Introduction to Data Science

STAT 220

Instructor Info

Deepak Bastola (he/him/his)

https://deepbas.io

dbastola@carleton.edu

Course Info

MW 12:30 - 01:40 PM
F 01:10 - 02:10 PM

CMC 102

https://deepbas.io/courses/stat220/

Office Hours

M 02:00-03:00 PM
T 01:30-02:30 PM
W 11:00-12:30 PM
Th 02:30-03:30 PM

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.

Key Topics

- Data acquisition and management
- Data wrangling and formatting
- Exploratory data analysis and data visualization
- Statistical modeling and inference
- Unsupervised machine learning
- Text mining

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 in the first week of the class and prior exposure to these is not required.

Learning Objectives

- Develop research questions that can be answered by data. Import/scrape data into R and reshape it to the form necessary for analysis.
- Manipulate common types of data, including numeric, categorical (factors), text, date-times, geo-location variables in order to provide insight into your data and facilitate analysis.
- Explore data using both graphical and numeric methods to provide insight and uncover relationships/patterns.
- Utilize fundamental programming concepts such as iteration, conditional execution, and functions to streamline your code.
- Build, tune, use, and evaluate basic statistical learning models to uncover clusters and classify observations.
- Draw informed conclusions from your data and communicate your findings using both written 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/
- Mastering Shiny - https://mastering-shiny.org/
- 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://maize.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 communications and materials used in this course will be accessible through the website at `https://deepbas.io/courses/stat220/`.

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 4a 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. I will also set aside 8-9 pm on Sundays and Thursdays for evening student hours on Slack each week.

Course Flow

The tentative course flow is:

1. Intro: Rstudio, Markdown and reproducibility, Git and GitHub, review of R structures and objects, writing simple functions
2. Getting data: importing options (including scraping table data), importing multiple tables
3. Reshaping data: 'tidyr', joining data tables, long vs wide
4. Cleaning data: separating columns, strings, dates and times
5. Data Transformations and EDA: 'dplyr' data wrangling (mutating, summarizing, counting, grouping), pipes and ggplot graphics
6. Visualization: 'ggplot2', simple maps and networks, simple interactive graphs via 'shiny'
7. Statistical Learning: basics of supervised and unsupervised learning, cross-validation

Grading Scheme

5%	Group assignments
20%	Individual assignments
10%	Paired projects
45%	Midterm Exams, 15% each • Each midterm grade will be 75% of your initial midterm score and 25% of your score after redos.
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. 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.

Assignments

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. There will be both weekly group assignments and individual assignments. The group assignments will be due on Thursdays and the individual assignments due on Mondays. Unexcused late work will not be accepted.

Group assignment

You will be assigned to a group this term to work on weekly group assignments. This assignment is graded for completion and effort and is practice for your individual assignment.

- Group recorder: This position will rotate among individuals in the group. Their job is to be the primary person to write-up the group homework assignment and make sure it is submitted on time. All others in the group need to contribute by joining in discussions, finding relevant material in resources or troubleshooting R code.
- Group grader: Shortly after the due date for group homework, I will post its solution. The grader in a group will check their group's answers against this solution key. They will record how well their responses align with the posted solution in a the group homework's Github repo. I won't consider a group assignment submitted until this entry is complete.

Individual assignment

You will also have homework problems that are 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.

Note-taking Activity

In addition, each class day we will have two designated note takers that will add notes to a Google Doc/Folder. 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. You can sign-up for this note taking activity here. Use the same link to upload your notes at your earliest convenience.

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 2-3 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 16.

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 201.

Stats Lab Hours

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

Make-up Policy

Late work

I will allow a couples hours of grace for each homework submission. If you experience significant technological problems that

limit your ability to participate, please contact the ITS Helpdesk at 507-222-5999 or helpdesk@carleton.edu. 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.

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.

Lending Library

If the price of your textbook is an obstacle for you, there are a number of campus resources to support you including a small CSA lending library and the Dean of Students Office. TRIO students can use the TRIO lending library. If you've exhausted these resources, and are not a TRIO student, then you can make use of the Math/Stat department lending library. We have a limited number of textbooks to lend out for Stat 120, and Math 111, 120, 210, 211, 232 and 236. Contact Sue Jandro in the Department of Mathematics and Statistics to reserve a book for the term.

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.

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`.