In this assignment, you are given a dataset of wine ratings, as assigned by human tasters, along with other pertinent characteristics of each wine. Your task is to build a regression model that can forecast the rating for novel, unseen wines. Begin by downloading the archive file provided on Moodle — this archive contains three files: two CSV files (you can just read them like plain text files, or open them in a spreadsheet program like Excel) with red wine and white wine data, and a third text file that describes the format of the data.

You have two options for performing the regression:

- You can use an off-the-shelf regression solver — I recommend the `scikit-learn` library for Python. This is already installed on the JupyterHub server and you'll find excellent online tutorials on how to use it.
- Or, you can implement stochastic gradient descent yourself, as described in class. This would be an excellent learning experience — and assuming you wrote it correctly, would be worth a few extra-credit points. (If you go this route, I will expect you to compare your results against `scikit-learn` for evidence that your implementation is correct.)

The Write-Up

Here's the overarching writing rule for this course: *you need to be sufficiently precise with your writing and include enough detail that a competent reader could reproduce your results.* Here are some specific things to address in your report, in no particular order. This is *not* meant to be an exhaustive list.

- What preprocessing did you perform on the data? Did you perform any exploratory data analysis? Generate any plots or charts? Describe these, along with any relevant findings, in your report.
- What regression models did you build? How do they compare in terms of performance? What was the best performing model, and how did it do? Optionally, you can go beyond the methods we've seen in class and try other linear regression variants — if you go this route, you should describe how your chosen algorithm works (and don't forget to include citations!).
- What was your model-building and tuning regime? How did you address overfitting? How did you make hyperparameter choices?

- Finally, what are the broader impacts of this work? How could the manner in which the data was collected affect the results? Could work of this kind (wine-quality prediction specifically, or the broader problem of predicting the quality of consumer products in these data-driven ways) have other knock-on impacts on society, either positive or negative? *I will expect serious engagement with these questions: while I don't expect this section to be longer than a paragraph in your report, it must nonetheless reveal considered reflection on your part.* One way to demonstrate this is by citing related work from reputable sources, though original reflections will also be accepted.

Your primary deliverable for this assignment is a PDF write-up composed in LaTeX and formatted using the AAAI template (see the LaTeX starter kit on Moodle). Submit a single zip file that contains all your LaTeX files and the code you wrote, via Moodle. Make sure that your final PDF is clearly identifiable in your zip archive.

Recommended Timetable

Here's a recommendation for how to budget your time over the next couple of weeks as you work on this assignment.

- **Jan. 27–29:** Explore the dataset, think about feature engineering, pick your modeling tool (Python + `scikit-learn` vs. your own implementation), build your first model.
- **Jan.30–Feb. 2:** Run more thorough experiments (hyperparameter tuning, further feature engineering, etc.), analyze your results and iterate, search the literature for related work on the problem, write your *Introduction* and *Background* sections, optionally meet with Dr. Ramanujan to get advice/feedback (both on technical issues and on writing).
- **Feb. 3–5:** Complete experiments, take a step back and think about your report's narrative and the broader impact of your work, write drafts of your *Experiments*, *Results* and *Broader Impacts* sections, consult with Dr. Ramanujan as appropriate.
- **Feb. 6:** Wrap-up any pending experiments, write the *Conclusions* section and the abstract, proofread the entire report and prepare for your check-in with Dr. Ramanujan.
- **Feb. 7–9:** Check-ins with Dr. Ramanujan about the status of your code, experiments and paper draft during class time.
- **Feb. 10–14:** Use the feedback to improve your report — proofread, (re)run experiments, submit your final paper.