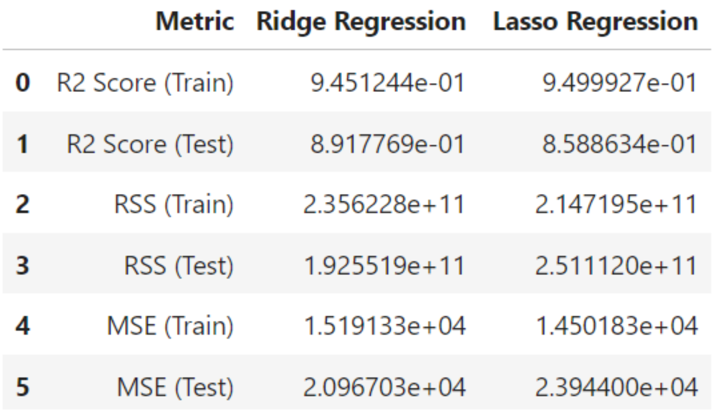
Question 1

1.1 What is the optimal value of alpha for ridge and lasso regression?

The optimal value of alpha for ridge is 6.0 and for lasso is 50

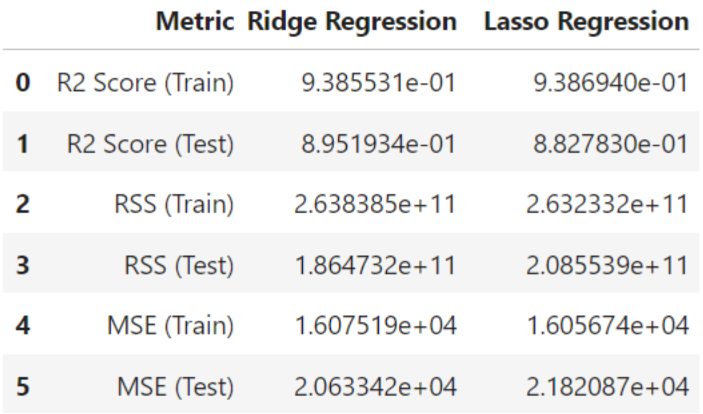
1.2 What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

Below table is for the optimal value of alpha for ridge 6.0 and for lasso 50



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.451244e-01	9.499927e-01
1	R2 Score (Test)	8.917769e-01	8.588634e-01
2	RSS (Train)	2.356228e+11	2.147195e+11
3	RSS (Test)	1.925519e+11	2.511120e+11
4	MSE (Train)	1.519133e+04	1.450183e+04
5	MSE (Test)	2.096703e+04	2.394400e+04

Below table is after doubling the values of optimal alpha

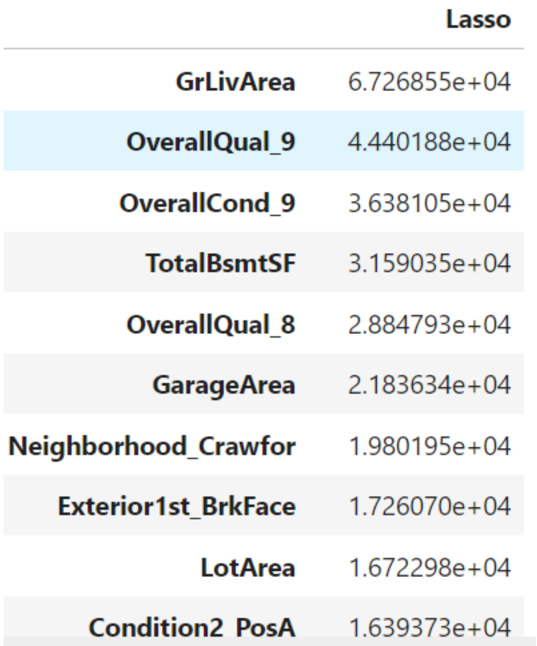


	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.385531e-01	9.386940e-01
1	R2 Score (Test)	8.951934e-01	8.827830e-01
2	RSS (Train)	2.638385e+11	2.632332e+11
3	RSS (Test)	1.864732e+11	2.085539e+11
4	MSE (Train)	1.607519e+04	1.605674e+04
5	MSE (Test)	2.063342e+04	2.182087e+04

So, after doubling the values of alpha the R2 for test for Lasso has improved a little bit. Rest of the value of R2 are mostly same.

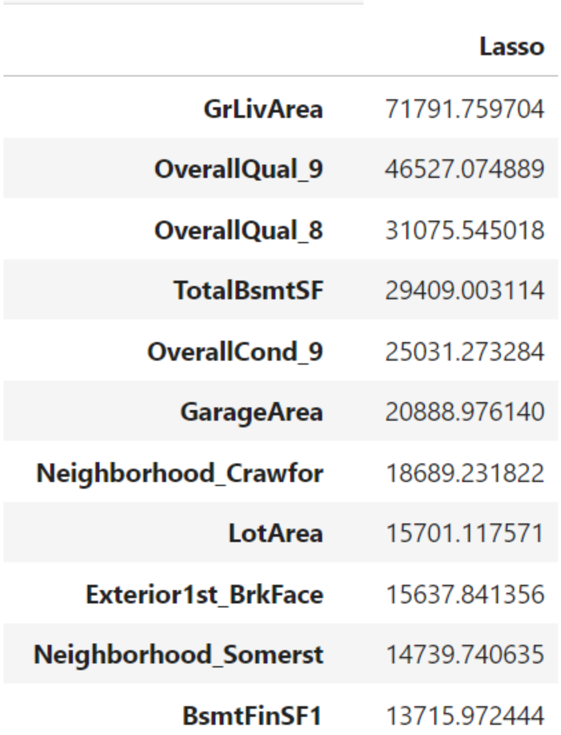
What will be the most important predictor variables after the change is implemented?

Below are the predictor variables for Lasso for optimal alpha value



	Lasso
GrLivArea	6.726855e+04
OverallQual_9	4.440188e+04
OverallCond_9	3.638105e+04
TotalBsmtSF	3.159035e+04
OverallQual_8	2.884793e+04
GarageArea	2.183634e+04
Neighborhood_Crawfor	1.980195e+04
Exterior1st_BrkFace	1.726070e+04
LotArea	1.672298e+04
Condition2 PosA	1.639373e+04

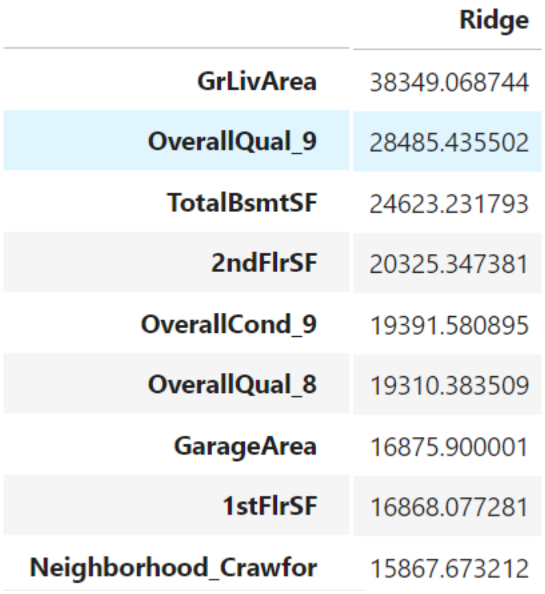
Below are the predictor variables for Lasso after doubling alpha value



	Lasso
GrLivArea	71791.759704
OverallQual_9	46527.074889
OverallQual_8	31075.545018
TotalBsmtSF	29409.003114
OverallCond_9	25031.273284
GarageArea	20888.976140
Neighborhood_Crawfor	18689.231822
LotArea	15701.117571
Exterior1st_BrkFace	15637.841356
Neighborhood_Somerst	14739.740635
BsmtFinSF1	13715.972444

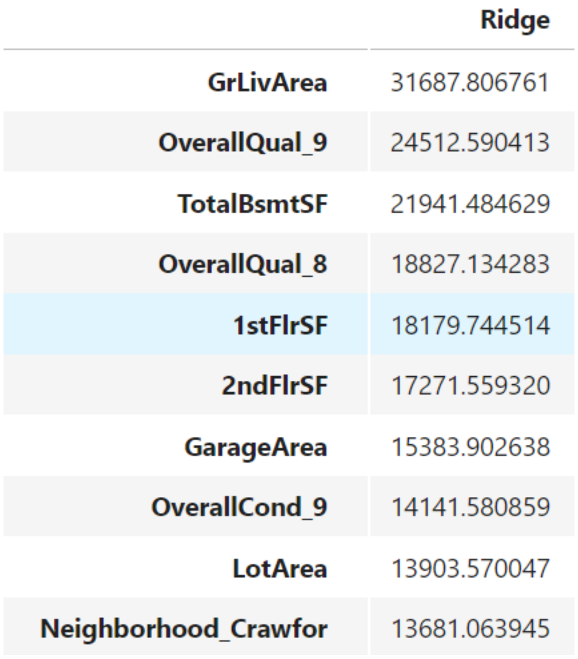
Looking at top 5 predictors, in case of Lasso ,after doubling values of Alpha the importance of OverallCond_9 is reduced while of OverallQual_8 has increased.

Below are the predictor variables for Ridge for optimal alpha value



	Ridge
GrLivArea	38349.068744
OverallQual_9	28485.435502
TotalBsmtSF	24623.231793
2ndFlrSF	20325.347381
OverallCond_9	19391.580895
OverallQual_8	19310.383509
GarageArea	16875.900001
1stFlrSF	16868.077281
Neighborhood_Crawfor	15867.673212

Below are the predictor variables for Ridge after doubling alpha value



	Ridge
GrLivArea	31687.806761
OverallQual_9	24512.590413
TotalBsmtSF	21941.484629
OverallQual_8	18827.134283
1stFlrSF	18179.744514
2ndFlrSF	17271.559320
GarageArea	15383.902638
OverallCond_9	14141.580859
LotArea	13903.570047
Neighborhood_Crawfor	13681.063945

Looking at top 5 predictors, in case of Ridge ,after doubling values of Alpha the importance of OverallQual_8 and 1stFlrSF has increased while of OverallCond_9 has decreased.

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 model we will choose to apply will depend on use case. If we have too many variables and one of our primary goal is feature selection, then we will use Lasso. If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use Ridge Regression.

In this case since there are many features, so to make the model simple we can choose Lasso.

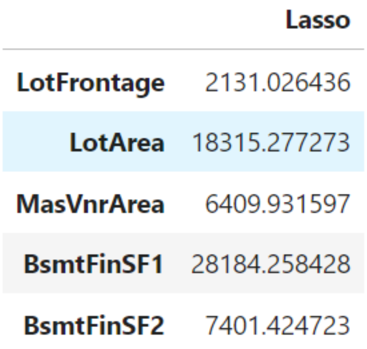
Question 3

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 top 5 variables for Lasso are

'GrLivArea','OverallQual_9','OverallCond_9','TotalBsmtSF','OverallQual_8'

After dropping them below are new top 5 predictors



	Lasso
LotFrontage	2131.026436
LotArea	18315.277273
MasVnrArea	6409.931597
BsmtFinSF1	28184.258428
BsmtFinSF2	7401.424723

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 when any variation in the data does not affect its performance much.

A model is generalizable if it's able to for new and unseen data, we can get the same distribution as the one used to create the model.

To make sure a model is robust and generalizable, we have to take care that it doesn't overfit. This is because an overfitting model has very high variance and smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.

In other words, the model should not be too complex to be robust and generalizable.

If we look at it from the prespective of Accuracy, a complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. The addition of bias means that accuracy will decrease.

In general, we must find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.