Economia, Finanza e Analisi dei Dati

Laurea Magistrale

Data Analysis for Economics

01 - Introduction

Marco Nieddu

Fall 2025

This course: what it is

- **New** (2nd edition) course on coding (mainly Python) to collect, organize, and analyze data in economics
- **Course goal**: provide you with knowledge to
  - go from a research idea to the final dataset that allows you to test it
  - write results of your analysis in an 'economic report/paper' structure, with appropriate **charts** and tables
- **Skills you will develop**: data visualization, cleaning, wrangling, programming, etc.
  - practical (and often neglected) skills useful for dissertation and career
  - important prerequisite for 'standard' metrics/policy evaluation course in next term(s)

This course: what it is

- Besides coding, this is also a **Topics in Applied Economics** course
- We will work with real data to answer economic questions such as:
  - Gender pay gap (Labor econ)
  - Prices and taxation (Public econ / IO)
  - Religious holidays and productivity (Political econ)
- At the beginning of each topic, we devote space to economic theory and related literature (before data work)

This course: what it is *not*

- Econometrics is **not** the core of this course
  - Although we may cover some econometrics/statistics concepts and intuitions
- We fill the gaps left by traditional econometrics and methods classes
  - e.g., how to actually find datasets in the wild and clean them

Course preliminaries I: People

- Me: Marco Nieddu
  - Associate Professor in Public Economics (*Scienza delle finanze*)
  - I work on applied microeconomics (mostly education/development)
  - I'm an (applied) economist using data science tools for economics research, **not a data scientist!**
- Teaching Assistant: Alessio Garau
  - He will help with projects (more later) + give extra classes on programming basics and syntax

Course preliminaries II: Course Material

- No single textbook; we borrow material from several sources
- The syllabus combines my own material with that of other authors (**which I fully acknowledge**)
- All teaching material will be available in the course shared folder*
  - *Instructions on setup in next slides
  - New material posted online before each class

Course preliminaries II: Course Material

- We use OneDrive and/or Dropbox
- Slides and material posted before class
- Files update live during classes, so you can follow scripts if you get lost
- **Please make sure you are added to the shared folder(s), and that you can see them on your computer (synced).** In case you cannot, please send me an email **ASAP**.

Course preliminaries II: Course Material

Some notes/books I borrow from:

- Pedro H.C. Sant'Anna (Emory University), "Data Science for Economics", link
- Jared Hutchins (University of Illinois), "Data Science for Applied Economics", link
- Grant McDermott (Amazon), "Data Science for Economists", link
- Tyler Ransom (University of Oklahoma), "Data Science for Economists", link
- Florian Heiss and Daniel Brunner, "Using Python for Introductory Econometrics", link
- Gábor Békés & Gábor Kézdi, "Data Analysis for Business, Economics, and Politics". Cambridge UP, 2021

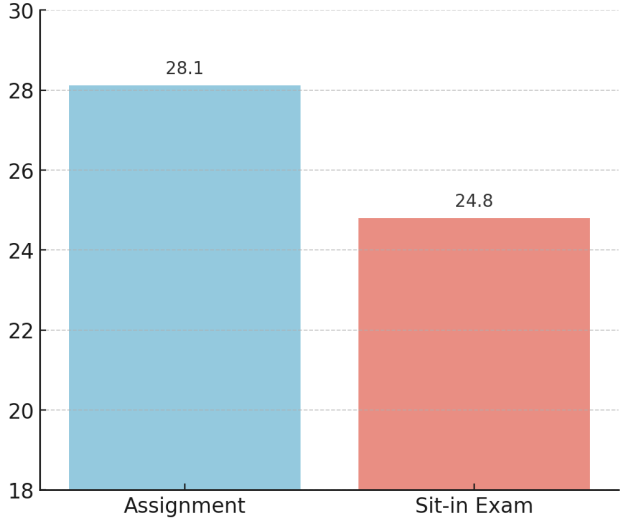
Exam/Grading

	Component	Weight	When	Rules
1	Individual project	50%	Nov. 16th	Open book (GPT ✓)
2	Sit-in exam (PC)	50%	Jan. 19th*	Closed book (GPT ✗)

OR

	Component	Weight	When	Rules
1	Sit-in exam (PC)	100%	Jan. 19th	Closed book (GPT ✗)

Which option to choose?



1. Individual project

- Output: 5-10 page essay on topic of your choice (economics-related)
- Includes finding data, collecting/cleaning, and performing brief analysis, possibly following this structure:
  1. Abstract (whole story in 150 words)
  2. Research question (why important)
  3. Brief review of literature (what we know)
  4. Data description (which data/why/how they look)
  5. Results: 3-4 objects to summarize findings (+ comments)
  6. Scripts (and source file) to reproduce* figures/tables/analyses

*you may be asked questions about the code (No "IDK, GPT did this")

1. Individual project (cont'd)

- To get the grade, you must also give short (10-15 min) presentation on your project at a later (but not final) stage
  - For me, needed to ensure substantial human contribution (I don't grade GPT)
  - For you, chance to receive feedback and improve project before submission
- **Important dates:**
  - Deadline to present research idea: October 8th
  - Presentations: November 3rd and 4th
  - Deadline to submit final project: November 15th

2. Sit-in Exam

The exam is a mini-version of the research project

- You are expected to download/import some data, analyze, and write a very short comment
- Task simple enough to be completed in three hours
- **Closed-book exam** (No GPT!), but you may look at course material
- Rationale: incentivize using GPT as learning tool (a complement), not as substitute

Course roadmap

- Lectures: Mon/Tue/Wed 12-14 (last class: November 5th)
- Additional TA classes (Thu, time TBA)

Week	Day	Title	Data
W1	29sep - 01oct	Intro and Python basic syntax	
W2	06oct - 08oct	Topic 1: Gender differences in labor market	Eurostat
W3	13oct - 15oct	Topic 2: Gas prices and tax policies	MISE
W4	20oct - 23oct	Topic 2: Gas prices and tax policies	MISE
W5	27oct - 29oct	Topic 3: Religious festivals and development	Wikipedia
W6	03nov - 05nov	Presentations and Mock Exam	

Our schedule is ambitious and may be adjusted during the course!

Getting started

- Download and install Anaconda:
  - https://www.anaconda.com/download
  - Anaconda includes Python, Spyder, and Jupyter
- Download STATA (you have access as UNICA students)

Why are we using Python?

We are going to focus on one language: **Python**. Why?

1. General purpose
2. Open source
3. Relatively easy to use
4. Bonus: GPT can help a lot when coding in Python

But we will also see the **STATA** equivalents of scripts we write for each project/topic

1. General Purpose

- **R and Stata** aimed mainly at data analysis and processing
- **Python** not specific to data analysis. Used for ML, web scraping, even apps

Advantage: Python can do nearly all tasks

Disadvantage: Its econometrics support is weaker than R and Stata

2. Open Source

- **Stata** costs a lot, sold by private company
- **Open Source languages** (R, Python) free to use
- Maintained by a community, packages exist for almost everything

Disadvantage: Packages do not always work well together

3. Easy to use

- Compared to other languages, **Python** is intuitive and easy
- High-level, object-oriented, interpreted $\rightarrow$ easy to debug

Disadvantage: Slow, not always memory efficient

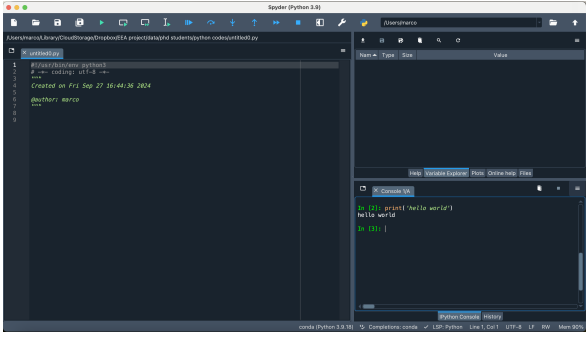
4. Increasingly popular

- https://pypl.github.io/PYPL.html
- https://r4stats.com/articles/popularity/

Getting started II

Once you have installed Anaconda

1. Launch Spyder
   - We are going to use Spyder (Python scripts) as it makes reproducibility easier
2. Once open, in the console (bottom right), type: `print('hello world')`



If you see 'hello world' below the command line, you are good to go. This is where we will start tomorrow.

Addendum: Data visualization

- When it is beautiful
  - https://www.reddit.com/r/dataisbeautiful/
- When it is not
  - https://www.reddit.com/r/dataisugly/