1) What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Optimal value of alpha for ridge: 4

Optimal value of alpha for ridge: 100

After doubling alpha for ridge and lasso i.e., 8 and 200

Changes in Ridge Regression metrics:

R2 score of train set decreased from 0.94 to 0.93

R2 score of test set remained same at 0.93

Changes in Lasso metrics:

R2 score of train set decreased from 0.92 to 0.91

R2 score of test set decreased from 0.93 to 0.91

Top Features:

- OverallQual
- OverallCond
- YearBuilt
- Neighborhood_StoneBr
- Exterior1st_BrkFace
- TotalBsmtSF
- LotArea

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?

Both Lasso and Ridge has almost same R2 score, But I do observe the test R2 score for Lasso is slightly higher than the Ridge. Hence, I choose Lasso.

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?

   New predictor variables, refer to Jupyter notebook (Advanced_Regression_Assignment) cells 192 – 198 for detailed analysis.

   a. -LotFrontage
   b. -MasVnrArea
   c. -BsmtFinSF1
   d. -GrLivArea
   e. -YearRemodAdd

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?

   The Regression model calculated should not decrease the value of the test data significantly. One should make sure the outliers are not provided more importance during building a model. It is also important to avoid overfitting the training data due to which the model might not work properly on the unseen test data. A generalized model will have a good trade-off between bias and variance.