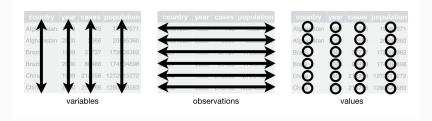
# Tidy data

Francisco Rodriguez-Sanchez

@frod\_san

https://frodriguezsanchez.net

### Tidy data



year	cases	country	1999	2000	
1999	745	Afghanistan	745	<b>2</b> 666	
2000	2666	Brazil	37737	80488	
1999	37737	China	212258	213766	
2000	80488				
1999	212258				
2000	213766		table4		
	1999 2000 1999 2000 1999	1999 745 2000 2666 1999 37737 2000 80488 1999 212258	1999 745 Agrandson 2000 2666 Brazil 1999 37737 China 2000 80488 1999 212258	1999 745 Afghanistan 745 2000 2666 Brazil 37737 1999 37737 China 212258 2000 80488 1999 212258	

#### Be careful with data entry and management in Excel!

COMMENT Open Access

# Gene name errors are widespread in the scientific literature



Mark Ziemann<sup>1</sup>, Yotam Eren<sup>1,2</sup> and Assam El-Osta<sup>1,3\*</sup>

#### Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (xls and.xks suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene sym-

· Put variables in columns (things you are measuring: height, weight, sex)

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- $\cdot$   $\,$  Avoid spaces, numbers, and special characters in column names.

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- Always  $\mbox{\it write zero values},$  to distinguish from blank/missing data.

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- · Use blank/empty cells, or NA, for missing data.
- · Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- · Use blank/empty cells, or NA, for missing data.
- · Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- · Use Data validation in Excel (or GForms) to constrain data entry to accepted values.

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- · Use blank/empty cells, or NA, for missing data.
- · Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- · Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- · Don't combine multiple pieces of information in one cell.

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- · Use blank/empty cells, or NA, for missing data.
- · Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- · Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- · Don't combine multiple pieces of information in one cell.
- · Don't touch raw data. Do all data manipulation through code.

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- · Use blank/empty cells, or NA, for missing data.
- · Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- · Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- · Don't combine multiple pieces of information in one cell.
- · Don't touch raw data. Do all data manipulation through code.
- · Export data as plain text (txt, csv).

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- · Use blank/empty cells, or NA, for missing data.
- · Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- · Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- · Don't combine multiple pieces of information in one cell.
- · Don't touch raw data. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- · Use blank/empty cells, or NA, for missing data.
- · Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- · Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- · Don't combine multiple pieces of information in one cell.
- · Don't touch raw data. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- http://kbroman.org/dataorg/

- · Put variables in columns (things you are measuring: height, weight, sex)
- · Each observation in one row (e.g. individuals).
- · Avoid spaces, numbers, and special characters in column names.
- · Always write zero values, to distinguish from blank/missing data.
- · Use blank/empty cells, or NA, for missing data.
- · Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- · Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- · Don't combine multiple pieces of information in one cell.
- · Don't touch raw data. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- · http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- http://kbroman.org/dataorg/
- · Broman & Woo: Data organization in spreadsheets

# Common spreadsheet errors

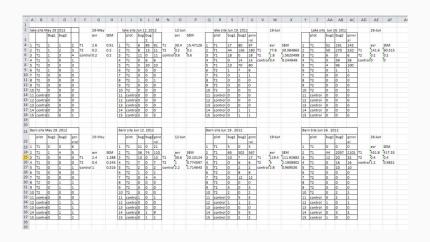
## More than one variable per column

Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146

Date collected	Plot	Species	Sex	Weight
1/9/78	1	DM	M	40
1/9/78	1	DM	F	36
1/9/78	1	DS	F	135
1/20/78	1	DM	F	39
1/20/78	2	DM	M	43
1/20/78	2	DS	F	144
3/13/78	2	DM	F	51
3/13/78	2	DM	F	44
3/13/78	2	DS	F	146

Source: Data Carpentry

#### Multiple tables



# Multiple tabs

Could you avoid new tab by adding a column to original spreadsheet?

# Using formatting, comments, etc to convey information

Plot: 2				
Date collecte	Species	Sex	Weight	
1/8/14	NA			
1/8/14		M	44	
1/8/14		M	38	
1/8/14				
1/8/14		M	22	
1/8/14	DM	M	38	
1/8/14	DM	M	48	
1/8/14		M	43	
1/8/14		F	35	
1/8/14		M	43	
1/8/14		F	37	
1/8/14	PF	F	7	
1/8/14		M	45	
1/8/14				
1/8/14	DS	M	157	
1/8/14	OX			
2/18/14	NA	M	218	
2/18/14		F	7	
2/18/14	DM	M	52	

measurement	device no	t calibrated
-------------	-----------	--------------

Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14		M	44	
1/8/14	DM	M	38	Υ
1/8/14				
1/8/14		M	22	
1/8/14		M	38	
1/8/14		M	48	
1/8/14		M	43	
1/8/14		F	35	
1/8/14		M	43	
1/8/14		F	37	
1/8/14		F		Υ
1/8/14		M	45	Υ
1/8/14	OT			
1/8/14		M	157	N
1/8/14				
2/18/14		M	218	
2/18/14		F		Υ
2/18/14	DM	M	52	Υ

Your turn: tidy up this messy dataset

https://ndownloader.figshare.com/files/2252083