CS 418: Introduction to Data Science (Fall, 2023)
Final Project Specifications

The goal of this project is to have an opportunity to develop a data science project that covers the entire life cycle of data science. By the end of the project, you will build a data-driven software that helps users analyze and visualize a set of data while discovering a set of previously unseen observations.

Important notes:

Github: The students need to create the github repository for your project. One person per team should be designated as admin and they can create a private github repository for your project where all team members can contribute and all progress can be tracked. All team members should have student github accounts and be added to the repository before the proposal is due. The github student developer pack has many advantages over a regular free github account (https://education.github.com/pack). You can make your repositories public after the final week of the semester. If you don't have experience with github, take a look at this introduction: https://guides.github.com/activities/hello-world/.

(NO to) Kaggle projects: Submitting a Kaggle competition as your group project is not acceptable. While Kaggle is being a valuable resource, Kaggle competitions are not typical data science projects because a lot of the thinking that goes into a regular data science project has already been done for you and packaged into the competition rules: 1) the problem has been defined, 2) the dataset has been figured out, 3) the framework for evaluation has been figured out. However, you can use a dataset for a different problem. Methods and findings are expected to be different as well.

Contribution: All team members are expected to contribute to the project, and will be graded on their individual efforts in addition to the group outcome (see the "Grading criteria"). The project will consist of five main deliverables and will be evaluated out of 100. The weight of the project towards your final grade is 30%.

Form your group and identify a leader by 15th September, 2023.

- **Five members per group**
- **The groups will be made by only graduate students or only undergraduate students**
- **Add the information about your group HERE: LINK**

1) Proposal (10 points), Due 1st October

The goal of the proposal is to start thinking about the final project. Good presentation skills are important in any field especially for a data scientist. Therefore, we will practice this skill throughout the semester including the proposal.

The format of the proposal is a PPT (or other type of) presentation. The proposal should include four/five presentation slides, converted to PDF:

Slide 1 - Project name and participants: The name of your proposed project, your team name, together with the names, UIC email handles, and github handles of all the team members. Include a link to your github project repository.

Slide 2 - Problem: What is your "big idea"? What is the problem you want to solve, question you want to answer, or decision making you want to support? Why should others care about it? How did you choose this problem? Do you have any specific hypotheses?

Slide 3 - Data: Describe the data. Do you currently have access to this data or do you need to collect it? How much effort is that data collection and can you complete it within a reasonable amount of time? Describe your data in terms of size (e.g., number of rows per table or number of images), type of data, type of features, and any other relevant details.

Slide 4 or (4-5)- Solution and Expected Deliverables/Findings: How do you plan to approach the problem? What is the proposed scope of your project and the next steps? What do you envision the end result to be? What techniques do you think you will use to analyze the data? Do you envision your system to be interactive or static? What do you hope to have achieved for the Progress report? Keep in mind that your direction may change as the course goes on: this is okay and why we are starting so early. Until the progress report, you are allowed to change your goals and discuss your evolving strategies by consulting with me.

Need to submit:
PDF of the slides to Gradescope by 11:59 pm on October 1st. Only one person per group needs to submit, tagging their teammates. No late submissions will be accepted.

Grading criteria:
presentation clarity, aesthetics, whether it includes all information requested.

2) Pitch (10 points), 10th and 12th October

The goal of the pitch is to present the proposal in front of the entire class and develop the presentation skill.

What you need to do for presentation:

- Each person in the group presents one slide (Ideally).

- The total presentation time for each group should be 7 minutes
- Presentation: 6 min, Q&A: 1 min
- Practice before the presentation

Need to submit: Nothing

Grading criteria:
presentation skill (mainly clarity), and timing

3) Progress Report (20 points), 10th November

The progress report is a chance for you to take stock of how far you have come and to reflect on whether or not you are comfortable with the substance or scope of your final project. **The format of the progress report will be a Jupyter notebook** that should be uploaded to the private github repository you have set up for your team. It should Include:

- **Project introduction:** an introduction that discusses the data and related problems that you are investigating.
- **Any changes since the proposal:** a discussion whether your scope has changed since the check-in proposal slides. List the parts that were removed from your plan as well as the parts that were added newly in your plan.
- **Data**: explain how you have prepared your data.
- **Exploratory data analysis:** explain what your data looks like (visualizations are often better). Include any interesting issues or preliminary conclusions you have about your data.
- **At least five visualizations** that shows an interesting hypothesis, along with an explanation about why you thought this was an interesting hypothesis to Investigate. Write the name of the member(s) who is responsible for each of them while explaining it.
- **At least two ML analyses** on your dataset, along with a baseline comparison and an interpretation of the result that you obtain. Write the name of the member(s) who is responsible for it while explaining it.
- **Reflection:** a discussion on the following aspects:
    - What is the most challenging part of the project that you've encountered so far?
    - What are your initial insights?
    - Are there any concrete results you can show at this point? If not, why not?
    - Going forward, what are the current biggest problems you're facing?
    - Do you think you are on track with your project? If not, what parts do you need to dedicate more time to?
    - Given your initial exploration of the data, is it worth proceeding with your project, why? If not, how will you move forward (method, data etc)?
- **Next Step:** Concrete plans and goals for the next month

Need to submit:
A PDF of your Jupyter notebook to Gradescope which includes a link to the notebook located in your repository (the two notebooks should look the same).

Grading Criteria:
The amount of progress that has been made.

4) Presentation (25 points), 28th and 30th November

For your presentation and final report, you will be outlining everything that you have done, explaining your results, and submitting your code. This should, in many ways, be a retrospective on the proposal and include the same components (project name and team members, problem, data, solution, findings). It should include results that show whether your solution worked well or not. If it didn't work well, discuss whether you tried anything to improve it and what you could try. Discuss the main takeaways from your project.

The presentations will happen during the last week of classes. Each team will be randomly assigned to present on either Tuesday (11/28) or Thursday (11/30), and given exactly 8 minutes to present their project, including slides and project demo (if applicable).

Need to submit:
A PDF of your presentation slides (location to be determined later). This is due at 11:59pm on 27th November for all teams, regardless of the day when your team presents.

Grading Criteria:
I will provide more details closer to the date. It will take into account the quality of the project as well the quality of the presentation along with other things.

5) Final project submission (35 points), 5th December

In addition to outlining everything that you have done, the final deliverables have concrete requirements with a short report (3 pages for undergraduate and 6 pages for graduate students):

- Data: Please submit your cleaned data or, if it's too large, a reference to the original data as well as the scripts you used to clean it.

- ML/Stats: Use at least one machine learning or statistical analysis techniques (per Member) to analyze your data, explain what you did, and talk about the inferences you uncovered. Mention the member's name accordingly.

- Visualization: Provide at least two (for undergraduates) or three (for graduate students) distinct visualizations (per Member) of your data or final results. Mention the member's

name accordingly.

- Additional work: In addition to the requirements in the ML and visualization sections above, we would like to see at least one extra from either category. That means a total of five deliverables.

- Results: Fully explain and analyze the results from your data, i.e. the inferences or correlations you uncovered, the tools you built, or the visualizations you created.

Need to submit:
All your code should be in your team's repository. The report will contain the link of the repository. I will provide more details on the format closer to the date.

Grading criteria:
There will be a grade assigned to the whole project, and a grade assigned to you individually based on peer assessment of your teammates (and/or your github code contributions). We will take an average of all of them for this part of the project.