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) <u>Best Alpha Values:</u>

a. Determined best alpha for Ridge Regression (Initial Model): 8

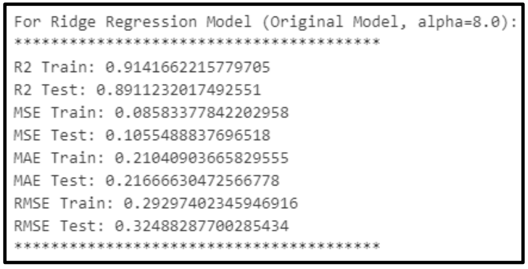
b. Determined best alpha for Lasso Regression (Initial Model): 0.001

(2) <u>Model Adjustments with Alpha Multiplied by Two:</u>

(Refer to the Jupyter notebook file for the code. The outcomes are detailed below)

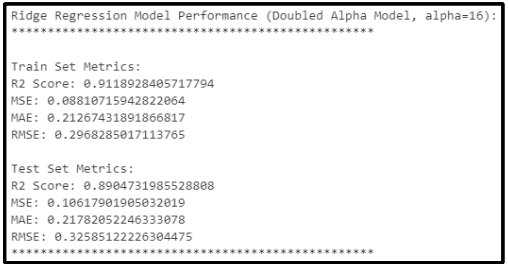
(i) Ridge Regression:

Initial Model (alpha=8)



```
For Ridge Regression Model (Original Model, alpha=8.0):
****************************************
R2 Train: 0.9141662215779705
R2 Test: 0.8911232017492551
MSE Train: 0.08583377842202958
MSE Test: 0.1055488837696518
MAE Train: 0.21040903665829555
MAE Test: 0.21666630472566778
RMSE Train: 0.29297402345946916
RMSE Test: 0.32488287700285434
****************************************
```

Model with Alpha Multiplied (alpha=16)



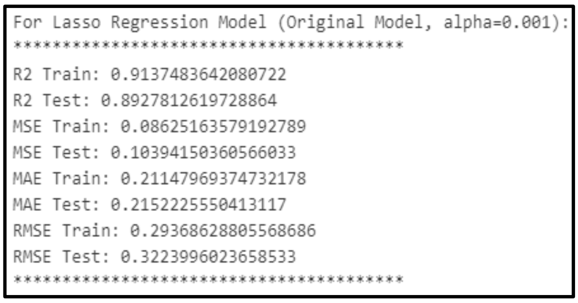
```
Ridge Regression Model Performance (Doubled Alpha Model, alpha=16):
**************************************************

Train Set Metrics:
R2 Score: 0.9118928405717794
MSE: 0.08810715942822064
MAE: 0.21267431891866817
RMSE: 0.2968285017113765

Test Set Metrics:
R2 Score: 0.8904731985528808
MSE: 0.10617901905032019
MAE: 0.21782052246333078
RMSE: 0.32585122226304475
**************************************************
```

<u>Observations:</u>

- The test accuracy for the ridge regression with an alpha of 8 is marginally better than that of the model with the alpha doubled to 16.
- When comparing MSE test scores for the same dataset between the original and the model with doubled alpha, the scores are somewhat lower for the original alpha model.
- The ridge regression with the original alpha value appears to outperform the model with the doubled alpha on both training and testing datasets.
- A higher alpha value results in a lower R2 score and a higher MSE, indicating more significant reduction of the coefficients. Therefore, the initial alpha model is the preferred option.

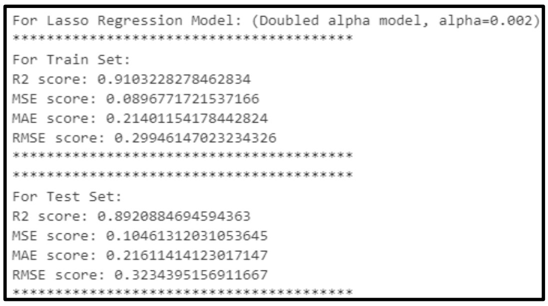
(ii) Lasso Regression:

Initial Model (alpha=0.001)



```
For Lasso Regression Model (Original Model, alpha=0.001):
****************************************
R2 Train: 0.9137483642080722
R2 Test: 0.8927812619728864
MSE Train: 0.08625163579192789
MSE Test: 0.10394150360566033
MAE Train: 0.21147969374732178
MAE Test: 0.2152225550413117
RMSE Train: 0.29368628805568686
RMSE Test: 0.3223996023658533
****************************************
```

Model with Alpha Multiplied (alpha=0.002)



```
For Lasso Regression Model: (Doubled alpha model, alpha=0.002)
****************************************
For Train Set:
R2 score: 0.9103228278462834
MSE score: 0.0896771721537166
MAE score: 0.21401154178442824
RMSE score: 0.29946147023234326
****************************************
****************************************
For Test Set:
R2 score: 0.8920884694594363
MSE score: 0.10461312031053645
MAE score: 0.21611414123017147
RMSE score: 0.3234395156911667
****************************************
```

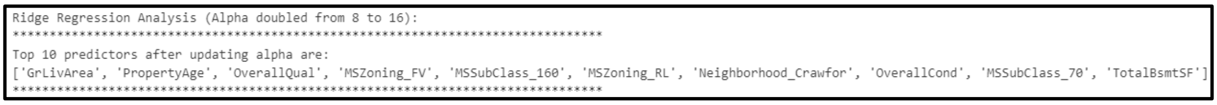
<u>Observations:</u>

- The test accuracy for the ridge regression with an alpha of 8 is marginally better than that of the model with the alpha doubled to 16.
- When comparing MSE test scores for the same dataset between the original and the model with doubled alpha, the scores are somewhat lower for the original alpha model.
- The ridge regression with the original alpha value appears to outperform the model with the doubled alpha on both training and testing datasets.

- A higher alpha value results in a lower R2 score and a higher MSE, indicating more significant reduction of the coefficients. Therefore, the initial alpha model is the preferred option.

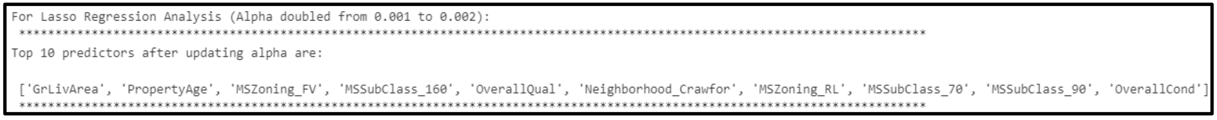
(3) <u>Key Predictor Variables Post-Implementation Change. The top 10 features are:</u>

a. Ridge Regression Model (alpha doubled to 16)



```
Ridge Regression Analysis (Alpha doubled from 8 to 16):
********************************************************************************
Top 10 predictors after updating alpha are:
['GrLivArea', 'PropertyAge', 'OverallQual', 'MSZoning_FV', 'MSSubClass_160', 'MSZoning_RL', 'Neighborhood_Crawfor', 'OverallCond', 'MSSubClass_70', 'TotalBsmtSF']
********************************************************************************
```

b. Lasso Regression Model (alpha doubled to 0.002)



```
For Lasso Regression Analysis (Alpha doubled from 0.001 to 0.002):
 *****************************************************************************************************************************
Top 10 predictors after updating alpha are:

 ['GrLivArea', 'PropertyAge', 'MSZoning_FV', 'MSSubClass_160', 'OverallQual', 'Neighborhood_Crawfor', 'MSZoning_RL', 'MSSubClass_70', 'MSSubClass_90', 'OverallCond']
 *****************************************************************************************************************************
```

****

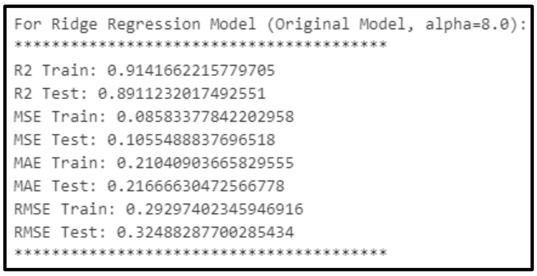
Question 2

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?

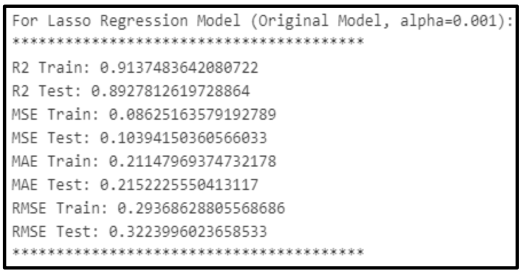
Answer:

Optimal Value of Alpha:

- For Ridge Regression (Baseline Model), the ideal alpha value is determined to be 8.0.
- For Lasso Regression (Baseline Model), the ideal alpha value is pinpointed at 0.001.



```
For Ridge Regression Model (Original Model, alpha=8.0):
****************************************
R2 Train: 0.9141662215779705
R2 Test: 0.8911232017492551
MSE Train: 0.08583377842202958
MSE Test: 0.1055488837696518
MAE Train: 0.21040903665829555
MAE Test: 0.21666630472566778
RMSE Train: 0.29297402345946916
RMSE Test: 0.32488287700285434
****************************************
```



```
For Lasso Regression Model (Original Model, alpha=0.001):
****************************************
R2 Train: 0.9137483642080722
R2 Test: 0.8927812619728864
MSE Train: 0.08625163579192789
MSE Test: 0.10394150360566033
MAE Train: 0.21147969374732178
MAE Test: 0.2152225550413117
RMSE Train: 0.29368628805568686
RMSE Test: 0.3223996023658533
****************************************
```

- The R2 test score for the Lasso Regression Model edges out the Ridge Regression Model, suggesting a slightly superior performance on unseen data. This is coupled with a modest dip in training accuracy, indicating a well-tuned model.
- The Mean Squared Error (MSE) on the test set for Lasso Regression is marginally lower compared to Ridge Regression, hinting at a more effective performance on test data. The ability of Lasso Regression to perform feature selection—zeroing out less significant predictor variables—gives it an advantage over Ridge Regression. Thus, the predictors identified by Lasso are more impactful for selecting key variables in house price prediction within this study.
- In practical scenarios, when selecting a regression model, analysts must navigate the hidden pitfalls of outliers, error non-normality, and overfitting, particularly in sparse datasets. Employing the L2 norm (Ridge) can leave the model vulnerable to these issues. Conversely, the L1 norm (Lasso) offers a robust defense, significantly mitigating these risks and leading to more resilient regression models.

****

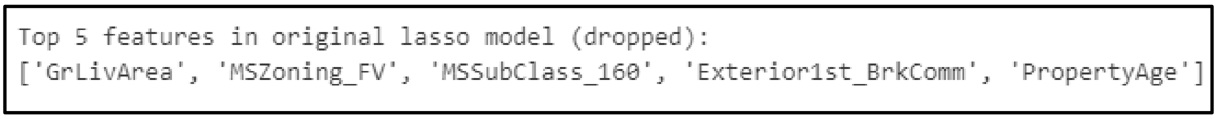
Question 3

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:

(Refer to the Jupyter notebook file for the code. The findings are as follows)

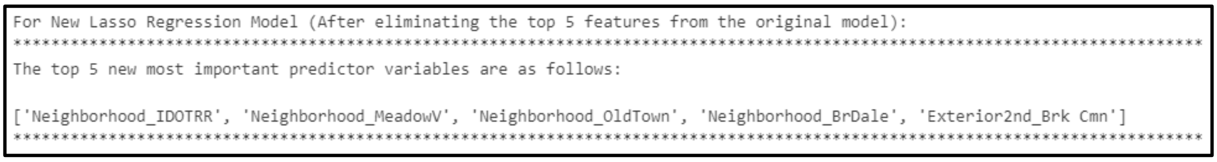
The initial Lasso Model highlighted the following top five features:



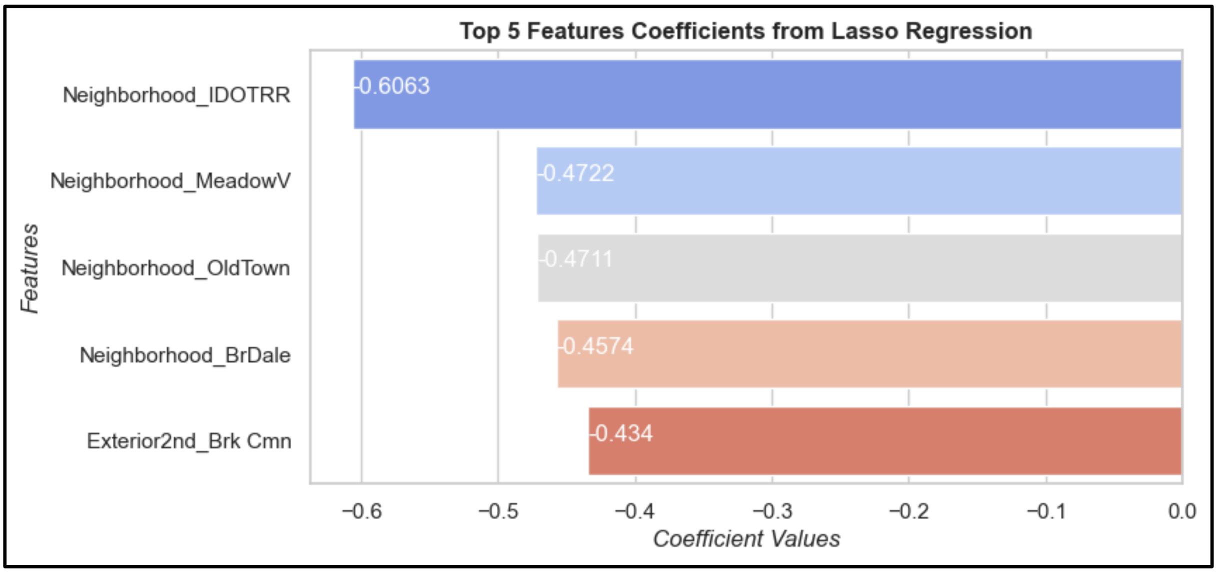
```
Top 5 features in original lasso model (dropped):
['GrLivArea', 'MSZoning_FV', 'MSSubClass_160', 'Exterior1st_BrkComm', 'PropertyAge']
```

The revised model's top five predictor variables

(This is after excluding the top five predictors identified in the original Lasso model):



```
For New Lasso Regression Model (After eliminating the top 5 features from the original model):
********************************************************************************************************************
The top 5 new most important predictor variables are as follows:

['Neighborhood_IDOTRR', 'Neighborhood_MeadowV', 'Neighborhood_OldTown', 'Neighborhood_BrDale', 'Exterior2nd_Brk Cmn']
********************************************************************************************************************
```



****

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:

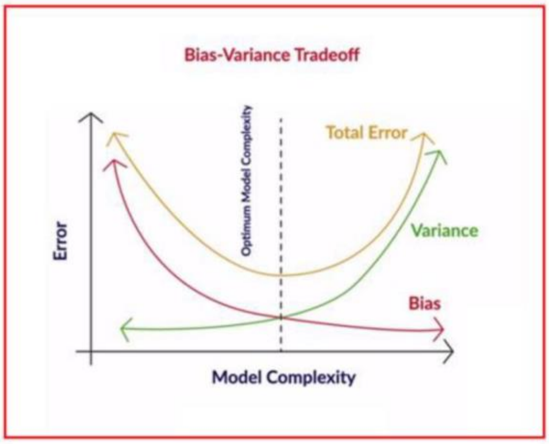
The robustness of a model refers to its ability to maintain consistent error rates between training and testing, as well as its stability when the dataset is slightly altered. Essentially, a model's robustness indicates how well it can be applied to new datasets beyond those it was trained and tested on.

Implementing regularization techniques allows us to manage the balance between the complexity of the model and its bias, which is intrinsically linked to the model's robustness. Regularization works by penalizing overly complex models, ensuring that only the necessary complexity is retained. This process simplifies the model optimally, enhancing its robustness and generalizability. Achieving this balance is crucial; a model should be simple enough to be robust but not so simple that it becomes ineffective.

The concept of the Bias-Variance Trade-off emerges when simplifying a model:

- A complex model is highly sensitive to minor changes in the dataset, leading to instability.
- A simpler model, which captures the underlying pattern of the data, remains stable even with the addition or removal of data points.

Bias measures the expected accuracy of the model on test data. While a complex model can predict accurately with sufficient training data, an overly simplistic model—one that predicts the same outcome for all test inputs—will have a high bias due to its consistently high error rates across all test inputs. Variance measures how much the model changes in response to variations in the training data.



Maintaining a balance between bias and variance is key to preserving a model's accuracy, as it minimizes overall error. However, there is a tension between accuracy and robustness; an overly accurate model may be prone to overfitting, performing well on training data but poorly on new data, and vice versa.

****