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 – The optimal value of alpha for **Ridge Regression = 10** & **Lasso Regression = 0.001**

Effect of doubling the value of alpha for the Ridge Regression, the model will apply more penalty on the curve, trying to make model more generalized therefore making model more simpler and not thinking of fitting every bit of the dataset.

Similarly for Lasso Regression, doubling the value of alpha will try to penalize more and more coefficients of the variables will be become zero eventually, along with the increase R2 square also decreases a bit.

So new alpha value for **Ridge = 20** & **Lasso = 0.002**

The most import variable after the changes has been implemented for are as follows -

GrLivArea

OverallQual_8

OverallQual_9

Functional_Typ

Neighborhood_Crawfor

Exterior1st_BrkFace

TotalBsmtSF

CentralAir_Y

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 – Regularizing coefficients and improving the prediction accuracy with decrease in the variance also making model more interpretable.

We can choose one depending upon the following conditions -

• The model which we need to be choose to apply solely depend upon the given use case.

• If there are too many predictor variables and one of the primary goal is feature variable selection, then we can go with Lasso Regression Model.

• If there aren't large no coefficients available and reducing the coefficient magnitude is one of our important task, in that case we choose Ridge Regression Model.

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?

Answer – Initially found following most important predictor variables –

GrLivArea

SaleCondition_Partial

Neighborhood_Crawfor

OverallQual

SaleCondition_Normal

After removing above predictor variables and building model again got the following top 5 predictor variables –

1stFlrSF

TotalBsmtSF

Neighborhood_Sawyer

BsmtFinSF2

Neighborhood_Timber

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?

Answer – A model is known to be robust when any variation in the dataset doesn't affect its performance at large scale.

• A model is known to be generalizable model when it is able to adapt properly to new dataset, previously unseen data, drawn from the same distribution as the one used for the creation of model.

• We need to take care of the model that it won't overfit to remain as robust and generalizable model. As the overfitting model have very high variance and the smallest change in the data affects model prediction drastically. Such a model may be able to identify all the patterns of a training data, but unable to determine the patterns in unseen test dataset.

• Thus the model shouldn't be too complex in order to be remain as robust and generalizable.

• From the Accuracy point of view, a too complex model will have a very high accuracy. Hence to make our model more robust and generalizable, we have to minimize variance which will lead to some bias. Addition of the bias means that the accuracy will decrease.

• Therefore in general we have to adjust some balance between model accuracy and it's complexity, which can be achieved by Regularization techniques like Ridge and Lasso Regression.