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 of alpha

1. Ridge : 'alpha': 1.0

2. Lasso: 'alpha': 20

If we double the Ridge and Lasso alpha it would then be Alpha =2 for Ridge and alpha= 40 for lasso

	Beta	Ridge	Lasso	new_rigdge	new_lasso
0	LotArea	50778.226674	53042.651179	47866.516319	49531.307458
1	OverallQual	81428.952314	82043.345395	80748.077435	84392.851318
2	OverallCond	42290.263972	44427.309501	39131.148124	42626.964127
3	YearBuilt	-41665.704991	-43355.314126	-39896.903699	-41905.295263
4	BsmtFinSF1	32403.084458	30529.761295	33774.254018	30983.988949
5	TotalBsmtSF	85959.968806	92539.056781	79497.796218	88211.689456
6	1stFlrSF	78784.597413	0.000000	76868.753555	0.000000
7	2ndFlrSF	49898.893671	6534.020721	47646.646084	4249.632396
8	GrLivArea	99065.127727	189687.968895	96118.509358	189796.706554
9	BedroomAbvGr	-17451.275076	-23137.371375	-11107.580777	-18672.697354
10	GarageArea	47967.518745	46023.903942	49048.614330	45560.160681
11	Neighborhood_StoneBr	29571.363589	29723.372474	28434.905805	28691.152018
12	Condition1_RRAe	-27330.602207	-29612.621699	-22690.618930	-24704.750420
13	Condition2_PosN	-17679.694750	-20035.828382	-10480.143125	-0.000000
14	RoofStyle_Gable	26974.181074	39392.414922	16605.224197	20215.937520
15	RoofStyle_Gambrel	41944.435938	59143.600787	29009.834440	36071.717939
16	RoofStyle_Hip	28074.090663	39762.637383	18660.691609	20781.315690
17	RoofStyle_Mansard	10992.077652	20337.645499	3082.138416	0.000000
18	RoofStyle_Shed	17666.772629	27245.917428	9479.011642	0.000000
19	Exterior1st_Stone	-22201.731493	-29332.217915	-14034.099365	-11259.755619
20	ExterQual_Fa	-33302.937250	-37545.696348	-27106.885310	-30575.707099
21	ExterQual_Gd	-42334.277206	-44392.334431	-39419.377808	-42268.764865
22	ExterQual_TA	-55485.663628	-56383.909207	-53943.988025	-54701.152976
23	Foundation_Slab	15184.776710	17963.985955	10912.526769	13847.889912

The beta values of the predictor variables shows definite change in doubling alpha values. Doubling of alphas has resulted in decreased beta values for the predictor parameters.

```
623001948.718489
```

Out[2354]:

	Metrics	Linear Regression	LR with RFE	Ridge	Lasso	Ridge_new	Lasso_new
0	R2_train	9.497507e-01	9.022725e-01	8.862064e-01	8.895629e-01	8.811263e-01	8.862091e-01
1	R2_test	-3.373567e+19	8.788528e-01	8.862064e-01	8.717670e-01	8.811263e-01	8.733231e-01
2	MSE_train	2.929137e+08	5.696733e+08	5.816344e+08	5.747401e+08	5.940347e+08	5.870210e+08
3	MSE_test	1.868178e+29	6.708758e+08	6.280170e+08	6.409570e+08	6.225577e+08	6.230619e+08

With regards to R2_score we can see a decrease in R_score of for both ridge and Lasso regression with increase in alpha.

The 10 most important variables arranged in the order of their absolute beta values

In [2751]:

```
# The most important 10 predictor variables are
sorted_df_lasso[:10]['Beta']
```

Out[2751]:

```
8             GrLivArea
5           TotalBsmtSF
1           OverallQual
22         ExterQual_TA
0               LotArea
10           GarageArea
2           OverallCond
21         ExterQual_Gd
3             YearBuilt
15     RoofStyle_Gambrel
Name: Beta, dtype: object
```

The values are arranged based on their absolute values of betas so as to understand and arrange in order based on their influence on dependent variable.

```
GrLivArea: Above grade (ground) living area square feet
TotalBsmtSF: Total square feet of basement area
OverallQual: Rates the overall material and finish of the house
ExterQual_TA:ExterQual: Evaluates the quality of the material on
the exterior, TA stands for average
LotArea: Lot size in square feet
```

```
GarageArea: Size of garage in square feet
OverallCond: Rates the overall condition of the house
ExterQual_Gd: Evaluates the quality of the material on the
exterior, Gd stands for good
YearBuilt: Original construction date
RoofStyle_Gambrel: Type of roof,Gambrel Gabrel (Barn)
```

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?

Out of Ridge and Lasso, although both the systems has given a very good R2_score. The lasso we a better score even when one of the variables beta value is zero. So we achieve a more generalised model with lesser predictor variables which further reduces the complexity of the model. So we go with Lasso regression.

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?

The predictor variables arranged in based on the absolute values of their betas and the most influential five values based on their beta values are as shown below

Out[2954]:

	Beta	Ridge	Lasso
8	GrLivArea	99065.127727	189687.968895
5	TotalBsmtSF	85959.968806	92539.056781
1	OverallQual	81428.952314	82043.345395
15	RoofStyle_Gambrel	41944.435938	59143.600787
22	ExterQual_TA	-55485.663628	-56383.909207

These are removed from the data frame and then lasso regression is performed on remaining variables.
The values arranged in order of their beta values are as shown below

	Beta	New_lasso
4	1stFlrSF	293399.248208
5	2ndFlrSF	122727.728734
15	Exterior1st_Stone	-71220.367201
2	YearBuilt	-66177.396212
20	BsmtQual_Fa	-64345.280340
22	BsmtQual_TA	-61972.014123
7	GarageArea	61947.351157
21	BsmtQual_Gd	-57214.013347
24	SaleType_Con	55882.520481
1	OverallCond	54637.899597
0	LotArea	42973.703283
8	Neighborhood_StoneBr	36462.976235
6	BedroomAbvGr	-35558.644106
9	Condition1_RRAe	-34104.420395
19	Foundation_Wood	-31771.381769
3	BsmtFinSF1	28896.815773
23	SaleType_CWD	22432.143325
12	RoofStyle_Hip	22292.575084
11	RoofStyle_Gable	17958.699619
18	Foundation_Slab	-14563.797146
14	RoofStyle_Shed	11885.066857
17	ExterQual_Gd	9698.373656
16	ExterQual_Fa	2970.370139
13	RoofStyle_Mansard	-1012.277693
10	Condition2_PosN	0.000000

Out this first 10 influential values are taken out and are shown as below.

```
In [2959]:
# The most important 10 predictor variables now are
sorted_df_lasso_n[:10]['Beta']

Out[2959]:
4              1stFlrSF
5              2ndFlrSF
15    Exterior1st_Stone
2             YearBuilt
20          BsmtQual_Fa
22          BsmtQual_TA
7            GarageArea
21          BsmtQual_Gd
24         SaleType_Con
1           OverallCond
Name: Beta, dtype: object
```

They are
1.1stFlrSF: First Floor square feet
2. 2ndFlrSF: Second floor square feet
3.Exterior1st_stone: Exterior covering on house, stone
4.YearBuilt: Original construction date
5.BsmtQual_Fa: Evaluates the height of the basement,Fa Fair (70-79 inches)
6.BsmtQual_TA: Evaluates the height of the basement,TA Typical (80-89 inches)
7. GarageArea: Size of garage in square feet
8. BsmtQual_Gd: Evaluates the height of the basement
9.SaleType_Con: Type of sale, con:Con Contract 15% Down payment regular terms
10.OverallCond: Rates the overall condition of the house

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?

The best way to make a model robust and generalisable are is to avoid overfitting and to remove the influence of noise. The outliers may cause a heavy price in the accuracy of the model. It's also a challenge to find the most influential variables if data is subdued by the outliers.

The outliers must be removed so as to lessen their influence on the target variable. The overall accuracy of model will be effected when the outliers are prominent as it can lead to anomaly in computation of coefficients.