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 in Ridge regression is 10.0 and the optimal value of alpha in Lasso regression is 0.001 after carrying out GridSearchCV.

When the value of alpha is doubled for Ridge regression and Lasso regression, the most important predictor variables change. The R2 score decreases slightly when optimum value of alpha is doubled. This is because, doubling the alpha value causes the penalty to be increased. This can cause the complexity of the model to decrease. This causes the bias of the model to increase, hence R2 score of the model will be less.

The R2 score of train data with optimum alpha is 94.9 (For Ridge regression) and 92.9 (For Lasso regression) and is 92.5 (For Ridge regression) and 91.6 (For Lasso regression) when alpha value is doubled.
The most important predictor variables when alpha is doubled (alpha=20) for Ridge regression are:

- GrLivArea
- OverallQual_8
- YearRemodAdd
- Neighborhood_Crawfor
- Neighborhood_Somerst

The most important predictor variables when alpha is doubled (alpha=0.002) for Lasso regression are:

- GrLivArea
- OverallQual_8
- OverallQual_9
- Neighborhood_Somerst
- YearBuilt

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?

The decision to choose between ridge regression and lasso regression depends on few factors. In case of Ridge regression, the penalty is the sum of squares of the coefficients. This causes the coefficient values to become small values (but not zeros) so as to decrease the penalty term.

Hence, Ridge regression is useful in cases where we have to retain all the predictor variables and not leave any predictor variable.

In case of Lasso regression, the penalty is the sum of the absolute values of the coefficients. This causes the coefficient values to become zero value (as per the model) so as to decrease the penalty term. Lasso regression is useful in cases where some variables which are not needed as per the model can be left out, since the coefficients will be zero for such variables in Lasso regression.

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 excluding the five most important variables in the original Lasso regression model, the most important parameters in the new Lasso model are:

- BsmtQual_No
- MSZoning_FV
- MSZoning_RH
- MSZoning_RL
- MSZoning_RM

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?

For a model to be robust and generalizable, it has to be simple model. A complex model can cause the data to be fit for the train dataset only and hence cause overfitting. A complex model can cause increase in variance. Hence, it is a tradeoff between bias and variance.

Bias shows how much error the model is likely to make. Variance shows how sensitive the model is to input data. If a model is too complex, it is more likely to change with changes in input data but it will have low bias. If the model is too simple, it will have high bias but low variance.
Bias defines how accurate the model is likely to be on future test data. Extremely simple model fails to predict complex and real-world phenomena.

If a model is robust and generalizable, it may cause a slight compromise on the accuracy of the model. However, it performs better when the input data changes.