Effective Programming Practices for Economists

Data management with pandas

Managing data with a complex structure

Janoś Gabler and Hans-Martin von Gaudecker

Motivation

- Real-world data often has complex structure
- Understanding how to organize data is crucial
- Proper data organization can save weeks of work
- Three principles, which come from database research

1. 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

2. 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

3. Variable names have no structure

- a.k.a. use long format if you can
- There should not be different variables with similar content referring to different time periods etc.

country	year	gdp_per_cap	pop
Cuba	2002	6341	11226999
Cuba	2007	8948	11416987
Spain	2002	24835	40152517
Spain	2007	28821	40448191

(long format)

country	gdp_per_cap_2002	gdp_per_cap_2007	pop_2002	pop_2007
Cuba	6341	8948	11226999	11416987
Spain	24835	28821	40152517	40448191

(wide format)

Complex data structures in practice

1. Do all data cleaning in a collection of tables, each of which satisfies the rules
   - You might get the data like that (different providers, smart providers)
   - You might need to break original data up (time-varying vs time-constant)
2. Store each of these tables in a separate file
3. Merge tables as needed for analysis
   - Could be storing one big table as after data management is done
   - Could be merging tables on the fly
   - DRY!