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?

Ridge and Lasso Regression Model are built with optimum alpha calculated in GridSearchCV method. Optimum alpha = 9.0 for ridge and 0.0001 for lasso model.

In case we choose double value of alpha for ridge from 9.0 to 18.0, we get below results:

Model Evaluation: Ridge Regression, alpha=18.0

- R2 score (train): 0.9161
- R2 score (test): 0.8711
- RMSE (train): 0.1134
- RMSE (test): 0.1535

In case we choose double value of alpha for lasso from 0.0001 to 0.0002, we get below results:

Model Evaluation: Lasso Regression, alpha=0.0002

- R2 score (train): 0.9163
- R2 score (test): 0.8712
- RMSE (train): 0.1133
- RMSE (test): 0.1534

In brief, as alpha increase R2 come down & the regression model becomes underfit.

And when we double alpha, training & testing R2 comes down.

Most important predictor variables after the change are implemented i.e., top 5 features in Lasso final model:

- 1stFlrSF
- 2ndFlrSF
- OverallQual
- OverallCond
- SaleCondition_Partial

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?

Model evaluation is done on basis of R2 score and Root Mean Square Error.

Optimum alpha for ridge is 10.000000, so results are:

==========

- R2 score (train): 0.9162727511799029
- R2 score (test): 0.8707483515191132
- RMSE (train): 0.11331675525908998
- RMSE (test): 0.15370583952669675

Optimum alpha for lasso is 0.001000, so results are:

==========

- R2 score (train): 0.9153116265683532
- R2 score (test): **0.8748566174477652**
- RMSE (train): 0.11396529411531482
- RMSE (test): 0.1512433424766017

We choose **Lasso Regression** as in final model for having slightly better R-square value on test data.

Question 3

After building the model, you realized 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?

If five most predictor variables are not available then need to create new model by dropping top five once, for new test data, below are next 5 important variables:

- GarageArea
- KitchenQual
- LotArea
- Fireplaces
- BsmtQual

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

We can make sure that a model is robust and generalizable by:

- Model shouldn't be underfit & overfit for that select hyperparameter alpha carefully in regression machine learning problem.
- Make sure test score efficiency should be always greater than train score efficiency.
- Avoid outliers' effect while building model.

Implications on accuracy:

- Outliers mostly affect on model accuracy so avoid as much outliers if those are irrelevant.
- Standard deviation of value 3-5 will make model enough confident so model results will be standardized as well as consistent for unseen test data.
- If model is of good predictive power, then only it's trusty & worthy to utilize for unseen test data.