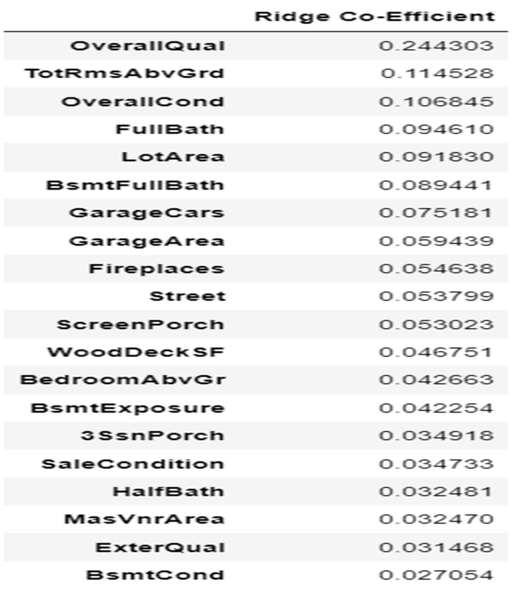
Question 1

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?

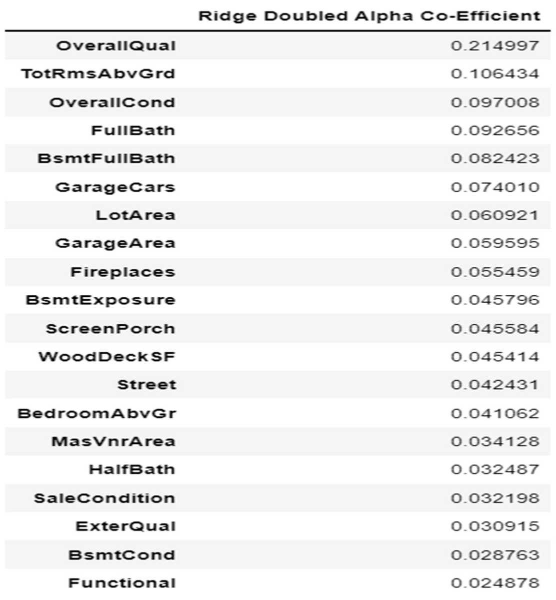
Answer: The optimal value of alpha of ridge and lasso regression is 2 and 0.0001. We use these alpha values in both regressions. 2 in ridge regression and 0.0001 in lasso regression. Then we find the R^2 value i.e.., 0.85 approximately.

We observed the two observations i.e., Lasso Model by doubling the value of alpha to 0.0002 and Ridge Model by doubling the value of alpha to 4. R^2 score is slightly decreasing but do not see huge change. Finally we observed that best alpha value of Ridge and lasso regression is 2 and 0.0001. But there is a small change in both regression coefficient values. (Using snipping tool. Jupyter notebook values are crop the values and insert below of the both regression coefficients).

Ridge Regression:

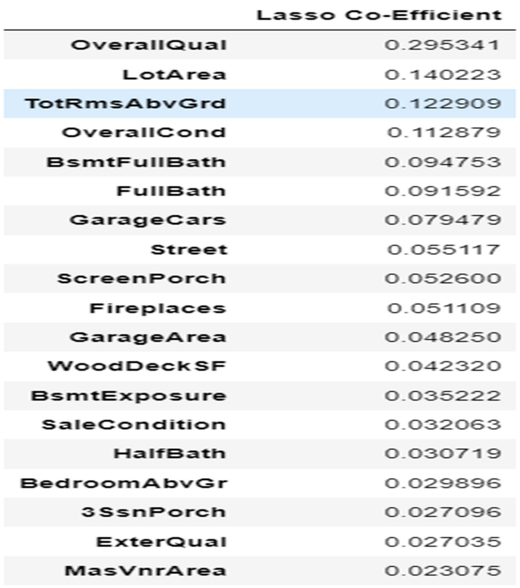


	Ridge Co-Efficient
OverallQual	0.244303
TotRmsAbvGrd	0.114528
OverallCond	0.106845
FullBath	0.094610
LotArea	0.091830
BsmtFullBath	0.089441
GarageCars	0.075181
GarageArea	0.059439
Fireplaces	0.054638
Street	0.053799
ScreenPorch	0.053023
WoodDeckSF	0.046751
BedroomAbvGr	0.042663
BsmtExposure	0.042254
3SsnPorch	0.034918
SaleCondition	0.034733
HalfBath	0.032481
MasVnrArea	0.032470
ExterQual	0.031468
BsmtCond	0.027054

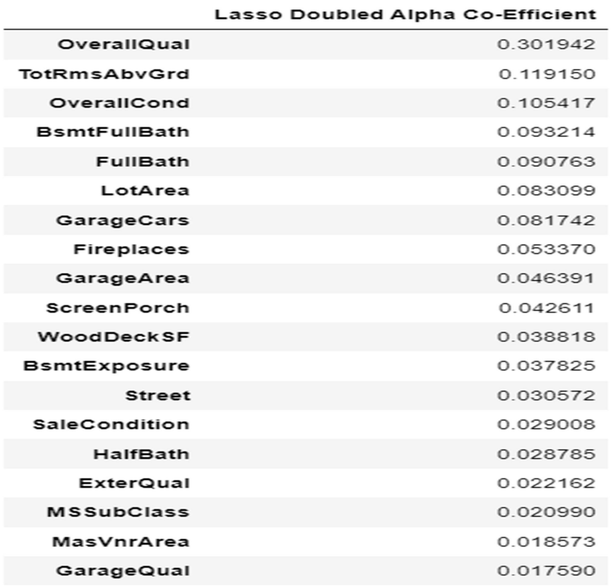


	Ridge Doubled Alpha Co-Efficient
OverallQual	0.214997
TotRmsAbvGrd	0.106434
OverallCond	0.097008
FullBath	0.092656
BsmtFullBath	0.082423
GarageCars	0.074010
LotArea	0.060921
GarageArea	0.059595
Fireplaces	0.055459
BsmtExposure	0.045796
ScreenPorch	0.045584
WoodDeckSF	0.045414
Street	0.042431
BedroomAbvGr	0.041062
MasVnrArea	0.034128
HalfBath	0.032487
SaleCondition	0.032198
ExterQual	0.030915
BsmtCond	0.028763
Functional	0.024878

Lasso Regression:



	Lasso Co-Efficient
OverallQual	0.295341
LotArea	0.140223
TotRmsAbvGrd	0.122909
OverallCond	0.112879
BsmtFullBath	0.094753
FullBath	0.091592
GarageCars	0.079479
Street	0.055117
ScreenPorch	0.052600
Fireplaces	0.051109
GarageArea	0.048250
WoodDeckSF	0.042320
BsmtExposure	0.035222
SaleCondition	0.032063
HalfBath	0.030719
BedroomAbvGr	0.029896
3SsnPorch	0.027096
ExterQual	0.027035
MasVnrArea	0.023075



	Lasso Doubled Alpha Co-Efficient
OverallQual	0.301942
TotRmsAbvGrd	0.119150
OverallCond	0.105417
BsmtFullBath	0.093214
FullBath	0.090763
LotArea	0.083099
GarageCars	0.081742
Fireplaces	0.053370
GarageArea	0.046391
ScreenPorch	0.042611
WoodDeckSF	0.038818
BsmtExposure	0.037825
Street	0.030572
SaleCondition	0.029008
HalfBath	0.028785
ExterQual	0.022162
MSSubClass	0.020990
MasVnrArea	0.018573
GarageQual	0.017590

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: The optimal value of lambda/alpha for ridge and lasso regression is:

Ridge regression:

The optimum alpha is 2
The R2 Score of the model on the test dataset for optimum alpha is 0.8527753412851128
The MSE of the model on the test dataset for optimum alpha is 0.002916540591270374

Lasso regression:

The optimum alpha is 0.0001
The R2 Score of the model on the test dataset for optimum alpha is 0.8564535203566637
The MSE of the model on the test dataset for optimum alpha is 0.002843675361642544

The MSE of the both models are not so much difference. Lasso should perform better in situations where only a few among all the predictors that are used to build our model have a significant influence on the response variable. So, feature selection, which removes the unrelated variables, should help. But Ridge should do better when all the variables have almost the same influence on the response variable. So Lasso helps in feature reduction, lasso has better over the ridge regression.

Question 3

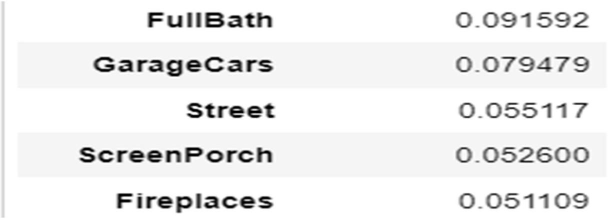
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer: The five most important predictor variables in the lasso model are:

1. Overall quality of the house
2. Total rooms above grade (does not include bathrooms)
3. Lot size in square feet
4. Overall Condition of the house
5. Basement full bathrooms in the house.

Excluding the five most important predictor variables in lasso regression analysis model. The R^2 value of without the above top five predictor variables is drops 0.71. And the mean squared error increase to 0.002943675361642.

After the new top 5 predictors are:



FullBath	0.091592
GarageCars	0.079479
Street	0.055117
ScreenPorch	0.052600
Fireplaces	0.051109

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:

The following unique point's model is robust and generalizable. Implications for the accuracy of the model explained below:

A simpler model is usually more generic than a complex model. This becomes important because generic models are bound to perform better on unseen data sets. A simpler model requires fewer training data points. This becomes extremely important because in many cases, one has to work with limited data points.

A simple model is more robust and does not change significantly if the training data points undergo small changes.

A simple model may make more errors in the training phase but is bound to outperform complex models when it views new data. This happens because of over fitting.

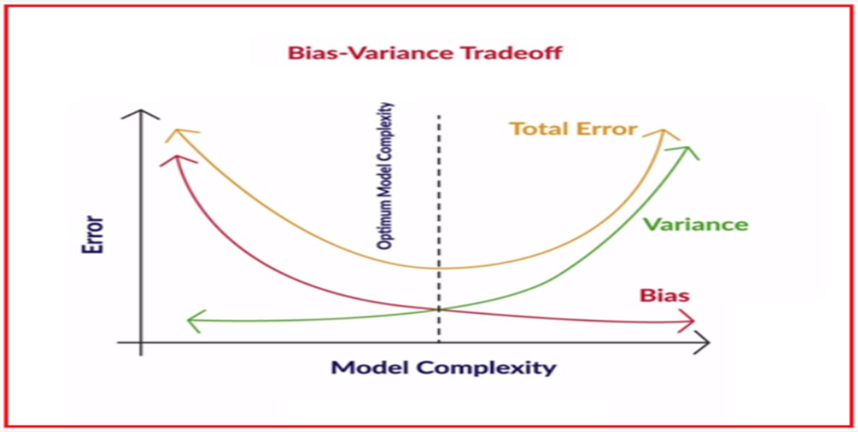
Simpler models are generic, i.e., they apply to a wider range of data. Complex models make assumptions about the data, which are likely to be wrong. Simpler models require less training data compared with complex models. Simpler models are more robust.

Bias and Variance: We considered the example of a model memorising the entire training data set. If you change the data set slightly, this model will also need to change drastically.

The model is, therefore, unstable and sensitive to changes in training data, and this is called high variance. The 'variance' of a model is the variance in its output on some test data with respect to the changes in the training data. In other words, variance here refers to the degree of changes in the model itself with respect to changes in the training data.

Bias quantifies how accurate the model is likely to be on future (test) data. Extremely simple models are likely to fail in predicting complex real-world phenomena. Simplicity has its own disadvantages. Imagine solving digital image processing problems using simple linear regression when much more complex models such as neural networks are typically successful for such problems.

We say that a linear model has a high bias because it is quite simple to be able to learn the complexity involved in the task. Ideally, we want to reduce both bias and variance because the expected total error of a model is the sum of the errors in bias and variance, as shown in the figure given below.



Bias Variance Tradeoff

Regularization helps with managing model complexity by essentially shrinking the model coefficient estimates towards 0. This discourages the model from becoming too complex, thus avoiding the risk of over fitting.