Data Organization in Spreadsheets

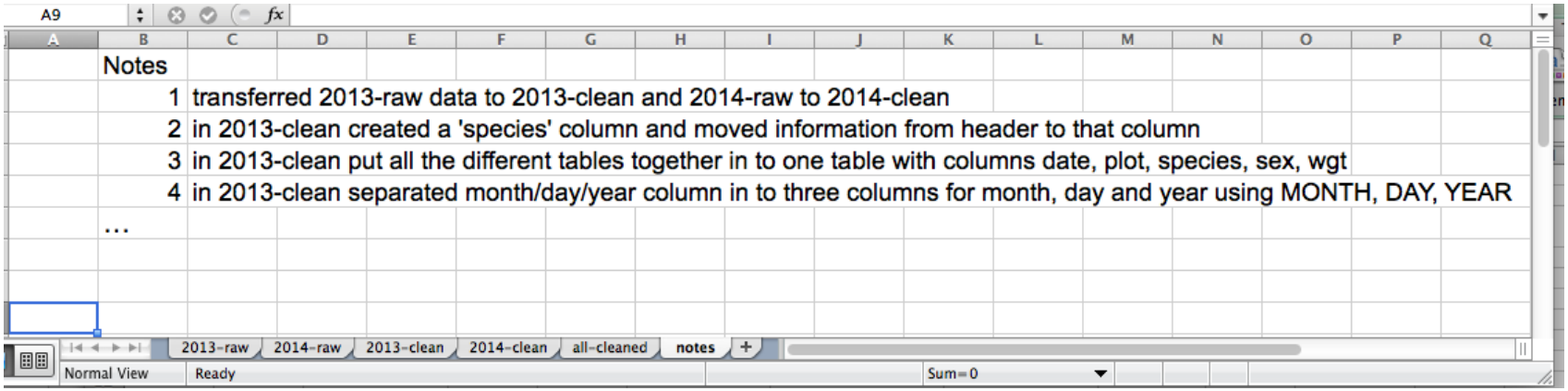
Morning Session I

Jacob F. Koehler, PhD. [1]

[1]Data Carpentry

Rutgers Newark, February 11, 2018

Keeping Track



	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
...	

Figure: Raw Copy and Notes

Structuring Data

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

Figure: What Could be Better?

Columns and Row Organization

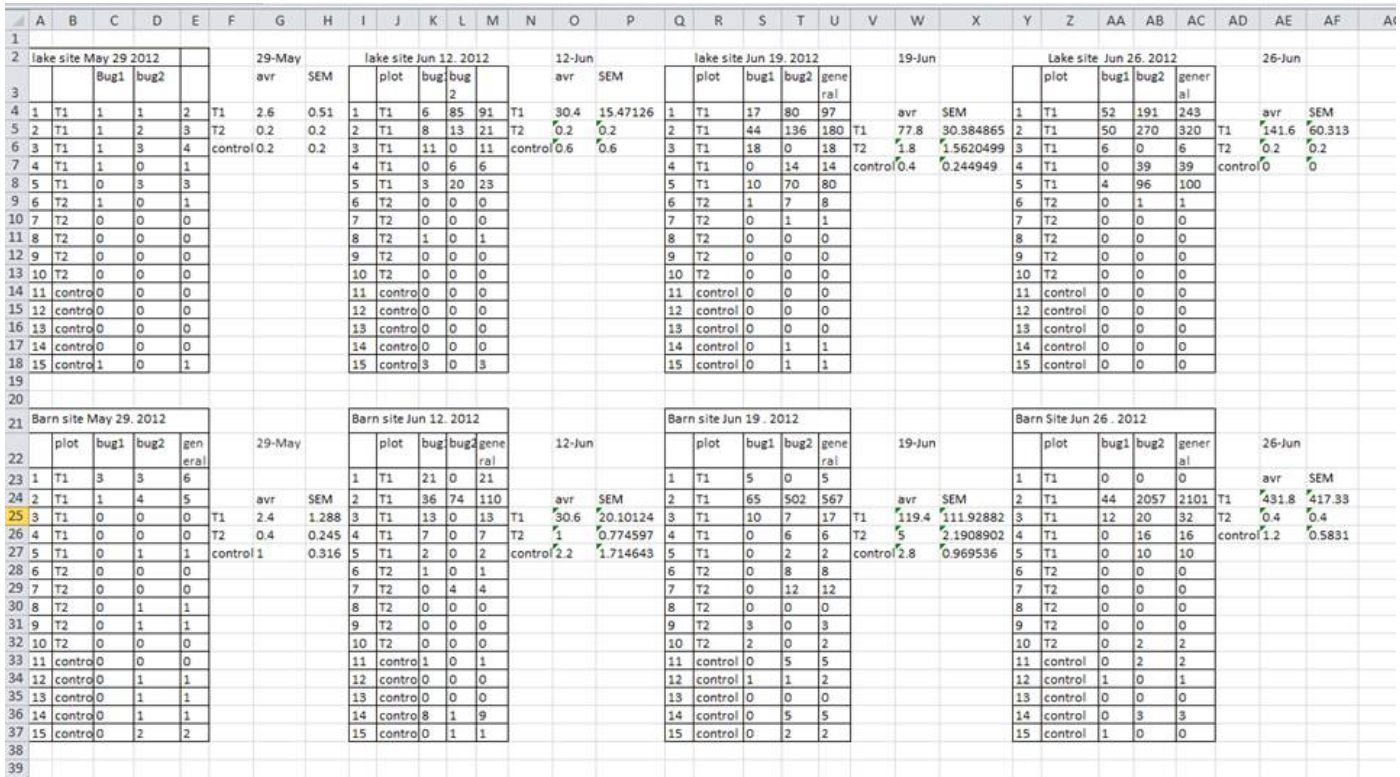
Date collected	Plot	Species	Sex	Weight
1/9/78	1	DM	M	40
1/9/78	1	DM	F	36
1/9/78	1	DS	F	135
1/20/78	1	DM	F	39
1/20/78	2	DM	M	43
1/20/78	2	DS	F	144
3/13/78	2	DM	F	51
3/13/78	2	DM	F	44
3/13/78	2	DS	F	146

Figure: Each Column is Variable, Each Row an Observation

Exercise I

1. Download the data by clicking link in lesson or in etherpad to get it from FigShare.
2. Open up the data in a spreadsheet program.
3. You can see that there are two tabs. 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. Now you're the person in charge of this project and you want to be able to start analyzing the data.
4. With the person next to you, identify what is wrong with this spreadsheet. Also discuss the steps you would need to take to clean up the 2013 and 2014 tabs, and to put them all together in one spreadsheet.

Multiple Tables



lake site May 29 2012

		Bug1	bug2	
1	T1	1	1	2
2	T1	1	2	3
3	T1	1	3	4
4	T1	1	0	1
5	T1	0	3	3
6	T2	1	0	1
7	T2	0	0	0
8	T2	0	0	0
9	T2	0	0	0
10	T2	0	0	0
11	control	0	0	0
12	control	0	0	0
13	control	0	0	0
14	control	0	0	0
15	control	1	0	1

29-May	avr	SEM
T1	2.6	0.51
T2	0.2	0.2
control	0.2	0.2

lake site Jun 12. 2012

	plot	bug1	bug2	
1	T1	6	85	91
2	T1	8	13	21
3	T1	11	0	11
4	T1	0	6	6
5	T1	3	20	23
6	T2	0	0	0
7	T2	0	0	0
8	T2	1	0	1
9	T2	0	0	0
10	T2	0	0	0
11	control	0	0	0
12	control	0	0	0
13	control	0	0	0
14	control	0	0	0
15	control	3	0	3

12-Jun	avr	SEM
T1	30.4	15.47126
T2	0.2	0.2
control	0.6	0.6

lake site Jun 19. 2012

	plot	bug1	bug2	general
1	T1	17	80	97
2	T1	44	136	180
3	T1	18	0	18
4	T1	0	14	14
5	T1	10	70	80
6	T2	1	7	8
7	T2	0	1	1
8	T2	0	0	0
9	T2	0	0	0
10	T2	0	0	0
11	control	0	0	0
12	control	0	0	0
13	control	0	0	0
14	control	0	1	1
15	control	0	1	1

19-Jun	avr	SEM
T1	77.8	30.384865
T2	1.8	1.5620499
control	0.4	0.244949

Lake site Jun 26. 2012

	plot	bug1	bug2	general
1	T1	52	191	243
2	T1	50	270	320
3	T1	6	0	6
4	T1	0	39	39
5	T1	4	96	100
6	T2	0	1	1
7	T2	0	0	0
8	T2	0	0	0
9	T2	0	0	0
10	T2	0	0	0
11	control	0	0	0
12	control	0	0	0
13	control	0	0	0
14	control	0	0	0
15	control	0	0	0

26-Jun	avr	SEM
T1	141.6	60.313
T2	0.2	0.2
control	0	0

Barn site May 29. 2012

	plot	bug1	bug2	general
1	T1	3	3	6
2	T1	1	4	5
3	T1	0	0	0
4	T1	0	0	0
5	T1	0	1	1
6	T2	0	0	0
7	T2	0	0	0
8	T2	0	1	1
9	T2	0	1	1
10	T2	0	0	0
11	control	0	0	0
12	control	0	1	1
13	control	0	1	1
14	control	0	1	1
15	control	0	2	2

29-May	avr	SEM
T1	2.4	1.288
T2	0.4	0.245
control	1	0.316

Barn site Jun 12. 2012

	plot	bug1	bug2	general
1	T1	21	0	21
2	T1	36	74	110
3	T1	13	0	13
4	T1	7	0	7
5	T1	2	0	2
6	T2	1	0	1
7	T2	0	4	4
8	T2	0	0	0
9	T2	0	0	0
10	T2	0	0	0
11	control	1	0	1
12	control	0	0	0
13	control	0	0	0
14	control	8	1	9
15	control	0	1	1

12-Jun	avr	SEM
T1	30.6	20.10124
T2	1	0.774597
control	2.2	1.714643

Barn site Jun 19 . 2012

	plot	bug1	bug2	general
1	T1	5	0	5
2	T1	65	502	567
3	T1	10	7	17
4	T1	0	6	6
5	T1	0	2	2
6	T2	0	8	8
7	T2	0	12	12
8	T2	0	0	0
9	T2	3	0	3
10	T2	2	0	2
11	control	0	5	5
12	control	1	1	2
13	control	0	0	0
14	control	0	5	5
15	control	0	2	2

19-Jun	avr	SEM
T1	119.4	111.92882
T2	5	2.1908902
control	2.8	0.969536

Barn Site Jun 26 . 2012

	plot	bug1	bug2	general
1	T1	0	0	0
2	T1	44	2057	2101
3	T1	12	20	32
4	T1	0	16	16
5	T1	0	10	10
6	T2	0	0	0
7	T2	0	0	0
8	T2	0	0	0
9	T2	0	0	0
10	T2	0	2	2
11	control	0	2	2
12	control	1	0	1
13	control	0	0	0
14	control	0	3	3
15	control	1	0	0

26-Jun	avr	SEM
T1	431.8	417.33
T2	0.4	0.4
control	1.2	0.5831

Figure: Recall Column Organization

Using Problematic Null Values

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

Figure: Null Suggestions

Using Formatting to Convey Information

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		

Figure: Highlighting Problematic Data

Using Formatting to Convey Information

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		

Figure: Highlighting Problematic Data

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

Figure: Solution: New Column

Problematic Field Names

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

Figure: Example Names

Dates as Data

Challenge: pulling month, day and year out of dates

- In the **dates** tab of your spreadsheet you have the data from 2014 plot 3. There's a **Date collected** column.
- Let's extract month, day and year from the dates to new columns. For this we can use the built in Excel functions **YEAR() MONTH() DAY()**

Hours, Minutes, and Seconds

Current time and date are best retrieved using the functions **NOW()**, which returns the current date and time, and **TODAY()**, which returns the current date. The results will be formatted according to your computer's settings.

1. Extract the year, month and day from the current date and time string returned by the **NOW()** function.
2. Calculate the current time using **NOW()-TODAY()**.
3. Extract the hour, minute and second from the current time using functions **HOUR(), MINUTE()** and **SECOND()**.
4. Press **F9** to force the spreadsheet to recalculate the **NOW()** function, and check that it has been updated.

Preferred date format

	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

Figure: Year, Month, Day in Separate Columns

Preferred date format

	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

Figure: Year, Month, Day in Separate Columns

Challenge: 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?

Quality Assurance

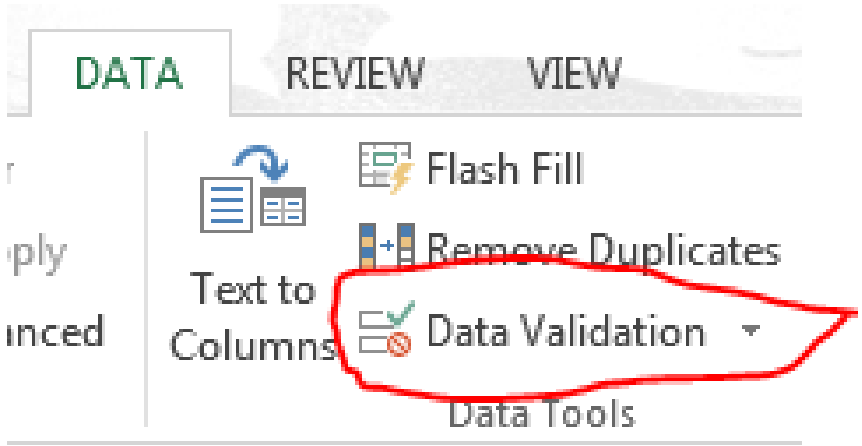


Figure: Checking Values are Valid During Data Entry

Quality Control

Sorting

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.

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

What do you notice?

Conditional Formatting

Challenge:

1. In the main Excel menu bar, click **Format** > **Conditional Formating**... Click the + to add a formatting rule.
2. Apply a **2-Color Scale** formatting rule with the lowest values set to orange and the highest values set to yellow.
3. Now we can scan through and different colors will stand out. Do you notice any strange values?

Exporting Data

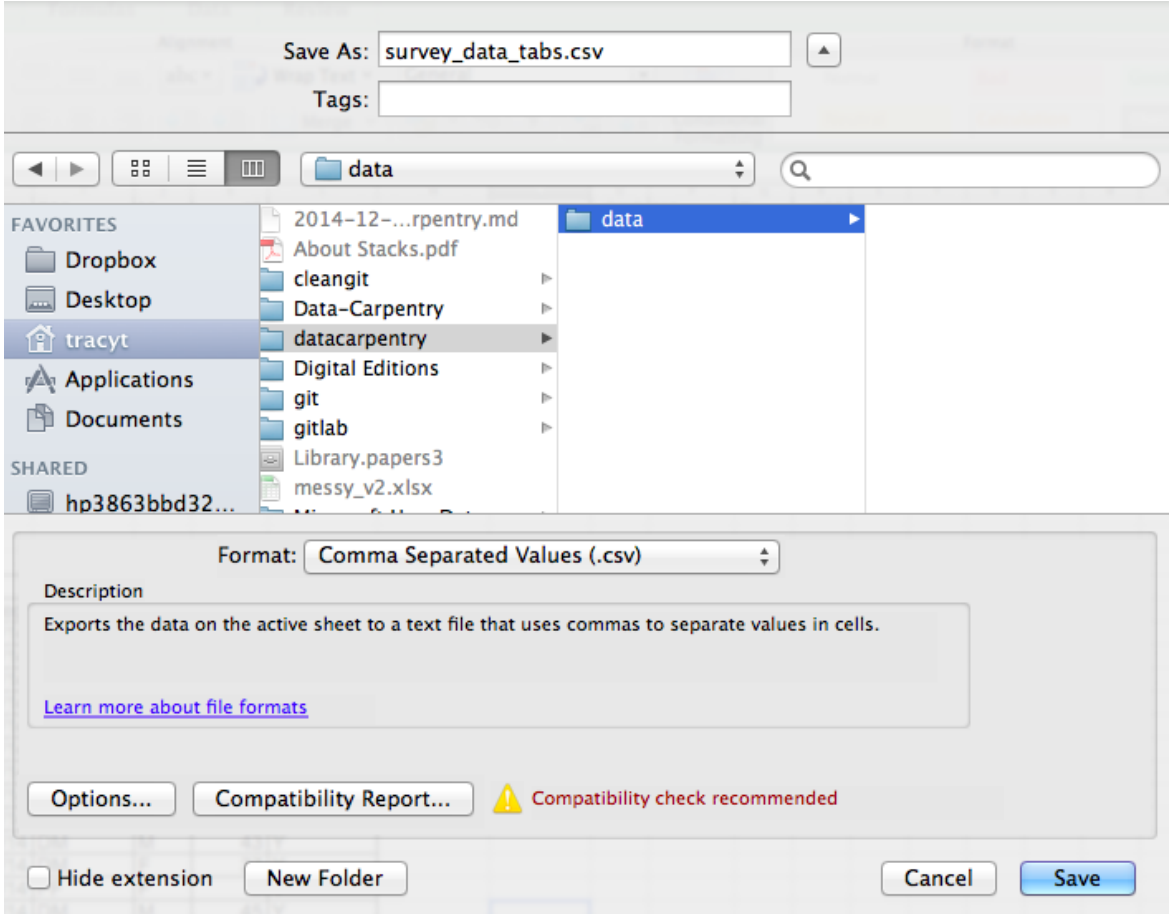


Figure: Saving as .csv