Understanding Life Expectancy, GDP, and Population Dynamics: A Comprehensive Overview of UN Dataset

Amit Kumar

2024-04-25

Introduction

The United Nations (UN) dataset provides comprehensive data on various socio-economic factors for 141 countries spanning from 1952 to 2007. The dataset includes essential metrics such as life expectancy, GDP per capita and population size, which are crucial for evaluating the development and well-being of nations over time.

In this report, we aim to conduct a thorough analysis of the UN dataset using a variety of statistical methods. Emphasising on exploratory data analysis (EDA) to gain insights into the overall trends and patterns. Subsequently, we will apply different statistical techniques such as Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), Multidimensional Scaling (MDS), Linear Discriminant Analysis (LDA), and Clustering to uncover hidden relationships, structures, and groupings within the dataset.

Through this analysis, we seek to contribute to a deeper understanding of the socio-economic dynamics across nations over time, ultimately striving towards the advancement of global development and welfare.

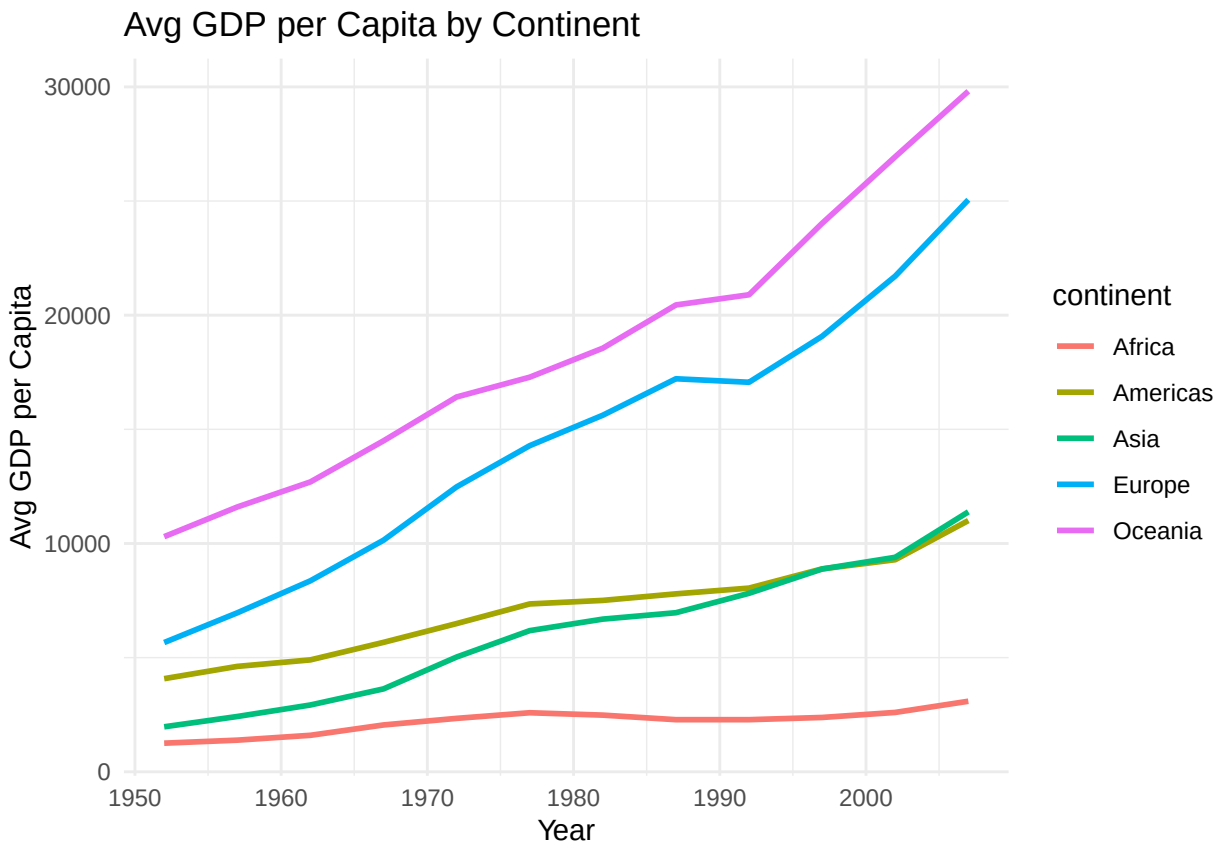
Exploratory Data Analysis

```
library(dplyr)
library(tidyr)
library(tibble)

#Pre-processing the data

gdp$continent <- UN$continent
lifeExp$continent <- UN$continent
popn$continent <- UN$continent

gdp_melt <- gdp %>%
  rownames_to_column(var = "country") %>%
  pivot_longer(cols = -c(country, continent), names_to = "Year",
               values_to = "GDP") %>% mutate(Year = as.integer(Year))

lifeExp_long_melt <- lifeExp %>%
  rownames_to_column(var = "country") %>%
  pivot_longer(cols = -c(country, continent), names_to = "Year",
               values_to = "Life_Expectancy") %>% mutate(Year = as.integer(Year))

popn_long_melt <- popn %>%
  rownames_to_column(var = "country") %>%
  pivot_longer(cols = -c(country, continent), names_to = "Year",
```

```
                values_to = "Population") %>% mutate(Year = as.integer(Year))
```

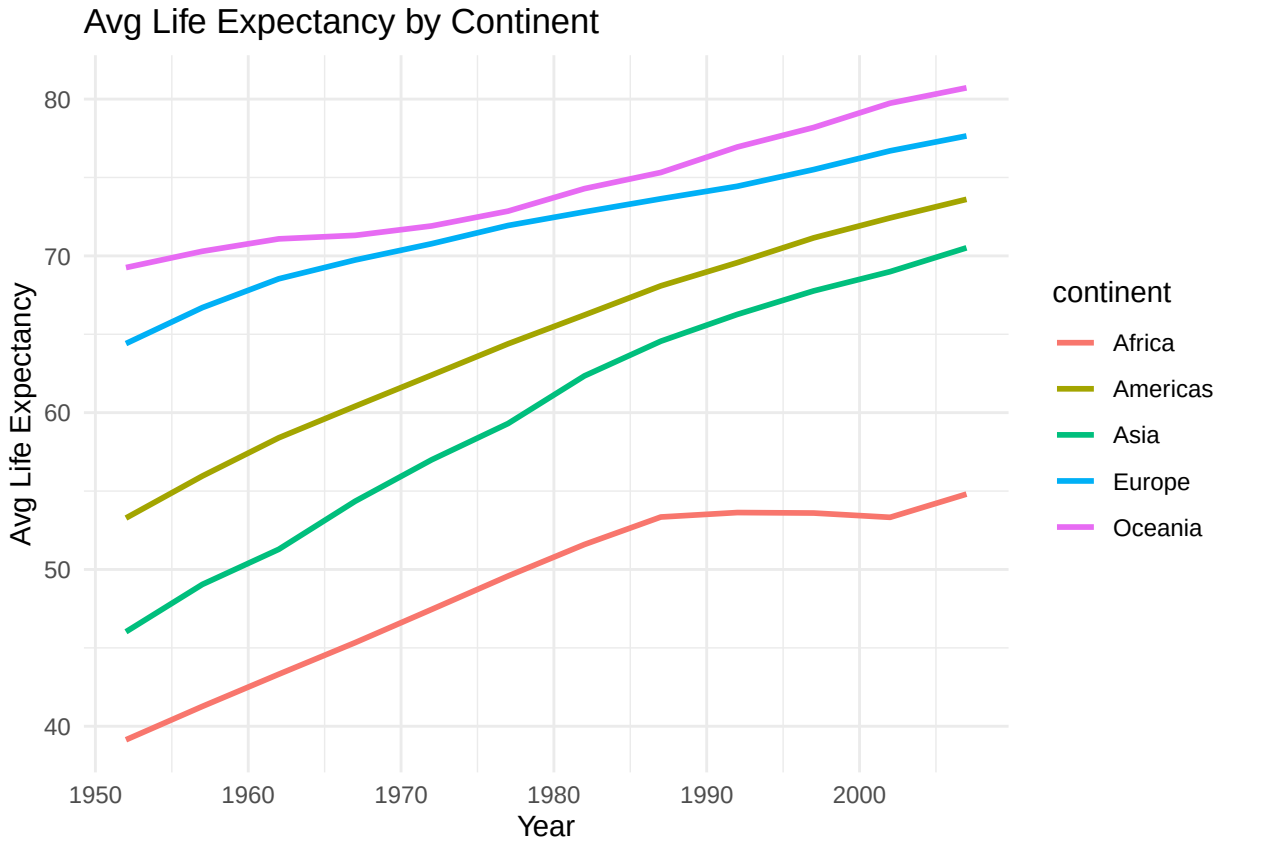
When employing a line plot to illustrate the avg gdp per capita by continents, it demonstrates that, over the specified period, there was a substantial rise in the average GDP per capita across all continents. In Africa, the average GDP per capita remained relatively low and experienced minimal growth in contrast to other continents. On the other hand, the Americas demonstrated a continuous ascent in GDP per capita, signifying consistent economic growth. Asia exhibited substantial expansion, characterized by a sharp incline indicative of rapid economic development. Europe showed consistent economic progress, contributing to its continually rising GDP per capita. Oceania stood out by having the highest average GDP per capita throughout the period, maintaining a consistently strong economic performance across its countries.

```
library(ggplot2)
library(GGally)

# Plot average GDP per continent over years
ggplot(gdp_melt, aes(x = Year, y = GDP, color = continent)) +
  geom_line(stat = "summary", fun = "mean", size = 1) +
  labs(x = "Year", y = "Avg GDP per Capita", title = "Avg GDP per Capita by Continent") +
  theme_minimal()
```

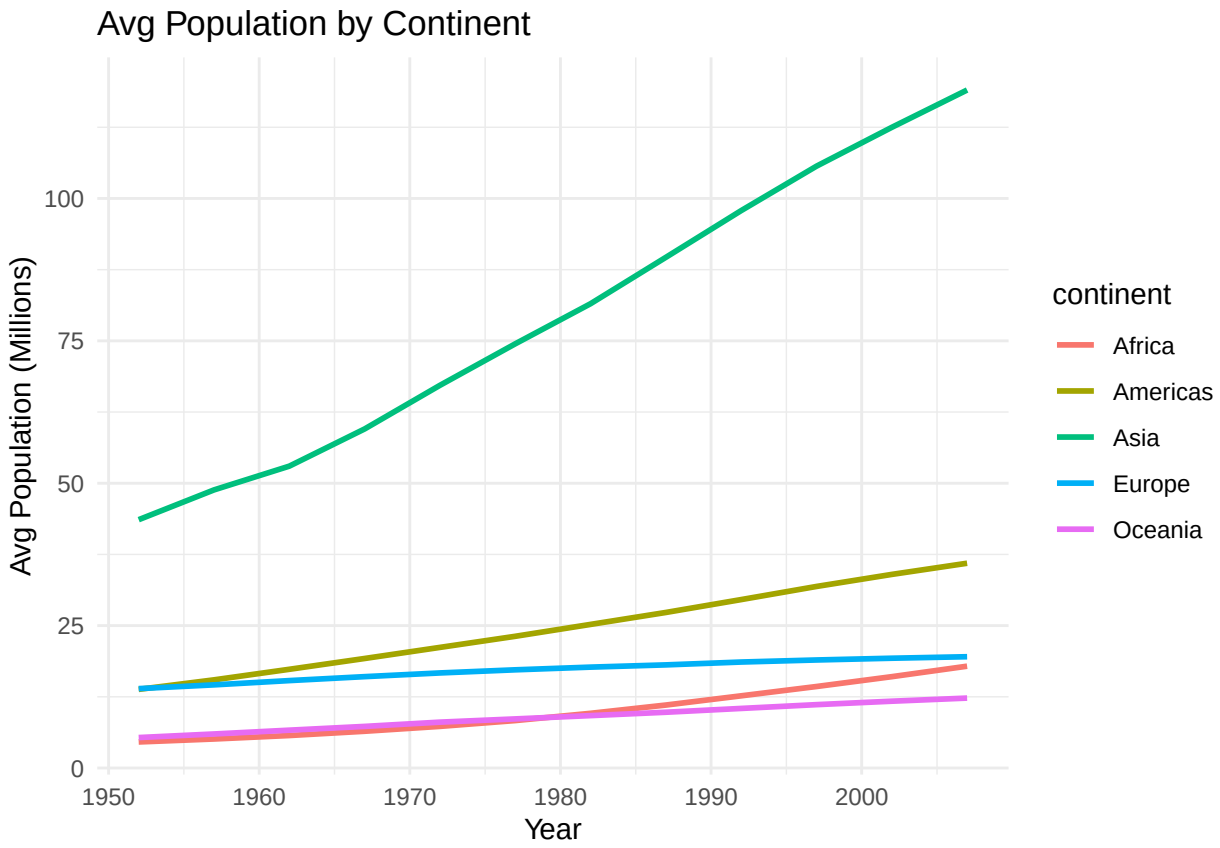


The line graph depicts a significant trend that extends from 1950 to 2000, displaying a steady increase in average life expectancy across all the continents. By the year 2000, Oceania and Europe had the highest life expectancies, both around 80 years, while the Americas and Asia followed closely behind, with roughly 70 years of life expectancy. On the other hand, Africa showed the lowest average life expectancy, and the rate of increase in life expectancy seems to have slowed down in Africa by the year 2000. This pattern indicates a positive relationship between a continent's level of development and its life expectancy.

```r
# Plot average life expectancy per continent over years
ggplot(lifeExp_long_melt, aes(x = Year, y = Life_Expectancy, color = continent)) +
  geom_line(stat = "summary", fun = "mean", size = 1) +
  labs(x = "Year", y = "Avg Life Expectancy", title = "Avg Life Expectancy by Continent") +
  theme_minimal()
```

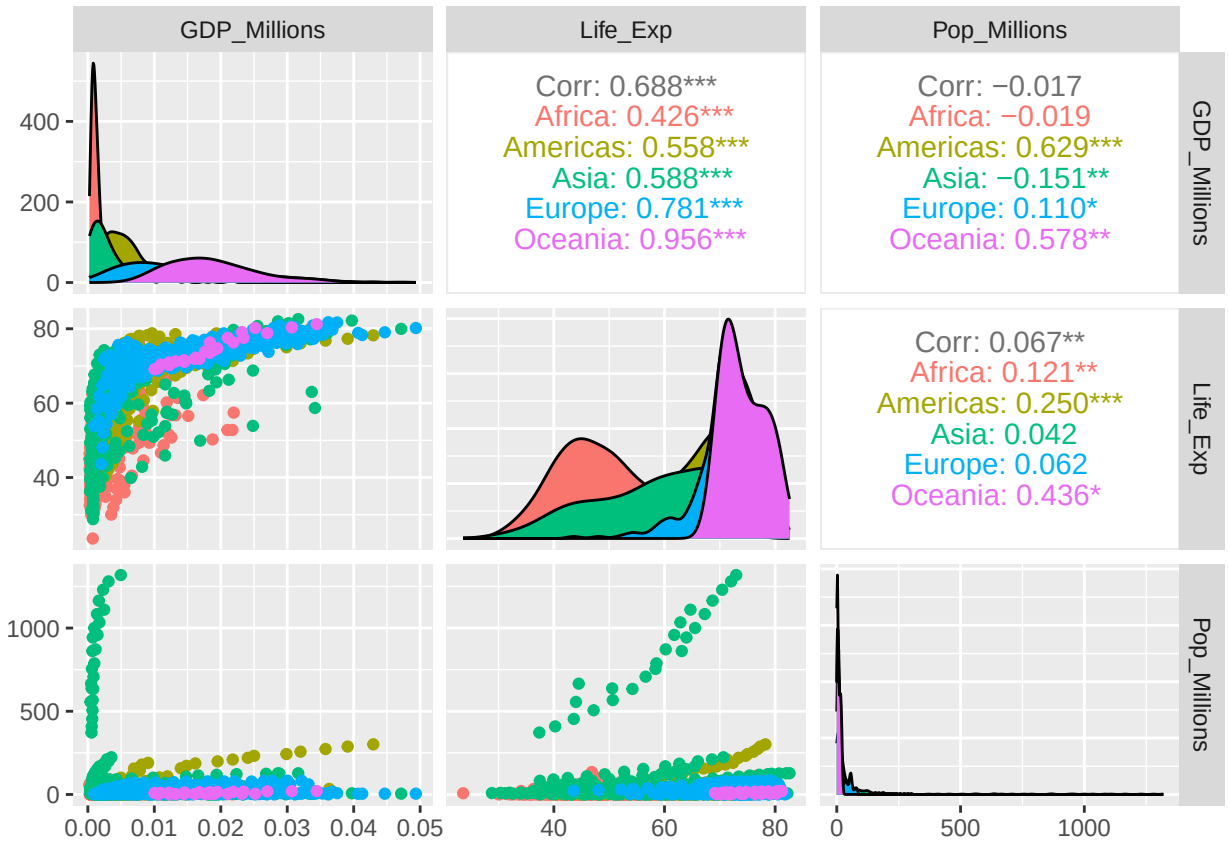


Asia displays the highest average population, commencing just below 50 million in 1950 and exhibiting a sharp ascent to surpass 100 million by 2000, indicating substantial population growth over the said period. Africa demonstrates moderate growth, starting slightly above zero and reaching almost 25 million by 2000. Conversely, the Americas exhibit steady growth in population, attaining nearly 30 million by 2000, while Europe experiences minimal growth and variation in population. Oceania sustains the lowest population levels over the entire period.



Following the analysing of the graph, it becomes evident that there exists a strong correlation between the life expectancy of the population and the GDP of the continent. As the GDP of a continent grows, the life expectancy also tends to increase. Furthermore, there is a discernible correlation between population size and the GDP of the continent. This correlation is particularly evident in the case of the Asian and Americans continents, where an increase in population is accompanied by a corresponding increase in GDP.

```
plot <- data.frame(
  GDP_Millions = (gdp_melt$GDP)/1e6,
  Life_Exp = lifeExp_long_melt$Life_Expectancy,
  Pop_Millions = (popn_long_melt$Population)/1e6
)
plot <- ggpairs(
  data = plot,
  mapping = aes(colour = gdp_melt$continent),
  label_size = 1
) +
theme(
  plot.title = element_text(size = 9),  # Adjust title size
  axis.text = element_text(size = 9),    # Adjust axis text size
  axis.title = element_text(size = 10)    # Adjust axis title size
)
plot
```



Principal component analysis

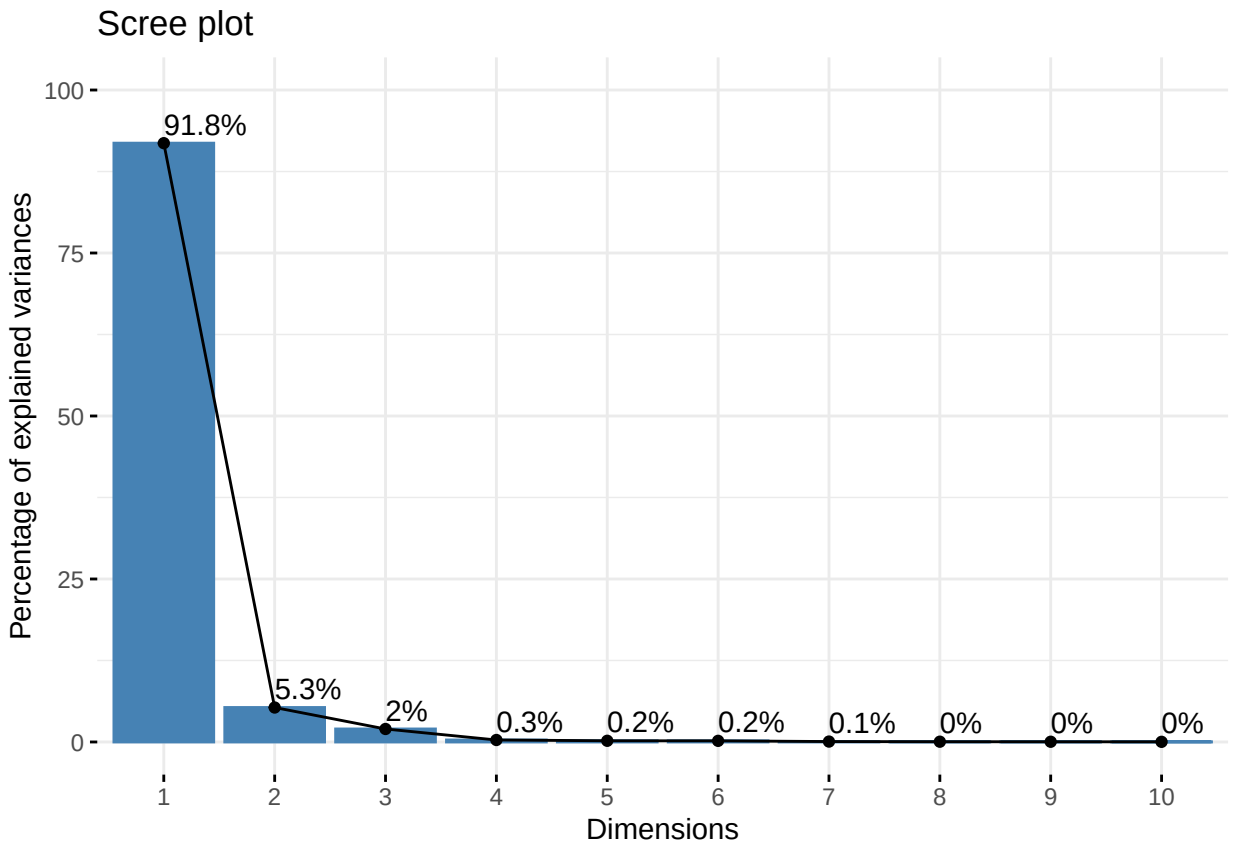
We will perform Principal Component Regression (PCR) analysis on three variables namely GDP, life expectancy, and population, and subsequently assess its results. Based on our exploratory data analysis, we intend to utilize a correlation matrix in our PCR analysis. This is necessary since the population sizes across countries vary considerably, with some possessing high populations and others possessing low populations. Furthermore, our analysis disclosed that African countries have the lowest average life expectancy, while Oceania exhibits the highest. Similarly, the GDP per capita in Asian countries is approximately 30,000, while African countries have a GDP per capita of around 2,000.

```
pca_gdp <- prcomp(gdp, scale = TRUE)
summary(pca_gdp)
```

```
## 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
```

Principal Component Regression (PCR) and scree plot reveals that the first two principal components PC1 and PC2 of GDP data account for approximately 97% of the total variance.

```
library(factoextra)
fviz_eig(pca_gdp, addlabels = TRUE, ylim = c(0, 100))
```

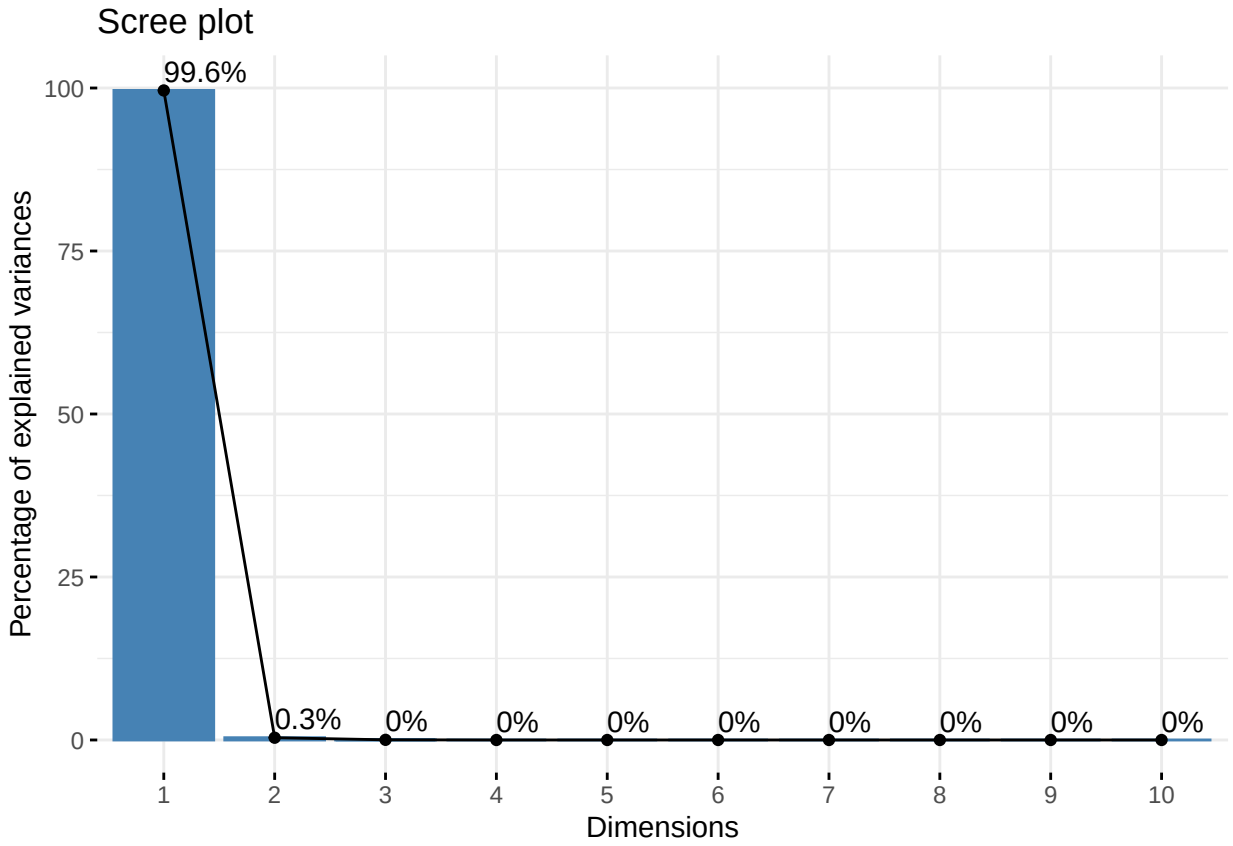



```
pca_popn <- prcomp(popn, scale = TRUE)
summary(pca_popn)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5      PC6      PC7
## Standard deviation     3.4576 0.20435 0.05580 0.01224 0.01163 0.005631 0.004665
## Proportion of Variance 0.9962 0.00348 0.00026 0.00001 0.00001 0.000000 0.000000
## Cumulative Proportion  0.9962 0.99971 0.99997 0.99998 0.99999 1.000000 1.000000
##                            PC8      PC9    PC10      PC11      PC12
## Standard deviation     0.002513 0.001916 0.00152 0.0008781 0.0008349
## Proportion of Variance 0.000000 0.000000 0.00000 0.0000000 0.0000000
## Cumulative Proportion  1.000000 1.000000 1.00000 1.0000000 1.0000000
```

Principal Component Regression (PCR) and scree plot for population data reveals that the first principal components PC1 of population account for approximately 99% of the total variance.

```
fviz_eig(pca_popn, addlabels = TRUE, ylim = c(0, 100))
```

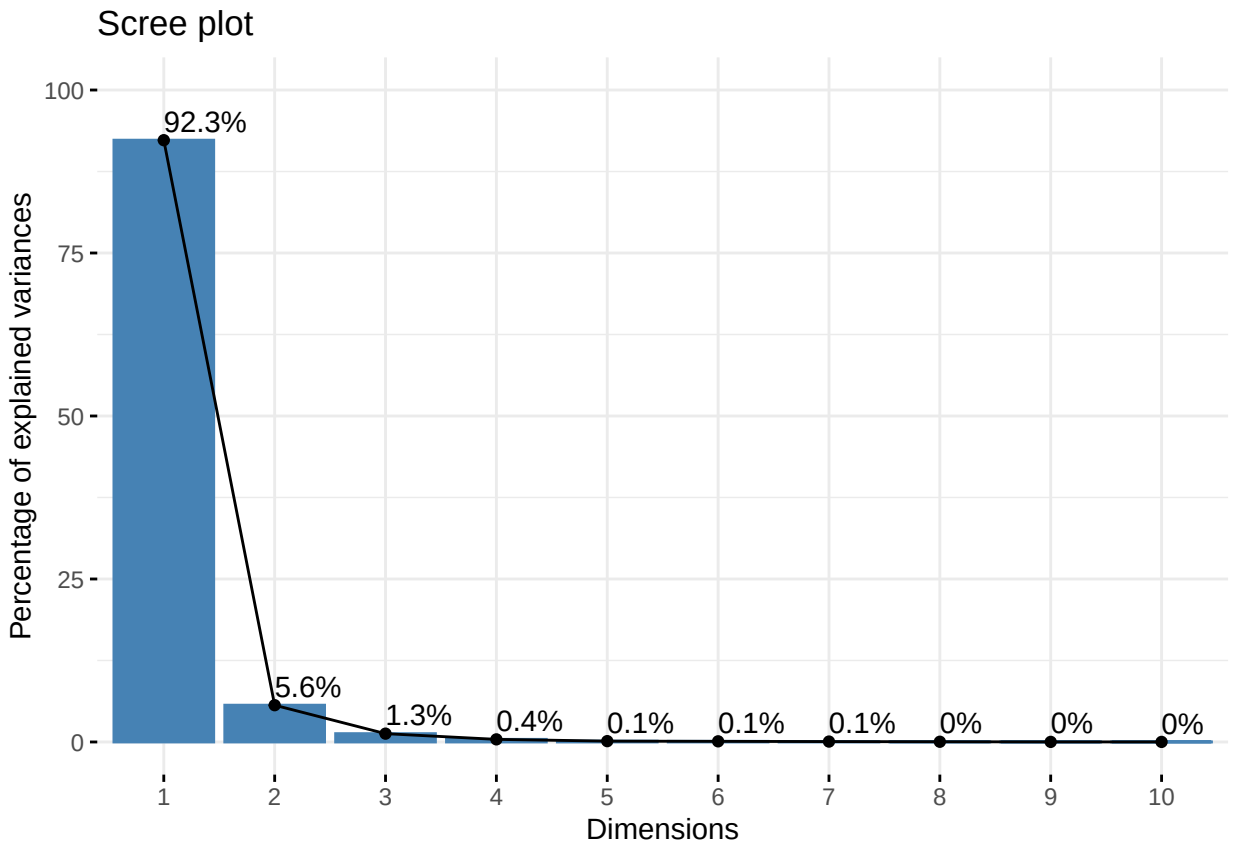



```
pca_lifeExp <- prcomp(lifeExp, scale = TRUE)
summary(pca_lifeExp)
```

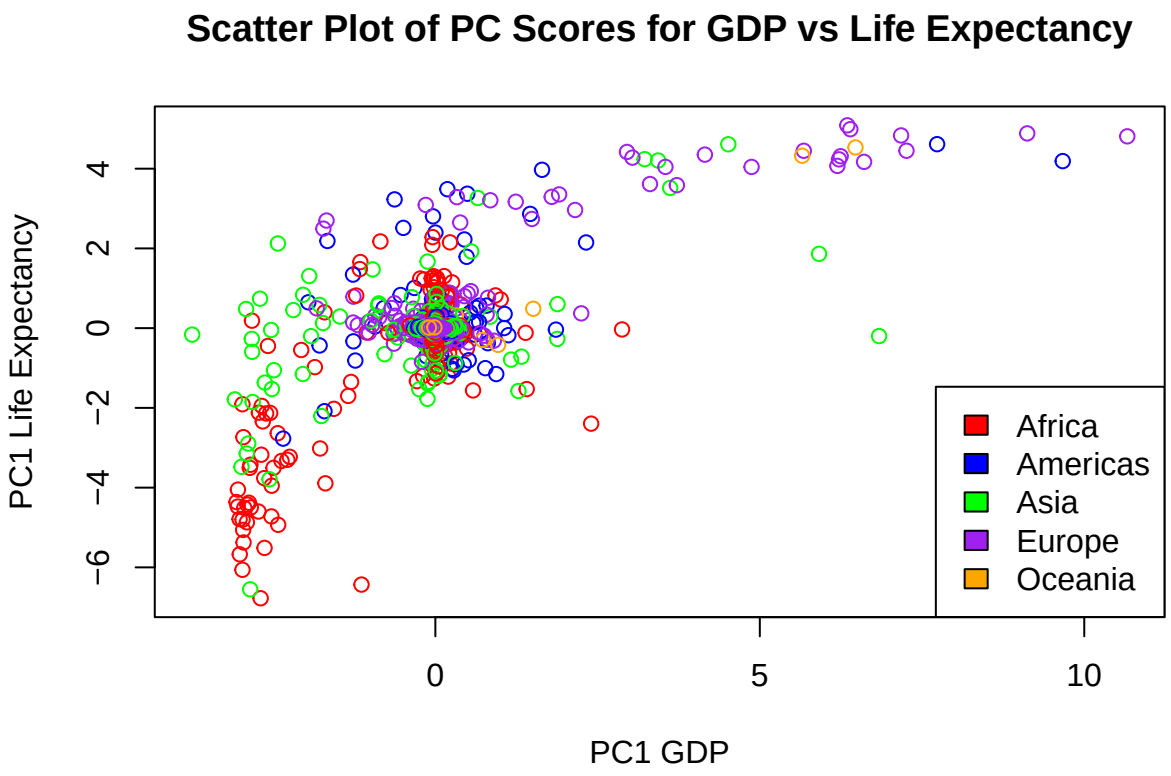
```
## Importance of components:
##                           PC1     PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.328 0.82287 0.3919 0.21989 0.12537 0.10995 0.08805
## Proportion of Variance 0.923 0.05643 0.0128 0.00403 0.00131 0.00101 0.00065
## Cumulative Proportion  0.923 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
```

Principal Component Regression (PCR) and scree plot for life expectancy data reveals that the first and second principal components PC1 and PC2 respectively accounts for approximately 98% of the total variance.

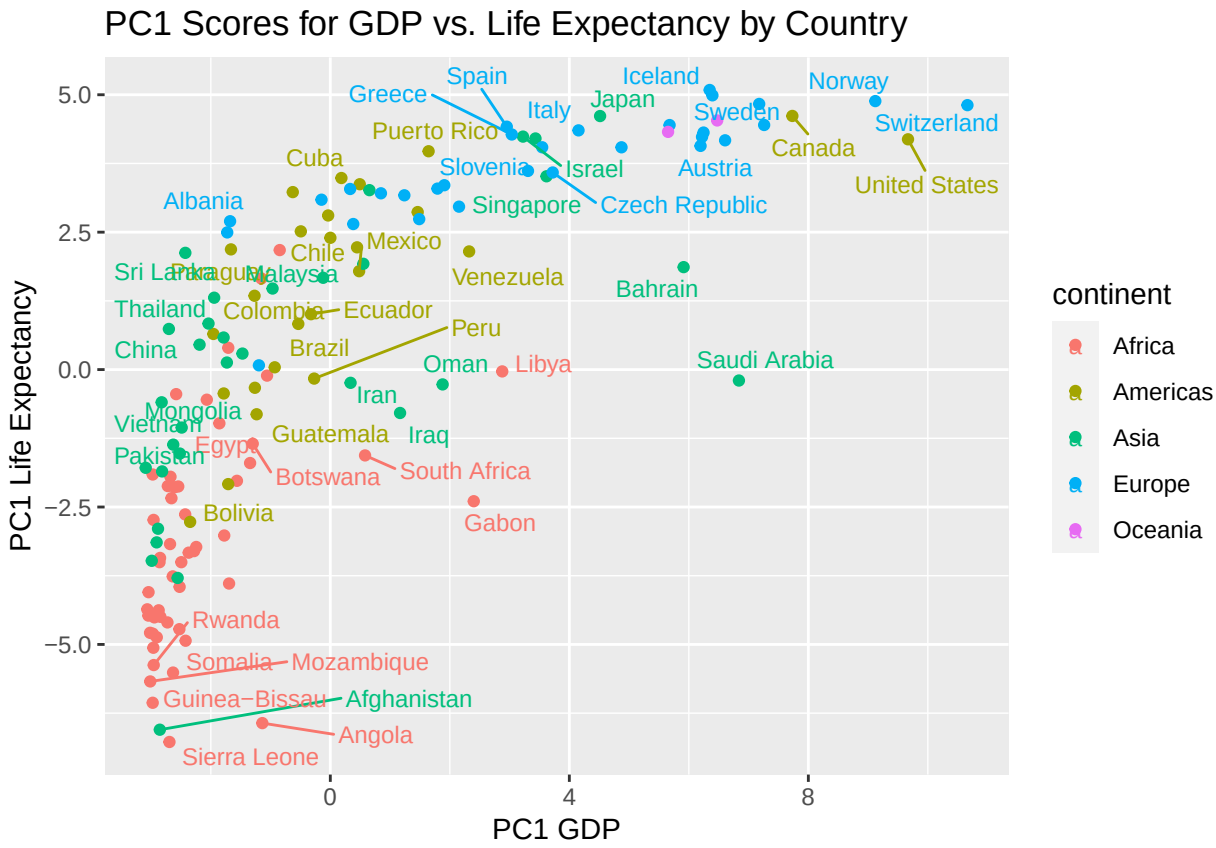
```
fviz_eig(pca_lifeExp, addlabels = TRUE, ylim = c(0, 100))
```




```
# Scatter plot of PC scores for GDP and life expectancy
continent_colors <- c("Africa" = "red", "Americas" = "blue",
                      "Asia" = "green", "Europe" = "purple",
                      "Oceania" = "orange")

plot(pca_gdp$x, pca_lifeExp$x,
     col = continent_colors[UN$continent],
     xlab = "PC1 GDP", ylab = "PC1 Life Expectancy",
     main = "Scatter Plot of PC Scores for GDP vs Life Expectancy")
legend("bottomright", legend = unique(UN$continent),
       fill = unique(continent_colors))
```




```
library(ggrepel)
# Scatter plot of PC1 scores for GDP and life expectancy

pc_scores <- data.frame(country = UN$country, PC1_gdp = pca_gdp$x[, 1],
                        PC1_lifeExp = pca_lifeExp$x[, 1], continent = UN$continent)

ggplot(pc_scores, aes(x = PC1_gdp, y = PC1_lifeExp, color = continent,
                      label = country)) +
  geom_point() +
  geom_text_repel(size = 3,max.overlaps = 15)+
  labs(x = "PC1 GDP", y = "PC1 Life Expectancy",
       title = "PC1 Scores for GDP vs. Life Expectancy by Country")
```



Analysing the PC scores for GDP and life expectancy against the first PC score provides informative insights into the data's variability. Upon visualizing, it becomes evident that there is considerable variation among countries, indicating diverse relationships between GDP and life expectancy across continents. The disperse nature of data points signifies substantial variation without discernible trends or groupings, underlining the heterogeneity in these relationships across countries.

Utilizing the first PC helps mitigate this variation, revealing a common pattern that exists across continents. This suggests that there is a more consistent relationship between GDP and life expectancy across countries, as captured by PC1. Additionally, analyzing the first PC enables us to quantify its contribution to the total variance, emphasizing its importance in explaining the overall variability within the dataset.

Canonical Correlation Analysis

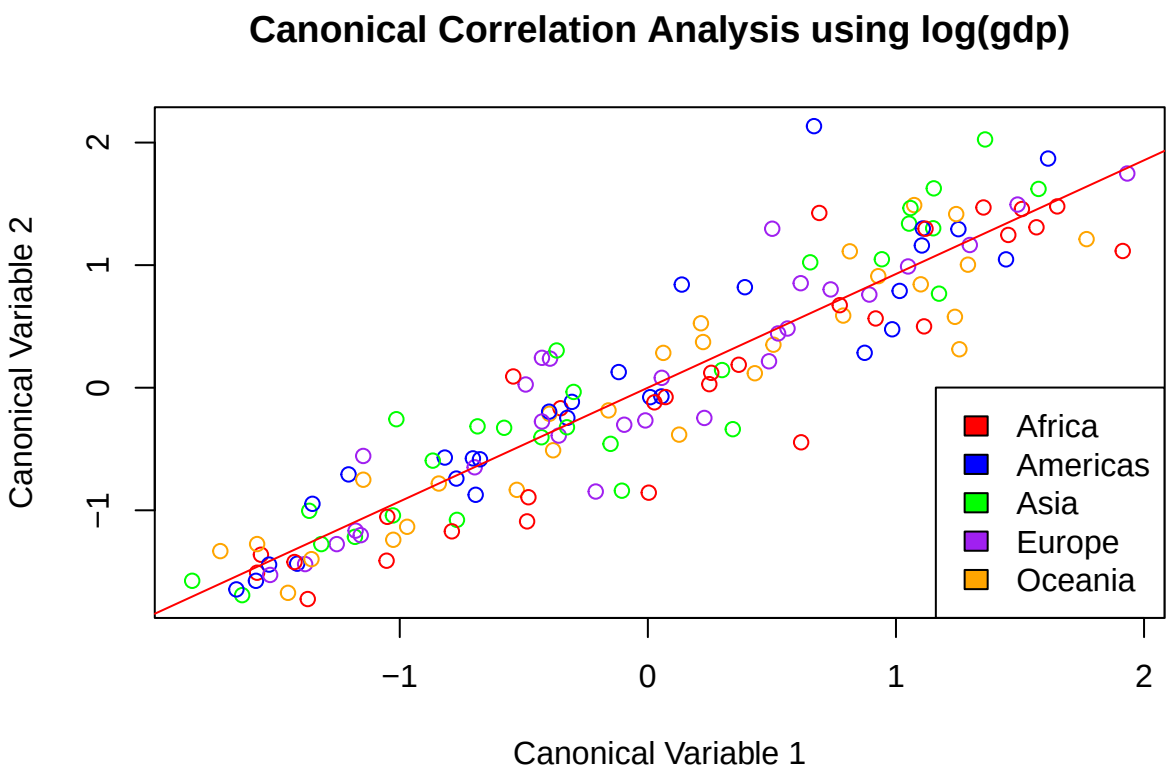
A scatter plot depicting the two canonical correlation variables exhibits a robust positive correlation between the first pair of CC variables when utilising log-transformed GDP and life expectancy, with a correlation value of 0.92. This significant positive correlation is visually apparent in the scatter plot, displaying a clear linear relationship between the two variables.

```
library(CCA)
# CCA using log(GDP) and life expectancy
cca_results_log <- cc(log(gdp), lifeExp)
cat("cor:",cca_results_log$cor[[1]])
```

```
## cor: 0.9277816
```

```
library(CCA)
# Scatter plot of first pair of CC variables
cc_log_var1 <- cca_results_log$scores$xscores[, 1]
cc_log_var2 <- cca_results_log$scores$yscores[, 1]
```

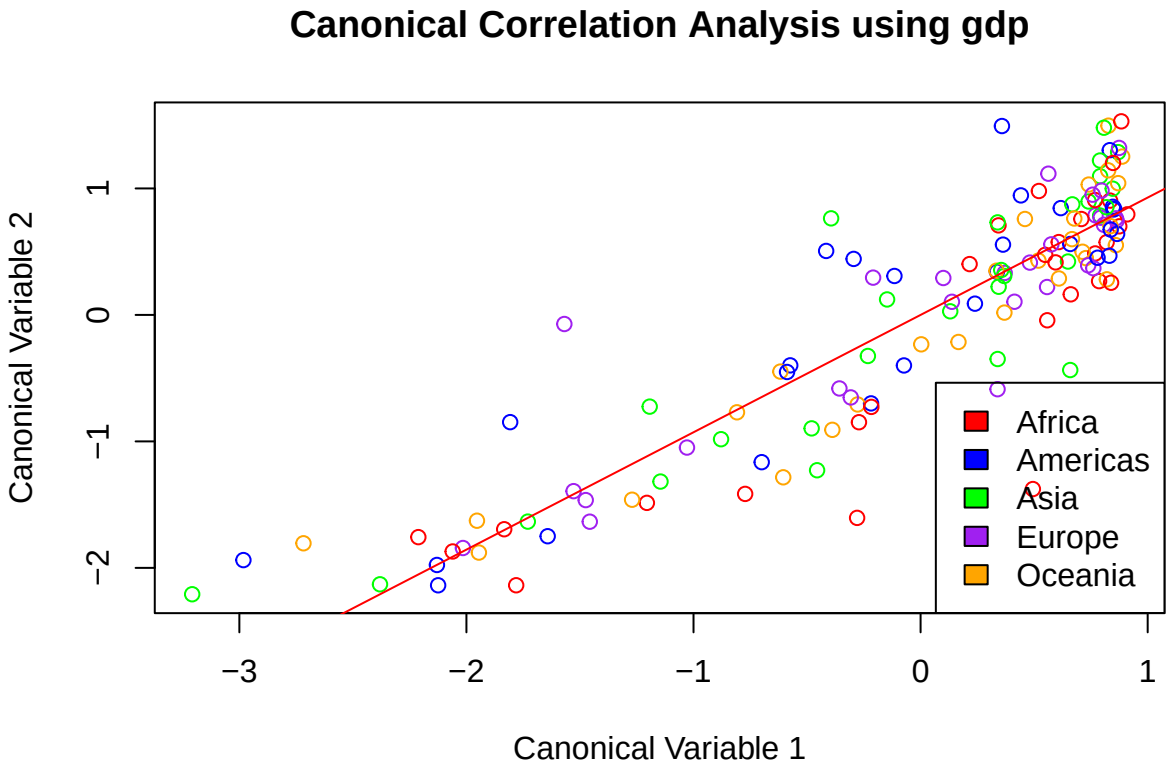
```r
plot(cc_log_var1, cc_log_var2,
     xlab = "Canonical Variable 1",
     ylab = "Canonical Variable 2",
     main = "Canonical Correlation Analysis using log(gdp)",
     col = continent_colors)
abline(lm(cc_log_var2 ~ cc_log_var1), col = "red")
legend("bottomright", legend = unique(UN$continent), fill = unique(continent_colors))
```




```r
cca_results <- cc(gdp, lifeExp)
cat("cor:",cca_results$cor[[1]])
```

```
## cor: 0.8911086
```

```r
cc_var1 <- cca_results$scores$xscores[, 1]
cc_var2 <- cca_results$scores$yscores[, 1]
plot(cc_var1, cc_var2,
     xlab = "Canonical Variable 1",
     ylab = "Canonical Variable 2",
     main = "Canonical Correlation Analysis using gdp",
     col = continent_colors)
abline(lm(cc_log_var2 ~ cc_log_var1), col = "red")
legend("bottomright", legend = unique(UN$continent), fill = unique(continent_colors))
```



On the other hand, when not using the log transformation, the first pair of CC variables displays a strong positive correlation, although slightly lower at 0.89 as compared to the log-transformed pair. This disparity in correlation values is evident in the graphical representation as well.

Multidimensional scaling

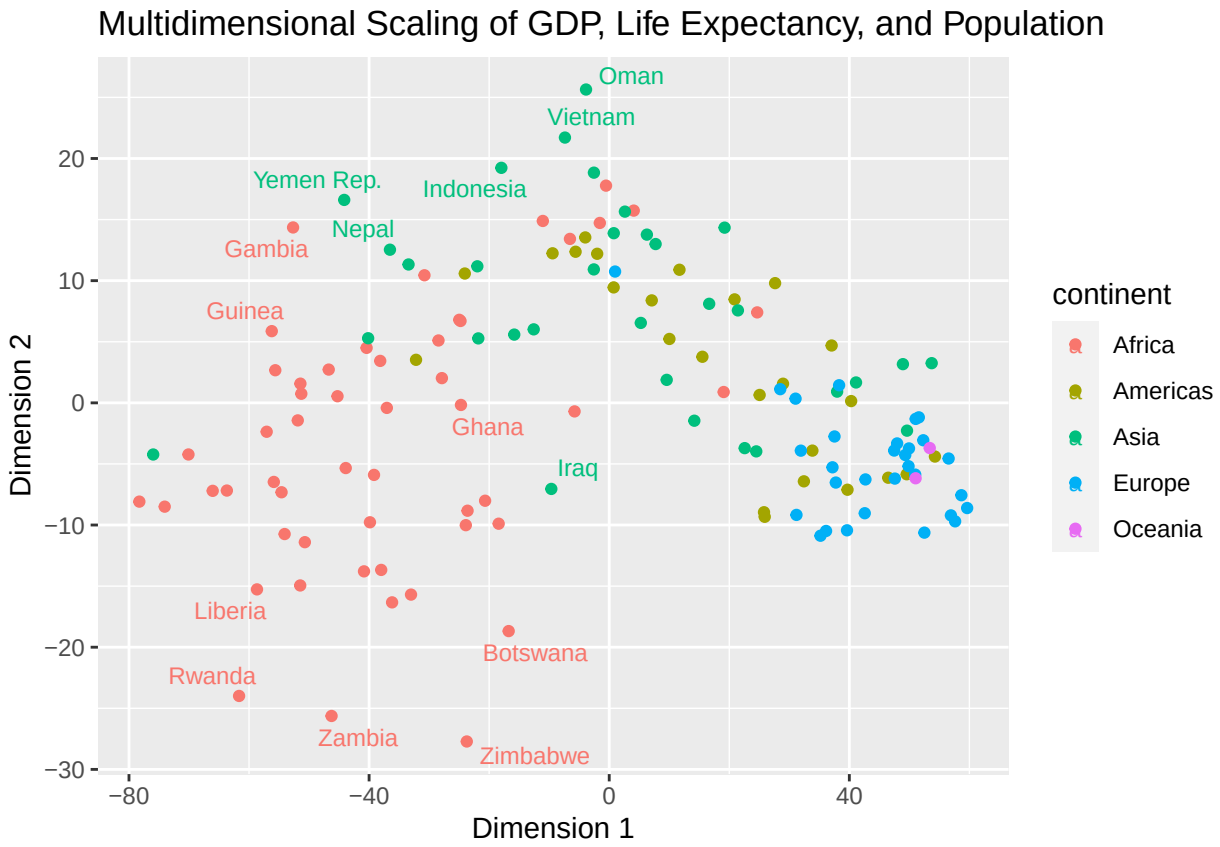
Multidimensional scaling (MDS) technique used to reduce high-dimensional data points to a lower-dimensional space, while maintaining the distances between the points as much as possible. The scatter graph below uses three variables: GDP, life expectancy, and population to form a high-dimensional space. MDS has mapped these variables into two dimensions, which are plotted on the x and y axes.

The placement of the points on the graph signifies the comparative resemblances and dissimilarities among countries with respect to the three variables. Countries that are closer to one another on the plot exhibit greater similarity in their GDP, life expectancy, and population, whereas countries that are farther apart reflect greater disparities.

Countries can be broadly grouped by their continents. For instance, numerous African countries are situated in the lower left quadrant of the plot, whereas numerous European nations are situated in the upper right quadrant. This indicates that there are some recurring patterns with respect to GDP, life expectancy, and population across continents. It is evident that countries in the upper right quadrant of the plot generally possess higher GDPs and life expectancies than those in the lower left quadrant. This suggests a positive correlation between GDP and life expectancy which matches the case with the exploratory data analysis.

```
UN.transformed <- cbind(log(gdp), lifeExp, log(popn))
mds_result <- cmdscale(dist(UN.transformed))
mds_data <- data.frame(x = mds_result[, 1], y = mds_result[, 2], continent = UN$continent)

ggplot(mds_data, aes(x = x, y = y, color = continent)) +
  geom_point() +
  labs(x = "Dimension 1", y = "Dimension 2",
       title = "Multidimensional Scaling of GDP, Life Expectancy, and Population")+
```

```
geom_text_repel(aes(label = UN$country), size = 3, max.overlaps = 5)
```



Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) model applied to the UN dataset achieved an accuracy of 85% in classifying countries into their respective continents based on socio-economic factors such as life expectancy, GDP per capita, and population. This accuracy indicates that the model correctly identified the continent for 85% of the countries in the test set, reflecting the effectiveness of LDA in capturing regional socio-economic patterns. The high accuracy suggests that countries within the same continent share similar socio-economic characteristics, reinforcing the importance of these factors in distinguishing between different regions. Additionally, the LDA model's ability to reduce dimensionality highlights the key socio-economic variables that contribute to these regional distinctions. The accuracy of the LDA model not only serves as a measure of its performance but also provides a benchmark for evaluating other statistical techniques used in the analysis.

```
library(caret)
flda_model <- lda(UN$continent ~ ., data = UN)
set.seed(123)  # for reproducibility
train_index <- createDataPartition(UN$continent, p = 0.7, list = FALSE)
train_data <- UN[train_index, ]
test_data <- UN[-train_index, ]
train_labels <- UN$continent[train_index]
test_labels <- UN$continent[-train_index]
predictions <- predict(flda_model, newdata = test_data)
accuracy <- sum(predictions$class == test_labels) / length(test_labels)
cat("Accuracy of the FLDA model on the test set:", accuracy*100,"%", "\n")
```

```
## Accuracy of the FLDA model on the test set: 85 %
```

Clustering

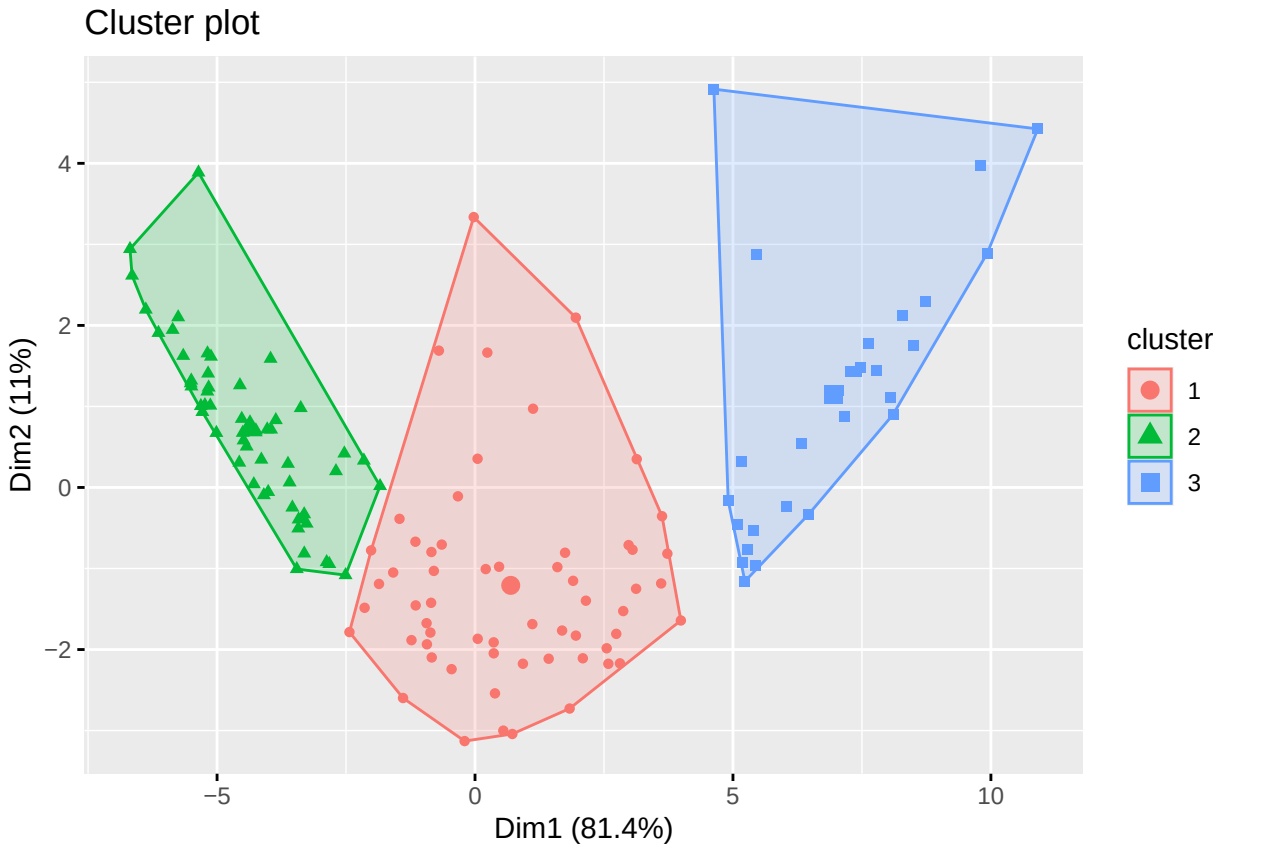
For finding the optimum number of cluster I applied K-means clustering, and Hierarchical clustering.

The k-means plot illustrates three distinct groupings, each distinguished by a unique color and form:

- Cluster 1 (Blue): This cluster signifies countries with a low GDP and low life expectancy. It comprises developing or low-income nations with lower life expectancy and GDP.
- Cluster 2 (Red): This cluster denotes countries with a medium GDP and medium life expectancy. It represents nations with moderate values for both GDP and life expectancy.
- Cluster 3 (Green): This cluster signifies countries with a high GDP and high life expectancy. It comprises developed or high-income nations with higher life expectancy and GDP.

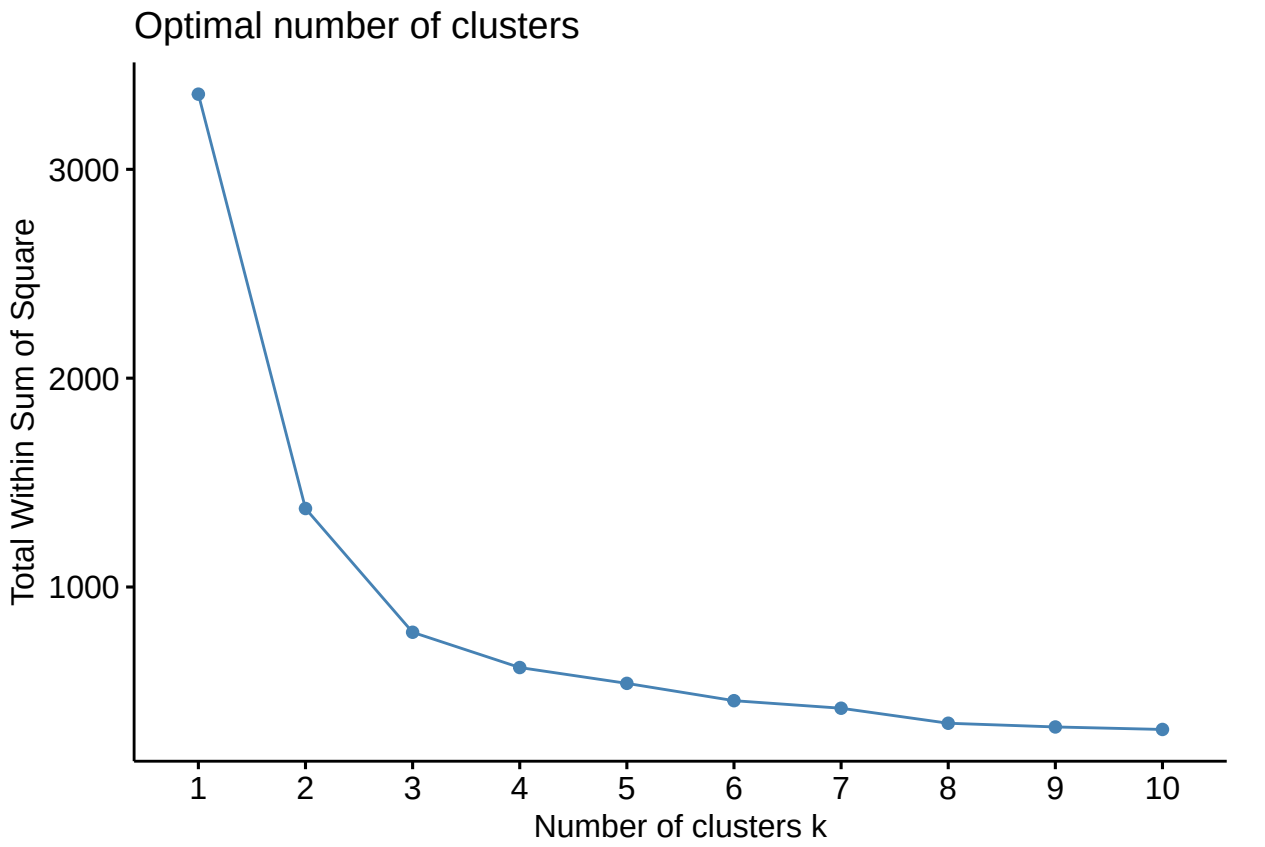
```
UN.scaled <- UN[, 1:26]
UN.scaled[, 3:26] <- scale(UN[, 3:26])
set.seed(123)
kmeans_result <- kmeans(UN.scaled[, 3:26], centers = 3,nstart=25)
```

```
fviz_cluster(kmeans_result, data = UN.scaled[, 3:26],
             geom = "point")
```



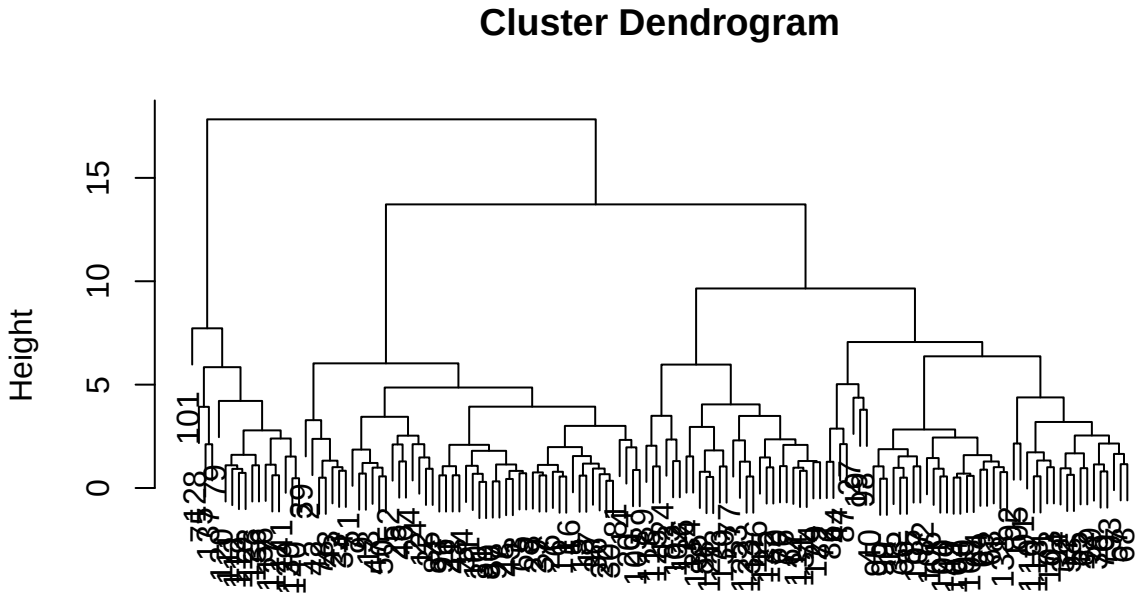
The K-means elbow method reveals a decrease in the sum of square errors as the number of clusters increases, and this trend continues until reaching three clusters. At this point, the errors become less pronounced, indicating that three clusters are a suitable choice. The three clusters can be utilized to categorize countries based on their GDP and life expectancy: first cluster cab comprises countries with a low GDP and life expectancy, second cluster includes countries with a moderate level of GDP and life expectancy and third cluster characterized by countries with a high GDP and life expectancy.

```
fviz_nbclust(UN.scaled[, 3:26], kmeans, method = "wss")
```

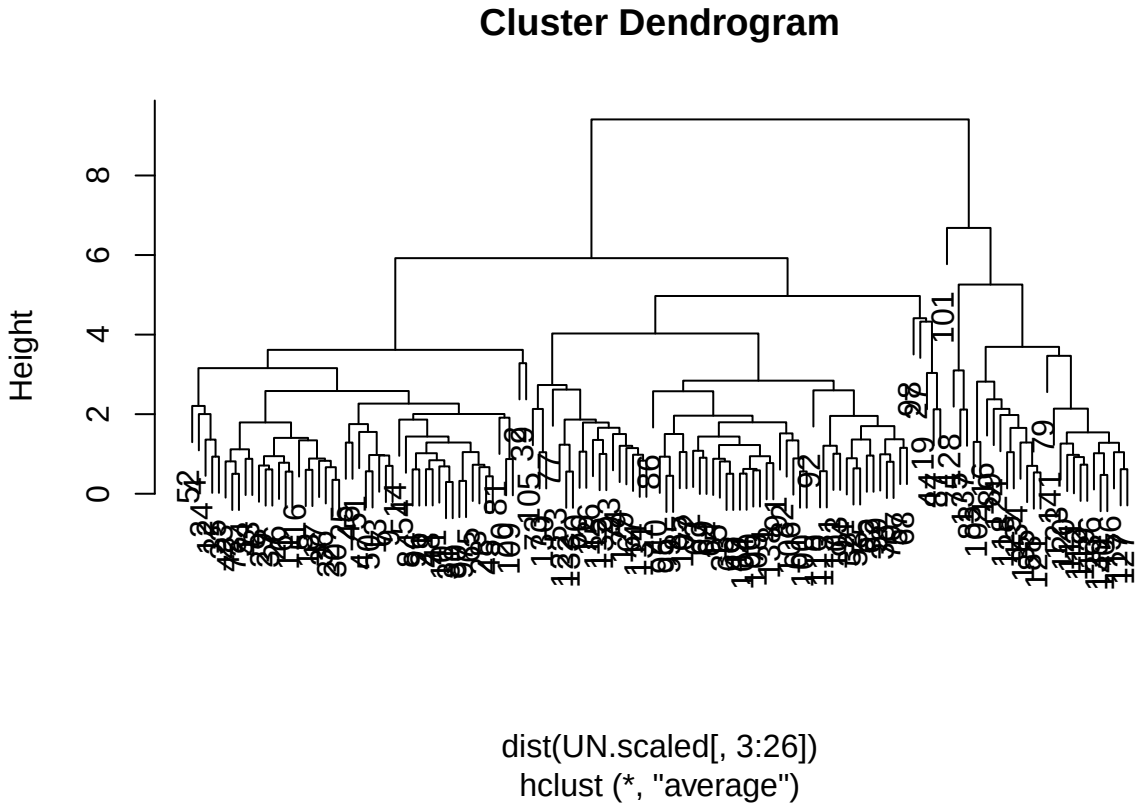


Furthermore, I utilized Hierarchical clustering, identifying three clusters with complete linkage and four clusters with average linkage.

```
hierarchical_result_com <- hclust(dist(UN.scaled[, 3:26]),method ="complete")
plot(hierarchical_result_com)
```




```
hierarchical_result_avg <- hclust(dist(UN.scaled[, 3:26]),method ="average")
plot(hierarchical_result_avg)
```



Linear regression

To predict life expectancy for the year 2007 using GDP for each country, I implemented three regression approaches: ordinary least squares (OLS), principal component regression (PCR), and ridge regression. OLS is applicable when a linear relation exists between predictor variables and the target variable, as indicated by the graph, which presents a more linear correlation between log GDP and life expectancy compared to raw GDP values.

```
#Preparing the data
# Convert data matrices to data frames
gdp_df <- as.data.frame(gdp)
lifeExp_df <- as.data.frame(lifeExp)

# Select GDP up to 2002 for predictors
gdp_predictors <- gdp_df[,-which(colnames(gdp_df) == "2007")]

#GDP in log
gdp_predictors_log <- gdp_predictors
log_data <- sapply(gdp_predictors, function(x) if(is.numeric(x)) log(x) else x)
gdp_predictors_log[] <- log_data

#life expectancy in 2007
lifeExp_response <- lifeExp_df[, which(colnames(lifeExp_df) == "2007")]

#Function to calculate RMSE
calculate_rmse <- function(actual, predicted) {
  sqrt(mean((actual - predicted)^2))
}
```

```
#Ordinary least square
lm_model <- lm(lifeExp_response ~ ., data = gdp_predictors)
summary(lm_model)
```

```
##
## Call:
## lm(formula = lifeExp_response ~ ., data = gdp_predictors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.416  -5.874   1.338   6.581  15.656
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.7991375  1.1059349  53.167   <2e-16 ***
## `1952`      -0.0015701  0.0024510  -0.641   0.5229
## `1957`       0.0051958  0.0032019   1.623   0.1071
## `1962`      -0.0046630  0.0026034  -1.791   0.0756 .
## `1967`      -0.0002082  0.0016311  -0.128   0.8986
## `1972`       0.0016026  0.0015693   1.021   0.3091
## `1977`      -0.0012547  0.0010032  -1.251   0.2133
## `1982`       0.0015816  0.0012202   1.296   0.1972
## `1987`       0.0003344  0.0008251   0.405   0.6860
## `1992`      -0.0024264  0.0011219  -2.163   0.0324 *
## `1997`       0.0026538  0.0013142   2.019   0.0455 *
## `2002`      -0.0004950  0.0006996  -0.708   0.4805
## ---
```

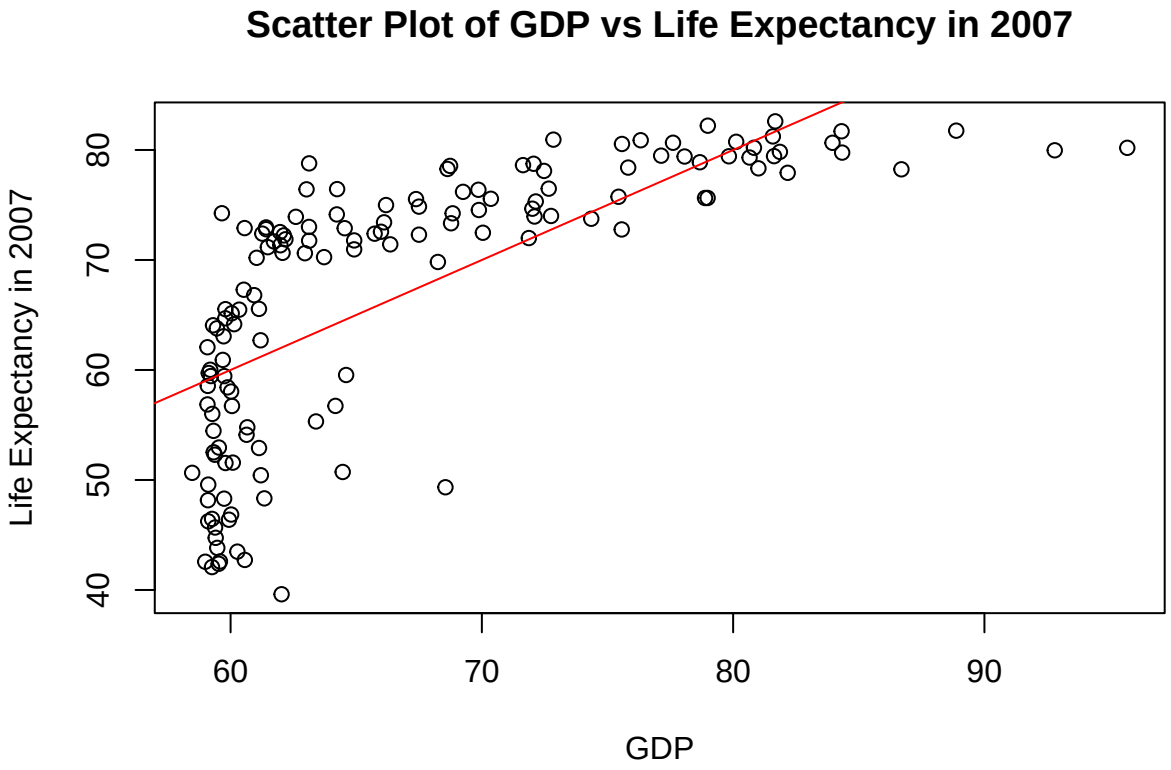
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.887 on 129 degrees of freedom
## Multiple R-squared:  0.5015, Adjusted R-squared:  0.459
## F-statistic:  11.8 on 11 and 129 DF,  p-value: 4.87e-15
```

```
#predictors <- as.matrix(gdp_predictors)
lm_predictions <- predict(lm_model, newdata = gdp_predictors)
rmse_lm <- calculate_rmse(lifeExp_response, lm_predictions)

#OLS with log-transformed GDP values as predictors
lm_model_log <- lm(lifeExp_response ~ ., data = gdp_predictors_log)
summary(lm_model_log)
```

```
##
## Call:
## lm(formula = lifeExp_response ~ ., data = gdp_predictors_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.4219  -2.1505   0.6691   3.8925  14.0263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.027      4.682   1.074   0.2849
## `1952`        -5.650      6.444  -0.877   0.3822
## `1957`        14.202      9.560   1.486   0.1398
## `1962`        -6.619      9.951  -0.665   0.5071
## `1967`         2.964      7.058   0.420   0.6752
## `1972`        -5.882      6.408  -0.918   0.3604
## `1977`        -2.278      6.047  -0.377   0.7070
## `1982`        -4.071      7.668  -0.531   0.5964
## `1987`        10.428      6.444   1.618   0.1081
## `1992`        -8.733      5.783  -1.510   0.1334
## `1997`        16.306      6.560   2.486   0.0142 *
## `2002`        -3.182      4.289  -0.742   0.4595
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.982 on 129 degrees of freedom
## Multiple R-squared:  0.6924, Adjusted R-squared:  0.6661
## F-statistic: 26.39 on 11 and 129 DF,  p-value: < 2.2e-16
```

```
predictors <- as.matrix(gdp_predictors_log)
lm_predictions_log <- predict(lm_model_log, newdata = gdp_predictors_log)
rmse_lm_log <- calculate_rmse(lifeExp_response, lm_predictions_log)
```

```
# Plotting the data and the regression line
plot(predict(lm_model),lifeExp_response, xlab = "GDP",
     ylab = "Life Expectancy in 2007",
     main = "Scatter Plot of GDP vs Life Expectancy in 2007")
abline(0,1, col = "red")
```




```
plot(predict(lm_model_log), lifeExp_response,
     xlab = "Log GDP", ylab = "Life Expectancy in 2007",
     main = "Scatter Plot of log GDP vs Life Expectancy in 2007")
abline(0,1, col = "red")
```

```r
#Principal Component Regression (PCR)
library(pls)

pcr_raw <- pcr(lifeExp_response ~ ., 3, scale = TRUE, data = gdp_predictors, validation = 'CV')
summary(pcr_raw)
```

```
## Data:    X dimension: 141 11
##  Y dimension: 141 1
## Fit method: svdpc
## Number of components considered: 3
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps
## CV           12.13    9.318    9.087    9.116
## adjCV        12.13    9.306    9.074    9.100
##
## TRAINING: % variance explained
##                   1 comps  2 comps  3 comps
## X                   92.40    97.13    99.20
## lifeExp_response    43.07    46.13    46.28
```

```r
pcr_predictions <- predict(pcr_raw, newdata = gdp_predictors)
rmse_pcr <- calculate_rmse(lifeExp_response, pcr_predictions)

#PCR with log-transformed GDP values as predictors
pcr_log <- pcr(lifeExp_response ~ ., 3, scale = TRUE,
               data = gdp_predictors_log, validation = 'CV')
summary(pcr_log)
```

```
## Data:    X dimension: 141 11
##  Y dimension: 141 1
## Fit method: svdpc
## Number of components considered: 3
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps
## CV           12.13    7.895     7.35    7.096
## adjCV        12.13    7.891     7.34    7.086
##
## TRAINING: % variance explained
##                   1 comps  2 comps  3 comps
## X                   94.88    98.63    99.41
## lifeExp_response    57.97    64.58    66.91
```

```r
pcr_predictions_log <- predict(pcr_log, newdata = gdp_predictors_log)
rmse_pcr_log <- calculate_rmse(lifeExp_response, pcr_predictions_log)
```

```r
library(glmnet)
#Ridge Regression
lambdas<- seq(0, 100, by = 0.1)
ridge_model_raw <- glmnet(gdp_predictors,lifeExp_response, alpha = 0, lambda = lambdas)

ridge_predictions_raw <- predict(ridge_model_raw, newx = as.matrix(gdp_predictors))
```

```
rmse_ridge <- calculate_rmse(lifeExp_response, ridge_predictions_raw)

#Ridge Regression with log-transformed GDP values as predictors
ridge_model_log <- glmnet(gdp_predictors_log,lifeExp_response, alpha = 0, lambda = lambdas)

ridge_predictions_raw_log <- predict(ridge_model_log, newx = as.matrix(gdp_predictors_log))
rmse_ridge_log <- calculate_rmse(lifeExp_response, ridge_predictions_raw_log)
```

```
#RMSE
cat("RMSE for Ordinary Least Squares (OLS) regression:", rmse_lm, "\n")
```

```
## RMSE for Ordinary Least Squares (OLS) regression: 8.500893
```

```
cat("RMSE for Ordinary Least Squares (OLS) with log GDP:", rmse_lm_log, "\n")
```

```
## RMSE for Ordinary Least Squares (OLS) with log GDP: 6.677916
```

```
cat("RMSE for Principal Component Regression (PCR):", rmse_pcr, "\n")
```

```
## RMSE for Principal Component Regression (PCR): 8.91611
```

```
cat("RMSE for Principal Component Regression (PCR) with log GDP:", rmse_pcr_log, "\n")
```

```
## RMSE for Principal Component Regression (PCR) with log GDP: 7.308454
```

```
cat("RMSE for Ridge Regression:", rmse_ridge, "\n")
```

```
## RMSE for Ridge Regression: 9.308866
```

```
cat("RMSE for Ridge Regression with log GDP:", rmse_ridge_log, "\n")
```

```
## RMSE for Ridge Regression with log GDP: 8.083264
```

In order to evaluate accuracy, I compared OLS with principal component regression, as well as ridge regression. These methods are capable of capturing intricate relationships between variables and can potentially improve predictive performance. After comparing the root mean square error (RMSE) for each model, I discovered that the OLS model employing log GDP achieved the lowest RMSE, amounting to 6.68. This suggests that the OLS model featuring log GDP offers the most precise predictions of life expectancy for the year 2007 when contrasted with the other methods.