SI 370 - Data Exploration

Fall 2020: TuTh 8:30-10:00am, via Zoom

Note: Some syllabus details may be subject to change.
Last updated: 8/27/2020, 3:00pm

Instructor: Chris Teplovs (cteplovs@umich.edu)
Office hours: Fridays, 1-2pm via Zoom

GSI: Mauli Pandey
Office hours: Mondays, 11am-1pm via Zoom

Instructional Aides: TBD

Communication

The best way to get your questions answered is by using Piazza, which is available via Canvas.

We try to answer questions that are sent via Piazza within about 48 hours, although our experience is that your peers are able to respond quicker. Responses on weekends and holidays may be slower.

If you have questions about course material, homework, or labs, please feel free to come and talk with us during Zoom-based office hours. Your GSI 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.

Quick Links

Communication
Course Description
Where this course fits in the curriculum
Learning Objectives

Competency:
Literacy:
Awareness:
Course Description

SI370 aims to help students get started with their own data acquisition and exploratory data analysis (EDA). EDA is crucial in evaluating and designing solutions and applications, as well as understanding information needs and use. Students in this course will learn basic concepts of information visualization and 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. Though the focus of the class is a high-level understanding of EDA much of the class will be using Python to implement solutions.
Where this course fits in the curriculum

This exploratory data analysis and visualization course aims to help students get started with their own data acquisition and exploratory analysis. Exploratory data analysis is crucial to evaluating and designing solutions and applications, as well as understanding information needs and use. Students in this course will learn basic concepts of information visualization and 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. In this course students will be able to work with the Pandas, seaborn, and scikit-learn packages of Python.

This course offers advanced material for those in the BSI beyond the 106/206 sequence. In SI330 you learned more about the data structures for analysis or how to obtain data. We will do our best to provide you with “clean” data so that manipulation is not strictly necessary (but in real world settings, both obtaining and manipulating as well as analysis skills are necessary). Both data manipulation and data analysis will be required for the group projects.

The focus of the class is on data exploration--how we can understand new datasets (with a focus on applications for people, by people, and about people) and ask high level questions. *Exploratory* Data Analysis is a critical piece of the “sensemaking” that goes into any analytical workflow: we often need to get a sense of what is going on with the data in order to better approach *Confirmatory* Data Analysis. That is, we want to understand what we have before we decide on the statistical analysis or other experiments that we might want to do. We apply ideas from statistics in this process, but often what we glean from our data will power the statistical analysis we do after.

Visualization plays a key role in this process as it provides a visual summary of the data in a way that we can understand it. Because of this, visualization will feature heavily in this class.

Though we largely focus on the exploration process of the analyst, where one or a small group is looking at the data, the tools you will learn in the class apply more broadly to the communication of information. For example, how you might visualize the data and convey it to others. You’ll find this useful in many of the other classes (as well as in “real-world” settings) where you have to present information or results to others.
Learning Objectives

**Competency:**

- Apply tools of EDA in new situations (specifically techniques/methods/workflows to: (a) maximize insight into a data set; (b) uncover underlying structure; (c) extract important variables; (d) detect outliers and anomalies; (e) test underlying assumptions; (f) develop parsimonious models; and (g) determine optimal factor settings. [http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm](http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm)
- Compute and visualize summary statistics of datasets (specific implementation in Python, but with general understanding of how to use other languages/systems)
- Combine the use of graphical aesthetics with data manipulation to visualize relationships between variables
- Use factors to analyze categorical data and exploratory clustering analysis for unlabeled data.
- Use classification techniques in unsupervised machine learning
- Produce polished information graphics for publication.

**Literacy:**

- Be familiar with basic concepts and design principles of information visualization
- Be familiar with basic analysis and visualization techniques for time series data, multidimensional data, network data, and text data

**Awareness:**

- Be aware of Python modules to process complex types of data such as networks, time series, and images.
- Be aware of advanced data visualization techniques.
Textbooks

There are several required books for this class; all are available via the O'Reilly Learning Portal.

- Joel Grus, *Data Science from Scratch* (2015)

All are available from for free via [https://www.lib.umich.edu/announcements/oreilly-safari-books-online](https://www.lib.umich.edu/announcements/oreilly-safari-books-online)

or


Schedule/Readings

Please see the [Schedule of Topics and Readings](https://www.lib.umich.edu/announcements/oreilly-safari-books-online).

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 team projects. Synchronous activities occur during regularly scheduled class time (i.e. Tuesdays and Thursdays from 8:30am to 9:50am) and include live coding, question-and-answer sessions, reporting on team project progress, and time dedicated to getting started with homework assignments.
The core part of the course consists of six modules. There is a homework assignment that is assigned at the end of each of those modules.

**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 contain embedded 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, most classes will start with a brief quiz based on the readings. You can accumulate points toward your overall score in the course by completing these quizzes.

**Giving and Receiving Assistance**

SI106 has this policy and I think it’s appropriate here as well (slightly modified from the SI106, W14 syllabus). Learning technical material is often challenging. We are going to cover a wide range of topics in the course and we will move quickly between topics. Because it is our goal for you to succeed in the course, we encourage you to get help from anyone you like.

All that said: You are responsible for learning the material, and you should make sure that all of the assistance you are getting is focused on gaining knowledge, not just on getting through the assignments. If you receive too much help and/or fail to master the material, you will crash and burn later in the semester.

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

If you are 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 Q&A site (Piazza), including posting (short!) code fragments. 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 allowed (and even encouraged) for assignments – you can get help from anyone as long as it is clearly acknowledged. Collaboration or outside help is not allowed on exams, though you will be allowed to use some materials that you bring with you. Use of solutions from previous semesters is not allowed. The authorship of any assignments must be in your own style.

Google/StackOverflow/etc.

Use these! Seriously… make an effort to discover the answers on your own by using the Web. You’ll be surprised how often you can find related code to help you. Give us the URLs of where you found help and even if you don’t totally solve the problem, if it looks like you were heading the right direction in finding the solution we’ll give you some credit!

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 most 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. 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 will be due no less than five days after they are released (see lateness policy below). You can expect that some of the material in the homework assignments will be covered in class so it’s in your interest to do as much of the class work as possible and make sure you understand the material. The goal of the assignments is to make sure that you understand how to use the material in class on practical/real-world problems. We will try to provide you with real or realistic data and problems as much as possible -- things you would actually encounter if you were doing data exploration in a professional context.

See above for information on Giving/Receiving help. Copying wholesale (and/or failing to acknowledge your source) will be considered cheating and will be turned over to the academic advising office. Note that I am extremely good at catching people doing this. If you’re good enough to hide the fact that you did it, you probably are better off just doing the assignment.

**Group Projects**

In this iteration of the course, you will form groups to tackle real-world data science problems (one per group). These projects will be centered around data supplied by clients that represent a broad array of sectors. A total of three class sessions will be dedicated to the presentation of your project after the Thanksgiving break. This will allow you, along with your group, to pursue something that is of interest to you and, presumably, to the rest of the class. More details will follow throughout the first half of the semester.

**Grading**

A total of 450 points are available in this course, apportioned as follows:

- In-video quizzes: 100 points
- Quizzes based on readings: 50 points
- Participation (individual & team, typically based on in-class notebooks) (20 x 5 points each): 100 points
- Homework assignments (best 5 of 6, 20 points each): 100 points
- Project (100 points total; see breakdown below):
  - Proposal (team): 10 points
Peer review of proposal (individual): 10 points
Client presentation (team): 10 points
Final report (team, to be shared with client): 70 points

The mapping between points and final grades is:

- 435+  A+
- 415   A
- 405   A-
- 390   B+
- 375   B
- 360   B-
- 345   C+
- 330   C
- 315   C-
- 300   D+
- 285   D
- 270   D-
- <270  you’re not going to get this, so let’s not talk about it

Late Policy

I realize that the occasional crisis might screw 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 370. 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 5pm Friday, with no late days left, penalties would be:

- Before 5pm Saturday: 25% deduction
- Before 5pm Sunday: 50% deduction
- Before 5pm Monday: 75% deduction
After 5pm 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 student-led topic component.**

**Original Work**

See above for information on Giving/Receiving help (which applies exclusively to homeworks). Unless otherwise specified in an assignment, all submitted work must be your own, original work. You may discuss general approaches with others on individual assignments, but may not copy code or answers wholesale and 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 Associate Dean for Academic Affairs.

One caveat in regards to SI370 (Data Exploration) or any other related class. You may not submit the same work to both classes for your final project. We are aware of the students in both and while it is ok to broadly tackle the same problem, the work you turn in for both classes must be significantly different. If you’re not sure about what that means, come talk to us.

**Accommodations for Students with Disabilities**

If you think you need an accommodation for a disability, please let us 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 us 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. We 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://www.umich.edu/~mhealth/