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:
Optimal Value of alpha for ridge and lasso regression is 500.
If we use 1000 as alpha value, R2 value for both training and test data set decreases.
See comparison below:

	Metric	Linear Regression	Ridge Regression_500	Lasso Regression_500	Ridge Regression_1000	Lasso Regression_1000
0	R2 Score (Train)	9.332834e-01	9.332830e-01	9.402480e-01	9.158972e-01	9.298193e-01
1	R2 Score (Test)	-1.145205e+20	8.862030e-01	9.040113e-01	8.725126e-01	9.015748e-01
2	RSS (Train)	-1.145205e+20	8.862030e-01	9.040113e-01	8.725126e-01	9.015748e-01
3	RSS (Test)	3.495264e+11	3.495264e+11	3.130389e+11	4.406124e+11	3.676750e+11
4	MSE (Train)	1.898253e+04	1.898253e+04	1.796442e+04	2.131290e+04	1.946911e+04
5	MSE (Test)	8.462573e+14	2.667633e+04	2.450027e+04	2.823543e+04	2.480927e+04

Predictor variables for alpha 500:

	Linear	Ridge	Lasso
GrLivArea	9.751602e+03	6964.654676	20485.309132
TotalBsmtSF	-1.698498e+15	6095.127578	9762.473642
OverallQual_9	-1.369992e+14	5149.517877	9744.089404
OverallQual_8	-2.636263e+14	3975.134738	7925.736399
BsmtFinSF1	1.755575e+15	5067.099075	7459.532391
OverallQual_10	-7.888043e+13	4176.500227	7086.040605
SaleCondition_Partial	8.817938e+03	2606.417507	6953.419706
LotArea	6.730153e+03	4103.704899	5234.248595
GarageArea	4.871688e+03	3983.708982	4931.660497
Neighborhood_NridgHt	1.853875e+03	4308.654129	4332.002711
BsmtExposure_Gd	4.347844e+03	3626.034904	4331.403050
Neighborhood_StoneBr	3.211125e+03	3541.563673	3913.525110
Neighborhood_Crawfor	5.612344e+02	2649.903439	3628.757371
HalfBath_1	3.825188e+03	3173.863531	3337.174631
TotRmsAbvGrd_11	9.007183e+13	3571.387557	2983.008991
MSSubClass_60	-5.230562e+03	1653.610546	2963.751551
OverallCond_7	-3.050079e+14	1578.336153	2832.937416
Foundation_PConc	2.267109e+03	1892.143537	2775.218304
GarageCars_3	1.309400e+15	3649.511181	2727.669578
KitchenAbvGr_1	-1.627896e+15	1360.987925	2713.549489
1stFlrSF	8.769092e+03	6413.936018	2644.254301
BsmtFinType1_NA	4.668000e+03	296.117777	2517.597614
OverallCond_9	-9.952402e+13	1730.002431	2464.526310
BsmtFinType1_GLQ	2.947703e+03	2612.294407	2463.552368
Functional_Typ	-7.656250e+02	949.639707	2365.323325
Neighborhood_NoRidge	1.614758e+03	2644.494303	2153.022876
Exterior1st_BrkFace	-9.912500e+02	1837.334624	2078.805834
OverallQual_7	-3.374445e+14	-71.812573	1971.733569
Condition1_Norm	3.335688e+03	1555.816119	1884.319317
OverallCond_8	-1.911931e+14	1230.980844	1725.411444
Exterior2nd_CmentBd	7.955188e+03	1372.601858	1554.029993
YearBuiltRange_2001-2010	1.592712e+04	1368.514027	1535.669572
Fireplaces_1	8.201250e+02	2035.526647	1500.584757
FullBath_3	6.106773e+03	2732.410516	1499.253028
MSZoning_FV	7.571500e+03	1065.217867	1476.882338
SaleCondition_Alloca	2.304086e+03	1149.852976	1365.956285
SaleCondition_Normal	3.133234e+03	701.928626	1318.967253
TotRmsAbvGrd_10	1.369992e+14	2632.680006	1313.540081
Neighborhood_BrkSide	-1.932062e+03	815.793926	1268.120152
Fireplaces_2	7.862188e+02	2352.497779	1251.148333
MasVnrArea	8.494834e+02	2717.782618	1239.938462

Predictor variables for alpha 1000:

	Linear	Ridge	Lasso
GrLivArea	9.751602e+03	5955.802685	20190.655174
OverallQual_9	-1.369992e+14	4436.336527	10168.090837
TotalBsmtSF	-1.698498e+15	5187.706661	8222.525052
OverallQual_8	-2.636263e+14	3433.754000	7776.030021
OverallQual_10	-7.888043e+13	3540.781126	7258.509728
BsmtFinSF1	1.755575e+15	4187.878882	7225.344726
SaleCondition_Partial	8.817938e+03	2254.125703	6115.847895
GarageArea	4.871688e+03	3812.126307	5875.678742
LotArea	6.730153e+03	3592.697496	5094.687722
1stFlrSF	8.769092e+03	5482.062314	4120.106817
BsmtExposure_Gd	4.347844e+03	3155.610624	4118.495764
Neighborhood_NridgHt	1.853875e+03	3780.343191	3678.840924
Neighborhood_Crawfor	5.612344e+02	2207.571259	3496.970265
Neighborhood_StoneBr	3.211125e+03	2937.535848	3328.410412
HalfBath_1	3.825188e+03	2659.023545	3301.674719
MSSubClass_60	-5.230562e+03	1571.872657	3118.631069
Foundation_PConc	2.267109e+03	1772.752712	3015.172737
GarageCars_3	1.309400e+15	3749.047580	2815.465667
KitchenAbvGr_1	-1.627896e+15	1121.567052	2702.688125
OverallCond_7	-3.050079e+14	1053.376592	2508.814154
TotRmsAbvGrd_11	9.007183e+13	3142.053472	2483.307937
BsmtFinType1_GLQ	2.947703e+03	2455.248590	2462.309287
Functional_Typ	-7.656250e+02	841.094269	2421.445559
OverallCond_9	-9.952402e+13	1298.007421	1890.183274
Condition1_Norm	3.335688e+03	1211.536455	1785.875398
Exterior1st_BrkFace	-9.912500e+02	1517.378478	1722.406068
Fireplaces_1	8.201250e+02	2017.025751	1576.983873
FullBath_3	6.106773e+03	2584.519782	1491.852237
YearBuiltRange_2001-2010	1.592712e+04	1349.062545	1417.479964
Neighborhood_NoRidge	1.614758e+03	2327.487762	1389.016844
OverallQual_7	-3.374445e+14	-105.574626	1258.535834
Fireplaces_2	7.862188e+02	2231.228781	1254.262795
MasVnrArea	8.494934e+02	2750.395794	1145.016035
OverallCond_8	-1.911931e+14	811.075319	1141.136089
MSZoning_FV	7.571500e+03	694.893139	883.727547
MasVnrType_Stone	2.133500e+03	1559.792339	864.460409
Exterior2nd_CmentBd	7.955188e+03	1244.469501	847.030209
BsmtFinType1_NA	4.668000e+03	-119.326325	828.185151
Neighborhood_BrkSide	-1.932062e+03	542.255932	766.875374
TotRmsAbvGrd_10	1.369992e+14	2341.210335	629.987241

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:
We would choose Lasso regression since we have better R2 score and better coefficient values.

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: Below are the new predictor variables

	Ridge	Lasso
1stFlrSF	17750.882992	27031.352352
2ndFlrSF	5854.823278	15059.298294
Neighborhood_NridgHt	7611.648230	8106.205568
LotArea	5850.997454	5840.099281
Neighborhood_StoneBr	5914.974911	5757.501376
BsmtExposure_Gd	5586.669032	5732.003524
GarageCars_3	3853.183182	5151.553714
GarageArea	5644.734482	5061.827363
BsmtFinType1_GLQ	4177.723181	4777.949108
SaleType_New	4427.228350	4724.232186
HalfBath_1	4645.556445	4253.882599
Neighborhood_NoRidge	4528.552814	3799.722332
Neighborhood_Crawfor	3061.785873	3336.215684
MSSubClass_60	1453.512317	3238.268966
Functional_Typ	596.241433	3014.348670
YearBuiltRange_2001-2010	2837.365282	2993.146088
TotRmsAbvGrd_11	4667.814928	2925.428050
Foundation_PConc	3296.477877	2844.630871
MasVnrArea	3598.255981	2758.287952
OverallCond_7	2337.462106	2589.425933
OverallCond_9	2651.256247	2582.926690
FullBath_3	3754.251289	2546.069029
TotRmsAbvGrd_10	4169.960789	2482.982188
Exterior1st_BrkFace	2804.659233	2458.682850
MSZoning_FV	2402.835104	2193.619216
Condition1_Norm	2866.888591	2191.941575
Fireplaces_2	3271.252794	2155.852936
OverallCond_8	2185.914611	1946.737560
BsmtFullBath_2	2767.477115	1929.288496
Exterior2nd_CmentBd	1958.851658	1889.019714
KitchenAbvGr_1	1853.702175	1827.901443
BsmtFullBath_1	2336.071288	1397.457056

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?

Ans:
Robustness and generalizability of the model is determined on how well the model behaves on the test data set. Unseen dataset which might be little more noisy and adverse.
To keep the model robust and generalisable it's better to split the data into train, validation and test datasets.
Model is run on the train data set and validated on the validation data set.
Here we need to check the accuracy on the train and the validation data set.
Accuracy should not have much variation for the model to be good.
Finally when this model is run on a test data set, accuracy should be similar to the train and the validation data set for the model to be called a good model and robust.