





Final Project – Principles and Techniques in Data Science (2025)

From Raw Data to Insight: A Full Data Science Pipeline in the Wild

Project Objective

- Students will **design and implement a full data science pipeline** on a **self-collected dataset**, with a focus on understanding, creativity, and model interpretability. The goal is to explore and reason through real-world data by applying techniques taught in the course, not just running code.
- Each project must be completed and presented by a **group of 2, 3, or 4** students — **no more, no less**.
- Each **student is required to present exactly 10 minutes during the oral defence:**

 2 students → 20 minutes total (10 minutes each)

 3 students → 30 minutes total (10 minutes each)

 4 students → 40 minutes total (10 minutes each)


- Groups exceeding or falling short of the allotted time per student will receive a **point deduction**.

Individual Responsibility in Group Presentations

While the project is completed in a group, **each student is expected to be fully familiar with all aspects of the project** — not just their individual part.

During the oral defence, each presenter should:

- **Demonstrate a clear understanding of the entire pipeline:** from data collection to modelling and evaluation.
- **Clearly present and explain their own contributions** within the project.

 The oral presentation may include **questions about any part of the project**, and may be **directed to any group member** — all group members should be equally prepared.



Submission Deadline

All project deliverables — including the **executed Jupyter Notebook, dataset file(s), and presentation slides** — must be submitted **at least 24 hours before your scheduled oral defence**.

There will be 2 dates for defence – July 31, 2025 and August 27, 2025.

Failure to submit on time may result in a deduction of points or disqualification from presenting.

Data Collection Requirements

- You must **scrape your own dataset** from **at least one** web source (e.g., public APIs, HTML pages, RSS feeds, public transcripts, or user-generated content).
- **No pre-curated datasets** (e.g., from Kaggle or similar) are allowed.
- Final dataset must include **at least 10,000 records** and should be **non-trivial** (contain sequences, time-based info, or structured text/numerical data).
- Include **documentation of your scraping process**, including tools, challenges, and any filtering or transformation applied.

Important Note on Data Approval

Before diving deep into data collection and analysis, it is **strongly recommended** that your group **submit a short description of your proposed dataset and project idea** for approval. This will ensure that:

- The dataset is rich and complex enough for the project goals
 - The idea is aligned with course expectations
 - You avoid wasted effort on unsuitable or overly simple data
-

Pipeline Requirements

You must include and explain the following steps in your project:

1. Preprocessing & Exploratory Data Analysis

- Data cleaning and transformation.
- Exploratory visualizations to understand key patterns.



- Justify your decisions: Why was a certain transformation or representation chosen?

2. Clustering or Anomaly Detection

Choose one and implement using at least **3 techniques**, **one of which must be hierarchical**:

- If clustering: Compare how different methods (e.g., DBSCAN, Agglomerative, K-Means) discover structure in your data.
- If anomaly detection: Identify and analyze outliers using both statistical and machine learning techniques.

3. Time Series Forecasting or Recommendation System

Choose **one** of the two and clearly define your problem:

- **Time Series**: Apply ARIMA/SARIMAX or similar to forecast a trend (e.g., user activity, event frequency).
- **Recommendation**: Build a rule-based or collaborative filtering system using association rules or similarity metrics.

4. Model Explainability and Fairness

- Train at least one **interpretable model** (e.g., decision trees, logistic regression).
- Apply **SHAP/LIME** or similar to **interpret feature importance and predictions**.
- Discuss **fairness or bias considerations** – does the data or model show any systematic skew? How would you mitigate it?

5. Model Optimization and Evaluation

- Perform **hyperparameter tuning** (e.g., Grid Search or Bayesian Optimization).
- Evaluate performance using **appropriate metrics** for clustering, forecasting, or recommendation.
- Include **visual and statistical comparisons** across methods.

Deliverables

1. Data –

- Your final dataset(S) as CSV, JSON, or other readable formats.



- Include a brief description of the fields and how data was cleaned / transformed.

2. Jupyter Notebook

- Must include **code + markdown cells + output (executed cells)** with thoughtful explanations.
- Clearly structured sections with interpretation of results.

3. 15-Minute Oral Presentation ("Project Defense")

- Present your project goals, scraping process, methodology, key insights, and takeaways.
- Use **slides with visualizations**, not just raw results. Focus on reasoning, decisions, and interpretation.

Grading Criteria

| Component | Details | Weight |
|-------------------------------------|---|--------|
| Data Originality & Scraping | Quality and complexity of data collection. Creativity, completeness, and documentation of scraping process. | 15% |
| Preprocessing & EDA | Thoughtfulness of data cleaning, transformations, and initial visual insights. | 15% |
| Clustering or Anomaly Detection | Correct application of 3 methods (including 1 hierarchical), comparison and interpretation. | 15% |
| Time Series / Recommendation Module | Proper implementation of one of the two tasks, with evaluation. | 15% |
| Explainability & Fairness | Application of SHAP/LIME, discussion of biases, and fairness awareness. | 15% |
| Model Tuning & Evaluation | Hyperparameter tuning, model comparison, appropriate metric use. | 15% |
| Oral Presentation | Clarity, insights, visual delivery, understanding of the project during oral defense. | 10% |



💡 **Need Inspiration?**

You are free to explore **any domain**, such as:

- User behavior on social platforms
- Activity logs or browsing patterns
- Product reviews, news trends, or discourse analysis
- IoT sensor logs, public transport data, or service transactions

What matters most is your **critical thinking, creativity, and reasoning** in how you handle data and extract insight.

Good luck!

Raz and Yarden.

Appendix: Suggested timeline checkpoints:

Week 1: Project kickoff, form topic ideas, review scraping tools.

Week 2–3: Complete data scraping (10,000+ records), submit 1-paragraph dataset description.

Week 4–5: Preprocessing + EDA; submit EDA summary with key visualizations.

Week 6–7: Clustering or anomaly detection analysis – submit initial clustering findings.

Week 8–9: Forecasting or recommendation task implementation.

Week 10: Explainability & fairness section draft (SHAP, bias discussion).

Week 11: Optimization, hyperparameter tuning, model evaluation.

Week 12: Prepare Jupyter Notebook & finalize visualizations.

Week 13: Oral presentation (15 minutes) + Q&A.