Assignment-based Subjective Questions

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?**

   Ans. Optimal value of alpha for Ridge and Lasso are:

   Ridge – 3.0

   Lasso – 0.0001

   When we double the value of alphas for ridge and lasso, the r2_score decreases little bit. Also, when we look at the value of coefficients, the values of coefficients are decreased. Also, lasso model penalises and makes the best feature of the model(with original alpha value) 0.

   R2 scores without doubling the value of alpha:

	Ridge	Lasso
Train R2	0.774	0.775
Test R2	0.786	0.785

   R2 Scores after doubling the value of alpha:

	Ridge	Lasso
Train R2	0.770	0.772
Test R2	0.785	0.784

   The values for our significant predictor variables are decreased after doubling the value of alpha.

	Feature	Lasso	lasso(double_alpha)	Ridge	ridge(double_alpha)
7	OverallQual_Excellent	0.194	0.000	0.183	0.022
14	GarageCars	0.136	0.046	0.131	0.056
2	Neighborhood_NoRidge	0.112	0.109	0.105	0.097
11	FullBath	0.095	0.043	0.093	0.040
12	BsmtFullBath	0.081	0.051	0.079	0.052

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?**

   Ans. First, we built model using all the features, then using lasso's chosen features, we built models for both Ridge and Lasso. Then, after repeating the process, at a point, lasso was not penalising any feature to make it 0. We consider those features and then using RFE we build the simpler model.

	Metric	Linear Regression	Ridge	Lasso
0	Train R2	0.776	0.774	0.775
1	Test R2	0.784	0.786	0.785
2	Train MSE	0.003	0.003	0.003
3	Test MSE	0.002	0.002	0.002
4	Train RMSE	0.053	0.053	0.053
5	Test RMSE	0.049	0.049	0.049

We can see, all the metrics are almost similar for both ridge and lasso. But, **as lasso does feature selection and avoids overfitting, we will choose lasso model**.

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?**

Ans. After dropping the five most significant predictor variables, we can see the huge drop in the R2 scores. Also, as we removed the most significant features, MSE and RMSE values are increased.

	Metric	Linear Regression	Ridge	Lasso
0	Train R2	0.524	0.511	0.523
1	Test R2	0.490	0.487	0.491
2	Train MSE	0.006	0.006	0.006
3	Test MSE	0.006	0.006	0.006
4	Train RMSE	0.077	0.078	0.077
5	Test RMSE	0.076	0.076	0.076

After removing the five significant predictor variables, new top 5 variables are:

- Neighborhood_NridgHt
- Fireplaces
- Neighborhood_StoneBr
- 2ndFlrSF
- Exterior2nd_ImStucc

And the coefficients for the new significant features are:

	Feature	LRM	Ridge	Lasso
1	Neighborhood_NridgHt	0.169	0.146	0.166
7	Fireplaces	0.146	0.133	0.146
0	Neighborhood_StoneBr	0.160	0.109	0.153
9	2ndFlrSF	0.103	0.092	0.103
2	Exterior2nd_ImStucc	0.146	0.066	0.130

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?**
   Ans. Simple models are more robust and generalisable. According to Occam's Razor, the model should be as simple as possible but not simpler as it will tend to underfit. The model should have the low bias and low variance. The model should be able to perform well on the unseen data i.e. test data. The difference between training accuracy and testing accuracy should be smaller. To make the model robust and generalisable, we need to do EDA part carefully, as it will handle the missing values, outliers, and skewness of the data. We will also use the regularisation to make the model simpler by choosing optimal model that will neither overfit nor underfit. It penalises the model parameters so that it won't overfit. And, Lasso, a type of regularisation does feature selection by making the non-significant features 0. We can use residual analysis to check whether our model is actually performing well or there are some patterns we missed. We can use the above techniques to make the model simpler because simpler models are more robust and can perform well on unseen data giving less errors on test data making it more generalisable. And when the model gives the less errors on unseen data, it impacts the accuracy positively. The accuracy of robust and generalisable model is always more than any model that is not generalisable.