DEPARTMENT OF STATISTICS & DATA SCIENCE

THE WHARTON SCHOOL

University of Pennsylvania

STAT 4770 Fall 2024, Q2

An Introduction to Python for Data Science

Syllabus

Instructor: Richard Waterman. waterman@wharton.upenn.edu

Classes meet:

M/W 345 JMHH from 12:00 pm to 1:30 pm.

Office hours:

Waterman: M 10:15am-11:45am. https://upenn.zoom.us/j/5136434021

TA: Zirui Fan (ziruifan@upenn.edu)

BACKGROUND

Python has become the most popular programming language for data science and competency in Python is a critical skill for students interested in this area. This course introduces Python within the context of the closely related areas of statistics and data science.

GOALS

At the end of the course, students will have a solid grasp of Python programming basics and have been exposed to the entire data science workflow. This includes interacting with SQL databases to query and retrieve data through data wrangling, reshaping, summarizing, analyzing and ultimately reporting results. The course will introduce and use popular

Python libraries such as pandas, numpy, seaborn and matplotlib and use the Jupyter notebooks framework for coding.

PREREQUISITES

No prior programming experience is expected, but statistics, through the level of multiple regression is required. This requirement may be fulfilled with undergraduate courses such as Stat 1020, Stat 1120, MBA courses such as Stat 6130/6210, or by waiving MBA statistics.

COURSE MATERIALS

CANVAS

All course materials, including class notes and assignments, will be available on Canvas. We will use the Piazza discussion forum environment.

COMPUTING PLATFORM

All students should install the Anaconda Distribution Platform with includes Python 3.9, available at https://www.anaconda.com/products/individual (there is no need to join the Anaconda Nucleus community). This distribution comes with Jupyter notebooks and the Spyder IDE, together with most of the libraries necessary for the class. Installing the software before the first class is an extremely good idea!

BOOKS

Though there is no required textbook for the class, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd Edition, Wes McKinney, would be an excellent text as a reference.

CLASS NOTES: these will be available from Canvas.

DATA SOURCES

The course will use real life data sets from a variety of disciplines, including health care, finance, cyber security, marketing and internet sources.

CODING BRICKS

Bricks are used for building. In this class, they build knowledge. There will be 4 coding "bricks" to complete during the quarter. These "bricks" will be prescriptive in nature and involve performing a set of programming and data analysis-related tasks in Python. Completing the bricks will teach you the essentials required to write and understand basic Python code. The deliverable will be in the form of a Jupyter notebook which is to be uploaded to Canvas.

You will receive feedback on the coding bricks from the TAs in the form of a score out of 50. The score out of 50 is purely for your own personal skill assessment. The actual score

toward your final grade depends on whether a good faith effort (GFE) as judged by the TAs has been made to answer the questions. A GFE for every question in the brick scores 5 points. A GFE for three quarters or more of the questions scores 3 points, and a GFE for less than three-quarters scores 1 point. It does not matter if the answers are incorrect. The bricks allow you to learn to code in a low-stakes environment. If you use an external source such as Google, StackOverflow, or ChatGPT to answer a question in a brick, you must note that fact with a comment (just as you would reference a primary source in an essay).

TAKE HOME FINAL PROJECT

There is no final exam but rather a take home final project. You can discuss the project with other students but **must write** your own code. If you use code from any outside source (Google, StackOverflow, ChatGPT, etc.), then it must be attributed in the project code itself.

Homework schedule

Deliverable	Due date
Homework 1	11/4, 11:59PM
Homework 2	11/13, 11:59PM
Homework 3	12/2, 11:59PM
Homework 4	12/11, 11:59PM
Take home final project	12/18, 11:59PM

Quizzes

There will be 5 ***in-class*** quizzes. They are closed book. Each quiz has 5 questions (multiple choice) and will take 8 minutes. You can drop the lowest quiz score. There will be no make-up quizzes. You have a single attempt for the quiz.

Quiz schedule

Deliverable	Date
Quiz 1	10/28
Quiz 2	11/6
Quiz 3	11/13
Quiz 4	11/20
Quiz 5	12/4

COURSE MODULES

Module 1	Python Bootcamp I: introduction to Python.
Module 2	Python Bootcamp II: Jupyter notebooks.
Module 3	Introducing pandas.
Module 4	SQL databases, retrieving and joining data. Portable data formats: csv and json. Dates and times in pandas.
Module 5	Writing basic functions. From transactions to behavior.
Module 6	Data visualization using Seaborn and Matplotlib.
Module 7	Statistical modeling with statsmodels.
Module 8	Machine learning with SKLearn.
Module 9	Scientific computing and numpy

Class content

MODULE 1. Python Bootcamp I

In this module we will start to get to know Python, its syntax, and capabilities. We will introduce the Spyder IDE and Jupyter notebooks, both of which come with the Anaconda Distribution.

MODULE 2. Python Bootcamp II

More Python fundamentals and more on Jupyter notebooks.

MODULE 3. Introducing pandas

Once data is accessible within Python, a key step in the data science pipeline is wrangling that data, which includes, cleaning, merging, reshaping, and summarizing/aggregating them. This module introduces the pandas library that facilitates this step.

MODULE 4. SQL databases, joining and retrieving data. Data formats, csv, json.

Most real business data sets are stored in relational databases, and this class introduces these databases and shows how to access them using Python. Data is also moved around in various formats, and we will illustrate some of these with a discussion of the csv, html and json formats, and again, how to import them into Python. We will also discuss the use of Beautiful Soup to scrape web data.

MODULE 5. Writing basic functions. From transactions to behavior

Reusing and organizing code is important. Functions help achieve these goals. We will also discuss topics such as handling missing data and common data-cleaning tasks within the panda's framework. Combing the "groupby" command with bespoke simple functions allows one to move from transactional to behavioral data with ease.

MODULE 6. **Data visualization using Seaborn and Matplotlib**

Visualization allows the analyst to gain insight to data as well as sharing their findings in a compelling and engaging way. We will use the popular Python libraries, seaborn and matplotlib for this step.

MODULE 7. **Statistical modeling with statsmodels**

Once a data set is suitably organized, modeling and data mining tools can be applied. We start by looking at hypothesis testing and multiple regression,

MODULE 8. **Machine learning with SKLearn**

This module will introduce and show the implementation of the foundational decision trees and then move on to the random forest, discussing the train/test paradigm and the hunt for good tuning parameters.

MODULE 9. **NumPy for simulation modeling**

NumPy is Python's popular platform for scientific computing. It also comes with computationally efficient data structures, in particular, arrays. We will explore this platform in the context of basic linear algebra, Monte Carlo simulation modeling, and optimization.

GRADING

The final grade will be weighted using 24% from the four assignments (each counts the same), 45% from the final project, 26% from the quizzes, and 5% for attendance and participation. You can drop the lowest quiz score. There will be **no** extra credit opportunities at the end of the course. Grades are not rounded up. Grade queries must be submitted within one week of the coding brick solutions being posted. Late homeworks (coding bricks) are penalized by 25% for up to 24 hours late, 50% for up to 48 hours late, and after 2 days late homework is not accepted.

CLASSROOM EXPECTATIONS

I expect you to conduct yourself in a way that does not distract other students' learning. Questions are strongly encouraged.

ACADEMIC INTEGRITY

All relevant University policies regarding Academic Integrity must be followed. Please consult the Code of Academic Integrity for details and clear descriptions of prohibited actions. Any violation of the Code will automatically lead to a FAIL or F grade. Violation of the MBA Code of Ethics may lead to additional sanctions.

COURSE CALENDAR Q2 FALL 2024

DATE	Class #	Activity
10/21/2024	1	
10/23/2024	2	
10/28/2024	3	Quiz 1 in class.
10/30/2024	4	
11/4/2024	5	HW 1 due.
11/6/2024	6	Quiz 2 in class.
11/11/2024	7	
11/13/2024	8	Quiz 3 in class. HW 2 due.
11/18/2024	9	
11/20/2024	10	Quiz 4 in class.
11/25/2024	11	
12/2/2024	12	HW 3 due.
12/4/2024	13	Quiz 5 in class.
12/9/2024	14	
12/11/2024		HW 4 due.
12/18/2024		Final project due.