This part of the mini-project is all about prediction: your goal is to find a model that predicts your outcome variable(s) well. You must build a model for both a **regression** task (i.e. continuous outcome variable) and a **classification** task (i.e., binary outcome variable) on your data set.

Important notes:

- DO NOT use the test set you previously created. For now we just want you to focus on *building* the best model you can. You will use whatever model you build on the test set in the final part of the project.
- Keep your report concise. We want you to turn in code for the best linear/logistic regression model that you built, as well as for the best alternative model that you constructed (if any – see below).

1 Baseline models

Before you launch into creating your predictive models, start with some baselines to give you some sense of what you are trying to achieve. Make sure you have an evaluation strategy in mind; recall that if you're not sure what to do, cross validation is always a reasonable way to compare model performance.

In particular, answer the following questions:

a) For your regression problem, predict the mean outcome; for your classification problem, predict whichever label (zero or one) occurs more frequently in the data. What is the performance of each approach?

b) For your classifiation problem, is 0-1 loss the right objective, or something else?

c) Fit a linear regression (resp., logistic regression) model with only a few covariates that you think are likely to be important. How does your model compare to the baseline model?

2 Building your model

Now you can set about building the best predictive model you can.

a) Start by improving on the linear regression (resp., logistic regression) baseline model you built above, however you can: transformations, interactions, regularization, etc. Keep track of the best model you can build for your chosen objective. Make sure you record the model you have selected.

b) Next, feel free to try any other methods you wish on your data; among other things, you can try k-nearest-neighbors, or naive Bayes classifiers, or any other methods you may be interested in investigating. You are welcome to do as much or as little as you like for this part of the project, but if you do try out other methods, please (concisely) describe the following: Which method(s) did you try? Which methods worked well, and which ones failed? Can you explain their performance, e.g., in terms of bias and variance?

c) Finally, report on the best of all the models you tried. Most importantly, give an estimate of what you think the test error will be when you run your model on the previously held out test set; explain how you arrived at your estimate.

d) Conclude with self-reflection: did you notice any practices in your model building procedure (even in the way we asked you to carry out the steps above) that you think might lead your estimate in the preceding step to be overly optimistic?