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:

Optimal Alpha value for Ridge : 0.3

Optimal Alpha value for Lasso : 0.0001

Doubling alpha of ridge: Below are the metrics of ridge regression with Alpha of 0.3 and 0.6

	Metric	Ridge Alpha-0.3	Ridge Alpha-0.6
0	R2 Score (Train)	0.815748	0.813152
1	R2 Score (Test)	0.801339	0.801875
2	RSS (Train)	2.267224	2.299168
3	RSS (Test)	1.079888	1.076971
4	MSE (Train)	0.047123	0.047454
5	MSE (Test)	0.049654	0.049587

Metric	Lasso Alpha-0.0001	Lasso Alpha-0.0002
R2 Score (Train)	0.808744	0.801297
R2 Score (Test)	0.799179	0.800867
RSS (Train)	2.353415	2.445047
RSS (Test)	1.091627	1.082453
MSE (Train)	0.048011	0.048936
MSE (Test)	0.049923	0.049713

From the above metrics comparison, we can observe that when the Alpha value is doubled the Train R^2 is decreased and Test R^2 is increased, this means that as the Alpha value increases Variance of the model will be decreased and Bias of the model will be increased

Top 10 Lasso predictor variables with 0.0001:

Feature	Coefficient	AbsCoef
Condition2_PosN	-0.280379	0.280379
OverallQual_10	0.278183	0.278183
OverallQual_9	0.182536	0.182536
LotArea	0.168969	0.168969
OverallQual_8	0.108809	0.108809
RoofMatl_WdShngl	0.100814	0.100814
TotRmsAbvGrd_4	-0.049531	0.049531
GarageCars_3	0.049241	0.049241
Fireplaces_2	0.043673	0.043673
TotRmsAbvGrd_10	0.043335	0.043335

Top 10 Lasso predictor variables with 0.0002:

	Feature	Coefficient	AbsCoef
	Condition2_PosN	-0.280379	0.280379
	OverallQual_10	0.278183	0.278183
	OverallQual_9	0.182536	0.182536
	LotArea	0.168969	0.168969
	OverallQual_8	0.108809	0.108809
	RoofMatl_WdShngl	0.100814	0.100814
	TotRmsAbvGrd_4	-0.049531	0.049531
	GarageCars_3	0.049241	0.049241
	Fireplaces_2	0.043673	0.043673
	TotRmsAbvGrd_10	0.043335	0.043335

As shown above from lasso Regression both before and after the change the top predictor variables remain the same.

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?

Answer:

Lasso regression with 39 independent variables is selected as final model. Even though Ridge R^2 is marginally better than Lasso I have **selected Lasso as the model is less complex with 39 variables** as compared to ridge with 54 variables.

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?

Answer:

Top 5 features in the final lasso model are 'Condition2_PosN', 'OverallQual_10', 'OverallQual_9', 'LotArea', 'OverallQual_8'.

After removing the above variables from X Train and rebuilding the lasso regression the below variables are obtained as top 5 features. Some of them are not same as top 6 to 10 features that are present in the model before removing top 5. So the importance of the predictor variables are changed.

	Feature	Coefficient	AbsCoef
24	RoofMatl_WdShngl	0.167012	0.167012
50	GarageCars_3	0.111349	0.111349
42	TotRmsAbvGrd_10	0.080597	0.080597
36	TotRmsAbvGrd_3	-0.076111	0.076111
37	TotRmsAbvGrd_4	-0.070767	0.070767

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?

Answer:

To make the model more robust and generalisable we should use Regularisation methods that balances the model between over fitting and under fitting. We should choose appropriate Alpha where there sum of the RSS and penalty is minimum. This value of Alpha is called optimum model complexity.

We can use cross validation to arrive at optimum alpha value.

As alpha increases the variance will decrease and Bias will increase