STARS@SLU
Data Science Research Project

Summer 2025

STARS students will apply data analysis and visualization methods to analyze a large, interesting data set. Teams will produce a conference style poster and present it at the end of the week.

Timeline

Monday	9-11:15	Data discovery.
Monday	1:00-2:30	**Proposal due.** Select project topic.
Tuesday	9-11:15	Data cleaning.
Tuesday	1:00-2:30	**Data overview due.** Data exploration.
Wednesday	9-11:15	Data exploration.
Wednesday	1-2:30	Poster concept.
Thursday	9-11:15	Create poster. **Poster draft due.**
Thursday	1-2:30	Poster presentation practice. **Final poster due.**
Friday	12-1	Prepare for poster session.
Friday	1-2	**Poster session.**

Data discovery & project proposal

In data science work, much of the effort involved is in the collection, arrangement, and cleaning of data prior to performing statistical analyses. This is an iterative process, as visualizations and computations often reveal subtleties in the original data that need to be addressed.

Identify at least three data sets which are:

- Available to you at no cost.
- Well documented.
- Contain at least 1000 observations.
- Contain a variety of interesting variables, both categorical and quantitative.
- Smaller than 100 megabytes.

Make sure you can access these data sets, read them into R, and at least do `str()` and `head()` on it. Your data may still require heavy manipulation and cleaning at this time.

Try to select data sets that will lead you to questions you care about. These could be related to your academic or career interests, hobbies or other interests. In the end, you will want to learn something from the data that matters to you.

Proposal

Create a Quarto document called “proposal” with one section for each of your proposals. Each section should include:

- A title and short (1 line) description of the proposal.
- A reference for the data: Who collected it, who distributes it, where is it documented? Include descriptions as well as links.
- Specifications. How large is the data (in bytes)? How many observations? How many variables? What are the interesting variables and what type are they (character, numeric, categorical, time, place, etc.)
- Interesting questions the data might be able to answer. You might consider single variables, where the answer is a summary statistic, chart, or list. For example: “Which drug companies sold the most oxycontin in Missouri?”. Consider relationships between variables, for example: “How do house prices depend on lot size and location?”

Data cleaning

Probably the most challenging step in this project is cleaning your data. Data, when loaded, is rarely in the format you need. Missing values might be coded in strange ways. There may be errors and outliers. Dates and numbers may be strings. Values may not sort properly. Cleaning your data is an iterative process. Look at the data with plots or tables, identify problems, recode or restructure the data, and repeat.

Your data cleaning should be reproducible, which means all the commands needed to clean the data should be in an R script that loads the raw data and produces a useable data frame. Don’t hand edit any data along the way.

Data overview

Create a data overview Quarto document that displays each variable you care about in some form. For example, as a table, top ten list, histogram, boxplot, bar chart, or scatterplot.

Data exploration

At this stage of the project, find the story you want to tell with your data. Which results are the most compelling, which are the most important? If needed, incorporate new sources of information for context, for example census information, geocoding, sentiment analysis. Refine your visualizations so that they are easy to understand and visually appealing.

At this stage, you should have a Quarto document that performs any analysis and creates any charts that will be in your poster.