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 to 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 optimum value of alpha for Ridge Regression: **6.0**

The optimum value of alpha for Lasso Regression: **0.0001**

If we double the value of alpha for both Ridge and Lasso: Ridge will lead to some of the coefficients very close to 0 and Lasso would cause some of the coefficients to be 0. Also, the test error might increase.

Most Important Predictors after the change has been implemented:

Ridge

	Variable	Coeff
3	OverallQual	0.063468
82	Neighborhood_NWAmes	0.059928
13	GrLivArea	0.052488
11	2ndFlrSF	0.049594
16	FullBath	0.041532

Lasso

	Variable	Coeff
13	GrLivArea	0.273423
3	OverallQual	0.125794
129	RoofMatl_WdShake	0.084417
82	Neighborhood_NWAmes	0.080138
23	GarageCars	0.057524

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:

Ridge regression keeps all the variables that are present in the data. Now, when the number of variables is very large and the data may have unrelated or noisy variables, it is not a good idea to keep those variables in the model. Lasso regression helps us by performing feature selection.

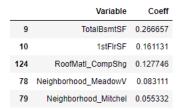In our case, lets see the R2 score for both train and test,

Comparison	Ridge	Lasso
R2 Score Train	0.8902962	0.8982089
R2 Score Test	0.8734002	0.880616

We can thereby say, Lasso is the best to apply in our case as it has less features and more R2 score without overfitting.

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 3: After creating another model after excluding the five most import predictors, below are the five most important predictors.

	Variable	Coeff
9	TotalBsmtSF	0.266657
10	1stFlrSF	0.161131
124	RoofMatl_CompShg	0.127746
78	Neighborhood_MeadowV	0.083111
79	Neighborhood_Mitchel	0.055332

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: A model is robust and generalisable when it not only gives good accuracy on Training data but also the test data which is unseen, robust model doesn't change prediction performance on new data. It doesn't by hearts the data points i.e., doesn't overfits rather able to genialize.

Accuracy doesn't drop much in test data vs training data; it means good model will perform equally good on test and train data.