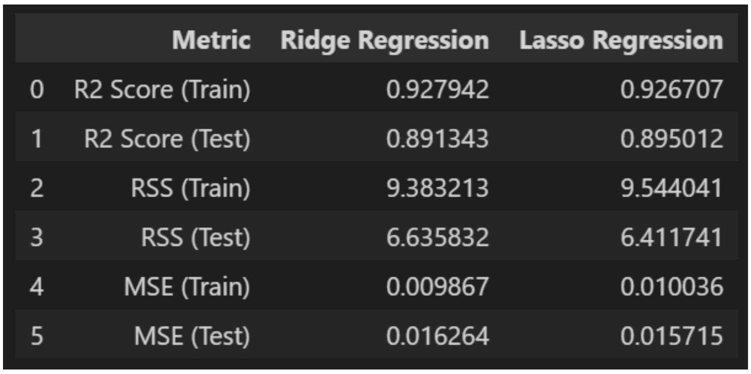
Question 1

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?

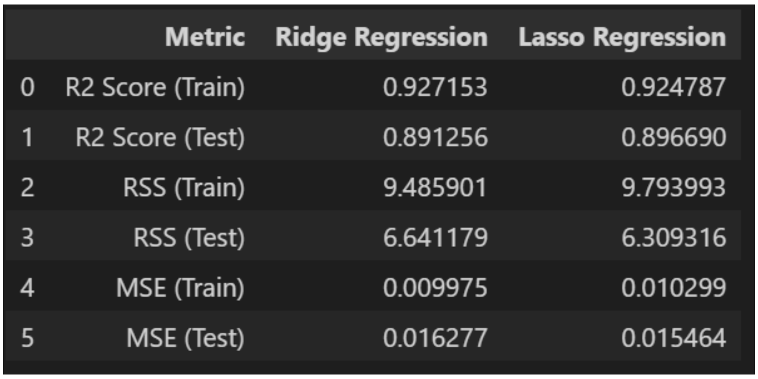
Ans: The optimal value of alpha for ridge and lasso regression are as follows
1. Ridge Regression – 20
2. Lasso Regression – 0.001

Result before doubling the value of alpha



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.927942	0.926707
1	R2 Score (Test)	0.891343	0.895012
2	RSS (Train)	9.383213	9.544041
3	RSS (Test)	6.635832	6.411741
4	MSE (Train)	0.009867	0.010036
5	MSE (Test)	0.016264	0.015715

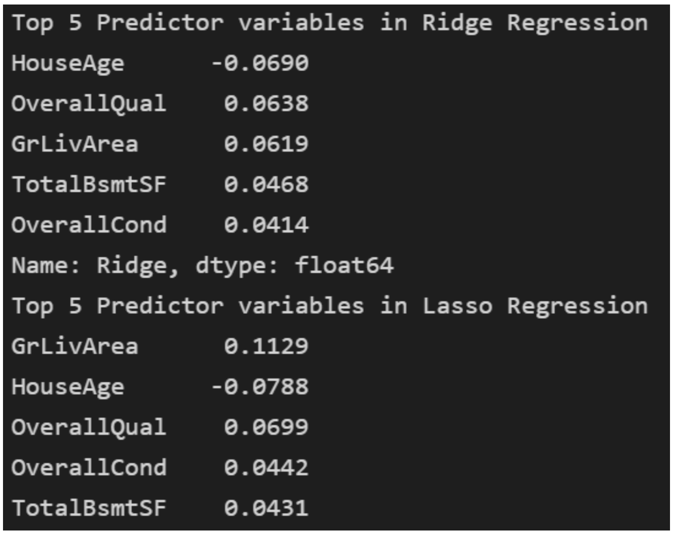
Result after doubling the value of alpha



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.927153	0.924787
1	R2 Score (Test)	0.891256	0.896690
2	RSS (Train)	9.485901	9.793993
3	RSS (Test)	6.641179	6.309316
4	MSE (Train)	0.009975	0.010299
5	MSE (Test)	0.016277	0.015464

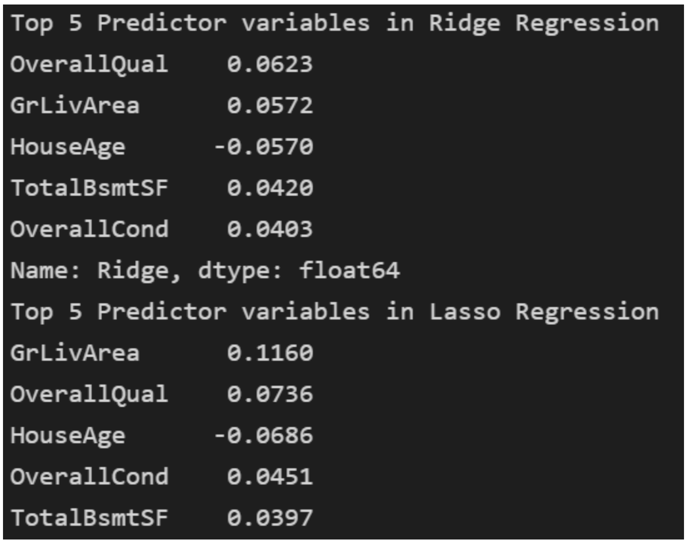
As we see in the above two figures, there is **no major significant change in model accuracy** on doubling the value of alpha for both Ridge and Lasso regression. However the values are better with optimal values of alpha and not the doubled value of alpha.

The most important variable before doubling the value of alpha



```
Top 5 Predictor variables in Ridge Regression
HouseAge      -0.0690
OverallQual    0.0638
GrLivArea      0.0619
TotalBsmtSF    0.0468
OverallCond    0.0414
Name: Ridge, dtype: float64
Top 5 Predictor variables in Lasso Regression
GrLivArea      0.1129
HouseAge      -0.0788
OverallQual    0.0699
OverallCond    0.0442
TotalBsmtSF    0.0431
```

The most important variable after doubling the value of alpha



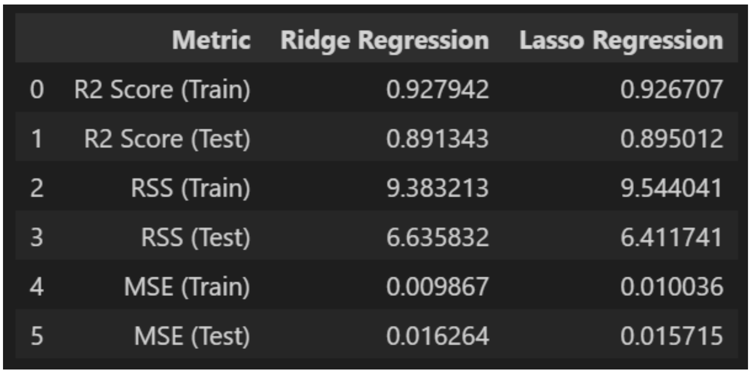
```
Top 5 Predictor variables in Ridge Regression
OverallQual    0.0623
GrLivArea      0.0572
HouseAge      -0.0570
TotalBsmtSF    0.0420
OverallCond    0.0403
Name: Ridge, dtype: float64
Top 5 Predictor variables in Lasso Regression
GrLivArea      0.1160
OverallQual    0.0736
HouseAge      -0.0686
OverallCond    0.0451
TotalBsmtSF    0.0397
```

As we see from the above two figures, the most important predictor in **Ridge regression** after doubling the value of alpha has changed to **OverallQual from HouseAge** whereas for **Lasso Regression** it remains same i.e **GrLivArea**

Question 2

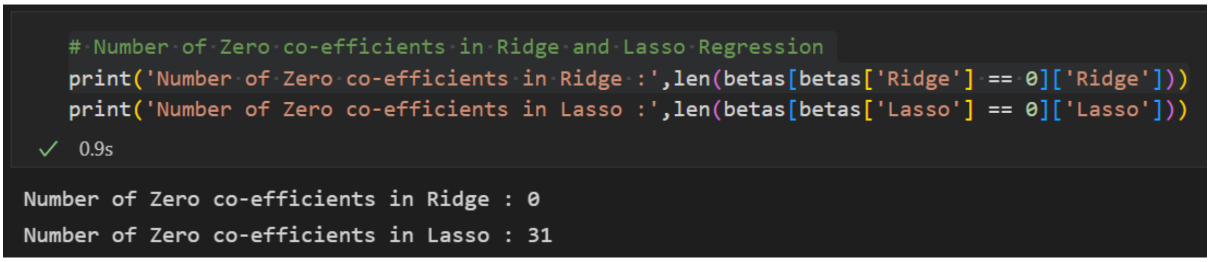
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Ans: The metrices for ridge and lasso regression is shown below



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.927942	0.926707
1	R2 Score (Test)	0.891343	0.895012
2	RSS (Train)	9.383213	9.544041
3	RSS (Test)	6.635832	6.411741
4	MSE (Train)	0.009867	0.010036
5	MSE (Test)	0.016264	0.015715

As we can see from the figure, r2 score and MSE is slightly better in Ridge when compared to Lasso. However Ridge regression, considers all features whereas in Lasso Regression it reduces co-efficient of 31 variables to zero thus making the model simpler.



```
# Number of Zero co-efficients in Ridge and Lasso Regression
print('Number of Zero co-efficients in Ridge :',len(betas[betas['Ridge'] == 0]['Ridge']))
print('Number of Zero co-efficients in Lasso :',len(betas[betas['Lasso'] == 0]['Lasso']))
```

✓ 0.9s

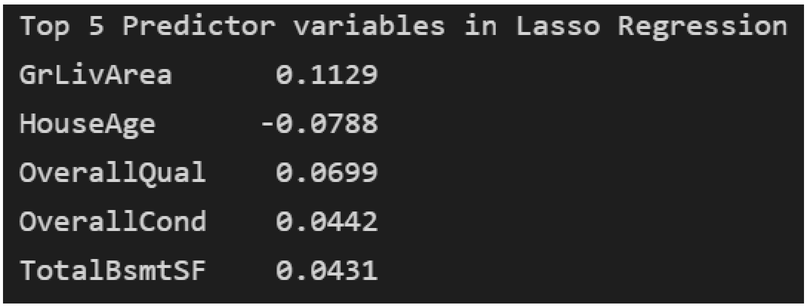
```
Number of Zero co-efficients in Ridge : 0
Number of Zero co-efficients in Lasso : 31
```

Hence, I will choose **lasso regression**.

Question 3

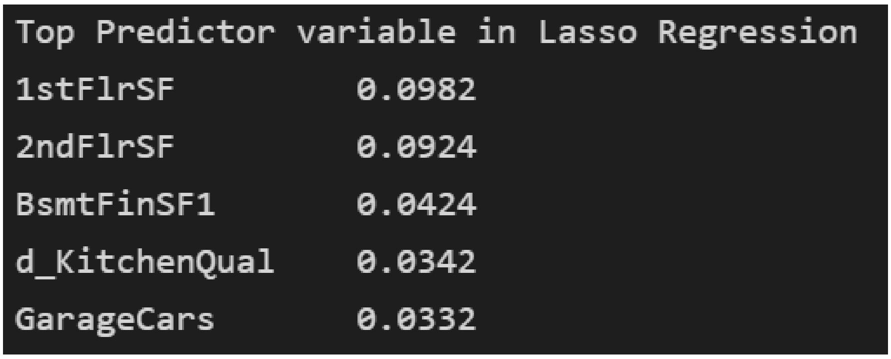
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Ans: The top five predictor variables in lasso model are



```
Top 5 Predictor variables in Lasso Regression
GrLivArea      0.1129
HouseAge      -0.0788
OverallQual    0.0699
OverallCond    0.0442
TotalBsmtSF    0.0431
```

After deleting them, the **new top 5 predictor variables** are shown below:



```
Top Predictor variable in Lasso Regression
1stFlrSF         0.0982
2ndFlrSF         0.0924
BsmtFinSF1       0.0424
d_KitchenQual    0.0342
GarageCars       0.0332
```

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: A model is said to be robust and generalisable only if

- It performs well on unseen data. Performance on training data will not give concreate idea about how model will perform when exposed to unseen data.
- Overfitting of the model on the train data needs to be avoided i.e. model does well on train data but performs poorly on test data.
- The model should be simpler enough to understand the pattern. It should try to keep balance between bias and variance.

The metrices like r2-score, RSS and MSE gives us an idea about the accuracy of the model. However for model to be robust and generalisable, the difference between these metrices for train and test data should be very less.