**Data Visualization  
140.610.79**

*Last updated June 6, 2025* **Final Project**

You will create and submit a scientific notebook in which you select a topic or research question, identify a relevant dataset (or simulate one), and develop a series of visualizations to explore and communicate insights from the data.

The goal is to demonstrate your ability to use data visualizations to explore, explain, and communicate insights from real or simulated data.

**Data**

You may use either a real-world dataset or a simulated dataset for your project.

Real Data

You are encouraged to choose a dataset that is relevant to your interests, research, or field of study. See the list of recommended sources at the end of this document for ideas.

Simulated Data

If you do not have access to a real dataset that fits your topic, you are welcome to simulate your own. Simulated data can be a powerful way to:

* Explore hypothetical scenarios
* Test visualization techniques
* Replicate common patterns seen in public health and epidemiology

If you choose to simulate your data, **be sure to include the code used to generate the data** with your project.

**Submission Requirements**

* You must submit a scientific notebook (Quarto, Rmarkdown, or Jupyter)
* You must include **10 total graphics**, with the following breakdown:
  + **Up to 5** can be univariate, exploratory visualizations
  + **At least 5** must involve **two or more variables**
  + **At least 2** of those multivariable graphics must show **three or more variables**
* Each visualization must include:
  + The **code** used to generate the plot
  + **A brief explanation** of why the visualization was chosen
  + The **rendered image** of the visualization
  + **Attention to detail**, including clear labels, legends, and annotations—**visuals should be polished to publication submission-ready quality**

Notebooks can be more challenging for users of certain statistical software. If you prefer to put together a document using word processing software (e.g., Word, Google Docs), that is fine, but you must still include code, images, and a justification.

**Choosing What to Visualize**

Given the varied nature of problems people in the class work on and the data relevant to those problems, it will be up to the student to figure out appropriate visualizations. It may help you to consider the following concerns:

* Data quality: missing data, outliers, etc.
* Exploratory data analysis
* Analysis results such as statistical models

**Grading and Evaluation**

Projects will be evaluated based on:

1. Clarity and appropriateness of visualizations
2. Quality of code and documentation
3. Effectiveness of visual explanations and justifications
4. Visual design: labeling, layout, color, legibility
5. Reproducibility and organization of the notebook

Each criterion will be assessed on a 1–5 scale, where:

* 5 – Excellent: Clear, polished, exceeds expectations
* 4 – Good: Meets expectations with minor issues
* 3 – Satisfactory: Adequate but with notable weaknesses
* 2 – Needs improvement: Major issues present
* 1 – Incomplete or unclear: Requirements not met

|  |  |
| --- | --- |
| **Category** | **Description** |
| Visualization Clarity & Appropriateness | Are the chosen visualizations well-suited to the data and research question? Do they aid understanding? |
| Code Quality & Documentation | Is the code organized, readable, and reproducible? Are comments or explanations clear and helpful? |
| Explanation & Justification | Does each visualization include a clear rationale and interpretation? Do justifications show insight? |
| Visual Design & Presentation | Are visual elements (labels, colors, layout) polished and effective? Are plots legible and well-designed? |
| Notebook Organization & Reproducibility | Is the notebook easy to follow? Can the results be reproduced from the code and data provided? |

**Due Date and Submission**

The assignment is due Thursday, July 3, three weeks after our last session. If you need an extension, please contact me.

Submit your project via CoursePlus. Be sure to upload both your code and the rendered document. Alternatively, you may provide a link to a GitHub repository containing your completed work.

**Possible Data Sources**

* **CDC Data**: <https://data.cdc.gov/browse>
* **CDC Wonder**: <https://wonder.cdc.gov/>
* **National Health and Nutrition Examination Survey (NHANES):** <https://wwwn.cdc.gov/nchs/nhanes/>
* **Behavioral Risk Factor Surveillance System (BRFSS):** <https://www.cdc.gov/brfss/annual_data/annual_data.htm>
* **WHO Global Health Observatory:** <https://www.who.int/data/gho>
* **Our World In Data:** <https://ourworldindata.org/>
* **USA Facts:** <https://usafacts.org/>
* **Demographic and Health Surveys:** https://dhsprogram.com/data/available-datasets.cfm
* **Stata and local data:** e.g., <https://opendata.maryland.gov/>
* **Sim data:** Make your own data; this is a great way to learn