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

With 20 features and selecting alpha from [0.0001, 0.001, 0.01,0.05, 0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9, 1, 10,20,50, 100,500, 1000] with 5 folds validation gives below two best values

Ridge: 0.1

Lasso: 20

If we choose to double the values of alpha, most important predictor variables will not change, but the coefficients will change. And hence R-squared and Cost Function values will change

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

I will us ridge regression, as it is giving me less R2 score difference between train and test sets. And also RSS is quite low in case of Ridge Regression

Metric	Linear Regression	Ridge Regression	Lasso Regression
R2 Score(Train)	8.58E-01	8.40E-01	8.53E-01
R2 Score(Test)	7.39E-01	7.57E-01	7.50E-01
RSS (Train)	9.03E+11	1.02E+12	9.39E+11
RSS (Test)	7.39E+11	6.88E+11	7.06E+11
MSE (Train)	2.97E+04	3.16E+04	3.03E+04
MSE (Test)	4.10E+04	3.96E+04	4.01E+04

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?

Answer3

Next 5 parameters

Parameter	Coef
1stFlrSF	436556.85
Condition2_PosN	-381412.5
OverallQual	169973.85
2ndFlrSF	94577.684
LotArea	91429.324
BsmtFinSF1	49700.344

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

We make sure that the model is robust and genralisable following below points

1. Not making the model too complex
2. Feature selection using RFE
3. Avoiding overfitting by regularization

When we use regularization, we basically penalize the co-efficients by introducing lamda. This helps in dealing with the problem overfitting. Hence the accuracy of the model will go down in training set, but the model will work better in unseen data