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?

Ans:
i)
Optimal value of ridge regression = 6
Optimal value of ridge regression = 0.0001

ii)
a) With alpha for Ridge changing from 6 to 12, R2 score dropped from
0.934303 to 0.92783 for train data
0.892292 to 0.88942 for test data.

b) With alpha for Lasso changing from 0.0001 to 0.0002, R2 score dropped from
0.944549 to 0.939237 for train data
0.895919 to 0.893725 for test data

iii) Most important Predictor variables after implementing change are:

Lasso

GrLivArea-0.369072
RoofMatl_WdShng-l0.214055
MSZoning_FV-0.177251
OverallQual_Excellent-0.159879
Neighborhood_StoneBr-0.151570

Ridge

GrLivArea-0.194211
OverallQual_Excellent-0.113637
1stFlrSF-0.099668
Neighborhood_StoneBr-0.093984
OverallQual_Very Good-0.088172

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?

Ans: I would go with Ridge regression as the deviation between R2 Score of train and test is lesser with ridge regression and hence provides better prediction as compared to Lasso. Although the number of predictor variables chosen by ridge is 292 whereas lasso is 218, we are choosing the top few predictors for our prediction. Hence choosing Ridge regression model.

Question 3

Ans: After building the model, you realized 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?

Top 5 variables rebuilding the model:

	Lasso
1stFlrSF	0.259284
OverallQual_Excellent	0.187333
OverallQual_Very Good	0.118594
Neighborhood_StoneBr	0.115292
Neighborhood_Crawfor	0.114022

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?

Ans: When building a model with many predictors, it is advisable to start with a simple model. The model can be rebuilt to increase the complexity as we go.

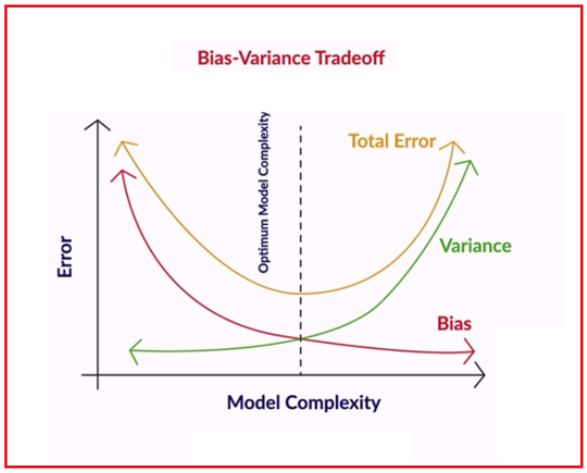
During model evaluation, several tests should be performed to make sure the model is not capturing random effects in the dataset. We can have separate set of unseen test/validation data as well as techniques like K-fold cross validation can be helpful to overcome it.

Simple model helps in:
➢ Prevents Overfitting: A high-dimensional dataset having too many features can sometimes lead to overfitting (model captures both real and random effects).

➢ Interpretability: An over-complex model having too many features can be hard to interpret especially when features are correlated with each other.

➢ Computational Efficiency: A model trained on a lower-dimensional dataset is computationally efficient (execution of algorithm requires less computational time).

Bias Variance Tradeoff:



If a model is simple and have a smaller number of features, then it may have high bias and low variance, in contrast, if a model has huge number of features, then it may have low bias and high variance. So, as the bias increases variance decreases and vice-versa. So, we need to get a model which has low bias as well as low variance. That is why the trade-off is required.

Bias can be minimized by training with more data and variance can be reduced by using Ridge/Lasso regularization methods.

An optimal balance of bias and variance would never overfit or underfit the model. Therefore understanding bias and variance is critical for understanding the behavior of prediction models.