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?

ANSWER: The optimal value of alpha for ridge regression is 4 and for lasso regression it is 0.0001 .Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model. This modification is done by adding a penalty parameter that is equivalent to the square of the magnitude of the coefficients.

Loss function = OLS + alpha * summation (squared coefficient values)

In the above loss function, alpha is the parameter we need to select. A low alpha value can lead to over-fitting, whereas a high alpha value can lead to under-fitting.

Lasso regression, or the Least Absolute Shrinkage and Selection Operator, is also a modification of linear regression. In Lasso, the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients .

The loss function for Lasso Regression can be expressed as below:

Loss function = OLS + alpha * summation (absolute values of the magnitude of the coefficients)

Therefore, doubling the value of alpha for both ridge and lasso regression would lead to under-fitting.

SalePrice is the most important variable.

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?

ANSWER: We would apply the one with lower lambda as lower the λ on the features, the model will resemble linear regression model and training and test score increases. For α = 0.0001, coefficients for Lasso regression and linear regression show close resemblance.

Question 3

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?

ANSWER: To overcome this issue, we can either change the model or metric, or we can make some changes in the data and use the same models. We have used IQR to identify outliers in our data and treat them . It tells how spread the middle values are. It can be used to tell when a value is too far from the middle. An outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile.