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 cloning the git repository onto your local machine. The data folder contains three files: two CSV files with red wine and white wine data, and a third text file that describes the format of the data. You should create a new folder called source which contains any source code for the project. Your Jupyter notebook can live in the main folder of the repository. As seen in class, the `pandas` package can be used to manage CSV files.

For performing the regression:

- You must implement stochastic gradient descent yourself, as described in class.
- You may verify your results to an off-the-shelf regression solver — I recommend the `scikit-learn`, `numpy`, or `statsmodels` libraries for Python. These should be installed on all campus machines and you'll find excellent online tutorials on how to use them.

**The Write-Up**

**Your primary deliverable for this assignment is a blog-like post (see `https://www.wired.com/author/rhett-allain/` for inspiration) written in a `jupyter notebook`. Your notebook, along with all other necessary files, will be hosted on you and your partner's fork of the GitHub Classroom repository.**

The goal is to communicate what you did to the reader in a manner that grabs the reader's interest while also communicating the work effectively. This work should appeal to a general reader with no machine learning background but with an interest in the topic.

Note that you may not want to show *all* your code to the reader in the notebook. You can create user-defined modules and import them in the notebook to hide certain functionality that may not help the reader understand the work.

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, and no matter the format of the report. 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?

### Recommended Timetable

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

- **Aug. 30–Sept 4:** Explore the dataset, think about feature engineering, build your first model.
- **Sept 5–7:** 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. Kuchera to get advice/feedback (both on technical issues and on writing)
- **Sept. 8–10:** Complete experiments, take a step back and think about your report's narrative, write drafts, consult with Dr. Kuchera as appropriate
- **Sept 11–13:** Wrap-up any pending experiments, revise and proof-read the blog post and submit.