Data, Data Storage, Data Collection

Project Guidelines

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

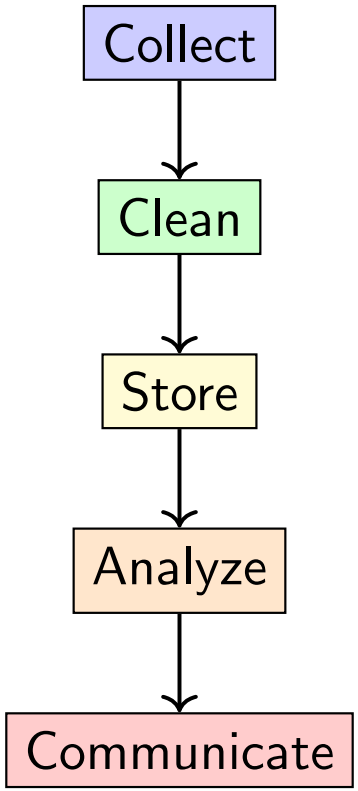
Project: End-to-End Data Lifecycle

The goal of the project is to follow the **full data lifecycle**: from collection and cleaning to analysis and communication, using **real datasets**.

This is your chance to bring together concepts from class in a practical, open-ended task.

Goal

1. Choose a **dataset**, either from open data or one you build yourself. Proprietary datasets are not allowed.
2. Apply the **lifecycle steps**: collect, clean, store, analyze, communicate. Even small choices (e.g., handling missing values) should be documented and explained..
3. The outcome should not just be numbers: your project should **tell a story** and generate new insights.



What You Should Do

Collection : select or acquire a dataset.

Cleaning : handle missing values, errors, and inconsistencies.

Storage : choose a format and justify your decision.

Analysis : compute descriptive statistics and derive insights.

Communication : produce visualizations and a short written report.

What You Should Not Do

You are not expected to build complex machine learning models.

You are not expected to use advanced statistics beyond what we cover in class.

The emphasis is on careful data handling, transparent decision-making, and clear communication.

Focus on reproducibility and clarity rather than complexity.

Process & Milestones

1. **23.09** (next week): Pitch session
2. **06.10** (two weeks later): One-page project proposal
3. **03.12** (one week before the end of the lecture): Project report
4. **08.12 or 09.12**: Group presentations

Pitch Session: 23.09 (next week)

Each student will pitch for 1–2 minutes. Include:

- The dataset you want to use (or how you plan to build it),
- The main question you want to explore,
- Why the question is interesting or important,
- Any challenges you anticipate.

After the pitches, we will finalize the groups.

One-Page Proposal: 06.10 (two weeks later)

Each group will submit a short one-page proposal.

The proposal should include:

- **Names** of group members,
- The **dataset** you will use (or plan to build),
- The central **question** you will explore,
- The **lifecycle steps** that are most relevant.

The proposal fixes the scope of your work and must to be approved (by me).

Project Report: 03.12 (one week before the last lecture)

Each group will submit a **report** of up to 5 pages alongside the **code** (Jupyter notebook) that can reproduce your analysis.

The report should clearly identify the **question** that you tried to solve, provide some **documentation about the dataset**, and present the **results** of the analysis. Additionnaly, the report should offer a **reflection** on the various steps of the data lifecycle, putting forward the **difficulties** encountered and the **lessons** learned.

Group Presentation 08.12 or 09.12

Each group presents for 15 minutes, followed by questions.

The presentation, report, and code must be consistent.

This is your chance to share results and highlight the most interesting aspects of your project.

Evaluation Criteria

Projects will be evaluated on:

- Understanding and application of the data lifecycle.
- Clarity and reproducibility of the codebase.
- Quality and readability of the report and presentation.
- Creativity in dataset choice and approach.
- Teamwork and fair contribution.

Evaluation Criteria

Projects will be evaluated on:

- Understanding and application of the data lifecycle.
- Clarity and reproducibility of the codebase.
- Quality and readability of the report and presentation.
- Creativity in dataset choice and approach.
- Teamwork and fair contribution.
- Quality of the pitch (next week).
- Quality of the questions asked to peers during their presentations.

Support

If you have questions:

- Ask at the end of lectures,
- Or contact me by email.

It is normal to refine your project idea as you go.

Use feedback from peers and from me to improve your work.

The aim is not perfection, but a clear, well-structured process.