Multivariate CW

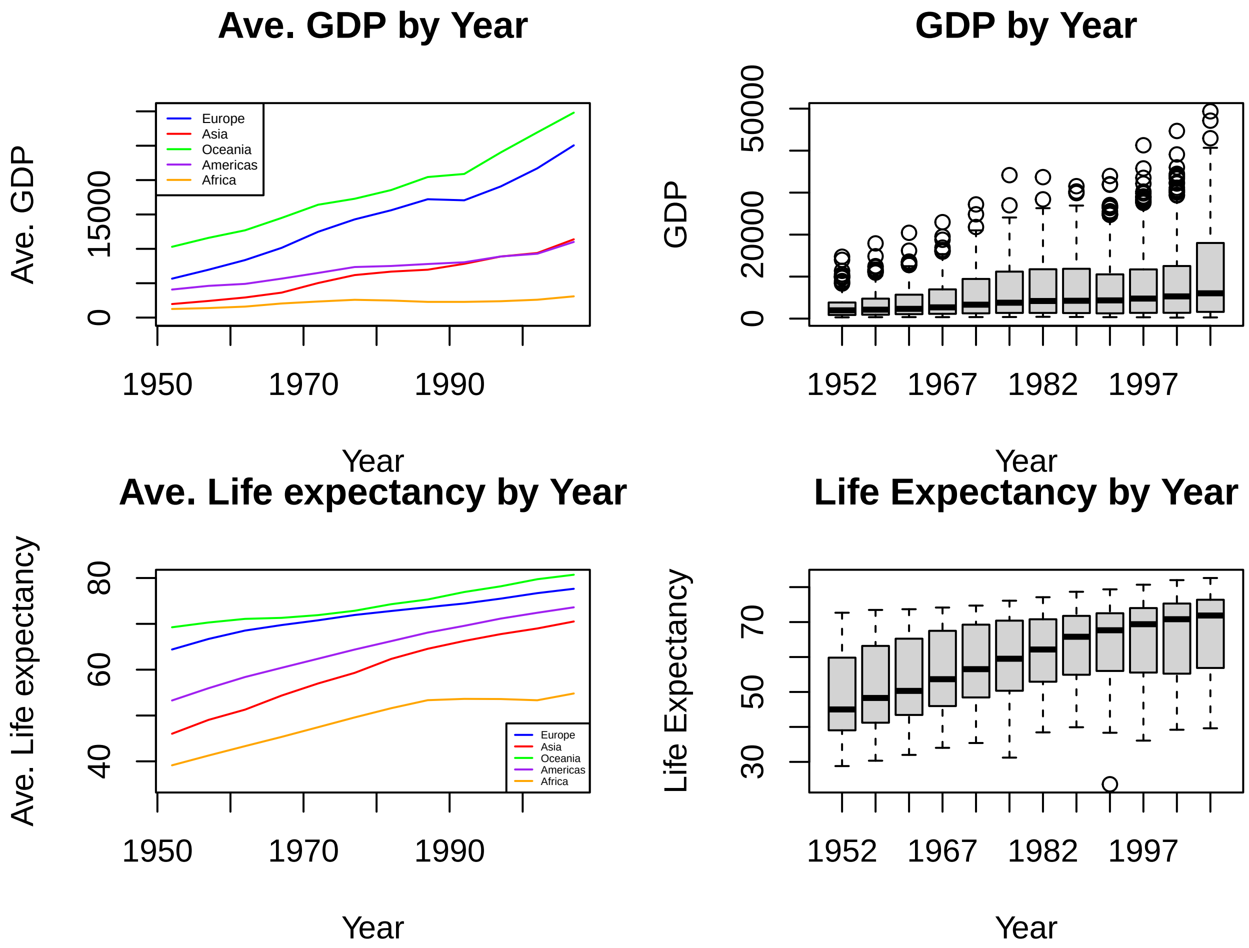
Liangxiao LI

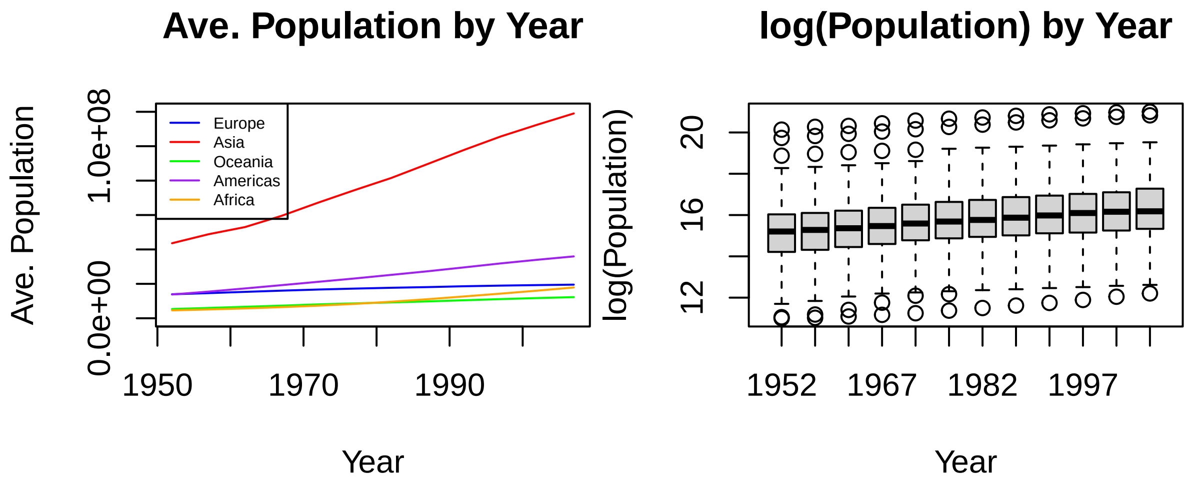
2024-04-13

Part 1: 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.





1) Line plots

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.

For both GDP and Life expectancy, the ranking of continents from highest to lowest is as follows: Oceania > Europe > Americas > Asia > Africa

For population, Asia shows tremendously higher population than other continents, followed by Americas, Europe, Africa and Oceania.

2) Box plots

Here I plot the box plots for gdp,life expectancy and population(without catagorizing by continents).

The box plot again shows the average GDP, life expectancy and population are growing steadily across the globe as years goes by.

The GDP plot shows that there exist lots of outliers among the data due to skewness, therefore we may consider using log(gdp) in later calculation.

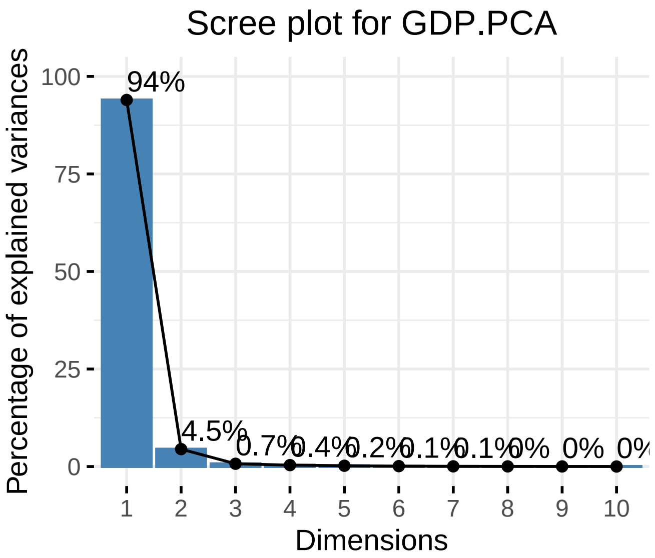
Part 2: Principal component analysis

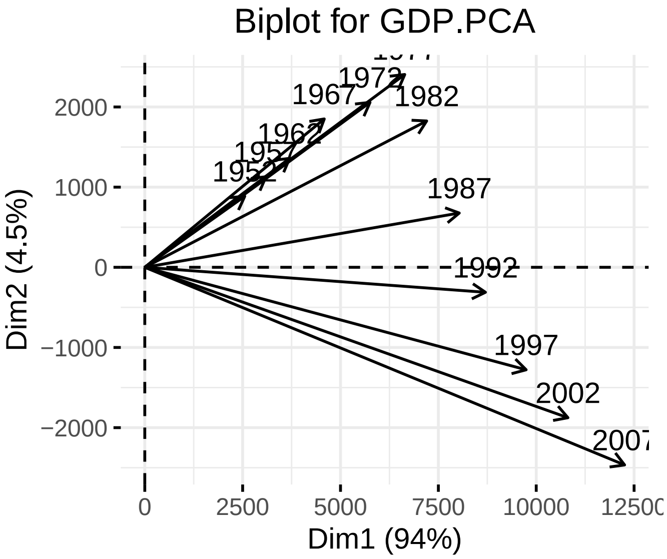
Here for all three different datasets, I perform PCA horizontally, treating each country as a data point and each year as a feature.

Here I perform PCA on three different datasets: gdp, life expectancy and population, where each dataset contains the value in different years. Since each dataset shares same type of data from different years, I decide to perform PCA based on **S(sample covariance matrix)**.

Part 2.1: Number of PCs to retain

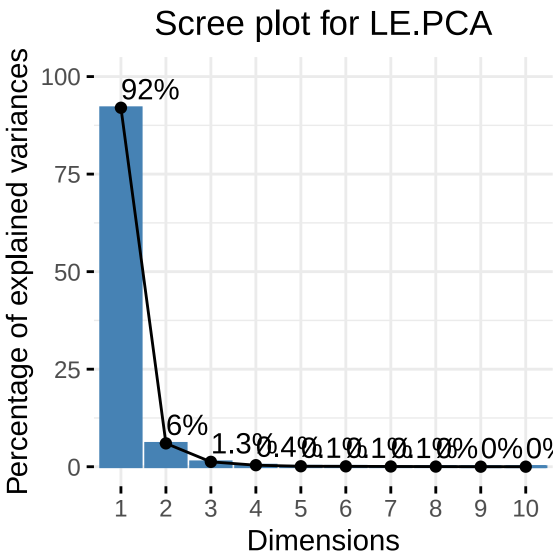
First we plot the scree plot and the corresponding circle plots to decide how many PCs we should retain.

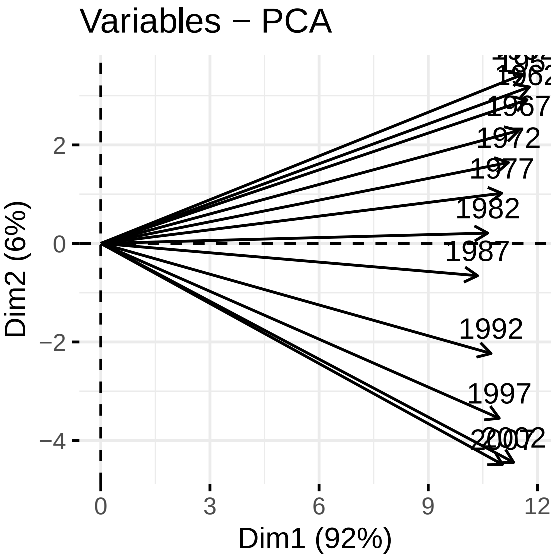




From the scree plot, we'd retain PC1 and PC2 as they explained 94% and 4.5% of the variance within the data. Which covers most of the variability of the data.

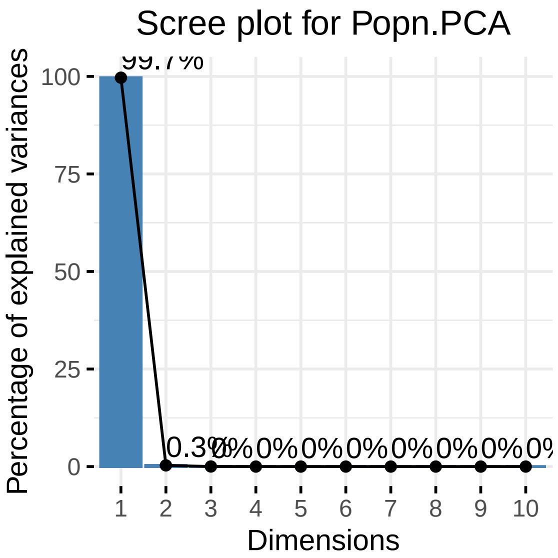
From the covariance plot, we notice that all GDP variables are positively correlated with PC1, column 1952 - 1987 are positively correlated with PC2 and column 1992 - 2007 are negatively correlated with PC2

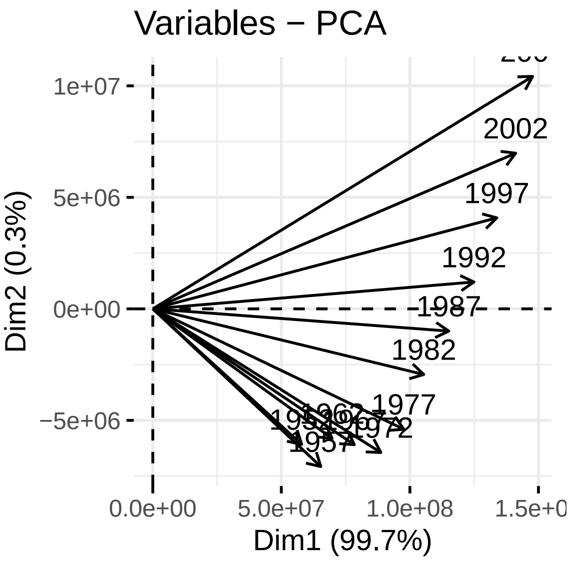




From the scree plot, we'd retain PC1 and PC2 for life expectancy as they explained 92% and 6% of the variance within the data.

From the covariance plot, we notice that all variabels are positively correlated with PC1, column 1952 - 1982 are positively correlated with PC2 and column 1987 - 2007 are negatively correlated with PC2

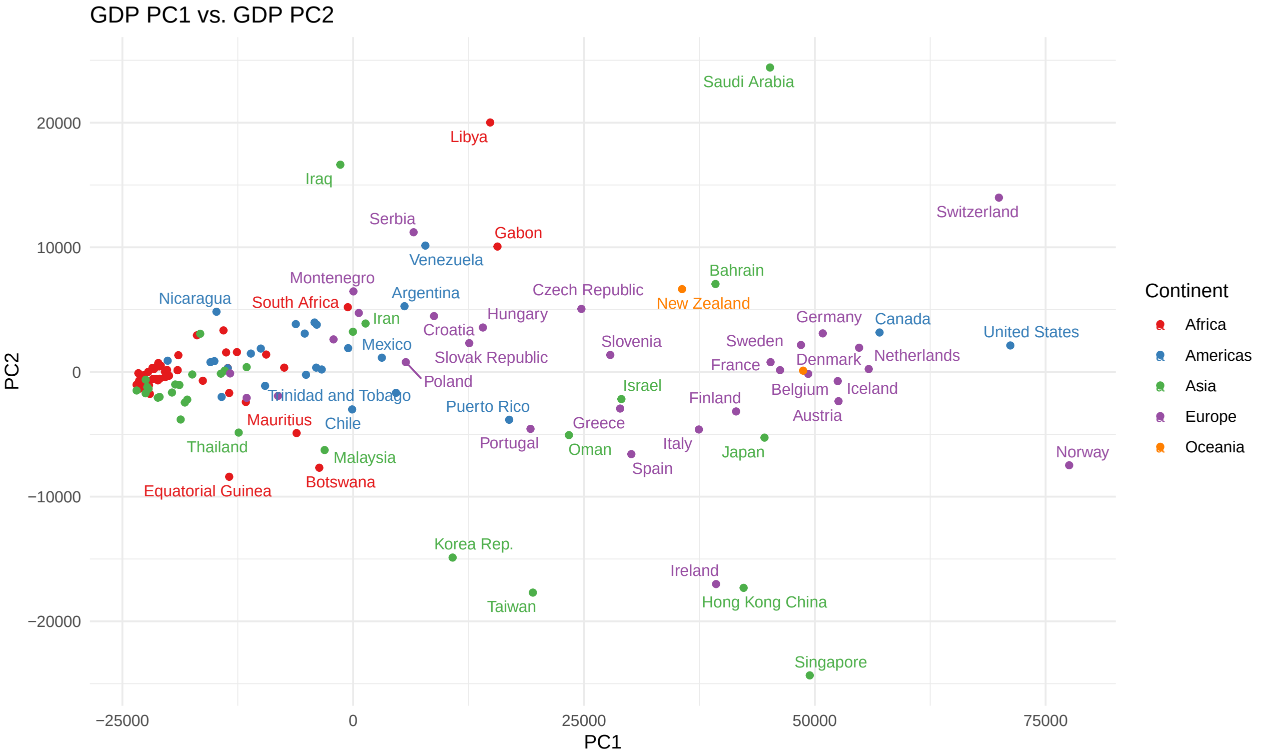




From the scree plot, we'd retain PC1 for population as it explained 99.7% of the variance within the data.

From the covariance plot, we notice that all variabels are positively correlated with PC1, column 1952 - 1987 are negatively correlated with PC2 and column 1992 - 2007 are positively correlated with PC2

Part 2.2: Scatter plots for PCs and interpretations



```
##       1952 1957 1962 1967 1972 1977 1982 1987  1992  1997  2002  2007
## PC1 0.10 0.12 0.14 0.18 0.22 0.25 0.28 0.31  0.33  0.37  0.41  0.47
## PC2 0.16 0.20 0.24 0.32 0.36 0.42 0.32 0.12 -0.05 -0.22 -0.33 -0.43
```

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 positive). Higher PC1 indicate higher overall GDP.

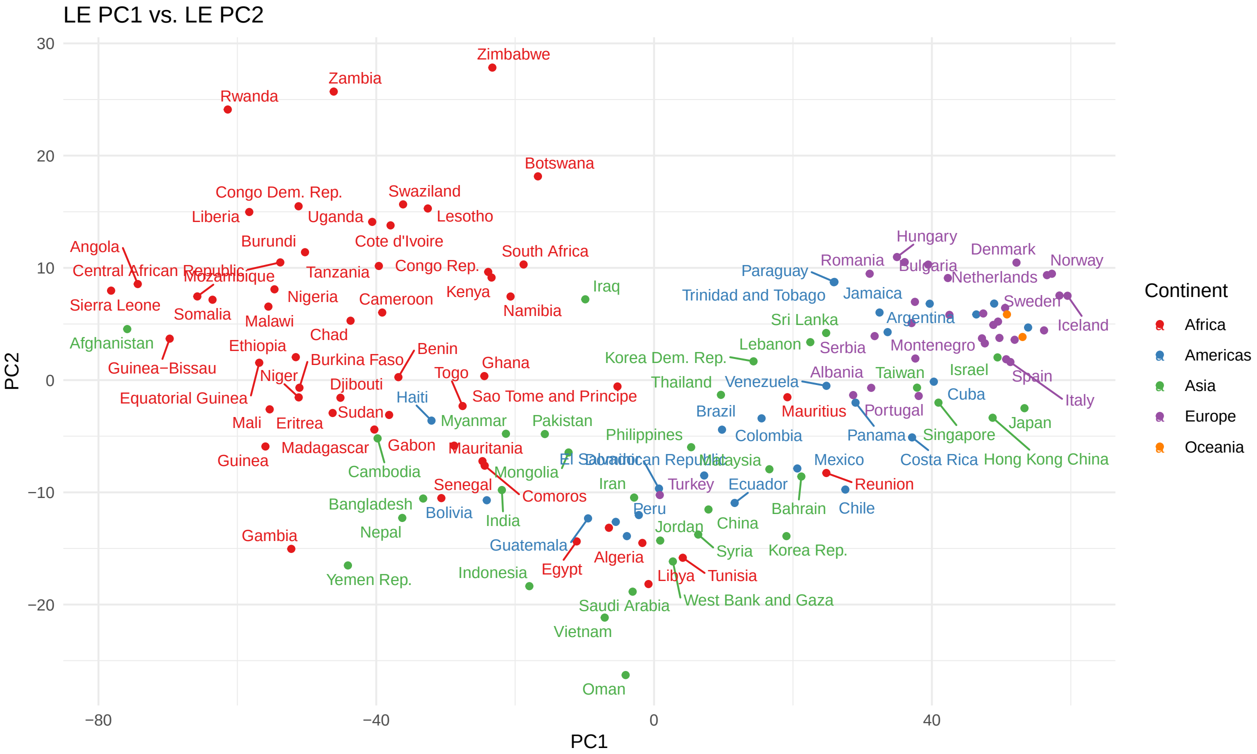
Continents interpretation: **European** countries have generally higher GDP while **African** countries have generally lower GDP.

Countries interpretation: Switzerland/United States has generally the highest GDP due to high PC1, meaning they have generally highest GDP around the globe.

b) **PC2** Might represent a cyclic variation such as economic fluctuations. As higher PC2 indicate higher GDP between 1952-1987, while lower PC2 indicate higher GDP between 1992-2007.

Continents interpretation: **Asia** countries tend to have higher GDP after 1992 due to its low PC2.

Countries interpretation: Singapore's extremely low PC2 means it started its economic growth after 1992 and thrives in 2002-2007.



```
##      1952 1957 1962 1967 1972 1977 1982  1987  1992  1997  2002  2007
## PC1 0.30 0.30  0.3 0.30 0.29 0.28 0.27  0.27  0.28  0.28  0.29  0.29
## PC2 0.35 0.32  0.3 0.23 0.17 0.10 0.02 -0.07 -0.23 -0.36 -0.45 -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 ~ +0.3). Higher PC1 indicate higher overall life expectancy.

Continents interpretation: **European** countries have the highest overall life expectancy while **African** countries have the lowest overall life expectancy.

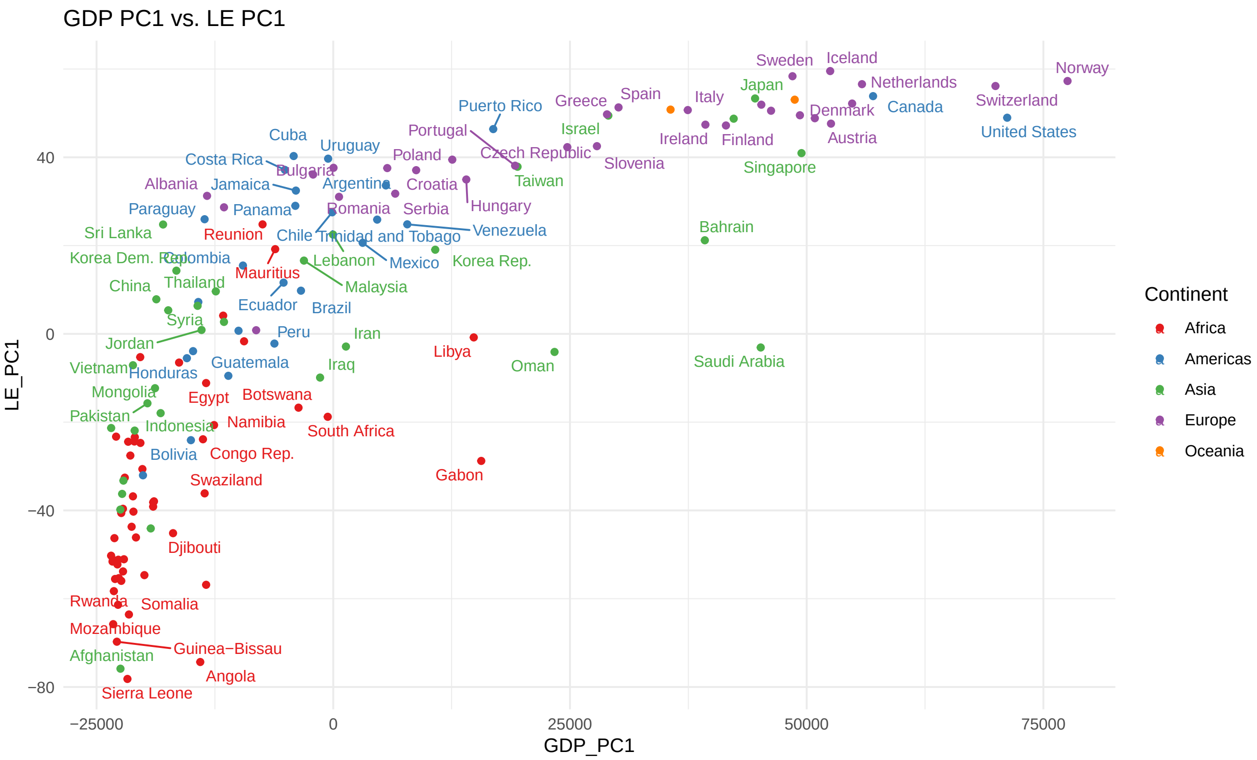
Countries interpretation: **Sierra Leone** has the lowest PC1, indicating an overall lowest life expectancy.

b) **PC2** Might represent a cyclic variation. As higher PC2 indicate higher life expectancy between 1952-1982, while lower PC2 indicate higher life expectancy after 1987. (PC2 loadings are negative after 1982 and positive before 1982)

Continents interpretation: Therefore **Asian** countries tend to have higher life expectancy after 1987. This means that Asian countries generally started developing quickly after 1987.

Countries interpretation: **Oman** has the lowest PC2, indicating higher life expectancy after 1987.

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



```
##          1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 2002 2007
## LE_PC1    0.1 0.12 0.14 0.18 0.22 0.25 0.28 0.31 0.33 0.37 0.41 0.47
## GDP_PC1   0.3 0.30 0.30 0.30 0.29 0.28 0.27 0.27 0.28 0.28 0.29 0.29
```

Above is the scatter plot of first PC score for life expectancy against first PC score for GDP. (The loadings for them are attached as well)

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.

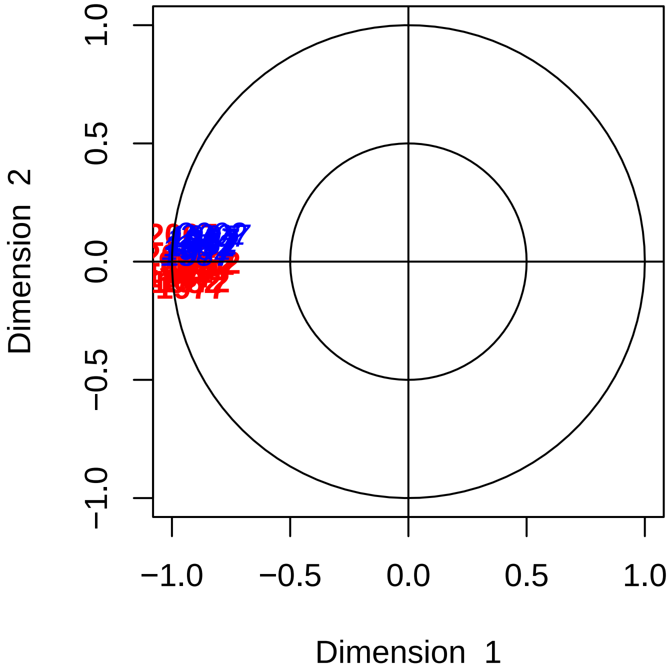
Continents interpretation: **African** tend to have both the lowest overall GDP/life expectancy. **European** and **American** have the higher overall GDP/life expectancy.

Countries interpretation: **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.

Extra: The reason why the scatters are located on the left-upper side is due to the difference between scales of GDP and life expectancy. As life expectancy has a range of 0-100 while GDP can reach a value of 5e+05.

Part 3: Canonical correlation analysis

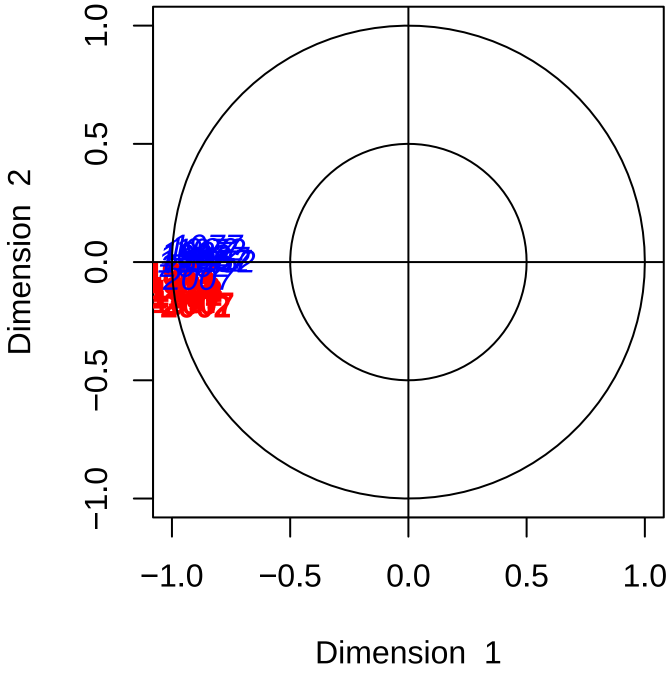
```
cca<-cc(log(gdp),lifeExp)
plt.cc(cca, var.label=TRUE,type='v')
```



First I plot the correlation between the first two log(GDP) CC scores η_1 and η_2. The original variables are log(GDP) (red) and life Expectancy(blue).

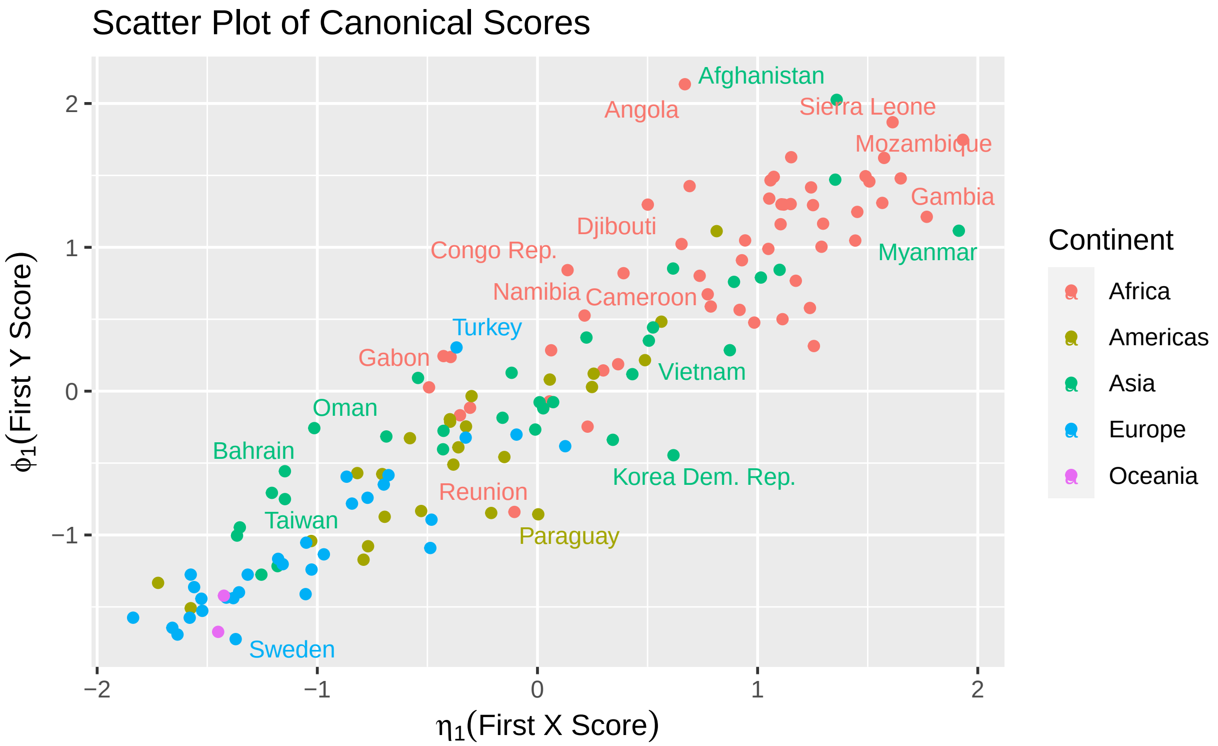
We see that η_1 is highly negatively correlated with all variables. But η_2 is less interesting here as it's uncorrelated with all other variables.

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



Then I plot the correlation between the first two life expectancy CC scores ψ_1 and ψ_2. The original variables are log(GDP) (blue) and life Expectancy(red).

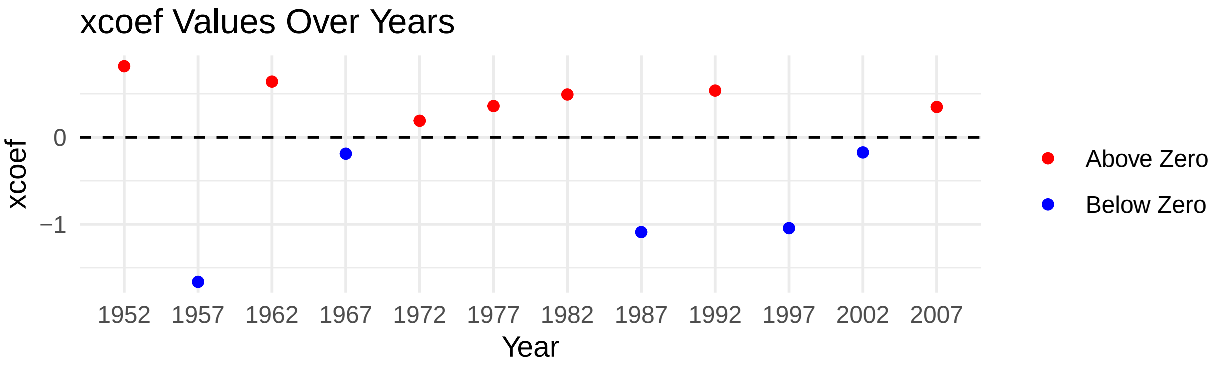
We see that ψ_1 is highly negatively correlated with all variables. But ψ_2 is less interesting here as it's uncorrelated with all other variables.

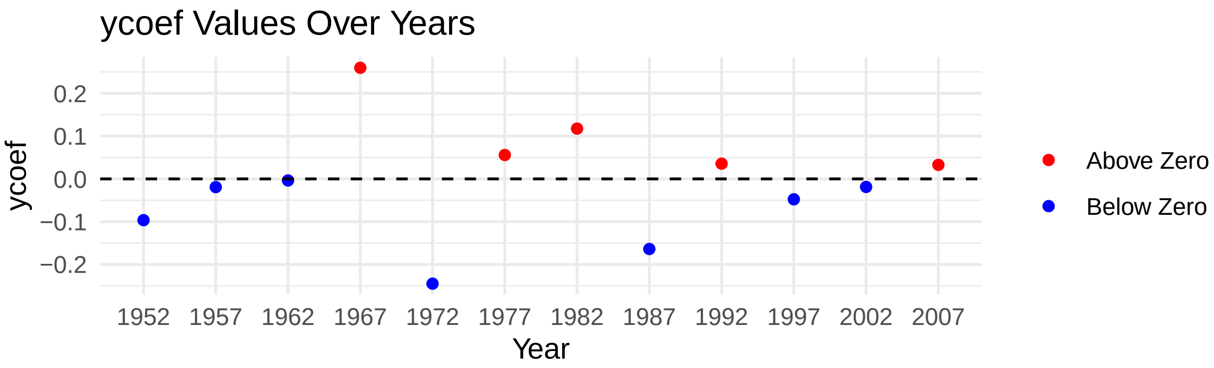


Above is the scatter plot of the first pair of CC variables. From this plot we can conclude that there are strong correlation between the first pair of CC variables η_1(First X score) and ψ_1(First Y score).

Part 3.1: Interpretation

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





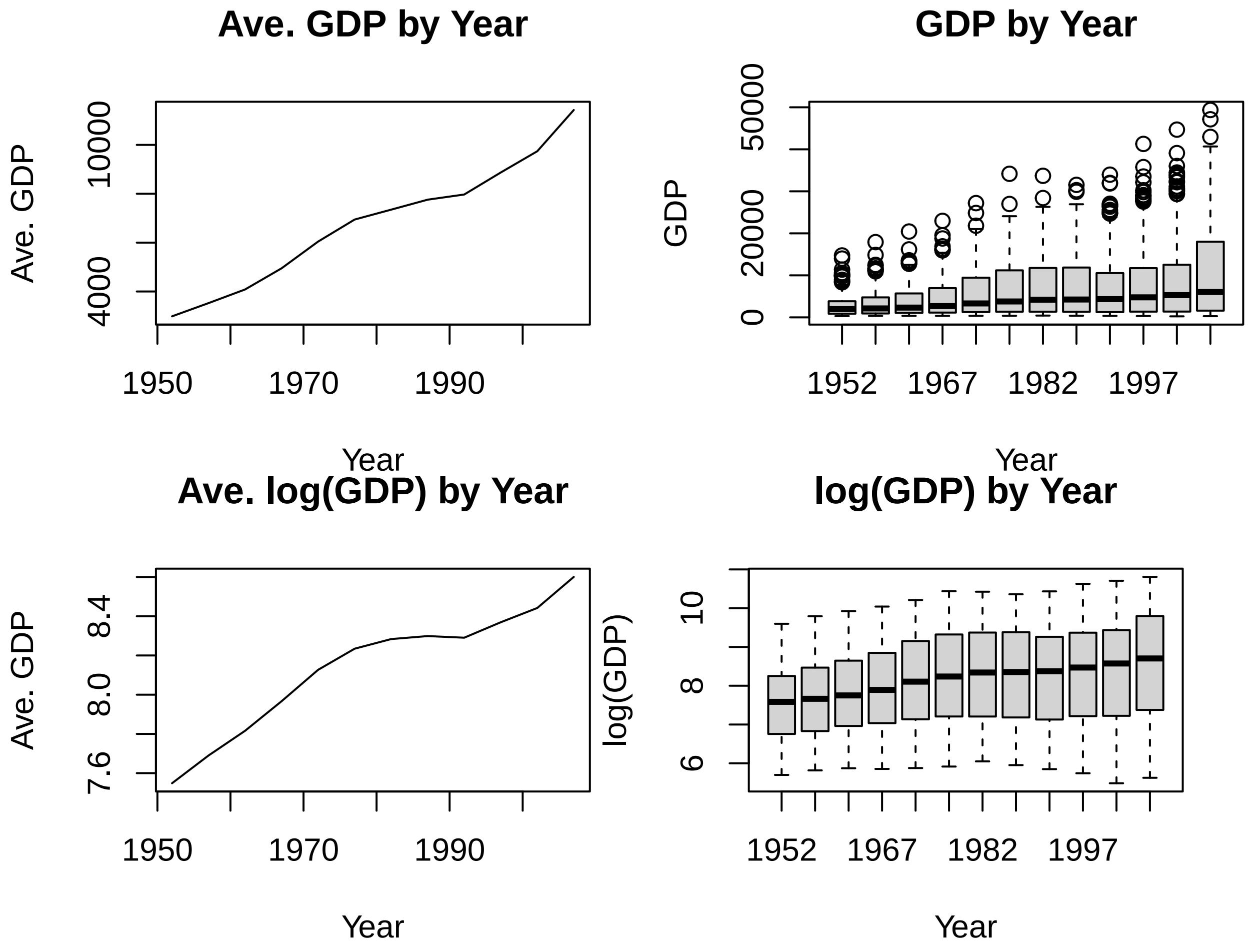
1) η_1: The higher the η_1, the higher the values of log(gdp) in 1952, 1962, 1972-1982, 1992, 2007.

2) ψ_1: The higher the ψ_1, the higher the life expectancy in 1967, 1977,1982,1992 and 2007.

We see that Europe countries generally has lower η_1, meaning that it yields high gdp value overall since the x coefficient has a mean of -0.06(less than 0). Europe countries also have lower ϕ_1, meaning that it yields higher life expectancy since the y coefficient has a mean of -0.008(less than 0). The African countries yields the opposite conclusion.

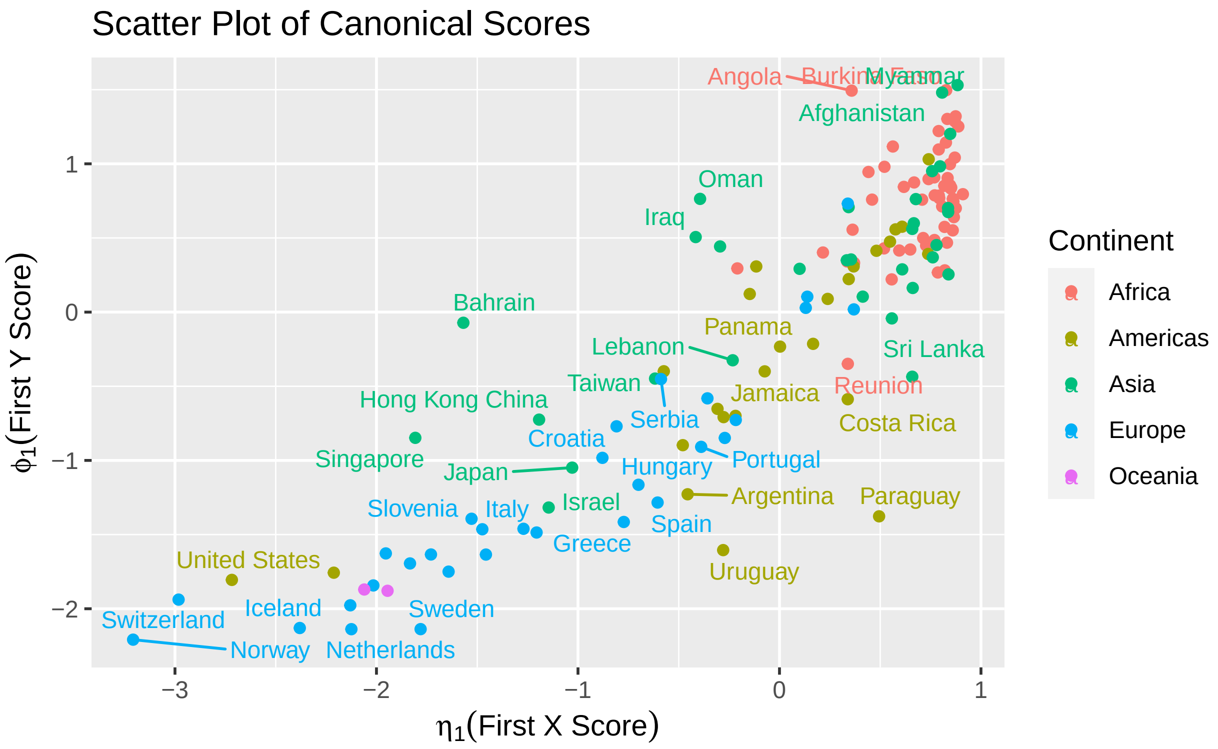
Part 3.2: Why log(gdp)?

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 the distribution less skewed and evenly distributed, which is helpful for the CCA.

Furthermore, if we apply cca on raw GDP instead of the log(GDP), we'll obtain the following η_1 and ψ_1:



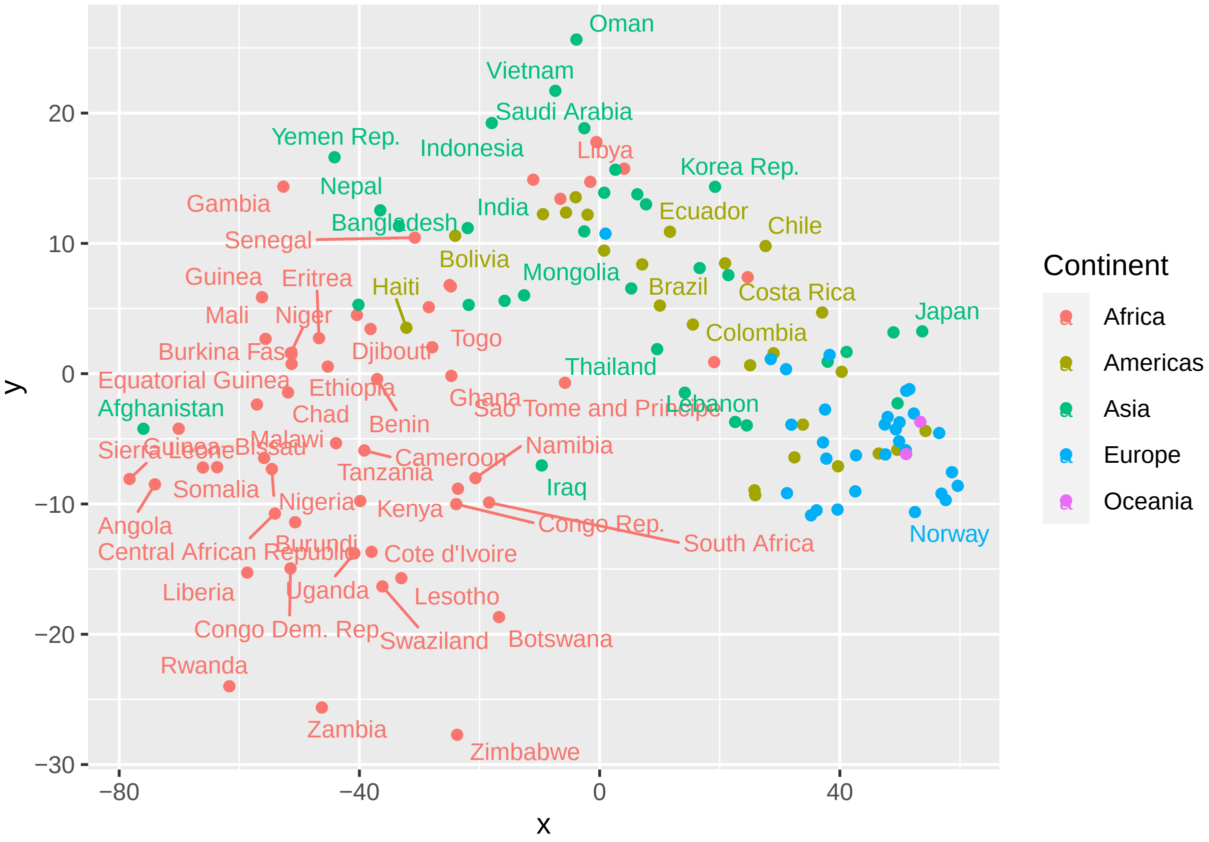
Therefore this is another reason why we need to apply log(gdp), since the raw gdp cca plot shows weak correlation between η_1 and ψ_1.

Part 4: Multidimensional scaling

Part 4.1: Interpretation of MDS

```
UN.transformed <- cbind(log(UN[,3:14]), UN[,15:26], log(UN[,27:38]))
UN.transformed <- dist(UN.transformed)
UN.transformed <- cmdscale(UN.transformed)
```

For this part I calculated the distance matrix of the dataset, and then performed multidimensional scaling for the distance matrix.



MDS can identify patterns, trends, and potential anomalies among the data provided. The plot above contains the lower dimensional representation of the original data which contains the distance information.

In this plot, shorter distance indicates higher similarity between data points(countries), on the contrary longer distance indicate minor similarity.

As a result, data points are clustered based on continents, African countries are mostly located at the left bottom while the European countries are located at the right center. This shows that countries from the same continent tends to share similar properties as they clustered together. (This also means the MDS successfully captures the trend between different countries)

Part 4.2: Comparison between MDS and PCA/CCA

All three methods are dimensionality reduction methods with different aims. PCA focus on describing the principle components that maximize transformed data's variance, CCA focus on describing the canonical components that maximize transformed data's correlation, and MDS creates a set of R^2 dataset that contains the information of the distance matrix.

From the colored figures for all three methods, we see that they all successfully captured the trend inside the data, clustering the data from Africa and Europe, while countries in Asia/America/Oceania spread evenly among the figure.

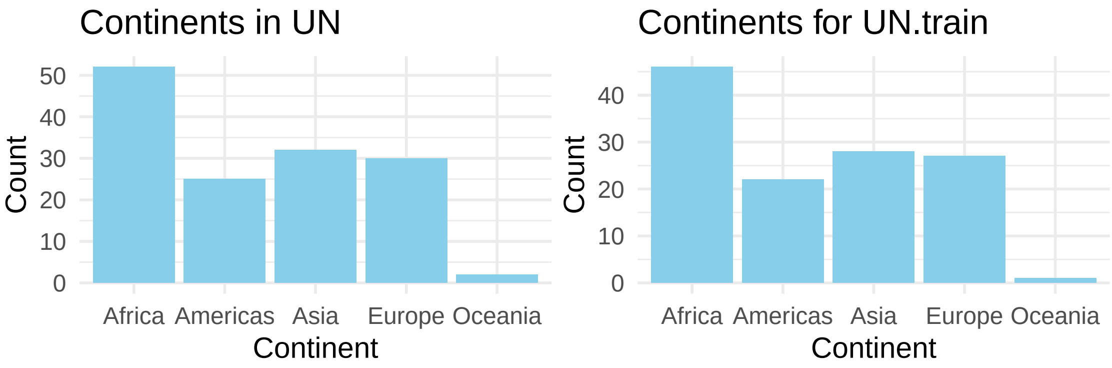
Part 5: Linear Discriminant Analysis

Part 5.1: Train-Test split

Since we are predicting the continent of each country, let's first visualize out sample data.

```r
set.seed(123)  # for reproducibility
# Creating indices for a stratified sample
test.index <- createDataPartition(UN$continent, p = 0.1, list = FALSE)

UN.test <- UN[test.index,]
UN.train <- UN[-test.index,]
```



```
## [1] "The number of occurrences of Oceania in UN is 2"
```

```
## [1] "The number of occurrences of Oceania in UN.train is 1"
```

This plot shows that we have only 2 observation of Oceania. Therefore I implement stratified sampling technique from the caret package to ensure that the train-test split is well partitioned and guarantees at least one observation of Oceania is included into the training set

Part 5.2: LDA fitting (Without PCA)

Now we fit the lda predictor and the result is given as follows:

```
## [1] "The predictive accuracy is  64.7058823529412 %"
```

```r
table(UN.pred$class, UN.test$continent)
```

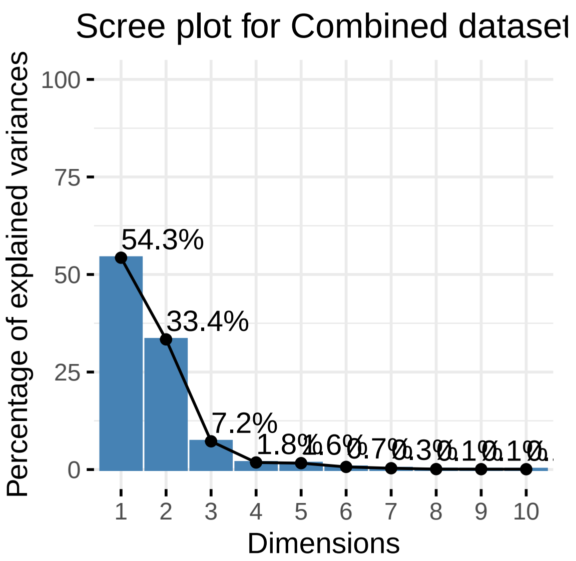
```
##
##             Africa Americas Asia Europe Oceania
##   Africa         6        0    1      0       0
##   Americas       0        1    1      0       0
##   Asia           0        2    2      0       0
##   Europe         0        0    0      2       1
##   Oceania        0        0    0      1       0
```

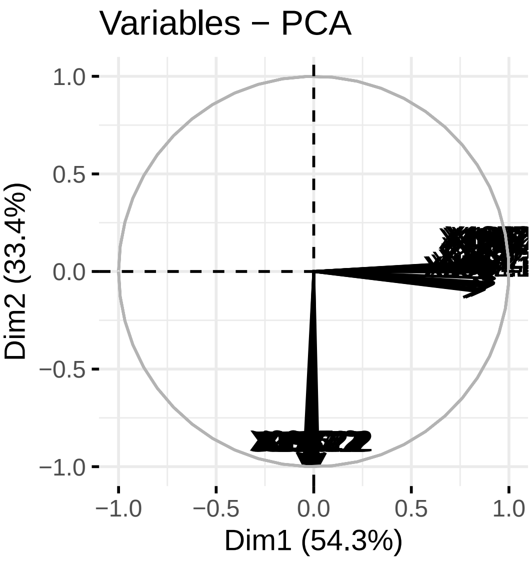
In this table, the rows are the number of predicted continents, where the columns indicates the number of actual continents.

Therefore 1 Oceania country has been mis-classified as Europe, 1 Europe country is mis-classified as Oceania, 2 Asia country is misclassified as Africa/America and 2 American country are mis-classified as Asia.

Part 5.3: LDA fitting (With PCA)

64% prediction accuracy is unacceptable for a predictive model. Therefore I decide to perform PCA on the dataset before performing LDA.





Based on the scree plot above, I decided to train the LDA predictor based on the first 6 principle components to include more information. Using the same train_test split obtained earlier, the PCA based lda predictor shows an increased accuracy of 76%, which is improved compared with the previous non-PCA measure.

```
## [1] "The predictive accuracy is  76.4705882352941 %"
```

```
table(all.pred$class, all.test$continent)
```

```
##
##            Africa Americas Asia Europe Oceania
##   Africa        5        0    0      0       0
##   Americas      0        2    1      0       0
##   Asia          1        1    3      0       0
##   Europe        0        0    0      3       1
##   Oceania       0        0    0      0       0
```

In this table, the rows are the number of predicted continents, where the columns indicates the number of actual continents.

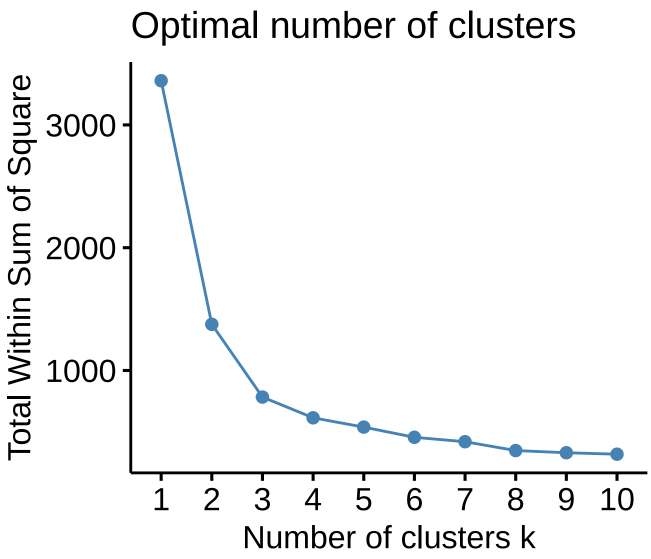
Therefore 1 Oceania country has been mis-classified as Europe, 1 Asia country is misclassified as America, 1 American country is mis-classified as Asia and 1 Africa country is mis-classified as Asis.

Part 6: Clustering

In the following section, I'm going to perform three different clustering techniques, and they are K-means clustering, Model-based clustering and Hierarchical clustering.

Part 6.1: K-means clustering

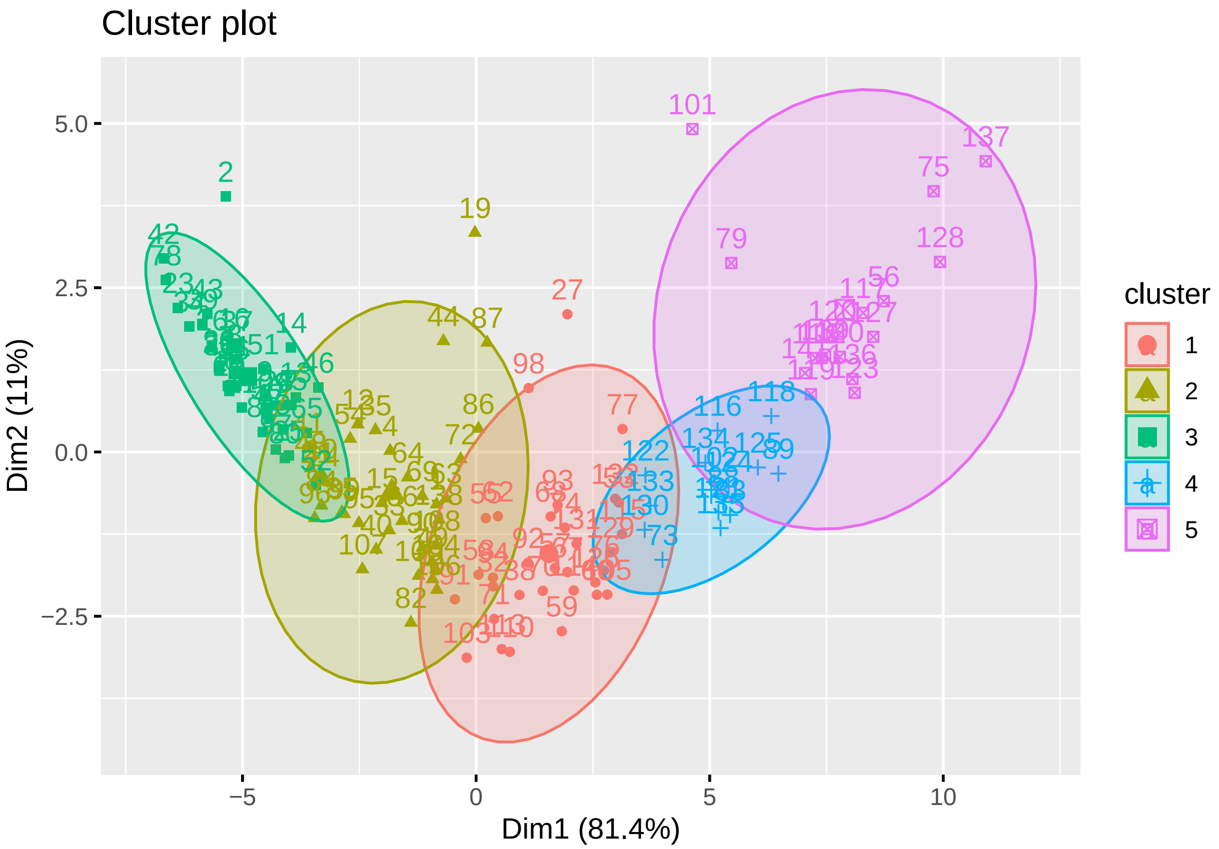
Based on the elbow method, it seems to have 3-5 natural clusters in the data, because the Wss decrease rapidly from 2 to 5, and later parts only yields minor decreases.



```
library(factoextra)
set.seed(123)
UN.k <- kmeans(UN.scaled[,3:26], centers = 5, nstart=25)
```

Therefore I perform K-means cluster with center number K=5, and the clustering result is included below. It seems that African countries has similar data pattern and was clustered mostly in cluster 2. Furthermore, Americas and Asia are mostly clustered into cluster 1,2.

	Africa	Americas	Asia	Europe	Oceania
1	3	14	7	8	0
2	16	7	14	1	0
3	33	1	5	0	0
4	0	1	4	10	0
5	0	2	2	11	2



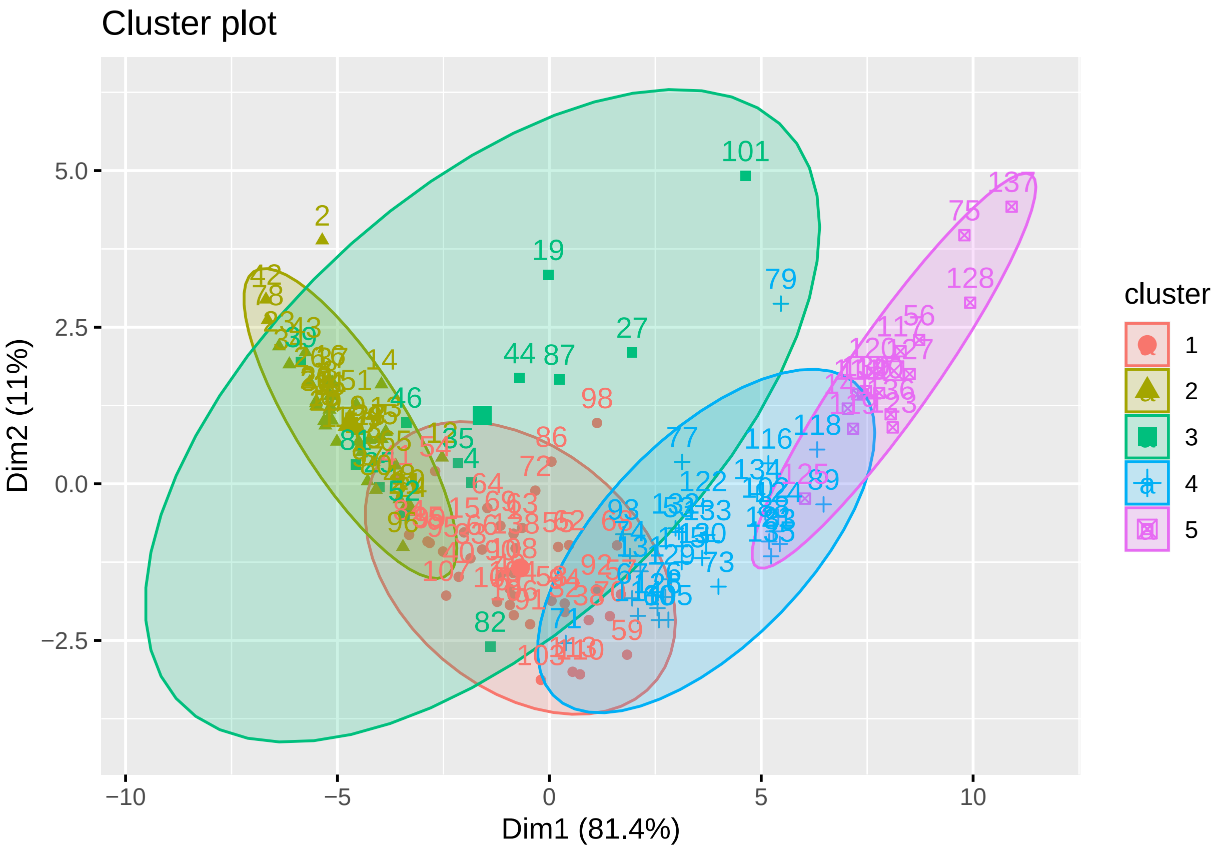
From the above PC component plot, it can be seen that the data has been clustered into 5 different natural sets with minor coincide between clusters.

Part 6.2: Model-based clustering : Gaussian clusters

Since how to choose K for model-based clustering is beyond the scope of this module, I randomly choose G = 5 ^_^ (Because we got five different continents and hopefully it can cluster the sets into different continents)

	Africa	Americas	Asia	Europe	Oceania
1	8	14	16	3	0
2	35	1	5	0	0
3	9	0	4	0	0
4	0	8	7	15	0
5	0	2	0	12	2

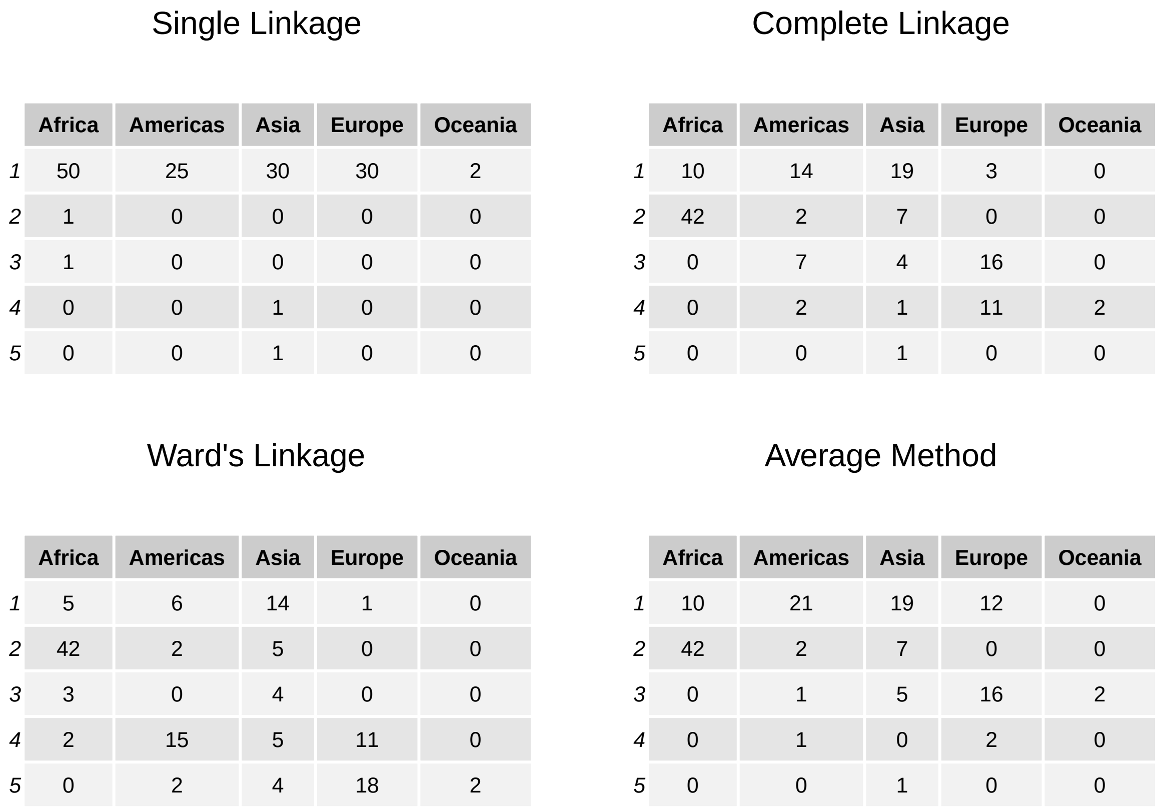
The result is similar to the previous part, cluster 2 seems to contain most of the Africa countries and cluster 1 seems to contain most of the Americas and Asia countries. Cluster 4/5 contains most of the Europe countries.



From the above PC component plot, it can be seen that the model-based Gaussian clustering doesn't provide a good cluster, as there are plenty of overlaps between different clusters(especially cluster 3).

Part 6.3: Hierarchical clustering

For Hierarchical clustering, since we only have five different continents, therefore we choose to cut off the tree at depth=5 and check which clustering technique can distinguish the continents the best.



Single Linkage

	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

Complete Linkage

	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

Ward's Linkage

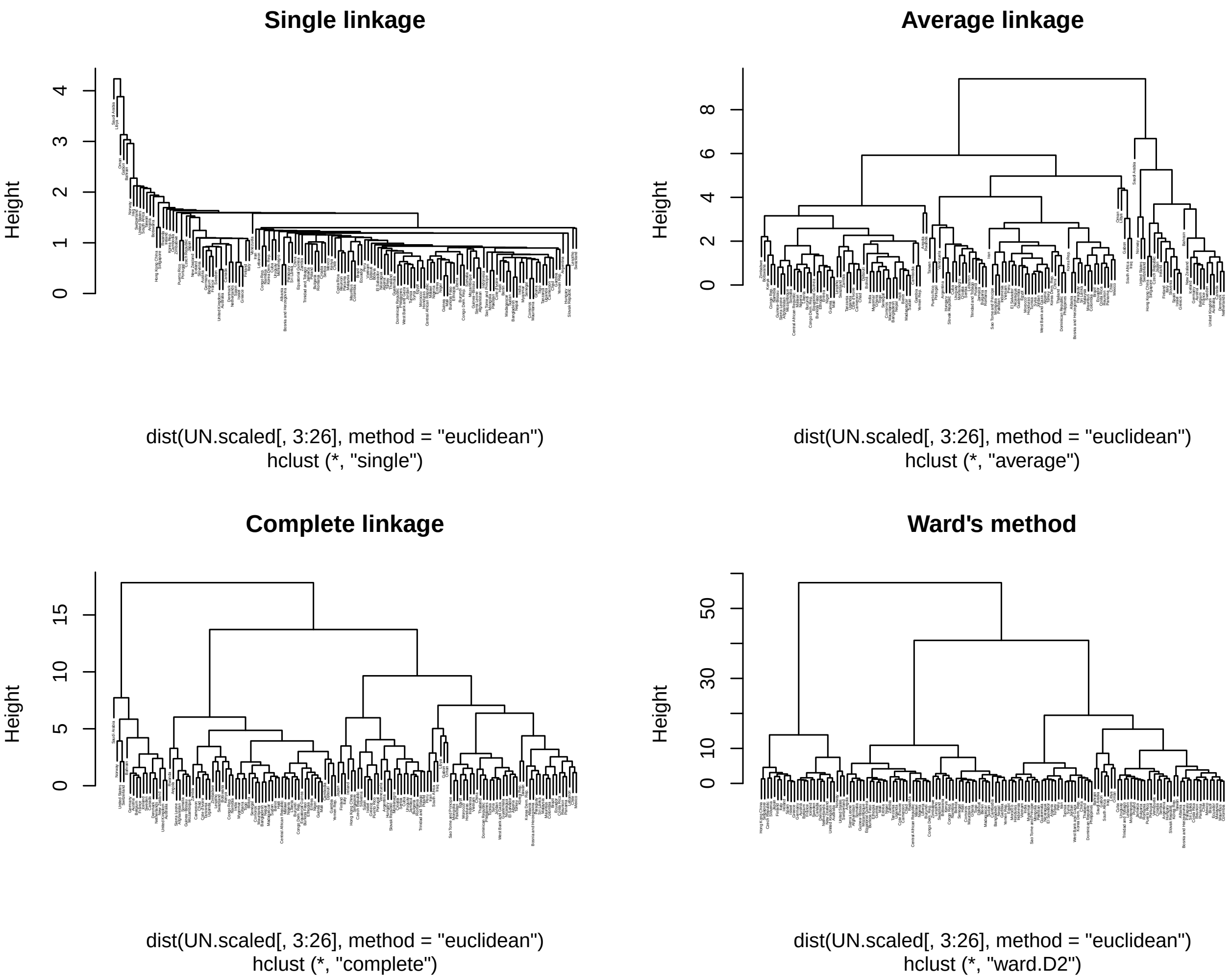
	Africa	Americas	Asia	Europe	Oceania
1	5	6	14	1	0
2	42	2	5	0	0
3	3	0	4	0	0
4	2	15	5	11	0
5	0	2	4	18	2

Average Method

	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

From the above tables, it can be checked that the Hierarchical clustering with Ward's method is the best, as group_2 is dominanted with **Africa** countries, while group_5 is dominated by **Europe** countries.

This means that **Africa** and **Europe** contries tend to be closer to each other under the **Ward's** method.



From the above Dendrograms, it can be observed that African countries are successfully clustered together in each method, this reflects that African countries data shares a lot in common.

Furthermore, **Saudi Arabia** can be observed at upper part in many dendrograms, this means it doesn't join into cluster until later stage of clustering. This observation reflects **Saudi Arabia** has special data property when compared with other countries.

Part 6.4: Clustering Conclusion

Different methods find different clusters, and we naturally interpret the clusters as continents.

Therefore, we evaluate the performance of clustering based on whether it successfully split the data into clusters that represent continents. By visually examining the clustering results, the Hierarchichal clustering with **Ward's Method** proves to be the best method as it successfully obtained groups that are dominated by Africa, Asia, Europe and America.

Part 7: Linear regression

Part 7.0: Train Test split

```r
set.seed(123)  # for reproducibility
# Creating indices for a stratified sample
test.index <- createDataPartition(UN$continent, p = 0.1, list = FALSE)
```

```
UN.test <- UN[test.index,]
UN.train <- UN[-test.index,]
```

Before fitting any models, I first split the data into training dataset and testing dataset using stratified sampling technique, one of the training set is GDP and another training set is log(GDP). In later part I'll fit the **OLS**, **PCR** and **Ridge regression** on raw GDP and log(GDP), then compare their testing accuracy on the testing dataset to determine which model is the best.

Part 7.1: OLS(Ordinary Least square)

```
##
## Call:
## lm(formula = y ~ ., data = data_train)
##
## Coefficients:
##    (Intercept)  gdpPercap_1952  gdpPercap_1957  gdpPercap_1962  gdpPercap_1967
##      5.782e+01      -1.932e-03       5.712e-03      -5.067e-03      -3.870e-06
## gdpPercap_1972  gdpPercap_1977  gdpPercap_1982  gdpPercap_1987  gdpPercap_1992
##      1.346e-03      -1.399e-03       2.011e-03       7.629e-05      -3.283e-03
## gdpPercap_1997  gdpPercap_2002  gdpPercap_2007
##      4.491e-03      -2.114e-03       7.499e-04
```

```
##
## Call:
## lm(formula = y ~ ., data = data_train_log)
##
## Coefficients:
##    (Intercept)  gdpPercap_1952  gdpPercap_1957  gdpPercap_1962  gdpPercap_1967
##          4.253          -5.488          12.945          -5.008           1.442
## gdpPercap_1972  gdpPercap_1977  gdpPercap_1982  gdpPercap_1987  gdpPercap_1992
##         -4.477          -1.221          -6.514          11.791          -7.399
## gdpPercap_1997  gdpPercap_2002  gdpPercap_2007
##         17.137         -12.360           6.615
```

The above are the two OLS models I fitted for GDP and log(GDP), and their coefficients are listed above.

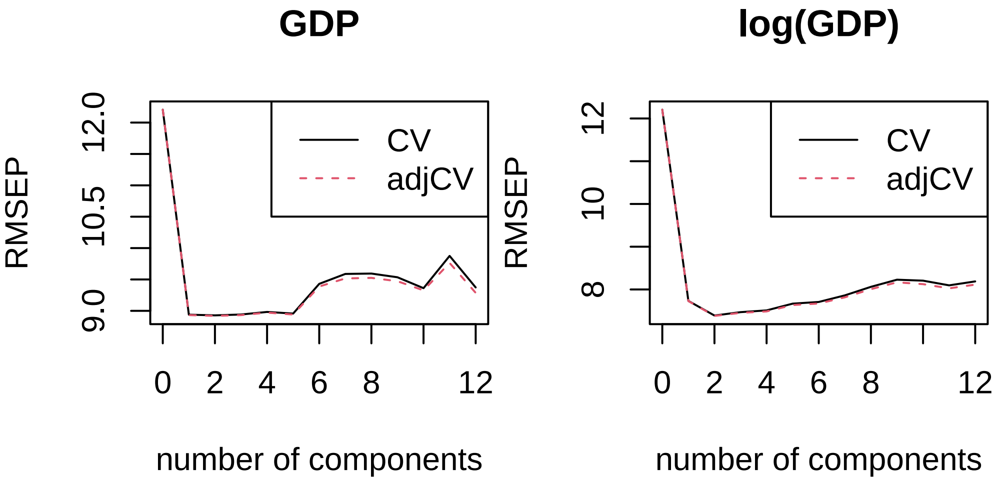
```
## [1] "The RMSEP for ols prediction on test set is 11.49"
```

```
## [1] "The RMSEP for ols_log prediction on test set is 4.89"
```

First I fit the OLS regression model on both gdp and log(gdp) model. The RMSEP(root of mean square error for prediction) are listed above. I'll compare the accuracy between all models after all model accuracies are calculated, and I'll apply the same train_test split for each model trainings.

Part 7.2: PCR (principle component regression)

```
## Data:    X dimension: 124 12
##  Y dimension: 124 1
## Fit method: svdpc
## Number of components considered: 12
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           12.21    8.939    8.928    8.941    8.983    8.957    9.430
## adjCV        12.21    8.932    8.919    8.932    8.971    8.942    9.388
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
## CV       9.588    9.594    9.535     9.361     9.875     9.373
## adjCV    9.516    9.525    9.472     9.326     9.763     9.286
##
## TRAINING: % variance explained
##    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X    93.71    98.23    99.04    99.46    99.71    99.84    99.90    99.94
## y    47.00    47.57    47.60    47.69    48.41    48.41    49.56    50.19
##    9 comps  10 comps  11 comps  12 comps
## X    99.97     99.98    100.00    100.00
## y    50.80     51.01     52.76     54.13
```



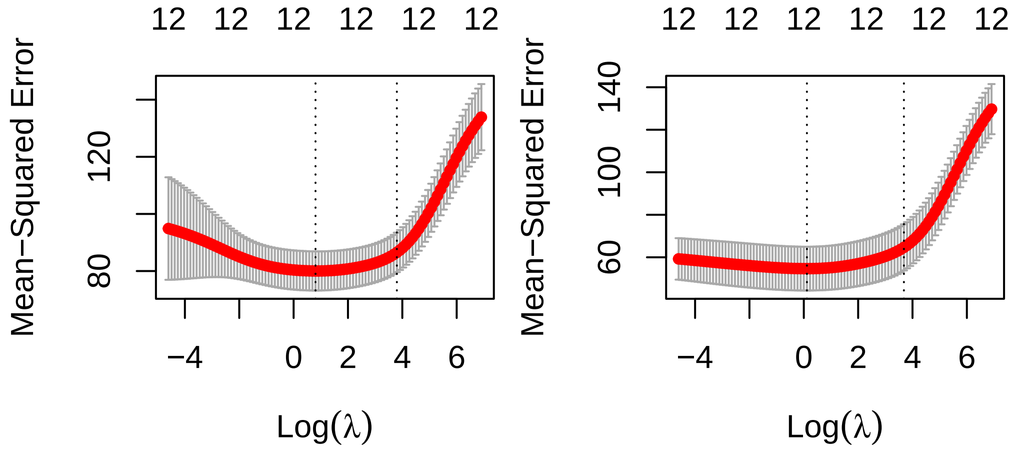
From the above figure, it looks like for GDP 1 component is sufficient, as using more than 1 component doesn't lead to big decreases in the cross-validation error and even lead to larger prediction errors. But for log(GDP), 2 components are needed, since the lowest RMSEP is achieved at components = 2.

```
## [1] "The RMSEP for pcr prediction on test set is 8.88"
```

```
## [1] "The RMSEP for pcr prediction on test set is 5.54"
```

Here I fit the PCR regression model on both gdp and log(gdp) model. The RMSEP(root of mean square error for prediction) are listed above.

Part 7.3: Ridge regression



Here I choose the lambda which minimise the mean-square error in the cross validation.

```
## [1] "The RMSEP for ridge on test set is 8.9"
```

```
## [1] "The RMSEP for ridge_log on test set is 5.56"
```

Here I fit the Ridge regression model on both gdp and log(gdp) model. The RMSEP(root of mean square error for prediction) are listed above.

Part 7.4: Conclusion

```
##    Model RMSEP_Not_Logged RMSEP_Logged
## 1    OLS        11.493051     4.887370
## 2    PCR         8.882820     5.536546
## 3 Ridge          8.901709     5.556581
```

Above table summarizes the final RMSEP(Root of Mean Square Error for Prediction) based on different prediction model and whether using log(gdp) (second column) or raw gdp(first column).

Therefore we reach the following conclusions:

1) For dataset, using **log(gdp)** as predictors produces better prediction result than using **raw gdp**.
2) For model, the **OLS** regression produce the lowest RMSEP, which is better than PCR and Ridge regression.