Data Science - Home Assignment

Document Comparison System

As construction projects scale and more stakeholders are involved in the project, data is often produced by separate teams. Without a single source of truth or a systematic way to validate that information, minor discrepancies can cascade into costly conflicts, especially when drawings, schedules, and material specs must align perfectly across disciplines. For example, for subcontractors, checking submittals against specifications is critical to staying contract-compliant, avoiding delays, and protecting profit. Specs are legally binding, and installing non-compliant materials can lead to rejected work, extra costs, or liability. Verifying compliance early prevents rework and keeps approvals moving smoothly. It also opens the door for value-engineered substitutions that save money while meeting requirements. Ultimately, accurate submittals build trust with the GC and design team, boosting your reputation and chances for future work.

Your task is to design and implement a generic comparison engine system that helps verify whether two documents that should contain equivalent information actually do.

Task

I/O

- **Input**: Two files uploaded. Start by supporting CSV and PNG (you may add others if you wish). You can add additional input values if it will improve results (explain what and why)
- **Output**: A JSON report containing:
  - Matched records: List of information/fields that appear in both documents
  - Field-level differences (old, new, delta, and a boolean *is_within_tolerance*)
  - Overall pass/fail flag

Development Guidelines

- **Language**: Python
- **Libraries and API:** any of your choices (feel free to request API key for one of the services)
- **Correctness Evaluation:** Demonstrate *how you know* your comparison is correct. Provide at least one clear automatic validation strategy
- **Storage**: Can be in-memory dict, file system, or any of your choices
- **Optional**: Docker / lightweight UI to preview results

Examples: **Data Science - Home Assignment**

Example 1

Verify that every room that appears in both documents the area difference is ≤ 1 m².
floor_plan_example.png vs floorplan_takeoff.csv

Example 2

Compare the cost items and bar chart of planned vs actual totals.
construction_budget_summary.png vs construction_budget.csv:

Deliverables

- Code repository (GitHub or zip)
- Clear report or README with instructions and sample comparison results.
- (Optional) Architecture diagram or design notes

There is no "right answer"

The goal is to see how you structure the problem, apply data processing skills, and design for flexibility. There is also no "right" approach, nor is a perfect solution expected (especially not in 2 hours).

Good luck! 🙂