Week 6 Assignment

PA 434

2/16/2020

Data

For this assignment, you are going to use a simulated dataset from the stevedata package, which I modified to better fit the purpose of this assignment.

You can download the csv file called "api_data" from Blackboard.

About the data

Source: This is a simulated dataset based off the apipop data in the survey package, which were drawn from survey sampling help pages of UCLA Academic Technology Service. Data were simulated Steven V. Miller, an associate professor in the Department of Political Science at Clemson University.

Population: Data represents an hypothetical universe (i.e, full population) of 10,000 schools in an hypothetical state.

Variables

There are 7 columns in the dataset:

uid = unique identifier for schools

county = a character vector for the county, named after an Ohio State All-American.

community_schooltype = a character vector for the school's community, either rural, suburban, or urban, and the school type, whereas E = elementary school, M = middle school, and H = high school

api = a numeric vector vector an academic performance index for the school

variable_name = a character vector include information on:

1. **meals** = school students eligible for subsidized meals
2. **colgrad** = school parents with college degrees
3. **fullqual** = fully qualified teachers (i.e., completed all qualification requirements)

percentage = percentage referred to the variable in the variable_name column.

Step 1 - Before starting your analysis

Before starting your exploratory analysis, consider these questions

1. What is the unit of analysis? In other words, what should each row represent?
2. Is the data "tidy"? Why so? Why not? If not, make the appropriate modifications to tidy your data.
3. Briefly report anything that the reader should now about missing data in the dataset.

Step 2 - Exploratory data analysis using plots

When creating graphs for this section, make sure to double-check your axes and properly use labels and titles.

There is no need to attache your plots. Just send the code and we will run it to check out the graph (your script needs to be very clean).

1. Use plots to explore the distribution or frequency of variables in your dataset. The variable *uid* and *county* can be excluded from the analysis. Report your findings and considerations.
2. Plot the correlation between the performance index and parents' education.

2a. State your expectations before drawing the graph and discuss what you have learned from the graph.

2a bis.[OPTIONAL] If you are confortable with it, add a regression line to your graph.

2b. Finalize your graph by showing the breakdown by community type and then split the graph according to the type of school.

2c. Summarize your findings.

3. Create a bar plot representing the 10 schools with the highest academic performance index and the 10 schools with the lowest academic performance. Make sure that they are colored differently and ranked from the highest API to the lowest. Add a line representing the overall API average across all schools. **Hint:** you first need to use dplyr functions to create the data you need.
4. A state policymakers is trying to idenfity which counties should receive funding priority in order to improve their average performance index.

4a. Think about a couple of criteria to select priority counties

4b. Create a plot showing which countries should be prioritized and why.

Note: You are welcome to explore anything else that captures your attention and attach the code + explanation for review.