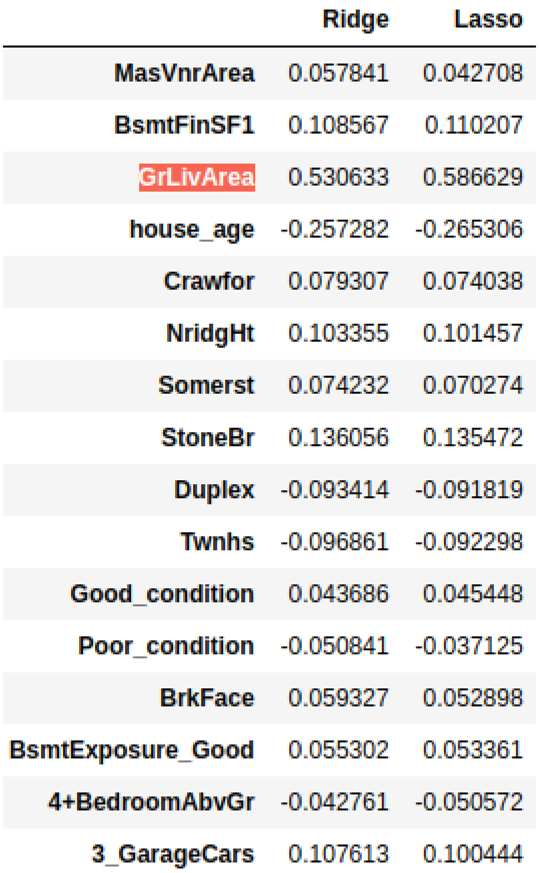
1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans:

The optimal value are -
1. Ridge = 1.0
2. Laso = 0.0001

If I double the parameters
Ridge = 2.0
Laso = 0.0002

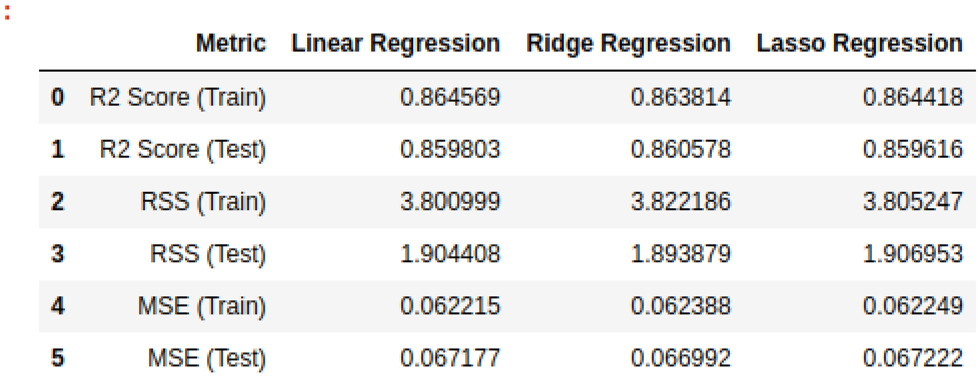
GrLivArea is the most important predictor.



	Ridge	Lasso
MasVnrArea	0.057841	0.042708
BsmtFinSF1	0.108567	0.110207
GrLivArea	0.530633	0.586629
house_age	-0.257282	-0.265306
Crawfor	0.079307	0.074038
NridgHt	0.103355	0.101457
Somerst	0.074232	0.070274
StoneBr	0.136056	0.135472
Duplex	-0.093414	-0.091819
Twnhs	-0.096861	-0.092298
Good_condition	0.043686	0.045448
Poor_condition	-0.050841	-0.037125
BrkFace	0.059327	0.052898
BsmtExposure_Good	0.055302	0.053361
4+BedroomAbvGr	-0.042761	-0.050572
3_GarageCars	0.107613	0.100444

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?

I have choosed Ridge because It has less variance and lesss bias on data.



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.864569	0.863814	0.864418
1	R2 Score (Test)	0.859803	0.860578	0.859616
2	RSS (Train)	3.800999	3.822186	3.805247
3	RSS (Test)	1.904408	1.893879	1.906953
4	MSE (Train)	0.062215	0.062388	0.062249
5	MSE (Test)	0.067177	0.066992	0.067222

MSE is lower for Ridge. (Less Bias)
Difference between R2 Score of Test and Train is lower in case of Ridge. (Low Variance)

Laso is overfitted comparative to Ridge.

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?

For model to be generalisable and robust.Model should not be overfitted. It should have low vaiance and low bias.

If we make model robust and generalisable, then accuracy of the model is decreased. As, accuracy means how well variables can define the complete dataset (More accuracy causes overfitting, model learns the complete data.)

To make the model robust and generalisable, we need a model which learns the pattern of data (not complete data itself). Hence we need perfrom traid off between accuracy and model generalisation.

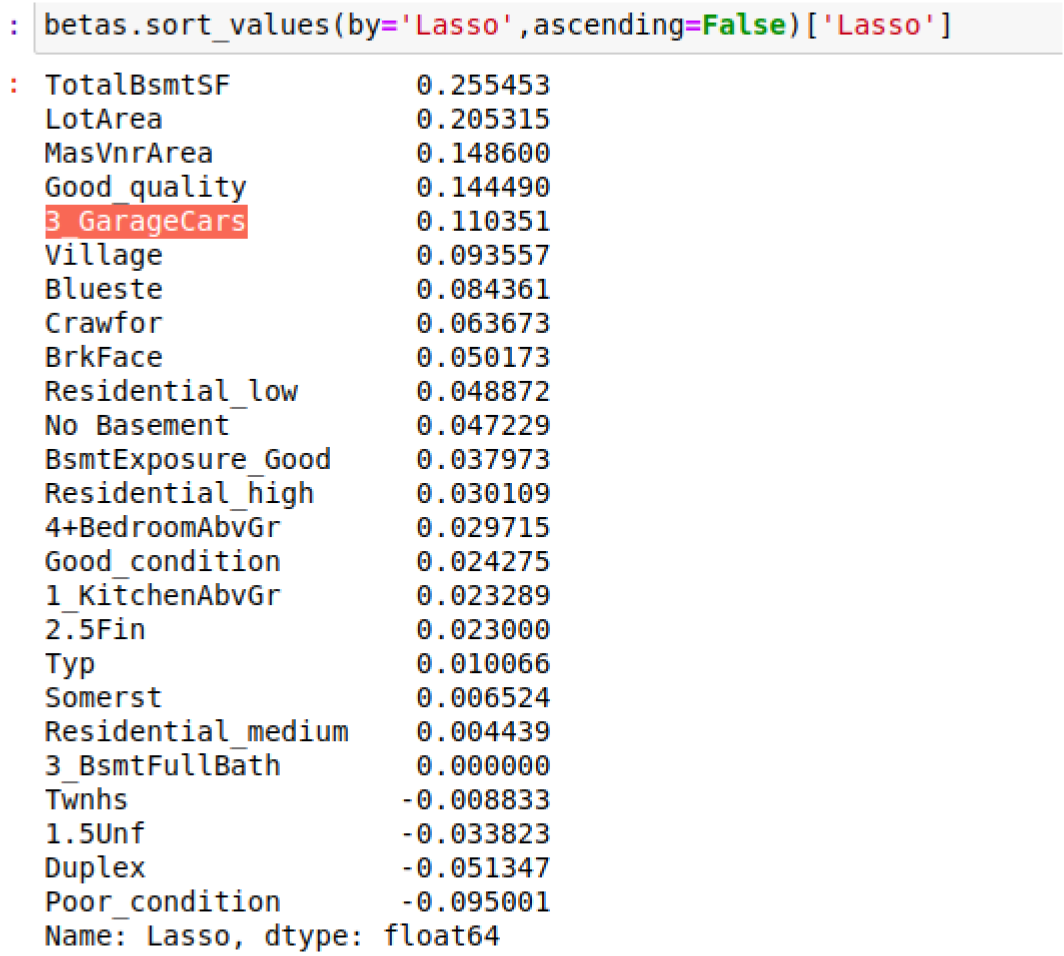
Question 3

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?

Initial top 5 = 'GrLivArea', 'house_age', 'BsmtFinSF1', 'StoneBr','NridgHt'

After Dropping these columns. Now the top 5 are:

TotalBsmtSF
LotArea
MasVnrArea
Good_quality
3_GarageCars



```
betas.sort_values(by='Lasso',ascending=False)['Lasso']
```

```
TotalBsmtSF           0.255453
LotArea               0.205315
MasVnrArea            0.148600
Good_quality          0.144490
3_GarageCars          0.110351
Village               0.093557
Blueste               0.084361
Crawfor               0.063673
BrkFace               0.050173
Residential_low       0.048872
No Basement           0.047229
BsmtExposure_Good     0.037973
Residential_high      0.030109
4+BedroomAbvGr        0.029715
Good_condition        0.024275
1_KitchenAbvGr        0.023289
2.5Fin                0.023000
Typ                   0.010066
Somerst               0.006524
Residential_medium    0.004439
3_BsmtFullBath        0.000000
Twnhs                -0.008833
1.5Unf               -0.033823
Duplex               -0.051347
Poor_condition       -0.095001
Name: Lasso, dtype: float64
```

Note: Please refer Code in jupyter notebook for code.