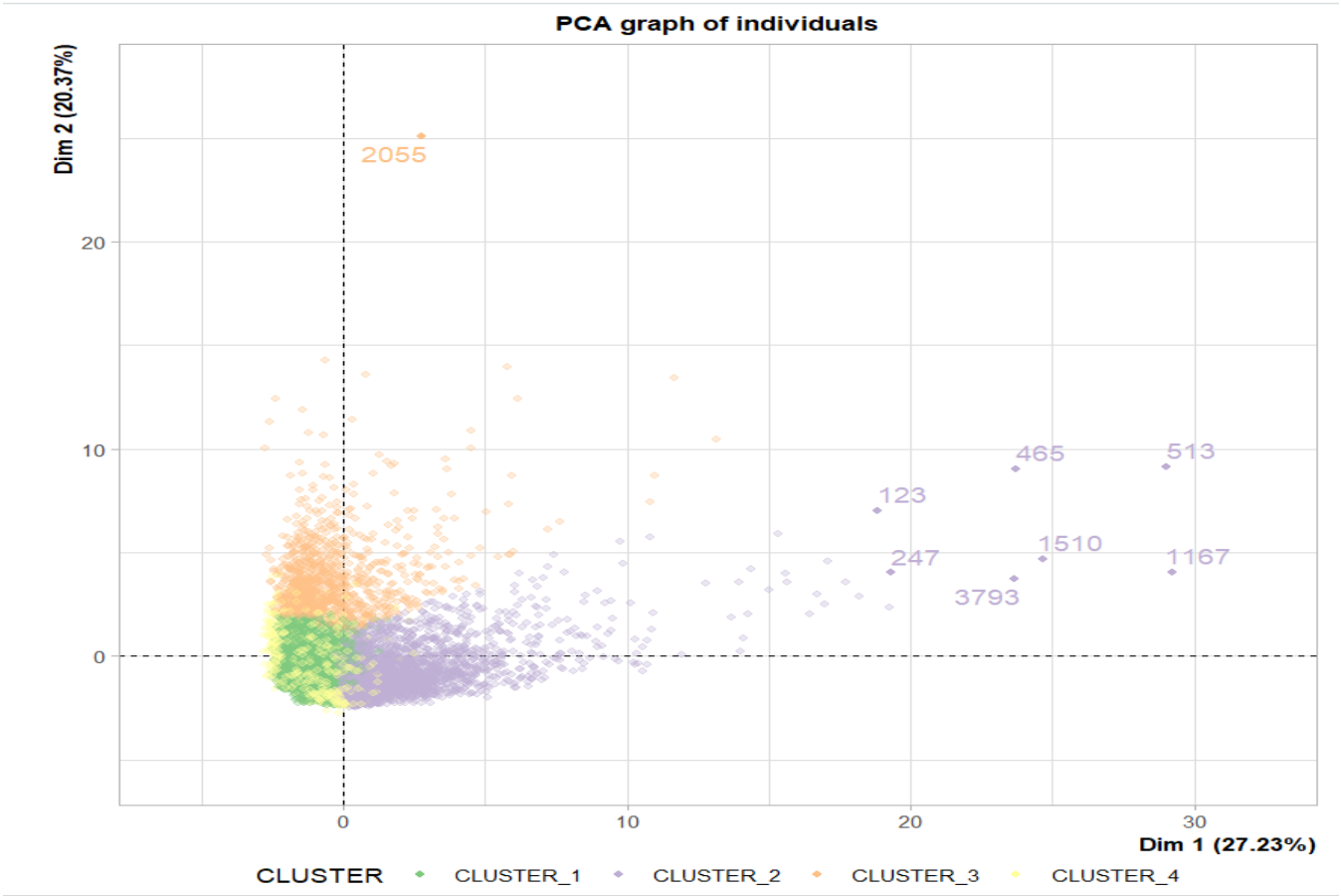

PURCHASES
factor(CLUSTER)
50000
40000
30000
20000
10000
0
1
2
3
4
CLUSTER
1
2
3
4

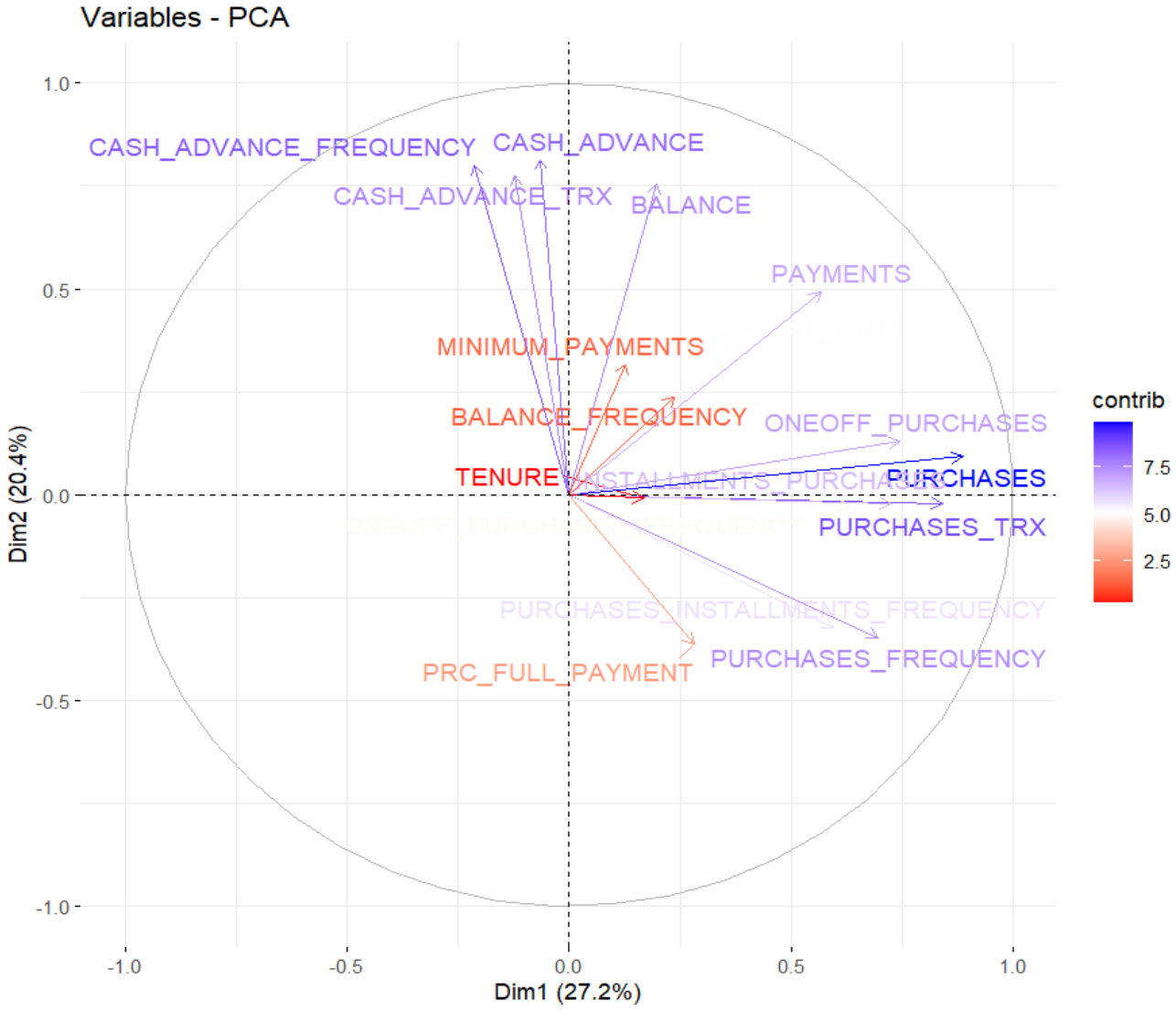

50000
40000
30000
20000
10000
0
PAYMENTS
1
2
3
4
factor(CLUSTER)
CLUSTER
1
2
3
4

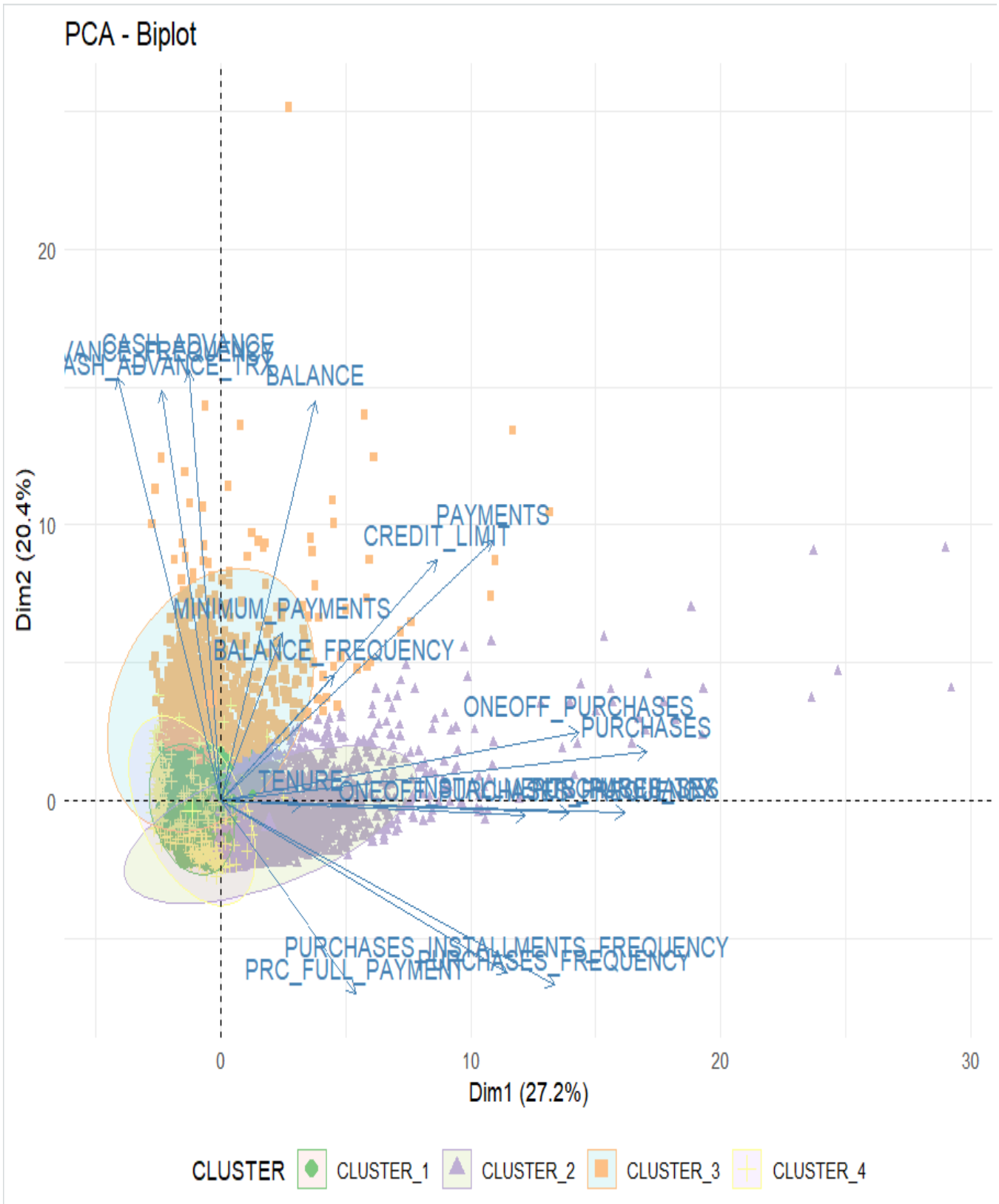

Variance explained by each dimensions
Percentage of explained variances
20
10
0
27.2%
20.4%
8.9%
7.6%
6.3%
5.7%
4.9%
4.2%
3.7%
3.1%
2.4%
1.8%
1.4%
1.2%
1%
0.3%
0%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Dimensions


Contribution of variables to Dim-1
Contributions (%)
15
10
5
0
PURCHASES
PURCHASES_TRX
ONEOFF_PURCHASES
INSTALLMENTS_PURCHASES
PURCHASES_FREQUENCY
ONEOFF_PURCHASES_FREQUENCY
PURCHASES_INSTALLMENTS_FREQUENCY
PAYMENTS
CREDIT_LIMIT
PRC_FULL_PAYMENT
BALANCE_FREQUENCY
CASH_ADVANCE_FREQUENCY
BALANCE
TENURE
MINIMUM_PAYMENTS
CASH_ADVANCE_TRX
CASH_ADVANCE


Contribution of variables to Dim-2
Contributions (%)
20
15
10
5
0
CASH_ADVANCE
CASH_ADVANCE_FREQUENCY
CASH_ADVANCE_TRX
BALANCE
PAYMENTS
CREDIT_LIMIT
PRC_FULL_PAYMENT
PURCHASES_FREQUENCY
PURCHASES_INSTALLMENTS_FREQUENCY
MINIMUM_PAYMENTS
BALANCE_FREQUENCY
ONEOFF_PURCHASES
PURCHASES
ONEOFF_PURCHASES_FREQUENCY
PURCHASES_TRX
INSTALLMENTS_PURCHASES
TENURE


PCA graph of individuals
Dim 2 (20.37%)
Dim 1 (27.23%)
2055
465
513
123
1510
247
1167
3793
CLUSTER
CLUSTER_1
CLUSTER_2
CLUSTER_3
CLUSTER_4


Variables - PCA
Dim2 (20.4%)
Dim1 (27.2%)
CASH_ADVANCE_FREQUENCY
CASH_ADVANCE
CASH_ADVANCE_TRX
BALANCE
PAYMENTS
MINIMUM_PAYMENTS
BALANCE_FREQUENCY
ONEOFF_PURCHASES
TENURE
PURCHASES
PURCHASES_TRX
PURCHASES_FREQUENCY
PRC_FULL_PAYMENT
contrib


PCA - Biplot
BALANCE
PAYMENTS
CREDIT_LIMIT
MINIMUM_PAYMENTS
BALANCE_FREQUENCY
ONEOFF_PURCHASES
PURCHASES
TENURE
PURCHASES_INSTALLMENTS_FREQUENCY
PRC_FULL_PAYMENT
Dim2 (20.4%)
Dim1 (27.2%)
20
10
0
0
10
20
30
CLUSTER
CLUSTER_1
CLUSTER_2
CLUSTER_3
CLUSTER_4

13. Conclusion

Based on the results obtained and analysis that we have done, we can say that:

- Cluster 1 & 4 have the lowest amount of purchases compared with the other clusters, hence if there are offers such as reward programs, discounts using credit card, they could be the best target.
- Cluster 2 & 3 also has the highest number of payments compared with the other clusters, indicating how aware they are of their credits. Hence, if there are offers such as loyalty points, they could be the best target.
- Cluster 4 relatively has the lowest number of purchases and payments compared with the other clusters. They also can be offered to zero interest program to increase theirs purchase and payments.

14. Scope for future work

- Although K-Means is the best approach for the clustering purpose but as a part of the future work, we can also try the segmentation work by trying out different types of clustering algorithms like, Agglomerative Hierarchical Clustering, Density-based Clustering, Fuzzy Clustering, etc.
- Also, in this project we have omitted the observations that had missing values. We can also do the data imputation and try again with the same algorithm or a different one so that if we may get some different results.

15. References

- https://www.researchgate.net/publication/313737530_Review_on_Customer_Segmentation_Technique_on_Ecommerce
- https://ieeexplore.ieee.org/document/5974496
- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4031925
- https://rdrr.io/
- https://developers.google.com/machinelearning/clustering/clustering-algorithms
- https://www.qualtrics.com/au/experiencemanagement/brand/customersegmentation/?rid=ip&prevsite=en&newsite=au&geo=IN&geomatch=au