ps5

Alex Han

2023-02-18

INFO201 Problem Set: rmarkdown and plotting library(tidyverse)

1 Load and check data (5pt)

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?

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)


df <- read_delim("data/gapminder.csv")
```

```
## Rows: 13055 Columns: 25
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (6): iso3, name, iso2, region, sub-region, intermediate-region
## dbl (19): time, totalPopulation, fertilityRate, lifeExpectancy, childMortali...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
x <- nrow(df)
y <- ncol(df)

cat("It has", x, "rows, and",y, "columns")
```

```
## It has 13055 rows, and 25 columns
```

3. (2pt) Print a small sample of data. Does it look OK?

```r
df %>%
  sample_n(10)
```

```
## # A tibble: 10 x 25
##    iso3  name         iso2  region sub-r~1 inter~2  time total~3 ferti~4 lifeE~5
##    <chr> <chr>        <chr> <chr>  <chr>   <chr>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 PAK   Pakistan     PK    Asia   Southe~ <NA>     2006 1.64e8    4.31    64.2
##  2 AUS   Australia    AU    Ocean~ Austra~ <NA>     2001 1.94e7    1.74    79.6
##  3 MEX   Mexico       MX    Ameri~ Latin ~ Centra~  1973 5.63e7    6.23    62.8
##  4 BLR   Belarus      BY    Europe Easter~ <NA>     1983 9.84e6    2.11    70.1
##  5 PSE   Palestine, ~ PS    Asia   Wester~ <NA>     1965 NA        NA       NA
##  6 NER   Niger        NE    Africa Sub-Sa~ Wester~  2010 1.65e7    7.47    57.3
##  7 NRU   Nauru        NR    Ocean~ Micron~ <NA>     1988 9.06e3   NA       NA
##  8 NGA   Nigeria      NG    Africa Sub-Sa~ Wester~  1962 4.70e7    6.35    37.9
##  9 PRI   Puerto Rico  PR    Ameri~ Latin ~ Caribb~  2007 3.78e6    1.65    78.4
## 10 LBY   Libya        LY    Africa Northe~ <NA>     1986 3.99e6    6.02    67.0
## # ... with 15 more variables: childMortality <dbl>, youthFemaleLiteracy <dbl>,
## #   youthMaleLiteracy <dbl>, adultLiteracy <dbl>, GDP_PC <dbl>,
## #   accessElectricity <dbl>, agriculturalLand <dbl>, agricultureTractors <dbl>,
## #   cerealProduction <dbl>, fertilizerHa <dbl>, co2 <dbl>,
## #   greenhouseGases <dbl>, co2_PC <dbl>, pm2.5_35 <dbl>, battleDeaths <dbl>,
## #   and abbreviated variable names 1: `sub-region`, 2: `intermediate-region`,
## #   3: totalPopulation, 4: fertilityRate, 5: lifeExpectancy
```

The data looks OK.

2 Descriptive statistics (15pt)

1. (3pt) How many countries are there in the dataset? Analyze all three: iso3, iso2 and name.

```r
count(unique(df['iso3']))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   253
```

```r
count(unique(df['iso2']))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   249
```

```r
count(unique(df['name']))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   250
```

2. If you did this correctly, you saw that there are more names than iso-2 codes, 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.