Multivariate CW

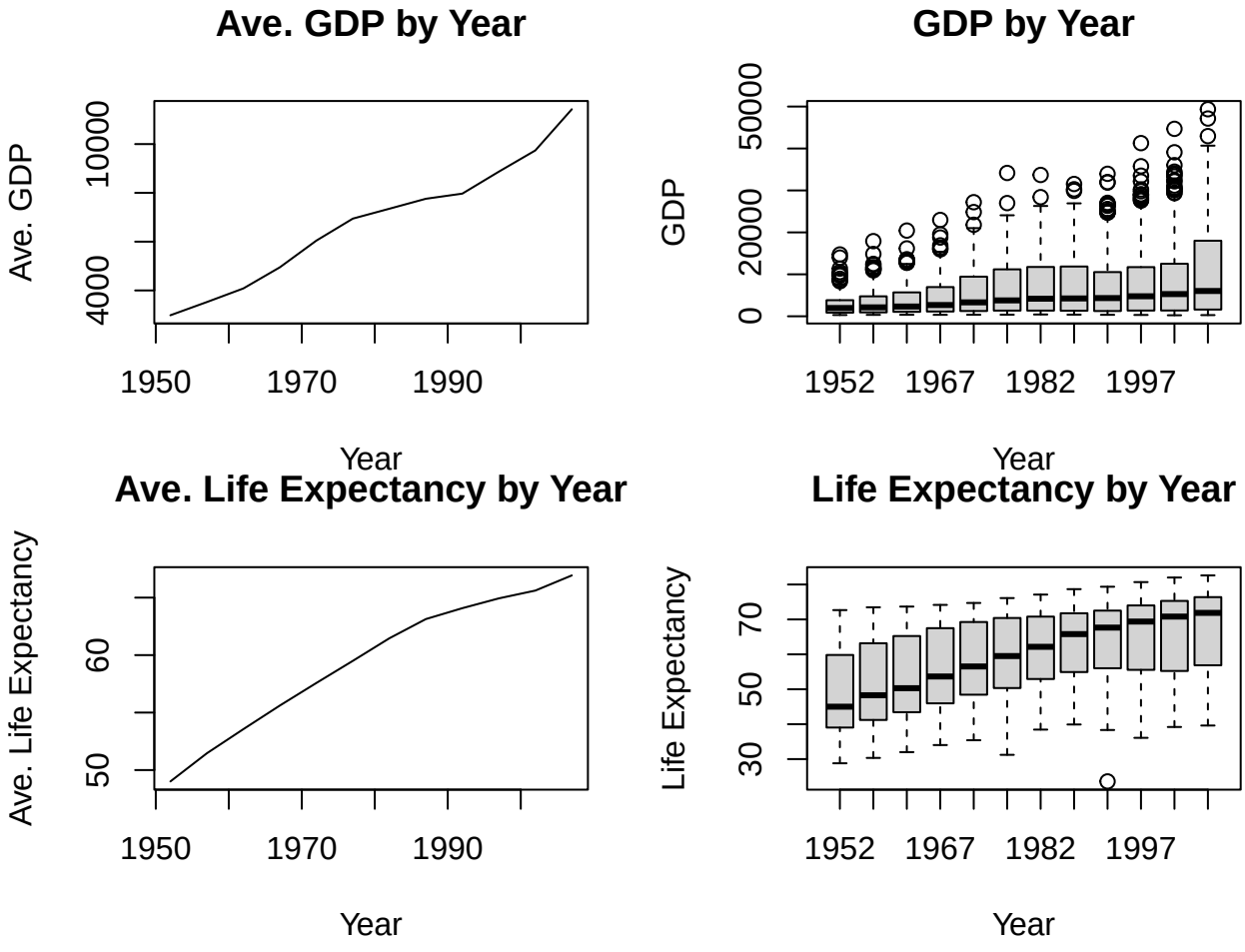
Liangxiao LI

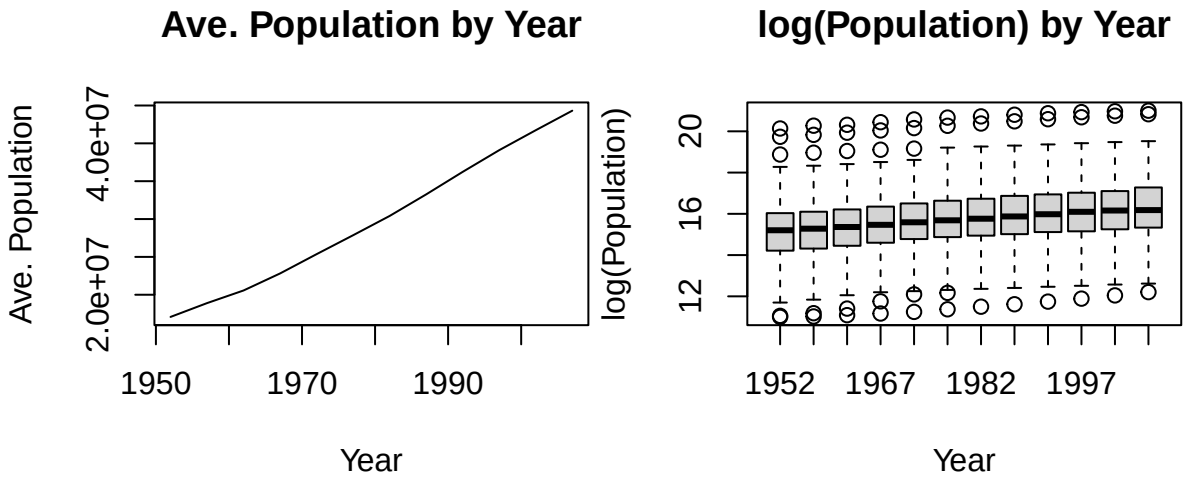
2024-04-13

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

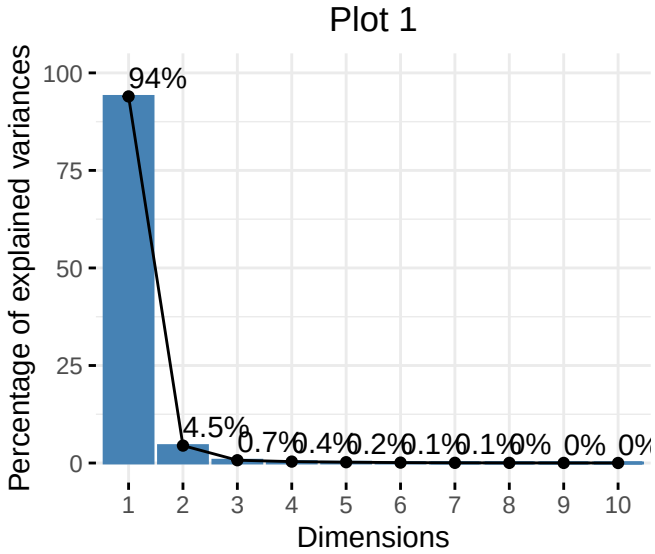
Principal component analysis

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

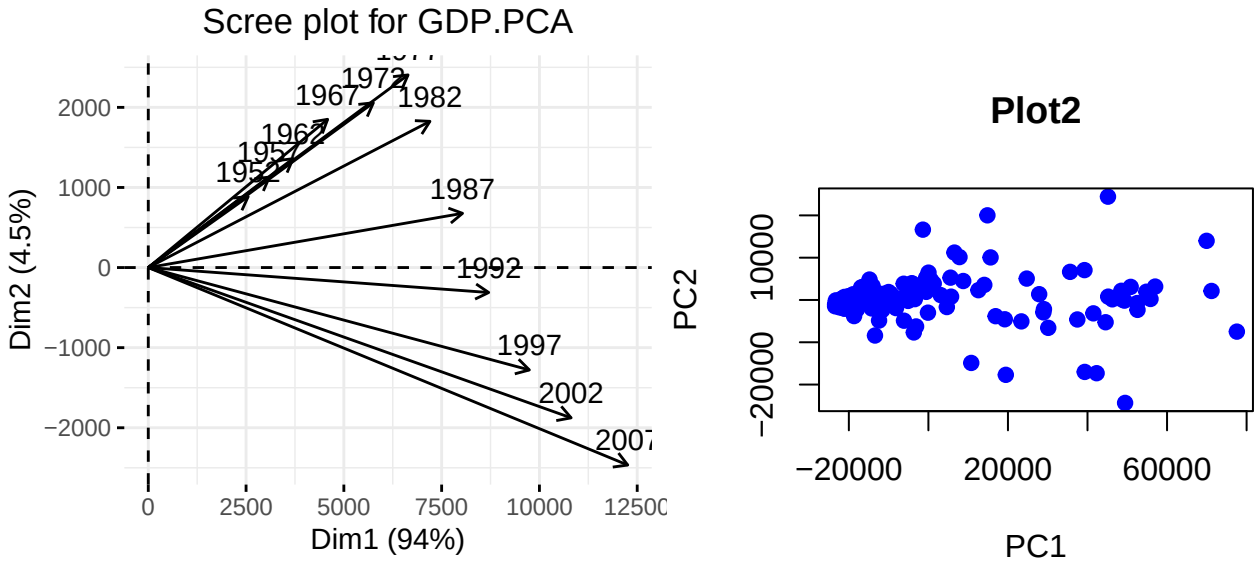
Since the columns are 12 different years, this means the features are measuring similar entities, therefore we should perform PCA based on S(sample covariance matrix) for gdp, life expectancy and population.

GDP

From the scree plot we see that the first principal component explains 94% of the variance within the data, while the second principal component explains 4.5% of the data. We therefore draw the conclusion that we choose the first two principal components for gdp.



Now that we have reduced the dimension to p = 2, we draw the following biplot/scatter plot to show the scores of different years.



1) Plot 1: This is a plot of the contributions of PC1/2 in the PCA.

PC1 is likely to capture an underlying pattern that increases with each year, such as economic growth or inflation. Because the years are aligned along PC1 with an ascending order from left to right,

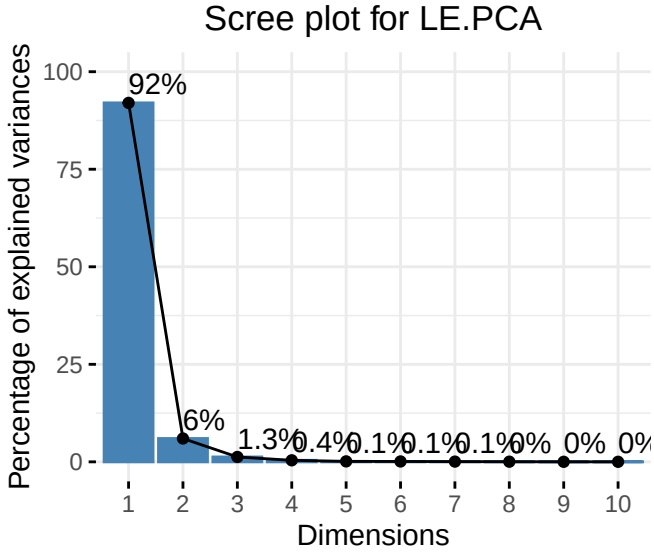
The distribution of years along PC2 is less clear, but it might represent a cyclic variation which such as economic fluctuations.

2) Plot 2: This is a scatter plot of the first two Principal Components for GDP

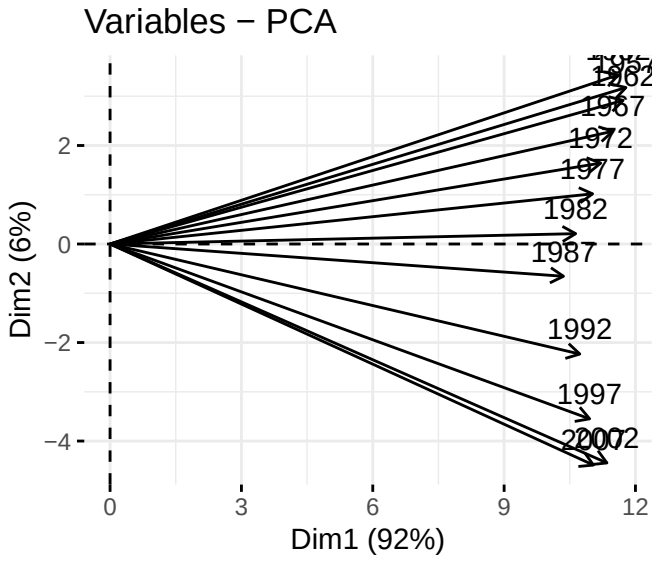
This shows that most observations are located around (PC1,PC2) = (-20000,0)

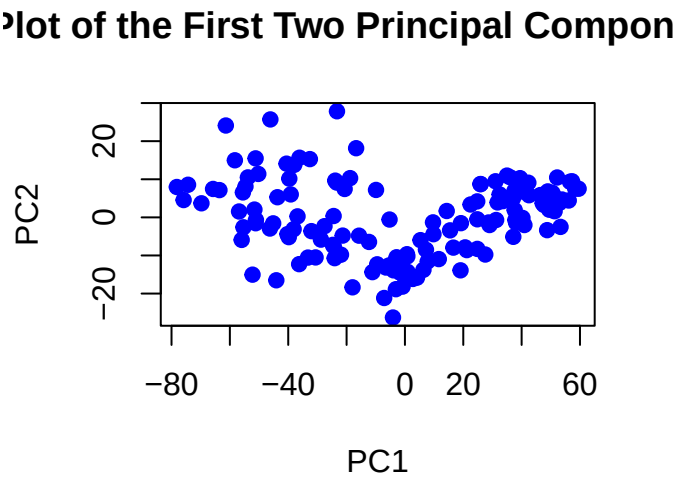
Life Expectancy

From the scree plot we see that the first principal component explains 92% of the variance within the data, while the second principal component explains 6% of the data. We therefore draw the conclusion that we choose two principal components for life expectancy.



Now that we have reduced the dimension to p = 2, we draw the following plots to show the scores of different years.





1) Plot 1: This is a plot of the contributions of PC1/2 in the PCA.

PC1 is likely to represent a general trend in life expectancy that changes across all the years, because the arrows for consecutive years are pointing in the same direction.

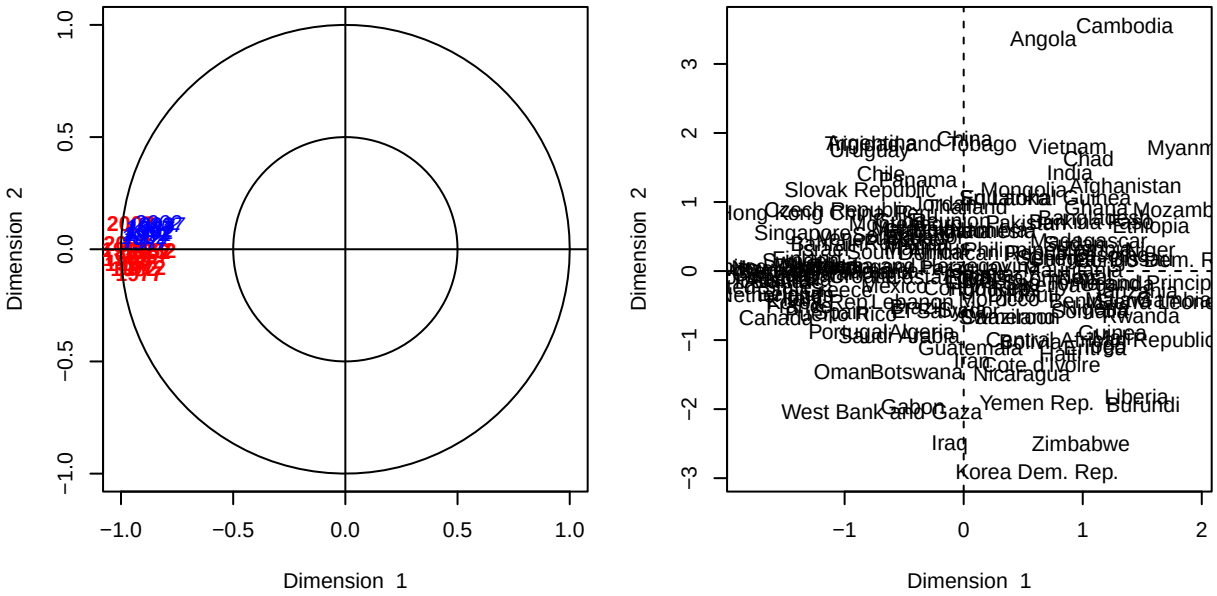
PC2 is suggesting another pattern in data which is orthogonal to the trend captured by PC1, this could represent fluctuations or deviations from the overall trend of life expectancy.

2) Plot 2: This is a scatter plot of the first two Principal Components for Life Expectancy

This shows that most observations are located around (PC1,PC2) = (-20000,0)

Population

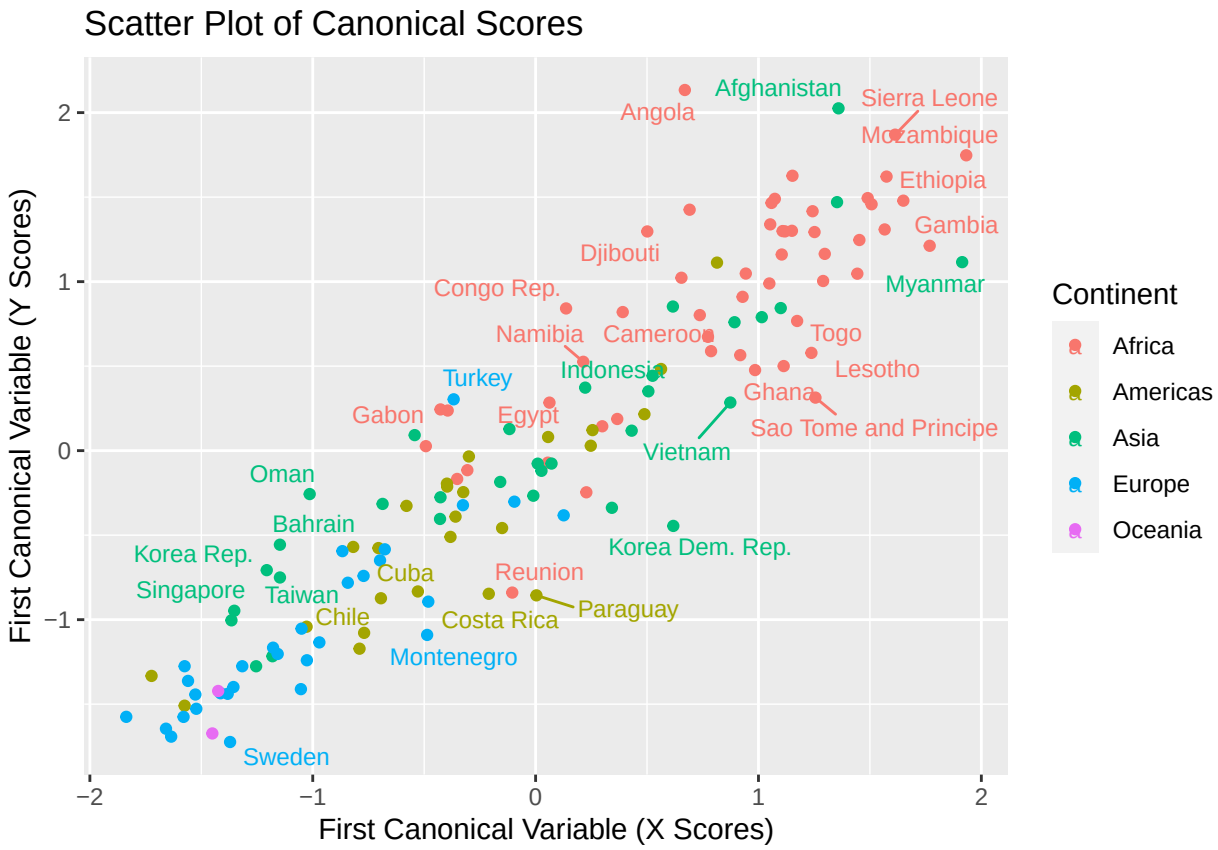
Canonical correlation analysis



1) Left plot:

2) Right plot: the plot of cca*scores*xscores first,second dimension

```
## Warning: ggrepel: 108 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



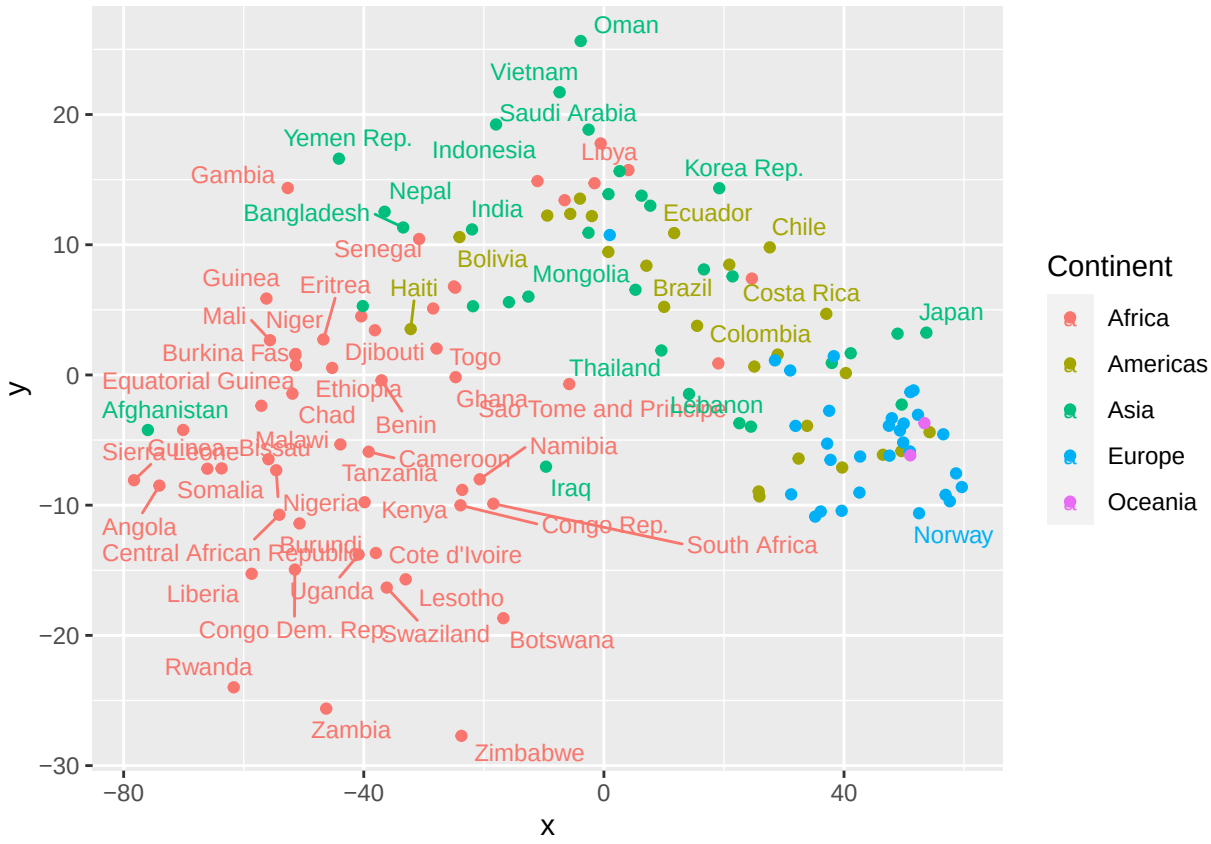
$\phi = -0.1(1952) - 0.2(1957) - 0.003(1962) + 0.26(1967) - 0.24(1972) + 0.06(1977) + 0.12(1982) - 0.16(1987) + 0.04(1992) - 0.05(1997) - 0.02(2002) + 0.03(2007)$

The Canonical correlation analysis transforms the original data into new ones that are maximally correlated.

The plot shows the scatter plot of the first pair of CC variables, this shows that there's high correlation between η_1 and ϕ_1, where η is the first set of transformed data and ϕ is the second set of transformed data.

Multidimensional scaling

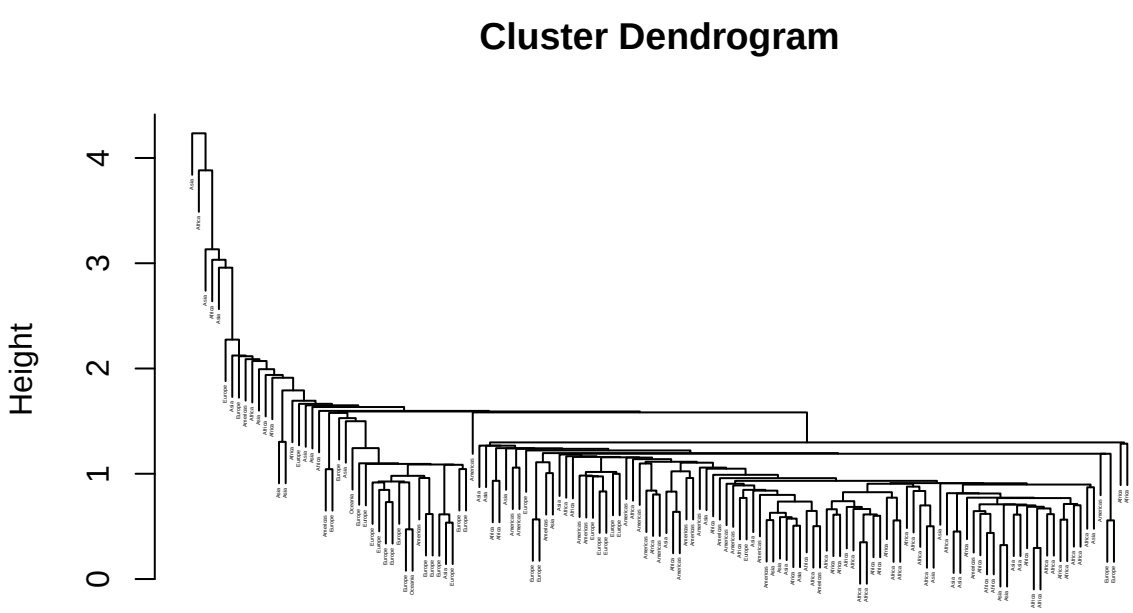
```
## Warning: ggrepel: 78 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Linear discriminant analysis

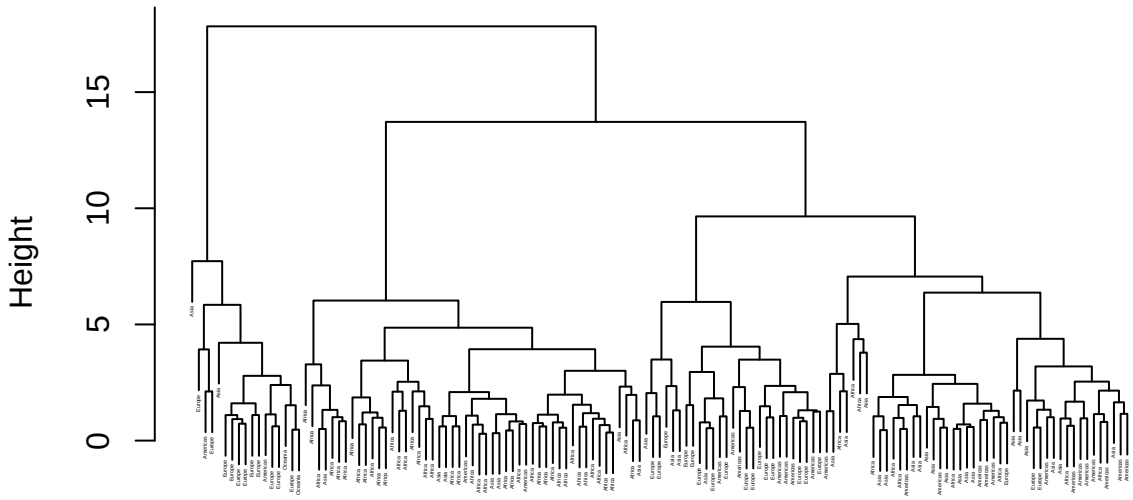
```
## [1] "The predictive accuracy is  60 %"
```

Clustering



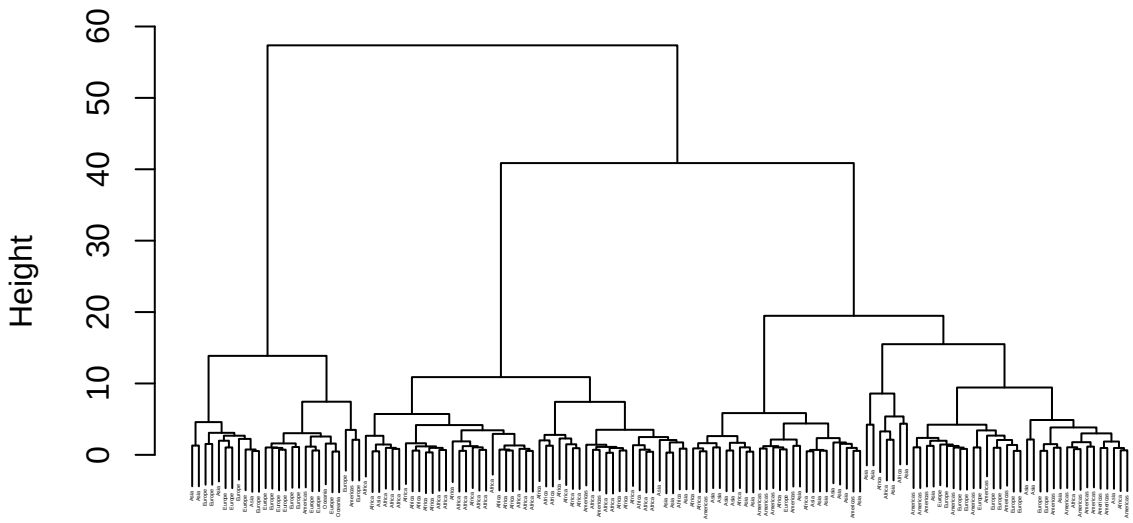
dist(UN.scaled[, 3:26], method = "euclidean")
hclust (*, "single")

```
##
## 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
```

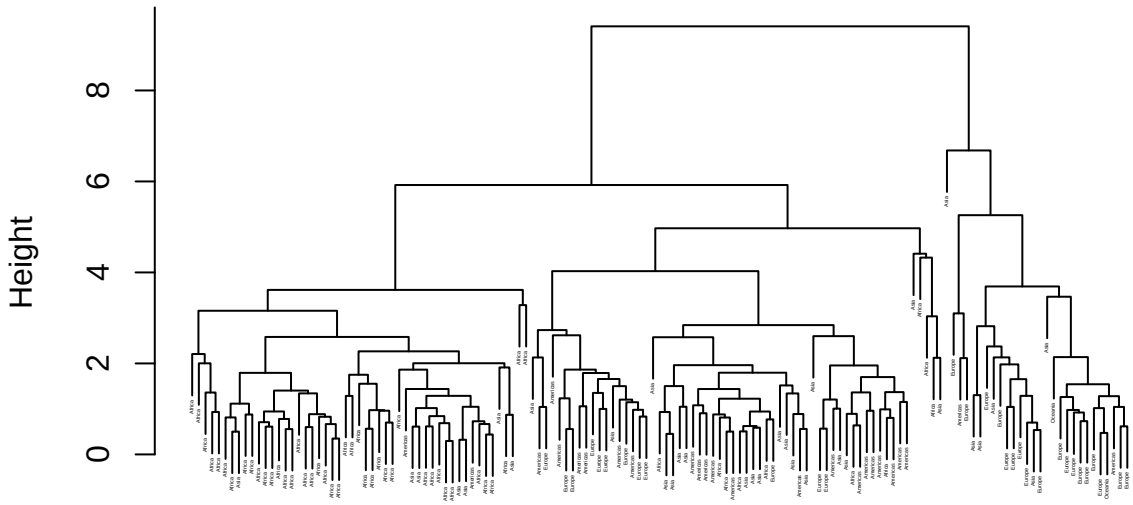


dist(UN.scaled[, 3:26], method = "euclidean")
hclust (*, "complete")

```
##
## 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]
```

```r
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
gdp.pca <- prcomp(gdp)
le.pca <- prcomp(lifeExp)
popn.pca <- prcomp(popn)
#summary(gdp.pca)
#gdp.pca$rotation # the loadings/eigenvectors
#gdp.pca$center  # the sample mean
library(factoextra)
#fviz_eig(gdp.pca, addlabels = TRUE, ylim = c(0, 100)) #Scree plot

plot <- fviz_eig(gdp.pca, addlabels = TRUE, ylim = c(0, 100)) + ggtitle("Plot 1")+ theme(plot.title = el
print(plot)

#fviz(gdp.pca, element='var')#Interpretation of leading PC
plot <- fviz(gdp.pca, element = "var") + ggtitle("Scree plot for GDP.PCA")+ theme(plot.title = element_
print(plot)

plot(gdp.pca$x[,1], gdp.pca$x[,2],
     xlab = "PC1", ylab = "PC2",
     main = "Plot2",
     pch = 19, col = "blue")
#fviz_eig(le.pca, addlabels = TRUE, ylim = c(0, 100)) #Scree plot

plot <- fviz_eig(le.pca, addlabels = TRUE, ylim = c(0, 100)) + ggtitle("Scree plot for LE.PCA")+ theme(p
print(plot)
fviz(le.pca, element='var') #Interpretation of leading PC

plot(le.pca$x[,1], le.pca$x[,2],
     xlab = "PC1", ylab = "PC2",
     main = "Scatter Plot of the First Two Principal Components for GDP",
     pch = 19, col = "blue")
library(CCA)
library(ggplot2)  # Make sure ggplot2 is loaded
library(ggrepel)

cca<-cc(log(gdp),lifeExp)
plt.cc(cca, var.label=TRUE)
```

```r
# 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 Canonical Variable (X Scores)",
       y = "First Canonical Variable (Y Scores)",
       title = "Scatter Plot of Canonical Scores",
       color = "Continent")
cca$cor # the canonical correlations
cca$xcoef[,1] # the canonical correlation vectors for eta
cca$ycoef[,1]  # the canonical correlation vectors for phi
head(cca$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))
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)
```