R script:

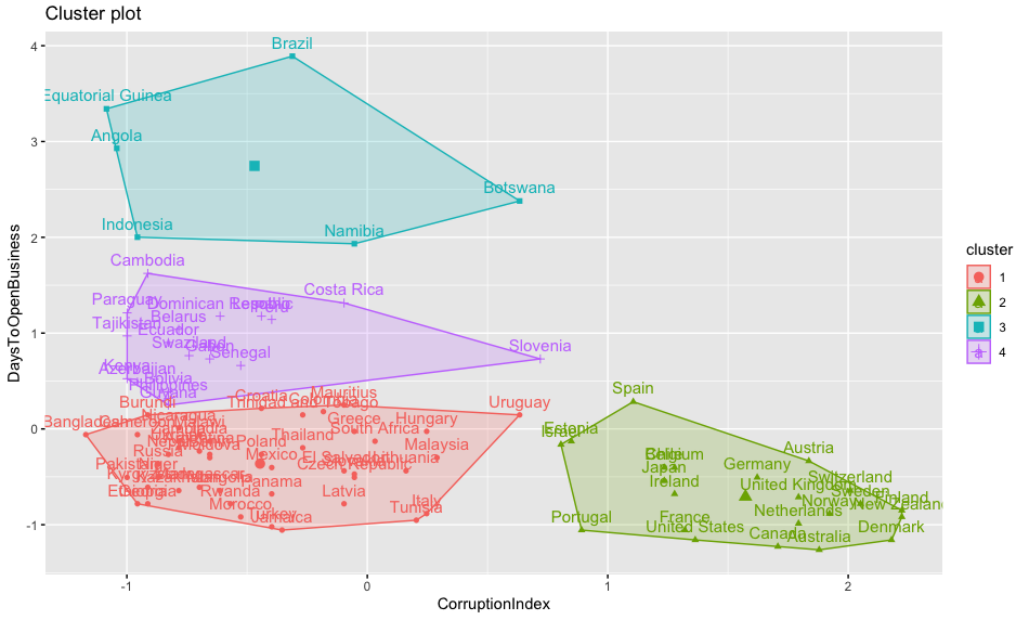
```
# Aadhithya Dinesh
# MIS 545 Section 02
# Lab11DineshA.R
# To import a dataset of country-level data and generate clusters
using the
# k-means clustering method.
# We will be importing csv files, assigning data types, generating
clusters,
# and interpreting clusters.

# install.packages("tidyverse")
# install.packages("factoextra")

library(tidyverse)
library(stats)
library(factoextra)
library(cluster)
library(gridExtra)

# set the working directory
setwd("~/MIS/Classes/MIS545/Assignments/Lab11")

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

# print the countries tibble
print(countries)

# print the structure of countries
print(str(countries))

# print the summary of countries
print(summary(countries))

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

# removing countries with missing data in any feature
countries <- countries %>% drop_na()
```

```
# print the summary of countries again to ensure no NA values are
present
print(summary(countries))

# scaling both features in the tibble so they have equal impact
countriesScaled <- countries %>%
  select(CorruptionIndex, DaysToOpenBusiness) %>% scale()

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

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

# display cluster sizes
print(countries4Clusters$size)

# display cluster centers (z-scores)
print(countries4Clusters$centers)

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

# optimizing the value for k using the methods below

# 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 analysis using 3 as the optimal number of clusters
countries3Clusters <- kmeans(x = countriesScaled,
```

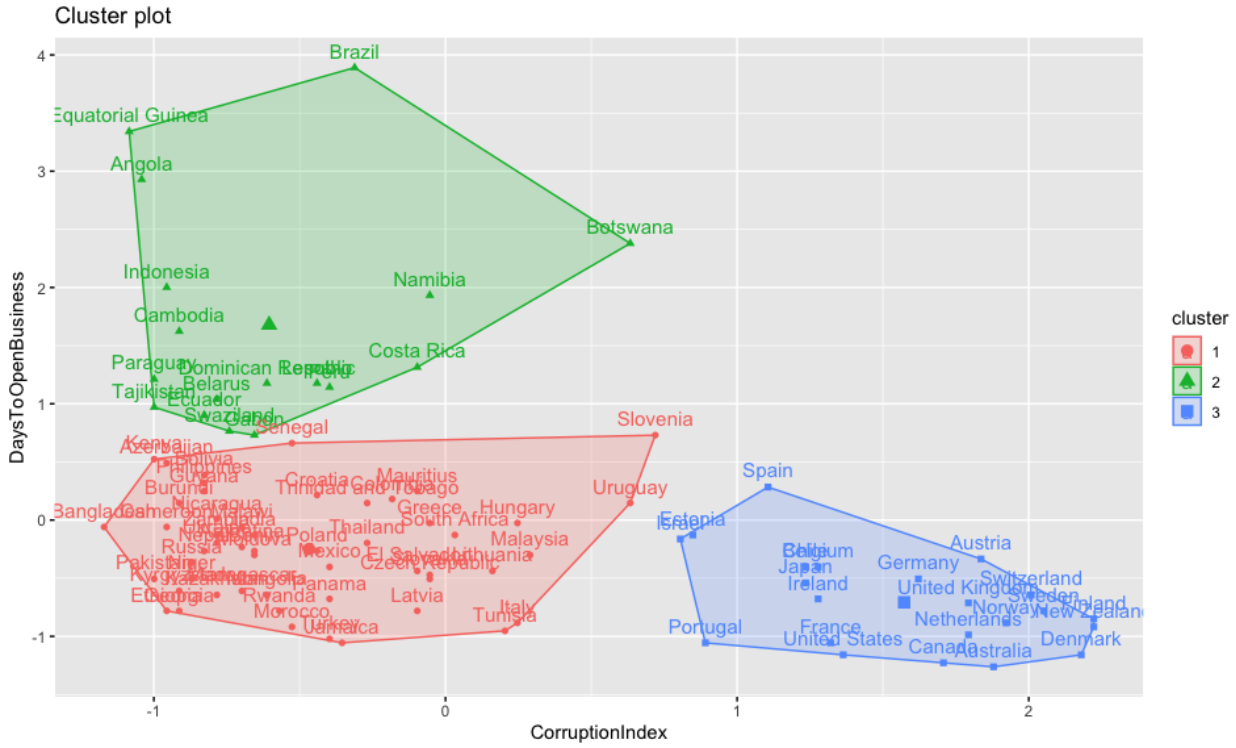
```
                              centers = 3,
                              nstart = 25)

# display cluster sizes
print(countries3Clusters$size)

# display cluster centers (z-scores)
print(countries3Clusters$centers)

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

# determining similarities and differences among all the features
countries %>%
  mutate(cluster = countries3Clusters$cluster) %>%
  select(cluster,
         CorruptionIndex,
         CompulsoryEducationYears,
         GiniCoefficient,
         GDPPerCapita,
         EduPercGovSpend,
         EduPercGDP,
         CompulsoryEducationYears,
         DaysToOpenBusiness) %>%
  group_by(cluster) %>%
  summarise_all("mean")
```

Cluster Plot visualizations:

k = 4

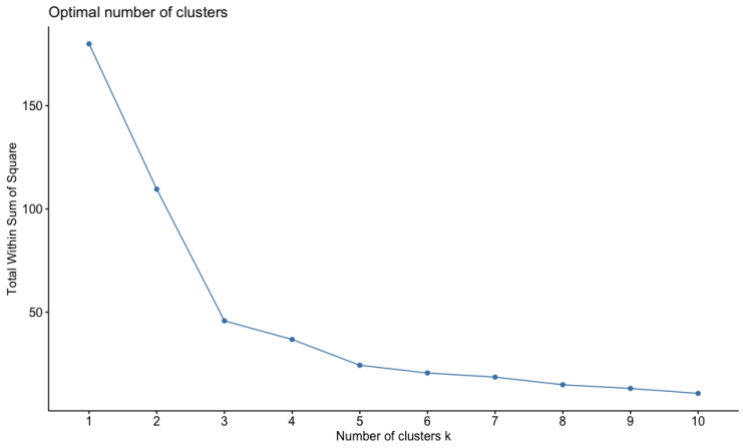


k=3

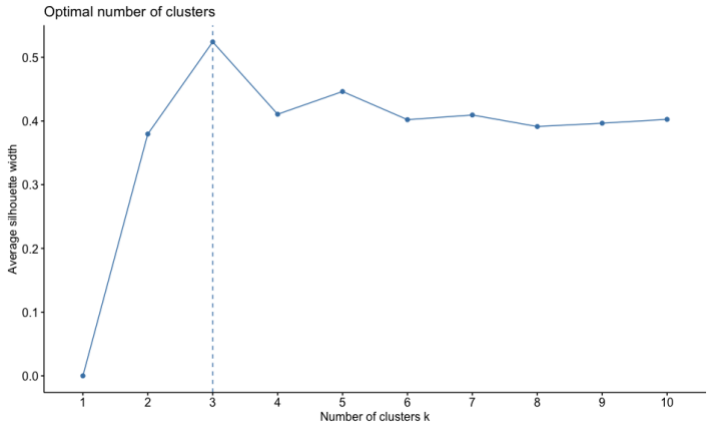


Z-optimization plots:

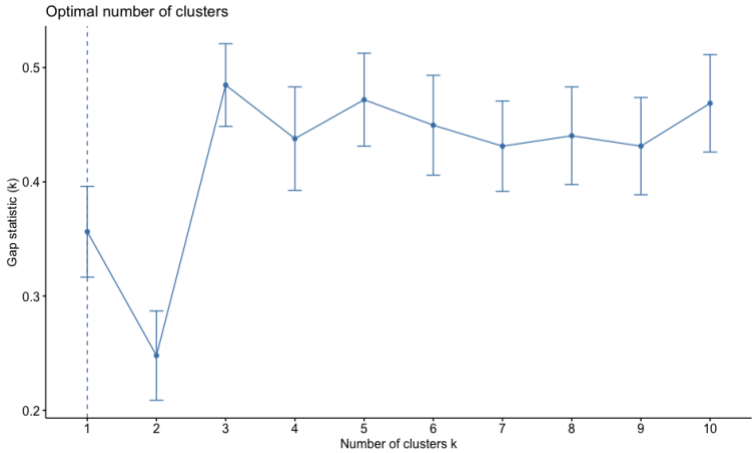
1. Elbow method



2. Silhouette



3. Gap Statistic



Answers:

1. DaysToOpenBusiness seems to be having a higher z-score if the corruptionIndex of a country seems to be having a lower z-score. This means that in countries which have high corruption rates (lower z-scores), the number of days needed to open a business are usually high.

2. Higher the corruption index of a country (meaning lower corruption), the number of compulsory years of education in these countries seem to be higher than the average. Furthermore, the Education percentage of GDP also seems to be higher in countries where the corruption index is higher.

3. Here are the similarities and differences based on different attributes for all the countries:

	cluster	CorruptionIndex	CompulsoryEducationYears	GiniCoefficient	GDPPerCapita	EduPercGovSpend	EduPercGDP	DaysToOpenBusiness
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	3.34	8.17	39.9	4921.	0.158	0.0452	31.5
2	2	3.02	8.53	48.8	3903.	0.139	0.0388	87.6
3	3	8.09	10.7	33.2	36173.	0.133	0.0581	18.1

From this matrix we can see the following differences:

a. Cluster 3 has almost 5 points above the other two clusters in terms of corruptionIndex and also has compulsoryEducationYears of over 2 years greater than the other two clusters.
b. The GDPPerCapita of the countries where the corruption is the least (cluster 3) seems to be extremely high. So, lower the corruption, higher the GDPPerCapita.
c. The EduPercGOvSPend is almost similar, so there are no differences between clusters based on this attribute.
d. The percentage of GDP spent on Education is over 1% more for cluster 3 as compared to both the other clusters. This shows that with more number of people being educated, the chances of corruption reduces.
e. The GiniCoefficient is similar as well.
f. The number of days to open a business drastically falls as the corruption index average value increases. In cluster 3 where the corruption index is 8.09, the number of days on average is the least (18.1) This shows that with less corruption, the process to open a business is streamlined and transparent.