

Exploring data

Chapter 24

This chapter covers

- Python's advantages for handling data
- Jupyter Notebook
- pandas
- Data aggregation
- Plots with matplotlib

Python's advantages for exploring data

- Python has become one of the leading languages for data science and continues to grow in that area.
- However, Python isn't always the fastest language in terms of raw performance.
- Conversely, some data-crunching libraries, such as NumPy, are largely written in C and heavily optimized to the point that speed isn't an issue.
- In addition, considerations such as readability and accessibility often outweigh pure speed; minimizing the amount of developer time needed is often more important.
- Python is readable and accessible, and both on its own and in combination with tools developed in the Python community, it's an enormously powerful tool for manipulating and exploring data.

Python can be better than a spreadsheet

- Spreadsheets have been the tools of choice for ad-hoc data manipulation for decades.
- People who are skilled with spreadsheets can make them do truly impressive tricks: spreadsheets can combine different but related data sets, pivot tables, use lookup tables to link data sets, and much more.
- But although people everywhere get a vast amount of work done with them every day, spreadsheets do have limitations, and Python can help you go beyond those limitations;
 - Most spreadsheet software has a row limit—currently, about 1 million rows.
 - Spreadsheets are two-dimensional grids, rows and columns, or at best stacks of grids, which limits the ways you can manipulate and think about complex data.
- With Python, you can code your way around the limitations of spreadsheets and manipulate data the way you want.

Python and pandas

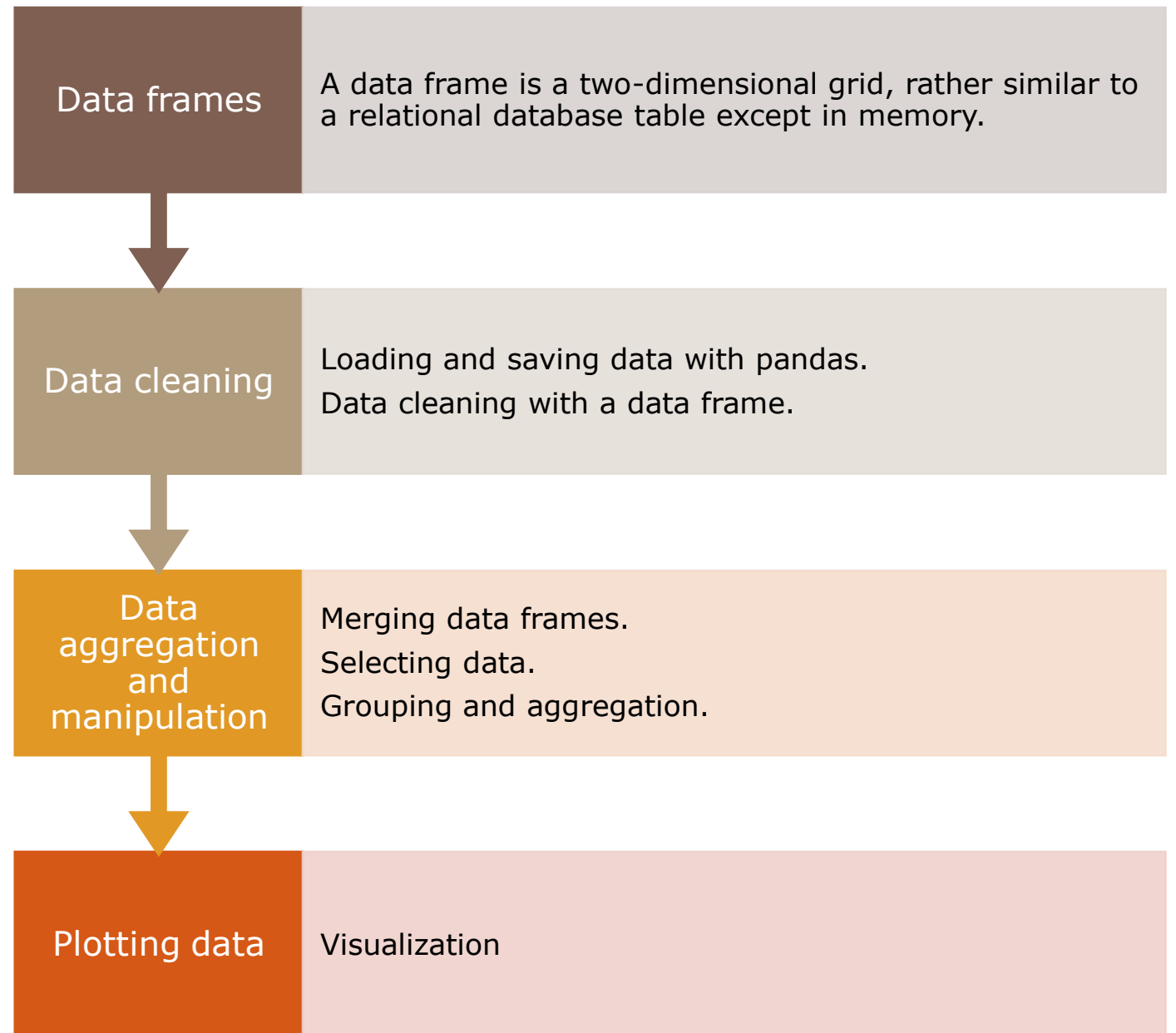
- One of the now-standard tools for handling data in Python – pandas - was created to automate the boring heavy lifting of handling data sets.
- pandas was created to make manipulating and analyzing tabular or relational data easy by providing a standard framework for holding the data, with convenient tools for frequent operations.
- As a result, it's almost more of an extension to Python than a library, and it changes the way you can interact with data.
- The plus side is that after you grok how pandas work, you can do some impressive things and save a lot of time.

Installing pandas

- pandas is easy to install with pip.
- It's often used along with matplotlib for plotting, so you can install both tools from the command line with this code:

```
pip install pandas matplotlib
```

Next steps
– let's
crunch it!



Summary

- Python offers many benefits for data handling, including the ability to handle very large data sets and the flexibility to handle data in ways that match your needs.
- Jupyter notebook is a useful way to access Python via a web browser, which also makes improved presentation easier.
- pandas is a tool that makes many common data-handling operations much easier, including cleaning, combining, and summarizing data.
- pandas also makes simple plotting much easier.