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?

Answer1:

Ridge :
Optimal value of alpha: 0.3

Lasso:
Optimal value of alpha: 0.0001

If we double the value of alpha, the complexity of the model will reduce which may over simplify the model and it will not perform well for train or test data. Specifically for this case, when I doubled the alpha for both ridge and lasso regression and checked the R2_score, there was a slight decline in accuracy for both train and test data.

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?

Answer2:
I chose to apply the Lasso regression because the performance of both ridge and lasso are quite similar in terms of R2_score. Lasso has some advantages over ridge like feature selection and ease of understanding due to less number of features. Hence, I chose Lasso.

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: The five most important features initially were:

Feature

'GrLivArea', 'OverallQual', 'MasVnrArea', 'GarageCars', 'BsmtFullBath'

I removed the above from the dataset and create the model again using Lasso, now the top five important features are

Feature Coefficient:

1stFlrSF	0.316191
2ndFlrSF	0.162403
TotalBsmtSF	0.084614
BsmtFinSF1	0.077619
GarageArea	0.075036

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 a model to be robust and generalisable , it should not be overfitting the training data. An overfitted model and highly complex model has a very high variance and a small change in data can affect the performance heavily. Such a model will have high accuracy on training data but will have a very low accuracy for test data.

In other words, we need to reduce the complexity of a model to increase the robustness and generalisability , in order to do so , we can use regularisation techniques.