Final Project

Assessing the Predictability of Life Expectancy

Jonah Zembower, Ben Nicholson, Andrew Smith, Maria Morales Laguarda

December 11, 2024

Dr. Jared Burns

SMA 235 Probability and Statistics II

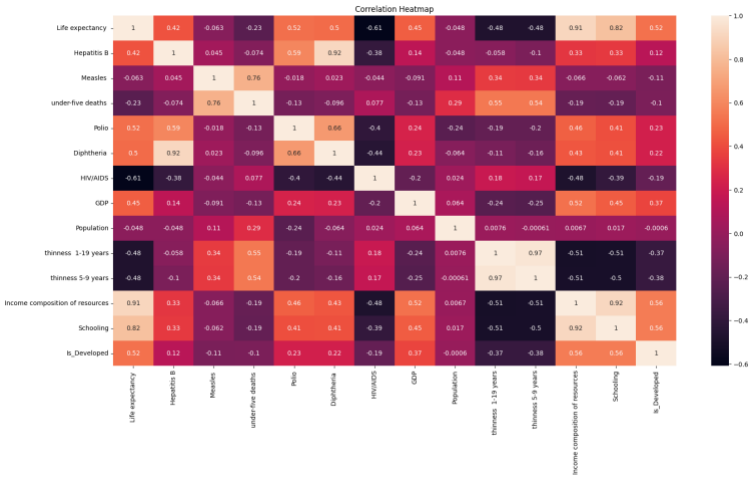
Abstract:

- This dataset was found on Kaggle from the World Health Organization (WHO) and United Nations (UN) data. It includes various variables such as Country, Year, Status, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles , BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources, and Schooling (*Life expectancy (WHO),* 2018).
- There are 20 total predictor variables related to income, disease rates, population rates, and mortality rates.
- Data was collected from 2000-2015 for 183 countries.
- Missing data was found a lot in countries that were lesser known.
- The goal for the future is to see if these predictor variables are effective in predicting the life expectancy of a given country in a single year or seeing the trend across multiple years.

EDA:

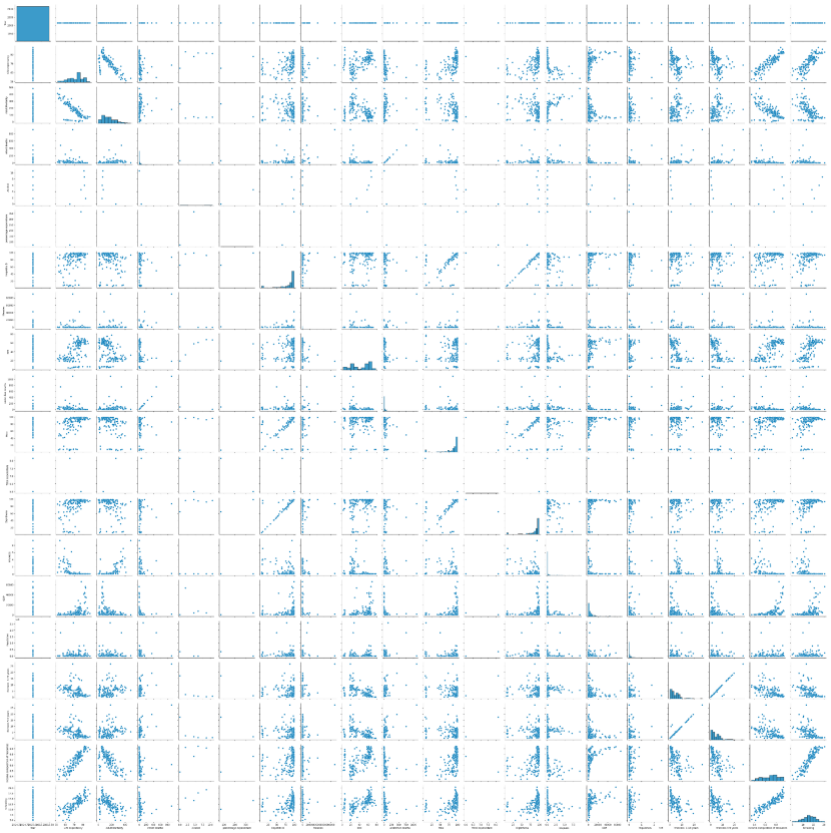
- We are planning to focus on the life expectancy being our predicted variable with multiple variables to assess for predictiveness of that value.
- Let’s look at the given data visualizations to see how the data looks as a whole.

Correlation Matrix:

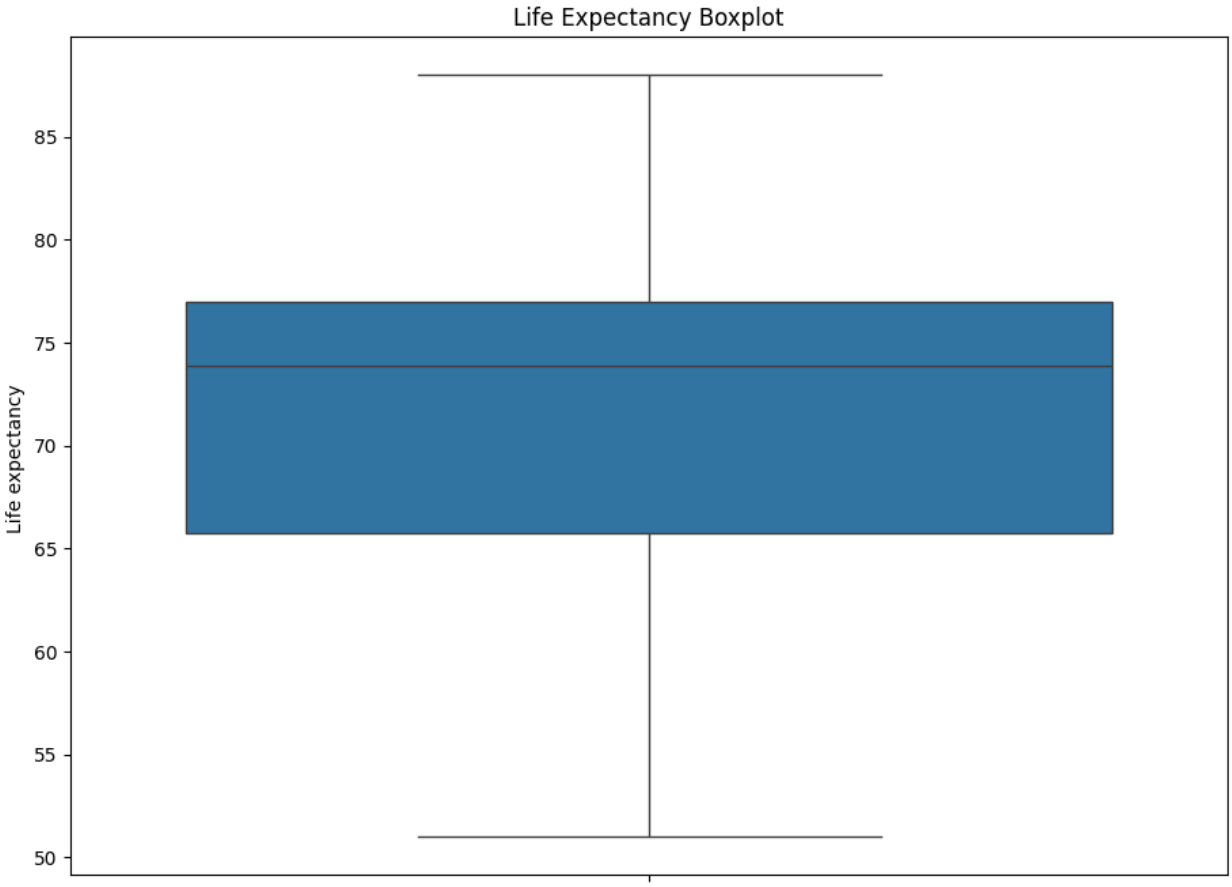


- Showing the individual correlation values both positive and negative. Positive meaning as one increases so does the other and negative meaning as one increases the other has an opposite effect.

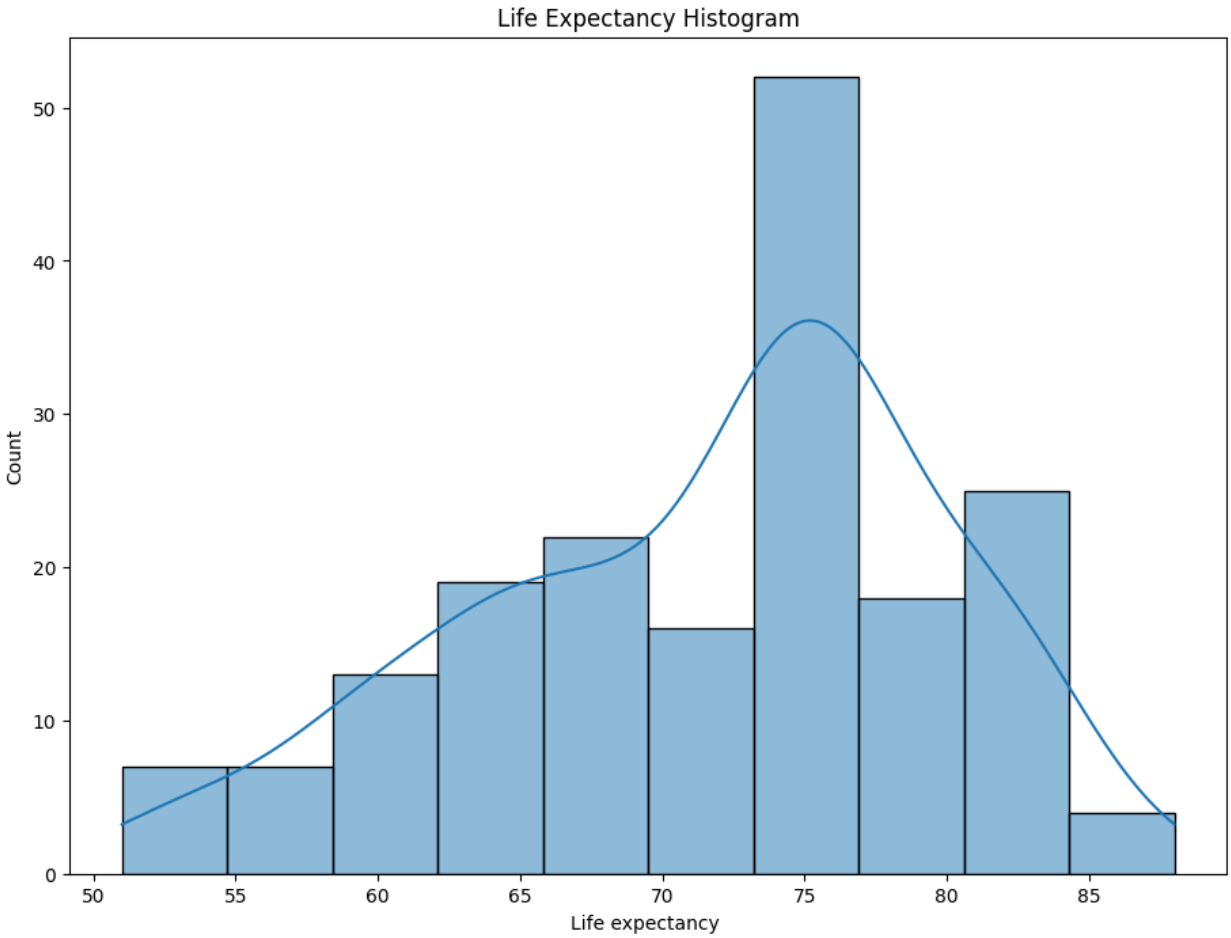
Pairplot:



- This pairplot is showing the relationship between each of the variables in a scatter plot. Similar to what we would see from the correlations.



- A box plot showing the values that are life expectancy in the dataset.
- We see the general median is around 74 and the Q1 is about 66 and Q3 is about 77. The IQR appears to be about 11. There is a large range as well of values from 52 to 87.



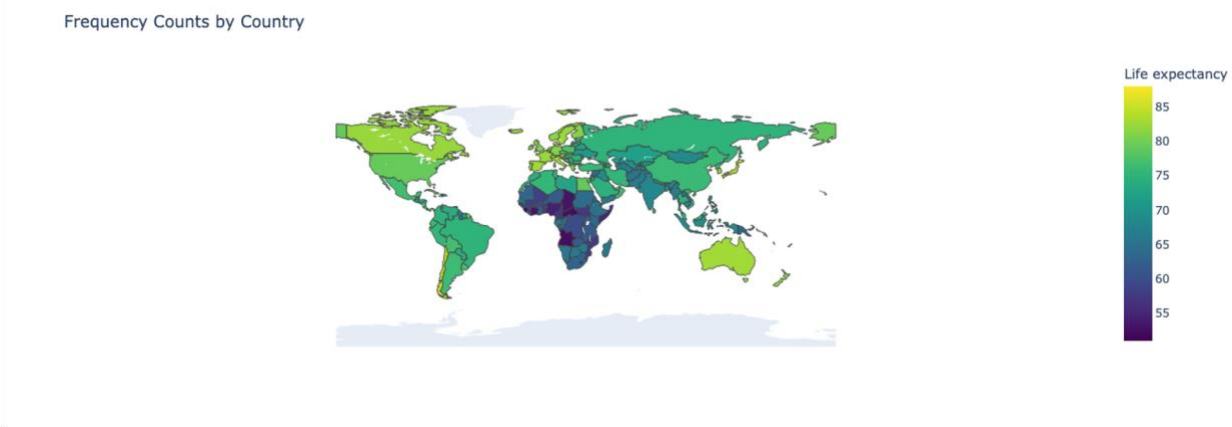
- A histogram of the data for life expectancy.
- It appears that 75 is the average life expectancy for most countries.
- Doesn’t appear to be exactly the normal distribution.
- We found from this the highest and lowest life expectancy values:

-

Country	Life expectancy
Sierra Leone	51.0
Angola	52.4
Central African Republic	52.5
Chad	53.1
Côte d'Ivoire	53.3

Country	Life expectancy
Slovenia	88.0
Denmark	86.0
Cyprus	85.0
Chile	85.0
Japan	83.7

- Furthermore, we also wanted to look at a map of all the countries, and their specific life expectancies based on their location in the world. This also adds to what we are hoping to understand about how certain countries tend to live longer than others.



Hypothesis Test:

- H_O: We assume that no variables can aid in the prediction of life expectancy for a given country.
- H_A: There are some variables that can predict life expectancy for a given country.

- We are assessing this with a p-value of 0.05 for statistical significance.
- Skewness Results of the data:
  - 

	Skewness
Life expectancy	-0.494597
Hepatitis B	-2.067015
Measles	7.398626
under-five deaths	6.323516
Polio	-2.188802
Diphtheria	-2.349014
HIV/AIDS	3.389411
GDP	2.867168
Population	6.168962
thinness 1-19 years	1.954519
thinness 5-9 years	1.968618
Income composition of resources	-0.340756
Schooling	-0.225799
Is_Developed	1.726101

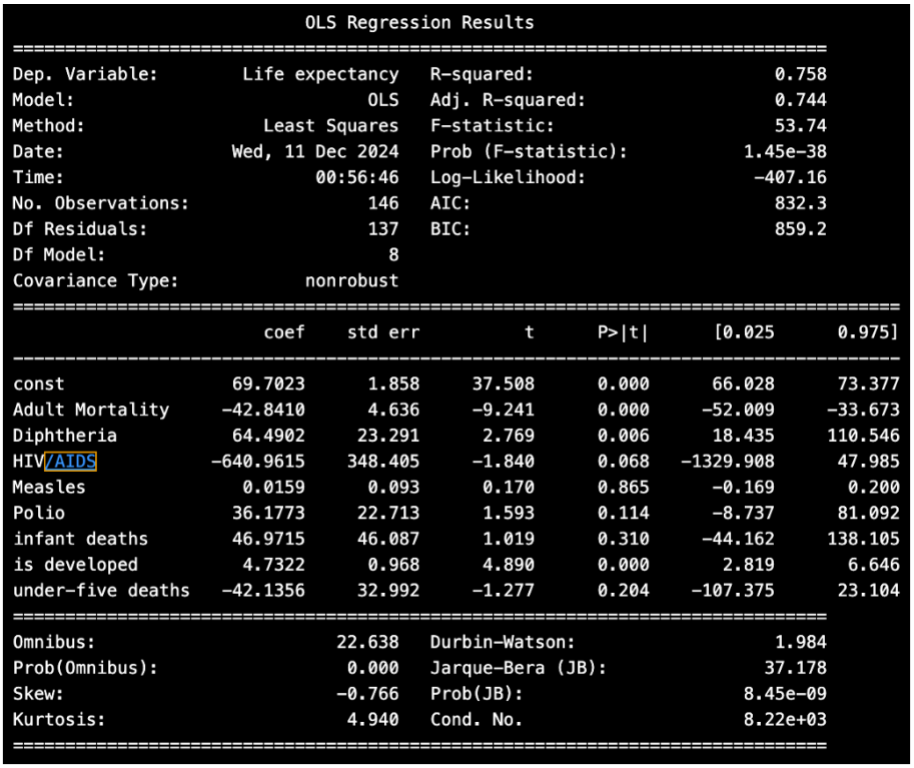
    - The general skewness showcases very high positive and negative values. When we look at skewness, we look for values that are between -0.5 and 0.5 as low skewness and values that are less than -1 and greater than 1 as high skewness.
    - The highest skewness variables are Hepatitis B, Measles, under-five deaths, Polio, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, and Is_Developed.
    - The low skewness variables are Life expectancy, income composition of resources, and schooling.
    - Overall, this data appears to be very skewed which may affect the results of life expectancy potentially.
- Multiple Linear Regression Standardized Values 2015 data:

- We first attempted to do the 2015 data to focus on that and what predictor variables are most meaningful for it.
- Initial results showcased a lot of multicollinearity. Here are the steps that were taken.
-

Feature	VIF
Hepatitis B	8.231820
Measles	3.240833
under-five deaths	5.071519
Polio	2.818256
Diphtheria	9.757366
HIV/AIDS	1.874652
GDP	2.028845
Population	1.730861
thinness 1-19 years	13.995553
thinness 5-9 years	13.265933
Income composition of resources	9.178848
Schooling	9.228119
Is_Developed	1.711962
const	76.120879

- Any VIF greater than 10 is indicative of high multicollinearity and it gets stronger the higher you go. This was found according to a study from the NIH (Kim, 2019).
- So, let's consider what we can remove here then to focus on a certain amount of predictors.
- We have assessed the multicollinearity and we decided to make some changes to the data for 2015 that we could focus on. First, we dropped all of the columns which had NA values (Hepatitis B, GDP, Population, Thinness 5-9, Thinness 10-19, Income Composition of Resources, Schooling). Also, we standardized the predictor variables to make them similar across the board for this initial analysis.

We had a multiple linear regression with an $R^2 = 0.758$, but there were some high p-values for some variables:

- 

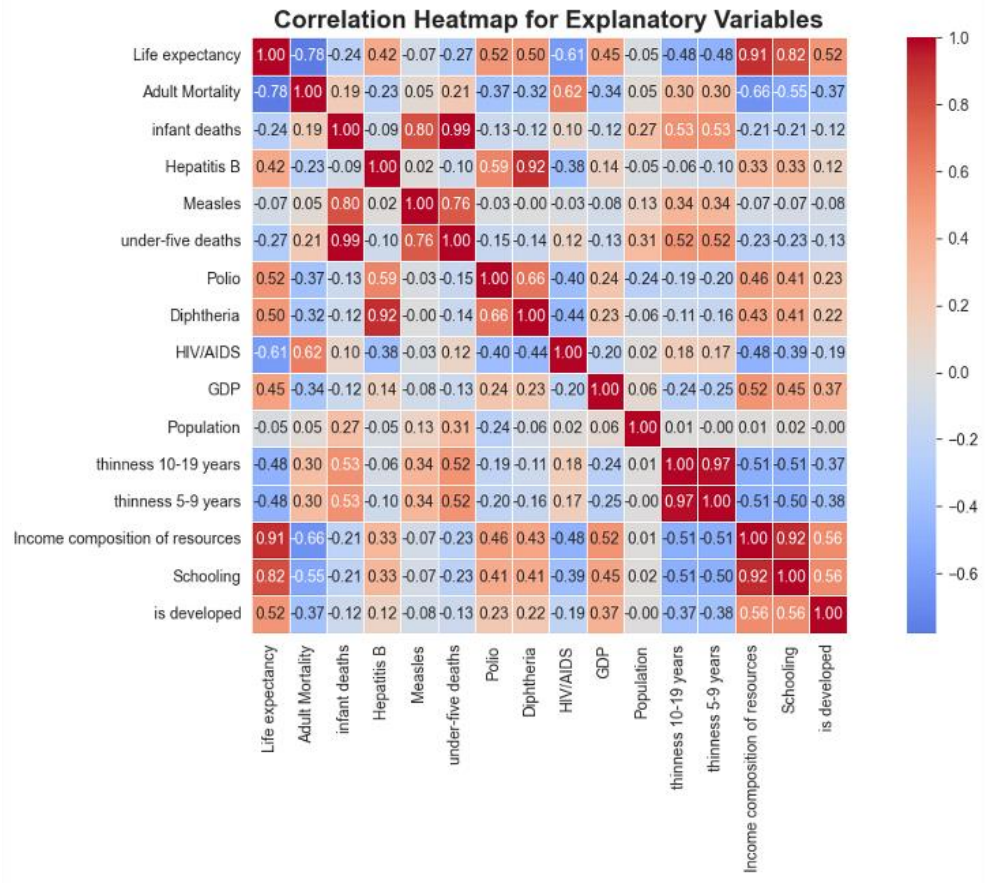
OLS Regression Results

Dep. Variable:	Life expectancy	R-squared:	0.758
Model:	OLS	Adj. R-squared:	0.744
Method:	Least Squares	F-statistic:	53.74
Date:	Wed, 11 Dec 2024	Prob (F-statistic):	1.45e-38
Time:	00:56:46	Log-Likelihood:	-407.16
No. Observations:	146	AIC:	832.3
Df Residuals:	137	BIC:	859.2
Df Model:	8		
Covariance Type:	nonrobust		

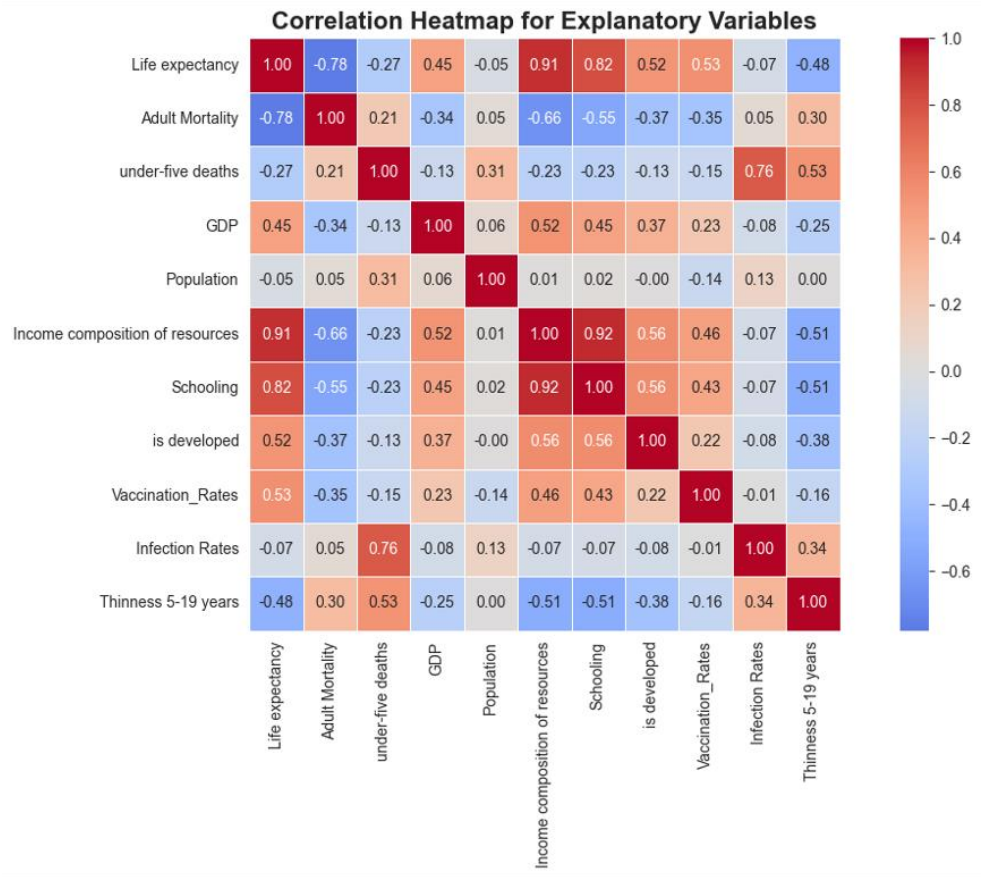
	coef	std err	t	P>|t|	[0.025	0.975]
const	69.7023	1.858	37.508	0.000	66.028	73.377
Adult Mortality	-42.8410	4.636	-9.241	0.000	-52.009	-33.673
Diphtheria	64.4902	23.291	2.769	0.006	18.435	110.546
HIV/AIDS	-640.9615	348.405	-1.840	0.068	-1329.908	47.985
Measles	0.0159	0.093	0.170	0.865	-0.169	0.200
Polio	36.1773	22.713	1.593	0.114	-8.737	81.092
infant deaths	46.9715	46.087	1.019	0.310	-44.162	138.105
is developed	4.7322	0.968	4.890	0.000	2.819	6.646
under-five deaths	-42.1356	32.992	-1.277	0.204	-107.375	23.104

Omnibus:	22.638	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37.178
Skew:	-0.766	Prob(JB):	8.45e-09
Kurtosis:	4.940	Cond. No.	8.22e+03

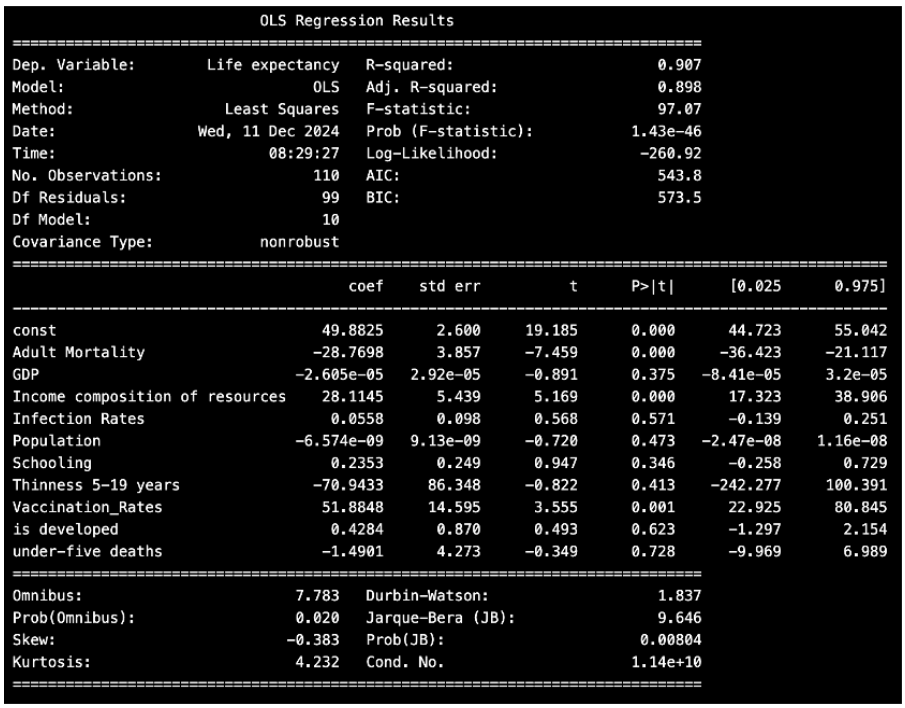
- We planned on performing some dimension reduction. Here were some things we noted:
  - Some higher correlations
    - under-five deaths and infant deaths (0.99)
    - Diptheria and hepatitis B (0.92)
    - measles and infant deaths (0.8)
    - schooling and income composition of resources (0.92)
    - thinness 5-9 and thinness 10-19 (0.97)
  - Remove infant deaths



-
  - Then we looked into combining variables of the dataset.
    - Vaccine Variables - Hepatitis B,Diphtheria,Polio
    - Infection Variables – HIV/AIDS, Measles
    - Thinness 5-19 – Thinness 5-9, Thinness 10-19
    - Drop the above columns in place for their averaged values.
    - The new heatmap has lower correlations between variables.



- Now, we wanted to assess the same multiple linear regression but with the reduced variables.
  - Remove the rows with NA values to train the model (41 countries out of 183)
  - R^2 of 0.907
  - High p-values of GDP, Infection Rates, Population, Schooling, Thinness 5-19, is developed, under-five deaths

- 

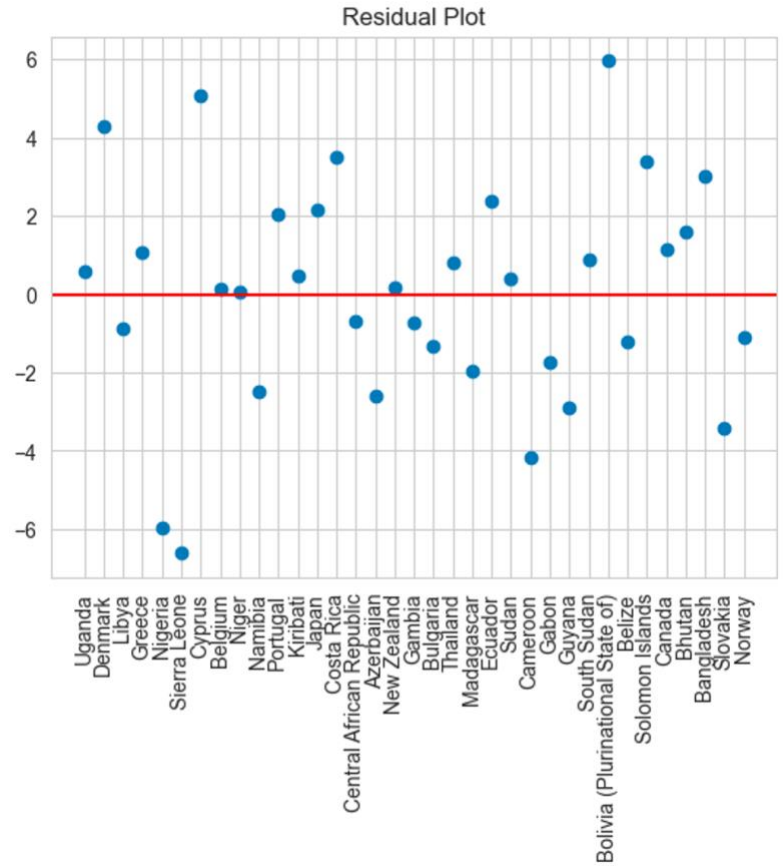
OLS Regression Results

Dep. Variable:	Life expectancy	R-squared:	0.907
Model:	OLS	Adj. R-squared:	0.898
Method:	Least Squares	F-statistic:	97.07
Date:	Wed, 11 Dec 2024	Prob (F-statistic):	1.43e-46
Time:	08:29:27	Log-Likelihood:	-260.92
No. Observations:	110	AIC:	543.8
Df Residuals:	99	BIC:	573.5
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	49.8825	2.600	19.185	0.000	44.723	55.042
Adult Mortality	-28.7698	3.857	-7.459	0.000	-36.423	-21.117
GDP	-2.605e-05	2.92e-05	-0.891	0.375	-8.41e-05	3.2e-05
Income composition of resources	28.1145	5.439	5.169	0.000	17.323	38.906
Infection Rates	0.0558	0.098	0.568	0.571	-0.139	0.251
Population	-6.574e-09	9.13e-09	-0.720	0.473	-2.47e-08	1.16e-08
Schooling	0.2353	0.249	0.947	0.346	-0.258	0.729
Thinness 5-19 years	-70.9433	86.348	-0.822	0.413	-242.277	100.391
Vaccination_Rates	51.8848	14.595	3.555	0.001	22.925	80.845
is developed	0.4284	0.870	0.493	0.623	-1.297	2.154
under-five deaths	-1.4901	4.273	-0.349	0.728	-9.969	6.989

Omnibus:	7.783	Durbin-Watson:	1.837
Prob(Omnibus):	0.020	Jarque-Bera (JB):	9.646
Skew:	-0.383	Prob(JB):	0.00804
Kurtosis:	4.232	Cond. No.	1.14e+10

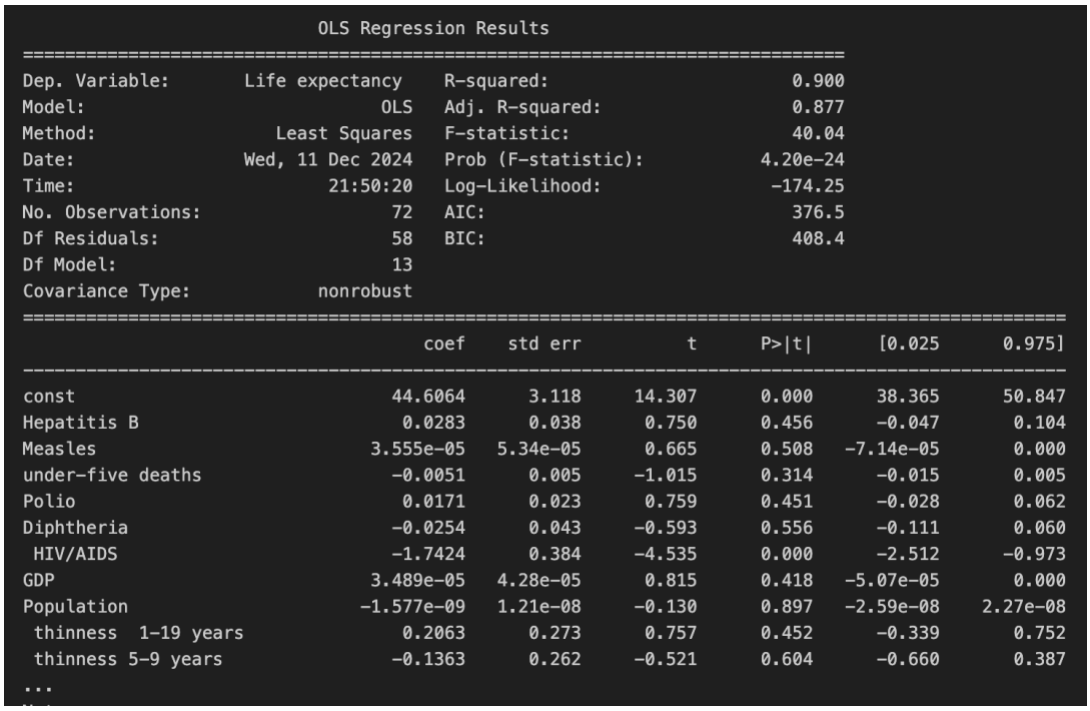
- Finally, we wanted to further simplify the model to only utilize the variables that had statistically significant predictability for the life expectancy.
  - R^2 of 0.869
  - Three variables – adult mortality, income composition of resources, vaccination rates
  - This was our chosen final model.
  - Final Equation: Life Expectancy = 47 – 20.1265 (Adult Mortality) + 34 (Income Composition of Resources) + 47 (Vaccination Rates)
- This model performed really well for what we had hoped but there are some key things to note.
  - There are some high residuals, but generally the values are well predicted.



-
    - Also, the standardized values would mean that you would need to standardize all future collected values the same way to hopefully find meaning. Also, this may not be true for future unstandardized values.
- Therefore, considering this model, we also wanted to look at the actual values.

We looked at the 2015 data unstandardized and found some of the following results.

- We initially assessed all of the variables from the data to see if there were any particular variables to focus on:



OLS Regression Results

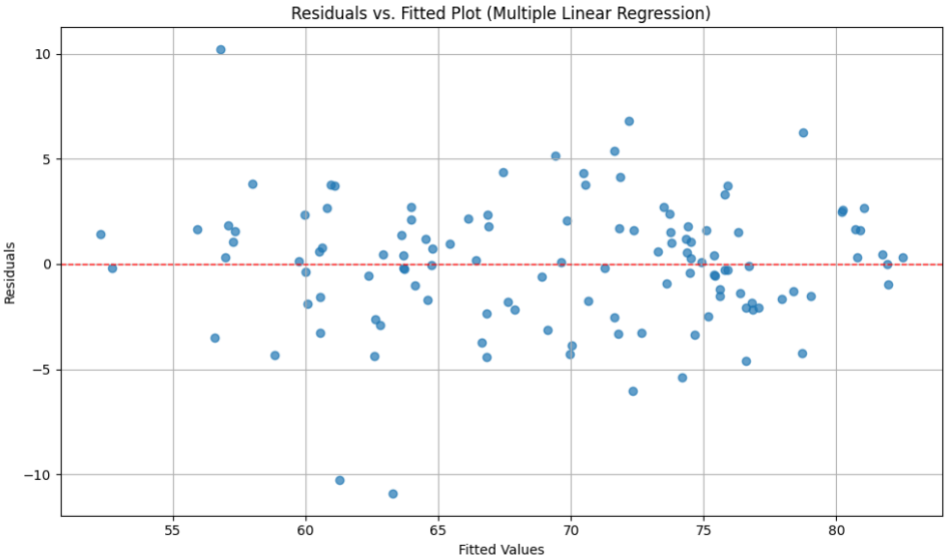
Dep. Variable:	Life expectancy	R-squared:	0.900
Model:	OLS	Adj. R-squared:	0.877
Method:	Least Squares	F-statistic:	40.04
Date:	Wed, 11 Dec 2024	Prob (F-statistic):	4.20e-24
Time:	21:50:20	Log-Likelihood:	-174.25
No. Observations:	72	AIC:	376.5
Df Residuals:	58	BIC:	408.4
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	44.6064	3.118	14.307	0.000	38.365	50.847
Hepatitis B	0.0283	0.038	0.750	0.456	-0.047	0.104
Measles	3.555e-05	5.34e-05	0.665	0.508	-7.14e-05	0.000
under-five deaths	-0.0051	0.005	-1.015	0.314	-0.015	0.005
Polio	0.0171	0.023	0.759	0.451	-0.028	0.062
Diphtheria	-0.0254	0.043	-0.593	0.556	-0.111	0.060
HIV/AIDS	-1.7424	0.384	-4.535	0.000	-2.512	-0.973
GDP	3.489e-05	4.28e-05	0.815	0.418	-5.07e-05	0.000
Population	-1.577e-09	1.21e-08	-0.130	0.897	-2.59e-08	2.27e-08
thinness 1-19 years	0.2063	0.273	0.757	0.452	-0.339	0.752
thinness 5-9 years	-0.1363	0.262	-0.521	0.604	-0.660	0.387

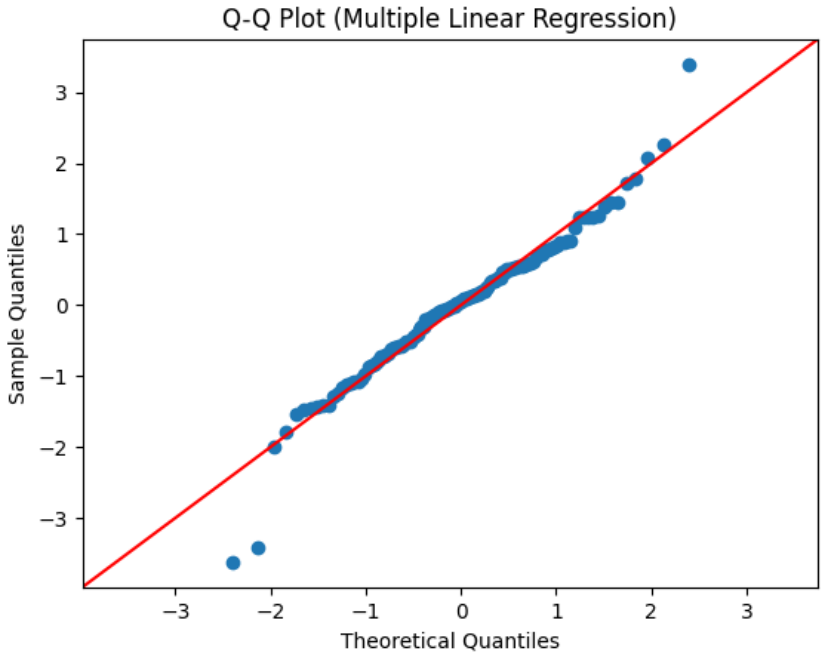
...

- These regression results gave two statistically significant variables
  - HIV/AIDS
  - Income composition of resources
- Also, there was one other variable that was closer than others that we decided to keep to assess for the final model (under-five deaths).

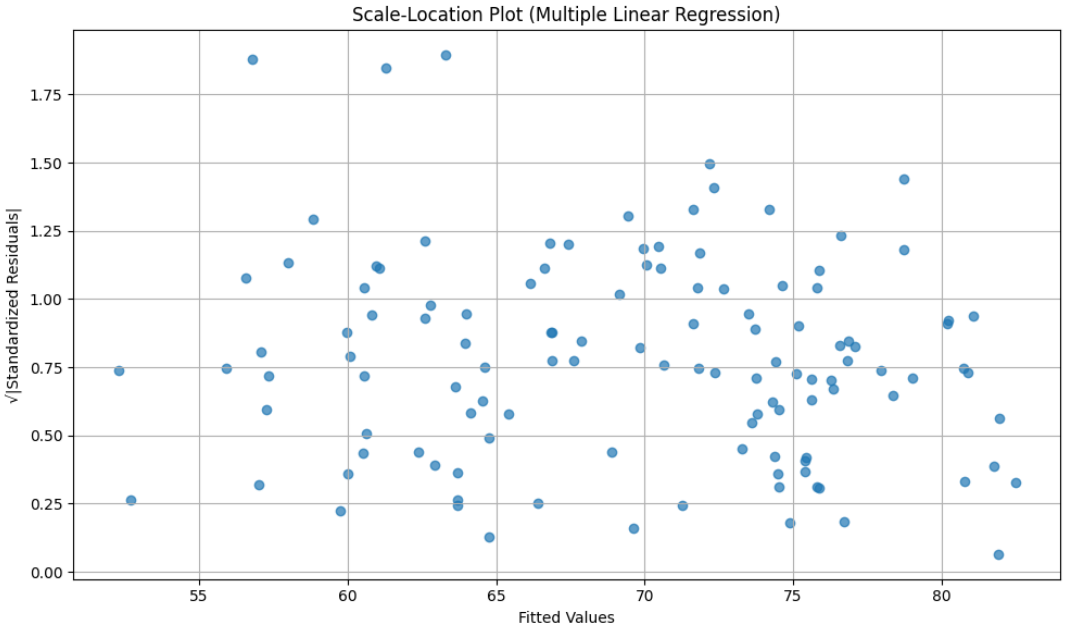
- First, we computed the four plots to assess the assumptions of linearity, normality, homoscedasticity, and leverage.



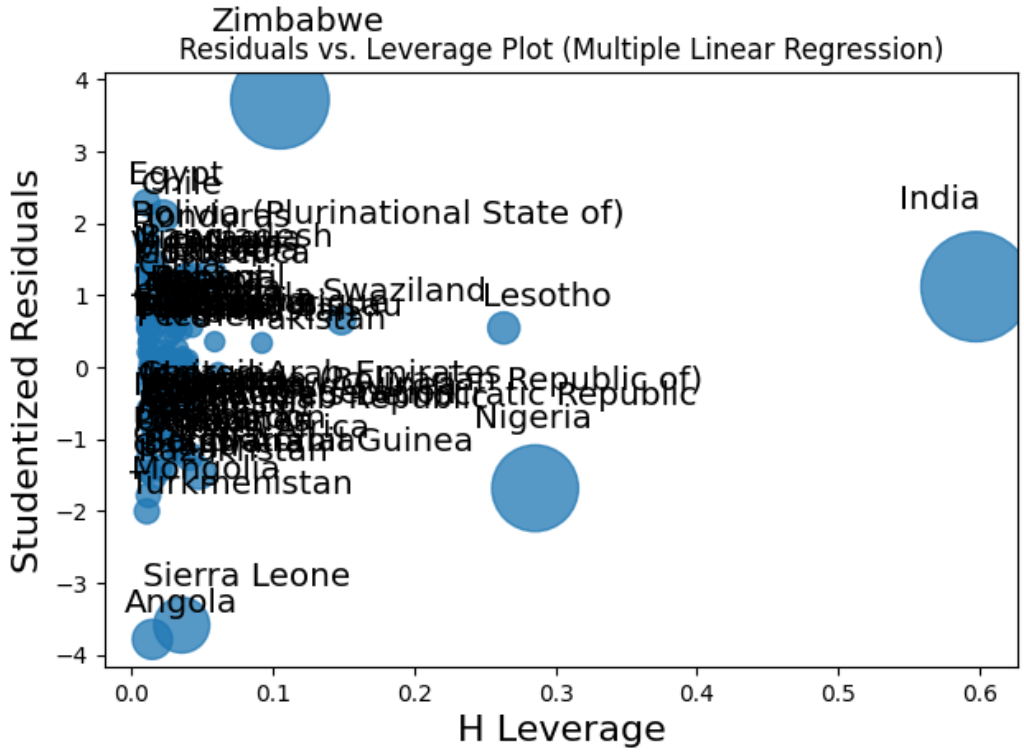
- The residuals vs. fitted plot allowed us to assess the potential linearity. Generally the values appear to be equally distributed about the line and follow it. Therefore, we can accept the assumption of linearity.



- Then, we assessed the Q-Q plot and found that it also generally followed the line. Therefore, we were able to accept the assumption of normality for the data.



- Here is the scale location plot. We can see that the residuals appear to be very similarly spread; therefore, we will decide to accept the assumption of Homoscedasticity.



        - The final graph above allows us to assess the leverage of data points. Here, it appears that there are influential points in the graph. Therefore, we will reject the assumption that there are no influential points in the graph.
    - Overall, the model performed well in accounting for the general assumptions of these plots by allowing us to accept Linearity, Normality, and Homoscedasticity. So, let's move forward by showcasing the exact regression results.
- Here are the following results:
    -

OLS Regression Results

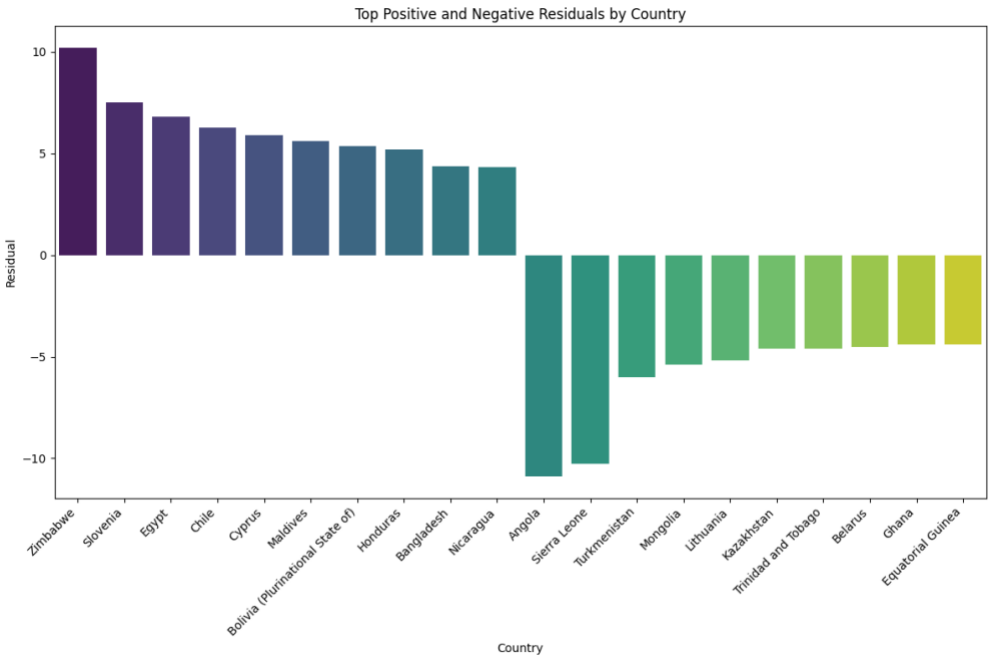
Dep. Variable:	Life expectancy	R-squared:	0.858
Model:	OLS	Adj. R-squared:	0.855
Method:	Least Squares	F-statistic:	234.5
Date:	Wed, 11 Dec 2024	Prob (F-statistic):	4.52e-49
Time:	22:08:21	Log-Likelihood:	-302.42
No. Observations:	120	AIC:	612.8
Df Residuals:	116	BIC:	624.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	44.4743	1.505	29.547	0.000	41.493	47.456
under-five deaths	-0.0028	0.002	-1.268	0.207	-0.007	0.002
HIV/AIDS	-1.3264	0.198	-6.706	0.000	-1.718	-0.935
Income composition of resources	40.7082	2.116	19.239	0.000	36.517	44.899

Omnibus:	11.579	Durbin-Watson:	1.981
Prob(Omnibus):	0.003	Jarque-Bera (JB):	23.605
Skew:	-0.322	Prob(JB):	7.49e-06
Kurtosis:	5.075	Cond. No.	1.25e+03

        - These results showcase an R^2 of 0.858.
        - Furthermore, the variables that were used did a very good job in predicting life expectancy, and they appear to have a large association with the life expectancy of a given country.
        - We could remove under-five deaths since it isn't statistically significant, but we decided to keep it in the equation due to leaving the model with only two variables.

- We assessed the residuals of this graph and here are the largest residuals positive and negative to help us understand what countries have a smaller chance of predicting life expectancy correctly.



Furthermore, we attempted to look at the whole data not just focusing on one year to see what the best predictor variables are associated with that. We found:

- OLS Regression Results

Dep. Variable:	Life expectancy	R-squared:	0.799
Model:	OLS	Adj. R-squared:	0.798
Method:	Least Squares	F-statistic:	1386.
Date:	Fri, 06 Dec 2024	Prob (F-statistic):	0.00
Time:	22:12:24	Log-Likelihood:	-7062.5
No. Observations:	2450	AIC:	1.414e+04
Df Residuals:	2442	BIC:	1.419e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.7291	0.510	91.714	0.000	45.730	47.728
Polio	0.0278	0.005	5.298	0.000	0.017	0.038
Diphtheria	0.0312	0.005	5.954	0.000	0.021	0.041
HIV/AIDS	-0.6632	0.017	-39.716	0.000	-0.696	-0.630
GDP	6.132e-05	6.98e-06	8.782	0.000	4.76e-05	7.5e-05
thinness 5-9 years	-0.1487	0.022	-6.778	0.000	-0.192	-0.106
Income composition of resources	10.1603	0.713	14.245	0.000	8.762	11.559
Schooling	1.0629	0.047	22.596	0.000	0.971	1.155

Omnibus:	177.866	Durbin-Watson:	0.445
Prob(Omnibus):	0.000	Jarque-Bera (JB):	913.760
Skew:	0.043	Prob(JB):	3.80e-199
Kurtosis:	5.991	Cond. No.	1.33e+05

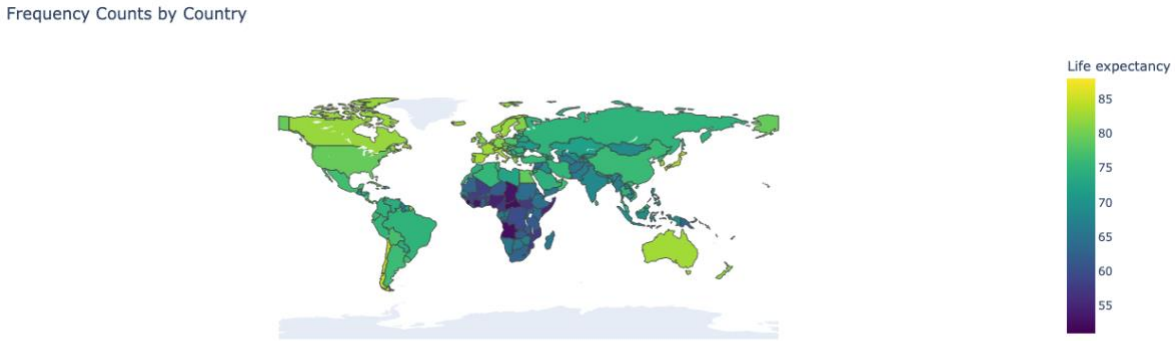
- This has a lot more instances and the predictors above seemed to work pretty accurately with an R^2 value of about 0.8. This also is in conjunction with statistically significant predictors that appear to affect the response variable of life expectancy. These predictors include Polio, Diphtheria, HIV/AIDS, GDP, thinness 5-9 years, Income consumption of resources, and Schooling. Now, since we did this with many different instances looking outside of just one year, it helped the model tremendously. The R^2 lost predictability but there were increased variables that could be used to predict the response variable because they were noted as statistically significant.

Overall, all of the ways that we performed the multiple linear regression are appropriate, but it will depend on what you would want to look at for future predictions whether that be standardized values for 2015, actual values for 2015, or finally the entire dataset including the years from 2000-2015 to see if we can make a more general trend prediction of what was most important to life expectancy over many years.

Now, let's begin to look at the ANOVA of different continents and their life expectancies.

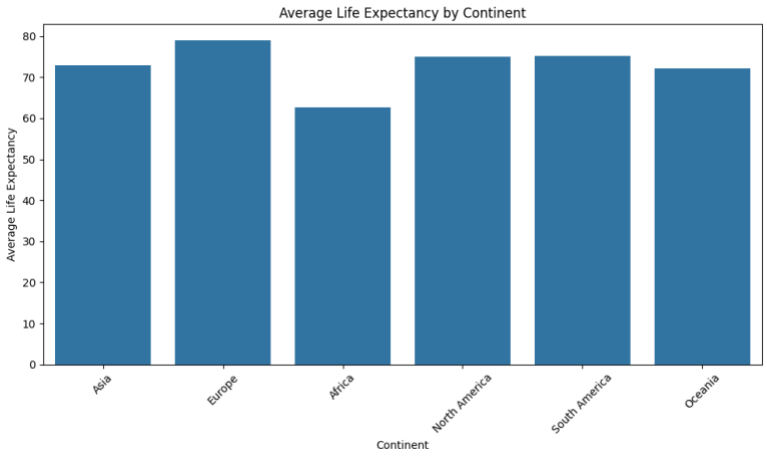
ANOVA:

- We first split all of the countries into their respective continents. Then, we performed an ANOVA model on the Life Expectancy values of the different continents to see if they were statistically significant.
- Here is a heatmap for the life expectancy of different countries across the continents:



-
    - We can see that the higher life expectancy appears to be in North America, Europe, and Australia.
- Here is the ANOVA analysis of the life expectancy values.
    -

	sum_sq	df	F	PR(>F)
C(Continent)	7066.209785	5.0	50.58717	2.263679e-32
Residual	4944.807701	177.0	NaN	NaN

    - 

    - Continents Averages:
        - Africa: 62.666667
        - Asia: 72.852273
        - Europe: 78.971429
        - North America: 75.014286

    - Oceania: 72.160000
    - South America: 75.225000

- This plot showcases the difference in the averages isn't too drastic, but it is enough to showcase that it is statistically significant in the difference of life expectancy from continent to continent. We could look at the larger dataset again, but I believe it is unnecessary due to how statistically significant the p-value was for the ANOVA in 2015.

Non Parametric Test:

- We performed the Mann-Whitney U Test in order to see the pairwise combinations of each of the continents and their median values for life expectancy respectively against one another.
- Since life expectancy appeared to be normally distributed, we can assess this further.
- Here is the table with the computed values:
    -

Continent 1	Continent 2	p-value
Asia	Europe	1.308203e-06
Asia	Africa	2.411076e-11
Asia	North America	8.070586e-02
Asia	South America	1.621046e-01
Asia	Oceania	6.479647e-01
Europe	Africa	2.345740e-15
Europe	North America	1.237563e-03
Europe	South America	1.365551e-02
Europe	Oceania	1.718215e-03
Africa	North America	1.157660e-08
Africa	South America	2.638222e-06
Africa	Oceania	3.844587e-04
North America	South America	8.515173e-01
North America	Oceania	7.589741e-02
South America	Oceania	7.485800e-02

- The statistical significance was present in all continent relations except Asia, North America; Asia, South America; Asia, Oceania; Europe, South America; North America, South America; North America, Oceania; South America, Oceania
  - This signified there was no difference between the two groups in this population.
- The statistical significance was present in the following continent groups Asia, Europe; Asia, Africa; Asia, Africa; Europe, Africa; Europe, North America; Europe, Oceania; Africa, North America; Africa, South America; Africa, Oceania
  - This signified there was a difference between the two groups in this population.
- Overall, we would still conclude the data to be normally distributed despite some slight skewness from the earlier inspection of the data.

Summary:

- There are a lot of key insights we found from this code. First, we looked at many different ways of displaying the data and cleaning it. It was most important for the data to be available for prediction by removing variables that had high null values.
- We also focused on ways of combining variables, assessing correlations, and even standardizing if necessary.
- Furthermore, we looked at multiple linear regression to understand the relationships between the potential predictor variables and the response variable of life expectancy. We did this for 2015 data and also for the entire dataset.
  - Effective predictors in 2015 Standardized Data: Adult Mortality, Income Composition of Resources, Vaccination Rates
  - Effective predictors in 2015 Actual Data: under-five deaths, HIV/AIDS, and Income Composition of Resources

    - Effective predictors for long-term data: Polio, Diphtheria, HIV/AIDS, GDP, thinness 5-9 years, Income Composition of Resources, Schooling
    - Interestingly, in all these models, population was not a heavy indicator of life expectancy.
- Next, we performed an ANOVA grouping all the countries into continents and then we assessed if there was any difference in life expectancy values for the data of each continent. There were thus statistically significant differences.
- Finally, we assessed some nonparametric techniques to see if this was also showcasing statistical significance.
    - Mann-Whitney U Test: Found statistically significant differences in 9 of the 16 paired combinations of continents.

Interpretation:

- We successfully identified predictor variables for both the year 2015 specifically and the entire dataset that demonstrate strong predictive accuracy for the response variable, life expectancy. These findings highlight the potential to estimate a country's life expectancy based on certain predictor variables.
- Additionally, conducting an ANOVA analysis provided insights into the relationships between the mean life expectancies across different continents in our dataset. The results revealed significant differences, indicating that life expectancy varies by continent. These variations underscore the need for targeted healthcare interventions tailored to the specific needs of each region.

- The initial analysis of the standardized data showcased that adult mortality, income composition of resources, and vaccination rates were the highest contributors to predicting life expectancy. These likely affect life expectancy through an increase or decrease in deaths of adults, the states of income within the population, and the vaccination rates which can contribute to disease rates.
- Our analysis also identified key predictors of life expectancy in 2015 for the actual data, including under-five deaths, income composition of resources, and HIV/AIDS prevalence. These variables likely influence life expectancy through mechanisms such as the adverse impact of healthcare in saving babies that are lost in pregnancy, incomes that affect the taxes and healthcare potential, and HIV/AIDS that attacks the immune system. These insights are critical for organizations like the World Health Organization (WHO) to proactively monitor and address trends that could adversely affect life expectancy.
- When examining the entire dataset, the most significant predictors included Polio, Diphtheria, HIV/AIDS, GDP, thinness in children aged 5-9 years, Income Composition of Resources, and Schooling. These predictors influence life expectancy either positively or negatively, depending on the direction and magnitude of their coefficients. Identifying these relationships provides actionable information for prioritizing interventions in countries facing the most critical challenges.
- Our findings also highlighted disparities across continents, with Africa showing a significantly lower average life expectancy (approximately 63 years) compared to other continents, which range between 72 and 80 years. This stark contrast underscores the urgency for increased focus and intervention in African countries to address the underlying factors contributing to lower life expectancy. By leveraging insights from the

identified predictors, healthcare organizations can implement targeted strategies to mitigate negative trends and improve health outcomes.

- Finally, a non-parametric test revealed that life expectancy values for some continents were not significantly different from one another, suggesting similarities in the underlying relationships influencing life expectancy. This finding has practical implications for designing streamlined interventions that could be applied effectively across multiple regions, enabling organizations such as the WHO to enhance global health outcomes and increase life expectancy more broadly.

References

Kim, J. H. (2019, July 15). *Multicollinearity and misleading statistical results*. Korean journal of anesthesiology. https://pmc.ncbi.nlm.nih.gov/articles/PMC6900425/#:~:text=Condition%20Number%20and%20Condition%20Index&text=The%20largest%20condition%20index%20is,multicollinearity%20is%20regarded%20as%20strong

Life expectancy (WHO). Kaggle. (2018, February 10). https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who