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?

Answer1:

- Somewhere around 0.0001 to 100 should look decent value. Lower the best.
- The difference between testing data and training data R^2 value will decrease. Bias of the model will increase and the variance of the model will decrease. It will also change the co-efficient values.
- GrLivArea, OverallQual, TotalBsmtSF, GarageArea, LotArea, OverallCond, Neighborhood_Crawfor.

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?

Answer:

Looking at the r2 figures, the difference in the test & test scores are not large (just in fraction), so choosing Lasso as it will be more robust. And it has feature elimination that some of the co-efficients will be 0.

Question 3

After building the model, you realised 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?

Answer 3:

FullBath, GarageArea, LotArea, Fireplace, HalfBath

Question 4

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

The model should not be memorized all the data completely. The more data is clean (EDA is done correctly) then more generalizable, robust it can be. Trade-off should be done. If the outliers are handled properly, multi-collinearity variable is been treated, then we can make sure that the model is robust and generalizable.

Test and training accuracy should have no much difference. If the data is biased then we ignore accuracy score. If not, the model should be cleaned the important variables outliers accordingly. Variables that doesn't contribute to the target variable should be ignored from the dataset. The accuracy on the training data should not be 100% so that on the test data it will score lower.