DATA 606 Data Project Proposal

Shaya Engelman

2023-10-28

Data Preparation

```
#load required packages
library(tidyverse)
library(RCurl)
library(Hmisc)
```

```
#load data
gdp_malnutrition <- read.csv('https://raw.githubusercontent.com/Shayaeng/Data606/main/Final%20Project/ma

gdp_malnutrition <- gdp_malnutrition[complete.cases(gdp_malnutrition), ]
```

Research question

What is the linear relationship between malnutrition rates and Gross Domestic Product (GDP) across different countries, and to what extent do changes in GDP predict changes in malnutrition rates?

Cases

What are the cases, and how many are there?

The cases are represented in the rows. They are a specific country's GDP and deaths due to malnutrition in a specific year.

Data collection

Describe the method of data collection.

The malnutrition data was collected from the Global Health Data Exchange and can be found here: http://ghdx.healthdata.org/gbd-results-tool. The GDP data was taken from The World Bank website and can be found here: https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators

Type of study

What type of study is this (observational/experiment)?

This study is observational.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

I found the data here: Hunger and Undernourishment - Our World in Data

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The dependent variable is the count of of deaths due to malnutrition in any given year. Since it is a count, it is a quantitative variable.

Independent Variable(s)

The independent variable is the GDP of any given country in any given year. It is also a quantitative variable.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
#for basic summary statistics
summary(gdp_malnutrition$Deaths_Protein.energy.malnutrition)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.010   0.330   2.830   8.721  11.200 142.530
```

```
summary(gdp_malnutrition$GDP_per_capita_PPP_.2017)
```

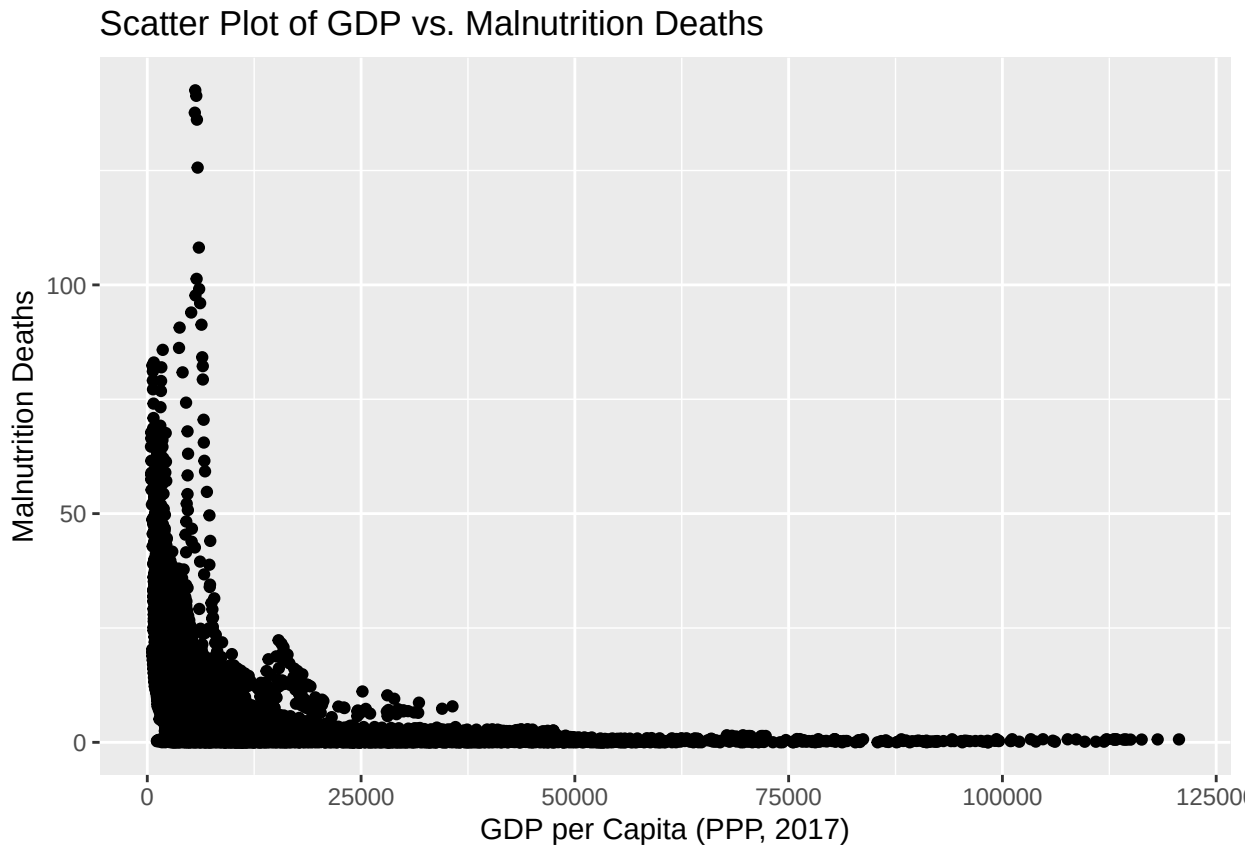
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    436.4   3456.2   9815.1  17087.2  23954.8 120647.8
```

The smallest amount of malnutrition deaths in the observed data is 0.01 per 100,00 people while the largest amount was 142.53 per 100,000 people. The median amount was 2.83 while the mean amount was 8.72. The IQR was 10.87 ranging from 0.33 to 11.2.

The smallest GDP per capita in the observed data was 436 while the largest GDP was 120648. The median GDP was 9815 while the mean was 17087. The IQR was 20499 ranging from 3456 to 23955.

In both variables the large gap between the median and mean imply a very uneven distribution of the data. This is also hinted at by the huge discrepencies between the maximum datapoint and the end of the third quartile. This can be visualized with the following scatterplot.

```
ggplot(gdp_malnutrition, aes(x = GDP_per_capita_PPP_.2017., y = Deaths_Protein.energy.malnutrition)) +
  geom_point() +
  labs(x = "GDP per Capita (PPP, 2017)", y = "Malnutrition Deaths", title = "Scatter Plot of GDP vs. Mal
```



This scatterplot does indeed show a much higher concentration of datapoints in the bottom left section. This means a much larger amount of lower values for both variables. However, it also does seem to be showing a relationship between the two variables.