

Organización y manejo de datos

Francisco Rodriguez-Sanchez

@frod_san

Tidy data

country	year	cases	population
Afghanistan	1999	37737	172206362
Afghanistan	2000	80488	174204898
Brazil	1999	212258	1272015272
Brazil	2000	213766	128201583

variables

country	year	cases	population
Afghanistan	1999	37737	172206362
Afghanistan	2000	80488	174204898
Brazil	1999	212258	1272015272
Brazil	2000	213766	128201583

observations

country	year	cases	population
Afghanistan	1999	37737	172206362
Afghanistan	2000	80488	174204898
Brazil	1999	212258	1272015272
Brazil	2000	213766	128201583

values

country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

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

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 (.xls and .xlsx 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-

Hojas de cálculo: recomendaciones

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

Hojas de cálculo: recomendaciones

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

Hojas de cálculo: recomendaciones

- 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.

Hojas de cálculo: recomendaciones

- 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.

Hojas de cálculo: recomendaciones

- 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.

Hojas de cálculo: recomendaciones

- 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.

Hojas de cálculo: recomendaciones

- 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.

Hojas de cálculo: recomendaciones

- 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.

Hojas de cálculo: recomendaciones

- 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.

Hojas de cálculo: recomendaciones

- 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).

Hojas de cálculo: recomendaciones

- 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/>

Hojas de cálculo: recomendaciones

- 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/>

Hojas de cálculo: recomendaciones

- 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	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

Múltiples tablas

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	
1																																		
2	lake site May 29 2012						29-May			lake site Jun 12 2012					12-Jun		lake site Jun 19 2012					19-Jun		Lake site Jun 26 2012								26-Jun		
3			bug1	bug2					SEM			plot	bug1	bug2	general				plot	bug1	bug2	general					plot	bug1	bug2	general				
4	1	T1	1	1	2	T1	2.6	0.51		1	T1	6	85	95	T1	30.4	15.47126	1	T1	17	80	97		avr	SEM	1	T1	52	191	243		avr	SEM	
5	2	T1	1	2	3	T2	0.2	0.2		2	T1	8	13	21	T2	0.2	0.2	2	T1	44	136	180	T1	77.8	30.384865	2	T1	50	270	320	T1	141.6	50.313	
6	3	T1	1	3	4		control	0.2	0.2	3	T1	11	0	11		control	0.6	0.6	3	T1	18	0	18	T2	1.8	1.5620499	3	T1	6	0	6	T2	0.2	0.2
7	4	T1	1	0	1					4	T1	0	6	6				4	T1	0	14	14		control	0.4	0.244949	4	T1	0	39	39	control	0	0
8	5	T1	0	3	3					5	T1	3	20	23				5	T1	10	70	80				5	T1	4	96	100				
9	6	T2	1	0	1					6	T2	1	0	1				6	T2	1	7	8				6	T2	0	1	1				
10	7	T2	0	0	0					7	T2	0	0	0				7	T2	0	1	1				7	T2	0	0	0				
11	8	T2	0	0	0					8	T2	1	0	1				8	T2	0	0	0				8	T2	0	0	0				
12	9	T2	0	0	0					9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0				
13	10	T2	0	0	0					10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0				
14	11	control	0	0	0					11	control	0	0	0				11	control	0	0	0				11	control	0	0	0				
15	12	control	0	0	0					12	control	0	0	0				12	control	0	0	0				12	control	0	0	0				
16	13	control	0	0	0					13	control	0	0	0				13	control	0	0	0				13	control	0	0	0				
17	14	control	0	0	0					14	control	0	0	0				14	control	0	1	1				14	control	0	0	0				
18	15	control	1	0	1					15	control	3	0	3				15	control	0	1	1				15	control	0	0	0				
19																																		
20																																		
21	Barn site May 29 2012						29-May			Barn site Jun 12 2012					12-Jun		Barn site Jun 19 2012					19-Jun		Barn Site Jun 26 2012								26-Jun		
22		plot	bug1	bug2	general						plot	bug1	bug2	general					plot	bug1	bug2	general					plot	bug1	bug2	general				
23	1	T1	3	3	6					1	T1	21	0	21				1	T1	5	0	5				1	T1	0	0	0		avr	SEM	
24	2	T1	1	4	5		avr	SEM		2	T1	36	74	110		avr	SEM	2	T1	65	502	567		avr	SEM	2	T1	44	2057	2101	T1	431.8	417.33	
25	3	T1	0	0	0	T1	2.4	1.288		3	T1	13	0	13	T1	30.6	10.10124	3	T1	10	7	17	T1	119.4	11.92882	3	T1	12	20	32	T2	0.4	0.4	
26	4	T1	0	0	0	T2	0.4	0.245		4	T1	7	0	7	T2	5	0.774597	4	T1	0	16	6	T2	5	1.1908902	4	T1	0	16	16	control	1.2	0.5831	
27	5	T1	0	1	1	control	1	0.316		5	T1	2	0	2	control	2.2	1.714643	5	T1	0	2	2	control	2.8	0.969536	5	T1	0	10	10				
28	6	T2	0	0	0					6	T2	1	0	1				6	T2	0	8	8				6	T2	0	0	0				
29	7	T2	0	0	0					7	T2	0	4	4				7	T2	0	12	12				7	T2	0	0	0				
30	8	T2	0	1	1					8	T2	0	0	0				8	T2	0	0	0				8	T2	0	0	0				
31	9	T2	0	1	1					9	T2	0	0	0				9	T2	3	0	3				9	T2	0	0	0				
32	10	T2	0	0	0					10	T2	0	0	0				10	T2	2	0	2				10	T2	0	2	2				
33	11	control	0	0	0					11	control	1	0	1				11	control	0	5	5				11	control	0	2	2				
34	12	control	0	1	1					12	control	0	0	0				12	control	1	1	2				12	control	1	0	1				
35	13	control	0	1	1					13	control	0	0	0				13	control	0	0	0				13	control	0	0	0				
36	14	control	0	1	1					14	control	8	1	9				14	control	0	5	5				14	control	0	3	3				
37	15	control	2	2	2					15	control	0	1	1				15	control	0	2	2				15	control	1	0	0				
38																																		
39																																		

Múltiples tablas

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

Usar comentarios, formato de celdas, etc

Plot: 2					
Date collect	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		
	measurement device not calibrated				

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

Tu turno: organiza mejor este dataset

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

Happy writing!



<https://github.com/Pakillo/ReproducibleScience>