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 : optimal value of alpha for ridge is 50. optimal value of alpha for lasso is 0.01. The r2 for train and test both has reduced after double the value of alpha. Ridge Alpha:50, r2 train 82, r2 test: 77. Lasso Alpha:0.001 r2 train:90,r2 test: 81. These are most important predictor variables after the change is implemented OverallQual, GrLivArea, Fireplaces, 2ndFlrSF, 1stFlrSF, GarageArea, FullBath, KitchenQual, ExterQual, BsmtFinSF1

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: I will choose Ridge because Ridge looks better here since difference between test and train is less than Lasso. But R2 of Ridge is less than Lasso but that will be ok since R2 is not suppose to determine overall fit of the model

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:

OverallQual, GrLivArea, Fireplaces, 2ndFlrSF, 1stFlrSF

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: The model should be generalisable so that the test accuracy is not less than the training score. The model should be accurate for datasets other than the ones which were used during training. Missing value imputation with mean or median should be taken properly, this will impact the r2 score the model in test data. R2 score should not be perfect since this will overwrite all data point and will not perform well on unseen data. If Accuracy is high, model will not perform well on unseen data. If the model is not robust , it cannot be trusted for predictive analysis.