Subjective Assignment Questions

Q1. 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. Below are the optimal alpha value for the ridge lasso regression model.

	R2 Train Score	R2 Test Score
Ridge (alpha=20)	0.8846830175183199	0.8688378418646683
Lasso (alpha=0.001)	0.9064342411024804	0.8563465968424706

After the alpha value is doubled below are the R2 score.

	R2 Train Score	R2 Test Score
Ridge (alpha=40)	0.8733224535950143	0.8660272636096509
Lasso (alpha=0.002)	0.8938316817294861	0.8635862956225242

When alpha value is doubled there is slight difference in R2 score for ridge and lasso regression model.

Below changes are seen for the important predictor variables, for ridge it remains same but for lasso it changes.

```
Ridge max col = Neighborhood_NoRidge
Ridge max coef = 0.2996929317374142

Ridge_double max col = Neighborhood_NoRidge
Ridge_double max coef = 0.21362590514802876

Lasso max col = RoofMatl_WdShngl
Lasso max coef = 0.9371314350834653

Lasso_double max col = Neighborhood_NoRidge
Lasso_double max coef = 0.5461195806394026
```

Q2: 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: Lasso regression, with alpha value 0.01 as in lasso regression the feature selection becomes easy as in lasso the coefficients of features are equated to zero and it becomes easy to choose the features when the number of features are large.

Q3. 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: The most important predictor variables are:

1. RoofMatl_WdShngl(0.937131435)
2. Neighborhood_NoRidge(0.54646509)
3. Neighborhood_NridgHt(0.360928649)
4. GrLivArea(0.354855141)
5. Neighborhood_Crawfor(0.291513759)

Top 5 new variables after creating new model are:

1. 2ndFlrSF(0.402004523)
2. 1stFlrSF(0.273898024)
3. BsmtExposure_Gd (0.230974145)
4. Exterior1st_BrkFace(0.213795032)
5. Exterior2nd_ImStucc(0.192449674)

Q4. 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.To make the model robust and generalisable, we need to get rid of outliers and use robust error metrics. To overcome the problem of outliers we could remove the outliers manually, do data transformation like use Log,exp etc. and we could also do capping of value at certain threshold.

The model should not overfit on train data set and should perform equally well with test data set.