Final Activity: Stanford Open Policing Project

SOC Methods Camp

September 6th, 2019

Introduction

We've covered a lot of math topics this week that will be helpful to you as you go into your first year stats sequence. A lot of the foundational R skills we covered will also be useful to you throughout your career. The goal of this activity is to give you a chance to creatively apply what you've learned to actual data.

We'll be working with the Stanford Open Policing Project's data. The Project collects data on vehicle and pedestrian stops throughout the United States, tidies it up, and makes it readily available for public download. You can find the data at https://openpolicing.stanford.edu/data/.

Your task

Your job is to clean, analyze, and make a small PDF report on the data. The possibilities are endless, so get creative! You could decide to look at racial differences in stops over time, seasonal changes in stops and stop outcomes, whether or not cops stop more people near donut shops, or anything else you're curious about. We just have a few requirements:

1. **Merge at least two data files together.** There are lots of data files available. Use at *least* two in your work (though you can always use more). Be thoughtful and intentional about what you're choosing to use. For example, you might merge all the California files because you're interested in that state, or combine NYC and LA because you want to look at big cities.

2. **Use a for loop and if else statements for data cleaning.** We've seen this week that dplyr makes it really easy to transform variables without having to rely on complicated for loops and if else statements. Still, we're going to make you practice these skills at least one more time. At some point in your data exploration, use for loops and if else statements to transform at least one variable. For example, you might transform the Stop Time variable from clock time to time of day (morning, afternoon, night, etc.).

3. **Use dplyr for data manipulation and investigation.** Dplyr is a very powerful tool. We just want to see that you're using it throughout your exploration of the data. Some place it might be helpful include making new data variables with mutate, using summarise to spit out nice summary statistics, etc.

4. **Write at least one function.** As part of your data cleaning and analysis, you're probably going to need to do some kind of math at some point. Maybe you want to find a stop's mean distance from the city center, or the average duration of a stop. Now, obviously R has a built in mean() command that can be helpful here. But go ahead and write your own function for finding means (or doing something else!) and use it instead. A more advanced function you could write would be one that translates the date of the stop into a day of the week, to look at when in the week cops are most likely to stop people.

5. **Produce at least three difference ggplot figures.** Data visualization is an essential skill. Come up with at least three different figures that show something interesting about your data. Now, a lot of questions you might be exploring might seem like they only lend themselves to one graph at first. For example, if you're looking at how the number of daily stops changes throughout the year you might default to using a line graph showing average number of stops each day of the year in a given area. But you could also dig deeper into the data by making a second graph with separate lines to show what that average looks like for white people versus people of color.

6. **Write a nice report.** In the end, we'd like to see a *short* PDF printout of your work. You should include your code, and comment it all clearly so we can easily reproduce your work. In the PDF, make sure to write about what you explored and why. comment on your findings and figures, and offer some sort of conclusion.
7. **Talk about probability.** Somewhere in your report, tell us a little about probabilities. For example, tell us what the probability is of an arrest being made given that the person stopped was driving (as opposed to walking).

One possible idea...

Sometimes creativity is hard. Feel free to use this suggestion to get you going if you're not sure what you want to explore in the data.

The *Los Angeles Times* reported in 2019 that the LAPD's elite Metropolitan Division, which was expanded in 2015, stops African American drivers "at a rate more than five times their share of the city's population." [1] Unfortunately, Stanford doesn't provide data for LA. However, we can look at whether or not there's similar inequality in other major cities in California.

Some ways we could do this:

-Pull and merge data for San Francisco and San Diego (make sure to make a variable indicating which city a data point came from!)

-Create a new data frame with one observation per city, per year (or month) showing what proportion of drivers stopped were white, black, Hispanic, etc. (Can you use a function to generate the proportions data? Can you use a for loop?)

-Graph the new data to look at stops over time.

-Explore your data more deeply! What happens when we break it down by sex and race? What happens if we look at race, sex, and age? What if we look at race and reason for stop?

[1] " 'Stop-and-frisk in a car:' Elite LAPD unit disproportionately stopped black drivers, data show", Chang and Poston, Jan. 24th 2019.