Introduction to Data Analysis in R

Malaika Aggarwal

April 1 2023

Final Project

The World Bank defines "Life Expectancy" as the number of years a person is expected to live. Life Expectancy is measured for countries to check the level of their development and their healthcare system. This variable is used to measure various Indexes like the Human Development Index, Poverty Index etc.

I used this data set to analyse causal relationships between Life Expectancy and various dependent variables like schooling, Aids, Polio, GDP etc. The data set shows a 15-year (2000-2015) trend for 193 countries and islands.

The objective was to study how factors like expenditure, infant mortality rate, and immunization against diseases affect Life Expectancy. This data was collected for the World bank.

I expected to find that a greater schooling (number of years of education), expenditure, GDP increase Life Expectancy. On the contrary, variables like Alcohol intake, infant mortality reduce number of years from a persons life. The data set also includes immunisation programs for various life threatening diseases in these countries.

This report is divided into 4 sections: Raw Data Set, Data Cleaning, Analysis and Code.

The data set contains 2938 observations with 22 variables. In my R code I, first, clean my data set by checking for duplicate values and dropping NA values. I also plot the distribution along with box plots of various variables as can be seen in the next section.

Next, I perform some regressions to check for and understand the casual relationships between the variable. I take Life expectancy as the dependent variable and the other variables as the independent variables.

More health expenditure, immunization, schooling and GDP all show a positive effect on life expectancy, while alcohol consumption and population have a negative effect.

The panel regression with fixed effects shows that the effect of these variables on life expectancy is consistent over time with a R square of 21.2 percent.

The last section explains the code for one of the graphs (graph 4). This shows the relationship between Life Expectancy and GDP in the year 2014. To create this graph, I used the ggplot package, specifying the data, the x and y variables, and the color and size of the points. I also included a title, axis labels, and use the geom smooth function to show the trend in the data.

The link to my git hub repository is: `https://github.com/malaikaaggarwal/DataAnalysisRProject.git`

1 Raw Data-set

The data, I used, represents Life expectancy and the factors that influence it. It is a panel data set representing 193 countries and islands for a period of 15 years. I extracted this data set from kaggle online server.

The main variable of interest is Life Expectancy. The other variables include Adult Mortality, Infant Deaths, Alcohol Intake, Body Mass Index (BMI), Health Expenditure, Schooling, and several other variables that represent immunisation of various diseases. The dataset has 2938 observations with 22 variables including:

- Country- Country name
- Year- 2000-2015
- Life Expectancy- Y variable
- Adult Mortality- the probability of people who have reached age 15 but will die before reaching the age of 60 years
- infant deaths- number of infant deaths
- Alcohol- alcohol intake
- BMI- Body mass index, IV for Exercising, walking, taking care of one's body.
- Expenditure- health expenditure
- Schooling - number of years of education
- The other variables are represent diseases.

The data set had various missing values:

```
> colSums(is.na(Lifedata))
                         Country                             Year                           Status
                               0                                0                                0
                  LifeExpectancy                     infantdeaths                        Alcohol1
                              10                               10                                0
                  expenditureper                        HepatitsB                         Measles
                             194                                0                              553
                             BMI                      Under5deaths                            polio
                               0                               34                                0
                          totexp                Total expenditure                       Diphtheria
                              19                              226                               19
                           HAIDS                              GDP                       Population
                               0                              448                              652
           thinness  1-19 years                thinness 5-9 years Income composition of resources
                              34                               34                              167
                        Schooling
                             163
>
```

So, I dropped the NA values for analysis proposes.

2 Data Cleaning

To clean the data, I used various commands.

- I checked for duplicate rows using the command: duplicatrows <- duplicated(Lifedata)

  Lifedata[duplicatrows,] This showed me that there were no duplicated rows.
- I checked for how the variables are stored in the data set and if they are correctly stored by using str(Lifedata)

- I checked for missing values and decided to remove them from the data set with the commands:

  colSums(is.na(Lifedata))

  Lifedata$_{f}ull < -na.omit(Lifedata)$

  colSums(is.na(Lifedata$_{f}ull)$)

Now, I use Lifedatafull for all the analysis below.

- I decided to make new columns for GDP and Population to use their log values by using the command:

  Lifedata$_{f}ull < -Lifedata_{f}ull| >$

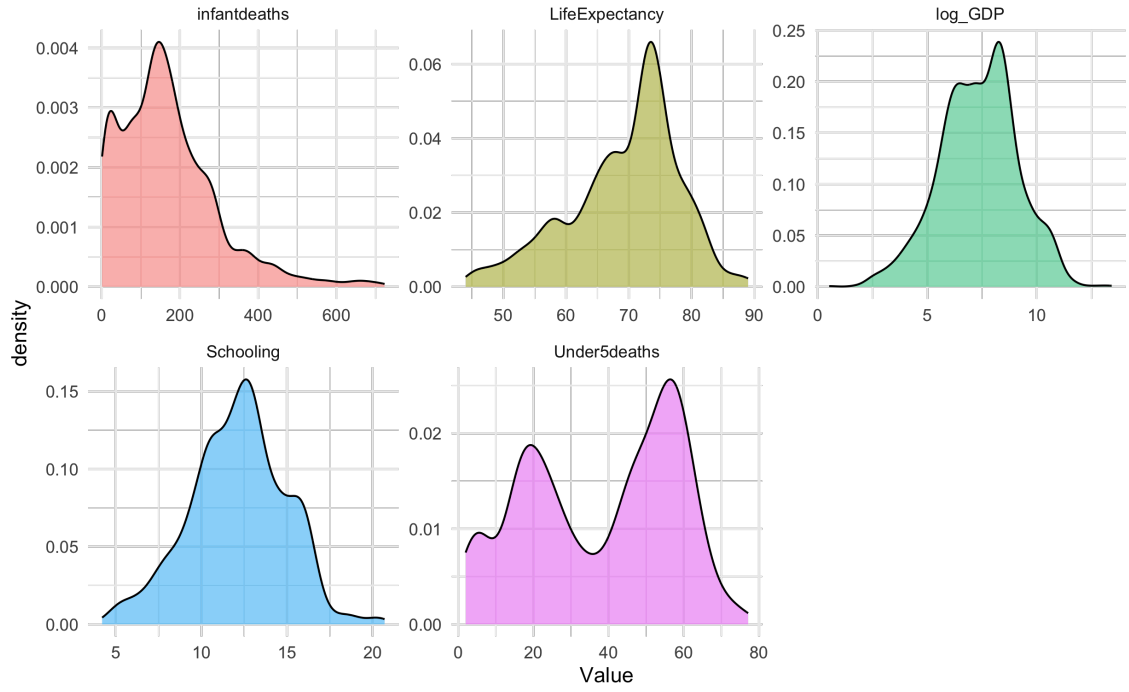
  mutate(log$_{G}DP = log(GDP))| >$

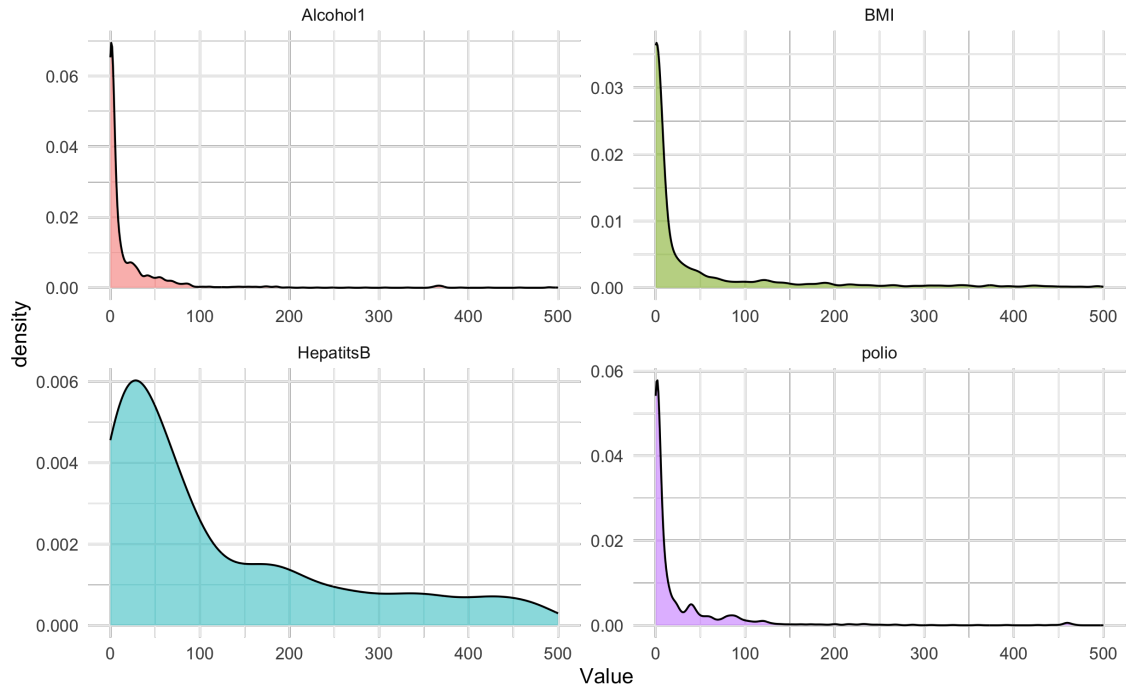
  mutate(log$_{p}op = log(Population)$)

The data set did not have any duplicate rows, so, I left it there. There were multiple missing value sin the variables, So, I dropped all the NA values and was left with 1649 observations. I made columns for log GDP and log population to simplify the analysis and its interpretation.

Overall, these data cleaning techniques ensured that the data set becomes suitable for further analysis.

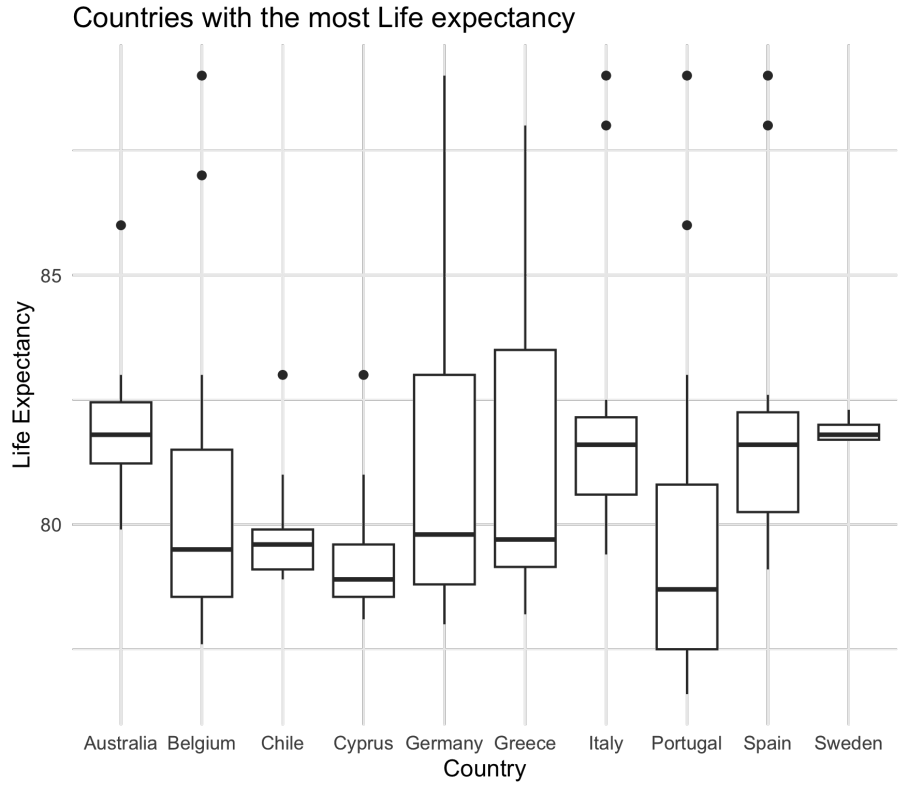
I then plotted the distribution of a few variables:

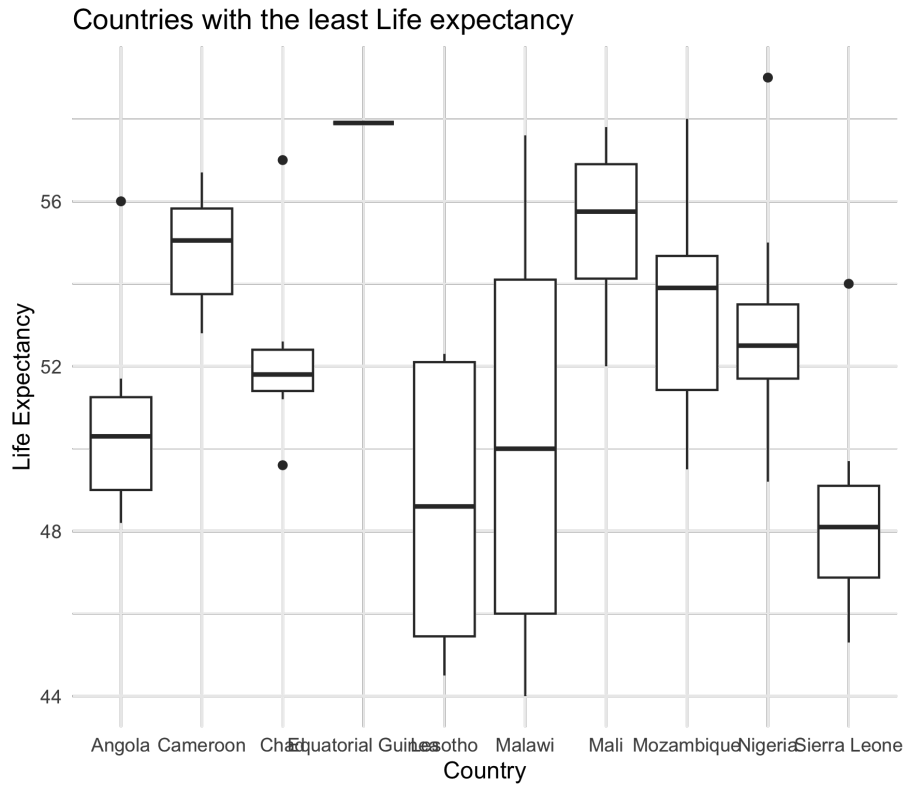




We can see that infant deaths is skewed to the left whereas Life expectancy is skewed more to the right. Log GDP and Schooling have a distribution more closer to normal distribution.

I also included box plots for the 10 countries with the most and the least for the year of 2014 :





The plots suggest that there are some outliers in the data set, for example, Spain has two outliers.

3 Data Analysis

I perform a few regressions for analysis taking Life expectancy as the y variable. I wanted to see how many of the variables can explain the variation in y and if they shared a casual relationship.

The first regression I perform takes Alcohol consumption, expenditure, Measles, Under5 deaths, polio and BMI as the x variables. The results suggest that for more health expenditure, and immunisation life expectancy increases. However, this regression also suggests that for more alcohol consumption Life expectancy also increases.

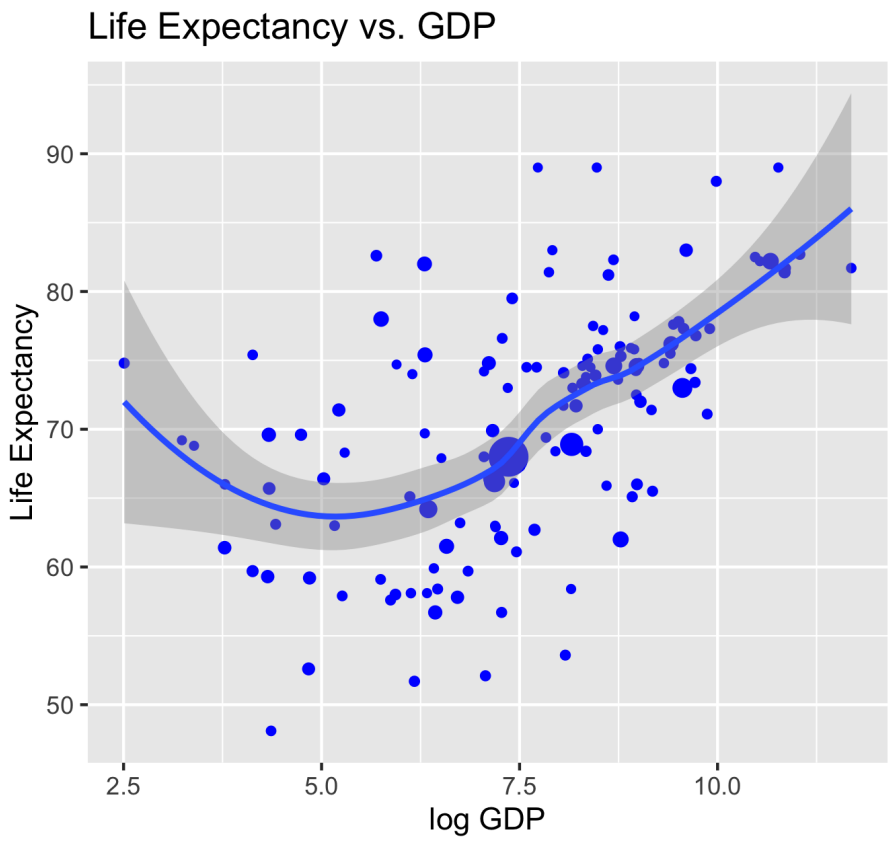
The second regression takes Alcohol, expenditure, log GDP, under 5 deaths, Log population and BMI as the dependent variables. This regression gives results to support my hypothesis. More alcohol consumption decreases Life Expectancy by 0.001 units, More expenditure on health increases Life expectancy, more GDP and BMI increases Life expectancy and more population decreases the y variable.

The third regression takes into account Schooling, population and GDP as x variables. The results suggest that for an additional year of schooling, life expectancy increases by 2 units. Additional population decreases Life expectancy and more GDP increases the y variable.

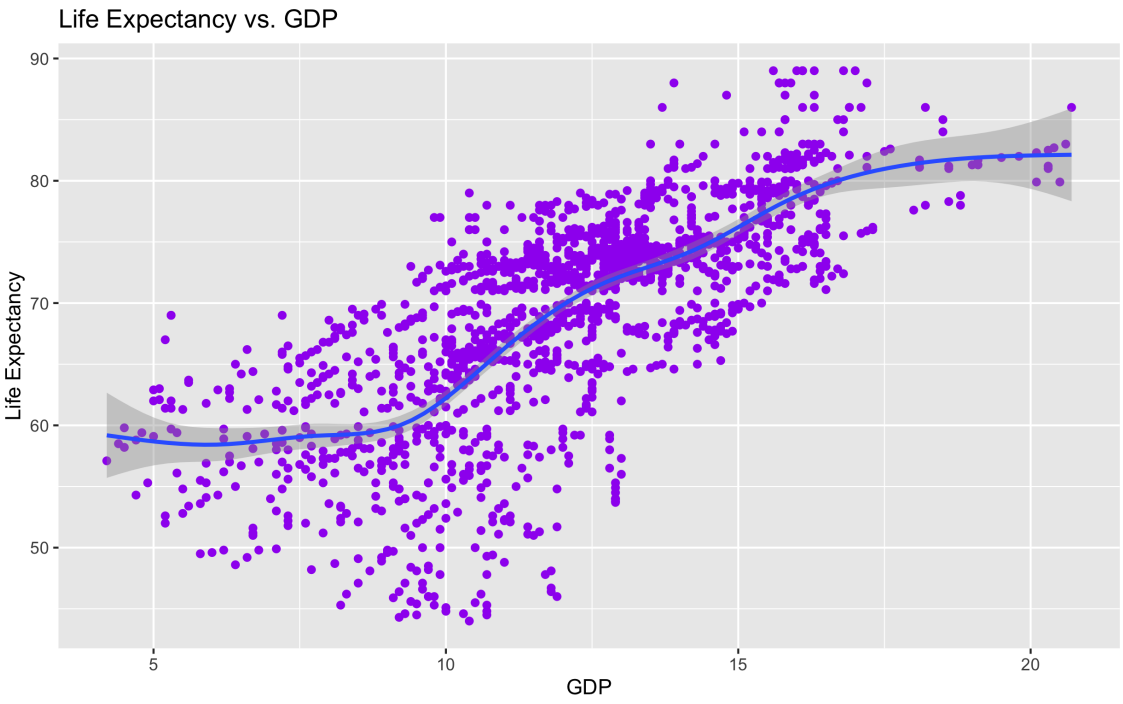
I also perform a panel regression taking into account fixed effects. This regression takes the same variables as the third regression but includes fixed effects for years. The results suggest that all three variables have a positive effect on the dependent variable. The model explains approx. 21.2 percent of variation.

To get visual representation of the relationships, I plot scatter plots.

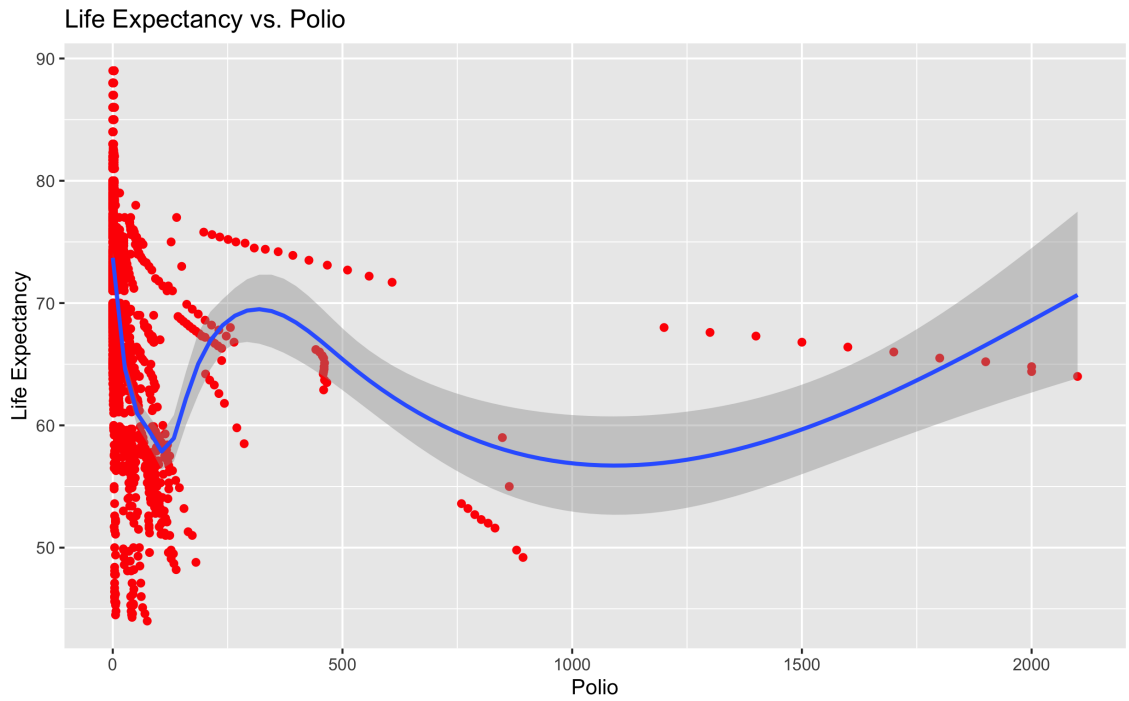
This graph shows the relationship between Life expectancy and GDP in the year 2014. We can see an upward trend suggesting that they share a positive relationship.



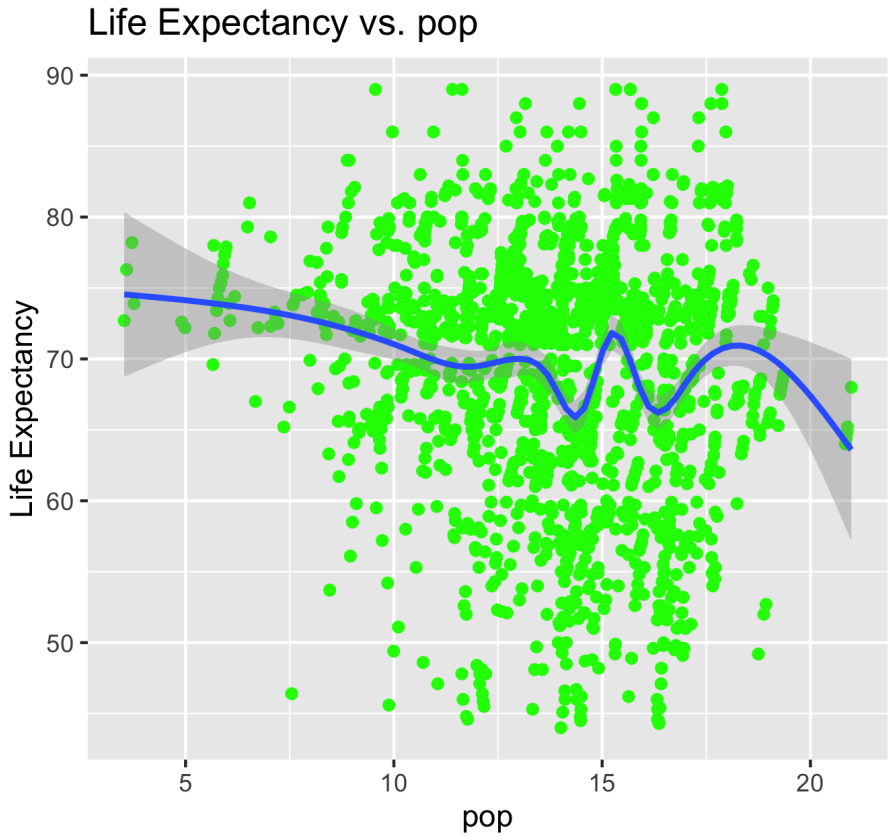
The next graph shows the relationship between Life expectancy and GDP for all years and all countries. We can see an upward trend suggesting that they share a positive relationship.



This graph shows the relationship between Life Expectancy and Polio in all years and all countries. It is hard to tell the relationship between these two variables looking at this graph because a lot of points are clustered around 0. This is because, there are a few outliers so the x axis scale is in a big range, therefore clustering the countries that have lower values.



The next graph shows the relationship between log population and Life expectancy. We see that the points are clustered in the middle, again, making it hard to decipher a relationship between them.



Therefore, We can conclude from the results of the regression models and visualisation that various factors can influence life expectancy, and the relationships between the independent variables and the dependent variable are complex.

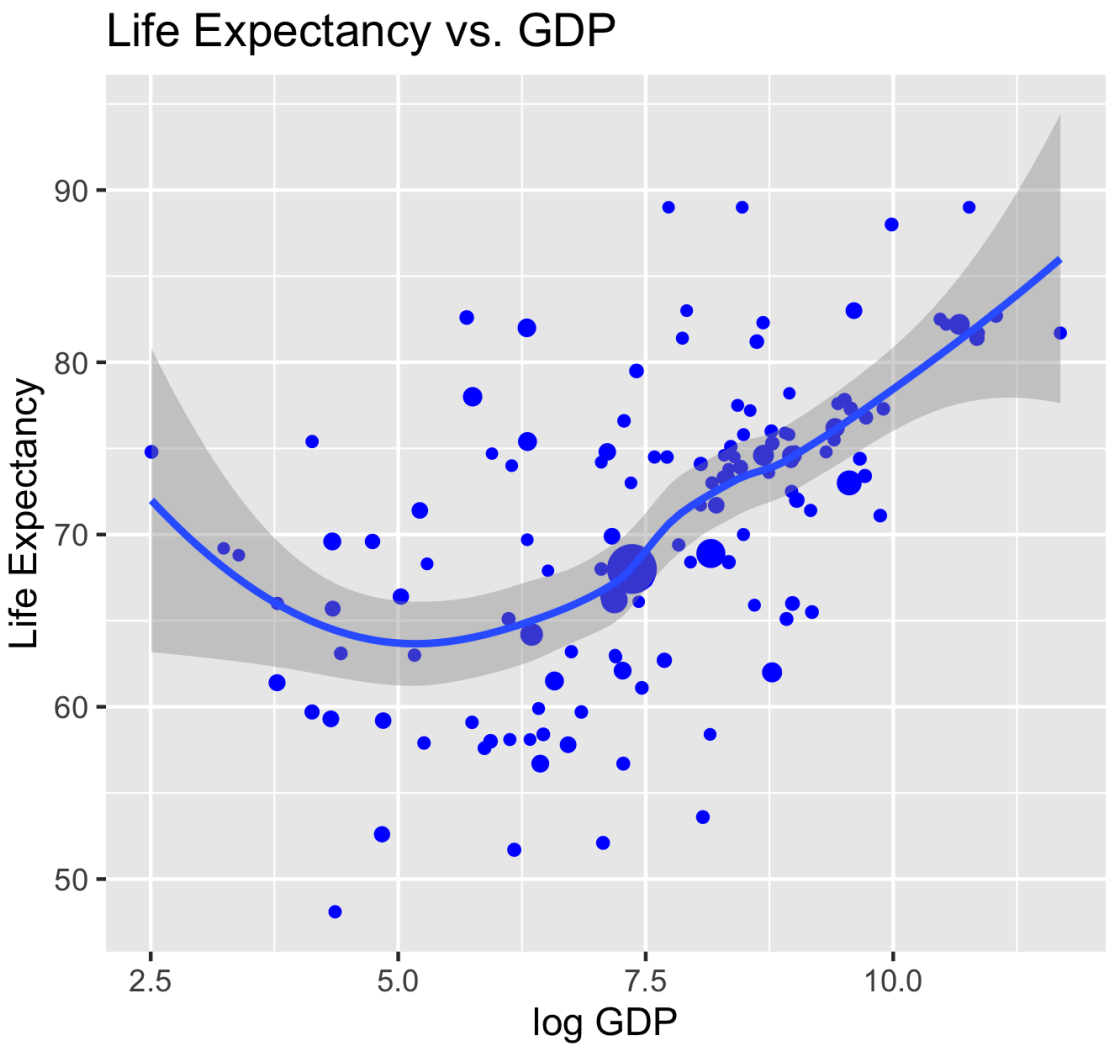
More health expenditure, immunization, schooling, GDP, and BMI all show a positive effect on life expectancy, while alcohol consumption and population have a negative effect.

The panel regression with fixed effects shows that the effect of these variables on life expectancy is consistent over time. The visual representation of the relationships through scatter plots provides an illustration of the relationships between some of the independent variables and life expectancy. Some plots show a clear relationship while other represent an unclear relationship because of various outliers in the data set.

Overall, the results of the regression models tell us insights into the factors that influence life expectancy and can help improve policies to get better public health.

4 Code

My favourite code is the one I did for the following graph:



The code for this graph is as follows: graph4 <- ggplot(data=data$_2$014, $aes(x = log_GDP, y = LifeExpectancy))+$

geom$_p oint(aes(size = Population), color = "blue")+$

ggtitle("Life Expectancy vs. GDP") +

xlab("log GDP") + ylab("Life Expectancy") +

guides(color=FALSE, size=FALSE) +

geom$_s mooth()$

This creates a scatter plot using the ggplot2 package. The plot shows the relationship between log GDP and Life Expectancy.

The data argument uses the data frame to use for the plot (data 2014 in this case). This tibble selects all variables only for the year 2014 (this I do to get a more specified relationship). Aes() is used for mapping between the variables and to specify the visual properties of the plot.Log GDP is on the x-axis and Life Expectancy is on the y-axis. I wanted the points of the plot to represent the population size, so I set size equal to the population of the country.

The geom point() adds the points to the plot, with the size of each point equal to the Population. The color of the points is set to blue using "color".

The ggtitle(), xlab(), and ylab() add a title and axis labels to the plot. The guides() function removes the color and size legend from the plot.

Lastly, the geom smooth() function is used to add a line to the plot to better see the trend of the two variables.

5 Conclusion

The reason to study this regression is that it creates a big policy implication. As we know, that Life expectancy is a check for a country's development as it is usually checked for the country's health infrastructure. Studying this, gives countries around the world, the ways to increase their Life Expectancy. Increasing the number of years in schooling, we can see has a positive effect on Life Expectancy. Similarly, increasing vaccine immunisation, expenditure on healthcare, Taking care of oneself that means alcohol intake or exercising (BMI) increases Life expectancy. This policy implication has many fold effects. With a greater Life Expectancy, the workforce is expanded which would ensure greater revenue and GDP. This also works for the overall happiness of the society. This is a big step towards countries getting developed. To unlock a country's human resources, this is extremely important.