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 :

For Ridge Regression alpha value turns out to be 6

For Laao Regression the alpha is 0.0001

For alpha =6

	FOR alpha = 6	For Alpha =12
R2 Score Train	0.8369537881840629	0.8368985821965664
R2 Score Test	0.8014470687988775	0.8027142565258067
RSS TRAIN	1040348767993.2339	1040701020771.3762
RSS Test	559663635114.389	556091797190.1873
MSE Train	1018950801.1686914	1019295808.7868524
MSE Test	1277770856.4255457	1269615975.3200622

The metrics don't change with change in alpha

There is a considerable change in the values of coefficients if the alpha is doubled in ridge regression , however there's is not much change in lasso regression , the consecutive changes in the values of coefficients is tabulated below.

For lasso the most important variable remains unchanged which is

1st Floor SF - First Floor square feet

2nd Floor SF- Second floor square feet

Overall Quality

Year Built

	Ridge	Ridge		Lasso	Lasso	
	alpha =6	*alpha =12*	*chage in coeff*	*alpha = 0.0001*	*aplha=0.0002*	*change in coeff*
LotArea	5270.92	5329.00	58.07	5209.07	5209.07	6E-06
OverallQual	16488.01	16626.09	138.07	16334.62	16334.62	0.00023
OverallCond	5968.28	5895.41	72.87	6043.84	6043.84	0.000174
YearBuilt	10761.72	10637.67	124.05	10893.43	10893.43	0.000192
MasVnrArea	6596.77	6594.87	1.90	6598.65	6598.65	0.000161
ExterQual	-5003.55	-5082.78	79.23	-4922.07	-4922.07	5E-06
BsmtQual	-9644.58	-9641.08	3.50	-9649.12	-9649.12	6E-06
BsmtFinSF1	5022.20	5074.82	52.62	4965.85	4965.85	1.7E-05
1stFlrSF	28400.63	28018.85	381.78	28799.78	28799.78	0.000185
2ndFlrSF	23034.95	22749.78	285.17	23332.37	23332.37	0.000241
KitchenAbvGr	-5015.95	-4925.65	90.30	-5111.32	-5111.32	9.4E-05
KitchenQual	-8239.00	-8263.59	24.59	-8213.79	-8213.79	8E-06
Condition2_PosN	-14709.98	-14526.86	183.12	-14898.15	-14898.15	0.000231
RoofStyle_Hip	1294.32	1301.23	6.91	1287.55	1287.55	0.000116
RoofMatl_Membran	1649.24	1634.31	14.93	1665.04	1665.04	8.8E-05
RoofMatl_Metal	1097.43	1083.26	14.17	1112.26	1112.26	7.7E-05
RoofMatl_Roll	-367.54	-366.91	0.63	-368.09	-368.09	7.8E-05
RoofMatl_Tar&Grv	440.07	439.58	0.49	440.31	440.31	0.000157
RoofMatl_WdShake	-1381.74	-1335.38	46.37	-1429.94	-1429.94	0.000111
RoofMatl_WdShngl	5649.86	5629.03	20.83	5671.23	5671.23	9.3E-05
MasVnrType_BrkFace	-3968.30	-3866.52	101.78	-4073.58	-4073.58	0.000268
Functional_Typ	4232.87	4165.64	67.23	4303.79	4303.79	0.000158
SaleType_New	2299.25	2314.20	14.95	2283.53	2283.53	6.1E-05

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?

Lasso Regression as theoreticaaly it removes some of the unimportant features , in this cae here in lasso regression even of we double the value of alpha these not much change in the model , so it will turn out to be a better technique

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?

OverallCond
RoofMatl_WdShngl
LotArea
BsmtFinSF1
Functional_Typ

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?

Removal of outliers , handling of missing data

For Generalizaition : The model should learn from the underlying pattern of the data rather than learning from noise if the latter happens it may result in model overfitting and resultant accuracy will be impacted , the test acccuarcy would be less than training accuracy.

For Robustness: Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis.