Data Science for Everyone

NYU Center for Data Science

Professor Jones-Rooy
andrea.jonesrooy@nyu.edu

Spring 2019

1 Overview

This course will change your life. It will empower you to understand and use data in a principled way to better explain, make decisions in, and predict the world. By the end of this course, you will be able to:

- Access and interpret many publicly available datasets in your area(s) of interest.
- Conduct original statistical analyses of data to test your own hypotheses about the world, as well as draw meaningful, transparent, and scientific inferences from them.
- Assess the quality, usefulness, and limitations of a dataset.
- Evaluate data-driven conclusions or arguments made in the news or other outlets.
- Apply data to make informed predictions about possible outcomes in the world.

Overall, this course will transform you from a passive consumer of conclusions about data that other people have made to an informed, empowered, and critical reader, evaluator, and producer of data-driven insights. This course will also set you up for further advanced study in data science. In addition, data science is practically a prerequisite for many professions, so this course may be a powerful investment in your career.

2 Skills You'll Learn

In the process of understanding these concepts, you'll specifically learn to:

- Program in Python, a widely used data science computer language.
- Conduct a wide range of statistical tests.
- Understand the various buzzwords and jargon surrounding data science, like machine learning and deep learning, artificial intelligence, algorithms, big data, and more.

The only prerequisite for this course is high school algebra. We do not expect you to have any prior calculus or computer programming experience. If you already have some relevant experience, especially programming, please see me to discuss if this is the right course for you.

3 Expectations

I believe everyone can be a data scientist. But, it takes hard work and dedication.

I expect you to attend all lectures and labs, to come to class having prepared the readings, and to complete all assignments and surveys. These are necessary but not sufficient requirements for success in this course. There is no forced distribution over grades in this course. You can earn an A if you put in the work.

While correlation does not imply causation, students who attend class and do the readings tend to get the highest grades. As with most things in life, you'll get out what you put in.

Finally, everything must be turned in on time. There are no exceptions. This is good training for the real world.

4 Advice

Learning programming and statistics is 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 concept review. If you take the time to complete all surveys, labs, and assignments, you will be a far more skilled data scientist by the end of this course (and you'll do better on the exams).

During lectures, I will present key concepts, skills, and examples. I do not allow electronics in lecture. Please take notes by hand (there are lots of good reasons to do this). Laptops are required during labs, as you will be doing exercises using the skills of the week.

As with learning a language, you will experience periods of frustration and confusion. Keep trying. Learn from each other. Practice finding solutions on your own. Soon, you shall also experience the joy of accomplishment and mastery.

5 Attendance Policy

Come to class.

Everyone has a buffer of three classes you can miss before I mark off for attendance. I do not distinguish between excused and unexcused before this with the exception of a documented prolonged illness or family emergency. Unless you have a situation that requires you to miss class for an extended period of time, I do not offer make-up surveys, labs, homework, or exams, so please do not ask.

6 Grading

You can accumulate up to 100 points during this course as follows:

- Attendance: 10 points (1/4 point each lecture)
- Labs: 10 points (1 point each; I only grade count the top 10 of 12)
- Homework: 10 points (2 point each)
- Midterm: 30 points
- Final: 40 points

Course letter grades will be assigned according to the following scale:

- A = 94-100, A- = 90-93
- B+ = 87-89, B = 84-86, B- = 80-83,
- C+ = 77-79, C = 74-76, C- = 70-73
- D+ = 67-69, D = 65-66
- F = below 65

7 Contacting me

My office hours are Wednesdays from 2-4p or by appointment (can be one-on-one or a small group). Do not drop by without an appointment. I also try to stay on top of email, but don't expect me to reply at all hours. If I haven't replied after 48 hours, please feel free to ping me again. I will be annoyed by anyone who does so more frequently than that. You've been warned.

8 Academic Accommodations

Academic accommodations are available for students with disabilities. Please contact the Moses Center for Students with Disabilities (212-998-4980; mosescsd@nyu.edu) for further information. Students who are requesting academic accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance.

9 Reading

Please prepare all readings listed below in parentheses before the class meets for that day. Our course text is *Inferential Thinking*, and is available for free online. Any chapters listed are from that book unless otherwise specified. Occasionally a few additional readings are added from outside the course text. After each class, I will post the slides from that lecture on our courses site for your reference. There are no readings for labs. .

The datasets we'll use in class can be found on our course GitHub page. For some assignments, I will also ask you to find datasets of your own.

10 Course Outline: Summary

- **Weeks 1-2**: What it means to think like a scientist
- **Weeks 3-6**: Programming and working with data
- **Week 7**: Midterm Exam
- **Weeks 8-12**: Statistics and prediction
- **Week 13**: Ethics and the practice of science

11 Course Outline: Full

Each week has three parts, including two lectures and one lab. Unless otherwise specified, any part ending in 0.1 is a Monday lecture and 0.2 is a Wednesday lecture. Labs are on Fridays.

Week 1: What is data science?

Welcome to the first day of the rest of your life! We're going to get acquainted with what we mean by data science, why it's important, and why you're here (in this classroom; broader philosophical questions are unlikely to be resolved, at least not this week).

By the end of this week, you will be able to: understand and explain what we mean by data science, access and use all the necessary computing tools for the semester, and think about some of the big questions in data science that will guide the course.

- **1.1**: Introduction (no reading)
- **1.2**: What is data science? (Ch. 1: Data Science)
- **Lab 1**

Week 2: Thinking like a scientist

This week we put the "science" in data science. While the phrase "data science" tends to make people think of data, this focus obfuscates the fact that data on its own is just a tool, and it's science where real discovery takes place. It takes principled thinking, including theory development, hypothesis formation, and rigorous testing to actually learn anything from data. Learning to be a scientist is a lifelong pursuit (welcome!), but causal inference, the focus of this week, is a wonderful and important place to start. This week will shape your thinking for the rest of the semester – and your life.

By the end of this week, you will learn: what it means to think like a scientist, how and why to form theories and hypothesis, what we mean by "causal inference" plus several ways to think about causality, and just how hard it is to actually identify causality, the holy grail of science.

- **2.1**: Causality (Ch. 2: Causality and Experiments)
- **2.2**: Causal inference, hypotheses, and experiments ("Causal Reasoning", "Mill's Methods of Induction" (Mill's Methods is optional – for those extra excited!))
- **Lab 2**; HW 1 handed out

Week 3: Programming

We're finally doing some programming! This may be a review or refresher for some, or the very first time you're programming ever (besides our exciting Lab 1.2). You'll learn an overview of thinking like a programmer, as well as some data science-specific techniques that will be building blocks for much future work (in this class and beyond). Welcome!

By the end of this week, you will be able to: Write expressions in Python, think about and work with different data types, and think about and work with sequences.

- **3.1**: Introduction to Python (Ch. 3: Programming in Python)
- **3.2**: Data types and sequences (Ch. 4: Data Types, Ch. 5: Sequences)
- **Lab 3**

Week 4: Organizing data

Tables are the fundamental way data is stored. This week we build them, import them, manipulate them, and start to analyze them!

By the end of this week, you will be able to: build tables in python, import tables from external datasets into python, and conduct early-stage analyses of tables of data.

- **4.1**: No class (President's Day) (Optional: "Viewing Holidays with a Data Modeler's Eyes", "World Public Holidays Calendar")

- **4.2**: Tables (Ch. 6: Tables)
- **Lab 4**; HW 1 due, HW 2 handed out

Week 5: Visualizing data

This week we will learn how to turn tables of data into visualizations to help more quickly understand and communicate patterns in the data! This is COOL and FUN. And will also feel very satisfying to generate. Like you're creating knowledge. Which you are!

By the end of this week, you will be able to understand, generate, and interpret: scatterplots, line graphs, bar charts, histograms, and overlaid graphs.

- **5.1**: Scatterplots, lines, and bars (Ch. 7: Visualization, through 7.1)
- **5.2**: Histograms (Ch. 7 Visualization, 7.2-7.3)
- **Lab 5**

Week 6: Functions & randomness

This week we get into some deeper and more powerful techniques to manipulate data as well as generate insights about it!

By the end of this week, you will be able to: understand and write functions, classify data into categories, join two tables by columns, use randomness to think about treatments and controls, as well as conduct simulations.

- **6.1**: No class (snow day)
- **6.2**: Matthew Daniels lecture
- **Lab 6**; HW 2 due

Week 7: Midterm

It's midterm week! Congratulations on making it this far, and get ready to show off off what you've learned!

By the end of this week, you will be able to summarize, discuss, and demonstrate all the amazing things we've learned so far this semester, including how to: design and conduct research from which you can actually learn things, think through how data might help you ask and answer good questions, conceptualize the kinds of data and manipulations you would need to answer your questions, and actually write a program where you import and analyze data of interest and draw conclusions.

- **7.1**: Midterm review
- **7.2**: Midterm exam in class
- No lab! Congratulations on making it halfway!

Week 8: Pandas, statistics, and the big picture

Pandas and statistics! That classic pair! We're going to incorporate coding from the pandas package for python, which is widely used in data science. You'll find that the coding skills and syntax you already learned with the data science package will help you more easily adopt this new vocabulary. Then, we jump right into the statistics portion of this course, pulling the key insights from our textbook, then adding some useful applications.

By the end of this week, you will be able to: import and begin to analyze data in pandas, a commonly used data science package for python, work with functions and randomness in datasets, and start thinking about and applying foundational statistical methods for inference from data.

- **8.1**: Key concepts in data science, part 1 (Ch. 8: Functions and Tables (skim), Pandas overview and documentation, "Pandas Cheat Sheet")
- **8.2**: Key concepts in data science, part 2 (Ch. 9: Randomness (skim), Ch. 10: Sampling and Empirical Distributions)
- **Lab 8**

Week 9: Statistics, part 1

We did it! We start really thinking about (and acting on) how we can understand (hopefully) real things about the world using data – and understand the limits of what we can know (but why it's still worth trying). This is a next step after the observation work we did with tables and visualizations, specifically: Hypotheses, sampling, and testing!

By the end of this week, you will be able to: understand why the Facebook algorithm (or any information filtering algorithm) is terrifying (short version: Even though we understand the inputs (code), we don't totally understand its outputs or its effects), start doing statistics for making inferences about the world from data(!), and actually define "statistics" and understand both its power and limitations.

- **9.1**: Core concept (Pandas Cheat Sheet, "Who Controls Your Facebook Feed?"
- **9.2**: Hypotheses and comparing samples (Ch. 11: Testing Hypotheses (skim), Ch. 12: Comparing Two Samples)
- **Lab 9**; HW 3 handed out

Week 10: Statistics, part 2

This week is about some key concepts and methodologies that are the backbone of much of statistics – and sets us up for making predictions!

By the end of this week, you will be able to: understand, identify, and describe biases in samples of all kinds, think about how to handle their inevitable presence in your datasets, evaluate biases in published empirical work, and think about the tradeoffs, promise, and unknowns surrounding AI, what it is, and some latest applications.

- **10.1**: Sample biases (Ch. 13: Estimation, "YouTube's Product Chief on Online Radicalization and Algorithmic Rabbit Holes"
- **10.2**: AI and biases (Ch. 14: Why the Mean Matters, Ch. 15: Prediction, just through 15.1)
- **Lab 10**; HW 3 due

Week 11: Prediction, part 1

We're going to learn a few different ways to find out whether two variables are meaningfully different from one another, as well as whether we can learn anything about what one is likely to do based on what the other is doing!

By the end of this week, you will be able to: test hypotheses, conduct correlation analysis (and talk in (even more) profound, awe-inspiring detail about why correlation does not equal causation!), understand and conduct regression analysis, and be very good at talking about the limits of these types of analyses!

- **11.1**: Hypothesis testing and thresholds (Ch. 15: Prediction, from 15.2 onward, "Python for Data Analysis: Hypothesis Testing and the T-Test)
- **11.2**: Correlation and regression (Ch. 16: Inference for Regression (skim), "Common Pitfalls in Statistical Analysis: Linear Regression Analysis")
- **Lab 11**

Week 12: Prediction, part 2

We break down regression in two big ways: How to interpret results and how to think about its limitations (especially in light of the assumptions behind it). We discuss broader ranges of regression models beyond OLS. And, classifiers!

By the end of this week, you will be able to: interpret OLS regression estimation results, think about the assumptions and other limitations of inference from regression results, and understand the basics of classification in machine learning.

- **12.1**: Regression interpretation and limitations ("Interpreting Regression Output", "Introduction to Correlation")
- **12.2**: Regression assumptions and an example (Ch. 17: Classification, Ch. 18: Updating Predictions)
- **Lab 12**

Week 13: Fun with Python

We've been living in some theoretical and abstract statistics the past few weeks – but this week we return to our roots and get back to python! The focus of this week is to practice using python to conduct a full analysis from start to finish (for both the last lab and homework of the semester), to learn a few new fun things in python, and to pick up some final tips on self-teaching after the course is over. Finally, ok, we'll have a little more abstraction when we talk about classification and discover once and for all how to be happy.

By the end of this week, you will be able to: conduct an analysis from start through regression and diagnostics in python, discuss classification at length at any social gathering/job interview and explore building it in python, do a few more nifty things in python and go out and figure out even more things on your own after the semester is over, and explain the major factors associated with happiness (sort of).

- **13.1**: Classification, text as data, and other fun things; HW 4-5 handed out ("Seaborn Cheat Sheet", Machine Learning: Classification", both optional)
- **13.2**: No class; use this time to work on Lab 12
- **Lab**: Review lab results, intro. to data ethics

Week 14: Data & ethics

We wrap it all up with a discussion of how to be ethical while conducting data science, including how to know what's "right" and also remembering the importance of theory-led research and honesty (in data science and ... you know ... all things).

By the end of this week, you will be able to:think about the major areas and debates surrounding data ethics and privacy, think about concrete steps you can take to make sure you're doing "good" and minimally harmful data science, and connect our concrete data science programming and statistical skills to good scientific practices for generating new knowledge and discoveries!

- **14.1**: Data ethics and privacy ("What is Data Ethics?")
- **14.2**: Ethics, data, theory, and models ("The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", "Big Data and the End of Theory?", and if you're interested, see even more:
  - "Why Theory Matters More than Ever in the Age of Big Data"
  - "Big Data Need Big Theory, Too")
- **Lab 14**; HW 4-5 due

Week 15: Conclusion

We review for the cumulative final exam! No readings, no lab, no homework. The only learning objective is to review and practice! I will still hold office hours this week, Wednesday, May 15, 2-4p (usual time), as well as by appointment (these can be one-on-one or in groups, or both, as you prefer). Email me for an appointment outside of office hours.

- **15.1**: Review for final exam
- Final exam during the exam period