Problem Statement

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?

The optimal value of alpha for Ridge is 100 and Lasso is 0.01, If we double the value of alpha for ridge and lasso the variance increases where r2 value slowly goes on decreasing and in lasso the coefficients of important predictor variables increased but as the value of alpha increase all the predictor value tends to zero

import predictor variable

Lasso : - OverallQual, GrLivArea, , GarageCars, OverallCond, TotalBsmtSF.

Ridge : - OverallQual, OverallCond, GrLivArea, 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?

I choose Lasso because as Lasso gives me significant coefficients value and other remaining the tends zero and its RMSE value is okay with test data as compared to Ridge.

Lasso RMSE: Train- 0.011263837206319304, Test - 0.011723196243577996

Ridge RMSE: Train- 0.008574982914362565, Test-0.010485780656134683

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?

The other five important variables excluding the first important predictors are

0.03647798072575451, 'TotRmsAbvGrd'

0.045329589756128684, 'Fireplaces'

0.06796569055620952, 'GarageArea'

```
 0.07680565208861224, '2ndFlrSF'

 0.10471058447242536, '1stFlrSF'
```

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?

Ensuring that a model is robust and generalizable is crucial to its performance and reliability. A robust and generalizable model performs well on both the training data and unseen or new data.

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.