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 optimal value of alpha for lasso is 1.0023052380778996 and ridge is 0.5473213937810939.

With the above alpha we get the following output:

| | Metric | Linear Regression | Ridge Regression | Lasso Regression |
|---|---|---|---|---|
| 0 | R2 Score (Train) | 9.357343e-01 | 9.271426e-01 | 9.357187e-01 |
| 1 | R2 Score (Test) | 7.484327e-01 | 8.338757e-01 | 7.211874e-01 |
| 2 | RSS (Train) | 4.100599e+11 | 4.648809e+11 | 4.101595e+11 |
| 3 | RSS (Test) | 7.090960e+11 | 4.682566e+11 | 7.858925e+11 |
| 4 | MSE (Train) | 2.004060e+04 | 2.133821e+04 | 2.004304e+04 |
| 5 | MSE (Test) | 4.023606e+04 | 3.269677e+04 | 4.235888e+04 |

So the following are our parameters with coefficients.

```
                          Ridge          Lasso
RoofMatl_WdShngl      84794.635768   77288.919986
OverallQual_10        82293.592349  112845.866055
Neighborhood_NoRidge  50952.182357   47292.339759
FullBath_3            45065.809628   40054.557886
GrLivArea             43081.213723   44661.059760
                          Ridge          Lasso
PoolQC_Gd          -188351.889525 -499887.605520
Condition2_PosN    -178410.067029 -314533.234742
PoolQC_Not present  -49086.392892 -130050.838092
Heating_OthW        -34719.500130  -55821.621266
MSSubClass_75       -30853.222105  -63421.946976
```

**So from the above we have sorted out that out 315 parameters which includes dummy encoding parameters too, a range of optimal parameters above 100-150 which we can confirm with our grid search CV gives us all sorts of output, but using ridge regression & lasso we involve Standardisation and we have compared the metrics:**

- The 5 most positive coefficients which we got are:
  - Roof made with wood shingles.
  - Overall material and finish of the house comes to be best rated: 10.
  - Neighborhood is Northridge.
  - Full bathrooms above grade 3.
  - Above grade (ground) living area square feet.
- The 5 factors which negatively affect the sale price:
  - Pool quality is just good.
  - Proximity to various conditions: Adjacent to postive off-site feature
  - Hot water or steam heat other than gas
  - dwelling involved in the sale: 2-1/2 STORY ALL AGES

If we double the alpha, we get the following:

| | Metric | Linear Regression | Ridge Regression | Lasso Regression |
|---|---|---|---|---|
| 0 | R2 Score (Train) | 9.357343e-01 | 9.210625e-01 | 9.356762e-01 |
| 1 | R2 Score (Test) | 7.484327e-01 | 8.489141e-01 | 7.238685e-01 |
| 2 | RSS (Train) | 4.100599e+11 | 5.036763e+11 | 4.104305e+11 |
| 3 | RSS (Test) | 7.090960e+11 | 4.258678e+11 | 7.783353e+11 |
| 4 | MSE (Train) | 2.004060e+04 | 2.221073e+04 | 2.004966e+04 |
| 5 | MSE (Test) | 4.023606e+04 | 3.118174e+04 | 4.215473e+04 |

```
                            Ridge          Lasso
OverallQual_10        73033.266650  117829.744202
RoofMatl_WdShngl      69342.088003   76797.979945
Neighborhood_NoRidge  51066.740442   47333.423890
FullBath_3            45056.241081   40084.925705
GrLivArea             41718.269526   44314.665281
                             Ridge          Lasso
Condition2_PosN    -124806.948360 -313452.038516
PoolQC_Gd          -121251.811118 -494396.342889
PoolQC_Not present  -31647.083471 -127600.425523
BsmtQual_TA         -28193.440570  -25993.754245
BsmtQual_Fa         -25519.155461  -24172.146769
```

Generally there is not much change in the predictors, there has been some change in the negative relation pre dictors but there has been a change in the top predictor both on the positive coefficient and negative coefficie nt.

Earlier it was related to the roof material but now it is related to the overall quality rated 10 and the same way the most impact worthy negative predictor was Pool quality: good, now it is all about Condition 2: Proximity to various conditions: Near positive off-site feature--park, greenbelt, etc.

Also we see the metrics shift a bit down and up but not so much of vivid changes.

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: I am going to apply the Ridge regression as it doesn't straightaway reduces the parameters to 0 but try to minimize the coefficients down which will not abolish the impact of the parameters on the target variable but will reduce the noise/variance by introducing the bias, the penalty term of the Ridge is far better than that of l asso as It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squa red coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a re sult, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regr ession.
Although Lasso has an advantage of auto removal of less impacting betas, in our model we can see that Ridge has worked better than the Lasso which even fell short in front of regular RFE linear regression.

Also, Lasso tends to do well if there are a small number of significant parameters and the others are close to ze ro (ergo: when only a few predictors actually influence the response).

Ridge works well if there are many large parameters of about the same value (ergo: when most predictors imp act the response).

As we have so many parameters around 315, ridge actually gives out better results if we compare the metrics:

| | Metric | Linear Regression | Ridge Regression | Lasso Regression |
|---|---|---|---|---|
| **0** | R2 Score (Train) | 9.357343e-01 | 9.271426e-01 | 9.357187e-01 |
| **1** | R2 Score (Test) | 7.484327e-01 | 8.338757e-01 | 7.211874e-01 |
| **2** | RSS (Train) | 4.100599e+11 | 4.648809e+11 | 4.101595e+11 |
| **3** | RSS (Test) | 7.090960e+11 | 4.682566e+11 | 7.858925e+11 |
| **4** | MSE (Train) | 2.004060e+04 | 2.133821e+04 | 2.004304e+04 |
| **5** | MSE (Test) | 4.023606e+04 | 3.269677e+04 | 4.235888e+04 |

Q3)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: As per the below list, we have sorted the 10 most impactful predictors from the lasso:

```
                            Ridge          Lasso
OverallQual_10        82293.592349  112845.866055
RoofMatl_WdShngl      84794.635768   77288.919986
TotRmsAbvGrd_12       10350.297361   48485.983163
HouseStyle_2.5Unf     12335.570785   47366.016619
Neighborhood_NoRidge  50952.182357   47292.339759
                             Ridge          Lasso
PoolQC_Gd          -188351.889525 -499887.605520
Condition2_PosN    -178410.067029 -314533.234742
PoolQC_Not present  -49086.392892 -130050.838092
PoolQC_Fa           -14035.739970 -101044.650389
Condition2 RRAe     -13580.413319  -72478.895096
```

Out of which we can take 5 which can be like this:

Positively impacting:
==========
OverallQual_10 : Overall material and finish of the house comes to be best rated: 10.
RoofMatl_WdShngl : Roof made with wood shingles.
TotRmsAbvGrd_12 : Total rooms above grade (does not include bathrooms)

Negatively impacting:
==========
PoolQC_Gd : Pool quality is just good.
Condition2_PosN : Proximity to various conditions: Near positive off-site feature--park, greenbelt, etc.

After removing these columns from the predictor list, we get the following predictors:

Positively impacting:
==========
Neighborhood_NoRidge : Neighborhood is Northridge.
GrLivArea : Above grade (ground) living area square feet.
FullBath_3 : 3 full bathrooms above grade.

Negatively impacting:
==========
BsmtCond_Po : the general condition of the basement is Poor - Severe cracking, settling, or wetness.
OverallQual_2 : the overall material and finish of the house is poor.

Q) 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: Through standardisation we try to decrease the variance and then add a bit bias to make the model less o verfitted and through that it will fall for less noise but more general patterns which it should learn from the tra in data which in fact make the model generalizable too.
the test accuracy should not be too less than the training score. The model should be working for the data whi ch is unseen more like the test data.
Too much weightage should not be given to the outliers so that the accuracy predicted by the model is high, in short we should not try to learn the noise as mentioned before.
The outlier analysis needs to be done like we have done capping, it can be done via IQR or 3 sigma technique t o reduce outlier consideration.
If the model is not robust , it cannot be automated for predictive analysis of unknown data.