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 Regression:

Optimal Value: 0.001

R2 Score: 0.8874218919494045

Important Predictor Var: RoofMatl, TotalBsmtSF, Condition2_PosN, 1stFlrSF, LotArea

Note: The first few features are in infact derived from RoofMatl namely RoofMatl_Membran, RoofMatl_Metal, RoofMatl_Tar&Grv, RoofMatl_WdShngl, RoofMatl_CompShg, RoofMatl_Roll, RoofMatl_WdShake

After doubling the alpha value, we still see the same set of important predictor variables.

Only differences is in coefficients of these features and a very very slight increase in R2 score to 0.8874279859172068

Lasso Regression:

Optimal Value: 0.00001

R2 Score: 0.8862645807375794

Important Predictor Var: RoofMatl, TotalBsmtSF, Condition2_PosN, 1stFlrSF, LotArea

Note: The first few features are in infact derived from RoofMatl namely RoofMatl_Membran, RoofMatl_Metal, RoofMatl_Tar&Grv, RoofMatl_WdShngl, RoofMatl_CompShg, RoofMatl_Roll, RoofMatl_WdShake

After doubling the alpha value, we still see the same set of important predictor variables.

Only differences is in coefficients of these features and a very very slight increase in R2 score to 0.8874279859172068

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?

We will choose the value of lambda which gives a good R2 Score. Hence as mentioned in previous answer, we chose lambda of 0.001 for Ridge and 0.00001 for Lasso Regression. Though a little higher still gives a good R2 Score, the more higher like more than 10 times gives poor R2 Score and hence are not to be chosen.

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?

Considering RoofMatl as one predictor variable (ignoring the dummies created from it), and exluding the other 4 top predictors namely, TotalBsmtSF, Condition2_PosN, 1stFlrSF, LotArea, we now have to find the next 5 important predictor variables. From the experiment or jupyter code, we see them to be 2ndFlrSF, (ignored Condition2_RRAe – derived dummy of Condition2) BsmtFinSF1, OverallQual, Functional_Sev, RoofStyle_Shed

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?

We have considered RFE for selecting only the top features with removal of features using trial and error of feature count using VIF as reference. In other words, we started with feature count 50 and kept coming down till we did not see any infinity (Inf) VIF value. That value came to be 38. Thus using RFE itself, we have ensured a robust and general model. With Ridge and Lasso Regressions, we can ensure that there is no overfitting of the data. For the same, we select the model with the best R2 Score.