Solutions to the Assignment 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?

Solution:

The optimal value of alpha for ridge and lasso regression are 0.8 and 0.00001 respectively.

```
alpha for ridge regression     0.8

r2_score for training data  0.840304336177706
r2_score for testing data  0.8237512180407998


doubling the alpha for ridge regression 1.6

r2_score for training data  0.838385789960715
r2_score for testing data  0.8224503253735966

Most important predictor variable after the change:
TotalBsmtSF with the coefficient 0.11229135

(TotalBsmtSF: Total square feet of basement area)
```

```
alpha  for lasso regression      0.0001

r2_score (Training data)   0.8357552577963208
r2_score (Testing data)    0.8246434662442312

doubling the alpha for lasso regression 0.0002

r2_score for training data  0.8291908383641476
r2_score for testing data  0.8160356691011377


Most important predictor variable after the change:
TotalBsmtSF with the coefficient 0.12053198

(TotalBsmtSF: Total square feet of basement area)
```

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:

Following is the comparison table for both the Ridge and the Lasso models:

	Metric	Ridge_Regression	Lasso_Regression
0	r2 Score (Train)	0.838386	0.829191
1	r2 Score (Test)	0.822450	0.816036
2	RSS (Train)	1.988667	2.101812
3	RSS (Test)	0.965128	0.999997
4	MSE (Train)	0.044133	0.045372
5	MSE (Test)	0.046941	0.047782

As per the above table, Ridge Regression is better. It will be preferred for the application in the above case for the following reasons (as per the analysis):

r2_score is a bit better for the Ridge Regression model (0.838386 and 0.822450 respectively) as compared to that of Lasso model (0.822450 and 0.816036 respectively).

Even RSS and MSE values are lower for Ridge Regression Model as shown in the above table which are preferred.

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?

Solution:

The five most important predictors (according to the Ridge Regression)

```
SaleCondition_Partial     0.112291
SaleCondition_AdjLand     0.092485
GarageFinish_Unf          0.086860
GarageFinish_RFn          0.069409
Foundation_Wood           0.063716
```

Now, after removing the above predictors from the given data and rebuilding the model, a new model is generated with the following top five predictors:

```
Neighborhood_NridgHt    0.032039
Neighborhood_NoRidge    0.031134
BsmtFullBath            0.018468
MasVnrArea              0.017436
GarageCars              0.016616
```

Note: The entire process is demonstrated in the ipynb file.

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?

Solution:

A model is considered robust when the performance of the model does not change significantly from the training to testing stage. The accuracy that is obtained for a trained model should be nearly maintained even with the testing data. In many cases, a model works extremely well with the training data i.e. overfits itself. It considers each and every data point of the dataset and adapts itself to it (e.g. in Polynomial approach).

In many such cases, it doesn't perform well with the test data.

Generalisation means how well is a trained model able to classify or work with unseen data. Training a generalized machine learning model means, in general, it works for all subset of unseen data.

Variance and bias are two important terms in relation to machine learning. Variance refers to the variety of predictions values made by a machine learning model . Bias gives the difference of the predictions from the actual values. A high-biased model is a model in which prediction values are quite far from the actual values. A low-biased, high-variance model is called overfit model and a high-biased, low-variance model is called underfit model.

By generalisation, the best trade-off between underfitting and overfitting is obtained so that the trained model gives the best performance in all the circumstances. An overfit model obtains a high prediction score on the training data and low one from the testing (unseen) data. An underfit model has low performance in both training (seen) and testing (unseen) datasets.