Data Organization in Spreadsheets

June 18, 2019

Instructors

- Ryan Clement, Data Services Librarian
    - he/him pronouns
    - Librarian for economics, geography, sociology, anthropology, philosophy
    - Find me at go/ryan/
- Wendy Shook, Science Data Librarian
    - Librarian for physics, chemistry, biology, computer science, geology, environmental studies
    - Find me at go/sciencelibrarian/

Setup

Spreadsheet software

To interact with spreadsheets, we can use LibreOffice, Microsoft Excel, Gnumeric, OpenOffice.org, Google Sheets, or other programs. Commands may differ a bit between programs, but the general ideas for thinking about spreadsheets are the same.
For this lesson, if you don't have a spreadsheet program already, you can use LibreOffice. It's a free, open source spreadsheet program. You can also use Google Sheets; if you don't have your own Google account, you can use your Middlebury login.

Data files

Below are the necessary files for the workshop. Please download them when instructed.

- https://ndownloader.figshare.com/files/2252083

Exercises

1. We're going to take a messy version of the survey data and describe how we would clean it up.
    a. Download the data by clicking here to get it from FigShare.
    b. Open up the data in a spreadsheet program.

    c. 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.
    d. 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.
    e. **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.
    f. After you go through this exercise, we'll discuss as a group what was wrong with this data and how you would fix it.
2. Challenge: pulling month, day and year out of dates
    a. In the dates tab of your spreadsheet you have the data from 2014 plot 3. There's a Date collected column.
    b. 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()]
    c. (Make sure the new column is formatted as a number and not as a date.)
3. Challenge: pulling hour, minute and second out of the current time
    a. 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.
    b. Extract the year, month and day from the current date and time string returned by the NOW() function.
    c. Calculate the current time using NOW()-TODAY().
    d. Extract the hour, minute and second from the current time using functions HOUR(), MINUTE() and SECOND().
    e. Press F9 to force the spreadsheet to recalculate the NOW() function, and check that it has been updated.
4. 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
    a. Once downloaded, sort the Weight_grams column in your spreadsheet program from Largest to Smallest.
    b. What do you notice?
5.

Notes

1. Use data validation.
    a. https://datacarpentry.org/spreadsheet-ecology-lesson/04-quality-control/index.html
2. Conditional formatting > color scales

3. Save data as CSV UTF-8 (Comma delimited) (.csv)