Đại học Quốc gia TP. HCM
Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

BÁO CÁO LAB 03
TRỰC QUAN HÓA DỮ LIỆU VỚI NUMPY, PANDAS, MATPLOTLIB

Trực quan hoá dữ liệu (CSC10108)

Nhóm 17

TP Hồ Chí Minh, ngày 14/06/2021

Contents

1 Thông tin nhóm

STT	MSSV	Họ tên	Email	SĐT
1	18120078	Ngô Phù Hữu Đại Sơn	18120078@student.hcmus.edu.vn	0919070940
2	18120201	Nguyễn Bảo Long	18120201@student.hcmus.edu.vn	0981850699
3	18120227	Phạm Văn Minh Phương	18120227@student.hcmus.edu.vn	0343049359
4	18120253	Mai Ngọc Tú	18120253@student.hcmus.edu.vn	0981850699
5	1712424	Hàn Văn Gia Hiên	1712424@student.hcmus.edu.vn	0911572108

Table 1: Bảng danh sách thành viên nhóm

2 Phân tích hoàn thiện yêu cầu

2.1 Tổng quan mức độ hoàn thành mỗi yêu cầu

STT	Yêu cầu	Công việc	Hoàn thành (%)
1	Analysis of Car MPG data	- Trả lời câu hỏi - Vẽ biểu đồ và nghiên cứu lịch sử	100/100
2	Story of EPC data	- Trực quan các biến, vẽ biểu đồ giải thích ý nghĩa, viết story	100/100

Table 2: Bảng phân tích hoàn thành yêu cầu

2.2 Mức độ hoàn thành của thành viên nhóm

STT	Họ tên	Công việc tham gia	Hoàn thành (%)
1	Ngô Phù Hữu Đại Sơn	- Vẽ biểu đồ cho phần Story of EPC data	100/100
2	Nguyễn Bảo Long	- Vẽ biểu đồ câu hỏi 4, 5, 6 phần Exploratory analysis of Car MPG data Nhận xét biểu đồ	100/100
5	Phạm Văn Minh Phương	- Tìm hiểu lịch sử xe hơi giải thích dữ liệu phần Exploratory analysis of Car MPG data - Viết báo cáo	100/100
5	Mai Ngọc Tú	- Vẽ biểu đồ câu hỏi 1, 2, 3 phần Exploratory analysis of Car MPG data - Nhận xét biểu đồ	100/100
5	Hàn Văn Gia Hiên	- Vẽ biểu đồ câu hỏi 7, 8, 9 phần Exploratory analysis of Car MPG data - Nhận xét biểu đồ	100/100

3 Mô tả các thư viện liên quan

1. ***Matplotlib***: Thư viện hỗ trợ các công cụ vẽ các biểu đồ trực quan(biểu đồ cột, biểu đồ tròn, biểu đồ đường, Histogram, . . .).

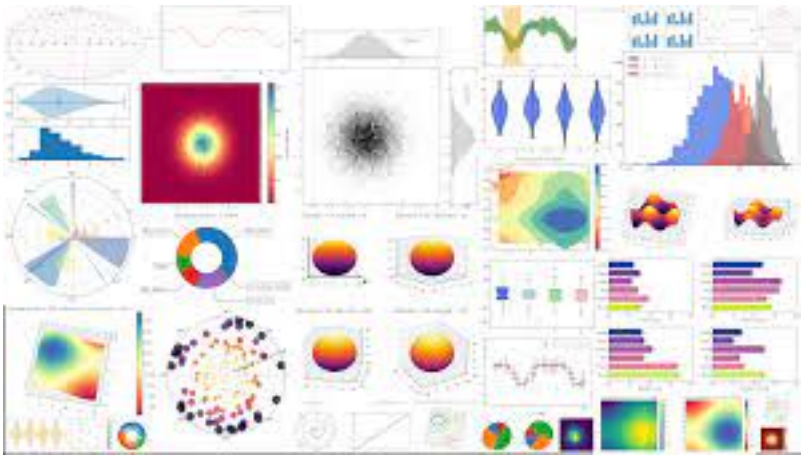

Figure 1: Một số biểu đồ vẽ bằng thư viện matplotlib

2. ***Pandas***: Hỗ trợ các lệnh, hàm và cấu trúc dữ liệu (Dataframe) để tải lên, xử lý và ghi ra các file dữ liệu bản như xls, csv,. . .

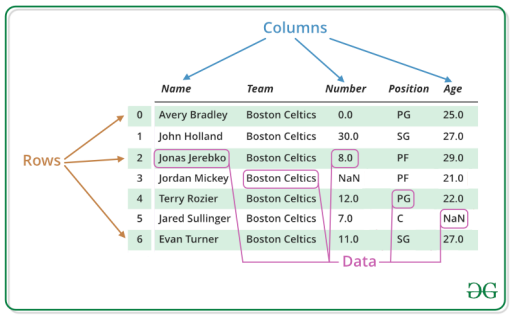

Figure 2: Cấu trúc một Dataframe trong Pandas

3. ***Numpy***: Hỗ trợ các hàm và cấu trúc dữ liệu dạng mảng (một hay nhiều chiều)để thực hiện các phép tính tiện lợi hơn so với cấu trúc mảng mặc định của python (các phép tính đại số tuyến tính, các phép cắt mảng, nối mảng,. . .).

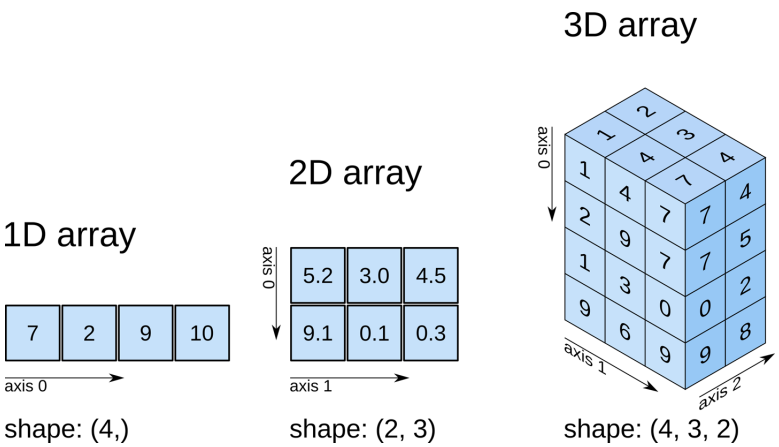

Figure 3: Minh họa mảng 1, 2, 3 chiều trong numpy

4 Exploratory analysis of Car MPG data

4.1 Trả lời câu hỏi

4.1.1 Có bao nhiêu xe và bao nhiêu thuộc tính trong bộ dữ liệu

- Có 312 xe và 9 thuộc tính trong bộ dữ liệu

4.1.2 Có bao nhiêu công ty xe hơi riêng biệt xuất hiện trong bộ dữ liệu? Tên của chiếc xe có MPG cao nhất là gì? Công ty nào sản xuất nhiều xe 8 xi-lanh nhất? Tên của những chiếc xe 3 xi-lanh? Tìm kiếm trên internet thông tin về lịch sử và độ phổ biến của những chiếc xe 3 xi-lanh này.

- Có 38 công ty xe hơi xuất hiện trong bộ dữ liệu
- Xe có MPG cao nhất: Mazda glc
- Công ty sản xuất nhiều xe 8 xi-lanh nhất: Ford
- Tên những xe có 3 xi lanh: Mazda rx2 coupe, Maxda rx3, Mazda rx-4, Mazda rx-7 gs
- Lịch sử và độ phổ biến của những chiếc xe 3 xi-lanh trên:
  - Mazda RX-2 là một chiếc ô tô cỡ vừa ra mắt vào năm 1970 và được sản xuất tới năm 1979. Mazda RX-2 được định hướng là một chiếc ô tô cho gia đình và sử dụng động cơ quay. Là một phiên bản cực kì thành công và có độ phổ biến rất cao tại thời điểm đó.
  - Mazda RX3 được bán từ năm 1971 tới năm 1978. Vào năm đầu ra mắt, doanh số toàn cầu của RX-3 là khoảng 200,000 chiếc. Từ 1972 trở đi, doanh số RX-3 đã vượt hẳn RX-2 và đạt kỉ lục bán ra 105,819 chiếc chỉ trong năm 1972, giúp tổng xe động cơ xoay đạt 500,000 chiếc. Từ 1974 trở đi doanh số của RX-4 cao hơn nhưng RX-3 vẫn được các tay đua tin dùng trong các giải đua xe chuyên nghiệp
  - Mazda RX-4 là mẫu xe điều hành được sản xuất tại Nhật từ 1966 tới 1991. Được marketing là một chiếc xe thể thao sang trọng.
  - Mazda RX-7 là mẫu xe thể thao được sản xuất từ năm 1978 tới 2002. Là dòng xe động cơ quay bán chạy số 1 mọi thời đại với 811,634 chiếc được bán ra.

4.1.3 Khoảng, trung bình, độ lệch chuẩn của từng thuộc tính? Chú ý tới những giá trị tiềm ẩn còn thiếu.

	mpg	cyclinder	displacement	horsepower
Range	37.6	5	387	184
Mean	23.504433	5.475369	194.779557	105.081281
STD	7.738736	1.712160	104.922458	38.480533

	weight	acceleration	model	origin
Range	3527	16.8	12	2
Mean	2979.413793	15.519704	75.921182	1.568966
STD	847.004328	2.803359	3.748737	0.797479

4.1.4 Vẽ histogram cho từng thuộc tính. Chú ý việc chọn số lượng bin phù hợp. Viết 2-3 câu tóm tắt những khía cạnh thú vị trong dữ liệu bằng cách nhìn vào histogram

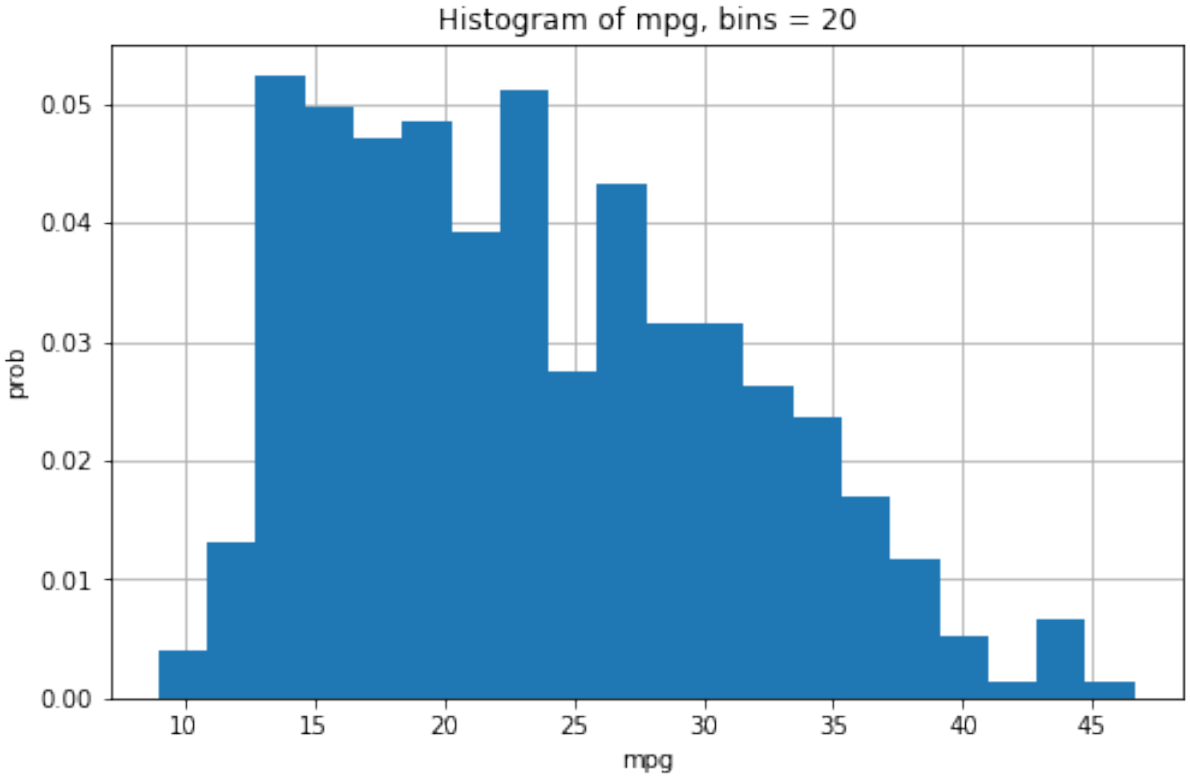

Figure 4: Histogram của mpg

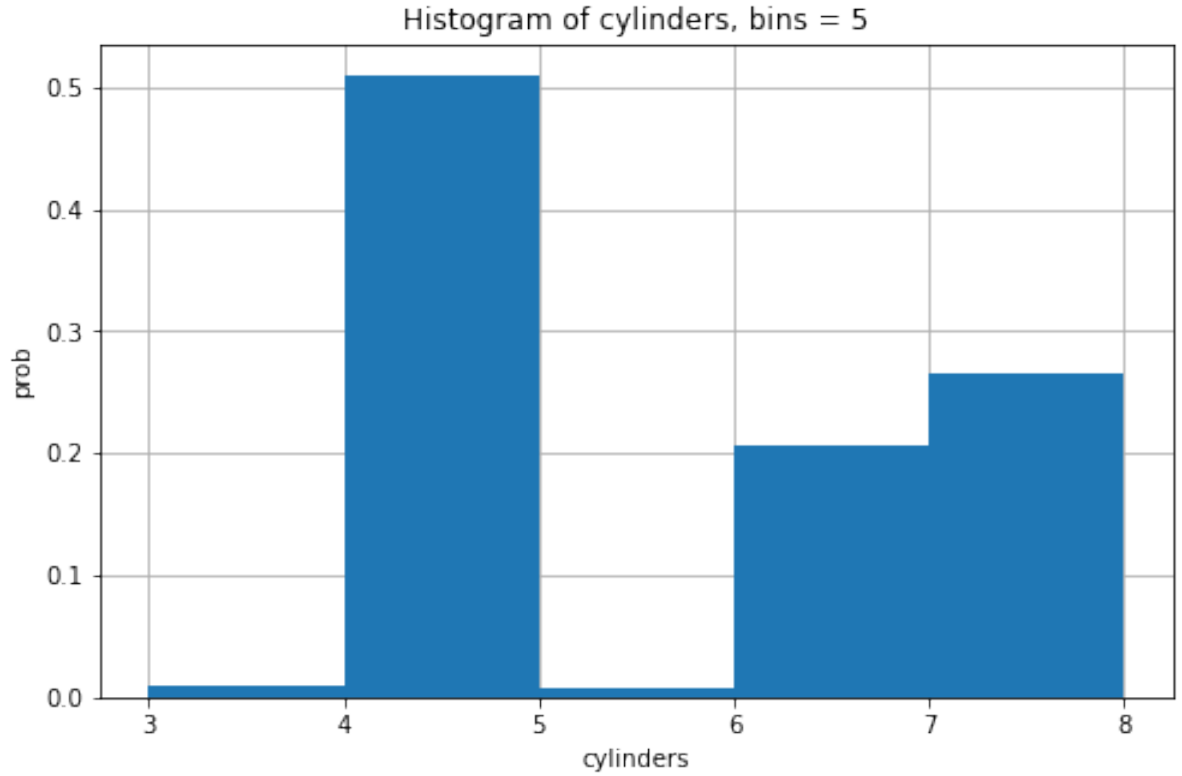

Figure 5: Histogram của cylinders

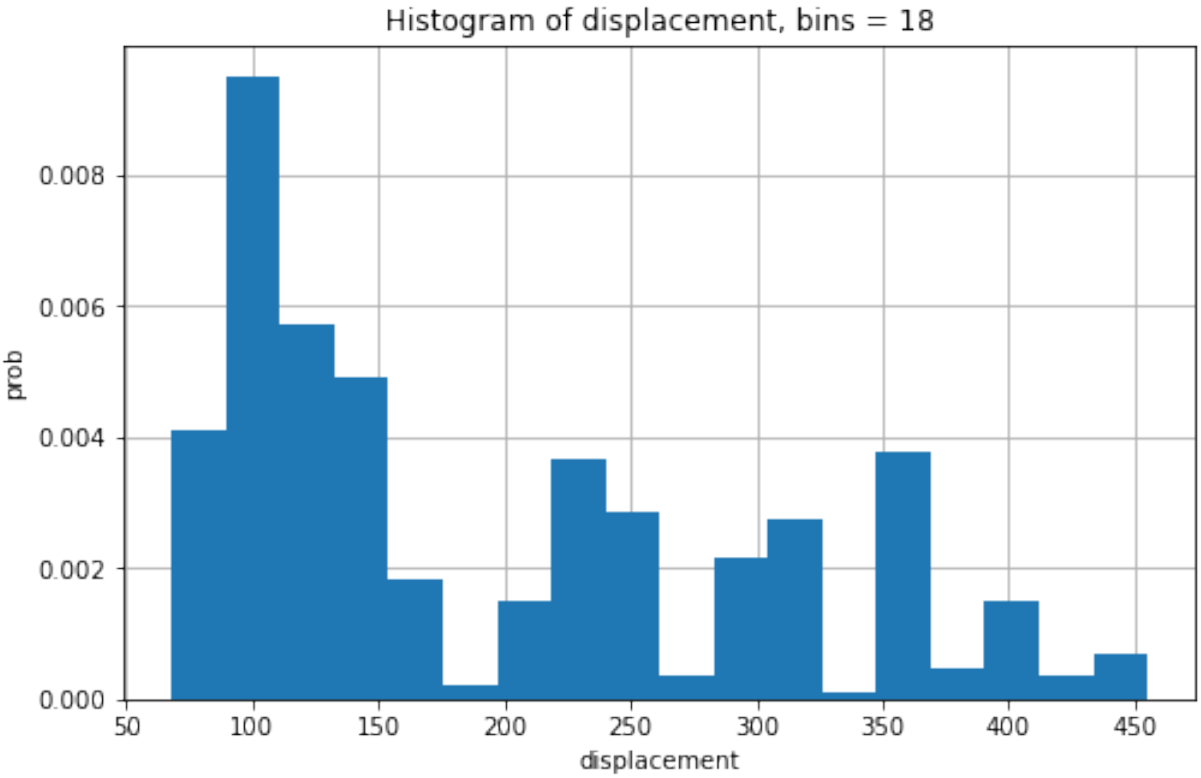

Figure 6: Histogram của displacement

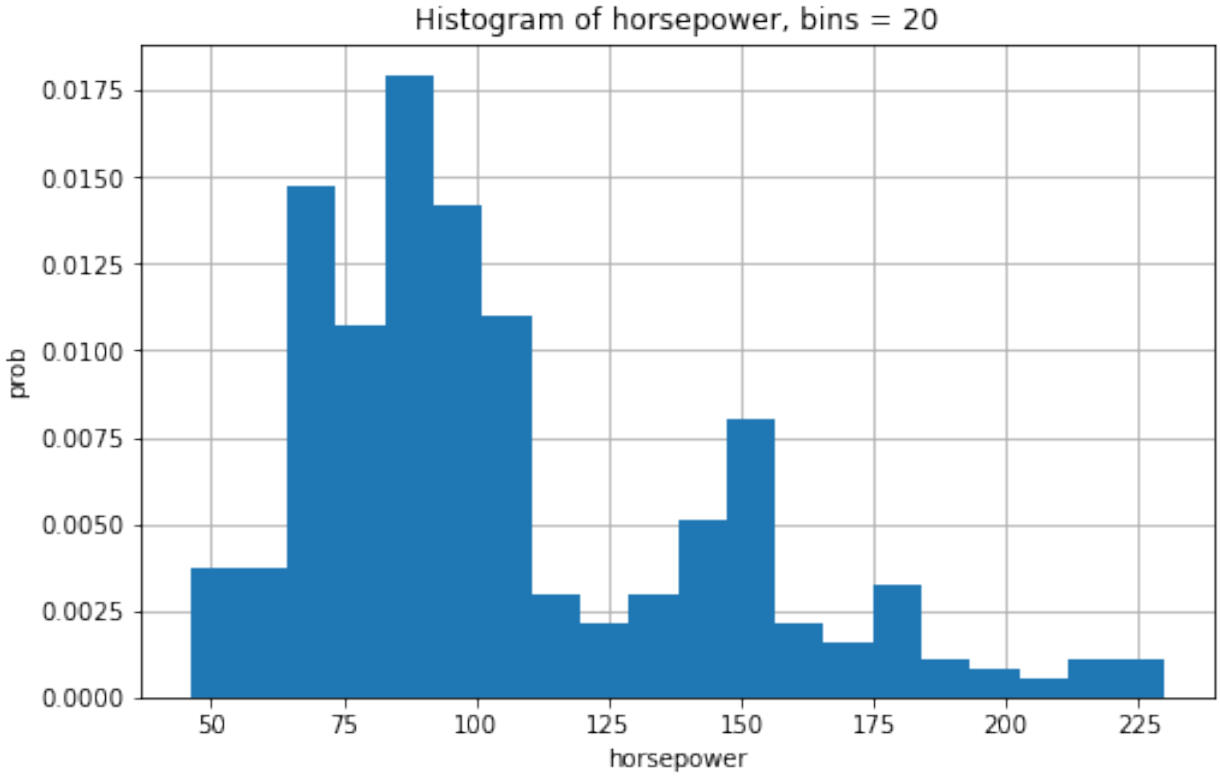

Figure 7: Histogram của horsepower

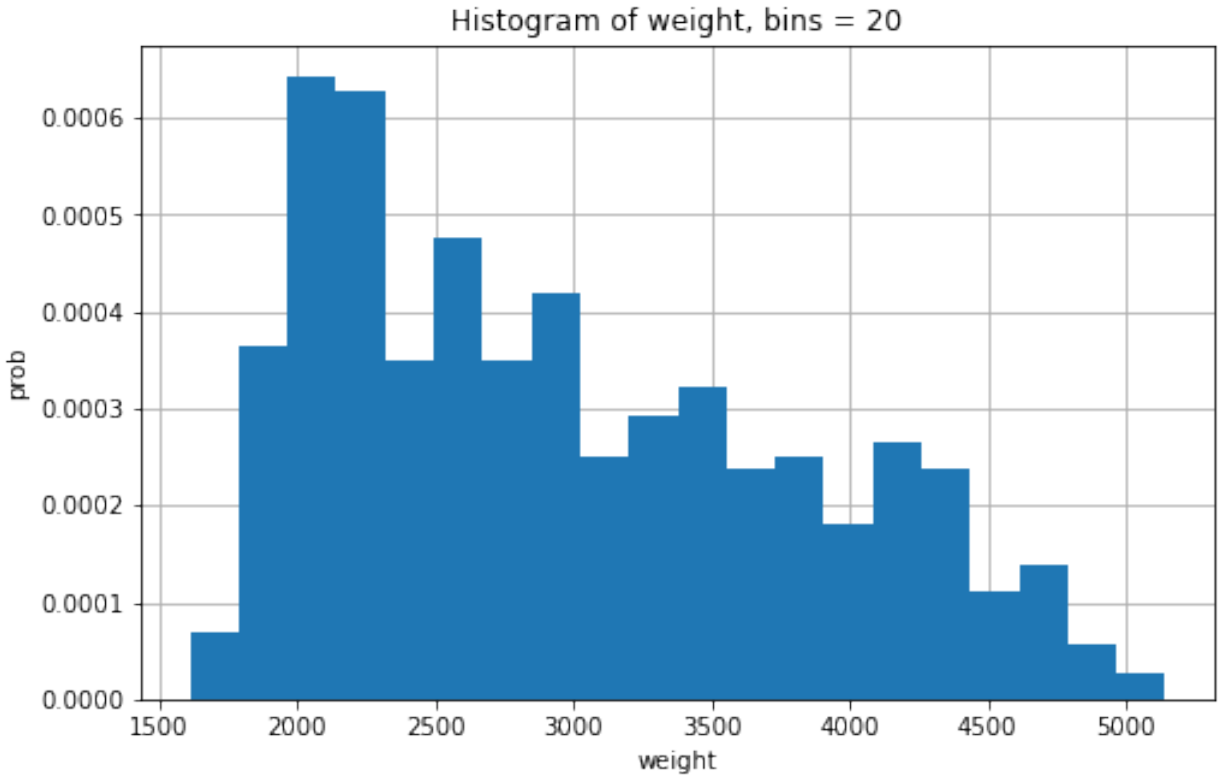

Figure 8: Histogram của weight

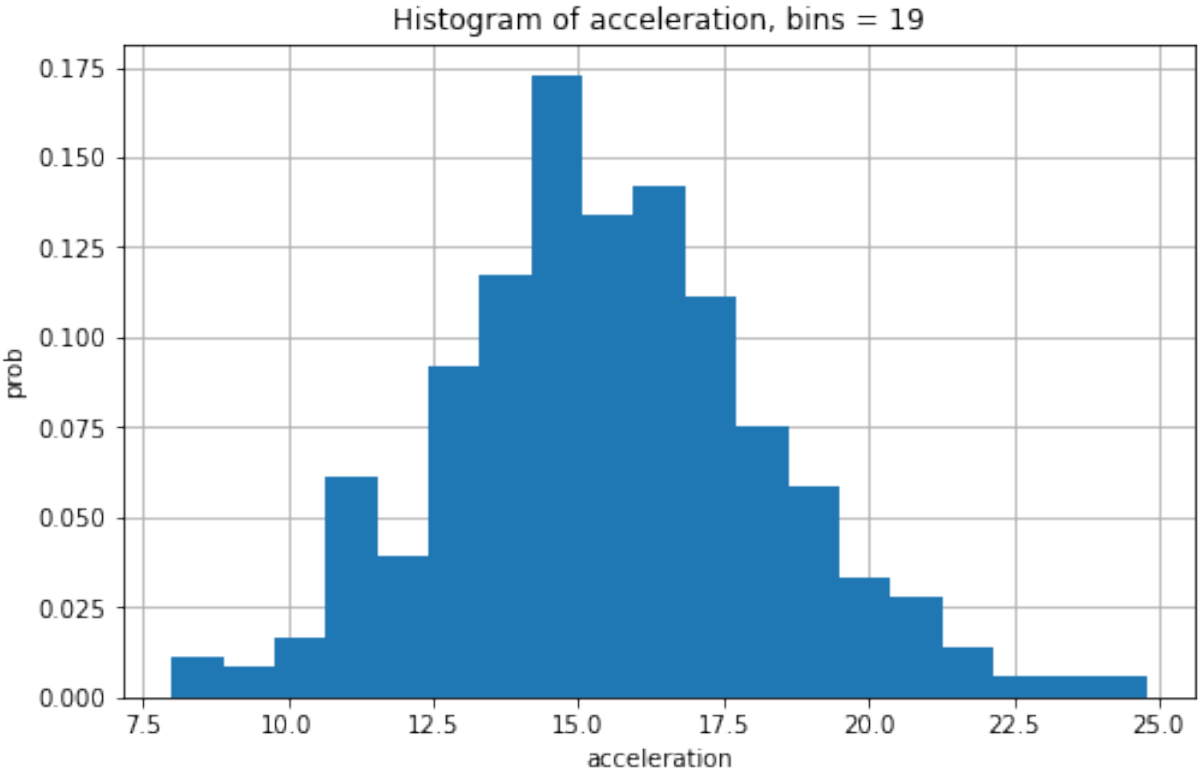

Figure 9: Histogram của acceleration

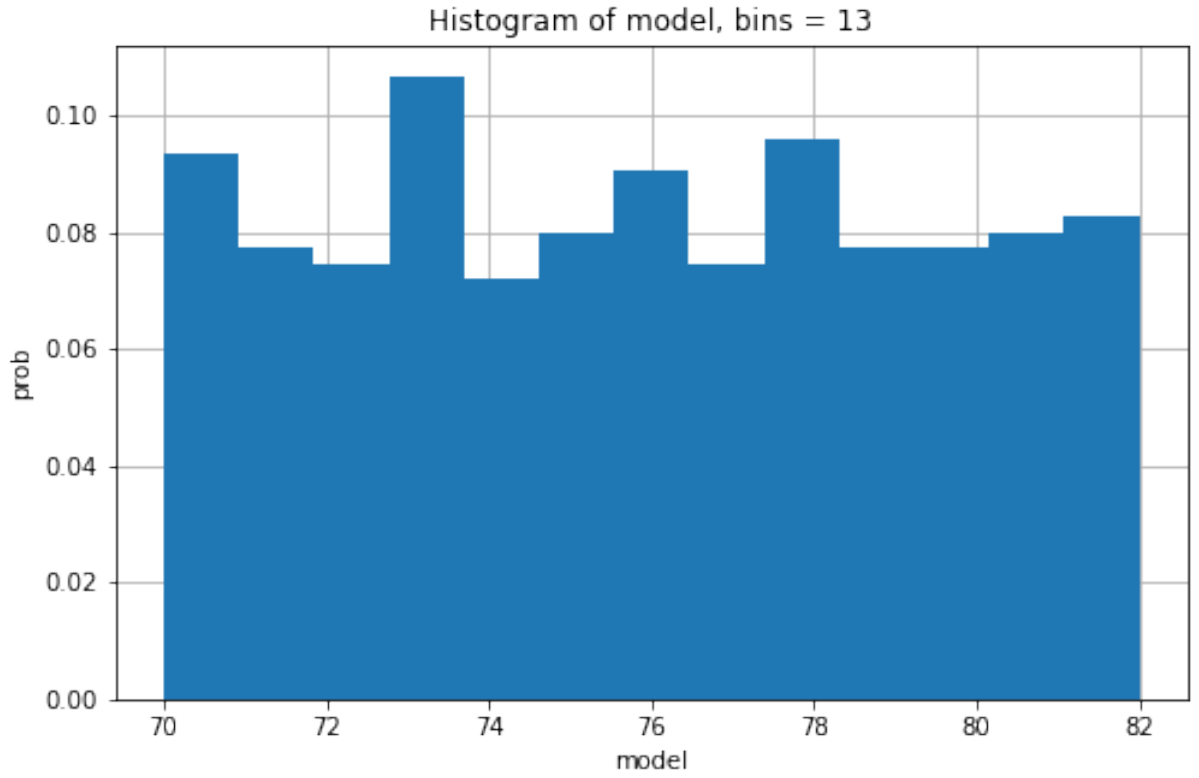

Figure 10: Histogram của model

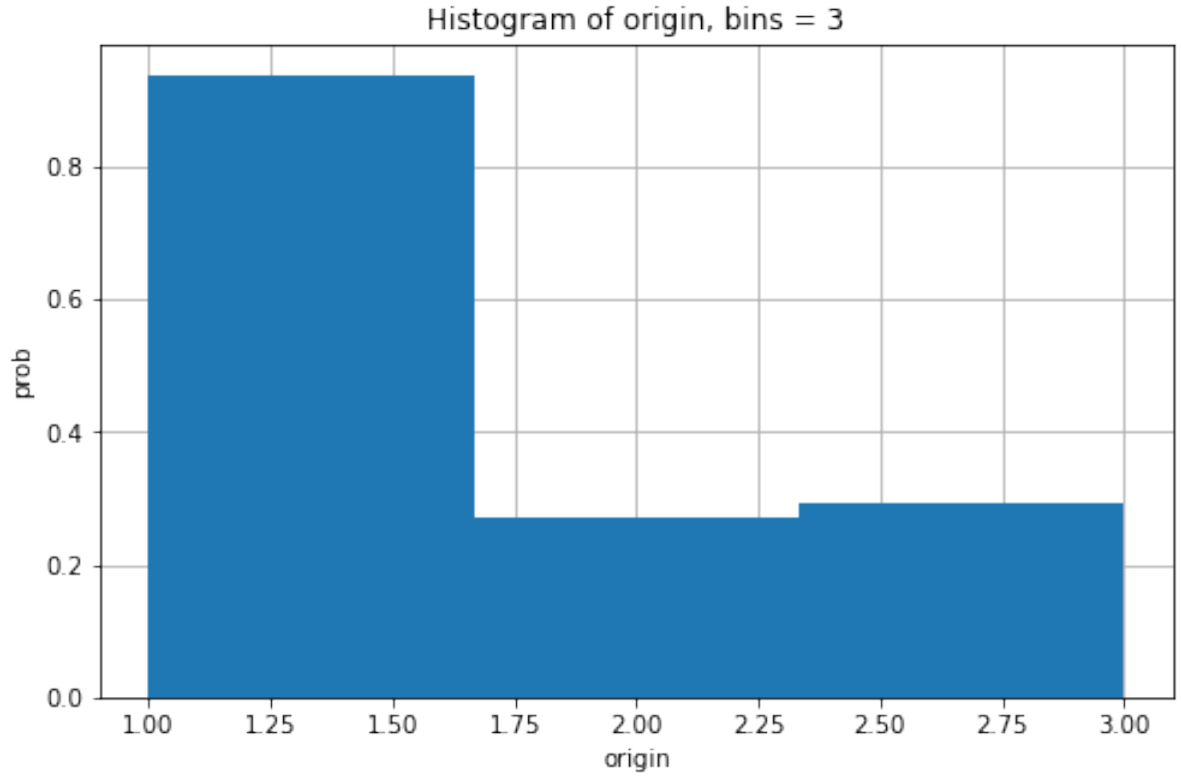

Figure 11: Histogram của origin

- Biểu đồ histogram cung cấp cái nhìn tổng quát về phân bố dữ liệu
- Các khía cạnh rút ra được từ histogram:
  - 'model' của các hãng xe được sản suất khá đồng đều, trung bình mỗi năm, các hãng đều có ra mẫu xe mới. Trong đó, năm 1973 cho ra đời nhiều mẫu xe nhất
  - Số lượng xi-lanh phổ biến nhất là 4, kế đến là 7 và 6
  - Thông thường, với 1 gallon nhiên liệu, 1 xe có thể đi được từ 12 đến 25 dặm

4.1.5 Vẽ scatterplot của thuộc tính weight với MPG. Có kết luận gì về quan hệ giữa hai thuộc tính này? Hệ số tương quan giữa hai thuộc tính là bao nhiêu?

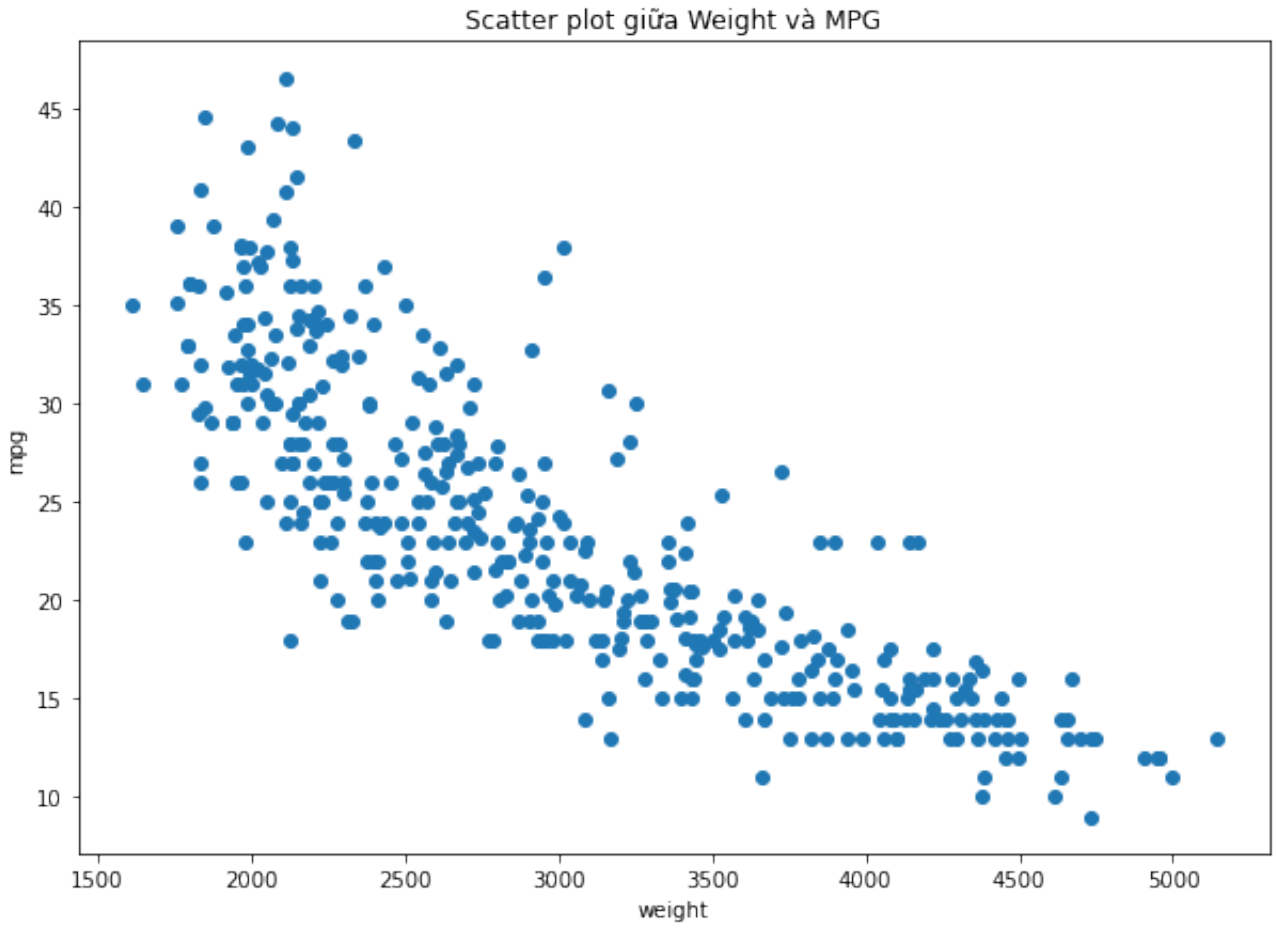

Figure 12: Scatter plot giữa weight và MPG

- Hệ số tương quan rất cao và âm chứng tỏ giữa 2 thuộc tính này tồn tại quan hệ nghịch biến
- Điều này là dễ hiểu vì khi tăng khối lượng xe và giữ nguyên lượng xăng thì xe sẽ phải tốn nhiều năng lượng hơn để di chuyển dẫn tới MPG (Số dặm đi được theo mỗi đơn vị nhiên liệu gallon) giảm.

4.1.6 Vẽ scatterplot của thuộc tính year với cylinders. Thêm một lượng noise ngẫu nhiên nhỏ vào giá trị để làm scatterplot đẹp hơn. Có thể kết luận được gì? Thực hiện tìm kiếm trên internet về lịch sử của công nghiệp xe hơi trong thập niên 70 có thể giải thích được kết quả

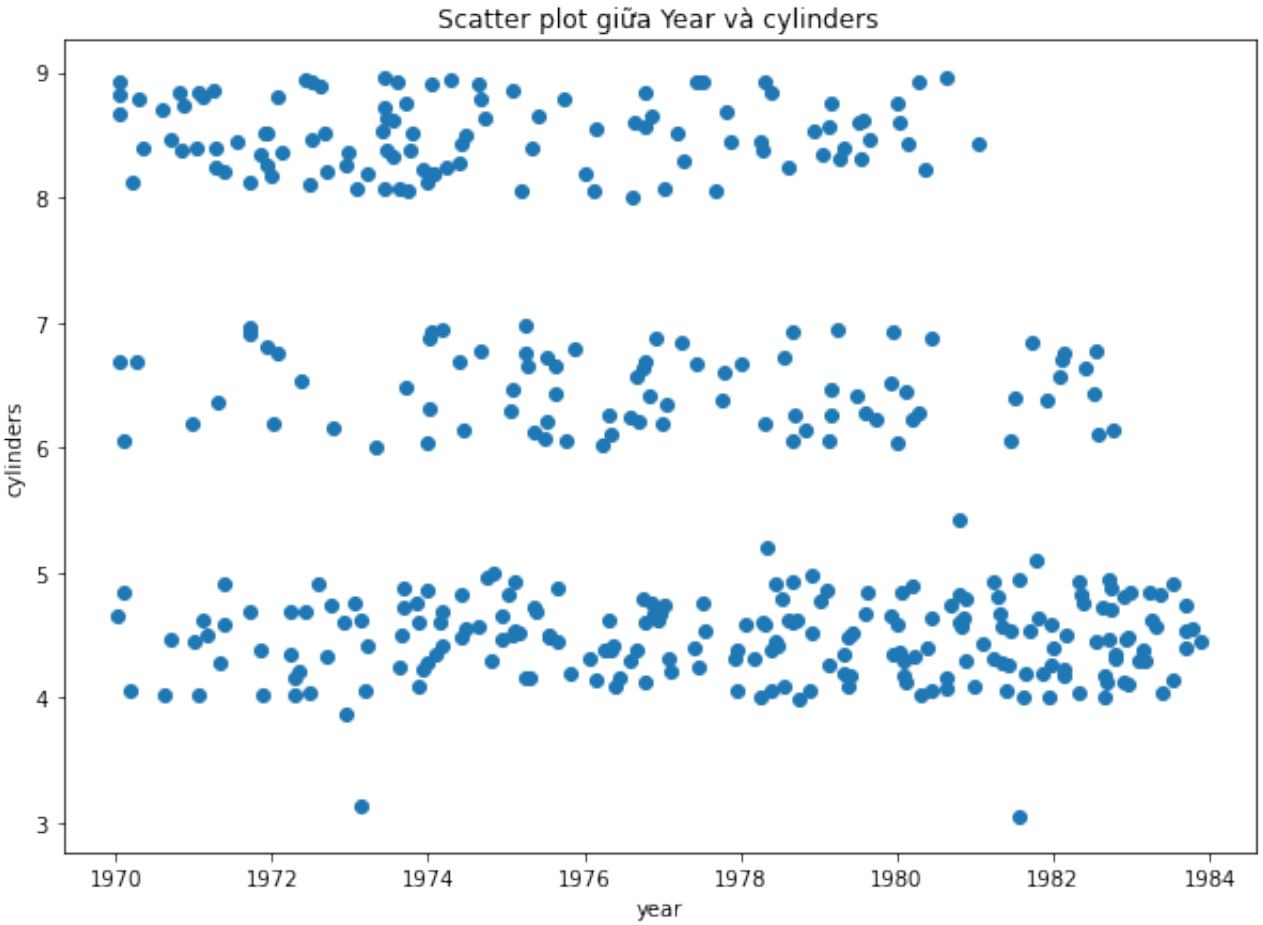

Figure 13: Scatter plot giữa year và cylinders

- Việc bỏ thêm nhiễu vào dữ liệu giúp tăng hiệu suất trực quan (dễ quan sát được mật độ cũng như phân bố của dữ liệu)
- Ngoài ra, còn có thể dễ dàng xác định các điểm dữ liệu outlier
- Nhiều mẫu 8 xi lanh được sản xuất nhiều ở giai đoạn đầu những năm 70 (70-74) và giảm dần theo thời gian và biến mất hẳn vào các năm cuối
- Ngược lại, xe 4 xi lanh vẫn luôn là mẫu xe phổ biến nhất và được sản xuất ngày càng nhiều.
- Lịch sử công nghiệp xe hơi những năm 70:
  - Khoảng thời gian đầu, mọi người đều thích sử dụng những chiếc xe có 8 xi lanh do đặc tính của nó: nhanh, mạnh, kiểu dáng đẹp, hầm hố. Đồng thời vào khoảng thời gian này, việc xe 8 xi-lanh thống trị ở các giải đua xe cũng góp phần định hướng thị trường xe hơi ngày đó.
  - Đây là khoảng thời gian mà thế giới xảy ra khủng hoảng dầu khí nên có rất nhiều đạo luật được ban hành (Emergency Petroleum Allocation Act - 1973; Energy Policy

and Conservation Act - 1975...) , ảnh hưởng trực tiếp tới việc sản xuất công nghiệp. Ngoài ra, khoảng thời gian này cũng là lúc hàng loạt đạo luật về khí thải ra đời. Các hãng xe hơi buộc phải đưa ra giải pháp để giảm 2 thứ: nhiên liệu tiêu thụ và khí thải. Điều này dẫn tới kết quả hiển nhiên trong việc các hãng xe gia tăng việc sử dụng các động cơ 4 xi-lanh để chế tạo xe hơi vì chúng nhỏ hơn, nhẹ hơn và thân thiện môi trường hơn và công suất của chúng cũng phù hợp cho nhu cầu phổ thông.

- Cũng trong khoảng thời gian này, bộ tăng áp (turbocharger) cũng được sử dụng ngày càng phổ biến vì nó giúp những động cơ 4, 6 xi lanh có sức mạnh như những động cơ có 8 xi lanh nhưng vẫn giữ được kích cỡ nhỏ gọn và sử dụng ít nhiên liệu và xả ít khí thải hơn. Đột phá thực sử xảy ra vào năm 1978 khi động cơ diesel tăng áp đầu tiên được ra đời theo chiếc Mercedes Benz 300SD, theo sau là động cơ VW Golf Turbodiesel vào năm 1981. Động cơ tăng áp sử dụng diesel có hiệu suất cao hơn và xả thải ít hơn hẳn động cơ chạy bằng gas ở trên các mẫu xe trước đó
- Các điểm outlier trên biểu đồ là bốn mẫu xe 3 xi lanh ra đời vào các năm 1972, 1973, 1977, 1980. Tất cả đều đến từ hãng xe Mazda vì họ muốn giữ sự khác biệt của mình với các hãng xe khác và loại động cơ xoay 3 xi-lanh mà họ sử dụng cũng có hiệu suất và lợi thế ổn định.

4.1.7 Thể hiện 2 scatterplot thú vị với bạn. Bàn luận về những gì bạn thấy

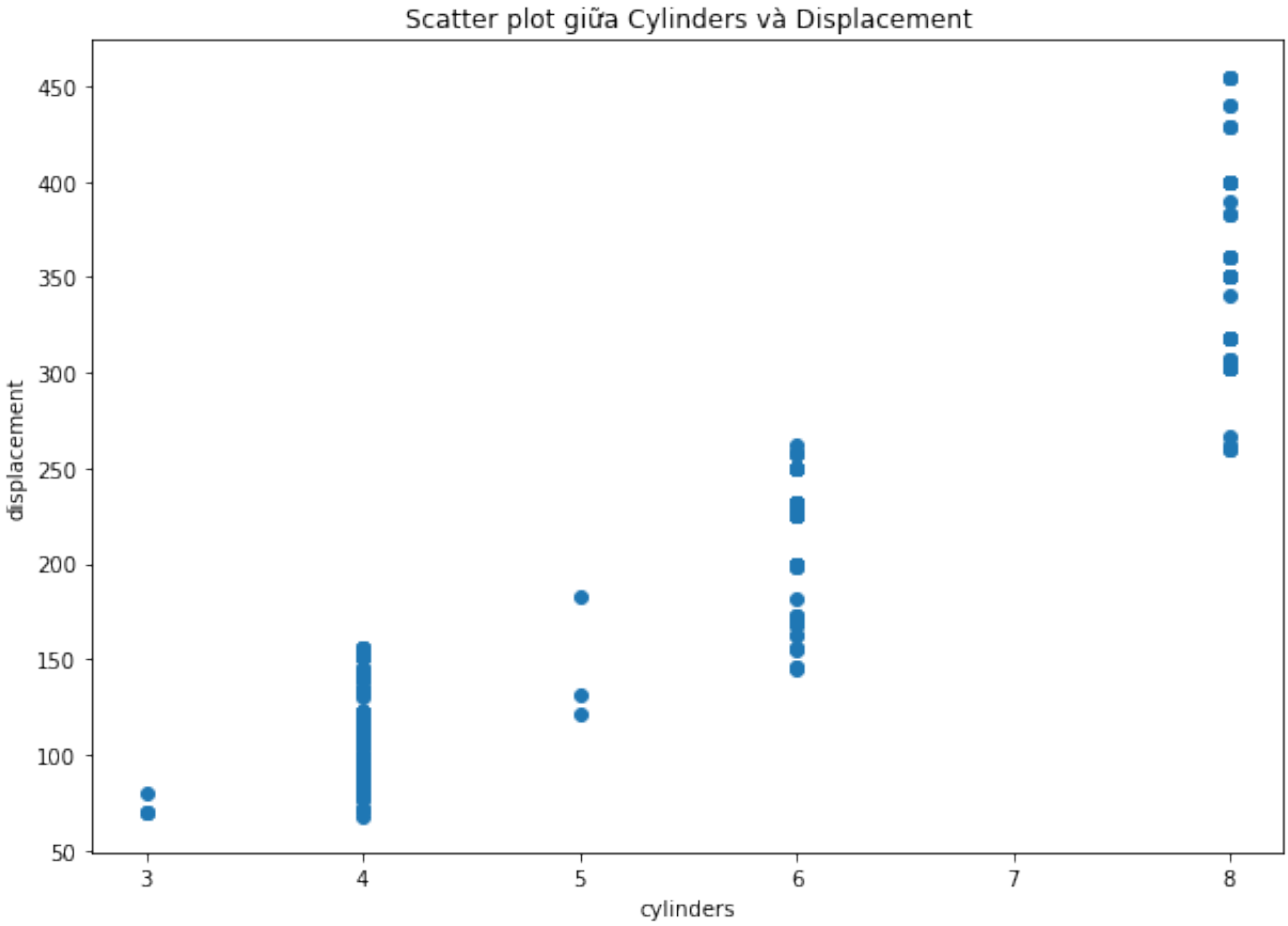

Figure 14: Scatter plot giữa cylinders và displacement

- Ta có thể quan sát rõ được sự tương quan thuận giữa số cylinder và displacement.

- Đúng như trong thực tế, những cỗ máy có nhiều xi-lanh thường đi với một dung tích lớn để tăng công suất cho xe.
- Động cơ 4 xi-lanh, dung tích nhỏ thường được sử dụng cho xe tầm trung, đa dụng nên có số lượng tương đối nhiều.

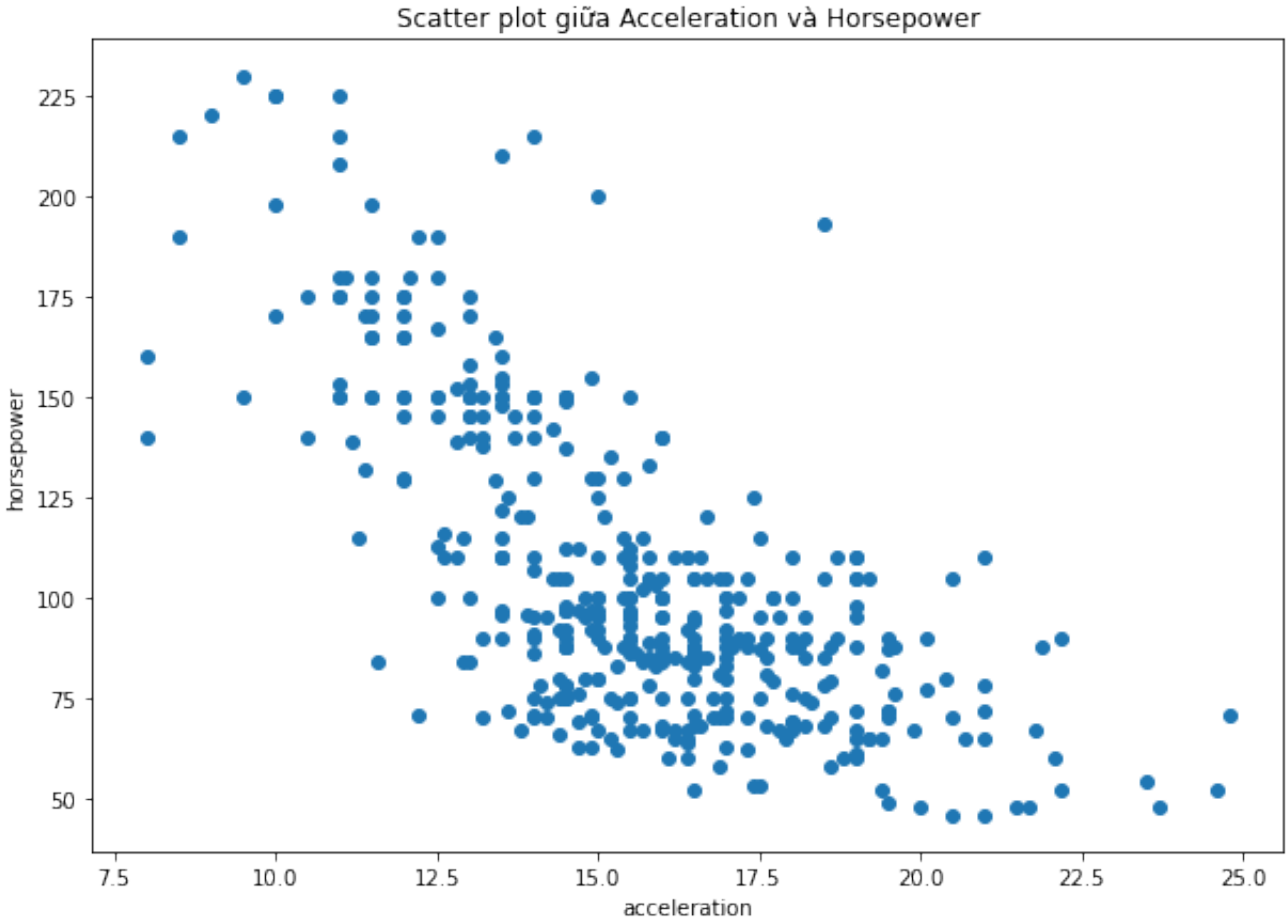

Figure 15: Scatter plot giữa acceleration và horsepower

4.1.8 Vẽ biểu đồ thời gian cho tất cả các công ty chỉ ra số xe hơi họ giới thiệu mỗi năm. Bạn có thấy xu hướng nào thú vị không?

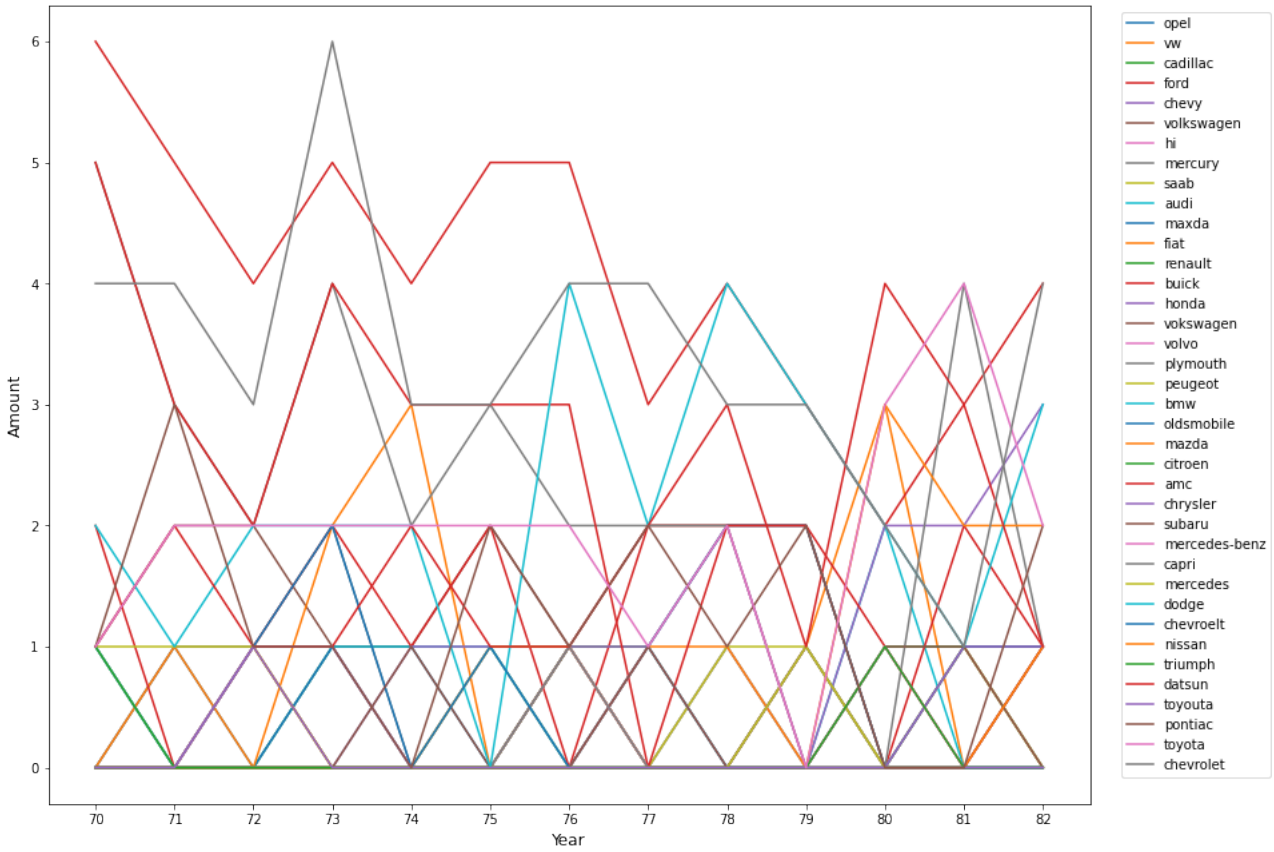

Figure 16: Scatter plot giữa acceleration và horsepower

- Các hãng lớn sẽ ra mắt từ 2 - 6 xe mỗi năm, còn các hãng còn lại thì chỉ từ 1 - 2 xe.
- Các hãng xe có xu hướng ra mắt giảm dần số lượng dần về những năm cuối thập kỷ và cho ra mắt hàng loạt số lượng xe mới vào nhứng năm đầu của thập kỷ kế tiếp.
- Biểu đồ khá khó nhìn. Nếu có chức năng tương tác để chọn 1 số công ty cụ thể thì sẽ cho kết quả trực quan tốt hơn.

4.1.9 Tính toán tương quan theo cặp, và vẽ heatmap với Matplotlib. Bạn có thấy tuông quan nào thú vị không?

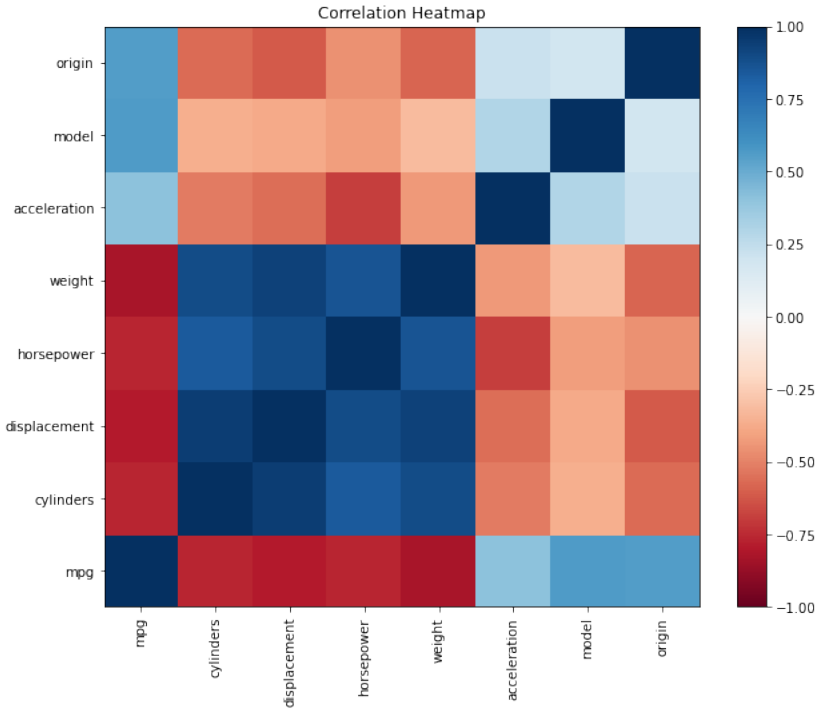

Figure 17: Heatmap thể hiện tương quan giữa các thuộc tính

- Ta có thể thấy sự tương quan thuận cao giữa các thông số cơ khí với nhau (cylinders, displacement, horsepower, weight), những thông số này đại diện cho phân khúc, công suất của từng chiếc xe.
- Thông số mpg tương quan nghịch cao với các thông số cơ khí trên thể hiện nếu xe càng mạnh, phân khúc càng cao thì sẽ càng tiêu hao nhiên liệu (mpg thấp).
- Thông số mpg tương quan thuận với model và origin thể hiện đời xe càng cao thì càng tiết kiệm nhiên liệu.
- Thông số acceleration tương quan nghịch với các thông số cơ khí trên thể hiện nếu xe càng to, công suất cao thì khả năng tăng tốc sẽ yếu.

5 Story of Electric Power Compsumtion Data

5.1 Plot 1

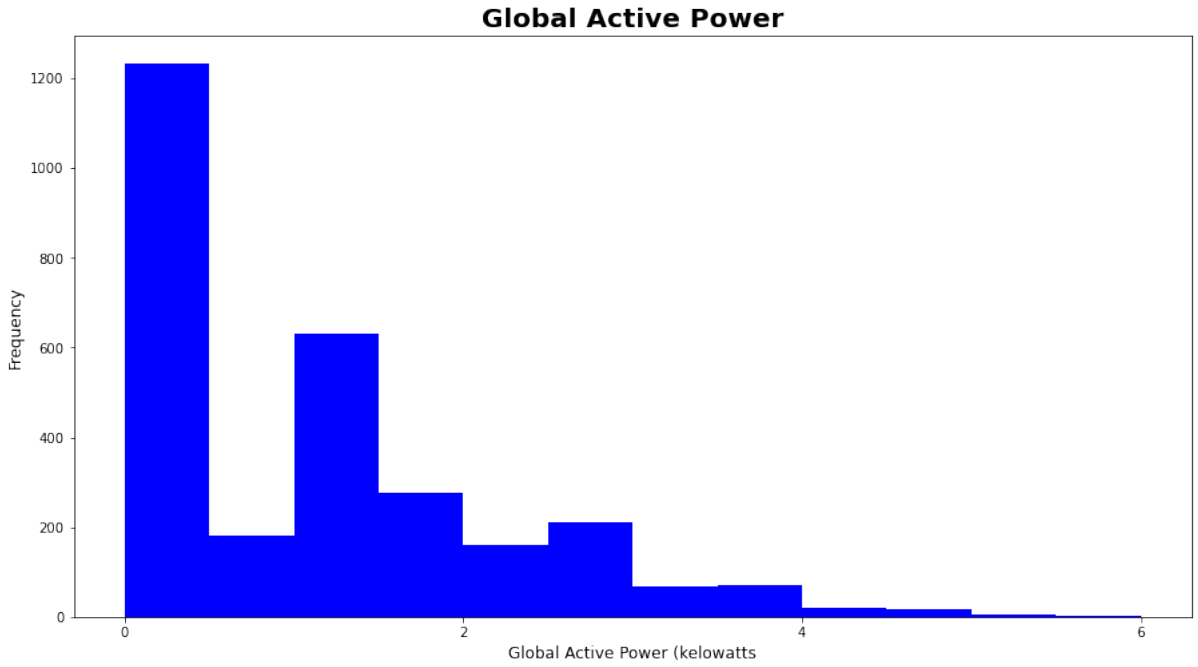

Figure 18: Tần suất xuất hiện của các mức công suất

- Hộ gia đình này sử dụng điện rất liên tục, luôn luôn có thiết bị điện được sử dụng trong nhà. Phần lớn thời gian sử dụng ở mức thấp (bật đèn...).

5.2 Plot 2

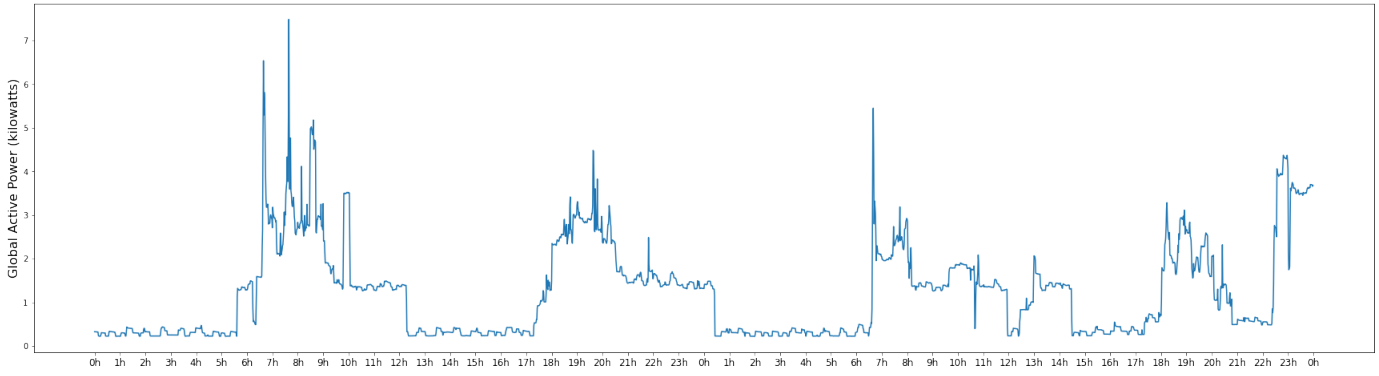

Figure 19: Mức công suất theo giờ

- Hộ gia đình này có những khoảng sử dụng điện cao điểm. Cụ thể hơn là vào từ 6h - 12h và 17h-0h.

5.3 Plot 3

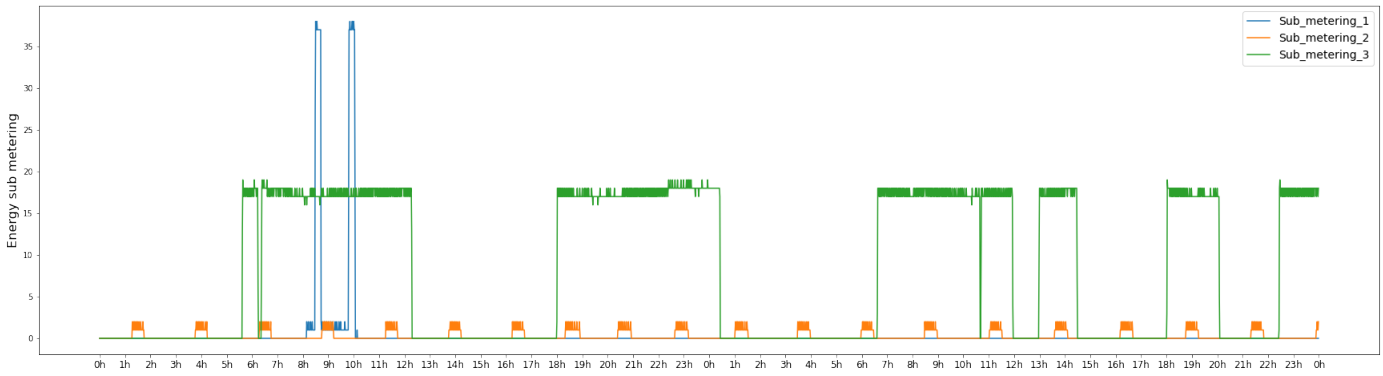

Figure 20: Công đo được từ 3 điện kế

- Có thể nhận biết rằng gia đình này đã nấu ăn vào khoảng 8h-10h ngày 01-02 nhưng lại không sử dụng bếp vào thời gian này ngày hôm sau, có thể là họ đã đặt thức ăn bên ngoài. Ngoài ra họ cũng không nấu ăn vào buổi chiều hay ban đêm.

- Ở phòng bếp luôn có một chiếc tủ lạnh sử dụng điện theo chu kì theo chu kì 2 tiếng thì sẽ làm lạnh 30p.

- Máy sưởi và điều hòa được sử dụng khi trời sáng và khi đi ngủ(vào 2 khung giờ: 5h30 12h và 6h 0h), chỉ không sử dụng vào thời điểm trưa - chiều, khi trời ấm hơn. Một điểm bất thường là vào ngày 02-02 họ không dùng điều hòa vào lúc 20h-23h. Có thể lúc đấy gia đình họ đã đi ra ngoài chơi hoặc dự tiệc.

5.4 Plot 4

Figure 21: Một số biểu đồ khác

- Khảo sát các biểu đồ khác cũng cho thấy một số kết quả tương tự về khoảng giờ cao điểm. Hiệu điện thế sẽ giảm khi có nhiều thiết bị sử dụng điện. Mức công phản kháng cũng có sự tăng giảm đồng đều với mức công suất sử dụng.

6 Tham khảo

- https://classiccars.fandom.com/wiki/Mazda_RX-7
- https://classiccars.fandom.com/wiki/Mazda_RX-4
- https://en.wikipedia.org/wiki/Mazda_RX-7
- https://en.wikipedia.org/wiki/Mazda_Luce
- https://en.wikipedia.org/wiki/Mazda_Grand_Familia
- 2021: 50TH ANNIVERSARY OF THE MAZDA RX-3
- https://en.wikipedia.org/wiki/Mazda_Capella#RX-2
- http://www.turbos.bwauto.com/en/products/turbochargerHistory.aspx
- The History of the Mazda 3: A Look Back Through Time