



CC BY, Icons by <https://icons8.com>

## 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 spatial 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!
- 4) Export and store your data as a text-based file (csv, tsv...)

## Some common errors

[illegible]

## 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 allow 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 collect	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

RootDiameter

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

You'll need an annotation file to track the meaning of your codes!



# 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!



# Data Organization in Spreadsheets

Do your exercise!

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