Name- Ayush Mehta (ayush10mehta@gmail.com)

Advance Linear Regression Assignment

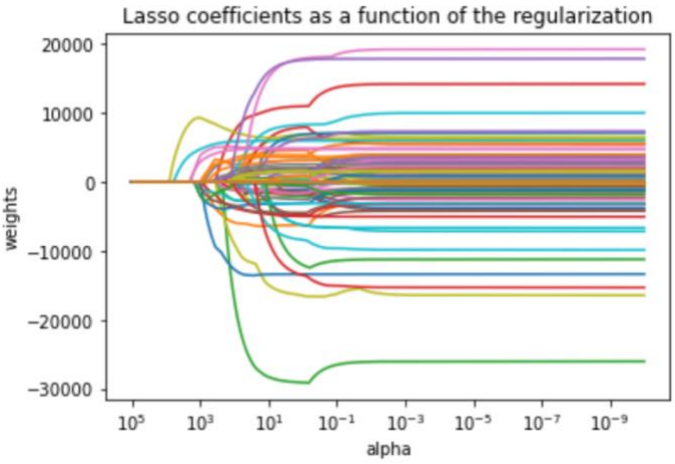
Assignment-based Subjective Questions

Question 1

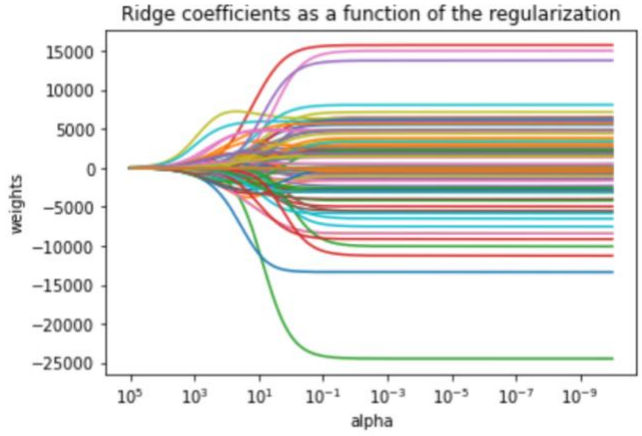
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Part 1

Optimum value for lasso regression: {'alpha': 62.06206206206207}



Optimum value for ridge regression: {'alpha': 8.0}



Part 2

Changes in the model if we choose double the value of alpha for both ridge and lasso

Lasso regression:

R2score on training data has decreased but it has slightly increased on testing data.

For value = alpha

```
Number of non-zero Coefficients 52
MSE Train 218273158.61031085
MAE Score Train 11061.373652187236
R2 Score Train 0.882094499492115

MSE Test 248944540.7579655
MAE Score Test 11067.029384621992
R2 Score Test 0.8459570913920883
```

For value = 2 x alpha

```
Number of non-zero Coefficients 45
MSE Train 225624273.60287145
MAE Score Train 11267.784238422619
R2 Score Train 0.8781236177858752

MSE Test 247015359.2128869
MAE Score Test 11047.963163203476
R2 Score Test 0.8471508381419941
```

Ridge regression:

R2score on training data has slightly decreased but it has slightly increased on testing data.

For value = alpha

```
MSE Train 216988695.224888
MAE Score Train 11022.107102290322
R2 Score Train 0.8827883333070768

MSE Test 249377520.83516553
MAE Score Test 11133.49503065163
R2 Score Test 0.8456891702307803
```

For value = 2 x alpha

```
MSE Train 220821582.38157606
MAE Score Train 11124.320967112148
R2 Score Train 0.8807179070509269

MSE Test 248416346.22743532
MAE Score Test 11125.42616207275
R2 Score Test 0.8462839297374721
```

Part 3

Most important predictor variables after the change is implemented are:

Lasso regression:

	Coefficient
OverallQual	7367.245376
MSZoning_FV	6447.701525
OverallCond	5850.291963
BsmtFullBath	5090.206360
Fireplaces	4601.336977

Ridge regression:

	Coefficient
MSZoning_FV	7898.298567
OverallQual	7023.122013
OverallCond	5905.664622
BsmtFullBath	4813.838495
Fireplaces	4664.553687

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 r2_score of lasso is slightly higher than ridge for the test dataset so we will choose lasso regression to solve this problem.

Lasso Regression

```
Number of non-zero Coefficients 52
MSE Train 218273158.61031085
MAE Score Train 11061.373652187236
R2 Score Train 0.882094499492115

MSE Test 248944540.7579655
MAE Score Test 11067.029384621992
R2 Score Test 0.8459570913920883
```

Ridge Regression

```
MSE Train 216988695.224888
MAE Score Train 11022.107102290322
R2 Score Train 0.8827883333070768

MSE Test 249377520.83516553
MAE Score Test 11133.49503065163
R2 Score Test 0.8456891702307803
```

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 model is again created excluding this 5 most important predictor variable:

	Coefficient
MSZoning_FV	8607.973313
KitchenQual_Ex	8229.472637
OverallQual	6916.956486
OverallCond	5901.526845
BsmtFullBath	4992.269562

The 5 important predictor variables know are:

	Coefficient
RoofStyle_Flat	23744.350482
BsmtQual_Ex	10763.344980
HouseStyle_1.5Unf	10737.101263
BsmtFinType1_GLQ	7144.991804
Fireplaces	6207.415139

Evaluation

```
Number of non-zero Coefficients 68
MSE Train 263788744.53388563
MAE Score Train 12190.285279088861
R2 Score Train 0.857508160185005

MSE Test 353399472.8999559
MAE Score Test 13968.093642785516
R2 Score Test 0.7813220465077815
```

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?

The model needs to be generalised to ensure that test accuracy does not fall short of training results. For datasets other than the ones that were used during training, the model should be accurate. The outliers shouldn't be given an excessive amount of weight in order to maintain the high level of model accuracy. Only those outliers that are pertinent to the dataset should be preserved after conducting the outliers analysis to verify that this is not the case. The dataset must be cleaned up of any outliers that don't make sense to preserve. Predictive analysis cannot be believed if the model is not robust.