.

SOCIOL 690S: Data Wrangling and Data Visualization with R

Kieran Healy

Duke University

kieran.healy@duke.edu

- Instructor: Kieran Healy, Sociology Department, 268 Reuben-Cooke.
- Time and Place: Wed at 1:25pm-3:55pm in Reuben-Cooke 329 from January 10th–April 17th, 2024.

About this course

This course will teach you the elements of data wrangling and data visualization, mostly in R.

For the data wrangling side, we will not not focus on particular statistical methods or modeling techniques. Rather, we will learn how to accomplish everyday tasks that have to happen before you get to that part. These include topics such as getting your own data into R, rearranging and recoding it, exploring its structure, munging and reshaping tables, and presenting summary tabulations and graphs of this work. We will also examine some more advanced versions of these topics such as managing large datasets, parallelizing tasks, and some of the rudiments of writing functions and maintaining code that any social scientist working with quantitative data should know a bit about.

For the visualization side we we will emphasize the importance of being able to look at and learn from your data yourself and also the best way to present it visually to others. Throughout the course we will emphasize how R and the tidyverse "thinks". Every dataset is different, especially at stage where it still needs further cleaning or arranging before it can be easily analyzed or effectively presented. This course will teach you the logic and implicit "flow of action" behind the tidyverse's tools, giving you the ability to apply and extend this way of thinking when working with your own data and its particular challenges.

Should I take this course?

You should take this course if you are interested in answering questions like these:

- How can I properly get my data into R?
- How should I deal with different types of data?
- How can I explore the structure of my data?
- How can I manipulate, summarize, and tabulate my data?
- How can I efficiently clean my data?
- How can I reshape or reconfigure my data?
- How can I graph or report on my data?

The course does not presume any prior experience with R. However, if you are an R user and have been annoyed with questions like these:

- How can I get these 50 CSV files into R?
- Why can't I get the right answer when summarizing this grouped data?
- How can I tell R that my categorical measure is ordered?
- How can I clean up this textual data?
- How can I neatly calculate summary statistics for all the measures in my data?
- How can I arrange this table to print in a nice way?
- How can I get my Stata data files into R?
- What do I do if I have more data than RAM?
- How can I make this code run faster?
- How can I write my own functions?
- Why doesn't the answer I found on Stack Overflow work properly?
- Why *does* the answer I found on Stack Overflow work properly?
- Why does R keep telling me "Object of type 'closure' is not subsettable"?

… then this course will be worthwhile for you, too.

Core texts

I recommend (but do not require you buy) the following books. Draft or full versions of all of them are available for free online.

- Hadley Wickham, Garrett Grolemund, and Mine Çetinkaya-Rundel, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, Second. (Sebastopol, CA: O'Reilly Media, 2023), https://r4ds.hadley.nz.
- Kieran Healy, *Data Visualization: A Practical Introduction* (Princeton: Princeton University Press, 2019), http://socviz.co/.
- Claus E. Wilke, *Fundamentals of Data Visualization* (Sebastopol, California: O'Reilly Media, 2019), https://serialmentor.com/dataviz/.
- Hadley Wickham and Jennifer Bryan, *R Packages* (Sebastopol, CA: O'Reilly, 2023), https://r-pkgs.org.

Software

We will do all of our visualization work in this class using R and use RStudio to manage our code and projects. R is a freely-available programming language that is designed for statistical computing and widely used across the natural and social sciences, as well as in the rapidly-growing world of "data science" generally. RStudio is an integrated development environment, or IDE, for R, a kind of control center from which you can manage the engine-room of R itself. It is also freely available. If you haven't used these tools before, don't worry. The course does not presuppose any familiarity with them. We will get up and running with them during the first week.

Schedule

The weekly schedule can be viewed on its own page, which has more details on readings, examples, and problem sets.

Week	Date	Topic
Week 1	Jan 10 / -	No class (Weds schedule in effect)
Week 1	Jan 17 / -	How R thinks
Week 2	Jan 24 / -	Manipulating tables with dplyr
Week 3	Jan 31 / -	Tidying data
Week 4	Feb 7 / -	Import, clean, and recode
Week 5	Feb 14 / -	Iterating with functionals
Week 6	Feb 21 / -	(Class cancellation)
Week 7	Feb 28 / -	APIs and scraping

Week	Date	Topic
Week 8	Mar 6 / -	Databases
Week 9	Mar 13 / -	Spring Break
Week 10	Mar 20 / -	Parallel processing and furrr
Week 11	Mar 27 / -	ggplot Geoms geoms geoms
Week 12	Apr 3 / -	Scales, guides, themes
Week 13	Apr 10 / -	Maps and spatial data
Week 14	Apr 17 / -	Polishing and presenting

Course policies

- Attendance is required, and important. I am a reasonable person; if you need to be absent please *let me know in advance* insofar as that is possible.
- Do the assigned readings in advance of class.
- Submit problem sets, or other assignments, on time.

Required work and grading

Weekly **Class Participation** and **Problem Sets** will let you reflect on the reading and practice your coding and visualization skills. Problem sets are due by end of day the *Monday* after they are assigned.

How you should approach this course

The material covered in the course has a lot of *continuity* and it is *cumulative*. You will be learning a set of practical skills. This means that techniques we learn early on will be necessary for understanding things that come later. It also means that regular practice will help you a lot. So, this is not a "Topic of the week" course where you can tune out for a few weeks while expecting to be able to easily drop back in later. The material we cover each week will not be overwhelming. If you participate during class and keep up with the weekly assignments you'll be in a very strong position to do well in the class. If you don't, it'll be harder than you expected.

Duke community standard

Like all classes at the university, this course is conducted under the Duke Community Standard. Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and nonacademic endeavors, and to protect and promote a culture of integrity. To

uphold the Duke Community Standard you will not lie, cheat, or steal in academic endeavors; you will conduct yourself honorably in all your endeavors; and you will act if the Standard is compromised.