CS 210: Data Management for Data Science

Jeff Ames (jeff.ames@rutgers.edu)

Fall 2022

Prerequisite: CS 111 (Intro to Computer Science) or CS 143 (Data Literacy)

Course details, including a list of teaching assistants (TAs), office hours, and recitations, will be posted on Canvas. There will be no office hours or recitations the first week.

Overview

This course is designed to provide students with the knowledge and skills needed to acquire and curate real word data, to explore the data to discover patterns and distributions, and to manage large datasets with databases.

Students will learn the minimal aspects of Python as needed to acquire and curate datasets, using Python libraries to get data from various online data sources online, detect which aspects of data are uncurated or unreliable and understand why it is so, learn various domain independent and domain dependent ways to curate the data, and get the curated data into a form that can be explored, managed and analyzed. Students will also learn how to get datasets into database-ready form and do basic analysis of such datasets using relational databases and SQL.

Textbook

There is no required textbook for this course.

Grading

Grades will be weighted as follows:

Homework	55%
Exams	45% (15% each)

Any regrading request must be raised within one week of grades being returned, after which they are considered final.

Exam dates

- Exam 1: 10/25
- Exam 2: 11/15
- Exam 3: 12/13

Homework

There will be about 4 homework assignments. You can resubmit homeworks any number of times. Grading will be based on the last submission. You may submit homework up to 24 hours late with a penalty of 1 point per hour.

Homework must be submitted on Canvas; emailed submissions are not accepted. You are responsible for ensuring the submitted files are correct.

Tentative schedule of topics

Week	Topic
1	Data management overview, setting up Python
2	Getting started with basic Python elements.
3	Building simple program logic with Python
4	Storing and managing data with lists and dictionaries
5	Curating and extracting data with Python
6	Regular expressions, file handling in Python
7	Working with CSV and JSON data formats
8	Managing numeric data with Numpy
9	Storing and analyzing data with Pandas
10	Data curating and management with Numpy and Pandas
11	Visualizing and exploring data with matplotlib
12	Creating relational databases
13	Relational data management with SQL
14	Relational data management with SQL

Academic integrity

Rutgers University takes academic dishonesty very seriously. By enrolling in this course, you assume responsibility for familiarizing yourself with the Academic Integrity Policy and the possible penalties (including suspension and expulsion) for violating the policy. As per the policy, all suspected violations will be reported to the Office of Student Conduct. Please review the Academic Integrity Policy at: https://nbacademicintegrity.rutgers.edu/.

Accomodations

Rutgers University welcomes students with disabilities into all of the University's educational programs. In order to receive consideration for reasonable accommodations, a student with a disability must contact the appropriate disability services office at the campus where you are officially enrolled, participate in an intake interview, and provide documentation: https://ods.rutgers.edu/students/documentation-guidelines.

If the documentation supports your request for reasonable accommodations, your campus's disability services office will provide you with a Letter of Accommodations. Please share this letter with your instructors and discuss the accommodations with them as early in your courses as possible.

To begin this process, please complete the registration form (https://webapps.rutgers.edu/student-ods/forms/registration).