SUBJECTIVE QUESTIONS

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: Optimal value of alpha for ridge regression is **5** and for lasso regression is **0.0001**. If we choose double the value of alpha the bias in the model will increase and train score will reduce. In our case in python notebook we see that train accuracy has dropped by small amount in case of both ridge and lasso. The number of parameters in ridge regression will remain same, however the number of parameters in lasso regression may decrease.

   Most important predictor variable after the change is implemented in case of ridge regression is as follows

   GrLivArea
   1stFlrSF
   OverallQual_10
   TotalBsmtSF
   2ndFlrSF

   In case of Lasso are as follows

   GrLivArea
   OverallQual_10
   TotalBsmtSF
   OverallQual_9
   Neighborhood_NoRidge

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: ridge regression uses square of betas as the penalty term which decreases the value of betas in case of overfitting, however it doesn't remove any predictor variables.

   Lasso regression chooses absolute value of betas as the penalty term and hence it can remove some predictor variables by making betas equal to zero. Since our aim here is to interpret the variables and decide which ones are better, we go with lasso regression since it is more simple and has almost same accuracy as ridge in our case.

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: For this we create lasso regression excluding the five most important variables we identified. The five most important variables in earlier model are (with alpha 0.0001)

GrLivArea
OverallQual_10
TotalBsmtSF
OverallQual_9
2ndFlrSF

When we remove these and build the model, the most important predictors are,

1stFlrSF → positively related
OverallQual_3 → negatively related
OverallQual_5 → negatively related
OverallQual_4 → negatively related
OverallQual_6 → negatively related

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: To make sure the model is more robust and generalizable we apply regularization so that the model doesn't overfit the train data and perform poorly on test data. The accuracy of train data may fall but the accuracy of test data will greatly increase. That is we compromise a small amount of bias for a large decrease in variance.