Effective Programming Practices for Economists

Data management with pandas

Rules for data management

Janoś Gabler and Hans-Martin von Gaudecker

Motivation

- So far we have shown you the mechanics of using pandas
- Now we talk about general best practices
  - Can save weeks of work in large projects
  - Grounded in database research

1. Never ever change source data

- **Source data**: Original dataset as downloaded or collected
- Commit the source data to git and never change it
- All modified datasets should be stored under different names
- Modified datasets should not be under version control!

2. Separate data mgm’t and analysis

- **Data management**: Converting source data to formats your analysis programs need
- Separate data management code from analysis code
- Never modify the content of a variable outside the data management code!

3. Values have no internal structure

- a.k.a. the **first normal form**
- I.e., no need for parsing values before using them
- E.g. store first names and last names separately
- Not too often a problem in economic data
  - X-digit industrial or educational classifiers
  - Store each digit level you need in a separate variable

4. No redundant information in tables

- a.k.a. the **second normal form**
- In a panel structure: Store time-constant characteristics in a separate table
- Violations make things much harder and error-prone:
  - Changes to data
  - Consistency checks
  - Selecting observations

5. No structure in variable names

a.k.a. use long format if you can

There should not be different variables with similar content referring to different time periods etc.

If you need wide format for regressions, still do your data management in long format

	country	year	gdp_per_cap	pop
0	Cuba	2002	6340.65	11226999
1	Cuba	2007	8948.10	11416987
2	Spain	2002	24835.47	40152517
3	Spain	2007	28821.06	40448191

(long format)

	gdp_per_cap_2002	gdp_per_cap_2007	pop_2002	pop_2007
country				
Cuba	6340.65	8948.10	11226999.00	11416987.00
Spain	24835.47	28821.06	40152517.00	40448191.00

(wide format)