

# Data Organisation in Spreadsheets

Karl W. Broman & Kara H. Woo (2017) Data Organization in Spreadsheets,  
The American Statistician, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989)

# Be consistent

- codes - “male”, “MALE”, “m”, “Male”
- missing values - “NA”, “\*”, “-“, not “-999”
- column names - “glucose\_10\_wk”, “gluc\_10\_week”
- identifiers - “mouse153”, “153”
- data layout in multiple sheets
- file names
- data format - e.g. dates (see later)
- use of spaces - “male” vs “ male”

# Column names

- Don't use spaces
  - Use underscores - glucose\_week\_01
  - Use CamelCase - glucoseWeek01
- Avoid special characters (\$ % & \* etc)
- As short as possible whilst retaining meaning...

Use ISO 8601 for dates...

# 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.  $27\frac{1}{2}$ -13 2013.158904109

MMXIII-II-XXVII MMXIII  $\frac{\text{LVII}}{\text{CCCLXV}}$  1330300800

$((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$  ~~2013~~ 

10/11011/1101 02/27/20/13  $\begin{array}{cccc} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ 5 & 6 & 7 & 8 & \end{array}$

No empty cells

# Just one thing per cell

	A	B	C
1	<b>id</b>	<b>sex</b>	<b>medication</b>
2	A01	MN	pen
3	A02	FN	pen, strep
4	A03	ME	strep
5	A04	FE	pen, strep, NSAID
6	A05	ME	strep, pen

Don't merge any cells



# Keep data rectangular

	A	B	C	D	E	F	G
1				<b>week 1</b>		<b>week 2</b>	
2	<b>id</b>	<b>sex</b>	<b>medication</b>	<b>glucose</b>	<b>TP</b>	<b>glucose</b>	<b>TP</b>
3	A01	MN	pen	3	23	5	23
4	A02	FN	pen, strep	4	12	4	24
5	A03	ME	strep	5	13	4	24
6	A04	FE	pen, strep, NSAID	2	14	4	23
7	A05	ME	strep, pen	4	15	1	12

# Keep data rectangular

	A	B	C	D	E	F	G
1	id	sex	medication	glucose_wk1	TP_wk1	glucose_wk2	TP_wk2
2	A01	MN	pen	3	23	5	23
3	A02	FN	pen, strep	4	12	4	24
4	A03	ME	strep	5	13	4	24
5	A04	FE	pen, strep, NSAID	2	14	4	23
6	A05	ME	strep, pen	4	15	1	12

Create a data-dictionary  
(that's machine readable)

	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 in-sheet sums or  
formulae

Don't store data in  
colours or fills (alone)