Proposal

Milestone 2

Goals

The goals of this milestone are as follows:

- Discuss topics you're interested in investigating and find data sets on those topics.
- Identify 3 data sets you're interested in potentially using for the project.
- Get these datasets into R.
- Write up reasons and justifications for why you want to work with these datasets.
- Review your team contract.

> ! Important
>
> You must use one of the data sets in the proposal for the final project, unless instructed otherwise when given feedback.

Finding a dataset

Criteria for datasets

The data sets should meet the following criteria:

- At least 500 observations
- At least 8 columns
- At least 6 of the columns must be useful and unique explanatory variables.

  - Identifier variables such as “name”, “social security number”, etc. are not useful explanatory variables.
  - If you have multiple columns with the same information (e.g. “state abbreviation” and “state name”), then they are not unique explanatory variables.

- You may not use data that has previously been used in any course materials, or any derivation of data that has been used in course materials.

- You can curate one of your datasets via web scraping.

Please ask a member of the teaching team if you’re unsure whether your data set meets the criteria.

If you set your hearts on a dataset that has fewer observations or variables than what’s suggested here, that might still be ok; use these numbers as guidance for a successful proposal, not as minimum requirements.

Resources for datasets

You can find data wherever you like, but here are some recommendations to get you started. You shouldn’t feel constrained to datasets that are already in a tidy format, you can start with data that needs cleaning and tidying, scrape data off the web, or collect your own data.

- Awesome public datasets
- Bikeshare data portal
- CDC
- Data.gov
- Data is Plural
- Durham Open Data Portal
- Edinburgh Open Data
- Election Studies
- European Statistics
- CORGIS: The Collection of Really Great, Interesting, Situated Datasets
- General Social Survey
- Google Dataset Search
- Harvard Dataverse
- International Monetary Fund
- IPUMS survey data from around the world
- Los Angeles Open Data
- NHS Scotland Open Data
- NYC OpenData
- Open access to Scotland’s official statistics
- Pew Research
- PRISM Data Archive Project

- Statistics Canada
- TidyTuesday
- The National Bureau of Economic Research
- UCI Machine Learning Repository
- UK Government Data
- UNICEF Data
- United Nations Data
- United Nations Statistics Division
- US Census Data
- US Government Data
- World Bank Data
- Youth Risk Behavior Surveillance System (YRBSS)
- FRED Economic Data

Components

For each data set, include the following:

Introduction and data

For each data set:

- Identify the source of the data.
- State when and how it was originally collected (by the original data curator, not necessarily how you found the data).
- Write a brief description of the observations.
- Address ethical concerns about the data, if any.

Research question

Your research question should contain at least three variables, and should be a mix of categorical and quantitative variables. When writing a research question, please think about the following:

- What is your target population?
- Is the question original?
- Can the question be answered?

For each data set, include the following:

- A well formulated research question. (You may include more than one research question if you want to receive feedback on different ideas for your project. However, one per data set is required.)
- Statement on why this question is important.
- A description of the research topic along with a concise statement of your hypotheses on this topic.
- Identify the types of variables in your research question. Categorical? Quantitative?

Glimpse of data

For each data set:

- Place the file containing your data in the `data` folder of the project repo.
- Use the `glimpse()` function to provide a glimpse of the data set.

Grading

Total	10 pts
Introduction and data	3
Research question	3
Glimpse of data	3
Workflow and formatting	1

Each component will be graded as follows:

- **Meets expectations (full credit)**: All required elements are completed and are accurate. The narrative is written clearly, all tables and visualizations are nicely formatted, and the work would be presentable in a professional setting.
- **Close to expectations (half credit)**: There are some elements missing and/or inaccurate. There are some issues with formatting.
- **Does not meet expectations (no credit)**: Major elements missing. Work is not neatly formatted and would not be presentable in a professional setting.

It is critical to check feedback on your project proposal. Even if you earn full credit, it may not mean that your proposal is perfect.