

Group Project

Data, Data Storage, Data Collection

Overview

The group project is the main assignment of this course. You will work in groups of three students to complete a small end-to-end data workflow: from collection to cleaning, storage, analysis, and communication of results. At the end, you will present your findings to the class. This is your chance to bring together concepts from the course in a practical, open-ended task.

Objectives

- Apply the concepts of the **data lifecycle** in practice.
- Gain experience with real-world datasets and their challenges.
- Practice teamwork, documentation, and reproducibility.
- Communicate results effectively through reports and presentations.

What You Should Do

- **Collection:** select or acquire a dataset (public/open or self-built). Proprietary data cannot be used.
- **Cleaning:** handle missing values, errors, and inconsistencies.
- **Storage:** choose a format and explain your choice.
- **Analysis:** compute descriptive statistics and extract insights.
- **Communication:** produce visualizations and a short written report.

What You Should Not Do

- No machine learning models.
- No advanced statistics beyond the scope of the class.
- Do not aim for complexity: focus on clarity, reproducibility, and transparent decision-making.

Milestones

1. **30.09: Pitch session.** Each student pitches for 1 minute: dataset idea, main question, why it matters, and expected challenges. Groups will be finalized afterwards.
2. **14.10: One-page proposal.** Each group submits a short proposal listing group members, dataset, central question, and key lifecycle steps. This fixes the scope of your work and must be approved.
3. **03.12: Project report + code.** Submit a written report (up to 5 pages) together with a reproducible Jupyter notebook, or share a link to your code repository.
4. **09.12: Presentations.** Each group presents for 15 minutes, followed by questions.

Deliverables

Each group must submit:

1. A **Jupyter notebook** (clean, documented, runnable) covering collection, cleaning, storage, analysis, and communication. You can provide a link to a repository.
2. A **report** (up to 5 pages) including your research question(s), dataset documentation, results (findings and limitations), a reflection on the lifecycle (key methodological choices and justifications), and some comments on the lessons learnt.
3. A **presentation** during the final session (you need to send the slides prior to the presentation).

Grading Scheme (in Percent)

The project will be graded based on three main categories: lifecycle application (50%), deliverables (45%), and questions to your peers (5%). Some students who did an interesting pitch at the beginning of the lecture will also receive some bonus points.

Lifecycle Application

Each step of the data lifecycle is equally weighted (10% of the final grade). Besides the analysis step, every step has some bonus points if you have successfully used a difficult method or something not covered in the lecture.

1. **Collection:** adequacy of the dataset with the question, justification of choice, sourcing, and consideration of ethical aspects.
2. **Cleaning:** identification and handling of missing values, outliers, and other data issues, along with justification of choices.
3. **Storage:** choice of data storage format and model, documentation of metadata, and justification of decisions.
4. **Analysis:** correctness of analysis, insightfulness of results, transparency of the methods, and consistency with cleaning and storage.
5. **Communication:** clarity and adequacy of visualizations, integration of visuals with narrative, transparency about limitations.

Deliverables

Every deliverable is also equally weighted (15% of the final grade).

1. **Report:** structure, dataset documentation, results and insights, reflection on the data lifecycle, and overall writing quality.
2. **Code:** reproducibility, organization, documentation, consistency with the report, and code quality.
3. **Presentation:** content clarity, structure and timing, slide design, delivery, handling of questions.

Questions to Peers

Besides the quality of the work and deliverables, part of the grade (5%) is attributed based on the quality of the questions that you ask your colleagues (this part is graded individually).

Support

You can ask questions after lectures or contact me by email. It is normal to refine your project as you go. Use feedback from your peers and me to improve your work. The aim is not perfection, but a well-structured and transparent process.

Practical Notes

- All code must run without private APIs or closed data.
- Document clearly any preprocessing or external tools used.
- Respect ethical considerations: anonymize if needed, cite sources.