STAT 8330: **Comments: Project 1**

A good project should include the following:

- A brief introduction that gives the context of the problem, the number of variables, samples, and predictive/classification goals and metrics used to evaluate performance. Potential challenges should be mentioned. If space allows, a brief overview of the rest of the report.
- Data Analysis: what is unique about these data (e.g., missingness, lots of categorical variables, skewness). Potential relationships (e.g., some plots to suggest some variables are more important than others, and some predictors are highly correlated with each other).
- Data Cleaning: How to deal with the missingness in this problem. Important to note that in some cases, the missingness is because a house doesn't have a particular feature - that isn't really missing. In other cases, a covariate might actually be missing. Possible fixes: in the first case, combine categories; in the second, consider imputing values (see the `mice` package that has multiple imputation strategies).
- Feature Engineering: Combine variables in novel ways to make new predictors. E.g., create new variables that calculate the total number of bathrooms and the total number of square feet; create a categorical variable that bins the neighborhood into three groups (rich, poor, and neutral), etc. One can also use cluster analysis, PCA, or nonlinear dimension reduction to make new variables.
- Candidate Models: Describe your candidate models; in particular, **why** you considered the models you did and why you didn't consider others that we have talked about. Then, describe how you train these models (what is your procedure for splitting the data for your training evaluation; did you use cross-validation? how did you select your tuning parameters, etc.). Based on your analysis what are your best models that you will use for the final prediction?
- Final Predictions: Present clearly the prediction result from Kaggle and make sure you clearly identify the model you used (inlcuding the tuning parameters).
- Discussion: how did your models perform? How did these results compare to your training results? Why might they be different?
- Conclusion: Brief conclusion stating your results and (importantly!) what modifications you could try to improve your results.

Some Additional Suggestions for the Future:

- It is a good idea to seriously think about the workflow of your modeling (see the flowchart from Wonjae's project that I have included below for an excellent example!)
- It is good to familiarize yourself with imputation methods such as in `mice`
- When fitting many different classification or regression models, consider using the `caret` package.
- Be careful when removing variables from your analysis; most of the methods we use for large p problems can deal with a large number of variables through regularization or, in the tree context, by construction.
- Be careful when using transformations - they may not be needed. Remember, it is the errors in a linear regression that have to be normal (after conditioning on the predictors) not the response!

- It is quite often the case that an *ensemble* of models works better than one single model. This notion of *model averaging* has been noted in Statistics for quite a while and is increasingly being used in the best predictions in machine learning (see `https://en.wikipedia.org/wiki/Ensemble_learning` and you can just google on *ensemble learning wth r* to see practical implementations). It is a simple idea, and I deliberately don't talk about this in Data III up to this point, but note that the best class predictions over the last 4 times I've taught the course have used ensembling.

** Example Flowchart: (Wonjae Lee)