EXAMINING THE IMPACT OF HEALTH INDICATORS, SOCIOECONOMIC FACTORS, AND EDUCATION ON GLOBAL LIFE EXPECTANCY.

Group 14:

SRINATH KALEPU
DEEEPTHI MALLEPALLY
SAI MAHIDHAR REDDY BYREDDY
SREE UMA MAHESHWAR VANGAPATY
SAI KUMAR REDDY YALAGALAPALLI

INTRODUCTION

- **Influence of Socioeconomic Status (SES):** Lower SES is linked to reduced life expectancy due to higher mortality rates, regardless of age and sex.
- **Educational and Occupational Disparities:** Significant disparities in life expectancy are associated with levels of education and type of occupation, with lower education and manual occupations linked to shorter life spans (Singh & Lee, 2020).
- **Economic Development:** Higher GDP per capita is associated with increased life expectancy, as economic growth provides resources for better nutrition, healthcare, and living conditions (Freeman et al., 2020; Rogoz et al., 2022).

RESEARCH QUESTION &HYPOTHESIS

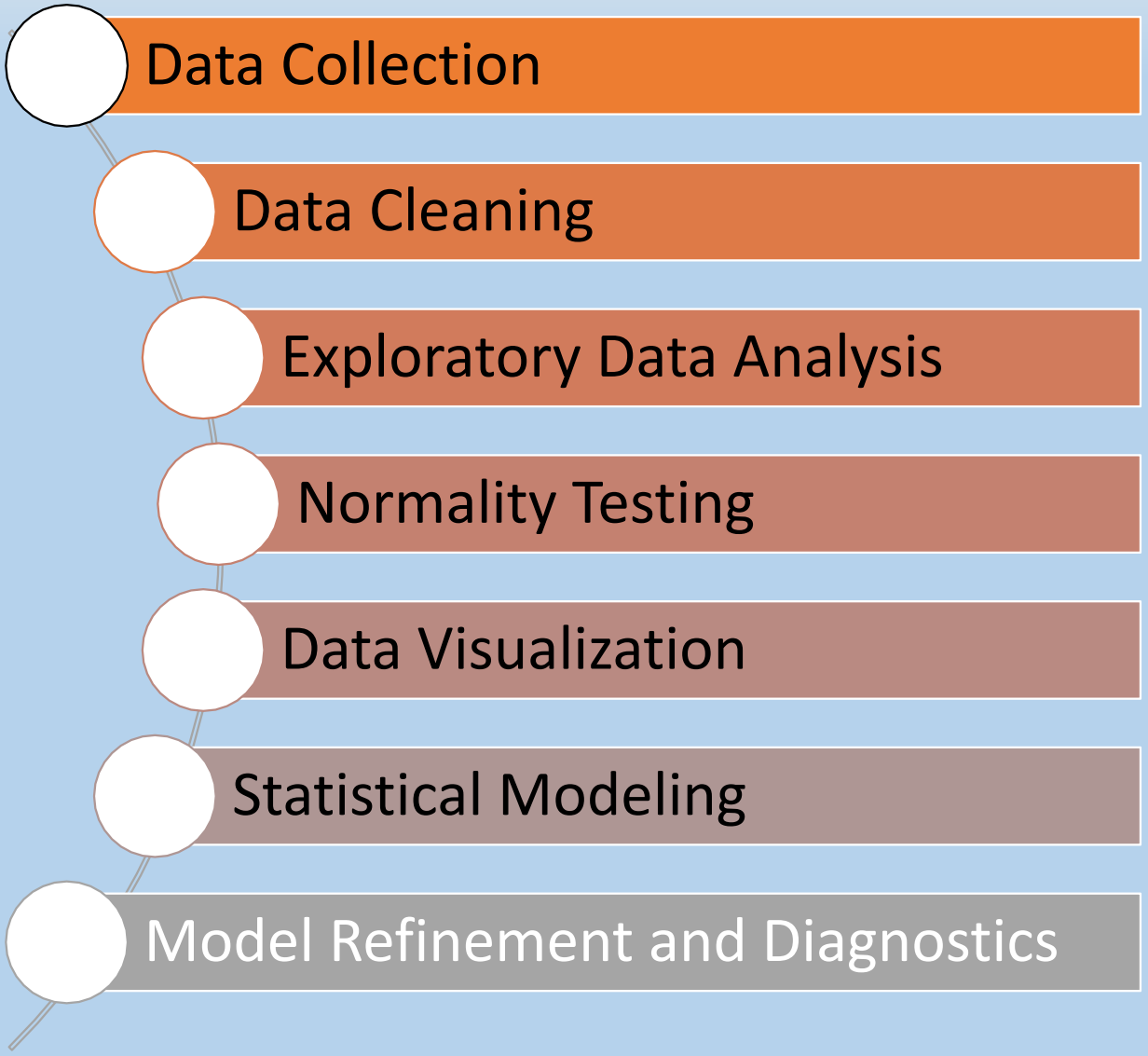
- **Research Question1:** What are the primary health indicators and socio-economic factors that significantly influence life expectancy across different countries?
- **Null Hypothesis (H0):** The health indicators and socio-economic factors (such as adult mortality, infant deaths, health expenditure, Hepatitis B coverage, under-five deaths, Diphtheria immunization coverage, HIV/AIDS prevalence, income composition of resources, thinness in children aged 5-9 years) do not have a significant impact on life expectancy.
- **Alternative Hypothesis (Ha):** At least one of the health indicators and socio-economic factors has a significant impact on life expectancy.
- **Research Question2:** How does the level of education impact life expectancy when controlling for economic status and health indicators?
- **Null Hypothesis (H0):** The level of education, as represented by the "Schooling" variable, does not have a significant impact on life expectancy after accounting for economic status and health indicators.
- **Alternative Hypothesis (Ha):** The level of education has a significant impact on life expectancy, even after controlling for economic status and health indicators

DATA SET

- Our project uncovers the determinants of life expectancy by analyzing global health data from 2000-2015. Through rigorous data analysis using R, we aim to understand how factors such as health indicators, economic status, and education levels impact life expectancy.
- The dataset spans from 2000 to 2015, covering 193 countries, and includes variables related to life expectancy and various health factors obtained from the WHO Global Health Observatory data repository.
- Economic data was sourced from the United Nations website. This comprehensive dataset allows for a detailed analysis of trends and factors influencing life expectancy over a significant period.
- Data collected from Kaggle .The final dataset used for modelling comprises 2938 rows and 22 columns, with 20 predictor variables categorized into immunization, mortality, economic, and social factors.

Link:
https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

STEPS


Data Collection
Data Cleaning
Exploratory Data Analysis
Normality Testing
Data Visualization
Statistical Modeling
Model Refinement and Diagnostics

Analyzing Data Variables

Numerical Variables

- Adult Mortality
- Infant deaths
- Alcohol
- Percentage Expenditure
- Hepatitis B
- Measles
- BMI
- Under five deaths
- Polio
- Total Expenditure
- Diphtheria
- HIV. AIDS
- GDP
- Thinness..1..19.years
- Thinness.5.9.years
- Income Composition of Resources
- Schooling

Categorical Variables

- Status

Data Importing & Data Description

- The CSV file containing Life Expectancy data was uploaded to R using the "read.csv" code.
- We examined the first few rows of the modified dataset with the command "head(data)" and obtained a string type of the columns dataset using “str(data).”

```r
data <- read.csv("D:/Life Expectancy Data.csv")
```

	Status <chr>	Life.expectancy <dbl>	Adult.Mortality <dbl>	infant.deaths <dbl>	Alcohol <dbl>	percentage.expenditure <dbl>	Hepatitis.B <dbl>	Measles <dbl>	BMI <dbl>
1	Developing	65.0	263	45	0.01	71.279624	65	730	19.1
2	Developing	59.9	271	45	0.01	73.523582	62	492	18.6
3	Developing	59.9	268	45	0.01	73.219243	64	430	18.1
4	Developing	59.5	272	45	0.01	78.184215	67	730	17.6
5	Developing	59.2	275	45	0.01	7.097109	68	730	17.2
6	Developing	58.8	279	45	0.01	79.679367	66	730	16.7

6 rows | 1-10 of 19 columns

```
'data.frame':   1853 obs. of  19 variables:
 $ 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                : num  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : num  45 45 45 45 45 45 45 45 45 45 ...
 $ 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                    : num  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : num  730 492 430 730 730 730 730 730 730 730 ...
 $ 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              : num  55 55 55 55 55 55 55 55 55 55 ...
 $ Polio                          : num  62 62 62 67 68 66 63 64 63 62 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73
7.43 ...
 $ Diphtheria                     : num  65 62 64 67 68 66 63 64 63 62 ...
 $ 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 ...
 $ thinness..1.19.years           : num  14.4 14.4 14.4 14.4 14.4 ...
 $ thinness.5.9.years             : num  14.3 14.3 14.3 14.3 14.3 14.3 14.3 14.3 14.3
14.3 ...
 $ 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 ...
```

```r
head(data)
str(data)
```

Defining the selected variables

```r
selected_data <- data[, !(names(data) %in% c("Country", "Year", "Population"))]
str(selected_data)
```

- We removed extra spaces in column names and later we selected relevant variables from a dataset by excluding the columns “Country”, “Year”, and “Population”.

```r
names(data)
```

```
 [1] "Country"                        "Year"                  "Status"                    "Life.expectancy"
 [5] "Adult.Mortality"                "infant.deaths"         "Alcohol"                   "percentage.expenditure"
 [9] "Hepatitis.B"                    "Measles"               "BMI"                       "under.five.deaths"
[13] "Polio"                          "Total.expenditure"     "Diphtheria"                "HIV.AIDS"
[17] "GDP"                            "Population"            "thinness..1.19.years"      "thinness.5.9.years"
[21] "Income.composition.of.resources" "Schooling"
```

Data Cleaning

- We calculated the number of missing values in each column of the "`selected_data`" by applying a function that sums up the NA values in each column and stores the result in the "`missing_values`" object.

```{r}
missing_values <- sapply(selected_data, function(x) sum(is.na(x)))
missing_values
```

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

- We removed rows with missing values from the "`selected_data`", stored the result in "`selected_data1`", and then verified the removal by calculating and displaying the number of missing values in each column of the "`selected_data1`".

```{r}
missing_values <- sapply(selected_data1, function(x) sum(is.na(x)))
missing_values
```

```
                     Status               Life.expectancy               Adult.Mortality                  infant.deaths                       Alcohol
                          0                            0                            0                              0                             0
      percentage.expenditure                  Hepatitis.B                      Measles                           BMI             under.five.deaths
                          0                            0                            0                              0                             0
                    Polio                Total.expenditure                  Diphtheria                      HIV.AIDS                           GDP
                          0                            0                            0                              0                             0
       thinness..1.19.years             thinness.5.9.years Income.composition.of.resources                  Schooling
                          0                            0                            0                              0
```

Exploratory Data Analysis

Summary Statistics:

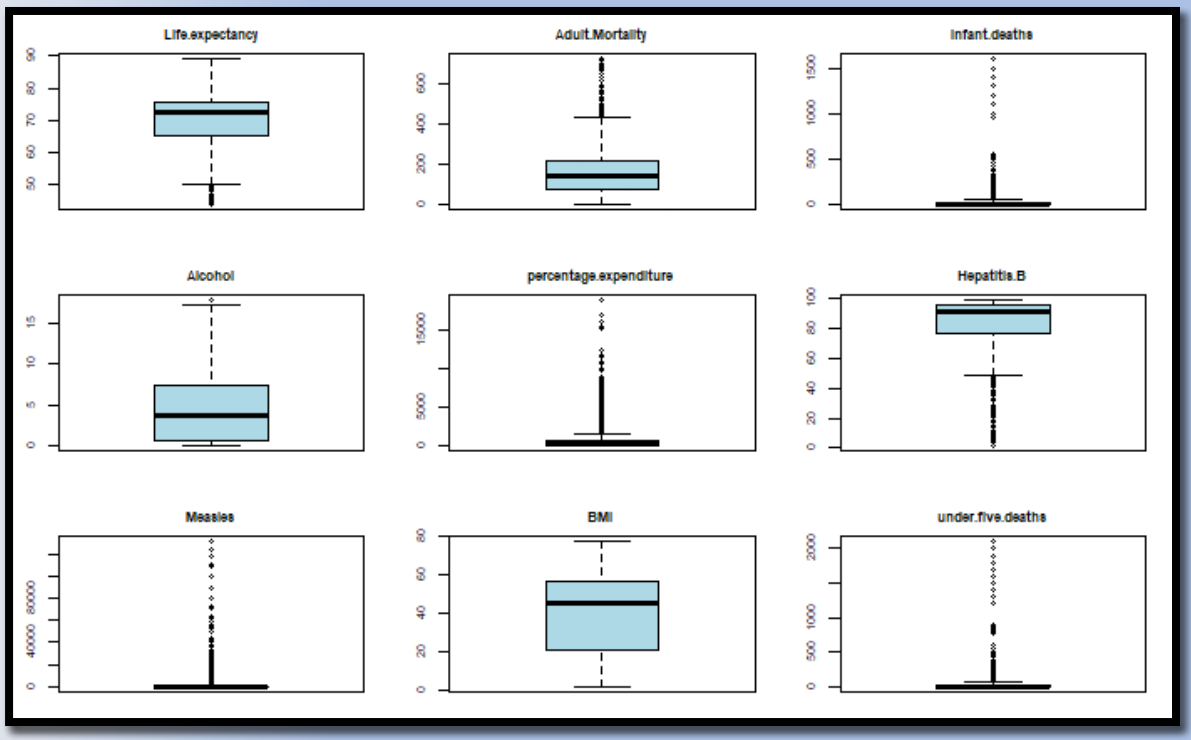
- We generated a summary of the "`selected_data1`", providing key statistical measures such as mean, median, minimum, maximum, and quartiles for each variable in the dataset. This exploration provided a detailed insight into the central tendencies and distribution characteristics of the dataset.

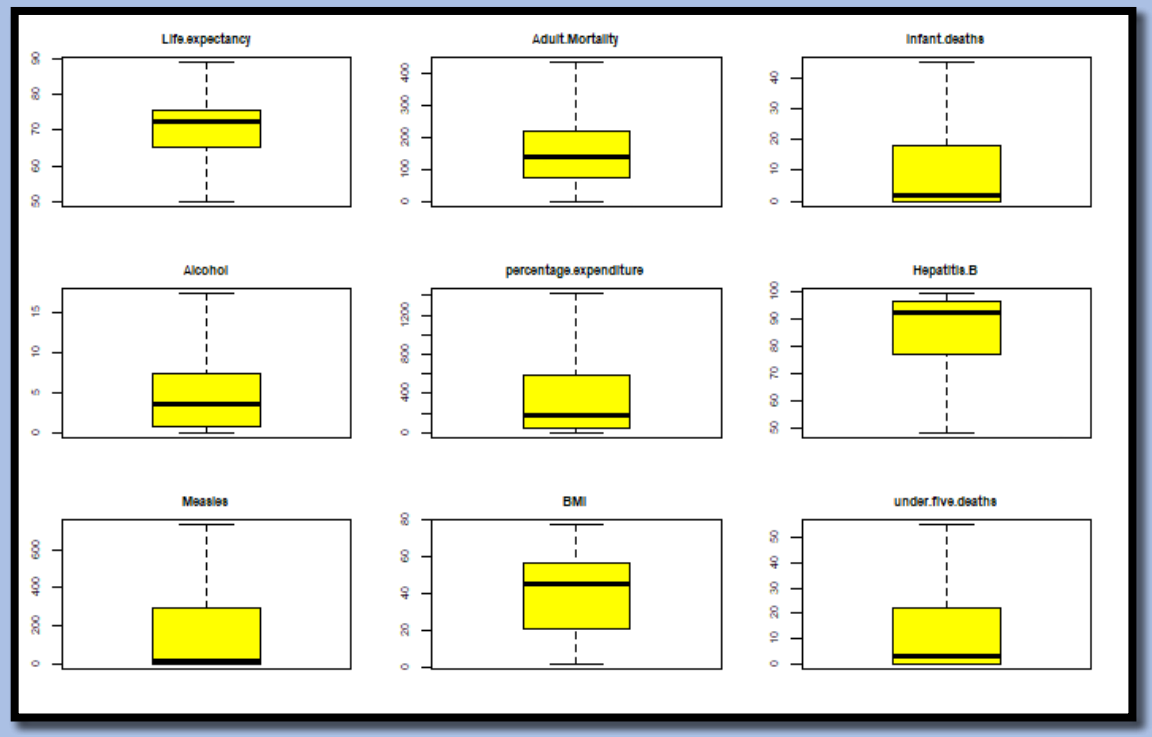
```r
summary(selected_data1)
```

```
   Status           Life.expectancy Adult.Mortality infant.deaths       Alcohol       percentage.expenditure  Hepatitis.B       Measles
 Length:1853        Min.   :44.00   Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.00       Min.   : 2.00   Min.   :     0
 Class :character   1st Qu.:65.30   1st Qu.: 74.0   1st Qu.:   0.00   1st Qu.: 0.660   1st Qu.:   41.91       1st Qu.:77.00   1st Qu.:     0
 Mode  :character   Median :72.50   Median :138.0   Median :   2.00   Median : 3.570   Median :  169.20       Median :92.00   Median :    13
                    Mean   :70.03   Mean   :158.7   Mean   :  29.08   Mean   : 4.393   Mean   :  764.91       Mean   :80.81   Mean   :  1994
                    3rd Qu.:75.50   3rd Qu.:217.0   3rd Qu.:  18.00   3rd Qu.: 7.380   3rd Qu.:  591.78       3rd Qu.:96.00   3rd Qu.:   292
                    Max.   :89.00   Max.   :723.0   Max.   :1600.00   Max.   :17.870   Max.   :18961.35       Max.   :99.00   Max.   :131441
      BMI        under.five.deaths     Polio       Total.expenditure   Diphtheria      HIV.AIDS           GDP             thinness..1.19.years
 Min.   : 2.00   Min.   :   0.00   Min.   : 3.00   Min.   : 0.740    Min.   : 2.0   Min.   : 0.100   Min.   :     1.68   Min.   : 0.100
 1st Qu.:21.20   1st Qu.:   0.00   1st Qu.:83.00   1st Qu.: 4.200    1st Qu.:83.0   1st Qu.: 0.100   1st Qu.:   526.53   1st Qu.: 1.700
 Median :44.90   Median :   3.00   Median :94.00   Median : 5.660    Median :93.0   Median : 0.100   Median :  1938.00   Median : 3.400
 Mean   :39.13   Mean   :  39.48   Mean   :84.83   Mean   : 5.784    Mean   :85.2   Mean   : 1.778   Mean   :  6762.02   Mean   : 4.809
 3rd Qu.:56.40   3rd Qu.:  22.00   3rd Qu.:97.00   3rd Qu.: 7.350    3rd Qu.:97.0   3rd Qu.: 0.500   3rd Qu.:  5836.18   3rd Qu.: 6.800
 Max.   :77.10   Max.   :2100.00   Max.   :99.00   Max.   :14.390    Max.   :99.0   Max.   :50.600   Max.   :119172.74   Max.   :27.200
 thinness.5.9.years Income.composition.of.resources   Schooling
 Min.   : 0.100     Min.   :0.0000                   Min.   : 0.00
 1st Qu.: 1.800     1st Qu.:0.5270                   1st Qu.:10.60
 Median : 3.400     Median :0.6910                   Median :12.50
 Mean   : 4.844     Mean   :0.6433                   Mean   :12.32
 3rd Qu.: 6.800     3rd Qu.:0.7760                   3rd Qu.:14.20
 Max.   :28.200     Max.   :0.9360                   Max.   :20.70
```

Outliers Detection

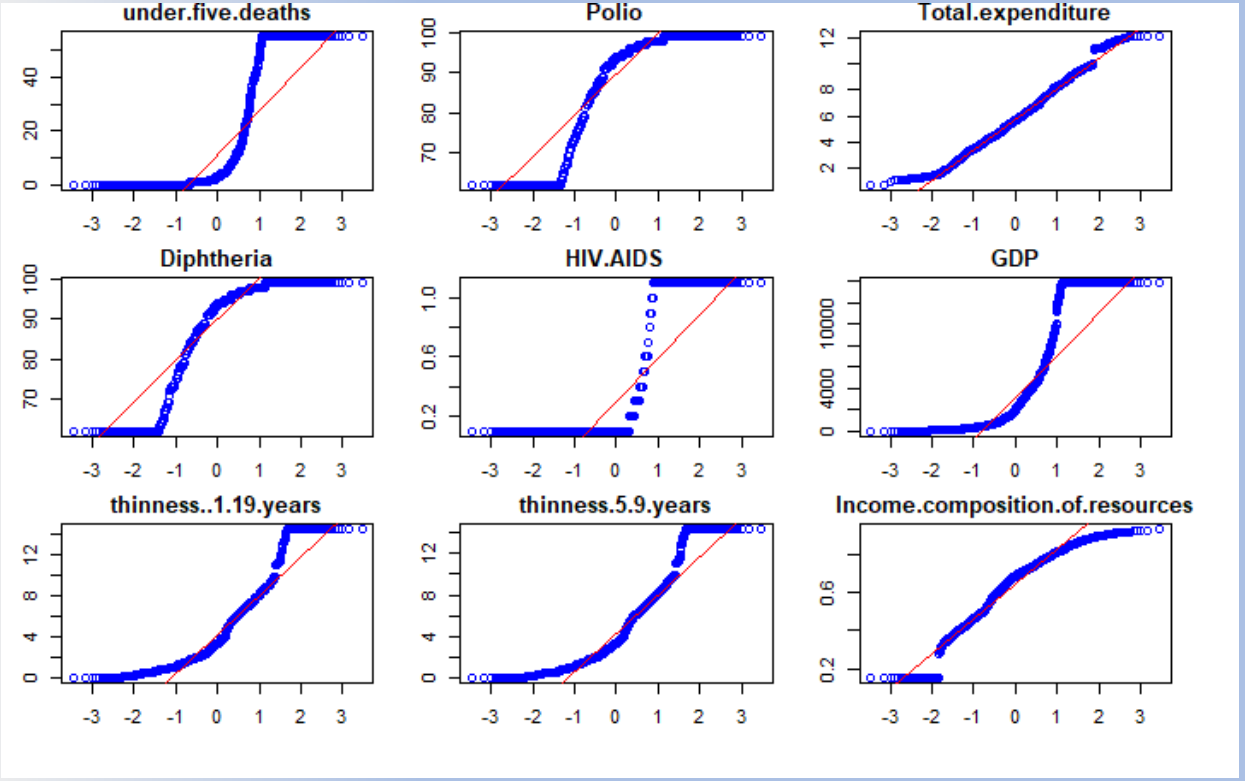
- We generated boxplots for each numeric column in the "`selected_data1`". to visualize outliers and then defined a function to cap outliers at 1.5 times the interquartile range below the first quartile and above the third quartile and apply this function to each numeric column in the "`selected_data1`", and finally checked the number of outliers in each numeric column after the capping process.

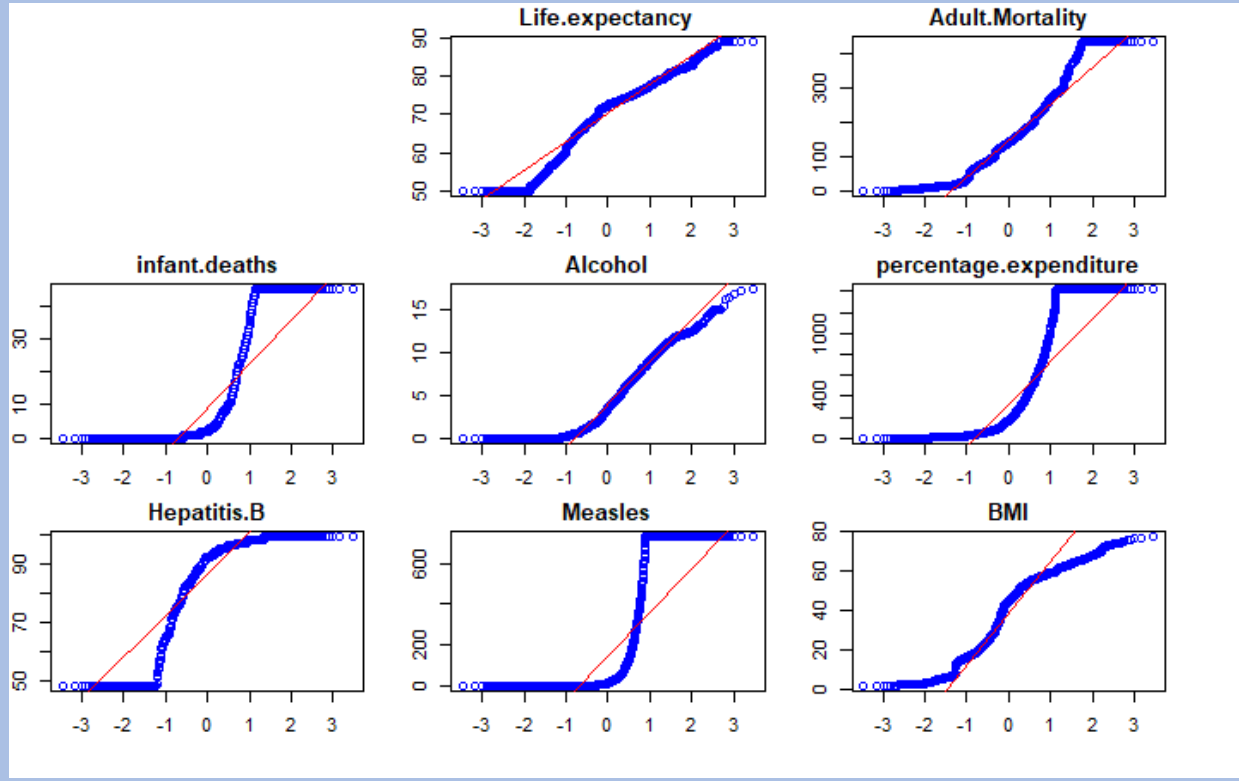




Checking for Normality

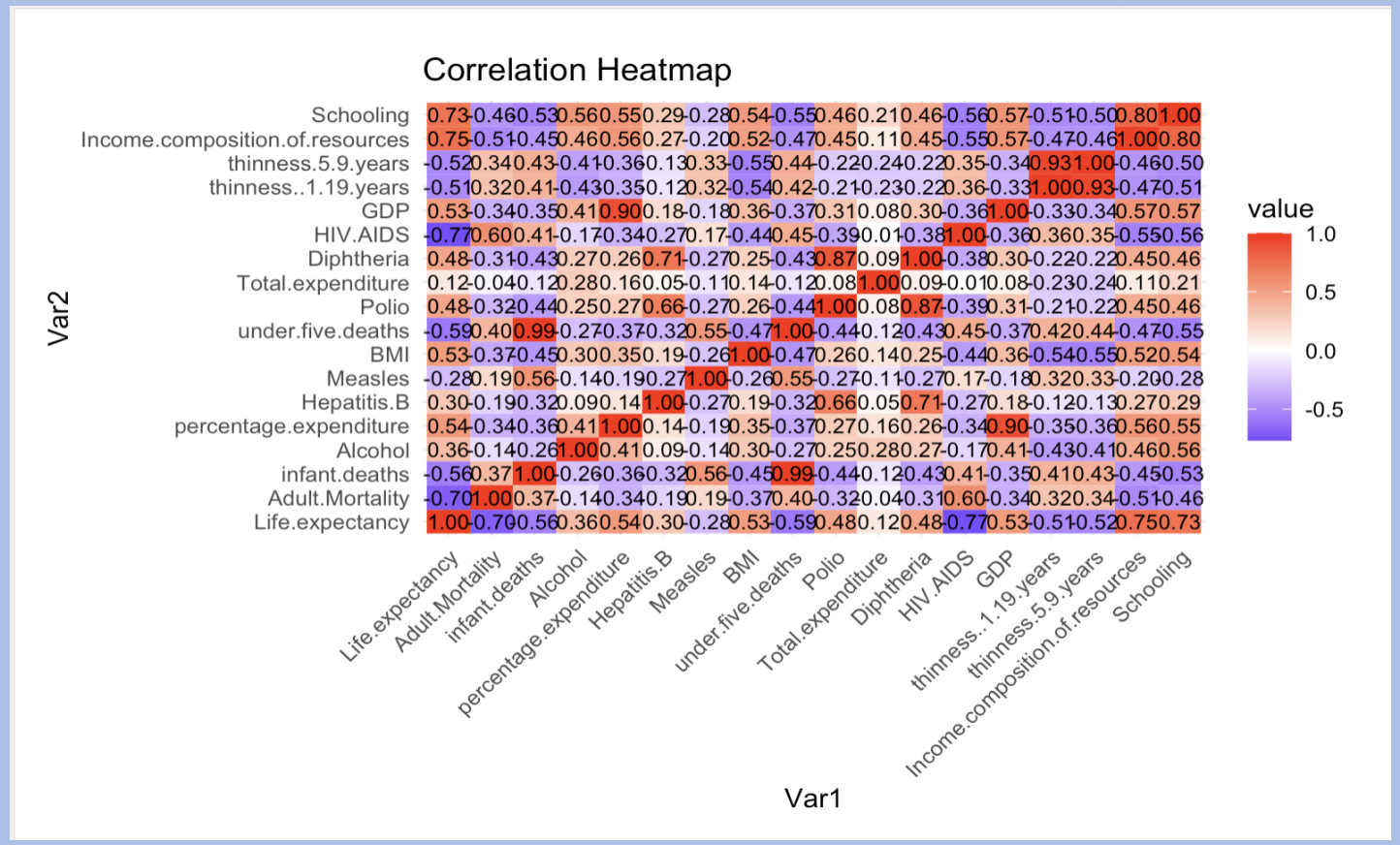
- The code generates Quantile-Quantile (Q-Q) plots for each numeric column in the `"selected_data1"`. dataset to assess if the data in these columns follows a normal distribution, with the plots organized in batches of up to nine per page for easy visualization





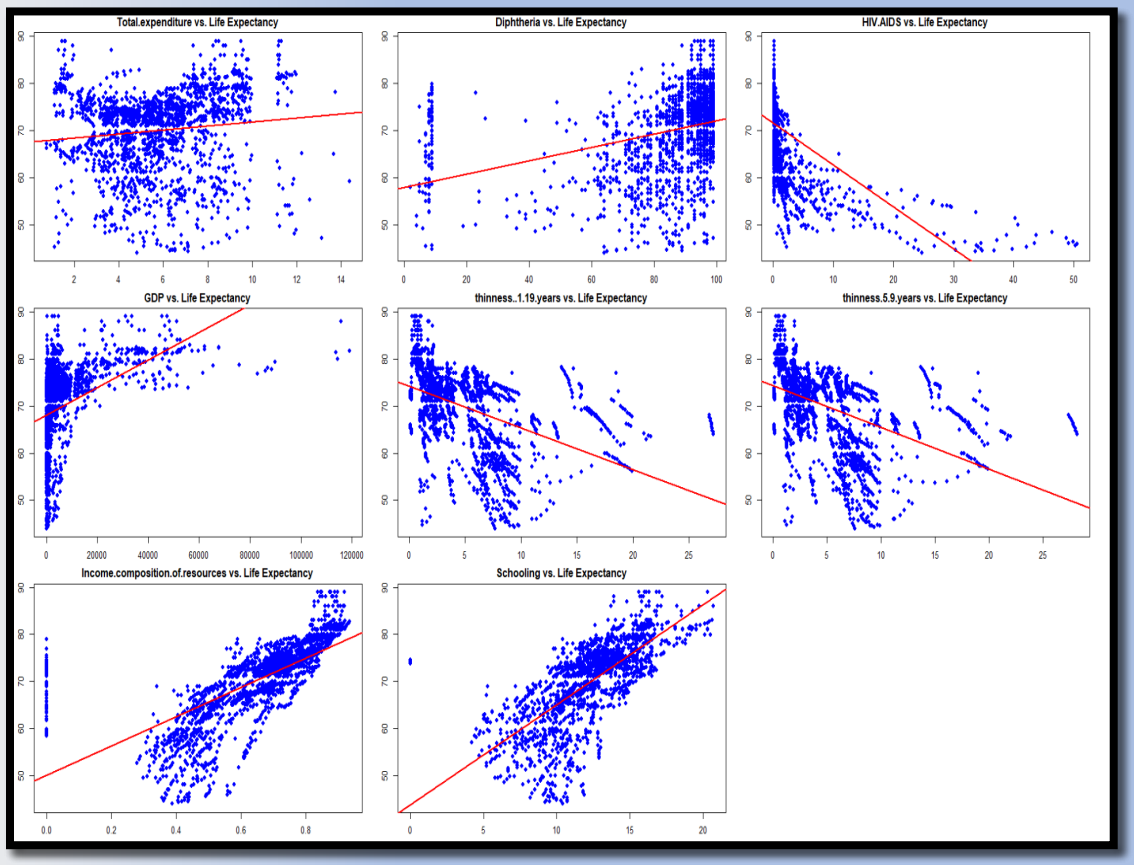
Correlation Heatmap

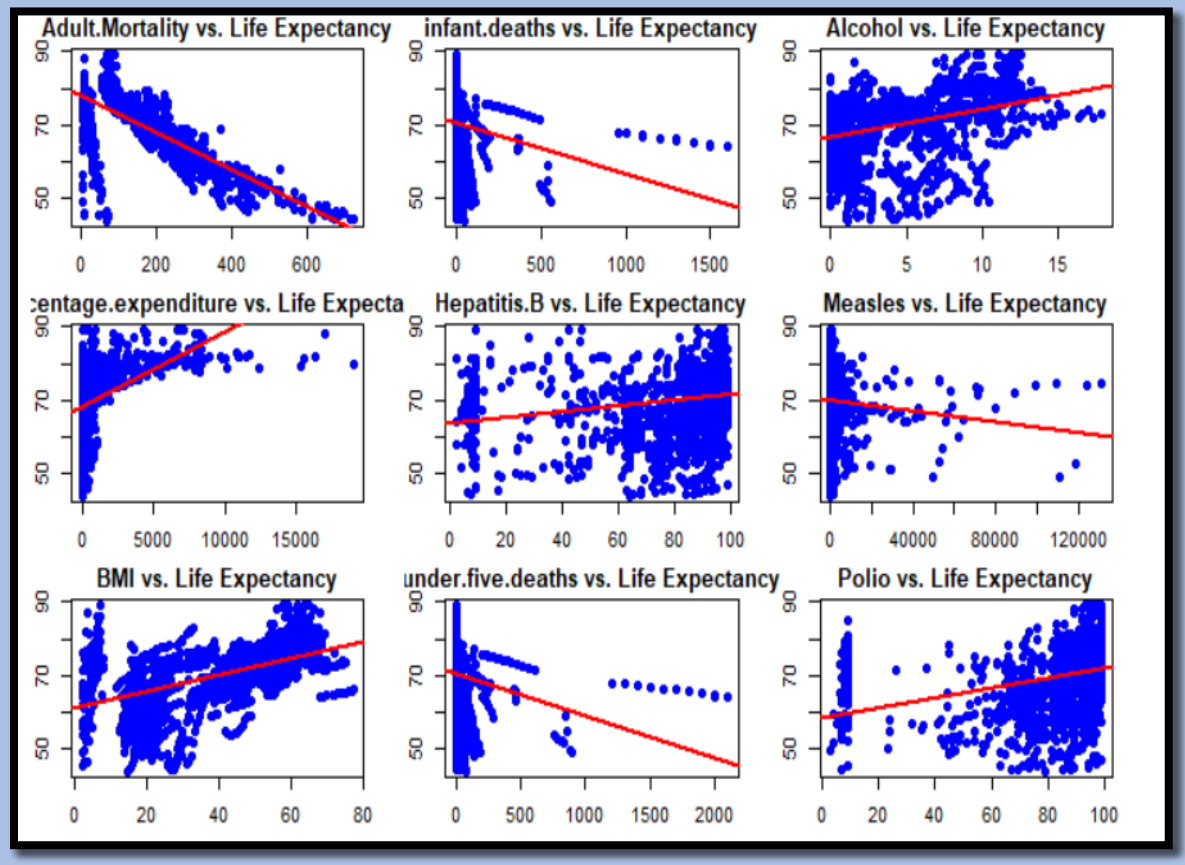
- We calculated the correlation matrix for the numeric columns for the data, reshaped this matrix into a format suitable for visualization, and then created a heatmap to visualize the correlations, with correlation coefficients displayed on the heatmap.



Statistical Modelling

- We generated scatter plots for each numeric column (excluding “Life. expectancy”) in the dataset against “Life.expectancy”, fit a linear regression line to each plot, and organize the plots for easy visualization.





Multiple Linear Regression Model

- We conducted a multiple linear regression model to visualize the data in a linear distribution. The findings highlight the variables that significantly impact life expectancy, such as developing status, adult mortality, under-five deaths, increased BMI, and vaccination coverage. The HIV/AIDS rate is a major concern that dramatically decreases life expectancy. Additionally, higher levels of human development, education, and income composition of resources show strong positive associations, indicating the importance of targeted health interventions, improved education, and resource allocation for enhancing life expectancy in various countries.

```r
full_model <- lm(Life.expectancy ~ ., data = selected_data1)
summary(full_model)
```

```
Call:
lm(formula = Life.expectancy ~ ., data = selected_data1)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9024  -1.8961   0.1136   1.8703  11.3512

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     6.386e+01  1.020e+00  62.629  < 2e-16 ***
StatusDeveloping               -1.292e+00  2.992e-01  -4.319 1.65e-05 ***
Adult.Mortality                -1.759e-02  9.679e-04 -18.176  < 2e-16 ***
infant.deaths                   2.209e-01  5.482e-02   4.029 5.82e-05 ***
Alcohol                        -4.704e-02  2.870e-02  -1.639 0.101351
percentage.expenditure          2.164e-03  3.981e-04   5.437 6.15e-08 ***
Hepatitis.B                    -2.052e-02  7.170e-03  -2.861 0.004265 **
Measles                        -1.369e-04  3.502e-04  -0.391 0.695893
BMI                            -3.668e-03  5.435e-03  -0.675 0.499893
under.five.deaths              -2.218e-01  4.407e-02  -5.033 5.30e-07 ***
Polio                           1.592e-03  1.402e-02   0.114 0.909638
Total.expenditure               5.138e-02  3.729e-02   1.378 0.168391
Diphtheria                      5.845e-02  1.542e-02   3.790 0.000155 ***
HIV.AIDS                       -6.650e+00  3.021e-01 -22.014  < 2e-16 ***
GDP                            -7.873e-05  4.078e-05  -1.930 0.053712 .
thinness..1.19.years            4.250e-02  5.890e-02   0.722 0.470663
thinness.5.9.years             -2.166e-01  5.949e-02  -3.641 0.000279 ***
Income.composition.of.resources 1.038e+01  8.450e-01  12.285  < 2e-16 ***
Schooling                       2.569e-01  5.740e-02   4.477 8.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.438 on 1834 degrees of freedom
Multiple R-squared:  0.8339,    Adjusted R-squared:  0.8322
F-statistic: 511.4 on 18 and 1834 DF,  p-value: < 2.2e-16
```

Model Refinement

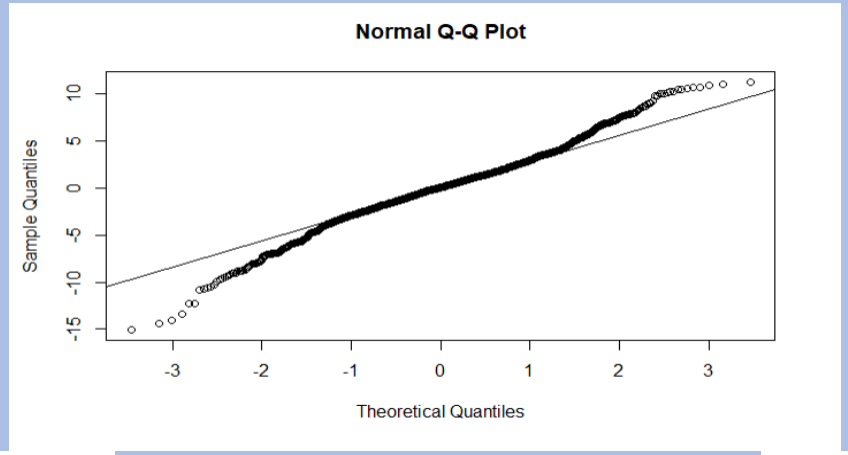
- Building Linear regression model with only significant predictors

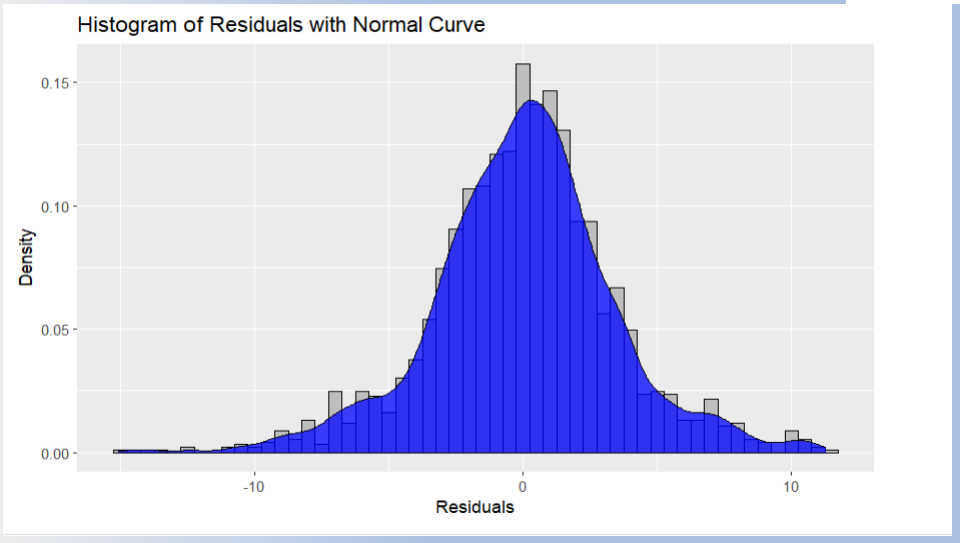
```r
significant_model <- lm(Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
                        percentage.expenditure + Hepatitis.B + under.five.deaths +
                        Diphtheria + HIV.AIDS + Income.composition.of.resources +
                        thinness.5.9.years + Schooling, data = selected_data1)

summary(significant_model)
```

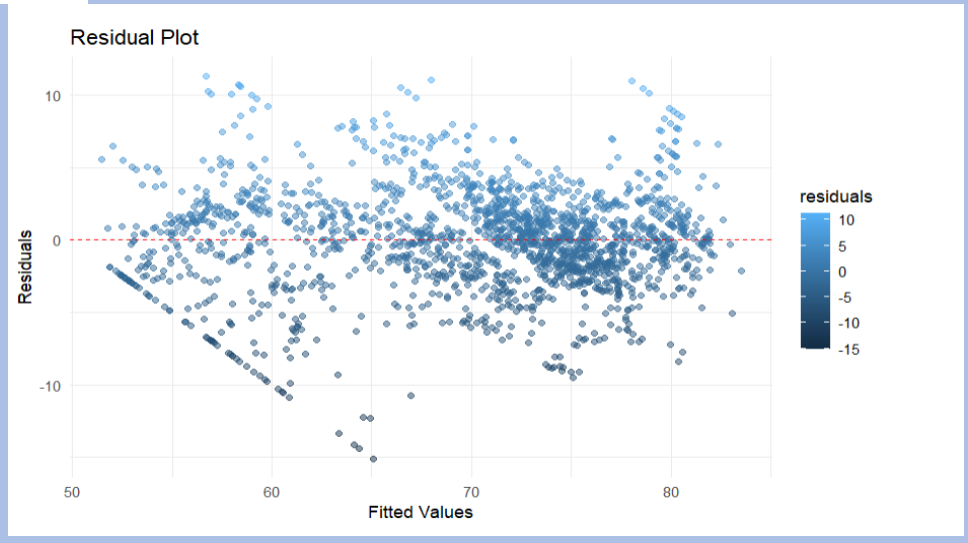
```
Call:
lm(formula = Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
    percentage.expenditure + Hepatitis.B + under.five.deaths +
    Diphtheria + HIV.AIDS + Income.composition.of.resources +
    thinness.5.9.years + Schooling, data = selected_data1)

Residuals:
     Min       1Q   Median       3Q      Max
-15.0782  -1.8938   0.0626   1.8866  11.2892

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    64.3279092  0.9595132  67.042  < 2e-16 ***
StatusDeveloping               -1.1436538  0.2716369  -4.210 2.67e-05 ***
Adult.Mortality                -0.0178572  0.0009557 -18.685  < 2e-16 ***
infant.deaths                   0.2084594  0.0540567   3.856 0.000119 ***
percentage.expenditure          0.0015092  0.0002109   7.157 1.19e-12 ***
Hepatitis.B                    -0.0191129  0.0069817  -2.738 0.006249 **
under.five.deaths              -0.2129692  0.0437080  -4.873 1.20e-06 ***
Diphtheria                      0.0574492  0.0108815   5.280 1.45e-07 ***
HIV.AIDS                       -6.6333603  0.2931874 -22.625  < 2e-16 ***
Income.composition.of.resources 10.0671978  0.8305172  12.122  < 2e-16 ***
thinness.5.9.years             -0.1735087  0.0262310  -6.615 4.87e-11 ***
Schooling                       0.2261686  0.0539281   4.194 2.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.441 on 1841 degrees of freedom
Multiple R-squared:  0.8329,    Adjusted R-squared:  0.8319
F-statistic: 834.1 on 11 and 1841 DF,  p-value: < 2.2e-16
```

Model Diagnostics

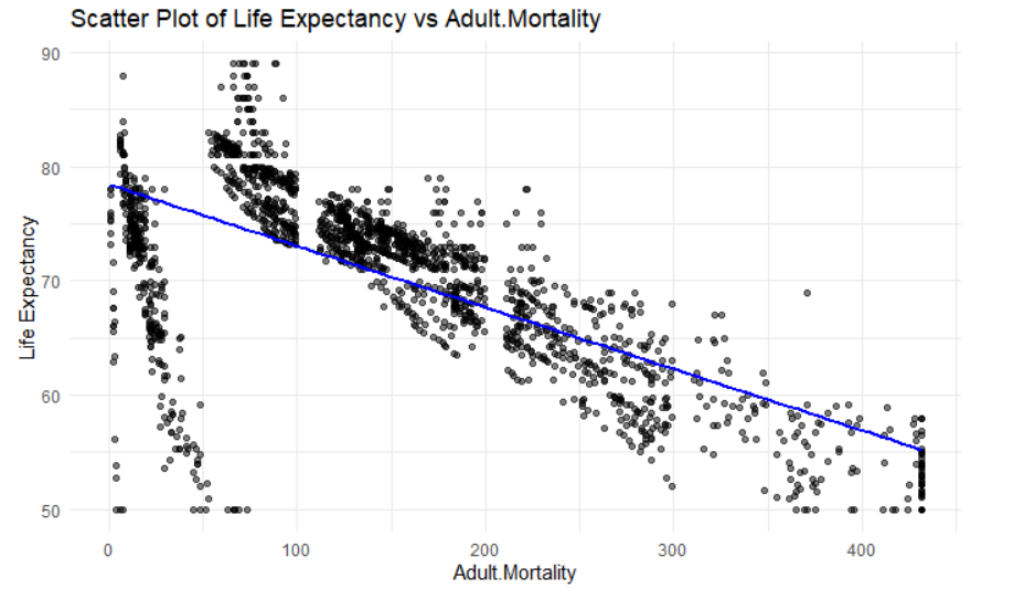
- We generated a Quantile-Quantile (Q-Q) plot of the residuals from the "significant_model" to assess if the residuals follow a normal distribution, which is an important assumption in linear regression models.
- The code creates a histogram and a density plot of the residuals from the "significant_model" to visualize their distribution, which helps in assessing the goodness of fit of the model.

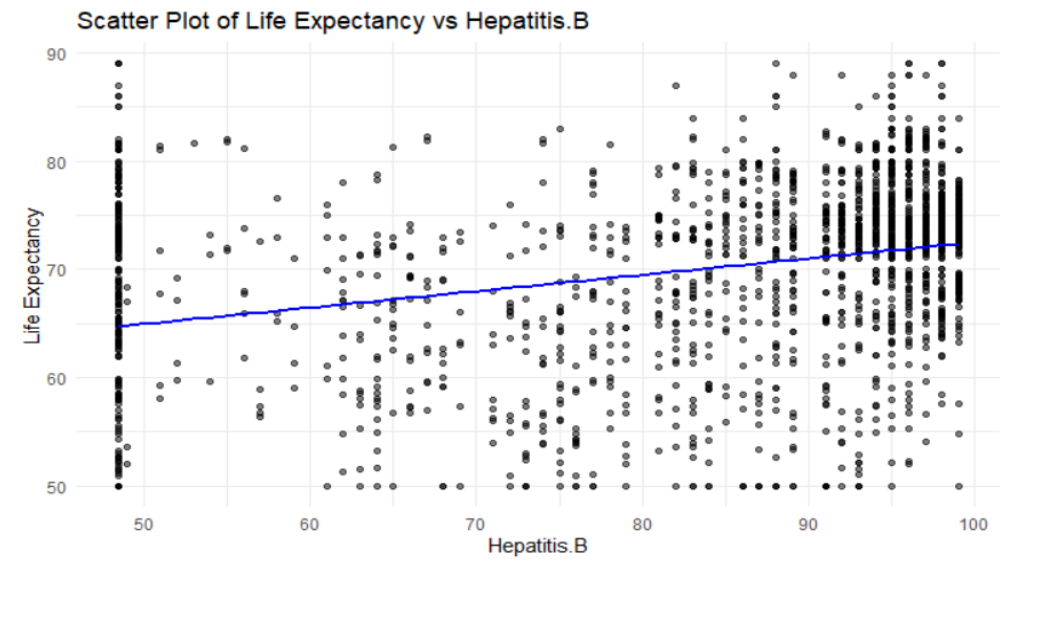


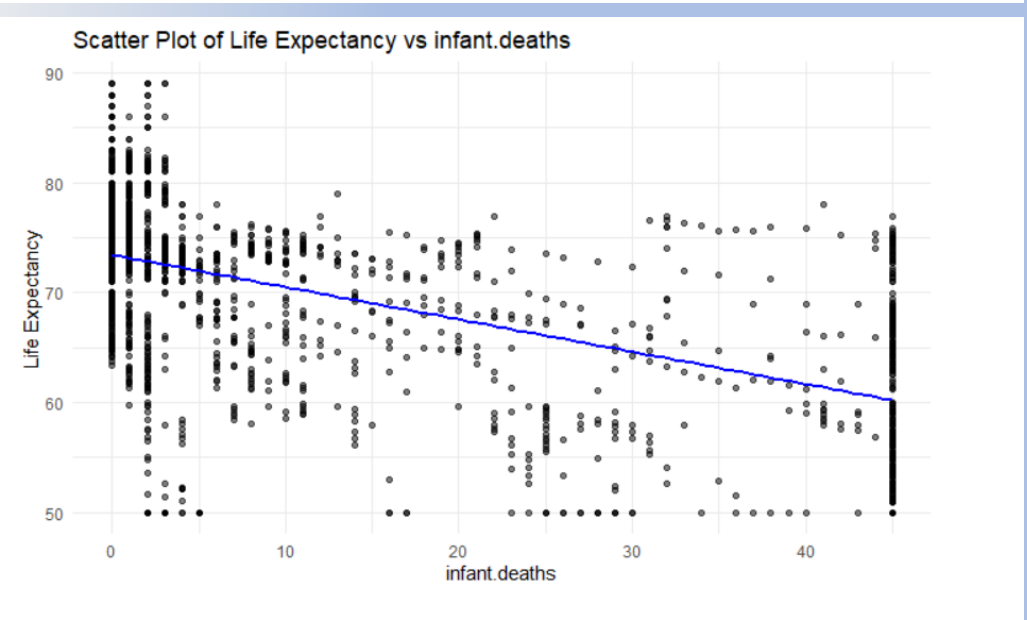


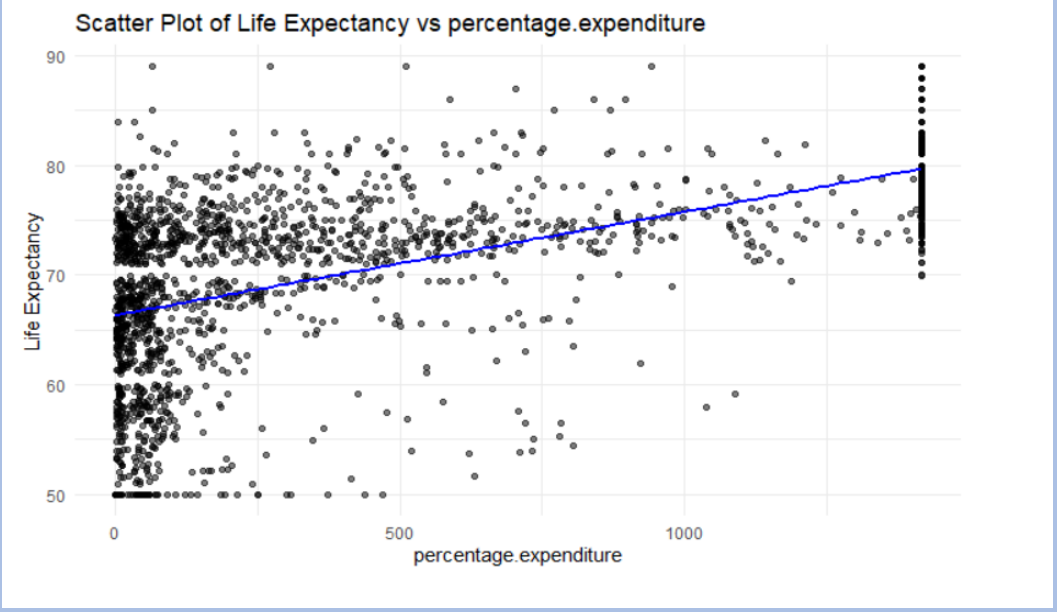


Data Visualization

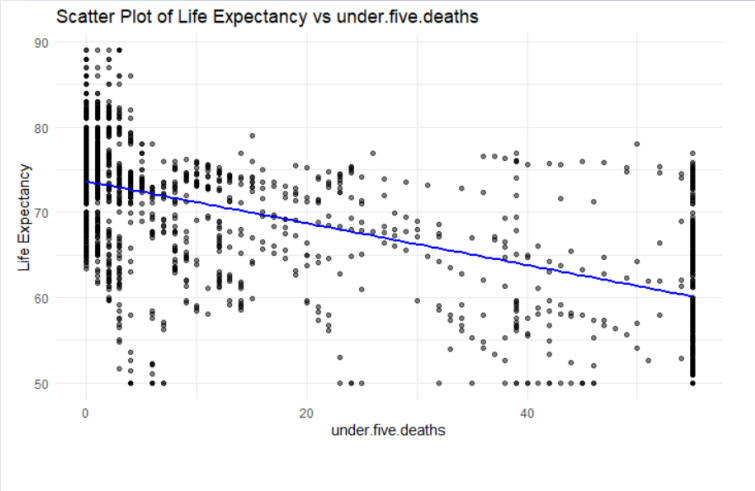


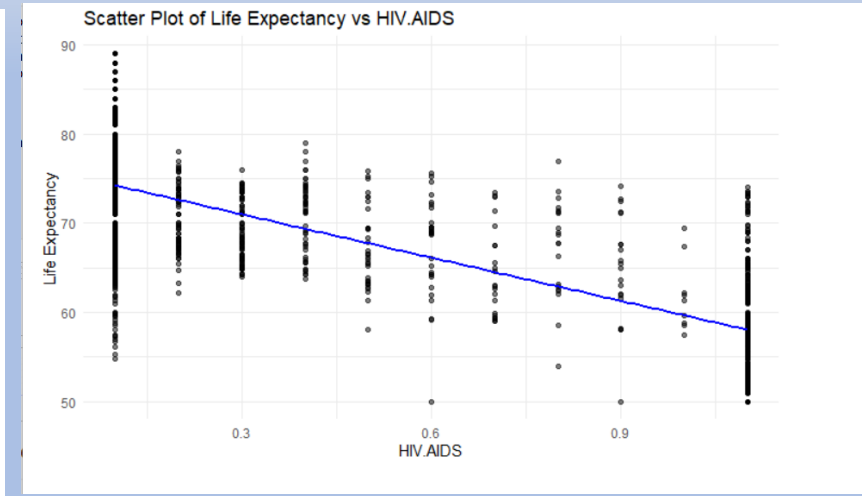


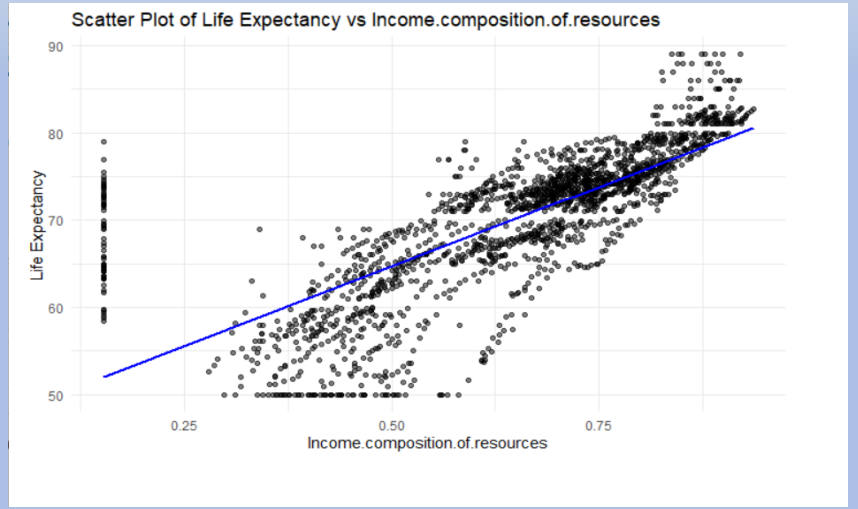


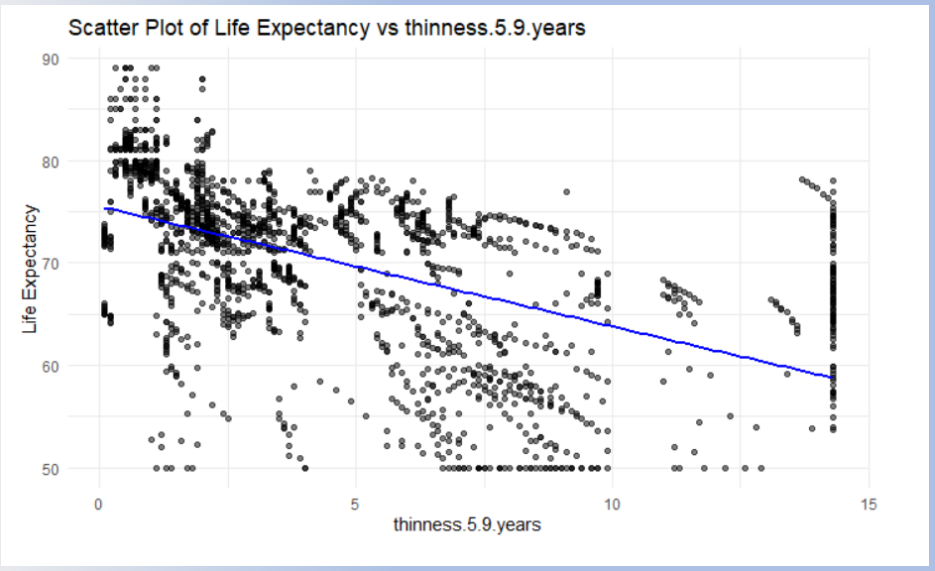


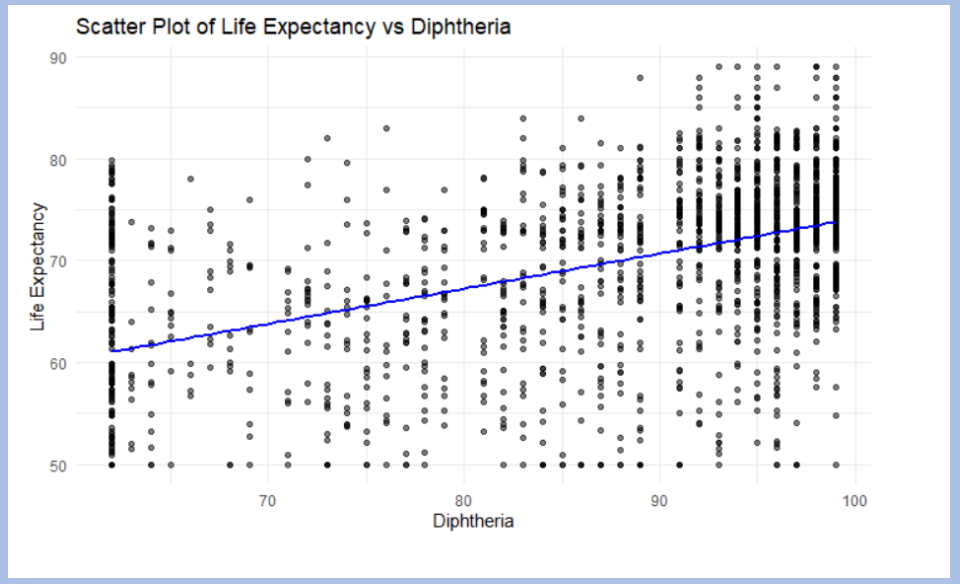
Data Visualization



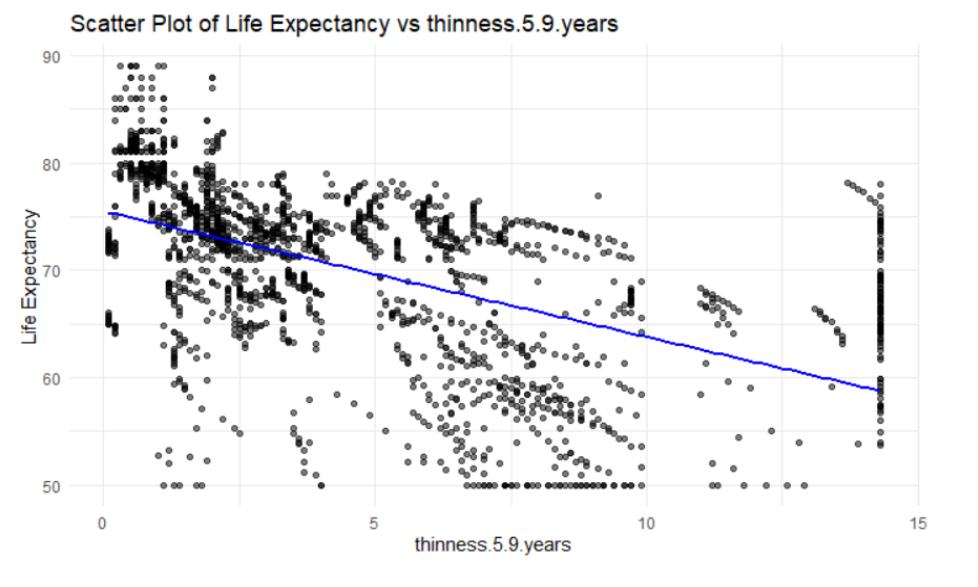


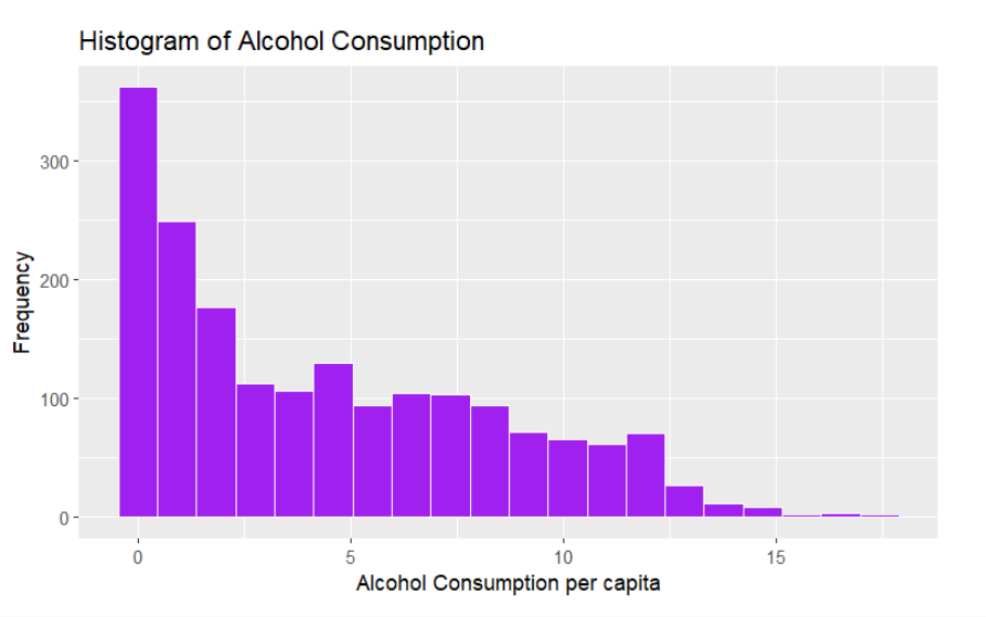


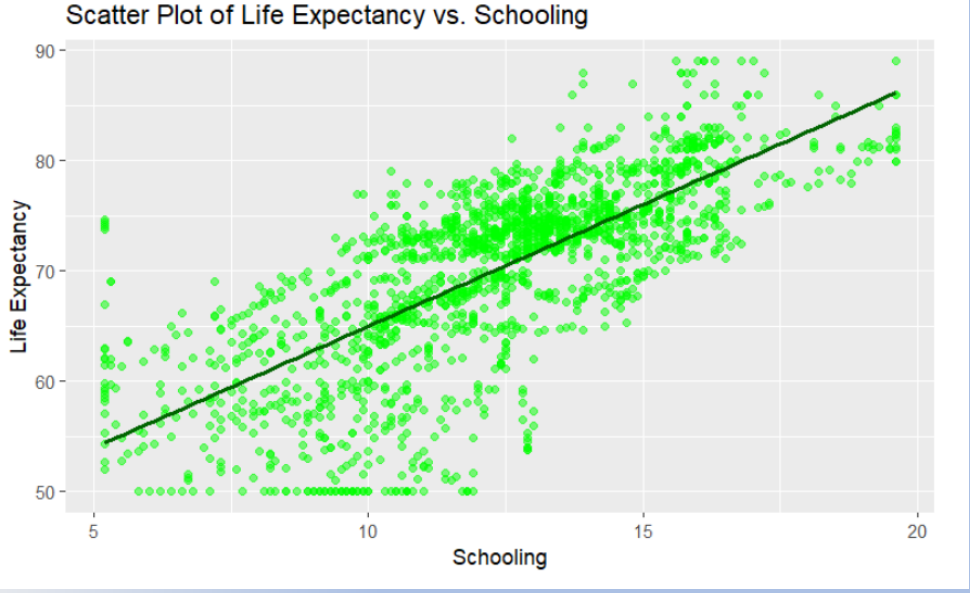


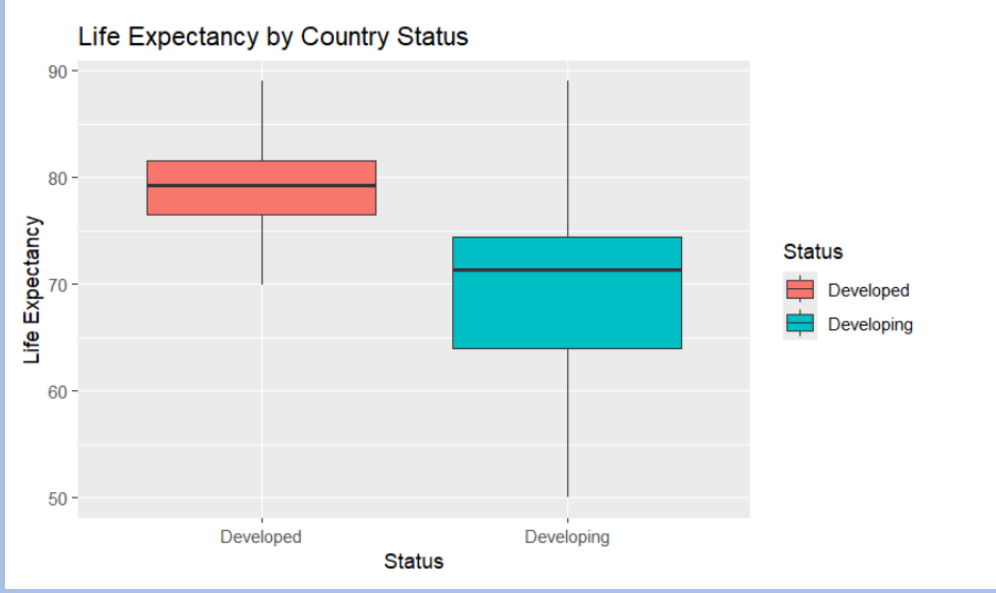


Data Visualization









CONCLUSION

- From the statistical analysis it can be concluded that the best model for determining life expectancy is a linear regression model with the status of development, adult mortality, infant deaths, health expenditure as a percentage of GDP, Hepatitis B coverage, under-five deaths, Diphtheria immunization coverage, HIV/AIDS prevalence, income composition of resources, thinness in children aged 5-9 years, and schooling as significant explanatory variables.

- For the first research question, small p-values indicate that the observed relationships between these variables and life expectancy are highly unlikely to have occurred by chance, the evidence strongly suggests that these health indicators and socio-economic factors are indeed significant determinants of life expectancy across different countries.

- For the second research question, the level of education, represented by the "Schooling" variable, has a significant positive impact on life expectancy, even after controlling for economic status and health indicators. The p-value of 2.87e-05 for the "Schooling" variable indicates strong evidence against the null hypothesis, suggesting that education plays a crucial role in determining life expectancy, beyond the effects of other factors

REFERENCE

Freeman, T., Gesesew, H. A., Bambra, C., Giugliani, E. R. J., Popay, J., Sanders, D., Macinko, J., Musolino, C., & Baum, F. (2020). Why do some countries do better or worse in life expectancy relative to income? An analysis of Brazil, Ethiopia, and the United States of America. *International Journal for Equity in Health*, *19*(1). https://doi.org/10.1186/s12939-020-01315-z

Rogoz, A. T. M., Sart, G., Bayar, Y., & Gavriletea, M. D. (2022). Impact of economic freedom and educational attainment on life expectancy: evidence from the new EU member states. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.907138

Singh, G. K., & Lee, H. (2020). Marked disparities in life expectancy by education, poverty level, occupation, and housing tenure in the United States, 1997-2014. *International Journal of MCH and AIDS*, *10*(1), 7–18. https://doi.org/10.21106/ijma.402

APPENDIX

```r
options(repos = c(CRAN = "https://cloud.r-project.org/"))
install.packages(c("dplyr", "ggplot2", "tidyr", "corrplot", "car", "lmtest", "broom", "GGally"))
```

```
Error in install.packages : Updating loaded packages
```

Importing Libraries

```r
library(dplyr)
library(ggplot2)
library(reshape2)
library(tidyr)
library(corrplot)
library(car)
library(lmtest)
library(MASS)
library(broom)
library(GGally)
```

1. Data Collection

Data Loading and Initial Exploration

```r
data <- read.csv("Life Expectancy Data.csv")
head(data)
str(data)
```

```r
dim(data)
```

```r
names(data)
```

2. Data Cleaning

2.1 Remove extra spaces in column names

```r
names(data) <- gsub("\\s+", "", names(data))
print(names(data))
```

2.2 Selecting Relevant Variables and Checking the structure of the selected data

```r
selected_data <- data[, !(names(data) %in% c("Country", "Year", "Population"))]
str(selected_data)
```

2.3 Handling Missing Values

```r
missing_values <- sapply(selected_data, function(x) sum(is.na(x)))
missing_values
```

#Removing Rows with Missing Values

```r
selected_data1= na.omit(selected_data)
selected_data1
```

#Verifying Missing Values After Removal

```r
missing_values <- sapply(selected_data1, function(x) sum(is.na(x)))
missing_values
```

3. Exploratory Data Analysis

#Data Summary

```r
summary(selected_data1)
```

3.1 Outlier Detection

```r
count_outliers <- function(x) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = TRUE)
  iqr <- IQR(x, na.rm = TRUE)
  lower <- qnt[1] - 1.5 * iqr
  upper <- qnt[2] + 1.5 * iqr
  return(sum(x < lower | x > upper, na.rm = TRUE))
}
outlier_counts <- sapply(selected_data1[, sapply(selected_data1, is.numeric)], count_outliers)
outlier_counts
```

Percentage of outliers

```r
# Calculate percentage of outliers for each column
outlier_percentages <- (outlier_counts / sapply(selected_data1[, sapply(selected_data1, is.numeric)], length)) * 100
outlier_percentages
# Calculate total percentage of outliers
total_outliers <- sum(outlier_counts)
total_non_missing_values <- sum(sapply(selected_data1[, sapply(selected_data1, is.numeric)], length))
total_outlier_percentage <- (total_outliers / total_non_missing_values) * 100
total_outlier_percentage
```

3.2 Visualizing Outliers

```r
num_cols <- length(names(selected_data1)[sapply(selected_data1, is.numeric)])
batches <- ceiling(num_cols / 9)  # Adjust the denominator to change batch size

for (i in 1:batches) {
  par(mfrow=c(3, 3), mar=c(2, 2, 2, 2))  # Adjust layout and margins as necessary
  start_col <- (i - 1) * 9 + 1
  end_col <- min(i * 9, num_cols)
  for (col in names(selected_data1)[sapply(selected_data1, is.numeric)][start_col:end_col]) {
    boxplot(selected_data1[[col]], main=col, col="lightblue", ylab="Values", xlab=col, cex.axis=0.7, cex.main=0.8)
  }
  par(mfrow=c(1, 1))
}
```

APPENDIX

```
# 3.3 Defining and Applying the Capping Function

```{r}
cap_outliers <- function(x) {
  qnt <- quantile(x, probs = c(0.25, 0.75), na.rm = TRUE)
  iqr <- IQR(x, na.rm = TRUE)
  lower <- qnt[1] - 1.5 * iqr
  upper <- qnt[2] + 1.5 * iqr
  x[x < lower] <- lower
  x[x > upper] <- upper
  return(x)
}


numeric_cols <- sapply(selected_data1, is.numeric)
selected_data1[, numeric_cols] <- lapply(selected_data1[, numeric_cols], cap_outliers)
selected_data1
```

```{r}
# Function to count outliers in a variable
count_outliers <- function(x) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = TRUE)
  iqr <- IQR(x, na.rm = TRUE)
  lower <- qnt[1] - 1.5 * iqr
  upper <- qnt[2] + 1.5 * iqr
  return(sum(x < lower | x > upper, na.rm = TRUE))
}

outlier_counts <- sapply(selected_data1[, sapply(selected_data1, is.numeric)], count_outliers)
outlier_counts
```
```

```
# 3.4 Verifying Outliers After Capping

```{r}
num_cols <- length(names(selected_data1)[sapply(selected_data1, is.numeric)])
batches <- ceiling(num_cols / 9)

for (i in 1:batches) {
  par(mfrow=c(3, 3), mar=c(2, 2, 2, 2))
  start_col <- (i - 1) * 9 + 1
  end_col <- min(i * 9, num_cols)
  for (col in names(selected_data1)[sapply(selected_data1, is.numeric)][start_col:end_col]) {
    boxplot(selected_data1[[col]], main=col, col="yellow", ylab="Values", xlab=col, cex.axis=0.7, cex.main=0.8)
  }
  par(mfrow=c(1, 1))
}
```

4 Checking for Normality

```{r}
all_cols <- names(selected_data1)
num_cols <- length(all_cols)
batches <- ceiling(num_cols / 9)  # Adjust the denominator to change batch size

for (i in 1:batches) {
  par(mfrow=c(3, 3), mar=c(2, 2, 2, 2))  # Adjust layout and margins as necessary
  start_col <- (i - 1) * 9 + 1
  end_col <- min(i * 9, num_cols)
  for (col in all_cols[start_col:end_col]) {
    if (is.numeric(selected_data1[[col]]) && !any(is.na(selected_data1[[col]]))) {
      qqnorm(selected_data1[[col]], main=col, col="blue")
      qqline(selected_data1[[col]], col="red")
    } else {
      plot.new()
    }
```

```
#Histogram for Alcohol Consumption

```{r}
# Histogram for Alcohol Consumption
ggplot(selected_data1, aes(x = Alcohol)) +
  geom_histogram(bins = 20, fill = "purple", color = "white") +
  ggtitle("Histogram of Alcohol Consumption") +
  xlab("Alcohol Consumption per capita") +
  ylab("Frequency")
```

5.2 Boxplots and Histograms for Distribution Analysis

#Boxplot for Life Expectancy by Status (Developed vs Developing)

```{r}
ggplot(selected_data1, aes(x = Status, y = Life.expectancy, fill = Status)) +
  geom_boxplot() +
  ggtitle("Life Expectancy by Country Status") +
  xlab("Status") +
  ylab("Life Expectancy")
```

#Density plot for Life Expectancy by Status

```{r}
ggplot(selected_data1, aes(x = Life.expectancy, fill = Status)) +
  geom_density(alpha = 0.5) +
  ggtitle("Density Plot of Life Expectancy by Status") +
  xlab("Life Expectancy") +
  ylab("Density")
```
```

```
#4.1 Data Transformation

```{r}
selected_data1_log <- selected_data1
selected_data1_log[, sapply(selected_data1_log, is.numeric)] <- lapply(selected_data1_log[, sapply(selected_data1_log, is.numeric)], log)
str(selected_data1_log)
```

5. Data Visualization

5.1 Exploring Distributions and Relationships

Scatter plot of Life Expectancy vs. GDP

```{r}
library(ggplot2)
ggplot(selected_data1, aes(x = GDP, y = Life.expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  ggtitle("Scatter Plot of Life Expectancy vs. GDP") +
  xlab("GDP") +
  ylab("Life Expectancy")
```

#Scatter plot of Life Expectancy vs. Schooling

```{r}
ggplot(selected_data1, aes(x = Schooling, y = Life.expectancy)) +
  geom_point(alpha = 0.5, color = "green") +
  geom_smooth(method = "lm", se = FALSE, color = "darkgreen") +
  ggtitle("Scatter Plot of Life Expectancy vs. Schooling") +
  xlab("Schooling") +
  ylab("Life Expectancy")
```
```

```
5.3 Advanced Relationship Analysis with Faceted and Colored Plots

#Faceted scatter plot for Life Expectancy vs. BMI, separated by Status

```{r}
ggplot(selected_data1, aes(x = BMI, y = Life.expectancy)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~Status) +
  ggtitle("Life Expectancy vs. BMI by Country Status") +
  xlab("BMI") +
  ylab("Life Expectancy") +
  theme_minimal()
```

#Scatter plot of Life Expectancy vs. Alcohol Consumption with regression line

```{r}
ggplot(selected_data1, aes(x = Alcohol, y = Life.expectancy)) +
  geom_point(aes(color = Status), alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Life Expectancy vs. Alcohol Consumption") +
  xlab("Alcohol Consumption (per capita)") +
  ylab("Life Expectancy") +
  theme_minimal()
```

5.4 Histogram for BMI

```{r}
ggplot(selected_data1, aes(x = BMI)) +
  geom_histogram(bins = 20, fill = "orange", color = "white") +
  ggtitle("Histogram of BMI") +
  xlab("BMI") +
  ylab("Frequency")
```
```
```

APPENDIX

```
# 5.5 Correlation Heatmap

```{r}
# Correlation matrix
cor_matrix <- cor(selected_data1[, sapply(selected_data1, is.numeric)], use = "complete.obs")

library(reshape2)
melted_cor_matrix <- melt(cor_matrix)

# Heatmap with correlation coefficients
library(ggplot2)
ggplot(melted_cor_matrix, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = sprintf("%.2f", value)), color = "black", size = 3) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_text(angle = 0, hjust = 1)) +
  ggtitle("Correlation Heatmap")

```
```

```
# 6.Statistical Modeling

# 6.1 Model Selection

```{r}
numeric_cols <- names(selected_data1)[sapply(selected_data1, is.numeric)]

numeric_cols <- numeric_cols[numeric_cols != "Life.expectancy"]
num_cols <- length(numeric_cols)
batches <- ceiling(num_cols / 9)

for (i in 1:batches) {
  par(mfrow = c(3, 3), mar = c(2, 2, 2, 2))
  start_col <- (i - 1) * 9 + 1
  end_col <- min(i * 9, num_cols)

  for (col in numeric_cols[start_col:end_col]) {
    plot(selected_data1[[col]], selected_data1$Life.expectancy,
         main = paste( col, "vs. Life Expectancy"),
         xlab = col,
         ylab = "Life Expectancy",
         pch = 16,  # Set point character
         col = "blue")  # Set point color

    abline(lm(Life.expectancy ~ ., data = selected_data1[, c(col, "Life.expectancy")]),
           col = "red", lwd = 2)
  }

  par(mfrow = c(1, 1))
}
```
```

```
# 6.2 Building Linear multiple regression model

```{r}
full_model <- lm(Life.expectancy ~ ., data = selected_data1)
summary(full_model)

```

6.3 Model Diagnostics

```{r}
# Fit the model
full_model <- lm(Life.expectancy ~ ., data = selected_data1)

# Q-Q plot of residuals to check normality
qqnorm(full_model$residuals, main="Q-Q Plot of Residuals")
qqline(full_model$residuals, col="red")

```

6.4 Model Refinement

#Building Linear multiple regression model with only significant predictors

```{r}
significant_model <- lm(Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
                        percentage.expenditure + Hepatitis.B + under.five.deaths +
                        Diphtheria + HIV.AIDS + Income.composition.of.resources +
                        thinness.5.9.years + Schooling, data = selected_data1)

summary(significant_model)

```
```

```
```{r}
# Q-Q plot of residuals
qqnorm(residuals(significant_model))
qqline(residuals(significant_model))
```

```{r}
data <- selected_data1
significant_predictors <- c("Status", "Adult.Mortality", "infant.deaths",
                            "percentage.expenditure", "Hepatitis.B", "under.five.deaths",
                            "Diphtheria", "HIV.AIDS", "Income.composition.of.resources",
                            "thinness.5.9.years", "Schooling")
for (predictor in significant_predictors) {
  p <- ggplot(data, aes_string(x = predictor, y = "Life.expectancy")) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    labs(title = paste("Scatter Plot of Life Expectancy vs", predictor),
         x = predictor,
         y = "Life Expectancy") +
    theme_minimal()

  print(p) # Display the plot
}

```
```

```
#6.5 Checking the Assumptions of the Significant Model

#Histogram of Residuals

```{r}
residuals_data <- data.frame(residuals = residuals(significant_model))
ggplot(residuals_data, aes(x = residuals)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5, color = "black", fill = "gray") +
  geom_density(alpha = 0.75, fill = "blue") +
  ggtitle("Histogram of Residuals with Normal Curve") +
  xlab("Residuals") +
  ylab("Density")
```

#Residual Plot

```{r}
ggplot(residuals_data, aes(x = fitted(significant_model), y = residuals)) +
  geom_point(aes(color = residuals), alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residual Plot", x = "Fitted Values", y = "Residuals") +
  theme_minimal()
```
```