GENERAL ASSEMBLY Data Science

FINAL PROJECT REQUIREMENTS:

For our final project, students will work individually and choose one of the following options:

Option I:
Address a data-related problem in your field or a field you're interested in. Pick a subject you're passionate about; if you're interested in the subject matter it'll be more fun and you'll probably produce a better project! Apply modeling techniques (regression, classification, etc.) and data analysis principles (cross-validation, caution against overfitting, etc.) and report your results.

**Your project will need to be vetted by Alfred or Patrick to ensure the scope is appropriate.*

Option II:
Choose from the following suggested Kaggle competitions or choose one of your own and apply modeling techniques and data analysis principles, and then report your results.

- Yelp's Recruiting Competition: Given training data in the form of 229k reviews of 19k businesses and check-ins from 43k users, the goal is to predict the number of "Useful" votes a review will receive. A lot of the data is unstructured and messy, but there's a lot of good signal in textual analysis, and I think someone who runs an LDA will go far in this competition.

- Random Acts of Pizza-Predicting Altruism: This data covers 5,671 requests from a Reddit community called ‘Random Acts of Pizza’ in which people tell the group why they need a pizza right now. The goal is to predict whether or not someone ended up getting a free pizza delivered. It contains both text data and continuous data like upvotes and downvotes.

****For this option, if you choose something other than the recommended competitions please check with the instructional team to make sure the competition is suitable for this course.*

OUTLINE (Due Aug 7th)

- What problem are you solving?
- Description of data set: Where is it coming from? What is your target feature?
- Hypothesis?
- Statistical methods you plan to use and why
- What business applications do you think your findings will have and why?

PRESENTATIONS (August 17th):

On the last day, all students are required to give a 5 – 7 minute presentation summarizing their data results. The presentations should target a non-technical audience and serve the purpose of having students practice the highly sought after communication skills data scientists need.

What to cover in presentation:

- Overview of problem and hypothesis
- Overview of data
- Modeling techniques used and why
- What decisions your findings allow you to make.

GRADING:

EXCELLENT	Student's presentation is engaging, clear, and informative. It describes the project, approach, and conclusions, and is suitable for a non-technical audience.
GOOD	Student's presentation is as above but is either inadequately engaging, clear, or informative.
FAIR	Student's presentation fails on two out of three of engaging, clear, and informative.
POOR	Student's presentation fails on all three or is off-topic with respect to their paper.

***Additional open-ended feedback will be provided to each student

PAPER: (3-5 PAGES)

Students are also required to submit a 3 – 5 page paper that describes the project's technical details. The paper should target a technical audience.

What to cover in paper:

- Description of problem and hypothesis.
- Detailed description your data set.
- How did you decide what features to use in your analysis?
  - What challenges did you face in terms of obtaining and organizing the data?

- Describe what kinds of statistical methods you used, and perhaps others you considered but did not use, and how you decided what to use.
- What business applications do your findings have?

GRADING

EXCELLENT	Student's paper demonstrates thorough understanding of techniques, data management, and the application of these in programming. It is clearly communicated to a reasonably technical audience.
GOOD	Student's paper demonstrates above knowledge, but lacks some necessary rigor, detail, and/or exploratory depth or is not well communicated.
FAIR	Student's paper demonstrates some learning of principles taught in class, but is clearly lacking in rigor and/or depth.
POOR	Student's paper is incomplete or does not conclusively demonstrate understanding of statistics or programming.

***Additional open-ended feedback will be provided to each student