R code:

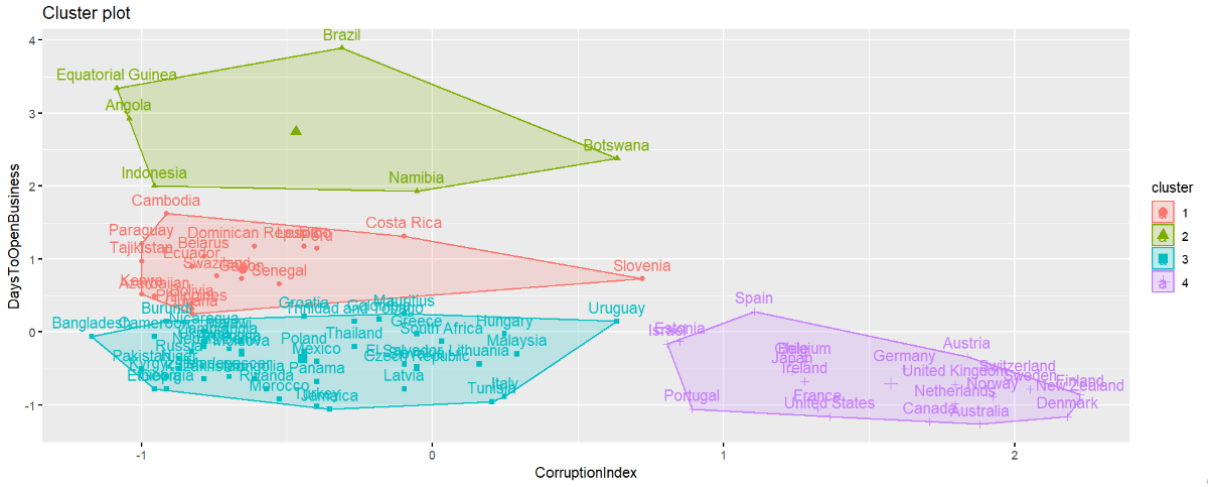
```
# Prasann Patil, Vihari Reddy Tummuru
# MIS545 Section 1
# Lab11Group8PatilTummuru.R
# Import CountryData.csv and generate clusters to discover patterns
# in the dataset

# installing the tidyverse package and factoextra package
# install.packages("tidyverse")
# install.packages("factoextra")

# Loading the tidyverse, stats, factorextra, cluster, gridextra libraries
library(tidyverse)
library(stats)
library(factoextra)
library(cluster)
library(gridExtra)

# setting working directory to Lab10 folder
setwd("C:/Users/91740/OneDrive/Desktop/Lab11")
getwd()

# Reading csv into tibble countries
countries <- read_csv(file = "CountryData.csv",
                      col_types = "cnnnnini", col_names =  TRUE)

# Displaying tibble on the console
print(countries)

# Displaying structure of tibble on the console
str(countries)

# Displaying summary on the console
summary(countries)
```

```
# Converting the column containing the country name to the
# row title of the tibble
countries  <- countries %>% column_to_rownames(var = "Country")

# Removing countries from the tibble with missing data in any feature
countries <- countries %>% drop_na()

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

# Scaling two features in the tibble so they have equal impact
# on the clustering calculations
countriesScaled <- countries %>%
  select(CorruptionIndex, DaysToOpenBusiness) %>% scale()

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

# Generate the k-means clusters
countries4Clusters <- kmeans(x = countriesScaled,
                             centers = 4,
                             nstart = 25)

# Displaying cluster sizes on the console
countries4Clusters$size

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

# Visualizing the clusters
fviz_cluster(object = countries4Clusters,
             data = countriesScaled,
```

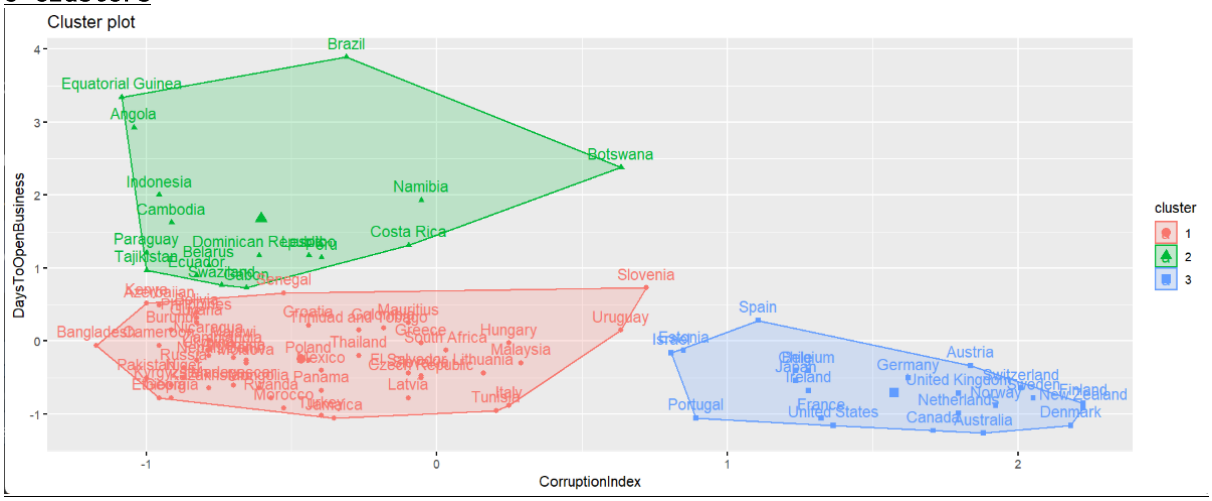
```
             repel = FALSE)

# Optimizing the value of k
# Elbow method
fviz_nbclust(x = countriesScaled,
             FUNcluster = kmeans,
             method = "wss")

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

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

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

# Displaying cluster sizes on the console
countries3Clusters$size

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

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

Both generated cluster plot visualizations

4 clusters

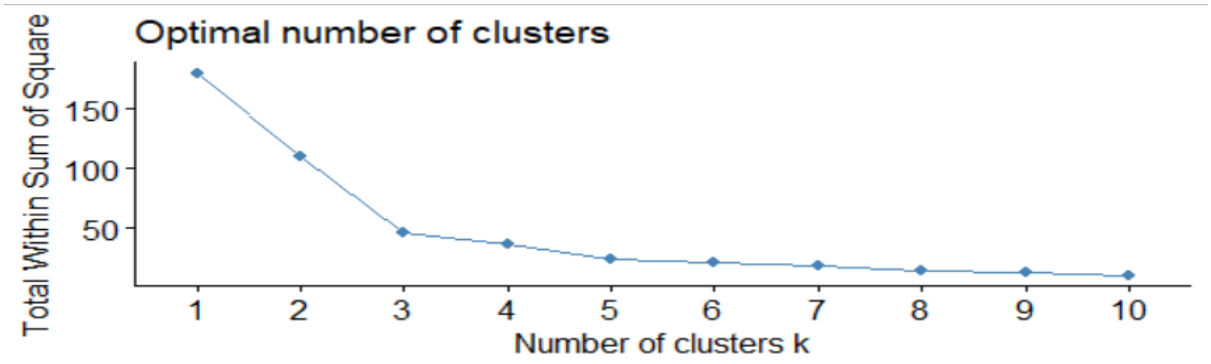


3 clusters

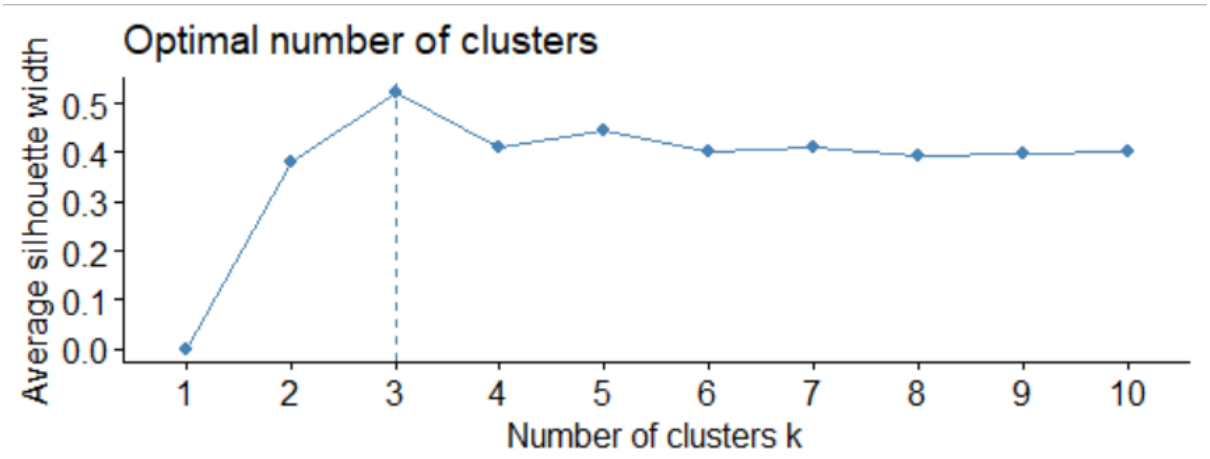


All 3 z-optimization plots (elbow method, average silhouette method, and gap statistic method)

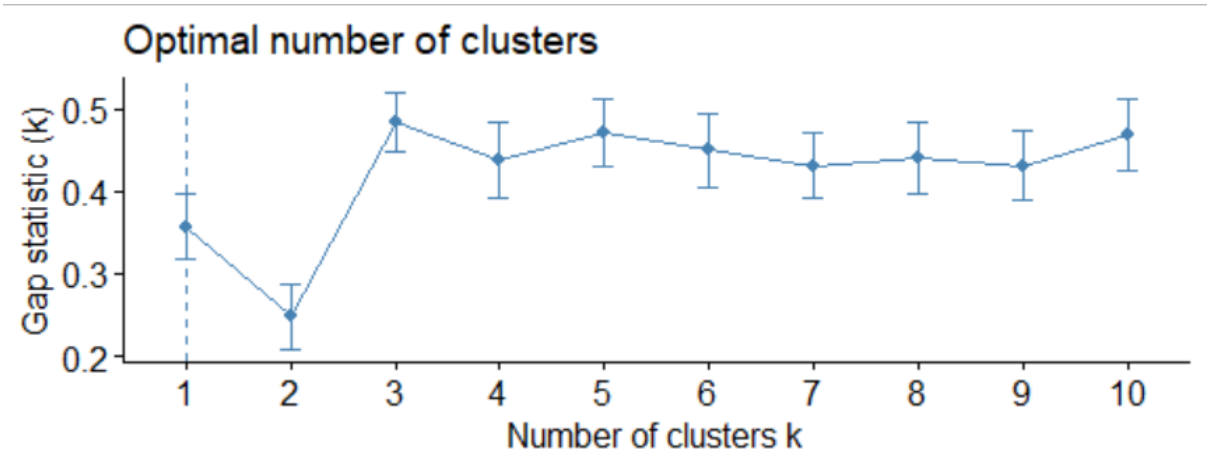
elbow



Average silhouette



Gap statistic method



Q) For each cluster, how would you describe it given your analysis?

Cluster 3 contains the countries which have the highest corruption index which suggests that there is less corruption here. In these countries the avg number of days to open a Business is on the lower side

Cluster 1 has a lower corruption index when compared to Cluster 3 which suggests that there is high corruption and in these countries the avg number of days to open a business is also lesser.

Cluster 2 has a low corruption index which suggests that there is higher corruption but the avg number of days to open a business is on the higher side.Cluster 3 has the highest GDP per capita and has the least corruption(higher corruption index)

Q) Based on your analysis, what is the relationship between education and corruption?

Cluster 3 which has a higher Compulsory Education years and higher Education spending as a percentage of GDP has lower corruption (high corruption index)

Cluster 1 and 2 has a lower compulsory education years and lower education spending and have higher corruption.

As education spending or education years increases corruption decreases correspondingly.