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 the ridge is 3 and for the lasso, it is 0.0001

After I double the alpha values, the coefficient values got changed

Important predictors remain the same even after the change is implemented

Below are the most important predictors and their coefficient values for alphas(3,0.0001) and modified alphas(6,0.0002)

	Ridge Regression	Ridge Regression_2	Lasso Regression	Lasso Regression_2
OverallQual	0.100583	0.084664	0.158631	0.169342
GrLivArea	0.096835	0.084872	0.279545	0.285704
1stFlrSF	0.084712	0.074936	0.000000	0.000000
OverallCond	0.071265	0.060536	0.094467	0.088494
TotalBsmtSF	0.064805	0.058177	0.100774	0.097297
BsmtFinSF1	0.060402	0.053420	0.047846	0.048981
2ndFlrSF	0.048790	0.041791	0.000000	0.000000
Neighborhood_StoneBr	0.045725	0.039699	0.041744	0.033276
LotArea	0.045430	0.038603	0.051703	0.044649
GarageCars	0.040674	0.039035	0.052185	0.052530

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 prefer to use lasso model. For Lasso model

1. Test accuracy is improved slightly than ridge
2. Difference between test and train is less than ridge

	Metrics	Linear Regression	RFE Regression	Ridge Regression	Ridge Regression_2	Lasso Regression	Lasso Regression_2
0	R2_Score_Train	9.546069e-01	0.847304	0.945058	0.939629	0.942741	0.937661
1	R2_Score_Test	-7.073013e+08	0.869542	0.928888	0.927880	0.929958	0.932468
2	RSS_Train	7.488734e-01	2.519106	0.906409	0.995976	27863.856638	27584.162214
3	RSS_Test	5.156401e+09	0.951072	0.518427	0.525771	5338.135558	5254.621209
4	MSE_Train	2.918836e-02	0.053534	0.032112	0.033661	0.032782	0.034205
5	MSE_Test	3.693409e+03	0.050160	0.037034	0.037295	0.036754	0.036089

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?

After excluding 5 variables from data . For LASSO the important top 5 predictors are

1. 1stFlrSF
2. 2ndFlrSF
3. BsmtFinSF1
4. GarageArea
5. BsmtUnfSF__

	Features	coefficients
8	1stFlrSF	0.260614
9	2ndFlrSF	0.150491
5	BsmtFinSF1	0.108632
20	GarageArea	0.070398
7	BsmtUnfSF	0.063641

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 make the model robust and generalizable by keeping variance low by inducing little bias to the model. This can be controlled by using penalizing concepts Ridge and Lasso regression models. The test accuracy should not be much lower than the training score. Test accuracy should be close to the training accuracy. Overall, we should have optimal bias and variance.

The implications of this are taking the correct data that is within the expected range. Outliers have to be detected and should be removed if required. This prevents overfitting. Check whether the linear assumptions are satisfied or not. If not, try to transform data in an appropriate manner and then do analysis