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?

The Optimal value of alpha for Ridge is: 500
The Optimal value of alpha for Lasso is: 0.01

After doubling the value of Alpha for both Lasso and Ridge regression

The Optimal value of alpha for Ridge is: 1000

- The r2 score increased from 0.893 to 0.92 for the train data set
- The r2 score dropped from 0.861 to 0.836 for the test data set

The Optimal value of alpha for Lasso is: 0.02

- The r2 score dropped from 0.896 to 0.877 for the train data set
- The r2 score dropped from 0.86 to 0.85 for the test data set

The most relevant variables after the change of the value of alpha are:

- 'GrLivArea'
- 'OverallQual_9'
- 'OverallQual_8'
- 'OverallQual_10'
- 'GarageCars_3'
- 'FullBath_3'
- 'Neighborhood_NoRidge'
- 'Fireplaces_2'
- 'Neighborhood_NridgHt'
- 'BsmtExposure_Gd'
- 'GarageArea'
- 'YearBuilt_cls_After_1980'

- 'TotalBsmtSF'
- 'Neighborhood_Crawfor'
- 'TotRmsAbvGrd_10'
- 'Condition1_Norm'
- 'CentralAir_Y'
- 'BsmtFullBath_1'
- 'SaleType_New'
- 'LotArea'
- 'OverallQual_7'
- 'BsmtFinType1_GLQ'
- '1stFlrSF'
- 'Neighborhood_Somerst'
- 'Functional_Typ'
- 'Fireplaces_1'
- 'HalfBath_1'
- 'LotConfig_CulDSac'
- 'KitchenAbvGr_1'
- 'GarageType_Attchd'

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?

Lasso Regression Model would be the model of choice because even though it has a slightly lower R squared value, it would be preferred since it actually does feature selection by eliminating all variables that do not contribute significantly to the model, therefore allowing the model to be more generic without no much compromise on bias.

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?

After remodeling without the first five variables part of the data set, we have these top 5 variables.

- '2ndFlrSF'
- '1stFlrSF'
- 'GarageArea'
- 'YearBuilt_cls_After_1980'
- 'Fireplaces_2'

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?

A model is very generalizable if it is not overly complex and does not overfit. To check if a model is generalizable you have to test it with data it was not trained with as see how it performs. A non generic model would perform very well on a training data set but not on test data. A very high variance will be observed in the prediction of an overfitting model.

To make an overfitting model generalizable, we need to regularize by adding a hyperparameter that will have an alpha value that makes all the beta values that are part of the model to eliminate overfitting. The implication of regularization is that there is a significant reduction in the variance with a little compromise on

bias of the model. For a little more bias in the model we can eliminate a significant amount of variance in the model when we find the optimal value of the alpha.