CE 49X: Introduction to Computational Thinking and Data Science

Lab 02: Soil Test Data Analysis

Dr. Eyuphan Koc
Department of Civil Engineering
Boğaziçi University

October 10, 2025

1 Objective

In this lab, you will:

- Reinforce Python fundamentals by working with a small soil test dataset
- Practice reading, cleaning, and transforming data from a CSV file
- Compute and display descriptive statistics (minimum, maximum, mean, median, and standard deviation) for a numeric column
- Organize your code into functions and add appropriate error handling
- Notify the instructor upon completion

2 Overview

Building on Lab 1, this lab focuses on data cleaning and transformation. You will work with a soil test dataset containing values like soil pH, nitrogen, phosphorus, and moisture. Your task is to load the data, handle missing values, optionally remove outliers, and compute basic descriptive statistics.

3 Prerequisites

Before starting, ensure that you have:

- Completed Lab 1 and set up your Python environment
- Python 3.10 or later installed
- Required Python libraries:

```
pip install pandas numpy
```

- A text editor or IDE for Python development

4 The Dataset

The file `soil_test.csv` is located in the `/lab02/` directory. A sample of its content is shown below:

```
sample_id,soil_ph,nitrogen,phosphorus,moisture
1,6.5,20,15,30
2,7.0,25,20,35
3,5.8,18,NaN,32
4,6.2,22,17,28
5,6.8,NaN,16,33
```

Note: "NaN" indicates missing values.

5 Instructions

5.1 Create the Python Script

Complete the tasks on the file named `lab02_soil_analysis.py`.

5.2 Script Requirements

Your script should include the following functionality:

5.2.1 Load the Dataset

- Use Pandas to load `soil_test.csv`
- Gracefully handle the case when the file is not found

5.2.2 Data Cleaning

- Handle missing values (e.g., fill with the column mean or drop the rows)
- *(Optional)* Remove outliers for a chosen column (e.g., soil_ph values more than 3 standard deviations away from the mean)

5.2.3 Compute Descriptive Statistics

For at least one numeric column (e.g., `soil_ph`), compute:

- Minimum
- Maximum
- Mean
- Median
- Standard Deviation

Print these statistics in a clear, formatted manner.

5.2.4 Modular Code

- Organize your code into functions (e.g., `load_data()`, `clean_data()`, `compute_statistics()`)
- Add comments and docstrings for clarity

5.2.5 Error Handling

- Use try/except blocks to manage potential errors during file I/O and data processing

6 Running Your Script

From your working directory, run your script by executing:

```
python lab02_soil_analysis.py
```

Ensure that the script runs without errors and displays the computed statistics.

7 Submission Instructions

Due Date: October 16, 2025

When you have completed lab 02, follow these steps:

1. Complete all TODO items in your Python script
2. Test your script to ensure it runs without errors
3. Answer the reflection questions in comments
4. Save your file as `lab02_soil_analysis.py`
5. Upload your Python file to Moodle before the due date

7.1 Submission Checklist

Before submitting, make sure:

- ☐ Your script runs without errors and shows outputs
- ☐ All TODO items have been completed
- ☐ Code includes comments explaining your logic
- ☐ Output is formatted clearly with appropriate decimal places
- ☐ Reflection questions are answered in comments
- ☐ File is saved as `lab02_soil_analysis.py`
- ☐ File is uploaded to Moodle before the lecture on October 16, 2025

8 Additional Resources

- **Pandas Documentation:** `https://pandas.pydata.org/docs/`
- **Numpy Documentation:** `https://numpy.org/doc/`
- **Git and GitHub Guides:** `https://guides.github.com/`

Good luck with lab 02! If you have any questions or encounter issues, please post on the class discussion board or reach out to the instructor.