Data Mining

ENSIA 2025-2026

Lab sheet N°3: Exploratory Data Analysis (EDA)

Objectives

- Introduction to Exploratory Data Analysis (EDA) techniques (statistical and graphical)
- Understand the data and summarize its key statistical properties
- Comprehend the distribution and dispersion of the data
- Analysis of relationships between attributes
- Choose/apply statistical methods for a given dataset or attributes using **NumPy** and **Pandas**
- Choose/apply graphical methods for a given dataset or attributes using **Matplotlib** and **Seaborn**
- Decide which data preprocessing method to use following the EDA phase

Part 1: Exercise on the Chapter Introduction (30 minutes)

Think of possible useful data mining tasks (classification, clustering, association rules, anomaly detection, etc.) for the three datasets of the previous lab: Breast cancer, Airbnb, and Daily precipitation.

Part 2: Exercises on the Chapter Data - Part 1 (60 minutes)

1. Classify the following attributes as binary, discrete, or continuous. Also, classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio).

Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. Answer: discrete, quantitative, ratio

A. Time in terms of AM or PM.
B. Brightness, as measured by a light meter.
C. Brightness, as measured by people's judgments.
D. Angles, as measured in degrees between 0 and 360.
E. Bronze, Silver, and Gold medals, as awarded at the Olympics.
F. Height above sea level.
G. Number of patients in a hospital.
H. ISBNs for books. (Check the format here: ISBN Format)
I. Ability to pass light in terms of the following values: opaque, translucent, transparent.
J. Military rank.
K. Distance from the center of campus.

2. Calculate the indicated similarity or distance measures for the following vectors, X and Y.

A.	x = (1, 1, 1, 1),	y = (2, 2, 2, 2)	cosine, correlation, Euclidean
B.	x = (0, 1, 0, 1),	y = (1, 0, 1, 0)	cosine, correlation, Euclidean, Jaccard
C.	x = (0, − 1, 0, 1),	y = (1, 0, − 1, 0)	cosine, correlation, Euclidean

D. x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard

E. x = (2, − 1, 0, 2, 0, − 3), y = (− 1, 1, − 1, 0, 0, − 1) cosine, correlation

- For each distance/similarity metric, say if the vectors X and Y are objects (rows) or attributes (columns)

- For case (A), why is the correlation undefined between the two vectors X and Y?

- When should we use cosine similarity instead of correlation and vice versa?

Part 3: Exploratory Data Analysis with Matplotlib and Seaborn (90 minutes)

- Activate **DM_ENV** and install the required Python libraries: **Matplotlib** and **Seaborn**
- Alternatively, students who find problems setting up the environment can use Google Colab
- 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

Required tools

Programming language: Python 3

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

Python libraries:

- **Matplotlib:** A plotting library, used for creating static, animated, and interactive visualizations.
- **Seaborn:** A data visualization library based on Matplotlib, providing high-level functions for creating attractive statistical graphics.

Resources

Matplotlib:

- **Official Reference cards:** Matplotlib cheatsheets — Visualization with Python
- **Official documentation:** Matplotlib 3.10.7 documentation
- **Tutorial:** Plot types — Matplotlib 3.10.7 documentation

Seaborn:

- **Reference card:** Cheat sheet Seaborn.indd
- **Official documentation:** Seaborn
- **Tutorial:** User guide and tutorial — seaborn 0.13.2 documentation

Criteria to choose a statistical/visualization technique:

- The type of analysis (comparison, relationship, composition, distribution, dispersion, etc.)
- The number of attributes (one, two, multiple attributes, etc.)
- The types of attributes (categorical, numerical, etc.)

  The summary statistics: Summary statistics.pdf

  The chart chooser: The chart chooser.pdf

Notebooks

- **Guided Notebook:** Data_visualization (Guided).ipynb

- **Non-guided Notebook:** Data_visualization (Non-Guided).ipynb