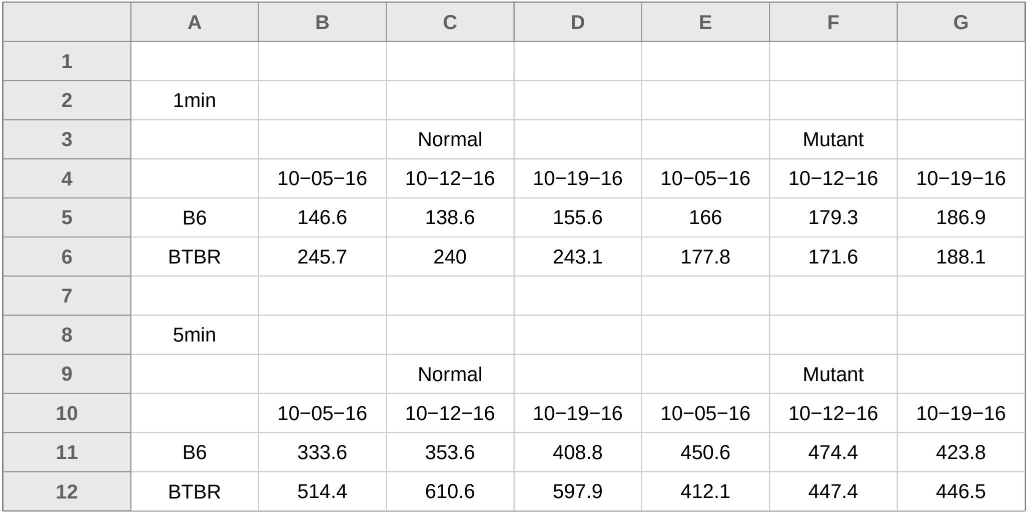
data organization in spreadsheets

Karl Broman

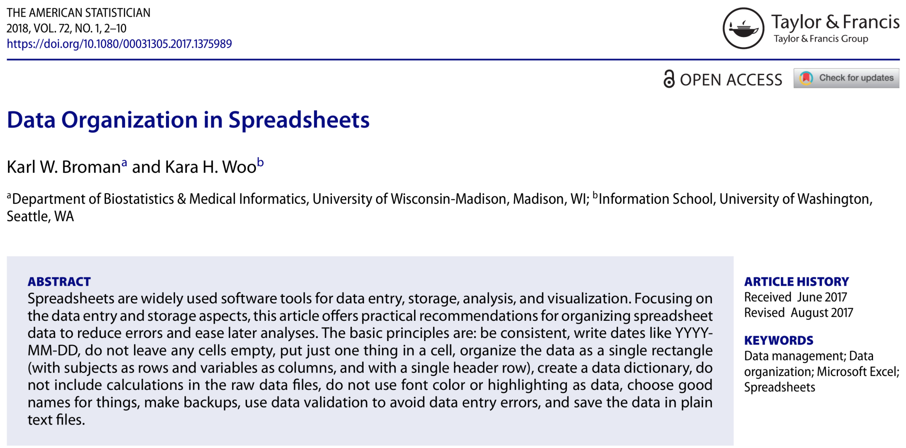
Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman

Slides: kbroman.org/Talk_DataOrg



	A	B	C	D	E	F	G
1							
2	1min						
3			Normal			Mutant	
4		10−05−16	10−12−16	10−19−16	10−05−16	10−12−16	10−19−16
5	B6	146.6	138.6	155.6	166	179.3	186.9
6	BTBR	245.7	240	243.1	177.8	171.6	188.1
7							
8	5min						
9			Normal			Mutant	
10		10−05−16	10−12−16	10−19−16	10−05−16	10−12−16	10−19−16
11	B6	333.6	353.6	408.8	450.6	474.4	423.8
12	BTBR	514.4	610.6	597.9	412.1	447.4	446.5



THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

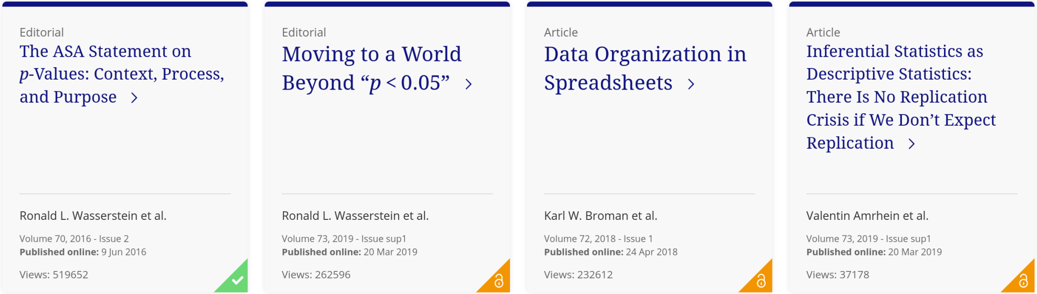
ABSTRACT
Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

doi.org/gdz6cm

American Statistician



Editorial

The ASA Statement on *p*-Values: Context, Process, and Purpose ›

Ronald L. Wasserstein et al.

Volume 70, 2016 - Issue 2

Published online: 9 Jun 2016

Views: 519652

Editorial

Moving to a World Beyond "$p < 0.05$" ›

Ronald L. Wasserstein et al.

Volume 73, 2019 - Issue sup1

Published online: 20 Mar 2019

Views: 262596

Article

Data Organization in Spreadsheets ›

Karl W. Broman et al.

Volume 72, 2018 - Issue 1

Published online: 24 Apr 2018

Views: 232612

Article

Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication ›

Valentin Amrhein et al.

Volume 73, 2019 - Issue sup1

Published online: 20 Mar 2019

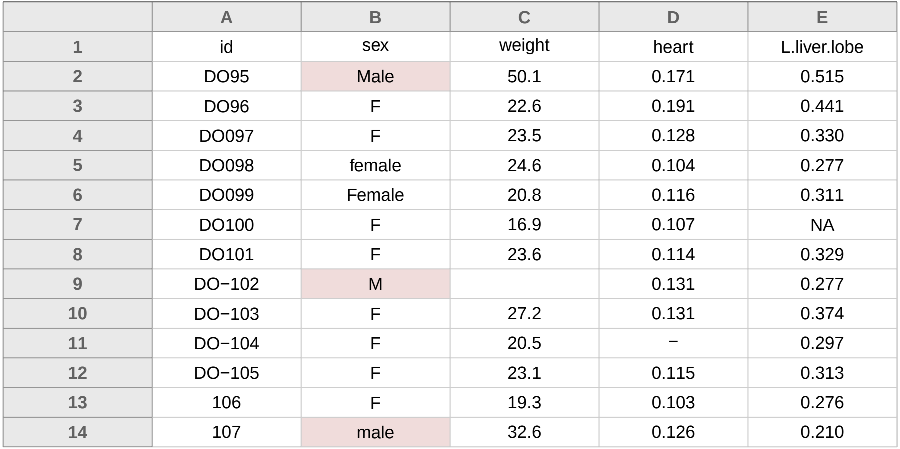
Views: 37178

`bit.ly/amstat_most_read`

Be consistent

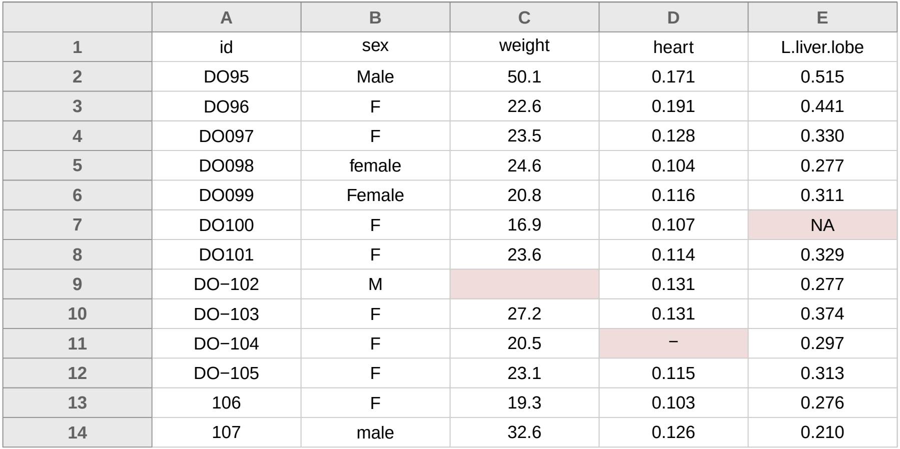
Be consistent

Consistent categories



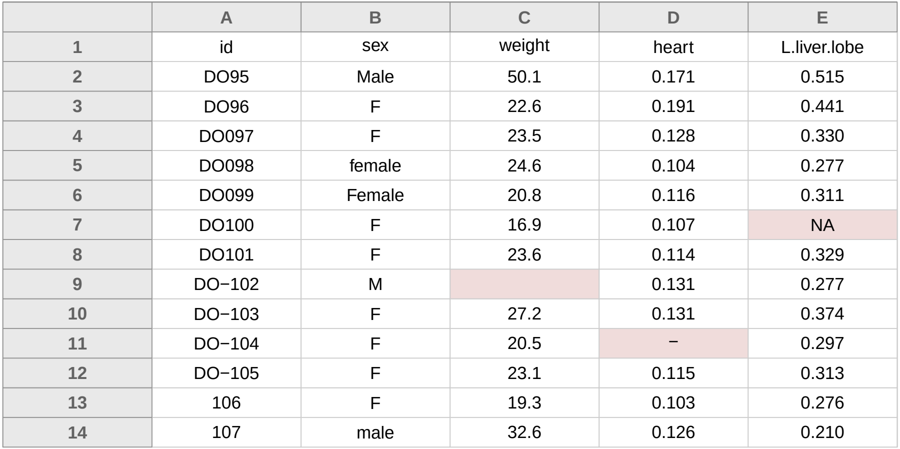
	A	B	C	D	E
1	id	sex	weight	heart	L.liver.lobe
2	DO95	Male	50.1	0.171	0.515
3	DO96	F	22.6	0.191	0.441
4	DO097	F	23.5	0.128	0.330
5	DO098	female	24.6	0.104	0.277
6	DO099	Female	20.8	0.116	0.311
7	DO100	F	16.9	0.107	NA
8	DO101	F	23.6	0.114	0.329
9	DO−102	M		0.131	0.277
10	DO−103	F	27.2	0.131	0.374
11	DO−104	F	20.5	–	0.297
12	DO−105	F	23.1	0.115	0.313
13	106	F	19.3	0.103	0.276
14	107	male	32.6	0.126	0.210

Consistent missing values



	A	B	C	D	E
1	id	sex	weight	heart	L.liver.lobe
2	DO95	Male	50.1	0.171	0.515
3	DO96	F	22.6	0.191	0.441
4	DO097	F	23.5	0.128	0.330
5	DO098	female	24.6	0.104	0.277
6	DO099	Female	20.8	0.116	0.311
7	DO100	F	16.9	0.107	NA
8	DO101	F	23.6	0.114	0.329
9	DO−102	M		0.131	0.277
10	DO−103	F	27.2	0.131	0.374
11	DO−104	F	20.5	−	0.297
12	DO−105	F	23.1	0.115	0.313
13	106	F	19.3	0.103	0.276
14	107	male	32.6	0.126	0.210

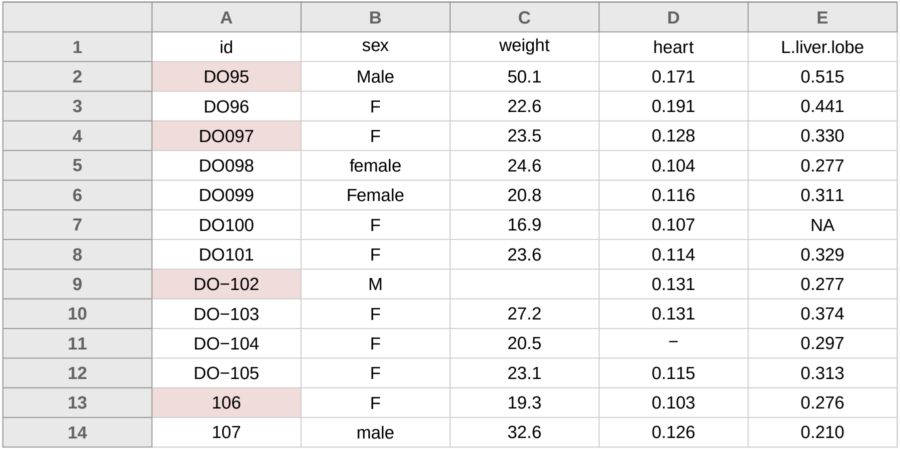
Consistent missing values



	A	B	C	D	E
1	id	sex	weight	heart	L.liver.lobe
2	DO95	Male	50.1	0.171	0.515
3	DO96	F	22.6	0.191	0.441
4	DO097	F	23.5	0.128	0.330
5	DO098	female	24.6	0.104	0.277
6	DO099	Female	20.8	0.116	0.311
7	DO100	F	16.9	0.107	NA
8	DO101	F	23.6	0.114	0.329
9	DO−102	M		0.131	0.277
10	DO−103	F	27.2	0.131	0.374
11	DO−104	F	20.5	–	0.297
12	DO−105	F	23.1	0.115	0.313
13	106	F	19.3	0.103	0.276
14	107	male	32.6	0.126	0.210

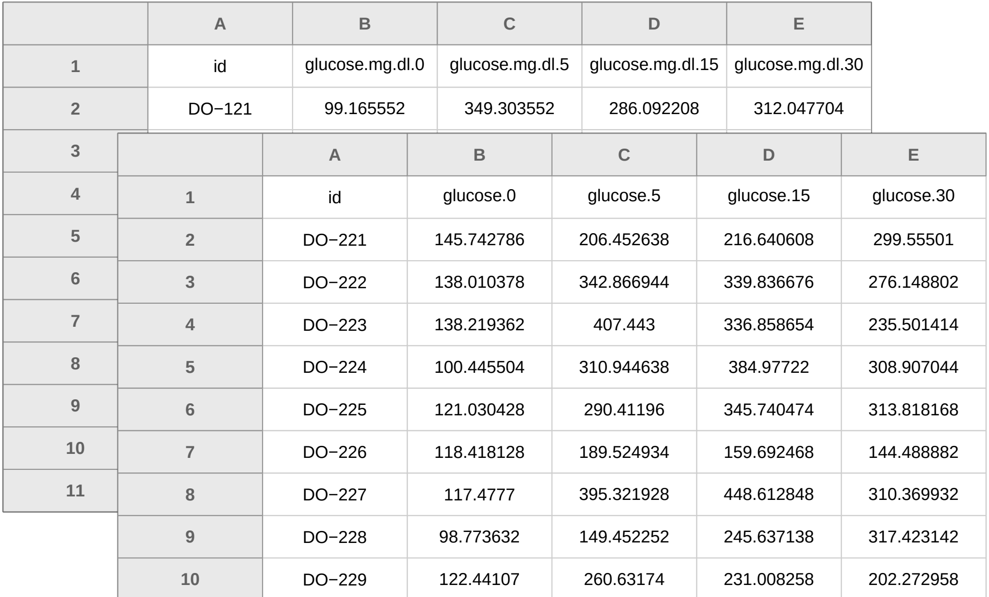
And no 999 or −999!

Consistent subject IDs



	A	B	C	D	E
1	id	sex	weight	heart	L.liver.lobe
2	DO95	Male	50.1	0.171	0.515
3	DO96	F	22.6	0.191	0.441
4	DO097	F	23.5	0.128	0.330
5	DO098	female	24.6	0.104	0.277
6	DO099	Female	20.8	0.116	0.311
7	DO100	F	16.9	0.107	NA
8	DO101	F	23.6	0.114	0.329
9	DO−102	M		0.131	0.277
10	DO−103	F	27.2	0.131	0.374
11	DO−104	F	20.5	−	0.297
12	DO−105	F	23.1	0.115	0.313
13	106	F	19.3	0.103	0.276
14	107	male	32.6	0.126	0.210

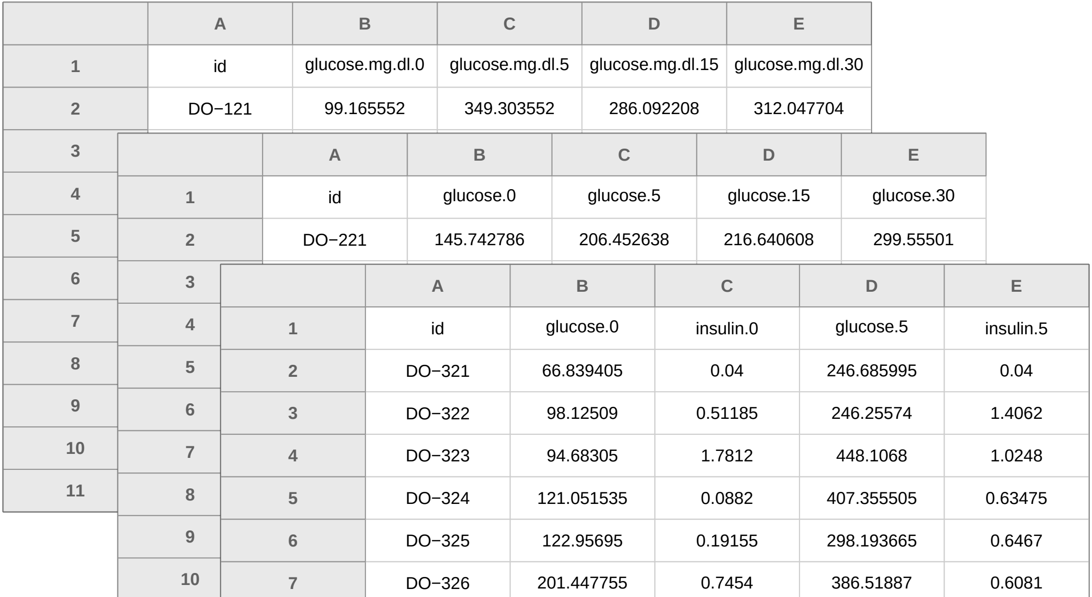
Consistent column names



	A	B	C	D	E
1	id	glucose.mg.dl.0	glucose.mg.dl.5	glucose.mg.dl.15	glucose.mg.dl.30
2	DO−121	99.165552	349.303552	286.092208	312.047704
3					
4					
5					
6					
7					
8					
9					
10					
11					

	A	B	C	D	E
1	id	glucose.0	glucose.5	glucose.15	glucose.30
2	DO−221	145.742786	206.452638	216.640608	299.55501
3	DO−222	138.010378	342.866944	339.836676	276.148802
4	DO−223	138.219362	407.443	336.858654	235.501414
5	DO−224	100.445504	310.944638	384.97722	308.907044
6	DO−225	121.030428	290.41196	345.740474	313.818168
7	DO−226	118.418128	189.524934	159.692468	144.488882
8	DO−227	117.4777	395.321928	448.612848	310.369932
9	DO−228	98.773632	149.452252	245.637138	317.423142
10	DO−229	122.44107	260.63174	231.008258	202.272958

Consistent layout



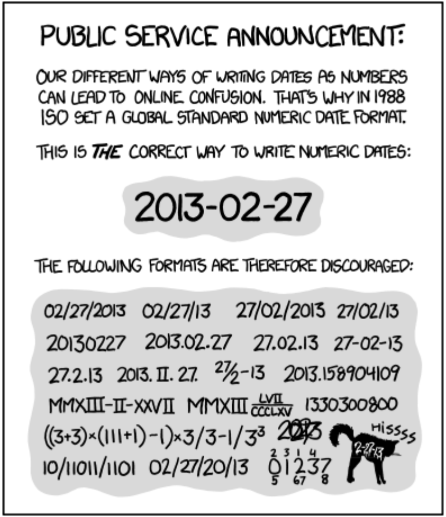
	A	B	C	D	E
1	id	glucose.mg.dl.0	glucose.mg.dl.5	glucose.mg.dl.15	glucose.mg.dl.30
2	DO−121	99.165552	349.303552	286.092208	312.047704

	A	B	C	D	E
1	id	glucose.0	glucose.5	glucose.15	glucose.30
2	DO−221	145.742786	206.452638	216.640608	299.55501

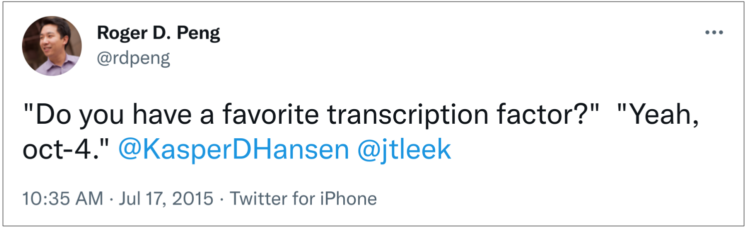
	A	B	C	D	E
1	id	glucose.0	insulin.0	glucose.5	insulin.5
2	DO−321	66.839405	0.04	246.685995	0.04
3	DO−322	98.12509	0.51185	246.25574	1.4062
4	DO−323	94.68305	1.7812	448.1068	1.0248
5	DO−324	121.051535	0.0882	407.355505	0.63475
6	DO−325	122.95695	0.19155	298.193665	0.6467
7	DO−326	201.447755	0.7454	386.51887	0.6081

Consistent date format

	A	B	C
1	Date	Assay date	Weight
2		12/9/05	54.9
3		12/9/05	45.3
4	12/6/2005	e	47
5		e	45.7
6		e	52.9
7		1/11/2006	46.1
8		1/11/2006	38.6


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27⁄2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
HISSsss
2-27-13
10/11011/1101 02/27/20/13
2 3 1 4
01237
5 67 8

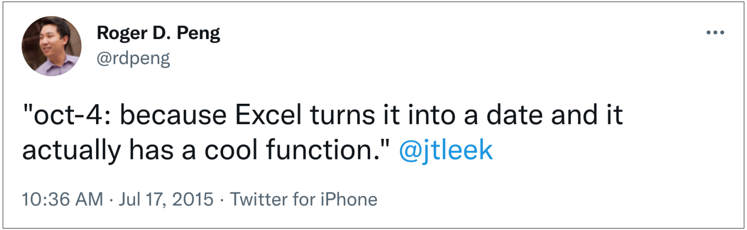
xkcd.com/1179



Roger D. Peng
@rdpeng

"Do you have a favorite transcription factor?" "Yeah, oct-4." @KasperDHansen @jtleek

10:35 AM · Jul 17, 2015 · Twitter for iPhone



Roger D. Peng
@rdpeng

"oct-4: because Excel turns it into a date and it actually has a cool function." @jtleek

10:36 AM · Jul 17, 2015 · Twitter for iPhone

Consistent file names

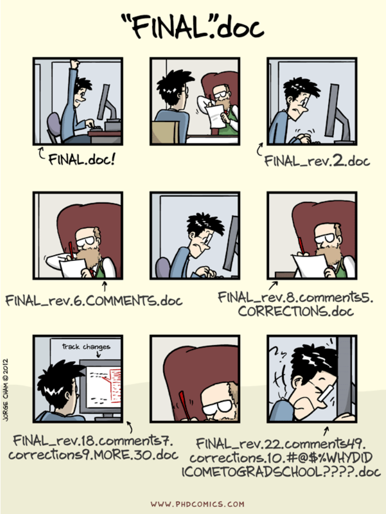
```
Complete F2 Liver TG Set.xls
CPL Rosetta Lipids FINAL.xls
D2O Summary of All F2 Samples MF 30July2009.xls
FINAL RBM Data 102989 26Sept2007.xls
Mapped Urine Plasma Data to Statgen.xls
Necropsy Tracking Report rk 2011-04-26.xls
Necropsy Tracking Report rk61412.xls
Necropsy_Tracking_Report_rk_052912_atb.xls
Original Necropsy Tracking Report rk.xls
RBM Tube Number Key.xls
```

Consistent file names

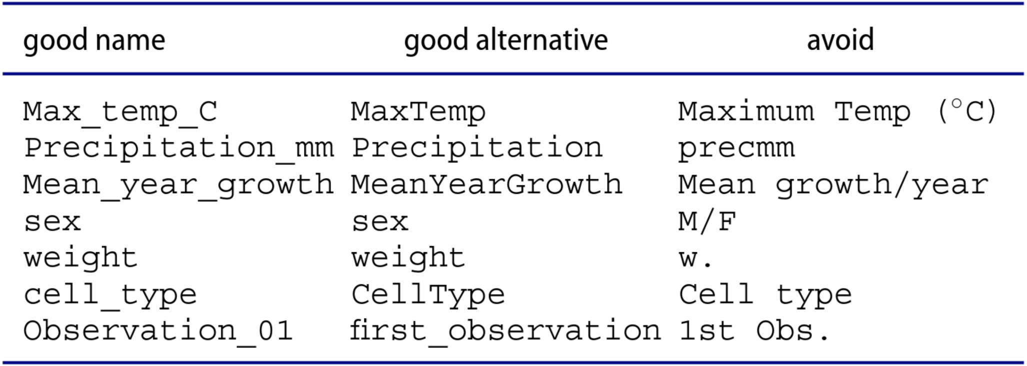
```
Complete F2 Liver TG Set.xls
CPL Rosetta Lipids FINAL.xls
D2O Summary of All F2 Samples MF 30July2009.xls
FINAL RBM Data 102989 26Sept2007.xls
Mapped Urine Plasma Data to Statgen.xls
Necropsy Tracking Report rk 2011-04-26.xls
Necropsy Tracking Report rk61412.xls
Necropsy_Tracking_Report_rk_052912_atb.xls
Original Necropsy Tracking Report rk.xls
RBM Tube Number Key.xls
```

- No spaces or special characters
- Short but descriptive
- Consistent scheme
- Take advantage of computer sorting

No “final” in file names



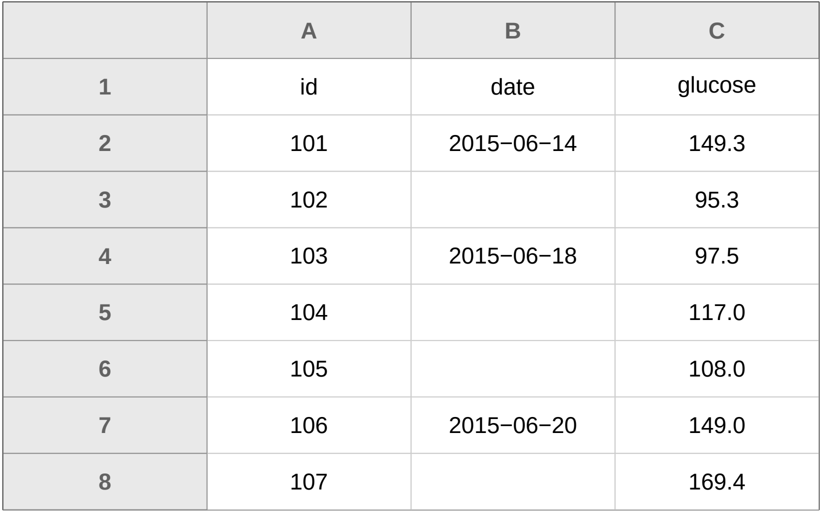
Choose good names for things



good name	good alternative	avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.

DataCarpentry.org

No empty cells

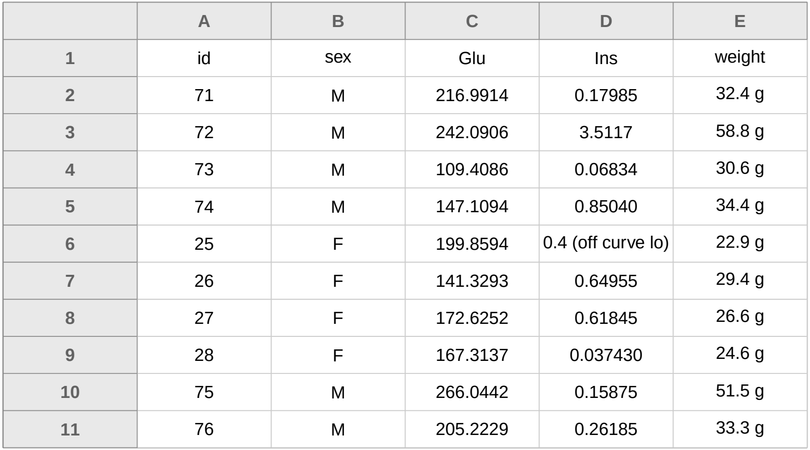


	A	B	C
1	id	date	glucose
2	101	2015−06−14	149.3
3	102		95.3
4	103	2015−06−18	97.5
5	104		117.0
6	105		108.0
7	106	2015−06−20	149.0
8	107		169.4

No empty cells

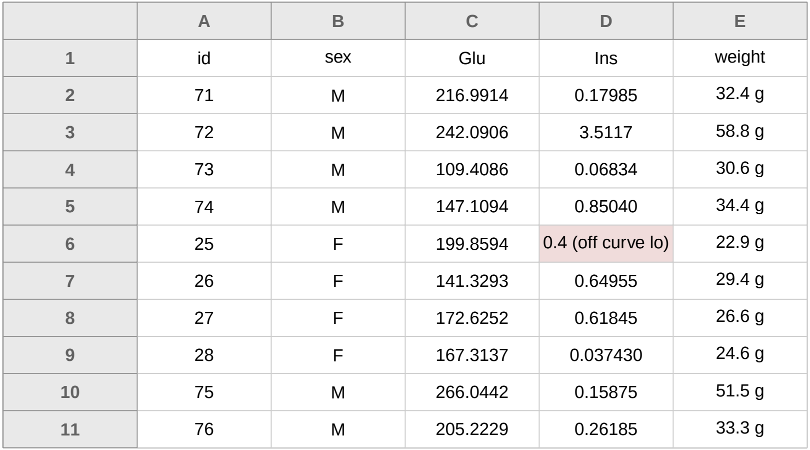
	A	B	C
1	id	date	glucose
2	101	2015−06−14	149.3
3	102	2015−06−14	95.3
4	103	2015−06−18	97.5
5	104	2015−06−18	117.0
6	105	2015−06−18	108.0
7	106	2015−06−20	149.0
8	107	2015−06−20	169.4

One thing per cell



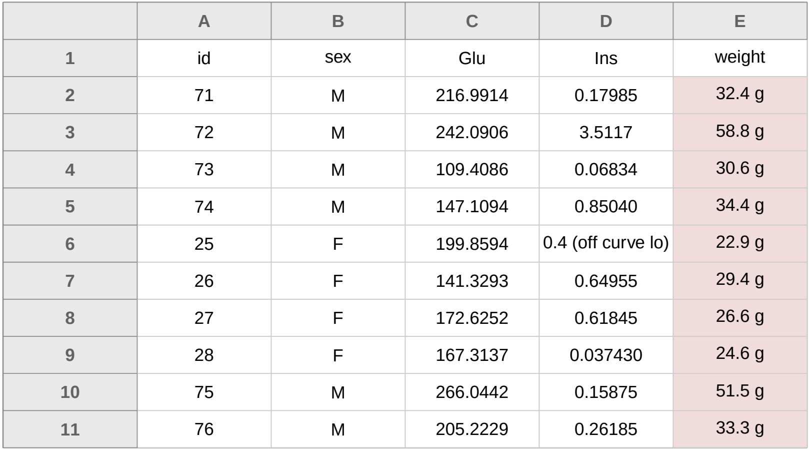
	A	B	C	D	E
1	id	sex	Glu	Ins	weight
2	71	M	216.9914	0.17985	32.4 g
3	72	M	242.0906	3.5117	58.8 g
4	73	M	109.4086	0.06834	30.6 g
5	74	M	147.1094	0.85040	34.4 g
6	25	F	199.8594	0.4 (off curve lo)	22.9 g
7	26	F	141.3293	0.64955	29.4 g
8	27	F	172.6252	0.61845	26.6 g
9	28	F	167.3137	0.037430	24.6 g
10	75	M	266.0442	0.15875	51.5 g
11	76	M	205.2229	0.26185	33.3 g

One thing per cell



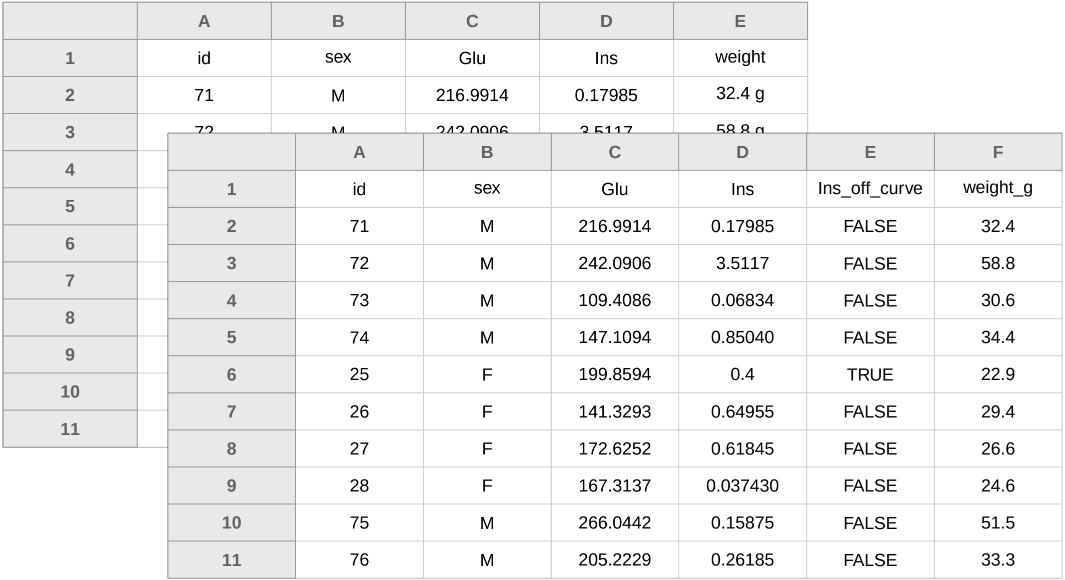
	A	B	C	D	E
1	id	sex	Glu	Ins	weight
2	71	M	216.9914	0.17985	32.4 g
3	72	M	242.0906	3.5117	58.8 g
4	73	M	109.4086	0.06834	30.6 g
5	74	M	147.1094	0.85040	34.4 g
6	25	F	199.8594	0.4 (off curve lo)	22.9 g
7	26	F	141.3293	0.64955	29.4 g
8	27	F	172.6252	0.61845	26.6 g
9	28	F	167.3137	0.037430	24.6 g
10	75	M	266.0442	0.15875	51.5 g
11	76	M	205.2229	0.26185	33.3 g

One thing per cell



	A	B	C	D	E
1	id	sex	Glu	Ins	weight
2	71	M	216.9914	0.17985	32.4 g
3	72	M	242.0906	3.5117	58.8 g
4	73	M	109.4086	0.06834	30.6 g
5	74	M	147.1094	0.85040	34.4 g
6	25	F	199.8594	0.4 (off curve lo)	22.9 g
7	26	F	141.3293	0.64955	29.4 g
8	27	F	172.6252	0.61845	26.6 g
9	28	F	167.3137	0.037430	24.6 g
10	75	M	266.0442	0.15875	51.5 g
11	76	M	205.2229	0.26185	33.3 g

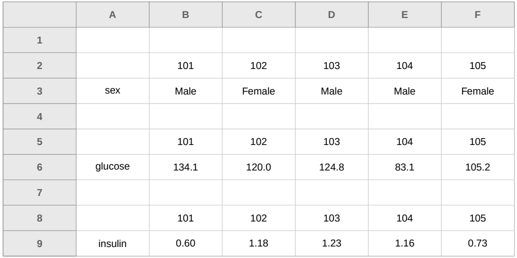
One thing per cell



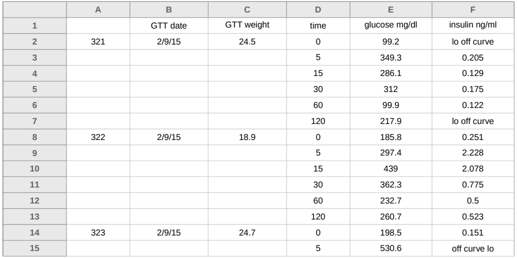
	A	B	C	D	E
1	id	sex	Glu	Ins	weight
2	71	M	216.9914	0.17985	32.4 g
3	72	M	242.0906	3.5117	58.8 g
4					
5					
6					
7					
8					
9					
10					
11					

	A	B	C	D	E	F
1	id	sex	Glu	Ins	Ins_off_curve	weight_g
2	71	M	216.9914	0.17985	FALSE	32.4
3	72	M	242.0906	3.5117	FALSE	58.8
4	73	M	109.4086	0.06834	FALSE	30.6
5	74	M	147.1094	0.85040	FALSE	34.4
6	25	F	199.8594	0.4	TRUE	22.9
7	26	F	141.3293	0.64955	FALSE	29.4
8	27	F	172.6252	0.61845	FALSE	26.6
9	28	F	167.3137	0.037430	FALSE	24.6
10	75	M	266.0442	0.15875	FALSE	51.5
11	76	M	205.2229	0.26185	FALSE	33.3

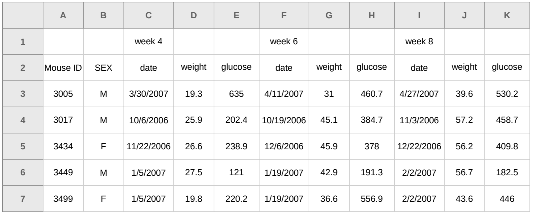
Make it a rectangle



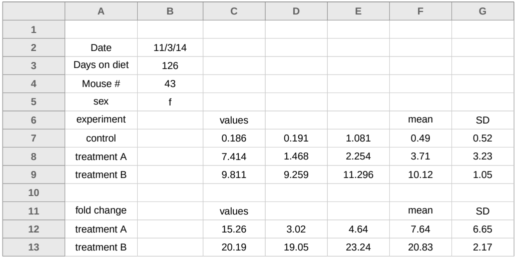
	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Female
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73



	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	286.1	0.129
5				30	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/15	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/15	24.7	0	198.5	0.151
15				5	530.6	off curve lo



	A	B	C	D	E	F	G	H	I	J	K
1			week 4			week 6			week 8		
2	Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3	3005	M	3/30/2007	19.3	635	4/11/2007	31	460.7	4/27/2007	39.6	530.2
4	3017	M	10/6/2006	25.9	202.4	10/19/2006	45.1	384.7	11/3/2006	57.2	458.7
5	3434	F	11/22/2006	26.6	238.9	12/6/2006	45.9	378	12/22/2006	56.2	409.8
6	3449	M	1/5/2007	27.5	121	1/19/2007	42.9	191.3	2/2/2007	56.7	182.5
7	3499	F	1/5/2007	19.8	220.2	1/19/2007	36.6	556.9	2/2/2007	43.6	446

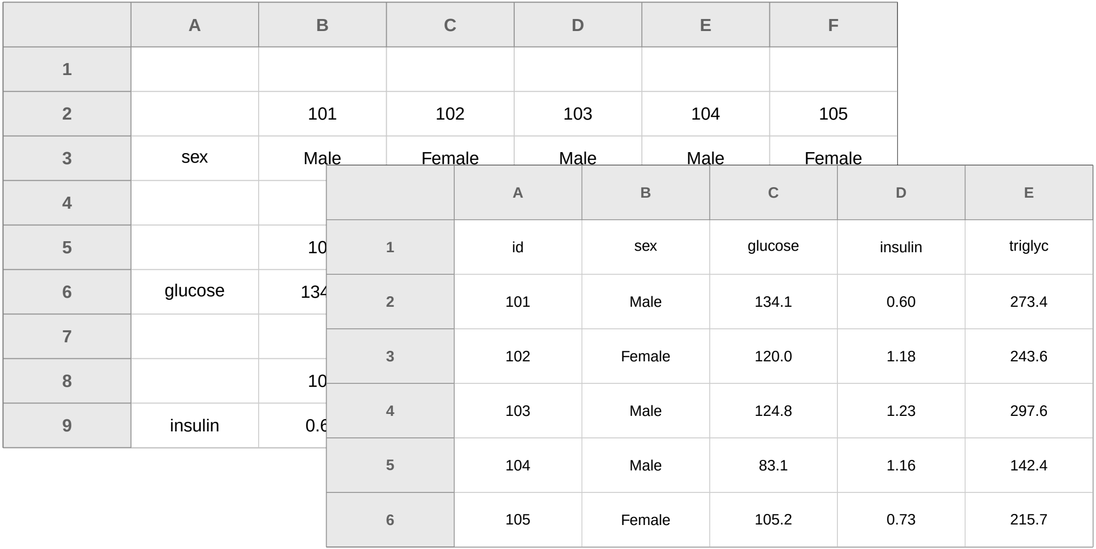


	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

Make it a rectangle

	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Female
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73

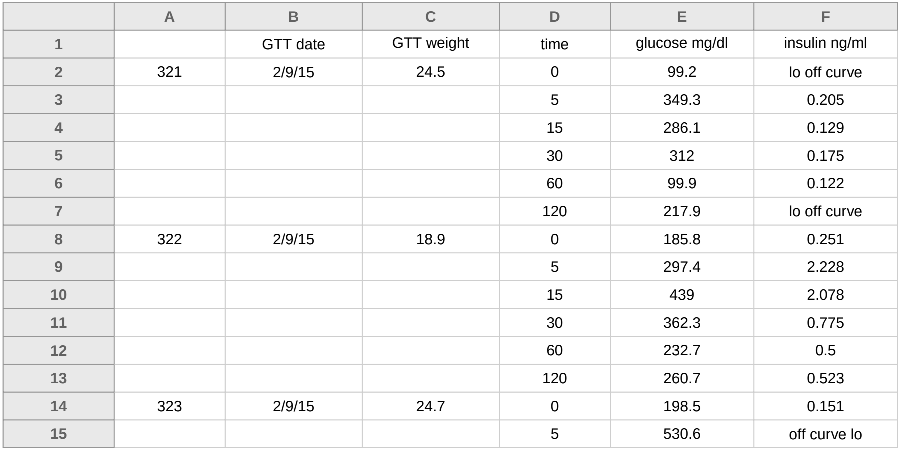
Make it a rectangle



	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Female
4						
5		10				
6	glucose	134				
7						
8		10				
9	insulin	0.6				

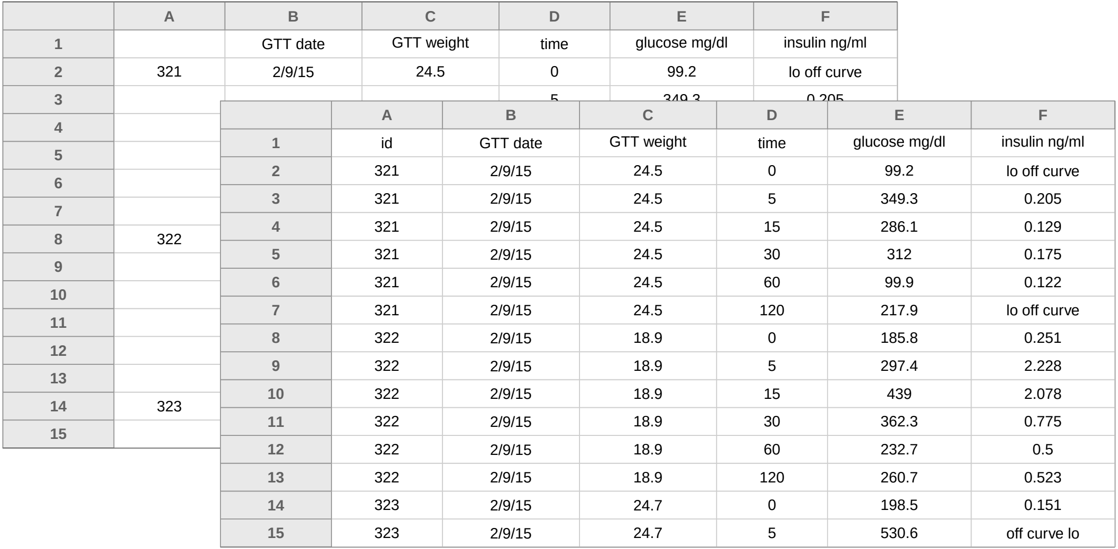
	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Female	105.2	0.73	215.7

Make it a rectangle



	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	286.1	0.129
5				30	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/15	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/15	24.7	0	198.5	0.151
15				5	530.6	off curve lo

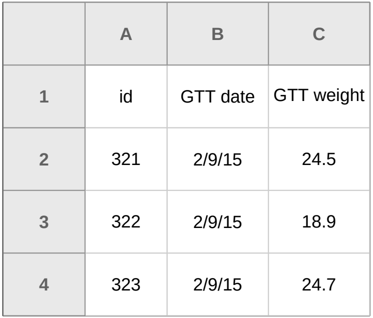
Make it a rectangle



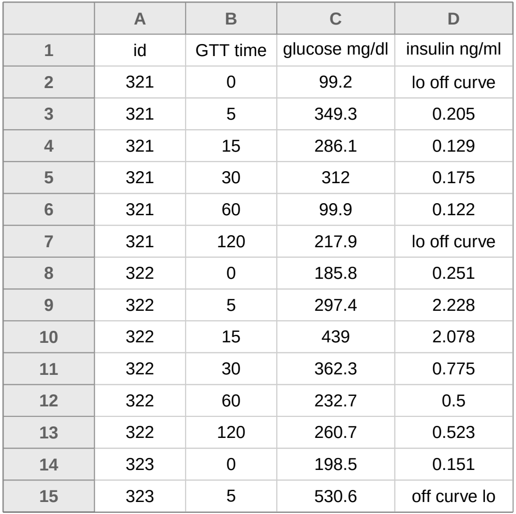
	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4						
5						
6						
7						
8	322					
9						
10						
11						
12						
13						
14	323					
15						

	A	B	C	D	E	F
1	id	GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3	321	2/9/15	24.5	5	349.3	0.205
4	321	2/9/15	24.5	15	286.1	0.129
5	321	2/9/15	24.5	30	312	0.175
6	321	2/9/15	24.5	60	99.9	0.122
7	321	2/9/15	24.5	120	217.9	lo off curve
8	322	2/9/15	18.9	0	185.8	0.251
9	322	2/9/15	18.9	5	297.4	2.228
10	322	2/9/15	18.9	15	439	2.078
11	322	2/9/15	18.9	30	362.3	0.775
12	322	2/9/15	18.9	60	232.7	0.5
13	322	2/9/15	18.9	120	260.7	0.523
14	323	2/9/15	24.7	0	198.5	0.151
15	323	2/9/15	24.7	5	530.6	off curve lo

Make it a rectangle

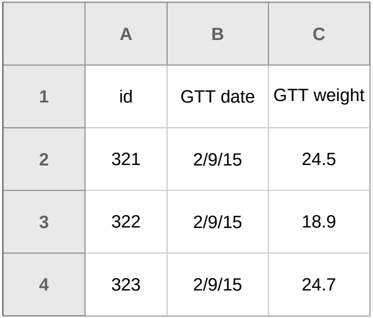


	A	B	C
1	id	GTT date	GTT weight
2	321	2/9/15	24.5
3	322	2/9/15	18.9
4	323	2/9/15	24.7

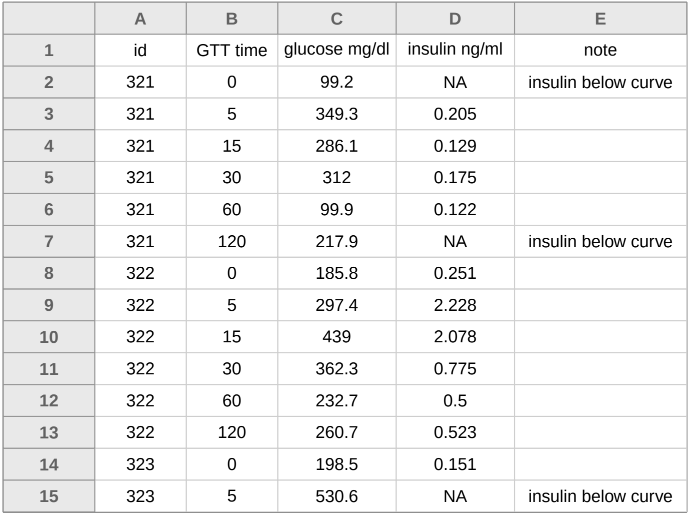


	A	B	C	D
1	id	GTT time	glucose mg/dl	insulin ng/ml
2	321	0	99.2	lo off curve
3	321	5	349.3	0.205
4	321	15	286.1	0.129
5	321	30	312	0.175
6	321	60	99.9	0.122
7	321	120	217.9	lo off curve
8	322	0	185.8	0.251
9	322	5	297.4	2.228
10	322	15	439	2.078
11	322	30	362.3	0.775
12	322	60	232.7	0.5
13	322	120	260.7	0.523
14	323	0	198.5	0.151
15	323	5	530.6	off curve lo

Make it a rectangle

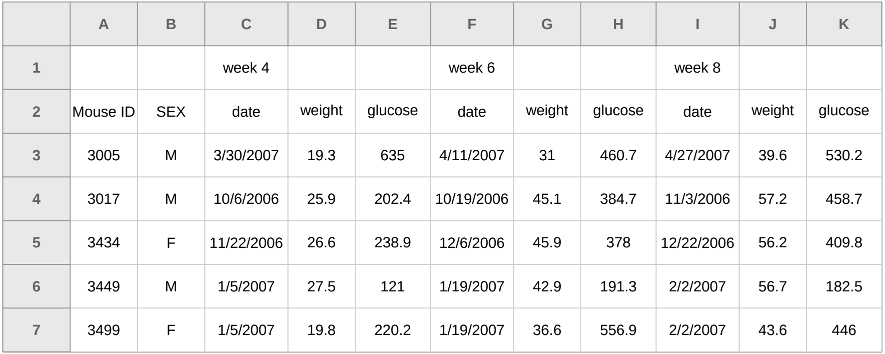


	A	B	C
1	id	GTT date	GTT weight
2	321	2/9/15	24.5
3	322	2/9/15	18.9
4	323	2/9/15	24.7



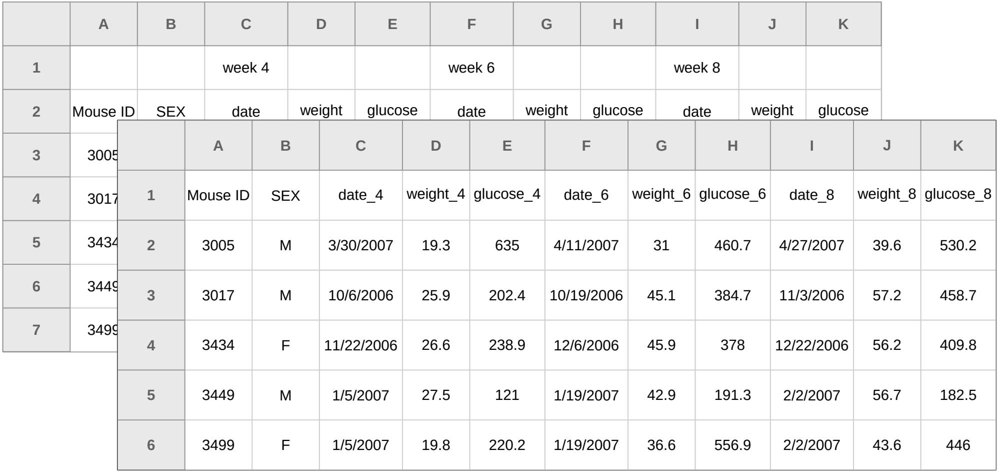
	A	B	C	D	E
1	id	GTT time	glucose mg/dl	insulin ng/ml	note
2	321	0	99.2	NA	insulin below curve
3	321	5	349.3	0.205	
4	321	15	286.1	0.129	
5	321	30	312	0.175	
6	321	60	99.9	0.122	
7	321	120	217.9	NA	insulin below curve
8	322	0	185.8	0.251	
9	322	5	297.4	2.228	
10	322	15	439	2.078	
11	322	30	362.3	0.775	
12	322	60	232.7	0.5	
13	322	120	260.7	0.523	
14	323	0	198.5	0.151	
15	323	5	530.6	NA	insulin below curve

Make it a rectangle



	A	B	C	D	E	F	G	H	I	J	K
1			week 4			week 6			week 8		
2	Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3	3005	M	3/30/2007	19.3	635	4/11/2007	31	460.7	4/27/2007	39.6	530.2
4	3017	M	10/6/2006	25.9	202.4	10/19/2006	45.1	384.7	11/3/2006	57.2	458.7
5	3434	F	11/22/2006	26.6	238.9	12/6/2006	45.9	378	12/22/2006	56.2	409.8
6	3449	M	1/5/2007	27.5	121	1/19/2007	42.9	191.3	2/2/2007	56.7	182.5
7	3499	F	1/5/2007	19.8	220.2	1/19/2007	36.6	556.9	2/2/2007	43.6	446

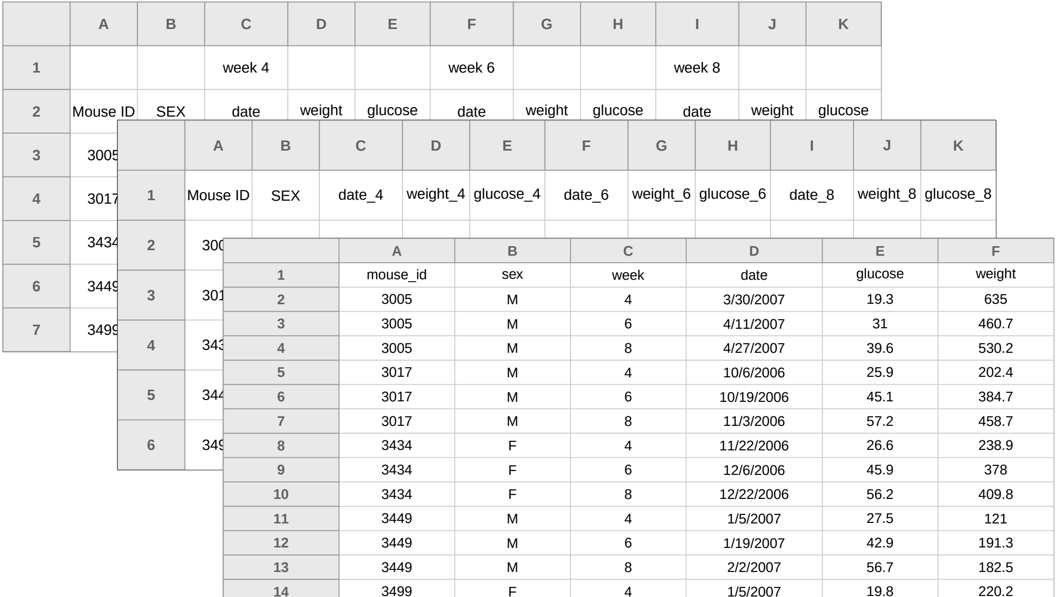
Make it a rectangle



	A	B	C	D	E	F	G	H	I	J	K
1			week 4			week 6			week 8		
2	Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3	3005										
4	3017										
5	3434										
6	3449										
7	3499										

	A	B	C	D	E	F	G	H	I	J	K
1	Mouse ID	SEX	date_4	weight_4	glucose_4	date_6	weight_6	glucose_6	date_8	weight_8	glucose_8
2	3005	M	3/30/2007	19.3	635	4/11/2007	31	460.7	4/27/2007	39.6	530.2
3	3017	M	10/6/2006	25.9	202.4	10/19/2006	45.1	384.7	11/3/2006	57.2	458.7
4	3434	F	11/22/2006	26.6	238.9	12/6/2006	45.9	378	12/22/2006	56.2	409.8
5	3449	M	1/5/2007	27.5	121	1/19/2007	42.9	191.3	2/2/2007	56.7	182.5
6	3499	F	1/5/2007	19.8	220.2	1/19/2007	36.6	556.9	2/2/2007	43.6	446

Make it a rectangle

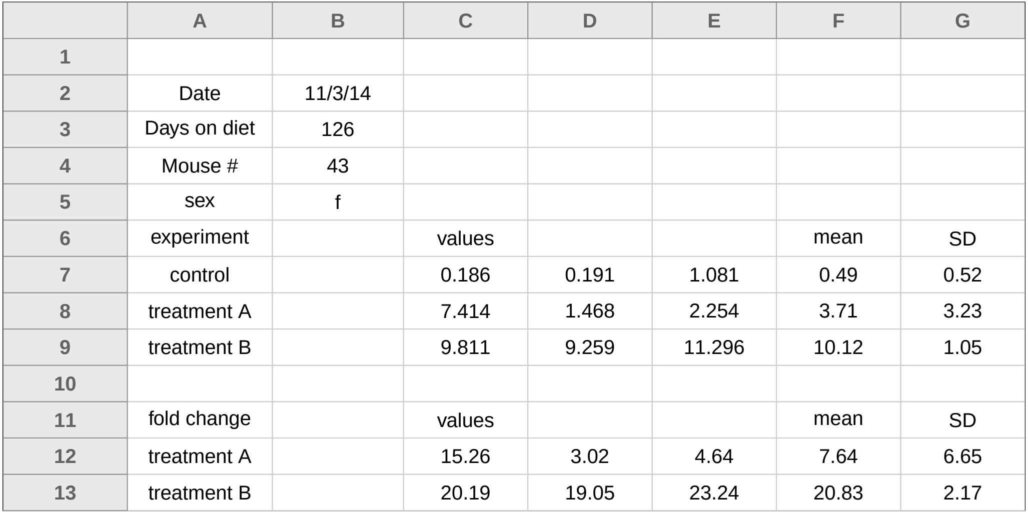


	A	B	C	D	E	F	G	H	I	J	K
1			week 4			week 6			week 8		
2	Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3	3005										
4	3017										
5	3434										
6	3449										
7	3499										

	A	B	C	D	E	F	G	H	I	J	K
1	Mouse ID	SEX	date_4	weight_4	glucose_4	date_6	weight_6	glucose_6	date_8	weight_8	glucose_8
2	300										
3	301										
4	343										
5	344										
6	349										

	A	B	C	D	E	F
1	mouse_id	sex	week	date	glucose	weight
2	3005	M	4	3/30/2007	19.3	635
3	3005	M	6	4/11/2007	31	460.7
4	3005	M	8	4/27/2007	39.6	530.2
5	3017	M	4	10/6/2006	25.9	202.4
6	3017	M	6	10/19/2006	45.1	384.7
7	3017	M	8	11/3/2006	57.2	458.7
8	3434	F	4	11/22/2006	26.6	238.9
9	3434	F	6	12/6/2006	45.9	378
10	3434	F	8	12/22/2006	56.2	409.8
11	3449	M	4	1/5/2007	27.5	121
12	3449	M	6	1/19/2007	42.9	191.3
13	3449	M	8	2/2/2007	56.7	182.5
14	3499	F	4	1/5/2007	19.8	220.2

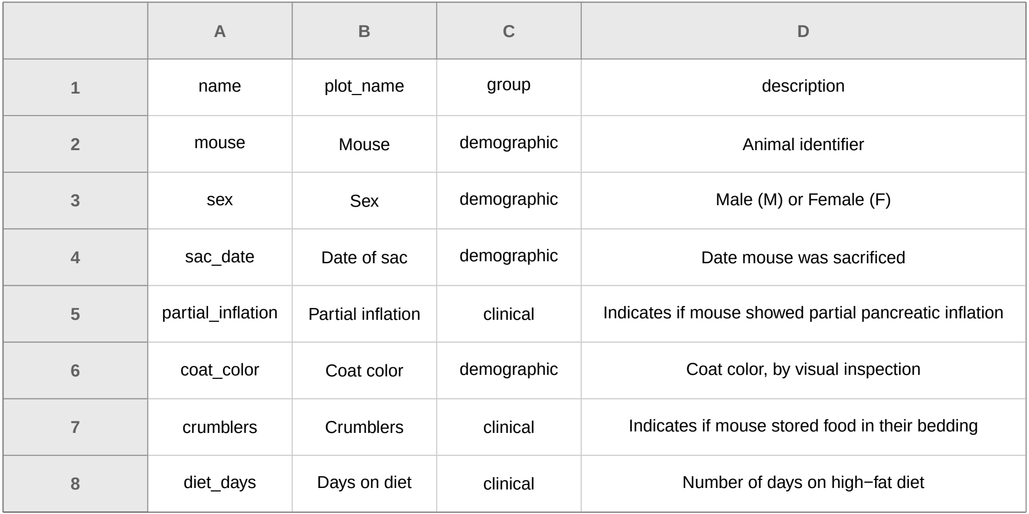
Make it a rectangle



	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

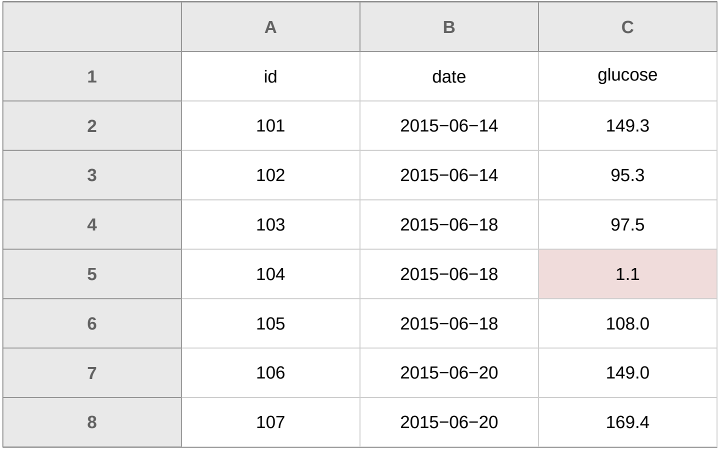
No calculations in the data file

Make a data dictionary



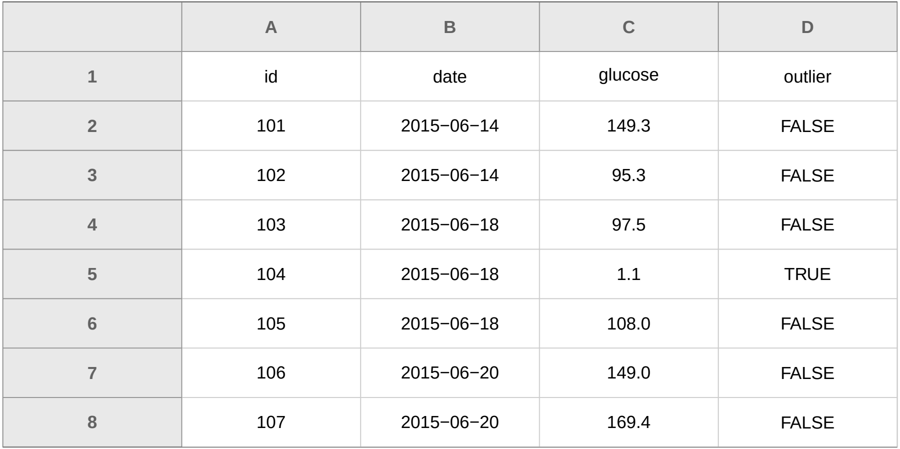
	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding
8	diet_days	Days on diet	clinical	Number of days on high−fat diet

No color/formatting as data



	A	B	C
1	id	date	glucose
2	101	2015−06−14	149.3
3	102	2015−06−14	95.3
4	103	2015−06−18	97.5
5	104	2015−06−18	1.1
6	105	2015−06−18	108.0
7	106	2015−06−20	149.0
8	107	2015−06−20	169.4

No color/formatting as data

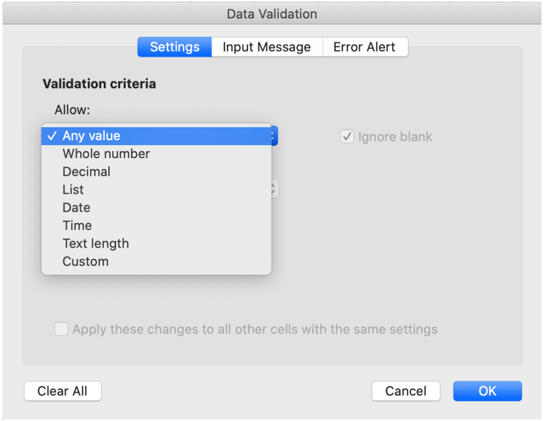


	A	B	C	D
1	id	date	glucose	outlier
2	101	2015−06−14	149.3	FALSE
3	102	2015−06−14	95.3	FALSE
4	103	2015−06−18	97.5	FALSE
5	104	2015−06−18	1.1	TRUE
6	105	2015−06−18	108.0	FALSE
7	106	2015−06−20	149.0	FALSE
8	107	2015−06−20	169.4	FALSE

Make backups

- Automatic
- Multiple locations (including off site)
- Consider formal version control

Use data validation



Data Validation

Validation criteria

Allow: ✓ Any value

☑ Ignore blank

□ Apply these changes to all other cells with the same settings

Save as plain text

- Don’t rely on a proprietary format
- Save as a comma-delimited (CSV) or tab-delimited (TSV) file

  Or vertical-bar-delimited?

Summary

1. Be consistent
2. Write dates as `YYYY-MM-DD`
3. Choose good names for things
4. No empty cells
5. One thing per cell
6. Make it a rectangle
7. Make a data dictionary
8. No calculations in the data file
9. No color/formatting as data
10. Make backups
11. Use data validation
12. Save as plain text

Acknowledgements

Kara Woo

Jenny Bryan

Hadley Wickham

All of my past scientific collaborators

Further reading

- Broman KW, Woo KH (2018) Data organization in spreadsheets. Am Stat 72:2-10 `doi.org/gdz6cm`
- White EP et al. (2013) Nine simple ways to make it easier to (re)use your data. Ideas Ecol Evol 6:1-10 `doi.org/10.4033/iee.2013.6b.6.f`
- Briney K (2015) Data management for researchers. Pelagic Publishing. ISBN: 9781784270117
- Ziemann et al (2016) Gene name errors are widespread in the scientific literature. Genome Biol 17:177 `doi.org/10.1186/s13059-016-1044-7`
- Ellis SE, Leek JT (2018) How to share data for collaboration. Am Stat 72:53-57 `doi.org/10.1080/00031305.2017.1375987`
- Wilson SL et al. (2021) Sharing biological data: why, when, and how. FEBS Letters 595:847-863 `doi.org/10.1002/1873-3468.14067`

Slides: kbroman.org/Talk_DataOrg

kbroman.org

github.com/kbroman

@kwbroman