Exploring Determinants and Constructing a Predictive Model for Global Life Expectancy: A Comprehensive Analysis of Health Factors

Helena Blumenau

Department of Biostatistics
University of Kansas, USA

December 1, 2023

Contents

List of Tables

List of Figures

Appendix Figures

Title

Exploring Determinants and Constructing a Predictive Model for Global Life Expectancy: A Comprehensive Analysis of Health Factors

Abstract

This study examines the web of factors correlated with global life expectancy through the construction of a comprehensive predictor model. Utilizing multiple linear regression, an array of predictor variables were explored, resulting in a final model that reflects crucial aspects of health and socio-economic status. Adult mortality, alcohol consumption, measles cases, body mass index, polio vaccination rates, diphtheria vaccination rates, the country's developmental status, and gross domestic product were included in the final model. .The selected predictors underscore the complex nature of factors that ultimately are related to life expectancy, capturing both health related, lifestyle and socio-economic indicators. The study acknowledges limitations, including anomalies in the data and challenges with the predictive model, but despite these constraints, the study contributes valuable insights into global life expectancy determinants, ultimately offering actionable information for public health policies. This study sets the stage for future research.

Introduction

Life expectancy is a critical indicator of the overall health and well-being of populations, ultimately reflecting the effectiveness of healthcare systems, social policies, and individual behavioral patterns. It stands as a comprehensive reflection of the interconnectedness between various socio-economic, environmental and health care factors. There have been numerous studies that explored factors influencing overall life expectancy. Previous research has identified GDP, adult mortality rate and percentage expenditure as having the greatest impact (Wang, 2021). Additionally, the impact of vaccinations on public health and life expectancy is invaluable, despite them being under-utilized globally (Ehreth, 2003; Rughiniș et al., 2022). For instance, the implementation of measles vaccination in countries like the United States has led to a substantial decline in measles cases, while countries without such measures reported an increase (Ehreth, 2003). Furthermore, investigations into lifestyle factors, such as alcohol consumption, have demonstrated that alcohol consumption contributes to premature death and a reduction in life expectancy (Ranabhat et al., 2019).

This study aims to delve deeper into the determinants associated with life expectancy, examining a diverse set of predictors across different countries. The dataset, sourced from the World Health Organization's Global Health Observatory (GHO) data repository, spans 193 countries and encompasses life expectancy and various health-related factors. Given the exploratory nature of this project, a comprehensive set of independent variables will be considered to evaluate their associations with life expectancy.

These variables include:

1. **Adult mortality rate:** Adult mortality rates of both sexes per 1,000 population
2. **Infant deaths:** Number of infant deaths per 1,000 population
3. **Alcohol consumption:** Alcohol, recorded per capita consumption in liters of pure alcohol
4. **Hepatitis B, Polio, and Diphtheria vaccination rate:** Immunization coverage among 1-year-olds (percentage)
5. **Measles:** Number of reported cases of measles per 1,000 people
6. **Body mass index:** Average body mass index of entire population
7. **Each country's developmental status:** Developed or developing status
8. **Gross domestic product:** Gross domestic product per capita in USD

The preliminary goal is to discern which variables are significantly associated with life expectancy, exploring potential multicollinearity among different vaccines and investigating the impact of lifestyle factors such as alcohol consumption and BMI. Ultimately, the study aims to construct a predictive model for life expectancy based on the variables in the dataset.

By pursuing this research, we seek not only to enhance our understanding of factors linked to life expectancy but also to provide a tangible forecasting tool. The final predictive model, grounded in empirical data, holds the potential to guide targeted policies for improving global population health outcomes.

Primary Analysis Objectives

To investigate the linear association between life expectancy and key predictor variables, including adult mortality rate, infant deaths, alcohol consumption, various vaccination rates, body mass index, developmental status and gross domestic product. Among these predictors, the objective of this study is to identify which variables are related to life expectancy.

Secondary Analysis Objectives

Alongside creating the predictive model, this study will investigate whether lifestyle factors, such as alcohol consumption and BMI, are correlated with life expectancy. Additionally, this study will examine the presence of multicollinearity among different predictor variables, including the several vaccination rates, anticipating an overall negative association with life expectancy. I also expect to find similar patterns for GDP, adult mortality rate, and percentage expenditure. Similarly, I anticipate that increased alcohol consumption and BMI will correlate with a decrease in average life expectancy.

Materials and Methods

Data Collection

This data set originates from the World Health Organization's Global Health Observatory (GHO) data repository. It examines life expectancy and various health-related factors across 193 countries. Data collection spans the years 2000 to 2015, and a unified dataset was created by merging the public data files.

Data Preprocessing

An examination of the dataset was undertaken to identify essential preprocessing steps. This process involved a thorough assessment of data quality, including the identification of missing or duplicate values, as well as the verification of data types to ensure R loaded the data types correctly. The data appeared to be well-structured, but missing values were present. To address missing values in the dataset, the mean value for each specific column was used.. The resulting dataset comprises 22 columns and 2938 observations, each representing health-related data for a particular country each year. Additionally, there were several outliers identified in exploratory analysis, which after using Cook's distance were deemed highly influential, and were removed from the final dataset. This can be seen in Appendix C.

Final Dataset

The names, type of data, data format, as well as a description and example for each of the variables included in the dataset is included in the table below.

Table 1: Final Dataset

Variable Name	Data Type	Data Format	Description	Example
Country	Character	tttt	Country from which the data was taken from	Afghanistan
Year	Number	123	Year in which the data was collected	2015
Status	Factor	{Developing, Developed}	Factor whether the country has a status of developed or developing	Developing
Life Expectancy	Number	123	Average life expectancy in age for each country in a given year	59.9

Adult Mortality	Number	123	Adult mortality rates of both sexes, measured in terms of the probability of dying between 16 and 60 years per 1,000 people	263
Infant Deaths	Number	123	Number of infant deaths per 1,000 people	62
Alcohol	Number	123	Alcohol consumption per capita in liters of pure alcohol, taken from those age 15+	0.01
Hepatitis B	Number	123	Percentage of Hepatitis B immunization coverage among 1 year olds	65
Polio	Number	123	Percentage of Polio immunization coverage among 1 year olds	58
Diphtheria	Number	123	Percentage of Diphtheria Tetanus Toxoid and Pertussis immunization coverage among 1 year olds	65
Measles	Number	123	Number of reported cases of measles per 1,000 people	492
BMI	Number	123	Average body mass index for the entire population	19.1
GDP	Number	123	Gross domestic product per capita, reported in USD	584.25921

Statistical Analysis

The data is available in .xlsx (excel) format. The data analysis is done using the statistical software R, and the project focuses on using multiple linear regression. Each of the predictor variables was explored individually. The missing values in the dataset were addressed as specified previously, and there is a relatively large sample size (2938 total observations), which ensures better predictability in the sample. Automatic model selection methods were used to arrive at the final model. The model assumptions were assessed and confirmed, ending with a suggestion on the final predictive model.

Primary Objective Analysis

It is crucial to explore individual predictors and the response variable to ensure that model assumptions are met later on. This exploration helps identify any potential skewness in the data, or outliers. If, later, it is observed that the model does not fit the data well, this initial exploration serves as a reference point for identifying where transformations may be necessary to improve the overall model fit. After exploring the dataset variables, the next step involves checking the linearity between the predictor variables and the response variable. Additionally, predictor variables were also examined for potential multicollinearity, as this can impact the overall accuracy of predictions made from the model. Addressing these issues will contribute to fitting a more reliable model. The methodology for this analysis includes selecting a subset of predictor variables that best predict life expectancy. Statistical tests will be conducted with a significance level of 0.05, considering p-values less than 0.05 as statistically significant.

Secondary Objective Analysis

The assessment of correlations among predictor variables with the response variable, as well as with one another, includes the use of statistical measures such as R-squared values, Pearson correlation coefficient, scatterplot matrices, and the variance inflation factor (VIF). These metrics provide insights into the relationships occurring within the dataset.

Results

Summary Statistics

To start the analysis, a preliminary exploration of the dataset was conducted using both graphical and descriptive methods. The mean, median, and standard deviation were examined for each variable to gain insights into central tendencies and variation in the data. The average life expectancy across the dataset is 69.225 years, ranging from 36.3 to 89 years.The average adult mortality rate was 164.796 deaths per 1,000 people, ranging from 1.00 to 723.00 deaths. The average infant deaths in the dataset was 30.303, ranging from 0.00 to 1800.00. The average alcohol consumption per capita in liters of pure alcohol is 4.603, ranging from 0.010 to 17.87 liters. The average Hepatitis B immunization coverage among 1 year olds is 80.940%, and ranges from 1.00 % to 99.00%. Polio immunization coverage in the dataset averages at 82.55%, but ranges from 3.00% to 97.00%. The average Diphtheria immunization coverage among 1 year olds is 82.324%, ranging from 2.00% to 97.00%. The average number of reported cases of measles per 1,000 people is 2419.592, ranging from 0.00 to 212183.000. The average body mass index across the dataset is 38.32, ranging from 1.00 to 87.30. The average gross domestic product in USD is $7483.16, but ranges from $1.681 to $119,172.74 (Table 2).

Table 2: Summary Statistics for Dataset Variables

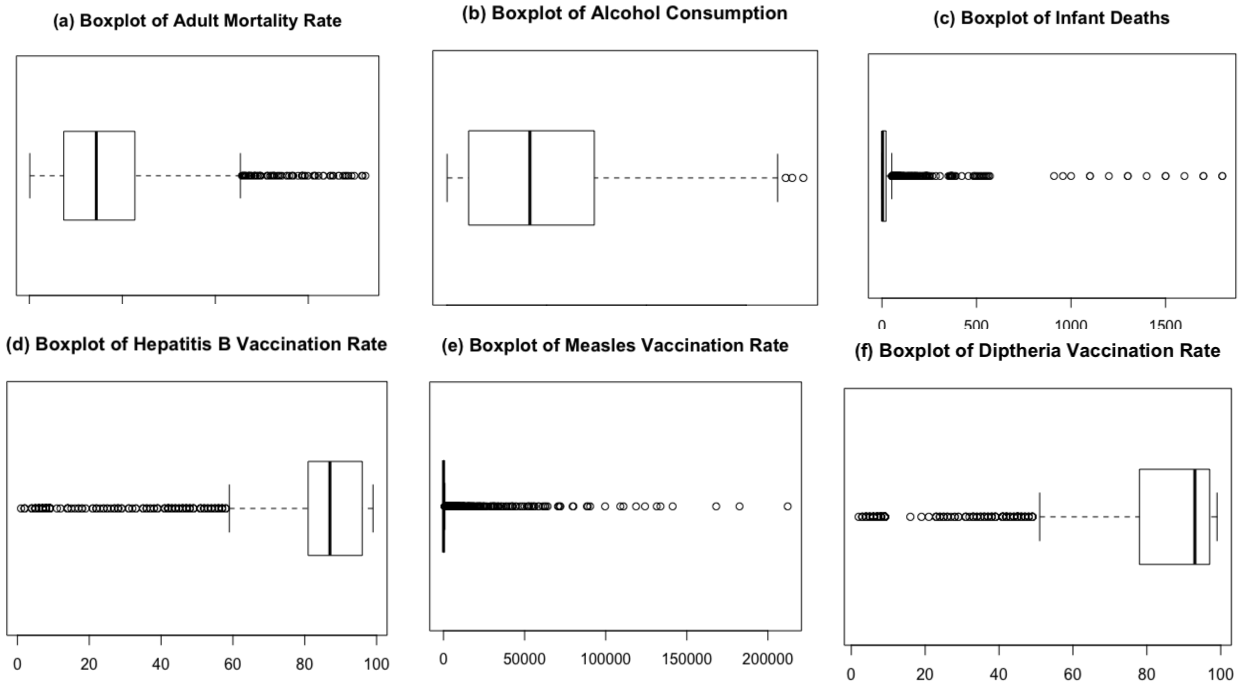
Variable	Min	Q1	Mean	Q3	Max	SD
Life Expectancy	36.300	63.200	69.225	75.600	89.000	9.508
Adult Mortality	1.00	74.00	164.796	227.00	723.00	124.08
Infant Deaths	0.00	0.00	30.303	22.00	1800.00	117.93
Alcohol	0.010	1.092	4.603	7.390	17.87	3.916
Hepatitis B	1.00	80.940	80.940	96.00	99.00	22.586
Polio	3.00	78.00	82.55	97.00	99.00	23.352
Diphtheria	2.00	78.00	82.324	97.00	99.00	23.641

Measles	0.00	0.00	2419.592	360.250	212183.0	11467.27
BMI	1.00	19.40	38.32	56.1	87.30	19.927
GDP	1.681	580.487	7483.16	7483.158	119172.74	13136.80

Data Visualization

Graphical methods, including boxplots, histograms and scatter plots were utilized in visualizing the distributions of each variable included in the final dataset. These were used to identify any potential patterns or trends in the data that may need to be addressed. Figure 1 and Table 2 show the distributions of each potential predictor variable. Figure 1(i) appears to be symmetrical, while figure 1(a), (b) and (g) show right, or positive, skewness. Figure 1(c), (d), (e), (f), and (h) show left, or negative skewness. Figure 1(a-h) also appears to have outlying or extreme data points. More on the distribution of the predictor variables can be found in Appendix A. Preliminary data analysis on life expectancy shows that the response variable is negatively skewed as the whiskers to the left are longer than the ones on the right. This is shown in figure 2(b), as well as in the histogram (Figure 2a). Presence of outlying values are observed in 2(a).

Figure 1: Analysis of Predictor Variables



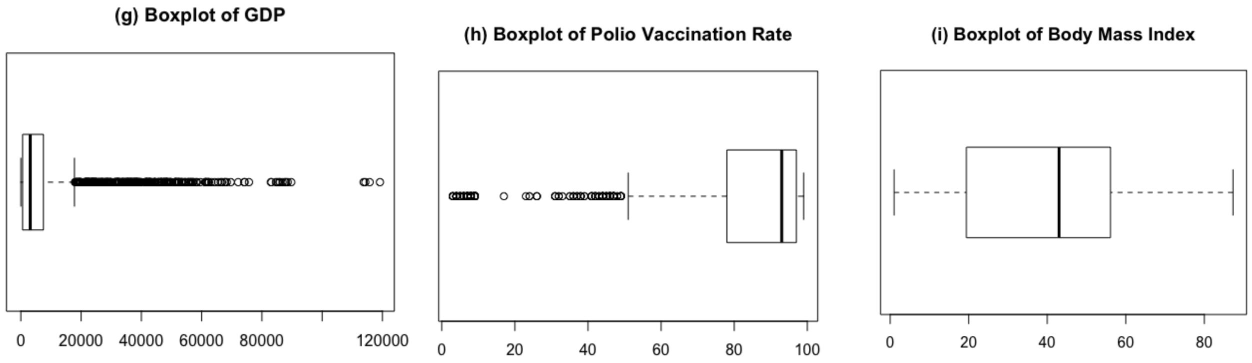
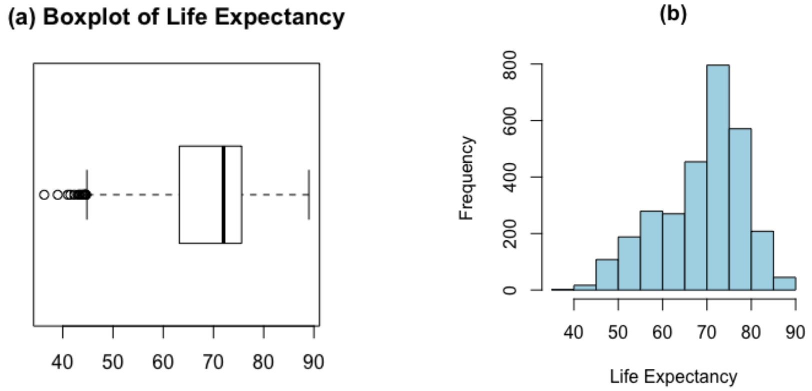


Figure 2: Distribution of Life Expectancy



Correlation Analysis

The bivariate relationships are also examined visually with a scatterplot matrix. Seen below are the scatter plots with linear relationships. The full scatterplot matrix can be found in Appendix A. Figure 3a shows the relationship between Hepatitis B vaccination rate and Diphtheria vaccination rates, where a positive linear relationship can be seen visually. Figure 3b shows the relationship between Hepatitis B vaccination rate and Polio vaccination rates, where a positive linear relationship can also be seen visually. The other scatterplots in Appendix A do not show as strong of a linear association between variables.

Correlation coefficients were calculated to statistically assess the relationships between the previously identified predictor variables, as well as with the response variable. In this case, to try and identify any linear associations between these variables. This served as the foundation for identifying potential influential factors to form a predictive model. The analysis of correlations between various predictor variables is shown below in Figure 3. There is only one relationship with a correlation coefficient greater than 0.5. The strongest relationship is seen between developmental status and alcohol, with a correlation coefficient of -0.579. The next strongest relationship is seen between Diphtheria and Hepatitis B vaccination rates, with a correlation coefficient of 0.4999, indicating some multicollinearity in these relationships. Other than these two identified, the correlation coefficients are all relatively small, indicating no multicollinearity present. To examine the influence of these relationships, the variance inflation factor was utilized in the next section.

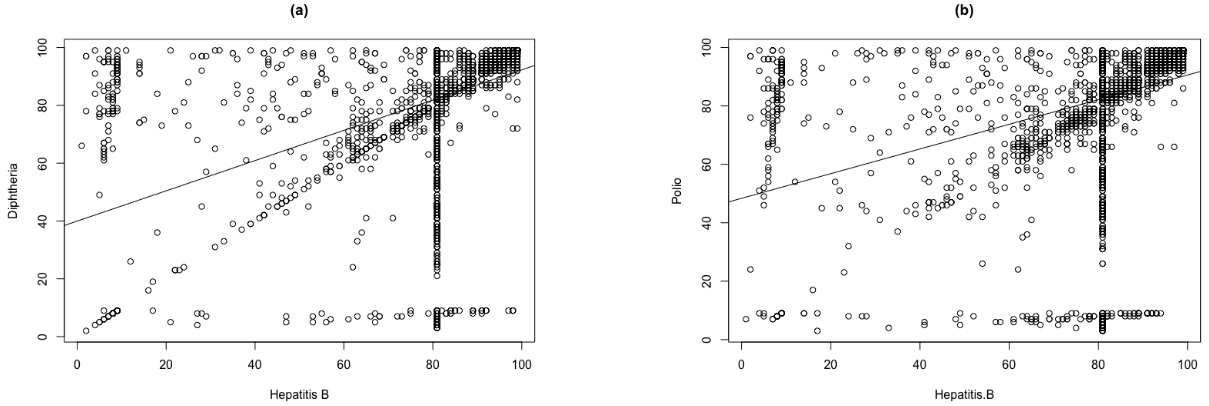
Table 3A: Correlation Matrix of Predictor Variables

	Adult Mortality	Infant Deaths	Alcohol	Hepatitis B	Measles	BMI	Polio	Diphtheria	Status	GDP
Adult Mortality	1.000	0.079	-0.190	-0.139	0.031	-0.381	-0.273	-0.273	0.315	-0.278
Infant Deaths	–	1.000	-0.114	-0.179	0.501	-0.227	-0.171	-0.175	0.112	-0.107
Alcohol	–	–	1.000	0.0754	-0.051	0.318	0.214	0.215	-0.579	0.319
Hepatitis B	–	–	–	1.000	-0.090	0.135	0.408	0.499	-0.09	0.062
Measles	–	–	–	–	1.000	-0.176	-0.136	-0.142	0.077	-0.068
BMI	–	–	–	–	–	1.000	0.282	0.281	-0.311	0.276
Polio	–	–	–	–	–	–	1.000	0.673	-0.220	0.194
Diphtheria	–	–	–	–	–	–	–	1.000	-0.217	0.183
Status	–	–	–	–	–	–	–	–	1.000	-0.446
GDP	–	–	–	–	–	–	–	–	–	1.000

Table 3B: Correlation Matrix of Response Variable

Life Expectancy	Adult Mortality	Infant Deaths	Alcohol	Hepatitis B	Measles	BMI	Polio	Diphtheria	Status	GDP
Correlation Coefficient	-0.696	-0.196	0.391	0.204	-0.158	0.559	0.461	0.475	0.482	0.430

Figure 3: Scatterplot Matrices



Variable selection

Our model incorporates multiple predictor variables, creating the need to assess potential multicollinearity between variables. Figure 3 reveals the correlational relationships among these predictor variables. Including variables that have significant correlations with one another can increase the variance of the model coefficients, which would limit their ability to accurately represent the underlying patterns in the data. The correlation matrix indicates that there may be potential multicollinearity between alcohol and developmental status, as well as between hepatitis B vaccination rates and Diphtheria vaccination rates, as well as with Polio vaccination rates. To quantify these relationships, the variance inflation factor (VIF) was utilized for each predictor variable within the model. We find that all VIF values fall within the range 1 - 2.2, with the highest, Diphtheria vaccination rates, at 2.126. All other predictors are within the 1-2 range. These values signify the coefficients are slightly inflated due to linear dependence with one another, but not to such an extent that we would consider it problematic and exclude them from the predictive model. Given that the variables are all below a common threshold of 5, the conclusion is drawn that multicollinearity will not have a significant influence on the regression estimates.

Model Selection

Model Choice

Multiple linear regression was selected as the primary method due to its capability to handle several predictor variables simultaneously. The regression analysis was conducted to quantify the impact of each predictor variable on life expectancy, and to provide coefficients to indicate the strength and direction of these relationships.

The initial analysis used the Bayesian Information Criterion, BIC, to establish the optimal linear model. Other methods, such as Mallow's C_p, R^2, and residual sum of squares, were employed to validate the model choice from the BIC, and information on this can be found in Appendix B. As the number of predictor variables increases, the significance of each additional variable decreases. The BIC is calculated using the following formula:

***BIC* = -2 x log-likelihood + *k* x log(*n*)**

Where:

- **Log-likelihood** is the maximum log-likelihood of the model
- ***k*** is the number of parameters in the model
- ***n*** is the sample size

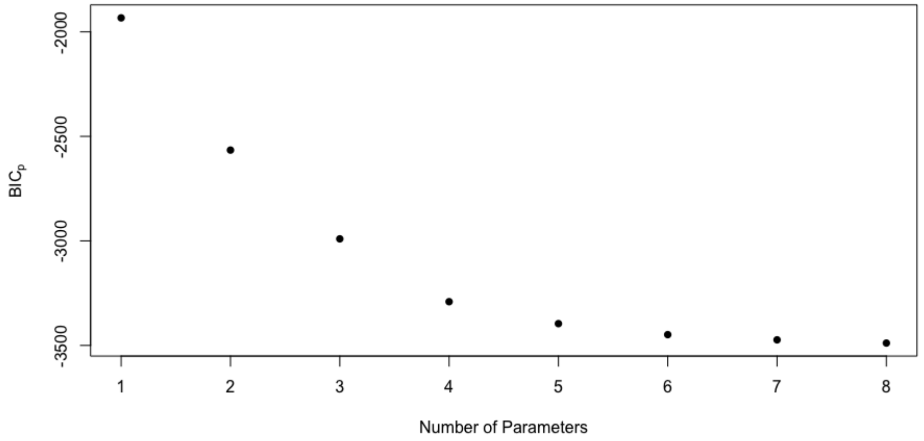
The BIC is a tool used for model selection, and is especially useful when comparing models with many parameters. It's derived from Bayesian principles and balances the model fit with model complexity, penalizing models that have more parameters. The goal is to minimize this value, so the model with the lowest BIC is considered the best-fitting model. Again, the BIC strikes a balance between how well it fits and how complex the model is, and this penalty is proportional to the number of parameters. This criterion more heavily penalizes for the number of parameters, as well as for sample size, than the Akaike's Information Criterion (AIC), which is why BIC was opted for in this model.

Table 4: Model Selection using BIC

Model	Included Variables								Size	BIC
1	Adult Mortality								1	-1933.2
2	Adult Mortality	BMI							2	-2565.7

3	Adult Mortality	BMI	Diphtheria						3	-2990.3
4	Adult Mortality	BMI	Diphtheria	Status					4	-3290.9
5	Adult Mortality	BMI	Diphtheria	Status	GDP				5	-3995.9
6	Adult Mortality	BMI	Diphtheria	Status	GDP	Polio			6	-3448.6
7	Adult Mortality	BMI	Diphtheria	Status	GDP	Polio	Alcohol		7	-3473.4
8	Adult Mortality	BMI	Diphtheria	Status	GDP	Polio	Alcohol	Measles	8	-3488.9

Figure 4: Model Selection using BIC



Based on the BIC criteria, the model with the lowest BIC value is one that includes adult mortality, BMI, Diphtheria vaccination rate, country status, GDP, Polio vaccination rate, Alcohol consumption and Measles cases (Table 4, Figure 4). The other model selection techniques confirmed this choice (Appendix B). Despite having a large number of predictor variables, the model selection tools concluded this was the best fit to the data, while penalizing for complexity. Predictor variables infant deaths and Hepatitis B vaccination rates were dropped from the best eight models.

Model Assumptions

Residual Analysis

The evaluation of homoscedasticity and the identification of systematic variance in the residuals can be seen in the residuals vs. fitted value plot (Figure 5). There are no strong visual patterns in the plot, suggesting that the assumption of constant variance of the residuals is met. Although, there does appear to be a subtle clustering of points, which is further discussed in the conclusion section. The linear model relies on key assumptions, including normality, and the appropriateness of the linear model to explain the data. Normality is examined in the Q-Q plot, Figure 6. The majority of residuals closely align with the best fit line, however a distinct departure can be seen in the lower tail of the plot, indicating a heavier distribution. This deviation from normality is further explored in the conclusion section, where its impact on the validity of the linear regression model is addressed.

Figure 5: Residuals vs. Fitted Values Scatterplot

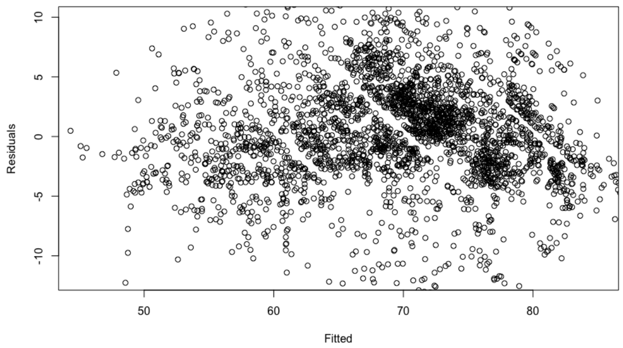
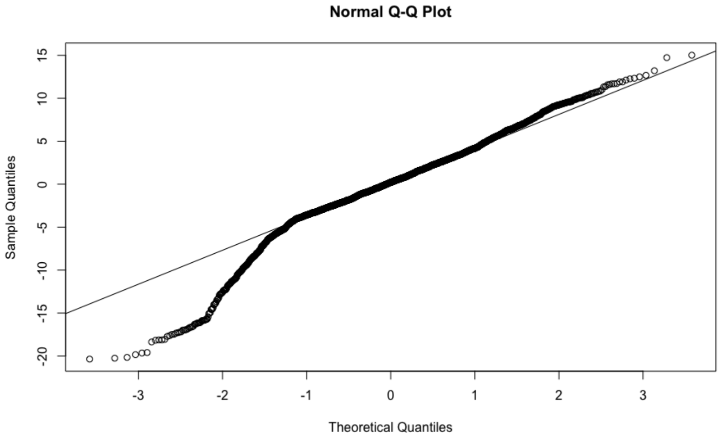


Figure 6: Normality Q-Q Plot



Final Model

The final regression model is summarized in table 4. The coefficients presented in the table quantify the impact of each predictor variable on the response variable, life expectancy, offering insights into the strength and direction of these associations. Notably, the statistical significance of each coefficient is important for determining the reliability of $\hat{Y}_i$. Additionally, measures such as the R-squared value, 0.7328, tell us that 73.28 % of the variance in life expectancy is explained by the model.

The estimated regression function from this data analysis would be:

$$\hat{Y}_i = 59.40 - 0.03772X_1 + 0.2092X_2 - 0.00003X_3 + 0.09086X_4 + 0.04239X_5 + 0.005691X_6 + 2.603X_7 + 0.0008X_8$$

where,

$\hat{Y}_i$ is the predicted life expectancy

X_1 is the adult mortality

X_2 is the alcohol consumption

X_3 is the measles cases

X_4 is the body mass index

X_5 is the polio vaccination

X_6 is the diphtheria vaccination

X_7 is the country's developmental status

X_8 is the gross domestic product

$i = 1,2,3,...,2913$

The model was evaluated using Leave-One-Out- Cross-Validation (LOOCV) to assess its predictive performance. The LOOCV results provide valuable insights into the model's ability to generalize to new data. The Root Mean Squared Error (RMSE), calculated as the square root of the mean squared differences between predicted and observed life expectancy values, provides a measure of the model's precision. The overall RMSE is 4.834755. Additionally, the relative RMSE, normalized by the range of life expectancy values, offers a scale independent assessment of the model's performance. The relative RMSE is 0.09174. The R squared value was reported in this table, estimating that approximately 73.06% of the variability in life expectancy is accounted for by the model. The Mean Absolute Error (MAE) is another measure of average magnitude of errors that is less sensitive to outliers compared to RMSE. The MAE is 3.557, which tells us on average, the absolute difference between predicted and observed values is 3.56 units, or years in this model (Table 7). The provided RMSE and MAE values suggest that the model's predictions deviate by around 4.84 years and 3.56 years, respectively. The R squared value of 0.73057 indicates that the model explains most of the variability in life expectancy on the included predictor variables (73.06%). The model appears to have a reasonably good fit to the data based on these metrics. Predictions were calculated and compared to observed values as well (Table 8).

Table 5: Statistics Table for the Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	59.40	0.5483	108.327	0.0000
Adult Mortality	-0.03772	0.000826	-45.667	0.0000
Alcohol	0.2092	0.02869	7.292	0.0000
Measles	-0.00003	0.00000993	-3.400	0.000683
BMI	0.09086	0.005235	17.355	0.0000
Polio	0.04239	0.00527	8.042	0.0000
Diphtheria	0.05691	0.005211	10.921	0.0000
Status	2.603	0.3133	8.306	0.0000
GDP	0.00008	0.000007767	10.301	0.0000
Observations	2913			
R^2	0.7328			
Adjusted R^2	0.732			
Residual Std. Error	4.82 (df = 2904)			
F Statistic	995.3 (df = 8, 2904)			

Table 6: ANOVA Table for the Regression Model

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Adult Mortality	1	135478	135478	5807.737	0.0000
Alcohol	1	17888	17888	766.813	0.0000
Measles	1	2134	2134	91.467	0.0000
BMI	1	12276	12276	526.268	0.0000
Polio	1	9387	9387	402.410	0.0000
Diphtheria	1	2899	2899	124.270	0.0000
Status	1	3202	3202	137.267	0.0000
GDP	1	2475	2475	106.105	0.0000
Residuals	2904	67742	23		

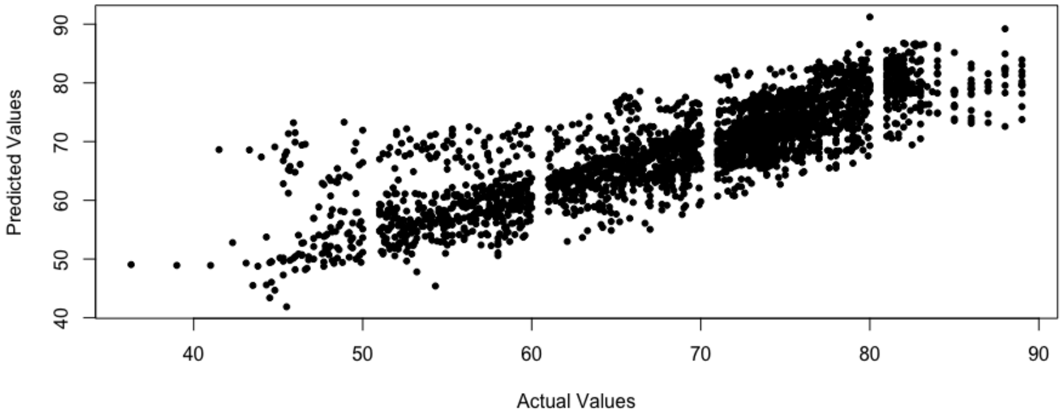
Table 7: Validating Model Prediction (LOOCV)

RMSE	R Squared	MAE
4.841972	0.7305752	3.557004

Table 8: Testing Model Prediction

Observation	Observed Life Expectancy	Predicted Life Expectancy	Predicted Range
800	78	66.067	$55.872 \leq \hat{Y} \leq 76.261$
1500	56	53.305	$43.108 \leq \hat{Y} \leq 63.503$
2005	73.8	76.257	$66.066 \leq \hat{Y} \leq 86.447$

Figure 7: Actual vs. Predicted Values



Discussions and Conclusions

This study aimed to investigate the factors influencing life expectancy through the creation of a predictor model. We explored a wide range of predictor variables, including adult mortality rates, infant deaths, alcohol consumption, Hepatitis B immunization, measles cases, BMI, Polio and DIphtheria immunization rates, developmental status, and gross domestic product. Using multiple linear regression, the model was trained and evaluated, with a focus on predictive accuracy while maintaining the interpretability of the identified relationships. Adult mortality, alcohol consumption, measles cases, body mass index, polio vaccination rates, diphtheria vaccination rates, the country's developmental status, and gross domestic product were included in the final model. The findings revealed several associations between life expectancy and predictor variables, specifically adult mortality, BMI and diphtheria vaccination rates, as they were selected to be incorporated into the predictive models each time when using various model selection criteria. The variables which were included later were alcohol consumption and measles cases. Infant deaths and Hepatitis B vaccination rates were not included in any of the models. For the primary objective analysis, we were able to create a predictive model for life expectancy which included the majority of variables examined going into the study.

In regards to the secondary analysis objectives, this study investigated whether lifestyle factors are correlated with life expectancy. The original hypothesis was that increased alcohol consumption and BMi will correlate with a decrease in average life expectancy. As seen in table 3b, the correlation coefficient between alcohol consumption and life expectancy was 0.391, which indicates the presence of a moderate positive linear association. This means that an increase in alcohol consumption is related to an increase in life expectancy, not very strongly. The correlation coefficient between BMI and life expectancy was much higher, 0.559, indicating a relatively strong association between body mass index and life expectancy. Additionally, this study investigated the multicollinearity between vaccination rates. Table 3a did find correlations between vaccination rates. Hepatitis B and polio have a correlation coefficient 0.408, and between Hepatitis B and Diphtheria have a correlation coefficient 0.499, both indicating strong positive linear associations. While these were not found to be influential in terms of VIF, it is still interesting to note.

In drawing these conclusions from the study on life expectancy, there were several limitations and challenges that should be addressed. There were some peculiarities in the reported data. For example, instances of extreme values, such as the maximum of infant deaths out of 1,000 live births being 1,800 infant deaths. There were similar anomalies in Measles incidence, which raises questions about the accuracy and consistency of data reporting. Future studies may benefit from a deeper exploration of data quality and the impact of these values on model outcomes.

The predictive model, while exhibiting strong performance in terms of fitting the data well, and an R squared value of 0.7306, explaining 73.06% of the variation in life expectancy, also had some challenges. Specifically, this is referring to the achievement of the normality of residuals. While the normality was somewhat okay, it wasn't exactly how we want to see the spread of residuals. Even after many attempts at variable transformations, both of the response and predictor variables, this normality did not improve significantly. It was the ultimate decision to keep the model as is, because while this still did not fix the normality of residuals, it did compromise the interpretability of the model. This challenge is inherent in balancing the statistical assumptions of linear regression, where there is very much a practical need for a model that remains interpretable in real-world scenarios.

In conclusion, the study provides valuable insights into the many factors that relate to life expectancy globally. The identified determinants offer actionable information for public health and policy interventions. By acknowledging the limitations and nuances in this study, it lays the framework for future studies, with an emphasis on robust methodologies that may better suit this complex dataset.

References

Ehreth, J. (2003). The value of vaccination: A global perspective. *Vaccine*, *21*(27–30), 4105–4117. https://doi.org/10.1016/s0264-410x(03)00377-3

Ranabhat, C., Park, M.-B., & Kim, C.-B. (2019). Influence of alcohol and red meat consumption on life expectancy: Results of 164 countries from 1992 to 2013. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3377523

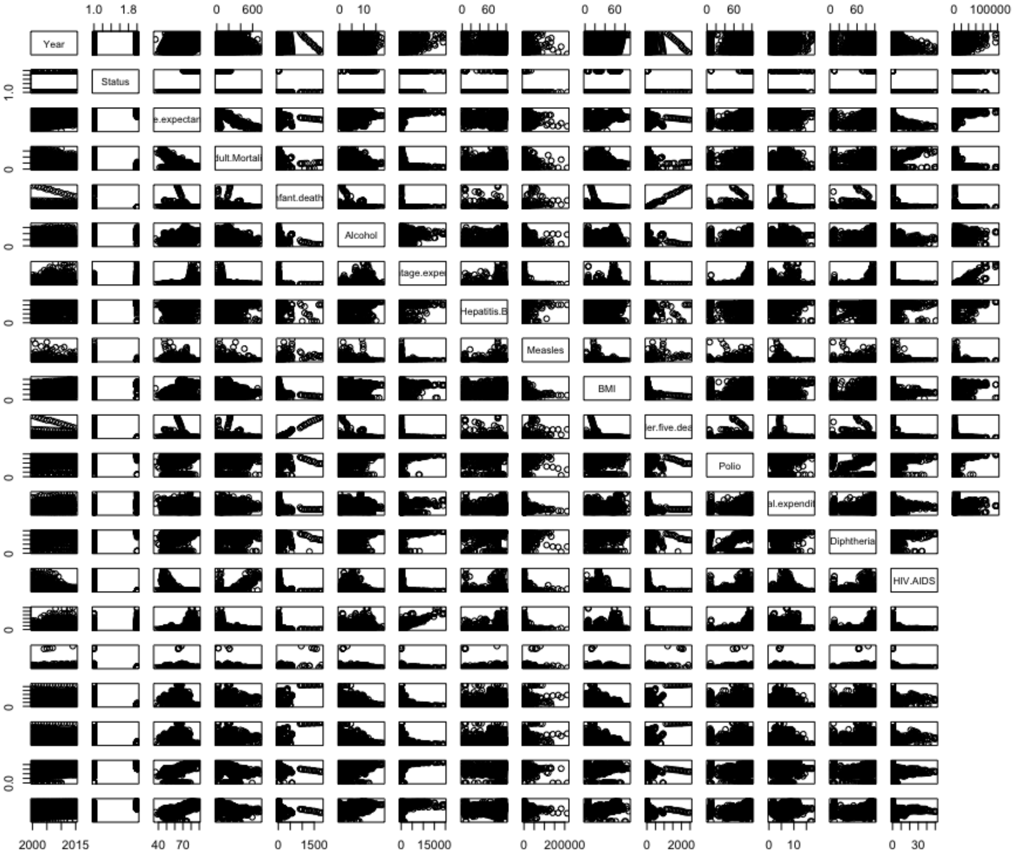
Rughiniș, C., Vulpe, S.-N., Flaherty, M. G., & Vasile, S. (2022). Vaccination, life expectancy, and trust: Patterns of COVID-19 and measles vaccination rates around the world. *Public Health*, *210*, 114–122. https://doi.org/10.1016/j.puhe.2022.06.027

Wang, Y. (2021). The greatest factors affecting life expectancy: A research based on different continents and countries. *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. https://doi.org/10.1109/mlbdbi54094.2021.00107

Appendix

Appendix A: Descriptive Statistics

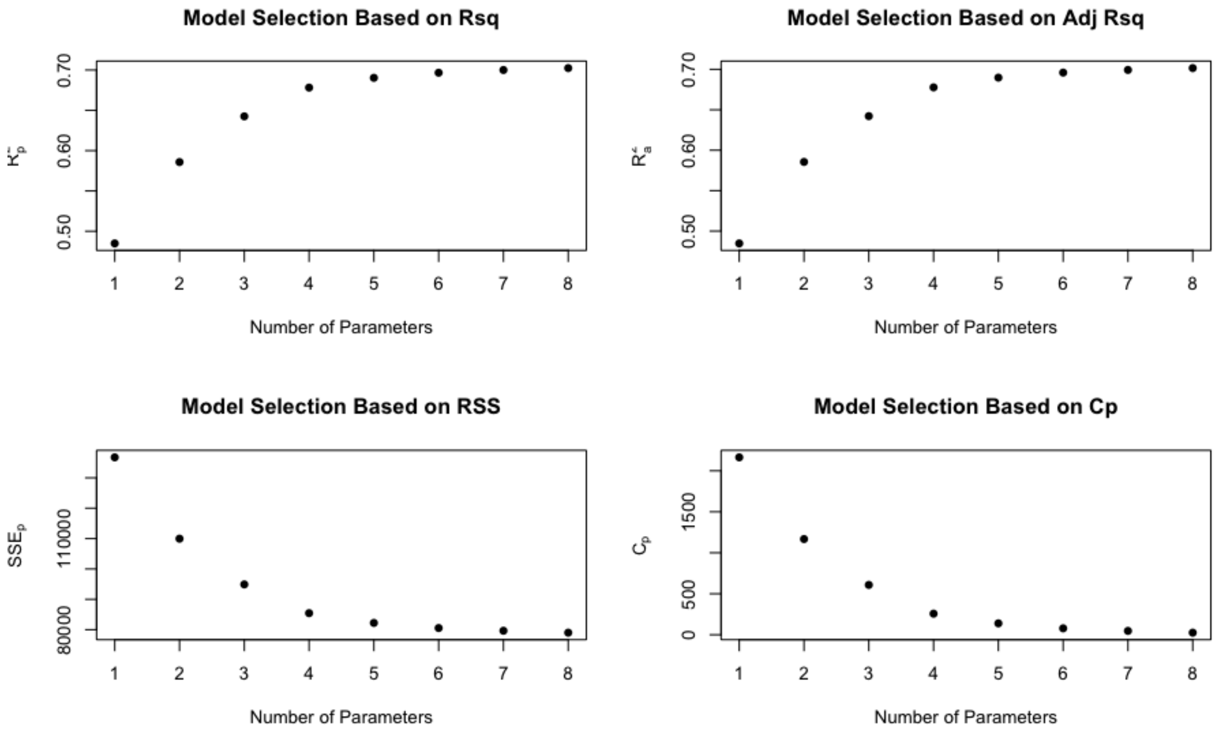
Figure A1: Full Scatterplot Matrix



Appendix B: Model Selection Criteria

Additional model selection criteria were employed to validate and confirm the outcomes derived from the Bayesian Information Criterion (BIC). Specifically, the results from the tests, including R^2 (rsq), residual sum of squares (rss), and Mallows' Cp, all favor model 8. The rsq values were 0.4849163, 0.5858158, 0.6425239, 0.6781659, 0.6903037, 0.6966308, 0.7000079, 0.7023856, where we can see model 8 has the highest (0.7024). The rss values were 136749.96, 109962.08, 94906.59, 85443.98, 82221.50, 80541.72, 79645.12, 79013.86, and again model 8 has the lowest (79013.86). For the adjusted rsq values 0.4847409, 0.5855336, 0.6421584, 0.6777270, 0.6897756, 0.6960098, 0.6992912, 0.7015728, again the highest is model 8 (0.7015728). For Mallows' Cp, the values are 2163.99045, 1167.34753, 608.08433, 257.32129, 139.18854, 78.56710, 47.14201, 25.60883, where model 8 is again favored with the lowest value (25.60883).

Figure B1: Model Selection Criteria Plots



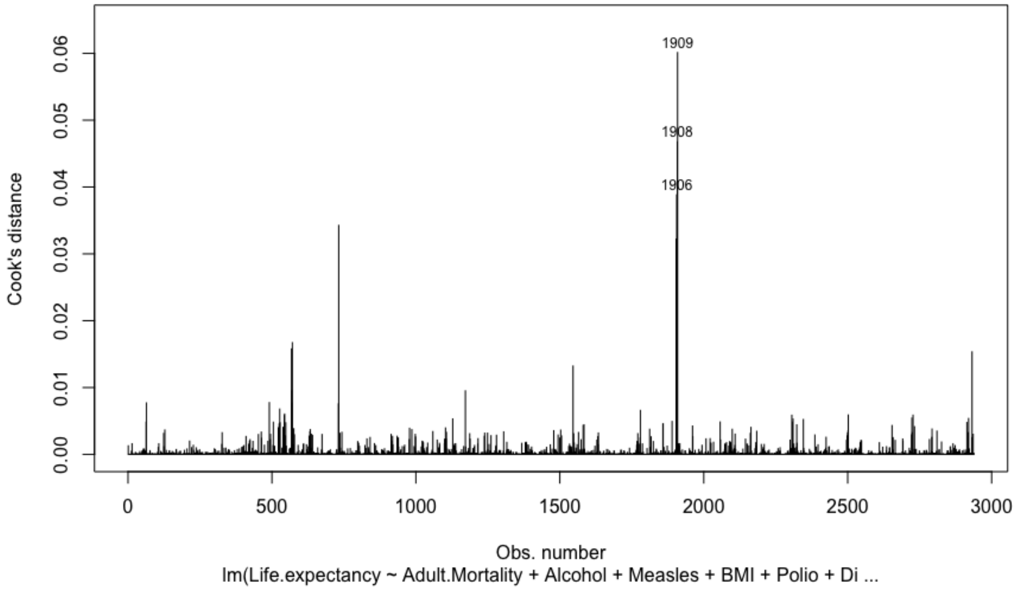
Appendix C: Addressing Outliers

In addressing outliers that were visually apparent in the initial plots of the predictor variables (Figure 1), an investigation was taken to determine their overall influence on the analysis. Cook's distance, a metric used for detecting influential data points, was employed to quantify the influence level (Figure C1). The threshold was calculated using Cook's distance formula:

Threshold = 4/n

In an effort to mitigate the influence of these outliers, they were removed from the dataset. This process aimed to make the overall analysis more robust to outliers, and to try and result in a more accurate representation of the underlying patterns in the variables of interest.

Figure C1: Outliers by Cook's Distance



Appendix D: R Code

```
############################################################
# Required R Packages
############################################################
library(tidyverse)
library(ggplot2)
library(dplyr)
library(leaps)
library(knitr)
library(car)
library(caret)

############################################################
# Load Data into R
############################################################

data <- read.csv("Life Expectancy Data.csv", na = "NA")

############################################################
# Filling in Missing Data
############################################################

data <- data %>%
  mutate(across(everything(), ~ifelse(is.na(.), mean(., na.rm = TRUE), .)))
data$Status <- factor(data$Status, levels = c("Developing", "Developed"))

view(data)

############################################################
# Analysis of potential predictors
############################################################
# Dependent  Variable: Life Expectancy
# Independent Variables: Adult Mortaliy, Infant Deaths, Alcohol Consumption, Hepatitis B,
# Measles, Polio & Diptheria Vaccination rates, BMI, Country Status, GPD

par(mfcol = c(2,2))

boxplot(data$Adult.Mortality, horizontal = T, main = "(a) Boxplot of Adult Mortality Rate",
        col = "white")

boxplot(data$infant.deaths, horizontal = T, main = "(c) Boxplot of Infant Deaths",
        col = "white")

boxplot(data$Alcohol, horizontal = T, main = "(b) Boxplot of Alcohol Consumption",
        col = "white")

boxplot(data$Hepatitis.B, horizontal = T, main = "(d) Boxplot of Hepatitis B Vaccination Rate",
        col = "white")

par(mfcol = c(2,3))

boxplot(data$Measles, horizontal = T, main = "(e) Boxplot of Measles Vaccination Rate",
        col = "white")
```

```
boxplot(data$Polio, horizontal = T, main = "(h) Boxplot of Polio Vaccination Rate",
       col = "white")

boxplot(data$Diphtheria, horizontal = T, main = "(f) Boxplot of Diptheria Vaccination Rate",
       col = "white")

boxplot(data$BMI, horizontal = T, main = "(i) Boxplot of Body Mass Index",
       col = "white")

boxplot(data$GDP, horizontal = T, main = "(g) Boxplot of GDP",
       col = "white")

predictor_names <- colnames(model.matrix(mlm))

# Calculate correlation coefficients with the response variable
cor(data$Life.expectancy, data$Adult.Mortality)
cor(data$Life.expectancy, data$infant.deaths)
cor(data$Life.expectancy, data$Alcohol)
cor(data$Life.expectancy, data$Hepatitis.B)
cor(data$Life.expectancy, data$Measles)
cor(data$Life.expectancy, data$BMI)
cor(data$Life.expectancy, data$Polio)
cor(data$Life.expectancy, data$Diphtheria)
cor(data$Life.expectancy, as.numeric(data$Status))
cor(data$Life.expectancy, data$GDP)

# Basic Statistics for Predictors
dependent_vars <- c('Adult.Mortality',
                    'infant.deaths',
                    'Alcohol',
                    'Hepatitis.B',
                   'Measles',
                   'Polio',
                   'Diphtheria',
                   'BMI',
                   'GDP')

summary_data <- data.frame()

for (var in dependent_vars) {
  summary_stats <- data %>%
    summarise(
      Variable = var,
      Mean = mean(.data[[var]], na.rm = TRUE),
      Median = median(.data[[var]], na.rm = TRUE),
      SD = sd(.data[[var]], na.rm = TRUE),
      Q1 = quantile(.data[[var]], 0.25, na.rm = TRUE),
      Q3 = quantile(.data[[var]], 0.75, na.rm = TRUE),
      Min = min(.data[[var]], na.rm = TRUE),
      Max = max(.data[[var]], na.rm = TRUE),
      N = sum(!is.na(.data[[var]]))
    )
  summary_data <- bind_rows(summary_data, summary_stats)
}
```

```
kable(summary_data, format = "html")

# Response Variable
boxplot(data$Life.expectancy, horizontal = T, main = "(a) Boxplot of Life Expectancy",
        col = "white")
hist(data$Life.expectancy, col = "lightblue",
     xlab = "Life Expectancy",
     ylab = "Frequency",
     main = "(b)")

##############################################################
# Model Selection
##############################################################
# Independent Variables: Adult Mortaliy, Infant Deaths, Alcohol Consumption, Hepatitis B,
# Measles, Polio & Diptheria Vaccination rates, BMI, Country Status, GPD

numeric_data <- data[sapply(data, is.numeric)]

mlm <- lm(Life.expectancy~Adult.Mortality + infant.deaths + Alcohol
                + Hepatitis.B + Measles + BMI + Polio + Diphtheria + Status + GDP, data = data)
summary(mlm)
m1 <- lm(life.expectancyAdult.Mortality + infant.deaths + Alcohol
+ Hepatitis.B + Measles + BMI + Polio + Diphtheria + Status + GDP, data = data)
cor(model.matrix(mlm))
plot(Diphtheria~Hepatitis.B, data = data, xlab = "Hepatitis B", main = "(a)")
abline(lm(Diphtheria~Hepatitis.B, data = data))
plot(Polio~Hepatitis.B, data = data, main = "(b)")
abline(lm(Polio~Hepatitis.B, data = data))
qqnorm(residuals(mlm))
qqline(residuals(mlm))

ma <-
  regsubsets(Life.expectancy~Adult.Mortality + infant.deaths + Alcohol
            + Hepatitis.B + Measles + BMI + Polio + Diphtheria + Status + GDP, data = data)
sma <- summary(ma)
par(mfcol = c(2,2))

sma$rsq  # Model 8
plot(1:8,sma$rsq, pch=16, xlab="Number of Parameters",ylab=expression(R[p]^2), main = "Model Selection Based
on Rsq")

sma$rss # Model 8
plot(1:8,sma$rss, pch=16, xlab="Number of Parameters",ylab=expression(SSE[p]), main = "Model Selection Based
on RSS")

sma$adjr2 # Model 8
plot(1:8,sma$adjr2, pch=16, xlab = "Number of Parameters", ylab = expression(R[a,p]^2), main = "Model Selection
Based on Adj Rsq")

sma$cp # Model 8
plot(1:8, sma$cp,  pch=16, xlab = "Number of Parameters", ylab = expression(C[p]), main = "Model Selection
Based on Cp")
abline(0,1)

sma$bic # Model 8
```

```
plot(1:8, sma$bic,  pch=16, xlab = "Number of Parameters", ylab = expression(BIC[p]))

full.model <- lm(Life.expectancy~Adult.Mortality + Alcohol + Measles + BMI + Polio + Diphtheria + Status +
GDP, data = data)
cor(numeric.data)
summary(full.model)$adj.r.squared

residuals <- residuals(full.model)

# R squared = 0.702, accounts for ~70.2% of variance in life expectancy

model.variables <- all.vars(formula(full.model))
model.data <- newdata[,model.variables]
str(model.data)
model.data$Measles <- as.numeric(model.data$Measles)
model.data$Status <- as.factor(model.data$Status) # 2 = developing, 1 = developed
model.data$Status <- as.numeric(model.data$Status)

cor_matrix <- cor(model.data)
pairs(numeric_data)

plot(full.model)

## Summary:
# R squared is good
# Appears to be som issues with fit -- may need to transform some variables, check outliers

# Model Assumptions ##

# OUTLIERS
outlierTest(full.model)
# Identified 2502, 2310, 2501, 2500, 2308, 2425, 2936, 1581, 1584, 2912
threshold <-  4/(nrow(data)
all.cooks.distance <- cooks.distance(full.model)
distance.1909 <- all.cooks.distance[1909]
distance.1908 <- all.cooks.distance[1908]
distance.1904 <- all.cooks.distance[1904]
distance.2502 <- all.cooks.distance[2502]
distance.2310 <- all.cooks.distance[2310]
distance.501 <- all.cooks.distance[501]
distance.2501 <- all.cooks.distance[2501]
distance.2500 <- all.cooks.distance[2500]
distance.2308 <- all.cooks.distance[2308]
distance.2425 <- all.cooks.distance[2425]
distance.2936 <- all.cooks.distance[2936]
distance.1581 <- all.cooks.distance[1581]
distance.1584 <- all.cooks.distance[1584]
distance.2912 <- all.cooks.distance[2912]
plot(full.model, which=4, cook.levels=threshold)
## Points 1909, 1908 and 1906 exceed this threshold
outliers <- all.cooks.distance > threshold
influential.points <- c(571, 525, 732, 2502, 522, 2310, 2164, 2501, 2500, 2793, 2308, 2425, 2936, 1581,2920, 2932,
2498, 2306,2931,1584, 2912, 1906,1908,1909, 2919)
# points that are highly influential according to cook's distance
```

```
newdata <- data[-influential.points, ]

full.model2 <- lm(Life.expectancy~Adult.Mortality + Alcohol + Measles + BMI + Polio + Diphtheria + Status + GDP, data = newdata)
plot(full.model2, which = 2)

plot(residuals(full.model2)~fitted(full.model2), xlab = "Fitted", ylab = "Residuals", xlim = c(45,85), ylim = c(-12,10))

## Issues with non-linearity
# shapiro wilk

log_Alcohol <- log(newdata$Alcohol)
sqrt_BMI <- sqrt(newdata$BMI)
boxcox_results <- boxcox(Life.expectancy ~ Adult.Mortality + log_Alcohol + Measles + sqrt_BMI + Polio + Diphtheria + Status + GDP, data = newdata)
lambda_value <- boxcox_results$x[which.max(boxcox_results$y)]
transformed_Alcohol <- (newdata$Alcohol^lambda_value - 1) / lambda_value
interaction_BMI_Polio <- newdata$BMI * newdata$Polio
squared_BMI <- newdata$BMI^2

full_model_new <- lm(Life.expectancy ~ Adult.Mortality + transformed_Alcohol + Measles + sqrt_BMI + Polio + Diphtheria + Status + GDP + interaction_BMI_Polio + squared_BMI, data = newdata)
plot(full_model_new)
residuals_new <- residuals(full_model_new)

shapiro.test(residuals_new)

## The transformations are barely improving, indicating that maybe another model may better fit the data

## LINEARITY
plot(Life.expectancy~Adult.Mortality, data = newdata)
abline(lm(Life.expectancy~Adult.Mortality, data = newdata))
# Looks slightly curvilinear

plot(Life.expectancy~Alcohol, data = newdata)
abline(lm(Life.expectancy~Alcohol, data = newdata))

plot(Life.expectancy~Measles, data = data)
abline(lm(Life.expectancy~Measles, data = data))
# Very skewed, transformation
plot(Life.expectancy~log(Measles), data = data)
abline(lm(Life.expectancy~Measles, data = data))
# Looks better

plot(Life.expectancy~BMI, data = data)
abline(lm(Life.expectancy~BMI, data = data))

plot(Life.expectancy~Polio, data = data)
abline(lm(Life.expectancy~Polio, data = data))

plot(Life.expectancy~Diphtheria, data = data)
abline(lm(Life.expectancy~Diphtheria, data = data))

plot(Life.expectancy~Status, data = data)
```

```
abline(lm(Life.expectancy~Status, data = data))

plot(Life.expectancy~GDP, data = data)
abline(lm(Life.expectancy~GDP, data = data))

# maybe curvilinear, transformation
data$GDP.squared <- data$GDP^2
plot(Life.expectancy~GDP.squared, data = data)
# Worse
plot(Life.expectancy~log(GDP), data = data)
abline(lm(Life.expectancy~log(GDP), data = data))

## Independence of Errors
plot(residuals(full.model2)~fitted(full.model2), data = newdata)
# looks a bit clustered

# Normality
residuals <- residuals(full.model2)
qqnorm(residuals)
qqline(residuals)
# left tail is heavy

vif(full.model2)
vif(mlm)
# no perfect multicollinearity

summary(full.model2)
anova(full.model2)

# I would say that while this model violates slightly the normality assumption,
# making too many transformations would skew the interpretability of the model given the circumstances

############################################################
# Model Validation
############################################################
ctrl_loocv <- trainControl(method="LOOCV")
model_loocv <- train(Life.expectancy~Adult.Mortality + Alcohol + Measles + BMI + Polio + Diphtheria + Status +
GDP
                    , data = newdata, method = "lm", trControl = ctrl_loocv)

summary(model_loocv)

predictions <- predict(full.model, newdata = newdata)

# Calculate RMSE
rmse <- sqrt(mean((newdata$Life.expectancy - predictions)^2))

# Calculate relative RMSE
relative_rmse <- rmse / (max(newdata$Life.expectancy) - min(newdata$Life.expectancy))

## Prediction1

row.index <- 1500
data.point <- newdata[row.index, ]
```

```
predicted.value <- predict(full.model, newdata = data.point)
prediction.interval <- predict(full.model, newdata = data.point,
                        interval = "prediction", level = 0.95)

# Using this we can see the prediction for life expectancy
# Is within the 95% prediction interval

## Prediction2

row.index2 <- 800
data.point2 <- newdata[row.index2, ]

predicted.value <- predict(full.model, newdata = data.point2)
prediction.interval <- predict(full.model, newdata = data.point2,
                        interval = "prediction")

## Prediction3

row.index3 <- 2005
data.point3 <- newdata[row.index3, ]

predicted.value <- predict(full.model, newdata = data.point3)
prediction.interval <- predict(full.model, newdata = data.point3,
                        interval = "prediction")

jittered.actual <- jitter(actual.values, factor = 0.3)
actual.values <- data$Life.expectancy

predicted.values.full <- predict(full.model)

plot(actual.values, predicted.values.full,
    main = "Actual vs. Predicted Values",
    xlab = "Actual Values",
    ylab = "Predicted Values",
    pch = 20)

hist(residuals, main = "Histogram of Residuals", xlab = "Residuals", col = "lightblue",
    breaks = 15, xlim = c(-15,15))

plot(residuals(full.model)~predicted.values.full, ylim = c(-10,10))
abline(h=0)
```