Stat 220: Introduction to Data Science

Spring 2025

Meetings	MWF 2a 9:50-11 MW | 9:40-10:40 F	CMC 102
Professor	Amanda Luby aluby@carleton.edu	CMC 223
Office Hours	M 4:15-5:15, T 10:30-11:30, W 2:15-3:15, F 11-12 By appt through my Google Calendar	CMC 307
Organization	github.com/stat220-s25	
Website	stat220-s25.github.io	
Texts	*R for Data Science* (2nd Ed) https://r4ds.hadley.nz/ Wickham, Çetinkaya-Rundel, Grolemund *Modern Data Science with R* (3rd Ed) https://mdsr-book.github.io/mdsr3e/ Baumer, Kaplan, Horton *Fundamentals of Data Visualization* https://clauswilke.com/dataviz/ Wilke	
Software	Maize RStudio Server maize.mathcs.carleton.edu R (optional) free from r-project.org RStudio (optional) free from rstudio.com/downloads	

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 rmarkdown or 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. These are either partner or individual projects.

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.

Grading Policies

Grades are an imperfect measure, especially when we are all coming to class with different backgrounds and experience levels. This course is designed to reward you for consistently participating, staying on top of the course material, and trying your hardest. The grading scheme in this class is designed to reflect how we learn "in the real world": by making mistakes, receiving feedback, and making changes.

How each assignment is graded

Each assignment will include a short rubric with the specifications of how I will evaluate your work. In general, these are the "marks" that you can receive on each course component:

	How it's evaluated	How it's recorded
Homework Problems	Completeness, Correctness, and Effort	Successful or Not Successful
In-class activities	Attendance, effort, and completeness	Successful or Not Successful
Lab Quiz Problems	Completeness and Correctness	Successful or Not Successful
Portfolio Projects	Completeness, effort, correctness, and communication quality	Exceptional, Successful, or Not Successful
Final Project	Completeness, effort, correctness, and communication quality	Exceptional, Succesful, or Not Successful

Earning a course grade

Your course grade is assigned using the table below. Each row indicates the minimum percentage of "Successful" results needed to satisfy the requirement of that grade. To earn a grade, complete *all* requirements listed in the row for that grade.

	Homework Problems	In-class activties	Lab Quiz Problems	Portfolio Projects (4 total)	Final Project
A	85%	90%	90%	2 Exceptional, 2 Successful	Exceptional
B	75%	80%	80%	4 Successful	Successful
C	65%	70%	70%	3 Successful	Successful
D	55%	50%	50%	2 Successful	Successful

Example: Ben finishes the course with 82% of homework problems successfully completed, 90% of in-class activities completed, 85% of lab quiz problems successfully completed, 1/4 portfolio projects marked "excellent" and 3/4 marked "successful", and a "Successful" final project. Ben satisfies everything in the "B" row and earns a "B" in the course.

Plus/minus grades

"Plus" and "minus" grades will be given if you complete all the requirements for a base letter grade **and** make sufficient progress toward the next grade. Below is an overview:

- If "B" base grade:
  - *and* A in three bins: "B+"
  - *and* A in four bins: "A-"
- If "C" base grade:
  - *and* at least B in three bins: "C+"
  - *and* at least B in four bins: "B-"
- If "D" base grade:
  - *and* at least C in three bins: "D+"
  - *and* at least C in four bins: "C-"
- If "F" base grade:
  - *and* at least D in three bins: "D-"

Example: Mira finishes the course with 88% of homework problems successfully completed, 92% of in-class activities successfully completed, 85% of lab quiz problems successfully completed, 3/4 portfolio projects marked "excellent", and a "Successful" final project. Mira satisfies everything in the "B" row, and meets the "A" threshold for three bins (homework, in-class activities, and portfolio projects). Mira earns a B+. If Mira instead receives "excellent" marks on the final project, Mira earns an A-.

How lab quiz resubmissions work

When you resubmit a lab quiz, you change the *denominator* for your quiz bin. Let's say you earn a successful mark on 7/10 problems on the in-class version and 9/10 on the resubmission. Your ultimate score is (7+9)/(10+10). You are not required to do resubmissions.

What if I miss an in-class lab quiz?

If you need to miss a lab quiz, please make arrangements with me at least two weeks in advance. If you must miss a lab quiz due to an illness or last-minute emergency, you must notify me in advance of the quiz to arrange an alternative time. If you do not notify me in advance, you will earn 0/10 points on the in-class portion, but can still submit the revision as usual.

Important points about this grading system

- Different categories of coursework do not "average together": you can't make up for less-than-great work on portfolio projects by doing very well on quizzes, for instance. Each course grade requires *consistent quality and effort across all bins* to earn the grade
- You do not have to do everything. If you want an "A" in the class, for example, you don't have to complete *every* quiz question correctly, only 90% of them.

Tokens

- Turning in a token provides either (a) a 72-hour extension or (b) a revision on a portfolio project
- Tokens may be used for *extensions* on the final project milestone check-ins, but not for the final due-date. Milestone check-ins cannot be revised
- You may use a token for an *extension* on the lab quiz resubmission, but not on the in-class portion. You do not need to use a token to resubmit each lab quiz.
- You can revise the same portfolio project multiple times, if needed, but you must spend a token each time.
- A portfolio project must be completed with a good-faith effort to be eligible for revision. If I deem a submission to be "not assessable" due to a lack of effort, then it cannot be revised.
- All revisions must be submitted by 11:59pm on the last day of class.

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. You have two options for using RStudio:

1. The server version of RStudio on the web at https://maize.mathcs.carleton.edu. The advantage of using the server version is that all of your work will be stored in the cloud, where it is automatically saved and backed up. This means that you can access your work from any computer on campus using a web browser. The downside is that you have to share limited computational resources with each other!
2. A local version of RStudio installed on your machine.The downside to this approach is that your work is only stored locally, but I get around this problem by keeping all of my work on GitHub. You will learn how to use GitHub throughout the course.

Note that you do not have to choose one or the other, you may use both. However, it is important that you understand the distinction so that you can keep track of your work. Both R and RStudio are free and open-source.

Academic Integrity

You are expected to follow Carleton's policies regarding 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 help with R error messages, 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. You should never copy and paste any course materials *into* a large-language model, and you should never copy and paste anything *out* of a large-language model into your course materials. 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.

	Collaboration Allowed
Homework Problems	You are allowed and encouraged to collaborate on homework. You may also use outside resources, but your submitted work must be your own and reflect your own understanding (i.e. it can't be copied). All resources used must be cited. If you cannot explain each line of code to me with just your brain, you should not submit it as your own work.
Lab Quiz Problems	**No collaboration is allowed at all** (including online forums like StackExchange or Reddit). You may use your own notes for resubmissions, but should not use outside resources.
Portfolio Projects	You are expected to collaborate with your group, but cannot rely on external sources other than to help motivate the questions or provide other background information (including online forums like StackExchange or Reddit). You may use any resources from class and package documentation, but getting answers on significant parts of solutions from outside resources is not allowed.
Final Project	You are expected to collaborate with your group, but cannot rely on external sources other than to help motivate the questions or provide other background information. Any outside resources should be properly cited.

Commitment to an Inclusive and Collaborative Atmosphere

We all come to class with different backgrounds and experiences, and this diversity makes our class environment richer. I value diversity and inclusion, and am committed to a climate of mutual respect and full participation in and out of the classroom. This class strives to be a learning environment that is usable, equitable, inclusive and welcoming, regardless of race, ethnicity, religion, gender and gender identities, sexual orientation, ability, socioeconomic background, and nationality. If you anticipate or experience any barriers to learning, please discuss your concerns with me.

Resources

ACCOMMODATIONS: Carleton College is committed to providing equitable access to learning opportunities for all students. The Office of Accessibility Resources (Henry House, 107

Union Street) is the campus office that collaborates with students who have disabilities to provide and/or arrange reasonable accommodations. If you have, or think you may have, a disability, please contact OAR@carleton.edu to arrange a confidential discussion regarding equitable access and reasonable accommodations. You are also welcome to contact me privately to discuss your academic needs. However, all disability-related accommodations must be arranged, in advance, through OAR.

STATS LAB: The Stats Lab (CMC 304) offers drop-in help R/RStudio help sessions run by friendly and knowledgeable lab assistants on most weekday evenings and some weekend times. Stat Lab is primarily intended for Stat120, but most Lab Assistants can also help with Stat220.

TUTORS: If you find you need more support than office hours and the stats lab can provide, the Academic Support Center offers peer tutoring on the basis of referrals, requests, and availability of tutors. You can request tutoring through a form on their website, or discuss your needs with me and I can submit a referral.

TITLE IX: Please be aware that all faculty are "responsible employees", which means that if you tell me about a situation involving sexual harassment, sexual assault, dating violence, domestic violence, or stalking, I must share that information with the Title IX Coordinator. Although I have to make this notification, you will control how your case will be handled, including whether or not you wish to meet with the Title IX coordinator or pursue a formal complaint.

And finally....

Take care of yourself. Do your best to maintain a healthy lifestyle this term by wearing a mask when you're sick, eating a vegetable every now and then, exercising, avoiding excessive drug and alcohol use, getting enough sleep, and taking some time to relax. Your physical and mental health is more important than your grade in this course. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. For more information, see Student Health and Counseling (SHAC), the Office of Health Promotion, or the Office of the Chaplain. If you are experiencing physical or mental health symptoms as a result of coursework, please speak with me so we can address the problem together.