Problem Statement 2.

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:-
In our case for:
Ridge regression we see that the best/optimal value for alpha is 5.0
Lasso regression we see that the best/optimal value for alpha is 0.001

Doubling the value of alpha for ridge regression:-
The model will apply more penalty on the curve, and try to make the model more generalized which is making the model more simpler.

Doubling the value of alpha for lasso regression:-
When we try to increase the value of alpha the model try to penalize more and try to make most of the coefficient value 0

Predictor variables when alpha for Ridge is 10:-
BedroomAbvGr
HalfBath
FullBath
LandSlope
Neighborhood_BrkSide
BsmtFullBath

Predictor variables when alpha for lasso is 0.002
BedroomAbvGr
FullBath
HalfBath
BsmtFinSF1
Neighborhood_Gilbert
LandSlope
BsmtFullBath

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:-
Lasso regression would be a better option as it would help in feature elimination and the model will be more robust.

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:-
The five most important variables are -
GrLivArea
Street_Pave
RoofMatl_Metal
RoofStyle_Shed
11stFlrSF

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:-
The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.