INFO201 Problem Set: rmarkdown and plotting

February 12, 2023

Instructions

This is a problem set about rmarkdown and plotting (using ggplot). Unlike the previous problem sets, this one does not give you a ready-made GH repo with a code file–it is now your taks to create a repo and include your rmarkdown file in there.

You should answer the questions below in that file, knit it, and submit both the compiled html and link to your repo on canvas.

- This problem sets asks you to write extensively when commenting your results. Please write clearly! Answer questions in a way that if the code chunks are hidden then the result is still readable!
  - All substantial questions need explanations. You do not have to explain the simple things like "how many rows are there in data", but if you make a plot of life expectancy, then you should explain what does the plot tell you.
  - Write explanations as markdown and use the styles like bold and italic as appropriate.
- Do not print too much results. It is all well to print a few lines of data for evaluaton/demonstration purposes. But do not print dozens (or thousands!) of lines–no one bothers to look at that many numbers. You will lose points for annoying others (here your graders, but later potentially your boss).
- Do not make code lines too long. 80-100 characters is a good choice. Your grader may not be able to follow all the code if the line is too long–most of us are using small laptop screens! (And again–you *want* to keep your graders happy!)

Gapminder data

We use gapminder dataset, downloaded from https://www.gapminder.org/data/, however, the data structure there is quire complex, please use the dataset provided on canvas (in files/data).

The variables are:
name country name
iso3 3-letter country code
iso2 2-letter country code
region broad geographic region
sub-region more precise region
intermediate-region
time year

totalPopulation total population
GDP_PC GDP per capita (constant 2010 US$)
accessElectricity Access to electricity (% of population)
agriculturalLand Agricultural land (sq. km)
agricultureTractors Agricultural machinery, tractors (count)
cerealProduction Cereal production (metric tons)
feritilizerHa Fertilizer consumption (kilograms per hectare of arable land)
fertilityRate total fertility rate (births per woman)
lifeExpectancy Life expectancy at birth, total (years)
childMortality Mortality rate, under-5 (per 1,000 live births)
youthFemaleLiteracy Literacy rate, youth female (% of females ages 15-24)
youthMaleLiteracy Literacy rate, youth male (% of males ages 15-24)
adultLiteracy Literacy rate, adult total (% of people ages 15 and above)
co2 CO2 emissions (kt)
greenhouseGases Total greenhouse gas emissions (kt of CO2 equivalent)
co2_PC CO2 emissions (metric tons per capita)
pm2.5_35 PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-1 value 36ug/m3 (
battleDeaths Battle-related deaths (number of people)

1 Load and check data (5pt)

You first task is to do a very simple data check:

1. (1pt) For solving the problems, and answering the questions, create a new rmarkdown document with an appropriate title. See https://faculty.washington.edu/otoomet/info201-book/r-markdown.html#r-markdown-rstudio-creating.
2. (2pt) Load data. How many rows/columns do we have?
3. (2pt) Print a small sample of data. Does it look OK?

---

2 Descriptive statistics (15pt)

1. (3pt) How many countries are there in the dataset? Analyze all three: *iso3*, *iso2* and *name*.
2. If you did this correctly, you saw that there are more iso-2 codes than names, and there are even more *iso3*-codes. What is going on? Can you find it out?
   (a) (5pt) Find how many names are there for each iso-2 code. Are there any iso-2 codes that correspond to more than one name? What are these countries?
   (b) (5pt) Now repeat the same for name and iso3-code. Are there country names that have more than one iso3-code? What are these countries?
   Hint: two of these entitites are *CHANISL* and *NLD CURACAO*.
3. (2pt) What is the minimum and maximum year in these data?

---

3 CO2 emissions (30pt)

Next, let's analyze CO2 emissions.

1. (2pt) How many missing co2 emissions are there for each year? Analyze both missing *CO2* and *co2_PC*. Which years have most missing data?

2. (5pt) Make a plot of total CO2 emissions over time for the U.S, China, and India. Add a few more countries of your choice. Explain what do you see.

3. (5pt) Now let's analyze the CO2 emissions per capita (*co2_PC*). Make a similar plot of the same countries. What does this figure suggest?

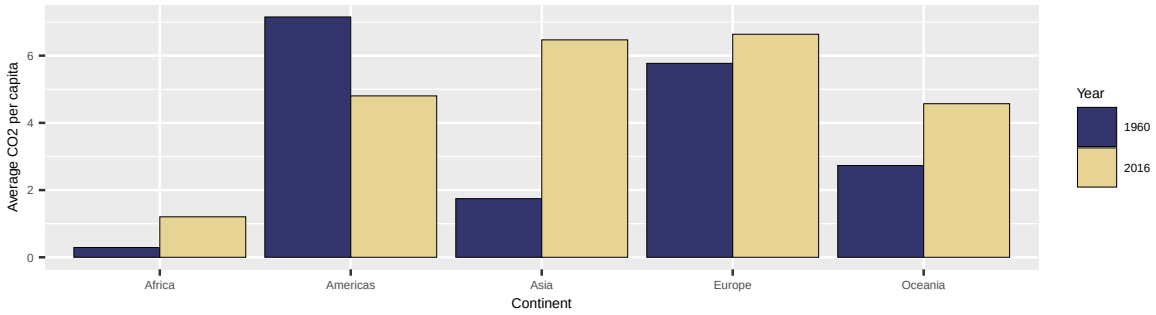
4. (6pt) Compute average CO2 emissions per capita across the continents (assume *region* is the same as continent). Comment what do you see.

   Note: just compute averages over countries and ignore the fact that countries are of different size.

   Hint: Americas 2016 should be 4.80.

5. (7pt) Make a barplot where you show the previous results–average CO2 emissions per capita across continents in 1960 and 2016.

   Hint: it should look something along these lines:



6. Which countries are the three largest, and three smallest CO2 emitters (in terms of CO2 per capita) in 2019 for each continent? (Assume *region* is continent).

---

4 GDP per capita (50pt)

Let's look at GDP per capita (*GDP_PC*).

1. (8pt) Make a scatterplot of GDP per capita versus life expectancy by country, using data for 1960. Make the point size dependent on the country size, and color those according to the continent. Feel free to adjust the plot in other ways to make it better.

   Comment what do you see there.

2. (4pt) Make a similar plot, but this time use 2019 data only.

3. (6pt) Compare these two plots and comment what do you see. How has world developed through the last 60 years?

4. (6pt) Compute the average life expectancy for each continent in 1960 and 2019. Do the results fit with what do you see on the figures?

   Note: here as *average* I mean just average over countries, ignore the fact that countries are of different size.

5. (8pt) Compute the average LE growth from 1960-2019 across the continents. Show the results in the order of growth. Explain what do you see.

   Hint: these data (data in long form) is not the simplest to compute growth. But you may want to check out the `lag()` function. And do not forget to group data by continent when using `lag()`, otherwise your results will be messed up! See https://faculty.washington.edu/otoomet/info201-book/dplyr.html#dplyr-helpers-compute.

6. (6pt) Show the histogram of GDP per capita for years of 1960 and 2019. Try to put both histograms on the same graph, see how well you can do it!

7. (6pt) What was the ranking of US in terms of life expectancy in 1960 and in 2019? (When counting from top.)

   Hint: check out the function `rank()`!

   Hint2: 17 for 1960.

8. (6pt) If you did this correctly, then you noticed that US ranking has been falling quite a bit. But we also have more countries in 2019–what about the relative rank divided by the corresponding number of countries that have LE data in the corresponding year?

   Hint: 0.0904 for 1960.

Finally tell us how many hours did you spend on this PS.