Assignment 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?

For Ridge Regression, Optimum Alpha = 2

For Lasso Regression, Optimum Alpha = 0. 0001

When alpha values are doubled; Ridge Regression coefficients are:

	Ridge newd Alpha Co-Efficient
Total_Usable_Living_Area	0.119740
TotRmsAbvGrd	0.068665
GarageArea	0.056238
OverallCond	0.049476
LotArea	0.039918
Neighborhood_Crawfor	0.036387
Total_Porch_Area	0.036020
Neighborhood_StoneBr	0.032988
Total_Bathroom_Count	0.027077
LotFrontage	0.024392
HouseStyle_2.5Unf	0.022526
Neighborhood_NoRidge	0.022149
Exterior1st_BrkFace	0.021089
Functional_Typ	0.020985
Alley_Pave	0.020823
CentralAir_Y	0.019180
Heating_GasW	0.019023
ExterCond_Ex	0.017339
KitchenQual_Ex	0.016540
RoofStyle_Mansard	0.015877
HouseStyle_1.5Fin	0.015870
RoofMatl_WdShngl	0.015666
Neighborhood_NridgHt	0.015375
Condition1_Norm	0.015003
Condition2_PosA	0.014848

When alpha values are doubled; Lasso Regression coefficients are:

	Lasso newd Alpha Co-Efficient
Total_Usable_Living_Area	0.215147
TotRmsAbvGrd	0.071788
GarageArea	0.071011
OverallCond	0.058840
Total_Porch_Area	0.038290
Neighborhood_Crawfor	0.035055
LotArea	0.027104
Functional_Typ	0.024699
Neighborhood_StoneBr	0.022809
CentralAir_Y	0.018640
Exterior1st_BrkFace	0.017408
Neighborhood_Somerst	0.014113
Alley_Pave	0.013762
Condition1_Norm	0.013277
BsmtFinType1_Unf	0.013128
BsmtFinType2_Unf	0.012286
BsmtExposure_Gd	0.012063
KitchenQual_Ex	0.012036
BsmtQual_Ex	0.010782
Neighborhood_NoRidge	0.010446

The most important predictor variable will be 'Total_Usable_Living_Area'

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 lambda value in case of Ridge and Lasso is as follows:-

- Ridge – 2
- Lasso – 0.0001

The MSE values are

Ridge - 0.0013971780207398894

Lasso - 0.0014675904134552114

Lasso will be preferred over Ridge as the final model due to feature reduction

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?

Top 5 predictor variables after the changes

	Lasso Co-Efficient
LotArea	0.117865
Total_Bathroom_Count	0.085034
HouseStyle_2.5Unf	0.057746
LotFrontage	0.057387
Neighborhood_Crawfor	0.045244

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?

Given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons:-

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex

ones and hence are easier to train.

Simpler models are more robust.

Complex models tend to change wildly with changes in the training data set

Simple models have low variance, high bias and complex models have low bias, high variance

Simpler models make more errors in the training set. Complex models lead to overfitting

they work very well for the training samples, fail miserably when applied to other test

samples

Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

Also, Making a model simple leads to Bias-Variance Trade-off:

• A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.

• A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.