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) For Ridge regression below are the important predictor variables

	columns	Coeff
89	Very Excellent	0.058301
9	GrLivArea	0.056035
15	TotRmsAbvGrd	0.053322
53	NoRidge	0.051101
8	2ndFlrSF	0.050901

2) For Lasso regression below are the important predictor variables

	columns	Coeff
9	GrLivArea	0.325070
89	Very Excellent	0.132622
85	Excellent	0.088968
53	NoRidge	0.058984
60	StoneBr	0.053362

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?

```
R2 score of Ridge Train 0.8851788704694146
R2 score of Ridge Test 0.8050518965198912
==========================================================
RSS of Ridge Train 1.717365330226838
RSS of Ridge Test 1.55492073696985
==========================================================
Mean squared Error of Train for Ridge 0.0014703470293037995
Mean squared Error of Test for Ridge 0.005325071017020034
==========================================================
RMSE of Train for Ridge 0.03834510437205511
RMSE of Test for Ridge 0.07297308419561306
```

```
R2 score of lasso Train 0.9048506044611213
R2 score of lasso Test 0.8777973519344571
==========================================================
RSS of lasso Train 1.4231376555739523
RSS of lasso Test 0.9746975128133477
==========================================================
Mean squared Error of Train for lasso 0.001218439773607836
Mean squared Error of Test for lasso 0.0033380051808676293
==========================================================
RMSE of Train for lasso 0.03490615667196599
RMSE of Test for lasso 0.057775472138855165
```

Lasso produces best R2 score and RMSE. So I will use Lasso Regression.

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?

Below are the 5 most important predictor variables.

	columns	Coeff
6	1stFlrSF	0.270273
7	2ndFlrSF	0.147574
5	TotalBsmtSF	0.122767
51	NoRidge	0.064043
58	StoneBr	0.060383

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?

Keep the test dataset away from the model training and evaluate the model on validation set. So that training set will tested only on final model (final model will be tested with unseen data)

(OR)

Change the folds into different size in the final model and verify the R2 score and means squared error values don't have much variance with the previous results.

Implications:

Training and Test R2 score values should be close. (avoid overfitting)

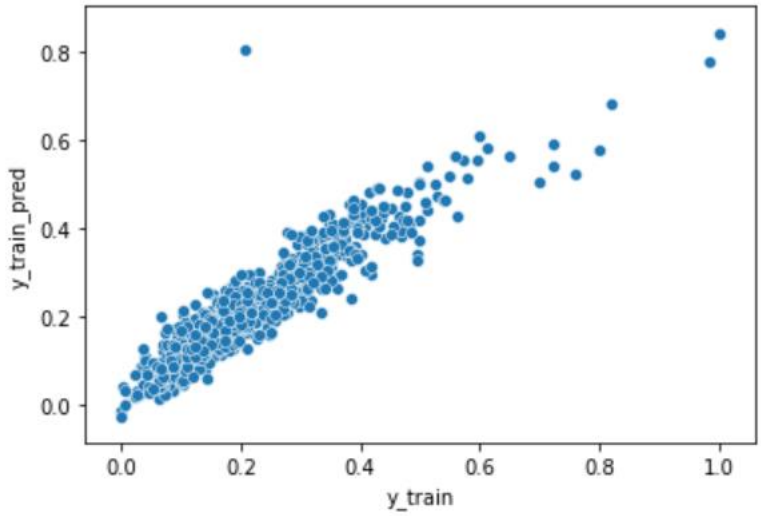
Keep the model as simple as possible (always use feature selection)

Avoid multicollinearity columns in the dataset.

Ensure the RMSE value should be in reasonable range.

Linear relationship between your actual vs predicted elements.

```python
# y_train vs y_train_pred
sns.scatterplot(x=y_train, y=y_train_pred)
plt.xlabel("y_train")
plt.ylabel("y_train_pred")
plt.show()
```




```python
# y_test vs y_test_pred
sns.scatterplot(x=y_test, y=y_test_pred)
plt.xlabel("y_test")
plt.ylabel("y_test_pred")
plt.show()
```