PSCI 1800: Introduction to Data Science (for the Social Sciences)
Fall 2022Term

Lecture Time/Location:
MW 10:15 – 11:14 AM
LRSM Auditorium
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 TAs:
Jon Griffiths
jongrif@sas.upenn.edu

Tyler Jenkins-Wong
tjwong@sas.upenn.edu

Kira Wang
kirawang@sas.upenn.edu

Please contact the TAs 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.

---

[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.

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 first course to help you do just that.

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.

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 TAs.

I often tell students that learning data science is no different from learning a 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.

You'll also need a Dropbox account. 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).

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

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

---

[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.

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:

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. **Slack**: We have a class slack channel (PSCI 1800 Fall 2022). The link to join our Slack channel is: https://join.slack.com/t/sas-amd4396/shared_invite/zt-1dibtaxk2-57ynIkXt2o6Bg3zqfjzC1g. This is an informal way of talking to the professor and TAs, as well as your fellow students. You can ask questions in the Slack channel, and you can also answer/respond to them as well.
3. **Canvas**: We'll send notices (mostly emails, and also sometimes announcements) via Canvas.
4. **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 and Garrett Grolemund. *R for Data Science: Visualize, Model, Transform, Tidy and Import Data.* Available online at: http://r4ds.had.co.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 the most of the same content, but in a slightly different format, with different emphases, etc.

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.

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.

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: 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!

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.

Masking

Because we will be spending 1 hour together in a tightly spaced classroom twice per week, and because said classroom has not great ventilation, we recommend that you wear a mask to class. The evidence suggests that we're all much safer if everyone wears a mask.

It is inevitable that someone will get COVID-19 during the term. If you have COVID, or suspect you might have COVID, please do not come to class (instead go take a test and isolate yourself). We will record the class lectures (via a screencast) and make them available to students who are ill.

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

---

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

August 31st: What's the value of data science for the social sciences? And why would you want to take this class?

September 5th: No Class, Enjoy the Labor Day Holiday

September 7th: Beginning to Use R & An Introduction to RMarkdown
Reference: RDS, Chapters 1, 4, and 6

Recitation section begin the week of September 7th

By September 9th, all students should have successfully installed R, RStudio, and Dropbox on their computer to continue in the class.

Topic 1: Basics of Data Visualization and Cleaning

September 12th and 14th: Visualizing Data: An Introduction to ggplot
Reference: RDS, Chapter 3.
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 19th and 21st: Wrangling and Tidying Data, Part 1
Reference: RDS, Chapter 5

Homework #1 due 9/21

September 26th and 28th: Wrangling and Tidying Data, Part 2

October 3rd and 5th: Exploratory Data Analysis & Tables
Reference: RDS, Chapter 7

Homework #2 due 10/3

Topic 2: Reading Your Own Data into R

October 10th: Combining Data
Reference: RDS, Chapter 13

October 12th and 17th: Importing and Pivoting Data
Reference: RDS, Chapter 11 and 12
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 10/17

October 19th: Catch-Up Day
I've built in two "catch up" days into the syllabus. This gives us a space if we're running behind schedule, want extra practice (or to do a joint coding exercise in class), or something else.

The midterm exam goes out 10/21, is due back 10/26 via Canvas

Topic 3: Some Useful Data Skills

October 24th and 26th: 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 31st and November 2nd: A Very Brief Introduction to Text-Based Data (aka Strings)
Reference: RDS, Chapter 14

Homework #4 is due 11/9

November 7th and 9th: Functions and Loops
Reference: RDS, Chapters 19 and 21

November 14th: Catch-Up Day
This is our second "catch-up" day.

Homework #5 [Wordle] is due 11/21

Topic 4: An Incredibly Brief Introduction to Regression

November 16th and 21st: Conditioning and the Idea of Inference: Univariate and Bivariate
Reference: Tom Holbrook, *An Introduction to Political and Social Data Analysis Using R*, Chapter 8: Sampling and Inference [https://bookdown.org/tomholbrook12/bookdown-demo/]

Wednesday, November 23rd: No Class, Enjoy Thanksgiving. Recitation sections will not meet the week of November 21st.

November 28th and 30th: The Idea of Regression
Reference: Kosuke Imai, *Quantitative Social Science: An Introduction* (Princeton, NJ: Princeton University Press, 2017). Chapter 4: Prediction

December 5th: Can Correlation Ever Be Causality?
Reference: Kosuke Imai, *Quantitative Social Science: An Introduction* (Princeton, NJ: Princeton University Press, 2017). Chapter 2: Causality

December 7th: The Ethics of Working with Data

References:

Part 1: Data Creation

+ Derek Willis, “Professors’ Research Project Stirs Political Outrage in Montana,” *New York Times*, 28 October 2014.

+ Thomas Leeper, In Defense of the Montana Experiment

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 7th

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 12th: Wrap Up & Some Concluding Thoughts

The final project is due via Canvas at the date and time assigned by the registrar.