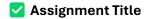
# Data Analysis and Visualization

Prof. Antonio Liotta, Academic year 2025-26



"From Raw Data to Features: Tackling Non-Tabular Data Challenges in a Data Science Pipeline"

Assignment Specification

#### **Objective**

Students will work in groups to create a **20-minute lecture** and a **Jupyter Notebook demo** that explain and demonstrate how to process **non-tabular data** (text, graphs, images, or audio) from **data ingestion** to **feature extraction** and preparation for a **simple classification task**.

#### Scope

Each group will:

#### 1. Start from the Problem

- o Define a real-world classification problem for their data type.
- Explain why this problem matters.

#### 2. Identify Challenges

- Discuss typical issues for their data type:
  - Data quality (missing values, noise, imbalance).
  - Complexity of feature extraction.

#### 3. Introduce Relevant Methodologies

- Data quality verification techniques.
- Cleaning and transformation strategies.
- Feature extraction methods specific to their data type.
- o Explain why certain features are valuable for this type of data.

#### 4. Practical Aspects

- $\circ$  Which **Python libraries** are available for each step (ingestion → cleaning → feature extraction → EDA → feature table).
- Which sample datasets are available (Kaggle, Hugging Face, Open Data).

#### 5. **Demo**

- Jupyter Notebook showing:
  - Data ingestion.
  - Cleaning and verification.
  - Feature extraction.
  - EDA (exploratory data analysis).
  - Building a feature table (CSV/Excel).
  - Simple classifier which completes the data science pipeline (low emphasis on evaluating the classifier)

#### 6. **Group Organization**

- Assign roles:
  - Project Coordinator (manages timeline and integration).
  - Data Cleaning Team.
  - Feature Engineering Team.
  - **Demo Team** (aligns notebook with lecture).
  - Presenters.
- o Decide who will present and how tasks will be divided.

#### **Deliverables**

- PowerPoint lecture (fundamentals + workflow + challenges + methodologies).
- Jupyter Notebook demo (aligned with lecture).
- Discussion question for peers.
- File: group\_contributions\_table.xlsx (filled with group member and their contributions)

#### **Presentation Format**

• 20 min lecture + 10 min discussion.

#### Deadline

• One week from assignment date.

### **☑** Suggested Lecture Template (may be customised by the group)

- 1. Problem Definition & Importance.
- 2. Challenges for this data type.
- 3. Methodologies for cleaning, verification, transformation.
- 4. Valuable features for this data type.
- 5. Python libraries for each step.
- 6. Example datasets.
- 7. Workflow diagram (ingestion → cleaning → features → EDA → feature table).
- 8. Demo overview.
- 9. Discussion question.
- 10. References.

## Grading Rubric (Total: 100 points)

Criteria	Points
Problem Definition & Relevance	10
Clearly explains the real-world problem and why it matters.	
Challenges Analysis	15
Identifies key data-specific issues (quality, complexity).	
Methodologies & Theory	20
Introduces relevant techniques for cleaning, verification, feature extraction.	
Feature Explanation	15
Explains why selected features are valuable for this data type.	
Practical Tools & Datasets	10
Lists appropriate Python libraries and datasets.	
Jupyter Notebook Demo	20
Demonstrates ingestion, cleaning, feature extraction, EDA, and feature table creation	
Presentation Quality	10
Clear, structured slides; effective communication; discussion question.	