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?

Ans: The optimal value of alpha for ridge and lasso regression is 0.001 and 0.001.

The changes in the model if we choose double the value of alpha for both ridge and lasso then we are basically controlling the penalty term means higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients are reduced and it shrinks the parameters, therefore it is mostly used to prevent multicollinearity, also it reduces the model complexity by coefficient shrinkage.

The most important predictor variables after the change is implemented are TotalBsmtSF', 'Age of YearBuilt', 'Age of YearRemodAdd', 'Age of GarageYrBlt' and 'Age of YrSold'

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?

Ans: Lasso regression would be a better option it would help in feature elimination and the model will be more robust.

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?

Ans: The five most important predictor variables are xterQual', 'CentralAir', '1stFlrSF', 'GrLivArea' and 'BldgType_Duplex'

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?

Ans: A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalizable so that the test accuracy is not lesser than the training score. The model should also be made standard so that the accuracy of the test is not less than the training points. The model must be accurate to the data sets other than those used during the training. Too much weight should not be given to exporters so that the accuracy predicted by the model is high. To ensure that this is not the case, external analysis needs to be performed and only those that are compatible with the database need to be maintained. That last nonsense output should be deleted from the database. This will help to increase the accuracy of the predictions made by the model. Confidence intervals can be used (usually 3-5 deviations). This will help to measure the predictions made by the model. If the model is weak, it cannot be trusted with predictive analysis.