Final Project

In lieu of a final exam, you are required to complete a final project report. The report will be submitted to Canvas by the final exam day, March 17th, end of day.

You are encouraged to work with a partner ont his project but are not required to. Group sizes larger than 2 are not allowed.

Course project guidelines

Your assignment for the course project is to formulate and answer a question of your choosing based on one of the following datasets:

1. Fertility and infant mortality data: births and infant deaths in the U.S. 2010-present across some demographic subgroups.
2. ClimateWatch historical emissions data: greenhouse gas emissions by U.S. state 1990-present
3. World Happiness Report 2023: indices related to happiness and wellbeing by country 2008-present
4. A dataset of your choosing: must get instructor approval (request approval via email).

You can combine these datasets with other external datasets if it helps you answer you question. Please cite all sources.

A good question is one that you want to answer. It should be a question with contextual meaning, not a purely technical matter. It should be clear enough to answer, but not so specific or narrow that your analysis is a single line of code. It should require you to do some nontrivial exploratory analysis, descriptive analysis, and possibly some statistical modeling. You aren't required to use any specific methods, but it should take a bit of work to answer the question. There may be multiple answers or approaches to contrast based on different ways of interpreting the question or different ways of analyzing the data. If your question is answerable in under 15 minutes, or your answer only takes a few sentences to explain, the question probably isn't nuanced enough.

Deliverable

Prepare and submit a pdf and corresponding quarto document that summarizes your work. Your quarto document should contain the following sections/contents:

Report requirements

- Your report must be at least 8 pages and no more than 20 pages (excluding any appendices).
- No more than 2 of your total pages can consist of figures (i.e. you are allowed roughly 4 half-page figures). You can include appendices for as many extra figures tables that you woudl like if needed (no length limit). Your report must include *at least* two figures.
- **Important:** Unlike homework assignments, no raw R output should be included in your report. Your report should include nicely formatted figures *and tables!* Reports that include extensive raw R output will receive deductions. You may want to try the gt package or related packages for formatting any relevant output.

Required sections

- **Introduction**: write up a short summary of the dataset you chose to work and the relevant context. You can follow the conventions introduced in previous assignments. Cover the sampling if applicable and data semantics, but focus on providing high-level context and not technical details; don't report preprocessing steps or describe tabular layouts, etc.
- **Question of interest**: motivate and formulate your question; explain what a satisfactory answer might look like.
- **Data analysis**: provide a walkthrough with commentary of the steps you took to investigate and answer the question. You should try to focus on presenting the analysis clearly and succinctly.
- **Summary of findings an discussion**: answer your question by interpreting the results of your analysis, referring back as appropriate. Include a discussion of some limitations and future directions.

Evaluation

Your work will be evaluated on the following criteria:

1. Thoughtfulness: does your question reflect some thoughtful consideration of the dataset and its nuances, or is it more superficial?

2. Thoroughness: is your analysis an end-to-end exploration, or are there a lot of loose ends or unexplained choices?
3. Mistakes or oversights: is your work free from obvious errors or omissions, or are there mistakes and things you've overlooked?
4. Clarity of write-up: is your report well-organized with commented codes and clear writing, no raw R output, or does it require substantial effort to follow?