

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

ENCS5341

Machine Learning and Data Science

Assignment No.1

Studying the dataset and analyzing aspects

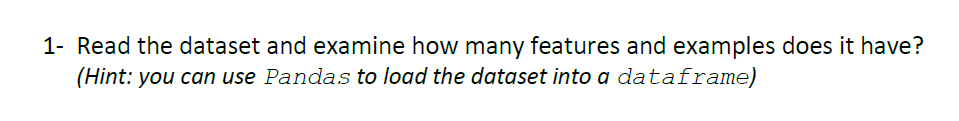
**Student’s Name:** Lojain Abdalrazaq. **ID:** 1190707.

**Instructor’s Name: Dr.** Yazan Abu Farha.

Section: 2.

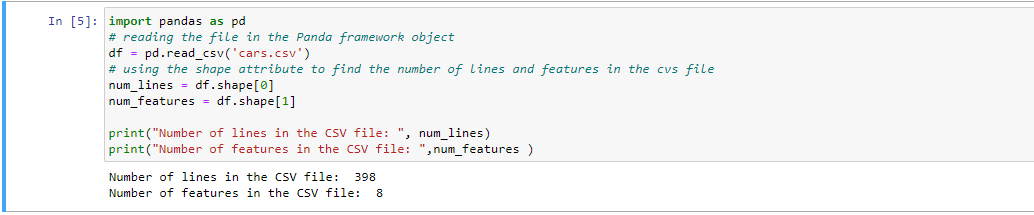
November 29, 2023

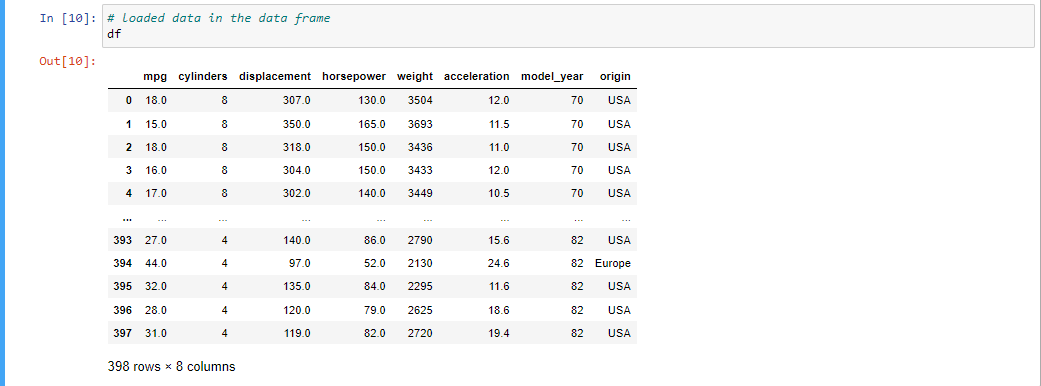
**Summarization the Results**

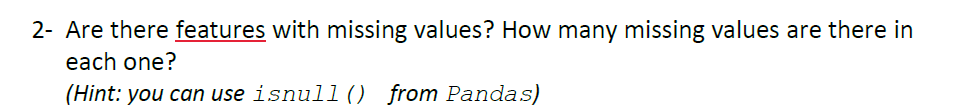
*Question 1:*

In this part the Panda was used , and the shape[0], and shape[1] were used to find the number of lines (records), and the number of features (columns) in the dataset.

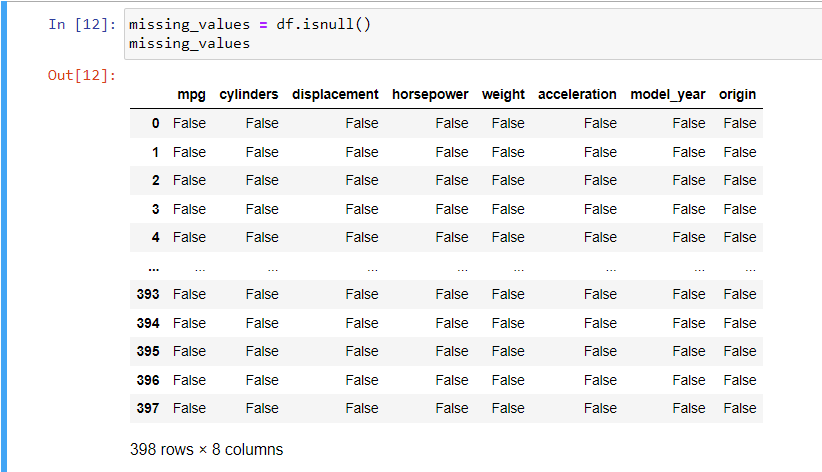
Number of lines in the CSV file: 398

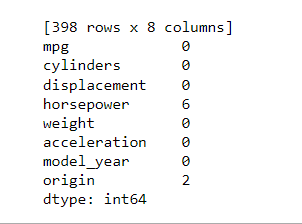
Number of features in the CSV file: 8

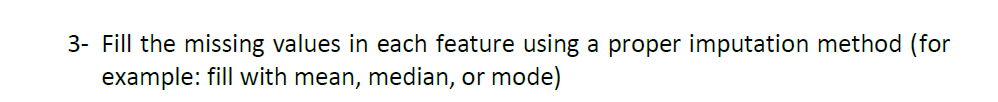
And here is the loaded dataset (398 records, and 8 features).

*Question 2:*

Now, in this part we want to know the total number of missing values in the dataset, this done by using the *isnull() function.*

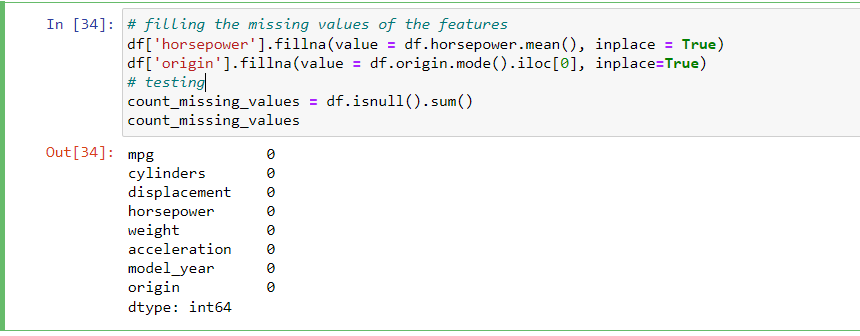
The following table represents the Boolean table such that the *false* value means the index is not null or missed, while the *true* value means the index has a missing value.

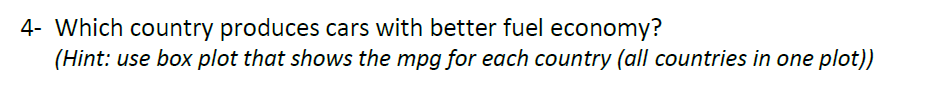
* And the number of missing values in 8.
* Horsepower: 6 missing values.
* **Origin: 2 missing values.

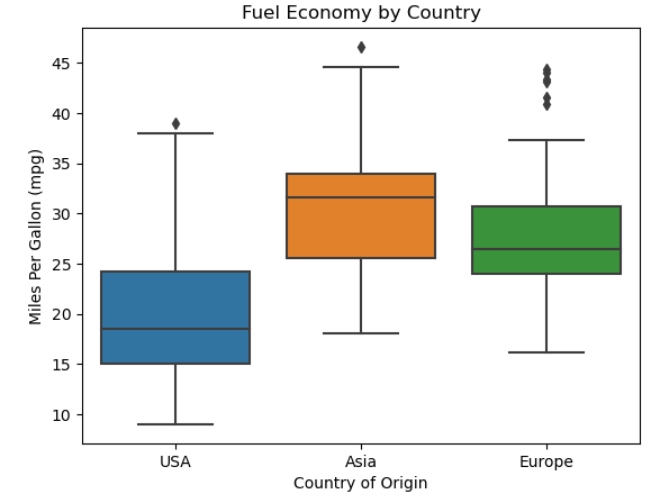
*Question 3:*

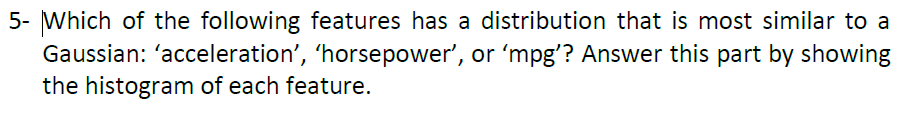
For the first feature (*Horsepower*) I used the mean to fill the missing values. While the (*origin*) feature, the mode was used which represents the most frequent value of the feature.

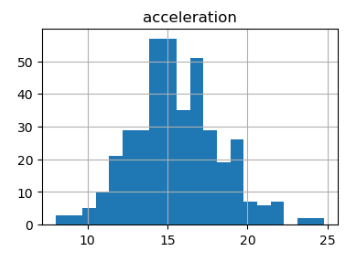
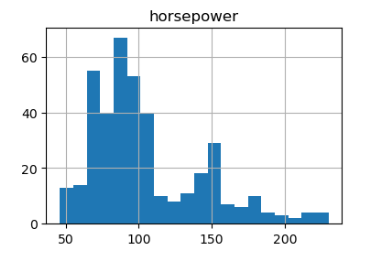
In addition, it was checked for a second time, and the number of missed values were 0 instead of 8 values.

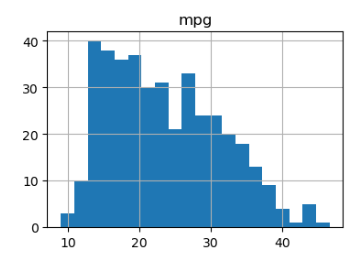


*Question 4:*

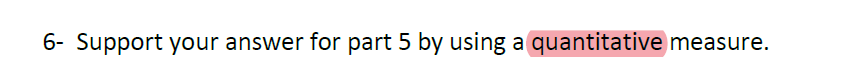
In this part, the following figure is the result from plotting the ‘country’ vs ‘mpg’ features in X and Y axis. And from the plot, it is notices that the Asia has the most economic cars, then Europe, and finally America.

*Question 5:*

The histogram of each feature is as the following:



**From the given histograms, the plot of the “acceleration” feature is the most similar to the Gaussian distribution.**

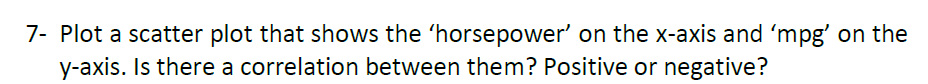
*Question 6:*

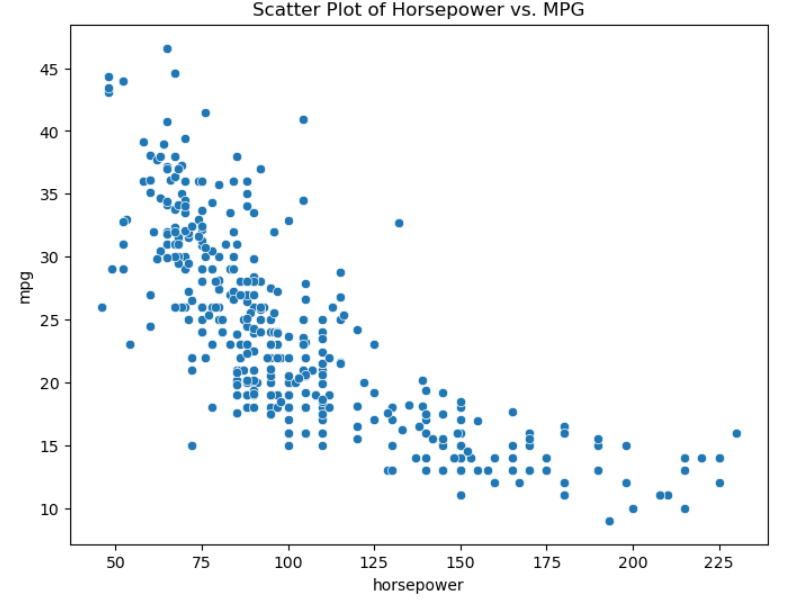
To check the normalization of the required features in quantitative measure, I used the p-value measure, such that of the P>=0.05, then the feature is normally distributed, while P<0.05 has a less chance to be normally distributed.

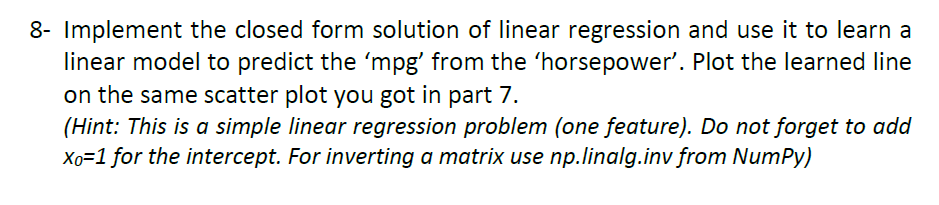
The P-Value of the 3 features is as the following:

* Acceleration: pvalue=0.039872437715530396
* Horsepower: pvalue=4.5655601017434086e-15
* Mpg: pvalue=1.1833407853600875e-07.

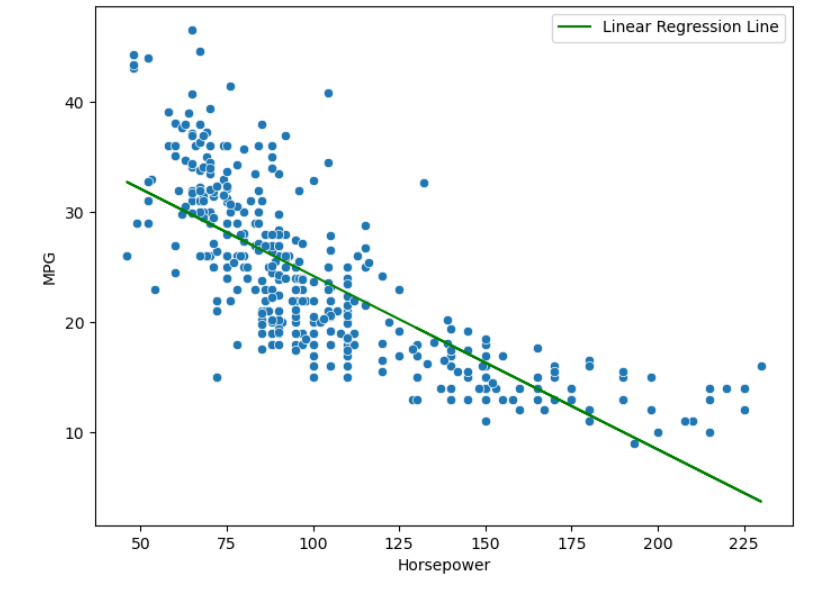
And it is noticed that the last two values are too small, while the Acceleration is the most close to the 0.05 value, so it is the most likely to be similar to the Gaussian distribution.

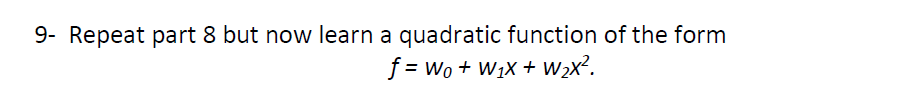
*Question 7:*

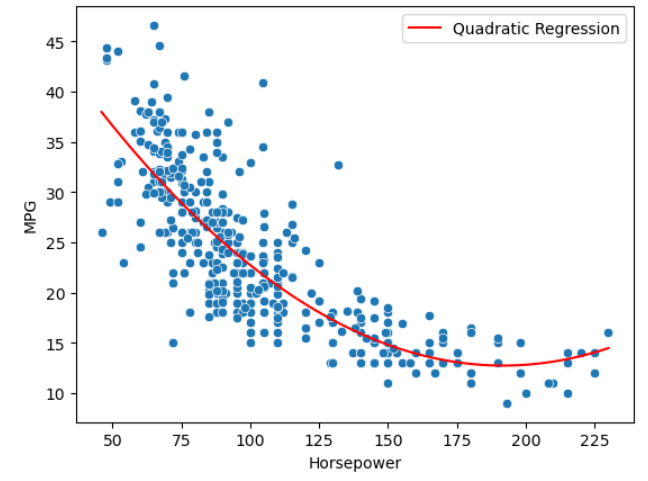
From the figure, it is noticed that there is a negative correlation between the “mpg” and the “horsepower”.

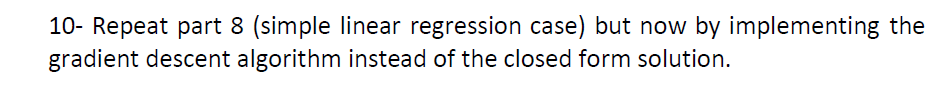
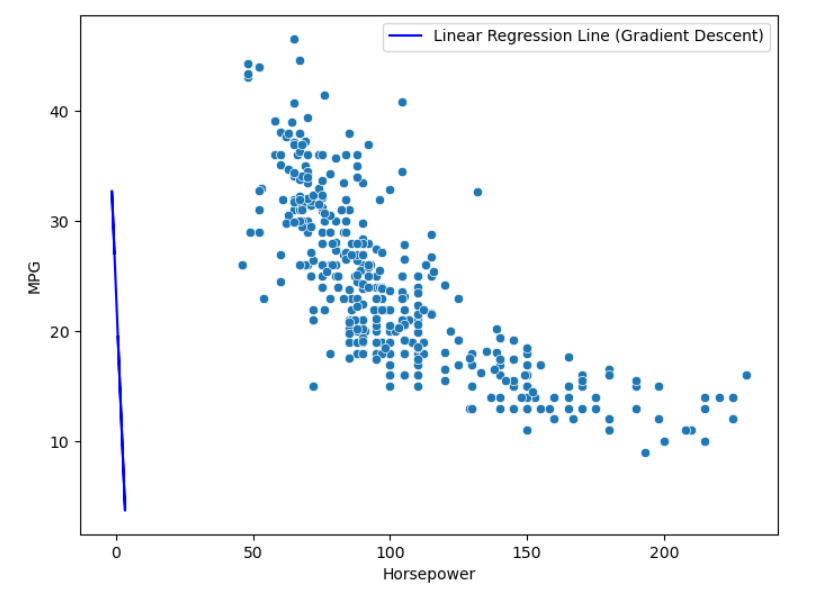
*Question 8:*

Using the linear system equation, the learned line of the simple linear regression is as the following:



*Question 9:*



*Question 10:*