Advanced Linear Regression
Subjective questions

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?

   - The optimal value of alpha for ridge and lasso regression are 240 and 0.0158 respectively.
   - With doubling the alpha value, the r-square metric for train and test, for both lasso and ridge regression, dropped considerably. It shows the model underfitting both the train samples, and consequently results in sub-optimal score for test dataset.
   - Coefficients of the top 3 (important) predictor variables increased after doubling alpha. This phenomenon observed for both lasso and ridge regression.

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?

   Lasso regression will be selected for the following two reasons:
   1. Optimal test r-square was better at 0.829 in comparison to 0.824 for ridge.
   2. Feature selection was done better with removing those that have zero coefficients. With fewer features, the model becomes simple and yet outperform on the r2 score.

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?

   Five most important predictor variables are:
   a. First floor sq. ft. : 1stFlrSF
   b. Second floor sq. ft. : 2ndFlrSF
   c. Garage area : GarageArea
   d. # of full bathrooms : FullBath
   e. Year remodeled : YearRemodAdd

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?

   The original data is split into train and test at 0.7/0.3 ratio. The model was trained with Lasso (Ridge) on the train dataset. To evaluate the robustness and generalizability of the model, predictions were performed on unseen data, test samples, unseen by the model. The metric used to evaluate the model performance was r-square. The optimal model is the one where the divergence between the train and test value is minimum, while ensuring the test performance is near maximum. These conditions ensure that the model is generalizable and robust in the test phase.