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:

Ridge	100
Lasso	0.01

On doubling the alpha, the R2s and coefficient reduce further. Also some features become 0 using lasso.

Below become the top 5 predictors after the change

Ridge

Predictor	Coefficient
BsmtFullBath	0.138842
OverallCond	0.125666
LowQualFinSF	0.091336
2ndFlrSF	0.087570
Neighborhood_NridgHt	0.077062

Lasso

Predictor	Coefficient
BsmtFullBath	0.385049
OverallCond	0.211536
CentralAir	0.098347
YearRemodAdd	0.085330
BsmtFinSF2	0.083436

Whereas earlier the predictors were:

Ridge

Predictor	Coefficient
BsmtFullBath	0.165847
OverallCond	0.131659
LowQualFinSF	0.118057
2ndFlrSF	0.094847
CentralAir	0.086876

Lasso

Predictor	Coefficient

BsmtFullBath	0.407378
OverallCond	0.168601
YearRemodAdd	0.112075
CentralAir	0.108771
BsmtFinSF2	0.081902

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:

Considering alpha I would select Lasso. It also helps in feature selection by make making coefficients 0.

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:

Ridge

Predictor	Coefficient
BsmtHalfBath	0.168272
YearBuilt	0.142319
LowQualFinSF	0.117198
2ndFlrSF	0.099313
1stFlrSF	0.087850

Lasso

Predictor	Coefficient
BsmtHalfBath	0.407020
YearBuilt	0.186451
BsmtUnfSF	0.097607
1stFlrSF	0.096168
MasVnrArea	0.087828

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?

1. Eliminate some variables based on collinearity. There are some parameters which are related. So probably few predictors can be eliminated. Some analysis on non-linearity can help in performing data transformation.
2. There are many predictors with lot of outliers. Eliminating/treating outliers will give better results.