Lisa Song, ProPublica
Oct 7, 2017
lisa.song@propublica.org

SEJ 2017
Data Journalism: How to Find It, Mine It, Animate It

Outline of intro to google sheets session

For a quick intro to data reporting, see http://bit.ly/2ciYEMZ

--it's a modified version of a document I wrote last year, after attending the 2016 ProPublica Data Institute (http://bit.ly/2chCaed), where they taught us the basics of data reporting, data visualization, web design and coding.

--I was working for InsideClimate News at the time. We didn't have a data reporter, and I wanted to do more simple data stories, so I wrote the memo to encourage my colleagues to help me think of data stories, and to learn the basics of data reporting if they wanted to.

--the document includes resources for data journalism training and links to websites where you can teach yourself data analysis, coding, etc

What we learned/talked about during the session

The TCEQ dataset we worked with at the conference is at http://bit.ly/2wLGjoZ
--please copy this into your own google drive before attempting to edit

Terminology: workbook, worksheet, column, row, cell, header row

Data dictionary/record layout/code sheet: most databases should have one of these. It's a document that explains what the columns mean. Always ask for the data dictionary that accompanies the dataset when you're FOIAing for a database.

Data format: numbers, text, date, time, etc. Remember to change zip codes and other numerical codes into the text format, so you don't accidentally try to do math on them. In the TCEQ database, the columns SIC and REGION should both be in the text format.

Data cleaning: most databases are "dirty" and need to be cleaned and checked for typos, spelling variations, missing data, etc. You can use OpenRefine to clean your data. If there's a lot of missing data, you may not be able to use the dataset, since your results won't be valid.

Functions: for example, =SUM(A2:A20) means add up all the values from cell A2 to cell A20

--there are many other google sheet functions, such as =AVERAGE(cell1:cell2), =PRODUCT(cell1:cell2), etc. Just google for Excel functions and you'll find many more

Sorting: when you place the rows in order according to a particular column

Filtering: you when filter the worksheet down to a particular subset of rows

Best practices
--always make a copy of the original dataset, and work off the copy. You may need to consult the original later for fact checking

--keep a data diary--a record of every step of your data analysis. You'll need this later on to be able to retrace your steps and check your work

--have someone else check your work if you can, for extra bulletproofing