DEPARTMENT OF STATISTICS

THE WHARTON SCHOOL

University of Pennsylvania

STAT 477X/777X & OIDD 477X/777X Quarter 1, Fall 2020

An Introduction to Python for Data Science

Syllabus

Instructor: Richard Waterman waterman@wharton.upenn.edu 443 JMHH

Classes meet: TBD

Office hours: TBD

Teaching Assistant: TBD

Office hours: TBD

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, starting from interacting with SQL databases to query and retrieve data, through data wrangling, reshaping, summarizing, analyzing and ultimately reporting their results. The course will introduce and use popular Python libraries such as pandas and NumPy, and all analyses will be performed using Jupyter notebooks. Students will be expected to maintain a GitHub repository for their code,

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 102, Stat 112, MBA courses such as Stat 613/621, or by waiving MBA statistics.

Course Materials

Canvas

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

Computing Platform

All students should install the Anaconda Distribution of Python 3.7 available at https://www.anaconda.com/distribution/ . It comes pre-installed with the majority of the libraries necessary for the class.

Books

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

Class notes: these will be available from Canvas.

Homework

There will be 5 homeworks during the quarter. These homeworks will be prescriptive and involve performing a set of programming related tasks in Python. The deliverables will be in the form of Jupyter notebooks. There is no final exam but rather a take home final project. You may discuss the homeworks 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.

Data sources

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

Homework schedule

Deliverable	Due date
Homework 1	
Homework 2	
Homework 3	
Homework 4	
Homework 5	
Take home final project	

Quizzes

There will be 5 in-class quizzes. Each quiz has 5 questions (mainly multiple choice) and will take 10 minutes (you can leave after 5 minutes). You can drop the lowest quiz score. There will be no make-up quizzes.

Quiz schedule

Deliverable	Date
Quiz 1	
Quiz 2	
Quiz 3	
Quiz 4	
Quiz 5	

CLASS SCHEDULE

Table 1 Class schedule

Module 1	Python Bootcamp I
Module 2	Python Bootcamp II: Jupyter notebooks and GitHub
Module 3	SQL databases, joining and retrieving data. Portable data formats: csv, json.
Module 4	Data wrangling, reshaping and summarizing with pandas
Module 5	Data visualization using Seaborn
Module 6	NumPy for simulation modeling
Module 7	Statistical modeling and machine learning with SKLearn

Class content

MODULE 1. Python Bootcamp I, Jupyter notebooks

In this class we will get to know Python, its syntax and capabilities. This includes installing Python and creating a first Jupyter notebook.

MODULE 2. Python Bootcamp II, and GitHub

Maintaining a code repository is a good way of organizing, revising and sharing code. We will discuss these ideas using a GitHub repository.

MODULE 3. 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 these with a discussion of the csv and json formats, and again, how to import them into Python. We will also discuss the use of API's to automate data retrieval from web-based sources, for example, the Twitter API.

MODULE 4. Data wrangling, reshaping and summarizing with 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 5. Data visualization using Seaborn

Visualization allows the analysts to gain insight to data as well as sharing their findings in a compelling and engaging manner. We will use the popular Python library, Seaborne, for this step.

MODULE 6. 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 Monte Carlo simulation modeling.

MODULE 7. Statistical modeling and machine learning with SKLearn

Once a data set is suitably organized, modeling and data mining tools can be applied. We will demonstrate these tools, using hypothesis tests, multiple regression, and classification trees.

GRADING

The final grade will be weighted using 60% from the five assignments (each counts as 12%), 20% 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. Grade queries must be submitted within one week of the homework solutions being posted.

CLASSROOM EXPECTATIONS

There is no formal participation component to the final grade but questions are strongly encouraged. Phones, laptops and other electronic devices are not to be used in class except for Python related activities.