Data Science : final project outline

Your final data project has these core pieces:

- Your work due Monday, December 16, at midnight.
- Final project includes two parts:
  - A written report summarizing your project
  - A brief video presentation (5 slide max)
- Due by end of finals: Five short responses to videos.
- The written report determines the bulk of your grade.

Complexity	Writing	EDA	Code/citations	Testing	Video
15%	20 %	15 %	10%	15 %	20%

Core elements

- **Data context:** An explanation of the context of your data: what is happening that means you're interested in this data? What background does the reader need to understand your analysis? This includes an explanation of your data's contents.
- **What are your questions?** Make clear to the reader what questions you might ask of the data and how those questions connected to your explorations of the data.
- **Visualizations:** Show the reader what is happening within the data through a series of visualizations that connect to your core questions.
- **Probabilistic reasoning:** As much as the data allows, explain hypotheses and perspectives through the lens of probability. This may include visualizations, calculating probabilities by hand, or simualtion results.
- **Testing:** At least two **relevant** examples where you translate a question into a statement about parameters underlying the data (e.g. $\mu = 0$). This include executing the test and finding the result.

A word on testing...

While testing is a required part of the project, that doesn't mean you have to do a classical hypothesis test.

Confidence intervals, bootstrap tests, or other simulation-based approaches are great for this as well. However, in a few places you will have to make clear that you can calculate a p-value, understand the probabilistic construction of it, and how one can make decisions using them.

Style guide

Style counts!

- Must be in Markdown (or better: Quarto, LaTeX)
- Well-written paragraphs:
  - Each paragraph should be a clear set of grouped sentences, with minimal repetition. This should discuss some point necessary to your broader argument. No filler!
- Communication should be through text.
  - Put the appropriate content into the document, not the code. If the code provides values (e.g. ‘cat‘), make sure that is used sparingly.
- Integrated figures and code.
  - Your goal is to lead the reader through the data; your code and figures should be in the service of that, not a distraction.
- Code commented.
- Citations as needed [link for citation help].
- Choose a style and stick with it.
- Appendix as appropriate.

This goes for your problem sets as well!

- From now on, I'll grade 20% of your problem sets on style.

...

There is no fixed length for the project. Quality over quantity.

- That said, I'd be surprised if a well-done project took fewer than 10 pages. But I've been wrong before.

- **Complexity:** Data sets vary wildly in complexity. This section rates how much effort your explication of the data required: if it is a relatively tame data set, I expect that you've really covered all the reasonable perspectives on the data; if it is more wild, I'm looking for a thoroughness commensurate with the difficulty of the data. More effort on more complex data that yields less than a simple analysis of simple data is still more valuable. (That said, complete is better than complex.)

- **Writing:** This is simply that your writing is clear, legible, proof-read, and coherent. Stylistically, I have no requirments but the text should scan: an interested reader should be able to clearly follow the progression of the work from the beginning, with context explained, through the EDA, analysis, testing, and conclusions.

- **EDA:** This section covers the quality, care, and sophistication of the figures, including the exploratory data analysis. We've done a lot with visualization: I expect figures with clear axes, appropriately sized and colored bars and lines, and the ability to understand by eye what is going on with each. The figure should be located in the text so that a reader knows where the corresponding text lies.

- **Citations/code:** The code (and your citations) should be clear and readable. This means comments as appropriate and For citations, I don't have a style; pick one and stick to it.

- **Testing:** This counts the hypothesis testing, confidence intervals, bootstrap tests or other elements of probabilistic reasoning that are present in your project. I'm looking to see that tests or their alternatives are used in an appropriate, balanced way that provides the reader with a clear sense of the uncertainties (or not) throughout the project.

- **Video:** Completion of the video and review of other students' projects.