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?

A: The Optimal value of alpha for **ridge** is 4

Train data accuracy score for alpha = 4 is 89.7%

Test data accuracy score for alpha = 4 is 88.6%

The Change in accuracy score when alpha is 8 are:

Train Score = 89.6%

Test Score = 88.6%

Alpha = 4

	Features	Coefficients
11	MSZoning_RL	0.377
7	GrLivArea	0.360
12	MSZoning_RM	0.269
2	OverallQual	0.233
9	MSZoning_FV	0.199
6	TotalBsmtSF	0.193
3	OverallCond	0.158
10	MSZoning_RH	0.103
8	GarageArea	0.097
1	LotArea	0.091

Alpha = 8

	Features	Coefficients
7	GrLivArea	0.358
11	MSZoning_RL	0.326
2	OverallQual	0.236
12	MSZoning_RM	0.224
6	TotalBsmtSF	0.191
9	MSZoning_FV	0.172
3	OverallCond	0.158
8	GarageArea	0.098
1	LotArea	0.091
10	MSZoning_RH	0.088

The Optimal value of alpha for **lasso** is 0.0001

Train data accuracy score for alpha = 0.0001 is 89.7%

Test data accuracy score for alpha = 0.0001 is 88.6%

The Change in accuracy score when alpha is 0.0002 are:

Train Score = 89.71%

Test Score = 88.6%

Alpha = 0.0001

	Parameters	Coefficients
11	MSZoning_RL	0.436
7	GrLivArea	0.362
12	MSZoning_RM	0.321
2	OverallQual	0.231
9	MSZoning_FV	0.229
6	TotalBsmtSF	0.195
3	OverallCond	0.158
10	MSZoning_RH	0.119
8	GarageArea	0.097
1	LotArea	0.091

Alpha = 0.0002

	Parameters	Coefficients
11	MSZoning_RL	0.443
7	GrLivArea	0.363
12	MSZoning_RM	0.327
9	MSZoning_FV	0.232
2	OverallQual	0.231
6	TotalBsmtSF	0.195
3	OverallCond	0.157
10	MSZoning_RH	0.121
8	GarageArea	0.097
1	LotArea	0.091

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: The test accuracy for both lasso and ridge models are almost similar. Lasso has small optimal alpha value of 0.0001 when compared to Ridge and Lasso also penalises any unnecessary variables. Thus, helping in reduction of variables. The lasso regression model will be a good model when compared to Ridge by slight margin because of lasso's small optimal alpha value to the final model and also helps in feature elimination. Thus, helps in building a robust model.

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?

Answer: The five most important variables now after building the lasso model by dropping the top 5 predictor variables are:

	Parameters	Coefficients
5	TotalBsmtSF	0.298
6	GarageArea	0.247
2	OverallCond	0.205
1	LotArea	0.204
8	Neighborhood_Crawfor	0.131

Question 4

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?

Answer:

We can make sure a model is robust and generalizable by following these

- By imposing penalty on the number of features. As the number of features increases i.e. the model becomes more and more complex.
- To reduce this complexity, we can use regularization that introduces a regularization term while calculating the total error made by the model. This concept of regularization is used by both ridge and lasso regression algorithms.
- There is also a concept of hyper parameter called lambda which can be tuned to control the complexity of a model in ridge and lasso regression. The implications of using regularization is that we may sometimes have to be satisfied with less accuracy than other multiple linear

regression models but we have an advantage of being less complex and more generalizable. The model accuracy is affected as regularization keeps a balance between complexity and generalizability thus prohibiting the model to have more than required columns.