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?

Solution:

(1) Optimal Value of Alpha:

- The computed optimal value of alpha for Ridge Regression (Original Model): *8*
- The computed optimal value of alpha for Lasso Regression (Original Model): *0.001*

(2) Changes in the model, if you choose double the value of alpha for both ridge and lasso regression:

(Please refer the jupyter (.ipynb) file for the code. Results are mentioned below)

(i) Ridge Regression:

Original Model (alpha=8), Doubled Alpha Model(alpha=16)

```
For Ridge Regression Model (Original Model, alpha=8.0):
 ****************************************

For Train Set:
R2 score: 0.9141662215779705
MSE score: 0.08583377842202958
MAE score: 0.21040903665829555
RMSE score: 0.29297402345946916

For Test Set:
R2 score: 0.8911232017492551
MSE score: 0.1055488837696518
MAE score: 0.21666630472566778
RMSE score: 0.32488287700285434
 ****************************************
```

```
For Ridge Regression Model (Doubled alpha model, alpha=8*2=16):
 ****************************************

For Train Set:
R2 score: 0.9118928405717794
MSE score: 0.08810715942822064
MAE score: 0.21267431891866817
RMSE score: 0.2968285017113765

For Test Set:
R2 score: 0.8904731985528808
MSE score: 0.10617901905032019
MAE score: 0.21782052246333078
RMSE score: 0.32585122226304475
 ****************************************
```

Observations:

- The test accuracy of the ridge regression model (alpha=8) is slightly higher in comparison to the test accuracy of the doubled alpha model (doubled alpha=16).
- MSE test scores comparing similar data of the original dataset and doubled alpha model gives us an idea that it is slightly smaller for the single alpha model than the doubled alpha model.
- Ridge Regression model (single alpha model) seems to perform better on the train and test data in comparison to the doubled alpha Ridge Regression model.
- Increase in the value of alpha in the model lead to a decrease in R2 score but an increase in the MSE (causing more shrinkage of coefficient values). Thus, making the original (single) alpha model a better choice.

(ii) Lasso Regression:

Original Model (alpha=0.001), Doubled Alpha Model(alpha=0.002)

```
For Lasso Regression Model (Original Model: alpha=0.001):
 ****************************************

For Train Set:
R2 score: 0.9137483642080722
MSE score: 0.08625163579192789
MAE score: 0.21147969374732178
RMSE score: 0.29368628805568686

For Test Set:
R2 score: 0.8927812619728864
MSE score: 0.10394150360566033
MAE score: 0.2152225550413117
RMSE score: 0.3223996023658533
 ****************************************
```

```
For Lasso Regression Model: (Doubled alpha model: alpha:0.001*2 = 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
 ****************************************
```

Observations:

- The test accuracy of the lasso regression model (alpha=0.001) is slightly higher in comparison to the test accuracy of the doubled alpha model (doubled alpha=0.002).
- MSE test scores comparing similar data of the original dataset and doubled alpha model gives us an idea that it is slightly smaller for the single alpha model than the doubled alpha model.
- Lasso Regression model (single alpha model) seems to perform better on the train and test data in comparison to the doubled alpha Lasso Regression model.
- Increase in the value of alpha in the model lead to a decrease in R2 score but an increase in the MSE (causing more shrinkage of coefficient values). In Lasso, the insignificant coefficients that have their values near to 0 correspond to 0 values; performing feature selection in the model. Thus, making the original (single) alpha model a better choice.

(3) The most important predictor variables after the change is implemented. Top 10 features are as follows:

(i) Ridge Regression Model (doubled alpha=16)

```
For Ridge Regression (Doubled alpha model, alpha=8*2=16):
****************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['GrLivArea', 'AgeofProperty', 'OverallQual', 'MSZoning_FV', 'MSSubClass_160', 'MSZoning_RL', 'Neighborhood_Crawfor', 'Overall
Cond', 'MSSubClass_70', 'TotalBsmtSF']
****************************************************************************************************************
```

(ii) Lasso Regression Model (doubled alpha=0.002)

```
For Lasso Regression (Doubled alpha model: alpha:0.001*2 = 0.002):
****************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['GrLivArea', 'AgeofProperty', 'MSZoning_FV', 'MSSubClass_160', 'OverallQual', 'Neighborhood_Crawfor', 'MSZoning_RL', 'MSSubCl
ass_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?

Solution:

Optimal Value of Alpha:

- The computed optimal value of alpha for Ridge Regression (Original Model): *8.0*
- The computed optimal value of alpha for Lasso Regression (Original Model): *0.001*

```
For Ridge Regression Model (Original Model, alpha=8.0):
 ****************************************

For Train Set:
R2 score: 0.9141662215779705
MSE score: 0.08583377842202958
MAE score: 0.21040903665829555
RMSE score: 0.29297402345946916

For Test Set:
R2 score: 0.8911232017492551
MSE score: 0.1055488837696518
MAE score: 0.21666630472566778
RMSE score: 0.32488287700285434
 ****************************************
```

```
For Lasso Regression Model (Original Model: alpha=0.001):
 ****************************************

For Train Set:
R2 score: 0.9137483642080722
MSE score: 0.08625163579192789
MAE score: 0.21147969374732178
RMSE score: 0.29368628805568686

For Test Set:
R2 score: 0.8927812619728864
MSE score: 0.10394150360566033
MAE score: 0.2152225550413117
RMSE score: 0.3223996023658533
 ****************************************
```

- The R2 test score on the Lasso Regression Model is slightly better than that of Ridge Regression Model. Moreover, the training accuracy is slightly reduced; hence, making the model an optimal choice as it seems to perform better on the unseen data.
- The MSE for Test set (Lasso Regression) is slightly lower than that of the Ridge Regression Model; implies Lasso Regression performs better on the unseen test data. Also, since Lasso helps in feature selection (the coefficient values of some of the insignificant predictor variables became 0), implies Lasso Regression has a better edge over Ridge Regression. Therefore, the variables predicted by Lasso can be applied in order to choose significant variables for predicting the price of a house in this analysis.

Moreover, while choosing a type of regression in the real world, an analyst has to deal with the lurking and confounding dangers of outliers, non-normality of errors and overfitting especially in sparse datasets among others. Using L2 norm (Ridge) results in exposing the analyst to such risks. Hence, use of L1 norm (Lasso) could be quite beneficial as it is quite robust to fend off such risks to a large extent, thereby resulting in better and robust 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?

<u>Solution:</u>

<u>(Please refer the jupyter (.ipynb) file for the code. Results are mentioned below)</u>

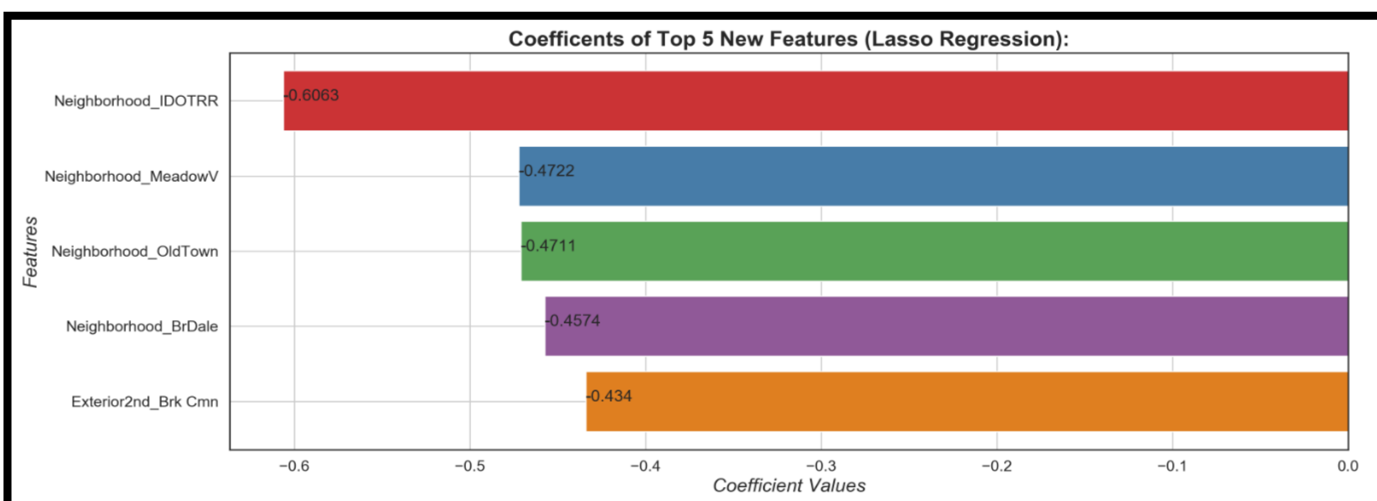
Top five features in original Lasso Model (before removing) were as follows:

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

<u>Top five predictor variables in the new model:</u>
(After removing the aforementioned top 5 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_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?

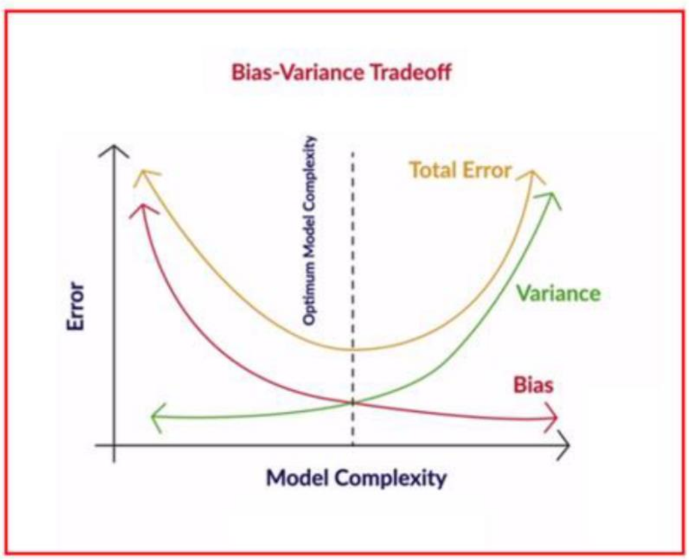
Solution:

Robustness of a model implies, either the testing error of the model is consistent with the training error, the model performs well with enough stability even after adding some noise to the dataset. Thus, the robustness (or generalizability) of a model is a measure of its successful application to data sets other than the one used for training and testing.

By the implementing regularization techniques, we can control the trade-off between model complexity and bias which is directly connected the robustness of the model. Regularization, helps in penalizing the coefficients for making the model too complex; thereby allowing only the optimal amount of complexity to the model. It helps in controlling the robustness of the model by making the model optimal simpler. Therefore, in order to make the model more robust and generalizable, one need to make sure that there is a delicate balance between keeping the model simple and not making it too naive to be of any use. Also, making a model simple leads to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias helps you quantify, how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there has to be enough training data. Models that are too naïve, for e.g., one that gives same results for all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high. Variance is the degree of changes in the model itself with respect to changes in the training data.
Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.



Thus, accuracy and robustness may be at the odds to each other as too much accurate model can be prey to over fitting hence it can be too much accurate on train data but fails when it faces the actual data or vice versa.

****