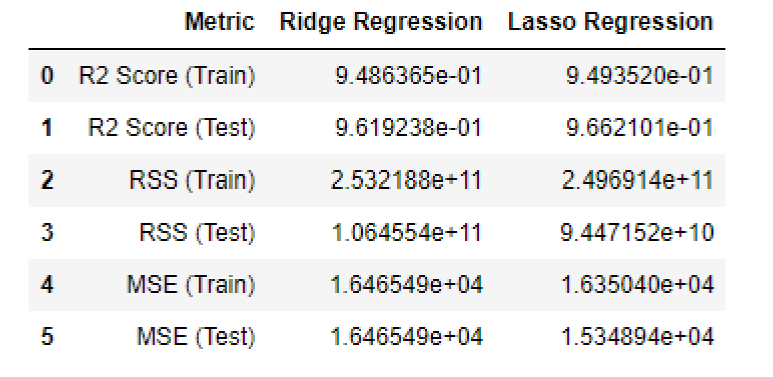
Assignment Part-I

Q1-Which variables are significant in predicting the price of a house?

- GrLivArea
- YearBuilt
- OverallQual_9
- OverallQual_8
- TotalBsmtSF
- BsmtFinSF1
- SaleType_New
- Functional_Typ
- OverallQual_10
- LotArea

Q2 -How well those variables describe the price of a house



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.486365e-01	9.493520e-01
1	R2 Score (Test)	9.619238e-01	9.662101e-01
2	RSS (Train)	2.532188e+11	2.496914e+11
3	RSS (Test)	1.064554e+11	9.447152e+10
4	MSE (Train)	1.646549e+04	1.635040e+04
5	MSE (Test)	1.646549e+04	1.534894e+04

Attributes	Ridge	Lasso
MSSubClass	-1440.21509	-2677.95
LotFrontage	1388.961764	704.5802
LotArea	4237.009361	4817.778
YearBuilt	4649.447228	9769.289
YearRemodAdd	2132.734279	1362.468
BsmtFinSF1	5943.468665	6564.843
BsmtFinSF2	552.722998	683.513
BsmtUnfSF	435.174331	0
TotalBsmtSF	7052.447656	7245.852
CentralAir	597.617358	0
1stFlrSF	7049.475689	0
2ndFlrSF	5841.000047	616.6708

LowQualFinSF	0	0
GrLivArea	10515.08494	25205.58
BsmtFullBath	1880.234999	1490.398
BsmtHalfBath	-64.802937	0
FullBath	3565.507285	2450.223
HalfBath	2426.606671	1061.701
BedroomAbvGr	-1291.775366	-1970.12
KitchenAbvGr	0	0
TotRmsAbvGrd	3703.066701	1841.014
Fireplaces	1587.134294	815.3802
GarageCars	2014.609531	1031.86
GarageArea	4527.492315	4478.204
WoodDeckSF	1699.585668	1535.793
OpenPorchSF	2689.54157	2159.877
EnclosedPorch	1199.232839	1394.26
ScreenPorch	2134.685688	2299.353
MoSold	-257.970774	-529.841
YrSold	-422.296394	-245.389
MSZoning_FV	1447.325179	3575.051
MSZoning_RH	326.252839	1070.906
MSZoning_RL	1274.663138	4640.138
MSZoning_RM	1160.295294	3687.396
Street_Pave	528.478976	157.6299
LotShape_IR2	878.638259	608.4521
LotShape_IR3	-61.416852	-41.3911
LotShape_Reg	-143.635937	263.0023
LandContour_HLS	1302.119539	1508.229
LandContour_Low	-466.406529	-182.242
LandContour_Lvl	-124.640473	0
Utilities_NoSeWa	-1101.866774	-971.062
LotConfig_CulDSac	1773.703066	1748.396
LotConfig_FR2	-947.404877	-842.457
LotConfig_FR3	-904.233118	-743.592
LotConfig_Inside	162.245155	0
LandSlope_Mod	1179.623651	1003.498
LandSlope_Sev	-3045.243813	-3917.47
Neighborhood_Blueste	237.522574	308.4645
Neighborhood_BrDale	669.993051	1233.39
Neighborhood_BrkSide	1385.824674	2420.015
Neighborhood_ClearCr	-444.516276	-209.479
Neighborhood_CollgCr	-493.424873	0
Neighborhood_Crawfor	3256.025176	4304.247
Neighborhood_Edwards	-2294.737456	-1567.29
Neighborhood_Gilbert	-1020.450315	-130.48
Neighborhood_IDOTRR	-1439.711454	-222.439

Neighborhood_MeadowV	-1720.901333	-860.461
Neighborhood_Mitchel	-1763.990712	-1634.5
Neighborhood_NAmes	-1283.605938	-117.023
Neighborhood_NPkVill	388.06599	563.1851
Neighborhood_NWAmes	-1657.144624	-1299.73
Neighborhood_NoRidge	4544.634848	4458.541
Neighborhood_NridgHt	4869.963367	4626.254
Neighborhood_OldTown	-1758.327604	-388.085
Neighborhood_SWISU	-806.085048	-358.069
Neighborhood_Sawyer	-409.950594	346.0708
Neighborhood_SawyerW	72.455801	271.3674
Neighborhood_Somerst	234.401292	188.668
Neighborhood_StoneBr	4092.584408	4461.441

Assignment Part-2

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 Value of Alpha -10 (Ridge and Lasso)

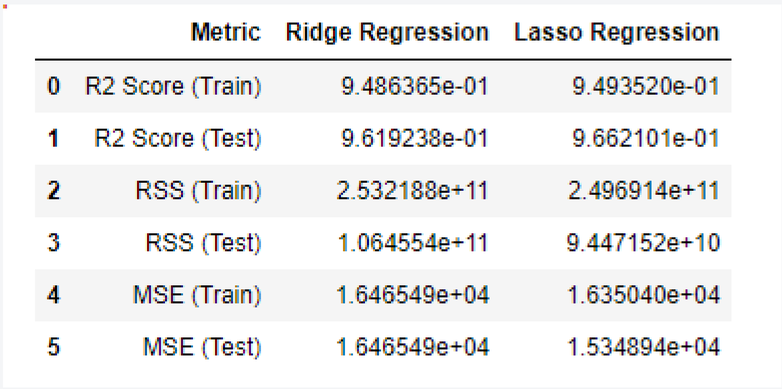
After doubling the value the (Alpha=100) the top predictor remains same

Top Predictor: GrLivArea

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?

Optimal Value of Alpha -10 (Ridge and Lasso)



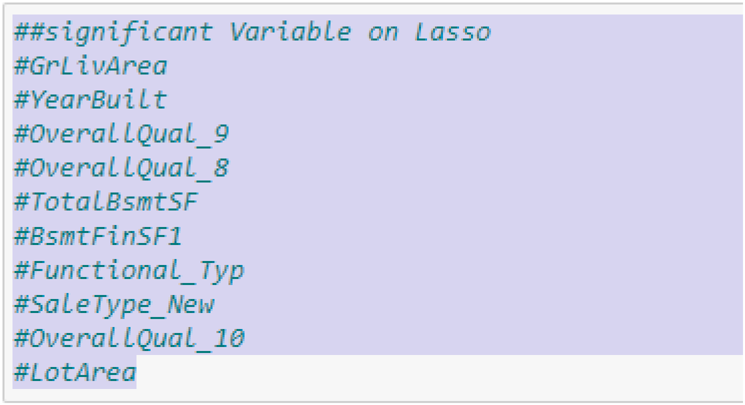
	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.486365e-01	9.493520e-01
1	R2 Score (Test)	9.619238e-01	9.662101e-01
2	RSS (Train)	2.532188e+11	2.496914e+11
3	RSS (Test)	1.064554e+11	9.447152e+10
4	MSE (Train)	1.646549e+04	1.635040e+04
5	MSE (Test)	1.646549e+04	1.534894e+04

We see that R2 score for Lasso is slightly better than Ridge. So I'll 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?

Below are the top 10 predictor values for Lasso regression,



```
##significant Variable on Lasso
#GrLivArea
#YearBuilt
#OverallQual_9
#OverallQual_8
#TotalBsmtSF
#BsmtFinSF1
#Functional_Typ
#SaleType_New
#OverallQual_10
#LotArea
```

Top 5 Predictor Now(After removing the top5):

2ndFlrSF
MSZoning_RL
MSZoning_RM
1stFlrSF
BsmtFinSF1

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?

- Outliners should be removed to keep the Model more robust. So preference to have the outliner analysis done during EDA .
- Model should give the nearly same result if we change the data.
- Test accuracy should not be less than training accuracy.
- If we are not getting the same level of accuracy then the model is not a robust one and cannot be used for predictive analysis.