Data Analysis with Python and SQL
Winter 2025

Course Description

TL;DR:

- This course requires/expects absolutely no background in coding
- You'll learn to use ChatGPT and similar tools to write code
- You'll be writing code yourself every week, in class and on homeworks
- You can use ChatGPT on homeworks and on the exams
- Exams will be in-class and 2 hours (midterm in week 5, final during finals week)

Let's get this out of the way: **A lot of code today is written by ChatGPT.** This isn't a bad thing per se, and it can be a useful adjunct. But ChatGPT is not error-free, so knowing something about coding is still important. In this course, you'll learn enough computer science, statistics, data analysis, and, yes, generative AI to do data analysis at your job.

We'll use two of the most popular programming languages today, Python and SQL. **No prior experience in coding or data analytics is required.** If you have any Python background, let me know and we'll work on finding more challenging practice problems.

You will be writing code every single week. If that scares you a little, that's ok, that's what the instructor and TAs are here to help with! We will carry out analytics, including data cleaning, computing summary statistics, basic visualizations and basic linear regression. No math or statistics background is required, we'll learn everything we need.

This course could be taken as a preparation for Business Statistics (41000), Applied Regression (41100), and Advanced Decision Models with Python (36109). Understanding how to work with data will also be helpful in other coursework even if programming is not required, such as Healthcare Data Analytics (40205).

Learning Outcomes:

1. **Write basic code in Python and SQL** to carry out fundamental data analysis tasks like data import, data cleaning, and data visualization
2. **Prompt ChatGPT and similar genAI tools** to optimize your coding workflow
3. **Define data analytics fundamentals** to enable you to talk intelligently with programmers and analysts on your team

Learning Format

You won't get better at coding by thinking about coding. You have to write code yourself. We'll do most of our learning during class, where you'll work on many hands-on activities, so **attendance is mandatory**.

- **Class format:** Our three-hour weekly sessions will include the instructor demonstrating concepts, followed by students practicing the material in groups or individually. There's not much theory in this course, and most of our time will be spent getting comfortable with hands-on analytics by doing it.

- **Canvas and GitHub:** All materials will be shared via Canvas. Homework assignments will be downloaded and turned in via GitHub and Canvas. We'll explain this in more detail in class.

- **Question and answer forum:** We understand that coding seems much easier when someone is standing in front of you, showing you how to do it. Working on the homework independently will feel like a huge jump, so be prepared for that. To support you, we will be using Slack as a forum for asynchronous questions. TAs and the instructor will be monitoring it closely, and we promise a response in 24 hours (often much faster).

- **Attendance:** You'll check in for attendance yourself via Canvas every week during class time. It will count for 20% of your final grade. I understand things happen sometimes, and I don't want to be the arbiter of what's a valid excuse for missing a session, so everyone gets **one free pass on attendance** for the quarter. This pass will be applied automatically at the end of the quarter.

- **Exams:** There will be an **in-class midterm and in-class final**. The midterm will be during week 5. Both exams will be individual work only. The midterm will use Python and the final will include both Python and SQL. The final will be during finals week (week 10).

Syllabus

Week 1: Introduction to Programming through Python

- What is programming
- What is Python and why do we use it for data analytics
- What is an IDE (integrated development environment) and setting up **Jupyter Lab**
- Introduction to fundamental programming concepts:
    - Variables, data types, container data types (lists, dictionaries, JSON)
    - Control flow: functions, loops, and if/else statements

Week 2: Python for Data Analysis using Pandas

- Introduction to libraries, including the Python scientific stack **(Pandas, numpy)**
- Introduction to object-oriented programming (OOP) concepts (lightly theoretical)
- How to think like a programmer to enable you to continue learning on the job!
- Importing data with Pandas
- Basic data manipulations in Pandas (info(), describe(), value_counts(), isna())

Week 3: Data Cleaning (incl. Missing Data) and Exploratory Data Analysis in Python

- What is exploratory data analysis
- Introduction to **tidy data principles** (hands-on examples)
- Intermediate Pandas (groupby, pivots, missing data and data imputation)
- Hands-on practice cleaning and tidying a real-world dataset

Week 4: Data Visualization and Predictive Modeling in Python

- Plotting using **seaborn** and **matplotlib**
- Basic predictive modeling (linear and logistic regression) using **sklearn**
    - No specific stats or math background is required -- we'll focus on the code
- Metrics for predictive models (R-squared, RMSE, accuracy, and confusion matrix)

Week 5: **In-class midterm** and short lecture on time-series data

- Introduction to the datetime data type in Pandas
- Plotting time-series data in Pandas and seaborn
- **Midterm is in-class, open notes, open Internet, open ChatGPT, etc. The only thing you can't do is use your neighbor's solution.**

Week 6: Tour of Advanced Topics in Python (APIs, Web Scraping, Optional Cloud Lab)

- **APIs** and how to use them, including current use cases in business (hands-on)
- Hands-on experience with several current APIs (including Yelp, Google Maps, and others requested by students)
- Introduction to web scraping and use cases in business
- Hands-on experience with simple web scraping using **BeautifulSoup**
- Introduction to cloud computing concepts with **optional take-home lab in AWS** (using free-tier AWS resources to set up a Python script in the cloud!)

Week 7: Introduction to Databases and Basic SQL Queries

- Database and relational data concepts (lightly theoretical)
- Data architecture and data normalization (lightly theoretical)
- What is SQL and SQL use-cases in business (includes a discussion of current use cases of SQL in enterprise, including in banking/finance)
- Hands-on basic SQL queries (SELECT, WHERE, GROUP BY, HAVING, ORDER BY)

Week 8: Intermediate SQL Programming

- More SQL queries, including UPDATE, CREATE TABLE, DROP and CASE statements
- Intermediate SQL programming, including **sub-queries** and **window functions**

Week 9: Putting Together a Data Project in Python and SQL

- This week's lecture will be a flipped classroom
- You'll do some prep work (~1 hour) prior to class, then class time will be used for a guided, hands-on end-to-end data analysis project! You will work with classmates to complete a data project, from exploration through a simple predictive model, in three hours

Week 10: No instruction for Booth courses

- In-class final scheduled exam time
- The final will require both Python and SQL programming
- The final is open notes, open Internet, open ChatGPT, etc. The only thing you can't use is your neighbor

Course Materials

There are no required textbooks for the course, but I do include some recommended resources below if you'd like to dive deeper into any topics we cover. Where it makes sense, I'll provide chapter excerpts on Canvas.

- *Python for Data Analysis* by William McKinney
- *Automate the Boring Stuff with Python* (2nd edition) by Al Sweigart
- *Learn Python the Hard Way* by Zed Shaw
- W3 Schools SQL tutorials
- Four-hour video introduction to SQL from freeCodeCamp

Prerequisites

None!

Class Meetings

- **Monday** 8.30-11.30am @ Harper
- **Monday** 1.30-4.30pm @ Harper
- **Wednesday** 6-9pm @ Zoom

Instructor and TAs

Instructor: Lara Kattan

- Email: lara.kattan@chicagobooth.edu
- Offce hours: Slack message or email for appointment (zoom during the week or Gleacher/Harper right before/after class)

TAs: Jian Park (Harper), Brandon Ferguson (Gleacher)

- Jian email: jianp@uchicago.edu
- Brandon emmail: bfergus0@chicagobooth.edu
- Office hours:
    - Slack message anytime, 24-hour promised turnaround
    - Zoom by appointment only

Grades

- Homework - 10%
  - Homework will be assigned every week for total of 7 assignments
  - I'll *automatically* drop your lowest HW grade at the end of the quarter.
- Midterm (in-class week 5, individual work) - 20%
- Final (in-class finals week, individual work) - 50%
- Participation/attendance - 20%
  - Attendance will be self-reported through Canvas. Lying about attendance is a violation of the Honor Code.
  - Participation will be tracked through in-class un-graded quizzes.

Generative AI use policy

GenAI is ubiquitous, and being skilled in using GenAI tools is essential to success in the workplace. Recognizing this, you will be instructed in GenAI best practices and allowed to use GenAI tools for all assignments and exams. Any use of GenAI must be cited for the midterm and final exam only. **Un-cited use of GenAI on any exam violates the Booth Honor Code.**

Software and Technology Requirements

All software used in this course is free to use. Python is open-source software and is freely available. For SQL, we will be locally installing a PostgreSQL server on your machine.

For Python, you may use Booth-hosted servers that don't require you to install anything and can be accessed through any browser with an internet connection. You will, however, need a machine that you have permission to install software on for SQL. We will provide instructions for Windows and MacOS operating systems. We will attempt to help with Linux installs, but can't guarantee it will work. Your machine should have at least 4GB of RAM (newer machines sold today will have closer to 16G RAM, but you don't need anything that powerful yet).

Python (Jupyter Lab) server: https://jupyter-class.chicagobooth.edu/hub/login

A short how-to document will be posted on Canvas with links to what you need to install.

Accommodations for Disabilities

The University of Chicago is committed to ensuring the full participation of all students in its programs. If you have a documented disability (or think you may have a disability) and, as a result, need a reasonable accommodation to participate in class, complete course requirements, or benefit from the University's programs or services, please contact Student Disability Services as soon as possible. To receive a reasonable accommodation, you must be appropriately registered with Student Disability Services. Please contact the office at 773-702- 6000/TTY 773-795-1186 or disabilities@uchicago.edu, or visit the website at disabilities.uchicago.edu. Student Disability Services is located at 5501 S. Ellis Avenue.

Honor Code

Students are required to adhere to the standards of conduct in the Chicago Booth Honor Code and the Chicago Booth Standards of Scholarship. These documents are available on the Web at: http://programs.chicagobooth.edu/parttime/docs/handbook.pdf. All graded assignments and exams are subject to the Chicago Booth honor code. In accordance with the Honor Code, students will be required to sign the following pledge on every examination: I pledge my honor that I have not violated the Honor Code during this examination.