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 1:

Optimal value of alpha for ridge: 0.0001
Optimal value of alpha for lasso: 0.001
These optimal values are achieved when a table of split train and split test r2 score is printed for various values of alpha. Out of them the best or optimal score is obtained for the highest mean train score and mean test score.

When we double the values of alpha for ridge:
The value of coefficients get lesser tending towards zero but not exactly zero.
When we double the values of alpha for lasso:
The less important features get eliminated by turning their coefficient to be exactly 0.

The 10 most important feature after the change is implemented for ridge are:
GrLivArea, OverallQual, GarageCars, BsmtQual_5, Neighborhood_NoRidge, KitchenQual_5, HouseStyle_1Story, Foundation_PConc, BsmtExposure_4, BsmtFinType1_6 .

```
ridge_coef.sort_values(by='Coef',ascending=False)
```

	Feaure	Coef
4	GrLivArea	0.378458
1	OverallQual	0.263569
6	GarageCars	0.164892
12	BsmtQual_5	0.134350
7	Neighborhood_NoRidge	0.114845
17	KitchenQual_5	0.108814
8	HouseStyle_1Story	0.108222
10	Foundation_PConc	0.096011
13	BsmtExposure_4	0.090241
16	BsmtFinType1_6	0.086943
9	Foundation_CBlock	0.077984
2	OverallCond	0.068882
18	GarageType_NA	0.052158
14	BsmtFinType1_4	0.045094
15	BsmtFinType1_5	0.044163
0	const	0.000000
3	LowQualFinSF	-0.036250
11	BsmtQual_3	-0.041169

The 10 most important feature after the change is implemented for lasso are:
GrLivArea, OverallQual, 2ndFlrSF, HouseStyle_1Story, Neighborhood_NoRidge, GarageCars, Neighborhood_NridgHt, BsmtQual_5, Neighborhood_Somerst, BsmtExposure_4

```
lasso_coef.sort_values(by='Coef',ascending=False).head(25)
```

	Feaure	Coef
13	GrLivArea	0.251239
2	OverallQual	0.162522
11	2ndFlrSF	0.147913
84	HouseStyle_1Story	0.147000
73	Neighborhood_NoRidge	0.135977
23	GarageCars	0.121451
74	Neighborhood_NridgHt	0.121207
139	BsmtQual_5	0.117538
79	Neighborhood_Somerst	0.095601
143	BsmtExposure_4	0.095090
156	KitchenQual_5	0.083319
64	Neighborhood_Crawfor	0.066293
3	OverallCond	0.066023
14	BsmtFullBath	0.059174
1	LotArea	0.057627
16	FullBath	0.055666
80	Neighborhood_StoneBr	0.054626

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 2 :

Ridge

Optimal Alpha: 0.0001

Train score: `0.8367045391243007`

Test score: `0.8322687243377265`

```
No. of columns passed (that get selected after rfe):
19
```

```
model_cv.best_params_
```

```
{'alpha': 0.0001}
```

```
ridge = Ridge(alpha = 0.0001)
ridge.fit(X_train_rfe,y_train)

y_pred_train = ridge.predict(X_train_rfe)
print(r2_score(y_train,y_pred_train))

y_pred_test = ridge.predict(X_test_new)
print(r2_score(y_test,y_pred_test))
```

```
0.8367045391243007
0.8322687243377265
```

Lasso

Optimal Alpha: 0.001

Train score: 0.8885357943016787

Test score: 0.8595570736810942

No. of columns passed (all X_train features): 171

```
print(lasso_cv.best_params_)
print(lasso_cv.best_score_)
```

```
{'alpha': 0.001}
-0.25510590668625366
```

```
lasso = Lasso(alpha=0.001)
lasso.fit(X_train,y_train)

y_train_pred = lasso.predict(X_train)
y_test_pred = lasso.predict(X_test)

print(r2_score(y_true=y_train,y_pred=y_train_pred))
print(r2_score(y_true=y_test,y_pred=y_test_pred))
```

```
0.8885357943016787
0.8595570736810942
```

As we got nearly good score for both ridge and lasso, we will go with lasso as we don't need to do feature selection manually as we have to do for ridge. And it takes care of feature selection by itself and no need to worry if there are large dimensions. Also for us, the train and test score for lasso outperforms ridge.

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 3:

For Ridge next 5 significant predictor variables are:

17	KitchenQual_5	0.108814
8	HouseStyle_1Story	0.108222
10	Foundation_PConc	0.096011
13	BsmtExposure_4	0.090241
16	BsmtFinType1_6	0.086943

For Lasso next 5 significant predictor variables are:

2	OverallQual	0.155742
10	1stFlrSF	0.145163
73	Neighborhood_NoRidge	0.143438
152	HeatingQC_4	0.142330
74	Neighborhood_NridgHt	0.137903

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 4:

1. A model that is resistant to outliers like tree based rather regression models

and doing a non parametric test than a parametric test does help making a model robust

2. A more robust error metric like mean absolute difference, negative mean absolute error , Huber Loss etc reduces the effect of outliers

3. Using median as a measure of central tendency instead of mean as median performs better when there are outliers

4. Removing missing values/outliers or imputing them using best measures

5. If our data is skewed towards left or right we need to transform the data This will help obtain a better normal bell curve and hence improve accuracy.