

Importing, summarizing and visualizing data

Jordi Villà i Freixa

Universitat de Vic - Universitat Central de Catalunya
Study Abroad

jordi.villa@uvic.cat

course 2023-2024

- 1 Python
- 2 Dealing with data
- 3 Bibliography

Introduction to the course

The material in these slides is strongly based on [1]. When other materials are going to be used, they will be cited accordingly.

Python as the programming language to learn

- Easy to learn and powerful
- High-level efficient data structures
- Effective approach to object-oriented programming
- Interpreted
- Elegant syntax and **dynamic typing**

The choice for rapid application development in many areas on most platforms:

- Official Python tutorial
- Interactive Python tutorial at [LearnPython.org](https://learnpython.org)

Learning Python through Jupyter notebook and JupyterLab

<https://jupyter.org>

- Jupyter Notebook:
 - web application
 - creating and sharing computational documents
 - several programming languages, including Python
 - interactive output
- JupyterLab
 - interactive development environment (IDE) for notebooks, code and data
 - flexible interface
 - modularity

Installing Jupyter

```
conda install jupyter  
jupyter notebook
```

Code 1: installing and executing Jupyter notebook

How is data stored?

- Data can be thought of as being the result of some random experiment.
- We are interested in analysing such data.
- The format is typically a set of variables or *features* as **columns** while the different items are given as **rows**.
- Typically the first columns represents a unique identifier or index.
- Some columns refer to the experimental settings and others are real variables.
- Many times variables and experimental designs are stored in two different files. The we call the experimental desgins file as the **metadata** file, describing the details of the different experiments (or columns).

Data vs Metadata

NAME	AGE	GENDER	HEIGHT (CM)
A	20	MALE	172
B	21	MALE	168
C	19	FEMALE	160
D	20	MALE	163

Diagram illustrating the distinction between Metadata and Data:

- The first row (headers) is labeled **METADATA**.
- The subsequent rows (records) are labeled **DATA**.

Training datasets

There exist several datasets repositories that one can use to test the methods that are being developed. Some of them are going to be used in this course:

- Machine Learning Repository at University of California (<https://archive.ics.uci.edu>)
- Vincent Arel-Bundock repository (<https://vincentarelbundock.github.io/Rdatasets/>)
- Data from Pierre Lafaye de Micheaux and collaborators in their book "The R Software. Fundamentals of Programming and Statistical Analysis" [2]

EX 1 I

Exercise 1 Data visualization

Import the *EuStockMarkets* dataset from the Vincent Arel-Bundock repository. The data set contains the daily closing prices of four European stock indices during the 1990s, for 260 working days per year.

- 1 Create a vector of times (working days) for the stock prices, between 1991.496 and 1998.646 with increments of 1/260.
- 2 Reproduce Figure 1.10. [Hint: Use a dictionary to map column names (stock indices) to colors.]

EX 1 II

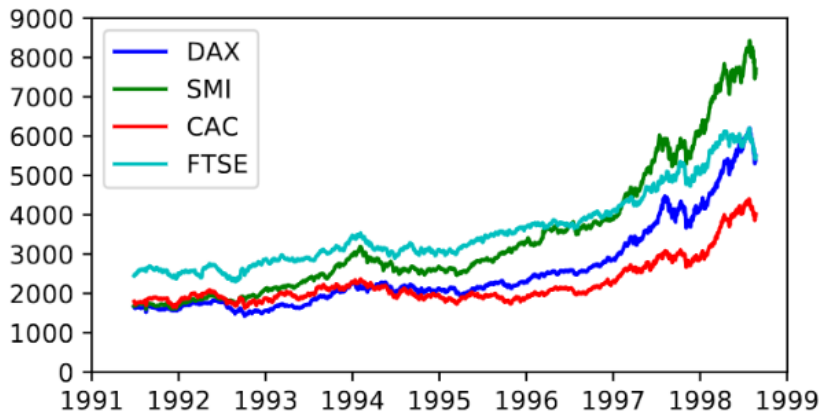


Figure 1.10: Stocks



Dirk P. Kroese, Zdravko Botev, Thomas Taimre, and Radislav Vaisman.

Data Science and Machine Learning: Mathematical and Statistical Methods.

Machine Learning & Pattern Recognition. Chapman & Hall/CRC, 2020.



Pierre Lafaye De Micheaux, Rémy Drouilhet, and Benoit Liquet.

The R Software: Fundamentals of Programming and Statistical Analysis, volume 40 of *Statistics and Computing*.

Springer, New York, NY, 2013.