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.

- 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.

- Titanic: Machine Learning from Disaster: The data is highly structured, and there's little preprocessing that needs to be done to create a classifier or a logit. The goal here is to predict which passengers on the Titanic survived the sinking, given features about their demography (gender, age, socioeconomic class), and their position in the boat (where they stayed, the fare they had, the port they left from, etc.)

- Your choice from kaggle.com

****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 21ST OCTOBER)

- 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?

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 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: (4 -6 PAGES)

Students are also required to submit a 4 – 6 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 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	Oct 21st (1 week after assigning)
Meet with GA instructional team to discuss project idea	Oct 21st – Oct 28th
Final Presentations/Paper	Nov 25th/27th

The instructor and TAs 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.