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?

Ans:

The alpha value for Ridge Regression is 3 and for the Lasso Regression it is 0.001. After doubling the alpha values there is not much difference in r2 scores and slight difference in the coefficient values both in ridge and lasso regression.

The following are the top coefficient results before and after changes

Ridge regression:

Before:

	Ridge Coefficient
Neighborhood_Crawfor	0.677208
OverallQual	0.373262
Foundation_Slab	0.367079
Neighborhood_ClearCr	0.265230
Condition1_PosN	0.263547
Neighborhood_StoneBr	0.259757
Exterior1st_WdShing	0.227270
Foundation_PConc	0.219011
GarageArea	0.211640
Functional_Min2	0.210317
LotConfig_CulDSac	0.182492
Neighborhood_NridgHt	0.165728
TotalBsmtSF	0.155863
BsmtCond_Gd	0.154967
BsmtExposure_Gd	0.137094
OverallCond	0.104482
Neighborhood_SWISU	0.103609
SaleType_CWD	0.098623
Condition1_RRAn	0.083947
BsmtFinSF1	0.069810

After:

	Ridge Doubled Alpha Coefficient
Neighborhood_Crawfor	0.581877
OverallQual	0.382382
Foundation_Slab	0.276825
GarageArea	0.214223
Neighborhood_ClearCr	0.212332
Foundation_PConc	0.207092
Condition1_PosN	0.201313
Neighborhood_StoneBr	0.199539
Functional_Min2	0.188876
LotConfig_CulDSac	0.171273
Exterior1st_WdShing	0.158702
TotalBsmtSF	0.149968
BsmtCond_Gd	0.146179
Neighborhood_NridgHt	0.142002
BsmtExposure_Gd	0.126723
OverallCond	0.104747
Neighborhood_SWISU	0.078483
Condition1_RRAn	0.073962
BsmtFinSF1	0.071090

Lasso regression:

Before:

	Lasso Coefficient
Neighborhood_Crawfor	0.764159
Foundation_Slab	0.419823
OverallQual	0.372493
Neighborhood_ClearCr	0.280890
Condition1_PosN	0.279832
Neighborhood_StoneBr	0.278532
Exterior1st_WdShing	0.216466
GarageArea	0.212833
Foundation_PConc	0.212324
Functional_Min2	0.199030
LotConfig_CulDSac	0.175418
TotalBsmtSF	0.158677
Neighborhood_NridgHt	0.155202
BsmtCond_Gd	0.143245
BsmtExposure_Gd	0.125730
OverallCond	0.102046
Neighborhood_SWISU	0.083312
BsmtFinSF1	0.069096

After:

	Lasso Doubled Alpha Coefficient
Neighborhood_Crawfor	0.581877
OverallQual	0.382382
Foundation_Slab	0.276825
GarageArea	0.214223
Neighborhood_ClearCr	0.212332
Foundation_PConc	0.207092
Condition1_PosN	0.201313
Neighborhood_StoneBr	0.199539
Functional_Min2	0.188876
LotConfig_CulDSac	0.171273
Exterior1st_WdShing	0.158702
TotalBsmtSF	0.149968
BsmtCond_Gd	0.146179
Neighborhood_NridgHt	0.142002
BsmtExposure_Gd	0.126723
OverallCond	0.104747
Neighborhood_SWISU	0.078483
Condition1_RRAn	0.073962
BsmtFinSF1	0.071090

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?

Ans:

The optimal value of lambda in Ridge regression is 1 and in case of Lasso regression is 0.001

The Mean square error in case of Ridge is 0.1538 and in case of Lasso regression Mean square error is 0.1544. The MSE of both models is almost same.

Lasso model has more near zero coefficients than Ridge. So, I prefer Lasso over Ridge even though the validation scores are almost equal.

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?

Ans:

The following are the top five features of Lasso model.

1. Neighborhood_Crawfor
2. Foundation_Slab
3. OverallQual
4. Neighborhood_ClearCr
5. Condition1_PosN

After removing above features from dataset and rebuild the lasso model. Following are the top five features:

	Lasso Coefficient
SaleType_CWD	1.421923
Neighborhood_StoneBr	0.604184
Exterior1st_WdShing	0.439305
Neighborhood_NridgHt	0.407928
Foundation_PConc	0.353640

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?

Ans:

To make sure the model is robust and generalizable it should output and forecast consistently accurate results even if one or more variables or assumptions are drastically changed due to unforeseen circumstances. In terms of accuracy a robust and generalisable model will perform equally well on both training and test data. That means accuracy does not change much for training and test data.

Complex models have high accuracy but low bias, they cause overfitting. Less complex models have low variance and high bias(underfitting). Hence, we need to find an optimal value of bias and variance to get a more robust and generalizable model.