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:
The optimal value of alpha is 10 for Ridge Regression and 0.001 for Lasso Regression.

If I double the value of alpha,the R2 score is reduced a bit for Ridge Regression whereas it reduced significantly for the Lasso Regression.
With double the alpha,Train and Test accurecy are similar incase of Ridge Regression but for Lasso, model performance on test data reduced significantly.
R2_Score_Train_Data R2_Score_Test_Data
Ridge(alpha=10) 83.28% 81.04%
Ridge(alpha=20) 83.27% 80.50%
Lasso(alpha=0.001) 81.08% 76.88%
Lasso(alpha=0.002) 79.22% 72.66%

For ridge regression, even if we double the alpha, the top 5 variable and their coeff value remain same but in lasso, with higher alpha, coeff value of the variable changes a lot.
So basically, as we increase the alpha, the bias will steadily increase and the variance will steadily decrease since the model is becoming less flexible as model coefficients are restricted.

After the change, please find below the top 5 important variable out of both Ridge and Lasso Regression.
Ridge:
OverallQuall
TotalFloor_excluding_basement
TotalBathroom (It is a calculated column which sum up all the bathrooms across all the floors)
TotalFloor_excluding_basement
KitchenQual

Lasso:
OverallQuall
TotalFloor_excluding_basement
KitchenQual
TotalBathroom
BsmtQual_5

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?

For Ridge Regression, we have the following R2 scores on the training set for different value of alpha.
Alpha R2 Score
0 86.04%
0.001 86.05%
0.01 86.04%
0.1 86.03%
1 85.77%
10 83.28%
100 67.20%

1000 21.94%
We can see as the value of the alpha increases, R2 score decreases as well. I will choose the alpha=10 as
this gives R2 score as 83.28% which 2.77% lesser than the best R2 score but at the same time it has alpha 10000 times higher. Higher the alpha, simpler the model would be.
For Lasso Regression, we have the following R2 scores on the training set for different value of alpha.
Alpha R2 Score
0 86.05%
0.001 81.08%
0.01 30.64%
0.1 0%
1 0%
10 0%
100 0%
1000 0%
For Lasso, R2 Score is highest when no regularization applied. With a little alpha still we have R2 score more than 80% whereas further increasing alpha decreasing R2 score to 0. So I will choose alpha=0.001.

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?

TotalBsmtSF
GarageCars
BsmtQual_5
MasVnrArea
Neighborhood_Crawfor

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?

A model is robust when its model parameters or model coefficients do not change much when encountered with new sample and generalization is something to do with the predictive power of the model for different samples from the same population.
A model with good generalization should have a consistent predictive power on the unseen samples.
For our model to be simple at the same time should have a good predictive power.
The goal of an ideal model should be low bias and low variance though it is difficult to achieve.
However,Regularization helps us to find that sweet spot by balancing bias and variance. It enables a model to perform well on unseen data at the same time to identify the underlying patterns present in the data. We try to maximize alpha without sacrificing model performance much so that we can get a model with decent predictive capabilities.

In order to make it more robust,as we increase alpha, it shrinks coeffcients towards zero hence the variance starts decreasing and the bias start increasing. More the simpler the model, less accurate it would be. So the implication of regularization on model results in less accuracy.

And if we want to make it understand the data pattern more accurately, it will be high variance with low
bias and model accuracy will be higher whereas it may perform bad on the unseen data.