DS 5110 Project Guidelines

Overview

The purpose of the project is to gain hands-on experience working on real-life data, in a collaborative team. Although a majority of the data management and processing for the project should be done using methods learned in class, you are free to incorporate techniques and tools that have not been covered in lecture.

The project should be done in teams of 3 - 5 class members *from the same class section*. The project is worth a total of 30% of class grade and includes a project proposal, an in-class team presentation, and a final written report. Each team should work independently on their project, but may collaborate on technical questions via Piazza. *Make sure to proofread your proposal, slides, and final report – it should appear polished and professional.*

Projects with a very high degree of similarity with other past or current students' work, or with other pre-existing work (e.g., found online) that is not clearly cited, will be considered plagiarism, and will be treated accordingly.

Data

Your team is responsible for finding a real-life dataset of your choice. You may either use a pre-existing publicly-available dataset, or collect data yourself (e.g., via web scraping or web API). Use of non-public data may be permitted with prior approval. There are no restrictions on the scale of the dataset, but it should be sufficiently complex to allow for an in-depth analysis using the techniques discussed in this class. You should have some specific analytic goals in mind for the dataset you select.

Team members

Please form teams of 3 - 5 class members who will work together on the project. You may browse class members' miniposters on Piazza for ideas of what types of data interest your classmates. You may reach out on Piazza when seeking out other project team members.

You must submit the names of all members of your project team via Canvas before the project proposal is assigned.

Project proposal

The project proposal will be submitted on Canvas as a PDF.

The project proposal should be no more than one page, not including any references or supplementary figures. It does *not* need to contain code or be generated from R Markdown. The proposal should contain the following:

1. **Title:** A descriptive title of the project
2. **Authors:** List all team members' full names
3. **Summary:** 2-3 paragraphs summarizing the problem you wish to solve, including a description of the dataset and project goals, and a very brief, non-technical description of methods.
4. **Proposed plan:** 2-3 paragraphs describing in more detail the methods you will use to solve the problem. These may be processing, visualization, and analytic methods already discussed in class, or it

may be your design for a software tool for working with a particular type of data or solving a common data analysis challenge, etc. Also discuss any anticipated challenges and plans for overcoming them.

5. **Preliminary results:** 1 paragraph describing any preliminary results you have. This is to demonstrate that the project goals are feasible, and should at least show that you are able to load and explore the dataset satisfactorily.
6. **References:** Cite any references used in the proposal, including any sources of data and associated publications. Use a consistent format and numbering scheme.

Project presentation

During the last two weeks of classes, each team will make a short in-class presentation of their project. You may use the project report format (below) as a guideline for formatting the presentation. All team members are expected to speak during the in-class presentation.

The slides should be submitted on Canvas as a PDF before the presentation.

Project report

The project report will be submitted on Canvas as a PDF.

The project report should be no more than eight pages, not including references or appendices. The main text should not include code or raw console output unless relevant for demonstration purposes and does *not* need to be generated from R Markdown. The report should contain the following:

1. **Title:** A descriptive title of the project
2. **Authors:** List all team members' full names
3. **Summary:** Summarize the background of the project (i.e., what is the problem to be solved or the question to be answer), any related work, a description of the data and project goals, and a brief, non-technical description of your methods and results.
4. **Methods:** This should be a technical description of the methods you used in the project (e.g., the data tidying, transformation visualization, and modeling you performed). It should be sufficiently detailed that others could attempt to reproduce your results.
5. **Results:** All projects must demonstrate results on real-life data. Show the most relevant results from your analysis, using appropriate figures or tables. Describe only the most interesting and useful results.
6. **Discussion:** What is the meaning and impact of the results? Who is able to benefit from this project? How can the results be used to make better-informed decisions? What could be improved about the project in future work?
7. **Statement of contributions:** List the full names of the authors and how each team member contributed to the completion of the project.
8. **References:** Cite any references used in the report, including any sources of data and associated publications. Use a consistent format and numbering scheme
9. **Appendix:** Optional. Any supplementary text and material that does not warrant inclusion in the main text. This can include relevant code, additional technical details, or supplementary figures. If included, the main text should cite the additional information that can be found in the appendix where relevant.