

# DATA 11900: Introduction to Data Science II

Instructor: Dr. Amanda R. Kube Jotte

Winter 2024

E-mail: akube@uchicago.edu

Office: Ryerson 256

Class Meetings: Section 2: TTh 11:00 am - 12:20 pm

Location: Ryerson Phys Lab 276

Class Meetings: Section 3: MWF 9:30 - 10:20 am

Location: Stuart Hall 101

Instructor Office Hours: WTh 1:00-2:00 pm in Ryerson 257G

---

## Course Description

**(From Course Listings)** This course is the second quarter of a two-quarter systematic introduction to the foundations of data science, as well as to practical considerations in data analysis. A broad background on probability and statistical methodology will be provided. More advanced topics on data privacy and ethics, reproducibility in science, data encryption, and basic machine learning will be introduced. We will explore these concepts with real-world problems from different domains.

The only prerequisite for this course is Data 118. Students who have not taken Data 118 should have an equivalent expertise in Python and Statistics. You should talk to the instructor if you have questions about this.

This course will cover topics such as data cleaning, reproducibility, regression, classification, clustering, and SQL. These topics will be taught and reinforced through lectures, coding activities, labs, homework assignments, exams, and a project.

## Course Objectives

The objectives of this course are similar to (and extensions of) those of Data 118.

1. To introduce students to the data analysis process.
2. To help students to be informed and critical readers of quantitative data-based arguments.

3. To provide skills that enable students to perform data analysis using Python and integrate code, background information, results and interpretation in a Jupyter Notebook.
4. To help students gain flexible problem-solving skills applicable to a large variety of problems.
5. To introduce students to more advanced topics in machine learning and databases.

## Student Learning Outcomes

1. Students will be able to view and analyse in Python using Jupyter Notebooks and packages such as NumPy, Pandas, Matplotlib, and scikit-learn.
2. Students will be able to think critically about the use and cleaning of data.
3. Students will be able to clean, filter, analyse, and visualize datasets and use these methods to better understand and explain their data.
4. Students will gain an understanding of statistical and machine learning methods including regression, classification, and clustering.
5. Students will be able to use concepts of statistical inference and machine learning to engage with research questions.
6. Students will be able to use basic commands from the database language SQL
7. Students will be informed and critical readers of data-based arguments.

## Course Policies

### Course Materials and Announcements

**Textbooks** You do not need to purchase a printed textbook as I will be assigning readings from free online textbooks:

- “Python Data Science Handbook” by Jake VanderPlas:  
<https://jakevdp.github.io/PythonDataScienceHandbook/>
- “Computational and Inferential Thinking” available at <https://www.inferentialthinking.com/chapters/intro>. This textbook was developed at UC Berkeley for use in their "Data 8" introductory Data Science course. The first few weeks of DATA 11900 will cover the last chapters of the textbook.

For a more advance introduction to Python, we suggest: Wes McKinney. (2017) “Python for Data Analysis”, 2nd edition. O’Reilly Media.

### Software:

You will need the ability to work with Jupyter Notebooks to complete assignments and view

lectures for this course, so you must have access to a computer with Python3 and Jupyter Notebooks installed or access to Google Collab to open and edit the .ipynb files.

### Discussion:

We will use Ed and Canvas for all questions and discussions related to the class. Please post questions on Ed rather than sending an email. This serves multiple purposes. First, others may have the same question. Posting to Ed allows us to clarify the issue for everyone at the same time. Second, we are much more likely to respond in a timely manner if you post on Ed as both instructors and TAs will see the post. Lastly, your fellow students may be able to answer your question. One of the best ways to study is to think about how you would teach the material to someone else. We learn a lot by teaching others [Guerrero and Wiley, 2021, etc].

Before posting to Ed, **please check that no one else has already asked the question.** We will not respond to duplicate questions on Ed. The Ed Discussion for this course is shared between Section 2 and Section 3 of this course, meaning that there are **about 140 students**. Using Ed and using it properly allows us to be much more responsive to all students than if we had to answer questions individually.

All announcements related to the class will be made in class, on Canvas, or on Ed. **I will assume that any announcement made on Canvas or on Ed is known to everyone in class within one business day of it being posted.** It is important to check Ed and Canvas regularly! You are responsible for all announcements made in lecture or online.

### Emails:

If you have something that you want to talk to me about individually, you are encouraged to send me an email. However, I ask that you please include these things in every email that you send:

1. Your full name as it appears on Canvas
2. The number and/or name of the course you are in
3. The section number of the course you are in
4. The name or number of the assignment you are referring to if applicable

If you do not include these, **I will likely not respond.** I get an incredible number of emails every day. I also teach at 2 different schools and have three separate classes at the same time. It saves an incredible amount of time if I do not have to search for your name across all of my courses to figure out who you are and what you need.

### Labs

There will be weekly lab assignments posted to Canvas. These labs are not required, but they are your main source of practice for concepts we cover in class. Not completing your lab will not have a negative effect on your grade, however **completing and submitting a week's lab session will earn you extra credit.** If you complete all labs, you will earn an extra 4% (about a half letter grade) toward your final grade. Completing any portion of the labs will earn you that portion of the extra 4% (ie if you complete half of the labs you will earn 2% or half of the extra credit). We

encourage you to complete your lab during one of the weekly TA office hours so that you can work together with peers and get help from the TAs when you need.

If you are worried about being on the edge of a grade cutoff and email about wanting to be bumped up, the first thing I will ask is whether you completed the labs. These are your opportunity to bump your grade yourself and to show me your commitment to the course.

## Grading Policy

Your course grade will be calculated as follows:

- Homework - 30%
- Project - 30%
- Final Exam - 40%

Letter grades will be assigned as follows:

```
def letter_grade(mark):
    if mark >=93:
        grade = 'A'
    elif mark >= 90:
        grade = 'A-'
    elif mark >= 87:
        grade = 'B+'
    elif mark >= 83:
        grade = 'B'
    elif mark >= 80:
        grade = 'B-'
    elif mark >= 77:
        grade = 'C+'
    elif mark >= 73:
        grade = 'C'
    elif mark >= 70:
        grade = 'C-'
    elif mark >= 67:
        grade = 'D+'
    elif mark >= 63:
        grade = 'D'
    elif mark >= 60:
        grade = 'D-'
    else:
        grade = 'F'
    return grade
```

A Pass/Fail grade may be given upon written request to the instructor before the reading period starts. 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 instructors

and Academic Adviser. The grade of W needs to be requested from and discussed with your Academic Adviser.

Please submit all assignments on time! You will want feedback on your work before completing the next assignments as most topics in this course build on one another. For this reason, **late assignments will not be accepted**. I do not allow extensions. However, I welcome you to talk to me if you are having issues or if an unusual circumstance arises. This is my policy because I do not think it is my place to choose who's reason meets the threshold for an extension.

Because I do not allow late assignments, **I will drop your lowest homework grade**. This is meant to account for any unforeseen issues in the submission process or other problems you may encounter. This includes any reasons you may have asked for an extension. It is best if you save this dropped submission for situations like this! I will not drop additional homework grades or grant additional extensions.

**Side-note:** With the way grades are set up on Canvas, the extra credit will not calculate correctly until all grades are in. For this reason, it may be useful for you to calculate your grade yourself to see where you stand during the course. You can do so using the following equation:  $\text{Current Grade} = 4(\% \text{ of completed labs}) + .3(\text{Current Homework Score}) + .3((\text{Projected})\text{Project Score}) + .4((\text{Projected}) \text{Exam Score})$ .

## Topics to be Covered

Topics we will discuss throughout the quarter will include but are not limited to:

- Data Cleaning
- Exploratory Data Analysis
- Linear Regression
- Multiple Linear Regression
- Feature Importance and Model Selection
- More Methods for Classification and Regression
- Unsupervised Learning and Clustering
- Introduction to SQL

A more detailed calendar will be posted to Canvas and updated as needed. This calendar is subject to change.

## Use of Generative AI

You will not be allowed to use ChatGPT, Google Bard, or any similar large language models in this class unless told otherwise by me. Doing so will be treated as a violation of academic integrity. The reason for this, is that it is important you learn to understand and write code on your own in order to be able to properly use such resources in the future.

## Academic Integrity

Acting with academic integrity means, in brief, not submitting the statements, work, or ideas of others as one's own. Students are expected to comply with University regulations regarding honest work. If you are in doubt about what constitutes academic dishonesty, speak with me before the assignment is due. Failure to maintain academic integrity on an assignment will result in a penalty befitting the violation, up to and including failing the course and further University sanctions. For more information, consult the student manual <https://studentmanual.uchicago.edu/academic-policies/academic-honesty-plagiarism/>.

## Accommodations

**Accessibility:** Students with disabilities who have been approved for the use of academic accommodations by Student Disability Services (SDS) and need a reasonable accommodation(s) to participate fully in this course should follow the procedures established by SDS for using accommodations. Timely notifications are required in order to ensure that your accommodations can be implemented. Please meet with me to discuss your access needs in this class after you have completed the SDS procedures for requesting accommodations. For more information, visit [disabilities.uchicago.edu](https://disabilities.uchicago.edu).

**Accommodations based upon sexual assault:** The University is committed to offering reasonable academic accommodations to students who are victims of relationship or sexual violence, regardless of whether they seek criminal or disciplinary action. If a student comes to us to discuss or disclose an instance of sexual assault, sex discrimination, sexual harassment, dating violence, domestic violence or stalking, or if we otherwise observe or become aware of such an allegation, we will keep the information as private as we can, but as faculty members of University of Chicago, we are required to immediately report it to a Department Chair or Dean or directly to the University's Title IX Coordinator. If you would like to speak with the Title IX Coordinator directly, Bridget Collier can be reached at [bcollier@uchicago.edu](mailto:bcollier@uchicago.edu) or 773-834-6367. Additionally, you can report incidents or complaints to the Sexual Assault Dean-on-Call (SADoC) by calling 773-834-HELP, or by contacting UCPD at (773)702-8181 or your local law enforcement agency. See <https://studentmanual.uchicago.edu/university-policies/the-university-of-chicago-policy-on-title-ix-sexual-harassment/>.

**Bias Reporting:** The University has a process through which students, faculty, staff and community members who have experienced or witnessed incidents of bias, prejudice or discrimination against a student can report their experiences to the University's Bias Education and Support (BEST) team. See: <https://diversityandinclusion.uchicago.edu/resources/reporting-incidents/>

**Mental Health:** Student Wellness' Mental Health professional staff members work with students to resolve personal and interpersonal difficulties, many of which can affect the academic experience. These include conflicts with or worry about friends or family, concerns about eating or drinking patterns, and feelings of anxiety and depression. See: <https://wellness.uchicago.edu/mental-health/>

**Preferred Name and Gender Inclusive Pronouns:** In order to affirm each person's gender iden-

---

tity and lived experiences, it is important that we check in with others about pronouns. This simple effort can make a profound difference in a person's experience of safety, respect, and support. See: <https://inclusion.uchicago.edu/lgbtq-student-life/lgbtq-resources/>

## References

[Guerrero and Wiley, 2021] Guerrero, T. A. and Wiley, J. (2021). Expecting to teach affects learning during study of expository texts. *Journal of Educational Psychology*, 113(7):1281.