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: The optimal value of alpha for ridge and lasso regression is 7.0 and 0.0001 respectively. If we double the value of alpha for both ridge and lasso then R2_score for training set and testing set is more near to each other, as well as model coefficients are reduced more towards zero in ridge, while in lasso more coefficients shrunk to zero.
Most important 5 predictor variables for ridge is: GrLivArea, OverallQual, 1stFlrSF, Neighborhood_StoneBr, TotalBsmtSF

Most important 5 predictor variables for lasso is: GrLivArea, OverallQual, Condition2_PosN, Neighborhood_StoneBr, 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?

Answer: we will apply lasso regression in our case, because we have total 255 columns (with dummy variables) and we know in lasso regression model coefficients are shrunk to exactly zero which provides feature selection. So, it will reduce features and will be more interpretable. In our case, alpha for lasso is 0.0001, after training we got R2_score for training set is 93.0 %, while for testing set is 91.5%. We can use lasso for sure.

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: After retraining of lasso model top 5 predictor variables are: 1) 1stFlrSF, 2) 2ndFlrSF, 3) TotalBsmtSF, 4) Condition2_Norm & 5) LandContour_HLS

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: we can use R2_score or mean sq. error to find that whether model is generalized or not, value of R2_score or any other metric for training set and testing set should be very near to each other. If model is not generalized then whatever prediction we will have it will be wrong. So, to make model more robust and generalized we require optimal trade-off between variance and bias. To make this we can use regularization or cross-validation.