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. ClimateWatch historical emissions data: greenhouse gas emissions by U.S. state 1990-present
2. World Happiness Report 2023: indices related to happiness and wellbeing by country 2008-present
3. Any dataset from the class assignments or mini projects

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 jupyter notebook that summarizes your work. Your notebook should contain the following sections/contents:

- **Data description**: write up a short summary of the dataset you chose to work with following 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. This section can and should include code cells and text cells, but you should try to focus on presenting the analysis clearly by organizing cells according to the high-level steps in your analysis so that it is easy to skim. For example, if you fit a regression model, include formulating the explanatory variable matrix

and response, fitting the model, extracting coefficients, and perhaps even visualization all in one cell; don't separate these into 5-6 substeps.

- **Summary of findings**: answer your question by interpreting the results of your analysis, referring back as appropriate. This can be a short paragraph or a bulleted list.

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, or does it require substantial effort to follow?