

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 do look for

- Rows with values that are _____
- Toddler criminals
- Extreme outliers that aren't like other values
- Inconsistencies
- Truncated data

Fairfax County arrests

- We start with a basic question: what is this?
- Sorting and filtering to get the basics
- Pivot tables for finding distinct values

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

Pivot Tables

- When you need to look at distinct values
- Distinct means exactly that - any difference means a different value
- One pivot table at a time

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