Organizing Data in Spreadsheets

Author: Karl Browman

Presentation By: Morgan, Lauren, Gary and Dakota

Who is Karl Browman?

Karl Browman

- Professor in the Department of Biostatistics & Medical Informatics at the University of Wisconsin–Madison
- Researcher in statistical genetics
- Developer of R/qtl, an interactive environment for mapping quantitative trait loci for R.
- BS in mathematics in 1991, from the University of Wisconsin–Milwaukee,
- PhD in statistics in 1997, from the University of California, Berkeley; his PhD advisor was Terry Speed.

Interest in Organizing Data in Spreadsheets

- Proponent of data analysts being able to handle any data files they receive from others.

- In spreadsheets, data can be a sloppy mess requiring serious reorganization efforts (to be avoided)
    - Data analysts have to spend time reorganizing data from spreadsheets, instead of spending more time on analyses

- Writing scripts to rearrange the layout of data to prepare it for analysis is tedious.

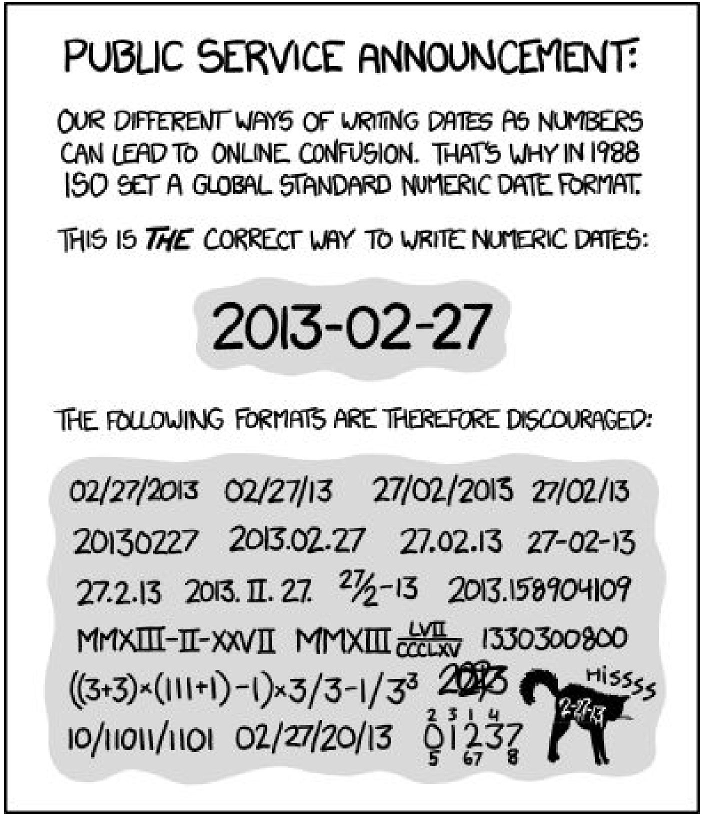
How to Organize Data in Spreadsheets

Be consistent

- Keep one naming convention for each type of object
- Use the same conventions across all files in the project
- Ex:
    - camelCase for all variables in code
    - YYYY-MM-DD for all dates
    - snake_case for all file names
- Store data in the same layout when possible
- Avoid white spaces at all costs, use underscores/hyphens/periods/ect.

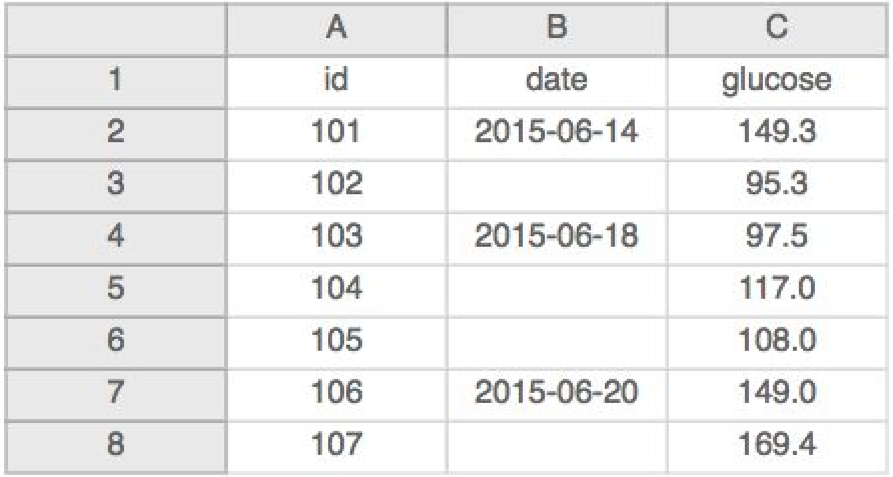
Write dates as YYYY-MM-DD

- When writing dates, the most common convention is to use: YYYY-MM-DD
- This format is easily legible and used by most operating systems

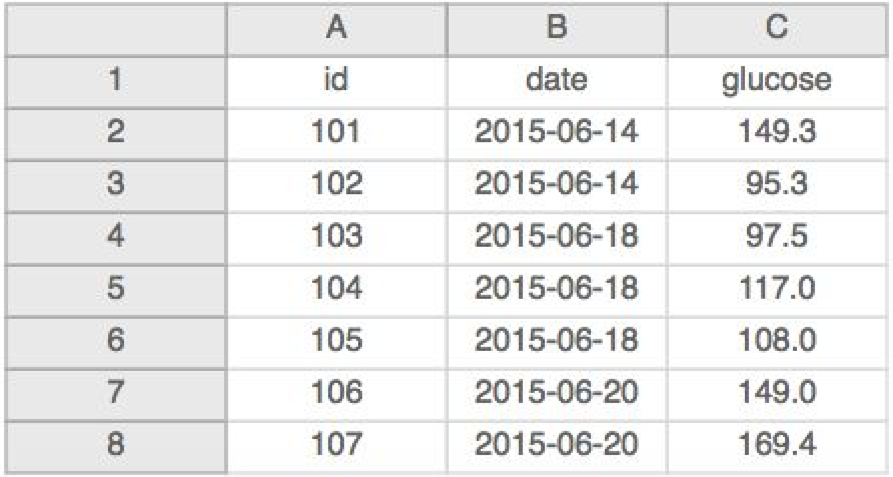

PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 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
10/11011/1101 02/27/20/13 01237
2-27-13
Hissss

No empty cells

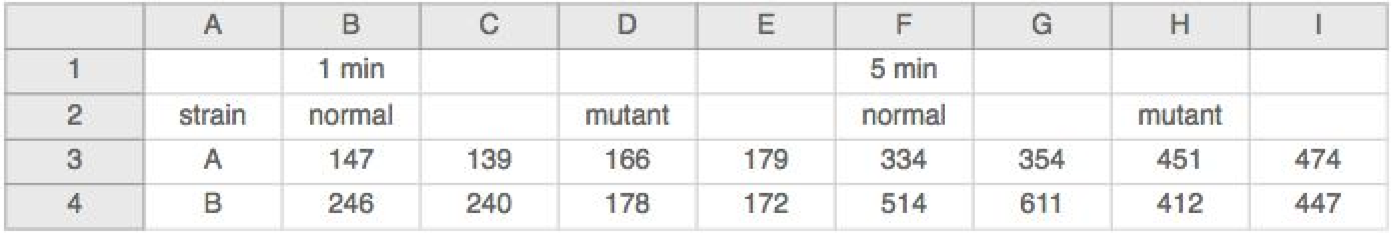
- When writing data tables we always prefer to leave no empty cells
- Empty cells can lead to complications when running code scripts on the data
  - Some programs read empty cells as special values: NA, NaN, None, 0, ect.
- If possible, edit/clean the data before running any analysis



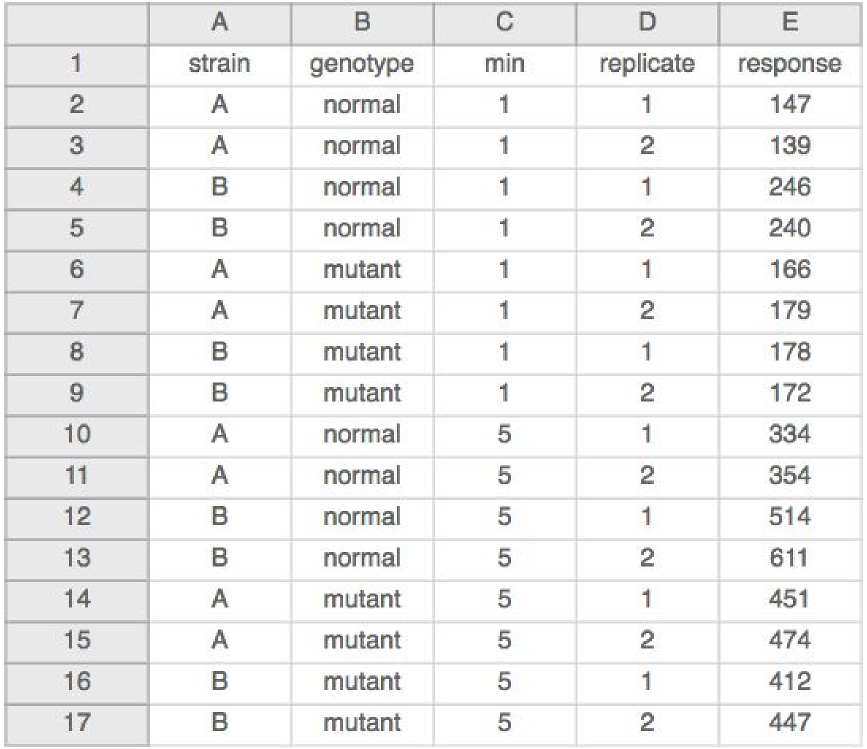
	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102		95.3
4	103	2015-06-18	97.5
5	104		117.0
6	105		108.0
7	106	2015-06-20	149.0
8	107		169.4



	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



	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447



	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

Put just one thing in each cell

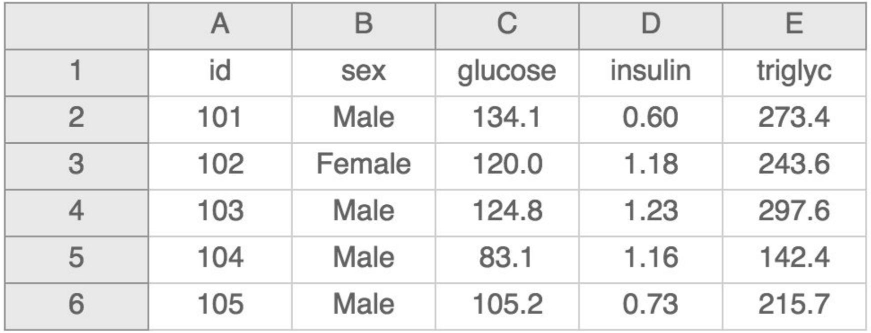
- It’s best practice to only include one piece of information in each cell
- Some data may be entered with two factors, if this occurs we always try to separate the column into two columns
  - Ex: Sex and age could be recorded as in one cell as “M75”
  - We would prefer to create two separate columns and delete the original, now the row in question will have one cell for sex (M) and another for age (75)
- Another common issue is recording units of measure in a cell
  - Ex: Weights could be recorded as: “150lb”, or “68kg”
  - We would prefer to create two separate columns and delete the original, now the row in question will have one cell for weight (150/68) and another for unit of measure (lb/kg)
- A final piece of advice it to never merge cells, the aesthetic is not worth the potential coding issues that arise with blank space creation

Discussion Question 1:

Why do we care about variable naming conventions within code files that aren’t shown in the final report?

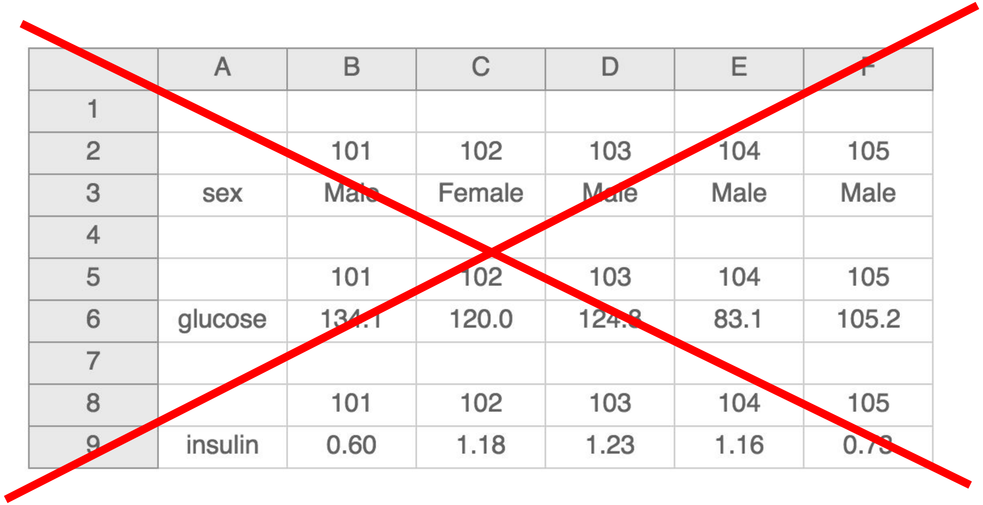
Make it a rectangle

- The best layout for your data within in a spreadsheet is as a big rectangle with rows corresponding to subjects and columns corresponding to variables.

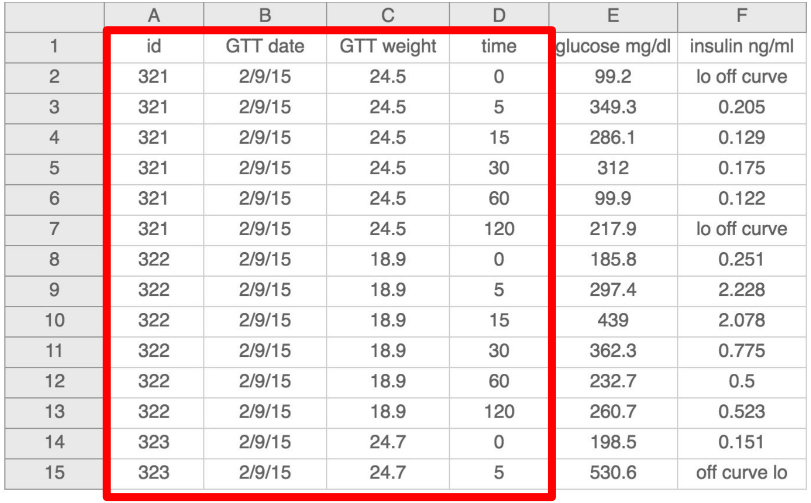


	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

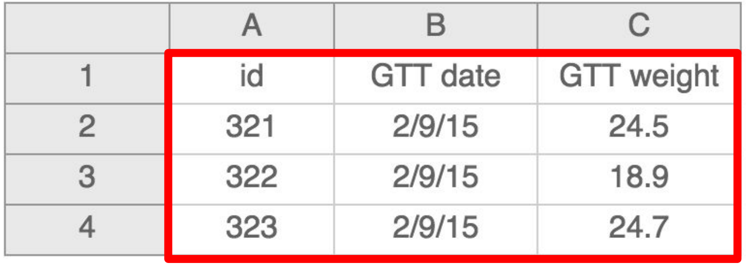
Good example 👍



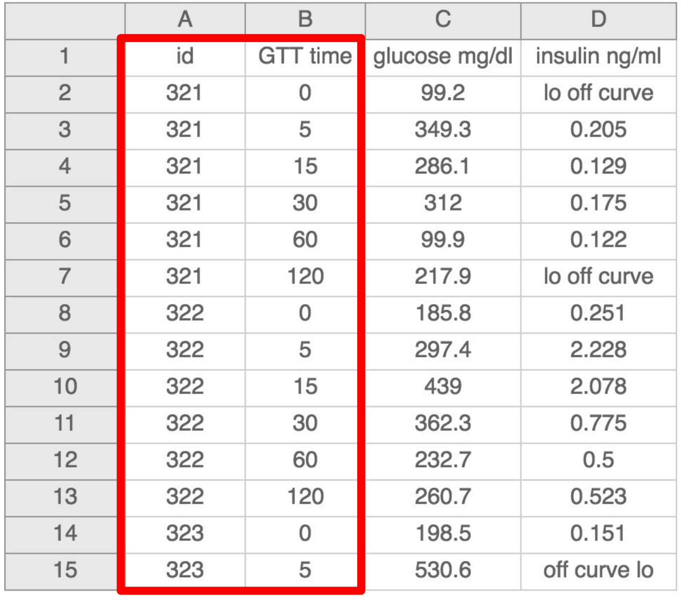
	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Male
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73



	A	B	C	D	E	F
1	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



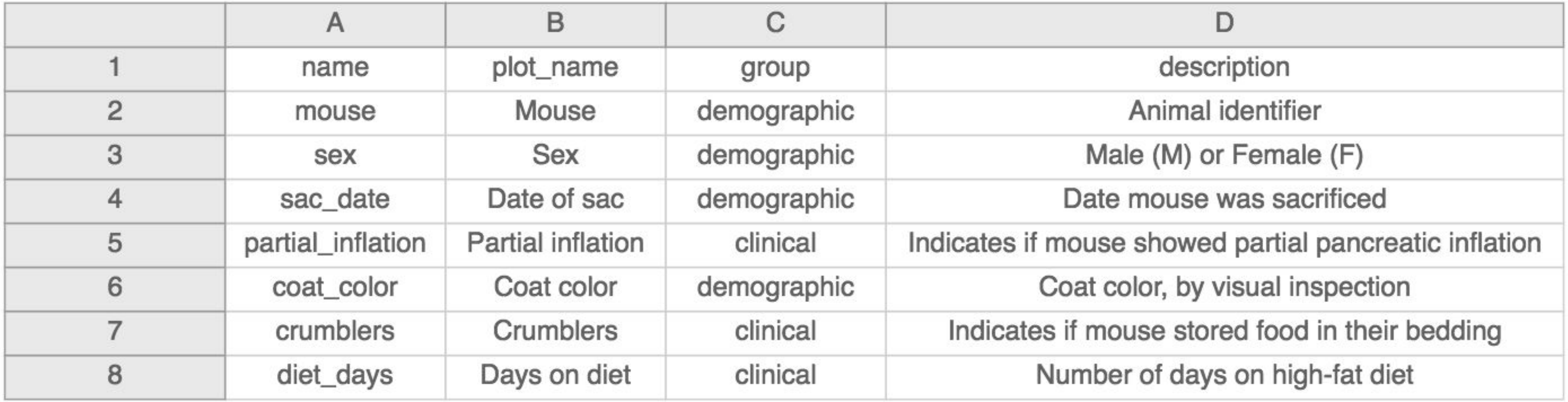
	A	B	C
1	id	GTT date	GTT weight
2	321	2/9/15	24.5
3	322	2/9/15	18.9
4	323	2/9/15	24.7



	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

Create a data dictionary

- A data dictionary is essentially part of the *metadata* (information *about* the data)



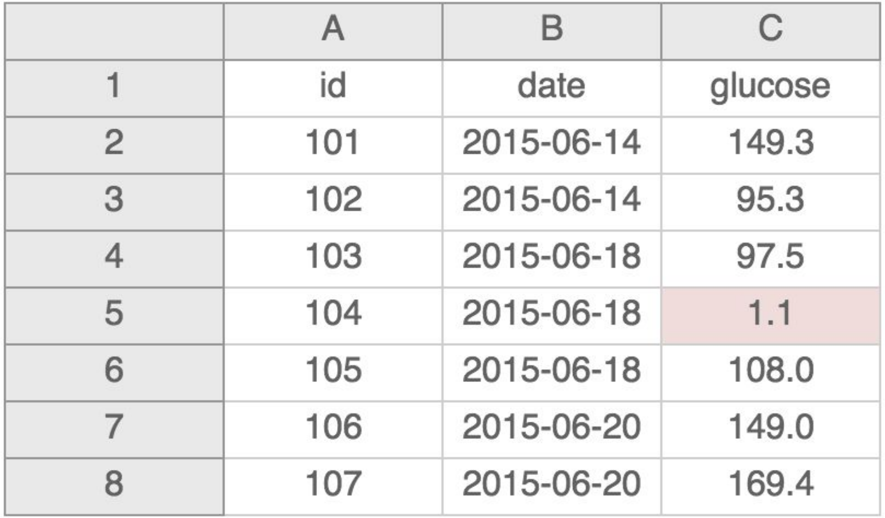
	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	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 calculations in the raw data files

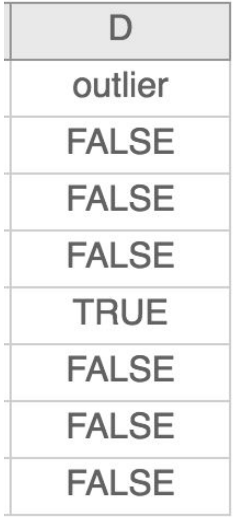
- Your primary data file should contain *just the data* and nothing else: no calculations, no graphs.
- There's a way higher risk of deleting things and messing things up if you're doing calculations
- Write it, protect it, back it up.
- If you want to do some analyses in Excel, make a copy of the file and do your calculations and graphs in the copy.

Don’t use font, color or highlighting as data

- You might be tempted to highlight particular cells with suspicious data, or rows that should be ignored.



	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



D
outlier
FALSE
FALSE
FALSE
TRUE
FALSE
FALSE
FALSE

Discussion Question 2:

What’s another reason for not manipulating your raw data file other than typos or accidentally deleting data

Choose good names for things

- Don’t use spaces for variable or file names (“glucose 6 weeks”)
- Be careful not to include extraneous spaces (“glucose “)
- Avoid special characters (“$per-gallon”)
- Make names short but meaningful (“weight” vs “w.”)
- Don’t include “final” in a filename...you will inevitably have a “final_rev2”, “final_rev3”, etc.

Make backups

- Don’t let a burning building destroy your life’s work. Making backups using systems such as git or dat will prevent this.
- Keep all versions of data files in case you make an error and want to return to a prior version
- Write-protect a data file once you’ve finished compiling and cleaning the data so no changes can be made (this is done by making the file “read only”)

Use data validation to avoid data entry mistakes

1) Use data validation Excel feature if applicable

- Select a column
- In the menu bar, choose Data → Validation
- Choose appropriate validation criteria. For example:
    - A whole number in some range
    - A decimal number in some range
    - A list of possible values
    - Text, but with a limit on length

2) Select data type of column to prevent data from being construed

- Select the column
- In the menu bar, select Format → Cells
- Choose “Text” on the left

Save the data in plain text files

- Saving the data as a comma or tab delimited plain text file to increase the reproducibility of your work
- These files never require any kind of software
- Note: if your file contains special features that would be compromised by saving as a plain text file, DON’T save the data as a plain text file OR make the data simpler so that it can be saved in that format without losing information

Other things to avoid

- Be careful of automatic data changes (such as “100,000” changed to “1e6”)
- “Freeze Panes” is handy for seeing the column headers while scrolling through an Excel file
- File in blank cells with zeros. Zeros are data!

Discussion Question 3:

Since it's recommended that most analysis be done in R or Python rather than Excel, is it important to teach how to organize data in spreadsheet?

Final Thoughts

Final Thoughts

- It is recommended to follow recommended rules for organizing spreadsheet data
- By following the recommended rules, you will decrease errors in the dataset as well as increase the reproducibility of all analysis done using the dataset