Foundations of Data Science

CSC 10800
Department of Computer Science
The City College of New York, CUNY

Course Number: 61813
Room: NAC 1/340
Time: 12:30 PM - 1:45 PM
Schedule: Mondays / Wednesdays
Credits/Hours: 3 Credits/3 Hours

Dr. Patrick Smyth
patricksmyth01@gmail.com
First-Day Information Collection Form
Link to Lab Journal Template

Course Description

This course introduces the fundamental concepts and computational techniques of data science to all students, including those majoring in the Arts, Humanities, and Social Sciences. Students engage with data arising from real-world phenomena—including literary corpora, spatial datasets, and social networks data—to learn analytical skills such as inferential thinking and computational thinking. The competencies learned in this course will provide students with skills that will be of use in their professional careers, as well as tools to better understand, quantitatively and qualitatively, the social world around them. Finally, by teaching critical concepts and skills in computer programming and statistical inference, the course prepares students for further coursework in technology-dependent subjects, such as Digital Humanities.

This course is based on the Fall 2021 iteration of the Data 8: The Foundations of Data Science syllabus and has been further developed under a Creative Commons license.

Course Learning Outcomes:

- Interpret and draw appropriate inferences from quantitative representations, such as formulas, graphs, or tables.
- Use algebraic, numerical, graphical, or statistical methods to draw accurate conclusions and solve mathematical problems.
- Represent quantitative problems expressed in natural language in a suitable mathematical format.
- Effectively communicate quantitative analysis or solutions to mathematical problems in written or oral form.
- Evaluate solutions to problems for reasonableness using a variety of means, including informed estimation.
- Apply mathematical methods to problems in other fields of study.

Course Texts

Adhikari, Andi; John DeNero; and David Wagner. 2021. Computational and Inferential Thinking: The Foundations of Data Science. 2nd Edition.

This textbook is a free online textbook that includes interactive Jupyter notebooks and public data sets for all examples. The textbook source is maintained as an open source project.

Other readings will be assigned from *Data Science from Scratch* by Joel Grus and will be provided as excerpts in digital form.

You may wish to purchase a primer on the Python programming language. Two possibilities here are *Head First Python* and *Learn Python 3 the Hard Way*.

Accessibility and Accomodations

If a reasonable accommodation by myself or City College would help you to achieve your goals for this course, you may either approach me for a discussion of possible accommodations or speak with City College's Disability Services Office. You can connect with Disability Services staff by emailing disabilityservices@ccny.cuny.edu. Universal access is a critical part of my pedagogy, and I am always willing to discuss accessibility improvements, whether or not the discussion is predicated on a formal recognition of a disability by the university.

Assessment Overview

In this course, you'll be evaluated on your course contribution, notes taken in a lab journal, a take-home midterm, and a final project.

- `Course Contribution: 15%`
- `Lab Journal: 15%`
- `Assignment #1 (Data Discovery): 10%`
- `Assignment #2 (Data Exploration and Cleaning): 10%`
- `Midterm Exam (Data Investigation): 20%`
- `Assignment #3 (Data Response OR Technical Tutorial): 10%`
- `Final Project (White Paper): 20%`

Class Contribution

This portion of your grade is based on your participation in weekly discussions and technical work performed during class.

Lab Journal

Your personal lab journal is an honest reflection of your own individual technical study. Every week, you will write the amount of time you spent on technical learning and experimentation and thoughts on your progress. The journal is intended as a mechanism to provide credit to you for time-consuming technical

work, as well as to allow you to calibrate your habits and find out what approaches to study are most effective for you. To encourage the honesty needed for the journal to be useful, you may take two health weeks where you record no entries. Your lab journal can be kept on Google Docs or on GitHub.

Assignment Descriptions

Assignment #1 (Data Discovery)

In this assignment, you will survey an area of research or field of inquiry to find and evaluate existing data sets on the topic. You will find and provide high-level descriptions of three data sets, and choose one data set to describe more fully. The goal of this assignment is to familiarize yourself with various formats in which data can appear, to learn to access data programatically, and to evaluate what kinds of research questions can be addressed with specific forms of data. You will also familiarize yourself with the most common platforms and entities where data can be found.

Assignment #2 (Data Exploration and Cleaning)

In this assignment, you will perform a preliminary analysis on a chosen data set, trying out and answering small research questions, attending curiously to missing or nonsensical data, and cleaning and transforming the data to make it suitable for answering larger research questions.

Midterm Exam (Data Investigation)

For your midterm, you will make a series of observations and extract insights by performing an open-ended analysis of a data set. In the process, you will need to use a certain number of specific techniques and methodologies taught in class up to the point of the exam.

Assignment #3 (Data Response OR Technical Tutorial)

In this assignment, you will critically examine an argument offered in a research paper, white paper, or (data-informed) blog post and respond to the specific claims presented. In responding, you will offer your own write-up based on analysis of a data set.

Alternately, you may write a technical tutorial, with examples, that shows how to perform a task related to data analysis or programming. This can be related to a topic we covered in class, but should represent substantial self-learning outside of lessons in class. This is a good option if you feel you need to learn a new approach or method for your final project that is not otherwise covered in class.

Final Project (White Paper)

In this assignment, you will formally propose a research question, advance hypotheses, and draw a conclusion, all while backing up your arguments with supporting data. In order to fully investigate your research question, you will need to survey a number of data sets, find one or more appropriate to your question, and integrate an analysis of the data into your paper. You will have the option to collaborate in groups of up to four on this assignment, though groups will be expected to produce commensurately substantial results.

Class #	Topic	Reading
1	Welcome, overview, and Expressions	
2	Python: Types and Conditionals	
3	Python: Collections and Iteration	
4	Pandas: Series, Data Frames, and Numeric Methods	
5	Practical Skill: Filesystem Structure and Input/Output	
6	Practical Skill: Finding Data	
7	Pandas: Indexing, Selecting, and Comparison	
8	Python: Writing a function	
9	Pandas: Sorting, Grouping, and Concatenation	
10	Python: Introspection, Objects, and Nested Data Structures	
11	Pandas: Cleaning and Derived Columns	
12	Python: Comprehensions	
13	Ethics: Anonymity and Epistemic Humility	
14	Visualization with Matplotlib: The Humble Pie Chart	
15	Visualization with Matplotlib: Scatterplot and Histogram	
16	Midterm	
17	Statistical Concepts: The Normal Distribution	
18	Statistical Concepts: Correlation	
19	Statistical Concepts: Linear Regression	
20	Statistical Concepts: Regression Inference	
21	Spring Cleaning and Catch Up	
22	NLP: Processing Text	
23	NLP: Extracting Information from Text	
24	NLP: Ngrams and Prediction	
25	NLP: Classifying Text	
26	Case Study #1	
27	Case Study #2	
28	Conclusion: Community and Identifying as Technical	