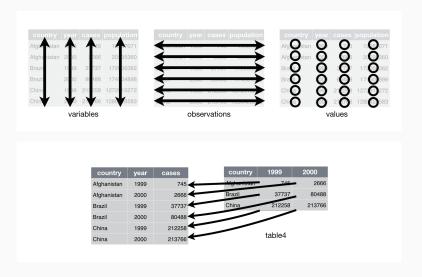
Organización y manejo de datos

Francisco Rodriguez-Sanchez

 $@frod_san$

Tidy data



http://r4ds.had.co.nz/tidy.html

Mucho cuidado con Excel!

COMMENT Open Access

Gene name errors are widespread in the scientific literature



Mark Ziemann¹, Yotam Eren^{1,2} and Assam El-Osta^{1,3*}

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 (xds 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-

Cada columna es una variable (altura, peso, sexo, etc)

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.
- Fechas: pon año, mes, día en columnas separadas, o YYYY-MM-DD como texto.

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.
- Fechas: pon año, mes, día en columnas separadas, o YYYY-MM-DD como texto.
- Usa Data validation en Excel (or GForms) para restringir valores de entrada de datos.

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.
- Fechas: pon año, mes, día en columnas separadas, o YYYY-MM-DD como texto.
- Usa Data validation en Excel (or GForms) para restringir valores de entrada de datos.
- No combines información dentro de las celdas.

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.
- Fechas: pon año, mes, día en columnas separadas, o YYYY-MM-DD como texto.
- Usa Data validation en Excel (or GForms) para restringir valores de entrada de datos.
- No combines información dentro de las celdas.
- Nunca alteres los datos brutos. Modifica copia en Excel, o usa código.

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.
- Fechas: pon año, mes, día en columnas separadas, o YYYY-MM-DD como texto.
- Usa Data validation en Excel (or GForms) para restringir valores de entrada de datos.
- No combines información dentro de las celdas.
- Nunca alteres los datos brutos. Modifica copia en Excel, o usa código.
- Guarda los datos como texto plano (txt, csv).

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.
- Fechas: pon año, mes, día en columnas separadas, o YYYY-MM-DD como texto.
- Usa Data validation en Excel (or GForms) para restringir valores de entrada de datos.
- No combines información dentro de las celdas.
- Nunca alteres los datos brutos. Modifica copia en Excel, o usa código.
- Guarda los datos como texto plano (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.
- Fechas: pon año, mes, día en columnas separadas, o YYYY-MM-DD como texto.
- Usa Data validation en Excel (or GForms) para restringir valores de entrada de datos.
- No combines información dentro de las celdas.
- Nunca alteres los datos brutos. Modifica copia en Excel, o usa código.
- Guarda los datos como texto plano (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- http://kbroman.org/dataorg/

- Cada columna es una variable (altura, peso, sexo, etc)
- Cada fila es una observación (p. ej. individuos).
- Evita espacios, números, y caracteres especiales en nombres de columna.
- Anota siempre los valores cero, para distinguirlos de los datos ausentes.
- Datos ausentes: celda en blanco, o NA.
- Fechas: pon año, mes, día en columnas separadas, o YYYY-MM-DD como texto
- Usa Data validation en Excel (or GForms) para restringir valores de entrada de datos.
- No combines información dentro de las celdas.
- Nunca alteres los datos brutos. Modifica copia en Excel, o usa código.
- Guarda los datos como texto plano (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- http://kbroman.org/dataorg/
- Broman & Woo: Data organization in spreadsheets

Errores frecuentes

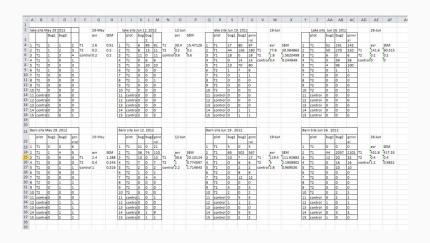
Más de una variable por columna

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	М	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

Múltiples tablas



Múltiples tablas

Mejor añadir una columna especificando el sitio, fecha, etc de muestreo

Usar comentarios, formato de celdas, etc

Plot: 2				
Date collecte	Species	Sex	Weight	
1/8/14	NA			
1/8/14	DM	M	44	
1/8/14	DM	M	38	
1/8/14	OL			
1/8/14	PE	M	22	
1/8/14	DM	M	38	
1/8/14	DM	M	48	
1/8/14	DM	M	43	
1/8/14	DM	F	35	
1/8/14	DM	M	43	
1/8/14	DM	F	37	
1/8/14	PF	F	7	
1/8/14	DM	M	45	
1/8/14	OT			
1/8/14	DS	M	157	
1/8/14	OX			
2/18/14	NA	M	218	
2/18/14	PF	F	7	
2/18/14	DM	M	52	
	measurem	ent dev	vice not o	alibrated

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

Tu turno: organiza mejor este dataset

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

Happy writing!



https://github.com/Pakillo/ReproducibleScience