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:
I have used both the neg_mean_absolute_err and neg_mean_squared_error as the scoring mechanism for the grid search CV. The models with neg_mean_absolute_err performed better.

Ridge Regression:

The best alpha for ridge regression came out to be 7. If we double the values of alpha, the more variables will be pushed towards 0.

Top 5 predictors with alpha = 7:

```
OverallQual              0.073090
Neighborhood_NoRidge     0.066436
GrLivArea                0.064088
2ndFlrSF                 0.060807
RoofMatl_WdShngl         0.050541
```

Top 5 predictors with alpha = 14:

```
OverallQual              0.060449
Neighborhood_NoRidge     0.057540
GrLivArea                0.049766
2ndFlrSF                 0.045364
FullBath                 0.039910
```

So, the most important predictor variables after the change is implemented are: OverallQual, Neighborhood_NoRidge, GrLivArea, 2ndFlrSF, FullBath.

Lasso Regression:

The best alpha for ridge regression came out to be 0.0001. If we double the values of alpha, the more variables will be pushed towards 0.

Top 5 predictors with alpha = 0.0001:

```
GrLivArea                0.300112
RoofMatl_WdShngl         0.120086
OverallQual              0.118400
Neighborhood_NoRidge     0.074586
GarageCars               0.058907
```

Top 5 predictors with alpha = 0.0002:

```
GrLivArea                0.279060
OverallQual              0.129071
RoofMatl_WdShngl         0.084388
Neighborhood_NoRidge     0.078688
GarageCars               0.056879
```

So, the most important variables after the changes are implemented are: GrLivArea, Overall Qual, RoofMatl_WdShngl, Neighborhood_NoRidge, GarageCars.

Question 2:
You have determined the optimal value of lambda for ridge and lasso regression during the a ssignment. Now, which one will you choose to apply and why?

Answer:

I chose the Lasso regression with alpha = 0.0001. This is because the lasso regression produc es a model with fewer predictors. Also, I get an R-squared value of 0.899516 on the train set and 0.880663 on the test set which is satisfactory.

Following is the table:

	ridge_mean_squared_error	ridge_mean_absolute_error	lasso_mean_squared_error	lasso_mean_absolute_error
train	0.88279	0.889238	0.822164	0.899516
test	0.867786	0.870975	0.824092	0.880663

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 mod el excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:
After rebuilding the model using lasso regression with the original 5 most important predicto rs dropped, the next important predictors are:

```
1stFlrSF         0.314410
2ndFlrSF         0.194550
LotArea          0.073468
MasVnrArea       0.070746
OverallCond      0.051709
```

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:
We can say that the model is robust if it:

- Does not overfit (the R-squared value on the train and test dataset does not vary massively)
- All the assumptions of linear regression hold.

We checked both these conditions and they are true. Following is the R-squared on the train and test dataset:

	ridge_mean_squared_error	ridge_mean_absolute_error	lasso_mean_squared_error	lasso_mean_absolute_error
train	0.88279	0.889238	0.822164	0.899516
test	0.867786	0.870975	0.824092	0.880663

Following is the distribution of the residuals: