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?

'Ridge Regression' and 'Lasso Regression' are fundamental regularization techniques used.

In our model, the Optimal value of alpha for ridge regression=10 and for Lasso=0.001

When I doubled the value for alpha for Lasso, I saw that the accuracy of train and test scores reduced slightly for lasso and in return the error terms have increased a bit. For Ridge, train score decreased while test score increase slightly. The Error term remained similar.

Original Model -

Lasso :

```
The R2 Score of the model on the TRAIN dataset : 0.9206709153542387
The R2 Score of the model on the TEST dataset : 0.9067784241823188
The RMSE of the model on the test dataset : 0.11546718823765788
The MSE of the model on the test dataset : 0.013332671559510716
The MAE of the model on the test dataset : 0.08327377518449156
```

Ridge :

```
The R2 Score of the model on the TRAIN dataset : 0.9364594823911133
The R2 Score of the model on the TEST dataset : 0.9077597079466582
The RMSE of the model on the test dataset : 0.1148578559506099
The MSE of the model on the test dataset : 0.013192327073571054
The MAE of the model on the test dataset : 0.08168110546842879
```

New Model with double alpha:

Lasso :

```
The R2 Score of the model on the TRAIN dataset : 0.9122051184016314
The R2 Score of the model on the TEST dataset : 0.9031718080740088
The RMSE of the model on the test dataset : 0.11767962655897
The MSE of the model on the test dataset : 0.01384849450705864
The MAE of the model on the test dataset : 0.08569888516670239
```

Ridge :

```
The R2 Score of the model on the TRAIN dataset : 0.9272757940609146
The R2 Score of the model on the TEST dataset : 0.9100087351460884
The RMSE of the model on the test dataset : 0.11344896766981026
The MSE of the model on the test dataset : 0.012870668265345653
The MAE of the model on the test dataset : 0.08166994099572968
```

The predictor variables after double the value of alpha are -

Lasso:

	Variable	Coeff
0	constant	11.887
13	GrLivArea	0.127
4	OverallQual	0.092
210	SaleCondition_Partial	0.085
50	Neighborhood_Crawfor	0.068
5	OverallCond	0.057

Ridge:

	Variable	Coeff
0	constant	11.845
4	OverallQual	0.075
50	Neighborhood_Crawfor	0.073
13	GrLivArea	0.072
5	OverallCond	0.055
209	SaleCondition_Normal	0.053

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?

The test accuracy is near about similar for Lasso & Ridge, though train accuracy was slightly lesser compared with Ridge.

Lasso :

```
The R2 Score of the model on the TRAIN dataset : 0.9206709153542387
The R2 Score of the model on the TEST dataset : 0.9067784241823188
The RMSE of the model on the test dataset : 0.11546718823765788
The MSE of the model on the test dataset : 0.013332671559510716
The MAE of the model on the test dataset : 0.08327377518449156
```

Ridge :

```
The R2 Score of the model on the TRAIN dataset : 0.9364594823911133
The R2 Score of the model on the TEST dataset : 0.9077597079466582
The RMSE of the model on the test dataset : 0.1148578559506099
The MSE of the model on the test dataset : 0.013192327073571054
The MAE of the model on the test dataset : 0.08168110546842879
```

With the similar accuracy, I would choose Lasso since it brings and assigns a zero value to insignificant features. It is always advisable to choose simpler model.

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?

Top 5 features of Rebuilt Lasso model after dropping the Top 5 Variables of the original model :

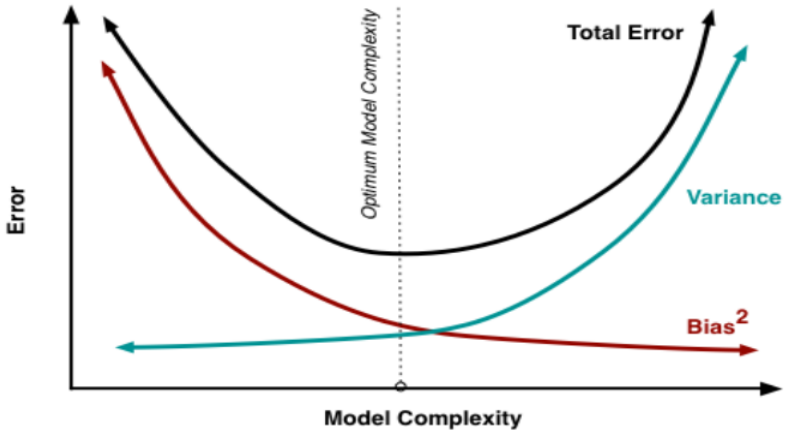
1. ScreenPorch
2. PoolArea
3. MSZoning_FV
4. 1stFlrSF
5. MiscFeature_Shed

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?

Model should be as simple as necessary. Hence a simple model has an few advantages over other complex model. Generalizability and robustness arising due to a simple model is the biggest advantages of simple model.

Robust model is not sensitive to training data. Robust models have low variance and high bias
• Variance = How sensitive is model to the training data. This refers to consistency of the model.
• Bias = Accuracy of the data on unseen future data.



Making a model simple lead 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.

Hence a simple model may have a lower accuracy but since it would be more generalized we can expect it to perform consistently when productionised.