Rbasics

PhD toolbox - 39th PhD cycle

Part II - How to manage data in spreadsheets

Data Organization in Spreadsheets

Now you should know the basic R syntax and you’re ready to start to import real datasets in R!

but

Your data requires a clear structure

Spreadsheets (mostly Excel) are useful tools for data entry but not suitable for reproducible research

Example: statistical procedures in Excel are manual. If you need to change one parameter of your analysis you’ll have to redo all your job.

Data Organization in Spreadsheets

Do not treat your data spreadsheet as your lab book!

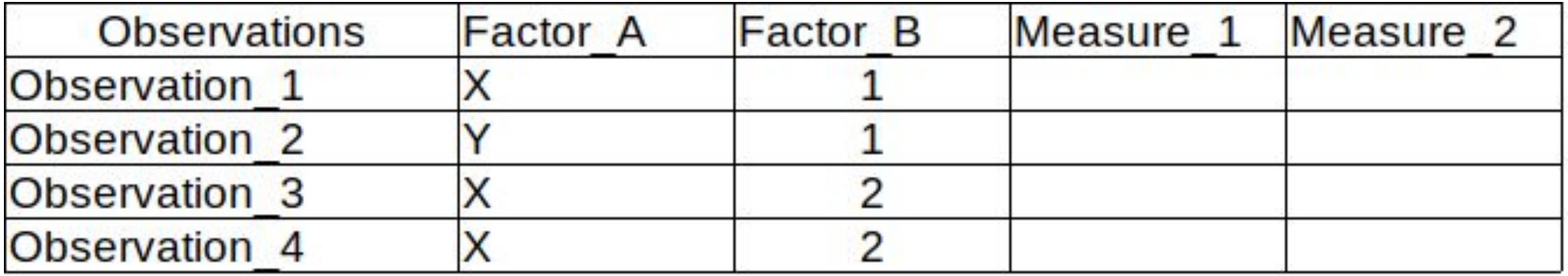
- Your data needs to be correctly read and interpreted by your Computer (not by your supervisor!)
- Additional notes and graphic layout of your data are useless most of the time
- keep your spreadsheet as tidy as possible

Some operative TIPS according to https://datacarpentry.org/spreadsheet-ecology-lesson/

Data Organization in Spreadsheets

Some cardinal rules to correctly compile your data spreadsheet

1) variables in column, observations in rows

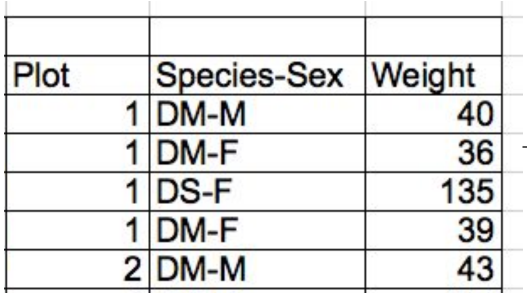


Observations	Factor_A	Factor_B	Measure_1	Measure_2
Observation_1	X	1		
Observation_2	Y	1		
Observation_3	X	2		
Observation_4	X	2		

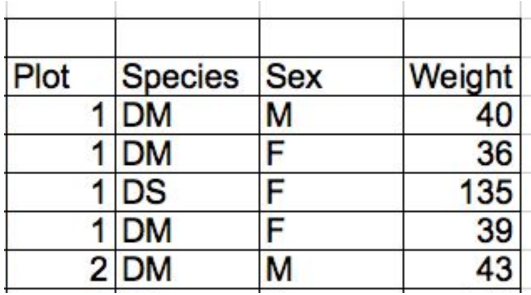
Data Organization in Spreadsheets

Some cardinal rules to correctly compile your data spreadsheet

2) Don’t mix multiple information in one cell



Plot	Species-Sex	Weight
1	DM-M	40
1	DM-F	36
1	DS-F	135
1	DM-F	39
2	DM-M	43



Plot	Species	Sex	Weight
1	DM	M	40
1	DM	F	36
1	DS	F	135
1	DM	F	39
2	DM	M	43

Data Organization in Spreadsheets

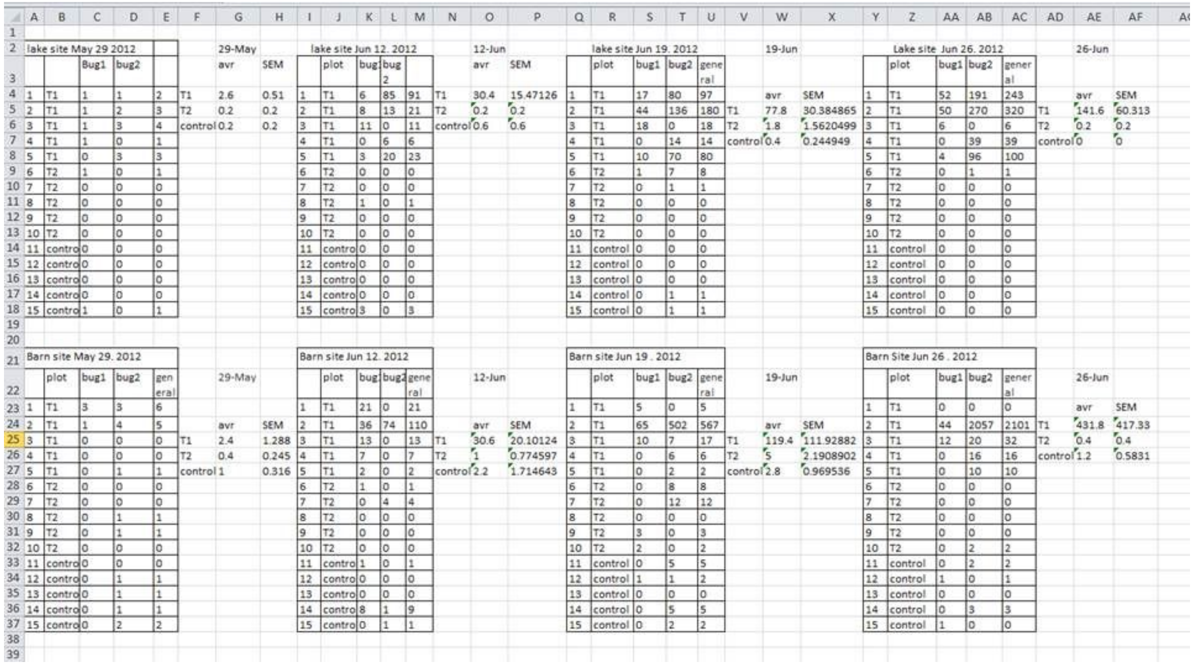
Some cardinal rules to correctly compile your data spreadsheet

3) **NEVER** touch the raw data! If needed make a copy and modify it.

4) Export and store your data as a text-based file (csv, tsv…)

Data Organization in Spreadsheets

Some common **errors**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2	lake site May 29 2012						29-May		lake site Jun 12. 2012						12-Jun	
3			Bug1	bug2			avr	SEM		plot	bug	bug2			avr	SEM
4	1	T1	1	1	2	T1	2.6	0.51	1	T1	6	85	91	T1	30.4	15.47126
5	2	T1	1	2	3	T2	0.2	0.2	2	T1	8	13	21	T2	0.2	0.2
6	3	T1	1	3	4	control	0.2	0.2	3	T1	11	0	11	control	0.6	0.6
7	4	T1	1	0	1				4	T1	0	6	6			
8	5	T1	0	3	3				5	T1	3	20	23			
9	6	T2	1	0	1				6	T2	0	0	0			
10	7	T2	0	0	0				7	T2	0	0	0			
11	8	T2	0	0	0				8	T2	1	0	1			
12	9	T2	0	0	0				9	T2	0	0	0			
13	10	T2	0	0	0				10	T2	0	0	0			
14	11	contro	0	0	0				11	contro	0	0	0			
15	12	contro	0	0	0				12	contro	0	0	0			
16	13	contro	0	0	0				13	contro	0	0	0			
17	14	contro	0	0	0				14	contro	0	0	0			
18	15	contro	1	0	1				15	contro	3	0	3			
19																
20																
21	Barn site May 29. 2012								Barn site Jun 12. 2012							
22		plot	bug1	bug2	general		29-May			plot	bug	bug2	general		12-Jun	
23	1	T1	3	3	6				1	T1	21	0	21			
24	2	T1	1	4	5		avr	SEM	2	T1	36	74	110		avr	SEM
25	3	T1	0	0	0	T1	2.4	1.288	3	T1	13	0	13	T1	30.6	20.10124
26	4	T1	0	0	0	T2	0.4	0.245	4	T1	7	0	7	T2	1	0.774597
27	5	T1	0	1	1	control	1	0.316	5	T1	2	0	2	control	2.2	1.714643
28	6	T2	0	0	0				6	T2	1	0	1			
29	7	T2	0	0	0				7	T2	0	4	4			
30	8	T2	0	1	1				8	T2	0	0	0			
31	9	T2	0	1	1				9	T2	0	0	0			
32	10	T2	0	0	0				10	T2	0	0	0			
33	11	contro	0	0	0				11	contro	1	0	1			
34	12	contro	0	1	1				12	contro	0	0	0			
35	13	contro	0	1	1				13	contro	0	0	0			
36	14	contro	0	1	1				14	contro	8	1	9			
37	15	contro	0	2	2				15	contro	0	1	1			
38																
39																

	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1																
2		lake site Jun 19. 2012					19-Jun			Lake site Jun 26. 2012					26-Jun	
3		plot	bug1	bug2	general					plot	bug1	bug2	general			
4	1	T1	17	80	97		avr	SEM	1	T1	52	191	243		avr	SEM
5	2	T1	44	136	180	T1	77.8	30.384865	2	T1	50	270	320	T1	141.6	60.313
6	3	T1	18	0	18	T2	1.8	1.5620499	3	T1	6	0	6	T2	0.2	0.2
7	4	T1	0	14	14	control	0.4	0.244949	4	T1	0	39	39	control	0	0
8	5	T1	10	70	80				5	T1	4	96	100			
9	6	T2	1	7	8				6	T2	0	1	1			
10	7	T2	0	1	1				7	T2	0	0	0			
11	8	T2	0	0	0				8	T2	0	0	0			
12	9	T2	0	0	0				9	T2	0	0	0			
13	10	T2	0	0	0				10	T2	0	0	0			
14	11	control	0	0	0				11	control	0	0	0			
15	12	control	0	0	0				12	control	0	0	0			
16	13	control	0	0	0				13	control	0	0	0			
17	14	control	0	1	1				14	control	0	0	0			
18	15	control	0	1	1				15	control	0	0	0			
19																
20																
21	Barn site Jun 19 . 2012								Barn Site Jun 26 . 2012							
22		plot	bug1	bug2	general		19-Jun			plot	bug1	bug2	general		26-Jun	
23	1	T1	5	0	5				1	T1	0	0	0		avr	SEM
24	2	T1	65	502	567		avr	SEM	2	T1	44	2057	2101	T1	431.8	417.33
25	3	T1	10	7	17	T1	119.4	111.92882	3	T1	12	20	32	T2	0.4	0.4
26	4	T1	0	6	6	T2	5	2.1908902	4	T1	0	16	16	control	1.2	0.5831
27	5	T1	0	2	2	control	2.8	0.969536	5	T1	0	10	10			
28	6	T2	0	8	8				6	T2	0	0	0			
29	7	T2	0	12	12				7	T2	0	0	0			
30	8	T2	0	0	0				8	T2	0	0	0			
31	9	T2	3	0	3				9	T2	0	0	0			
32	10	T2	2	0	2				10	T2	0	2	2			
33	11	control	0	5	5				11	control	0	2	2			
34	12	control	1	1	2				12	control	1	0	1			
35	13	control	0	0	0				13	control	0	0	0			
36	14	control	0	5	5				14	control	0	3	3			
37	15	control	0	2	2				15	control	1	0	0			
38																
39																

1) Using multiple tables

The computer reads your table "by row".

Here, a computer will assign to the same sample values from 4 different samples!

Data Organization in Spreadsheets

2) **Using multiple tabs**

This can look tidy but does not allows you to make data communicating in different tabs. Sooner or later you’ll need to collapse all your data in a single table.

3) **Do not properly indicate real zeros and missing data**

- write always all the real zeros
- leave blank (or fill with **NA** values) if data is missing

Data Organization in Spreadsheets

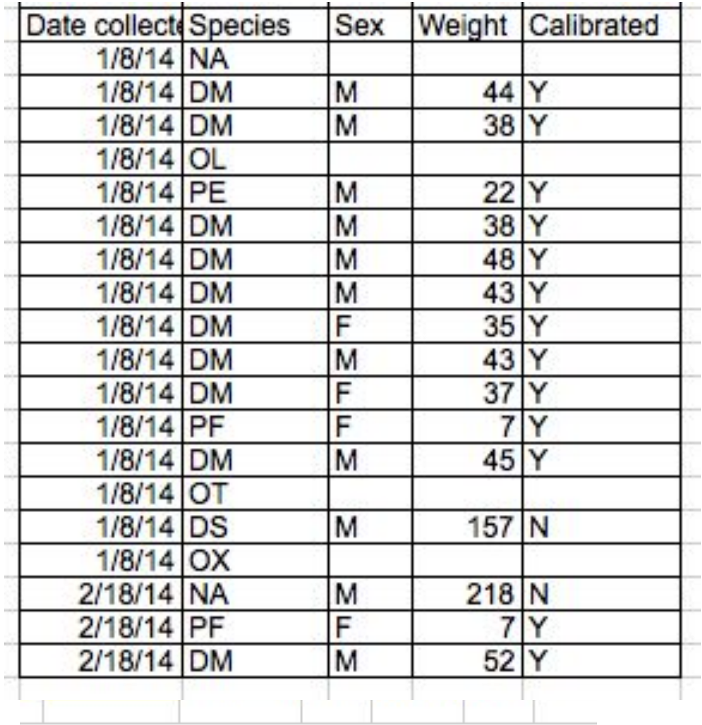
4) **Do not use formatting to convey information!**

- it will be lost when exporting your table in a text file

Solution:

Add a new variable encoding which observation will need to be excluded from the analysis.

More in general:
Don't be afraid to add as much as variables are needed to properly annotate your sample



Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Data Organization in Spreadsheets

5) Do not merge cells!

It will create artifacts or issues when exporting into a text file.

Solution: re-structure your data such as merging cells is not required

- In my experience this is commonly used in table headers!

Data Organization in Spreadsheets

6) **Headers should be one line**

- see the previous point
- column names should avoid problematic characters
  - symbols (°, ?, %, !, +,[], ())
  - spaces
- use underscore (_) or **camel case** notations

Example:

Root diameter (mm) -> Root_diameter or RootDiameter

- keep it as simple as possible: e.g. RD.

<u>You'll need an annotation file to track the meaning of your codes!</u>

Data Organization in Spreadsheets

6) do not includes measure units in your data spreadsheet

Measure units are essential, but:

- do not include in your data (your observations can have all the same measure unit).

  If not so: can you convert them to the same unit? Otherwise add a variable indicating the measure unit for each of your observation.

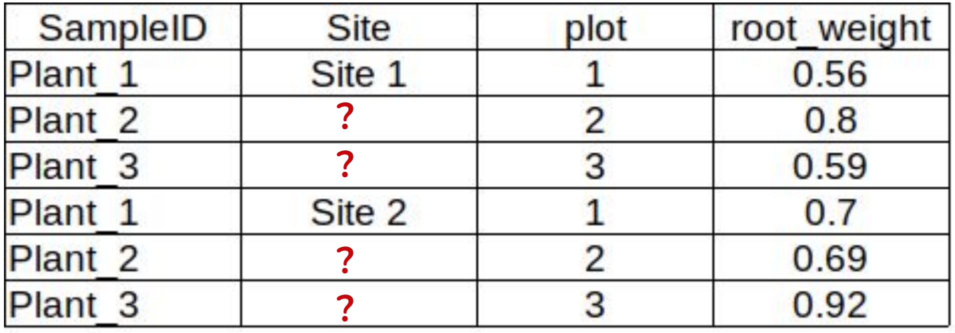
- do not include in your column header.

  Compile e README file writing annotation of your column names.

Data Organization in Spreadsheets

7) **Write your annotations for every sample**

- Computers are very literal. If you do not write in each row sample information, your computer won't understand where is the sample from



SampleID	Site	plot	root_weight
Plant_1	Site 1	1	0.56
Plant_2	?	2	0.8
Plant_3	?	3	0.59
Plant_1	Site 2	1	0.7
Plant_2	?	2	0.69
Plant_3	?	3	0.92

Each row must be unique!

8) **Include your replicate number, but only for tracking purposes**

Most of the analyses do not require a replicate number!

Often they are stored along with the sample name -> split in a new variable!

Some notes about date/hour formatting

- Storing dates/times in one field in the format (“15/01/2024”) can cause compatibility issues between softwares

- Storing dates as YEAR, MONTH, DAY in separate columns eliminates any ambiguities!
- as a single string YYYYMMDDhhmmss format (or YYYYMMDD for date only)

- as YEAR, DAY-OF-YEAR (**DOI**):

“=A1-DATE(YEAR(A1);1;0)” where A2 is the date”

see

> library(anytime) # in R for format conversion!

Data Organization in Spreadsheets

Do your exercise!

```
>download.file("https://ndownloader.figshare.com/files/2252083",
"survey_data_spreadsheet_messy.xls")
```

Solution

```
>download.file("https://raw.githubusercontent.com/mchialva/PhDToolbox2024/main/Datasets/survey_data_spreadsheet_tidy.xlsx", "survey_data_spreadsheet_tidy.xlsx")
```