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 for alpha for Ridge is 3 and Lasso is 0.006

Doubling the value of alpha does not lead to a significant change in the accuracy but small change in the value of the features' coefficients. A new model was created and the following are the top most important predictors:

RIDGE:

	Ridge Co-Efficient
Total_sqr_footage	0.279384
TotRmsAbvGrd	0.166088
GarageCars	0.113928
GarageArea	0.113621
Fireplaces	0.096714
MasVnrArea	0.076461
SaleCondition_Partial	0.075530
LotArea	0.073585
Neighborhood_StoneBr	0.070150
Neighborhood_Crawfor	0.068255
LotFrontage	0.060788
CentralAir_Y	0.051740
RoofStyle_Mansard	0.047604
Neighborhood_Veenker	0.046093
SaleCondition_Alloca	0.045744
GarageQual_Gd	0.045742
BsmtCond_Gd	0.042199
BsmtCond_TA	0.037549
Neighborhood_ClearCr	0.037542
SaleCondition_Normal	0.036566

	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.217854
TotRmsAbvGrd	0.155639
GarageCars	0.118477
GarageArea	0.112434
Fireplaces	0.103577
MasVnrArea	0.075150
SaleCondition_Partial	0.071627
LotArea	0.067600
Neighborhood_Crawfor	0.064019
Neighborhood_StoneBr	0.063731
LotFrontage	0.053740
CentralAir_Y	0.050061
BsmtCond_Gd	0.041546
BsmtCond_TA	0.038227
BedroomAbvGr	0.036248
Neighborhood_Veenker	0.036017
GarageQual_Gd	0.035999
BsmtFinType1_GLQ	0.035654
RoofStyle_Mansard	0.033477
Neighborhood_ClearCr	0.033046

Lasso:

	Lasso Co-Efficient
Total_sqr_footage	0.481359
TotRmsAbvGrd	0.156685
GarageCars	0.130103
GarageArea	0.084203
Fireplaces	0.081504
SaleCondition_Partial	0.078460
Neighborhood_Crawfor	0.063131
CentralAir_Y	0.056859
MasVnrArea	0.052245
Neighborhood_StoneBr	0.051203
SaleCondition_Normal	0.041380
BsmtCond_Gd	0.038437
LotFrontage	0.033577
BsmtCond_TA	0.032668
LotArea	0.031003
Condition1_Norm	0.028776
GarageQual_Gd	0.026241
BsmtFinType1_GLQ	0.025442
BsmtFinType1_Unf	0.024867
LotConfig_CulDSac	0.022090

	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.475258
GarageCars	0.149792
TotRmsAbvGrd	0.149687
Fireplaces	0.088986
GarageArea	0.074501
SaleCondition_Partial	0.073487
CentralAir_Y	0.059248
Neighborhood_Crawfor	0.055009
MasVnrArea	0.037555
SaleCondition_Normal	0.035610
Neighborhood_StoneBr	0.032437
BsmtCond_Gd	0.031741
BsmtCond_TA	0.029571
BsmtFinType1_GLQ	0.028402
Condition1_Norm	0.024404
BsmtFinType1_Unf	0.022780
LotConfig_CulDSac	0.019036
KitchenQual_Gd	0.017550
PavedDrive_Y	0.016643
BsmtExposure_Gd	0.012937

Since alpha values are low, there is not any significant change to the model.

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?

The optimal value are:

Ridge – 3

Lasso – 0.006

The Mean Squares Error (MSE) for the models are:

Ridge – 0.0051

Lasso – 0.0047

The MSE for models are almost similar, but since Lasso helps in feature reduction and we have over 200 features for the model, it is better to use Lasso regression for our business case.

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 removing the top five most important predictors, a Lasso Model was built and the following are the most important predictors:

1. Lot size in square feet
2. Masonry veneer area in square feet
3. Bedroom: Bedrooms above grade
4. Linear feet of street connected to property
5. Neighborhood

```
The R2 Score of the model on the test dataset is 0.703655275236539
The MSE of the model on the test dataset is 0.009513709753541263
```

```
The most important predictor variables are as follows:
```

	Lasso Co-Efficient
LotArea	0.297323
MasVnrArea	0.249949
BedroomAbvGr	0.222297
LotFrontage	0.214203
Neighborhood_StoneBr	0.132450

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.

Occam's Razor states that:

Given that two models show a similar performance, the model that makes fewer errors on the test data should be picked, thus keeping it simple and robust.

1. Simpler models are more generic in nature and are easily applicable.
2. Simpler models require fewer training data to be effective.

To make sure that a model is robust and generalizable, we can use the following methods:

1. Removal of Outliers
2. Trimming down features and using appropriate features.
3. Regularization techniques like L1 (lasso) and L2 (Ridge) to penalize the cost function.

Bias Variance Trade Off

Although it is good to have a high accuracy, a model which scores very high on accuracy on training data might not give appropriate results on the test data.

Trying for a high accuracy might lead to overfitting of the model, which will cause the model to be very complex and thus likely to give errors on test data. Thus, it is feasible to go for a lower accuracy value on the training dataset and make the model more flexible.

This compromise on the accuracy is call the Bias Variance Trade off as shown in the picture below.