SI 618 - Data Manipulation & Analysis

Fall 2023: Thursday, 1:00-4:00pm AUD A

Note: Some syllabus details may be subject to change.

Version 2023.1.17.1.CT

Instructor: Chris Teplovs (he/him)
Email: cteplovs@umich.edu (see section labeled “Communication” below)
Office hours: Thursdays, 10:00-11:00 am, UMSI Loft, 715 N. University Ave. (above Hola Seoul): NOTE: my office hours start in Week 2 (i.e. no office hours on Thursday, January 5, 2023)

GSIs: Arjun Padmanabhan
Niloufar Sedarati
Wengran Xiao

Office Hours:

Arjun Ananda Padmanabhan: Tuesday, 1:00 - 2:00 pm , Room 1277 Northquad (School of Information)

Niloufar Sedarati : Monday, 8:00 am to 9:00 am , Room 1270 NorthQuad

Wengran Xiao : Fridays 4pm-5pm NQ 1278

Instructional Aides:

TDB

Communication

The best way to contact any member of the teaching team for clarification about course content is by using Slack via Canvas.

We try to answer questions that are sent via Slack within about 48 hours. Responses on weekends and holidays may be slower. Your GSIs should be your first choice for technical questions; conceptual questions are best directed to Dr. Teplovs.

Personal matters should be communicated via email to Dr. Teplovs. Please include “[SI 618]” in your subject line to receive a timely response.

Course Description

This course aims to help students get started with their own data harvesting, processing, aggregation, and analysis. Data analysis is crucial to evaluating and designing solutions and applications, as well as understanding user's information needs and use. In many cases the data we need to access is distributed online among many web pages, stored in a database, or available in a large text file. Often these data (e.g. web server logs) are too large to obtain and/or process manually. Instead, we need an automated way of gathering the data, parsing it, and summarizing it, before we can do more advanced analysis. Therefore, students will learn to use Python and its modules to

accomplish these tasks in a 'quick and easy' yet useful and repeatable way. Next, students will learn techniques of exploratory data analysis, using scripting, text parsing, structured query language, regular expressions, graphing, and clustering methods to explore data. Students will be able to make sense of and see patterns in otherwise intractable quantities of data. The skills students will learn include the following: Big data processing; Converting messy data into a form that can be analyzed using Pandas; Compute and visualize summary statistics of datasets; Master the specification of graphical displays using Seaborn and matplotlib; Combine the use of graphics with data manipulation to visualize relationships between variables; Use machine learning techniques including clustering and classification. Use dimension reduction techniques.

Textbooks

All required materials are available via O'Reilly Safari Books Online, available via https://www.lib.umich.edu/announcements/oreilly-safari-books-online

We have created a "playlist" of all the readings in the course:

https://learning.oreilly.com/playlists/2f86a3dc-9073-40ae-86e3-461bdd997021

Schedule

The course will combine pre-recorded lectures with live hands-on coding sessions. The general idea is that you will learn the concepts and techniques and then practice them by writing code in class. Much of the pre-recorded material is shared with the undergraduate equivalents of this course, so you may encounter references to other course numbers (in particular, SI 370). In addition, there will be regular programming and analysis assignments to be done as homework. You will use Python for all of the in-class work and homework assignments.

Learning Outcomes

At the end of the course, students should be able to:

- C: use python (via Jupyter) for data analysis
- C: load and manipulate data in a variety of formats (CSV, JSON, unstructured text, results of SQL queries)
- C: filter, sort, select columns, etc.
- L: create visualizations using matplotlib and seaborn
- L: perform parallel and distributed analysis (dask, spark)
- ~~L: extract data using SQL~~

- L: construct a machine learning pipeline using scikit-learn
- L: use a scikit-learn classifier
- ~~A: describe networks and basic network analysis~~
- ~~A~~L: describe common techniques used in natural language processing

Course Outline

Note 1: Some syllabus details and timing may be subject to change.
Note 2: NB refers to the in-class notebook
Note 3: Please see "playlist" above for links to readings

Date	**Topic**	**Pre-class preparation**	**What's due before class (note: check due dates in Canvas)**	**What's due within 3h of the end of class**
8/31	Course Introduction, Jupyter intro, python review			NB 1
9/7	Data manipulation I: pandas DataFrames	NOTE: HEAVY READINGS THIS WEEK! McKinney: Chapter 4 Chapter 5 Chapter 6 *Recommended (not required):* McKinney: Chapters 1,2,3 *Also recommended (not required):* Chen: Chapter 1 Chapter 2 Chapter 3		NB 2

9/14	Data manipulation II: pandas	McKinney 7, 8, 10 Pivot tables	Homework 1 (Pandas data manipulation)	NB 3
9/21	Data analysis I: univariate stats, visualization, seaborn, intro to correlation	McKinney section 9, Downey chapter 7,	Homework 2 (more data manipulation)	NB 4
9/28	Data analysis II: ANOVA, t-test, linear models	Statistics in Python Gedeck, Bruce, & Bruce, chapters on ANOVA, t-test, linear models	Homework 3 (Visualization)	NB 5
10/5	Categorical data (contingency tables, chi-square, mosaic plots) Introduction to text processing (regular expressions)	Chi-square Mosaic plots Mastering Python Regular Expressions	~~Homework 4 (Linear models)~~	NB 6
10/12	Natural language processing (nltk, spaCy, gensim)	Speech and Language Processing-Ch1 Grus, chapter 21 (optional, extra reading)	Homework 4 (Linear models)	NB 7
10/19	Machine Learning: Intro & Regression	VanderPlas chapter 5 Géron chapter 1 Géron chapter 2 Grus, chapter 11 intro to machine learning scikit-learn	Homework 5 (NLP) Project proposal	NB 8

10/26	Machine Learning: Dimension Reduction	Géron chapter 8 Géron chapter 9 Grus chapter 20 (optional, extra reading) https://medium.com/@luckylwk/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b		NB 9
11/2	Machine Learning: Clustering		~~Project check in~~	NB 10
11/9	Machine Learning: Classification	Grus chapter 13 Géron chapter 3, 4, 6, 7 https://blog.datarobot.com/classification-with-scikit-learn	Homework 6 (Clustering)	NB 11
11/16	Big Data (Dask + Spark)	Daniel chapters 1 & 2 (https://learning.oreilly.com/library/view/data-science-with/9781617295607/), also added to class playlist in O'Reilly; Karau et al.	Homework 7 (Classification)	NB 12
11/23	**Break - No classes**			
11/30	Big Data (Spark)			NB 13

12/7	Platforms for Data Science Closing thoughts, jobs, etc.	(none)	Homework 8 (Big Data) Project Final Report + Code	

Quizzes

There will be quizzes available to you almost every week. The quizzes will be based on material in the pre-recorded lectures and the readings. I use readings to help you learn the material: you can typically keep trying them until you get 100%, so there's no reason to not get full points for them!

In-class Notebooks

Our synchronous meetings are centered around co-constructed Jupyter notebooks. I write these notebooks to support learning by providing hands-on exercises that allow you to practice applying the techniques we are learning. At the end of our synchronous meetings you will be expected to turn in your notebook with the completed exercises. **These notebooks will be due 3 hours after the end of each class; in almost all cases you will submit these notebooks at the end of our synchronous time together.** These notebooks are not meant to be homework assignments (see next section). Whereas you will be working in groups during the hands-on segments, you will be required to submit your own notebooks for credit.

Homework Assignments

Homework assignments typically will be released on Thursday before class and due the following Thursday before class. Please see the late policy section for what happens if you can't make this deadline. As much as possible, homework assignments will be based on real-world datasets and focused on realistic problems.

Please make sure you read the section below on Giving/Receiving Assistance. If you copy someone else's homework solution completely, or almost completely (and/or fail to acknowledge your source), then this will be considered cheating, and I'll refer your case to the academic advising office for disciplinary action. I have long experience and a particular talent for catching people, so just don't do it.

Please contact the instructor if you have any uncertainty or questions about this policy.

Final Project

The goal of the final project is to further apply what you've learned in class to real-world datasets. The deliverables for the final project will include an initial proposal, final report, and code. Whereas the proposal contributes only a tiny proportion to this component of the grade, **you cannot submit a final project without having completed the proposal**. Details on the final project will be available as we get closer to the project proposal date.

Class Format

This class consists of two different "modes" of instruction: asynchronous and synchronous. Asynchronous instruction is work that you do on your own time. Asynchronous activities include doing assigned readings, watching pre-recorded video lectures, completing homework assignments; and working on your projects. Synchronous activities occur during regularly scheduled class time and include live coding, question-and-answer sessions, reporting on project progress, and time dedicated to getting started with homework assignments.

Attendance

Attendance and participation during synchronous sessions is mandatory. Repeatedly missing class or failing to participate will likely lead to a failing grade.

Pre-recorded Lecture Videos

You will have access to pre-recorded video lectures approximately 5 days prior to the topics being covered in class. Most of these videos will have associated quizzes, which provide both an opportunity to check your understanding as well as contribute points to your overall score in the course.

Readings

There are readings assigned almost every week. These are intended to supplement the face-to-face classes and you will get much more out of class if you read these before you attend. To encourage you to complete the readings before class, there will be quizzes based on the readings. You can accumulate points toward your overall score in the course by completing these quizzes.

Giving and Receiving Assistance

I'm going to state a policy used in the other SI programming courses (the following text is taken in modified form from the SI370 syllabus). Learning technical material is often challenging, and a course like this one covers a range of topics and can move quickly. I want you to succeed in the course and I encourage you to get help from anyone you like.

However: In the end, **you** are responsible for learning the material – so you need to make sure that the help you get is focused on gaining knowledge, not just on getting through the assignments. If you rely too much on help so that you fail to master the material, especially the basics earlier in the semester, you will crash and burn later in the course.

The final submission of each homework exercise must be in your own words. If you get help on an assignment, please indicate the nature and amount of help you received. If the assignment involves computer code, add a comment indicating who helped you and how. Any excerpts from the work of others must be clearly identified as such (e.g. quotation with citation, or with comments in the code if it is a code fragment you have borrowed).

If you're a more advanced student and are willing to help other students, please feel free to do so. Just remember that your goal is to help teach the material to the student receiving the help. It is acceptable for this class to ask for and provide help on an assignment via the class Q&A platform, including posting short code fragments (e.g. 4-5 lines). Just don't post complete answers. If it seems like you've posted too much, one of the instructional staff will contact you to let you know, so don't worry about it. When in doubt, err on the side of helping your fellow students. To reiterate, the collaboration policy is as follows. Collaboration in the class is encouraged for assignments – you can get help from anyone as long as it is clearly acknowledged. Use of solutions from previous semesters is not allowed. The authorship of any assignments must be in your own style.

Grading

This course uses a points system to determine your final grade. Point distribution for the different components is as follows:

Assignment	Overall Weight	Number	Points each	Total
Quizzes[1]	10%	variable	variable	200
In-class notebooks	25%	13	variable	500

Homework	40%	8	100	800
Project ——————————— Proposal ~~Check-in~~ Final report Code	25% ——— 5% ~~3%~~ 10% 10%			500
Total points available				2000
[1] Quizzes are not "optional" – they are a great source of easy points!				

Conversion from points to letter grades will use the following mapping:

A+	1950
A	1900
A-	1850
B+	1800
B	1750
B-	1700
C+	1650
C	1600
C-	1550
D+	1500
D	1450
D-	1400

So, to get an A you will need to accumulate 1900 points. **DO NOT TRUST THE PERCENTAGE (%) SCORES IN CANVAS!** Always compare your earned points to the table above to determine your grade (or your projected grade).

Late Policy

I realize that the occasional crisis might mess up your schedule enough to require a bit of extra time in completing a course assignment. Thus, I have instituted the following late policy that gives you a limited number of flexible “late day” credits.

You have **three (3)** free late days to use during SI 618. One late day equals exactly one 24-hour period after the due date of the assignment (including weekends). No fractional late days: they are all or nothing. Once you have used up your late days, **25% penalty** for each subsequent 24h period after the deadline that an assignment is late. For example, if the due date is 1pm Thursday, with no late days left, penalties would be:

Before 1pm Friday:: 25% deduction

Before 1pm Saturday: 50% deduction

Before 1pm Sunday: 75% deduction

After 1pm Monday: 100% deduction

You don't need to explain or get permission to use late days, and we will track them for you. In cases where late days can be assigned in multiple ways (e.g. you have only one late day left but hand in two late assignments) we will always allocate late days in a way that maximizes your grade. Note that resubmissions after the deadline will be counted as late submissions. **Also, late days may not be applied to the project final report.**

If you are submitting your work late and you believe you have a ***valid*** excuse to not use your free late days, please complete the SI 618 WN 23 Late Form. Completion of the form does not guarantee that the late penalty will be waived.

Classroom policy

Students are asked to attend class on time and remain through the entire class. Please mute your IMs during class – it can be embarrassing if a member of the teaching team is helping you and you get a very personal IM.

Original Work

Unless explicitly specified, all submitted work must be your own, original work. You may discuss general approaches with others on individual assignments, but you should work on the code by yourself. It is a violation of the original work policy to copy code or other work wholesale. If you did work closely with any other students that helped you with any assignment, you must indicate on your turned-in assignment who you worked with, and how. Any excerpts from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the School's policy on Academic and Professional Integrity will result in severe penalties, which might range from failing an assignment, to failing a course, to being expelled from the program, at the discretion of the instructor and the Senior Associate Dean for Academic Affairs.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. I will treat any information you provide as private and confidential.

Student Mental Health and Wellbeing

The University of Michigan is committed to advancing the mental health and wellbeing of its students, while acknowledging that a variety of issues, such as strained relationships, increased anxiety, alcohol/drug problems, and depression, directly impacts students' academic performance.

If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (732) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources.

For a more comprehensive listing of the broad range of mental health services available on campus, please visit: http://umich.edu/~mhealth/