CS 418 — Introduction to Data Science — Spring 2024
Group Project Requirements

The goal of the project is to give students an opportunity to develop an **end-to-end** data science project of their choice. By the end of the course you will have engineered a piece of data-driven software that helps users analyze and visualize a set of data while discovering a set of previously unseen correlations.

There is a Github Classroom where you 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. To create the team and add your teammates using their github usernames, follow this link: https://classroom.github.com/a/CKzhb5b4. 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 finals week of the semester. If you don't have experience with github, take a look at this introduction: https://docs.github.com/en/get-started/quickstart/hello-world. This should be completed before the Proposal due date (Feb. 23rd).

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 "How this part will be graded" for the Progress report and the Final Project).

The project will consist of five main deliverables whose total is 30% of your course grade:

Proposal (4%) – due 11:59pm on February 23rd

The goal of the proposal is to get groups thinking about what they want to do for their final project.

It is important for a data scientist to have good presentation skills. Therefore, we will practice this skill throughout the semester, including with the proposal itself. The format of the proposal is a PPT (or other type of) presentation. The proposal should include exactly four presentation slides, converted to PDF:

- **Project name and participants** (slide 1): 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 located in the Github Classroom designated for this course. *We will check whether your github repository is created in the classroom set up for the course and whether all team members have been added there.*

- **Problem** (slide 2): 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?
- **Data** (slide 3): What is the data that you plan to use? 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.
- **Solution** (slide 4): 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.

Some things to consider: submitting a Kaggle competition as your group project is not acceptable. While 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.

What you need to submit: PDF of the slides to Gradescope by 11:59pm on February 23rd. Only one person per group needs to submit, tagging their teammates. No late submissions will be accepted.

How this part will be graded: presentation clarity, aesthetics, whether it includes all information requested.

Check-in with Professor (1%) – March 5th, March 8th

After the proposals have been submitted, you will be given the opportunity to schedule time with me on March 5th to discuss anything that would help set your project for success, such as 1) challenges that you have encountered and need advice on, 2) help with further refining your plan, 3) getting feedback whether the scope of your proposal is appropriate for the class or needs to be adjusted. Be prepared to discuss what you have done so far. It is expected that by that point you have collected and cleaned your data, and have ideas of the next steps with your data, i.e., either how to begin integration and/or analysis. More details on specific meeting times will be provided later.

What you need to submit: As an output of this meeting, I expect you to submit to Gradescope a refined set of four proposal slides (same format) to reflect the check-in discussion and additional insight from cleaning your data by 11:59pm on March 8th.

How this part will be graded: whether the slides were improved/adjusted based on the feedback.

Progress report (5%) – due 11:59pm on March 29

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 you are analyzing, and the question or questions you are investigating.
- **Any changes**: a discussion whether your scope has changed since the check-in proposal slides. What did you aim to do that you will not do and what have you added to the project?
- **Data cleaning**: show clearly how you cleaned your data.
- **Exploratory data analysis**: explain what your data looks like (words are fine, but visualizations are often better). Include any interesting issues or preliminary conclusions you have about your data.
- **At least one visualization** that tests an *interesting hypothesis*, along with an explanation about why you thought this was an interesting hypothesis to investigate.
- **At least one ML analysis** on your dataset, along with a baseline comparison and an interpretation of the result that you obtain.
- **Reflection**: a discussion of the following:
    - What is hardest 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 are you going to change your project and why do you think it's better than your current results?
- **Next steps**: What you plan to accomplish in the next month and how you plan to evaluate whether your project achieved the goals you set for it.

What you 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).

How this part will be graded: the amount of progress that has been made, clarity of exposition. There will be a grade assigned to the whole progress report that everyone receives, and a grade assigned to you individually based on your github code contributions.

Presentation (5%) – due 9am on April 22

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 retrospection on the proposal and include the same four components (project name and team members, problem, data, solution) though you can use more slides. For everything we asked you to plan, we now want you to explain what you did and how you did it. Additionally, it should include an evaluation that shows 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 Monday (4/22) or Wednesday (4/24), and given exactly 8 minutes to present their project, including slides and project demo (if applicable).

What you need to submit: A PDF of your presentation slides (location to be determined later). This is due at 9am on April 22 for all teams, regardless of the day when your team presents.

How this part will be graded: I will provide more details closer to the date.

Final project (15%) – due 3pm on May 1

In addition to outlining everything that you have done, the final deliverables have concrete requirements:

- **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 ***two*** machine learning or statistical analysis techniques to analyze your data, explain what you did, and talk about the inferences you uncovered.
- **Visualization:** Provide at least ***two*** distinct visualizations of your data or final results. This means two *different* techniques. If you use bar charts to analyze one aspect of your data, while you may use bar charts again, the second use will not count as a *distinct* visualization.
- **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.

What you need to submit: All your code should be in your team's repository. I will provide more details on the format closer to the date.

How this part will be graded: there will be a grade assigned to the whole project that everyone receives, and a grade assigned to you individually based on peer assessment of your teammates and your github code contributions.