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', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'YearBuilt' are the most important predictor variables

Important predictors Lasso:

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. 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?

The top predictor variables are: 'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'YearBuilt'

GrLivArea 0.128692
OverallQual 0.065629
TotalBsmtSF 0.048505
OverallCond 0.043846
YearBuilt 0.043804

After removing them, the predictor variables are:

GrLivArea 0.146928
MSZoning_RL 0.083526
MSZoning_RM 0.065606
TotalBsmtSF 0.056324
GarageCars 0.043311

The predictor variables remain the same, with a slightly different order for Lasso applied to the RFE created variables.
GrLivArea 0.136432
MSZoning_RL 0.083509
TotalBsmtSF 0.073250
MSZoning_RM 0.069320
GarageCars 0.041585

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?

Generalisation is important and test accuracy needs to be higher than the training score. However, the difference should not be exceedingly high. The model should generalise during training, but if the score is very high in training, and lower in testing, it means that the model has memorised the data, meaning that it is overfitting. Overall, there should not be large differences between the results. Low test scores could result from splitting the data set too early in the preprocessing step, so that some steps may be missed on the test data. Robustness of a model is generally not solely based on high test scores, but also depends on the assumption that the train scores are higher than the test scores. Both scores have to be high enough to be acceptable for the specific business case and expectations of the model. It is also important to consider the values obtained for train and test, so that the model will perform well on unseen data. This means that the data should retain some outliers to help with predictions.