EDA - Exploratory Data Analysis

STAT 140SL

WINTER 2021

What is EDA?

- EDA assumes that data has an underlying structure and the goal of EDA is to uncover that structure.
- EDA is like digging for gold and seeking value in data (leveraging data)
- EDA requires practice, you need to examine a lot of data to begin understanding its value.

It's all about

- Starting the process of converting data into useful information
- Laying the foundation for providing answers to questions being asked by a client or by a researcher
- Note - Data analysis is not the same as data quality:
  - A large, high quality dataset is not useful if poorly analyzed
    A small, low quality dataset can be useful if correctly analyzed.

It's all about (cont'd)

- Possessing a solid set of tools (e.g., R, SQL, stats training)
- Keeping focused / being guided by the questions you have been asked to answer.
- Having a desire to learn about the data and subject

Example - Kaggle San Francisco Crimee

- https://www.kaggle.com/c/sf-crime
- The training set is useful for EDA, the original competition was to explore the data and get the most upvotes.
- This data set is large but not too large.

Example

- Example Question: “What is the relationship between time and crime?”
- Analysis : Explore the occurrence of crime conditioning on the time variable. (What does this mean to you?)
- Interpretation : At first glance, is the pattern of crime over time random or not?

EDA Characteristics

- Descriptive — distributions & patterns, conditioned over aspects of the question of interest (e.g., time, groups, geography, etc.)
- Not focused on causality – descriptive analysis tells you the “what”, “where”, “when”, and “who” about the data, not the why/how.
- Example – average number of crimes per month, week, day, hour etc.

Basic Descriptive Stats

- Mean, median, mode,
- quartiles, percentiles, minimum, maximum
- Standard deviation, IQR, skewness, kurtosis, range

Data Types

- Binary - has only two values
- Nominal - can have many values, no intrinsic order
- Ordinal - can have many values, order matters
- Discrete - numerical whole
- Continuous - numerical

Example Analysis

- Exit this slide and go to the folder in Week 7

Missing Values

- Data may be missing for numerous reasons. Input may be incorrect (e.g., typographical error) or design generated missingness (e.g., a long survey may bore/irritate respondents so they don't answer the last page).
- Missing data is often represented by blanks or NA. However, sometimes missing data might be represented by 0, 9, 99. Why might these be unwise representations?
- R functions will sometimes fail when missing values are present - is this a good thing or?

Missing Values

- Many studies/experiments assume “random sampling” and/or “random assignment”. When data are MCAR (missing completely at random), there is no violation of these important assumptions. MCAR means missing is independent of measured and unmeasured variables.

- Not missing at random (NMAR) can introduce errors in analysis. If we are studying bullying in high school and the students who are the victims of bullying are more likely to be non-respondents than the bullies, this creates problems for our analysis.