STA 141C - Big Data & High Performance Statistical Computing **Winter 2023**

Final Project

Lecturer: Bo Y.-C. Ning

Disclaimer: *My notes may contain errors, distribution outside this class is allowed only with the permission of the Instructor.*

1 Description

You should work in a group of 3-5 people (no more than 5). The purpose of the project is to provide you with real data science experience, which includes:

- Posing questions
- Finding data sources
- Exploring your data
- Statistical analysis
- Optimize your code
- Summary and visualize the result
- Working collaboratively with a team
- Presenting your findings through writing and speaking

2 Getting Started

I recommend you get started by brainstorming project topics or themes that interest your group. To narrow down your project to just one topic, think about:

- What questions does your topic address or what problems does your topic solve? Why and to whom are these meaningful?
- What's challenging about your topic?
- Are there credible, public datasets available to explore the topic?
- See below for some suggested data sources.
- Is a 5∼6-week project long enough to explore the topic reasonably well?

Make sure that everyone in the group agrees on the topic. You will need a certain curiosity about your topic in order to stay motivated throughout the quarter.

Once you've selected a project topic, you can start working on the project proposal and the project itself.

The final project has four components:

Name	Due
Proposal	Feb 17 11:59pm
Presentation	Week 10⋆
Report	Finals Week, March 23 11:59pm
Group Evaluation	Finals Week, March 24 11:59am

⋆ Final presentation: March 14 for groups with odd number and March 16 for groups with even number. Each group has 8-10 mins

Final presentation holds in person/on Zoom?.

3 Proposal

Your group should submit a 1-2 page project proposal (1 per group) by Feb 17 11:59pm. Your proposal should address:

- What's the topic of your project? What question(s) will you attempt to answer or what problems will you attempt to solve? Why and to whom are these meaningful?
- What data source(s) will your team use? Briefly describe each data source and Provide a link for each data source. This is a check to make sure that there is actually data available for your topic. If you ultimately decide not to use some of the data sources, or find additional data sources later, that's okay.
- What statistical methods will your team use? Make sure the method you choose is not too complicated and you are capable of writing the code
- What makes your project challenging? Consider that you will have ∼ 5 weeks to work on the project. Do not pick a project that is too hard!

The proposal is your best opportunity to get feedback on your project. Make sure it's clear and addresses the questions above. You can also use the proposal to tell us about any other comments or concerns you have about your project topic. You do not need to present any data analyses in the proposal.

The proposal will be graded satisfactory/unsatisfactory (5 points). Your priority should be working on the project itself; don't spend more than a few hours working on the proposal. Make sure all group members have read the proposal and agree on what it says.

Submit your proposal on Canvas. Each group only need to submit one file. The proposal should be .pdf (consider it as writing a statistical paper, we do not accept ".doc" file!)

4 Presentation

In week 10, each group will present preliminary results from their project to the class. Each group can either elect a leader to present results or choose to let everyone in the team to present. More details about presentations will be released in week 8.

5 Report (grading criteria)

The final report is due in finals week. There is no page limit for the final report (∼ 10-page seems a reasonable range). But the final report is graded based on quality not quantity!

The final project should be a ".pdf" file and its format should be similar to a statistical paper.

The report should at least contain the following components:

- Cover letter (3 points): **summarize** the main findings of the report and **highlight** how you applied the techniques/methods you learned from the class to complete final project. The format of the cover letter can be similar to the cover letter for journal submission.
- Introduction (4 points): describe your problem, introducing dataset and the problem(s) you try to solve
- Proposed method (5 points): describe the method you are using: providing necessary details. If you write equations, define every mathematical symbol
- data analysis study (5 points): you can run simulation to check if your code and/or the method works; this is also the place you test the method if the model assumption(s) do(es) not hold. You should make plots and tables to summarize your finding based on a real dataset; check model assumptions
- Conclusion or summary (3 points)
- Reference and acknowledgement (1 point):
- Appendix/Code (5 points). This is the place you put technical stuffs and code. You can put mathematical proofs (if you have) into this section. You should put your code in this section. Please follow the same principal as you do in your homework. **Important:** At the beginning of your code, please describe how do you optimize your code using the methods you learned from the class.

To be clear is more important than to be technical, make sure even your friends who do not take the class can understand your writing! Putting a lot of jargons and equations without explaining them is bad!

6 Potential projects and data source

- Any project your team wants to work with
- Replicate a paper result and apply to a different dataset. You can find a paper on arXiv `https://arxiv.org/archive/stat`
- Kaggle competition `https://www.kaggle.com/`
- Yahoo dataset `https://webscope.sandbox.yahoo.com/`
- Yahoo finance dataset `https://finance.yahoo.com/quotes/OCR,dataset/view/v1/`
- Implement one of the algorithm described in this course, and try to scale it to large datasets.