Wharton
UNIVERSITY *of* PENNSYLVANIA

DEPARTMENT OF STATISTICS & DATA SCIENCE

THE WHARTON SCHOOL
University of Pennsylvania

STAT 7770 Spring 2023, Q3

An Introduction to Python for Data Science

Syllabus

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

TA: TBD

This class is online. Classes will be recorded, and recordings made available to all students. There is **no attendance requirement**, so a student may choose to cover the material asynchronously or attend the lectures live.

Classes meet:

MW 7-8:30PM.

Office hours:

Waterman: MW 1:30PM – 3:00PM. https://upenn.zoom.us/j/5136434021

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 to 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.

HOMEWORK

There will be 4 homeworks. These homeworks will be prescriptive in nature and involve performing a set of programming and data analysis related tasks in Python. The deliverables will be in the form of Jupyter notebooks which will be uploaded to Canvas. There is no final exam but rather a take home final project. You can discuss the homework with other students, but you **must write** your own code. If you use code from any outside source, then it must be attributed in your homework code itself. When appropriate, homeworks will be run through Turnitin to determine originality. Late homeworks are penalized by 25% up to 1 day late and 50% up to two days late. Homeworks more than two days late will receive a 0.

Homework schedule

Deliverable	Due date
Homework 1	2/1, 11:59PM
Homework 2	2/10, 11:59PM
Homework 3	2/20, 11:59PM
Homework 4	3/3, 11:59PM
Take home final project	3/15, 11:59PM

Quizzes

There will be 5 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. Quizzes need to be taken during the assigned day.

Quiz schedule

Deliverable	Date
Quiz 1	1/30
Quiz 2	2/6
Quiz 3	2/13
Quiz 4	2/20
Quiz 5	2/27

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	NumPy scientific computing and simulation modeling.

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 data bases, 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 pandas 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 50% from the four assignments (each count as 12.5%), 30% from the final project and 20% from the quizzes. All assignments will be included in the final grade. There is **no** "drop the lowest score" policy for the assignments, but 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. Undergraduate and MBA sections will be graded separately. Grade queries must be submitted within one week of the homework solutions being posted.

For undergraduates you can expect that a final score in the 90's will receive some form of A, 80's some form of B and 70's some form of C. I reserve the right to curve the grades. There is no pre-specified percentage of students that will receive a particular grade – everyone could get an A, or everyone could get a C, but that's unlikely!

For MBAs I will simply follow the standard grading constraints of an average GPA of 3.5.

CLASSROOM EXPECTATIONS

There is no formal participation component to the final grade, but questions are strongly encouraged.

COURSE CALENDAR Q3 SPRING 2023

DATE	MBA class #	Activity
1/18	1	
1/23	2	
1/25	3	
1/30	4	Quiz 1 in class.
2/1	5	HW 1 due.
2/6	6	Quiz 2 in class.
2/8	7	HW 2 due 2/10.
2/13	8	Quiz 3 in class.
2/15	9	
2/20	10	Quiz 4 in class. HW 3 due.
2/22	11	
2/27	12	Quiz 5 in class.
3/4		HW 4 due.
3/15		Final project due.