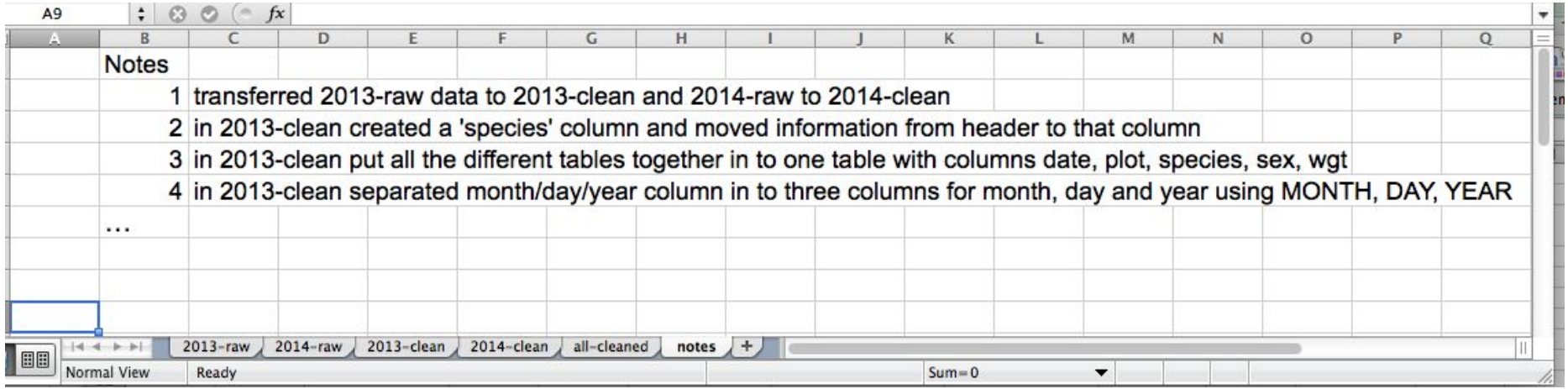
How many people have accidentally done something that made them frustrated or sad?

What kind of operations do you do in spreadsheets?

Which ones do you think spreadsheets are good for?



Separate Sheets tabs for information

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

Combined variables in a single column

What is one problem you see with the recorded data in the example?

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

separate variables in unique column

Columns = variables, rows = observations, cells = data (values)

Exercise 1

Exercise

We're going to take a messy version of the survey data and describe how we would clean it up.

1. Download the data by clicking here 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.

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

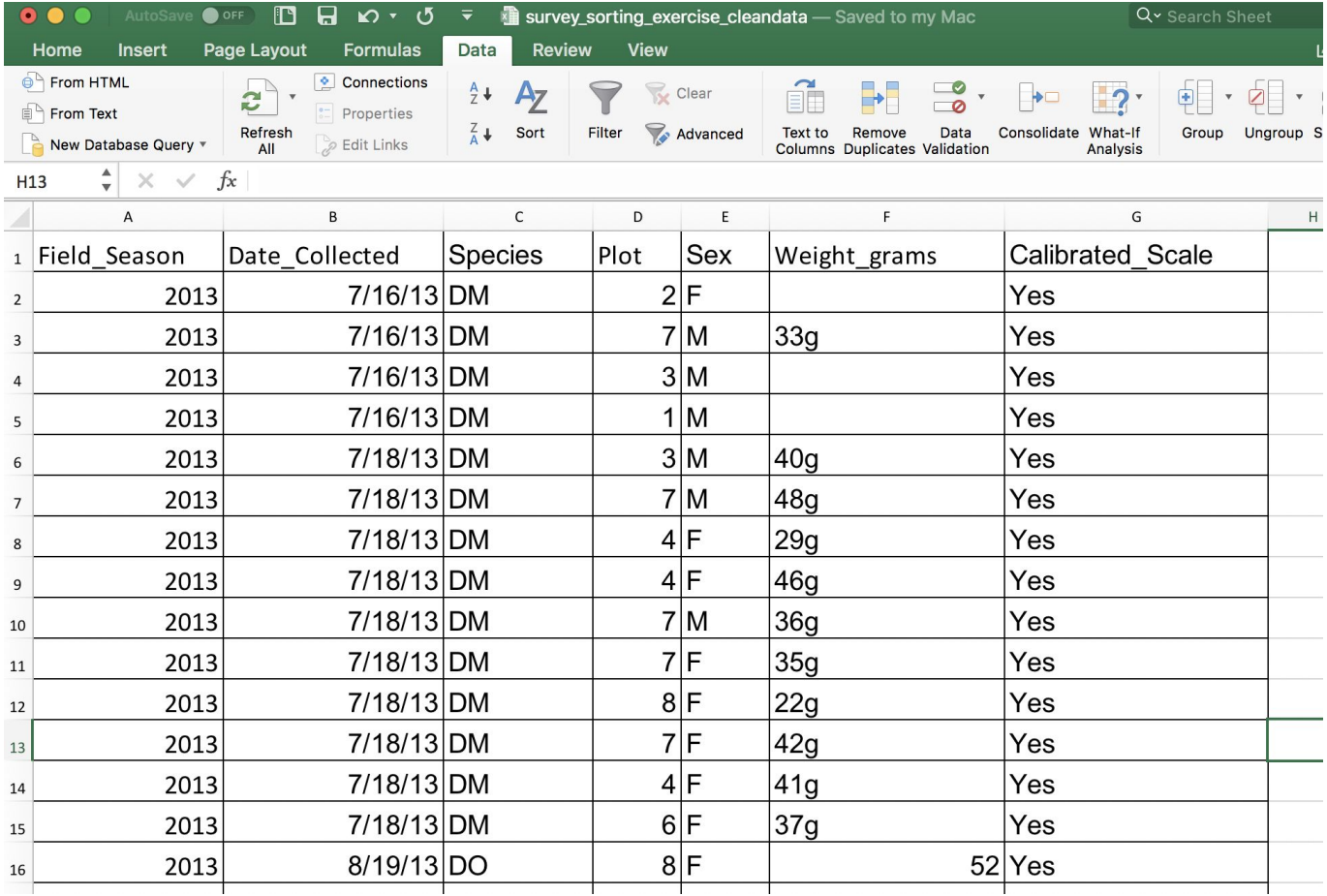
After you go through this exercise, we'll discuss as a group what was wrong with this data and how you would fix it.

Solution

Exercise 1 - issues list - solutions

- Using multiple tables
- Using multiple tabs
- Not filling in zeros
- Using problematic null values
- Using formatting to convey information
- Using formatting to make the data sheet look pretty
- Placing comments or units in cells
- Entering more than one piece of information in a cell
- Using problematic field names
- Using special characters in data
- Inclusion of metadata in data table
- Date formatting

Cleaned worksheet example:



	A	B	C	D	E	F	G
1	Field_Season	Date_Collected	Species	Plot	Sex	Weight_grams	Calibrated_Scale
2	2013	7/16/13	DM	2	F		Yes
3	2013	7/16/13	DM	7	M	33g	Yes
4	2013	7/16/13	DM	3	M		Yes
5	2013	7/16/13	DM	1	M		Yes
6	2013	7/18/13	DM	3	M	40g	Yes
7	2013	7/18/13	DM	7	M	48g	Yes
8	2013	7/18/13	DM	4	F	29g	Yes
9	2013	7/18/13	DM	4	F	46g	Yes
10	2013	7/18/13	DM	7	M	36g	Yes
11	2013	7/18/13	DM	7	F	35g	Yes
12	2013	7/18/13	DM	8	F	22g	Yes
13	2013	7/18/13	DM	7	F	42g	Yes
14	2013	7/18/13	DM	4	F	41g	Yes
15	2013	7/18/13	DM	6	F	37g	Yes
16	2013	8/19/13	DO	8	F	52	Yes

Exercise 2

Exercise

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()`

(Make sure the new column is formatted as a number and not as a date.)

You can see that even though you wanted the year to be 2014, your spreadsheet program automatically interpreted it as 2015, the year you entered the data.

Solution

Exercise 2 solution

Solution

	A	B	C	D	E	F	G
1	Plot: 3						
2	Date collected	Species	Sex	Weight	Month	Day	Year
3	1/8	PF	M	7	=MONTH(A3)	=DAY(A3)	=YEAR(A3)
4	2/18	OT	M	24	2	18	2015
5	2/19	OT	F	23	2	19	2015
6	3/11	NA	M	232	3	11	2015
7	3/11	OT	F	22	3	11	2015
8	3/11	OT	M	26	3	11	2015
9	3/11	PF	M	8	3	11	2015
10	4/8	NA	F		4	8	2015
11	5/6				5	6	2015
12	5/18	NA	F	182	5	18	2015
13	6/9	OT	F	29	6	9	2015
14	7/8	NA	F	115	7	8	2015
15	7/8	NA	M	190	7	8	2015

Exercise 3

Exercise

Challenge: pulling hour, minute and second out of the current time

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.

Exercise 3 solution

	A	B	C	D	E	F	G	H
1		**month**	**year**	**day**	**current_time**	**hour**	**min**	**sec**
2	8/10/18 14:09	8	2018	10	0.590197338	14	9	53
3								
4								
5								

	A	B	C	D	E	F	G	H
1		**month**	**year**	**day**	**current_time**	**hour**	**min**	**sec**
2	=NOW()	=MONTH(NOW())	=YEAR(NOW())	=DAY(NOW())	=NOW()-TODAY()	=HOUR(NOW()-TODAY())	=MINUTE(NOW()-TODAY())	=SECOND(NOW()-TODAY())
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								

Date formats in spreadsheets example:

	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
5									

Exercise 4: saving dates as CSV

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?

Solution

Exercise 4: solution

Solution

- Click to the “dates” tab of the workbook and double-click on any of the values in the `Date collected` column. Notice that the dates display with the year 2015.
- Select `File -> Save As` in Excel and in the drop down menu for file format select `CSV UTF-8 (Comma delimited) (.csv)`. Click `Save`.
- You will see a pop-up that says “This workbook cannot be saved in the selected file format because it contains multiple sheets.” Choose `Save Active Sheet`.
- Navigate to the file in your finder application. Right click and select `Open With`. Choose a plain text editor application and view the file. Notice that the dates display as month/day without any year information.
- Now right click on the file again and open with Excel. Notice that the dates display with the current year, not 2015.
  As you can see, exporting data from Excel and then importing it back into Excel fundamentally changed the data!

	A	B	C
1	DATE	Number	How it was interpreted
2	Jul-10	40360	1-Jul-10
3	Jul-14	41821	1-Jul-14
4	Jul-15	42186	1-Jul-15
5	Jul-19	43647	1-Jul-19

Spreadsheet date interpretation

	A	B	C	D
1	**Date**	**Year**	**DOY**	**Convert back to date**
2	July 2, 2014	=YEAR(A2)	=A2-DATE(YEAR(A2),1,0)	=DATE(B2,1,C2)
3	2-Jul	2014	183	7/2/2014
4				

Storing dates as Year, Day-of-Year

Or as a string: YYYYMMDDhhmmss

March 24, 2018 17:25:35 = 20180324172535

Exercise 5 - Sorting

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?

Solution

Exercise 5 - Sorting Solution 1

	A	B	C	D	E	F	G
1	Field_Season	Date_Collected	Species	Plot	Sex	Weight_grams	Calibrated_Scale
2	2013	7/18/13	DM	7	M	48g	Yes
3	2013	7/18/13	DM	4	F	46g	Yes
4	2013	7/18/13	DM	7	F	42g	Yes
5	2013	7/18/13	DM	4	F	41g	Yes
6	2013	7/18/13	DM	3	M	40g	Yes
7	2013	7/18/13	DM	6	F	37g	Yes
8	2013	7/18/13	DM	7	M	36g	Yes
9	2013	7/18/13	DM	7	F	35g	Yes
10	2013	7/16/13	DM	7	M	33g	Yes
11	2013	7/18/13	DM	4	F	29g	Yes
12	2013	7/18/13	DM	8	F	22g	Yes
13	2014	2/18/14	NA	1	M	218	No
14	2014	1/8/14	DS	7	M	157	No
15	2014	3/13/14	DS	3	F	146	Yes
16	2014	1/20/14	DS	4	F	144	Yes
17	2014	1/9/14	DS	7	F	135	Yes
18	2013	11/13/13	DS	17	M	132	No
19	2014	1/8/78	DS	3	F	128	Yes
20	2013	11/13/13	DS	11	F	126	Yes
21	2013	11/13/13	DS	11	F	122	Yes
22	2013	11/12/13	DS	1	F	121	Yes
23	2013	11/12/13	DS	9	F	120	Yes

	A	B	C	D	E	F	G
1	Field_Season	Date_Collected	Species	Plot	Sex	Weight_grams	Calibrated_Scale
58	2014	1/8/78	DM	1	F	37	Yes
59	2014	1/8/78	DM	7	M	37	Yes
60	2014	1/9/14	DM	18	F	36	Yes
61	2014	1/8/14	DM	5	F	35	Yes
62	2014	1/8/78	OL	18	M	35	Yes
63	2013	10/17/13	DO	3	F	33	Yes
64	2013	10/17/13	DO	17	F	31	Yes
65	2014	1/8/14	PE	3	M	22	Yes
66	2014	1/8/14	PF	17	F	7	Yes
67	2014	2/18/14	PF	4	F	7	Yes
68	2013	11/13/13	DS	17	F		Yes
69	2013	7/16/13	DM	2	F		Yes
70	2013	7/16/13	DM	3	M		Yes
71	2013	7/16/13	DM	1	M		Yes
72	2014	1/8/14	NA	3			Yes
73	2014	1/8/14	OL	6			Yes
74	2014	1/8/14	OT	18			Yes
75	2014	1/8/14	OX	1			Yes
76	2014	1/8/78	DM	1	M		Yes

Exercise 6 - Conditional formatting

Exercise

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?

Solution

Exercise 6 - Conditional formatting Solution

Cells that contain non-numerical values are not colored. This includes both the cells where the letter “g” was included and the empty cells.

	A	B	C	D	E	F	G
1	Field_Season	Date_Collected	Species	Plot	Sex	Weight_grams	Calibrated_Scale
2	2013	7/18/13	DM	7	M	48g	Yes
3	2013	7/18/13	DM	4	F	46g	Yes
4	2013	7/18/13	DM	7	F	42g	Yes
5	2013	7/18/13	DM	4	F	41g	Yes
6	2013	7/18/13	DM	3	M	40g	Yes
7	2013	7/18/13	DM	6	F	37g	Yes
8	2013	7/18/13	DM	7	M	36g	Yes
9	2013	7/18/13	DM	7	F	35g	Yes
10	2013	7/16/13	DM	7	M	33g	Yes
11	2013	7/18/13	DM	4	F	29g	Yes
12	2013	7/18/13	DM	8	F	22g	Yes
13	2014	2/18/14	NA	1	M	218	No
14	2014	1/8/14	DS	7	M	157	No
15	2014	3/13/14	DS	3	F	146	Yes
16	2014	1/20/14	DS	4	F	144	Yes
17	2014	1/9/14	DS	7	F	135	Yes
18	2013	11/13/13	DS	17	M	132	No
19	2014	1/8/78	DS	3	F	128	Yes
20	2013	11/13/13	DS	11	F	126	Yes
21	2013	11/13/13	DS	11	F	122	Yes
22	2013	11/12/13	DS	1	F	121	Yes
23	2013	11/12/13	DS	9	F	120	Yes

As such, when exporting to CSV using Excel, your data in text format will look like this:

> data1,data2\r\n1,2\r\n4,5\r\n

When opening your CSV file in Excel again, it will parse it as follows:

	A	B
1	data1	data2
2	1	2
3	4	5

However, if you open your CSV file on a different system that does not parse the “\r” it will interpret your CSV file differently:

Your data in text format then look like this:

> data1
> data2\r
> 1
> 2\r
> …

This will then in turn parse as:

	A	B
1	data1	data2\r
2	1	2\r
3	4	5\r