Problem Statement Part 2

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 both ridge and lasso regression
Lasso Alpha 10
Ridge Alpha 1

Code -

```
Lambdas <- 10^seq(2, -3, by = -.1)

# Setting alpha = 1 implements lasso regression
Lasso_reg <- cv.glmnet(x, y_train, alpha = 1, lambda = lambdas,
Standardize = TRUE, nfolds = 5)

Lambda_best <- lasso_reg$lambda.min
Lambda_best
```

Output: -

```
0.001
```

If we double the value of alpha for both lasso and ridge regression then you should see that the optimal value of alpha is 20, with a negative MSE of -3.07267. This is a slight improvement upon the basic multiple linear regression.

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 – Lasso regression would be better option it would help in feature elimination and the model will be more robust.

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?

Answer-

Lotarea,overallqual,yearbuilt,bsmtfinsf1,totalbsmtsf are the top 5 important predictors.

Let's drop these columns

```
X_train2 = X_train1.drop(['lotarea','overallqual','yearbuilt','bsmtfinsf1','totalbsmtsf'],axis=1)
X_test2 = X_test1.drop(['lotarea','overallqual','yearbuilt','bsmtfinsf1','totalbsmtsf'],axis=1)
X_train2.head()
```

After proper analyzation we found
11stflrsf-----------First Floor square feet
Grlivarea-----------Above grade (ground) living area square feet
Street_Pave---------Pave road access to property
Roofmatl_Metal------Roof material_Metal
Roofstyle_Shed------Type of roof(Shed)

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?

Answer - The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.