data cleaning principles

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

@kwbroman
kbroman.org
github.com/kbroman

kbroman.org/Talk_DataCleaning

These are slides for a talk for the csv,conf,v6 (https://csvconf.com/) on May 4-5, 2021.

Data analysts spend a lot of time organizing and cleaning data, but few of us have been trained to do so. Why is that?

Some say that data cleaning is difficult to generalize. But I think there are some general principles. Moreover, I think we have an important shared experience in data cleaning that we can commiserate about, and through which we can learn from each other.

Tidy data are all alike,
but every messy dataset
is messy in its own way.

– Hadley Wickham

Hadley's talking more about data organization than data cleanliness. And his point is that if you make data tidy, it simplifies all the downstream analyses.

But is every messy dataset uniquely messy?

For sure, my collaborators have shown impressive creativity in their organization and management of data. But we do see the same sorts of problems over and over.

If I clean up [Medicare] data ...
does any of the knowledge I gain ...
apply to the processing of RNA-seq data?

– Roger Peng

In his discussion of David Donoho's paper about data science, Roger Peng wrote about how data cleaning is frustratingly difficult to generalize.

But my answer to his question is absolutely!

A person with experience cleaning one dataset has important experience to draw upon when moving to another dataset even if it's of a totally different nature.

Data Mishaps Night

Join us for the first inaugural Data Mishaps Night! We will feature a lineup of data mistake stories with a focus on the human aspect of data work and lessons learned the hard way.

Caitlin Hudon & Laura Ellis

dataMishapsNight.com

In February, 2021, Caitlin Hudon and Laura Ellis organized an Friday evening conference where 16 people gave short presentations on data mishaps.

Many of the stories concerned mistakes in data cleaning, and while these weren't necessarily the most amusing stories, they did seem to bring out a strong sense of shared experience. We have suffered and struggled through very similar data problems.

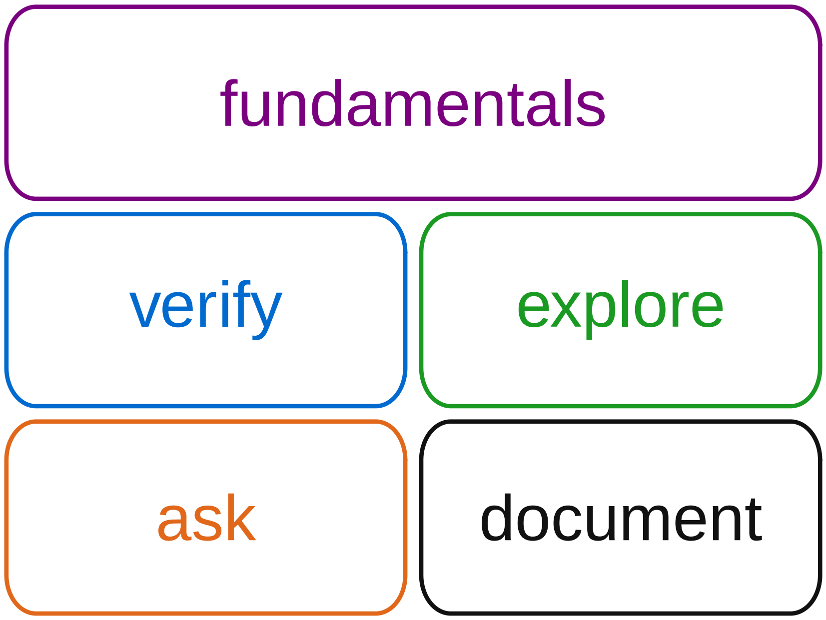
Data cleaning

- tedious
- embarrassing
- needs context
- doesn't feel like progress
- requires creativity
- requires coding prowess
- source of most problems

Really, I think we don't usually teach data cleaning because it's something we prefer to keep private.

We're shy about it.

And data cleaning code is our ugliest code.



I'm proposing a set of basic principles for data cleaning, and splitting them into five groups. There are some fundamental principles, followed by four basic ideas: verify things that you expect, explore to find further oddities, ask questions, and document what you've done.

fundamentals

1. Don't clean data when you're tired or hungry.

(paraphrasing Ghazal Gulati)

At her talk at the Data Mishaps night, Ghazal Gulati emphasized this point, of not cleaning data when you're tired or hungry.

Data cleaning requires considerable concentration, and you need to allow sufficient time to do the work. If you're in a hurry, you'll miss things.

fundamentals

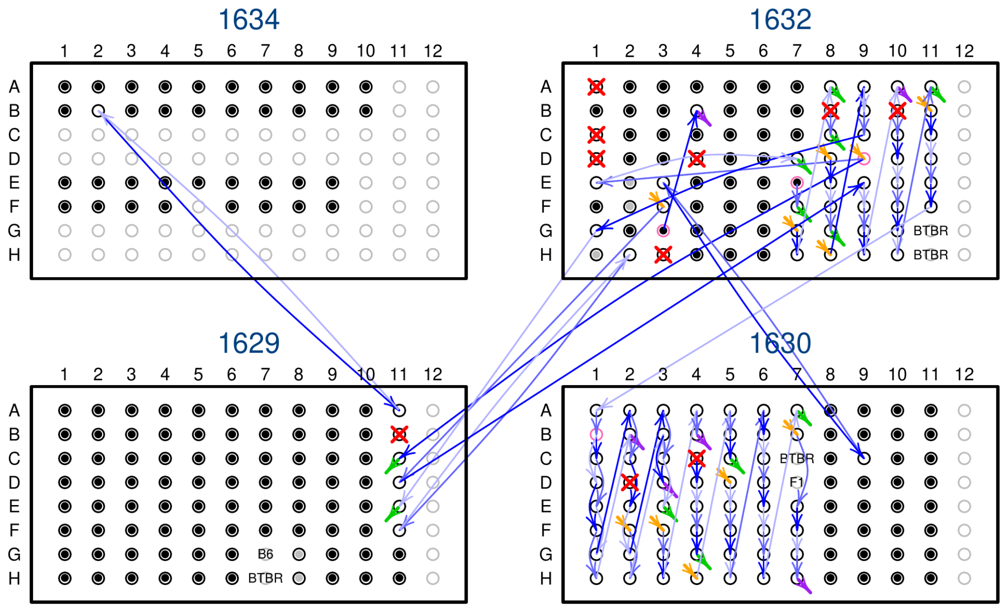
2. Don't trust anyone (even yourself)

"my motto is 'trust no one'
...except maybe @kwbroman?"

– Jenny Bryan

fundamentals

3. Think about what might have gone wrong and how it might be revealed



9

Personally, I think this is the most important principle for data cleaning. It has been central in guiding my approach.

I need a good figure for this.

fundamentals

4. Use care in merging

	A	B	C	D	E	F	G
1	id	glucose.0	glucose.5	glucose.15	glucose.30	insulin.0	insulin.5
2	DO−221	145.742786	206.452638	216.640608	299.55501	0.74455	2.0264
3	DO−222						
4	DO−223						
5	DO−224						
6	DO−225						
7	DO−226						
8	DO−227						
9	DO−228						
10	DO−229						
11	DO−230						

	A	B	C	D	E	F	G
1	id	glucose.0	insulin.0	glucose.5	insulin.5	glucose.15	insulin.15
2	DO−321	66.839405	0.04	246.685995	0.04	305.26214	0.04
3	DO−322	98.12509	0.51185	246.25574	1.4062	301.8201	2.828
4	DO−323	94.68305	1.7812	448.1068	1.0248	521.61894	1.02725
5	DO−324	121.051535	0.0882	407.355505	0.63475	470.541525	0.8195
6	DO−325	122.95695	0.19155	298.193665	0.6467	323.148455	0.40515
7	DO−326	201.447755	0.7454	386.51887	0.6081	654.99799	1.07225
8	DO−327	130.025425	0.0509	477.302675	0.166	610.49733	0.4842
9	DO−328	143.60919	0.23435	438.88705	0.70505	406.249135	0.2498
10	DO−329	125.29262	0.04	543.74634	1.7366	520.205245	0.8498
11	DO−330	135.61874	0.91275	393.03416	3.73095	454.62209	1.7325

Many problems arise due to mistakes when merging data from multiple files. A common problem is a change in the data arrangement, such as in the order of columns.

Focus on the labels (which are more likely correct), rather than the position of variables in a file (which are more likely to change).

fundamentals

5. Dates & categories suck

You may ask, “How is that a principle?”

Principle:

a fundamental truth that guides our thinking

I was thinking the same thing. Was I drifting away from principles and more to just stuff to know or do? This seems a pretty good definition.

fundamentals

5. Dates & categories suck

So yeah, this counts as a principle.

Much of the pain will be in the dates and categorical data; you should be ready for that.

verify

6. Check that distinct things are distinct

verify

7. Check that matching things match

verify

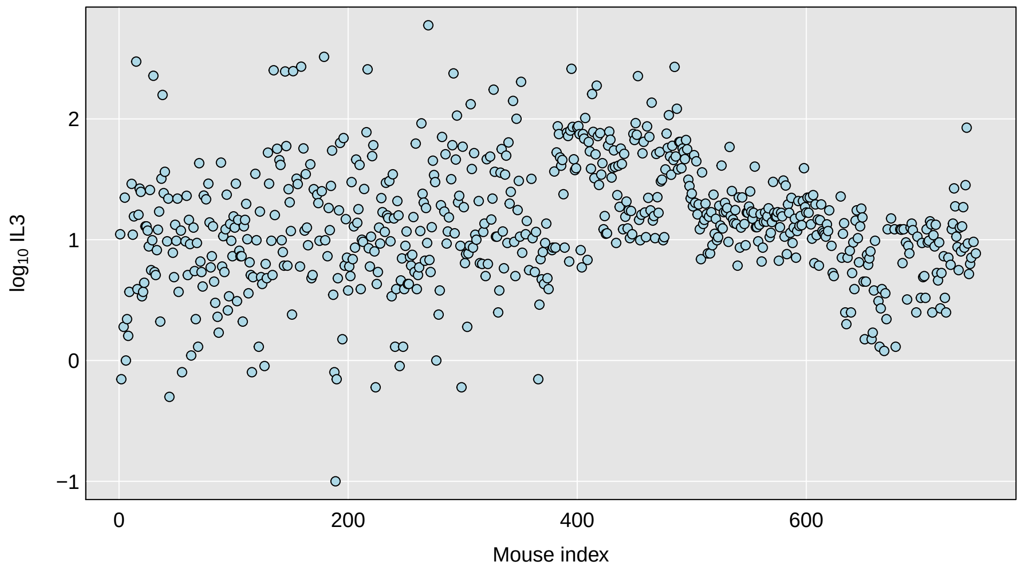
8. Check calculations

verify

9. Look for other instances of a problem

explore

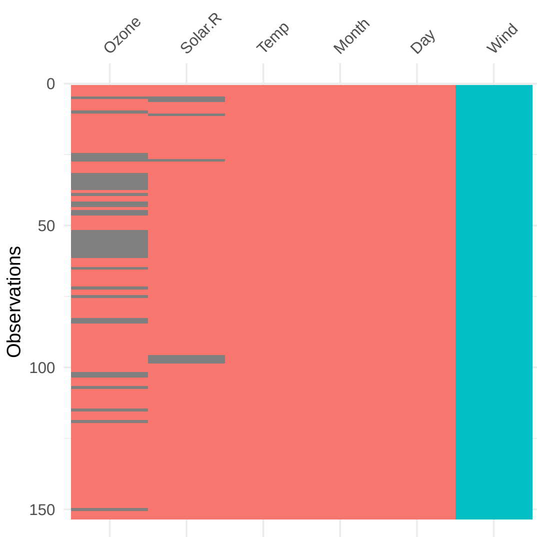
10. Make lots of plots



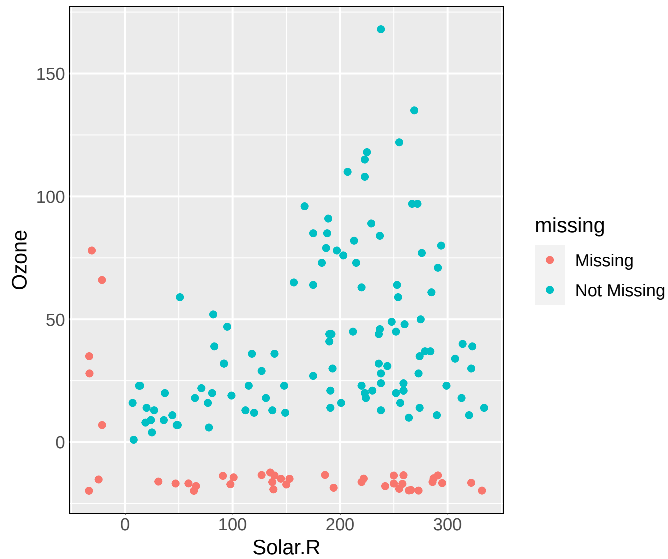
explore

11. Look at missing value patterns

{visdat}



{naniar}



visdat: https://docs.ropensci.org/visdat/

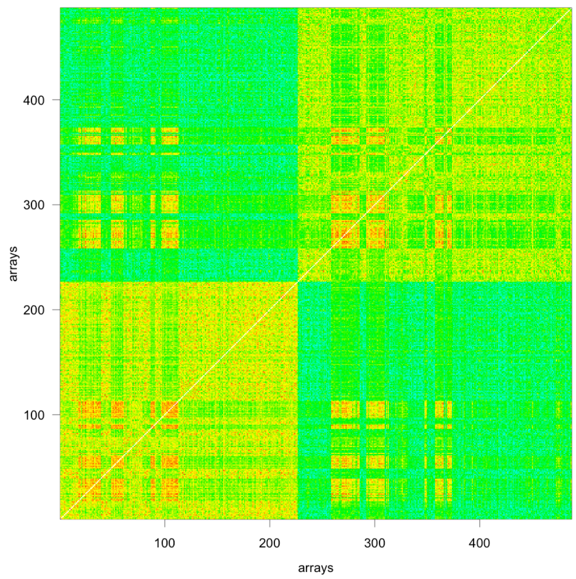
naniar: http://naniar.njtierney.com/

explore

12. With massive data,
    make more plots not fewer

explore

13. Follow up all artifacts



kbroman.org/blog/2012/04/25/microarrays-suck

Wow the clash of those colors is particularly bad.

This is a heat map of the correlation matrix for a set of gene expression microarrays. The plaid pattern was a shock to me, and was caused by a set of bad arrays that we hadn't noticed previously.

My point here is simply to follow up all artifacts.

If you see something weird, follow through and try to figure out the underlying cause. If could be an error, or a set of bad assays, or it could be the most interesting finding in the study.

ask

14. Ask questions
15. Ask for the primary data
16. Ask for metadata
17. Ask why data are missing

document

18. Create checklists & pipelines
19. Document not just what but why
20. Expect to recheck

fundamentals

1. Don't clean data when tired or hungry
2. Don't trust anyone (even yourself)
3. Think about what might have gone wrong
4. Use care in merging
5. Dates & categories suck

verify

6. Distinct things are distinct
7. Matching things match
8. Check calculations
9. Look for other instances

explore

10. Make lots of plots
11. Look at missing value patterns
12. With big data make more plots
13. Follow up all artifacts

ask

14. Ask questions
15. Ask for the primary data
16. Ask for metadata
17. Ask why data are missing

document

18. Create checklists & pipelines
19. Document not just what but why
20. Expect to recheck

In summary...

Slides: kbroman.org/Talk_DataCleaning

kbroman.org

github.com/kbroman

@kwbroman