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:

Optimal value of alpha for ridge regression is - 10.0

Optimal value of alpha for ridge regression is - 0.001

There will very minor change in model if we choose double the value of alpha for both ridge and lasso.

If we multiply or divide the lambda by 10/100/1000 then there will be big effect.

If lambda value is too high, model will be simple, but there is risk of *underfitting* data. Our model won't learn enough about the training data to make useful predictions.
If lambda value is too low, model will be more complex, and we run the risk of *overfitting* data. Our model will learn too much about the particularities of the training data, and won't be able to generalize to new data.

The most important predictor variables after the change is implemented are:

OverallQual

CentralAir

KitchenAbvGr

GarageCars

FullBath

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:

The result provided by Lasso and ridge regression are almost same in terms of R-squared, root mean squared and root mean squared error.

Lasso regression result on test data:

```
R2 = 0.8412076852196503
RSS = 11.79906616804327
RMSE = 0.0269385072329755
```

Ridge regression result on test data:

```
R2 = 0.8424580958940577
RSS = 11.706154377538986
RMSE = .026726379857394945
```

As test results are almost same we can choose any one.

I will choose the Lasso regression model as there are a large number of features in the provided dataset. And Lasso makes a feature selection too along with regularization by making some coefficients to zero. And hence eliminating some features.

Lasso predicts with less number of identifiers.

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:

After dropping the 5 most predictor variables from the dataset, if we rebuild the model then its performance is significantly dropped.

Lasso regression result on test data:

```
R2 = 0.6976040536011043
RSS = 22.469536925915314
RMSE = 0.05130031261624501
```

The 5 most important predictor variables now are:

Fireplaces

BsmtQual

BsmtCond

PoolQC

BsmtFullBath

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:

For making the model robust and generalisable, there should be a bias-variance trade off.

The 'variance' of a model is the variance in its output on some test data with respect to changes in the training dataset. Variance here refers to the degree of changes in the model itself with respect to changes in the training data.

Bias quantifies how accurate is the model likely to be on test data. Complex models, assuming we have enough training data available, can do a very accurate job of prediction.
Models that are too naive are very likely to do badly.

The best model for a task is one that balances both - achieves reasonable degree of predictability (low variance) without compromising too much on the accuracy (bias).

There should be no underfitting and no overfitting.

We should make model simple, not complex. And to make model simple but not too simple, we can use regularization.
In linear regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.