data cleaning principles

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

@kwbroman

kbroman.org

github.com/kbroman

kbroman.org/Talk_DataCleaning

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

– Hadley Wickham

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

– Roger Peng

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

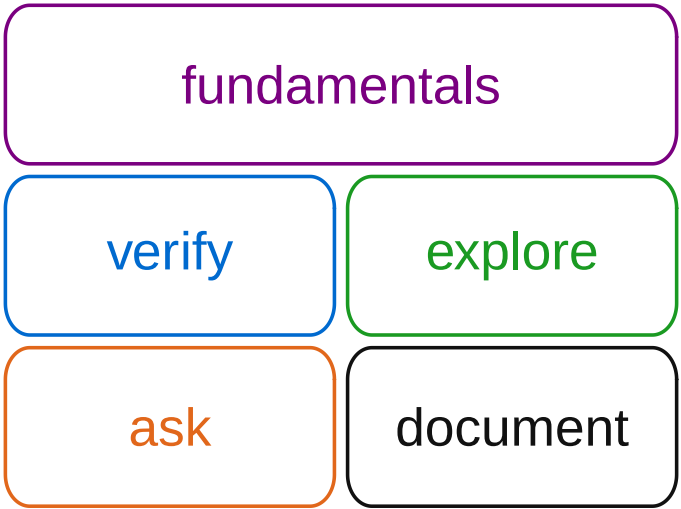
Data cleaning

- tedious
- embarrassing
- needs context
- doesn’t feel like progress

Data cleaning

- tedious
- embarrassing
- needs context
- doesn't feel like progress

- requires creativity
- requires coding prowess
- source of many problems


fundamentals
verify
explore
ask
document

fundamentals

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

(paraphrasing Ghazal Gulati)

fundamentals

2. Don’t trust anyone (even yourself)

fundamentals

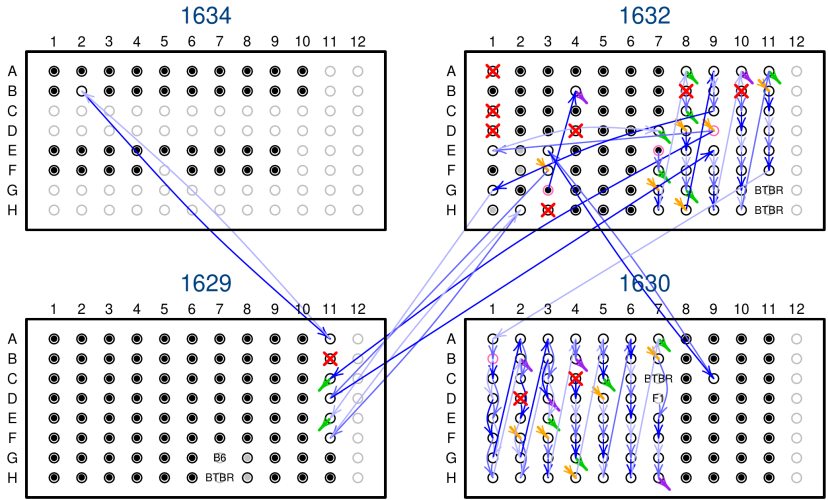
2. Don’t trust anyone (even yourself)

“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



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

fundamentals

5. Dates & categories suck

Principle:

a fundamental truth that guides our thinking

fundamentals

5. Dates & categories suck

verify

6. Check that distinct things are distinct

	A	B	C	D	E	F	G
1	WiscID	ID	NEOID	Fem_CA	Fem_Imax	Fem_Imin	Fem_J
2	F2.C1W.F.1248	1248	NEO183	0.7524	0.1427	0.1006	0.2433
3	F2.C1W.M.1250	1250	NEO184	0.7669	0.1556	0.09652	0.2521
4	F2.C1W.F.1251	1251	NEO185	0.7613	0.1549	0.09659	0.2515
5	F2.C1W.F.1254	1254	NEO186	0.7475	0.1503	0.08603	0.2363
6	F2.C1W.M.1257	1257	NEO187	0.8197	0.1849	0.1056	0.2905
7	F2.___.F.715	715	NEO764	0.6017	0.09662	0.05969	0.1563
8	F2.___.F.751	751	NEO765	0.7273	0.1304	0.08735	0.2178
9	F2.___.F.1251	1251	NEO766	0.6675	0.1157	0.07814	0.1938
10	F2.___.M.1340	1340	NEO768	0.6656	0.1387	0.08122	0.2199
11	F2.C1W.M.739	739	NEO779	0.9336	0.2828	0.1628	0.4456

verify

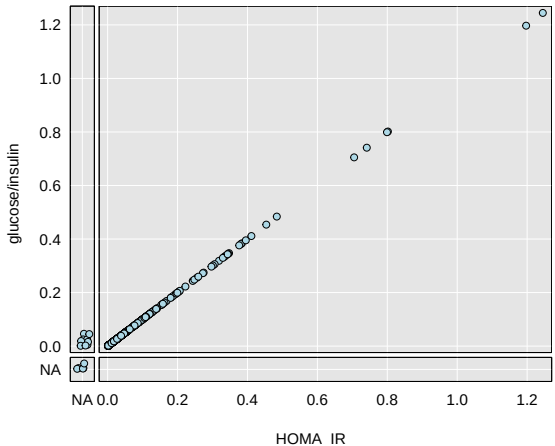
7. Check that matching things match

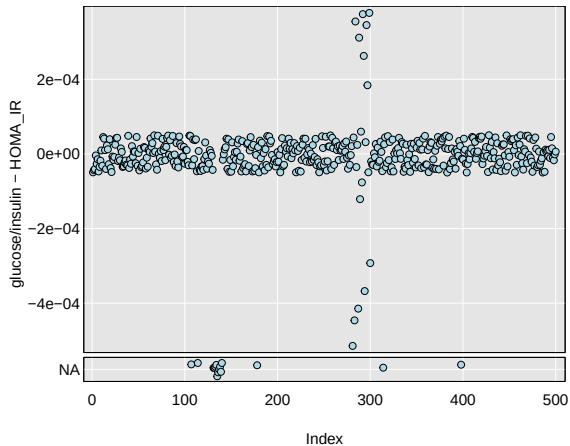
	A	B	C	D
1	id	sex	n_gen	age_days
2	F20.25	M	20	75
3	F21.30	M	21	75
4	F21.68	M	21	71
5	F22.52	M	22	73
6	F21.71	F	22	63
7	F22.116	F	22	57
8	F21.F20.9.M5	M	20	82
9	F21.F20.18.M5	M	20	77
10	F20.26	M	20	75
11	F21.62	M	21	72

	A	B	C	D
1	id	sex	age_at_dosing	n_gen
2	F22.69	F	67	22
3	F22.106	F	69	22
4	F22.70	F	67	22
5	F22.107	F	69	22
6	F21.71	F	65	21
7	F22.116	F	62	22
8	F22.73	F	65	22
9	F22.117	F	62	22
10	F21.108	F	62	21
11	F22.118	F	59	22

verify

8. Check calculations



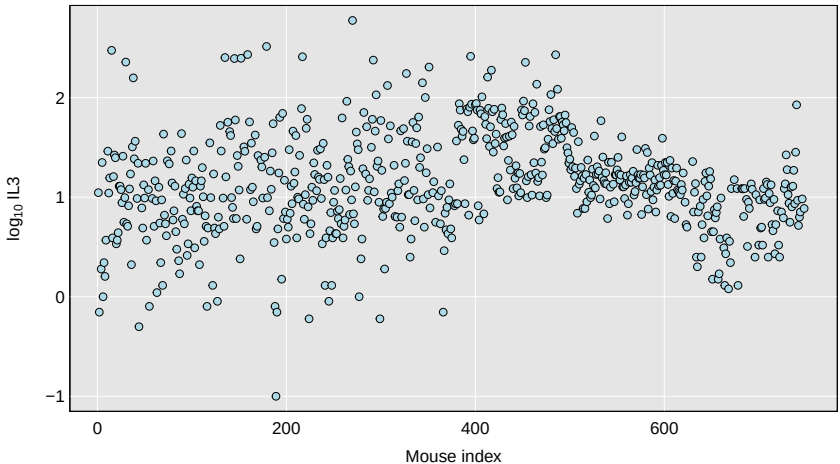


verify

9. Look for other instances of a problem

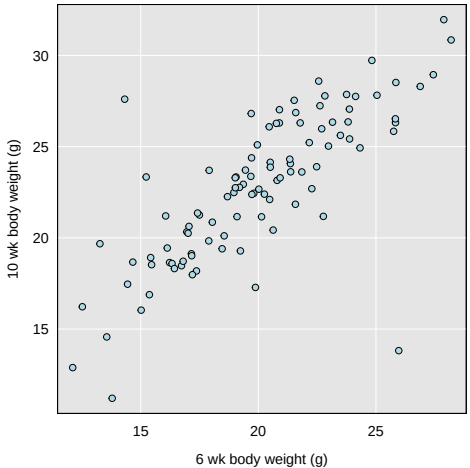
explore

10. Make lots of plots



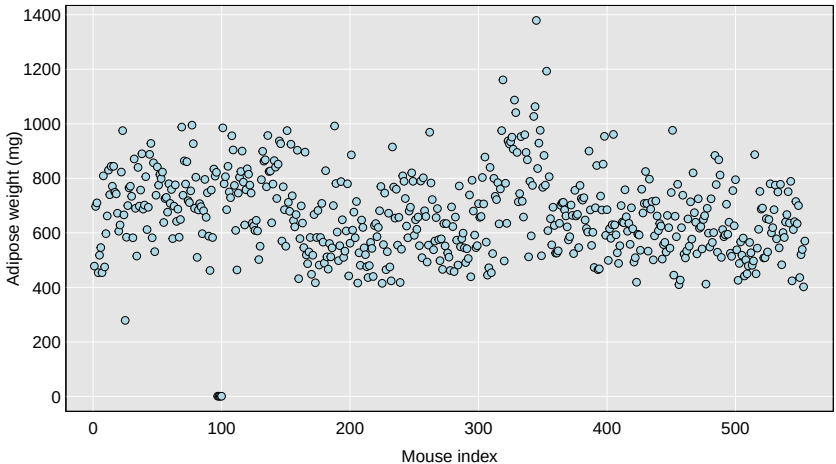
explore

10. Make lots of plots



explore

10. Make lots of plots



explore

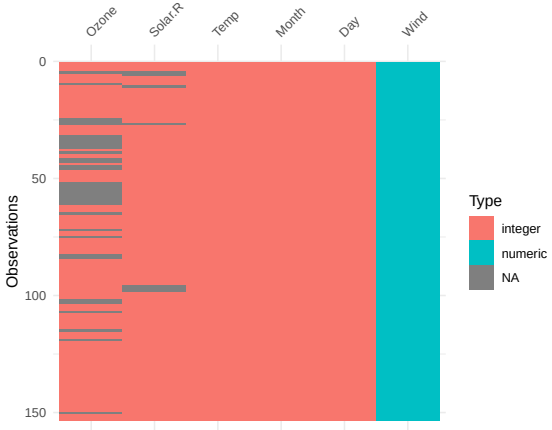
10. Make lots of plots

	A	B	C	D
1	id	Rt Kidney wt	Rt Adipose wt	Liver wt
2	DO−121	294	757	930
3	DO−122	296	583	439
4	DO−123	NA	834	527
5	DO−124	513	808	600
6	DO−125	381	780	493
7	DO−126	225	1.066	355
8	DO−127	262	1.03	512
9	DO−128	231	0.687	497
10	DO−129	263	0.932	580
11	DO−130	266	985	906

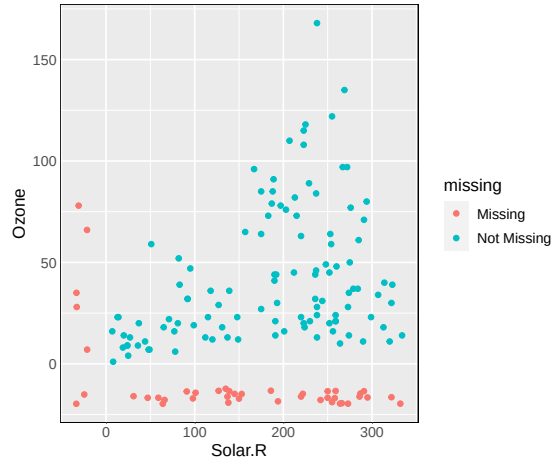
explore

11. Look at missing value patterns

{visdat}

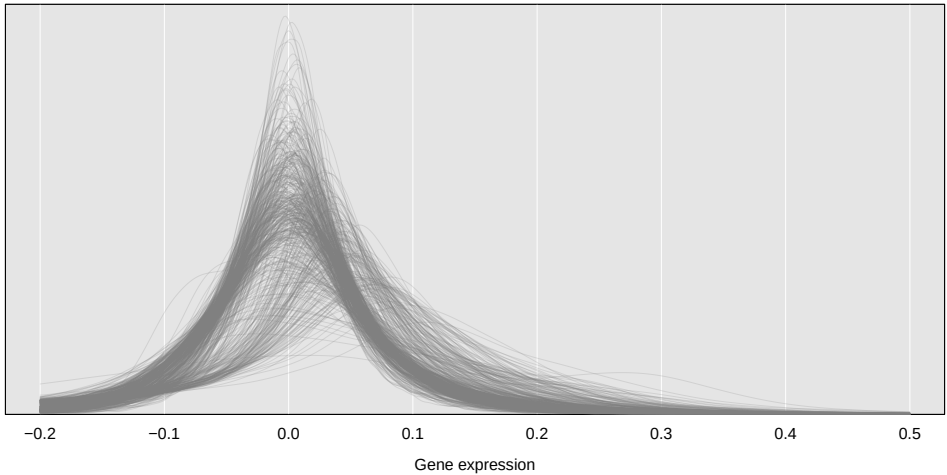


{naniar}



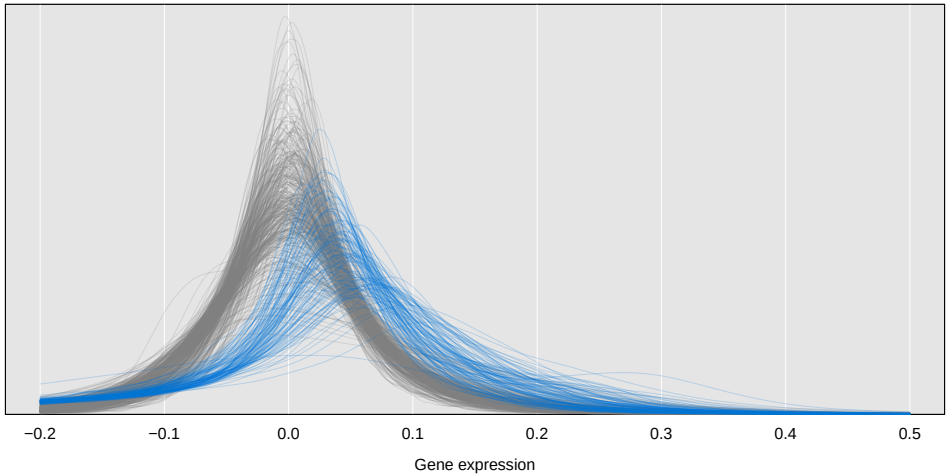
explore

12. With massive data,
make more plots not fewer



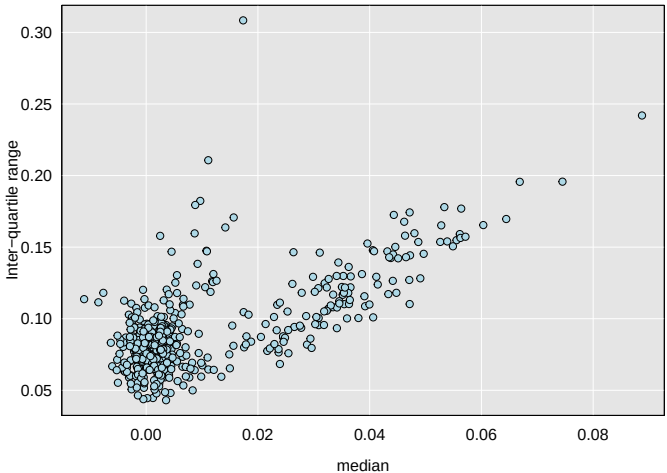
explore

12. With massive data,
make more plots not fewer



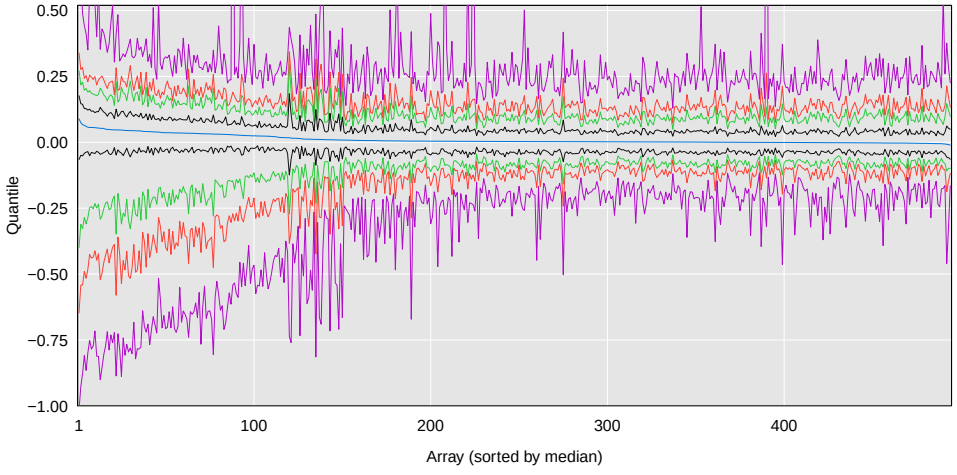
explore

12. With massive data,
make more plots not fewer



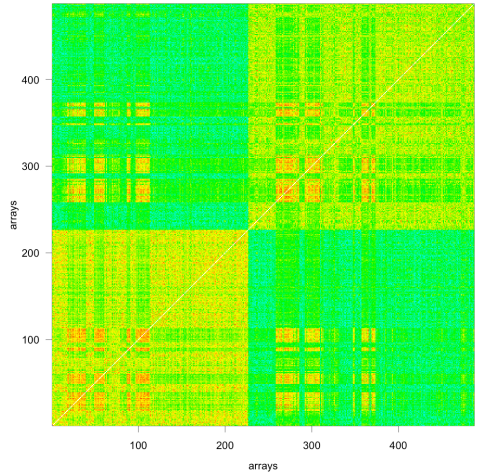
explore

12. With massive data,
make more plots not fewer



explore

13. Follow up all artifacts



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

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. Verify that distinct things are distinct
7. Verify that matching things match
8. Check calculations
9. Look for other instances of problems

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

I will let the data speak for itself
when it cleans itself.

– Allison Reichel

Slides: kbroman.org/Talk_DataCleaning

kbroman.org

github.com/kbroman

@kwbroman