SI 618 Exploratory Data Analysis Project

Kim Vuong

Motivation

(a) With an interest in public health and health disparities, the overall goal for this project is to explore whether factors such as sanitation services and health infrastructure available may have correlation with global life expectancy. I also would like to see if there are any potential links between a country's GDP spending more money on health and life expectancy. The data table I will be creating will contain rates for these factors and by visualizing them I hope to extract any potential relationships between these factors.

(b) **Questions to Explore:**

1. Is there any potential relationship and/or trend between how much a country's GDP spending on health and life expectancy?
2. Is there any potential relationship and/or trend between countries with better health infrastructure in place (e.g. availability of more hospitals) and having longer life expectancy?
3. Is there any potential relationship and/or trend between sanitation conditions such as drinking water and sanitation services with life expectancy?

Data Source

All 5 datasets were from The World Health Organization (The WHO) as the source. The format was all in Comma Separated Values (CSV).

1) **Life Expectancy Data by Country:** http://apps.who.int/gho/data/node.main.SDG2016LEX?lang=en

   Variables and Type:

   - Country - Character
   - Year (2000-2016) - Integer
   - Life Expectancy at Birth (Years) for: *Both sexes - Numeric* Male - Numeric *Female - Numeric
   - Life Expectancy at age 60 (Years) for: *Both sexes - Numeric* Male - Numeric *Female - Numeric
   - Healthy Life Expectancy (HALE) at Birth (Years) for: *Both sexes - Numeric* Male - Numeric *Female - Numeric
   - Healthy Life Expectancy (HALE) at 60 (Years) for: *Both sexes - Numeric* Male - Numeric *Female - Numeric

   Number of Observations: 3111

2) **Percent Gross Domestic Product Health Expenditure:** https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS (URL from The World Bank but data source is The WHO)

   Variables and Type:

   - Country - Character
   - Country Code - Character
   - Indicator Name - Character
   - Indicator Code - Character
   - Year (1960-2019, 59 variables) - Numeric

   Number of Observations: 264

3) **Health Infrastructure by Country:** http://apps.who.int/gho/data/node.main.506?lang=en

   Variables and Type:

- Country - Character
- Year (2010, 2013) - Integer

- Total density per 100,000 population: Health posts - Numeric
- Total density per 100,000 population: Health centres - Numeric
- Total density per 100,000 population: District/rural hospitals - Numeric
- Total density per 100,000 population: Provincial hospitals - Numeric
- Total density per 100,000 population: Specialized hospitals - Numeric
- Total density per 100,000 population: Hospitals - Numeric

Number of Observations: 283

4) **Basic and Safely Managed Drinking Water Services Data by Country:** http://apps.who.int/gho/data/node.main.WSHWATER?lang=en

Variables and Type:

- Country - Character

- Year (2000-2017, 102 variables)
  - Total Population using at least basic drinking-water services (%) - Integer
  - Urban Population using at least basic drinking-water services (%) - Integer
  - Rural Population using at least basic drinking-water services (%) - Integer
  - Total Population using safely managed drinking-water services (%) - Integer
  - Urban Population using safely managed drinking-water services (%) - Integer
  - Rural Population using safely managed drinking-water services (%) - Integer

Number of Observations: 194

5) **Basic and Safely Managed Sanitation Services Data by Country:** http://apps.who.int/gho/data/node.main.WSHSANITATION?lang=en

Variables and Type:

- Country - Character

- Year (2000-2017, 102 variables)
  - Total Population using at least basic sanitation services (%) - Integer
  - Urban Population using at least basic sanitation services (%) - Integer
  - Rural Population using at least basic sanitation services (%) - Integer
  - Total Population using safely managed sanitation services (%) - Integer
  - Urban Population using safely managed sanitation services (%) - Integer
  - Rural Population using safely managed sanitation services (%) - Integer

Number of Observations: 194

Methods

Q1. Is there any potential relationship and/or trend between how much a country's GDP spending on health and life expectancy?

1. Manipulation

For this question, I worked with the Percent Gross Domestic Product Health Expenditure dataset and the Life Expectancy dataset. By looking at the datasets, I decided to choose 2016 data because it is the most complete and most current data for these two variables and subsetted each dataset. For my life expectancy dataset, I chose to use the variable 'life expectancy at birth (years) for both sexes' because it had the most complete data compared to the other categories and I was interested in overall life expectancy(not looking at differences in sexes). I also changed the variable 'life expectancy at birth (years) for both sexes' from a factor to a numeric type, so that analysis can be done. I subsetted the three columns I needed: 'country', 'year', and ‘life expectancy at birth (years) for both sexes’. I subsetted the Percent Gross Domestic Product Health Expenditure to just the country and data for the year 2016. Then, I merged the two datasets, which are in data.table format on ‘country’ to get one table with values for both variables matching on country.

2. Missing, Incomplete, Noisy Data

I omitted any missing data after subsetting my three columns for my life expectancy dataset so as to not lose any row observations that may have ‘NA’ for the other life expectancy categories I was not using but had the value I needed for my life expectancy category. For the GDP health expenditure dataset, after subsetting the data, I omitted any missing values for the year 2016.

3. Challenges

The Life Expectancy data had two rows as column headers so when I initially read the file in, some of the rows were the headers and not the observations I needed. I had to combine the column headers by “collapsing” them to get just one row as the column headers. First, I read the first two rows, combined them, and saved it as a vector. Then I read in the rest of the data and inserted the saved vector for the column names parameter. I renamed the column names after subsetting the data so that the column names were shorter and clearer than the combined column headers.

Q2. Is there any potential relationship and/or trend between countries with better health infrastructure in place (e.g. availability of more hospitals) and having longer life expectancy?

1. Manipulation

For this question, I worked with the Health Infrastructure dataset and the Life Expectancy dataset. By looking at the datasets, I decided to choose 2013 data because it is the most complete and most current data for these two variables and subsetted each dataset. For my life expectancy dataset here, I subsetted the initial life expectancy dataset to the year 2013 for ‘life expectancy at birth (years) for both sexes’. For my health infrastructure dataset, I chose the variable ‘Total density per 100,000 population: Hospitals’ and subsetted this to a data table containing country and the density of hospitals for the year 2013. I also renamed the column for the density to a shorter name to refer to ‘Density.of.Hospitals’ when I perform the analysis. Then, I merged the two data tables, life expectancy for 2013 and the hospital density, on country.

2. Missing, Incomplete, Noisy Data

I already omitted all missing data for the life expectancy dataset in question 1, so the subsetted 2013 data here is already complete. I did omit ‘NA’ values for the health infrastructure dataset that was subsetted by the year 2013 for question 2 prior to merging the two data tables.

3. Challenges

I did not encounter any challenges with the datasets for this question. It was all very straight-forward to work with.

Q3. Are there any potential relationships and/or trends between sanitation conditions such as drinking water and sanitation services with life expectancy?

1. Manipulation

For this question, I worked with the drinking water dataset and the sanitation services dataset with the Life Expectancy dataset. By looking at all three datasets, I decided to choose 2016 data because it is the most complete and most current data for these three variables. I already had the 2016 data for life expectancy subsetted from question 1. For both the drinking water dataset and the sanitation services dataset, the format of the columns were the same. I chose to look at the ‘Total Population using at least basic drinking-water services (%)’, ‘Total Population using safely managed drinking-water services (%)’, ‘Total Population using at least basic sanitation services (%), and ’Total Population using safely managed sanitation services (%)’ because these variables had the most complete data compared to the subcategories for urban and rural population. I subsetted the data to include the columns for country, total population using at least basic services, and total population using safely managed services for drinking water and sanitation services for the year 2016.

2. Missing, Incomplete, Noisy Data

For the drinking water and life expectancy table, I omitted missing values after merging the two variables. I also did the the same procedure with the sanitation services and life expectancy table.

3. Challenges

The Drinking Water and Sanitation Services datasets each had three rows as column headers so when I initially read the file in, some of the rows were the headers and not the observations I needed. I had to combine the column headers by "collapsing" them to get just one row as the column headers. First, I read the first three rows, combined them, and saved it as a vector for each dataset. Then I read in the rest of the data and inserted the saved vector for the column names parameter to the respective dataset. I renamed the column names after subsetting the data so that the column names were shorter and clearer than the combined column headers.

Analysis and Results

Q1. Is there any potential relationship and/or trend between how much a country's GDP spending on health and life expectancy?**

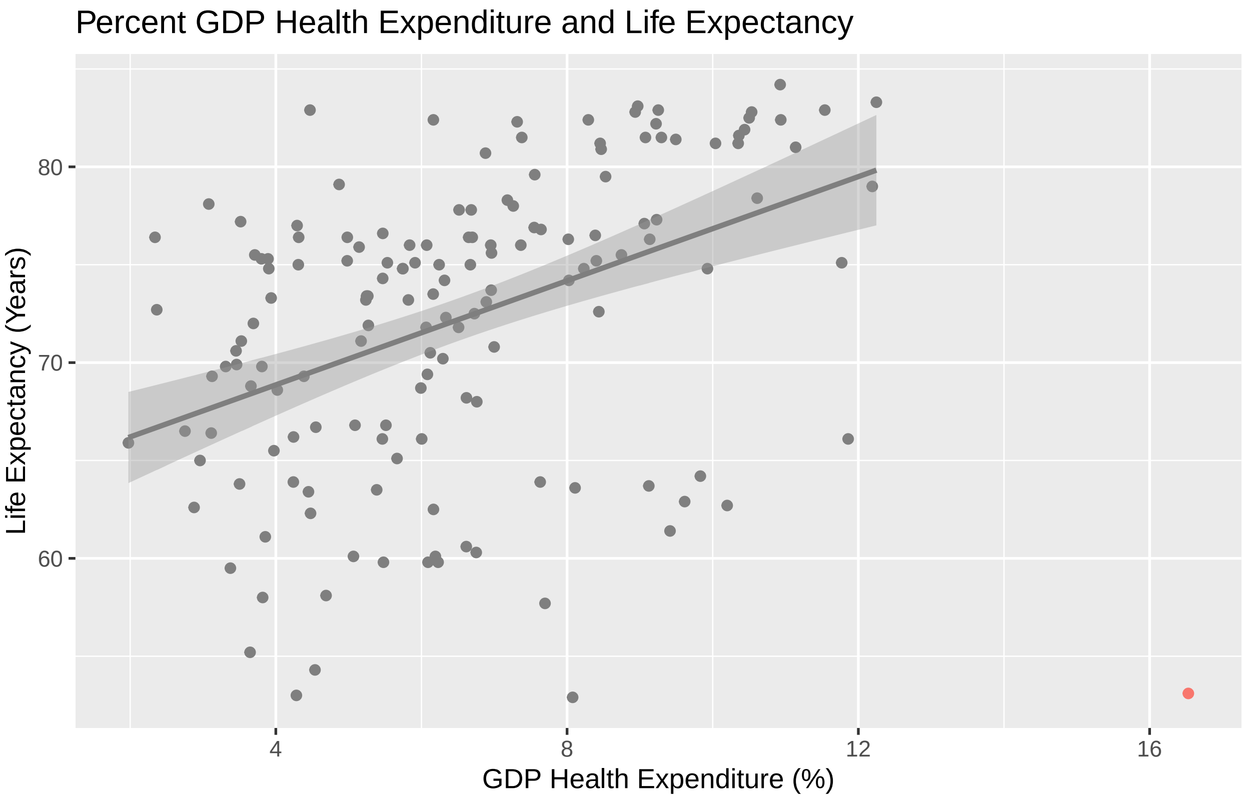
1. Code Workflow

After loading my data and prepping it for analysis, I calculated the correlation between percent GDP on health expenditure with life expectancy. I also used ggplot2() and plotted a scatterplot and added a trend line using the smoothing technique.

2. Results, Relationship, Insights

The correlation between percent GDP on health expenditure and life expectancy using the Pearson method was 0.33. This is a low correlation for a potential positive linear relationship between these two variables. As for trends, after plotting the smoothing line using the Loess method, the data seems to show that as GDP health expenditure increases, so does life expectancy. Note from the visualization below, there is a very prominent outlier from this trend coded in red.

3. Data Visualization



Q2. Is there any potential relationship and/or trend between countries with better health infrastructure in place (e.g. availability of more hospitals) and having longer life expectancy?

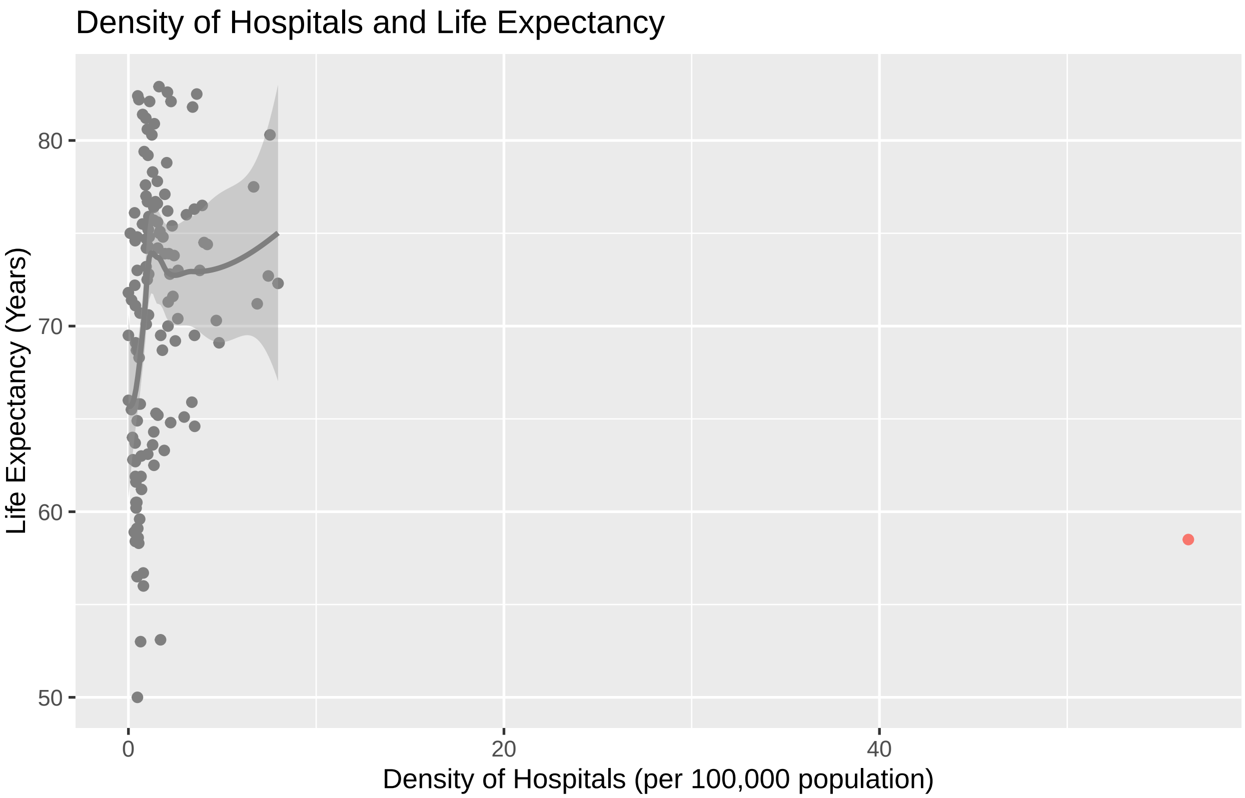
1. Code Workflow

After loading my data and prepping it for analysis, I calculated the correlation between hospital density per 100,000 population and life expectancy. I also used ggplot2() and plotted a scatterplot and added a trend line using the smoothing technique.

2. Results, Relationship, Insights

The correlation between hospital density per 100,000 population and life expectancy using the Pearson method was -0.05. These two variables have no linear relationship. As for trends, after plotting the smoothing line using the Loess method, the data seems to be in a vertical line with the majority of the points and then slightly increases. Note from the visualization below, there is a very prominent outlier from this trend coded in red. Overall, there does not seem to be a relationship or trend between these two variables.

3. Data Visualization



Q3. Are there any potential relationships and/or trends between sanitation conditions such as drinking water and sanitation services with life expectancy?

1. Code Workflow

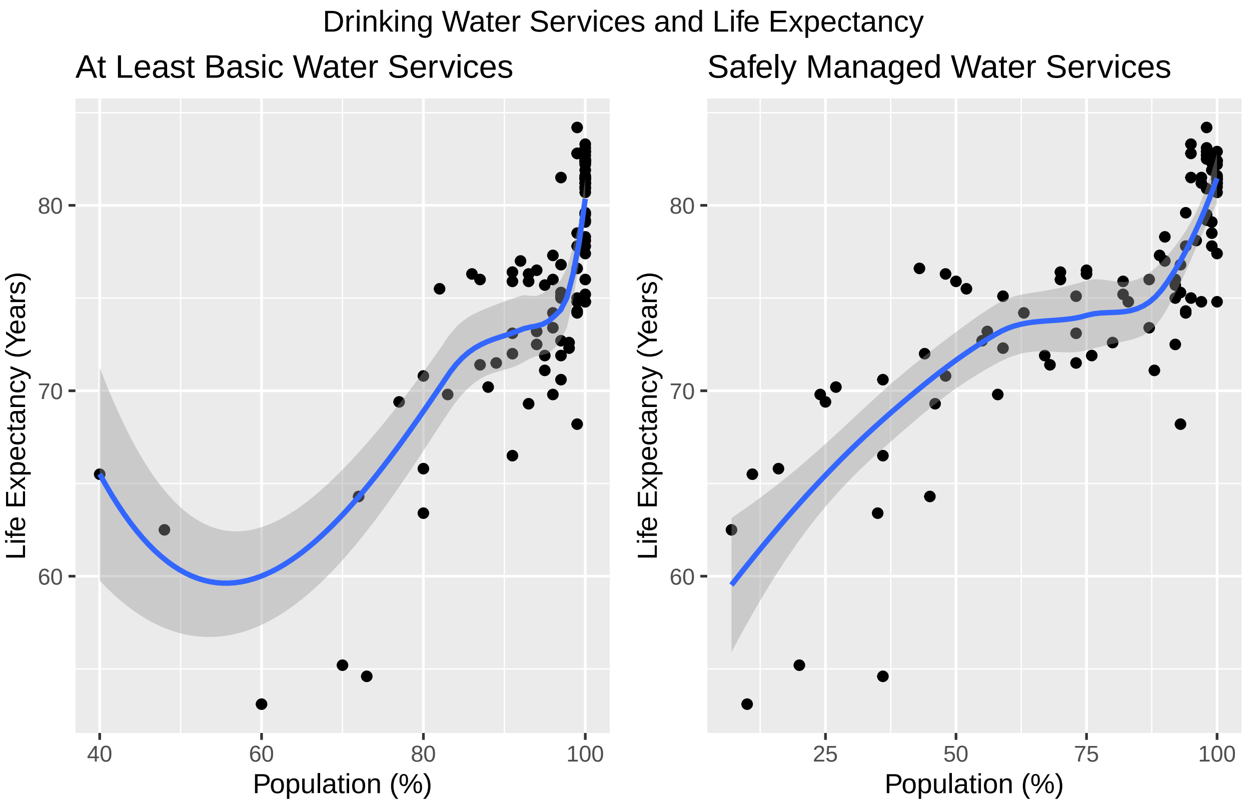
After loading my data and prepping it for analysis, I calculated the correlation between drinking water and life expectancy for both at a basic management level and at a safe management level. I also repeated this calculation for sanitation services and life expectancy. Then, I used ggplot2() and plotted a scatterplot and added a trend line using the smoothing technique. I arranged the plots for both levels for each variable side by side for comparison.

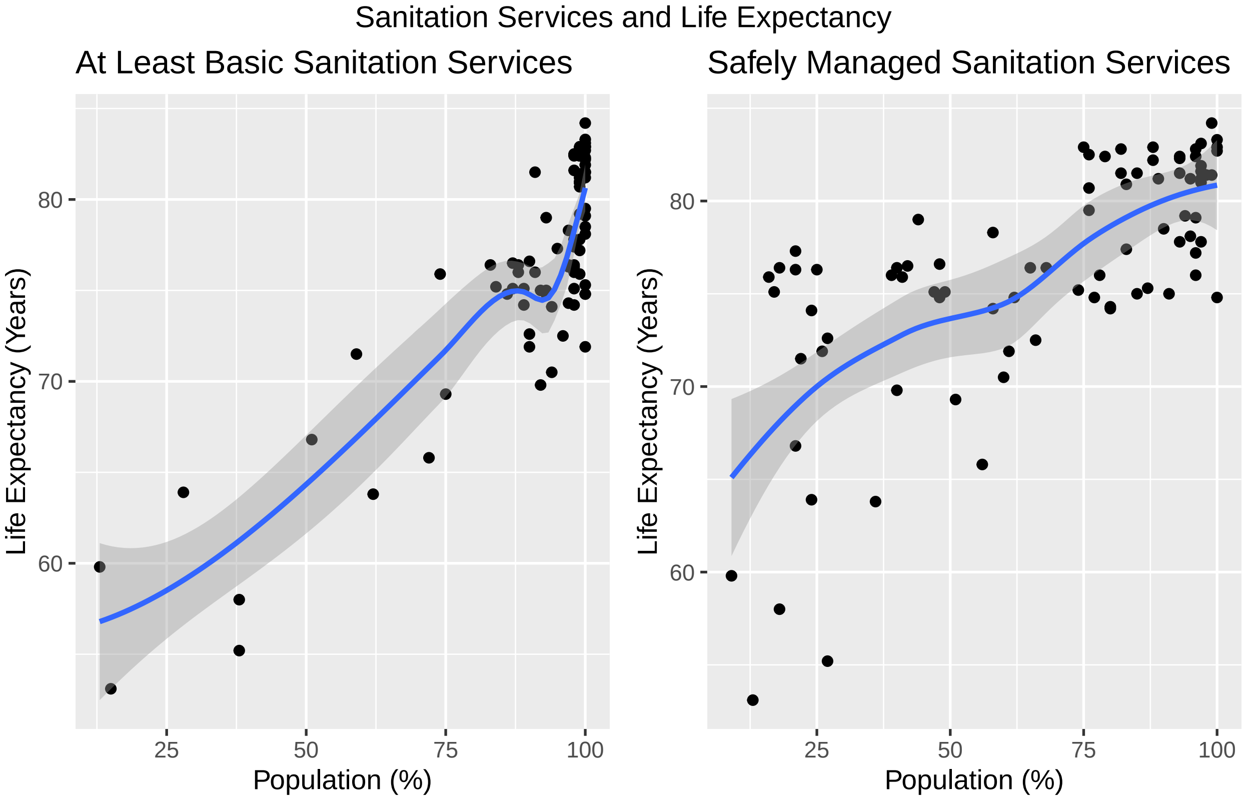
2. Results, Relationship, Insights

For drinking water services, at both management levels, there is a strong positive correlation with life expectancy. At the at least basic drinking water level the correlation was 0.75 while at the safely managed drinking water level, the correlation was 0.82. The smoothing using the loess method on the scatterplot below indicates an overall increasing trend as the greater the percent of the total population has access to the drinking water at either level of management, the life expectancy tends to increase as well.

For sanitation services, at least basic sanitation services had a very strong positive correlation with life expectancy with a 0.85 value. As for safely managed sanitation services, there is a moderately positive correlation with life expectancy with a 0.69 value. The smoothing using the loess method on the scatterplot below seems to indicate an overall upward direction suggesting that as the greater the percent of the total population has access to sanitation services at either level of management, the life expectancy tends to increase as well.

3. Data Visualization


Drinking Water Services and Life Expectancy
At Least Basic Water Services
Safely Managed Water Services
Life Expectancy (Years)
80
70
60
40
60
80
100
25
50
75
100
Population (%)


Sanitation Services and Life Expectancy
At Least Basic Sanitation Services
Safely Managed Sanitation Services
Life Expectancy (Years)
80
70
60
25
50
75
100
Population (%)