# Semester Project Assignment

Jonathan Gilligan

2025-03-17

## Contents



## Semester Project Assignment

### Overview

You will investigate a data set of your choice. The project will involve several steps, and will culminate in a written report, produced using R and Quarto (if you strongly prefer to use a different set of tools, that is fine. Python can work with Quarto too, and will allow you to continue to use RStudio to develop your project.)

You will use R to import data, analyze it, and answer some questions about it.

The GitHub Classroom assignment for the project is https://classroom.github.com/a/pGhugW_l.

### Details

#### Choosing a data set

You can use data from any source. This could be research data from a project you have worked on or are working on, or it could be a publicly available set of data from some other source, such as the US Geological Survey, the National Oceanic and Atmospheric Administration, NASA, etc. It does not need to be geological or environmental data.

#### Major Parts of the Project:

The project will include the following parts:

1. Importing the data into R, and converting it into one or more `data.frames`, `tibbles`, or other data structures.

2. Study the descriptive statistics of your data and identify possible probability distributions you think it may have come from.

    - Report the mean and standard deviation of each variable in your data (e.g., temperature, rainfall, isotopic compositions, mineral content, etc.).
    - Plot the distribution of each variable using histograms and kernel density estimates.
    - Using common probability distribution functions that we have studied, determine whether any of them seems to describe the data well. Q-Q plots will probably be helpful to you.
    - If you are interested in relationships between different variables in a multivariate data set, report correlations and optionally produce a correlation plot.

3. Come up with some questions you want to ask about the data.

    This could be whether data from one set of observations is different from another (e.g., the in-class exercise about whether droughts were different in the Little Ice Age and the Mediaval Warm Period).

    Another kind of question is whether there is a relationship between two or more variables in a multivariate data set (e.g., analyze correlations and perform regression analysis).

    A different kind of question would be to apply principal components analysis (PCA) to a multivariate data set and then use either regression or cluster analysis to the result to identify interesting patterns or relationships in the data.

    These are examples, and you should feel free to discuss your data with me, by email or in my office hours, as you explore it, graph it, and think about what could be cool to do with it.

    **Don't be overly ambitious.** This is a one-semester class. Don't make your statistical questions too complicated. The point is to practice using some of the techniques we've learned. Keep your questions at a level where you can answer them in a week or two.

4. Using the basic logic of hypothesis testing, choose appropriate statistical tests and use them to help answer the questions you came up with in step (3).

5. Write a report about your data, how you analyzed it, and what you learned.

    - You should use the features of Quarto to include figures and tables, and use R to insert numbers into the text as necessary.
    - You can use author-date format for citing publications, or if you're adventurous you can use Quarto's tools for automatically managing citations and references (see https://r4ds.hadley.nz/quarto.html#visual-editor for an example)
    - You will turn in your final report by rendering the Quarto document to HTML, PDF, or Word format and pushing it to GitHub.
        - I have created a template for you on GitHub Classroom at https://classroom.github.com/a/pGhugW_l

    There isn't a definite length for the report. I expect that the report will be somewhere in the range of 10–20 pages, but if you put a lot of figures and tables into your report, it could go longer.

    I will not grade on length, but on how clearly you present your work (writing, clear use of graphics, etc.), finding interesting questions to ask and interesting things to say about your data, and effectively using the statistical methods we have studied, which are appropriate to your questions.

## Timeline:

- By **Thursday March 27**, submit a short description of the data set you want to work with for your project. This is not a formal thing, like a paper, but send me a page or two telling me what the data set is (what kind of data it has, where the data comes from, what variables are measured, roughly how big the data set is (how many rows of data in a spreadsheet or a similar measure of how many observations it includes))

  - I expect that the specific research quesitons you will be asking will develop as you play with the data, but for this initial assignment, give me an idea of what kinds of questions you might want to ask about this data.
  - I encourage you to come to my office hours or send me email to ask questions or discuss what you are interested in and what you might want to do with data.

    It's fine to say, "Here's the kind of thing I'm interested in. Can you help me find a good data set to work with?" or "Here's a data set I'm interested in. Can you help me come up with some good questions to ask about it, using statistics?"

- The final project is due **Tuesday April 22** before midnight.

## Resources for finding environmental data

- A primary source of data on the world climate is the National Oceanic and Atmospheric Administration (NOAA), which has a Climate Data Online page (https://www.ncei.noaa.gov/cdo-web/) that can help you identify and download data.
  - There is also a more general Data Access page (https://www.ncei.noaa.gov/access/search/index) that can help you explore even more data sets.
  - If you're interested in data on hurricanes, there is a page on hurricane data sets at https://www.nhc.noaa.gov/data/ and the "Best Track" (HURDAT2) data set, which lists data on the full track of each named tropical storm from 1851–2023 may be especially rich. If you want to dig into hurricane data, I would also recommend the book, *Hurricane Climatology* by James B. Elsner and Thomas H. Jagger (Oxford University Press, 2013), which includes R code for all sorts of statistical analysis of hurricanes. You can access this book online through the VU libraries.
- Records of carbon dioxide and other greenhouse gases in the atmosphere can be downloaded from the Scripps $CO_2$ Program at Scripps Institution of Oceanography at https://scrippsco2.ucsd.edu/. Measurements include measurements of multiple gases taken at sites around the world, going back as far as 1958. Some data also has been merged with records from glacial ice cores to estimate $CO_2$ concentrations going back thousands of years. Some data also includes analysis of different isotopes of carbon and oxygen in atmospheric $CO_2$, which can give evidence for the source (biological versus volcanic) and age (e.g., burning wood versus fossil fuels) responsible for the changing $CO_2$ concentration of the atmosphere.
- The U.S. Geological Survey collects all sorts of great geological data, which you can access through its data portal https://www.usgs.gov/tools/geo-data-portal. The USGS also actively develops many tools for analyzing geological and hydrological data in R, including the `dataRetrieval` package, which lets you automate retrieving data using R. USGS also has a searchable catalog of all its data products at https://www.usgs.gov/products/data/all-data
- NOAA's National Centers for Environmental Information https://www.ngdc.noaa.gov/ also hosts data on natural hazards https://www.ngdc.noaa.gov/hazard/hazards.shtml, including volcanoes, earthquakes, wildfires, tsunamis, and storms.
- The Smithsonian Institution's Volcanos of the World databases https://volcano.si.edu/gvp_votw.cfm includes an extensive collection of volcanic activity spanning the entire Holocene.
- Columbia University has built an extensive archive of climate data of all sorts at the International Research Institute for Climate and Society/Lamont-Doherty Earth Observatory (IRI/LDEO) Climate Data Library https://iridl.ldeo.columbia.edu/index.html?Set-Language=en
- The R Open Science project has produced hundreds of R packages for acquiring and analyzing all sorts of scientific data. It is worth browsing or searching their list of data access packages for downloading sicentific data from public sources on the web https://ropensci.org/packages/data-access/