Homework 1: Data Visualization

February 3, 2022

Answer the following questions in a .pdf or .docx, explaining all of your answers and putting any tables and figures in the document as necessary. When data is called for to answer applied questions, I will provide it in bblearn. Turn in your R code that created all of the tables and figures separately, and be sure that it runs from source in such a way that it loads the data and performs all the tests without me fiddling with it. Make sure to document your R source code using # comments if you want partial credit.

For these questions, we will be using some cross-sectional data from the human development report in 2013. I have also provided a .txt codebook that briefly describes the data and its source.

1. Please report the mean and median of life expectancy and gross national income for each country. What do the values of the mean and median of gross national income per capita mean about the distribution of the variable in the data? What do the values of the mean and median of life expectancy mean about the data?

2. Based on the mean value you calculated for life expectancy, can we conclude that the average person in the world lives that long? Why or why not?

3. What is the 4th highest country with respect to average years of schooling in this data from 2012? How about the country reporting the 5th lowest average years of schooling?

4. What are the standard deviation and variance of average years of school and expected years of schooling? Since both are measured in years, what do these numbers mean about the dispersion of the variables underlying relative to one another?

Now, we want to see if some of these basic HDI indicators vary by geographical region. Code the data by geographic region as defined by the United Nations (see `https://unstats.un.org/unsd/methodology/m49/#geo-regions`, and use the lowest regional level of aggregation (e.g. "Northern Africa"). I suggest doing it carefully by hand with excel, but there may be more automated solutions if you choose.

You do not need to turn in this new data file and the bits of code in your .R script for the questions that follow will be an exception to my requirement that your code work seamlessly on the dataset(s) I provide to you.

1. Report a table of the mean human development score index by geographic region. What is the lowest average human development score index? What is the highest?

2. Among the regions, which region has the highest standard deviation in life expectancy? In words, what does that mean practically?

3. Finally, I want to know what the average women's labor force participation is per country. I have provided a data file with women's labor force participation and some identifier variables I downloaded from the World Bank. Using your coding work from before, merge this data into the human development data. Make a table of the regional average women's labor force participation.

   (Remember the R coding lessons from your book with respect to missing data and the summarize commands—the new data has missing values that will need to be ignored. In addition, one of your regions has no women's labor participation data at all so won't be calculated.)