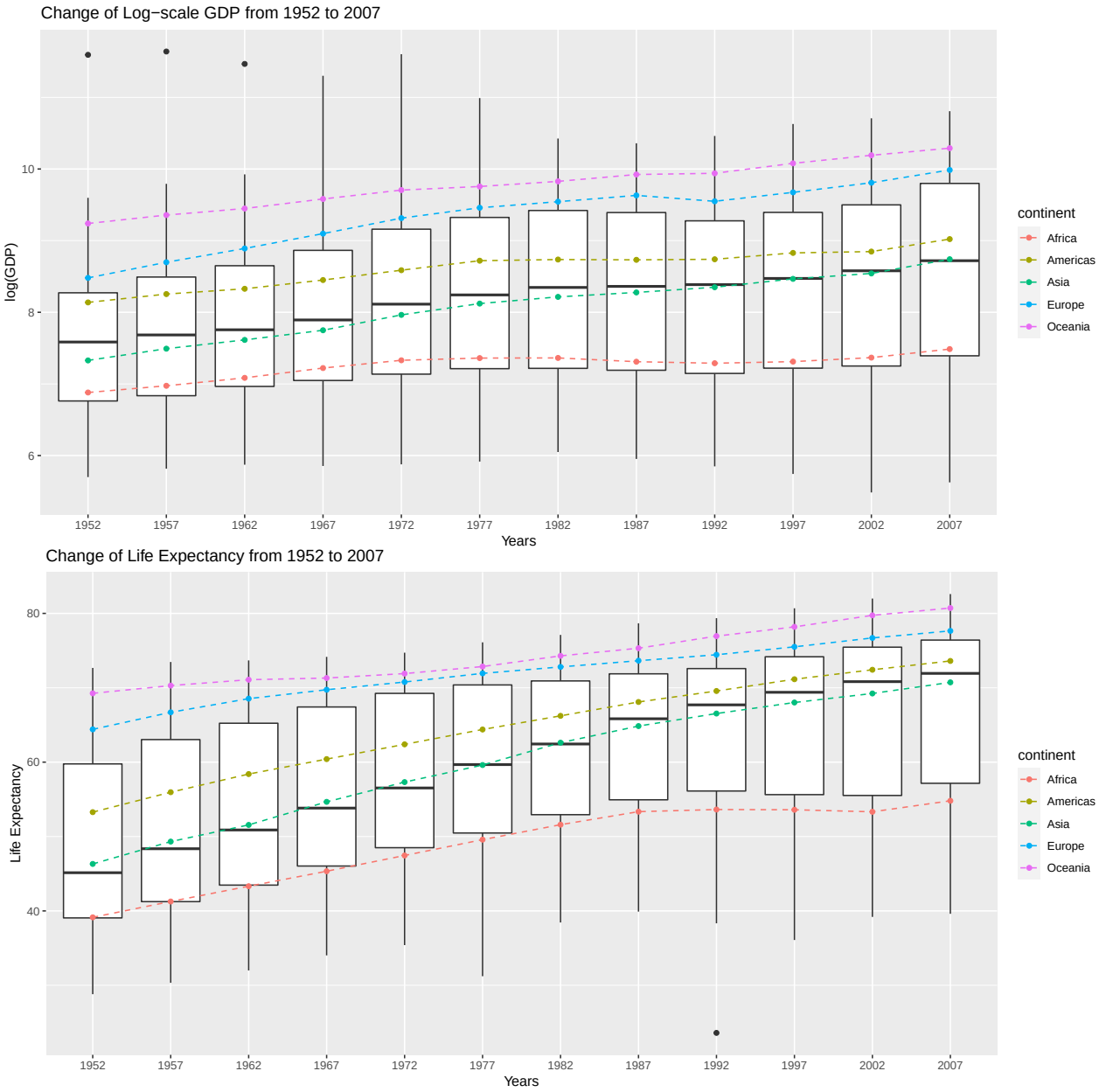
20129999 MATH3030 Coursework

```
# input the data
gap.raw <- read.csv("gap.csv")
gap <- gap.raw
gap[,3:14] <- log(gap.raw[,3:14])
years <- seq(1952, 2007, 5)
```

1) Exploratory Data Analysis

```
# install the packages
library(dplyr)
library(reshape)
library(ggplot2)

# EDA for the GDP data
df1 <- gap %>%
  dplyr::select(c(2:14)) %>%
  `colnames<-` (c("country", years)) %>%
  melt(id = c("country"))
df2 <- aggregate(gap[,3:14], list(gap$continent), FUN = mean) %>%
  `colnames<-` (c("continent", years)) %>%
  melt(id = c("continent"))
# plot
ggplot() +
  geom_boxplot(df1, mapping = aes(variable, value)) +
  geom_line(df2, mapping = aes(variable, value, group = continent, color = continent),
            linetype = "dashed") +
  geom_point(df2, mapping = aes(variable, value, color = continent)) +
  ggtitle("Change of Log-scale GDP from 1952 to 2007") + xlab("Years") + ylab("log(GDP)")

# EDA for the life expectancy data
df3 <- gap %>%
  dplyr::select(c(2, 15:26)) %>%
  `colnames<-` (c("country", years)) %>%
  melt(id=c("country"))
df4 <- aggregate(gap[,15:26], list(gap$continent), FUN = mean) %>%
  `colnames<-` (c("continent", years)) %>%
  melt(id=c("continent"))
# plot
ggplot() +
  geom_boxplot(df3, mapping = aes(variable, value)) +
  geom_line(df4, mapping = aes(variable, value, group = continent, color = continent),
            linetype = "dashed") +
  geom_point(df4, mapping = aes(variable, value, color = continent)) +
  ggtitle("Change of Life Expectancy from 1952 to 2007") +
  xlab("Years") + ylab("Life Expectancy")
```



The boxplots present the spread of the log-scale GDP and the life expectancy of all countries at every five year from 1952 to 2007; the lineplots present the average log-scale GDP and the life expectancy of each continent over time. As we can see, first, there were obvious upward trends regarding to both the GDP and the life expectancy over years. Second, we notice that the Oceanian countries had the highest average GDP and life expectancy among five continents; contrarily, the African countries had the lowest values of these two indexes. Finally, there might be a positive relationship between the GDP and the life expectancy as both of them increased from 1952 to 2007.

2) Principal Component Analysis

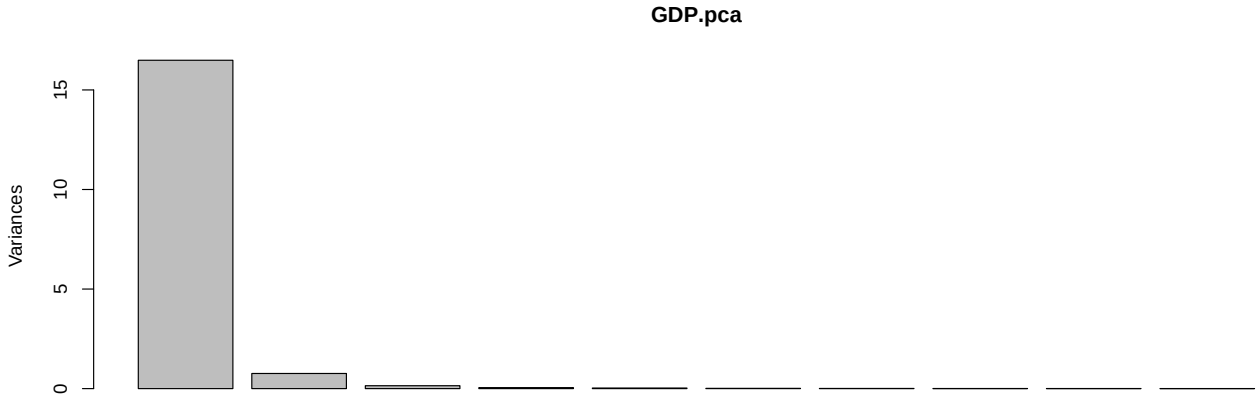
a)

We carry out PCA based on the covariance matrix since the variables present the same type of quantity.

```r
# PCA of log(GDP)
gdp <- gap[,3:14] %>%
  `colnames<-` (years) %>%
  `rownames<-` (gap[,2])
GDP.pca <- prcomp(gdp)

#PCA of Life Expectancy
lifeExp <- gap[,15:26] %>%
  `colnames<-` (years) %>%
  `rownames<-` (gap[,2])
LE.pca <- prcomp(lifeExp)
```

b)

```r
# the srcee plot
plot(GDP.pca)
```

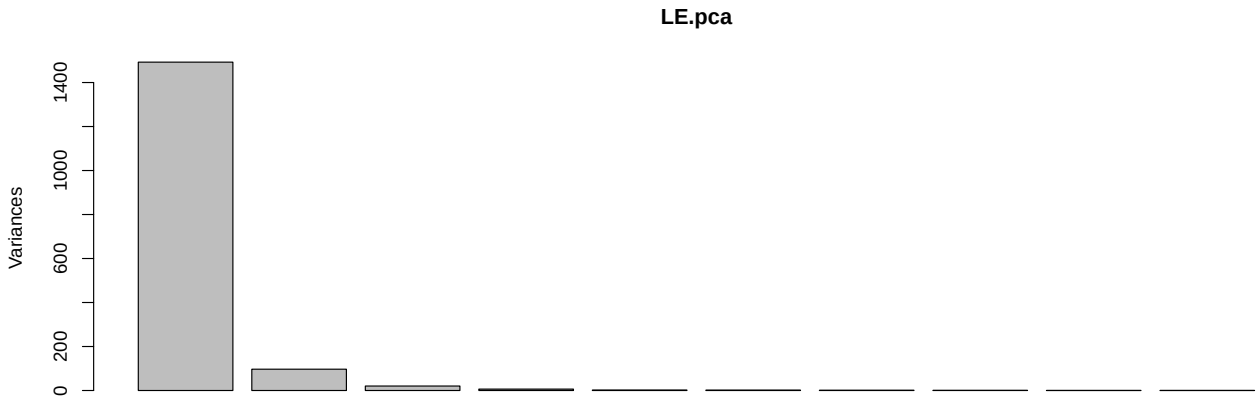



```r
# the proportion of variance
summary(GDP.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     4.0608 0.8718 0.38139 0.2215 0.16930 0.11366 0.09388
## Proportion of Variance 0.9416 0.0434 0.00831 0.0028 0.00164 0.00074 0.00050
## Cumulative Proportion  0.9416 0.9850 0.99328 0.9961 0.99771 0.99845 0.99895
##                           PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.07325 0.06644 0.05813 0.05649 0.04419
## Proportion of Variance 0.00031 0.00025 0.00019 0.00018 0.00011
## Cumulative Proportion  0.99926 0.99951 0.99971 0.99989 1.00000
```

The first principal component of the log-scale GDP data is capable to explain 94% of variance and the first three principal components can explain up to 99% of variance. We choose to retain the first two principal components which are informative enough.

```r
# the scree plot
plot(LE.pca)
```




```
# the proportion of variance
summary(LE.pca)
```

```
## Importance of components:
##                             PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     38.6350 9.83955 4.50873 2.52549 1.39085 1.27733 1.01896
## Proportion of Variance  0.9203 0.05969 0.01253 0.00393 0.00119 0.00101 0.00064
## Cumulative Proportion   0.9203 0.97994 0.99247 0.99641 0.99760 0.99860 0.99924
##                             PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.79691 0.48377 0.43319 0.33836 0.23486
## Proportion of Variance 0.00039 0.00014 0.00012 0.00007 0.00003
## Cumulative Proportion  0.99964 0.99978 0.99990 0.99997 1.00000
```

The first principal component of the life expectancy data is able to explain 92% of variance and the first three principal components can explain up to 99% of variance. Again, we choose to retain the first two principal components which are fairly informative.

c)

```
# the first loading of GDP.pca
GDP.pca$rotation[,1]
```

```
##       1952       1957       1962       1967       1972       1977       1982
## -0.2374495 -0.2476188 -0.2565166 -0.2689206 -0.2851636 -0.2940603 -0.2961380
##       1987       1992       1997       2002       2007
## -0.3049343 -0.3092904 -0.3144214 -0.3158196 -0.3185089
```

The first principal component measures the overall negative log(GDP) which puts a bit more weights on the recent variables. Thus, low values of PC1 indicate the country had a higher GDP and high values present a lower GDP.

```
# the second loading of GDP.pca
GDP.pca$rotation[,2]
```

```
##        1952        1957        1962        1967        1972        1977
##  0.37554442  0.36407931  0.33016476  0.28195358  0.22398302  0.12145189
##        1982        1987        1992        1997        2002        2007
##  0.01868930 -0.08938367 -0.20505847 -0.31986881 -0.38510113 -0.41470527
```

The second principal component measures the change of the log-scale GDP over time. It puts positive weights on the early variables and reduces to negative weights on the recent variables. As a result, low values of PC2 indicate the GDP of the country increased a lot in the past 55 years and high values present a decrease in the GDP.

```
# the first loading of LEP.pca
LE.pca$rotation[,1]
```

```
##      1952      1957      1962      1967      1972      1977      1982      1987
## 0.2995501 0.3041007 0.3024818 0.2967688 0.2898377 0.2850260 0.2751025 0.2681294
##      1992      1997      2002      2007
## 0.2777391 0.2835186 0.2937106 0.2856875
```

The first principal component measures the overall life expectancy in the past 55 years so, in general, high values of PC1 indicate the country had a longer life expectancy and the low values present a short life expectancy.

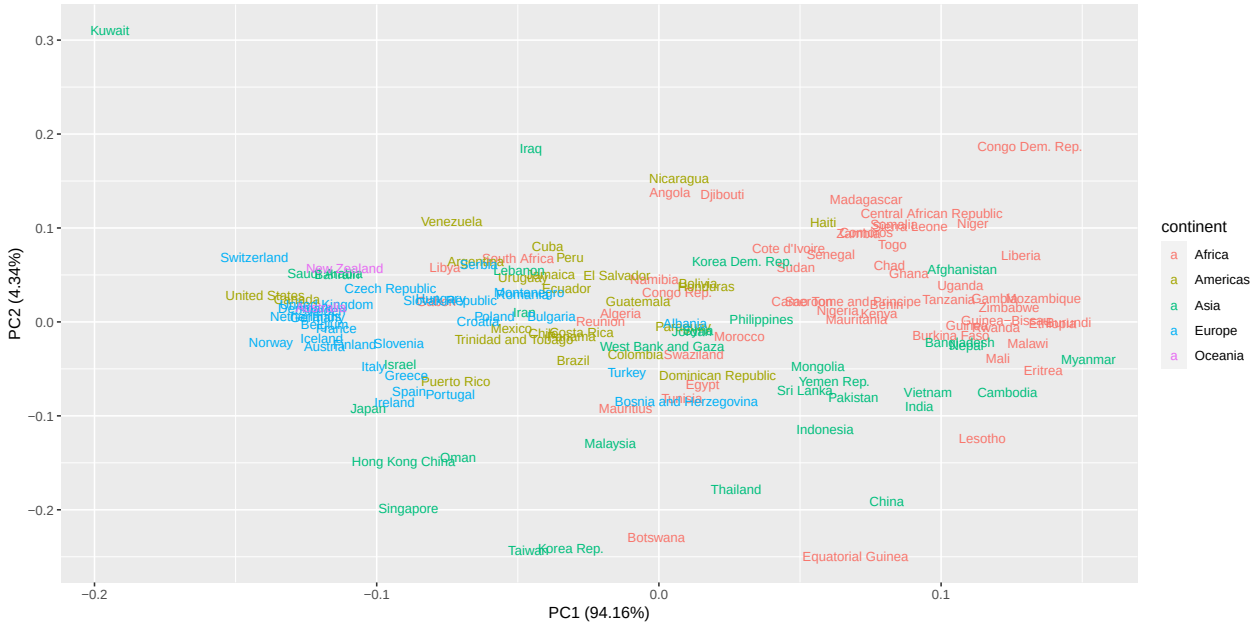
```
# the second loading of LEP.pca
LE.pca$rotation[,2]
```

```
##        1952        1957        1962        1967        1972        1977
##  0.34902226  0.32308013  0.29646127  0.23213961  0.16609597  0.10290633
##        1982        1987        1992        1997        2002        2007
##  0.02102816 -0.06709016 -0.22660990 -0.35942503 -0.44973534 -0.45398620
```

The second principal component measures the change of the life expectancy. It puts positive weights on the early variables and reduces to negative weights on the recent variables. Namely, low values of PC2 indicate there was an increase in the life expectancy of the country from 1952 to 2007; high values present a decrease in the life expectancy.

d)

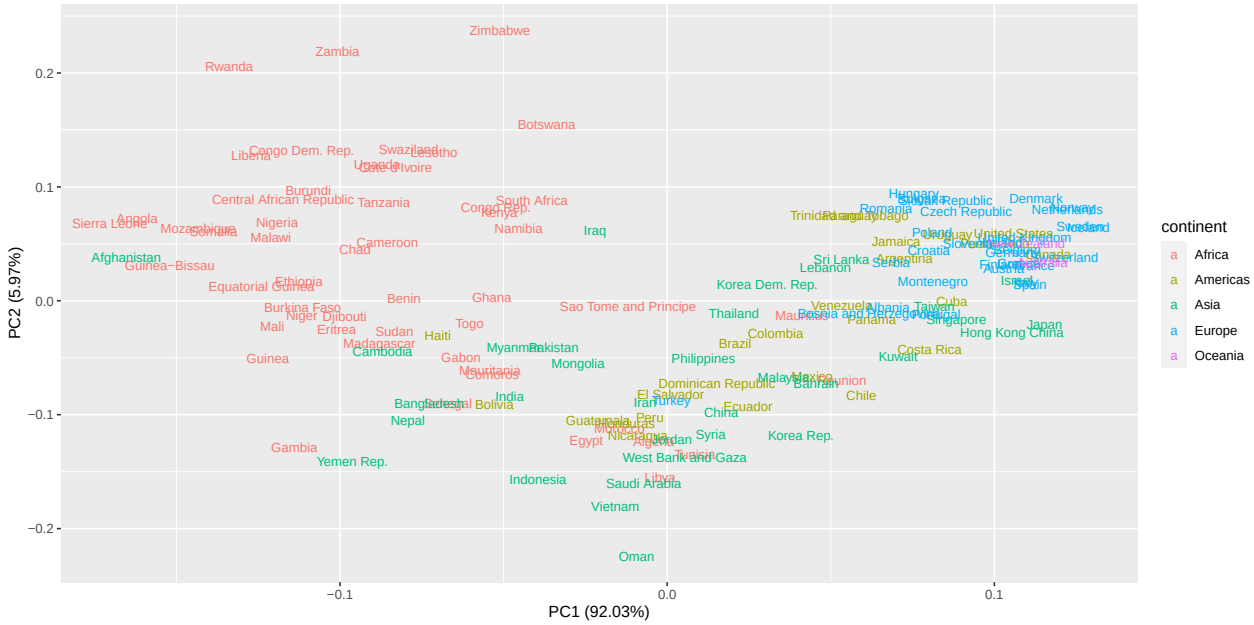
```
# install the package
library(ggfortify)

# plot
autoplot(GDP.pca, data = gap.raw, label = TRUE, label.size = 3, shape = FALSE, colour = "continent")
```



We notice that Kuwait has the lowest PC1 score but the highest PC2 score which may present that Kuwait had a higher GDP compared to other countries in the past 55 years but its GDP was decreasing. Indeed, we

can see from the data that the GDP of Kuwait was 108382.3529 in 1952. Although it dropped to 47306.9898 in 2007, it was still one of the highest in the world.

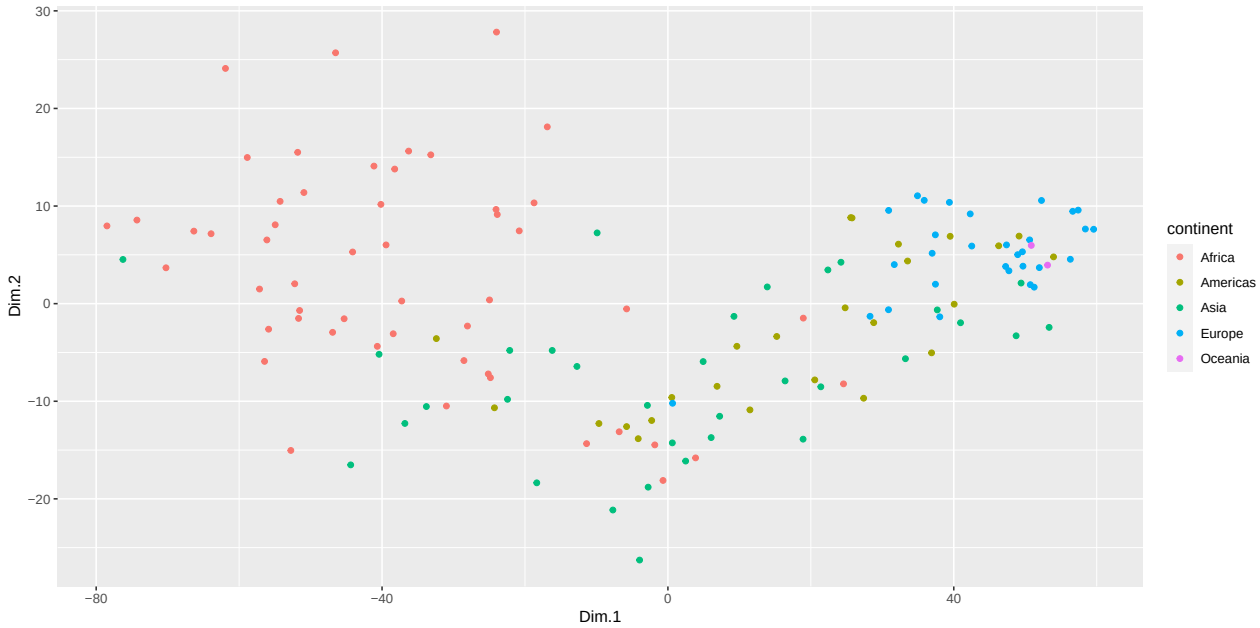
```
# plot
autoplot(LE.pca, data = gap.raw, label = TRUE, label.size = 3, shape = FALSE, colour = "continent")
```



The country, Omen, has a PC1 score of approximate zero but the lowest PC2 score which indicates there was a great improvement in the life expectancy in Omen. The life expectancy there was 37.578 in 1952 and it increased to 75.64 by 2007.

3) Multidimensional Scaling

```
# MDS of gap
gap.mds <- gap[,3:26] %>%
  dist() %>%
  cmdscale() %>%
  data.frame() %>%
  `colnames<-` (c("Dim.1","Dim.2")) %>%
  mutate(continent = gap[,1])

# plot
ggplot() +
  geom_point(gap.mds, mapping = aes(Dim.1, Dim.2, color = continent))
```



This plot is as same as the previous plot of PCA on the life expectancy. Most of the African countries are on the top-left corner of the figure; Asian countries are spread in the bottom; American, European and Oceanian countries are in the middle-right part.

4) Hypothesis testing

We first conduct a multivariate hypothesis test to see whether there was a statistically significant difference between the mean log(GDP) and the life expectancy of Asian and European countries in 2007.

```
# install the package
library(ICSNP)

# log(GDP) and the life expectancy of Asian countries in 2007
Asia.2007 <- gap %>%
  filter(continent == "Asia") %>%
  dplyr::select(c(14,26))

# log(GDP) and the life expectancy of European countries in 2007
European.2007 <- gap %>%
  filter(continent == "Europe") %>%
  dplyr::select(c(14,26))
```

Let $\mathbf{x}_1, ..., \mathbf{x}_{33}$ be the log-scale GDP and the life expectancy of Asian countries in 2007 and let $\mathbf{y}_1, ..., \mathbf{y}_{30}$ be the log-scale GDP and the life expectancy of European countries in 2007.

Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be the population means for Asian and European data respectively. Our hypotheses are

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

We will assume that

$$\mathbf{x}_1, ..., \mathbf{x}_{33} \sim N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$
$$\mathbf{y}_1, ..., \mathbf{y}_{30} \sim N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

```
HotellingsT2(Asia.2007, European.2007)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  Asia.2007 and European.2007
## T.2 = 12.681, df1 = 2, df2 = 60, p-value = 2.55e-05
## alternative hypothesis: true location difference is not equal to c(0,0)
```

The p-value, 0.0000255, is less than 0.05. This suggests us to reject the null hypothesis that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Namely, there was a significant difference between the mean log-scale GDP and the life expectancy of Asian and European countries in 2007.

We now check if the continents were similar in 1952.

```
# log(GDP) and the life expectancy of Asian countries in 1952
Asia.1952 <- gap %>%
  filter(continent == "Asia") %>%
  dplyr::select(c(3,15))

# log(GDP) and the life expectancy of European countries in 1952
European.1952 <- gap %>%
  filter(continent == "Europe") %>%
  dplyr::select(c(3,15))
```

Let $\mathbf{x}_1, ..., \mathbf{x}_{33}$ be the log-scale GDP and the life expectancy of Asian countries in 1952 and let $\mathbf{y}_1, ..., \mathbf{y}_{30}$ be the log-scale GDP and the life expectancy of European countries in 1952

Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be the population means for Asian and European data respectively. Our hypotheses are

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

We will assume that

$$\mathbf{x}_1, ..., \mathbf{x}_{33} \sim N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$
$$\mathbf{y}_1, ..., \mathbf{y}_{30} \sim N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

```
HotellingsT2(Asia.1952, European.1952)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  Asia.1952 and European.1952
## T.2 = 39.347, df1 = 2, df2 = 60, p-value = 1.21e-11
## alternative hypothesis: true location difference is not equal to c(0,0)
```

The p-value, 1.21×10^{-11}, is less than 0.05. Again, this suggests us to reject the null hypothesis that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Thus, there was a difference between the mean log-scale GDP and the life expectancy of Asian and European countries in 1952. However, The p-value of this test based on the data of 1952 is less than the p-value of the previous test based on the data of 2007. This indicates that the difference of the mean log-scale GDP and the life expectancy of these two continents is actually becoming smaller.

5) Linear Discriminant Analysis

a)

```
# install the package
library(MASS)
```

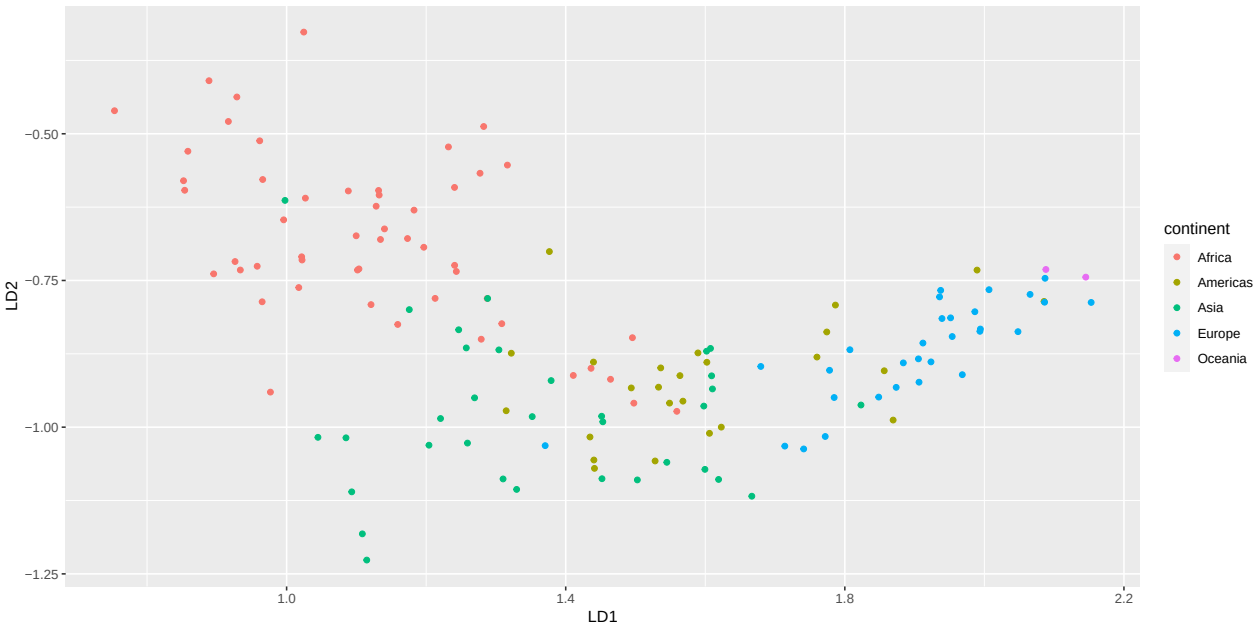
```r
# set the random seed
set.seed(1357)

# remove "country"
gap.Q4 <- gap %>% dplyr::select(!c(2))

# split the data into the training set and the testing set
test.ind <- sample(142, size = 71)
gap.test <- gap.Q4[test.ind,]
gap.train <- gap.Q4[-test.ind,]
```

```r
# LDA of the gap training dataset
gap.lda <- lda(continent ~ ., data = gap.train)
# prediction
gap.pred <- predict(gap.lda, gap.test)
# print the result
print(paste("The predictive accuracy is ",
            round(sum(gap.pred$class == gap.test$continent)/dim(gap.test)[1]*100, 2), "%"))
```

```
## [1] "The predictive accuracy is  71.83 %"
```

b)

```r
# install the package
library(vcvComp)

# between-class variance
B <- cov.B(gap[,3:26], gap[,1])
# within-class variance
W <- cov.W(gap[,3:26], gap[,1])
# eigenvalue/vector of W^-1*B
gap.eig <- eigen(solve(W)%*%B)
# take the first two eigenvectors
V <- gap.eig$vectors[,1:2]
# 2-D fisher discriminant
Z <- data.frame(Re(as.matrix(gap[,3:26])%*%V)) %>%
  mutate(continent = gap[,1])

# plot
ggplot() +
  geom_point(Z, mapping = aes(X1, X2, color = continent)) + xlab("LD1") + ylab("LD2")
```

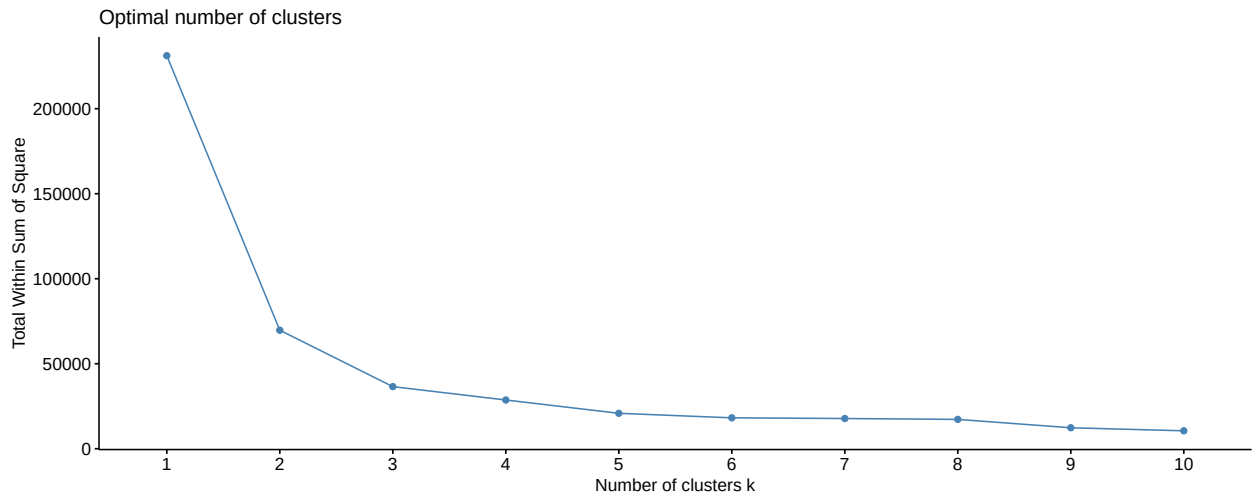


This plot is as same as the previous plot of PCA on the life expectancy. The first projected coordinate is able to separate Africa, America and Europe and the second projected coordinate is doing the work in separating the Asia from Africa.

6) Clustering

a)

```
# install the package
library(factoextra)

# prepare the data
gap.c <- gap[,3:26] %>%
  `rownames<-` (gap[,2])

# elbow plot
fviz_nbclust(gap.c, kmeans, method = "wss")
```



We use the elbow method to determine the optimal number of clusters. As shown, since there is a reasonable decrease in the total within sum of square when moving from two to three clusters but moving to four clusters only yields a minor improvement, we decide to choose three clusters.

```
# K-means of gap
gap.k <- kmeans(gap.c, centers = 3, nstart = 20)
# plot
fviz_cluster(gap.k, data = gap.c, pointsize = 1, labelsize = 8)
```

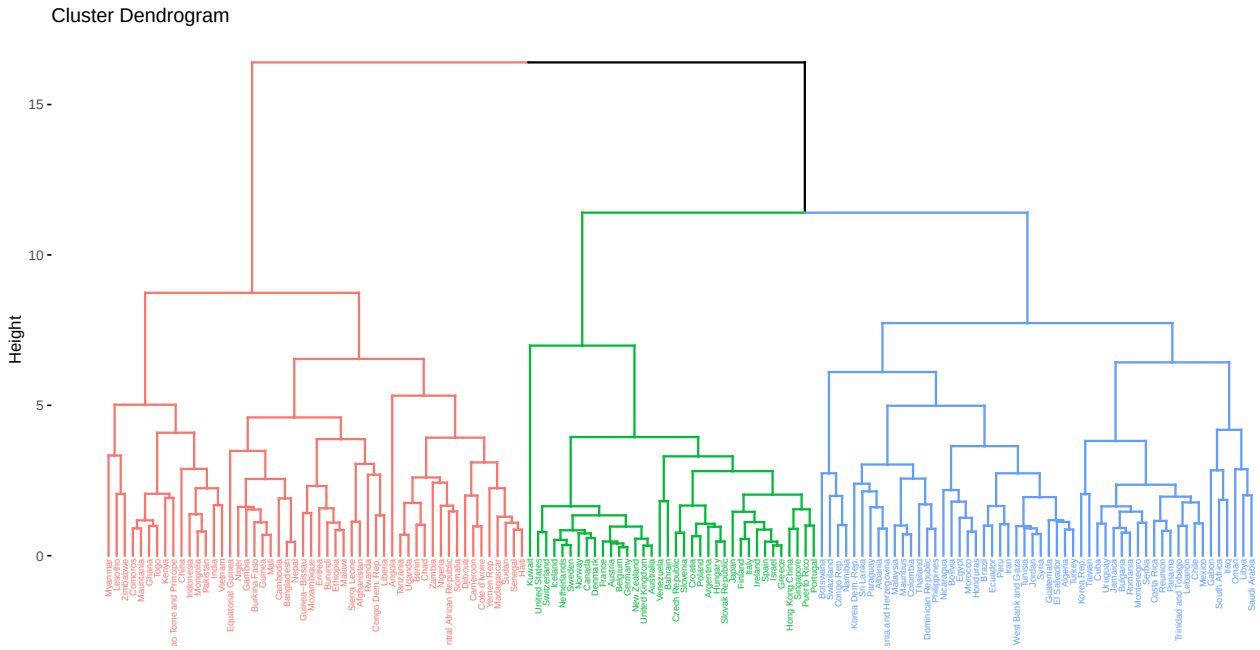
b)

```
# scaling the data
gap.scaled <- scale(gap.c)

# distance matrix
d <- dist(gap.scaled, method = "euclidean")
```

```
# single linkage
D.sl <- hclust(d, method = "single")

# complete linkage
D.co <- hclust(d, method = "complete")

# group linkage
D.ga <- hclust(d, method = "average")
```

After trying the single, complete and group average methods, we choose to retain the result using complete linkage.

```
# dendrogram of complete method
fviz_dend(D.co, cex = 0.4, k = 3, color_labels_by_k = TRUE)
```



c)

```
# confusion table
table(kmeans = gap.k$cluster, Hierarch = cutree(D.co, 3))
```

```
##       Hierarch
## kmeans  1  2  3
##      1 33 12  0
##      2  1 40  0
##      3 20  0 36
```

As shown, there are 109 countries classified into the same cluster.

```
# continent data
continent <- data.frame(gap[,1], row.names = gap[,2]) %>%
  `colnames<-` ("continent")
# kmean
gap.kmean <- merge(data.frame(gap.k$cluster), continent, by = 0)
# hierarch
gap.hierarch <- merge(data.frame(cutree(D.co, 3)), continent, by = 0)
```

```
# table of gap.kmean
table(kmeans = gap.kmean$gap.k.cluster, Continent = gap.kmean$continent)
```

```
##       Continent
## kmeans Africa Americas Asia Europe Oceania
##      1     15       10   19      1       0
##      2     35        1    5      0       0
##      3      2       14    9     29       2
```

```
# table of gap.hierarch
table (Hierarch = gap.hierarch$cutree.D.co..3., Continent = gap.hierarch$continent)
```

```
##         Continent
## Hierarch Africa Americas Asia Europe Oceania
```

```
##         1     13        19   15      7        0
##         2     39         1   12      0        0
##         3      0         5    6     23        2
```

While comparing above two tables, we can see the first and second cluster are similar regardless the methods. There are over ten countries from Africa, Americas and Asia each in the first cluster and majority of the African countries are classified into the second cluster. In the third cluster, the hierarchical clustering method brings Oceanian countries, most of European countries and a few of Asian and American countries together. However, there are more countries from Americas and Asia and even two African countries classified into the third cluster while using the k-mean method.

According to our finding, countries do not cluster by the continents naturally. In the previous PCA, MDS and LDA parts, we hardly can separate Oceanian and European countries and the feature of some Asian and American countries are similar. As clustering is an unsupervised learning, we are not able to differentiate the continents by clustering precisely based on the data we have.

7) Linear Regression

```
# response y: life expectancy data in 2007
y <- as.matrix(gap[,26])
# covariates x: GDP data from 1952 to 2007
x <- as.matrix(exp(gap[,3:14]))
# combine data
gap.lm <- data.frame(y,x)

# split the data into the training set and the testing set
set.seed(2468)
test.ind <- sample(142, size = 71)
gap.lm.test <- gap.lm[test.ind,]
gap.lm.train <- gap.lm[-test.ind,]
```

We first fit the normal linear model with the OLS method.

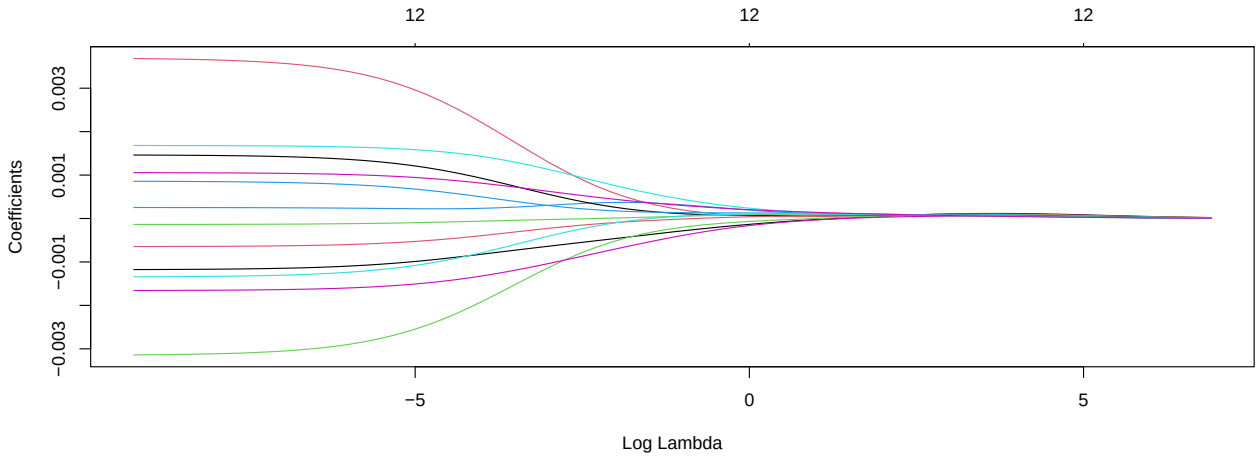
```
# linear model
m1 <- lm(y ~ ., data = gap.lm.train)
# mean squared predicted error
predicted1 <- predict(m1, newdata = gap.lm.test[,-1])
mean((gap.lm.test$y - predicted1)^2)
```

```
## [1] 368.8025
```

The mean square prediction error of this model is 368.8025. In the following part, we fit a ridge regression to see whether we can decrease the mean square prediction error by using the biased estimator.

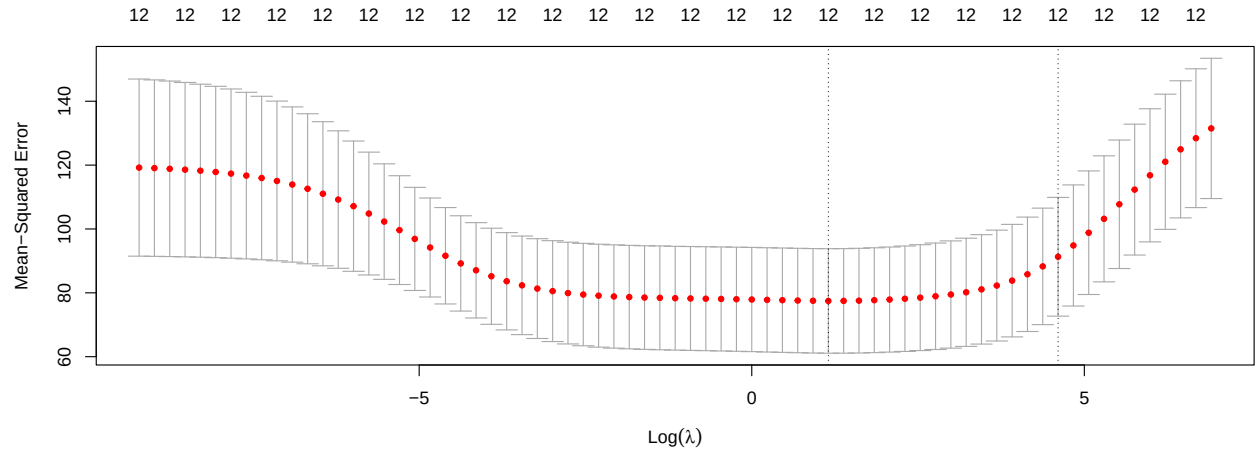
```
# install packages
library(glmnet)

# set lambda
lambdas <- 10^seq(3, -4, by = -0.1)
# ridge regression
lm.rigde <- glmnet(gap.lm.train[,-1], gap.lm.train$y, alpha = 0, lambda = lambdas)
```

```
# parameter estimates change as the size of lambda changes
plot(lm.rigde, xvar = "lambda")
```



We can see that as lambda grows large, the parameters all shrink to zero as expected.

```
cv_fit <- cv.glmnet(as.matrix(gap.lm.train[,-1]), as.matrix(gap.lm.train$y),
                    alpha = 0, lambda = lambdas)
plot(cv_fit)
```




```
# value of lambda that minimises the prediction error
cv_fit$lambda.min
```

```
## [1] 3.162278
```

We find that when the value of lambda is 3.162278, we can minimise the prediction error. We then fit a model with the lambda and access the performance of the model.

```
# fit a ridge regression again
m2 <- glmnet(as.matrix(gap.lm.train[,-1]), as.matrix(gap.lm.train$y),
             alpha = 0, lambda = cv_fit$lambda.min)
# mean squared predicted error
predicted2 <- predict(cv_fit,
                      newx = as.matrix(gap.lm.test[,-1]),
                      s = "lambda.min")
mean((gap.lm.test$y - predicted2)^2)
```

```
## [1] 97.56985
```

As we can see, the mean squared predicted error reduces to 97.56985 which shows that the ridge regression is a better method to fit the dataset.