ECO225

Big Data Tools for Economists

University of Toronto

Department of Economics

Fall 2025

Instructor: Marlène KOFFI

Mail: marlene.koffi@utoronto.ca

Office Hours (online unless otherwise specified; in person GE311):

- Tuesdays 2:10 p.m.-3:10 p.m.

Teaching Assistants (TA):

- Derek Caughy :
  - derek.caughy@mail.utoronto.ca
  - Office hour 1: Mondays 12:10 pm-1 pm in room SS 1085 (subject to change if the time is used for class)
  - Office hour 2: Fridays 12 pm-1 pm in GE 213
- Manahil Malik
  - minahil.malik@mail.utoronto.ca
  - Office hours 1: Mondays 2 p.m.-3 p.m. (Room TBD)
  - Office hours 2: Thursdays 3 p.m.-4 p.m. (Room TBD)

I. Overview

This course offers an introduction to Python programming with a focus on applications to both structured and unstructured data. It is designed primarily for students who are new to coding, while also being suitable for those with some prior experience. The course provides a practical foundation in using Python as a tool for applied analysis, equipping students with the skills to work with data in research and real-world contexts.

Students will begin by learning the essentials of the language—data types, loops, functions, and libraries—and then see how these building blocks can be applied to a range of real-world data sources. Examples will include traditional structured datasets as well as unstructured materials such as text files, webpages, and social media data. Through guided exercises, students will learn how to transform raw information into usable datasets, manage and link data, create informative visualizations, and analyze the data.

The goal by the end of the course is for students to not only be comfortable writing Python code, but also to understand how to apply programming skills to research questions, particularly in economics and the social sciences.

II. Course delivery

Lectures are held on Mondays from 10:10 a.m. to 12:00 p.m. (Toronto time) in room SS 1085. However, it will not be uncommon for us to use the 3 hours of class: 10:10 a.m. to 1:00 pm. If this is the case, you will be notified via Quercus.

This course might have some online components. The lectures may be a combination of online and in-person lectures. In case we adopt the online mode for a given lecture, you will be notified via Quercus.

All course materials will be posted on Quercus. I recommend that you check it regularly. Students are **expected to attend all class sessions and actively participate in discussions and activities**. Given the structure of the class, students must bring their laptops to the different lectures so they can practice coding during the lectures.

Please note that the materials posted on Quercus will not necessarily cover all of the topics discussed during class. Class time is an interactive space between the instructor and students, and additional content—such as explanations, recommendations, practical tips, expectations and grading criteria for projects—may be provided beyond what appears in the written materials. All such content is considered part of the course, and students may be assessed on it through exercises, assignments, projects, and exams. **If a student misses a class, it is their responsibility to catch up on the material covered.**

If the University suggests it, you are also strongly encouraged and requested to wear a mask during in-person classes, tutorials, and Data Camp.

Online course: This course is designed to be interactive, mimicking an in-person session as closely as possible when we are online. Therefore, students must have a usable microphone and are encouraged to activate their camera. For the presentations, the presenters are required to activate their cameras.

Tutorials

Unlike traditional lecture-based tutorials, this course uses tutorial time for individualized support. Tutorials will be structured as opportunities for one-to-one interaction, where you can clarify

questions about the material, exercises, or your projects. These sessions are intended to provide personalized guidance, help you work through challenges, and ensure you are making steady progress. By making regular use of tutorials, you will be able to deepen your understanding of the course content and strengthen your skills in a focused setting.

To ensure that every student receives adequate support and is well-positioned to succeed in this course, attendance at tutorials is required and will be monitored. Each student must attend at least one tutorial session every two weeks. You could choose between this traditional tutorial time and the office hour time. Again, those times are designed to provide hands-on practice, opportunities to ask questions, and guidance on assignments and projects. You will also report in a biweekly mode on your advancement on your research project. Regular participation will help reinforce the material covered in lectures and give students the chance to address challenges early, ensuring steady progress throughout the term. This is just a guideline. You are welcome to attend as many as needed.

NB: Please note that because we have three hours allocated for class and tutoring combined, we may need to adjust the schedule as necessary. This could mean more class time, more tutoring time, or a complete remodel. You will be notified if any changes occur.

Time zone

All times posted will be in local Toronto time. Errors in calculations are not an acceptable reason to miss deadlines.

III. Evaluation

Task		Weight	(Due) Date
Class Participation		10%	Every Lecture
Assignments	Weekly exercises	10%	During every Lecture
	Take-home assignments	10%	October 6, 2025 10 AM EST November 10, 2025 10 AM EST
Research project	Part 1: Submission of ideas	12%	October 20, 2025 10 AM EST

	Part 2: Slides submission (selection of top projects)	12%	November 17, 2025 10 AM EST
	Part 3: Final submission *Code *Report	16%	December 2, 2025 10 AM EST
	Bi-weekly monitoring	5%	At least one "tutorial"/office hour session every two weeks
	Oral presentation	Up to 10 bonus points on your research project grade	Lecture of December 1rst, 2025
Final Exam		25%	Final Assessment Period

III.1. Class Participation:

Participation marks will be based on your level of engagement throughout the course. These are determined by your attendance at lectures, asking live questions, and engaging in chat discussions during the course (in the online format). You are also encouraged to use the discussion platform on Quercus. We might use the Quizzes section on Quercus as well. I will keep track of all these interactions during the course.

To monitor attendance, students will be required to submit their Python code from the class activities at the end of each session. This submission serves both as a record of participation and as a way to ensure that students are actively engaging with the material during class.

III.2. Assignments:

Take-home assignments will be posted one week before the due date. Late assignments will receive a grade of zero (see late policy). You will have to upload the assignments via Quercus.

Take-home assignments will also provide an opportunity to explore concepts that we will not have time to cover in class through exercises that utilize the tools you are already familiar with.

Take-home assignments are individual. You can help each other, but each student will have to submit their own "copy" and their own "code" in conformity with the Student Academic Integrity Code (See below).

Each week, short in-class exercises (weekly exercises) will be used to assess your understanding of the course content. These activities are designed to be brief, ranging from 1 to 15 minutes, and will take place during lecture time. The purpose of these exercises is twofold: first, to give you regular opportunities to apply the concepts introduced in class, and second, to provide the instructors with feedback on which topics may require further clarification.

Exercises are completed individually and must be submitted through Quercus, either by uploading a file or through the Quercus Quiz section, depending on the format of the activity.

III.3. Research Project:

This assignment will serve as an introduction to the process of conducting research. It is designed to give you firsthand experience with formulating a question, applying analytical tools, and interpreting results—much like in real-world economic research. The project will be completed individually, allowing each student to develop their own ideas and demonstrate independent work.

You need to find a relevant question and apply the tools seen in class to answer that question. Your question may draw on any area of economics, including but not limited to: Macroeconomics, Microeconomics, Labor Economics, Health Economics, Economics of Education, Economics of Science, Economics of Innovation, Economics of Sport, Development Economics, International Economics/Trade, Monetary Economics, Public Economics, Urban and Regional Economics, Environmental and Resource Economics, Industrial Organization, Behavioral Economics, Financial Economics, Political Economy, Economics of Technology and Digital Markets, Economics of Inequality and Distribution, Cultural Economics, Agricultural Economics, Experimental and Computational Economics, etc.

Once you have identified your research question, you will be required to collect the relevant data. The data may be structured, unstructured, or a combination of both, depending on the nature of your question.

You will then carry out a series of tasks, including:

- Cleaning and preparing the data.
- Generating descriptive statistics.
- Summarizing and visualizing patterns in the data.
- Estimating a regression model.
- Developing a machine learning model.
- Drawing conclusions from your analysis, written as if you were advising a policymaker.

We will also evaluate your creativity (e.g., identifying a novel dataset, formulating an original question), effort (e.g., undertaking intensive data collection or detailed analysis), and relevance of your analysis (e.g., producing meaningful insights with an interesting policy conclusion).

Please note that datasets must not be taken from sources where the associated code and analysis are already provided (for example, Kaggle).

a. *Part 1: Submission of ideas*

You will prepare a short slide deck (maximum 7 slides, excluding the title slide) and follow the structure below. The goal is to communicate your project clearly, concisely, and reproducibly.

- Slide 1 — Title & Identifiers
  - Project title
  - Full name
  - UTORid, University of Toronto email address
- Slide 2 — Motivation
  - Research question: state it precisely.
  - Why it matters: policy or scientific relevance; who should care and why?
- Slide 3 — Data: At this stage, your dataset must be collected (for very large datasets, collection should be nearly complete).
  - Source(s): origin, access method (API, bulk download, web scraping, etc.), links/citations.
  - Collection steps (detailed): describe exactly how you obtained the data (e.g., endpoints, query parameters, scraping logic, file formats).
  - Python commands used: show the key commands/snippets for extraction and reading (e.g., requests, BeautifulSoup, pandas.read_csv, json, geopandas). Briefly justify why these tools were appropriate.
  - Data characteristics: time span/coverage, unit of observation, number of observations and variables, key variables, basic data issues (missingness, outliers, duplicates).
- Slide 4 — Analysis Plan

Provide a clear, step-by-step plan for the remaining work:

  - Data preparation: cleaning steps, feature engineering, linking/merging.
  - Descriptives & visualization: tables/figures you will produce.
  - Timeline & risks: milestones, anticipated challenges, and mitigations.

Code Submission (required): Along with your slides, submit:

- Python code for data extraction and reading, and the code used to produce the data characteristics shown on Slide 3.
- Provide the code as a Jupyter notebook (.ipynb) or script (.py).
- Submissions are via Quercus (file upload or Quercus Quiz, as specified)

b. Part 2: Slides submission (selection of top projects)

This stage is a more complete version of Part 1. You are expected to have addressed and incorporated all feedback received from your initial submission. Slides 1 to 3 (Title, Motivation, and Data) will remain unchanged, but you will add the following components:

- Descriptive Statistics: well-organized tables summarizing key variables.
- Data Visualization and Summarization: plots, charts, or figures that highlight important patterns in the data.
- Correlational Analysis: results showing relationships between key variables.
- Regression Analysis: specification(s) and results with clear interpretation.
- Machine Learning Analysis (if applicable): preliminary results from one model.

You should not exceed 20 slides in total. At this stage, your project should be close to completion. While we will be lenient if the machine learning component is not yet included, it must be incorporated in the final version of your project.

This submission will also serve as the basis for selecting the top research projects that will be presented to the entire class. Depending on time availability, we will select 7 to 10 projects that demonstrate originality, rigor, and clear presentation of results.

c. Part 3: Final submission

For the final submission, you are required to prepare a **written report of no more than 15 pages, along with your code and data**. The report should incorporate feedback received on your earlier submissions and must follow the structure outlined below:

Report Structure

- Introduction: motivation for the study, research question(s), and an overview of the data and key results.
- Data and Descriptive Analysis: description of the dataset, key characteristics, descriptive statistics, and visualizations.
- Regression Analysis and Machine Learning Models: presentation of specifications, results, and interpretation.
- Conclusion: summary of findings, limitations, and implications (e.g., policy relevance).

Writing Guidelines

- Maximum 15 pages, including graphs and tables.
- Font size 12, double-spaced.
- A separate title page (not counted in the 10 pages) should include: project title, full names, UTORids, and email addresses.

d. Oral presentation

As the final step, you will present your project in front of the entire class (conditional on being selected). The presentation will take the form of slides, and selected students will receive guidance during office hours on how to prepare an effective presentation.

Bonus Grade

The oral presentation is graded as a bonus component to encourage high-quality work. By being selected to present, you may earn up to 10 bonus points on your research project grade. Given that the research project accounts for 45% of the overall course grade, this translates into a possible 4.5-point boost to your final course grade. This bonus can significantly improve your final standing. For example, if you earn 10 points, this will usually move you to an upper grade regimen (B to B+, B+ to A-, A- to A, A to A+).

The incentive is therefore substantial and provides strong motivation to produce your best possible research proposal.

In addition to the grade bonus, being selected to present offers professional benefits. Producing and delivering a high-quality research presentation is an excellent experience for future academic or professional settings. You may also highlight this achievement on your CV or in job and graduate school interviews as evidence of your ability to carry out independent research, communicate results clearly, and present in front of an audience.

III.4. *Final Exam*

The final exam will be administered in accordance with the University's regulations for final assessments. No course materials, external resources, or online tools will be permitted during the examination. The purpose of the exam is to evaluate your knowledge of Python and the material covered throughout the course. In light of the growing use of Generative AI platforms, the exam will place particular emphasis on assessing your understanding of coding concepts rather than the mechanical use of such tools. Several in-class exercises will reflect the type of questions you can expect on the exam.

Important Note: You are expected to work on the final exam by yourself, proctored or not. You may not use unauthorized aids or communicate with others about the test.

III.5. Other considerations

Late penalty:

All the evaluations are due at the "Due Date" indicated in the table.

Late assignments will receive a grade of zero unless a **valid** reason for absence is provided.

Late research projects will be downgraded by 10 percentage points for each day of delay.

If you do not take the final exam within the specified windows, you will receive a score of zero unless otherwise notified by the administration.

Infraction:

All submitted work may be subject to peer review and plagiarism checks. Ensure that your project, presentation and all submitted assignments reflect your own work and proper academic standards (see below).

Missed Work Policy

Missed work will not be accepted due to technical difficulties, confusion about deadlines, internet or hardware problems, or submission errors (e.g., sending the wrong file, incomplete files, or missing sections of an assignment).

You are permitted to miss one weekly assignment over the course of the semester without penalty. This allowance should be used wisely and reserved for unforeseen circumstances.

Remarking Policy

Students wishing to request a re-mark must do so no later than two weeks after the work has been returned. Please note that a request triggers a re-evaluation of the entire assignment or exam, not just the specific section in question.

If a re-mark is granted, the resulting grade will become the new official grade, whether it is higher, lower, or unchanged. By proceeding with a re-mark request or subsequent appeal, the student acknowledges and accepts this condition.

IV. Course Material

Given the structure of this course, we will not rely on a single textbook. Instead, we will draw selectively from a set of useful references, which will serve as supporting materials throughout the course.

- Provided lecture notes and Python codes.
- Online courses and problems from https://www.datacamp.com.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer. (JWHT): http://faculty.marshall.usc.edu/gareth-james/ISL/.
- Selected book sections posted on Quercus

Software

All programming activities in this course will be conducted exclusively in Python, which is open-source and freely available. Students are required to install Python on their personal computers. We will primarily use either Spyder or Jupyter Notebook as development environments. Students are strongly encouraged to avoid using alternative Python interfaces, as doing so may significantly hinder code debugging and the ability to receive effective assistance.

Prerequisites

The professor is not able to modify or waive the prerequisites for this course. If you have any questions, please contact the undergraduate administrative staff in the Economics Department.

V. Topics and Required Readings

Chapter 0: Introduction to Python

Chapter 1: Managing Databases in Python

Chapter 2: Scientific Computing with Python

Chapter 3: Data Visualization

Chapter 4: Web Scraping

Chapter 5: Text as data: an introduction

Chapter 6: An introduction to other unstructured data

Chapter 7: Linear regression

Chapter 8: An Introduction to Machine Learning

Important notices:

(1) The schedule is tentative and may be adjusted depending on the relative length and complexity of certain chapters.
(2) We will aim to cover as much material as time permits.
(3) The schedule should be viewed as a road map of the fundamental concepts that students are expected to learn and review in preparation for each assignment.
(4) Please reserve December 1st for the presentation of the final project.

Email Policy

Email should be reserved for private matters or to notify me of technical issues (e.g., broken links, typos). Questions about course content will be addressed in person during lectures or office hours.

Before sending an email, please follow these steps:

1. Check whether your question has already been answered in the syllabus or on Quercus announcements.
2. For coding questions, remember that Python is open-source; most answers can be found online (e.g., via Google).
3. Post your question on our discussion platform to engage with classmates.
4. Consult the teaching assistants (TAs).
5. If you have tried all of the above and still need clarification, you may contact me by email.

Guidelines for Email Communication:

- Emails should be sent from your University of Toronto email address.
- The subject line must include the course number and your UTORid.
- Always use a proper salutation and maintain a polite, professional tone.
- If your message requires only a short response (e.g., one sentence), I will reply within three business days. More complex questions should be raised in class or office hours.
- I will not respond to questions that are already addressed on Quercus or in the syllabus.
- Requests for grades, solutions to problem sets, or midterm answers are not appropriate by email.
- Emails that do not follow these guidelines may not receive a response.

Academic Integrity

Academic integrity is essential to pursuing learning and scholarship in a university, and ensuring that a degree from the University of Toronto is a strong signal of each student's academic achievement. As a result, the University treats cases of cheating and plagiarism very seriously. The University of Toronto's *Code of Behaviour on Academic Matters* (http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/ppjun011995.pdf) outlines the behaviors that constitute academic dishonesty and the processes for addressing academic offenses. Potential offenses include, but are not limited to:

In papers and assignments:

- Using someone else's ideas or words without appropriate acknowledgment.
- Submitting your own work in more than one course without the permission of the instructor in all relevant courses
- Making up sources or facts
- Obtaining or providing unauthorized assistance on any assignment

On tests and exams:

- Using or possessing unauthorized aids
- Looking at someone else's answers during an exam or test
- Misrepresenting your identity

In academic work:

- Falsifying institutional documents or grades
- Falsifying or altering any documentation required by the University, including (but not limited to) doctor's notes

All suspected cases of academic dishonesty will be investigated following procedures outlined in the *Code of Behaviour on Academic Matters*. Please have a look at these sections on Perils and Pitfalls https://www.academicintegrity.utoronto.ca/perils-and-pitfalls/ and Smart Strategies https://www.academicintegrity.utoronto.ca/smart-strategies/. Also, see the U of T writing support website at https://writing.utoronto.ca/. We may use Turnitin or Ouriginal for the final submission of the research project. More details will be given during class.

University disclaimer concerning Turnitin:

"Normally*, students will be required to submit their course essays to Turnitin.com for a review of textual similarity and detection of possible plagiarism. In doing so, students will allow their essays to be included as source documents in the Turnitin.com reference database, where they will be used solely for the purpose of detecting plagiarism. The terms that apply to the University's use of the Turnitin.com service are described on the Turnitin.com website."*

Class Materials Policy:

Class materials are subject to the University's policy on intellectual property. Without the instructor's explicit permission, it is strictly forbidden to copy, share, or distribute any class materials except for current academic use purpose.

Code of Conduct in an online environment

- The first thing to recall is that we are in a learning environment. Mistakes, discussions, exchange of ideas, etc., are acceptable as long as they are made in total respect for the person and individuality.
- Please mute yourself unless you need to talk for the class's benefit (ask a question, answer a question, etc.).
- To avoid unpleasant interruptions, when you want to ask a question, please use the chat function or wait for the time allowed to do so. As you will notice, I frequently ask if there are any questions. When asked, you are welcome to unmute yourself and ask any question you may have.
- Again, I make a point of honor to have a respectful environment during class. So please, respect your peers. Use proper and respectful language and refrain from any insults, threats, or bad jokes.
- Finally, adhere to the same standards as you would in the classroom.

Academic Accommodations

The University is committed to accessibility. If a student requires accommodations for a disability or has any accessibility concerns about the course, please contact Accessibility Services as soon as possible. Their website is http://www.studentlife.utoronto.ca/as.

Other important notice

The students are expected to comply with all University policies even if not expressly mentioned above.

Absence

Please refer to the University guidelines: https://www.artsci.utoronto.ca/faculty-staff/teaching/academic-handbook#MissedTermWork

Generative AI

The University has created sample statements for instructors to include in course syllabi and course assignments to help shape the message to students about what AI technology is, or is not, allowed. Please read the following material: https://www.viceprovostundergrad.utoronto.ca/wp-content/uploads/sites/275/2023/04/Syllabus-Language-for-Gen-AI-April-2023.pdf