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?

Ans: -

What is the optimal value of alpha for ridge and lasso regression?

Ridge	Lasso
0.005	16.372746

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

```
# Before change
df_final
```

	Parameter	Ridge	Lasso
0	Train_R2	0.84906	0.844945
1	Test_R2	0.84173	0.840276
2	Alpha	0.00500	16.372746

```
# After change
df_final_change
```

	Parameter	Ridge	Lasso
0	Train_R2	0.845818	0.836047
1	Test_R2	0.843546	0.832647
2	Alpha	0.100000	32.745492

R2 value for ridge regression have increased and R2 value for lasso regression have reduced. But the change is R2 is not significant. Moreover top 5 variable remain the same but after that variable are changing as you can see in the below picture.

If we are increasing the alpha value more features will have coefficient of 0 and in return dependency on many variables will be reduced.

```
print(len(final_lasso_variables))
print(len(final_lasso_variables_change))
print(len(final_ridge_variables))
print(len(final_ridge_variables_change))
```

```
79
56
101
101
```

What will be the most important predictor variables after the change is implemented?

```
final_lasso_variables.head(20)
```

	Variable_Name	Coef	Abs_Coef
0	GrLivArea	282326.735800	282326.735800
1	OverallQual	155680.715684	155680.715684
2	RoofMatl_WdShngl	105220.020957	105220.020957
3	Neighborhood_NridgHt	68035.511961	68035.511961
4	Heating_OthW	-66122.237825	66122.237825
5	Functional_Sev	-62274.959880	62274.959880
6	Neighborhood_StoneBr	56939.762751	56939.762751
7	Exterior2nd_ImStucc	39006.338067	39006.338067
8	Neighborhood_NoRidge	38249.509442	38249.509442
9	BldgType_Twnhs	-33406.240230	33406.240230
10	Exterior2nd_Stucco	-23765.920575	23765.920575
11	BsmtExposure_Gd	23712.262644	23712.262644
12	BldgType_TwnhsE	-23330.268569	23330.268569
13	Neighborhood_SWISU	-21655.253710	21655.253710
14	MSZoning_FV	20345.663245	20345.663245

```
final_lasso_variables_change
```

	Variable_Name	Coef	Abs_Coef
0	GrLivArea	268604.605706	268604.605706
1	OverallQual	171277.036647	171277.036647
2	RoofMatl_WdShngl	97355.498567	97355.498567
3	Neighborhood_NridgHt	63272.495078	63272.495078
4	Heating_OthW	-57546.205632	57546.205632
5	Neighborhood_StoneBr	50500.711665	50500.711665
6	Functional_Sev	-48647.970198	48647.970198
7	Exterior2nd_ImStucc	34576.336650	34576.336650
8	Neighborhood_NoRidge	34259.245883	34259.245883
9	BsmtExposure_Gd	23896.399209	23896.399209
10	BldgType_Twnhs	-23371.645514	23371.645514
11	WoodDeckSF	18622.262597	18622.262597
12	Exterior2nd_Stucco	-18108.290216	18108.290216
13	BldgType_TwnhsE	-16370.029073	16370.029073
14	Exterior2nd_CmentBd	16032.022451	16032.022451

```
final_ridge_variables.head(14)
```

	Variable_Name	Coef	Abs_Coef
0	GrLivArea	274944	274944
1	OverallQual	141407	141407
2	RoofMatl_WdShngl	119304	119304
3	Functional_Sev	-83633.6	83633.6
4	Neighborhood_NridgHt	73129.8	73129.8
5	Heating_OthW	-68347	68347
6	Neighborhood_StoneBr	64854.8	64854.8
7	Utilities_NoSeWa	-52870.3	52870.3
8	Exterior2nd_ImStucc	49330.6	49330.6
9	BldgType_Twnhs	-47864.5	47864.5
10	Neighborhood_NoRidge	41179.9	41179.9
11	GarageQual_Po	-40512.4	40512.4
12	Neighborhood_NPkVill	36506.4	36506.4
13	Functional_Maj2	-35411.4	35411.4

```
final_ridge_variables_change.head(14)
```

	Variable_Name	Coef	Abs_Coef
0	GrLivArea	242658	242658
1	OverallQual	129681	129681
2	RoofMatl_WdShngl	115001	115001
3	Functional_Sev	-68914.5	68914.5
4	Neighborhood_NridgHt	63762.5	63762.5
5	Neighborhood_StoneBr	55835	55835
6	Heating_OthW	-54300.3	54300.3
7	Utilities_NoSeWa	-50876	50876
8	Exterior2nd_ImStucc	48569.9	48569.9
9	BldgType_Twnhs	-37547.1	37547.1
10	Neighborhood_NoRidge	36792.1	36792.1
11	Functional_Maj2	-35555.5	35555.5
12	MasVnrArea	33518.8	33518.8
13	Neighborhood_Blueste	25943.8	25943.8

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?

Ans: - On increase of lambda value more features will have 0 coefficient, but we need to look at R2 as well. There should be very less difference between train and test. On looking at the combination of this two we need to take decision. We have used high lambda is second iteration we are able to see there is no much difference between R2 value of both lasso and ridge. Even top 5 variables are same with both the values of lambda. But we are getting an advantage of using less variables in final model in case of high lambda. With less variable chances of changing CSI will be low. In our case lambda is very less but we are getting very less difference between R2 value of test and train almost negligible. In second case we have used high value of lambda and we are getting good R2 value as well, and difference between test and train is low as well. So, with high value of lambda with same R2 (almost) value we are able to reduce variable, so second iteration is good as comparative to first iteration.

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?

Ans: -

```
# Dropping First 5 Variable and taking the same alpha
a.head(20)
```

	Variable_Name	Coef	Abs_Coef
12	MasVnrArea	138707.059036	138707.059036
1	Neighborhood_StoneBr	65327.306717	65327.306717
63	Neighborhood_MeadowV	-56451.859493	56451.859493
3	Neighborhood_NoRidge	55932.594864	55932.594864
13	Functional_Maj2	-52032.622612	52032.622612
2	Exterior2nd_ImStucc	51227.200371	51227.200371
11	Exterior2nd_CmentBd	45493.290176	45493.290176
10	WoodDeckSF	41462.564399	41462.564399
6	BsmtExposure_Gd	34500.663036	34500.663036
20	Neighborhood_Blueste	33895.724565	33895.724565
14	Neighborhood_NPkVill	32302.125222	32302.125222
16	Foundation_Slab	-31857.690646	31857.690646
15	Neighborhood_Crawfor	28138.558887	28138.558887
7	BldgType_TwnhsE	-27970.498715	27970.498715
67	RoofMatl_Membran	25162.473370	25162.473370

```
# Dropping First 5 Variable and tuning alpha
a.head(20)
```

	Variable_Name	Coef	Abs_Coef
12	MasVnrArea	139381.290958	139381.290958
1	Neighborhood_StoneBr	66479.093875	66479.093875
13	Functional_Maj2	-58420.279490	58420.279490
63	Neighborhood_MeadowV	-57951.735450	57951.735450
3	Neighborhood_NoRidge	56297.013748	56297.013748
2	Exterior2nd_ImStucc	53395.413568	53395.413568
11	Exterior2nd_CmentBd	46683.234349	46683.234349
20	Neighborhood_Blueste	43550.001314	43550.001314
10	WoodDeckSF	41629.526640	41629.526640
14	Neighborhood_NPkVill	36597.961531	36597.961531
6	BsmtExposure_Gd	34662.991610	34662.991610
16	Foundation_Slab	-33641.898918	33641.898918
15	Neighborhood_Crawfor	30260.781311	30260.781311
7	BldgType_TwnhsE	-29890.909294	29890.909294
67	RoofMatl_Membran	29219.615358	29219.615358

R2 value is dropping drastically.

Train: - 0.6689850515889246

0.7175182901130661

Test: -

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Ans: -

1. To make the model more robust and generalizable we can split the event into buckets of 10 or 15 and based on business requirement we can combine them in later stage. After converting the event into buckets we can then build a multi-class classification model on the top of it.
2. As in the current scenario we can convert the predicted output and actual output into buckets then we can look at over, under and same band prediction and try to make it more accurate, more robust and generalizable.
3. Bucket prediction we cater for robustness and generalization.
4. But when we do a bucket prediction we loose on our accuracy when compare to actual value. As output can be consider as mean value of that bucket and giving more loop holes for incorrect prediction.