CSc 10800: Foundations of Data Science

Department of Computer Science, The City College of New York, CUNY

Course Number: 21866
Class Location: NAC 5/109 (North Academic Center)
Meeting Days: Tues / Thurs @ 5:00 - 6:15 pm
Credits/Hours: 3 credits / 3 hours

Instructor: Di Yoong
Office: NAC 8/202D (North Academic Center)
Office Hours: 4:00 - 5:00 pm or by appointment
Email:

Course Description

This course introduces the fundamental concepts and computational techniques of data science to all students, including those majoring in the Arts, Humanities, and Social Sciences. Students engage with data arising from real-world phenomena—including literary corpora, spatial datasets, and social networks data—to learn analytical skills such as inferential thinking and computational thinking. The competencies learned in this course will provide students with skills that will be of use in their professional careers, as well as tools to better understand, quantitatively and qualitatively, the social world around them. Finally, by teaching critical concepts and skills in computer programming and statistical inference, the class prepares students for further coursework in technology-dependent subjects, such as Digital Humanities. The course is designed for students who are new to statistics and programming. Students will make use of the Python programming language, but no computer science pre-requisites are required.

This course does **not** satisfy degree requirements for Computer Science students. Computer Science students should **not** be enrolled in this course.

Learning Goals

At the end of the course, you will:

- Gain foundational python skills, including creating functions and writing loops
- Explore different data science methods, including linear regression and text analysis
- Practice communicating and translating data and analysis for broad audience
- Critically consider data and projects in context

Course Expectations

Classmate's Contact Information

Name	Email	Phone Number

Classroom Policies

Respect and accountability are crucial to productive class discussions. As co-producers of knowledge, I am expecting that we will practice respect for each other and be accountable to our words and actions. The classroom space is a learning space that can be, at times, uncomfortable, especially as we speak through our different perspectives and experiences. As long as we strive to be respectful to each other and accountable to the opinions, comments, questions, and concerns we share, this learning space will become a great place for us to nudge our boundaries. The college also has a formal policy for the code of conduct that is shared below.

Student Code of Conduct. All student members of the College community are expected to conduct themselves in a manner that demonstrates mutual respect for the rights and personal/academic well-being of others, preserves the integrity of the social and academic environment, and supports the mission of the College. The College has an inherent right to address behavior that impedes, obstructs, or threatens the maintenance of order and attainment of the aforementioned goals by violating the standards of conduct set forth in the University student conduct policies noted below as well as other policies that may established by the respective Schools, Global Sites, and administrative offices of the University.

The goals of the CCNY Community Standards are:

- To promote a campus environment that supports the overall educational mission of the University
- To protect the University community from disruption and harm
- To encourage appropriate standards of individual and group behavior
- To foster ethical values and civic virtues
- To foster personal learning and growth while at the same time holding individuals and groups accountable to the standards of expectations established by the Code of Conduct (Article XV)

Email and communications. *Please include CSc 10800 and an appropriate subject line (e.g. CSc 10800: Appointment with you) in your email.* I check my email once a day during the weekdays. Emails sent over the holidays and weekends will be read on the next working day. I

usually respond within 48 hours after receiving your email. If I did not respond after 48 hours, please send a follow-up message.

Attendance and Grading Policies

Attendance and participation is required. Learning a new programming language requires consistent practice and your understanding of the material will be greatly facilitated by your participation in class. Students are expected to come to class prepared, which includes completing the assigned reading before class and being ready to engage in class discussions.

Absences. If you are unable to attend class, please email me in advance. If you are unable to email me in advance, please let me know as soon as it is possible. I do not require proof and would just need to know if you are not going to be in class. You may miss up to 3 classes. Missing more than 3 classes may impact your grades in class.

Technology. In-class lessons and homeworks are done in Jupyter notebooks. The notebooks assume a Python 3 installation with the standard modules from an Anaconda installation such as NLTK, Pandas, Numpy and Matplotlib.

Grading policy. While learning a new programming language, it is inevitable that we will end up with mistakes and "fail" to obtain the desired output. In addition, there are also several ways to reach the desired output. For activities and assignments related to programming, grades are awarded on process and effort rather than accuracy. For most other activities and assignments you will have multiple opportunities to resubmit.

Late Work Policy: Extensions may be offered on a case-by-case basis. If you require an extension you must reach out to me at least 1 class session before the due date unless it is an emergency. Late submissions may be penalized.

Plagiarism and academic integrity. Plagiarism is copying and using other people's words without proper acknowledgment or citation as it is indicated in the CUNY Policy on Academic Integrity. All writing submitted for this course is understood to be your original work written. Plagiarism is unacceptable and has serious consequences that can include a failing grade. In cases where I detect academic dishonesty (the fraudulent submission of another's work, in whole or part, as your own), you may be subject to a failing grade for the project or the course, and in the worst case, to academic probation or expulsion. You are expected to read, understand, and adhere to CCNY's Policy on Academic Integrity.

Using generative AI applications such as ChatGPT and Bard in your assignment may be flagged as an academic integrity issue. If you are interested in using such applications for your assignment, they should be listed as a co-author and you will also need to specify how you used them in the assignment. Usage of such softwares can be useful to check for errors but are rarely useful for producing entire assignments.

CCNY Resources

The **AccessAbility Center/Student Disability Services** ensures equal access and full participation to The City College of New York's programs, services, and activities by coordinating and implementing appropriate accommodations. If you are a student with a disability who requires accommodations and services, please visit the office in NAC 1/218, or contact AAC/SDS via email (disabilityservices@ccny.cuny.edu), or phone (212-650-5913 or TTY/TTD 212-650-8441).

The **CCNY Service Desk** is IT's point of contact for students who need help with services such as Blackboard, CUNYfirst, and Citymail.

Laptop Loaner Program. The City College Office of Information Technology provides a laptop loaner program for current CCNY students. The program is funded by the CCNY Student Technology Fee. The laptops are internally equipped with WiFi for use where wireless access exists. Wireless networking is available throughout much of the campus. All laptops are loaded with MS Office, Adobe Acrobat, as well as other CCNY-approved software. This program is designed for experienced computer users who are able to use the installed applications. Please note: you must be logged into WiFi for some software to be fully operational.

The **Counseling Center** provides free and confidential services to all undergraduate and graduate students who are currently enrolled at City College. Services provided include screening and assessment, crisis intervention, individual short-term counseling, group counseling, referral and case management, and workshops.

The Psychological Center is a community-based sliding fee scale mental health clinic located in the North Academic Center at the City College of New York. They are open to the college and the community at large, and provide children, adolescents and adults with psychological treatment in the following modalities: individual psychotherapy, group psychotherapy, family and couple psychotherapy. Additionally, they conduct psychological evaluations as well as psychological/neuropsychological assessments. At this time, The Psychological Center provides short- and longer-term empirically supported treatments which include: psychodynamic psychotherapy, Transference-Focused Psychotherapy, Dialectical Behavior Therapy, Emotion-Focused Therapy and Motivational Interviewing.

The **Emergency Grants Program** provides assistance to students in good academic standing who are facing unforeseen events, resulting in a financial emergency that jeopardizes their ability to persist at City College. The goal of the fund is to help students remain in school without interruption so they successfully complete their degrees.

Grading and Assignments Overview

Grading

Participation and Content Check-Ins	20%
Activity 1: Data types, variables, and string methods	10%
Activity 2: Foundations of python	15%
Activity 3: Preparing the *Trans-Atlantic Slave Trade of Americas* dataset	10%
Activity 4: Exploring the *Trans-Atlantic Slave Trade of Americas* dataset	15%
Activity 5: Term Frequency–Inverse Document Frequency	15%
Semester Reflection	15%

Assignment prompts will be made available on Blackboard and are due by class session.

Grading Scale (%)

97 - 100	A+	77 - 79.9	C+
94 - 96.9	A	74 - 76.9	C
90 - 93.9	A-	70 - 73.9	C-
87 - 89.9	B+	67 - 69.9	D+
84 - 88.9	B	60 - 66.9	D
80 - 83.9	B-	0 - 59.9	F

Course Schedule and Materials

All materials (e.g. readings and datasets) in this course are provided to you on Blackboard. Please refer to the course schedule for the due dates. Materials are due before class.

While I do not anticipate major changes to the syllabus, please be aware that the schedule may change at the discretion of the instructor. Changes will be announced on Blackboard.

	Date	Topics	Readings	Technical Readings	Due by Class
1	25/01 (Thu)	Introductions			
	30/01 (Tue)	What is data science?	Data 8: What is data science (Chapter 1.1 and 1.2) Ted Underwood: Seven ways humanists are using computers to understand text	DHRI: Installing GitBash	
Command Line and Python Basics					
2	01/02 (Thu)	Command Line		Melanie Walsh: The Command Line	Download GitBash for Windows user
	06/02 (Tue)	Whose data science?	Data Feminism: Introduction: Why data science needs feminism Feminist data manifest-no	DHRI: Installing Python (and Anaconda)	
3	08/02 (Thu)	Stages of data, Jupyter notebook and Interacting with Python	DHRI: Data literacies 01 - 03	Melanie Walsh: How to use Jupyter notebooks; Anatomy of a python script DHRI: Interacting with Python	Download Anaconda
	13/02 (Tue)	Data types, Comparisons, and Variables		Melanie Walsh: Python data types; Variables Data 8: Comparisons	
4	15/02 (Thu)	String method and Debugging	Aditya Mukerjee: I Can Text You A Pile of Poo, But I Can't Write My Name	Melanie Walsh: String methods; Common python error	
	20/02 (Tue)	Lists and For loops		Melanie Walsh: Python Lists and Loops	
5	22/02 (Thu)	**No Class. CUNY Monday.**			
	27/02 (Tue)	Lists, For loops II, and Conditionals		DHRI: Conditionals Melanie Walsh: Python Lists and Loops (Cont.)	Activity 1
6	29/02 (Thu)	Functions		Melanie Walsh: Functions	

	05/03 (Tue)	Foundations of Python review			
Data Analysis with Pandas					
7	07/03 (Thu)	Data in context; Introducing Pandas - Getting started with data analysis	Heather Krause: Data biographies - Getting to know your data Data Feminism: The numbers don't speak for themselves	Melanie Walsh: Panda Basics -- Part 1 (Dataset to Calculate Summary Stastics) Melanie Walsh: Missing data	
	12/03 (Tue)	Pandas: Data cleaning		DRI: Basic data cleaning; Rename, select, drop, and add new columns	Activity 2
8	14/03 (Thu)	Pandas: Data cleaning II		DRI: Sort columns, Groupby Columns, & Count values; Basic data visualizations; Write to csv	
	19/03 (Tue)	Preparing *Trans Atlantic Slave Trade of Americas* dataset	Jamelle Bouie: We Still Can't See American Slavery for What It Was		
Statistics in the Humanities					
9	21/03 (Thu)	Statistics in the humanities	John Canning: Statistics for humanities, pp. 7-22	W3 Schools: Mean, Median, Mode	
	26/03 (Tue)	Understanding the spread and sampling	John Canning: Statistics for humanities, pp. 23-36		Activity 3
10	28/03 (Thu)	Testing hypotheses	Data 8: Assessing a model Data 8: Multiple categories (Read untill A New Statistic: The Distance between Two Distributions)		
	02/04 (Tue)	Causality and experiments	Data 8: Causality and experiments		
11	04/04 (Thu)	Understanding relationships	John Canning: Statistics for humanities, pp. 75-85	The PyCoach: A Simple Guide to Linear Regression using Python	
	09/04 (Tue)	Exploring *Trans Atlantic Slave Trade of Americas* dataset	Jessica M. Johnson: Markup bodies - Black [life] studies and slavery [death] studies at the digital crossroads, pp. 57 - 65		
Text Analysis					
12	11/04 (Thu)	Text analysis: Beginning with NLTK	Lauren Klein: Distant Reading After Moretti	DHRI: Text analysis (Sections 01 to 08)	

	16/04 (Tue)	Text analysis: Data cleaning		DHRI: Text analysis (Sections 09 to 12)	Activity 4
13	18/04 (Thu)	Term-Frequency Inverse Document Frequency	Rohit Maden: TF-IDF/Term Frequency Technique	Melanie Walsh: TF-IDF with Scikit-Learn	
	25/04 (Thu)	**No Class. Spring break.**			
14	30/04 (Tue)	**No Class. Spring break.**			
	*02/05 (Thu)	Term-Frequency Inverse Document Frequency II		Melanie Walsh: TF-IDF with Scikit-Learn	
15	*07/05 (Tue)	Sentiment Analysis		Melanie Walsh: Sentiment Analysis	
		Data Storytelling			
	*09/05 (Thu)	Data storytelling	Jackie Mansky: W.E.B. Du Bois' visionary infographics come together for the first time in full color Ben Schmidt: Gendered language in teacher reviews, FAQ, Article in Chronicle Hannah Anderson and Matt Daniels: Film dialogue from 2,000 sceenplays, broken down by gender and age		Activity 5
16	*14/05 (Tue)	Buffer day			
	*21/05 (Thu)	**Finals week.**			Semester Reflection

*Class sessions in May will be conducted online. Further information will be made available closer to the date on Blackboard.

This course was developed with the following resources:

- Melanie Walsh, *Introduction to Cultural Analytics & Python*, Version 1 (2021), https://melaniewalsh.github.io/Intro-Cultural-Analytics/welcome.html,
- John Canning, *Statistics for the Humanities*, (2014), http://statisticsforhumanities.net/book/,
- Digital Humanities Research Institute, *DHRI-Curriculum*, https://github.com/DHRI-Curriculum