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. The variables are fully described in **Appendix B, Table XXXXX**. Some important variables are called out below:

- Life.expectancy: Continuous variable with the estimated life expectancy in a given year for the country
- Status: Categorical variable identifies a country as 'Developing' or 'Developed' in the given year.
- GDP: Continuous variable showing GDP per capita in the given year.
- Population: Continuous variable with the total estimated population for the country in a given year.
- BMI, Alchohol, Measles, Polio, and other health and disease-related variables.

Sample Dataset Characteristics For this research proposal, the data set will be truncated to 2014 and reduced to certain variables based on their potential relevance to the research questions. In doing so, the dataset is reduced in size to 183 observations of the following variables:

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

The research team estimates that these variables will lead to the best model fit for the dual-purpose nature of this analysis. Table 1 describes summary information about each of the variables that the team believes have a potential to be included in a final model.

This is where we need to put in plots, or details of some of the justifications for our 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?

Table 1: Summary of Variables

Statistic	N	Mean	St. Dev.	Min	Max
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
Hepatitis.B	173	83.116	23.357	2	99
Measles	183	1,831.208	8,770.077	0	79,563
Polio	183	84.727	20.869	8	99
Diphtheria	183	84.082	23.033	2	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
Population	142	21,062,964.000	112,170,596.000	41.000	1,293,859,294.000
Income.composition.of.resources	173	0.688	0.154	0.345	0.945
Schooling	173	12.887	2.912	4.900	20.400

Primary Relationship of Interest

Present descriptive statistics and exploratory plots in whichever format you think is best (tables, figures) for your primary relationship of interest (dependent and primary independent variables, if applicable). Describe your findings.

Other Characteristics

Briefly describe other variables in the data. If there are many, do not list them all. Instead, describe the types of variables present (e.g., "demographic information").

Potential Challenges

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.

Appendix A

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

Appendix B

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 (%)room)
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)

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