Ranjith Ramaswamy

Introduction:

In this reprt, we will examine data on GDP per capita and life expectancy in different countries between 1952 and 2007. We strive to identify insights and trends that can throw light on the socioeconomic development and health outcomes of countries by using a variety of statistical techniques.

Data Description:

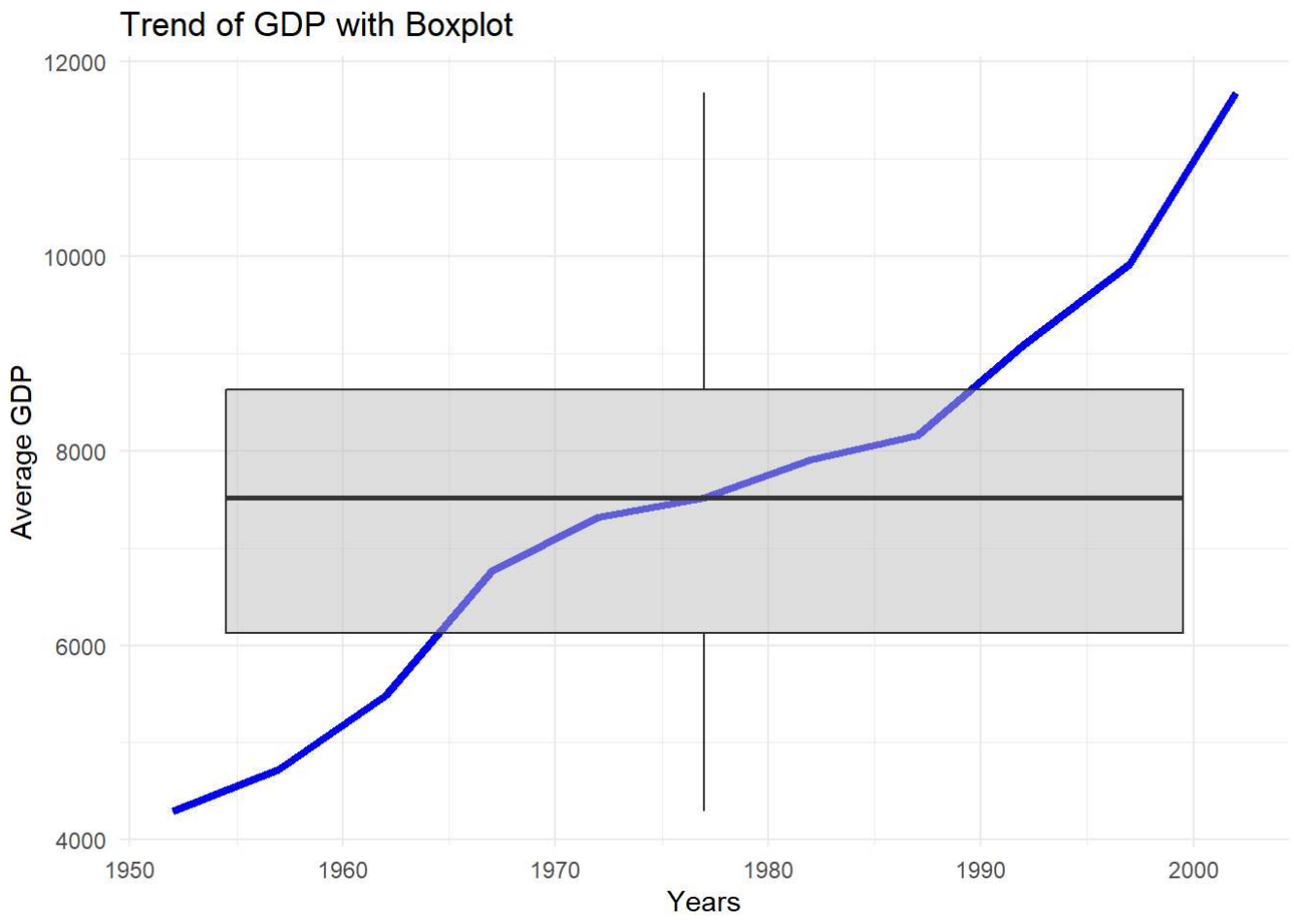
The “gap.csv” file contains information on life expectancy and GDP for 142 nations between 1952 and 2007. We will divide the dataset into two parts:GDP and life expectancy, and take the logarithm of GDP values to simplify the data. This will make it possible for us to compare the life span and economic success of various nations across time.

```
gap.raw <- read.csv('D:/Downloads/gap.csv')
gap <- gap.raw
gap[,3:14]<- log(gap.raw[,3:14])
gdp <- exp(gap[,3:14])
years <- seq(1952, 2007,5)
colnames(gdp) <- years
rownames(gdp) <- gap[,2]
lifeExp <- gap[,15:26]
colnames(lifeExp) <- years
rownames(lifeExp) <- gap[,2]
```

1.Exploratory Data Analysis

let’s explore the changes in GDP and life expectancy over the past years.

```
avg_gdp <- colMeans(gdp[, -1])
avg_lifeExp <- colMeans(lifeExp[, -1])
trend_gdp <- data.frame(years = years[-length(years)], average_gdp = avg_gdp)
trend_lifeExp <- data.frame(years = years[-length(years)], average_lifeExp = avg_lifeExp)
ggplot(data = trend_gdp, aes(x = years, y = average_gdp)) +
  geom_line(color = "blue", size = 1.5) +
  geom_boxplot(aes(y = average_gdp), width = 0.2, fill = "gray", alpha = 0.5, outlier.shape = NA) +
  labs(title = "Trend of GDP with Boxplot", x = "Years", y = "Average GDP") +
  theme_minimal()
```

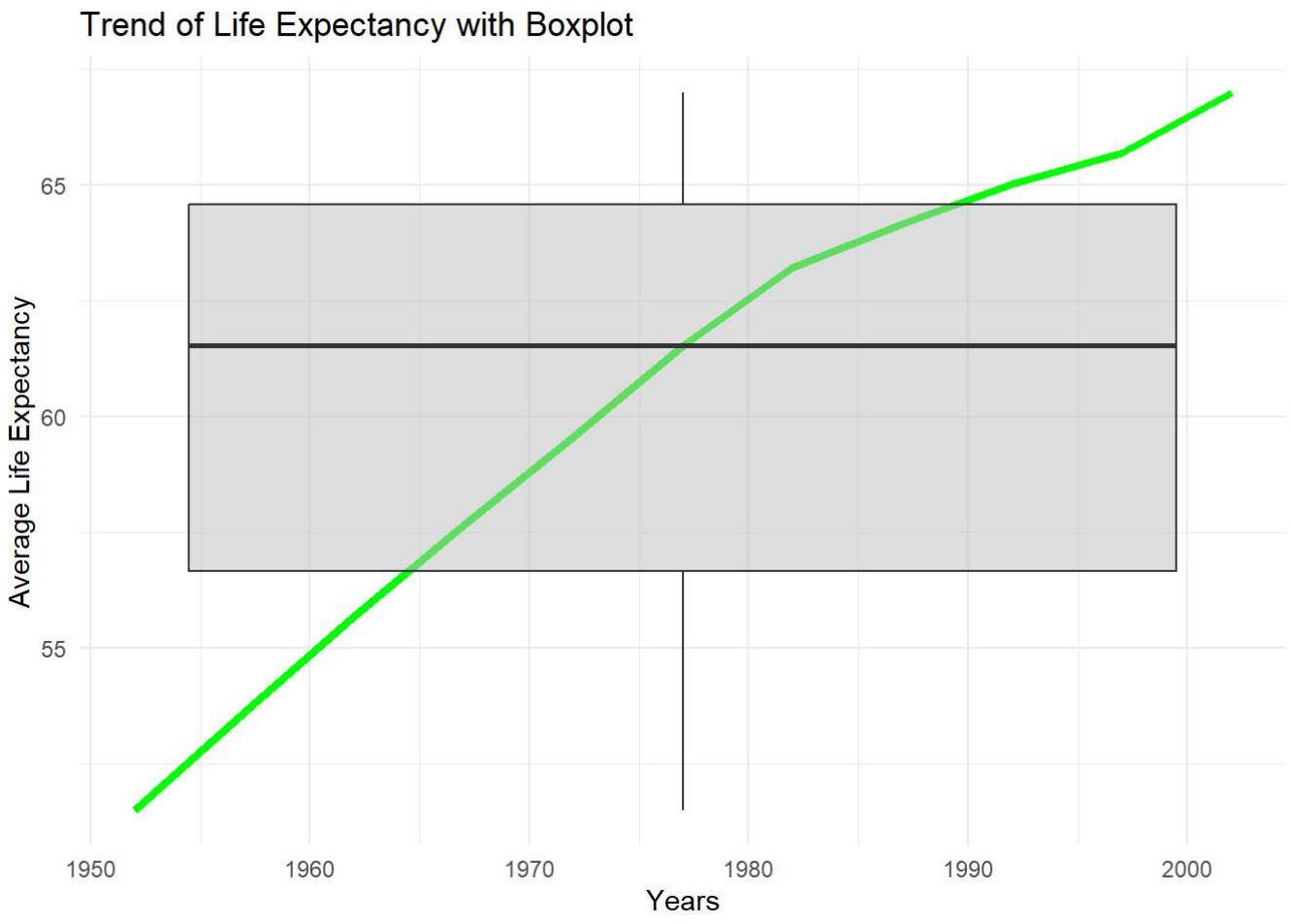


The graph shows a steady increase in the average GDP of 142 countries with minor fluctuations, from 4299 in 1952 to 11680 in 2007.

The highest GDPs were recorded for Kuwait (108382.3) in 1952 and Norway (49357.1) in 2007, while the lowest were for Lesotho (298.84) in 1952 and Congo Dem. Rep. (277) in 2007.

This indicates a generally positive trend in global economic growth, with notable disparities among different nations.

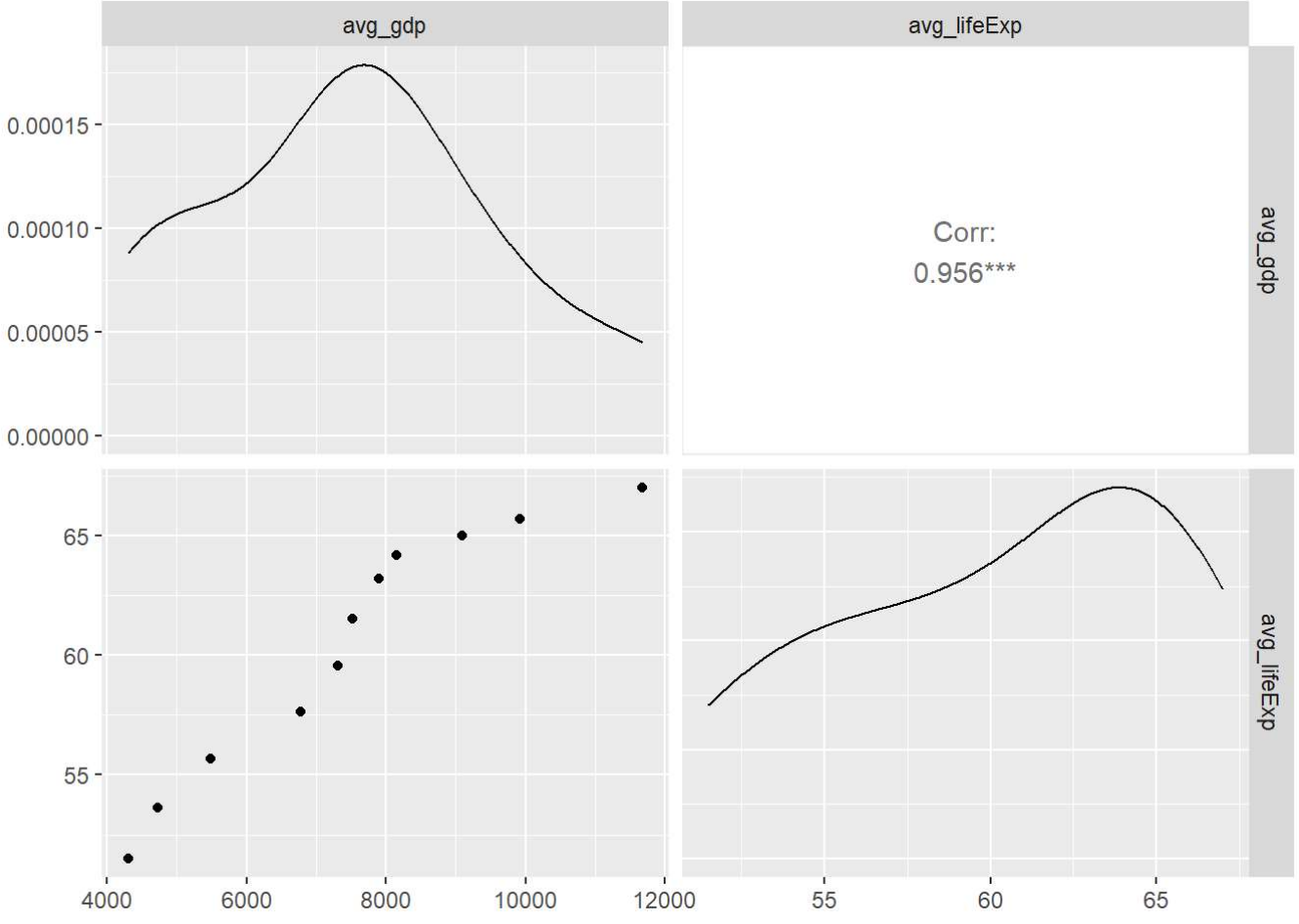
```
ggplot(data = trend_lifeExp, aes(x = years, y = average_lifeExp)) +
  geom_line(color = "green", size = 1.5) +
  geom_boxplot(aes(y = average_lifeExp), width = 0.2, fill = "gray", alpha = 0.5, outlier.shape = NA) +
  labs(title = "Trend of Life Expectancy with Boxplot", x = "Years", y = "Average Life Expectancy") +
  theme_minimal()
```



The life expectancy graph spanning 1952 to 2007 reveals a remarkable rise in the average life expectancy worldwide, from about 48 years in 1952 to around 68 years in 2007. In 1952, Norway had the highest life expectancy of about 73 years, whereas Afghanistan had the lowest of around 29 years. By 2007, Japan reported the highest life expectancy of around 83 years, while Swaziland had the lowest, around 40 years.

The variations in life expectancy among countries highlight substantial differences in their healthcare and social welfare systems. This showcases the need for continued efforts to improve global health and bridge the gaps in life expectancy between nations.

```
data <- data.frame(years = trend_gdp$years, avg_gdp = trend_gdp$average_gdp, avg_lifeExp = trend_lifeExp$average_lifeExp)
ggpairs(data, columns = c("avg_gdp", "avg_lifeExp"))
```



Based on the scatterplot matrix analysis, we can conclude that there is a robust positive correlation(0.956) between average GDP and average life expectancy for the 142 countries analyzed between 1952 and 2007.

The distributions of average GDP and life expectancy are right-skewed and slightly left-skewed, respectively. The scatterplot shows a clear trend of increasing life expectancy with higher GDP, indicating a positive impact of economic development and healthcare improvements on lifespan.

2.Principal component analysis

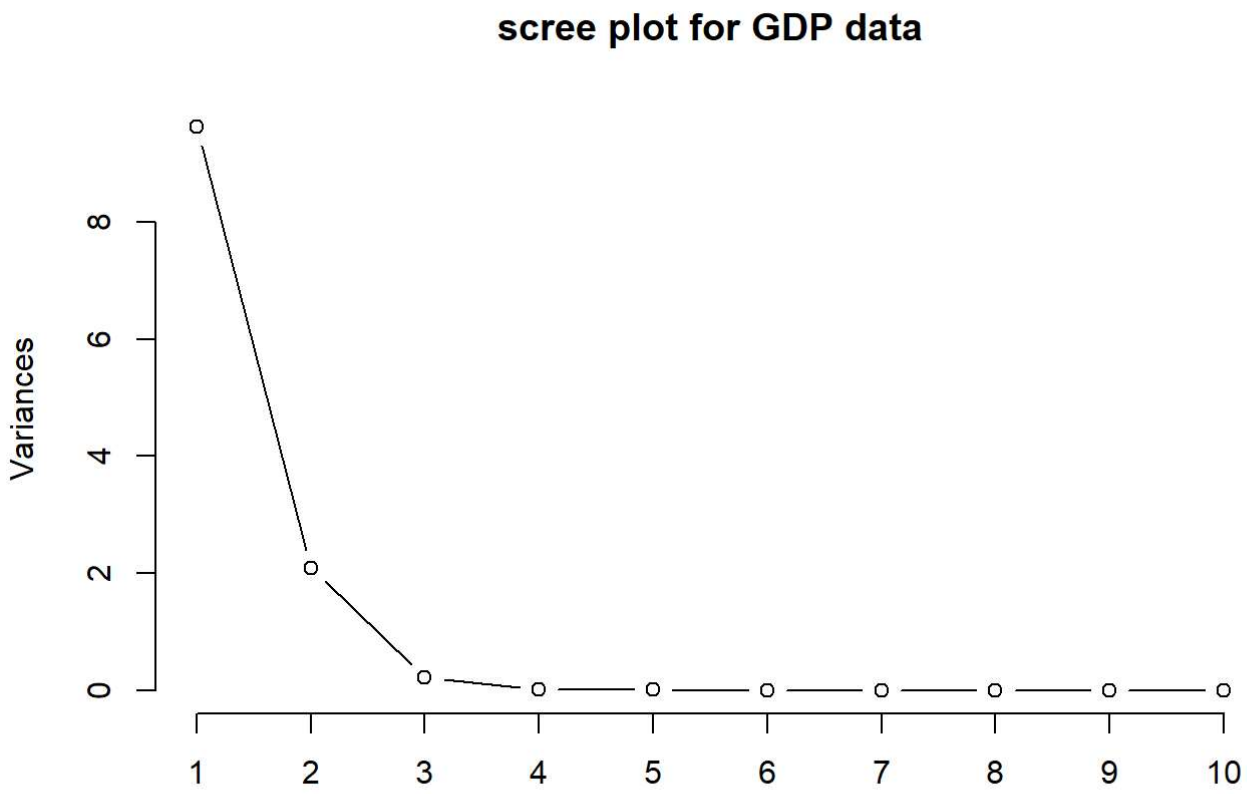
```
gdp.pca <- prcomp(gdp, scale = TRUE)
# Performing PCA on life expectancy data
lifeExp.pca <- prcomp(lifeExp, scale = TRUE)
summary(gdp.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.1002 1.4452 0.4762 0.16486 0.14463 0.10219 0.07150
## Proportion of Variance 0.8009 0.1741 0.0189 0.00226 0.00174 0.00087 0.00043
## Cumulative Proportion  0.8009 0.9750 0.9939 0.99613 0.99788 0.99875 0.99917
##                            PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.06442 0.04977 0.04188 0.03468 0.01842
## Proportion of Variance 0.00035 0.00021 0.00015 0.00010 0.00003
## Cumulative Proportion  0.99952 0.99973 0.99987 0.99997 1.00000
```

```
summary(lifeExp.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5    PC6     PC7
## Standard deviation     3.3284 0.8219 0.39169 0.21950 0.1250 0.1098 0.08778
## Proportion of Variance 0.9232 0.0563 0.01279 0.00402 0.0013 0.0010 0.00064
## Cumulative Proportion  0.9232 0.9795 0.99229 0.99630 0.9976 0.9986 0.99925
##                            PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.06709 0.04413 0.03673 0.02815 0.02024
## Proportion of Variance 0.00038 0.00016 0.00011 0.00007 0.00003
## Cumulative Proportion  0.99963 0.99979 0.99990 0.99997 1.00000
```

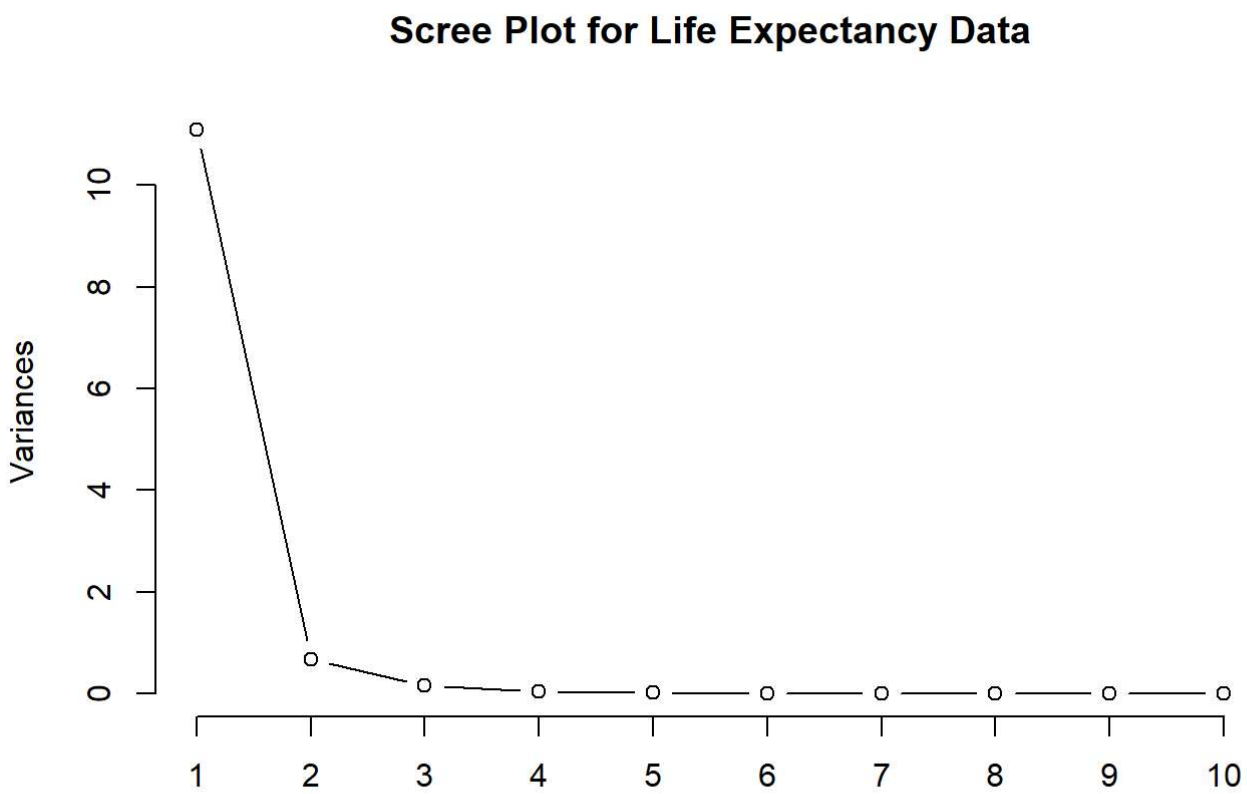
```
plot(gdp.pca, type = "l",main="scree plot for GDP data")
```



Based on the scree plot of GDP data, we will retain the first two principal components:

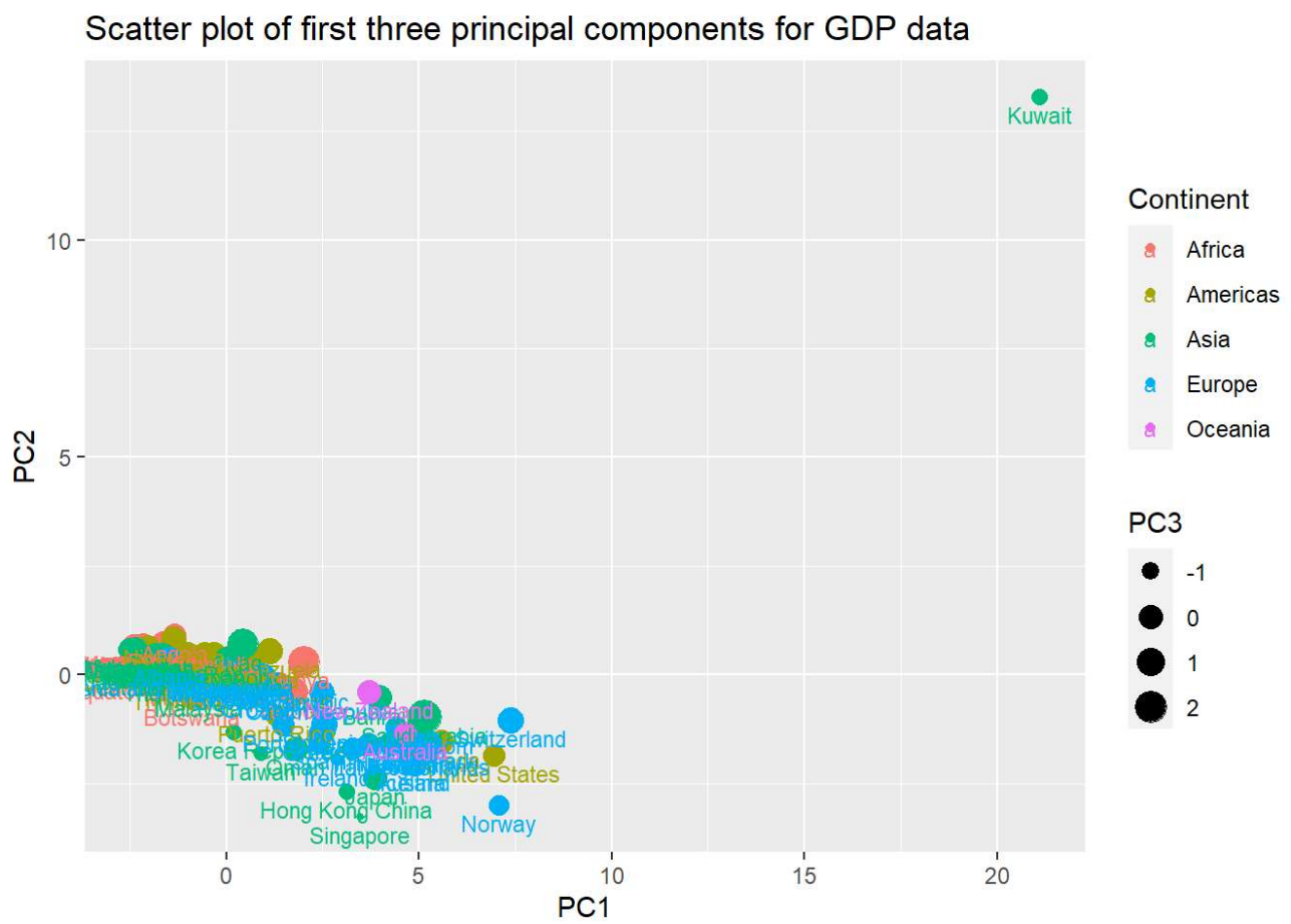
The first component represents a combination of all the variables and captures the general level of economic development, while the second component represents differences in economic growth rates over time.

```
plot(lifeExp.pca, type = "l", main = "Scree Plot for Life Expectancy Data")
```

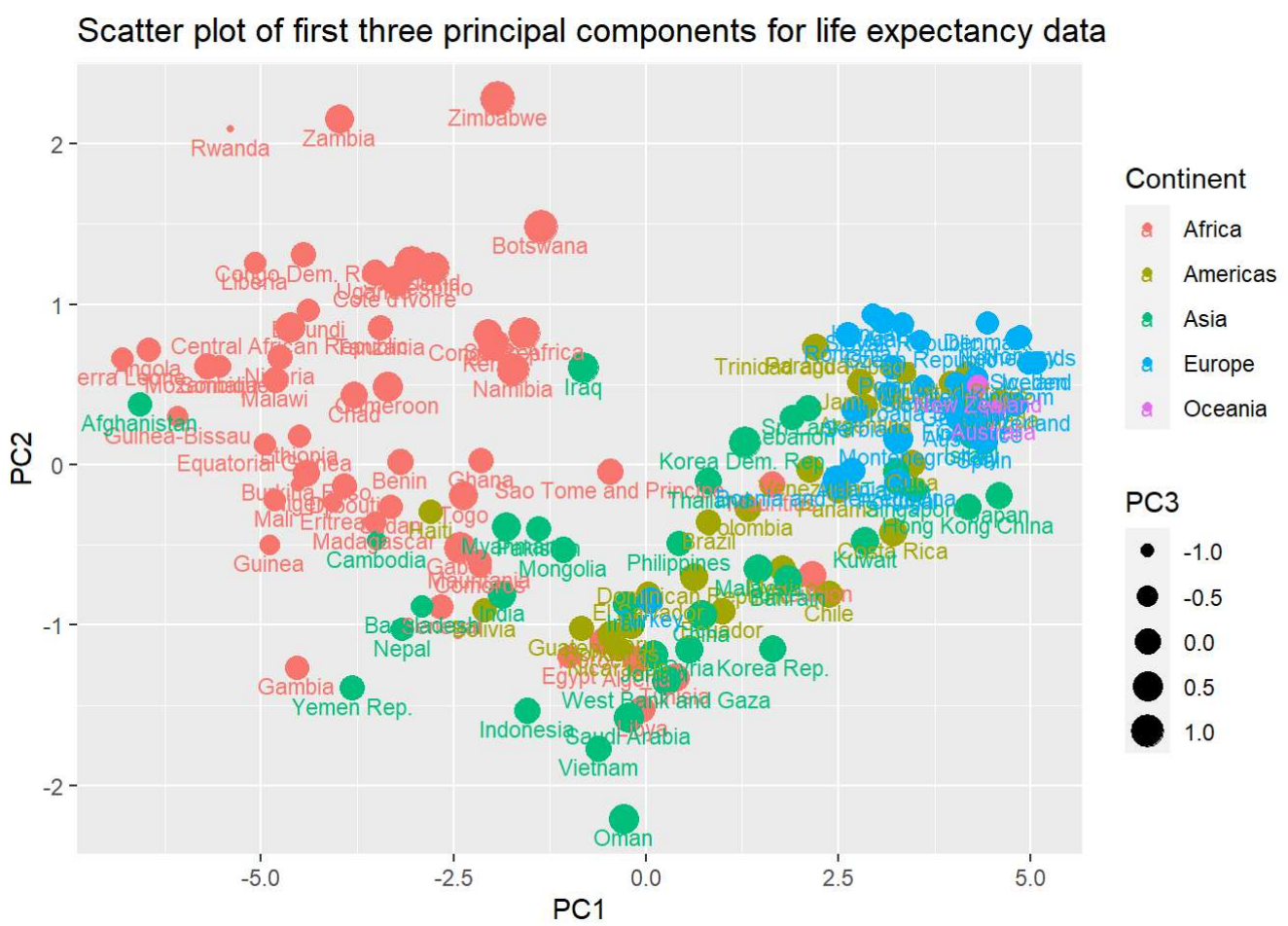


Based on the scree plot of life expectancy data, we will retain the first principal component. This component represents a combination of all the variables and captures the general level of health and healthcare access.

```
gdp.scores <- data.frame(Country = rownames(gdp.pca$x), gdp.pca$x[,1:3])
lifeExp.scores <- data.frame(Country = rownames(lifeExp.pca$x), lifeExp.pca$x[,1:3])

# Adding continent information
gdp.scores$Continent <- gap$continent[match(gdp.scores$Country, gap$country)]
lifeExp.scores$Continent <- gap$continent[match(lifeExp.scores$Country, gap$country)]
ggplot(gdp.scores, aes(x = PC1, y = PC2, color = Continent, size = PC3)) +
  geom_point() +
  geom_text(aes(label = Country), size = 3, vjust = 1.5) +
  ggtitle("Scatter plot of first three principal components for GDP data ")
```




```
ggplot(lifeExp.scores, aes(x = PC1, y = PC2, color = Continent, size = PC3)) +
  geom_point() +
  geom_text(aes(label = Country), size = 3, vjust = 1.5) +
  ggtitle("Scatter plot of first three principal components for life expectancy data")
```



The first two components are plotted on the x and y axis respectively, and the third component is represented by the size of the points.

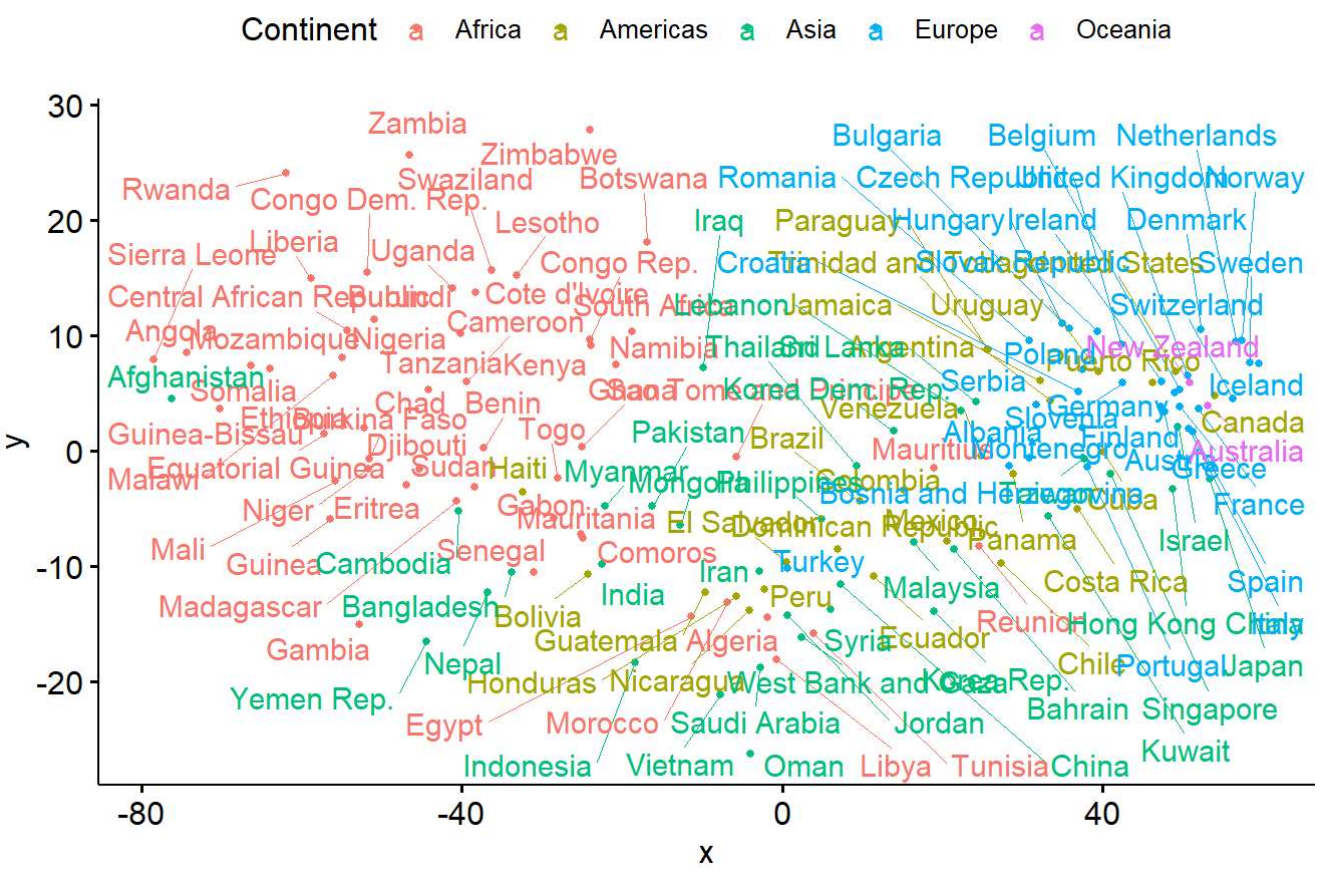
Based on GDP scatterplot, we can say that Kuwait has high values for the first two principal components, indicating that it has a high level of economic development compared to other countries in the dataset.

The negative value for the third principal component suggests that Kuwait's economic development is different from other countries in a way that is not captured by the first two components.

This could be due to factors such as the structure of the economy, the role of the government in the economy, or other unique features of Kuwait's economic system.

3.Multidimensional scaling

```
data <- gap[,3:26]
dist_matrix <- dist(data, method = "euclidean")
# Performing MDS
mds <- cmdscale(dist_matrix, k = 2)
mds_df <- data.frame(
  Country = gap$country,
  x = mds[, 1],
  y = mds[, 2],
  Continent = gap$continent
)
# Plotting MDS
ggscatter(mds_df, x = "x", y = "y",
          label = "Country",
          color = "Continent",
          size = 1,
          repel = TRUE) +
  ggtitle("Multidimensional Scaling of GDP and Life Expectancy Data:")
```



Based on the plots generated above and previously, we can observe some similarities, like: African countries tend to cluster together, as do Asian and European countries, suggesting that there may be underlying patterns in the data related to geography, culture, or other factors that differentiate countries from different regions. This is consistent with what we might expect based on prior knowledge of the world, and it is encouraging that these patterns are visible in both the MDS and PCA plots.

It is worth noting, however, that the MDS plot and PCA plot are not identical and may highlight slightly different aspects of the data.

4.Hypothesis test

A multivariate hypothesis test employing a MANOVA can be used to determine whether there is a significant difference in mean GDP and life expectancy between Asian and European countries.

If the overall MANOVA test produces a p-value smaller than the selected significance level (typically 0.05), we can reject the null hypothesis and conclude that there is a significant difference between Asian and European countries' mean GDP and life expectancy.

This test can reveal important information on future economic and social differences between the two continents.

```
gdp_2007 <- gdp[, "2007"]
lifeExp_2007 <- lifeExp[, "2007"]

data_2007 <- data.frame(gdp_2007, lifeExp_2007)
data_2007$Continent <- gap$continent
data_2007_subset <- subset(data_2007, Continent %in% c("Asia", "Europe"))
manova_results <- manova(cbind(gdp_2007, lifeExp_2007) ~ Continent, data = data_2007_subset)
summary(manova_results)
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## Continent  1 0.25699   10.377      2     60 0.0001348 ***
## Residuals 61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value from the multivariate analysis of variance (MANOVA) for the mean GDP and life expectancy of Asian and European countries in the year 2007 is 0.0001348, which strongly rejects the null hypothesis.

This suggests that there is a significant difference between the average GDP and life expectancy of the two regions. Possible reasons for this disparity could include differences in healthcare systems, economic policies, or cultural factors.

```
gdp_1952 <- gdp[, "1952"]
lifeExp_1952 <- lifeExp[, "1952"]
data_1952 <- data.frame(gdp_1952, lifeExp_1952)
data_1952$Continent <- gap$continent
data_1952_subset <- subset(data_1952, Continent %in% c("Asia", "Europe"))

manova_results_1952 <- manova(cbind(gdp_1952, lifeExp_1952) ~ Continent, data = data_1952_subset)
summary(manova_results_1952)
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## Continent  1 0.58016   41.456      2     60 4.927e-12 ***
## Residuals 61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The small p-value of 4.927e-12 for the year 1952 indicates that we have strong evidence to reject the null hypothesis and conclude that there is a significant difference between the mean GDP and life expectancy of Asian and European countries during that year.

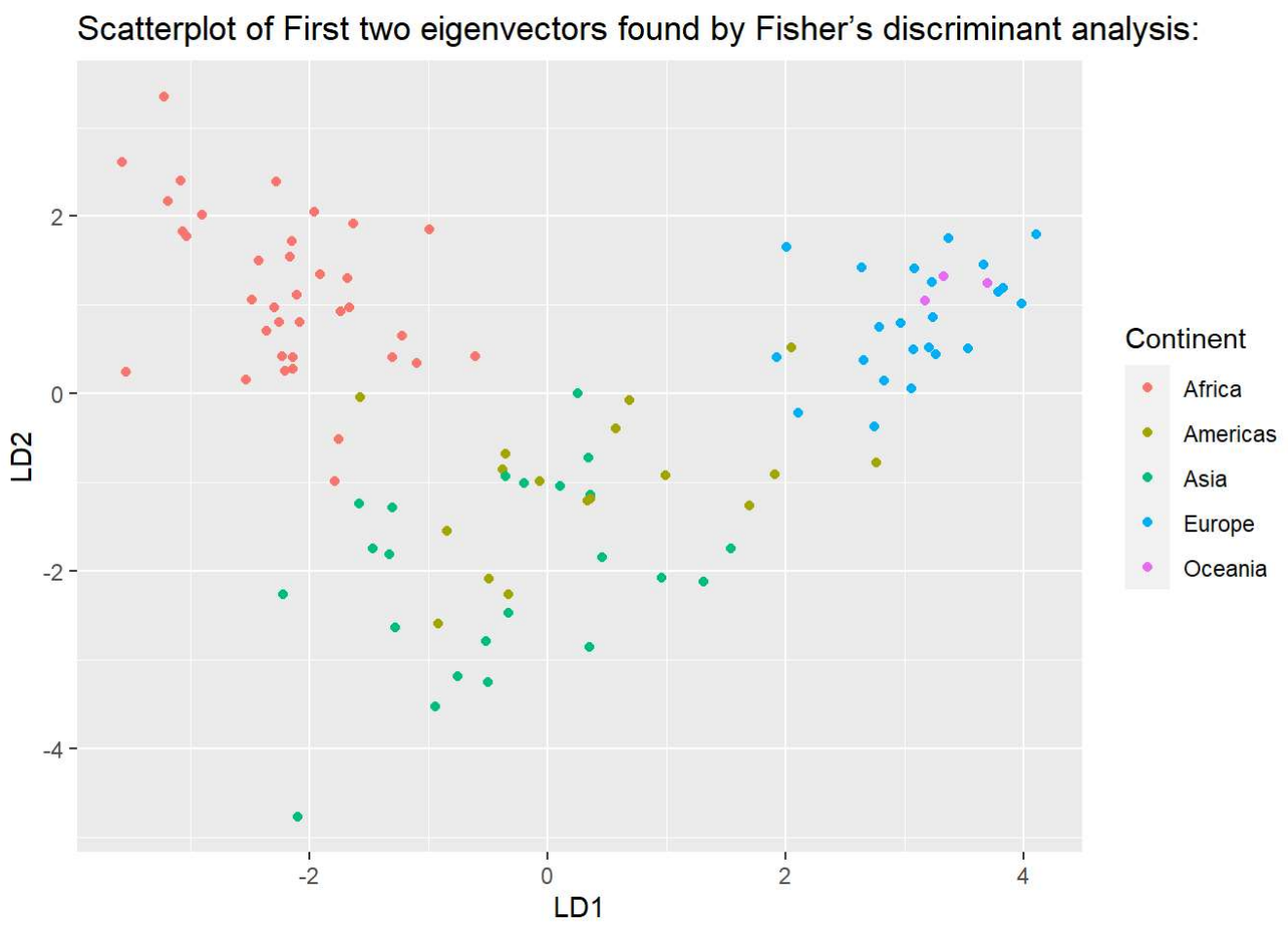
This finding suggests that the two continents were more dissimilar in terms of GDP and life expectancy in 1952 compared to 2007, indicating potential shifts in economic and social development over time.

5.Linear discriminant analysis

```
data <- gap[,3:26]
data <- as.data.frame(data)
data$Continent <- gap$continent

set.seed(123)
inTrain <- createDataPartition(data$Continent, p = 0.7, list = FALSE)
training <- data[inTrain, ]
testing <- data[-inTrain, ]
lda.fit <- lda(Continent ~ ., data = training)
lda.pred <- predict(lda.fit, newdata = data, type = "lda")
lda_accuracy <- mean(lda.pred$class == data$Continent)
print(paste("The predictive accuracy is ", lda_accuracy*100, "%"))
```

```
## [1] "The predictive accuracy is  78.8732394366197 %"
```

```
lda.projection <- predict(lda.fit)$x[, 1:2]
lda.projection_df <- data.frame(lda.projection, Continent = predict(lda.fit)$class)

ggplot(lda.projection_df, aes(x = LD1, y = LD2, color = Continent)) +
  geom_point() +
  ggtitle("Scatterplot of First two eigenvectors found by Fisher's discriminant analysis:")
```



Compared to the PCA plot we produced earlier, the LDA plot displays a clearer separation between the continents than the PCA plot, indicating that the LDA approach may be better suited for predicting the continent of each country using the GDP and life expectancy data.

While the PCA plot revealed some clustering by continent, it was not as apparent as the LDA plot's clustering.

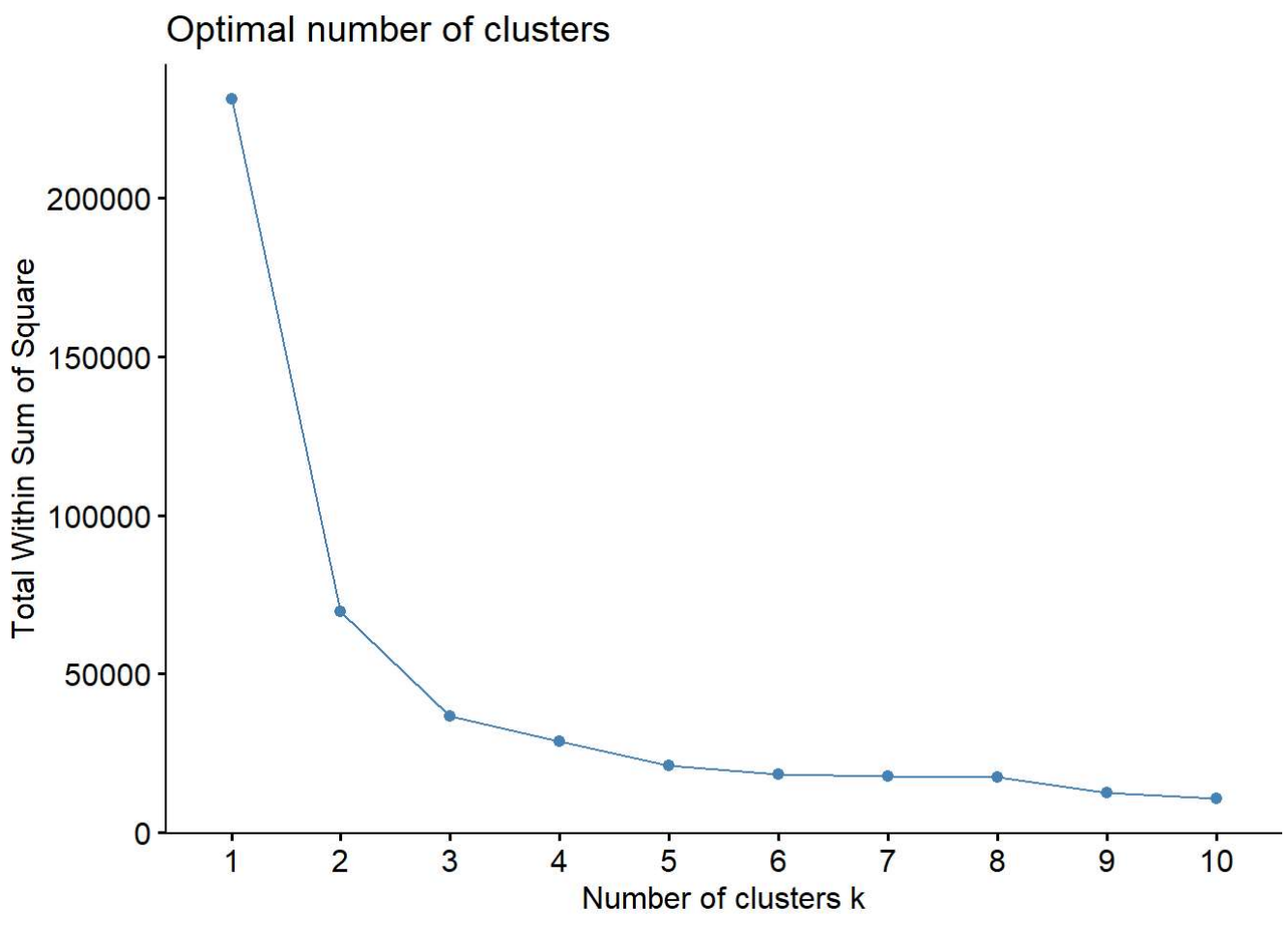
This is because LDA is explicitly meant to find linear combinations of variables that maximise class separation, whereas PCA just aims to maximize data variance.

6.Clustering

K-means clustering

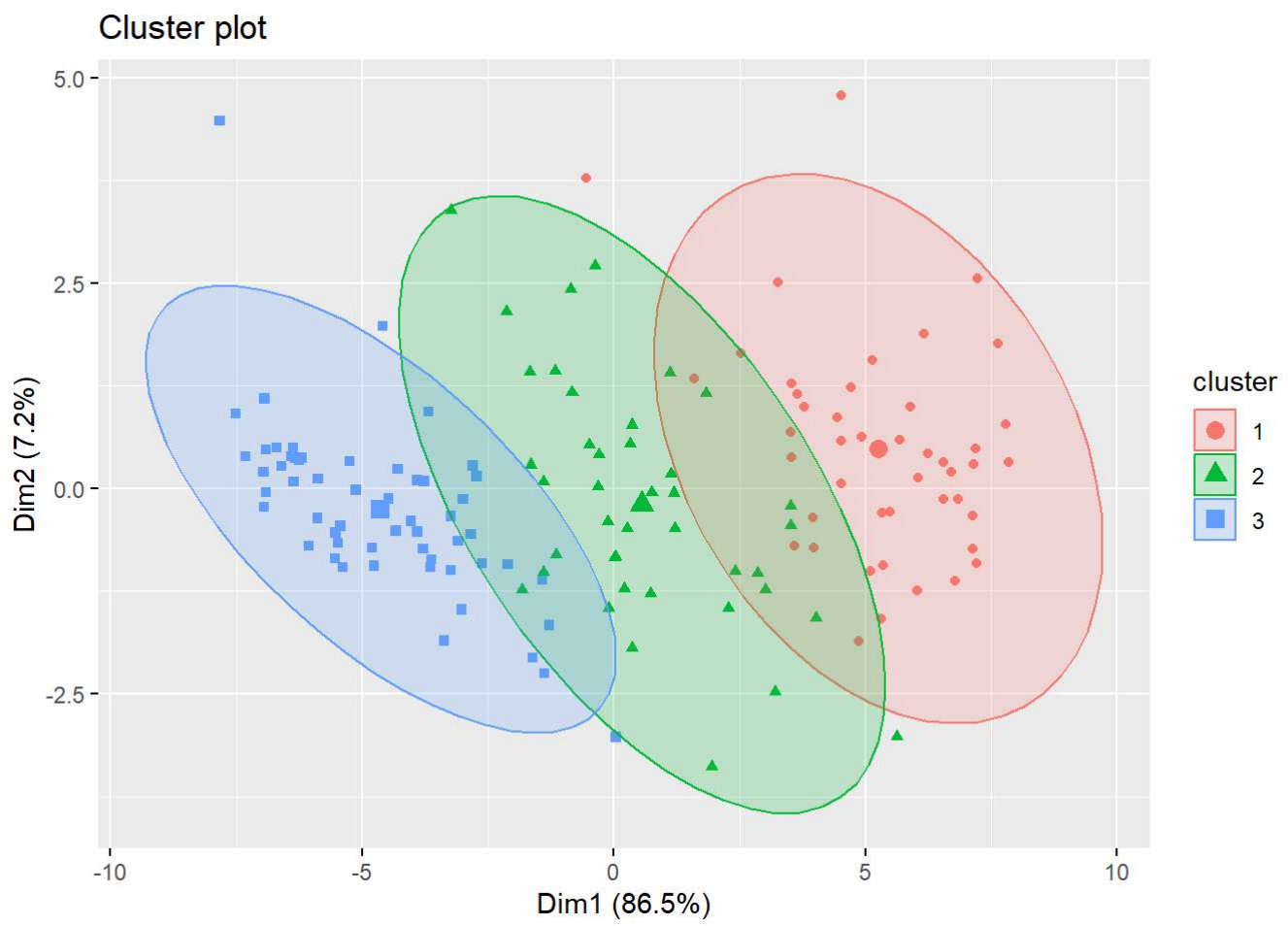
Choosing the number of clusters using elbow method:

```
fviz_nbclust(gap[3:26], kmeans, method = "wss")
```



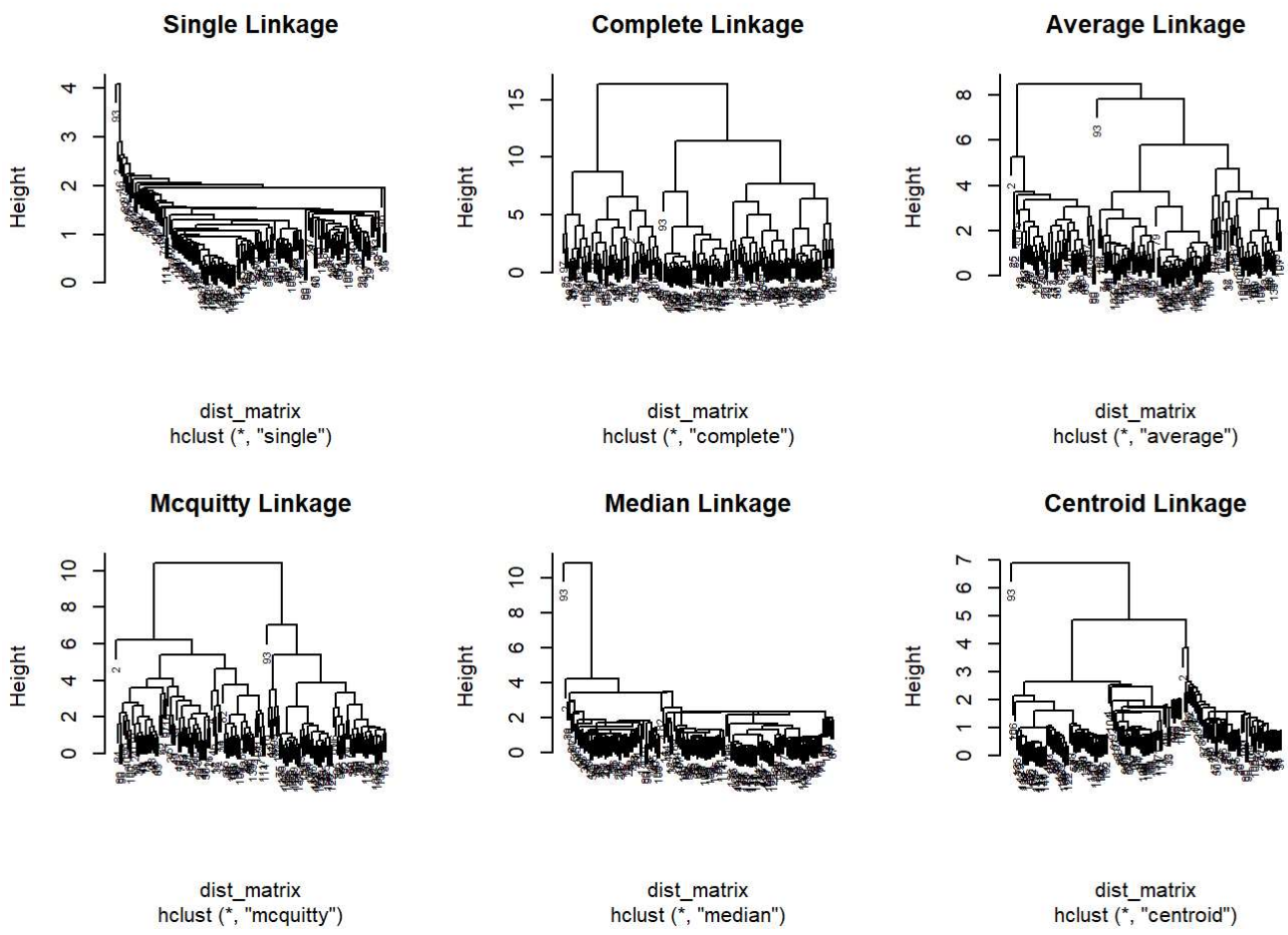
Based on the above plot, three clusters are chosen as the best fit, as there is a noticeable drop in W going from two to three clusters, but only a small improvement is gained by adding more clusters.

```
# Applying k-means clustering
kmeans_clusters <- kmeans(gap[,3:26], centers = 3)
fviz_cluster(kmeans_clusters, data = gap[,3:26], geom = "point", frame.type = "norm")
```



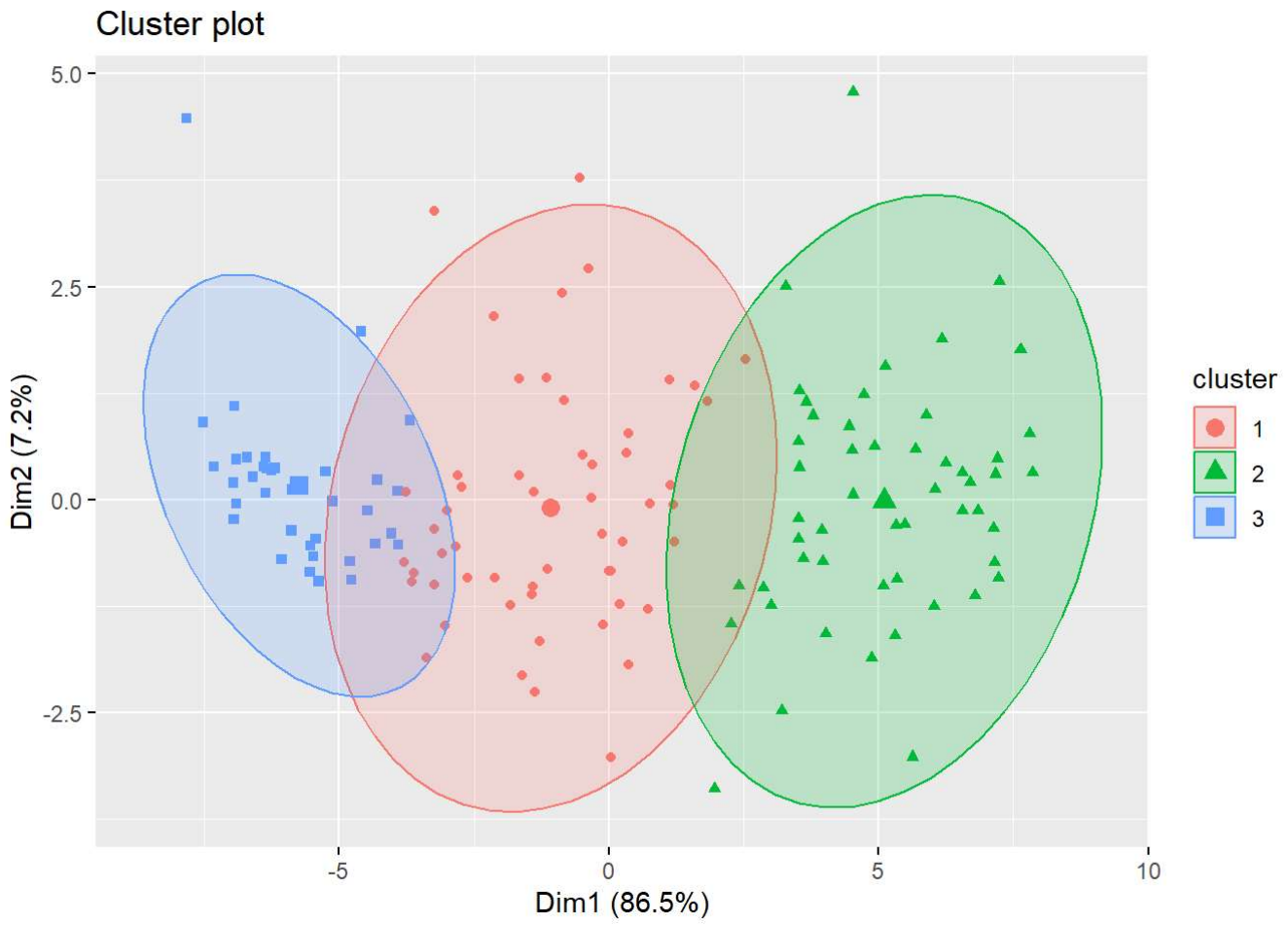
Agglomerative hierarchical clustering

```
gap.scaled <- gap
gap.scaled[,3:26] <- scale(gap[,3:26])

# distance matrix
dist_matrix <- dist(gap.scaled[,3:26])

# hierarchical clustering using different linkage methods
hc_single <- hclust(dist_matrix, method = "single")
hc_complete <- hclust(dist_matrix, method = "complete")
hc_average <- hclust(dist_matrix, method = "average")
hc_mcquitty <- hclust(dist_matrix, method = "mcquitty")
hc_median <- hclust(dist_matrix, method = "median")
hc_centroid <- hclust(dist_matrix, method = "centroid")

# Visualizing the dendrograms
par(mfrow=c(2,3))
plot(hc_single, cex = 0.5, main = "Single Linkage")
plot(hc_complete, cex = 0.5, main = "Complete Linkage")
plot(hc_average, cex = 0.5, main = "Average Linkage")
plot(hc_mcquitty, cex = 0.5, main = "Mcquitty Linkage")
plot(hc_median, cex = 0.5, main = "Median Linkage")
plot(hc_centroid, cex = 0.5, main = "Centroid Linkage")
```



Based on the dendrograms, we chose the complete linkage method and 3 clusters.

```
cutree_hc_complete <- cutree(hc_complete, k = 3)
fviz_cluster(list(data = gap.scaled[,3:26], cluster = cutree_hc_complete), geom = "point", frame.type = "norm")
```




```
# adjusted Rand index
adjusted_rand_index <- adjustedRandIndex(kmeans_clusters$cluster, cutree_hc_complete)
adjusted_rand_index
```

```
## [1] 0.4745206
```

After analysing the results of K-means and hierarchical clustering, it is evident that the clusters obtained from both methods are somewhat similar but not identical.

We also calculated the adjusted Rand index to measure the similarity between the two clustering solutions, which yielded a value of 0.4745206. This value suggests that there is some agreement between the two clustering solutions, but not a strong one.

The clusters found using K-means and hierarchical clustering also show that countries do not naturally cluster by continent, as we observe countries from different continents in each cluster.

Therefore, these clustering methods do not provide strong evidence for continental groupings of countries based on the chosen variables.

7.Linear regression:

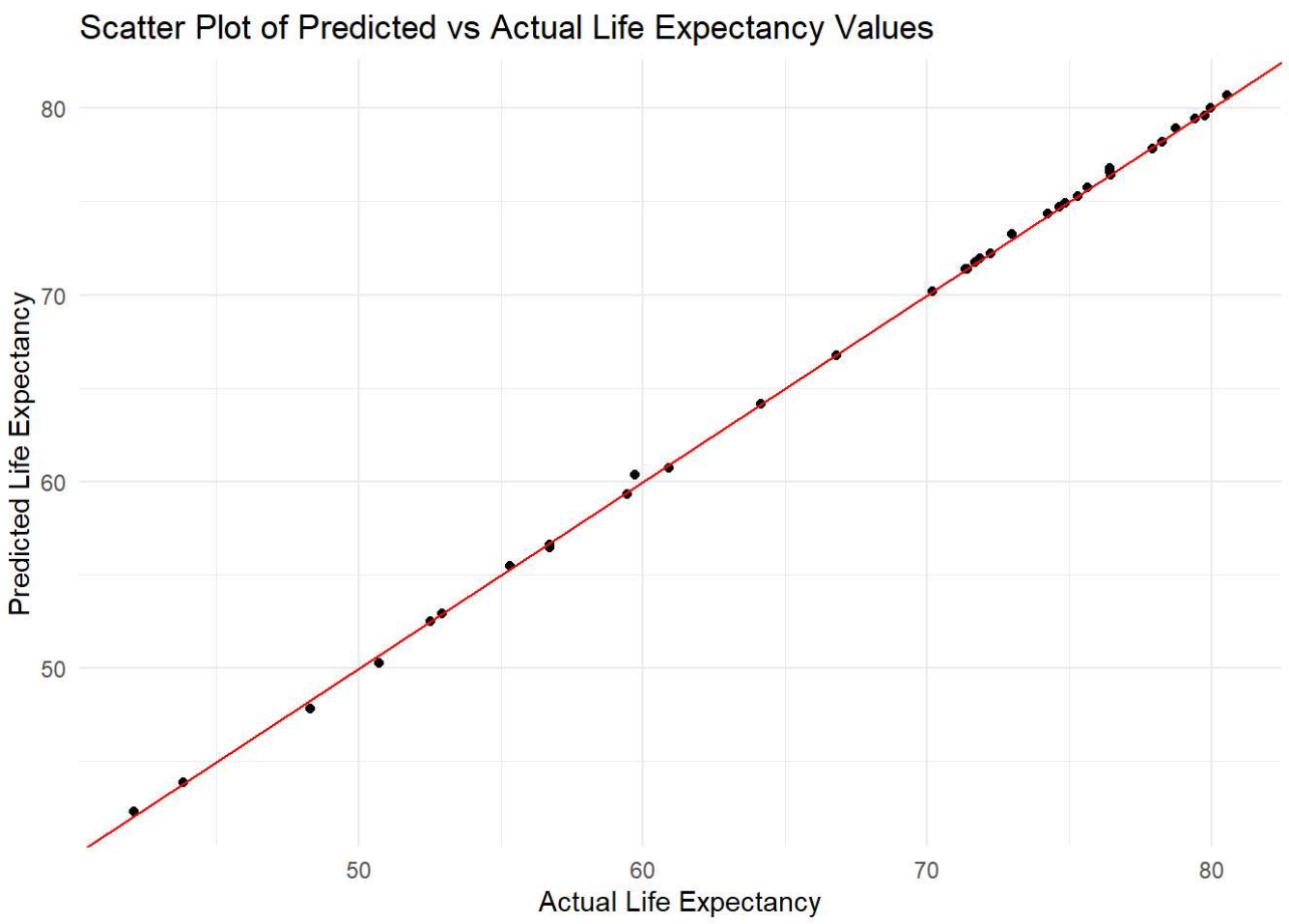
```
lifeExp_2007=gap$lifeExp_2007
gdp1=gap[,3:14]
data1=cbind(gdp1,lifeExp_2007)

model1 <- lm(lifeExp_2007 ~ ., data = data1)
summary(model1)
```

```
##
## Call:
## lm(formula = lifeExp_2007 ~ ., data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3801  -2.1474   0.6636   3.6927  13.5548
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.332      4.481   1.190  0.23619
## gdpPercap_1952  -5.601      6.398  -0.875  0.38301
## gdpPercap_1957  12.953      9.504   1.363  0.17531
## gdpPercap_1962  -5.724      9.908  -0.578  0.56447
## gdpPercap_1967   2.712      6.909   0.392  0.69536
## gdpPercap_1972  -6.033      6.084  -0.992  0.32329
## gdpPercap_1977  -2.736      6.018  -0.455  0.65018
## gdpPercap_1982  -2.603      7.559  -0.344  0.73111
## gdpPercap_1987   9.867      6.414   1.538  0.12644
## gdpPercap_1992  -8.113      5.715  -1.420  0.15810
## gdpPercap_1997  18.464      6.724   2.746  0.00689 **
## gdpPercap_2002 -12.441      7.872  -1.580  0.11648
## gdpPercap_2007   6.554      4.670   1.404  0.16286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.936 on 129 degrees of freedom
## Multiple R-squared:  0.698,  Adjusted R-squared:  0.6699
## F-statistic: 24.85 on 12 and 129 DF,  p-value: < 2.2e-16
```

```
p <- prcomp(data[, 1:length(data) - 1])
p_scores <- p$x[, 1:12]

train_index <- sample(1:nrow(p_scores), nrow(p_scores) * 0.75)
train <- p_scores[train_index, ]
test <- p_scores[-train_index, ]
train_data <- data.frame(y = data[train_index, ]$lifeExp_2007, x = train)
fit1 <- lm(y ~ ., data = train_data)

testing1 <- data.frame(x = test)
test_pred <- predict(fit1, testing1)
errors <- test_pred - data[-train_index, ]$lifeExp_2007
acc=sqrt(mean(errors^2))
print(paste("Accuracy: Average difference between the predicted and actual life expectancy values is about",acc," years" ))
```

```
## [1] "Accuracy: Average difference between the predicted and actual life expectancy values is about 0.19790534834044  year
s"
```

```
#Plotting
test_df <- data.frame(actual = data[-train_index, ]$lifeExp_2007, predicted = test_pred)
ggplot(test_df, aes(x = actual, y = predicted)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(x = "Actual Life Expectancy", y = "Predicted Life Expectancy",
       title = "Scatter Plot of Predicted vs Actual Life Expectancy Values") +
theme_minimal()
```



We evaluated the accuracy of various linear regression models in predicting life expectancy from GDP figures after fitting and testing them.

Based on the R-squared value, RMSE, and AIC values of the models, our findings show that the multiple linear regression model is more accurate than other linear regression models. As a result, it appears that using all years of GDP as independent variables is a stronger predictor of life expectancy.