DATA CARPENTRY WORKSHOP

DATA ORGANIZATION IN SPREADSHEETS

Nathalia Graf-Grachet
10/18/2018

SOURCES:

- Data Carpentry lesson material: https://datacarpentry.org/spreadsheet-ecology-lesson/
- Data for this lesson comes from Ernest, M., Brown, J., Valone, T., and White, E.P. (2017). Portal Project Teaching Database. Version 6. Figshare. DOI: 10.6084/m9.figshare.1314459.v6
- #1 https://ndownloader.figshare.com/files/2252083
- #2 https://github.com/datacarpentry/spreadsheet-ecology-lesson/blob/gh-pages/data/survey_sorting_exercise.xlsx?raw=true

INTRODUCTION

- Spreadsheet programs like Excel are great for data entry, designing data tables, and handling basic functions.
- People format data differently, use different software to format/analyze data, etc.
- Well-formatted tables is the foundation of a pain-free data analysis.

GOAL

- To learn best recommended data practices.
- You'll learn to:
    - format data tables in spreadsheets
    - avoid common formatting mistakes
    - control data entries/spot bad entries
    - export data in .csv or .tsv for any analysis software and data repositories.

TOPICS:

I. Formatting data tables in spreadsheets

II. Formatting problems

III. Dates as data

IV. Quality control

V. Exporting data

I. Formatting data tables in spreadsheets

1. Variables in columns, observations in rows, data/value in cells.
2. Don't combine multiple pieces of information in one cell.
3. Leave the original data raw.
4. Create a new file with your cleaned or analyzed data.
5. Keep track of the steps in clean up or analysis in a plain test file 'README.txt'.
6. Avoid using the spreadsheet for notes because computers don't view our notes in the same way.

DIFFERENT PEOPLE ENTERED DATA INTO THE SPREADSHEET…

Example #1

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

DIFFERENT PEOPLE ENTERED DATA INTO THE SPREADSHEET...

Recommendation

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

survey_data_2013

Search Sheet

Home | Insert | Page Layout | Formulas | Data | Review | View | Share

Paste | Calibri (Body) | 12 | Number | Conditional Formatting | Format as Table | Cell Styles | Cells | Editing

F2

	A	B	C	D	E	F	G	H	I	J	K
1	Date Collected	Plot	Species	Sex	Weight						
2	1/9/13	1	DM	M	40						
3	1/9/13	1	DM	F	36						
4	1/9/13	1	DS	F	135						
5	1/20/13	1	DM	F	39						
6	1/20/13	2	DM	M	43						
7	1/20/13	2	DS	F	144						
8	3/13/13	2	DM	F	51						
9	3/13/13	2	DM	F	44						
10	3/13/13	2	DS	F	146						
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											

README_survey_data_2013 — Edited

```
Notes on survey_data_2013.xlsx

19/04/2013

1. Received survey data from field season 2013 from John.


22/04/2013

2. In 2013-raw: raw data as received.

3. In 2013-clean: separated species and sex into different
columns.
```

2013-raw | 2013-clean | +

Ready

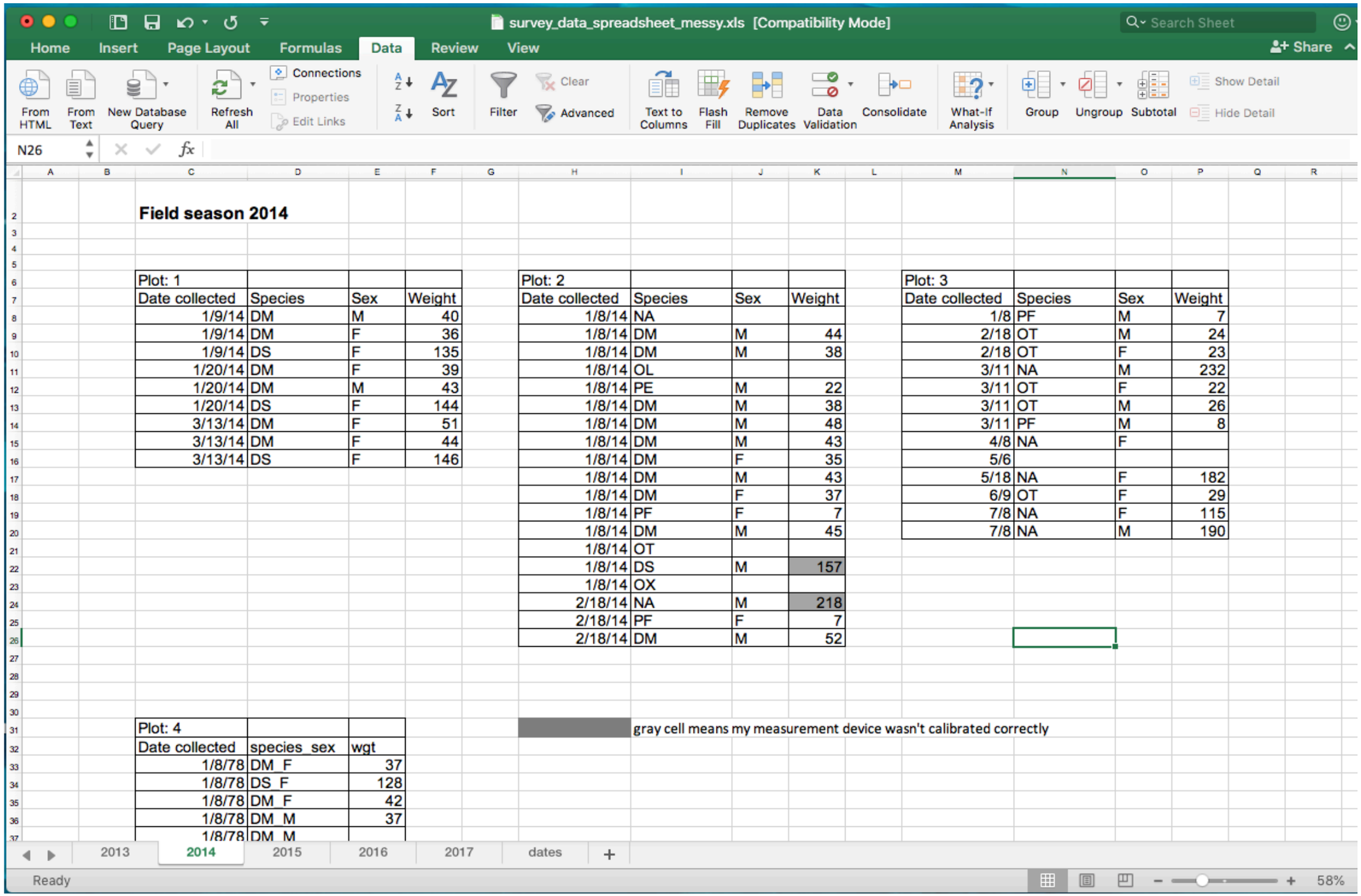
120%

II. Formatting problems

- https://ndownloader.figshare.com/files/2252083
- Download and open up in Excel
- Observations of a small mammal community in southern Arizona to study the effects of rodents and ants on the plant community that has been running for almost 40 years.
- 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.

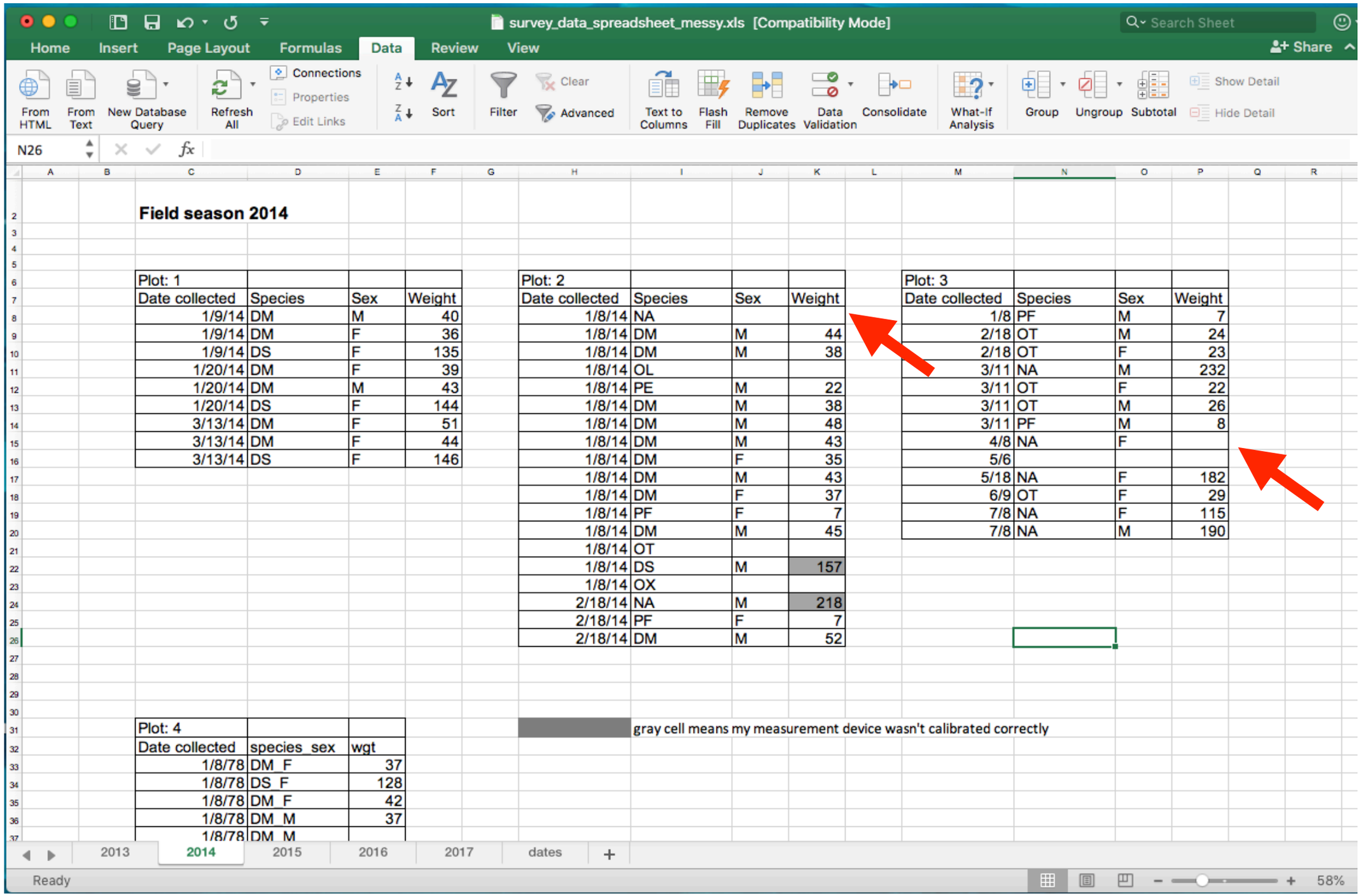
II. Formatting problems

➤ Creating multiple tables, and multiple tabs



II. Formatting problems

- Not filling up the zeros, and null values



II. Formatting problems

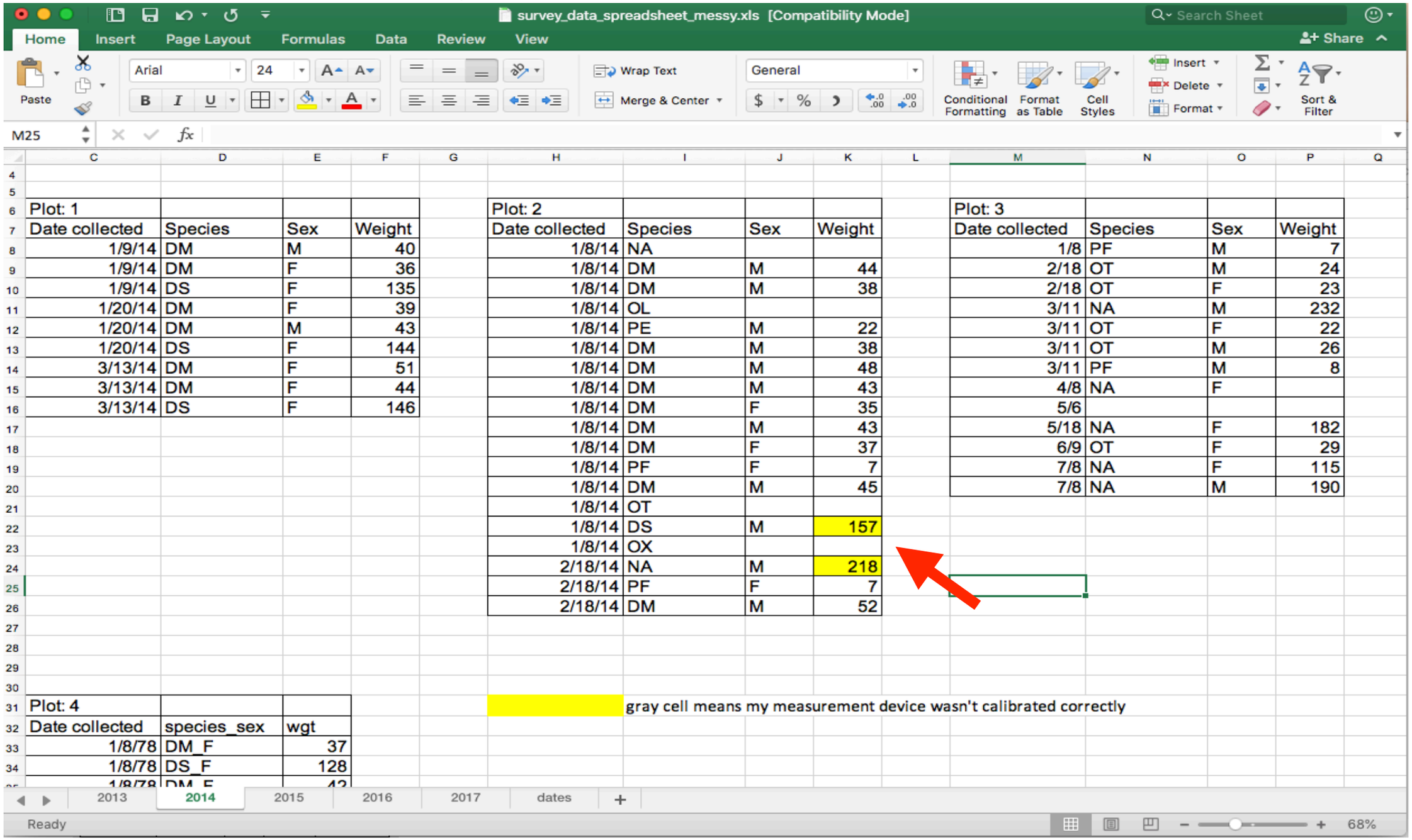
➤ Using problematic null values

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

II. Formatting problems

- Using formatting to convey information or make the spreadsheet pretty (merged cells)



Plot: 1			
Date collected	Species	Sex	Weight
1/9/14	DM	M	40
1/9/14	DM	F	36
1/9/14	DS	F	135
1/20/14	DM	F	39
1/20/14	DM	M	43
1/20/14	DS	F	144
3/13/14	DM	F	51
3/13/14	DM	F	44
3/13/14	DS	F	146

Plot: 2			
Date collected	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

Plot: 3			
Date collected	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

Plot: 4		
Date collected	species_sex	wgt
1/8/78	DM_F	37
1/8/78	DS_F	128

gray cell means my measurement device wasn't calibrated correctly

II. Formatting problems

- Using problematic field names:
    - Choose descriptive field names, but do not include spaces, numbers, or special characters.
    - Spaces can be misinterpreted by parsers that use whitespace as delimiters
    - Some programs don’t like field names that are text strings that start with numbers.
    - Underscores are a good_alternative_to_spaces.
    - Consider writing names in camel case ‘FileNameOne’
    - Abbreviations may not be so obvious after 6 months…

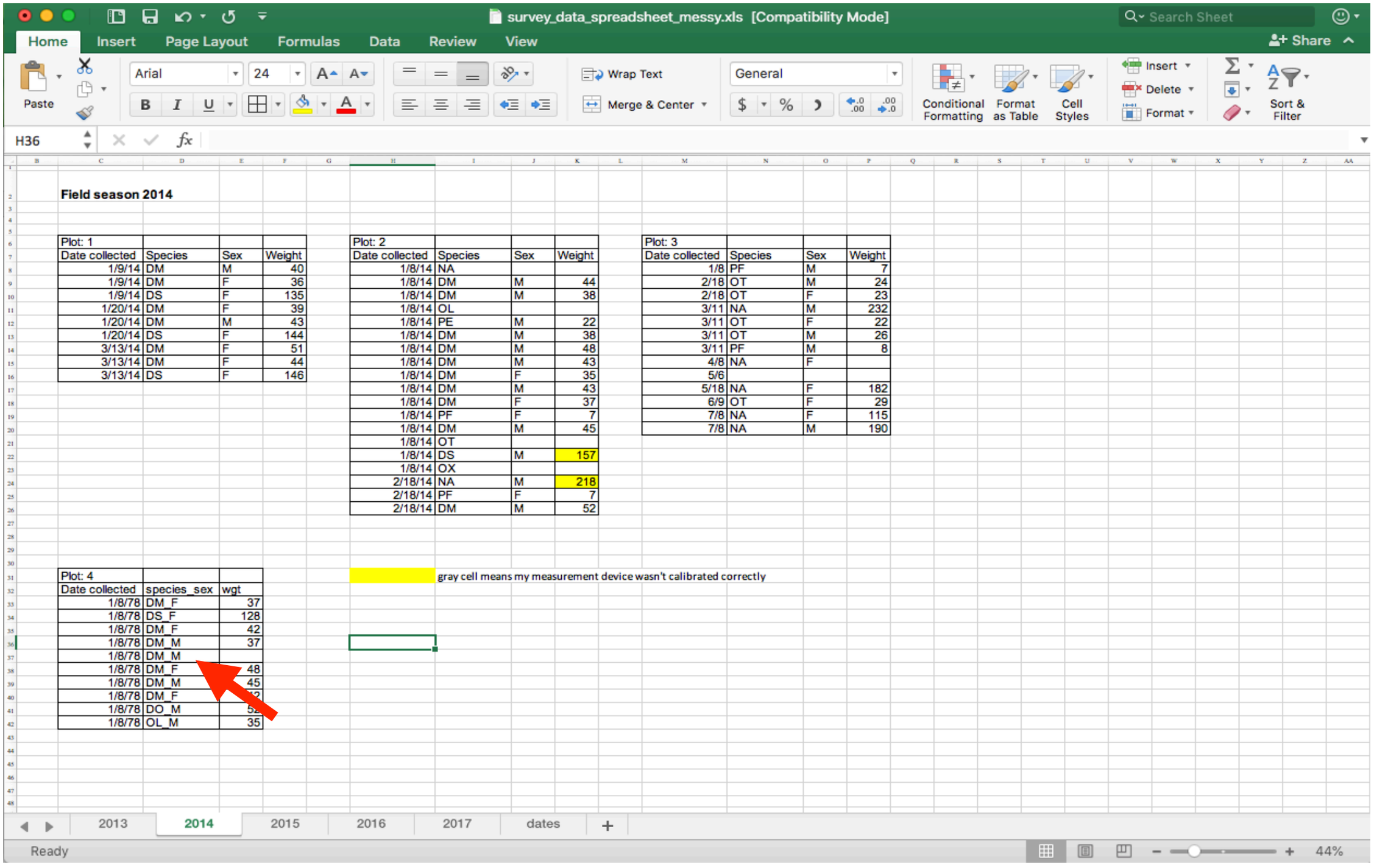
II. Formatting problems

➤ Recommendation for problematic field names

Good Name	Good Alternative	Avoid, please
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

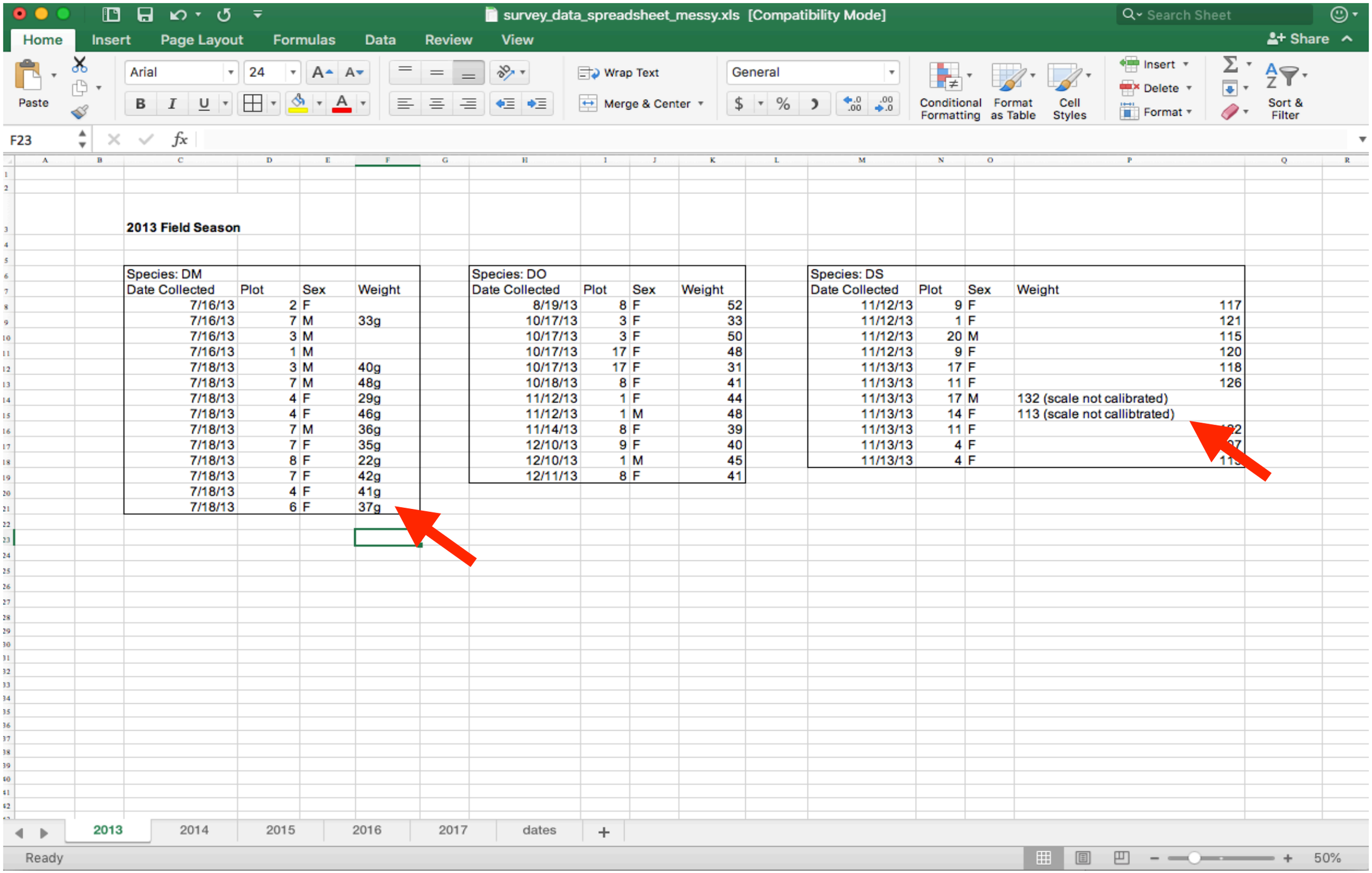
II. Formatting problems

➤ Adding more than one information in a cell



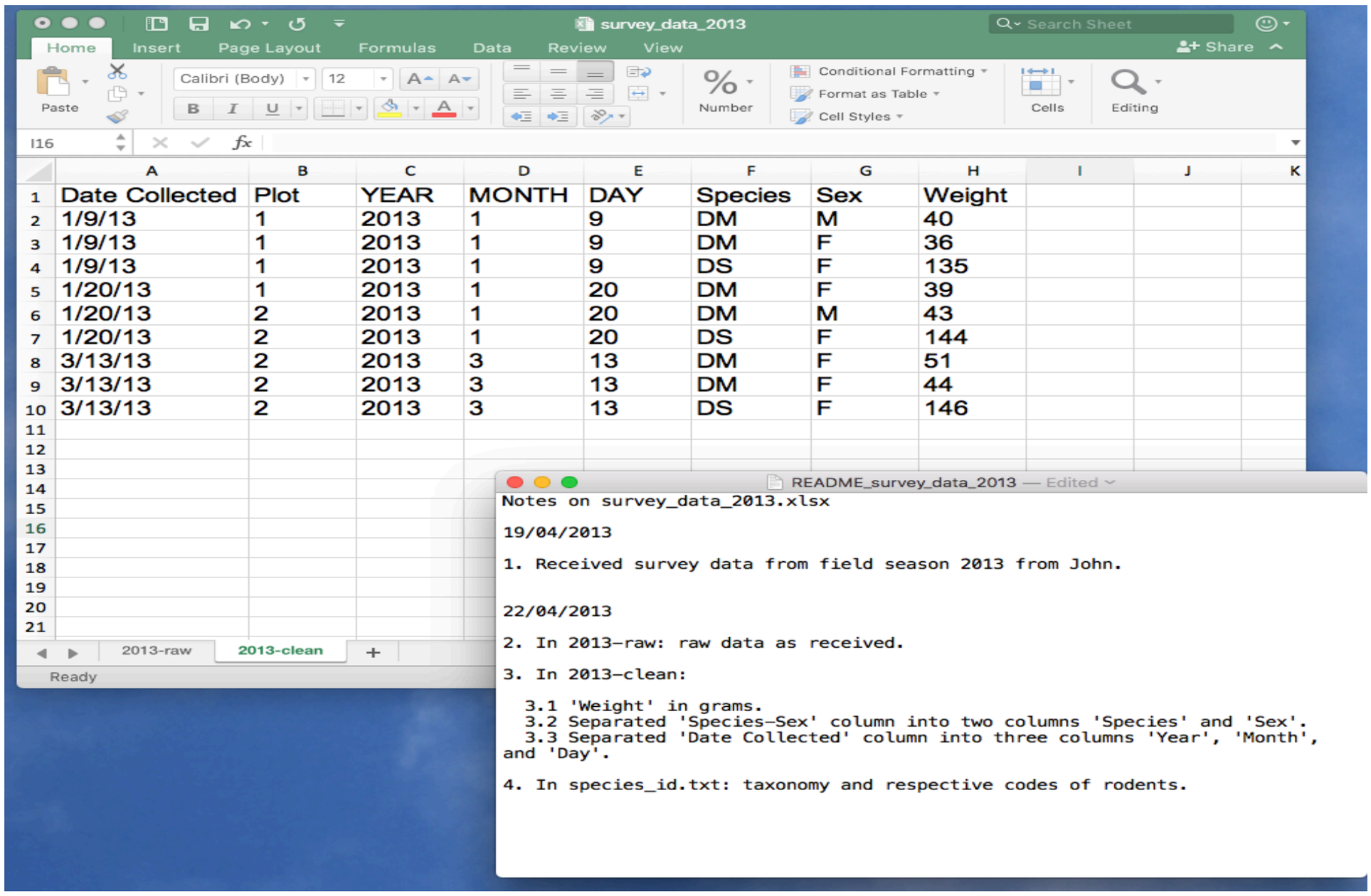
II. Formatting problems

- Adding meaning of data, of column, comments, or units



III.Dates as data

- Storing date as a single entry is not the best practice
- YEAR, MONTH, and DAY in separate cells.



	A	B	C	D	E	F	G	H
1	Date Collected	Plot	YEAR	MONTH	DAY	Species	Sex	Weight
2	1/9/13	1	2013	1	9	DM	M	40
3	1/9/13	1	2013	1	9	DM	F	36
4	1/9/13	1	2013	1	9	DS	F	135
5	1/20/13	1	2013	1	20	DM	F	39
6	1/20/13	2	2013	1	20	DM	M	43
7	1/20/13	2	2013	1	20	DS	F	144
8	3/13/13	2	2013	3	13	DM	F	51
9	3/13/13	2	2013	3	13	DM	F	44
10	3/13/13	2	2013	3	13	DS	F	146

IV. Quality control

- Avoid bad data from being entered in the first place.
- Spot bad entries in the data set.
- Quality control —> always keep an original spreadsheet untouched, make separate copies with modified data and rename appropriately, and ensure data is stored as values.
- Create README files (:
- Download this file (https://github.com/datacarpentry/spreadsheet-ecology-lesson/blob/gh-pages/data/survey_sorting_exercise.xlsx?raw=true) for this demonstration.

V. Exporting data

- Excel default file format `.xls` or `.xlsx`.
- Different versions of Excel and different spreadsheet programs handle data differently leading to inconsistencies.
- Journals and data repositories might require `.tsv` or `.csv`.
- Tab-delimited (tab separated values or TSV) or comma-delimited (comma separated values or CSV).
- CSV files are plain text files where the columns are separated by commas.
- Advantages: open and read using any software, easily imported by SQL, R and Python, not tied to a software version.

V. Exporting data

To save a file you have opened in Excel in CSV format:

1. From the top menu select 'File' and 'Save as'.
2. In the 'Format' field, from the list, select 'Comma Separated Values' (*file_name*.csv).
3. Double check the file name and the location where you want to save it and hit 'Save'.

IMPORTANT: If you are working with data that contains commas, consider using tabs as your delimiter and working with TSV files. TSV files can be exported from spreadsheet programs in the same way as CSV files.

SUMMARY

- Variables in columns, observations in rows, and one value per cell.
- Focus on the data.
- Keep track of clean up, comments, units, etc. in a README.txt.
- When using large tables: Sort & Filter, add conditional formatting to check for bad data.
- When entering values: add data validation criteria to columns.
- Opt for saving data tables in .csv or .tsv.