SI 618 - Data Manipulation & Analysis

Winter 2021: Mondays, 1:00-4:00 via Zoom

Note: Some syllabus details may be subject to change.

Version 2021.01.20.1.CT

Instructor: Chris Teplovs (he/him)
Email: cteplows@umich.edu (see section labelled “Communication” below)
Office hours: TBD

GSIs:
Yijing Chen (she/her)
Email: yijingch@umich.edu
Office hour: 9am-10am on Fridays

Qiaoning Zhang (she/her)
Email: qiaoning@umich.edu
Office hour: 4pm-5pm on Tuesdays

Communication

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

We try to answer questions that are sent via Piazza 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. In addition, there will be regular programming and analysis assignments to be done as homework. You will use Python for most of the in-class work and homework assignments.
# 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

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Pre-class preparation</th>
<th>What’s due before class</th>
<th>What’s due within 24h of the end of class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/25</td>
<td>Course Introduction, Jupyter intro</td>
<td></td>
<td>NB 1</td>
<td></td>
</tr>
</tbody>
</table>
| 2/1   | Data manipulation I: pandas DataFrames                                | NOTE: HEAVY READINGS THIS WEEK!  
McKinney:  
Chapter 4  
Chapter 5  
Chapter 6  

 Recommended (not required):  
Chen:  
Chapter 1  
Chapter 2  
Chapter 3 |                       |                       |                         |                                          |
| 2/8   | Data manipulation II: pandas                                          | McKinney 7, 8, 10  
Pivot tables  
Homework 1 (Pandas data manipulation: Olympics) |                       | NB 3                      |
| 2/15  | Data analysis I: univariate stats, visualization, seaborn, intro to correlation | McKinney section 9,  
Downey chapter 7, Statistics is easy  
Homework 2 (more data manipulation) |                       |                           |
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Book/Website</th>
<th>Homework/Project</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/22</td>
<td>Data analysis II: ANOVA, t-test, linear models</td>
<td>Statistics in Python Gedeck, Bruce, &amp; Bruce, Chapters XXXXXXX</td>
<td>Homework 3 (Visualization and correlation)</td>
<td>5</td>
</tr>
<tr>
<td>3/1</td>
<td>Categorical data (contingency tables, chi-square, mosaic plots)</td>
<td>Chi-square</td>
<td>Homework 4 (Linear models)</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Text processing: Regular Expressions</td>
<td>Mosaic plots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/8</td>
<td>Natural language processing (nltk, gensim)</td>
<td>Speech and Language Processing-Ch1</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>3/15</td>
<td>Machine Learning: Clustering</td>
<td>intro to machine learning</td>
<td>Homework 5 (NLP)</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>scikit-learn clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/5</td>
<td>Spark I</td>
<td>Karau et al.</td>
<td>Homework 7 (Classification)</td>
<td>11</td>
</tr>
<tr>
<td>4/12</td>
<td>Spark II</td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>4/19</td>
<td>Platforms for Data Science</td>
<td></td>
<td>Homework 8 (Spark); Project report</td>
<td></td>
</tr>
</tbody>
</table>
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! Note, however, that they are all time-limited (i.e. they will disappear), so you need to keep up with them. Make sure you keep an eye on the due dates.

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 24 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 Tuesday before 11pm and due the following Monday 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.
**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 (i.e. Mondays from 1pm to 4pm) 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.

A note about scheduling conflicts and time zone impossibilities

Some of you have approached me asking if you can still take this course if you have a time conflict with another one or if you are currently residing in a timezone that makes attending the synchronous meetings extremely difficult (like 1:00am!). Whereas this is a remote course this semester due to COVID, we will meet via Zoom from 1:00pm to 4:00pm Ann Arbor time. As mentioned above, students will be expected to hand in their work from the synchronous sessions within 24h of the end of the class so you will need to commit to doing the work on your own later that same day. The challenge for you will be to complete the work without having the opportunity to interact with your peers and the teaching team.

My general sense is that missing the synchronous sessions will make the class more challenging than it has to be, but the decision is ultimately yours and I'll try to support you no matter what you choose to do.

**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 SI106 and SI370 (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 Slack, 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. There are a total of 1600 points available. Point distribution for the different components is as follows:

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Overall Weight</th>
<th>Number</th>
<th>Points each</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quizzes</td>
<td>20%</td>
<td>variable</td>
<td>variable</td>
<td>320</td>
</tr>
<tr>
<td>In-class notebooks</td>
<td>30%</td>
<td>12</td>
<td>40</td>
<td>480</td>
</tr>
<tr>
<td>Homework</td>
<td>50%</td>
<td>8</td>
<td>100</td>
<td>800</td>
</tr>
</tbody>
</table>

1600

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

<table>
<thead>
<tr>
<th>Grade</th>
<th>Points Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1525+</td>
</tr>
<tr>
<td>A-</td>
<td>1475-1525</td>
</tr>
<tr>
<td>B+</td>
<td>1400-1474</td>
</tr>
<tr>
<td>B+</td>
<td>1350-1399</td>
</tr>
<tr>
<td>B-</td>
<td>1300-1349</td>
</tr>
<tr>
<td>C+</td>
<td>1250-1299</td>
</tr>
<tr>
<td>C</td>
<td>1200-1249</td>
</tr>
<tr>
<td>C-</td>
<td>1150-1199</td>
</tr>
<tr>
<td>D+</td>
<td>1100-1149</td>
</tr>
<tr>
<td>D</td>
<td>1050-1099</td>
</tr>
<tr>
<td>D-</td>
<td>1000-1049</td>
</tr>
</tbody>
</table>

So, to get an A you will need to accumulate 1525 points. I do not award A+ grades in this class.

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 Monday, with no late days left, penalties would be:

Before 1pm Tuesday: 25% deduction
Before 1pm Wednesday: 50% deduction
Before 1pm Thursday: 75% deduction
After 1pm Thursday: 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 final project.

Classroom policy

Students are asked to attend class on time and remain through the entire class. It is difficult to teach to a bunch of named boxes, so you are encouraged to leave your video on if possible.

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/