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: *Optimal value of alpha for ridge and lasso regression are 6.0 and 0.0005 respectively*.

Ridge: Optimal value of alpha = 6.0
R-squared on Training Dataset: 0.9163167171724649
R-squared on Test dataset: 0.8181232114733891
Top 5 features: ['GrLivArea', 'OverallQual', 'OverallCond', 'GarageCars', 'Age_YearBuilt']

Lasso: Optimal value of alpha = 0.0005
R-squared on Training Dataset: 0.9162512324085729
R-squared on Test dataset: 0.8185467152326872
Lasso Regression picked 47 features and eliminated the other 3 features
Top 5 features: ['GrLivArea', 'OverallQual', 'OverallCond', 'GarageCars', 'Age_YearBuilt']

After doubling the optimal value of alpha,
Ridge:
R-squared on Training Dataset: 0.9162512324085729
R-squared on Test dataset: 0.8185467152326872

Lasso:
R-squared on Training Dataset: 0.9159651703802929
R-squared on Test dataset: 0.81897349485103

If we choose double the value of alpha for both ridge and lasso, then there is not much significant change noticed, but the R-squared for training for both the models reduced a bit as we have doubled the value of regularized hyperparameter. Also the R-squared for testing for both the models increased a bit as the models has become more generalized.

After the change is implemented, there is no change in the most important predictor variables noticed. The top five most important predictor variables are 'GrLivArea', 'OverallQual', 'OverallCond', 'GarageCars' and 'Age_YearBuilt' (Derived variable)

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: *As per model developed, there is no significant difference in R-squared for training and testing between the models.*

Ridge: Optimal value of alpha = 6.0
R-squared on Training Dataset: 0.9163167171724649
R-squared on Test dataset: 0.8181232114733891

Lasso: Optimal value of alpha = 0.0005
R-squared on Training Dataset: 0.9162512324085729
R-squared on Test dataset: 0.8185467152326872
Lasso Regression picked 47 features and eliminated the other 3 features

As per the figure mentioned above, it is noticed that the R-squared for training for lasso is less than ridge where as the R-squared for testing for lasso is greater than ridge. I'll like to go with Lasso Regression as Lasso Regression picked 47 features and eliminated the other 3 features and the model is helping in feature selection as well and dropping the insignificant features hence making the model more robust.

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 five most important predictor variables are 'TotalBsmtSF', '2ndFlrSF', 'KitchenQual', 'Full Bath' and 'Fireplaces'*

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: *A model is considered to be robust if the model does not change drastically upon changing the training set, i.e. the model is stable. The model is considered generalisable if it does not overfits the training data, and works well with new data. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.*