Causal Inference

Mixtape Session

Roadmap

- Hidden curriculum
  - Background
  - Empirical workflow
  - Hierarchical folder structure
  - Naming conventions
  - Version control
  - Soft skills


©Robert Del Tredici

Do Not Touch the Original Data

- Empirical workflows require data manipulation
- It is **imperative** that you always and only work with **copies**
- Never save over the original dataset – be careful of Excel which may do it automatically
- Avoid this by storing the raw data separate from your copies
- Do not alter the raw data – you will ruin it and may not get it back

SPOT THE DIFFERENCE DAILY

January 2, 2022

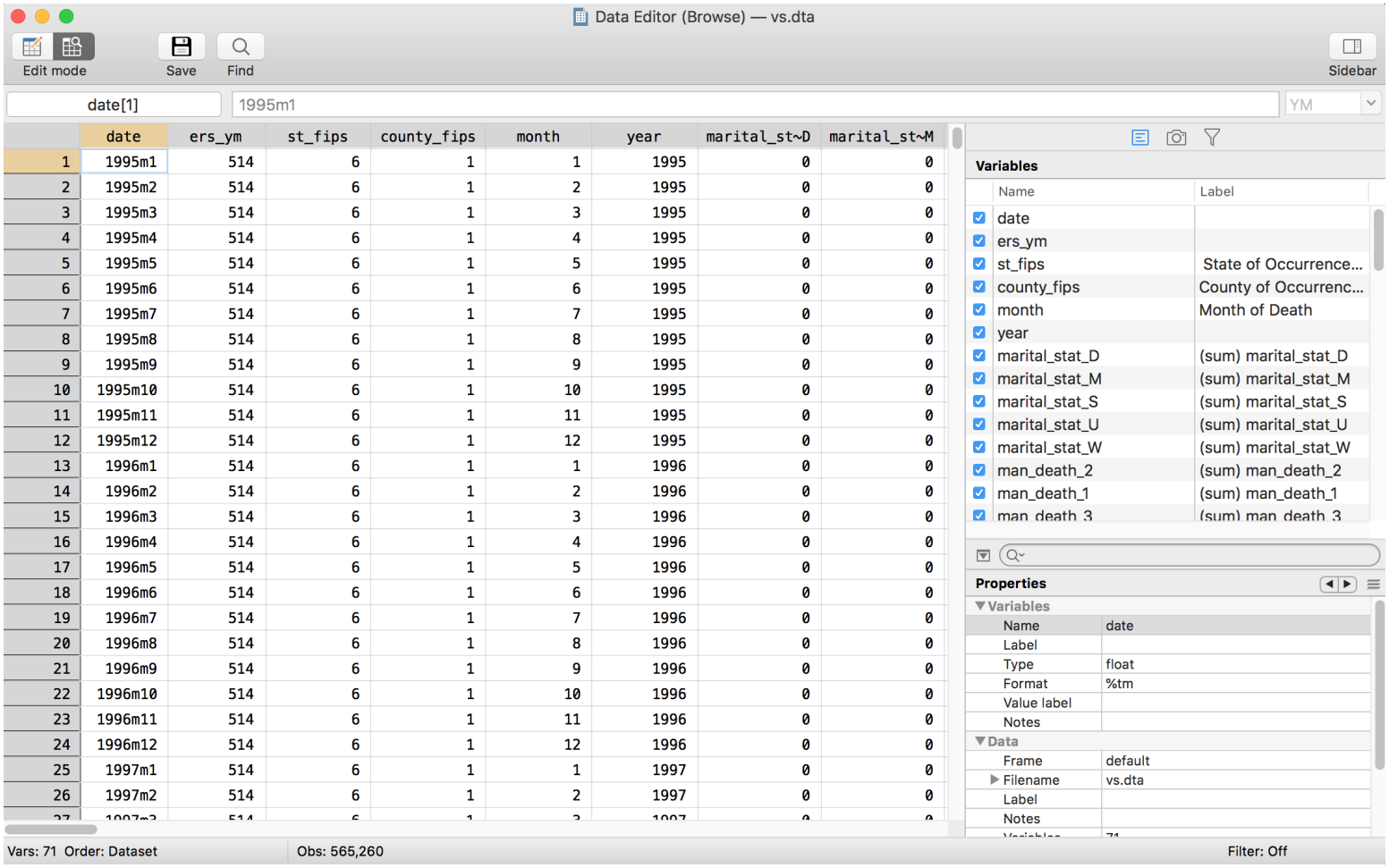
HINT

Spot the 10 differences

00:03

Try and spot the problems

- Our eyeballs evolved to spot patterns
- We can therefore use it to find things that belong but also things that don't
- First just scan the data in its spreadsheet form – get comfortable with what you're going to bre using
- Use `browse` or excel to just read the spreadsheet with your eyes.
- See if anything jumps out



	date	ers_ym	st_fips	county_fips	month	year	marital_st~D	marital_st~M
1	1995m1	514	6	1	1	1995	0	0
2	1995m2	514	6	1	2	1995	0	0
3	1995m3	514	6	1	3	1995	0	0
4	1995m4	514	6	1	4	1995	0	0
5	1995m5	514	6	1	5	1995	0	0
6	1995m6	514	6	1	6	1995	0	0
7	1995m7	514	6	1	7	1995	0	0
8	1995m8	514	6	1	8	1995	0	0
9	1995m9	514	6	1	9	1995	0	0
10	1995m10	514	6	1	10	1995	0	0
11	1995m11	514	6	1	11	1995	0	0
12	1995m12	514	6	1	12	1995	0	0
13	1996m1	514	6	1	1	1996	0	0
14	1996m2	514	6	1	2	1996	0	0
15	1996m3	514	6	1	3	1996	0	0
16	1996m4	514	6	1	4	1996	0	0
17	1996m5	514	6	1	5	1996	0	0
18	1996m6	514	6	1	6	1996	0	0
19	1996m7	514	6	1	7	1996	0	0
20	1996m8	514	6	1	8	1996	0	0
21	1996m9	514	6	1	9	1996	0	0
22	1996m10	514	6	1	10	1996	0	0
23	1996m11	514	6	1	11	1996	0	0
24	1996m12	514	6	1	12	1996	0	0
25	1997m1	514	6	1	1	1997	0	0
26	1997m2	514	6	1	2	1997	0	0

Missing observations

- Check the size of your dataset in Stata using `count`
- Check the number of observations per variable in Stata using `summarize`
  - → String variables will always report zero observations under `summarize` so `count if X==""` will work
- Use `tabulate` also because oftentimes missing observations are recorded with a -9 or some other illogical negative value