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:

After carefully cleaning the housing data set, a linear model is applied to predict which variables are significant in predicting the price of a house. There were 67 features that contributed to the price of the model, which does not allow the model to perform without a bias and performed poorly on the test data. And the linear model is perfected by using ridge and lasso regression for regularization.

Here are the alpha values considered the best fit for the sale price prediction -

Alpha for ridge: 2.0

```
# Printing the best hyperparameter alpha
print(model_cv.best_params_)
```

```
{'alpha': 2.0}
```

Alpha for ridge: 0.0001

```
# Printing the best hyperparameter alpha
print(model_cv.best_params_)
```

```
{'alpha': 0.0001}
```

When the alpha values are doubled for both lasso and ridge operations, the co-efficient received is further penalized and may end up being too simple to accurately predict the sale price of the assets.

As predicted by lasso, the most significant predictor GrLivArea: Above grade (ground) living area square feet and has the highest correlation with the key indicator variable- SalePrice.

```
pd.set_option('display.max_rows', None)
df = pd.DataFrame( data = lasso.coef_, columns = ['Coefficient'] ,index = X_train.columns)
df.sort_values(by=["Coefficient"],ascending=False)
```

	Coefficient
GrLivArea	0.377698

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 to and why?

Answer:

After performing both ridge and lasso, the lasso is surely the preferred method to apply regularization.

Following is a comparative table to discuss the options-

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.535559e-01	-0.802262	0.920869
1	R2 Score (Test)	-1.374235e+23	-0.842005	0.824342
2	RSS (Train)	8.372672e-01	32.490121	1.426528
3	RSS (Test)	1.109051e+24	14.865561	1.417613
4	MSE (Train)	2.765190e-02	0.172254	0.036094
5	MSE (Test)	5.512256e+10	0.201811	0.062321

If observed closely, the R2 Score ie R squared value after lasso has surprisingly transformed, and the prediction is quite comparable.

Plus, if analyzed intuitively, the square feet of the living area should be a contributing factor for the sale price.

NB: In this model, to check the efficacy of regularization for linear regression, no other feature selection method is applied. This makes lasso stand out which allowed 68 features to make their contributions and allowing lasso to deduct and pick the best features.

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?

Answer:

	Coefficient
GrLivArea	0.377698
GarageArea	0.066842
OverallQual_Excellent	0.064074
YearBuilt	0.059178
BsmtFullBath	0.052252
FullBath	0.050643
MSZoning_RL	0.045846
RoofMatl_WdShngl	0.044398
Neighborhood_Crawfor	0.043781
MSZoning_RH	0.042068
TotRmsAbvGrd	0.040656

As per the coefficient table from the table after regularization with lasso, these are the ranked coefficients that contributed as per efficiency.

Hence if we remove the 1st, five features then the next 5 contributing factors would be-

1. **MSZoning_RL:** Identifies the general zoning classification of the sale, Residential Low Density
2. **RoofMatl**: Roof material with Wood Shingles
3. **Neighborhood_Crawfor:** Physical locations within Ames city limits, especially Crawford
4. **MSZoning_RH:** Residential high Density
5. **TotRmsAbvGrd:** Total rooms above grade (does not include bathrooms)

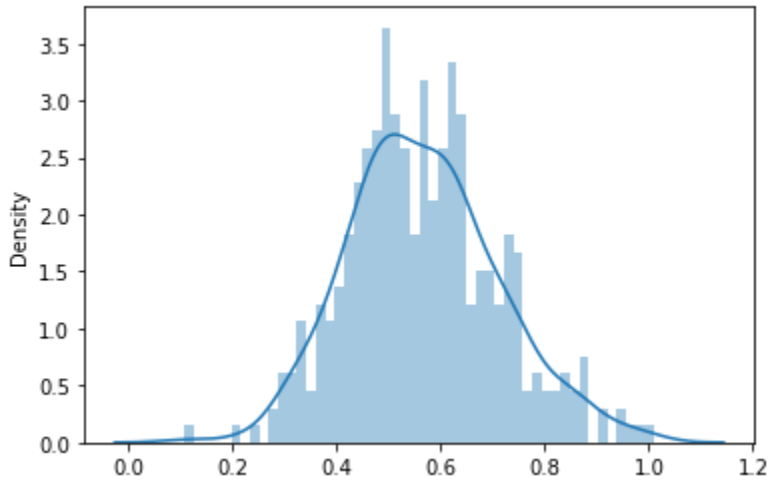
Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:

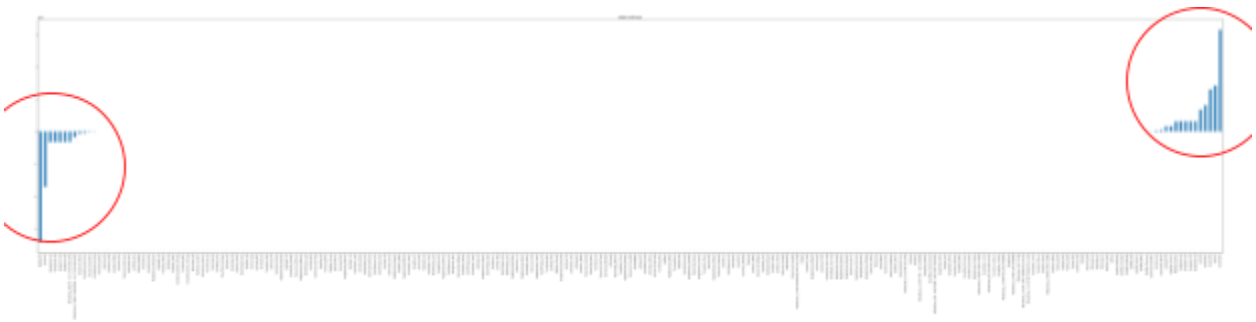
The first indication is the comparable R2 value when the model is performed on the test set. Plus the features selected by lasso as sale price predictor are very relevant in the practical sense and surely qualifies the sanity check made by the linear regression after lasso operation.

On the other hand, as per residual analysis, the error terms show a decent normalized curve.



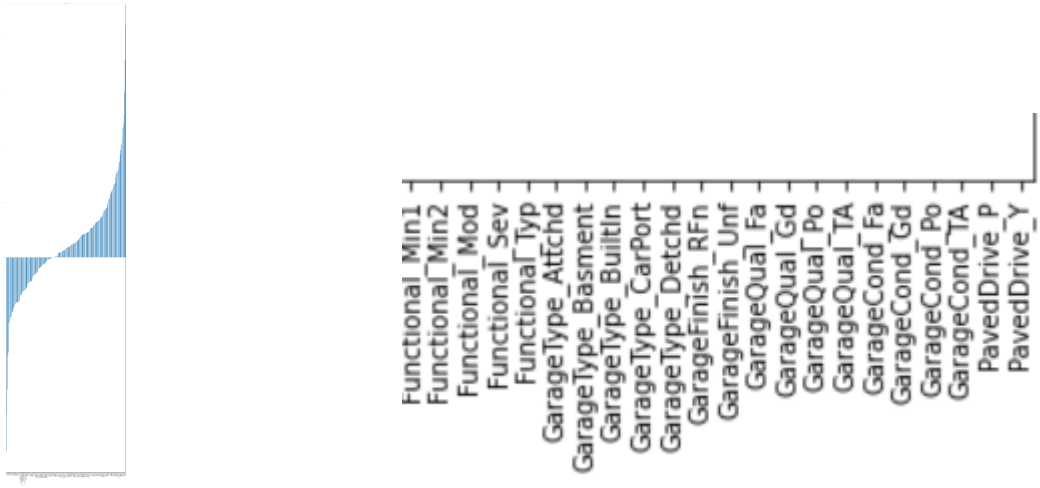
Magnitudes of Coefficients with regularizations are also checked to which speaks how the model has effectively performed the feature selection.

Simple Linear model coefficients against the features are plotted



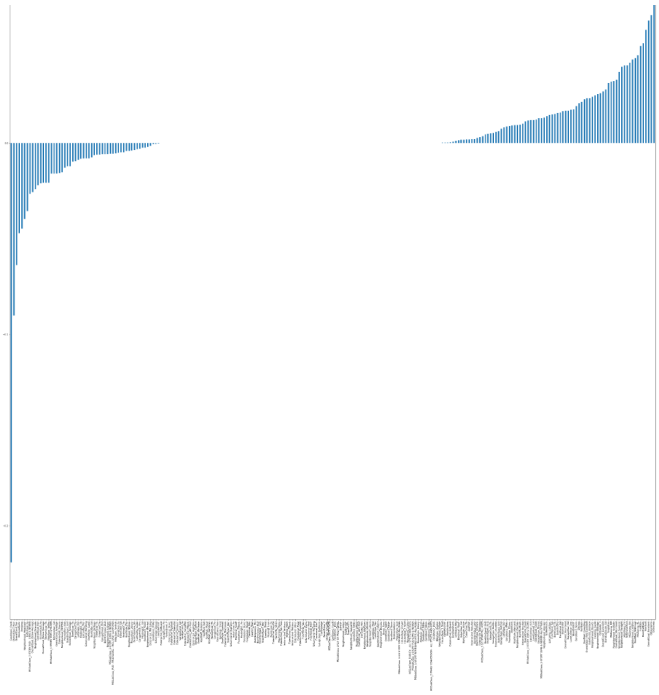
The red circles indicate the extreme variation of predicted Beta values

Ridge regularized Linear model coefficients against the features are plotted



Ridge helped the coefficients to some extent but failed to eliminate the redundant features due to its property that it cannot turn the coefficients to zero.

Lasso regularized Linear model coefficients against the features are plotted



Where lasso performed excellently in the regularization of the model.

- It fixed the beta extremes
- And, eliminating the redundant features

This approves that the model won't fail upon an unknown test set is asked to accurately predict the sale price.