Final Project

MAT219 Data Science I

Overview

You will undertake a data science project on a topic of your choice. The project is an opportunity to show off what you've learned about data science. Your task is to use data to tell us something interesting. This project is deliberately open-ended to allow you to fully explore your creativity. Here are the main rules that must be followed:

1. **Your project must be centered around real data that is sufficiently large, complex, and/or messy**. While staying within the scope of what you can handle, the more challenging your data set is, the more you will be able to use the tools learned in this class. For example, one thing that will make your data science project more ambitious is combining two or more data sets that are not directly related.
2. **Your project must tell us something interesting**. An example of a project that doesn't tell us anything would be one where you obtain a single dataset and summarize the variables with some perfunctory tables and visualizations. You want to go beyond this by posing good questions that require you to dig into the dataset (look at subsets of the dataset, create new variables, etc.) to formulate answers. Think about the questions that we asked for the data projects and how you worked with the data to provide answers.
3. **Cleverness and creativity will be rewarded**. Going above and beyond what we did in class will be rewarded.
4. **Collaboration**. Teams can range from 1 to 3 people. All team members are expected to contribute equally to the completion of this assignment and group assessments will be given at its completion - anyone judged to not have sufficiently contributed to the final product will have their grade penalized. While different teams members may have different backgrounds and abilities, it is the responsibility of every team member to understand how and why all code and approaches in the assignment works.

Project Due Dates

- Proposal due: End Of Week 4, July 23rd, 11:59pm
- Final Submission: Final Day Of Class, August 4th, at 1:59pm (Paper and Video Presentation)

Components

Proposal

Your proposal is to be submitted via D2L. Spend time deciding on a topic that interests you. Think about compelling questions that can be answered with data. Here is one way to approach this project: The website FiveThirtyEight used to have an "Ask Mona" column run by Mona Chalabi where people would send her questions, and she would use data to provide an answer. Imagine sending yourself such a question, and then doing some research to find where relevant data can be obtained. Once you have an idea of the data out there, try to picture your end product. What type of visualizations, tables, and/or statistical methods would be helpful. Don't think about coding, a particular dataset, or what you know how to do at first. This will increase the likelihood that you will come up with something ambitious and original, and you will be more motivated to learn new things as you work to accomplish your goal. The topic is completely open to your choice, but keep in mind the rules listed above.

Your proposal should contain the following content:

- **Title**: The title of your project
- **Purpose**: Describe the general topic/phenomenon you want to explore, as well some carefully considered questions that you hope to address. You should make an argument motivating your work. Why should someone be interested in what you are doing? What do you hope people will learn from your project?
- **Data**: As best you can, describe where you will find your data, and what kind of data it is. Where will you be accessing your data? Be as specific as you can, listing URLs and file formats if possible.
- **Variables**: As best you can, list and briefly describe each variable that you plan to incorporate. Be specific about units, scale, etc. where possible.
- **End Product**: Describe what you hope to deliver as a final product. What type of visualizations, tables, and/or output from statistical methods are you aiming for?

Write-Up

Your write-up should be a reproducible Quarto PDF document that when printed is of length no more than 30 pages (including code). In your write-up, you should layout your data analysis work-flow similar to how we did in the midterm project. In particular it should include the following sections first:

1. Introduction
2. Load Packages
3. Load the data
4. Tidy the data
5. Analyze the tidy data

Your write-up should then continue with answering the questions you have posed. This will likely include a few more sections with some additional code. Be sure to address the following:

- **Why should anyone care about this?**
- **What is this about?** Do not assume that your readers have any domain knowledge! The burden of explanation as to what you are talking about is on you! For example, if your project involves phylogenetic trees, do not assume that your audience has anything other than a basic, lay understanding of genetics.
- **Where did your data come from?** What kind of data was it? Is there a link to the data or some other way for the reader to follow up on your work?
- **What are your findings?** What kind of statistical computations (if any) have you done to support those conclusions? Again, while the R code will show you performing the calculation, it is up to you to interpret, in English sentences, the results of these calculations. Do not forget about units, axis labels, etc.
- **What are the limitations of your work?** Be clear so that others do not misinterpret your findings. To what population do your results apply? Do they generalize? How could your study be improved? Suggesting plausible extensions does not weaken your work, it strengthens it by connecting it to future work.

Content Do not present all of the R code that you wrote throughout the process of working on this project. However,

- Include the minimal amount of concisely written R code that produces the computations, visualization, tables, etc. that are used in your write-up.
- If you make a claim, it must be justified by explicit computation. A knowledgeable reviewer should be able to reproduce your analysis without doing more work or writing new code.

Evalution

Overall Your final project will be evaluated based on:

- **Originality/Interest**: Is the topic original, interesting, and substantial, or is it trite, pedantic, and trivial? How much creativity, initiative, and ambition did you demonstrate? Is the basic question driving the project worth investigating, or is it obviously answerable without a data-based study?
- **Degree of Difficulty**: How challenging was the project? Were the data particularly large, complex, and/or messy? Did the data come in an obscure format? Was a challenging visualization constructed? Were any elements from outside the coursework necessary to complete the project?
- **Design**: How well were the graphical elements of the project designed? Were they clunky or elegant? Was a truly original view of the data presented? Were any interactive elements usable?
- **Meaning/Analysis/Statistical Understanding**: Did we learn anything meaningful from this project? Are the chosen analyses appropriate for the variables/relationships under investigation, and are the assumptions underlying these analyses met? Are the analyses carried out correctly? Did you make appropriate conclusions from the analyses, and are these conclusions justified?
- **Write-Up**: How effectively does the write-up communicate the goals, procedures, and results of the study? Are the claims adequately supported? Does the writing style enhance what you are trying to communicate? How well is it edited? Are the statistical claims justified? Are text and analyses effectively interwoven?