Examining Life Expectancy and Developmental Status

Statistical analysis and visualization of the significant factors in predicting life expectancy and developmental status for countries around the world

Jaiden Atterbury and Tanner Huck

CSE 163

Kevin Lin

5 June 2023

Research Questions and Results Summary

In this project we decided to use some of our statistical and programming abilities to analyze a data set on life expectancy around the globe from 2000-2015. In particular we decided to answer the three following research questions. Before moving on, note that the variables described in these questions will be explained in detail in the "Variable explanation" section of this report.

1. What are significant predictors of life expectancy? Given that our data has many relevant covariates that can impact life expectancy such as diseases, geographic location, economic conditions, and so on, how many of these variables are significant predictors of the life expectancy of a nation in a given year? In other words, which combination of variables creates the most significant/best model for predicting the life expectancy of a given country?
    - **Answer:** As found from fitting a multiple linear regression model to our data set, the significant predictors of life expectancy at the 5% level of significance are: infant deaths, under five deaths, adult mortality, alcohol consumption, Hepatitis B immunization rates, body mass index scores, HIV incidents, GDP per capita, "thinness" of individuals aged 10 to 19 years old, and schooling. After finding the variance inflation factor between these variables to address the multicollinearity problem present between the significant predictors, the final variables that were used for predicting the life expectancy of a country were: adult mortality, alcohol consumption, HIV incidents, and GDP per capita, and thinness of individuals aged 10 to 19 years old. Overall, the latter 5 variables were extremely significant at the 5% level, unlike some of the other "significant" variables before multicollinearity was addressed. As a brief side note, the constant term in the regression model was significant in both of the models that were fit. Even though this constant term isn't a predictor of life expectancy, its inclusion in the model helps improve the overall accuracy.

2. What are significant predictors of developmental status? In the data set, the binary "status" variable takes on values of developed and developing to label each country in the data set. This label is created in order to signify a country's status as either a first or third world country. With that being said, what variables in the data set are significant predictors for classifying a given country as developed or developing? Just as in research question 1, which combination of variables creates the most significant/best model for classifying if a nation is developing or developed? Furthermore, in terms of accuracy, how does a logistic regression compare to a decision tree classifier when fit onto this data set?
    - **Answer:** As found from fitting a logistic regression model to our data set, the significant predictors of the developmental status variable at the 5% level of significance are: infant deaths, under five deaths, adult mortality, alcohol

consumption, HIV incidents, GDP per capita, “thinness” of individuals aged 10 to 19 years old, and lastly schooling. Secondly, after fitting a DecisionTreeClassifier from the sklearn package, we found that the logistic regression model is less accurate than the DecisionTreeClassifer in both training and testing accuracy, but only marginally. In particular, the logistic regression model has a training accuracy score of 97.4% and a testing accuracy score of 97.6%, while the DecisionTreeClassifier has a training accuracy score of 100% and a testing accuracy score of 98.1%. Thus, the DecisionTreeClassifier is only slightly better than the logistic regression model for this specific split of training and testing data. It is important to note that all models in this report would change depending on which data points made the training set versus the testing set and vice versa.

3. How does life expectancy vary across different areas of the world? Given that our data set contains data on life expectancy and other forms of health data for many years and for almost every country in the world, how does life expectancy and other variables in this data set change and differ between different geographical locations and change over time? In particular, where in the world are these aspects of life expectancy most and least prevalent? How is life expectancy different between countries, continents, and regions? With this knowledge, why do these trends occur, and what are some steps to solving these problems?
    - **Answer:** From our different interactive plots, we can see that life expectancy has some variation from location to location. First looking at the different regions of the world, we can see that the average life expectancies from 2000-2015 are all quite similar. Africa seems to be a bit lower with a life expectancy of about 57, but all the other regions are in the range of the upper sixties to lower seventies. Then looking at the life expectancy of continents over time, we can conclude that life expectancy is increasing regardless of continent at a similar rate. We can once again see that Africa has shorter life expectancies on average, but is quickly catching up to the rest of the continents. Finally on the country level, this is where we can see the largest differences in life expectancy. We can see that some countries in the Americas, Europe, even countries like Australia tend to have large life expectancies, whereas certain countries in Africa and Asia are lower on average. Overall, we can see that life expectancy is fairly similar around the world both in terms of average rates and rates over time. Countries that are close to each other tend to have similar life expectancies and furthermore, we have the ability to target these specific areas with lower life expectancy and may conduct further research into why they may have a lower life expectancy in general.

Motivation:
The topic of life expectancy has significant real world importance due to its implications for public health and overall social well-being. Investigating and gaining insight into factors that impact life expectancy, we can make informed decisions and help improve global health outcomes, which in turn will improve the quality of life for both individuals and populations across the globe. In identifying the significant predictors of life expectancy, we can determine which variables have the highest impact and attempt to mitigate and/or understand why they may lead to longer and/or shorter life expectancies in general. Predicting and classifying a country's development status may help us identify areas of the globe that need the most outreach and support from government and nonprofit agencies. By finding which variables are significant predictors in this classification task, the use of a classification model using these variables may assist in assessing the progress and challenges faced by certain underdeveloped nations. On the contrary, we can learn what is working in these already developed nations and assess the feasibility of implementing them in underdeveloped countries. Furthermore, by identifying the significant predictors of the developed label this may help us develop strategies to help support and aid countries with improving their life expectancies and becoming more technological and sustainable countries as a whole. Finally, understanding the geographical data and how life expectancies change over time can help us understand what factors have been shown to impact life expectancy over time and where they may occur most frequently. We can analyze which countries have successfully raised their life expectancies and what trends led to this increase.

Data Setting:
Data source: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

"Life Expectancy (WHO)" by Kumar Rajarshi seeks to analyze life expectancy considering demographic variables, income composition and mortality rates. Rajarshi explains that although similar data analysis has been completed in the past, some important factors like immunization and human development were not taken into consideration. This new data set compiled by Rajarshi includes a range of data from 2000 to 2015 for almost all the countries in the world. Additionally, important immunization data about certain diseases and viruses such as Hepatitis B, Polio and Diphtheria will also be considered. Hopefully, this data can help determine which factors contribute to lower life expectancies and see which geographic areas need to be given more attention in order to improve their life expectancies.

More specifically, the data set includes information on several variables such as country, year, various demographic and socioeconomic factors, disease prevalence, immunization coverage, and healthcare expenditure. In analyzing the data set, we can gain insights into the impact of factors such as economic, social, and healthcare indicators on the overall life expectancy of a population. Furthermore, we can utilize this data set to perform data analysis, statistical modeling, and machine learning tasks to gain a better understanding of the factors influencing

life expectancy and if a country is a developed nation or not. This research may potentially lead to targeting specific regions and help find interventions and policies aimed at improving public health and extending life expectancy globally.

From our datasheet, we can learn many key aspects of the data analysis and motivation behind the data collection. One complication of the data set is its analysis of cultural, social, and economic backgrounds. The main goal of the data set was to analyze different factors to see how they affect life expectancy. However, depending on how this data was collected, it may lead to some biased or incomplete conclusions. A specific complication of the data set is its lack of political and policy factor data. An aspect that may greatly impact healthcare and life expectancy are political governance, healthcare funding, social welfare, etc. Without taking these factors into consideration, we may overlook key insights and limit the generalizability. On the other hand, one way the datasheet may deepen our understanding is through variable definitions. Our datasheet does a great job of explaining what each factor is, how it was collected, and how it affects life expectancy. For example for the Polio variable, it is actually measuring the immunization rate, not the percentage of adults with Polio like one may expect, without the datasheet we may have misinterpreted this value.

Variable explanation:
There are a total of 21 variables in this data set, but using R software one additional variable was created, which will be described and highlighted below. Other variable names were also changed to make it easier to understand, again using R denoted below. We do assume readers have some knowledge of diseases such as Measles or HIV. If not, we have provided links next to terms which help define the terms from the World Health Organization themselves.

Variables (variable names bolded):

- **Country**: Country observed.
- **Region**: The region in the world where the country is located (e.g., Asia, Africa, Oceania).
- **Year**: Year observed.
- **Infant Deaths**: Infant deaths per 1000 population. (Infant defined as 0-1 years old).
- **Under Five Deaths**: Deaths of children under 5 years old per 1000 population.
- **Adult Mortality**: Deaths of adults per 1000 population.
- **Alcohol Consumption**: Alcohol consumption recorded in liters of pure alcohol per capita for those 15+ years old.
- **Hepatitis B**: Percent coverage of Hepatitis B immunization among 1-year-olds.
- **Measles**: Percent of first dose of Measles-containing vaccine immunization among 1-year-olds. Measles Definition
- **BMI**: Body Mass Index (BMI), a measure of nutritional status in adults, calculated as weight in kilograms divided by the square of height in meters ($\frac{weight}{height^2}$).

- **Polio**: Percent coverage of polio immunization among 1-year-olds. Polio Definition
- **Diphtheria**: Percent coverage of Diphtheria tetanus toxoid and pertussis immunization among 1-year-olds. Diphtheria Definition
- **Incidents_HIV**: Incidents of HIV per 1000 population for those aged 15-49. HIV Definition
- **GDP_per_capita**: GDP per capita of a country in USD, calculated by dividing the value of an economy's GDP by the number of inhabitants.
- **Population_mln**: Total population in millions.
- **Thinness_ten_nineteen_years**: Prevalence of thinness among adolescents aged 10-19 years, defined as those with a BMI < -2 standard deviations below the median.
- **Thinness_five_nine_years**: Prevalence of thinness among adolescents aged 5-9 years, defined as those with a BMI < -2 standard deviations below the median.
- **Schooling**: Average years that people aged 25+ spent in formal education.
- **Economy_status_Developed**: Indicator for whether the country is classified as developed (1) or not (0).
- **Economy_status_Developing**: Indicator for whether the country is classified as developing (1) or not (0).
- **Life_expectancy**: Average life expectancy for both genders in different years.
- **Develop_Status**: Newly added variable that denotes whether a country is classified as developed or developing based on the variables Economy_status_Developed and Economy_status_Developing. Developed and developing are broad classifications used to categorize countries based on their economic and social development. The UN has a list for analytical purposes that classifies countries as either developed or developing.

Method:

Since randomness is involved in this report, the first thing we must do is set the random seed for reproducibility, in our case we set this seed to 10 using np.random.seed(10).

Data Exploration and Preprocessing:

To start the data exploration and preprocessing, we will first load in the data set and examine the variables related to life expectancy. This will help us in the understanding of different variables, the types of data present in the data set, as well as identify missing values, outliers, and inconsistencies in the data. Next we want to clean the data and handle any of the found missing values through the process of imputation or complete removal. We must be careful when choosing to remove or fix any of the data set, because missing data may provide valuable insights or tell a story about the data. It may also be necessary to find any data that is incorrectly imputed in the data set and make decisions about whether or not to keep them.

Once the data cleaning process is complete, we can perform exploratory data analysis. Our goal is to gain insights into the distribution and relationships among certain variables of interest. This

involves examining patterns, trends, and different variations in the data. We may choose to make histograms, scatter plots, or other types of graphs to visually understand these relationships. Similarly, we can also look at different summary statistics to understand the central tendencies and variabilities of these same variables.

Specifically, we will first look at the summary statistics of the life expectancy variable. Then we will create a histogram of the life expectancy with a density plot overlaid on top. Third, a line plot that shows the average life expectancy over time. Next, a bar chart that will show the top five highest and lowest countries with respect to life expectancy. Also, five scatter plots with lines of best fit to examine the Infant Death rate, Under Five death rate, Adult Mortality, BMI, and GDP per capita variables vs life expectancy. Lastly, to get a better understanding of the linear relationship between life expectancy and the other continuous variables in the data set, we will make a correlation matrix. All of these will be used to guide our decision making for the models, and more importantly to get a better understanding of life expectancy as a whole.

Research Question 1: Predicting Life Expectancy:

Unlike the data pre-processing and exploration, the methodology for this research question changed a lot from the proposal. Before fitting any model to the data we must first analyze the distribution of the life expectancy variable to assess the normality of the dependent variable, which will allow us to critique if fitting this kind of model is appropriate. If the life expectancy variable is not normal we will attempt to make a transformation to the data to obtain thus sought after normality. The next step to answering this question is by fitting an initial multiple regression model using the statsmodels package on all of the continuous variables in the data set. Once this is done it will be important to drop all of the insignificant variables at the 5% level of significance and check the variance inflation factor (VIF) of each variable to address/assess the presence of multicollinearity. We will then drop all of the variables that have a VIF score of more than 5, however, it is important to note that if many similar variables are correlated to each other we will pick one of these variables since the chosen variable will probably have a low VIF score once the other variables are removed. Also, we will choose to keep the constant term of the model depending on if it is significant or not. After this we will split the data into a training and testing set (80/20 split) and re-fit the model to the data using only these newly found significant and uncorrelated variables using the training set. Once this model is obtained we will check the following assumptions: linearity, no autocorrelation between the model residuals, homoscedasticity of the residuals, normality of the residuals (centered around zero), and lastly no outliers. We will test these assumptions using a rainbow test, Durbin-Watson test, a het breuschpagen test, a histogram of the residuals with a density plot overlaid on top, and lastly a studentized residual plot. After the model assumptions are assessed and the model validity is put into question, the last step in answering this research question is to find the training and testing accuracy of the model itself. Lastly, we will display the model equation and interpret what these coefficients mean in the context of life expectancy.

Research Question 2: Classifying Development Status:

Just like the first research question, the second research question drastically changed from the project proposal. This happened mainly due to the length of the coding file itself, and due to the fact that testing logistic regression assumptions itself is much harder than for linear regression. Instead of checking the logistic regression assumptions we instead compared our logistic regression model to a sklearn DecisionTreeClassifier, since this is an already robust and well-tested model. Unlike linear regression, we can jump right into fitting the logistic regression model since we aren't dealing with any categorical variables. Again we will fit this model using the statsmodels package on all of the continuous variables in the data set. Once this is done we will drop all of the insignificant variables at the 5% significance level, including the constant term of the model. Once this is done we will split the data into a training and testing set (80/20 split) and re-fit the model to the data using only these newly found significant variables. The last step in the logistic regression portion of this research question is to find the training and testing accuracy of the model itself. After this we will create a DecisionTreeClassifier on the same training data from above, and similarly, we will find the training and testing accuracy of the model. Lastly, we will compare these accuracies and decide which one is the better model and why that might be so.

Research Question 3: Temporal and Geographical Analysis:

This question aims to explore the geographical trends of life expectancy. We will begin by grouping the data by year and calculate different statistics about life expectancy like the average. Then these life expectancy details will be graphed in interactive visualizations. We will create a line plot that measures the mean life expectancy per continent over time, a bar graph that displays the mean life expectancy per region, and a choropleth of the globe that displays the average life expectancy per country. Using these different visualizations, we will analyze how life expectancy changes for different locations and identify regions where life expectancy and disease prevalence are most and least prevalent. This will help us determine potential factors contributing to these trends, such as socioeconomic indicators, healthcare expenditure, or immunization coverage. We hope to find the reasons behind these trends and propose steps for addressing the identified problems.

Result Validity:

The validity of our results will be assessed by evaluating the model's performance using appropriate metrics and statistical tests. We will use metrics such as AIC, BIC, or adjusted R-squared to assess the quality of the models as well as compare models to each other. Then we can select models based on performance on these metrics. Finally, we can reassess and ensure that the chosen model is reliable and provides meaningful insights.

Results:

In this section we will dive into the main part of our project; describing the results of our analysis/code.

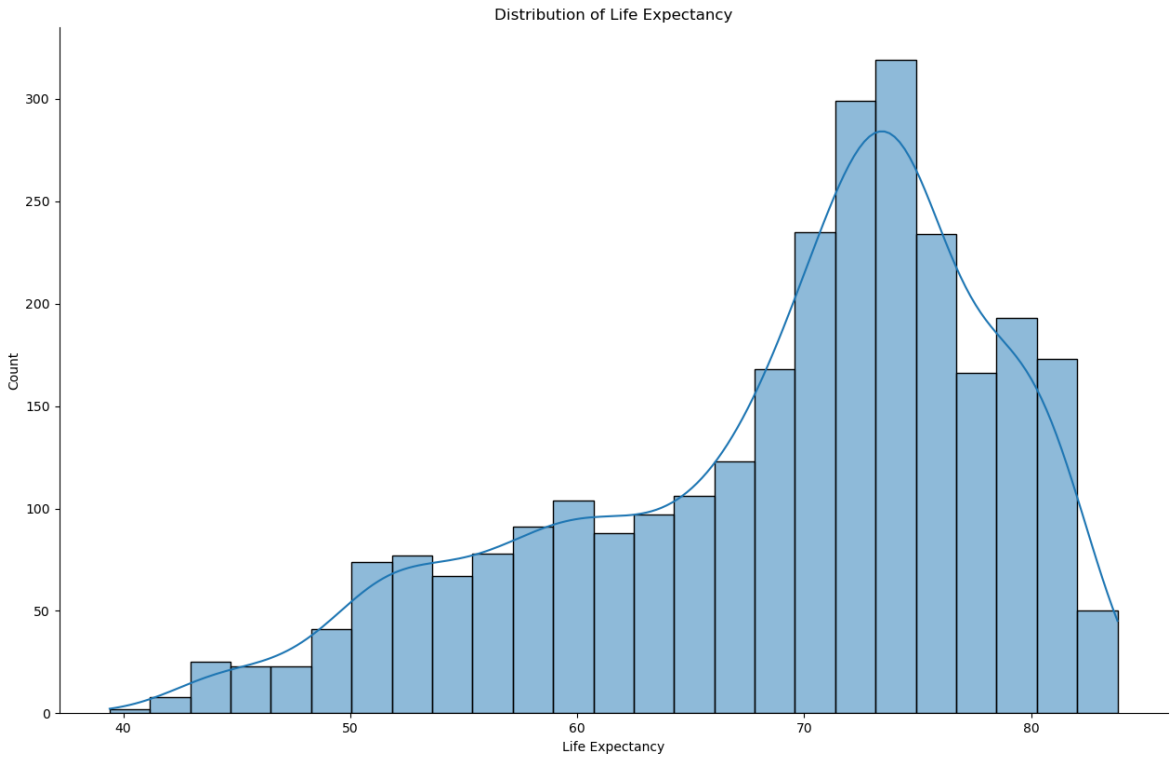
Data Exploration and Preprocessing:

a) Summary statistics of life expectancy

```
count    2864.000000
mean       68.856075
std         9.405608
min        39.400000
25%        62.700000
50%        71.400000
75%        75.400000
max        83.800000
Name: Life_expectancy, dtype: float64
```

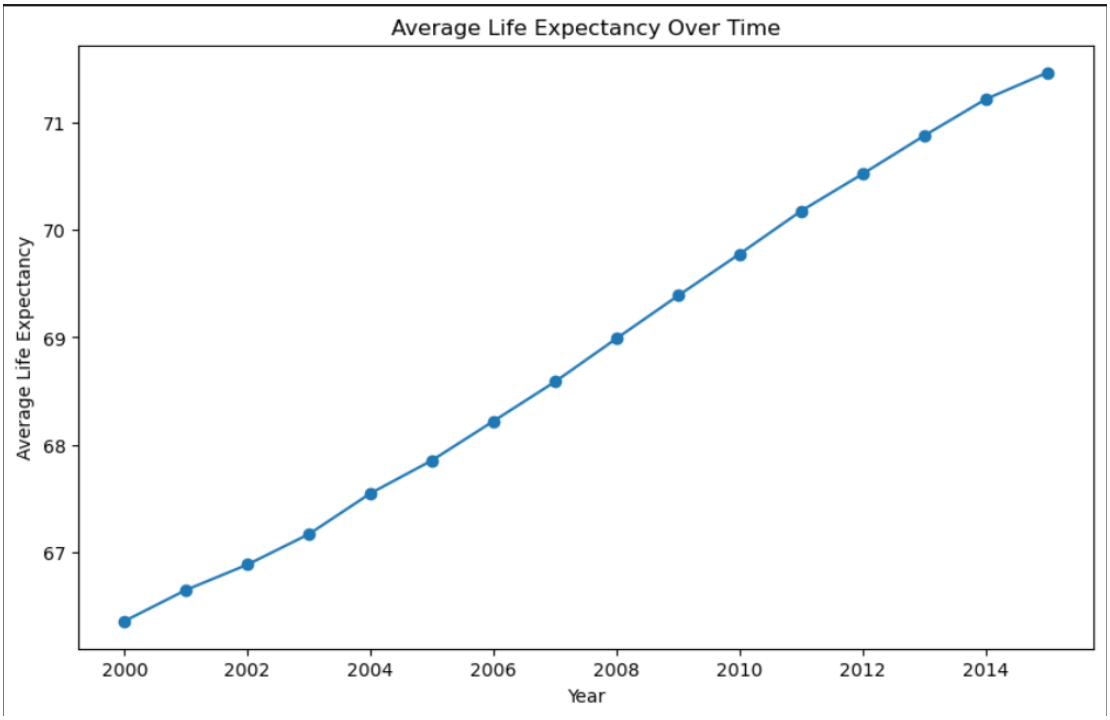
Using the summary statistics shown above, we can take a glimpse at the life expectancy variable and try to get a base understanding of its distribution. We can learn that we have 2864 observations of life expectancy values with the average being about 69 years. This tells us that the average life expectancy across our data set (and the world if generalizable) is about 69 years of age. Some other key observations are a minimum life expectancy of about 39 years and a max of about 84 years. Now that we have more information about life expectancy, we can better compare specific regions together and see if different locations are higher or lower than the average life expectancy.

b) Life expectancy distribution – Histogram



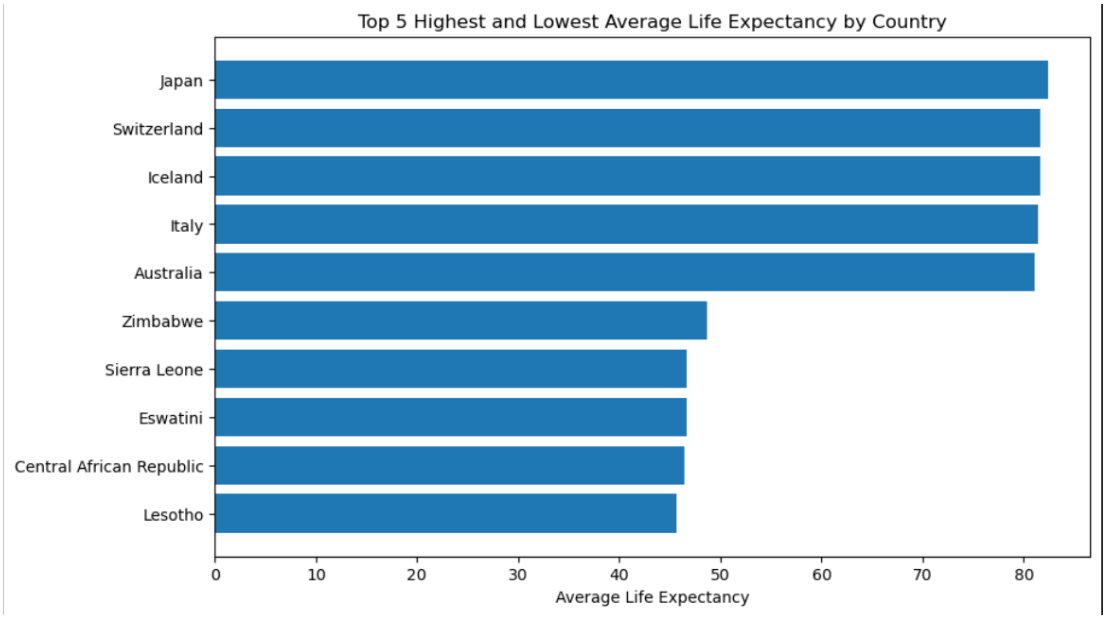
To gain a more complete picture of the distribution of life expectancy, we created a histogram of the life expectancy variable with its density plot overlaid on top. At first glance, we can see that the distribution is decently skewed to the left. This tells us that the majority of the life expectancies are higher, with smaller amounts of life expectancy observations being on the lower side. This tells us that it is more common to see a life expectancy around the mid-seventies (mode) than any other value and it is more uncommon to see life expectancies fifty or less. In terms of society, we want this distribution to be as skewed to the left as possible. However, as will be described in the multiple linear regression portion of the report, this distribution may cause us problems when fitting a model and finding the significant predictors of life expectancy.

c) Life expectancy over time – Line/dot plot



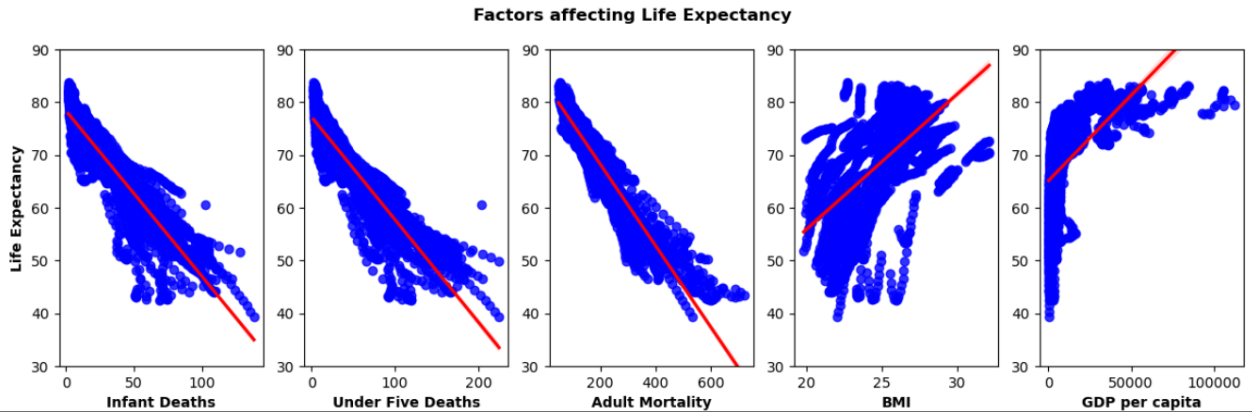
The next important detail we want to analyze about life expectancy is how it is changing over time. This line plot shows the average life expectancy over time and clearly shows us an upward trend. We can observe that from 2000-2015, the average life expectancy is always increasing at a pretty constant rate. This means that average life expectancies of humans around the globe are getting larger and larger as time goes on, which is great news for society as a whole. Hopefully with further research we can see why there is a positive slope and what factors lead to this increase.

d) Countries with the highest/lowest life expectancy – Bar graph



The next topic that we are interested in is what countries have the highest or lowest life expectancies. This bar plot shows the five highest and five lowest countries in regards to average life expectancy. This plot makes it easy to compare these countries to each other and see how large the difference in life expectancies may be. The country with the highest average life expectancy is Japan at around eighty two years and the country with the smallest average life expectancy is Lesotho at around forty seven years. Although this difference does not seem that large on the graph, a forty year difference is very large in the real world. Furthermore, we can see that all of the countries in the bottom five are African countries. This is a cause for concern and should be investigated further by policy makers and nonprofits looking to better society as a whole.

e) Different variables vs. Life expectancy – Scatter plots



The above plot starts to explore the relationship between different variables and life expectancy, which will be further extended below in the correlation matrix section. We chose to look at five different variables that we expected to have some sort of linear relationship with life expectancy, those being: Infant Death rate, Under Five death rate, Adult Mortality, BMI, and GDP per capita. Each variable is plotted in a scatter plot against life expectancy with a line of best fit included to help emphasize the general trend of the plot/relationship between the variables. From the graphs, we can see that each variable besides GDP per capita has a fairly strong linear relationship with life expectancy, as shown by many of the points clustered around the line of best fit. GDP per capita seems to have a more complicated relationship with life expectancy, which we will explore later. Overall this graph tells us that some variables like Infant Death rate, Under Five death rate, and Adult Mortality may be used as good predictors of life expectancy in linear regression whereas GDP per capita might not be.

f) Different variables vs Life expectancy part 2 - Correlation matrix

	Life_expectancy
Infant Deaths	-0.9200319194470860
Under Five Deaths	-0.920419133640263
Adult Mortality	-0.9453603642730650
Alcohol Consumption	0.39915910757917200
Hepatitis B	0.41780443201507800
Measles	0.49001858940944100
BMI	0.5984233246973870
Polio	0.6412174553454280
Diphtheria	0.6275413923742570
Incidents_HIV	-0.5530274644851240
GDP_per_capita	0.5830897215324400
Population_mln	0.026297879724181600
Thinness_ten_nineteen_years	-0.4678244950192930
Thinness_five_nine_years	-0.45816622746008500
Schooling	0.7324844688915010
Life_expectancy	1.0

As can be seen from this relevant excerpt of the correlation matrix that we made, we can see that Infant Deaths, Under Five Deaths, Adult Mortality, and Schooling all have strong correlations with Life expectancy. Furthermore, Alcohol Consumption, Hepatitis B, Measles, BMI, Polio, Diphtheria, Incidents_HIV, GDP_per capita, both types of Thinness, all have moderate correlations with life expectancy. Lastly, Population_mln is the only variable that has a weak correlation with life expectancy. All of these assessments come from the correlation rule of thumb that states any absolute correlation between 0 and 0.3 is considered weak, between 0.3 and 0.7 is considered moderate, and between 0.7 to 1 is considered strong. Furthermore, all variables that are negatively correlated show negative association, meaning when one of the variables increases the other decreases. On the other hand, all of the variables with a positive correlation show positive association, meaning when one of the variables increases the other increases as well. It should come as no surprise that the variables in the above scatterplot have fairly high correlations due to the fact that they all showed relatively strong linear trends, except that of GDP_per_capita which showed more of a logarithmic or polynomial relationship with life expectancy. Due to its high correlation we will still include it in the initial model, however with further research it could be determined that the use of this variable in a linear model is not valid.

Research Question 1: Predicting Life Expectancy:

The first part of answering this question was understanding the distribution of the response variable life expectancy. As can be seen in the above exploratory analysis, the distribution of life

expectancy is not approximately normal, in fact it is heavily left skewed. Thus, fitting a linear regression model wouldn't be appropriate and we had to perform a data transformation. However, after applying transformations such as a log base 10, natural log, square root, square, and cube root transformation, we had no luck in turning life expectancy into one that was approximately normal. Thus we had to accept that our model may be flawed and inaccurate and fit the regression anyways. After fitting the multiple linear regression model onto all of our continuous variables, we got the following model summary:

OLS Regression Results

Dep. Variable:	Life_expectancy	R-squared:	0.979
Model:	OLS	Adj. R-squared:	0.979
Method:	Least Squares	F-statistic:	7038.
Date:	Sun, 28 May 2023	Prob (F-statistic):	0.00
Time:	17:59:01	Log-Likelihood:	-3961.7
No. Observations:	2291	AIC:	7955.
Df Residuals:	2275	BIC:	8047.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.9860	0.698	121.814	0.000	83.618	86.354
Infant Deaths	-0.0569	0.007	-8.071	0.000	-0.071	-0.043
Under Five Deaths	-0.0503	0.004	-11.420	0.000	-0.059	-0.042
Adult Mortality	-0.0488	0.001	-71.233	0.000	-0.050	-0.047
Alcohol Consumption	0.0839	0.010	8.366	0.000	0.064	0.104
Hepatitis B	-0.0102	0.003	-3.562	0.000	-0.016	-0.005
Measles	0.0011	0.002	0.553	0.580	-0.003	0.005
BMI	-0.1567	0.022	-7.255	0.000	-0.199	-0.114
Polio	0.0039	0.007	0.584	0.559	-0.009	0.017
Diphtheria	0.0005	0.007	0.077	0.939	-0.013	0.014
Incidents_HIV	0.1012	0.020	4.969	0.000	0.061	0.141
GDP_per_capita	3.068e-05	2.45e-06	12.536	0.000	2.59e-05	3.55e-05
Population_mln	-0.0002	0.000	-0.959	0.338	-0.001	0.000
Thinness_ten_nineteen_years	-0.0414	0.021	-1.971	0.049	-0.083	-0.000
Thinness_five_nine_years	0.0029	0.021	0.139	0.890	-0.038	0.044
Schooling	0.1042	0.019	5.590	0.000	0.068	0.141

Omnibus:	7.295	Durbin-Watson:	2.046
Prob(Omnibus):	0.026	Jarque-Bera (JB):	8.068
Skew:	0.078	Prob(JB):	0.0177
Kurtosis:	3.245	Cond. No.	4.95e+05

After analyzing the p-values in the middle part of the summary table we came to the conclusion that the following variables were significant predictors of life expectancy at the 5% level of significance: Infant Deaths, Under Five Deaths, Adult Mortality, Alcohol Consumption, Hepatitis B, BMI, Incidents_HIV, GDP_per_capita, Thinness_ten_nineteen_years, and Schooling. The constant term was also very significant at the 5% level of significance. After we found these significant predictors, we tested our first model assumption and assessed if there was multicollinearity present between our significant predictors, finding the variance inflation factor (VIF) of our significant predictors yielded the following results:

	Feature	VIF
0	Infant Deaths	96.465101
1	Under Five Deaths	81.674510
2	Adult Mortality	27.296996
3	Alcohol Consumption	4.632607
4	Hepatitis B	36.083576
5	BMI	71.449627
6	Incidents_HIV	2.970468
7	GDP_per_capita	2.844661
8	Thinness_ten_nineteen_years	3.192527
9	Schooling	28.247637

As common statistical practice states, any VIF higher than 5 (or 10 depending on who you ask) should be removed from the model. However since Infant Deaths, Under Five Deaths, and Adult Mortality are all so highly correlated, removing two should make the VIF of the other variable decrease drastically. With that said we will keep the Adult Mortality, Alcohol Consumption, Incidents_HIV, GDP_per_capita, and Thinness_ten_nineteen_years variables and calculate their VIFs again:

	Feature	VIF
0	Adult Mortality	5.211281
1	Alcohol Consumption	2.815981
2	Incidents_HIV	1.869550
3	GDP_per_capita	1.907810
4	Thinness_ten_nineteen_years	2.670264

Hence after recalculating the VIF for all of our independent variables we see that all of them don’t suffer from any multicollinearity. Thus we split the data into a training and test set and ran another linear regression model, instead just on these 5 significant predictors. Doing so yielded these results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life_expectancy   R-squared:                       0.942
Model:                            OLS   Adj. R-squared:                  0.942
Method:                 Least Squares   F-statistic:                     7374.
Date:                Sun, 28 May 2023   Prob (F-statistic):               0.00
Time:                        20:46:42   Log-Likelihood:                -5144.2
No. Observations:                2291   AIC:                         1.030e+04
Df Residuals:                    2285   BIC:                         1.033e+04
Df Model:                           5
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           83.0503      0.175    473.252      0.000      82.706      83.394
Adult Mortality                 -0.0821      0.001   -114.372      0.000      -0.084      -0.081
Alcohol Consumption              0.2915      0.014     20.440      0.000       0.264       0.319
Incidents_HIV                    0.6726      0.030     22.620      0.000       0.614       0.731
GDP_per_capita                1.264e-05    3.6e-06      3.514      0.000    5.59e-06    1.97e-05
Thinness_ten_nineteen_years      -0.1144      0.012     -9.196      0.000      -0.139      -0.090
==============================================================================
Omnibus:                      169.936   Durbin-Watson:                   2.026
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              518.151
Skew:                          -0.363   Prob(JB):                    3.05e-113
Kurtosis:                       5.214   Cond. No.                     7.77e+04
==============================================================================
```

Again looking at the p-values in the middle part of the summary we see that all of these variables are extremely significant at the 5% level of significance. Hence our final model equation, rounded to one decimal place where possible, was: Life Expectancy = 83.1 - 0.1*Adult Mortality + 0.3*Alcohol Consumption + 0.7*Incidents_HIV + 0.00001*GDP_per_capita - 0.1*Thinness_ten_nineteen_years. These coefficients can be interpreted as the amount life expectancy changes given a 1 unit increase in the independent variables, holding all other variables constant. In this case the constant term doesn't hold a meaningful interpretation so we won't share it. Once we fit our model it was time to check the model assumptions. In particular we checked 5 model assumptions: linearity, no autocorrelation between the model residuals, homoscedasticity of the residuals, normality of the residuals (centered around zero), and lastly no outliers.

a) Linearity assumption - Rainbow test:

In order to check the assumption that the relationship between life expectancy and the above predictor variables is linear, we will run a rainbow test from the statsmodel package. In particular this test has two competing hypotheses: the null hypothesis, which states that the relationship is linear, and the alternative hypothesis, which states the relationship is non-linear. After running this test on our model we obtained a p-value of 0.9947393758103786, hence we fail to reject the null hypothesis and assume that there persists a linear relationship between life expectancy and the predictor variables.

b) No Autocorrelation between the Residuals - Durbin-Watson test:

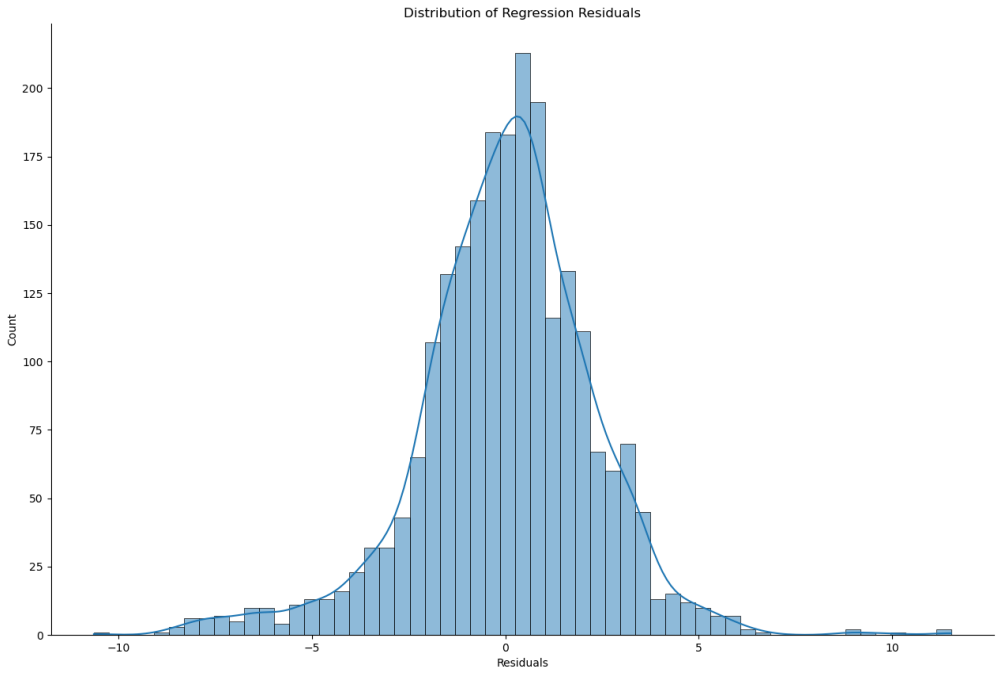
In order to check the assumption that there is no autocorrelation between the residuals, we will run a Durbin-Watson test from the statsmodel package (although it is also shown in the above model output). In particular, this test has two competing hypotheses: the null hypothesis, which states that there is no evidence of autocorrelation, and the alternative hypothesis, which states that there is evidence of autocorrelation. After running this test on our model's residuals we obtained a test statistic value of 2.02 (same as in the above model summary). As the test states, since the test statistic is in the range of 1.5 to 2.5, we fail to reject the null hypothesis, and assume there is no autocorrelation among our residuals.

c) Homoscedasticity of the Residuals - het breuschpagen test:

In order to check the assumption that each residual has equal variance (the distribution of the residuals has constant variance), we will again use the statsmodel package to run a het breuschpagen test. In particular, this test has two competing hypotheses: the null hypothesis, which states that the distribution of the residuals has constant variance, and the alternative hypothesis, which states that the distribution of the residuals has non-constant variance. After running this test we obtained a p-value of 5.2602233700414695e-90, which is way past machine precision and is thus 0. Hence we reject the null hypothesis and conclude that there is definitely unequal variance between the residuals.

d) Normality of the Residuals (centered around zero) - Displot:

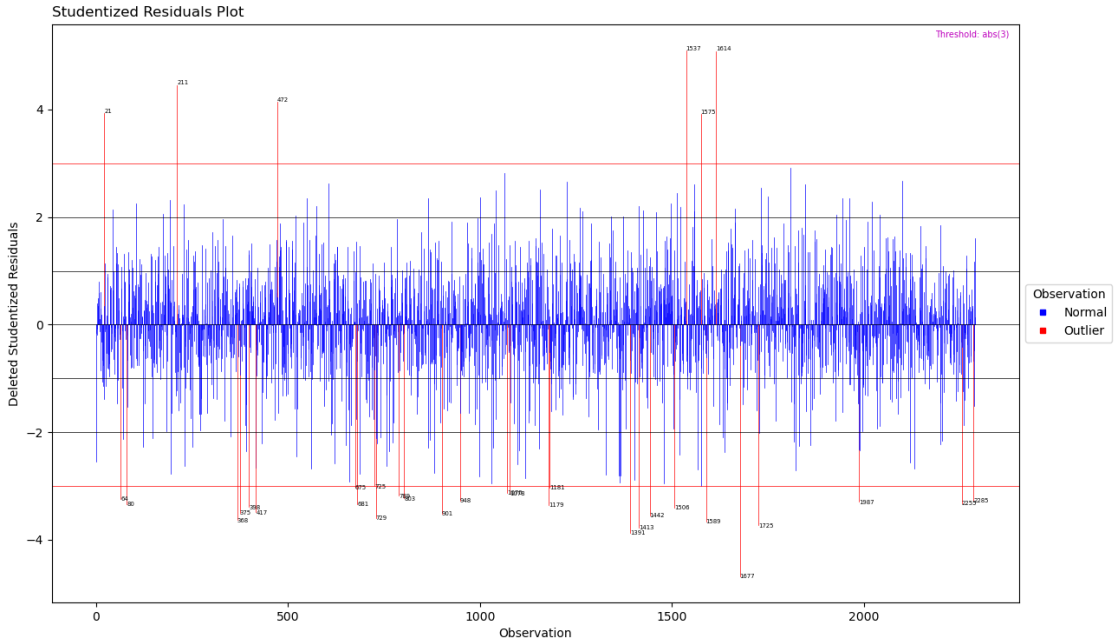
In order to check the assumption that the distribution of the residuals is approximately normal with an expected value of zero, we will use seaborn to create a histogram with its density plot overlaid.



As can be seen from the above histogram of the residuals with its corresponding density plot overlaid, we can see that the shape of the histogram is approximately normal and is centered around 0. Thus we will assume that the residuals are approximately normally distributed with an expected value of zero.

e) No outliers - Studentized Residual Plot:

In order to check that there are no outliers/influential points in our data set/training set, we created a function that creates a studentized residual plot that resembles the ols_plot_resid_stud plot from the R programming language. In particular we will look for studentized residuals that have an absolute value greater than 3, these will appear as red lines in the following plot:



As can be seen from the above studentized residual plot, there are 34 potential outliers/influential points in our data set. Hence we have violated the no outliers assumption. However, given that there were thousands of data points in our data set, relative to the size of the data set itself, this isn't a large amount of outliers.

Since we only passed 3 out of the 5 residual assumptions, and given the fact that the dependent variable isn't normally distributed, we have to question the reliability of this model. In a real life scenario, in order to fit this model, we would need to do a lot more work to make sure these assumptions are passed or at the very least only slightly violated.

The last thing that needed to be done to assess the result validity of this model was to compute the training and testing mean squared error, which is the golden standard for model accuracy testing of regression models. In our case, the training set mean squared error was

5.222004318670649, and the testing mean squared error was 5.489100109185608. These values aren't bad at all considering that many of the model assumptions failed and given that there are many data points in our data set.

Conclusion:

After fitting the model we can see that some of the variables such as BMI, GDP per capita, Mortality rates, etc. were all significant predictors as expected. However, it may come as a surprise that most of the immunization rates were not significant predictors of life expectancy. One reason why this may be the case is that immunizations are usually given to prevent rarer diseases/infections in the first place, hence in the long run, these types of diseases usually don't impact the overall prediction of a country's life expectancy for that given year anyway, so having a high or low value for these immunizations don't play as a big of a role as something like adult mortality rates. Furthermore, the variable doesn't track immunization rates for those who receive immunizations after the age of 1, hence missing a lot of important information, which could've played a role in predicting life expectancy.

Also, despite the fact that the model violates many of its key and non-negotiable assumptions, it is important to note that there is an inherent oversimplification of this model in the first place. To model such a complex phenomenon such as life expectancy with only a few independent variables doesn't do justice to the complexity of the problem itself, and in turn reduces the ability for smaller mean squared error values and thus leads to inaccurate predictions when introduced to new unseen data that has no relation to this dataset. Thus, if researchers were to use this model they would need to be very skeptical of the results obtained. Although, when accurate, these predictions are a great way to be able to see how life expectancy is expected to change in the future barring any unforeseen circumstances such as ones similar to COVID-19.

Research Question 2: Classifying Development Status:

Since we decided to not test the logistic regression assumptions and instead compare two different models on the basis of accuracy, we can jump right in to fitting the model to our data. Even though logistic regression can have categorical variables as independent variables since there aren't many useful categorical variables in this data set, we are instead going to use the same continuous variables we used for our multiple linear regression model, but instead of an equation that predicts life expectancy, we now shift our focus to an equation that predicts/classifies if a country is a developed or developing nation. To do this we will use the Economy_status_Developed variable which encodes the Devlop_status variable as either 1 if the nation is developed, or 0 if the nation is developing. Using the statsmodels package again, we ran a logistic regression model on the data and here were the results:

Logit Regression Results

Dep. Variable:	Economy_status_Developed	No. Observations:	2291
Model:	Logit	Df Residuals:	2275
Method:	MLE	Df Model:	15
Date:	Sun, 28 May 2023	Pseudo R-squ.:	0.8920
Time:	21:23:54	Log-Likelihood:	-127.01
converged:	True	LL-Null:	-1176.0
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-12.9646	6.919	-1.874	0.061	-26.526	0.597
Infant Deaths	-0.2858	0.120	-2.380	0.017	-0.521	-0.050
Under Five Deaths	0.2811	0.077	3.644	0.000	0.130	0.432
Adult Mortality	-0.0201	0.006	-3.215	0.001	-0.032	-0.008
Alcohol Consumption	0.9303	0.108	8.608	0.000	0.718	1.142
Hepatitis B	0.0308	0.019	1.650	0.099	-0.006	0.067
Measles	-0.0723	0.020	-3.668	0.000	-0.111	-0.034
BMI	-0.1936	0.191	-1.013	0.311	-0.568	0.181
Polio	0.0323	0.085	0.380	0.704	-0.134	0.199
Diphtheria	0.0332	0.087	0.383	0.702	-0.137	0.203
Incidents_HIV	-14.0964	2.558	-5.512	0.000	-19.109	-9.084
GDP_per_capita	0.0001	2.5e-05	4.946	0.000	7.46e-05	0.000
Population_mln	-0.0100	0.005	-1.863	0.062	-0.021	0.001
Thinness_ten_nineteen_years	-0.4614	0.535	-0.863	0.388	-1.510	0.587
Thinness_five_nine_years	-0.5202	0.509	-1.022	0.307	-1.517	0.477
Schooling	1.0912	0.193	5.655	0.000	0.713	1.469

As can be seen from the above model summary, the significant predictors of developmental status at the 5% level of significance are: Infant Deaths, Under Five Deaths, Adult Mortality, Alcohol consumption, Measles, Incidents_HIV, GDP_per_capita, and schooling. Unlike what we did with the linear regression model, we won't be checking any of the model assumptions for this model. Hence we can split the data into a training and testing set (80/20 split) and refit the model onto only the significant predictors. Notice that, unlike last time, the constant term is insignificant, so we won't have a constant in this model. Refitting the model onto the significant predictors we obtain the following results:

Logit Regression Results

Dep. Variable:	Economy_status_Developed	No. Observations:	2291
Model:	Logit	Df Residuals:	2283
Method:	MLE	Df Model:	7
Date:	Sun, 28 May 2023	Pseudo R-squ.:	0.8741
Time:	21:23:54	Log-Likelihood:	-144.46
converged:	True	LL-Null:	-1147.5
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Infant Deaths	-0.6300	0.104	-6.056	0.000	-0.834	-0.426
Under Five Deaths	0.3268	0.064	5.095	0.000	0.201	0.453
Adult Mortality	-0.0206	0.005	-4.401	0.000	-0.030	-0.011
Alcohol Consumption	0.7785	0.075	10.320	0.000	0.631	0.926
Measles	-0.0787	0.013	-6.099	0.000	-0.104	-0.053
Incidents_HIV	-12.1606	2.203	-5.520	0.000	-16.478	-7.843
GDP_per_capita	6.73e-05	1.57e-05	4.276	0.000	3.65e-05	9.82e-05
Schooling	0.5031	0.103	4.863	0.000	0.300	0.706

Hence our final model equation, rounded to one decimal place where possible, was: $log(\frac{P(Developed)}{P(Developing)})$ = -0.6*Infant Deaths + 0.3*Under Five Deaths - 0.02*Adult Mortality + 0.8*Alcohol Consumption - 0.1*Measles - 12.2* Incidents_HIV + 0.00007*GDP_per_capita + 0.5*Schooling. What this equation is modeling is the log odds ratio between the probability of being a developed country, and the probability of being a developing country. Thus each coefficient represents the change in the log odds ratio for a one unit increase in that given variable, holding every other variable constant. Although log odds can be confusing, its power comes in when we do classification, because it turns an output range of [0, 1], into an output range of [-∞, ∞]. Thus, any data point that is positive we can classify as developed, and any data point that is 0 or negative we can classify as developing. Instead of manually doing these computations we will instead use the accuracy_score function from the sklearn package to compute the training and testing accuracy of the model for us. After testing the model for accuracy we obtain a training accuracy of 97.4%, and a testing accuracy of 97.6%. As can be seen the model works well on unseen data and provides us some assurance that the model is doing a good job overall, although we must be skeptical about the results since we didn't test model assumptions.

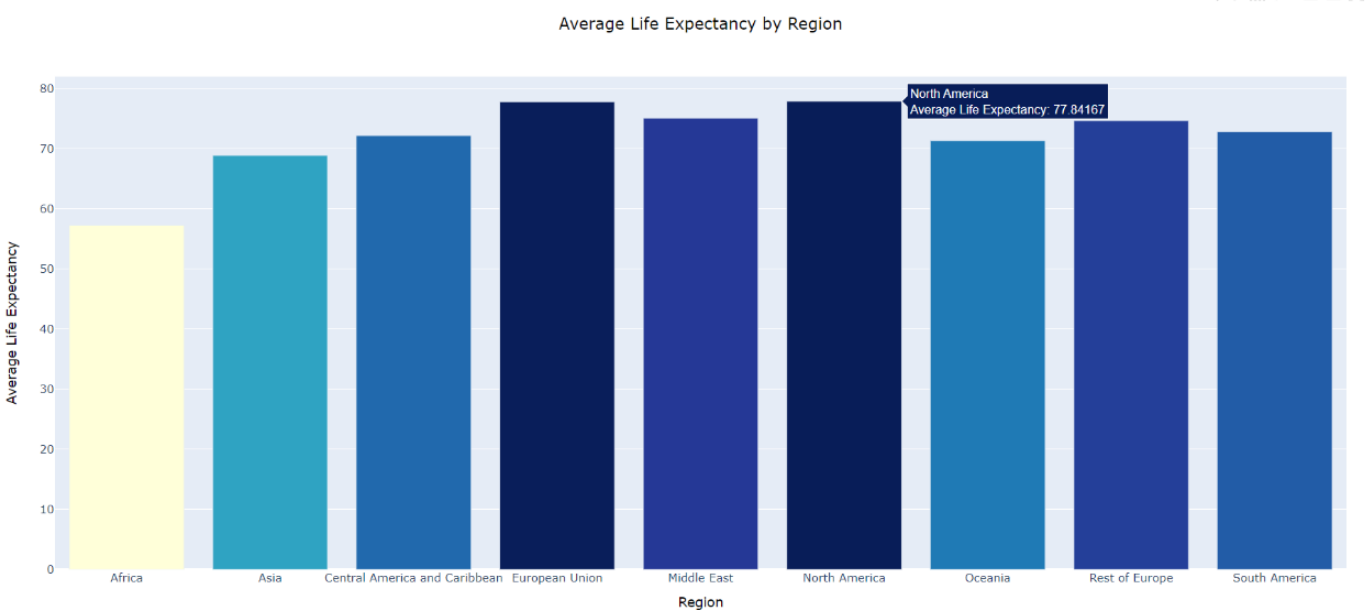
The last part of this question is, even though the logistic regression model had very high accuracies, is there a similar simple model that works even better? To answer this question we decided to fit a DecisionTreeClassifier from sklearn onto the same exact data we used for our logistic regression model. After fitting the model we tested its training and testing accuracy and found a training accuracy of 100% and a testing accuracy of 98.1%. As can be seen the DecisionTreeClassifier is slightly better in both training and testing accuracy.

Conclusion:

After fitting the model we can see that some of the expected variables such as BMI, GDP, Mortality rates, etc. were all significant predictors as expected. It also turns out that this model worked very well on unseen data, although we can't be too confident in our results because we didn't check the validity of the underlying model assumptions. The main limitation to the use/accuracy of the model and model accuracy scores themselves is that we used one of if not the most simple tools in validation, a training and testing set. With that said, we can't be sure that our model would always produce the same accuracy for different testing sets because models themselves are subject to the data they are given, and random chance can skew the results in ways that differ from reality. Thus, before being used in any commercial setting the model would need to be tested from top to bottom, and the results would need to be validated in a more rigorous way. Lastly, one reason why the DecisomnTreeClassifier performed better than our logistic regression model on the same data, is that the classes in sklearn have been optimized by higher level programmers with much more experience than us, thus it makes sense that their results are less error prone than the ones we came up with, and hence why they have higher accuracies in both training and testing.

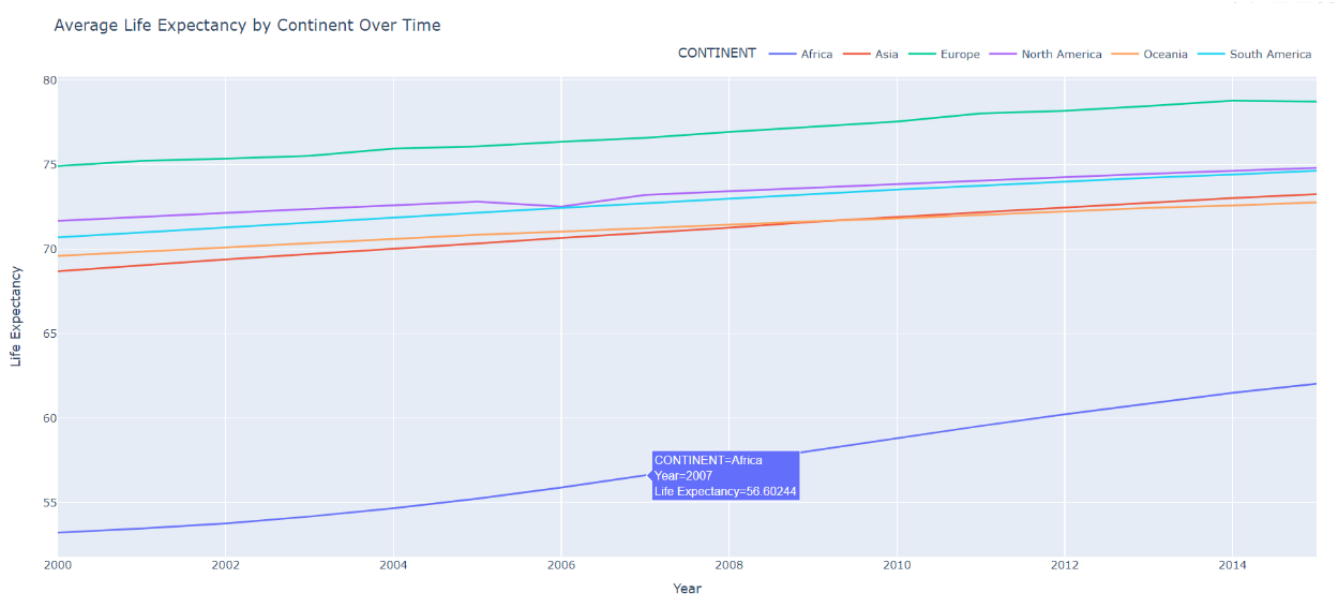
Research Question 3: Temporal and Geographical Analysis:

a) Life expectancy by region – Interactive bar chart



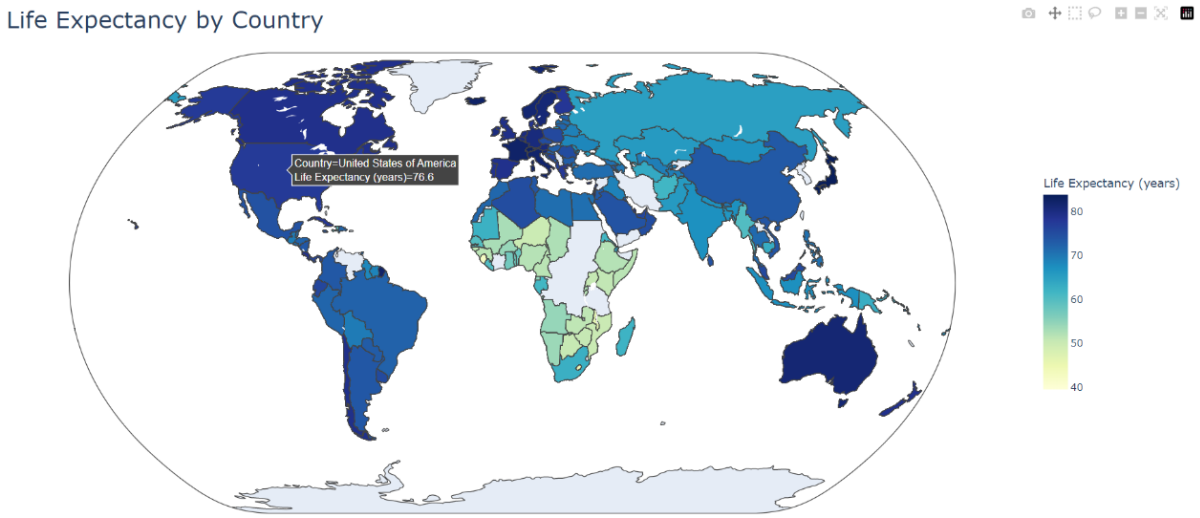
In this bar chart, we can see the average life expectancy for each region as specified by the data set. This allows us to more directly compare how each region compares with respect to life expectancy and which regions have the highest or lowest average life expectancy. The interaction allows the user to hover over each bar to more easily see the region's name and its exact (rounded) average life expectancy. We can observe that North America has the highest average life expectancy at 77.84 years and Africa has the lowest at 57.22 years. We can also observe that the majority of the regions are fairly similar to each other, the only region that is notably different from the rest is Africa, with a smaller average life expectancy as described above. This breakdown of average life expectancy by region further emphasizes what we have already been seeing through the report already; that Africa has the lowest life expectancies on average and that most of the focus from researchers and social scientists should be on improving the quality of life and technological advancements for the citizens living in African countries.

b) Life expectancy by continent – Interactive line graph



Now switching gears and looking at the scope of continents, we want to analyze how the average life expectancy is changing over time. Above is an interactive line plot that shows how the average life expectancy per continent is changing over time. The user interaction allows you to hover over each line to more easily see the continents name and its exact (rounded) average life expectancy. Similarly to the line graph in the Data Exploration section, we can see that each line is increasing over time. Just like the interactive bar graph for regions, we can see that most continents are fairly similar, with only Africa being notably different. We can conclude that all continents have an increasing average life expectancy over time.

c) Life expectancy by country – Interactive choropleth



Finally, here is an interactive choropleth that shows each country colored by its average life expectancy. The countries in darker blue have a higher average life expectancy and the lighter blues represent shorter life expectancies. The interaction allows the user to zoom in and out as well as see the country's name and its exact (rounded) average life expectancy. This allows the user to more directly compare how life expectancy is different per country and how life expectancy varies on a map. Recall that we found North America to have higher average life expectancies, hence it has darker blue coloring and Africa has smaller average life expectacies, hence the lighter colors. Also note that we have some missing data in Africa, this reveals that our analysis for Africa may be inconclusive or incorrect, maybe these missing countries would increase the average life expectancy.

Conclusion:

In conducting analysis of life expectancy across different geographical levels and locations, we can make several conclusions. One takeaway is that the average life expectancy is relatively similar regardless of location, at around the high sixties to low seventies. The only exception to this is Africa, who's average life expectancy is a bit lower at around fifty seven years. However, this difference may be due to incomplete data. These trends are most apparent when considering continents and regions around the world. We can also learn that the average life expectancy is always increasing, even for Africa. Although Africa stands out with lower life expectancies we can see that it is slowly catching up to the rest of the world. Understanding these geographical differences in average life expectancy is crucial in identifying what leads to higher or lower life expectancies. This can help us identify which countries, continents, and regions have higher life expectancies and learn what aids these locations in life expectancy. It also helps us see which locations may need help raising their life expectancy and we can work towards providing them with the necessary resources.

Impacts:

An important impact and one of our main motivations are the possible public health interventions that can be made with the aid of our analysis. We have seen what parts of the world have lower life expectancies and what type of factors may lead to lower life expectancies. We can use this to our advantage and develop targeted public health interventions that may improve life expectancy in the areas that need help. Governments, healthcare organizations, and policymakers may use our findings to design strategies to improve health and extend life expectancy.

Another possible impact of our research is the advocacy and awareness of life expectancy. Our research highlights the difference in life expectancy per location and the factors that may lead to lower life expectancies. This may raise attention to these factors and help people avoid them. Additionally, it may help empower advocacy groups or other organizations to help advocate for longer life expectancies.

Limitations:

One limitation of this project is the data set. Our analysis heavily relies on the quality and completeness of the data set. We found that some countries are missing, which you can clearly see in the interactive map plot. With those missing countries, the map looks incomplete and we are unable to see how those missing countries may have affected the continent and region they are in. If those countries had a much higher or lower life expectancy than its surrounding countries, that would be a key detail that is overlooked. Additionally, we may not have all the relevant factors that may have an impact on life expectancy. The author of our data set stated that one of the main motivations they had for creating it was the inclusion of different immunization and health related factors, however, there is no possible way to account for every possible factor. This means that we will never have a perfect model for life expectancy. Another possible problem that arises in the data set is the data collection process and when the data was collected. The data is from the Global Health Observatory (GHO) data repository under World Health Organization (WHO), which is a trustworthy site, however, we do not have the specifics of how this data was collected or by who. This may have some unforeseen problems like bias or data manipulation. We also know that the data set only contains data from 2000-2015, thus it may not be relevant to making conclusions about the current conditions and life expectancies around the world.

Another limitation of our analysis are the ethical considerations. When interpreting why life expectancy is different per region or country, it is important to also consider the inclusion of different factors that may bring up ethical concern. Factors like privacy and confidentiality, bias and discrimination, or social and environmental aspects may have an impact on life expectancy. It must be decided if factors like these are appropriate and should be included in analysis.

The last key limitation of our analysis is our lack of mastery in the subject of regression and machine learning itself. As mentioned numerous times above, many of our model assumptions failed, including the all important normality of the response variable. Without knowledge of how to use the right transformation, we had to run the model without fulfilling one of the most critical parts of the model itself, hence we had to be skeptical of every result we got because it was on the basis of a flawed assumption. Furthermore, our lack of mastery in the subject of regression made it so that we couldn't create the best and most accurate model possible. Without this expertise, we wouldn't be able to give researchers the ability to have any assurance in the predictions they obtain from our models and thus they wouldn't be able to use those predictions to try and enhance/improve life expectancy in areas of concern or be able to forecast if current implementations are paying off. Lastly, with the flawed assumptions and our lack of ability to fix them, we run the risk of misrepresenting the actual reality of the world and thus our model could serve to harm people in need. For example, if our model projects a country is on the right track and leads to governments not giving countries additional aid when they in fact need it, this could lead to unwanted loss of life in countries who need the most help.

Challenge Goals:

In this project, we planned to meet 4 challenge goals: Multiple data sets, External Library, Result Validity, and Machine Learning. In reality we ended up also doing the data cleaning challenge goal as well, although this doesn't show up much in our code as it was mainly done before we even started coding such as fixing column names, etc. Although all of these challenge goals were intertwined, we mainly focused on result validity and machine learning.

Challenge Goal 1: Machine Learning:

For this project we plan on creating one or more of the following models: Multiple Linear Regression, Logistic Regression, and Principal Component Analysis in order to predict or classify certain aspects of our data that we find interesting. We will probably use sklearn or some other library to create this model. However, there is a possibility that we make our own class for these models, in order to fully show our understanding and growth in this challenge goal.

Reality:

The main parts of this challenge goal that differed from what was actually done in the project itself was fitting a principal component analysis model and the fact that we didn't create any classes ourselves. The main reason we didn't implement PCA into our project was due to the fact that we had already implemented multiple linear regression and logistic regression models and adding an even more complicated model would've been too much to do. Instead we decided to add a simpler third model, that being a DecisionTreeClassifier. Furthermore, we decided not to create our own classes simply due to the fact that we didn't have the knowledge on linear and logistic regression in order to do so. With that said, we were still able to implement regression models using an outside package, statsmodels, as well as use sklearn to create a third model.

Challenge Goal 2: Result Validity:

In the world of modeling there are many ways to assess which of many models is better at doing some sort of prediction or classification task. Some of these assessments/tests include Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R squared to name a few. We plan on implementing these tests ourselves or using relevant packages in python in order to compute these values and find the best model for the task at hand.

Reality:

Although result validity was still a major part of our project, and one that ended up taking a lot of time to do, the majority of what we wrote down for this challenge goal in the proposal drastically changed. In particular, we ended up not implementing these testing functions ourselves due to the fact that we just didn't have the statistical knowledge to create these functions from scratch. Instead we used statistical tests from the statsmodel package, as well as creating plots using seaborn, and lastly creating our own functions to make some of these testing plots. Furthermore, our testing shifted from comparing models to testing the assumptions of the models themselves. We decided to make this change because testing and assessing the assumptions of a model itself allows us to justify if a model should be used or not, which falls under the title of "result

validity" more than comparing models does. Another aspect of result validity that we didn't account for was testing our models training and testing accuracy.

Plan Evaluation:

In this section we will assess our tentative step-by-step plan for the project that we proposed in the project proposal portion of this project. Before discussing the plan from the project proposal it is important to note that things such as turning/refactoring our code into the main method pattern, finding relevant packages, sifting through documentation, and surfing stackexchange for help took up much more time than was anticipated and won't really be talked about in the following evaluation.

1. Finalize exactly what we want to do. (2 hours)
    a. This part of the plan is hard to give an exact time estimate because we didn't really have an idea of exactly what we wanted to do until we found out that our original plan probably wasn't feasible. So technically, this part of the plan took around 8 hours to complete. However, in reality we didn't discuss this for 8 hours, instead it took us about 8 hours into coding/working on the project to realize exactly what it is we were going to do for the project itself. This took more time than initially thought because we weren't 100% locked into what we put down in the proposal. In particular, we decided to scale back the work in some places, while increasing the work in other places. Had we known exactly what we wanted to do from the start this part of the project wouldn't have taken as much time as it did.
2. Exploratory data analysis to fully understand what we're working with. (2 hours)
    a. This part of the plan is fairly accurate, although similar to the last part, this part of the project was one that was worked on in many different sessions so it is hard to get an accurate time estimate. In total we spent around 2-3 hours working on this part of the project. This took more time than initially thought because we decided to expand the amount we did in the exploratory analysis itself.
3. Create mock visualizations. (3 hours)
    a. This part of the project only took us about 1 hour maximum. The reason why the estimated time spent was an overestimate was because most of this work actually fell into the "Exploratory data analysis" portion of the project. Thus, creating these visualizations didn't take as long as initially expected.
4. Learn plotly/bokeh/other packages to make our visualizations interactive. (6 hours)
    a. This part of the project took about 6 hours to complete, although it didn't pan out exactly how we expected it to go. What we mean by this is that we intended to learn plotly and then put these skills to use by making our visualizations. Instead it actually turned out that we learned bits and pieces of plotly while creating our visualization instead of learning it entirely. The reason why this part of the project

was so close to reality is because it was one of our toughest parts of the project and really challenged us since neither of us had much experience with anything related to HTML before.

5. Create and test multiple linear regression model(s). (6 hours)
    a. This part of our project was one of the most difficult things to do due to the fact that there were so many parts to it and that there were so many different tests/functions we had to create in order to get the results we wanted. In particular, this part of the project took upwards of 8 hours to complete, with most of the time spent looking for and creating functions for model testing and debugging. The main reason why this part took longer than expected was because we underestimated how long it would take us to find the relevant functions and packages needed to perform our statistical tests, as well as underestimating the amount of bugs we would run into while completing this part of the project.
6. Create and test logistic regression model(s). (6 hours)
    a. This part of our project is identical to the last part except we ended up not testing the model assumptions and instead compared the logistic regression model to a DecisionTreeClassifier in terms of accuracy. This part of the project in reality only took us about 2 hours to complete. The main reason why this part of the project took way less time than anticipated was because we decided to alter our challenge goal due to the fact that it was out of our realm of knowledge, and furthermore, we worked out most of the kinks of working with the statsmodels package and overall model testing in the previous part.
7. Write report/final presentation deliverables (10 hours)
    a. Lastly, this part of the project, like many of the others in this list, is very hard to estimate because ironically it contains parts of the project that we haven't done yet such as the final presentation slides and video. However, the estimated time seems to be fairly accurate to the time we have spent so far, and how much time is remaining in finishing up the deliverables. The time it actually took us to complete this part of the project is surprisingly accurate because writing reports is always the most time consuming part of any project, hence why we gave it the highest time allotment.

Testing:

Given that our project was much like assessment 4 and 6 in which testing was not possible, testing our code proved to be a challenge for our group. Thus, with a little inspiration from statistics and the TAs in office hours, we found different ways to ensure that our code was accurate. The main ways we tested were through the use of statistical tests (as explained in the multiple linear regression results section), and through the use of matching the expected value of our plots with the data they were plotting to ensure our plotting functions were working correctly. Below we will give an example of the testing we did in each relevant section of our

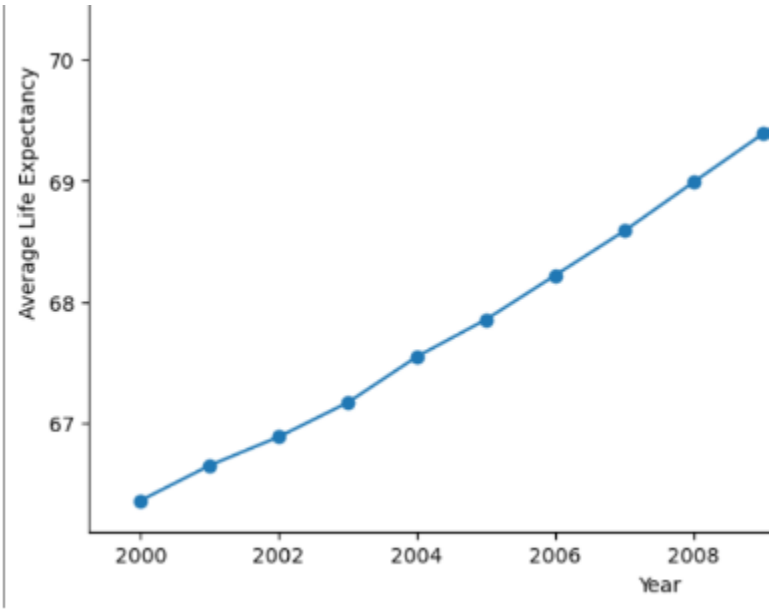
results page, these examples include testing our line plot, studentized residual plot, and choropleth map.

1. Testing Life expectancy over time – Line/dot plot:

To test the graphs, we found expected values from our data set and checked to see if the graph was outputting the correct value. For example, running the following code,

```
avg_2007 = life_expect[life_expect['Year'] == 2007]
print(avg_2007['Life_expectancy'].mean())
```

Returns the value 68.58826815642458. This tells us that the average life expectancy for the year 2007 is about 69 years. Then if we look at the Line/dot plot in the exploratory section, we can see that this is plotted correctly.



We know this because we can see that the value on the line that corresponds to 2007 is at about 69 years. Similar reasoning can be applied to any value on any of our graphs. Although it is not possible to directly use asset statements to see if our graphs are correct, we can intuitively understand when the graphs are incorrect and use code to check certain values to double check ourselves.

2. Testing studentized residual plot:

Following a similar strategy as the line/dot plot, we can make sure the red lines in the data set are corresponding to actually large studentized residuals.

```
print(len(resid[abs(resid) >= 3]))
```

The output of this line of code is 34, which exactly matches the count we got when we "eyeballed" the plot back in the research question #1 results section. Thus, we have some sense of security that the studentized residual plot is actually plotting the data we want it to.

3. Testing interactive choropleth:

Again, Following a similar testing strategy as the line/dot plot, we can also test to make sure that the interactions are working for our third research question. This is even easier than before because the interaction shows us exact values. Using this code to find the average life expectancy for Canada,

```
avg_Canada = merged_data[merged_data['Country'] == 'Canada']
print(avg_Canada['Life_expectancy'].dropna().mean())
```

We get the value 80.6125. Then comparing this with our interactive choropleth

Since these values are the same, we know that our graphs are working as intended.

Collaboration:

In this project we collaborated with no other groups.

Works cited:

Here is a list of resources we used to write the code for this report, as well as some of the inspiration for the project in general.

- Plotly documentation: https://plotly.com/python/
- Plotly express documentation: https://plotly.com/python/plotly-express/
- Inspiration for making interactive choropleth map using plotly express: https://stackoverflow.com/questions/75980836/i-made-a-plotly-express-choropleth-mapbox-of-us-zip-codes-can-i-add-a-choroplet

- Inspiration for making multiple scatter plots in a single figure: https://stackoverflow.com/questions/55126088/scatter-plot-grid-faceted-by-columns-in-matplotlib-or-seaborn
- Matplotlib documentation: https://matplotlib.org/stable/index.html
- Seaborn distplot documentation: https://seaborn.pydata.org/generated/seaborn.displot.html\
- Statsmodel documentation: https://www.statsmodels.org/stable/index.html
- Inspiration for making studentized residual plot: https://rpubs.com/pfr088883/1033107
- Rule of thumb for the Durbin-Watson test: https://www.statology.org/durbin-watson-test-python/