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?

The optimal values of alpha for:

- Ridge Regression is 20
- Lasso Regression is 100

The R^2 value of the both Ridge and Lasso will get decreased when the alpha value is doubled, meaning the better fit for the model decreases

	Alpha	Alpha Doubled
Ridge Regression	OverallQual_9 Neighborhood_NoRidge OverallQual_10 Neighborhood_NridgHt Neighborhood_Crawfor	OverallQual_9 Neighborhood_NoRidge Neighborhood_NridgHt BsmtExposure_Gd OverallQual_10
Lasso Regression	OverallQual_10 OverallQual_9 Neighborhood_NoRidge OverallQual_8 Neighborhood_Crawfor	OverallQual_10 OverallQual_9 Neighborhood_NoRidge OverallQual_8 Neighborhood_Crawfor

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?

We will choose the Lasso Regression because apart from being robust it will also help in feature selection by setting the coefficient to zero if they are not relevant.

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 first top 5 predictor variables

- OverallQual_9
- Neighborhood_NoRidge
- Neighborhood_NridgHt
- OverallQual_10
- BsmtExposure_Gd

The top 5 predictor variables after excluding the above 5 are:

- OverallQual_6
- Neighborhood_NWAmes
- BsmtQual_TA
- Neighborhood_CollgCr
- Fireplaces

The top variables were taken from the Elastic Net Regression

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 is robust and generalisable when they the training data is not affected by the outliers i.e. proper outlier analysis has to be performed and only those are relevant to the dataset should be maintained and the irrelevant outliers should be discarded. This will gradually be increasing the accuracy in the predictions made by the model. The model is to be generalised so that there is no much of a difference in test and training dataset accuracy