University of Pennsylvania
The Wharton School
Operations, Information and Decisions Department
Department of Statistics and Data Science

OIDD/STAT 4770
Fall 2022
Introduction to Python for Data Science

Syllabus

0. Logistics **2**

1. Background **2**
1.1 Goals 3
1.2 Prerequisites 3

2. Course materials **3**
2.1 Canvas 3
2.2 Computing platform 3
2.3 Books 3
2.4 Class notes 3
2.5 Data sources 4

3. Homework, Quizzes, and Final Project **4**
3.1 Homework 4
3.2 Quizzes 4
3.3 Final project 5

4. Class participation **5**

5. Grading **6**
5.1 Grade breakdown 6
5.2 Grade query procedure 6
5.3 Grade cutoffs 6

6. Course structure **7**

7. Key dates **8**

0. Logistics

Instructor:
Dean Knox <dcknox@wharton.upenn.edu>

Head TA:
Guilherme Duarte <gjduarte@wharton.upenn.edu>

TAs:
Haoran Liu <haoranl@wharton.upenn.edu>
Lavanya Neti <lvneti@wharton.upenn.edu>
Asha Pereira <ashapere@sas.upenn.edu>

Canvas link:
https://canvas.upenn.edu/courses/1675132

Classes meet:
Tuesdays and Thursdays
8:30–10:00am (401)
10:15–11:45am (402)
JMHH G55

Office hours:

Instructor or TA	Day	Time	Location
Knox	Tuesdays	4:00–6:00pm	JMHH 551
Duarte	TBA	TBA	TBA
Liu	TBA	TBA	TBA
Neti	TBA	TBA	TBA
Pereira	TBA	TBA	TBA

1. 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.

1.1 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`, `plotnine`/`ggplot`, `statsmodels`, and `scikit-learn`; it will use the Jupyter notebooks framework for coding.

1.2 Prerequisites

No prior programming experience is expected, but a familiarity with statistics through the level of multiple regression is required. This requirement may be fulfilled with undergraduate courses such as STAT 102 or 112, MBA courses such as STAT 613 or 621, or by waiving MBA statistics.

2. Course materials

2.1 Canvas

All course materials—including class notes, forum, quizzes, and assignments—will be available through Canvas. We will use the Ed Discussion discussion forum environment.

2.2 Computing platform

All students should install the Anaconda Distribution Platform containing Python 3.9, available at <https://www.anaconda.com/distribution>. It comes with Jupyter notebooks the Spyder integrated development environment, and most of the libraries necessary for the class. It is highly recommended that you install the software before the first class.

2.3 Books

There is no required textbook for the class. However, the following free references may be helpful for learning more.

> Charles Severance. 2016. *Python for Everybody: Exploring Data Using Python 3*. <https://www.py4e.com/book>
>
> Wes McKinney. 2022. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter* (3rd edition). <https://wesmckinney.com/book/>

2.4 Class notes

Lecture slides and other notes will be available on Canvas.

2.5 Data sources

The course will use real life data sets from a variety of disciplines. Most notably, we will use the BetterUp-Wharton CANDOR data set, a multi-modal corpus of naturalistic conversation. Other data sets used may touch on health care, finance, marketing and technology.

3. Homework, Quizzes, and Final Project

3.1 Homework

There will be six homework assignments. These assignments 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 code itself. Assignments may be run through Turnitin to determine originality. Late assignments are penalized by 25% when received up to 1 day late and by 50% up to two days late. Assignments more than two days late will receive a 0.

Deliverable	**Due date**
Homework 1	Sep. 13, 11:59PM EDT
Homework 2	Sep. 29, 11:59PM EDT
Homework 3	Oct. 11, 11:59PM EDT
Homework 4	Oct. 25, 11:59PM EDT
Homework 5	Nov. 8, 11:59PM EDT
Homework 6	Nov. 29, 11:59PM EDT

3.2 Quizzes

There will be six in-class quizzes. They are open book. Each quiz will be multiple-choice, take place in the first 10 minutes of class, and be submitted through Canvas. You can drop the lowest quiz score. There will be no make-up quizzes, as solutions will be discussed immediately after the quiz. You have a single attempt for the quiz.

Deliverable	**Due date**
Quiz 1	Sep. 6, ten minutes after class start

Quiz 2	Sep. 20, ten minutes after class start
Quiz 3	Oct. 4, ten minutes after class start
Quiz 4	Oct. 18, ten minutes after class start
Quiz 5	Nov. 1, ten minutes after class start
Quiz 6	Nov. 15, ten minutes after class start

3.3 Final project

The final project will give you a chance to use the various data skills acquired during the course by analyzing a dataset and presenting the results. It is an open-ended creative project in which you will tell a compelling story based on the data, statistical analysis, and graphics. You should prepare a high-level analysis pitch comprising about 20 slides that would take roughly 15 minutes to deliver.

Deliverable	**Due date**
Final project	Dec. 8, 11:59PM EDT

4. Class participation

Two forms of contributions to collective learning will be considered. The first is participation in-class discussion: students may receive credit by asking questions that clarify difficult concepts or by drawing connections between course material and real-world data-science applications. The second is participation in Ed Discussions: students may receive credit by answering peers' questions. This is highly encouraged, as the best way to attain mastery of a subject is by explaining it to others. There is no prespecified division of credit between in-person and online participation; for example, students may obtain full credit by answering questions in Ed Discussion only, as in-class time is constrained by the material that we will cover.

Name tents are required. However, attendance for its own sake will not be considered in grading; it is necessary but not sufficient for in-class participation.

The in-class use of electronic devices for the express purposes of responding to quizzes, following along in course material, and coding in Python is encouraged. However, this use should not distract from your awareness of and ability to participate in the course. Other uses of electronic devices are not permitted. All devices must be silenced at the start of class.

5. Grading

5.1 Grade breakdown

Grades will be assigned as follows:

- 60% will come from the assignments (10% each, with all assignments included)
- 20% will come from quizzes (4% each, with the lowest dropped)
- 15% will come from the final project
- 5% will come from class participation
-

There will be no extra credit opportunities at the end of the course.

5.2 Grade query procedure

Grade queries must be submitted within one week of the homework solutions being posted. To submit a query, you must send an email:

- To the full teaching staff (email addresses duplicated below for convenience)
  - dcknox@wharton.upenn.edu
  - gjduarte@wharton.upenn.edu
  - haoranl@wharton.upenn.edu
  - lvneti@wharton.upenn.edu
  - ashapere@sas.upenn.edu
- Indicating the specific question or sub-question of concern in the subject line
- Describing the nature of the concern in the body
- Attaching the original homework as submitted on Canvas (note that file integrity will be verified by comparing the SHA-1 hash of the attachment and original submission)

Queries that do not adhere to these requirements (e.g., received after one week, directed to an individual teaching assistant or instructor, attaching homework that has been subsequently modified from the original submission) will not be considered.

5.3 Grade cutoffs

I reserve the right to curve grades upwards if needed. Grades will not be curved downwards.

Grade	**Lower cutoff**
A+	97%
A	93%
A-	90%
B+	87%

B	83%
B-	80%
C+	77%
C	73%
C-	70%
D+	67%
D	63%
D-	60%

6. Course structure

MODULE 1
Python Bootcamp
- Interacting with Python: using Jupyter notebooks, getting help
- Variables and containers
- Importing modules
- Control flow

MODULE 2
Data manipulation and analysis with the `pandas` library
- `pandas` syntax
- Cleaning, merging, reshaping data

MODULE 3
Data visualization
- Basic graphics with `matplotlib`
- Grammar of graphics with `ggplot` and `plotnine`
- Principles of visualization from Tufte onwards

MODULE 4
Data collection
- Interacting with CSV, JSON, SQL data storage formats in Python
- Querying relational databases
- Scraping web data with `requests` and `BeautifulSoup`

MODULE 5
Data wrangling
- Common data cleaning tasks

- Handling missing data
- Aggregating transactional data into
- Reusing and organizing code

MODULE 6
Basic statistical analysis with `statsmodels`
- Hypothesis testing
- Multiple regression

Machine learning with `scikit-learn`
- Variable selection with LASSO
- Classification with trees, random forests, and support vector machines
- Dimension reduction with principal component analysis

MODULE 7
Numerical methods with `numpy` and `pyomo` (time permitting)
- Efficient arrays and linear algebra operations
- Monte Carlo simulation modeling
- Optimization

7. Key dates

Quiz 1	Sep. 6
Homework 1 due	Sep. 13
Quiz 2	Sep. 20
Homework 2 due	Sep. 29
Quiz 3	Oct. 4
Homework 3 due	Oct. 11
Quiz 4	Oct. 18
Homework 4 due	Oct. 25
Quiz 5	Nov. 1
Homework 5 due	Nov. 8
Quiz 6	Nov. 15
No class	Nov. 22
Homework 6 due	Nov. 29
Final project due	Dec. 8