DS5110 Homework 3

Kylie Ariel Bemis

5 February 2024

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 1–3 use data from the US Department of Education's Civil Rights Data Collection. It was downloaded from the zipped 2017-2018 data available at https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2017-18.html. The Public Use Data File User's Manual and a spreadsheet describing the file structure are included in the zipped files, or can be downloaded at the same location. Use these as a reference to help you understand the dataset. The CRDC data is supplemented by statistical data from EDFacts (not included). We will use only the CRDC data. Import all CRDC reserve codes as missing values.

Problem 1

We would like to know the distribution of students by race and gender across all schools. Calculate and visualize the overall proportions of enrolled students of every race and gender combination out of the total number of students across all schools. Describe the distribution.

Problem 2

We would like to know the distribution of Advanced Placement (AP) students (i.e., students enrolled in at least one AP course) by race and gender across all schools. Filter the data to include only schools with AP programs. Calculate and visualize the overall proportions of AP students of every race and gender combination out of the total number of AP students across all schools. Describe the distribution. How does it compare to the distribution from Problem 1?

Problem 3

We would like to visualize whether there is a trend of students of color (i.e., non-white students) being underrepresented in AP programs at schools. For each school, calculate (1) the proportion of students of color out of all enrolled students at the school and (2) the proportion of students of color in at least one AP class out of all students in at least one AP class. Visualize the former as an independent variable against the latter as a dependent variable (you may choose to include a smooth line as well). Are students of color typically underrepresented in AP classes?

Hint: It may help to include a reference line with slope 1 using `geom_abline()`.

Part B

Problems 4–5 use a subset of the DBLP database of bibliographic information on major computer science journals and proceedings, originally available from https://data.mendeley.com/datasets/3p9w84t5mr. The dataset has been processed to include predictions of the author's genders (male or female only) using the open-source Genderize API. The processed data has been made available in a zipped folder that includes the SQLite database and accompanying documentation. We are primarily interested in the "general" and "authors" tables.

Problem 4

Filter the data to include only the authors for whom a gender was predicted as 'male' or 'female' with a probability of 0.90 or greater, and then visualize the total number of *distinct* male and female authors published each year. Comment on the visualization.

Problem 5

Still including only the authors for whom a gender was predicted with a probability of 0.90 or greater, create a stacked bar plot showing the *proportions* of distinct male authors vs. distinct female authors published each year. (The stacked bars for each year will sum to one.) Comment on the visualization.