Subjective Questions

Question 1: What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

The optimal values of alpha determined are 100 (Ridge) and 500 (Lasso). Doubling the value of alpha would result for Ridge, on a better R2 score on the test set and reduction of the residual error and for Lasso R2 score would improve on test set but the residual error would increase. The change would not alter the most important predictor variables on Ridge but on Lasso would reduce the importance of the variable YearBuilt.

Question 2: You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Based on the results (present at the jupyter notebook file) I would use Ridge Regression, because the R2 score on the test set is much better while the residual error is similar. Ridge and Lasso R2 scores on the training data are very similar but Lasso model is clearly overfitting resulting on a R2 score of 0.47 on the test set.

Question 3: After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Question 4: How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

A model is robust and generalized if it performs well on unknown data, meaning that it performs well on training and test data. Models with high number of coefficients tend to be more complex and tend to overfit, hence have a bad R2 score on test data when comparing with R2 score on training data. A robust and generalized model is a model with a good fitting, thus identifying all the patterns. Regularization can be used to ensure the model is more robust and generalizable, it brings the coefficients to zero or removes them sacrificing bias but reducing the variance and this as implications on the accuracy of the model but reduces the overall error.