1. Name: Kyle Jones (Project Group 109)

2. Project Description: The Ames housing dataset consists of 2,891 property sales records in Ames, Iowa between 2006 and 2010. Each sales record has up to 80 variables that describe the quality and quantity of physical attributes of the property, which could be directly related to the sale price. The primary goal of this project is to construct a "best fitting" model for predicting housing prices, with "best fit" evaluated by Root-Mean-Squared-Error (RMSE) between the logarithm of the prediction and the logarithm of the actual price. (Taking logarithms ensures that errors on cheap and expensive houses have equal weight on model selection.) A secondary goal of this analysis is to determine which of the attribute variables hold substantial weight in the prediction of sales price. Perhaps the most unique characteristic of this housing dataset relative to other housing datasets is the plethora of predictors available for each sale record. It will be interesting to see how the more unique characteristics (e.g., zoning classification) stack up relative to more traditional characteristics (e.g., square footage).

3. Dataset: This is a Kaggle competition dataset, see link <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>.

4. Scientific Research Questions:

a) Which statistical learning method discussed from class results in lowest log RMSE?
   - As mentioned in the project description, the primary goal of this analysis is to produce a model with low prediction error. I plan to explore the following statistical learning methods from class: linear and regularized regression (lasso and ridge), KNN, decision tree, support vector regression, random forest, and AdaBoosting. Each individual technique will be optimized using cross validation on the train set. The final model may end up being one of these optimized techniques, or a combination (ensemble method).

b) Which predictors are associated most with sales price?
   - A secondary goal of this project is to identify important predictors of sales price among the large set (80) of predictors available, with special attention to those not already known to be great predictors of home price (e.g., square footage). I plan to use lasso regression and random forest methods to measure feature importance.
   - I also hope to understand the relationship between these important predictors (linear or not linear) and sales price.

c) If time permits, I'd also like to explore potential interaction effects between predictors. In other words, is the effect that a predictor has on sales price dependent on the value of another predictor?

   - An obvious variable in this dataset that may have an interaction effect with many predictors is location. For example, perhaps sales price is more sensitive to square footage in communities where property space is limited? I plan to create interaction terms in a regression model to test for possible interaction effects.