HOUSING PRICE CASE STUDY

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?

- In my final mode, the optimal value for Ridge is 100 and Lasso is 0.01.
- If we double the value of alpha for both Ridge and Lasso, there will be no changes in the model.
- In Ridge regression, the important variables are:
    1. **KitchenQual**
    2. **MSSubClass**
    3. **Neighborhood**
    4. **GarageArea**
    5. **MSZoning**
- In Lasso regression, the important variables are:
    1. **MSSubClass**
    2. **OverallQual**
    3. **MasVnrArea**
    4. **BsmtQual**
    5. **BsmtFinSF1**

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?

I would prefer to apply Lasso regression for the below mentioned reason:

- The regularisation value i.e. alpha is 0.01 which is much lower compared to alpha value of Ridge. This helps us to achieve low variance along with low bias and hence is better.
- R2 value for both the model is very similar and have difference of 2 points only. Though R2 value for Lasso is slightly lower, but it has feature elimination property which helps us to predict with low number of variables and similar performance.

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.883812	0.863882
1	R2 Score (Test)	0.863985	0.852896
2	RSS (Train)	18.647309	21.845858
3	RSS (Test)	9.802399	10.601602
4	MSE (Train)	0.135144	0.146276
5	MSE (Test)	0.149599	0.155578

Hence, Final model is Lasso with R2 score of 86 in train data and 85 in test data.

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?

The five most important predictor variables in Lasso model are –

1.	OverallQual
2.	OverallCond
3.	Fireplaces
4.	BsmtFinType1
5.	SaleCondition

If we remove these and rebuild the model, the five most important predictor variables now are –

1.	HeatingQC
2.	GarageType
3.	Functional
4.	BsmtQual
5.	YrSold

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?

To ensure the model is robust and generalizable, the bias and variance should be low for the model. Also, it should be overfitting and it should be simple.

If the model over fits, the accuracy will be high on training set but low on testing set and it is no longer generalizable. Hence accuracy should be good and similar in both training and testing set and then only it will be generalizable and robust.