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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans) After performing GridSearchCV() for ridge and lasso optimal values found are

For Lasso – 0.001

For Ridge – 2.0

If we increase the value of alpha the coefficients will be penalised high and their magnitude gets reduced. For ridge it becomes smaller and smaller and for lasso they become zero.

After implementing the changes-

Lasso important features-

1.GrLivArea

2.OverallQual

3.GarageArea

4.OverallCond

5.BsmtQual

6.LotArea

Ridge important features-

1.GrLivArea

2.OverallQual

3.1stFlrSF

4.OverallCond

5.2ndFlrSF

6.LotArea

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) We need to choose an optimal value of alpha as if we choose a higher value of alpha the model becomes very simple and most of the features will be zero. So, the model wont perform well.

For Ridge regression the tuning parameter lambda which is square of magnitude of coefficients. The optimal value can be identified by using cross validation technique. The penalty is lambda times the sum of squares of coefficients. So higher the coefficient value the penalty for that feature will be high.

For Lasso the tuning parameter called lambda as the penalty of absolute of magnitude of coefficients. If increase the lambda value, the coefficients will become zero. Unlike Ridge, Lasso also does feature elimination.

I have chosen lasso model as it does feature elimination also.

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)

Lasso Top 5 important features-

1.GrLivArea

2.OverallQual

3.GarageArea

4.OverallCond

5.BsmtQual

After Removing them important features-

1.1stFlrSF

2.2ndFlrSF

3.LotArea

4.BsmtCond

5.KitchenQuality

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) If a model is simple then it is robust and generalisable. There is a trade of between bias and variance. Simpler models will have high bias and less variance and hence more generalisable. In terms of accuracy a robust model performs same on test and training data (there won't be much difference in RSquare and Adjusted RSquare)

Variance is how sensitive the model is to input data. If input data changes the model varies a lot. This is overfitting and happens if it is a complex model and complex models are not generalisable. Bias is error term high bias more wrong prediction both on training and testing data. So we need to trade off between these two to get a robust and generalisable model (simple but not to simple).