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?

The optimal values for Ridge and Lasso regression using top 120 variables selected by RFE are :

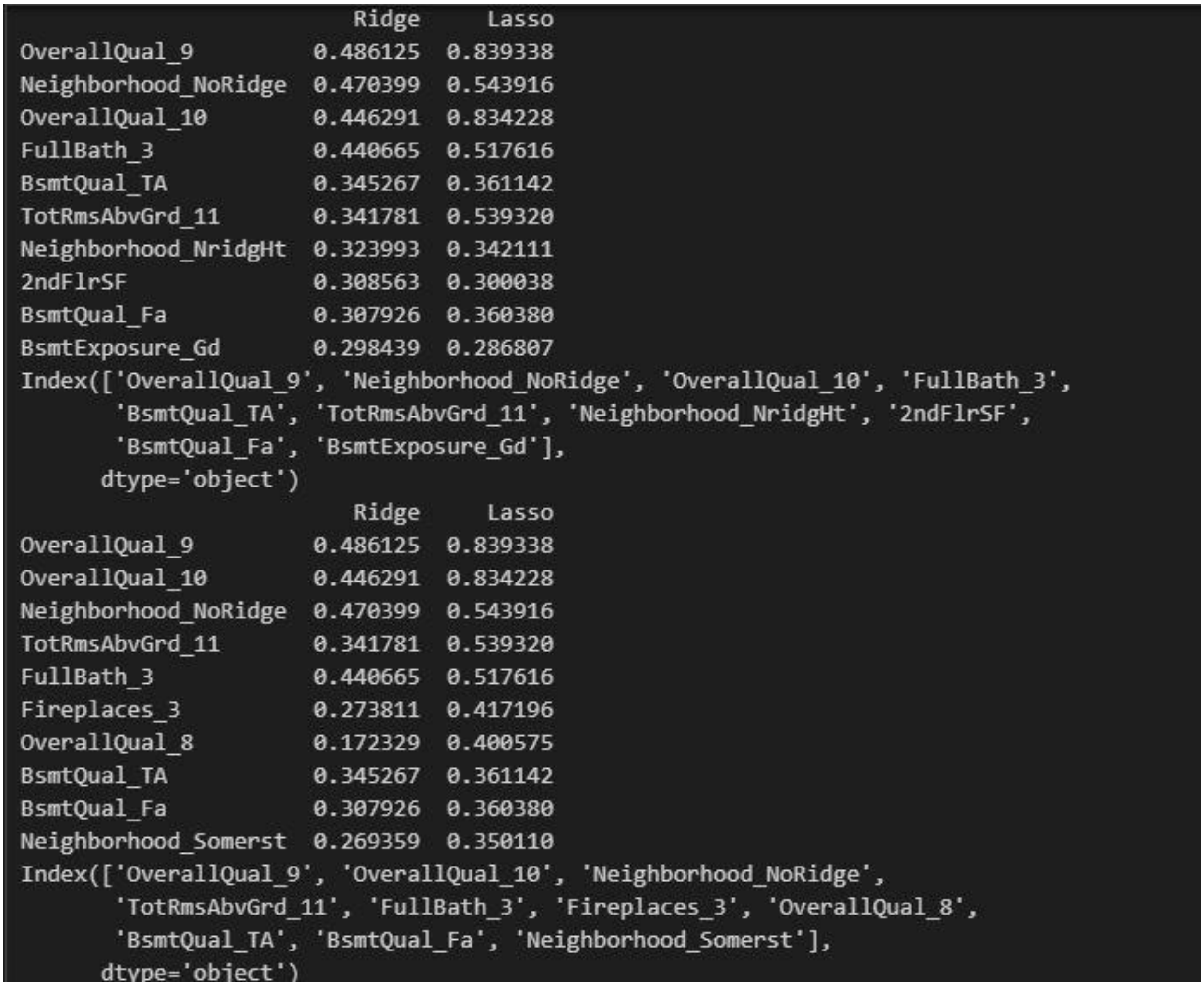
- Ridge Alpha : 6.0
- Lasso Alpha : 0.001

As we increase the value of alpha for ridge and lasso the model will become simpler i.e., the bias will increase and variance will decrease

As we increase the value of λ then the magnitude of the coefficients decreases.

Most Important Predictor Variables Before Change :

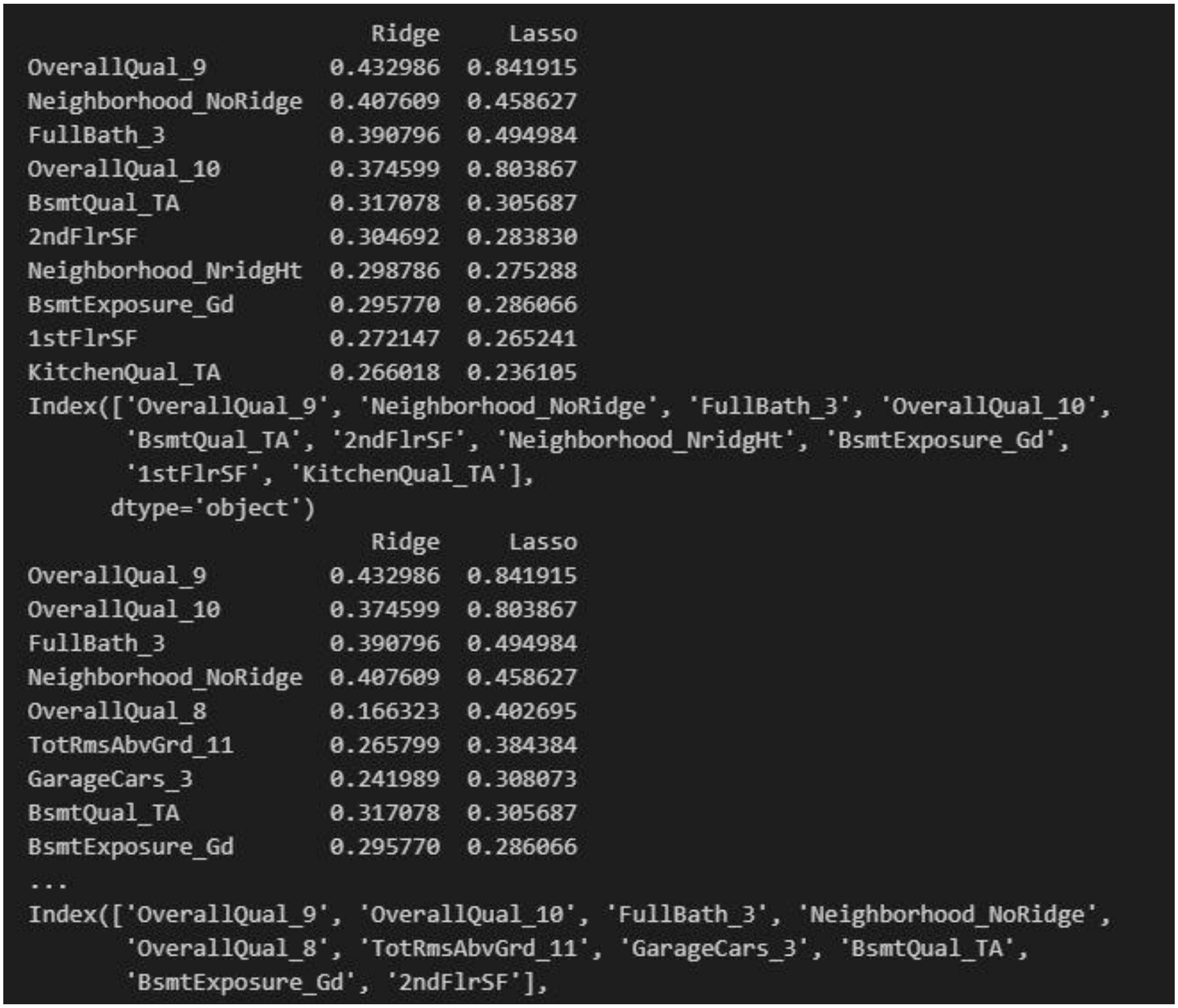
First Half Variables are the sorted using absolute ridge coefficients Second Half Variables are the sorted using absolute lasso co-efficients



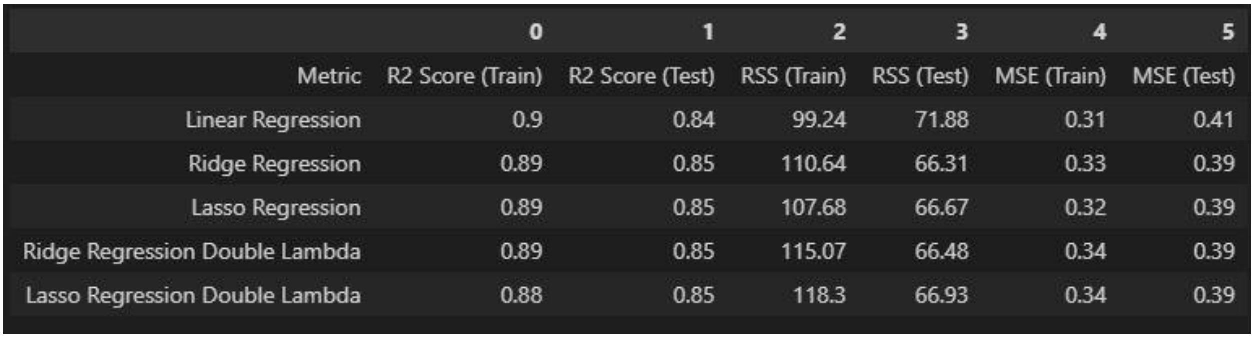
```
                         Ridge     Lasso
OverallQual_9         0.486125  0.839338
Neighborhood_NoRidge  0.470399  0.543916
OverallQual_10        0.446291  0.834228
FullBath_3            0.440665  0.517616
BsmtQual_TA           0.345267  0.361142
TotRmsAbvGrd_11       0.341781  0.539320
Neighborhood_NridgHt  0.323993  0.342111
2ndFlrSF              0.308563  0.300038
BsmtQual_Fa           0.307926  0.360380
BsmtExposure_Gd       0.298439  0.286807
Index(['OverallQual_9', 'Neighborhood_NoRidge', 'OverallQual_10', 'FullBath_3',
       'BsmtQual_TA', 'TotRmsAbvGrd_11', 'Neighborhood_NridgHt', '2ndFlrSF',
       'BsmtQual_Fa', 'BsmtExposure_Gd'],
      dtype='object')
                         Ridge     Lasso
OverallQual_9         0.486125  0.839338
OverallQual_10        0.446291  0.834228
Neighborhood_NoRidge  0.470399  0.543916
TotRmsAbvGrd_11       0.341781  0.539320
FullBath_3            0.440665  0.517616
Fireplaces_3          0.273811  0.417196
OverallQual_8         0.172329  0.400575
BsmtQual_TA           0.345267  0.361142
BsmtQual_Fa           0.307926  0.360380
Neighborhood_Somerst  0.269359  0.350110
Index(['OverallQual_9', 'OverallQual_10', 'Neighborhood_NoRidge',
       'TotRmsAbvGrd_11', 'FullBath_3', 'Fireplaces_3', 'OverallQual_8',
       'BsmtQual_TA', 'BsmtQual_Fa', 'Neighborhood_Somerst'],
      dtype='object')
```

Most Important Predictor Variables After Change :

First Half Variables are the sorted using absolute ridge coefficients Second Half Variables are the sorted using absolute lasso coefficients



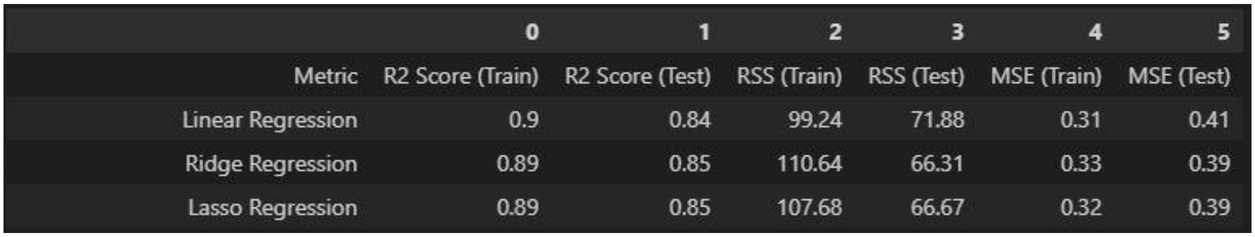
```
                          Ridge     Lasso
OverallQual_9          0.432986  0.841915
Neighborhood_NoRidge   0.407609  0.458627
FullBath_3             0.390796  0.494984
OverallQual_10         0.374599  0.803867
BsmtQual_TA            0.317078  0.305687
2ndFlrSF               0.304692  0.283830
Neighborhood_NridgHt   0.298786  0.275288
BsmtExposure_Gd        0.295770  0.286066
1stFlrSF               0.272147  0.265241
KitchenQual_TA         0.266018  0.236105
Index(['OverallQual_9', 'Neighborhood_NoRidge', 'FullBath_3', 'OverallQual_10',
       'BsmtQual_TA', '2ndFlrSF', 'Neighborhood_NridgHt', 'BsmtExposure_Gd',
       '1stFlrSF', 'KitchenQual_TA'],
      dtype='object')
                          Ridge     Lasso
OverallQual_9          0.432986  0.841915
OverallQual_10         0.374599  0.803867
FullBath_3             0.390796  0.494984
Neighborhood_NoRidge   0.407609  0.458627
OverallQual_8          0.166323  0.402695
TotRmsAbvGrd_11        0.265799  0.384384
GarageCars_3           0.241989  0.308073
BsmtQual_TA            0.317078  0.305687
BsmtExposure_Gd        0.295770  0.286066
...
Index(['OverallQual_9', 'OverallQual_10', 'FullBath_3', 'Neighborhood_NoRidge',
       'OverallQual_8', 'TotRmsAbvGrd_11', 'GarageCars_3', 'BsmtQual_TA',
       'BsmtExposure_Gd', '2ndFlrSF'],
```



	0	1	2	3	4	5
Metric	R2 Score (Train)	R2 Score (Test)	RSS (Train)	RSS (Test)	MSE (Train)	MSE (Test)
Linear Regression	0.9	0.84	99.24	71.88	0.31	0.41
Ridge Regression	0.89	0.85	110.64	66.31	0.33	0.39
Lasso Regression	0.89	0.85	107.68	66.67	0.32	0.39
Ridge Regression Double Lambda	0.89	0.85	115.07	66.48	0.34	0.39
Lasso Regression Double Lambda	0.88	0.85	118.3	66.93	0.34	0.39

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?



	0	1	2	3	4	5
Metric	R2 Score (Train)	R2 Score (Test)	RSS (Train)	RSS (Test)	MSE (Train)	MSE (Test)
Linear Regression	0.9	0.84	99.24	71.88	0.31	0.41
Ridge Regression	0.89	0.85	110.64	66.31	0.33	0.39
Lasso Regression	0.89	0.85	107.68	66.67	0.32	0.39

The optimal values for Ridge and Lasso regression using top 120 variables selected by RFE are :

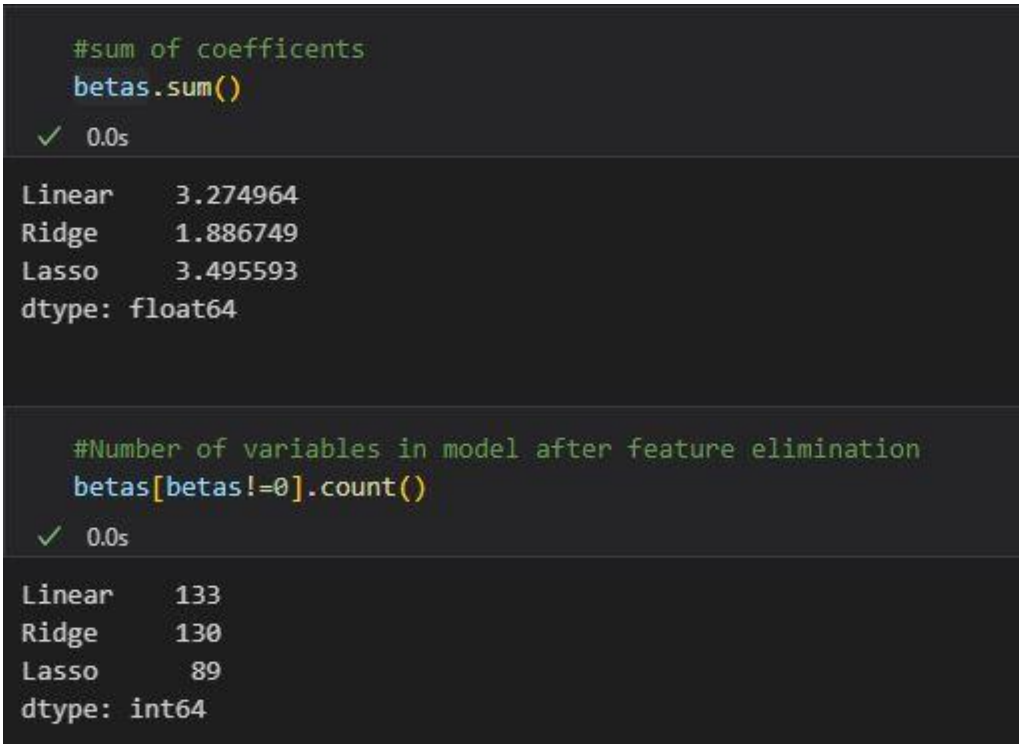
- Ridge Alpha : 6.0
- Lasso Alpha : 0.001

Difference in r2 values for train and test data have improved from 0.6 to 0.4.

The RSS for the test data has reduced from 71.88 to 66.31 and 66.67 for Ridge and Lasso respectively (lower the value better)

The MSE for the test data has reduced from 71.88 to 66.31 and 66.67 for Ridge and Lasso respectively (lower the value better)

Also, the Lasso performs feature elimination by making their coefficient 0 indirectly making the model simple, robust and generalized in nature.



```
#sum of coefficents
betas.sum()
```

✓ 0.0s

```
Linear    3.274964
Ridge     1.886749
Lasso     3.495593
dtype: float64
```

```
#Number of variables in model after feature elimination
betas[betas!=0].count()
```

✓ 0.0s

```
Linear    133
Ridge     130
Lasso      89
dtype: int64
```

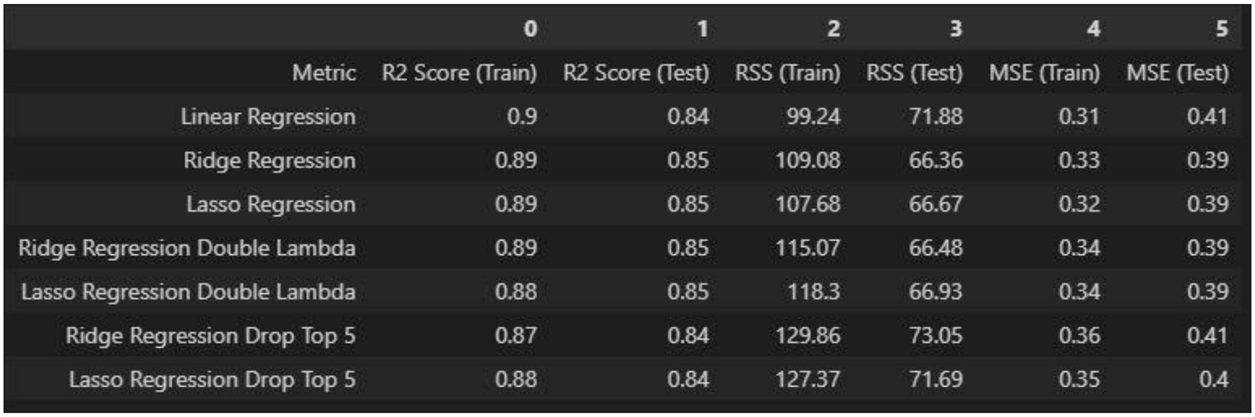
In the given data set Lasso eliminated 3 features from 120 selected by RFE.

In the given data set Lasso eliminated 40 features from 120 selected by RFE.

Clearly the Lasso Regularization will produce a simple model and yield better results on unseen data.

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?



	0	1	2	3	4	5
Metric	R2 Score (Train)	R2 Score (Test)	RSS (Train)	RSS (Test)	MSE (Train)	MSE (Test)
Linear Regression	0.9	0.84	99.24	71.88	0.31	0.41
Ridge Regression	0.89	0.85	109.08	66.36	0.33	0.39
Lasso Regression	0.89	0.85	107.68	66.67	0.32	0.39
Ridge Regression Double Lambda	0.89	0.85	115.07	66.48	0.34	0.39
Lasso Regression Double Lambda	0.88	0.85	118.3	66.93	0.34	0.39
Ridge Regression Drop Top 5	0.87	0.84	129.86	73.05	0.36	0.41
Lasso Regression Drop Top 5	0.88	0.84	127.37	71.69	0.35	0.4

There is a drop in r2 values for train and test data

There is increase in RSS and MSE values for train and test data

The Top 5 predictors after changes

Using Ridge :

- 'OverallQual_5'
- 'OverallQual_6'
- 'OverallQual_4'
- 'KitchenQual_TA'
- 'KitchenQual_Fa'

Using Lasso :

- 'OverallQual_5’
- 'OverallQual_4'
- 'OverallQual_6'
- 'OverallQual_3'
- 'Fireplaces_3'

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?

We can use L1 and L2 Regularization to make sure that the model is robust and generalisable.

A model is considered to be robust if its output is consistently accurate even if one or more of the input variables changed.

A model is considered to be generalisable if it is able to predict on the complete population using a model trained by a subset of the data

Simpler models are usually more ’generic’.

Simpler models are more robust.

Typical ways of looking the complexity of a model.

1. Number of parameters required to specify the model completely.
2. The degree of the function, if it is a polynomial
3. Size of the best-possible representation of the model.
4. The depth or size of a decision tree.

The difference in r2 for train and test is small

The variance and bias should be as low as possible (Bias Variance Trade-off)

The model should not be complex i.e., the degree of polynomial or depth of the tree should be as small

Implications:

- The accuracy of the model on train set might go down
- The accuracy of the model on test set might go up