1. Question Number 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?

   Solution:
   (a) The optimal Value of alpha for Ridge Regression is 20
   (b) The optimal Value of alpha for Lasso Regression is 0.003

From the changes made in the jupyter notebook, we can see the following changes for Ridge Regression model,

i. The test accuracy of the original model is better than the doubled model.
ii. When alpha is increased R2 decreases
iii. The original model performs better than the model for which alpha is doubled

```
For Ridge Regression Model (Original Model, alpha=20.0):
 ****************************************

For Train Set:
R2 score: 0.8575943956994172
MSE score: 0.14240560430058286
MAE score: 0.2227237363333244
RMSE score: 0.37736667089262516

For Test Set:
R2 score: 0.8523231804725492
MSE score: 0.1520745125426828
MAE score: 0.24770984480962793
RMSE score: 0.3899673224036635
 ****************************************
```

```
For Ridge Regression Model (Doubled alpha model, alpha=20*2=40):
 ****************************************

For Train Set:
R2 score: 0.8497925887475557
MSE score: 0.15020741125244433
MAE score: 0.22629763161307567
RMSE score: 0.38756600889712234

For Test Set:
R2 score: 0.8506943511998044
MSE score: 0.15375184699815483
MAE score: 0.2503549435424015
RMSE score: 0.392112033732905
 ****************************************
```

In case of Lasso Model, we can observe the following changes:

```
For Lasso Regression Model (Original Model: alpha=0.003):
 ****************************************

For Train Set:
R2 score: 0.8540112695899239
MSE score: 0.1459887304100761
MAE score: 0.22653243394644185
RMSE score: 0.38208471627386

For Test Set:
R2 score: 0.8481650142241087
MSE score: 0.1563565055279494
MAE score: 0.24755630114301627
RMSE score: 0.39541940459207287
 ****************************************
```

```
For Lasso Regression Model: (Doubled alpha model: alpha:0.003*2 = 0.006)
 ****************************************

For Train Set:
R2 score: 0.8408634254613427
MSE score: 0.15913657453865734
MAE score: 0.23646689090161446
RMSE score: 0.398919258169692

For Test Set:
R2 score: 0.8398659632267915
MSE score: 0.1649026953702169
MAE score: 0.25670474123831694
RMSE score: 0.4060821288486073
 ****************************************
```

i. The R2 score of original model is better than the doubled alpha model
ii. Original model seems to perform better than the doubled alpha model
iii. Even though the R2 score seem to reduce for doubled alpha model, the MSE score is increasing.

From this we can say that the original model is better for both Lasso and Ridge Regression model.

The most important predictor variable after the change is implemented are as below: Only Top 10

1. Ridge Model (alpha = 40)
    a. Neighborhood_NoRidge
    b. Neighborhood_NridgHt
    c. OverallQual,
    d. GrLivArea
    e. MSSubClass_160
    f. Neighborhood_Edwards
    g. 1stFlrSF
    h. MSSubClass_120
    i. GarageCars
    j. Neighborhood_Crawfor

2. Lasso Model (alpha = 0.006)
    a. Neighborhood_NoRidge
    b. Neighborhood_NridgHt
    c. GrLivArea
    d. OverallQual
    e. GarageCars
    f. MSZoning_RM
    g. MSSubClass_160
    h. BsmtExposure
    i. KitchenQual
    j. Neighborhood_Edwards

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 Value of alpha for Ridge Regression is 20

The optimal Value of alpha for Lasso Regression is 0.003

```
For Lasso Regression Model (Original Model: alpha=0.003):
 ****************************************

For Train Set:
R2 score: 0.8540112695899239
MSE score: 0.1459887304100761
MAE score: 0.22653243394644185
RMSE score: 0.38208471627386

For Test Set:
R2 score: 0.8481650142241087
MSE score: 0.1563565055279494
MAE score: 0.24755630114301627
RMSE score: 0.39541940459207287
 ****************************************
```

```
For Ridge Regression Model (Original Model, alpha=20.0):
 ****************************************

For Train Set:
R2 score: 0.8575943956994172
MSE score: 0.14240560430058286
MAE score: 0.2227237363333244
RMSE score: 0.37736667089262516

For Test Set:
R2 score: 0.8523231804725492
MSE score: 0.1520745125426828
MAE score: 0.24770984480962793
RMSE score: 0.3899673224036635
 ****************************************
```

- From the two models above, we can see that the R2 of Ridge model is a bit better that that of Lasso.
- The MSE for the test set in Lasso Regression is a bit higher that that of Ridge Regression model, which tells that Ridge is performing better on

unseen data. Hence Ridge is a better model for the purpose of predicting prices than Lasso 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?

The 5 most most important features are as follows:

- Neighborhood_NoRidge
- Neighborhood_NridgHt
- GrLivArea
- MSSubClass_160
- OverallQual

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?

A robust model must follow certain criteria as mentioned below:

- A model with higher accuracy I.e. Greater that 70 %.
- VIF should be less than 5
- P-value should be less than 0.05.
- The model must be simple since simpler models are more robust
- Simpler models require less training samples than the complex ones
- Simpler models are generic in nature
- When making a model simple, it leads to Bias-Variance Tradeoff

- Bias helps to quantify the accuracy of the model, Variance is the degree of changes in the model with respect to training data.
- From the graph below, we can see that with a balance in bias and variance the error can be minimalized.