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?

- Optimal value of alpha for ridge: 10
- Optimal value of alpha for ridge: 0.001

After choose double the value of alpha for ridge, alpha will be 20

The comparison between the result of the Ridge model with alpha =10 and alpha=20:

Ridge with alpha=10

score	Train_ridge	Test_ridge
r2_score	0.9003	0.8946
MSE	0.0157	0.0173
MAE	0.0826	0.0889
RMSE	0.1252	0.1317

Ridge with alpha=20

score	Train_ridge	Test_ridge
r2_score	0.8957	0.8939
MSE	0.0164	0.0175
MAE	0.084	0.089
RMSE	0.128	0.1321

Observations:

The train r2_score of the ridge regression model alpha=10 is higher in comparison to the train r2_score of the ridge regression model alpha=20

The test r2_score of the ridge regression model alpha=10 is almost equal to the test r2_score of the ridge regression model alpha=20

In case of other metrics like (MSE, MAE, RMSE) ,there are very small change in train result if alpha value will increase but we can see the value of all the metrics are near about same for test result even if alpha value will increase.

The comparison between the result of the Lasso model with alpha =0.001 and alpha=0.002:

Lasso with alpha=0.001

score	Train_lasso	Test_lasso
r2_score	0.8918	0.8924
MSE	0.017	0.0177
MAE	0.0858	0.0896
RMSE	0.1304	0.133

Lasso with alpha=0.002

score	Train_lasso	Test_lasso
r2_score	0.882	0.8858
MSE	0.0186	0.0188
MAE	0.0897	0.0923
RMSE	0.1362	0.1371

Observations:

The train r2_score of the lasso regression model alpha=0.001 is higher in comparison to the train r2_score of the lasso regression model alpha=0.002

The test r2_score of the lasso regression model alpha=0.001 is also higher in comparison to the test r2_score of the lasso regression model alpha=0.002

Increase in the value of alpha in the model leads to decrease the r2_score and increase the MSE value and RMSE value. This situation indicates that the coefficient of more number of features which may be significant for the model, became zero which means they became insignificant if the alpha value became double. So that we can conclude the original alpha value will be the good choice for Lasso model

The most important predictor variables after the increasing alpha for both model:

Ridge with alpha=20

	Features_ridge	Coefficient_ridge
6	GrLivArea	0.134885
1	OverallQual	0.101585
8	AgeofProperty	0.091861
19	Neighborhood_Edwards	0.076530
18	Neighborhood_Crawfor	0.070924
32	Condition1_Norm	0.067972
56	Exterior1st_BrkFace	0.059652
9	MSZoning_FV	0.059490
11	MSZoning_RL	0.059041
2	OverallCond	0.058244
75	SaleType_New	0.055193
0	LotArea	0.053588
26	Neighborhood_NridgHt	0.053478
41	Condition2_PosN	0.053454
5	CentralAir	0.051666

Lasso with alpha=0.002

	Features_lasso	Coefficient_lasso
6	GrLivArea	0.134718
1	OverallQual	0.113424
8	AgeofProperty	0.095212
18	Neighborhood_Crawfor	0.070281
32	Condition1_Norm	0.066498
2	OverallCond	0.059224
26	Neighborhood_NridgHt	0.051293
0	LotArea	0.048206
19	Neighborhood_Edwards	0.048100
7	GarageCars	0.045753
5	CentralAir	0.034962
11	MSZoning_RL	0.034518
3	BsmtFinSF1	0.030935
9	MSZoning_FV	0.030000
56	Exterior1st_BrkFace	0.029208

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?

- Optimal value of alpha for ridge: 10
- Optimal value of alpha for ridge: 0.001

score	Train_linear	Test_linear	Train_ridge	Test_ridge	Train_lasso	Test_lasso
r2_score	0.9391	0.8724	0.9003	0.8946	0.8918	0.8924
MSE	0.0096	0.021	0.0157	0.0173	0.017	0.0177
MAE	0.0714	0.0919	0.0826	0.0889	0.0858	0.0896
RMSE	0.0978	0.1449	0.1252	0.1317	0.1304	0.133

Referring to the above result we can see, the r2_score and MSE are near about same in case of test data for both Ridge and Lasso regression. So that we can choose Lasso over Ridge because Lasso gives feature selection option means it results in model parameters such that lesser important features coefficients become zero which leads to remove unwanted features from model without affecting the model accuracy and makes the model more generalized and simple and accurate.

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?

Five most important predictor variables in original Lasso Model were as follows:

```
array(['GrLivArea', 'OverallQual', 'AgeofProperty', 'Condition2_PosN',
       'Neighborhood_Crawfor'], dtype=object)
```

Five most important predictor variables in Lasso Model after creating another model excluding the five most important predictor variables in original Lasso Model were as follows:

```
array(['Neighborhood_IDOTRR', 'Neighborhood_MeadowV',
       'Neighborhood_OldTown', 'Neighborhood_Edwards',
       'Neighborhood_BrDale'], dtype=object)
```

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?

Check the below points to predict a model is robust and generalizable

1. If the train accuracy is very high compare to test accuracy, then we can conclude the model is overfit. By the implementing regularization techniques, we can control the trade-off between model complexity and bias which is directly connected the robustness of the model. Regularization helps in penalizing the coefficients for making the model too complex thereby allowing only the optimal amount of complexity to the model. It helps in controlling the robustness of the model by making the model optimal simpler.

2. We can use a model that's resistant to outliers. Tree-based models are generally not affected by outliers, while regression-based models are

3. We can use a robust error metric mean absolute difference instead of mean squared error to reduces the influence of outliers

4. We need to transform the data - If data has a very pronounced right tail, try a log transformation.

5. We need to reduce co-linearity between independent variable.