Analysis of National Life Expectancies

STAT 4355.001

May 13, 2021

Group 11 Abraca-Data

Abraca-Data Team Members: Prachi Patel, Evan Meade, Alejandro De La Cruz

The Dataset:

Though this data was originally collected by WHO and the United Nations (UN), we sourced it from Kaggle. This was done for convenience, as a user had already downloaded and combined the data into a convenient .csv form. As a result, we will have more time for analysis as we will be spending less time hunting data.

As stated in the goal, this dataset contains 2938 observations of 22 variables. The data contains large spatial and temporal baselines; it includes observations of **193 countries ranging from 2000-2015**. Each country is additionally classified as "developed" or "developing," providing opportunity for comparative analysis of these two groups. While not all (country, year) pairs have complete data, approximately half of them do, and many more have partially complete data. This allows us to assess the effects of time and place on life expectancy.

The main content of the dataset is a collection of health and economic indicators that measure societal factors which are potentially relevant to predicting life expectancy. The health factors include mortality rates (adult, children, infant, etc.), disease rates (HIV/AIDS, polio, measles, etc.), and national medical averages (life expectancy, BMI, alcohol consumption, etc.). Together, these figures provide a snapshot of the leading ailments in a country and the average person's health. Meanwhile, the economic factors track GDP, government health expenditures, and population, as well as average income and education levels. These factors give insights into the resources available to the government as well as the average citizen to face the medical ailments captured by the health indicators. So, taken altogether, this dataset provides a coarse view of the resources invested by countries to address public health issues. As such, by analyzing it, we can assess what resources/ailments have the highest impact on life expectancies.

Goal:

This dataset containing **22 variables** and **2938 entries** is about the life expectancy in different countries. The dataset has gathered information from **193 countries**, which has been collected from the World Health Organization (WHO) data repository website, who keep track of health status and many other factors for all countries. Our goal with this project is to investigate the key factors that help improve a countries' overall life expectancy. We want to find out what happens to the life expectancy in a country when certain statuses change. For instance, if the amount of schooling a person gets in a country increases or decreases, how will that affect the life expectancy in the country? By understanding why a certain country has a high or low life expectancy, we can figure out how to improve life for them using the information found.

Data Cleaning:

We began our cleaning by identifying which factors were most highly correlated with life expectancy, which led us to a set of 10 predictors with correlation coefficients greater than 0.4 in magnitude (this process is detailed below in Question 1). This was a practical consideration as we wanted to use our time investigating the most promising predictors in such a large set. After feedback from the professor, we decided to exclude adult mortality, as it is actually a component of the life expectancy calculation, and thus an unfair predictor. So, we took our remaining 9 predictors along with year, country, development status, and life expectancy to be our initial data subset.

Having selected our variables of interest, we examined missing values. Unfortunately, they were not randomly distributed, so any attempt at interpolation would be extremely difficult. In other words, some countries didn't have any measurements of a given predictor in the dataset. So interpolation could not be done at a national level, it would have to be done at a global level, which would lead to more significant biases than we were comfortable with. In the end, we decided to simply exclude the ~600 entries with missing predictor values out of our ~2,900 observation subset. We believe it was the fairest decision given our level of statistical education, though we note such an action introduces a potential selection bias into the analysis. We must recognize that we may be analyzing not all countries in general, but those with more developed infrastructure and open governments which facilitate easier recordkeeping. However, given that we still had over 150 countries in our final dataset, we believe this bias to be not unreasonable.

Question 1: *Which predictors are most correlated with life expectancy?*

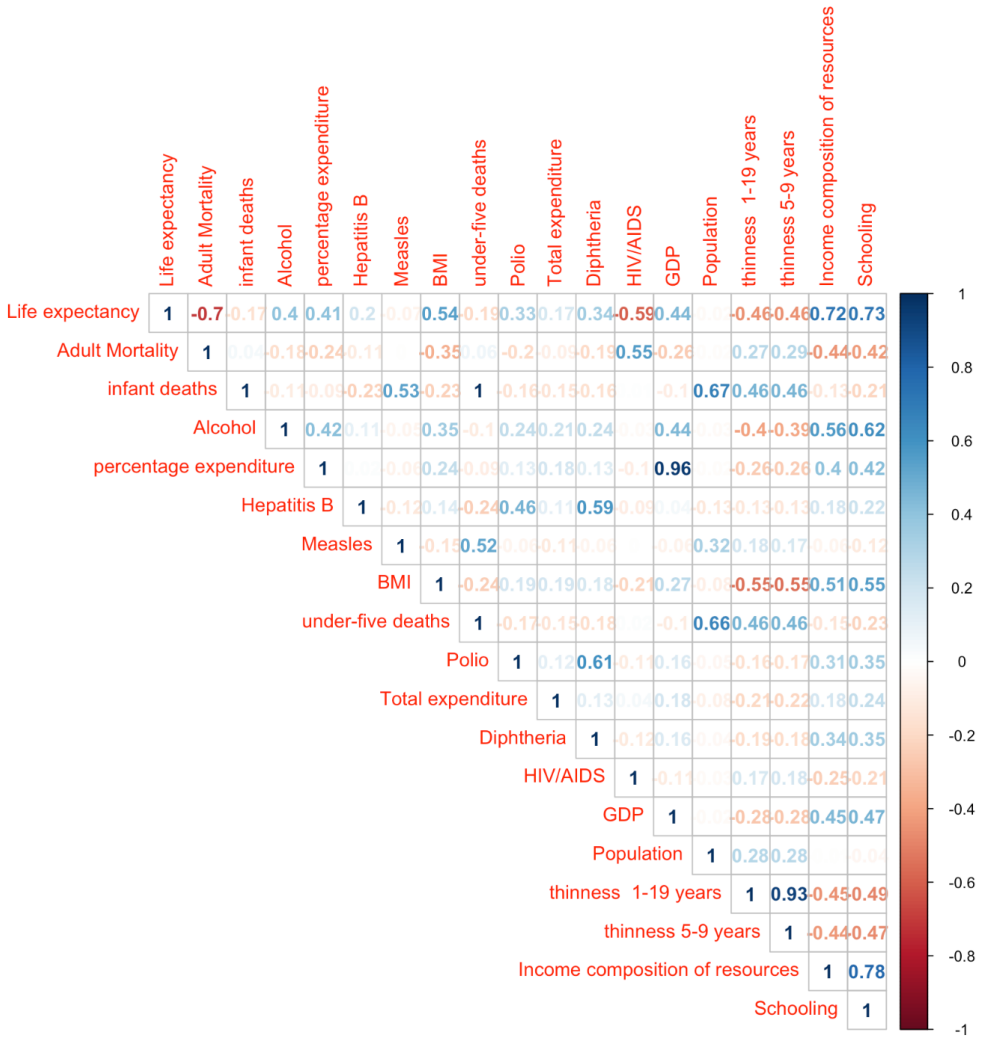

Figure 1.1

For this analysis there are ten features within the dataset that we would like to choose to predict the **Life Expectancy**, which is our response variable. These ten features represent the predictors with the highest magnitude correlations with life expectancy ($R^2 > 0.4$). As such, they are promising candidates for developing a predictive model, but are obviously subject to change as our analysis progresses. We have performed some minor "*cleaning*" since there is about **1%** to **5%** of missing data in certain columns, where we just automatically dropped them for the sake of time. The amount of observations have significantly dropped by nearly half from **2938** to **1649 observations**. Later on, we would do some more preprocessing by doing imputation once we have chosen features to work with for this analysis.

Response: **Life Expectancy**

Potential Features/Predictors: **Schooling**, **GDP**, **Alcohol**, **BMI, Percentage Expenditure, Income composition of resources, HIV/AIDS, Thinness 10-19 years (and 5-9 years), Adult Mortality**.

- There is a strong positive correlation between **Schooling** and **Life Expectancy** of **0.73**. This could be due to education being well established and widespread in certain developed/wealthier countries. This means countries have less corruption, more developed infrastructure, access to healthcare, etc.
- Moderate correlation between **GDP** and **Life Expectancy** of **0.44**, which are likely due to the same reasons as the first bullet point.
- Within the correlation plot there is a moderate positive correlation between **Alcohol** and **Life Expectancy** of **0.40**. We could assume this is due to the fact that certain wealthier countries are able to afford the consumption of alcohol is more common among the wealthy.
- Moderate positive correlation between **BMI** and **Life Expectancy** of **0.54**. This could be due to having healthier lifestyles in developed/wealthier countries compared to developing countries.
- There is a moderate positive correlation between **Percentage Expenditure** and **Life Expectancy** of **0.41**. This could possibly be due to countries having the monetary funds where they could spend on healthcare which reflects on government spending on healthcare, resulting in the correlation of **Life Expectancy**.
- There is a strong positive correlation between **Income composition of resources** and **Life Expectancy** of **0.72**. A human development index of how developed a certain country is in terms of income composition of resources.
- Strong negative correlation between **Adult Mortality** and **Life Expectancy** of **-0.70**. Which indicates that there is a higher mortality rate across from both sexes which would result in having a lower life expectancy.
- Since there is no correlation for this feature, **Status**, it **could serve as an indicator** of which countries are developed or not, which could possibly give an accurate precision of the Linear Regression model.

Question 2: *Is there a linear relationship between any of the predictors and life expectancy? Or is there some nonlinear relationship at play?*

The first step in determining whether any of the predictors have a linear relationship with life expectancy is to determine whether they have any predictive power at all. We can test for this by performing an F-test on a linear model fit to all 9 predictors. This corresponds to a null hypothesis that all of the predictors have slopes of 0, meaning that they are not significant for prediction. The alternative hypothesis is that at least one of the predictors is significant in predicting life expectancy.

```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
    percentage.expenditure + Income.composition.of.resources +
    HIV.AIDS + thinness.5.9.years + thinness.10.19.years, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-27.5813  -2.5470  -0.0207   2.6412  25.4956

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.980e+01  4.985e-01  99.895  < 2e-16 ***
Schooling                        1.078e+00  5.206e-02  20.702  < 2e-16 ***
GDP                              4.434e-05  1.709e-05   2.594 0.009537 **
Alcohol                         -1.009e-01  2.961e-02  -3.407 0.000668 ***
BMI                              5.476e-02  6.336e-03   8.642  < 2e-16 ***
percentage.expenditure           1.773e-04  1.111e-04   1.596 0.110670
Income.composition.of.resources  9.957e+00  7.498e-01  13.279  < 2e-16 ***
HIV.AIDS                        -6.609e-01  1.768e-02 -37.389  < 2e-16 ***
thinness.5.9.years              -4.548e-02  5.638e-02  -0.807 0.419871
thinness.10.19.years            -7.606e-02  5.757e-02  -1.321 0.186593
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.522 on 2301 degrees of freedom
Multiple R-squared:  0.7837, Adjusted R-squared:  0.7829
F-statistic: 926.5 on 9 and 2301 DF,  p-value: < 2.2e-16
```

Figure 2.1

Given that the F-test resulted in an exceptionally low p-value below 10^{-15}, we can clearly reject the null hypothesis and conclude that at least one of the predictors is useful in predicting life expectancy. As for which combination of predictors is most useful, that will be addressed in Question 3.

To test for linearity, we begin by plotting each predictor against life expectancy to look for visual evidence of linearity. While a less quantitative approach than significance tests, visualizations play a key role in the data exploration and modeling process by granting the analysts (us) a better grasp of the information at play.

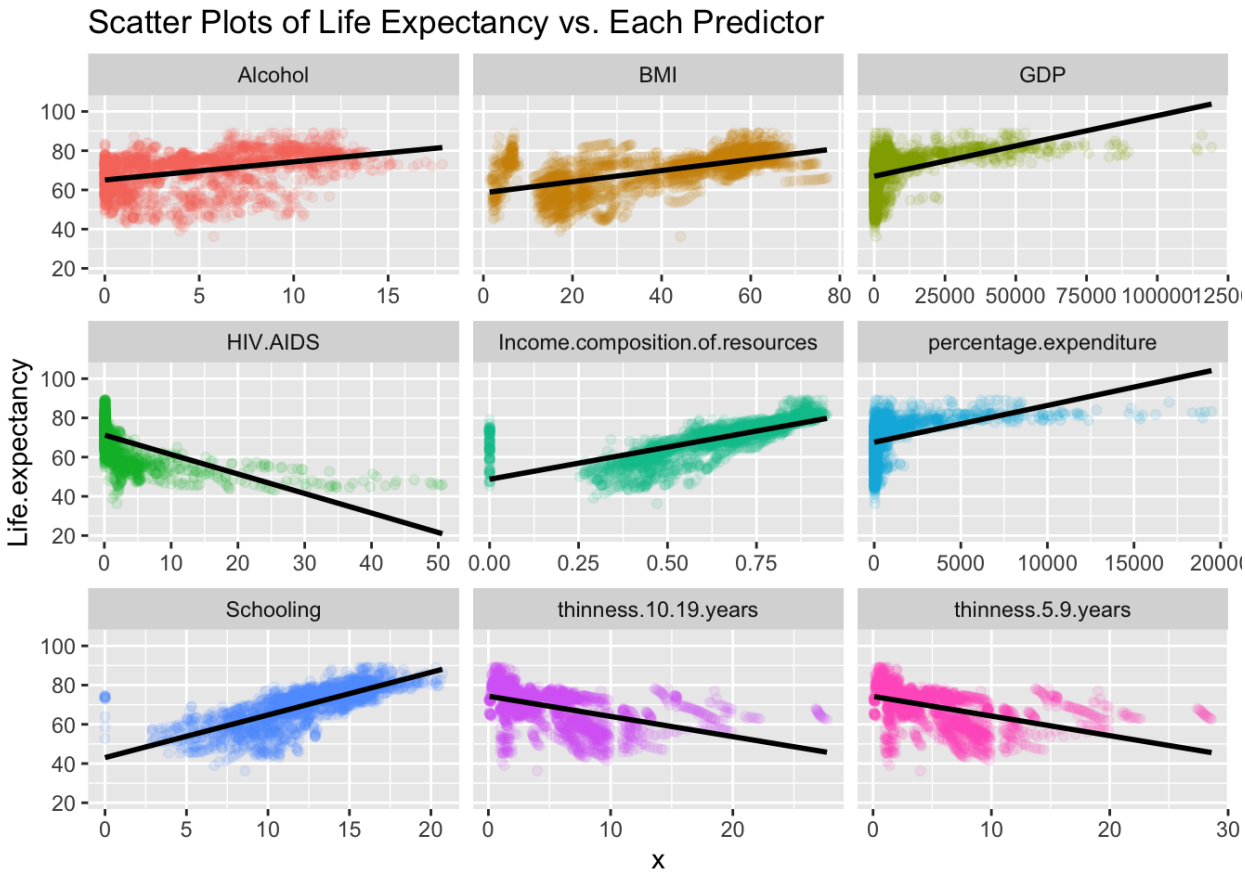

Figure 2.2

Based on scatterplots of each variable, we can clearly identify linear relationships between many of the predictors and life expectancy. In particular, "Alcohol," "BMI," and "Schooling" show clear linear patterns. The thinness variables also show linearity, although even visually they look highly collinear, something that will be addressed later. The remaining variables, such as "GDP," "HIV.AIDS," "Income.composition.of.resources," and "percentage.expenditure" show linearity at higher x values, but have wide ranges at lower x values. This may be evidence of non-constant error variance, and is certainly something we examine for later on in our analysis.

Indeed, the issue of linearity is not confined to this small question; it is addressed in more detail later in the report as we fit models and perform residual analysis. The goal of this question is simply to ask if applying linear models to this predictive problem at all seems reasonable.

Based on the exceptionally low p-value of our F-test and the obvious linear behaviors in the scatterplots above, we can confidently conclude that there is some linear relationship here, and possible non-linear relationships for points with low predictor values. So, we continue with our linear model fitting and analysis, returning to the question of linearity as appropriate.

Question 3: *Given these predictors, how accurately can one predict life expectancy with a linear model?*

For this analysis there are nine features within the data that we would like to choose to predict the **Life Expectancy**, which is our response variable. Now, we are going to perform a linear model and evaluate its accuracy to determine if it is a good regression model or not based on our assumptions of having a high magnitude correlation with the response variable, otherwise the Life Expectancy variable. After evaluating the linear model, we can observe that the accuracy of this model is **78.29%** from the Adjusted R-Squared from the predictor variable that we have selected. From the linear model, we can also observe that the linear has a few variables that are not that significant since they have a higher p-value which suggests that the changes in the predictor variables are not associated with the changes in the response variable. And the predictor variables that have no significance to the linear mode are **Percentage Expenditure**, **GDP**, **Thinness 1 to 19 years**, and **Thinness 5 to 9 years**. In the next part of the analysis, we might need to consider reducing the dimensionality to observe if the model improves in terms of predictive power.

```
Call:
lm(formula = Life.expectancy ~ (Alcohol + percentage.expenditure +
    BMI + GDP + Schooling + Income.composition.of.resources +
    HIV.AIDS + thinness.10.19.years + thinness.5.9.years), data = life)

Residuals:
     Min       1Q   Median       3Q      Max
-27.5813  -2.5470  -0.0207   2.6412  25.4956

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.980e+01  4.985e-01  99.895  < 2e-16 ***
Alcohol                         -1.009e-01  2.961e-02  -3.407 0.000668 ***
percentage.expenditure           1.773e-04  1.111e-04   1.596 0.110670
BMI                              5.476e-02  6.336e-03   8.642  < 2e-16 ***
GDP                              4.434e-05  1.709e-05   2.594 0.009537 **
Schooling                        1.078e+00  5.206e-02  20.702  < 2e-16 ***
Income.composition.of.resources  9.957e+00  7.498e-01  13.279  < 2e-16 ***
HIV.AIDS                        -6.609e-01  1.768e-02 -37.389  < 2e-16 ***
thinness.10.19.years            -7.606e-02  5.757e-02  -1.321 0.186593
thinness.5.9.years              -4.548e-02  5.638e-02  -0.807 0.419871
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.522 on 2301 degrees of freedom
Multiple R-squared:  0.7837,    Adjusted R-squared:  0.7829
F-statistic: 926.5 on 9 and 2301 DF,  p-value: < 2.2e-16
```

Figure 3.1

Now, after we have our linear model, we are going to check for its residuals plot (Figure 3.2) to see if there are not any non-linear relationships. Here, we see that the linearity seems to hold reasonably well, as the blue line is close to the dashed line, specifically near 0 in the x-axis (or horizontal axis). We can also note the homoscedasticity, as we move to the right on the x-axis, it is “equally” spread residuals around the horizontal line without any distinct patterns, which is a good indication that we do not have any non-linear relationships. Finally, we can also observe that there are some points within the residual plot where they could possibly be outliers. But overall, our linear model is good and reasonable. In addition, we have constructed a histogram of the residuals (Figure 3.3) from the linear model and as we can observe we can see that the variance is normally distributed.

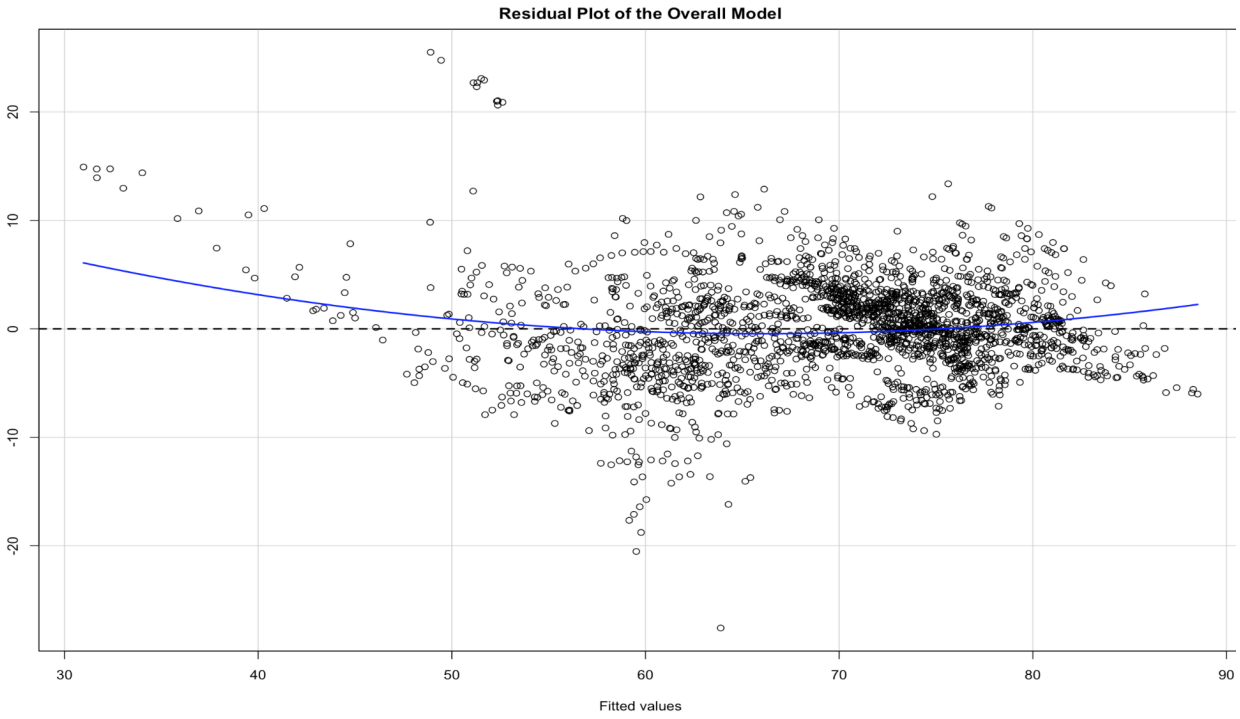

Figure 3.2

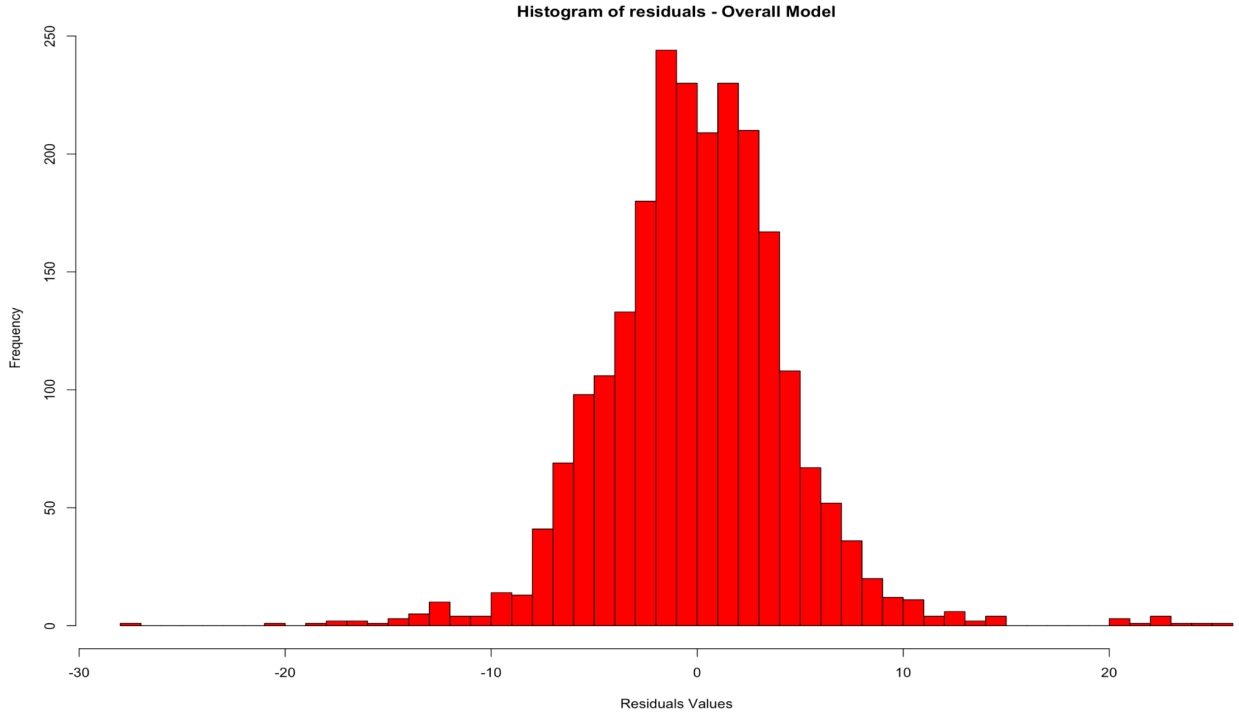

Figure 3.3

Question 4: *How has national life expectancy changed over time?*

One of the advantages of this dataset is that it has such a long temporal baseline. Namely, it contains annual data for most of the 155 countries from 2000 to 2015 (inclusive). We did not include year as a predictor in our model because we wanted to capture the impacts of only a country's functional factors, such as GDP, health expenditures, etc. While year may be correlated with life expectancy, there should be no inherent impact of the year 2007 on life expectancy, though year may correlate with predictors which have an impact. To test this theory, we can simply look at a plot of the residuals against year, using a model of all 7 prediction parameters (2 of the original 9 removed in prior model fitting and analysis).

```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
    Income.composition.of.resources + HIV.AIDS + thinness.10.19.years,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-27.5862  -2.5642  -0.0208   2.6258  25.4375

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.976e+01  4.959e-01 100.344  < 2e-16 ***
Schooling                       1.078e+00  5.208e-02  20.710  < 2e-16 ***
GDP                             6.913e-05  7.475e-06   9.247  < 2e-16 ***
Alcohol                        -9.308e-02  2.927e-02  -3.180  0.00149 **
BMI                             5.495e-02  6.281e-03   8.749  < 2e-16 ***
Income.composition.of.resources 9.872e+00  7.485e-01  13.189  < 2e-16 ***
HIV.AIDS                       -6.614e-01  1.767e-02 -37.426  < 2e-16 ***
thinness.10.19.years           -1.186e-01  2.616e-02  -4.533 6.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.523 on 2303 degrees of freedom
Multiple R-squared:  0.7834,    Adjusted R-squared:  0.7828
F-statistic:  1190 on 7 and 2303 DF,  p-value: < 2.2e-16
```

Figure 4.1

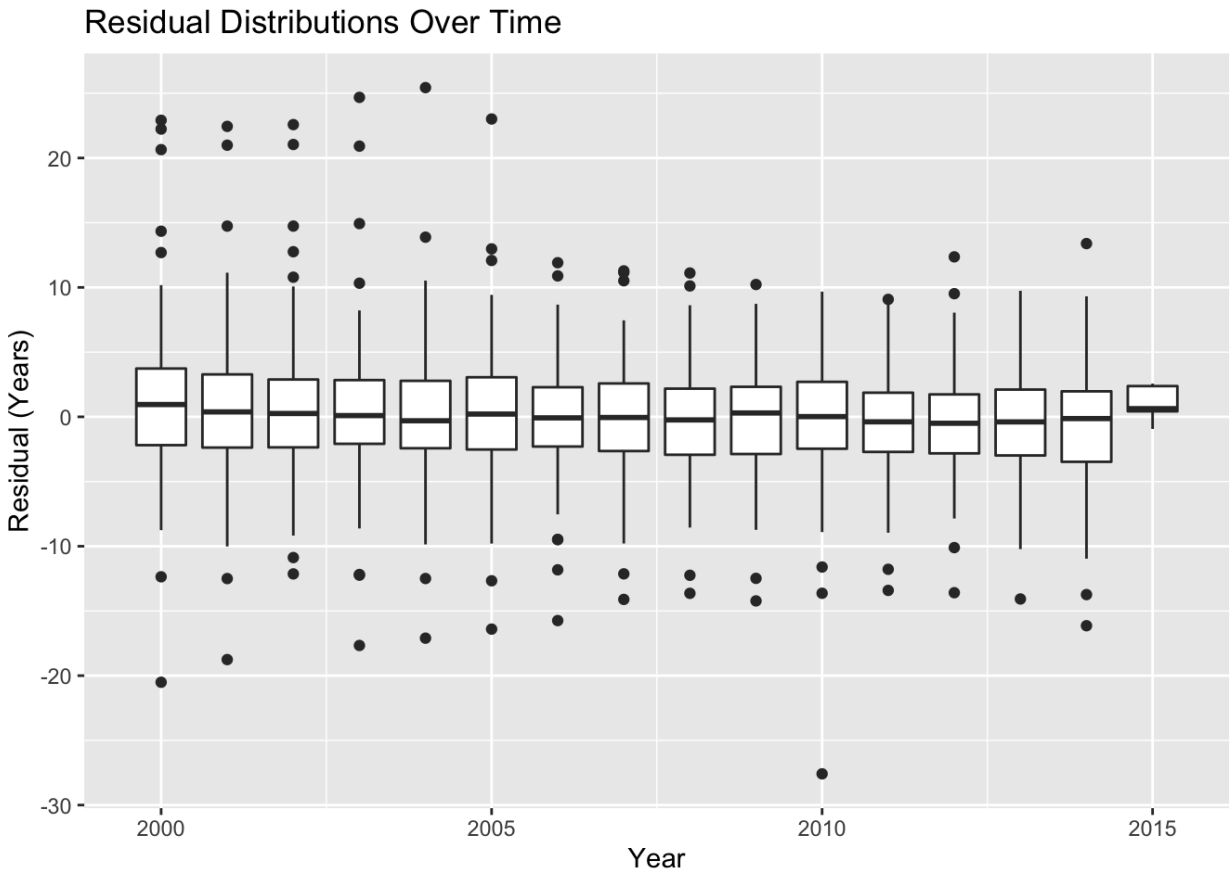

Figure 4.2

Even without normalizing, we can see that the structure of the residual distributions does not change in an obvious way with year. The quantiles are basically constant, though there are fewer extreme outliers after 2005. Lastly, we see a decidedly squashed distribution in 2015, which upon investigation, is the result of that group only having data for 5 countries.

With the independence of the year out of the way, we now explore the change in average national life expectancy over time. Below, we can see the average of all surveyed nations' life expectancies in each year. Note that this is not normalized for population, as we are interested in the average *nation's* life expectancy, not the average *person's* life expectancy.

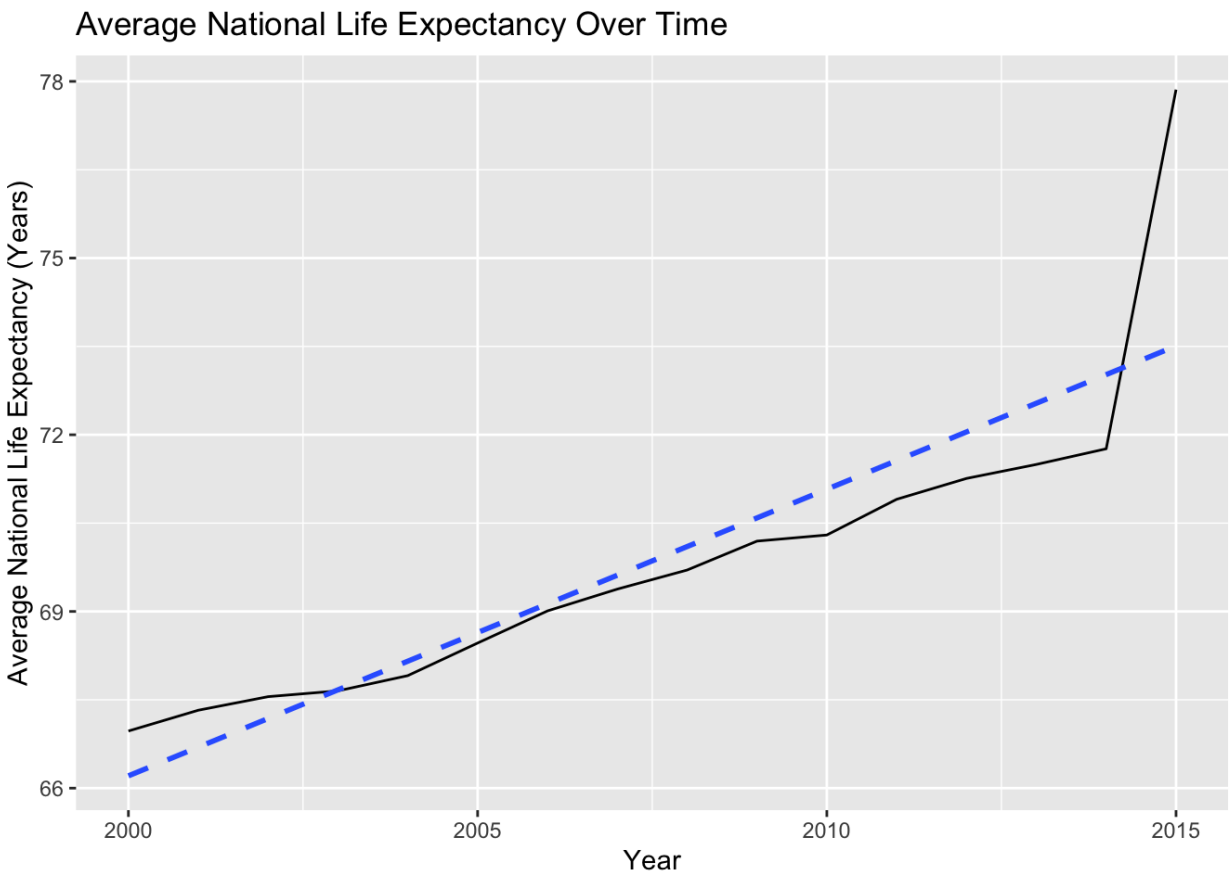

Figure 4.3

```
Call:
lm(formula = Life.expectancy ~ Year, data = year_averages)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2597 -0.6871 -0.2405  0.0802  4.3518

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -907.02839  145.05142  -6.253 2.12e-05 ***
Year           0.48662    0.07225   6.735 9.55e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.332 on 14 degrees of freedom
Multiple R-squared:  0.7641, Adjusted R-squared:  0.7473
F-statistic: 45.36 on 1 and 14 DF,  p-value: 9.551e-06
```

Figure 4.4

Though for the reasons above we would not attribute causation to the relationship between life expectancy and year, there is clearly a correlation present. In fact, by a simple model of life expectancy vs. year, we obtain a slope estimate of 0.48 year/year with a p-value of $< 10^{-5}$, making the relationship very significant. Again, as noted before, 2015 only has 5 observations, and so it is not unexpected for its average to be skewed from the rest of the data.

The next dimension upon which to investigate change over time is development status. We can see that average national life expectancy is increasing over time, but how are the benefits being distributed? Are developed or developing countries benefiting more? We segment those averages below, and then fit a linear model to year and a proxy variable which equals 0 for developing nations and 1 for developed nations.

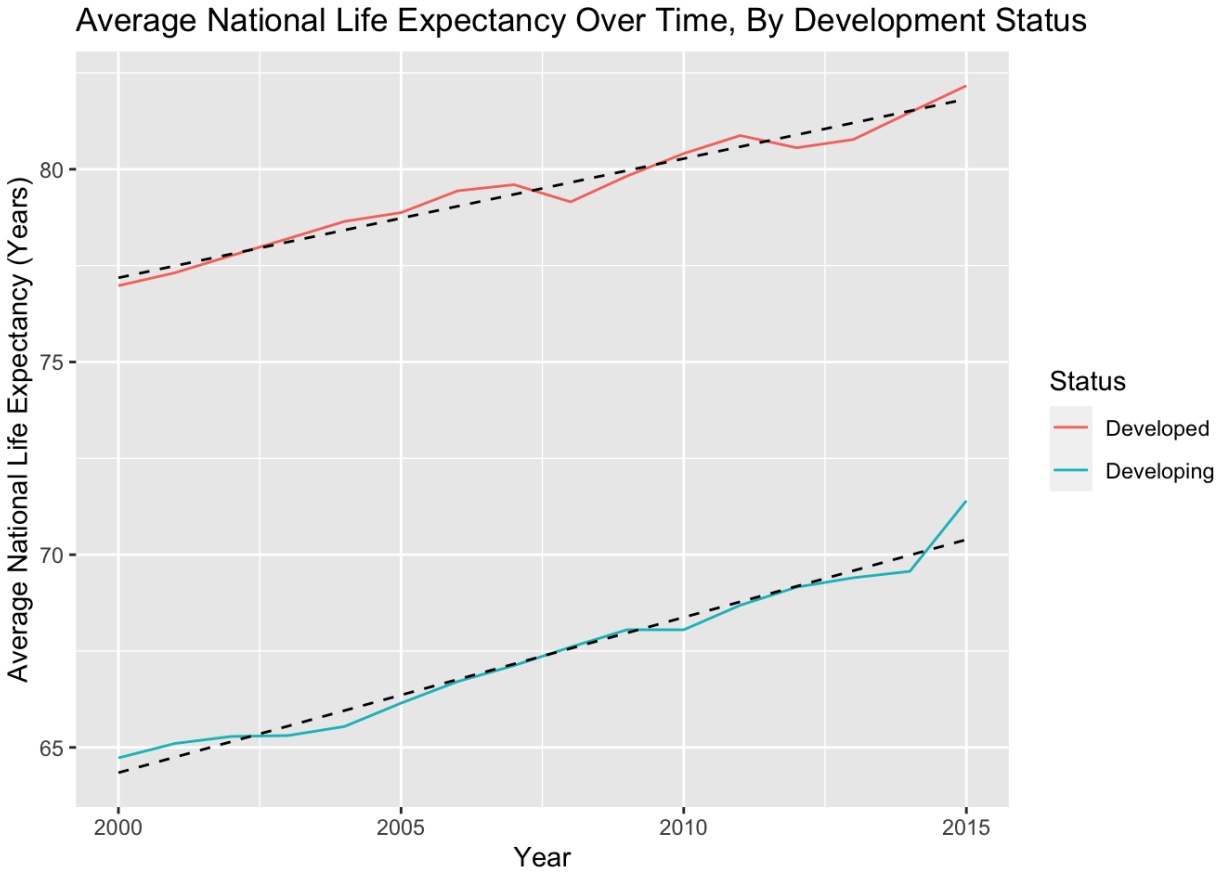

Figure 4.5

```
Call:
lm(formula = Life.expectancy ~ Status.proxy + Year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-31.883  -5.658   1.239   6.384  21.901

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -657.67632   80.43631  -8.176 4.77e-16 ***
Status.proxy   12.22552    0.44800  27.289  < 2e-16 ***
Year            0.36112    0.04008   9.011  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.328 on 2308 degrees of freedom
Multiple R-squared:  0.2642, Adjusted R-squared:  0.2636
F-statistic: 414.4 on 2 and 2308 DF,  p-value: < 2.2e-16
```

Figure 4.6

Clearly then, development status is highly correlated with life expectancy over time. On average, developed countries have life expectancies over 12 years higher than developing countries. Additionally, every year raises the global life expectancy by about a third of a year. Interestingly enough, these rates are basically equal for both developed and developing countries, which means proportionally, life expectancies are increasing faster for developing nations. These coefficients are highly significant and graphically the trends look relatively constant. While there is high correlation between life expectancy and development status, we still do not include it in our model because, like the year, it is too correlated with other predictors and reflects variables that are already included. For instance, GDP factors into classifying a country's development, so having development as a variable does not bring any significant new context to the problem.

For fun, if we extrapolate this linear pattern forwards, that means each person can expect to live, on average, 1.5 times as long as the life expectancy when they were born. For people born in developed countries around 2000, that's about 116 years old. While this is certainly too great of an extrapolation to have any real confidence in it, we thought it was an interesting fact and sanity check on our modeling. It's the kind of estimate which feels high, yet achievable in our lifetimes.

Question 5: *Which countries over- or under-perform in life expectancy relative to what our linear model would predict? What might account for this difference?*

While our dataset includes a nice selection of ~20 predictors (which we trimmed down to 9), one .csv file can hardly contain all possible information which is relevant to predicting a country's life expectancy. Policies, wars, geography, and more are all likely relevant features which are not totally reflected in the data. It seems reasonable to imagine that some nations may have combinations of these external variables which are particularly helpful or harmful to life expectancy. For that reason, we hypothesized that not all nations' residuals would follow the same normal distribution estimated by the linear regression because the residuals would include the impacts of unincorporated relevant variables.

We begin our residual analysis by visualizing the distribution of residuals for all observations. We leave these in years here for interpretability.

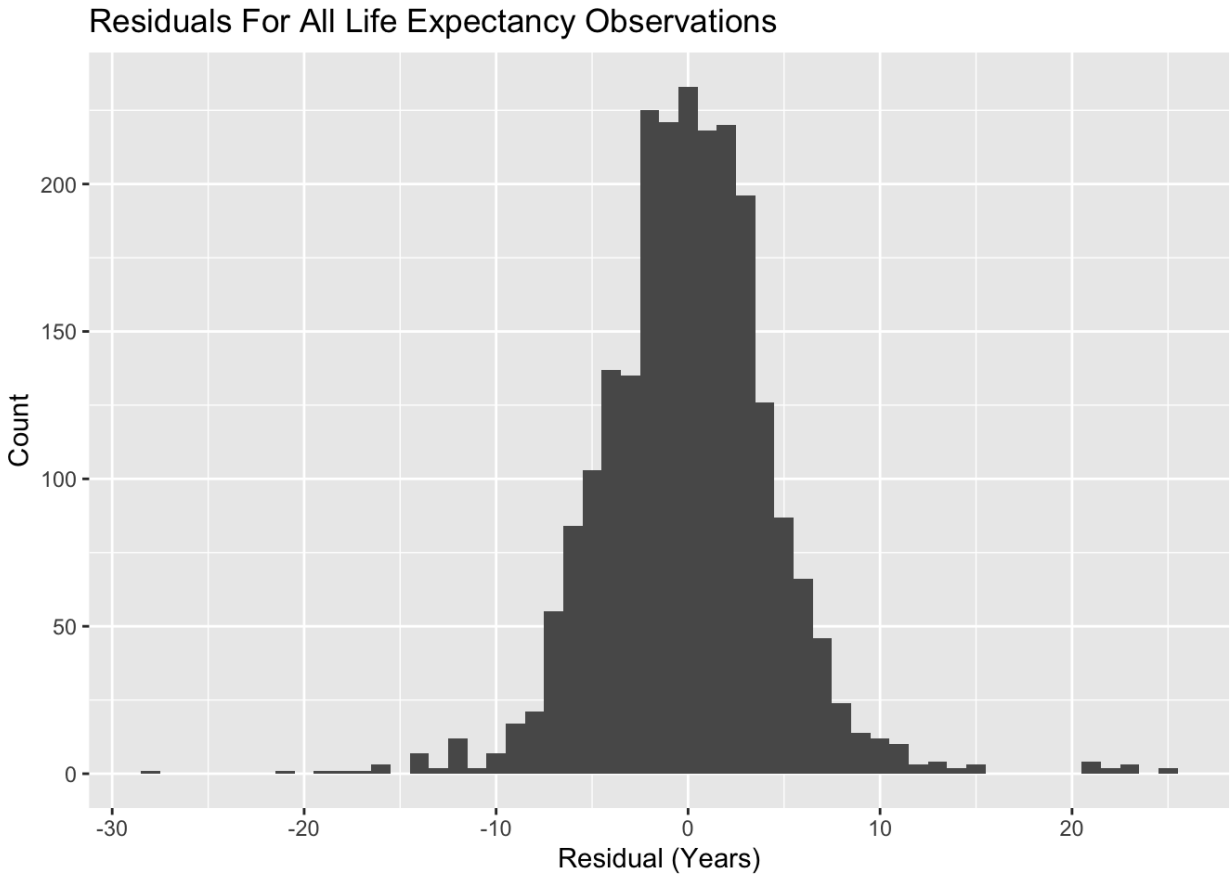

Figure 5.1

Upon plotting the distribution, we may be impressed to find that most estimates are within 3-5 years of the actual life expectancies, and almost all are within 10 years. Additionally,

the distribution is approximately normal, as it is unimodal and symmetric about 0. This gives us reassurance that our nearly normal residual condition is satisfied.

Now, we group the residuals by nation and take their average. These average residuals should give us a sense of whether a country is over- or under-performing in life expectancy relative to our predictions. This may be an indication of some wildly effective or ineffective policy or situational circumstance which accounts for a particular country's life expectancy. Such a metric is also useful as a measure of resource efficiency; which countries can achieve a better outcome than most other countries with similar resources, and what can account for these differences? We plot the national averages of the residuals below.

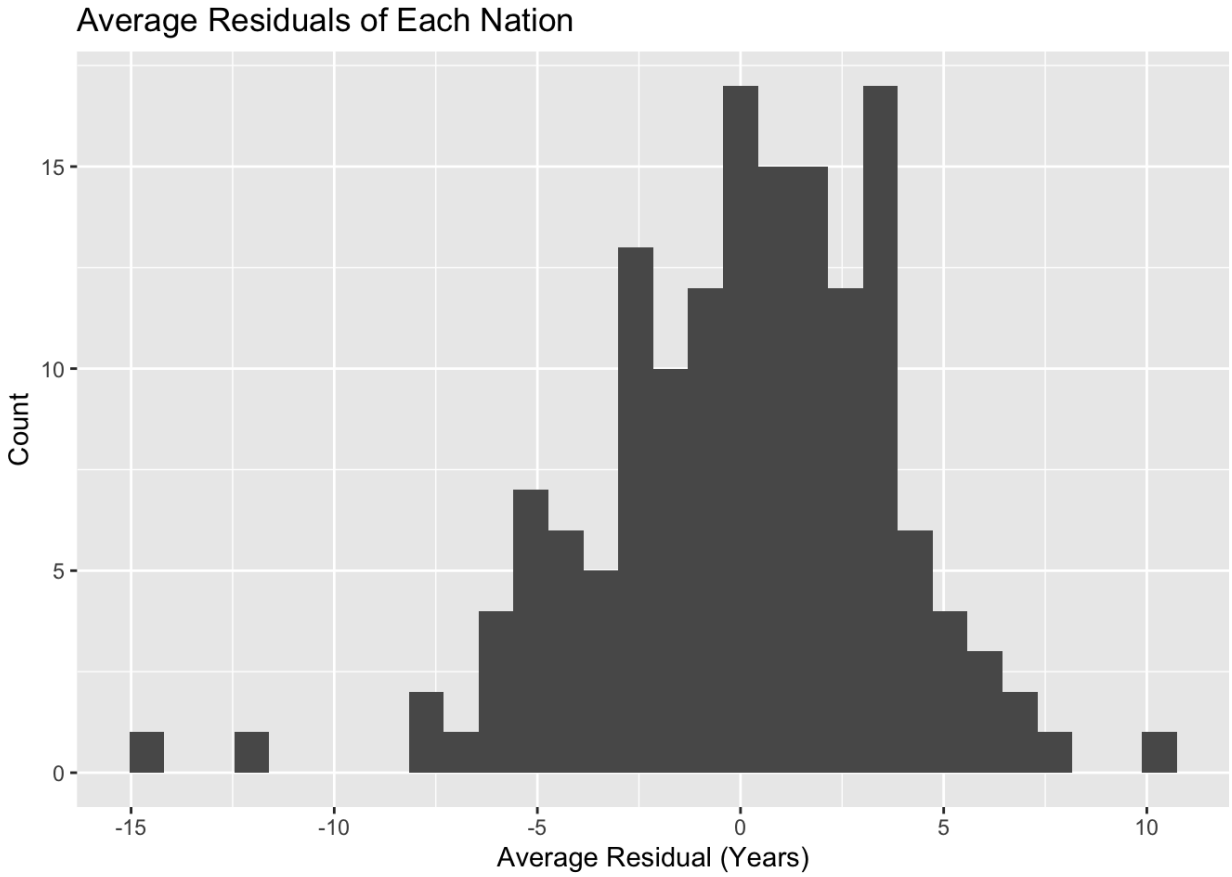

Figure 5.2

While this distribution is still approximately normal, one thing to note is that the spread has not seemed to change that much. Though the central limiting theorem would tell us that standard deviation should go down with sample size, as this is essentially a sample mean of error, we still see that most countries' averages are within 3-5 years of 0, not less. This is evidence that some countries may have intrinsic factors at play which alter the distribution of their errors.

To test this, we construct the hypotheses below, noting that the sum of all residuals for a country should follow the derived normal distribution given below. Using this, we can calculate

p-values for each country having a particular average error value. If it is significant, it would indicate that the country's actual error distribution is different from the overall error distribution, and the country has unincorporated intrinsic factors at play. Because we have almost 200 different countries in the dataset, we will use a restrictive significance threshold of 0.001, so that we should expect no significant outliers if our null hypothesis is true.

$$H_0 : \epsilon \sim N(0, \sigma^2) \text{ for all countries}$$

$$H_1 : \epsilon \nsim N(0, \sigma^2) \text{ for at least one country}$$

$$z_j = \sum_{i=1}^{n_j} \epsilon = \sum_{i=1}^{n_j} N(0, \sigma^2) \sim N(0, n_j\sigma^2) \text{ for each country } j$$

Figure 5.3

```
(signif_error <- av_error %>%
    filter(p_val < 0.001) %>%
    arrange(p_val))
```

```
## # A tibble: 43 x 5
##    Country                count total_error mean_error    p_val
##    <chr>                  <int>       <dbl>      <dbl>    <dbl>
##  1 Sierra Leone              15      -226.      -15.0  3.19e-38
##  2 Angola                    15      -182.      -12.1  1.29e-25
##  3 Antigua and Barbuda       15       148.        9.89 1.24e-17
##  4 Nigeria                   15      -115.       -7.64 3.10e-11
##  5 Swaziland                 15       113.        7.51 6.49e-11
##  6 Kazakhstan                15      -113.       -7.51 6.51e-11
##  7 Maldives                  15       105.        7.03 8.62e-10
##  8 Bosnia and Herzegovina    15       102.        6.82 2.67e- 9
##  9 Russian Federation        15      -101.       -6.75 3.82e- 9
## 10 Togo                      15       -95.3      -6.36 2.62e- 8
## # … with 33 more rows
```

Figure 5.4

We can see here that even with such a restrictive significance threshold of 0.001, we have 43 countries whose error distributions clearly differ from the overall error distribution. Now, we can plot the residuals over time of the top 5 over- and under-performing nations.

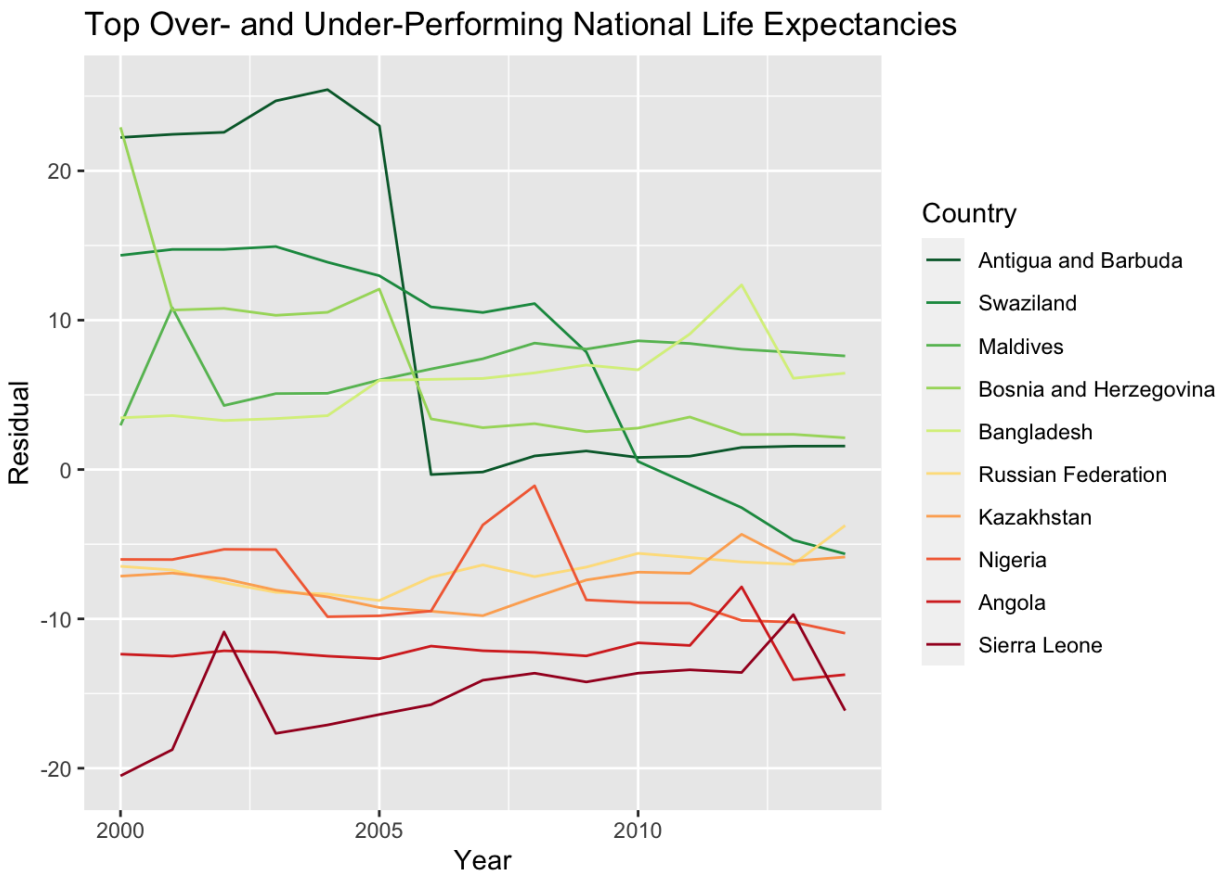

Figure 5.5

Here, the greenest countries represent the greatest over-performers, while the red-countries represent the greatest under-performers. All of them are statistically significant by our test. Visually, the under-performers, like Sierra Leone, are more consistent in their errors over time, dipping over a decade below what the model would predict. This would indicate these countries each have some significant circumstance which is a threat to public health. Meanwhile, the over-performers fluctuate more, such as Antigua and Barbuda, who goes from being the greatest over-performer (by about 20 years) to almost neutral in 2006. However, these countries still over-perform by almost a decade.

Question 6: *Are any of the predictors correlated with each other? If so, can the dimensionality of the dataset be reduced while retaining predictive power?*

We performed the same procedure as we did for Question 3, which was performing the residual analysis. This time we are going to remove one variable for the new model, which is the thinness.5.9.years variable since it did not have any significance to the overall model and we also had an extra variable that is similar to the thinness variable, so it should suffice. After making

certain adjustments to the model we can observe after making the removal of certain insignificant variables from the model we can conclude that it did not affect the accuracy of the linear model that much and still retains its predictive power of **78.28%**, even though we lost about **0.01%** of predictive power. From that, we could easily tell that there was some over-fitting going on with the model where it needed to be removed. As from the problem that we have previously answered, our linear model is reasonable and good.

```
Call:
lm(formula = Life.expectancy ~ (Alcohol + BMI + GDP + Schooling +
    Income.composition.of.resources + HIV.AIDS + thinness.10.19.years),
    data = life)

Residuals:
     Min       1Q   Median       3Q      Max
-27.5862  -2.5642  -0.0208   2.6258  25.4375

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.976e+01  4.959e-01 100.344  < 2e-16 ***
Alcohol                         -9.308e-02  2.927e-02  -3.180  0.00149 **
BMI                              5.495e-02  6.281e-03   8.749  < 2e-16 ***
GDP                              6.913e-05  7.475e-06   9.247  < 2e-16 ***
Schooling                        1.078e+00  5.208e-02  20.710  < 2e-16 ***
Income.composition.of.resources  9.872e+00  7.485e-01  13.189  < 2e-16 ***
HIV.AIDS                        -6.614e-01  1.767e-02 -37.426  < 2e-16 ***
thinness.10.19.years            -1.186e-01  2.616e-02  -4.533 6.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.523 on 2303 degrees of freedom
Multiple R-squared:  0.7834,    Adjusted R-squared:  0.7828
F-statistic:  1190 on 7 and 2303 DF,  p-value: < 2.2e-16
```

Figure 6.1

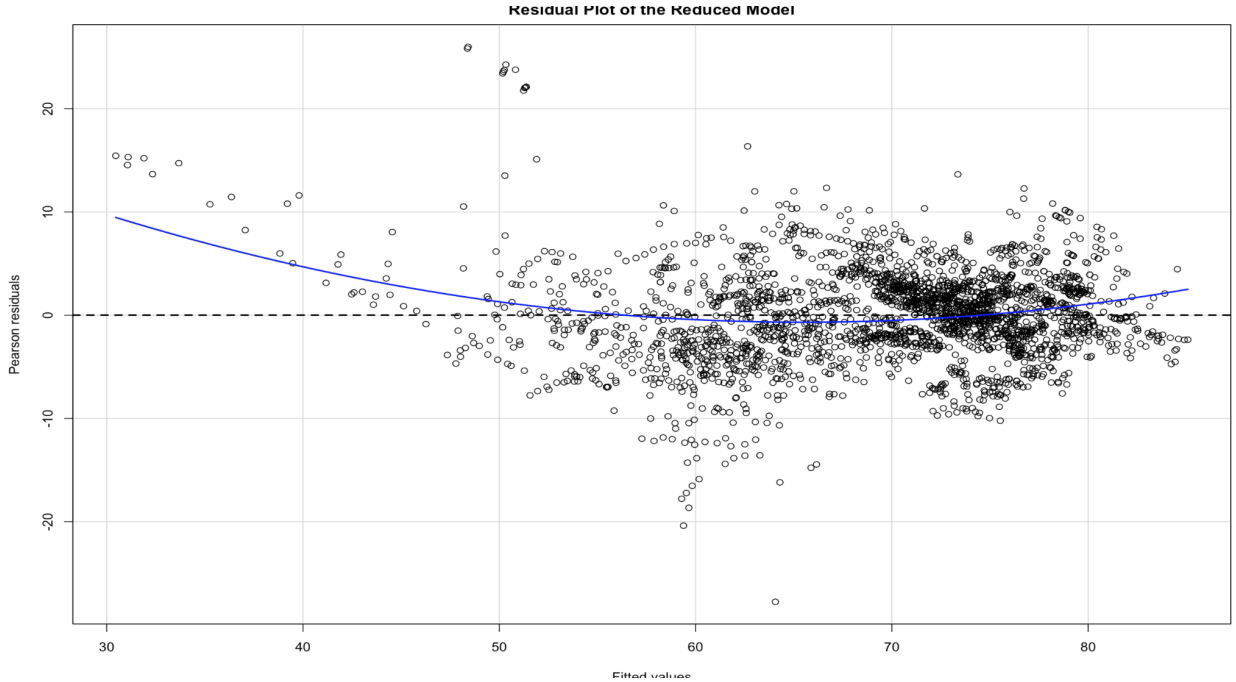

Figure 6.2

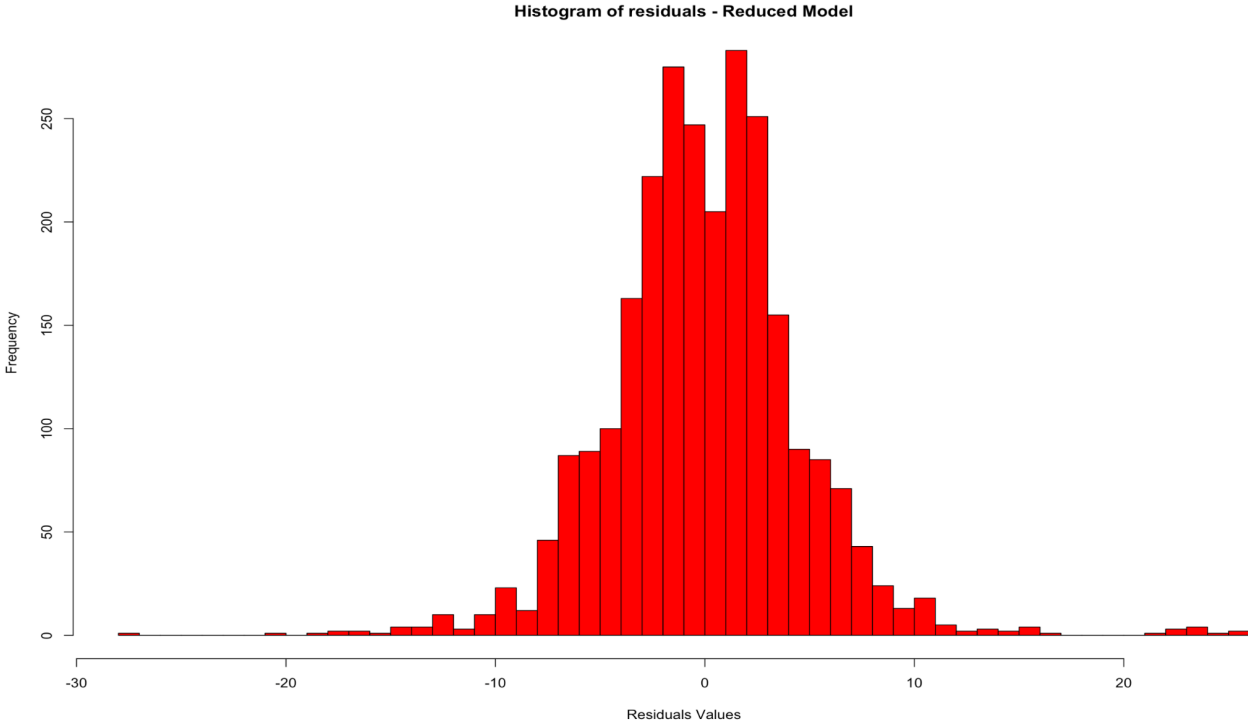

Figure 6.3

Question 7: *To improve the lifespan of a country with low life expectancy(<65), should they improve on their healthcare expenditure? Does it differ between developing and developed countries?*

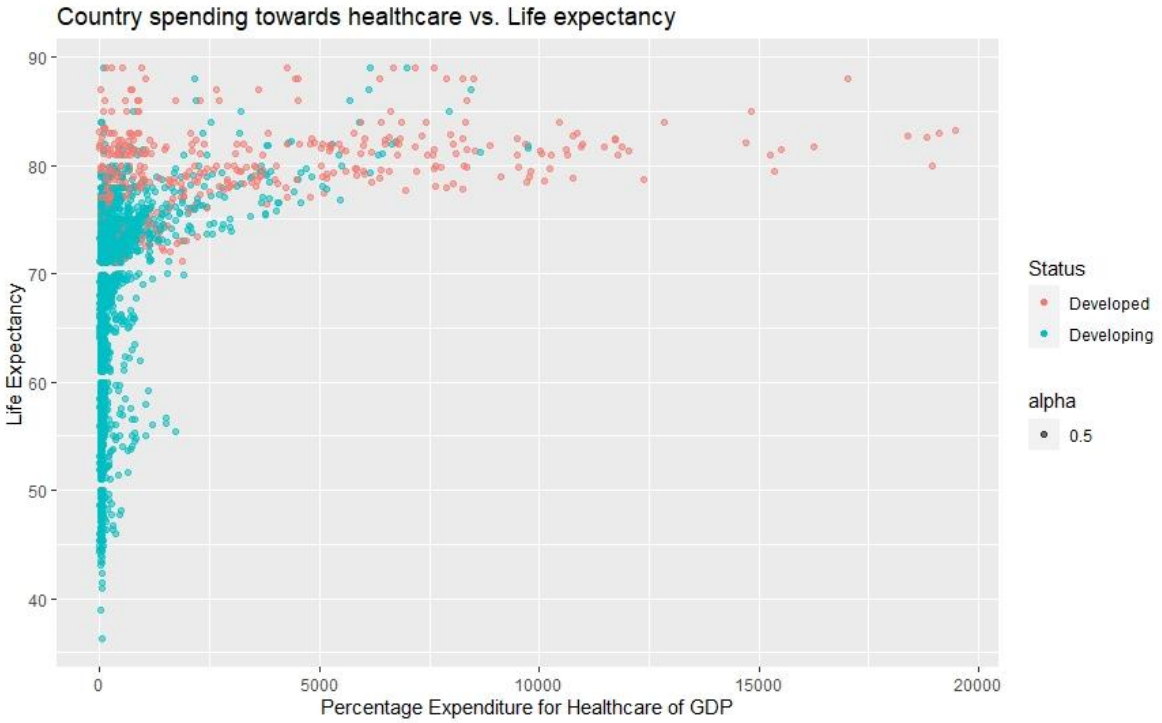

Figure 7.1

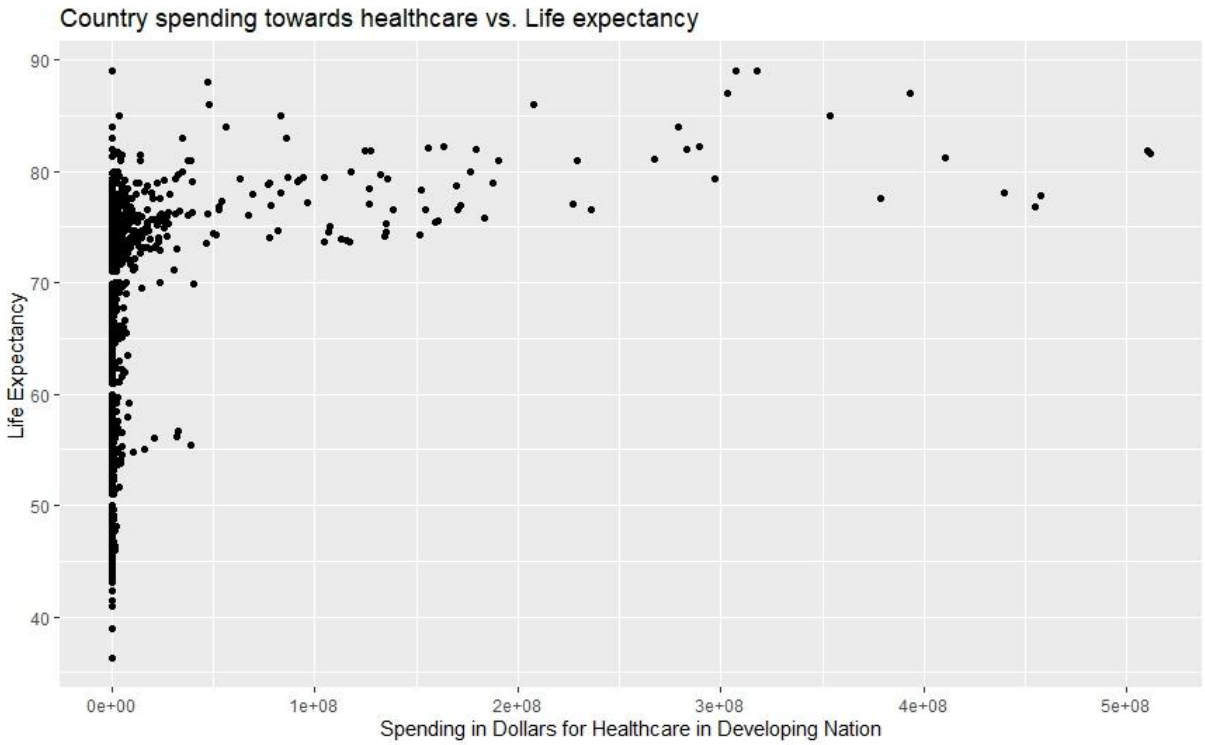

Figure 7.2

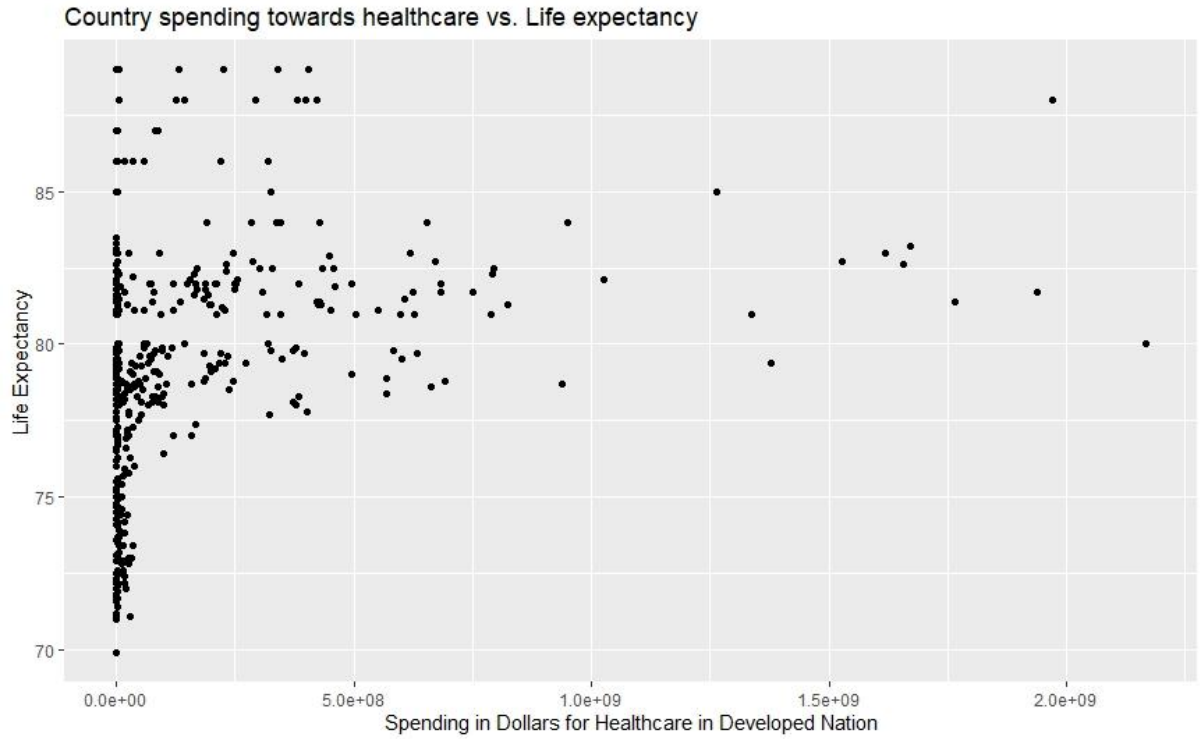

Figure 7.3

This question was one taken from the dataset and expanded upon. Before doing any analysis we felt that the answer would be obvious. That improving upon healthcare expenditure would increase life expectancy. We felt that this was the obvious choice because by improving

healthcare expenditure, the country would be investing more money into saving the lives of their people, so this would increase life expectancy. Our results were mixed and told us otherwise. In the graph above, you can see the comparison between the percentage expenditure of the GDP for healthcare and the life expectancy. It is very obvious that after a certain point of spending seen in **Figure 7.1**, the country's life expectancy is going to be high(>70). However, the question then arises is that if a country is spending a low amount, why do they still have a high life expectancy. The explanation comes in two forms: one that the country's GDP is high so that even a smaller percentage would go a long way in helping out, and the second and more likely reason, is that there are always other factors that play into making a country's life expectancy higher. Further analysis shows that all the developed countries have high life expectancies and all the countries with low life expectancies are still developing. This means that improving on healthcare will most likely help improve a country's life expectancy. After a certain point in spending, it seems as if it is almost guaranteed to improve life expectancy. To answer the question posed, yes the country should improve on their healthcare expenditure to improve the lifespan of a country with low life expectancy. However, that is not the only thing they must improve upon, because a country may not be able to spend that much money towards healthcare. This is true especially since the price that seemingly guarantees a high life expectancy is exorbitant, as seen when comparing spending by developing/developed countries in **Figure 7.2** and **Figure 7.3**. In the case of a developing vs developed nation, developed nations don't need to improve on healthcare spending to get a high life expectancy because they already have it. A developing nation, however, will need to improve in healthcare amongst other factors to improve life expectancy. This is later proven by the model seen in **Figure 8.3**.

Question 8: *In a developing country, what should they do to increase life expectancy versus a developed country?*

```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
    Income.composition.of.resources + HIV.AIDS + thinness.10.19.years,
    data = life)

Residuals:
     Min       1Q   Median       3Q      Max
-27.5862  -2.5642  -0.0208   2.6258  25.4375

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.976e+01  4.959e-01 100.344  < 2e-16 ***
Schooling                        1.078e+00  5.208e-02  20.710  < 2e-16 ***
GDP                              6.913e-05  7.475e-06   9.247  < 2e-16 ***
Alcohol                         -9.308e-02  2.927e-02  -3.180  0.00149 **
BMI                              5.495e-02  6.281e-03   8.749  < 2e-16 ***
Income.composition.of.resources  9.872e+00  7.485e-01  13.189  < 2e-16 ***
HIV.AIDS                        -6.614e-01  1.767e-02 -37.426  < 2e-16 ***
thinness.10.19.years            -1.186e-01  2.616e-02  -4.533 6.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.523 on 2303 degrees of freedom
Multiple R-squared:  0.7834,    Adjusted R-squared:  0.7828
F-statistic:  1190 on 7 and 2303 DF,  p-value: < 2.2e-16
```

Figure 8.1

```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
    percentage.expenditure + Income.composition.of.resources +
    HIV.AIDS + thinness.5.9.years + thinness.10.19.years, data = developed)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8284 -1.6045 -0.4987  0.7789  9.2844

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.327e+01  3.349e+00  15.908  < 2e-16 ***
Schooling                       -4.088e-01  1.008e-01  -4.058 5.92e-05 ***
GDP                             -1.913e-05  1.554e-05  -1.231   0.2191
Alcohol                         -2.498e-01  4.716e-02  -5.298 1.91e-07 ***
BMI                             -1.170e-02  7.807e-03  -1.498   0.1348
percentage.expenditure           1.082e-04  8.939e-05   1.211   0.2266
Income.composition.of.resources  4.473e+01  4.367e+00  10.241  < 2e-16 ***
HIV.AIDS                                NA         NA      NA       NA
thinness.5.9.years               1.348e+00  1.139e+00   1.184   0.2372
thinness.10.19.years            -3.189e+00  1.234e+00  -2.584   0.0101 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.536 on 414 degrees of freedom
Multiple R-squared:  0.6094,    Adjusted R-squared:  0.6018
F-statistic: 80.74 on 8 and 414 DF,  p-value: < 2.2e-16
```

Figure 8.2

```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
    percentage.expenditure + Income.composition.of.resources +
    HIV.AIDS + thinness.5.9.years + thinness.10.19.years, data = developing)

Residuals:
     Min       1Q   Median       3Q      Max
-27.3764  -2.6856   0.1231   2.6903  26.0064

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.951e+01  5.388e-01  91.898  < 2e-16 ***
Schooling                        1.103e+00  5.777e-02  19.094  < 2e-16 ***
GDP                             -4.149e-05  2.562e-05  -1.619    0.106
Alcohol                         -1.667e-01  3.670e-02  -4.542 5.92e-06 ***
BMI                              7.683e-02  7.533e-03  10.199  < 2e-16 ***
percentage.expenditure           1.327e-03  2.399e-04   5.531 3.63e-08 ***
Income.composition.of.resources  7.797e+00  7.882e-01   9.892  < 2e-16 ***
HIV.AIDS                        -6.494e-01  1.821e-02 -35.667  < 2e-16 ***
thinness.5.9.years               6.692e-03  5.792e-02   0.116    0.908
thinness.10.19.years            -5.918e-02  5.895e-02  -1.004    0.316
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.624 on 1878 degrees of freedom
Multiple R-squared:  0.7473,    Adjusted R-squared:  0.7461
F-statistic: 617.1 on 9 and 1878 DF,  p-value: < 2.2e-16
```

Figure 8.3

To increase life expectancy, not every country can or will follow the same path. Certain factors that help improve a developed country may not hold true for a developing country because they don’t have the same effect. In a developed country, the significant factors that affect life expectancy are schooling, alcohol consumption, and income composition of resources. In a developing country, those factors are schooling, alcohol consumption, average BMI, percentage expenditure on healthcare, income composition of resources, and HIV/AIDS cases. These are the factors that a developed vs developing country should improve upon. As previously guessed, not all factors are significant for both a developing and developed country. It is necessary for example, to improve upon the average body mass index in a developing country than a developed country, maybe because the body mass index on average is lower. The factors that hold true for both developing and developed countries are improving on schooling, alcohol consumption, and the income composition of resources.The developing countries have added factors to improve upon because they are still growing countries.

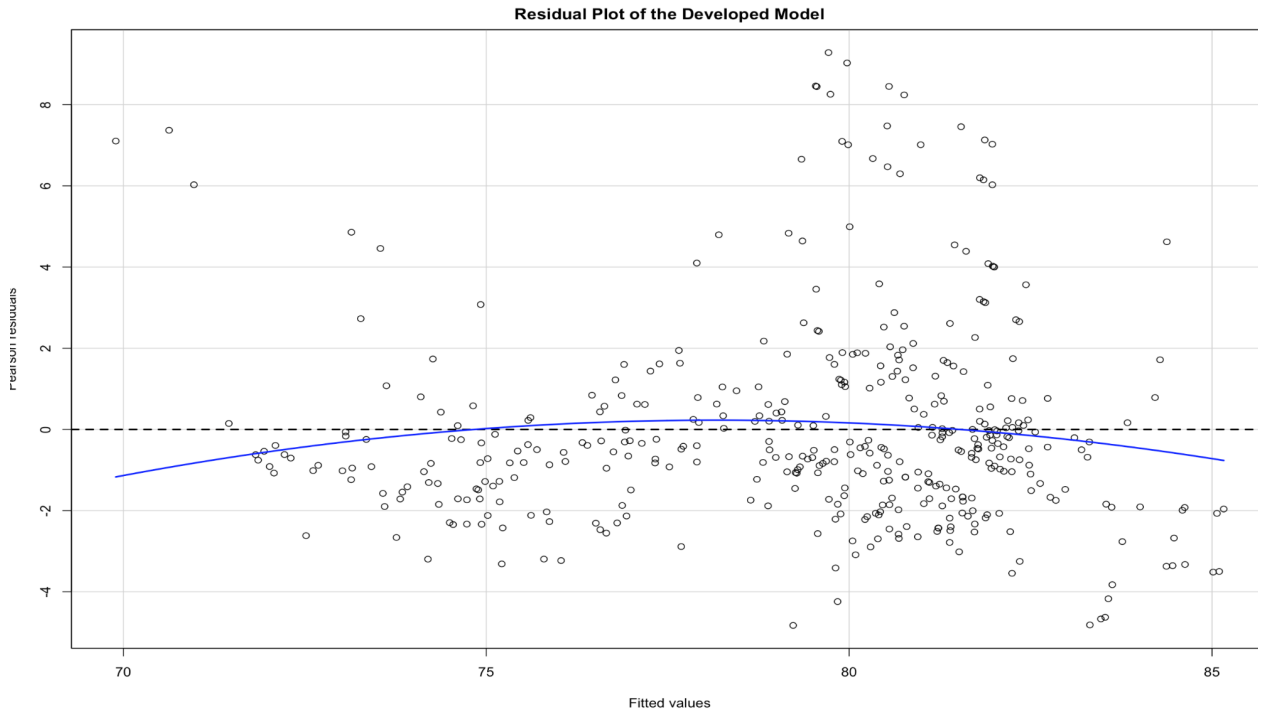

Figure 8.4

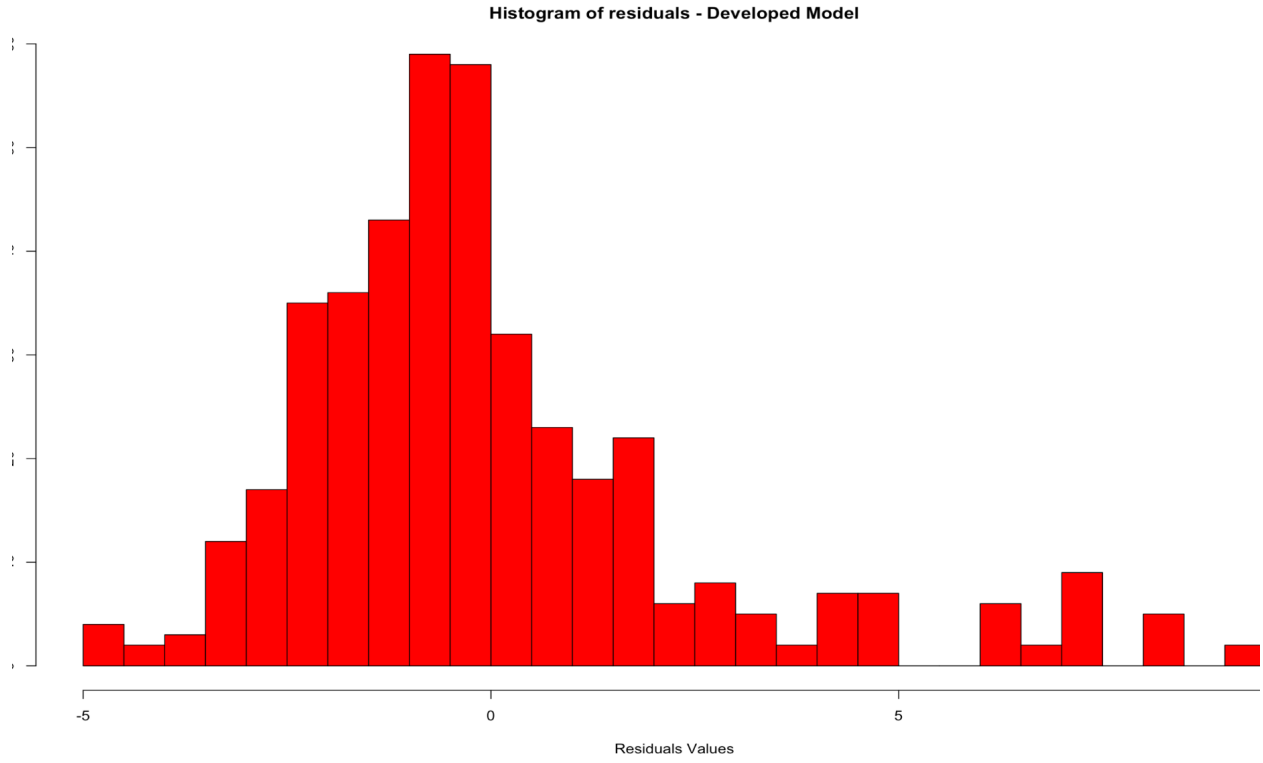

Figure 8.5

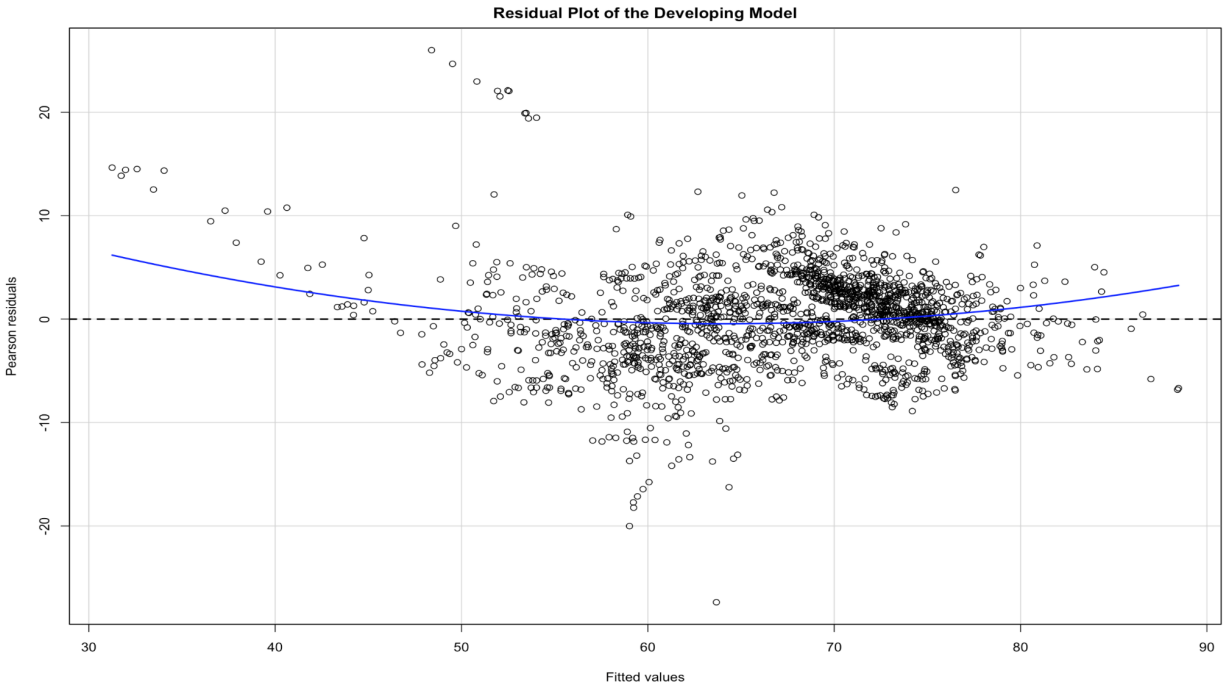
Residual Plot of the Developing Model
Pearson residuals
Fitted values
20
10
0
-10
-20
30
40
50
60
70
80
90

Figure 8.6

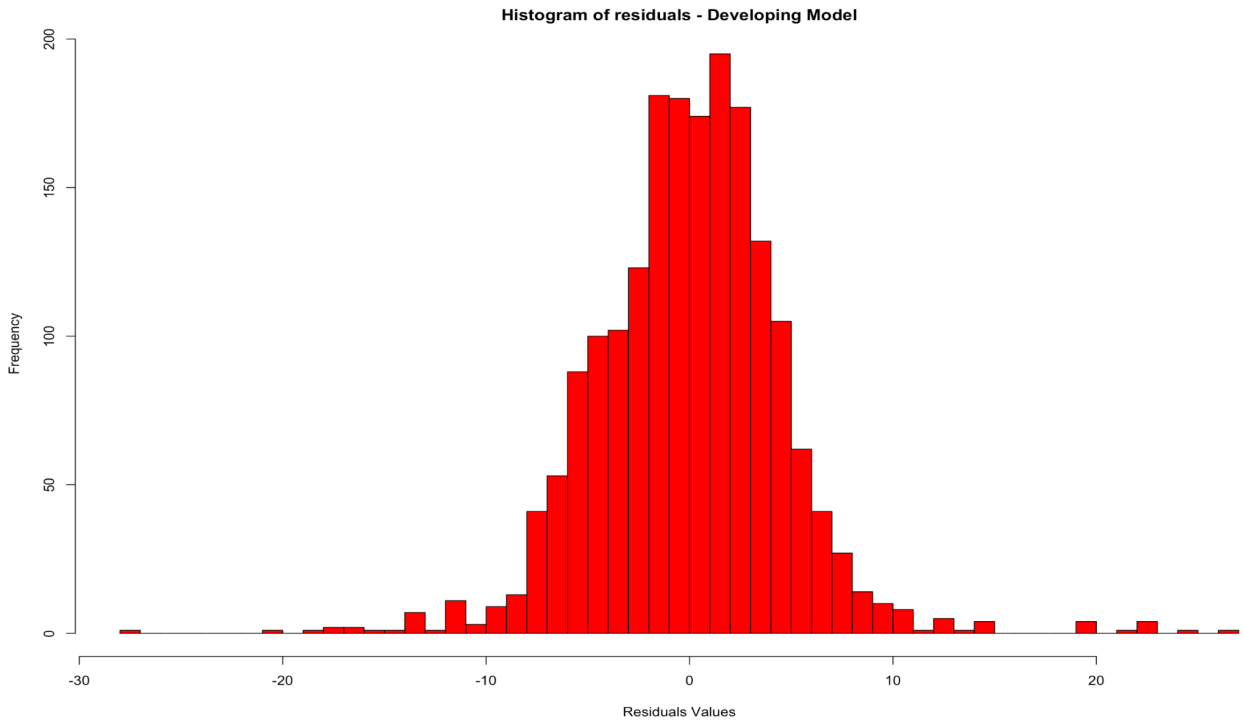

Figure 8.7

After that, we conducted a residual analysis for both of these models. For the developed model, we can see that there are many outliers, seen in **Figure 8.4**, which are due to several variables (such as the BMI, Percentage Expenditure, GDP, Thinness 10-19 Years, and Thinness 5-9 Years). Due to the disproportion in the number of residual points between the developing and developed model, we can see that they are slightly not distributed well across the x-axis (or the horizontal line) since most of the points are concentrated quite further into the fitted values axis. From this, we created the histogram model for the points to see if they are or not normally distributed. In **Figure 8.5**, we can see that the distribution is skewed to the right where there are some outliers happening. Even though the blue line is near the zero-line we can safely say that it does not show any signs of having any non-linear relationships for this model. But due to not having observations and outliers into this model, it is slightly a good model to work with.

Then we perform the same procedure for the developing model. In contrast, for **Figure 8.6**, there are many residual points within the graph giving us a good distribution across the x-axis near the 0 line, which therefore holds the linearity of the model to hold quite reasonably well. As for the histogram of residuals (**Figure 8.7**) for we can see that the variance is normally distributed as opposed to the developed model's histogram. In comparison to having an abundance of observations and few outliers we can say that this is a reasonably good model.

Conclusion:

To conclude, we answered our main goal which is to investigate the key factors that help improve a country's overall life expectancy. When we constructed our linear model from the 7 selected variables derived from the 10 original variables selected, we found that overall countries can improve on certain factors such as schooling, GDP, alcohol consumption, income composition of resources, and cases of HIV/AIDS. From the reduced model (refer to **Figure 8.1**) we have found great results with an R^2 value of 78.28% which is a decently good model since we assume that the model isn't being overfitted. As for residual analysis of the model (refer to **Figure 6.1** and **Figure 6.2**) we mentioned how there weren't any non-linear relationships on the residual plot and the histogram of residual showed to be normally distributed, thus it is a reasonably good model. Even though it has a good R^2 score there is a bias occurring within the model since we have removed all instances of missing values which doesn't give us the specifics of improving life expectancy. As previously mentioned in question 8, we have decided to separate the data into two models (developed and developing countries) since not all factors may apply for both country statuses. And based on our findings we found that there are a few commonalities in terms of key factors which were schooling, consumption of alcohol and income composition of resources which were common in the developed and developing countries. The key factors that both developed and developing countries should improve on is the body mass index, and health care expenditure. Essentially, depending on the country's status those are the key factors that a country should focus on in order to improve their overall life expectancy.

Reflection on Project & Future Expansion:

In this project we had parts work well for us, and some parts where we hit a roadblock. One of the main issues we had in this project had to deal with missing values and bias. In this dataset we had many missing values, and as a result we were put in a dilemma when deciding between trying to fill in these missing values or to delete them. There were pros and cons, but ultimately we decided that filling in the missing values would take too much time because we would need to scour the internet to find the missing information. Furthermore, we understood that by deleting the missing values we would introduce a bias into the dataset, but we felt that this would be easier and simpler considering the time constraint. Thus we decided to just delete the missing values. From there the project became easier. We decided to split up the questions to finish up in our own time, so that coordinating meetings to do the project together would not create problems. We found it easier to do the project this way and at the end of the week, we held meetings and reported our findings, so that everyone understood what was going on. After the cleaning of the data (removing unnecessary columns and missing values), we found that we were able to work smoothly and efficiently to finish our project in a timely manner.

For future expansion, we have many possibilities to explore. First and foremost would be to figure out the missing values and where we could get that information from. By doing so we could get a more complete result because a lot of the countries with missing values were developing countries where they were in the very early stages of development. Some of the information might be available in different sources. In this case, we would have to dig a little for the information but it would prove to be helpful in the long run. By completing the dataset and filling in any missing values, we can provide a more accurate model and conclusion. Another aspect that we can consider is looking at additional information, such as how much a government

spends in a year for the public. Through this kind of information we would see if there were other factors that were not included in the dataset that were crucial to a high life expectancy. We could also add in more recent information that would give some more insight as to how the life expectancy has changed in the last five years, or if it has changed at all. We could investigate if the life expectancy becomes stagnant after a certain stage.

Appendix A - Member Roles:

Prachi Patel- Goal, Question 1, 2, 7, & 8 (Developed vs. Developing), Reflection/Future Expansion
Evan Meade- The Dataset, Data Cleaning, Question 1, 2, 4, & 5
Alejandro De La Cruz- Question 1, 2, 3, 6, & 8 (Added after comments in presentation: Residual Analysis), Conclusion

Appendix B - References:

Life Expectancy (WHO) | Kaggle :
https://www.kaggle.com/kumarajarshi/life-expectancy-who

Appendix C - R Code:

```
# Abraca-Data Code
# Team 11
# STAT 4355 - Spring 2021
# Prachi Patel, Evan Meade, Alejandro De La Cruz


# Imports
library(tidyverse)
library(corrplot)
library(janitor)
library(gganimate)
library(ggpubr)
library(ggthemes)
library(scales)
library(car)


#
```

```
# Cleaning data
#

# Reading in .csv
data <- read.csv("Life Expectancy Data.csv")

# Trimming to columns of interest (R^2 > 0.4, excluding adult
mortality)
predictors <- c("Schooling", "GDP", "Alcohol", "BMI",
"percentage.expenditure",
                "Income.composition.of.resources", "HIV.AIDS",
                "thinness.5.9.years", "thinness..1.19.years")
cols <- c("Country", "Year", "Status", "Life.expectancy", predictors)
data <- data[cols]

# Fixing original source's typo in naming the columns
data$thinness.10.19.years <- data$thinness..1.19.years
data <- data[-(ncol(data) - 1)]
predictors <- c("Schooling", "GDP", "Alcohol", "BMI",
"percentage.expenditure",
                "Income.composition.of.resources", "HIV.AIDS",
                "thinness.5.9.years", "thinness.10.19.years")
cols <- c("Country", "Year", "Status", "Life.expectancy", predictors)

# Removing rows with any NA values
data <- data[-which(rowSums(is.na(data)) > 0), ]

# Writing cleaned data to a separate .csv file
write.csv(data, "life_CLEAN.csv", row.names = FALSE)


# Reading in cleaned data
data <- read_csv("life_CLEAN.csv")
life <- read_csv("life_CLEAN.csv")
data



#
# Question 1 - Alejandro
#

corrplot(cor(life[ , -c(1:3)]),
         type="upper",
         method = "number",
```

```
         addCoefasPercent = FALSE)


#
# Question 2 - Evan
#

# Plotting each predictor against life expectancy
data %>% gather(Predictor, x, Schooling:thinness.10.19.years) %>%
  ggplot() +
  geom_point(aes(x, Life.expectancy, color = Predictor), alpha = 0.1)
+
  geom_smooth(aes(x, Life.expectancy, group = Predictor),
              color = "black", method = "lm", se = FALSE) +
  facet_wrap(~ Predictor, scales = "free_x") +
  theme(legend.position = "none") +
  labs(title = "Scatter Plots of Life Expectancy vs. Each Predictor")


#
# Question 3 - Alejandro
#

model <- lm(Life.expectancy ~
              (Alcohol + percentage.expenditure + BMI + GDP +
Schooling +Income.composition.of.resources + HIV.AIDS
               + thinness.10.19.years + thinness.5.9.years),
            data = life)

summary(model)
residualPlot(model, main = "Residual Plot of the Overall Model")
hist(model$residuals, breaks = 50, col = 'red', main = "Histogram of
residuals - Overall Model", xlab = "Residuals Values")


#
# Question 4 - Evan
#

# Calculating average life expectancies for each year
year_averages <- aggregate(Life.expectancy ~ Year, data = data, FUN =
mean)

# Plotting average national life expectancy over time
```

```
ggplot(data = year_averages, mapping = aes(x = Year, y =
Life.expectancy)) +
  geom_line() +
  labs(title = "Average National Life Expectancy Over Time",
       x = "Year",
       y = "Average National Life Expectancy (Years)") +
  stat_smooth(method = "lm", se = FALSE, linetype = "dashed")

# Summarizing fit to year
summary(lm(Life.expectancy ~ Year, data = year_averages))

# Average life expectancies per year, with term for development status
status_averages <- aggregate(Life.expectancy ~ Year + Status, data =
data, FUN = mean)

# Plotting developed vs. developing average life expectancies over
time
ggplot(data = status_averages, mapping = aes(x = Year, y =
Life.expectancy)) +
  geom_line(mapping = aes(color = Status)) +
  labs(title = "Average National Life Expectancy Over Time, By
Development Status",
       x = "Year",
       y = "Average National Life Expectancy (Years)") +
  stat_smooth(mapping = aes(group = Status), method = "lm", se =
FALSE,
              color = "black", size = 0.5, linetype = "dashed")

# Creating proxy variable for status
data$Status.proxy <- as.integer(factor(data$Status, levels =
c("Developing",

"Developed"),
                                       labels = c(0, 1))) - 1

# Summary of fit to status proxy variable and year
summary(lm(Life.expectancy ~ Status.proxy + Year, data = data))



#
# Question 5 - Evan
#

# Fitting full model to the 7 selected predictors
```

```
model <- lm(Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
              percentage.expenditure + Income.composition.of.resources
+
              HIV.AIDS + thinness.5.9.years + thinness.10.19.years,
data = data)
summary(model)

# Refining that model to significance
# Removing thinness.5.9.years due to high collinearity
vif(model)
model <- lm(Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
              percentage.expenditure + Income.composition.of.resources
+
              HIV.AIDS + thinness.10.19.years, data = data)
summary(model)

# Removing percentage expenditure for low significance
vif(model)
model <- lm(Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
              Income.composition.of.resources +
              HIV.AIDS + thinness.10.19.years, data = data)
summary(model)

# Calculating predicted values and residuals for all entries
data$y.hat <- predict(model, data)
data$error <- data$Life.expectancy - data$y.hat

# Plotting histogram of residuals
ggplot(data = data) +
  geom_histogram(mapping = aes(x = error), binwidth = 1) +
  labs(title = "Residuals For All Life Expectancy Observations",
       x = "Residual (Years)",
       y = "Count")

# Plotting boxplots of residual distributions over time
ggplot(data = data) +
  geom_boxplot(mapping = aes(x = Year, group = Year, y = error)) +
  labs(title = "Residual Distributions Over Time",
       x = "Year",
       y = "Residual (Years)")

# Calculating p-value for each country's sample mean of the residual
distribution,
# assuming a null hypothesis of following N(0, sigma^2)
```

```
rse <- summary(model)$sigma
av_error <- data %>%
  group_by(Country) %>%
  summarize(count = n(), total_error = sum(error), mean_error =
total_error / count) %>%
  mutate(p_val = pnorm(abs(total_error), mean = 0, sd = sqrt(count) *
rse, lower.tail = FALSE))

# Printing countries with statistically significant deviations from a
mean residual
# distribution of 0
(signif_error <- av_error %>%
    filter(p_val < 0.001) %>%
    arrange(p_val))

# Plotting histogram of each nation's average residual
ggplot(data = av_error) +
  geom_histogram(mapping = aes(x = mean_error)) +
  labs(title = "Average Residuals of Each Nation",
       x = "Average Residual (Years)",
       y = "Count")

# Identifying top over- and under-performers by mean residual
outlier_order <- c(head(order(av_error$mean_error, decreasing = TRUE),
5),
                   tail(order(av_error$mean_error, decreasing = TRUE),
5))
av_error$over.rank <- order(av_error$mean_error, decreasing = TRUE)
outliers <- av_error$Country[outlier_order]
print(paste0(c("Top Over-perfomers: ", outliers[1:5])))
print(paste0(c("Top Under-perfomers: ", outliers[10:6])))
outlier_data <- data[which(data$Country %in% outliers), ]
outlier_data$Country <- factor(outlier_data$Country, levels =
outliers)

# Plotting top 5 over- and under-performers by mean error
ggplot(data = outlier_data) +
  geom_line(mapping = aes(x = Year, y = error, color = Country)) +
  scale_color_brewer(palette = "RdYlGn", direction = -1) +
  labs(title = "Top Over- and Under-Performing National Life
Expectancies",
       x = "Year",
       y = "Residual")
```

```
#
# Question 6 - Alejandro
#

model2 <- lm(Life.expectancy ~ (Alcohol + BMI + GDP + Schooling +
Income.composition.of.resources + HIV.AIDS +   thinness.10.19.years),
             data = life)

summary(model2)
residualPlot(model2, main = "Residual Plot of the Reduced Model")
hist(model2$residuals, breaks = 50, col = 'red', main = "Histogram of
residuals - Reduced Model", xlab = "Residuals Values")


#
# Question 7 - Prachi
#

ggplot(data = life, aes(x = percentage.expenditure, y =
Life.expectancy, color = Status, alpha = 0.5 )) +
  geom_point(stat = "identity") +
  labs(x = "Percentage Expenditure for Healthcare of GDP", y = "Life
Expectancy",
       title = "Country spending towards healthcare vs. Life
expectancy")


#
# Question 8 - Prachi (and Alejandro)
#

developed <- data.frame(life[which(life['Status'] == "Developed"),])
developing <- data.frame(life[which(life['Status'] != "Developed"),])

developed['Health'] <- developed['GDP'] *
developed['percentage.expenditure']
developing['Health'] <- developing['GDP'] *
developing['percentage.expenditure']


ggplot(data = developed, aes(x = Country,y = Health, color = Year)) +
geom_point()
```

```
ggplot(data = developed, aes(x = Country,y = Life.expectancy, color =
Year)) + geom_boxplot()

ggplot(data = developing, aes(x = Country,y = Life.expectancy, color =
Year)) + geom_boxplot()


ggplot(data = developing, aes(x = Health,y = Life.expectancy)) +
geom_point()+
  labs(x = "Spending in Dollars for Healthcare in Developing Nation",
y = "Life Expectancy",
       title = "Country spending towards healthcare vs. Life
expectancy")
ggplot(data = developed, aes(x = Health,y = Life.expectancy)) +
geom_point()+
  labs(x = "Spending in Dollars for Healthcare in Developed Nation", y
= "Life Expectancy",
       title = "Country spending towards healthcare vs. Life
expectancy")


ggplot(data = life, aes(x = Status,y = Life.expectancy, color = Year))
+ geom_boxplot()



fit <-
lm(Life.expectancy~Schooling+GDP+Alcohol+BMI+Income.composition.of.res
ources+
            HIV.AIDS+thinness.10.19.years, data = life)
summary(fit)

fit0 <-
lm(Life.expectancy~Schooling+GDP+Alcohol+BMI+percentage.expenditure+

Income.composition.of.resources+HIV.AIDS+thinness.5.9.years+thinness.1
0.19.years, data = developed)

summary(fit0)

fit1 <-
lm(Life.expectancy~Schooling+GDP+Alcohol+BMI+percentage.expenditure+
```

```
Income.composition.of.resources+HIV.AIDS+thinness.5.9.years+thinness.1
0.19.years, data = developing)
summary(fit1)

# Alejandro
model_developed <- lm(Life.expectancy ~ (Alcohol + BMI + Schooling +
percentage.expenditure + GDP + Income.composition.of.resources  +
HIV.AIDS + thinness.10.19.years + thinness.5.9.years),
                      data = developed)
summary(model_developed)
residualPlot(model_developed, main = "Residual Plot of the Developed
Model")
hist(model_developed$residuals, breaks = 50, col = 'red', main =
"Histogram of residuals - Developed Model", xlab = "Residuals Values")

model_developing <- lm(Life.expectancy ~ (Alcohol + BMI + Schooling +
percentage.expenditure + GDP + Income.composition.of.resources +
HIV.AIDS +   thinness.10.19.years + thinness.5.9.years),
                       data = developing)
summary(model_developing)
residualPlot(model_developing, main = "Residual Plot of the Developing
Model")
hist(model_developing$residuals, breaks = 50, col = 'red', main =
"Histogram of residuals - Developing Model", xlab = "Residuals
Values")
```