THIS IS THE FALL 2023 SYLLABUS. WE WILL ADJUST SOME ASSIGNMENTS AND ASSIGNMENT WEIGHTINGS, BUT THE BASIC STRUCTURE WILL REMAIN EXTREMELY SIMILAR.

PSCI 1800: Introduction to Data Science (for the Social Sciences)
Fall 2023 Term

Lecture Time/Location:
MW 10:15 – 11:14 AM
Leidy Labs 109
To find your recitation section time/location, please use Courses@Penn.

Instructor:
Professor Matt Levendusky
mleven@sas.upenn.edu
Office Hours: Mondays & Wednesdays, 11:15 – Noon (right after class). I'm also available almost every day on campus. Send me an email if you want to meet.
Office Location: 402 PCPSE (133 S. 36th St)

Course TA:
Nick Pangakis
njpang@sas.upenn.edu

Please contact the TA directly for their office hour times and locations.

Background on the Course:

We live in a data-driven world. We are bombarded with stories every day about how data is yielding new information about our lives. For example, a recent study of racial profiling by the police analyzed more than 20 million traffic stops from the state of North Carolina.[1] Another analysis looked at four million such stops to examine whether there are differences between how often male and female police officers search suspects when they stop them (and how often each finds contraband when they do so).[2] Less than a generation ago, it would have been impossible to analyze that much data without a supercomputer and highly specialized programming knowledge. Today, you can easily analyze that sort of data on a standard desktop computer with some relatively basic skills.

But in a world of "big data," there is more need than ever to know how to analyze it. All of the data in the world is useless if you don't know to use it to draw meaningful conclusions. This is the Political Science course to help you begin your journey to doing just that.

[1] Frank Baumgartner, Derek Epp, and Kelsey Shoub. 2018. *Suspect Citizens: What 20 Million Traffic Stops Tell Us about Policing and Race*. New York: Cambridge University Press.
[2] Kelsey Shoub, Katelyn Stauffer, and Miyeon Song. 2021. "Do Female Officers Police Differently? Evidence from Traffic Stops." *American Journal of Political Science* 65 (3): 755-69.

Data science is just a buzzword that means data analysis, so these courses will teach you how to analyze and interpret the types of data that speak to social and political questions. For example, in this course, we'll use datasets that allow us to analyze many questions, including: which counties moved from Trump to Biden, and what does that tell us about U.S. electoral politics? How has the cost of college varied over time across different states? How does your college major influence your future earnings? And what do all those traffics stops tell us about race and policing?

In this first course, we will introduce the basic principles of data analysis using R. This class will prepare you to do both future coursework, as well as introduce you to the general process of analyzing data. Many of you may not go on to become data analysts, but in an increasingly data-driven world, you will encounter it as a consumer and as a citizen. Understanding what's happening "under the hood" will make you a better, more informed consumer of that data. Indeed, even with generative AI (or maybe especially with these tools), knowing a bit about what happens in R makes you a much more informed consumer.

We have one over-arching learning goal for this class:

To learn how to analyze a substantive problem of interest to you using data.

To support that meta-goal, we have four sub-goals:

1. Learn how to import, manipulate, and clean data so that it can be analyzed using the R statistical computing language
2. Learn the basics of descriptive analysis, and how you can use the data to explore important real-world patterns
3. Learn core data visualization skills, so that you can easily present key results to non-technical audiences
4. Learn to think systematically, so that you can understand how to tackle a problem step-by-step. This is crucial for working R, but also for life.

Course Prerequisites

This course is a first introduction to data analysis. We don't assume any statistical or technical background, but if you are familiar with how a computer works, the concepts will come more easily.

But that said, you will work hard in this class. Doing data analysis requires "learning by doing," so this is not something where you can sit back and learn passively. You'll need to bring your laptop to class, ready to engage with the exercises we'll do in class. You'll need to work hard on the homework assignments, and be prepared to Google things that you don't understand (like the error messages from R).

Also, it is imperative that you do not fall behind. The class is cumulative, and not understanding something today will prohibit you from learning the material in the weeks to come. If you're struggling, it is crucial that you reach out to the instructor and the TA. We want to help you, but you need to let us know when you need more help.

I often tell students that learning data science is no different from learning a foreign language. R, the data analysis program we'll use, is a language for communicating with your computer no different than Spanish or Italian. At first, it all seems totally overwhelming: what is the pipe, and why do I need to run library(tidyverse) every time I start a session? But eventually, it becomes more natural.

Computing

Data analysis requires knowing a computer programming language. In this course sequence, we'll use R Statistical Computing Language (https://www.r-project.org). R is one of, if not the, most popular tools for doing data analysis. Indeed, in a recent survey of data science jobs ads, more than 50% mentioned R skills![3]

R is a free open-source program, and we'll interact with it via the RStudio IDE (Integrated Development Environment), which you can download at https://www.rstudio.com. An IDE is just a fancy term for a package that makes it easier to interact with a computer program. So RStudio just makes it easier to work with R.

We'll also show you how to set up cloud storage on Dropbox. Dropbox is the leading cloud storage platform, its basically a way to share files across multiple machines. You can get a free account at https://www.dropbox.com/. Happily, all three of these tools—R, RStudio, and Dropbox—are free (at least for what we need). If you don't like Dropbox, that's fine. You can use Box (to which Penn subscribes), or just store the files locally on your computer. If you do this, though, please do be sure to back your work up regularly.

You should have these installed on a laptop so that you can work on some short assignments that we will do together in lecture and recitation. If you don't have a laptop, you can still take the class, but it will be more challenging and require a larger time commitment on your part. Please note that the College Houses allow students to rent a laptop for up to 5 days (https://www.collegehouses.upenn.edu/equipment), and the library also allows for brief rentals as well (https://guides.library.upenn.edu/computingindex/computing/public/laptops). Support to cover course costs (such as a laptop) is also available via Student Financial Services.

Notetaking

[3] Jeff Hale, "The Most In-Demand Skills for Data Scientists," *Toward Data Science* [Blog], Available online : https://towardsdatascience.com/the-most-in-demand-skills-for-data-scientists-in-2021-4b2a808f4005. R ranked as the third most common skill behind Python and SQL. Python is a more general programming language (while you can use it for statistical analyses, it is much clunkier than R for doing so). SQL is a database querying language. If you are interested in them, there courses in both at Penn.

Because we do so much work with code, we will provide you with access to an RMarkdown file with the code from class via Canvas, as well as the course slides (that will have the code chunks we run embedded in them, along with the output). We recommend using these Markdown files to take notes in class.

Communication

We'll communicate with you in several different ways. We'll try for duplication between methods, but we will undoubtedly fail in that goal sometimes, so please pay attention to all of these channels:

1. **Email**: we use email a lot, so please check your account regularly. It's old fashioned, but someday, your boss will be a middle-aged person like Professor Levendusky. Middle-aged people like email, so learn to use it now.
2. **Canvas**: We'll send notices (mostly emails, and also sometimes announcements) via Canvas.
3. **In-Class**: We'll also try to announce things in class and recitation as well.

Textbook

Really, this class is much more about what we do in class than the textbook. But a good guide to what we'll do is this one:

Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data,* 2nd edition. Available online at: https://r4ds.hadley.nz

I abbreviate this book as RDS below. It is available for free online (if you prefer a physical version, you can order one on Amazon, but that's completely unnecessary). View the textbook as another resource to help supplement and deepen what we do in class. It covers (mostly) the same content, but in a slightly different format, with different emphases, etc. This is the second edition of the book, but honestly, it's not much different from the first in what we'll discuss. You can use the old first edition link as well, just note that they re-numbered the chapters somewhat.

Assessment

We'll assess you by four means in this class:

Homework Assignments (40%): There will be 6 homework assignments, but we will only count your best 5 scores toward your final grade (i.e., we will drop your lowest score).

Please note that you must submit every homework assignment. Because the material is cumulative, you need to submit all of the assignments to master the material. Not submitting one homework will harm your ability to learn future material.

Note that part of your homework score on each assignment is determined by solving the problem, and part is determined by the quality of the writeup and explanation. Earning a high score on the homework requires not just figuring out the code, but *clearly communicating what the code is doing*. We want to help you learn not just to code but to communicate that to a non-technical audience. That latter skill is even more valuable than knowing how to code, especially in a post chat-GPT world.

Midterm Exam (20%): In late October, we'll have a (take-home) midterm exam, covering the first part of the class. Please note that the midterm exam will require you to work on your own, so make sure you're doing your own work on the homework assignments.

Final Project (35%): At the end of the term, in lieu of a final exam, you'll have a final project. In this project, you'll create an report that takes a dataset of your choice and uses the skills you learn in this class to tell us something about it. We'll explain this more as we go through the term.

Note that there are various "check-in" points built into the homework assignments to help students stay on track and not leave this until the last moment.

Section Participation and Engagement (5%): Attendance in lecture is not required, but attendance in recitation section is. You should attend your weekly recitation section and come prepared to engage actively in it. Recitation section will be devoted to covering the code we discuss in more detail.

Nota Bene #1: We encourage you to work together in groups to tackle the homework assignments. But note that the homework only has value if you understand it! A long-term study found that people who just copy homework from others unsurprisingly did worse in the class (and then obviously worse in other classes).[4] Also, the midterm exam will be on your own, so you'll need to solve the homework assignments to do well on it!

Nota Bene #2: You must turn in all assignments for this course—including all homework assignments—or you risk a failing grade in the class. Completing all course assignments is a requirement for the class.

Code of Academic Integrity

All students at Penn are required to uphold the university's Code of Academic Integrity, which you can find online at https://catalog.upenn.edu/pennbook/code-of-academic-integrity/. Please read and familiarize yourself with the code.

You are not permitted to use any generative AI technology, including tools such as Chat-GPT, for your work in this class. Using such tools will be considered a violation of

[4] Arnold Glass and Mengxue Kang, "Fewer Students Are Benefitting from Doing Their Homework: An Eleven-Year Study," *Educational Psychology* 42(2): 185-99.

Penn's Code of Academic Integrity and suspected used will be reported to the Center for Community Standards and Accountability. Please contact me if you have questions about this policy.

Course Schedule

This is a rough timeline of the topics we'll cover in class. It is, of course, subject to change as we go through the semester.

Topic 0: Introduction to the Course

August 30th: What's the value of data science for the social sciences? And why would you want to take this class?

September 4th: No Class, Enjoy the Labor Day Holiday

September 6th: Beginning to Use R & An Introduction to RMarkdown
Reference: RDS, Chapters 1, 3, and 7.1 (don't read about projects)

Recitation section begin the week of September 4th

By September 11th, all students should have successfully installed R, RStudio, and Dropbox (or some other storage method) on their computer to continue in the class. If you're struggling with this, please visit the R tutoring hours or office hours.

Topic 1: Basics of Data Visualization and Cleaning

September 11th and 13th: Visualizing Data: An Introduction to ggplot
Reference: RDS, Chapter 2.
For a "real-world" example of these skills, see: https://medium.com/bbc-visual-and-data-journalism/how-the-bbc-visual-and-data-journalism-team-works-with-graphics-in-r-ed0b35693535

September 18th and 20st: Wrangling and Tidying Data, Part 1
Reference: RDS, Chapter 4

Homework #1 due 9/20

September 25th and 29th: Wrangling and Tidying Data, Part 2

October 2nd and 4th: Exploratory Data Analysis & Tables
Reference: RDS, Chapters 10 and 11

Homework #2 due 10/4

Topic 2: Reading Your Own Data into R

October 9th and 11th: Combining Data
Reference: RDS, Chapter 20

October 11th and 16th: Importing and Pivoting Data
Reference: RDS, Chapters 6, 8, and 21
Note: If you're unfamiliar with file structures, the following video is a good introduction to them: https://www.youtube.com/watch?v=NG7Y0kkGR8g

Homework #3 due 10/18

October 18th: Final Project Discussion and "Catch Up"
We'll first discuss the final project. In any remaining time, we'll catch up with any remaining material.

The midterm exam goes out 10/20, is due back 10/25 via Canvas

Topic 3: Some Useful Data Skills

October 23rd and 25th: Mapmaking in R, using the maps() and sf() packages
Reference: Kieran Healy, *Data Visualization: A Practical Introduction* (Princeton, NJ: Princeton University Press, 2019). Chapter 7: Drawing Maps.

October 30th: Short final project blurb due on Canvas

October 30st and November 1st: A Very Brief Introduction to Text-Based Data (aka Strings)
Reference: RDS, Chapter 15

Homework #4 is due 11/8

November 6th and 8th: Functions and Loops
Reference: RDS, Chapters 26 and 27

November 13th: Catch-Up Day
This is our second "catch-up" day.

Topic 4: An Incredibly Brief Introduction to Regression

November 15th: Conditioning and the Idea of Inference: Univariate and Bivariate Statistics
Reference: Tom Holbrook, *An Introduction to Political and Social Data Analysis Using R*, Chapter 8: Sampling and Inference [https://bookdown.org/tomholbrook12/bookdown-demo/]

November 20th and 27th: The Idea of Regression

Reference: Elena Llaudet and Kosuke Imai, *Data Analysis for Social Science: A Friendly and Practical Introduction* (Princeton, NJ: Princeton University Press, 2022). Chapter 4: Predicting Outcomes Using Linear Regression.

Homework #5 is due 11/20.

Wednesday, November 22rd: No Class, Enjoy Thanksgiving. Recitation sections will not meet that week.

November 29th and December 4th: Can Correlation Ever Be Causality?
Reference: Elena Llaudet and Kosuke Imai, *Data Analysis for Social Science: A Friendly and Practical Introduction* (Princeton, NJ: Princeton University Press, 2022). Chapter 2: Estimating Causal Effects with Randomized Experiments.

December 6th: The Ethics of Working with Data
References:
Part 1: Data Creation
+ Natasha Singer, "LinkedIn Ran Social Experiments on 20 Million Users Over Five Years." New York Times, 24 September 2022.
Part 2: Data Privacy
+ V. Joseph Holtz, Christopher Bollinger, Tatiana Komarova, and Bruce Spencer. 2022. "Balancing Data Privacy and Usability in the Federal Statistical System." *Proceedings of the National Academy of Sciences* 119(31): e2104906119.
Part 3: Algorithmic Bias
+ Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 266(6464): 447-53.

Homework #6 is due December 6th

In your final recitation section this week, you are required to present one visualization from your final project and give a 3-minute presentation about it.

December 11th: Wrap Up & Some Concluding Thoughts

The final project is due via Canvas on December 18th at 10:15 AM