Public Syllabus

Note: This is a partial syllabus designed to be public-facing. Carleton students should see the version on Moodle for all course details.

Course Description

Stat 220 will cover the computational side of statistics that is not typically taught in an intro or methodology focused course like regression modeling. Most of the data you encountered in your first (or second, or third, ..) stats course were contained in small, tidy .csv files with rows denoting your cases and columns containing your variables. Most of the messiness to these data may have been some missing values (NAs). In this course, we'll learn how to extract information from data in its "natural" state, which is often unstructured, messy and complex. To do this, we will learn methods for manipulating and merging data in standard and non-standard formats, data with date, time, or geolocation variables, text processing and regular expressions, and scraping the web for data. To effectively communicate the information contained in these data, we will cover advanced data visualization methods, including methods for creating interactive graphics. We will primarily use the statistical software R in this course, and cover best practices for reproducible analyses and sharing code.

Course Objectives

After completing this course, you should be able to demonstrate your competency in each of the following areas:

- **Develop** research questions that can be answered by data
- **Acquire** data by importing different file types into R, accessing data through API's, and scraping data from the web
- **Wrangle** common types of data into the form that is needed for analysis
- **Visualize** data to provide insight and uncover relationships and patterns
- **Communicate** your findings to stakeholders in written or oral format
- **Document** your code and **collaborate** across coding projects.

Course Components

Meetings

There will be three course meetings per week (Mondays, Wednesdays, and Fridays). Daily attendance and active participation is expected. Course meetings will combine demonstrations/lecture and in-class group exercises. On most days, I'll ask you to complete a reading or watch a short video before class.

Assignments

Homework will be assigned once-ish per week, distributed via GitHub. You will submit homework assignments via gradescope. You will use `quarto` for all assignments and submit all necessary work for each assignment on GitHub.

Portfolio Projects

Portfolio project require you to integrate several smaller computational tasks and require clear communication of the proposed solution or findings to a broader audience. You will typically work in pairs or triples.

Lab Quizzes

Part of being proficient in data science is being able to do basic data analysis "on the fly", without access to class resources. There will be 3 short (~30 minute) in-class lab quizzes to assess your ability to do basic tasks in R. I recognize that "in the real world", you will almost always have access to your resources, so you will also have 48 hours to re-submit.

Final Project

The final project is a capstone experience synthesizing everything you've learned over the course of the term. This is an opportunity for you to exercise your creativity and create something meaningful. The final project is wildly open-ended and more details will follow.

Communication

Assignments and slides will be shared publicly on our course website. Grades will be posted on Moodle. Please use our github discussion page for any homework or course content questions; email me privately with any personal matters (grade discussions, illness, emergency, etc.). Any time-sensitive announcements will be sent via email. It is your responsibility to make sure that your notification settings allow time-sensitive announcements to reach you.

Materials

Textbook

There is no "perfect" data science textbook. We will use excerpts from the following texts:

- R for Data Science 2e
- Modern Data Science with R 3e
- Fundamentals of Data Visualization

These books are all freely available online. If you prefer a hard copy, they are also available for purchase through the publisher.

Software

The use of the R programming language, with the RStudio interface is an essential component of this course.

Academic Integrity

I encourage you to discuss the homework problems with others and use the resources available to you to try to figure out tough problems. You should code and write up your solutions on your own. Lab quizzes must be done by yourself without communicating with others; all work must be your own. You should collaborate with your teammates on projects, and should use external resources for background research and debugging, but all work should be original. The use of textbook solution manuals (physical or online), course materials from other students, or materials from previous versions of this course are not allowed.

Large-language models (e.g. ChatGPT, Gemini, etc.) should only be used for coding or debugging help after you've attempted to solve the problem on your own, and you should never type homework problems directly into a prompt. Copying, paraphrasing, summarizing, or

submitting work generated by anyone but yourself without proper attribution is considered academic dishonesty (this includes output from LLMs).

I also have a few rules in place to protect my intellectual property. You may not record my lectures using tools such as Otter.ai or upload any video or audio recordings to generate transcripts or study notes. You may not upload my course materials (slides, assignment prompts, note sets, etc.) into AI tools or homework help sites (such as chegg).

"AI" tools are new for all of us and it's OK to have questions about what is and isn't appropriate. Please ask if you are unsure of whether or not your actions are complying with the assignment/quiz/project instructions. Always default to acknowledging any help received. Cases of suspected academic dishonesty are handled by the Provost's Office and I am obligated to report any suspected violations of this policy.