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?

Answer1: for Ridge, optimal value of alpha is 20.

For lasso, optimal value of alpha is 0.001

If we double the value of alpha: the model metrics are as below

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.882995	0.907818
1	R2 Score (Test)	0.856503	0.857926
2	RSS (Train)	119.462004	94.117603
3	RSS (Test)	64.721899	64.079970
4	MSE (Train)	0.342060	0.303615
5	MSE (Test)	0.384405	0.382494

- **-** Changes in Ridge Regression metrics:

R2 score of train set decreased from 0.897 to 0.882

R2 score of test set changed from 0.861 to 0.856

- Changes in Lasso metrics:

R2 score of train set changed to 0.907 from 0.922

R2 score of test set changed to 0.857 from 0.850

the most important predictor variables in ridge with original alpha:

```
OverallQual_9           0.216777
GrLivArea               0.200850
Fireplaces_2            0.190282
Neighborhood_NridgHt    0.179888
GarageCars_3            0.176775
```

the most important predictor variables in lasso with original alpha:

```
OverallQual_10      0.917842
OverallQual_9       0.772296
RoofMatl_WdShngl    0.685467
FullBath_3          0.553067
GrLivArea           0.372001
```

The top 10 important predictor variables in ridge with the change of doubled alpha :

```
OverallQual_9           0.216777
GrLivArea               0.200850
Fireplaces_2            0.190282
Neighborhood_NridgHt    0.179888
GarageCars_3            0.176775
TotRmsAbvGrd_10         0.173527
FullBath_3              0.169123
BsmtExposure_Gd         0.159164
Neighborhood_NoRidge    0.145086
2ndFlrSF                0.140329
```

The top 10 important predictor variables in lasso with the change of doubled alpha :

```
OverallQual_10          0.745474
OverallQual_9           0.723331
FullBath_3              0.506430
GrLivArea               0.347364
Neighborhood_NoRidge    0.338145
RoofMatl_WdShngl        0.314195
OverallQual_8           0.305218
Neighborhood_Crawfor    0.258519
Neighborhood_NridgHt    0.236482
GarageCars_3            0.233474
```

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?

Answer 2:

Since the metrics and model performance of both Ridge and lasso are close to each other having close r2 score, RSS, MSE etc., the model we will choose to apply will depend on the use case.

If we have too many variables and one of our primary goal is feature selection, then we will use Lasso.

If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use Ridge 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?

Answer 3:

The top 5 most important variables for lasso (alpha = 0.001) are:

```
OverallQual_10          0.917842
OverallQual_9           0.772296
RoofMatl_WdShngl        0.685467
FullBath_3              0.553067
GrLivArea               0.372001
```

If the data for top 5 are not available, we will build another model excluding those 5 variables. The five most important variable after this are below:

```
Neighborhood_NoRidge    0.371424
2ndFlrSF                0.356895
1stFlrSF                0.296808
Neighborhood_NridgHt    0.296788
GarageCars_3            0.277917
```

Note: after removing the top 5 variables also, the optimal value of alpha for lasso came to be 0.001

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 4:

- A model is robust when any variation in the data does not affect its performance much.
- A generalizable model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.
- To make sure a model is robust and generalizable, we have to take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.
- In other words, the model should not be too complex in order to be robust and generalizable.
- If we look at it from the perspective of Accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.
- In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge Regression and Lasso.