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?

The optimal value of alpha for ridge and lasso regression are as following,

Alpha for Ridge Regression = 100

Alpha for Lasso Regression = 500

The metrics we got by these alpha values are as follows,

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.407341e-01	9.280144e-01	9.273976e-01
1	R2 Score (Test)	-4.535714e+21	7.302949e-01	5.673580e-01
2	RSS (Train)	4.017076e+11	4.879227e+11	4.921033e+11
3	RSS (Test)	1.099510e+34	6.537969e+11	1.048775e+12
4	MSE (Train)	1.983545e+04	2.186063e+04	2.195408e+04
5	MSE (Test)	5.010287e+15	3.863531e+04	4.893326e+04

Suppose, if we choose to double the value of alpha for both Ridge and Lasso, then the values are as follows,

Alpha for Ridge Regression = 200

Alpha for Lasso Regression = 1000

The metrics we got by these alpha values are as follows,

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.407341e-01	9.183924e-01	9.149281e-01
1	R2 Score (Test)	-4.535714e+21	7.852522e-01	5.993207e-01
2	RSS (Train)	4.017076e+11	5.531414e+11	5.766223e+11
3	RSS (Test)	1.099510e+34	5.205739e+11	9.712939e+11
4	MSE (Train)	1.983545e+04	2.327583e+04	2.376473e+04
5	MSE (Test)	5.010287e+15	3.447499e+04	4.709104e+04

The most important predictor variables after this change of alpha value are,

For Ridge Regression,

Variables	Value
OverallQual	9.536537e+03
Condition1_RRAn	9.619532e+02
MasVnrType_Stone	9.577810e+02
BsmtFinSF2	9.495807e+02
Condition2_Norm	9.394410e+02
GarageType_BuiltIn	9.231467e+02
SaleCondition_Normal	9.167678e+02

For Lasso Regression,

Variables	Value
GrLivArea	29281.682681
OverallQual	13669.770915
YearBuilt	8181.314350
TotalBsmtSF	7629.696866
BsmtFinSF1	7345.736461
Neighborhood_NridgHt	7297.406544

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?

I will choose Lasso over Ridge Regression because the R2_Score of Lasso is lesser compared to Ridge.

And by using lasso we can make the model simple by reducing the number of features.

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?

The five most important predictor variables after making exclusion changes are,

1stFlrSF	27220.667914
2ndFlrSF	24821.638395
Neighborhood_NridgHt	8136.879689
Neighborhood_NoRidge	7572.215362
BsmtExposure_Gd	7152.132438

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?

We can say that a model is more robust and generalisable when the scores of the test and train data are near to similar. Sometimes even if the training data score is lesser, the overall difference between the test and train data should be lesser in order to be generalisable. The accuracy of a robust and generalisable model will always be somewhat lesser than the overfit model.