Final Project Proposal Guidelines

Proposal Due 11/7

Choose one of the following templates. Deliverables for the proposal are in **bold.**

Your submission should also include the **names of your group members. You must have 2-4 people in your group.** You can't do a solo project because this is a Teamwork Hub course. **For 4-person teams, give a breakdown of what you plan each team member will do.**

Only one proposal per team needs to be submitted.

Template 1: Self-Collected Data and Visualization

Decide on two hypotheses that you can test with a dataset you collect yourself. **Describe in your proposal how you will get your data.** (You will eventually turn in this data with the project, as a .CSV file.). For this template, you should be either collecting the data yourself or doing significant cleaning or manipulation of the data; if you want to work with an existing dataset, see template 2.

Examples of hypotheses from previous iterations of the course: Students who live on one side of campus party more than those who live on the other side; Lego buyers who are young get sets that are more fantastical than older buyers; there are more McDonalds' in poorer counties in Massachusetts than wealthier counties. *Don't forget that you need two of these,* and they should ideally be related enough that you can collect the data for both simultaneously (e.g. as part of the same survey).

Describe how you plan to evaluate your hypotheses with statistics and/or visualization. Statistics could include t-tests, correlations, regressions, and/or chi-square tests. Visualization could include boxplots, scatterplots, and/or bar charts, or things like word clouds and geolocated data. If you lack a statistical background, you can skip the statistics and plan to include more ambitious visualization. You must include at least some visualization.

You must do all your analysis and visualization in Python.

Template 2: Publicly Available Dataset and Machine Learning

Find a publicly available dataset (give a link in your proposal) that has some data you can use for the final project, and **decide on 2 approaches that you could take to predict some variable with machine learning.** (Approaches include k-nearest neighbors, decision trees, and random forests; you can also experiment with anything else you find in scikit-learn or elsewhere.) You don't need to implement the machine learning algorithm yourself - you can make use of scikit-learn and other libraries. **Identify the columns you will use for prediction.** You should also **plan to vary some parameter in each approach** to achieve the best possible performance - for example, vary *k* for *k*-nearest neighbors, or vary maximum depth for decision trees.

As with template 1, you must code everything in Python.

Template 3: Propose Your Own

If you don't want to follow the templates here, you may propose your own project. Describe in your proposal the approaches you'll try and what you expect to put in your final paper. The problem should involve Python and data somehow. Try to plan for an amount of work equivalent to one problem set per group member.

Deadlines

11/7 Project proposals due
11/14 Peer assessments out
11/30 Peer assessments due (Wednesday)
12/7,12/9,12/12 Final project presentations
12/12 Final papers and code due

Project Grading Breakdown

Proposal	10%
Peer review	10%
Ambition	25%
Execution	30%
Presentation	10%
Paper	15%

Notes:
Proposals are graded based on whether all the boldface pieces are present, and not on the ambition (or anticipated execution) of the proposal.

Peer review, an assignment in which you evaluate your teammates' progress, is just graded on whether you answer all the questions - points aren't subtracted for reporting trouble with the group.

Ambition refers to the scope of the project that is turned in, not the scope of the project that was proposed.

Execution refers to how well executed the plan was, such as whether techniques and parameters are chosen appropriately. This is how the code is graded, but also could include things like choice of statistics or visualization.

Presentation is a "lightning" talk delivered on one of the last 3 days of class. It is graded mostly on whether all the parts are there, but could have deductions for being unclear or misleading.

Paper is graded mostly on whether all the parts are there, but could have deductions for saying something incorrect or being misleading. The final paper will include the pieces of a scientific paper: introduction in which you explain how you came to the problem, methodology where you describe exactly what you did, results and conclusions.