Life Expectancy Analysis

George Matlis, Xaris Moutafidis

Aristotle University of Thessaloniki

February 14, 2024

Presentation Contents

Project Overview

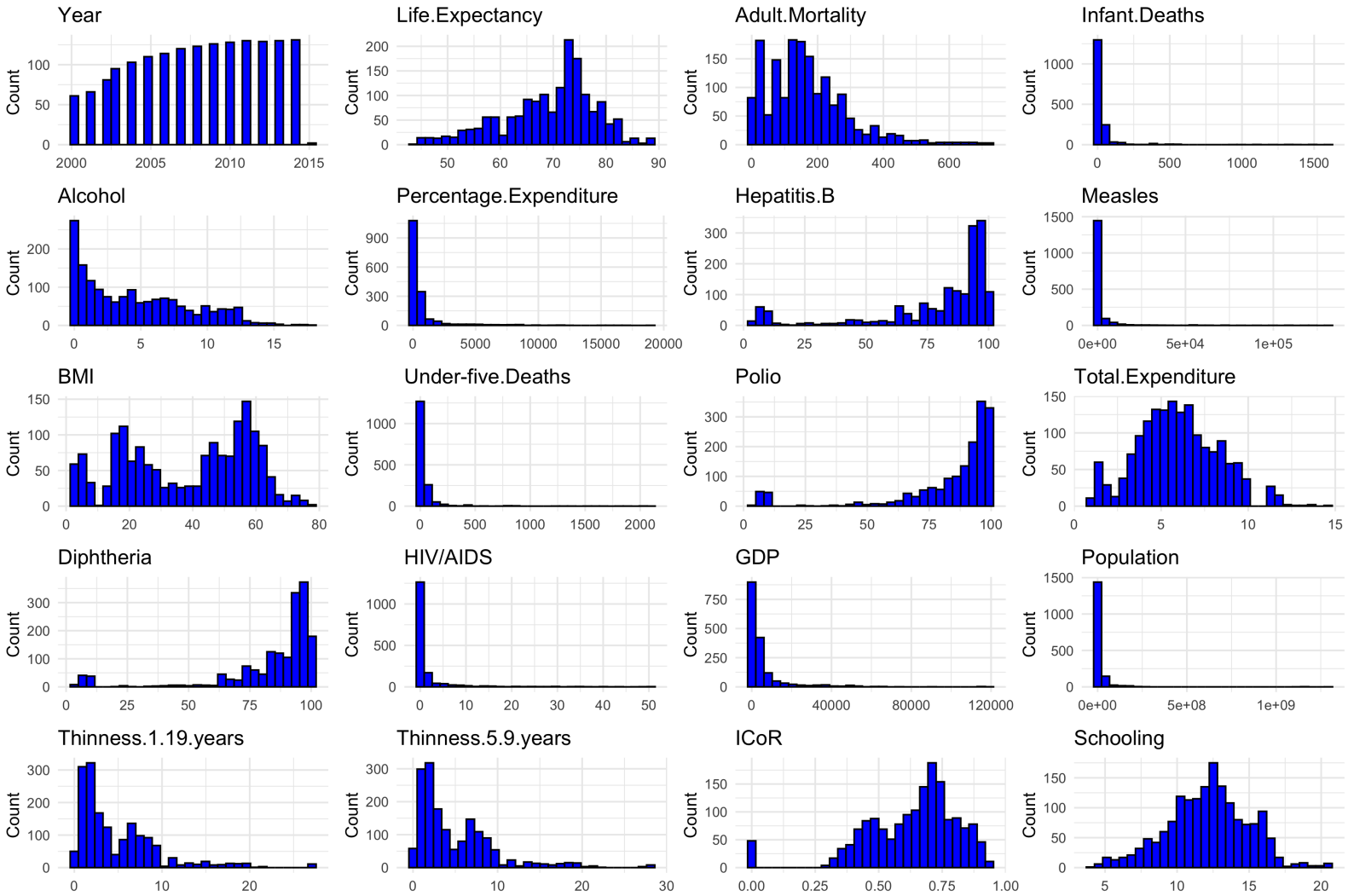
This project aims to

- Describe the variables
- Define theoretical assumptions about the variables and the data
- Test the theoretical assumptions using the proper statistical testing methods
- Compute the correlation of the independent variables with the target variable
- Build models using the target variable and the independent variables

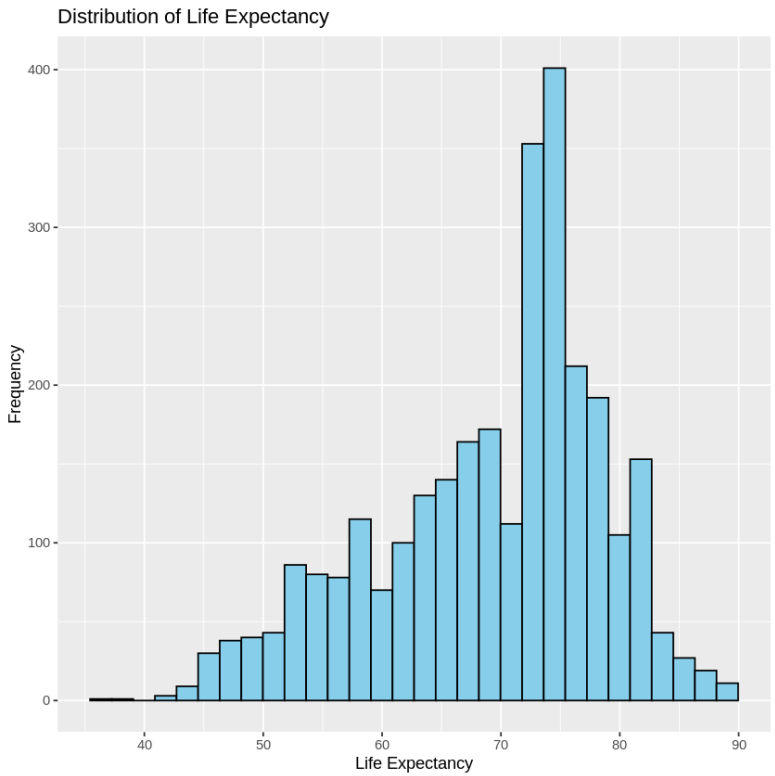
Variables Description

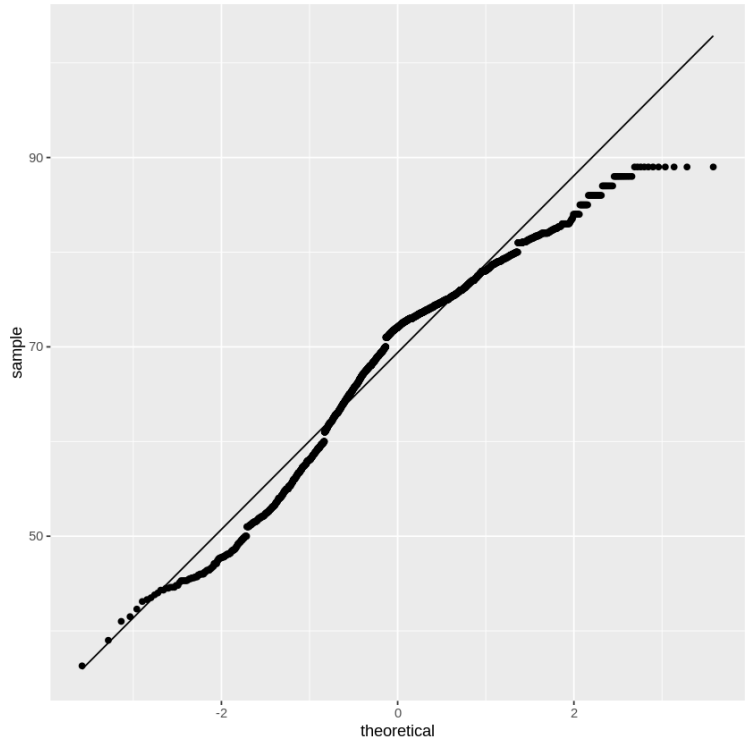
Name	Definition	Categories
Country	Country name	
Year	Year	
Status	Status	Developed Developing
Life Expectancy	Life Expectancy in age	
Adult Mortality(both sexes)	Probability of dying between 15 and 60 years per 1000 population	
Infant Deaths	Number of Infant Deaths per 1000 population	
Alcohol	Alcohol, 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 immunization coverage among 1-year-olds (%)	
Measles	Measles - number of reported 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 1 000 live births 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	1=Low (≤8) 2=Medium (>8 & ≤12) 3=High (>12)

Independent Variables Distribution



Life Expectancy

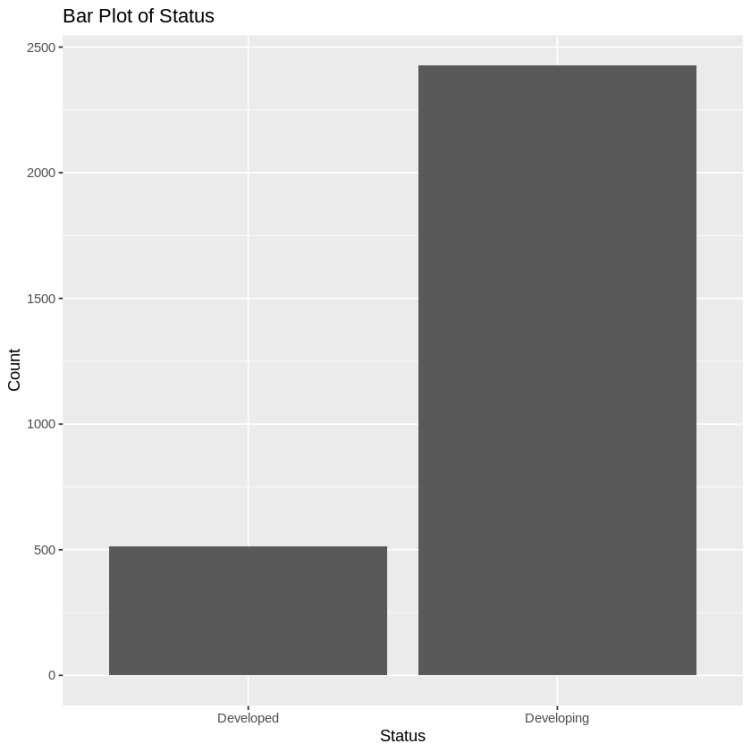


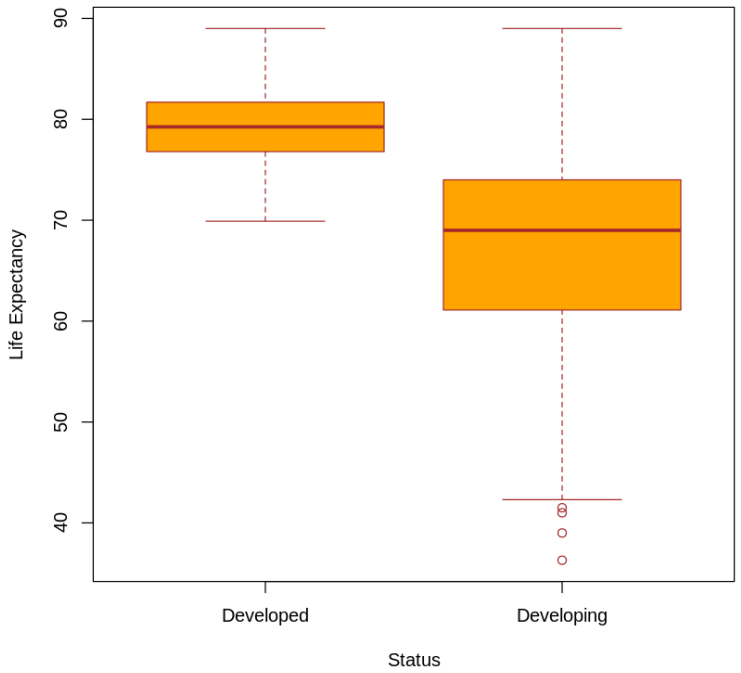


vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2928	69.22493	9.523867	72.1	69.91493	8.59908	36.3	89	52.7	-0.6379506	-0.2380132	0.1760061

Status Variable

Hypothesis: The level of development in a country is expected to influence life expectancy



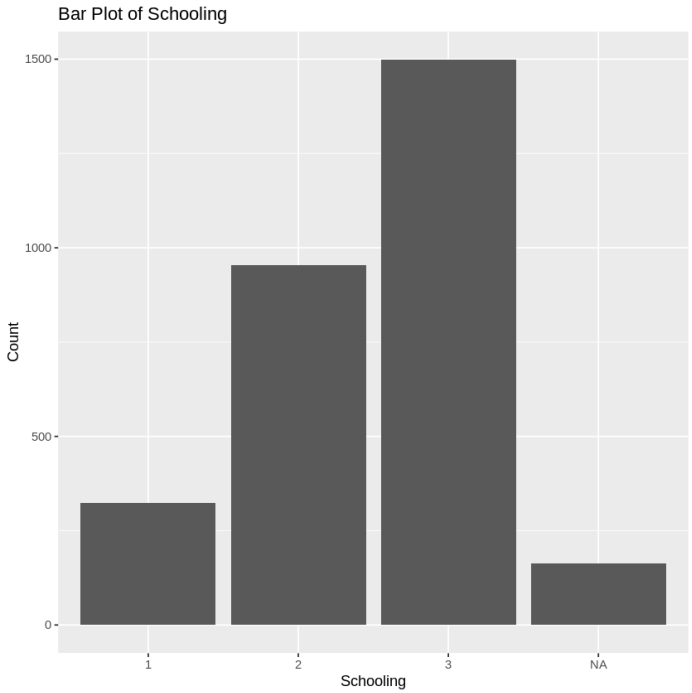


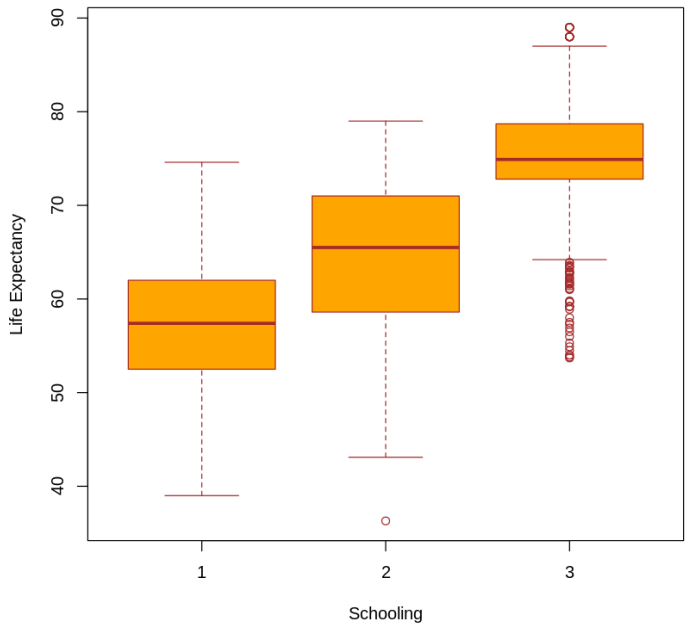

```
 Descriptive statistics by group
group: Developed
   vars   n mean   sd median trimmed  mad  min max range skew kurtosis   se
X1    1 512 79.2 3.93  79.25   79.15 3.63 69.9  89  19.1 0.09    -0.14 0.17
------------------------------------------------------------
group: Developing
   vars    n  mean   sd median trimmed  mad  min max range  skew kurtosis   se
X1    1 2416 67.11 9.01     69   67.81 8.45 36.3  89  52.7 -0.62    -0.37 0.18
```

```
data:  Life Expectancy by Status
W = 1131521, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Schooling Variable

Hypothesis: The number of years of schooling has an impact on life expectancy.






```
 Descriptive statistics by group
group: 1
   vars   n  mean  sd median trimmed  mad min  max range skew kurtosis   se
X1    1 321 57.29 6.7   57.4   57.22 6.97  39 74.6  35.6 0.13    -0.03 0.37
------------------------------------------------------------
group: 2
   vars   n  mean   sd median trimmed  mad  min max range  skew kurtosis   se
X1    1 953 64.13 8.04   65.5   64.73 9.04 36.3  79  42.7 -0.58    -0.32 0.26
------------------------------------------------------------
group: 3
   vars    n  mean   sd median trimmed mad  min max range  skew kurtosis   se
X1    1 1494 75.27 5.28   74.9   75.44 4.3 53.7  89  35.3 -0.47     1.59 0.14
```

```
        Kruskal-Wallis rank sum test

data:  Life Expectancy by Schooling
Kruskal-Wallis chi-squared = 1479.3, df = 2, p-value < 2.2e-16
```

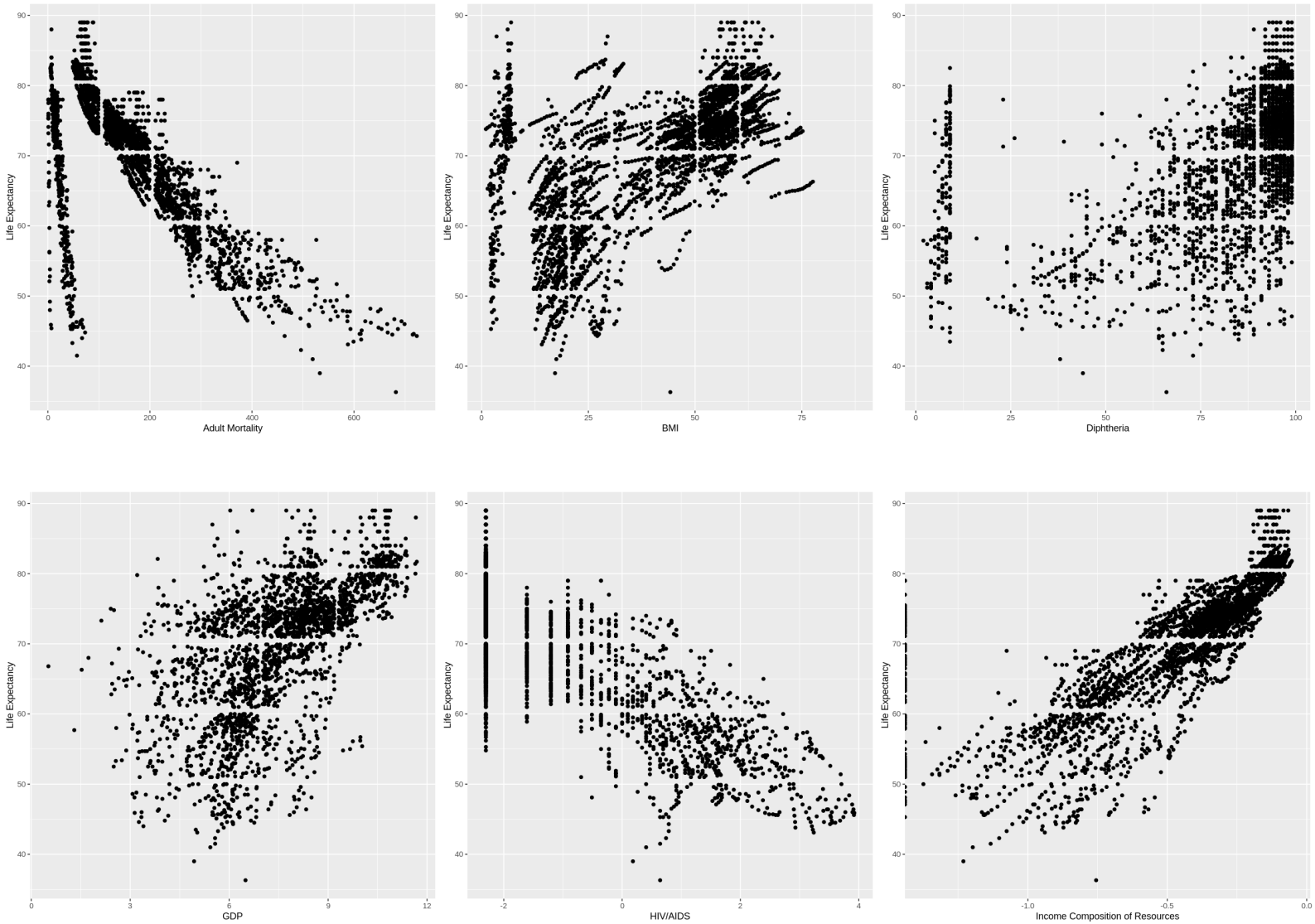
```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  le$`Life Expectancy` and le$Schooling

       Low    Medium
Medium <2e-16 -
High   <2e-16 <2e-16

P value adjustment method: BH
```

Correlation Analysis

We assume that the following variables will have a significant impact to *Life Expectancy*:

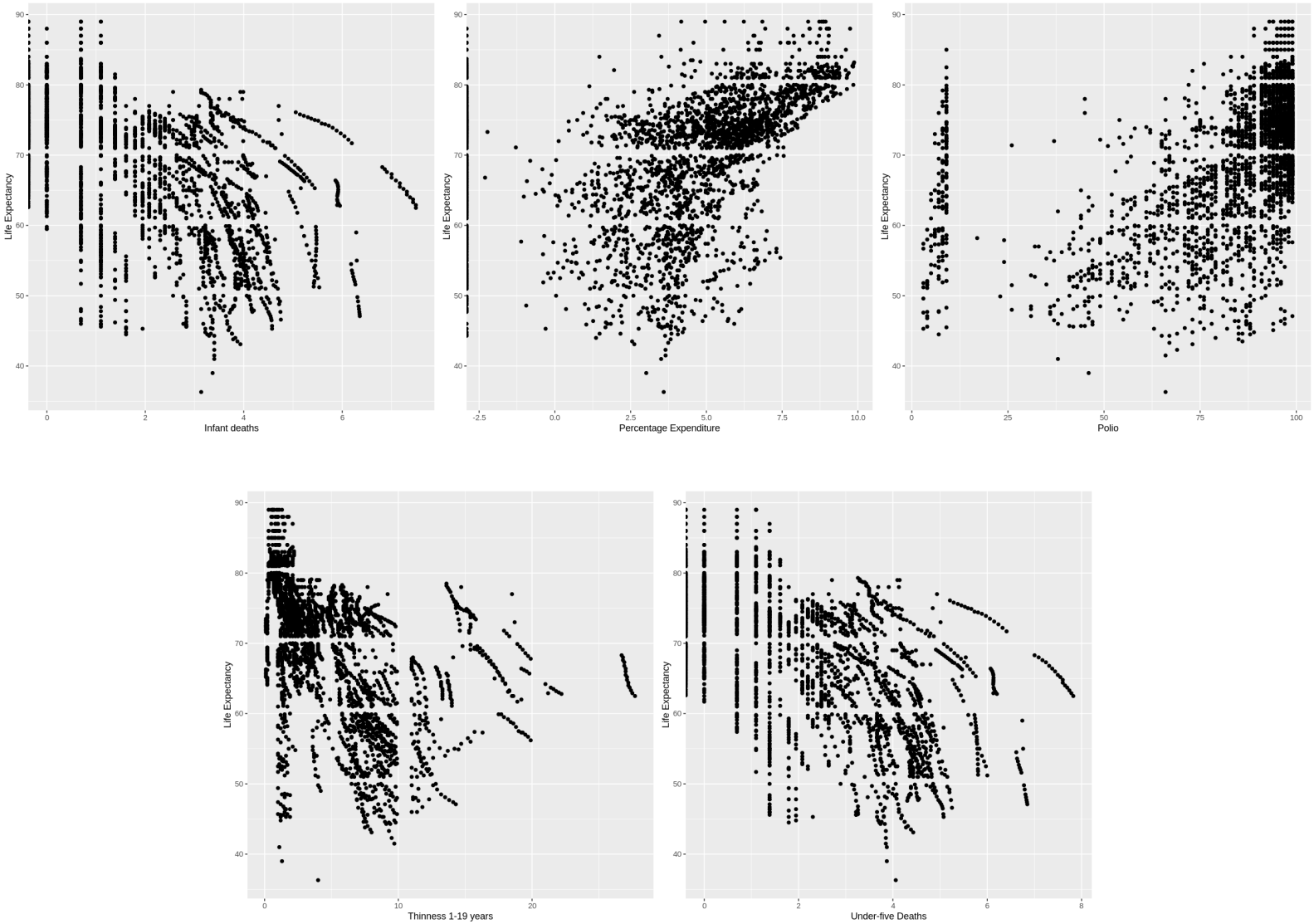
- **Adult Mortality:** High mortality probability in both genders lowers life expectancy, lacking insights into specific causes
- **Percentage Expenditure:** A substantial allocation of a country's GDP per capita to health-related activities suggests a potential link to increased life expectancy
- **Body Mass Index (BMI):** A higher average body mass suggests an unhealthy lifestyle, potentially contributing to lower mortality
- **Total Expenditure:** Government investment in healthcare is anticipated to boost life expectancy
- **GDP:** A higher GDP correlates with improved life expectancy through increased investments in healthcare, education, and living standards
- **Income Composition of Resources:** Higher income links to better healthcare, living standards, and education

Correlation Analysis



1

Correlation Analysis



Correlation Analysis

Table: Correlation table for a limited number of indicators that demonstrated a meaningful and strong correlation with life expectancy

Indicator	**r**
Adult Mortality	-0.65
Infant Deaths	-0.60
Percentage Expenditure	0.43
BMI	0.58
Under-five Deaths	-0.62
Polio	0.53
Diphtheria	0.54
HIV/AIDS	-0.75
GDP	0.64
Thinness 1-19	-0.61
Thinenss 5-9	-0.62
Income Composition of Resources	0.86

Linear Regression Model

Among the indicators showing a correlation with Life Expectancy, only the *Thinness 5-9 years* variable was excluded from the linear regression model using the forward insertion method.

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
<I<chr>>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
	NA	NA	2449	227595.19	11104.127
+ `Income Composition of Resources`	-1	124704.85642	2448	102890.33	9161.060
+ `HIV/AIDS`	-1	34988.01000	2447	67902.32	8144.856
+ Schooling	-2	12145.13925	2445	55757.18	7666.050
+ `Adult Mortality`	-1	8985.83780	2444	46771.34	7237.498
+ Status	-1	2584.40399	2443	44186.94	7100.236
+ Diphtheria	-1	2380.59323	2442	41806.35	6966.552
+ BMI	-1	1393.95677	2441	40412.39	6885.469
+ `Percentage Expenditure`	-1	1055.56633	2440	39356.82	6822.624
+ Polio	-1	434.68216	2439	38922.14	6797.414
+ ThinnessAyears	-1	340.43582	2438	38581.71	6777.891
+ `Under-five Deaths`	-1	49.43720	2437	38532.27	6776.750
+ `Infant Deaths`	-1	1996.95007	2436	36535.32	6648.369
+ GDP	-1	64.19707	2435	36471.12	6646.060

Linear Regression Model

However, the model summary indicated that *GDP* was not statistically different from zero, thereby removing it from the final model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-19.2955  -2.1240  -0.0719   2.1326  15.0807

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.944e+01  5.607e-01 106.017  < 2e-16 ***
StatusDeveloping               -2.239e+00  2.479e-01  -9.031  < 2e-16 ***
`Adult Mortality`              -1.704e-02  7.914e-04 -21.536  < 2e-16 ***
`Infant Deaths`                 9.304e-02  7.907e-03  11.766  < 2e-16 ***
`Percentage Expenditure`        3.891e-04  4.217e-05   9.229  < 2e-16 ***
BMI                             3.651e-02  4.929e-03   7.407 1.72e-13 ***
`Under-five Deaths`            -6.954e-02  5.821e-03 -11.946  < 2e-16 ***
Polio                           2.465e-02  4.455e-03   5.533 3.45e-08 ***
Diphtheria                      2.491e-02  4.489e-03   5.549 3.14e-08 ***
`HIV/AIDS`                     -4.761e-01  1.694e-02 -28.111  < 2e-16 ***
ThinnessAyears                 -8.200e-02  2.255e-02  -3.636 0.000282 ***
`Income Composition of Resources`  8.107e+00  5.905e-01  13.729  < 2e-16 ***
SchoolingMedium                 3.777e+00  2.877e-01  13.131  < 2e-16 ***
SchoolingHigh                   6.794e+00  3.559e-01  19.087  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.859 on 2714 degrees of freedom
  (210 observations deleted due to missingness)
Multiple R-squared:  0.8296,    Adjusted R-squared:  0.8288
F-statistic:  1016 on 13 and 2714 DF,  p-value: < 2.2e-16
```

Linear Regression Model Unique Countries

After keeping only one entry per country in our dataset, we have created the following model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.3340 -1.5518  0.1779  1.5454  8.8663

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      48.992050   2.026358  24.177  < 2e-16 ***
`Income Composition of Resources` 35.930381   2.201826  16.318  < 2e-16 ***
`HIV/AIDS`                       -1.267662   0.262873  -4.822 4.92e-06 ***
`Adult Mortality`                -0.023041   0.004908  -4.695 8.25e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.623 on 103 degrees of freedom
Multiple R-squared:  0.9158,    Adjusted R-squared:  0.9133
F-statistic: 373.4 on 3 and 103 DF,  p-value: < 2.2e-16
```

Evaluating The Model

Table: Model error evaluation for a training set of 80%

Metric	**Model1 Error**	**Model2 Error**
ME	0.09	-0.48
MdE	0.03	-0.083
MAE	2.79	2.31
MdAE	2.22	1.63
MMRE	0.04	0.04
MdMRE	0.03	0.02
MMER	0.041	0.034
MdMER	0.03	0.023

Thank You!