For the first part of the course project there are two objectives:

a) Choose a dataset.

b) Investigate and explore the dataset.

1 Choosing a dataset

You can either choose a dataset we have selected, or find (or construct) 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, here are the requirements:

- The dataset should be 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 $20 - 25$ columns. Of course, larger is welcome.
- The dataset should have a reasonable mix of both continuous and categorical variables. Every column you use in your analysis must be either categorical or continuous; and you will need at least one outcome variable that is continuous. If you start with a dataset which has few columns which appear to be neither categorical or continuous, you will either need to exclude them from your analysis, or delete or transform them before analysis.

If you choose to use a dataset that we have provided, you have two choices: 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).

2 Investigating and exploring the dataset

Once you've chosen your dataset, work out the following steps.

a) Describe the dataset you have selected. Explain how the data was collected, and explain the meaning of the columns. 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.

b) 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? (You don't need to report anything to us for this part.)

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

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.

d) Compute the mean and variance for each of the columns (you don't need to report this to us). Are there any columns that appear to be random noise?

e) Suggest at least one possibility for a continuous response variable. Explain your choice. Compute the mean and variance of this variable.

f) Suggest at least one possibility for a binary response variable. Compute the mean of this variable (i.e., the fraction of rows for which this variable is 1.) 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.

g) Find the five covariates that are most strongly positively correlated, as well as most strongly negatively correlated, with your choice of continuous response variable. (You should do the same for your choice of binary response variable, but you don't need to report the results.)

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

h) Now find mutual correlations between the ten variables you identified in the last part. Create a scatterplot for every pair 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?

i) 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? (You don't need to report your answer to this part.)

j) Suggest one or two population models that you think might be relevant for your chosen continuous response variable. (You should do the same for your chosen binary response variable, but don't need to report the results.) 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!