R: A Hitchhikers Guide to Reproducible Research

- My favourite mistake

Brendan Palmer,
Clinical Research Facility - Cork &
School of Public Health
@B_A_Palmer

CRF-C
HRB Clinical Research Facility Cork

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh
School of Public Health

Some of my best friends use spreadsheets

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], 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-

cough We've known for a long time *cough*

BMC Bioinformatics

BioMed Central

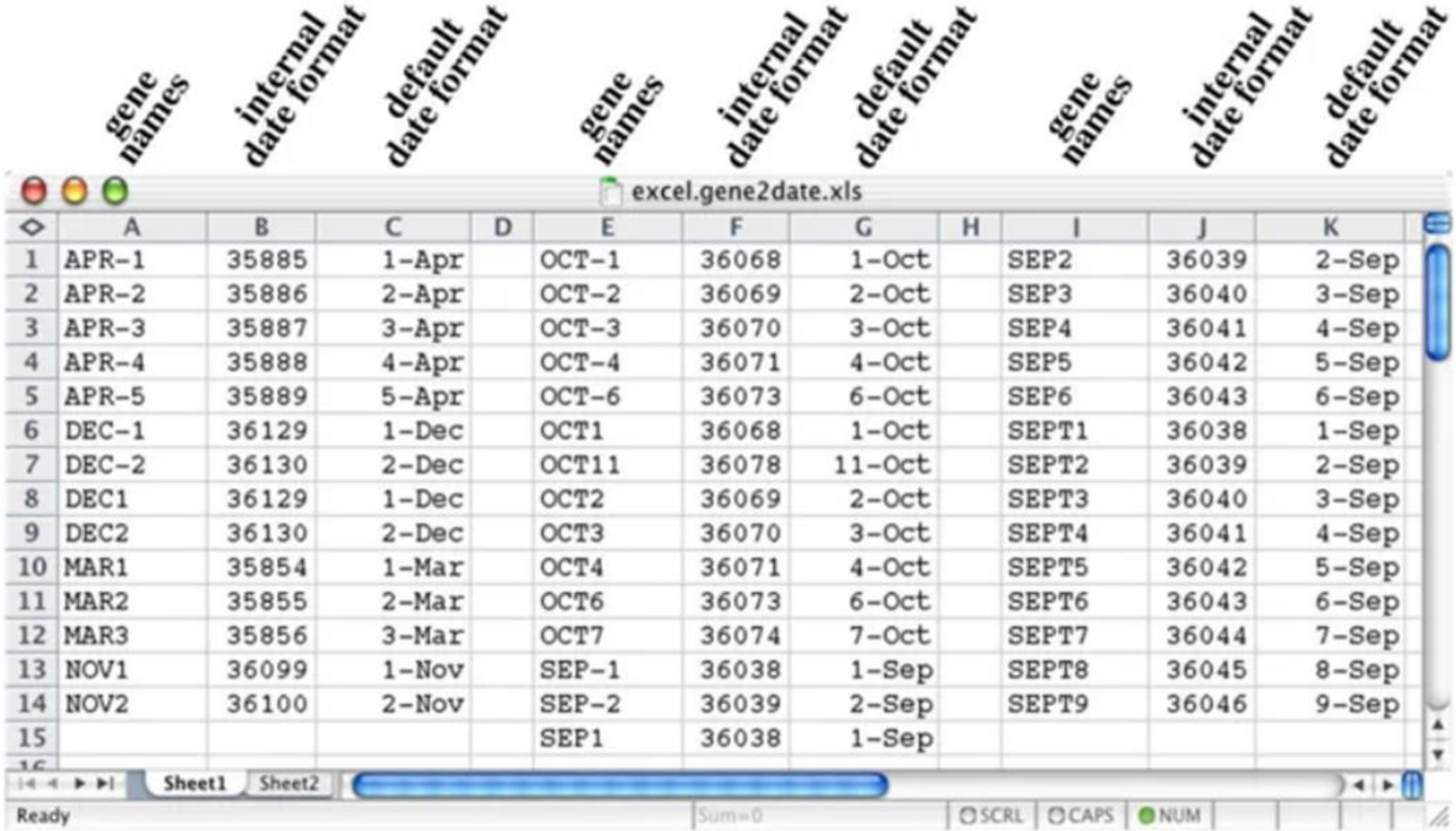
Correspondence

Open Access

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg[†1], Joseph Riss[†2], David W Kane[3], Kimberly J Bussey[1], Edward Uchio[4], W Marston Linehan[4], J Carl Barrett[2] and John N Weinstein[*1]

2004



	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

Excel also frequently gets clipboard amnesia

25

The answer, unfortunately, is **no**, you can't stop this from happening.

As described by Joel Spolsky, developer and program manager for excel:

> The official reason is that Excel doesn't really have cut and paste, it has move and copy. That's necessary because Excel automatically does reference fix up. For example, if cell A2 is defined as =A1, and you move cell A1 to A3, cell A2 will be updated to =A3.
>
> If Excel actually cut things to the clipboard you would somehow need to have a reference pointing >into< the clipboard which is bizarre and for which there is no reasonable syntax. In other words, Excel doesn't want to leave you with dangling references during a move operation and isn't confident that it would be able to fix them up correctly when you completed the move by selecting "Paste."
>
> Joel Spolsky 3/9/2004

source

What this means is that because of the difficulty inherent in the way excel maintains *references*, at the time of development there was no good way to store these references outside of excel and have them remain dynamic to be re-inserted. Once you change *focus* excel's ability to retain your original references is lost.

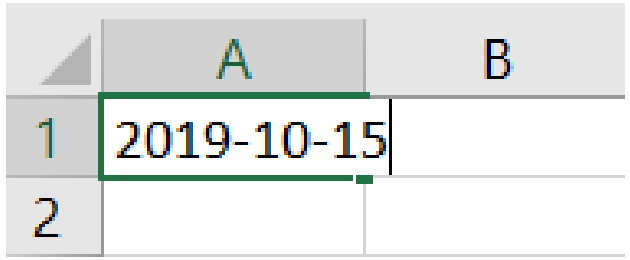
Unfortunately, MS does not consider this a bug.

But it doesn't end there

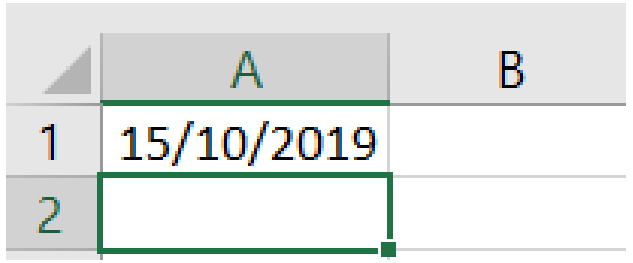
Date and time expressed according to ISO 8601 [refresh]

Date	2019-10-15
Date and time in UTC	2019-10-15T19:49:52+00:00 2019-10-15T19:49:52Z 20191015T194952Z
Week	2019-W42
Date with week number	2019-W42-2
Date without year	--10-15[1]
Ordinal date	2019-288

- YYYY-MM-DD or YYYYMMDD
- Type this into Excel

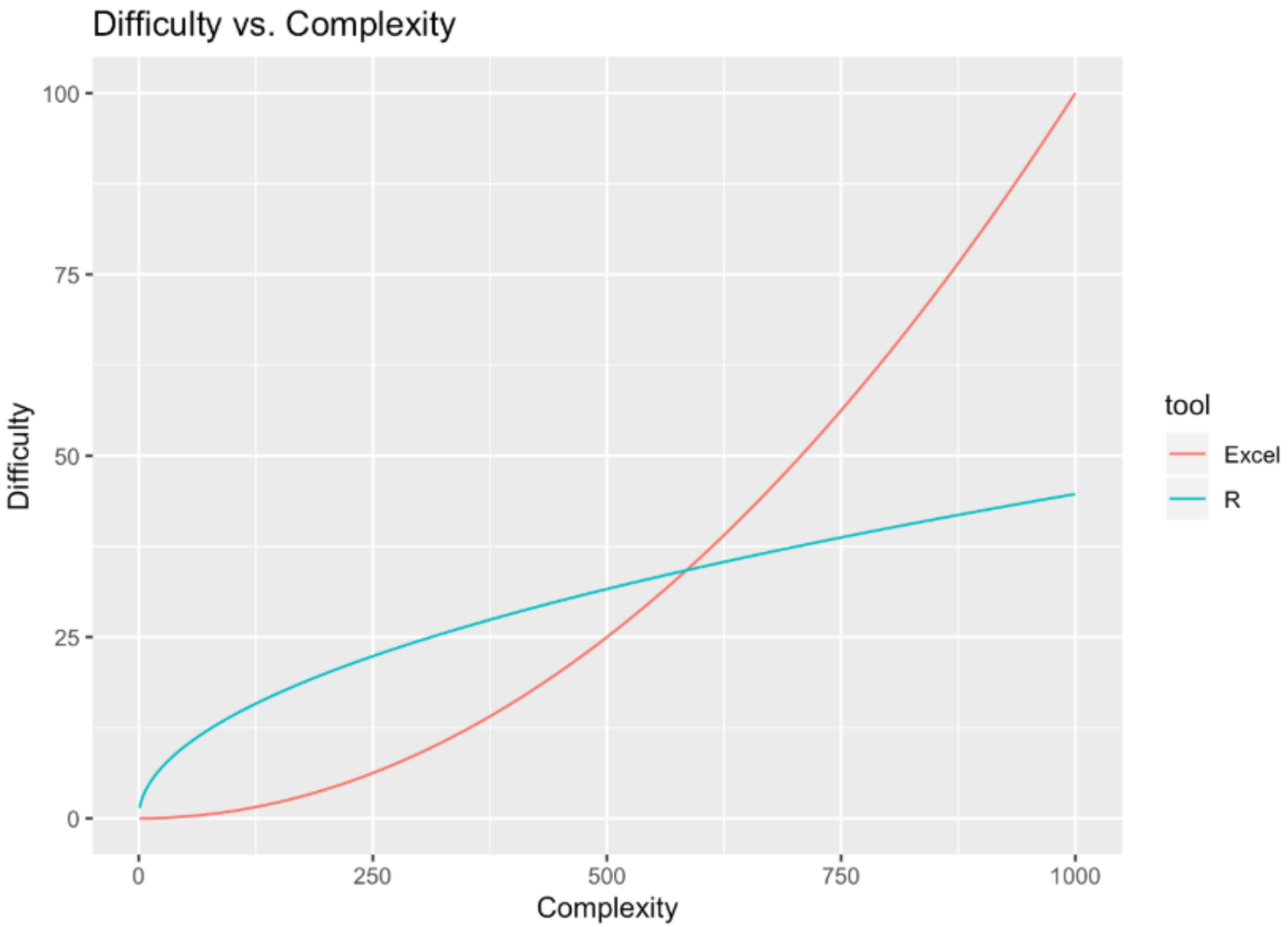


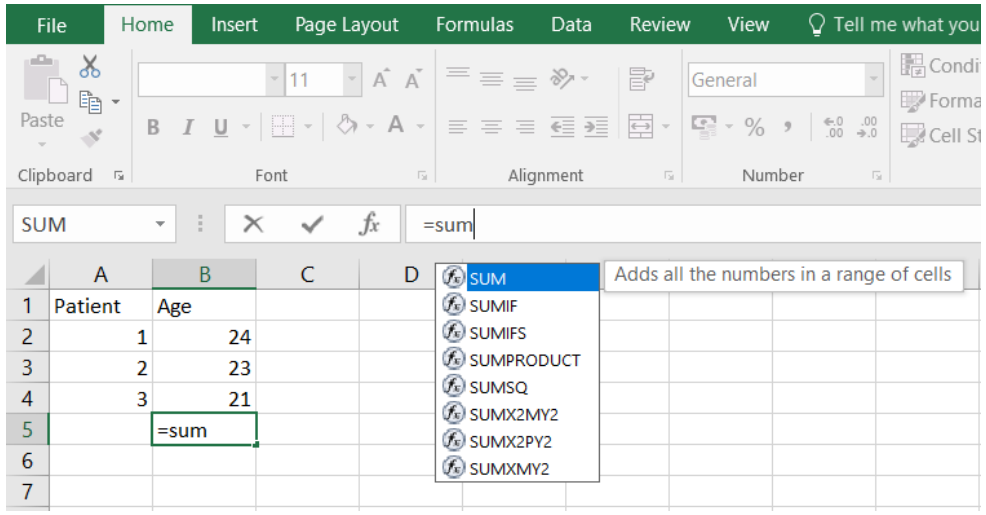
- And hit return



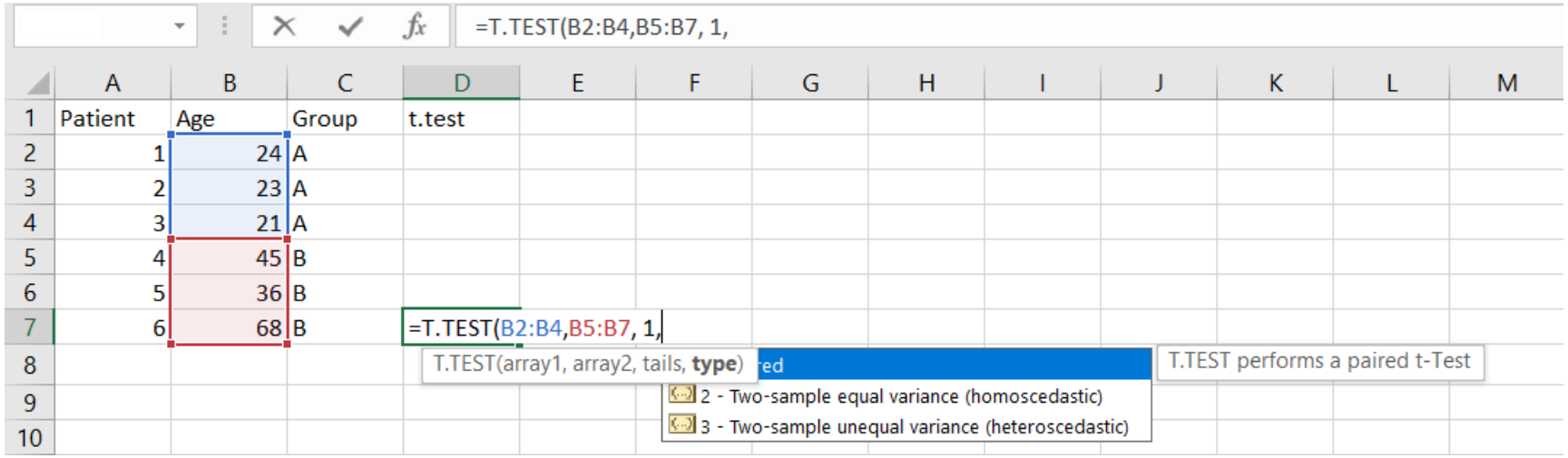
- DD/MM/YYYY

Excel is intuitive to use

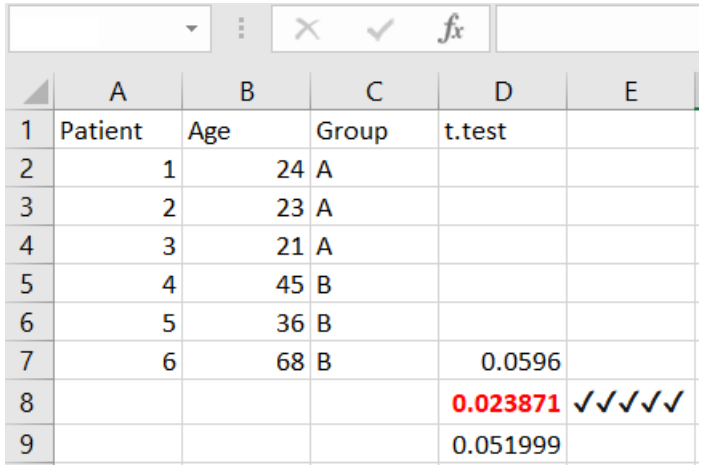




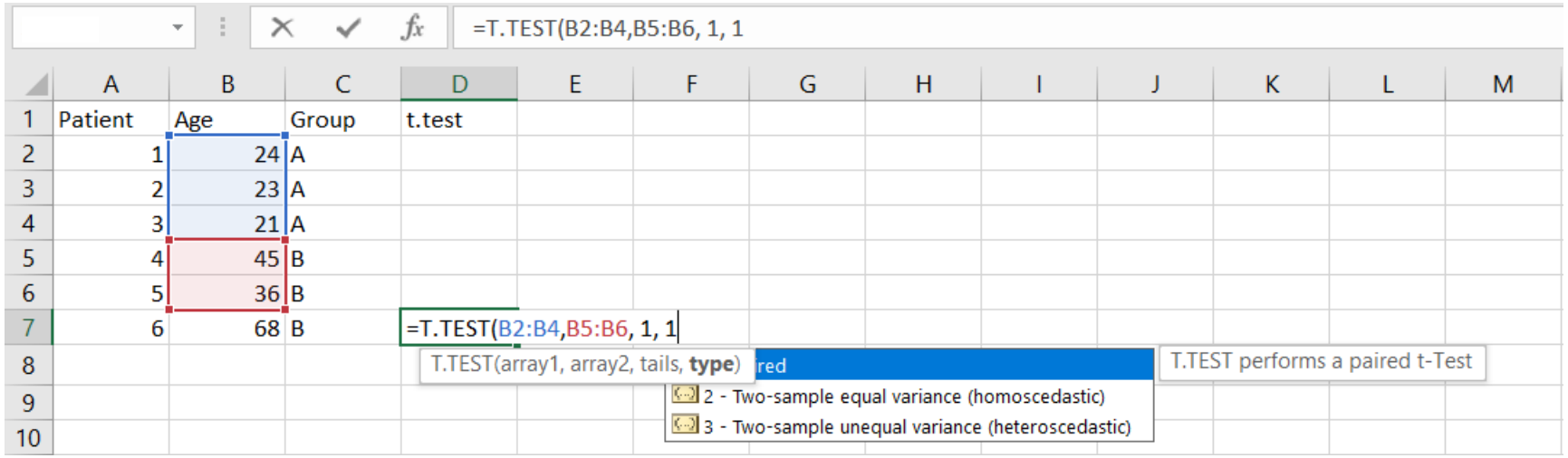
But a breeding ground for errors



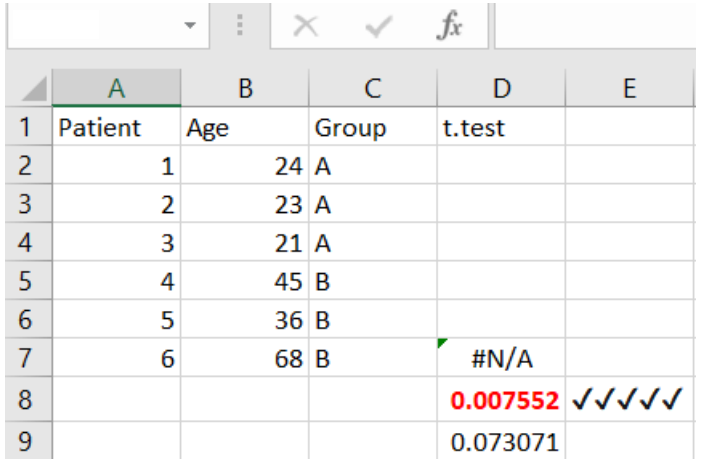
	A	B	C	D
1	Patient	Age	Group	t.test
2	1	24	A	
3	2	23	A	
4	3	21	A	
5	4	45	B	
6	5	36	B	
7	6	68	B	=T.TEST(B2:B4,B5:B7, 1,
8				
9				
10				



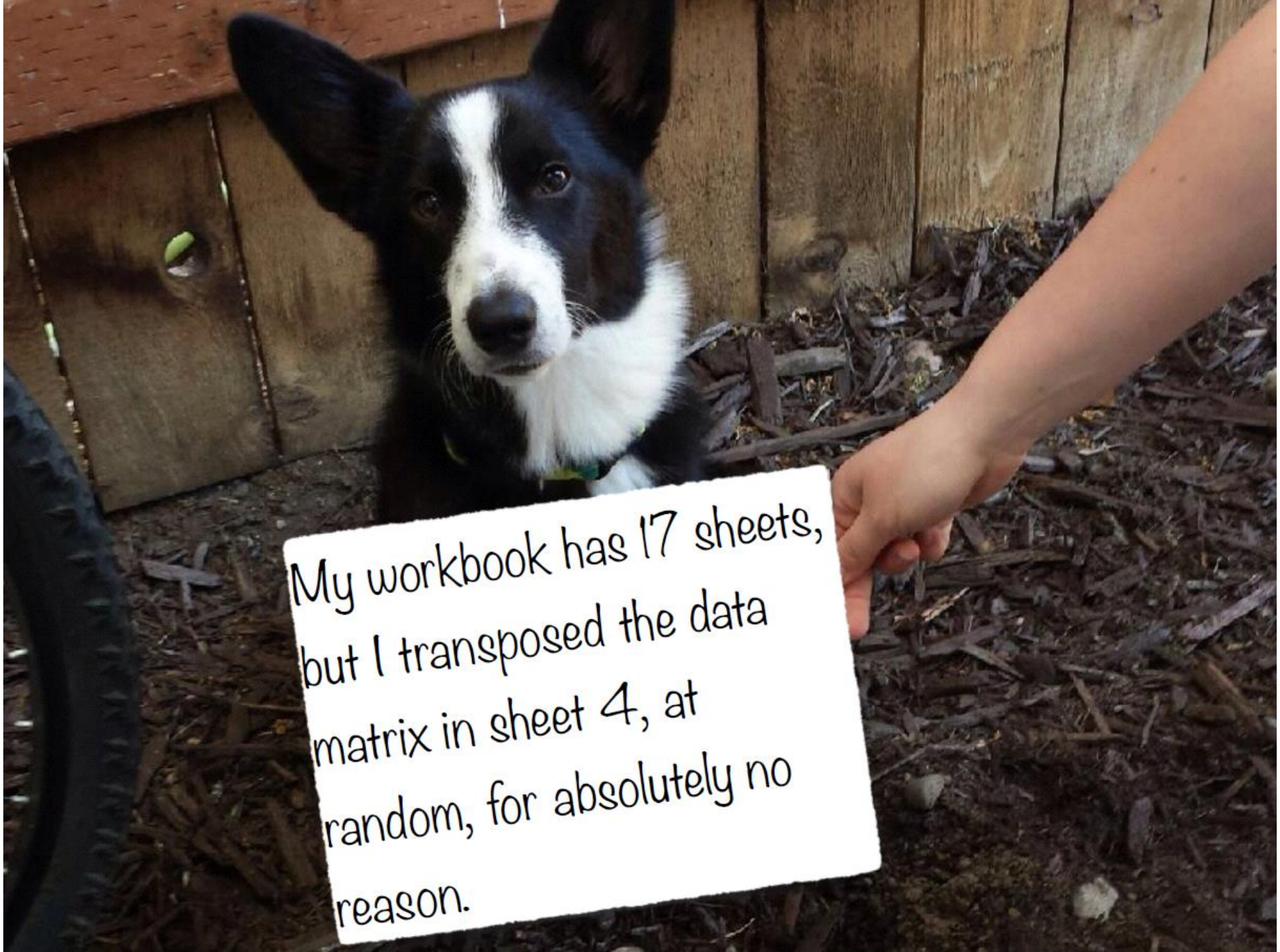
	A	B	C	D	E
1	Patient	Age	Group	t.test	
2	1	24	A		
3	2	23	A		
4	3	21	A		
5	4	45	B		
6	5	36	B		
7	6	68	B	0.0596	
8				0.023871	✓✓✓✓✓
9				0.051999	

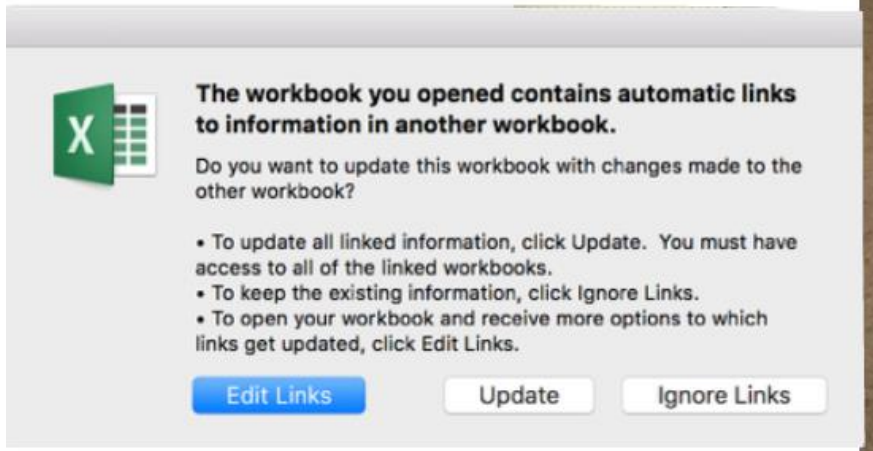


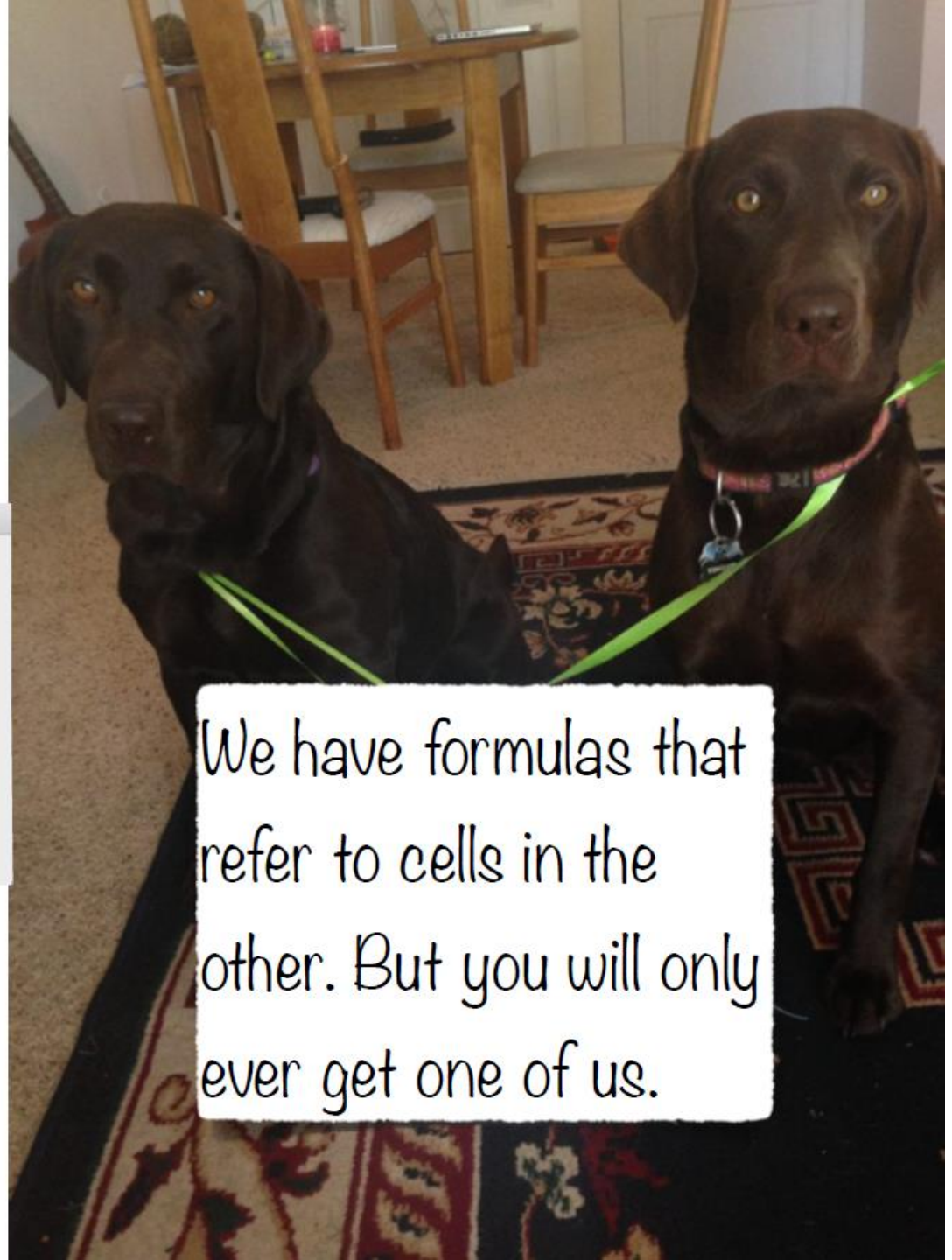
	A	B	C	D
1	Patient	Age	Group	t.test
2	1	24	A	
3	2	23	A	
4	3	21	A	
5	4	45	B	
6	5	36	B	
7	6	68	B	=T.TEST(B2:B4,B5:B6, 1, 1
8				
9				
10				

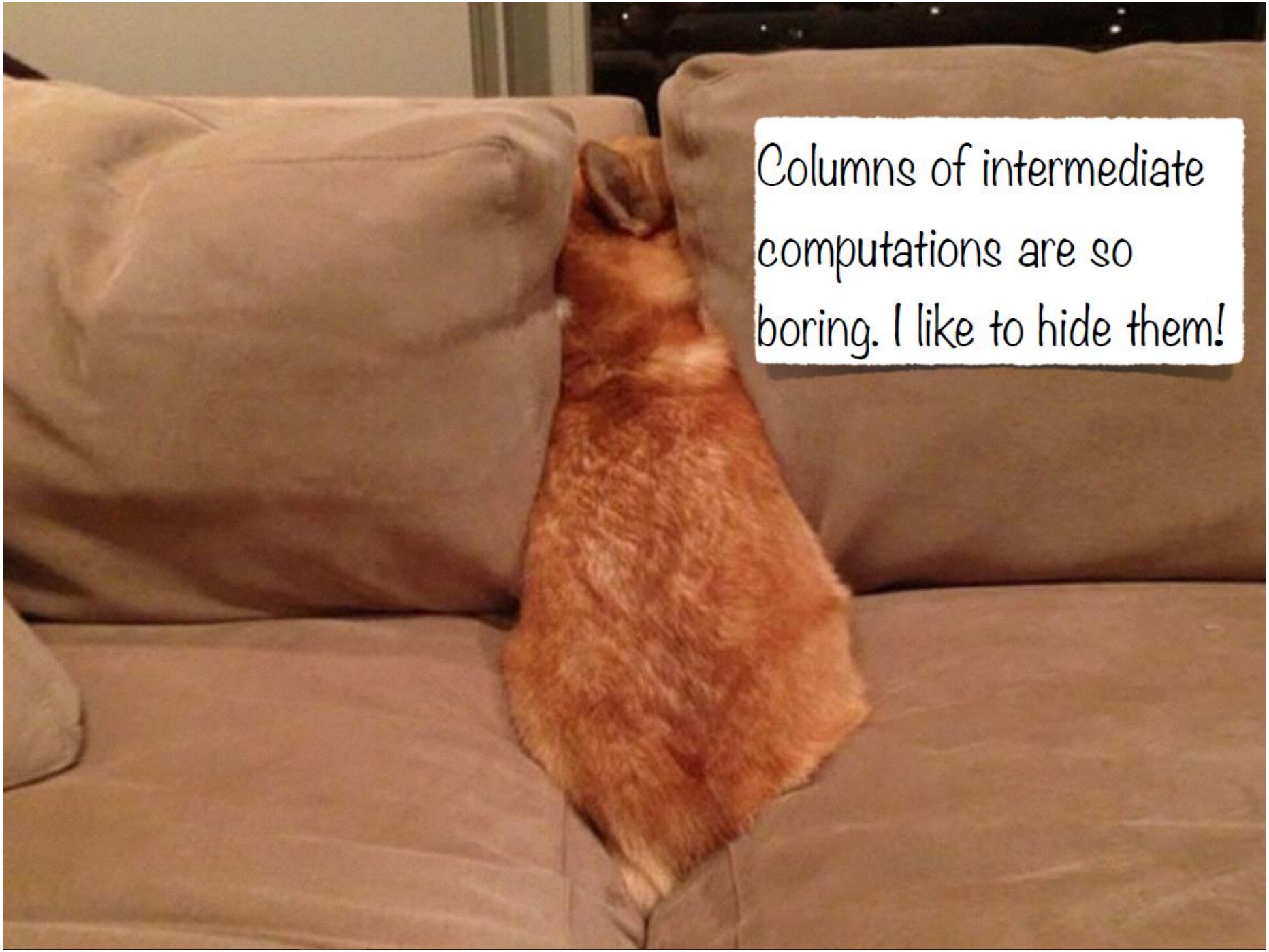


	A	B	C	D	E
1	Patient	Age	Group	t.test	
2	1	24	A		
3	2	23	A		
4	3	21	A		
5	4	45	B		
6	5	36	B		
7	6	68	B	#N/A	
8				0.007552	✓✓✓✓✓
9				0.073071	


My workbook has 17 sheets, but I transposed the data matrix in sheet 4, at random, for absolutely no reason.


The workbook you opened contains automatic links to information in another workbook.
Do you want to update this workbook with changes made to the other workbook?
• To update all linked information, click Update. You must have access to all of the linked workbooks.
• To keep the existing information, click Ignore Links.
• To open your workbook and receive more options to which links get updated, click Edit Links.
Edit Links
Update
Ignore Links

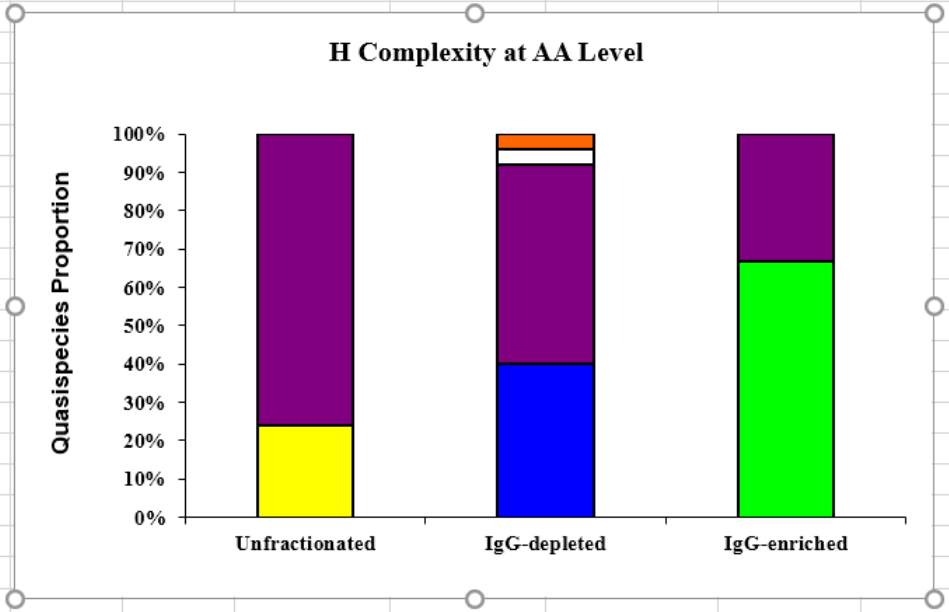

We have formulas that refer to cells in the other. But you will only ever get one of us.


Columns of intermediate computations are so boring. I like to hide them!

Amino Acid Complexity Across Fractions

Species Name	Fraction	Sequence Length (bp)	Original Clone
HUF-1	UF	318	HUF-1/HUF-2/HUF-3/HUF-4/HUF-5/HUF-6/HUF-7/HUF-8/HUF-9/HUF-10
HUF-2	UF	321	HUF-13/HUF-15/HUF-16/HUF-17/HUF-18/HUF-19/HUF-20/HUF-21/HUF-22/HUF-23/HUF-24/HUF-7/HUF-11/HUF-18/HUF-10/HUF-16/HUF-17/HUF-23/HUF-24/HUF-1/HUF-3/HUF-6/HUF-8/HUF-12/HUF-14
HU-1	U1	318	HU-1/HU-3/HU-4/HU-10/HU-12/HU-13/HU-14/HU-17/HU-18/HU-22
HU-2	U1	321	HU-2/HU-5/HU-6/HU-7/HU-11/HU-19/HU-22/HU-24
HU-3	U1	321	HU-20
HU-4	U1	321	HU-8
HB-1	B	321	HB-2/HB-4/HB-5/HB-6/H-8/HB-10
HB-2	B	321	HB-3/HB-7/HB-9

	Unfractionated	IgG-depleted	IgG-enriched	Colour
HUF-1	10			
HUF-2	32			
HU-1		10		
HU-2		13		
HU-3		1		
HU-4		1		
HB-1			6	
HB-2			3	



Taking small steps to achieve big changes

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

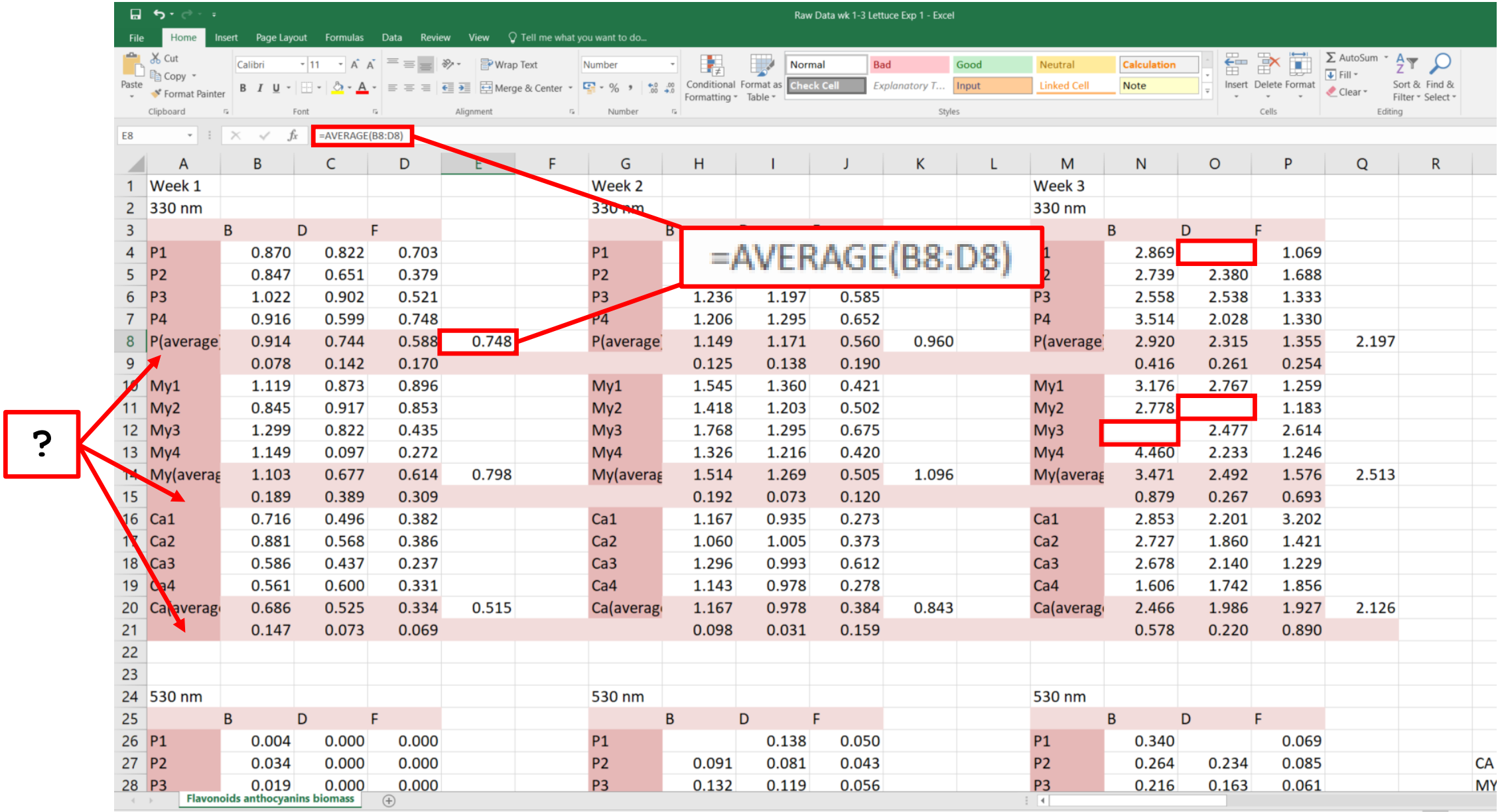
KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

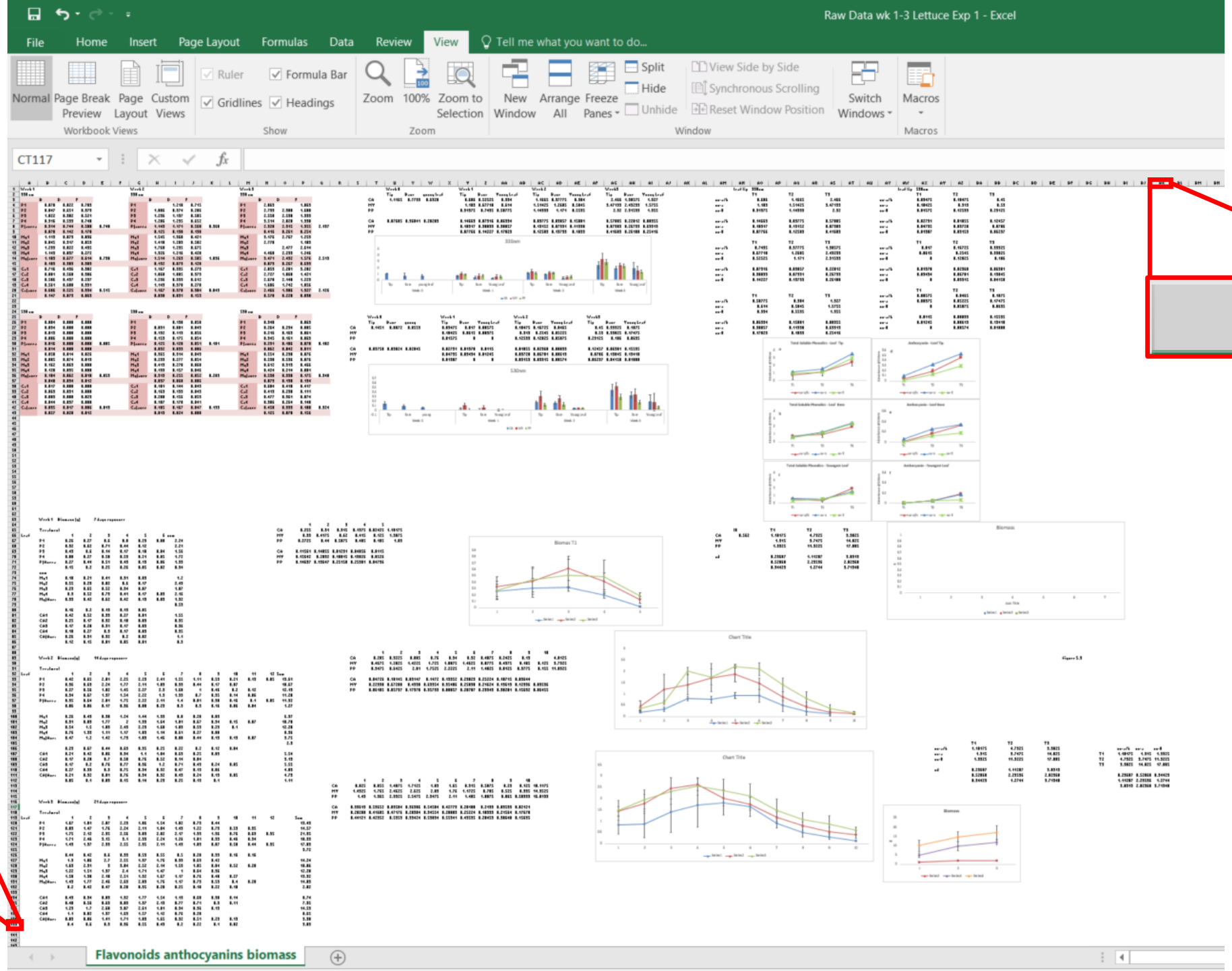
Our real life experiment…

- UV light has potential to change the secondary metabolite composition (colour) of bronze/red lettuce

- Experimental setup:
  - 3 lettuce varieties
  - 3 UV filter conditions
  - 3 week duration

Real data comes with real problems




Raw Data wk 1-3 Lettuce Exp 1 - Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Tell me what you want to do...
Normal
Page Break Preview
Page Layout
Custom Views
Workbook Views
Ruler
Gridlines
Formula Bar
Headings
Show
Zoom
100%
Zoom to Selection
Zoom
New Window
Arrange All
Freeze Panes
Split
Hide
Unhide
View Side by Side
Synchronous Scrolling
Reset Window Position
Window
Switch Windows
Macros
CT117
Flavonoids anthocyanins biomass

BK

139

Raw Data wk 1-3 Lettuce Exp 1 - Excel

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

Normal Page Break Preview Page Layout Custom Views — Workbook Views

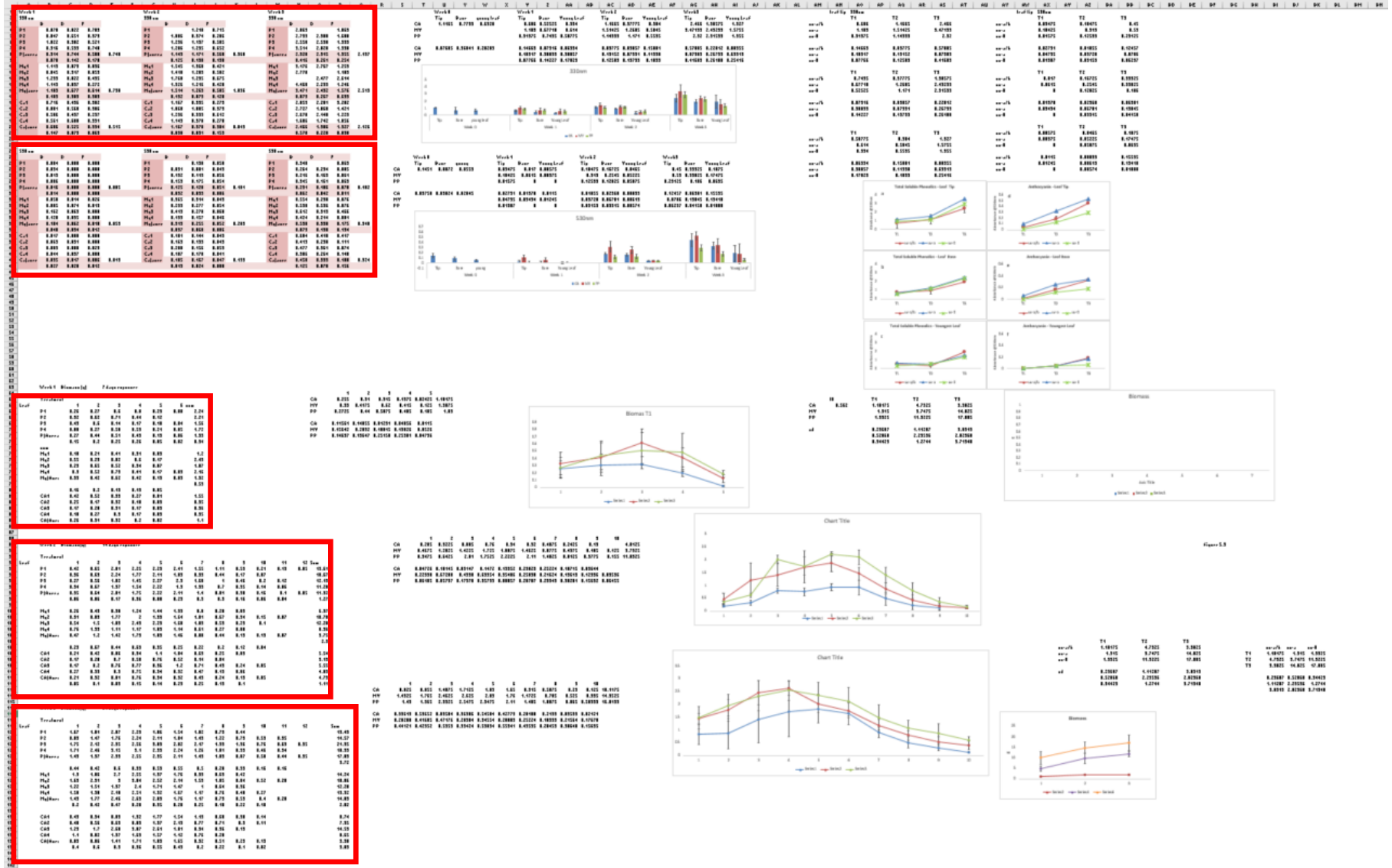
Ruler Formula Bar Gridlines Headings — Show

Zoom 100% Zoom to Selection — Zoom

New Window Arrange All Freeze Panes Split Hide Unhide View Side by Side Synchronous Scrolling Reset Window Position Switch Windows — Window

Macros — Macros

CT117



Flavonoids anthocyanins biomass

Less stress, more success

	A	B	C	D	E	F	G	H	I	J	K
1	id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator	
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly	
3	2	0	ptp	nofilter	2	1.1805	0.42	cos	2019/04/01	Darren Dahly	
4	3	0	ptp	nofilter	3	1.0345	0.62	cos	2019/04/01	Darren Dahly	
5	4	0	ptp	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer	
6	5	0	my	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer	
7	6	0	my	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer	
8	7	0	my	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer	
9	8	0	my	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer	
10	9	0	ca	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer	
11	10	0	ca	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer	
12	11	0	ca	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer	
13	12	0	ca	nofilter	4	1.094	0.63	cos	2019/04/01	Darren Dahly	
14	13	1	ptp	filter	1	0.87	0.76	cos	2019/04/08	Darren Dahly	
15	14	1	ptp	filter	2	0.847	0.95	cos	2019/04/08	Darren Dahly	
16	15	1	ptp	filter	3	1.022	0.95	cos	2019/04/08	Darren Dahly	
17	16	1	ptp	filter	4	0.916	0.95	cos	2019/04/08	Darren Dahly	
18	17	1	my	filter	1	1.119	1.55	cos	2019/04/08	Darren Dahly	
19	18	1	my	filter	2	0.845	3.16	cos	2019/04/08	Darren Dahly	
20	19	1	my	filter	3	1.299	4.9	cos	2019/04/08	Brendan Palmer	
21	20	1	my	filter	4	1.149	5.5	cos	2019/04/08	Brendan Palmer	
22	21	1	ca	filter	1	0.716	5.5	cos	2019/04/08	Brendan Palmer	
23	22	1	ca	filter	2	0.881	7.94	cos	2019/04/08	Brendan Palmer	
24	23	1	ca	filter	3	0.586	8.71	cos	2019/04/08	Brendan Palmer	
25	24	1	ca	filter	4	0.561	8.71	cos	2019/04/08	Brendan Palmer	
26	25	2	ptp	filter	1	0	14.45	cos	2019/04/15	Brendan Palmer	
27	26	2	ptp	filter	2	1.006	2.14	cos	2019/04/15	Brendan Palmer	
28	27	2	ptp	filter	3	1.236	1.86	cos	2019/04/15	Brendan Palmer	
29	28	2	ptp	filter	4	1.206	1.2	cos	2019/04/15	Brendan Palmer	
30	29	2	my	filter	1	1.545	2.45	cos	2019/04/15	Brendan Palmer	

data | dictionary | values

Less stress, more success

	A	B	C	D	E	F	G	H	I	J	K
1	id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator	
2	1	0	ptp	no							
3	2	0	ptp	no							
4	3	0	ptp	no							
5	4	0	ptp	no							
6	5	0	my	no							
7	6	0	my	no							
8	7	0	my	no							
9	8	0	my	no							
10	9	0	ca	no							
11	10	0	ca	no							
12	11	0	ca	no							
13	12	0	ca	no							
14	13	1	ptp	fil							
15	14	1	ptp	fil							
16	15	1	ptp	fil							
17	16	1	ptp	fil							
18	17	1	my	fil							
19	18	1	my	fil							
20	19	1	my	fil							
21	20	1	my	fil							
22	21	1	ca	fil							
23	22	1	ca	fil							
24	23	1	ca	fil							
25	24	1	ca	fil							
26	25	2	ptp	fil							
27	26	2	ptp	fil							
28	27	2	ptp	fil							
29	28	2	ptp	fil							
30	29	2	my	fil							

data | dictionary | v

	A	B	C	D	E	
1	field_name	data_type	data_format	example	standard_units	description
2	id	numeric	integer	23	NA	Unique identifier applied to each observation
3	week_no	numeric	integer	1	NA	Week number, 1 = 7 days exposure, 2 = 14 days exposure
4	filter_name	character	NA	my	NA	3 filter types; 'ptp' = polytunnel plastic blocks all UV light
5	treatment	character	NA	filter	NA	Presence or absence of a filter at the time of sampling
6	replicate_no	numeric	integer	1	NA	The number of replicates in each treatement
7	flavonoids	numeric	double	0.3421	parts per million (ppm)	Leaf disc taken from the tip of the most mature leaf at th
8	biomass	numeric	double		gram (g)	Above ground biomass on the day of harvest
9	variety	character	NA	cos	NA	3 commerical varieties of red lettuce used; 'cos' = Cos Di
10	date	date	YYYY/MM/DD	2019/06/28	ISO 8601	Experiment date
11	investigator	character	Firstname Lastname	Aoife Coffey	NA	Primary researcher who performed the experiment
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						

data | dictionary | values

Less stress, more success

id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator
1	0	ptp	no filter	1	1.061	0.29	cos	2019/04/01	Darren Dahly
2	0	ptp	no						
3	0	ptp	no						
4	0	ptp	no						
5	0	my	no						
6	0	my	no						
7	0	my	no						
8	0	my	no						
9	0	ca	no						
10	0	ca	no						
11	0	ca	no						
12	0	ca	no						
13	1	ptp	fil						
14	1	ptp	fil						
15	1	ptp	fil						
16	1	ptp	fil						
17	1	my	fil						
18	1	my	fil						
19	1	my	fil						
20	1	my	fil						
21	1	ca	fil						
22	1	ca	fil						
23	1	ca	fil						
24	1	ca	fil						
25	2	ptp	fil						
26	2	ptp	fil						
27	2	ptp	fil						
28	2	ptp	fil						
29	2	my	fil						

data | dictionary | v

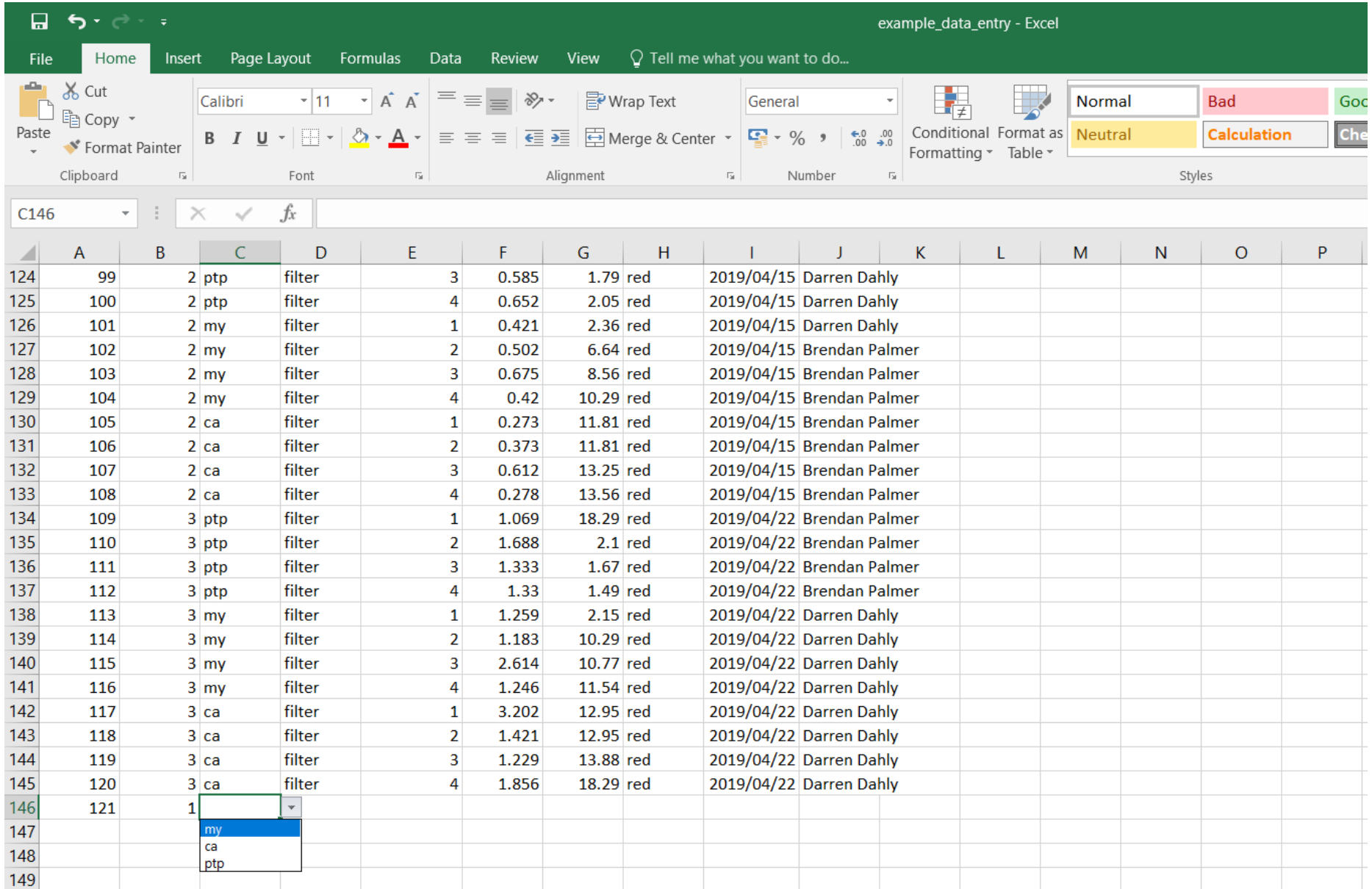
field_name	data_type	data_format	example	standard_units	description
id	numeric	integer	23	NA	Unique identifier applied to each observation
week_no	numeric	integer			
filter_name	character	NA	n		
treatment	character	NA	fi		
replicate_no	numeric	integer			
flavonoids	numeric	double			
biomass	numeric	double			
variety	character	NA	c		
date	date	YYYY/MM/DD			
investigator	character	Firstname Lastname	A		

data | **dictionary** | values

id	week_no	filter_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator
	0	my	filter	1			cos		Brendan Palmer
	1	ca	no_filter	2			sky		Darren Dahly
	2	ptp		3			red		
	3			4					

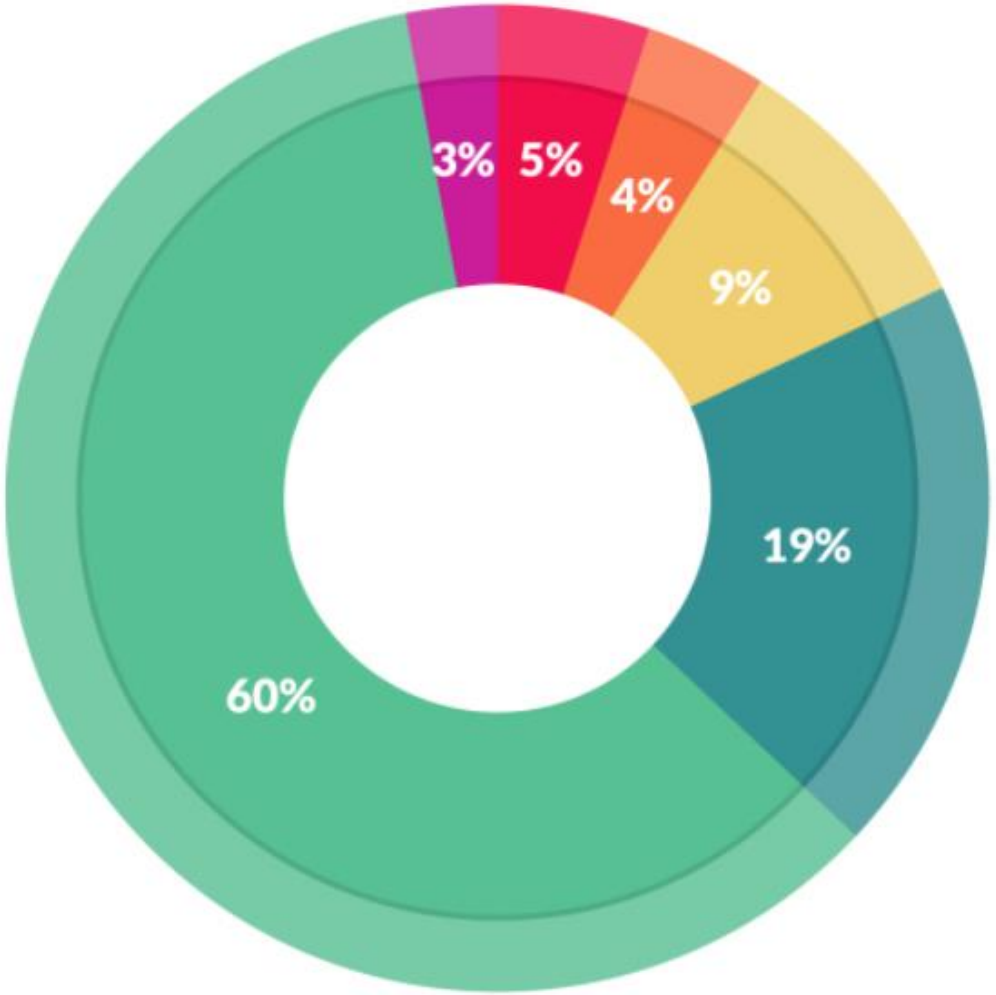
data | dictionary | **values**

Less stress, more success



	A	B	C	D	E	F	G	H	I	J
124	99	2	ptp	filter	3	0.585	1.79	red	2019/04/15	Darren Dahly
125	100	2	ptp	filter	4	0.652	2.05	red	2019/04/15	Darren Dahly
126	101	2	my	filter	1	0.421	2.36	red	2019/04/15	Darren Dahly
127	102	2	my	filter	2	0.502	6.64	red	2019/04/15	Brendan Palmer
128	103	2	my	filter	3	0.675	8.56	red	2019/04/15	Brendan Palmer
129	104	2	my	filter	4	0.42	10.29	red	2019/04/15	Brendan Palmer
130	105	2	ca	filter	1	0.273	11.81	red	2019/04/15	Brendan Palmer
131	106	2	ca	filter	2	0.373	11.81	red	2019/04/15	Brendan Palmer
132	107	2	ca	filter	3	0.612	13.25	red	2019/04/15	Brendan Palmer
133	108	2	ca	filter	4	0.278	13.56	red	2019/04/15	Brendan Palmer
134	109	3	ptp	filter	1	1.069	18.29	red	2019/04/22	Brendan Palmer
135	110	3	ptp	filter	2	1.688	2.1	red	2019/04/22	Brendan Palmer
136	111	3	ptp	filter	3	1.333	1.67	red	2019/04/22	Brendan Palmer
137	112	3	ptp	filter	4	1.33	1.49	red	2019/04/22	Brendan Palmer
138	113	3	my	filter	1	1.259	2.15	red	2019/04/22	Darren Dahly
139	114	3	my	filter	2	1.183	10.29	red	2019/04/22	Darren Dahly
140	115	3	my	filter	3	2.614	10.77	red	2019/04/22	Darren Dahly
141	116	3	my	filter	4	1.246	11.54	red	2019/04/22	Darren Dahly
142	117	3	ca	filter	1	3.202	12.95	red	2019/04/22	Darren Dahly
143	118	3	ca	filter	2	1.421	12.95	red	2019/04/22	Darren Dahly
144	119	3	ca	filter	3	1.229	13.88	red	2019/04/22	Darren Dahly
145	120	3	ca	filter	4	1.856	18.29	red	2019/04/22	Darren Dahly
146	121	1								
147										
148										
149										

Resources are being wasted by not doing this



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

CrowdFlower (2016) Data Science Report

Tidy data is clean data

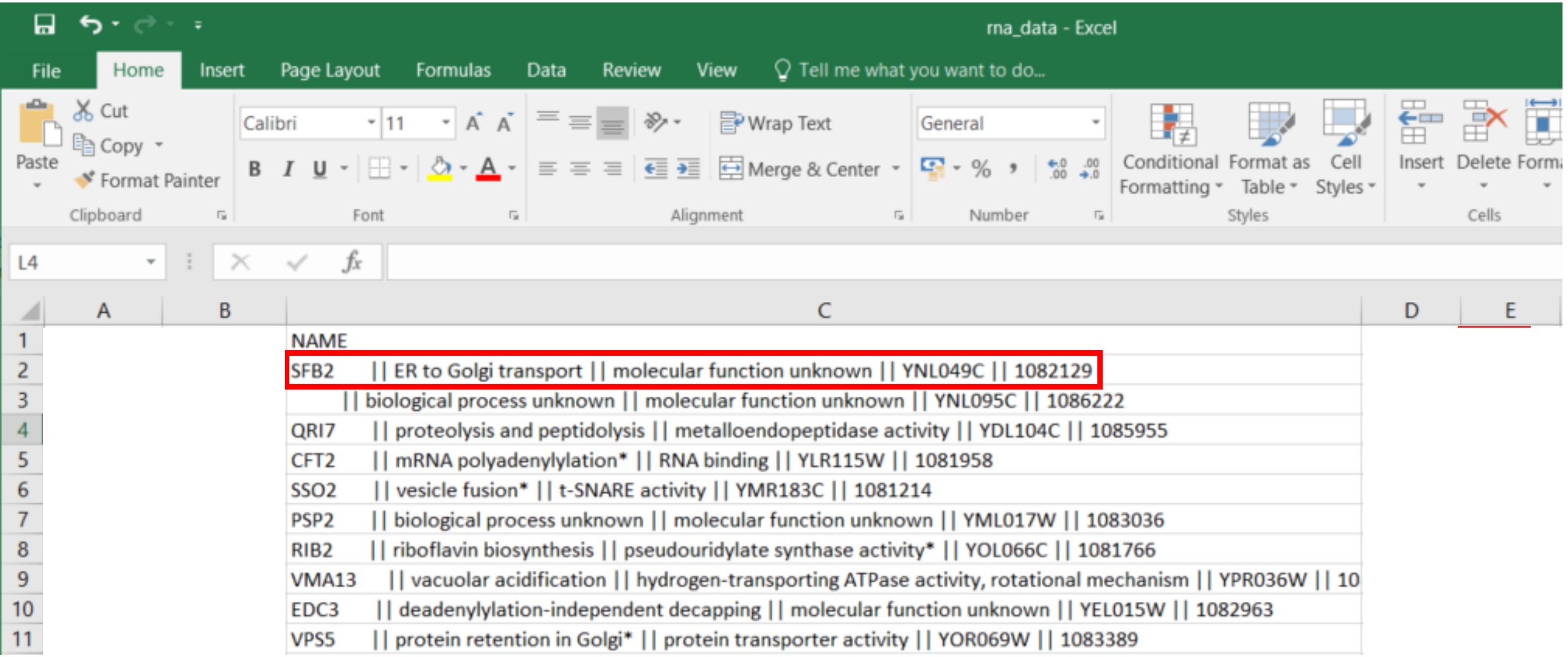
Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

- Each variable forms a column
- Each observation forms a row
- Each cell contains a value


rna_data - Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Tell me what you want to do...
Cut
Copy
Format Painter
Paste
Clipboard
Calibri
11
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Cells
L4
A
B
C
D
E
1 NAME
2 SFB2 || ER to Golgi transport || molecular function unknown || YNL049C || 1082129
3 || biological process unknown || molecular function unknown || YNL095C || 1086222
4 QRI7 || proteolysis and peptidolysis || metalloendopeptidase activity || YDL104C || 1085955
5 CFT2 || mRNA polyadenylylation* || RNA binding || YLR115W || 1081958
6 SSO2 || vesicle fusion* || t-SNARE activity || YMR183C || 1081214
7 PSP2 || biological process unknown || molecular function unknown || YML017W || 1083036
8 RIB2 || riboflavin biosynthesis || pseudouridylate synthase activity* || YOL066C || 1081766
9 VMA13 || vacuolar acidification || hydrogen-transporting ATPase activity, rotational mechanism || YPR036W || 10
10 EDC3 || deadenylylation-independent decapping || molecular function unknown || YEL015W || 1082963
11 VPS5 || protein retention in Golgi* || protein transporter activity || YOR069W || 1083389

The need for greater research reproducibility

The Reinhart-Rogoff error – or how not to Excel at economics

April 22, 2013 9.40pm BST

Data and computer code should be made publicly available at an early stage – or else ... esarastudillo

Email
Twitter 88
Facebook 453
LinkedIn
Print

Last week we learned a famous 2010 academic paper, relied on by political big-hitters to bolster arguments for austerity cuts, contained significant errors; and that those errors came down to misuse of an Excel spreadsheet.

Sadly, these are not the first mistakes of this size and nature when handling data. So what on Earth went wrong, and can we fix it?

Harvard's Carmen Reinhart and Kenneth Rogoff are two of the most respected and influential academic economists active today.