Data Wrangling and Visualization 101 - Overview

Florian Jaeger

10/25/2020

Our goals for the next two weeks

This git repository contains a number of R and Matlab tutorials originally developed for an undergraduate research course in the brain and cognitive sciences at the University of Rochester (BCS 206/207). Taken together the lectures cover about six meetings worth of materials. Lecture files are numbered in the recommended order of going through them. Topics covered include the following:

- Preliminaries: Version control, code style guides, and code documentation.
- Data wrangling: Turning the data into the form you need (e.g., *dplyr* in R)
- Data visualization:
  - General principles
  - How to plot in R (*ggplot*, *plotly*) / Matlab
- Documenting your code:
  - R Markdown / MatLab notebook
- Introduction to confidence intervals

The course focuses on learning what tools are available and on *examples* of use (rather than an in-depth tutorial). There are great online tutorials and cheat sheets that contain further information. Specifically:

- **By Friday 10/30:**
  - Send your initial data (e.g., excel, matlab, or csv file) to both Sabya and Florian (you can use Slack). If you won't have your own data by then (which might well happen), that's ok. In that case, please ask your PI whether they have an old data set with data similar to the one that you would be analyzing.
- Monday 11/2:
  - **Prepare before class:** You will receive a data set ahead of class (described below). Load it into Matlab/R and familiarize yourself with its structure. Quick primers are available online for both R (https://rstudio.cloud/learn/primers/1) and Matlab (https://www.mathworks.com/help/matlab/getting-started-with-matlab.html), as are tutorials on how to load data. Go through them *before* class.
  - In class, we will use that data to illustrate how we can visualize our data at various levels of summarization.
- **By Friday 11/6:**
  - Complete all exercises from the Monday class that weren't completed during class and send them to your group's R/Matlab instructor.
- Wednesday 11/4:
  - **Prepare before class:** Load your group's data. For at least one subject in your data, try to repeat the different plots we've introduced on Monday for your own data. You will be asked to present your efforts in class (to go through your script while sharing your screen). It's ok to get stuck, but please use Slack to ask for help prior to class.
  - In class, we will also go through problems/errors you might have encountered while trying to create visualizations of your data.

- Monday 11/9:
  - **Prepare before class:** Watch this great video on confidence intervals by Dr. Nic. Then watch this simulation video by the Khan Academy on the meaning of confidence intervals and what happens to those confidence intervals, if we repeat (replicate) an experiment. Some of you might prefer to instead (or also) have a manuscript that covers this topic in writing. Cumming's (2013) suggestions for a 'new' statistics is a worthwhile read and has a self-contained section on "Replication, p values, and CIs" that I recommend. You can just read those 2 pages.
  - Summarizing variability in your data—and thus the researcher's *uncertainty* about the central tendencies in the data
  - Introduction to standard deviations, standard errors, and confidence intervals
- Wednesday 11/11:
  - **Prepare before class:** If you haven't already, please read Ten simple rules for better figures
  - Preparing your visualizations for presentation (captions, axis titles, legends, and other annotations)
  - Saving your visualizations (format, dimensions)