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?
A :
Ridge :

R2 score for training data is 91.00%
R2 score for testing data is 90.04%
MSE for training data is 0.013853065834995766
MSE for testing data is 0.016969995795888454
optimum alpha value for alpha is 0.9

Lasso :
R2 score for training data is 91.23%
R2 score for testing data is 90.02%
MSE for training data is 0.013498806998017927
MSE for testing data is 0.01699533660197611
optimum alpha value for lasso is 0.0001

After doubling :
#alpha for ridge = 1.8
#alpha for lasso = 0.0002

Lasso :
training score is 90.94%
testing score is 90.14%

Ridge :
training score is 90.77%
testing score is 90.06%

Most important predictor variables after the change is implemented is :
GrLivArea
OverallQual
1stFlrSF
PoolQC
OverallCond
LotArea
GarageCars
Neighborhood
ScreenPorch
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?

A : Training and testing scores for lasso and ridge are almost similar , lasso performs slightly better than ridge , since lasso eliminates certain features , it's better to go with 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?

A: TotalBsmtSF,TotRmsAbvGrd,Neighborhood,LotArea,OverallCond

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?
Model should always be robust and generalisable so that it can work well and predict on unseen data too .In order to have test accuracy close to training accuracy , the model should perform well on unseen data. To achieve this , we should ensure that too much weightage is not given to outliers , if there are outliers then it needs to be analysed well and only non-significant ones has to be removed from datasets and the significant ones needs to be retained. If the model is trained well on outliers then it will not perform well on unseen data leading to low test accuracy score compared to training score.