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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans:

Lasso regression optimal value of alpha is 50

Ridge regression optimal value of alpha is 4

Doubling alpha -LASSO: -

Alpha = 50,

R2 for train: 0.9372405328256925

R2 for test: 0.92546641 23086983

Double alpha in lasso i.e., 100.

R2 for train: 0.9314089232702133

R2 for test: 0.9231590633923132

Doubling alpha -RIDGE:

alpha = 4,

R2 for train: 0.9371096095852764

R2 for train test: 0.9253982765709685

double alpha = 8,

R2 for train: 0.9254099264105871

R2 for test 0.9182570406142743

The most important predictor variables after we double the alpha values are:

- GrLivArea
- OverallQual

- Functional_Typ
- Neighborhood
- Exterior1st_BrkFace
- TotalBsmtSF
- CentralAir

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:

- Observing the result of both the models, Lasso has better values of R2.
- If we have too many variables and one of our primary goals is feature selection, then we will use Lasso.

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?

Ans:

Will remove - 'OverallCond','GarageCars','BsmtFinSF2','OverallQual','CentralAir'

Predictors after removing – 'OverallCond', 'house_age', 'TotalBsmtSF' , 'OverallQual', 'GrLivArea'

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:

- A model is robust when any variation in the data does not affect its performance much.
- A generalizable model can adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

- To make sure a model is robust and generalizable, we must take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data but fail to pick up the patterns in unseen test data.

- In other words, the model should not be too complex to be robust and generalizable.If we look at it from the perspective of Accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.
- In general, we must find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.