Project Phase #1

EAS 587

Issued: February 2, 2026
GitHub Repo Invite Due: February 9, 2026
Workshop: February 20, 2026
Due: February 27, 2026 @ 11:59PM

Content Covered
Problem Statements, Data Acquisition, Data Processing, Data Cleaning, EDA

---

1 Project Overview

The course project forms the hands-on practical learning component of the course, and will have students putting into practice each step of the data science pipeline (depicted in Figure 1, adapted from [1]). The project is broken into four phases, with Phase 1 covering the first five steps of the pipeline. Your project should be motivated by an issue in an application domain of your interest, and your work in Phase 1 will establish the foundation for all later phases.

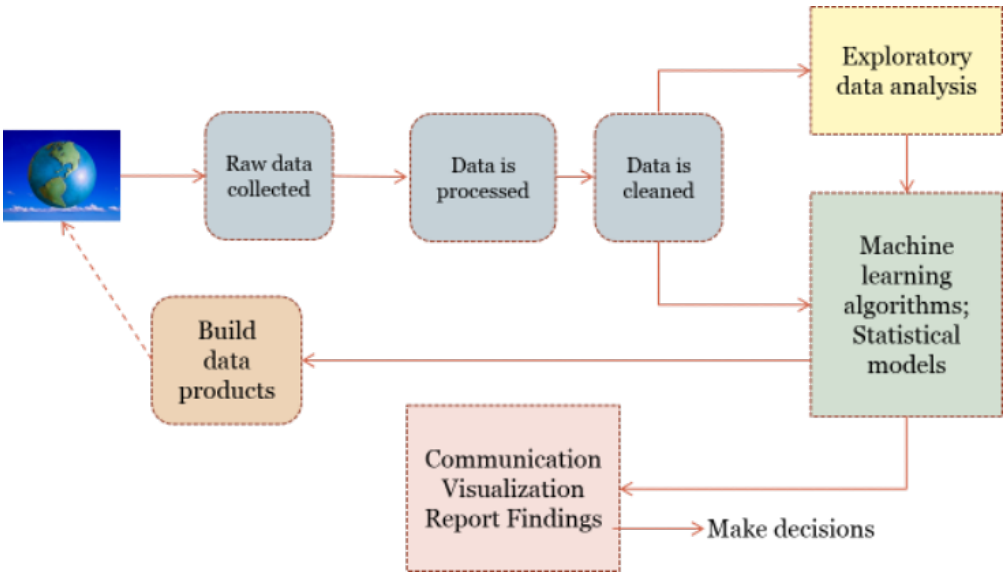

Figure 1: The Data Science Pipeline

Learning Outcomes for Phase 1:

1. Identify problems prevalent in public application domains. (Task 1)
2. Research and identify structured datasets to address the problem and collect the relevant data. (Task 2)
3. Clean and provision the data for downstream exploration and analytics. (Tasks 3–4)
4. Understand the basic characteristics of the data by performing John Tukey's exploratory data analysis (EDA) [2]. (Task 5)

Description

Data collection is a critical phase of the data-science process. Many organizations—including federal agencies (e.g., data.gov, Pew Research), social networks (e.g., Twitter, Facebook), and commercial platforms (e.g., Amazon, NYTimes)—make subsets of their data available to the public. Some data is downloadable as files (CSV, XLSX), some as databases (DB, DB3), and some must be scraped from web pages.

You will research data-generating organizations, select a domain and dataset(s), and define the questions you aim to answer. For this project, you will work with structured data.

General Project Requirements

1. **Work Environment:** Python is required. You may use Jupyter, IPython, or any Python environment.
2. **Programming:** Prepare to write and debug Python code using course materials and online resources.
3. **Academic Integrity:** Violations result in an automatic F. See the syllabus for details.
4. **Project Phases:** Each phase builds on the last. You must complete Phase 1 before beginning Phase 2. During Phase 1 you may change your problem or dataset; after Phase 1, changes are not allowed.
5. **Teams:** Work in groups of two or three. Register your team on Piazza before requesting project guidance. One submission per team.

Submission Requirements

1. **Deadlines:** Due February 27, 2026 at 11:59 PM. One-day late submissions incur a 15% penalty. No submissions accepted after one day late.
2. You must submit Phase 1 to begin Phase 2.
3. **GitHub Repository:** Each team must maintain a **private** GitHub repository. Invite the instructor, grader, and TA by February 9. Repositories must show sustained, individual contribution: at least 20 commits total, 5 per member, across at least 5 days. Commit messages must be meaningful. Commits exceeding 500 lines changed may be flagged.

4. **Workshop Slides (New Requirement):** A short slide deck (1–3 slides) will be required for the February 20 workshop. **Content details will be provided later.** These slides are worth **2 marks** and are part of your Phase 1 deliverables.

5. **Submission:** Submit a `.txt` file to UBLearns containing:

   - URL to your private GitHub repository
   - Link to your Google Doc report

   Your Google Doc must be shared with instructors (Commenter access). Version History will be reviewed. Bulk-paste patterns may be flagged. Your report must include a **Design Decisions** section documenting key choices.

   Failure to follow submission guidelines results in a 3-mark penalty.

Required Repository Structure

```
project-repo/
|-- README.md
|-- requirements.txt
|-- data/
|   |-- raw/
|   +-- processed/
+-- src/
    |-- data_collection.py    (illustrative)
    |-- data_cleaning.py      (illustrative)
    +-- eda.py                (illustrative)
```

File/Directory	Description
`README.md`	Overview, setup instructions, and how to run the code. Must run in a fresh environment. **Colab notebooks are not accepted**.
`requirements.txt`	All Python dependencies. Install via `pip install -r requirements.txt`.
`data/raw/`	Original, unmodified data files.
`data/processed/`	Cleaned and processed data files.
`src/`	Python source code. Filenames above are illustrative; organize as appropriate.

Note: If data files exceed 50MB, provide download instructions instead of committing them.

2 Report Deliverables [75 marks]

2.1 Workshop Slides [2 marks]

A short slide deck (1–3 slides) supporting your project will be required for the February 20 workshop. **Content details will be provided later.**

2.2 Problem Statement

Form a clear title and problem statement. Include:

b. **[8 marks] Problem Context & Significance**

- Background motivating your objectives
- At least **3 real-world examples** where the problem impacts stakeholders
- Quantification of significance (cost, people affected, frequency, etc.)

b. **[7 marks] Project Impact & Use Cases**

- At least **4 distinct use cases**
- Target user group and benefit for each
- How your approach differs from existing solutions
- Scalability and broader applicability

2.3 Data Sources [8 marks]

Your dataset(s) must contain at least **50,000 rows**. You may use multiple sources (e.g., medical, financial, sports, Kaggle, Amazon reviews, Twitter, YouTube, Reddit). Your data must contain enough columns to support cleaning and EDA requirements. If your data is already clean or too narrow, you must find additional sources.

> **Note:** Examples of acceptable datasets can be found in the UBlearns entry showing datasets used in last semester's projects.

You must cite and link all data sources.

2.4 Data Cleaning/Processing [20 marks]

Document at least **10** distinct cleaning/processing operations. Explain why each was necessary and how it was implemented.

Verification Requirement: Run your cleaning pipeline in a fresh environment (e.g., teammate's machine) to confirm reproducibility. Include a brief statement confirming this.

2.5 Dead Ends [5 marks]

Document at least three approaches that did not work and what you learned from each.

2.6 Exploratory Data Analysis (EDA) [20 marks]

Perform EDA following the principles of John Tukey [2]. Demonstrate at least **10** significant and relevant EDA operations. Include figures and tables with proper labels. Explain how EDA informed your cleaning and feature decisions.

2.7 Surprise Findings [5 marks]

Describe what surprised you in the data and how it changed your approach. Emphasize domain-specific reasoning.

3 Code Deliverables [25 marks]

- **Code Organization & Readability (10 marks):** Clear structure, meaningful names, consistent style, appropriate comments.
- **README Quality (5 marks):** A grader should be able to reproduce your results in under 10 minutes (excluding processing time).
- **Git Practices (5 marks):** Meets commit requirements.
- **Dependencies & Reproducibility (5 marks):** Code runs in a fresh environment using `requirements.txt`.

4 Deliverables Summary

Deliverable	**Marks**
Workshop Slides	2
Problem Statement (Context + Use Cases)	15
Data Sources	8
Data Cleaning/Processing	20
Dead Ends	5
Exploratory Data Analysis (EDA)	20
Surprise Findings	5
Report Subtotal	**75**
Code Deliverables	25
Total	**100**

Appendix A: Required Report Sections

Your Google Doc report must include the following sections in this order:

1. **Title and Team Members**
2. **Workshop Slides (attached or linked)**
3. **Problem Statement**
    - Problem Context & Significance
    - Project Impact & Use Cases
4. **Data Sources**
5. **Data Cleaning/Processing**
6. **Exploratory Data Analysis (EDA)**
7. **Dead Ends**
8. **Surprise Findings**
9. **Design Decisions**
10. **References**

References

[1] Adapted from standard data science pipeline diagrams. [2] John W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.