Multivariate Analysis Coursework

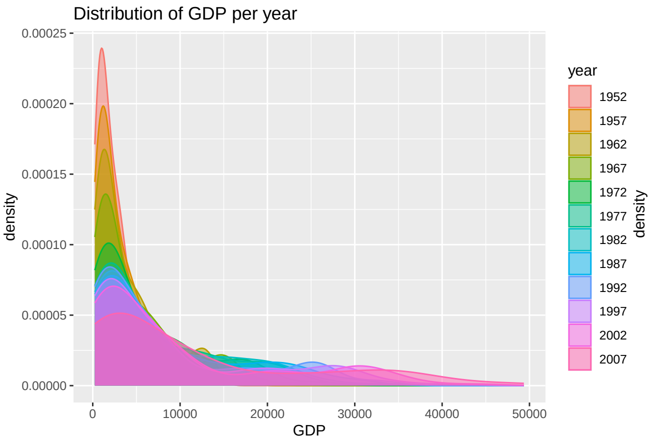
Thomas Sharples

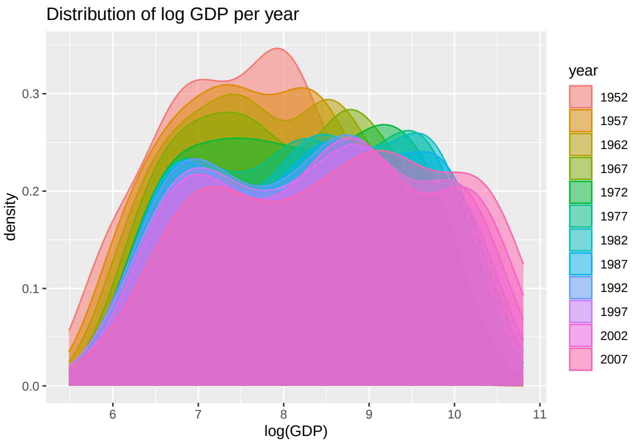
2024-04-09

In this document we will be investigating United Nations data on 141 different coutries from 1952 to 2007. We will be performing analysis on the variables GDP per capita, life expectancy and population using various methods such as PCA, linear discriminant analysis, clustering and more.

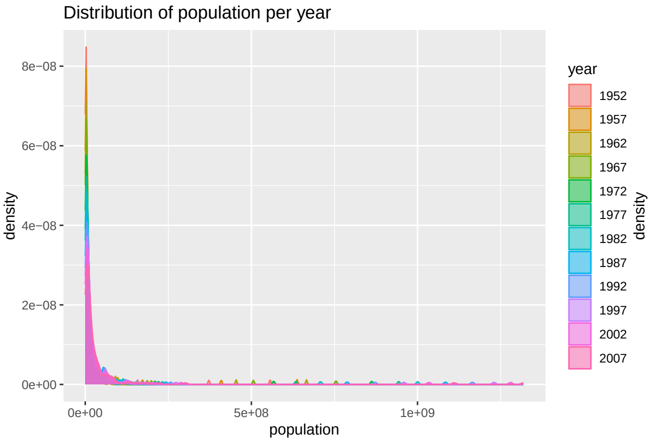
Exploratory Analysis

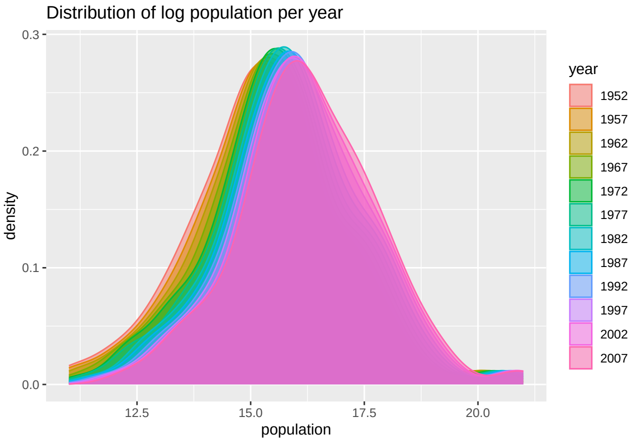
First we will take a look how each of the variables have changed from 1952 to 2007and see if we can find some common trends between them.



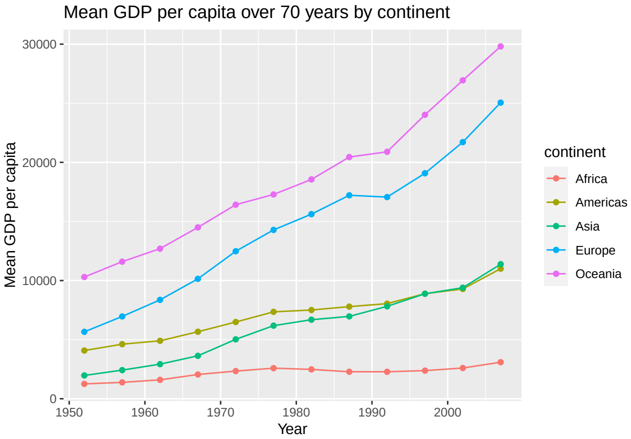


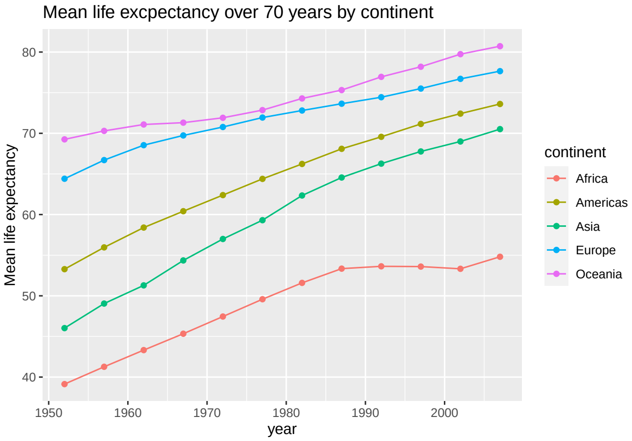
This density plot shows us overall trend for GDP per capita over the 70 years, GDP per capita has increases every year quite uniformly with the proportion of people below $10,000 reducing every year. The GDP data is positively skewed and may suggest that using a transformation would be beneficial. The plot on the right shows a log transformation and this transformation reduces this skewness.

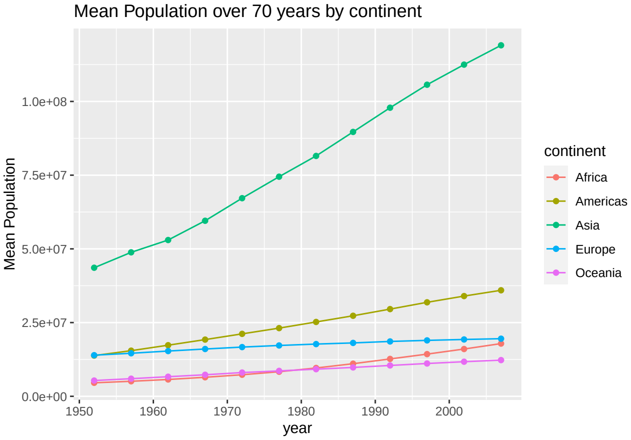




The density plots here shows that the population data is very positively skewed and a log transformation significantly reduces this.



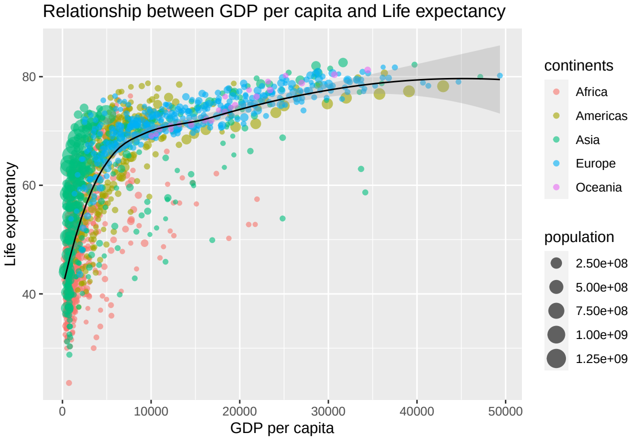


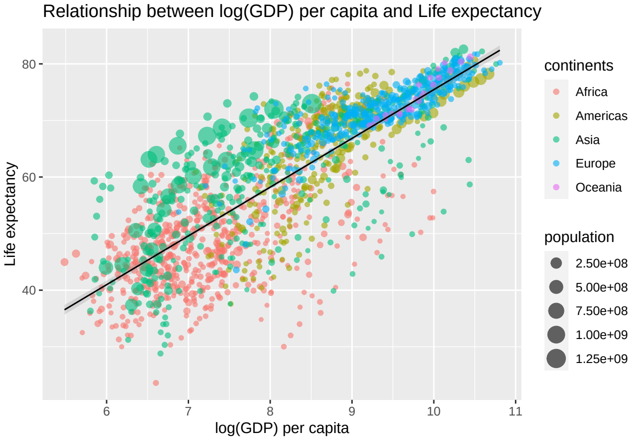


All continents follow a increasing trend from the years 1952 to 2007 where Oceania and Europe have remained to be the continents with the highest GDPs per capita. Africa's GDP per capita did increase over the 70 years but it didn't see the same augmentation as the 1st world countries like Oceania and Europe.

Similarly to GDP life expectancy increases linearly for all the continents. The rate of increase in Oceania and Europe is smaller than Asia and Americas showing that life expectancy has rapidly increased for these two continents. From 1952 to 1985 Africa had a steady linear increase in life expectancy but plateaued from 1985-2000, this may have been due to the AIDs epidemic.

The mean population of all the countries has increased linearly in the last 70 years where Asia has had the most rapid and drastic augmentation of them all. Oceania has had quite little growth over this time period even though it is leading for GDP per capita and life expectancy.





Plotting GDP per capita against life expectancy we can see the influence of the positive skewness in the GDP data. This suggests a non-liner relationship between GDP and life expectancy with clustering around less than $10,000. Applying a log transformation we see that the data spreads out and we now have a linear relationship between the two variables. The transformation makes the graph significantly more interpretable.

Europe, Oceania and large populations in Americas follow this linear relationship with the smallest deviations and they are grouped around the 70-80 life expectancy range. Africa on the other hand is very spread out deviating from the fitted line the most and they also have the lowest life expediencies.

Principal Component Analysis

In this section we will be running PCA on the variables GDP per capita and Life expectancy to see what conclusions we can draw from them!

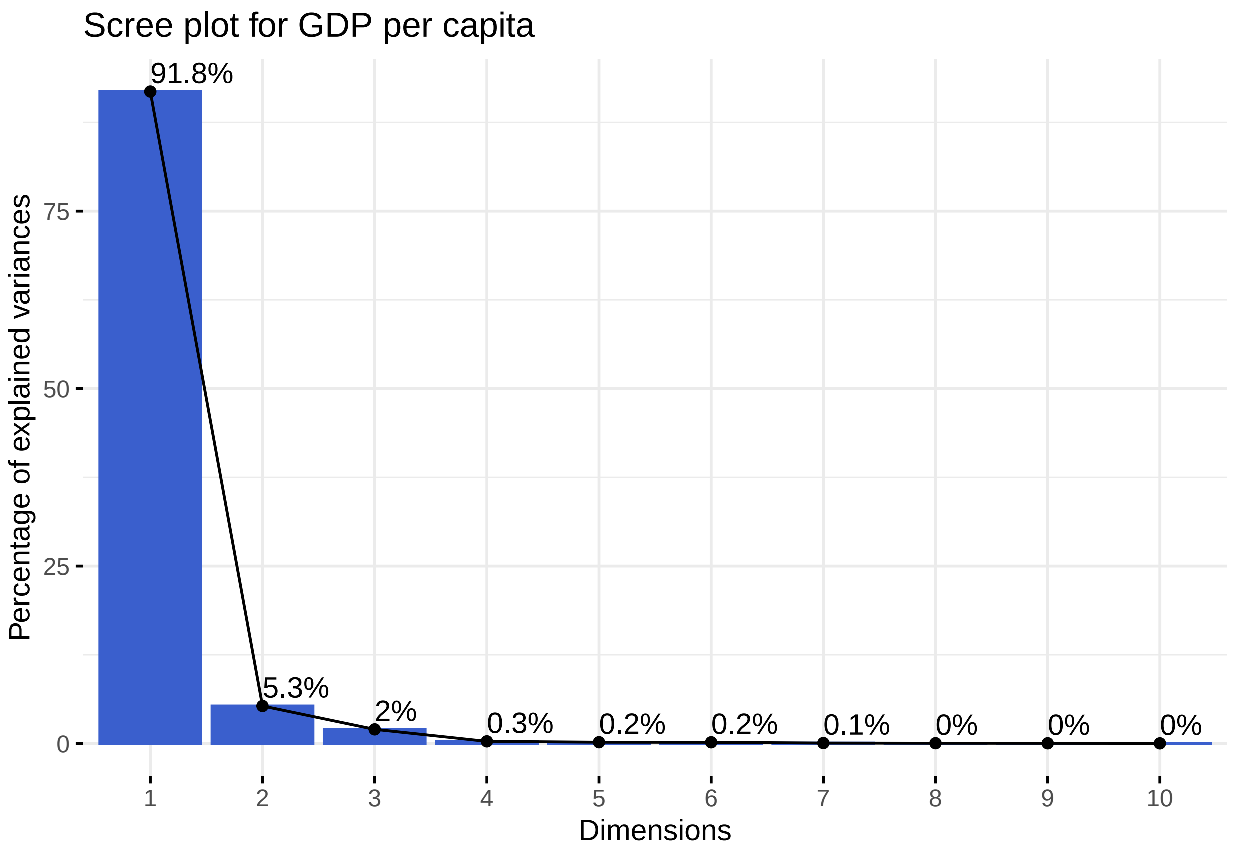
For this PC analysis we will be using correlation (R) because I want to focus on the relationships between variables rather than the absolute values of them, correlation is also less sensitive to skewness and from our EDA we can see that GDP is positively skewed. GDP and life expectancy use different scales so when we compare the first principal components of them we want them to be normalised.

PCA on GDP per capita

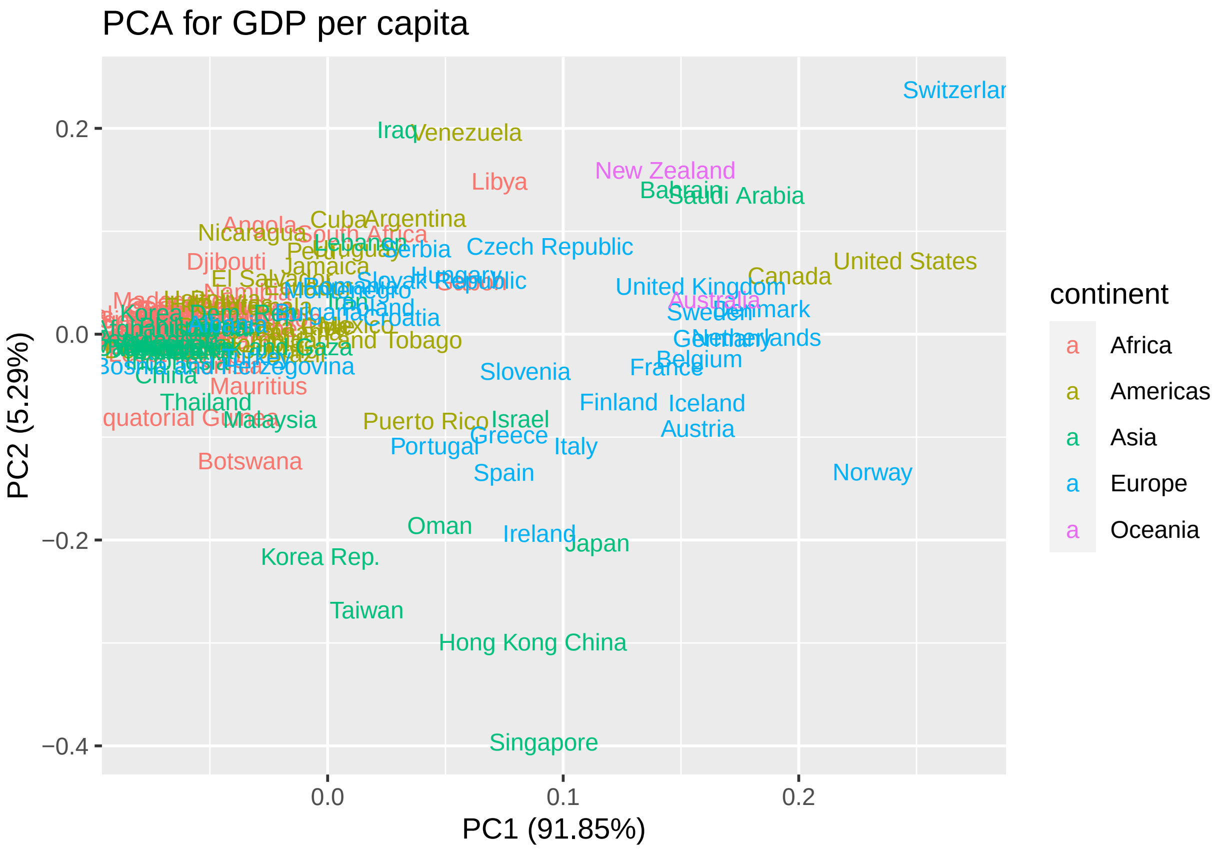
```r
gdp.pca <- prcomp(gdp[,2:13], scale=TRUE) # TRUE = corr
summary(gdp.pca)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.3199 0.79665 0.48936 0.19220 0.14866 0.14530 0.08930
## Proportion of Variance 0.9184 0.05289 0.01996 0.00308 0.00184 0.00176 0.00066
## Cumulative Proportion  0.9184 0.97134 0.99130 0.99437 0.99622 0.99797 0.99864
##                            PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.07253 0.06473 0.05548 0.04551 0.04159
## Proportion of Variance 0.00044 0.00035 0.00026 0.00017 0.00014
## Cumulative Proportion  0.99908 0.99943 0.99968 0.99986 1.00000
```

The summary output shows us that 97.1% of the variance is explained by the first two components suggesting that we only need to compute the first two PC scores. The components after this account for 2% or less of the variability in the data and there is no need to compute their principal component scores. A scree plot can help us visualise the summary output.



So we have successfully reduced the dimensions from p=12 down to p=2.

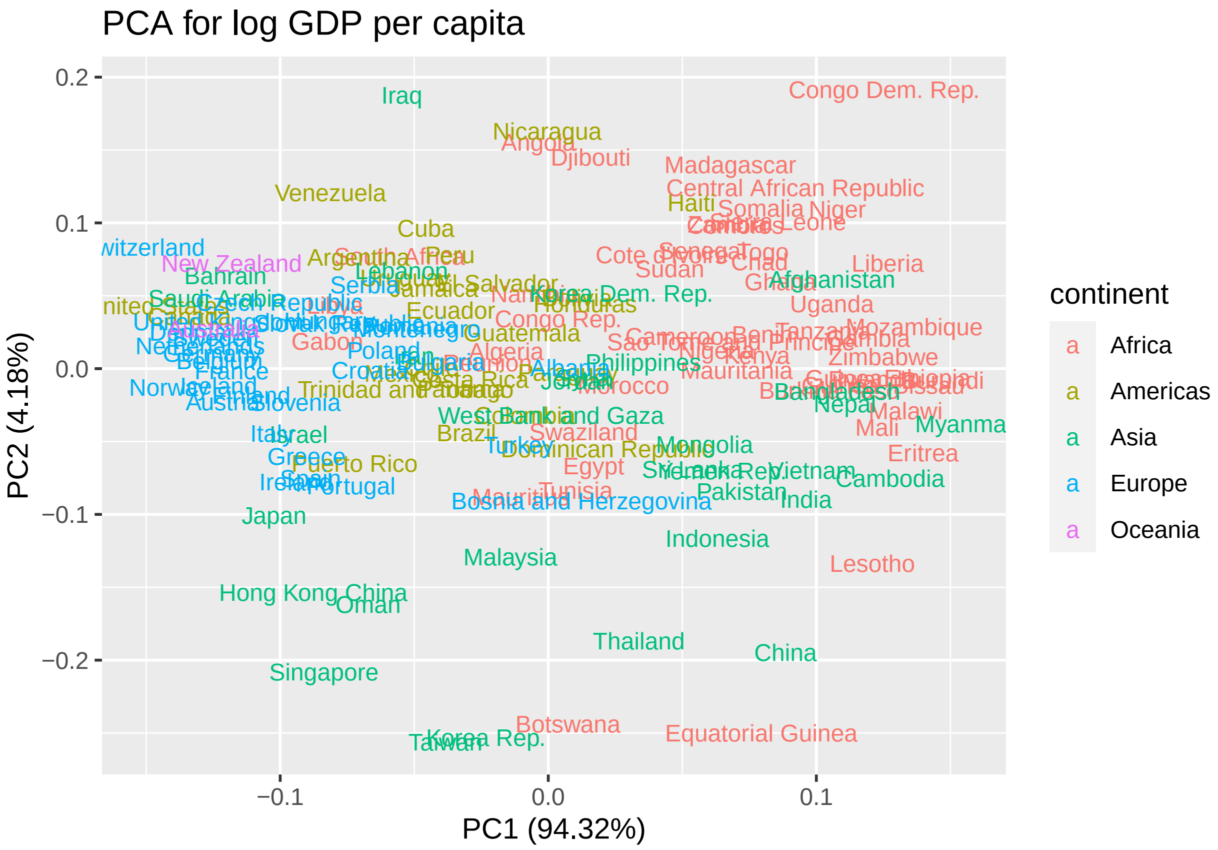


We may interpret the first principal component is the measure of average wealth per country, the more positive the first PC value the more wealthy the country is, however I feel that skewness in the GDP data is influencing our plot by clustering around the origin so a log transformation may make it more readable.

Applying log transformation to GDP,

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.3643 0.70842 0.31393 0.18516 0.13471 0.09757 0.07505
## Proportion of Variance 0.9432 0.04182 0.00821 0.00286 0.00151 0.00079 0.00047
## Cumulative Proportion  0.9432 0.98502 0.99323 0.99609 0.99760 0.99840 0.99886
##                            PC8    PC9    PC10    PC11    PC12
## Standard deviation     0.06369 0.0597 0.04794 0.04525 0.04068
## Proportion of Variance 0.00034 0.0003 0.00019 0.00017 0.00014
## Cumulative Proportion  0.99920 0.9995 0.99969 0.99986 1.00000
```

Applying the log transformation we see that the first two components now explain 98.5% of the variance which is an improvement. Plotting PC1 against PC2,

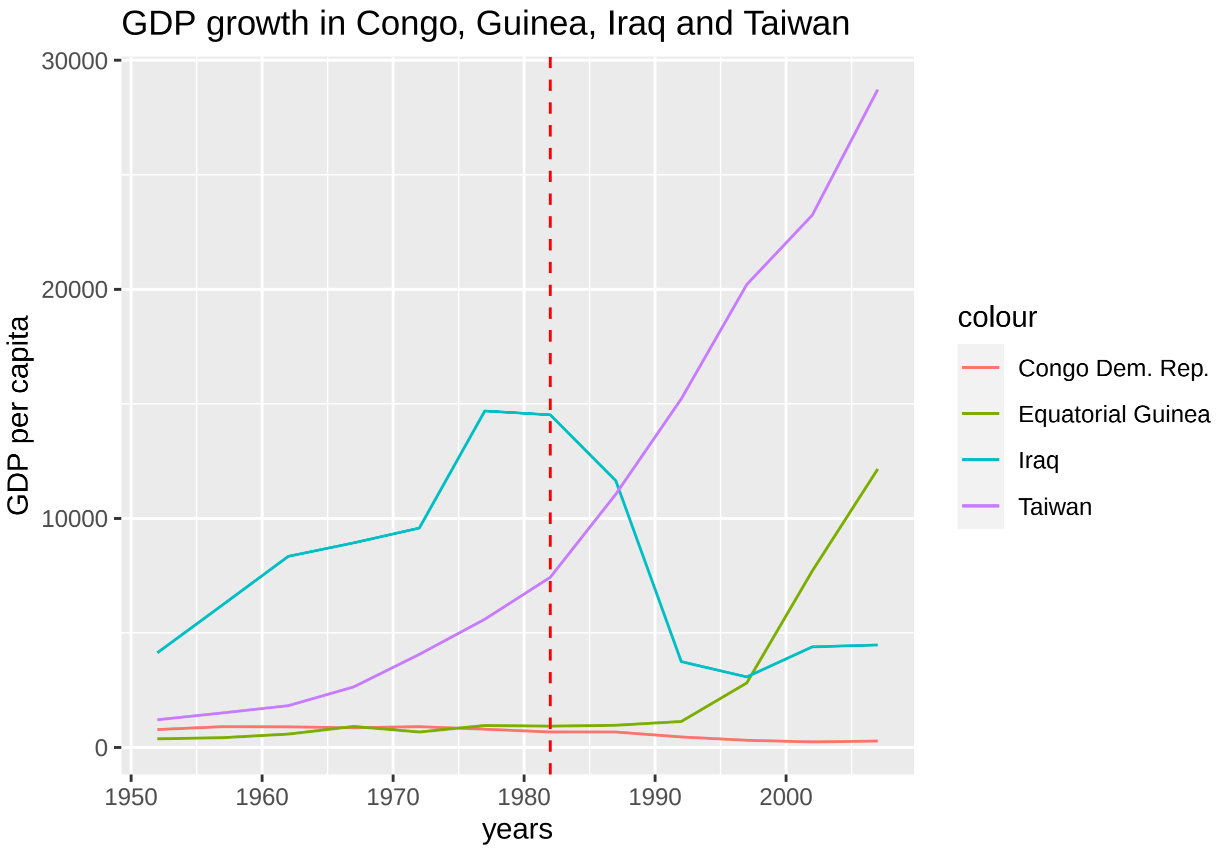


The log transformation spreads out the data and allows us to see more clusters rather than one large cluster around the origin. Now we have European and Oceanic countries are on the left with negative PC1 scores and African countries are on the right with positive PC1 values. I would interpret the first principal component as the average wealth per country the more negative the PC1 value is the higher the average wealth. We see that the log transformation creates clusters somewhat by continent but not explicitly with the more well known wealthy countries on the left and the poorer African and Asian on the right.

```
gdp_log.pca$rotation[,1:2]
```

```
##             PC1         PC2
## 1952 -0.2817154  0.38605152
## 1957 -0.2855489  0.36017954
## 1962 -0.2889846  0.31529866
## 1967 -0.2907042  0.24985118
## 1972 -0.2926581  0.16291613
## 1977 -0.2939384  0.06866022
## 1982 -0.2942796 -0.02039763
## 1987 -0.2935572 -0.12534807
## 1992 -0.2908985 -0.24360796
## 1997 -0.2868638 -0.34792157
## 2002 -0.2839805 -0.39614363
## 2007 -0.2805256 -0.41795686
```

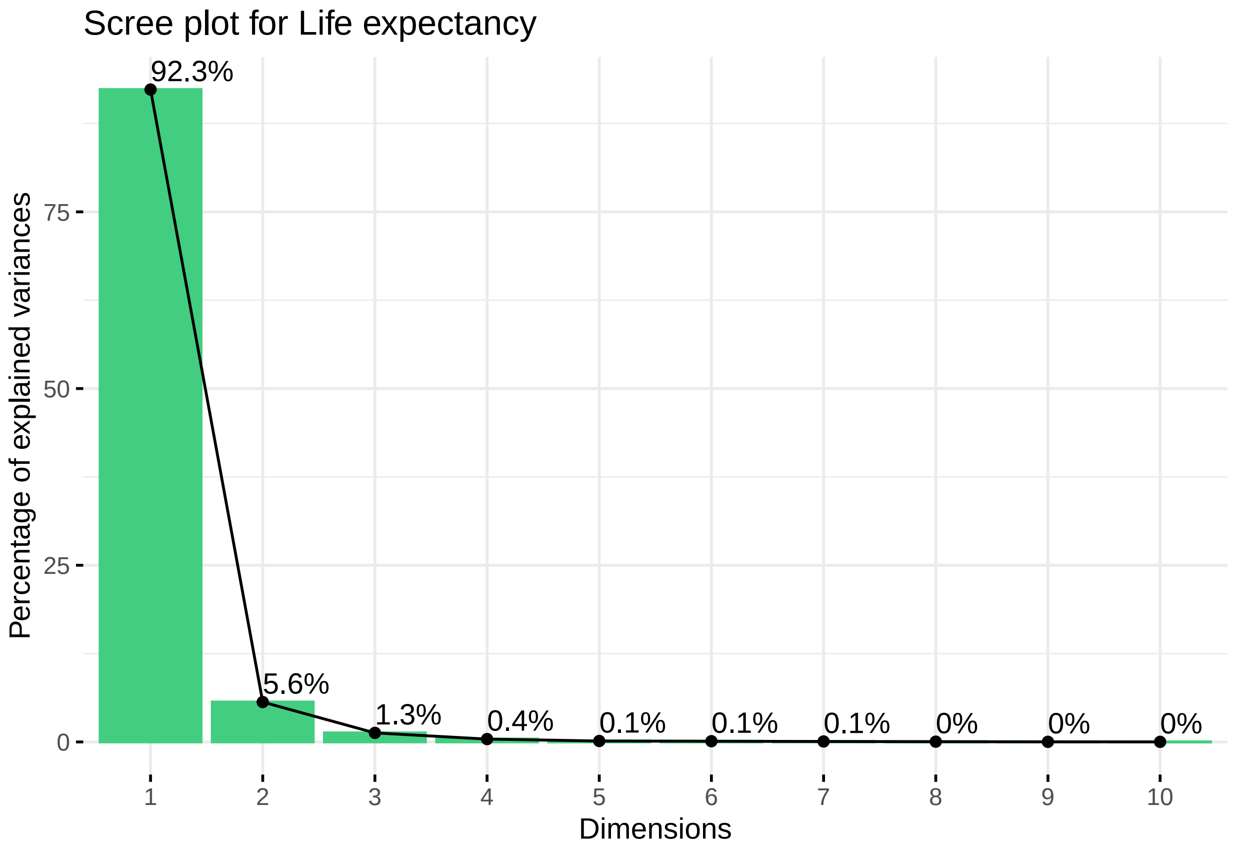
Using the eigen vectors for PC2, I interpret the second principal component as an indicator of GDP growth after 1982, if the PC2 score is large and negative then it indicates significant GDP growth and if it is positive it suggests GDP decline.



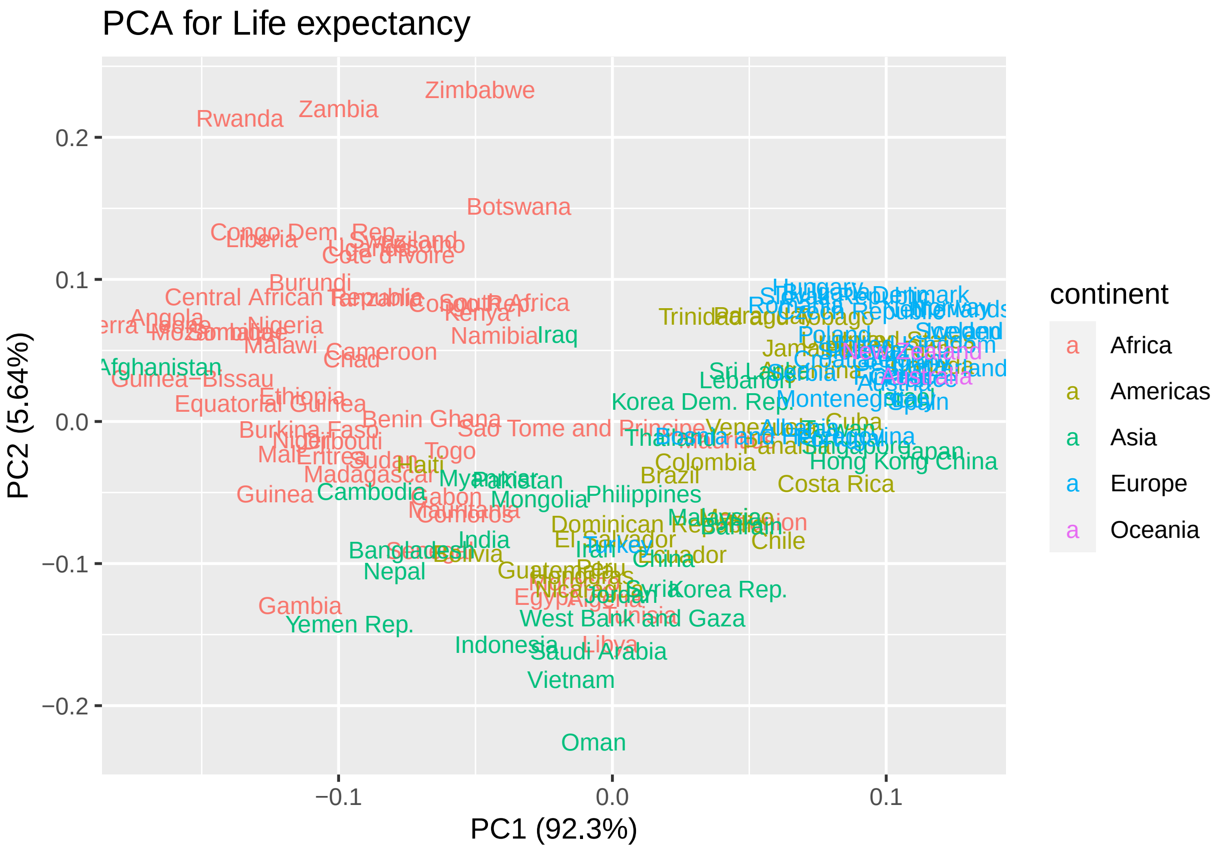
Plotting the countries' GDPs with the most extreme PC2 scores we can see that the countries with negative PC2 scores Taiwan and Guinea had large GDP growth and Iraq and Congo who have positive PC2 scores had GDP decline post 1982.

PCA on Life expectancy

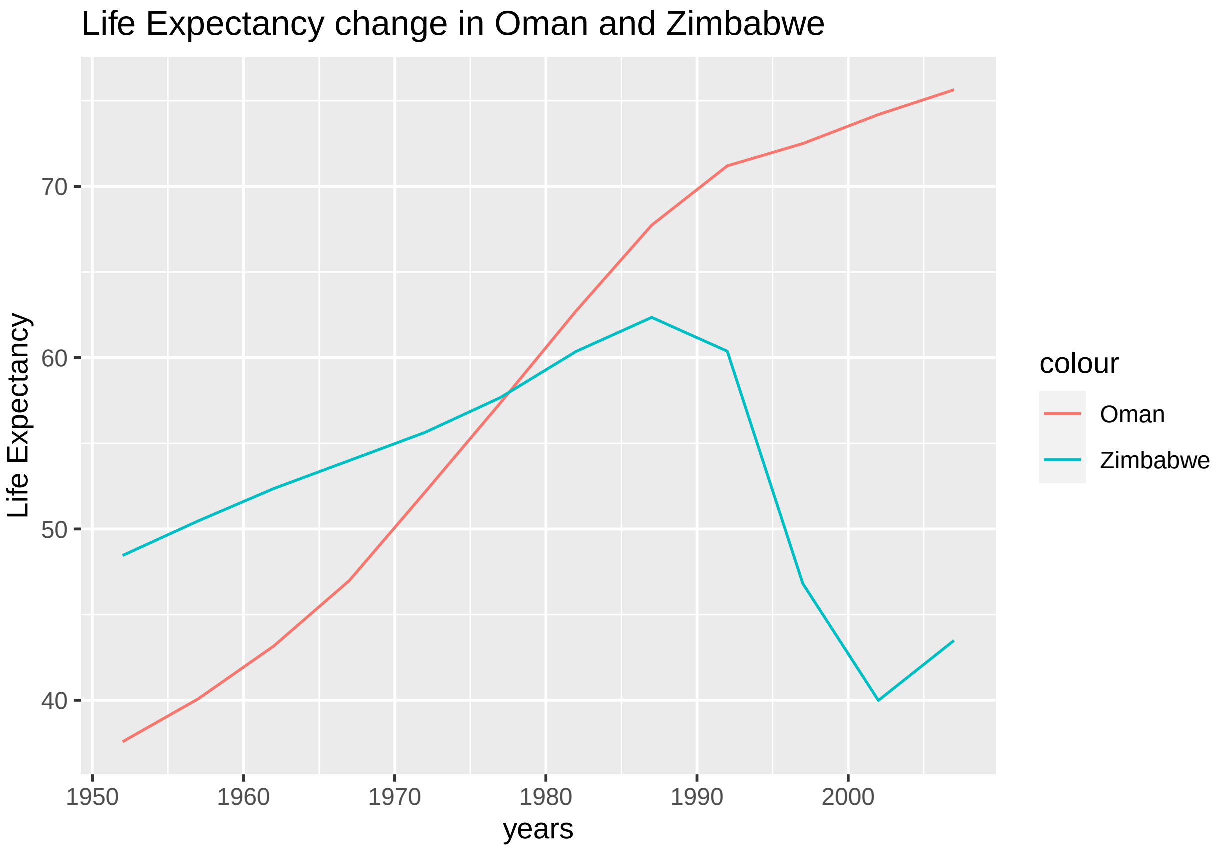
```
## Importance of components:
##                          PC1     PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.328 0.82287 0.3919 0.21989 0.12537 0.10995 0.08805
## Proportion of Variance 0.923 0.05643 0.0128 0.00403 0.00131 0.00101 0.00065
## Cumulative Proportion  0.923 0.97947 0.9923 0.99629 0.99760 0.99861 0.99926
##                            PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.06680 0.04373 0.03657 0.02820 0.02028
## Proportion of Variance 0.00037 0.00016 0.00011 0.00007 0.00003
## Cumulative Proportion  0.99963 0.99979 0.99990 0.99997 1.00000
```



97.9% of the variance is explained in the first two components so we will only keep these two components to compute PC scores. The variance explained by the components after this is 1.3% or less and thus there is no need to continue with these components.



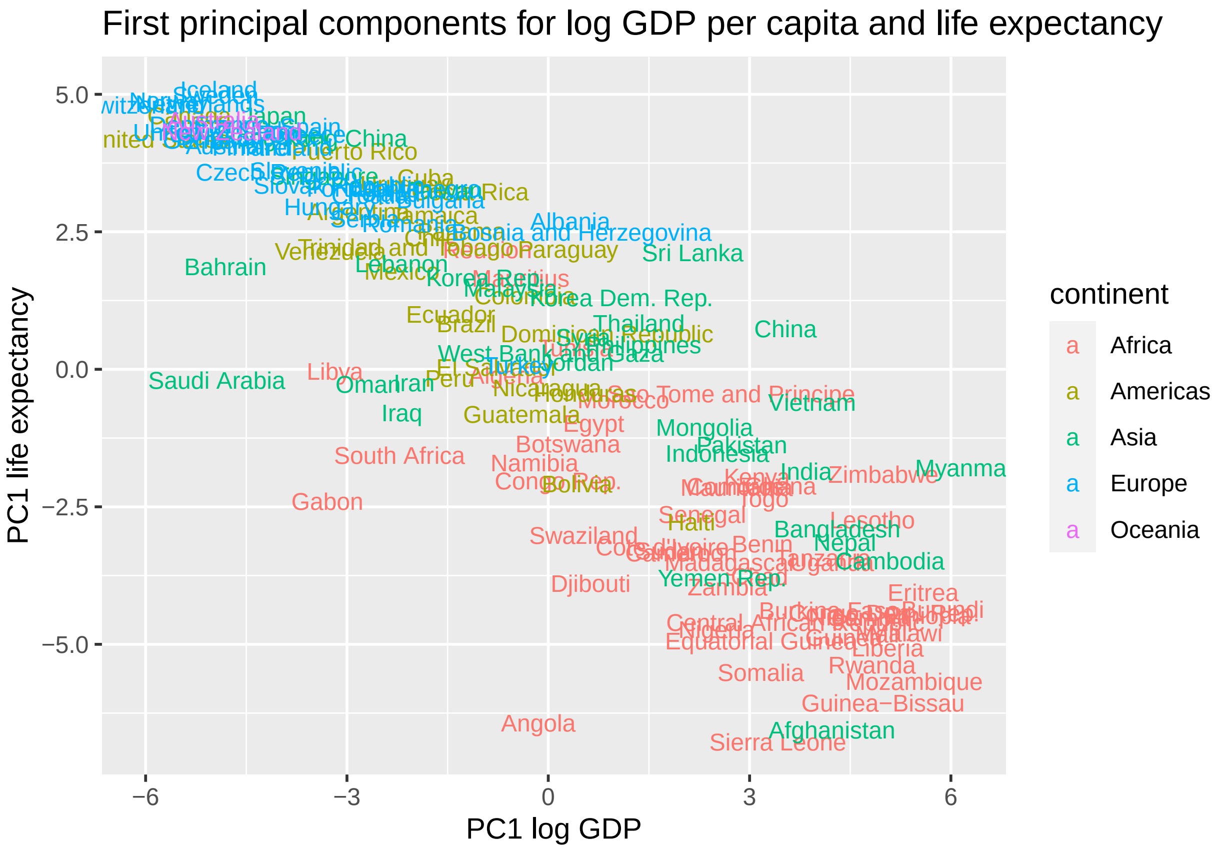
This PCA splits the life expectancy data into clusters roughly by continent. We see that African countries have negative PC1 scores and European and Oceanic countries have positive scores. I would interpret the first principal component as a measure of average life expectancy, where negative scores indicate low life expectancy and positive scores suggest high life expectancy. I interpret the second PC score as a measure for growth of life expectancy, negative scores indicate a increase in life expectancy and positive scores imply a decrease in life expectancy.



Plotting the countries with the most extreme PC2 value we see that Oman had a significant increase in life expectancy whereas Zimbabwe's life expectancy decreased.

PC1 GDP vs PC1 Life expectancy

Decided to continue with log GDP we plot the PC1 scores for GDP vs life expectancy,



Earlier we interpreted the first principal components for life expectancy and GDP as an average measure of each. Plotting these against each other creates a linear relationship indicating that large positive life expectancy PC scores and large negative GDP PC score implies that the higher a country's GDP the higher the life expectancy is. The countries are clustered into continents where Europe and Oceania have the countries with the highest combined GDP and life expectancy scores whereas Africa has the lowest.

Canonical Component Analysis

Next performing CCA on log GDP and life expectancy. First looking at the eta and psi coefficients,

Eta

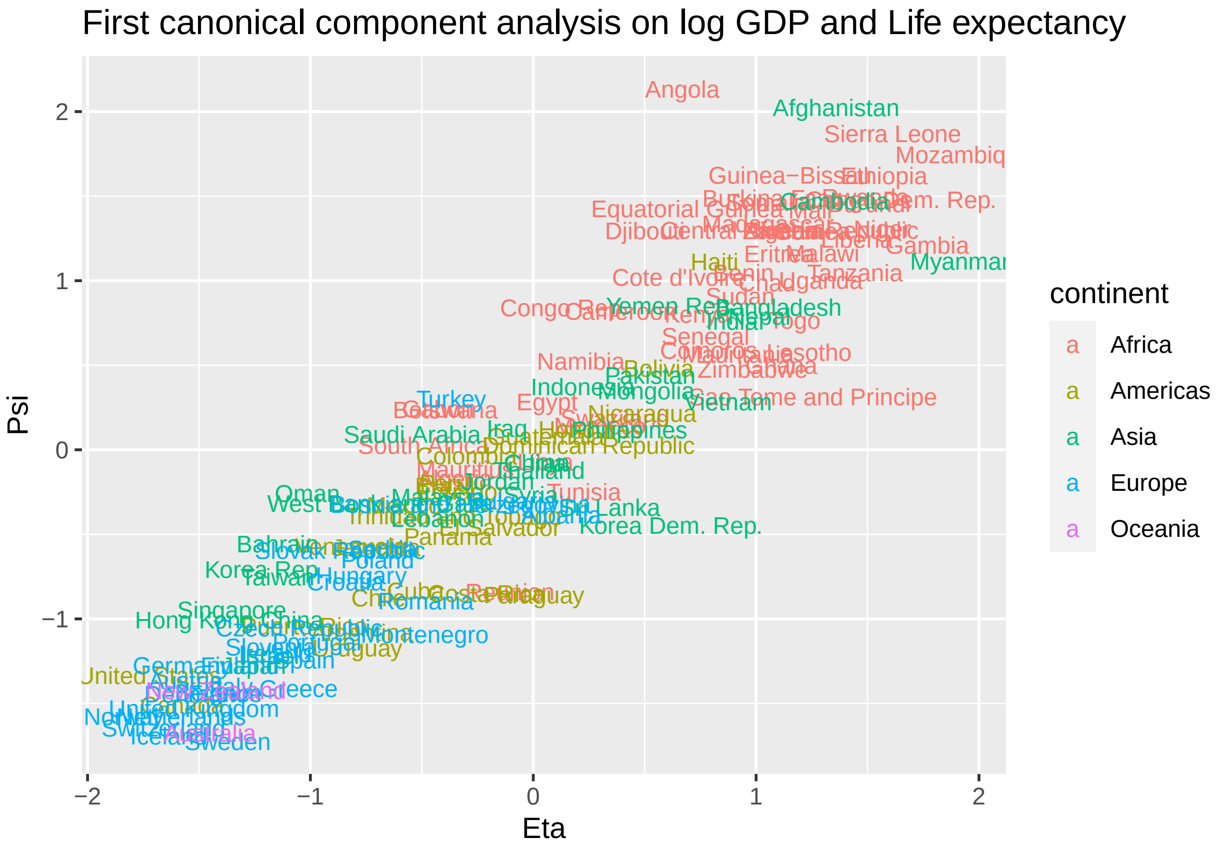
```
##       1952       1957       1962       1967       1972       1977       1982
##  0.8165860 -1.6623324  0.6398426 -0.1891125  0.1897061  0.3589767  0.4915870
##       1987       1992       1997       2002       2007
## -1.0905478  0.5365836 -1.0452679 -0.1744903  0.3482162
```

We see here that for eta there are large negative coefficients in the years 1957, 1987 and 1997.

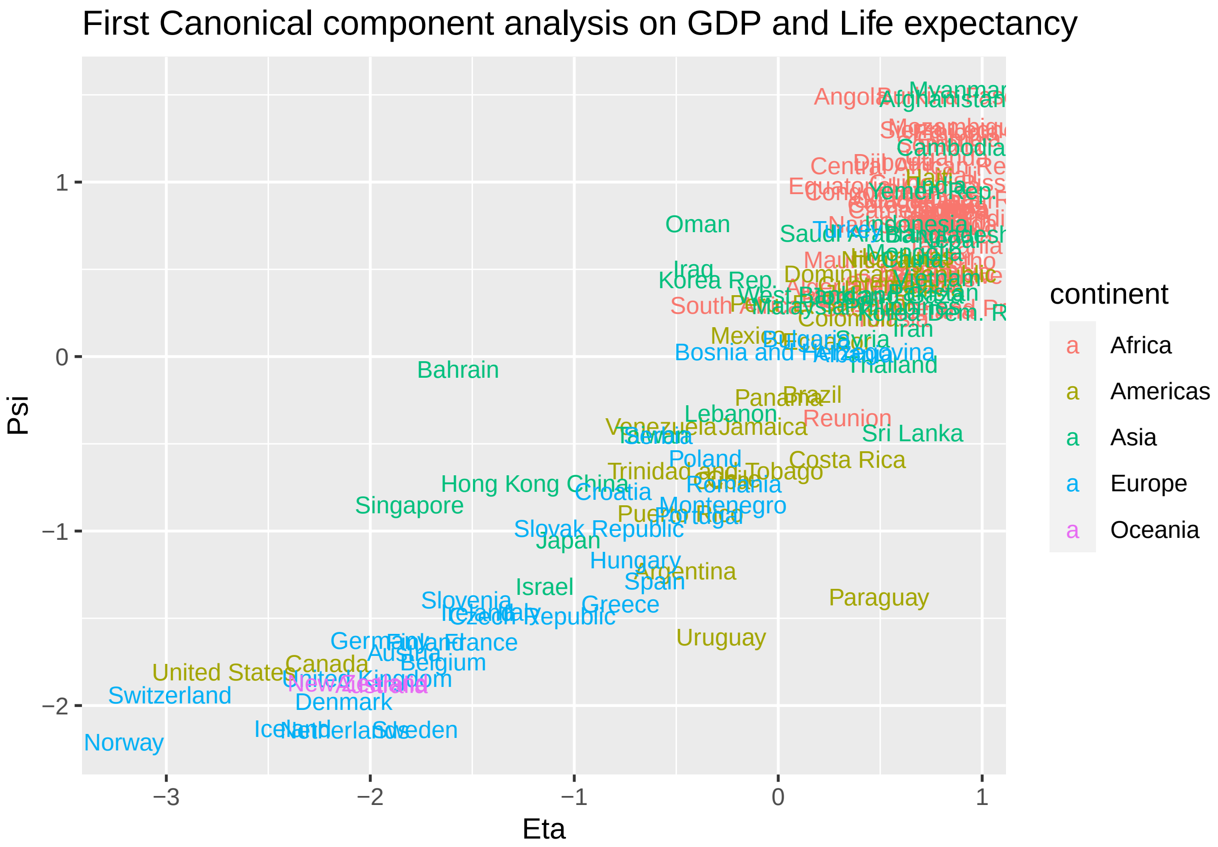
Psi

```
##         1952         1957         1962         1967         1972         1977
## -0.096441526 -0.019209519 -0.003757894  0.259642979 -0.244863502  0.055790699
##         1982         1987         1992         1997         2002         2007
##  0.117480442 -0.163825557  0.035611001 -0.047727753 -0.018807336  0.032752075
```

For psi the majority of the coefficients are negative.

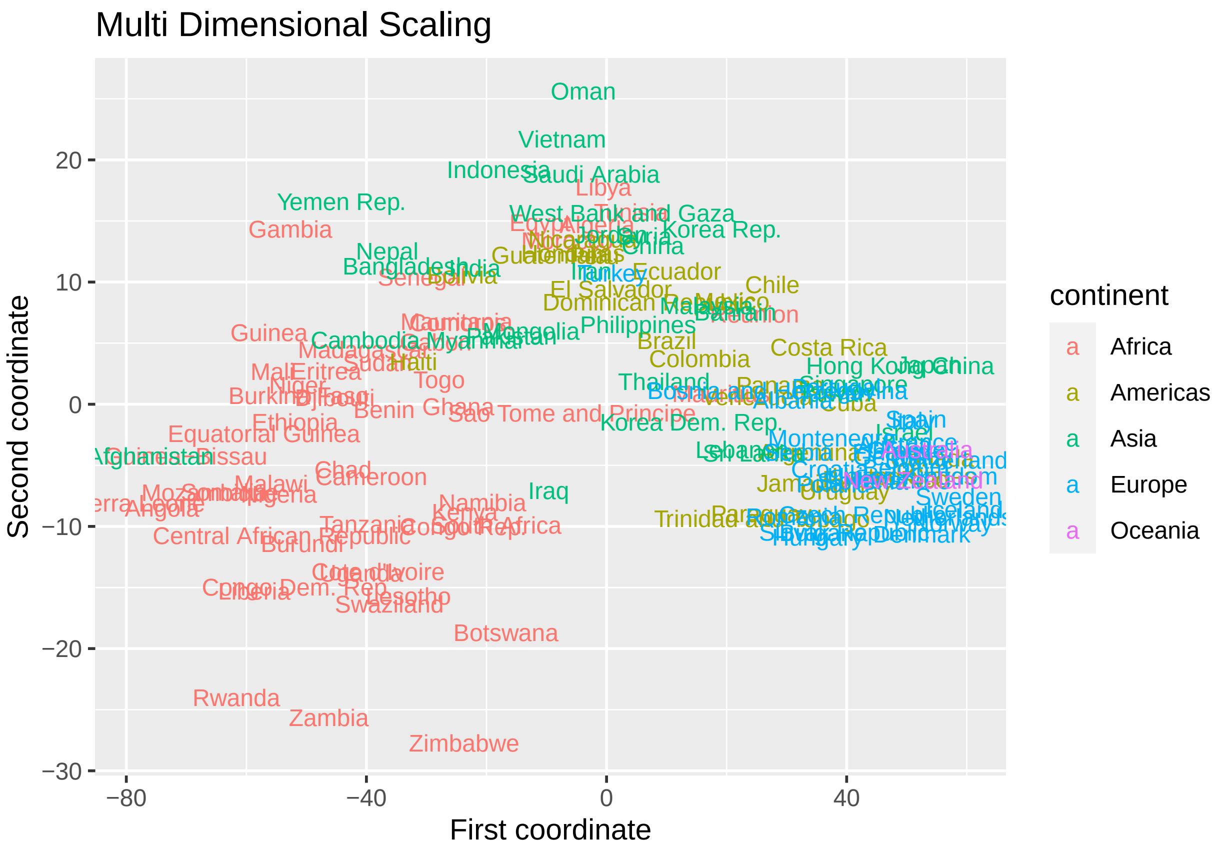


This plot shows us that there is a very strong positive correlation between log GDP and life expectancy. We see that the European and Oceanic countries have large negative eta and psi values indicating that they have large GDP contributions in the years 1957, 1987 and 1997 and that they had high life expectancy throughout. However African countries have lower GDPs and life expectancy.



Without applying the log transformation we see that there is a large cluster of countries in the top right corner of the plot, under this there is a positive correlation between GDP and life expectancy however it is weaker than when using log GDP. The plot suggests that a lot of Asian countries have similar relationships between GDP and life expectancy than African countries do but we see that the log transformation spreads this information out and differentiates the continents more. This aggressive clustering is most likely caused by the skewness in the GDP data.

Multi Dimensional Scaling



The MDS plot shows us a nice representation of similarities between countries, we see a cluster of European, Oceanic and some American countries on the right. This indicates that they have similar development levels, however on the left we see the African countries grouped together in a less dense cluster. This suggests that African countries have generally lower development levels as a whole but there is more variance in these levels, for example Libya is closer to in development to Saudi Arabia than Rwanda is. We see a distinctive difference in European and 1st world countries to African countries again and there is a bridge of Asian and American countries in between.

These results are very similar to what our principal and canonical component analysis found that European countries are more developed than African ones.

Linear Discriminant Analysis

Now we will perform linear discriminant analysis to predict continent based on GDP, life expectancy and population.

```
## [1] "Predictive accuracy is 50%"
```

The predictive accuracy of the test set is quite weak at 50%.

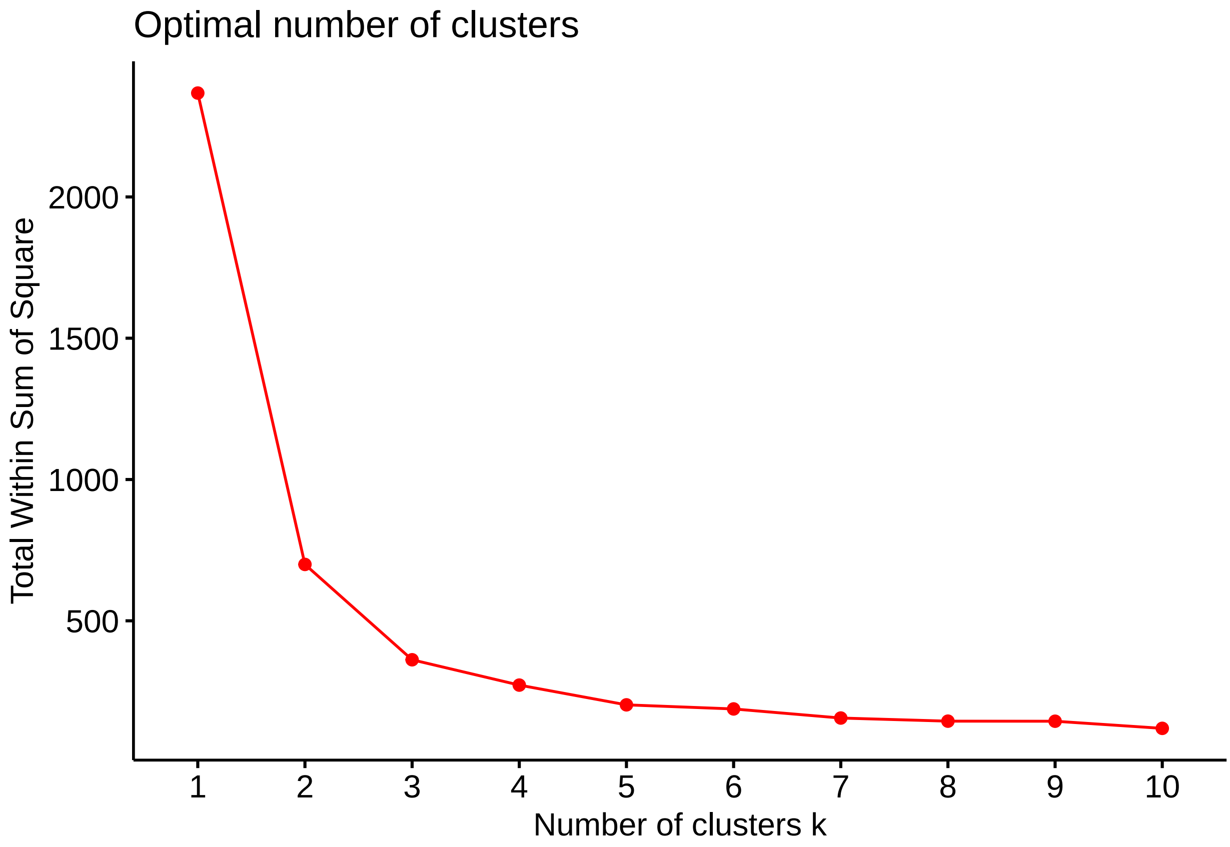
```
## 
##             Africa Americas Asia Europe Oceania
##   Africa      12        2    2      0       0
##   Americas     0        4    2      1       0
##   Asia         6        5    3      1       0
##   Europe       0        1    2      6       2
##   Oceania      0        0    0      1       0
```

The table shows us that 6 African countries misclassified as Asian; American countries are misclassified more than they are classified correctly often predicting them as Asian countries; similarly Asian countries are misclassified as African, American and European; European countries on the whole are classified correctly and the Oceanic countries are classified completely correctly.

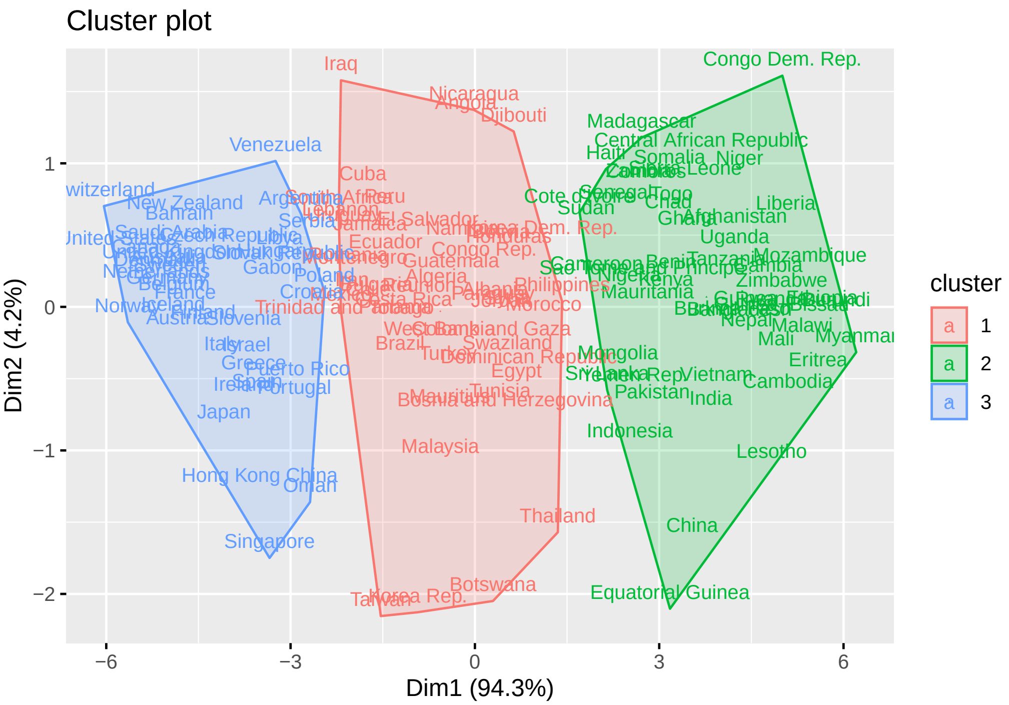
Clustering

GDP clustering

First we will take a look at K-means clustering, we must first assess the optimal number of clusters. I will be using log GDP data here instead of raw GDP because it reduces the skewness.

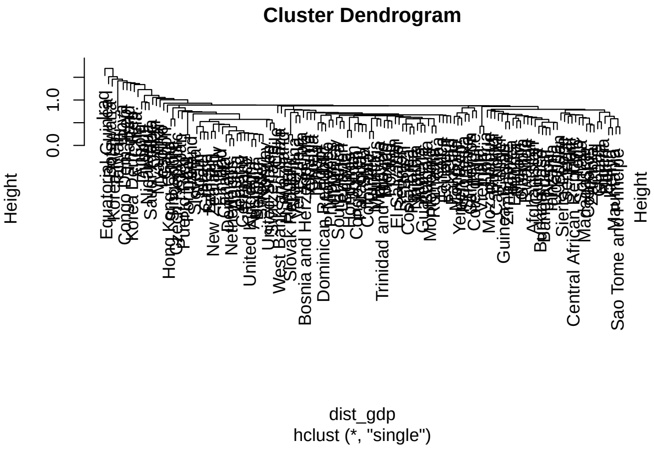


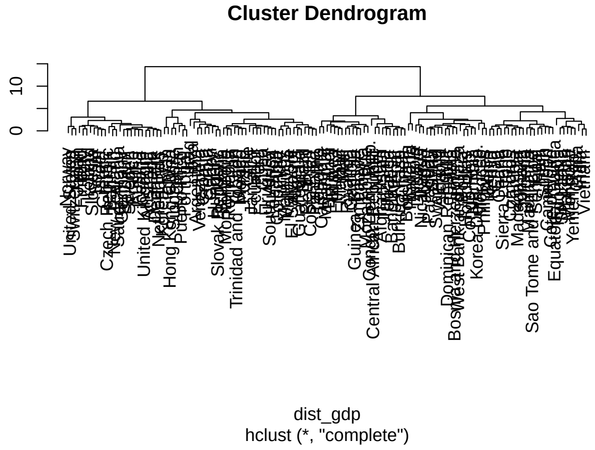
Analysing GDP using the elbow test we can see that there is a significant difference from 2-3 but after it levels off, so we will continue with 3 clusters.

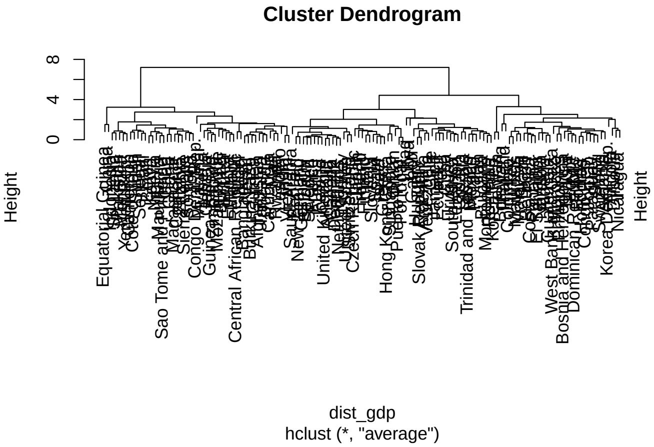


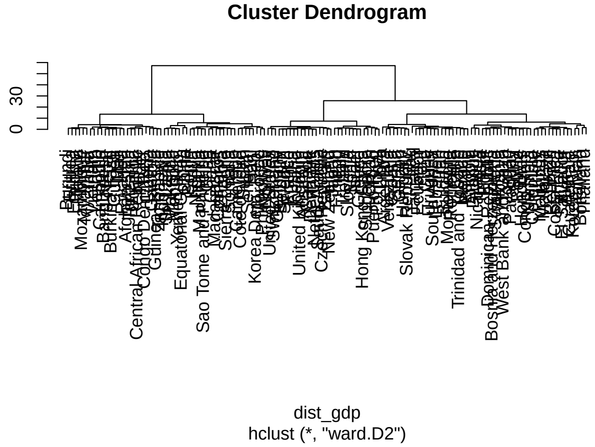
K-means clustering splits the GDP data into 3 distinctive clusters with the 3rd clustering primarily containing European countries, the second cluster is mainly Asian countries and the 1st is African countries.

Comparing this to hierarchical methods,



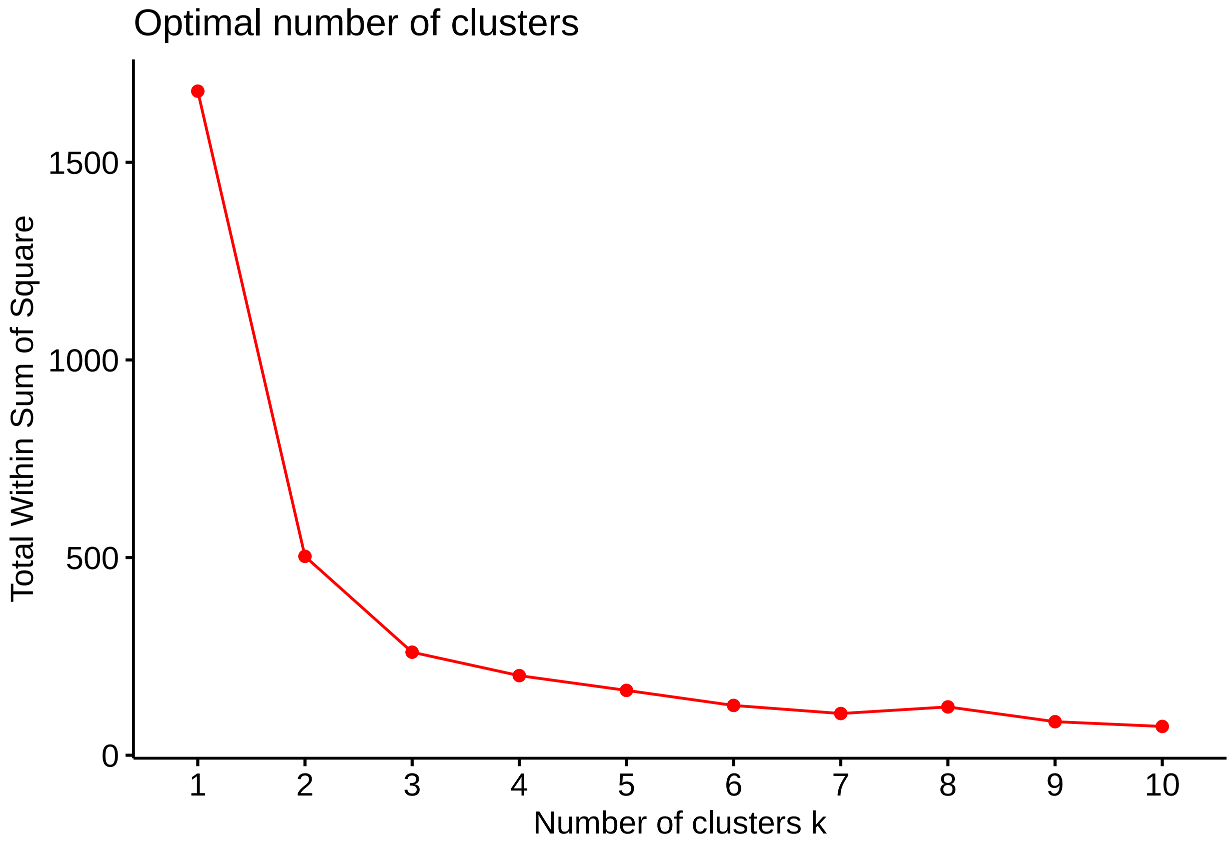




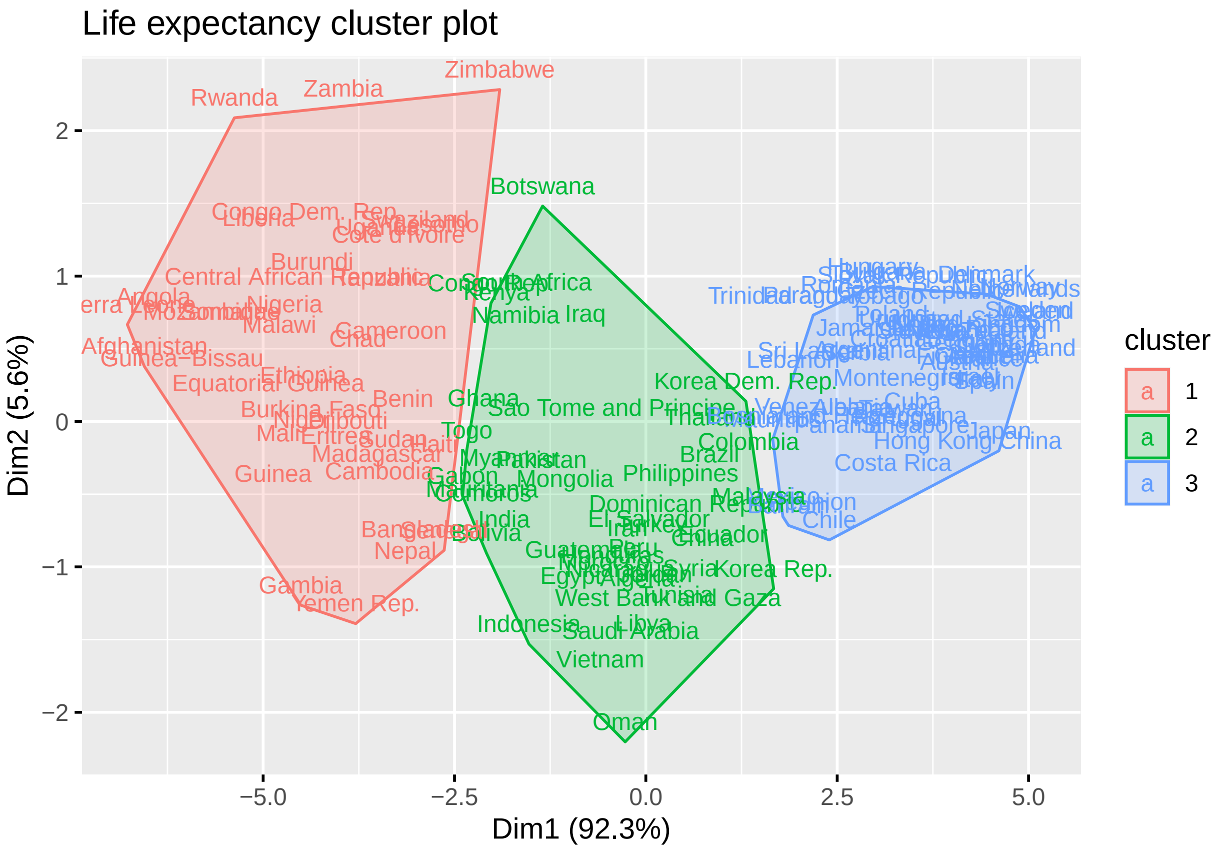


The cluster dendrograms show that complete linkage, group average and ward's method all find 3 clusters which agrees with the kmeans clustering. The natural interpretation of these clusters is that they are grouped by their level of economic development, we see the commonly known 1st world countries clustered in blue and the less developed countries in green. We could consider the red cluster as newly emerging countries that are bridging the gap from 3rd world status to 1st world.

Life expectancy clustering

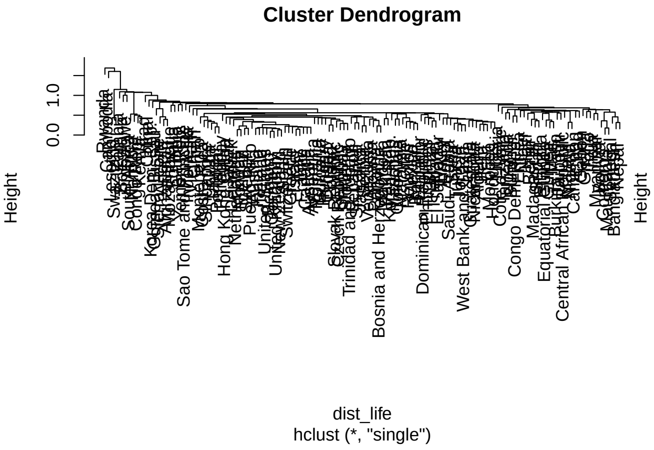


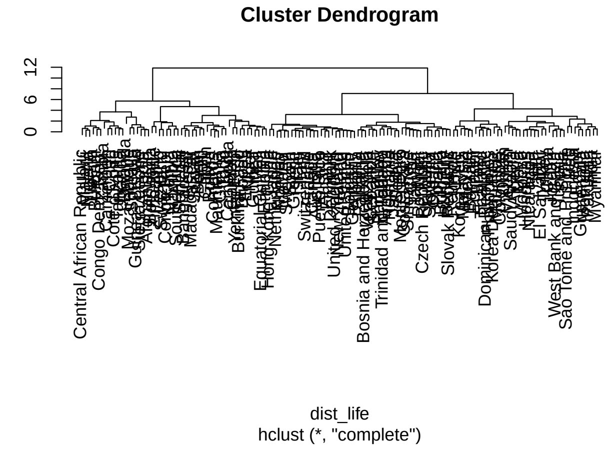
There is a significant change from 2-3 and after that each change in number of clusters is approximately the same, so continuing with 3 clusters.

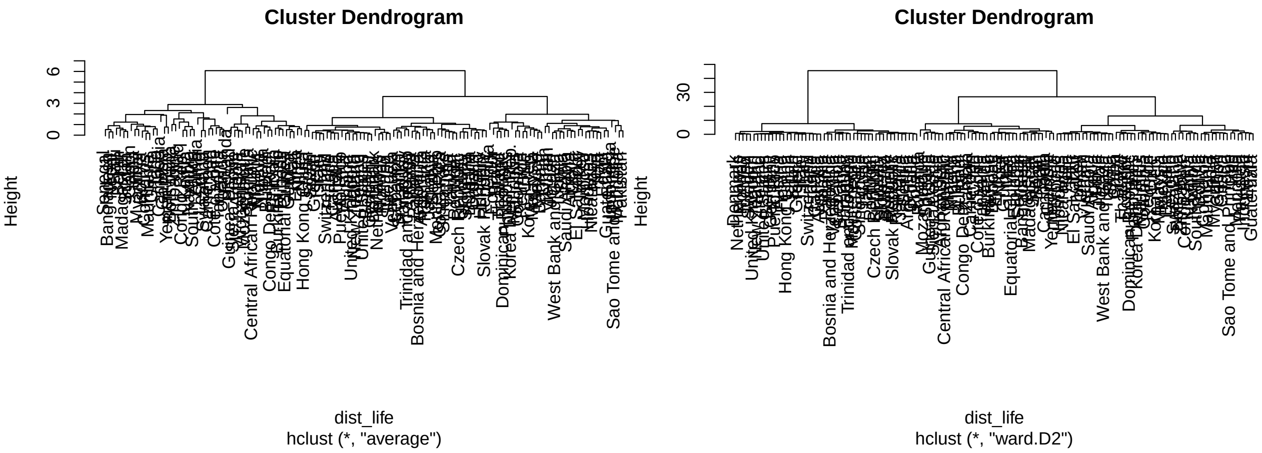


Kmeans clustering finds 3 distinct clusters with the 1st cluster mainly containing African countries, the second containing the upper bound of African countries and Asian countries, and the 3rd containing the more commonly known 1st world countries.

We now compare this to hierarchical clustering methods,







The hierarchical method are good at finding 3 clusters, complete linkage, group average and ward's method again find 3 clusters however single linkage is not able to find a nice solution. I would interpret these clusters as groups of countries that have similar levels of development contributing to life expectancy such as health care, environment, access to food and education. Cluster 1 is primarily made up of African countries who have the lowest life expectancy on average (found in EDA) whereas cluster 3 contains a mix of European, American and Asian countries.

I decided to exclude model based clustering from this analysis because there are 12 different years of data the graph would be very difficult to read.

Linear Regression

To fit a linear regression model for life expectancy in 2007 we need to first consider the effect of using raw GDP data or using log GDP data, from our EDA we know that the GDP data is positively skewed so I am inclined to use log GDP. Let's take a look,

OLS Regression

```
lm.ols <- lm(lifeExp_2007 ~ ., data=linear)
summary(lm.ols)$r.squared # log GDP
```

```
## [1] 0.6975618
```

```
lm.ols1 <- lm(lifeExp_2007~., data=linear_no_log)
summary(lm.ols1)$r.squared # raw GDP
```

```
## [1] 0.5065578
```

We see that the model using log GDP has a significantly higher R squared by 0.191 which suggests that log GDP leads to a more accurate linear model.

```
mean(lm.ols$residuals^2) # log GDP
```

```
## [1] 43.84091
```

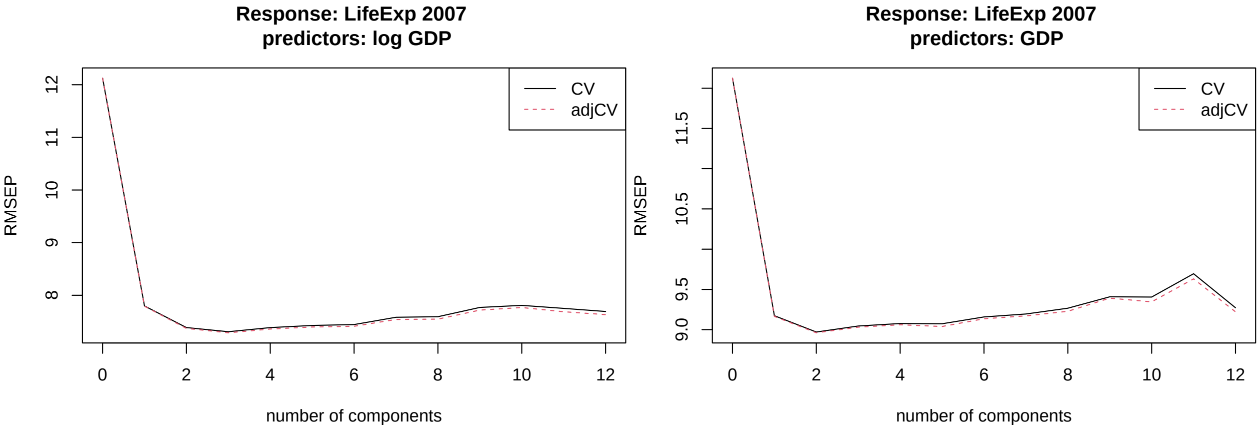
```
mean(lm.ols1$residuals^2) # raw GDP
```

```
## [1] 71.52851
```

Log GDP data yields a lower MSE and is suggests that it is more accurate.

Principal component regression

Since we are using 12 different years of GDP data we will have 12 components, so first we need to find the optimal number of components to use for our PCR,



Raw GDP suggests that using 2 components yields the lowest cross validation error whereas log GDP data suggests 3 components, after this as the number of components increases the error also increases.

```
summary(lm.pcr2)$r.squared # log GDP
```

```
## [1] 0.6674861
```

```
summary(lm.pcr3)$r.squared # raw GDP
```

```
## [1] 0.4660761
```

The R squared when using log GDP as the predictor is again significantly higher than when using the raw GDP data.

```
mean(lm.pcr2$residuals^2) # log GDP
```

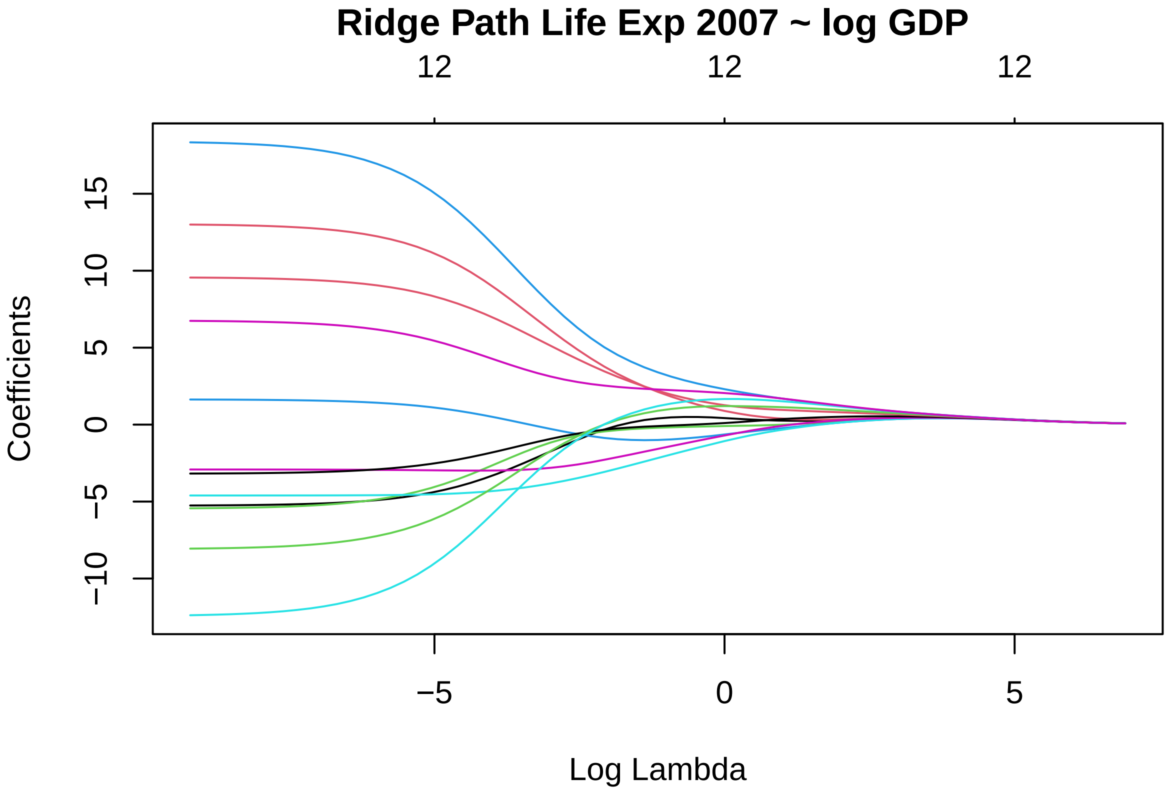
```
## [1] 48.20062
```

```
mean(lm.pcr3$residuals^2) # raw GDP
```

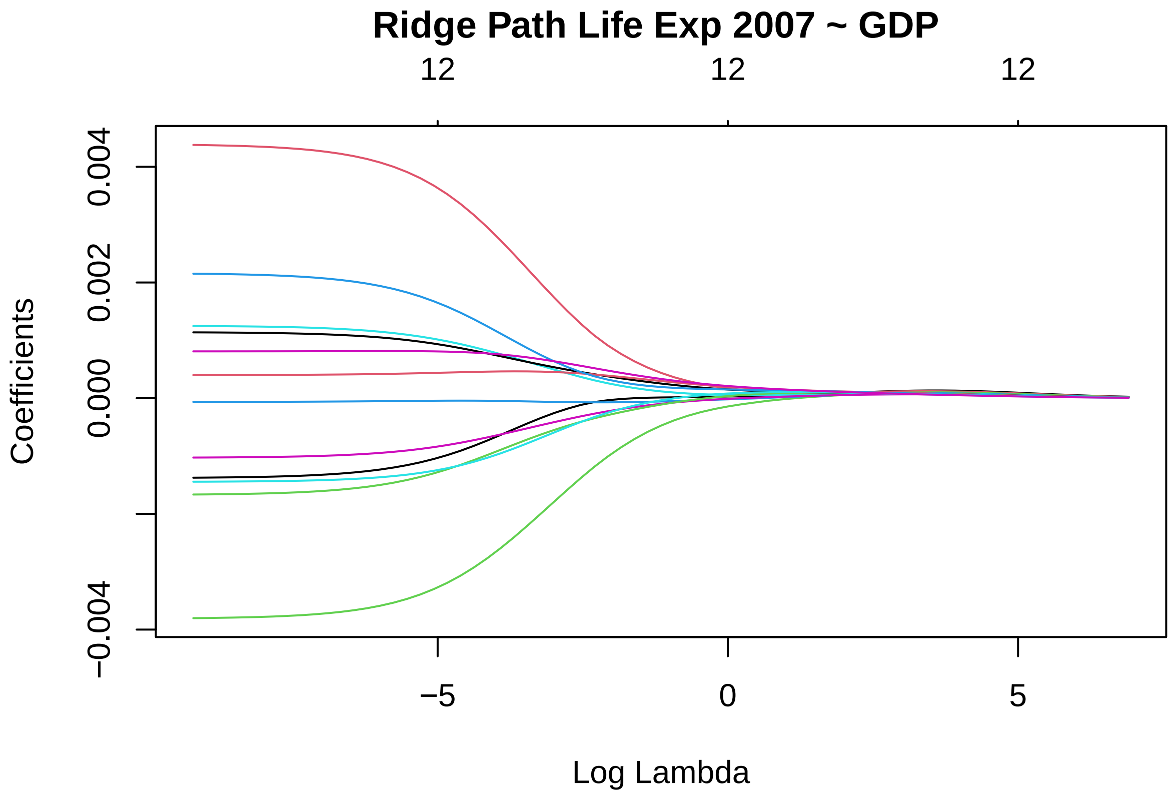
```
## [1] 77.39667
```

Investigating the mean square error shows us that using log GDP has a lower error and thus is more accurate than using raw GDP data.

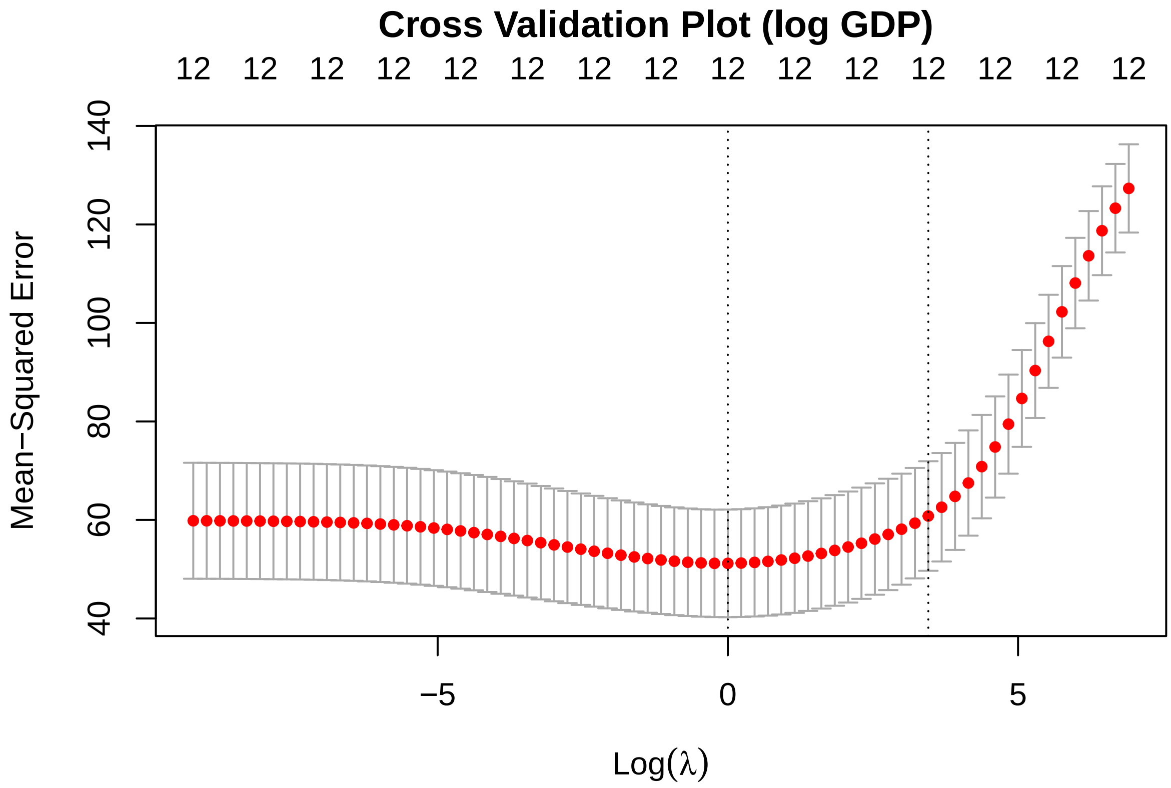
Ridge regression



For log GDP data we see that the ridge path is very noisy for negative log lambda values but as it increases it becomes more stable to tends to 0. Around log lambda = 0 there is a funnel of values which look symmetrical.



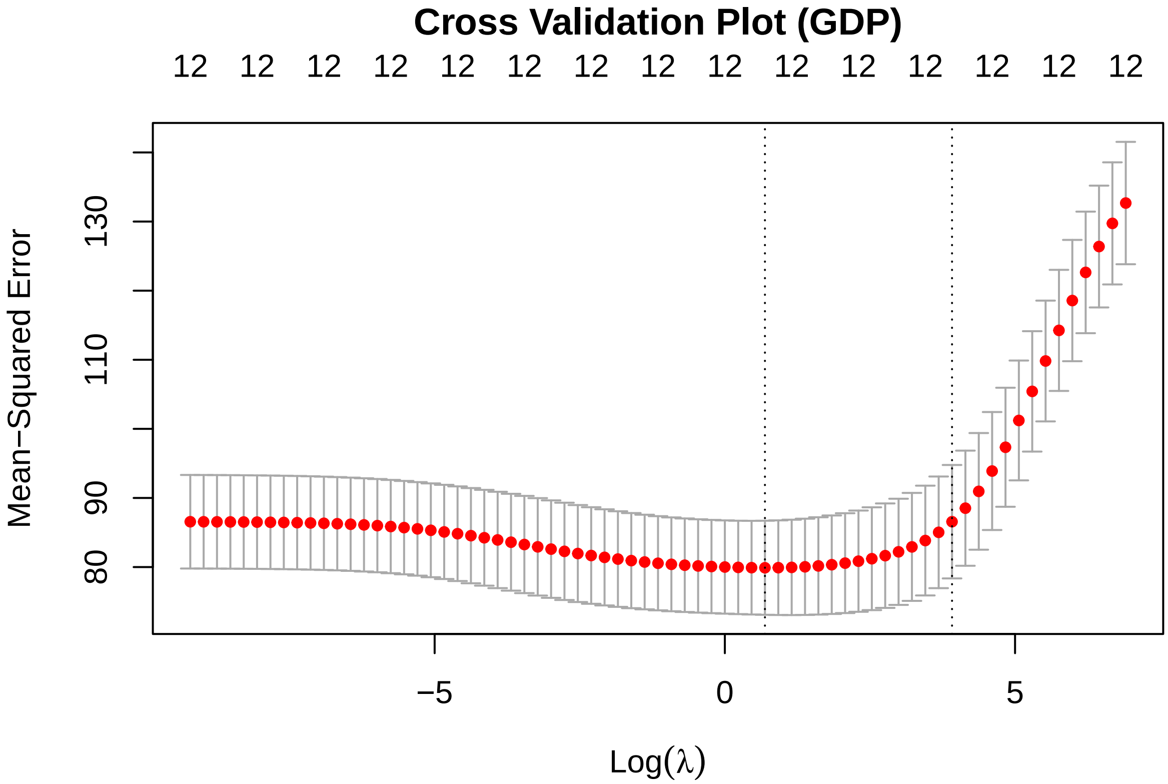
The model using raw GDP data tends to 0 a lot faster with coefficients joining at 0 at different lambda values. Now taking a look at the cross validation plots to determine which model has the lowerest error,



```
## lambda = 1
```

```
## MSE = 51.17825
```

This plot shows us that the value of lambda that yields the lowest MSE is lambda = 1 and the mean square error is 51.176. For the smallest lambda values the MSE is around 60 and as it starts to get closer to 1 it dips however after this it rapidly increases in inaccuracy.



```
## lambda = 1.995262
```

```
## MSE = 79.9016
```

The model using GDP data has a slightly higher lambda value giving minimum MSE of 1.99 however the error here is significantly larger at 79.9 and tells us that the raw GDP data is less accurate.

Comparing Regression methods

Now that we have formed all the models we need to assess which model is the best, below are tables comparing MSE and R squared values for models using raw and log GDP data respectively,

Raw GDP data

```
##                   OLS        PCR      Ridge
## MSE       71.5285070 77.3966668 79.9016013
## R squared  0.5065578  0.4660761  0.4678837
```

Log GDP data

```
##                  OLS        PCR      Ridge
## MSE       43.8409072 48.2006249 51.1782516
## R squared  0.6975618  0.6674861  0.6675334
```

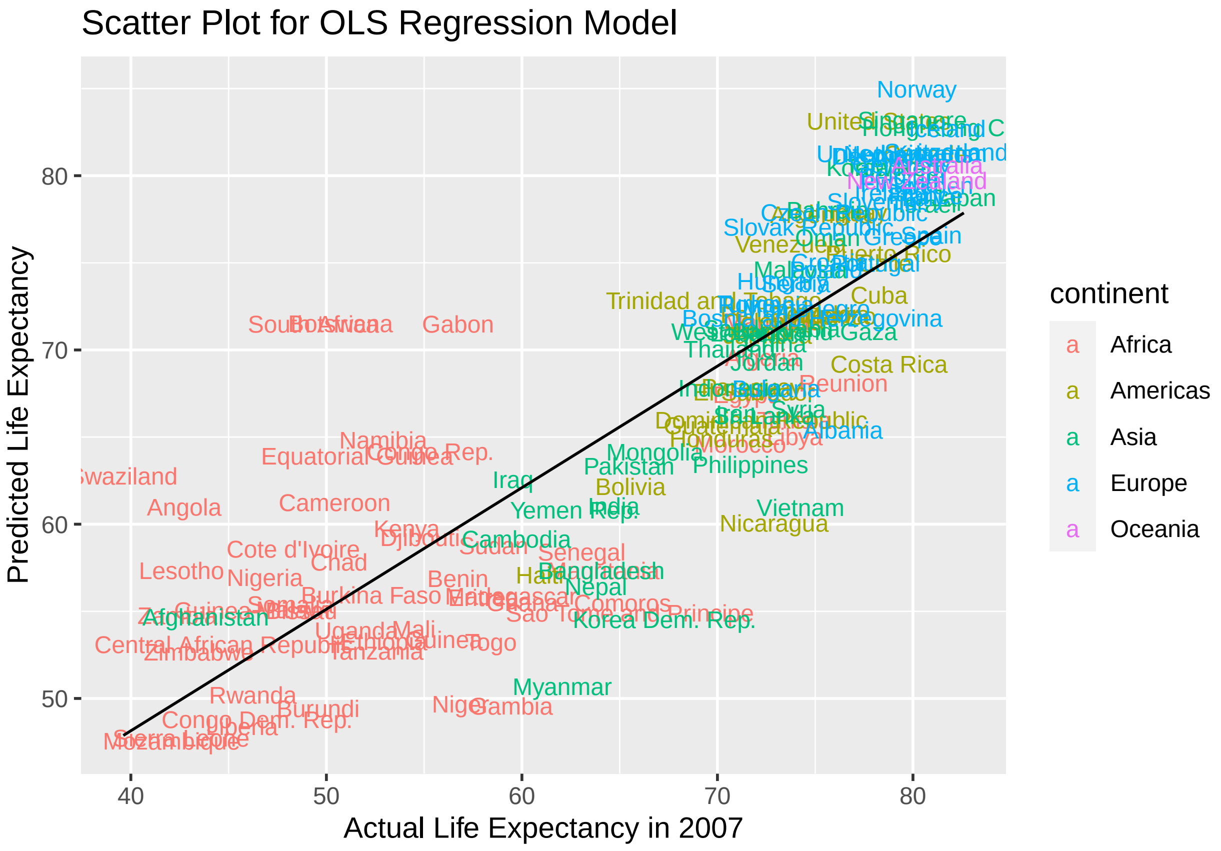
We can see that across all the models raw GDP data performs worse, in each model it has a significantly larger mean squared error and the R squared is lower. This leaves us with the models using log GDP data, PCR and Ridge regression models have similar R squareds with ridge giving slightly higher however the MSE for PCR is much better.

The best model for predicting 2007 life expectancy data is OLS regression model since it has the smallest MSE and highest R squared. Here are the coefficients for the model below,

```
coef(lm.ols)
```

```
##  (Intercept) log_gdp_1952 log_gdp_1957 log_gdp_1962 log_gdp_1967 log_gdp_1972
##     4.392543    -5.404620    13.302313    -5.629509     1.655186    -4.551923
## log_gdp_1977 log_gdp_1982 log_gdp_1987 log_gdp_1992 log_gdp_1997 log_gdp_2002
##    -2.882553    -3.357105     9.838070    -8.542309    19.130456   -13.121482
## log_gdp_2007
##     7.003482
```

Finally here is a scatter plot of our countries with our regression line layered over.


Scatter Plot for OLS Regression Model
Predicted Life Expectancy
Actual Life Expectancy in 2007
continent
a Africa
a Americas
a Asia
a Europe
a Oceania
Norway
United States
Singapore
Hong Kong
Korea
Australia
New Zealand
Japan
Israel
Slovenia
Czech Republic
Bahrain
Argentina
Slovak Republic
Venezuela
Oman
Greece
Spain
Puerto Rico
Malaysia
Croatia
Portugal
Hungary
Serbia
Trinidad and Tobago
Cuba
Montenegro
Bosnia and Herzegovina
West Bank and Gaza
Thailand
Jordan
Costa Rica
Reunion
Egypt
Syria
Dominican Republic
Guatemala
Honduras
Libya
Albania
Morocco
Philippines
Mongolia
Pakistan
Bolivia
South Africa
Botswana
Gabon
Namibia
Equatorial Guinea
Congo Rep.
Swaziland
Angola
Cameroon
Iraq
Yemen Rep.
India
Vietnam
Nicaragua
Kenya
Djibouti
Cambodia
Sudan
Senegal
Cote d'Ivoire
Chad
Lesotho
Nigeria
Benin
Haiti
Bangladesh
Nepal
Burkina Faso
Madagascar
Eritrea
Ghana
Comoros
Sao Tome and Principe
Korea Dem. Rep.
Zambia
Guinea-Bissau
Somalia
Malawi
Afghanistan
Uganda
Mali
Central African Republic
Ethiopia
Guinea
Togo
Zimbabwe
Tanzania
Myanmar
Rwanda
Burundi
Niger
Gambia
Congo Dem. Rep.
Liberia
Sierra Leone
Mozambique