

Effective Programming Practices for Economists

# Data management with pandas

What is (modern) pandas?

Janoš Gabler and Hans-Martin von Gaudecker

7/21/2024

# What is pandas?

- Industry standard DataFrame library in Python
- Covers all you need for data management
  - Loading datasets in many formats
  - Cleaning data
  - Generating variables
  - Reshaping datasets
- Compatible with all plotting and statistics libraries

# What is a DataFrame?

|   | country | continent | year | life_exp |
|---|---------|-----------|------|----------|
| 0 | Cuba    | Americas  | 2002 | 77.16    |
| 1 | Cuba    | Americas  | 2007 | 78.27    |
| 2 | Spain   | Europe    | 2002 | 79.78    |
| 3 | Spain   | Europe    | 2007 | 80.94    |

- Tabular data format
- Variables are columns
- Observations are rows
- Can be manipulated in Python

# What is **modern** pandas?

- Pandas was created in 2008 and has some baggage
- With version 3.0 many things will improve
- Those features can already be enabled now:
  - More speed and less memory usage through better dtypes
  - Less confusion through copy-on-write
  - Better handling of missing values
  - Removal of the `inplace` argument

# How to use modern pandas

- Install version 2.1 or higher of pandas
- Install version 13.0 or higher of pyarrow
- Set some options after import

```
import pandas as pd

pd.options.mode.copy_on_write = True
pd.options.future.infer_string = True
```

- When loading datasets, use `engine="pyarrow"` if available