Examining the Relationship of GDP on Life Expectancy: A Global Perspective

Group 06

1 Introduction

The analysis is aimed at investigate the relationship between Life Expectancy at birth and the Gross Domestic Product (GDP) of nations in the year 2021. Data acquisition and consolidation were performed using information sourced from the website ourworldindata.org. The designated topic for our group is 'Poverty and Economic Development'. This research aims to examine the linear association between Life Expectancy at birth and the Gross Domestic Product of countries in 2021, while also exploring the potential influence of continents.

In particular, this report presents numerical and graphical summaries of the life expectancy and GDP of the nations for the year 2021 along with their continents.

```
# Import the data sets
life_expectancy <- read_csv("life-expectancy.csv")
gdp_per_capita_worldbank <- read_csv("gdp-per-capita-worldbank.csv")

#Exporting first data set to excel file#
write.csv(life_expectancy, file = "Group_06_Data_01.csv")

#Exporting second data set to excel file#
write.csv(gdp_per_capita_worldbank, file = "Group_06_Data_02.csv")
```

2 Data Wrangling Methods

Before we begin the analysis of our data, let's transform the data using various tools.The process below describes the detailed data wrangling techniques that are used to get the desired data set.

On further examination of the two data sets, the latest year was considered as the year of interest.

```
gdp <- gdp_per_capita_worldbank %>%
  group_by(Entity) %>%
  slice(which.max(Year)) #gdp has 208 rows and 4 columns#
unique_year_gdp <- unique(gdp$Year)
```

Let's examine a subset of the gdp_per_capita_worldbank data set. We use the *group_by( )* function to group the 'Entity' and *slice( )* function to choose only the observations pertaining to the latest year. Using the *unique( )* function,the latest years for entities are found to be 2018, 2019, 2020 and 2021.

```
lifeExp <- life_expectancy %>%
  group_by(Entity) %>%
  slice(which.max(Year)) #This data set has 261 rows and 4 columns#
unique_year_lifeExp<- unique(lifeExp$Year)
lifeExp_1949_1913 <- lifeExp[lifeExp$Year== '1949' | lifeExp$Year == '1913', ]
```

Similarly, for life_expectancy data set, the latest years for entities were found to be 2021, 1949 and 1913. It can be observed that 1949 and 1913 are not recent years. The countries corresponding to these years are checked and found to be England and Wales, Northern Ireland, Scotland an USSR.

```
lifeExp_2021 <- lifeExp %>%
  filter(Year == 2021)#This data set has 257 rows and 4 columns#
```

We use *filter( )* function to choose only the observations pertaining to 2021.

```
df1 <- gdp %>%
  inner_join(lifeExp_2021, by = c("Entity", "Code", "Year"))
#This data set has 196 rows and 5 columns#
unique_year_df1 <- unique(df1$Year)
```

It is seen that the 'Entity' variable in the two data sets have certain clusters (like 'Europe and Central Asia (WB)', 'South Asia' etc.). We merge the two data sets using *inner_join( )* function. We use 'Code' as one of the key variables to address the aforementioned anomaly.

```
rows_with_na <- df1[is.na(df1$Code), ] #This data set has 196 rows and 5 columns#

df2 <- na.omit(df1) #This data set has 191 rows and 5 columns#
```

Let's check if the merged data set has any missing values. Here, we can see rows with entities 'High-income countries', 'Low-income countries', 'Lower-middle-income countries', 'Middle-income countries' and 'Upper-middle-income countries'. Since our research pertains to analyzing countries, we omit these rows from the data set.

```
years_of_interest <- c('2020', '2018', '2019')
rows_of_interest <- gdp[gdp$Year %in% years_of_interest, ]
```

However, the first data set gdp also includes years 2018, 2019, and 2020. The result shows rows for Entities such as 'Kuwait', 'San Marino', 'Sint Maarten (Dutch part)', and 'Turkmenistan'. The aim is analyse the data sets for year 2021, so these rows are omitted.

```r
mycodes <- df2$Code

final_df <- data.frame(country = df2$Entity,
                       country_code = mycodes,
                       GDP = df2$`GDP per capita, PPP (constant 2017 international $)`,
                       life_exp = df2$`Period life expectancy at birth - Sex: all - Age: 0`,
                       continent = countrycode(sourcevar = mycodes, origin = "iso3c",
                                               destination = "continent"))
check_na <- colSums(is.na(final_df)) #Two rows have missing values for 'Continent'#
missing_continents <- final_df[is.na(final_df$continent), ]
lifeExp_gdp <- na.omit(final_df) #This data set has 189 rows and 5 columns#

write.csv(lifeExp_gdp, file = "Group_06_Data.csv") #Exporting data to excel file#
```

A new data frame is formed to list all the Countries by their respective continents. A new column containing the respective continent is added. Upon further inspection, 'Kosovo' and 'World' have missing values for the variable 'Continent'. Due to reasons, beyond the scope of this research, data for Country 'Kosovo' is excluded from the data set. To keep the data structure consistent, we remove 'World' as well.

3 Exploratory Data Analysis

Let's have a look at the first five rows of the data frame

Table 1: Glimpse of the first five rows in the data set

country	country_code	GDP	life_exp	continent
Afghanistan	AFG	1,516.31	61.98	Asia
Albania	ALB	14,518.91	76.46	Europe
Algeria	DZA	11,039.81	76.38	Africa
Angola	AGO	5,908.57	61.64	Africa
Antigua and Barbuda	ATG	19,124.43	78.50	Americas

The Table 1 provides us with the information as follows:

- **country** : Represents the country.
- **country_code** : country codes of the corresponding country. It is defined by International Organization for Standardization.
- **GDP** : Gross domestic product.This value is adjusted for inflation and for differences in the cost of living between countries. This data is expressed in international-$ at 2017 prices.
- **life_exp** : Period life expectancy at birth. It is expressed in years.
- **continent** : Represents the continent of the corresponding country.

3.1 Summary Statistics

```
lifeExp_gdp |>
  summarize('Mean' = mean(life_exp),
            'Median' = median(life_exp),
            'St.Dev' = sd(life_exp),
            'Min' = min(life_exp),
            'Max' = max(life_exp),
            'IQR' = quantile(life_exp,0.75)-quantile(life_exp,0.25),
            'Sample_size' = n(),
            .by = continent) |>
  gt() |>
  fmt_number(decimals=2) |>
  cols_label(
    Mean = html("Mean"),
    Median = html("Median"),
    St.Dev = html("Std. Dev"),
    Min = html("Minimum"),
    Max = html("Maximum"),
    IQR = html("IQR"),
    Sample_size = html("Sample Size")
  )
```

Table 2: Summary statistics on life expectancy by continent of 189 countries.

continent	Mean	Median	Std. Dev	Minimum	Maximum	IQR	Sample Size
Asia	74.07	73.47	5.77	61.98	85.47	9.13	45.00
Europe	78.24	80.11	4.42	68.85	83.99	7.66	39.00
Africa	62.86	61.65	5.62	52.53	76.38	6.09	52.00
Americas	73.21	72.83	4.18	63.19	82.66	4.95	39.00
Oceania	70.11	68.88	6.33	63.62	84.53	5.40	14.00

The Table 2 provide us with the Summary statistics of Life Expectancy respectively:

- **continent** : Represents the continent of the corressponding country.
- **Mean** : Mean or average (in years)
- **Median** : Median or the 2^{nd} quartile or 50^{th} percentile (in years)
- **Std.Dev** : the standard deviation (in years)
- **Minimum** : the minimum value or the 0^{th} percentile (in years)
- **Maximum** : the maximum value or 100^{th} percentile (in years)
- **IQR** : Interquartile Range or IQR is the measure of spread of middle 50% of values (in years). It is calculated by subtracting the 1^{st} quartile from 3^{rd} quartile.
- **Sample Size** : the total number of countries in that particular continent.

Europe has the highest average life expectancy (78.24 years), followed by Asia (74.07 years), the Americas (73.21 years), Oceania (70.11 years), and Africa (62.86 years). This suggests that, on average, people tend to live longest in Europe and shortest in Africa.

Africa has higher standard deviation (5.62), indicating relatively greater variability in life expectancy across African countries. Europe has comparatively low standard deviation (4.42), suggesting more consistent life expectancy across European countries.

In summary, Europe generally exhibits higher average and median life expectancy with lower variability compared to other continents. Africa, on the other hand, shows lower average and median life expectancy with higher variability, indicating disparities in healthcare access and socio-economic factors. The Americas and Asia show moderate average life expectancy with low variability for Americas and high variability for Asia. Oceania, with a smaller sample size, shows moderate average life expectancy.

The Table 3 provide us with the Summary statistics of GDP respectively:

- **continent** : Represents the continent. This constitutes that countries that belong to that continent.
- **Mean** : Mean or average (in $)
- **Median** : Median or the 2^{nd} quartile or 50^{th} percentile (in $)
- **Std.Dev** : the standard deviation (in $)
- **Minimum** : the minimum value or the 0^{th} percentile (in $)
- **Maximum** : the maximum value or 100^{th} percentile (in $)
- **IQR** : Interquartile Range is the measure of spread of middle 50% of values (in years). It is calculated by subtracting the 1^{st} quartile from 3^{rd} quartile.
- **Sample Size** : the total number of countries in that particular continent.

Table 3: Summary statistics on GDP by continent of 189 countries.

continent	Mean	Median	Std. Dev	Minimum	Maximum	IQR	Sample Size
Asia	24,687.91	14,193.12	24,807.56	1,516.31	106,032.23	33,049.55	45.00
Europe	41,075.95	38,717.69	21,844.46	12,943.61	115,683.49	23,705.53	39.00
Africa	5,737.47	3,279.17	6,000.30	705.03	28,760.52	4,494.70	52.00
Americas	21,873.94	18,512.46	17,080.51	2,870.14	80,271.13	12,546.79	39.00
Oceania	11,818.82	5,747.97	15,124.30	1,937.09	49,774.34	8,144.56	14.00

Europe has the highest average GDP among the continents, with a mean GDP of $41,075.95. Africa has the lowest average GDP ($5,737.47), indicating lower economic development compared to other continents. The Americas ($21,873.94) and Asia ($24,687.91) have similar average GDP, both higher than Africa but lower than Europe. Oceania's average GDP ($11,818.82) falls between the lower GDP of Africa and the Americas/Asia.

Africa exhibits the highest variability in GDP ($6,000.30), indicating significant differences in economic output among their countries. Oceania also displays high GDP variability ($15,124.30), reflecting economic disparities within the region.The Americas show lower GDP variability compared to Africa and Europe ($21,844.46), suggesting more uniform economic performance across countries within these continents.

In summary, Europe generally exhibits the highest economic output with the low variability among its nations. Africa shows the lowest economic output with considerable variability among countries. The Americas and Asia fall in between, with moderate economic output and variability. Oceania, while having a smaller sample size, shows moderate economic output and variability similar to the Americas and Asia.

3.2 Correlation

```
cor <- lifeExp_gdp %>%
  get_correlation(formula = life_exp ~ GDP)
#0.7371373 which shows strong positive linear relationship#
```

The correlation coefficient is approximately 0.74 indicating a strong positive linear relationship between life expectancy and GDP.

3.3 Visualization

From the histogram represented in Figure 1 below, it is evident that the "GDP" data exhibits skewness. To reduce this skewness and facilitate more robust analysis, a logarithmic transformation is used.

Figure 1: Understanding Data Structures

```r
scatterplot_1 <- ggplot(lifeExp_gdp, mapping= aes(x=GDP,
                                                  y = life_exp,color=continent))+
  geom_point(size = 2, shape = 19) +
  labs(x="GDP", y = "Life Expectancy (years)", color = "Continent") +
  scale_color_manual(values = c("#339900","#CC0000","#3399CC","orange","#9933CC"))+
  theme_minimal()
scatterplot_2 <- ggplot(lifeExp_gdp, mapping= aes(x=log(GDP),
                                                  y = life_exp,color=continent))+
  geom_point(size = 2, shape = 19) +
  labs(x="log(GDP)", y = "Life Expectancy (years)", color = "Continent")+
  scale_color_manual(values = c("#339900","#CC0000","#3399CC","orange","#9933CC"))+
  theme_minimal()
grid.arrange(scatterplot_1, scatterplot_2)
```

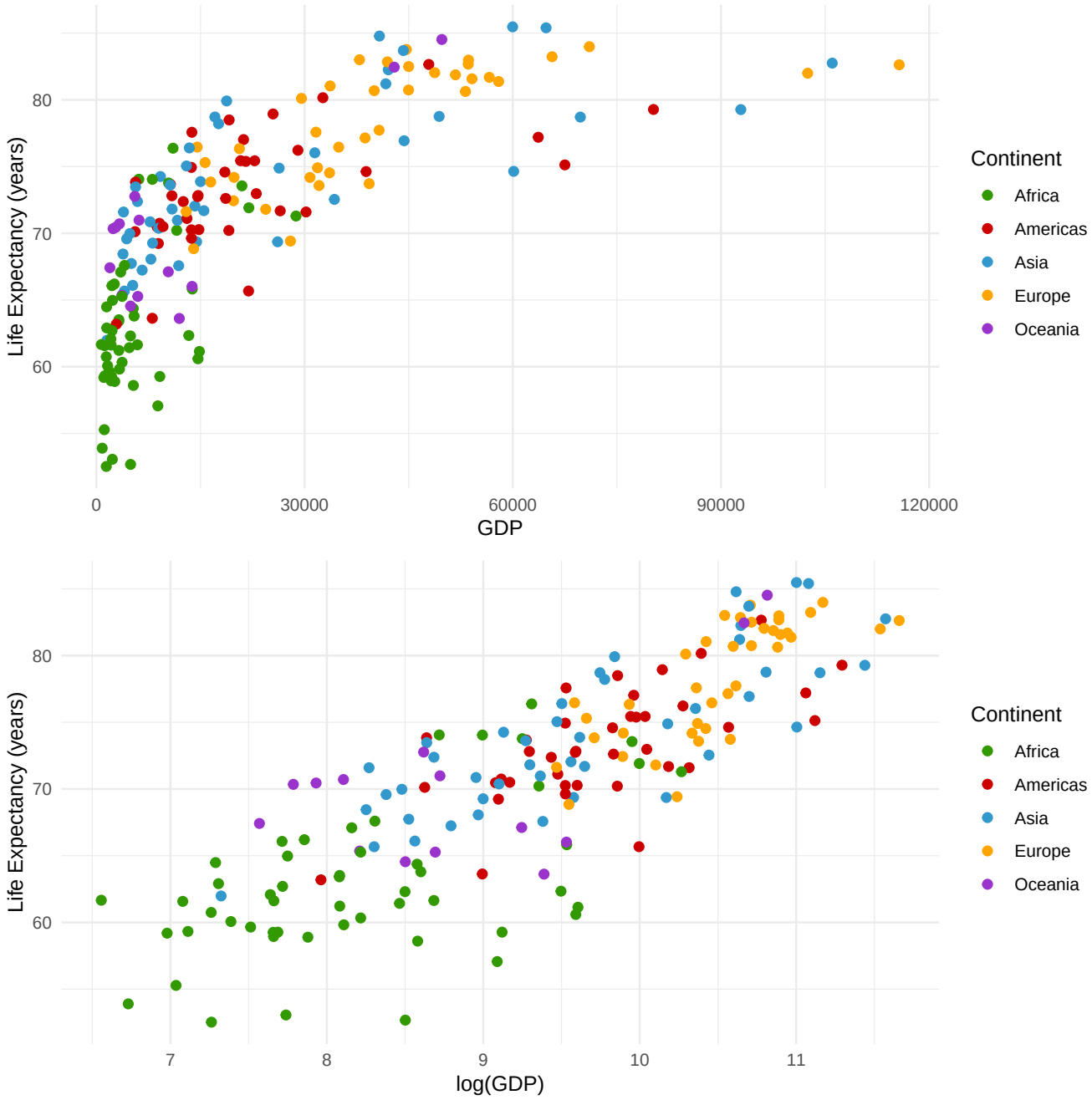

Figure 2: Relationship between Life Expectancy and GDP of countries.

Figure 2 shows a strong positive linear relationship between life expectancy and log(GDP) as suggested by the correlation coefficient (0.74). The comparison of the two scatter plots shows that the transformation of the data using log exhibits a linear relationship.

```r
life_exp_outliers <- lifeExp_gdp %>%
  group_by(continent) %>%
  mutate(is_outlier = life_exp > quantile(life_exp, 0.75) + 1.5 * IQR(life_exp) |
           life_exp < quantile(life_exp, 0.25) - 1.5 * IQR(life_exp)) %>%
  filter(is_outlier) %>%
  ungroup()
p1 <- ggplot(lifeExp_gdp, aes(x = continent, y = life_exp))+
  geom_boxplot(fill=c("#339900","#CC0000","#3399CC","orange","#9933CC")) +
  geom_text(data = life_exp_outliers, aes(label = country), vjust=0.1 ,hjust=1.1)+
  labs(x = "Continent", y = "Life expectancy (years)")+
  theme_minimal()
print(p1)
```

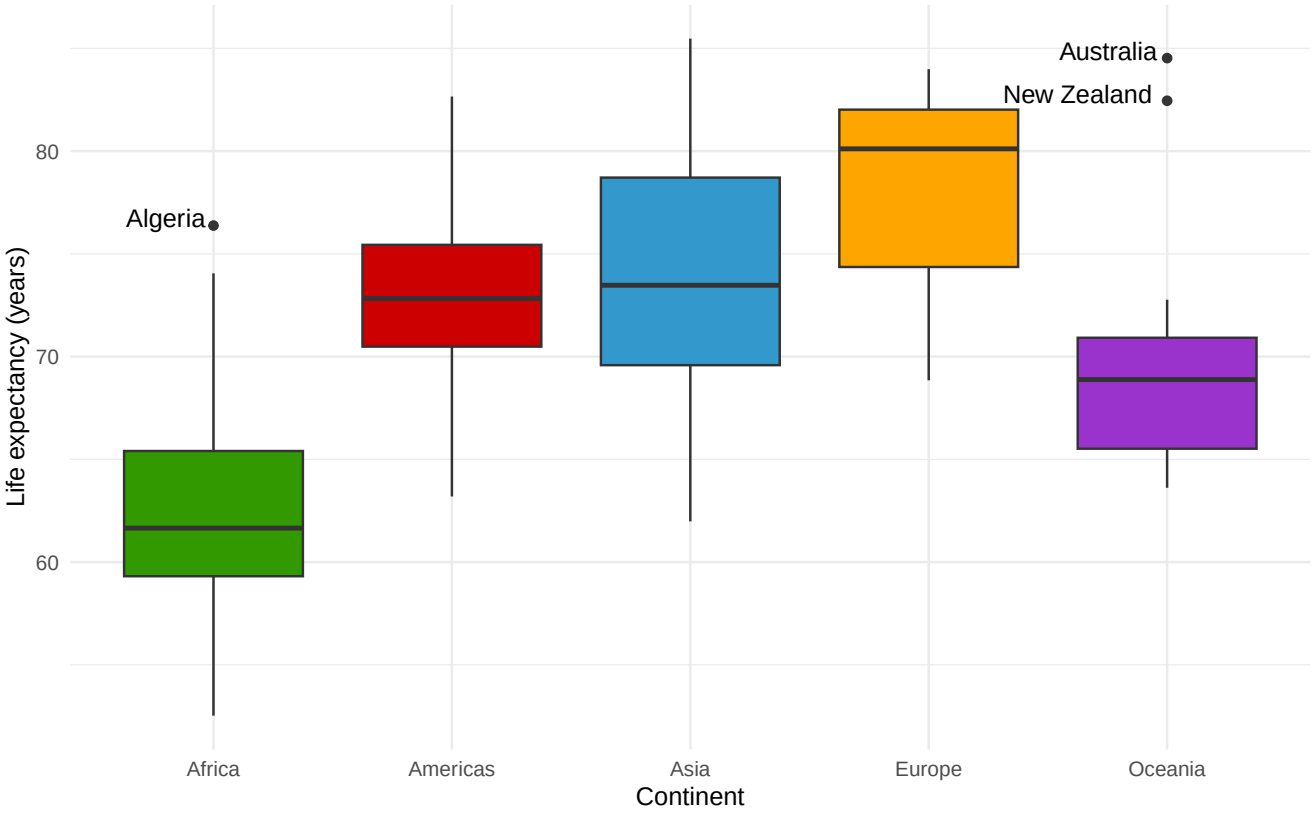

Figure 3: Boxplot of Life Expectancy of the continents

In Figure 3 it is seen that the middle 50% of the life expectancy distribution of Africa is much smaller than the middle 50% of the remaining four continents. There is more variability in life expectancy in the continents of Africa, Asia and Europe. In the continent Europe, due to presence of longer whisker(below) indicates skewness in direction below. Africa has an Algeria as an outlier. Oceania has two outliers namely Australia and New Zealand.

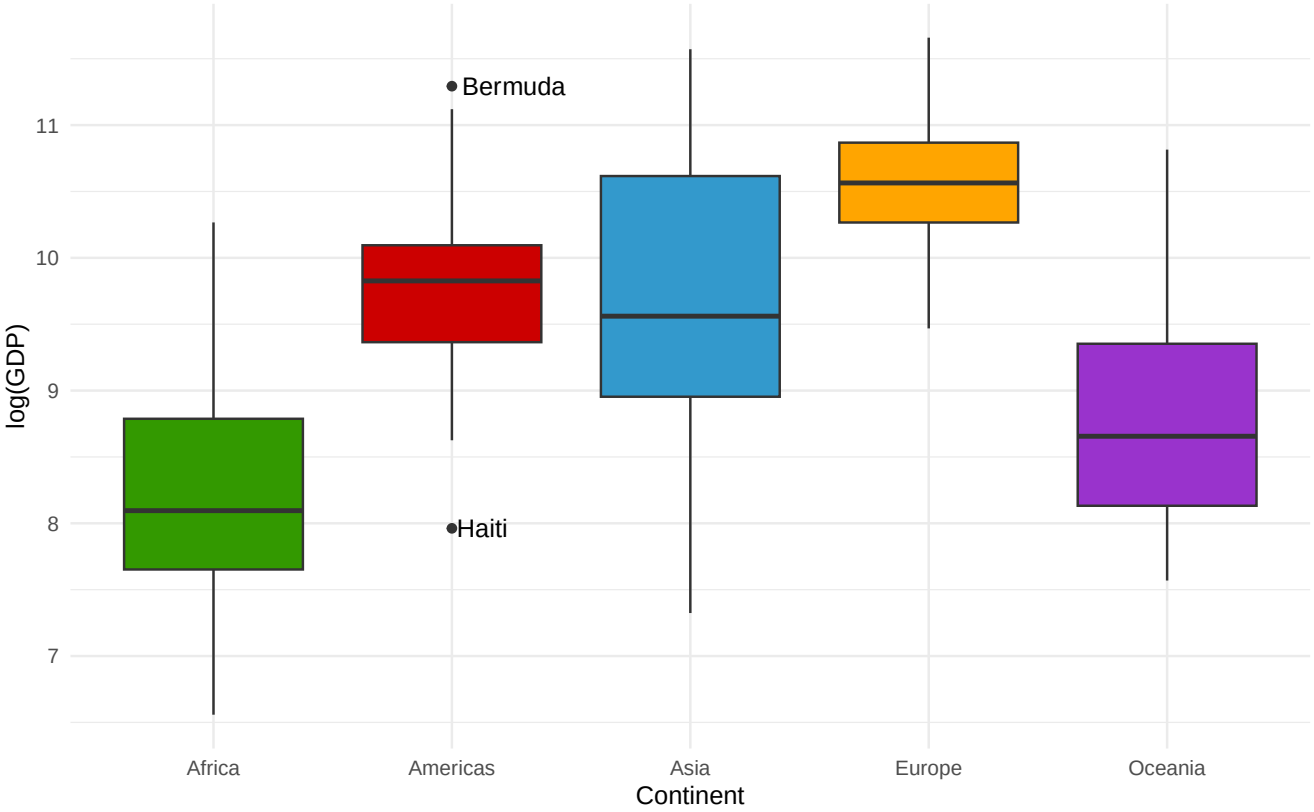

Figure 4: Boxplot of log(GDP) of the continents

Similarly, in Figure 4, it is seen that the middle 50% of the distribution of Africa is much smaller than the middle 50% Americas, Asia and Europe. However, there is a small difference in comparison to Oceania. There is more variability in the continents of Africa, Asia and Oceania. Here, Americas have two outliers namely Bermuda and Haiti.