**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 - : In the case of ridge regression:- When we plot the curve between negative mean absolute error and alpha, As we can see as alpha increases accuracy of train decreases and test data decreases but after a certain value(8) both decreses.So we have choosen that optimal value which gives us a balanced result i.e error in both training and test data is optimal.The alpha value is choosen is such a way so that the model do not underfit or overfit.

For lasso regression as we can see as alpha increases accuracy of train decreases and test data decreases but after a certain value(.0001) both decreses.So we have choosen that optimal value which gives us a balanced result i.e error in both training and test data is optimal.The alpha value is choosen is such a way so that the model do not underfit or overfit. When we double the value of alpha for our ridge regression no we will take the value of alpha equal to 16 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set .from the graph we can see that when alpha is 16 we get more error for both test and train. Similarly when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

The most important variable after the changes has been implemented for ridge regression are as follows:-

1. MSZoning\_FV 2. MSZoning\_RL 3. Neighborhood\_Crawfor 4. MSZoning\_RH 5. MSZoning\_RM 6. SaleCondition\_Partial 7. Neighborhood\_StoneBr 8. GrLivArea 9. SaleCondition\_Normal 10. Exterior1st\_BrkFace .
2. The most important variable after the changes has been implemented for lasso regression are as follows:- 1. GrLivArea 2. OverallQual 3. OverallCond 4. TotalBsmtSF 5. BsmtFinSF1 6. GarageArea 7. Fireplaces 8. LotArea 9. LotArea 10. LotFrontag

**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 - In Ridge we got alpha as 8 with .881 accuracy on train data set and .861 R2 on test dataset with 245 feature variables In Lasso we got alpha as 0.0001 with .888 and .865 R2 on test and train dataset with 132 feature variables. The accuracy of both the models is similar. To be precise Lasso performs better than ridge. I will choose lasso instead of ridge because lasso has achieved the same accuracy as ridge with less number of feature variables thereby reducing the complexity of the model. From a business perspective if we see the company would want to get a better result by considering as few factors as possible.More features mean more data is required thus more complex the model. It is simple to explain model to stakeholders and policy makers with as min variables as possible especially when the accuracy of the model is same in both cases

**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 - Those 5 most important predictor variables that will be excluded are :- 1. GrLivArea 2. OverallQual 3. OverallCond 4. TotalBsmtSF 5. GarageArea

**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-We train our model on train data set but whether our model is good or not is determined by how good the model performs on test or unseen data.For that a model needs to flexible. If should try to make model in such a way so that changes in training data do not affect the performance of the model to a large extent.Thus we should try to make our model simple. Extremely complex model reduces the generality of the model.The complex model will perform very good on the training data but very poorly on the test data.On the other hand an extremely simple model would not give us good results in both train and test data.This is where the concept of bias variance tradeoff comes into picture.We want to build a good model but we do not want to build it very complex. Hence we should always try to build a model which is complex enough to accommodate the variance of the data and accurate too but at the same time not too complex so that it overfits. The above task can be achieved using regularization. It is a process which forces the model to simplify itself by penalizing the model if the complexity of the model increases too much. But at the same time allows us to build just enough complex model which balances the overfitting and underfitting.