Stat 220: Introduction to Data Science

Adam Loy

Fall 2021

E-mail: aloy@carleton.edu
Office: CMC 218
Student Hours: M 2:30-3:30 pm, T 12:30-1:30 pm, W 1-2 pm, T 12:30-1:30 pm & by appointment
Class Room: CMC 102
Class Hours: 1a (8:30-9:40 MW & 8:30-9:40 F)
Web: stat220-fa21.github.io

What is this course about?

This class provides a practical foundation for you to compute with data in order to gain useful insights. Topics focus on elementary programming concepts in the R language and the necessary tools to handle, analyze, and interpret data. Specifically, we will grapple with the key aspects of a data-scientific investigation:

- Forming a data-driven question
- Acquiring and importing data
- Tidying/wrangling data
- Exploring data
- Modeling data
- Communicating results

What will you learn?

After successfully completing this course, you will be able to

1. Develop research questions that can be answered by data.
2. Import/scrape data into R and reshape it to the form necessary for analysis.
3. Manipulate common types of data, including numeric, categorical (factors), text, and date-times in order to provide insight into your data and facilitate analysis.
4. Explore data using both graphical and numeric methods to provide insight and uncover relationships/patterns.
5. Utilize fundamental programming concepts such as iteration, conditional execution, and functions to streamline your code.
6. Build, tune, use, and evaluate basic statistical learning models to uncover clusters and classify observations.
7. Collaborate on a data science project.
8. Draw informed conclusions from your data and communicate your findings to stake holders in written or oral format.
9. Consider the ethical implications of data science products.

Are you prepared?

If you took Stat 120, 230, or 250 at Carleton, then you are in good shape. I expect you all to have some basic experience in R. Specifically, I expect that everyone can load a data set into R, calculate basic summary statistics, and create basic exploratory graphics.

Classroom culture

All people in this class deserve to feel safe, respected, and valued. That means that all members of our classroom community are responsible to each other to make sure that all voices get heard, all comments are considered respectfully, and everyone has a chance at success. Determination, cooperation, and hard work are highly valued in this class; helping your neighbor understand the material is more important than trying to be the first to answer. We flourish as a community when every individual participates and learns.

What do I expect of you?

I expect you to

- prepare for class through review and reading/viewing;
- attend class meetings and participate actively in class activities;
- be proactive in completing course work and avoid procrastination;
- maintain awareness of course announcements and calendar events at all times, by checking email, Slack, and the course webpage on a several-times-a-day basis;
- take initiative to seek out help when you are stuck or have a question using office visits, Slack posts, study groups, and whatever else works for you;
- interact collegially and respectfully with other members of the course;
- complete assignments on time and in accordance with the college academic integrity policy.

What can you expect of me?

I want you to be successful in this course. I will do my utmost to help you do this, by creating and maintaining a learning environment based on challenge and support and giving my highest professional commitment to your success and well-being. I will

- encourage and support your efforts to learn more about R and the fundamental ideas about data-scientific computing;
- organize class materials and activities to promote the learning goals for the course;
- be available to help you when difficulties arise;
- provide prompt feedback on assignments, and update you on your overall progress;
- keep information up-to-date on the course website/GitHub repository;
- check e-mail and Slack frequently and provide timely responses. (See "How can you contact me?" for what this means.)

How can you contact me?

- **Come to class:** Our class meets in person during 1a in CMC 201. You can always talk to be briefly before class.
- **Student hours:** This term I'll hold drop-in student hours in CMC 307 at the following times:

Day	Time
Monday	2:30-3:30 pm
Tuesday	12:30-1:30 pm
Wednesday	1-2 pm
Thursday	12:30-1:30 pm

  I'll also hold two evening student hours on Slack each week. These will be on 8-9 pm on Mondays and Thursdays.
- **Appointments:** You don't need an appointment for student hours, just drop in. If you need a face-to-face meeting outside of student hours, there will be special times set up for appointments during the week. You can schedule a meeting via Calendly: https://calendly.com/aloy-meetings/15min.
- There are three main **ways to reach me outside of student hours or appointments**: through email, through voice, and through the messaging app known as Slack that we will use in the course. **Slack is preferred**. (Use a direct message if you wish the message to be private.)
- **A note on my availability: I do not typically check email or Slack messages during the hours of 9pm and 7am on weekdays, and I check them infrequently over the weekend** in order to devote time to family, rest, and finding balance. Messages received during these times will receive attention once I am back online. Otherwise you can expect to receive a response to your message within 6 hours, often much sooner. If you post questions to Slack, you are likely to receive responses faster.

Materials

Textbook

There is no "perfect" data science textbook, so we will use excerpts from the following e-books:

- R for Data Science - https://r4ds.had.co.nz/
- Introduction to Data Science - https://ubc-dsci.github.io/introduction-to-datascience/
- Introduction to Data Science - https://rafalab.github.io/dsbook/
- Fundamentals of Data Visualization - https://serialmentor.com/dataviz/
- An Introduction to Statistical Learning - http://faculty.marshall.usc.edu/gareth-james/ISL/

Required software

- The use of the R programming language with the RStudio interface (downloadable from rstudio.org) 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. This is recommended due to the computational resources this course demands. 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.

What types of assignments will you have?

This course requires four types of graded work.

1. **Homework (25%).** Homework assignments will be assigned regularly from GitHub. You will use R Markdown on all assignments and submit all necessary work (.Rmd and compiled .md files) for each assignment on GitHub. You may work with classmates but you must write up your assignment on your own. Please reference Using Git and GitHub in Stat 220 on the resources course webpage. **Unexcused late work will not be accepted.**

   In addition, each class day we will have two designated note takers that will add notes to a Google Doc. These notes will be shared with the entire class to create a crowd-sourced resource that everyone can benefit from. Your contribution to these notes will count as a homework assignment.
2. **Quizzes (25%).** There will be four quizzes checking your understanding of key concepts and tools throughout the course. These quizzes will be given during class.
3. **Mini projects (25%).** There will be four mini projects assigned throughout the course, roughly every 2 weeks. You will be assessed for your ability to complete the data-scientific task as well as your ability to communicate your results. You will use GitHub to collaborate and submit these assignments and you will work with a new partner on each assignment. **Unexcused late work will not be accepted.**
4. **Final Project (25%).** Much of this course is a build up to the final project, which is a capstone experience synthesizing everything you have learned over the course of the term. This is an opportunity for you to exercise your creativity and create something meaningful. This project will be wildly open-ended, and its evaluation will emphasize originality and ingenuity in addition to sophistication and complexity. More details about the project will follow.

This course also requires some ungraded work.

1. **Class attendance and engagement.** Regular class attendance is the easiest way to learn the course material, practice new skills, ask questions, and get immediate feedback; however, attendance in class is not mandatory. When you attend class, I expect you to be fully engaged.

2. **Preparation.** You are expected to read or watch the assigned material prior to class. There will be some topics that we do not discuss in depth during class, but you will be expected to understand. Come prepared to ask questions and engage actively with each day's activities. In addition, review your class notes after each class, carefully reconstructing for yourself the ideas. Coming to class three times a week without this extra work will not lead to deep learning.

Tokens

I recognize that life happens. We each have our individual situations and priorities, and these may not align with the course schedule and requirements. In those moments, the nature of the course may be overwhelming. The best way to prevent this is to start assignments early in order to provide an opportunity to ask questions when you need help. An alternative is to make use of "tokens."

Each student starts the term with 4 "tokens." Tokens can be used as follows:

- A 48-hour extension on a homework assignment (the request must be submitted before the deadline)
- A 48-hour extension on a mini project (each group member will need to spend a token)
- Revision of a quiz problem or mini project (for mini projects, each group member will need to spend a token)

To spend a token, fill out a Token Spending Form found under the resources tab of the course website. Once you have submitted the form, the item you "purchased" is yours—you don't need to wait for me to give permission or to respond to you. Token totals will be updated weekly in the Moodle gradebook.

There are a few opportunities for you to **earn up to 3 additional tokens**. These include:

- Come to student hours prepared with a significant question more than 24 hours before a homework is due. The token will be given only if you ask for it!
- Turn in a Mistake Report about a significant mistake from written homework or a quiz problem. See the Mistake Report handout for directions.
- After reading the syllabus, drop by student hours and introduce yourself/let me know what you did this summer. Be sure to ask for the token!
- Attend a statistics talk, make sure I see you there, and write a brief reflection (one paragraph) on what you learned. You can submit your reflection on Moodle. I'll be sure to announce the talks in class.

A few limitations on tokens:

To keep the logistics of grading reasonable and to help you get the most out of your revisions, there are some limitations to this revision policy:

- Quiz problem rewrites must be submitted within one week after the quiz is returned to you.
- All rewrites must be submitted by 4:30 p.m. on the last day of classes, unless otherwise noted.

What support is available to help me meet my goals?

There are many resources to help you better understand the course material. These include:

- **My office.** I have scheduled student hours listed at the top of this syllabus, please feel free to attend these even if you don't have a clearly formulated question. I also welcome you to contact me outside of class and student hours. The easiest way to do this is to post a message/question on Slack. You can also DM me on Slack or schedule an appointment on calendly.
- **Your peers.** I strongly encourage you to get to know and to work with your peers in this class. They will be a valuable asset in helping you master this material. Note that working together does not mean copying solutions. All homework problems should be initially attempted by you and write-ups for the problems must be expressed in your own words. I suggest completing your write-up away from others working on the problems. Only after you have thought about the problems yourself is it valid to seek help from your peers, the tutors, or myself. You may not work with your peers on Problems.
- **Dean of students.** If your personal situation, due to COVID-19 illness or other circumstances, begins to impact your ability to engage with the course, please contact the Dean of Students Office and myself.
- **Campus resources on wellbeing:** I urge you to make yourself—your own health and well-being—your priority throughout this fall, academic year, and your career at Carleton. It's important to recognize stressors you may be facing, which can be personal, emotional, physical, financial, mental, or academic. Sleep, exercise, and connecting with others can be strategies to help you flourish at Carleton. If you are having difficulties maintaining your well-being, feel free to contact me and/or pursue other resources, such as Student Health and Counseling or the Office of Health Promotion.

COVID-19 related policies

What are the health and safety protocols?

Carleton's culture of accountability and respect remains essential as we move into the 2021-22 academic year. We have an obligation to protect one another and the members of the Northfield community, and we all must continue to take that responsibility seriously—especially amidst a global pandemic. To ensure the well-being of each other and the broader community, we will

- stay home when sick. (Even if you don't have COVID-19, you should stay home if you aren't feeling well.)
- follow CDC guidance on testing, quarantine, and isolation.
- follow the College mask-wearing policy.
- not eat in class while the college is requiring mask wearing.

What happens if you (the instructor) cannot be in class?

While I am vaccinated, my three young children are not and breakthrough infections can still occur. There is a possibility that I will need to miss class this term either due to COVID symptoms or to care for one of my children. If I cannot be physically present for a class meeting, then you will receive a notification through both Slack and email on or before that day. In most cases, class

will still be held, but in a synchronous online format using Zoom. Please check Slack and email daily.

What happens if I (the student) cannot be in class?

Similarly, it is likely some folks will need to miss class for illness throughout the term. Please reach out to me as soon as possible if you need to miss class because of COVID-19 symptoms, the need to quarantine, or the need to isolate so that we can make arrangements for your continued engagement in the course. While each persons needs may be unique, there are a few guiding principles we'll use to help you remain engaged in the course.

- All materials (slide, handouts, etc.), assignments, and announcements will be posted on the course webpage.
- There are options for virtual student hours (see How can you contact me?). In addition, you can set up an appointment via Calendly.
- Each class day, we will have two designated note takers that will add notes to a Google Doc. These notes will be shared with the entire class so that you can "get the notes from class" without any additional hassle. This has the added benefit of creating a community of learning and caring, and also allows me to check on class understanding.
- I will not record individual lectures or replicate lectures during office hours. I understand that this might help you engage in the course; however, this would be a huge time investment that would detract from the energy I have to meet all of our course goals—I'd be teaching the course twice. The collaborative daily notes provides a rich alternative and allow me to provide the best course experience for the entire group.
- If, after working through the collaborative notes and readings, you feel that you need a "lecture" to truly learn the material, please talk to me. There are alternative video formats and videos from previous courses that I can suggest on a case-by-case basis.

Academic Integrity

You are expected to follow the policies regarding academic integrity established by Carleton College. All work on quizzes must be your own. You may collaborate on homework, but you must submit your own assignment that reflects your own thinking, work, and organization. Any work you submit for a grade should be your own, not a facsimile of a classmate's work or an online solution, which would constitute academic dishonesty. To check if your homework meets this standard, imagine I asked you to explain your reasoning for each problem—you should be able to do so with ease using language and code similar to your submission. Finally, cell phones will be prohibited during quizzes.

Cases of academic dishonesty will be dealt with strictly. Each such case will be referred to the Academic Standing Committee via the Associate Dean of Students or the Associate Dean of the College. A formal finding of responsibility can result in disciplinary sanctions ranging from a censure and a warning to permanent dismissal in the case of repeated and serious offenses.

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 (e.g., mental health, attentional, learning, autism spectrum disorders, chronic health, traumatic brain injury and concussions, vision, hearing, mobility, or speech impairments), please contact OAR@carleton.edu or call Sam Thayer ('10), Director of the Office of Accessibility Resources (x4464), to arrange a confidential discussion regarding equitable access and reasonable accommodations.