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?

Optimal Alpha value :

Lasso : 0.2 with High R2 Train 80% and R2Test 72%

Ridge : 0.2 with High R2 Train 91% and R2Test 87%

If we double the Alpha Model will underfit in both cases.

Most important feature as per the model is "**OverallQual**"

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?

Lasso is the one I feel to apply, Lasso can be used to identify the important features and not important features as well. All Zero marked Features are not important for this Model.

```
lr.coef_

Training R2: 0.8584758800065541
Testing R2: 0.7839201869657713

ut[248]: array([-6.77770732e-04,  0.00000000e+00,  2.17664934e-06, -0.00000000e+00,
                -0.00000000e+00,  0.00000000e+00,  0.00000000e+00, -0.00000000e+00,
                -0.00000000e+00, -0.00000000e+00,  2.79667619e-04, -0.00000000e+00,
                -0.00000000e+00,  0.00000000e+00,  0.00000000e+00,  6.03054178e-02,
                 2.70917109e-02,  3.47248502e-03,  2.15982041e-03,  0.00000000e+00,
                -0.00000000e+00,  0.00000000e+00, -0.00000000e+00, -0.00000000e+00,
                 1.56769075e-06,  0.00000000e+00,  0.00000000e+00,  3.87044309e-03,
                -0.00000000e+00,  0.00000000e+00,  0.00000000e+00, -4.94389951e-03,
                 1.13036399e-04,  0.00000000e+00,  8.13171920e-05,  5.00738309e-05,
                 3.73168044e-05, -0.00000000e+00, -0.00000000e+00, -0.00000000e+00,
                 0.00000000e+00,  2.10107955e-04,  2.47445238e-04,  2.53790783e-04,
                 3.84570660e-05,  0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
                 0.00000000e+00,  0.00000000e+00, -0.00000000e+00,  0.00000000e+00,
                 0.00000000e+00,  0.00000000e+00,  0.00000000e+00, -1.22947026e-02,
                -1.60083092e-03, -7.15030148e-04, -0.00000000e+00,  0.00000000e+00,
                 2.39458958e-04, -0.00000000e+00,  0.00000000e+00, -0.00000000e+00,
                 1.21703626e-04, -7.32864735e-05,  2.07856519e-04,  1.96402817e-04,
                 3.03431553e-04, -1.74571263e-03,  0.00000000e+00,  0.00000000e+00,
                 0.00000000e+00, -2.47588251e-06,  0.00000000e+00, -0.00000000e+00,
                 0.00000000e+00, -0.00000000e+00])
```

Using Lasso its easy to find out important features as

```
In [254]: X_cols = house.drop(["Id", "SalePrice", "TransformedPrice"], axis=1)
          print(pd.DataFrame({'feature':list(X_cols.columns),
                              'coef'    :abs(lr.coef_)}).sort_values('coef',ascending=False)[:10].sort_index())
```

```
         feature      coef
15   OverallQual  0.060305
16   OverallCond  0.027092
17     YearBuilt  0.003472
18  YearRemodAdd  0.002160
27    Foundation  0.003870
31  BsmtFinType1  0.004944
55   FireplaceQu  0.012295
56    GarageType  0.001601
57   GarageYrBlt  0.000715
69      PoolArea  0.001746
```

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?

```
In [264]: # Lets find the five most important features now
          X_cols = house.drop(["Id", "SalePrice", "TransformedPrice", "OverallQual", "OverallCond", "Foundation", "
          print(pd.DataFrame({'feature':list(X_cols.columns),
                              'coef'    :abs(lr.coef_)}).sort_values('coef',ascending=False)[:5].sort_index())
```

```
         feature      coef
15     YearBuilt  0.003585
16  YearRemodAdd  0.003653
49    Functional  0.013723
50    Fireplaces  0.007006
51    GarageType  0.004092
```

So five most important features after dropping earlier most important features are

- YearBuilt
- YearRemodAdd
- Functional
- Fireplaces
- GarageType

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?

Ideal goal is to have a best fit model which is not overfitting or underfitting.
In terms of regularization , zero alpha is always an unregularised model and high alpha is underfitting. Target to get a model with lower total error that mean low bias and low variance. Robust and generalized, its model is always less complex so it's not overfitting .