PSYC1495 Week 3: Data sources

September 17th, 2025

Agenda

- I made another syllabus switch – sorry, I think I'm done now

- Publicly available data
  - Where to find?
  - Find your own

- Coding tutorial

Secondary data: General repositories

- Tidy Tuesday
  - Online community for plotting in R
- Awesome Public Datasets
  - List of datasets in tons of different categories
- Google Dataset Search
  - Search by keyword, find related data sources from across the internet
- ProPublica data
  - Data that ProPublica has used in their articles

Secondary data: Governments

- NYC OpenData
  - Datasets from NYC city agencies
  - Each dataset is downloadable, but also possible to wrangle data and make plots on their platform directly
- Most city governments have something similar
- Data.gov
  - Combines datasets (350k+) from across government sources, searchable
- HealthData.gov
  - Data from Health and Human Services

Secondary data: Justice-specific

- Datasets for Justice Innovation
  - Collection of justice-related datasets from a group at Stanford Law
- Prison Policy Initiative
  - Datasets used in briefings, most are related to incarceration
- The Vera Institute
  - Data for our first day of coding came from here

Accessing data more easily with R packages

- There are some R packages that are designed for accessing data from specific sources
  - Tidycensus: Accessing US census data
  - Eurostat: Accessing EU statistics
  - blscrapeR: Bureau of Labor Statistics
  - ngramR: Google n-gram
- R packages can also make it easy to access data from online platforms

Secondary data: Online platforms

- Social media
  - Reddit: RedditExtractoR
  - Youtube: TubeR

- Large databases
  - Google trends: gtrendsR
  - Wikipedia: getwiki or wikipediatrend

- News
  - New York Times: nytimes

- Rvest can help you get data from almost any HTML-based website

Finding data on your own

- With a partner...
    - Look up what data is available from each of your hometowns
        - Could be city-level, county, state, etc.
    - Together, familiarize yourself with the websites
    - Try to identify one or two cool datasets from each

More resources

- Here are some resources I've come across since posting these slides

- Data Collaborative for Justice

- Pew Research Center

- General Social Survey
  - There's an associated R package called gssr

- Opportunity Insights