**STA6704 - Final project R-code**

***# Load required libraries***

library(DataExplorer)

library(ggplot2)

library(dplyr)

library(tidyr)

library(BioStatR)

library(corrplot)

library(qgraph)

library("factoextra")

library(psych)

library(dendextend)

library(cluster)

library(gridExtra)

#library(Hmisc)

#library("FactoMineR")

***# Load the dataset***

country\_data <- read.csv(file = "C:\\Users\\divya\\Desktop\\Country-data.csv" , header = TRUE)

head(country\_data)

dim(country\_data)

var\_definition <- read.csv(file = "C:\\Users\\divya\\Desktop\\data-dictionary.csv" , header = TRUE)

var\_definition

***# Standard statistical measures***

summary(country\_data)

str(country\_data)

***# Missing values***

sum(is.na(country\_data))

plot\_missing(country\_data, ggtheme = theme\_bw(), title = "Missing Values in Country Dataset") +

theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

***# Univariate analysis: Histogram***

numeric\_data <-country\_data[, sapply(country\_data, is.numeric)]

hist(numeric\_data)

title("Histogram of variables", line = -1, outer = TRUE, font.main=4)

***# Univariate analysis: Boxplot***

boxplot\_data <- country\_data %>% gather(Attributes, values, c(2:10))

box\_plot <- ggplot(boxplot\_data, aes(x=Attributes, y=values)) + geom\_boxplot(aes(fill=Attributes)) + ggtitle("Boxplot of Varibles")

box\_plot + facet\_wrap( ~ Attributes, scales="free") + theme\_bw() +guides(fill = "none") + theme(

plot.title = element\_text(hjust = 0.5, face="bold.italic"))

***# Boxplot – comparing variables:***

data\_compare <- country\_data %>% gather(Attributes, values, c(2:5, 7:9))

ggplot(data\_compare, aes(x=reorder(Attributes, values, FUN=median), y=values,

fill=Attributes)) +

geom\_boxplot(show.legend=FALSE) +

labs(title="Boxplots of Variables - Comparison") +

theme\_bw() +

theme(axis.title.y=element\_blank(), axis.title.x=element\_blank(), plot.title = element\_text(hjust = 0.5, face="bold.italic")) +

ylim(-10, 220) + coord\_flip()

***# scatter plot – bivariate:***

upper.panel <- function(x, y){

points(x,y, pch=21, col ="blue")

r <- round(cor(x, y), digits=2)

txt <- paste0("R = ", r)

usr <- par("usr"); on.exit(par(usr))

par(usr = c(0, 1, 0, 1))

text(0.5, 0.8, txt)

}

pairs(country\_data[,2:10], lower.panel = NULL, upper.panel = upper.panel, diag.panel = panel.hist , main = "Bivariate distribution - Country Data" , font.main=4)

***# Correlation Matrix:***

country\_corr <- cor(country\_data[,2:10])

corrplot(country\_corr, type="upper", tl.col = "black", tl.srt = 45, addCoef.col ='black', number.cex = 0.8)

title("Correlation Matrix - Country Data", line = -1, outer = TRUE, font.main=4)

qgraph(country\_corr, labels = colnames(country\_data[,2:10]),

layout = "spring",

label.cex = 1, normalize=TRUE, edge.width=0.65 )

title("Correlation graph - Country Data", line = -2, outer = TRUE, font.main=4)

***# Scale dataset:***

country\_df <- as.data.frame(country\_data)

country\_scaled <- scale(country\_df[,2:10])

rownames(country\_scaled) <- country\_df[,1]

***# PCA using prcomp***

country\_pca <- prcomp(country\_scaled)

country\_pca

summary(country\_pca)

country\_eigen <- get\_eigenvalue(country\_pca)

country\_eigen

factoextra::fviz\_eig(country\_pca, addlabels = TRUE) + theme\_classic() + labs(title = "Variances - PCA", x = "Principal Components", title.size = 4) +theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

#screeplot(country\_pca)

factoextra::fviz\_pca\_var( X = country\_pca, col.var = "contrib", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE, title = "Variables Contribution to Principal Components") + theme\_classic() + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

barplot(country\_pca$rotation[,"PC1"], main = "Barplot of PCA - PC1", font.main = 4)

barplot(country\_pca$rotation[,"PC2"], main = "Barplot of PCA - PC2", font.main = 4)

***#***fviz\_contrib(country\_pca, choice = "var", axes = 1)

#fviz\_contrib(country\_pca, choice = "var", axes = 2)

***#PCA using princomp***

country\_prin1 <- principal(country\_scaled, nfactors=2, rotate="none")

country\_prin1

country\_prin\_var1 <- loadings(country\_prin1)

country\_prin\_var1

qgraph(country\_prin\_var1, labels = colnames(country\_data[,2:10]), label.cex = 1, normalize=TRUE, edge.width=0.65 )

title("Variables contribution to principal components, PC1 & PC2", line = -3, outer = TRUE, font.main=4)

country\_prin2 <- principal(country\_scaled, nfactors=2, rotate="varimax")

country\_prin2

country\_prin\_var2 <- loadings(country\_prin2)

country\_prin\_var2

qgraph(country\_prin\_var2, labels = colnames(country\_data[,2:10]), label.cex = 1, normalize=TRUE, edge.width=0.65 )

title("Variables contribution to principal components(rotated), PC1 & PC2", line = -3, outer = TRUE, font.main=4)

***# Hierarchical clustering:***

distance\_country <- dist(country\_scaled) # default euclidean

hiercluster\_comp <- hclust(distance\_country, method = "complete")

dend\_country <- as.dendrogram(hiercluster\_comp)

plot(dend\_country, main = "Dendogram - Hierarchical Clustering(euclidean, complete)", font.main = 4)

dendogram\_3 <- color\_branches(dend\_country, k =3)

plot(dendogram\_3, main = "Dendogram - Hierarchical Clustering (3 clusters color)", font.main = 4)

cutree\_hierclust <- cutree( tree = hiercluster\_comp, k = 3)

prin\_comp\_hier <- data.frame(prcomp(x = country\_scaled, center = FALSE,scale. = FALSE )$x[,1:2], Name = rownames(country\_scaled), Cluster = as.character(cutree\_hierclust), stringsAsFactors = FALSE)

ggplot(prin\_comp\_hier) +

aes(x = PC1,y = PC2,color = Cluster,fill = Cluster,label = Name ,group = Cluster) +

geom\_point() +

ggrepel::geom\_text\_repel(color = "black",size = 3) +

ggtitle("Scatter plot – hierarchical cluster (euclidean, complete)") +

theme\_bw() +

theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

dist\_country\_canberra <- dist(country\_scaled, "canberra")

hiercluster\_wardd <- hclust(distance\_country, method = "ward.D")

dend\_ward\_country <- as.dendrogram(hiercluster\_wardd)

plot(dend\_ward\_country, main = "Dendogram - Hierarchical Clustering(canberra, ward.D)", font.main = 4)

dendogram\_ward\_3 <- color\_branches(dend\_ward\_country, k =3)

plot(dendogram\_ward\_3, main = "Dendogram - Hierarchical Clustering (3 clusters color)", font.main = 4)

cutree\_hierclust\_ward <- cutree( tree = hiercluster\_wardd, k = 3)

princomp\_hier\_ward <- data.frame(prcomp(x = country\_scaled, center = FALSE,scale. = FALSE )$x[,1:2], Name = rownames(country\_scaled), Cluster = as.character(cutree\_hierclust\_ward), stringsAsFactors = FALSE)

ggplot(princomp\_hier\_ward) +

aes(x = PC1,y = PC2,color = Cluster,fill = Cluster,label = Name ,group = Cluster) +

geom\_point() +

ggrepel::geom\_text\_repel(color = "black",size = 3) +

ggtitle("Scatter plot – hierarchical cluster (Canberra, ward.D)") +

theme\_bw() +

#theme(legend.position = "none")+

theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

cutree\_hierclust\_ward

***# Kmeans Clustering:***

kmeans2\_country <- kmeans(country\_scaled, centers = 2)

factoextra::fviz\_cluster( object = kmeans2\_country, data = country\_scaled, main = "Cluster Plot (Kmeans - 2 clusters)") + theme\_classic() + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

kmeans3\_country <- kmeans(country\_scaled, centers = 3)

factoextra::fviz\_cluster( object = kmeans3\_country, data = country\_scaled, main = "Cluster Plot (Kmeans - 3 clusters)") + theme\_classic() + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

kmeans4\_country <- kmeans(country\_scaled, centers = 4)

factoextra::fviz\_cluster( object = kmeans4\_country, data = country\_scaled, main = "Cluster Plot (Kmeans - 4 clusters)") + theme\_classic() + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

kmeans5\_country <- kmeans(country\_scaled, centers = 5)

factoextra::fviz\_cluster( object = kmeans5\_country, data = country\_scaled, main = "Cluster Plot (Kmeans - 5 clusters)") + theme\_classic() + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

set.seed(423)

factoextra::fviz\_nbclust(x = country\_scaled, FUNcluster = kmeans, method = "wss") + labs(title = "Elbow Method") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

set.seed(423)

factoextra::fviz\_nbclust(x = country\_scaled, FUNcluster = kmeans, method = "silhouette") + labs(title = "Silhouette Method") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

set.seed(423)

factoextra::fviz\_nbclust(x = country\_scaled, FUNcluster = kmeans, method = "gap\_stat") + labs(title = "Gap Statistics") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"))

par(mfrow = c(2, 2))

plot(silhouette(kmeans2\_country$cluster, daisy(country\_scaled)),

col = c("#cc3300", "#006600"), main = " ", font.main = 4)

plot(silhouette(kmeans3\_country$cluster, daisy(country\_scaled)),

col = c("#cc3300", "#006600", "#0099FF"), main = " ", font.main = 4)

plot(silhouette(kmeans4\_country$cluster, daisy(country\_scaled)),

col = c("#cc3300", "#006600", "#0099FF","#AA4371"), main = " ", font.main = 4)

plot(silhouette(kmeans5\_country$cluster, daisy(country\_scaled)),

col = c("#cc3300", "#006600", "#0099FF","#AA4371", "#2C728EFF"), main = " ", font.main = 4)

title("Silhouette plot - kmeans Clusters", line = -1.5, outer = TRUE)

country\_df["kmeans3"] = kmeans3\_country$cluster

***# Visualizing clusters in boxplots***

par(mfrow = c(3, 3))

boxplot(child\_mort~kmeans3,data=country\_df, col = rainbow(3), xlab = "", main = "Child Mortality Rate")

boxplot(exports~kmeans3,data=country\_df, col = rainbow(3) , xlab = "", main = "Exports")

boxplot(health~kmeans3,data=country\_df, col = rainbow(3) , xlab = "", main = "Health")

boxplot(imports~kmeans3,data= country\_df, col = rainbow(3), xlab = "", main = "Imports")

boxplot(income~kmeans3,data=country\_df, col = rainbow(3), xlab = "", main = "Income")

boxplot(inflation~kmeans3,data=country\_df, col = rainbow(3), xlab = "", main = "Inflation")

boxplot(life\_expec~kmeans3,data=country\_df, col = rainbow(3), xlab = "", main = "Life Expectancy")

boxplot(total\_fer~kmeans3,data=country\_df, col = rainbow(3),xlab = "", main = "Total Fertility")

boxplot(gdpp~kmeans3,data=country\_df, col = rainbow(3), xlab = "", main = "GDP per capita")

title("Kmeans 3 Clusters Vs Input variables", line = -1, outer = TRUE)

#par(mfrow=c(1,1))

dev.off()

mean\_kmeans3 <- country\_df %>%

group\_by(kmeans3) %>%

summarise(n = n(),

child\_mort = mean(child\_mort),

exports = mean(exports),

health = mean(health),

imports = mean(imports),

income = mean(income),

inflation = mean(inflation),

life\_expec = mean(life\_expec),

total\_fer = mean(total\_fer),

gdpp = mean(gdpp))

mean\_kmeans3

country\_df["country\_category"] = NA

country\_df$country\_category[country\_df$kmeans3 == 1] <- "Under-Developed"

country\_df$country\_category[country\_df$kmeans3 == 2] <- "Developing"

country\_df$country\_category[country\_df$kmeans3 == 3] <- "Developed"

***#Barplot of input variables and clusters:***

ggplot(data = country\_df, aes(x = reorder(country,-child\_mort), y = child\_mort, fill = country\_category)) + geom\_bar(stat="identity") + xlab("Country") + ylab("Child Mortality") + theme\_classic() + theme(axis.text.x=element\_blank(), axis.ticks.x=element\_blank()) + scale\_fill\_discrete(name = "Countries Clusters") + ggtitle("Barplot of Child Mortality - Clusters") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic")) + theme(legend.position = c(0.9, 0.9))

ggplot(data = country\_df, aes(x = reorder(country,-gdpp), y = gdpp, fill = country\_category)) + geom\_bar(stat="identity") + xlab("Country") + ylab("GDP") + theme\_classic() + theme(axis.text.x=element\_blank(), axis.ticks.x=element\_blank()) + scale\_fill\_discrete(name = "Countries Clusters") + ggtitle("Barplot of GDP per capita - Clusters") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic")) + theme(legend.position = c(0.9, 0.9))

ggplot(data = country\_df, aes(x = reorder(country,-income), y = income, fill = country\_category)) + geom\_bar(stat="identity") + xlab("Country") + ylab("Income") + theme\_classic() + theme(axis.text.x=element\_blank(), axis.ticks.x=element\_blank()) + scale\_fill\_discrete(name = "Countries Clusters") + ggtitle("Barplot of Income - Clusters") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic")) + theme(legend.position = c(0.9, 0.9))

ggplot(data = country\_df, aes(x = reorder(country,-life\_expec), y = life\_expec, fill = country\_category)) + geom\_bar(stat="identity") + xlab("Country") + ylab("Life expectancy") + theme\_classic() + theme(axis.text.x=element\_blank(), axis.ticks.x=element\_blank()) + scale\_fill\_discrete(name = "Countries Clusters") + ggtitle("Barplot of Life Expectancy - Clusters") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic")) + theme(legend.position = c(0.9, 0.9))

***#Scatter plot of input variables and clusters:***

p1 <- ggplot(data = country\_df, aes(x = child\_mort, y = life\_expec, color = country\_category)) + geom\_point() + xlab("Child Mortality") + ylab("Life Expectancy") + theme\_bw() + theme( axis.text.x=element\_blank(), axis.ticks.x=element\_blank()) + scale\_fill\_discrete(name = "Countries Clusters") + ggtitle("Child Mortality Vs Life Expectancy - Clusters") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic")) + theme(legend.position = c(0.85, 0.895))

p2 <- ggplot(data = country\_df, aes(x = income, y = gdpp, color = country\_category)) + geom\_point() + xlab("Income") + ylab("GDP") + theme\_bw() + theme( axis.text.x=element\_blank(), axis.ticks.x=element\_blank()) + scale\_fill\_discrete(name = "Countries Clusters") + ggtitle("Income Vs GDP - Clusters") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic")) + theme(legend.position = c(0.85, 0.895))

grid.arrange(p1, p2, ncol = 2)

***# Countries list***

display <- ttheme\_minimal( core = list(bg\_params = list(fill = blues9[1:4], col=NA), fg\_params = list(fontface=3)), colhead = list(fg\_params=list(col="navyblue", fontface=4L)), rowhead = list(fg\_params = list(col = "blue", fontface=4L)))

dev <- subset(country\_df, kmeans3 == 3)$country

dim(dev) <- c(6,6)

plot.new()

title("Developed Countries", line = -8, outer = TRUE, font.main=4)

grid.table(dev, theme = display, rows = NULL)

developing <- subset(country\_df, kmeans3 == 2)$country

dim(developing) <- c(14,6)

plot.new()

title("Developing Countries", line = -2, outer = TRUE, font.main=4)

grid.table(developing, theme = display, rows = NULL)

under\_dev <- subset(country\_df, kmeans3 == 1)$country

under\_dev <- under\_dev[1:48]

dim(under\_dev) <- c(8,6)

plot.new()

title("Under-Developed Countries", line = -2, outer = TRUE, font.main=4)

grid.table(under\_dev, theme = display, rows = NULL)

***#Visualizing clusters in map***

world\_map <- map\_data('world')

colnames(world\_map)[5] <- "country"

#data\_merger <- left\_join(country\_df, world\_map, by = "country")

#data\_merger %>% filter(is.na(long))

#subset(country\_data, country %in% (unique(subset(data\_merger,kmeans4 == 4)$country)))

world\_map["country"][world\_map["country"] == "Barbuda"] <- "Antigua and Barbuda"

world\_map["country"][world\_map["country"] == "USA"] <- "United States"

world\_map["country"][world\_map["country"] == "UK"] <- "United Kingdom"

world\_map["country"][world\_map["country"] == "Republic of Congo"] <- "Congo, Rep."

world\_map["country"][world\_map["country"] == "Democratic Republic of the Congo"] <- "Congo, Dem. Rep."

world\_map["country"][world\_map["country"] == "Laos"] <- "Lao"

world\_map["country"][world\_map["country"] == "Ivory Coast"] <- "Cote d'Ivoire"

world\_map["country"][world\_map["country"] == "Kyrgyzstan"] <- "Kyrgyz Republic"

world\_map["country"][world\_map["country"] == "Macedonia"] <- "Macedonia, FYR"

world\_map["country"][world\_map["country"] == "Micronesia"] <- "Micronesia, Fed. Sts."

world\_map["country"][world\_map["country"] == "Slovakia"] <- "Slovak Republic"

world\_map["country"][world\_map["country"] == "Saint Vincent"] <- "St. Vincent and the Grenadines"

world\_map["country"][world\_map["country"] == "Grenadines"] <- "St. Vincent and the Grenadines"

data\_merger <- left\_join(world\_map, country\_df, by = "country")

ggplot(data\_merger, aes(x = long, y = lat, group = group, fill = factor(kmeans3))) + geom\_polygon(color = 'gray') + scale\_fill\_manual(values = c("#1F968BFF", "#FDE725FF","#481B6DFF"), limits = c(1,2,3), labels = c("Under-Developed","Developing","Developed"), name = "Countries Clusters") + theme\_bw() + ggtitle("Kmeans Country Clusters on World Map") + theme(plot.title = element\_text(hjust = 0.5, face="bold.italic"), legend.position = c(.9,.1))