Exploratory Data Analysis (EDA)

(20-ish minute crash course)

DSSG Solve for Good Summer Academy

Why EDA?

“All data is wrong!”

-- somebody wise (and possibly annoyed)

- Data has ***limitations*** depending on your use case (resolution, anonymization, omitted variables), need to make sure data makes sense.
- There are likely ***biases*** on the data based on how it was collected (e.g., response biases, social desirability, selection bias)
- There can be ***missingness*** in the data; it’s important to understand why.
- Data was probably not collected for your explicit purpose leads to a different type of curation.
- Develop ***domain expertise*** to understand the underlying generative process, limitations, and idiosyncrasies (easier to do in partnership with a domain expert).

How to do EDA?

- Everything happens somewhere and at sometime:
  - Understand trends over time
  - Understand trends over space
- Plot Summary Statistics of Variables to understand where they sit and deviation (mean, median, standard deviation, percentiles) lots of good resources about metrics.
  - Central Tendency/Variability
  - Correlations between variables
  - Cross-Tabs between different groups
- How rapidly do distributions change over time (seasonality, exogenous shocks, different definitions of labels)
- Understand missing data and outliers

Bias in Data

- Response Bias
- Selection Bias
- Social Desirability Bias
- Concept Drift
- Omitted Variables
- Missingness in the data.

Tales from the Data

- Syracuse (Prevent Water Main Breaks)
  - At first, the water mains that broke the most were the newest. It turns out those were the ones that were recently replace and the date of installation was overwritten.

- Recidivism Prediction
  - Entry and Exit Dates were flipped for several months which means that having negative days in jail was highly predictive of recidivating back to jail.

- HIV Retention Early Warning System.
  - ICD-10 codes for billing extremely inconsistent. Cases where emergency contact is missing is a measure of social vulnerability and predictive of dropping out-of-care.

Tools

- Databases and SQL (Use databases and learn SQL, scalable and efficient)
- Python
  - Pandas (Not as efficient as a DB or SQL. Not scalable, not efficient, clunky)
  - Visualization (Matplotlib, Seaborn, other exotic libraries)
- R
  - Ggplot2
- Tableau (Proprietary but very powerful)