Data Analysis for Economics

01 - Introduction

Marco Nieddu

Fall 2024

This course: what it is

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

This course: what it is *not*

- Econometrics is **not** the core of this course.
  - Although we will (re-)cover some econometrics/statistics concepts, mostly through simulations.
- We are going to fill in 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

- Lectures: Mon/Tue/Wed 4-6pm (the last class is on November 6th)

Mentimeter code:
1398 6831

Course preliminaries II: People

- Me: Marco Nieddu
  - Assistant professor (RTD/B) in Public Economics (*Scienza delle finanze*).
  - I work on applied microeconomics topics (mostly education/development).
  - I'm an (applied) economist who uses data science tools for research projects in economics **not a data scientist!**.
- Teaching Assistant: Alessio Garau
  - He will help you with the projects (more on this later).

Course preliminaries III: Course Material

- We will not follow a single textbook in a chapter-by-chapter manner but rather borrow material from several textbooks/lecture notes, as no ideal textbook is available.
- The syllabus is the result of combining my own material with material from different authors (**which I fully acknowledge**).
- All the relevant material for this class is (going to be) available on the UNICA repository:
  - UNICA DAEc2024
  - New material will be posted online before each class.

Course preliminaries III: Course Material

Some notes/books I borrowed from:

- Pedro H.C. Sant'Anna's (Emory University) notes on "Data Science for Economics", freely available at: https://psantanna.com/Econ520/index.html.
- Jared Hutchins's (University of Illinois) notes on "Data Science for Applied Economics", freely available at: https://github.com/jphutch/ACE-592-SAE.
- Grant McDermott's (Amazon) notes on "Data Science for Economists", freely available at: https://github.com/uo-ec607/lectures.
- Tyler Ransom's (University of Oklahoma) notes on "Data Science for Economists", freely available at: https://github.com/tyleransom/DScourseS24.
- Florian Heiss and Daniel Brunner's book "Using Python for Introductory Econometrics", freely available at: https://www.urfie.net.
- Gábor Békés & Gábor Kézdi, Data Analysis for Business, Economics, and Politics. Cambridge University Press, 2021.

Exam/Grading

	Component	Weight	When	Rules
1	Individual project	50%	Nov. 15th	Open book (GPT ✓)
2	Sit-in exam (PC)	50%	Nov. 7th	Closed book (GPT ✗)

1. Individual project

- Expected output: 5-10 page essay on a topic of your choice (but economics-related).
- This includes finding the appropriate data, collecting and cleaning them, and performing a (brief) analysis to say something on this topic, possibly following this structure/checklist:
  1. Research question (why it is important),
  2. Brief review of the relevant literature (what we know already),
  3. Data description (which data/why/how do they look),
  4. Results: 3-4 objects to summarize your findings (+ comments),
  5. Scripts (and source file) to reproduce all figures/tables/analyses.

1. Individual project (cont'd)

- To get full points, the submitted material should allow me to replicate your results smoothly.
- ... and you should be able to interpret/explain the code during the presentation
  - No "IDK, GPT did this"
- **Important dates:**
  - Deadline to present your research idea: October 15th.
  - Presentations: November 5th and 6th.
  - Deadline to submit the final project: November 15th.

2. Sit-in Exam

The exam is going to be a mini-version of the research project.

- You are expected to download/import some data, analyze them, and write a very short comment.
- The task is simple enough to be performed in a two-hour time.
- **Closed-book exam** (No GPT!), but you are allowed to look at the course material.
- Rationale: incentivize using GPT as a learning tool (a complement), rather than as a substitute for your work.
  - In line with the course AI policy framework (credits to P. H. Sant'Anna).
- **Important dates:**
  - Mock exam: November 4th.
  - Real exam: November 7th.

Course roadmap

	Week	Day	Title	Data
1	W1	30-set	Intro	
2	W1	01-ott	Python basic syntax	
3	W1	02-ott	Data visualization	
4	W1	03-ott	Metrics through simulations	
5	W2	07-ott	P1: working with ready-to-go data	Eurostat
6	W2	08-ott	P1: working with ready-to-go data	Eurostat
7	W2	09-ott	P2: from ready-to-go to DIY	Gas prices (MISE)
8	W3	14-ott	P2: from ready-to-go to DIY	Gas prices (MISE)
9	W3	15-ott	P2: from ready-to-go to DIY	Gas prices (MISE)
10	W3	16-ott	P3: more on API/web scraping	Wikipedia
11	W4	21-ott	P3: more on API/web scraping	Wikipedia
12	W4	22-ott	P3: more on API/web scraping	Wikipedia
13	W4	23-ott	P4: geo data and mapping	OpenStreetMap
14	W5	28-ott	P4: geo data and mapping	OpenStreetMap
15	W5	29-ott	P4: geo data and mapping	OpenStreetMap
16	W6	04-nov	Mock exam	
17	W6	05-nov	Presentations	
18	W6	06-nov	Presentations	

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

Getting started

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

Why are we using Python?

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

1. It is general purpose.
2. It is open source.
3. It is relatively easy to use.
4. Bonus: GPT can help you significantly when it comes to coding with Python.
   - Not as reliable when it comes to STATA

Although from time to time we are going to see the STATA equivalents.

1. General Purpose

- **R and Stata** are aimed primarily at data analysis and data processing.
- **Python** is not specific to data analysis. It is used for machine learning, web scraping, or even writing applications.

Advantage: Since Python is not specific to any one task, it can do nearly all of them.
Disadvantage: Its "econometrics" support is lacking compared to R and Stata.

2. Open Source

- **Stata and Matlab** cost a lot of money, being sold by private companies.
- **Open Source languages** (e.g., R, Python) are free to use.
- Maintained by a community, there are packages for basically everything.

Disadvantage: Packages do not always work well together.

3. Easy to use

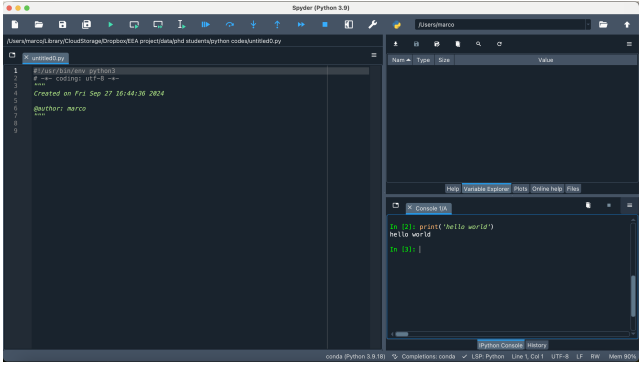
- Compared to other languages, **Python** is intuitive and easy to understand.
- Being high-level, object-oriented, and interpreted makes it easy to use and debug.

Disadvantages: It is slow, and not always memory efficient.

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 the following: `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/