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?

Ans:
Optimal Value for Ridge regression is 20. If I double this value to 40, there is slight impact on Test score, but more importantly, we get additional feature as significant one i.e. Kitchen Quality in addition to other 3 which appeared for value 20
For Lasso regression, optimal value of alpha is 100 where Train and test scores are improving. If we double the values, too many variables become significant based on P-values. So it is not optimum.

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?

Ans: I will choose 20 for Ridge and 100 for Lasso. These values provide correct number of predictors 3-4 which can be used to explain the model. Test and Train scores are also optial at this stage.

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?

Ans : As per Ridge, Condition2 , Neighborhood_NridgHt, GarageArea , GarageFinish, RoofMatl_WdShngl are next 5 significant features
As per Lasso, 1stFlrSF, GarageArea, BsmtQual and GarageFinish and OverallQual are next best significant features

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?

Ans: EDA and Feature engineering is important, if I am able to combine few features model will be more accurate. Currently test score is hovering at 80% which is lower as compared to typical industry observations. More data is needed. There are only 1400 observations. Test score should improve with more training and test data.