FINAL PROJECT REQUIREMENTS:

For the Data Science final project, students will work individually and can choose from one of the following two options:

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

****For this option, you will need to vet your project with the instructional team to make sure the scope is suitable for this course.*

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.

****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.*

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 Feb 20 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?

NOTE the following Home Work Assignments require submission of a paper version through Schoology in addition to participation in class:

- *HW3, HW5, HW6, and HW8*

ELEVATOR PITCH (HW3) (DUE IN CLASS+SCHOOLOGY SAT. 6/28)

- A one-paragraph write up of the project you propose

- A concise (<90 seconds) verbal pitch of your project to the rest of the class • Include:

    - A concise statement of the goal of your project
    - Where you will get the data / what dataset you plan to use
    - What type of machine learning problem this is (from our 2x2)
    - Why you think this is a cool project

NOTE: This does not lock you into that project. You can change your idea and your ML technique as we learn more. The point is to get moving on it :-)

PROJECT PROPOSAL (HW5) (DUE IN CLASS+SCHOOLOGY SAT. 7/12)

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

PROJECT MILESTONES (HW6) (DUE IN CLASS+SCHOOLOGY SAT. 7/19)

What to cover in the outline:

- Description of problem and hypothesis.
- Detailed description your data set.
    - What is the nature of the data you are working with?
    - What are the feature engineering methods that you are experimenting with.
- Algorithms you will be implementing in the course of model development.
- Tuning methodology based on the data set?
- Testing evaluation metrics you will use to measure success.
- What business applications do your findings have?

PEER FEEDBACK (HW7) (DUE IN CLASS SAT. 7/26 AND 8/2)

On 7/26 and 8/2 We will split into teams of 2-3 people, review each other's final projects and progress to date, and provide peer feedback.

PAPER/NOTEBOOK (HW8): (DUE ON SCHOOLOGY 8/16)

Students are 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?
- Describe what kinds of machine learning and 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?

PRESENTATIONS (LAST DAY OF CLASS):

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
- Modeling techniques used and why
- What decisions your findings allow you to make.

GRADING:

EXCELLENT:	Student's paper/presentation 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/presentation demonstrates above knowledge, but lacks some necessary rigor, detail, and/or exploratory depth or is not well communicated.
FAIR:	Student's paper/presentation demonstrates some learning of principles taught in class, but is clearly lacking in rigor and/or depth.
POOR	Student's paper/presentation is incomplete or does not conclusively demonstrate understanding of statistics or programming.

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