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. As given by the model, the best value for ridge regression was 3.0 and the best value of alpha for Lasso Regression was 0.0001.

I built the model again with the alpha values doubled to 6 and 0.0002 respectively. The r2 score on train and test sets changed. Following are the values:
Original r2 score on train data (for ridge regression) : 91.33%
Original r2 score on test data (for ridge regression) : 79.15%
Changed r2 score on train data (for ridge regression) : 90.26%
Changed r2 score on test data (for ridge regression) : 80.72%
Original r2 score on train data (for lasso regression) : 91.72%
Original r2 score on test data (for lasso regression) : 76.23%
Changed r2 score on train data(for lasso regression) : 90.73%
Changed r2 score on test data(for lasso regression): 78.36%

As we can see, in both the cases, the r2 dropped on the train sets but increased on the test sets. This shows that the problem of overfitting was solved by increasing the alpha values. In case of lasso regression, more variables were set to zero coefficients and in case of ridge regression, the coefficients became smaller.

Let's look at the most important predictor variables after the implementation of the change:
For Ridge Regularization:

1. GrLivArea : Above grade (ground) living area square feet
2. OverallQual : Rates the overall material and finish of the house
3. 2ndFlrSF : Second floor square feet
4. 1stFlrSF : First floor square feet
5. Neighbourhood_NoRidge : North Ridge Neighbourhood location

For Lasso Regression

1. GrLivArea : Above grade (ground) living area square feet
2. OverallQual : Rates the overall material and finish of the house
3. BsmtFinSF1 : Type 1 finished square feet
4. RoofMatl_WdShngl : Roof Material is Wood Shingles
5. Garage Area : Size of garage in square feet

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 would choose to apply Ridge Regression. This is because of the difference in the r2 scores between the two models.
Given below are the r2 scores for comparison:
r2 score on train data (for ridge regression) : 91.33%
r2 score on test data (for ridge regression) : 79.15%
r2 score on train data (for lasso regression) : 91.72%
r2 score on test data (for lasso regression) : 76.23%

The r2 scores on the train sets are nearly same but those the test sets tells us that ridge regression handles the problem of overfitting by performing better on unseen data well.

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. The next five most variables will be

1. TotalBsmtSF : Total square feet of basement area
2. GarageArea : Size of garage in square feet
3. OverallCond : Rates the overall condition of the house
4. Neighborhood_NoRidge : North Ridge Neighbourhood
5. MasVnrArea : Masonry veneer area in square feet

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. To make a model robust and generalisable, here are some steps we can take

a. Split the model into train and test : This helps us in evaluating the model better. When we build a model, it is with the sole purpose that it would perform effectively on unseen data. This is the role of the test set.
b. Use of cross-validation method like k-fold cross validation : One of the major issues that we want to avoid while building a model is overfitting, i.e the model shouldn't learn too much of the training set. This will hinder its performance on unseen data. Cross -validation allows us to use different subsets of the train set to train the model. This gives a more robust fitting model.
c. Regularization methods like Ridge and Lasso Regularization : These techniques penalise the terms with large coefficients. While Ridge regression reduces the coefficients of the terms significantly with the optimal value of alpha, Lasso regression puts some terms to zero entirely, i.e it performs feature selection.
d. Model Selection : Building two or three different models using different techniques and then comparing them to prevent over and underfitting is the best method to ensure that the model is robust and generalisable.

Models built using the above methods may not perform very well on the training data, it tends to perform well on the test set. This shows that the model hasn't overfit. The models need not be too complex, with lots of feature but must have an optimum number of features to make accurate predictions.