**README**

**Dataset**

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.

The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes.

**Data Attributes :-**

1. mpg: continuous   
2. cylinders: multi-valued discrete   
3. displacement: continuous   
4. horsepower: continuous   
5. weight: continuous   
6. acceleration: continuous   
7. model year: multi-valued discrete   
8. origin: multi-valued discrete   
9. car name: string (unique for each instance)

**Libraries Used**

1. Pandas
2. NumPy
3. Scikit Learn
4. Matplotlib

**Data Preprocessing**

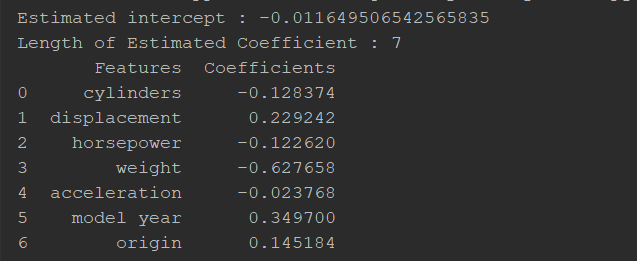
1. Dropped car name column from data frame as it car name doesn’t plays any significant role in predicting mpg.
2. Scaled data using Scikit learn preprocessing module because horsepower attribute has significantly larger values than other attributes.
3. Data split using Sklearn’s train\_test\_split() of model\_selection module in the ratio of 65: 35.

**Compiling the Code**

To compile the code just run ```python LinearRegression\_part2.py```

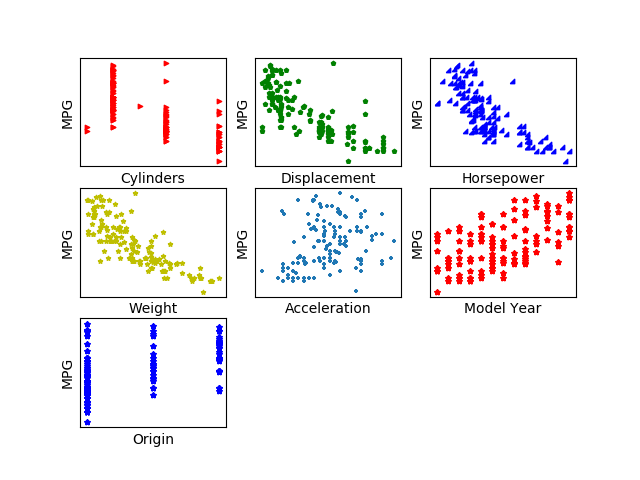
**Data Training**

Trained model with training data using Scikit Learn’s linear regression module. Computed estimated intercepts and coefficients or weights ( attached below screenshot of estimated intercept and coefficient generated by the program ).



By the above image, it is clear that displacement, model year and origin attributes of data are positively correlated with mpg i.e, on increase of any of these values mpg value will also increase. Cylinders, horsepower, weight and acceleration attributes of data are negatively correlated with mpg i.e, on increase of any of these values mpg value will decrease.

Below is the scatter plot generated using mpg and other dependent attribute of data.

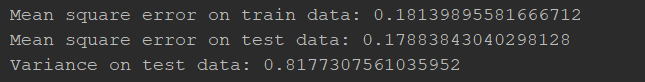


With the above scatter plot, we can easily see the correlation between input and output attribute.

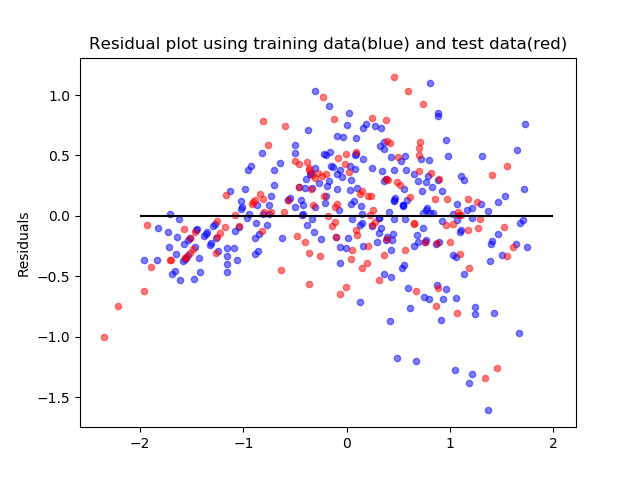
**Question:**

**Are you satisfied that the package has found the best solution. How can you check. Explain.**

From the attached screenshot of program output, we can see that mse value on test data is low and is comparable with mse value on train data. And variance on test data is also approaching 1.

****

In the residual plot also, we can see that the train data and test data is randomly scattered around line zero, indicating a good model.



Therefore, from mse value, variance value and residual plot we can say that the package has found best solution.