

HW02: Ingesting Baltimore City Data

STAT 370

Assigned: February 17, 2017; Due: February 24, 2017

1. Open your class repository project in R Studio. (On a lab computer you will create a New Project and download from GitHub; on your own computer you will Open Project—if the project does already exist, and isn't already open).
2. Unless you have just downloaded it, pull the latest changes from GitHub to get the latest version from the cloud. You should always do this first to avoid conflicts, which can arise if you edit your project on two computers. For the same reason, when finished, and before you log out of a lab computer, and before you edit on another computer, **remember to COMMIT, PUSH, and CHECK GITHUB to verify that your changes have arrived in the cloud.** Conflicts can be resolved, but they are a pain.
3. Find a data set you are interested in exploring on the open Baltimore data website: <http://data.baltimorecity.gov/>
4. Figure out how to use R to download the data set you want. Discover the right command with Google and talk to me if you are having trouble. Use the browser to discover the URL for the file on the baltimorecity.gov website. CSV will be easiest (that stands for Comma Separated Values), although CSV may not be available for all data sets (talk to me if that is the case, or find another). The appropriate R command will download the file from the internet and put it on your local file system in the location you specify. Later, you will want to use this command within a dynamic document to automatically download the latest version of the data, if it changes, or is updated, regularly. You will download the file automatically with R, instead of manually using a browser.
5. Load your data set from where you put it in your file system into R. Google how to load a CSV file into R (or if not CSV, Google whatever you downloaded).
6. Verify that the data loaded correctly.
7. Figure out a way to display some of the data. If you have loaded in a data frame named “df”, you could type the command “head(df)” to display the first 6 rows. In addition to data frames, that command should work with many other types of data, as well.
8. Alternatively, if you have time, you could try to plot the data. I recommend that you download the book *The Hitchhiker's Guide to GGPlot2 in R* by Jodie Burchell and Mauricio Vargas from the publisher **Leanpub** (Google “Leanpub”). The publisher asks for a donation, but if you select “\$0.00” for your donation amount, the publisher gives you the book for free. Give generously, if you are able—but don't feel bad if you can't—I must confess that got the book for free, too. We will go over this material in class, but get a head start if you have time. (By the way: Leanpub has many other relevant selections available for the same arrangement.)
9. In a future class, we will wrap all of this material in a dynamic document, which when **Knit** is pressed, will download the latest data from the website and graph it. See if you can find a data set here which changes, regularly, to use with the second part of this assignment! If you can't find regularly changing data on this site, look elsewhere; weather data?
10. **Remember to COMMIT, PUSH, and CHECK GITHUB to verify that your changes have arrived in the cloud.**