Q1

Question 1

What is the optimal value of alpha for ridge and lasso regression?

- Ridge & Lasso:
  - For both ridge and lasso,
    - Selecting the optimal alpha value is crucial, since in case the alpha is too less, the model will be overfitting and if it is too high, model will be under-fitting.
- Ridge : Optimal alpha value = 0.3
- Lasso : Optimal alpha value = 0.0001

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

- Ridge :
  - In ridge regression, if the value of lambda is 0 then the cost function will not be having the penalty term, means there wont be any shrinkage of model coefficients.
  - As the alpha value increases, shrinkage penalty will increase and the model coefficients will be reduced and tends to be zero (near to zero and will not be exactly zero) as alpha is infinity.
- Lasso :
  - The lambda characteristics of lasso is exactly similar to ridge, but the only difference is that, the coefficients of ridge will be tends to zero (meaning nearly zero and not exactly zero), but in lasso, the coefficients of the in-significant variable will be zero. Thus, lasso is also performing variable selection.

	Metric	Linear Regression	Ridge Regression	Lasso Regression	Double Ridge Regression	Double Lasso Regression
0	R2 Score (Train)	0.901366	0.901091	0.900657	0.900643	0.899585
1	R2 Score (Test)	0.858757	0.861306	0.862465	0.862432	0.864984
2	RSS (Train)	13.897916	13.936630	13.997768	13.999695	14.148841
3	RSS (Test)	10.672933	10.480296	10.392726	10.395194	10.202420
4	MSE (Train)	0.120698	0.120866	0.121131	0.121139	0.121783
5	MSE (Test)	0.150533	0.149168	0.148544	0.148561	0.147177

After doubling the alpha values,

- There is minor variation in the R2 values,
- Upon increasing the value of alpha, the R2 value of train data will be reduced, since this is basically making the data more general.
- Upon continuing increasing the alpha value, the model will become under-fitted and R2 of both train and test will be reduced

What will be the most important predictor variables after the change is implemented?

Both before doubling and after doubling,
The important predictor variable remains the same.

1) OverallQual
2) TotRmsAbvGrd
3) **YearBuilt**
4) **TotalBsmtSF**
5) **OverallCond**

Q2

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?

- Ridge:
    - As mentioned earlier, as the value of the alpha increases, the shrinkage penalty will increase and the model coefficients will be reduced and tends to be zero (near to zero and will not be exactly zero) as alpha is infinity
    - Number of coefficients will be same but with simpler coefficients
    - Features which are not influencing the output will also be present
- Lasso:
    - As the value of the alpha increases, the coefficients of the features which are not influencing the output will become zero, means that, the variables will be eliminated. In other words, Lasso also does variable elimination.
    - The value of RSS and MSE is also comparatively less than Ridge.
    - Thus, Applying Lasso is preferable than applying ridge since in-significant variable will be eliminated.

Q3

Question 3

After building the model, you realized 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 removing the top 5 important predictor variable and developing the lasso model, the following are the five most important predictor variable:

	Lasso
OverallQual	0.817756
TotRmsAbvGrd	0.583019
YearBuilt	0.471267
TotalBsmtSF	0.362510
OverallCond	0.338498

Q4

<u>Question 4</u>

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

- Robust and generalizable :
  - These are the terms used for using the model developed for the unseen data.
  - As the model is more robust and generalized, unseen data can be used to predict more reliable output
- Train data :
  - If the model is not robust and generalized, it means that the model built have memorized the train data. This can be measured by using the R square value.
  - For a over-fitted model, the R square value of the train data set will be very high than the test data set.
  - The bias will be less and variance of the model will be more.
- Test data :
  - If the R squared value of the test data is very less than the training data, then the model is over-fitted and can be generalized.
- Regularization:
  - If the model is more general, the bias will be high and causing the output to be unreliable.
  - So, trade-off between bias and variance has to be achieved for getting a robust and generalized model.
  - Regularization helps in achieving a less complex model by shrinking the model coefficients. Thus, preventing the model from overfitting.
- Accuracy:
  - As we are generalizing the model, the accuracy of the train data will reduce while the accuracy of the test data will increase.