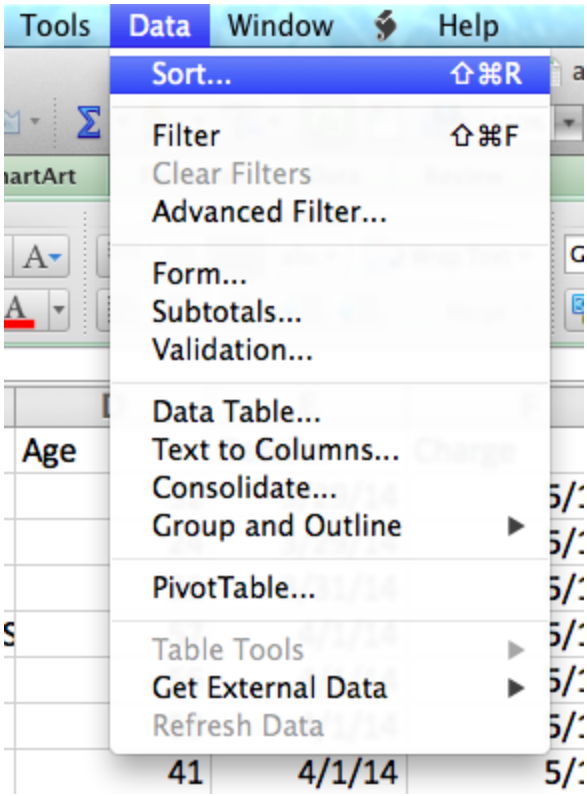
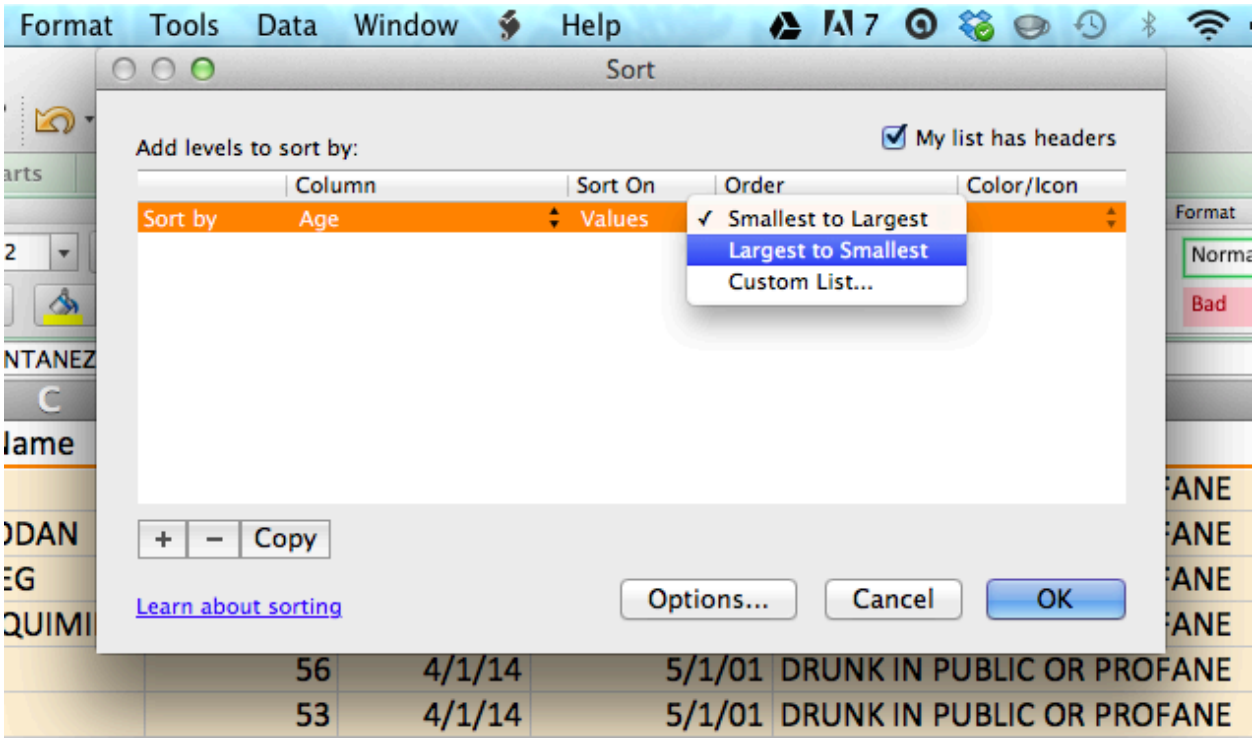
Interviewing data, by Derek Willis

Journalism Interactive, April 5, 2014
Derek's outline: http://dwillis.github.io/interviewing-data/
File: arrest.csv (download this & open it in Excel, then follow along)

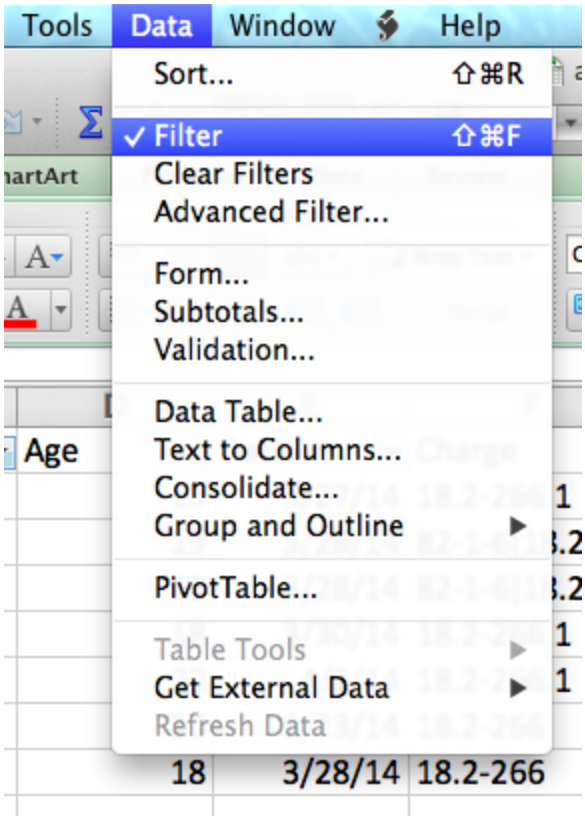
These are notes taken in Derek's session by Mindy McAdams @macloo
They are published with Derek's permission.

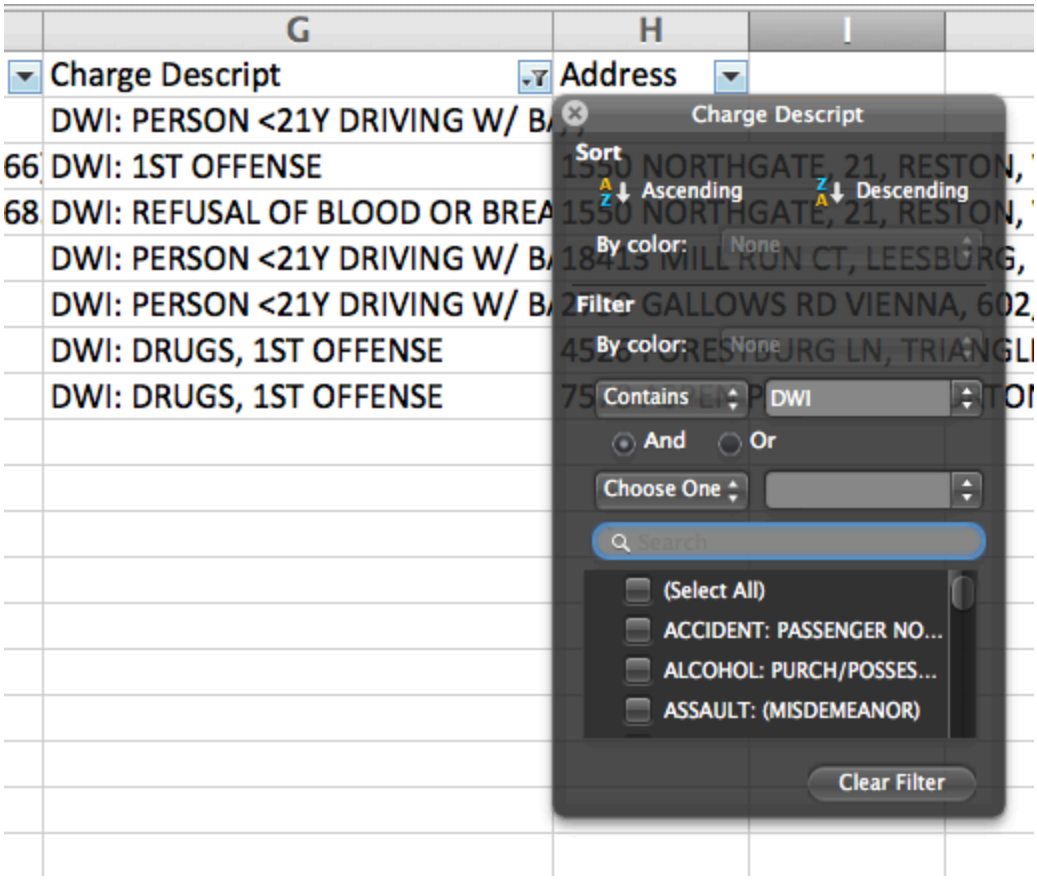
1. Get data.
2. Look it over, see what you've got.
3. Be very skeptical (it probably has things wrong with it: not merely cosmetic, but things missing, or things that don't make sense).
4. Figure out what's wrong with it (perfect for them, whoever made this dataset; NOT perfect for YOU).
5. **Sort** (Note: Never sort on data w/o headers! Add headers if they're not there).



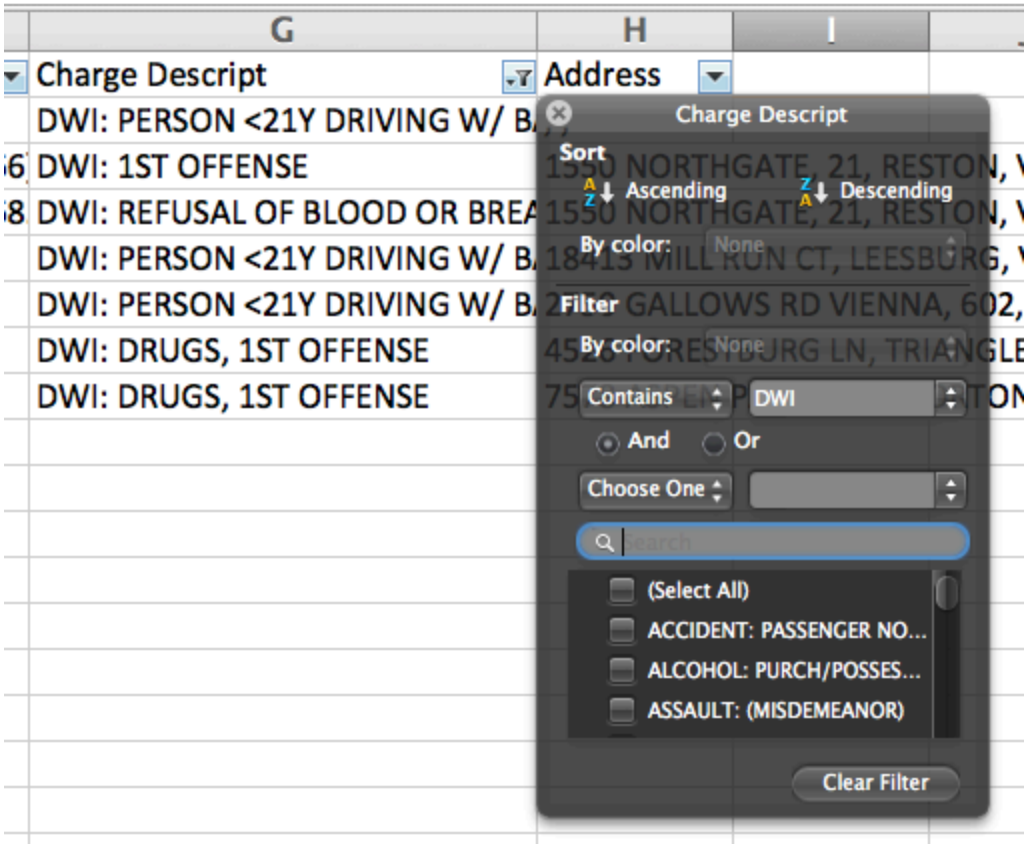


6. Sort arrests by age: What do you find? No one under age 18? What do you know? Juveniles not included. Thus not all people arrested in that time period. Example of a sort: Arrests, sort by age. See if any ages are missing. (If ages wrong, probably MORE is also wrong.)
7. Sort on date: Maybe there are outliers. In this dataset, two arrests dated December.
8. Names that repeat: This is not a list of people; it's a list of CHARGES (count of all lines does not give you how many people were arrested, b/c of duplicates).
9. Sort on charges: Excel error, converts to date on Charge (5/1/01) - How to fix: re-download, do IMPORT, tell that column to be TEXT not GENERAL.
10. Sort on address: Some people have none, but having no address is treated inconsistently: e.g., 2 commas, or "no fixed," or "unknown."
11. Now we know — there are some issues with these data.
12. Address is for the person, not the crime. So we do not know WHERE the crime happened.
13. This interview with the data should help you figure out WHAT TO DO NEXT.
14. Arrest records: What do we care about? Violent crimes. Which is most violent? Murder.
15. **FILTER.** Data menu - turn on Filter there, on menu. Widget appears on each column header. Widget opens dark gray menu where we do magical things.
    Turn filters on:

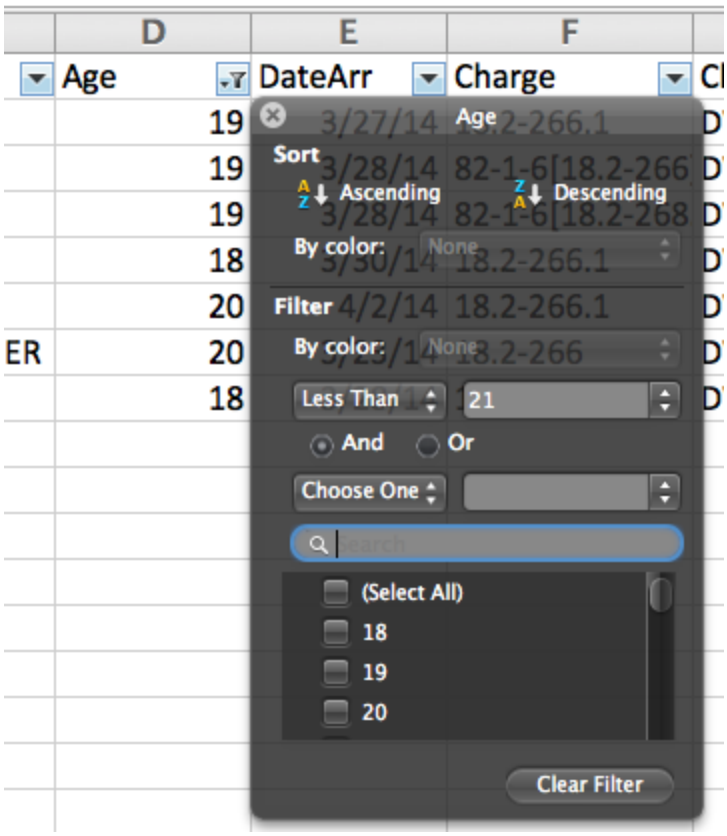

Tools
Data
Window
Help
Sort... ⇧⌘R
Filter ⇧⌘F
Clear Filters
Advanced Filter...
Form...
Subtotals...
Validation...
Data Table...
Text to Columns...
Consolidate...
Group and Outline
PivotTable...
Table Tools
Get External Data
Refresh Data
Age
18 3/28/14 18.2-266


G
H
I
Charge Descript
Address
DWI: PERSON <21Y DRIVING W/ BA
DWI: 1ST OFFENSE
DWI: REFUSAL OF BLOOD OR BREA
DWI: PERSON <21Y DRIVING W/ BA
DWI: PERSON <21Y DRIVING W/ BA
DWI: DRUGS, 1ST OFFENSE
DWI: DRUGS, 1ST OFFENSE
Charge Descript
Sort
Ascending
Descending
By color:
None
Filter
By color:
None
Contains
DWI
And
Or
Choose One
Search
(Select All)
ACCIDENT: PASSENGER NO...
ALCOHOL: PURCH/POSSES...
ASSAULT: (MISDEMEANOR)
Clear Filter

16. In arrests list: Filter on Charge Descript. Type in “murder”? Nothing. Type “homicide”? Nothing. (In "Contains.") But type in "monument": This is Fairfax, Va., data, and they've been having a lot of monument defacement lately. So you get a list of monument damage!
17. Combining filters. Example: filter “Charge Descript” on “Contains” DWI
18. **Tip:** Start broad w/ filters, narrow it down, then go narrow. Otherwise you might come up with nothing.
19. Then: Add Boolean AND or OR (in the filter menu); next parameter can be "Does not contain." E.g. DWI and 2nd (for 2nd offense).



20. **Tip:** Good to always start with "Contains."
21. Or — we can add a second filter on a different column. That is, one filter on each of two columns. Note: The order matters. E.g. first filter on DWI charge, THEN filter on age (Less than 21). (Not the opposite way.) **Note:** Must CLEAR a filter to make it go away. Clear is on the dark gray filter menu, near bottom. (see images above and below)



22. **Count:** Bottom middle of sheet shows you the count (from your filter). It's a pop-up menu. Get count, sum, etc., for sheet here. (see image below)

CONCLUSION

Sort and **filters** cover about 90 percent of what you want to do when interviewing the data, Derek said. If you can't sort and filter, you're not actually using a spreadsheet.

“Get students to think about ways they can start interviewing data. Ways they can incorporate data into their reporting.” —Derek Willis