BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department

Machine Learning and Data Science - ENCS5341

Assignment #1

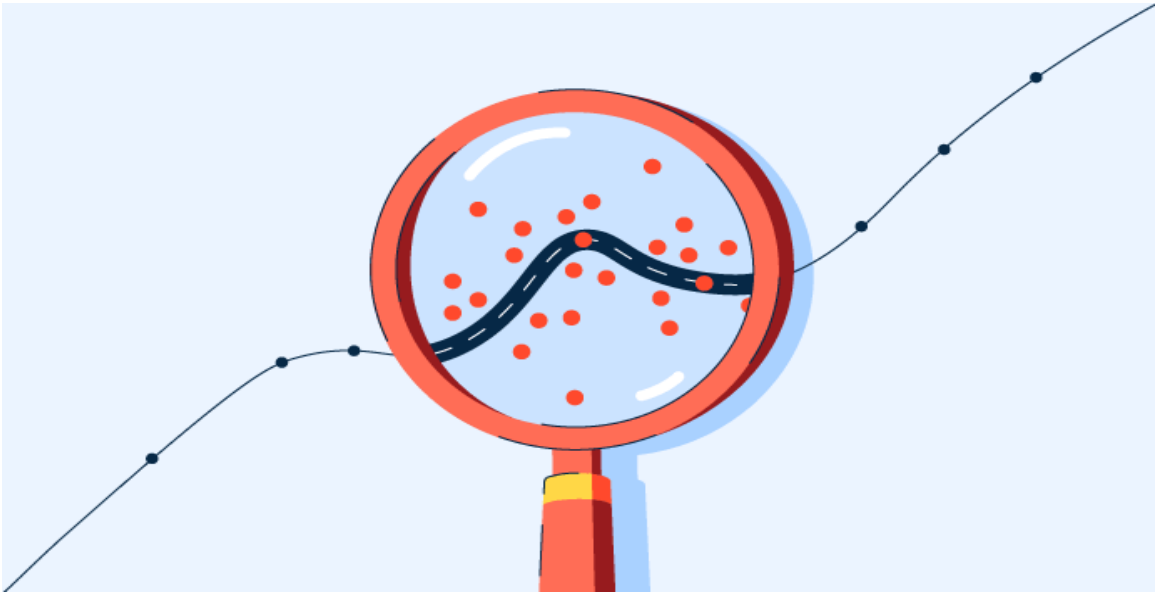
Prepared by: Dana Bornata.

ID: 1200284

Instructor: Dr. Yazan Abu Farha

Date :23/11/2023

Section #1



1)Read the dataset and examine how many features and examples does it have?

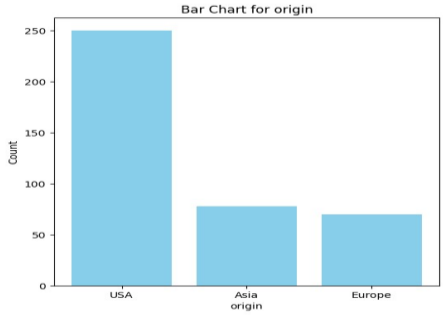
I took the data from the file and then calculated the number of examples through the number of colum and the number of features through a number of row

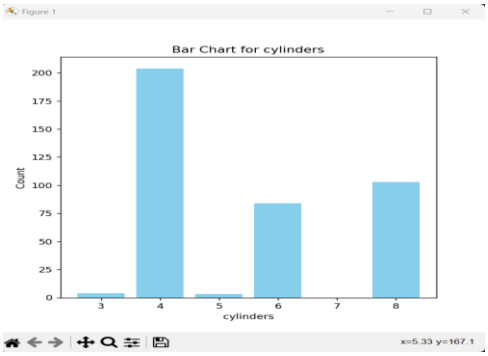
```
                1)
Hello, this is some of the data in the file:

    mpg  cylinders  displacement  horsepower  weight  acceleration  model_year origin
0  18.0          8         307.0       130.0    3504          12.0          70    USA
1  15.0          8         350.0       165.0    3693          11.5          70    USA
2  18.0          8         318.0       150.0    3436          11.0          70    USA
3  16.0          8         304.0       150.0    3433          12.0          70    USA
4  17.0          8         302.0       140.0    3449          10.5          70    USA

Number of features>>>>    8   >>>>>>>
Number of examples>>>>  398   >>>>>>
```

- **The number of features is 8 and the number of data is**





398

- **I used bar charts for some features to help understand the distribution of data quickly and simply.**

2) Are there features with missing values? How many missing values are there in each one?

When we do Forms of data pre-processing, we first check the data to see if there is missing data or not to fix this problem Because when there is a lack of data, big problems arise when using it. To find missing values using function in Python isnull () , We check if the number of values is greater than zero, then there are missing values, and if not, then it returns that there are no missing values.

```
*****************************************

                2)

Features with missing values and the num of missing values:
horsepower: 6
origin: 2
```

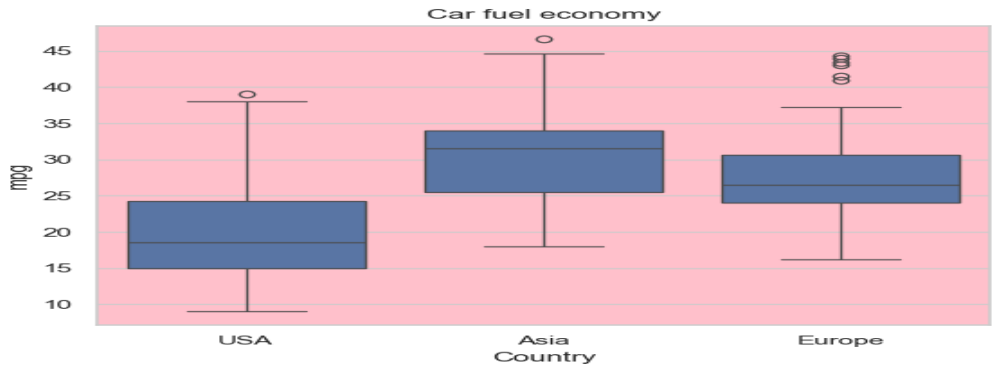
3) Fill the missing values in each feature using a proper imputation method

This happens in the data cleaning phase - losing values in pre-processing the data so that the data is usable and operations can be performed on it. Here we used the measure of central tendency of the attributes (e.g., mean or median) to fill in the missing items with a value, here the mean values were used to fill in the missing values in the numeric columns, and also if the columns are non-numeric, we fill them using the more common value method. I filled it using fillna.

So here I have compensated for missing values using the average of the numeric values for the numerical columns and the most frequent value for the non-numeric columns.

```
**Missing values after imputation **
mpg             0
cylinders       0
displacement    0
horsepower      0
weight          0
acceleration    0
model_year      0
origin          0
dtype: int64
```

4) Which country produces cars with better fuel economy?

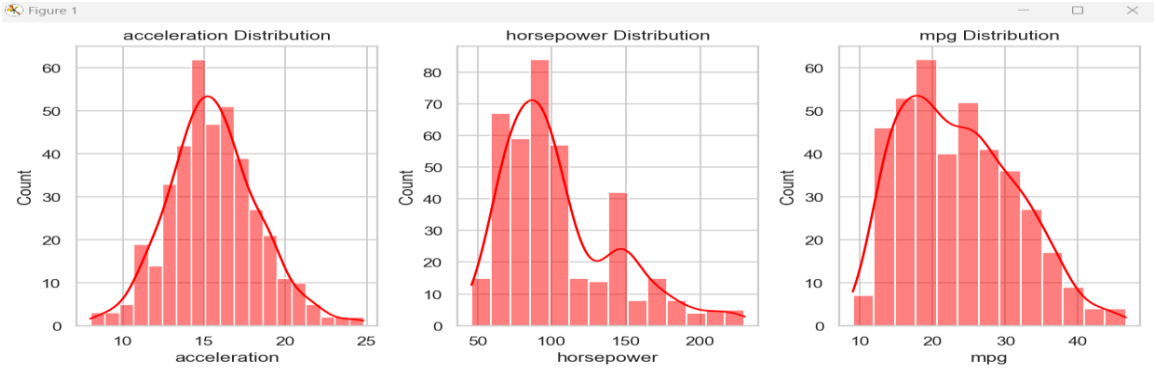



```
                4)
Median fuel economy by country:
USA: 18.55 mpg
Europe: 26.5 mpg
Asia: 31.55 mpg

The country with the best fuel economy is: Asia (31.55 mpg)
```

First: Box plots are a statistical visualization tool that provides a summary of the distribution of a data set. It displays key statistical measures, including median, quartile, and potential outliers. We found out which country produces cars with the best fuel economy, calculated the average for each country, and bigger is better.

5) Which of the following features has a distribution that is most similar to a Gaussian: ‘acceleration’, ‘horsepower’, or ‘mpg’?



A Gaussian distribution, also known as a normal distribution, is characterized by a bell-shaped curve. It has a mean and median at the center, a defined spread measured by the standard deviation, and a prevalence in various natural and statistical phenomena. In determining which feature ('acceleration', 'horsepower', 'mpg') resembles a Gaussian distribution, one would assess visual characteristics and statistical properties like mean and standard deviation. If a feature displays a bell-shaped curve and aligns with Gaussian distribution properties, it can be considered to have a distribution similar to Gaussian.

- Here, by using the histogram function to draw the curve ,The result was that the acceleration was closest to the shape of a Gaussian curve.

6) When we find out who is close to the Gaussian using a quantitative measure.

The purpose of calculating skewness and kurtosis is to examine the shape of the data distribution and understand how it deviates from a Gaussian (normal) distribution.

Skewness: Objective: Measures the asymmetry of the data distribution.

Kurtosis: Objective: Measures the thickness of the tails and the peakiness of the distribution.

To be close to a Gaussian distribution (Normal): Skewness close to zero suggests a symmetric distribution. Kurtosis around 3 indicates a distribution similar to Gaussian

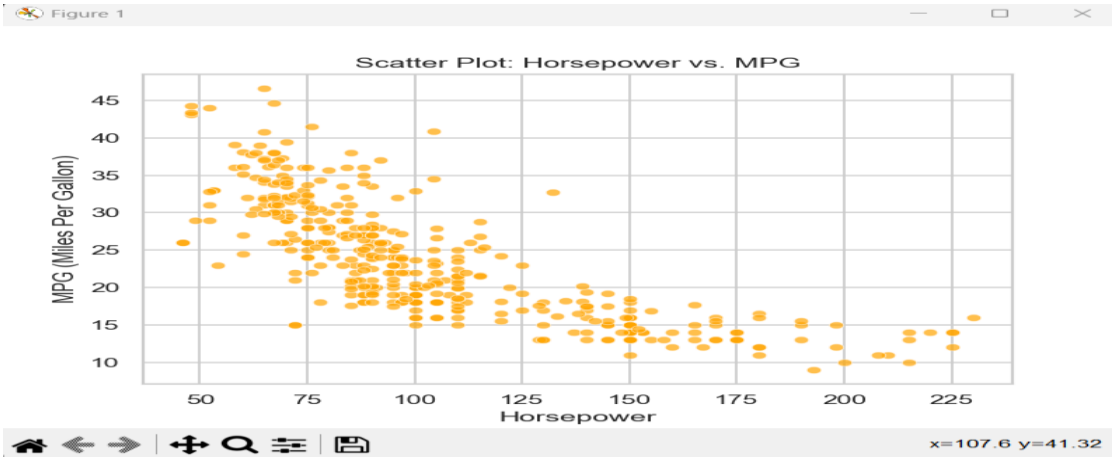
```
                6)

acceleration:
Skewness: 0.27772507624356363
Kurtosis: 0.3992077323931644
The distribution deviates from a Gaussian shape.
horsepower:
Skewness: 1.0914191838332945
Kurtosis: 0.7290385466123319
The distribution deviates from a Gaussian shape.
mpg:
Skewness: 0.45534192556309266
Kurtosis: -0.5194245405990445
The distribution deviates from a Gaussian shape.

The feature closest to a Gaussian distribution is:>>>>>>> acceleration
```

7) Plot a scatter plot that shows the 'horsepower' on the x-axis and 'mpg' on the y-axis. Is there a correlation between them? Positive or negative?

• Directions: All correlation coefficients between 0 and 1 represent positive correlations, while all coefficients between 0 and -1 are negative correlations. Positive relation means if one increase(decrease), then other will increase(decrease). Negative relation means if one increase(decrease), then other will decrease(increase).



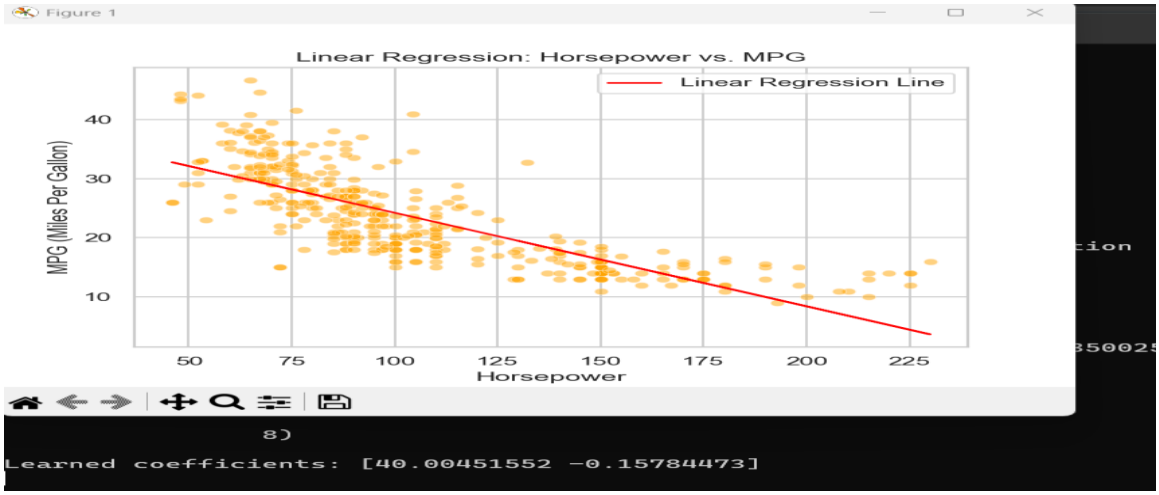

```
                7)

Correlation coefficient between 'horsepower' and 'mpg': -0.7714371350025521
There is a negative correlation between 'horsepower' and 'mpg'.
```

We calculated the correlation coefficient through the corr function. If it is between zero and negative one, it is negative, and if it is between zero and one, it is positive.

- In the data we had, the correlation was negative

8) Implement the closed form solution of linear regression and use it to learn a linear model to predict the ‘mpg’ from the ‘horsepower’.



We Used the closed form solution of linear regression to find the coefficients

$\theta=(XTX)-1XTy$

and we Adding a Regression Column values of 1) for the intercept to apply linear regression.

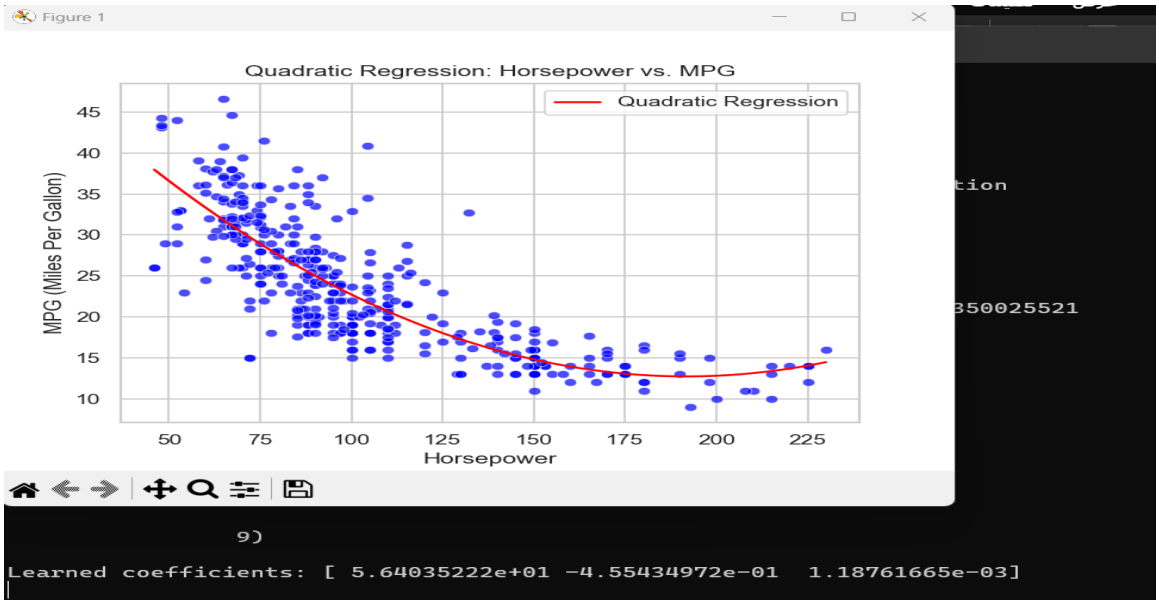
I implemented a linear regression model using a closed-form solution to understand the relationship between “horsepower” and “miles per gallon” and determined the parameters (weights) that determine the optimal line to represent this relationship. I then plotted the original data points and the learned linear line to illustrate the discovered relationship.

Intercept (Bias): 40.00451552 , Coefficient for 'horsepower': -0.15784473

The intercept represents the expected 'mpg' when 'horsepower' is zero.

The negative coefficient for 'horsepower' indicates an inverse relationship with 'mpg,' suggesting that as 'horsepower' increases, 'mpg' tends to decrease.

9) When we Repeat part 8 but now learn a quadratic function of the form f = w0 + w1x + w2x 2 .



We added two columns of ones and x ^2, Utilize the closed-form solution for quadratic regression using the equation: $\theta=(XTX)-1Xty$

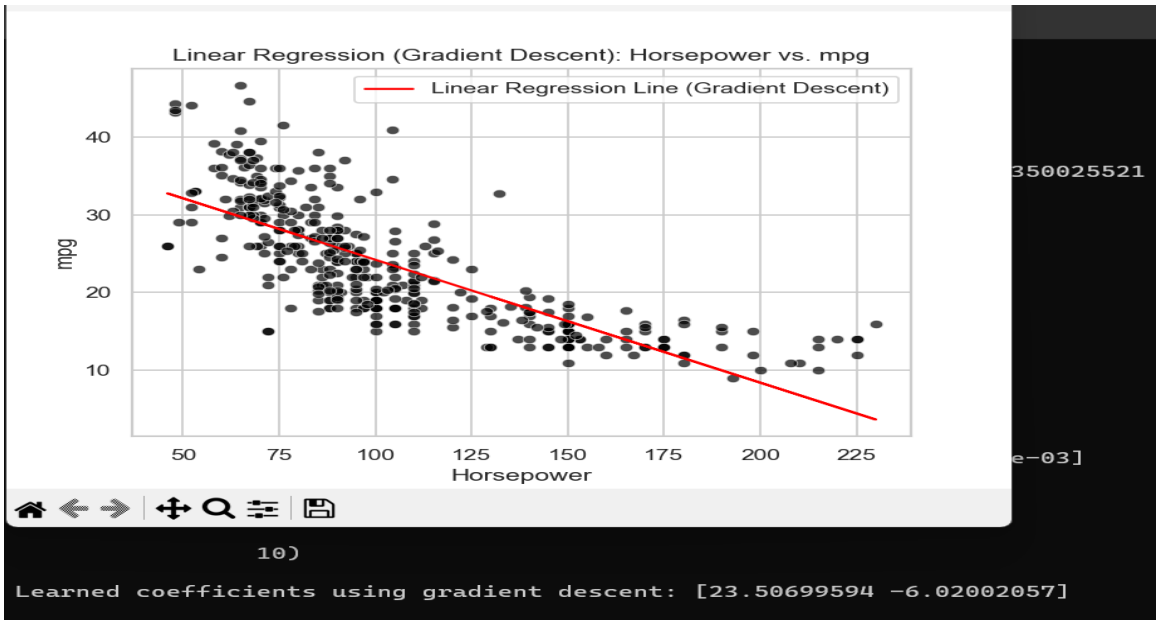
Earned transactions: [56.40352222, -0.45543497, 0.00118761665]

Intercept (bias): 56.40352222 , Horsepower coefficient: -0.45543497 , 'Horsepower^2' factor: 0.00118761665

The quadratic function has the form: f = w0 + w1 * x + w2 * x^2.

>>> The quadratic term (w2 * x^2) allows the model to capture nonlinear relationships.

10)by implementing the gradient descent algorithm instead of the closed form solution.



The intercept (bias) is 23.50699594, which is the expected 'mpg' when 'horsepower' is zero.

The coefficient for 'horsepower' (-6.02002057) represents the impact of 'horsepower' on 'mpg.' It indicates that for each unit increase in 'horsepower,' 'mpg' is expected to decrease by approximately 6.02 units.

The scatter plot with the linear regression line using gradient descent is showing the relationship between 'horsepower' and 'mpg.'

The red line represents the learned linear regression line based on the gradient descent optimization.

The similarity to a linear regression plot is expected because it is essentially learning a linear relationship between 'horsepower' and 'mpg.'