Assignment-based Subjective Questions
(Advanced Linear Regression)

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 1
Optimal value of alpha obtained for ridge regression = 6
Optimal value of alpha obtained for lasso regression = 0.0001

Regularization technique	**r2_score**	**MSE**	**Top 5 features**
Ridge with alpha = 6	0.88	0.0014	1. Overall Quality 2. Wood shingles roof material 3. Neighbourhood near stone brook 4. Above ground living area 5. Total rooms above ground
Lasso with alpha = 0.0001	0.89	0.0013	1. Above ground living area 2. Wood shingles roof material 3. Overall Quality 4. Lot Area 5. Neighbourhood near stone brook

Results obtained after using alpha = 12 for ridge regression and alpha = 0.0002 for lasso regression:

Regularization technique	**r2_score**	**MSE**	**Top 5 features**
Ridge with alpha = 12	0.87	0.0016	1. Overall Quality 2. Neighbourhood near stone brook 3. Total rooms above ground 4. Garage Cars 5. Above ground living

			area
Lasso with alpha = 0.0002	0.88	0.0014	1. Above ground living area 2. Overall Quality 3. Wood shingles roof material 4. Neighbourhood near stone brook 5. Garage cars

We can see that there is not much difference between the errors and top features. One thing to note is that r2_score has increased and MSE has increased for double the value of alpha. This means that the values obtained initially were indeed optimal and moving away from that increases the error values.

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
Optimal value of lambda obtained for ridge regression = 6
Optimal value of lambda obtained for lasso regression = 0.0001

Both values have been obtained through hyperparameter tuning done through GridSearchCV(sklearn's cross validation technique API). r2_score and MSE obtained on training data and test data are approximately the same. So, we can choose either one of them. However, lasso regression does feature elimination by reducing some of the feature coefficients to zero, so I would choose 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?

Answer 3
After dropping the top 5 features, model was built again.

For lasso regression new alpha value is also 0.0001 and the top 5 features are:

1. First floor square feet
2. Second floor square feet
3. Garage Cars
4. Overall condition
5. Masonry veneer area in square feet

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
To construct a generic ML model we need to ensure that it does not overfit on the training data. If a model is overfitting it means that it has not only recognized the patterns in the data but also the noises. Such a model will perform really well on training/seen data but will fail miserably on test/unseen data. Therefore we need to strike a balance between the bias and variance of the model.

Such a model will only recognize the generic patterns of the test data and will give more accurate results on unseen data than an overfitted model.