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:

Ridge: -

Optimal value = **10**

Top 5 features:

Feature	Coefficient
Neighborhood_Crawfor	0.105
Condition2_PosN	-0.095
Neighborhood_NridgHt	0.084
Neighborhood_Edwards	-0.079
Neighborhood_Somerst	0.071

```
r2 score on train data with ridge 0.9202295667375724
r2 score on test with ridge 0.8875514063482818
```

Now if we double the alpha i.e., 10*2 = **20**

Feature	Coefficient
Neighborhood_Crawfor	0.081
Neighborhood_Edwards	-0.068
Neighborhood_NridgHt	0.066
OverallQual	0.061
Neighborhood_Somerst	0.054

```
r2 score on train data with ridge 0.914084268937706
r2 score on test with ridge 0.8863727567551977
```

```
Most important predictor remains same but we see there is inclusion of
another variable "OverallQual"
```

Lasso: -

Optimal value = **0.001**

Top 5 features:

Feature	Coefficient
Condition2_PosN	-0.206
Neighborhood_Crawfor	0.116
Neighborhood_Somerst	0.097
Neighborhood_NridgHt	0.087
OverallQual	0.065

```
r2 score on train data with lasso 0.9062122102037385
r2 score on test data with lasso 0.8811710295777795
```

Now if we double the alpha i.e., 0.001*2 = **0.002**

Feature	Coefficient
Neighborhood_Crawfor	0.074
OverallQual	0.072
Neighborhood_Somerst	0.063
BsmtFullBath	0.053
Neighborhood_NridgHt	0.053

```
r2 score on train data with lasso 0.8905643372018106
r2 score on test data with lasso 0.8733664414166025
```

Most important predictor is now “**Neighborhood_Crawfor**”

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: We will prefer to choose Lasso because as we see the R2 score for both is almost close but in case of Lasso the model complexity will be less as it **uses only 43 variables**. Hence the model can be more generalized and will be less sensitive to new 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?

Answer:

After removing top 5 features of the existing model from Lasso these are the new top 5 predictors.

Feature	Coefficient
Neighborhood_Edwards	-0.116
MSZoning_FV	0.093
Neighborhood_IDOTRR	-0.086
Neighborhood_MeadowV	-0.07
Condition1_Norm	0.07

And these are new r2 scores:

```
r2 score on train data with lasso 0.8905645216501836
r2 score on test data with lasso 0.8650848278496368
```

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: A good model is always more generalisable and robust. If a model is more generalised it will be less sensitive to test data and will make generalised prediction, it will not behave very abruptly in case of extrapolation and outliers in test data.

If we focus more on reducing Bias then variance will go very high and if we focus more on variance then bias will be very high. So, this goes hand in hand and we choose a value where both are low so that the model is more stable and robust. This is called bias variance trade-off.

So now when we try to make the model more generalised than our accuracy might reduce but if we try to fit all the train data points then in case of any outliers our model will drastically fail to predict even a value which is close.

Hence, we will compromise on the accuracy but, in this case, there will not be significant difference between train and test data.