Practical Data Science and Data Management

ICPSR Summer Program

July 29-August 1, 2019

Instructor: Matt Denny

The most challenging aspect of conducting empirical social science research is often not running the final analysis, but instead collecting, cleaning, managing, and visualizing the data used in the study. This course is intended to arm you with the tools you will need to go from data collection up until the point at which you perform your analysis. Some topics we will cover include web scraping, using the Twitter API, the basics of working with large datasets, generating publication quality plots, and tools for working with text, network and panel data. This workshop will be taught using R, but will assume no previous experience managing data, or programming in R.

You can email me at matthewjdenny@gmail.com with any questions. There are lots of additional materials available on my website at: http://www.mjdenny.com/teaching.html, but you will only need to look at the stuff linked to from this syllabus in order to be successful in this course. To download all of the materials associated with this course, you will want to start by downloading a GUI client for Git.

- For Windows: https://windows.github.com/
- For Mac: https://mac.github.com/
- For Linux, you may have to rely on the command line, although https://git-scm.com/downloads/guis has some options (depending on your distro).

About a week before the workshop starts, you will want to visit the course website: https://github.com/matthewjdenny/ICPSR_PDSaDM_2019. You may then want to `clone` the git repo for this course onto your computer by clicking the "Clone in Desktop" button on the right hand side of the page. If you want to directly edit the files posted on the course repo and track your changes, you can copy individual files into another directory and create your own Git repo with the files in it. If you are not sure what any of the above meant, don't worry! We will go over using Github at the beginning of the first workshop, so there is no need to spend too much time trying to figure Github out. If you are excited to get started, I suggest you check out this **Github pictorial**. Welcome to the workshop!

Course Overview

- **Overview:** In this four day short course, participants will be introduced to core data management concepts and techniques using the R programming language. We will focus on reading data into R (and writing data from R) from a wide variety of sources, and manipulating and cleaning it for use in statistical and descriptive analyses. We will place an emphasis on providing participants with the skills to deal with poorly formatted source data, and multiple datasets at once. We will then apply these basic skills to automated data collection (both from websites and the Twitter API), and will go over some of the basics of manipulating and working with text data in R. The workshop will also include a mini-unit on social network data management, as well as a mini-unit on producing publication quality plots and graphics.
- **Goals:** By the end of the workshop, participants should possess the basic skills necessary to perform most data collection and management tasks they are likely to encounter over the course of a research project intended for publication as a scholarly journal article.
- **Prerequisites:** No previous experience with R is required, and a number of tutorial resources will be provided before the beginning of the course. However, participants are expected to have some very basic experience working with data using software such as R, Stata, SAS, SPSS or Excel.
- **Exercises:** There will be nightly exercises (with solutions provided) for those who want to apply what they learn in the course on their own, or for those who want some example code to work off of after the workshop. These exercises will be completely optional.

Schedule

This is a draft outline of the workshop schedule, it will likely change over the course of the workshop depending on how fast we end up going.

Before the workshop

Please download R and RStudio before the workshop. I provide directions below, as well as a screen-cast tutorial. If you have never used R before, you may want to look through some of the introductory examples listed below:

- If you are using a Windows computer, start by downloading and installing RTools. It is very important that you do this first. If you already have R installed on your computer, download RTools and then reinstall R. You can download RTools here: https://cran.r-project.org/bin/windows/Rtools/index.html You will want to get "RTools35.exe" (or the newest version). You will also want to check the box to edit your system PATH variable while you are installing RTools. You can see an example of how this might look (although you should expect the screen to have a different path variable on your computer) by looking at the top picture in this post: https://github.com/stan-dev/rstan/wiki/Install-Rtools-for-Windows. Again, you will need to do this before installing R on a windows computer.
- Download R: https://cran.r-project.org/. Right at the top of the page, you will see links to download the latest version of R (currently 3.5.0, but there may be a newer version by the time you download it).
- Download RStudio: https://www.rstudio.com/products/rstudio/download/. Not that as a researcher, you can select the free license.
- This section of Quick R provides a basic overview of the R interface. You can navigate between pages by clicking on the links on the top left – http://www.statmethods.net/interface/index.html. I still end up finding useful examples here, but do not find the website to be in the most conversational form. This is a useful starting point for basic example code, particularly for plotting and statistical tests.
- A nice place to start learning R interactively is Swirl. Note that this is probably the best "teach yourself" option for just messing around with code, but you will want to actually get R installed on your computer to do serious work.

To make things easier, I have created a video tutorial that will walk you through installing R and RStudio on your computer. You can check it out by clicking on the following video link: https://www.youtube.com/watch?v=0FWXWnPuxrs

Monday 7/29/19

On the first day of the workshop, we will go over installing and using R and RStudio, and on the basics of R programming and data I/O. This will serve as a common foundation for the rest of the course and will be targeted at participants with little to no experience with R or programming.

Tuesday 7/30/19

On the second day of the workshop, we will shift to developing the programming techniques to manipulate larger and more complex datasets, to work with multiple datasets at once, and automate tasks and reuse code. Tim permitting, we will begin to dive into basic text processing in R.

Wednesday 7/31/19

On the third day of the workshop, we will pick up with more advanced text processing. We will then discuss the ethics and legal issues around web scraping before diving into some web scraping examples. Time permitting, we will also begin to cover using the Twitter API.

Thursday 8/1/19

On the final day of the workshop, we will wrap up our coverage of the twitter API. We will then move on to creating publication quality plots in R using ggplot and base R graphics. Finally, time permitting, we will cover the basics of social network data management.

Resources

- A nice place to start learning R interactively is Swirl.
- Quick-R has a bunch of easy to read tutorials for doing all sorts of basic things – http://www.statmethods.net/.
- Hadley Wickham wrote a book that covers a bunch of advanced functionality in R, titled **Advanced R** – which is available online for free here – http://adv-r.had.co.nz/.
- Hadley Wickham has an R package `rvest` for web scraping that is detailed in this blog post.
- A blog post by Charles Dimaggio that I have referred to in the past: blog post.
- Another blog post by Zev Ross that I have referred to in the past: blog post.