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: The optimal value of alpha for ridge regression is 100 . If we double or keep on increase the value of alpha the model is simplified but the negative mean square error increases and the accuracy decrease

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: Based on the metrics, Ridge and Lasso Regression exhibit similar performance with high R2 scores, low RSS, and MSE. The choice depends on priorities: Lasso for feature selection and interpretability, Ridge for a balanced approach. Given their comparable results, considerations like model simplicity and interpretability may guide the selection.

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: The top five variables excluding the ones not available in the incoming data:

1. Street (Coefficient: 0.107373)
2. CentralAir(Coefficient: 0.064317)
3. OverallQual (Coefficient: 0.033311)
4. GarageCars(Coefficient: 0.032937)
5. BsmtFullBath(Coefficient: 0.036703)

These variables, based on their absolute Lasso coefficients, are considered the most important in the revised model excluding the five variables not available in the incoming data.

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 5: As the model is more generalizable the model becomes simple robust and explainable but there is a tradeoff that it may not be so accurate. that is there is always a tradeoff between the bias/variance. More generalized the model is the better it can used in different domains and can perform fairly well with an unlearnt dataset.it also ensures robustness and stability If the model has high variance it can be highly accurate for a specific dataset and a specific domain And the model does not perform well for unseen data and cannot be used for general use. The idea of regularization based regression techniques is to balance between the simplicity/explainabilty of the model and its accuracy and variance