Question1

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?

Ans: The optimal value of alpha for ridge and lasso regression are,

Ridge	Lasso
10	0.001

The R2 scores are,

Ridge (10)		Lasso (0.001)	
Train	Test	Train	Test
0.908	0.873	0.912	0.878

The RMSE values are,

Ridge		Lasso	
Train	Test	Train	Test
0.302	0.364	0.295	0.357

The top features are,

Ridge	Lasso
YrSold	LotFrontage
LotFrontage	YrSold
LotArea	LotArea
MasVnrArea	MasVnrArea
BsmtFinSF1	BsmtFinSF1
BsmtFinSF2	BsmtFinSF2
MoSold	BsmtUnfSF
BsmtUnfSF	MoSold
TotalBsmtSF	TotalBsmtSF
1stFlrSF	1stFlrSF
2ndFlrSF	2ndFlrSF
GrLivArea	GrLivArea
BsmtFullBath	BsmtFullBath
FullBath	GarageFinish

If the alpha, is doubled, the R2 scores gets decreased negligibly.

Ridge (20)		Lasso (0.002)	
Train	Test	Train	Test
0.9	0.868	0.902	0.874

The RMSE values are increase slightly,

Ridge		Lasso	
Train	**Test**	**Train**	**Test**
0.315	0.371	0.312	0.363

The top features after doubling are as mentioned in the below table. They are almost the same, but the coefficients slightly varied giving different weightage to couple of independent variables.

Ridge	Lasso
YrSold	LotFrontage
LotFrontage	YrSold
LotArea	LotArea
MasVnrArea	MasVnrArea
BsmtFinSF1	MoSold
MoSold	BsmtFinSF1
BsmtFinSF2	BsmtFinSF2
BsmtUnfSF	BsmtUnfSF
TotalBsmtSF	TotalBsmtSF
1stFlrSF	1stFlrSF
2ndFlrSF	2ndFlrSF
GrLivArea	GrLivArea
BsmtFullBath	BsmtFullBath
FullBath	FullBath

Question2

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?

Ans: We will choose Lasso regression as it performed slightly better compared to ridge regression. The feature elimination kicked in for the lasso regression which can be observed with many independent variables having a coefficient of zero. We observed a better R2 score, RMSE and bias-variance trade off with Lasso regression.

Question3

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 in lasso model are

LotFrontage
YrSold
LotArea
MasVnrArea

BsmtFinSF1

If these variables are not available in the incoming data and we create a model excluding these, we end up with the following model stats.

R2 score almost remains the same

Before		After	
Train	**Test**	**Train**	**Test**
0.912	0.878	0.911	0.877

RSME almost remains the same

Before		After	
Train	**Test**	**Train**	**Test**
0.295	0.357	0.297	0.357

The new predictor variables are

BsmtFinSF2
MoSold
BsmtUnfSF
TotalBsmtSF
1stFlrSF

Question4

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?

A robust and generalisable model will have to satisfy these conditions

1. A good R2 Score. It should be in the range of 0.75 to 0.95.
2. Be simple. A simple model is usually generalisable. It should have a fine trade-off between bias and variance.
    a. P-value should be < 0.05. This indicates the variable is significant
    b. VIF should be < 5. This removes multi-collinearity
3. Have no outliers in the predictors.
4. Correlation to exist between predictor and target variable. Especially regression models
5. Robust error metrics. Low mean square error and root mean square errors.