DS5110 Homework 3

Kylie Ariel Bemis

5 October 2022

Instructions

Your solutions should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. *Make sure that you answer all parts of the problem.*

Submit your solutions on Canvas by the deadline displayed online. For full credit, your submission must include exactly two files:

- R Markdown (.Rmd)
- Knitted PDF report (.pdf)

Problems must appear in order, and problem numbers must be clearly marked. Any written responses should appear outside of code blocks and use Markdown for text formatting. Code comments are encouraged, but will be ignored for grading purposes. Solutions that are especially difficult to grade due to poor formatting will not receive full credit.

All solutions to the given problems must be your own work. If you use third-party code for ancillary tasks, you **must** cite them.

Part A

Problems 1–2 investigate air quality in the United States from 1980-2019. Download the data files from "epa-aqi-data-annual.zip" on Piazza. The original datasets can be found on the EPA website (https://www.epa.gov/outdoor-air-quality-data). The data comes from air monitoring stations that record the different air quality measures, and calculate an overall air quality index (AQI). Higher values of AQI indicate worse air quality, with AQI values greater than 200 being considered "very unhealthy" for anyone outdoors. The provided dataset summarizes the data from each monitoring station by year.

Problem 1

Import and combine the data files for all years from 1980-2019. Then visualize how the "Median AQI" values have changed over time from 1980 through 2019. What do you observe about how AQI has changed over time?

Hint: The `dplyr::bind_rows()` function may be useful.

Problem 2

We would like to create chloropleths (color-coded maps) showing the average AQI in each state for each decade (1980-1989, 1990-1999, etc).

Create a new variable for "decade", and then calculate the mean "Median AQI" values for each state and decade.

Create a chloropleth for each decade, each showing the average AQI in each state using fill color. How has AQI changed over the decades and across states?

Hint: The `ggplot2::map_data()` function may be useful.

Part B

Problems 3–5 use the complete Gapminder dataset. Download the data files from "ddf–gapminder–systema_globalis-master.zip" on Piazza. The original datasets can be found on Github (https://github.com/open-numbers/ddf--gapminder--systema_globalis). The data is divided into tables of "entities" and "datapoints". The "entities" tables provide dictionaries of terms, countries, regions, etc. The "datapoints" tables contain single variables measured for each country and year.

Problem 3

We would like to investigate differences between different regions of the world, but there are too many countries to do this by country. Instead, we will use the regions provided by the "world_4region" table.

Import and join the "country" and "world_4region" entity tables, keeping only countries that have been assigned a region. Visualize the count of countries in each region.

Note: In all problems in this part, use the "name" column from the "world_4region" table to label the regions in plots, and do not include countries without an assigned region in your analysis.

Problem 4

We would like to see how worldwide infant mortality rates have changed over time. Import the "infant_mortality_rate_per_1000_births" datapoints table. Then visualize infant mortality rate over time, by world region.

How has infant mortality rate changed over time in each of the four world region? Is there anything else notable that you see in the visualization?

Problem 5

We would like to see the impact that infant mortality rate has on life expectancy. Import the "life_expectancy_years" datapoints table. Then visualize the relationship between infant mortality and life expectancy, by world region.

Is there a relationship between infant mortality rate and life expectancy? If so, describe the relationship. Is this consistent across regions?