Multivariate CW

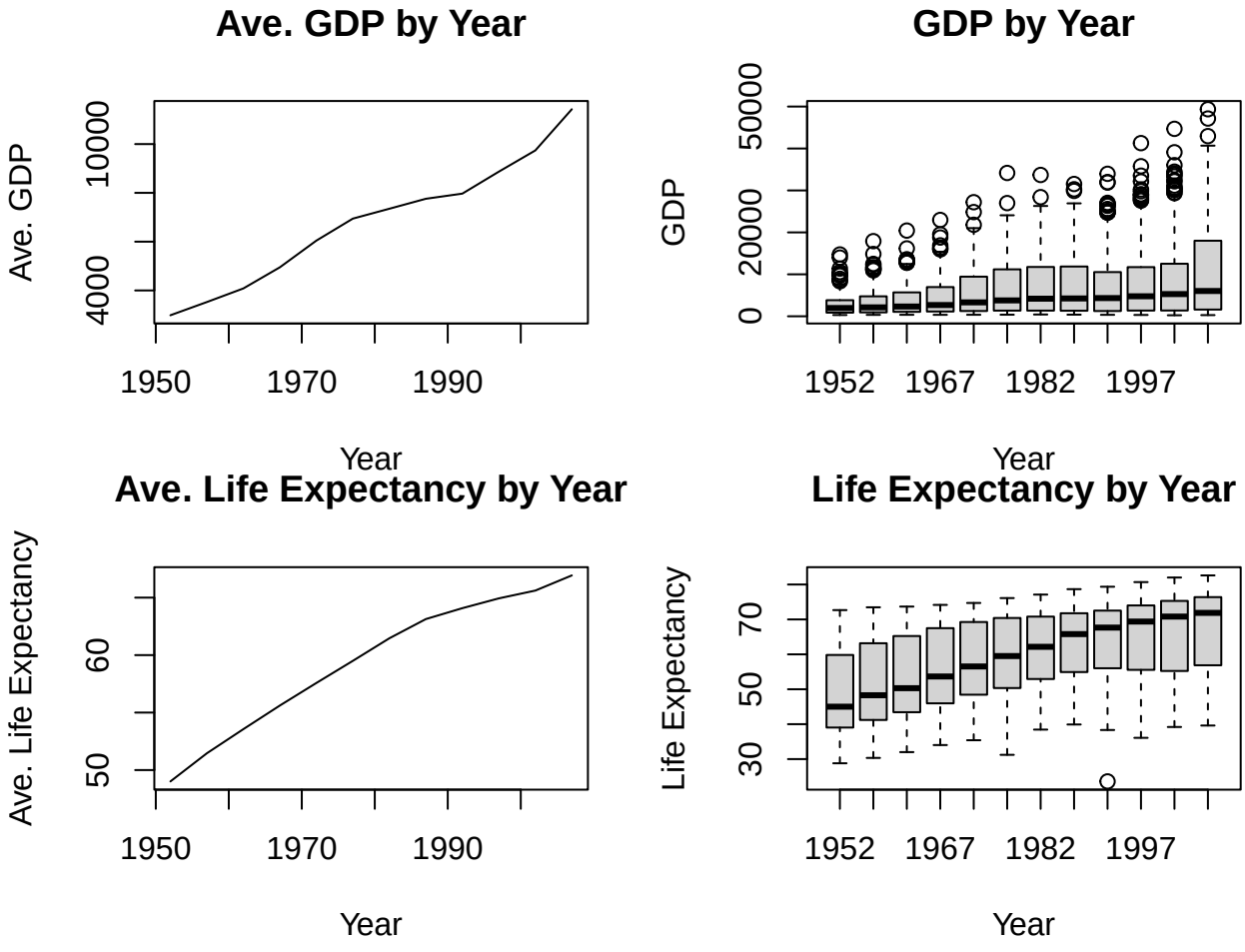
Liangxiao LI

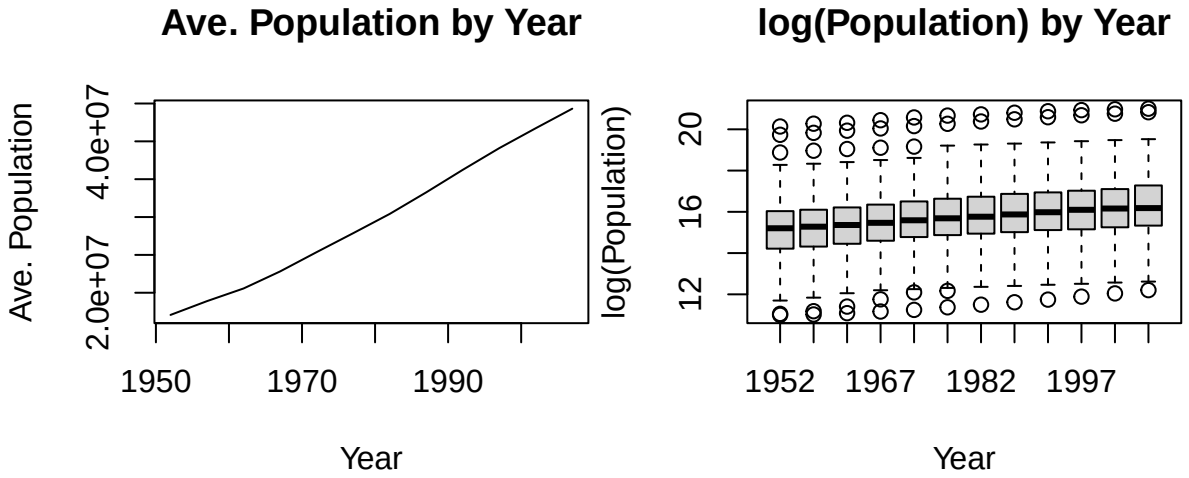
2024-04-13

Part1: Exploratory Data Analysis

First we load the data

Since we have 141 rows of different countries, therefore visulizing individual line plots for each country would result in a cluttered figure. For such a large number of states, I'll focus on aggregate plots by calculating the average GDP, life expectancy, population accross the United State.





From the above line plots, it can be concluded that the average GDP, life expectancy and population are growing steadily across the globe as years goes by.

In the following section we plot the box plots for each dataframe, since the population value exceed the R integer boundary, we'll plot 'years x log(population)' for the population box plot

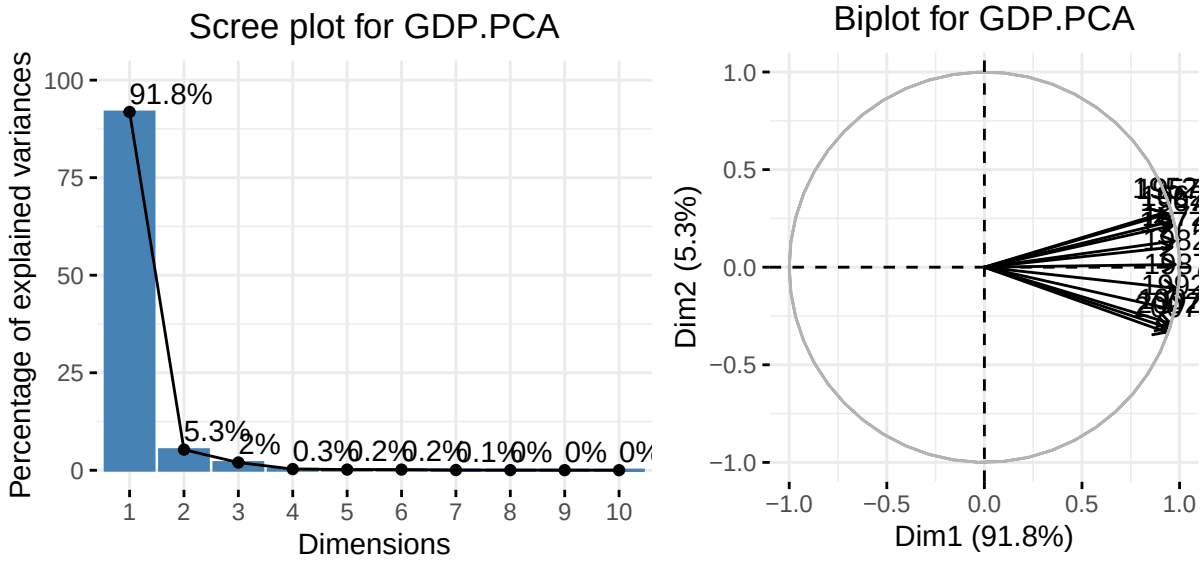
Part2: Principal component analysis

Here for all three different datasets, I perform PCA horizontally, treating each country as a data point and each year as a feature.

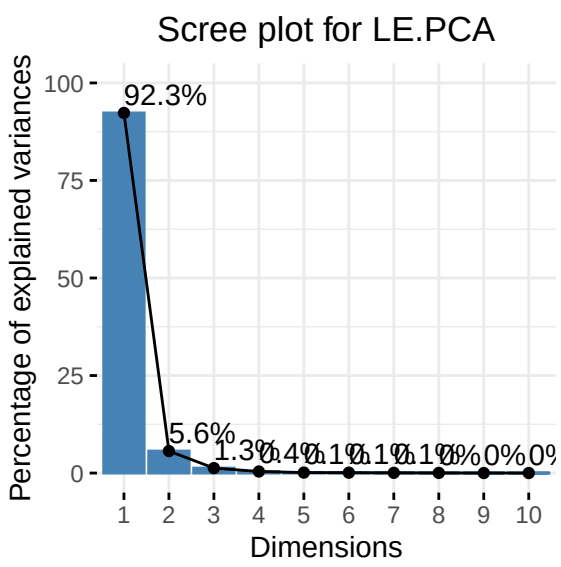
Since the columns are 12 different years, this means the features are measuring similar entities, therefore we should perform PCA based on *S(sample covariance matrix)* for gdp, life expectancy and population.

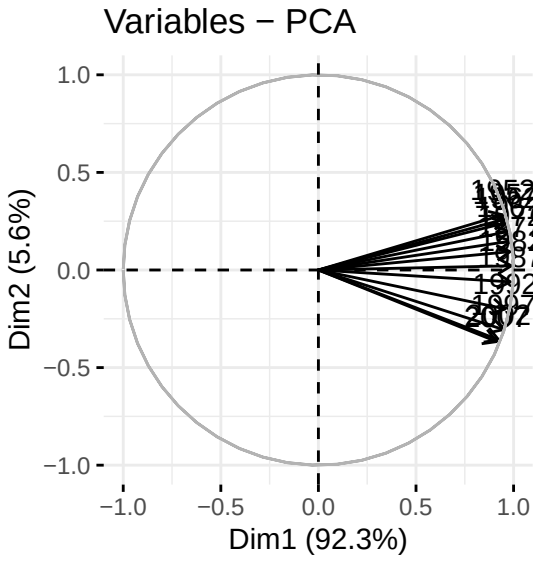
Part2.1: Number of PCs to retain

First we plot the scree plot and the corresponding biplots to decide how many PCs we should retain.

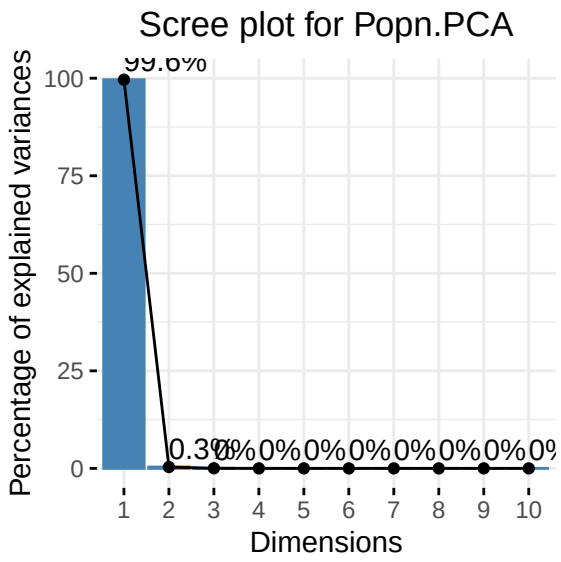


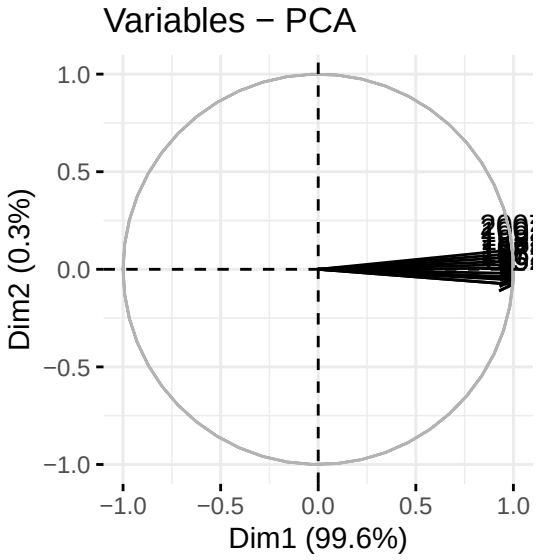
From the scree plot, we'd retain PC1 and PC2 as they explained 91.8% and 5.3% of the variance within the data.





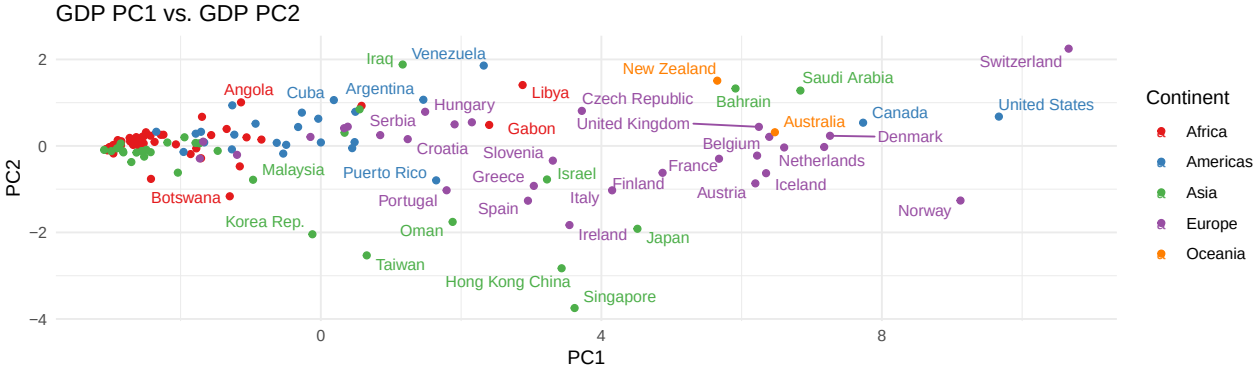
From the scree plot, we'd retain PC1 and PC2 for life expectancy as they explained 92.3% and 5.6% of the variance within the data.





From the scree plot, we'd retain PC1 for population as it explained 99.6% of the variance within the data.

Part2.2: Scatter plots for PCs and interpretations



```
##      1952 1957 1962 1967 1972 1977 1982  1987  1992  1997  2002  2007
## PC1 0.28 0.28 0.29 0.29 0.29 0.29 0.29  0.30  0.29  0.29  0.28  0.28
## PC2 0.35 0.34 0.30 0.27 0.17 0.13 0.02 -0.14 -0.27 -0.36 -0.39 -0.42
```

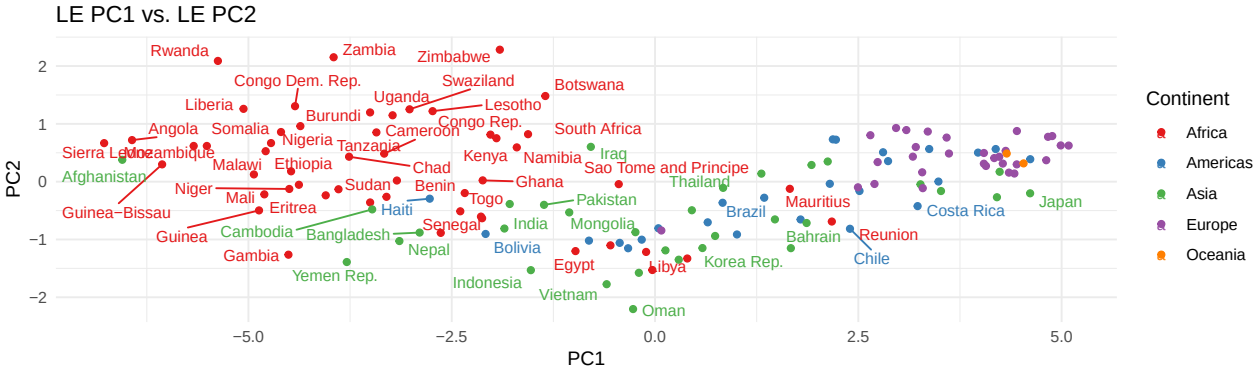
Above is the scatter plot of PC scores and the loadings for GDP PC1/2:

a) PC1 seems to measure a general trend for GDP, which can be seen from the loadings above (All loadings are approximately $+0.28 \sim +0.3$). Higher PC1 indicate higher overall GDP.

Therefore **European** countries have generally higher GDP while **African** countries have generally lower GDP.

b) PC2 Might represent a cyclic variation which such as economic fluctuations. As higher PC2 indicate higher GDP between 1952-1982, while lower PC2 indicate higher GDP between 1987-2007.

Therefore **Asia** countries tend to have higher GDP after 1987, especially Singapore.




```
##      1952 1957 1962 1967 1972 1977 1982  1987  1992  1997  2002  2007
## PC1 0.28 0.29 0.29 0.29 0.30 0.29 0.30  0.30  0.29  0.28  0.28  0.27
## PC2 0.34 0.32 0.30 0.24 0.18 0.11 0.03 -0.07 -0.24 -0.37 -0.43 -0.45
```

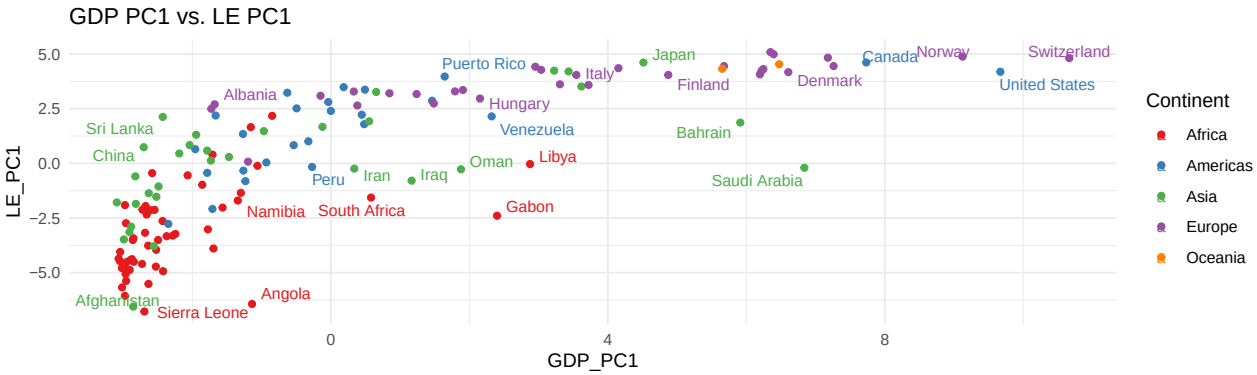
Above is the scatter plot of PC scores and the loadings for life expectancy PC1/2:

a) PC1 seems to measure a general trend for life expectancy, which can be seen from the loadings above (All loadings are approximately $+0.28 \sim +0.3$). Higher PC1 indicate higher overall life expectancy.

Therefore **European** countries have the highest overall life expectancy while **African** countries have the lowest overall life expectancy.

b) PC2 Might represent a cyclic variation which such as . As higher PC2 indicate higher life expectancy between 1952-1982, while lower PC2 indicate higher life expectancy after 1987.

Therefore **Asian** countries tend to have higher life expectancy after 1987, especially Oman. This means that Asian countries started developing quickly after 1987.




```
##          1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 2002 2007
## LE_PC1   0.28 0.28 0.29 0.29 0.29 0.29 0.29  0.3 0.29 0.29 0.28 0.28
## GDP_PC1 0.28 0.29 0.29 0.29 0.30 0.29 0.30  0.3 0.29 0.28 0.28 0.27
```

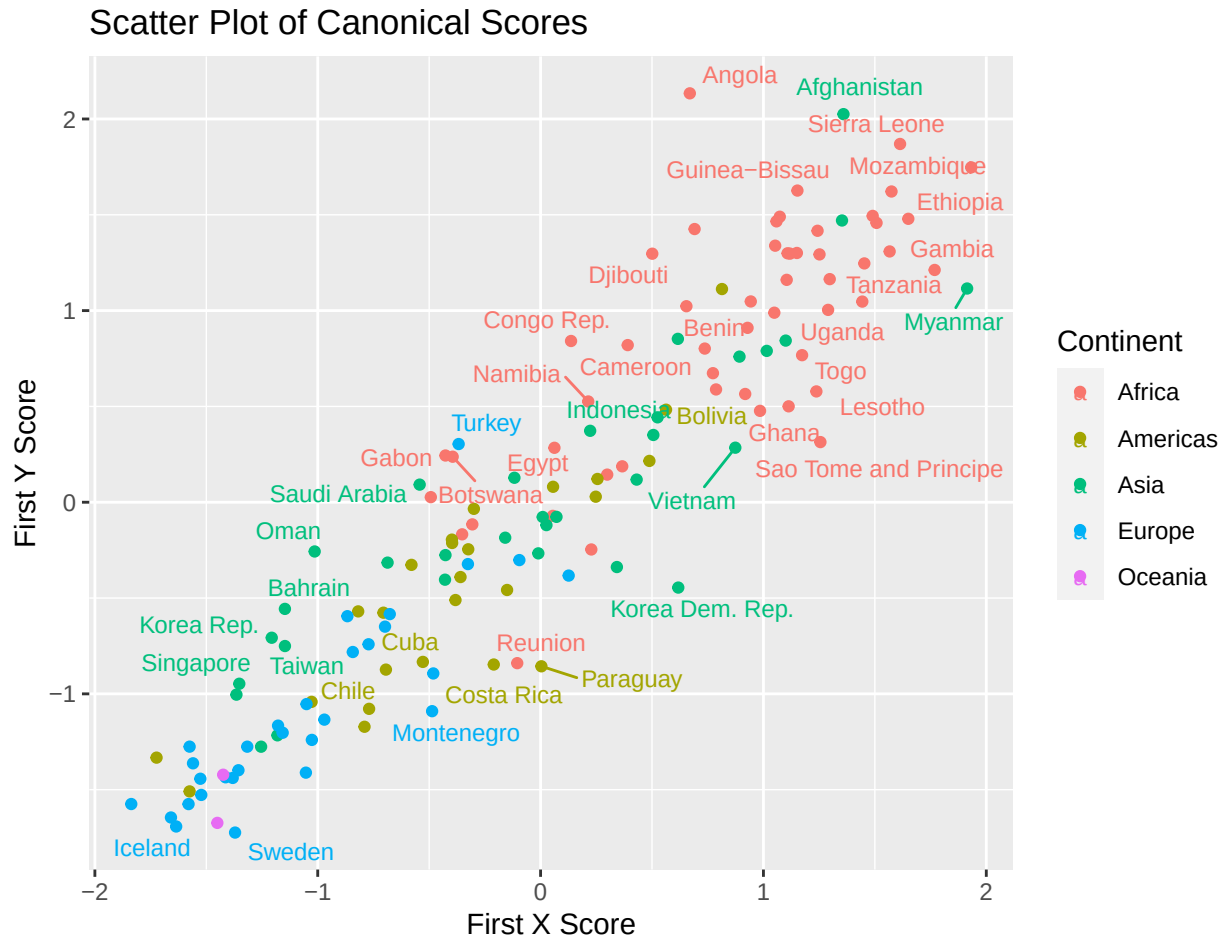
Above is the scatter plot of PC scores and the loadings for the first PC score for life expectancy against first PC score for GDP:

As mentioned in previous sections, GDP_PC1 and LE_PC1 both measure general trends for GDP and LE. Higher GDP_PC1 indicate higher overall GDP, higher LE_PC1 indicate higher overall life expectancy.

Therefore **African** tend to have both the lowest overall GDP/life expectancy. **European** and **American** have the higher overall GDP/life expectancy.

For individual countries, **United States**, **Switzerland** and **Norway** have both high GDP_PC1 and LE_PC1, indicating best economic situation and citizen health condition. **Sierra Leone** and **Afgharistan** have low GDP_PC1 and LE_PC1, indicating worst economic situation and citizen health condition.

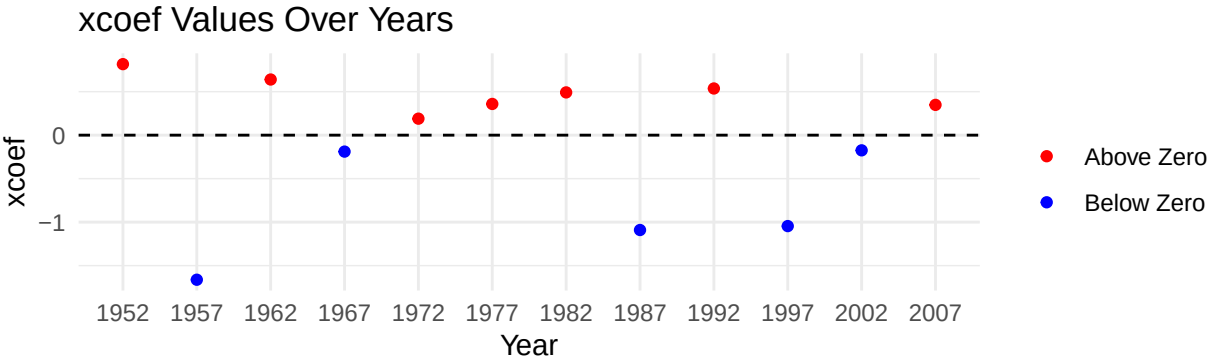
Canonical correlation analysis

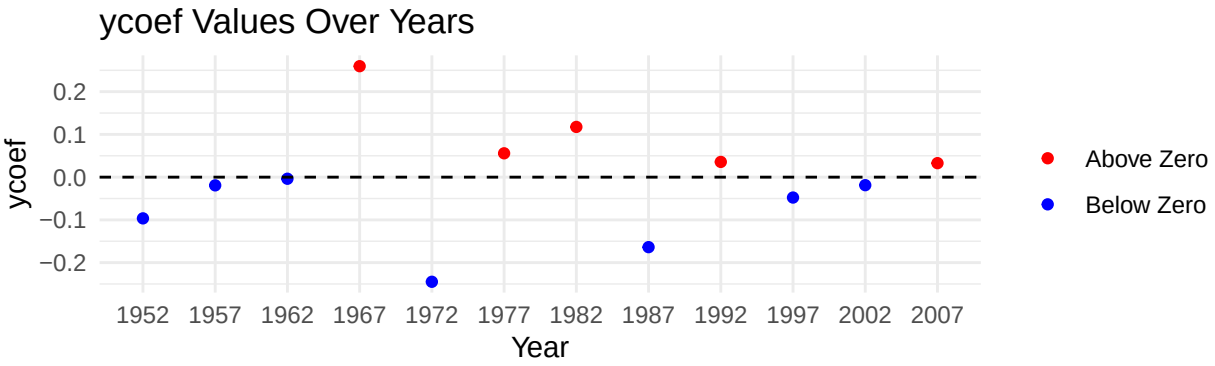


Above is the scatter plot of the first pair of CC variables. The observations are colored based on the continent information.

From this plot we can conclude that there are strong correlation between the first pair of CC variables.

To help interpret the first pair of canonical variables, I ploted the their loadings(x/y coefficients) in the following graph.





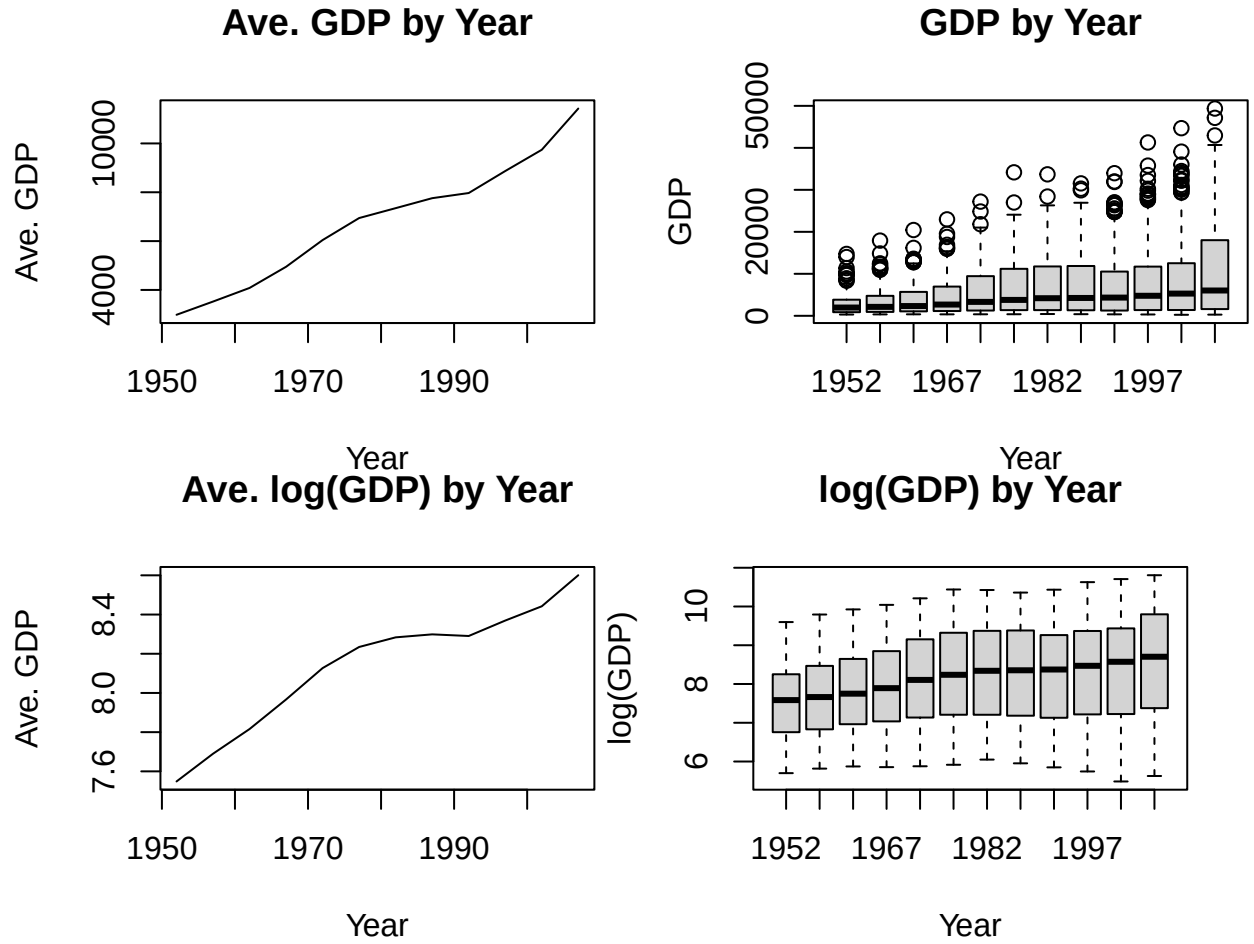
These coefficients contributes to the x score and y score.

1) X

We see that for log(gdp), the higher the value in the red dots mean higher value in the x score. This means that Europe countries has higher gdp in 1957, 1967,1987, 1997 and 2002, while the Africa countries has the opposite

Why log(gdp)?

To explain why we apply log(gdp) instead of gpd here, we need to look back at the exploratory analysis.

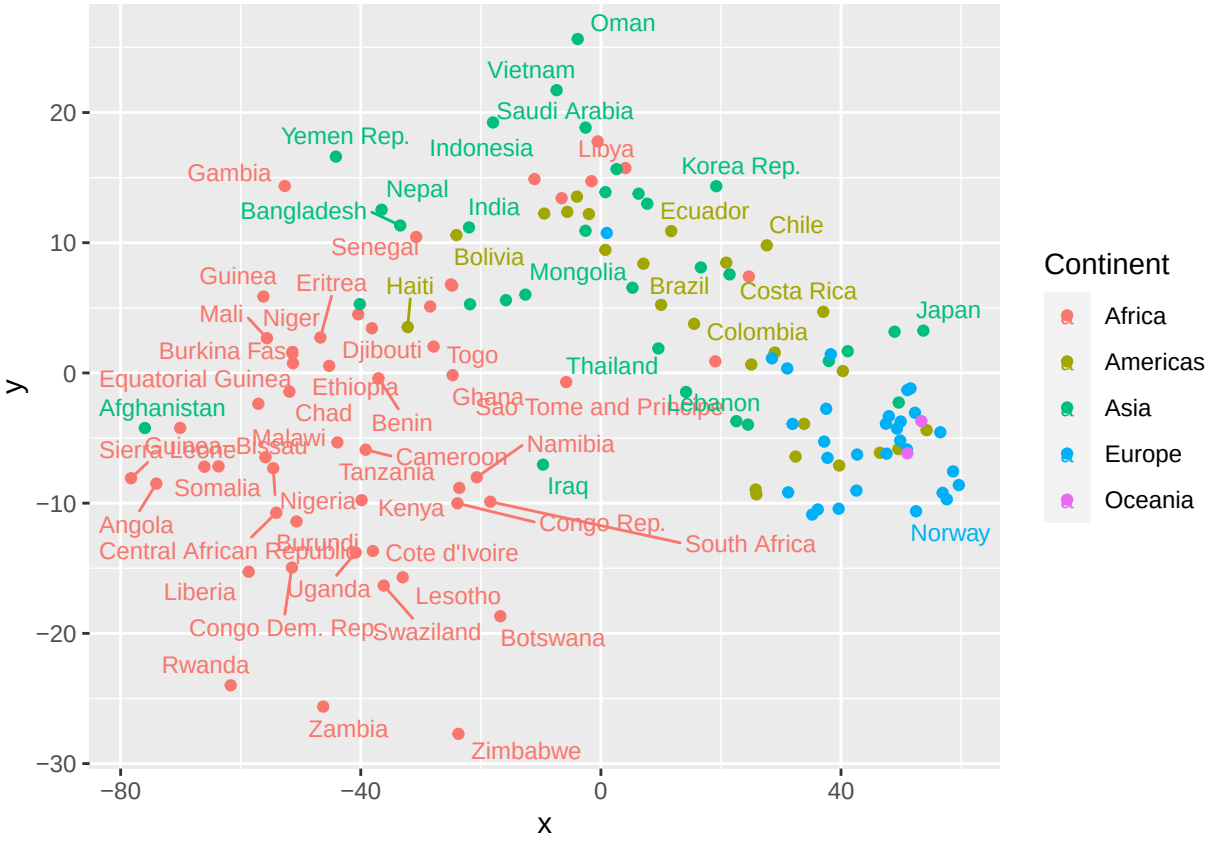


The above boxplots and lineplots show that log(gdp) removes the extreme values and the spread of data doesn't increase too dramatically. This helps make reduce the distribution less skewed, which is helpful for

the CCA.

Multidimensional scaling

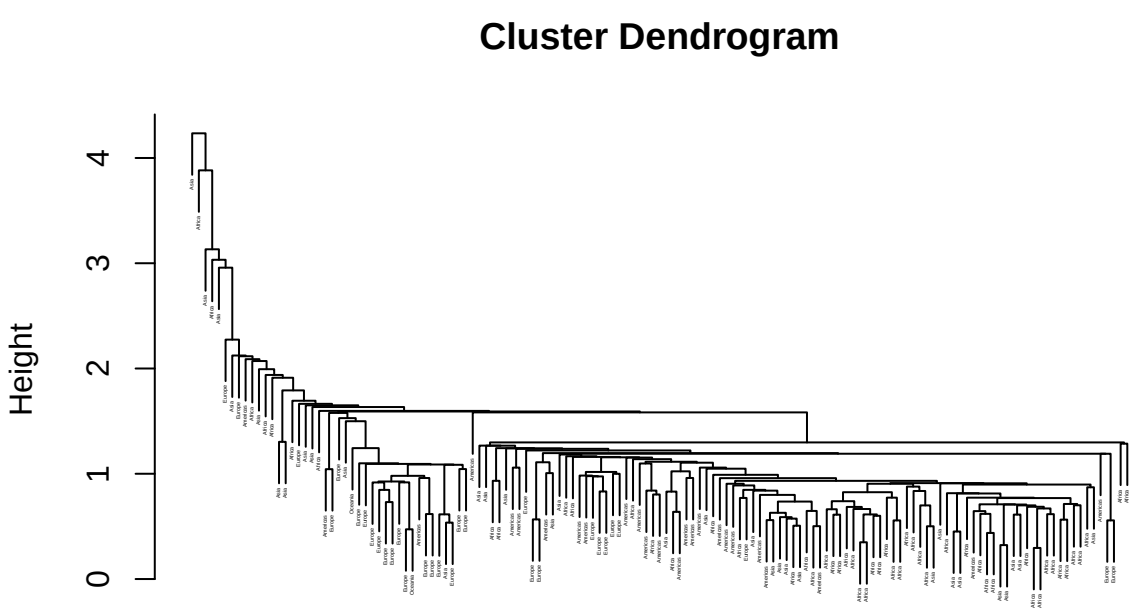
```
## Warning: ggrepel: 78 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Linear discriminant analysis

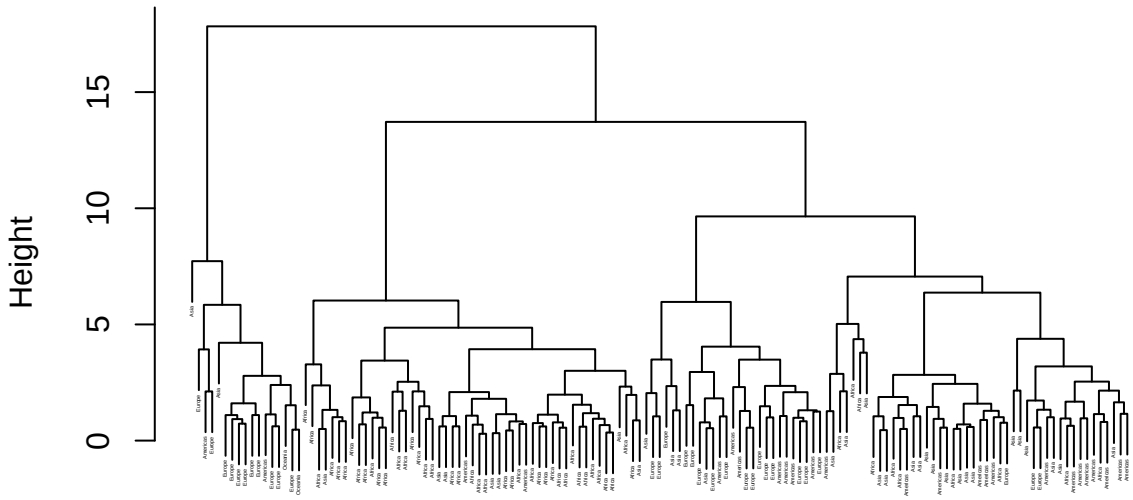
```
## [1] "The predictive accuracy is  60 %"
```

Clustering



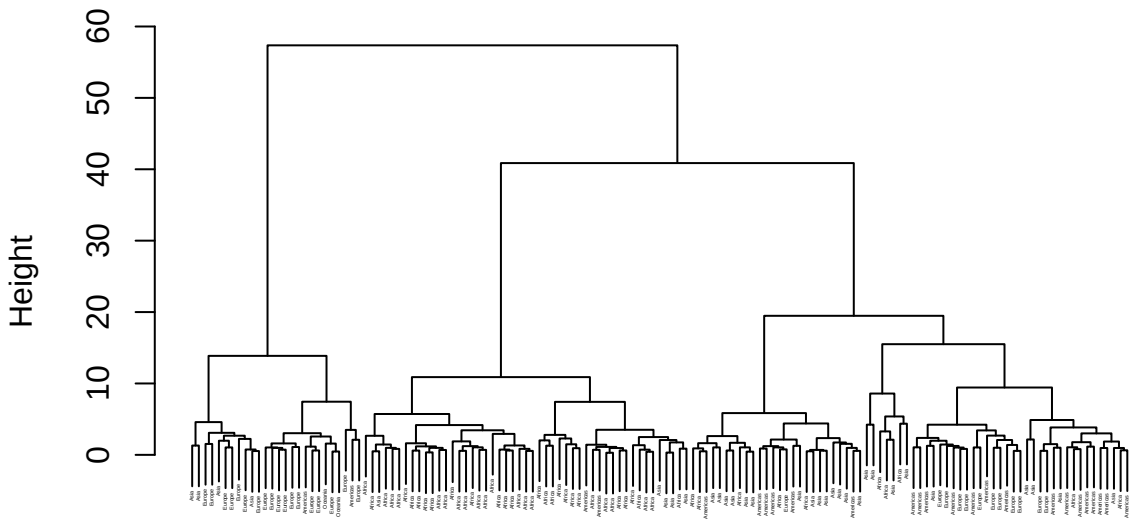
dist(UN.scaled[, 3:26], method = "euclidean")
hclust (*, "single")

```
##
## result Africa Americas Asia Europe Oceania
##      1     50       25   30     30       2
##      2      1        0    0      0       0
##      3      1        0    0      0       0
##      4      0        0    1      0       0
##      5      0        0    1      0       0
```

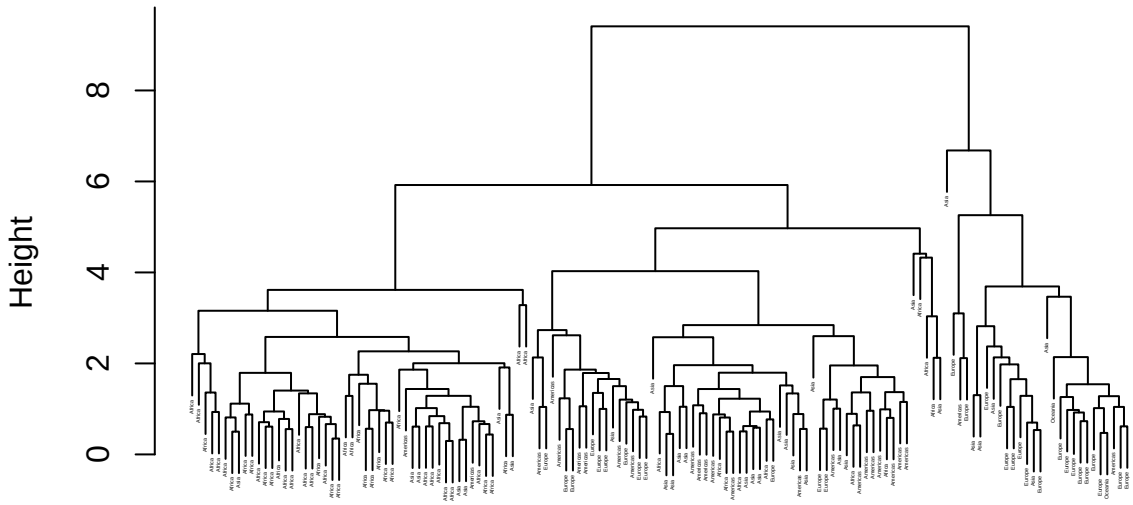


dist(UN.scaled[, 3:26], method = "euclidean")
hclust (*, "complete")

```
##
## result Africa Americas Asia Europe Oceania
##      1     10       14   19      3       0
##      2     42        2    7      0       0
##      3      0        7    4     16       0
##      4      0        2    1     11       2
##      5      0        0    1      0       0
```




```
##
## result Africa Americas Asia Europe Oceania
##      1      5        6   14      1       0
##      2     42        2    5      0       0
##      3      3        0    4      0       0
##      4      2       15    5     11       0
##      5      0        2    4     18       2
```




```
##
## result Africa Americas Asia Europe Oceania
##      1     10       21   19     12       0
##      2     42        2    7      0       0
##      3      0        1    5     16       2
##      4      0        1    0      2       0
##      5      0        0    1      0       0
```

Linear regression

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
UN <- read.csv('UN.csv')

gdp <- UN[,3:14] # The GDP per capita.
years <- seq(1952, 2007,5)
colnames(gdp) <- years
rownames(gdp) <- UN[,2]

lifeExp <- UN[,15:26] # the life expectancy
colnames(lifeExp) <- years
rownames(lifeExp) <- UN[,2]

popn <- UN[,27:38] # the population size
colnames(popn) <- years
rownames(popn) <- UN[,2]
```

```
library(ggplot2)
library(reshape2)
average_gdp <- apply(gdp, 2, mean)
plot(years, average_gdp, type = "l", xlab = "Year", ylab = "Ave. GDP", main = "Ave. GDP by Year")

boxplot(gdp, names = years, main = "GDP by Year", xlab = "Year", ylab = "GDP")

average_le <- apply(lifeExp, 2, mean)
plot(years, average_le, type = "l", xlab = "Year", ylab = "Ave. Life Expectancy", main = "Ave. Life Expe

boxplot(lifeExp, names = years, main = "Life Expectancy by Year", xlab = "Year", ylab = "Life Expectancy


average_popn <- apply(popn, 2, mean)
plot(years, average_popn, type = "l", xlab = "Year", ylab = "Ave. Population", main = "Ave. Population

boxplot(log(popn), names = years, main = "log(Population) by Year", xlab = "Year", ylab = "log(Populati
library(ggplot2)  # Make sure ggplot2 is loaded
gdp.pca <- prcomp(gdp, scale=TRUE)
le.pca <- prcomp(lifeExp, scale=TRUE)
popn.pca <- prcomp(popn, scale=TRUE)
#summary(gdp.pca)
#gdp.pca$rotation # the loadings/eigenvectors
#gdp.pca$center  # the sample mean
library(factoextra)
library(ggrepel)
#fviz_eig(gdp.pca, addlabels = TRUE, ylim = c(0, 100)) #Scree plot

plot <- fviz_eig(gdp.pca, addlabels = TRUE, ylim = c(0, 100)) + ggtitle("Scree plot for GDP.PCA")+ theme
print(plot)

plot <- fviz(gdp.pca, element = "var") + ggtitle("Biplot for GDP.PCA")+ theme(plot.title = element_text
print(plot)

#fviz_eig(le.pca, addlabels = TRUE, ylim = c(0, 100)) #Scree plot

fviz_eig(le.pca, addlabels = TRUE, ylim = c(0, 100)) + ggtitle("Scree plot for LE.PCA")+ theme(plot.titl

fviz(le.pca, element='var') #Interpretation of leading PC

#fviz_eig(le.pca, addlabels = TRUE, ylim = c(0, 100)) #Scree plot

fviz_eig(popn.pca, addlabels = TRUE, ylim = c(0, 100)) + ggtitle("Scree plot for Popn.PCA")+ theme(plot

fviz(popn.pca, element='var') #Interpretation of leading PC
#fviz(gdp.pca, element='var')#Interpretation of leading PC
pca_data <- data.frame(PC1 = gdp.pca$x[,1], PC2 = gdp.pca$x[,2], Continent = UN$continent)

ggplot(pca_data, aes(x = PC1, y = PC2, color = Continent)) +
  geom_point() + # This adds the scatter plot points
  geom_text_repel(aes(label = UN[,2]), size = 3)+
  labs(title = "GDP PC1 vs. GDP PC2",
       x = "PC1", y = "PC2") +
```

```
  theme_minimal()+ # Use a minimal theme for the plot
  scale_color_brewer(palette = "Set1")
#scale(gdp, center = TRUE) %*% gdp.pca$rotation[,1]
pc_loadings <- data.frame(PC1 = round(gdp.pca$rotation[,1],digits = 2), PC2 = round(gdp.pca$rotation[,2
# If you want to include the variable names as a row names in the table
rownames(pc_loadings) <- rownames(gdp.pca$rotation)

# Display the table
t(pc_loadings)

pca_data <- data.frame(PC1 = le.pca$x[,1], PC2 = le.pca$x[,2], Continent = UN$continent)

ggplot(pca_data, aes(x = PC1, y = PC2, color = Continent)) +
  geom_point() + # This adds the scatter plot points
  geom_text_repel(aes(label = UN[,2]), size = 3)+
  labs(title = "LE PC1 vs. LE PC2",
       x = "PC1", y = "PC2") +
  theme_minimal()+ # Use a minimal theme for the plot
  scale_color_brewer(palette = "Set1")

pc_loadings <- data.frame(PC1 = round(le.pca$rotation[,1],digits = 2), PC2 = round(le.pca$rotation[,2],
# If you want to include the variable names as a row names in the table
rownames(pc_loadings) <- rownames(le.pca$rotation)

# Display the table
t(pc_loadings)

pca_data <- data.frame(PC1 = gdp.pca$x[,1], PC2 = le.pca$x[,1], Continent = UN$continent)

ggplot(pca_data, aes(x = PC1, y = PC2, color = Continent)) +
  geom_point() + # This adds the scatter plot points
  geom_text_repel(aes(label = UN[,2]), size = 3)+
  labs(title = "GDP PC1 vs. LE PC1",
       x = "GDP_PC1", y = "LE_PC1") +
  theme_minimal()+ # Use a minimal theme for the plot
  scale_color_brewer(palette = "Set1")

pc_loadings <- data.frame(LE_PC1 = round(gdp.pca$rotation[,1],digits = 2), GDP_PC1 = round(le.pca$rotat
# If you want to include the variable names as a row names in the table
rownames(pc_loadings) <- rownames(popn.pca$rotation)

# Display the table
t(pc_loadings)
library(CCA)
#How the scores are calculated
#temp <-  scale(log(gdp), center = TRUE, scale =FALSE) #Centering the matrix
#print(temp %*% cca$xcoef[,1])

cca<-cc(log(gdp),lifeExp)
#plt.cc(cca, var.label=FALSE)

# Convert cca scores to a dataframe
scores_df <- data.frame(xscores = cca$scores$xscores[,1],
```

```
                        yscores = cca$scores$yscores[,1],
                        row.names = rownames(cca$scores$xscores))
# Assuming you have a dataframe `UN` with a column `continent` that matches the rows of your CCA analys
scores_df$continent <- UN$continent

ggplot(scores_df, aes(x = xscores, y = yscores, color = continent)) +
  geom_point() +
  geom_text_repel(aes(label = rownames(scores_df)), size = 3) +
  labs(x = "First X Score",
       y = "First Y Score",
       title = "Scatter Plot of Canonical Scores",
       color = "Continent")
# Plot the data using ggplot2
data <- data.frame(cca$xcoef[,1])
ggplot(data, aes(x = rownames(cca$xcoef), y = cca$xcoef[,1])) +
  geom_point(aes(color = ifelse(cca$xcoef[,1] > 0, "Above Zero", "Below Zero"))) + # Color points based
  scale_color_manual(values = c("Above Zero" = "red", "Below Zero" = "blue")) + # Assign colors
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") + # Add horizontal line at y=0

  labs(title = "xcoef Values Over Years",
       x = "Year",
       y = "xcoef") +
  theme_minimal()+ # Use a minimal theme+ # Use a minimal theme
  theme(legend.title = element_blank())

data <- data.frame(cca$ycoef[,1])
ggplot(data, aes(x = rownames(cca$ycoef), y = cca$ycoef[,1])) +
  geom_point(aes(color = ifelse(cca$ycoef[,1] > 0, "Above Zero", "Below Zero"))) + # Color points based
  scale_color_manual(values = c("Above Zero" = "red", "Below Zero" = "blue")) + # Assign colors
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") + # Add horizontal line at y=0
  labs(title = "ycoef Values Over Years",
       x = "Year",
       y = "ycoef") +
  theme_minimal()+ # Use a minimal theme+ # Use a minimal theme
  theme(legend.title = element_blank())
average_gdp <- apply(gdp, 2, mean)
plot(years, average_gdp, type = "l", xlab = "Year", ylab = "Ave. GDP", main = "Ave. GDP by Year")

boxplot(gdp, names = years, main = "GDP by Year", xlab = "Year", ylab = "GDP")

average_gdp <- apply(log(gdp), 2, mean)
plot(years, average_gdp, type = "l", xlab = "Year", ylab = "Ave. GDP", main = "Ave. log(GDP) by Year")

boxplot(log(gdp), names = years, main = "log(GDP) by Year", xlab = "Year", ylab = "log(GDP)")

head(cca$scores$xscores[,1]) # the canonical correlation variables
library(dplyr)
library(ggpubr) # repels figure labels
UN.transformed <- cbind(log(UN[,3:14]), UN[,15:26], log(UN[,27:38]))
UN.transformed <- dist(UN.transformed)
UN.transformed <- cmdscale(UN.transformed)
UN.transformed <- data.frame(UN.transformed,
                        row.names = rownames(cca$scores$xscores))
```

```r
colnames(UN.transformed) <- c("x", "y")

UN.transformed$continent <- UN$continent

ggplot(UN.transformed, aes(x = x, y = y, color = continent)) +
    geom_point() +  # This will color the points based on continent
    geom_text_repel(aes(label = row.names(UN.transformed)), size = 3) +
    labs(color = "Continent")  # Labeling the color legend as "Continent"
set.seed(123) # so that I get the same results each time.
#UN.scaled <- UN[,1:38]
#UN.scaled[,3:38] <- scale(UN[,3:38])
test.index <- sample(1:141, size=20)
UN.test <- UN[test.index,]
UN.train <- UN[-test.index,]

UN.lda<-lda(continent ~ gdpPercap_1952+gdpPercap_1957+gdpPercap_1962+gdpPercap_1967+gdpPercap_1972+gdpP
UN.pred <- predict(UN.lda, UN.test)
print(paste("The predictive accuracy is ",
sum(UN.pred$class== UN.test$continent)/dim(UN.test)[1]*100, "%"))
UN.scaled <- UN[,1:26]
UN.scaled[,3:26] <- scale(UN[,3:26])
UN.single <- hclust(dist(UN.scaled[,3:26],method="euclidean"),method="single")
plot(UN.single, labels=UN$continent,cex=0.2)
result <- cutree(UN.single, k=5)
table(result, UN$continent)
UN.complete <- hclust(dist(UN.scaled[,3:26],method="euclidean"),method="complete")
plot(UN.complete, labels=UN$continent,cex=0.2)
result <- cutree(UN.complete, k=5)
table(result, UN$continent)
UN.ward <- hclust(dist(UN.scaled[,3:26],method="euclidean"),method="ward.D2")
plot(UN.ward, labels=UN$continent,cex=0.2)
result <- cutree(UN.ward, k=5)
table(result, UN$continent)
UN.average <- hclust(dist(UN.scaled[,3:26],method="euclidean"),method="average")
plot(UN.average, labels=UN$continent,cex=0.2)
result <- cutree(UN.average, k=5)
table(result, UN$continent)
```