Question 2 Report (Linear Regression)

1. Analysis Of Data:

- At first, imported all the necessary libraries namely pandas, numpy, matplotlib,seaborn.
- Downloaded the automobile.csv data and read it through pandas. Found out the output column(price) and the input columns.
- Also roughly understood which are the columns on which the output heavily depends and which ones are least used by watching corelation in heatmap.
- Also observed that in some columns there are values as '?'.

2. Preprocessing Steps:

- At first, removed the rows which got output or 'price' as '?', I did not replace them with any value because there were only four rows like that.
- Then I selected numerical columns as input & I checked for the duplicate rows, and dropped them except the first instance.
- For univariate linear regression I used 'engine-size' as input feature because of its high corelation with 'price' column, which I saw from the heatmap.
- Then I spiltted the data into training and testing datasets. After this step, we will have 4 datasets, input_training (train_x), input_testing(test_x), output_training (train_y), and output_testing(test_y).
- For mutiple linear regression I selected four columns with high co-relation namely 'width', 'curb-weight', 'engine-size', 'compression-ratio' with 'price' column. Then splitted the dataset like above.

3. Code Approaches:

Univariate Linear Regression:

- At first, I did it using gradient descent, using learning rate as.0000001 and ran the for loop 2000 times.
- Then plotted the cost values with respect to iterations.
- Finally tested the model on test data.
- Then I also used the closed form approach as shown by sir in the class.(lecture 3 slides)
- Then also I tested the model on test data.

Multiple Linear Regression :

- At first I did it in closed form usig the normal equation
- Then tested the model and got better result than univariate.
- Again I did it using gradient descent & ran the for loop for 2000 times with learning rate .00000001

- Then plotted the cost values with respect to iterations.
- Note: I have done array reshaping in the code.
- Finally I tested the model with respect to test data.

4. Test Result of the Model:

Univariate Gradient Descent: Root Mean Squared Error - `8892.721526208605`

Univariate Closed Form: Root Mean Squared Error - `8870.149305866245`

Multivariate Gradient Descent: Root Mean Squared Error - `1283.958653181554`

Multivariate Closed Form: Root Mean Squared Error - `1186.6923022870083`