Project 3 Instructions

Please use the project template Quarto document to complete your project. The rendered document (as a pdf) *and* the raw Quarto file (as .qmd) must be submitted to Canvas by 11:00pm on Thurs., **Apr. 17, 2025.** These two documents will be graded jointly, so they must be consistent (as in, don't change the Quarto file without also updating the pdf document!).

All results presented *must* have corresponding code, and the code should be visible in the final generated pdf for ease of grading. **Any answers/results given without the corresponding R code that generated the result will be considered absent.** All code reported in your final project document should work properly. Please do not include any extraneous code or code which produces error messages. (Code which produces warnings is acceptable, as long as you understand what the warnings mean and explain this.)

For this project, you will be working with a dataset of your own choosing. **Important:** The dataset needs to be picked from the TidyTuesday project (link), and it needs to be one that has been released between May 28, 2024 and December 31, 2024 (both dates inclusive).

The project structure will be similar to Project 2. However, this time you will choose your own dataset and your own question, that you will then answer. Also, you will have to do some data analysis in addition to data visualization. The final project should be structured as follows:

- Questions (1 specific question you will answer)
- Introduction (1–2 paragraphs)
- Approach (2–3 paragraphs)
- Analysis (2–4 code blocks, 2 figures total, text/code comments as needed)
- Discussion (1–3 paragraphs)

We encourage you to be concise. A paragraph should typically not be longer than 5 sentences. (But don't count sentences, if you need a few more that's fine!)

Instructions

First state the question you will answer. Project 3 will have a single, broad question that requires more than one visualization to answer. The question should be conceptual and open-ended and not prompt a specific analysis.

In the Introduction section, write a brief introduction to the dataset, the questions, and what parts of the dataset are necessary to answer the questions. You may repeat some of the information about the dataset provided above, paraphrasing on your own terms. Imagine that your project is a standalone document and the grader has no prior knowledge of the dataset. You do not need to describe variables that are never used in your analysis.

In the Approach section, describe what type of data analysis you will perform and what kind of plots you will generate to address your questions. For each plot, provide a clear explanation as to

why this plot (e.g. boxplot, barplot, histogram, etc.) is best for providing the information you are asking about. (You can draw on the materials provided here for guidance.)

In the Analysis section, provide the code that performs required data wrangling and then generates your plots. You may find it helpful to compute and output summary tables in addition to making plots. Use scale functions to provide nice axis labels and guides. Also, use theme functions to customize the appearance of your plot. **For full points, you will have to apply some unique styling to your plots;** you cannot rely exclusively on preexisting theme functions. All plots must be made with ggplot2. Do not use base R plotting functions.

In the Discussion section, interpret the results of your analysis. Identify any trends revealed (or not revealed) by the plots. Speculate about why the data looks the way it does.

Statistical modeling requirement: Your analysis must use one of the following: (i) statistical modeling on subsets of data, as discussed on Mar 13, 2025; (ii) principal components analysis, as discussed on Mar 27, 2025; or (iii) clustering, as discussed on Apr 3, 2025.

Visualization requirement: You have to use faceting or color mapping in at least one figure. You have to use at least two different geoms across all your figures. You have to make at least one compound figure that consists of two separate pieces, as discussed on Mar 4; note that this is separate from the faceting or color mapping requirement.