Problem Set #7 Instructions

November 9, 2018

Reading for next week

- [Recommended but not required]GW Chapter 13 (relational data)

General instructions for R script

In this homework, you will submit an `R script`. Please make sure that your R script runs without any errors before submitting.

- General Instructions for Problem Sets Here
- Open R script with RStudio (may open with R or other application)
- Can create "comments" by using "#"
- Shortcuts for executing commands
- Cmd/Ctrl + Enter: execute highlighted line(s)
- Cmd/Ctrl + Shift + Enter (without highlighting any lines): run entire script
- Output from commands executed from R script
- Output will appear in the "console"

Broad overview and learning goals

Broad overview of problem set

The goal of this problem set is to create an analysis dataset that contains: (A) NCES data on the characteristics of public high schools; (B) Census data on characteristics of the zip-code the high school is located in; and (C) data on how many visits each high school received from one of the universities in the "off campus recruiting project." After you merge all these data sources together, you will conduct some exploratory data analyses of your own choosing to investigate the characteristics (e.g., school, community) associated with receiving or not receiving a visit from the university. Below we will provide specific instructions about the steps we expect you to take to complete the problem set.

Data sources you will work with in this problem set

- Off-campus recruiting data
  - one observation per university recruiting event
- American Community Survey
  - one observation for each zip-code
- Common Core of Data
  - one observation for each public school

The R script we provided reads in these data sources, labels variables, and does some data cleaning. You will run all this code as is. After you run this code, you will have the following data frame objects:

- `events_per_pubschool`
  - For each public university in the sample, this data frame has one observation for each public high school that received a visit. The variable `num_events` counts the number of visits that high school received
  - Note that we deleted visits to other location types (e.g., private high schools, community colleges)
- `zip_data`
  - Census data from the American Community survey on the characteristics of each zip-code, with one observation per zip-code
- `ccd`
  - Data on the characteristics of public schools in the US, with one observation per school (the variable `ncessch` uniquely identifies schools)

These objects will be the basis of your problem set answers. We labelled all variables, but if you need additional resources, we provided codebooks and links to some of the data. If you get stuck, consult your group, post on Piazza, or stop by at our office hours.

General guidelines for merges/joins

In this problem set you will perform several joins. For each join you **must** follow these steps

- Prior to merging, investigate the two data frames
  - Try to identify variable(s) that uniquely identify object or dataframe
  - Based on this investigation, decide which variable (or variables) is the "key" variable for the merge
  - You may find it helpful to print a few observations of the data frame (or view with the `View()` function) to help you understand data structure
- Join the two data frames
  - Note: for some joins, you may have to rename a "key" variable; you can do this prior to the join or within the join code [see section 3.2.2 of lecture 7]
- After joining, spend some time investigating the quality of the join
  - You **must** use an `anti_join()` to identify observations that did not merge
  - Beyond that, just conduct any exploratory data analyses you think might provide insight about why some observations did not "match"
  - You may find it helpful to print selected observations or view object with `View()` function
  - Do not spend more the 15-20 minutes investigating a single join

Specific instructions (steps to complete the problem set)

1. Run all the code we created for you
   - It may be helpful to run one line at a time, just to see what we are doing to read-in and clean data
   - Note: some variables we create, we ended up deleting because we wanted to make homework shorter so some of the code you run is superfluous to the problem set.
2. Do some preliminary investigations of the three data frames to become acquainted with them. Whatever helps you feel comfortable with the data (no need to spend more than 10-20 minutes on this)
   - e.g., identify variable names, variable labels, print some observations, etc.
   - Note: In completing subsequent steps of the problem set, you may find it helpful to conduct additional investigations of the data
3. You will create an object based on `ccd` (the data frame with one observation per public school) called `ccd_hs` that only contains observations for high schools that meet **all** of the following (admittedly arbitary) criteria:
   - Grade 12 is offered (g12offered)
   - Enrolls at least 10 students in the 12th grade (g12)

- Is not a virtual school (virtual)
- Located in the 50 U.S. states, the District of Columbia, or land regulated by the Bureau of Indian Affairs (fipst)
  - Hint: `fipst < 52`
- Is a regular school or vocational school (sch_type)
  - Hint: `sch_type %in% c(1,3)`
- Updated status is open, new, or reopened (updated_status)
  - Hint: `updated_status %in% c(1,3,8))`

4. Using `ccd_hs` as the `x` table `zip_data` as your `y` table, perform a `left_join()`, assigning the resulting object the name `ccd_hs_zip`. Basically, the resulting data frame will have one observation per high school and will have additional variables for characteristics of the zip-code that high school is located in.
5. By applying a `filter()` to the data frame `events_per_pubschool`, keep only the observations for one university (e.g., UC Berkeley, University of Alabama) and assign this object the name `events_nameofinstitution` (e.g., `events_ucberkeley`)
   - Hint on how to filter one university: `filter(univ_id == number)`
   - Choose any university you are interested in
   - This will be the university that you conduct exploratory data analyses on after merging with data on the characteristics of public high schools.
6. Using `ccd_hs_zip` as your `x` table and `events_nameofinstitution` as your `y` table, perform a `left_join()`, assigning the resulting object the name `ccd_hs_zip_events`
7. Create a variable called `num_visits` that identifies the number of visits each high school received from the university you chose to keep.
   - **Hints**: The variable `num_visits` will be based on the input variable `num_event`. Essentially `num_visits` should have the same value as `num_event` for observations where `num_event` is not equal to `NA`. For observations where `num_event` is equal to `NA` the variable `num_event` should equal `0`. Why? Observations where `num_event` equals `NA` are high schools that did not receive a visit from the university (i.e., no observation for these high schools in the data frame `events_zip_nameofinstitution`)
8. Based on the variable `num_visits` you just created, create a 0/1 variable `got_visit` that equals `1` if the high school got at least one visit and equals `0` if the high school did not receive any visits.
9. Perform exploratory data analysis on variables you find interesting, with the general focus of identifying characteristics associated with getting visit(s) versus not getting visit(s) (Spend between 45 minutes to 1 hour on this).

Prior to submitting your homework, please do the following:

- Make sure that your R script runs without any errors before submitting
- Make sure to include your name on line three of the R script
- Use this naming convention for your R script "lastname_firstname_ps7"
- Comment out or delete code that you don't want Patricia to see when grading
  - e.g., you might choose to comment out (but not delete) lines of code that provided helpful insights about the data to avoid having many pages of output.
  - e.g., you might decide to delete (as opposed to comment out) lines of code that were part of your investigations but did not provide big insights about the data
  - These decisions are subjective; don't spend much time worrying about which lines to delete or comment out.

Codebooks

ACS codebook
Common Core of Data codebook