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:

- Status: Categorical variable identifies a country as 'Developing' or 'Developed' in the given year.
- Population: Continuous variable with the total estimated population for the country in a given year.
- Life.expectancy: Continuous variable with the estimated life expectancy in a given year for the country
- percentage.expenditure: (How is it calculated, what are the units)
- Hepatitis.B: (How is it calculated, what are the units)
- Measles: (How is it calculated, what are the units)
- Polio: (How is it calculated, what are the units)
- HIV.AIDS: (How is it calculated, what are the units)
- Diphtheria: (How is it calculated, what are the units)
- GDP: Continuous variable showing GDP per capita in the given year.
- Schooling: (How is it calculated, what are the units)
- Income.composition.of.resources: (How is it calculated, what are the units)

Sample Dataset Characteristics

For this research, the data set will be specific to 2014 and reduced to certain variables based on their potential relevance to the research questions (for more information, refer **Appendix A, Table XXXXX**). 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.

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.

After conducting a priori variable selection of the independent variables, below are the 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. In the boxplot, we can see that the mean life expectancy is much lower for developing countries than for developed countries.

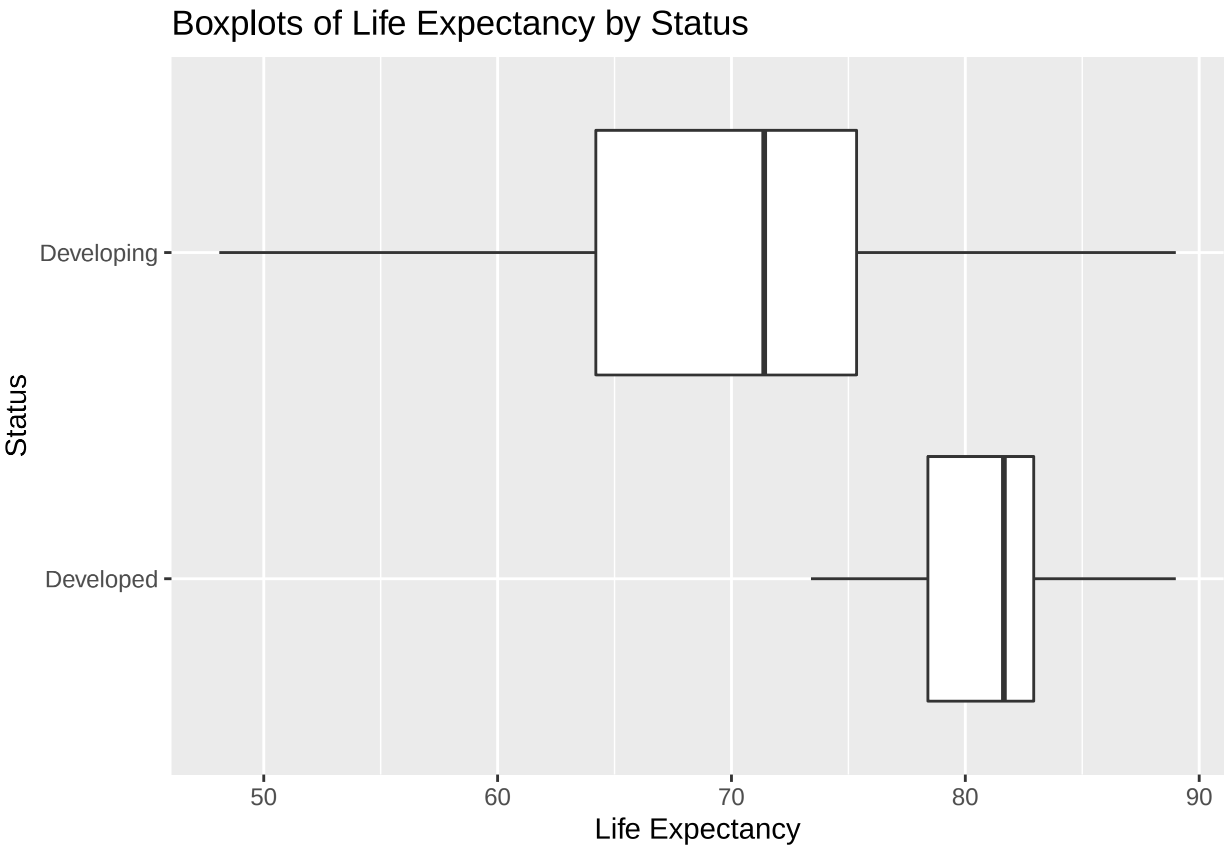
From the correlation matrix, we can see that some predictors are highly correlated with each other. One of each correlated predictor pair will be excluded from the analysis / model: * infant deaths are highly correlated with under five deaths so only one of these variables will be included * percent expenditure is highly correlated with GDP so only one of these variables will be included * Hepatitis is highly correlated with Polio and Diphtheria so only one of these variables will be included * BMI is highly correlated with thinness 5-9 and thinness..1.19.years so only one of these variables will be included

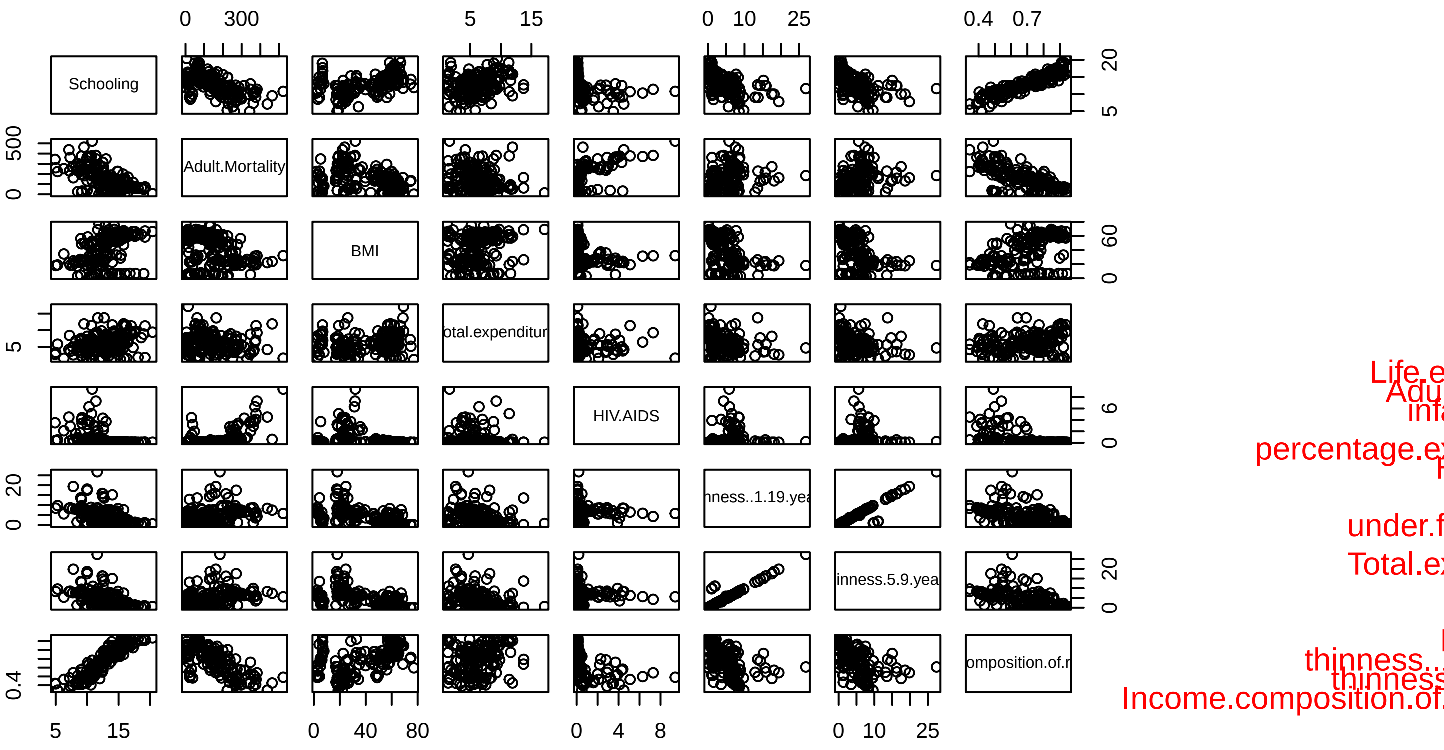
Since our research question focuses on the effect of social, economic and major disease factors over life expectancy, we will keep some variables that are of interest regardless of the future model performance. These are schooling, income composition of resources, total expenditure, status, population,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

For the rest of the exploratory data analysis, we will only consider these variables.

Table One

Due to the size of the grouping variable 'Country', we will plot the Table 1 with one developed and one developing country.


Boxplots of Life Expectancy by Status
Status
Developing
Developed
50
60
70
80
90
Life Expectancy


Schooling
Adult.Mortality
BMI
otal.expenditur
HIV.AIDS
nness..1.19.yea
inness.5.9.yea
omposition.of.r
0
300
5
15
0
10
25
0.4
0.7
20
5
500
0
60
0
5
6
0
20
0
20
0
0.4
5
15
0
40
80
0
4
8
0
10
25
Life.e
Adu
infa
percentage.ex
H
under.f
Total.ex
I
thinness..
thinness
Income.composition.of.

```
## % latex table generated in R 4.2.1 by xtable 1.8-4 package
## % Thu Oct 20 02:11:25 2022
## \begin{table}[ht]
## \centering
## \begin{tabular}{rlllllll}
##   \hline
##  &   Schooling & Income.composition.of.resources & Total.expenditure &    Status &   Population &  H
##   \hline
## X & Min.   : 4.90   & Min.   :0.3450   & Min.   : 1.210   & Length:183         & Min.   :4.100e+01
##   X.1 & 1st Qu.:10.80   & 1st Qu.:0.5700   & 1st Qu.: 4.480   & Class :character   & 1st Qu.:2.869e+0
##   X.2 & Median :13.00   & Median :0.7220   & Median : 5.840   & Mode  :character   & Median :1.568e+0
##   X.3 & Mean   :12.89   & Mean   :0.6884   & Mean   : 6.201   &  & Mean   :2.106e+07   & Mean   :83.
##   X.4 & 3rd Qu.:14.90   & 3rd Qu.:0.7960   & 3rd Qu.: 7.740   &  & 3rd Qu.:8.080e+06   & 3rd Qu.:97.0
##   X.5 & Max.   :20.40   & Max.   :0.9450   & Max.   :17.140   &  & Max.   :1.294e+09   & Max.   :99.0
##   X.6 & NA's   :10   & NA's   :10   & NA's   :2   &  & NA's   :41   & NA's   :10   & NA's   :28   \\
##    \hline
## \end{tabular}
## \end{table}
```

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.