Question 1:

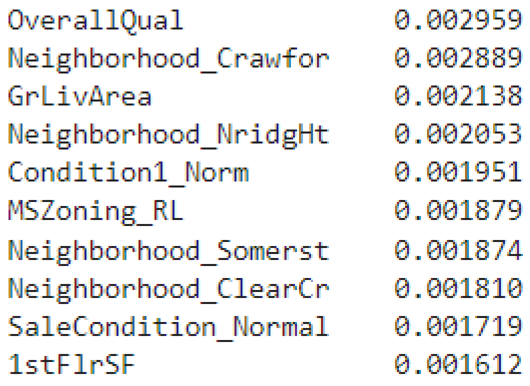
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans:

The Optimal Value of alpha for the Ridge and Lasso Regression are as follows:

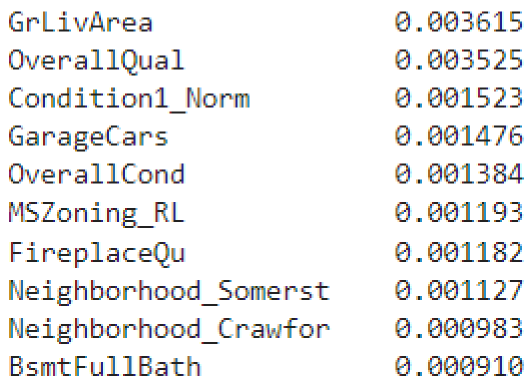
Ridge Regression: 20
Lasso Regression: 0.0001

Ridge

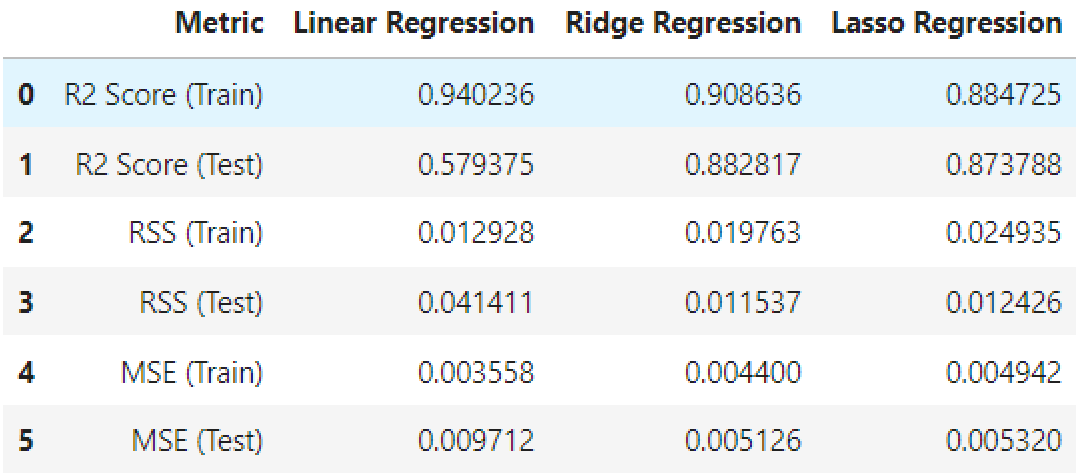


```
OverallQual             0.002959
Neighborhood_Crawfor    0.002889
GrLivArea               0.002138
Neighborhood_NridgHt    0.002053
Condition1_Norm         0.001951
MSZoning_RL             0.001879
Neighborhood_Somerst    0.001874
Neighborhood_ClearCr    0.001810
SaleCondition_Normal    0.001719
1stFlrSF                0.001612
```

Lasso



```
GrLivArea               0.003615
OverallQual             0.003525
Condition1_Norm         0.001523
GarageCars              0.001476
OverallCond             0.001384
MSZoning_RL             0.001193
FireplaceQu             0.001182
Neighborhood_Somerst    0.001127
Neighborhood_Crawfor    0.000983
BsmtFullBath            0.000910
```

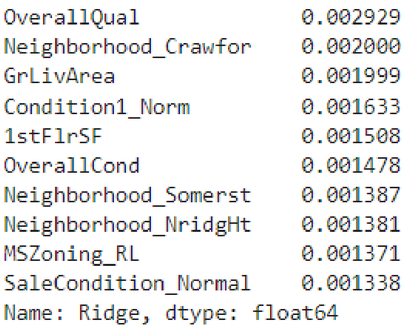


	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.940236	0.908636	0.884725
1	R2 Score (Test)	0.579375	0.882817	0.873788
2	RSS (Train)	0.012928	0.019763	0.024935
3	RSS (Test)	0.041411	0.011537	0.012426
4	MSE (Train)	0.003558	0.004400	0.004942
5	MSE (Test)	0.009712	0.005126	0.005320

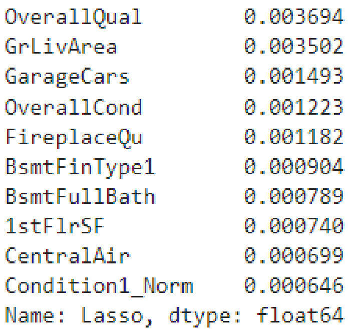
When we double the alpha values for both Ridge and Lasso, the R2 score almost remain the same, but there is a change in the important predictor variables compared to the optimum alpha value.

New Predictor variables and their coefficients based on the new alpha values are below.

Ridge   Lasso



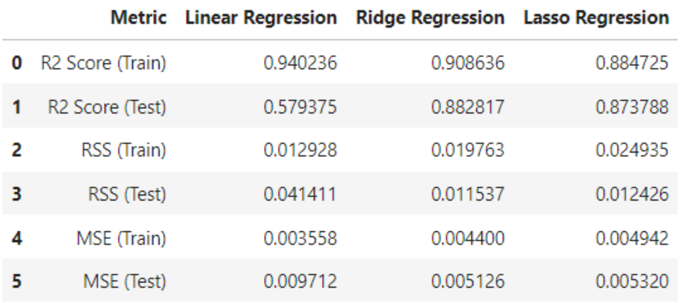
```
OverallQual               0.002929
Neighborhood_Crawfor      0.002000
GrLivArea                 0.001999
Condition1_Norm           0.001633
1stFlrSF                  0.001508
OverallCond               0.001478
Neighborhood_Somerst      0.001387
Neighborhood_NridgHt      0.001381
MSZoning_RL               0.001371
SaleCondition_Normal      0.001338
Name: Ridge, dtype: float64
```



```
OverallQual        0.003694
GrLivArea          0.003502
GarageCars         0.001493
OverallCond        0.001223
FireplaceQu        0.001182
BsmtFinType1       0.000904
BsmtFullBath       0.000789
1stFlrSF           0.000740
CentralAir         0.000699
Condition1_Norm    0.000646
Name: Lasso, dtype: float64
```

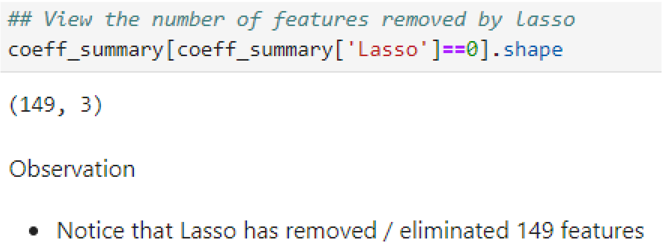
Question 2:
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Ans :



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.940236	0.908636	0.884725
1	R2 Score (Test)	0.579375	0.882817	0.873788
2	RSS (Train)	0.012928	0.019763	0.024935
3	RSS (Test)	0.041411	0.011537	0.012426
4	MSE (Train)	0.003558	0.004400	0.004942
5	MSE (Test)	0.009712	0.005126	0.005320

Prefer to select Lasso Regression.
Both Ridge and Lasso Regression are performing well in terms of predicting the Sale Price. However, Ridge regression does have one obvious disadvantage. It would include all the predictors in the final model. This may not affect the accuracy of the predictions but can make model interpretation challenging when the number of predictors is very large. The number of feature variables is very large (201) and the data may have unrelated or noisy variables, we may not want to keep such variables in the model. Lasso regression helps us here by performing feature selection.



```
## View the number of features removed by lasso
coeff_summary[coeff_summary['Lasso']==0].shape
```

```
(149, 3)
```

Observation

- Notice that Lasso has removed / eliminated 149 features

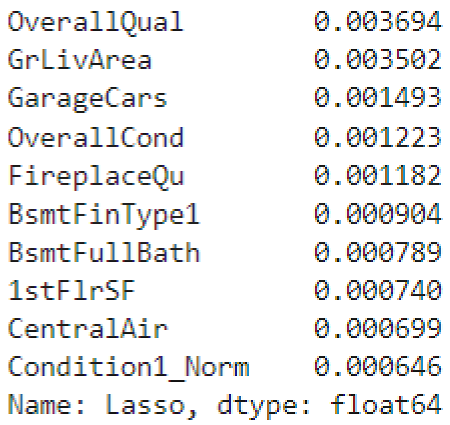
Lasso regression able to nullify close to 149 features, which is very a important feature and make the model more simple compared to Ridge.

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?

Ans:

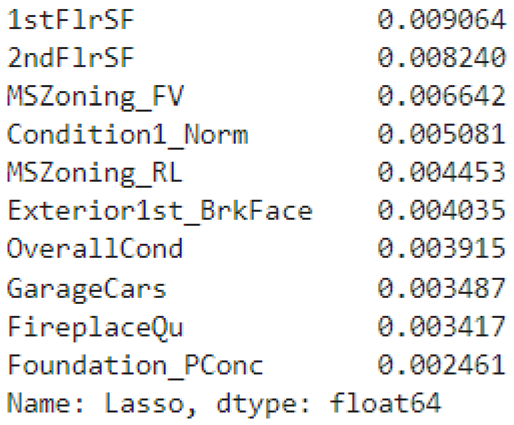
The optimal value of Lasso was 0.0001 , then its double to 0.0002 for the question 2 to measure the impact. Below are the top 10 predictors before dropping it.



```
OverallQual        0.003694
GrLivArea          0.003502
GarageCars         0.001493
OverallCond        0.001223
FireplaceQu        0.001182
BsmtFinType1       0.000904
BsmtFullBath       0.000789
1stFlrSF           0.000740
CentralAir         0.000699
Condition1_Norm    0.000646
Name: Lasso, dtype: float64
```

Dropped “'OverallQual','GrLivArea', 'GarageCars','OverallCond','FireplaceQu'” these variables from the model and rebuild it.

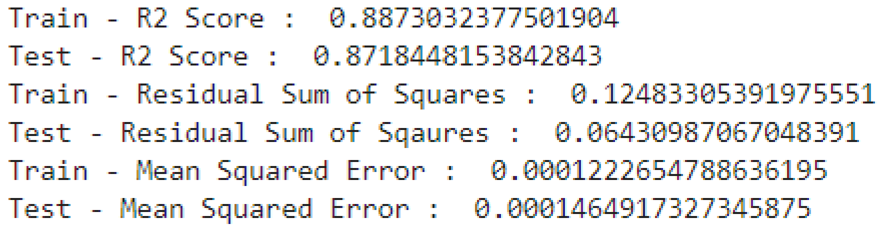
Below are new top 10 predictor variables after removing the 5 .



```
1stFlrSF                0.009064
2ndFlrSF                0.008240
MSZoning_FV             0.006642
Condition1_Norm         0.005081
MSZoning_RL             0.004453
Exterior1st_BrkFace     0.004035
OverallCond             0.003915
GarageCars              0.003487
FireplaceQu             0.003417
Foundation_PConc        0.002461
Name: Lasso, dtype: float64
```

Top 5 predictor variables - 1stFlrSF , 2ndFlrSF, MSZoning_FV , Condition1_Norm , MSZoning_RL

New R2 score after removing top 5 predictors.



```
Train - R2 Score :  0.8873032377501904
Test - R2 Score :  0.8718448153842843
Train - Residual Sum of Squares :  0.12483305391975551
Test - Residual Sum of Sqaures :  0.06430987067048391
Train - Mean Squared Error :  0.0001222654788636195
Test - Mean Squared Error :  0.0001464917327345875
```

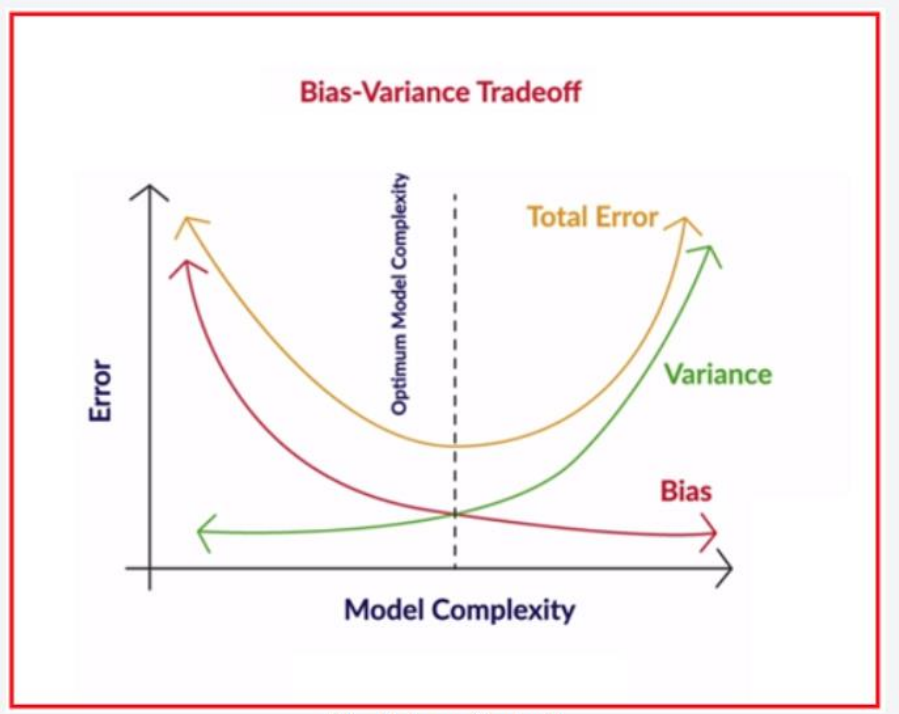
Question 4

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?

Ans:

For a model to be more generic and robust it needs to be simple and not complex. By simple it means minimum number of driving factors that can provide an impact on the decision.

1. The model may not trace all the data points in the training data set however it has a higher chance of detecting the test data.
2. There must be a good trade-off between bias and trade off
3. The model should not be too simple else it will fall under high bias and high variance. Not able to perform both on training and test data set
4. Regularization can help us to achieve a simpler model The robust model has a good accuracy value which satisfies a trade off between bias and variance. The complex model tends to have a good accuracy score however it tends to overfit the model and fail to predict the test data.



Bias Variance Tradeoff

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.

Bias measures how accurately a model can describe the actual task at hand.
Variance measures how flexible the model is with respect to changes in the training data.
As complexity increases, bias reduces and variance increases, and we aim to find the optimal point where the total model error is the least.