Effective Programming Practices for Economists

Data management with pandas

Data management: Definitions and example

Janoś Gabler and Hans-Martin von Gaudecker

Definitions

- **Data management**: Convert source data to formats your analysis programs need
  - Could include many different datasets
  - These could have a complex internals structure
  - Different datasets could be collected for different levels (e.g., a household survey and county-level unemployment rates from a statistical office)
- **Data cleaning**: Transform the contents of a table, leaving its structure unchanged
  - Coding missing values
  - Removing typos from strings
  - Setting proper data types

Remainder of the chapter

- *(Imperative and)* Functional data cleaning
- Handling complex data structures
- Tips, tricks, and advanced topics

Survey of course participants

	Q001	Q002	Q003
0	strongly disagree	agree	python
1	strongly agree	strongly agree	Python
2	-77	disagree	R
3	agree	-77	Python
4	-99	-99	Python
5	nan	strongly agree	Python
6	neutral	strongly agree	Python
7	disagree	agree	python
8	strongly agree	-99	PYTHON
9	agree	-99	Ypthon

- Q001: I am a coding genius
- Q002: I learned a lot
- Q003: What is your favourite language?

- -77 not readable
- -99 no reply

```
>>> df = pd.read_csv("survey.csv")
>>> df.dtypes
Q001     str
Q002     str
Q003     str
dtype: object
```