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?

Answer:
- Optimal value of alpha for Ridge is 10 and for Lasso is 0.0004.
- If value of alpha is doubled in Lasso then, more features will have a coefficient of 0.
  If Alpha is doubled in ridge, the impact of shrinkage grows and the coefficients tends to zero but not exactly 0.
- The most important predictor variables after the change is implemented will be GrLivArea

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:

• The optimal lambda value in case of Ridge and Lasso is as below:

♣ Ridge-10
♣ Lasso-0.01

• The Mean Squared error in case of Ridge and Lasso are:

- The Mean Squared Error of Lasso is slightly lower than that of Ridge.
- Also, since Lasso helps in feature reduction (as the coefficient value of one of the feature became 0), Lasso has a better edge over Ridge.

Therefore, the variables predicted by Lasso can be applied to choose significant variables for predicting the price of a house.

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: We have to re-created model and test it. We cannot generalize saying, if first 5 are not available then next 5 will be most significant predictor variables. But most probably they will be next important predictor variables

BsmtFinSF1	6885.556965
TotalBsmtSF	6850.114544
d_BsmtExposure	5970.459291
Neighborhood_Crawfor	5245.345842
d_ExterQual	

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?

Answer:

According to Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons:-

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust.

   Complex models tend to change wildly with changes in the training data set
   Simple models have low variance, high bias and complex models have low bias, high

   variance

- Simpler models make more errors in the training set. Complex models lead to overfitting —

they work very well for the training samples, fail miserably when applied to other test samples

Therefore to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

Also, Making a model simple leads to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naive, for

Variance refers to the degree of changes in the model itself with respect to changes in the training data.

Thus accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph