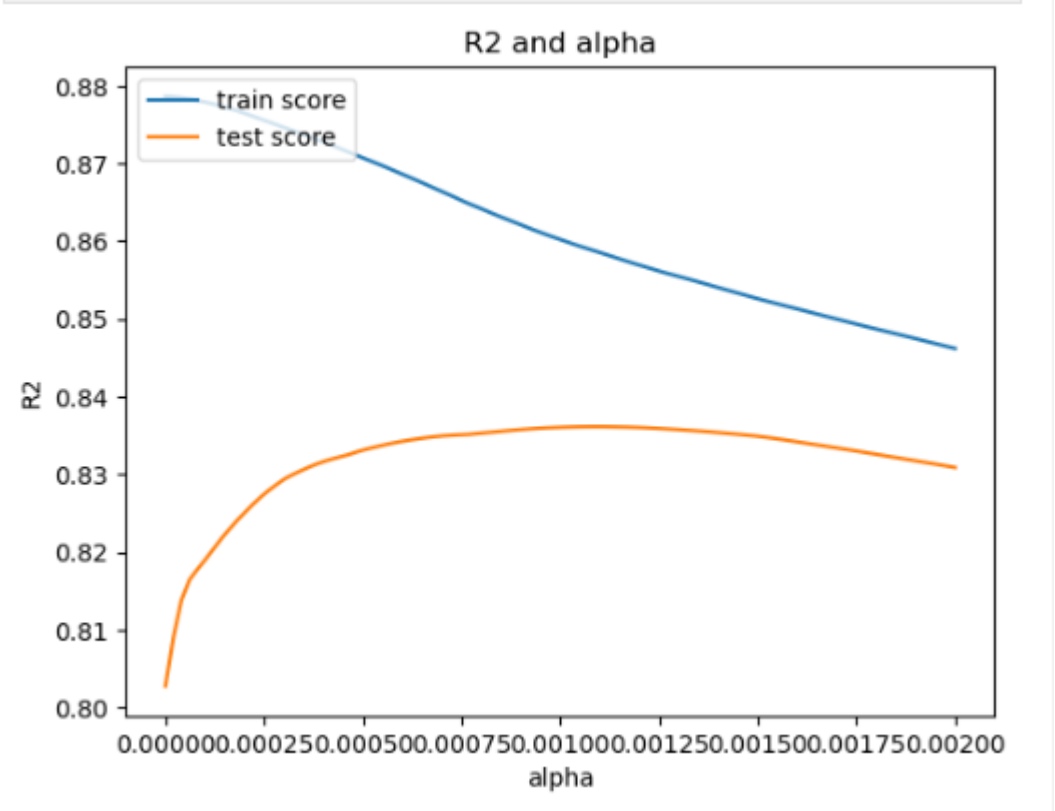
Question 1

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?

0.00075 for ridge and 0.001 for lasso were the optimal values for ridge and lasso respectively.



Upon doubling the alpha, the r2 scores decrease more for both train and test data. ***OverallQual***(0.77) and ***GrLivArea***(0.72) are the most significant variables.

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?

Lasso seems to give better results in terms of r2 score and also have a fairly explainable significance for the coefficients.

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 top most important variables are:
['PoolQC', 'OverallQual', '1stFlrSF', 'GrLivArea', 'OverallCond']

After removing these 5, the next top 5 are these:
['RoofStyle', 'BsmtCond', 'MSSubClass', 'MSZoning', 'LotArea']

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?

Having the training data be a good representation of the possible real world scenarios is a good starting point to make sure we have a robust and generalisable model. Making sure model doesn't tend to overfit the training data, Have a train, test and validation dataset to iteratively train and not fully expose the test data can be more ways to do so.