|  |  |
| --- | --- |
|  | ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  KHOA CÔNG NGHỆ THÔNG TIN |

**TRỰC QUAN HÓA DỮ LIỆU**

LAB 03

**TRỰC QUAN HOÁ DỮ LIỆU VỚI**

**NUMPY, PANDAS, MATPLOTLIB**

**GVHD: Thầy Bùi Tiến Lên**

**Thầy Lê Ngọc Thành**

|  |  |
| --- | --- |
| Mã nhóm : 05 |  |
| 1. Võ Thế Minh | * 18120211 |
| 1. Lê Đức Thành | * 18120238 |
| 1. Nguyễn Thị Ngọc Trâm | * 18120246 |
| 1. Nguyễn Huy Tú | * 18120254 |

**MỤC LỤC**

Thành phố Hồ Chí Minh, tháng 06/2021

[1. Danh mục hình 3](#_Toc74668340)

[2. Danh mục bảng 5](#_Toc74668341)

[3. Danh sách thành viên 6](#_Toc74668342)

[4. Đánh giá 6](#_Toc74668343)

[5. Giới thiệu về Numpy, Matplotlib, Pandas 7](#_Toc74668344)

[5.1. Numpy 7](#_Toc74668345)

[5.2. Matplotlib 7](#_Toc74668346)

[5.3. Pandas 8](#_Toc74668347)

[6. Task 03: Exploratory analysis of Car MPG data 10](#_Toc74668348)

[6.1. Câu hỏi 1 10](#_Toc74668349)

[6.2. Câu hỏi 2 10](#_Toc74668350)

[6.3. Câu hỏi 3 12](#_Toc74668351)

[6.4. Câu hỏi 4 13](#_Toc74668352)

[6.5. Câu hỏi 5 16](#_Toc74668353)

[6.6. Câu hỏi 6 17](#_Toc74668354)

[6.7. Câu hỏi 7 19](#_Toc74668355)

[6.8. Câu hỏi 8 21](#_Toc74668356)

[6.9. Câu hỏi 9 24](#_Toc74668357)

[7. Task 04: Electric power consumption data 26](#_Toc74668358)

[7.1. Thực hiện vẽ lại plot1 26](#_Toc74668359)

[7.1.1. Các bước tiến hành 26](#_Toc74668360)

[7.1.2. Story rút ra từ biểu đồ 29](#_Toc74668361)

[7.2. Thực hiện vẽ lại plot 2 29](#_Toc74668362)

[7.2.1. Các bước thực hiện 29](#_Toc74668363)

[7.2.2. Story rút ra từ biểu đồ 30](#_Toc74668364)

[7.3. Thực hiện vẽ lại plot 3 31](#_Toc74668365)

[7.3.1. Các bước thực hiện 31](#_Toc74668366)

[7.3.2. Story rút ra từ biểu đồ 32](#_Toc74668367)

[7.4. Thực hiện vẽ lại plot 4 33](#_Toc74668368)

[7.4.1. Các bước thực hiện 33](#_Toc74668369)

[7.4.2. Story rút ra từ biểu đồ 34](#_Toc74668370)

[8. Tham khảo 35](#_Toc74668371)

1. Danh mục hình

[Hình 1 Sử dụng Numpy (np) để tạo độ nhiễu cho số liệu để thực hiện vẽ biểu đồ Scatter 7](#_Toc74668304)

[Hình 2 Vẽ Scatter Plot sử dụng thư viện Matplotlib 8](#_Toc74668305)

[Hình 3 Sử dụng pandas để import data vào dataframe để tiến hành xử lý 9](#_Toc74668306)

[Hình 4 Thực hiện yêu cầu Task 3 câu 1 10](#_Toc74668307)

[Hình 5 Thực hiện yêu cầu Task 3 câu 2 11](#_Toc74668308)

[Hình 6 Thực hiện Task 3 câu 3 13](#_Toc74668309)

[Hình 7 Thực hiện Task 3 câu 4 (phần code) 14](#_Toc74668310)

[Hình 8 Thực hiện Task 3 câu 4 (kết quả) 15](#_Toc74668311)

[Hình 9 Thực hiện Task 3 câu 5 (phần code) 16](#_Toc74668312)

[Hình 10 Thực hiện Task 3 câu 5 (kết quả) (1) 16](#_Toc74668313)

[Hình 11 Thực hiện Task 3 câu 5 (kết quả) (2) 17](#_Toc74668314)

[Hình 12 Thực hiện Task 3 câu 6 18](#_Toc74668315)

[Hình 13 Scatter (1) giữa biến displacement và horsepower Task 3 câu 7 20](#_Toc74668316)

[Hình 14 Scatter (2) giữa biến xi lanh và khối lượng Task 3 câu 7 21](#_Toc74668317)

[Hình 15 Thực hiện Task 3 câu 8 (phần code) 22](#_Toc74668318)

[Hình 16 Biểu đồ số mẫu xe ra hàng năm của từng hãng trong khoảng thời gian 1970-1982 23](#_Toc74668319)

[Hình 17 Biểu đồ tổng mẫu xe ra mắt ở các năm 1970 đến 1982 24](#_Toc74668320)

[Hình 18 Thực hiện Task 3 câu 9 (phần code) 25](#_Toc74668321)

[Hình 19 Thực hiện Task 3 câu 9 (phần kết quả) 25](#_Toc74668322)

[Hình 20 Include các thư viên để tái tạo plot1 26](#_Toc74668323)

[Hình 21 Plot 1. Xoá các dòng dữ liệu sai 26](#_Toc74668324)

[Hình 22 Plot 1. Ghép cột thời gian và cột ngày tháng 26](#_Toc74668325)

[Hình 23 Plot 1. Thực hiện lọc data theo ngày hợp lệ 27](#_Toc74668326)

[Hình 24 Plot 1. Chuyển dữ liệu cột thành dạng datetime sử dụng hàm của pandas 27](#_Toc74668327)

[Hình 25 Plot 1. Thay thế dữ liệu ? thành 0 27](#_Toc74668328)

[Hình 26 Plot 1. Kết quả sau khi xử lý dữ liệu 27](#_Toc74668329)

[Hình 27 Plot 1. Đoạn code để vẽ biểu đồ 28](#_Toc74668330)

[Hình 28 Plot 1. Kết quả tái tạo 28](#_Toc74668331)

[Hình 29 Plot 2. Đoạn code vẽ plot 2 29](#_Toc74668332)

[Hình 30 Plot 2. Kết quả tái tạo 30](#_Toc74668333)

[Hình 31 Plot 3. Tiến hành import thư viện, data 31](#_Toc74668334)

[Hình 32 Plot 3. Kết quả sau khi xử lý dữ liệu 31](#_Toc74668335)

[Hình 33 Plot 3. Đoạn code vẽ plot 3 31](#_Toc74668336)

[Hình 34 Plot 3. Kết quả tái tạo 32](#_Toc74668337)

[Hình 35 Plot 4. Đoạn code vẽ hình 33](#_Toc74668338)

[Hình 36 Plot 4. Kết quả tái tạo 33](#_Toc74668339)

1. Danh mục bảng

Bảng 1 Danh sách thành viên 6

Bảng 2 Danh sách yêu cầu 6

Bảng 3 Kết quả tự đánh giá 6

Bảng 4 Lịch sử phát triển NumPy 7

1. Danh sách thành viên

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **MSSV** |
| SV1 | Võ Thế Minh | 18120211 |
| SV2 | Lê Đức Thành | 18120238 |
| SV3 | Nguyễn Thị Ngọc Trâm | 18120246 |
| SV4 | Nguyễn Huy Tú | 18120254 |

Bảng 1 Danh sách thành viên

1. Đánh giá

|  |  |  |
| --- | --- | --- |
| **STT** | **Yêu cầu** | **Đánh giá (%)** |
| 1 | Task 3: Exploratory analysis of Car MPG data gồm 9 câu hỏi nhỏ | 100% |
| 2 | Task 4: Electri power consumption data gồm 4 câu hỏi nhỏ | 100% |
| 3 | Giới thiệu về Numpy, Matplotlib, Pandas | 100% |

Bảng 2 Danh sách yêu cầu

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Yêu cầu** | **Thực hiện** | **Đánh giá (%)** |
| 1 | Task 3 + câu hỏi phụ | SV2, SV3 | 100% |
| 2 | Task 4 | SV1,SV4 | 100% |
| 3 | Giới thiệu về Numpy, Matplotlib, Pandas | SV2 | 100% |
| 4 | Tổng hợp và chỉnh sửa báo cáo | Cả nhóm | 100% |

Bảng 3 Kết quả tự đánh giá

1. Giới thiệu về Numpy, Matplotlib, Pandas
   1. Numpy

Numpy là thư viện phổ biến sử dụng cho ngôn ngữ lập trình python bổ sung nhiều tính năng về ma trận và mảng đa chiều cộng với một số lớn các hàm về toán học. Tiền thân của Numpy là Numeric được tạo bởi Jim Hugunin với sự cộng tác thêm của một số lập trình viên khác. Vào năm 2005, Travis Oliphant tạo ra Numby bằng cách kết hợp các tính năng của Numarray với Numeric với thêm một số tinh chỉnh.

Lịch sử phát triển Numpy:

|  |  |
| --- | --- |
| Mốc thời gian | Sự kiện diễn ra |
| 1995 | Nhóm matrix-sig được thành lập với mục tiêu tạo 1 gói chuyên về xử lý mảng trong đó có người thiết kế ra Python Guido van Rossum  Nhóm hoàn thành gói package và đặt tên là Numeric |
| 2005 | Version cuối cùng của Numeric (v24.2) được phát hành |
| 2006 | Version cuối cùng của numarray (v1.5.2) được phát hành |
| Đầu 2005 | Travis Oliphant bắt đầu quá trình kết hợp 2 phần mềm Numarray và Numeric |
| 2006 | NumPy phiên bản 1.0 được phát hành |
| 2011 | NumPy ra mắt phiên bản 1.5.0 hỗ trợ Python 3 |

Bảng Lịch sử phát triển NumPy

Một số tính năng chính:

NumPy có các tính năng tương tự như MATLAB cho phép người dùng xử lý nhanh các công việc liên quan đến ma trận, vector, mảng, đại số tuyến tính, giải tích và thị giác máy tính



Hình Sử dụng Numpy (np) để tạo độ nhiễu cho số liệu để thực hiện vẽ biểu đồ Scatter

* 1. Matplotlib

Matplotlib là thư viện dùng để vẽ đồ thị cho ngôn ngữ lập trình Python có tương tác mật thiết với NumPy. Matplotlib cung cấp API hướng đối tượng phục vụ việc đính kèm biểu đồ vào các ứng dụng sử dụng các toolkit như Tkinter, wxPython, Qt, GTK. Ngoài ra còn có pylab ( một ngôn ngữ thủ tục) dựa trên state machine để mô phỏng lại Matlab

Matplotlib được viết bởi John D Hunter. Từ khi được ra mắt, matplotlib được cập nhật liên tục bởi cộng ddoongf lớn. Matplotlib 2.0.x hỗ trợ Python phiên bản 2.7 đến 3.10. Matplotlib đã quyết định dừng hỗ trợ Python 2 sau 2020 sau khi ký điều khoản với Python 3

Một số toolkit của Matplotlib:

* Basemap: vẽ map với các tính năng về đường biên giới, đường bờ biển, …
* Cartopy: thư viên mapping sử dụng các kỹ thuật về arbitrary point, line, polygon,…
* Excel tools: tiện tích để trao đổi dữ liệu với Excel
* GTK tools: interface thư viện GTK
* Mplot3d: dùng để vẽ biểu đồ 3D

Chart, scatter chart

Description automatically generated

Hình Vẽ Scatter Plot sử dụng thư viện Matplotlib

* 1. Pandas

Pandas là thư viện được viết cho ngôn ngữ lập trình python nhằm mục đích xử lý và phân tích dữ liệu. Nó cung cấp cấu trúc dữ liệu và các phương thức để xử lý các bảng số liệu và chuỗi số. Tên pandas được lấy cảm hứng từ Panel Data ( 1 thuật ngữ kinh tế nói về một bộ dữ liệu được quan sát ở nhiều khoảng thời gian khác nhau).

Wes McKinney bắt đầu phát triển Pandas vào năm 2008 khi còn ở AQR đang cần một công cụ mạnh mẽ để giúp họ phân tích dữ liệu tài chính. Trước khi rời AQR, anh được cho phép biến dự án thành open source

Năm 2012, một nhân viên khác của AQR rời công ty và gia nhập pandas trở thành người đóng góp quan trọng thứ 2 cho dự án

Một số tính năng của thư viện:

* Đối tượng Dataframe để xử lý dữ liệu và index ( xác định vị trí)
* Công cụ để đọc và ghi dữ liệu trong cấu trúc dữ liệu và các loại file khác nhau
* Xử lý các dữ liệu bị thiếu
* Xử lý lại bộ dữ liệu bằng cách tái định hình hoặc tạo pivot
* Xử lý cắt nhỏ bộ dữ liệu thành các bộ nhỏ hơn theo tên cột, index
* Xử lý chèn thêm cột dữ liệu hoặc xoá bỏ
* Group by xếp loại data
* Trộn các dataset lại với nhau
* Các chức năng về xử lý thời gian như: tạo khoảng thời gian

Text

Description automatically generated

Hình Sử dụng pandas để import data vào dataframe để tiến hành xử lý

1. Task 03: Exploratory analysis of Car MPG data
   1. Câu hỏi 1

Câu hỏi:

* How many cars and how many attributes are in the data set
* Tạm dịch: có bao nhiêu phương tiện và bao nhiều biến trong bộ dữ liệu)

Kết quả:

Graphical user interface, text, application

Description automatically generated

Hình Thực hiện yêu cầu Task 3 câu 1

* 1. Câu hỏi 2

Câu hỏi:

* 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 produced 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
* Tạm dịch: có bao nhiêu nhà sản xuất ô tô trong bộ dữ liệu? Phương tiện nào có MPG tốt nhất? Nhà sản xuất nào sản xuất nhiều xe V8 nhất? Top 3 xe V3? Khảo sát một số thông tin về 3 chiếc xe trên

Kết quả:

Text

Description automatically generated with low confidence

Hình Thực hiện yêu cầu Task 3 câu 2

Khảo sát 4 xe 3 xi lanh:

1. Mazda Rx2

Chiếc mazda rx2 coupe được sản xuất từ năm 1970-1976. Xe được sản xuất ở phiên bản bao gồm Deluxe sedan, super deluxe sedan và coupe. Mazda Rx2 được sản xuất phần lớn tại Hiroshima Nhật Bản

Một số thông số của Mazda Rx2:

* 12A rota đôi
* hộp số 4 cấp
* 4 chỗ (sedan) hoặc 2 chỗ (coupe)
* 130 mã lực

Giá một chiếc Mazda Rx2 giao động từ $2998 đến $3958

1. Mazda RX3

Chiếc mazda RX3 được sản xuất từ năm 1971 đến năm 1978. Xe được sản xuất ở 3 dạng chính: coupe (2 chỗ ngồi), sedan(4 chỗ ngồi), và dạng honeycomb 7 chỗ ngồi. Ngoài ra, mazda RX3 còn có phiên bản thể thao GT với động cơ có mã lực lớn hơn, hệ thống giảm xóc nhỏ tiếng hơn

Chiếc mazda RX3 sử dụng dây chuyên mới của mazda trong việc sản xuất động cơ xe (TCP - Transplant Coating Process). Dây chuyền mới giúp giải quyết vấn đề về việc mạ rotor.

Một số thông số của Mazda RX3:

* 10A rotor đôi dung tích 982cc
* 125 mã lực với phiên bản GT, 109 mã lực với các phiên bản thường
* Giảm xóc phía trước sử dụng lò xo cuộc và bộ hấp thụ
* Hộp số 5 cấp

1. Mazda RX4

RX4 hay còn được gọi là Mazda Luce được sản xuất từ năm 1966 tới năm 1991. Nó được xuất khẩu dưới tên Mazda 929. Rx4 được đánh giá cao về thiết kế nội thất giúp người lái thoải mái. Rx4 cũng được sản xuất ở 2 phiên bản 4 chỗ ngồi và 2 chỗ ngồi

Một số thông số của Mazda RX4:

* Hộp số 4 cấp
* Sử dụng 3 loại động cơ tuỳ phiên bản: 1.5L UB l4, 1.8L VB L4, 1.3L 13A
* Tạo tối đa 135 mã lực

1. Mazda Rx7

là mẫu xe thể thao (sport car) của mazda được sản xuất từ năm 1978 đến năm 2002. Đã có hơn 800.000 chiếc được chế tạo. RX-7 còn được đánh giá là 1 trong 5 chiếc xe tốt nhật bởi tạp chí Car and Driver 5 lần khác nhau.

Một số thông số của Mazda RX7

* Tạo tối đa 250 mã lực
* Sử dụng động cơ 12A hoặc 13B
* Sử dụng hộp số 4 cấp hoặc 5 cấp
  1. Câu hỏi 3

Câu hỏi:

* What is the range, mean, and standard deviation of each attribute? Pay attention to potential missing values.
* Tạm dịch: Tìm khoảng (biên độ), trung vị và độ lệch chuẩn của từng biến

Kết quả:

Table

Description automatically generated

Hình Thực hiện Task 3 câu 3

* 1. Câu hỏi 4

Câu hỏi:

* 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.
* Tạm dịch: Vẽ biểu đồ histogram cho từng biến ? Lưu ý về việc lựa số bin phù hợp. Viết từ 2 đến 3 câu tóm tắt ý một số phương diện thú vị của dữ liệu khi quan sát histogram

Kết quả:

A picture containing text

Description automatically generated

Hình Thực hiện Task 3 câu 4 (phần code)

Chart, histogram

Description automatically generated

Hình Thực hiện Task 3 câu 4 (kết quả)

Nhận xét:

* Các biến MPG, Displacement, Horsepower, Weight có cùng chiều hướng lệch sang trái và phần sườn bên phải khá thoải. Điều này có thể là 1 biểu hiện cho việc các biến có mối liên quan với nhau (cần phân tích thêm). - Cylinder: Số lượng xe 4-cylinder gần như gấp đôi số lượng 6-cylinder và 8-cylinder, biểu đồ này cũng lệch trái (có thể quan sát về mối quan hệ với các biến ở trên cùng lệch trái)
* Model: số lượng các dòng xe được sản xuất qua các năm có sự thay đổi, nhưng nhìn chung không quá khác biệt
* Origin: đây là 1 biến đặc biệt (đánh dấu khu vực xuất xứ của các xe bằng số), do chưa xác định được cách phân loại nên khó phân tích thêm
  1. Câu hỏi 5

Câu hỏi:

* 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?
* Tạm dịch: Vẽ biểu đồ Scatter giữ biến weight và biến MPG và rút ra kết luận. Cho biết thêm về hệ số tương quan giữa 2 biến

Kết quả:

Chart, scatter chart

Description automatically generated

Hình Thực hiện Task 3 câu 5 (phần code)

Chart, scatter chart

Description automatically generated

Hình Thực hiện Task 3 câu 5 (kết quả) (1)

Chart, scatter chart

Description automatically generated

Hình Thực hiện Task 3 câu 5 (kết quả) (2)

Nhận xét:

* Weight và MPG có chiều thay đổi ngược nhau: MPG giảm dần khi Weight tăng dần
* Độ tương quan giữa 2 biến tương đối cao (giá trị tuyệt đối xấp xỉ 0.7 hoặc hơn), có nghĩa là phần lớn sự thay đổi của MPG có thể giải thích bởi sự thay đổi của Weight và ngược lại
* Tuy nhiên giá trị MPG tăng/giảm không đều khi Weight thay đổi: khi Weight càng giảm về 0 thì tốc độ tăng của MPG càng nhanh, do đó ta có thể xem xét đến việc dùng một mô hình dự đoán (ví dụ polynomial) thay cho đường thẳng, và xem xét sự ảnh hưởng của các biến khác tới biến Weight
  1. Câu hỏi 6

Câu hỏi:

* Plot a scatterplot of year vs. cylinders attributes. Add a small random noise to the values to make the scatterplot look nicer. What can you conclude? Do some internet search about the history of car industry during 70’s that might explain the esults. (Hint: data.mpg + np.random.random(len(data.mpg)) will add small random noise)
* Tạm dịch: Vẽ biểu đồ Scatter giữa biến năm và biến số lượng xi lanh. Thêm một số nhiễu để biểu đồ trong đẹp hơn và rút kết luận. Khảo sát về lịch sử công nghiệp ô tô vào thập niên 1970 để hiểu thêm về dữ liệu

Kết quả:

A picture containing graphical user interface

Description automatically generated

Hình Thực hiện Task 3 câu 6

Nhận xét:

Trong giai đoạn 1970 - 1982

* Có các loại động cơ: 3 / 4 / 5 / 6 / 8-cylinder (xi-lanh) đã được sản xuất

Các nhóm động cơ chiếm thiểu số

* 3-cylinder: xuất hiện rải rác qua các năm với số lượng không nhiều
* 5-cylinder: xuất hiện ở cuối thập niên 70 và biến mất

Các nhóm động cơ chiếm đa số: 4 / 6 / 8-cylinder

* 4-cylinder: số lượng dòng xe 4-cylinder ở đầu giai đoạn tuy không nhiều bằng 8-cylinder nhưng tăng trưởng dần qua từng năm và chiếm ưu thế tuyệt đối ở cuối giai đoạn 1982
* 6-cylinder: ở giai đoạn đầu số lượng dòng xe không nhiều (riêng năm 1972 không có giá trị nào); ở giai đoạn giữa, 6-cylinder đã vượt các dòng xe 8-cylinder và xấp xỉ bằng số lượng xe của 4-cylinder; tuy nhiên đến những năm từ 1980 trở đi, 6-cylinder có sự suy giảm và yếu thế hơn nhiều so với 4-cylinder.
* 8-cylinder: chiếm ưu thế ở giai đoạn đầu, các dòng xe có 8-cylinder chiếm đa số xe ra đời trong năm 1970 và tới năm 1973 vẫn chiếm thị phần đáng kể; từ năm 1974 đến 1979, các dòng xe 8-cylinder giảm sút so với 4-cylinder và chỉ tương đương với 6-cylinder; từ năm 1980 trở đi, 8-cylinder hầu như không có dòng xe nào nữa và hoàn toàn biến mất trong thị phần.

Nhận xét dựa trên lịch sử:

Việc sử dụng các động cơ V8 và V6 giúp xe mạnh mẽ hơn, sản xuất nhiều mã lực hơn. Trong giai đoạn 1960-1970 là cuộc đua giữa các nhà sản xuất xe để sản xuất xe nào có công suất mạnh hơn cho tốc độ cao hơn tiêu biểu là cuộc đua giữ Ferrari và Ford. Nơi chứng kiến sự đối đầu của các thiết kế xe mới nhanh hơn và mạnh hơn. Dần dần, các hãng xe khác cũng tham gia cuộc đua này tạo thị trường xe v8 đa dạng và đạt nhiều số lượng. Các xe 4 xi lanh được cho là dành cho các xe nhỏ, tiết kiệm năng lượng và có giá thành rẻ. Tuy nhiên 1 biến cố đã xảy ra vào năm 1979 khi khủng hoảng dầu mỏ xảy ra. Các xe v8 v6 sử dụng năng lương rất nhiều nên số lượng sản xuất được chững lại nhường chỗ cho dòng xe V4 giá thành rẻ sử dụng ít năng lượng đáp ứng hoàn cảnh giá xăng dầu tăng cao vào lúc đó + với tình trạng thất nghiệp mạnh diễn ra

* 1. Câu hỏi 7

Câu hỏi:

* Show 2 more scatterplots that are interesting do you. Discuss what you see.
* Tạm dịch: Vẽ thêm 2 biểu đồ Scatter mà bạn thích và đưa nhận xét

Kết quả

Chart, scatter chart

Description automatically generated

Hình Scatter (1) giữa biến displacement và horsepower Task 3 câu 7

Nhận xét:

* Mối quan hệ giữa Displacement (dung tích xi-lanh) và Horsepower(mã lực) là gần như tuyến tính
* Displacement và Horsepower có cùng chiều thay đổi: Displacement tăng thì Horsepower tăng và ngược lại
* Mối tương quan giữa 2 biến tương đối cao (>70%)

Chart, scatter chart

Description automatically generated

Hình Scatter (2) giữa biến xi lanh và khối lượng Task 3 câu 7

Nhận xét:

* Nhìn chung khi tăng số cylinders sẽ làm tăng khối lượng xe.
* Tuy nhiên khối lượng xe của mỗi dòng xe(phân theo số cylinder) có sự phân tán. Do đó có thể (và thực tế) là biến Weight sẽ chịu ảnh hưởng bởi nhiều biến khác ngoài cylinder
  1. Câu hỏi 8

Câu hỏi:

* Plot a time series for all the companies that show how many new cars they introduces during each year. Do you see some interesting trends? (Hint: data.car name.str.split()[0] returns a vector of the first word of car name column.)
* Tạm dịch: Vẽ biểu đồ thời gian cho mỗi nhà sản xuất thể hiện số loại xe mới mà họ ra mắt mỗi năm. Đưa ra nhận xét về xu hướng phát triển

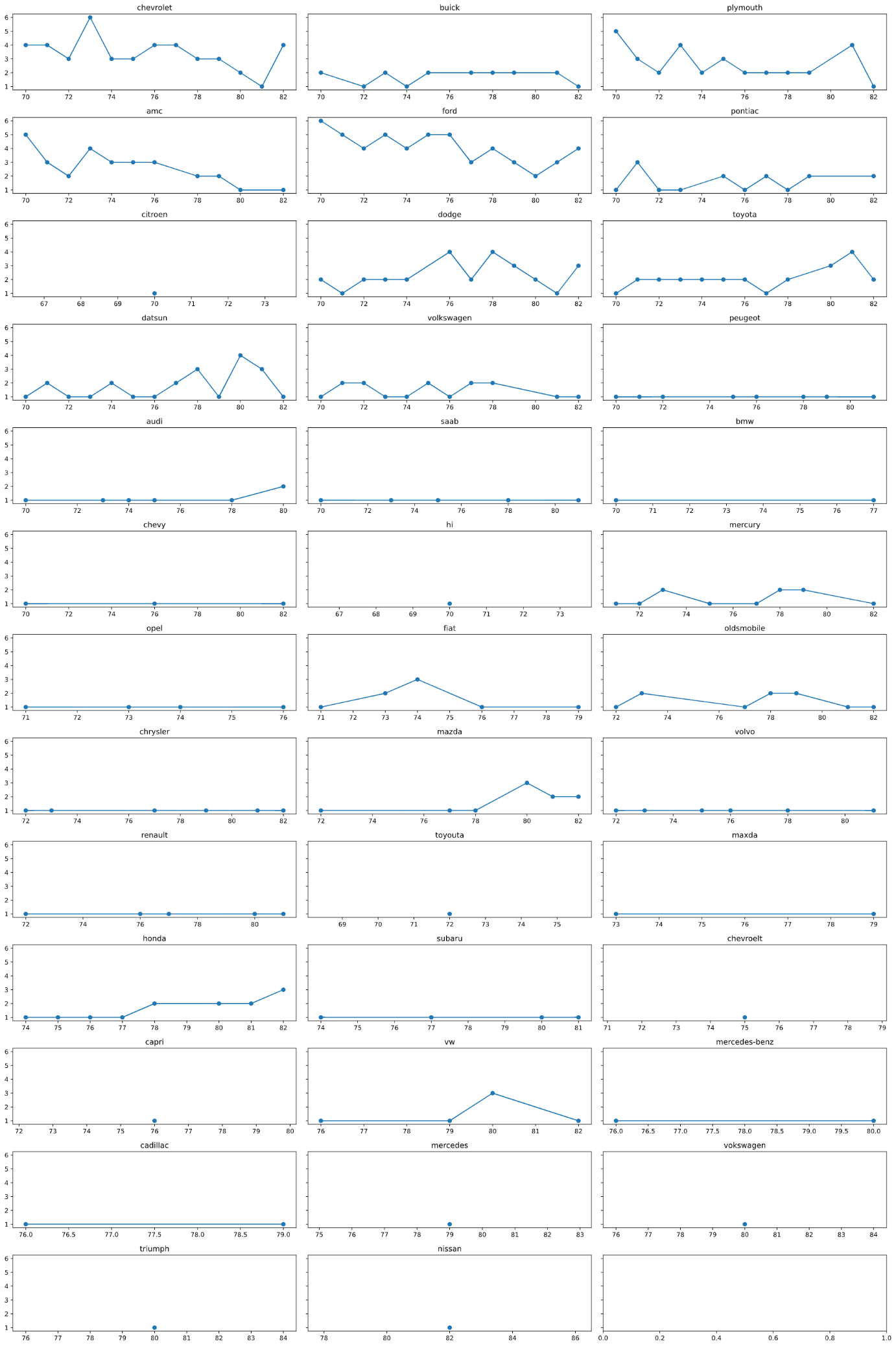
Kết quả:

(xem thêm ở file đính kèm)

Graphical user interface, text, application

Description automatically generated

Hình Thực hiện Task 3 câu 8 (phần code)



Hình Biểu đồ số mẫu xe ra hàng năm của từng hãng trong khoảng thời gian 1970-1982

Chart, line chart

Description automatically generated

Hình Biểu đồ tổng mẫu xe ra mắt ở các năm 1970 đến 1982

Nhận xét:

* Ta có thể thấy số lượng mỗi dòng xe mà mỗi hãng sản xuất biến động liên tục theo thời gian
* Số lượng dòng xe ra mắt nhiều nhất vào năm 1973 và rơi xuống đáy ngay sau đó - năm 1974
* Xét trên tổng thể (tất cả các hãng), chu kì thay đổi số lượng dòng xe được ra mắt rơi vào khoảng từ 2-3 năm. Số lượng dòng xe ra mắt mỗi năm không thể duy trì ổn định mà sẽ đạt đỉnh và rơi xuống đáy (cực tiểu địa phương) trong vòng 1-2 năm tiếp theo
  1. Câu hỏi 9

Câu hỏi:

* Calculate the pairwise correlation, and draw the heatmap with Matplotlib. Do you see some interesting correlation? (Hint: data.iloc[:,0:8].corr(), plt.pcolor() draws the heatmap.)
* Tạm dịch: tính hệ tương qua và vẽ heatmap bằng thư viện Matplotlib và rút nhận xét

Kết quả:

Chart, treemap chart

Description automatically generated

Hình Thực hiện Task 3 câu 9 (phần code)

Graphical user interface, application

Description automatically generated

Hình Thực hiện Task 3 câu 9 (phần kết quả)

Nhận xét: Có thể chia bảng giá trị tương quan này thành 3 khu vực lớn

* Giữa các biến (MPG, cylinders, displacement, horsepower, weight) với nhau có độ tương quan rất cao (corr>0.7). Do đó tìm hiểu mối quan hệ của các biến này cần xem xét đến sự thay đổi của các biến còn lại.
* Mối quan hệ giữa (acceleration, model, origin) và (MPG, cylinders, displacement, horsepower, weight) phần lớn ở mức trung bình (0.3 ~ 0.7). Do đó sự thay đổi của 1 biến này vẫn có thể chịu ảnh hưởng của các biến khác, nhưng mức ảnh hưởng không cao.
* Mối quan hệ giữa các biến (acceleration, model, origin) với nhau là khá thấp và không đáng kể.

1. Task 04: Electric power consumption data
   1. Thực hiện vẽ lại plot1
      1. Các bước tiến hành
2. Import các thư viện cần thiết:



Hình Include các thư viên để tái tạo plot1

1. Thực hiện xử lý dữ liệu trước khi vẽ gồm các quá trình sau:

* Thực hiện xoá các dòng dữ liệu sai

Logo

Description automatically generated with low confidence

Hình Plot 1. Xoá các dòng dữ liệu sai

* Ghép cột thời gian và cột ngày tháng lại

Text

Description automatically generated with medium confidence

Hình Plot 1. Ghép cột thời gian và cột ngày tháng

* Lọc data từ ngày hợp lệ



Hình Plot 1. Thực hiện lọc data theo ngày hợp lệ

* Thực hiện chuyển đổi loại dữ liệu của cột thành datetime

A picture containing text

Description automatically generated

Hình Plot 1. Chuyển dữ liệu cột thành dạng datetime sử dụng hàm của pandas

* Thực hiện chuyển các cột dữ liệu ? sang thành 0

Graphical user interface, text, application

Description automatically generated

Hình Plot 1. Thay thế dữ liệu ? thành 0

* Kết quả sau khi xử lý

Table

Description automatically generated

Hình Plot 1. Kết quả sau khi xử lý dữ liệu

1. Tiến hành trực quan hoá dữ liệu

Graphical user interface, text, application, email

Description automatically generated

Hình Plot 1. Đoạn code để vẽ biểu đồ

Kết quả tái tạo plot1:

Chart, histogram

Description automatically generated

Hình Plot 1. Kết quả tái tạo

* + 1. Story rút ra từ biểu đồ

Biểu đồ trên sử dụng loại biểu đồ phân bố tần số (Histograms), dùng để đo lường tần số xuất hiện của một vấn đề nào đó. Qua đó, cho ta thấy rõ sự thay đổi cũng như biến động của tập dữ liệu.

Trong biểu đồ trên, ta có thể thấy:

* Trục hoành mang đơn vị đo của Công suất hoạt động Trung bình theo phút Toàn cầu của Hộ gia đình (tính theo kW).
* Trục tung mang đơn vị đo Tần số (hay số lần xuất hiện của dòng dữ liệu trong tệp dữ liệu).
* Chiều cao của cột thể hiện chi tiết tần số tương ứng với mỗi kW điện sử dụng.
* Có thể thấy biểu đồ thuộc dạng phân bố lệch trái, có 2 đỉnh phân bố độc lập. Lượng điện tiêu thụ nhiều nhất là tầm 4,3 kW.
* Trung bình lượng điện tiêu thụ trong khoảng thời gian khảo sát nằm trong khoảng gần 0,4 kW.
* Ở đỉnh nhỏ thì có gần 500 mốc thời gian là xài lượng điện khoảng 1,4 kW. Giá trị ở đỉnh nhỏ cũng có thể là giá trị bất thường của bộ dữ liệu.
  1. Thực hiện vẽ lại plot 2
     1. Các bước thực hiện

1. Thực hiện các bước xử lý dữ liệu đã làm với plot 1 ( xem lại các bước xử lý ở plot 1)
2. Tiến hành trực quan hoá dữ liệu

Graphical user interface, text, application

Description automatically generated

Hình Plot 2. Đoạn code vẽ plot 2

1. Kết quả

A picture containing diagram

Description automatically generated

Hình Plot 2. Kết quả tái tạo

* + 1. Story rút ra từ biểu đồ

Biểu đồ trên sử dụng loại biểu đồ đường, dùng để vẽ sự thay đổi của các đại lượng khi mà đơn vị thời gian nhiều và tương đối liên tục. Từ đó ta có thể thấy tốc độ tăng trưởng của đơn vị đại lượng đó.

Trong biểu đồ trên, ta có thể thấy:

* Trục hoành mang đơn vị đo *Thời gian trong khoảng giai đoạn 2 ngày* (từ ngày 01/02/2007 đến 02/02/2007, tính theo giờ).
* Trục tung mang đơn vị đo của *Công suất hoạt động Trung bình theo phút Toàn cầu của Hộ gia đình* (tính theo kW).
* Đường biểu diễn sự liên tục của năng lượng điện hoạt động trên toàn theo từng giờ.
* Hầu hết lượng điện tiêu thụ trung bình trong khoảng 0,4 kW. Vào một số khoảng thời gian gần cuối ngày thì lượng điện tăng vọt lên 2-3 kW.
* Lượng điện sử dụng cao nhất ghi nhận là vào khoảng 6 giờ ngày 2/2. Sau đó lượng điện biến động giảm dần ở mức 3 kW, vào đến vào cuối ngày thì lại tăng vọt lên mức 4 kW. Có thể thấy lượng điện vào ngày 2/2 có xu hướng biến động hơn là vào ngày 1/2.
  1. Thực hiện vẽ lại plot 3
     1. Các bước thực hiện

1. Import các thư viện cần thiết và import data vào dataframe

Graphical user interface, application

Description automatically generated with medium confidence

Hình Plot 3. Tiến hành import thư viện, data

1. Tiến hành xử lý dữ liệu tương tự ở bước 7.1 vẽ Plot 1

Table

Description automatically generated

Hình Plot 3. Kết quả sau khi xử lý dữ liệu

1. Tiến hành trực quan hoá

Text

Description automatically generated with medium confidence

Hình Plot 3. Đoạn code vẽ plot 3

1. Kết quả sau khi chạy

Chart, bar chart, histogram

Description automatically generated

Hình Plot 3. Kết quả tái tạo

* + 1. Story rút ra từ biểu đồ

Biểu đồ trên là loại biểu đồ đường, biểu đồ này được dùng để thể hiện sự thay đổi giá trị theo thời gian của các đại lượng đặc biệt Trong biểu đồ trên ta có thể thấy:

* Trục hoành mang đơn vị đo thời gian trong khoảng giai đoạn 2 ngày (từ ngày **01/02/2007 đến 02/02/2007**, tính theo giờ).
* Trục tung mang đơn vị đo lượng điện năng tiêu thụ theo Watt-h được đo bởi các đồng hồ điện
* Ba đường thẳng liên tục biểu diễn lượng điện năng tiêu thụ của hộ gia đình được tính hàng giờ
* Trong hai ngày này hộ gia đình dường như đã không ăn ở nhà khi lượng điện năng tiêu thụ trong phòng bếp cả hai ngày là bằng 0
* Vì tủ lạnh là thiết bị sử dụng điện chính và luôn phải hoạt động liên tục nên lượng điện năng tiêu thụ được đo ở phòng giặt đồ ở các mốc thời gian riêng biệt luôn có giá trị gần tương đương nhau
* Hộ gia đình đã sử dụng máy điều hòa và máy nước nóng ở các khung giờ nhất định trong ngày, họ thường sử dụng máy điều hòa vào khoảng thời gian 15h-24h trong ngày và máy nước nóng vào khoảng sáng sớm, Khi nhìn vào biểu đồ ta biết được lượng điện năng tiêu thụ của các thiết bị điện này là rất lớn và hộ gia đình sử dụng các thiết bị điện này ở một khoảng thời gian xác định trong ngày.
  1. Thực hiện vẽ lại plot 4
     1. Các bước thực hiện

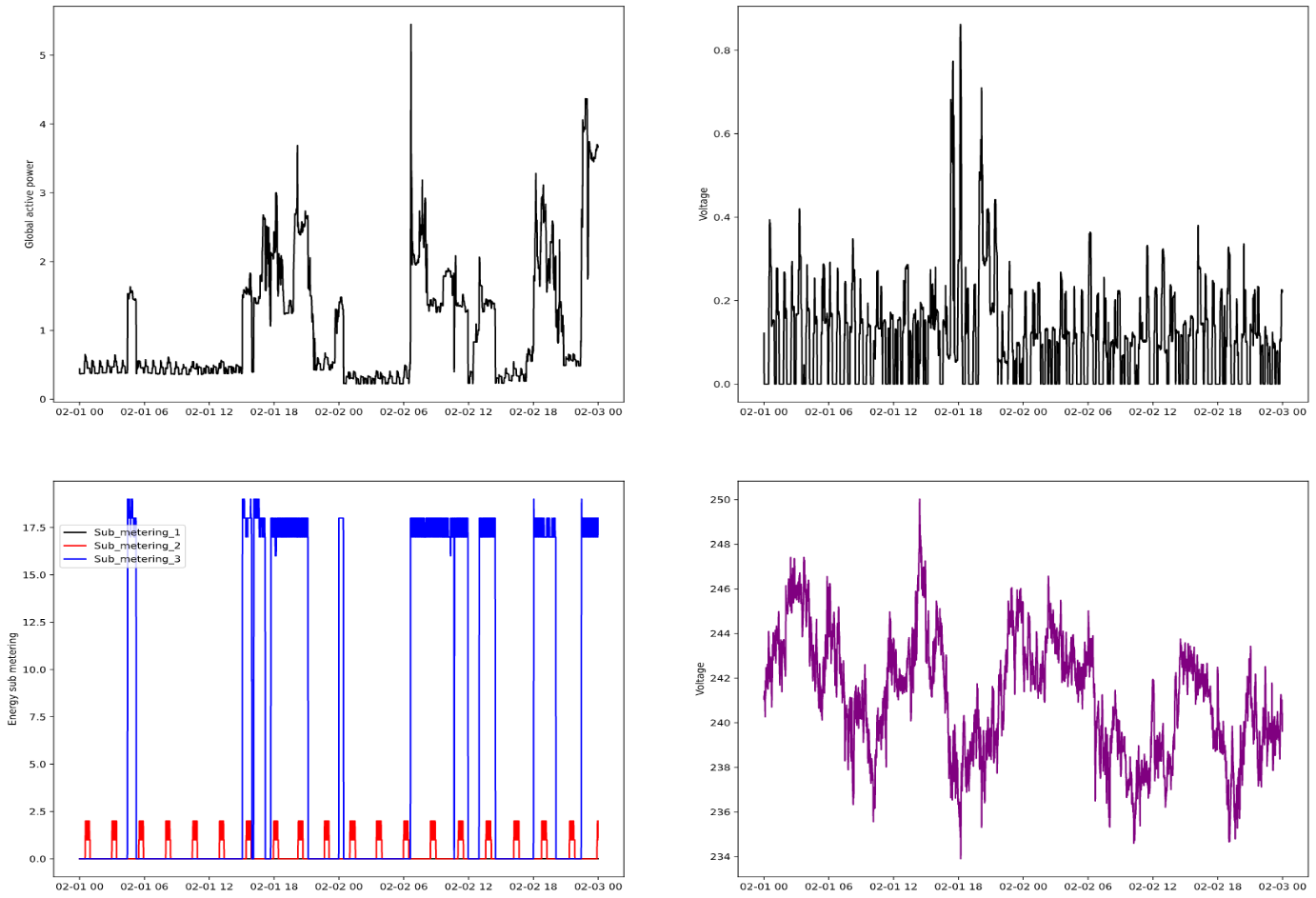
1. Thực hiện các bước xử lý giống với vẽ plot 3
2. Tiến hành vẽ plot

Text

Description automatically generated with low confidence

Hình Plot 4. Đoạn code vẽ hình

1. Kết quả sau khi thực hiện



Hình Plot 4. Kết quả tái tạo

* + 1. Story rút ra từ biểu đồ

Phân tích biểu đồ Biểu đồ trên được cấu thành từ 4 biểu đồ nhỏ, tất cả đều là biểu đồ đường và có trục hoành biểu diễn các khoảng thời gian đều và liên tục trong giai đoạn **01/02/2007 đến 02/02/2007** trong đó:

Biểu đồ **Công suất tiêu thụ** thể hiện công sức sử dụng điện của hộ gia đình

Biểu đồ **Công suất phản kháng** thể hiện công sức phản kháng của điện năng tiêu thụ

Biểu đồ **Lượng điện năng tiêu thụ** thể hiện lượng điện năng tiêu thụ của hộ gia đình

Biểu đồ **Mức điện áp** thể hiện mức điện áp của điện năng được hộ gia đình sử dụng

Câu chuyện Biểu đồ mức Công suất điện, công suất phản và Lượng điện năng tiêu thụ có mối quan hệ chặt chẽ, lượng điện năng tiêu thụ của hộ gia đình là cao nhất trong khoảng thời gian 18h-22h mỗi ngày chứng tỏ các thiết bị điện đã được sử dụng nhiều nhất trong khoảng thời gian này. Hộ gia đình đã sử dụng máy nước nóng, máy điều hòa và các thiết bị điện khác vào buổi tối

Máy điều hòa, đèn điện và tủ lạnh là các thiết bị điện được hộ gia đình sử dụng liên tục nên biểu đồ thể hiện mức công suất tiêu thụ, công suất phản kháng, lượng điện năng tiêu thụ và mức điện áp đều ở các khoảng thời gian nhất định trong ngày có xu hướng ít thay đổi

Máy nước nóng được hộ gia đình sử dụng có công suất tiêu thụ, công suất phản kháng và điện áp cao khi mỗi lần được sử dụng thì biểu đồ đều có xu hướng tăng mạnh. Trong khi đó tủ lạnh và máy điều hòa có lượng điện áp, công suất phản kháng ổn định

Ngoài các thiết bị điện được sử dụng liên tục trong ngày như đèn điện, tủ lạnh hộ gia đình có thói quen sử dụng các thiết bị điện nhất định trong các khoảng thời gian nhất định trong ngày. Họ sử dụng máy điều hòa vào buổi trưa và buổi tối, máy nước nóng vào buổi sáng và buổi chiều.

1. Tham khảo
2. **Determining Histogram Bin Width using the Freedman-Diaconis Rule**

<http://www.jtrive.com/determining-histogram-bin-width-using-the-freedman-diaconis-rule.html>

1. **Histogram for discrete values with matplotlib**

<https://stackoverflow.com/questions/30112420/histogram-for-discrete-values-with-matplotlib>

1. **Python draws a heat map**

<https://www.programmersought.com/article/27941083578/>

1. **Biểu đồ phân bố tần số (Histograms)**

<https://anhkiet1189.blogspot.com/2013/09/bieu-o-phan-bo-tan-so-histograms.html>