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 to 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: If considering all the variables at once and without implementing RFE, the optimal values are :

Ridge Regression :-0.05

Lasso Regression : - 20

The r2 score looks good for both train and test data not much difference is there. In lasso regression few features might have been eliminated means coefficient almost 0. But MSE still is more.

On doubling the values,

Ridge Regression: -0.1

Lasso Regression:- 40

	Metric	Ridge Regression	Lasso Regression	Ridge Regression After Double	Lasso Regression After Double
0	R2 Score (Train)	9.004698e-01	8.977562e-01	8.945318e-01	8.828201e-01
1	R2 Score (Test)	7.913883e-01	8.036510e-01	8.250005e-01	8.550250e-01
2	RSS (Train)	6.416331e+11	6.591263e+11	6.799130e+11	7.554140e+11
3	RSS (Test)	5.759738e+11	5.421167e+11	4.831710e+11	4.002740e+11
4	MSE (Train)	2.506862e+04	2.540806e+04	2.580559e+04	2.720067e+04
5	MSE (Test)	3.626305e+04	3.518109e+04	3.321340e+04	3.023024e+04

From this we can see that there is not much difference between training and test data r2 score, though it has reduced in comparison to above scenario But only r2 shouldn't be considered. It is important to check for MSE too. MSE is a little close in both the ridge and lasso regression but difference between training and test data is considerable.

Predictor variables are:-

From Lasso:

Roofmatl,Grlivarea,OverallQual,BsmtFinSF1,Lotarea,MasVnrArea, BsmtQual

In negative correlation- Condition 2, Functional , BsmtCondition, Neighbourhood

From Ridge:

Here we have got RoofMtl,GrLivArea,BsmtFinSF1 , OverallQual, SaleType_Con

In case of negative correlations Condition2,Functional,Exterior,Heating is also important

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 2: - If we see the optimal values we should go for ridge but Lasso is better since it performs elimination too.

Lasso Regression:- Here feature selection is possible. It shrinks the coefficients towards 0 to handle high variance. As we go on increasing the alpha value the coefficient become 0 . So larger the value more aggressive is the penalization.

Ridge Regression:- Higher the value of alpha higher the penalty but unlike Lasso it doesn't make the coefficient 0. It would include all the predictors in the final model. If the variables are too many then it causes issues.

Based on our requirement we can decide between Ridge and lasso. Better is lasso in case of too many features.

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 3: Building the model after removing the important 5 variables we found that

LotFrontage, FullBath, HeatingQC, YearRemote and BsmtFinType2 are 5 most important predictor variables now.

0	LotFrontage	181126.182174
56	Neighborhood_BrDale	73156.353938
62	Neighborhood_Gilbert	47740.695544
60	Neighborhood_Crawfor	46239.290097
8	BsmtUnfSF	34882.428877
61	Neighborhood_Edwards	27718.988845
58	Neighborhood_ClearCr	19294.140723
59	Neighborhood_CollgCr	19245.803410
84	Condition1_RRAn	17578.496671
88	Condition2_Norm	16055.879165
20	GarageQual	12187.876375
13	GrLivArea	11911.539166
89	Condition2_PosA	11692.701076
5	BsmtFinType1	10719.780704
9	TotalBsmtSF	10548.791656
11	1stFlrSF	9842.706940
21	GarageCond	9825.771879
6	BsmtFinSF1	8821.272381
2	BsmtQual	8456.988090

The below table I got when I used other variables and alpha values(not shown in notebook but derived earlier)

Featuere	Coef	
0	LotFrontage	180357.124113
19	FullBath	32167.350652
15	HeatingQC	10450.426291
4	YearRemodAdd	9605.309992
12	BsmtFinType2	9573.357327
163	OldTown	7848.543920
169	Timber	5918.681338
162	NridgHt	5560.262282
3	YearBuilt	5055.821762

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?

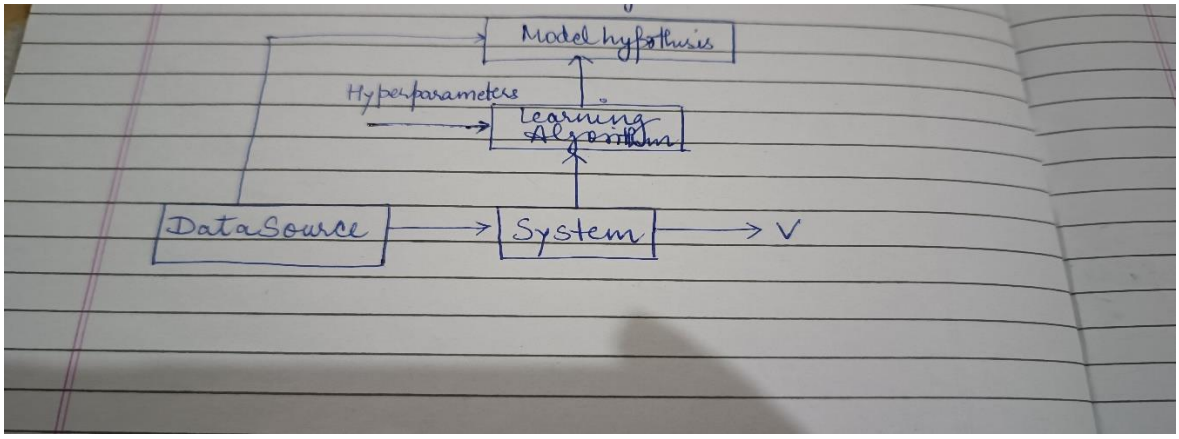
Answer 4:- We claim that a model is more robust and generalizable if it follows these conditions:

1) **No Overfitting**: - The model should not perform too well on training data and poorly on test data. It should not have high variance and should be able to perform well on unseen data even at the cost of training data.

2) **Low MSE** (Mean squared error):- i.e. the difference between the predicted values and the actual values should not be too high . Low MSE ensures low bias which is good for a model to perform well.

3) **Handling outliers**: - Too many outliers impact the data modelling. Because a few exceptions may train the data wrongly. Tree based models are not affected by outliers but regression models are. But we cannot use tree models always so we can perform statistical tests for this.

4) **Hyperparameters**:- Applying hyperparameters to regularize the model is another way to keep a check on the model. For eg:- In tree models the depth of a tree can be a hyperparameter other options are- minimum number of samples in a leaf or internal node etc. In regression adding a penalty to the error term handles it often denoted as alpha or lambda.



5) **Simple yet not naïve model: -** The model should not be too complex. In other words the polynomial degree shouldn't be too high so as to fit each and every feature and hence overfit it. Neither it should be too simple that it underfits the data (perform poorly even on training data).