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?

Ans:

For ridge regression: The curve between negative mean absolute error and alpha show that the value of alpha increase from 0 the error term decreases and the train error is showing increasing trend when value of alpha increases. For alpha value 2 the test error is minimum so the value of alpha equal to 2 for our ridge regression is best fit. For lasso regression: The small value that is 0.01 is set in my analysis, when we increase the value of alpha the model penalizes more and make most coefficient value zero. Initially value is 0.4 in negative mean absolute error and alpha.

When we double the value of alpha for our ridge regression no we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set .from the graph we can see that when alpha is 10 we get more error for both test and train. Similarly when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

The most important variable after the changes has been implemented for ridge regression are as follows:-

1. MSZoning_FV 2.
2. MSZoning_RL

3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial \
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The most important variable after the changes has been implemented for lasso regression are as follows: -

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontag

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?

Ans:

Lasso regression, because it tending towards simple model which is good at longer run and allows to select features more effectively by a tuning parameter called lambda as the penalty. With increase in lambda value Lasso shrinks the coefficient towards 0 and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the 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?

Ans:

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

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?

Ans:

The model should be as simple, accuracy may be decrease but it is more robust and generalizable. Considering the Bias-Variance trade-off. The simpler the model have more bias but less variance and more generalizable. Model perform well on both training and test data which is like. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data