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 1

Lasso with alpha 0.0001 shows the best R2 on Test Set. Optimal alpha for Ridge is 0.7

Observations after doubling Lasso alpha:

The top 5 predictors remain same though their odering has changed. Unlike Ridge where odering remained same

All the beta coeffecients have reduced.

R2 for both Test and Train set have reduced. This shows we moved away from optimal alpha.

Top predictor variables selected by Lasso regression with double the optimal alpha:

OverallQual	0.219316
1stFlrSF	0.210510
GarageArea	0.093166
OverallCond	0.091291
2ndFlrSF	0.087327

Observations after doubling Ridge alpha:

Top 5 predictors remain same

Coefficients of the variables reduce. As expected, increasing alpha shrinks the coefficients towards zero.

R2 for both Test and Train set reduces. This shows we moved away from the optimal aplha value.

Top predictor variables selected by Ridge regression with double the optimal alpha:

OverallQual	0.192935
1stFlrSF	0.155383
OverallCond	0.100637
GarageArea	0.091311
2ndFlrSF	0.081972

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 2

We will apply Lasso as its R2 score in the test set [0.8722] is highest, indicating that the Lasso model is able to explain the variance in Sale Price set better then Linear and Ridge models. We have also verified the predicted Sale Price from the model vs actual and observed a linear graph with positive gradient. Error terms are normally distributed with mean at zero. These observations give us confidence that the Lasso model with alpha 0.0001 is the best model.

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?

Answer 3
After removing the original top 5 predictor variables and re running the lasso model we get the following new top 5 variables:

```
TotRmsAbvGrd 0.173301
BsmtFinSF1 0.102449
FullBath 0.093912
LotArea 0.088911
GarageQual 0.084999
```

Observation: The new top 5 are not exactly the next 5 from the original prediction though we still have three out of the next 5. This shows removal of variable impacts coefficients of the remaining variables. R2 scores have reduced indicating that model is now underfitting.

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 4
We first clean up the data to remove irrelevant data. We converted Ordered variables into numerics so we can see the correlations. We converted nominal variables in to dummies so we have binary representation of otherwise text data. We also salted the numerics data to avoid bloating in the coefficients.

We ran a linear regression to get R2 scores. We ran RFE to see if auto selection of parameters improves the scores. We found that the original linear regression was giving better results so we moved to advanced regressions techniques to regularise the model.

We used k-fold cross-validation in the modelling to assess the model's performance across different subsets of the data. This helps in estimating how the model might perform on unseen data and reduces the risk of overfitting. We also used regularisation techniques like Ridge and Lasso to penalise large coefficients and avoid overfitting with the train data. We tried a wide range of hyper parameter, Alpha on both Ridge and Lasso techniques to come up with the optimal value. We then compared the R2 scores of Linear , Ridge and Lasso models to finalise the model.

We train the model on train set, generate the predictions on the test set and then compare the predicted values with actual value sin the test set. If we get a linear graph we know that our predicted values are in line with the actual values in the test set. This despite the fact that the model was never trained on the test set. This tells us that the model is able to do reasonable predictions on an unseen set of data. So the processes mentioned above helped us create an accurate model.