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 1:
The optimal values of alpha for ridge and lasso regression are:

- Lasso: 0.001
- Ridge: 0.9

In [63]:
```
## What if alpha is doubled?
### Lasso
lasso = Lasso(alpha=0.002)
lasso.fit(X_train,y_train)

y_train_pred = lasso.predict(X_train)
y_test_pred = lasso.predict(X_test)

print(r2_score(y_true=y_train,y_pred=y_train_pred))
print(r2_score(y_true=y_test,y_pred=y_test_pred))
```

```
0.8985698622551942
0.8620223561693784
```

In [64]:
```
### Ridge
ridge = Ridge(alpha = 1.8)
ridge.fit(X_train,y_train)

y_pred_train = ridge.predict(X_train)
print(r2_score(y_train,y_pred_train))

y_pred_test = ridge.predict(X_test)
print(r2_score(y_test,y_pred_test))
```

```
0.9050902361445005
0.859613643141553
```

Ridge Previous Values (alpha 0.9):

0.9084582758574987 0.85557107117059

Ridge New Values (alpha: 1.8):

0.9050902361445005 0.859613643141553

Lasso Previous Values (alpha 0.001):

0.9046654994563071 0.8583873386096132

Lasso New Values (alpha: 0.002):

0.8985698622551942 0.8620223561693784

We can conclude that lasso is affected more with the alpha change.

Predictor variables:

Ridge Previously (alpha - 0.9)

	Feaure	Coef
43	PoolArea	1.454440
27	HalfBath	0.525808
30	KitchenQual	0.432617
67	Neighborhood_Gilbert	0.341084
26	FullBath	0.326894
23	GrLivArea	0.322173
38	WoodDeckSF	0.307462
19	CentralAir	0.294278
68	Neighborhood_IDOTRR	0.261283
17	TotalBsmtSF	0.254831

Ridge after doubling alpha (alpha - 1.8)

	Feaure	Coef
43	PoolArea	1.203934
27	HalfBath	0.504115
30	KitchenQual	0.395646
26	FullBath	0.325126
23	GrLivArea	0.310162
67	Neighborhood_Gilbert	0.268174
19	CentralAir	0.232578
33	GarageFinish	0.218103
44	MiscVal	0.214884
55	LandContour_Lvl	0.204914

—

Lasso previously (alpha - 0.001)

	Featuere	Coef
43	PoolArea	1.473640
27	HalfBath	0.483000
30	KitchenQual	0.325463
26	FullBath	0.297824
23	GrLivArea	0.293612
14	BsmtFinType2	0.291309
67	Neighborhood_Gilbert	0.261915
2	LotShape	0.191473
44	MiscVal	0.189479
22	LowQualFinSF	0.176111

Lasso after doubling alpha (alpha - 0.002)

	Featuere	Coef
43	PoolArea	1.141892
27	HalfBath	0.427726
14	BsmtFinType2	0.291887
26	FullBath	0.268383
23	GrLivArea	0.259778
30	KitchenQual	0.234510
67	Neighborhood_Gilbert	0.200943
2	LotShape	0.200612
44	MiscVal	0.162882
22	LowQualFinSF	0.155019

We can clearly see how the features have changed in both cases after doubling the alpha value.

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 2:
In the assignment, we observed that the r2_scores are similar for both models but we went with lasso regression as it would penalize more on the dataset and would be helpful in feature elimination thus, being a robust model for the use case.

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 3:
The five most important predictor variables in the lasso model are:

- PoolArea
- HalfBath
- KitchenQual
- FullBath
- GrLivArea

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?

Answer 4:
For a model to be unaffected by outliers in the training data, it must be made robust and generalizable. It should be general to account for datasets which were not in the training dataset, and also the test data accuracy should not be significantly less than the train data score. The outliers shouldn't be given an excessive amount of weight in order to maintain a high level of model accuracy. Only those outliers that are pertinent to the dataset should be kept once the outlier analysis is completed to ensure that this is not the case. The dataset must be cleaned up of any outliers that don't make sense to preserve. We need to improve the model's ability to forecast outcomes accurately by employing confidence intervals. For predictive analysis, a weak model cannot be relied upon, thus it is extremely important for the model to be robust.