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?

Answer:

- The optimal value for Lasso regression model is 0.00025 and the optimal value for ridge regression model is 6.0.
- With initial alpha values, performance of Lasso model on train data is 91.9 % and ridge model is 92.0 %. When we double the values of alpha, performance of Lasso model on train data is 89.2% and ridge model is 90.9%. So, we observe that there is no much difference on the performance when we double the value of alpha.
- After the change is implemented the most important predictor variables are as below:

	Ridge	Lasso	abs_value_coeff
GrLivArea	0.056417	0.250914	0.250914
OverallQual	0.075004	0.190163	0.190163
GarageCars	0.041570	0.081807	0.081807
OverallCond	0.048961	0.063175	0.063175
TotRmsAbvGrd	0.054161	0.052203	0.052203
BsmtFullBath	0.031257	0.040670	0.040670
Neighborhood_NridgHt	0.030700	0.034664	0.034664
FullBath	0.044725	0.034461	0.034461
Neighborhood_Crawfor	0.037078	0.030615	0.030615
MSSubClass	-0.020926	-0.028679	0.028679

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?

Answer:

- With Lasso model, R2 score on training set is 91.9% and R2 score on testing set is 85.6% which is good. Model has performed well on train and test data.
- With Ridge model, R2 score on training set is 92.0% and R2 score on testing set is 87.8% which is good. Model has performed well on train and test data.
- Model evaluation results for both lasso and ridge model on train data as well as test data are at par. Ridge model is performing slightly better.
- We have evaluated models based on 257 features. Since most of the features are eliminated by Lasso, hence we will select Lasso as the best model.

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?

Answer:

Five most important predictor variables in the previous selected lasso model are **GrLivArea**, **PoolQC_Gd**, **OverallQual**, **OverallCond**, **Condition2_PosN**. After excluding these variables, if we rebuild the model then we get below top 5 most important predictor variables.

	Lasso	abs_value_coeff
1stFlrSF	0.260714	0.260714
2ndFlrSF	0.127221	0.127221
GarageCars	0.068352	0.068352
TotRmsAbvGrd	0.062453	0.062453
Neighborhood_NridgHt	0.049279	0.049279

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?

Answer:

If the model is overfitted then it may perform very well on training data with highest accuracy but its performance on test data may not be as good as on training data. To avoid this, we use regularization techniques such as lasso and ridge by allowing some error in the model which will avoid overfitting and make model more robust and generalizable.

By making model robust and generalizable, accuracy of model will go slightly down but it will perform well on training data and testing data as well.