Advanced Regression Assignment

House price prediction

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?

For Ridge alpha is **7.0**
For Lasso alpha is **0.0001**

There are slight changes like
R-squared (test and train) for ridge and lasso are slightly decreased.
RSS (train and test) for ridge and lasso are slightly increased.
No effect on MSE (ridge and lasso)
Also coefficients are slightly changed

	Metric	Linear Regression	Ridge regression	Lasso regression
0	R2 Score (Train)	9.528242e-01	0.895080	0.912206
1	R2 Score (Test)	-1.876279e+23	0.856918	0.848948
2	RSS(Train)	5.806048e-01	1.291272	1.080499
3	RSS(Test)	1.610540e+24	1.228174	1.296583
4	MSE(Train)	5.686628e-04	0.001265	0.001058
5	MSE(Test)	3.668655e+21	0.002798	0.002953

alpha_ridge = 7.0
alpha_lasso = 0.0001

	Metric	Linear Regression	Ridge regression	Lasso regression
0	R2 Score (Train)	9.528242e-01	0.881095	0.897867
1	R2 Score (Test)	-1.876279e+23	0.842200	0.849579
2	RSS(Train)	5.806048e-01	1.463399	1.256974
3	RSS(Test)	1.610540e+24	1.354503	1.291168
4	MSE(Train)	5.686628e-04	0.001433	0.001231
5	MSE(Test)	3.668655e+21	0.003085	0.002941

alpha_ridge = 14.0
alpha_lasso = 0.0002
(doubled)

The predictor variables are mostly same.

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?

We will choose Lasso regression with alpha 0.0001.
This somewhat helps in feature selection by forcing some coefficients.
Also R-Squared is higher and RSS is low.

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?

GirLivArea
OverallQual_VeryExcel
RoofMatl_WdShngl
OverallQual_Excel
Neighborhood_NoRidge

These are 5 most important features.

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?

The model accuracy for Training set is **89.78%** and for Test set is **84.95%**