Northeastern University

DS 3000 04: Foundations of Data Science

Fall 2022 Course Syllabus (CRN: 19856)

Instructor:	Dr. Eric Gerber	**Times:**	TF 1:35 pm – 3:15 pm
Office:	ME 331	**Class Room:**	WVG 104
Email:	e.gerber@northeastern.edu	**Email prefix:**	[DS 3000]:
Help Zoom:	Help Zoom Link	**Piazza:**	Piazza Signup Link
Gradescope:	Gradescope Link	**Piazza Code:**	fall2022

Communication: This course is entirely in person. There will be three main forms of communication between me and you, included but not limited to: (1) in-class announcements, (2) Canvas, and (3) email. Please check your email regularly, and also be aware that attendance is mandatory, though I will be flexible with excuses for missing. See the attendance section below for more details.

Course Canvas Page(s): This will serve as the main website, where all relevant information and assignments will be posted, where you can access both Gradescope and Piazza, and where you can keep track of your grades. Log in to Canvas at `http://canvas.northeastern.edu` and find our class; you should automatically be added to the course.

Student Help Hours: To facilitate your learning and engagement, I will be available several hours throughout the week at the Help Zoom Link listed at the top of the page, as well as occasionally in-person in my office, Meserve 331. My hope is for you to come ask questions, discuss problems, or simply chat whenever you feel you need help/are bored. Unless informed of a change, the hours will be:

- **Online** Tuesday through Friday: 12:00 pm to 1:15 pm
- **In-person** Tuesday and Friday: 12:00 pm to 1:15 pm

Additionally, you may **always** email me, or ask after class, to set up a time for us to meet that works best for you.

Finally, your teaching assistants will also host weekly help sessions **TBD**.

Prerequisites:

- CS 2510 or DS 2500 (strong background in programming assumed)

Description:
Data Science (DS) is an interdisciplinary field of inquiry concerned with the study and application of systematically extracting generalizable knowledge from data and using this knowledge to draw useful conclusions. This course is an introductory DS class focusing on the foundations of DS as an emerging field. The course introduces core modern DS technologies and methods that provide a foundation for subsequent DS classes. Therefore, the focus of this class is on the breadth of various DS skills, rather than the depth of the topics covered. As a skills-based course, DS 3000 will cover the use of Python for DS and will introduce some of the widely-used essential Python libraries, such as NumPy, pandas, matplotlib, and scikit-learn. More specifically, this class covers:

- working with applied linear algebra in standard numerical computing libraries (e.g. NumPy)
- loading, processing and integrating data from a variety of structured and unstructured soures using Python libraries (e.g. pandas)
- visualizing data using basic techniques and tools (e.g. matplotlib, seaborn, plotly)
- applying introductory concepts in probability, statistics, and machine learning using Python libraries (e.g. scikit-learn)
- using a standard DS tool (e.g. Jupyter Notebook).

Materials: The majority of class time will be spent writing and running computer code. Each student should bring their laptop to each class and be prepared to submit in-class assignments, when they occur, at the conclusion of the class.

Textbook and References: There is **NO** required textbook for this course. Material and lecture notes will be posted and updated on Canvas every week. If you are interested in having a free textbook as a resource, the below is quite useful:

- *Python Data Science Handbook*, by Jake VanderPlas.
  https://jakevdp.github.io/PythonDataScienceHandbook/

Tentative Course Outline: There is a tentative schedule (updated as needed) on the Canvas page. This is meant to give a rough idea of the topics covered over the course of the semester, and of tentative due dates, but is by no means set in stone.

Lectures: You are expected to attend every class period, if able. Be prepared to type and run code in every class and to save your work. There will be a Qwickly Check-In PIN at the beginning of every class that you will have 30 minutes to enter via Canvas and indicate your attendance. See next page for details.

Lecture Notes: All Jupyter Notebooks and relevant in-class material will be posted to Canvas after each class. These materials are meant to help you review, and catch-up should you miss a class. They are not a replacement for attending class.

Push-up Promise: To encourage you to attend and pay attention, I will make a promise to you that has worked out for all parties in the past. For every mistake **YOU** catch me make in lecture over the course of the semester, I will do **ONE** push-up at the end of the last day of class. Every tally will occur only after one of you points out a mistake I made (and after we discuss why/how I made the mistake). This not only gives you an incentive to pay close attention in lecture, but gives me an incentive to limit my mistakes. Don't be afraid to point out my mistakes! We all make them, and it really does help learning!

Attendance (5%): Attendance is worth 5% of your overall grade, and is measured via Qwickly Check-In on Canvas. At the beginning of each class, I will write a PIN for the class on the board and you will have 30 minutes from the start of class to enter it in Qwickly to log your attendance.

Life happens and sometimes you will be unable to attend class. All students will have up to three "no-question" absences they may use. Any further absences **must** be properly justified to me, via email, within 24 hours of the class being completed. For example:

Alice tests positive for COVID and, because she is overwhelmed, forgets to email me for the week of classes she must miss. This is fine, as she only misses two classes. Later in the semester, she spends a week at the hospital because her partner had surgery, and forgets to let me know; this results in a 5% deduction from her attendance grade. Although the first day of the week is excused no-question, the second is not.

Mandatory Office Hours/Piazza Check-In: Additionally, each student is expected to either attend help hours (either in-person or via Zoom) or post to Piazza about something substantive **at least three times** at various points of the semester. Each project team (see below) will also be required to meet (either in-person or via Zoom) with the instructor after the project groups have been created. These instances will count torwards students' attendance grade.

In-class assignments/Labs (5%): We will occasionally practice our python fluency with small group work in-class. These labs will be graded on completeness, not correctness, and (given time) we will review a solution together at the end of class time. These labs are meant to be low-stakes practice sessions, and will be due at the end of class via Canvas submission.

Visitors: My hope is that lab days will start with a short question and answer session with a visitor who works in the Data Science industry. This is a tentative plan, which will rely on the availability of the visitors, but currently I have asked people who work at Meta, for baseball teams, and pharmaceutical companies (among others).

Homework (40%): Homework will typically be assigned and due every week. Assignments will be done through Gradescope. Your TAs and I am very happy to help with homework questions, but you are expected to show independent thought on each assignment. Programming code requires thorough commenting, which must be done independently. Interpretations of generated output must be your own work. If there is evidence that you directly copied answers from **ANY** source, you will receive a score of 0 for the assignment, and further instances of academic misconduct will lead to disciplinary action.

Late Homework: There will be a 25% point penalty for every day an assignment is late. In order to be respectful of your TAs time (since they will be doing the grading), homework will not be accepted if it is more than 48 hours late. The **only** exceptions will be for extreme situations. However, within those 48 hours, each student as three "late day passes" which will automatically be applied for the first three late days. How this works based on when an assignment is submitted:

- Before deadline: Graded in full.
- Within 24 hours after deadline passes: 25% point deduction (or 1 late day pass used)
- Between 24 and 48 hours after deadline passes: 50% point deduction (or 2 late day passes used)
- More than 48 hours after deadline: Not accepted.

Advice on Homework: Make sure you give yourself plenty of time (i.e. multiple days) to complete the homework assignments. They are called programming "languages" for a reason; learning to code is like learning an entirely new language (and a strange one at that). Keep in mind that learning how to search the web for help in coding can be a great aid; your TAs and I will do our best to help guide your intuition in this.

Quizzes (20%): There will four out-of-class quizzes done through Gradescope which will assess your understanding of the course material. They will be similar to homework in difficulty and style. Quizzes will be written to take about 45 minutes, though up to two hours will be given to complete the quiz. You may choose when to take the quiz within the window given to best suit your schedule.

Warning: Under no circumstances may one student observe the quiz before they sign in to Gradescope to officially start their quiz. Observations include both observing the quiz documents as well as receiving any sense of the content or difficulty from other students. Sharing this information will result in academic integrity consequences.

Final Project (30%): You will be required to conduct a final group project focusing on finding a large dataset of interest, cleaning and analyzing the data with appropriate statistical techniques, and interpreting the results in an appropriate manner. The following policies are in place to ensure an equitable share of effort is done by all team members on the project:

- To receive any credit for the final project, you must earn an average of 60% on **both** the homework and quizzes separately.
- At the completion of the project, each group member will be asked to describe the contribution of all other group members with a **Statement of Contributions**. With this information, the instructor may adjust grades of individual group members, including failing students who have not made meaningful contributions to the final project.
- There will be a presentation aspect of the project where each group member will be required to present on something of substance.
- If there are life circumstances preventing your adequate participation in the group project, please reach out to me as soon as possible. The earlier in the semester we're aware of a situation, the more options we have to put supports in place and remedy it.

More details on the project will be shared after the first two weeks of the semester.

Total Grading Policy: Attendance (5%), Labs (5%), Homework (40%), Quizzes (20%), Final Project (30%).

The grading scale will be the traditional 60-70-80-90 for D-C-B-A, but + and - grades will be given based on cut-offs determined after the final.

Important Tentative Dates:

Project Proposals Due	October 14
Veteran's Day	November 11
Thanksgiving Break	November 25
Final Project Due	December 2

Academic Integrity: Under no circumstances may one student view or share their ungraded homework or quiz with another student Sharing or viewing another students ungraded work will result in a failing course grade. This does not extend to discussion of concepts or ideas, but prohibits any sharing of personal code. Academic dishonesty is not tolerated and in addition to course failure, all violations will be reported to OSCCR: http://www.northeastern.edu/osccr/academic-integrity.

Like every computer scientist, you are encouraged to borrow code you find online, so long as it was not written for this class in any semester. Doing so requires that you attribute credit to the source:

- a quick url link comment (e.g. stackoverflow) will suffice.
- you need not cite any python module you import

Disabled Resource Center (DRC): The office is available to assist students who have a legally documented disability or students who suspect that they may have a disability. If you have a disabling condition that may interfere with your ability to successfully complete this course, please contact the Disability Resource Center: http://www.northeastern.edu/drc/.

Student Resources: Your health and wellness is more important than any assignment. Please use the resources below if you are struggling, and don't hesitate to ask for my help!

- **Counseling Center (24/7 support): 877-233-9477** or https://www.northeastern.edu/uhcs/

Title IX: I am a mandatory reporter under Title IX, which means that I am required to report any and all allegations of discrimination to the Title IX coordinator.

Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance.

Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty and staff.

If you or someone you know has been a survivor of a Prohibited Offense, confidential support and guidance can be found through University Health and Counseling Services staff and the Center for Spiritual Dialogue and Service clergy members. By law, those employees are not required to report allegations of sex or gender-based discrimination to the University.

Alleged violations can be reported non-confidentially to the Title IX Coordinator within The Office for Gender Equity and Compliance at and/or through NUPD (Emergency 617-373-3333; Non-Emergency 617-373-2121). Reporting Prohibited Offenses to NUPD does **NOT** commit the victim/affected party to future legal action.

Changes to Syllabus: This course syllabus is intended as a guide. The instructor reserves the right to revise any part of the syllabus during the course. Any changes will be announced during class time and via email.

Some "Encouragement" (via xkcd.com/844/):