Introduction to Data Analysis with Python

Cecilia Graiff

October 9, 2025

ALMAnaCH, Inria Paris - École d'Affaires Publiques, SciencesPo
cecilia.graiff@sciencespo.fr

Homework Correction

Course recap

Course contents (adjusted)

- **Week 1**: Introduction to the course, Python installation, environment setting, basic Git tutorial, Jupyter Notebook tutorial.
- **Week 2**: Python coding basics (data structures, loops, functions, basic operations).
- **Week 3**: Data cleaning and basic manipulation (`numpy, pandas`)
- **Week 4**: Data visualization (`matplotlib, plotly`).
- **Week 5**: Exploratory Data Analysis (EDA) and more advanced manipulation (advanced `pandas` and `numpy`).
- **Week 6**: Basic statistic and linear models with `sklearn`; introduction to machine learning.

What is expected from you

Project Description

Due: Delayed to November 15, 2025, 23:59

What to upload: PDF with project proposal (2-3 research questions, planned pipeline) and group members' names

Final Project

Due: December 20, 2025, 23:59

What to upload: Complete project report in form of a paper (5-8 pages), commented code

Deadline cannot be delayed!

Resources for scientific writing

- Writing a scientific paper (ETH Zurich, 2019)
- Writing a scientific article: A step-by-step guide for beginners
- How to write your first research paper (NIH 2011)
- ... and many others!

Project checklist

- **Code**:
  - Comment each function to explain to me what it does
  - Upload the code on **GitHub** and share the repository with me
  - Document your repository structure in the **README** file

If you do not know how to use GitHub, you can refer to the guide I uploaded on Moodle.

Project checklist

- **Paper**:
  - 4-8 pages in English
  - Structured as a **research paper**:
    - **Introduction**: introduce your research questions (RQs) and their motivation
    - **Related Work**: ground in the literature your choice of RQs and methods
    - **Methods**: explain your pipeline and how you implemented it
    - **Discussion**: present your results (e.g. in form of graphs or tables) and interpret them **qualitatively** and **quantitatively**
    - **Conclusion**: sum up your work

Getting started with data

Raw data

- **Raw data** comes in many formats and has many uses
  - Tests like the PISA analysis help assess students' performance and evaluate the need for resources, teachers, etc
  - Scanning your public transportation subscription helps designing transportation policy
  - Wage data help design tax policies
  - And much more!

What to do with the data

The two first steps of a data analysis pipeline are the following:

- **Collect data**: Python offers many ways of reading existent datasets, or even of gathering data from Internet sources.
- **Clean data**: Datasets are often messy, with wrong and missing entries.

To perform a good analysis, the first rule is to have **good quality data**, so the cleaning step is fundamental.

Lab

- We will focus on `numPy and Pandas` for reading datasets and performing basic manipulation
- This course will focus more on the analysis of existing datasets; techniques to gather data will only be briefly mentioned on the side during the 6 lectures.