DS5110 Homework 1 - Solutions

Kylie Ariel Bemis

1/14/2018

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 1–3 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 a bar plot showing the number of flights flown out of New York airports by each carrier in 2013. Which airline carrier flew the most flights?

Problem 2

For each destination, calculate the proportion of flights that arrived at their destination earlier than scheduled. Also calculate the median distance flown to each destination.

Plot the proportion of early arrivals (on the y-axis) against the median distance flown (on the x-axis) for each destination. Add a smooth line to the plot. Based on the smooth line, at what distances are flights most likely to arrive early? Describe the relationship between early arrivals and flight distance.

Problem 3

Create two bar plots that characterize each carrier by how early their flights arrive. One should show the proportion of flights that arrive early for each carrier, and the other should show the median number of minutes early that flights arrivee for each carrier.

Which airlines are the most consistently ahead of schedule? Which airlines arrive the most early?

Which airlines are most consistently behind schedule? Which airlines arrive the latest?

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 (use faceting). Do you notice anything odd? (Besides the fact that the water samples are radioactive in the first place?)

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 with 0, and then create the histograms again, using a different combination of `ggplot2` functions this time.

Problem 5

Filter the dataset to remove any sites with "Unknown Risk" for the EPA risk rating.

Count the number of sites of each EPA risk rating in each EPA section, and then calculate the mean concentration of Uranium-238 in the water samples for each EPA risk rating in each EPA section.

Plot the number of sites at each EPA section using a bar plot, using the fill color of the bars to indicate the risk rating, and then plot the mean concentrations of Uranium-238 for each EPA section using a bar plot, using the fill color of the bars to indicate the risk rating.

Which EPA section(s) have the most sites with "More Risk"? Which EPA section(s) have the sites with the highest concentration of Uranium-238 on average?

Problem 6

Install the `maps` package (you do not need to load it) and use the `ggplo2::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.

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. Use color to indicate the EPA Section and size to indicate 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 expulsions under zero-tolerance policies.

Create a new `data.frame` or `tibble` 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 an expulsion under zero-tolerance policies
- The number of Black students who received an expulsion under zero-tolerance policies
- The proportion of students at each school who are Black
- The proportion of students expelled under zero-tolerance policies who are Black

Filter the data to include only those schools in which at least one student received an expulsion under zero-tolerance policies.

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

What do you observe in the plot? Does the plot indicate an over- or under-representation of Black students in expulsions under zero-tolerance policies?

Calculate the overall proportion of Black students across all schools and the overall proportion of students expelled under zero-tolerance policies who are Black across all schools.

Problem 8

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

Create a new `data.frame` or `tibble` 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 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 Hispanic
- The proportion of students at each school who are Hispanic
- The proportion of students in the GT program who are Hispanic

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

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

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

Problem 9

We would like to investigate whether disabled students are more often referred to a law enforcement agency or official.

Create a new `data.frame` or `tibble` 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

Filter the data to include only those schools without errors in data entry (i.e., remove all schools with more disabled students enrolled than the total number of enrolled students).

Plot the proportion of disabled students at each school (on the x-axis) versus the proportion of students referred to law enforcement 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 referred to law enforcement?

Calculate the overall proportion of disabled students across all schools and the overall proportion of students referred to law enforcement who are disabled across all schools.

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.