SI 370 - Data Exploration

Fall 2018: TuTh 9:00-10:30am

Note: Some syllabus details may be subject to change.

Instructor: Chris Teplovs (cteplovs@umich.edu)
Office hours: By appointment only (please sign up on my calendar)

GSI: Johan Mosquera (jmosquer@umich.edu)
Office hours: Tuesday 2:00-4:00pm, NQ 1286

Instructional Aide: Xinchun Li (xincli@umich.edu)

Communication

The best way to contact any member of the teaching team is by using Slack.

We try to answer questions that are sent via Slack within about 48 hours. 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 office hours. 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.

Quick Links

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 course offers advanced material for those in the BSI beyond the 106/206 sequence. You will find that SI330 (Data Manipulation) useful to take in advance of this course (or at the same time) but this is not required this year. In SI330 you'll learn 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).

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 are 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 (Links to an external site.)Links to an external site.)

- Compute and visualize summary statistics of datasets (specific implementation in Python, but 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.
- 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 will be no required text for this class; we will be providing PDFs as necessary. That said, there are a number of books that expand on the topics we will be talking about and can be useful for you as you make progress through your homework.

Recommended:

- Foster Provost and Tom Fawcett, *Data Science for Business* (2013)
- Joel Grus, *Data Science from Scratch* (2015)
- Wes McKinney, *Python for Data Analysis* (2012)

- Tamara Munzner, Visualization Analysis and Design (2014)

All are available from the Proquest site for free (on campus: http://proquest.safaribooksonline.com/., offcampus: http://proquest.safaribooksonline.com.proxy.lib.umich.edu

Schedule/Readings

Please see the Schedule of Topics and Readings.

Class format

We meet face-to-face twice a week. Typically, we will spend the first 10-15 minutes of class on a warm-up exercise, about the same amount of time discussing confusing parts of the readings, and then the remainder of the class in an interactive format that combines mini-lecture segments with hands-on lab exercises.

The core part of the course consists of five modules. There is a homework assignment that is assigned at the end of each module and we will spend part of the class getting started on the homework assignment together.

Attendance

Attendance in the class and participation in labs is mandatory. Repeatedly missing class (> 3 times, unexcused) or failing to participate will likely lead to a failing grade.

Readings

There are readings assigned almost every week. These are intended to supplement the face-to-face classes. You will get much more out of class if you read these before you attend. These will be posted in advance of class and made available in the course canvas site. You may find it useful to print (at least the shorter ones) and bring them to class/lab.

Electronics Policy

All the research, and my personal experience, suggests that everyone learns better and enjoys the course more when the distraction of electronics is removed. Clearly, putting away technology isn't possible in this course but I can guarantee that you will do much better in this course if you temporarily silence IMs and other distractions.

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. 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 (Slack), *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 face-to-face meetings are centered around co-constructed Jupyter notebooks. I write these notebooks to support mini-lectures that are interspersed with hands-on exercises that allow you to practice applying the techniques we are learning. At the end of each face-to-face class you are expected to turn in your notebook with the completed exercises. 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 homeworks 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.

Midterm

There will be a 90-minute written test in late October (by default during the class period, but depending on class size we may need to switch to a bigger room: specific date and room to follow shortly), covering all previous material in the course to that date. Except in cases of serious illness or other emergencies, you are expected to take the midterm on the scheduled date. We'll hold a review session prior to the midterm. You will be allowed to bring certain materials into the exam with you. Details will be provided closer to the test date.

Student-led Topics

In this iteration of the course, a total of three class sessions will be dedicated to student-led, group topics. 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. This could be a project similar to that which you completed in SI 330: an in-depth investigation of a particular dataset. A better choice would be to tackle a technique or tool in more depth. You will be responsible for distributing materials to the rest of the class to engage them in during your 20-minute "live" segment. More details, along with suggestions for possible topics, will follow throughout the first half of the semester.

Grading

A total of 600 points are available in this course, apportioned as follows:

In-class notebooks (best 12 of 13, 10 points each):	120 points
Homework (best 5 of 6, 30 points each):	150 points
Student-led topics (60 points group project & presentation	60 points
9 x 5 points each providing peer feedback):	45 points
Midterm:	150 points
Participation (Slack and in-class):	50 points
Bonus (opportunities will be announced):	25 points

The mapping between points and final grades is:

575+	A+
550	A
525	A-
500	B+

475	B
450	B-
425	C+
400	C
375	C-
350	D+
325	D
300	D-
<300	E

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.**

There will be multiple opportunities for extra credit during the course, although not enough to make up for missing multiple assignments.

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://umich.edu/~mhealth/