DSC532 Statistical Learning

Course Professor: Dr Xenia Miscouridou

Life Expectancy - Regression Problem

April 2024

Team members:
Fivos Lympouras
Constantinos Constantinou
Andreas Papadopoulos
Iosif Pintirishis

Contents

1 Introduction

This research will delve into various determinants of life expectancy, including immunization rates, mortality statistics, economic indicators, social considerations and additional health-related metrics. Given that the data contains a diverse range of countries, it will facilitate the identification of key factors that negatively affect life expectancy. Consequently, this analysis will enable countries to pinpoint specific areas requiring attention and intervention, in order to provide a roadmap for effectively improving the longevity of their citizens. Our aim is to predict the life expectancy at birth (response variable) using the other variables as predictors.

2 Our Data Set

The data comes from many sources. Our main file is the Life-Expectancy-Data-Update, which is from Kaggle (Life Expectancy dataset on Kaggle). There are some features that we decided to replace from the orginal source files in order to avoid any wrong inputs. Thus, the features `Adult mortality`, `Alcohol`, `Diptheria`, `HepatitisB`, `Measles`, `Polio` and `Mean_BMI` come from The Global Health Observatory (GHO) data repository under World Health Organization (WHO) which keeps track of the health status as well as many other related factors for all countries. The Population and GDP per capita are from databank.worldbank.

This dataset offers many factors which influence life expectancy (at birth) across 179 countries, from 2010 to 2015. The variables include:

- **Region** and **Country**: Categorical variables that provide geographic classification for comparative regional analysis.
- **Economy_status_Developed** and **Economy_status_Developing**: Encoded variables stating if the country is Developed or Developing.
- **Year**
- **Infant_deaths** and **Under_five_deaths**: Quantitative measures of child mortality, reflecting the quality of pediatric healthcare and general health conditions (infant and children under five deaths per 1000 population).
- **Adult_mortality**: Indicates the probability of dying between the ages of 15 and 60 per 1000 individuals, a marker of adult health in the population.
- **Alcohol**: Records per capita (15+) consumption of alcohol (in litres of pure alcohol), a factor with known health implications.
- **HepatitisB, Diphtheria, Polio**: These variables represent immunization coverage rates among 1-year-olds for various diseases, indicating the reach and effectiveness of public health immunization programs.
- **Measles**: Percentage coverage of the measles vaccine among 1-year-olds, crucial for preventing outbreaks.
- **Mean_BMI**: Provides an age-standardized estimate of the populations average body mass index, relating to nutritional status and health risks.
- **Incidents_HIV**: Reflects the rate of new HIV cases (per 1,000 uninfected population ages 15-49).
- **Thinness_five_nine_years** and **Thinness_ten_nineteen_years**: Represent the prevalence of underweight among children and adolescents, indicating potential nutritional deficiencies.
- **Schooling**: Measures the average number of years of education received for individuals aged 15-64, a socioeconomic factor linked to health literacy.
- **GDP**: Stands for Gross Domestic Product per capita, an economic indicator that correlates with a countrys health infrastructure and services (current US$).
- **Population**: Population of each country for a specific year.

3 Data Pre-processing

3.1 Convert Data types

In the data preprocessing, a new column `Economy_status` was created to categorize countries based on their economic status as either 'Developed' or 'Developing', using the two columns of `Economy_status_Developed` and `Economy_status_Developing`. Additionally, this new column was converted to a factor, and the data types of the `Population` and `GDP` columns were converted to numeric to facilitate further analysis. This kind of preprocessing is essential for ensuring that the dataset's variables are correctly formatted and meaningful for statistical modeling and data analysis tasks.

3.2 Missing values

Furthermore, we checked for missing values and found that the features `Adult_mortality`, `Alcohol`, `Diphtheria`, `HepatitisB`, `Measles`, `Polio`, `Mean_BMI`, `Population` and `GDP` had a lot of missing values. Some of the missing values are due to the fact that there are some countries which they have different names in the Global Health Observatory (GHO) data repository compare to our main file. So we created a function, changed the names of these countries and imported again some files. After doing that, the missing values reduced a lot. We have the following number of missing values for each column:

- **HepatitisB**: 550
- **Alcohol**: 38
- **GDP**: 37
- **Polio**: 26
- **Diphtheria**: 24
- **Measles**: 24
- **Mean_BMI**: 16

- **Adult_mortality**: 16
- **Population**: 16

All the predictors except `HepatitisB` have missing values less than 2% which is a small percentage. We have 19.2% of missing values for `HepatitisB`. We usually remove the feature if the missing values are above 10%. At first we kept it in order to see if it is among the important features. If that was the case we were going to fit our model with and without `HepatitisB`. Later, we will see that `HepatitisB` isn't among the important predictors.

We assume that the missing values above follow the Missing Completely at Random (MCAR) mechanism, meaning that the missingness of data is completely random and does not depend on any observed or unobserved data.

Due to the big percentage of missing data for `HepatitisB` we choose to fill the missing values of `HepatitisB` using Predictive Mean Matching from mice package. Predictive Mean Matching (PMM) is a method for imputing missing values that is widely used in statistical analysis, particularly when dealing with continuous variables. It is a non-parametric imputation technique that falls under the umbrella of multiple imputation methods. The core idea behind PMM is to impute missing values by matching them with observed (non-missing) values from similar cases. For a better imputation we will use the `PMM` in a scaled dataset. As a result our variable `HepatitisB` will be in a scaled form.

In addressing the small percentage of missing values within our dataset for the other variables who have missing values, we choose to impute using the median of the training set. The median is a robust measure of central tendency that is less influenced by outliers and skewed distributions, making it particularly suitable for our dataset as we are going to see later in our Exploratory Data Analysis (EDA). By using only the training set for the imputation, we ensure that our model remains uninformed by the test set, preserving the validity of our predictive performance assessments.

3.3 Split of the data set

Here, it is important to mention that the split of the data into training and test set was based on the year. Our train set contains the data from 2000 to 2012 and our test set from 2013 to 2015. Since we had to do with time we prefer to split the data this way to test how well our model can predict future cases. The division of our dataset into training and test sets has yielded a nearly ideal ratio of 81.25% to 18.75%, which is close to the preferred 80-20 split commonly used in model validation.

4 Exploratory Data Analysis (EDA)

4.1 Life Expectancy

Life expectancy is our response variable. Human life expectancy is a statistical measure of the estimate of the average remaining years of life at a given age. In our project, Life expectancy refers to life expectancy at birth meaning the average number of years a newborn is expected to live, assuming that current mortality rates at each age will remain constant throughout the life of that newborn. This feature is used as a measure of the overall health of a population.

The minimum life expectancy recorded is 39.40 years, and the maximum is 83.80 years. The median, indicating the middle value of the dataset, is 71.40 years, which is relatively close to the mean life expectancy of 68.86 years.

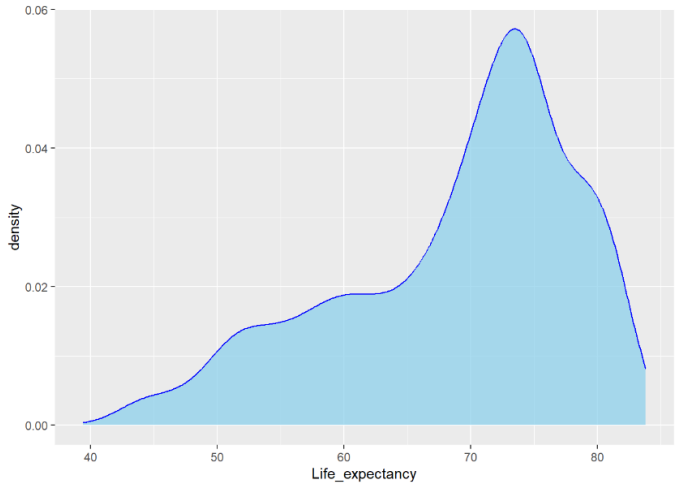


Figure 1: The density of Life Expectancy

From our calculations Life Expectancy has a left skewed distribution as it has a negative skewness and $Mode > Median > Mean$. We can confirm this in figure 1.

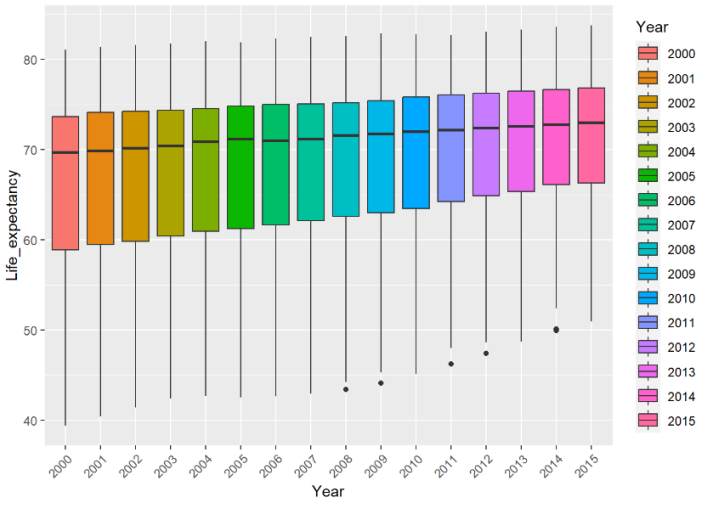


Figure 2: Boxplots of Life Expectancy from 2000 to 2015

The figure 2 presents a series of box plots of life expectancy from 2000 to 2015, showing a general upward trend in median life expectancy over the years. The spread within each year remains consistent, suggesting that while improvements are being made, the relative differences in life expectancy across countries persist. The indicated outliers in the figure 2 come from the country Lesotho in Africa. The life expectancy in Lesotho is significantly below the global average.

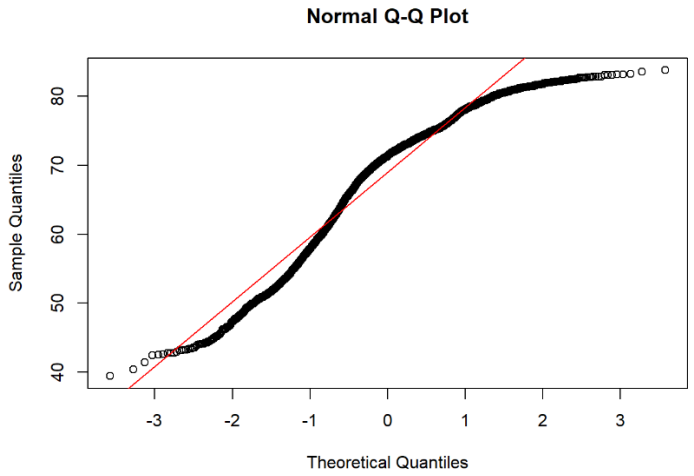


Figure 3: Life Expectancy QQ-plot

The density graph (figure 1) and the QQ-plot (figure 3) suggests that our respone variable does not follow Normal Distribution. We can confirm that because on both Shapiro and Lilliefors tests we reject the Null Hypothesis ($p-value < 0.05$, with $\alpha = 0.05$ the significance level). The Null Hypothesis says that it follows a Normal Distribution.

This deviation from normality is important to consider in the next steps of our statistical analyses and model selection. Thats why, we will try to make the distribution of our response variable at least roughly Normal. After trying many transformations we selected $Life\_expectancy^3$. The below density graph (figure 4) shows an improvement of our response variable regarding if it follows a Normal distribution.

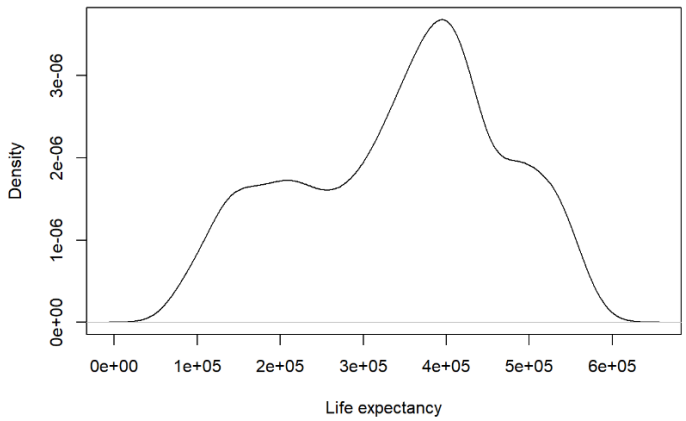


Figure 4: Density of $Life\_expectancy^3$

4.2 Economy Status

The database has one variable that categorizes countries into two groups: Developed vs Developing countries. According to World Trade Organization, each country defines itself as `Developed` or `Developing`. Therefore, it is challenging to categorize countries. UN has a list dated 2014 that for analytical purposes classifies countries as `developed`, `in transition` and `developing` economies. Countries that have economies in transition have similar characteristics to the countries that are categorized as developed or developing countries. Countries have been grouped according to their Gross National Income per capita.

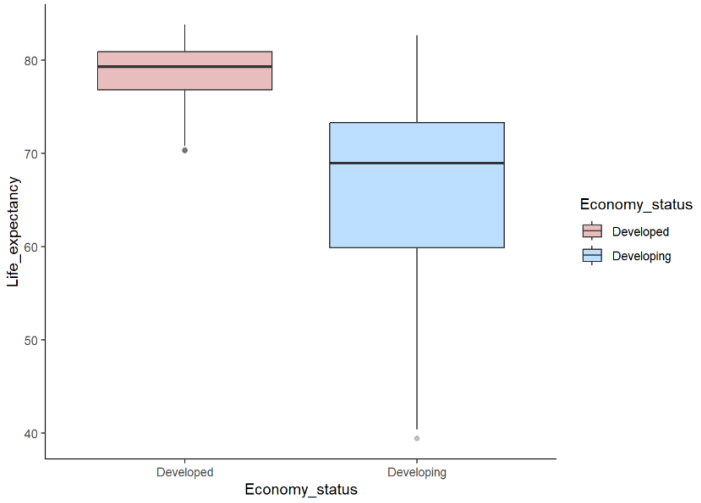


Figure 5: Life expectancy - Economy status Box plot

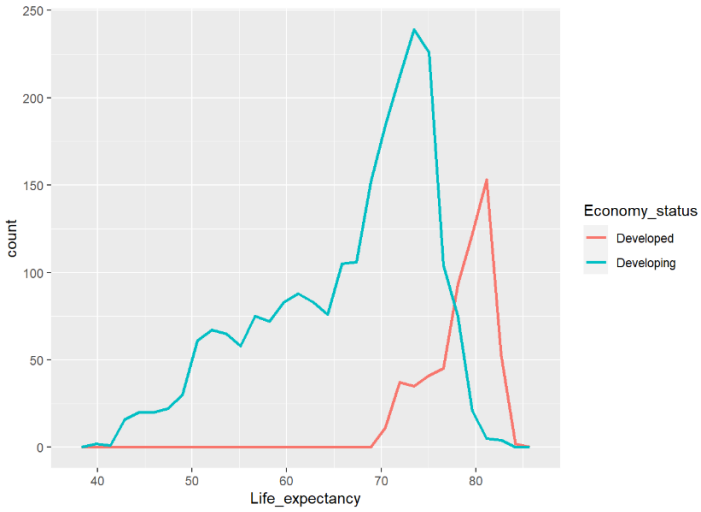


Figure 6: Life expectancy - Economy status

From the graphs above 5 and 6 we can see that the people from Developed countries seem to have higher Life expectancy compare to people who come from Developing countries. We can also observe from the graph 6 that we have more people from developing countries than people from developed. One way to see the proportion is the bar plot below (figure 7).

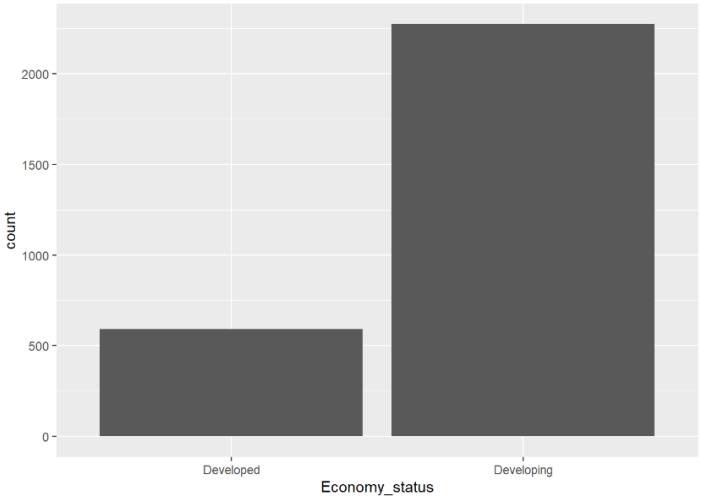


Figure 7: Economy status Bar plot

The people from Developing countries are 4 times the people from Developed countries in our dataset.

We want to check with hypotheses tests that indeed the Life expectancy is higher for people from Developed countries compare to people from Developing countries. To be able to be sure for the results of t-test we should check the normality, as t-test assumes that the variables follow a Normal Distribution. From both shapiro and lillieforse test we reject the Null Hypotheses so the variables does not follow Normal distribution. The t-test suggests that the mean of Life expectancy developed is greater than the mean of Life expectancy developing as we reject the Null Hypothesis for a significance level $\alpha = 0.05$ ($p-value < 0.05$). Despite the result of the t-test we did a permutation test which does not have any assumptions.

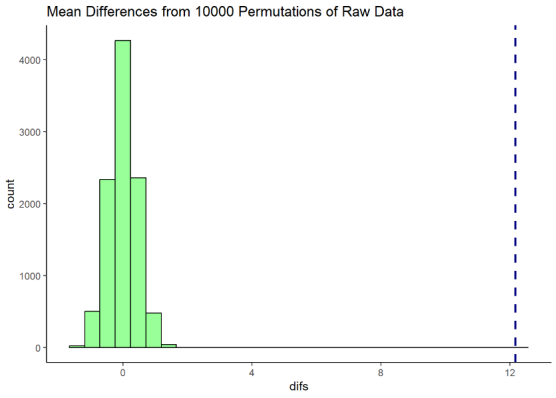


Figure 8: Permutation test graph

The vertical line on the histogram (figure 8) represents the observed mean difference in life expectancy between developed and developing countries as calculated in the original data, against which the distribution of mean differences from the permutation test is compared.

For the permutation test, we reject the Null Hypothesis ($p - value = 0 < 0.05$) so we have that indeed the mean of Life expectancy of people from Developed countries is greater than the Life expectancy of people from Developing countries. We expect `Economy_status` to stay after feature selection.

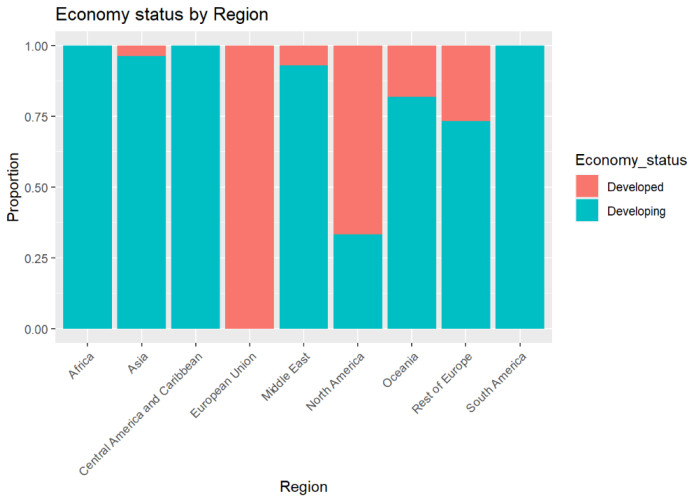


Figure 9: Proportions of Economy status by Region

From the figure 9 we see that all European Union countries are developed and all Africa, Central America and Caribbean and South America countries are developing.

4.3 Region

The dataset categorizes the world into distinct regions: Asia, Middle East, South America, Central America and the Caribbean, European Union, Rest of Europe, Africa, Oceania, and North America, providing a broad geographic framework for the analysis of life expectancy. We find Japan in Asia Region as the country with the highest life expectancy. Conversely, the country with the lowest recorded life expectancy is Sierra Leone, located in Africa.

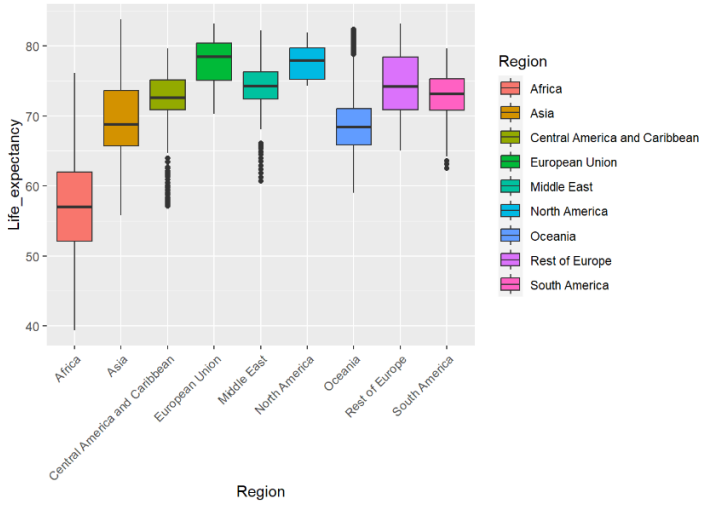


Figure 10: Boxplots of Life Expectancy for each Region

In the graph above (figure 10) we can observe the boxplots of life Expectancy for each Region. Africa shows the lowest median life expectancy with a wide interquartile range, indicating significant variability within the continent. In contrast, the European Union and North America have higher median life expectancies with narrower interquartile ranges, suggesting less variability and generally higher life expectancy. Outliers can be observed in several regions, which could represent countries with life expectancies that deviate significantly from the regional average. According to this graph (figure 10) we expect Region to be among the important features that affect Life Expectancy.

4.4 Alcohol

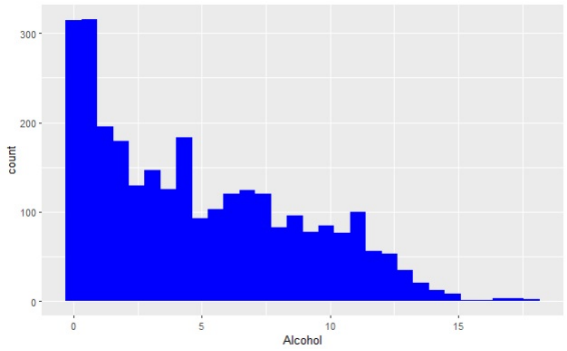


Figure 11: Histogram of Alcohol consumption

From the graph we see that `Alcohol` follows a right-skewed distribution and we would like to check if it follows an exponential distribution with $\lambda = \frac{1}{\mu_{\text{Alcohol}}}$. After investigating further, by performing a Kolmogorov-Smirnov test we conclude that it does not follow the exponential distribution with $\lambda = \frac{1}{\mu_{\text{Alcohol}}}$.

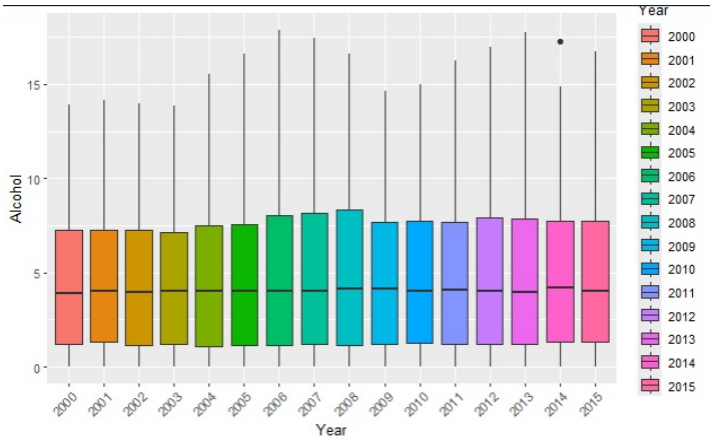


Figure 12: Boxplots for Alcohol consumption by Year

The boxplots in figure 12 reveal that through out the years alcohol consumption has more or less remained the same, with the exception of Estonia 2014 with enough consumption to become an outlier.

Moreover, from our analysis we observed that European Union countries have the highest median for Alcohol consumption.

4.5 Population

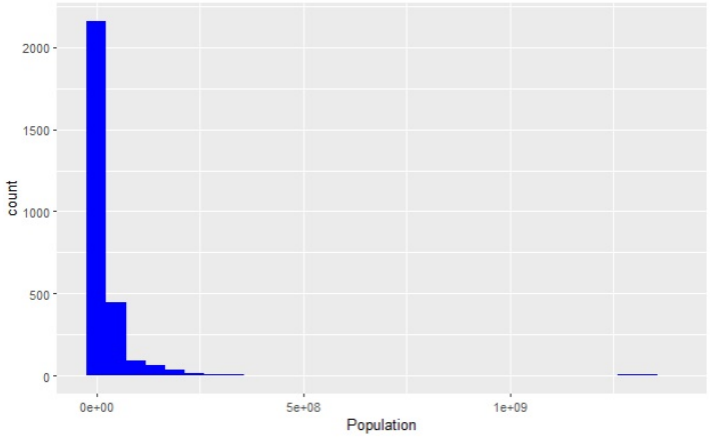


Figure 13: Histogram of Population

In the initial analysis of our dataset, as depicted in figure 13, the histogram of the population variable exhibited a highly right-skewed distribution. This skewness can often lead to challenges in modeling, as many statistical methods assume that the data follows a normal distribution, or at least one that is symmetric and with a constant spread. Using Log-transformation the distribution of `Population` is closer to the Normal, which may be useful later when we fit our model.

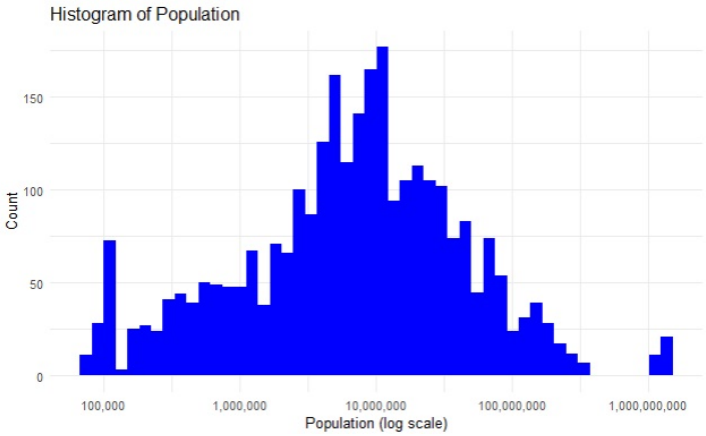


Figure 14: Histogram of log Population

4.6 GDP

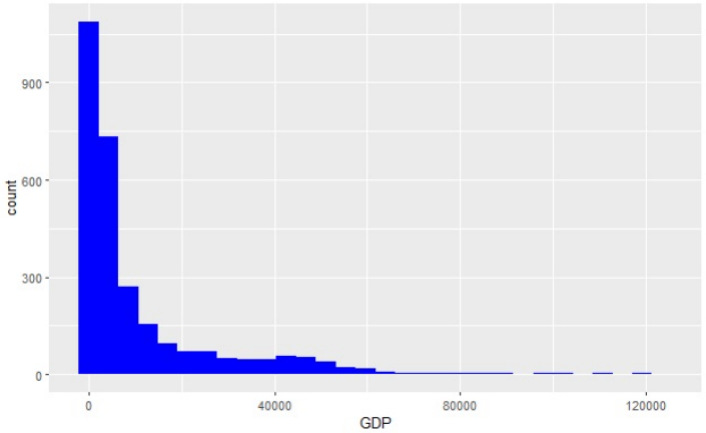


Figure 15: Histogram of GDP

As previous, we use Log-transformation to have a distribution closer to the Normal.

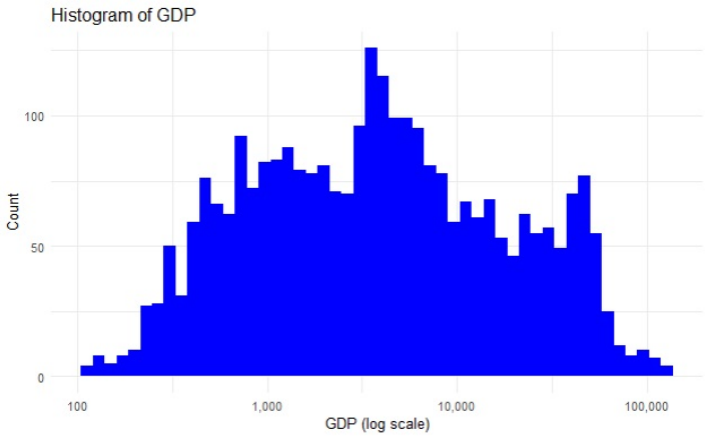


Figure 16: Histogram of log GDP

4.7 Under Five & Infant Deaths

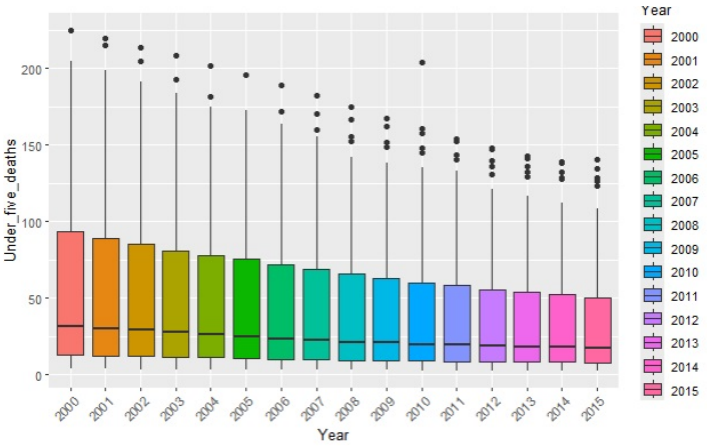


Figure 17: Boxplots of Under Five Deaths each Year

Most countries have a relatively low `Under Five Deaths` index which is only heightened in the African and Asian Regions. That being said we can see the numbers slowly decreasing throughout the years. In figure 17 we can see a specific data point in Year 2010 which is abnormally large even for an outlier, That point is Haiti which in 2010 was struck by an earthquake. Almost the exact trends can be found in the `Infant Deaths` Column.

4.8 Thinness

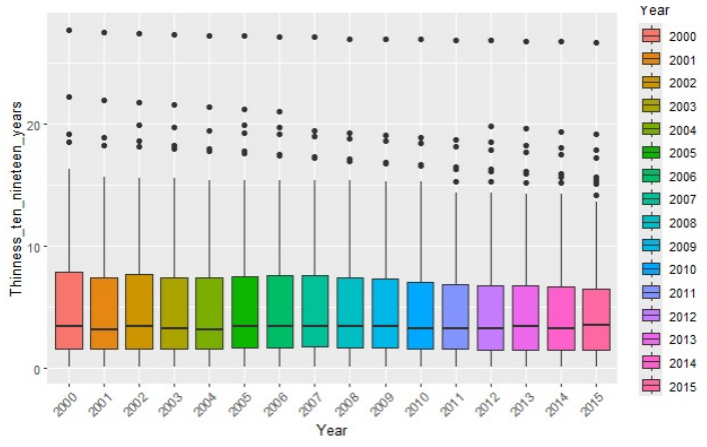


Figure 18: Boxplots of Thinness (10-19 Years) Index in each Year

From our analysis, we found that the maximum outliers which we see in the graph (figure 18) for each year are all from India, a trend which holds true also for `Thinness` 5-9 Years. Moreover, the African and Asian regions have the highest levels of thinness of all other regions.

4.9 Adult Mortality Rates

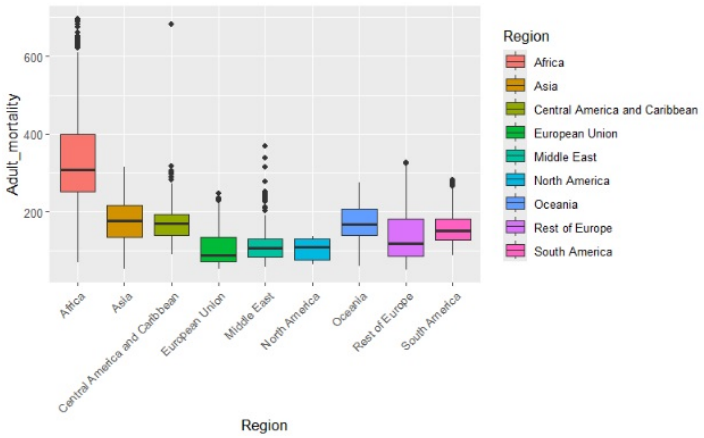


Figure 19: Boxplots of Adult Mortality Rates by Region

Adult mortality indicates the probability of dying between the ages of 15 and 60 per 1000 individuals. Africa region has the highest median again.

In our analysis we observed like `Population` and `GDP` that if we log transform `Adult_Mortality`, then the feature is closer to follow a Normal distribution.

Due to the large number of features, the exploratory analysis in this report has been shortened and we only made references to any important notes.

4.10 Correlations

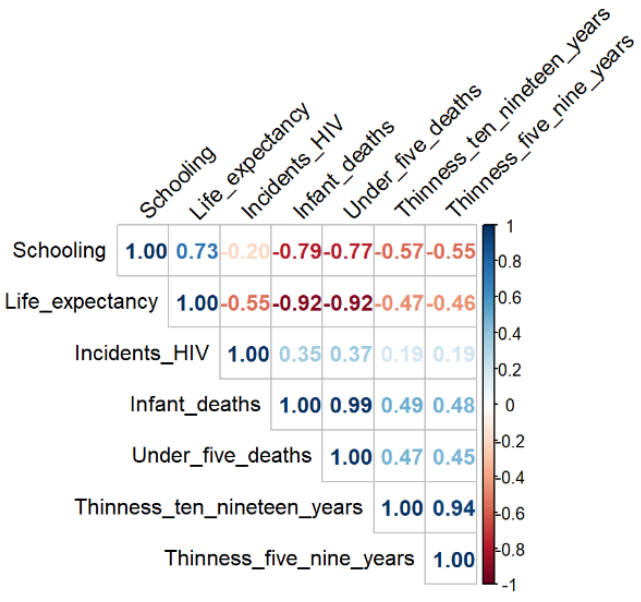


Figure 20: Correlation Plot between some of our Features

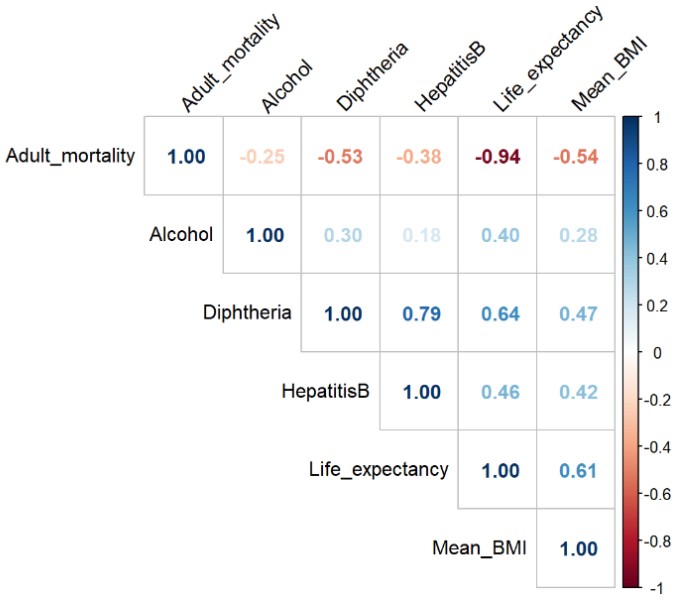


Figure 21: Correlation Plot 2 between some of our Features

From the correlation matrix in figure 20 we noticed that the most significant features were `Infant deaths`, `Under five deaths` and `Adult mortality`. They were highly correlated with Life expectancy. `Infant deaths` are highly correlated with `Under five deaths`. We expected to see such correlation because `Under five deaths` also include `Infant deaths`. Also, from the correlation matrix (figure 21) we have noticed that `Diphtheria` and `HepatitisB` are highly correlated. After feature selection, we expect that the predictors which are highly correlated with each other only one of them will be among the important features.

5 Feature Selection

Firstly, we fit a linear model using all the variables. To be precise we removed the column `Economy_status`, which we created before and the column `Economy_status_Developing` as they provide the exact same information as `Economy_status_Developed`.

```
##
## Residual standard error: 24320 on 2301 degrees of freedom
## Multiple R-squared:  0.962,  Adjusted R-squared:  0.9616
## F-statistic:  2332 on 25 and 2301 DF,  p-value: < 2.2e-16
```

Figure 22: Linear regression model using all the predictors

The models F-statistic is highly significant, as evidenced by a p-value less than 2.2e-16, indicating that the model as a whole has strong predictive power and at least one of the predictors is statistically significant. Individual t-tests (in the r markdown the full summary) for the coefficients reveal that variables such as `RegionAsia`, `RegionCentral America and Caribbean`, `Economy_statusDeveloping`, `Infant_deaths`, and `Incidents_HIV`, among others, are statistically significant, with p-values well below the conventional alpha level of 0.05. This suggests that these factors have a significant impact on life expectancy. Conversely, variables like `Thinness_ten_nineteen_years` and `Alcohol` show a higher p-value, indicating that their relationship with life expectancy is not statistically significant in this model. Because of the large value of predictors we cant be sure for the results of the t-tests, but the important hypothesis test here is the F-test to make sure that we will be able later to predict Life expectancy using these variables.

5.1 Best subset selection

First, we use Best subset selection with complexity penalty criterion. Best subset selection can suffer from overfitting when p is large and is computationally expensive when $p > 30$. We have $p = 25$ and $n = 2327$ so n is much greater than p.

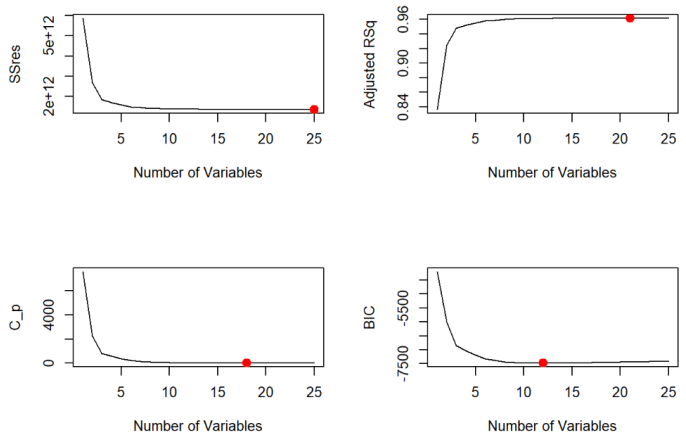


Figure 23: Best subset selection

The suite of graphs 23 illustrates the outcomes of a best subset selection approach to identify the optimal model for predicting life expectancy. The Sum of Squared Errors graph indicates improved fit with more variables, while the Adjusted R-squared graph identifies the model that achieves the best balance of fit and simplicity. The Cp and BIC graphs both penalize model complexity, with their respective red dots highlighting the models that minimize these criteria. In statistical modeling, it is generally preferred to have fewer predictors to avoid overfitting, where the model performs well on the training data but poorly on unseen data. A simpler model with fewer predictors is easier to interpret and usually more generalizable. We can see that approximately at $p = 5$ and after we have the same values. Thus, after checking the most important predictors for $p = 5$ and $p = 8$ we see that, they share key variables such as `Region`, `Infant_deaths`, `Economy_status`, `Adult_Mortality`, and `GDP`. The difference is that for $p = 8$ it adds more dummy variables of `Region`, but for the fitting of the model it does not make a difference since we fit the model using the variable `Region` (meaning that it includes all of its categorical values).

5.2 K-Folds Cross Validation

After, we tried K-Folds Cross Validation (direct way). K-folds cross-validation is a method used to evaluate the predictive performance of a model, where the dataset is divided into K equally sized folds, the model is trained on K-1 folds and validated on the remaining fold, and this process is repeated K times with each fold used once as the validation set. We used $K = 10$.

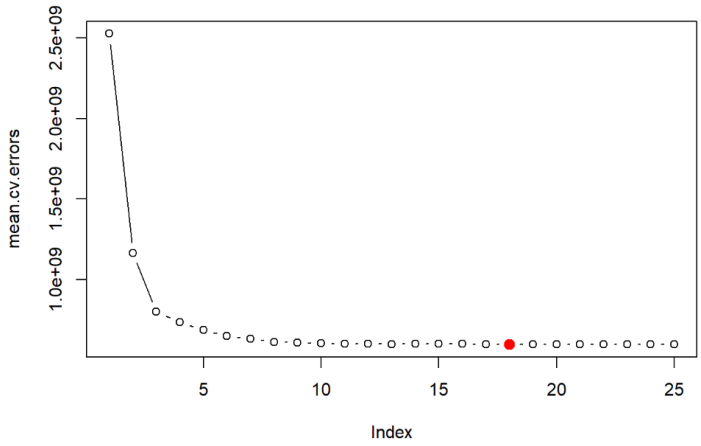


Figure 24: 10-Folds Cross Validation

The graph 24 shows the mean squared errors for each model with different numbers of predictors, with the red dot indicating the model with the lowest cross-validation error, suggesting an optimal balance between complexity and prediction error.

Applying the one standard error rule to the cross-validation results has identified a model with five predictors as optimal, balancing complexity and predictive accuracy. The selected model includes the same 5 predictors like best subset selection. So the 5 predictors are Region, Infant deaths, Economy status, GDP and Adult Mortality. As we have seen before from the correlations, `Adult mortality` and `Infant deaths` were high correlated with our response variable. Furthermore, as we have seen in our EDA, `Region` and `Economy status` were affecting Life expectancy. `GDP` was also positevely correlated with Life expectancy.

6 Models

In our analysis we fitted our model using Linear and Random Forest regression algorithms.

6.1 Random Forest

For the Random Forest regression algorithm we used the important features we found in Feature Selection. The 5 features are `Region`, `Infant_deaths`, `Economy_status`, `GDP`, `Adult_mortality`. Below in table 1 we can see the metrics of the test set (unseen data).

Metric	Value
Root Mean Square Error (RMSE)	28798.74
Mean Absolute Error (MAE)	22766.24
R-squared	0.9583656

Table 1: Random Forest Regression Metrics

6.2 Linear Regression

For Linear Regression with the same 5 features we get the results in table 2. The equation of the specific model is the Equation 1.

Metric	Value
Root Mean Square Error (RMSE)	27962.78
Mean Absolute Error (MAE)	22408.43
R-squared	0.9514785

Table 2: Linear Model Regression Metrics

In the subsequent steps, we aim to enhance the linear regression model. One of the possible problems that we have already dealt with via feature selection is the multicollinearity. Some assumptions of linear regression are that the error terms should be independent (therefore, uncorrelated) and the variance of error terms should not be constant.

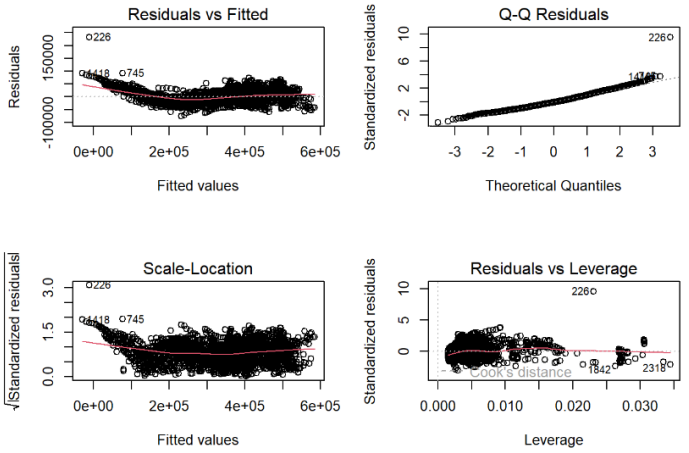


Figure 25: Regression Analysis plots

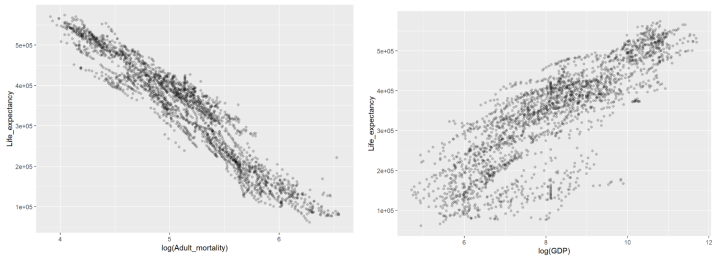
The diagnostic plots from the regression analysis indicate potential issues in the model. The Residuals vs Fitted and Scale-Location plots suggest non-linearity and heteroscedasticity, implying that the assumptions of linearity and equal variance are not fully met. The Q-Q plot shows some deviations from normality in the tails. Additionally, the Residuals vs Leverage plot identifies several influential points that might be overly impacting the models predictions, as evidenced by outliers lying beyond Cooks distance lines (The dashed Cooks distance lines suggest cutoffs for what might be considered an influential point).

In our data analysis, we took significant steps to mitigate the influence of outliers and leverage points that could potentially skew our linear regression model. Outliers were identified as observations with studentized residuals exceeding a thresholdspecifically, those greater than 2 and 3 in absolute value. Similarly, leverage points were determined based on their leverage values, with particular attention to those exceeding 0.2, indicating a substantial influence on the model estimation process.

We created 2 new training datasets where these observations were removed. We proceeded with the updated 2 training sets, which excluded the identified outliers and high leverage points, to develop a two refined regression models. However, the refined models did not surpass the test performance of our best model.

Another important assumption of linear regression is the linearity of the response-predictors relationship. As we have seen previously on our EDA we may improve our model if we log transform the predictors of Adult mortality and GDP.

Below in figure 26 we can see that when we log transform the predictors, they have a linear relationship with our response.



(a) $log(Adult\_Mortality)$ (b) $log(GDP)$

Figure 26: Scatter Plots of Life Expectancy

Indeed, we have an improvement on the metrics calculated on unseen data (table 3) using the equation 2.

Metric	Value
Root Mean Square Error (RMSE)	21600.57
Mean Absolute Error (MAE)	16161.29
R-squared	0.9668497

Table 3: Linear Model Performance Metrics

Next, we checked the optimal complexity of Infant deaths.

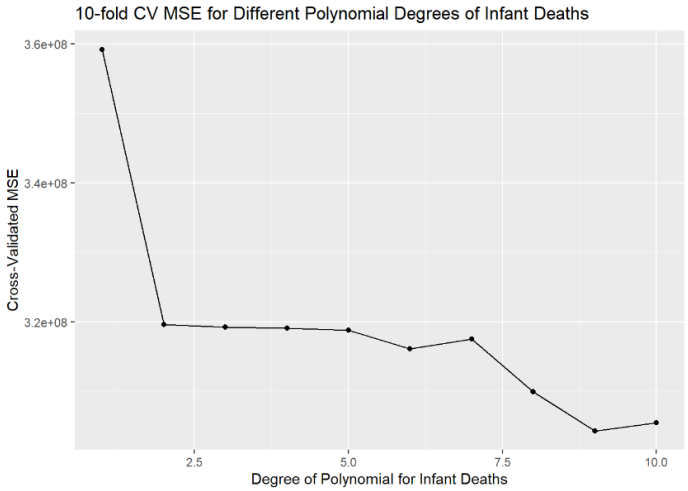


Figure 27: Complexity of Infant Deaths

From the figure 27 and after doing some anova hypotheses tests we conclude that the best equation of our model is with the equation 3 (Infant deaths polynomial degree 2).

Metric	Value
Root Mean Square Error (RMSE)	21195.52
Mean Absolute Error (MAE)	15885.05
R-squared	0.9683997

Table 4: Performance Metrics of the Improved Linear Model

Using the best model we plotted again the plots below.

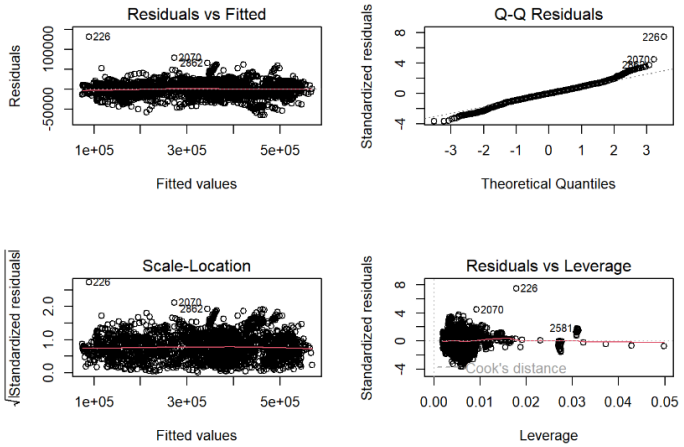


Figure 28: Regression Analysis plots

The diagnostic plots in figure 28 indicate moderate improvements in the regression models assumptions. The Residuals vs Fitted and Scale-Location plots suggest a slight reduction in non-linearity and heteroscedasticity, indicating better compliance with the assumptions of equal variance and linearity. The Q-Q plot shows a better alignment of residuals with the normal distribution, although deviations at the tails persist, suggesting the presence of outliers.

Then, we used the bootstrap approach to assess the variability of the coefficient estimates and the predictions.

Coefficient	Original	Bootstrap
(Intercept)	9220.9801	10209.4647
RegionAsia	1399.3967	1458.1152
RegionCen. Amer. and Carib.	1599.3726	2157.7485
RegionE. U.	2442.0790	2160.6409
RegionM.E.	1851.6111	2489.5517
RegionN.A.	3301.7948	3806.1180
RegionOceania	1857.2134	2203.5634
RegionRest of Eur.	1943.0865	1919.7969
RegionS.A.	1833.8277	1700.5248
Infant_deaths	45508.8268	50010.2154
$Infant\_deaths^2$	24393.0259	23290.4293
E.s. Developing	2159.5661	1913.2789
log(GDP)	517.9888	555.8683
log(Adult_mortality)	1356.8241	1491.6040

Table 5: Comparison of Original and Bootstrap Standard Errors

Comparing the bootstrap results with the original regression output in table 5, we see that for most coefficients, the bootstrap standard errors are reasonably close to the original standard errors, which suggests that the model estimates are relatively stable. However, the coefficients have some bias according to the bootstrap results, which could warrant further investigation.

Concurrently, considering the potential non-linearity in the relationship between a country's GDP and its adult mortality rates, an interaction term was incorporated to explore whether it could enhance the model's predictive power, as detailed in the equation 4.

Metric	Value
Root Mean Square Error (RMSE)	21245.88
Mean Absolute Error (MAE)	15969.35
R-squared	0.9682709

Table 6: Regression Model Metrics with Interaction Term

The metrics on unseen data were slightly worse. Thus, our best model is without an interaction term, equation 3.

7 Conclusion

In this comprehensive study on the determinants of life expectancy, we embarked on a thorough exploration of various factors that could potentially influence the average number of years a newborn is expected to live.

Our analysis began with data pre-processing to address issues such as missing values and data type conversions, ensuring a solid foundation for our investigation. The EDA phase highlighted the influence of factors such as GDP, region and adult mortality rates on life expectancy, amongst others. Through feature selection techniques like best subset selection and k-folds cross-validation, we identified key predictors: region, infant deaths, economy status, GDP, and adult mortality.

We then applied both linear regression and Random Forest models to predict life expectancy. The linear regression model, after several refinements including logarithmic transformations, provided an improved understanding of the relationship between the predictors and life expectancy.

Our findings underscore the multifaceted nature of life expectancy, which is influenced by a complex interplay of economic, regional, and health-related factors. For instance, the distinction between developed and developing countries in terms of life expectancy highlights the critical role of socio-economic status. Moreover, the significance of health indicators such as adult mortality rates and infant deaths points to the importance of healthcare access and quality in determining life expectancy.

In conclusion, this project not only provides valuable insights into the factors affecting life expectancy across different countries but also offers a roadmap for policymakers and healthcare providers to target interventions effectively. By focusing on the identified key predictors, countries can strive to enhance healthcare delivery, address socio-economic disparities and ultimately improve the health and well-being of their citizens. The predictive models developed herein serve as powerful tools for forecasting life expectancy, enabling stakeholders to make informed decisions in their efforts to enhance public health and extend human longevity.

8 Equations

$$\begin{aligned}\hat{Life\_expectancy}^3 =& \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Asia} + \hat{\beta}_2 \cdot \text{Central America\&Caribbean} + \hat{\beta}_3 \cdot \text{European Union} \\ &+ \hat{\beta}_4 \cdot \text{Middle East} + \hat{\beta}_5 \cdot \text{North America} + \hat{\beta}_6 \cdot \text{Oceania} \\ &+ \hat{\beta}_7 \cdot \text{Rest of Europe} + \hat{\beta}_8 \cdot \text{South America} + \hat{\beta}_9 \cdot \text{Infant\_deaths} \\ &+ \hat{\beta}_{10} \cdot \text{Economy\_status Developing} + \hat{\beta}_{11} \cdot \text{GDP} + \hat{\beta}_{12} \cdot \text{Adult\_mortality}\end{aligned} \tag{1}$$

$$\begin{aligned}\hat{Life\_expectancy}^3 =& \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Asia} + \hat{\beta}_2 \cdot \text{Central America\&Caribbean} + \hat{\beta}_3 \cdot \text{European Union} \\ &+ \hat{\beta}_4 \cdot \text{Middle East} + \hat{\beta}_5 \cdot \text{North America} + \hat{\beta}_6 \cdot \text{Oceania} \\ &+ \hat{\beta}_7 \cdot \text{Rest of Europe} + \hat{\beta}_8 \cdot \text{South America} + \hat{\beta}_9 \cdot \text{Infant\_deaths} \\ &+ \hat{\beta}_{10} \cdot \text{Economy\_status Developing} + \hat{\beta}_{11} \cdot log(\text{GDP}) + \hat{\beta}_{12} \cdot log(\text{Adult\_mortality})\end{aligned} \tag{2}$$

$$\begin{aligned}\hat{Life\_expectancy}^3 =& \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Asia} + \hat{\beta}_2 \cdot \text{Central America\&Caribbean} + \hat{\beta}_3 \cdot \text{European Union} \\ &+ \hat{\beta}_4 \cdot \text{Middle East} + \hat{\beta}_5 \cdot \text{North America} + \hat{\beta}_6 \cdot \text{Oceania} \\ &+ \hat{\beta}_7 \cdot \text{Rest of Europe} + \hat{\beta}_8 \cdot \text{South America} + \hat{\beta}_9 \cdot \text{Infant\_deaths} + \hat{\beta}_{10} \cdot \text{Infant\_deaths}^2 \\ &+ \hat{\beta}_{11} \cdot \text{Economy\_status Developing} + \hat{\beta}_{12} \cdot log(\text{GDP}) + \hat{\beta}_{13} \cdot log(\text{Adult\_mortality})\end{aligned} \tag{3}$$

$$\begin{aligned}\hat{Life\_expectancy}^3 =& \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Asia} + \hat{\beta}_2 \cdot \text{Central America\&Caribbean} + \hat{\beta}_3 \cdot \text{European Union} \\ &+ \hat{\beta}_4 \cdot \text{Middle East} + \hat{\beta}_5 \cdot \text{North America} + \hat{\beta}_6 \cdot \text{Oceania} \\ &+ \hat{\beta}_7 \cdot \text{Rest of Europe} + \hat{\beta}_8 \cdot \text{South America} + \hat{\beta}_9 \cdot \text{Infant\_deaths} + \hat{\beta}_{10} \cdot \text{Infant\_deaths}^2 \\ &+ \hat{\beta}_{11} \cdot \text{Economy\_status Developing} + \hat{\beta}_{12} \cdot log(\text{GDP}) + \hat{\beta}_{13} \cdot log(\text{Adult\_mortality}) \\ &+ \hat{\beta}_{14} \cdot log(\text{GDP}) \cdot log(\text{Adult\_mortality})\end{aligned} \tag{4}$$

Please note that in the regression model for forecasting life expectancy within the Africa Region, all regional predictors are set to 0. Same for predicting Life Excpectancy for Developed countries the `Economy_statusDeveloping` is set to 0. Furthermore, the coefficients of each equation are not equal.