Group Project

Data, Data Storage, Data Collection

Overview

The group project is the main assignment of this course. You will work in groups of five students to design and implement a small, end-to-end data workflow to answer a concrete question. You will start from a question, determine what data is needed to answer it, acquire and prepare the data, choose how to store it, analyze it, and communicate your findings. The project is intentionally open-ended: you choose the topic, question, data sources, and technical approach. There is no single correct technical solution. What matters is that your choices are appropriate for the question, that you can justify them, and that you identify their limitations.

Project Requirements

Start from a Question

Your project must start from a well-defined data-driven question. The question should determine what data you need and what kind of analysis could provide an answer. Your question does not need to be scientifically novel. It should be specific enough that you can determine what evidence would support an answer.

Build a Reproducible Data Workflow

Starting with your question, build a workflow that turns raw data into evidence to answer it.

Collection: You must acquire or generate data as part of your project. Identify and justify your data source(s), document their provenance and the acquisition or generation procedure, and discuss relevant limitations and possible sources of bias. Possible approaches include using public APIs, building a simulation, scraping the web, or reusing public datasets. You are not required to use a particular acquisition method. However, **at least one source must involve an acquisition or generation process that you carry out and document**. Respect applicable ethical and legal constraints, and anonymize or otherwise protect personal data when necessary.

Cleaning: Assess the quality of your collected data and produce data suitable for your question. Identify relevant data-quality problems, explain how they affect your intended use, and apply appropriate cleaning or transformation operations. Important decisions should be justified, and the operations performed should be documented.

Storage: Make an explicit decision about how your data is stored. Choose an appropriate storage format, model, and, where appropriate, technology. Justify your choices based on the data and intended operations: consider relevant requirements such as structure, queries, and scale. Depending on the project, an appropriate solution may involve files, a relational database,

a NoSQL database, or a combination of storage solutions. Your justification should explain why the alternatives you considered were less appropriate. **A detailed comparison with at least one other solution is expected**.

Analysis: Use the data to provide an evidence-based answer to your original question. Select appropriate methods based on the question and the data, justify important analytical choices, and distinguish observations from interpretations. Discuss relevant limitations of the analysis and its conclusions. The answer may be inconclusive if the available evidence does not support a clear conclusion.

Communication: The goal of the project should be to propose an answer to the initial question. The presentation and the report should use appropriate visualizations to communicate your findings to someone who has not worked on the project.

Ensure Reproducibility

Reproducibility is a requirement across the whole project. Another student should be able to understand where the data came from, how it was acquired and transformed, how it was stored, and how the reported results were obtained. Document important assumptions, external resources, and technical choices. Keep the raw data, or provide clear instructions for obtaining it again. If any part of the workflow cannot be reproduced exactly, document the limitation and explain how you addressed it.

What You Should Not Do

The project is open-ended, but the following are *not* the intended approach:

- Start from an existing analysis-ready dataset and build a question around it without carrying out a meaningful acquisition step.
- Proprietary or inaccessible data cannot be used.
- Introduce artificial data-quality problems to exhibit complex cleaning.
- Choose a storage technology because it is technically complex.
- Use machine-learning models or advanced statistics without discussing it with me first, and be ready to justify why simpler methods are insufficient.
- Prioritize a flashy presentation over the quality and reproducibility of the underlying workflow.

Project Milestones

1. **16.09: Project proposal.** Each group submits a one-page proposal including the group composition, the main question, a description of the data they expect to need, and the expected challenges. The proposal fixes the scope of your work and must be approved.
2. **29.09: Raw data acquisition.** Submit your initial dataset together with the material to reproduce its acquisition and a short document justifying your choices. This is the first peer-reviewed checkpoint.
3. **06.10: Peer-review round 1.** Submit the reviews for the other groups.
4. **03.11: Intermediate version.** Submit the current version of your project, containing at least the cleaning part of the workflow, and ideally some comments about the storage process. Include a short document that presents the assessment of the raw data quality and a justification for the main cleaning procedures. This is the second peer-reviewed checkpoint.

5. **10.11: Peer-review round 2.** Submit the reviews for the other groups.
6. **27.11: Final deliverables.** Submit the written report together with the code of your project.
7. **02.12: Presentations.** Each group presents for 15 minutes, followed by questions. The presentation should also be submitted.

Final Deliverables

Each group will submit:

1. A **reproducible data workflow**: clean, documented, and runnable code covering collection, cleaning, storage, analysis, and communication. A Jupyter notebook or a runnable software project is acceptable.
2. A **report**: at most 5 pages describing the question, data, methodology, results, and limitations. The report should explain the important choices made throughout the data workflow.
3. A **presentation** the final sessions. The slides must be submitted at the latest by the end of the presentation.

Evaluation

The project will be evaluated on the appropriateness and justification of its choices, not on technical complexity for its own sake.

Dimension	What I look for
Collection	Appropriate data, justified acquisition, provenance, awareness of bias and limitations
Cleaning	Appropriate diagnosis and transformations, with justified decisions
Storage	Appropriate representation/storage solution and justification
Analysis	Appropriate analysis providing evidence relevant to the question
Communication	Clear presentation of findings, evidence, interpretation, and limitations
Reproducibility	A workflow that another student can understand and reproduce

The project will also be evaluated through the report, code/workflow, and presentation. Peer reviews will also contribute to the final grade as described below.

Peer Review and Questions to Peers

During the project, you will review other groups' work using a structured review form. The purpose of the peer review is not to assign a grade. Instead, you should assess whether the submitted work provides convincing evidence that the project requirements have been met. Your review should identify, in particular:

- something that is particularly well done;
- something that should be improved;
- a question or issue that remains unclear.

You will be asked to reflect on the feedback you received and explain which changes you made as a result.

Part of the grade is also 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.