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:
The optimal value for alpha is

Ridge regression → **1**
Lasso regression → **0.0001**

When we double the value of alpha, the coefficients of the model features change a little and we see a moderate rise in r2 score for both ridge and lasso models.

r2 score for Ridge(alpha=1) → ~88%
r2 score for Ridge(alpha=2) → ~88.4%

r2 score for Lasso(alpha=0.0001) → ~88.25%
r2 score for Lasso(alpha=0.0002) → ~88.6%

The most important features after the change are :-

1. GrLivArea (Above ground living area)
2. OverallQual (Overall Quality)
3. PoolQC_Gd (Pool quality - good)
4. Condition2_PosN (Near positive off-site feature--park, greenbelt, etc.)

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:
Aside from having a somewhat higher r2 score, the Lasso model has the extra benefit of pushing certain features to 0, resulting in a less complicated and more robust model.

Therefore, we 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?

Answer:
The new most significant features are :-

1. GarageArea (Size of garage)
2. Neighborhood_NoRidge (Near Northridge)
3. Neighborhood_StoneBr (Near Stone Brook)
4. 2ndFlrSF (Second Floor square feet)
5. 1stFlrSF (First Floor 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:
Regularisation is used to ensure that the model is resilient and generalizable.

Regularisation may result in poorer training accuracy, but it aids in improving the model's accuracy on unknown data.

This occurs because regularisation puts a penalty on "high" weight coefficients, which inhibits dataset overfitting. When a model overfits, its accuracy is high on training data but fails on unknown data since it learns even noise from data.