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:

1. Optimal value of alpha:
    a. Ridge regression: 0.1
    b. Lasso regression: 0.0002

2. Changes in the model when doubling the value of alpha:
   <u>Metrics before doubling alpha:</u>

Ridge Train Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	5.539625	0.881175	0.006015	0.077555

Ridge Test Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	4.614343	0.801684	0.011682	0.108083

Lasso Train Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	7.44416	0.840323	0.008083	0.089904

Lasso Test Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	3.657384	0.842812	0.009259	0.096225

<u>Metrics after doubling alpha:</u>

Ridge Train Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	6.163785	0.867787	0.006692	0.081808

Ridge Test Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	3.951345	0.830178	0.010003	0.100017

For Ridge regression,

- train R^2 score decreased from 0.88 to 0.86.
- test R^2 score increased to 0.83 from 0.80

Lasso Train Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	8.458646	0.818562	0.009184	0.095834

Lasso Test Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	3.875335	0.833445	0.009811	0.09905

Both train and test R^2 scores have decreased for Lasso regression.

3. Top 5 predictor variables after doubling alpha:

	Ridge (alpha=0.2)
Condition2_PosN	0.879829
RoofMatl_Tar&Grv	0.768722
RoofMatl_WdShngl	0.738253
RoofMatl_Membran	0.735579
RoofMatl_CompShg	0.717727

Top predictor for Ridge regression: Condition2_PosN (House near positive off-site feature--park, greenbelt, etc.)

	Lasso (alpha=0.0004)
GrLivArea	0.678989
Condition2_PosN	0.482891
OverallQual	0.364254
LotArea	0.266483
GarageCars	0.225172

Top predictor for Lasso regression: GrLivArea (Above grade (ground) living area square feet)

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:

For the optimal value of alpha that was determined using GridSearchCV, looking at the test metrics Lasso regression performs better than ridge regression. The R^2 score is high and MSE and RMSE are low for Lasso. Also, Lasso regression penalizes the features even more as compared to ridge regression and some will be completely eliminated by making their coefficients zero. Hence, my preference is Lasso regression.

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:

The top 5 predictors for the Lasso regression model are:

	Lasso (alpha=0.0002)
Condition2_PosN	0.764465
GrLivArea	0.751140
OverallQual	0.340240
LotArea	0.330049
RoofMatl_Tar&Grv	0.310086

After removing these 5 features from the training set, the new model's top 5 predictors are:

Lasso Train Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	9.813501	0.7895	0.010655	0.103224

Lasso Test Metrics:

	RSS	R2 Score	MSE	RMSE
metrics	4.38329	0.811614	0.011097	0.105342

The top 5 features of the new model are:

	New Lasso
1stFlrSF	0.705217
2ndFlrSF	0.297066
GarageCars	0.291193
house_age	0.288948
TotalBsmtSF	0.259122

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:

To make a model robust and generalisable, the model must not be complex. i.e., we must limit the number of features that are used to build the model. Unnecessary or redundant features must be removed. Small changes in the data must not impact the model's performance (low variance). The train accuracy or other metrics of a complex model will be high due to low bias and the test accuracy will be low due to high variance. This is opposite for a simpler model which has a higher bias and low variance. Complex models can also suffer from overfitting the training dataset. This can be overcome by applying regularization on the model which will bring down the model's coefficients which in-turn will reduce the variance of the model.