Data Analysis with Hypothesis Testing and Linear Regression: Exploring Factors Affecting Life Expectancy

..., Alina Yildir, ...

2024-09-30

```r
library(ggplot2)
library(ggtext)
library(dplyr)
library(tidyr)
library(reshape2)

# We will use the following color codes for our visualizations:

# "#2e6a57" => Life Expectancy (Target Variable)
# "#9b2542" => Health Risk Factors and Mortality Indicators
# "#6F4E37" => Vaccination and Disease Control Indicators
# "#2b457e" => Healthcare Expenditure and Resource Indicators
# "#d59d3f" => Socio-Economic and Educational Indicators
```

```r
dataset <- read.csv("life_expectancy_data_raw.csv")
head(dataset)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
## 6 Afghanistan 2010 Developing            58.8             279            74
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                83
## 2    0.01              73.523582          62     492 18.6                86
## 3    0.01              73.219243          64     430 18.1                89
## 4    0.01              78.184215          67    2787 17.6                93
## 5    0.01               7.097109          68    3013 17.2                97
## 6    0.01              79.679367          66    1989 16.7               102
##   Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 1     6              8.16         65      0.1 584.25921   33736494
## 2    58              8.18         62      0.1 612.69651     327582
## 3    62              8.13         64      0.1 631.74498   31731688
## 4    67              8.52         67      0.1 669.95900    3696958
## 5    68              7.87         68      0.1  63.53723    2978599
## 6    66              9.20         66      0.1 553.32894    2883167
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
```

```
## 1                17.2              17.3                           0.479
## 2                17.5              17.5                           0.476
## 3                17.7              17.7                           0.470
## 4                17.9              18.0                           0.463
## 5                18.2              18.2                           0.454
## 6                18.4              18.4                           0.448
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
## 6       9.2
```

Exploratory Data Analysis

Univariate Analysis

```
str(dataset)
```

```
## 'data.frame':    2938 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

The dataset contains 2,938 observations and 22 variables, capturing a range of demographic, health, and economic indicators for various countries.

```
summary(dataset)
```

```
##    Country               Year         Status          Life.expectancy
```

```
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  Adult.Mortality infant.deaths       Alcohol        percentage.expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##   Hepatitis.B       Measles               BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##  NA's   :19      NA's   :226       NA's   :19
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
##  1st Qu.:   463.94   1st Qu.:1.958e+05   1st Qu.: 1.60
##  Median :  1766.95   Median :1.387e+06   Median : 3.30
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  5910.81   3rd Qu.:7.420e+06   3rd Qu.: 7.20
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  NA's   :448         NA's   :652         NA's   :34
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.50      1st Qu.:0.4930                  1st Qu.:10.10
##  Median : 3.30      Median :0.6770                  Median :12.30
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7790                  3rd Qu.:14.30
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
##  NA's   :34         NA's   :167                     NA's   :163
```

```
colSums(is.na(dataset))
```

```
##                         Country                            Year
##                               0                               0
##                          Status                 Life.expectancy
##                               0                              10
```

```
##                  Adult.Mortality                      infant.deaths
##                               10                                  0
##                          Alcohol          percentage.expenditure
##                              194                                  0
##                     Hepatitis.B                            Measles
##                              553                                  0
##                              BMI                 under.five.deaths
##                               34                                  0
##                            Polio               Total.expenditure
##                               19                                226
##                       Diphtheria                          HIV.AIDS
##                               19                                  0
##                              GDP                          Population
##                              448                                652
##             thinness..1.19.years              thinness.5.9.years
##                               34                                 34
## Income.composition.of.resources                         Schooling
##                              167                                163
```

```
num_duplicates <- nrow(dataset[duplicated(dataset), ])
num_duplicates
```

```
## [1] 0
```

The dataset has no duplicate rows (i.e., every row is unique based on the combination of all columns).

We will briefly overview each column, highlighting its potential influence on life expectancy, the target variable.

The `Country` variable contains 2,938 entries representing different countries over various years, with each country listed multiple times based on the year. The dataset includes 193 distinct countries; this variable has no missing values. The `Country` variable allows for **analyzing variations in life expectancy among nations**.

```
num_unique_countries <- length(unique(dataset$Country))
num_unique_countries
```

```
## [1] 193
```

The `Year` variable ranges from 2000 to 2015, covering 15 years. There are no missing values in this variable. The `Year` variable allows for **examining how life expectancy has changed over time**.

The `Life.expectancy` variable, **the target variable in this analysis**, ranges from 36.3 to 89 years, with a mean of 69.22 and a median of 72.1. Since the median is higher than the mean, this suggests a slight left skew in the distribution, meaning some younger ages are pulling the mean down. There are 10 missing values in this variable.

```
life_expectancy_cleaned <- dataset %>%
  filter(!is.na(Life.expectancy))

ggplot(life_expectancy_cleaned, aes(x = "", y = Life.expectancy)) +
  geom_boxplot(fill = "#2e6a57", color = "black") +
  labs(title = "Distribution of Life Expectancy",
```

```
      x = "", y = "Life Expectancy (Years)") +
theme_minimal() +
theme(
  plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
  axis.title.x = element_blank(),
  axis.title.y = element_text(size = 12)
)
```

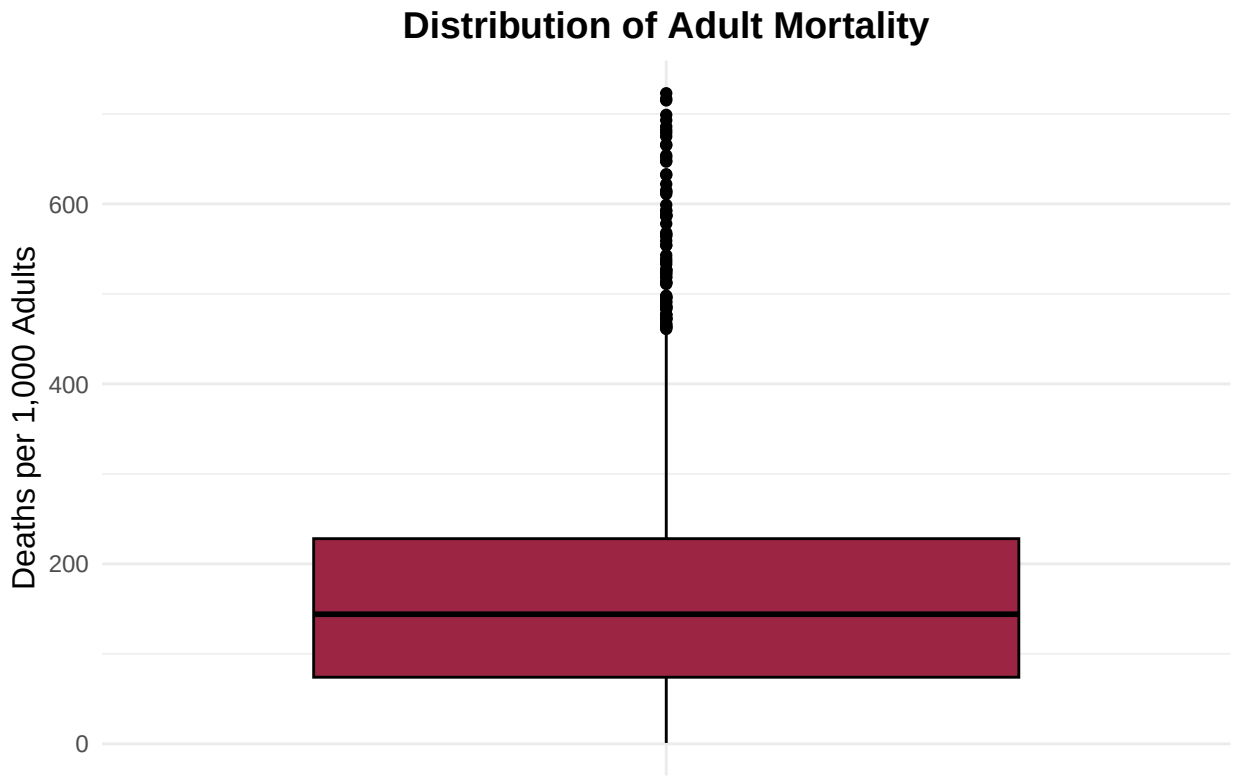
We have grouped the rest of the dataset variables into four categories:

- Health Risk Factors and Mortality Indicators
- Vaccination and Disease Control Indicators
- Healthcare Expenditure and Resource Indicators
- Socio-Economic and Educational Indicators

Health Risk Factors and Mortality Indicators

The `Adult.Mortality` variable represents the number of deaths of individuals between 15 and 60 years old per 1,000 population. The values range from 1 to 723 deaths per 1,000 adults, with a mean of 164.8 and a median of 144. This suggests that the distribution is skewed to the right, indicating the presence of outliers at the higher end of the scale. There are 10 missing values in this variable. The `Adult.Mortality` variable allows for **assessing the impact of adult mortality on life expectancy**.

```r
adult_mortality_cleaned <- dataset %>%
  filter(!is.na(Adult.Mortality))

ggplot(adult_mortality_cleaned, aes(x = "", y = Adult.Mortality)) +
  geom_boxplot(fill = "#9b2542", color = "black") +
  labs(title = "Distribution of Adult Mortality",
       x = "", y = "Deaths per 1,000 Adults") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```



The `under.five.deaths` variable measures the number of deaths of children under five per 1,000 population. The mean value is 42.04, with a median of 4. The much higher mean compared to the median indicates a positively skewed (right-skewed) distribution. This means most regions have low child mortality rates, but a few have much higher rates. There are no missing values in this variable. The `under.five.deaths` variable allows for **analyzing the relationship between child mortality and life expectancy**.

The `infant.deaths` variable represents the number of infant deaths per 1,000 population, with a mean of 23.94 and a median of 3. This significant difference between the mean and median suggests a highly right-skewed distribution, indicating the presence of outliers at the higher end of the scale. There are no missing values in this variable. The `infant.deaths` variable allows for **assessing the impact of infant mortality on life expectancy**.

```r
infant_deaths_cleaned <- dataset %>%
  filter(!is.na(infant.deaths),
         infant.deaths <= 1000)

mean(infant_deaths_cleaned$infant.deaths)
```

```
## [1] 23.94291
```

```r
median(infant_deaths_cleaned$infant.deaths)
```

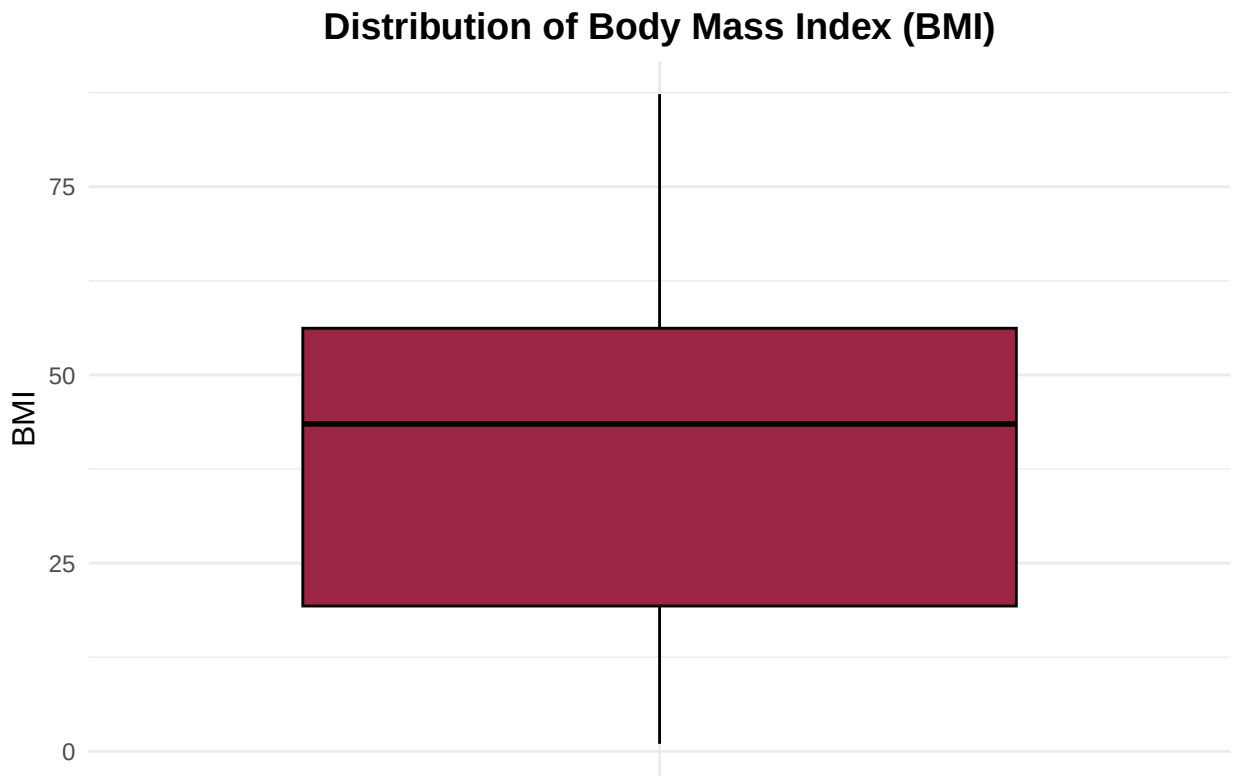
```
## [1] 3
```

```r
ggplot(infant_deaths_cleaned, aes(x = "", y = infant.deaths)) +
  geom_boxplot(fill = "#9b2542", color = "black") +
  labs(title = "Distribution of Infant Mortality",
       x = "", y = "Deaths per 1,000 Infants") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```

The `BMI` variable represents the entire population's average Body Mass Index (BMI). The mean BMI is 38.32, with a median of 43.5. The median being higher than the mean suggests a left-skewed distribution. This

indicates that while many individuals have higher BMI values, there are some with significantly lower BMI values, which pull the mean down. There are 34 missing values in this variable. The `BMI` variable allows for **exploring how BMI levels in a population correlate with life expectancy**.

```
bmi_cleaned <- dataset %>%
  filter(!is.na(BMI))

ggplot(bmi_cleaned, aes(x = "", y = BMI)) +
  geom_boxplot(fill = "#9b2542", color = "black") +
  labs(title = "Distribution of Body Mass Index (BMI)",
       x = "", y = "BMI") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```

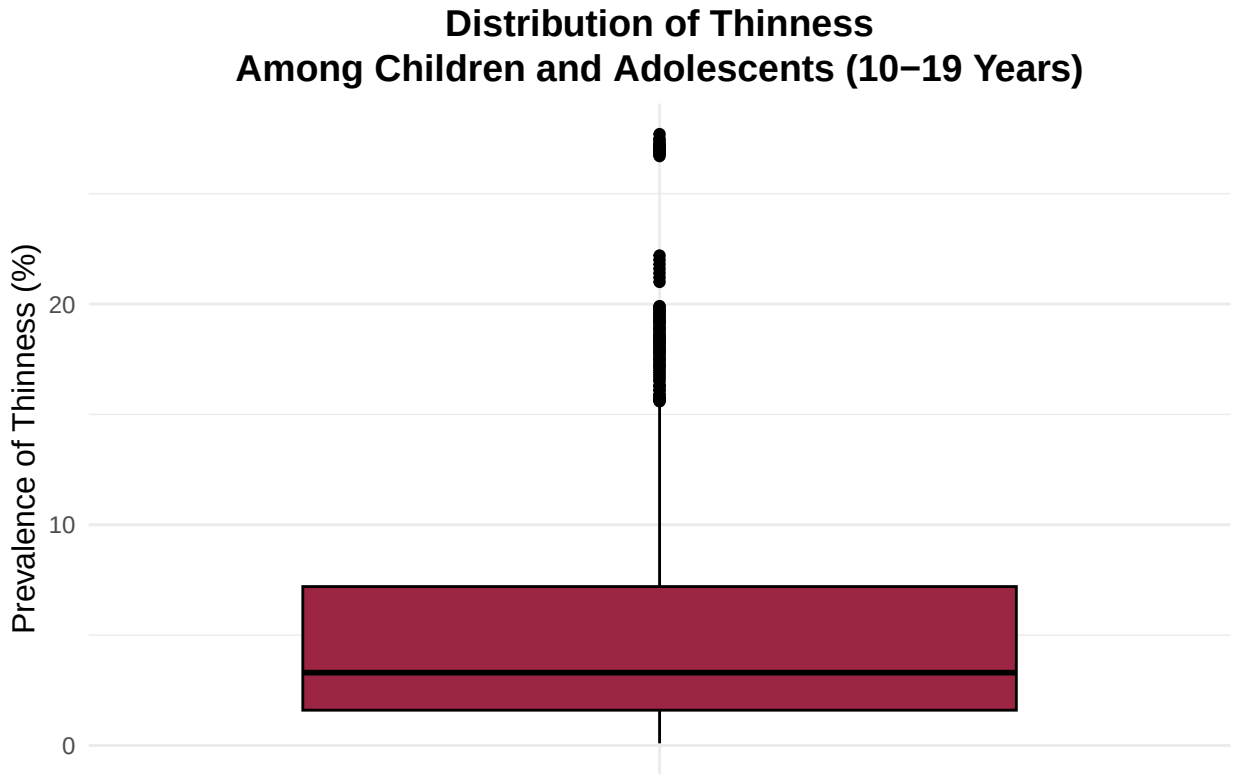


The `HIV.AIDS` variable represents deaths per 1,000 live births among children aged 0-4 due to HIV/AIDS. The minimum value is 0.1, with the first quartile at 0.1, the median also at 0.1, and the third quartile at 0.8. There are no missing values in this variable. The `HIV.AIDS` variable allows for **assessing how the HIV/AIDS epidemic affects child survival rates and, consequently, contributes to overall life expectancy**.

The `thinness..1.19.years` variable captures the prevalence of thinness (underweight) among children and adolescents aged 10 to 19, expressed as a percentage. The mean value is 4.84%, with a median of 3.3%. The higher mean compared to the median suggests the presence of outliers with high prevalence rates of thinness.

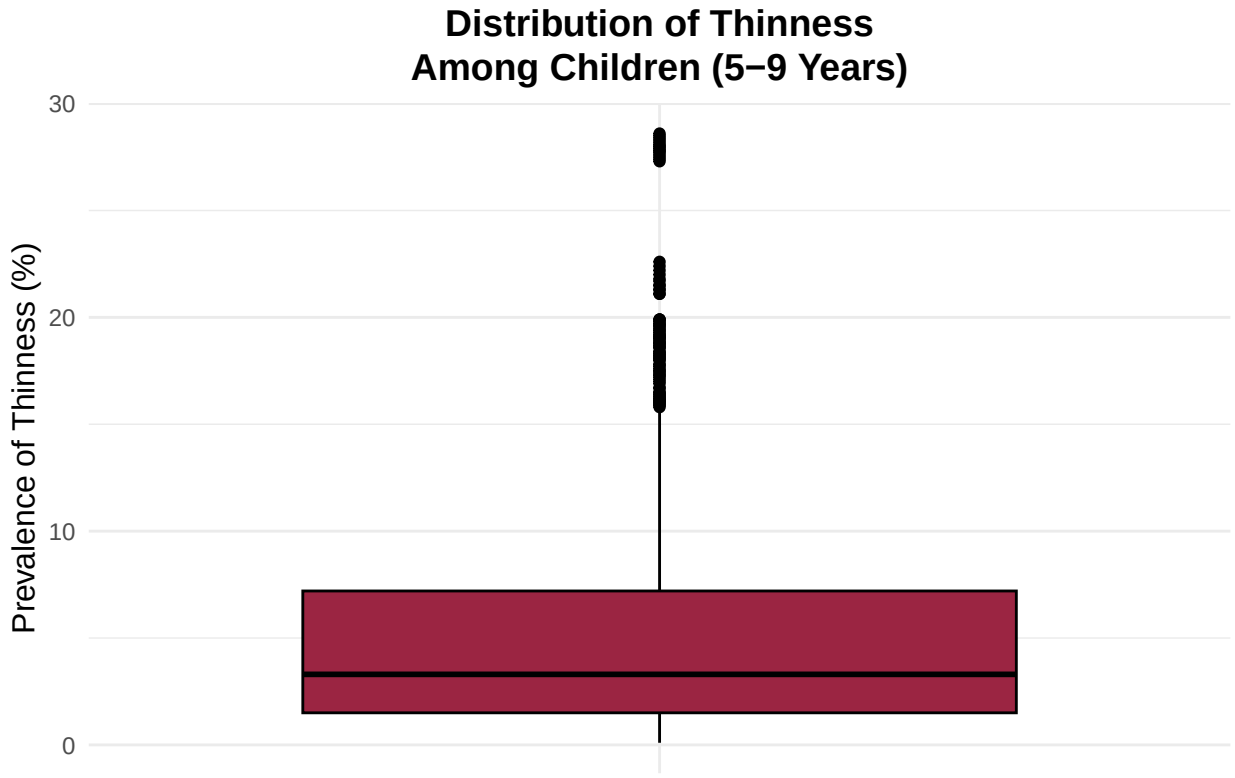
These outliers significantly affect the overall average. There are 34 missing values in this variable. The `thinness..1.19.years` variable allows for **assessing the effect of malnutrition on life expectancy**.

```
thinness_10_19_years_cleaned <- dataset %>%
  filter(!is.na(thinness..1.19.years))

ggplot(thinness_10_19_years_cleaned, aes(x = "", y = thinness..1.19.years)) +
  geom_boxplot(fill = "#9b2542", color = "black") +
  labs(title = "Distribution of Thinness<br>
       Among Children and Adolescents (10-19 Years)",
       x = "", y = "Prevalence of Thinness (%)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```



The `thinness.5.9.years` variable measures the prevalence of thinness (underweight) among children aged 5 to 9, expressed as a percentage. The mean value is 4.87%, with a median of 3.3%. The difference between the mean and median suggests variability in the data. The presence of regions with higher prevalence rates contributes to this spread. There are 34 missing values in this variable. The `thinness.5.9.years` variable allows for **evaluating the relationship between childhood malnutrition and life expectancy**.

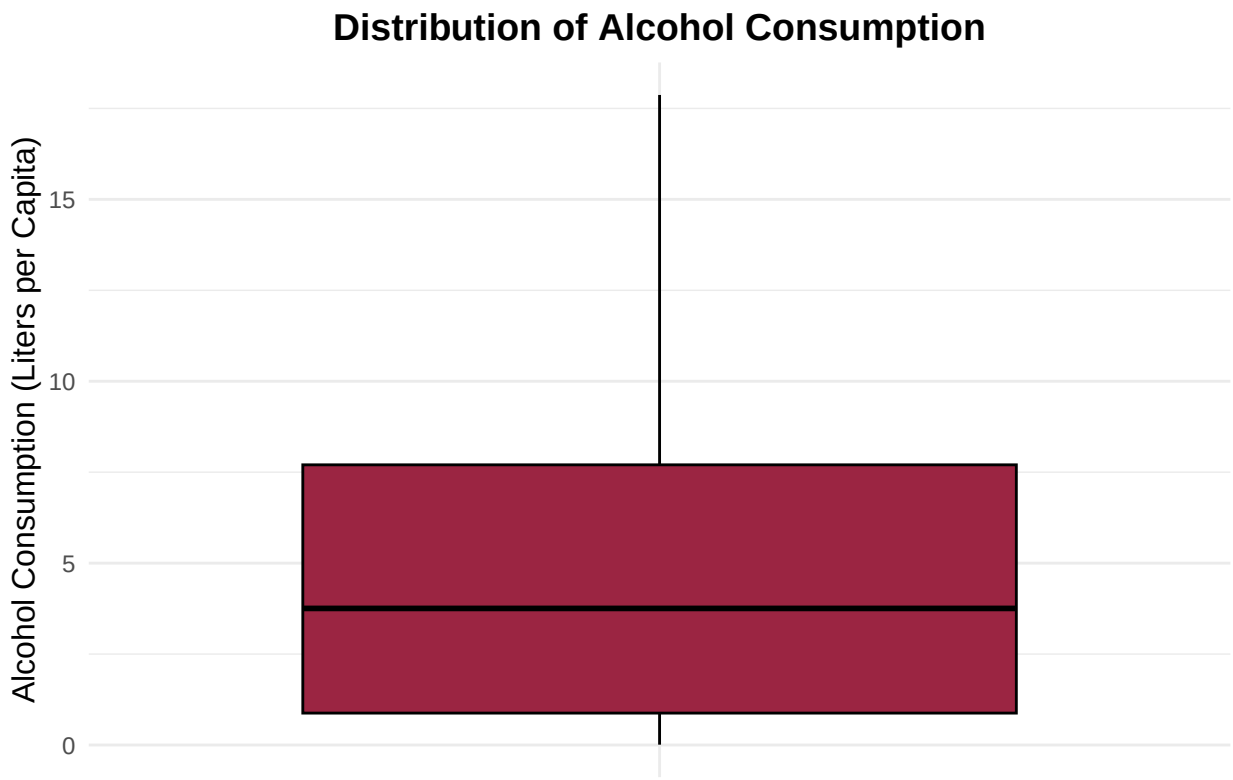
```r
thinness_5_9_years_cleaned <- dataset %>%
  filter(!is.na(thinness.5.9.years))

ggplot(thinness_5_9_years_cleaned, aes(x = "", y = thinness.5.9.years)) +
  geom_boxplot(fill = "#9b2542", color = "black") +
  labs(title = "Distribution of Thinness<br>
       Among Children (5-9 Years)",
       x = "", y = "Prevalence of Thinness (%)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```



The `Alcohol` variable represents the recorded per capita (age 15+) consumption of pure alcohol in liters, with a mean alcohol consumption of 4.6 liters and a median of 3.76 liters. The mean is higher than the median, suggesting a right-skewed distribution, where a few countries have significantly higher alcohol consumption, pulling the mean upwards. There are 194 missing values in this variable. The `Alcohol` variable allows for **examining the influence of alcohol consumption on life expectancy**.

```r
alcohol_cleaned <- dataset %>%
  filter(!is.na(Alcohol))

ggplot(alcohol_cleaned, aes(x = "", y = Alcohol)) +
```

```
  geom_boxplot(fill = "#9b2542", color = "black") +
  labs(title = "Distribution of Alcohol Consumption",
    x = "", y = "Alcohol Consumption (Liters per Capita)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```

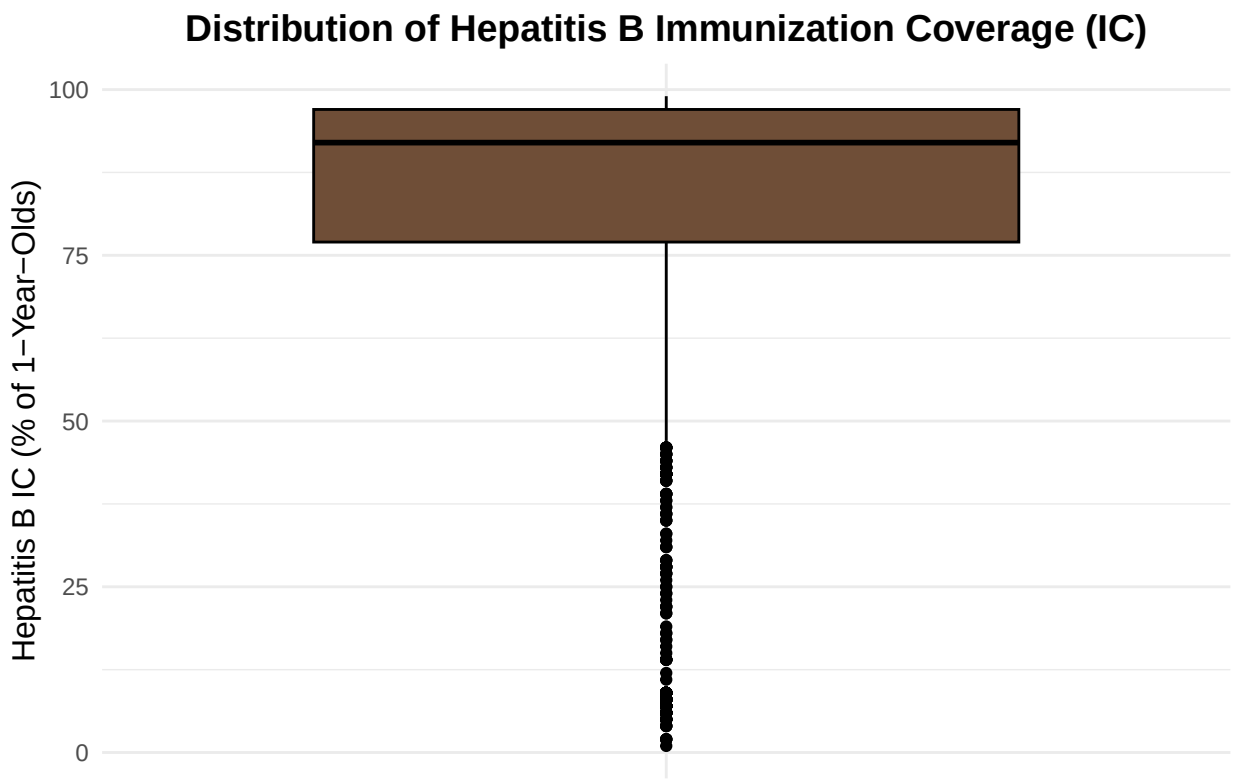


Vaccination and Disease Control Indicators

The `Hepatitis.B` variable shows the immunization coverage rate among 1-year-olds for Hepatitis B, expressed as a percentage. The mean coverage is 80.94%, with a median of 92%. The median is higher than the mean, indicating a left-skewed distribution. This suggests that while many countries have high immunization rates, there are some with significantly lower rates, which pull the mean down. There are 553 missing values in this variable. The `Hepatitis.B` variable allows for **analyzing the impact of Hepatitis B immunization on life expectancy**.

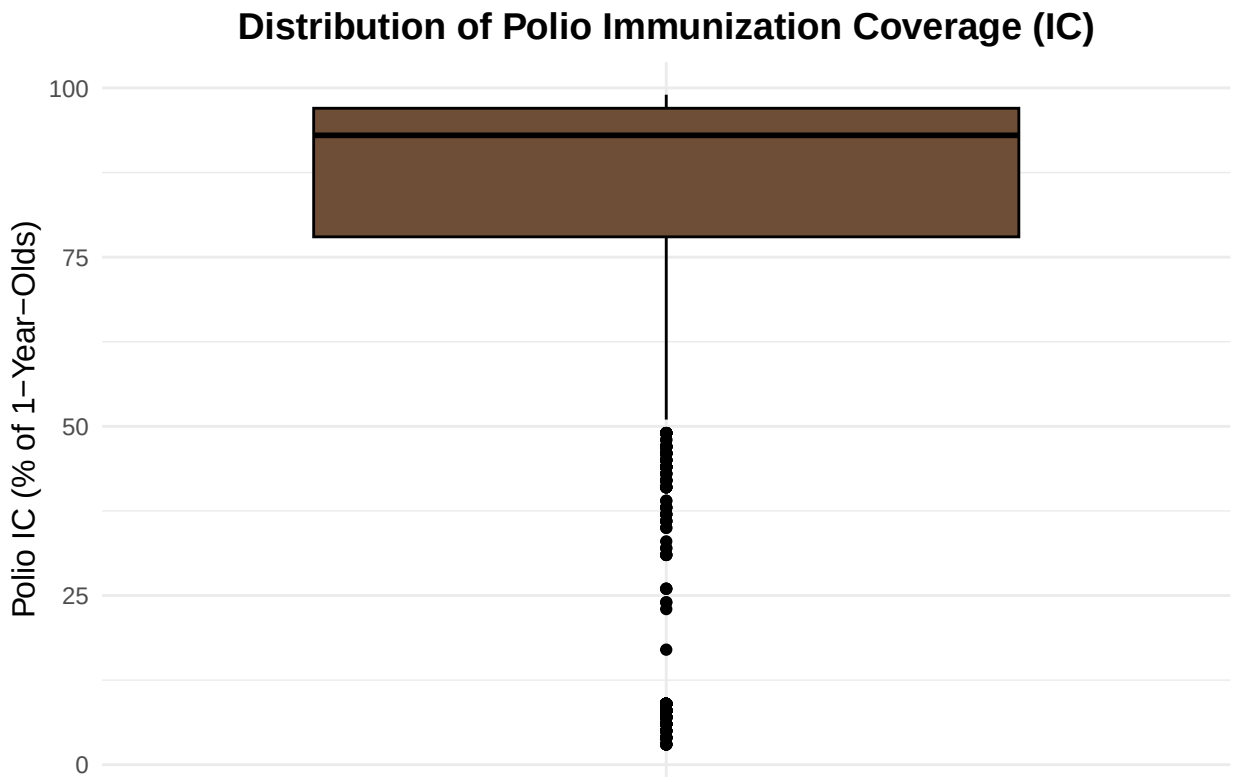
```
hepatitis_b_cleaned <- dataset %>%
  filter(!is.na(Hepatitis.B))

ggplot(hepatitis_b_cleaned, aes(x = "", y = Hepatitis.B)) +
  geom_boxplot(fill = "#6F4E37", color = "black") +
  labs(title = "Distribution of Hepatitis B Immunization Coverage (IC)",
```

```
      x = "", y = "Hepatitis B IC (% of 1-Year-Olds)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```



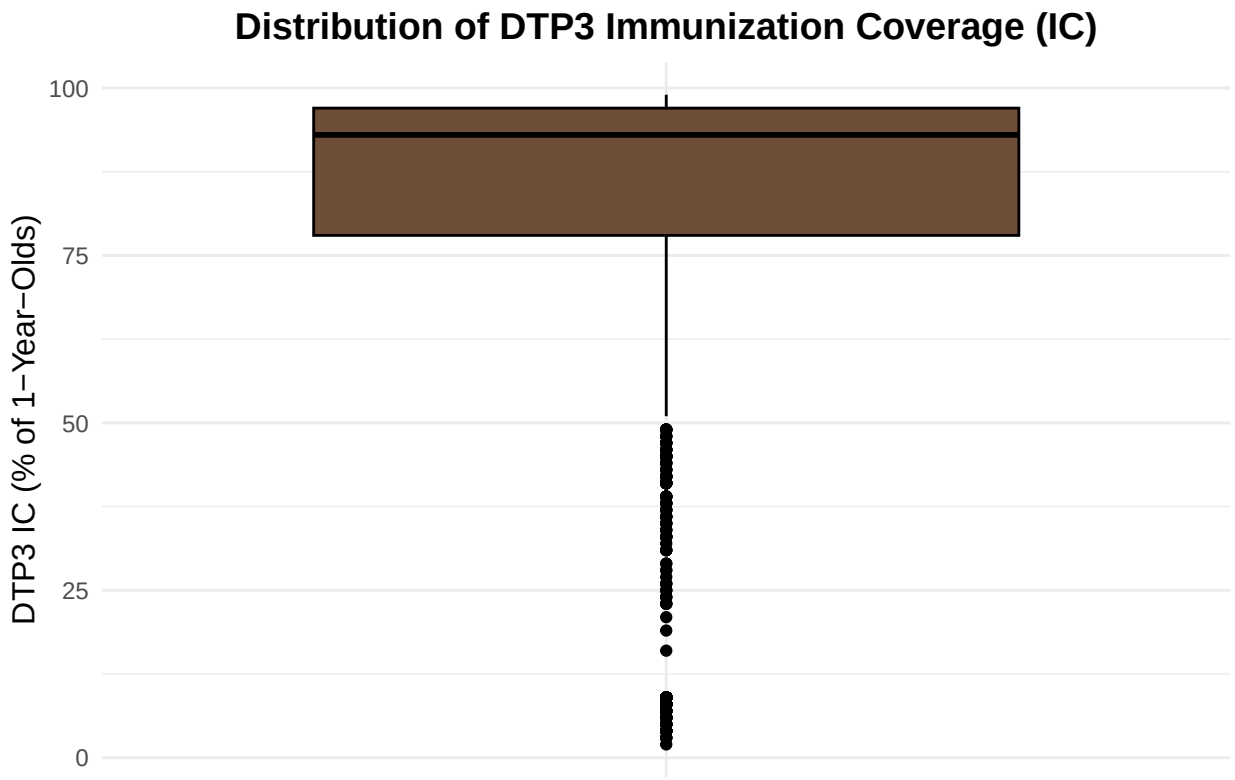
The `Polio` variable represents the immunization coverage rate for the Polio (Pol3) vaccine among 1-year-olds as a percentage. The mean coverage is 82.55%, with a median of 93%. The lower mean compared to the median suggests the presence of outliers with low immunization coverage rates. These outliers significantly affect the overall average. There are 19 missing values in this variable. The `Polio` variable allows for **exploring how Polio immunization rates impact life expectancy**.

```
polio_cleaned <- dataset %>%
  filter(!is.na(Polio))

ggplot(polio_cleaned, aes(x = "", y = Polio)) +
  geom_boxplot(fill = "#6F4E37", color = "black") +
  labs(title = "Distribution of Polio Immunization Coverage (IC)",
       x = "", y = "Polio IC (% of 1-Year-Olds)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```



The `Diphtheria` variable shows the immunization coverage rate for Diphtheria, Tetanus, and Pertussis (DTP3) among 1-year-olds, as a percentage. The mean coverage is 82.32%, with a median of 93%. The higher median compared to the mean indicates a negatively skewed distribution. This means most countries have high immunization rates, but a few have much lower rates. There are 19 missing values in this variable. The `Diphtheria` variable allows for **analyzing the impact of DTP3 immunization on life expectancy**.

```
diphtheria_cleaned <- dataset %>%
  filter(!is.na(Diphtheria))

ggplot(diphtheria_cleaned, aes(x = "", y = Diphtheria)) +
  geom_boxplot(fill = "#6F4E37", color = "black") +
  labs(title = "Distribution of DTP3 Immunization Coverage (IC)",
       x = "", y = "DTP3 IC (% of 1-Year-Olds)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```



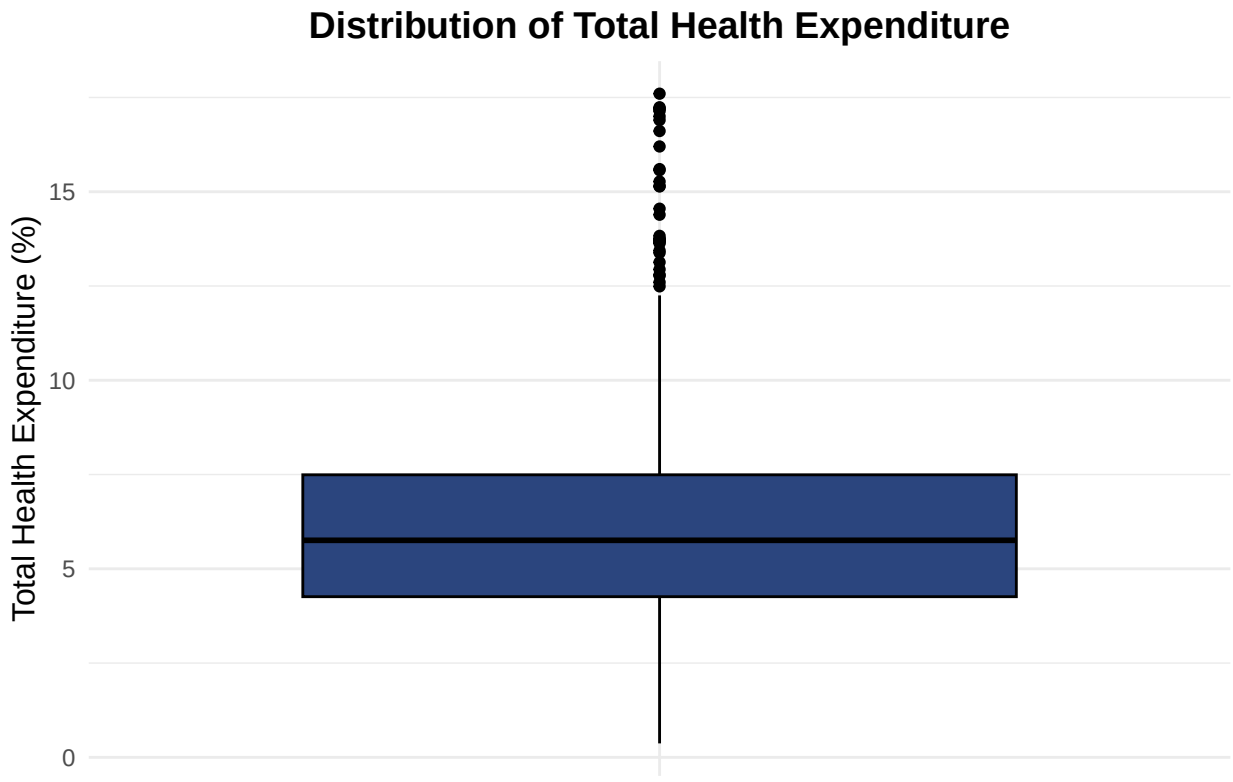
The `Measles` variable represents the number of reported measles cases per 1,000 population. The minimum number of cases reported is 0, with the first quartile (25th percentile) showing 0 cases, the median (50th percentile) reporting 17 cases, and the third quartile (75th percentile) showing 360.2 cases. There are no missing values in this variable. The `Measles` variable allows for **evaluating the impact of infectious disease outbreaks on life expectancy**.

Healthcare Expenditure and Resource Indicators

The `Total.expenditure` variable represents the percentage of government expenditure allocated to healthcare. The mean value is 5.94%, with a median of 5.76%. The mean and median are pretty close, suggesting a relatively symmetrical distribution. This indicates that the percentage of government expenditure allocated to healthcare is relatively consistent across different countries. There are 226 missing values in this variable. The `Total.expenditure` variable allows for **examining the role of government health spending in influencing life expectancy**.

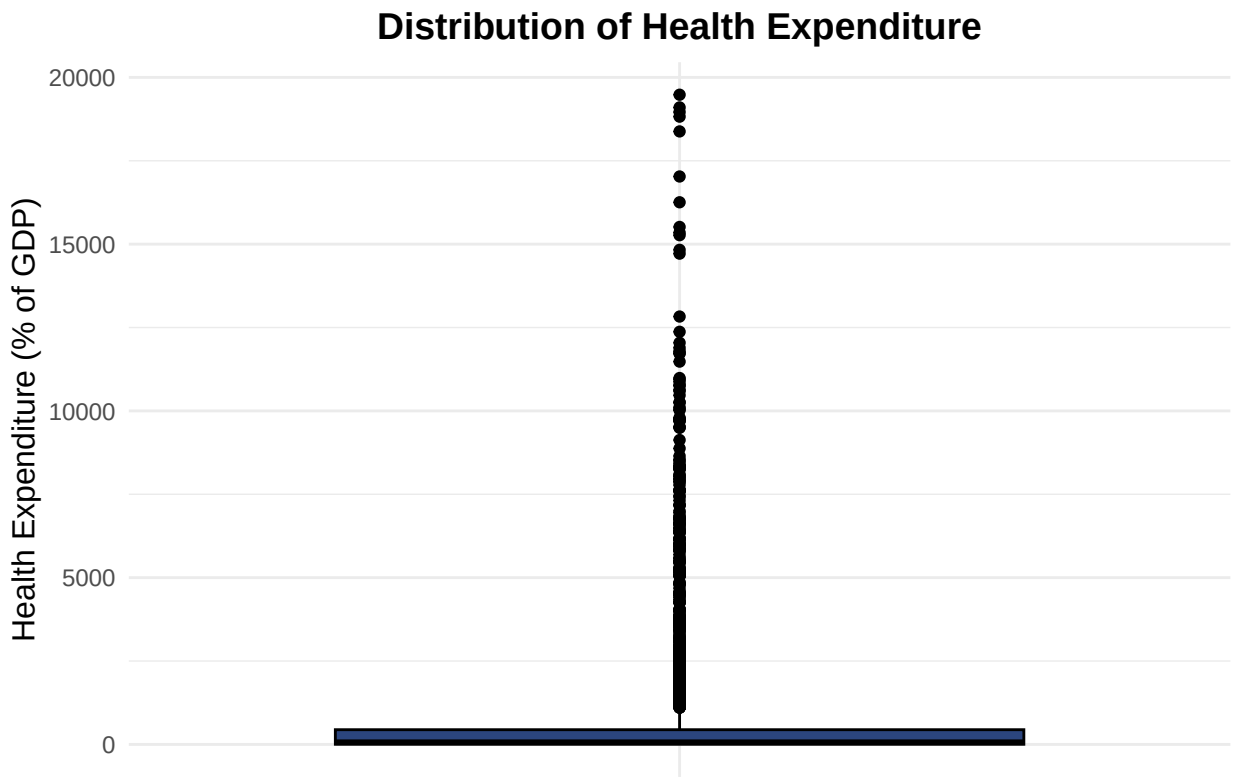
```
total_expenditure_cleaned <- dataset %>%
  filter(!is.na(Total.expenditure))

ggplot(total_expenditure_cleaned, aes(x = "", y = Total.expenditure)) +
  geom_boxplot(fill = "#2b457e", color = "black") +
  labs(title = "Distribution of Total Health Expenditure",
       x = "", y = "Total Health Expenditure (%)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
```

```
    axis.title.y = element_text(size = 12)
  )
```



The `percentage.expenditure` variable reflects the expenditure on health as a percentage of Gross Domestic Product (GDP) per capita, with a mean value of 738.25% and a median of 64.91%. The extremely high mean compared to the median indicates the presence of outliers. These outliers are countries with exceptionally high health expenditures as a percentage of GDP per capita. There are no missing values in this variable. The `percentage.expenditure` variable allows for **understanding the relationship between health expenditure and life expectancy**.

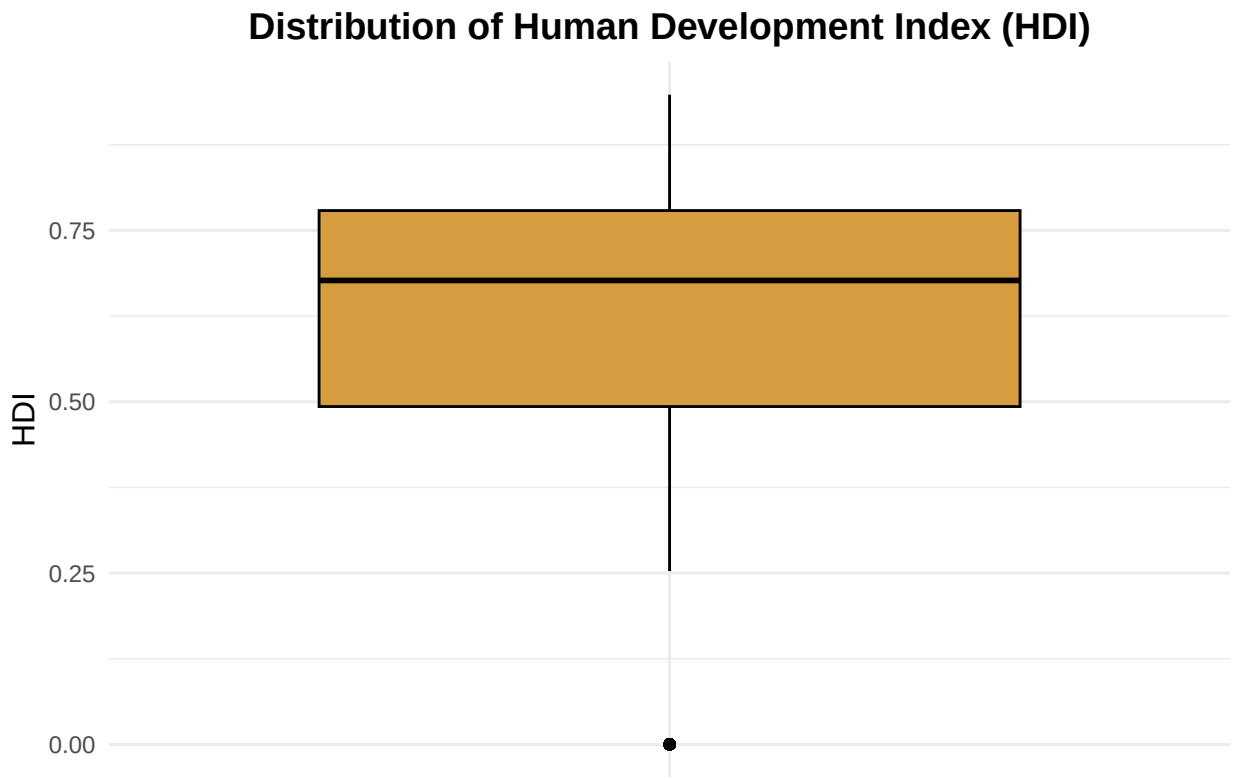
```
percentage_expenditure_cleaned <- dataset %>%
  filter(!is.na(percentage.expenditure))

ggplot(percentage_expenditure_cleaned,
       aes(x = "", y = percentage.expenditure)) +
  geom_boxplot(fill = "#2b457e", color = "black") +
  labs(title = "Distribution of Health Expenditure",
       x = "", y = "Health Expenditure (% of GDP)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.y = element_text(size = 12)
  )
```



Socio-Economic and Educational Indicators

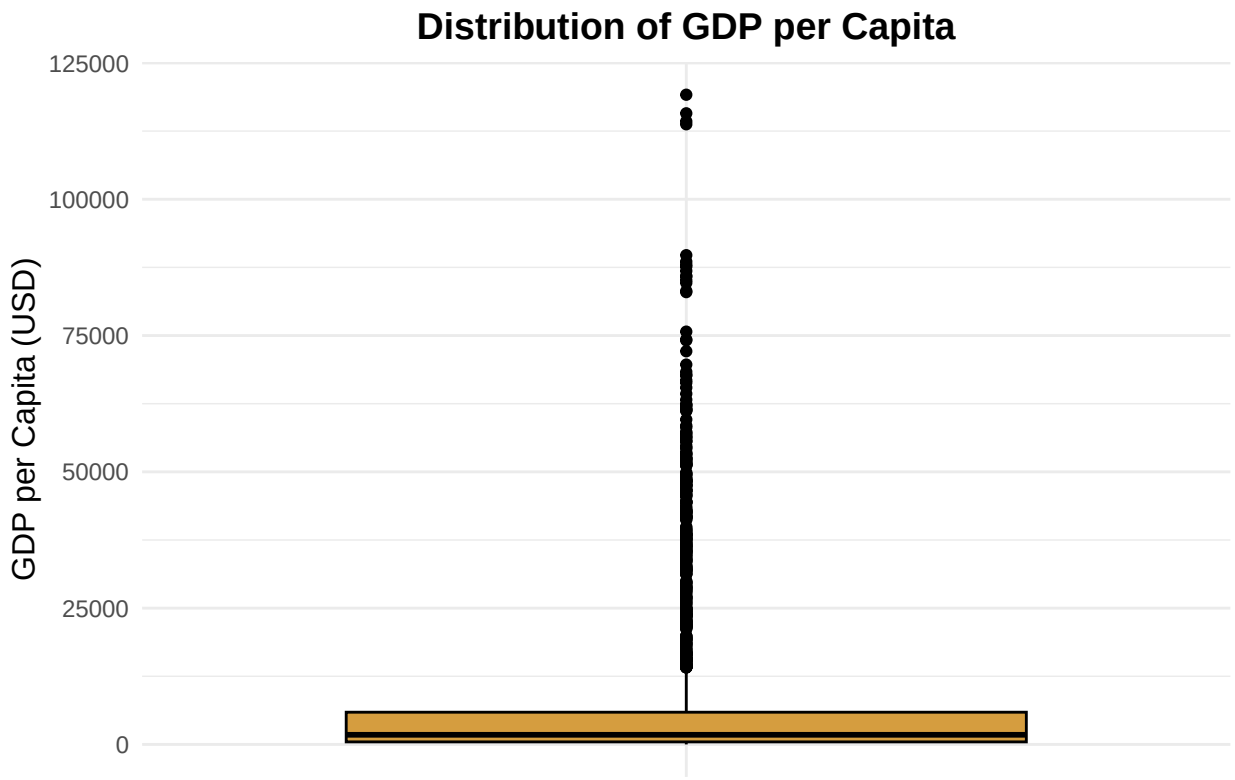
The `Income.composition.of.resources` variable represents the Human Development Index (HDI) in terms of income composition, with values ranging from 0 to 1. The mean value is 0.6276, and the median value is 0.677. The lower mean compared to the median suggests the presence of outliers with low HDI values. There are 167 missing values in this variable. The `Income.composition.of.resources` variable allows for **analyzing the relationship between HDI and life expectancy**.

```
income_composition_of_resources_cleaned <- dataset %>%
  filter(!is.na(Income.composition.of.resources))

ggplot(income_composition_of_resources_cleaned,
       aes(x = "", y = Income.composition.of.resources)) +
  geom_boxplot(fill = "#d59d3f", color = "black") +
  labs(title = "Distribution of Human Development Index (HDI)",
       x = "", y = "HDI") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```



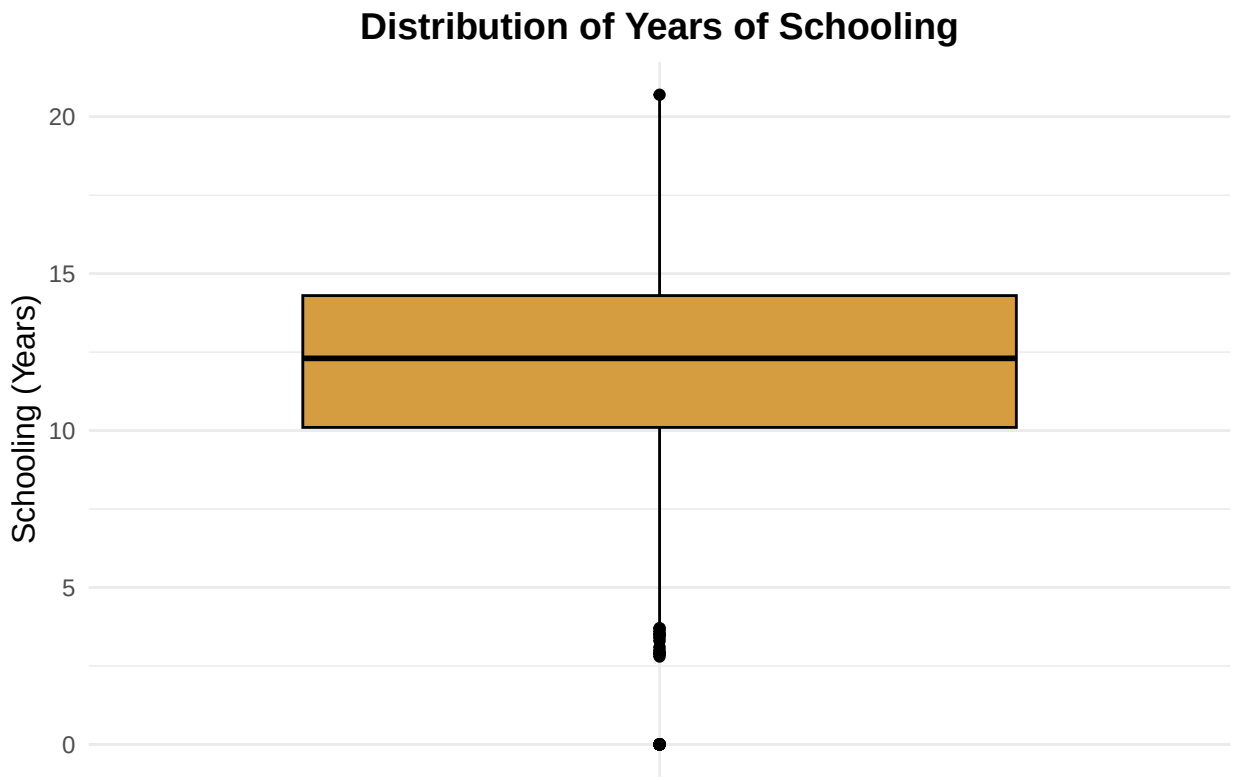
The `GDP` variable represents Gross Domestic Product per capita in USD. The average GDP per capita is $7,483.16, with a median of $1,766.95. The mean being significantly higher than the median suggests a right-skewed distribution. This indicates that while many countries have lower GDP per capita, a few countries with very high GDP per capita pull the mean upwards. There are 448 missing values in this variable. The `GDP` variable allows for **examining the relationship between economic prosperity and life expectancy**.

```
gdp_cleaned <- dataset %>%
  filter(!is.na(GDP))

ggplot(gdp_cleaned, aes(x = "", y = GDP)) +
  geom_boxplot(fill = "#d59d3f", color = "black") +
  labs(title = "Distribution of GDP per Capita",
       x = "", y = "GDP per Capita (USD)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```



The `Schooling` variable measures the average years of schooling in a country. The mean value is 11.99 years, and the median is 12.3 years. The close values of the mean and median suggest that the data is relatively concentrated around the central values, with moderate variability. There are 163 missing values in this variable. This variable allows for **exploring the impact of education on life expectancy**.

```
schooling_cleaned <- dataset %>%
  filter(!is.na(Schooling))

ggplot(schooling_cleaned, aes(x = "", y = Schooling)) +
  geom_boxplot(fill = "#d59d3f", color = "black") +
  labs(title = "Distribution of Years of Schooling",
       x = "", y = "Schooling (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.y = element_text(size = 12)
  )
```



The `Population` variable represents the total population of each country. The mean population is approximately 12.75 million, with a median of around 1.39 million. The much higher mean compared to the median indicates a positively skewed (right-skewed) distribution. This means most regions or countries have smaller populations, but a few have much larger populations. There are 652 missing values in this variable. The `Population` variable allows for **exploring how population size might relate to life expectancy**.

The `Status` variable categorizes countries as "Developed" or "Developing." The dataset has 32 developed, and 161 developing countries and no missing values are in this variable. The `Status` variable allows for **comparing life expectancy between developed and developing countries**.

```
num_developed_countries <- length(unique(dataset$Country[dataset$Status ==
                                                           "Developed"]))
num_developed_countries
```

```
## [1] 32
```

```
num_developing_countries <- length(unique(dataset$Country[dataset$Status ==
                                                            "Developing"]))
num_developing_countries
```

```
## [1] 161
```

Multivariate Analysis

```r
average_life_exp <- dataset %>%
  select(Country, Life.expectancy) %>%
  filter(!is.na(Life.expectancy)) %>%
  group_by(Country) %>%
  summarise(avg_life_exp = mean(Life.expectancy))
head(average_life_exp)
```

```
## # A tibble: 6 x 2
##   Country             avg_life_exp
##   <chr>                      <dbl>
## 1 Afghanistan                 58.2
## 2 Albania                     75.2
## 3 Algeria                     73.6
## 4 Angola                      49.0
## 5 Antigua and Barbuda         75.1
## 6 Argentina                   75.2
```

```r
mean(as.numeric(average_life_exp$avg_life_exp), na.rm = TRUE)
```

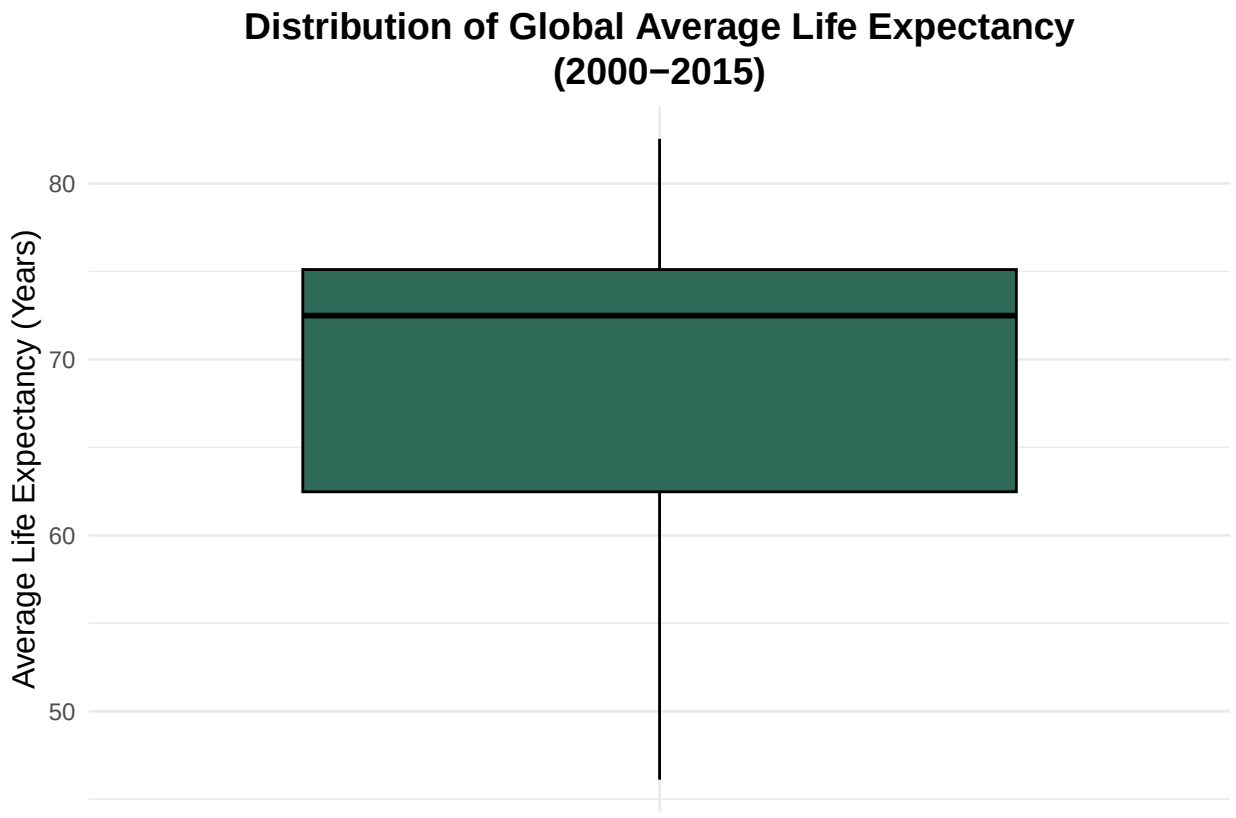
```
## [1] 69.22493
```

```r
median(as.numeric(average_life_exp$avg_life_exp), na.rm = TRUE)
```

```
## [1] 72.4875
```

Since the mean is less than the median, the distribution is likely left-skewed. This suggests that some lower life expectancy values are pulling the mean down. The left skewness implies the presence of outliers or a longer tail on the lower end of the life expectancy spectrum. These could be countries with significantly lower life expectancies due to various factors such as health crises, conflicts, or economic challenges.

```r
ggplot(average_life_exp, aes(x = "", y = avg_life_exp)) +
  geom_boxplot(fill = "#2e6a57", color = "black") +
  labs(title = "Distribution of Global Average Life Expectancy<br>
       (2000-2015)", x = "", y = "Average Life Expectancy (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.x = element_blank(),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```



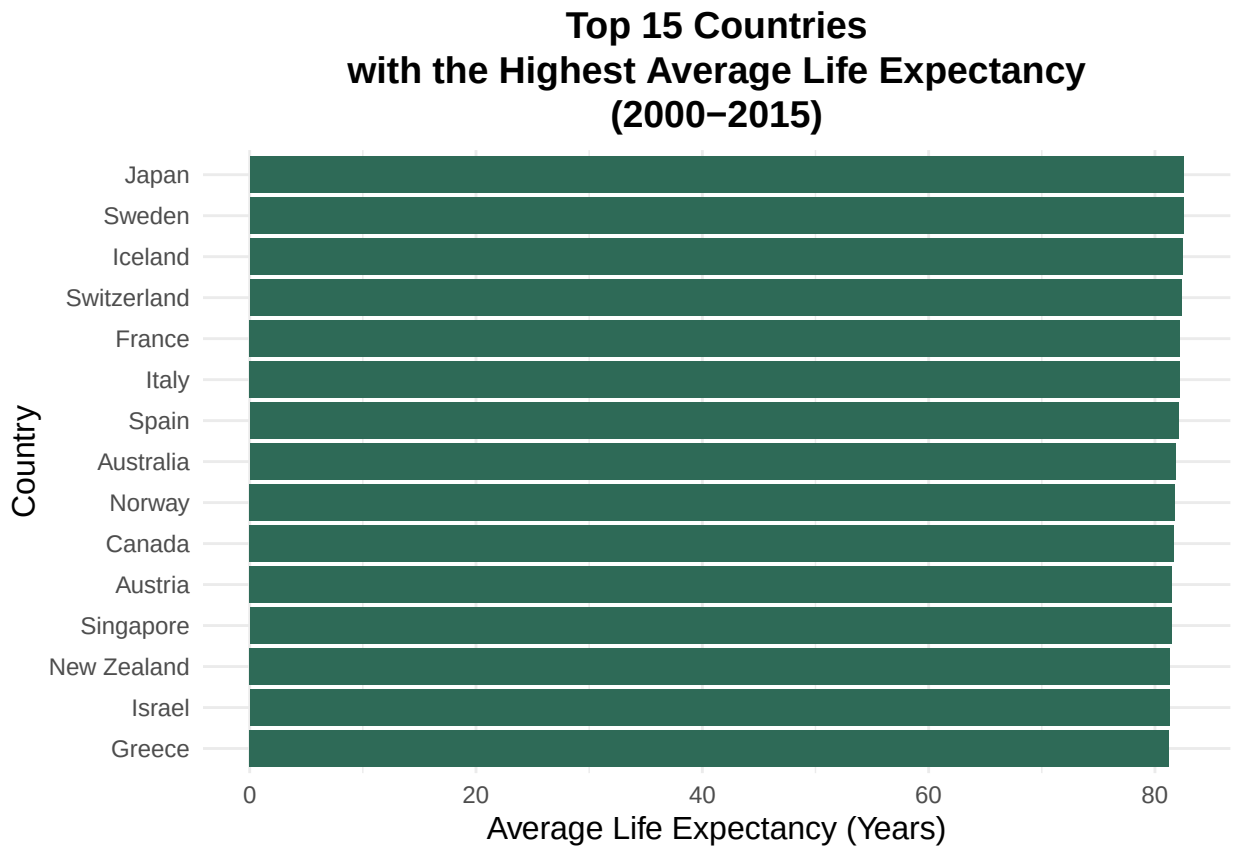
The top 15 countries with the highest average life expectancy are Japan, Sweden, Iceland, Switzerland, France, Italy, Spain, Australia, Norway, Canada, Austria, Singapore, New Zealand, Israel, and Greece.

```
average_life_exp_desc <- dataset %>%
  select(Country, Life.expectancy) %>%
  filter(!is.na(Life.expectancy)) %>%
  group_by(Country) %>%
  summarise(average_life_expectancy = mean(Life.expectancy)) %>%
  arrange(desc(average_life_expectancy)) %>%
  slice(1:15)
head(average_life_exp_desc)
```

```
## # A tibble: 6 x 2
##   Country     average_life_expectancy
##   <chr>                         <dbl>
## 1 Japan                          82.5
## 2 Sweden                         82.5
## 3 Iceland                        82.4
## 4 Switzerland                    82.3
## 5 France                         82.2
## 6 Italy                          82.2
```

```
ggplot(average_life_exp_desc,
       aes(x = reorder(Country, average_life_expectancy),
           y = average_life_expectancy)) +
```

```
  geom_bar(stat = "identity", fill = "#2e6a57") +
  coord_flip() +
  labs(title = "Top 15 Countries<br>
       with the Highest Average Life Expectancy<br>
       (2000-2015)", x = "Country", y = "Average Life Expectancy (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```

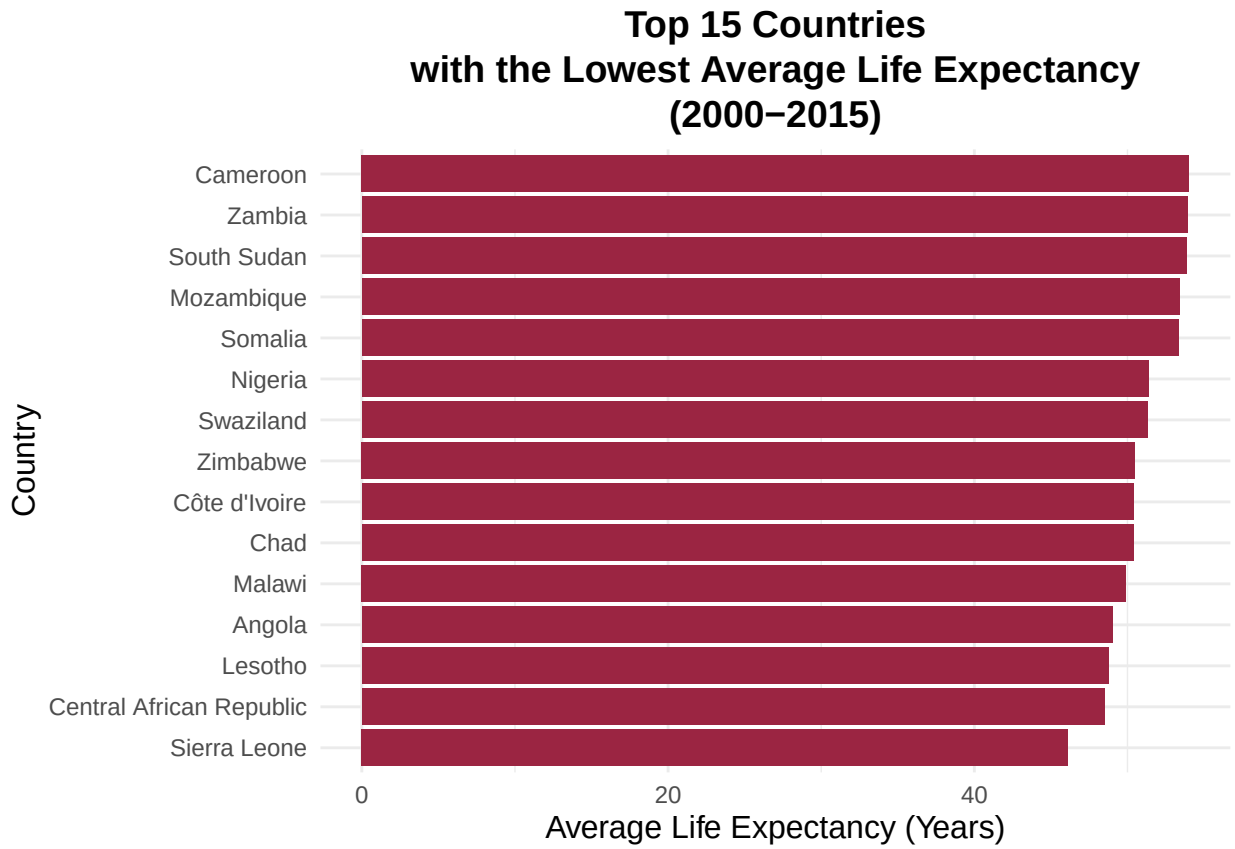


The top 15 countries with the lowest average life expectancy are Cameroon, Zambia, South Sudan, Mozambique, Somalia, Nigeria, Swaziland, Zimbabwe, Côte d'Ivoire, Chad, Malawi, Angola, Lesotho, Central African Republic, and Sierra Leone.

```
average_life_exp_asc <- dataset %>%
  select(Country, Life.expectancy) %>%
  filter(!is.na(Life.expectancy)) %>%
  group_by(Country) %>%
  summarise(average_life_expectancy = mean(Life.expectancy)) %>%
  arrange(average_life_expectancy) %>%
  slice(1:15)
head(average_life_exp_asc)
```

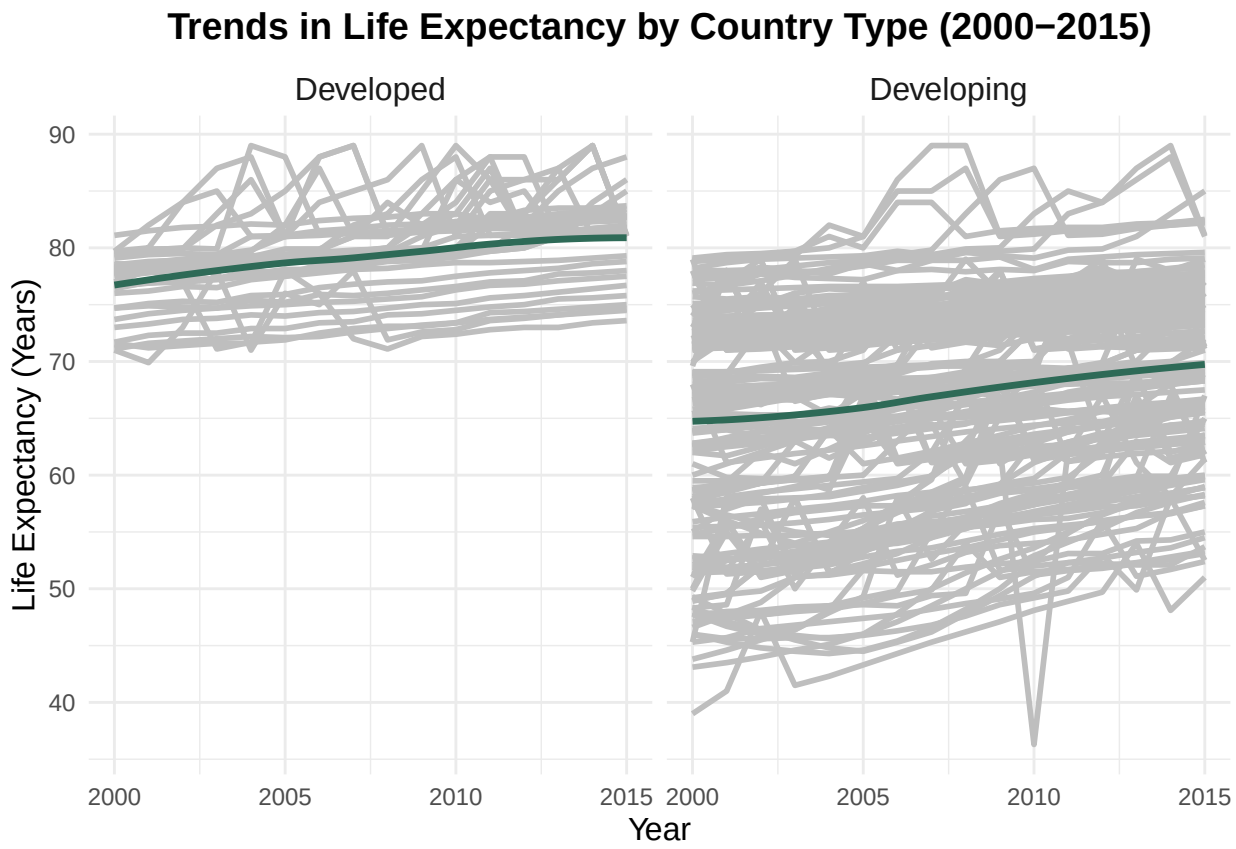
```
## # A tibble: 6 x 2
##   Country                   average_life_expectancy
##   <chr>                                       <dbl>
## 1 Sierra Leone                                 46.1
## 2 Central African Republic                     48.5
## 3 Lesotho                                      48.8
## 4 Angola                                       49.0
## 5 Malawi                                       49.9
## 6 Chad                                         50.4
```

```
ggplot(average_life_exp_asc,
       aes(x = reorder(Country, average_life_expectancy),
           y = average_life_expectancy)) +
  geom_bar(stat = "identity", fill = "#9b2542") +
  coord_flip() +
  labs(title = "Top 15 Countries<br>
       with the Lowest Average Life Expectancy<br>
       (2000-2015)", x = "Country", y = "Average Life Expectancy (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```

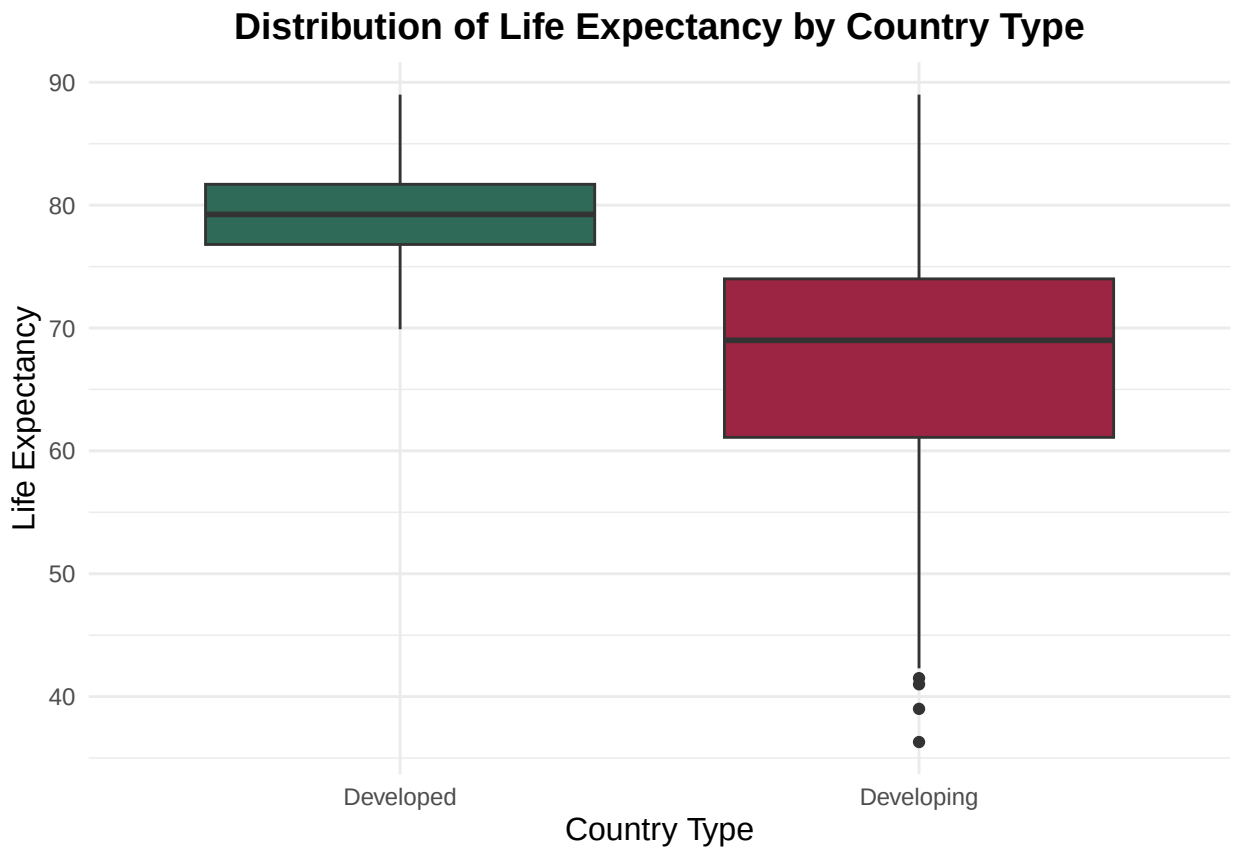


There has been a slight increase in life expectancy for both developed and developing countries in general.

```
life_expectancy_over_time_by_country_type <- dataset %>%
  select(Life.expectancy, Country, Year, Status) %>%
  filter(!is.na(Life.expectancy)) %>%
  filter(Status %in% c("Developed", "Developing"))

ggplot(life_expectancy_over_time_by_country_type,
       aes(x = Year, y = Life.expectancy, group = Country)) +
  geom_line(color = "grey", size = 1) +
  geom_smooth(aes(group = 1), method = "loess", se = FALSE,
              color = "#2e6a57", linewidth = 1.2) +
  labs(title = "Trends in Life Expectancy by Country Type (2000-2015)",
       x = "Year", y = "Life Expectancy (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    strip.text = element_text(size = 12),
    legend.position = "none"
  ) +
  facet_wrap(~ Status)
```




```
life_exp_comp_dev_dev <- dataset %>%
  select(Life.expectancy, Status) %>%
  filter(!is.na(Life.expectancy))

ggplot(life_exp_comp_dev_dev,
       aes(x = Status, y = Life.expectancy, fill = Status)) +
  geom_boxplot() +
  scale_fill_manual(values = c("Developed" = "#2e6a57",
                               "Developing" = "#9b2542")) +
  labs(title = "Distribution of Life Expectancy by Country Type",
       x = "Country Type", y = "Life Expectancy") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    legend.position = "none"
  )
```



The boxplot compares the life expectancy between **developed** and **developing** countries. The median life expectancy for developed countries is higher than for developing countries, as indicated by the higher median line within the box. The interquartile range (IQR) for developed countries is narrower, suggesting less variability in life expectancy compared to developing countries. The whiskers extend further for developing countries, indicating a wider range of life expectancy values. Additionally, there are more outliers in the developing countries' data, suggesting the presence of extreme values.

Health Risk Factors and Mortality Indicators

To visualize the impact of the **Health Risk Factors and Mortality Indicators** on life expectancy,

```
health_risk_factors_mortality_indicators_part_1 <- dataset %>%
  select(Life.expectancy, Adult.Mortality, under.five.deaths,
         infant.deaths, HIV.AIDS)
head(health_risk_factors_mortality_indicators_part_1)
```

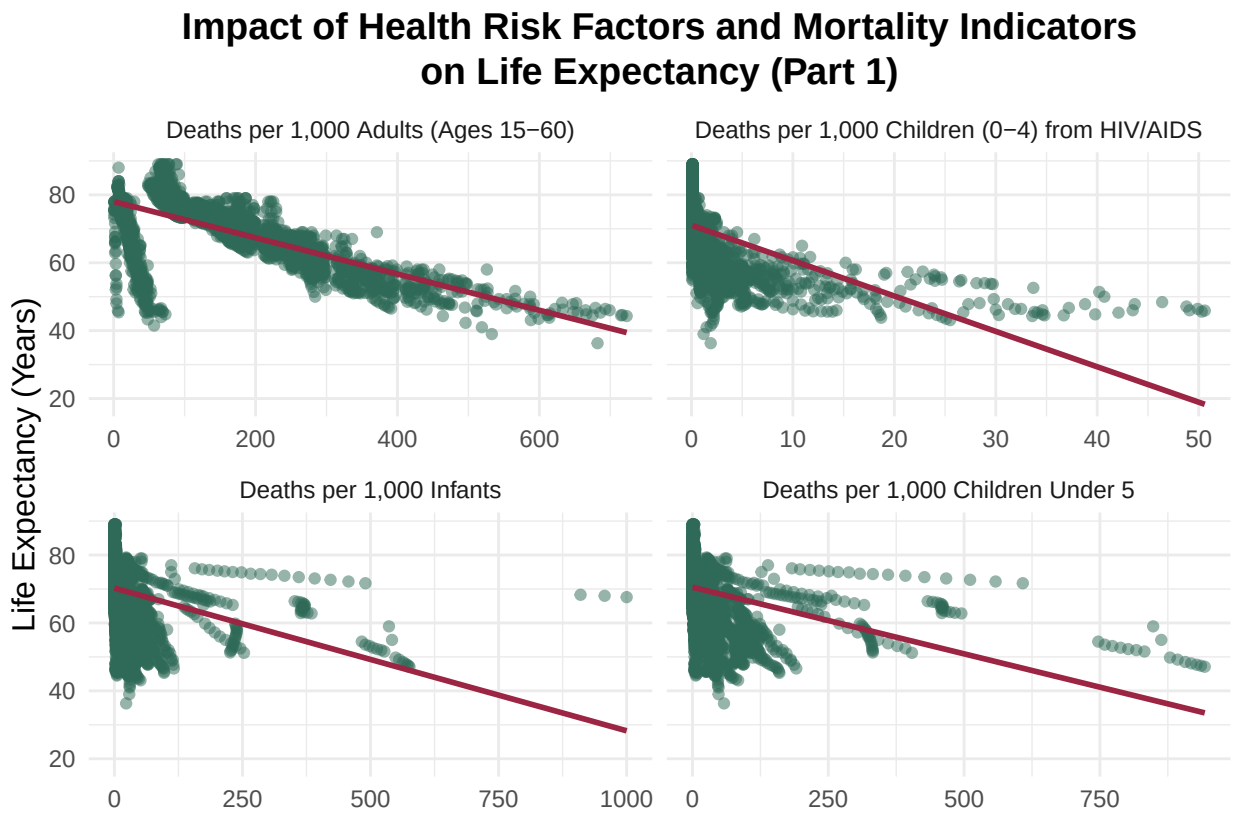
```
##   Life.expectancy Adult.Mortality under.five.deaths infant.deaths HIV.AIDS
## 1            65.0             263                83            62      0.1
## 2            59.9             271                86            64      0.1
## 3            59.9             268                89            66      0.1
## 4            59.5             272                93            69      0.1
## 5            59.2             275                97            71      0.1
## 6            58.8             279               102            74      0.1
```

```r
health_risk_factors_mortality_indicators_part_1_long_format <-
  health_risk_factors_mortality_indicators_part_1 %>%
  pivot_longer(cols = -Life.expectancy,
               names_to = "health_risk_factors_1",
               values_to = "values") %>%
  filter(!(health_risk_factors_1 == "infant.deaths"
           & values > 1000),
         !(health_risk_factors_1 == "under.five.deaths"
           & values > 1000),
         !(health_risk_factors_1 == "HIV.AIDS"
           & values > 1000),
         !is.na(Life.expectancy),
         !(health_risk_factors_1 == "Adult.Mortality"
           & is.na(values)),
         !(health_risk_factors_1 == "under.five.deaths"
           & is.na(values)),
         !(health_risk_factors_1 == "infant.deaths"
           & is.na(values)),
         !(health_risk_factors_1 == "HIV.AIDS"
           & is.na(values)))
head(health_risk_factors_mortality_indicators_part_1_long_format)
```

```
## # A tibble: 6 x 3
##   Life.expectancy health_risk_factors_1 values
##             <dbl> <chr>                 <dbl>
## 1            65   Adult.Mortality       263
## 2            65   under.five.deaths      83
## 3            65   infant.deaths          62
## 4            65   HIV.AIDS                0.1
## 5            59.9 Adult.Mortality       271
## 6            59.9 under.five.deaths      86
```

```r
custom_labels_1 <- c(
  "Adult.Mortality" = "Deaths per 1,000 Adults (Ages 15-60)",
  "infant.deaths" = "Deaths per 1,000 Infants",
  "under.five.deaths" = "Deaths per 1,000 Children Under 5",
  "HIV.AIDS" = "Deaths per 1,000 Children (0-4) from HIV/AIDS"
)

ggplot(health_risk_factors_mortality_indicators_part_1_long_format,
       aes(x = values, y = Life.expectancy)) +
  geom_point(alpha = 0.5, color = "#2e6a57") +
  geom_smooth(method = "lm", se = FALSE, color = "#9b2542") +
  facet_wrap(~ health_risk_factors_1, scales = "free_x",
             labeller = labeller(health_risk_factors_1 = custom_labels_1)) +
  labs(title = "Impact of Health Risk Factors and Mortality Indicators<br>
       on Life Expectancy (Part 1)", x = "", y = "Life Expectancy (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.y = element_text(size = 12)
  ) +
```

```
  theme(plot.title = element_markdown())
```




```
health_risk_factors_mortality_indicators_part_2 <- dataset %>%
  select(Life.expectancy, BMI, Alcohol,
         thinness..1.19.years, thinness.5.9.years)
head(health_risk_factors_mortality_indicators_part_2)
```

```
##   Life.expectancy  BMI Alcohol thinness..1.19.years thinness.5.9.years
## 1            65.0 19.1    0.01                 17.2                17.3
## 2            59.9 18.6    0.01                 17.5                17.5
## 3            59.9 18.1    0.01                 17.7                17.7
## 4            59.5 17.6    0.01                 17.9                18.0
## 5            59.2 17.2    0.01                 18.2                18.2
## 6            58.8 16.7    0.01                 18.4                18.4
```

```
health_risk_factors_mortality_indicators_part_2_long_format <-
  health_risk_factors_mortality_indicators_part_2 %>%
  pivot_longer(cols = -Life.expectancy,
               names_to = "health_risk_factors_2",
               values_to = "values") %>%
  filter(!is.na(Life.expectancy),
         !(health_risk_factors_2 == "BMI"
           & is.na(values)),
         !(health_risk_factors_2 == "thinness..1.19.years"
```

```
          & is.na(values)),
        !(health_risk_factors_2 == "thinness.5.9.years"
          & is.na(values)),
        !(health_risk_factors_2 == "Alcohol"
          & is.na(values)))
head(health_risk_factors_mortality_indicators_part_2_long_format)
```

```
## # A tibble: 6 x 3
##   Life.expectancy health_risk_factors_2 values
##             <dbl> <chr>                  <dbl>
## 1            65   BMI                    19.1
## 2            65   Alcohol                 0.01
## 3            65   thinness..1.19.years   17.2
## 4            65   thinness.5.9.years     17.3
## 5            59.9 BMI                    18.6
## 6            59.9 Alcohol                 0.01
```

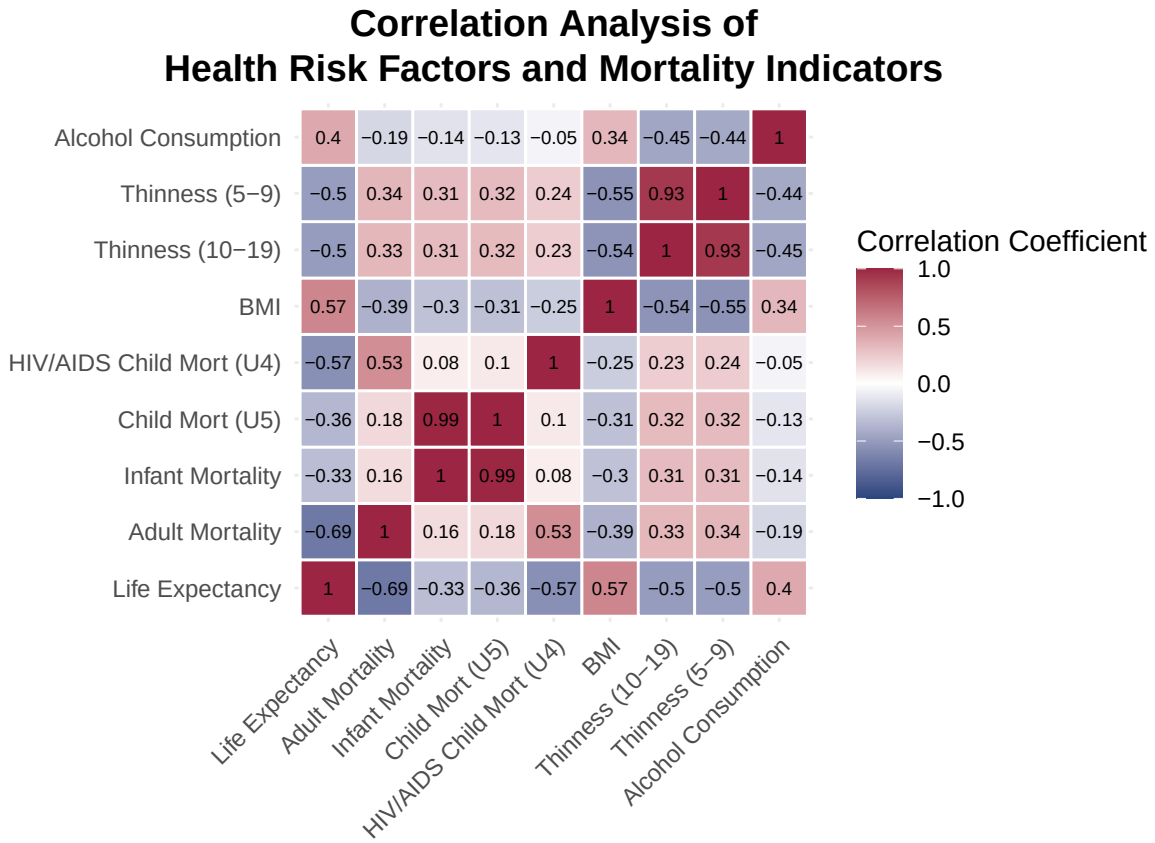
```
custom_labels_2 <- c(
  "BMI" = "Body Mass Index (BMI)",
  "thinness..1.19.years" = "Thinness (Ages 1-19)",
  "thinness.5.9.years" = "Thinness (Ages 5-9)",
  "Alcohol" = "Alcohol Consumption"
)

ggplot(health_risk_factors_mortality_indicators_part_2_long_format,
       aes(x = values, y = Life.expectancy)) +
  geom_point(alpha = 0.5, color = "#2e6a57") +
  geom_smooth(method = "lm", se = FALSE, color = "#9b2542") +
  facet_wrap(~ health_risk_factors_2, scales = "free_x",
             labeller = labeller(health_risk_factors_2 = custom_labels_2)) +
  labs(title = "Impact of Health Risk Factors and Mortality Indicators<br>
       on Life Expectancy (Part 2)", x = "", y = "Life Expectancy (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```


```
health_risk_factors_mortality_ind <- dataset %>%
  select(
    `Life Expectancy` = Life.expectancy,
    `Adult Mortality` = Adult.Mortality,
    `Infant Mortality` = infant.deaths,
    `Child Mort (U5)` = under.five.deaths,
    `HIV/AIDS Child Mort (U4)` = HIV.AIDS,
    `BMI` = BMI,
    `Thinness (10-19)` = thinness..1.19.years,
    `Thinness (5-9)` = thinness.5.9.years,
    `Alcohol Consumption` = Alcohol
  ) %>%
  filter(`Infant Mortality` <= 1000,
         `Child Mort (U5)` <= 1000,
         `HIV/AIDS Child Mort (U4)` <= 1000)
health_risk_factors_mortality_ind_cor <-
  cor(health_risk_factors_mortality_ind, use = "complete.obs")
health_risk_factors_mortality_ind_cor_long_format <-
  melt(health_risk_factors_mortality_ind_cor)
head(health_risk_factors_mortality_ind_cor_long_format)
```

```
##                       Var1            Var2      value
## 1          Life Expectancy Life Expectancy  1.0000000
## 2          Adult Mortality Life Expectancy -0.6920548
## 3         Infant Mortality Life Expectancy -0.3287908
```

```
## 4          Child Mort (U5) Life Expectancy -0.3589880
## 5 HIV/AIDS Child Mort (U4) Life Expectancy -0.5650514
## 6                      BMI Life Expectancy  0.5699343
```

```r
ggplot(health_risk_factors_mortality_ind_cor_long_format,
       aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white", size = 0.5) +
  geom_text(aes(label = round(value, 2)), color = "black", size = 2.5) +
  scale_fill_gradient2(low = "#2b457e", high = "#9b2542", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation Coefficient") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    axis.title.x = element_blank(),
    axis.title.y = element_blank()
    ) +
  coord_fixed() +
  labs(title = "Correlation Analysis of<br>
       Health Risk Factors and Mortality Indicators") +
  theme(plot.title = element_markdown())
```



The correlation between **Adult Mortality** and **Life Expectancy** is **-0.69**. This indicates a strong negative relationship, meaning higher adult mortality is associated with lower life expectancy.

The correlation between **Infant Mortality** and **Life Expectancy** is **-0.33**. This shows a moderate negative relationship, suggesting that higher infant mortality is associated with lower life expectancy.

The correlation between **Child Mortality (Under 5)** and **Life Expectancy** is **-0.36**. This indicates a moderate negative relationship, meaning higher child mortality is associated with lower life expectancy.

The correlation between **HIV/AIDS Child Mortality (Under 4)** and **Life Expectancy** is **-0.57**. This shows a strong negative relationship, suggesting that higher HIV/AIDS child mortality is associated with lower life expectancy.

The correlation between **BMI** and **Life Expectancy** is **0.57**. This indicates a strong positive relationship, meaning higher BMI is associated with higher life expectancy.

The correlation between **Thinness (Ages 1-19)** and **Life Expectancy** is **-0.5**. This shows a strong negative relationship, suggesting that higher thinness in the 10-19 age group is associated with lower life expectancy.

The correlation between **Thinness (Ages 5-9)** and **Life Expectancy** is **-0.5**. This indicates a strong negative relationship, meaning higher thinness in the 5-9 age group is associated with lower life expectancy.

The correlation between **Alcohol Consumption** and **Life Expectancy** is **0.4**. This shows a moderate positive relationship, suggesting that higher alcohol consumption is associated with higher life expectancy.

In summary, variables like adult mortality, infant mortality, child mortality, HIV/AIDS child mortality, and thinness have negative correlations with life expectancy, indicating that higher values in these variables are associated with lower life expectancy. On the other hand, BMI and alcohol consumption have positive correlations with life expectancy, suggesting that higher values in these variables are associated with higher life expectancy (just because two variables are correlated doesn't mean that one causes the other, correlation simply indicates that there is a relationship or pattern between the two variables).

Vaccination and Disease Control Indicators

To visualize the impact of the **Vaccination and Disease Control Indicators** on life expectancy,

```
vaccination_disease_control_indicators <- dataset %>%
  select(Life.expectancy, Hepatitis.B, Polio,
         Diphtheria, Measles)
head(vaccination_disease_control_indicators)
```

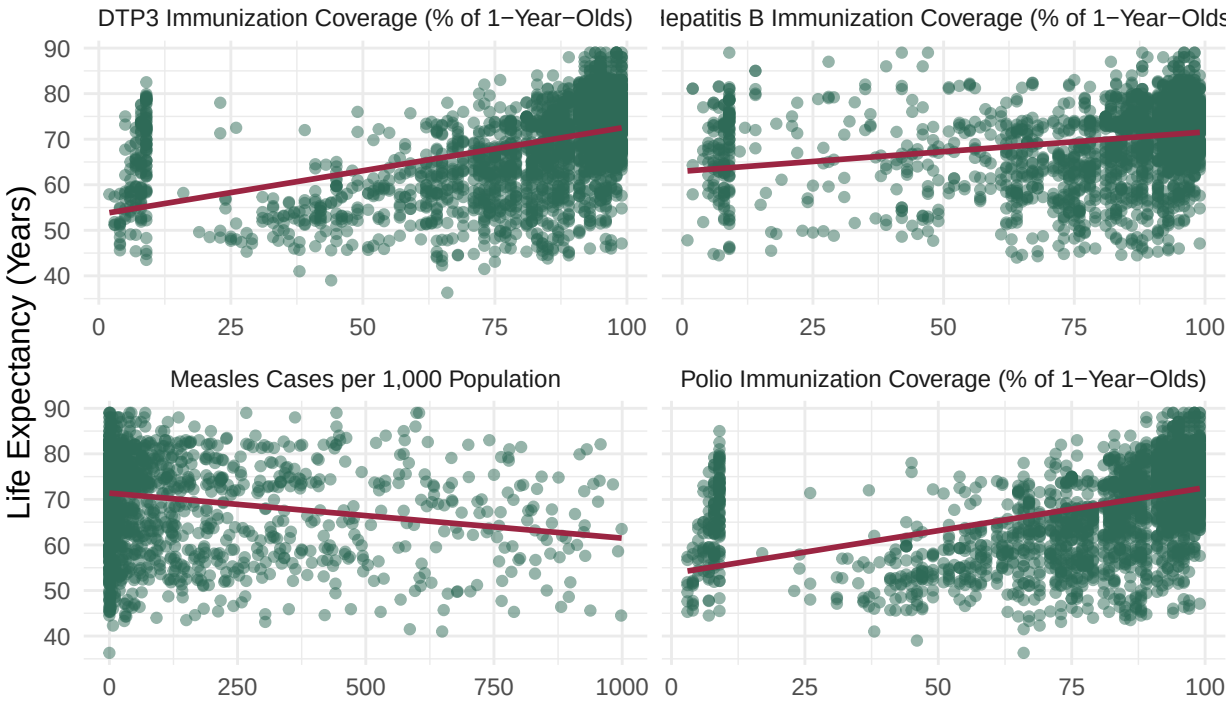
```
##   Life.expectancy Hepatitis.B Polio Diphtheria Measles
## 1            65.0          65     6         65    1154
## 2            59.9          62    58         62     492
## 3            59.9          64    62         64     430
## 4            59.5          67    67         67    2787
## 5            59.2          68    68         68    3013
## 6            58.8          66    66         66    1989
```

```
vaccination_disease_control_indicators_long_format <-
  vaccination_disease_control_indicators %>%
  pivot_longer(cols = -Life.expectancy,
               names_to = "vaccination_disease_indicators",
               values_to = "values") %>%
  filter(!(vaccination_disease_indicators == "Measles"
           & values > 1000),
         !is.na(Life.expectancy),
         !(vaccination_disease_indicators == "Hepatitis.B"
```

```
          & is.na(values)),
        !(vaccination_disease_indicators == "Polio"
          & is.na(values)),
        !(vaccination_disease_indicators == "Diphtheria"
          & is.na(values)),
        !(vaccination_disease_indicators == "Measles"
          & is.na(values)))
head(vaccination_disease_control_indicators_long_format)
```

```
## # A tibble: 6 x 3
##   Life.expectancy vaccination_disease_indicators values
##             <dbl> <chr>                           <int>
## 1            65   Hepatitis.B                        65
## 2            65   Polio                               6
## 3            65   Diphtheria                         65
## 4            59.9 Hepatitis.B                        62
## 5            59.9 Polio                              58
## 6            59.9 Diphtheria                         62
```

```
custom_labels_3 <- c(
  "Hepatitis.B" = "Hepatitis B Immunization Coverage (% of 1-Year-Olds)",
  "Polio" = "Polio Immunization Coverage (% of 1-Year-Olds)",
  "Diphtheria" = "DTP3 Immunization Coverage (% of 1-Year-Olds)",
  "Measles" = "Measles Cases per 1,000 Population"
)

ggplot(vaccination_disease_control_indicators_long_format,
       aes(x = values, y = Life.expectancy)) +
  geom_point(alpha = 0.5, color = "#2e6a57") +
  geom_smooth(method = "lm", se = FALSE, color = "#9b2542") +
  facet_wrap(~ vaccination_disease_indicators, scales = "free_x",
             labeller = labeller(vaccination_disease_indicators =
                                   custom_labels_3)) +
  labs(title = "Impact of Vaccination and Disease Control Indicators<br>
       on Life Expectancy", x = "", y = "Life Expectancy (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```

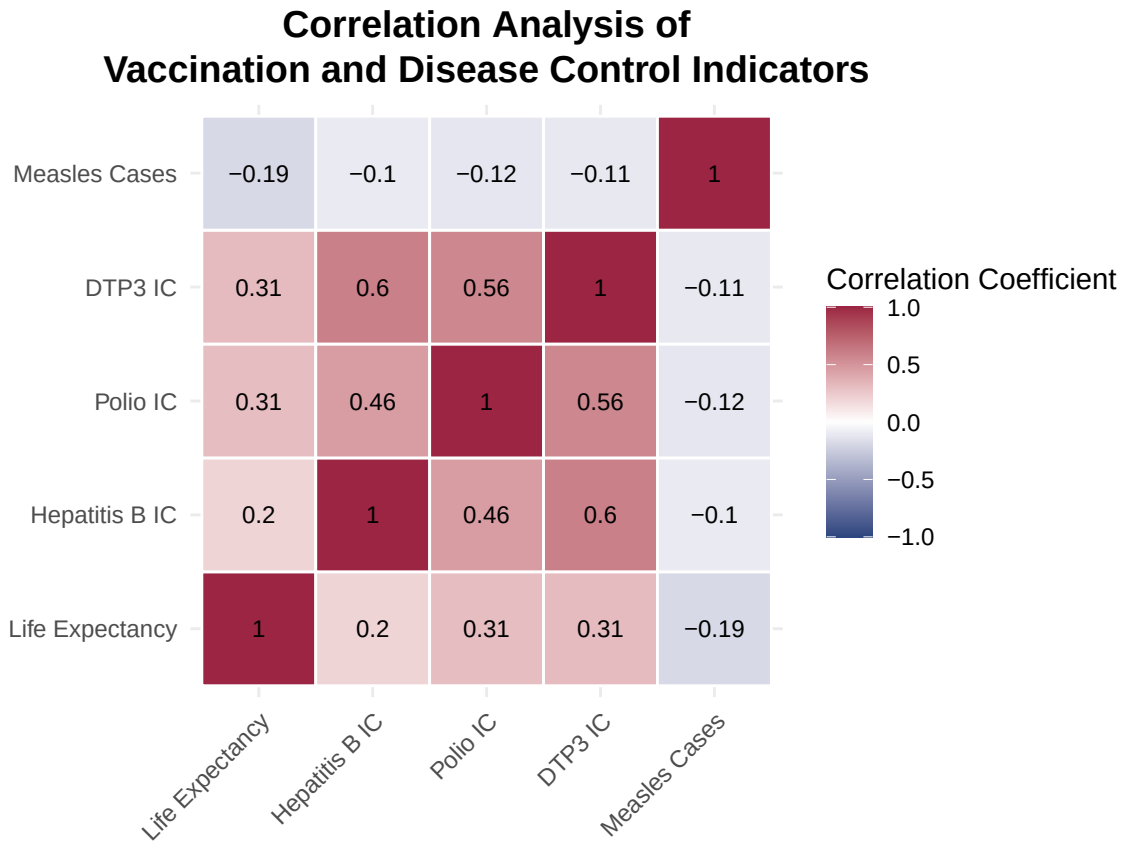



```
vaccination_disease_control_ind <- dataset %>%
  select(
    `Life Expectancy` = Life.expectancy,
    `Hepatitis B IC` = Hepatitis.B,
    `Polio IC` = Polio,
    `DTP3 IC` = Diphtheria,
    `Measles Cases` = Measles
  ) %>%
  filter(`Measles Cases` <= 1000)
vaccination_disease_control_ind_cor <-
  cor(vaccination_disease_control_ind, use = "complete.obs")
vaccination_disease_control_ind_cor_long_format <-
  melt(vaccination_disease_control_ind_cor)
head(vaccination_disease_control_ind_cor_long_format)
```

```
##              Var1            Var2      value
## 1 Life Expectancy Life Expectancy  1.0000000
## 2  Hepatitis B IC Life Expectancy  0.1979131
## 3        Polio IC Life Expectancy  0.3075591
## 4         DTP3 IC Life Expectancy  0.3133858
## 5   Measles Cases Life Expectancy -0.1890528
## 6 Life Expectancy  Hepatitis B IC  0.1979131
```

```
ggplot(vaccination_disease_control_ind_cor_long_format,
       aes(Var1, Var2, fill = value)) +
```

```
  geom_tile(color = "white", size = 0.5) +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  scale_fill_gradient2(low = "#2b457e", high = "#9b2542", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation Coefficient") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                                  lineheight = 1.2),
        axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.title.x = element_blank(),
        axis.title.y = element_blank()
        ) +
  coord_fixed() +
  labs(title = "Correlation Analysis of<br>
       Vaccination and Disease Control Indicators") +
  theme(plot.title = element_markdown())
```



The correlation between **Hepatitis B Immunization Coverage (IC)** and **Life Expectancy** is **0.2**. This indicates a weak positive relationship, suggesting that higher Hepatitis B immunization coverage is slightly associated with higher life expectancy.

The correlation between **Polio IC** and **Life Expectancy** is **0.31**. This shows a moderate positive relationship, meaning higher Polio immunization coverage is associated with higher life expectancy.

The correlation between **DTP3 IC** and **Life Expectancy** is **0.31**. This indicates a moderate positive relationship, suggesting that higher DTP3 (Diphtheria, Tetanus, and Pertussis) immunization coverage is associated with higher life expectancy.

The correlation between **Measles Cases** and **Life Expectancy** is **-0.19**. This shows a weak negative relationship, meaning higher numbers of measles cases are slightly associated with lower life expectancy.

In summary, immunization coverage for Hepatitis B, Polio, and DTP3 shows positive correlations with life expectancy, indicating that higher immunization rates are associated with higher life expectancy. The number of measles cases has a negative correlation with life expectancy, suggesting that more measles cases are associated with lower life expectancy.

Healthcare Expenditure and Resource Indicators

To visualize the impact of the **Healthcare Expenditure and Resource Indicators** on life expectancy,

```
healthcare_expenditure_resource_indicators <- dataset %>%
  select(Life.expectancy, Total.expenditure, percentage.expenditure)
head(healthcare_expenditure_resource_indicators)
```

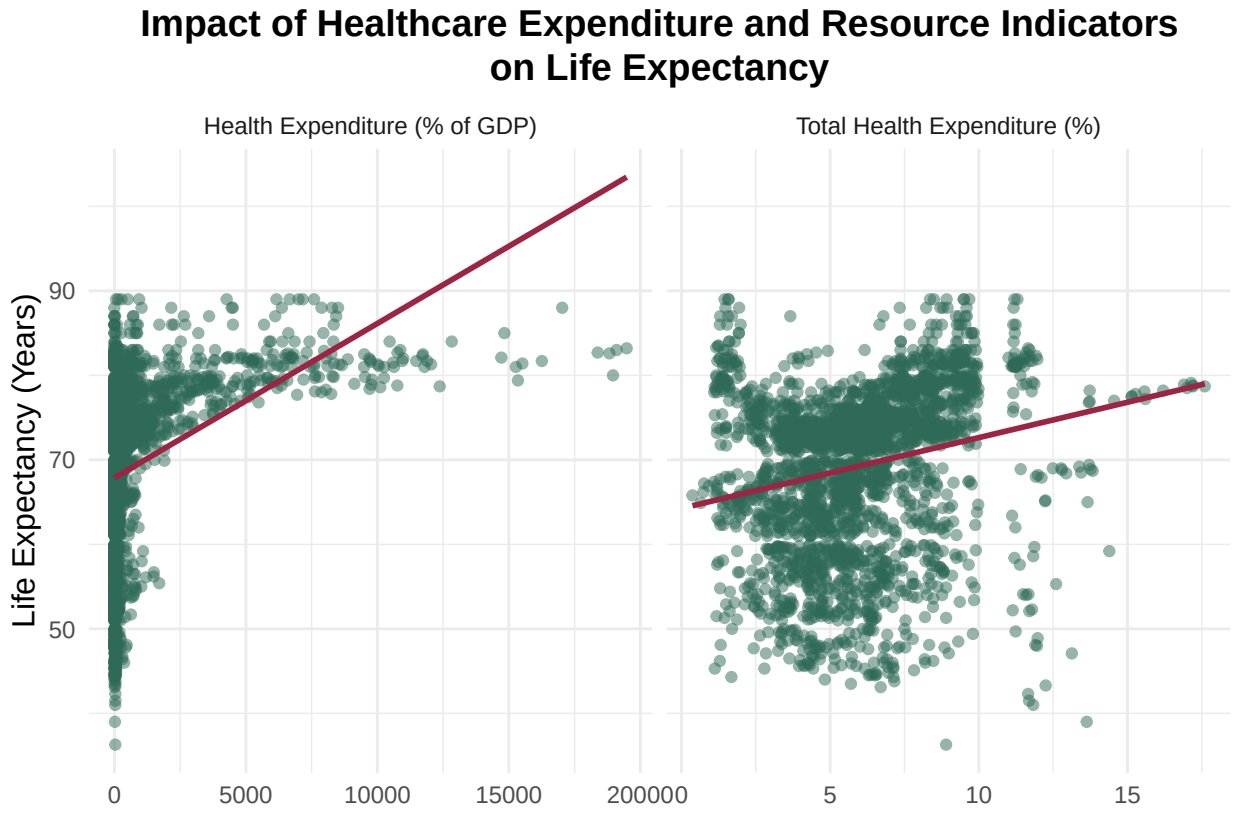
```
##   Life.expectancy Total.expenditure percentage.expenditure
## 1            65.0              8.16              71.279624
## 2            59.9              8.18              73.523582
## 3            59.9              8.13              73.219243
## 4            59.5              8.52              78.184215
## 5            59.2              7.87               7.097109
## 6            58.8              9.20              79.679367
```

```
healthcare_expenditure_resource_indicators_long_format <-
  healthcare_expenditure_resource_indicators %>%
  pivot_longer(cols = -Life.expectancy,
               names_to = "healthcare_expenditure_indicators",
               values_to = "values") %>%
  filter(!is.na(Life.expectancy),
         !(healthcare_expenditure_indicators == "Total.expenditure"
           & is.na(values)),
         !(healthcare_expenditure_indicators == "percentage.expenditure"
           & is.na(values)))
head(healthcare_expenditure_resource_indicators_long_format)
```

```
## # A tibble: 6 x 3
##   Life.expectancy healthcare_expenditure_indicators values
##             <dbl> <chr>                              <dbl>
## 1            65   Total.expenditure                   8.16
## 2            65   percentage.expenditure             71.3
## 3            59.9 Total.expenditure                   8.18
## 4            59.9 percentage.expenditure             73.5
## 5            59.9 Total.expenditure                   8.13
## 6            59.9 percentage.expenditure             73.2
```

```
custom_labels_4 <- c(
  "Total.expenditure" = "Total Health Expenditure (%)",
  "percentage.expenditure" = "Health Expenditure (% of GDP)"
)

ggplot(healthcare_expenditure_resource_indicators_long_format,
```

```
      aes(x = values, y = Life.expectancy)) +
  geom_point(alpha = 0.5, color = "#2e6a57") +
  geom_smooth(method = "lm", se = FALSE, color = "#9b2542") +
  facet_wrap(~ healthcare_expenditure_indicators, scales = "free_x",
             labeller = labeller(healthcare_expenditure_indicators =
                                   custom_labels_4)) +
  labs(title = "Impact of Healthcare Expenditure and Resource Indicators<br>
       on Life Expectancy", x = "", y = "Life Expectancy (Years)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```

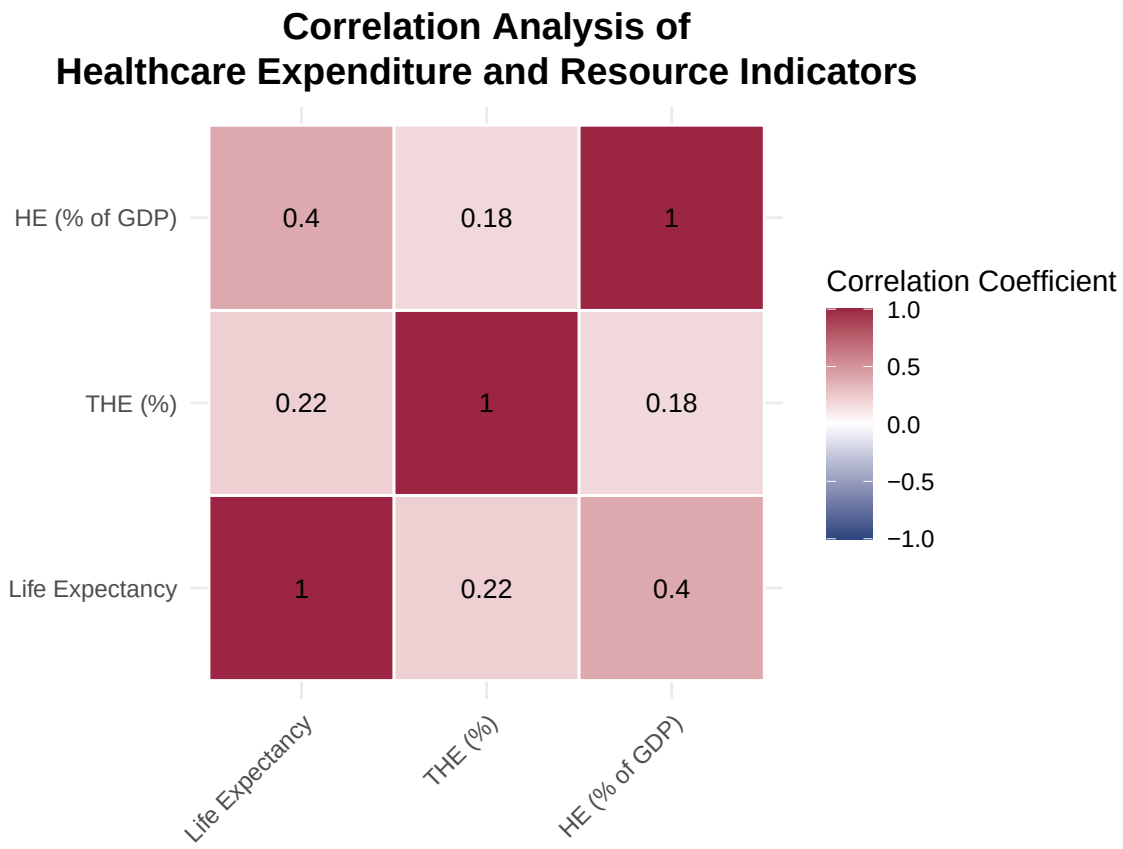



```
healthcare_expenditure_resource_ind <- dataset %>%
  select(
    `Life Expectancy` = Life.expectancy,
    `THE (%)` = Total.expenditure,
    `HE (% of GDP)` = percentage.expenditure
  )
healthcare_expenditure_resource_ind_cor <-
  cor(healthcare_expenditure_resource_ind, use = "complete.obs")
healthcare_expenditure_resource_ind_cor_long_format <-
```

```
  melt(healthcare_expenditure_resource_ind_cor)
head(healthcare_expenditure_resource_ind_cor_long_format)
```

```
##              Var1            Var2     value
## 1 Life Expectancy Life Expectancy 1.0000000
## 2         THE (%) Life Expectancy 0.2180864
## 3   HE (% of GDP) Life Expectancy 0.3995772
## 4 Life Expectancy         THE (%) 0.2180864
## 5         THE (%)         THE (%) 1.0000000
## 6   HE (% of GDP)         THE (%) 0.1762450
```

```
ggplot(healthcare_expenditure_resource_ind_cor_long_format,
       aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white", size = 0.5) +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3.5) +
  scale_fill_gradient2(low = "#2b457e", high = "#9b2542", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation Coefficient") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                                  lineheight = 1.2),
        axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.title.x = element_blank(),
        axis.title.y = element_blank()) +
  coord_fixed() +
  labs(title = "Correlation Analysis of<br>
       Healthcare Expenditure and Resource Indicators") +
  theme(plot.title = element_markdown())
```



The correlation between **Total Health Expenditure (%)** and **Life Expectancy** is **0.22**. This indicates a weak positive relationship, suggesting that higher total health expenditure as a percentage of total expenditure is slightly associated with higher life expectancy.

The correlation between **Health Expenditure (% of GDP)** and **Life Expectancy** is **0.4**. This shows a moderate positive relationship, meaning that higher health expenditure as a percentage of GDP is associated with higher life expectancy.

In summary, both total health expenditure and health expenditure as a percentage of GDP have positive correlations with life expectancy. This suggests that increased spending on health is associated with higher life expectancy, with health expenditure as a percentage of GDP showing a stronger relationship than total health expenditure.

Socio-Economic and Educational Indicators

To visualize the impact of the **Socio-Economic and Educational Indicators** on life expectancy,

```
socio_economic_educational_indicators <- dataset %>%
  select(Life.expectancy, Income.composition.of.resources,
         GDP, Schooling, Population)
head(socio_economic_educational_indicators)
```

```
##   Life.expectancy Income.composition.of.resources       GDP Schooling
## 1            65.0                           0.479 584.25921      10.1
## 2            59.9                           0.476 612.69651      10.0
## 3            59.9                           0.470 631.74498       9.9
```

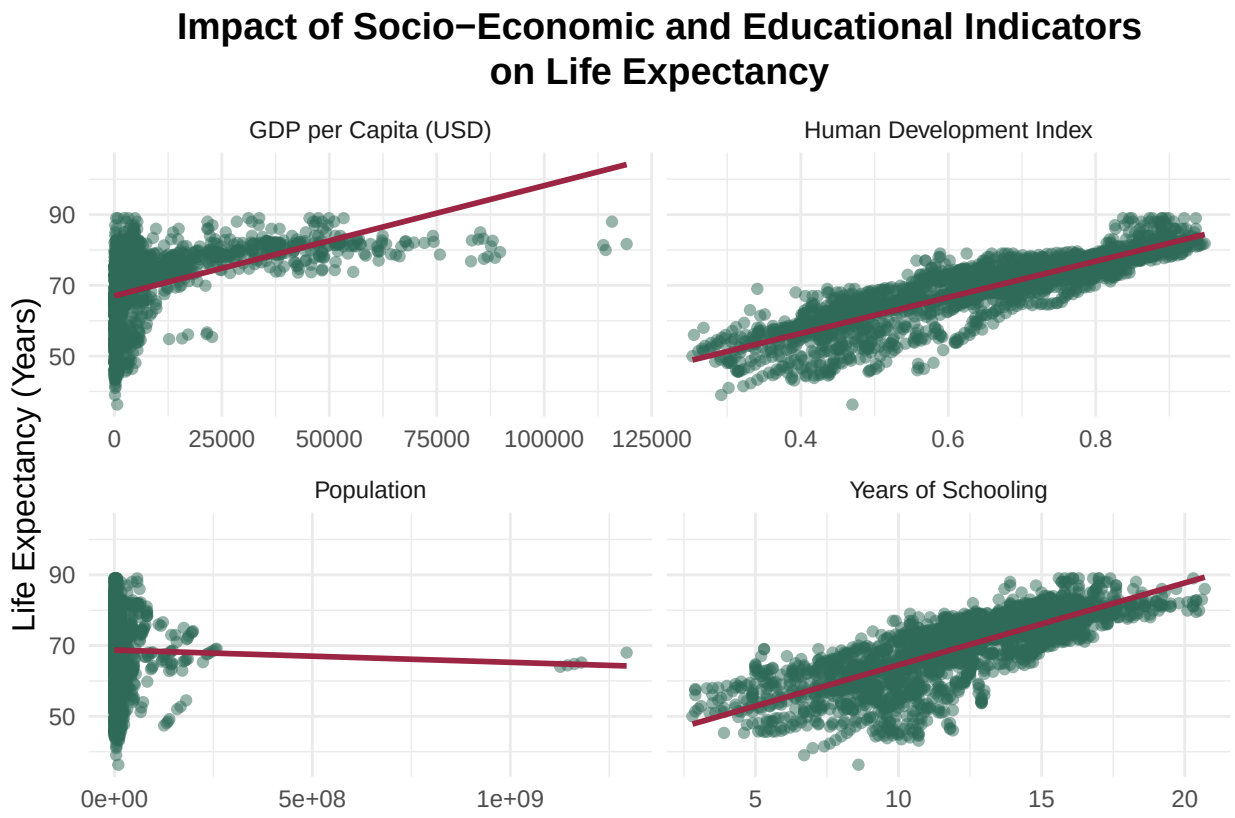
```
## 4           59.5                          0.463 669.95900       9.8
## 5           59.2                          0.454  63.53723       9.5
## 6           58.8                          0.448 553.32894       9.2
##   Population
## 1   33736494
## 2     327582
## 3   31731688
## 4    3696958
## 5    2978599
## 6    2883167
```

```
socio_economic_educational_indicators_long_format <-
  socio_economic_educational_indicators %>%
  pivot_longer(cols = -Life.expectancy,
               names_to = "socio_economic_indicators",
               values_to = "values") %>%
  filter(!is.na(Life.expectancy),
         !(socio_economic_indicators == "Income.composition.of.resources"
           & values == 0),
         !(socio_economic_indicators == "GDP"
           & values == 0),
         !(socio_economic_indicators == "Schooling"
           & values == 0),
         !(socio_economic_indicators == "Population"
           & values == 0))
head(socio_economic_educational_indicators_long_format)
```

```
## # A tibble: 6 x 3
##   Life.expectancy socio_economic_indicators             values
##             <dbl> <chr>                                  <dbl>
## 1            65   Income.composition.of.resources        0.479
## 2            65   GDP                                  584.
## 3            65   Schooling                             10.1
## 4            65   Population                        33736494
## 5            59.9 Income.composition.of.resources        0.476
## 6            59.9 GDP                                  613.
```

```
custom_labels_5 <- c(
  "Income.composition.of.resources" = "Human Development Index",
  "GDP" = "GDP per Capita (USD)",
  "Schooling" = "Years of Schooling",
  "Population" = "Population"
)

ggplot(socio_economic_educational_indicators_long_format,
       aes(x = values, y = Life.expectancy)) +
  geom_point(alpha = 0.5, color = "#2e6a57") +
  geom_smooth(method = "lm", se = FALSE, color = "#9b2542") +
  facet_wrap(~ socio_economic_indicators, scales = "free_x",
             labeller = labeller(socio_economic_indicators = custom_labels_5)) +
  labs(title = "Impact of Socio-Economic and Educational Indicators<br>
       on Life Expectancy", x = "", y = "Life Expectancy (Years)") +
  theme_minimal() +
```

```
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.title.y = element_text(size = 12)
  ) +
  theme(plot.title = element_markdown())
```

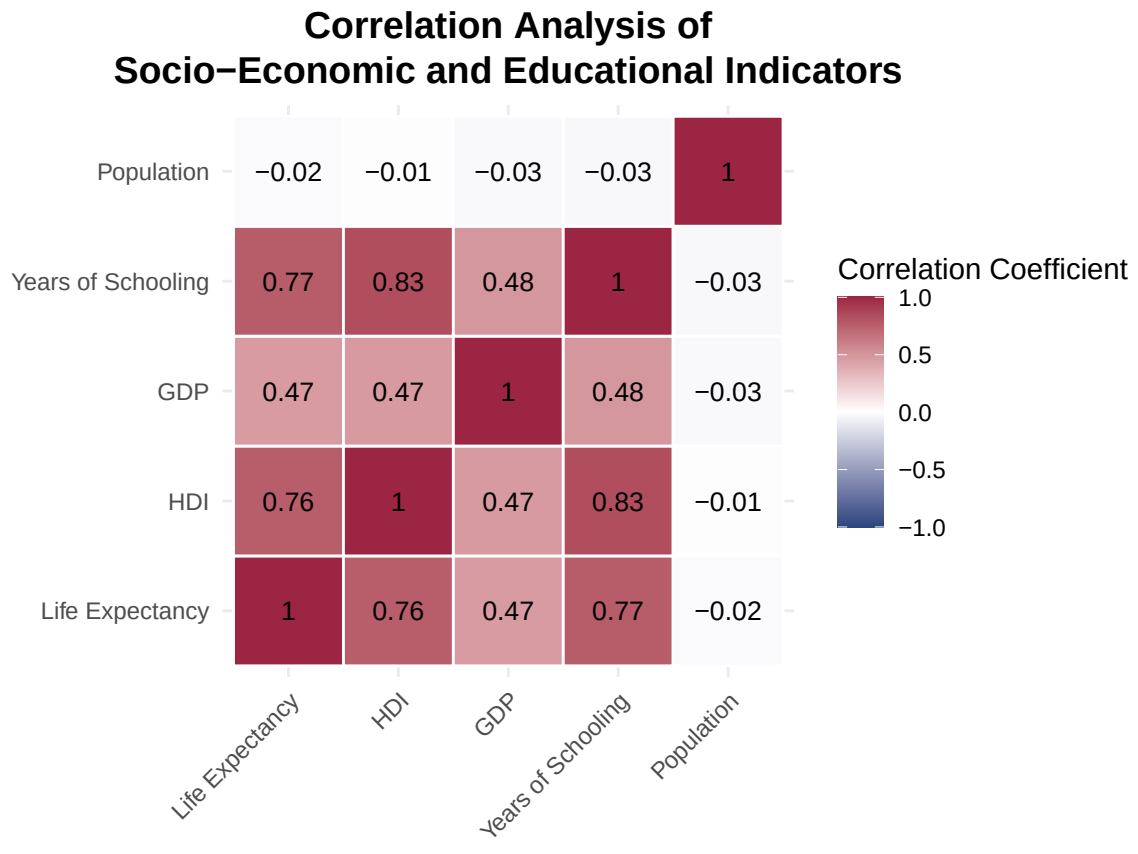



```
socio_economic_educational_ind <- dataset %>%
  select(
    `Life Expectancy` = Life.expectancy,
    `HDI` = Income.composition.of.resources,
    `GDP` = GDP,
    `Years of Schooling` = Schooling,
    `Population` = Population
    )
socio_economic_educational_ind_cor <-
  cor(socio_economic_educational_ind, use = "complete.obs")
socio_economic_educational_ind_cor_long_format <-
  melt(socio_economic_educational_ind_cor)
head(socio_economic_educational_ind_cor_long_format)
```

```
##                 Var1            Var2       value
## 1    Life Expectancy Life Expectancy  1.00000000
## 2                HDI Life Expectancy  0.75861711
## 3                GDP Life Expectancy  0.46566179
```

```
## 4 Years of Schooling Life Expectancy  0.76918472
## 5         Population Life Expectancy -0.02262795
## 6    Life Expectancy             HDI  0.75861711
```

```r
ggplot(socio_economic_educational_ind_cor_long_format,
       aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white", size = 0.5) +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3.5) +
  scale_fill_gradient2(low = "#2b457e", high = "#9b2542", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation Coefficient") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold",
                              lineheight = 1.2),
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    axis.title.x = element_blank(),
    axis.title.y = element_blank()) +
  coord_fixed() +
  labs(title = "Correlation Analysis of<br>
       Socio-Economic and Educational Indicators") +
  theme(plot.title = element_markdown())
```



The correlation between **HDI (Human Development Index)** and **Life Expectancy** is **0.76**. This indicates a strong positive relationship, meaning as HDI increases, Life Expectancy also tends to increase.

The correlation between **GDP (Gross Domestic Product)** and **Life Expectancy** is **0.47**. This shows a

moderate positive relationship, suggesting that higher GDP is associated with higher Life Expectancy, but the relationship is not as strong as with HDI.

The correlation between **Years of Schooling** and **Life Expectancy** is **0.77**. This indicates a strong positive relationship, similar to HDI, suggesting that more years of schooling are associated with higher Life Expectancy.

The correlation between **Population** and **Life Expectancy** is **-0.02**. This shows a very weak negative relationship, indicating that Population size has almost no linear relationship with Life Expectancy in this dataset.

In summary, HDI and Years of Schooling have strong positive correlations with Life Expectancy, while GDP has a moderate positive correlation. Population size, however, shows almost no correlation with Life Expectancy.

Hypothesis Testing

#Hypothesis Testing and Linear Regression for Schooling

Null hypothesis (H0) that there is no relationship between schooling and life expectancy. Alternative hypothesis (H1) More schooling leads to higher life expectancy

```
data1_filtered <- dataset[dataset$Schooling != 0, ]
data1_filtered <- data1_filtered[!(data1_filtered$Schooling == 0 | data1_filtered$Schooling == "" | is.
```

```
mean_value <- mean(data1_filtered$Schooling)
median_value <- median(data1_filtered$Schooling)
max_value <- max(data1_filtered$Schooling)
min_value <- min(data1_filtered$Schooling)
```

```
cat("Mean:", mean_value, "\n")
```

```
## Mean: 12.11342
```

```
cat("Median:", median_value, "\n")
```

```
## Median: 12.4
```

```
cat("Max:", max_value, "\n")
```

```
## Max: 20.7
```

```
cat("Min:", min_value, "\n")
```

```
## Min: 2.8
```

```
model <- lm(`Life.expectancy` ~ Schooling, data = data1_filtered)

model_summary <- summary(model)
model_summary
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling, data = data1_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.9571  -2.7750   0.7344   3.8731  15.4014
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.29864    0.44008   93.84   <2e-16 ***
## Schooling    2.32075    0.03516   66.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.8 on 2740 degrees of freedom
## Multiple R-squared:  0.6139, Adjusted R-squared:  0.6138
## F-statistic:  4356 on 1 and 2740 DF,  p-value: < 2.2e-16
```

Very low p value, can reject null hypothesis, is a relationship between schooling and life expectancy. Schooling has a statistically significant effect on life expectancy.

41.3 is estimated life expectancy when schooling is 0, life expectancy expected to increased by about 2.32 years for every year of schooling.

R-squared of 0.6139 suggests that 61.39% of the variance in life expectancy is explained by schooling

```
coef_summary <- summary(model)$coefficients
t_statistic <- coef_summary["Schooling", "Estimate"] / coef_summary["Schooling", "Std. Error"]

df <- model$df.residual

alpha <- 0.05
t_critical <- qt(1 - alpha, df)

cat("T-Statistic:", t_statistic, "\n")
```

```
## T-Statistic: 66.00378
```

```
cat("Critical T-Value:", t_critical, "\n")
```

```
## Critical T-Value: 1.64541
```

```
if (t_statistic > t_critical) {
  cat("Reject the null hypothesis: Schooling is associated with increased life expectancy.\n")
} else {
  cat("Fail to reject the null hypothesis: No significant association.\n")
}
```

```
## Reject the null hypothesis: Schooling is associated with increased life expectancy.
```

```r
estimate <- model_summary$coefficients["Schooling", "Estimate"]
std_error <- model_summary$coefficients["Schooling", "Std. Error"]

margin_of_error <- t_critical * std_error
lower_bound <- estimate - margin_of_error
upper_bound <- estimate + margin_of_error


cat("95% Confidence Interval for the Schooling coefficient:", lower_bound, "to", upper_bound, "\n")
```

```
## 95% Confidence Interval for the Schooling coefficient: 2.262899 to 2.378608
```

Very strong evidence that for each year of schooling expect life expectancy to increase between 2.26 and 2.38 years.

Linear Regression