DATA 11800 Introduction to Data Science

The University of Chicago, Winter 2023

Mon, Wed, Fri 1:30 – 2:20 pm

Ryerson Laboratory, Room 251

Professor Evelyn Campbell, Ph.D.
campbelle@uchicago.edu

Office Hours: Mondays 11:30 – 12:30 pm
Ryerson Laboratory, Room 256

- Sign up via Calendly

Description

Data science provides tools for gaining insight into specific problems using data, through computation, statistics, and visualization. This course introduces students to all aspects of a data analysis process, from posing questions, designing data collection strategies, management, storing and processing of data, exploratory tools and visualization, statistical inference, prediction, interpretation, and communication of results. Simple techniques for data analysis are used to illustrate both effective and fallacious uses of data science tools. Although this course is designed to be at the level of mathematical sciences courses in the Core, with little background required, we expect the students to develop computational skills that will allow them to analyze data. Computation will be done using Python and Jupyter Notebook.

Course Objectives

There are no prerequisites for this course, and students are not expected to have any prior knowledge on prograr or statistics. By the end of the course, students will be able to:

1. Introduce students to the data science pipeline.
2. Develop students' abilities to be informed and critical readers of quantitative, data-based arguments.
3. Enable students to perform data analysis using Python.
4. Develop students' abilities to integrate code, background information, results, and interpretation in a Jupyter Notebook to communicate data analyses effectively
5. Help students gain flexible problem-solving and programming skills applicable to a large variety of problems independently.
6. Understand foundational concepts of probability and statistics
7. Develop skills to apply probability and statistics for data analysis

Required Materials

- Access to a computer with Anaconda Navigator installed
- Data Science I online textbook
- Optional: Python Data Science Handbook by Jake VanderPlas, O'Reilly Media

Office Hours & Contact Information

I encourage all of you to stop by and attend office hours to discuss anything related to course. Feel free to come and talk about coursework and/or career and graduate school aspirations. Although office hours are a designated time for me to talk to students within the course, please schedule an appointment with me via Calendly (at least 24 hours in advance) to best utilize our time together. If you need more time to talk or an alternative meeting time, please email me. Also take advantage of labs/TA office hours if you have scheduling conflicts.

COURSE OVERVIEW AND POLICIES

Communications & Distribution of Course Content

We will use Canvas to post the slides and notebooks from the class, as well as the homework. Questions and discussions will be handled using Ed Discussion for which there is a link in Canvas. There is a file with information on Python and Jupyter Notebooks on Canvas. Other information, including contact for information for TAs, lab schedules, and other major communications will be shared via Canvas.

Class Format & Grading

This course consists of lectures and readings that discuss theories and concepts, labs that apply lecture content, and assessments that include weekly homework assignments, a midterm project, and a final exam. Files that correspond to each lecture will be uploaded to Canvas on a weekly basis. For each lecture, you should expect to refer to the lecture (in the format of a PowerPoint and/or PDF file) and/or a demo (in the format of a Python notebook). Please follow along with demos in class for explanation on how to use functions and data science methods that will be needed to complete labs and homework assignments. Accompanying files for demos will be provided on Canvas. Your final grade will be calculated in the following way:

- **Homework (30%)**: Every week, you will have a homework assignment that will require you to show cumulative proficiency in several skills covered in lectures and readings. Solutions will be discussed during lab sessions the day after the homework is due. Lab sessions will often practice methodologies that will be relevant to your homework assignments. You may discuss homework problems with other students, but you should code and write solutions independently. No late homework will be accepted for grading and no credit will be received. The lowest homework score will be dropped. Attendance of lab will add an additional 2.5 points to your homework for that week (not to exceed 100).
- **Midterm Project (30%)**: You will conduct a midterm project that will use topics covered in this course to analyze one or more real-world data sets. The project will include data exploration, visualization, and communication. Components of the midterm project include a detailed data analysis report, which will be further explained around the time of the project is released.
- **Final (40%)** Your final will be an in-person exam that will cover topics taught over the duration of the course. Information regarding the time, date, location, and other logistics will be provided in the second half of the course. A Pass/Fail grade may be given upon written request to the instructor before the reading period. The grade of P will be awarded only for work of C- quality or better. The grade of Incomplete will be only given in cases of emergency will require a conversation with the instructor and Academic Adviser. The grade of W needs to be requested from and discussed with your Academic Adviser. If you are receiving a degree this quarter, please inform the instructor.

Please refer to our Canvas page for information regarding due dates. Assignments will be posted on our Canvas page, and submission will be through **Gradescope** (link located on the left-hand sidebar in the Canvas page).

Academic Honesty & Misconduct

It is important that we maintain the academic integrity of the course by submitting original work, citing the ideas and work of others, and avoiding plagiarism and academic dishonesty of all forms. The University offers a formal statement regarding academic integrity:

> *Proper acknowledgment of another's ideas, whether by direct quotation or paraphrase, is expected. In particular, if any written or electronic source is consulted and material is used from that source, directly or indirectly, the source should be identified by author, title, and page number, or by website and date accessed. Any doubts about what constitutes "use" should be addressed to the instructor.*

Please familiarize yourself with the Academic Honesty & Misconduct Policy and the Academic Fraud Policy found on the University's webpage on academic policies. If you have any questions about what is or is not proper academic conduct, please ask. When in doubt, air on the side of caution and cite work/ideas that are not your own.

Accommodations & Support

This class respects and welcomes students of all backgrounds, identities, and abilities. Students that have a disability accommodation should provide a copy of your Accommodation Determination Letter from the Student Disability Services office at the beginning of the course. If there are circumstances that make our learning environment and activities difficult, please let me know. I am committed to creating an effective learning environment for all students, but I can only do so if you discuss your needs with me as early as possible. I promise to maintain the confidentiality of these discussions.

Diversity & Inclusion

The intention of this course is to serve students from different backgrounds and perspectives, address students' learning needs, and capitalize on diverse perspectives to enhance the learning of the community. Discussion and materials used from this course is meant to highlight the experiences of persons of various identities and cultures. I welcome any suggestions to improve these materials throughout the course. If you have any preferred names or pronouns, please make that known in the beginning of the course to me and your fellow classmates.

All students are expected to be respectful of each other to foster a comfortable and safe learning environment. If there comes a point within this course when something is said or done by myself or a classmate that is troubling, uncomfortable, or offensive to you, I welcome open and honest dialog to address it. This can be done in several ways:

1. Discuss the situation privately with me. I am always open to listening to students' experiences, and I want to work with students to find acceptable ways to process and address the issue.

2. Discuss the situation with your classmate(s). Discussion enhances the ability for all participants to have a fuller understanding of the context and impact the situation.

3. Notify me of the issue through another source such as your academic advisor, a trusted faculty member, or a peer. If, for any reason, you do not feel comfortable discussing the issue directly with me, I encourage you to seek out another, more comfortable avenue to address the issue.

Zoom & Virtual Lectures

To facilitate participation, discussion, and active learning, lectures will be held in person. Virtual attendance via Zoom is advised only for illness and emergencies. To obtain Zoom information, please reach out to the instructor in advance (at least 1 day). While students may attend lectures via Zoom, in person attendance is highly encouraged for the best learning experience. Lectures will not be recorded, and questions and discussion via Zoom cannot tended to during the lecture. To ask questions in real-time, students must be present in lecture.

Expected Course Timeline

Week	Topics	Relevant Readings
1	Course Intro, Data types, arithmetic operations, comparisons	Chapter 1 Chapter 2 Chapter 3
2	Data structures (lists, arrays, dictionaries, dataframes), NumPy functions	Chapter 4 Chapter 6
3	Pandas and Dataframes: groups and joins, visualization	Chapter 6 Chapter 7 Chapter 9
4	Functions, iteration, and conditionals	Chapter 5
5	Data collection, Intro to probability and probability distributions and random variables	Chapter 10 Chapter 11
6	Random choice, sampling, simulation, large sample behavior, Inference, point estimates, empirical distributions, center and spread, expectation	Chapter 11 Chapter 12
7	Bootstrap and CI Hypothesis testing	Chapter 13 Chapter 14
8	A/B testing Two sample inference	Chapter 13
9	Two-way tables Data ethics	Chapter 13 Chapter 15