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 A, 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.

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?

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

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