DS5110 Homework 1 - Due Sept. 26

Kylie Ariel Bemis

9/12/2017

Instructions

Create a directory with the following structure:

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

where `hw1-your-name.Rwd` 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 1–3 use the `flights` dataset from the `nycflights13` package, which includes data for all flights that departed New York City (JFK, LGA, or EWR) in 2013.

Problem 1

Create side-by-side boxplots of the distances flown between airports for each carrier. Do some airlines routinely fly longer routes than others?

Problem 2

For each destination, calculate the proportion of arriving flights delayed by one hour or more, and then plot these values against the mean distance flown to each destination.

Edit the plot to add a smoothed line and a visual representation of the total number of flights for each destination

What does the plot tell you about the relationship between flight distance and delays?

Problem 3

Create two bar plots that characterize each carrier by their delays. One should show the proportion of arriving flights delayed by one hour or more, and the other should show the mean arrival delay time for each carrier.

Which airlines have the worst delays? Which are the most consistently on time or ahead of schedule?

Part B

Problems 4–6 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 `tidyverse` function `read_csv`, and use it to import the dataset into R.

Problem 4

Create histograms showing the distribution of the amount of Radium-228 in water samples for each EPA section. Do you notice anything odd?

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.

Filter the dataset to remove the zero values and create the histograms again, using a different combination of `ggplot2` functions this time.

Problem 5

Create bar plots showing the mean concentrations of Uranium-234, Uranium-235, and Uranium-238 in water samples at each EPA section for each EPA risk rating, using facets for each EPA section.

Create another set of bar plots with the same information, but without facets, using color to indicate EPA risk rating instead.

Problem 6

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

Plot a map of the region (you may want to adjust the plotting limits to an appropriate "zoom" level), and overlay the locations of the water sampling sites on the map, using appropriate visual representations to indicate the EPA risk rating and the amount of Uranium-238 measured at each site.

Part C

Problems 7–10 use data from the US Department of Education's Civil Rights Data Collection. Download the zipped 2013-2014 data from https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2013-14.html. Download the Public Use Data File User's Manual at the same location.

Read the documentation for the `tidyverse` function `read_csv`, and use it to import the dataset into R. Check the User's Manual for how missing values were reported, and handle them appropriately.

Problem 7

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 total number of Black students enrolled at each school
- The total number of students who received one or more in-school suspension
- The number of Black students who received one or more in-school suspension
- The proportion of students at each school who are Black
- The proportino of students who received one or more in-school suspension who are Black

Plot the proportion of Black students at each school versus the proportion of suspended students who are Black. 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.

Problem 8

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 total number of disabled students (under IDEA and/or 504) at each school
- The total number of students who were disciplined with corporal punishment
- The number of disabled students who were disciplined with corporal punishment
- The proportion of students at each school who are disabled
- The proportion of students who were disciplined with corporal punishment who are disabled

Plot the proportion of disabled students at each school versus the proportion of disciplined students who disabled. 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.

Problem 9

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 total 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
- The proportion of students in the GT program who are Black or Hispanic

Plot the proportion of Black and Hispanic students at each school versus the proportion of GT students who Black or Hispanic. 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.

Problem 10

Develop your own question about whether a particular demographic is over- or under-represented in a particular aspect of the education system.

State your question.

Process, plot, and summarise the data to answer your question. State what you observe in the plot and your conclusions based on the plot and the summary statistics.