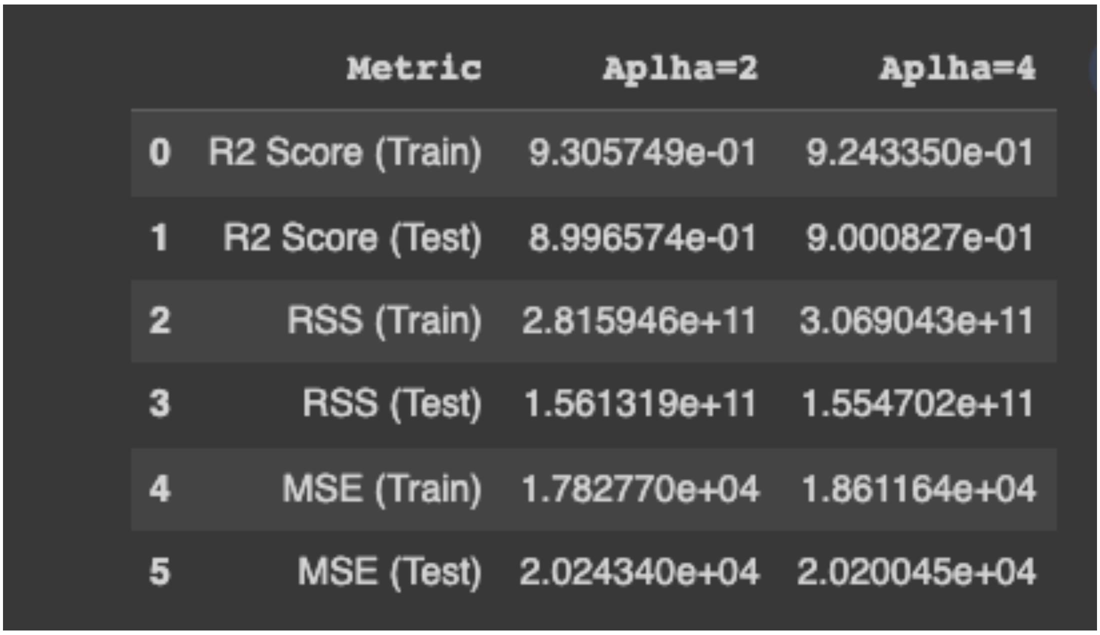
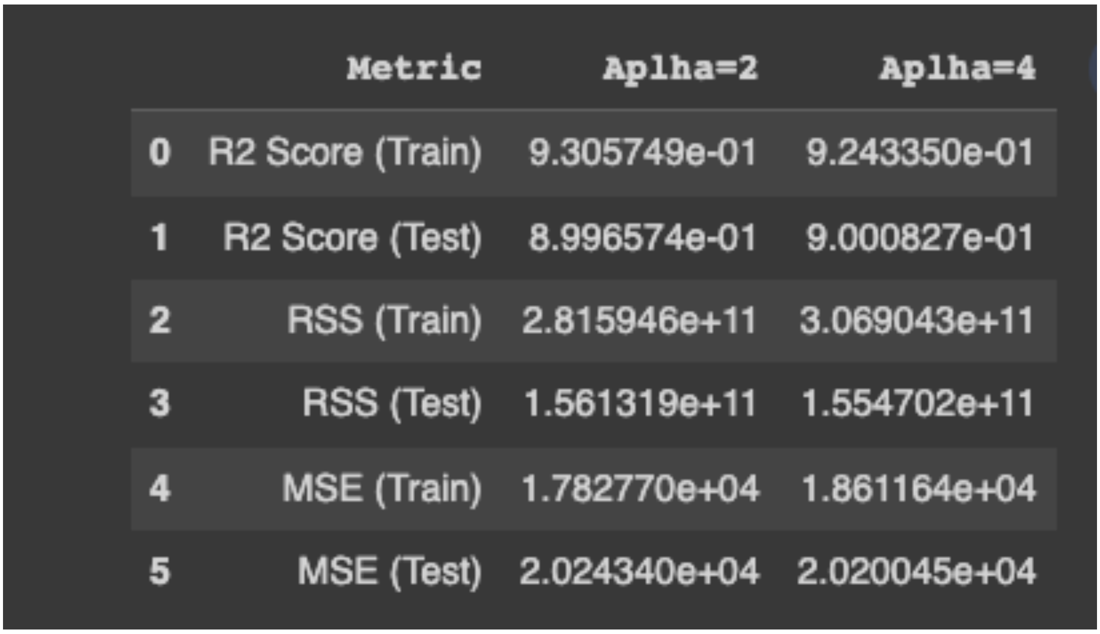
Question 1
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

- Optimal value of alpha
  - Ridge : 2.0
  - Lasso : 100
- On Doubling the value of alpha
  - Ridge :
    - Doubled Alpha : 4.0
    - Observation :
      - The R2 value for Train reduced by approx. 0.6% i.e. from 93.1% to 92.4%
      - The error Terms increased a little in Train data
    - 

	Metric	Aplha=2	Aplha=4
0	R2 Score (Train)	9.305749e-01	9.243350e-01
1	R2 Score (Test)	8.996574e-01	9.000827e-01
2	RSS (Train)	2.815946e+11	3.069043e+11
3	RSS (Test)	1.561319e+11	1.554702e+11
4	MSE (Train)	1.782770e+04	1.861164e+04
5	MSE (Test)	2.024340e+04	2.020045e+04

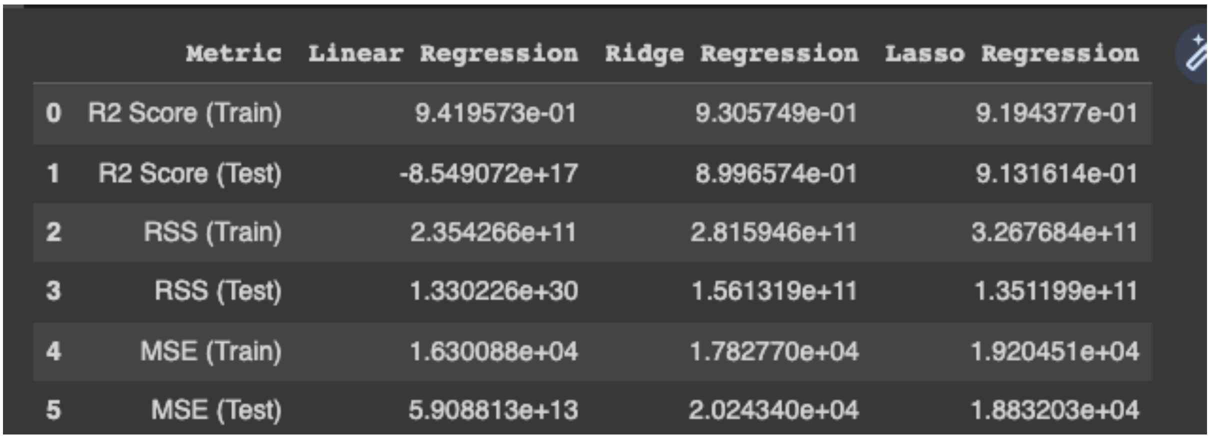
  - Lasso :
    - Doubled Alpha : 200
    - Observation :
      - The R2 value for Train reduced by approx. 1% i.e. from 91.94% to 90.88%
      - The error terms increased in Train data
    - 

	Metric	Aplha=2	Aplha=4
0	R2 Score (Train)	9.305749e-01	9.243350e-01
1	R2 Score (Test)	8.996574e-01	9.000827e-01
2	RSS (Train)	2.815946e+11	3.069043e+11
3	RSS (Test)	1.561319e+11	1.554702e+11
4	MSE (Train)	1.782770e+04	1.861164e+04
5	MSE (Test)	2.024340e+04	2.020045e+04

- Most Important Predictor variables:
  - Ridge : Even after doubling the alpha value, the top 4 predictor variables remains the same GrLivArea, OverallQual, TotalBsmtSF, BsmtFinSF1)
  - Lasso : Even after doubling the alpha value, the top 4 predictor variables remains the same (GrLivArea, OverallQual, TotalBsmtSF, OverallCond)

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?

- 

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.419573e-01	9.305749e-01	9.194377e-01
1	R2 Score (Test)	-8.549072e+17	8.996574e-01	9.131614e-01
2	RSS (Train)	2.354266e+11	2.815946e+11	3.267684e+11
3	RSS (Test)	1.330226e+30	1.561319e+11	1.351199e+11
4	MSE (Train)	1.630088e+04	1.782770e+04	1.920451e+04
5	MSE (Test)	5.908813e+13	2.024340e+04	1.883203e+04

- According to the above metrics, the R2 value of Ridge is better than that of Lasso in train data, but Lasso is more accurate when it comes to ridge.
- Even for Calculating MSE Ridge has lesser value compared to Lasso, but Lasso is more accurate.
- Considering that Lasso's accuracy is better in this particular use case, I would be using Lasso.

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?

- Dropped the following top 5 predictor variables in lasso : 'GrLivArea','OverallQual','TotalBsmtSF','OverallCond','Neighborhood_StoneBr'
- The latest top most important predictor variables are : ‘GarageType_No Garage’,GarageCond_Po’,’ GarageCond_TA’,’’GarageCond_Fa’,’ GarageCond_Gd’

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 model is robust and generalizable when it’s not too complex and avoid overfitting, i.e a simple model should have low variance and high bias.

- When we make model simpler, since it has low variance, the variance in it’s output on test data w.r.t training data will also be less.
- And since it has high bias the accuracy on the future test data will also be high.