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?

- Optimal Values:
  - Ridge Regression : 2
  - Lasso Regression : 0.001
- If the alpha values are doubled, the coefficients will vary and some of them will be pushed even further to zero, both in case of ridge & lasso regression
- Important predictors after the change:
  - Ridge – RoofMatl_ClyTile, RoofMatl_Roll, BsmtFinSF2Sqrt, GrLivAreaLog, RoofMatl_CompShg
  - Lasso – GrLivAreaLog, BsmtQual_Ex, 1stFlrSFLog, GarageAreaLog, Neighborhood_NoRidge

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?

- For the optimal values of lambda for ridge and lasso regression, their respective models deliver R2 scores of 0.87 and 0.88 on unseen test data respectively.
- However, the lasso model is performing as good as the ridge model, even with a significantly less number of predictors (coeff > 0).
- The lasso model, hence, is relatively simpler, more generic and easier to interpret and is the optimal choice.

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?

- Top 5 predictors are : ['GrLivAreaLog', 'RoofMatl_ClyTile', '1stFlrSFLog', 'MSZoning_C (all)', '2ndFlrSF']
- The corresponding data features are : ['GrLivArea', 'RoofMatl', '1stFlrSF', 'MSZoning', '2ndFlrSF']
- The model built after dropping these features has the following 5 important predictors:
  - 'TotalBsmtSFLog', 'LotAreaLog', 'GarageAreaLog', 'Neighborhood_Crawfor', 'Neighborhood_NoRidge'

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?

- By making sure the model is not overfitting, i.e., the train & test scores are fairly similar
- By making sure the model does not need too many predictors relative to the dataset
- By using ridge/lasso regressions, that employ a cost function relative to the coefficients, in additional to RSS
- The implication is the accuracy of the model in the training data is compensated for an increase in performance on the unseen test data.
    - The reason is that the model is stopped from memorizing the train data and only learns the overall patterns that can be generalized.