This version: January 30, 2023.

Data Science for Everyone
Syllabus

DS-UA 111, Spring 2023

Lectures: Tues. & Thurs. 2:00-3:15p
Cantor, 36 E. 8th St., rm. 200
Labs: Fridays (details on Albert)

Prof. Andrea Jones-Rooy

ajr348@nyu.edu
Office hours: Thursdays, 12-1p
Center for Data Science (CDS)
60 Fifth Ave., rm. 214

Prof. Marco Morucci

mm7936@nyu.edu
Office hours: Thursdays, 4-5p
CDS, 60 Fifth Ave., rm. 765

Teaching Assistants

Name	Email	Office hours	Location (CDS)
Aniket Bote	ab9114	Tues., 4-5p	rm. 244
Cristina Mac Gregor Vanegas	cm5940	Tues., 11a-12p	rm. 763
Srushti Pawar	sxp8182	Mon., 4-5p	rm. 763
Maitri Shah	mns9841	Wed., 11a-12p	TBC

1 Welcome

Welcome to Data Science for Everyone! This is the flagship undergraduate course of the NYU Center for Data Science and the first course in the sequences for the data science major and minor.

This course is for all students, and is especially geared toward those with no prior coding or statistics experience – or at least not both (if you have **both** coding and stats experience, please see us about course placement). Provided you are ready to **work hard**, in just one semester you will learn to conduct your own thoughtful, rigorous, and ethical data science research from start to finish.

We hope to inspire you to continue with your studies in data science and to use your newfound data science powers for good. Welcome to the first day of the rest of your life as a data scientist!

2 Course goals & skills

This course will empower you to understand and use data in a principled way to better explain, make decisions in, and predict outcomes in the world. This course will transform you from a passive consumer of conclusions about data that other people have made into an informed, empowered, and critical reader, evaluator, and producer of discoveries about the world using data. Overall, DS4E will set you up for further advanced study in data science, as well as equip you to conduct original analyses in your area(s) of interest.

By the end of this course, you will be able to:

- Access and analyze public datasets in your areas of interest.
- Assess the quality, usefulness, and limitations of a dataset.
- Conduct statistical analyses to make scientific inferences about the world.
- Evaluate "data-driven" conclusions in the news and other outlets.
- Use data to make predictions about possible outcomes.

In the process, you'll learn to:

- Program in Python, a widely used data science programming language.
- Conduct, evaluate, and interpret a wide range of statistical tests.
- Evaluate the strengths and weaknesses of a dataset and their implications for inference.
- Demystify buzzwords in data science.
- Consider and improve the ethical implications of data science research.

3 Expectations & advice

We believe everyone can be a data scientist **if** they're willing to put in the work.

We expect you to attend all lectures and lab sections, submit all homeworks on time, and complete both exams during the scheduled times (details below). These are necessary but not sufficient requirements for success in this course. In our experience, students who complete all aspects of the course, not just the graded components, tend to earn the highest grades.

Learning data science – including scientific reasoning, programming, and statistics – is a lot like learning a new language. It requires much practice and contact with the material. This course is designed to encourage and reward ongoing skills practice and regular concept review. As with learning a new language, you will experience periods of frustration and confusion. Keep practicing. Soon you will experience the joy of accomplishment and mastery.

We know that some students already have prior programming or statistics experience. This is fine but not necessary for this course. If you do have familiarity with programming and/or statistics, this means that some weeks may be a bit of a review for you. That said, we have also observed over the years that students with prior related experience still find that the data science approach to both programming and statistics is sufficiently novel that there is much to learn.

4 Grading & grading policies

This course provides twice-weekly lectures and a weekly lab section in which you enrolled when you registered for the course. **Students are expected to attend all three**. The information and skills that you need to complete your homework assignments and exams will be provided by the professors and TAs.

Your final grade will be based on:

- **Homeworks (40%)**. Students will have 1-2 weeks to complete each of the five homeworks, which will go out approximately every 1-2 weeks. We reserve the right to change this schedule.
- **Midterm exam (25%):** The midterm exam is on **Thursday, March 9**, in lecture.
- **Final exam (35%):** The cumulative final exam is on **Thursday, May 4**, in lecture.

The exam times are **firm**. If you miss an exam due to some **unavoidable** illness or other serious circumstance, we will not administer a make-up but will give you a waiver from the exam. The term "unavoidable" is key here: we will not issue waivers for voluntary vacations, trips, etc. In addition to not administering make-up exams, we also do not provide sample

exams. Do not ask.

You must turn in all homework **on time**. We will inform you how to turn in your homework, as it will not be accepted by email. Every day late results in a grade level drop: an A becomes an $A-$, an $A-$ becomes a $B+$, and so on. If you are sick or have some other **unavoidable** problem that interferes with your ability to submit your homework on time, please let us know **before** the deadline.

Regrade requests should be handled with care: if you believe there has been a grading error on your assignment or exam, you may submit a regrade request through the online form linked on Brightspace. We will then review your case and, if necessary, update your grade.

Note that we also reserve the right to **lower** your grade if, upon review of your assignment, we determine you were graded too generously on any question. You must submit all grade appeals in writing via email within **one week** of the grade being returned to you.

There is no **fixed** *a priori* curve for final grades, but our distribution is fair and in keeping with other introductory courses at NYU. The best way to do well in this course is to do your work to the best of your ability: doing better will always help your grade. You will be given feedback on your work, which we are happy to discuss with you. Please **avoid** asking about how everyone else is doing relative to you.

There are no exceptions to the above policies. They are here to help and protect you as students as much as they are for us as instructors.

5 Academic integrity policy

The University is very clear that students work is expected to be their own and that plagiarism is not tolerated. The same rules apply here:

1. No assignment on which you receive a grade is collaborative.
2. You may consult with others but all work handed in must be your own.
3. Do not copy another individual's work, answers or ideas.
4. Do not allow another individual to copy your work, answers or ideas.

We will pursue the highest possible punishment for those who violate these policies.

6 Academic accommodations

Academic accommodations are available for student accessibility. Please contact the Moses Center for Student Accessibility (212-998-4980; mosescsa@nyu.edu) for further information.

We recommend that students requesting academic accommodations reach out to the Moses Center **as early as possible in the semester** for assistance.

7 Readings

There are assigned readings for each week of lecture. We strongly recommend completing readings ahead of each week's lectures. All readings are from our course textbook, *Data Science for Everyone: The Book!*. We will post the relevant chapter(s) under Brightspace > Content for each lecture.

After each lecture, we will post the slides from that day to the relevant Content location as well.

8 Course schedule

- **Week 1: Introduction to data science**
  - Lecture 1a (1/24): Introductions, what we mean by "data science"
  - Lecture 1b (1/26): Core skills in data science, the scientific method
- **Week 2: Thinking like a scientist**
  - Lecture 2a (1/31): Causality, experiments
  - Lecture 2b (2/2): Observational data, types of causality, causal mechanisms
- **Week 3: Programming**
  - Lecture 3a (2/7): Mathematical operations, naming, built-in functions, data types, sequences, dictionaries, libraries
  - Lecture 3b (2/9): User-defined functions, control-flow statements, loops, conditional statements
- **Week 4: Statistics 1**
  - Lecture 4a (2/14): Populations & samples, probability distributions & simulation, the Law of Large Numbers (LLN), testing hypotheses
  - Lecture 4b (2/16): Test statistics, p-values, two-tailed tests, comparing two samples, p-hacking

- **Week 5: Statistics 2**
  - Lecture 5a (2/21): Bootstrapping, mean, standard deviation, normal distribution, Central Limit Theorem (CLT)
  - Lecture 5b (2/23): Percentiles, sampling distributions, confidence intervals
- **Week 6: Working with data**
  - Lecture 6a (2/28): Measurement, random errors, systematic errors, validity, exclusion
  - Lecture 6b (3/2): Importing, inspecting, cleaning, & organizing data
- **Week 7: Midterm exam**
  - Lecture 7a (3/7): Midterm review
  - Lecture 8b (3/9): Midterm exam in lecture; no sections this week
- **Spring break**
  - No class 3/14 (Spring break)
  - No class 3/16 (Spring break); no sections this week
- **Week 8: Prediction 1**
  - Lecture 8a (3/21): Correlation, regression, least squares
  - Lecture 8b (3/23): The linear regression model, hypothesis testing for regression
- **Week 9: Prediction 2**
  - Lecture 9a (3/28): Classification, k-Nearest Neighbors (k-NN)
  - Lecture 9b (3/30): Examples of "learners", assessing performance, implementing k-NN in Python
- **Week 10: Machine learning**
  - Lecture 10a (4/4): Supervised vs. unsupervised learning, regression via machine learning, implementing machine learning regression in Python
  - Lecture 10b (4/6): k-means clustering algorithm, unsupervised model validation, implementing k-means in Python

- **Week 11: Frontiers**
  - Lecture 11a (4/11): Reinforcement learning (RL)
  - Lecture 11b (4/13): Natural language processing (NLP)
- **Week 12: Inferences in the real world**
  - Lecture 12a (4/18): Conditioning on a collider, base rate fallacy
  - Lecture 12b (4/20): Check-in on material since midterm
- **Week 13: Ethics in data science**
  - Lecture 13a (4/25): Principles of ethical research
  - Lecture 13b (4/27): Machine bias, algorithmic "fairness"
- **Week 14: Final exam**
  - Lecture 14a (5/2): Final exam review
  - Lecture 14b (5/4): Final exam in lecture; no sections this week