Homework 1: Data Collection and Analysis using REST APIs

Proff. Alessio Martino & Lorenzo Ariemma & Mattia Cervellini

December 22, 2025

Objective

The goal of this assignment is to give hands-on experience with:

- Accessing real-world data using REST APIs
- Parsing and cleaning data
- Integrating heterogeneous datasets
- Exploring data using data science techniques
- Applying basic machine learning models to real data

Students will work with environmental, weather, and urban mobility data collected from public APIs.

Data Sources

Each student must use **at least two** of the following REST APIs[1].

Air Quality Data (OpenAQ)

- API endpoint:

  `https://docs.openaq.org/resources/measurements`

- Data includes: PM2.5, PM10, NO_2, O_3, CO
- Filters: city, date range, pollutant
- A free API key is required

[1] If the student prefers to use other APIs for the same data, ask the professors.

Weather Data (OpenWeather)

- API endpoint:

  `https://openweathermap.org/api/one-call-3`

- Data includes: temperature, humidity, wind speed, precipitation
- A free API key is required

Urban Mobility / Traffic Data

Choose one official open data portal providing REST access, for example:

- City open data portals (e.g., New York City, London, Rome, Milan)
- Public transportation or traffic flow datasets

1 Task 1: Data Collection via REST APIs

1. Select one city and a time period of at least 6 months.
2. Query the selected APIs programmatically using Python.
3. Save the API responses locally as `.json` or `.csv` files.

Important: Once the data has been downloaded and stored locally, all subsequent processing must use the local files. Repeated API queries should be avoided.

2 Task 2: Data Cleaning and Integration

Using Python and suitable libraries (e.g., Pandas, Scikit):

1. Parse all files into DataFrames.
2. If necessary, clean each DataFrame for empty values or useless columns.
3. Merge the DataFrames into a single unified table using proper matching between DataFrames.

The final DataFrames should include:

- Temporal variables (date, hour, day of week)
- Environmental variables (air quality and weather)
- Mobility or traffic-related variables

3 Task 3: Exploratory Data Analysis

Perform an exploratory data analysis including:

- Summary statistics
- Time series plots
- Correlation analysis

Examples of questions to investigate:

- Is traffic volume correlated with pollution levels?
- How do weather conditions influence air quality?
- Are there weekly or seasonal patterns in the data?

Important: the plots should output in an elegant way, with titles, axes titles, and correct boundaries.

4 Task 4: Machine Learning

Implement **at least one** of the following machine learning tasks.

Option A: Regression

Predict a continuous air quality variable (e.g., PM2.5) using:

- Weather variables
- Mobility or traffic variables

Possible models include linear regression, random forests, or gradient boosting.

Option B: Classification

Discretize air quality into categories (e.g., Good / Moderate / Poor) and predict the category using the available features.

Option C: Time Series Forecasting

Forecast future pollution levels using historical data. External variables such as weather or traffic may be included.

5 Task 5: Evaluation and Interpretation

1. Split the data into training and testing sets.
2. Evaluate the model using appropriate metrics (e.g., RMSE, MAE, accuracy, F1-score).
3. Interpret the results and discuss:
   - Model performance
   - Feature importance or coefficients
   - Limitations of the analysis

6 Deliverables

Students are required to submit a single Jupyter Notebook (.ipynb) as the final outcome of their work. The notebook must serve both as a code submission and as a written report, making appropriate use of Markdown cells.

Throughout the notebook, reflect on the challenges encountered at each stage of the work and discuss possible improvements or additional steps that could enhance the approach or results.

Code and narrative should be interleaved to clearly justify the methodological choices made during the assignment and to explain and interpret each of the obtained results.