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 to 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
0.934390 to 0.927772 for train data
0.896722 to 0.895274 for test data.
b) With alpha for Lasso changing from 0.0001 to 0.0002, $R2$ score dropped from train but increased slightly for test
0.944886 to 0.93943 for train data
0.896294 to 0.897006 for test data

iii) Most important Predictor variables after implementing change are:
Lasso

	Lasso
GrLivArea	0.376833
MSZoning_FV	0.252234
MSZoning_RL	0.208553
RoofMatl_WdShngl	0.193319
MSZoning_RM	0.184348

Ridge

	Ridge
GrLivArea	0.208251
OverallQual_Excellent	0.106975
1stFlrSF	0.106616
Neighborhood_StoneBr	0.098718
OverallQual_Very Good	0.086406

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 293 whereas lasso is 294, we are choosing the top few predictors for our prediction. Hence choosing Ridge regression model.

Question 3

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?

Ans:

Top 5 variables rebuilding the model:

	Lasso
1stFlrSF	0.283976
OverallQual_Excellent	0.164377
Neighborhood_StoneBr	0.120867
Neighborhood_Crawfor	0.112007
OverallQual_Very Good	0.109797

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.

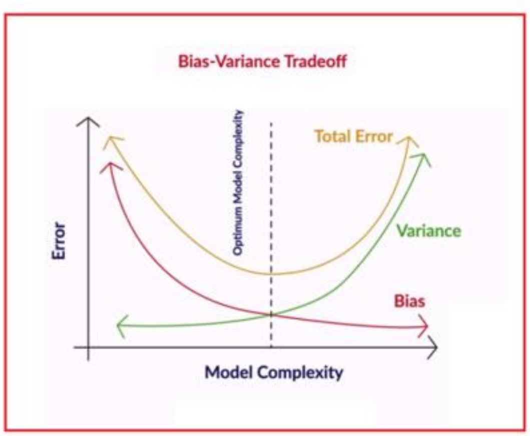
Simple model helps in:

➢ 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.