STAT 5000 Project Guidelines

STAT 5000

Statistical Methods and Applications I

For this project, you may do one of two things:

(1) conduct an analysis of a dataset using the exploratory data analysis and statistical techniques developed in this class. Specifically, you should ask questions about a dataset, and use statistical techniques to explore them, or

(2) explain and apply some non-trivial statistical method that we have not discussed in class (e.g., Bayesian linear regression, kriging, principal component analysis, etc. If you choose this option, you must give an in-depth explanation of the mathematics involved in the technique, in addition to applying it to some dataset.)

You may work in small groups of at most 3 people. I hope that you choose a topic or data set which is based on your interests; if you are doing research, feel free to use a dataset which is relevant to that. I expect you to describe your results in a short **formal paper** which will be due at the end of the semester. The project grade will be determined on the basis of the accuracy of the statistical analysis and the quality of the paper. Your final paper will be due **Thursday, December 12th at midnight**, and I will not accept anything after that time.

To reiterate, your document should be a formal paper, not a notebook like we use for homework and classwork. To this end, you can use any word processor you like to generate the document, but I recommend using LaTeX either through a local installation or on a site like Overleaf. You *should not* write your final paper in one long Markdown cell in a notebook! If you have been using LaTeX all semester, it should not be too difficult for you to write a formal paper using it! Additionally, it makes displaying figures and citing references much easier. I am happy to help you with this process.

Note: Many research questions involve more sophisticated techniques than we will cover in this course. For example, many datasets have variables that have complicated relationships to one another, and will require multivariate techniques like regression, generalized linear models, etc. These multivariate techniques are taught in the follow-up course, STAT 5010. *However, there are many datasets that either (1) we can learn from using the techniques in this course (exploratory data analysis, hypothesis tests, confidence intervals, bootstrapping), or (2) we can explore using the techniques we know as a starting point, and note that more sophisticated techniques would be required to fully answer the research question of interest.*

Grading

The project grade will be determined on the basis of the quality of the paper and the statistical analysis therein. Your R or Python code that was used to generate the statistical analysis will be graded out of 10 points. In total, your grade will be out of 50 points, divided in the following way:

- **20 points** will be related to written communication; you will receive a grade of $0-5$ in each of the following categories: (1) context and purpose of writing; (2) content development, (3) sources and evidence, and (4) control of syntax, mechanics, and research article conventions.

- **20 points** will be related to problem solving. You will receive a grade of $0-5$ for how well you (1) define the problem/question of interest, and (2) propose solutions to or answer the question. You will receive a grade of $0-10$ on implementing statistical solutions to answer your questions.

- **10 points** will be related to your codebase that was used to perform the statistical analyses in your paper. Include your code in a notebook as a separate file. Make sure that the code can be executed from top to bottom.

Sections and Topics for the Paper

Your finished paper probably should be at least 5 pages in length, given that you may have figures and explanatory mathematical calculations, but this is not a set rule. I suggest that you include the following sections and topics in your paper. If you are not analyzing a specific dataset (e.g., if you are explaining a modeling method), you should be able to adjust these sections accordingly.

(1) **Introduction and Background**

- Motivating the problem: what is the "research" question, and why is it interesting or worth answering?
- What is the relevant background information for readers to understand your project? Assume that your audience is not an expert in the application field.
- Is there any prior research on your topic that might be helpful for the audience?
- What is the source of the data? Is this an experiment or observational study? Who collected the data? Why was the data collected (if you weren't the one doing the collecting)?

(2) **Methods and Results**

- Describe your exploratory data analysis methods. What needed to be done to the dataset to make it amenable to analysis?
- What analyses are most appropriate to answer the question of interest?
- Describe the analyses used, and check your assumptions.
- Present relevant graphics and interpret results.
- Explicitly connect your technical results to your research questions.

(3) **Conclusions**

- What are your conclusions? What did you learn?
- How would you extend this research? What future research ideas come to mind based on your results and experience with this analysis?

Draft Checkpoint (optional)

In order to give you some feedback before the final due date of **Thursday, December 12th at midnight**, you may turn in drafts to the appropriate Canvas assignments on **Friday November 15th at midnight**. For the draft, you may want to treat it like a "project proposal", and give me the following information instead of a formal paper:

(1) What are the data that you plan to work with?

(2) Where did the data come from? Are they experimental or observational?

(3) Why is this data interesting to you? What questions do you hope to answer about it?

(4) What are the relationships between the variables? Does this theory suggest that they are related in some way?

(5) What random components are present (e.g., measurement error)?

(6) What prior research on your topic might be helpful to consider?

(7) What methods might be useful in analyzing this data?