DS6372 WHO Project

Yusi Zhang

Aurian Ghaemmaghami

Sreeni Prabhala

**Data Description**

The 2014 WHO dataset was collected from the WHO and United Nations website with the help of Deeksha Russell and Duan Wang for use in data science education. We have 183 observations with 22 explanatory variables, and it is collected from various countries around the world. We will bring up NA filtration in our “Model Build” section depicted below.

**Objective #1, Model 1**

1. **Problem Statement:** We must build a model with the main goal to identify relationships and is highly interpretable with regard to our response variable of interest: Life Expectancy.

**Exploratory Data Analysis**

1. **Initial EDA**
   1. We first began our data exploration process by viewing the different types of categorical and numerical variables of interest. Variables of “Country Name” and “Year” were deemed unnecessary to include in any further analysis, so we went ahead and filtered those variables from our model.
   2. Next, we built out a correlation matrix (Appendix: 1A) to assess any relationships we see prior to doing any filtering of NA values in our data set. The reason we started with a correlation matrix is because 20/22 of our variables were either integer or numerical values. Based off a correlation cut off point of 0.5 (moderate relationships with Life Expectancy), the variables of interest we wanted to focus on were Adult Mortality, Alcohol, BMI, HIV/AIDS, Income Composition of Resources, and Schooling.
   3. We then built out a scatterplot matrix (Appendix: 1B) of our predictors of interest to view any sort of collinearity between predictors or irregular distributions. We see that both Income Composition of Resources and Schooling show high correlation with each other, as well as, correlation with other predictors. We also observe that BMI seems to have a quadratic relationship with Life Expectancy. Lastly, both Alcohol and HIV/AIDS variables show evidence of a possible log transformation to help better explain the Life Expectancy Predictor
   4. Prior to addressing the collinearity show between Income Comp. and Schooling, we wanted to observe the individual relationship of BMI vs Life Expectancy (Appendix: 1C). We see visual evidence of a quadratic relationship and will move forward without this predictor in our model since we are trying to achieve the most interpretable model. Including quadratic terms may muddy up the interpretation of our problem of interest.
   5. Since the consideration of log transforming Alcohol and HIV seemed applicable, we first observed the correlation coefficients of our Alcohol and HIV variables as depicted in appendix: 1A (0.53,-0.62, respectively). We then log transformed both variables and saw an increase of correlation with both variables on Life Expectancy (0.56,-0.78, respectively – Appendix: 1D). With this information, we re-plotted our matrix of scatterplots (Appendix: 1E) and plotted each variable individually against Life Expectancy (Appendix: 1F/1G) to see if there is evidence of a linear relationship. Based off our graphics above, we will move forward with including both log transformed variables.
   6. Lastly, there is a categorical variable of interest named “Status” that shows the countries of being developed or developing. Our team believes this is a predictor that needs to be included in our model to create a more intuitive regression model for these subsets of countries. More specifically, we don’t want developed countries explaining levels of variability for the response in unison with developing countries. We believe this will create a disingenuous view of the analysis and will move forward with this categorical variable to include in our model.

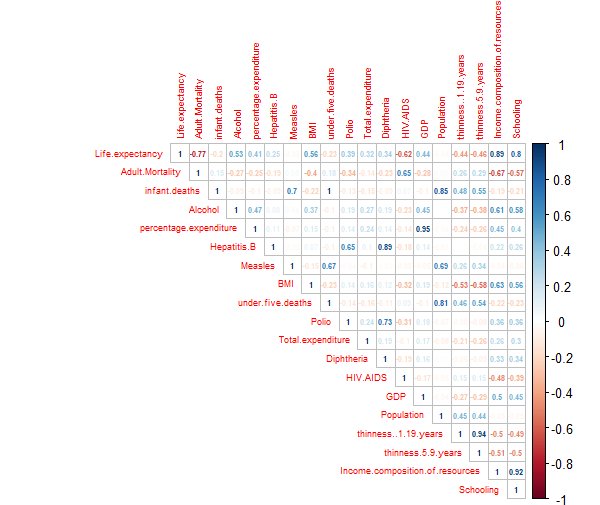
**Model Build**

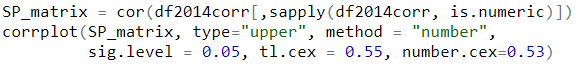
Predictors of interest: Adult Mortality, Log Alcohol, Log HIV/AIDS, Income Composition, and Schooling

1. **Dealing with NA Values:**
   1. With the predictors of interest listed above, we went ahead and filtered out the countries who have NA’s listed in those columns. Countries that were filtered out are listed below:
      1. South Sudan, Somalia, Coute d’ Ivor, Democratic Republic of Congo, Democratic Republic of Korea, Czechia, Republic of Korea, Tanzania, Republic of Moldova, Great Britain and Ireland. Below is the distribution of the countries by continent:
         1. Africa – 5/54 (9.3%)
         2. Europe – 4/44 (9.1%)
         3. Asia – 2/48 (4.2%)
      2. We are confident in moving forward with the filtered data set without these countries as it is only a small percentage of each of the world’s continents.
2. **Model Build:**
   1. After multiple trial and error (Appendix: Model Build) of dealing with correlated predictors and insignificant terms, we arrived at the following model:
   2. **Life Expectancy = 55.682 - 0.018Adult.Mortality -1.963StatusDeveloping + 26.827Income.comp.of.resources -1.423LogHIVAIDS + 0.226LogAlcohol**
3. **Assumptions:**

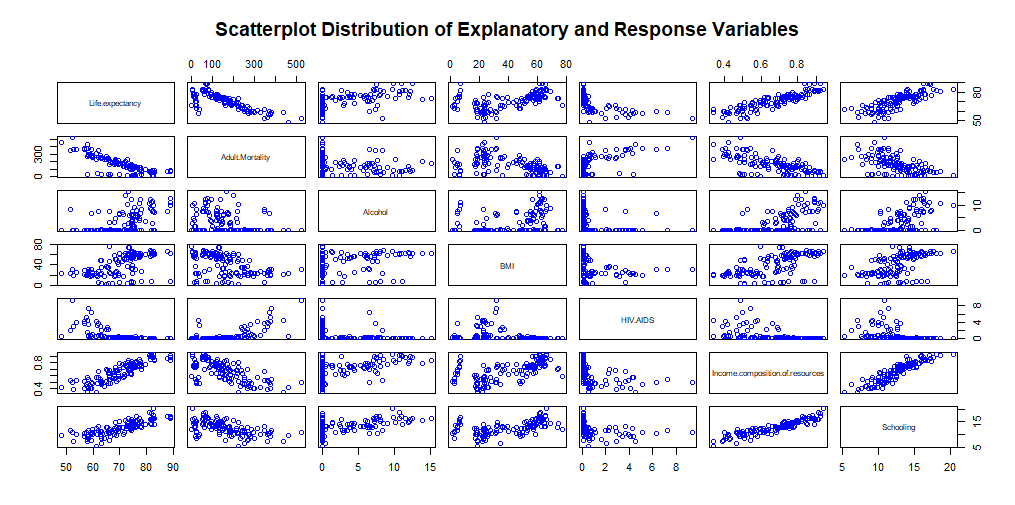
**Appendix:**

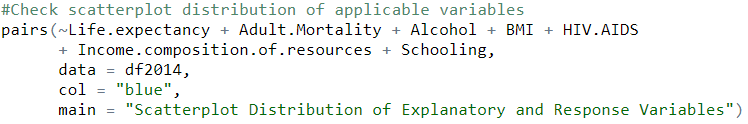
1A)



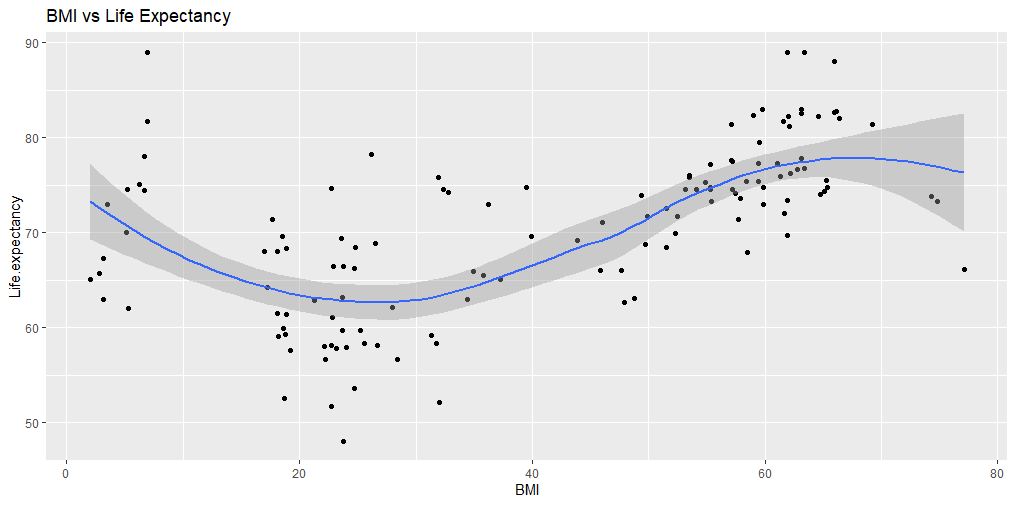


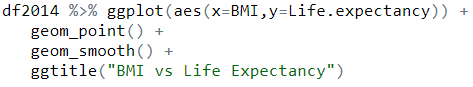
1B)



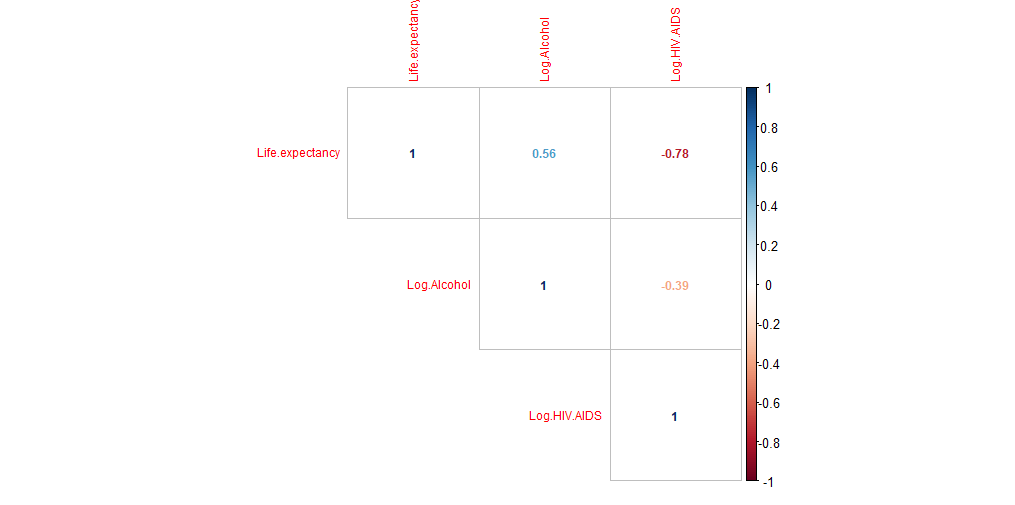


1C)



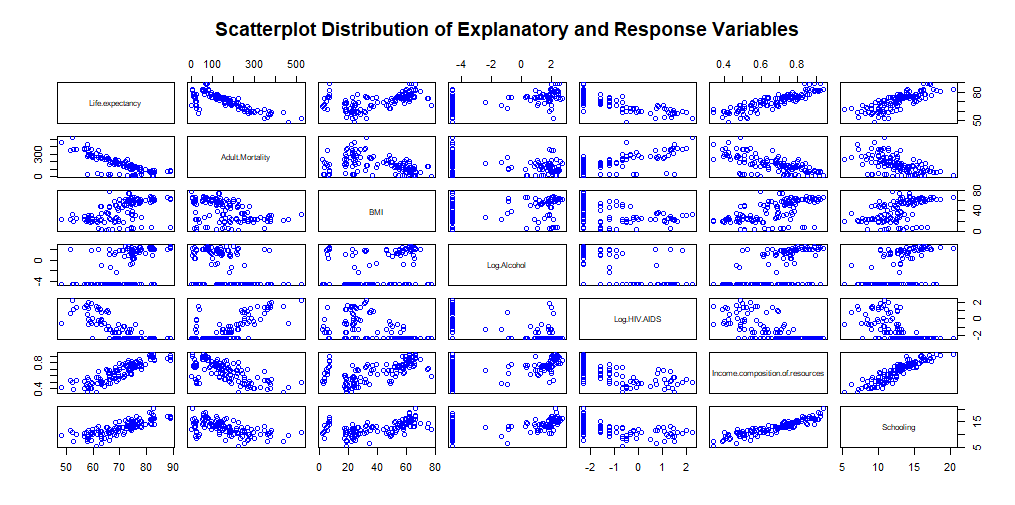


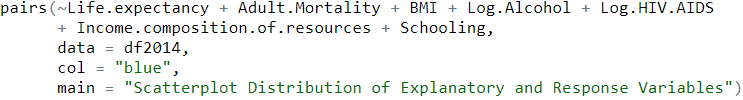
1D)



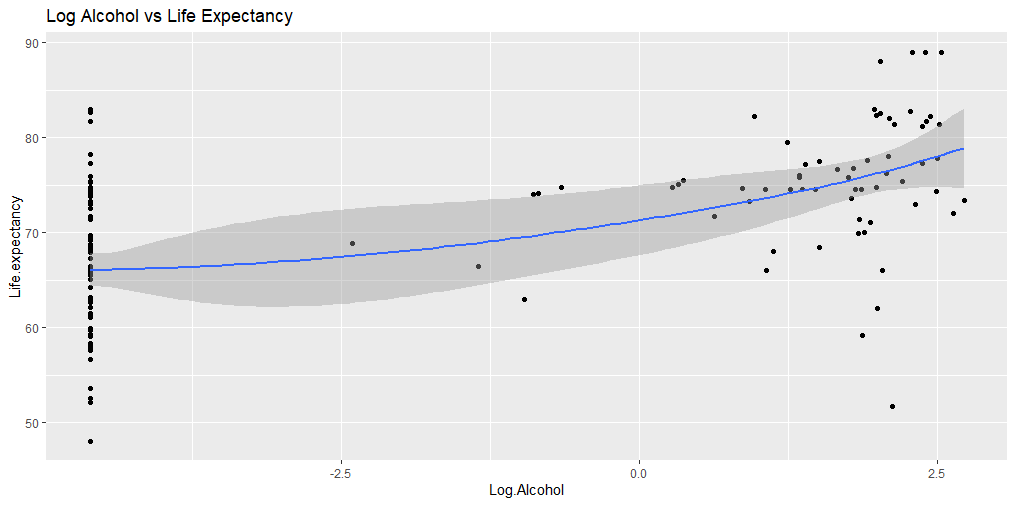


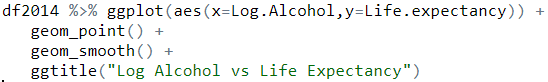
1E)





1F)





1G)