DS 5500 Homework 4 - Due Nov. 25

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 *hw4* folder with the title "[hw4] - your name name". Include in the body of the note a link to the Github repository with your solutions.

Overview

This homework assignment continues to use school data from the previous assignment.

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

The 2015-16 performance statistics on graduation rate and state assessments on mathematics and reading/language arts are available from the EDFacts website:

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

Additionally, the 2015-16 district-level "universe" survey data provides statistics about enrollment disaggregated by demographics like gender, race, disability status, etc:

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

These datasets can be linked based on the LEA IDs.

Problem 1

For the districts you selected for budget cuts in HW 3 Problem 4, calculate and visualize the proportion of each district's total funding that will be lost.

Which districts will be affected by your budget cuts the most?

Problem 2

A common problem with purely data-driven solutions is that they can inadvertently perpetuate hidden pre-existing biases in the data, and further disadvantage groups that are already disadvantaged.

Calculate the proportion of enrolled students by race for each district, then visualize the distributions of these for districts that received budget cuts versus districts that did not receive budget cuts.

Comment on whether the the distributions appear to be the same or different. Did your selection include any hidden biases, or manage to avoid them?

Problem 3

Calculate the proportion of enrolled students by disability status (students with an IEP under IDEA) for each district, then visualize the distributions of these proportions for districts that received budget cuts versus districts that did not receive budget cuts.

Comment on whether the the distributions appear to be the same or different. Did your selection include any hidden biases, or manage to avoid them?

Problem 4

Choose and critique one of your fellow classmates' selection of schools for budget cuts in HW 3 Problem 4 and Problem 5. What was the justification of their selection? Discuss any advantages or disadvantages of their approach.

Problem 5

Summarize and comment on what you learned from one the special topics lectures (MapReduce + Hadoop, Visualization, Causal Inference, or the Industry Panel) of your choice.