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:
After performing cross validation with 5 folds , the optimal values for ridge and Lasso are 1.00 and 0.0003 respectively , according to Assignment problem.
If alpha values are doubled there would a more penalty on the models and models will choose less features.
The important predictors chosen by RFE and Lasso model are

1) Linear feet of street connected to property
2) General shape of property
3) Rates the overall condition of the house
4) Original construction date
5) Remodel date

```
In [90]: ls = Lasso(alpha = 0.001, max_iter = 161, selection = 'cyclic', tol = 0.002, random_state = 101)
         rfecv = RFECV(estimator=ls, n_jobs = -1, step=1, scoring = 'neg_mean_squared_error' ,cv=5)
         rfecv.fit(X_train, y_train)

         select_features_rfecv = rfecv.get_support()
         RFEcv = cols[select_features_rfecv]
         print('{:d} Features Select by RFEcv:\n{:}'.format(rfecv.n_features_, RFEcv.values))

         72 Features Select by RFEcv:
         ['LotFrontage' 'LotShape' 'OverallCond' 'YearBuilt' 'YearRemodAdd'
          'MasVnrArea' 'ExterQual' 'BsmtQual' 'BsmtExposure' 'BsmtFinType1'
          'HeatingQC' 'CentralAir' 'FstFlrSF' 'BsmtFullBath' 'FullBath' 'HalfBath'
          'KitchenAbvGr' 'KitchenQual' 'TotRmsAbvGrd' 'Functional' 'Fireplaces'
          'GarageYrBlt' 'GarageFinish' 'PavedDrive' 'OpenPorchSF' 'EnclosedPorch'
          'MiscVal' 'YrSold' 'MSSubClass_120' 'MSSubClass_30' 'MSSubClass_50'
          'MSSubClass_60' 'MSSubClass_70' 'MSSubClass_75' 'MSZoning_FV'
```

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:
I will apply Lasso regression for this as Its gives 89.65 on Train set and also 88.121 on Test , this shows that model is not overfitting and generalizing well on both train and test set and its better than Ridge if we compare R2 score.

Lasso CV Result

	Scorer	Index	BestScore	BestScoreStd	MeanScore	MeanScoreStd
0	MEA	4	0.094048	0.003201	0.181185	0.008902
0	R2	28	89.653379	1.151589	59.257143	1.559886
0	RMSE	28	0.132608	0.024693	0.281730	0.082148

Lasso Prediction on Test Set

```
In [97]: r2_score(y_test,lasso.predict(X_test))
         Select 72 features
Out[97]: 0.8812140391528286
```

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:

After dropping Five most important predictors , the model has estimated below as important now.

1) Masonry veneer area in square feet
2) ExterQual - Evaluates the quality of the material on the exterior
3) ExterCond - Evaluates the present condition of the material on the exterior
4) the height of the basement
5) the general condition of the basement

```
3]: ls1 = Lasso(alpha = 0.001, max_iter = 161, selection = 'cyclic', tol = 0.002, random_state = 101)
    rfecv2 = RFECV(estimator=ls1, n_jobs = -1, step=1, scoring = 'neg_mean_squared_error' ,cv=5)
    rfecv2.fit(X_train1, y_train)

    select_features_rfecv_n = rfecv2.get_support()
    RFEcv = cols[select_features_rfecv_n]
    print('{:d} Features Select by RFEcv:\n{:}'.format(rfecv2.n_features_, RFEcv.values))

    64 Features Select by RFEcv:
    ['MasVnrArea' 'ExterQual' 'ExterCond' 'BsmtQual' 'BsmtCond' 'BsmtExposure'
     'BsmtFinType1' 'BsmtFinType2' 'HeatingQC' 'CentralAir' 'FstFlrSF'
     'BsmtFullBath' 'FullBath' 'HalfBath' 'KitchenAbvGr' 'KitchenQual'
     'TotRmsAbvGrd' 'Functional' 'Fireplaces' 'GarageFinish' 'PavedDrive'
     'OpenPorchSF' 'EnclosedPorch' 'YrSold' 'MSSubClass_120' 'MSSubClass_30'
     'MSSubClass_50' 'MSSubClass_60' 'MSSubClass_70' 'MSSubClass_75'
     'MSZoning_FV' 'MSZoning_RM' 'LandContour_Lvl' 'LotConfig_CulDSac'
     'LotConfig_FR2' 'Neighborhood_BrkSide' 'Neighborhood_ClearCr'
```

Lasso and RFE for feature selection

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?

ANSWER:
We can ensure that the model is robust and generalisable is by removing overfitting. Overfitting is a condition which may occur if we have too many features and the data is fitting good while training , but during testing it does not generalize well as on test set and it may have memorised the dataset. To handle such issue we need to apply regularisation to our model as it penalises us if we have more features in our dataset and thus helps reduce overfitting.
The model may not be able to generalise well if there is issue of multicollinearity or if model Is not a proper fit to the data.