Jae Kyung Jung

Qualitative Variables: Status, Country, Year

Quantitative Variables: Life Expectancy, Adult Mortality, Schooling, GDP

About the Dataset: This dataset is from the World Health Organization data. The data allows for an exploratory analysis on the effects of immunizations, and other factors that may or may not affect life expectancy. The GHO (Global Health Observatory) under WHO is relying on accurate measurement of data that has been kept track of within each population of countries worldwide constantly. Data consists of 2000-2015 looking over a 15 year period of record kept of the population. The dataset contains missing values but nothing substantial to allow us to infer it to be inaccurate.

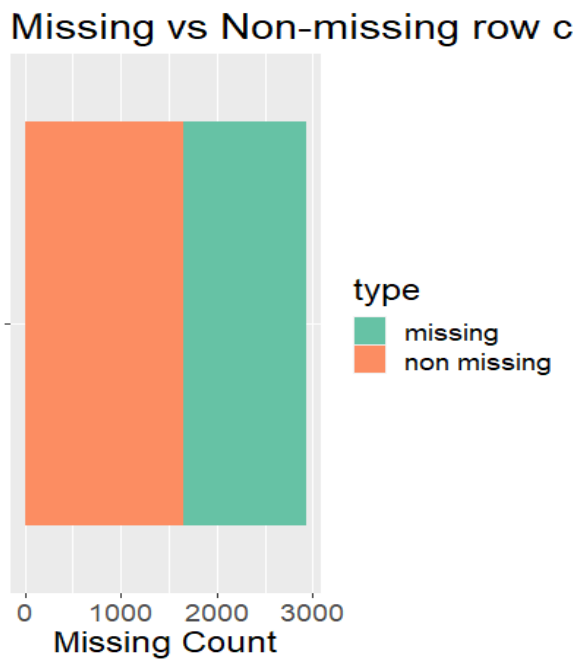
Observations: 2938

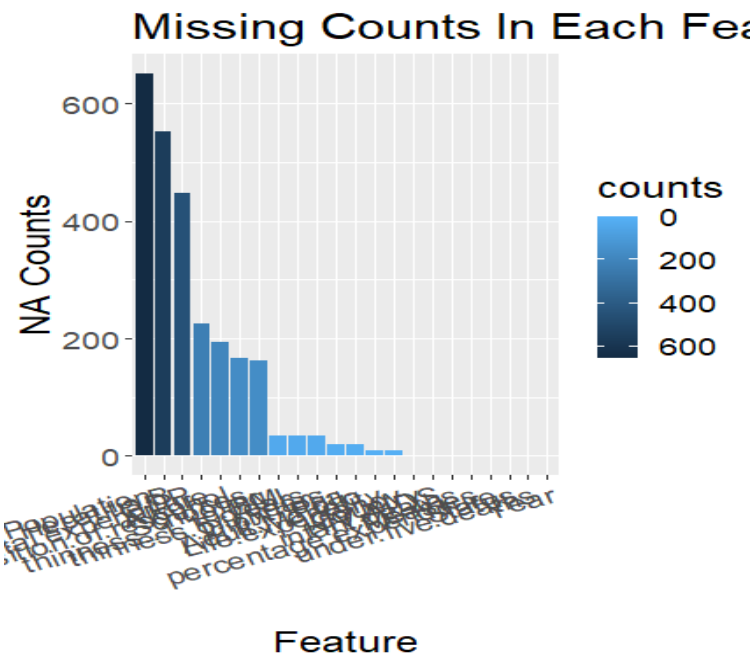
Variables: 22

Intro: This analysis will look over the life expectancy dataset to analyze what is significantly affecting the global life expectancy in humans. We will go in steps in order to find something substantial that could be correlated with life expectancy. Lastly, we will create a regression model in order to fit properly for the dependent variable Life Expectancy. We will see if the model will be able to accurately predict using the predictors imputed into the statistical regression model.

Cleaning the Data:

Since there are noticeable missing values in this dataset the first step would be to find how much of it is missing. We found that 43.87% are missing. The visual of the count is below and a visual count of each observation missing in each variable. With the substantial amount missing we need to apply data imputation in order to fill in the missing observations.





Used a box plot method in order to differentiate the high outliers vs low outliers.

(the graph was not useable for this document due exporting errors)

The 3 low outliers were Alcohol_mean BMI_mean and Income.composition.of.resources_mean

The rest were high outliers and medians were calculated instead.

Now that we have the mean and median for the variables we can apply imputation and fill in the missing observations.

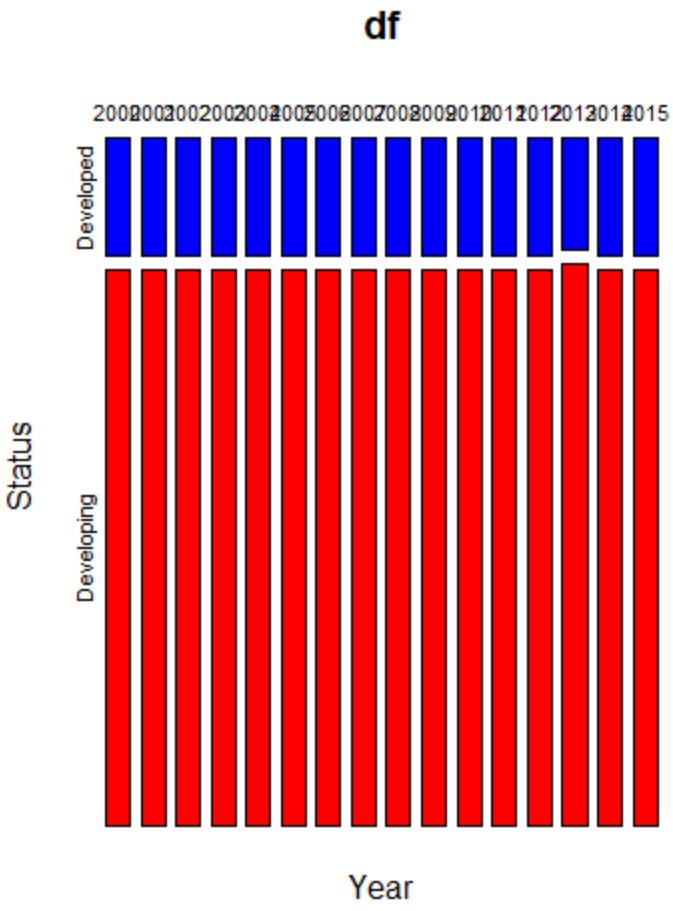
We will also change the variable status into a factor.

Exploratory Association Analysis

I. YEAR VS STATUS

Mosaic Plot

I used a mosaic plot to see any major discrepancy of the ratio between developed vs developing over the Years. From our visualization of the two variables, it is clear that no major changes has occurred in comparison to developed vs developing.]



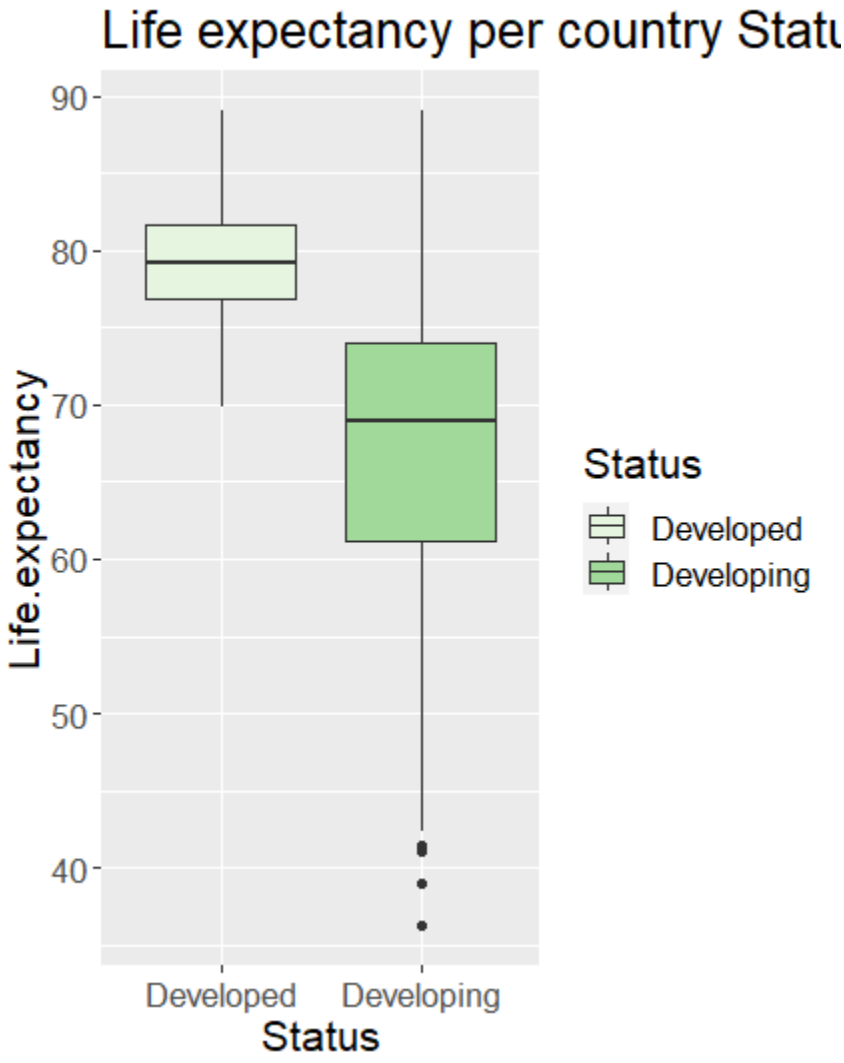
There seems to be no significant change within the year of 2000-2015. No strong associations. Next step, we used a chi square test in order to see the correlation between Status and Year. As we can see from a table, there are 513 developed vs 2426 developing countries. There are almost 5 times as many countries that are considered developing vs 513 countries that are developed. This could be mainly due to the categories WHO and the GHO put strictly in what could be considered developing.

From our Pearson's Chi Squared Test we are able to see that the p value was just a 1 and our chi squared value was .10287. Just from the p value it is safe to say that these two variables Status and Year do not have any significance and therefore no relationship as we can safely accept the null hypothesis.

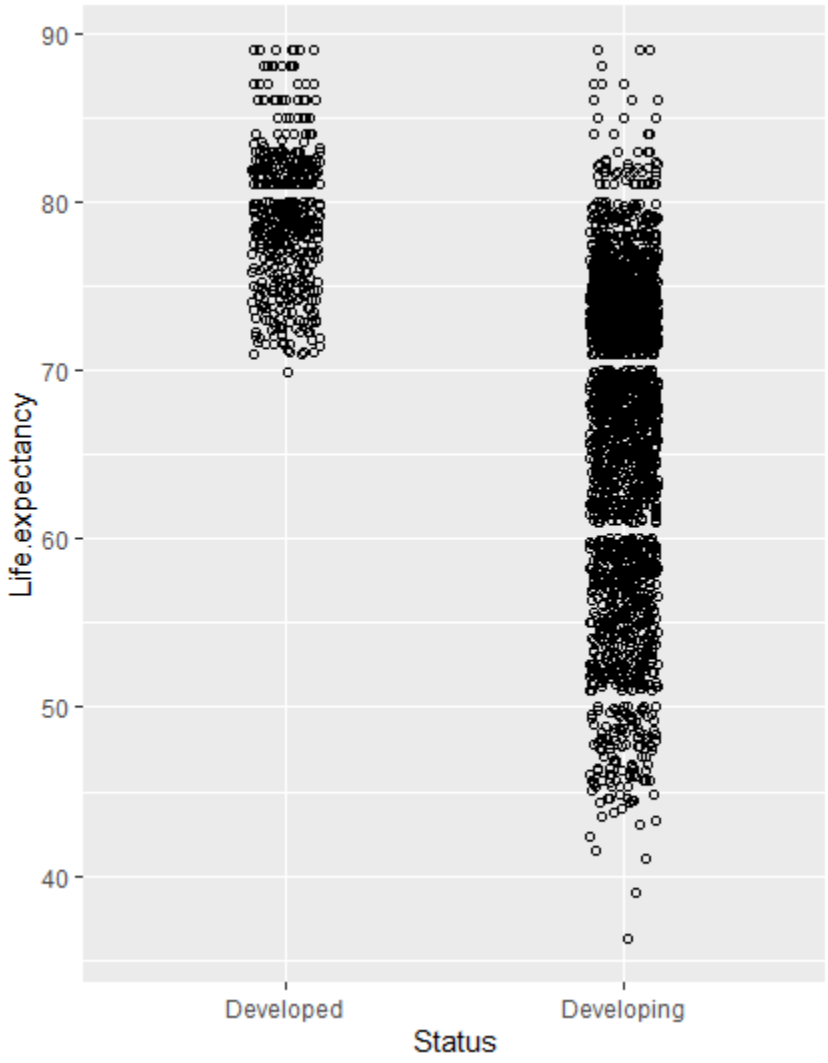
II. STATUS VS LIFE EXPECTANCY

Side by Side Boxplot

We will analyze the Life expectancy variable in relation to status. Below is the result of the boxplot.



From what we can see, it is clear that developed countries have higher life expectancies in comparison to developing ones. To get a better idea and see some of the outlines that could potentially be factoring in as well as the density of the data, we used a Tukey test plot.



We are able to see a clear picture of the nature of the data when comparing developed vs developing when it comes to life expectancy. We also have have the 95% confidence interval.

	diff	lwr	upr	p adj
Developing-Developed	-12.08639	-12.88249	-11.29028	0

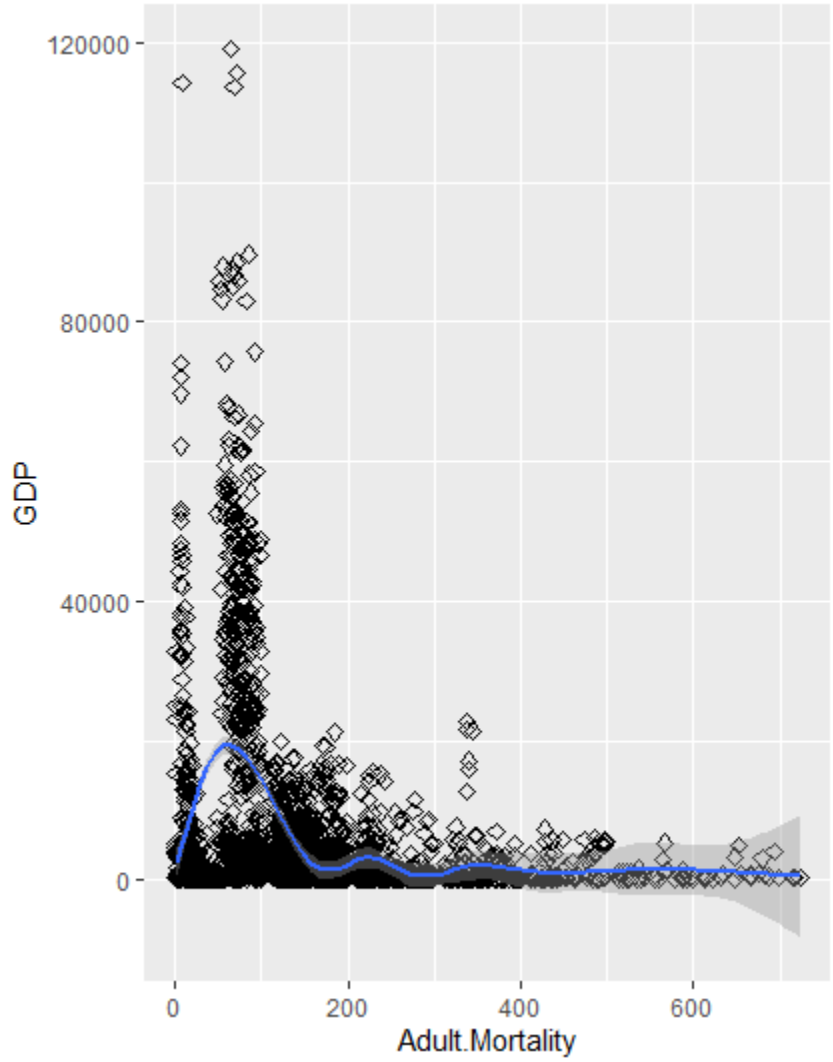
The anova to analyze the significance was also done with results below. The p value indicates significance.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Status	1	61715	61715	886.2	<2e-16
Residuals	2926	203776	70		

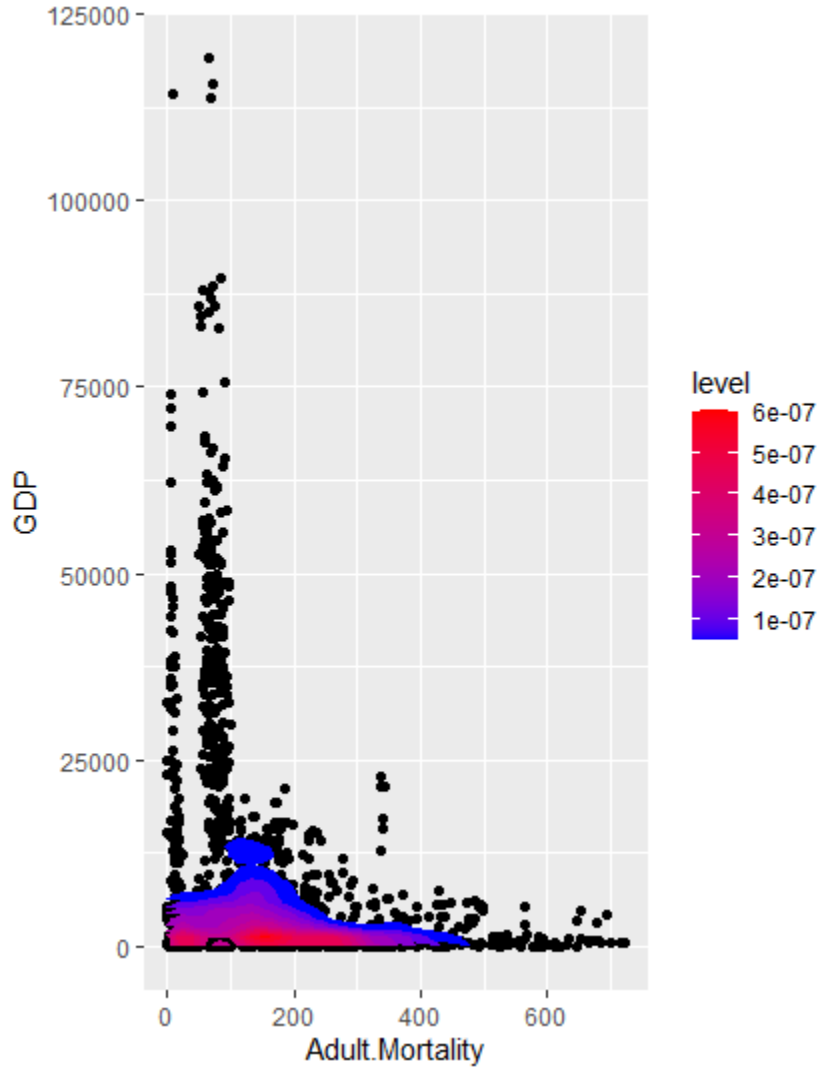
III. Adult Mortality VS GDP

Scatter Plot

We will now analyze the adult mortality in correlation to the GDP of the country. Using a scatter plot and line smoothing, we are able to see the nature of the two variables together.



We also used a density plot in order to see where the observations are clustered into and have a better sense of where the points are gravitating more towards.

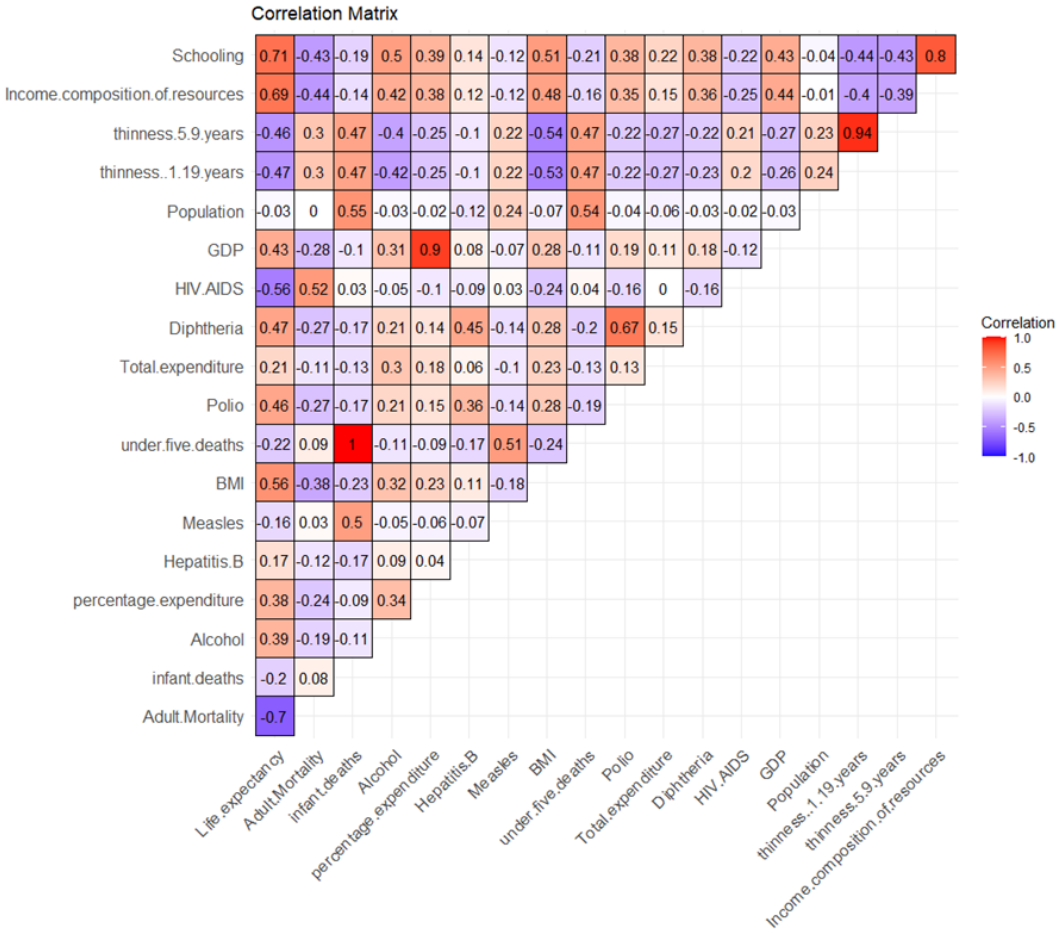


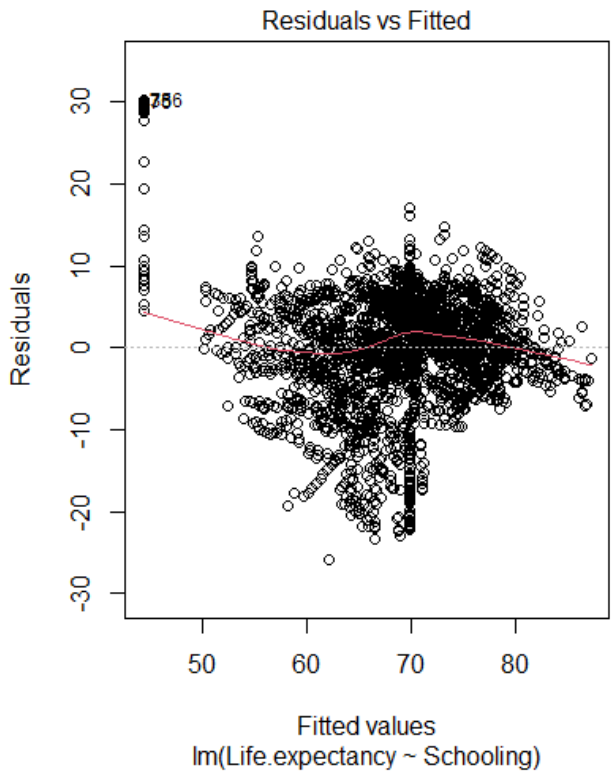
We can see that GDP affects adult mortality up to a certain point. With the correlation coefficient of -0.2814 we can see that this coefficient is below what could be considered having a significant correlation. Due to this point, we can say that GDP and Adult Mortality has no significant relationship. However, I only am scratching the surface and could probably clean and restructure the data frame further.

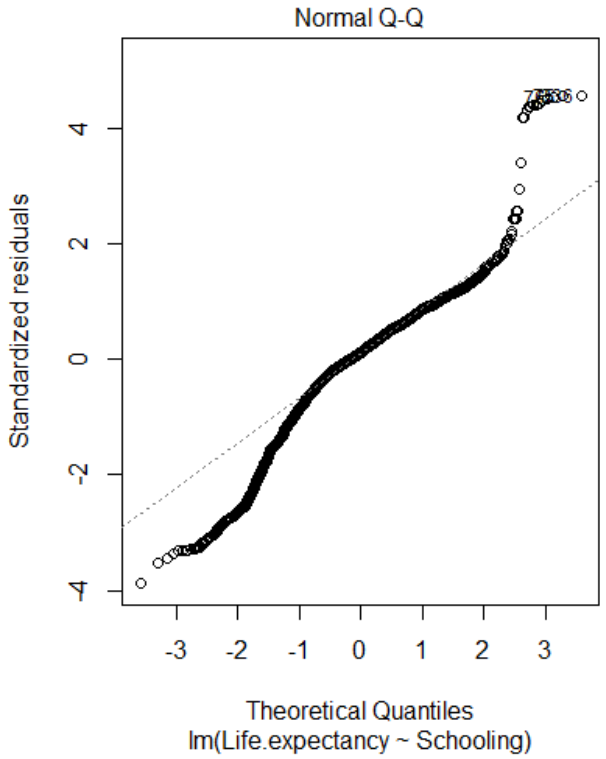
IV. Linear and Multilinear Regression Modeling

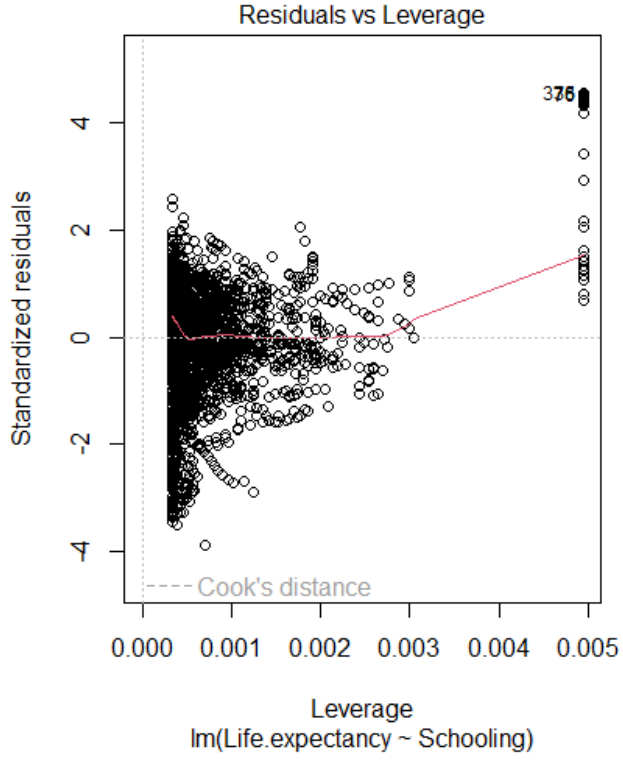
CORRELATION MATRIX

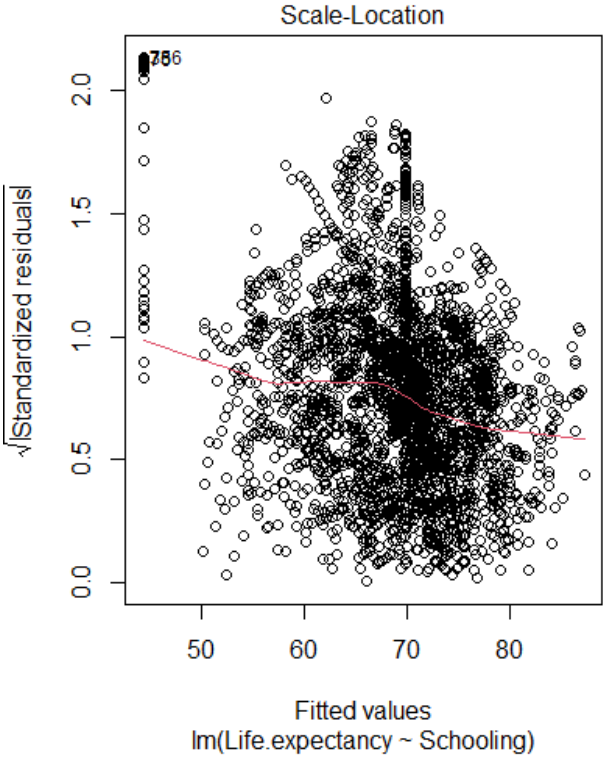
Created a correlation matrix in order to easily find interesting relationships between each variable. Selecting ones that might have significance. Because I am focusing on life expectancy, I chose to select schooling in relation life expectancy for my regression modeling.



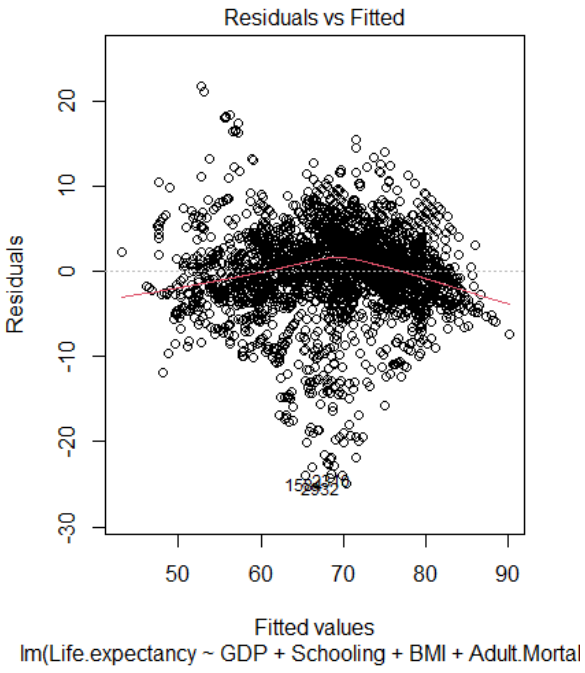

Residuals vs Fitted
Residuals
Fitted values
lm(Life.expectancy ~ Schooling)

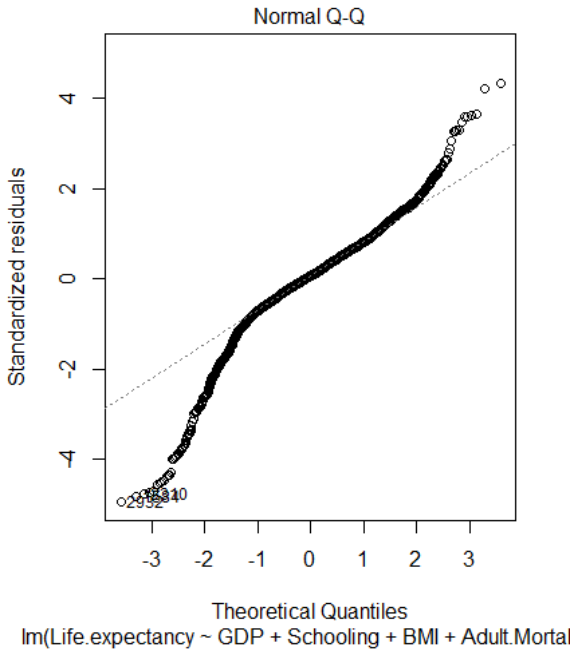

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Life.expectancy ~ Schooling)

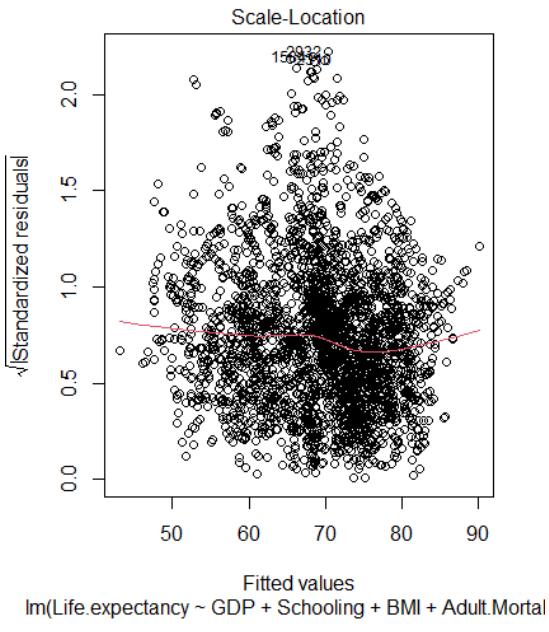

Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(Life.expectancy ~ Schooling)

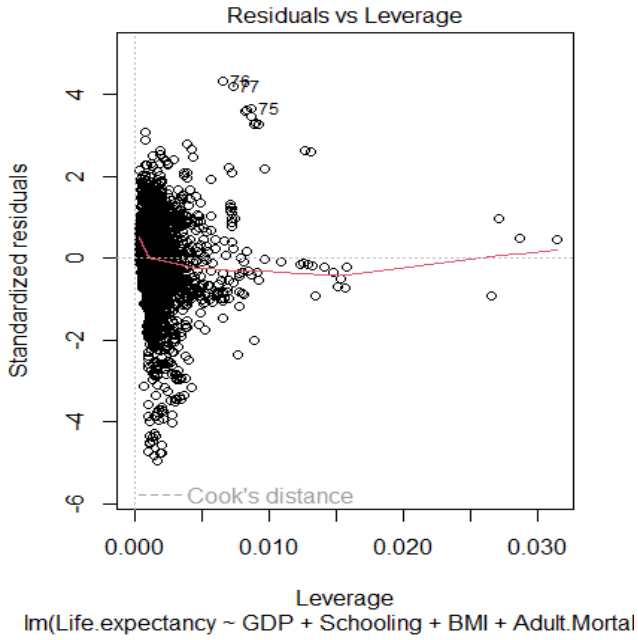

Scale-Location
√|Standardized residuals|
Fitted values
lm(Life.expectancy ~ Schooling)

The p value under .05 indicates significance. We can also see from our estimated regression coefficient that the positive indicates a direct positive relationship. We are also able to observe the Adjusted R squared being .5083 below what we would like the model to be. Therefore we tried a multilinear regression modeling with the variables GDP, schooling, BMI, adult mortality. This may be able to give us a better understanding of the variable life expectancy and a more accurate story.









From our multilinear regression model we are able to see that the testing is more accurate than the previous regression model. This is indicated by the R squared value of .749 and a p value that is less than .05. However, we also have to check the significance of each variable which all indicate a p value of less than .05 and therefore they are all ok to use for this model. One thing to indicate is the Sum of Sq value being highly different from each other but, that is probably due to the variance inflation factor in which I had to take into account but did not do in this analysis. Also there are other variables that could be part of the whole model that can make this much more accurate in order to achieve an R squared value of .9 and above.

V. Conclusion

From our analysis we were able explore some aspects about this dataset and find what affects life expectancy given all the other predictors. In the future we are able to play around with the dataset more and proceed to create a more accurate regression model and a better predicting model.