LLO8200 Data Science, Summer 2021: Final Projects

General Information

- We will do in-class presentations the last two weeks of class. EVERYONE should come prepared to ask questions of the presenters. The presentation should be planned carefully to take 12-15 minutes (MAX!). I will cut you off at 15 minutes, regardless of where you are in the presentation.
    - If the class does not have questions, I will... so do your classmates a favor and come with questions. I will expect EVERY group member to present - so figure out who is presenting what. No one-person shows here!
    - I will also expect an actual presentation (PPT, Prezi, whatever format you prefer, but it should be engaging for the class).
- Both the presentation and the paper should include:
    - An articulation of the problem in practice (or at least a general intro to the phenomena you studied)
        - This can/should include any references to pertinent literature - I am not talking about a full-blown literature review, but something that grounds your work in reality
    - A clear presentation of research question(s)
    - Descriptive statistics (visuals)
        - univariate, bivariate - introduce your data so that the class knows what you had available to you and what the data "looks" like
    - Analytical process (did you run a regression? maybe an ANOVA? How did you validate your model?)
    - Findings/key takeaways, limitations, and next steps
- After you present in class, I expect that you take feedback from me and your classmates and integrate that into your work products before turning them in.
- When you are ready to submit your final products, one group member should email me the following (CCing their group members) by the due date:
    - An .Rmd file that includes the code/text for your final report. Note, this will likely not include cleaning code.
        - If you want to share your cleaning code, you can give that to me in a separate .Rmd file(s) -- just label it "cleaning."
    - Your final presentation.
    - Your knit final report (see rubric below)
    - Any additional code files you want to include.

Presentation – Rubric

1. Articulation of the Problem in Practice/Relevance of the Study
    a. A strong presentation will clearly articulate why this study is important/relevant/interesting and how it adds to a body of knowledge.
    b. An acceptable presentation will attempt to articulate the importance of the matter at hand but will not properly link it to existing studies or make a compelling argument for the work's importance.
    c. A weak presentation will make no attempt to articulate the problem in practice or will not adequately argue for its importance/relevance.

2. Research questions
    a. A strong presentation will have clearly articulated research questions that stem from the identification of the problem in practice.
    b. An acceptable presentation will clearly articulate research question(s) but will fail to link them to the study rationale or problem in practice.
    c. A weak presentation will not clearly articulate research question(s).

3. Descriptive statistics and visuals
    a. A strong presentation will clearly identify the key variables of interest and include pertinent visuals, tailored to tell the story of the research study, and will not include superfluous information.
    b. An acceptable presentation will include visuals and descriptive statistics but will not use them to tell a cohesive/coherent story about the data.
    c. A weak presentation will not include descriptive statistics/visuals.

4. Data analysis/analyses
    a. A strong data analysis will use two of the three algorithms (conditional means, regression, logistic regression) that we discussed in class. The data analysis will include a measure of model fit and will describe which characteristics are closely related to the outcome. The analysis will include cross-validation, which will be correctly executed and described.
    b. An acceptable data analysis will include two algorithms, but there may be some mistakes or inaccuracies in how the results are presented. A cross-validation will be included but may not be correctly done.
    c. A weak data analysis will not use two algorithms or will use them inappropriately. It will either not include a measure of model fit or will misuse a measure of model fit. It will incorrectly describe relationships with predictors or not describe them at all. It will not include cross-validation, or the cross-validation will be done incorrectly.

5. Findings, limitations, and next steps
    a. A strong conclusion will summarize the key findings, briefly detail limitations, and discuss next steps for this work.
    b. An acceptable conclusion will include two of the three elements (findings, limitations, next steps), but will be missing the third.
    c. A weak conclusion will not adequately summarize the key findings, limitations, or next steps.

Final Paper Rubric

The general components to the final paper are the same as those for the presentation listed above, with the additional elements detailed below. The final paper will be graded on the following areas:

1. Data Analysis
    a. A strong data analysis will use two of the three algorithms (conditional means, regression, logistic regression) that we discussed in class. The data analysis will include a measure of model fit and will describe which characteristics are closely related to the outcome. The analysis will include cross-validation, which will be correctly executed and described.

b. An acceptable data analysis will include two algorithms, but there may be some mistakes or inaccuracies in how the results are presented. A cross-validation will be included but may not be correctly done.
c. A weak data analysis will not use two algorithms or will use them inappropriately. It will either not include a measure of model fit or will misuse a measure of model fit. It will incorrectly describe relationships with predictors or not describe them at all. It will not include cross-validation, or the cross-validation will be done incorrectly.

2. Graphical Presentation
    a. A strong final project will include nicely labeled, easy to understand graphics that describe exactly what is happening with the patterns in the data. The graphics will be complex, showing lots of numbers. The response could include (but does not have to include) interactive graphics.
    b. An acceptable final project will include graphics, but these may not be easy to read or may not be sufficiently detailed.
    c. A weak final project will include graphics that are poorly labeled and do not make much sense.

3. Written Description
    a. A strong final project will include a 1500–2000-word (~six page) description that is easily understandable by an interested layperson. Assume that your audience is your boss or a group who hired you to do this analysis-- not me. It will be much easier to write this if you have a perspective.
    b. An acceptable final project will be written well, but technical details may be poorly described or not described at all, and sentences will be hard to follow.
    c. A weak final project will be poorly written, with many mistakes regarding both the analysis and good writing practices.

4. Organization, Clarity, and Formatting
    a. A strong paper will have an .Rmd file that generates a very nicely formatted document, suitable for professional presentation. What kind of report would you want to give to a supervisor? That is what I want back from you. The organization should be very clear and easy to understand.
    b. An acceptable paper will have some formatting problems and may not look very nice but will still include all the relevant content.
    c. A weak paper will include code chunks, poor formatting, and will just be messy overall.

5. Coding
    a. A strong paper will have code that can generate results from the raw data in an easy-to-understand way. The code will be commented and will run on my computer without me having to tweak it in any way. (Easy test is to knit the document, with all related files in same directory)
    b. An acceptable paper will have code that is relatively clear, but has some problems, and may not be commented in a way that makes sense.
    c. A weak paper will have code that is messy, hard to understand and not commented. It will not run on my computer and cannot be easily debugged.