Logo

Description automatically generated

DATA VISUALIZATION REPORT

LAB 03: DATA VISUALIZATION BASICS

GROUP 07

20127449 – Tran Quoc Bao

20127452 – Ho Dang Cao

20127476 – Do Duc Duy

20127567 – Le Nguyen Binh Nam

Instructeurs

Dr. Bui Tien Len

Dr. Le Ngoc Thanh

Table of Contents

[EVALUATIONS 1](#_Toc134565311)

[LEVEL OF COMPLETION OF REQUIREMENTS 1](#_Toc134565312)

[LEVEL OF COMPLETION OF MEMBERS 1](#_Toc134565313)

[THEORY 2](#_Toc134565314)

[Data Visualization with Matplotlib 2](#_Toc134565315)

[DETAILED ALGORITHM 3](#_Toc134565316)

[Exploratory analysis of Car MPG data 3](#_Toc134565317)

[Electric power consumption data 17](#_Toc134565318)

# EVALUATIONS

## LEVEL OF COMPLETION OF REQUIREMENTS

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | TASK | MEMBER | LEVEL |
| 1 | Part 3 question 1 | 20127452 | 100% |
| 2 | Part 3 a half question 2 | 100% |
| 3 | Part 3 question 3 | 100% |
| 4 | Part 3 a half question 4 | 100% |
| 5 | Part 3 a half question 5 | 100% |
| 6 | Part 3 a half question 6 | 100% |
| 7 | Part 3 a half question 7 | 100% |
| 8 | Part 3 a half question 8 | 100% |
| 9 | Part 3 question 9 | 100% |
| 10 | Part 4 code 4 plot | 100% |
| 11 | Part 3 a half question 8 | 20127567 | 100% |
| 12 | Part 3 a half question 7 | 100% |
| 14 | Part 3 a half question 5 | 20127476 | 100% |
| 15 | Part 3 a half question 6 | 100% |
| 16 | Data Visualization with Matplotlib |  |
| **17** | Part 3 a half question 2 | 20127449 | 100% |
| 18 | Part 3 a half question 4 | 100% |
| 19 | A story in Part 4 | 100% |

## LEVEL OF COMPLETION OF MEMBERS

|  |  |  |
| --- | --- | --- |
| ID | FULL NAME | LEVEL |
| 20127449 | TRAN QUOC BAO | 100% |
| 20127452 | HO DANG CAO | 100% |
| 20127476 | DO DUC DUY | 100% |
| 20127567 | LE NGUYEN BINH NAM | 100% |

# THEORY

## Data Visualization with Matplotlib

* The matplotlib library is one of the most popular graphics libraries in Python used to create high-quality graphs and images. This library allows users to create many different types of charts including line charts, column charts, pie charts, scatter charts and more.
* The salient features of matplotlib include:
  + Ability to generate high quality graphs and images with high resolution and custom details.
  + Supports a wide variety of file formats, including PNG, PDF, EPS, and SVG.
  + Allows users to customize the properties of chart elements such as color, thickness, size, etc.
  + Supports multiple graph types, including line graph, column graph, pie chart, scatter plot, etc.
  + Ability to integrate with other libraries like NumPy, pandas, SciPy, etc.
  + Provides tools for creating animations and interactive charts.
  + With the support of matplotlib, Python users can create high-quality graphs and images to display and present data and information in a clearer and easier to understand manner. This library is a useful tool for data scientists, analysts, and researchers in analyzing data and displaying results visually and vividly.

# DETAILED ALGORITHM

## Exploratory analysis of Car MPG data

Description of data:

mpg: is Miles Per Gallon.

cylinders: the number of cylinders of each car (discrete valued attribute).

displacement: the engine displacement (quantitative continuous value).

horsepower: the horsepower of a car (continuous valued variable).

weight: the weight of each car (continuous valued quantitative variable).

acceleration: the acceleration of the car.

model: the year the model was released is mentioned in this attribute.

origin: has 3 discrete values. 1 refers to cars made in USA, 2 for Europe and 3 for Asia

car\_name: This is the name of the car in string.

1. How many cars and how many attributes are in the data set.

* There are 406 cars (number of rows) and 9 attributes (number of columns).



  
2. How many distinct car companies are represented in the data set? What is the name of the car  
with the best MPG? What car company produces the most 8-cylinder cars? What are the names  
of 3-cylinder cars? Do some internet search that can tell you about the history and popularity of  
those 3-cylinder cars.

**Note:** Companies are the first words of car names.

* There are 38 distinct car companies represented in the data set.





* Mazda glc is the car with the best MPG.





* Ford produced the most 8-cylinder cars.





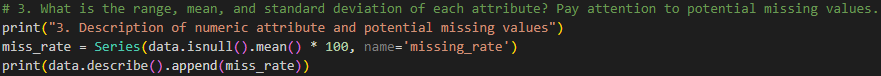
* Mazda rx2 coupe, Mazda rx3, Mazda rx-4 and Mazda rx-7 gs are 3-cylinder cars.





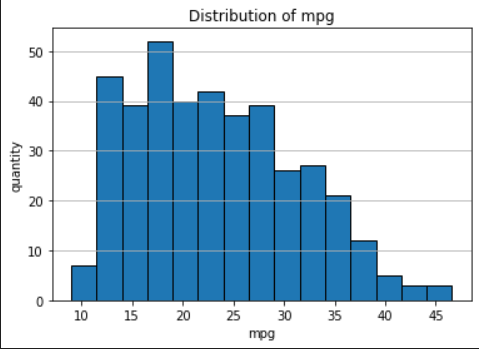
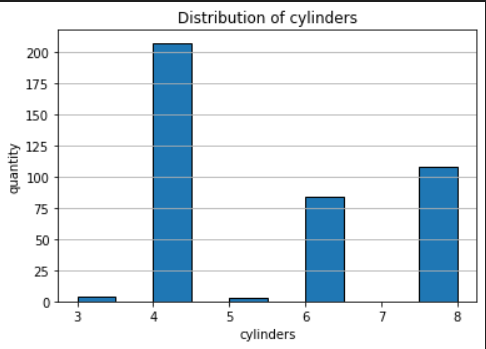
* The history and popularity of those 3-cylinder cars.
  + Mazda rx2 coupe: (1970-1978) The Mazda RX-2 was launched in 1970 and was sold between 1971 and 1974 in the US, when Mazda ceased production. However, it lived on in other parts of the world.
  + Mazda rx3:  Mazda launched the new Mazda RX-3 in September 1971, smaller and sportier than the Mazda RX-2, the rotary RX-3 was called the Mazda Savanna in Japan
  + Mazda rx-4: The Mazda RX-4 (called the Luce Rotary in Japan) is an automobile sold in the 1970s. It was a larger car than its rotary-powered contemporaries, the Capella-based RX-2 and Familia-based RX-3. It shared the Luce/929 chassis, replacing the R130 in October 1972, and was produced in October 1977.
  + Mazda rx-7 gs:

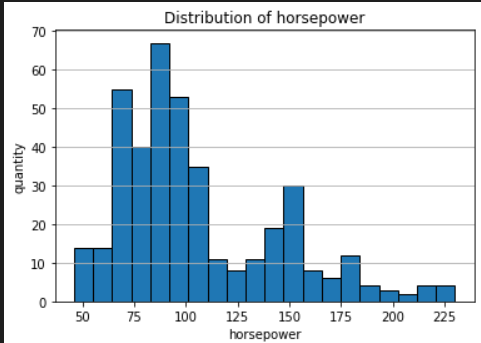
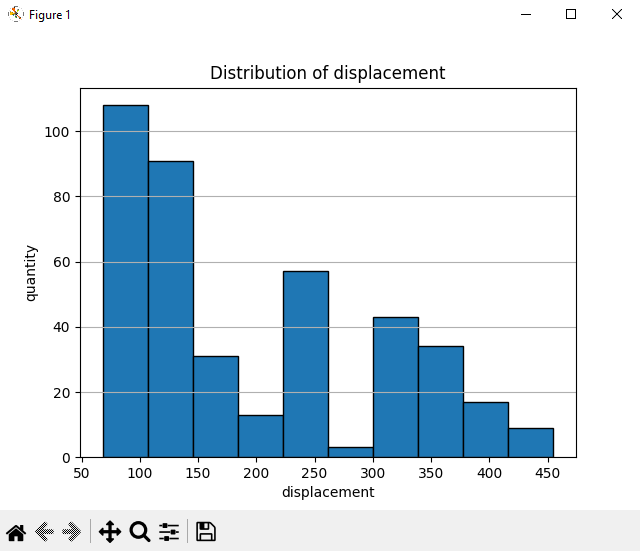
3. What is the range, mean, and standard deviation of each attribute? Pay attention to potential  
missing values.

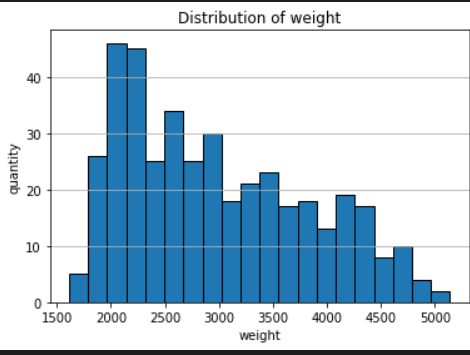
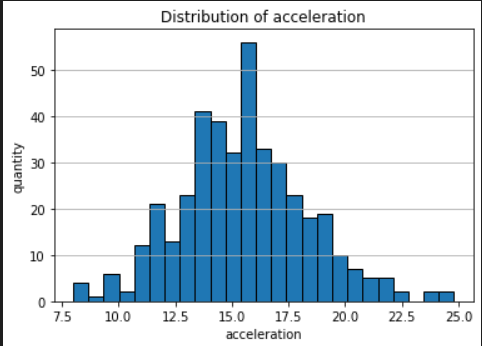


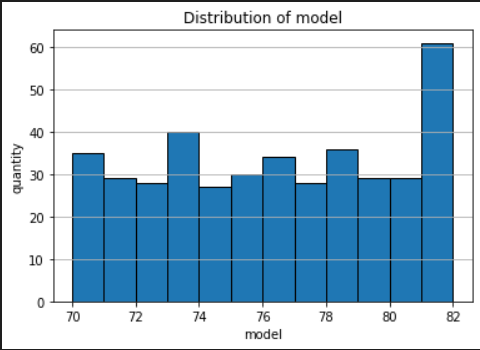
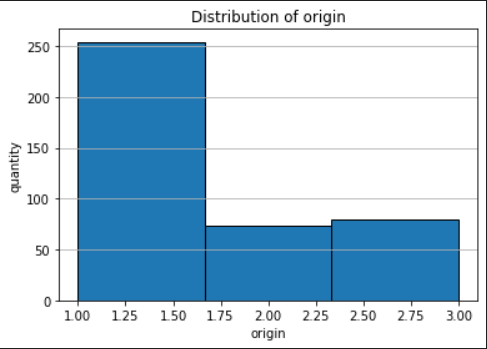
Graphical user interface, text

Description automatically generated  
4. Plot histograms for each attribute. Pay attention to the appropriate choice of number of bins.  
Write 2-3 sentences summarizing some interesting aspects of the data by looking at the histograms.



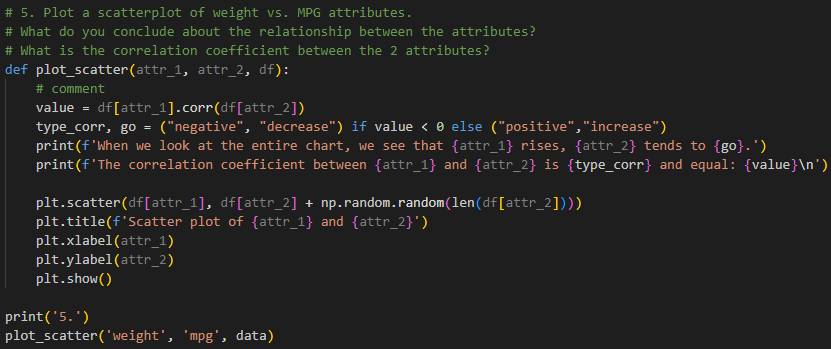
 

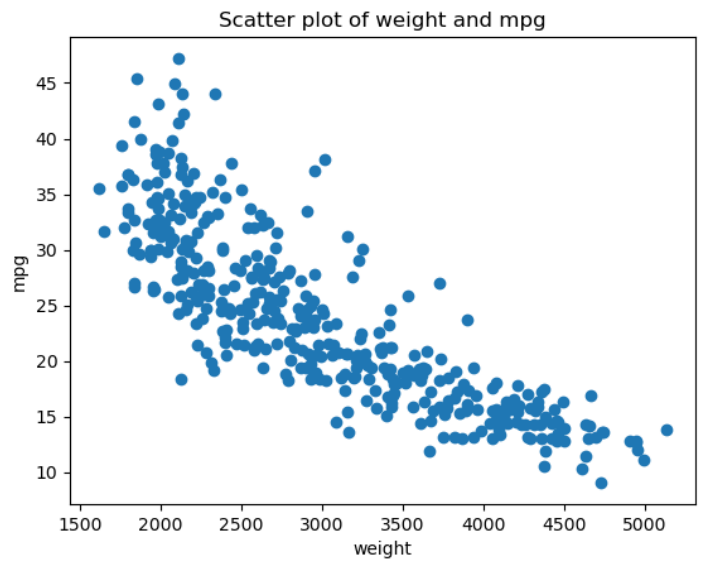
 

* Summarizing some interesting aspects of the data:
* Histogram of mpg data skewed to the left shows that most cars have low fuel consumption.
* Histogram of horsepower and weight is skewed to the left. Most cars have low horsepower and light weight.
* The number of 4-cylinder cars is the most.

5. Plot a scatterplot of weight vs. MPG attributes. What do you conclude about the relationship  
between the attributes? What is the correlation coefficient between the 2 attributes?

**Note:** Write a function to plot scatter between 2 attributes and to show coefficient.

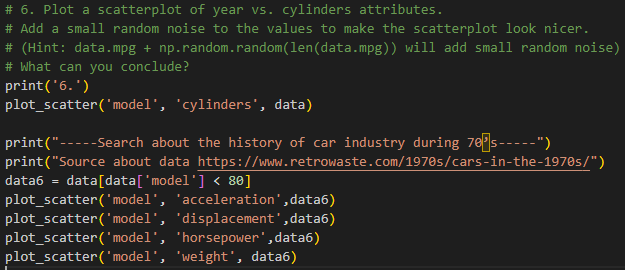


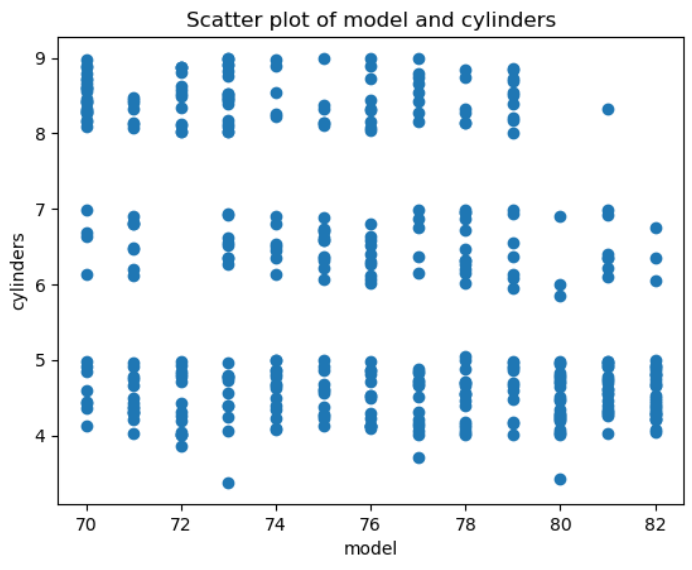


When we look at the entire chart, we see that weight rises, mpg tends to decrease.

The correlation coefficient between the 2 attributes is negative and equal: -0.8322442148315753.

6. Plot a scatterplot of year vs. cylinders attributes. Add a small random noise to the values to make the scatter plot look nicer. What can you conclude? Do some internet search about the history of the car industry during the 70's that might explain the results.

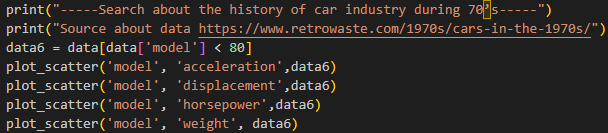




When we look at the entire chart, we see that as weight rises, cylinders tend to decrease.

The correlation coefficient between the 2 attributes is negative and equal: -0.3456474403380831

**Some things about the history of cars during the 70's in our data.**



‘model' with 'acceleration'

Chart, scatter chart

Description automatically generated

When we look at the entire chart, we see that the model rises, acceleration tends to rise.

The correlation coefficient between the 2 attributes is positive and equal: 0,24222651078606197.

'model' with 'horsepower'

Chart, scatter chart

Description automatically generated

When we look at the entire chart, we see that as the model rises, horsepower tends to decrease.

The correlation coefficient between the 2 attributes is negative and equal: -0.2809264283375341.

'model' with ‘displacement'

Chart, scatter chart

Description automatically generated

When we look at the entire chart, we see that the model rises, displacement tends to decrease.

The correlation coefficient between the 2 attributes is positive and equal: -0.20229459872935793

'model' with ‘'weight'

Chart

Description automatically generated

When we look at the entire chart, we see that as the model rises, weight tends to decrease.

The correlation coefficient between the 2 attributes is negative and equal: -0.1282099332787899

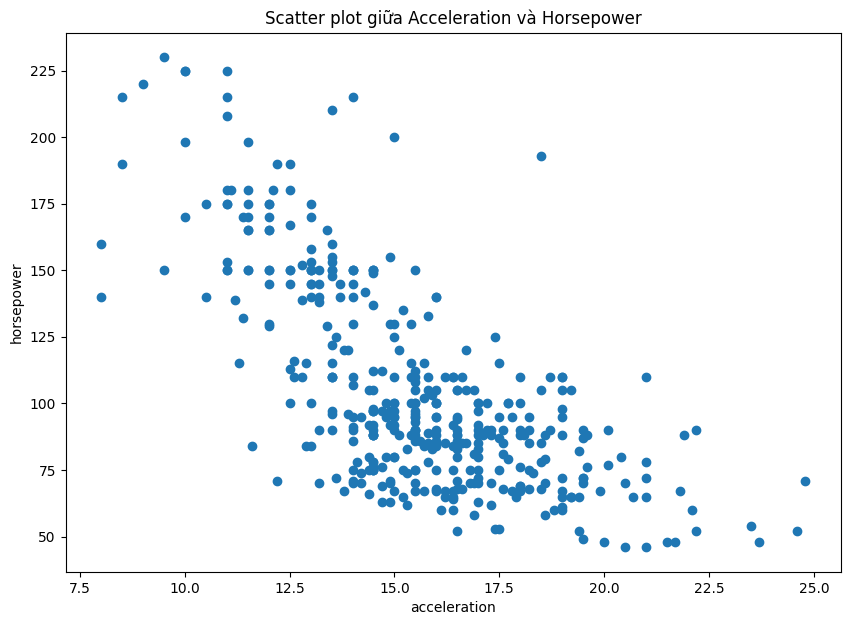
*The evidence on the internet has partly explained the change of properties in the data that we have visualized. The things that we realize after visualizing completely match the evidence on the internet.*

7. Show 2 more scattered plots that are interesting to you. Discuss what you see.

Chart, scatter chart

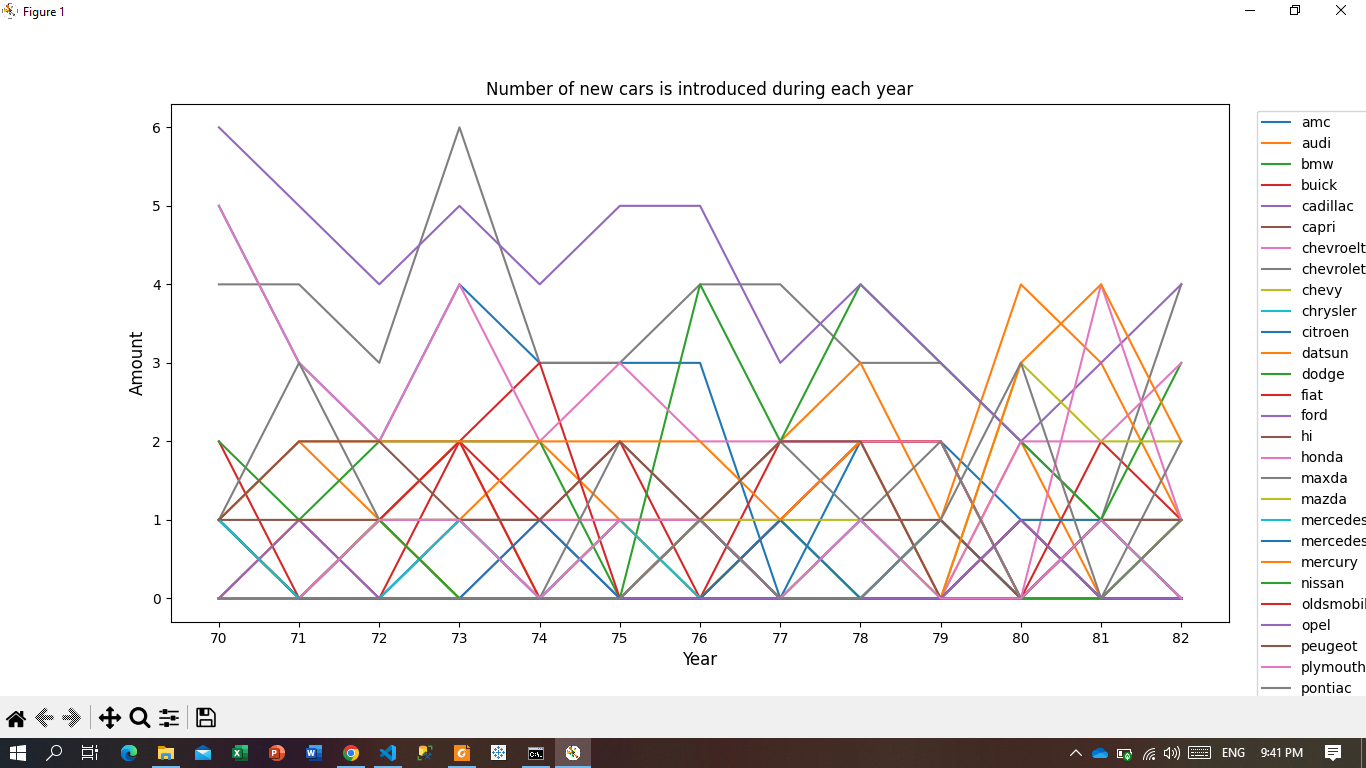
Description automatically generated

Based on the scatter plot between the cylinders and the displacement, we can see a clear positive correlation between these two variables. In fact, engines with many cylinders are often designed with a large capacity to provide greater capacity for the vehicle. In addition, small displacement 4-cylinder engines are often used in mid-range or multi-purpose vehicles, so they make up many the data set.



Based on the scatter plot between acceleration and horsepower, we can see a significant paradoxical relationship between the two. This means that if the car can accelerate at first (high acceleration), when accelerating to a higher level, it will be difficult for the car to achieve maximum power (low horsepower), and vice versa. The data set points in the middle of the graph show that the cars in this data set are geared towards a balance between the two states of acceleration and horsepower, i.e. have good acceleration and maximum performance. effectively. However, it is necessary to analyze each specific case more carefully to make more accurate evaluations about the relationship between these two indicators.

8. Plot a time series for all the companies that show how many new cars they introduce during each year. Do you see some interesting trends?



Based on the bar chart showing the number of cars launched by car manufacturer and year, we can see that the big car manufacturers usually launch 2 to 6 models per year, while the smaller car brands usually launch 2 to 6 models per year. launch only 1 to 2 models per year. In addition, we can also notice that the number of cars launched by car manufacturers tends to decrease in the last years of each decade and increase in the first years of the next decade. This shows the orientation and strategy of car manufacturers in product development and market access.

9. Calculate the pairwise correlation and draw the heatmap with Matplotlib. Do you see an interesting correlation?

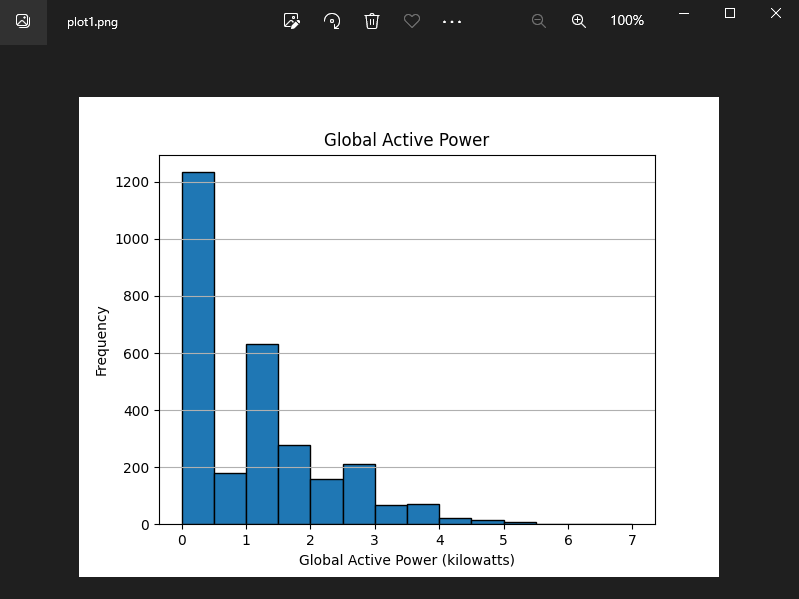


cylinders, displacement, horsepower, weight have a highly positive correlation with each other.

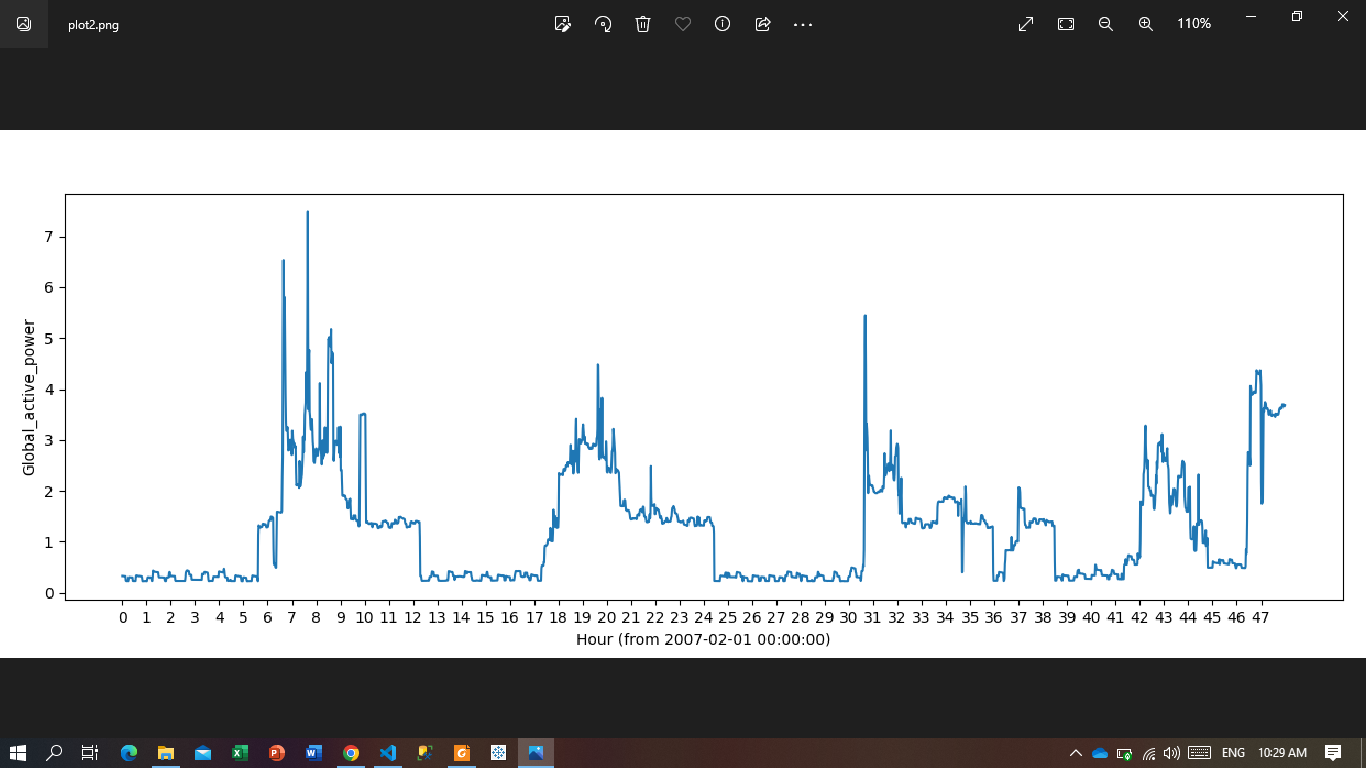
cylinders, displacement, horsepower, and weight have a highly negative correlation with mpg.

## Electric power consumption data

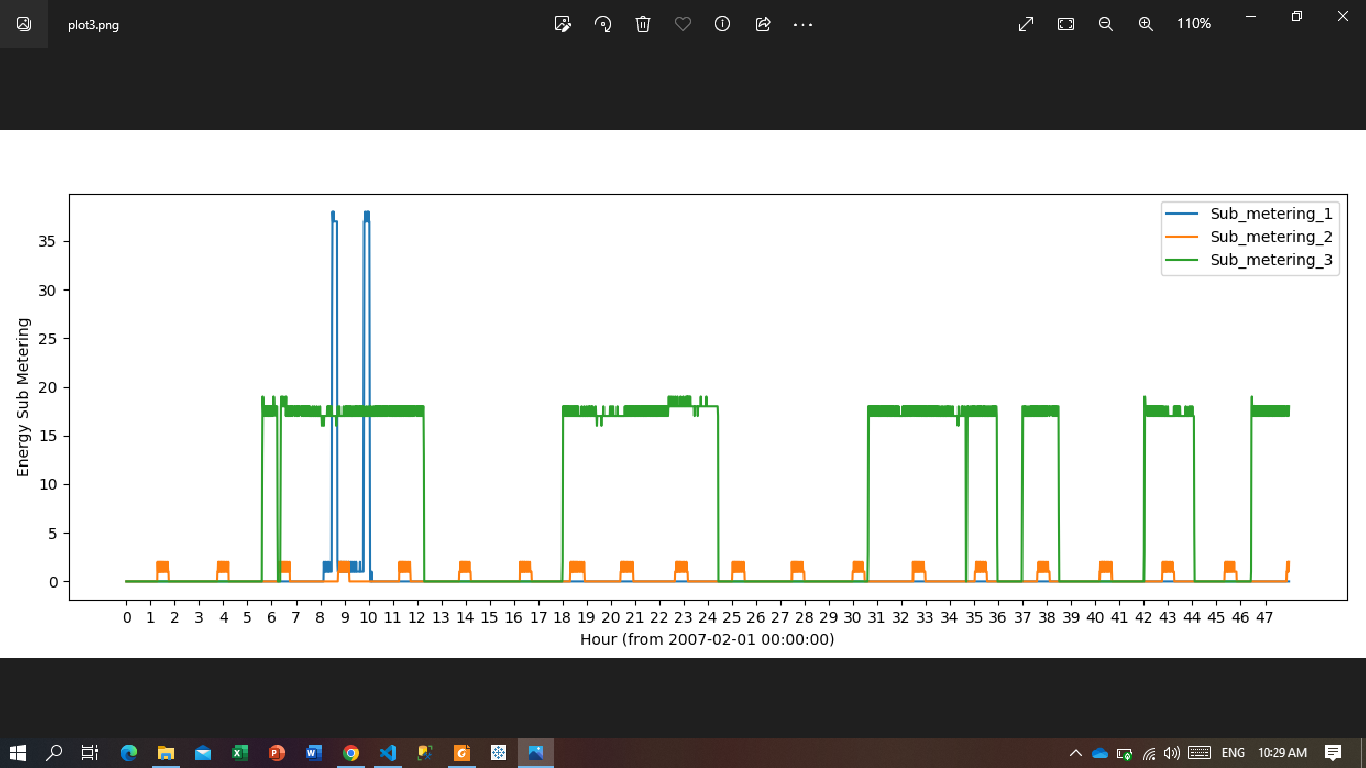
1. Plot 1



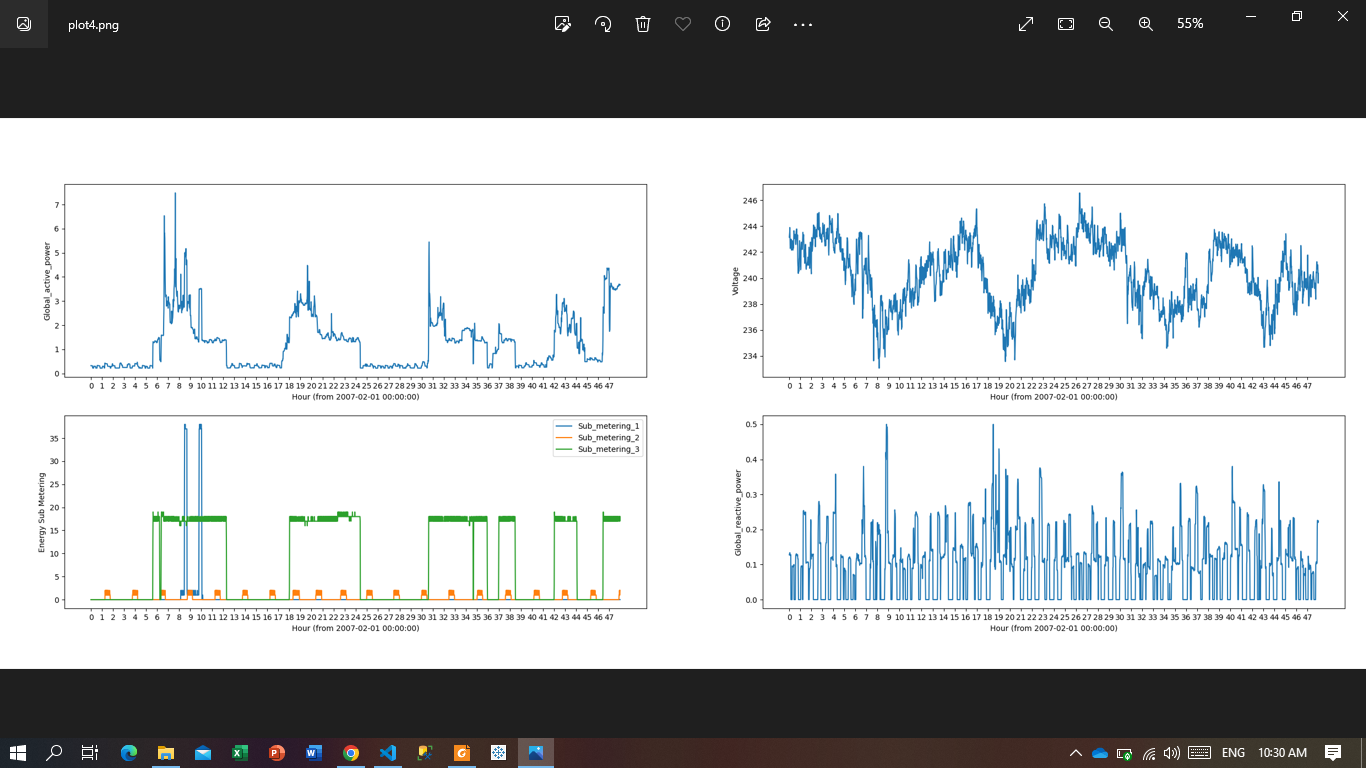
2. Plot 2



3. Plot 3



4. Plot 4



5. A story about the energy use of the household during the two days

* On the first day, from the 4 graphs above, the peak hours for household electricity use are:
  + First rush hour: 5h to 12h, the peak is 9h. Specifically, in chart 3, it shows that the increase in submeters 1, 2, and 3 of the same type shows that the household has used many electrical devices at the same time; Right now in the first chart, the amount of electricity used also spiked and in the second chart, the voltage dropped a lot (due to the amount of electricity used).
  + About Monday rush hour: 17h to 23h peak is 20h. In chart 3, submeters 2, 3 spike. The rest of the charts are similar to “First peak hour”.
* Second day: Peak hours: 7h - 12h, 13h-15h, 18h-21h. In the 3rd chart, the submeter 2 and 3 graphs spiked. Shows that use a lot of electrical appliances but not as much as the first day.