Project Submission #1

Pooja Kabber, Dingkun Yang, Echo Chen, Andrew Kroening

October 21st, 2022

Data Overview

The dataset used for this research is from the World Health Organization (WHO). The WHO compiles data on thousands of variables for as many countries as feasible and presents them for analysis. This particular dataset contains national-level observations of variables related to life expectancy around the globe for a period spanning the early portion of the 21st century.

Overall Characteristics The complete dataset contains observations beginning in the year 2000 and ending in the year 2015. As a complete dataset, there are 2,938 observations for 22 variables. In practical terms, each country has approximately one observation, averaging 183 for each of the 16 years encompassed by the data. The dataset effectively contains 20 variables for each country and year combination, covering major disease, economic, and social factors. All variables are fully described in **Appendix B, Table XXXXX**. Some important variables are called out below:

- Life.expectancy: Continuous variable of the adult mortality rates of both sexes (probability of dying between 15 and 60 years per 1000 population).
- Status: Categorical variable identifies a country as 'Developing' or 'Developed' in the given year.
- GDP: Continuous variable showing Gross Domestic Product per capita in the given year.
- Population: Continuous variable with the total estimated population for the country in a given year.
- percentage.expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%).
- Hepatitis.B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%).
- Measles: Continuous variable of the number of reported cases per 1000 population.
- Polio: Polio immunization coverage among 1-year-olds (%).
- HIV.AIDS: Continuous variable of the deaths per 1,000 live births due to HIV/AIDS for 0-4 year olds.
- Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).
- Schooling: Continuous variable expressing the average number of years of school completed among the country's population.
- Income.composition.of.resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1).

Sample Dataset Characteristics In the context of this analysis, the dataset will be reduced to include only the year 2014. This yields once observation for 183 countries across all of the variables.

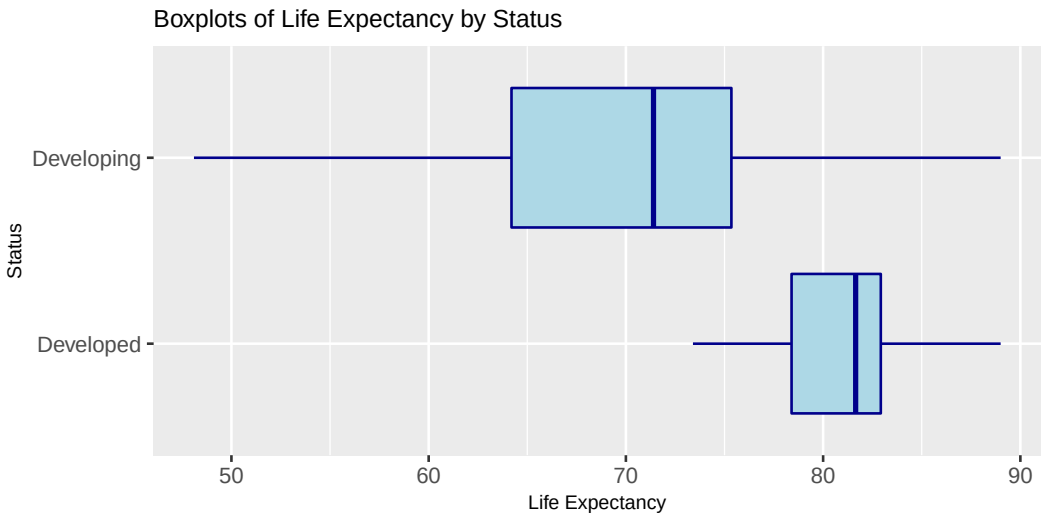
Research Questions:

- How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"
- How did disease and mortality rates, along with national economic factors, contribute to a country's development status in 2014?

Primary Relationship of Interest

We conduct a priori variable selection of the independent variables. Below are our findings:

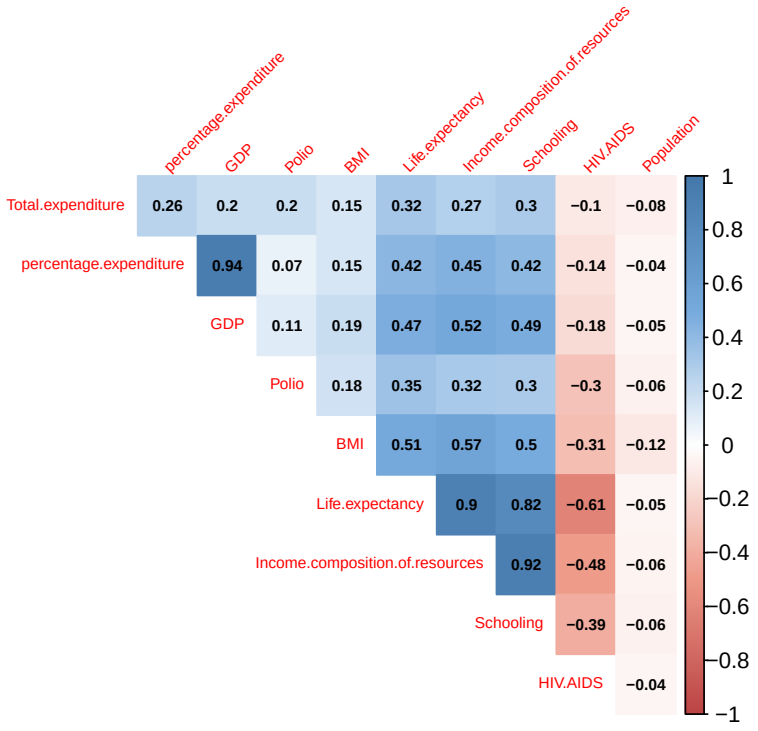
The predictors Schooling, Adult Mortality, BMI, Total Expenditure, HIV/AIDS, Thinness 1-19, Thinness 5-9, Income composition of resources and Status have a high correlation with Life Expectancy as can be seen from the scatterplot matrix and the boxplot below. From the boxplot, we can see that the mean Life Expectancy is much lower for developing countries than for developed countries.



From the below correlation matrix, we can see that some predictors are highly correlated with each other:

- "Adult.Mortality", "infant.deaths" and "under.five.deaths" variables are directly correlated to Life Expectancy, we choose to drop them from the predictor variable list; on the other hand, we are interested in the "HIV.AIDS" variable (Deaths per 1,000 live births HIV/AIDS (0-4 years). The rationale for this decision is that the mortality and death variables are measuring the fatalities of entire groups of the population, whereas the "HIV.AIDS" variable is a single measurement that might encompass several underlying societal indicators.
- We see extremely high correlation between GDP and percentage expenditure. However, since both of them could have meaningful interpretations and the categories are distinct, we may want to decide which one goes to our final model when we conduct model selection.
- To represent immunization coverage, among "Hepatitis.B","Polio", "Diphtheria tetanus toxoid and pertussis (DTP3)", we opted to use "polio", since it has the highest correlation with Life Expectancy.

- BMI is highly correlated with thinness 5-9 and thinness..1.19.years so only BMI will be included.



Since our research question focuses on the effect of social, economic and major disease factors over life expectancy, we will keep some variables that are of interest regardless of the future model performance. These are schooling, income composition of resources, total expenditure, status, population and gdp.

We will only consider these variables for the rest of the analysis. In doing so, the dataset is reduced in size to 183 observations of the following variables:

> 'Country', 'Status', 'Population', 'Life.expectancy', 'percentage.expenditure', 'Measles', 'Polio', ' HIV.AIDS', 'GDP', 'Schooling', 'Income.composition.of.resources', 'BMI', 'Total.expenditure'

Table One Table 1 describes summary information about each of these variables.

Table 1: Summary of Variables

Statistic	N	Mean	St. Dev.	Min	Max
Population	142	21,062,964.000	112,170,596.000	41.000	1,293,859,294.000
Life.expectancy	183	71.537	8.561	48.100	89.000
percentage.expenditure	183	1,001.913	2,553.290	0.000	19,479.910
Measles	183	1,831.208	8,770.077	0	79,563
Polio	183	84.727	20.869	8	99
HIV.AIDS	183	0.682	1.388	0.100	9.400
GDP	155	10,015.570	18,484.240	12.277	119,172.700
Schooling	173	12.887	2.912	4.900	20.400
Income.composition.of.resources	173	0.688	0.154	0.345	0.945
BMI	181	41.031	21.110	2.000	77.100
Total.expenditure	181	6.201	2.743	1.210	17.140

From the table we can see some interesting characteristics of the data. Life.expectancy has a standard deviation between countries as high as 8.56 years. There are some countries whose percentage.expenditure is 0, with a mean of 1001 and a max of 19,479 which gives us a sense of the skew in the various countries' spending power. GDP is as low as 12.28. We can also see that variables like Population, Measles and GDP are highly skewed. We will expand on these further in the Challenges section of the analysis.

Other Characteristics

Below is a brief summary of the other types of variables present in the dataset:

- Variables like Status and Population represent the social dimensions of the data collected.
- Adult.Mortality, infant.deaths and under.five.deaths represent the different kinds of rates that are consolidated into Life.expectancy.
- percentage.expenditure, GDP and Income.composition.of.resources give us a sense of the economic well-being of different countries.
- We can analyse the effect of major diseases using the variables Hepatitis.B, Measles, Polio, HIV.AIDS and Diphtheria and lifestyles of people using Alcohol and BMI.

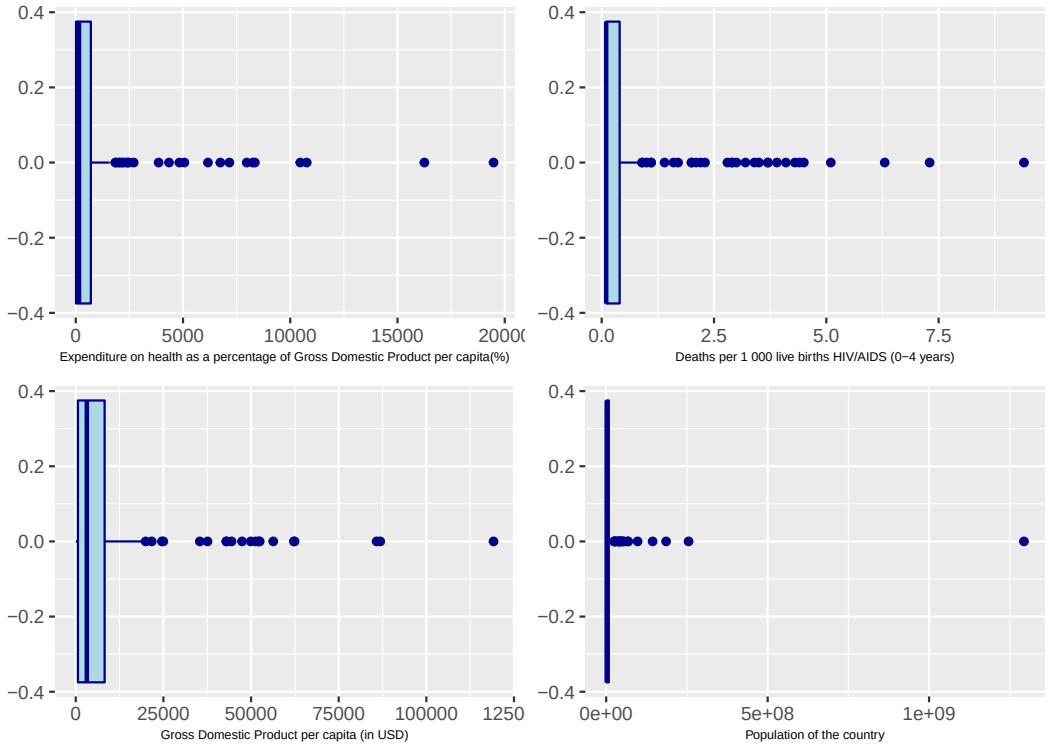
Potential Challenges

Missingness
The primary challenge facing the project is missing data. In the 2014 subset of the Life Expectancy dataset, there are 51 observations with a missing value in at least one of the potential predictors:

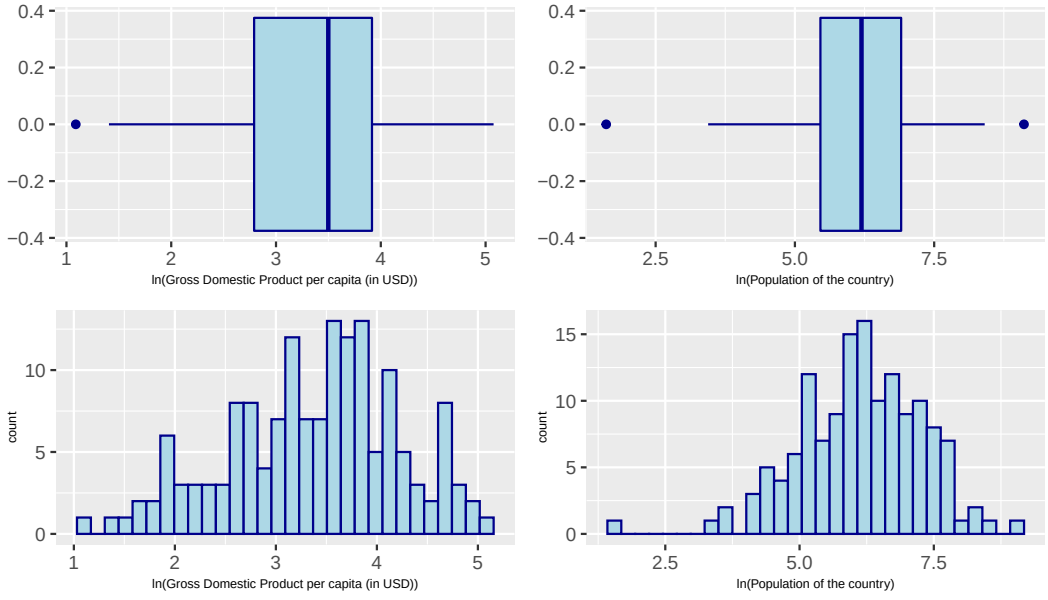
- 41x 'Population' observations
- 10x 'Hep.B' observations
- 28x 'GDP' observations
- 10x 'Schooling' observations
- 10x 'Income.composition.of.resources' observations

The team has proposed three potential avenues for mitigating the effects of these missing values. The observations in question may be dropped or filled with a mean or median value for that variable based on the observed skew.

Skew
There is a second challenge posed in some of the variables in the dataset by a skewed distribution. When checking box plots of all variables, we found 4 variables may need transformation before inclusion in the final model.



We may want to use log transformation for population and GDP, because of the magnitude of gaps. But for other two, we need more investigation due to the difficulty of interpretation. The box plots and histograms of the population and GDP variables after log transformation yield a better distribution:



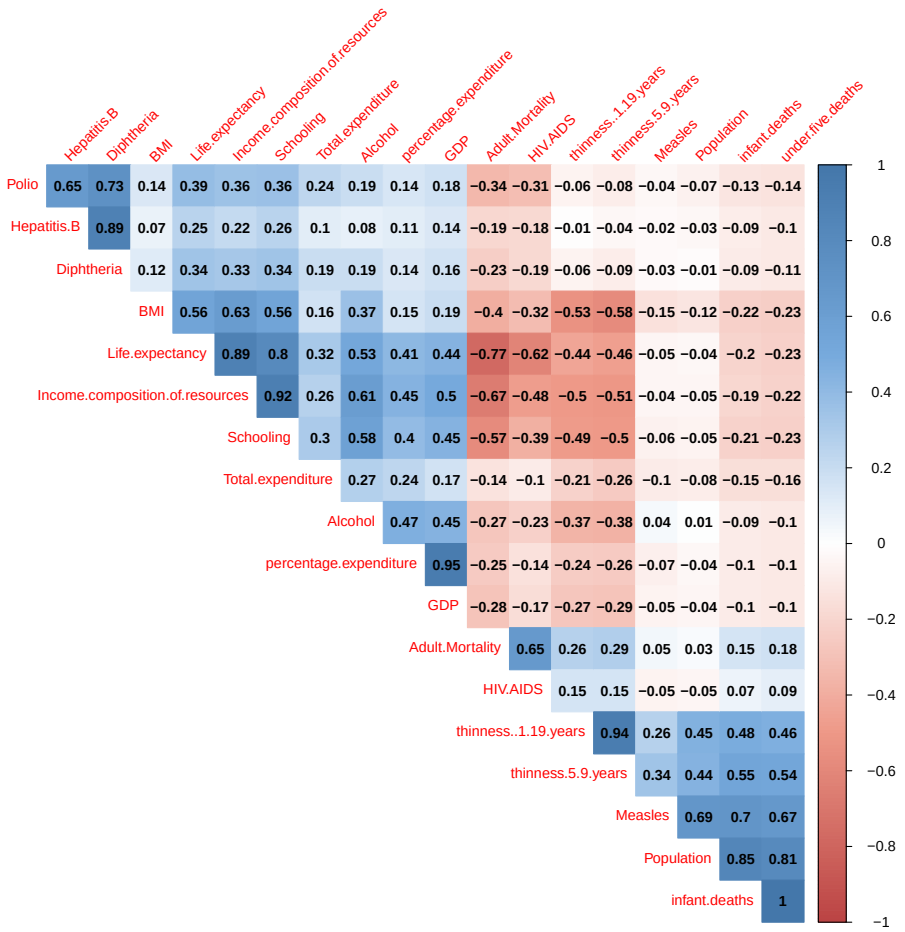
Potential solution

We may use log transformation on skewed variables as described above. We can also remove outliers or normalize(min-max) our dataset.

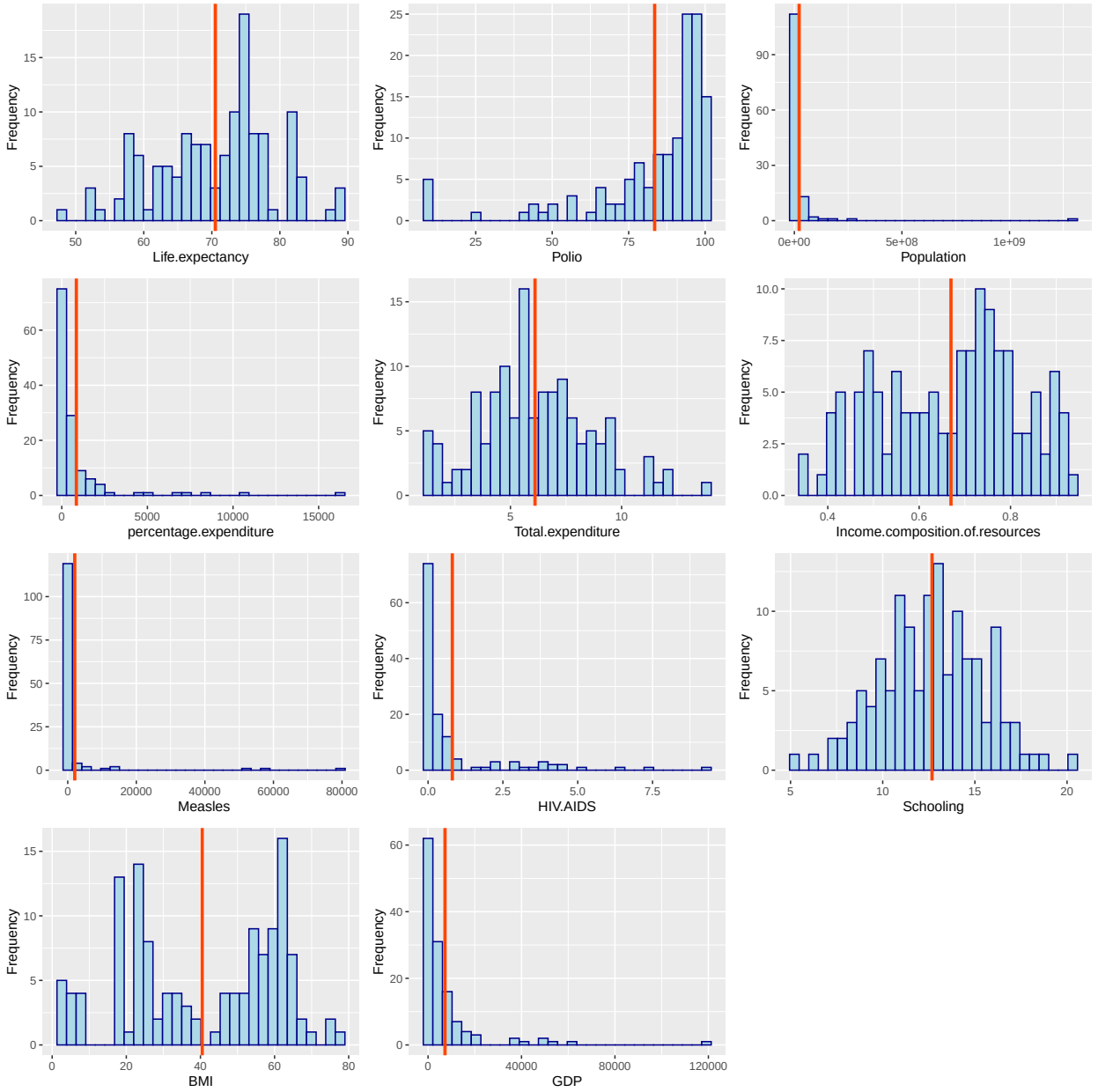
Appendix A - Supporting Tables and Figures

This appendix is for primary supporting information, tables, or plots that are *valuable for the model that we are keeping.*

"Figure XX - Correlation Matrix"



"Figure XXX - Histgrams for Potential Predictor Variables"



Appendix B - Variables Description

This appendix is for other information, tables, or plots that are less valuable to the models or will be removed.

Table XXXX - All Variables

Variable	Type	Description
Country	factor	Country name
Year	numeric	Year of the data
Status	factor	Country status of developed or developing
Life_Expectancy	numeric	Life expectancy in age
Adult_Mortality	numeric	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
infant.deaths	numeric	Number of Infant Deaths per 1000 population
Alcohol	numeric	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
percentage.expenditure	numeric	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hepatitis.B	numeric	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	numeric	number of reported cases per 1000 population
BMI	numeric	Average Body Mass Index of entire population
under.five.deaths	numeric	Number of under-five deaths per 1000 population
Polio	numeric	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total.expenditure	numeric	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	numeric	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV.AIDS	numeric	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	numeric	Gross Domestic Product per capita (in USD)
Population	numeric	Population of the country
thinness..1.19.years	numeric	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
thinness.5.9.years	numeric	Prevalence of thinness among children for Age 5 to 9(%)
Income.composition.of.resources	numeric	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	numeric	Number of years of Schooling(years)