GENERAL ASSEMBLY

FINAL PROJECT REQUIREMENTS:

For the Data Science final project, students will work individually to analyze in a problem in their field of interest using tools from the course.

Address a data-related problem in your professional field or in a field you're interested in. Pick a subject that you're passionate about; if you're strongly interested in the subject matter it'll be more fun for you and you'll probably produce a better project! (You can additionally choose a Kaggle competition)

In the course of the project, we expect you to complete the following tasks:
1) Gather, preprocess and **visualize** a dataset. What can you learn from a high-level analysis?
This will be the focus of the April 16th initial project presentations.

2) Apply modeling techniques (regression, recommendation, classification, etc.) and data analysis principles (cross-validation, caution against overfitting, etc.) and report your results.

3) Plan out how you would implement what you've done in (2) as a live system. Where would the data live? How would it represented? How would end-users access it? How often would you have to re-do your analysis?

You will need to vet your project with the instructional team to make sure the scope is suitable for this course. You can talk to us during class, office hours, or via email.

OUTLINE (DUE APRIL 14, PRESENT (3-5 MINUTES) AND DISCUSS APRIL16)

- Problem you are solving?
- Description of data set and how you will obtain it
- Hypothesis
- Statistical methods you plan to use and why
- What business / practical applications do you think your findings will have?

PRESENTATIONS (4/28 - 4/30):

On the last day of class, all students are required to give a 5 – 7 minute presentation that summarizes 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 that data scientists need.

What to cover in presentation:

- Overview of problem and hypothesis
- Overview of data

- Any visualizations or overview you created
- Modeling techniques used and why
- What decisions your findings allow you to make.
- Discuss your implementation plan (or any hurdles there would be)

GRADING:

EXCELLENT:	Student's presentation is engaging, clear, and informative, describing 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 his or her paper.

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

PAPER (OR WELL ANNOTATED IPYTHON NOTEBOOK)

Students are also required to submit a short paper with code or a well annotated IPython notebook 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?
    - What did you learn from the initial exploration phase
- 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?
- ***Describe the implementation plan in detail from the ingesting of data to how end-users would access it.***

GENERAL ASSEMBLY

GRADING:

EXCELLENT:	Student's paper demonstrates thorough understanding of statistical techniques, data management, and the application of these in programming, and 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

IMPORTANT DATES:

Deliverable:	**Deadlines:**
Outline of Project	April 14
Initial Data Processing Presentations	April 16
Final Presentations	April 28
Final Paper	Last Day of Class (April 30)

The instructor will be checking in with you periodically to make sure you are making good progress on your projects. Please use office hours to obtain additional help.