Data Organization in Spreadsheets

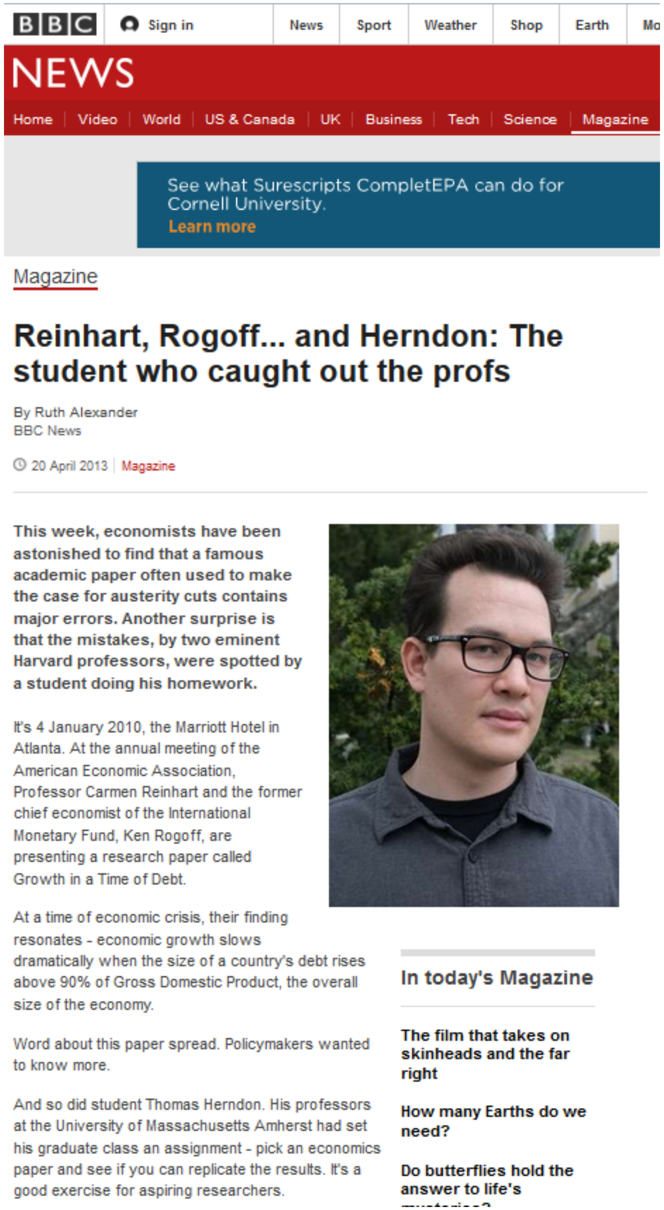
Learning Objectives

- Good data entry practices - formatting data tables in spreadsheets
- How to avoid common formatting mistakes
- Approaches for handling dates in spreadsheets
- Basic quality control and data manipulation in spreadsheets
- Exporting data from spreadsheets

Data Organization in Spreadsheets

What this lesson will not teach you

- How to do *statistics* in a spreadsheet
- How to do *plotting* in a spreadsheet
- How to *write code* in spreadsheet programs
- Why?
    - This requires a lot of manual work (lots of clicking!), is not very repeatable
    - It is also difficult to track or reproduce statistical or plotting analyses done in spreadsheet programs when you want to go back to your work or someone asks for details of your analysis.



Magazine

Reinhart, Rogoff... and Herndon: The student who caught out the profs

By Ruth Alexander
BBC News

20 April 2013 | Magazine

This week, economists have been astonished to find that a famous academic paper often used to make the case for austerity cuts contains major errors. Another surprise is that the mistakes, by two eminent Harvard professors, were spotted by a student doing his homework.

It's 4 January 2010, the Marriott Hotel in Atlanta. At the annual meeting of the American Economic Association, Professor Carmen Reinhart and the former chief economist of the International Monetary Fund, Ken Rogoff, are presenting a research paper called Growth in a Time of Debt.

At a time of economic crisis, their finding resonates - economic growth slows dramatically when the size of a country's debt rises above 90% of Gross Domestic Product, the overall size of the economy.

Word about this paper spread. Policymakers wanted to know more.

And so did student Thomas Herndon. His professors at the University of Massachusetts Amherst had set his graduate class an assignment - pick an economics paper and see if you can replicate the results. It's a good exercise for aspiring researchers.

In today's Magazine

The film that takes on skinheads and the far right

How many Earths do we need?

Do butterflies hold the answer to life's mysteries?

"… he'd spotted a basic error in the spreadsheet.

The Harvard professors had accidentally only included 15 of the 20 countries under analysis in their key calculation (of average GDP growth in countries with high public debt).

Australia, Austria, Belgium, Canada and Denmark were missing.

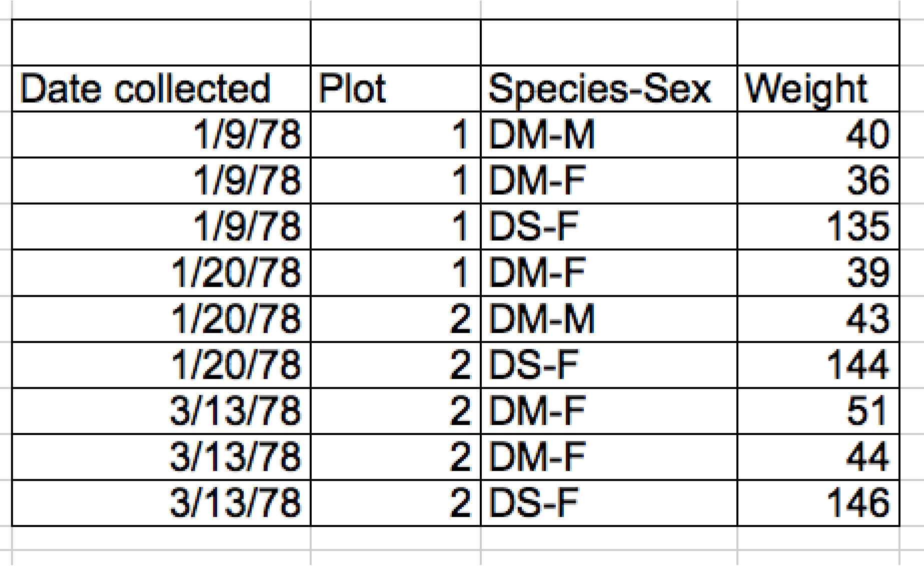
Oops."

http://www.bbc.com/news/magazine-22223190

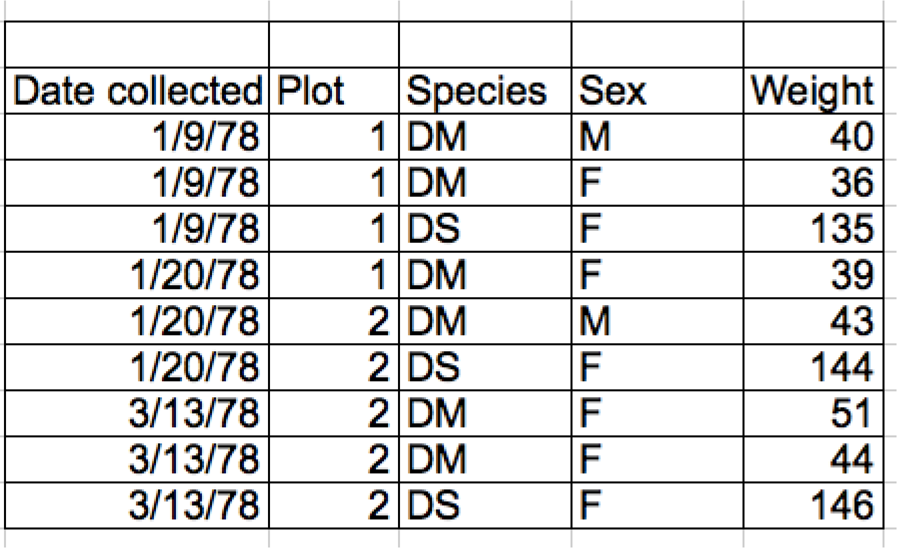
Structuring data in spreadsheets

The cardinal rules of using spreadsheet programs for data:

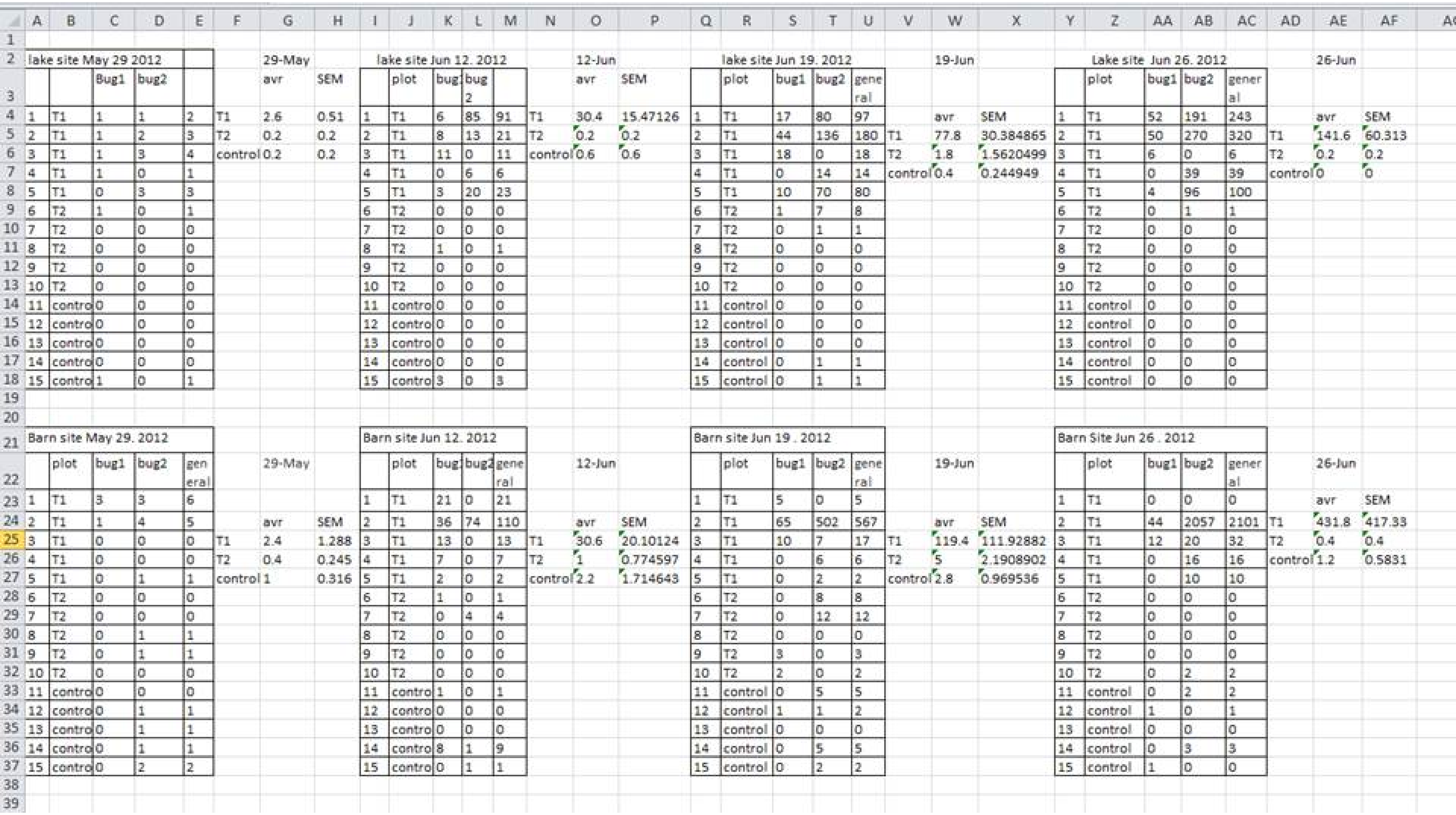
1. Put all your variables in columns - the thing you're measuring, like 'weight' or 'temperature'.
2. Put each observation in its own row.
3. Don't combine multiple pieces of information in one cell. Sometimes it just seems like one thing, but think if that's the only way you'll want to be able to use or sort that data.
4. Leave the raw data raw - don't mess with it!
5. Export the cleaned data to a text based format like CSV. This ensures that anyone can use the data, and is the format required by most data repositories.



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

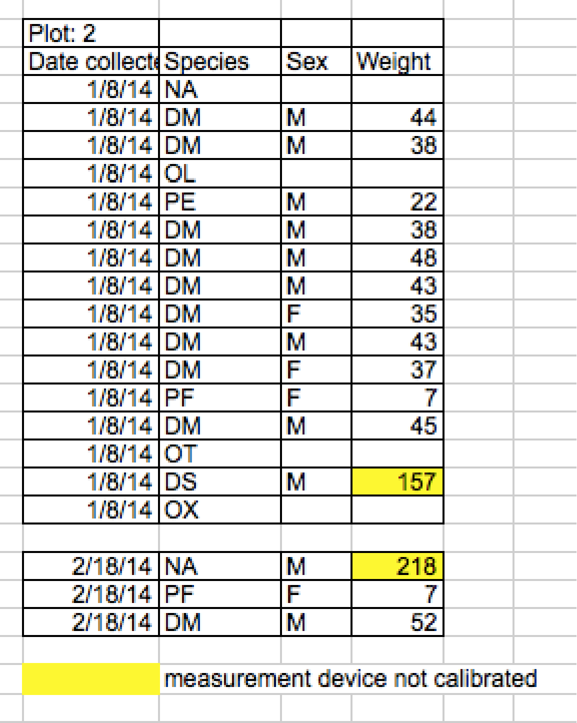


	A	B	C	D	E	F	G	H
1								
2	lake site May 29 2012						29-May	
3			Bug1	bug2			avr	SEM
4	1	T1	1	1	2	T1	2.6	0.51
5	2	T1	1	2	3	T2	0.2	0.2
6	3	T1	1	3	4	control	0.2	0.2
7	4	T1	1	0	1			
8	5	T1	0	3	3			
9	6	T2	1	0	1			
10	7	T2	0	0	0			
11	8	T2	0	0	0			
12	9	T2	0	0	0			
13	10	T2	0	0	0			
14	11	contro	0	0	0			
15	12	contro	0	0	0			
16	13	contro	0	0	0			
17	14	contro	0	0	0			
18	15	contro	1	0	1			
19								
20								
21	Barn site May 29. 2012							
22		plot	bug1	bug2	general		29-May	
23	1	T1	3	3	6			
24	2	T1	1	4	5		avr	SEM
25	3	T1	0	0	0	T1	2.4	1.288
26	4	T1	0	0	0	T2	0.4	0.245
27	5	T1	0	1	1	control	1	0.316
28	6	T2	0	0	0			
29	7	T2	0	0	0			
30	8	T2	0	1	1			
31	9	T2	0	1	1			
32	10	T2	0	0	0			
33	11	contro	0	0	0			
34	12	contro	0	1	1			
35	13	contro	0	1	1			
36	14	contro	0	1	1			
37	15	contro	0	2	2			
38								
39								

	I	J	K	L	M	N	O	P
1								
2	lake site Jun 12. 2012						12-Jun	
3		plot	bug1	bug2			avr	SEM
4	1	T1	6	85	91	T1	30.4	15.47126
5	2	T1	8	13	21	T2	0.2	0.2
6	3	T1	11	0	11	control	0.6	0.6
7	4	T1	0	6	6			
8	5	T1	3	20	23			
9	6	T2	0	0	0			
10	7	T2	0	0	0			
11	8	T2	1	0	1			
12	9	T2	0	0	0			
13	10	T2	0	0	0			
14	11	contro	0	0	0			
15	12	contro	0	0	0			
16	13	contro	0	0	0			
17	14	contro	0	0	0			
18	15	contro	3	0	3			
19								
20								
21	Barn site Jun 12. 2012							
22		plot	bug1	bug2	general		12-Jun	
23	1	T1	21	0	21			
24	2	T1	36	74	110		avr	SEM
25	3	T1	13	0	13	T1	30.6	20.10124
26	4	T1	7	0	7	T2	1	0.774597
27	5	T1	2	0	2	control	2.2	1.714643
28	6	T2	1	0	1			
29	7	T2	0	4	4			
30	8	T2	0	0	0			
31	9	T2	0	0	0			
32	10	T2	0	0	0			
33	11	contro	1	0	1			
34	12	contro	0	0	0			
35	13	contro	0	0	0			
36	14	contro	8	1	9			
37	15	contro	0	1	1			
38								
39								

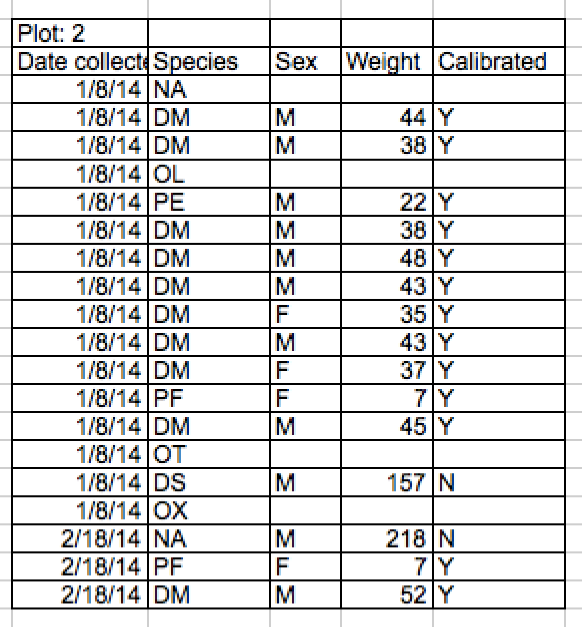
	Q	R	S	T	U	V	W	X
1								
2	lake site Jun 19. 2012						19-Jun	
3		plot	bug1	bug2	general			
4	1	T1	17	80	97		avr	SEM
5	2	T1	44	136	180	T1	77.8	30.384865
6	3	T1	18	0	18	T2	1.8	1.5620499
7	4	T1	0	14	14	control	0.4	0.244949
8	5	T1	10	70	80			
9	6	T2	1	7	8			
10	7	T2	0	1	1			
11	8	T2	0	0	0			
12	9	T2	0	0	0			
13	10	T2	0	0	0			
14	11	control	0	0	0			
15	12	control	0	0	0			
16	13	control	0	0	0			
17	14	control	0	1	1			
18	15	control	0	1	1			
19								
20								
21	Barn site Jun 19 . 2012							
22		plot	bug1	bug2	general		19-Jun	
23	1	T1	5	0	5			
24	2	T1	65	502	567		avr	SEM
25	3	T1	10	7	17	T1	119.4	111.92882
26	4	T1	0	6	6	T2	5	2.1908902
27	5	T1	0	2	2	control	2.8	0.969536
28	6	T2	0	8	8			
29	7	T2	0	12	12			
30	8	T2	0	0	0			
31	9	T2	3	0	3			
32	10	T2	2	0	2			
33	11	control	0	5	5			
34	12	control	1	1	2			
35	13	control	0	0	0			
36	14	control	0	5	5			
37	15	control	0	2	2			
38								
39								

	Y	Z	AA	AB	AC	AD	AE	AF
1								
2	Lake site Jun 26. 2012						26-Jun	
3		plot	bug1	bug2	general			
4	1	T1	52	191	243		avr	SEM
5	2	T1	50	270	320	T1	141.6	60.313
6	3	T1	6	0	6	T2	0.2	0.2
7	4	T1	0	39	39	control	0	0
8	5	T1	4	96	100			
9	6	T2	0	1	1			
10	7	T2	0	0	0			
11	8	T2	0	0	0			
12	9	T2	0	0	0			
13	10	T2	0	0	0			
14	11	control	0	0	0			
15	12	control	0	0	0			
16	13	control	0	0	0			
17	14	control	0	0	0			
18	15	control	0	0	0			
19								
20								
21	Barn Site Jun 26 . 2012							
22		plot	bug1	bug2	general		26-Jun	
23	1	T1	0	0	0		avr	SEM
24	2	T1	44	2057	2101	T1	431.8	417.33
25	3	T1	12	20	32	T2	0.4	0.4
26	4	T1	0	16	16	control	1.2	0.5831
27	5	T1	0	10	10			
28	6	T2	0	0	0			
29	7	T2	0	0	0			
30	8	T2	0	0	0			
31	9	T2	0	0	0			
32	10	T2	0	2	2			
33	11	control	0	2	2			
34	12	control	1	0	1			
35	13	control	0	0	0			
36	14	control	0	3	3			
37	15	control	1	0	0			
38								
39								



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

🟨 measurement device not calibrated



Plot: 2				
Date collecte	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

15 min

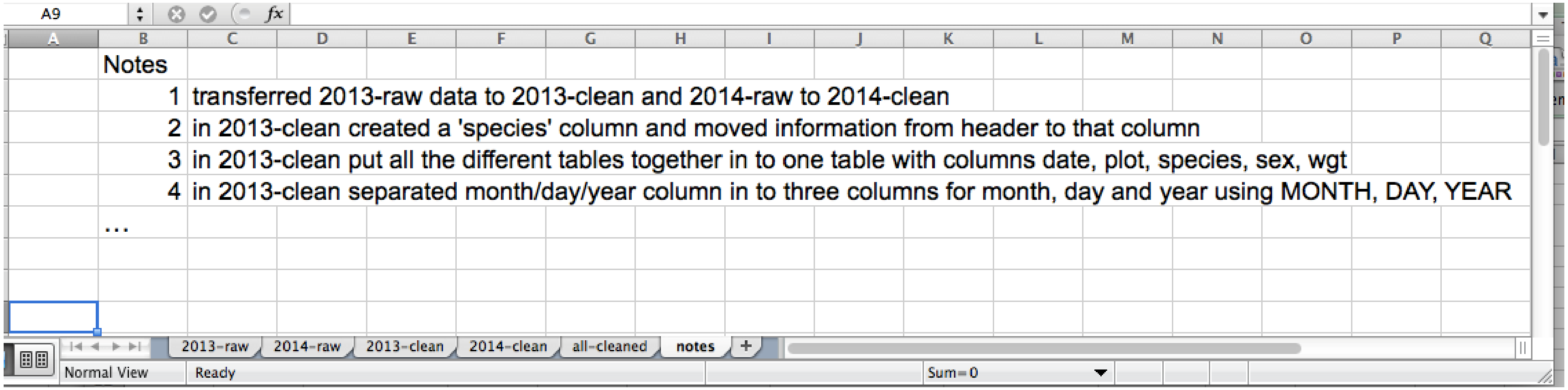
Exercise

- Download and open survey_data_spreadsheet_messy.xls

- Two field assistants conducted the surveys, one in 2013 and one in 2014, and they both kept track of the data in their own way.

- Clean the messy data so that a computer will be able to understand it. Clean up the 2013 and 2014 tabs, and put them all together in one spreadsheet.

*Do not forget of our first piece of advice, the **create a new file (or tab)** for the cleaned data, **never modify the original (raw) data**.*



Notes

1 transferred 2013-raw data to 2013-clean and 2014-raw to 2014-clean

2 in 2013-clean created a 'species' column and moved information from header to that column

3 in 2013-clean put all the different tables together in to one table with columns date, plot, species, sex, wgt

4 in 2013-clean separated month/day/year column in to three columns for month, day and year using MONTH, DAY, YEAR

…

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

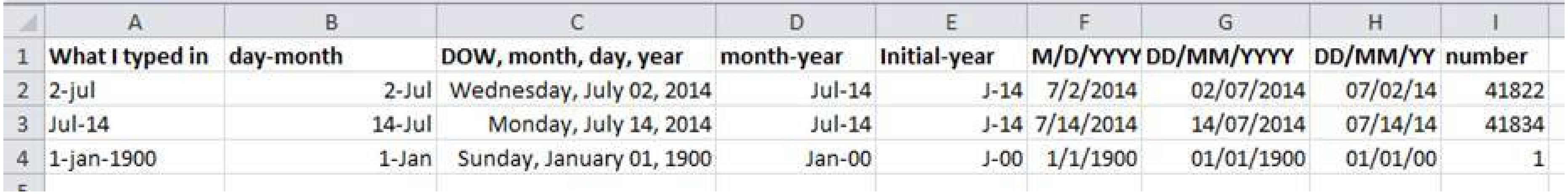
Field Names

good name	good alternative	avoid
Max_temp	MaxTemp	Maximum Temp (°C)
Precipitation	Precipitation_mm	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
first_observation	Observation_01	1st Obs.

Dates as data

Learning Objectives

- Understand how dates are handled and formatted in spreadsheets
- Manipulate dates stored in spreadsheets
- Understand the caveats of the default formatting of the dates



	A	B	C	D	E	F	G	H	I
1	**What I typed in**	**day-month**	**DOW, month, day, year**	**month-year**	**Initial-year**	**M/D/YYYY**	**DD/MM/YYYY**	**DD/MM/YY**	**number**
2	2-jul	2-Jul	Wednesday, July 02, 2014	Jul-14	J-14	7/2/2014	02/07/2014	07/02/14	41822
3	Jul-14	14-Jul	Monday, July 14, 2014	Jul-14	J-14	7/14/2014	14/07/2014	07/14/14	41834
4	1-jan-1900	1-Jan	Sunday, January 01, 1900	Jan-00	J-00	1/1/1900	01/01/1900	01/01/00	1

Exercise

- What happens to the dates in the “dates” tab of our workbook if we save this sheet in Excel (in csv format) and then open the file in a plain text editor (like TextEdit or Notepad)? What happens to the dates if we then open the csv file in Excel?

Exercise

- We’ve combined all of the tables from the messy data into a single table in a single tab. Download this semi-cleaned data file to your computer: survey_sorting_exercise

- Once downloaded, sort the Weight_grams column in your spreadsheet program from Largest to Smallest.

- What do you notice?