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?

In the current dataset, the Optimal value of alpha:

- For Ridge Regression : 0. 1
- For Lasso Regression: 0.00001

When we are preforming the Ridge and Lasso Regularization, we penalize the model in order to keep the most important features. It works by penalizing the magnitude of coefficients of features along with minimizing the error between predicted and actual observations. This can be particularly important when you have a dataset with high number of features. Both the Ridge and the Lasso regularization model, add penalty term with the cost function. The only difference in the penalty is, in Ridge Regression the penalty is the square of the magnitude while in the Lasso Regression, the penalty is absolute value of the Magnitude.

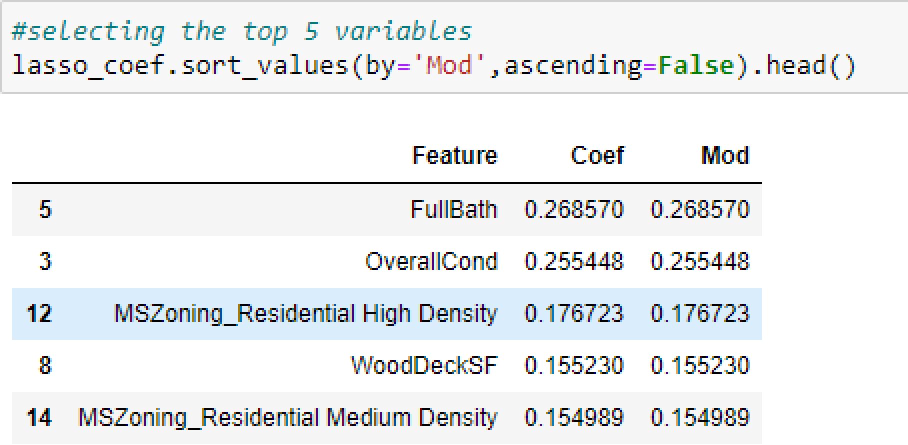
Ridge Regression:

RSS + α * (sum of square of coefficients)

Lasso Regression:

RSS + α * (sum of absolute value of coefficients)

The Ridge retains all the variables present in the data while lasso performs feature selection. It is important to choose the good optimal value for the regularization. We perform the hyper parameter tuning to get the good optimal value. After performing the 'GridSearchCV' with different parameters, we found out the best alpha value for the Ridge and Lasso Regularization as 0.1 and 0.00001. When applied, we found that Lasso is best model. Hence taking Lasso model for our further analysis. The top 5 features picked after performing Lasso regularization are:



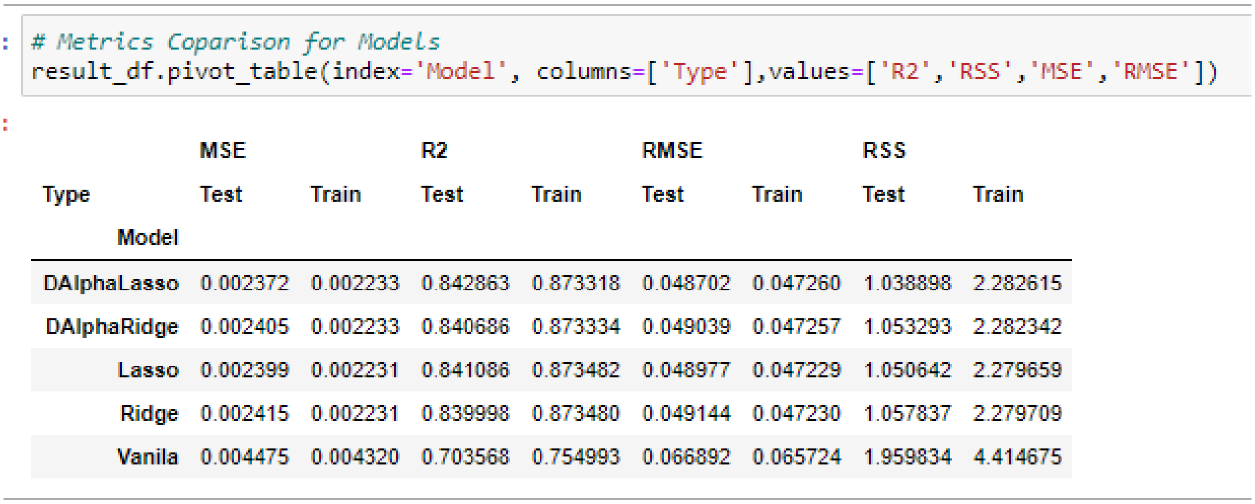
```
#selecting the top 5 variables
lasso_coef.sort_values(by='Mod',ascending=False).head()
```

	Feature	Coef	Mod
5	FullBath	0.268570	0.268570
3	OverallCond	0.255448	0.255448
12	MSZoning_Residential High Density	0.176723	0.176723
8	WoodDeckSF	0.155230	0.155230
14	MSZoning_Residential Medium Density	0.154989	0.154989

Double the alpha for Ridge and Lasso

The penalty in lasso, forces some of the estimates exactly to 0 if the lambda is large.

We can see the slight difference in both the R2 and the MSE value when the alpha is doubled in Ridge and Lasso methods.



```
# Metrics Coparison for Models
result_df.pivot_table(index='Model', columns=['Type'],values=['R2','RSS','MSE','RMSE'])
```

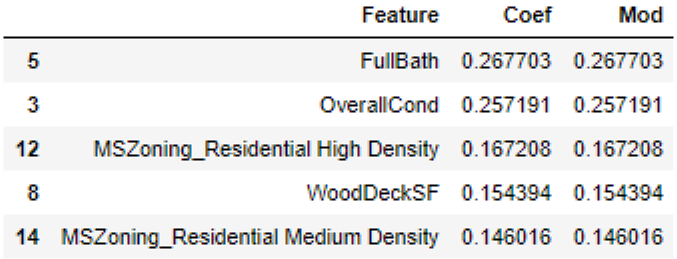
	MSE		R2		RMSE		RSS	
Type	Test	Train	Test	Train	Test	Train	Test	Train
Model								
DAlphaLasso	0.002372	0.002233	0.842863	0.873318	0.048702	0.047260	1.038898	2.282615
DAlphaRidge	0.002405	0.002233	0.840686	0.873334	0.049039	0.047257	1.053293	2.282342
Lasso	0.002399	0.002231	0.841086	0.873482	0.048977	0.047229	1.050642	2.279659
Ridge	0.002415	0.002231	0.839998	0.873480	0.049144	0.047230	1.057837	2.279709
Vanila	0.004475	0.004320	0.703568	0.754993	0.066892	0.065724	1.959834	4.414675

There is decrease in the beta coefficient value as well.

What will be the most important predictor variables after the change is implemented?

The most important predictor after the change is implemented in Lasso model are:

‘FullBath, ‘OverallCond, ‘MSZoning_Residential High Density’, ‘WoodDeckSF’, ‘MSZoning_Residential Medium Density’



	Feature	Coef	Mod
5	FullBath	0.267703	0.267703
3	OverallCond	0.257191	0.257191
12	MSZoning_Residential High Density	0.167208	0.167208
8	WoodDeckSF	0.154394	0.154394
14	MSZoning_Residential Medium Density	0.146016	0.146016

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?

The Optimal value of lambda are:

- For Ridge Regression : 0.1
- For Lasso Regression:0.00001

I am choosing to apply Lasso, since the Lasso model has higher R2 value than the ridge. The RMSE and RSS are slightly lower in the Lasso model.

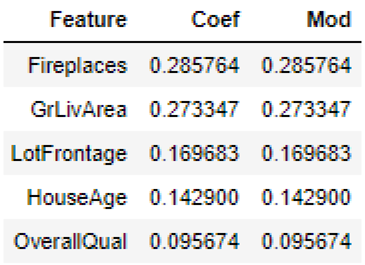
Question 3

After building the model, you realized 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?

The five important predictors of the Lasso model are mentioned above. Now removing these five important predictors and again the lasso model is built. The new five important predictors now are: (Please refer the workbook)

Fireplaces, GrLivArea, LotFrontage, HouseAge, OverallQual

The coefficient values picked for these predictors are given as: (Please refer the workbook)



Feature	Coef	Mod
Fireplaces	0.285764	0.285764
GrLivArea	0.273347	0.273347
LotFrontage	0.169683	0.169683
HouseAge	0.142900	0.142900
OverallQual	0.095674	0.095674

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 called Robust if the output variable is accurate even if the independent variable is changed. (i.e) when unforeseen value is tested. The model should work well on the unseen data without losing out on identifying the underlying patterns in the data. We can trade-off little bias to get better reduction in variance. The validation is about the Robustness of the model. Also we consider, R2, RMSE, MSE values to check the model performance. A good way to check the better validation process is using ‘Cross-Validation’. Cross-Validation takes multiple folds; it calculates different performance metrics and evaluate the variation between folds. This model helps in evaluating the performance of the model with different Train-Test combination of records.

The sensitivity allows us to explore the generalization of our model’s decision boundaries, to really see the impact of a lack of generalization.

We cannot have 100% accurate model, we need to consider the bias, variance and error factors.

We have to trade-off the accuracy of the model in the training data so that the model can work well with unseen data.

If we try to build more accurate model, then we might lead to overfitting.

The model has to identify the general pattern and has to work well on the unseen data.