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.

Possible Themes

You may propose your own dataset (with my approval), or choose any open one:

- Health and fitness data (e.g. activity trackers, public health surveys).
- Environmental data (e.g. weather, pollution, energy usage).
- Social data (e.g. mobility, cultural trends, Wikipedia activity).
- Open government data (transport, education, finance).

Milestones

1. **30.09: Pitch session**. Each student pitches for 1 minutes: dataset idea, main question, why it matters, 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.
4. **08.12 or 09.12: Presentations**. Each group presents for 15 minutes, followed by questions.

Deliverables

Each group must submit:

1. A **Jupyter notebook** (cleaned, documented, runnable) covering collection, cleaning, storage, analysis, and communication.
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).

Evaluation

Your grade will consider:

- Understanding and application of the data lifecycle (technical correctness and completeness of the workflow).
- Clarity and reproducibility of the codebase (provide additionnal information about needed dependencies if any)
- Quality and readability of the report and presentation (communication is one step of the lifecycle)
- Creativity in dataset choice and approach.
- Teamwork and fair contribution.
- Quality of your initial pitch.
- Quality of the questions you ask to peers during their presentations.

Support

You can always ask questions after lectures or contact me by email for guidance. It is normal to refine your project idea as you go. Use feedback from me and your peers 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.