Project 1: Present a Data Visualization

Agreement

This Project is meant to be an assessment of your ability to creatively execute several learning objectives at once by visualizing information contained in a data set of your choice.

While you will be scored individually, you are welcome to work with other students in the course to develop your project.

You may use any resource, either online or physical, to complete the work. This includes:

- Any help forum or website (e.g. StackOverflow) questions that already exist.
- Any notes, code, slides, papers, or previous feedback from the instructor.
- Any books, online or physical.
- Scholarly works such as papers.
- Help from generative artificial intelligence such as ChatGPT.

It DOES NOT include:

- Help from homework websites such as Course Hero or Chegg.
- Help from students or persons outside of those currently enrolled in the current semester of CPSC 292.

If you use work outside normal course resources (textbooks, lecture notes, slides, code, or instructor feedback), you are expected to cite the work by providing a URL to the source near the place that the code was used. If generative AI is used, you must include a section within the main documentation of the project detailing how and why you used generative AI including a link to the specific conversation(s) used to generate project content. **Failure to include this section or disclose the use of genAI will constitute unauthorized AI use and is subject to sanction as an academic integrity violation.**

Instructions

- **Step 1:** Choose a data set from the internet. This data set must have some numerical data associated with it, must be freely available to download, and must be approved by the instructor before **Oct. 3, 2025 at 5 pm**.
- **Step 2:** Create *two* visualizations of your data set that illustrate some hidden information or insight into the data set. Make sure that the visualizations are easy for your viewers to understand, follow best practices, and are in a reproducible RStudio project. Submit a compressed folder (.zip file) with your project to Canvas by the deadline **Oct. 10, 2025 at 9 am**. You will receive feedback from the instructor on improvements to your project.
- **Step 3:** Present your visualizations to the class and receive feedback on how to improve the visualizations. This will occur during class on **Oct. 17 or 20, 2025** (you will be assigned a day, but still must have completed your visualization by the deadline in Step 2).
- **Step 4:** Incorporate the feedback to your visualizations and submit a revised version of your visualization project by **Oct. 24, 2025 at 5 pm**. You will submit a compressed folder as a zip file to Canvas.

Additional Information

Selection of Data Set

Project 1 assesses your skills at creating visualizations based on real-world data sets. In order to smooth the process of loading and cleaning data sets, the instructor must approve your data set before the deadline listed in Step 1 above. Please allow at least one business day before the deadline for the instructor to review the data set and provide an approval or feedback.

You may use any data set on any topic, as long as it meets the following guidelines:

- The data set is freely available online under open source rules. Data sets that are behind paywalls, require a request to access, or are otherwise restricted are not eligible. Exceptions to this policy are if the set is a part of a research project that you are gathering yourself or is a data set from a professor at Chapman. Sites such as Kaggle that require a free account are OK.
- The data set must exist, or easily covert to, a text-based, tabular format such as CSV or TSV file. Excel spreadsheets and other proprietary formats are heavily discouraged.
- The data set cannot be a part of an existing R package (such as `datasets` or `ggplot2`).
- The data set has at least one continuous data type and another data type (e.g., categorical). Data sets containing only categorical data are very difficult to visualize, therefore you will be asked to choose a different set.
- The data set should be complex and/or large enough to use for Projects 2-4. The instructor may request a different data set because it has too few columns or data types which will make progressing to the next project difficult.

There are many places to find data sets on the internet. Some popular options are:

- Kaggle: a popular site that contains thousands of free data sets. Requires signing up for the website (free) to access.
- Data.gov: Open data sources from the US Government.
- CDC Open: List of resources containing data sets maintained by the US Centers for Disease Control (CDC).
- Data Rescue Tracker: backups of public data sets on from a variety of US government agencies. Click "backups list" to access data set list.
- Tableau: Data analysis site that has several free, public data sets.
- NY Open: Data sets from the New York Times.
- FiveThirtyEight: now defunct website that includes sets on political polls and sports teams. Data sets are available through 2023.
- Awesome Public Datasets (github): list of public data sets on a large variety of topics.

Submit your data set for approval by sending the URL as a slack to the instructor.

Required Components for the Project

It also assesses your ability to create an R Project that contains analysis and visualizations that are replicable, in other words, the figures and analysis can be replicated on other machines (most importantly, the instructor's!).

In order to do this, the project must be an organized, self-contained R Project in a single folder which will be compressed and submitted as a zip file. Within the folder should be:

- **Rproj file**: This file sets up the RStudio environment for the project and should be included.
- **Data set**: a separate file as a CSV or other text file of appropriate extension. Do not include XLSX or other file types of spreadsheet or database files. Data set files should be simple tables that start in the upper left of a spreadsheet program with data organized in columns and rows. They should include no extraneous text or information within the file.
- **RMD file**: this file containing the code for loading and cleaning the data set, creating the visualization, and any documentation. Code and regular text should be separated out into text areas and code chunks. RMDs should knit into either an HTML or DOC file successfully to get full credit for the project. The RMD file should contain:
  - A file **header** that includes a title describing the project based on the data set (do NOT put just "Project 1", describe the project!), your name, your course section (either 01 or 02), and the submission date. Note that this header needs to be formatted correctly, or the RMD will not knit!
  - A **setup code chunk** that contain library calls for any required packages to run the code.
  - A section titled **Loading and Cleaning Data** which should have a code chunk that loads the data set from the project and text descriptions of data cleaning steps taken.
  - A section titled **Visualizations** which contains both code that produces the data visualizations you have created and text description of what the visualizations tell you about the data.

  - A section titled **References** in which you list all external resources used to produce your project other than GenAI, including articles, Stack Overflow posts, books, etc.
  - (After the first submission) A section titled **Responses to Feedback** where you outline in text what were the major feedback received from the instructor and peers and how you changed the project to address the feedback.
  - (Optional) A section titled **Generative AI Reflection**. If you have used GenAI in creating your project, you must include a reflection (text only) which covers the following: which GenAI resource was used (Chat GPT, CoPilot, etc), how the resource was used to produce the code and how that code was integrated into your project, any complications you experienced using the resource, and a link to the conversation(s) used to generate code used in the project.

How Project 1 will be graded

As with all work in the course, Project 1 will be assessed for completeness, in other words, you will receive a complete (2 awarded course points) or incomplete (0 awarded course points). Completion will be scored on:

- Presence and completeness of each of the project items outlined above. Each file and section of the RMD file should be present to receive a complete.
- Meaningful improvement of the visualizations in response to feedback. There needs to be evidence that at least two items of feedback from the instructor and three items of feedback from peers were incorporated into the project. Please spell this out, the instructor can't keep track of everyone's changes!
- Successful replication of the Project by the instructor. The replication process is as follows:
  1. The zip file is downloaded and uncompressed.
  2. The RStudio project is opened using the RProj file.
  3. The RMD is opened and "Knit" button is pressed.

If the RMD fails to knit, or if there are errors produced in the code, the project will receive an incomplete.