Subjective Questions:-

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?

Answer-1:The optimal value of Ridge regression: 100
The optimal value of Ridge regression after RFE: 20
The optimal value of Lasso regression: 0.001

It has been observed that for all the models, the training score has decreased slightly as compared to the testing score. After doing RFE on the ridge regression, the changes occur more noticeable. If we double the value of alpha for both ridge and lasso then the gap between the training and test set decreases with the increase in the values.

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?

Answer-2: The optimal lambda value for ridge and lasso is as follows:
Ridge - 20
Lasso - 0.001

Out[73]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.559854e-01	0.949068	0.950742
1	R2 Score (Test)	-6.367209e+16	0.899484	0.901259
2	RSS (Train)	6.568165e+00	7.600360	7.350567
3	RSS (Test)	3.633004e+18	5.735246	5.633978
4	MSE (Train)	8.319341e-02	0.089492	0.088009
5	MSE (Test)	9.447909e+07	0.118708	0.117655

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?

Answer-3: The top 5 predictors variables are: ‘GrLivArea’, ‘OverallQual’, ‘TotalBsmtSF’,’OverallCond’, ‘YearBuilt

GrLivArea 0.128692

OverallQual 0.065629

TotalBsmtSF 0.048505

OverallCond 0.043846

YearBuilt 0.043804

After removing them, the predictor variables are:

GrLivArea 0.146928

MSZoning_RL 0.083526

MSZoning_RM 0.065606

TotalBsmtSF 0.056324

GarageCars 0.043311

,

These variables remain the same after having a slight difference order after applying Lasso to RFE created variables.

GrLivArea 0.136432

MSZoning_RL 0.083509

TotalBsmtSF 0.073250

MSZoning_RM 0.069320

GarageCars 0.041585

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?

Answer-4: As Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test
data, we should pick the one that makes fewer on the test data due to following reasons:-

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust

Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.
Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use.

Also, Making a model simple leads to Bias-Variance Trade-off:
• A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
• A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.
Bias quantifies how accurate the model is likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g. one that gives the same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.
Variance refers to the degree of changes in the model itself with respect to changes in the training data.
Thus accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph