PSCI 2300

Quantitative Political Science I: Computing

Vanderbilt University — Professor Brenton Kenkel

Fall 2024

How can we use polls to predict election outcomes? How can we detect evidence of discrimination? How can we predict the outbreak of political violence? PSCI 2300 will teach you how to address these and other social science questions through data analysis. The course will introduce the basic principles of statistical analysis and the baseline programming skills you need to analyze data. The goal is to give you the foundational tools you need to analyze data in your own research, as well as to be an informed and critical consumer of statistical claims made in the news media, policy reports, and academic research.

Here is a checklist to consider when deciding whether to take this class:

- □ I am a political science major or someone interested in quantitative social science.
- □ I want to learn about statistical concepts, as well as the programming skills needed to manipulate and analyze data in R.
- □ I am willing to spend time outside of class each week to keep up with the material.
- □ I would like to use data analysis in a paper (e.g., a political science honors thesis) and/or job in the future.
- □ I understand that this is an introductory course, so we will focus mainly on developing a conceptual understanding. We will not dive too deep into mathematical foundations or advanced computational methods.

General Information

Place and time. PSCI 2300 meets on Mondays and Wednesdays from 9:05–10:20am in Commons Center 335.

Email. You can email me at brenton.kenkel@gmail.com. I try to respond to all student questions within one business day, though be aware that I don't usually reply to emails at night or on weekends. And if you have a question about course material where the answer would be useful to others, please post it to the Brightspace discussion forum instead of emailing me!

Office hours and meetings. My office hours are Tuesdays from 2:00–4:00pm in Commons Center 326. You don't need to make an appointment—just show up during my office hours. If you have a question but can't meet during my office hours, email me first. From there we can decide if we can work out the issue over email or if we need to meet.

TA. The TA for PSCI 2300 is Nguyen Ha. You can email her at nguyen.t.ha@vanderbilt.edu. Her office hours are Wednesdays from 2:00–4:00pm in Commons Center 304K (within the LAPOP offices).

Logistics

Lectures. Lectures will be interactive and vary over the course of the semester. The typical class will consist of me introducing concepts and code motivated by real-world questions, then writing code and running the analysis in real time while you follow along on your own computer. This means **you should bring your laptop every day** (a real laptop, not an iPad or other tablet that likely cannot run RStudio).

All lecture notes and code, along with any data we will be using, will be posted on Brightspace before each class. Make sure to download all this before you come to class. I'll try to have everything up by the Friday before each week of class, and I'll email you when the material is ready.

Readings. There's no required textbook. Your main source will be the notes I distribute each week. Depending on the week's content, I may also ask you to read some academic articles or news sources too. Any such material will be posted on Brightspace.

Software. All the software we use is free and open source. We will focus on the programming language and statistical environment R (https://www.r-project.org/).[1] To make our lives easier working with R, we will run it inside the RStudio development environment (https://posit.co/download/rstudio-desktop/). I write the course notes—and you will complete your assignments—using the R Markdown language (https://rmarkdown.rstudio.com/).

Assessment

Your grade will be based on three components:

- *Problem sets (50%).* There will be six problem sets distributed throughout the semester. You will have at least a week to complete each one. These will be based on applying

[1] If you are thinking about a career in data science, I strongly encourage you to learn the programming language Python. Many of the programming skills and techniques you learn about R in this course will carry over to working with Python. We focus on R in this course because it is designed for statistics (whereas Python is a general purpose language), and R is the most commonly used statistical environment in political science.

concepts and code from the lectures to a new question or problem using different real-world data. Your lowest score among the six problem sets will be dropped, and then each of the five remaining will comprise 10% of your grade.

- *Exams (50%).* There will be midterms (each 15% of your grade) and a final (20%). These will be in-class, closed-book exams that test your understanding of data analysis concepts and principles, as well as your ability to interpret and debug R code snippets.
- *Participation (extra credit, up to 3%).* Programming is hard. There will be times that you get stuck, and you can't find (or comprehend) the answers you find on Google or ChatGPT. When you reach this point, post your question to the Brightspace discussion board for the course. And if someone else has posted a question that you can help with, write a response! You can earn up to 3% extra credit in the course for being active on the course discussion boards.

A general note on expectations. As one of the top-ranked universities in the country, Vanderbilt's mission is to pursue excellence in education, scholarship, and research. To advance that pursuit, this course will maintain high academic standards. I expect students to attend class, to engage thoughtfully with the material, and to do the hard intellectual work necessary to succeed on assignments and exams. Deadlines will be enforced, and only the best work will receive high grades. I will only grant exceptions to course policies for documented medical or emergency situations.

Submitting problem sets. You'll submit problem sets on Brightspace. For each problem set, submit both your R Markdown source file and the compiled PDF output. If you collaborate with peers on the problem set, include a brief note listing who you worked with. If you use ChatGPT or another AI chatbot (once permitted), include the relevant conversation as a PDF with your submission.

Late assignments. For problem sets, there is a one-time-use, no-questions-asked, 72-hour extension policy. The first time you miss a deadline on one of these assignments, you can turn in the assignment anytime within 72 hours for no penalty. You don't have to ask or inform me that you are taking the extension option; it will be applied automatically. After the first 72 hours, or if you have used the extension on a previous assignment, each day late is a 5 percentage point grade reduction.

Generative AI. In the first couple weeks of the course, while you are still learning the basics of programming in R, I ask that you completely refrain from using generative AI tools like ChatGPT. This is for your own benefit. A solid understanding of R will allow you to ask ChatGPT more useful, directed questions—and to better spot when it gives you an answer that won't actually work.

After we cover "Prompting ChatGPT" in class, you are free to use generative AI tools on your programming assignments.[2] I only ask that you use generative AI as a *complement* to learning R on your own, not as a *substitute* for your personal learning. Keep in mind that there will be in-class examinations where you don't have access to ChatGPT.

Academic integrity. As in all classes at Vanderbilt, your work in PSCI 2300 is governed by the Honor Code. I encourage you to discuss course material and assignments with your peers, but the work you turn in must be solely your own.

I have no tolerance for plagiarism. If you turn in plagiarized work, you will receive a failing grade for the course and be reported to the Honor Council. Ignorance of what constitutes plagiarism is not an excuse or defense.

Accommodations

I want to provide an effective learning environment for students of all backgrounds, identities, and abilities. If there are circumstances that make our learning environment and activities difficult, if you have medical information that you need to share with me, or if you need specific arrangements in case the building needs to be evacuated, please let me know.

I will do whatever it takes to make sure this class is a place where you can learn and thrive, but I can only do so if you discuss your needs with me as early as possible. I promise to maintain the confidentiality of these discussions. If appropriate, you should also contact Student Access Services to get more information about specific accommodations.

Schedule

Outline of topics

The exact number of lectures on each topic is subject to change, as I may slow things down if there's something important the class is stuck on.

1. **20th century computing.** Basics of R, RStudio, and R Markdown. Navigating directories and loading data from the R console. (1 lecture.)
2. **Data wrangling.** `select |> filter |> mutate |> group_by |> summarize |> pivot`. (2 lectures.)

[2] I'm not going to (and practically speaking, couldn't) forbid you from using generative AI on the writing portions of these assignments, where you interpret and explain your findings. However, I recommend against it—not because of any ethical objection, but because ChatGPT tends to write bullshit. I mean this in the technical sense of the term "bullshit": see the recent academic paper "ChatGPT Is Bullshit" (https://doi.org/10.1007/s10676-024-09775-5), as well as the late philosopher Harry Frankfurt's excellent short book *On Bullshit* (Princeton University Press).

3. **Univariate description and prediction.** Continuous versus categorical variables. Averages, medians, and tabulations. Resampling to estimate prediction error. (3 lectures.)

4. **Prompting ChatGPT.** "As a large language model, I cannot write the PSCI 2300 syllabus for you." (1 lecture.)

5. **Data visualization.** ggplot2. Bar charts, histograms, and density plots for univariate visualization. Box plots, violin plots, and scatterplots for conditional relationships. Making maps. (4 lectures.)

6. **Correlation and regression.** Correlation versus causation. The statistical idea of "control". Bivariate and multivariate regression. Prediction and resampling from regression models. Feature selection with large numbers of variables. (5 lectures.)

7. **Classification.** Loss functions for classification. Bivariate and multivariate methods. Prediction and cross-validation. (2 lectures.)

8. **Clustering.** Supervised versus unsupervised analysis. Pre-processing data. The k-means algorithm. (2 lectures.)

9. **Text analysis.** Processing text for programmatic analysis. Sentiment analysis. Term frequency–inverse document frequency. Clustering text. (4 lectures.)

Dates to remember

- **Friday, September 6.** Problem Set 1 due.
- **Friday, September 20.** Problem Set 2 due.
- **Monday, September 30.** Midterm 1 in class.
- **Wednesday, October 9.** Problem Set 3 due.
- **Friday, October 25.** Problem Set 4 due.
- **Monday, November 4.** Midterm 2 in class.
- **Friday, November 15.** Problem Set 5 due.
- **Wednesday, December 4.** Problem Set 6 due.
- **Saturday, December 7.** Final exam in person from 3:00–5:00pm.