Subjective Quiestions

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?

1.)

Ridge Regression
Optimal Value of alpha = 5
Important Predictor = OverallQual_Excellent - 0.701

	Features	CoefValues
87	OverallQual_Excellent	0.701
68	Neighborhood_NoRidge	0.445
75	Neighborhood_StoneBr	0.424
90	OverallQual_Very Good	0.305
69	Neighborhood_NridgHt	0.280
...	...	...
152	KitchenQual_Gd	-0.210
153	KitchenQual_TA	-0.234
60	Neighborhood_Edwards	-0.257
128	BsmtQual_TA	-0.279
126	BsmtQual_Gd	-0.291

```
ridge.score(X_train,y_train)
```

```
0.8902215924773625
```

```
ridge.score(X_test,y_test)
```

```
0.8787308696481549
```

Lasso Regression
Optimal Value of alpha = 0.001
Important Predictor = OverallQual_Excellent - 0.913

	Features	CoefValues
87	OverallQual_Excellent	0.913
68	Neighborhood_NoRidge	0.559
75	Neighborhood_StoneBr	0.520
90	OverallQual_Very Good	0.373
69	Neighborhood_NridgHt	0.337
...	...	...
60	Neighborhood_Edwards	-0.184
89	OverallQual_Others	-0.186
127	BsmtQual_Others	-0.233
128	BsmtQual_TA	-0.280
126	BsmtQual_Gd	-0.280

```
lasso.score(X_train,y_train)
```

```
0.8877848585657777
```

```
lasso.score(X_test,y_test)
```

```
0.8813335846015165
```

2.) If the value of alpha for both ridge and lasso becomes double then, there will be change in coefficients values (value become lower) in case of Ridge and there will be less important features in lasso (as some variable coeffiencts changes to 0)

3.) The most important predictor variables after the changes is implemented are:

Ridge Regression
Optimal Value of alpha = 10
Important Predictor = OverallQual_Excellent - 0.593

	Features	CoefValues
87	OverallQual_Excellent	0.593
68	Neighborhood_NoRidge	0.357
75	Neighborhood_StoneBr	0.334
90	OverallQual_Very Good	0.271
69	Neighborhood_NridgHt	0.247
...	...	...
152	KitchenQual_Gd	-0.192
153	KitchenQual_TA	-0.214
128	BsmtQual_TA	-0.230
60	Neighborhood_Edwards	-0.234
126	BsmtQual_Gd	-0.256

```
ridge.score(X_train,y_train)
```

```
0.8854643518791872
```

```
ridge.score(X_test,y_test)
```

```
0.8781140825312901
```

Lasso Regression

Optimal Value of alpha = 0.002

Important Predictor = OverallQual_Excellent - 0.933

	Features	CoefValues
87	OverallQual_Excellent	0.933
68	Neighborhood_NoRidge	0.481
75	Neighborhood_StoneBr	0.433
90	OverallQual_Very Good	0.374
69	Neighborhood_NridgHt	0.294
...	...	...
153	KitchenQual_TA	-0.148
81	BldgType_TwnhsE	-0.155
60	Neighborhood_Edwards	-0.170
128	BsmtQual_TA	-0.193
126	BsmtQual_Gd	-0.211

```
lasso.score(X_train,y_train)
```

```
0.8806337443508282
```

```
lasso.score(X_test,y_test)
```

```
0.8831260233667456
```

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?

Optimal values for both the models are as:

Ridge Regression : 5
Lasso Regression : 0.001

And also we got the good score for both the model :

```
ridge.score(X_train,y_train)
```

```
0.8902215924773625
```

```
ridge.score(X_test,y_test)
```

```
0.8787308696481549
```

```
lasso.score(X_train,y_train)
```

```
0.8877848585657777
```

```
lasso.score(X_test,y_test)
```

```
0.8813335846015165
```

Therefore, using Lasso Regression will be more useful for us as it will eliminate less useful feature by making coeffiencts 0, therefore incresing the performance of the overall 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?

New top 5 important predictor variables are:

ExterCond_TA	-0.400
Neighborhood_Edwards	-0.415
ExterCond_Others	-0.421
CentralAir_Y	-0.421
Foundation_CBlock	-0.432

```
lasso.score(X_trial,y_train)
```

```
0.8703791559233862
```

```
lasso.score(X_trial_test,y_test)
```

```
0.8604011099263565
```

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 robustness can be understood as - If a model has a testing error(on a new test set) equal to the training error, then the model is said to be robust i.e the model generalises well and doesn't overfit.

We can make the model robust and generalisable by:

1. Handling/Removing the outliers
2. By Regularisation i.e. Penalising the coefficients .
3. Cross Validation can be done on the model in order to detect overfitting.
4. Can use more robust error metric technique, i.e. instead of sum of square error, mean sqaured error can be used.
5. Transform the data in order to handle its huge variation.
6. PCA technique can be used to reduce the complexity of the model which helps in to develop more generalise model.

Accuracy of the model remains stable against the new data.
By making model robust and generalisable doesn't mean increasing accuracy of the model. Instead it helps in reducing the complexity of the model and deal with overfitting and underfitting (if occurs) and

keeps prevents the accuracy of the model from fluctuating significantly.