SUBJECTIVE QUESTIONS

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 value for ridge is 7 and lasso is 0.0001.

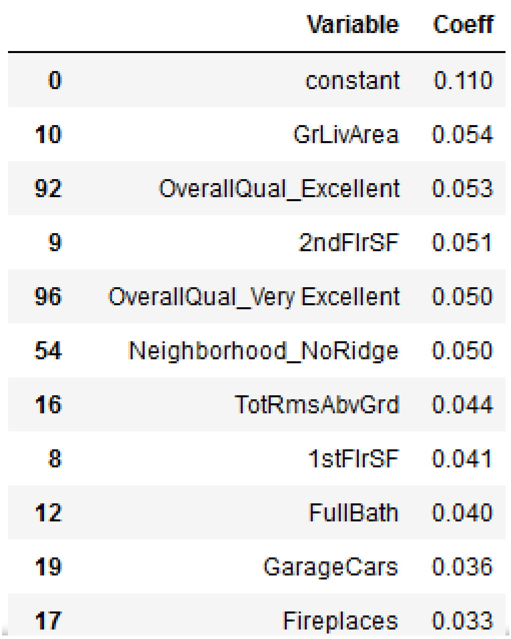
The impact of doubling the value of alpha is mentioned below:

a) Ridge:
R square of train data has decreased and test data has increased. The error terms of test data has decreased and train data has increased.

b) Lasso:
R square of train and test data has decreased. The error terms of train and test data has increased.

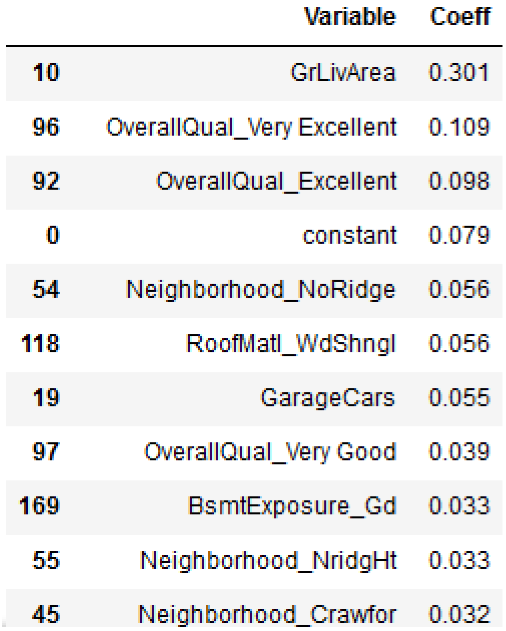
The most important predictor variables after the change is implemented are:

a) Ridge:



	Variable	Coeff
0	constant	0.110
10	GrLivArea	0.054
92	OverallQual_Excellent	0.053
9	2ndFlrSF	0.051
96	OverallQual_Very Excellent	0.050
54	Neighborhood_NoRidge	0.050
16	TotRmsAbvGrd	0.044
8	1stFlrSF	0.041
12	FullBath	0.040
19	GarageCars	0.036
17	Fireplaces	0.033

b) Lasso:



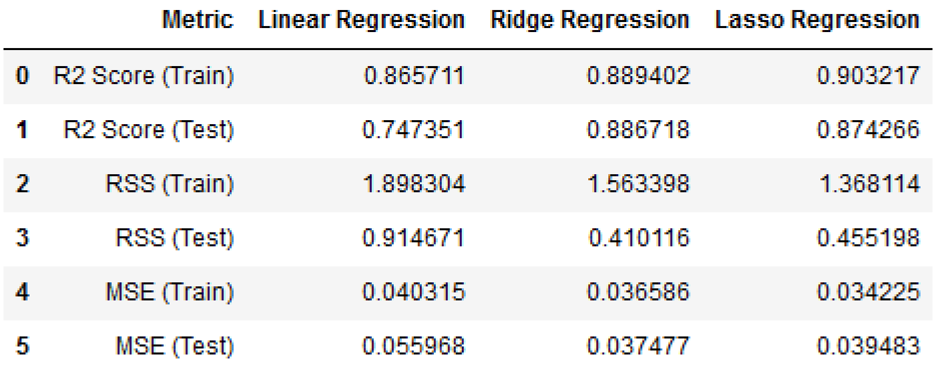
	Variable	Coeff
10	GrLivArea	0.301
96	OverallQual_Very Excellent	0.109
92	OverallQual_Excellent	0.098
0	constant	0.079
54	Neighborhood_NoRidge	0.056
118	RoofMatl_WdShngl	0.056
19	GarageCars	0.055
97	OverallQual_Very Good	0.039
169	BsmtExposure_Gd	0.033
55	Neighborhood_NridgHt	0.033
45	Neighborhood_Crawfor	0.032

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

To choose between ridge and lasso regression we need to compare the values of various parameters which are provided below in a tabular format.



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.865711	0.889402	0.903217
1	R2 Score (Test)	0.747351	0.886718	0.874266
2	RSS (Train)	1.898304	1.563398	1.368114
3	RSS (Test)	0.914671	0.410116	0.455198
4	MSE (Train)	0.040315	0.036586	0.034225
5	MSE (Test)	0.055968	0.037477	0.039483

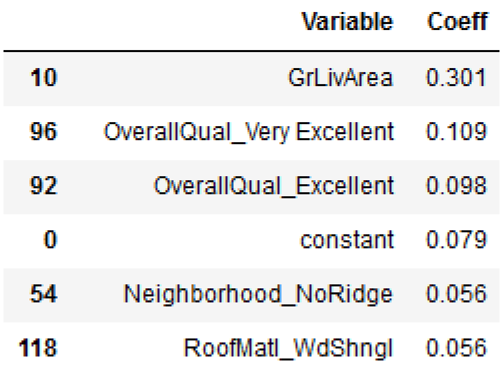
From the table above we can see that Ridge Regression has highest R-square score in test data and minimum error terms in test data. So we will choose Ridge Regression over Lasso Regression.

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

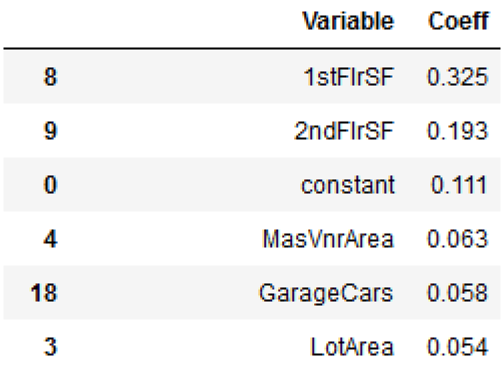
The top 5 predictor variables in the lasso model are listed below:



	Variable	Coeff
10	GrLivArea	0.301
96	OverallQual_Very Excellent	0.109
92	OverallQual_Excellent	0.098
0	constant	0.079
54	Neighborhood_NoRidge	0.056
118	RoofMatl_WdShngl	0.056

As per the question the incoming data does not have the above 5 variables. So we have to create another model excluding the above 5 variables.

After creating the new model the top 5 important predictor variables are ‘1stFlrSF’, ‘2ndFlrSF’, ‘GarageCars’, ‘LotArea’ and ‘MasVnrArea’. The table below provides the coefficients for these variables.



	Variable	Coeff
8	1stFlrSF	0.325
9	2ndFlrSF	0.193
0	constant	0.111
4	MasVnrArea	0.063
18	GarageCars	0.058
3	LotArea	0.054

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

A robust model is one where variation in the data does not have much impact on accuracy.

A generalisable model is one that can easily adapt to previously unseen data (test data).

To make sure a model is robust and generalisable, we have to consider the below points:

1. The model should not overfit. That is the model should not be complex so that it has high train accuracy but low test accuracy. It means that the model should not have high variance.
2. The model should not underfit. That is the model should not be so generalized that the line does not pass through most of the points. It means the model should not have high bias.

If we look at it from the prespective of Accuracy, a too robust model will have very high train accuracy but low test accuracy. If a model is highly generalizable then it will have low accuracy. So, to make our model more robust and generalizable, we have to strike some balance between model bias and variance. Addition of bias means that accuracy will decrease but it will provide the optimum solution.