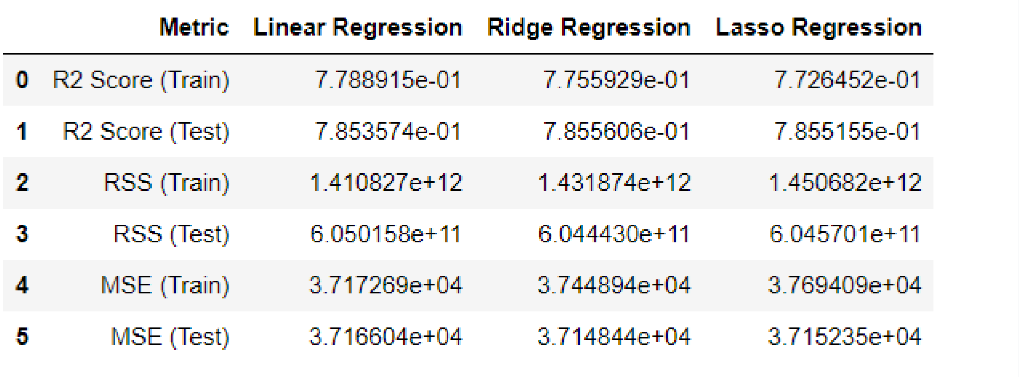
Q1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans. The optimal value of alpha for ridge regression is 100 and lasso regression is 1000.

Alpha = 100 for Ridge

Alpha = 1000 for Lasso

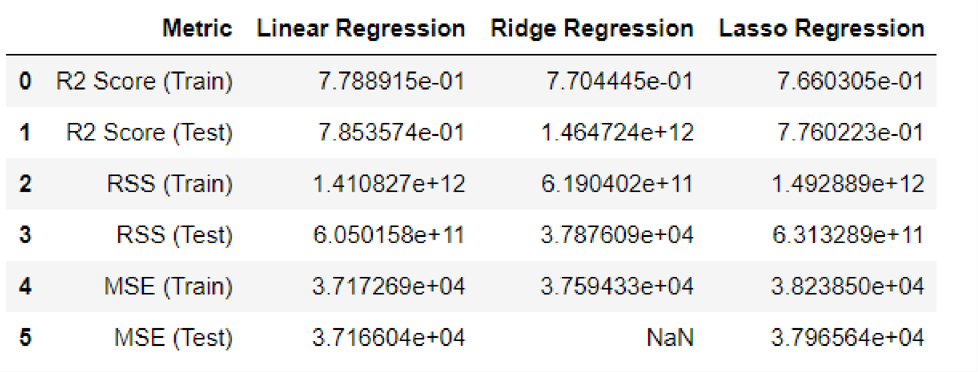


	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	7.788915e-01	7.755929e-01	7.726452e-01
1	R2 Score (Test)	7.853574e-01	7.855606e-01	7.855155e-01
2	RSS (Train)	1.410827e+12	1.431874e+12	1.450682e+12
3	RSS (Test)	6.050158e+11	6.044430e+11	6.045701e+11
4	MSE (Train)	3.717269e+04	3.744894e+04	3.769409e+04
5	MSE (Test)	3.716604e+04	3.714844e+04	3.715235e+04

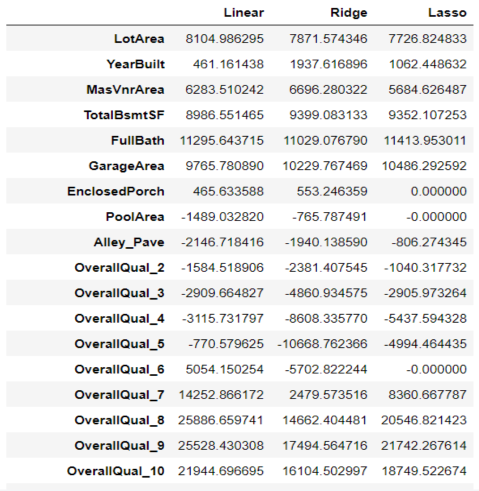
Alpha = 200 for Ridge

Alpha = 2000 for Lasso

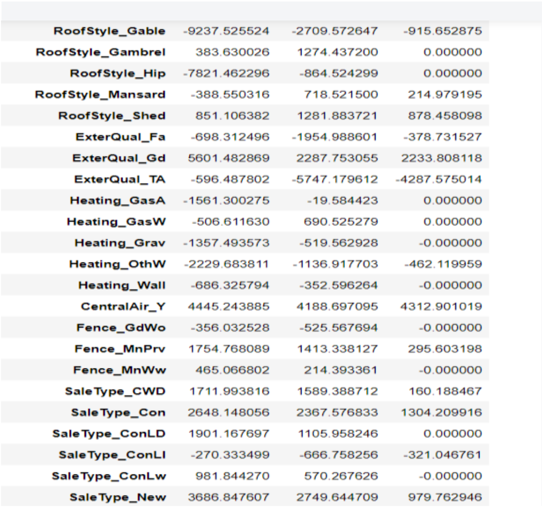
The MSE (test) for Ridge regression becomes NaN.



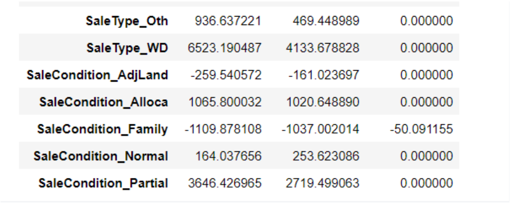
	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	7.788915e-01	7.704445e-01	7.660305e-01
1	R2 Score (Test)	7.853574e-01	1.464724e+12	7.760223e-01
2	RSS (Train)	1.410827e+12	6.190402e+11	1.492889e+12
3	RSS (Test)	6.050158e+11	3.787609e+04	6.313289e+11
4	MSE (Train)	3.717269e+04	3.759433e+04	3.823850e+04
5	MSE (Test)	3.716604e+04	NaN	3.796564e+04



	Linear	Ridge	Lasso
LotArea	8104.986295	7871.574346	7726.824833
YearBuilt	461.161438	1937.616896	1062.448632
MasVnrArea	6283.510242	6696.280322	5684.626487
TotalBsmtSF	8986.551465	9399.083133	9352.107253
FullBath	11295.643715	11029.076790	11413.953011
GarageArea	9765.780890	10229.767469	10486.292592
EnclosedPorch	465.633588	553.246359	0.000000
PoolArea	-1489.032820	-765.787491	-0.000000
Alley_Pave	-2146.718416	-1940.138590	-806.274345
OverallQual_2	-1584.518906	-2381.407545	-1040.317732
OverallQual_3	-2909.664827	-4860.934575	-2905.973264
OverallQual_4	-3115.731797	-8608.335770	-5437.594328
OverallQual_5	-770.579625	-10668.762366	-4994.464435
OverallQual_6	5054.150254	-5702.822244	-0.000000
OverallQual_7	14252.866172	2479.573516	8360.667787
OverallQual_8	25886.659741	14662.404481	20546.821423
OverallQual_9	25528.430308	17494.564716	21742.267614
OverallQual_10	21944.696695	16104.502997	18749.522674



RoofStyle_Gable	-9237.525524	-2709.572647	-915.652875
RoofStyle_Gambrel	383.630026	1274.437200	0.000000
RoofStyle_Hip	-7821.462296	-864.524299	0.000000
RoofStyle_Mansard	-388.550316	718.521500	214.979195
RoofStyle_Shed	851.106382	1281.883721	878.458098
ExterQual_Fa	-698.312496	-1954.988601	-378.731527
ExterQual_Gd	5601.482869	2287.753055	2233.808118
ExterQual_TA	-596.487802	-5747.179612	-4287.575014
Heating_GasA	-1561.300275	-19.584423	0.000000
Heating_GasW	-506.611630	690.525279	0.000000
Heating_Grav	-1357.493573	-519.562928	-0.000000
Heating_OthW	-2229.683811	-1136.917703	-462.119959
Heating_Wall	-686.325794	-352.596264	-0.000000
CentralAir_Y	4445.243885	4188.697095	4312.901019
Fence_GdWo	-356.032528	-525.567694	-0.000000
Fence_MnPrv	1754.768089	1413.338127	295.603198
Fence_MnWw	465.066802	214.393361	-0.000000
SaleType_CWD	1711.993816	1589.388712	160.188467
SaleType_Con	2648.148056	2367.576833	1304.209916
SaleType_ConLD	1901.167697	1105.958246	0.000000
SaleType_ConLI	-270.333499	-666.758256	-321.046761
SaleType_ConLw	981.844270	570.267626	-0.000000
SaleType_New	3686.847607	2749.644709	979.762946



SaleType_Oth	936.637221	469.448989	0.000000
SaleType_WD	6523.190487	4133.678828	0.000000
SaleCondition_AdjLand	-259.540572	-161.023697	0.000000
SaleCondition_Alloca	1065.800032	1020.648890	0.000000
SaleCondition_Family	-1109.878108	-1037.002014	-50.091155
SaleCondition_Normal	164.037656	253.623086	0.000000
SaleCondition_Partial	3646.426965	2719.499063	0.000000

The important predictors are:

- LotArea - Lot size in square feet
- MasVnrArea - Masonry veneer area in square feet
- TotalBsmtSF - Total square feet of basement area
- Fullbath - Full bathrooms above grade
- Central Air - Central air conditioning

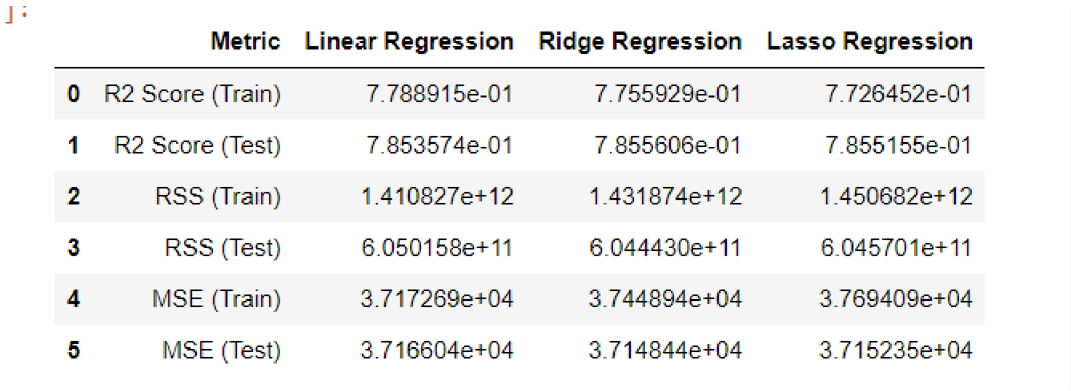
These house characteristics significantly affect the house prices. Overall quality being the categorical variable, as the overall quality in terms of material and finish increases with the reference to very poor category, house of price also increases.

Some of the variables negatively impact the house prices such area of the Pool (Pool area) and Type of alley access to property.

After changing the values of the alpha, there is no change in the important 5 predictors.

Q2. 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. As given below, the R^2 for train data for Ridge regression is slightly above the Lasso regression while R^2 for test data for Ridge regression is almost close to that of the Lasso regression. Thus, I will choose to apply Ridge regression.

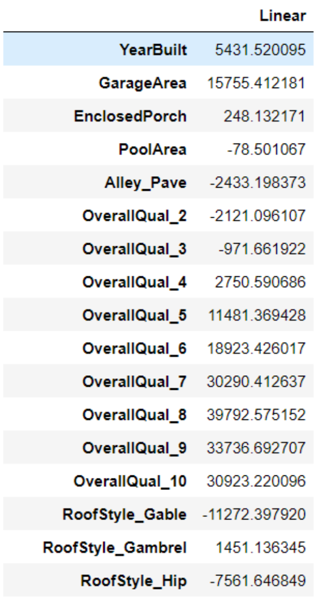


	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	7.788915e-01	7.755929e-01	7.726452e-01
1	R2 Score (Test)	7.853574e-01	7.855606e-01	7.855155e-01
2	RSS (Train)	1.410827e+12	1.431874e+12	1.450682e+12
3	RSS (Test)	6.050158e+11	6.044430e+11	6.045701e+11
4	MSE (Train)	3.717269e+04	3.744894e+04	3.769409e+04
5	MSE (Test)	3.716604e+04	3.714844e+04	3.715235e+04

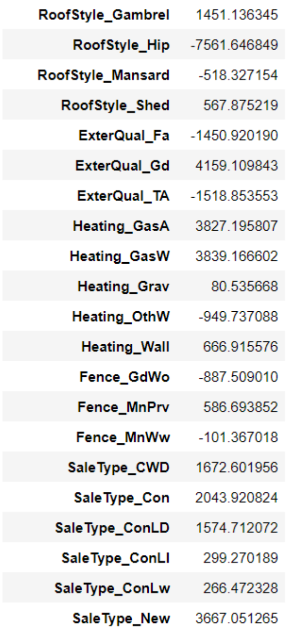
Q3. After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Ans. As per question, I have dropped 'LotArea ', 'MasVnrArea ', 'TotalBsmtSF ', 'CentralAir', 'FullBath. As given below the top five predictors are

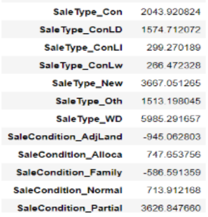
- YearBuilt- year in which it was built,
- GarageArea - Size of garage in square feet
- OverallQual- Rates the overall material and finish of the house.
- EnclosedPorch- Enclosed porch area in square feet
- PoolArea- Pool area in square feet



	Linear
YearBuilt	5431.520095
GarageArea	15755.412181
EnclosedPorch	248.132171
PoolArea	-78.501067
Alley_Pave	-2433.198373
OverallQual_2	-2121.096107
OverallQual_3	-971.661922
OverallQual_4	2750.590686
OverallQual_5	11481.369428
OverallQual_6	18923.426017
OverallQual_7	30290.412637
OverallQual_8	39792.575152
OverallQual_9	33736.692707
OverallQual_10	30923.220096
RoofStyle_Gable	-11272.397920
RoofStyle_Gambrel	1451.136345
RoofStyle_Hip	-7561.646849



RoofStyle_Gambrel	1451.136345
RoofStyle_Hip	-7561.646849
RoofStyle_Mansard	-518.327154
RoofStyle_Shed	567.875219
ExterQual_Fa	-1450.920190
ExterQual_Gd	4159.109843
ExterQual_TA	-1518.853553
Heating_GasA	3827.195807
Heating_GasW	3839.166602
Heating_Grav	80.535668
Heating_OthW	-949.737088
Heating_Wall	666.915576
Fence_GdWo	-887.509010
Fence_MnPrv	586.693852
Fence_MnWw	-101.367018
SaleType_CWD	1672.601956
SaleType_Con	2043.920824
SaleType_ConLD	1574.712072
SaleType_ConLI	299.270189
SaleType_ConLw	266.472328
SaleType_New	3667.051265



SaleType_Con	2043.920824
SaleType_ConLD	1574.712072
SaleType_ConLI	299.270189
SaleType_ConLw	266.472328
SaleType_New	3667.051265
SaleType_Oth	1513.198045
SaleType_WD	5985.291657
SaleCondition_AdjLand	-945.062803
SaleCondition_Alloca	747.653756
SaleCondition_Family	-586.591359
SaleCondition_Normal	713.912168
SaleCondition_Partial	3626.847660

R^2 of the new model is 73% and this has been decreased as compared to the original model.

Q4. 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. The R^2 of the both train and test are quite close approx. 78%. The model is robust and generalisable as predictors are able to explain 78% variation in the house prices. The model is robust and performing consistently on both train and test data.