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?

When we plot the curve between negative mean absolute error and alpha for ridge regression, we observe that the error term decreases as alpha increases from zero, and the train error shows a rising trend. We chose to use a value of alpha equal to 2 for our ridge regression since the test error is lowest when alpha is equal to 2.

We have chosen to maintain a very low value for lasso regression, which is 0.01. As alpha increases, the model attempts to punish more and attempt to make the majority of the coefficient values zero. Alpha and a negative mean absolute error of 0.4 were the initial values.

The model will apply more penalty on the curve and attempt to become more generic when the value of alpha for our ridge regression is doubled. By doing this, we make the model more straightforward and stop trying to fit all of the data in the data set. For ridge changing alpha 2 to 4 reduces accuracy from 92.29 to 92.05. Similar to how increasing the lasso's alpha penalizes our model more and causes more coefficients of the variable to be reduced to zero, the r2 square results is reduced in both. In lasso changing alpha from 0.01 to 0.02 reduces accuracy from 86.20 to 84.45. below are the 5 important variables post double of alpha value.

Lasso: 'OverallQual', 'GrLivArea', 'GarageArea', 'OverallCond', 'Fireplaces'

Ridge: Neighborhood_Crawfor, Neighborhood_StoneBr, MSZoning_RH, OverallQual, MSZoning_FV

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?

Ridge regression, which employs cross validation to identify the penalty is square of magnitude of coefficients, requires a tuning parameter called lambda. By applying the penalty, the residual sum or squares should be minimal. The coefficients with higher values are punished because the penalty is equal to lambda times the sum of the squares of the coefficients. The variance in the model is lost when we raise the value of lambda, but bias stays constant. In contrast to Lasso Regression, Ridge Regression incorporates all variables in the final model.

The penalty in lasso regression, which is determined through cross-validation, is the absolute magnitude of the coefficients. This tuning parameter is termed lambda. As lambda grows, Lasso decreases the coefficient toward zero, bringing the variables' values exactly to zero. Variable selection is also done via Lasso. When lambda is small, the model performs simple linear regression; however, as lambda rises, the model shrinks and ignores variables with a value of 0.

Regularizing coefficients is crucial for increasing prediction accuracy, reducing variation, and making the model understandable. As lasso model is simpler and more robust and has similar train and test accuracy that will be the choice

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?

In the first run we created the lasso using all variables and the first 5 most important variable where 'OverallQual', 'GrLivArea', 'OverallCond', 'GarageArea', 'BsmtFullBath'. We again created the model after removing these 5 columns from the dataset and observed reduced accuracy values also the most important variable post deleting above variables are MSSubClass, LotFrontage, LotArea, MasVnrArea, BsmtFinSF1.

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?

The model has to be extended to ensure that test accuracy is equal to or greater than training score. Other datasets beyond the ones used during training should yield valid results from the model. In order for the model to forecast data with high accuracy, the outliers shouldn't be given an excessive amount of weight. Only outliers that are pertinent to the dataset should be kept once the outlier’s analysis is completed to make sure that this is not the case. It is necessary to eliminate from the dataset any outliers that do not make sense to preserve. The model cannot be trusted for predictive analysis if it is not robust.