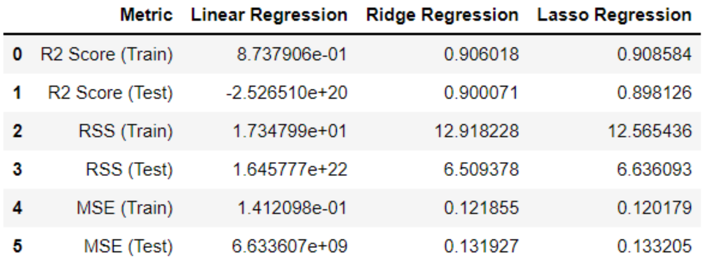
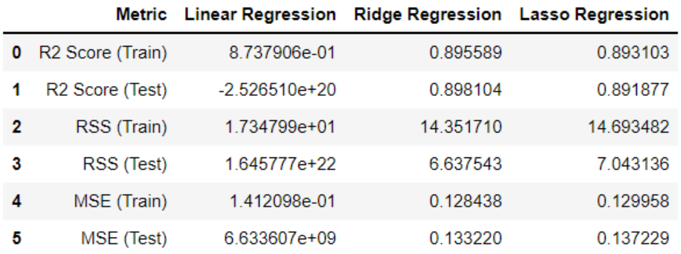
Q1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

A. Before doubling the optimal values for Ridge and Lasso, Ridge_alpha = 50, Lasso_alpha = 0.001



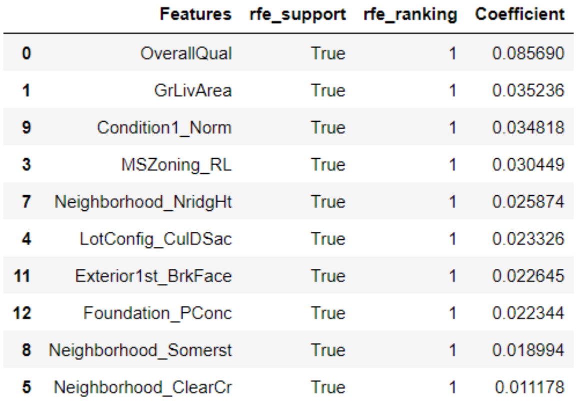
	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.737906e-01	0.906018	0.908584
1	R2 Score (Test)	-2.526510e+20	0.900071	0.898126
2	RSS (Train)	1.734799e+01	12.918228	12.565436
3	RSS (Test)	1.645777e+22	6.509378	6.636093
4	MSE (Train)	1.412098e-01	0.121855	0.120179
5	MSE (Test)	6.633607e+09	0.131927	0.133205

After doubling the optimal alpha values for Ridge and Lasso, Ridge_alpha = 100, Lasso_alpha = 0.002



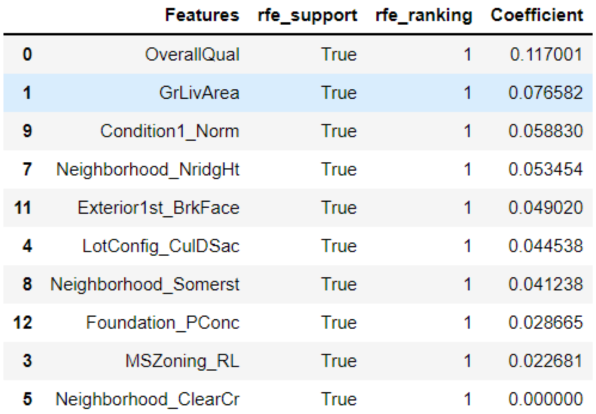
	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.737906e-01	0.895589	0.893103
1	R2 Score (Test)	-2.526510e+20	0.898104	0.891877
2	RSS (Train)	1.734799e+01	14.351710	14.693482
3	RSS (Test)	1.645777e+22	6.637543	7.043136
4	MSE (Train)	1.412098e-01	0.128438	0.129958
5	MSE (Test)	6.633607e+09	0.133220	0.137229

For ridge regression, the important predictors after doubling the value of alpha



	Features	rfe_support	rfe_ranking	Coefficient
0	OverallQual	True	1	0.085690
1	GrLivArea	True	1	0.035236
9	Condition1_Norm	True	1	0.034818
3	MSZoning_RL	True	1	0.030449
7	Neighborhood_NridgHt	True	1	0.025874
4	LotConfig_CulDSac	True	1	0.023326
11	Exterior1st_BrkFace	True	1	0.022645
12	Foundation_PConc	True	1	0.022344
8	Neighborhood_Somerst	True	1	0.018994
5	Neighborhood_ClearCr	True	1	0.011178

For Lasso regression, the important predictors after doubling the value of alpha



	Features	rfe_support	rfe_ranking	Coefficient
0	OverallQual	True	1	0.117001
1	GrLivArea	True	1	0.076582
9	Condition1_Norm	True	1	0.058830
7	Neighborhood_NridgHt	True	1	0.053454
11	Exterior1st_BrkFace	True	1	0.049020
4	LotConfig_CulDSac	True	1	0.044538
8	Neighborhood_Somerst	True	1	0.041238
12	Foundation_PConc	True	1	0.028665
3	MSZoning_RL	True	1	0.022681
5	Neighborhood_ClearCr	True	1	0.000000

Observations:

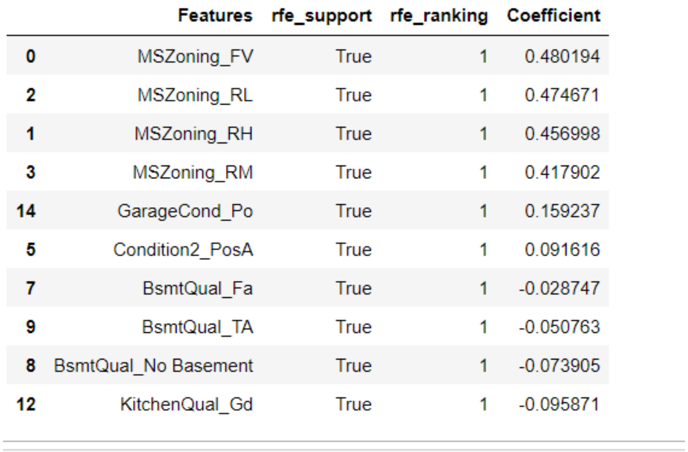
1. R2 Score on the train set got reduced for both Ridge and Lasso regression models when the optimal alpha value is doubled.
2. R2 Score on the test set reduced a bit for Lasso which might be due to zeroing down model coefficients for some predictors.

Q2. 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?

A. Though Ridge Regressor is performing little better compared to Lasso regressor, I prefer to choose Lasso as it helps in feature elimination

Q3. 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?

A.



	Features	rfe_support	rfe_ranking	Coefficient
0	MSZoning_FV	True	1	0.480194
2	MSZoning_RL	True	1	0.474671
1	MSZoning_RH	True	1	0.456998
3	MSZoning_RM	True	1	0.417902
14	GarageCond_Po	True	1	0.159237
5	Condition2_PosA	True	1	0.091616
7	BsmtQual_Fa	True	1	-0.028747
9	BsmtQual_TA	True	1	-0.050763
8	BsmtQual_No Basement	True	1	-0.073905
12	KitchenQual_Gd	True	1	-0.095871

The five most important predictors are MSZoning, Garage Condition, Bsmt Quality, Condition2 and Kitchen Quality

Q4. 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.

The model is considered more robust and generalizable when the model performs well on the unseen data and just not on the data with which it is trained

1. Detecting and treating the outliers in the predictors

2. Imputing the missing values appropriately after outlier treatment with the measures of central tendency based on the nature of predictor (numerical vs categorical)
3. Understand the feature variables in conjunction with the domain before dropping the features. Don't drop the predictors just because the data in the predictors are skewed (We need to bother only if the target variable is skewed as it might give inclined results). Dropping the predictors unnecessarily will reduce the predictive power of the model.
4. Scaling the predictors and target variable using scaling techniques
5. Perform Cross Validation using K-fold to detect if there is a model overfit

Implications:

1. The model accuracy reduces when we the outliers are not treated
2. The model accuracy reduces when missing values are not properly imputed
3. The model accuracy reduces when we drop the features which has predictive power
4. The model accuracy reduces when the predictors and target variable are not in the same scale. The model coefficients will be very large for few and very small for others which may lead to certain model coefficients (which is very small) to be insignificant
5. Apply Regularisation techniques if overfit is detected so that model accuracy improves