DS5110 Homework 2 - Due Feb. 1

Kylie Ariel Bemis

1/11/2019

Instructions

Create a directory with the following structure:

- `hw1-your-name/hw1-your-name.Rmd`
- `hw1-your-name/hw1-your-name.pdf`

where `hw1-your-name.Rmd` is an R Markdown file that compiles to create `hw1-your-name.pdf`.

Do not include data in the directory. Compress the directory as `.zip`.

Your solution 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`. Missing values and overplotting should be handled appropriately. Axes should be labeled clearly and accurately.

To submit your solution, create a new private post of type "Note" on Piazza, select "Individual Student(s) / Instructor(s)" and type "Instructors", select the folder "hw1", go to Insert->Insert file in the Rich Text Editor, upload your `.zip` homework solution. Title your note "[hw1 solutions] - your name" and post the private note to Piazza. **Be sure to post it only to instructors**

Part A

Problems 1–2 use data from the Navajo Nation Water Quality Project. Download the CSV file from http://navajowater.org/export-raw-data/.

Water quality is a major issue on American Indian reservations in the southwestern United States. The prevalence of uranium mines and uranium mill accidents mean that much of the water in the Navajo Nation is irradiated, and many homes are left without clean, drinkable water. Multiple environmental agencies routinely sample water in the region and report on contaminants.

Read the documentation for the function `read_csv()` from the `readr` package, and use it to import the dataset into R.

Problem 1

The concentration of radiactive elements in a sample is measured in rate of atomic disintegrations per volume, rather than mass per volume, as used for stable isotopes. This is done by counting the number of atomic disintegrations per minute and comparing it to the mass of the material involved. However, laboratory environments and instruments used for detection create some number of atomic emissions on their own, so background correction must be performed. Because this process involves sampling many times, and the background can be inconsistent, resulting in over-correction, sometimes negative values are reported for the concentration. For practical purposes, these values can be considered zero.

Mutate the dataset to replace the negative values of Radium-228 with 0, then filter the dataset to remove any sites with "Unknown Risk" for the EPA risk rating.

Visualize the distribution of Radium-228 within each EPA section and each risk level. State your observations.

Problem 2

Install the `maps` and `mapproj` packages (you do not need to load them) and use the `ggplot2::map_data()` function to get data for drawing the "Four Corners" region of the United States (i.e., Arizona, New Mexico, Utah, and Colorado).

Install the `measurements` package and use the `measurements::conv_unit()` function to convert the latitude and longitude information in the dataset to decimal degrees suitable to be used for plotting.

Create a map of the region (you may want to adjust the plotting limits to an appropriate "zoom" level) showing the locations of the water sampling sites, along with the EPA risk and the concentration of Radium-228 for each location (mapped to an appropriate aesthetic).

Part B

Problems 3–5 use data from the US Department of Education's Civil Rights Data Collection. Download the zipped 2015-2016 data from https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2015-16.html. The Public Use Data File User's Manual should be included in the zipped files, or can be downloaded at the same location. Use it as a reference to help you understand the dataset.

Read the documentation for the function `read_csv()` from the `readr` package, and use it to import the dataset into R. Check the User's Manual for how missing values were reported, and handle them appropriately. Treat all reserve codes as missing.

Problem 3

We would like to investigate whether Black students receive a disproportionate number of in-school suspensions.

Create a new `data.frame` with the following columns:

- The total number of students enrolled at each school
- The number of Black students enrolled at each school
- The total number of students who received one or more in-school suspension (including non-disabled students and disabled students served by IDEA)
- The number of Black students who received one or more in-school suspension (including non-disabled students and disabled students served by IDEA)
- The proportion of Black students at each school among all students
- The proportion of students who received one or more in-school suspension who are Black among all suspended students

Plot the proportion of Black students at each school (on the x-axis) versus the proportion of suspended students who are Black (on the y-axis). Include a smoothing line on the plot.

What do you observe in the plot? Does the plot indicate an over- or under-representation of Black students in in-school suspensions?

Calculate the overall proportion of Black students across all schools and the overall proportion of suspended students who are Black across all schools.

Are Black students over- or under-represented in in-school suspensions?

Problem 4

We would like to investigate whether disabled students are more often disciplined with corporal punishment.

Create a new `data.frame` containing only schools that use corporal punishment with the following columns:

- The total number of students enrolled at each school
- The number of disabled students (served by IDEA) at each school
- The total number of students who were disciplined with corporal punishment
- The number of disabled students (served by IDEA) who were disciplined with corporal punishment
- The proportion of disabled students (served by IDEA) at each school among of all students
- The proportion of students who were disciplined with corporal punishment who are disabled (served by IDEA) among all disciplined students

Plot the proportion of disabled students at each school (on the x-axis) versus the proportion of disciplined students who are disabled (on the y-axis). Include a smoothing line on the plot.

What do you observe in the plot? Does the plot indicate an over- or under-representation of disabled students among students who are disciplined with corporeal punishment?

Calculate the overall proportion of disabled students across all schools and the overall proportion of disciplined students who are disabled across all schools.

Are disabled students over- or under-represented in corporal punishment?

Problem 5

We would like to investigate whether Black and Hispanic students are over- or under-represented in Gifted & Talented programs.

Create a new `data.frame` containing only schools with a Gifted & Talented program with the following columns:

- The total number of students enrolled at each school
- The number of Black and Hispanic students at each school
- The total number of students in the school's GT program
- The number of students in the GT program who are Black or Hispanic
- The proportion of students at each school who are Black or Hispanic among all students
- The proportion of students in the GT program who are Black or Hispanic among students in the GT program

Plot the proportion of Black and Hispanic students at each school (on the x-axis) versus the proportion of GT students who Black or Hispanic (on the y-axis). Include a smoothing line on the plot.

What do you observe in the plot? Does the plot indicate an over- or under-representation of Black and Hispanic students in Gifted & Talented programs?

Calculate the overall proportion of Black and Hispanic students across all schools and the overall proportion of GT students who are Black or Hispanic.

Are Black and Hispanic students over- or under-represented in Gifted & Talented programs?