جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

MACHINE LEARNING AND DATA SCIENCE
ENCS5341

Assignment #1

Prepared by
Jana Abu Nasser
1201110

Instructor

Dr. Yazan Abu Farha

Section 1

BIRZEIT
30th November – 2023

Contents

1. Read the dataset and examine how many features and examples does it have? 3

2. Are there features with missing values? How many missing values are there in each one? . 3

3. Fill the missing values in each feature using a proper imputation method (for example: fill with mean, median, or mode) 4

4. Which country produces cars with better fuel economy? 5

5. Which of the following features has a distribution that is most similar to a Gaussian: ‘acceleration’, ‘horsepower’, or ‘mpg’? 6

6. Support your answer for part 5 by using a quantitative measure 6

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? 7

8. Implement the closed form solution of linear regression and use it to learn a linear model to predict the ‘mpg’ from the ‘horsepower’. Plot the learned line on the same scatter plot you got in part 7. 8

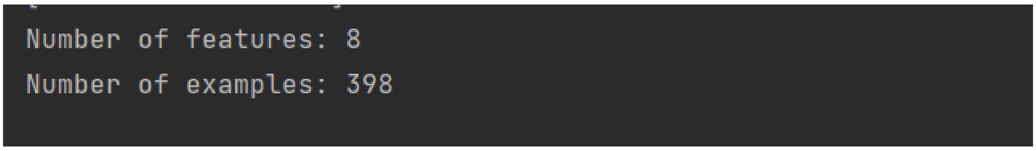
9. Repeat part 8 but now learn a quadratic function of the form 9

10. Repeat part 8 (simple linear regression case) but now by implementing the gradient descent algorithm instead of the closed form solution. 10

The "cars.csv" dataset contains details about cars, with a specific emphasis on fuel consumption. This assignment requires a comprehensive review and analysis of different elements within the dataset.

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

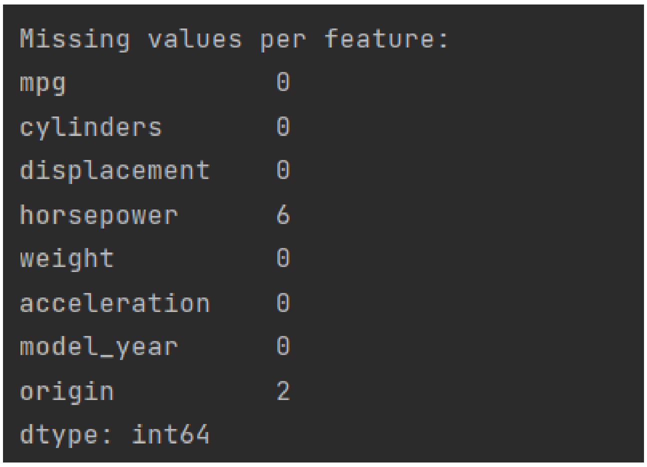
The data was loaded from a CSV file and stored in a Data Frame. The number of features is determined by the number of columns in the file, while the number of examples is matched to the number of rows.



```
Number of features: 8
Number of examples: 398
```

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

Certainly, missing values are present. Specifically, six missing values are observed in the "horsepower" column, and two missing values are displayed in the "origin" column, as indicated below:

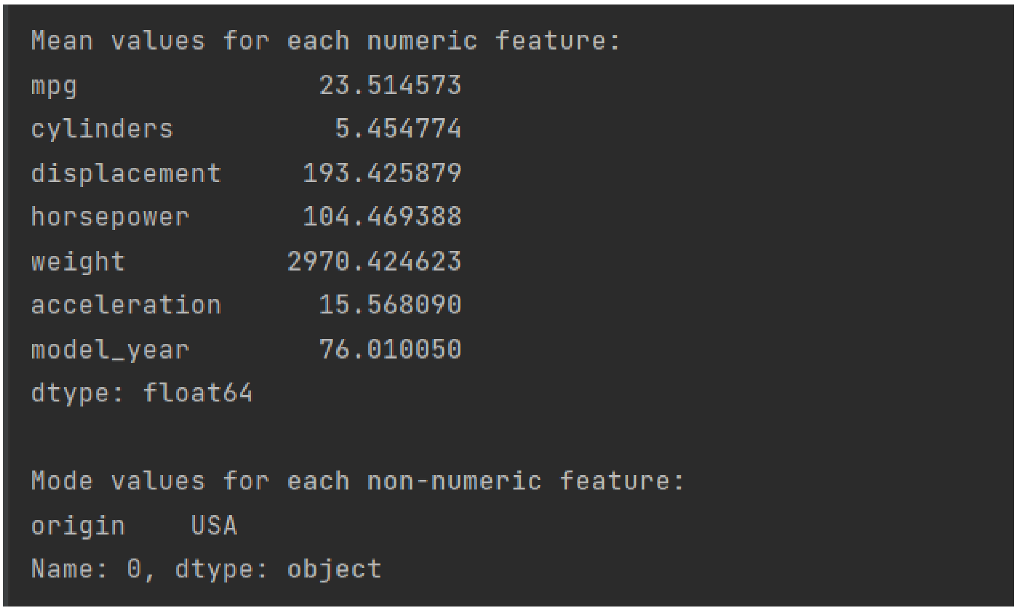


```
Missing values per feature:
mpg             0
cylinders       0
displacement    0
horsepower      6
weight          0
acceleration    0
model_year      0
origin          2
dtype: int64
```

"dtype" an abbreviation for "data type," denotes the kind of data stored within a pandas structure. When **"int64"** is specified, it indicates that the data type is a 64-bit integer. The "64" denotes the number of bits allocated for each integer, signifying the range of values it can encompass.

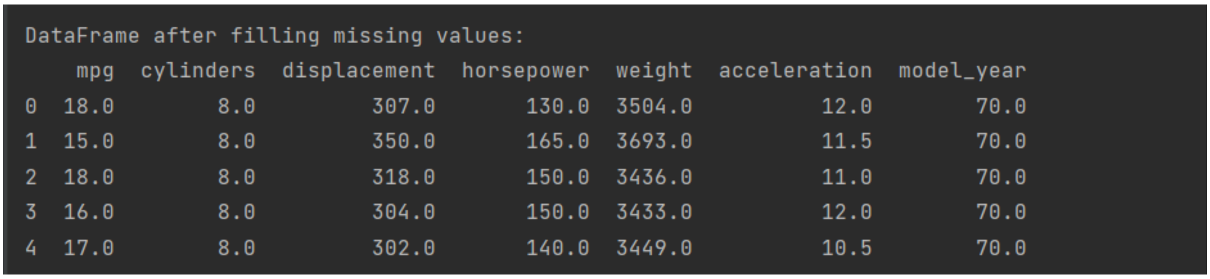
3. Fill the missing values in each feature using a proper imputation method (for example: fill with mean, median, or mode)

The missing values in the "horsepower" feature will be substituted with the computed mean. This approach minimizes the impact on the overall distribution of the data and furnishes a reasonable estimate for the absent values.



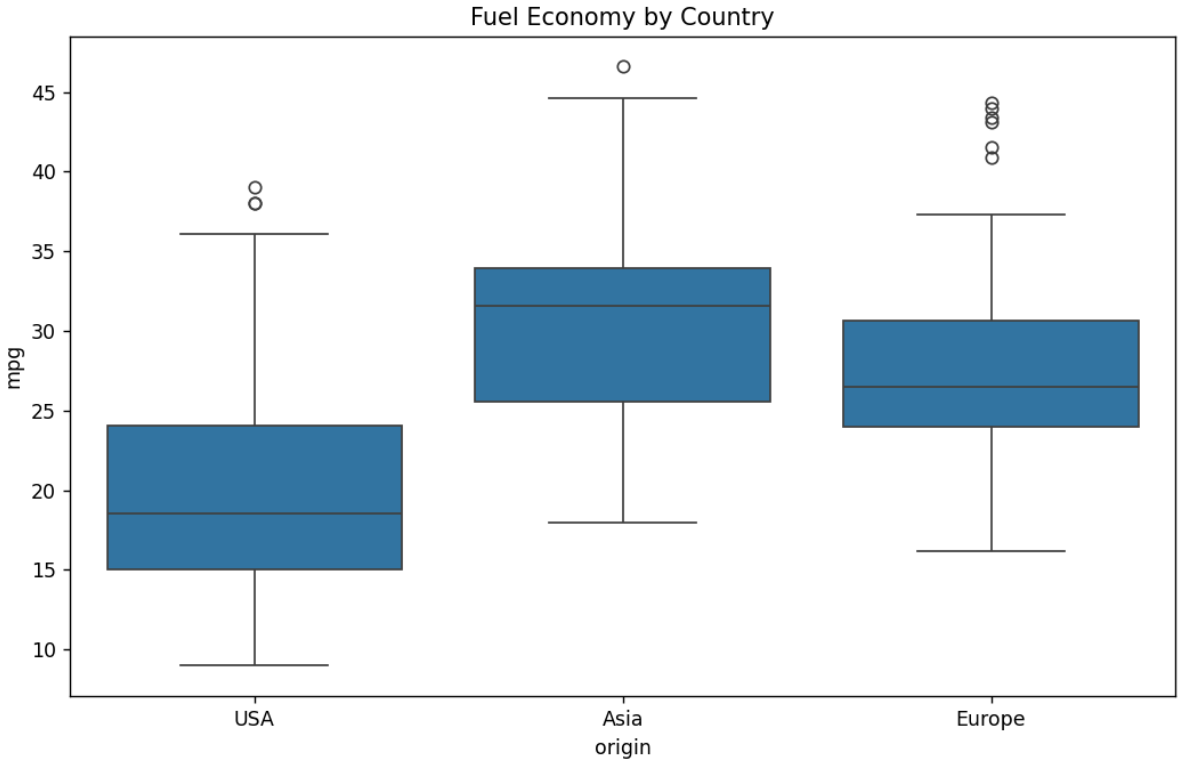
```
Mean values for each numeric feature:
mpg               23.514573
cylinders          5.454774
displacement     193.425879
horsepower       104.469388
weight          2970.424623
acceleration      15.568090
model_year        76.010050
dtype: float64

Mode values for each non-numeric feature:
origin    USA
Name: 0, dtype: object
```

The “origin” feature has two missing features. And this is part of the Data Frame after filling missing values:



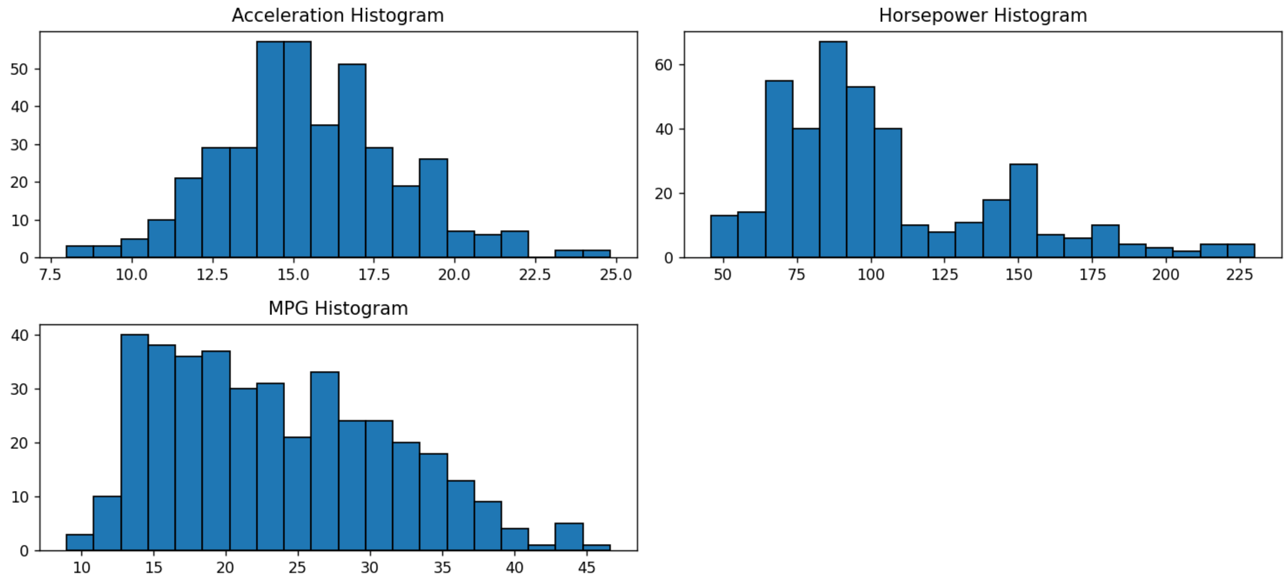
```
DataFrame after filling missing values:
    mpg  cylinders  displacement  horsepower  weight  acceleration  model_year
0  18.0        8.0         307.0       130.0  3504.0          12.0        70.0
1  15.0        8.0         350.0       165.0  3693.0          11.5        70.0
2  18.0        8.0         318.0       150.0  3436.0          11.0        70.0
3  16.0        8.0         304.0       150.0  3433.0          12.0        70.0
4  17.0        8.0         302.0       140.0  3449.0          10.5        70.0
```

4. Which country produces cars with better fuel economy?



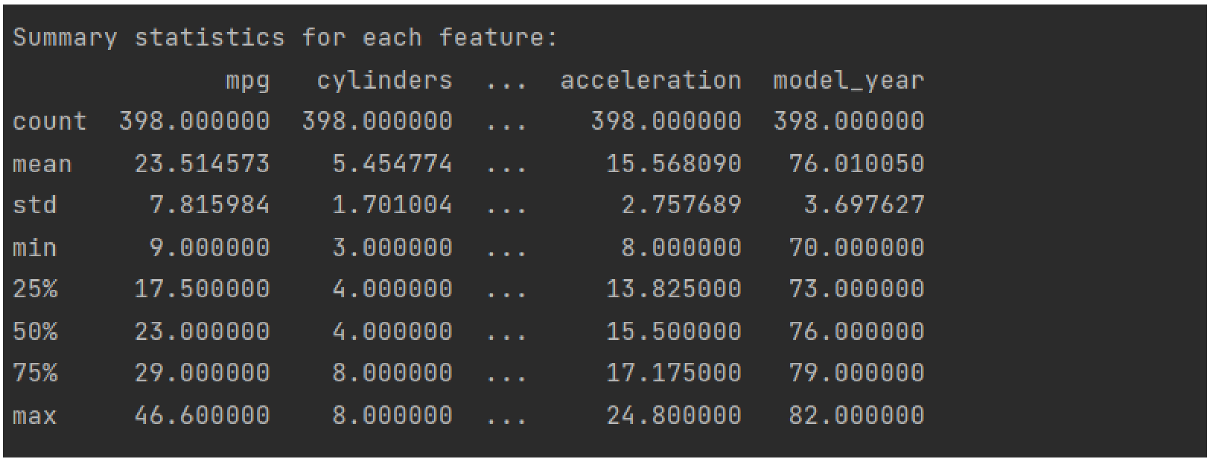
By utilizing the box plot, we can draw conclusions based on various indicators. A higher median, represented by the line inside the box, suggests better fuel economy. The height of the box, indicating the interquartile range, reflects more consistent fuel economy when taller. Longer whiskers on the plot may point to a wider range of fuel economy. Outliers, represented as points outside the whiskers, help identify any extreme values. Considering these criteria, it can be inferred that cars produced in Asia generally demonstrate superior fuel economy.

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



After visually analyzing the figure, it can be deduced that the distribution of 'acceleration' closely resembles a Gaussian distribution.

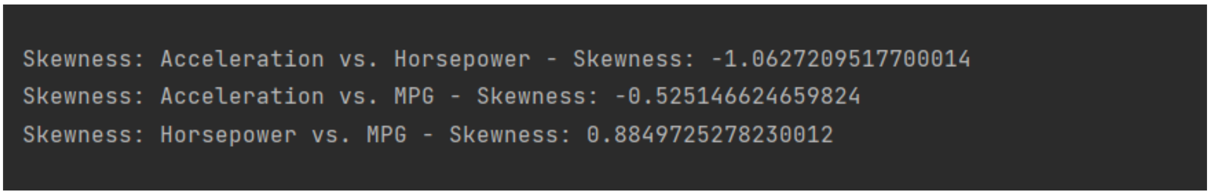
6. Support your answer for part 5 by using a quantitative measure



```
Summary statistics for each feature:
              mpg   cylinders  ...  acceleration  model_year
count  398.000000  398.000000  ...    398.000000  398.000000
mean    23.514573    5.454774  ...     15.568090   76.010050
std      7.815984    1.701004  ...      2.757689    3.697627
min      9.000000    3.000000  ...      8.000000   70.000000
25%     17.500000    4.000000  ...     13.825000   73.000000
50%     23.000000    4.000000  ...     15.500000   76.000000
75%     29.000000    8.000000  ...     17.175000   79.000000
max     46.600000    8.000000  ...     24.800000   82.000000
```

A Gaussian distribution is distinguished by its bell-shaped and symmetrical features. To validate these traits, an alternative method involves examining skewness and comparing the mean, median, and mode for each feature. If these central tendency measures demonstrate similar values, it indicates that the distribution closely resembles a Gaussian distribution.

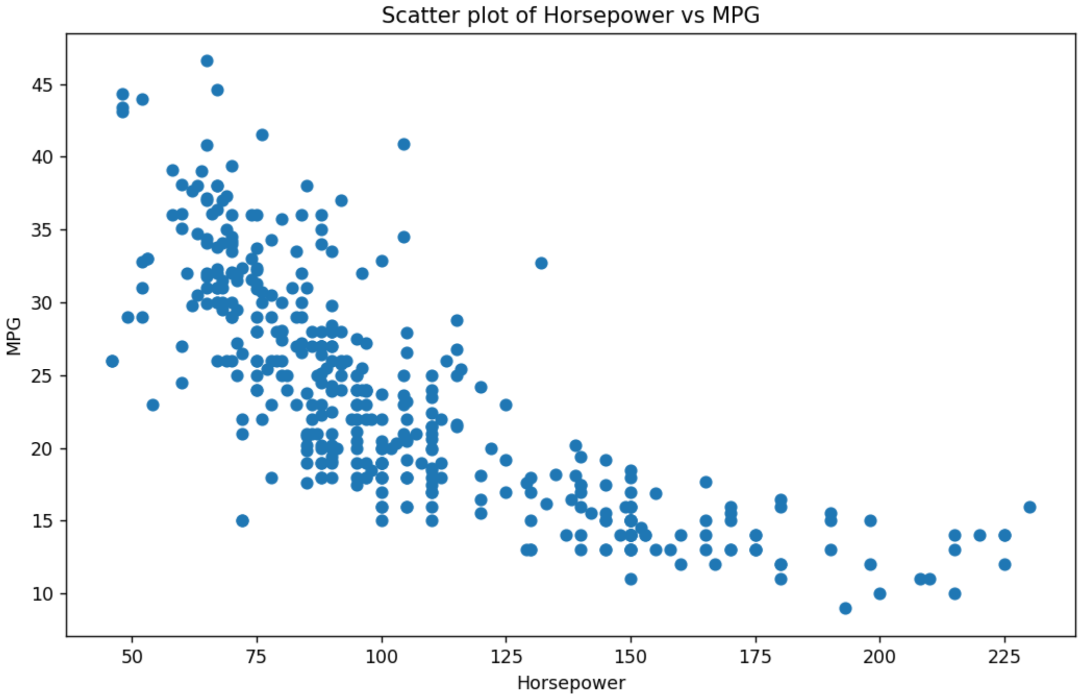
To calculate the skewness:

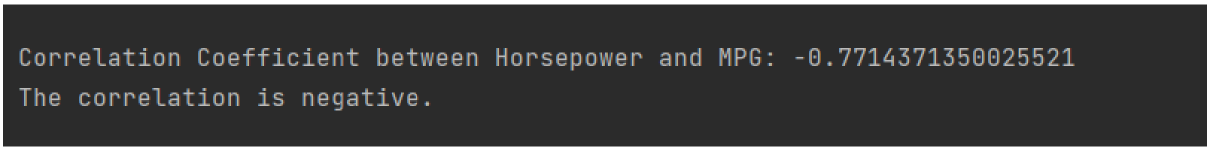


```
Skewness: Acceleration vs. Horsepower - Skewness: -1.0627209517700014
Skewness: Acceleration vs. MPG - Skewness: -0.525146624659824
Skewness: Horsepower vs. MPG - Skewness: 0.8849725278230012
```

- If S < 0, the distribution is negatively skewed (skewed to the left).
- If S = 0, the distribution is (not skewed).
- If S > 0, the distribution is positively skewed (skewed to the right).

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?

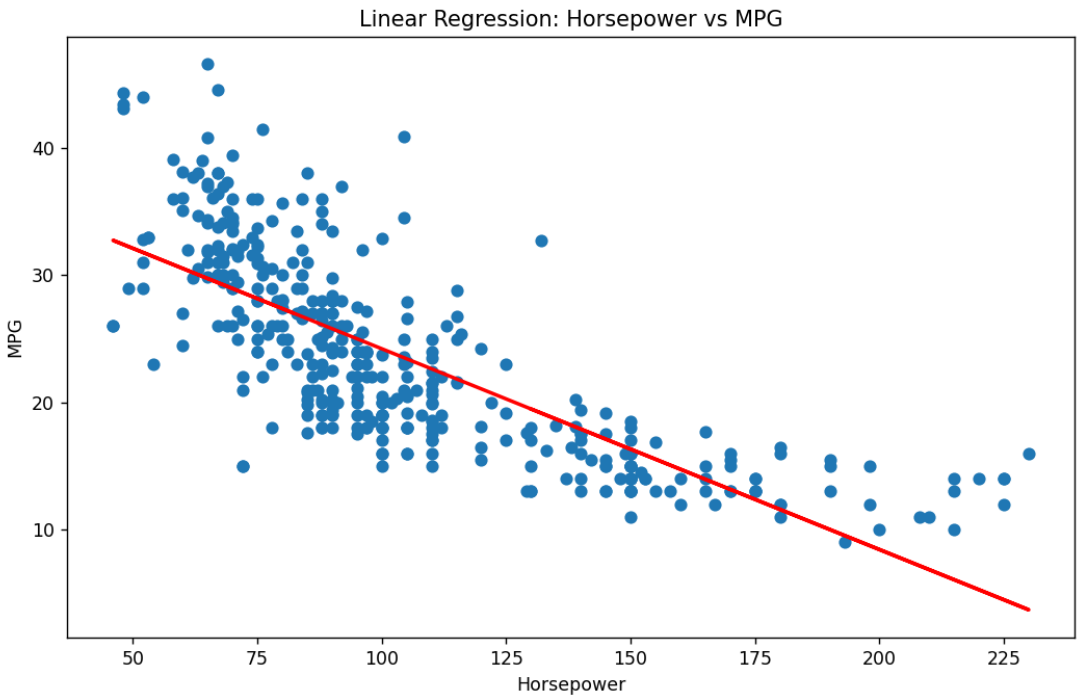


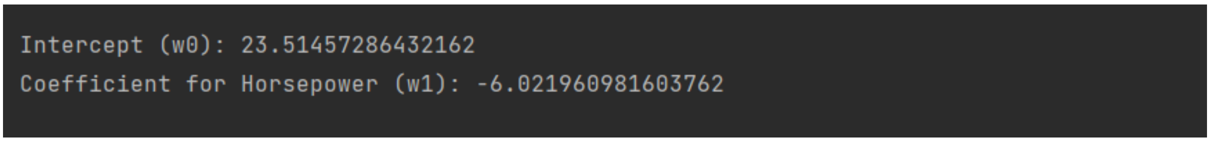


```
Correlation Coefficient between Horsepower and MPG: -0.7714371350025521
The correlation is negative.
```

Negative correlation is a statistical connection between two variables. In this scenario, when the value of one variable, like "Horsepower," rises, there is a tendency for the value of another variable, in this case, "MPG," to decrease.

8. Implement the closed form solution of linear regression and use it to learn a linear model to predict the 'mpg' from the 'horsepower'. Plot the learned line on the same scatter plot you got in part 7.





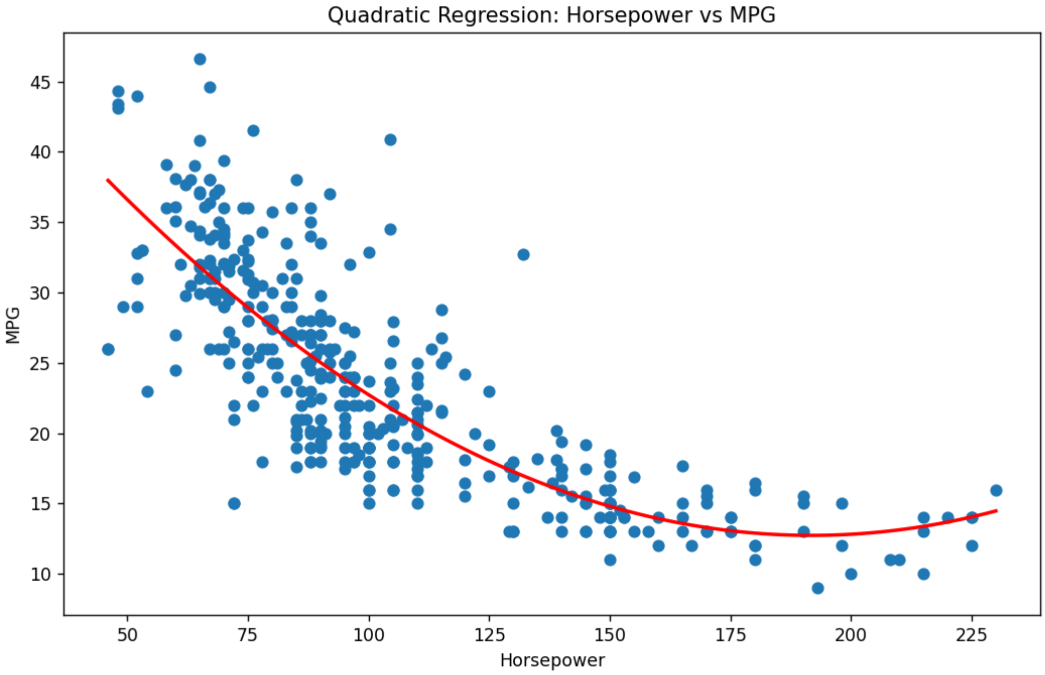
```
Intercept (w0): 23.51457286432162
Coefficient for Horsepower (w1): -6.021960981603762
```

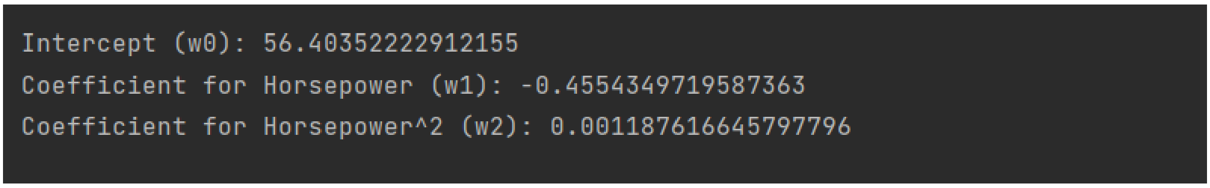
From the figure above :

$$y = -6.0219x + 23.5145$$

9. Repeat part 8 but now learn a quadratic function of the form

$$f = w_0 + w_1 x + w_2 x^2$$



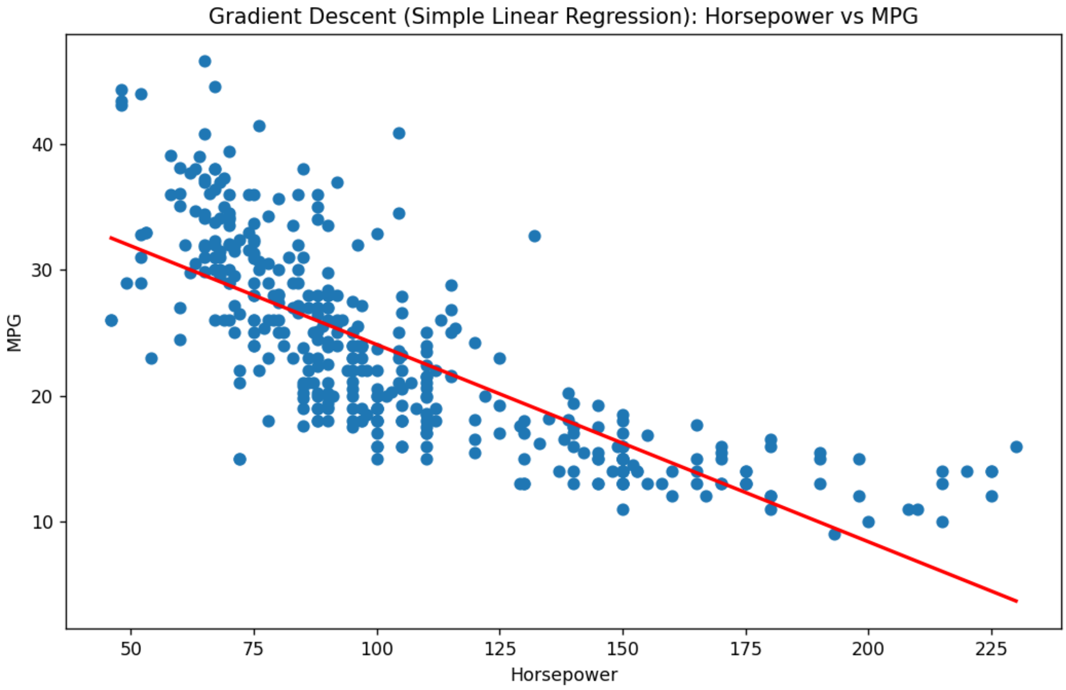


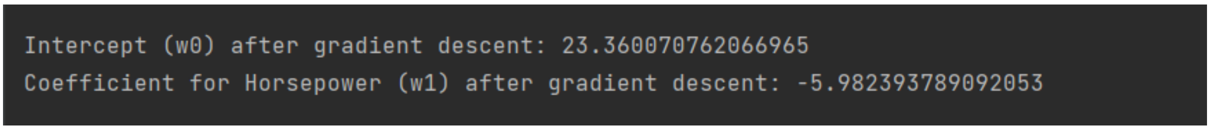
```
Intercept (w0): 56.40352222912155
Coefficient for Horsepower (w1): -0.4554349719587363
Coefficient for Horsepower^2 (w2): 0.0011876166457977
```

So:

$$y = \ 0.001187x^2 - 0.45543x + 56.4035$$

10. Repeat part 8 (simple linear regression case) but now by implementing the gradient descent algorithm instead of the closed form solution.





```
Intercept (w0) after gradient descent: 23.360070762066965
Coefficient for Horsepower (w1) after gradient descent: -5.982393789092053
```

So:

$$y = 23.3600 - 5.98239x$$