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?

There is no definitive optimal value of alpha (aka tuning parameter or hyper parameter) for RIDGE & LASSO regression. It's determined by trail & error method. Initially we assume a series of positive numbers between 0 & infinity eg. 0.0001, 0.001, 0.01, 0.1, 1, 10, 20, 30 etc. After that we pass these values one-by-one to GridSearchCV() function along with number of folds (typically it's 5). The best parameter is chosen from this exercise is used as 'alpha' in RIDGE or LASSO regression.

As the value of 'alpha' increases the model coefficients tends to decrease which makes the bias of the model to increase in other words the model's complexity reduces and consequently it may underfit the training dataset.

The importance of predictor variables in either of the regression models is directly proportional to the magnitude of the corresponding model coefficients. However, in case of LASSO regression, the model makes some of the model coefficients reduces to zero and hence the remaining predictors with non-zero model coefficients are significant.

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 optimum values for Ridge & Lasso regressions were obtained individually and are used for regression separately. However, in this assignment, in both the cases the 'lambda' was obtained as '20'.

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 5 predictors are as below:

GrLivArea
Condition2_PosN
OverallQual
RoofMatl_WdShngl
Neighborhood_NoRidge
```

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?

```
In general, the model should neither be too complex nor too simple. The
model complexity has to be at an optimum level such that the total erro
r i.e summation of BIAS and VARIANCE is least. A slight compromise on t
he BIAS to achieve significant reduction in variance is recommended to
get an optimum complexity while modelling.

During the process of optimizing the model, it is expected that the BIA
S is compromised slightly. To make sure the model is giving accurate re
sults on both the training & test data sets, it is advised to monitor t
he difference between R^2 values obtained for RIDGE & LASSO regressions
. The model for which this difference is less can be considered as best
model. Because the lesser the difference, fitter is the model. In other
words, if we find the R^2 value for test data is very low whereas the s
ame for training dataset is very much closer to 1 (e.g. 0.9876) that im
plies the model is overfitting to training data and throwing large erro
rs on unseen dataset.
```