Methods of Data Science I

Course project

In this project, you'll practice and synthesize your data science skills. The project is scaffolded into 6 stages, with due dates throughout the term. The project is progressive and iterative: you'll get feedback from me after each stage, and be able to rework previous stages if necessary as you move forward.

Proposal

What to submit: The proposal document, in the text of an email or as an attached plaintext, HTML, or PDF file

Your proposal should formulate a **research question** for your project and identify a **dataset** that you'll analyze to address your research question.

Your research question can be a more formal, academic question — like something from your home discipline — or a more causal, "data science-y" question — like you might examine in a hackathon, data science challenge, or detailed blog post.

The dataset should be relatively large and complex, and **not data that you have collected yourself or previously analyzed**. Part of the work of researching the data journey and exploring data analysis is familiarizing yourself with found or opportunistic data. It's okay if you're using the dataset for a project in another class or your lab group this semester, so long as you're asking a very different research question.

Because you're working with opportunistic data, it might be easier to find an interesting dataset first, then develop your research question.

If you're having trouble finding interesting datasets, try the TidyTuesday project or the Kaggle dataset library.

The **proposal document** should be 3 paragraphs long, with 75-150 words per paragraph. The paragraphs should be labeled "Intellectual Merit," "Broader Impacts," and "Data and Methods," in that order.

Intellectual Merit
What is your primary research question? Why is this question interesting and worth investigating, from an academic or intellectual perspective?

Broader Impacts
Why is your research question socially valuable? Why should the general public care about the answer to this question, and pay someone to investigate it?

Data and Methods

What dataset will you use to investigate your question? Why does this dataset seem appropriate for your question? What analytical methods will you apply? Because it's not the focus of our class, you don't have to use any fancy statistical methods (for whatever "fancy" means to you). You can even focus on exploratory methods, such as descriptive summaries and visualization.

Data journey narrative

What to submit: A document, by email, preferably as an attached plaintext, HTML, or PDF file.

Characterize the journey that your data took to get to you. Some relevant questions to answer might include

- Who generated these data? Why? How?
- What measurement instruments were used to generate these data? What infrastructure has been used to store and transmit the data?
- What was the original intended use of these data? What are some other ways the data have been used?
- How has the mutability/immutability of the data been important to its journey & use? How has its mutability/immutability been created and maintained?
- How will you need to transform the data further for your project? How do the physical and material properties of the data facilitate this, and/or create friction?
- What values have been crystallized in the data? How have power relations and values shaped what is/isn't included in the data?
- Based on these factors, do the data appear to be fit for purpose? That is, based on what you know so far, do the data appear to be appropriate for answering your research question?

Include references for the sources you use to answer these questions. You can work on this step of the project at the same time as the EDA step, but I will ask you to turn in the data journey narrative first. A good target length for your narrative is 500-1,000 words long, not counting any references.

Exploratory data analysis

What to submit: A link to your GitHub repo (with access if it's private). The EDA itself should have two analysis files: an `R` or `Rmd` script, and an `html` file generated from the script using knitr/Rmarkdown/Quarto. If your code requires anything more than simply running the script, include instructions in a README. *Your code from this step will go through peer review in the next step.*

Your exploratory analysis should cover the most relevant elements from the EDA checklists to validate your data. Be sure to identify and address limitations of your data for answering your research question.

Use a literate programming style, mixing code with expository text. Julia Silge has several good examples of blog tutorials written in this style:

- Sliding windows for #TidyTuesday rents in San Francisco

- Reordering and facetting for ggplot2

If your EDA concludes that these data are not fit for purpose, it's okay to continue forward with your project. Your report below should focus on explaining why your data are not fit for purpose, and then discuss how better data might be collected.

Code review and reproducibility check

What to submit: You'll submit your code review by adding comments to the author's code and then submitting a pull request. See the instructions below.

> **Code review step-by-step instructions**

As in the code review lab, your code review should cover the most relevant checklist elements. Specifically, be sure to address:

- Is it clear what steps you need to take to run the script?
- When you follow those steps, does the script run?
- Is the html file from the repository reproducible?
- How easy or difficult is it to read the script and understand what the author is doing and why?

(Reproducible) report

What to submit: A PDF of your report, a link or invitation to your GitHub repo with automatically reproducible code and data

The report for your analysis should structured somewhat like a traditional scientific paper: introduction, methods, results, discussion. But the content of these sections will be somewhat non-traditional.

The total length of this report should be 3,000-5,000 words, not including references.

As a stretch goal, your report should be **fully reproducible**. Using `renv` and a Makefile, I should be able to clone your repository, run `renv::restore()` and then `make`, and exactly reproduce the PDF of your report.

Introduction
As usual, motivate your topic, give some background, and clearly state the research question.

In addition, incorporate answers to the reflexivity questions: what did you already know about this topic before starting to work on the project, what did you expect to find, who's impacted by this topic, and how will your work respect them?

Methods
The methods section should focus on the data, and in particular should incorporate your data journey narrative.

Results

As usual, your findings should be framed as phenomena (in the sense of the Brown chapter) rather than causal claims, mechanisms, or theoretical results.

The first subsection of your findings should incorporate your EDA. The second subsection can address your research questions. Use visualizations along with or instead of tables. Keep your visualizations close to the data. For example, include both a scatterplot and a fitted regression curve.

Discussion

As usual, briefly (1-2 sentences) recap your major findings, then discuss limitations and directions for future research.

Resist the temptation to treat your findings as conclusive or as vindicating (or challenging) some larger theory. Instead, emphasize the ways future research can improve our understanding of the phenomena: how new data might be collected to mitigate the limitations of this dataset, how different kinds of data might complement the kind of data used here, and how further studies might trace out the scope of the phenomena found in this dataset.

Finally, come back to the reflexivity questions in the introduction. Were your expectations met? (Probably a mix of both yes and no; explain.) What implications might your findings have for people affected by the topic? How should these implications shape the trajectory of future research on this topic?

Flash talk presentation

What to submit: You won't submit anything, but you will give a presentation for the class

At the last class meeting, you'll give a 5-minute presentation of your project. Your target audience is the other students in the class, so you'll need to start by motivating your project and be thoughtful about the use of jargon from your home discipline. Depending on how much time we have, we may or may not have questions after the flash talks.

You should use slides for your presentation, but you don't have to turn them in. Google Slides, PowerPoint, Keynote, or whatever are fine, because you're responsible for arranging to project your slides and checking in advance that everything works as expected in our classroom.

Because flash talks are so short, I strongly recommend writing a script, checking the timing, and then memorizing the script. There *will* be a loud timer at 5 minutes.