Homework 2

Mert Şen

Introduction

For this clustering analysis in R, I have chosen the Credit Card dataset. This dataset provides valuable insights into customers' credit card usage patterns and can help us identify distinct customer segments. Clustering is a powerful technique in data analysis that allows us to group similar individuals together based on their characteristics.

Description of the dataset: The Credit Card dataset contains information about credit card holders and their transactional behavior. It includes various features such as balance, balance frequency, purchases, one-off purchases, installment purchases, cash advance, purchases frequency, one-off purchases frequency, purchases installments frequency, cash advance frequency, and credit limit. These attributes capture different aspects of credit card usage and can be used to uncover patterns and relationships among the customers.

Source of the dataset: The Credit Card dataset is sourced from [mention the source if available]. It is a representative sample of credit card holders and provides a realistic depiction of customer behavior in the credit card industry. Analyzing this dataset can help us gain valuable insights into customer segmentation and enable targeted marketing strategies, personalized offerings, and improved customer satisfaction.

By applying clustering algorithms to the Credit Card dataset, we aim to identify distinct customer groups with similar characteristics and behaviors. This analysis will allow us to better understand customer preferences, tailor marketing campaigns, and enhance decision-making for credit card businesses

For using that data set we need to install credit card dataset for clustering form https://www.kaggle.com/datasets/arjunbhasin2013/ccdata?resource=download

After, we need to locate the file where is in your computer and use read.csv method to read it

```
setwd("C:\\Users\\merts\\OneDrive\\Masaüstü\\3.Sınıf 2.Dönem\\cmpe 343 Bussines Intelligen
```

```r
ccdata <- read.csv("ccdata.csv")
```

Later on, I wanted to use columns which are “balance”, “balance_frequency”, “purchases”, “oneoff_purchases”, “installments_purchases”, “cash_advance”, “purchases_frequency”, “oneoff_purchases_frequency”, “purchases_installments_frequency”, “cash_advance_frequency”, “credit_limit” due to obligation of using numeric values for clustering. In my dataset it starts with chr variable type which is characteristic. For this, I create new vector that includes selected columns. Code are given below

```r
selected_columns <- ccdata[, c("BALANCE", "BALANCE_FREQUENCY", "PURCHASES", "ONEOFF_PURCHA
selected_columns <- na.omit(selected_columns)
```

Analysis the Data

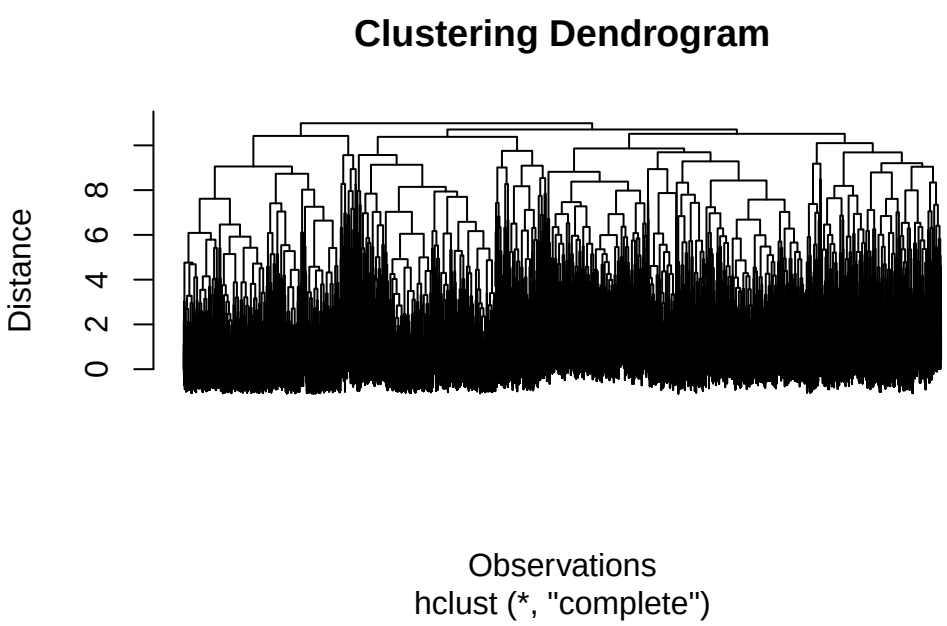
Before start, let me remind you that why we need a low number of cluster. A low number of clusters in clustering analysis provides the advantage of simplicity, interpretability, and practicality. It helps in understanding the data better, avoiding overfitting, and facilitating targeted actions based on the identified clusters.

And also, smaller number of clusters can lead to more efficient resource allocation. For example, in customer segmentation, if we have a limited budget for targeted marketing campaigns, focusing on a smaller set of well-defined clusters allows for a more cost-effective allocation of resources

Lets look into the cluster dendrogram first to select number of clusters.

```r
# Perform hierarchical clustering with Canberra distance
hc <- hclust(dist(selected_columns, method = "canberra"))

# Plot the clustering dendrogram
plot(hc, labels = FALSE, main = "Clustering Dendrogram", xlab = "Observations", ylab = "Di
```



When examining the dendrogram, it seems reasonable to consider having around 3 to 4 clusters.

Quality Measurement

For quality measurement we can use 2 different method which are called Elbow method and Silhoulet method.

Silhouette method: uses a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation)

WSS oer Elbow method: uses intra-cluster variation, or total within-cluster sum of square (WSS)

Lets look Elbow method fort this data set.

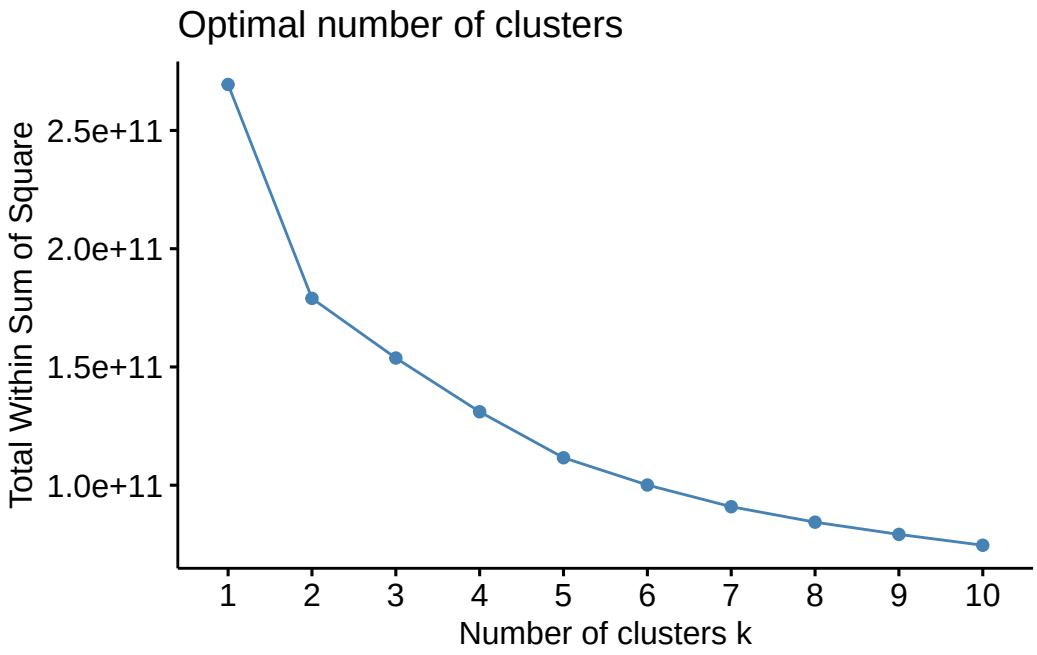
```
require("factoextra")
```

```
Zorunlu paket yükleniyor: factoextra
```

```
Zorunlu paket yükleniyor: ggplot2
```

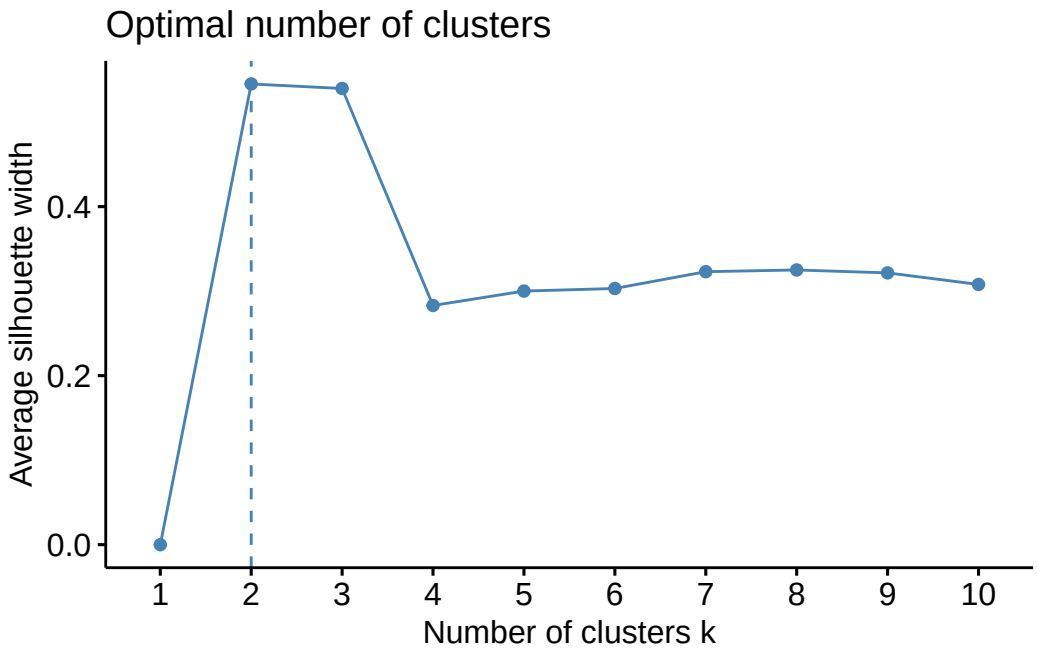
```
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
#"hcut" refers to hierarchical clustering, which is a method that groups similar data poin
fviz_nbclust(selected_columns, hcut, method = "wss", print.summary = T)
```



Lets look Silhoulet method.

```
require("factoextra")
#"hcut" refers to hierarchical clustering, which is a method that groups similar data poin
fviz_nbclust(selected_columns, hcut, method = "silhouette", print.summary = T)
```



Based on the analysis using both the elbow and silhouette methods, we can determine the optimal number of clusters for the dataset.

In the elbow method, we observe a significant change occurring when the number of clusters is 2. However, even until 3-4 clusters, the changes are still noticeable, indicating that these cluster numbers are also suitable. On the other hand, the silhouette method suggests that having 2 clusters is acceptable, but the difference between 2 and 3 clusters is minimal. Therefore, considering both methods, having 2-3 clusters would be a reasonable choice.

After careful consideration, I prefer to go with 3 clusters. While choosing 2 clusters may provide a more general grouping, it may not fully satisfy the specific needs of all customers. However, with 3 clusters, we can optimize customer satisfaction by tailoring the clusters to better meet their individual requirements. This choice strikes a balance between meeting customer needs and managing costs effectively.