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 1 - It was observed that most values of alpha either 5 or 1 is doing the best job in predicting the same results for test and train data.
Lazzo - higher the alpha results in most feature coefficients to be zero.
Ridge - Higher the alpha the more the regularization.
Important predictor -
GrLivArea, OverallQual, OverallCond, TotalBsmtSF, BsmtFinSF1, GarageArea, Fireplaces, LotArea, LotArea, LotFrontage

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 2 - : We must regularize coefficients and improve the prediction accuracy.
Decrease in variance, and making the model interpretably.

Ridge regression, uses a penalty parameter called lambda. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have higher values gets penalized.
Variance of model is dropped with increase with increase in lambda value and bias remains the same.
Ridge regression do not exclude any variables in final model.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation.
Lasso also does variable selection.
With increased value of lambda - Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0.

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 3 - Five most important predictor variables to be excluded are -
GarageArea, OverallQual, OverallCond, TotalBsmtSf, GrLivArea

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 4 - Model should be designed in such a way that it should not be too simple nor too complex. Model should be designed with minimum error considering the optimal values of bias and variance.
Simple model comes with more bias and less variance and vice-versa for complex model.
A robust and generalisable model will have best results for both training data and test data with a minimal difference which can be ignored.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

Bias: when the model is weak to learn from the data. High bias means model isunable to learn details in the data. Model performs poor on training and testing data.
Variance: Variance is error in model, when model tries to over learn from the data. High variance means
model performs exceptionally well on training data as it has very well trained on this of data but
performs very poor on testing data as it was unseen data for the model.