Introduction to Data Analysis with Python

Cecilia Graiff

February 5, 2026

ALMAnaCH, Inria Paris - École d'Affaires Publiques, SciencesPo

Agenda

1. Course overview
2. What is data analysis

Course Overview

Course objectives

- Learn how **data scientists** manage a project; learn **data analysis** workflows
- Get an introduction to Python programming
- Learn Python for data analysis: `pandas`, `NumPy`, `Matplotlib`, `seaborn`
- Learn basic **statistical concepts** that build the foundations of data analysis

Course Overview

- **Week 1 (05/02/2026)**: Installation, Python coding basics.
- **Week 2 (19/02/2026)**: Data collection and cleaning.
- **Week 3 (12/03/2026)**: Data manipulation (`numpy`, `pandas`).
- **Week 4 (20/03/2026)**: Data visualization (`matplotlib`, `plotly`).
- **Week 5 (03/04/2026)**: Exploratory Data Analysis (EDA)
- **Week 6 (23/03/2026)**: Introduction to statistical modeling

Catch-up session

- On March 26 and April 9 there will be **no lecture**. A catch-up session is planned respectively for:
- **Friday, March 20, from 19:15 until 21:15**
- **Friday, April 3, from 19:15 until 21:15**

What is expected from you

Attend Lectures:

In case of **more than two** unjustified absences, your course cannot be validated. You are also invited to participate actively and bring all of your questions to the class :)

Complete the assignment:

At the end of the course, you will have to submit your project and a paper that documents it thoroughly. You will also have to submit all the used materials and the code. The deadline for project submission is on **May 4, 2026.**

Assignment

- Each project should be submitted by groups of 3 people;
- Project ideas are due by **March 30, 2026**:
  - Individuate the dataset you want to work on
  - Individuate 2-3 research questions
  - Summarize the approach that you want to follow
- You will have to design a complete pipeline of data analysis. Do not worry: this will become more clear during the course!
- **The project description will not be evaluated.**

Homework

- **Not mandatory!**
- Homework will be uploaded on GitHub before or after every class
- They will be corrected together at the beginning of each class
- **You are strongly encouraged to complete your homework, because it will help you a lot!**

Datasets

- To complete the course, you will have to **submit your own project**. It is important that this project is based on **your interests and domain of specialization**.
- You can have a look at Datasets of the EU to individuate possible topics and research questions that are interesting to you. However, the choice is not restricted to those datasets.

Class delegate election!

Please raise your hands if you would like to be class delegate.

What is Data Analysis

What is Data?

Data

Collection of **values** *that convey information*,
that help us **analyze**, **interpret**, and **make decisions**.

Types of data

- **Structured:** Organized, easy to search (e.g., tables, databases)
- **Unstructured:** Raw, complex (e.g., images, text, videos)
- **Semi-structured:** Mixed form (e.g., JSON, XML files)

See also: IBM - Structured vs Unstructured Data

Types of variables

- **Quantitative data**: Quantitative variables consist of **numerical values**. They can be **measured** precisely and can be used to perform operations (addition, subtraction, statistics such as the mean...)

Types of variables

- **Quantitative data**: Quantitative variables consist of **numerical values**. They can be **measured** precisely and can be used to perform operations (addition, subtraction, statistics such as the mean...)
- **Qualitative data**: Qualitative variables are non-numerical, and usually describe qualities, characteristics, or categories rather than numerical amounts. They usually are **categories** or **labels** (example: country, gender). Qualitative variables are also referred to as **categorical**.

Disclaimer: For the purpose of this class, we will only deal with these two types of variables.

Data Analysis: Beginner vs Advanced

- **From descriptive to predictive:**
  - Beginner: "What happened?"
  - Advanced: "What will happen?" and "What if we change something?"
  - E.g., **observing** vs **predicting** GDP growth, election outcomes, or patient survival.
- **Coding and math foundations:**
  - Compute basic data manipulation and summary statistics vs implement models from scratch
- **Handling more and richer data:**
  - Smaller and easier datasets vs complex data types: text (political speeches), spatial data (disease spread), networks.
- **Model evaluation and generalization:**
  - When you use **predictive models**, you need to evaluate their quality

Adapted from IBM Research.

Example: Regression

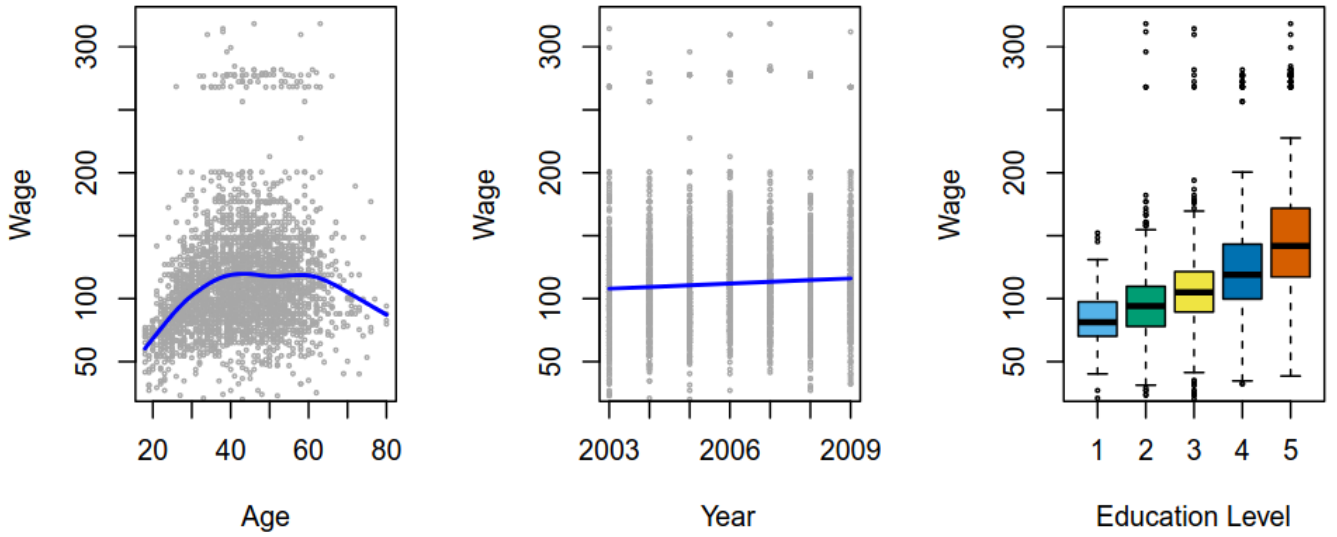


FIGURE 1.1. `Wage` *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: `wage` *as a function of* `age`*. On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`*. There is a slow but steady increase of approximately* $10,000 *in the average* `wage` *between* 2003 *and* 2009. Right: *Boxplots displaying* `wage` *as a function of* `education`*, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Example: Classification

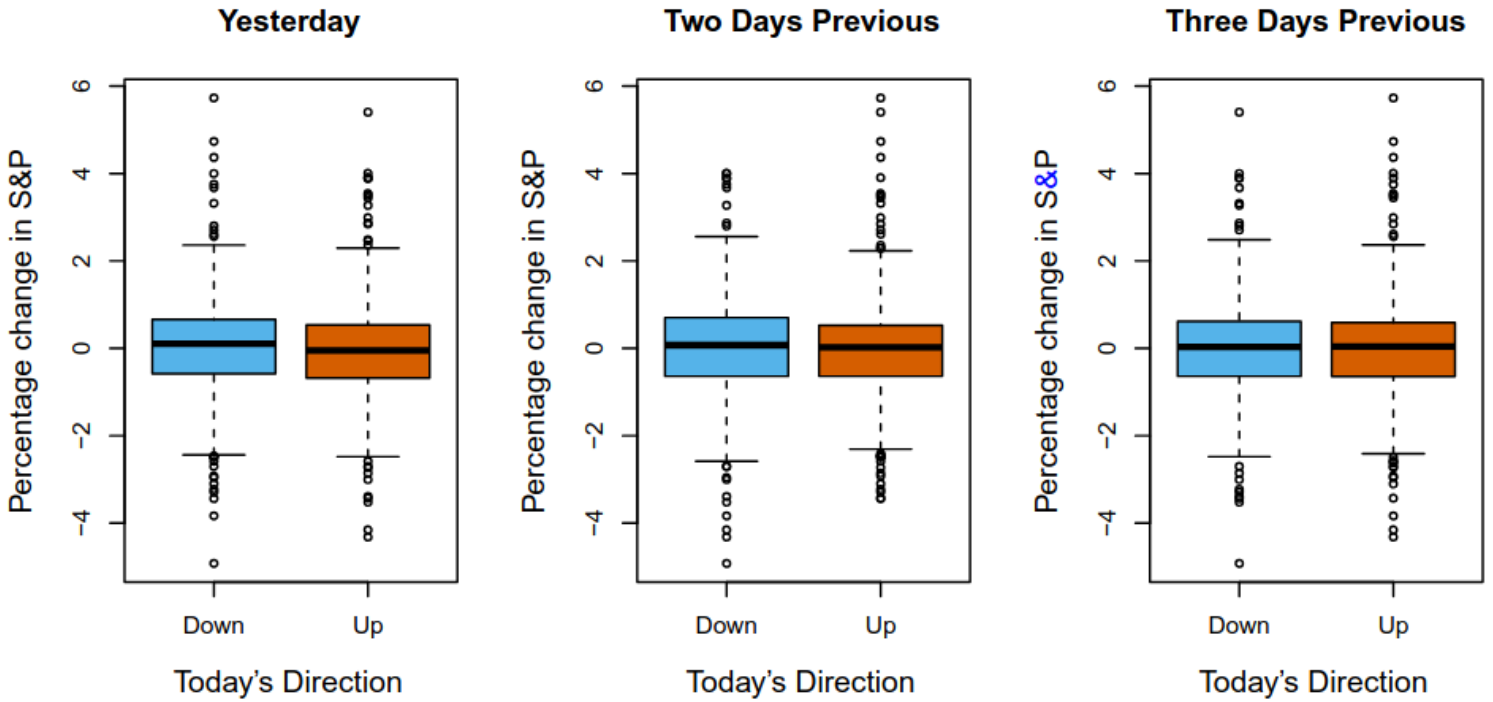
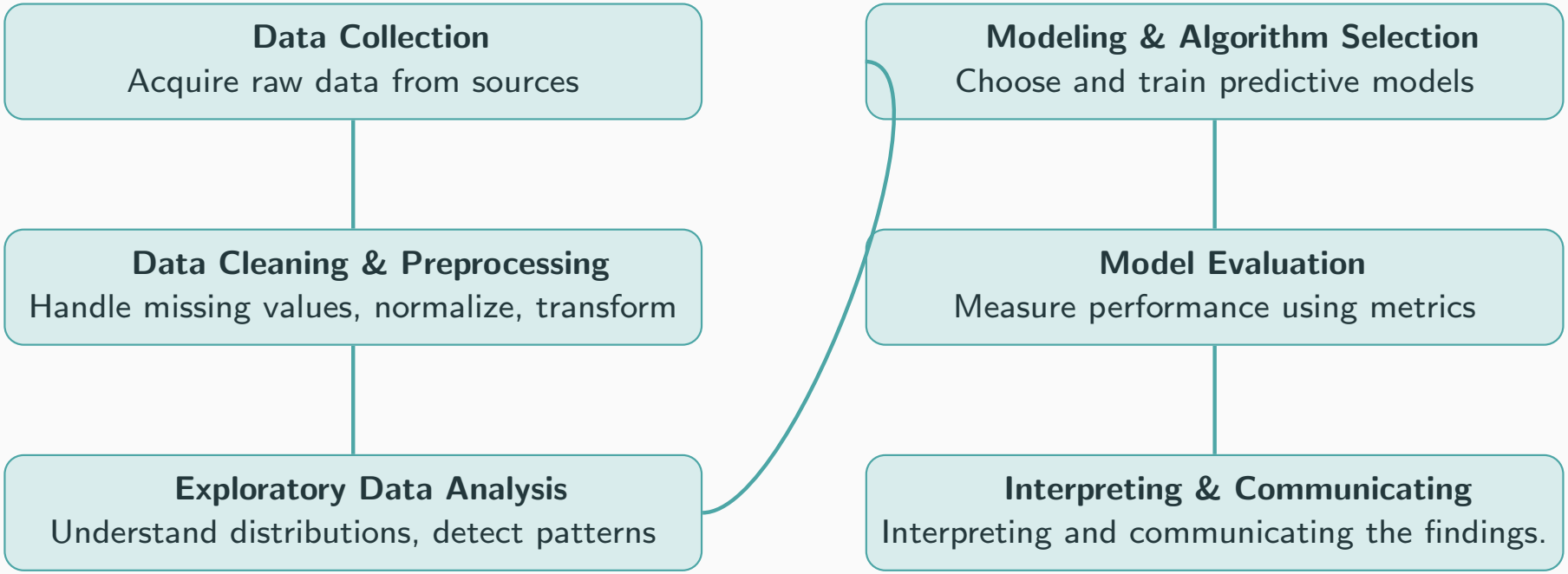


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Data Analysis Pipeline

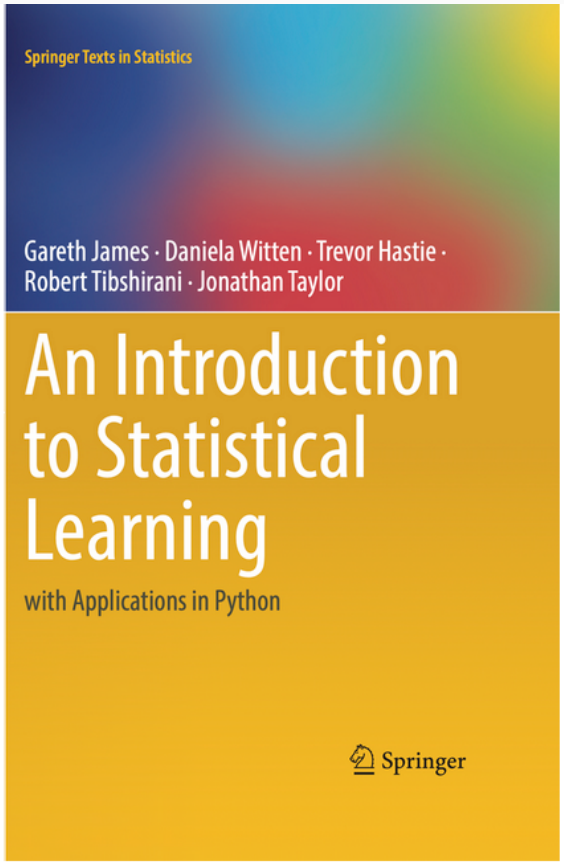


Conclusion

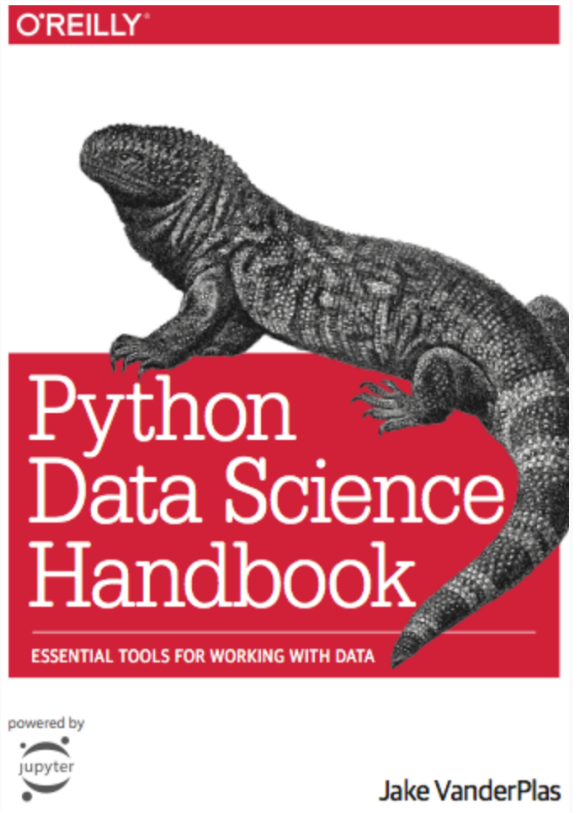
Take-home messages

- Course structure
- What is data and what is data analysis
- Now, we will focus on **learning a collaborative coding workflow** typical of data scientists

Recommended Readings



An Introduction to Statistical Learning



Python Data Science Handbook