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?

The Optimal values obtained are:

- **Ridge - 20.0**
- **Lasso - 0.001**

If we double the values of hyperparameter alpha for both Ridge and Lasso the values of coefficients will further reduce; for Lasso some of the coefficients may also attain the value of 0. As the value of alpha increases, the model complexity reduces, and the model tends to underfit.
If we double the value for Ridge and Lasso the coefficients will further reduce, e.g., below are the coefficients for top 10 Predictors after the change is implemented, we can see the values are further reducing.

LotArea	0.037389
LandSlope	0.007770
HouseStyle	-0.006454
OverallQual	0.080863
OverallCond	0.046911
YearBuilt	-0.041591
YearRemodAdd	-0.018736
MasVnrArea	0.000025
BsmtQual	0.019776
BsmtExposure	0.012965

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?

The Optimal values obtained are:

- **Ridge - 20.0**
- **Lasso - 0.001**

Ridge (20.0):

RMSE for Train: 0.12045961291211522
RMSE for Test: 0.14354572102869353

Lasso (0.001):

RMSE for Train: 0.12058278641557757
RMSE for Test: 0.1419645887272507

After comparing the values for RMSE, the Lasso model has slight advantage over Ridge, also for Lasso model many of coefficients e.g., Exterior2nd_VinylSd have become 0 which will help make the model simpler and generalize well on un-seen 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?

For the best model received during Lasso Regression if the top 5 variables are not available in the incoming data, we need to pick up next 5 based on the coefficient values.

YearBuilt	-0.044849
YearRemodAdd	-0.018355
MasVnrArea	0.002519
BsmtQual	0.017236
BsmtExposure	0.014130

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

1. Treat Outliers effectively so model can generalize well.
2. Make sure to scale the data before building the model.
3. Use a better value for hyperparameter so the Variance and Bias are low at the same time.
4. If predictors or target variable doesn't show any linear trend, see if transformation helps.
5. Use simpler Models with adequate number of features.
6. Use Regularization techniques to make sure the Model doesn't overfit.

If the model doesn't generalize well and is very complex it will not perform well in predicting values for un-seen data.