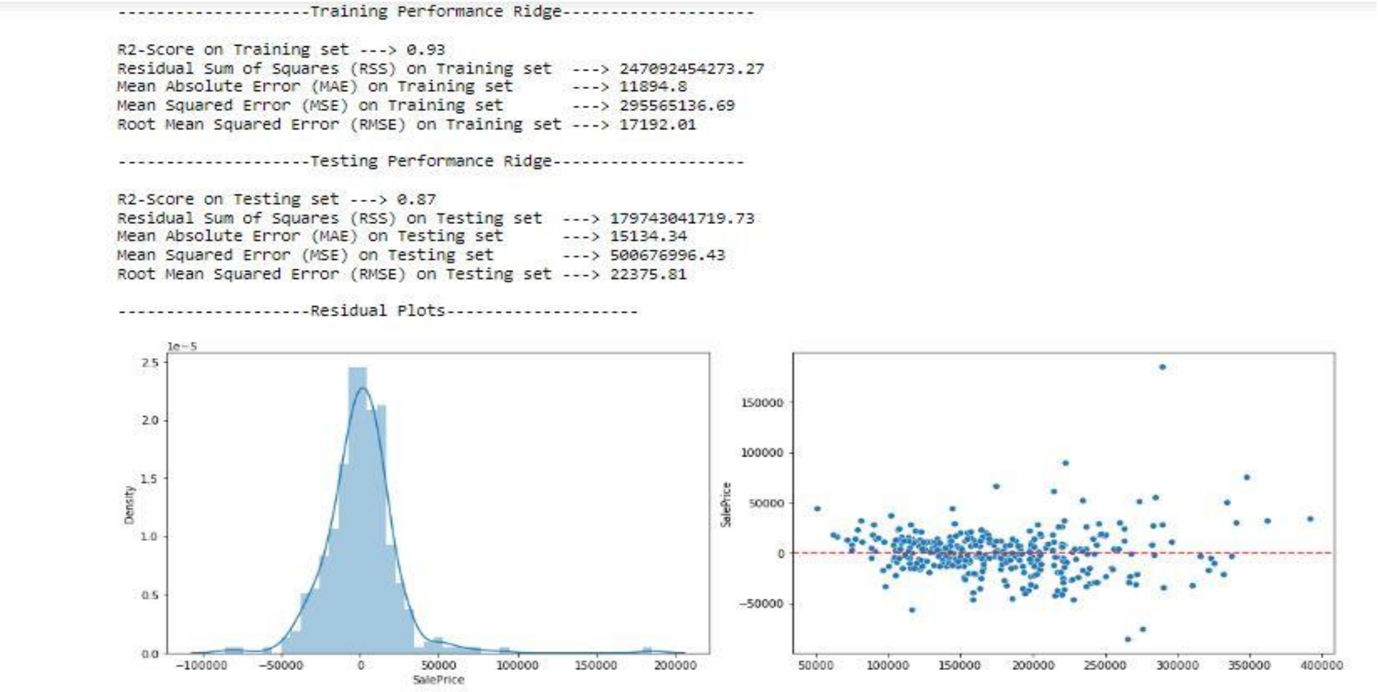
Question 1

<u>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?</u>

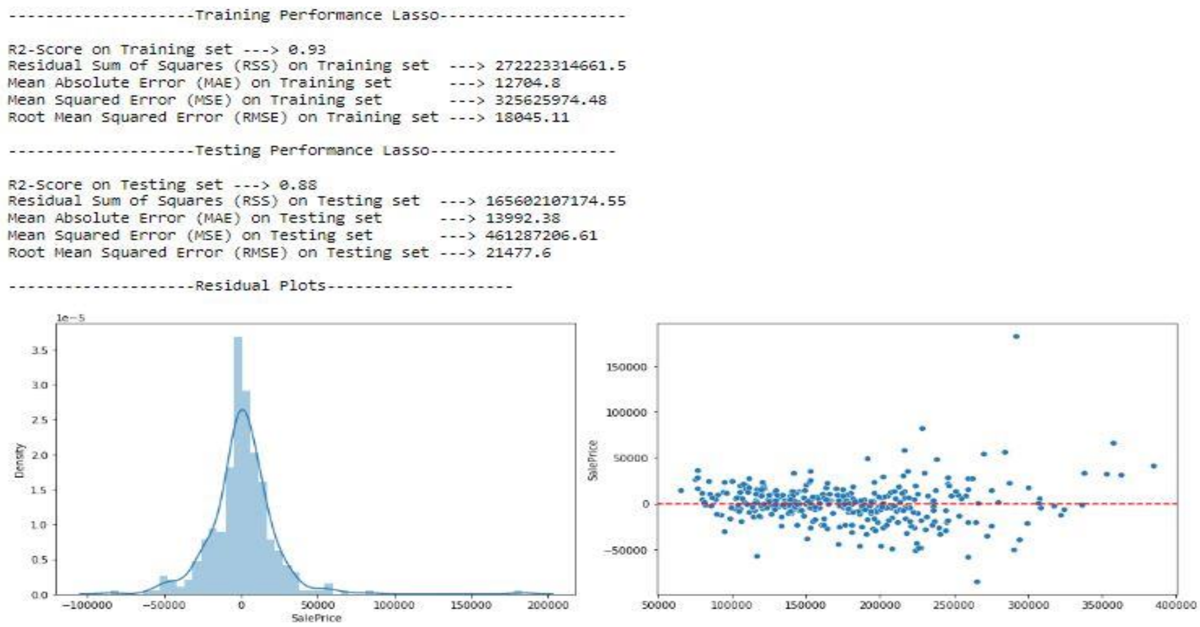
a) The optimal values of alpha/lambda for ridge & lasso regression are 159 & 560 respectively.

b) Upon doubling the alpha values, the model's performance decreased on the training set, while it performed a little better on the testing set, as it had increased the penalising effect. The detailed Results are shown in the below table.

c) We observe few changes in the ranking & overall reduction in the individual contributions towards the target variable. Additionally in case of Lasso, the overall number features are also reduced. It can be clearly observed in the below table.



```
--------------------Training Performance Ridge--------------------

R2-Score on Training set ---> 0.93
Residual Sum of Squares (RSS) on Training set  ---> 247092454273.27
Mean Absolute Error (MAE) on Training set      ---> 11894.8
Mean Squared Error (MSE) on Training set       ---> 295565136.69
Root Mean Squared Error (RMSE) on Training set ---> 17192.01

--------------------Testing Performance Ridge--------------------

R2-Score on Testing set ---> 0.87
Residual Sum of Squares (RSS) on Testing set  ---> 179743041719.73
Mean Absolute Error (MAE) on Testing set      ---> 15134.34
Mean Squared Error (MSE) on Testing set       ---> 500676996.43
Root Mean Squared Error (RMSE) on Testing set ---> 22375.81

--------------------Residual Plots--------------------
```



```
--------------------Training Performance Lasso--------------------

R2-Score on Training set ---> 0.93
Residual Sum of Squares (RSS) on Training set  ---> 272223314661.5
Mean Absolute Error (MAE) on Training set      ---> 12704.8
Mean Squared Error (MSE) on Training set       ---> 325625974.48
Root Mean Squared Error (RMSE) on Training set ---> 18045.11

--------------------Testing Performance Lasso--------------------

R2-Score on Testing set ---> 0.88
Residual Sum of Squares (RSS) on Testing set  ---> 165602107174.55
Mean Absolute Error (MAE) on Testing set      ---> 13992.38
Mean Squared Error (MSE) on Testing set       ---> 461287206.61
Root Mean Squared Error (RMSE) on Testing set ---> 21477.6

--------------------Residual Plots--------------------
```

Question 2

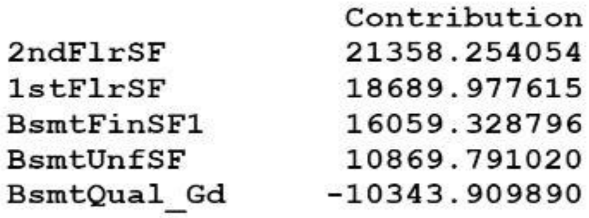
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Both Ridge and Lasso Regression gave us optimal predictions. But it is more preferred to pick **Lasso Regression**, as it also performed feature selection, by removing undesired features (originally from 262 to 122 features, dropping more than 50% unimportant features). This means that the model is more generalised & would perform significantly better on the unseen data (also demonstrated on the test set in the assignment notebook).

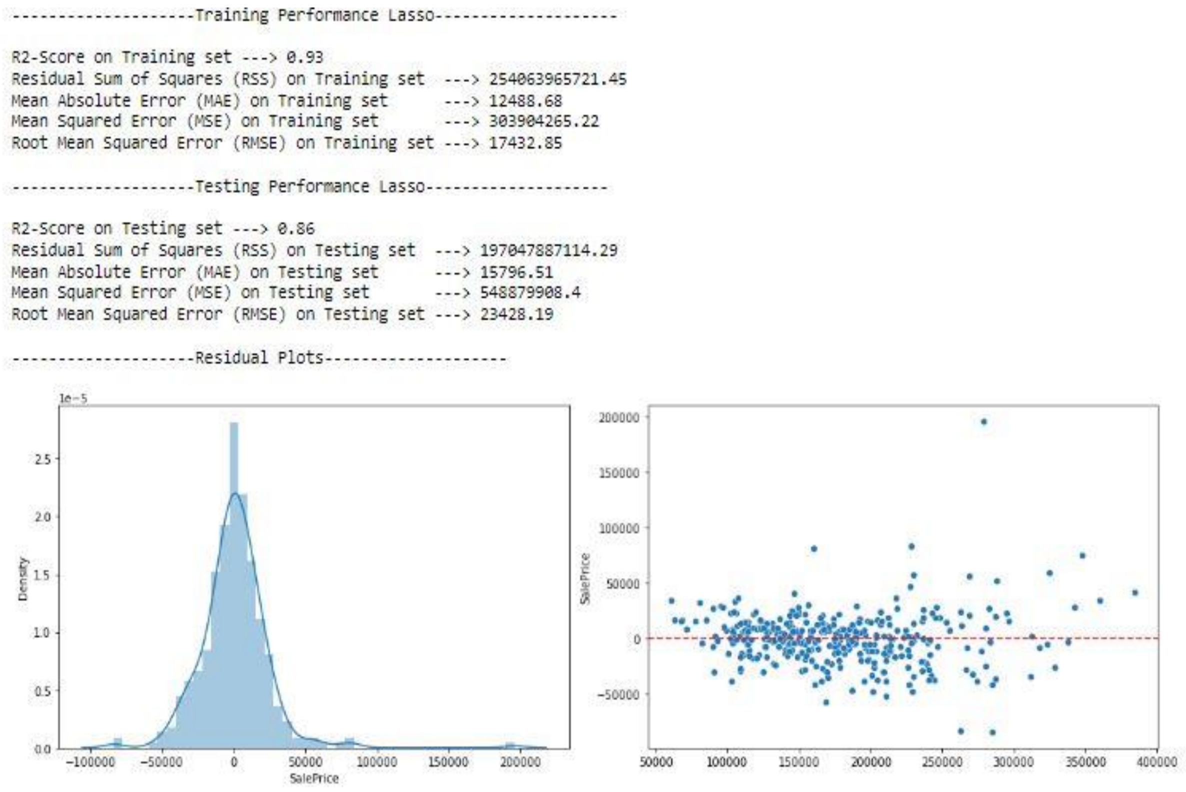
Question 3

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?

Upon dropping top 5 variables, we observe that model's performance deteriorated
& Their corresponding r2_scores decreased. Also, the five most import variables now are:



	Contribution
2ndFlrSF	21358.254054
1stFlrSF	18689.977615
BsmtFinSF1	16059.328796
BsmtUnfSF	10869.791020
BsmtQual_Gd	-10343.909890

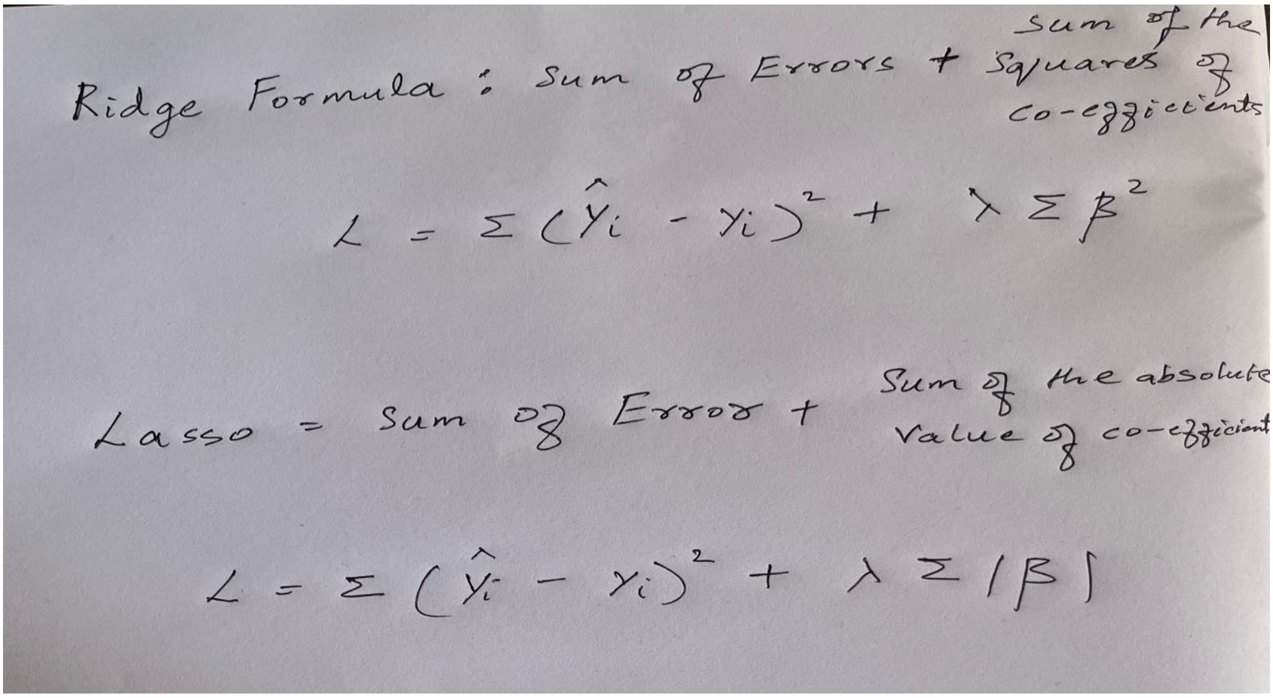


```
--------------------Training Performance Lasso--------------------

R2-Score on Training set ---> 0.93
Residual Sum of Squares (RSS) on Training set  ---> 254063965721.45
Mean Absolute Error (MAE) on Training set      ---> 12488.68
Mean Squared Error (MSE) on Training set       ---> 303904265.22
Root Mean Squared Error (RMSE) on Training set ---> 17432.85

--------------------Testing Performance Lasso--------------------

R2-Score on Testing set ---> 0.86
Residual Sum of Squares (RSS) on Testing set  ---> 197047887114.29
Mean Absolute Error (MAE) on Testing set      ---> 15796.51
Mean Squared Error (MSE) on Testing set       ---> 548879908.4
Root Mean Squared Error (RMSE) on Testing set ---> 23428.19

--------------------Residual Plots--------------------
```

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?

A simple linear regression model makes predictions taking into account of all the features/variables fed to it. But it is not recommended as it can often lead to overfitting problems. Hence Ridge & Lasso Regression were introduced. Both Ridge & Lasso Regression add a penalising term to avoid the model from overfitting.



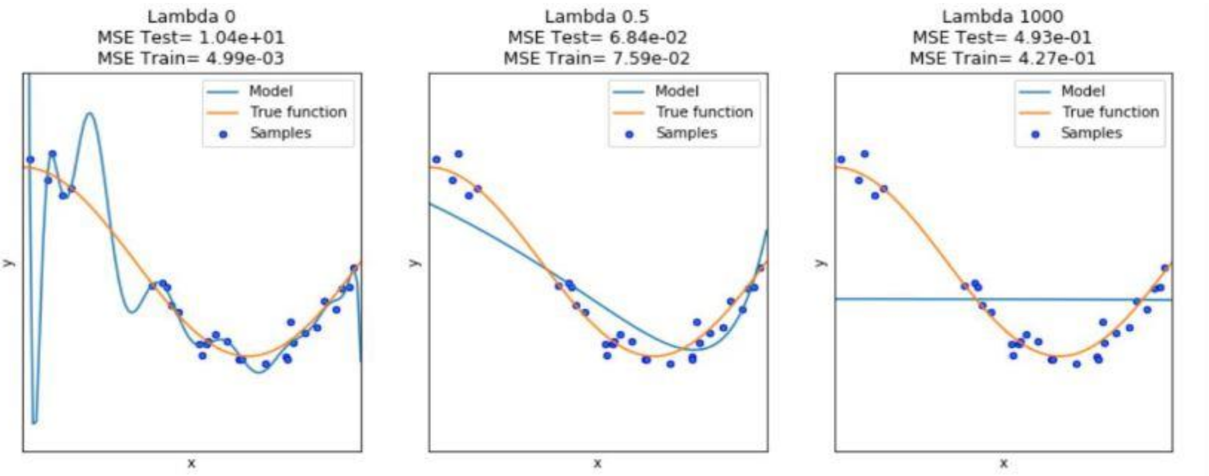
Ridge Formula : Sum of Errors + sum of the Squares of co-efficients

$$L = \Sigma(\hat{Y_i} - Y_i)^2 + \lambda \Sigma \beta^2$$

Lasso = Sum of Error + Sum of the absolute Value of co-efficient

$$L = \Sigma(\hat{Y_i} - Y_i)^2 + \lambda \Sigma |\beta|$$

a) To make the model more robust & generalizable, we can increase the penalising effect (higher alpha value), making it simpler, reducing the overall variance of the model, by small compromise in the bias.

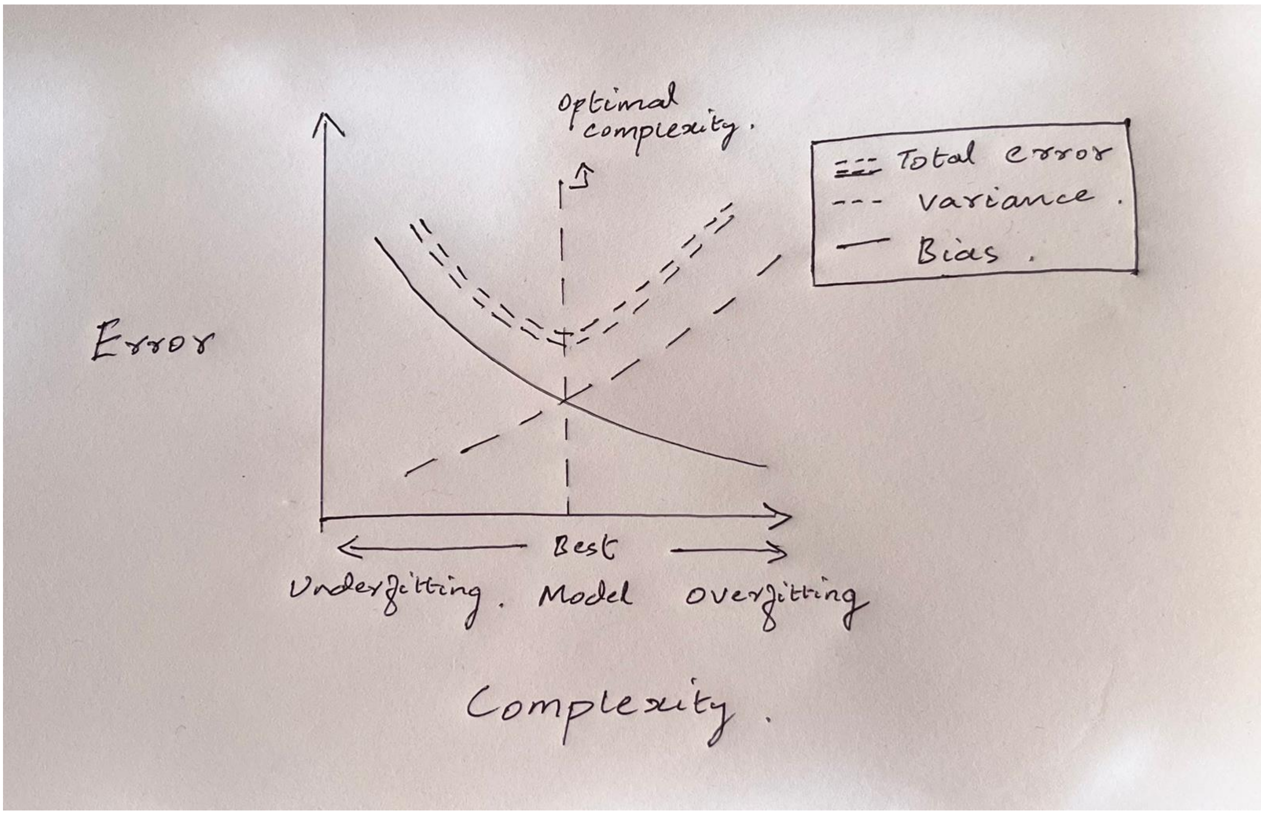


Lambda 0
MSE Test= 1.04e+01
MSE Train= 4.99e-03

Lambda 0.5
MSE Test= 6.84e-02
MSE Train= 7.59e-02

Lambda 1000
MSE Test= 4.93e-01
MSE Train= 4.27e-01

b) But the major implications of making the model more generalised, that it will reduce the model performance (RMSE) on the training set. This happens because the penalising effect of alpha, prevents the model from memorising the training data, making it less complex. This will lead to lower variance & slightly higher bias. Hence by iterating, we can find the optimal value for the same.


Optimal complexity.
Total error
Variance.
Bias.
Error
Best Model
Underfitting.
overfitting
Complexity.