Wrangling messy data files

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

In this lecture, we'll look at the problem of wrangling messy data files: A bit of data diagnostics, but mostly how to reorganize data files.

"In what form would you like the data?"

"In its present form!"

...so we'll have some messy files to deal with.

When collaborators ask me how I would like them to send me data, I always say: in its present form.

I cannot emphasize enough: If any transformation needs to be done, or if anything needs to be fixed, it is the data scientist who is in the best position to do the work, reproducibly and without introducing further errors.

But that means I spend a lot of time mucking about with some pretty messy files. In the lecture today, I want to impart some tips on things I've learned, doing so.

Challenges

Consistency

- file names
- file organization
- subject IDs
- variable names
- categorical data

Essentially all of the challenges come from inconsistencies: in file names, the arrangement of data within files, the subject identifiers, the variable names, and the categories in categorical data.

Code re-organizing data is the worst code.

Example file

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	B6 ob/ob x BTBR ob/ob						2.0mL RS-			2.0mL RS-			TT-1 bag			2.0mL RS-			TT-1 bag
2	Mouse ID	Date Born	Sac Date / Time	Sex	SVL Length (cm)	Hypothalamus(mg)	Hypothalamus weight(mg)	Hypothalamus Freezer Box	Brain	Brain weight(mg)	Brain Freezer Box	Left Liver	Liver weight(mg)	Liver Freezer Box	Rt. Kidney	Rt. Kidney weight(mg)	Rt. Kidney Freezer Box	Rt. Adipose	Rt. Ar
4	Mouse# 3002	6/2/05	8/15/05	F	10.0	RS-115943	8.2	1	RS-115942	391	1	RS-98275	413	1	RS-115948	246	1	RS-98271	530
5	Mouse# 3003	6/3/05	8/15/05	M	10.0	RS-115938	13.1	1	RS-115937	359	1	RS-98265	538	1	RS-115925	317	1	RS-98270	594
6	Mouse# 3004	6/3/05	8/15/05	M	9.3	RS-115815	13.5	1	RS-115814	365	1	RS-98277	654	1	RS-115820	324	1	RS-98272	670
7	Mouse# 3005	6/13/05	8/22/05	F	-	RS-115799	19.3	1	RS-115800	386	1	RS-98268	467	1	RS-115801	233	1	RS-98274	757
8	Mouse# 3006	6/13/05	8/22/05	F	9.5	RS-127305	11.7	1	RS-127304	384	1	RS-98258	498	1	RS-127303	233	1	RS-98257	676
9	Mouse# 3007	6/13/05	8/22/05	F	8.9	RS-127290	16.3	1	RS-127289	345	1	RS-98264	461	1	RS-127288	163	1	RS-98256	478
10	Mouse# 3008	6/13/05	8/22/05	F	10.3	RS-127275	19.7	1	RS-127274	422	1	RS-98259	465	1	RS-127273	299	1	RS-98255	742
11	Mouse # 3009	6/13/05	8/23/05	M	9.0	RS-126754	17.1	1	RS-126753	380	1	RS-98263	452	1	RS-126755	248	1	RS-98262	553
12	Mouse# 3010	6/13/05	8/23/05	M	10.2	RS-126744	20.6	1	RS-126745	395	1	RS-98261	657	1	RS-126740	331	1	RS-98276	496
13	Mouse# 3011	6/13/05	8/23/05	M	10.0	RS-127331	19.7	1	RS-127330	415	1	RS-98260	582	1	RS-127332	230	1	RS-98269	661
14	Mouse# 3012	6/13/05	8/23/05	M	10.7	RS-127341	17.6	1	RS-127340	418	1	RS-98273	431	1	RS-127338	278	1	RS-98254	629
15	Mouse# 3013	6/13/05	8/24/05	M	10.5	RS-126044	19	1	RS-126045	395	1	RS-97152	557	1	RS-126042	384	1	RS-97199	494
16	Mouse# 3014	6/13/05	8/24/05	M	9.4	RS-126024	16.6	1	RS-126022	362	1	RS-97189	401	1	RS-126020	214	1	RS-97196	604
17	Mouse# 3015	6/13/05	8/24/05	F	9.8	RS-126012	15.1	1	RS-126010	385	1	RS-97184	550	1	RS-126008	281	1	RS-97200	671
18	Mouse# 3016	6/13/05	8/24/05	F	9.0	RS-126000	15.1	1	RS-125998	386	1	RS-97194	463	1	RS-125996	223	1	RS-97195	693
19	Mouse# 3017	7/3/05	9/7/05	F	8.2	RS-125980	15.7	1	RS-125989	298	1	RS-97197	408	1	RS-125982	213	1	RS-97185	433
20	Mouse# 3018	7/3/05	9/7/05	F	9.0	RS-125979	15.1	1	RS-125977	363	1	RS-98278	591.3	1	RS-126168	199	1	RS-97201	676
21	Mouse# 3019	7/3/05	9/7/05	F	8.5	RS-126323	18.8	1	RS-126325	383	1	RS-97191	443.8	1	RS-126341	322	1	RS-97180	775

Here's an example file. Lots of work was done to prettify things, which means multiple header rows and a good amount of work to identify and pull out the essential columns.

Another example

	A	B	C	D	E	F	G	H	I	J	K	L
1									4 wk Orbital Eye Bleed			
2	Mouse ID	SEX	MHV status (+ or ?)	BIRTH DATE	SAC DATE	WEAN DATE	AGOUTI COAT (Y/N)	TUFT COAT (Y/N)	DATE	WEIGHT (g)	BODY LENGTH (cm)	GLUCOSE (mg/dl)
3	3001	F	Y	6/2/05	8/15/05	6/22/05	T	-	6/30/2005	23.1	75	637.351
4	3002	F	Y	6/2/05	8/15/05	6/22/05	T	-		22.8	80	261.842
5	3003	M	Y	6/3/05	8/15/05	6/22/05	T	-		24.1	80	124.065
6	3004	M	Y	6/3/05	8/15/05	6/22/05	B	-		21	78	254.393
7	3005	F	Y	6/13/05	8/22/05	6/30/05	T	Y	7/14/2005	22.3	78	116.15668
8	3006	F	Y	6/13/05	8/22/05	6/30/05	T	N		17.4	74	153.02296
9	3007	F	Y	6/13/05	8/22/05	6/30/05	T	N		13.6	68	99.39928
10	3008	F	Y	6/13/05	8/22/05	6/30/05	T	N		23.5	80	173.69042
11	3009	M	Y	6/13/05	8/23/05	6/30/05	T	N		19.3	75	123.41822
12	3010	M	Y	6/13/05	8/23/05	6/30/05	B	N		18.7	77	443.48456
13	3011	M	Y	6/13/05	8/23/05	6/30/05	B	N		24.6	79	162.51882
14	3012	M	Y	6/13/05	8/23/05	6/30/05	T	N		23.7	80	139.05846
15	3013	M	Y	6/13/05	8/24/05	6/30/05	T	N		28.5	80	226.75552
16	3014	M	Y	6/13/05	8/24/05	6/30/05	T	Y		13.6	68	96.0478
17	3015	F	Y	6/13/05	8/24/05	6/30/05	T	N				
18	3016	F	Y	6/13/05	8/24/05	6/30/05	T	N				
19	3017	F	Y	7/3/05	9/7/05	7/21/05	B	N	7/28/2005	9.8	66	234.7808
20	3018	F	Y	7/3/05	9/7/05	7/21/05	T	N		12.9	65	89.37385
21	3019	F	Y	7/3/05	9/7/05	7/21/05	T	N		12.5	65	155.8268
22	3020	F	Y	7/3/05	9/7/05	7/21/05	B	Y		15.9	70	80.8205
23	3021	F	Y	7/3/05	9/12/05	7/21/05	B	N		14.8	70	235.43875
24	3022	F	Y	7/3/05	9/12/05	7/21/05	T	N		19.9	71	469.66895
25	3023	M	Y	7/3/05	9/12/05	7/21/05	B	N		16.6	72	536.1219
26	3024	M	Y	7/3/05	9/12/05	7/21/05	T	Y		17.9	71	268.9942
27	3025	M	Y	7/3/05	9/13/05	7/21/05	T	N		16.6	71	230.17515
28	3026	M	Y	7/3/05	9/13/05	7/21/05	T	N		17.1	69	288.07475
29	3027	M	Y	7/3/05	9/13/05	7/21/05	B	N		13.1	69	124.2452
30	3028	M	Y	7/3/05	9/13/05	7/21/05	T	N		13.3	70	170.3017
31	3029	F	Y	7/9/05	9/20/05	7/27/05	T	N	8/4/2005	29	83	439.77188
32	3030	F	Y	7/9/05	9/20/05	7/27/05	T	N		28.1	83	438.51124
33	3031	M					T			30.2	85	664.79612
34	3032	M					T			30.4	85	403.21332
35	3033	F	Y	7/16/05	9/21/05	8/4/05	T	N	8/11/2005	19.5	77	274.8108
36	3034	F	Y	7/16/05	9/21/05	8/4/05	T	N		20.4	77	582.3402
37	3035	F	Y	7/16/05	9/21/05	8/4/05	T	N		18.6	75	461.0475
38	3036	F	Y	7/16/05	9/21/05	8/4/05	T	N		16.5	75	313.0132
39	3037	F	Y	7/16/05	9/22/05	8/4/05	T	N		18.3	78	121.5237

Here's a second worksheet from that file. Again, the header row has information on multiple lines that need to be merged. With the merged cells, it's hard to predict where they end up in a CSV file.

The format of individual identifiers is constantly changing.

Weird rounding

38.7	90	367.75144	12.271581150942	159.2311
37.5	89	404.04308	6.55818503449434	146.9497
41.9	90	218.343	9.55324086763758	101.9179
36	88	287.62704	4.65914900117792	91.0011
22.8	79	114.2122	32.46127	70.38872
20.8	75	166.4504	8.211126	60.96332
27.2	84	202.51284	13.1384923833842	105.07665
20.8	77	313.51314	11.1372217899707	93.32436
12.6	65	199.61718	16.7719514987531	66.61461
12.1	64	429.33954	18.9643060968415	49.52037
27.4	81	512.34846	4.31272238159915	101.51535
25.3	79	591.4965	9.70506442962546	186.98655
22	78	142.6692	14.9913480181089	53.79393
22.9	80	349.70889	17.0824838559225	180.93234
24.2	77	425.96127	5.77571495445421	151.72968
25.7	82	248.36079	14.3881991417965	99.37857
23.9	79	441.8874	17.1454129445892	70.17591
26.6	93	359.8437	11.3140598977232	152.79807
37.1	87	445.14312	10.4517	87.77684
35.3	85	183.7356	7.32103	67.86024
37.9	88	471.54792	11.8114	166.35688
37.4	87	142.80816	23.648	78.72284

Part of that file shows some weird rounding patterns. The font isn't even consistent. This suggests to me that there has been some copy-pasting of data, and that there may be some other set of files that is the primary source for the data.

Inconsistent IDs

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	mouse #	birthdate	sex	coat color	6 wk glu	6 wk ins	6 wk TG	10 wk glu	10 wk ins	10 wk TG	14 wk glu	14 wk ins	14 wk TG	GTT date	GTT weight	sac date	sac wk glu
2	121	10/15/14	F	agouti	149.37426	0.8442	139.2379	60.12283	0.6957333	120.88583	105.82285	0.2120998	211.87862	2/9/15	24.5		115.74088
3	122	10/15/14	F	agouti	95.326808	1.481575	202.05441	74.487115	0.7096667	132.7588	82.242928	0.5339661	121.14418	2/9/15	18.9		191.43122
4	123	10/15/14	F	agouti	97.490984	0.408725	79.373226	98.03989	0.7610667	142.69479	119.71168	0.6829993	93.352632	2/9/15	24.7		132.51577
5	124	10/15/14	F	agouti	116.96857	2.0537	143.44967	80.069995	1.3096333	145.20569	96.90912	1.4193986	141.42944	2/9/15	25.1		135.81992
6	125	10/15/14	F	white	108.0271	1.246475	125.88264	76.17361	0.6123667	98.07251	72.603664	0.5343661	101.70108	2/9/15	23.2		166.47222
7	126	10/15/															
8	127	10/15/															
9	128	10/15/															
10	129	10/15/															
11	130	10/15/															
12	131	10/15/															
13	132	10/15/															
14	133	10/15/															
15	134	10/15/															
16	135	10/15/															
17	136	10/15/															
18	137	10/15/															
19	138	10/15/															
20	139	10/15/															
21	140	10/15/															
22	141	10/15/															
23	142	10/15/															
24	143	10/15/															
25	144	10/15/															
26	145	10/15/															
27	146	10/15/															
28	147	10/15/															
29	148	10/15/															

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	mouse #	birthdate	sex	coat color	6 wk glu	6 wk ins	6 wk TG	10 wk glu	10 wk ins	10 wk TG	14 wk glu	14 wk ins	14 wk TG	GTT date	GTT weight	sac date	sac wk glu
2	DO-461	6/21/16	F	black	91.643808	0.35505	83.517195	93.594849	0.8989324	239.45556	80.501387	0.3877628	155.39943	10/17/16	20.2	11/14/16	88.70252
3	DO-462	6/21/16	F	agouti	111.6002	0.528125	138.46891	107.92265	0.3876329	114.35128	123.35268	0.2861638	185.66623	10/17/16	19	11/14/16	106.1972
4	DO-463	6/21/16	F	black	94.678414	0.934675	97.729902	99.024333	0.713366	113.64156	91.360961	1.1118889	119.85253	10/17/16	32.3	11/14/16	140.09932
5	DO-464	6/21/16	F	chinchilla	120.60115	2.176325	121.80574	111.79368	1.8336315	126.86816	142.72381	1.5440512	126.22905	10/17/16	40.3	11/14/16	129.6717
6	DO-465	6/21/16	F	agouti	90.820864	1.02785	95.218174	110.68767	2.4795309	173.7742	116.84672	2.436609	146.68582	10/17/16	37.4	11/16/16	142.96568
7	DO-466	6/21/16	F	agouti	112.16597	0.607675	80.270327	123.80892	0.7189993	106.12498	127.80413	0.5506278	64.195097	10/17/16	20.8	11/16/16	136.29398
8	DO-467	6/21/16	F	agouti	100.90193	1.07875	119.53906	114.65924	0.3764663	125.67454	104.07938	0.8151585	171.41285	10/17/16	24.7	11/16/16	117.46496
9	DO-468	6/21/16	F	agouti	93.701168	0.555725	73.163973	102.39262	0.641266	173.25804	105.20447	0.9074243	168.46984	10/17/16	24.6	11/16/16	121.61624
10	DO-469	6/21/16	F	black	100.90193	1.786925	183.68002	104.80573	2.303731	244.2623	105.0088	0.8191251	214.05758	10/17/16	21.6	11/18/16	118.00858
11	DO-470	6/21/16	F	agouti	98.587398	0.816475	97.178547	99.828701	0.3997663	84.8979	78.789292	0.3717629	80.323924	10/17/16	19.1	11/18/16	107.13618
12	DO-471	6/21/16	F	agouti	137.52294	1.016775	52.028698	107.67129	0.6544993	177.12924	113.22686	1.3451199	99.222639	10/19/16	33.8	11/18/16	144.2506
13	DO-472	6/21/16	F	white	102.80499	1.1494	109.36962	123.6581	0.5479661	229.48722	93.513309	1.2255211	284.14152	10/19/16	24.1	11/18/16	108.47052
14	DO-473	6/21/16	F	white	94.36981	0.76645	73.102711	143.46567	0.4791662	78.67172	141.59872	0.5927274	69.388637	10/19/16	20.6	11/22/16	128.13968
15	DO-474	6/21/16	F	agouti	110.98299	1.415925	62.320658	92.9413	0.8363658	86.41412	113.6182	0.4423956	74.582177	10/19/16	20.4	11/22/16	108.71762
16	DO-475	6/21/16	F	black	86.243238	0.78605	96.872239	95.052766	0.5956661	62.34816	93.611143	0.3843295	80.035394	10/19/16	18.5	11/30/16	134.91022
17	DO-476	6/21/16	F	agouti	136.90573	0.979725	117.51742	118.98271	0.3497997	134.35248	161.99711	0.836625	109.3789	10/19/16	25.2	11/30/16	119.19466
18	DO-477	6/21/16	F	agouti	128.31625	0.69315	249.35253	112.54777	0.7935992	233.4552	138.41912	0.8584914	234.08156	10/19/16	26.1	11/30/16	135.55268
19	DO-478	6/21/16	F	agouti	115.81779	0.4010329	48.843091	109.43084	0.2675997	95.02754	132.74474	0.2432976	91.6343	10/19/16	21.2	11/30/16	120.13364
20	DO-479	6/21/16	F	agouti	113.60613	1.382075	114.88317	105.00682	1.953098	141.25612	113.56928	1.3259534	132.37474	10/19/16	33.3	12/2/16	145.38726
21	DO-480	6/21/16	F	black	167.09749	2.2408	57.297201	123.80892	2.5369641	122.93244	136.3646	1.6026506	128.53729	10/19/16	31.4	12/2/16	156.95154
22	DO-481	6/21/16	F	agouti	105.53099	0.478775	64.893648	110.23521	0.1381332	73.7682	113.37361	0.6286604	74.495618	10/21/16	27.3	12/2/16	123.88956
23	DO-482	6/21/16	F	agouti	101.98204	0.820925	82.782055	90.829834	0.5752994	84.96242	103.34563	0.2304644	87.969969	10/21/16	21.7	12/2/16	113.70904
24	DO-483	6/21/16	F	agouti	82.951462	0.3453	78.493738	95.404677	0.5566661	101.44728	90.089119	1.5080183	107.67657	10/21/16	21.4	12/6/16	93.15032
25	DO-484	6/21/16	F	agouti	126.41319	0.67715	98.281257	100.18061	0.9220991	139.80442	114.44979	1.3265201	154.67811	10/21/16	28.2	12/6/16	132.3898
26	DO-485	6/21/16	F	agouti	93.752602	1.6095	90.868595	89.371917	0.675566	86.5109	83.045071	0.3703296	102.85812	10/21/16	25.4	12/6/16	98.98188
27	DO-486	6/21/16	F	agouti	100.90193	0.64165	83.578457	102.94563	0.7815659	80.31698	103.63913	0.6679933	88.200793	10/21/16	24.4	12/6/16	114.99396
28	DO-487	6/21/16	F	agouti	113.19465	0.318025	71.019815	96.108499	0.5215661	151.48254	125.26044	0.3840295	125.70969	10/21/16	24.7	12/8/16	128.73272
29	DO-488	6/21/16	F	agouti	91.695242	0.5937	115.12822	104.05163	0.8984324	205.51804	93.904645	0.5686943	129.22976	10/21/16	23.2	12/8/16	87.4176
30	DO-489	6/21/16	F	agouti	50.496608	0.385025	73.04145	72.932646	0.5427661	100.44722	98.796345	0.8198585	56.058551	10/21/16	21	12/8/16	86.77514

The format of the IDs is different between these files. Also in one of the files, there are missing dates that will need to be grabbed from some separate file.

Inconsistent layout

	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg	insulin ng/ml
2	DO-121	2/9/15	24.5	0	99.165552	lo off curve
3				5	349.30355	0.2052
4				15	286.09221	0.12895
5				30	312.0477	0.17545
6				60	99.871824	0.12165
7				120	217.93696	lo off curve
8	DO-122	2/9/15	18.9	0	185.80158	0.25145
9				5	297.39256	2.2281
10				15	439.0001	2.0778
11				30	362.25187	0.7746
12				60	232.65096	0.50015
13				120	260.72527	0.5234
14	DO-123	2/9/15	24.7	0	198.45562	0.15135
15				5	530.63889	lo off curve
16				15	614.15555	0.62425
17				30	647.46805	0.12085
18				60	531.05088	0.19775
19				120	388.0308	0.1853

	A	B	C	D
1	DO-221	0	145.74279	0.74455
2		5	206.45264	2.0264
3		15	216.64061	1.13205
4		30	299.55501	0.78475
5		60	242.65912	0.3326
6		120	186.23344	0.53575
7	DO-222	0	138.01038	0.70715
8		5	342.86694	1.1049
9		15	339.83668	0.8284
10		30	276.1488	0.5935
11		60	248.30168	0.4905
12		120	303.42121	1.0419
13	DO-223	0	138.21936	1.1223
14		5	407.443	2.1029
15		15	336.85865	1.8585
16		30	235.50141	1.50985
17		60	246.21184	0.86705
18		120	247.62249	0.89315

Another example of inconsistent layout. And messy, in both cases. You need to fill in the repeated values like mouse ID, and the column names are missing in the file on the right.

All kinds of inconsistencies

	A	B	C	D	E	F	G	H
1	date	mouse #	weight	heart	L liver lobe	remaining liver	R fat pad	L fat pad
2	3/9/15	121	26.7	0.136	0.325	0.655	0.383	0.317
3								
4								
5								
6	3/10/15							
7								
8								
9								
10	3/11/15							
11								
12								
13								
14	3/12/15							
15								
16								
17								

	A	B	C	D	E	F	G	H
1	mouse num	date	weight	heart	L liver lobe	remaining l	R fat pad	L fat pad
2	DO-221	7/20/15	24.1	0.136	0.339	0.743	0.289	0.262
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							
12	DO-231							
13	DO-232							
14	DO-233							
15	DO-234							
16	DO-235							
17	DO-236							

	A	B	C	D	E	F	G	H
1	mouse num	date	weight	heart	L liver lobe	remaining l	R fat pad	L fat pad
2	321	2/11/16	50.1	0.171	0.515	1.37	3.03	3.28
3	322							
4	323							
5	324							
6	325							
7	326							
8	327							
9	328							
10	329							
11	330							
12	331							
13	332							
14	333							
15	335							
16	336							
17	337							

	A	B	C	D	E	F	G	H
1	mouse nu	date	weight	heart	L liver lobe	remaining l	R fat pad	L fat pad
2	DO461	11/14/16	20.3	0.106	0.259	0.505	0.23	0.248
3	DO462	11/14/16	20.6	0.107	0.283	0.521	0.211	0.223
4	DO463	11/14/16	36.2	0.161	0.505	1.066	1.01	1.2
5	DO464	11/14/16	45.9	0.18	0.447	1.18	1.78	1.41
6	DO511	11/15/16	35.1	0.151	0.471	1.064	0.7	0.699
7	DO512	11/15/16	27.2	0.148	0.308	0.707	0.155	0.161
8	DO513	11/15/16	29.9	0.168	0.422	0.905	0.493	0.597
9	DO514	11/15/16	33.6	0.161	0.413	0.851	0.873	0.74
10	DO465	11/16/16	36.4	0.165	0.498	1.09	1.36	1.42
11	DO466	11/16/16	21.4	0.0989	0.254	0.601	0.375	0.39
12	DO467	11/16/16	26.3	0.154	0.47	0.936	0.291	0.225
13	DO468	11/16/16	25.9	0.151	0.311	0.88	0.244	0.212
14	DO515	11/17/16	45.9	0.156	0.474	1.09	2.09	1.81
15	DO516	11/17/16	34.5	0.197	0.502	1.1	0.856	0.861
16	DO517	11/17/16	41.6	0.184	0.561	1.12	1.15	0.981
17	DO518	11/17/16	41.8	0.185	0.497	1.14	1.26	1.25

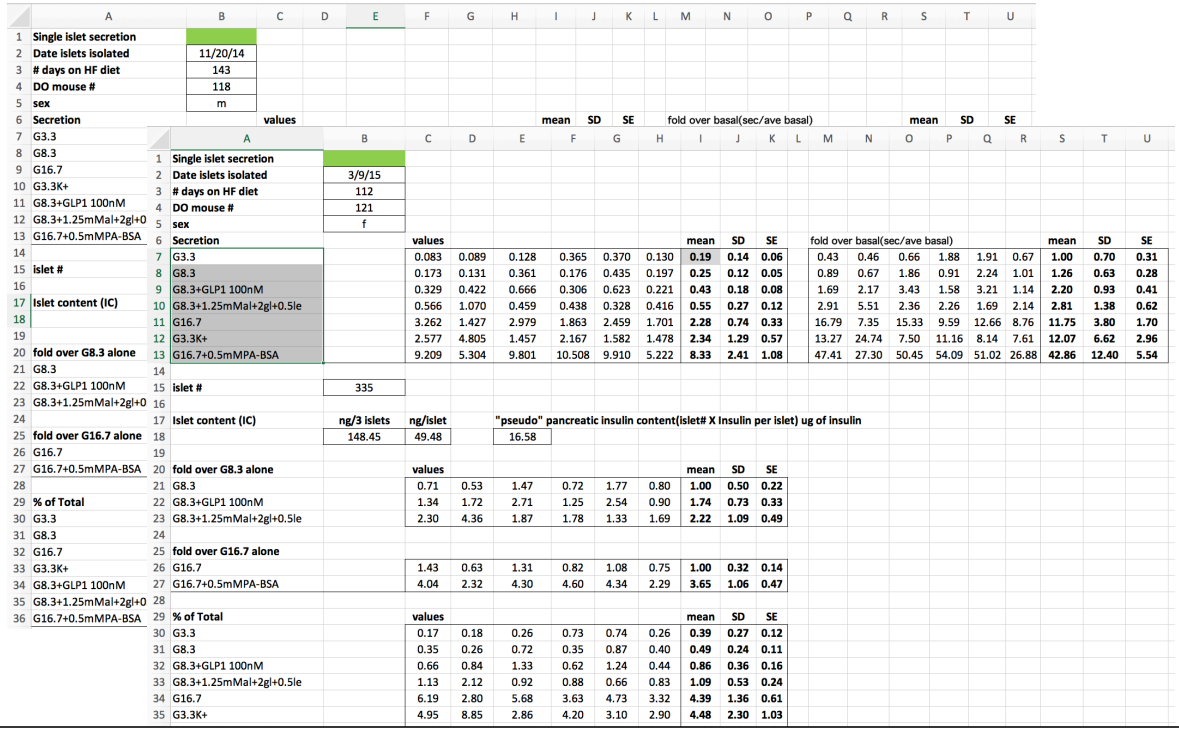
The layouts, IDs, and included information are all inconsistent here.

Multiple rectangles

	A	B	C	D	E	F	G	H	I	J	K	L
1	Wave 2 ID	Adiponectin (ug/mL)	collection date	BW	sex			Wave 1 ID	Adiponectin (ug/mL)	collection date	BW	sex
2	DO-121	25.28521548	3/9/15	26.7	F			DO-21	58.70791021	10/20/14	21.1	F
3	DO-122	8.589388212	3/9/15	19.3	F			DO-22	6.141839632	10/20/14	30.4	F
4	DO-123	16.45348107	3/9/15	28.2	F			DO-23	37.34270189	10/20/14	29.9	F
5	DO-124	22.86891765	3/9/15	26.4	F			DO-24	5.805316486	10/20/14	21.1	F
6	DO-125	37.13273594	3/11/15	24.6	F			DO-25	5.48942198	10/22/14	22.9	F
7	DO-126	18.76181517	3/11/15	31	F			DO-26	7.550740533	10/22/14	29.4	F
8	DO-128	11.50813114	3/11/15	23.9	F			DO-27	7.633411071	10/22/14	26.6	F
9	DO-129	7.447558701	3/11/15	22.6	F			DO-28	0.049261069	10/22/14	24.6	F
10	DO-130	10.48386039	3/13/15	25.9	F			DO-30	8.841227011	10/24/14		F
11	DO-131	8.471601718	3/13/15	25.6	F			DO-31	8.170986006	10/24/14	26.6	F
12	DO-132	3.04690223	3/13/15	27.4	F			DO-32	12.67835566	10/24/14	24.6	F
13	DO-133	0.099577938	3/13/15	24.8	F			DO-33	17.75682222	10/24/14	34.2	F
14	DO-137	11.20577459	3/17/15	27.7	F			DO-34	24.29713573	10/28/14	28.9	F
15	DO-138	12.72099796	3/17/15	20	F			DO-35	11.74448642	10/28/14	19.7	F
16	DO-140	23.68048642	3/17/15	22.3	F			DO-36	9.310303972	10/28/14	22.6	F
17	DO-141	14.64889349	3/17/15	26.2	F			DO-37	18.45679929	10/28/14	34.3	F
18	DO-142	42.30217756	3/19/15	37.8	F			DO-38	65.906108	10/30/14	34.1	F
19	DO-143	14.54807857	3/19/15	22.8	F			DO-39	55.95587133	10/30/14	30.8	F
20	DO-144	10.57159252	3/19/15	28.7	F			DO-40	20.5376597	10/30/14	29.6	F
21	DO-145	9.465243507	3/19/15	33.5	F			DO-41	26.11849635	10/30/14	21.4	F
22	DO-146	6.278729256	3/23/15	23.1	F			DO-42	14.58745555	11/3/14	27.4	F
23	DO-147	4.894797158	3/23/15	26.6	F			DO-43	21.77644658	11/3/14	33.3	F
24	DO-148	11.33704889	3/23/15	25.8	F			DO-44	12.48999428	11/3/14	25.4	F

Here's an example where they have a group of columns with one set of data, a few blank columns, then a group of columns with another set of data.

Stuff moving around



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	**Single islet secretion**																				
2	**Date islets isolated**	11/20/14																			
3	**# days on HF diet**	143																			
4	**DO mouse #**	118																			
5	**sex**	m																			
6	**Secretion**		**values**						**mean**	**SD**	**SE**		fold over basal(sec/ave basal)						**mean**	**SD**	**SE**
7	G3.3																				
8	G8.3																				
9	G16.7																				
10	G3.3K+																				
11	G8.3+GLP1 100nM																				
12	G8.3+1.25mMal+2gl+0																				
13	G16.7+0.5mMPA-BSA																				
14																					
15	**islet #**																				
16																					
17	**Islet content (IC)**																				
18																					
19																					
20	**fold over G8.3 alone**																				
21	G8.3																				
22	G8.3+GLP1 100nM																				
23	G8.3+1.25mMal+2gl+0																				
24																					
25	**fold over G16.7 alone**																				
26	G16.7																				
27	G16.7+0.5mMPA-BSA																				
28																					
29	**% of Total**																				
30	G3.3																				
31	G8.3																				
32	G16.7																				
33	G3.3K+																				
34	G8.3+GLP1 100nM																				
35	G8.3+1.25mMal+2gl+0																				
36	G16.7+0.5mMPA-BSA																				

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	**Single islet secretion**																				
2	**Date islets isolated**	3/9/15																			
3	**# days on HF diet**	112																			
4	**DO mouse #**	121																			
5	**sex**	f																			
6	**Secretion**		**values**										**mean**	**SD**	**SE**		fold over basal(sec/ave basal)						**mean**	**SD**	**SE**
7	G3.3		0.083	0.089	0.128	0.365	0.370	0.130					**0.19**	**0.14**	**0.06**		0.43	0.46	0.66	1.88	1.91	0.67	**1.00**	**0.70**	**0.31**
8	G8.3		0.173	0.131	0.361	0.176	0.435	0.197					**0.25**	**0.12**	**0.05**		0.89	0.67	1.86	0.91	2.24	1.01	**1.26**	**0.63**	**0.28**
9	G8.3+GLP1 100nM		0.329	0.422	0.666	0.306	0.623	0.221					**0.43**	**0.18**	**0.08**		1.69	2.17	3.43	1.58	3.21	1.14	**2.20**	**0.93**	**0.41**
10	G8.3+1.25mMal+2gl+0.5le		0.566	1.070	0.459	0.438	0.328	0.416					**0.55**	**0.27**	**0.12**		2.91	5.51	2.36	2.26	1.69	2.14	**2.81**	**1.38**	**0.62**
11	G16.7		3.262	1.427	2.979	1.863	2.459	1.701					**2.28**	**0.74**	**0.33**		16.79	7.35	15.33	9.59	12.66	8.76	**11.75**	**3.80**	**1.70**
12	G3.3K+		2.577	4.805	1.457	2.167	1.582	1.478					**2.34**	**1.29**	**0.57**		13.27	24.74	7.50	11.16	8.14	7.61	**12.07**	**6.62**	**2.96**
13	G16.7+0.5mMPA-BSA		9.209	5.304	9.801	10.508	9.910	5.222					**8.33**	**2.41**	**1.08**		47.41	27.30	50.45	54.09	51.02	26.88	**42.86**	**12.40**	**5.54**
14																					
15	**islet #**	335																			
16																					
17	**Islet content (IC)**	**ng/3 islets**	**ng/islet**		**"pseudo" pancreatic insulin content(islet# X Insulin per islet) ug of insulin**																
18		148.45	49.48		16.58																
19																					
20	**fold over G8.3 alone**		**values**										**mean**	**SD**	**SE**						
21	G8.3		0.71	0.53	1.47	0.72	1.77	0.80					**1.00**	**0.50**	**0.22**						
22	G8.3+GLP1 100nM		1.34	1.72	2.71	1.25	2.54	0.90					**1.74**	**0.73**	**0.33**						
23	G8.3+1.25mMal+2gl+0.5le		2.30	4.36	1.87	1.78	1.33	1.69					**2.22**	**1.09**	**0.49**						
24																					
25	**fold over G16.7 alone**																				
26	G16.7		1.43	0.63	1.31	0.82	1.08	0.75					**1.00**	**0.32**	**0.14**						
27	G16.7+0.5mMPA-BSA		4.04	2.32	4.30	4.60	4.34	2.29					**3.65**	**1.06**	**0.47**						
28																					
29	**% of Total**		**values**										**mean**	**SD**	**SE**						
30	G3.3		0.17	0.18	0.26	0.73	0.74	0.26					**0.39**	**0.27**	**0.12**						
31	G8.3		0.35	0.26	0.72	0.35	0.87	0.40					**0.49**	**0.24**	**0.11**						
32	G8.3+GLP1 100nM		0.66	0.84	1.33	0.62	1.24	0.44					**0.86**	**0.36**	**0.16**						
33	G8.3+1.25mMal+2gl+0.5le		1.13	2.12	0.92	0.88	0.66	0.83					**1.09**	**0.53**	**0.24**						
34	G16.7		6.19	2.80	5.68	3.63	4.73	3.32					**4.39**	**1.36**	**0.61**						
35	G3.3K+		4.95	8.85	2.86	4.20	3.10	2.90					**4.48**	**2.30**	**1.03**						

This has a super-complicated layout, has 500 worksheets with one mouse each, and the order of things aren't entirely consistent.

Being self-sufficient

- C
- Perl (or python or ruby or R)
- R

I've long said that for a data scientist to be self-sufficient, they should be savvy with multiple programming languages.

R for data analysis. C for when you need high-performance, and for like 15 years I used Perl for manipulating data files plus shell scripting.

But I'd now say use Python or Ruby instead of Perl; probably Python. And really, I think we can now do everything we want to do in R. It's a perfectly sufficient general programming language.

Key techniques

- stepping through a file
- regular expressions
  - search and replace patterns
- parsing individual lines in a file
- matching vectors
- construct meta data
- system calls

Here are some of the key techniques that I use to wrangle messy data files. I'll go through many of these in more detail.

Stepping through a file in R

```
filecon <- open("huge_data.txt", "r")
while(TRUE) {

  line <- readLines(filecon, n=1)
  if( grepl("^\\[Data\\]", line) )  break

}

data <- readLines(filecon)
close(filecon)
```


I often use `readLines()` to slurp up an entire file as a vector of character strings.

For really big files, I might want to read one or a few lines at a time instead, and throw out all but the stuff that matters. You can do this in R!

This particular example will skip over a header that ends with [Data] and then read in everything after that.

Regular expressions

`grep()`, `grepl()`, `sub()`, `gsub()`

- `^` and `$` match the beginning and end of a line
- `[034]` for any one of several things; `[0-9]` for a range
- `[^034]` for something other than this set of things
- `\s` for white space
- `.` match any one character
- `+` match the last bit 1 or more times
- `*` match the last bit 0 or more times
- parentheses to group bits for use with `+` and `*`
- when substituting, can use `\1`, `\2`, ... in place of matched groups
- In R, most backslashes need to be made double-backslashes.

Regular expression are hugely useful for the data wrangling work.

Parsing strings

- I use a lot of `strsplit()`
- The output is a list of vectors so is not pretty
- Also look at the `stringr` package
- To put things back together, use `paste()`, `paste0()`, or the `glue` package.

Messing about with strings is not as easy as in perl, python or ruby, but it is more and more do-able.

Matching vectors

- I spend a lot of time matching two vectors, say of subject IDs
- I mostly use `match()`, eg `match(old_ids, new_ids)`
- Check for `NA`s, which indicate unmatched values
- May want to check that the values on right are unique
- Often do something like `olddata[ match(new_ids, old_ids), ]`

Matching IDs is a constant; I mostly use `match()`. Don't assume that the values are unique or that they're all present.

I often use this to reorder the rows or columns in one data set to match the rows or columns in another data set.

Construct meta data

	A	B	C	D	E
1	short_name	file	from_column	id_column	column_offset
2	mouse	Attie_DO_mice_wave2_s	mouse #	1	0
3	sex	Attie_DO_mice_wave2_s	sex	1	0
4	sac_date	Attie_DO_mice_wave2_s	sac date	1	0
5	coat_color	Attie_DO_mice_wave2_s	coat color	1	0
6	oGTT_date	Attie_DO_mice_wave2_s	GTT date	1	0
7	diet_days	ex_vivo_waves1-3.csv	Days.on.Diet	1	0
8	num_islets	ex_vivo_waves1-3.csv	num_islets	1	0
9	Ins_per_islet	ex_vivo_waves1-3.csv	IC	1	0
10	Glu_0min	gtt2.csv	glucose.mg.dl.0	2	0
11	Ins_0min	gtt2.csv	insulin.ng.ml.0	2	0
12	Glu_tAUC	gtt2.csv	glucose.mg.dl.tAUC	2	0
13	Glu_iAUC	gtt2.csv	glucose.mg.dl.iAUC	2	0
14	Ins_tAUC	gtt2.csv	insulin.ng.ml.tAUC	2	0
15	Ins_iAUC	gtt2.csv	insulin.ng.ml.iAUC	2	0
16	Glu_6wk	Attie_DO_mice_wave2_s	6 wk glu	1	0
17	Ins_6wk	Attie_DO_mice_wave2_s	6 wk ins	1	0
18	TG_6wk	Attie_DO_mice_wave2_s	6 wk TG	1	0
19	Glu_10wk	Attie_DO_mice_wave2_s	10 wk glu	1	0
20	Ins_10wk	Attie_DO_mice_wave2_s	10 wk ins	1	0
21	TG_10wk	Attie_DO_mice_wave2_s	10 wk TG	1	0
22	Glu_14wk	Attie_DO_mice_wave2_s	14 wk glu	1	0
23	Ins_14wk	Attie_DO_mice_wave2_s	14 wk ins	1	0
24	TG_14wk	Attie_DO_mice_wave2_s	14 wk TG	1	0
25	oGTT_weight	Attie_DO_mice_wave2_s	GTT weight	1	0
26	Glu_sac	Attie_DO_mice_wave2_s	sac wk glu	1	0
27	Ins_sac	Attie_DO_mice_wave2_s	sac wk ins	1	0
28	TG_sac	Attie_DO_mice_wave2_s	sac wk TG	1	0
29	food_1wk	Attie_DO_mice_wave2_s	11/17/14	1	2
30	food_2wk	Attie_DO_mice_wave2_s	11/24/14	1	2
31	food_3wk	Attie_DO_mice_wave2_s	12/1/14	1	2
32	food_4wk	Attie_DO_mice_wave2_s	12/8/14	1	2

This is an example meta data file that I set up because each batch of subjects in a project had the data organized completely differently.

I identified the variables of interest and chose a fixed set of names.

Then for each batch, I worked out which file it was in, the name of the column to look for, and then I also needed a column "offset" because to find week 4 measurements for variable X, you might look for the column that is two to the right from the one labeled Y.

R challenges

- `stringsAsFactors`
- `check.names` in `read.csv()`
- dealing with factors
  - levels
  - converting to/from strings
- Consider the `forcats` package

Categorical data is a problem not just for R, but R does have a lot of pain points.

Further tips

- Avoid using numeric indices
  - refer to data by variable name and individual ID
  - this will be more robust
- `stopifnot()` to assert things that should be true
- `cbind` and `rbind`, but padding with missing values
- Sometimes converting excel $\rightarrow$ csv loses precision
- `get()` to grab an object from a character string with its name
- `eval(parse( ))` to evaluate a charcter string as R code

Here are some misc. further tips.

Verify everything

- subject IDs unique?
- identifiers that don't match the typical pattern?
- subjects in one file but not in another?
- re-calculate and verify any derived values (like ratios)
- data repeated in multiple files the same?

You really should verify everything. Don't trust that things match; everything that can go wrong, will.

Reproducible reports

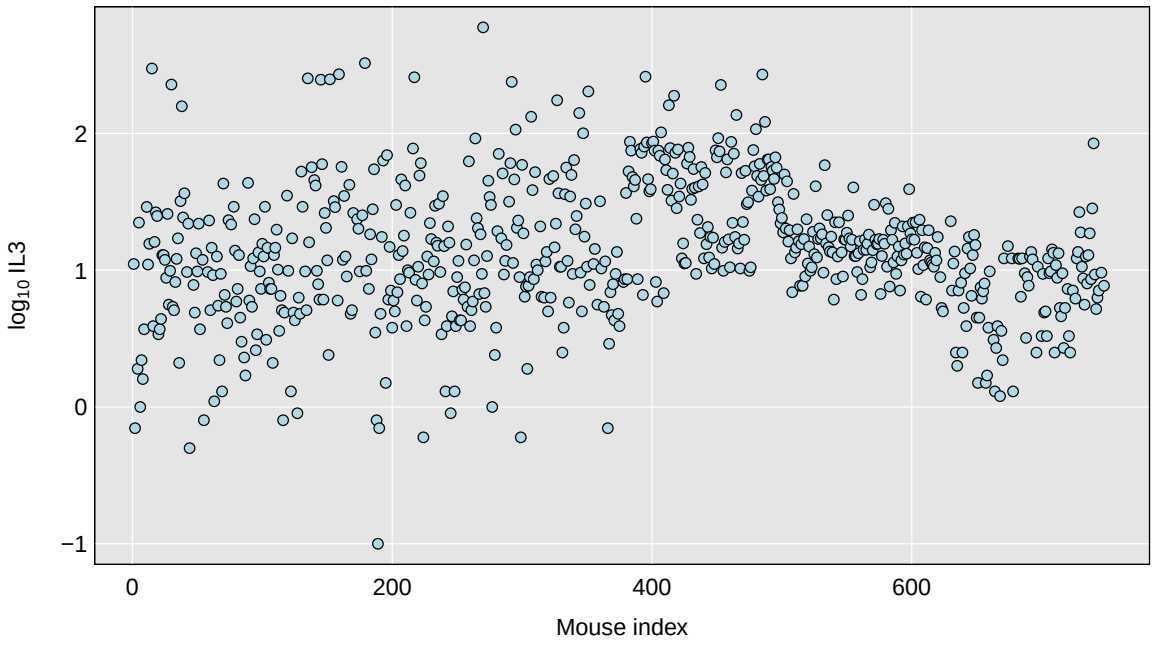
- You want all of this work to be reproducible
- Consider combining the data reorganization with the data cleaning
  - a lot of double-checking is happening when reorganizing
- Or clean each file one at a time
  - do the detailed diagnostics and cross-checks with data that are in a more convenient form
- Include diagnostic plots
  - Plot stuff vs time or by batch
  - Scatterplots of different variables
  - Consider taking logs
  - Look at missing data pattern
- Explain your thought process and observations

It is critical that the data re-formatting process be reproducible. You are very likely to be doing it more than once.

You often are doing data diagnostics and reformatting at the same time. Maybe have one long R Markdown or Jupyter document that does both? Or maybe it's best to rearrange each file, one at a time, and then do the serious diagnostics after you've gotten it all into a simpler form.

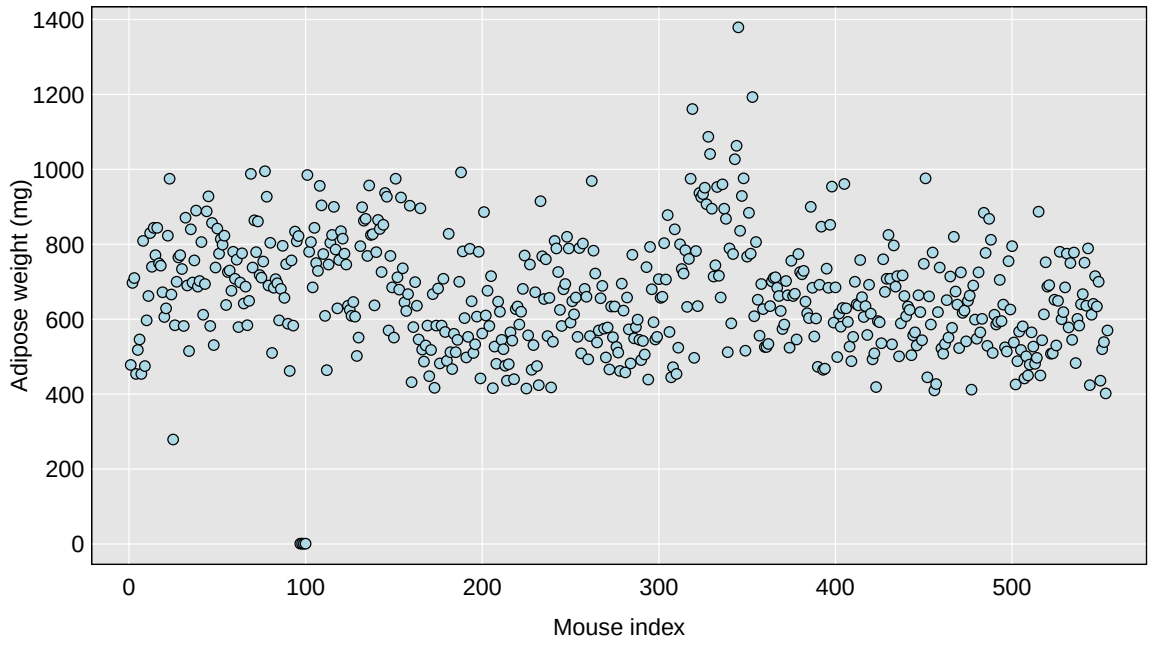
For diagnostics, you want to plot each variable against time, or by bath. And make scatterplots of different variables. The missing data pattern is also often quite informative about problems.

Batch effect



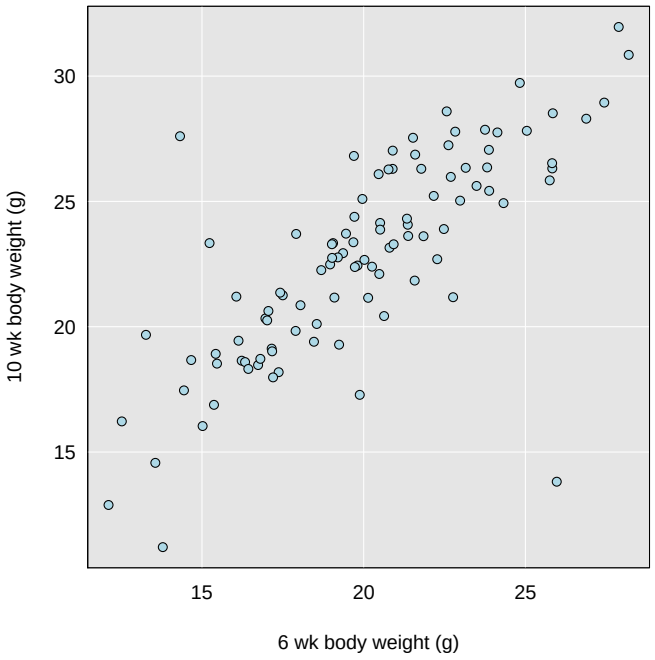
Here's an example of a clear batch effect. You can really only tell if you plot the variable by the order of measurement, and it's much more clear if you take logs.

Messed up units



Here's a case where a variable was recorded in the wrong units (g rather than mg) for a few individuals)

Outliers



In this particular case, it turned out that the day 10 weights for two subjects got swapped.

When you look at this sort of situation, ask yourself how you might find this problem if you have 20 weight measurements and 1500 individuals.

Summary

- Be prepared for anything
- Double-check everything
- Take your time and keep things organized
- Python is a good skill to have, but you can just do R

Summaries are always helpful.