Q1. 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 alpha value for ridge regression : 0.7
Training data r2 score = 0.838
Test data r2 score = 0.818

Changes in model if we double the value of alpha for ridge regression
Training data r2 score = 0.86175
Test data r2 score = 0.835

Optimal alpha value for lasso regression : 10.0
Training data r2 score = 0.8659
Test data r2 score = 0.8357

Changes in model if we double the value of alpha for lasso regression
Training data r2 score = 0.8650
Test data r2 score = 0.8369

Most important predictor variables after above changes: OverallQual, GrLivArea, LotArea, 1stFlrSF, TotalBsmtSF

Q2. 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:
The r2_score of lasso is slightly higher than lasso for the test dataset so we will choose lasso regression to solve this problem

Q3. 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:
Current 5 most important predictor variables are:
1stFlrSF, LotArea, OverallQual, GrLivArea, TotalBsmtSF

If we do not get these 5 predictor variables then we can use below 5 predictor variables:

After removing above variables we are left with below 5 most important predictor variables:
YearBuilt, BsmtFinSF1, Exterior2nd_CmentBd, 2ndFlrSF, OverallCond

Q4. 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: Model needs to made robust and generalisable so that they are not impacted by outliers in the training data. Model should be generalisable so that the test accuracy is not lesser than the training score. Model should be accurate for datasets other than the ones which were used during training. To ensure that only relevant data is part of model creation , outliers which are not relevant needs to be removed. Removing outliers will improve accuracy and predictions.