SUBJECTIVE QUESTIONS

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?

Ridge: 100

Ridge after RFE: 20

Lasso: 0.001

For all the models, the training score has decreased slightly and the testing score has increased slightly. The change is most noticeable for Ridge after RFE. Here, the changes were the largest, so that the gap between train and test data is the smallest.

'GrLivArea', 'YearBuilt', 'OverallQual', 'BsmtUnfSF', 'OverallCond' are the most important predictor variables

Important predictors Lasso: Double Alpha after removing most important predictor variables:

GrLivArea	0.152414
TotalBsmtSF	0.059183
GarageCars	0.043507
YearRemodAdd	0.031122
SaleCondition_Partial	0.030626

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?

The r2 score is slightly higher for Lasso and the gap between training and testing is slightly lower. Hence, I would choose lasso. Lasso helps in reducing the features in the model, helping to create a simpler final model. This is important for creating a robust and generalisable model, as discussed in question 4. It also has the lowest residual sum of squares of all of the created models.

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?

After the top 5 features are dropped the next most important predictors become
RoofMatl_CompShg,RoofMatl_Tar&Grv, RoofMatl_WdShngl, RoofMatl_WdShake, GrLivArea

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?

A model can be considered robust and generalizable if it is shows no drastic change in performance when the training set is changed, i.e. the model should not overfit on the training data and should be able to handle new/unseen data properly. When it comes to accuracy, a model which is robust and generalizable should perform equally well on both the training and test data