Global Life Expectancy Projections Group 1

Charitha Nallaka King, Joshua Taylor

Chilukuru, Kushwanth Kumar Abedalkanee, Ahamed Mohamed

# Introduction

When improving the quality of life around the world, it is important to know what factors affect life expectancy. This project aims to create a model that predicts how long people might live in different countries based on factors like health, wealth, and social status. The goal is to understand what influences life expectancy and use that knowledge to make better decisions about public health and social policies. affect life expectancy. We want to find which variables, such as diet, lifestyle, and government involvement, affect the life spans of people. Using this information, we can provide organizations with insights to make decisions that can improve the lives of people around the world.

# Data Preprocessing

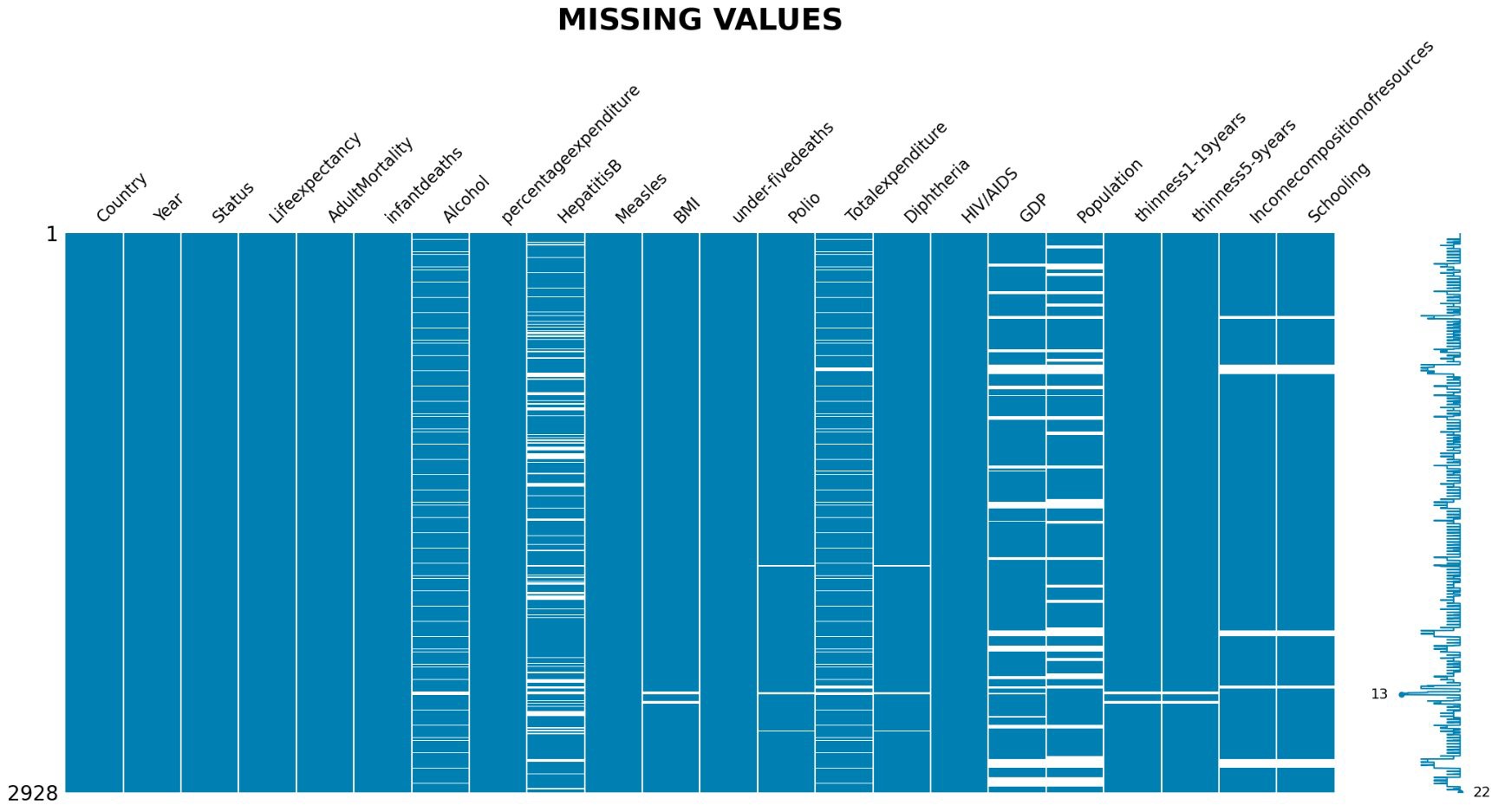
To carry out our goal, we used data collected by the World Health Organization for 193 countries from 2000 to 2015. In total, we had a dataset that included almost 3000 observations across 22 variables that cover health, social, and economic factors. The description of the different variables can be viewed below in **Table** 1.

|  |  |
| --- | --- |
| **Factor** | **Description** |
| **Name** | A country's name |
| **Status** | A country's economic class (Developed or Developing status) |
| **Year** | The year that the data was observed |
| **Life expectancy** | The number of years that an average person can be expected to  live |
| **Adult Mortality** | The probability that a person aged 15 years old will die before they  reach the age of 60 (in a population per 1000 people) |
| **Infant deaths** | The number of infant deaths (in a population per 1000 people) |
| **Alcohol** | Pure alcohol consumption per capita of persons aged over 15 in  Liters |
| **percentage**  **expenditure** | The percentage of a country's GDP per capita expended on  healthcare |
| **Hepatitis B** | The percentage of Hepatitis B immunization among children aged 1  year old |
| **Measles** | The number of Measles cases reported (in a population per 1000  people) |
| **BMI** | The average body mass index value of the country's population |
| **under-five deaths** | The number of deaths (in a population per 1000 people) of children  under 5 years of age |
| **Polio** | The percentage of Polio immunization among children aged 1 year  old |
| **Total expenditure** | The percentage of total government expenditure allocated to  general healthcare expenditure |
| **Diphtheria** | The percentage of Diphtheria, Tetanus, Toxoid, Pertussis (DTP3)  immunization among children aged 1 year old |

|  |  |
| --- | --- |
| **HIV/AIDS** | The number of HIV/AIDS deaths (per 1000 live births) of children  under 4 years of age |
| **GDP** | The Gross Domestic Product value per capita calculated in USD |
| **Population** | The country's population |
| **thinness 1-19 years** | The percentage thinness of children aged between 1 and 19 years |
| **thinness 5-9 years** | The percentage thinness of children aged between 5 and 9 years |
| **Income composition**  **of resources** | The Human Development Index value (0-1) in reference to a  country's configuration of resources |
| **Schooling** | The mean number of years of schooling |

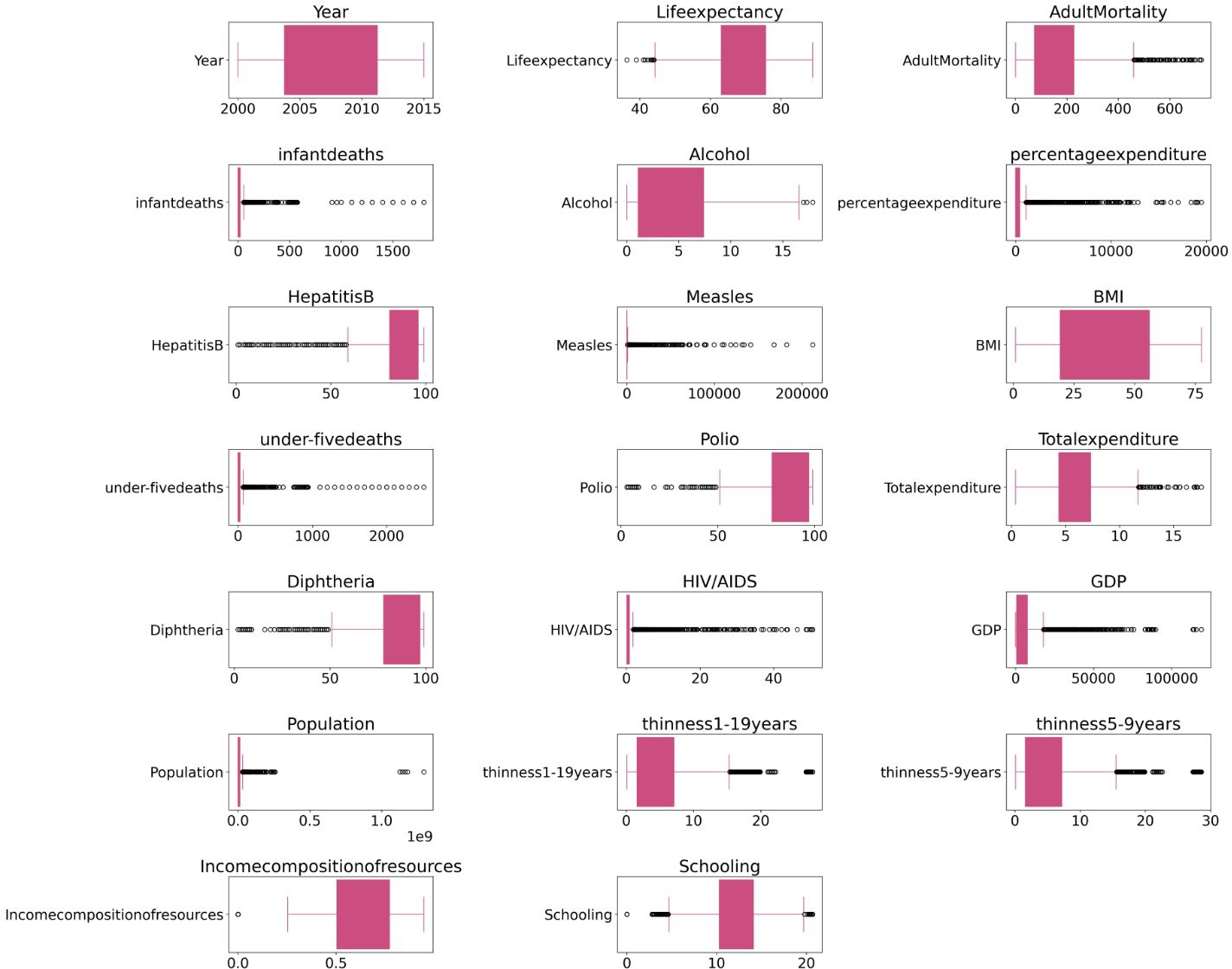
# Table 1

Before we could build our model, we needed to clean the data to eliminate any holes or errors that may exist. The first area we wanted to address was countries that did not have many observations. When analyzing the dataset, we saw most of the countries had 16 observations, corresponding to the 16 years for the period. However, there were some countries that only had one observation. To provide analysis that would reflect changing over time, we decided to remove the countries with only one observation from consideration.



# Figure 1

Once the countries were removed, we wanted to examine the dataset for any missing values or errors. As seen in **Figure 1,** the white spaces are any missing values for all the variables. For the values that had missing values, we imputed the average or median for the corresponding variable. If the variable had a normal distribution, we would impute with the average, and, if the variable had a skewed distribution, then we imputed with the median. This would let us fill out the data set without creating inconsistent numbers of observations across all countries.

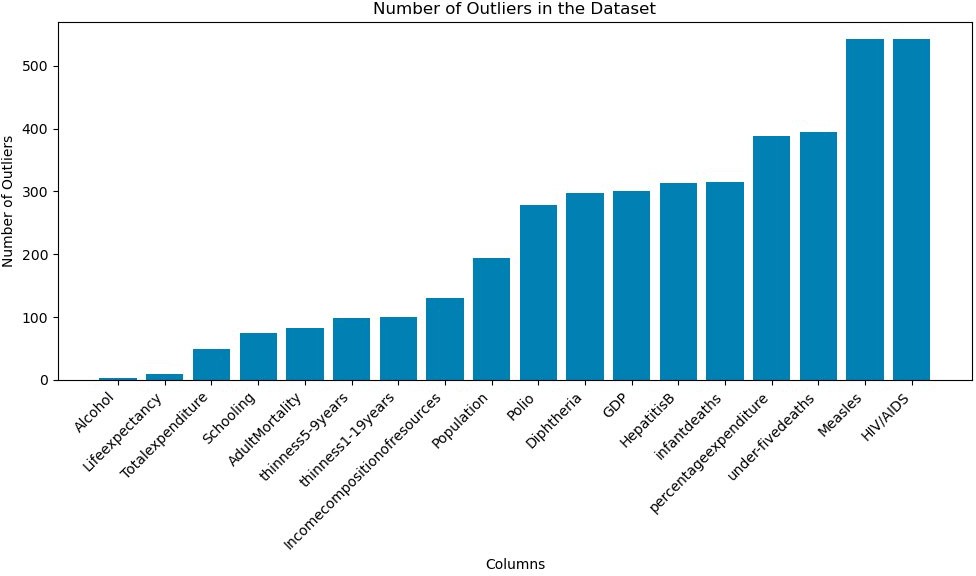


After the missing values were addressed, we wanted to address the outliers for each variable. Using the Inter Quartile Range, or IQR, we found all the outliers for each variable. The box plots in Figure 2 show each variable and note the outliers for all remaining countries. **Figure**

**3** shows the total number of outliers for each variable. To address the outliers, we replaced the outliers with the mean values until no outliers were remaining. Once the outliers were removed, we could begin performing exploratory data analysis.

**Exploratory Data Analysis**

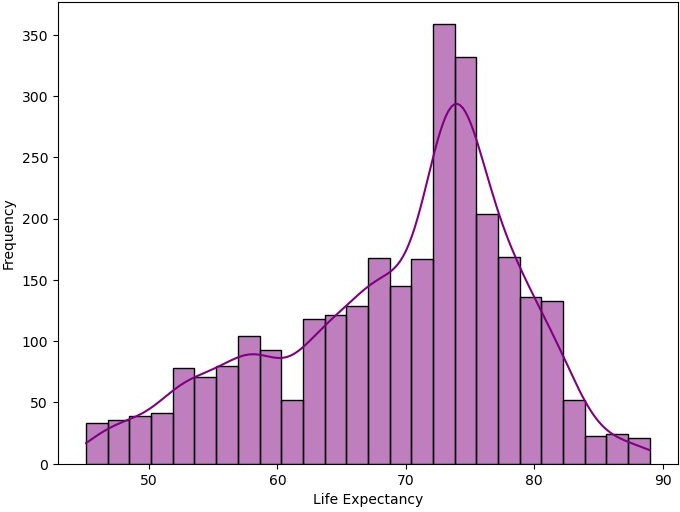
**Figure 2**



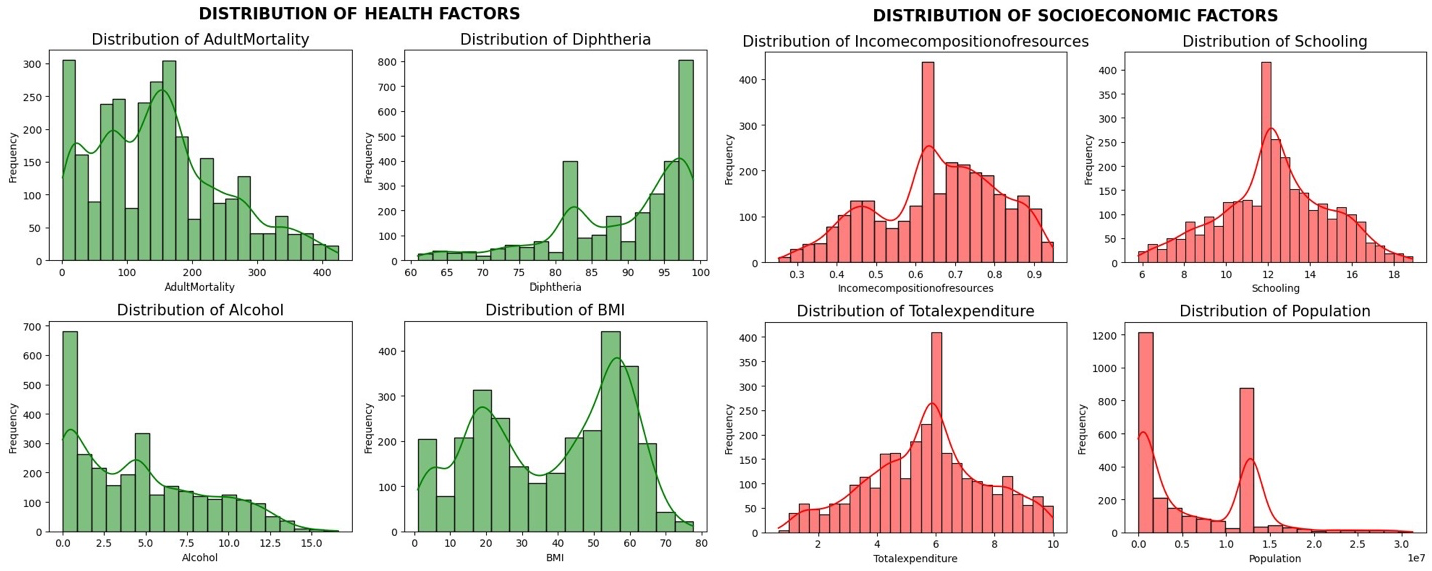
**Figure 3**

When performing the data analysis, we focused on two areas: univariate and bivariate. For the univariate analysis, we wanted to look at the distribution the individual variables. The first distribution we looked at was for our target variable, Life Expectancy. As seen below in **Figure 4,** the data was normal, with a peak between 70 and 80 years. When looking at the other variables, we see that there is a mix between normal and skewed distributions. Most of the socioeconomic variables were normally distributed, but the health factors had skewed

distributions. The variables with skewed distributions could negatively affect the accuracy of our model. These variables may need to be watched to see the effect on our model.

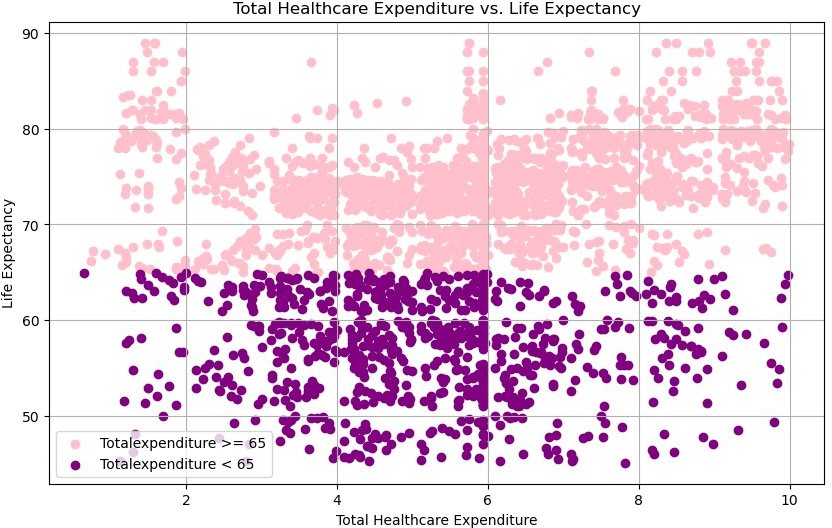
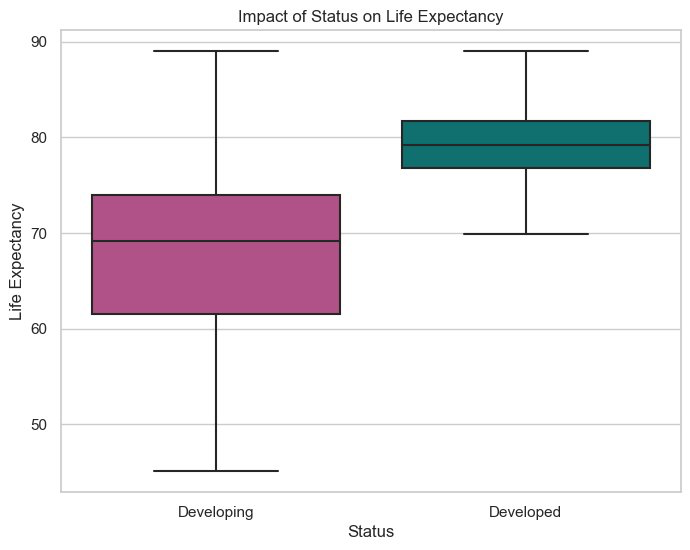


# Figure 4



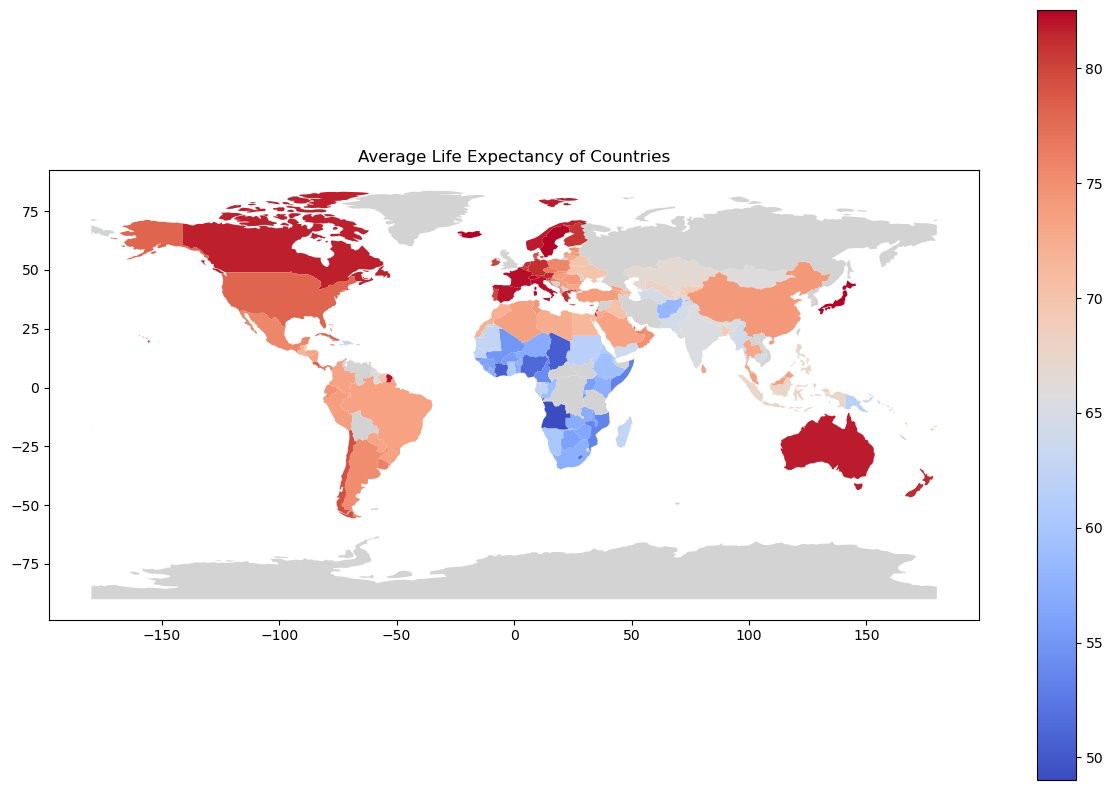
**Figure 5 Figure 6**

After addressing the imbalance, we performed bivariate analysis. The first variables we wanted to analyze were comparing “Life Expectancy” with “Status” and “Health Expenditure”. We noticed that the spread of the life expectancy for the developed country was a lot smaller than the life expectancy for the developing countries. However, the maximum values for the two different statuses were remarkably close. In the relationship between “Life Expectancy” and “Health Expenditure”, we noticed that the percentage that a government spends on healthcare has does not appear to have much of correlation with life expectancy. However, when looking at the percentage of a country’s GDP, there appears to be a clear line for life expectancy. The increasing the percentage of GDP spending on healthcare may be factor in increasing life expectancy.

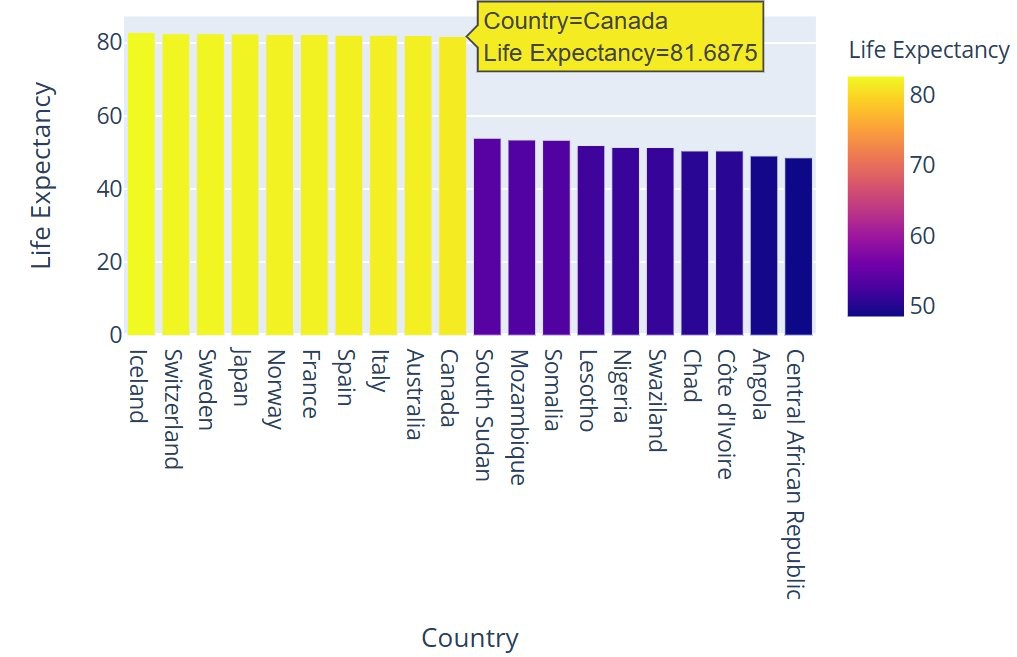


# Figure 7 Figure 8

In our analysis, we also compared the average life expectancy for the various countries.

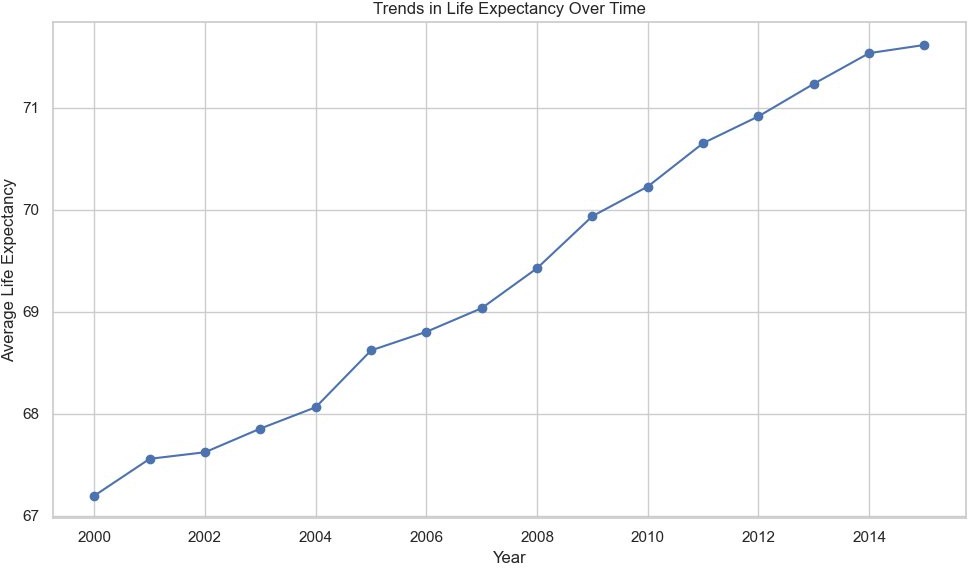
We noticed that most of the countries that had higher life expectancies were in Europe and North America. The ten countries with the highest life expectancy, except for Japan and Australia, were in Europe and North America. The ten countries that had the lowest life expectancy were in Africa. When looking at increasing life expectancy around the world, using the model could help countries in Africa to improve life expectancy.

# Figure 9



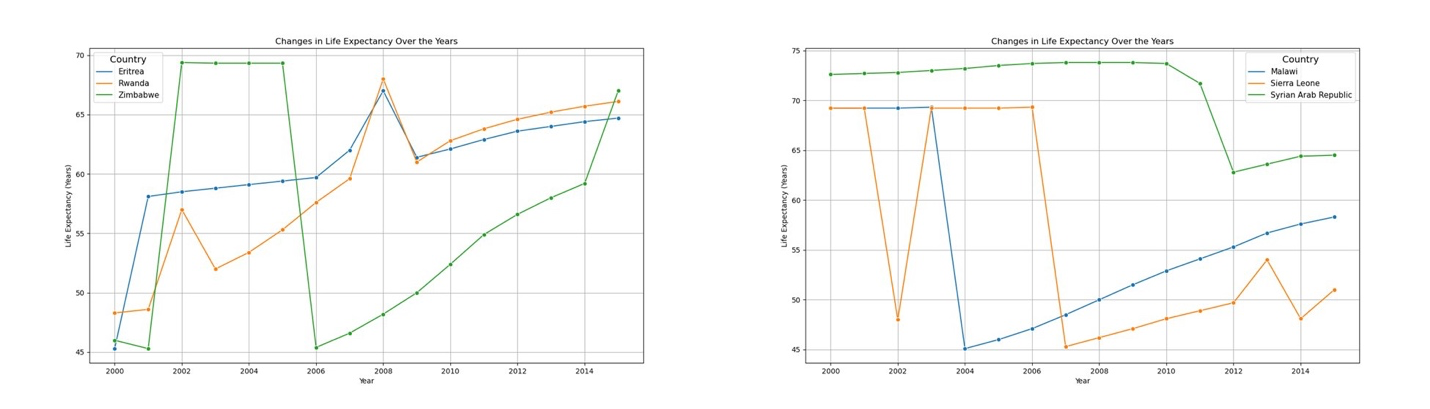
**Figure 10**

When analyzing the life expectancy in different countries, we also wanted to look at the change in life expectancy over time. As seen in **Figure 11**, the average life expectancy was increasing over the 16-year period. The increase over time appears to follow a linear pattern. This relationship could be used to predict average life expectancy if no added actions are taken. Collected more recent data could also offer more insight. More data could show if the linear pattern continues or if there will be a natural leveling. In either case, there could be useful information that can be applied to our model.



# Figure 11

In addition to analyzing the average life expectancy per country, we also looked at the change in life expectancy per year for each country. For the 16-year period, we found the percent change in life expectancy and the average yearly percent change in life expectancy for each country. Interestingly, the three countries with the highest average percent change in life expectancy were in Africa. The rapidly growing life expectancy could help provide insights and be reflective of recent changes. We also saw that the countries with the lowest average percent change for life expectancy were also from the same region. With the same region having both the top and bottom 3, we wanted to look further and see if there were any noticeable changes. We created graphs to display the life expectancy for each year for the 6 countries we found. The graphs can be seen in **Figure 12** and **13.**

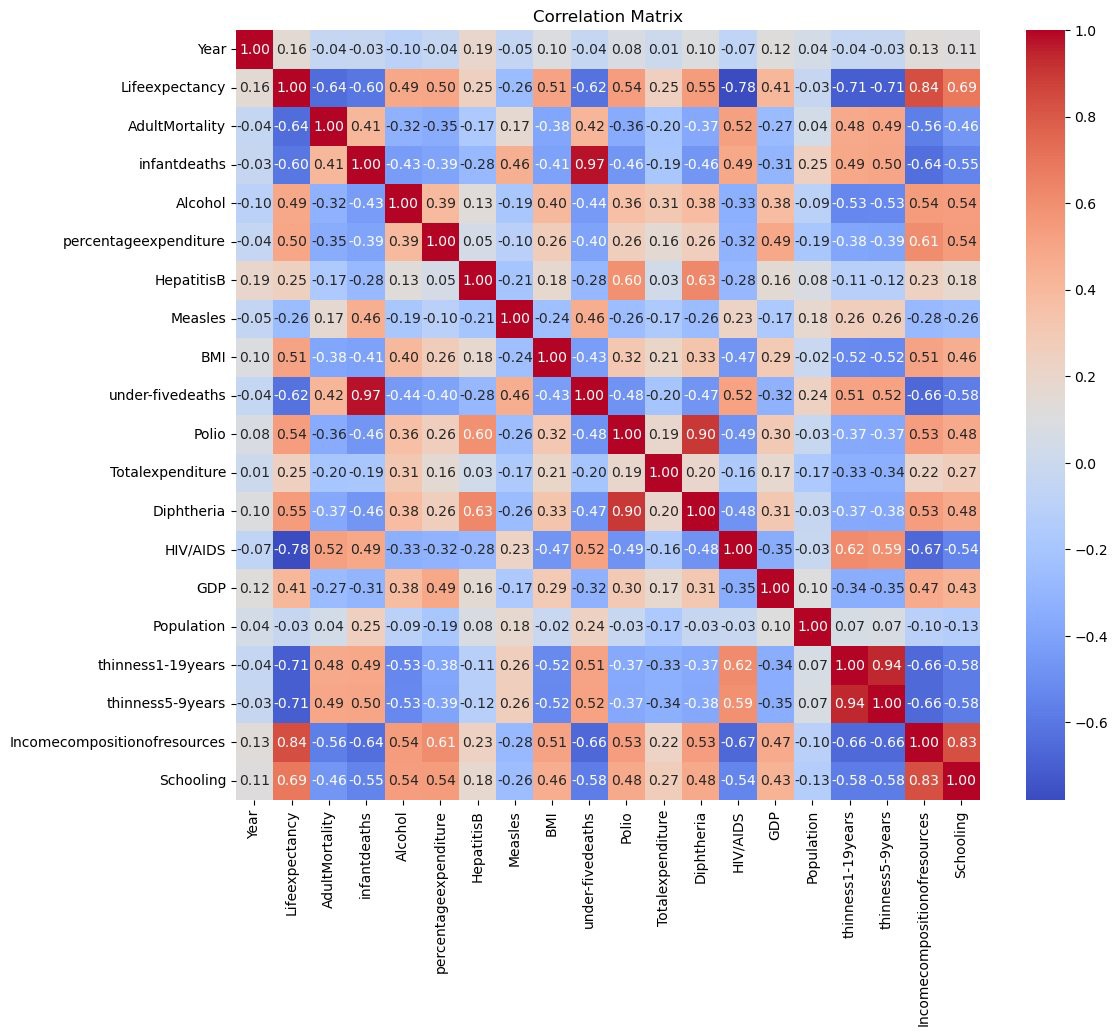


# Figure 12 Figure 13

When looking at the graph for the countries with the highest and lowest average percent change in life expectancy, we noticed some areas of concern. For each of the six countries, we saw that there were places where the life expectancy would drastically change in one year before continuing in a pattern. For these countries, drastic drops in life expectancy could be explained by a war or a natural disaster. However, these factors would not be considered in the model.

These rapid changes could impact the accuracy of our model. While these are just 6 countries, a deeper analysis may need to be performed to see if it occurs in other countries.

Before we built our model, the last bivariate analysis we wanted to perform was for a correlation matrix. This would let us compare all the variables against each other and see how correlated they are. Performing this analysis would allow us to see if there is any multicollinearity within our dataset. When we performed the correlation matrix, we were able to find a few variables that were highly correlated by having a correlation coefficient greater than 0.8. These variables could negatively affect the accuracy of our model, so we would need to handle any multicollinearity. If two variables were shown to be highly correlated, we would drop one of the variables. Once this was completed, we began to build the model.



# Figure 14

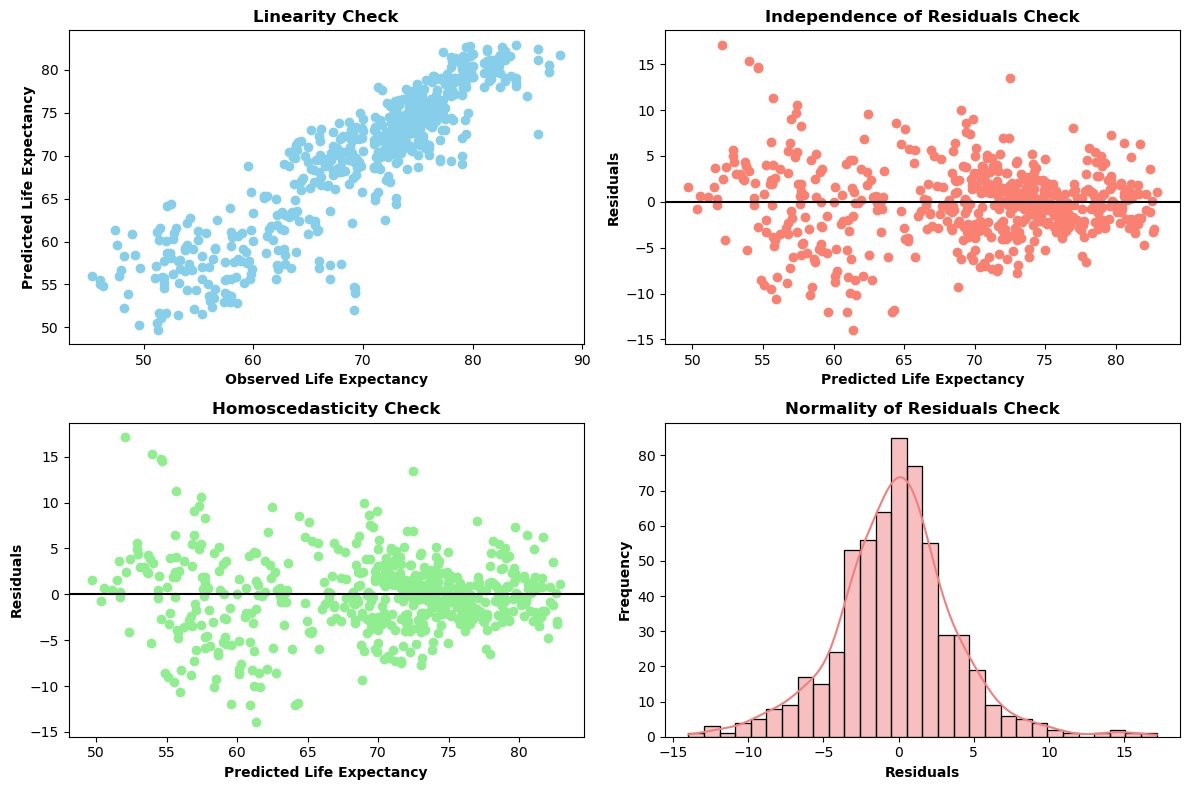
**Model**

For our model, we decided to perform a multilinear regression. We set up our target

variable as “Life Expectancy”. When the model was completed, we found the variables that had the greatest impact on life expectancy. These variables would have the greatest coefficients, therefore having the greatest impact on life expectancy. The top variables were “Income Composition of Resources”, “HIV/AIDS Prevalence”, and “Developing Status”. We saw that increasing the diversity of a country's resources would predict an increase in life expectancy.

We also saw that an increase in prevalence of HIV/AIDS would predict a decrease in life expectancy. We also saw that the increase in the percentage of GDP and government spending on healthcare would predict an increase in life expectancy. This information could be used to aid countries in increasing life expectancy.

After we built our model, we performed checks to ensure the linear regression was the best model for analysis. We analyzed the relationship between the predicted life expectancy and observed life expectancy to perform a linearity check. The graph showed a linear pattern, confirming linearity. We also checked the independence of residuals and homoscedasticity by comparing the residuals with the predicted values. In both residual plots, there was no noticeable pattern. The last check we performed was a distribution of residuals. The frequency chart displayed a normal histogram. With all these checks, we can say that linear regression is a good model of our analysis.



# Figure 15

To evaluate the model, we divided our dataset so 80% of observations were used to build the regression model and 20% was used to test the accuracy. We calculated the R2, correlation coefficient, Mean Square Error, or MSE, Root Mean Square Error, or RMSE, and Mean Absolute Erro, or MAE, to test evaluate the model. The R2 indicates how well the model explains the variance in the target variable and is a value between zero and 1, with 1 being the desired value. Our R2 value for the training dataset was 0.8193 and the testing dataset was 0.8047.

Since both values were greater than 0.8, the variables have a strong relationship with the target variable. The MSE is the average squared difference between predicted and actual values. The MSE for our training data was 15.76 and our testing data 15.99. The RMSE is the square root of MSE. Our RMSE for our two datasets were 3.97 and 3.99, respectively. The MAE is the average absolute difference between predicted and actual values. Our model for the training data had MAE of 2.87 and the testing data set was 2.92. For both MSE and MAE, the low value shows that our model predicted values within 4 years of the actual data. Our model would be able to provide accurate predictions to make informed decisions to improve life expectancy.

# Conclusion

After compiling our multiple linear regression model, we would be able to perform accurate predictions to help countries increase life expectancy. Our model suggests that increasing healthcare spending can improve life expectancy in countries with low averages. We also saw that immunization and other factors, like diet, lifestyle, and alcohol consumption, can affect life expectancy. Countries that had higher income, education, immunization, GDP, and lower disease rates had higher life expectancy. While we did see that increased education positively affected lifespan, densely populated areas may see lower life expectancies. Updating the dataset to include more years could help provide more information that could also aid countries with an already high life expectancy.

Despite the accurate analysis, there are some improvements that could be made to ensure the best model for improving life expectancy. One area that may need to be checked is looking more into the value provided. When looking at individual countries, we saw drastic changes in life expectancy. Understanding what factors could have caused this could provide more insight for analysis. There could be factors such as war or natural disasters that would not directly appear in the dataset but could have an impact on the model. Another improvement could be with the model itself. Looking at other models to ensure we are using the best one for analysis. Whether it is another type of regression or even other models, such as decision trees, random forest, or neural networks, could provide different results. Exploring these different models could provide a more accurate model, or even verify that our current linear regression model is the best way to analysis the effects of life expectancy. We could even combine multiple models for analysis.