PS 200X: Data Analysis Workshop

Fall 2023

Instructor: Doeun Kim (doeun2+TA@ucla.edu)
Office Hours: 10 - 11 a.m. on Fridays or by appointment
Sign up for office hours: https://doeunkim.org/officehour.html

Learning objectives

General:

- Learning R coding used in social science statistics coursework
- Gaining comfort with using R for basic data analysis and visualization
- Exposure to data and code used in published social science papers
- Learning how to get help

Specific:

- Getting a computer set up for scientific computation
- Importing data saved in a variety of formats and with various problems
- Exploring a new dataset in tabular format, including describing the format of variables, dimensions of the dataset, and what rows and columns represent
- Visualizing data, including histograms as well as line, point, and jittered-point plots; mapping elements of the data to plot features such as colors and symbols; annotating the plot to explain its axes, scale, important data points, etc.; and faceting
- Transforming data to prepare for analysis or visualization, including changing variable types, rescaling, combining variables, and tidying data
- Joining datasets by common variables, "fuzzy" comparisons, and spatial relations
- Finding and downloading replication data from social science articles
- Reproducing tables/figures from published studies by running their replication code
- Reporting results from analyses in code and in Quarto (quarto.org)

How we will meet these goals

- Reading textbook chapter and following along in code
- Chapter exercises, submitted after class
- Offline quiz on the board to start each class
- Student-led live coding to review
- In-class exercise in pairs that rotate three times in the quarter (switch in Week 4 and 7)
- Self evaluations and pair evaluations at end of Week 3 and 6

(Self-)evaluation[1]

There is little high-quality evidence that instructor-graded exams help students learn, and some evidence grading is harmful. Instead, you will *evaluate yourself* at several points during the quarter, in terms of your effort and your learning. In addition, you will *evaluate each other* within your pairs.

Your own evaluation, and the evaluation of you by your peers, will form the basis of your final grade. You will be provided with your peer evaluations at each point, and you will decide how to incorporate those into your evaluation. I reserve the right to change your grade, up or down, at the end of the quarter if I do not agree with your self-assessment (I will rely on the approximate breakdowns listed in the assignment section in making my own assessment). If I do decide to make a change, I will meet with you to talk about your performance before making a decision.

Auditing: auditing a class like this without completing the assignments will not be productive for you, so auditors will not be permitted. We encourage you to take the course for credit!

Getting help

This course is a lot of work! The activities are motivated by the idea that the most effective way to learn this material is to do it yourself. This means if you get behind, it will be hard to catch up. We don't want this to happen!

Prerequisites

This course is designed to enable pure beginners to succeed, that is, learners who have not ever tried coding and who do not have a lot of computing experience. For those who have used R but not extensively, have used R but not with the tidyverse, or have used Stata or another software tool before, this is a comprehensive introduction to the "tidy" way of using R (using dplyr, ggplot2, etc.).

Computation

You must have access to a laptop for each class session. Reach out to the instructor if this poses a problem.

Professional ethics

You are subject in this class to UCLA's academic honesty policies. You should not pass off others' work, words, or code as your own (you can avoid this by liberally citing and when relevant

[1] I draw on Jessica Calarco and Jesse Stommel's ideas on "ungrading."

including quotation marks or notes indicating what is directly taken from others; our greatest virtue is building off the past work of others).

Reading and resources

Main text: Grolemund, Garrett and Hadley Wickham. *R 4 Data Science*. (Free online, 2nd edition.)

Turn in all assignments in Quarto

Schedule

R and RStudio installation session (during math camp)

0. **Getting started I: Installation, visualization, and error messages**
   Ch2-3, Understanding error messages
1. **Getting started II: Transforming variables and tidying data**
   Ch4-7
2. **Getting started III: Importing and addressing common file type issues**
   Ch8-9, 21, add haven Stata and issues with ASCII vs UTF-8
   rio, readxl/writexl, haven, readr
3. **Visualizing data and the grammar of graphics**
   Ch10-12
4. **Transforming variables I: logicals, numbers, strings, factors, dates, and times**
   Ch13-15, 17-18
5. **Transforming variables II: regular expressions and missing values**
   Ch16, 19, add using ChatGPT for regular expressions
6. **Joining multiple data sources and brief intro to spatial data**
   Ch20, add st_join and fuzzy_join
7. **Fitting statistical models and reporting on results**
   TBD
8. **Functions and programming**
   Ch26
9. **Iteration (running an operation multiple times)**
   Ch27
10. **Navigating base R** 🤡
    Ch28 plus other resources