Life Expectancy Prediction

Vipul Gharde, Chaitanya Bachhav

2022-12-05

Abstract

We have implemented a Linear Regression model to predict the life expectancy of the human population with our best model having an adjusted R-squared value of 0.8239, RMSE of 3.71, and MAE of 2.85, from the processed dataset having ~1650 observations of ~20 variables related to life expectancy and health factors for 193 countries provided by the Global Health Observatory (GHO) data repository under the World Health Organization (WHO). Several model building techniques including Forward Selection, Backward Elimination, and Stepwise Regression were used to obtain the candidate models, which were then evaluated with K-Fold Cross Validation to yield the model with the lowest RMSE value. Our best model passes the normality assumption and has no issues with the multicollinearity of the variables.

Introduction

Life expectancy is an estimate of the expected average number of years of life (or a person's age at death) for individuals who were born into a particular population. It is one of the most used summary indicators for the overall health of a population. Its levels and trends direct health policies, and researchers try to identify the determining risk factors to assess and forecast future developments. (Luy et al. 2020)

The goal of this project is to build a Linear Regression model that can predict the life expectancy of the human population based on several factors such as the average Body Mass Index (BMI), the Gross Domestic Product (GDP) of a country, the amount of alcohol consumption in a country, immunization of various vaccines among 1-year-olds such as Hepatitis B, Polio, and Diphtheria vaccines, and more, and also derive insights into what factors are significant in determining a higher or lower life expectancy of the human population.

Materials and Methods

Software and Packages Used

We have used R with the RStudio Integrated Development Environment for our analysis and for building the Linear Regression models. We have also used the R packages `corrplot`, `ggplot`, `car`, `olsrr`, and `caret` to aid in our analysis and model building.

Dataset

The data related to life expectancy and health factors for 193 countries is taken from the Global Health Observatory (GHO) data repository under the World Health Organization (WHO). Its corresponding economic data was collected from the United Nations website for a period of 16 years (2000-2015).

The dataset is available at https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who.

`Life Expectancy Data.csv` contains the following fields:

- `Country` - Country Observed.
- `Year` - Year Observed.
- `Status` - Developed or Developing status.
- `Life.expectancy` - Life Expectancy in age.
- `Adult.Mortality` - Adult Mortality Rates on both sexes (probability of dying between 15-60 years/1000 population).
- `infant.deaths` - Number of Infant Deaths per 1000 population.
- `Alcohol` - Alcohol recorded per capita (15+) consumption (in litres of pure alcohol).
- `percentage.expenditure` - Expenditure on health as a percentage of Gross Domestic Product per capita (%).
- `Hepatitis.B` - Hepatitis B (HepB) immunization coverage among 1-year-olds (%).
- `Measles` - Number of reported Measles cases per 1000 population.
- `BMI` - Average Body Mass Index of entire population.
- `under.five.deaths` - Number of under-five deaths per 1000 population.
- `Polio` - Polio (Pol3) immunization coverage among 1-year-olds (%).
- `Total.expenditure` - General government expenditure on health as a percentage of total government expenditure (%).
- `Diphtheria` - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).
- `HIV.AIDS` - Deaths per 1000 live births due to HIV/AIDS (0-4 years).
- `GDP` - Gross Domestic Product per capita (in USD).
- `Population` - Population of the country.
- `thinness..1.19.years` - Prevalence of thinness among children and adolescents for Age 10 to 19 (%).
- `thinness.5.9.years` - Prevalence of thinness among children for Age 5 to 9 (%).
- `Income.composition.of.resources` - Human Development Index in terms of income composition of resources (index ranging from 0 to 1).
- `Schooling` - Number of years of Schooling (years).

In total, there are 2938 observations of 22 variables with 20 of them being numerical and 2 categorical (`Country` and `Status`).

We are using `Life.expectancy` to predict the life expectancy of the human population with the given independent variables in the dataset.

For data cleaning, we have dropped any observation that does not contain any value in any of its columns. This shrinks our dataset to 1649 observations.

We have plotted a boxplot and a histogram for all the numerical variables in the dataset. For categorical variables, we have plotted a barplot indicating the counts of each category of the variable. This can be viewed in Appendix A of the report.

Feature Selection

We have removed some of the variables for building the model due to the reasons mentioned below:

`Country` - Contains too many levels with no additional information to predict `Life.expectancy`.
`Year` - Contains time series data with no additional information to predict `Life.expectancy`.

We have also mutated `Hepatitis.B`, `Polio`, and `Diphtheria` for building the model since the range between their minimum values and their 1st Quartiles are too wide. We have mutated their values into 2 categorical values: '<90% Covered' and '>=90% Covered'.

This leaves us with 1649 observations of 20 variables with 16 of them being numerical and 4 categorical (`Status`, `Hepatitis.B`, `Polio`, and `Diphtheria`).

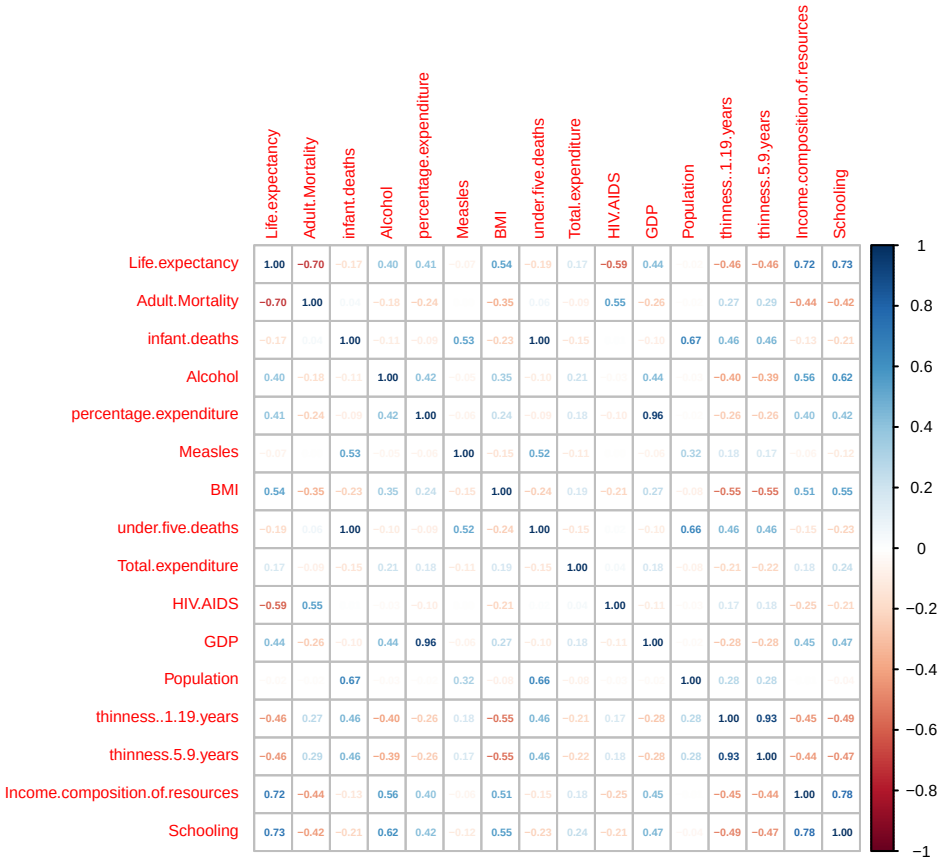
```
summary(life)
```

```
##     Status          Life.expectancy Adult.Mortality infant.deaths
##  Length:1649        Min.   :44.0    Min.   :  1.0   Min.   :   0.00
##  Class :character   1st Qu.:64.4    1st Qu.: 77.0   1st Qu.:   1.00
##  Mode  :character   Median :71.7    Median :148.0   Median :   3.00
##                     Mean   :69.3    Mean   :168.2   Mean   :  32.55
##                     3rd Qu.:75.0    3rd Qu.:227.0   3rd Qu.:  22.00
##                     Max.   :89.0    Max.   :723.0   Max.   :1600.00
##     Alcohol        percentage.expenditure Hepatitis.B          Measles
##  Min.   : 0.010   Min.   :    0.00       Length:1649        Min.   :     0
##  1st Qu.: 0.810   1st Qu.:   37.44       Class :character   1st Qu.:     0
##  Median : 3.790   Median :  145.10       Mode  :character   Median :    15
##  Mean   : 4.533   Mean   :  698.97                          Mean   :  2224
##  3rd Qu.: 7.340   3rd Qu.:  509.39                          3rd Qu.:   373
##  Max.   :17.870   Max.   :18961.35                          Max.   :131441
##       BMI        under.five.deaths    Polio           Total.expenditure
##  Min.   : 2.00   Min.   :   0.00   Length:1649        Min.   : 0.740
##  1st Qu.:19.50   1st Qu.:   1.00   Class :character   1st Qu.: 4.410
##  Median :43.70   Median :   4.00   Mode  :character   Median : 5.840
##  Mean   :38.13   Mean   :  44.22                      Mean   : 5.956
##  3rd Qu.:55.80   3rd Qu.:  29.00                      3rd Qu.: 7.470
##  Max.   :77.10   Max.   :2100.00                      Max.   :14.390
##   Diphtheria          HIV.AIDS           GDP              Population
##  Length:1649        Min.   : 0.100   Min.   :     1.68   Min.   :3.400e+01
##  Class :character   1st Qu.: 0.100   1st Qu.:   462.15   1st Qu.:1.919e+05
##  Mode  :character   Median : 0.100   Median :  1592.57   Median :1.420e+06
##                     Mean   : 1.984   Mean   :  5566.03   Mean   :1.465e+07
##                     3rd Qu.: 0.700   3rd Qu.:  4718.51   3rd Qu.:7.659e+06
##                     Max.   :50.600   Max.   :119172.74   Max.   :1.294e+09
##  thinness..1.19.years thinness.5.9.years Income.composition.of.resources
##  Min.   : 0.100       Min.   : 0.100     Min.   :0.0000
##  1st Qu.: 1.600       1st Qu.: 1.700     1st Qu.:0.5090
##  Median : 3.000       Median : 3.200     Median :0.6730
##  Mean   : 4.851       Mean   : 4.908     Mean   :0.6316
##  3rd Qu.: 7.100       3rd Qu.: 7.100     3rd Qu.:0.7510
##  Max.   :27.200       Max.   :28.200     Max.   :0.9360
##    Schooling
##  Min.   : 4.20
##  1st Qu.:10.30
##  Median :12.30
##  Mean   :12.12
##  3rd Qu.:14.00
##  Max.   :20.70
```

Correlations

We would want to look at the correlation matrix to see what variables are correlated with the target variable, and also to check if any independent variable is also correlated with another independent variable. Correlated independent variables in a model can negatively impact its performance, and so if any such pair of independent correlated variables is found, we would keep only one of them in the model.

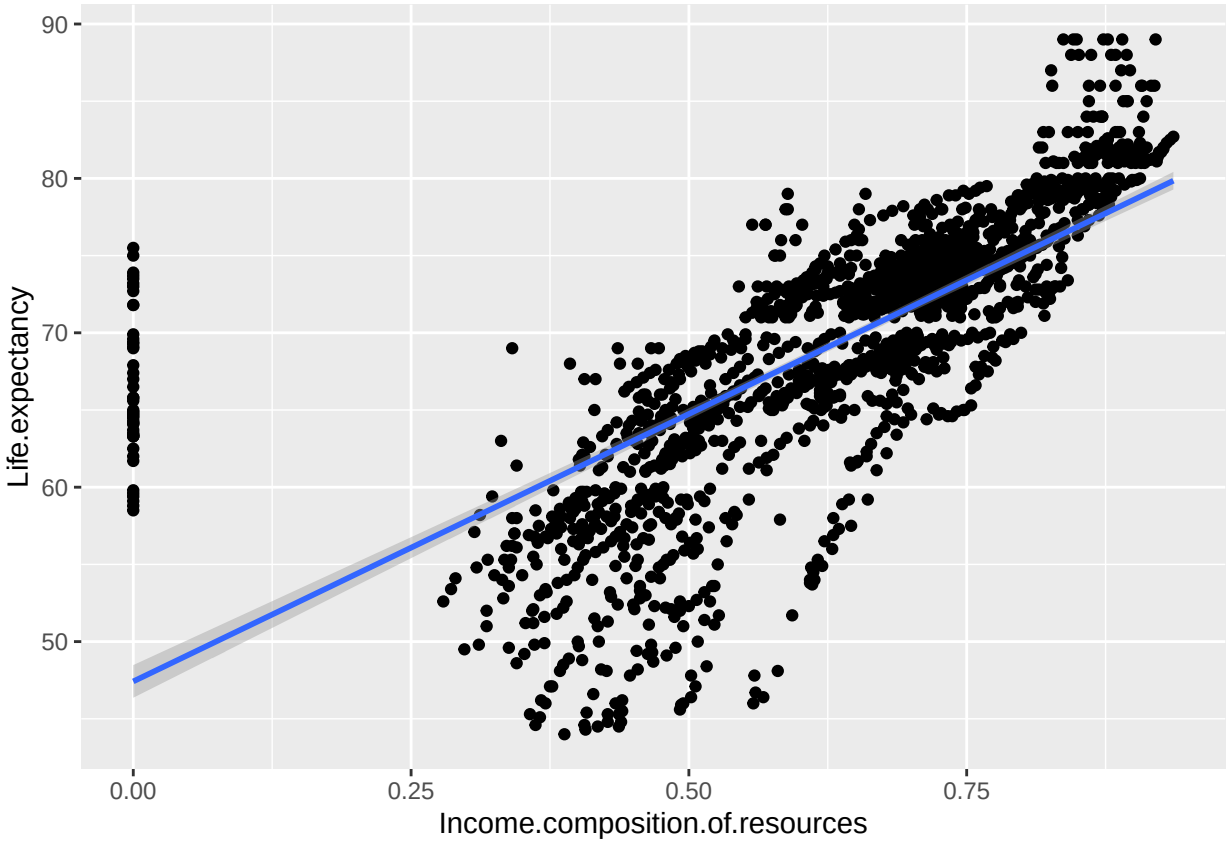
Since the number of variables is moderately large, we have plotted the correlation plot of the dataset rather than printing the correlation matrix by itself. The colors and their shades easily guide us to show what 2 variables are correlated.

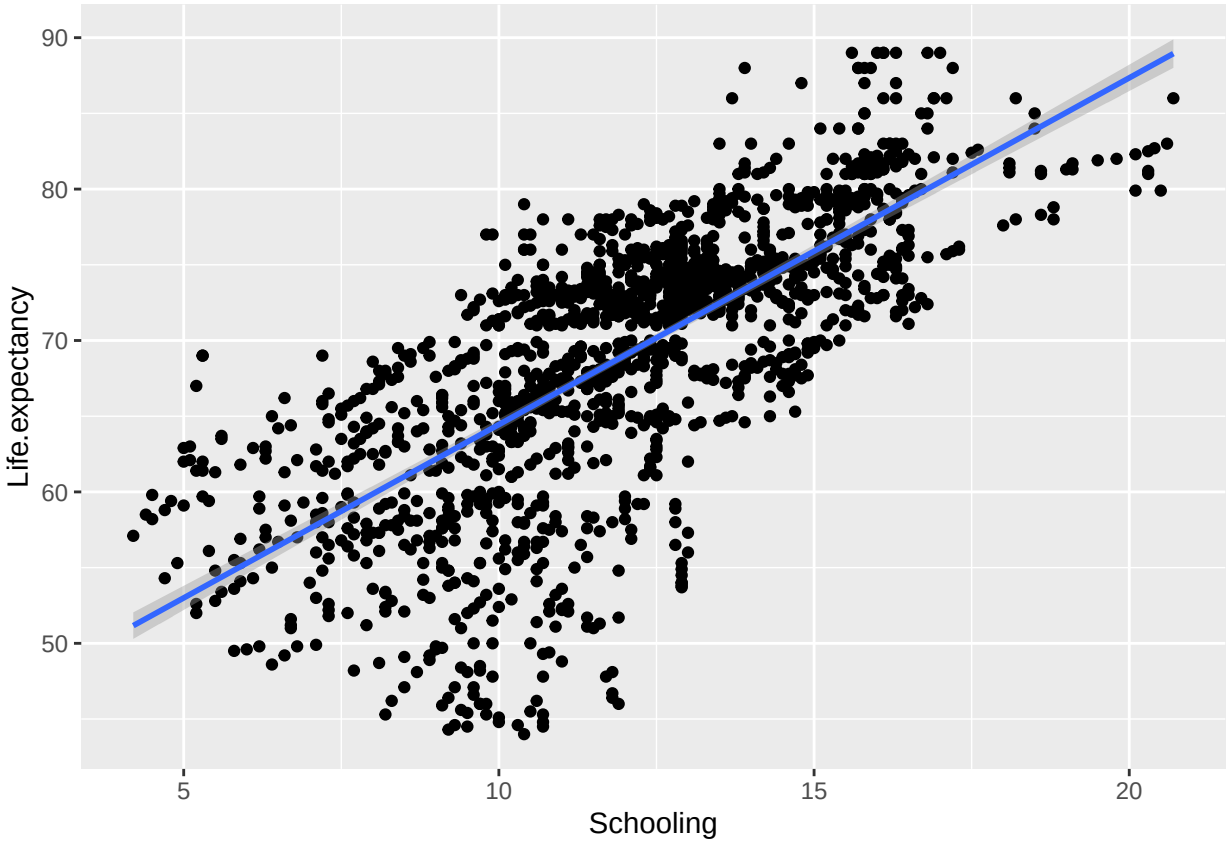


There are a few takeaways from this correlation plot:

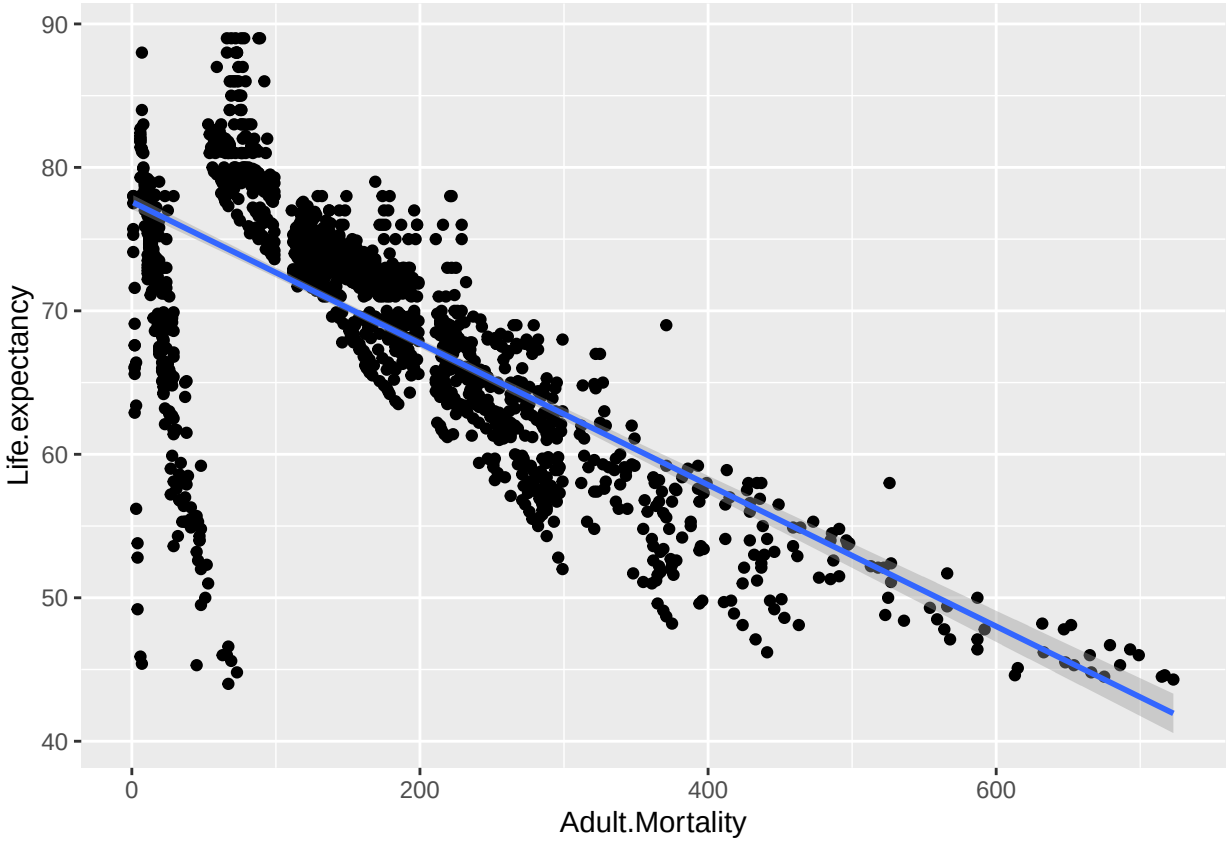
- `Life.expectancy` has a strong positive correlation with `Income.composition.of.resources` and `Schooling`.
- `Life.expectancy` has a negative correlation with `Adult.Mortality`, which makes sense since if the mortality rate of adult is high, then obviously the life expectancy will be low.
- `Life.expectancy` has a very weak correlation with `Measles` and `Population`.
- There is a very strong correlation between `infant.deaths` and `under.five.deaths`, `percentage.expenditure` and `GDP`, and `thinness..1.19.years` and `thinness.5.9.years`, indicating multicollinearity between them. Therefore, we have removed `under.five.deaths`, `percentage.expenditure`, and `thinness.5.9.years` for building the model.

It is evident from the scatterplot of `Life.expectancy` against `Income.composition.of.resources` and `Life.expectancy` against `Schooling` that there is a positive trend in the life expectancy of the human population with the increase in the values of these independent variables.

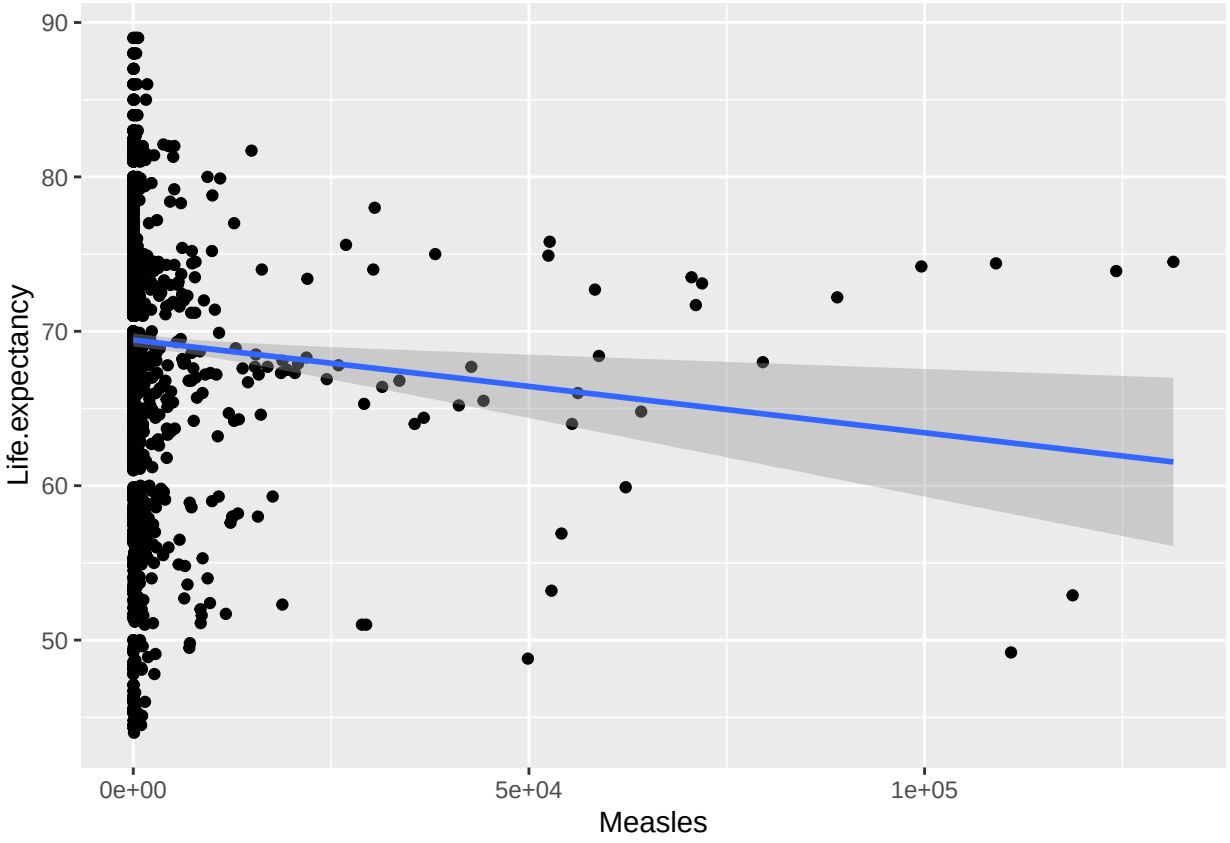


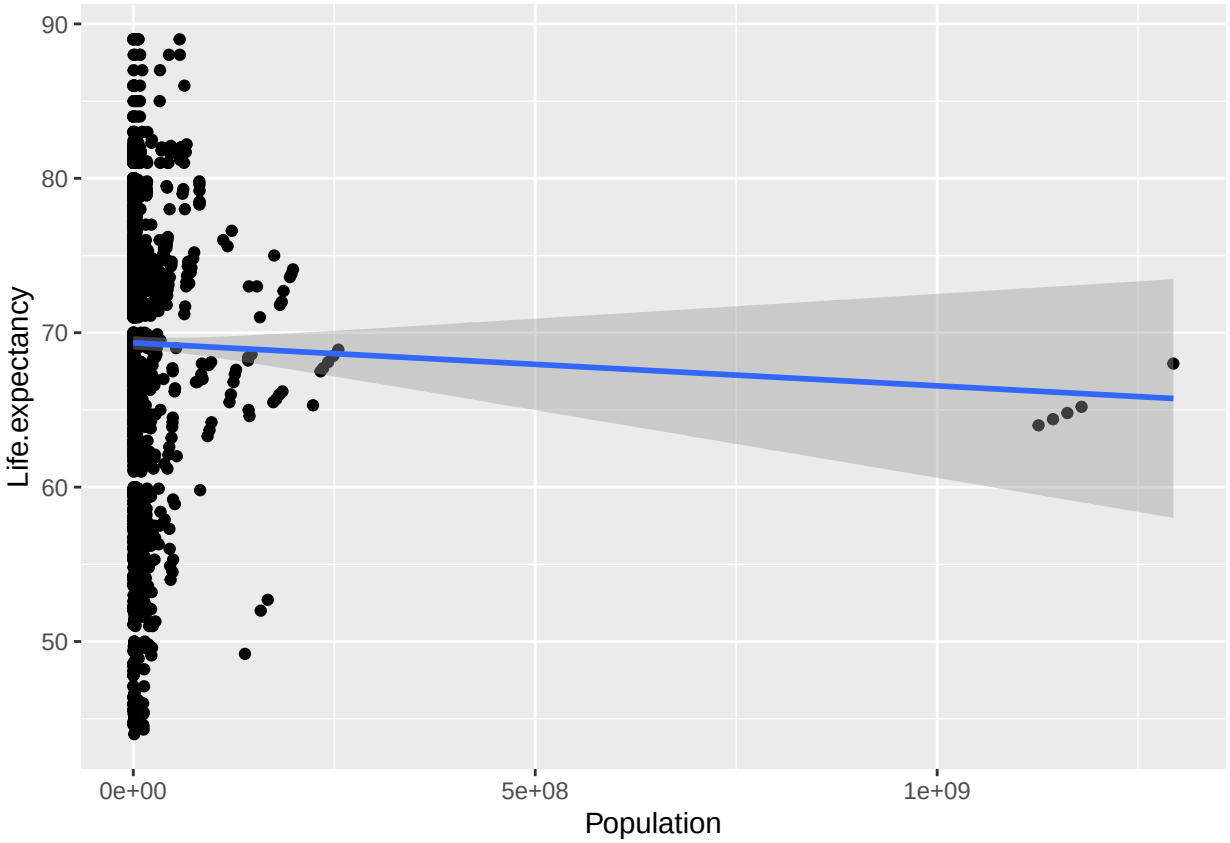

Life.expectancy
90
80
70
60
50
5
10
15
20
Schooling

It is evident from the scatterplot of `Life.expectancy` against `Adult.Mortality` that there is a negative trend in the life expectancy of the human population with the increase in the value of this independent variable.



It seems in the scatterplot of `Life.expectancy` against `Measles` and `Life.expectancy` against `Population` that there is a negative trend in the life expectancy of the human population with the increase in the values of these independent variables, but since the bulk of the data falls on the lower range of their values, there exist some high leverage and high influence points that appear to drive the regression line downward toward the negative. This is clearly evident in the scatterplot of `Life.expectancy` against `Population`.




90
80
70
60
50
Life.expectancy
0e+00
5e+08
1e+09
Population

Model Building

We now build a Linear Regression Model using all the remaining variables to predict the life expectancy of the human population. We will set the level of α to be 0.05 throughout the analysis.

```
lmod_all = lm(Life.expectancy ~ ., data = life)
summary(lmod_all)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0274  -2.1069   0.0579   2.3922  11.5489
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.506e+01  8.102e-01  67.965  < 2e-16 ***
## StatusDeveloping                 -9.930e-01  3.464e-01  -2.867   0.0042 **
## Adult.Mortality                  -1.785e-02  9.663e-04 -18.477  < 2e-16 ***
## infant.deaths                    -3.057e-03  1.260e-03  -2.425   0.0154 *
## Alcohol                          -1.539e-01  3.381e-02  -4.552 5.72e-06 ***
## Hepatitis.B>=90% Covered         -6.990e-01  3.174e-01  -2.202   0.0278 *
## Measles                           1.668e-05  1.078e-05   1.547   0.1220
## BMI                               3.591e-02  6.103e-03   5.884 4.85e-09 ***
## Polio>=90% Covered                5.337e-01  4.437e-01   1.203   0.2293
## Total.expenditure                 7.441e-02  4.170e-02   1.784   0.0745 .
## Diphtheria>=90% Covered           9.665e-01  4.888e-01   1.977   0.0482 *
## HIV.AIDS                         -4.278e-01  1.850e-02 -23.124  < 2e-16 ***
## GDP                               6.096e-05  9.637e-06   6.326 3.24e-10 ***
## Population                        2.558e-09  1.766e-09   1.449   0.1476
## thinness..1.19.years             -4.799e-02  2.791e-02  -1.720   0.0857 .
## Income.composition.of.resources   1.041e+01  8.503e-01  12.244  < 2e-16 ***
## Schooling                         8.790e-01  6.158e-02  14.274  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.688 on 1632 degrees of freedom
## Multiple R-squared:  0.826,  Adjusted R-squared:  0.8243
## F-statistic: 484.1 on 16 and 1632 DF,  p-value: < 2.2e-16
```

There are a few takeaways from this full model:

- The p-value of the model is 2.2e-16 < 0.05, indicating that it is significant.
- The adjusted R-squared value of the model is 0.8243, indicating that about 82.43% of the variability observed in the target variable can be explained by the independent variables in the model, which is quite a good result and can possibly be improved even further with model selection.
- `Adult.Mortality`, `Alcohol`, `BMI`, `HIV.AIDS`, `GDP`, `Income.composition.of.resources` and `Schooling` are the most significant variables with p-value < 0.05.
- From the model we can interpret that `Income.composition.of.resources` has a strong positive effect on life expectancy.
- From the model we can interpret that `StatusDeveloping`, `Adult.Mortality`, `infant.deaths`, `Alcohol`, `HIV.AIDS`, and `thinness..1.19.years` may have a negative effect on life expectancy.
- A peculiar result we can interpret from the model is that `Hepatitis.B>=90% Covered` may also have a negative effect on life expectancy.

We now generate models by using different techniques like Forward Selection Method, Backward Elimination Method and Stepwise Regression Method.

Forward Selection Method

The Forward Selection method involves building a model starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.

For building the model using Forward Selection method, we have used the default p-to-enter value of 0.3.

```
ols_step_forward_p(lmod_all)
```

```
##
##                                       Selection Summary
## ------------------------------------------------------------------------------------------------------------
##         Variable                                        Adj.
## Step                Entered                 R-Square    R-Square      C(p)          AIC         RMSE
## ------------------------------------------------------------------------------------------------------------
##    1    Schooling                                0.5294      0.5292    2767.6453    10612.7157    6.0362
##    2    HIV.AIDS                                 0.7304      0.7301     885.2760     9696.3271    4.5704
##    3    Adult.Mortality                          0.7871      0.7867     355.5223     9308.9473    4.0627
##    4    Income.composition.of.resources          0.8092      0.8087     150.4656     9130.3986    3.8474
##    5    BMI                                      0.8144      0.8138     103.6254     9086.7666    3.7957
##    6    GDP                                      0.8197      0.8190      55.8305     9040.9050    3.7422
##    7    Diphtheria                               0.8214      0.8206      42.0679     9027.4602    3.7258
##    8    Alcohol                                  0.8226      0.8217      32.8339     9018.3638    3.7144
##    9    thinness..1.19.years                     0.8235      0.8225      26.3197     9011.9039    3.7061
##   10    Status                                   0.8244      0.8233      19.8462     9005.4410    3.6977
##   11    Hepatitis.B                              0.8248      0.8236      17.7935     9003.3780    3.6943
##   12    Total.expenditure                        0.8251      0.8239      16.7491     9002.3193    3.6920
##   13    infant.deaths                            0.8254      0.8240      16.6117     9002.1684    3.6907
##   14    Measles                                  0.8256      0.8241      16.5671     9002.1083    3.6895
##   15    Population                               0.8258      0.8242      16.4465     9001.9689    3.6882
##   16    Polio                                    0.8260      0.8243      17.0000     9002.5080    3.6877
## ------------------------------------------------------------------------------------------------------------
```

The Forward Selection method included all the variables in the model.

Backward Elimination Method

The Backward Elimination method involves building a model starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.

For building the model using Backward Elimination method, we have used the default p-to-remove value of 0.3.

```
ols_step_backward_p(lmod_all)
```

```
## [1] "No variables have been removed from the model."
```

The Backward Elimination method did not eliminate any variables from the model.

Stepwise Regression Method

The Stepwise Regression method is a combination of the above two methods, starting with no variables in the model and testing at each step for variables to be included or excluded.

For building the model using Stepwise Regression method, we have used the default p-to-enter value of 0.1 and the default p-to-remove value of 0.3.

```
ols_step_both_p(lmod_all)
```

```
##
##                                       Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------
##                                        Added/                   Adj.
## Step               Variable           Removed     R-Square    R-Square      C(p)          AIC
## ----------------------------------------------------------------------------------------------
##    1               Schooling          addition       0.529       0.529    2767.6450    10612.715
##    2               HIV.AIDS           addition       0.730       0.730     885.2760     9696.327
##    3           Adult.Mortality        addition       0.787       0.787     355.5220     9308.947
##    4    Income.composition.of.resources addition     0.809       0.809     150.4660     9130.398
##    5                   BMI            addition       0.814       0.814     103.6250     9086.760
##    6                   GDP            addition       0.820       0.819      55.8310     9040.905
##    7             Diphtheria           addition       0.821       0.821      42.0680     9027.460
##    8               Alcohol            addition       0.823       0.822      32.8340     9018.363
##    9         thinness..1.19.years     addition       0.823       0.823      26.3200     9011.903
##   10                Status            addition       0.824       0.823      19.8460     9005.441
##   11             Hepatitis.B          addition       0.825       0.824      17.7930     9003.378
##   12          Total.expenditure       addition       0.825       0.824      16.7490     9002.319
## ----------------------------------------------------------------------------------------------
```

A total of 12 variables were included in the model built using Stepwise Regression method.

In summary, the variables chosen by the methods are indicated in the following table (x denotes the variable was chosen by the corresponding method):

Model Selection Method	Status	Adult.Mortality	infant.deaths	Alcohol
Forward Selection	x	x	x	x
Backward Elimination	x	x	x	x
Stepwise Regression	x	x		x

Model Selection Method	Hepatitis.B	Measles	BMI	Polio	Total.expenditure
Forward Selection	x	x	x	x	x
Backward Elimination	x	x	x	x	x
Stepwise Regression	x		x		x

Model Selection Method	Diphtheria	HIV.AIDS	GDP	Population
Forward Selection	x	x	x	x
Backward Elimination	x	x	x	x
Stepwise Regression	x	x	x	

Model Selection Method	thinness..1.19.years	Income.compostition.of.resources	Schooling
Forward Selection	x	x	x
Backward Elimination	x	x	x
Stepwise Regression	x	x	x

Both the Forward Selection method and Backward Elimination method have chosen the same set of variables.

K-Fold Cross Validation

We were left with 2 models - the full model and the model built using Stepwise Regression method - as our candidate models. To find out which model is the better one to pick as our final model, we ran K-Fold Cross Validation on both models and subsequently picked the model with the lowest mean RMSE value as our final model. We chose the value of K to be 5.

In K-Fold Cross-Validation, the original sample of the dataset is randomly partitioned into K equal sized subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining K-1 subsamples are used as training data. The cross-validation process is then repeated K times, with each of the K subsamples used exactly once as the validation data. The K results are then averaged to produce a single estimation, which in our case is the mean RMSE value.

```
# Define training control
set.seed(13245)
train.control = trainControl(method = 'cv', number = 5)
```

Cross-Validation for the full model:

```
# Train the model
CV_all = train(
  Life.expectancy ~ .,
  data = life,
  method = 'lm',
  trControl = train.control
)
```

```
# Summarize the results
CV_all
```

```
## Linear Regression
##
## 1649 samples
##   16 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1318, 1319, 1319, 1320, 1320
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.719367  0.8214392  2.853294
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Cross-Validation for model chosen by Stepwise Regression method:

```
# Train the model
CV_stepwise = train(
  Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources + BMI + GDP
  data = life,
  method = 'lm',
  trControl = train.control
)
```

```
# Summarize the results
CV_stepwise
```

```
## Linear Regression
##
## 1649 samples
##   12 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1319, 1320, 1318, 1320, 1319
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.718134  0.8211816  2.853578
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Results of 5-Fold Cross-Validation of the 2 models:

```
rbind(CV_all$results, CV_stepwise$results)
```

```
##   intercept     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 3.719367 0.8214392 2.853294 0.1466741 0.01735040 0.1035478
## 2      TRUE 3.718134 0.8211816 2.853578 0.2043897 0.01790453 0.1159896
```

Since the model chosen by Stepwise Regression method `lmod_stepwise` has a lower RMSE value, we have selected this model to be our final model.

Results

Out of the 2 candidate models, we have picked the model chosen by the Stepwise Regression method to be our final model. This decision was based on the fact that the Stepwise Regression model had the lowest mean RMSE value when evaluated with K-Fold Cross Validation (K = 5).

```
lmod_final = lm(
  Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources + BMI + GDP
  data = life
)
summary(lmod_final)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
##     Income.composition.of.resources + BMI + GDP + Diphtheria +
##     Alcohol + thinness..1.19.years + Status + Hepatitis.B + Total.expenditure,
##     data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9828  -2.1039   0.0534   2.3728  11.4923
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.505e+01  8.059e-01  68.312  < 2e-16 ***
## Schooling                        8.931e-01  6.129e-02  14.572  < 2e-16 ***
## HIV.AIDS                        -4.269e-01  1.846e-02 -23.131  < 2e-16 ***
## Adult.Mortality                 -1.781e-02  9.641e-04 -18.471  < 2e-16 ***
## Income.composition.of.resources  1.034e+01  8.471e-01  12.204  < 2e-16 ***
## BMI                              3.577e-02  6.082e-03   5.881 4.94e-09 ***
## GDP                              6.079e-05  9.646e-06   6.302 3.77e-10 ***
## Diphtheria>=90% Covered          1.452e+00  3.444e-01   4.215 2.64e-05 ***
## Alcohol                         -1.607e-01  3.355e-02  -4.789 1.83e-06 ***
## thinness..1.19.years            -6.820e-02  2.502e-02  -2.726  0.00648 **
## StatusDeveloping                -9.968e-01  3.465e-01  -2.877  0.00407 **
## Hepatitis.B>=90% Covered        -6.338e-01  3.140e-01  -2.018  0.04371 *
## Total.expenditure                7.257e-02  4.164e-02   1.743  0.08155 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.692 on 1636 degrees of freedom
## Multiple R-squared:  0.8251, Adjusted R-squared:  0.8239
## F-statistic: 643.3 on 12 and 1636 DF,  p-value: < 2.2e-16
```

The final model contains 12 variables: `Schooling`, `HIV.AIDS`, `Adult.Mortality`, `Income.composition.of.resources`, `BMI`, `GDP`, `Diphtheria`, `Alcohol`, `thinness..1.19.years`, `Status`, `Hepatitis.B`, and `Total.expenditure`.

There are a few takeaways from this final model:

- The p-value of the model is 2.2e-16 < 0.05, indicating that it is significant.
- The adjusted R-squared value of the model is 0.8239, indicating that about 82.39% of the variability observed in the target variable can be explained by the independent variables in the model.
- Pretty much all the variables in the model are the most significant variables with p-value < 0.05.
- From the model we can interpret that `Income.composition.of.resources` has a strong positive effect on life expectancy.
- From the model we can interpret that `HIV.AIDS`, `Adult.Mortality`, `Alcohol`, `thinness..1.19.years`, and `StatusDeveloping`, may have a negative effect on life expectancy.
- `Hepatitis.B>=90% Covered` may also have a negative effect on life expectancy, the same observation we had seen previously on the full model.

Model Error Estimation

We have primarily used the R-squared, the adjusted R-squared, the root-mean-square error (RMSE), and the mean absolute error (MAE) as the metric for evaluating our models.

The estimates for the final model are derived from the results of K-Fold Cross Validation (K = 5), and are summarized in the following table:

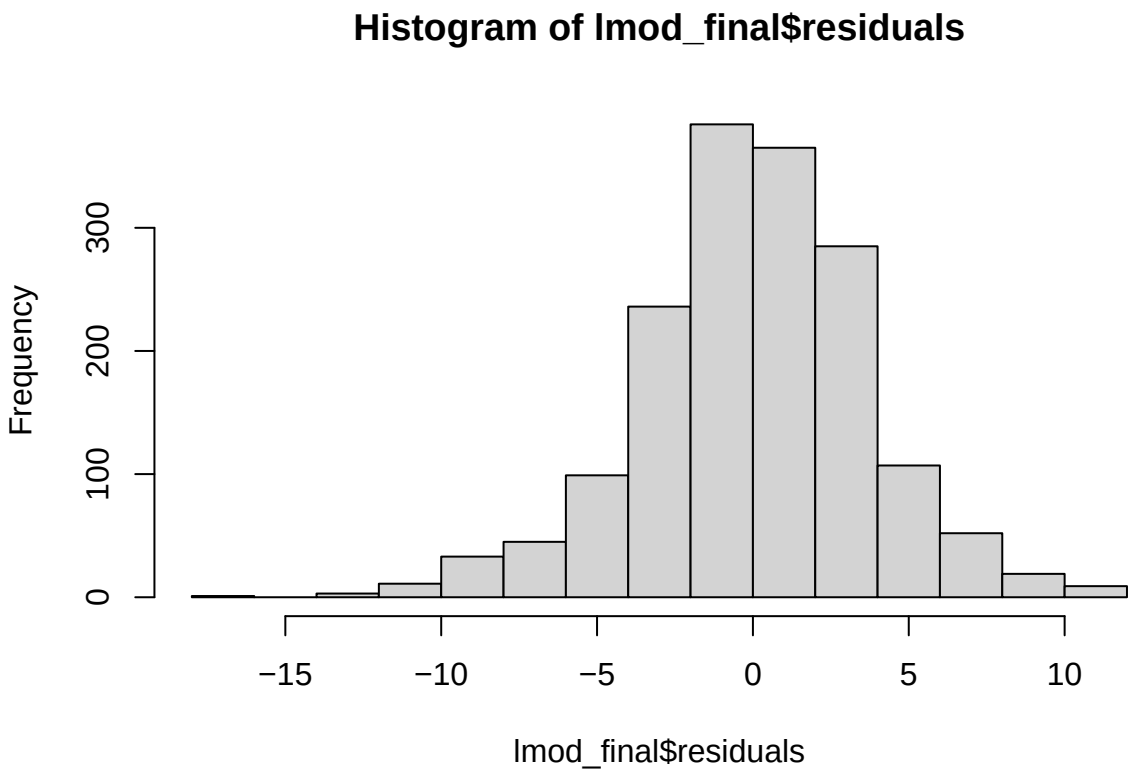
Metric	Estimate
R-squared	0.8219568
RMSE	3.711009
MAE	2.846318

Model Adequacy Checking

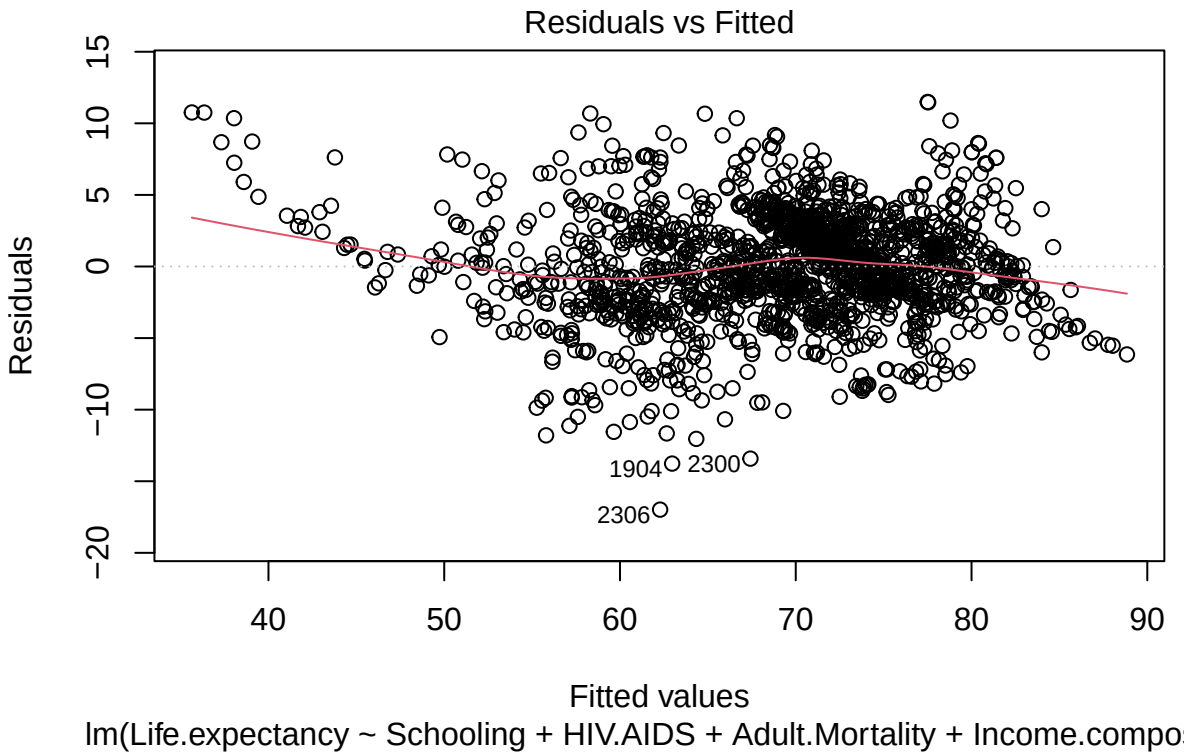
To make sure our final model behaves as expected in terms of prediction and inference, we need to test that the assumptions made in building the Linear Regression model are not broken. These assumptions are:

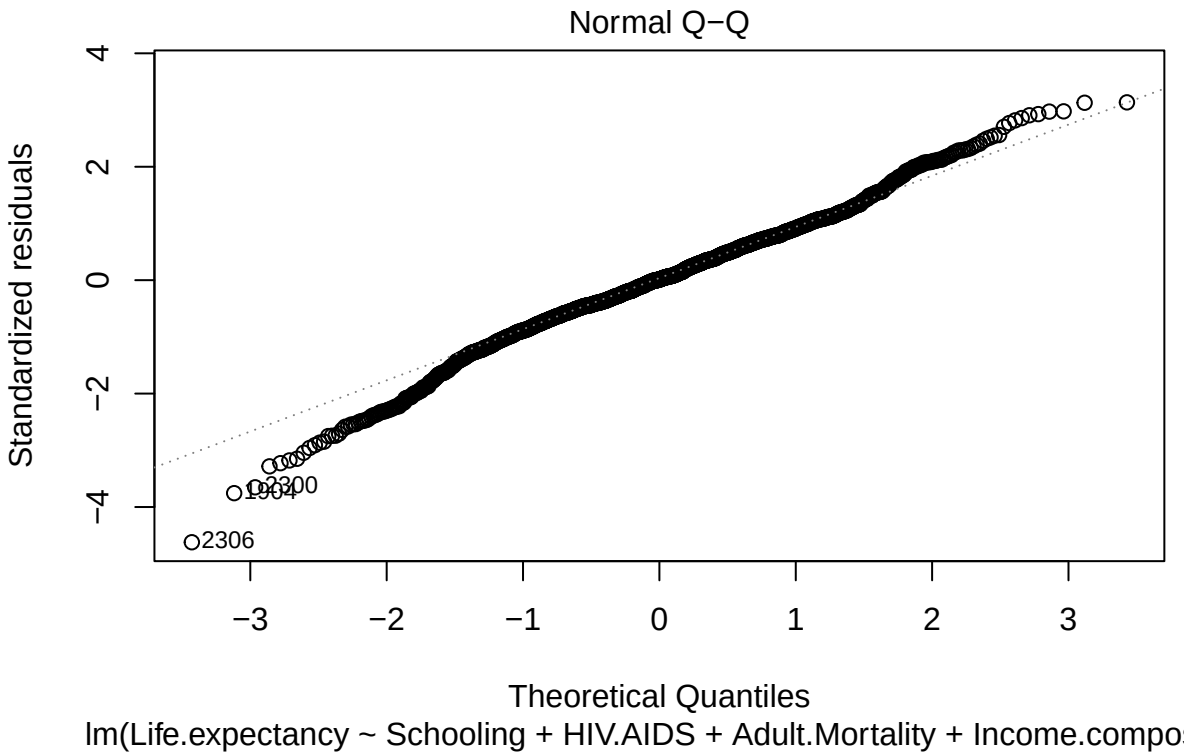
1. The relationship between the response y and the regressors is linear, at least approximately.
2. The error term ϵ has zero mean.
3. The error term ϵ has constant variance σ^2.
4. The errors are uncorrelated.
5. The errors are normally distributed. (Montgomery, Peck, and Vining 2021)

We will be testing these assumptions by checking whether the residuals of the model are normally distributed, whether the plot of the residuals against the fitted values show any pattern, and whether there are any anomalies in the Q-Q plot.



Most of the residuals seem to be distributed in the center, indicating that they are distributed normally.


Residuals vs Fitted
Residuals
15
10
5
0
−10
−20
40
50
60
70
80
90
1904
2300
2306
Fitted values
lm(Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.compos ...



There is no obvious observable pattern in the above plots, and so we conclude that our final model is appropriate.

We also run a multicollinearity test to see if there is any multicollinearity between the variables in the model. We have used the Variance Inflation Factor (VIF) to determine the multicollinearity between the variables.

```
vif(lmod_final)
```

```
##                       Schooling                        HIV.AIDS
##                        3.548828                        1.498703
##                Adult.Mortality Income.composition.of.resources
##                        1.764485                        2.908091
##                             BMI                             GDP
##                        1.745170                        1.481404
##                      Diphtheria                         Alcohol
##                        3.511705                        2.209220
##           thinness..1.19.years                          Status
##                        1.600520                        1.818933
##                     Hepatitis.B              Total.expenditure
##                        2.981721                        1.108283
```

A VIF > 10 implies serious problems with multicollinearity.
Since the VIF for all of the predictors is less than 10, there seems to be no issue with multicollinearity.

Discussion

We have implemented a Linear Regression model that predicts the life expectancy of the human population using 12 variables related to life expectancy and health factors. We have shown that several of these variables have some correlations with the dependent variable suggesting that fitting a Linear Regression model to this dataset would be tractable. Our model also passes the normality assumption and has no issues with the multicollinearity of the variables.

While all the variables in our model are significant, the inferences made from this model should be taken with a grain of salt - we do not believe that an increase in Hepatitis B (HepB) immunization coverage among 1-year-olds should be associated with a declining life expectancy in a country. This calls for further investigation.

We had planned to include the Best Subsets Regression method and the All Possible Regressions method for building our candidate models, but since the computational time for running these methods was too high on our machines, we have omitted these methods in our analysis.

Due to the limited amount of time for this analysis, we have not explored working with Polynomial Regression and including interaction terms in our models. We believe that in particular, a squared term for `BMI` in the model would be appropriate since it is well-known that too low or too high of a BMI poses health risks for individuals. It is also possible to convert this numerical variable into categorical values, for example, 'Underweight' for BMI < 18.5, 'Healthy' for $18.5 \leq$ BMI < 25, 'Overweight' $25 \leq$ BMI < 30, and 'Obese' for BMI > 30.

We are curious to know how our model compares to other regression techniques such as KNN Regression, Support Vector Regression, Decision Tree Regression, and even Randon Forest Regression. We have left this for future work.

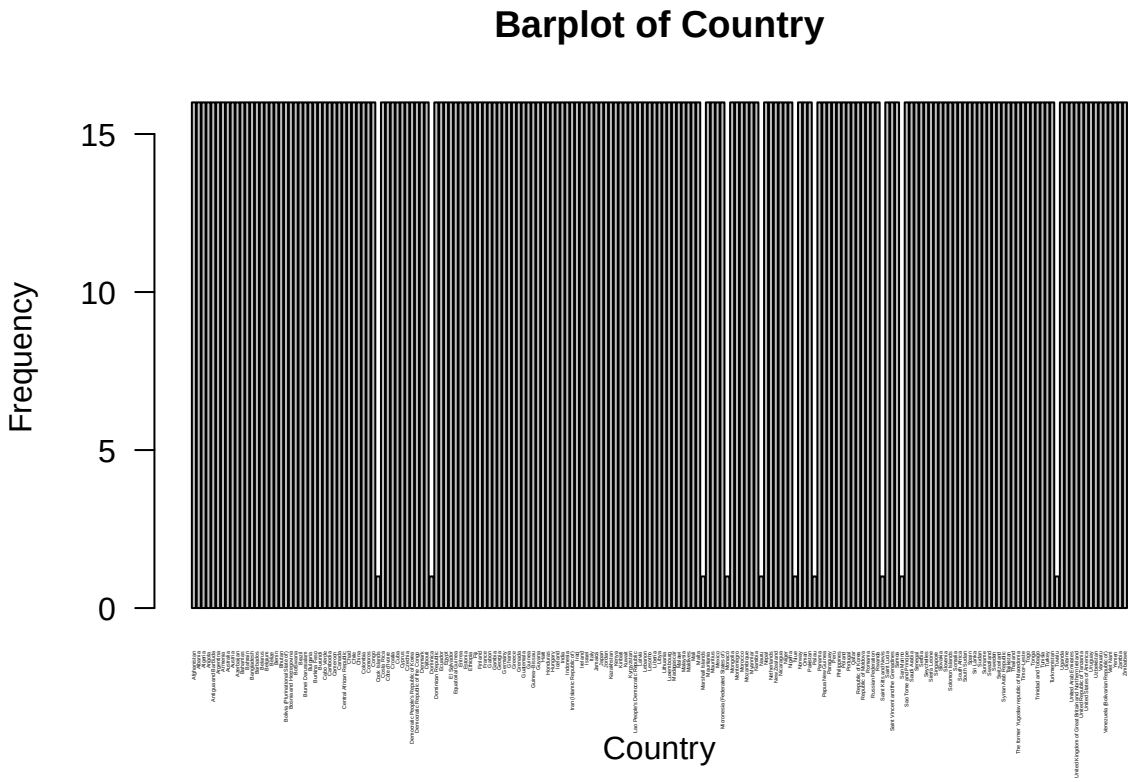
Though we believe that the variables we have in our dataset are relevant in predicting the life expectancy of the human population, we think that having more relevant variables like sex, exercise, smoking, and environment pollution would be even more helpful for prediction and inference. It also almost never harms in collecting more observations of data related to life expectancy and health factors for building our model.

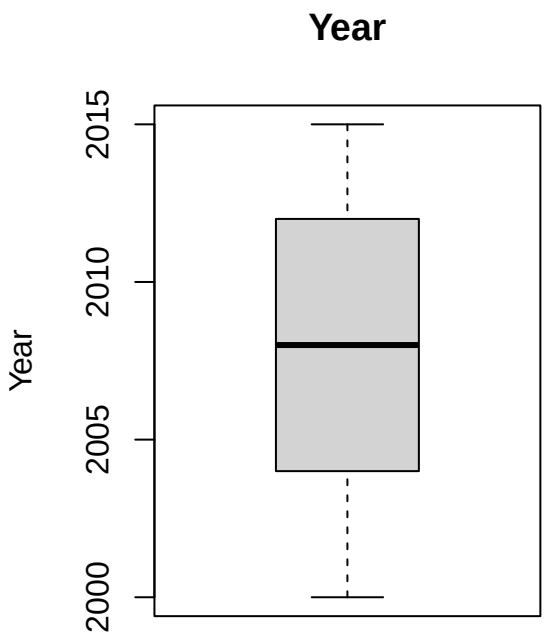
Literature Cited

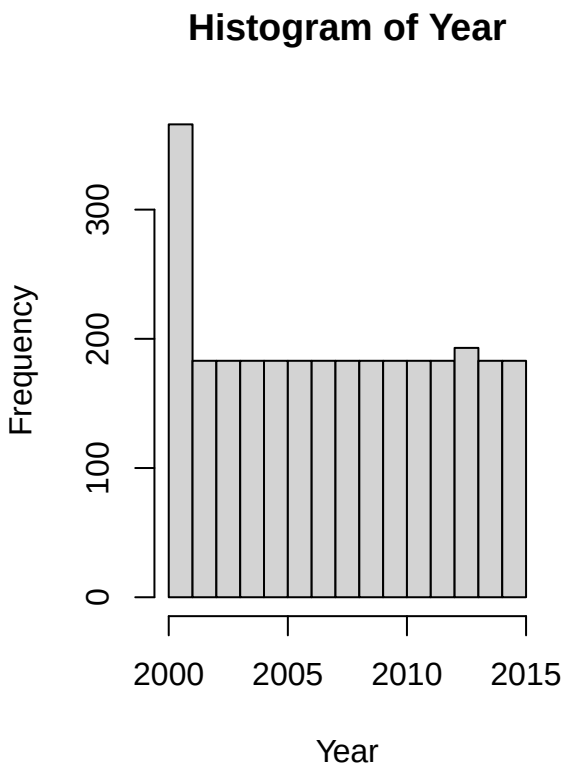
Luy, Marc, Paola Di Giulio, Vanessa Di Lego, Patrick Lazarevič, and Markus Sauerberg. 2020. "Life Expectancy: Frequently Used, but Hardly Understood." *Gerontology* 66 (1): 95–104.

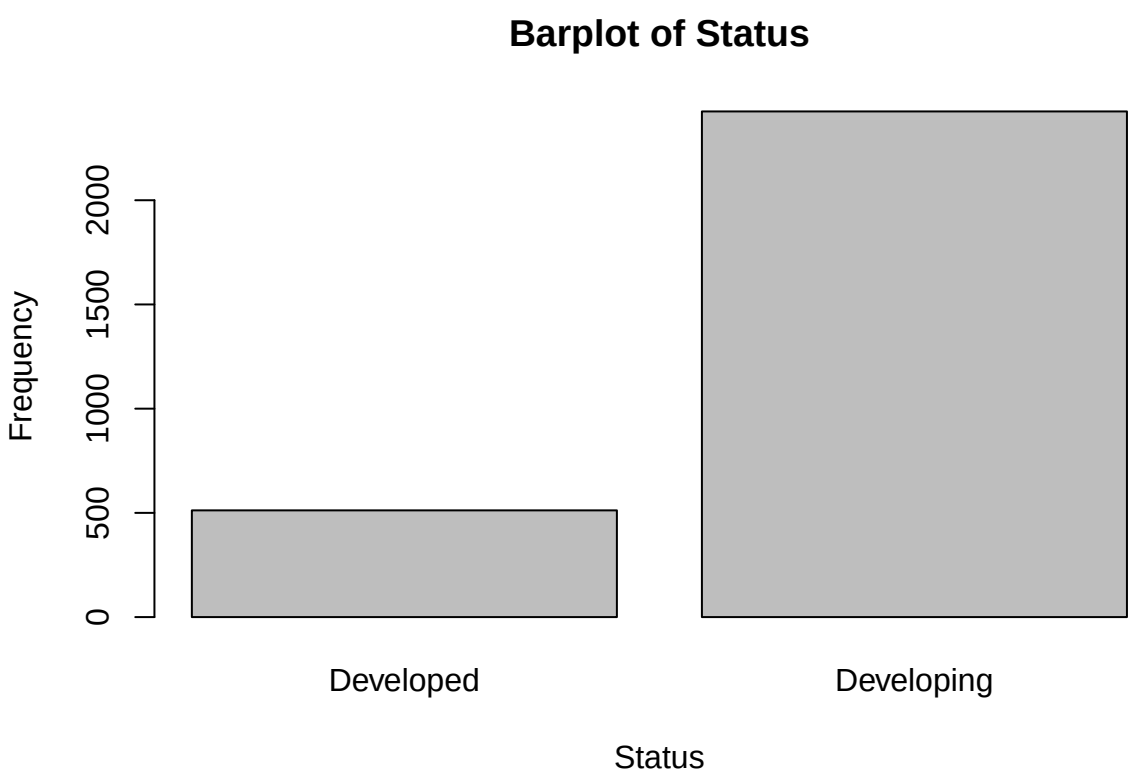
Montgomery, Douglas C, Elizabeth A Peck, and G Geoffrey Vining. 2021. *Introduction to Linear Regression Analysis*. John Wiley & Sons.

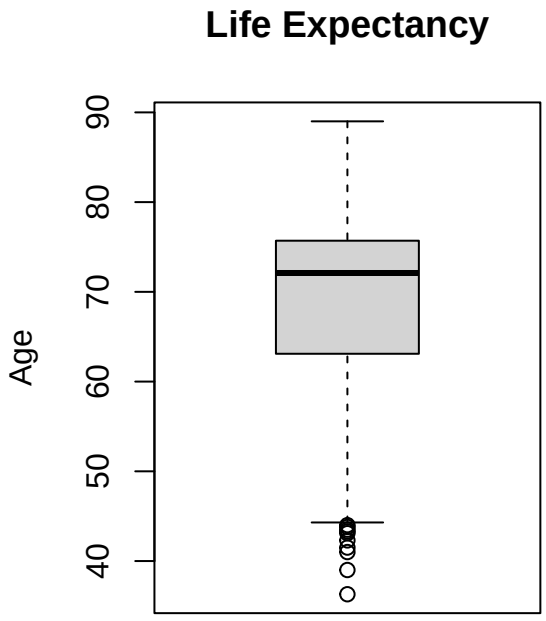
Appendix A - Variable Distribution Plots

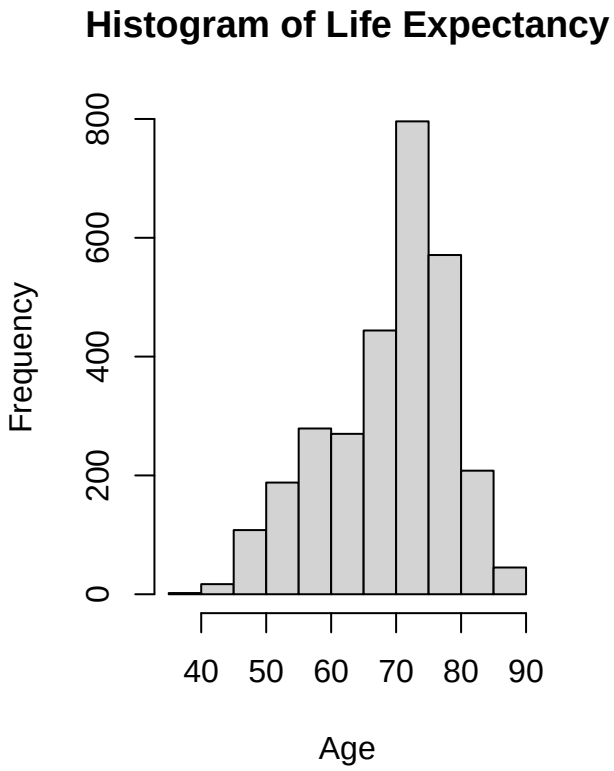


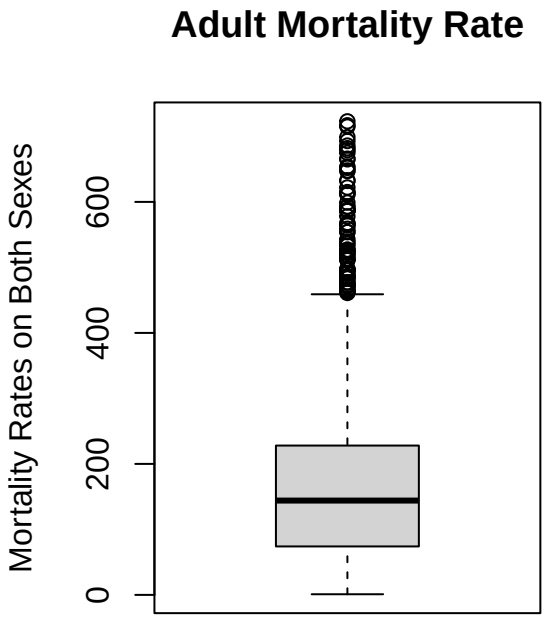

Year
Year
2000
2005
2010
2015

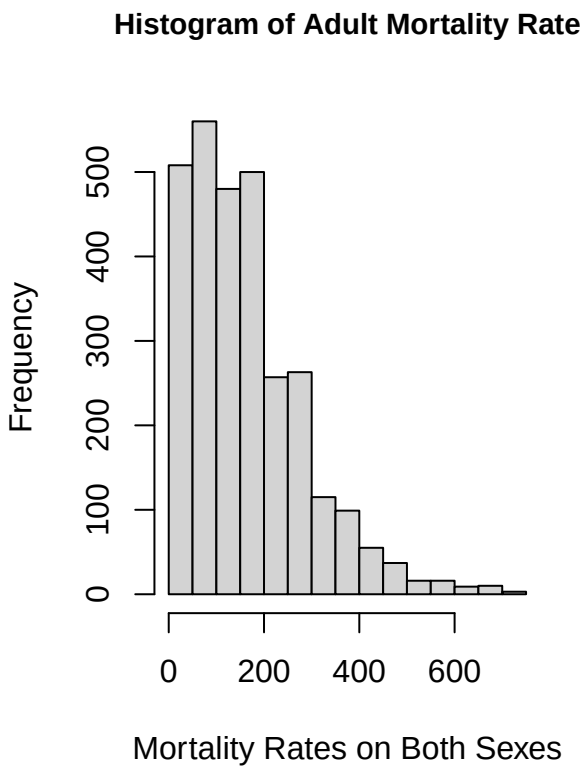

Histogram of Year
Frequency
0
100
200
300
2000
2005
2010
2015
Year

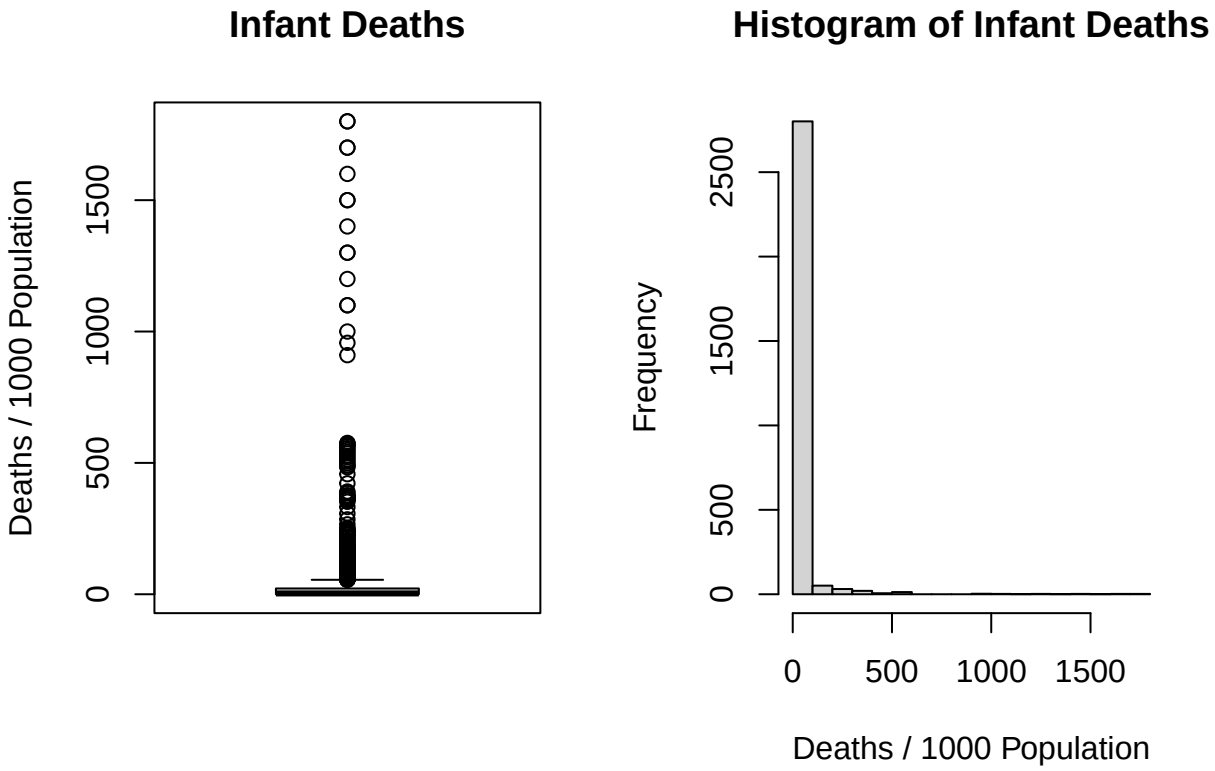

Barplot of Status
Frequency
0
500
1000
1500
2000
Developed
Developing
Status

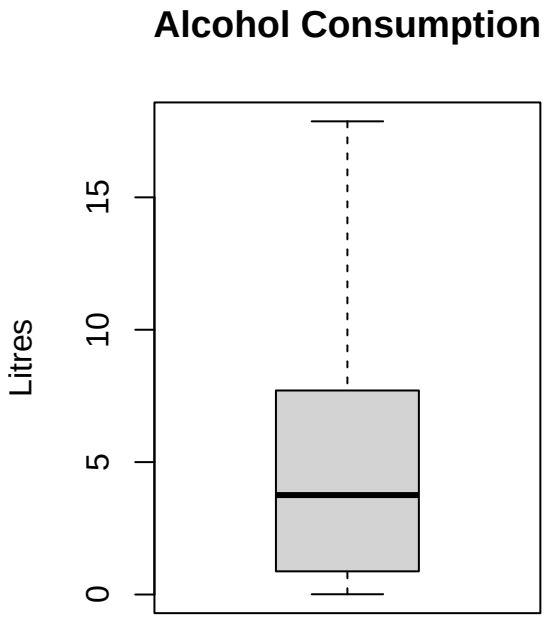

Life Expectancy
Age
40
50
60
70
80
90

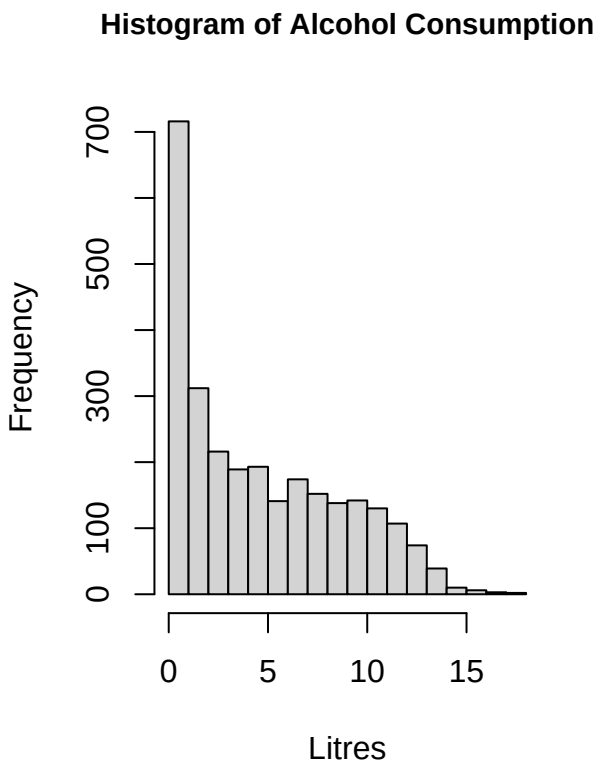

Histogram of Life Expectancy
Frequency
0
200
400
600
800
40
50
60
70
80
90
Age

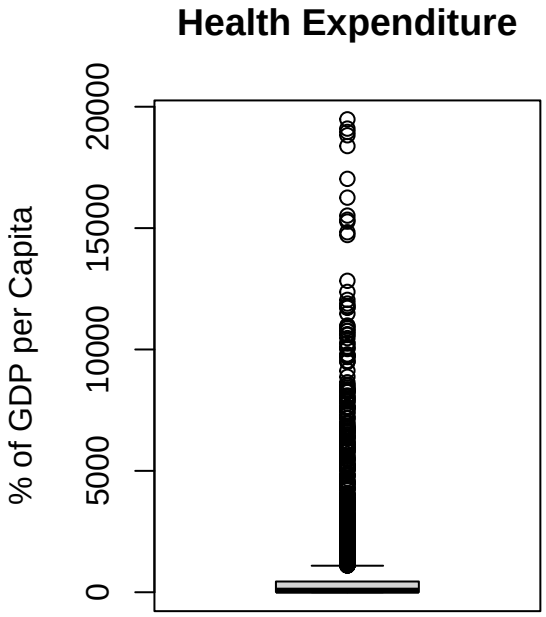

Adult Mortality Rate
Mortality Rates on Both Sexes
0
200
400
600

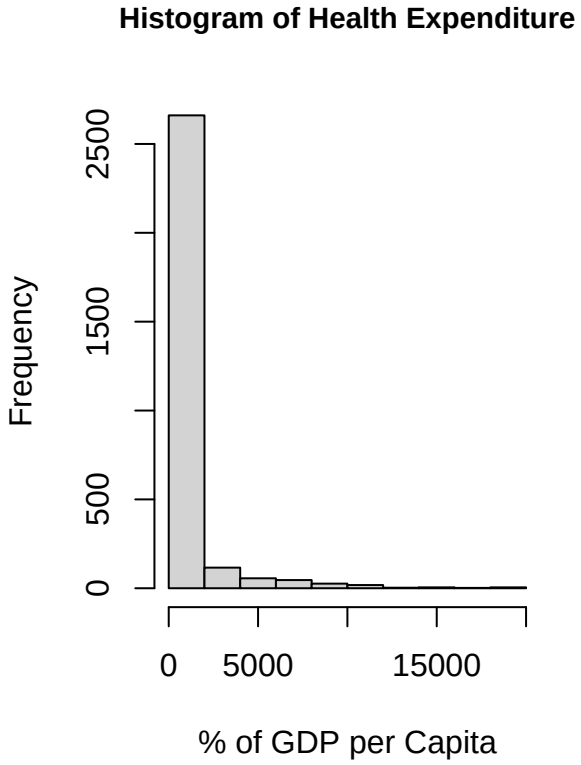

Histogram of Adult Mortality Rate
Frequency
0
100
200
300
400
500
0
200
400
600
Mortality Rates on Both Sexes

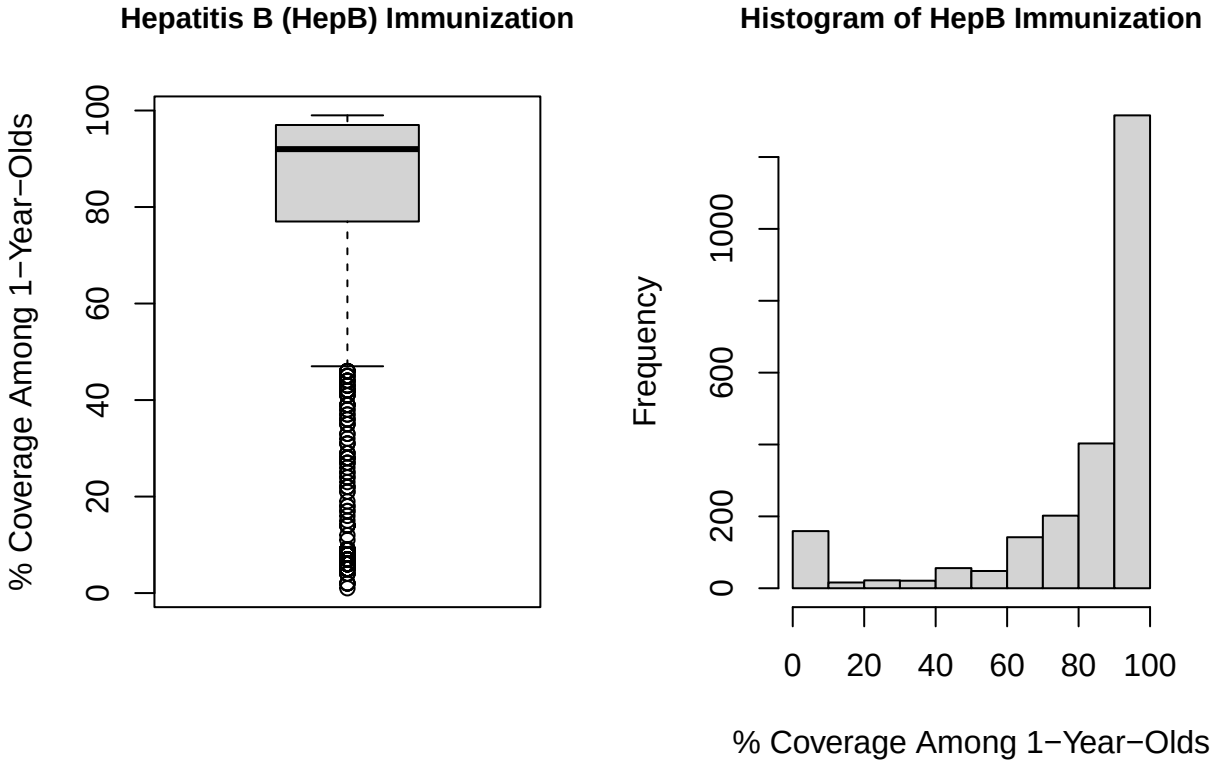

Infant Deaths
Deaths / 1000 Population
0
500
1000
1500
Histogram of Infant Deaths
Frequency
0
500
1500
2500
0
500
1000
1500
Deaths / 1000 Population

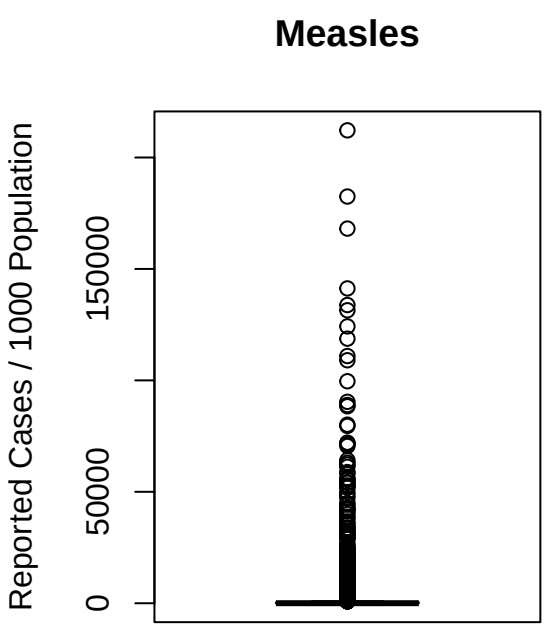

Alcohol Consumption
Litres
0
5
10
15

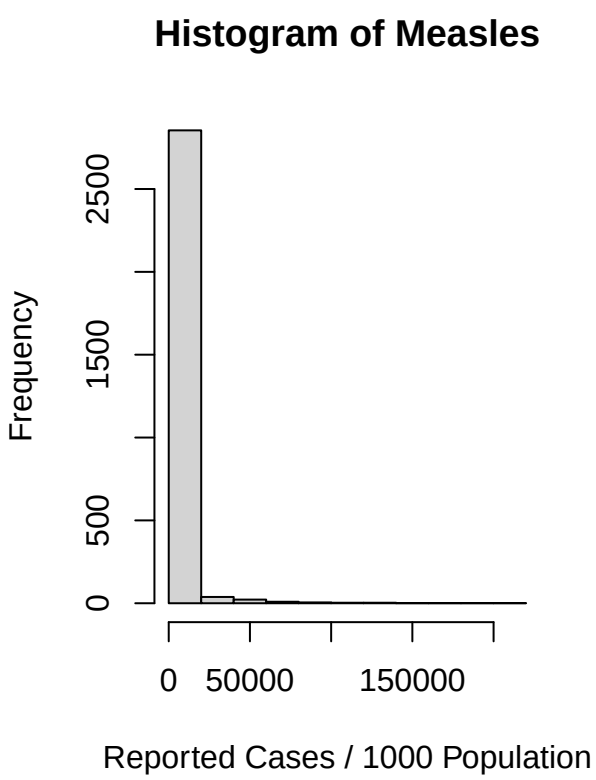

Histogram of Alcohol Consumption
Frequency
0
100
300
500
700
0
5
10
15
Litres

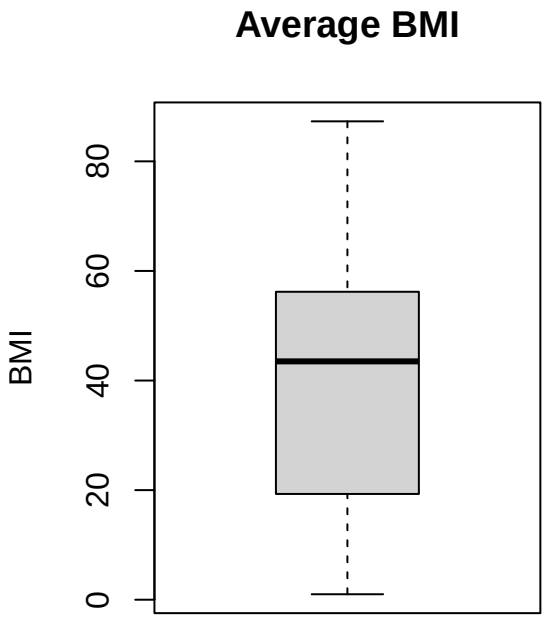

Health Expenditure
% of GDP per Capita
0
5000
10000
15000
20000

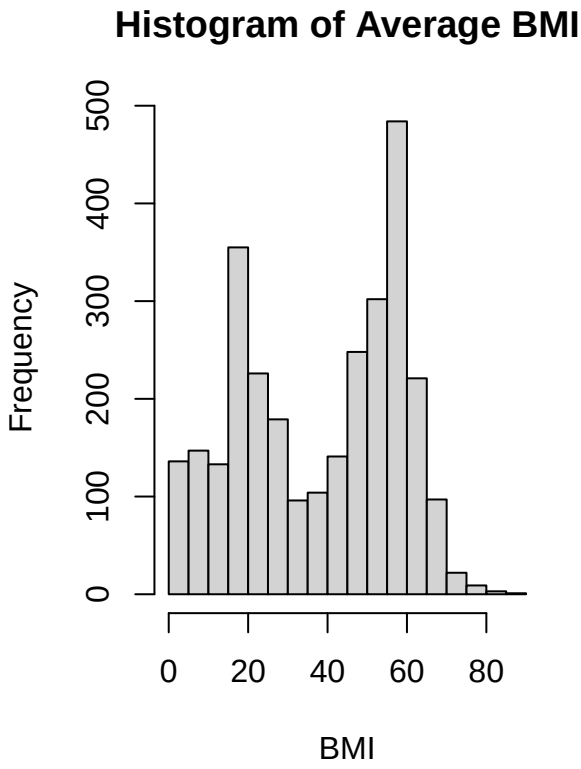

Histogram of Health Expenditure
Frequency
0
500
1500
2500
0
5000
15000
% of GDP per Capita

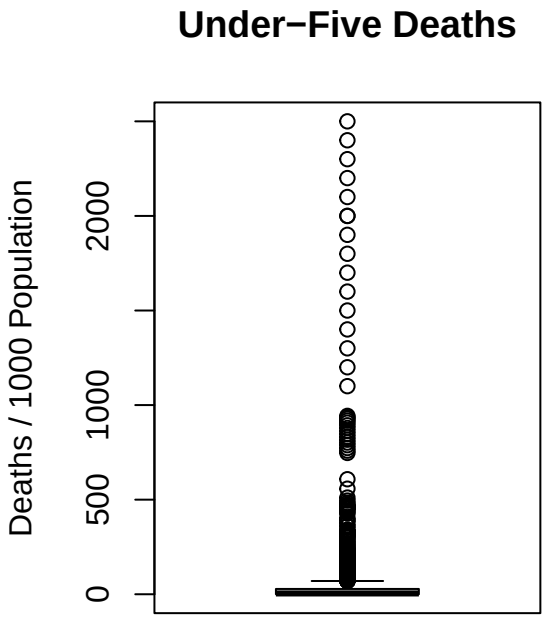

Hepatitis B (HepB) Immunization
% Coverage Among 1−Year−Olds
0
20
40
60
80
100
Histogram of HepB Immunization
Frequency
0
200
600
1000
0
20
40
60
80
100
% Coverage Among 1−Year−Olds

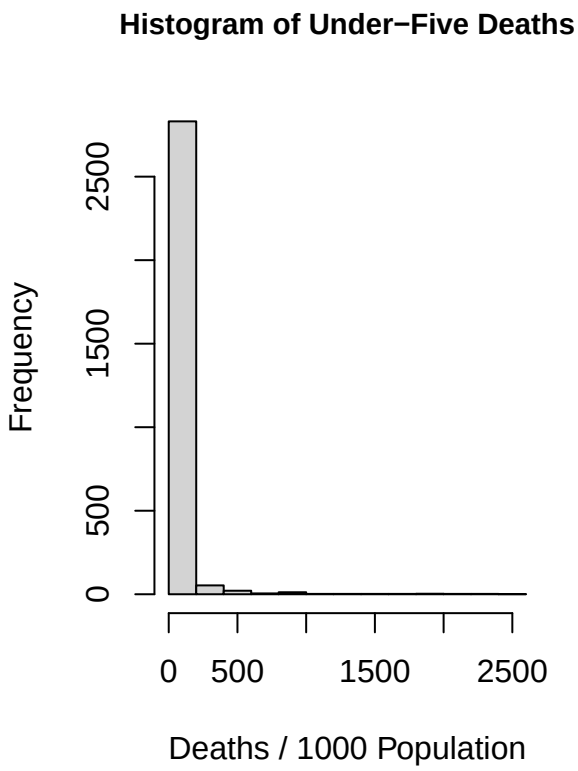

Measles
Reported Cases / 1000 Population
0
50000
150000

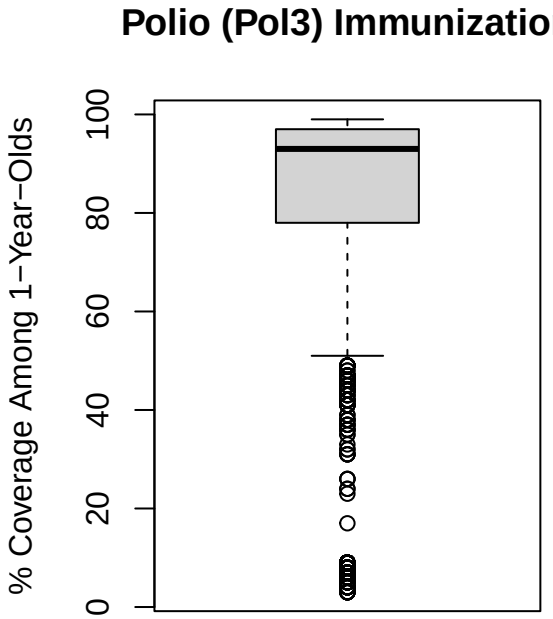

Histogram of Measles
Frequency
0
500
1500
2500
0
50000
150000
Reported Cases / 1000 Population

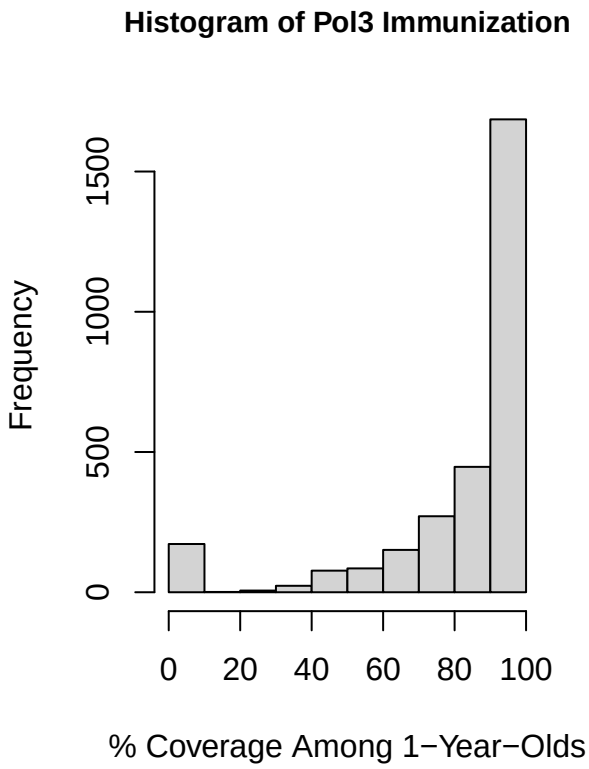

Average BMI
BMI
0
20
40
60
80

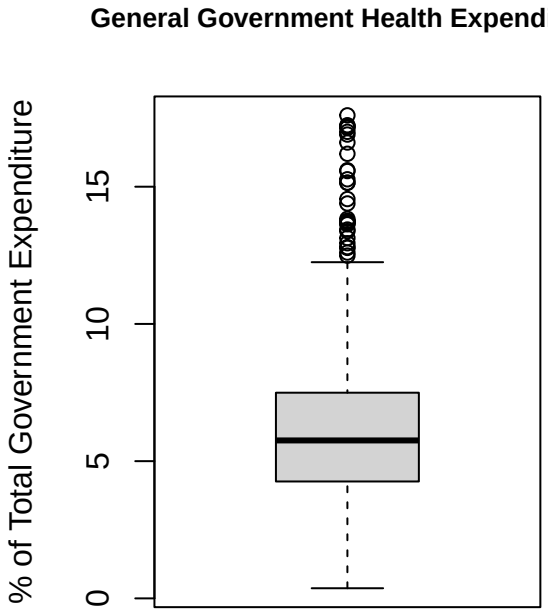

Histogram of Average BMI
Frequency
0
100
200
300
400
500
0
20
40
60
80
BMI

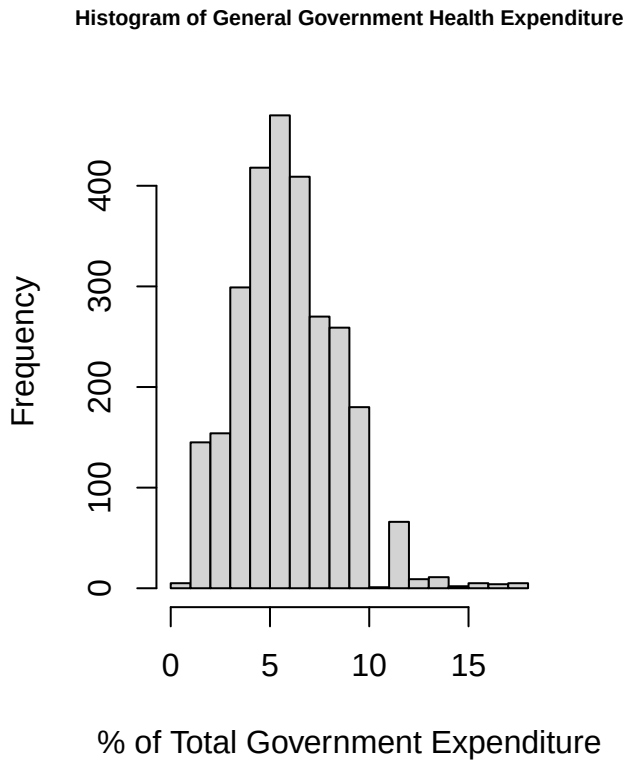

Under−Five Deaths
Deaths / 1000 Population
0
500
1000
2000

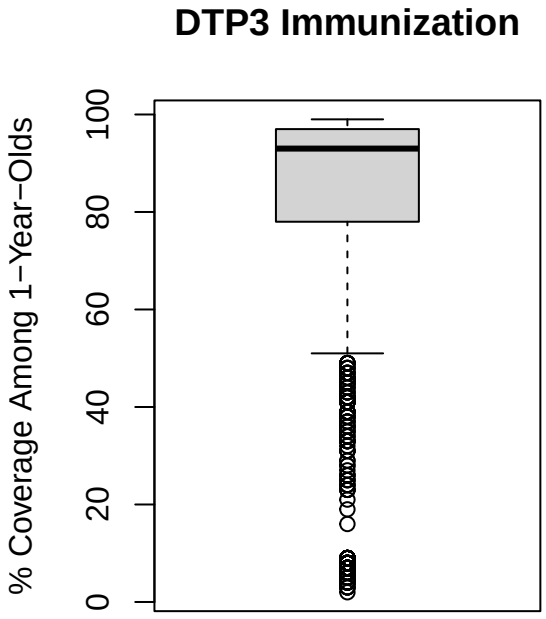

Histogram of Under−Five Deaths
Frequency
0
500
1500
2500
0
500
1500
2500
Deaths / 1000 Population

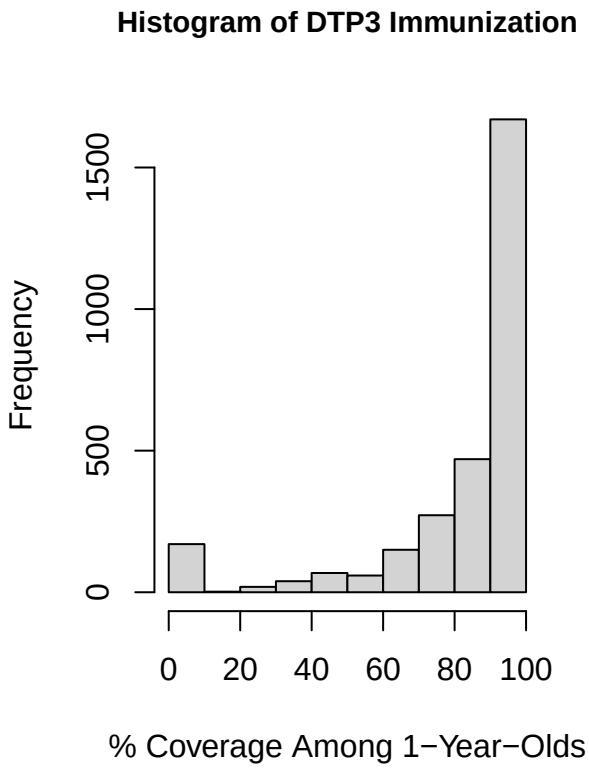

Polio (Pol3) Immunization
% Coverage Among 1−Year−Olds
0
20
40
60
80
100

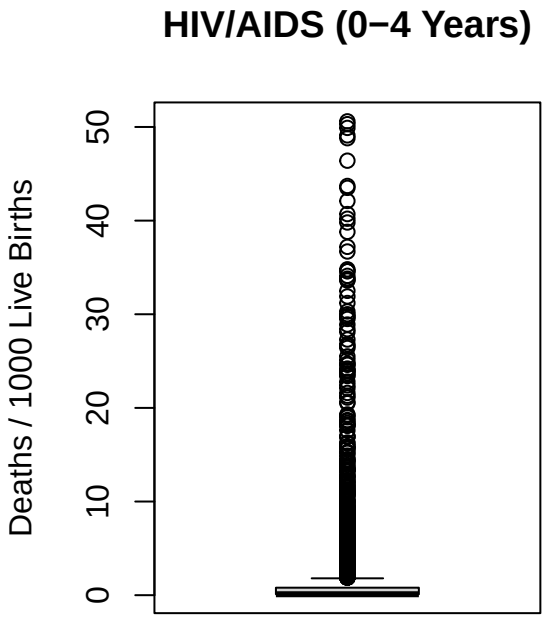

Histogram of Pol3 Immunization
Frequency
0
500
1000
1500
0
20
40
60
80
100
% Coverage Among 1−Year−Olds

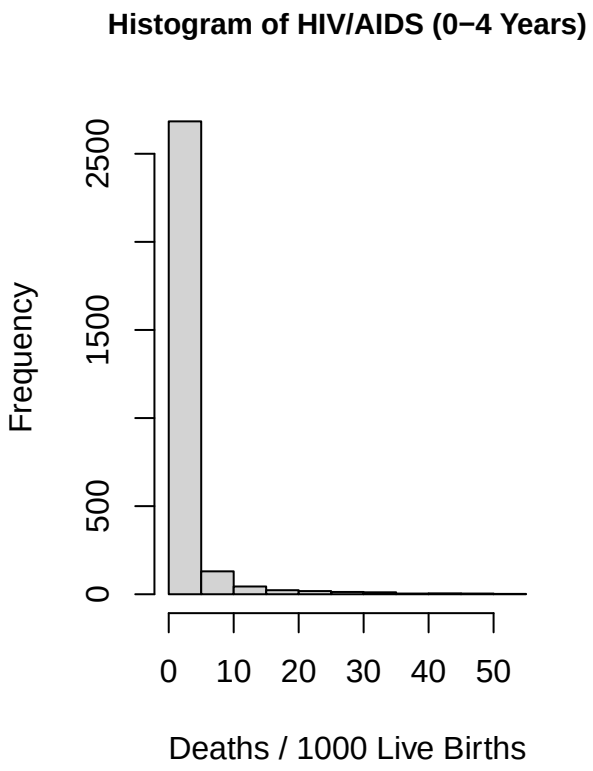

General Government Health Expenditure
% of Total Government Expenditure
0
5
10
15

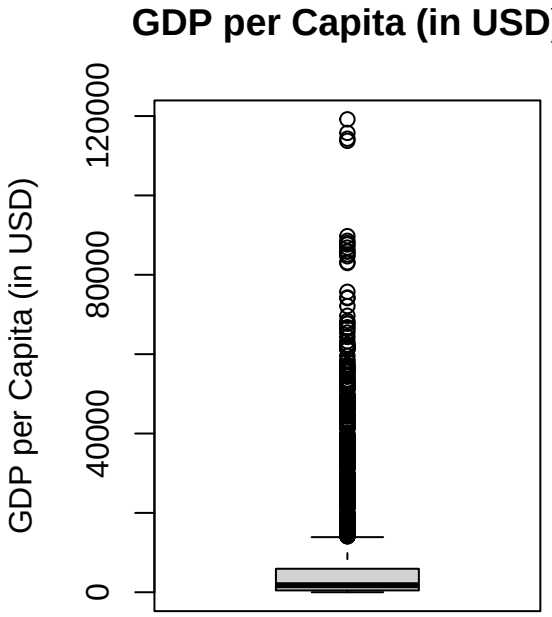

Histogram of General Government Health Expenditure
Frequency
0
100
200
300
400
0
5
10
15
% of Total Government Expenditure

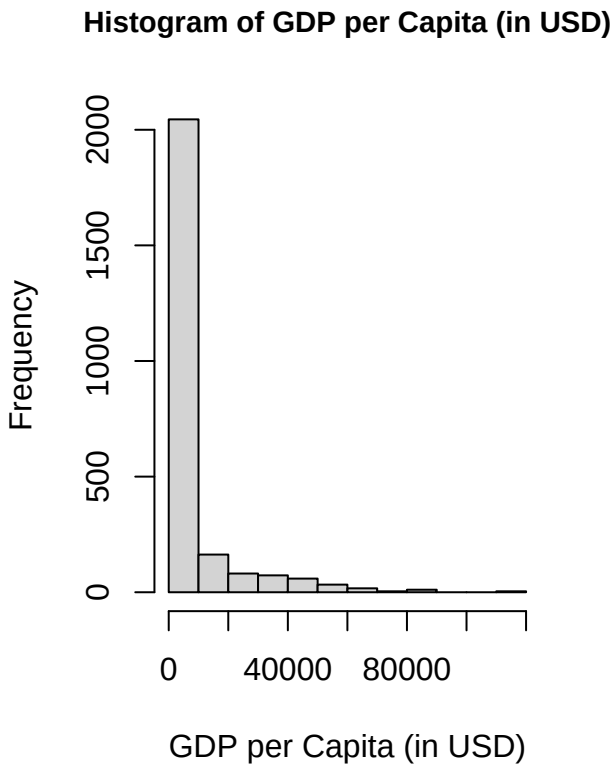

DTP3 Immunization
% Coverage Among 1−Year−Olds
0
20
40
60
80
100

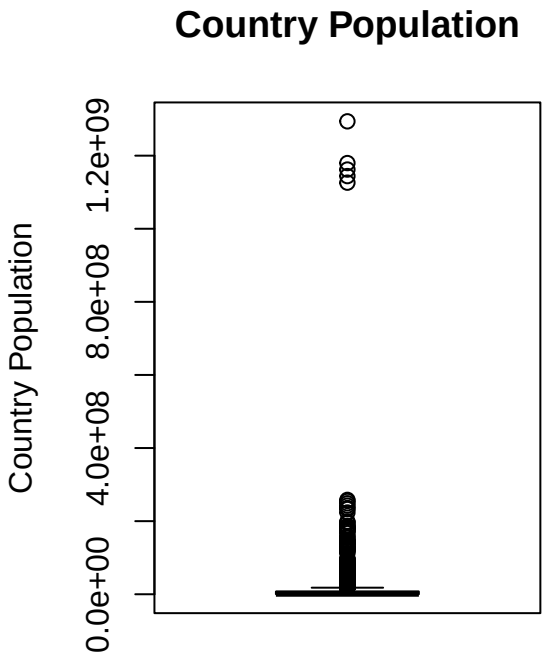

Histogram of DTP3 Immunization
Frequency
0
500
1000
1500
0
20
40
60
80
100
% Coverage Among 1−Year−Olds

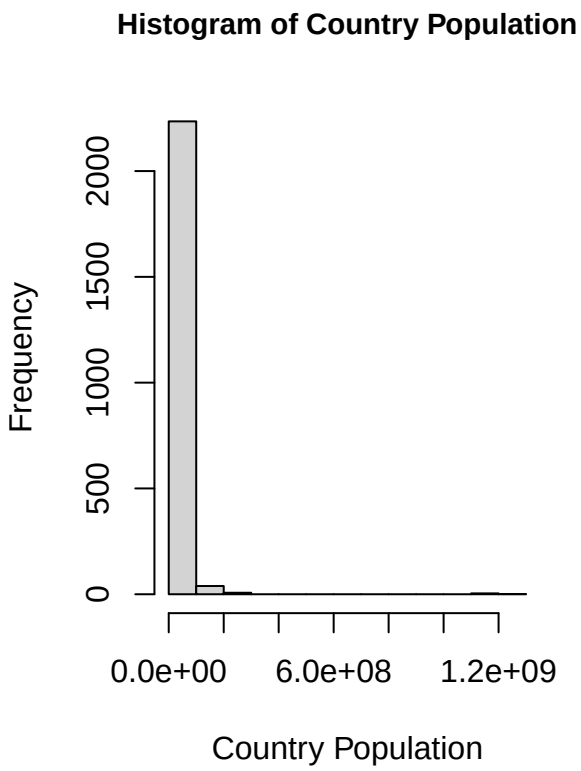

HIV/AIDS (0−4 Years)
Deaths / 1000 Live Births
0
10
20
30
40
50

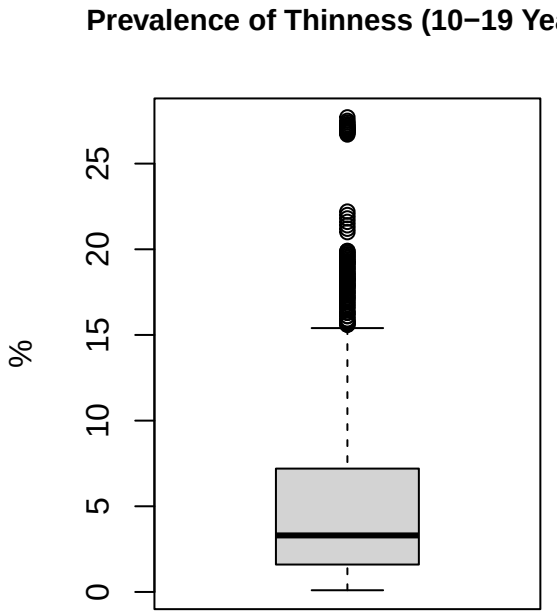

Histogram of HIV/AIDS (0−4 Years)
Frequency
0
500
1500
2500
0
10
20
30
40
50
Deaths / 1000 Live Births

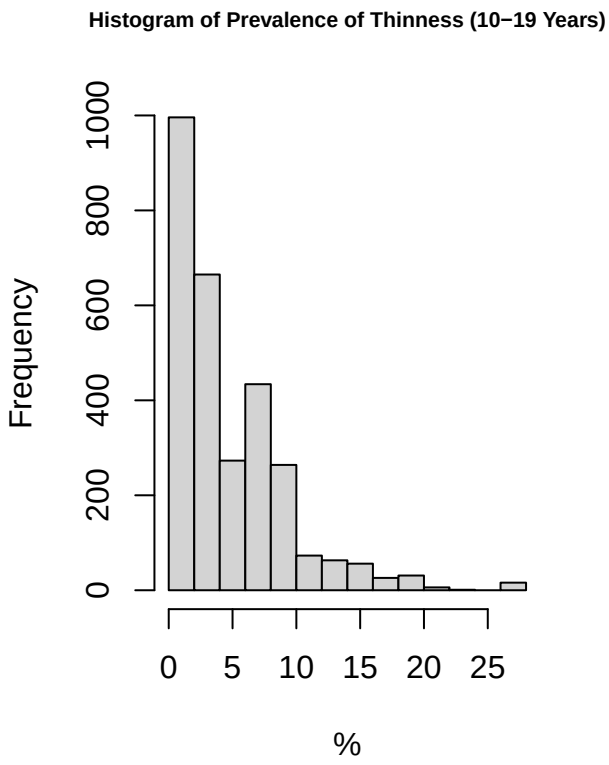

GDP per Capita (in USD)
GDP per Capita (in USD)
0
40000
80000
120000

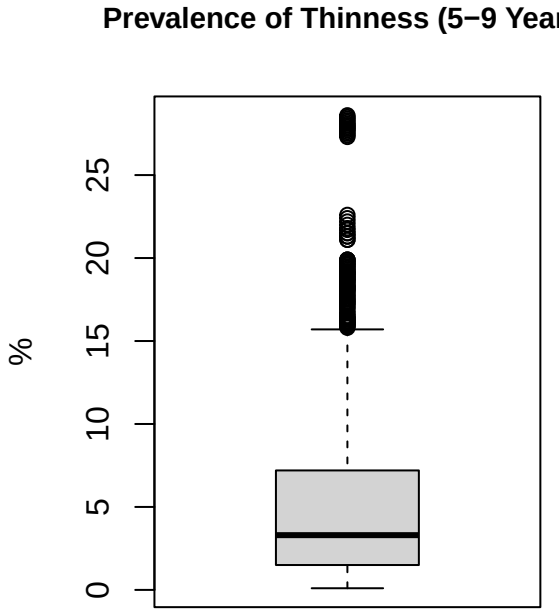

Histogram of GDP per Capita (in USD)
Frequency
0
500
1000
1500
2000
0
40000
80000
GDP per Capita (in USD)

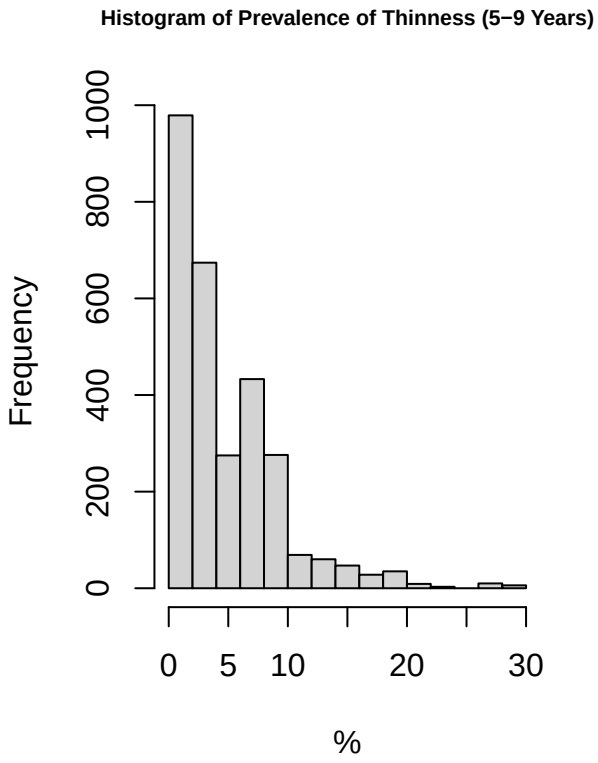

Country Population
Country Population
0.0e+00
4.0e+08
8.0e+08
1.2e+09

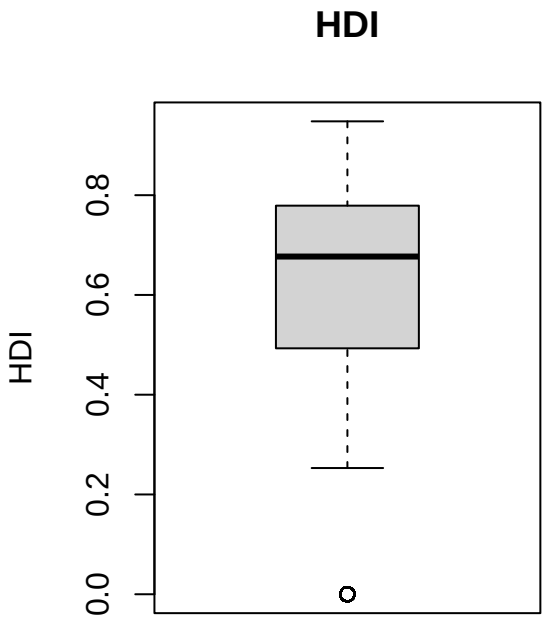

Histogram of Country Population
Frequency
0
500
1000
1500
2000
0.0e+00
6.0e+08
1.2e+09
Country Population

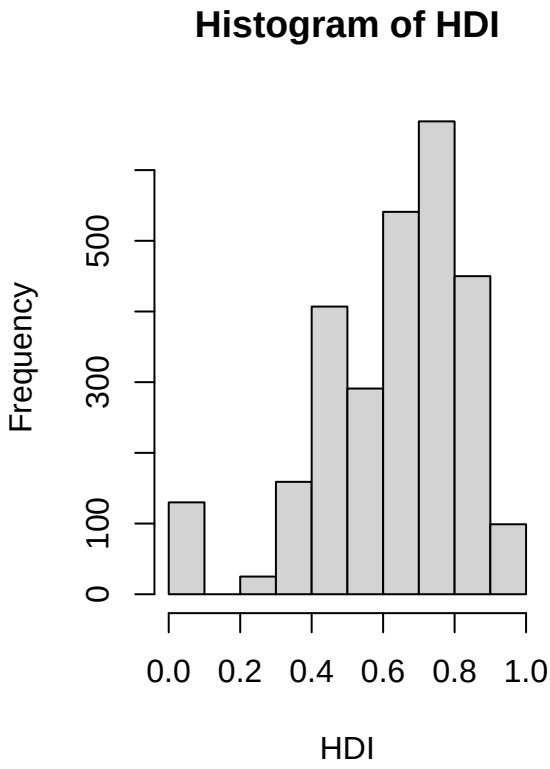

Prevalence of Thinness (10−19 Years)
%
0
5
10
15
20
25

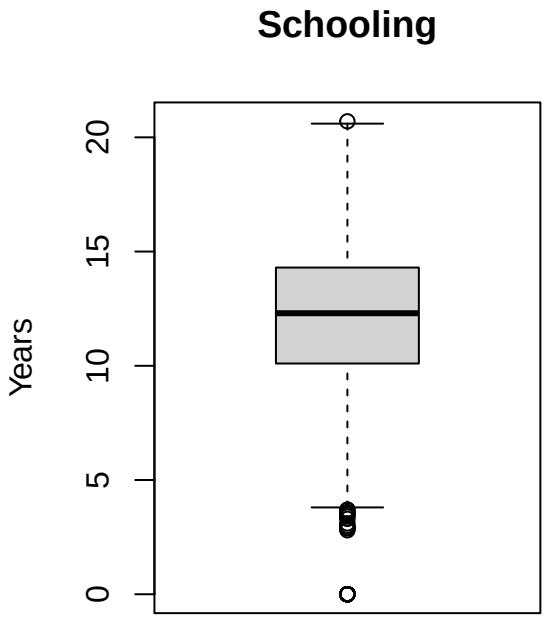

Histogram of Prevalence of Thinness (10−19 Years)
Frequency
0
200
400
600
800
1000
0
5
10
15
20
25
%

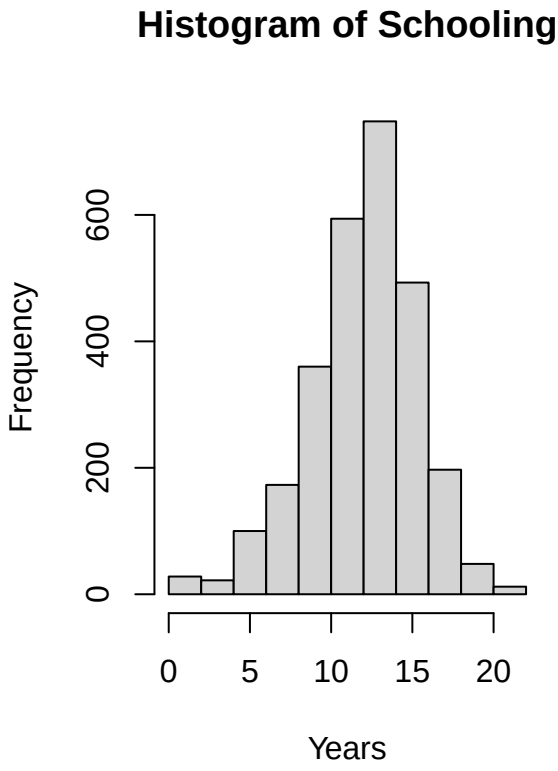

Prevalence of Thinness (5−9 Years)
%
0
5
10
15
20
25


Histogram of Prevalence of Thinness (5−9 Years)
Frequency
0
200
400
600
800
1000
0
5
10
20
30
%


HDI
HDI
0.0
0.2
0.4
0.6
0.8


Histogram of HDI
Frequency
0
100
300
500
0.0
0.2
0.4
0.6
0.8
1.0
HDI


Schooling
Years
0
5
10
15
20


Histogram of Schooling
Frequency
0
200
400
600
0
5
10
15
20
Years