For the first part of the course project there four steps:

a) Choose a dataset

b) Register a team (**due October 20, 2016 at 5pm**)

c) Investigate and explore the dataset

d) Write a report and submit on Gradescope and to your peer reviewers (**due October 27, 2016 at 5 PM**)

1 Introduction

This is the first part of the mini-project. You will be applying the concepts you learn in the class, while analyzing a dataset of your choosing. You can choose to do this alone, or in teams of two. (Working together with someone else is *strongly encouraged!*)

In this first part, the focus is on obtaining and understanding your data. The second and third part are focused on prediction and inference, respectively. An important part of the project is to provide peer-review on other projects; that way you can learn about other approaches, and receive feedback on your own work.

2 Choosing a dataset

You can either choose a dataset we have selected, or find a dataset of your own. You are strongly encouraged to find your own dataset, ideally in an area you find more interesting and are personally motivated to explore. You have to like your data; you're going to spend a lot of hours staring at it, so you should find it fun and interesting to work with the dataset you've chosen!

If you choose your own dataset, make sure the dataset is rich enough to let you play with it, and see some common phenomena. In other words, it must have at least a few thousand rows ($> 3.5 - 4K$), and at least $15 - 20$ columns. Of course, larger is welcome. Some data sets might have fewer columns but very rich structure and also be viable, in that case, please come talk to one of the TAs.

We have provided two datasets that you could use: data on real estate from Ames, Iowa; and data from the U.S. College Scorecard. Both datasets are available online (in the "Datasets" section of the course site), together with some information about their origin, as well as a data dictionary (that explains what the columns mean).

3 Register a team

Once you have formed a team and have picked your data set, **register your team** by filling out this form:
`http://tinyurl.com/mse226project`

We use the responses to assign TAs to projects, and to set up the peer grading. **In order to have everything run smoothly, the deadline to submit a team at the link above is 5pm on October 20, 2016.** Be sure to register before the deadline, as we will not adjust the peer review assignments to accomodate late submissions.

4 Setting aside a holdout set

Before doing anything with your data, *randomly* choose a test set (representing 20% of your rows), and keep it for later. (You don't need to report anything to us for this part.)

You will not touch this test set again until the end of the course! Fix this set from the beginning, and use the remaining 80% for exploration, model selection, and validation.

Note: It may be harder to do this properly with some type of datasets, like time series; if you have selected time series data, let us know and we can help you find a testing strategy. In general, for such data, you want to train on earlier data and test on later data.

5 Investigating and exploring your data

Once you've chosen your dataset, it is time to explore the data to get a better grasp of the structure, and to find interesting questions to explore.

Write a report of **1-2 pages** describing your dataset and interesting findings; your report must be submitted by **October 27, 2016 at 5 PM on Gradescope**. Below we provide some guidance on what we are expecting; however, data analysis is a dynamic process, rather than ticking of checkboxes. You will be judged on the overall quality of your analysis, rather than by a rubric of items that need to be satisfied. Furthermore, be succinct; the goal of the report is to highlight your findings, rather than describe everything you did. Finally, keep your audience in mind; your peer reviewers might not know anything about your dataset, make sure your report is accessible and interesting to them.

Required: Make sure your report addresses the following.

- Describe the dataset you have selected. Explain how the data was collected; Do you have any concerns about the data collection process, or about the completeness and accuracy of the data itself?

  Note: This is also a good time to go through some basic *data cleaning*: if there are columns that are obviously extraneous to the data analysis (e.g., IDs or metadata that have no bearing on your analysis), you can remove those now to make your life easier.

- Suggest at least one possibility for a continuous response variable. Explain your choice.

- Suggest at least one possibility for a binary response variable. Explain your choice. *Note*: You can always create a binary response variable by starting with a continuous variable Y, and then defining $Z = 1$ if Y exceeds a fixed threshold, and $Z = 0$ otherwise.

- Loosely speaking, what questions might you be able to answer using this dataset? What makes this dataset exciting?

Recommended: The following list serves as a guide of the kinds of questions your report might address; good reports will make decisions about which of these are most meaningful to include, or even other features of your data analysis that are not described below.

- Are any values in your dataset `NULL` or `NA`? Think of what you will do with rows with such entries: do you plan to delete them, or still work with the remaining columns for such rows?

- Are there any columns that appear to be irrelevant to the questions you would like to answer?

- Find covariates that are most strongly positively correlated, as well as most strongly negatively correlated, with your choices of response variable.

  Are there variables you think should affect your response variable, that nevertheless have weak correlation with your response variable?

- Look for mutual correlations between these variables you identified in the last part. Create scatterplots for pairs of covariates you believe correlates well to the response variable.

  Are correlations *associative* in your data? That is, if A is correlated strongly with B, and B with C, is A also correlated strongly with C in your data?

- Are there variables you would like to *add* to your dataset as you embark on your analysis? For example, are there interactions or higher order terms that might be relevant?

- How can you visualize interesting patterns in your data?

- Suggest one or two population models that you think might be relevant for your chosen continuous response variable. Does your suggestion depend on the desired goal (prediction vs. inference)?

  Note that there is no right answer to this question! We just want you to start thinking about what kinds of models might be reasonable to capture relationships between variables in your data. At this point you don't have to fit any regressions; it will just be useful to refer back to your answer to this question as you move forward and actually start building models in the next two problem sets.

Note: These steps are just the tip of the iceberg! Ideally, you will look at your data many different ways; for example, it's useful to look at means and variances of columns, grouped based on the level of a categorical variable. (E.g., in the College Scorecard data, you might look at how future earnings differ for public vs. private colleges.)

Try to play with and understand your data as much as you can *before* you start building models!

6 Submission and checklist

- Make sure you register your team by October 20, 2016, at 5 PM so that we can organize the reviews.
- To submit your report, upload it to Gradescope by October 27, 2016, at 5 PM. Furthermore, send it to the team(s) reviewing your project by email. You will receive the details when we have finalized the peer review pairings.
- Reach out using a private post on Piazza if there are any logistical issues, such as not receiving a project you have to review, as soon as possible.