ECO 395m: Python, Databases, and Big Data

Unique Numbers: 36415, 36420

Course website: https://github.com/edkrueger/eco395m

Class:

Class will meet Tuesday and Thursday **3:30pm - 5:00pm CT**.

TA sessions are optional and will meet Fridays **TBD**.

Class will be held in person in **BRB 1.118**.
TA sessions will be held in person in **BRB 1.118**.

Recordings of lectures will be posted for a limited time on Canvas through Lectures Online.

Instructor: Edward Krueger

- Email: edwardkrueger@utexas.edu
- Office Hours: Tuesdays and Thursdays 5:00pm - 5:45pm CT

- Office Hours Location: **BRB 1.118** (unless otherwise indicated)

Teaching Assistant: Shreeyesh Menon

- Email: shreeyesh.menon@utexas.edu
- TA In Person Office Hours: TBD
- TA Office Hours Location: TBD

Course Description and Requirements

Overview

This is a master's level course on Python, databases and big data. Our goal in this course is to learn the fundamental programming and data skills for working with data using a standard, modern tech stack. This is a foundational but fast-paced course; as such, we will move very quickly. We'll prefer practical applications over theoretical explorations.

Though we'll be learning many tools in this course, we'll focus on the big picture. By the end of this course, you should have a conceptual understanding of the data ecosystem as well as the practical skills required to programmatically access, load and manipulate data.

What this course is not:

- This is *not* a course in computer science. We won't cover topics like designing and evaluating algorithms or the inner working of databases.
- This is *not* a course on machine learning or AI. We won't cover machine learning algorithms. However, the material in this course is complementary: collecting, manipulating and maintaining data is a prerequisite to applied ML/AI.
- This is *not* a course on statistics. We won't cover the *statistical* analysis of data. However, you'll find the material complementary: once data is in the correct form, most statistical tests are available in open-source packages and extremely simple to apply.
- This is *not* a course on application development. However, we will incidentally cover many of the same fundamentals and serve as a strong foundation.

Reading and References

All reading and references for this course are optional but highly recommended. If there is a concept that you do not understand, refer to the recommended books. There are many resources, freely available and otherwise.

That said, the quality of resources, especially the freely available ones, varies tremendously. There is a tremendous amount of bad advice, bad practices and just plain wrong information about coding and data on the internet.

Here is a non-exhaustive list of resources that I recommend and will suggest readings from in class:

- *Unix for the Beginning Mage* by Joe Topjian (Freely available online.)
- *git - the simple guide* by Roger Dudler (Freely available here: https://rogerdudler.github.io/git-guide/)
- *Think Python : How to Think Like a Computer Scientist* By Downey, Allen B. (Get the 2nd or 3rd edition. The 1st edition is written for Python 2, which is rather unlike Python 3. I'll be using the 2nd edition.)
- *Effective Pandas* by Matt Harrison (I'll be using the 1st edition, but the 2nd edition will suffice as well.)
- *Practical SQL: A Beginner's Guide to Storytelling with Data* by DeBarros, Anthony (You may use either the 1st or 2nd edition. The 2nd edition covers some material I will cover that the 1st edition does not. I'll be using the 2nd edition.)

Other resources that I recommend:

- The Python Documentation: https://docs.python.org/3.7/
- The Pandas Documentation: https://pandas.pydata.org/docs/
- The Postgres Documentation: https://www.postgresql.org/docs/

Software

We will use a lot of free, open-source software in this course, including but not limited to Python, Git and Postgres. I'll indicate when you need to install new software in class and sometimes make specific recommendations. If needed, the TAs will assist during TA office hours.

In addition, we'll use GitHub. The free plan will be sufficient for this course.

We'll use Discord as our primary channel for communication for everything except for grades. Announcements will be made on Discord; it is your responsibility to check for them.

Finally, we will use Google Cloud Platform (GCP). GCP charges by use of services but offers a $300 credit to new accounts. This *should* be sufficient for this course. Costs *may*, in some circumstances, exceed the credit amount. We'll provide guidance in keeping costs under control in class. *Please wait to redeem the credit until we instruct you to.*

Prerequisites

There are no formal prerequisites for this course. However, success in this course requires computer literacy, including the ability to install, use and troubleshoot software.

We will start from the beginning with Python. However, some slight exposure to programming is recommended to understand what you are signing up for. Programming is quite different from other skills you have acquired through your academic career.

How to Succeed in This Course

During Lectures

Lectures will be fast-paced, and you must pay attention in class. It can be tempting to replicate every step I take and write the same code I write *while in the lecture*. I don't recommend this as you'll likely fall behind and miss out on new material. So it may be wise to consider not opening your computer at all during class.

Outside of Class

If you'd like to go back and try to replicate the work I do in class, you will have a recording of the lectures. Your homework and group projects will provide you with an opportunity to do it yourself.

After getting the big picture from lectures, the best way to learn to program is to do it. Attending lectures and completing the readings won't be sufficient. In fact, you should consider attempting the homeworks first and consulting the readings if you need assistance.

Please attend office hours.

During Group Projects

A large part of the learning during this course will occur during group projects. You'll harden the skills you learned in class, learn new skills and learn how to work on real-world open-ended projects.

We'll set aside class time during group projects to work on them. It's essential that you attend class during these times. It's the only opportunity you will have access to me, the TAs, and your group members *at the same time*. We'll provide advice on scoping your projects, help you make good technical choices and help you troubleshoot complex issues.

Course Policies

Grading

Grades will be based on:

- Participation (10%)
- Homework Assignments (20%)
- Midterm Group Project (30%)
- Final Group Project (40%)

Final grades will be determined on the basis of the following rubric.

Please note: to ensure fairness, all numbers are absolute, and will not be rounded up or down at any stage. Thus a B- will be inclusive of all scores of 80.000 through 83.999. A = 94-100 A- = 90-93 B+ = 87-89 B = 84-86 B- = 80-83 C+ = 77-79 C = 74-76 C- = 70-73 D+ = 67-69 D = 64-66 D- = 60-63 F = 0-60.

Grades will be *not* be curved. Please do not ask me about extra credit or extra work to improve your grade. None will be given.

Participation

Your participation grade will be based on the extent to which you attend class during project work weeks and participate in your groups' meetings and your attendance during the midterm and final presentations. Should you have exceptional circumstances where you cannot attend class during project work, you must make arrangements with your group and inform me of these arrangements.

Homework Assignments

- There will be 8-12 homeworks assigned during the semester.
- Homework will not be assigned during project weeks but may be due during them.
- Homework assignments from previous semesters are posted in the class repository but may change in part or entirely at any time before I assign them for this semester.
- Due dates will be announced as homeworks is assigned. Homeworks will typically be due a week from when they are assigned. Late homework will *not* be accepted.
- For each homework, you may make an additional attempt *once* after receiving your initial grade. (Effective 2025: In order to be eligible for a regrade, you must have created a repo and have sent an invite by the original due date.) I will assign the due dates for the additional attempt for each homework. Late additional attempts will *not* be accepted. If you make an additional attempt, its grade will replace your initial grade but will incur a 5% penalty. So, for example, if you score a 90% on an additional attempt, you will receive an 85% as your grade for the homework.
- The lowest-scoring homework will be dropped to allow you some flexibility throughout the semester.
- Homeworks will be submitted in *private* Github repositories. You must invite the TA(s) and me as collaborators. We'll be able to find your homeworks if you follow the instructions in the assignments. Homework solutions may not be made public. Grades will be posted on Canvas. You will not need to submit homeworks on Canvas.

Group Projects and Group Project Presentations

- Students will form groups of around five students. (Sizes are subject to change based on enrollment.)
- Students may opt to form different groups for the midterm and final projects.
- Projects will be developed and submitted in public GitHub repositories. You are

encouraged to share your projects.

- Projects will be graded based on *both* group and individual requirements. GitHub will allow us to see who makes which contributions.
- Midterm Projects will be presented in class. Each group will have 3 minutes to present a lightning talk and 2 minutes of Q & A. (Presentation times are subject to change based on enrollment.)
- Final Projects will be presented during the final exam time slot for the course. Each group will have 10 minutes to present their project and 5 minutes of Q & A. (Presentation times are subject to change based on enrollment.)
- You must attend every group's presentation on presentation days, not just your own group's.
- If you know in advance that you have a conflict with one of the presentation dates (e.g., due to travel or religious observance) or require accommodations, please see me as soon as possible to work out an alternative.
- If you will be absent from a project presentation, you must notify me prior to the presentation at edwardkrueger@utexas.edu. Where advanced notification is not feasible, **the notification must be given by the end of the second day after the absence.** Non-excused absences will result in a zero for that portion of your grade.

AI Policy

The creation of artificial intelligence tools for widespread use is an exciting innovation. The University encourages all students to engage with AI responsibly and to understand that there are important limitations to using generative AI for learning.

The use of generative artificial intelligence tools (or Large Language Models [LLMs]) such as CoPilot or ChatGPT in this class shall be permitted on a limited basis.

You may use AI to research *how* to do things (write code, install software, etc.).

You may not ask AI to *generate* code for assignments or projects.

For example, its acceptable to to ask:

- I'm using beautifulsoup to extract elements from a web page, and I'm unsure how to extract the href value from an element. Can you show me how?
- When I run my code, I receive the following error "{error here}". Can you help me understand it?

It is not acceptable to prompt:

- Here is some HTML, write code using beautifulsoup to extract all the url of all of the links.
- Here is my code "{code here}", when I run it it get the error "{error here}", fix my code.

You may not use AI-based code completion.

You can use AI to brainstorm ideas for your projects, but you may not use it to generate a project scope for you.

Citation does not remedy unacceptable use.

Using LLMs as part of your tech stack in your projects is acceptable and encouraged. Additionally, certain homeworks will require the use of AI through an API.

Using generative AI without authorization or failing to cite generative AI use according to the citation policy in this course, even where permitted, may constitute a violation of UT Austin's Institutional Rules on academic integrity and may be referred to student conduct for resolution.

University Disclosures and Policies

Statement on Academic Integrity and Conduct

The University of Texas Honor Code states:

> The core values of The University of Texas at Austin are learning, discovery, freedom, leadership, individual opportunity, and responsibility. Each member of the university is expected to uphold these values through integrity, honesty, trust, fairness, and respect toward peers and community.

Each student in this course is expected to abide by the UT Honor Code and uphold academic integrity.

What this means for this course:
You are encouraged to study together and to discuss information and concepts covered in lecture and the recitation sections with other students. You can work together on homework assignments. However, this cooperation should never involve one student having possession of or copying directly from another student's work. Should such copying occur, both students involved will receive zeros for the assignment. For group projects, the Honor Code means that the work you represent as your contributions must be your own. *Git and GitHub will allow us to see your individual contributions to group projects*.

Sharing of Course Materials is Prohibited

No materials used in this class, including, but not limited to, lecture hand-outs, assessments (homework assignments), in-class materials, and review sheets, may be shared online or with anyone outside of the class - or in future classes - unless you have my explicit, written permission. Unauthorized sharing of materials promotes cheating. It is a violation of the University's Student Honor Code and an act of academic dishonesty. I am well aware of the sites used for sharing materials, and any materials found online that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in sanctions, including failure in the course.

Class Recordings:

Class recordings are reserved only for students in this class for educational purposes and are protected under FERPA. The recordings should not be shared outside the class in any form. Violation of this restriction by a student could lead to Student Misconduct proceedings.

ADA Notice

If you are a student with a disability, or think you may have a disability, and need accommodations please contact Disability and Access (D&A). You may refer to D&A's website for contact and more information: http://disability.utexas.edu/. If you are already registered with D&A, please deliver your accommodation letter to me as early as possible in the semester so we can discuss your approved accommodations.

Harassment Reporting Requirements

Under Texas Senate Bill 212 (SB 212), the professor and TAs for this course are required to report for further investigation any information concerning incidents of sexual harassment, sexual assault, dating violence, and stalking committed by or against a UT student or employee. Federal law and university policy also requires reporting incidents of sex- and gender-based discrimination and sexual misconduct (collectively known as Title IX incidents). This means we cannot keep confidential information about any such incidents that you share with us. If you need to talk with someone who can maintain confidentiality, please contact University Health Services at https://healthyhorns.utexas.edu/ or the UT Counseling and Mental Health Center at https://cmhc.utexas.edu/. You can also make an appointment with a confidential advocate by emailing advocate@austin.utexas.edu or calling (512) 232-2860. We strongly urge you to make use of these services for any needed support and to report any Title IX incidents to the Title IX Office.

Wellbeing Resources:

Grad school is hard. Take care of yourselves and others.

- The Counseling and Mental Health Center serves UT's diverse campus community by providing high quality, innovative and culturally informed mental health programs and services that enhance and support students' well- being, academic and life goals. To learn more about your counseling and mental health options, call CMHC at (512) 471-3515.
- Check out the Longhorn Wellness Center, and these self-care Virtual Mindfulness and Stress Reduction Activities.
- If you are experiencing a mental health crisis, call the CMHC Crisis Line 24/7 at (512) 471-2255.
- If you have concerns about the safety or behavior of fellow students, TAs or Professors, call BCCAL (the Behavior Concerns and COVID-19 Advice Line): 512-232-5050. Your call can be anonymous. If something doesn't feel right – it probably isn't. Trust your instincts and share your concerns.

Course Outline

This is far less reading than it appears to be. These books are far less dense than your typical textbook and contain many examples and sample outputs. Further, these books are largely intended as references. You don't need to read them word for word, but a quick read or skim will make them far more valuable references for when you are working on homework, group projects and future work.

Topic 1 Bash and Git

Suggested Readings: Topjian: 2, 3, 4, 6 -- *not* 3 Advanced, 4 Advanced, 6 Advanced)
Dudler: All

Topic 2 Python and Pandas

Suggested Readings: Downey: 1, 2, 3, 5 (skip recursion topics), 6 (skip recursion topics), 7, 8, 10, 11, 12, 14, 19 (skip Named Tuples), 20 Harrison: 3-11, 14, 16-28, 32, 33 (except "SQL" section)

Midterm Project

Project Work Week -- *Thursday, Oct 9th* and *Tuesday, Oct 14th* will be used for in class project work.
Midterm Project Presentations will be on *Thursday, Oct 16, 3:30pm-5:00pm*.

Topic 3 SQL and Python with SQL

Suggested Readings: DeBarros (1st Edition): 1, 2, 3, 5, 6, 7, 8, 9
DeBarros (2nd Edition): 2, 3, 4, 6, 7, 8, 9, 10, 16
Harrison: 33 ("SQL" section)

Final Project

Project Work Days -- *Thursday, November 20th*, *Tuesday, December 2nd*, and *Thursday, December 4th* will be used for in-class project work.

Final Project Presentations will be on *Saturday, December 13, 8:00 am-10:00 am*.