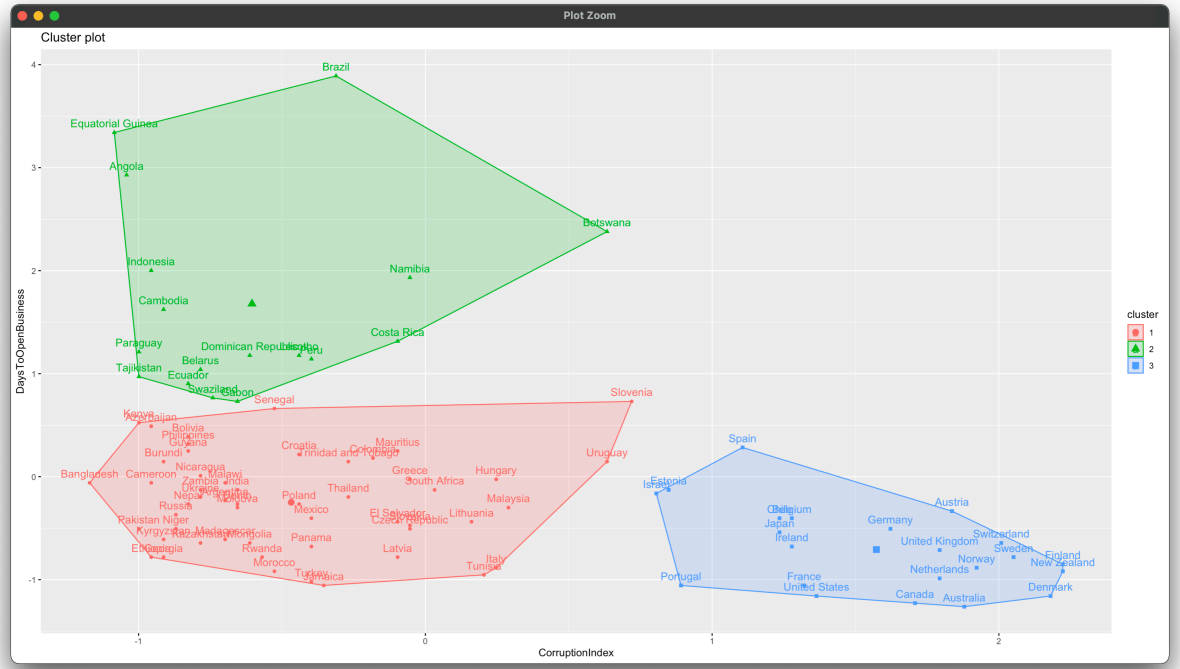
1. From R:
    1. All code from your R script (Code should be presented single-spaced in a fixed-width font. Adjust the font size so that no lines of code extend to the next line in the document)

```
# YUEH-TING WU
# MIS 545 Section 02
# Lab11WuY.R
# In this R programming, import a csv file and generate clusters
to discover
# patterns.

# Install the tidyverse and factoextra packages
# install.packages("tidyverse")
# install.packages("factoextra")

# Load the tidyverse, stats, factoextra, cluster, and gridExtra
libraries
library(tidyverse)
library(stats)
library(factoextra)
library(cluster)
library(gridExtra)

# Set the working directory to Lab11 folder
setwd("~/MIS 545/Lab11")

# Read CountryData.csv into an object called countries
countries <- read_csv(file = "CountryData.csv",
                      col_types = "cnnnnini",
                      col_names = TRUE)

# Display the countries tibble on the console
print(countries)

# Display the structure of the countries tibble
str(countries)

# Display a summary of the countries tibble
summary(countries)

# Convert the column containing the country name to the row
tittle of the tibble
# This is a requirement for later visualizing the clusters
countries <- countries %>% column_to_rownames(var = "Country")
```

```
# Remove countries from the tibble with missiing data in any
feature
countries <- countries %>% drop_na()

# View the summary of the countries tibble again to ensure there
are no NA
# values
summary(countries)

# We are going to cluster the countries based on their
corruption index and
# the number of days it takes to open a business.
# Create a new tibble called countriesScaled containing only
these two features
# and scale them so they have equal impact on the clustering
calculations.
countriesScaled <- countries %>%
  select(CorruptionIndex, DaysToOpenBusiness) %>% scale()

# Set the random seed to 679
set.seed(679)

# Generate the k-means clusters in an object called
countries4Clusters using
# 4 clusters and a value of 25 for nstart
countries4Clusters <- kmeans(x = countriesScaled,
                             centers = 4,
                             nstart = 25)

# Display cluster sizes on the console
countries4Clusters$size

# Display cluster centers (z-scores) on the console
countries4Clusters$centers

# Visualize the clusters
fviz_cluster(object = countries4Clusters,
             data = countriesScaled,
             repel = FALSE)

# Optimize the value for k
# Elbow method
fviz_nbclust(x = countriesScaled,
             FUNcluster = kmeans,
             method = "wss")
```

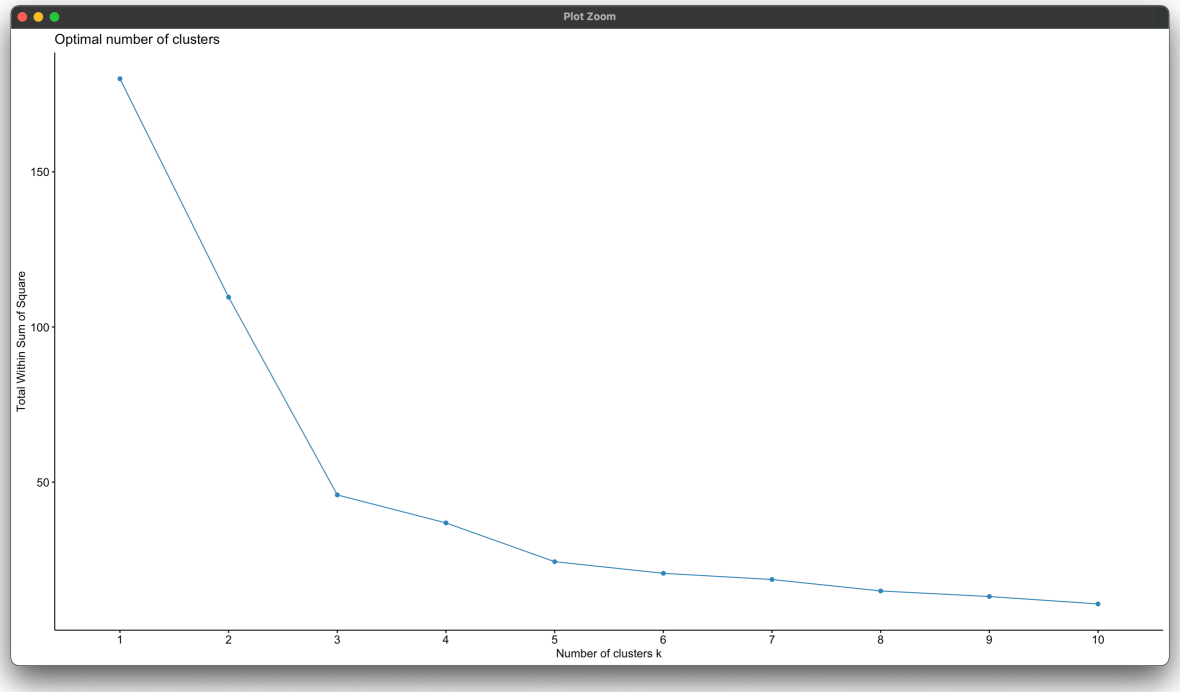
```
# Average silhouette method
fviz_nbclust(x = countriesScaled,
             FUNcluster = kmeans,
             method = "silhouette")

# Gap method
fviz_nbclust(x = countriesScaled,
             FUNcluster = kmeans,
             method = "gap")

# Regenerate the cluster analysis using the optimal number of
clusters
countries3Clusters <- kmeans(x = countriesScaled,
                             centers = 3,
                             nstart = 25)

# Display cluster sizes on the console
countries3Clusters$size

# Display cluster centers (z-scores) on the console
countries3Clusters$centers

# Visualize the clusters
fviz_cluster(object = countries3Clusters,
             data = countriesScaled,
             repel = FALSE)

# Determine similarities and differences among the clusters
using the remaining
# features in the dataset (GiniCoefficient, GDPPerCapita,
EduPercGovSpend,
# EduPercGDP, and CompulsoryEducationYears
countries  %>%
  mutate(cluster = countries3Clusters$cluster) %>%
  select(cluster, GiniCoefficient, GDPPerCapita,
EduPercGovSpend, EduPercGDP,
         CompulsoryEducationYears) %>%
  group_by(cluster) %>%
  summarise_all("mean")
```

2. Both generated cluster plot visualizations

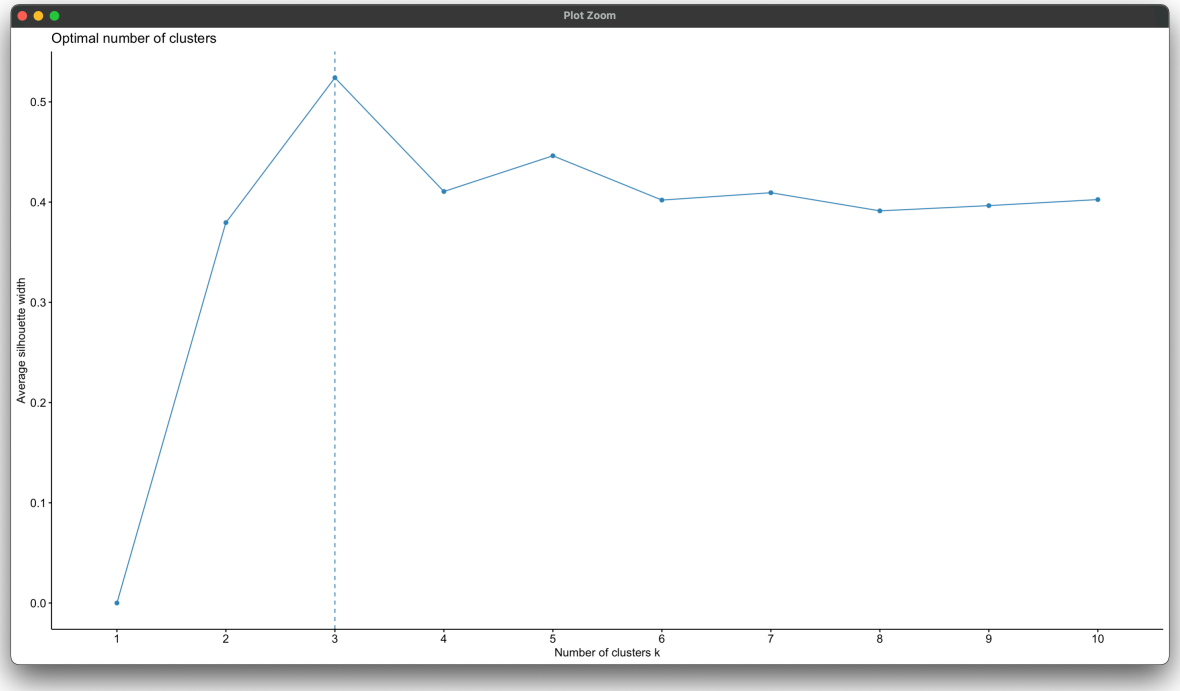


3. All 3 z-optimization plots (elbow method, average silhouette method, and gap statistic method)

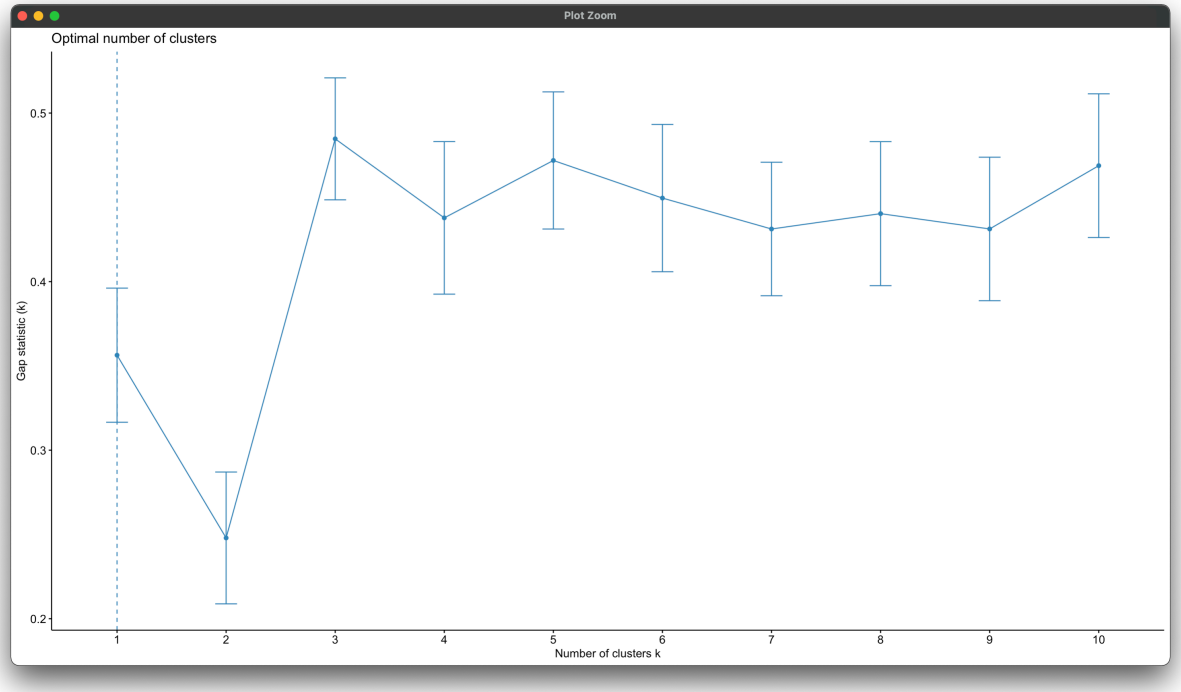
Elbow Method



Average Silhouette Method



Gap Method



2. Answer the following question in a few sentences: For each cluster, how would you describe it given your analysis?

Countries in cluster 1 have a lower corruption index and lower average number of days required to open a new business.
Countries in cluster 2 have a lower corruption index and higher average number of days required to open a new business.
Countries in cluster 3 have high corruption index and lower the average number of days required to open a new business.

3. Answer the following question in a few sentences: Based on your analysis, what is the relationship between education and corruption?

I assume that there is a direct relationship between education and corruption. In cluster 3, it shows that countries that have higher corruption index provide a more longer compulsory education. And other two clusters which have a lower corruption index provide the shorter compulsory education compare to cluster 3.