DEPARTMENT OF STATISTICS

THE WHARTON SCHOOL

University of Pennsylvania

STAT 177 920 Summer 2020

An Introduction to Python for Data Science

Syllabus

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

Classes meet: Tu/Th 6:40PM – 10PM EST

Office hours: Mo 6:00PM – 8:00PM EST

Teaching Assistant: TBA

Office hours: TBA

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 results. The course will introduce and use popular Python libraries such as Pandas and NumPy, and use the Jupyter notebooks framework.

PREREQUISITES

No prior programming experience is required, but some exposure to statistics would be helpful, though not a necessity as relevant concepts will be covered in class.

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 3.7, available at https://www.anaconda.com/distribution/ . It comes with Jupyter notebooks and the Spyder IDE, together with the majority 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 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 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 may not discuss the homeworks with other students, and 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 internet sources.

Homework schedule

Deliverable	Due date
Homework 1	July 13, 11:59PM EST
Homework 2	July 20, 11:59PM EST
Homework 3	July 27, 11:59PM EST
Homework 4	August 3, 11:59PM EST
Take home final project	August 12, 11:59PM EST

Quizzes

There will be 5 in-class online quizzes. They are open book. Each quiz has 5 questions (multiple choice) and will take 10 minutes. You can drop the lowest quiz score. There will be no make-up quizzes.

Quiz schedule

Deliverable	Date
Quiz 1	July 9
Quiz 2	July 16
Quiz 3	July 23
Quiz 4	July 30
Quiz 5	Aug 6

COURSE MODULES

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

Class structure

Classes meeting dates are July 2, 7, 9, 14, 16, 21, 23, 28, 30, and August 4 and 6. The class lasts three hours and twenty minutes, starting at 6:40PM EST. This will be broken into three segments with two ten minute breaks:

6:40PM – 7:40PM.

Break (10 minutes)

7:50PM – 8:50PM.

Break (10 minutes)

9:00PM – 10:00PM.

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. **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 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, and use Beautiful Soup to scrape web data.

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 and Matplotlib**

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 libraries, Seaborne and Matplotlib 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 four assignments (each counts as 15%), 25% from the final project and 15% 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.