Homework 1

Due by 9PM, September 11th

This homework is designed to get you a little more familiar with the functions we went over in class and to introduce you to the functions you read about for next class. Please submit R code, explanations, and / or plots for any section marked with the **[To submit]** heading by emailing jakeporway+itp@gmail.com.

Reading

- "What is Data Science?", Mike Loukides http://radar.oreilly.com/2010/06/what-is-data-science.html
- "The Scars of Stop and Frisk", New York Times http://www.nytimes.com/2012/06/12/opinion/the-scars-of-stop-and-frisk.html
- pp. 1-11 and pp 14-15 in http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf , excluding the sections titled "R Basics: Graphical Data Entry Interface" and "Example: Working with mathematics".
- Go through the code and the rest of the class notes at http://www.jakeporway.com/teaching/R/lecture1.R

Part One – Probing Stop and Frisk

Download the data at http://www.jakeporway.com/teaching/data/snf_11_2011_1.csv and load into R as we did in the notes.

[To submit] We saw in class and in the readings that we can interrogate vectors of data, making plots of stops by race and gender. Let's use the skills we picked up to answer the following questions (please submit the code you wrote to answer these questions):

- How many women were stopped? What percentage of the stops is this?
- How many different kinds of suspected crimes are there? What do you think about that? Is that what you expected?
- Which precinct had the most stops? How many were there? Which precinct had the least stops?
- How many people between 18 and 30 were stopped?

Notice that there are columns for "frisked", "searched", and "arrested". Every row represents a stop, but not everyone was frisked, searched, or arrested. Use your burgeoning data skills to:

- Find the number of people who were given the full treatment: frisked, searched, and then arrested.
- Make a histogram of their ages.

Part Two – Teach Me Something

This part of the assignment is fairly simple and open-ended. Your first task is to get yourself a data set that you like and teach me something about it. Anything. It doesn't have to be profound, it doesn't have to be earth changing, it should just use your skills from this lesson. Some thoughts on choosing your dataset:

- I'm assuming many of you have datasets that you're already working with for other projects (web traffic, Kinect output, Twitter feeds, biofeedback data, etc.), so feel free to use one of those.
- Don't have data already? No worries. The easiest place to get tabular data is from Google Fusion Tables http://www.google.com/fusiontables/search?source=ftlp&hl=en&q=. Search for data that's meaningful to you (try "poverty" or "environment" or "bieber") and then choose File > Download to save the file as a CSV.
- Uninspired? Go with the Stop and Frisk data we used in class. Just find something that interests you to talk about.
- Whichever route you go, think simple. My inclination when starting out was always to get a "rich" dataset, like genetics info or all of Wikipedia, a dataset with hundreds of columns and a trillion rows. Not only is that actually going to make your job much harder (for now, we might get into big datasets later), it's just overkill. Find something with 3-5 columns and you should be good to go.
- Not everything is a CSV (the only type of data we've loaded in yet), but if you can find tabular data, that's going to fit well with the course. E-mail me or talk to me if you have ideas for other formats that you really want to use (we'll be getting into JSON shortly if you can wait).

[To submit] Once you've got your dataset, your job is to do the following:

1. Write a couple of sentences about what your dataset contains (column names, types) and why you chose the dataset.
2. Teach me one thing about your dataset. This can (and should be) extremely basic. You don't have to find some amazing correlation in your data, just tell me one true thing. Or make one plot. You've learned how to look at maxes and mins, you can subset your data, you know how to plot it, so you should easily be able to find something to say about your data.