**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**

**Diagram

Description automatically generated with medium confidence**

**k=3**

**Chart

Description automatically generated with medium confidence**

**Z-optimization plots:**

1. **Elbow method**

**Chart, line chart

Description automatically generated**

1. **Silhouette**

**Chart, line chart

Description automatically generated**

1. **Gap Statistic**

**Chart, line chart, box and whisker chart

Description automatically generated**

**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:

A screenshot of a computer

Description automatically generated

From this matrix we can see the following differences:

1. 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.
2. 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.
3. The EduPercGOvSPend is almost similar, so there are no differences between clusters based on this attribute.
4. 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.
5. The GiniCoefficient is similar as well.
6. 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.