**Case Assignment #4**

This assignment is designed to have you explore a dataset with a large number of variables in order to create predictive models. You will work through the entire process by first performing some exploratory analysis and data manipulation, then creating some principal components, creating both predictive and classification models, and writing up the results with your conclusions.

**Examining a Housing Dataset**

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home. The [Ames Housing dataset](http://www.amstat.org/publications/jse/v19n3/decock.pdf) was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

I have made some variable modifications already - similar to what I’m asking you to do in the Data Manipulation section - and also have preemptively removed some variables that are not useful for our prediction models to focus your analysis on this project so that they are only 36 variables left, but the remaining variable descriptions are listed here: <http://jse.amstat.org/v19n3/decock/DataDocumentation.txt>. All variables with 0 or 1 values mean yes or no that they have that characteristic.

**CREATE A REPORT WITH THE FOLLOWING SECTIONS**

1. **INTRODUCTION**

Include a brief introduction to the problems you are exploring.

1. **DATA**
   1. **Data Manipulation**

What changes did you make to the data columns? Document them here. Also discuss that there is a training and test data set and why they are being used.

* 1. **Data Source**

Where did the data come from?

* 1. **Data Exploration**

Include a concise summary of the variables (for example, JMP Col Viewer). Include your summary explorations here focusing on the most useful variables. Do you see any issues with missing data?

* 1. **Feature Reduction**

Discuss the components that you created as potential new variables.

1. **ANALYSIS of PREDICTIVE MODELS**
   1. **Prediction for Sale Prices**

Create regression models for predicting sale prices, Include a summary of your two final models here. Compare the two models and determine which works the best and what variables most influence price. **Predicting Sales Price Analysis** – Create both a least squares regression and a regression tree model to predict sales price. As you examine different models, examine your error measures (talk about at least 3) to determine which model is best. Pick a best least squares regression model (remove variables that are not significant/have p-values greater than some threshold you set (e.g., .1 or .05)) and a best regression tree model that will be in your final report (remember simpler models are better if they have comparable error measures). Document the variables used in each model. Include your regression equation and the regression tree rules for these best models.

* 1. **Prediction for Overall Condition High**

Include a summary of your two final models here. Compare the two models and determine which works the best and what variables most influence a high overall condition. **Predicting Overall Condition High** – Create both a logistic regression and a classification tree model to predict overall condition high. As you examine different models, examine your accuracy and error rates (talk about at least 3 including error rate and sensitivity) to determine which model is best. Pick a best logistic regression (remove variables that are not significant/have p-values greater than some threshold you set (e.g., .1 or .05)) and a best classification tree model that will be in your final report (remember simpler models are better if they have comparable error and accuracy measures). Document the variables used in each model. Include your logit equation and the classification tree rules for these best models.

**ADDITIONAL DETAILS ABOUT THE ANALYSIS THAT MUST BE COMPLETED**

1. **Data Manipulation**
   1. Recode the GarageType variable to have values of either Attachd or Other.
   2. Create a new OverallCond\_High variable to be 1 if the OverallCond is greater than 5 and 0 otherwise. (Make sure NOT to use both of these variables in the same models) as they represent the same information. This new variable will be your Y/Response variable for your second set of prediction models.
   3. Feature Reduction (Principal Components Analysis) - Pick a subset of your variables – for example those that are related to extra features of the house, etc. and create a set of subcomponents (so if you have 4 variables – can you just use 1 or 2 components?) to represent this data. Rename the components based on the eigenvector loadings so they are more meaningful. You can try these new variables in your models to see if they perform any better than the original variables.
   4. Feel free to do any other variable manipulation that you think will be helpful for your models predicting price and overall condition high.
   5. Create two partition variables – both with 75% training and 25% validation. One will be fixed random and one will be stratified on OverallCond\_High.
2. **Exploratory Analysis**
   1. Include summaries of your data specifically examining those variables that you are using in your two models. These summaries should specifically try to show how the variables are related to Price and Overall Condition High (or Overall Condition as a proxy)