Exploratory Data Analysis

Exploratory Data Analysis (EDA)

- Promoted by John Tukey since the 1970s
- Summarizing the main characteristics of a data set through statistical summaries and visualizations
- A statistical model can but used but is not generally the purpose of EDA
- EDA is used as a way to find insights and possibly generate hypotheses

John Tukey on Visualization

The father of exploratory data analysis

- "The greatest value of a picture is when it forces us to notice what we never expected to see"
- "This is my favorite part about analytics: Taking boring flat data and bringing it to life through visualization"
- "When communication results to non-technical types there is nothing better than a clear visualization to make your point"
- "Visualization is often used for evil - twisting insignificant data changes and making them look meaningful. Don't do that crap if you want to be my friend. Present results clearly and honestly. If something isn't working - those reviewing results need to know."

The EDA Cycle

- EDA is an iterative cycle:
  - Generate questions about the data
  - Search for answers through visualizing, transforming, and possibly modeling the data
  - Use what is learned to refine question and/or generate new questions

Purpose of EDA

- Not a formal process with a strict set of rules
- EDA is a “state of mind” (from R4DS)
- The goal is to develop an understanding of the data
  - Ask questions of the data
  - Explore all ideas that occurs, realizing that some will not pan out and lead to dead ends.
  - Hopefully, a few ideas will emerge that will lead to insights to be communicated to others

Purpose of EDA

- EDA is an important part of any data analysis:
- Even if questions about the data are already determined (i.e., from a boss, teacher, client, etc)
  - Data cleaning is part of EDA
  - Determine if data meets expectations / can answer desired questions

Questions to ask

- Since EDA is a creative process, often the ***quantity*** of questions to start with is more important that the ***quality*** of questions
- The quality ideas will emerge through the EDA
- While there are no rules about the particular questions, two types of questions will (almost) always be useful
  - What type of variation occurs within the variables?
  - What type of covariation occurs between the variables?