The purpose of the course project is to give you an experience of working with real-world data. You will identify a dataset of your own choice, ask interesting questions, figure out the answers using the techniques we learned in the course, and present your work to the entire class.

You may either work on your own, or form a team with other students. Each team should contain at most 3 students. Remember: the larger the team, the harder the project needs to be, which will be enforced in the project proposal review time.

Each of you will get up to 100 points for completing the main components of the project. There are two opportunities where you can additionally get up to 10 bonus points:

- Choose to form a team with others.
- Provide meaningful improvement suggestions to other teams.

Details are elaborated below. You cannot get more than 100 points in the end: 95 points plus 10 bonus points means 100 total points in the end.

The late days policy applies to all project components with an explicit submission deadline, except the final report. Note that the total number of late days stays the same, so the more late days you use for the project, the less you have for your homeworks.

Project proposal (20 points, by Sunday 11:59 PM of week 4)

Identify a dataset that you feel curious about, and submit a project proposal that describes the following:

- Your dataset. If you have no idea where to find an interesting dataset, https://www.kaggle.com/datasets can be a good place to start, but you're free to use datasets from other places, or even datasets of your own.

- Questions you're asking. Explain why the questions are interesting to you. At least one question has to include a modeling component.
- A survey of related work, and elaborations on why your project is different from or complementary to existing research.
- Project plan. List the steps you plan to take to answer the questions you raised. The plan doesn't need to be very concrete; if there are uncertainties in some steps, please explain.

Project proposal review (week 5)

The class crew will review the proposals and provide feedback. The review includes an assessment of the feasibility and workload of the proposed projects given the team sizes. You may need to revise the proposal according to the feedback. The proposal should be finalized by the end of the week.

Elementary EDA (10 points, by Sunday 11:59 PM of week 6)

Submit a short Jupyter Notebook with at least 3 visualizations of your dataset. The visualizations should reveal interesting characteristics of the data, ideally any surprises you've found.

Presentation (30 points, first day of week 8)

You will present in turns to the entire class about your projects. The presentation should cover the motivation, the methodology, and the results of your projects.

[Bonus] Improvement suggestions (up to 5 points, during presentation)

At the end of every presentation, the audience is encouraged to propose ideas of improvements that are intended for the presenters to include in the final report. The ideas can be other questions to ask about the dataset or alternative methodologies. The suggestions cannot be too trivial or exceed 1 week's workload. The presenters don't need to agree / disagree at presentation time, but if the presenter incorporates the suggestion in the final report, and if the

improvements are assessed to be meaningful by the class crew, both the suggester and the presenter gets 1 point for each incorporated suggestion.

Final report (40 points, by Sunday 11:59 PM of week 8)

The report should be two Jupyter Notebooks, an analysis notebook and a narrative notebook. The narrative notebook should be in a research paper format, and include a title, list of authors, abstract, introduction, description of data, methods, results and discussions. The analysis notebook should contain all the analyses you performed to support your claims in the narrative notebook.

[Bonus] Team peer review (up to 5 points, same time as final report)

If you're not working alone, write a peer review for each of your teammates. The peer review should include the following three sections:

- Impact: What are XXX's contributions to the project?
- Recognition: What did XXX do really well?
- Growth Areas: How can XXX improve?

You also need to indicate on your review whether you want to share it with your teammate or not; it is your choice. The number of points you get will be assessed based on the peer reviews you get from your teammates.