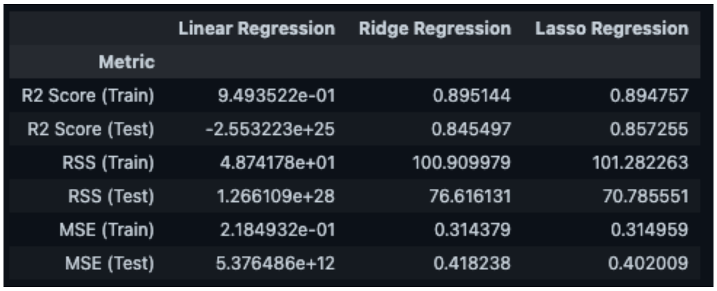
Question 1:

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

Answer:
The optimal value of alpha for ridge and lasso models are the best regulazation level that balance the bias-variance tradeoff, it plays as feature selectors which attempts to penalize any large value of coefficients; as a result, it counters the possible high-variance characteristics of linear regression. The optimal value of alpha calculated for ridge and lasso regression provided by this table:

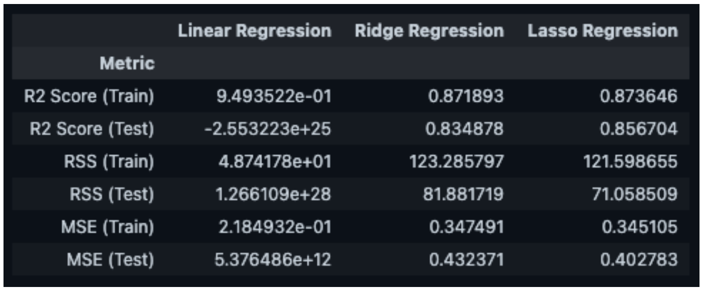
	Ridge	**Lasso**
Optimal alpha	500	0.01

The effect of doubling the value of alpha for both ridge and lasso can be described as below.



	Linear Regression	Ridge Regression	Lasso Regression
Metric			
R2 Score (Train)	9.493522e-01	0.895144	0.894757
R2 Score (Test)	-2.553223e+25	0.845497	0.857255
RSS (Train)	4.874178e+01	100.909979	101.282263
RSS (Test)	1.266109e+28	76.616131	70.785551
MSE (Train)	2.184932e-01	0.314379	0.314959
MSE (Test)	5.376486e+12	0.418238	0.402009

(Optimal alphas)



	Linear Regression	Ridge Regression	Lasso Regression
Metric			
R2 Score (Train)	9.493522e-01	0.871893	0.873646
R2 Score (Test)	-2.553223e+25	0.834878	0.856704
RSS (Train)	4.874178e+01	123.285797	121.598655
RSS (Test)	1.266109e+28	81.881719	71.058509
MSE (Train)	2.184932e-01	0.347491	0.345105
MSE (Test)	5.376486e+12	0.432371	0.402783

(Doubling alphas from its optimal values)

As we can see, doubling alphas from its optimal values slightly deteriorates the model's performance on both train set and test set

We can identify important predictors by looking at the magnitude of coefficients. Variables with great (negative and positive) coefficients including:

- OverallQual
- GrLivArea
- GarageCars
- Neighborhood_NridgHt
- RoofMatl_WdShngl

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:
I choose alpha = 500 for the Ridge model and alpha = 0.01 for the Lasso model because they are optimal in GridSearchCV with Mean Absolute Error.

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: To answer the question, some code is needed to first remove 5 variables in question 1, then re-train the model with remaining variables. The code can be found in the Jupyter Notebook. Here are the five most important predictor variables after the exclusion.

- 1stFlrSF
- 2ndFlrSF
- GarageArea
- FullBath
- TotRmsAbvGrd

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: Robustness and generalizability of a model can be evaluated base on the model's performance on train set and test set. If the model performs equivalently good on both sets, we can conclude the model is robust and generalizable. The accuracy of the model can be evaluated by the model performance on test set. If most of the unseen data fed into the model produces accurate predictions, we then can say the model has high accuracy