**I. Data Preparation Questions**

1. What specific data transforms did you perform prior to exploration and analysis, and why did you choose these?
   1. I deleted the PID column, since this is not useful to the data. I also used dummy variables for categorical values, without trimming the “NA” variables first since they can be useful information as well.

**II. Exploratory Analysis Questions**

Perform an exploratory analysis on your data by visualizing and/or applying other means of data exploration.

1. What (if any) insights jump out at you?
   1. Most of the variables seem to have a pretty strong correlation to the response variable. This makes sense since all elements of a house are important to the price.
2. Do you have any hypotheses about relationship of certain variables to the price?
   1. Street being gravel will likely decrease the price. Certain neighborhoods will obviously be cheaper than others. Conditional will have a large impact on price. Newer houses will be more expensive typically. Larger square footage houses will be more expensive. More bathrooms will indicate a larger house and therefore a higher price. Finished garages will be more expensive than unfinished garages. Pools will greatly increase the price of the house. Sale condition will have a reasonable influence on the price.

**III. Model Building**

First construct a baseline model (containing all predictors) to predict the price. Then build the best model you can devise. In this part use ONLY dataset A and DO NOT TOUCH dataset B. You will want to split this into training and test sets and apply error metrics/compare models only on the test data.

1. What approach did you use to arrive at the best model? Why did you select this approach?
   1. I used a lasso model with an alpha of 5 to create my model, since this will remove the most features that are not predictive of the response. I could have tuned the alpha for even higher R2 scores, but I decided this would likely be overfitting to the data.
2. Which error metric(s) are you using to compare performance? What is the value(s) of the error metric(s) for the baseline model and your best model?
   1. I am using the R2 score as well as the explained variance score, since I found the mean squared error and median absolute error to be too large for differences between them to be significant. The R2 score for the base model is .897, while the R2 score for the lasso model is .912.

**IV. Predicting and Validating**

Run your baseline and best models on dataset B. DO NOT do any further training. Remember to apply all transforms you used in building the model to this set (use the transform function on the preprocessors you created in part I).

1. What are the respective error metric values for each model on this set? How did your best model do on this data as compared to the baseline?
   1. I was unable to get my model to perform on the second data set without refitting it since the shapes were not the same? I am unsure why this happened.
2. Is your best model a good model? Why or why not?
   1. I would say it is not a bad model, since it seems to do well with the error metrics on the test data. However, it could be improved by dropping features that correspond to one another. For example, if Pool and Poolsize both exist, the Pool variable is essentially being counted twice.