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:

(Refer to the provided Jupyter notebook file for code details. Key findings are summarized below.)

(1) Determined Optimal Alpha Values:

- Optimal alpha for Ridge Regression: 8

- Optimal alpha for Lasso Regression: 0.001

(2) Impact of Doubling Alpha Value on Ridge and Lasso Regression Models:

(i) Ridge Regression Analysis:

```
Ridge Regression Model Performance (Alpha=8.0):
 ------------------------------------------------

Training Set Metrics:
R2 Score: 0.9135828257331216
MSE: 0.08641717426687842
MAE: 0.2064410669637679
RMSE: 0.2939679817035835

Testing Set Metrics:
R2 Score: 0.8977081065509518
MSE: 0.11070137654841027
MAE: 0.23087628551197853
RMSE: 0.332718163839022
 ------------------------------------------------
```

```
Ridge Regression Model Performance (Doubled Alpha=8*2=16):
 ---------------------------------------

Training Set Metrics:
R2 Score: 0.9111010762652046
MSE: 0.08889892373479537
MAE: 0.20892933701756003
RMSE: 0.29815922547322826

Testing Set Metrics:
R2 Score: 0.8975371403846054
MSE: 0.1108863979545022
MAE: 0.23120844486768852
RMSE: 0.33299609300185823
 ---------------------------------------
```

- Comparison between the original model (alpha=8) and the model with doubled alpha (alpha=16) reveals:

  - Slightly better test accuracy for the original model compared to the doubled alpha model.

  - Lower Mean Squared Error (MSE) for the test set in the original model than in the doubled alpha model, indicating the original model's superior performance on both training and testing datasets.

  - An increase in alpha leads to reduced R2 scores and higher MSE, suggesting the original model's preferable balance of coefficient shrinkage.

(ii) Lasso Regression Analysis:

```
Lasso Regression Model Performance (With alpha=0.001):
 ****************************************

Training Set Metrics:
R2 Score: 0.9135059399885377
MSE: 0.08649406001146226
MAE: 0.20691162463583912
RMSE: 0.29409872494021844

Testing Set Metrics:
R2 Score: 0.8984115923743538
MSE: 0.10994005669786232
MAE: 0.22847657954370804
RMSE: 0.3315720987928
 ****************************************
```

```
Lasso Regression Model Performance (With Doubled alpha:0.001*2 = 0.002):
 ****************************************

Training Set Metrics:
R2 Score: 0.909273732835048
MSE: 0.09072626716495198
MAE: 0.21075465161767734
RMSE: 0.3012080131154415

Testing Set Metrics:
R2 Score: 0.8987595121571955
MSE: 0.109563534203358
MAE: 0.22810198707416482
RMSE: 0.331003828079613
 ****************************************
```

- Evaluating the original model (alpha=0.001) against the doubled alpha model (alpha=0.002) shows:

  - Marginally higher test accuracy in the original model compared to the doubled alpha model.

  - Lower MSE for the test set in the original model, signifying its better performance on both datasets.

  - Doubling alpha results in decreased R2 scores and increased MSE, highlighting the effectiveness of the original model in selectively shrinking coefficients and performing feature selection.

(3) Top Predictor Variables Post-Adjustment:

- For the Ridge Regression model with doubled alpha (alpha=16), the following are the top 10 predictor variables:

```
Ridge Regression Analysis (With Alpha Doubled to 16):
 ----------------------------------------------------------------------------------------------------------------------------
Top 10 Most Significant Predictor Variables Post Adjustment:

 ['GrLivArea', 'PropertyAge', 'OverallQual', 'TotalBsmtSF', 'MSSubClass_90', 'MSZoning_FV', 'Neighborhood_Crawfor', 'Neighborhood_StoneBr', 'OverallCond', 'MSSubC
 ----------------------------------------------------------------------------------------------------------------------------
```

- For the Lasso Regression model with doubled alpha (alpha=0.002), the following are the top 10 predictor variables:

```
Lasso Regression Analysis (With Alpha Increased to 0.002):
 ----------------------------------------------------------------------------------------------------------------------------
Top 10 Key Predictor Variables Identified Post-Adjustment:

 ['GrLivArea', 'PropertyAge', 'OverallQual', 'MSSubClass_90', 'Neighborhood_Crawfor', 'MSZoning_FV', 'Neighborhood_StoneBr', 'TotalBsmtSF', 'MSSubClass_160', 'MSS
 ----------------------------------------------------------------------------------------------------------------------------
```

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 Alpha Values Determined:

For Ridge Regression, the optimal alpha value identified is 8.0.

For Lasso Regression, the optimal alpha value found is 0.001.

```
Ridge Regression Model Performance (Alpha=8.0):
 ------------------------------------------------

Training Set Metrics:
R2 Score: 0.9135828257331216
MSE: 0.08641717426687842
MAE: 0.2064410669637679
RMSE: 0.2939679817035835

Testing Set Metrics:
R2 Score: 0.8977081065509518
MSE: 0.11070137654841027
MAE: 0.23087628551197853
RMSE: 0.332718163839022
 ------------------------------------------------
```

```
Lasso Regression Model Performance (With alpha=0.001):
 ****************************************

Training Set Metrics:
R2 Score: 0.9135059399885377
MSE: 0.08649406001146226
MAE: 0.20691162463583912
RMSE: 0.29409872494021844

Testing Set Metrics:
R2 Score: 0.8984115923743538
MSE: 0.10994005669786232
MAE: 0.22847657954370804
RMSE: 0.3315720987928
 ****************************************
```

Comparative Analysis:

The Lasso Regression Model demonstrates a marginally superior R2 score for the test set compared to the Ridge Regression Model, indicating slightly better performance on unseen data. Additionally, the Lasso Regression Model shows a modest decrease in training accuracy, suggesting it might be a preferable choice due to its enhanced generalizability.

Performance Metrics:

The Mean Squared Error (MSE) for the test set is somewhat lower in the Lasso Regression Model than in the Ridge Regression Model. This suggests that Lasso Regression has a slight advantage in handling unseen test data. The ability of Lasso Regression to perform feature selection—reducing the coefficients of less significant predictors to zero—further underscores its superiority in this context. By identifying and utilizing only the most relevant variables, Lasso Regression provides a more focused approach for predicting house prices in this analysis.

Considerations for Regression Choice:

When selecting a regression model for practical applications, analysts must navigate challenges such as outliers, non-normally distributed errors, and the risk of overfitting, particularly in sparse datasets. The L2 norm utilized in Ridge Regression may leave models more susceptible to these issues. In contrast, the L1 norm employed by Lasso Regression offers greater resilience, mitigating such risks more effectively and leading to more robust and reliable models.

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:

(Refer to the provided Jupyter notebook file for code details. Key findings are summarized below.)

Top Five Features in the Original Lasso Model (Prior to Removal):

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

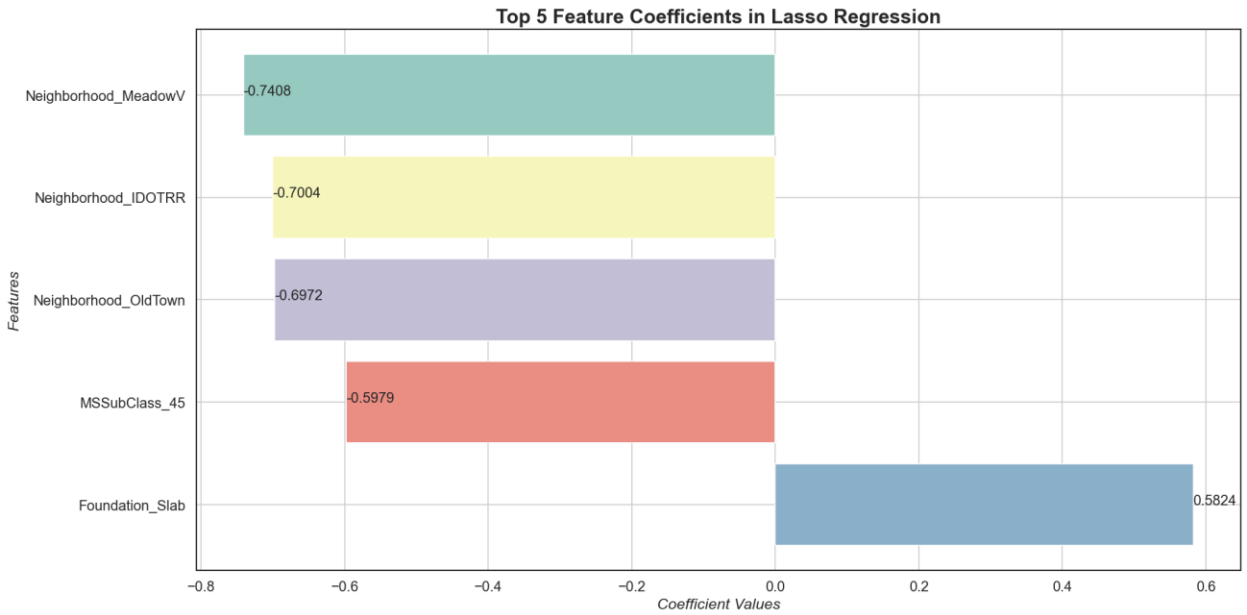
	LotFrontage	LotArea	OverallQual	OverallCond	MasVnrArea	BsmtFinSF1	BsmtUnfSF	TotalBsmtSF	BsmtFullBath	FullBath	HalfBath	BedroomAbvGr	Fireplaces	Garag
0	65.0	8450	7	5	196.0	706	150	856	1	2	1	3	0	
1	80.0	9600	6	8	0.0	978	284	1262	0	2	0	3	1	
2	68.0	11250	7	5	162.0	486	434	920	1	2	1	3	1	
3	60.0	9550	7	5	0.0	216	540	756	1	1	0	3	1	
4	84.0	14260	8	5	350.0	655	490	1145	1	2	1	4	1	

Top Five Predictor Variables in the Revised Model:

(After excluding the previously mentioned top five predictors from the original Lasso model)

```
For New Lasso Regression Model (After eliminating the top5 features from the original model):
 **********************************************************************************************************************************
The top5 new most important predictor variables are as follows:

 ['Neighborhood_MeadowV', 'Neighborhood_IDOTRR', 'Neighborhood_OldTown', 'MSSubClass_45', 'Foundation_Slab']
 **********************************************************************************************************************************
```



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:

Ensuring a model is both robust and generalizable involves creating a model that performs well not just on the training data but also on unseen data, maintaining consistent accuracy across different datasets. Robustness implies the model's ability to handle minor data variations, noise, and outliers without significant performance degradation. Generalizability refers to the model's capability to apply learned patterns to new, unseen data effectively.

One key strategy for enhancing robustness and generalizability is through cross-validation. This technique involves dividing the dataset into multiple subsets, using some for training and the remainder for testing, and rotating these roles among all subsets. This process helps identify models that perform well across various data samples, reducing the risk of overfitting to the training data.

Regularization (L1/Lasso and L2/Ridge) is another crucial method. It introduces a penalty term for complexity, discouraging overly complex models that fit the training data too closely. Lasso regression can also help in feature selection by shrinking less important feature coefficients to zero, simplifying the model. Simplifying the model helps balance the bias-variance trade-off, where a highly complex model (low bias) might have high variance, performing poorly on unseen data, and a too-simple model (high bias) might not capture underlying patterns well.

Furthermore, understanding and preprocessing data can enhance model robustness and generalizability. Techniques such as feature scaling, handling missing values appropriately, and removing irrelevant features can make the model more resilient to variations in data.

The implications of ensuring a model's robustness and generalizability directly impact its accuracy. A robust and generalizable model is likely to maintain high accuracy across different datasets, reflecting its ability to capture essential patterns without being misled by noise or specificities of the training data. However, there's a trade-off to consider: striving for robustness and generalizability might slightly reduce performance on the training set due to the model's simplicity. Yet, this compromise often results in better long-term accuracy and reliability when the model encounters real-world data, which is the ultimate goal of predictive modeling.