CS 4375
ASSIGNMENT 1. Report

Names of students in your group:

Haeun Kim

Number of free late days used: None

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

Dataset:
https://archive.ics.uci.edu/ml/datasets/Auto+MPG
Course recourse ("Scikit-Learn Lab")
https://scikit-learn.org/stable/modules/preprocessing.html
https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e

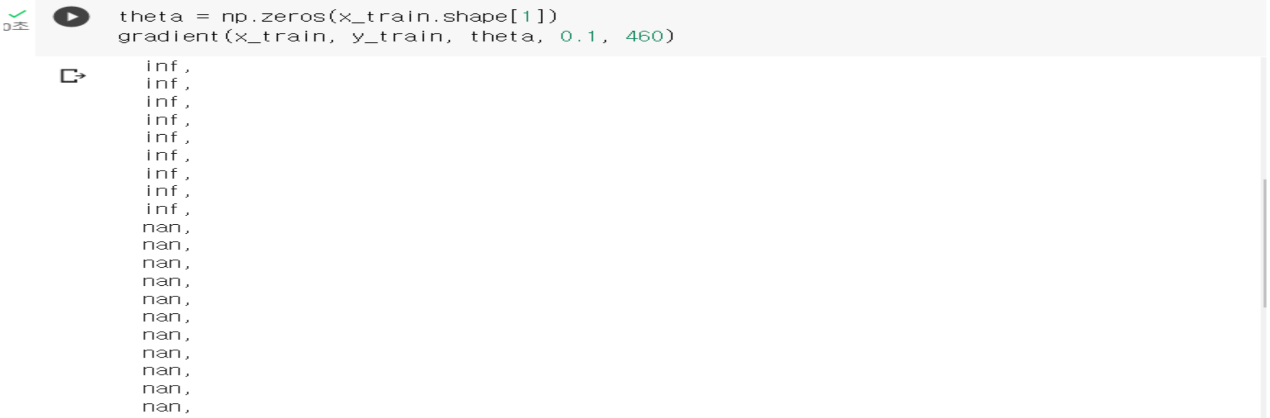
Part 1:

1.

I chose the dataset “Auto MPG” from the UCI ML Repository, and this dataset has 9 attributes (mpg, cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name) and 406 instances.

2.

According to the isna().sum() function, the ‘mpg’ has 8 NA and ‘horsepower’ has 6 NA, so I removed all the NAs and redundant rows by using ‘dropna()’ and ‘drop_duplicated’ respectively. The attribute ‘Car Name’ is string data type, and ‘Cylinders’, ‘Model Year’, and ‘ Origin’ are multi-valued discrete. Therefore, I removed all the attributes because I want to focus on the multi linear regression. I started this project without standarlization of the dataset, and I got the results below. So, I felt the need to standarliza and did the standarlization to this dataset.



```
theta = np.zeros(x_train.shape[1])
gradient(x_train, y_train, theta, 0.1, 460)
```

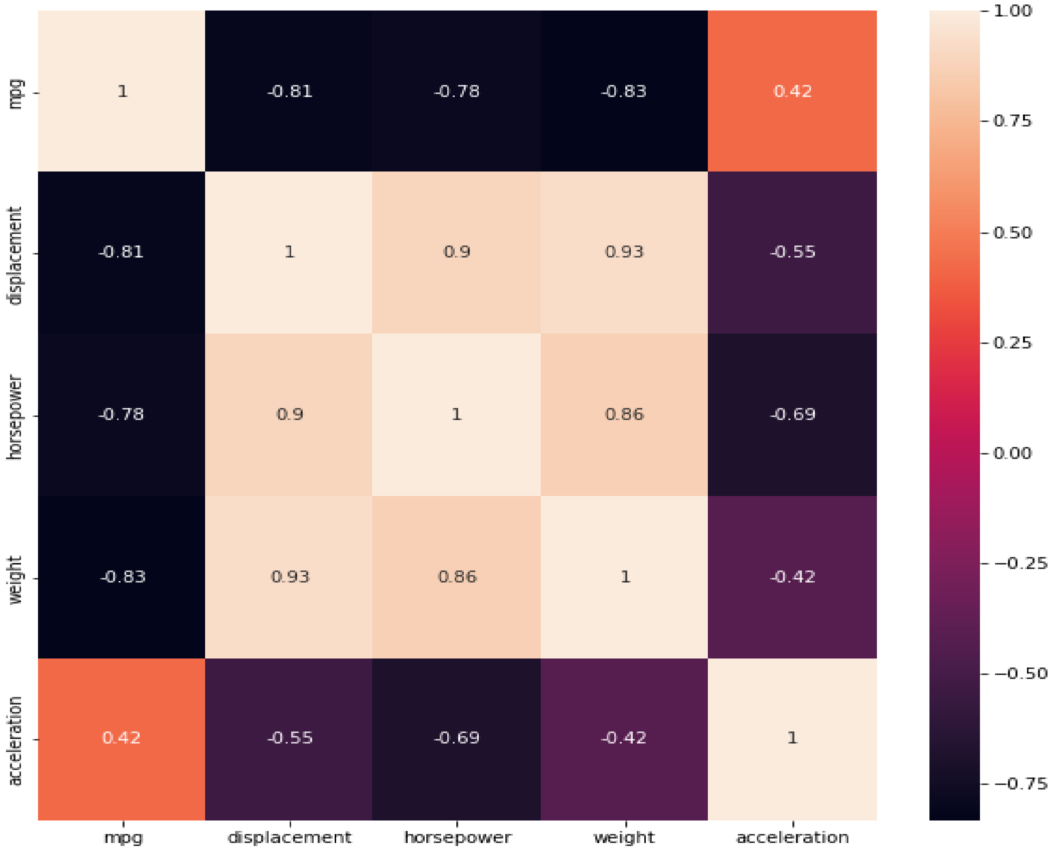
```
 inf,
 inf,
 inf,
 inf,
 inf,
 inf,
 inf,
 inf,
 inf,
nan,
nan,
nan,
nan,
nan,
nan,
nan,
nan,
nan,
nan,
nan,
```

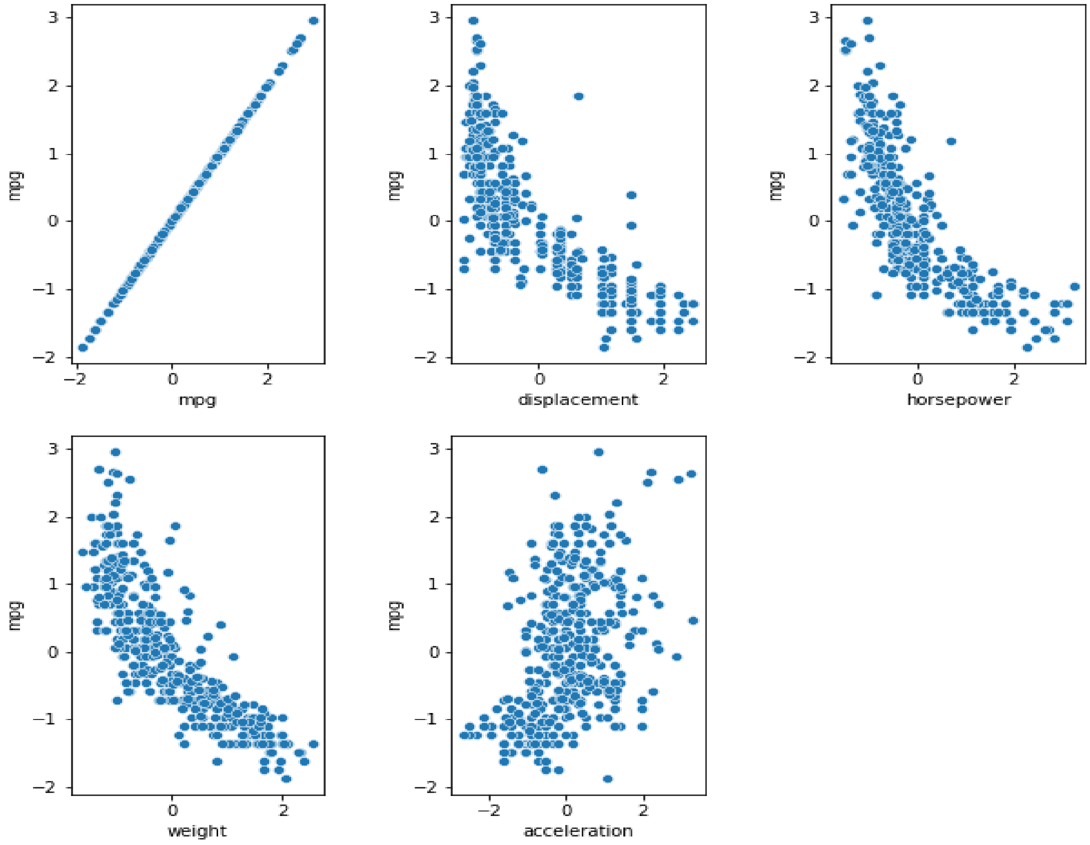
3.

I split the dataset into training and test part with 80/20 ratio.

Extra Work:

Before starting the linear regression, I wanted to figure out the relationship between the attributes. So, I made the heatmap with correlationship and the scatter plot between mpg and other attributes. I would like to set the ‘mpg’ as the Y response. With these plots, I could get the information that the correlationships ‘between displacement and mpg’ and ‘between weight and mpg’ have the negative relationship.


mpg
displacement
horsepower
weight
acceleration
1
-0.81
-0.78
-0.83
0.42
-0.81
1
0.9
0.93
-0.55
-0.78
0.9
1
0.86
-0.69
-0.83
0.93
0.86
1
-0.42
0.42
-0.55
-0.69
-0.42
1
mpg
displacement
horsepower
weight
acceleration
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


mpg
mpg
displacement
horsepower
weight
acceleration
3
2
1
0
−1
−2
−2
0
2

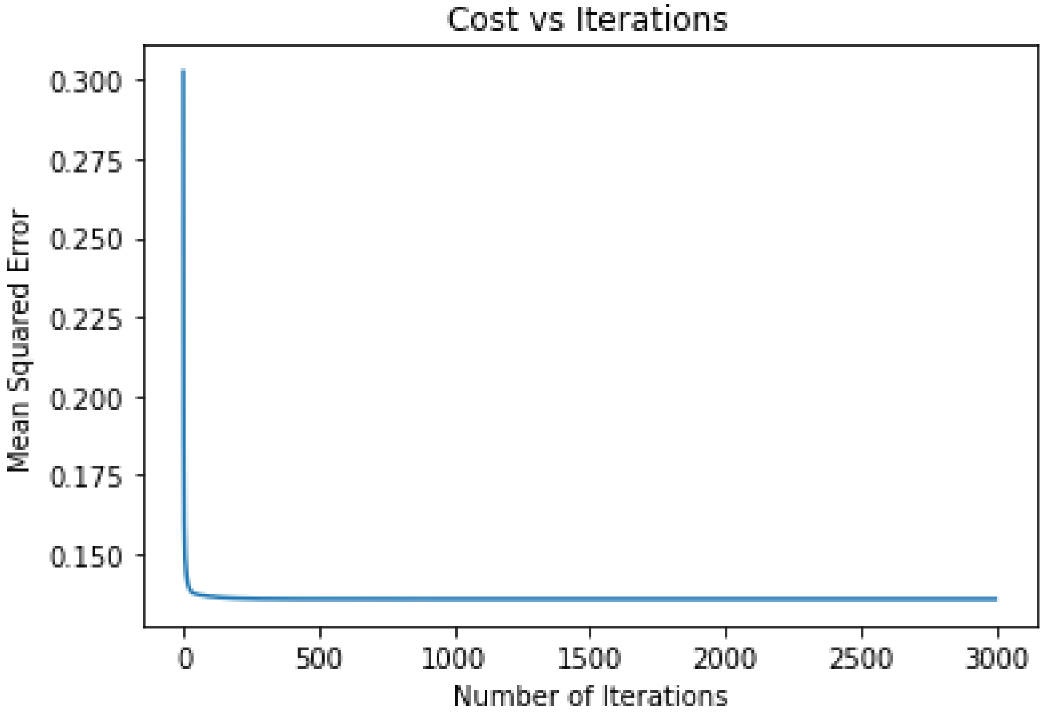
4. And I construct the linear regression model, and this below table is the output of RMSE and R^2 with each learning rate and the number of iteration, of training dataset. (I used round up function at 5 decimal point). The MSE and R^2 are converged to 0.27171 and 0.71029 respectively.

	MSE test	R^2 test	MSE train	R^2 train	Learning Rate	Iteration
1	1.20156	-0.09351	0.94288	-0.00536	0.000001	1000
2	1.1632	-0.05859	0.90745	0.03241	0.00001	1000
3	0.871	0.20733	0.64633	0.31083	0.0001	1000
4	1.01477	0.07649	0.77266	0.17613	0.0001	500
5	0.7623	0.30625	0.55458	0.40867	0.0001	1500
6	0.67966	0.38147	0.48776	0.47992	0.0001	2000
7	0.38523	0.64942	0.28043	0.70099	0.001	2000
8	0.38832	0.6466	0.27188	0.7101	0.01	3000
9	0.38762	0.64724	0.27171	0.71029	0.1	3000
10	0.38764	0.64722	0.27171	0.71029	0.01	10000

When I applied test dataset to the predicted theta, the MSE is 0.38764 and R^2 is 0.64722. The MSE is increased and R^2 is decreased in the test set when comparing to the training set, so I tried to adjust the learning rate and iteration, but whenever I changed the values, the MSE was more then 0.38762, and the R^2 was less than 0.64724. I found optimized the number of iteriation with numpy.argmin() is '2827'. So, I could conclude that I am satisfied that I have found the best solution with this dataset. Therefore, the optimized learning rate is 0.1 and the optimized iteration is 2827.

Part 2:

In the part 2, I did used the ML library that perfroms linear regression from Scikit Learn packages. I've tried this part with two method (one with SGDRegressor and the other with LinearRegression). And I could get the result that 'MSE is 0.27551' and 'R2 score is 0.70623' for training set and 'MSE is 0.39799' and 'R2 score is 0.63781' for test set with SGDRegressor. In addition, I could get the result that 'MSE is 0.27171' and 'R2 score is 0.71029' for training set and ‘MSE is 0.38762’ and ‘R2 score is 0.64724’ for test set with LinearRegression. The results of MSE and R^2 from part 1 and part 2 are similar to each other. So, I thought that I found the best solution.


Cost vs Iterations
Mean Squared Error
Number of Iterations