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:

Optimal value of alpha for Ridge is 3.0

Optimal value of alpha for Lasso is 0.001

R2 Score and Mean squared error for optimal alpha:

	R2_Test	R2_train	MSE_test	MSE_train
Ridge_alpha	0.925819	0.937046	3.486967e+08	2.746770e+08
Lasso_alpha	0.917612	0.921479	3.872763e+08	3.425995e+08

R2 Score and Mean squared error for double of optimal alpha:

	R2_Test	R2_train	MSE_test	MSE_train
Ridge_alpha2	0.921809	0.931156	3.675483e+08	3.003761e+08
Lasso_alpha2	0.907381	0.909459	4.353670e+08	3.950469e+08

R-score is better and mean squared error is lesser for optimal alpha. Hence, using optimal alpha gives a better model.

After implementing the change---

Top 5 predictors for Ridge:

	Ridge_Double_Alpha	L
OverallQual	0.250478	
GrLivArea	0.238504	
1stFlrSF	0.209177	
OverallCond	0.177226	
TotalBsmtSF	0.163323	

Top 5 predictors for Lasso:

	Lasso_Double_Alpha
GrLivArea	0.673333
OverallQual	0.474167
TotalBsmtSF	0.219081
OverallCond	0.187077
GarageCars	0.153244

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:

Though R score and mean squared error came out better using Ridge than Lasso, I would still choose Lasso, as it simplified the model by driving some co-efficient to zero. Lasso had 31 non-zero co-efficient in the final model, whereas Ridge used all 195 co-efficient.

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:

Using optimum alpha in lasso model, the 5 most important predictors were:

GrLivArea	0.710927
OverallQual	0.452359
OverallCond	0.246821
TotalBsmtSF	0.236818
YearBuilt	0.185968

Building new model after removing the above predictors, the 5 most important predictors now are:

1stFlrSF	0.805781
2ndFlrSF	0.360497
BsmtFinSF1	0.154489
GarageCars	0.142752

KitchenQual_Fa -0.122389

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:

R score and MAE (Mean absolute error) give us idea on how well the model is performing. Comparing these two factors for training and test data will give a fair idea on whether the model is stable.

If a model is not stable, it may perform very well with training data. But when it encounters unseen data, model scores will deteriorate drastically. The reason behind this is, unstable models tend to memorize all training data a.k.a overfit. Hence, it fails to generalize the solution and perform well with unseen data. We can use regularization techniques to ensure that the model does not overfit.

Regularization techniques like Ridge and Lasso penalize the model for the co-efficient used. Hence it discourages overfitting.