DS 5500 Homework 3 - Due Nov. 11

Instructions

Create a new public Github repository for this homework assignment. The repository should include all of the code necessary to reproduce your submitted solutions. Do not include the raw data in the repository.

Use the README.md of the Github repository to present your solutions. Answer all questions completely for full credit, including figures and tables where appropriate. For each problem, either provide relevant inline code snippets, or cite the source file where the relevant code lives (with line numbers if appropriate).

Describe any data processing steps (transformation, filtering, etc.) you perform when solving each problem, providing reasoning where appropriate. You may need to be creative when deciding how to approach each problem, as there may not be a single "correct" solution.

Your solutions should be posted as a ***public*** *note* on Piazza in the *hw3* folder with the title "[hw3] - your name name". Include in the body of the note a link to the Github repository with your solutions.

Overview

In this homework assignment, you are a data scientist working for the U.S. federal government. Due to budget cuts for education, your office has been tasked with cutting federal funding to some number of school districts. Your supervisor has asked you to develop a recommendation and objective justification for their decision using data analysis.

Due to related budget cuts, the most recent fiscal data you have to work with is from 2015-16. Download the 2015-16 district-level fiscal data from the National Center for Education Statistics' Common Core of Data:

- https://nces.ed.gov/ccd/f33agency.asp

For helping you make your decision, it may be helpful to have some performance metrics for each district. You can download the 2015-16 data for distrct-level statistics on graduation rate and state assessments on mathematics and reading/language arts from the EDFacts website:

- https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html

These datasets can be linked based on the LEA IDs.

Problem 1

Import and explore the district-level fiscal data from 2015-16.

Rank and visualize the states that take in the most federal funding (revenue).

Which states spend the most federal funding per student?

Problem 2

Visualize the relationship between school districts' total revenue and expenditures.

Which states have the most debt per student?

Problem 3

The district-level performance metrics from EDFacts may be useful in your decision.

However, to protect student privacy, the data in these datasets has been heavily "blurred" to prevent students from being identified. Therefore, most of the numeric metrics are presented as ranges in string format. In addition, censored and missing data must be imputed.

Write and explain a function for processing a single column of "blurred" metrics into usable numeric values.

Use it to process and then visualize the distribution of a performance metric of your choice.

Problem 4

You are tasked with cutting 15% of the U.S. federal budget currently being spent on funding school districts. How much money is this?

Choose which school districts will have their funding cut and how this will be done.

(You should produce a table of LEA IDs and the dollar amount by which their federal funding will be cut – you do *not* need print the entire table.)

Problem 5

Provide a statement for your supervisor justifying your decisions on which school districts will lose funding.