Evaluating Data

paranoia is a ~~valuable~~ necessary trait

poking holes is job one

- Do Not Trust Your Data.
  - Even (especially?) if you made it.
- Data comes from people. People make mistakes.
- First law of data entry: the more you type, the more mistakes you make.

what to look for

- missing values
- inconsistencies between data points
- out of range values
- suspicious repeats / duplicates
- misspellings / typos
- truncations within cells
- entire rows truncated from your dataset

https://twitter.com/bendystraw/status/736270952670081026

Start evaluating your data!

- First, basic question: what is this?
- Then, sort and filter to get familiar
- Next, pivot tables to see distinct values

All of the steps inform our judgement on whether this data is useful and can answer our questions.

Pivot Tables

- Great for when you need to look at distinct values
- Distinct means exactly that - any difference means a different value
- Use one pivot table at a time to stay sane

problems with names

- multiple versions of similar names (AT&T)
- suffixes, prefixes, titles
- marriages, divorces, births

problems with addresses

- abbrevs.
- partial addresses, no street number
- intersections
- quadrants
- PO boxes

When to walk away

Sometimes your data is simply not robust, detailed, accurate, or large enough to do what you want with it