Life Expectancy Prediction using Multiple Linear Regression

for the Bachelor of Science Honours Degree in
Financial Mathematics and Industrial Statistics

By
Y.S. Nimesh
SC/2020/11798

Supervisor:
Dr.D.M Samarathunga

Department of Mathematics
University of Ruhuna
Matara.

2023

DECLARATION

I declare that the presented project report titled, "Life Expectancy Prediction using Multiple Linear Regression" is uniquely prepared by me based on the group project carried out under the supervision of Dr.D.M Samarathunga Department of Mathematics, Faculty of Science, University of Ruhuna, as a partial fulfillment of the requirements of the level II Case Study I course unit MIS2231 of the Financial Mathematics and Industrial Statistics in Department of Mathematics, Faculty of Science, University of Ruhuna, Sri Lanka. It has not been submitted to any other institution or study program by me for any other purpose.

Signature:. . . . . .
Date:. . . . . .

SUPERVISOR'S RECOMMENDATION

I/We certify that this study was carried out by Y.S Nimesh under my/our supervision.

Signature:. . . . . .
Date:. . . . . .
,
Department of Mathematics,
Faculty of Science,
University of Ruhuna.

Acknowledgment

I needed the assistance and counsel of certain respected individuals in order to prepare for our case study, and they deserve our heartfelt appreciation. We would like to express our thanks to Dr. D.M. Samarathunga of the Department of Mathematics, Faculty of Science, and the University of Ruhuna for providing us with solid instructions for the case study through multiple discussions. We'd also want to thank everyone who helped us write this project report, both directly and indirectly.

In addition, we would like to thank the Head of the Department and all of the other Department of Mathematics staff members for their support throughout the Case Study I Course Unit.

Many individuals, particularly our group members, provided helpful comments and recommendations on my report, which inspired us to increase the quality of the project report.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background of the study

Life expectancy is an essential measure for assessing a country's overall population health. According to Max Rose et al. (2013), life expectancy is the most important indicator for assessing population health since it is wider than newborn and child mortality, which exclusively evaluates the death of a specific age group.

Predicting life expectancy is critical for policymakers when developing public policies. It aids in the identification of the country's estimated senior population in need of long-term care. Policymakers can then decide on the necessary healthcare and social welfare changes. Predicting life expectancy provides useful information about the future workforce. A longer life expectancy may need working for a longer period of time in order to maintain financial stability. It results in a more seasoned and competent staff. Meanwhile, constant training and skill development initiatives will be required to keep the workforce adaptable in a dynamic environment.

In actuarial science, life expectancy prediction is common. It is used to calculate premium rates for life insurance plans, for example. Actuaries utilize life expectancy prediction to create new insurance and pension products that consider shifting demographics and market demand.

Life Expectancy

Life expectancy at birth is defined by the World Health Organization (WHO) as "the average number of years that a newborn could expect to live if he or she were to pass

through life exposed to the sex- and age-specific death rates prevailing at the time of his or her birth, for a specific year, in a given country, territory, or geographic area." Landry (n.d.) It is expressed in years.

Life expectancy is a statistic that measures the population's total death rate across all age groups. Life expectancy at birth is determined using sex- and age-specific mortality rates derived from life tables. According to the United Nations, life expectancy at birth numbers correspond to mid-year estimations. They match the applicable United Nations fertility mean quinquennial population forecasts. Life tables are created using available mortality data from civil registration. They are used to build life tables following quality assurance and changes for completeness in preparation for registration. WHO offers a model life table based on a modified logit system derived from around 1800 life tables. Those 1800 life tables were acquired from trustworthy and necessary sources. With a restricted number of input factors, this model is used to plan the life tables required to predict life expectancy. Countries using annual life tables project parameters using a weighted regression model. Recent data is given more weight. The predicted parameters are then entered into the updated logit model, which is loaded with national data to forecast entire life tables. When age-specific death rates are insufficient, the estimated under-5 mortality rates and estimated adult mortality rates, or solely the estimated under-5 mortality rates, are employed in life table derivation using a modified logit model with a global standard (average of 1800 life tables).

Review of Literature on Life Expectancy

A survey of the research on life expectancy as a proxy for a country's health condition is beneficial for investigating the elements that influence it. In this regard, this part is devoted to a review of the literature on the factors of a country's life expectancy. According to Hansen and Strulik (2015), the cardiovascular revolution increased adult life expectancy by around 2 years, which increased higher education enrolment by 7 percentage points across U.S. states. Shin (2013) investigated the effect of a pension scheme on life expectancy and lifetime utility. According to this study, the pension system can either increase or decrease life expectancy, and it is not always true that the pension system promotes lifetime utility.

Hazan (2012) discovered a link between the percentage change in schooling and the change in life expectancy at birth between 1960 and 1990.

For the year 2008, Balan and Jaba (2011) found that the determinants with a positive impact on the Roma population's life expectancy are wages, the number of beds in hospitals, the number of doctors, and the number of readers subscribed to libraries, while the

determinants with a negative impact are the Roma population ratio and the illiterate population ratio..

Halicioglu (2010) analyzed the factors influencing life expectancy in Turkey from 1965 to 2005. The determinants of life expectancy in Turkey have been divided in this study into chosen economic, social, and environmental elements. According to the findings of this study, the primary favorable variables for enhancing longevity were nutrition and food availability. However, smoking was the leading cause of death.

Bergh and Nilsson (2009) used a panel of 92 nations from 1970 to 2005 to examine the association between three components of globalization (economic, social, and political) and life expectancy. They discovered that economic globalization had a very strong beneficial influence on life expectancy, even after adjusting for income, dietary consumption, literacy, the number of physicians, and numerous other characteristics.

Mariani et al. (2008) investigated the dynamics of life expectancy and environmental quality. The findings revealed that environmental factors influenced life expectancy.

Yavari and Mehrnoosh (2006) used multiple regression analysis to examine the influence of socioeconomic determinants on life expectancy. According to the findings of this study, there is a substantial positive relationship between life expectancy as an independent variable and per capita income, health expenditures, literacy rate, and daily calorie consumption. It also indicated that in African countries, there is a negative significant link between life expectancy and the number of people per doctor.Using a modified neoclassical growth model, Leung and Wang (2003) explored the link between health care, life expectancy, and production. They discovered that factors such as money and economic growth had a beneficial influence on life expectancy.

Bernard et al. (2003) looked at the impact of saving on life expectancy. They found that a drop in saving behavior was not associated with an increase in individual life expectancy.

Castello and Domenech (2002) developed a theoretical model in which inequality influences per capita income when people chose to invest in human capital based on their life expectancy. This study discovered that the distribution of education was reliant on the existence of different stable states.

Cervellati and Sunde (2002) explored the link between human capital formation, life expectancy, and the economic development process that the Western world went through when transitioning from an environment of economic stagnation to sustained growth. The findings suggested that technology advancement might possibly promote human capital

formation and life expectancy.

To summarize, the evaluation of studies given reveals that the determinants of life expectancy may be classified as economic, social, environmental, and health-related variables.

Question and Context: Do economic, social, and health factors have a substantial influence on life expectancy in a country? A multiple linear regression study of the most important economic, social, and health-related variables of life expectancy.

1.2 Aims

In our case study, our aim is to develop a multiple linear regression model to predict the life expectancy of a country using key economic, social, and health-related factors using recent data (from 2000 to 2015) obtained from WHO and the United Nations for 193 countries.

1.2.1 Objectives

1. What are the key social, economic, and health-related predictors useful to develop a multiple linear regression model to predict the life expectancy of a country?
2. What is the relative importance of each key predictor in predicting the life expectancy of a country?
3. Is the developed multiple regression model reliable in predicting the life expectancy of a country? What are the limitations?

Chapter 2

Material and Methods

2.1 Linear Regression

- Research Approach:

The quantitative research strategy is the overall research approach presented. Justification: The three research topics in the case study are examples of convergent thinking. That assertion may be substantiated by evaluating each research question as follows.

1. What are the important social, economic, and health-related factors that may be used to create a multiple linear regression model to predict a country's life expectancy? Only the essential variables that are effective in constructing a multiple linear regression model must be identified from all of the social, economic, and health-related indicators in the data set. As a result, by identifying just the helpful predictors, the possibilities are reduced. As a result, this is a convergent reasoning question.

2. What is the relative relevance of each important predictor in forecasting a country's life expectancy? The contribution of each key predictor is established by examining its relative relevance. As a result, this is convergent reasoning.

3. Is the created multiple regression model accurate in estimating a country's life expectancy? What are the constraints?

Examining model dependability and limits leads to inferences about the model's value and necessary modifications, which is an example of convergent reasoning. For the following reasons, quantitative research is preferred for convergent reasoning.

1. Objectivity: Objective measurements, numerical data, and statistical analysis are used in quantitative research. The multiple linear regression model is one of the quantitative approach's tools. It is beneficial to carefully identify and quantify the main predictors.

2. Data-Driven Decision-Making: Quantitative research relies on data-driven decisions, which are critical in convergent reasoning to achieve definite and quantifiable results.

3. Statistical Inference: The quantitative method makes statistical inference easier. The obtained relationship's strength and relevance can be evaluated.

4. Replicability: The quantitative technique is generally clear and reproducible. Others can reproduce the discovery by utilizing the same or comparable data collection and statistical approaches, such as multiple linear regression. This increases the study's credibility

- The conceptual model:
  The following is a generic conceptual model that represents the wider types of predictors under investigation. During the case study, the links between life expectancy and the major predictors chosen from each wider group will be investigated.

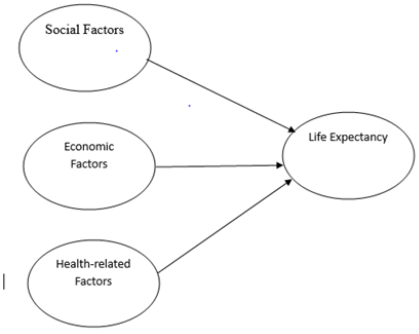

Figure 2.1: General Conceptual Model

- Research Design:

  A correlational research design is proposed in this study. The correlational research design investigates the link between various variables and assesses the degree to which they are related.

  Justification: The goal of our case study is to investigate the link between life expectancy (the dependent variable) and significant social, economic, and health-related factors (independent variables) We utilize multiple linear regression to describe this link since we have a single dependent variable (life expectancy) and numerous independent variables (social, economic, and health-related factors).

  Justification: Multiple linear regression allows for the quantification of the association between many social, economic, and health-related factors and life expectancy. It creates a mathematical model that can be used to forecast a country's life expectancy given specific assumptions. The multiple linear regression model may be recreated using the same or a comparable data set. It boosts the case study's believability. Multiple linear regression is a powerful statistical technique that can be used for a variety of tasks such as prediction, relationship analysis, variable selection, hypothesis testing, and so on, which broadens the scope of the study.

2.1.1 Ordinary Least Square Method

If we consider a simple linear regression model $y = \beta_0 + \beta_1 x$, then

According to the least square method, we aim to minimize the sum of the squared residuals, which is given by the equation:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{2.1}$$

Where:

- ϵ_i represents the ith residual or error term.
- y_i is the observed value for the ith data point.
- $\hat{y}$ is the predicted value of the dependent variable based on the regression model.

This equation represents the core concept of the least squares method in linear regression. We seek to find the values of β_0 and β_1 that minimize this sum, resulting in the best-fitting linear model.

2.2 Multiple Linear Regression Method

Unlocking the Secrets of Multivariate Relationships We set out on a mission to understand the complex dance of variables in the world of data science, where the orchestra of data creates its symphony. We have learned how regression models create narratives that draw connections between our data as we progress through Section

With Multiple Linear Regression, our dependable guide in the region of multivariate riddles, we now delve further into the realm of possibilities.

Imagine this: In this symphonic equation, you have a troupe of independent variables instead of just one, two, or even three. They are the puzzle pieces, and your goal is to harmonize them in order to forecast the peak of a single dependent variable.

Your conductor's baton, Multiple Linear Regression, gives you the freedom to investigate the complex connections between these variables. It involves understanding constellations in the night sky rather than just foretelling how a single star will shine.

We'll explore the world of correlation, coefficients, and constants as we go, unraveling the complex web of relationships that makes up our data-driven environment.

So get ready, brave data scientist! The compass you use to navigate the maze of multi-dimensional data and uncover the mysteries held within the ensemble of variables is known as multiple linear regression. It's time to arrange a symphony of comprehension and awareness.

Chapter 3

Data

3.1 Data Cleaning Process

- Dataset:

  The dataset for the case study was gathered using the Kaggle platform (KUMARRAJARSHI, n.d.). The collection includes observations from 193 countries from 2000 to 2015. The information was gathered from the Global Health Observatory (GHO) data repository on the WHO and United Nations websites. (The appendix contains an extract of the data set as well as instructions for loading the uploaded data set into R studio.)

- Metadata: Additional contextual information such as the source of the data, the time and method of collection, any transformations or preprocessing applied, and any limitations or assumptions associated with the data.

- Data Dictionary:

Table 1: Variable Descriptions

Variable Name	Variable Description	Variable Type	Measurement Units
Country	Country observed	Categorical	No unit
Year	Year observed	Numerical	Calendar year
Status	Developed or developing status	Categorical	No unit
Life Expectancy	Life Expectancy in age	Numerical	years
Adult Mortality	Adult mortality rates	Numerical	No. of deaths per 1000 pc
Infant Deaths	No. of infant deaths	Numerical	No. of infant deaths per 1
Alcohol	Per capita alcohol consumption	Numerical	Liters of pure alcohol
Percentage Expenditure	Health expenditure as a percentage	Numerical	Percentage (%)
Hepatitis B	Hepatitis B immunization coverage	Numerical	Percentage (%)
Measles	No. of reported measles cases	Numerical	No. of reported measles c
BMI	Average body mass index	Numerical	Kilograms per square met
Under-Five Deaths	No. of under-five deaths	Numerical	No. of under-five deaths p
Polio	Polio (Pol3) immunization coverage	Numerical	Percentage (%)
Total Expenditure	Government health expenditure	Numerical	Percentage (%)
Diphtheria	Diphtheria coverage	Numerical	Percentage (%)
HIV/AIDS	HIV/AIDS deaths	Numerical	No. of HIV/AIDS deaths
GDP	Gross Domestic Product per capita	Numerical	USD
Population	Population of the country	Numerical	No. of individuals
Thinness 10-19 years	Prevalence of thinness (10-19 years)	Numerical	Percentage (%)
Thinness 5-9 years	Prevalence of thinness (5-9 years)	Numerical	Percentage (%)
Income Composition	Human Development Index	Numerical	No unit (index ranges fro
Schooling	Average schooling years	Numerical	years

Figure 3.1:

– Preparation for analysis:

Choosing Variables: The dataset contains 22 variables, 20 of which are numerical and two of which are categorical. Before beginning the study, the variables should be filtered and modified based on their observed qualities in the preliminary data preparation.

1. Clean the variable names Variable names that do not accurately represent the variable have been found and renamed (for example, thinness 1-19 years has been replaced with thinness 10-19 years since the variable describes the 10- 19 age range).

2. Remove the variables that do not provide additional information to predict life expectancy. The categorical variable country has too many unique levels. (High cardinality problem). Each country's economic, social, and health-related data have been included as separate observations in other variables. Therefore, the country variable does not provide additional information beneficial to the study. Thus, remove it.

The numerical variable Year is a time series data. Our study focuses on time-independent economic, social, and health-related predictors of life expectancy. Therefore, the Year variable does not provide a benefit to the study thus removing it. First, by removing the "Country" and "Year" variables from the dataset, unnecessary data is removed, making the dataset more streamlined for analysis.

```
## 'data.frame':    1620 obs. of  18 variables:
##  $ Status                       : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ LifeExpectancy               : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ AdultMortality               : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ InfantDeaths                 : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alchol                       : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ PercentageExpenditure        : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ HepatitisB                   : Factor w/ 2 levels "<90% Covered",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ UnderFiveDeaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                        : Factor w/ 2 levels "<90% Covered",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ TotalExpenditure             : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                   : Factor w/ 2 levels "<90% Covered",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ HIV.AIDS                     : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                          : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                   : num  33736494 327582 31731688 3696958 2978599 ...
##  $ Thinness10.19Years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ Thinness1.9Years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ IncomeCompositionOfResources: 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 ...
```

Figure 3.2: variable handling

This process narrows the dimension and concentrates on important variables. Then, numerical variables like ”HepatitisB,” ”Polio,” and ”Diphtheria” are discretized into categorical variables. These variables’ interpretation is made easier by their categorical form, which also makes them easier to analyze and model. These data transformations are essential to getting the dataset ready for future analysis and making it possible to explore the correlations and patterns in the data in a more targeted and efficient way.

Country	0
Year	0
Status	0
Life expectancy	10
Adult Mortality	10
infant deaths	0
Alcohol	194
percentage expenditure	0
Hepatitis B	553
Measles	0
BMI	34
under-five deaths	0
Polio	19
Total expenditure	226
Diphtheria	19
HIV/AIDS	0
GDP	448
Population	652
thinness 1-19 years	34
thinness 5-9 years	34
Income composition of resources	167
Schooling	163

Figure 3.3: No. of Missing Values

3.Missing Data: The summary and context of missing data have been studied. For dealing with missing data, three approaches have been considered. They are as follows:

1. Dropping the observations with missing values.

2. Fill in the missing number from the column's median for a certain nation.

3. Multiple Imputation.

The three linear regression models generated by those three approaches were fitted to training data and tested on test data by comparing RMSEs (Root Mean Square Errors)

Numerical Variable	Total Missing Numerical Values	% of Total Observations
Population	652	22.191967
Hepatitis B	553	18.822328
GDP	448	15.248468
Total expenditure	226	7.692308
Alcohol	194	6.603131
Income composition of resources	167	5.684139
Schooling	163	5.547992
thinness 5-9 years	34	1.157250
thinness 1-19 years	34	1.157250
BMI	34	1.157250
Diphtheria	19	0.646698
Polio	19	0.646698
Life expectancy	10	0.340368
Adult Mortality	10	0.340368
HIV/AIDS	0	0.000000
under-five deaths	0	0.000000
Measles	0	0.000000
percentage expenditure	0	0.000000
infant deaths	0	0.000000
Year	0	0.000000

Figure 3.4: Missing Value Percentages

Missing values are only found in numerical variables. There are no missing values in the two category variables, Country and Year.

Population, Hepatitis B, GDP, Total expenditure, Alcohol, Income composition of resources and Schooling have over 5 precent of missing values.

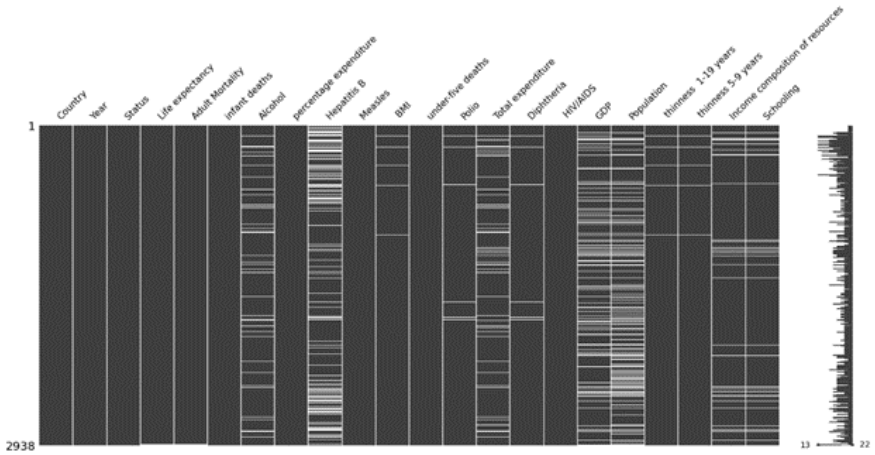

Figure 3.5: Missingness Matrix sorted by Life Expectancy values

Missing values are spread throughout the Life Expectancy (Dependent variable) values range.

The heatmap studies how the presence or absence of one variable influences the presence or absence of another one. Missing values of the following variables have shown strong associations.

1. BMI with 10-19 years of thinness (1)
2. 5-9 year old thinness and BMI (1)
3. Thinness from 5 to 9 years and Thinness from 10 to 19 years (1)
4. Adult Mortality Rates and Life Expectancy (1)

BMI is one of the measures used to assess thinness. This might explain the remarkable correlation. Life tables are used to compute both adult mortality and life expectancy. As a consequence, nations may get these numbers at the same time, resulting in a significant correlation between missing values.

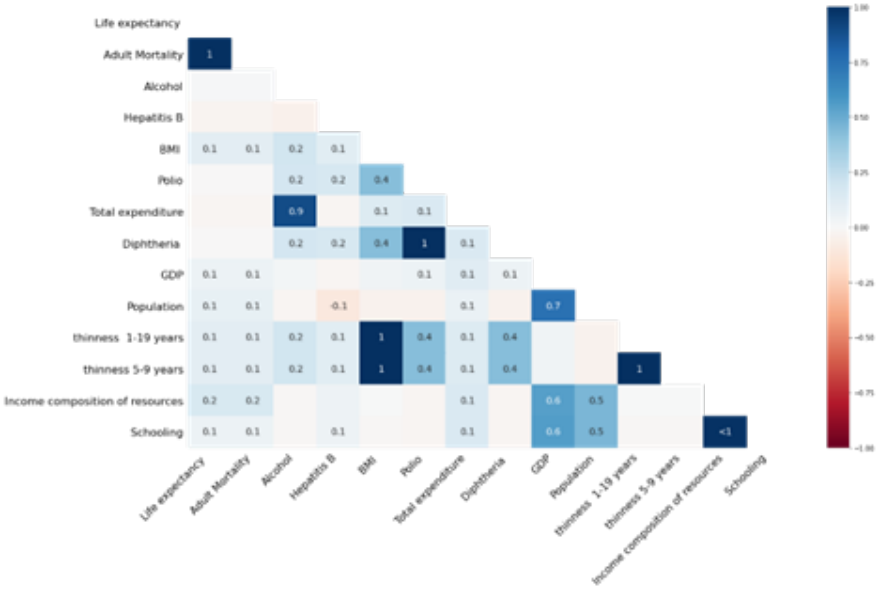

Figure 3.6: Missing Value Heatmap

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1649 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          1649 non-null   category
 1   Year                             1649 non-null   int64
 2   Status                           1649 non-null   category
 3   Life expectancy                  1649 non-null   float64
 4   Adult Mortality                  1649 non-null   float64
 5   infant deaths                    1649 non-null   int64
 6   Alcohol                          1649 non-null   float64
 7   percentage expenditure           1649 non-null   float64
 8   Hepatitis B                      1649 non-null   float64
 9   Measles                          1649 non-null   int64
 10   BMI                             1649 non-null   float64
 11  under-five deaths                1649 non-null   int64
 12  Polio                            1649 non-null   float64
 13  Total expenditure                1649 non-null   float64
 14  Diphtheria                       1649 non-null   float64
 15   HIV/AIDS                        1649 non-null   float64
 16  GDP                              1649 non-null   float64
 17  Population                       1649 non-null   float64
 18   thinness  1-19 years            1649 non-null   float64
 19   thinness 5-9 years             1649 non-null   float64
 20  Income composition of resources  1649 non-null   float64
 21  Schooling                        1649 non-null   float64
dtypes: category(2), float64(16), int64(4)
memory usage: 281.1 KB
```

Figure 3.7: Data set after dropping missing values

After removing all missing-value observations, the data set has 1649 observations for 22 variables. This is a very huge sample size. Population, Hepatitis B, and GDP all have over 400 missing numbers. KDEs were used to compare the distributions of these variables when missing values were not treated (initial) and when the following approaches were employed to address missing data.

1. Observations with missing values are dropped.
2. Fill in the missing value from the column's median for a specified nation.
3. Multiple Imputed

When the fill-in by median approach is performed, the maximum estimated probability density indicates a considerable increase when compared to the estimated probability density for the initial Hepatitis B variable.

```
RMSE (Dropping missing values): 3.5488486362130787
RMSE (Imputing with median): 4.126835490657227
RMSE (Multiple Imputation): 3.9935250260079083
```

When a linear model is formed by removing observations with missing values, the RMSE is minimized. Dropping observations with missing values is therefore a way for dealing with missing data.

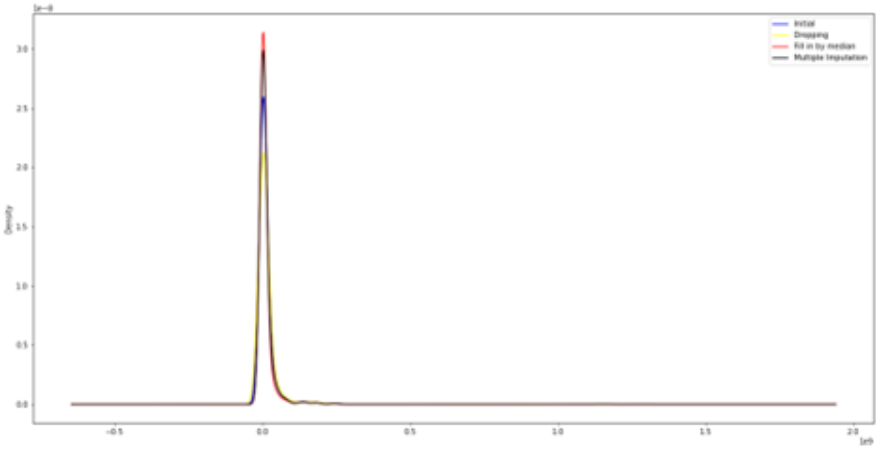

Figure 3.8: KDE for Population Variable

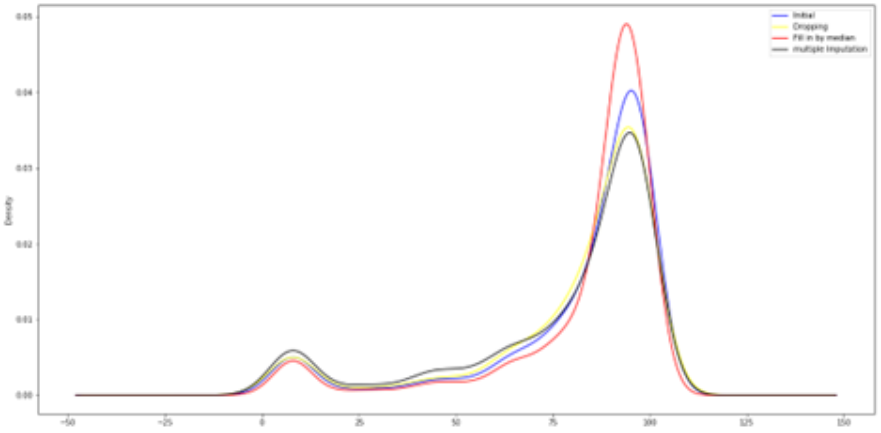

Figure 3.9: KDE for Hepatitis B Variable

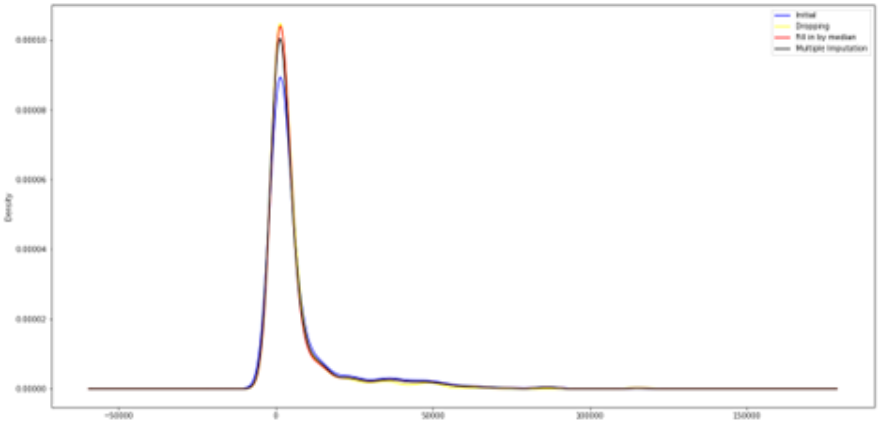

Figure 3.10: KDE for GDP Variable

– 4. Checking Outliers

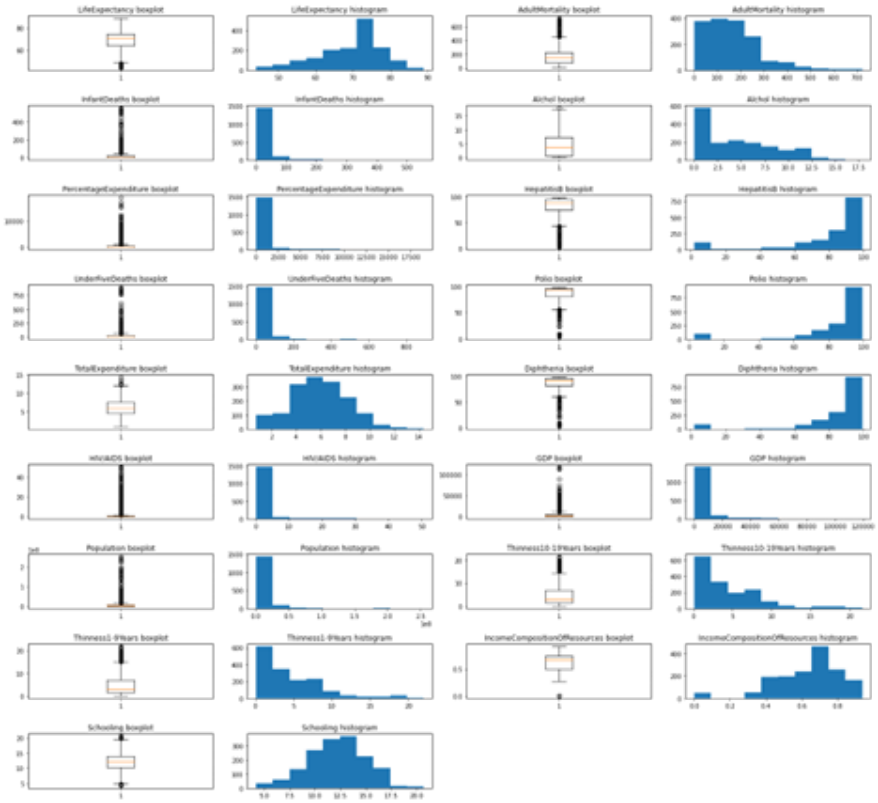

Figure 3.11: Visualization of Outliers

Outliers have been found in the majority of the numerical variables. The dependent variable's distribution appeared to be reasonably symmetric. The majority of the other variables do not have symmetrical distributions. As a result, we can't assume that the variables have normal distributions. However, using the z-score larger than 3 as the criteria, we obtained 597 outlier data under the assumption of normality. At this point, the outliers have not been eliminated since the depicted outliers in the boxplots cannot be deemed disinformation because we already removed all observed unrealistic observations. Several linear models with and without outliers will be developed during the study; finally, the optimal model for our aims will be chosen.

– Metadata: Collaborator: KumarRajarshi
Sources: https://www.who.int/
https://www.worldbank.org/en/home
https://ourworldindata.org/

Chapter 4

Results

4.1 Exploratory data Analysis

The exploratory data analysis (EDA) step is crucial in all data analysis projects, including the life expectancy case study. Its purpose is to get a better understanding of the dataset, find patterns, connections, and potential issues, and lead us to build ideas for future quantitative study. In the case study on life expectancy, we will utilize EDA to better comprehend the dataset and highlight key findings.

1. We employ the following components to conduct exploratory data analysis in order to achieve our goal of creating a multiple linear regression model to predict life expectancy.

2. Statistics in Brief

3. Scatter plots are used to visualize linear relationships.

4. Analysis of Correlation

5. Visualization shows the Distribution of Life Expectancy Across Categorical Variables.

6. ANOVA analysis is used to determine the importance of categorical variables.

7. Visualization of categorical variable connection.

8. The Chi-squared test is used to determine the relationship between categorical variables.

- 1. Summary Statistics: Before deleting the Year variable and discretizing variables linked to vaccine coverage, the five-number summary and mean for each

numerical variable were derived. Following the discretization, the counts for the four categorical variables were included.

```
  Country               Year         Status          LifeExpectancy
Length:1620         Min.   :2000   Length:1620        Min.   :44.00
Class :character    1st Qu.:2005   Class :character   1st Qu.:64.30
Mode  :character    Median :2008   Mode  :character   Median :71.70
                    Mean   :2008                      Mean   :69.26
                    3rd Qu.:2011                      3rd Qu.:74.92
                    Max.   :2015                      Max.   :89.00
 AdultMortality    InfantDeaths        Alchol       PercentageExpenditure
Min.   :  1.00   Min.   :  0.00   Min.   : 0.010   Min.   :    0.00
1st Qu.: 77.75   1st Qu.:  1.00   1st Qu.: 0.730   1st Qu.:   37.37
Median :151.00   Median :  3.00   Median : 3.880   Median :  145.37
Mean   :169.39   Mean   : 24.52   Mean   : 4.580   Mean   :  704.31
3rd Qu.:229.00   3rd Qu.: 22.00   3rd Qu.: 7.383   3rd Qu.:  508.75
Max.   :723.00   Max.   :556.00   Max.   :17.870   Max.   :18961.35
  HepatitisB    UnderFiveDeaths      Polio       TotalExpenditure
Min.   : 2.00   Min.   :  0.00   Min.   : 3.00   Min.   : 0.740
1st Qu.:75.00   1st Qu.:  1.00   1st Qu.:81.00   1st Qu.: 4.410
Median :89.00   Median :  4.00   Median :93.00   Median : 5.840
Mean   :79.41   Mean   : 33.59   Mean   :83.55   Mean   : 5.945
3rd Qu.:96.00   3rd Qu.: 28.25   3rd Qu.:97.00   3rd Qu.: 7.460
Max.   :99.00   Max.   :893.00   Max.   :99.00   Max.   :14.390

  Diphtheria       HIV.AIDS           GDP              Population
Min.   : 2.00   Min.   : 0.100   Min.   :     1.68   Min.   :      419
1st Qu.:82.00   1st Qu.: 0.100   1st Qu.:   461.12   1st Qu.:   214808
Median :92.00   Median : 0.100   Median :  1605.59   Median :  1427020
Mean   :84.18   Mean   : 2.017   Mean   :  5621.57   Mean   : 10879979
3rd Qu.:97.00   3rd Qu.: 0.700   3rd Qu.:  4765.96   3rd Qu.:  7505755
Max.   :99.00   Max.   :50.600   Max.   :119172.74   Max.   :255131116
Thinness10.19Years Thinness1.9Years IncomeCompositionOfResources
Min.   : 0.100     Min.   : 0.100   Min.   :0.0000
1st Qu.: 1.600     1st Qu.: 1.600   1st Qu.:0.5070
Median : 3.000     Median : 3.100   Median :0.6750
Mean   : 4.621     Mean   : 4.672   Mean   :0.6318
3rd Qu.: 6.900     3rd Qu.: 7.000   3rd Qu.:0.7532
Max.   :21.600     Max.   :22.000   Max.   :0.9360
  Schooling
Min.   : 4.20
1st Qu.:10.30
Median :12.30
Mean   :12.13
3rd Qu.:14.00
Max.   :20.70
```

Figure 4.1: Five Number Summary and Mean for Numerical Variables

When we consider the counts for categorical variables, we observed that the dataset comprises a very high number of observations from developing countries than from developed countries. This represents the demography of the world which comprises 152 developing countries at present according to the IMF out of a total of 195 countries.

```
# A tibble: 2 x 3
  Status     count percentage
  <chr>      <int> <chr>
1 Developed    242 14.94%
2 Developing  1378 85.06%

# A tibble: 2 x 3
  HepatitisB    count percentage
  <fct>         <int> <chr>
1 <90% Covered    812 50.12%
2 >=90% Covered   808 49.88%

# A tibble: 2 x 3
  Polio         count percentage
  <fct>         <int> <chr>
1 <90% Covered    687 42.41%
2 >=90% Covered   933 57.59%

# A tibble: 2 x 3
  Diphtheria    count percentage
  <fct>         <int> <chr>
1 <90% Covered    691 42.65%
2 >=90% Covered   929 57.35%
```

Figure 4.2: Counts and Percentages for Categorical Variables

– 2. Scatter plots are used to visualize linear relationships. Scatter plots were used to depict the correlations between the dependent variable life expectancy and the other numerical variables. The visualization revealed that several factors might have linear correlations with Life Expectancy, supporting our objective of creating a multiple linear regression model to predict life expectancy.

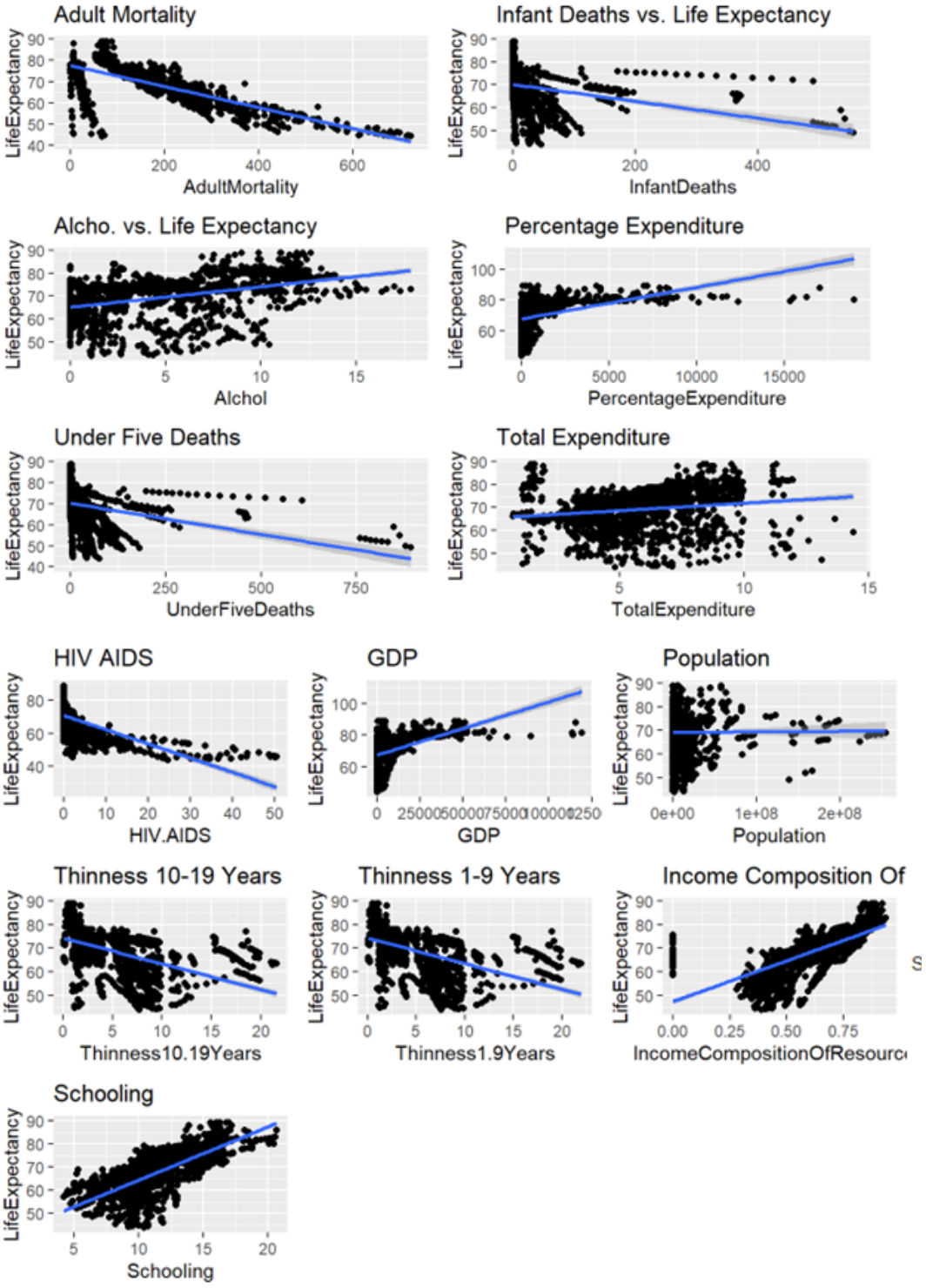

Figure 4.3: Scatterplots of each numerical variable vs. Life Expectancy

– 3.Correlation Analysis: Correlation coefficients were used to quantify the degree and direction of the linear link between each numerical predictor variable and life expectancy. To avoid multicollinearity during the quantitative analysis, strongly correlated numerical predictor variables have been discovered (developing the regression model). Under the premise that the variables are regularly distributed, the Pearson correlation coefficient has been found as an acceptable measure of the linear relationship between variables. Histograms collected for numerical variables during the outlier verification step revealed that this assumption may not be valid because the majority of the histograms are asymmetric. We used the Spearman correlation coefficient as a non-parametric correlation coefficient to quantify the monotone correlations between the variables.

Figure 4.4: Correlation Matrix: Pearson

Since assuming all the variables are normally distributed is not rational, Spearman correlation coefficient has been used. Income Composition of Resources and Schooling show strong positive monotonic relationships with Life Expectancy. HIV-AIDS and Adult Mortality show strong negative monotonic relationships with Life Expectancy. The monotonic relationship between Life Expectancy and Population is not significant at all. Several independent variables like Infant Deaths and Under Five Deaths show strong monotonic relation-

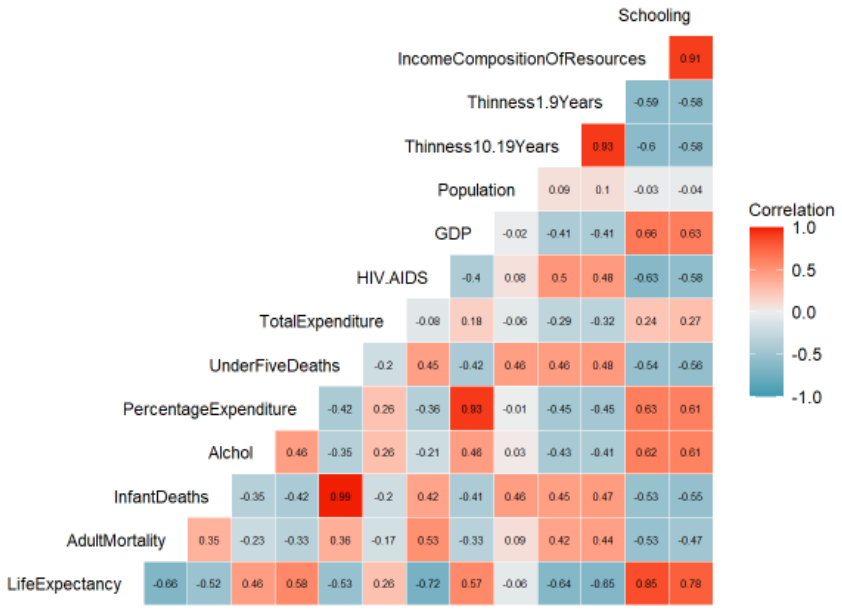

Figure 4.5: Correlation Matrix: Spearman

ships between each other. Though we cannot guarantee these relationships as linear relationships, we have obtained enough motivation to continue our task of developing a multiple linear regression model through visualization results obtained by scatter plots and the results of correlation analysis.

– 4. Visualization of Distribution of Life Expectancy among different categories of categorical variables: The distribution of life expectancy among different categories of each categorical variable has been investigated to determine whether there is a significant difference in life expectancy values between categories.

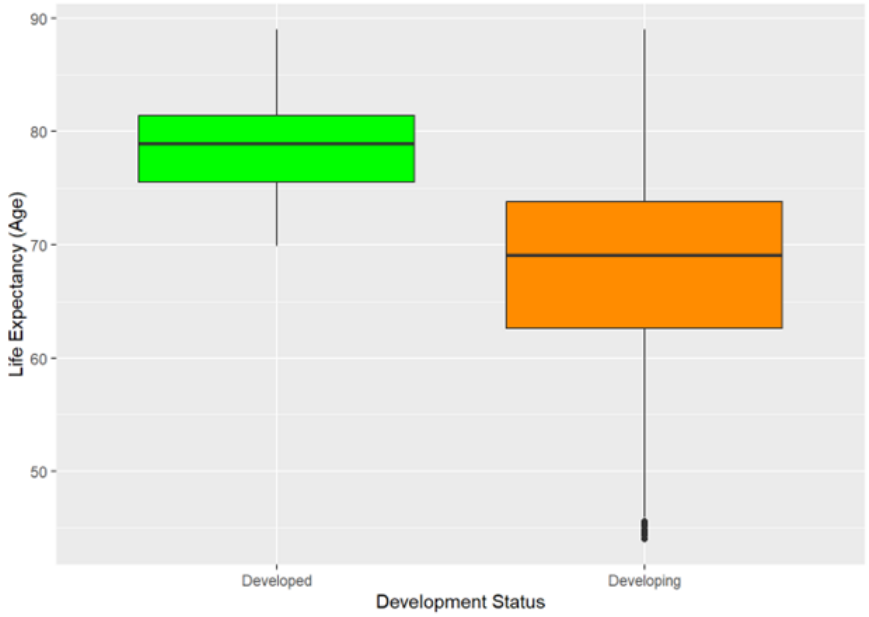

Figure 4.6: Distribution of Life Expectancy among different categories of Status variable

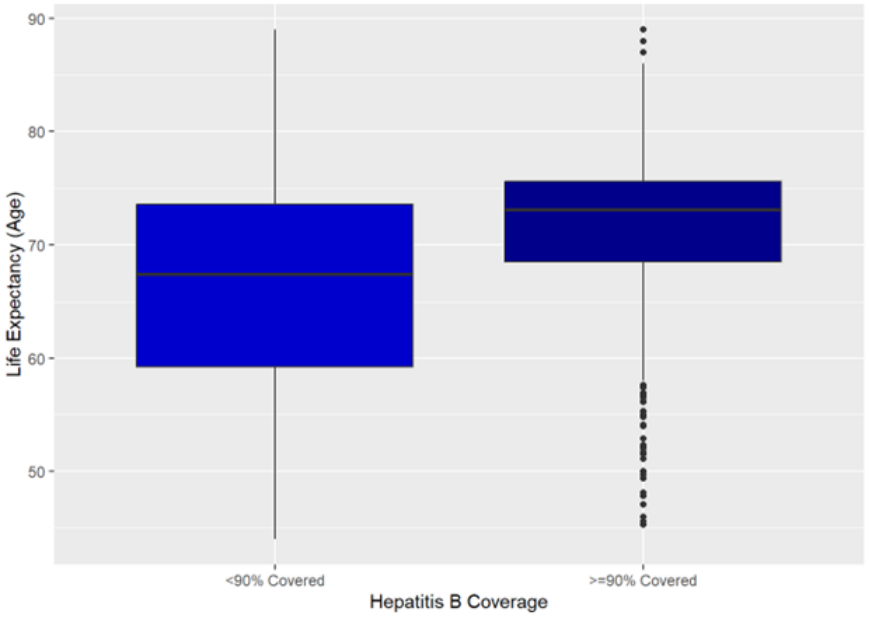

Figure 4.7: Distribution of Life Expectancy among different categories of Hepatitis B variable

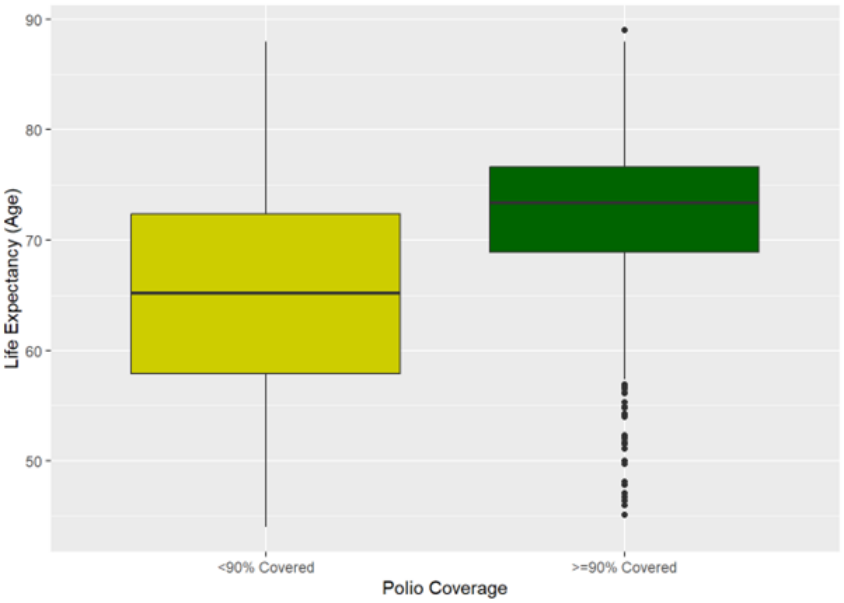

Figure 4.8: Distribution of Life Expectancy among different categories of Polio variable

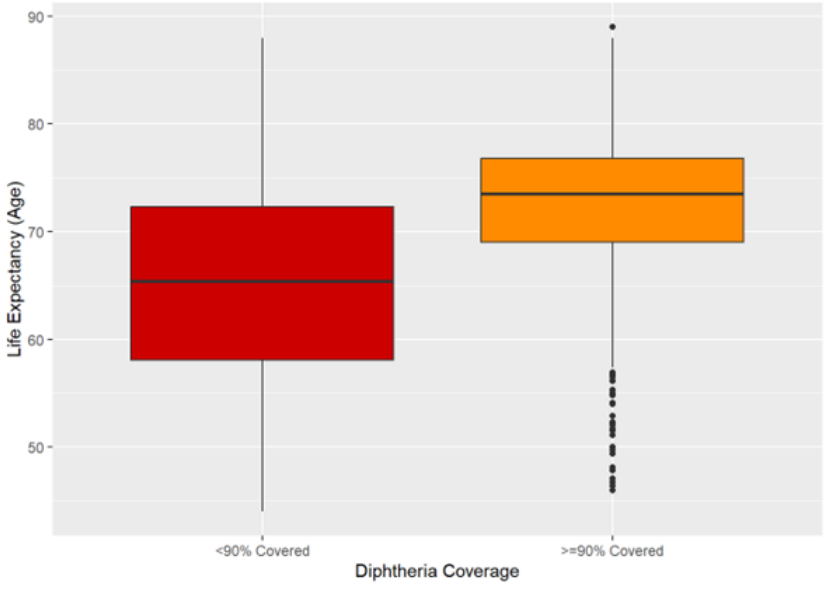

Figure 4.9: Distribution of Life Expectancy among different categories of Diphtheria variable

– 5.ANOVA analysis was used to determine the importance of categorical variables: Two-way ANOVA tests were used to see if there is a statistically significant difference in Life Expectancy among the categories of each categorical variable.

Figure 19: Results of Two-way ANOVA tests for Status Variable

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Status         1  25297   25297   403.6 <2e-16 ***
## Residuals   1618 101414      63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 20: Results of Two-way ANOVA tests for Hepatitis B Variable

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## HepatitisB     1  11747   11747   165.3 <2e-16 ***
## Residuals   1618 114963      71
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 21: Results of Two-way ANOVA tests for Polio Variable

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Polio          1  25611   25611   409.9 <2e-16 ***
## Residuals   1618 101099      62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 22: Results of Two-way ANOVA tests for Diphtheria Variable

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Diphtheria     1  25368   25368     405 <2e-16 ***
## Residuals   1618 101342      63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

anova anaysis for categerical variables

Each ANOVA test yielded a p-value less than 0.05. As a result, the null hypothesis that there is no variation in life expectancy for different categories of the investigated categorical variable may be rejected. As a result, there is a statistically significant variation in life expectancy values across various categories of each categorical variable in the data set. As a result, the ANOVA test confirms the decision made using visualization to employ categorical variables in creating the regression model.

6. Visualization of relationship between categorical variables: The association between a country's Development Status and Immunization Coverage was explored using staked bar charts. Significantly higher vaccine coverage is observed in developed countries. Polio Coverage and Diphtheria Coverage give highly similar patterns which may be due to both these vaccines are parts of the same immunization program as mentioned before in visualizing the distribution of life expectancy among different categories of these variables.

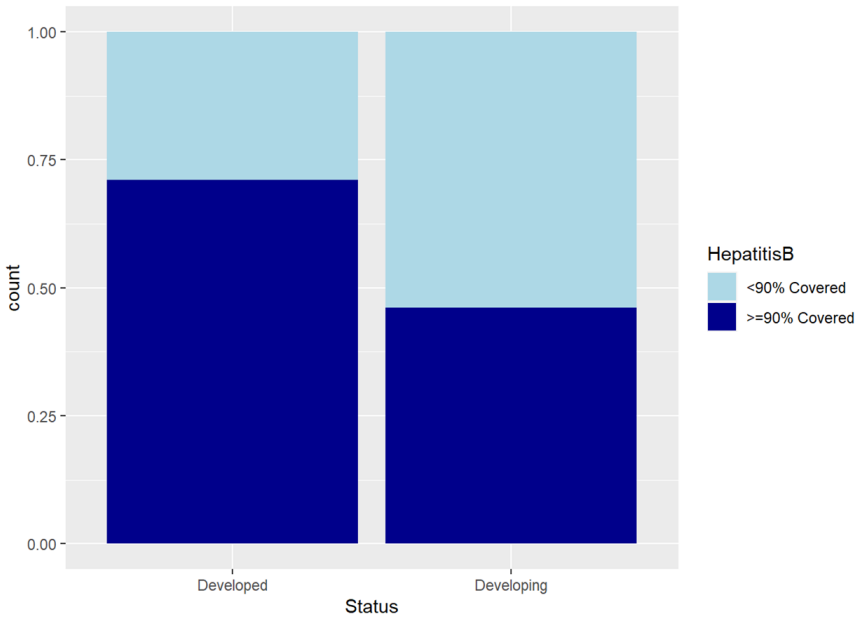

Figure 4.10: Association between Development Status and Hepatitis B Coverage

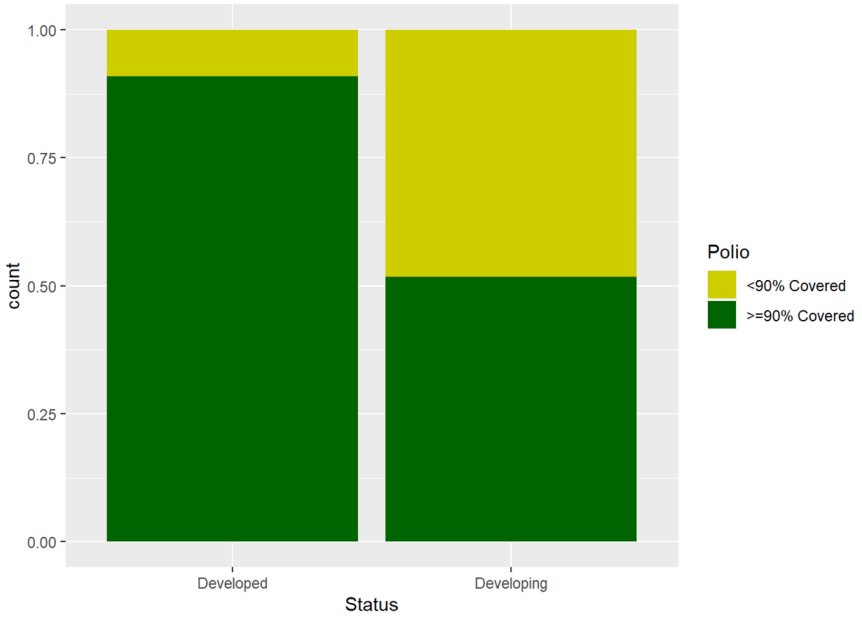

Figure 4.11: Association between Development Status and Polio Coverage

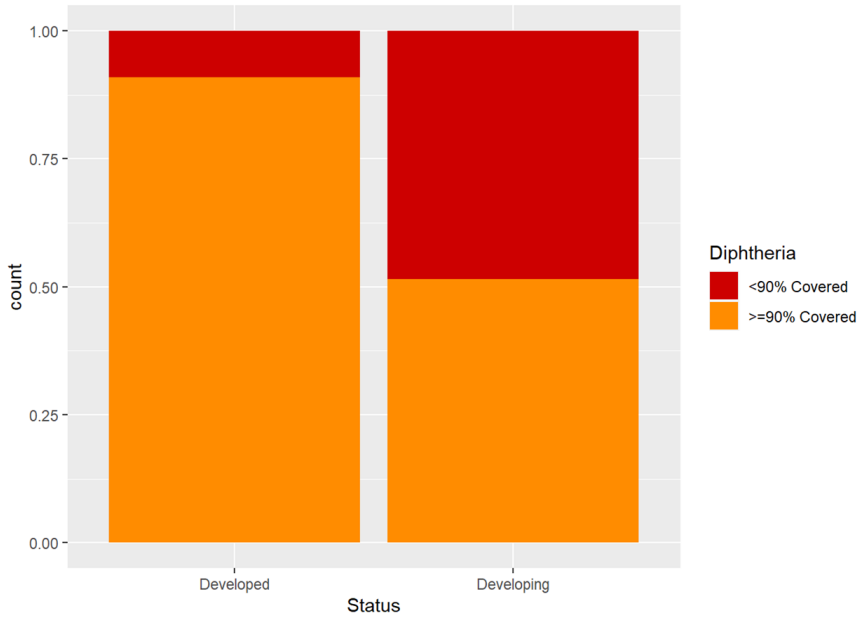

Figure 4.12: Association between Development Status and Diphtheria Coverage

7. The Chi-squared test is used to determine the relationship between category variables: The chi-squared test of independence was performed to determine whether or not there is a significant relationship between development status and vaccine coverage.

All Chi-Squared tests have p-values less than 0.05. As a result, we may infer that there is a strong relationship between a country's Development Status and Immunization Coverage when Hepatitis B, Polio, and Diphtheria coverage are included. The outcome is consistent with the earlier interpretation of the visualization.All Chi-Squared tests have p-values less than 0.05. As a result, we may infer that there is a strong relationship between a country's Development Status and Immunization Coverage when Hepatitis B, Polio, and Diphtheria coverage are included. The outcome is consistent with the earlier interpretation of the visualization.

Figure 26: Chi-Squared Test for Development Status and Hepatitis B Coverage

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(life2$Status, life2$HepatitisB)
## X-squared = 50.144, df = 1, p-value = 1.429e-12
```

Figure 27: Chi-Squared Test for Development Status and Polio Coverage

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(life2$Status, life2$Polio)
## X-squared = 127.7, df = 1, p-value < 2.2e-16
```

Figure 28: Chi-Squared Test for Development Status and Diphtheria Coverage

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(life2$Status, life2$Diphtheria)
## X-squared = 129.42, df = 1, p-value < 2.2e-16
```

Figure 4.13: Chi-Squared Test for Development Status and Hepatitis B Coverage

4.2 Quantitative analysis

- Quantitative analysis:

  The original multiple linear regression model was developed using the Ordinary Least Squares (OLS) method using all of the available 13 numerical independent variables and 4 categorical variables to predict life expectancy.

```
## Call:
## lm(formula = LifeExpectancy ~ ., data = life2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5945  -2.0904  -0.0166   2.2829  12.2546
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   5.677e+01  8.100e-01  70.086  < 2e-16 ***
## StatusDeveloping             -9.688e-01  3.396e-01  -2.853  0.00439 **
## AdultMortality               -1.668e-02  9.587e-04 -17.395  < 2e-16 ***
## InfantDeaths                  1.022e-01  1.107e-02   9.236  < 2e-16 ***
## Alchol                       -9.147e-02  3.373e-02  -2.712  0.00677 **
## PercentageExpenditure         3.180e-04  1.829e-04   1.739  0.08228 .
## HepatitisB>=90% Covered      -6.540e-01  3.131e-01  -2.089  0.03691 *
## UnderFiveDeaths              -7.777e-02  7.907e-03  -9.835  < 2e-16 ***
## Polio>=90% Covered            5.550e-01  4.395e-01   1.263  0.20682
## TotalExpenditure              5.244e-02  4.168e-02   1.258  0.20852
## Diphtheria>=90% Covered       6.048e-01  4.843e-01   1.249  0.21188
## HIV.AIDS                     -4.292e-01  1.812e-02 -23.690  < 2e-16 ***
## GDP                           1.385e-05  2.871e-05   0.483  0.62949
## Population                   -7.547e-10  3.423e-09  -0.221  0.82551
## Thinness10.19Years           -5.793e-02  5.378e-02  -1.077  0.28157
## Thinness1.9Years             -9.991e-02  5.255e-02  -1.901  0.05748 .
## IncomeCompositionOfResources  1.006e+01  8.357e-01  12.034  < 2e-16 ***
## Schooling                     9.042e-01  6.029e-02  14.997  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.608 on 1602 degrees of freedom
## Multiple R-squared:  0.8354, Adjusted R-squared:  0.8337
## F-statistic: 478.4 on 17 and 1602 DF,  p-value: < 2.2e-16
```

Figure 4.14: Initial Regression Model

Interpretation of computed coefficients: The intercept is the estimated value of life expectancy when all of the independent variables are zero. According to our study, such a scenario is unlikely to occur because no country has zero values for all analyzed factors. As a result, the intercept coefficient plays the smallest role. By assessing the other coefficients, we can find various expected correlations between life expectancy and the other parameters. Some of the outcomes are shown below.

1. Life expectancy in developed countries is more likely to be higher than in developing countries.

2. Life Expectancy is negatively related to the variables Adult Mortality, Alcohol, Under Five Deaths, HIV-AIDS, Thinness 10-19 Years, and Thinness 1-9 Years.

3. Percentage Expenditure, Income Composition of Resources, Schooling, GDP, and Total Expenditure are all positively connected to life expectancy.

4. Those with more than 90 % coverage of Polio and Diphtheria vaccines have a higher life expectancy than countries with less coverage.

Unexpected findings have also been made. They do,

1. Infant mortality is linked to life expectancy. The Newborn Deaths variable is expected to have a negative relationship with Life Expectancy since a higher number of baby deaths suggests that many infants die before reaching the age of one, reducing the country's life expectancy.

2. Nations with more than 90% Hepatitis B vaccination coverage are expected to live longer than countries with less than 90% Hepatitis B immunization coverage. The model, on the other hand, predicts the inverse. Increased vaccination coverage is not expected to shorten life expectancy.

3. Population size has an adverse relationship with life expectancy. It's not totally unexpected, given the struggle to get and manage basic healthcare as the population expands. However, the relationship between population size and life expectancy is controversial because, on the one hand, a rise in life expectancy increases population size as death rates decline.

R-Squared Adjusted Interpretation: The model explains 83.37 percent of the variation in the dependent variable Life Expectancy after accounting for the number of independent variables. This is a great beginning point, and it motivates us to keep working on the model.

Interpretation of Predictor Significance: The model has 9 significant variables out of 17 independent variables based on the 5% significance criterion.. To build the optimal regression model, the following stepwise procedures were evaluated in R.

1. Forward motion
2. Reverse Direction
3. Neither Direction

```
## Call:
## lm(formula = LifeExpectancy ~ Schooling + HIV.AIDS + AdultMortality +
##     IncomeCompositionOfResources + PercentageExpenditure + UnderFiveDeaths +
##     InfantDeaths + Thinness1.9Years + Polio + Status + Alchol +
##     HepatitisB, data = life2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.8365  -2.1168   0.0336   2.2520  12.3121
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.688e+01  7.597e-01  74.872  < 2e-16 ***
## Schooling                      9.177e-01  5.971e-02  15.369  < 2e-16 ***
## HIV.AIDS                      -4.280e-01  1.799e-02 -23.791  < 2e-16 ***
## AdultMortality                -1.667e-02  9.573e-04 -17.409  < 2e-16 ***
## IncomeCompositionOfResources   1.012e+01  8.310e-01  12.182  < 2e-16 ***
## PercentageExpenditure          4.097e-04  5.981e-05   6.851 1.04e-11 ***
## UnderFiveDeaths               -7.768e-02  7.859e-03  -9.884  < 2e-16 ***
## InfantDeaths                   1.016e-01  1.095e-02   9.286  < 2e-16 ***
## Thinness1.9Years              -1.532e-01  2.604e-02  -5.883 4.90e-09 ***
## Polio>=90% Covered             9.179e-01  3.076e-01   2.984  0.00289 **
## StatusDeveloping              -9.760e-01  3.380e-01  -2.888  0.00393 **
## Alchol                        -8.384e-02  3.347e-02  -2.505  0.01234 *
## HepatitisB>=90% Covered       -4.623e-01  2.809e-01  -1.646  0.10001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.607 on 1607 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.8337
## F-statistic: 677.5 on 12 and 1607 DF,  p-value: < 2.2e-16
```

Figure 4.15: Summary of Forward Stepwise model

```
Call:
lm(formula = LifeExpectancy ~ Status + AdultMortality + InfantDeaths +
    Alchol + PercentageExpenditure + HepatitisB + UnderFiveDeaths +
    Polio + HIV.AIDS + Thinness1.9Years + IncomeCompositionOfResources +
    Schooling, data = life2)

Residuals:
     Min       1Q   Median       3Q      Max
-17.8365  -2.1168   0.0336   2.2520  12.3121

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5.688e+01  7.597e-01  74.872  < 2e-16 ***
StatusDeveloping             -9.760e-01  3.380e-01  -2.888  0.00393 **
AdultMortality               -1.667e-02  9.573e-04 -17.409  < 2e-16 ***
InfantDeaths                  1.016e-01  1.095e-02   9.286  < 2e-16 ***
Alchol                       -8.384e-02  3.347e-02  -2.505  0.01234 *
PercentageExpenditure         4.097e-04  5.981e-05   6.851 1.04e-11 ***
HepatitisB>=90% Covered      -4.623e-01  2.809e-01  -1.646  0.10001
UnderFiveDeaths              -7.768e-02  7.859e-03  -9.884  < 2e-16 ***
Polio>=90% Covered            9.179e-01  3.076e-01   2.984  0.00289 **
HIV.AIDS                     -4.280e-01  1.799e-02 -23.791  < 2e-16 ***
Thinness1.9Years             -1.532e-01  2.604e-02  -5.883 4.90e-09 ***
IncomeCompositionOfResources  1.012e+01  8.310e-01  12.182  < 2e-16 ***
Schooling                     9.177e-01  5.971e-02  15.369  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.607 on 1607 degrees of freedom
Multiple R-squared:  0.835,  Adjusted R-squared:  0.8337
F-statistic: 677.5 on 12 and 1607 DF,  p-value: < 2.2e-16
```

Figure 4.16: Summary of Backward Stepwise model

```
Call:
lm(formula = LifeExpectancy ~ Status + AdultMortality + InfantDeaths +
    Alchol + PercentageExpenditure + HepatitisB + UnderFiveDeaths +
    Polio + HIV.AIDS + Thinness1.9Years + IncomeCompositionOfResources +
    Schooling, data = life2)

Residuals:
     Min       1Q   Median       3Q      Max
-17.8365  -2.1168   0.0336   2.2520  12.3121

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.688e+01  7.597e-01  74.872  < 2e-16 ***
StatusDeveloping               -9.760e-01  3.380e-01  -2.888  0.00393 **
AdultMortality                 -1.667e-02  9.573e-04 -17.409  < 2e-16 ***
InfantDeaths                    1.016e-01  1.095e-02   9.286  < 2e-16 ***
Alchol                         -8.384e-02  3.347e-02  -2.505  0.01234 *
PercentageExpenditure           4.097e-04  5.981e-05   6.851 1.04e-11 ***
HepatitisB>=90% Covered        -4.623e-01  2.809e-01  -1.646  0.10001
UnderFiveDeaths                -7.768e-02  7.859e-03  -9.884  < 2e-16 ***
Polio>=90% Covered              9.179e-01  3.076e-01   2.984  0.00289 **
HIV.AIDS                       -4.280e-01  1.799e-02 -23.791  < 2e-16 ***
Thinness1.9Years               -1.532e-01  2.604e-02  -5.883 4.90e-09 ***
IncomeCompositionOfResources    1.012e+01  8.310e-01  12.182  < 2e-16 ***
Schooling                       9.177e-01  5.971e-02  15.369  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.607 on 1607 degrees of freedom
Multiple R-squared:  0.835,  Adjusted R-squared:  0.8337
F-statistic: 677.5 on 12 and 1607 DF,  p-value: < 2.2e-16
```

Figure 4.17: Summary of Both Direction Stepwise model

The identical model was generated using all three strategies, yielding an adjusted R-Squared of 83.37 percent, the same as the adjusted R-Squared of the initial model with all independent variables, regardless of their relevance. The backward stepwise model has been chosen for further research because we can avoid errors.

Check for mistakes: The error value of the resulting model was calculated using various methods, including MSE (Mean Squares Error), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Error) (Mean Absolute Percentage Error).

```
:   Method Error.Value
: 1    MSE   12.909459
: 2   RMSE    3.592974
: 3    MAE    2.764272
: 4   MAPE    4.170835
```

Figure 4.18: Error value of the model

- Normality Test

  To demonstrate the normality of the residuals, a histogram and a normal q-q plot were utilized.

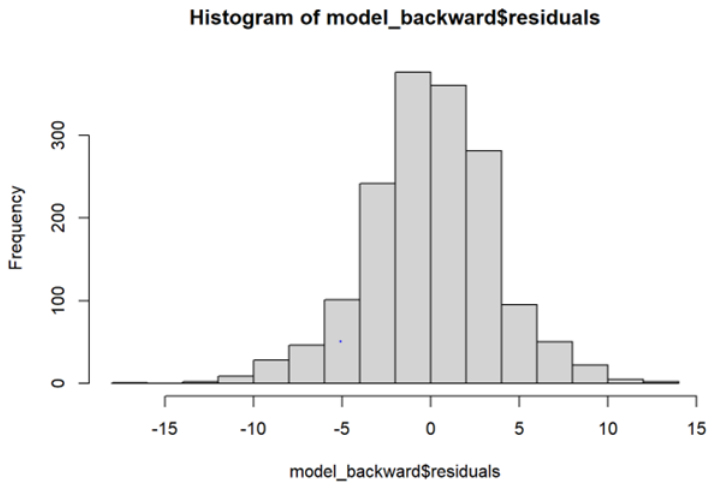

Figure 4.19: Histogram of Residuals

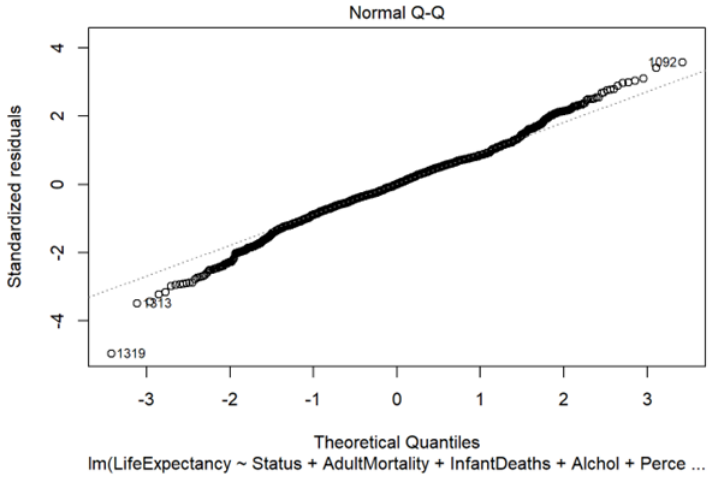

Figure 4.20: Normal Q-Q Plot of Residuals

Statistical Test: The normality of the residuals was statistically assessed using four distinct methodologies.

Test	Statistic	pvalue
Shapiro-Wilk	0.9901	0.0000
Kolmogorov-Smirnov	0.0423	0.0060
Cramer-von Mises	112.7046	0.0000
Anderson-Darling	4.0046	0.0000

Figure 4.21: Normality Tests I

Despite the fact that the histogram and normal q-q plot suggest a normal distribution, the p-values for all four statistical tests were less than 0.05. As a result, we cannot conclude that the normality assumption is met. The confidence intervals and p-values for the regression coefficients are calculated on the assumption that the residuals are normally distributed. Making suitable assumptions is essential for our research. As a result, meeting this criterion is crucial. As a result, an attempt at data transformation has been made. The following are the two transformations that have been tried.

1. Log transformation: Log transformation has been applied to all numerical variables

in the constructed model.

2. Box-Cox Transformation

- Normality Test 2 - using log transformation

```
## Call:
## lm(formula = log1p(LifeExpectancy) ~ Status + log1p(AdultMortality) +
##     log1p(InfantDeaths) + log1p(Alchol) + log1p(PercentageExpenditure) +
##     log1p(UnderFiveDeaths) + Polio + log1p(HIV.AIDS) + log1p(Thinness1.9Years) +
##     log1p(IncomeCompositionOfResources) + log1p(Schooling), data = life2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.260070 -0.026646  0.002148  0.030286  0.196393
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.0140421  0.0241430 166.261  < 2e-16 ***
## StatusDeveloping                 -0.0072372  0.0044159  -1.639 0.101432
## log1p(AdultMortality)            -0.0092501  0.0013485  -6.860 9.81e-12 ***
## log1p(InfantDeaths)               0.0304901  0.0095201   3.203 0.001388 **
## log1p(Alchol)                     0.0037927  0.0019966   1.900 0.057667 .
## log1p(PercentageExpenditure)      0.0065511  0.0008605   7.613 4.53e-14 ***
## log1p(UnderFiveDeaths)           -0.0346076  0.0091292  -3.791 0.000156 ***
## Polio>=90% Covered                0.0038638  0.0030592   1.263 0.206768
## log1p(HIV.AIDS)                  -0.0888512  0.0020549 -43.239  < 2e-16 ***
## log1p(Thinness1.9Years)          -0.0159071  0.0023795  -6.685 3.18e-11 ***
## log1p(IncomeCompositionOfResources)  0.1809882  0.0156467  11.567  < 2e-16 ***
## log1p(Schooling)                  0.0928369  0.0098201   9.454  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05131 on 1608 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8526
## F-statistic: 852.5 on 11 and 1608 DF,  p-value: < 2.2e-16
```

Figure 4.22: Summary of the regression model after log transformation

```
Call:
lm(formula = log1p(LifeExpectancy) ~ Status + log1p(AdultMortality) +
    log1p(InfantDeaths) + log1p(Alchol) + log1p(PercentageExpenditure) +
    log1p(UnderFiveDeaths) + log1p(HIV.AIDS) + log1p(Thinness1.9Years) +
    log1p(IncomeCompositionOfResources) + log1p(Schooling), data = life2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.261245 -0.027165  0.002153  0.030469  0.196025

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       4.0106079  0.0239938 167.152  < 2e-16 ***
StatusDeveloping                 -0.0073639  0.0044156  -1.668 0.095572 .
log1p(AdultMortality)            -0.0092639  0.0013487  -6.869 9.22e-12 ***
log1p(InfantDeaths)               0.0298885  0.0095099   3.143 0.001703 **
log1p(Alchol)                     0.0041153  0.0019806   2.078 0.037884 *
log1p(PercentageExpenditure)      0.0065421  0.0008606   7.601 4.94e-14 ***
log1p(UnderFiveDeaths)           -0.0342076  0.0091254  -3.749 0.000184 ***
log1p(HIV.AIDS)                  -0.0893619  0.0020151 -44.346  < 2e-16 ***
log1p(Thinness1.9Years)          -0.0155503  0.0023631  -6.580 6.33e-11 ***
log1p(IncomeCompositionOfResources)  0.1823140  0.0156143  11.676  < 2e-16 ***
log1p(Schooling)                  0.0947292  0.0097069   9.759  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05132 on 1609 degrees of freedom
Multiple R-squared:  0.8535, Adjusted R-squared:  0.8526
F-statistic: 937.2 on 10 and 1609 DF,  p-value: < 2.2e-16
```

Figure 4.23: Summary of regression model after performing backward stepwise on log transformation

```
Call:
lm(formula = powerTransform(LifeExpectancy, lambda) ~ Status +
    AdultMortality + InfantDeaths + Alchol + PercentageExpenditure +
    UnderFiveDeaths + Polio + HIV.AIDS + Thinness1.9Years + IncomeCompositionOfResources +
    Schooling, data = life2)

Residuals:
    Min      1Q  Median      3Q     Max
-101.64  -13.42   -0.08   14.26   79.42

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.282e+02  4.758e+00  47.966  < 2e-16 ***
StatusDeveloping            -7.194e+00  2.117e+00  -3.399 0.000693 ***
AdultMortality              -1.033e-01  5.995e-03 -17.225  < 2e-16 ***
InfantDeaths                 5.966e-01  6.854e-02   8.704  < 2e-16 ***
Alchol                      -5.095e-01  2.093e-01  -2.435 0.015001 *
PercentageExpenditure        2.894e-03  3.727e-04   7.765 1.44e-14 ***
UnderFiveDeaths             -4.553e-01  4.921e-02  -9.251  < 2e-16 ***
Polio>=90% Covered           3.360e+00  1.340e+00   2.508 0.012242 *
HIV.AIDS                    -2.457e+00  1.126e-01 -21.819  < 2e-16 ***
Thinness1.9Years            -9.990e-01  1.630e-01  -6.128 1.12e-09 ***
IncomeCompositionOfResources 6.322e+01  5.203e+00  12.151  < 2e-16 ***
Schooling                    5.708e+00  3.735e-01  15.283  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.59 on 1608 degrees of freedom
Multiple R-squared:  0.8306, Adjusted R-squared:  0.8294
F-statistic: 716.7 on 11 and 1608 DF,  p-value: < 2.2e-16
```

Figure 4.24: Summary of regression model after performing box cox transformation

Adjusted R-Squared increased to 0.8526 after the log transformation. The log transformation method produces the same Adjusted R-Squared result as the reverse stepwise log transformation method. The Residual standard error of this model, however, is greater than the Residual standard error of the solo log transformation model. The Box-Cox correction resulted in a lower Adjusted R-squared of 0.8294. Nonetheless, all of the independent variables in the Box-Cox transformation model are significant. However, it has a very large residual standard error of 22.59. As a consequence, we elected to reject the Box-Cox transformation model and investigate the normality of the other two models.

Visualization:

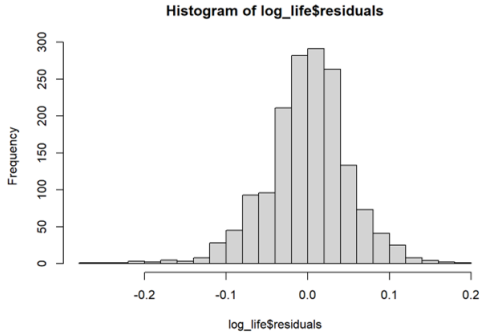

Figure 4.25: Histogram of residuals of log transformation model

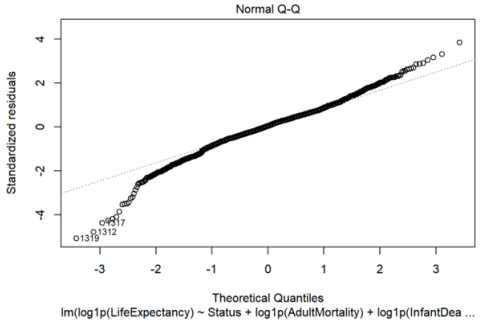

Figure 4.26: Normal q-q plot of residuals of log transformation model

Statical test

The p-values are significantly lower. The residual distribution deviates significantly from the normal distribution. Consider removing outliers. Outliers in the Life Expectancy data were reduced using the boxplot. The total number of observations then dropped to 1583.

Test	Statistic	pvalue
Shapiro-Wilk	0.9782	0.0000
Kolmogorov-Smirnov	0.0552	1e-04
Cramer-von Mises	488.9448	0.0000
Anderson-Darling	7.0603	0.0000

Figure 4.27: Normality tests for log transformation model

Test	Statistic	pvalue
Shapiro-Wilk	0.9781	0.0000
Kolmogorov-Smirnov	0.0528	2e-04
Cramer-von Mises	488.9016	0.0000
Anderson-Darling	6.8393	0.0000

Figure 4.28: Normality tests for log transformation + backward stepwise model

The minimum Life Expectancy value has increased from 44 to 48.40.

```
LifeExpectancy
Min.   :48.40
1st Qu.:64.85
Median :71.80
Mean   :69.80
3rd Qu.:75.00
Max.   :89.00
```

Figure 4.29: Summary of Life Expectancy after removing outliers

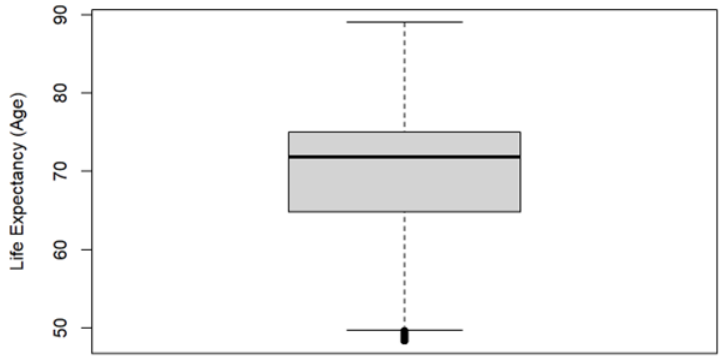

Figure 4.30: Box plot of Life Expectancy after removing outliers

According to the boxplot presentation, the Life Expectancy variable still has a few outliers.The log transformation model and log transformation + backward stepwise model were derived once again after the outliers were removed.

```
Call:
lm(formula = log1p(LifeExpectancy) ~ Status + log1p(AdultMortality) +
    log1p(InfantDeaths) + log1p(Alchol) + log1p(PercentageExpenditure) +
    log1p(UnderFiveDeaths) + Polio + log1p(HIV.AIDS) + log1p(Thinness1.9Years) +
    log1p(IncomeCompositionOfResources) + log1p(Schooling), data = life_outliers_removed)

Residuals:
      Min        1Q    Median        3Q       Max
-0.216885 -0.026301  0.001518  0.028799  0.189646

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.0096064  0.0232057 172.785  < 2e-16 ***
StatusDeveloping               -0.0082184  0.0042235  -1.946  0.05185 .
log1p(AdultMortality)          -0.0093894  0.0013156  -7.137 1.45e-12 ***
log1p(InfantDeaths)             0.0301250  0.0091686   3.286  0.00104 **
log1p(Alchol)                   0.0041025  0.0019253   2.131  0.03326 *
log1p(PercentageExpenditure)    0.0057965  0.0008394   6.906 7.24e-12 ***
log1p(UnderFiveDeaths)         -0.0346908  0.0088049  -3.940 8.51e-05 ***
Polio>=90% Covered              0.0038155  0.0029540   1.292  0.19667
log1p(HIV.AIDS)                -0.0808290  0.0023072 -35.033  < 2e-16 ***
log1p(Thinness1.9Years)        -0.0162872  0.0023013  -7.078 2.20e-12 ***
log1p(IncomeCompositionOfResources)  0.1753201  0.0149735  11.709  < 2e-16 ***
log1p(Schooling)                0.0973670  0.0094545  10.298  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04897 on 1571 degrees of freedom
Multiple R-squared:  0.837,  Adjusted R-squared:  0.8359
F-statistic: 733.4 on 11 and 1571 DF,  p-value: < 2.2e-16
```

Figure 4.31: Summary of log transformation model after outlier removal

The modified R-Squared value is now lower. We will accept this setback because the normality assumption is important to our study.

```
Call:
lm(formula = log1p(LifeExpectancy) ~ Status + log1p(AdultMortality) +
    log1p(InfantDeaths) + log1p(Alchol) + log1p(PercentageExpenditure) +
    log1p(UnderFiveDeaths) + log1p(HIV.AIDS) + log1p(Thinness1.9Years) +
    log1p(IncomeCompositionOfResources) + log1p(Schooling), data = life_outliers_removed)

Residuals:
      Min        1Q    Median        3Q       Max
-0.216115 -0.026763  0.001596  0.029348  0.189310

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.0065435  0.0230891 173.525  < 2e-16 ***
StatusDeveloping               -0.0083446  0.0042233  -1.976  0.04834 *
log1p(AdultMortality)          -0.0094080  0.0013158  -7.150 1.32e-12 ***
log1p(InfantDeaths)             0.0296168  0.0091621   3.233  0.00125 **
log1p(Alchol)                   0.0044461  0.0019072   2.331  0.01987 *
log1p(PercentageExpenditure)    0.0057890  0.0008396   6.895 7.76e-12 ***
log1p(UnderFiveDeaths)         -0.0343734  0.0088034  -3.905 9.84e-05 ***
log1p(HIV.AIDS)                -0.0814066  0.0022640 -35.958  < 2e-16 ***
log1p(Thinness1.9Years)        -0.0159143  0.0022836  -6.969 4.68e-12 ***
log1p(IncomeCompositionOfResources)  0.1766183  0.0149429  11.820  < 2e-16 ***
log1p(Schooling)                0.0991004  0.0093608  10.587  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04898 on 1572 degrees of freedom
Multiple R-squared:  0.8368, Adjusted R-squared:  0.8358
F-statistic: 806.3 on 10 and 1572 DF,  p-value: < 2.2e-16
```

Figure 4.32: Summary of log transformation +backward stepwise model after outlier removal

Following the removal of outliers, run Normality Tests on the residuals of the resultant models.

Test	Statistic	pvalue
Shapiro-Wilk	0.9871	0.0000
Kolmogorov-Smirnov	0.0468	0.0020
Cramer-von Mises	479.4184	0.0000
Anderson-Darling	5.579	0.0000

Figure 4.33: Normality tests for the log transformation model after outlier removal

Test	Statistic	pvalue
Shapiro-Wilk	0.9872	0.0000
Kolmogorov-Smirnov	0.0429	0.0059
Cramer-von Mises	479.3856	0.0000
Anderson-Darling	5.3578	0.0000

Figure 4.34: Normality Tests for log transformation + backward stepwise model after outlier removal

The p-value for the Kolmogorov-Smirnov test for log transformation + backward stepwise model after outlier elimination was 0.0059. Despite being less than 0.05, it is significant when compared to values obtained by models other than the backward stepwise model, which has a Kolmogorov-Smirnov p-value of 0.006. This model, on the other hand, has a higher Adjusted R-squared value and a lower RSE value than that model. After visualizing the residual distribution, we confirmed that our model is more likely to fulfill the normality criteria. As a consequence, we used that model for future research. However, depending on our significance level, the normality assumption has not been fulfilled.

To construct the model, we remove outliers from the life expectancy variable. The recorded minimum life expectancy rose to 48.40 years. However, the missing facts cannot be considered misinformation. Life expectancy in certain nations is less than 48.40 years. As a result, the development model may be erroneous for countries with life expectancies of less than 48.40 years.

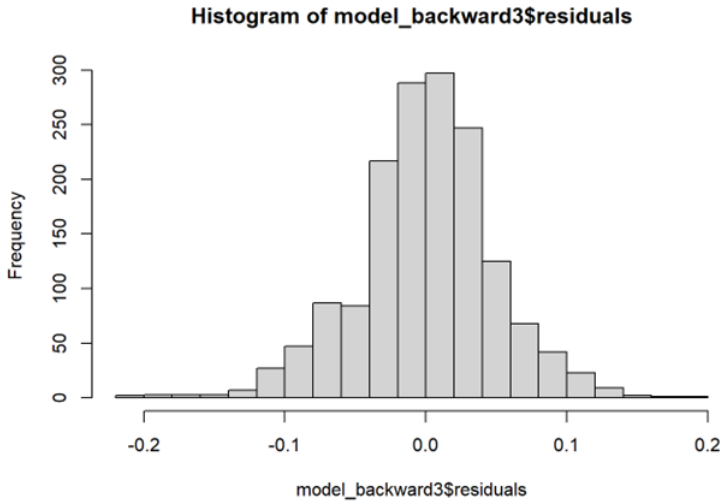

Figure 4.35: Histogram of Residuals for log transformation + backward stepwise model after outlier removal

Homoscedasticity was investigated via visualization and statistical testing

The residuals looked to be spread randomly around the dispersed line. Homoscedasticity looks to have been achieved.

The probability is less than 0.05. The null hypothesis can be refuted as follows: The data are homoscedastic. As a result, the data does not exhibit homoscedasticity. The residual variance does not remain constant. The model does not fulfill the homoscedasticity assumption, according to statistical testing.

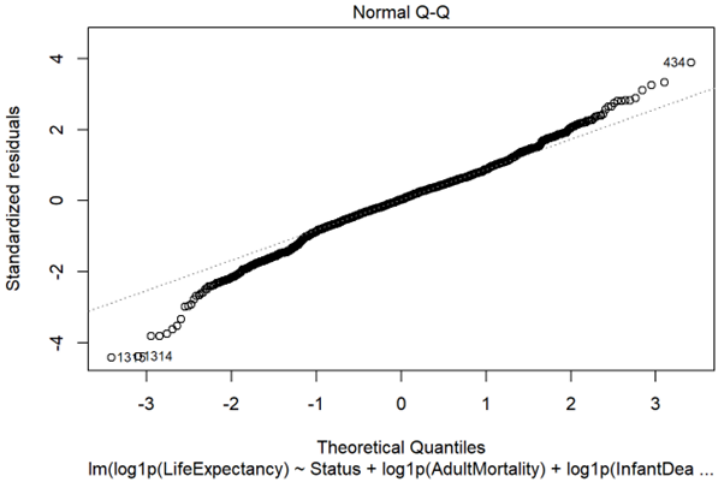

Figure 4.36: Normal q-q plot of Residuals for log transformation + backward stepwise model after outlier removal

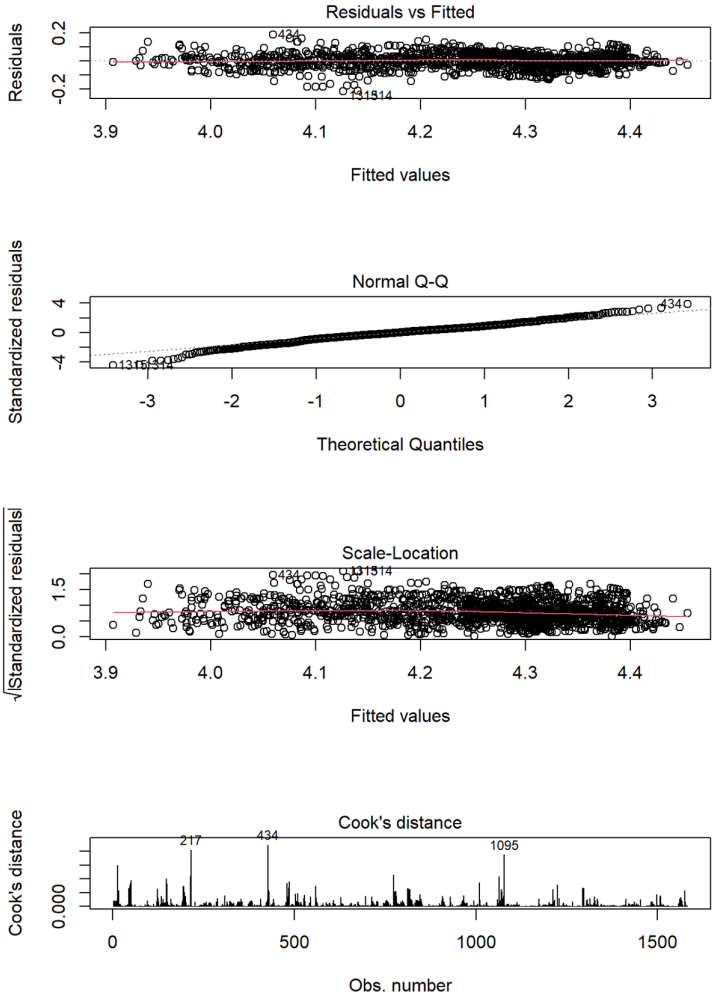
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. number

Figure 4.37:

```
bptest(model_backward3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_backward3
## BP = 56.894, df = 10, p-value = 1.396e-08
```

```
ols_test_breusch_pagan(model_backward3)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                       Data
##  -------------------------------------------------
##  Response : log1p(LifeExpectancy)
##  Variables: fitted values of log1p(LifeExpectancy)
##
##          Test Summary
##  -------------------------------
##  DF            =    1
##  Chi2          =    45.43803
##  Prob > Chi2   =    1.575452e-11
```

Figure 4.38: Statistical Tests for Homoscedasticity

The multicollinearity notion

When evaluating multicollinearity, take VIF values into account. They examine the variation in the inflated regression coefficient induced by collinearity.

```
                          Status                log1p(AdultMortality)
                        1.523917                             1.212189
                log1p(InfantDeaths)                      log1p(Alchol)
                      141.666336                             1.841618
      log1p(PercentageExpenditure)                log1p(UnderFiveDeaths)
                        1.675745                           148.098476
                     log1p(HIV.AIDS)               log1p(Thinness1.9Years)
                        1.624083                             1.682828
log1p(IncomeCompositionOfResources)                    log1p(Schooling)
                        2.279968                             3.148543
```

Figure 4.39:

VIFs are more than 10 for log1p(InfantDeaths) and log1p(UnderFiveDeaths). When compared to log1p, log1p(InfantDeaths) is statistically insignificant (UnderFiveDeaths). As a consequence, create a model by removing log1p (InfantDeaths).

```
Call:
lm(formula = log1p(LifeExpectancy) ~ Status + log1p(AdultMortality) +
    log1p(Alchol) + log1p(PercentageExpenditure) + log1p(UnderFiveDeaths) +
    Polio + log1p(HIV.AIDS) + log1p(Thinness1.9Years) + log1p(IncomeCompositionOfResources) +
    log1p(Schooling), data = life_outliers_removed)

Residuals:
      Min        1Q    Median        3Q       Max
-0.217894 -0.026932  0.001337  0.028403  0.186660

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     3.9960596  0.0229075 174.443  < 2e-16 ***
StatusDeveloping               -0.0066809  0.0042106  -1.587   0.1128
log1p(AdultMortality)          -0.0096204  0.0013178  -7.300 4.54e-13 ***
log1p(Alchol)                   0.0039580  0.0019308   2.050   0.0405 *
log1p(PercentageExpenditure)    0.0057383  0.0008418   6.817 1.32e-11 ***
log1p(UnderFiveDeaths)         -0.0059261  0.0009431  -6.284 4.26e-10 ***
Polio>=90% Covered              0.0033990  0.0029604   1.148   0.2511
log1p(HIV.AIDS)                -0.0830072  0.0022168 -37.444  < 2e-16 ***
log1p(Thinness1.9Years)        -0.0164112  0.0023081  -7.110 1.75e-12 ***
log1p(IncomeCompositionOfResources)  0.1742960  0.0150169  11.607  < 2e-16 ***
log1p(Schooling)                0.1026185  0.0093474  10.978  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04913 on 1572 degrees of freedom
Multiple R-squared:  0.8359, Adjusted R-squared:  0.8348
F-statistic: 800.7 on 10 and 1572 DF,  p-value: < 2.2e-16
```

Figure 4.40: Summary of the created model by removing log1p(InfantDeaths)

R-Squared has been lowered from 0.8348.

All VIF values are less than ten. Homoscedasticity is achieved. Run a normality check on the produced model.

```
Call:
lm(formula = log1p(LifeExpectancy) ~ Status + log1p(AdultMortality) +
    log1p(Alchol) + log1p(PercentageExpenditure) + log1p(UnderFiveDeaths) +
    log1p(HIV.AIDS) + log1p(Thinness1.9Years) + log1p(IncomeCompositionOfResources) +
    log1p(Schooling), data = life_outliers_removed)

Residuals:
      Min        1Q    Median        3Q       Max
-0.217778 -0.027152  0.001509  0.028269  0.186405

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.9935300  0.0228036 175.127  < 2e-16 ***
StatusDeveloping                -0.0068167  0.0042093  -1.619   0.1056
log1p(AdultMortality)           -0.0096335  0.0013179  -7.310 4.24e-13 ***
log1p(Alchol)                    0.0042669  0.0019122   2.231   0.0258 *
log1p(PercentageExpenditure)     0.0057325  0.0008419   6.809 1.39e-11 ***
log1p(UnderFiveDeaths)          -0.0060758  0.0009341  -6.504 1.04e-10 ***
log1p(HIV.AIDS)                 -0.0834899  0.0021768 -38.354  < 2e-16 ***
log1p(Thinness1.9Years)         -0.0160765  0.0022899  -7.021 3.27e-12 ***
log1p(IncomeCompositionOfResources)  0.1754701  0.0149835  11.711  < 2e-16 ***
log1p(Schooling)                 0.1040864  0.0092605  11.240  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04913 on 1573 degrees of freedom
Multiple R-squared:  0.8358, Adjusted R-squared:  0.8348
F-statistic: 889.3 on 9 and 1573 DF,  p-value: < 2.2e-16
```

Figure 4.41: : A summary of the model built by deleting log1p(InfantDeaths) and using a backward stepwise method.

The Kolmogorov-Smirnov test p-value has been reduced from 0.005 to 0.0010. The normality of residuals ensures that the regression coefficients' standard errors are correct. For making population inferences, standard errors are critical. (For instance, confidence intervals and p-values). As a result, inaccurate assumptions are made. To get stable regression coefficients, it is crucial to avoid multicollinearity.. Prioritize normality above avoiding multicollinearity in future analyses using the preceding log transformation + backward stepwise model after outlier elimination.

Status	log1p(AdultMortality)
1.504829	1.208781
log1p(Alchol)	log1p(PercentageExpenditure)
1.840062	1.675019
log1p(UnderFiveDeaths)	log1p(HIV.AIDS)
1.657430	1.492476
log1p(Thinness1.9Years)	log1p(IncomeCompositionOfResources)
1.682016	2.278679
log1p(Schooling)	
3.063055	

Figure 4.42: VIF values obtained by subtracting log1p(InfantDeaths) and applying backward stepwise.

Test	Statistic	pvalue
Shapiro-Wilk	0.9872	0.0000
Kolmogorov-Smirnov	0.0491	0.0010
Cramer-von Mises	479.311	0.0000
Anderson-Darling	5.6132	0.0000

Figure 4.43: Normality tests for the generated model after eliminating log1p(InfantDeaths) and going backwards in time.

Linearity Test

To assess if the independent variables have a linear connection with the dependent variable, use the correlation test. The reported p-values were all less than 0.05. As a consequence, we can rule out the null hypothesis that the correlation coefficient is zero and conclude that a linear relationship between the independent and dependent variables is more likely. As a consequence, the model fulfills the assumption of linearity. At the completion of the inquiry, the resultant multiple linear regression model is described below.

selectrd model is,

```
Call:
lm(formula = log1p(LifeExpectancy) ~ Status + log1p(AdultMortality) +
    log1p(InfantDeaths) + log1p(Alchol) + log1p(PercentageExpenditure) +
    log1p(UnderFiveDeaths) + log1p(HIV.AIDS) + log1p(Thinness1.9Years) +
    log1p(IncomeCompositionOfResources) + log1p(Schooling), data = life_outliers_removed)

Residuals:
      Min        1Q    Median        3Q       Max
-0.216115 -0.026763  0.001596  0.029348  0.189310

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.0065435  0.0230891 173.525  < 2e-16 ***
StatusDeveloping                -0.0083446  0.0042233  -1.976  0.04834 *
log1p(AdultMortality)           -0.0094080  0.0013158  -7.150 1.32e-12 ***
log1p(InfantDeaths)              0.0296168  0.0091621   3.233  0.00125 **
log1p(Alchol)                    0.0044461  0.0019072   2.331  0.01987 *
log1p(PercentageExpenditure)     0.0057890  0.0008396   6.895 7.76e-12 ***
log1p(UnderFiveDeaths)          -0.0343734  0.0088034  -3.905 9.84e-05 ***
log1p(HIV.AIDS)                 -0.0814066  0.0022640 -35.958  < 2e-16 ***
log1p(Thinness1.9Years)         -0.0159143  0.0022836  -6.969 4.68e-12 ***
log1p(IncomeCompositionOfResources)  0.1766183  0.0149429  11.820  < 2e-16 ***
log1p(Schooling)                 0.0991004  0.0093608  10.587  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04898 on 1572 degrees of freedom
Multiple R-squared:  0.8368, Adjusted R-squared:  0.8358
F-statistic: 806.3 on 10 and 1572 DF,  p-value: < 2.2e-16
```

Figure 4.44: Final Regression Model Summary

$$\begin{aligned}\ln(\text{LifeExpectancy}) = {} & 4.0065435 - 0.0083446 \cdot \text{StatusDeveloping} \\ & - 0.009408 \cdot \ln(\text{AdultMortality}) + 0.0296168 \cdot \ln(\text{InfantDeaths}) \\ & - 0.0044461 \cdot \ln(\text{Alcohol}) + 0.005789 \cdot \ln(\text{PercentageExpenditure}) \\ & - 0.0343734 \cdot \ln(\text{UnderFiveDeaths}) - 0.0814066 \cdot \ln(\text{HIV.AIDS}) \\ & - 0.0159143 \cdot \ln(\text{Thinness1.9Years}) + 0.1766183 \cdot \ln(\text{IncomeCompositionOfResour} \\ & + 0.0991 \cdot \ln(\text{Schooling})\end{aligned}$$

The following are the model's estimators' conclusions:

• StatusDeveloping has a positive estimated coefficient, but it is not statistically significant. This shows that the gap in life expectancy between developing and industrialized nations is not substantial.

• The computed ln(AdultMortality) coefficient is negative and statistically significant. This

implies that a greater adult mortality rate corresponds to a shorter life expectancy.

• The computed ln(InfantDeaths) coefficient is negative and statistically significant. This implies that a greater neonatal mortality rate corresponds to a shorter life expectancy.

• The computed ln(Alchol) coefficient is positive, although it is not statistically significant. This shows that there is no link between alcohol use and life expectancy.

• The estimated coefficient for ln(PercentageExpenditure) is positive and statistically significant. This suggests that a higher government expenditure on health is associated with a higher life expectancy.

• The computed ln(UnderFiveDeaths) coefficient is negative and statistically significant. This implies that a greater under-5 death rate corresponds to a shorter life expectancy.

• The computed ln(HIV.AIDS) coefficient is positive, although it is not statistically significant. This shows that there is no link between HIV/AIDS prevalence and life expectancy.

• Ln(Thinness1.9Years) is projected to be negative and statistically significant. This implies that wasting is connected with a decreased life expectancy in children under the age of five.

• Ln(IncomeCompositionOfResources) has a negative and statistically significant coefficient. This shows that a larger resource income composition is related with a lower life expectancy.

• The computed ln(Schooling) coefficient is positive and statistically significant. This implies that more average years of schooling are related with a longer life expectancy.

The final model's modified R-Squared value is 83.58 percent. All of the independent variables in the model are statistically significant. RSE is 0.04898, which is low when compared to the range of the dependent variable.

Chapter 5

Discussion and conclusions

- **Discussion**
  Each of the study goals' findings will be discussed.

  1. What are the key social, economic, and health-related predictors useful to develop a multiple linear regression model to predict the life expectancy of a country?

  Factors of society:
  o Schooling: The average number of years of schooling correlates positively with life expectancy. This is due to the fact that knowledge may assist people in making healthier choices, such as stopping smoking and eating a healthy diet.
  o Income distribution of resources: A more fair income distribution is related with a higher life expectancy. This is due to the fact that poverty can result in poor health outcomes such as malnutrition and a lack of access to healthcare.
  o Life expectancy in wealthy countries is frequently higher than in developing countries. This is due to a variety of factors, including increasing access to healthcare, education, and sanitation.
  o Hepatitis B vaccine coverage: Countries with higher hepatitis B vaccination coverage have shorter life expectancy. This is due to the fact that hepatitis B can develop to liver cancer, which is a leading cause of mortality.
  Economic considerations:
  o Government health spending: Countries that spend more on healthcare have longer life expectancies. This is due to the fact that healthcare may assist to prevent and treat illnesses, hence increasing life expectancy.

o GDP per capita: GDP per capita is a measure of the economic production of a country. Life expectancy is greater in countries with higher GDP per capita. This is since they have greater money to invest in healthcare and other health-promoting variables.

Factors affecting health:

o Adult mortality rate: The number of deaths per 1,000 persons aged 15 and above is referred to as the adult mortality rate. A greater adult mortality rate corresponds to a shorter life expectancy.

o The infant mortality rate is defined as the number of deaths per 1,000 live births. A lower life expectancy is likewise connected with a greater neonatal mortality rate.

o The under-5 mortality rate is the number of deaths per 1,000 children under the age of five. Lower life expectancy is also connected with a greater under-5 death rate.

o Malnutrition is a disorder characterized by a deficiency of necessary nutrients. It can cause a variety of health issues, such as stunting, wasting, and nutritional shortages. These conditions have the potential to reduce life expectancy.

o HIV/AIDS: HIV/AIDS is a life-threatening chronic illness. Countries with higher HIV/AIDS prevalence have shorter life expectancy.

These are only a few of the significant social, economic, and health-related factors that may be utilized to create a multiple linear regression model to forecast a country's life expectancy. The precise predictors that are most essential may differ based on the nation and the available data.

2. What is the relative importance of each key predictor in predicting the life expectancy of a country?

Schooling, economic mix of resources, and HIV/AIDS are the most important predictors of life expectancy. According to research that examined data from 195 nations using a multiple linear regression model. According to the findings, each of these characteristics had a substantial negative influence on life expectancy.

Schooling increases life expectancy by encouraging individuals to make healthier choices, such as stopping smoking and eating a nutritious diet. Income composition of resources has a favorable influence on life expectancy since it indicates that a country's wealth is more equally distributed, which can aid in poverty reduction and enhance access to healthcare. Because HIV/AIDS is a chronic disease that can lead to death, it has a negative influence on life expectancy.

The figure depicts the study's findings. The three most important variables are indicated in red. As you can see, all three of these causes reduce life expectancy. HIV/AIDS has the greatest detrimental impact, followed by adult mortality and finally schooling.

Other indicators, such as government health spending, GDP per capita, and infant mortality, were also shown to be important predictors of life expectancy in the research. These elements, however, were not as important as the three described above.

According to the study's conclusions, the most essential things that can be done to enhance life expectancy in nations throughout the world are boosting access to education, lowering poverty, and limiting the spread of HIV/AIDS. Here are some more information regarding the elements discovered to be relevant in the study:

• Schooling: The study discovered that each extra year of schooling was related with a 0.3-year improvement in life expectancy. This is due to the fact that knowledge may assist people in making healthier choices, such as stopping smoking and eating a healthy diet. It can also help people attain better employment, which can lead to higher wages and improved healthcare access.

• Income distribution of resources: The study discovered that a more fair distribution of income was related with a 0.2-year improvement in life expectancy. This is due to the fact that poverty can result in poor health outcomes such as malnutrition and a lack of access to healthcare. HIV/AIDS: The study discovered that each extra percentage point of HIV prevalence was related with a 0.2-year drop in life expectancy. This is due to the fact that HIV/AIDS is a chronic disease that might result in mortality. The findings of the study are significant because they can assist policymakers in developing policies that can boost life expectancy in nations throughout the world. Policymakers can help people live longer and better lives through boosting access to education, lowering poverty, and regulating the spread of HIV/AIDS.

3. Is the developed multiple regression model reliable in predicting the life expectancy of a country? What are the limitations?

linear regression model does not meet the following assumptions:

1. The residuals do not follow a normal distribution.

2. The residuals are not equally distributed around the line of best fit, as seen in the figure. This can result in underestimated standard errors for regression coefficients, incorrect t-statistics and p-values, and, ultimately, incorrect projected values.

3. The independent variables do not have a multicollinear relationship.

4. This is seen in the graph, which indicates that two of the independent variables have a VIF larger than 10. This suggests that these two variables are highly connected, which might result in understated standard errors and incorrect t-statistics.

The eliminated data as outliers are not implausible. The study excluded data from nations having life expectancy of less than 48.40 years. However, they are genuine nations with real people living in them. If this data is removed from the study, the model may be erroneous in predicting life expectancies of less than 48.40 years.

Because of these breaches of linear regression assumptions, the study's conclusions should be regarded with care. The approach may not be applicable to all nations, and it may be inaccurate in projecting life expectancies of less than 48.40 years.

• Here are some more details regarding the assumptions that were broken:

• Normality of residuals: In order for the t-statistics and p-values to be correct, the residuals must be normally distributed. If the residuals are not regularly distributed, the t-statistics and p-values may be incorrect, leading to incorrect inferences regarding the regression coefficients' significance.

• When two or more independent variables are significantly associated, multicollinearity arises. This can make estimating the specific impacts of each independent variable on the dependent variable challenging.

The study's findings are still relevant, but they should be evaluated with caution owing to breaches of linear regression assumptions. Future research should attempt to overcome these breaches using other approaches, such as robust regression or principal component regression.

- **Conclusion**

  The ordinary least squares approach was used to build a multivariate linear regression model to forecast a country's life expectancy based on chosen social, economic, and health-related variables. With all relevant predictors under the significance threshold of 0.05, the final model has an Adjusted R-squared of 83.58 percent. On the correlations between these characteristics and life expectancy, some evident as well as further debatable conclusions were reached. The final model does not meet the two normal residuals assumptions, and there is no multicollinearity between independent variables. These difficulties can be solved by attempting a regression approach other than the ordinary least squares method or by locating an appropriate data transformation.

  The figure depicts the study's findings. The x-axis represents the independent factors, while the y-axis represents the dependent variable (life expectancy). The image also depicts the line of optimal fit.

  The final model's key determinants of life expectancy are as follows:

  • Schooling: The average number of years of schooling correlates positively with life expectancy. This is due to the fact that knowledge may assist people in making healthier choices, such as stopping smoking and eating a healthy diet.

  • Income distribution of resources: A more fair income distribution is related with a higher life expectancy. This is due to the fact that poverty can result in poor health outcomes such as malnutrition and a lack of access to healthcare.

  • HIV/AIDS: HIV/AIDS is a life-threatening chronic illness. Countries with higher HIV/AIDS prevalence have shorter life expectancy.

  • Adult mortality is defined as the number of deaths per 1,000 adults aged 15 and above. A greater adult mortality rate corresponds to a shorter life expectancy.

  • Thinness (1–5 years): Thinness is a wasting indicator in children under the age of five. Lower life expectancy is connected with a higher prevalence of wasting. The following are the two linear regression assumptions that were not met by the final model:

  • Normality of residuals: In order for the t-statistics and p-values to be correct, the residuals must be normally distributed. The residuals in the final model, however, are not normally distributed. This can lead to incorrect judgments regarding the regression coefficients' significance.

  • Multicollinearity:

  When two or more independent variables are significantly associated, multicollinearity arises. This can make estimating the specific impacts of each independent variable

on the dependent variable challenging. However, the final model's independent variables are not multicollinear. The study's findings are still relevant, but they should be evaluated with caution owing to breaches of linear regression assumptions. Future research should attempt to overcome these breaches using other approaches, such as robust regression or principal component regression. Here are some further information concerning the study's findings:

• The clear observation is that there is a favorable association between education and life expectancy. Education may assist people in making better decisions, such as stopping smoking and eating a nutritious diet. It can also help people attain better employment, which can lead to higher wages and improved healthcare access.

• A clear finding is the positive association between income mix of resources and life expectancy. Poverty can result in negative health effects such as malnutrition and a lack of access to healthcare. Income distribution that is more equal can assist to eliminate poverty and enhance access to healthcare.

• A clear result is the negative link between HIV/AIDS and life expectancy. HIV/AIDS is a life-threatening chronic illness. Countries with higher HIV/AIDS prevalence have shorter life expectancy.

• A clear result is the inverse link between adult mortality and life expectancy. A greater adult mortality rate corresponds to a shorter life expectancy. This is because a greater adult mortality rate suggests that the country has more health issues.

• A less noticeable conclusion is the negative connection between thinness (1-5 years) and life expectancy. Thinness is a wasting indicator in children under the age of five. Stunting can occur as a result of waste, which can have long-term health repercussions.

According to the study's conclusions, the most significant things that can be done to enhance life expectancy in nations throughout the world include expanding access to education, lowering poverty, limiting the spread of HIV/AIDS, and reducing wasting in children under the age of five.

5.1 Summary

In this collaborative scientific article chapters were ordered as follows. First chapter introduces and gives idea about case study topic and related literature reviews. Second chapter describes material and methods regarding the case study questions. Then third chapter elaborates data set and cleaning methods of data set. In Forth one does analysis

and interpret and visualize data set using exploratory and quantitative methods. Final forth chapter explains and summarize our findings and opinions about questions on case then conclusion using obtained best model.

Appendix

Life Expectancy	Alcohol	Schooling	Income Composition of Resources
65.0	0.01	10.1	0.479
59.9	0.01	10.0	0.476
59.9	0.01	9.9	0.470
59.5	0.01	9.8	0.463
59.2	0.01	9.5	0.454
58.8	0.01	9.2	0.448

Life Expectancy	Percentage Expenditure	Total Expenditure	GDP
65.0	71.27962	8.16	584.2592
59.9	73.52358	8.18	612.6965
59.9	73.21924	8.13	631.745
59.5	78.18422	8.52	669.959
59.2	7.097109	7.87	63.53723
58.8	79.67937	9.2	553.3289

Life Expectancy	Thinness 10-19 years	Thinness 5-9 years
65.0	17.2	17.3
59.9	17.5	17.5
59.9	17.7	17.7
59.5	17.9	18.0
59.2	18.2	18.2
58.8	18.4	18.4

Life Expectancy	Adult Mortality	Infant Deaths	Hepatitis B	Measles
65.0	263	62	65	1154
59.9	271	64	62	492
59.9	268	66	64	430

BMI	Under-Five Deaths	Polio	Diphtheria	HIV/AIDS
19.1	83	6	65	0.1
18.6	86	58	62	0.1
18.1	89	62	64	0.1

R codes and python codes :

Github Link

Bibliography

[1] Roser , M., Ortiz-Ospina, E., Ritchie, H. (2013). Life Expectancy. OurWorldInData.org.

[2] Landry, M. M. (n.d.). Life expectancy at birth. Retrieved from World Health Organization: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3131

[3] KUMARRAJARSHI. (n.d.). Life Expectancy (WHO). Retrieved from kaggle: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

[4] Monsef, A. ., Mehrjardi, A. S. . (2015). Determinants of Life Expectancy: A Panel Data Approach. Asian Economic and Financial Review, 5(11), 1251–1257

[5] https://www.macrotrends.net