Introduction to Data Science

STAT 220

Instructor Info

Deepak Bastola (he/him/his)

https://deepbas.io

dbastola@carleton.edu

Course Info

MW 8:30 - 9:40 AM
F 8:30 - 9:30 AM

CMC 102

https://moodle.carleton.edu/course/view.php?id=43045

Office Hours

M 9:50-10:50AM (CMC 223)
T 1:00-2:00PM (CMC 328)
W 09:50-10:50AM (CMC 223)
Th 1:00 - 3:00 PM (Virtual)
F 1:30 - 2:30 PM (CMC 328)

CMC 223

https://calendly.com/dbastola/15min

Welcome!

Welcome to introduction to data science! This course will cover the computational side of data analysis, including data acquisition, management, and visualization tools. The course introduces principles of data-scientific, reproducible research and dynamic programming using the R/RStudio ecosystem.

Prerequisites

If you took Stat 120, 230, or 250 at Carleton, then you are in good shape. It is essential to recap your basic R and R-markdown skills by the first week of the class. Specifically, I expect that everyone can load a dataset into R, calculate basic summary statistics, and create basic exploratory data analysis. I will expose you to Git and GitHub version control throughout the first few weeks of the class and prior exposure to these is not required.

Learning Objectives

- Develop research questions answerable by data and acquire relevant data in R through importing or scraping.
- Manage, wrangle, and format various data types (numeric, categorical, text, date-times, geo-location) for streamlined analysis and insights.
- Perform exploratory data analysis using graphical and numeric methods to uncover relationships, patterns, and insights.
- Utilize programming concepts (iteration, conditional execution, functions) to optimize code efficiency and perform text mining using regular expressions.
- Implement and evaluate statistical modeling, inference techniques, and unsupervised machine learning algorithms to identify clusters and classify observations.
- Draw informed conclusions from data analyses and effectively communicate findings through written reports and interactive platforms.

Materials

Textbooks

We will use exerpts from the following e-books:

- R for data Science - https://r4ds.had.co.nz/
- Data Science: A First Introduction https://ubc-dsci.github.io/introduction-to-datascience/
- Introduction to Data Science - https://rafalab.github.io/dsbook/
- Fundamentals of Data Visualization - https://clauswilke.com/dataviz/
- An Introduction to Statistical Learning- https://statlearning.com/

Required Softwares

The use of the R programming language with the RStudio interface (downloadable from rstudio) is an essential component of this course. We will primarily be using the server version of RStudio on the web at https://maize2.mathcs.carleton.edu/auth-sign-in. You can access this from any computer on campus using a web browser. All of the computations and storage is done in the cloud and we can easily push our work to a remote Github server. If you are off campus, you will need to use the campus VPN (https://apps.carleton.edu/campus/its/services/accounts/offcampus/).

Classroom Culture

All people in this class deserve to feel safe, respected, and valued. That means that all members of our class 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. Please be prepared to take an active, patient and generous role in your learning and the learning of your classmates.

Course Communication

Course website

All of the essential materials used in this course will be accessible through our class moodle and the course helper website at `https://stat220-winter24.netlify.app/`.

Slack

Slack will be used for student hours, informal and urgent course communication. I can guarantee a reply within 24 hrs. You can join our course workspace here. You can use Slack right from a web browser, or you can download a standalone Slack application to your Mac, Windows, Linux and/or Android/iOS device. You can control whether you receive notifications on new posts by going to Preferences, as well as decide which 'channels' to subscribe to. A 'channel' is a discussion thread, which is used to organize communications into topics.

How can you contact me?

Our class meets in-person during 1a in CMC 102. I am open to chat briefly before class. You are always welcome to come to my office during student hours. You do not need to make an appointment, 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.

You can always use email to let me know about personal issues that arise during the term or specific technology issues that you are having. If you need a faster reply, direct message me on Slack.

Course Flow

The tentative course flow (in no particular order) is:

1. Introduction: Familiarize with RStudio, Markdown for reproducibility, Git and GitHub integration, review R structures and objects, and create basic functions.
2. Basic Data Visualization: Dive into ggplot2, create simple maps and networks.
3. Data Acquisition: Explore various data importing options (including web scraping), and learn to import multiple tables simultaneously.
4. Data Reshaping: Master the use of tidyr package, join data tables, and understand the differences between long and wide data formats.
5. Data Cleaning: Learn to separate columns, manipulate strings, and handle dates and times effectively.
6. Data Transformations and Exploratory Data Analysis (EDA): Utilize dplyr for data wrangling (mutating, summarizing, counting, grouping), implement pipes, and create visualizations with ggplot.
7. Advanced Data Visualization Develop basic interactive graphs using 'shiny' and 'leaflet'.
8. Statistical Learning: Understand the fundamentals of supervised and unsupervised learning techniques, and apply cross-validation for model evaluation.

Grading Scheme

5%	Class participation
15%	Individual assignments
15%	2 Paired Projects (7.5 % each)
45%	2 Midterm Exams, 22.5% each (Tentative: Week 4 and Week 9) • Each midterm grade will be 75% of your initial midterm score and 25% of your score after redos • If no redos then initial score counts as 100%
20%	Final Project

Your final grade will be the weighted average of the above.

Preparation and Participation

Data science is impossible to learn without doing. I expect you to come prepared to fully participate during lectures. You will also be expected to review and read any assigned readings/topics/codes.

In-class activities

There will be tailor-made .Rmd activity files made available to you that we will go over in most classes. These could be thought of as lab activities in tandem with the class flow. These will partially corroborate the lectures and is intended for your practice. Students will be expected to submit the class work to moodle in a .pdf format at the end of class to get participation credits. These activities will be graded merely for completion.

General awareness

It is your responsibility to maintain awareness of course announcements and calendar events at all times, by checking email, Slack, and the course web-page on a several-times-a-day basis. You are expected to be prudent and 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.

Individual Assignments

Homework assignments will be assigned regularly from GitHub. You will use R Markdown on all assignments and submit all necessary work (.Rmd, compiled .md and .pdf files) for each assignment on GitHub. These homework problems that are to be written up by yourself, though you can talk with classmates about these problems. But your R coding and explanations should be your own, and not shared between classmates. Follow the Homework Guidelines when writing up the assignments. You are allowed one “free” late assignment during the term. Unexcused late work will not be accepted.

Statistical Writing

Mini-projects

There will be two small, open-ended data projects assigned during the term that you will work on with an assigned partner. 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.

Final Project

You will work in a team of 2 students on a project of your choosing. The final project will be a culmination of everything you will learn in this course and its evaluation will emphasize originality and ingenuity in addition to sophistication and

complexity. More details about the project will come after midterm break. Your project submissions must be submitted to GitHub by noon Wednesday, March 13.

Technology

Expect to use a laptop or lab computer on most non-exam days. All labs on campus should have R and Rstudio, including the stats lab located in CMC 304.

Stats Lab Hours

Schedule for stats lab tutors can be accessed by clicking here, Stats lab Schedule.

Make-up Policy

Late assignments

I will allow a couples hours of grace for each homework submission. If your homework will be more than 2 hours late, please fill out this Late Assignment Form for any assignment that is more than a couple hours late:

- One “no excuse” late homework will be accepted if it is submitted within two days of the original due date.
- I will allow other exceptions on a case-by-case basis due to illness, family emergency, etc. Please fill out the late assignment form and email or talk to me about your situation.

Late projects

Please do not use the late homework form for late mini projects. As a general rule, late mini projects are not allowed, but I will allow for exceptions on a case-by-case basis. Email me if you need to discuss an extension on a project.

Make-up exams

No make-up exams will be given for any reason unless arrangements have been made with me at least 24 hours in advance of the scheduled release time. Exceptions will be made only for illness or emergencies.

Academic Honesty & Integrity

All assignments and exams must be done on your own. Note that academic dishonesty includes not only cheating, fabrication, and plagiarism, but also includes helping other students commit acts of academic dishonesty by allowing them to obtain copies of your work. You are allowed to discuss homework with classmates but you must write up your answers on your own. It is okay to get coding help from prefects, tutors, classmates, online resources, but extra care should be done to write your own versions of the codes. In short, all submitted work must be your own, in your own words.

ChatGPT Policy

In this data science class, the production of original code and work by students is encouraged, while using any available resources to support their learning for homework and projects as mentioned in the Homework Guidelines. ChatGPT, or some other variants, is acknowledged as a potential resource. However, effectively interpreting its results requires expertise. For this reason, using ChatGPT’s output verbatim for class assignments is strictly prohibited, unless explicitly approved in the assignment prompt. Seeking assistance from experts is recommended. This can be achieved by attending office hours, visiting the statistics lab, or utilizing the math skills center. Such an approach ensures a deeper understanding of the subject matter and upholds the academic integrity of the class.

Diversity and Inclusivity Statement

I strive to create an inclusive and respectful classroom that values diversity. Our individual differences enrich and enhance our understanding of one another and of the world around us. This class welcomes the perspectives of

all ethnicities, genders, religions, ages, sexual orientations, disabilities, socioeconomic backgrounds, regions, and nationalities.

Accommodations for Students with Disabilities

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.

Title IX

Carleton is committed to fostering an environment free of sexual misconduct. Please be aware all Carleton faculty and staff members, with the exception of Chaplains and SHAC staff, are "responsible employees." Responsible employees are required to share any information they have regarding incidents of sexual misconduct with the Title IX Coordinator. Carleton's goal is to ensure campus community members are aware of all the options available and have access to the resources they need. If you have questions, please contact Laura Riehle-Merrill, Carleton's Title IX Coordinator, or visit the Sexual Misconduct Prevention and Response website: `https://www.carleton.edu/sexual-misconduct`.