

Data Mining

ENSIA 2025-2026

Lab sheet N°1: Introduction to Numpy and Pandas



Objectives

- Introduction to data manipulation
- Create and manipulate one-dimensional and two-dimensional **Numpy** arrays
- Create and manipulate **Pandas** Series and Dataframes
- Describe how to index and "type" **Pandas** Series and Dataframes
- Transform data into a format suitable for analysis using **NumPy** and **Pandas**.

Required tools

Programming language: Python 3

Platforms: Anaconda, Jupyter Notebook, JupyterLab, Google Colab (cloud-based environment)

Python libraries:

- **Numpy:** A library for efficient numerical operations and multidimensional arrays, widely used in scientific computing and data analysis
- **Pandas:** A data manipulation and analysis library, providing data structures and functions to easily handle and process structured data.

Resources

Python:

- **Reference card:** [Python Cheatsheet](#)
- **Official documentation:** [Python Docs](#)
- **Official tutorial:** [The Python Tutorial](#)

Numpy:

- **Reference card:** [Cheat sheet Numpy Python copy.indd](#)
- **Official documentation:** [NumPy Documentation](#)
- **Tutorial (guided Notebook):** [numpy_tutorial.ipynb](#)
- **Exercise (non-guided Notebook):** [exercise_numpy.ipynb](#)

Pandas:

- **Official reference card:** [Data Wrangling - with pandas Cheat Sheet http://pandas.pydata.org](#)
- **Official documentation:** [pandas 2.3.2 documentation](#)
- **Tutorial (guided Notebook):** [pandas_tutorial.ipynb](#)
- **Exercise (non-guided Notebook):** [exercise_pandas.ipynb](#)

Part 1: Setting up the environment (10 minutes)

- Install all the required tools (Python, Anaconda, Jupyter Notebook, and JupyterLab)
- Create a virtual environment named "**DM_ENV**" for the Data mining course (see end of page)
- Activate the environment and install the required Python libraries: **Pandas** and **Numpy**
- Alternatively, students who find problems setting up the environment can use Google Colab

Part 2: Hands-on Data Analysis with Numpy (70 minutes)

- Execute and understand the **guided** Jupyter Notebook file, in local or on Google Colab
- Fill in the gaps and write the missing code in the **non-guided** Jupyter Notebook file
- Take a look at the provided resources (documentation, tutorial, reference card) for more info

Part 3: Hands-on Data Analysis with Pandas (70 minutes)

- Execute and understand the **guided** Jupyter Notebook file, in local or on Google Colab
- Fill in the gaps and write the missing code in the **non-guided** Jupyter Notebook file
- Take a look at the provided resources (documentation, tutorial, reference card) for more info

Create a virtual environment: To create a virtual environment named "**DM_ENV**" in Anaconda:

1. Open a terminal window.
2. Activate the Anaconda environment if it's not already activated.

```
conda activate anaconda3
```

3. Create the virtual environment

```
conda create --name DM_ENV python=3.12
```

4. Activate the virtual environment

```
conda activate DM_ENV
```

5. Install Jupyter notebook and IPython kernel

```
conda install notebook ipykernel
```

6. Install the IPython kernel for Jupyter with the environment name

```
ipython kernel install --user --name=DM_ENV
```

7. Check that the virtual environment is activated: the prompt should start with DM_ENV

8. Install packages in the virtual environment using the conda install command.

```
conda install numpy pandas
```

9. To deactivate the virtual environment, run the following command:

```
conda deactivate
```

You can now switch back and forth between the virtual environment and your base Anaconda environment by activating and deactivating the virtual environment.

More details about environments can be found in [Getting started with conda](#).