# CS 418 — Introduction to Data Science — Fall 2024
# Course 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.

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

## Proposal (4%) – due 11:59pm on September 12th

The goal of the proposal is to get everybody thinking about what they want to do for their final project. This is an individual submission so that we can assign teams based on interest, especially useful if you are not able to form full teams on your own.

It is important for a data scientist to have good communication skills. The format of the proposal is a document in Doc or Markdown format. The proposal should include four short paragraphs (4-5 lines per paragraph), converted to PDF:

- **Project name and Github** (Paragraph 1): The name of your proposed project, and one line describing the motivation, along with your github handle. If you have discussed the idea with classmates, then please include the names, UIC email handles, and github handles of all the team members.
- **Problem** (Paragraph 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** (Paragraph 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** (Paragraph 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, especially after team assignment: 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 or Ellen.

Some things to consider: submitting a Kaggle competition as your course 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 your write-up to Gradescope by 11:59pm on September 12th. Everyone enrolled in the course needs to submit. No late submissions will be accepted.

**How this part will be graded:** Writing clarity, aesthetics (using proper indentation, formatting and so on), mostly whether it includes all information requested.

## Check-in with Professor or TA (1%) – October 3rd, October 8th

After the proposals have been submitted, you will be assigned a team of 5 or 6 classmates who are enrolled in the class. If you provided details about classmates you discussed your idea with, then these preferences will be taken into account. After team assignment, the main goal is to "merge" individual project proposals to a single team proposal.

It is completely acceptable if any of the team member has to change the goal that was mentioned in the initial, individual proposal submission time. Here are some common reasons for this to happen: 1. one of my team member's proposal explained that their idea is more general so I decided to work on that, or 2. after discussion with my team members I realized that my proposal was simply hard to accomplish in a semester's time, or 3. we all collectively modified our ideas to make a single problem that I found to be far more interesting and so on.

First, 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, link will be shared soon (I have to renew my I-Card to authenticate...)

Second, all team members should have student github accounts and be added to the repository before the check-in. 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://guides.github.com/activities/hello-world.

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).

Finally, you may schedule time with me 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 each team member has collected and cleaned relevant data, and have ideas of the next steps with your data, i.e., either how to begin integration and/or analysis. If your project proposal goals align well with the data, then you are in a good position.

**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 (not paragraphs) to reflect the check-in discussion and additional insight from cleaning your data and fifth slide showing proof of Github repository creation with appropriate team member list by 11:59pm on October 1st.

**How this part will be graded:** whether the slides were adjusted based on the feedback, and Github setup.

## Progress report (5%) – due 11:59pm on November 7th

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:

- 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.

- **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 the 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?
- **Roles/Coordination (important)**: Who will be responsible for specific portions of the project (at least two for each portion is recommended): e.g., finding data sources, cleaning, statistical analysis, visualization, machine learning applications, etc.? What deadlines should various components of the project be completed by?
- **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 (10%) – due 2pm on December 03

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). 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 assigned to present on either Tuesday (12/03) or Thursday (12/05), and given roughly 10 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 2pm on December 3rd 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 notebook report (10%) – due 1pm on December 12, 2024

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 (or discovered if nobody has done the analysis before).
- **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. When printed to a PDF, your notebook should be no more than 10 pages. Anything beyond 10 pages will be ignored and not graded.

**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.