merging; due in 2 weeks

[version: Wednesday 10th February, 2021   16:56]

1. merge your dataset with at least 5 other datasets (including merge(s) used for ps2, so there must be 5 in total ; as always, start early and email listserv if you get stuck); you may merge on u/a or geography (eg state) or time, (eg year) or some characteristics eg occupation

2. at least one of the 5 merges has to be m:1 (or 1:m)

3. as always, after every merge need to investigate it, ie have a hard look at _merge, eg tab it with key/id/merge variable, edit if _merge!=3 etc AND you need to explain every non-merge—why it didn't merge and whether it's expected and why

4. it has to be 5 other real datasets; that is, you cannot generate "fake" new datasets as I did in the class by artificially splitting (collapsing, etc) my dataset; in short, data used for merge must not have been originally in the same dataset! Neither it can be different datasets from the same source! Use as varied sources as possible, say UN, OECD, WTO, WHO, IMF, WB, Census, CDC, survey data, NOAA (weather) etc

5. again, as always: what are these data, where it exactly come from? give URL! i need to be able to find the source!

6. reshape your data into something sensible at least twice

7. as we are getting closer to the final project, your writeup/description of research should get longer and include at minimum, why you use specific datasets, whats the research questions, hypotheses and models and specific variables you are interested in and why?

hints

1. you may reshape the same data several times to come up with something interesting

---

general directions (always the same):

- i will show your code in class and possibly post some of your code or link to it—again, as per our core values—opensource, transparency, sharing; but if you'd like to keep your code private, that's fine—just let me know, and i will keep your code secret (no penalty, except that you may get little less feedback—usually if we discuss your code in the class, you will benefit from it!)

- you must submit all the code that was executed from the very beginning starting with the very raw data as per replication principle; unless data is too big to fit online, then just start with a comment, eg "to fit data online i had to take a random sample of 10perc"

- all ps are mostly cumulative—you can, and should, include much of previous code you've written for this class; can also use code you've written outside of this class (other classes, projects, etc)—but you have to clearly mark the code that has not been written for this class—otherwise, scholastic dishonesty!

- use your own dataset; again if you do not have a dataset, ask for help finding it

- because you are only submitting code, it must load data from Internet—just put your data onto your own website, wordpress, google drive, etc; (when you put data into any public space, try not to violate data copyrights... I haven't heard of anyone having problems with that, but be careful—for instance you may subset dataset to few vars and smaller sample using **sample**); and it is also easier to experiment on small datasets

- keep it simple! at the beginning of your dofile drop unnecessary vars; and even retain only certain, say most important, observations; keep it manageable; it is much easier to learn using simple data; can always complicate later!; much better to do it right using simple data than do it wrong using complex data!

- have nice structure in your file: sections, subsections, etc; may also have multiple files

- great idea to submit ps as early as possible—we will probably give you some comments; if not, email us and ask for comments!

- it is great to copy code from others; again, one of the rules for this class is 'be lazy': don't reinvent the wheel, whatever you are coding, it has already been done, google things often; but of course you cannot submit 100% code by someone's else.

- if you do something extra/fancy that is relevant and closely related to the assignment questions, it will be extra credit

- use coding rules that we've learned so far

- submit (only) the code into git repo; ps are due by the beginning of the next class unless indicated otherwise, eg "due in 2 weeks"; late ps are not accepted; NOTE: push to github early and send email to listerv with the link to your submission and ask for comments and ask any questions—the surest way to get the ps right!

- we are on the way to developing the final project with these ps: as we progress, your ps should start resembling a coherent and logical project where you use data management techniques to build new a dataset that can be used to answer interesting questions– say in few sentences (as a comment) why are you doing what you are doing–that is, answer the "so what question": "ok, you're gonna run all that code, and so what?" what's the goal of all that, why are you doing this? you need a compelling justification for what you are doing; typically: to develop a new dataset (that has not existed before) that can be used to answer some exciting questions: say what are those questions you want to answer; be brief, say couple sentences, and definitely not more than say 100 lines, typically 10-50 lines is enough; related: even at the beginning, already in ps1, say why you use data you are using, is it best, does it serve the purpose; also, feel free to ask me questions in comments

- be prepared do present your code in class (if time), just briefly, key points, couple minutes