

# Evaluating Data

Why Paranoia is a ~~Valuable~~ Necessary Trait

# Poking Holes is Job One

- Don't 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

- 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 judgment 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

# The Problems with Names and Addresses

- Names

- Multiple versions of similar names (AT&T)
- Suffixes. Prefixes. Titles.
- Marriages. Divorces. Births.

- Addresses

- Abbrevs.
- Partial addresses (no street number)
- Intersections
- Quadrants
- PO Boxes