The Project:

There are many interesting areas of statistics that we won't have a chance to cover in this class. So, I want to give you a forum for investigating a topic of your choosing and relating your findings to me via a report.

For this project, you should work in **groups of two or three (unless you choose the "Application" option; see below)**. These groups are of your choosing. The deliverables for this project will be (the due time for all dates below is 9:30 am):

- March 29. Tell me your group members' names, topic of interest, and write up a paragraph giving a very brief overview of the topic and why you are interested in it. (2 percentage points).
- April 12. Write down an itemized list of what you plan on doing for the project (such as examining a method via simulation, analyzing a particular data set, describing how/why a method works, ...). Be specific about each of these items. Your goal is to convince me that you've thought through what you are planning on doing. (3 percentage points).
- April 26. Submit a progress report of your write-up. It can have as many figures/tables as you need, but can be no more than 500 words. (5 percentage points).
- May 10 Submit a final version of your write-up. It can have as many figures/tables as you need, but can be no more than 1000 words. (10 percentage points)

(for the word count limit, only count the "main text". Captions, references, code, footnotes, section headings, etc. do not count towards the limit)

This project, as per the syllabus, is worth 20% of your total grade in the class. The fraction of this 20% are listed parenthetically above.

Some examples of potential topics:

Log linear models and Poisson Regression: A particular form of a GLM in which the distribution of the response is assumed to be a Poisson distribution.

Contingency tables and/or A/B testing: Contingency tables are useful for analyzing data that is entirely categorical. A/B testing is a particular example of this that is commonly used at tech companies for running tests about different products that are offered.

False Discovery Rates (FDR): We have discussed controlling for multiple comparisons. All of the techniques we have covered involve inflating the quantiles, making larger confidence intervals. FDR controls for multiple comparisons via a conceptually different route. It is used commonly in applications where a huge number of tests need to be run simultaneously, such as in genomics.

Propensity Scores: In a designed experiment, we can control for confounding via randomization or blocking. If this is infeasible (or just didn't happen), propensity scores can be used to attempt to correct for systemic confounding in an observational study.

Density Estimation and/or cluster analysis: In some cases, there isn't a response variable Y or there aren't explanatory variables X in a particular problem. In this case, we can attempt to estimate the distribution of X or Y. If we are estimating the distribution of Y, this is commonly called density estimation and if we are estimating the distribution of X it is commonly called clustering.

Bootstrap: We can quantify uncertainty about an estimator using resampling instead of using the CLT. A major tool for this is called the bootstrap.

Application: If you have a project you are currently working on, you can make your report about this application. This project can be from work or research but not from a project in another class. If you choose this option and it is awkward to involve another person in the project (e.g. the data is proprietary), then you may work alone.

Other: Don't feel limited by this list. If there is another topic you are interested in, feel free to propose it.