**HUDM5001 -Programming for Data Science**

Fall 2024

Teachers College, Columbia University

---

**Lecture:** Tue. 11:00am-12:40pm **Credits:** 3

**Location:** GD 277 **Instructional Mode**: in-person

**Instructor:** Youmi Suk
**Email**: ysuk@tc.columbia.edu
**Office:** 552 Grace Dodge
**Office Hours**: Tue. 4:00-5:00pm, In-Person. Online sessions available upon request.

**Course Assistant:** Wenxuan Wang; Xiran Wen
**Email:** ww2681@tc.columbia.edu; xw2969@tc.columbia.edu
**Online Office Hours**:
Mon. 11:00am-12:00pm
Thr. 3:00-4:00pm

**Course Overview and Learning Outcomes**:

This course is an introduction to essential programming concepts, structures, and techniques for data science. Topics covered include data types, data structures, control statements, and functions, using the NumPy and Pandas libraries in the programming language Python. The course also covers version control using GitHub and database management using SQLite. Additionally, content on the development of interactive plots and dashboards using Plotly and Dash libraries will be included.

At the end of the course, students will

(1) Be able to confidently work in an appropriate programming environment (IDE).
(2) Correctly describe basic Python language constructs and develop Python codes and write basic programs.
(3) Understand the version control concepts and work on a data science project using GitHub and Python.
(4) Create a portfolio showcasing your visualization skills.

**Prerequisites:**

No prerequisites. But students should have some experience working with any programming or statistical analysis software, e.g., R, SPSS, STATA, or MATLAB.

**Textbook:**
McKinney, W. (2017) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd Edition). O'Reilly Media (available on Columbia Library EResources)

**Software or Development Tools**:

- Anaconda
- Spyder IDE (or other IDE tools)
- GitHub
- SQLite

**Topics Covered**

- Intro to GitHub
- Intro to SQL
- Python Programming
  - Intro to Spyder
  - variables and expressions
  - data types: int, float, bool, string, list, tuple, set, dict, range
  - operators
  - input/output
  - numpy
  - pandas
  - sqlite database
  - control structures
  - iterables and iterators
  - list comprehensions
  - functions
  - lambda functions
  - running scripts at the command line
  - classes
  - unit testing
- Python Data Visualization
  - Matplotlib, Plotnine, Plotly, and Dash

**Important Dates**

Date of First Live Session: Sep 3, 2024
Date of Last Live Session: Dec 10, 2024 (final meeting day)
In-class Quizzes: Oct 15 and Nov 19
Midterm Presentation Date: Nov 26, 2024
Semester Project Due Date: Dec 17, 2024 at 11:59 pm

**Course Schedule**

The following calendar is an outline of the course topics and assignments.
(A = Programming Assignment; Q = Take-home Quiz; "Q" = In-class Quiz)

| Week/Date | Topics | Reading | Assignments (due Mon) |
|---|---|---|---|
| Week 1<br>9/3 | Syllabus Review and the Shell | | Q<br>Pre-class S. |
| Week 2<br>9/10 | GitHub | [1] [2] [3] | A |
| Week 3<br>9/17 | Python Intro: Data Types, Variables, and Expressions | McKinney Chs 1 & 2<br>[4] | A, Q |
| Week 4<br>9/24 | Python Intro: Operators, Input/Output, Numpy | McKinney Ch 4<br>[5] [6] | A |
| Week 5<br>10/1 | Python Intro: Pandas | McKinney Ch 5<br>[7] [8] | A, Q |
| Week 6<br>10/8 | Pandas, SQLite Database | [8] [9] | A, "Q" |
| Week 7<br>10/15 | W6 "Q"; Control Structures and Iterables | [10] [11] | A, Q |
| Week 8<br>10/22 | Functions, Lambdas, List Comprehensions | McKinney Ch 3.2<br>[12] [13] [14] [15] | A, Q<br>Topic (10/23) |
| Week 9<br>10/29 | Recursion and Running from the Command Line | [16] [17] [18] | A, Q |
| Week 10<br>11/5 | College Holiday (Election Day) | | |
| Week 11<br>11/12 | Python Classes | [19] (up to 9.5)<br>[20] | A, "Q" |
| Week 12<br>11/19 | W11 "Q"; Unit Testing and Exception Handling | Unit test notes | A,<br>Mid. slides |
| Week 13<br>11/26 | Jupyter Notebook Presentation (Midterm) | | Q |
| Week 14<br>12/3 | Data Visualization: Matplotlib, Plotnine, and Plotly | McKinney Ch 9.1<br>[21] [22] | A, Q |
| Week 15<br>12/10 | Dash | [23] [24] | A, Q |
| Week 16<br>12/17 | Final Project | | Final report (due: 12/17) |

**Programming Assignments, Quizzes, the Midterm, and the Final Project**

*Programming Assignments:* The programming assignments consist of focused exercises related to each week's lectures. You are encouraged to first try to complete the homework by yourself. If you work with others, please make sure that you understand all of the work, and that your final submission is your own work. The assignments will be uploaded on GitHub no later than Tuesday and will be due the following Monday at 11:59pm, ET. The total possible points for each assignment will vary, and specific grading criteria will be provided with each assignment. Our CAs will view students' submissions and make comments on them. Assignments are expected to be completed by due dates. Assignments turned in late will be subject to the following penalty: 10% of the total score will be deducted for each day past the due date. An assignment with the lowest score will be dropped when computing the final letter grade at the end of the semester.

*Quizzes*: There will be several quizzes throughout the semester that will assess your knowledge of the various topics. Quizzes are based on the Jupyter Notebooks. All quizzes are mandatory for all students to take. Importantly, the quizzes should be done in a "closed book" format, which means you should not consult any resources including notes, books, the web, devices, or other external media. Quizzes will be administered either as take-home or in-class tests. In-class quizzes are scheduled for October 15 and November 19 from 11:00-11:20 am. Please arrive on time with your local machine, as late assignments will not be accepted for in-class quizzes except in cases of emergency. For take-home quizzes, a late assignment penalty of 10% of the total score will be applied for each day past the due date. If you know in advance that you will miss any of the scheduled quizzes, you must make arrangements with the instructor at least one month ahead of time.

*Midterm:* The instructor will place you into a group of 3-4 students. Your group will give a presentation about one topic we've discussed in the course. The instructor will provide a list of presentation topics, and each group needs to choose one topic. The sign-up sheet link will be available on Canvas Announcement, starting from Wednesday, October 23 at 8 am. The midterm presentation is a 7-min oral presentation about the chosen topic, and it should consist of explaining one or two concepts and demonstrating them with a few examples. You will also have to make three questions about the chosen topic for the quiz in the midterm week. You will use Jupyter Notebooks or PowerPoint slides for the presentation and all the group members will present together on the presentation date. Be sure to practice beforehand, and time yourselves before you give the midterm presentation.

*Final Project:* You will work with other students in the same group as for the midterm presentation. Pick a dataset that you and your group find interesting. Example sources are found below. Feel free to select your data from any other source as appropriate.

The final project should form a research question, and perform data pre-processing, data cleaning, outlier removal, and so on to sanitize your data as necessary. Explore your data to reveal interesting/useful information based on your project scenario, and create at least 2 visualizations that you find interesting/useful. Also, do at least one of the following, depending in your interests and background: (i) compute meaningful statistical quantities (e.g., means, correlations), (ii) perform a statistical test on the data (e.g., t-test), or (iii) fit a model to the data (e.g., regression).

*The final report* should cover the following sections: abstract, introduction, data, data processing methodology, results, and conclusions. Also, you should submit your *Python codes*, and make detailed annotations on the codes so that peers can easily reproduce your work. The files can be in Jupyter Notebooks or Python scripts. The maximum number of pages is limited to 10 pages (double spaced; excluding the appendix). The paper should be written as coherently as possible. More details about the final project will be announced on Canvas and GitHub.

**Data**

For your final project, you will analyze real data and draw meaningful conclusions with regard to your research questions. Here is a list of websites where you can find interesting data.

- kaggle
- AWS Open Data
- data.world
- ICPSR
- The Google Dataset Search
- The UCIML Repo
- The CMU data repository
- The datasets subreddit
- Tycho
- Data Portals

### Delivery Mode Expectations

Students complete assigned reading before live sessions.
In-person live sessions will consist of:

- the instructor gives code demos
- students work on small and larger coding assignments, with assistance from instructor/CAs/potentially their peers
- the instructor reviews coding solutions with the class
- students submit assignments through Canvas

Note that this course is conducted in-person. In-person lectures will not be recorded using Zoom or any recording tools. However, in the event of emergencies such as COVID-19 or natural disasters, in-person lectures may be recorded via Zoom.

### Electronic Submission of Assignments

All assignments must be submitted electronically through Canvas by the specified due dates and times. It is important to complete all assigned work—failure to do so will likely result in failing the class.

### Class Management

Email / Communication

- Email is the best way to get in touch with the teaching staff: professor and CAs.
- Please be sure to include the course number ("HUDM5001") in your email subject line when sending email to any of the teaching staff.

### Grading

Courses at Teachers College use the following grading system: A+, A, A-; B+, B, B-; C+, …, F. The symbol W is used when a student officially drops a course before its completion or if the student withdraws from an academic program of the University.

| | Requirement | weight for final grade |
|---|---|---|
| 1. | Programming assignments (drop one) | 35% |
| 2. | Quizzes | 40% |
| 3. | Midterm | 10% |
| 4. | Final Project | 15% |

| If your weighted total points are… | Your final letter grade is… |
|---|---|
| [93, 100] | A |
| [90, 93) | A- |
| [83, 90) | B |
| [80, 83) | B- |

| | |
|---|---|
| [73, 80) | C |
| [70, 73) | C- |
| < 70 | F |

Note that A+, B+, and C+ will be determined by the class curve and overall performance in the course.

**AI/ChatGPT**

Intellectual honesty is vital for an academic community and for the fair evaluation of your work by teaching staff. All work submitted in this course must be your own or that of your group, completed in accordance with the University's academic policies. You should not engage in unauthorized collaboration or make use of ChatGPT or any other AI composition software to complete any of the course assignments.

**Services for Students with Disabilities**

The College will make reasonable accommodations for persons with documented disabilities. Students are encouraged to contact the Office of Access and Services for Individuals with Disabilities for information about registration (166 Thorndike Hall). Services are available only to students who are registered and submit appropriate documentation." As your instructor, I am happy to discuss specific needs with you as well.

**IN Incomplete**

The grade of Incomplete is to be assigned only when the course attendance requirement has been met but, for reasons satisfactory to the instructor, the granting of a final grade has been postponed because certain course assignments are outstanding. If the outstanding assignments are completed within one calendar year from the date of the close of term in which the grade of Incomplete was received and a final grade submitted, the final grade will be recorded on the permanent transcript, replacing the grade of Incomplete, with a transcript notation indicating the date that the grade of Incomplete was replaced by a final grade.

If the outstanding work is not completed within one calendar year from the date of the close of term in which the grade of Incomplete was received, the grade will remain as a permanent Incomplete on the transcript. In such instances, if the course is a required course or part of an approved program of study, students will be required to re-enroll in the course including repayment of all tuition and fee charges for the new registration and satisfactorily complete all course requirements. If the required course is not offered in subsequent terms, the student should speak with the faculty advisor or Program Coordinator about their options for fulfilling the degree requirement. Doctoral students with six or more credits with grades of Incomplete included on their program of study will not be allowed to sit for the certification exam.

**Email**

Teachers College students have the responsibility for activating the Columbia University Network ID (UNI) and a free TC Gmail account. As official communications from the College – e.g., information on graduation, announcements of closing due to severe storm, flu epidemic, transportation disruption, etc. – will be sent to the student's TC Gmail account, students are responsible for either reading email there, or, for utilizing the mail forwarding option to forward mail from their account to an email address which they will monitor.

**Religious Holidays**

It is the policy of Teachers College to respect its members' observance of their major religious holidays. Students should notify instructors at the beginning of the semester about their wishes to observe holidays on days when class sessions are scheduled. Where academic scheduling conflicts prove unavoidable, no student will be penalized for absence due to religious reasons, and alternative means will be sought for satisfying the academic requirements involved. If a suitable arrangement cannot be worked out between the student and the instructor, students and instructors should consult the appropriate department chair or director. If an additional appeal is needed, it may be taken to the Provost.

**Academic Integrity**

Students who intentionally submit work either not their own or without clear attribution to the original source, fabricate data or other information, engage in cheating, or misrepresentation of academic records may be subject to charges. Sanctions may include dismissal from the college for violation of the TC principles of academic and professional integrity fundamental to the purpose of the College.

**Sexual Harassment and Violence Reporting**

Teachers College is committed to maintaining a safe environment for students. Because of this commitment and because of federal and state regulations, we must advise you that if you tell any of your instructors about sexual harassment or gender-based misconduct involving a member of the campus community, your instructor is required to report this information to the Title IX Coordinator, Janice Robinson. She will treat this information as private, but will need to follow up with you and possibly look into the matter. The Ombuds officer for Gender-Based Misconduct is a confidential resource available for students, staff and faculty. "Gender-based misconduct" includes sexual assault, stalking, sexual harassment, dating violence, domestic violence, sexual exploitation, and gender-based harassment. For more information, see http://sexualrespect.columbia.edu/gender-based-misconduct-policy-students

**Emergency Plan**

TC is prepared for a wide range of emergencies. After declaring an emergency situation, the President/Provost will provide the community with critical information on procedures and available assistance. If travel to campus is not feasible, instructors will facilitate academic continuity through Canvas and other technologies, if possible.

1. It is the student's responsibility to ensure that they are set to receive email notifications from TC and communications from their instructor at their TC email address.
2. Within the first two sessions for the course, students are expected to review and be prepared to follow the instructions stated in the emergency plan.
3. The plan may consist of downloading or obtaining all available readings for the course or the instructor may provide other instructions.