Modul 3 Tuần 1 - Advanced Data Visualization

Time-Series Team

Ngày 1 tháng 8 năm 2025

Buổi học thứ 6 (ngày 1/8/2025) được chia thành 4 phần chính nhằm giúp bạn hiểu được cách tìm đúng biểu đồ minh họa dựa trên 3 Case Study sử dụng Python.

- **Phần 1: Basic Data Visualization**
- **Phần 2 Case study: ETTH dataset**
- **Phần 3 Case study: Iris Dataset**
- **Phần 4 Case study: Student Performance**

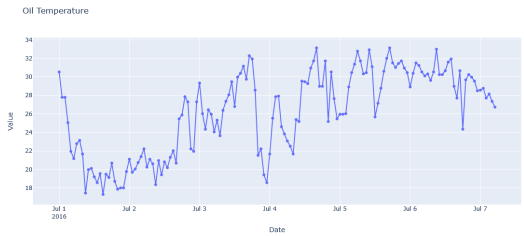
Phần 1: Basic Data Visualization

Phần này đưa ra khung toàn diện để hiểu và áp dụng các kỹ thuật trực quan hóa dữ liệu cơ bản bằng Python, được tổ chức xoay quanh ba case study và hướng dẫn chọn biểu đồ.

1.1 Giới thiệu các loại biểu đồ phổ biến

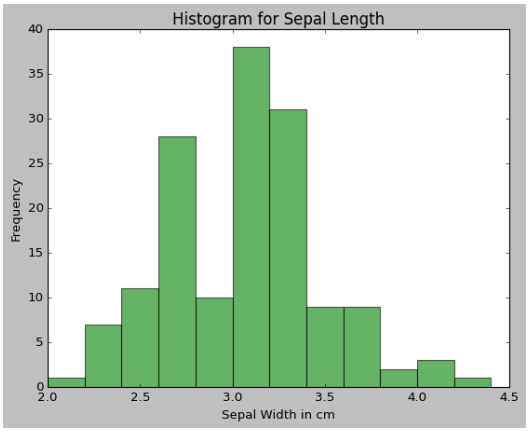
Trước khi vào phần code, đây là giải thích chi tiết hơn về từng loại biểu đồ bạn sẽ gặp — cách hoạt động, khi nào dùng, điểm mạnh/nhược, và ví dụ nhỏ để minh hoạ.

1. Biểu đồ xu hướng (Trend / Time-series) Line Chart: Line chart nối các điểm dữ liệu theo thứ tự (thường là thời gian). Dùng để nhận diện xu hướng chung, chu kỳ lặp, và các điểm đột biến (peaks/valleys). *Cách đọc:* nhìn dốc của đường để biết xu hướng (tăng/giảm), quan sát biên độ để thấy biến động. *Ưu điểm:* trực quan cho chuỗi thời gian, dễ so sánh nhiều series bằng nhiều đường. *Hạn chế:* với dữ liệu rời rạc không có thứ tự rõ ràng thì không phù hợp; nhiều đường quá sẽ gây rối. *Ví dụ:* nhiệt độ dầu trong ETTH theo giờ để thấy giờ đỉnh nhiệt.

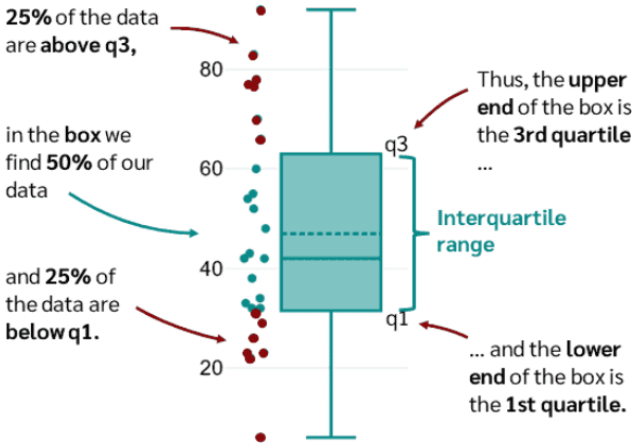


2. Biểu đồ phân phối dữ liệu (Distribution)

Histogram: Chia dải giá trị liên tục thành các **bin** (khoảng/nhóm giá trị), sau đó hiển thị tần suất trong mỗi bin. *Ghi chú chuyên ngành: bin* = một khoảng giá trị; số bin ảnh hưởng tới chi tiết/độ mượt của biểu đồ. *Ví dụ:* phân phối chiều dài cánh hoa trong Iris.



Box Plot (Hộp và râu): Dùng để tóm tắt và trực quan hóa phân phối dữ liệu thông qua các phân vị (quartiles) và giá trị ngoại lai (outliers). Cấu tạo như sau:



- **Râu Đỉnh (Upper Whisker):** Giá trị lớn nhất nhưng không phải ngoại lai.
- **Râu Đáy (Lower Whisker):** Giá trị nhỏ nhất nhưng không phải ngoại lai.
- **Upper Part - Q3 (Upper Quartile):** 25% dữ liệu lớn hơn giá trị này.
- **Median - Q2:** Giá trị trung vị, chia dữ liệu thành 2 nửa bằng nhau.
- **Lower Part - Q1 (Lower Quartile):** 25% dữ liệu nhỏ hơn giá trị này.
- **Outlier:** Các điểm bất thường, được định nghĩa là nhỏ hơn $Q_1 - 1.5 \times IQR$ hoặc lớn hơn $Q_3 + 1.5 \times IQR$.

IQR (Interquartile Range): Là khoảng giữa Q1 và Q3, công thức:

$$IQR = Q_3 - Q_1$$

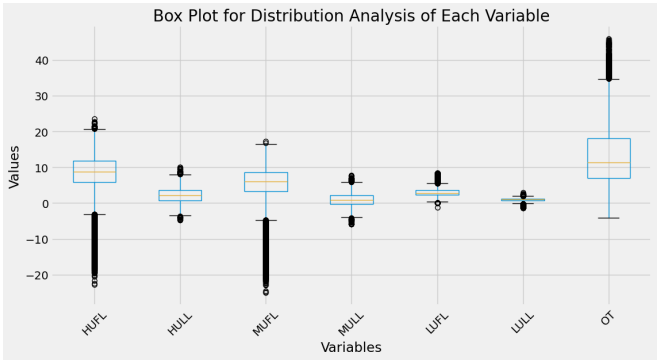
$\rightarrow$ thể hiện trực quan 50% dữ liệu tập trung ở giữa (middle 50%).

Ví dụ minh họa: Với dãy dữ liệu $[5, 7, 8, 12, 14, 18, 21, 23, 24, 26, 30]$:

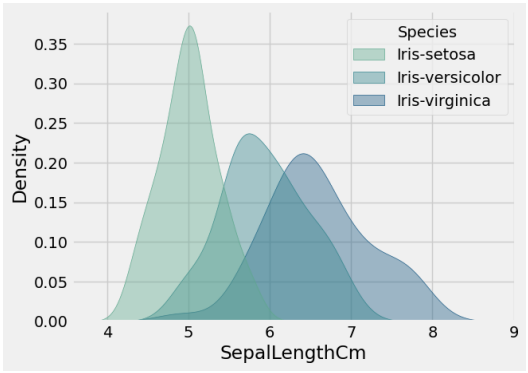
- Median (Q2) = 18
- Lower half = $[5, 7, 8, 12, 14] \Rightarrow$ Q1 = 8

- Upper half = $[21, 23, 24, 26, 30] \Rightarrow$ Q3 = 24
- $IQR = Q_3 - Q_1 = 24 - 8 = 16$
- **Upper Bound:** $Q_3 + 1.5 \times IQR = 24 + 24 = 48$
- **Lower Bound:** $Q_1 - 1.5 \times IQR = 8 - 24 = -16$

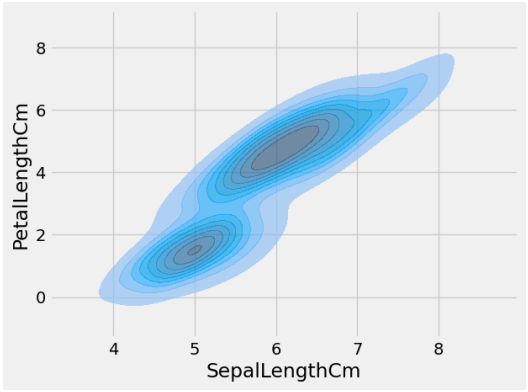
Vì toàn bộ dữ liệu đều nằm trong khoảng $[-16, 48]$, nên không có ngoại lai trong ví dụ này. Minh họa khác sử dụng Boxplot để kiểm tra giá trị ngoại lai và phân bổ giá trị của từng Lớp giá trị.



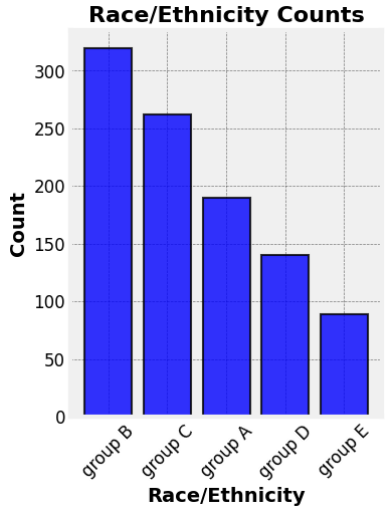
KDE Plot: KDE = *Kernel Density Estimate* (ước lượng mật độ bằng kernel), là một cách mượt hoá histogram để ước lượng mật độ xác suất liên tục. *Ghi chú chuyên ngành: bandwidth* (băng thông) điều chỉnh mức mượt — nhỏ quá gây nhiễu, lớn quá che đi chi tiết. *Ví dụ:* mật độ chiều dài cánh hoa theo loài.



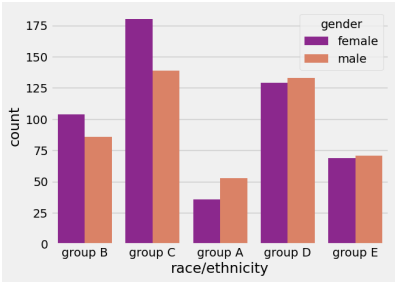
Distribution / DistPlot: Kết hợp histogram và KDE (và đôi khi *rug plot*) để đồng thời thấy tần suất và mật độ. *Ghi chú chuyên ngành: rug plot* = các dấu nhỏ trên trục x để chỉ từng quan sát; hữu ích để thấy mật độ thô.



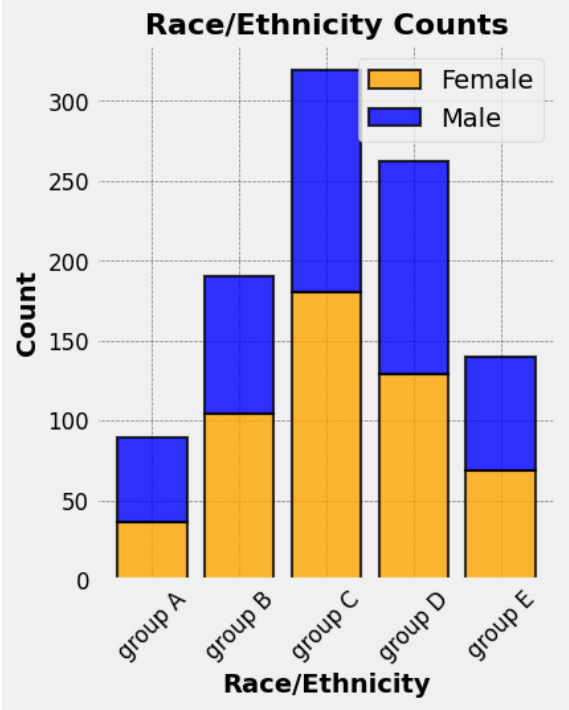
3. Biểu đồ so sánh nhóm (Comparison) Bar Chart: Mỗi cột biểu diễn giá trị (tổng, trung bình, tỷ lệ) cho một nhóm rời rạc. *Ghi chú chuyên ngành:* có thể thêm *error bars* (thanh sai số) để thể hiện độ tin cậy/độ biến thiên. *Ví dụ:* doanh số theo từng chi nhánh.



Count Plot: Là bar chart chuyên cho biến phân loại, hiển thị tần suất từng nhãn. Dùng để kiểm tra cân bằng lớp (class balance). *Ví dụ:* số sinh viên theo nhóm ôn luyện/không ôn luyện.

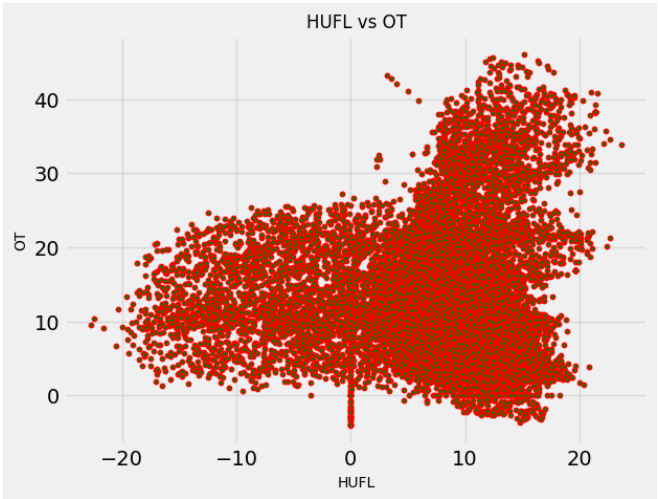


Stacked Bar Chart: Mỗi cột chia thành các phần con (các lớp nhỏ trong nhóm). *Ghi chú chuyên ngành: normalized stacked bar* = chuẩn hoá theo phần trăm để so sánh cấu trúc (tỉ lệ) giữa các nhóm. *Ví dụ:* doanh số online vs offline theo tháng.

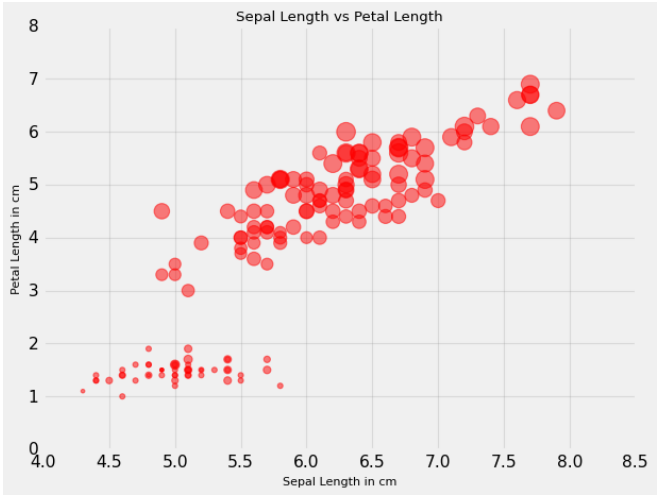


4. Biểu đồ quan hệ giữa các biến (Relationship)

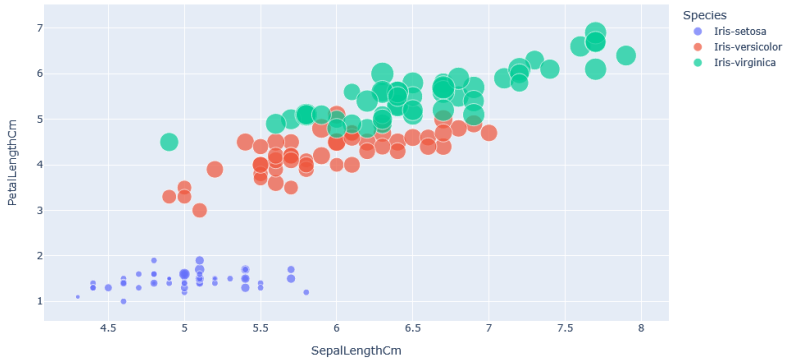
Scatter Plot: Mỗi điểm là một cặp (x, y), dùng để phát hiện xu hướng, cluster (cụm), và outlier. *Ghi chú chuyên ngành: overplotting* = hiện tượng chồng điểm khi có quá nhiều điểm; xử lý bằng *alpha/transparency* (độ trong suốt), *jitter* (dời nhẹ điểm) hoặc *hexbin* (đếm mật độ trong ô lục giác). *Trendline/regression* = đường ước lượng mối quan hệ (dùng để định lượng). *Ví dụ:* thời gian giao hàng vs tỷ lệ huỷ đơn.



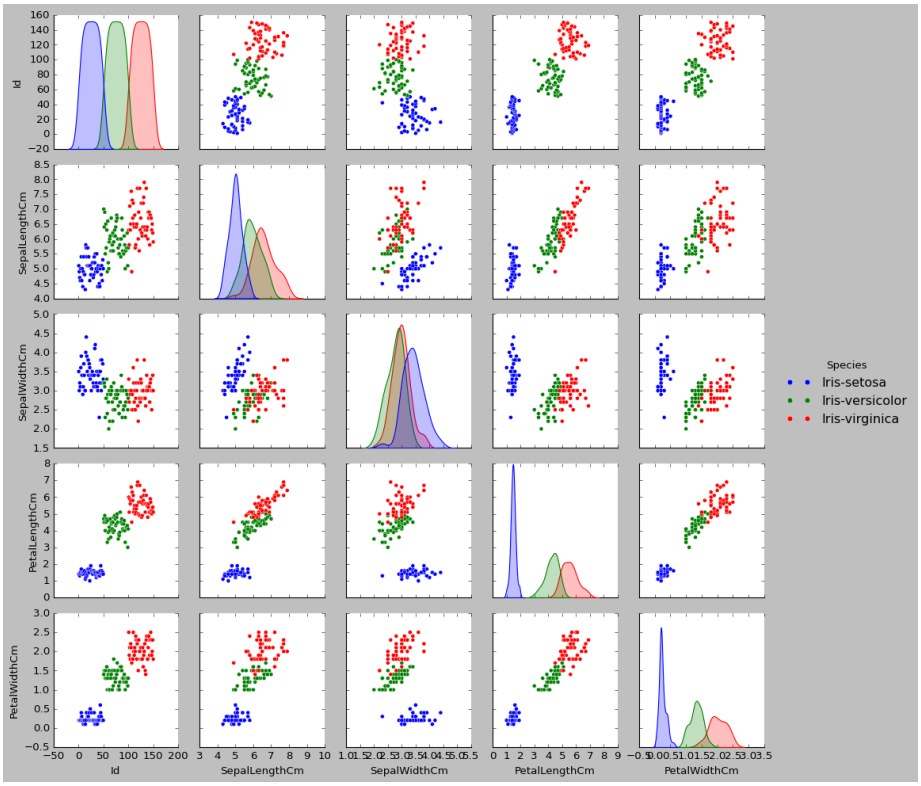
Bubble Chart: Scatter mở rộng thêm biến thứ ba qua kích thước bong bóng (size). *Ghi chú chuyên ngành:* cần *scale* (chuẩn hoá) kích thước để tránh hiểu nhầm về diện tích. *Ví dụ:* doanh thu (y), số đơn (x), thị phần (kích thước).

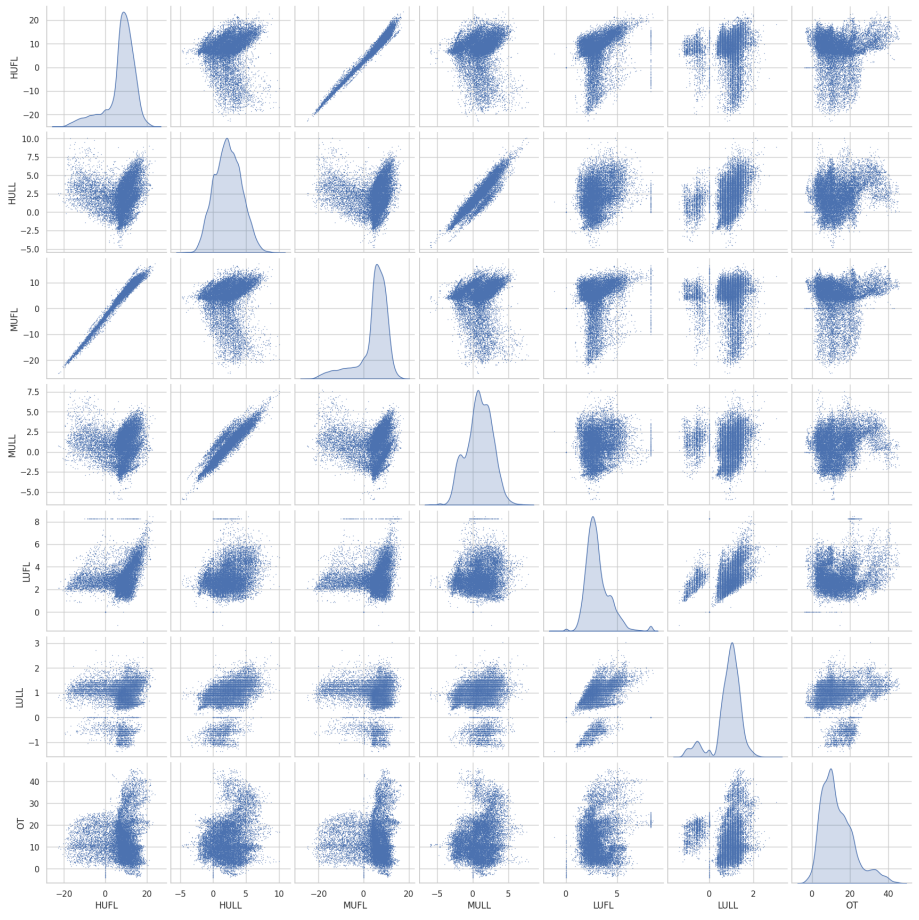


Interactive Scatter (Plotly): Scatter hỗ trợ tương tác (hover = di chuột xem chi tiết, zoom, lọc). Hữu ích khi khám phá dữ liệu lớn hoặc trình bày cho người dùng tương tác. *Ví dụ:* bản đồ pickup Uber với tooltip chi tiết.

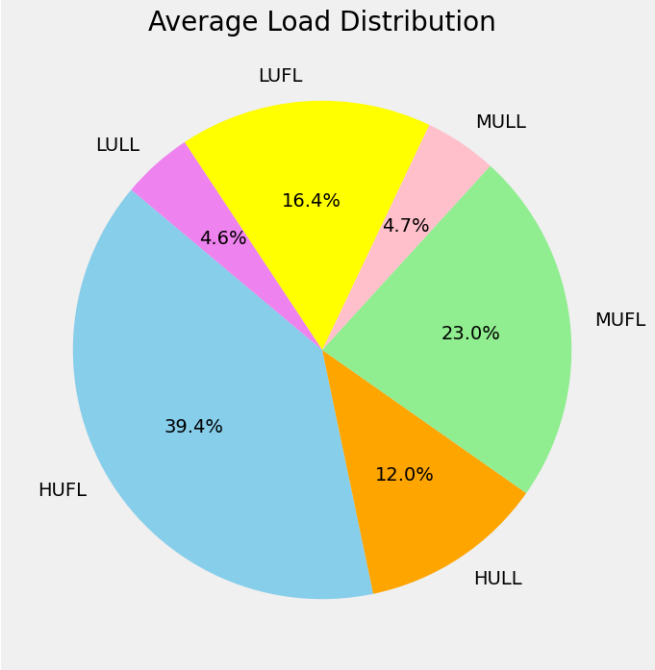


Pair Plot (Scatter matrix): Ma trận scatter cho mọi cặp biến, kèm biểu đồ phân phối (histogram/KDE) trên đường chéo; tiện để rà soát mối quan hệ cặp đôi và phân biệt nhóm theo *hue* (màu phân nhóm). *Ghi chú chuyên ngành: hue* = biến dùng để tô màu điểm theo nhóm. *Ví dụ:* 4 đặc trưng Iris để quan sát sự tách biệt giữa các loài.


Id
SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Species
Iris-setosa
Iris-versicolor
Iris-virginica

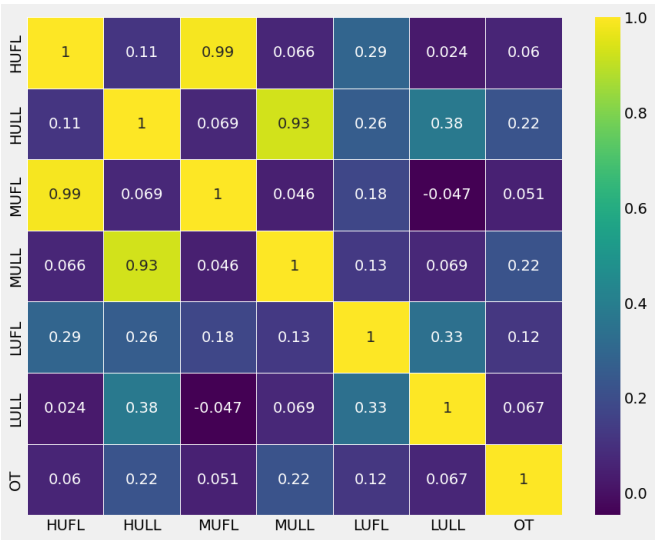

HUFL
HULL
MUFL
MULL
LUFL
LULL
OT

5. Biểu đồ thành phần - tổng thể (Part-to-Whole) Pie Chart: Hiển thị tỷ lệ phần trăm các phần trong tổng thể; dùng khi số phần nhỏ (thường <7). *Ghi chú chuyên ngành:* tránh lát mảnh quá nhỏ, luôn thêm nhãn phần trăm hoặc legend để người đọc dễ hiểu. *Ví dụ:* tỷ lệ tải HUFL/MUFL/LUFL.

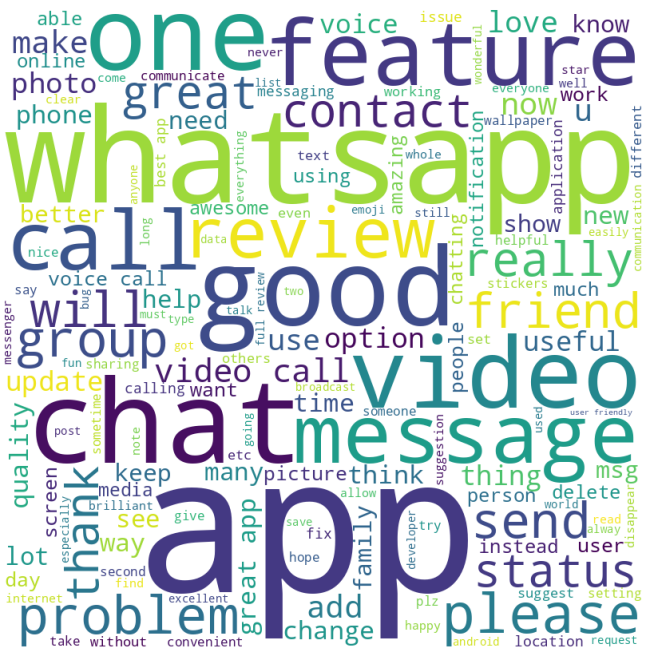


6. Biểu đồ đặc biệt

Heatmap: Ma trận màu (color matrix) hiển thị giá trị theo hai chiều — thường dùng cho ma trận tương quan hoặc bản đồ mật độ. Giúp nhanh nhận diện *hotspot* (vùng giá trị cao). *Ví dụ:* ma trận tương quan giữa các cảm biến ETTH.



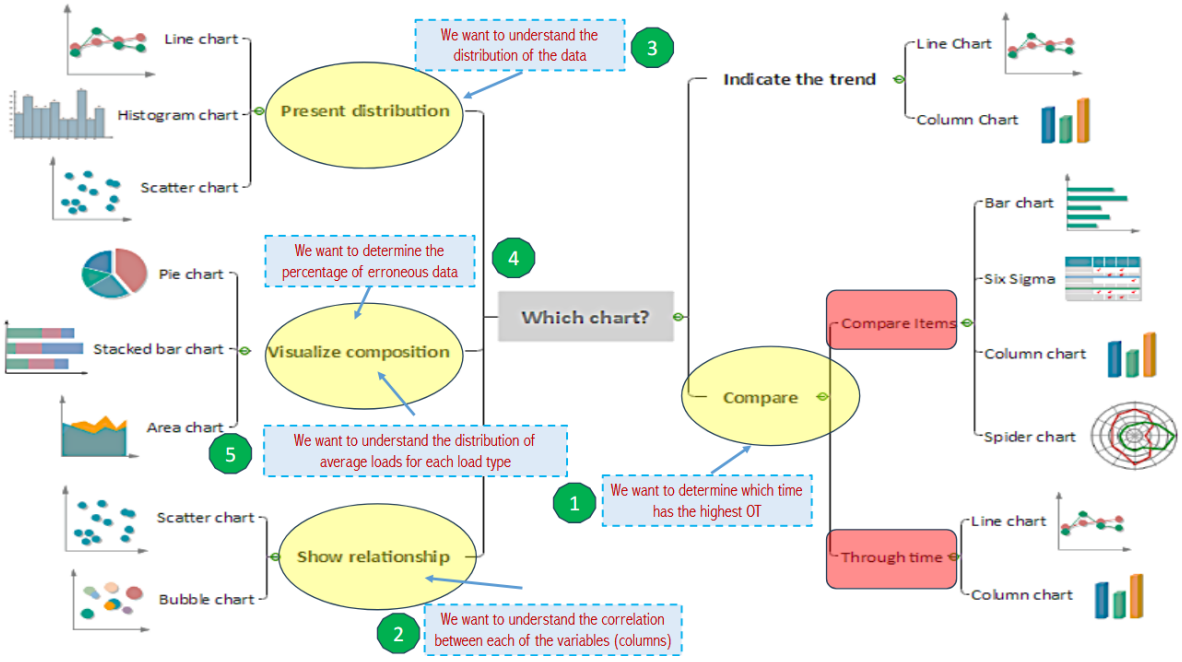
Word Cloud: Vẽ từ theo kích thước tương ứng tần suất hoặc trọng số; hữu ích cho khám phá nhanh các từ khóa nổi bật trong dữ liệu văn bản. *Ghi chú chuyên ngành:* trước khi tạo, cần loại bỏ *stopwords* (từ dừng — các từ phổ biến không mang ý nghĩa như ”và”, ”là”), và cân nhắc *stemming/lemmatize* (rút gốc từ) để nhóm các dạng từ tương tự. *Ví dụ:* từ khoá tìm kiếm nhiều nhất trên YouTube.



Để hiểu rõ hơn về code của từng biểu đồ, hãy truy cập: Code Visualization

1.2 Giới thiệu quy trình chọn biểu đồ

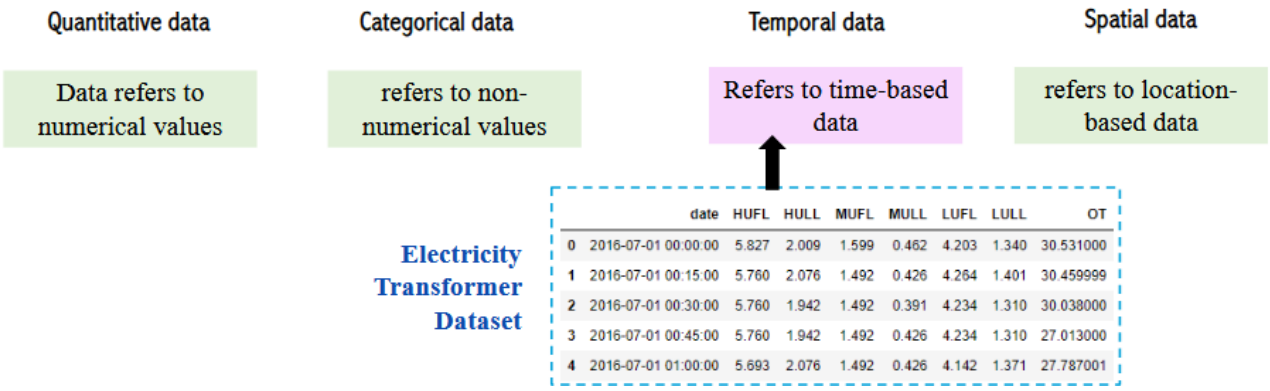
Khi bắt đầu trực quan hóa dữ liệu, có hai câu hỏi quan trọng cần đặt ra:
(1) Thông tin hay vấn đề nào cần minh họa?
(2) Loại biểu đồ nào thể hiện rõ nhất thông tin đó? Việc lựa chọn đúng phụ thuộc vào loại dữ liệu, mối quan hệ giữa các biến, mục tiêu trực quan hóa và đối tượng người xem. Không có một biểu đồ duy nhất phù hợp cho tất cả trường hợp, do đó đôi khi cần thử nghiệm nhiều loại khác nhau để tìm lựa chọn tối ưu.



Hình 1: Nguyên tắc chọn loại biểu đồ

Quy trình chọn biểu đồ có thể đi theo các bước sau:

- Bước 1: Xác định loại dữ liệu → dữ liệu định lượng (số đo, tính toán), dữ liệu phân loại (ví dụ Yes/No, Nam/Nữ), dữ liệu chuỗi thời gian (time series), hoặc dữ liệu không gian (gắn với vị trí địa lý).



	date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
0	2016-07-01 00:00:00	5.827	2.009	1.599	0.462	4.203	1.340	30.531000
1	2016-07-01 00:15:00	5.760	2.076	1.492	0.426	4.264	1.401	30.459999
2	2016-07-01 00:30:00	5.760	1.942	1.492	0.391	4.234	1.310	30.038000
3	2016-07-01 00:45:00	5.760	1.942	1.492	0.426	4.234	1.310	27.013000
4	2016-07-01 01:00:00	5.693	2.076	1.492	0.426	4.142	1.371	27.787001

- Bước 2: Xác định thông tin muốn truyền tải → nếu cần thể hiện xu hướng theo thời gian thì dùng line chart hoặc area chart (ví dụ: mức tiêu thụ điện trong một ngày). Nếu cần so sánh giữa các nhóm thì dùng bar chart hoặc column chart. Nếu muốn xem phân phối dữ liệu thì dùng histogram hoặc box plot.

	date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
0	2016-07-01 00:00:00	5.827	2.009	1.599	0.462	4.203	1.340	30.531000
1	2016-07-01 00:15:00	5.760	2.076	1.492	0.426	4.264	1.401	30.459999
2	2016-07-01 00:30:00	5.760	1.942	1.492	0.391	4.234	1.310	30.038000
3	2016-07-01 00:45:00	5.760	1.942	1.492	0.426	4.234	1.310	27.013000
4	2016-07-01 01:00:00	5.693	2.076	1.492	0.426	4.142	1.371	27.787001

Electricity Transformer Dataset

- A comparison chart is useful for showing the differences between two or more data points, such as a bar chart or a column chart.
- A distribution chart is useful for showing how data is spread out, such as a histogram or a box plot.
- A relationship chart is useful for showing how two or more variables are related, such as a scatter plot or a bubble chart.

- Bước 3: Xác định mục tiêu trực quan hóa → bạn muốn làm nổi bật xu hướng, phân phối, hay mối quan hệ giữa các biến?

	date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
0	2016-07-01 00:00:00	5.827	2.009	1.599	0.462	4.203	1.340	30.531000
1	2016-07-01 00:15:00	5.760	2.076	1.492	0.426	4.264	1.401	30.459999
2	2016-07-01 00:30:00	5.760	1.942	1.492	0.391	4.234	1.310	30.038000
3	2016-07-01 00:45:00	5.760	1.942	1.492	0.426	4.234	1.310	27.013000
4	2016-07-01 01:00:00	5.693	2.076	1.492	0.426	4.142	1.371	27.787001

Electricity Transformer Dataset

- if you want to show a trend over time, a line chart or an area chart might be more appropriate.
- If you want to compare data points, a bar chart or a column chart might be a better choice.
- If you want to show a distribution, a histogram or a box plot might be more useful.

- Bước 4: Xác định đối tượng người xem → nếu đối tượng là chuyên gia có thể dùng biểu đồ phức tạp (heatmap, scatter matrix), còn với công chúng thì nên chọn biểu đồ trực quan, dễ hiểu (bar chart, pie chart).

	date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
0	2016-07-01 00:00:00	5.827	2.009	1.599	0.462	4.203	1.340	30.531000
1	2016-07-01 00:15:00	5.760	2.076	1.492	0.426	4.264	1.401	30.459999
2	2016-07-01 00:30:00	5.760	1.942	1.492	0.391	4.234	1.310	30.038000
3	2016-07-01 00:45:00	5.760	1.942	1.492	0.426	4.234	1.310	27.013000
4	2016-07-01 01:00:00	5.693	2.076	1.492	0.426	4.142	1.371	27.787001

Electricity Transformer Dataset

- If your audience is expert, you might be able to use more complex charts, such as heat maps.
- If your audience is less familiar with data visualization, simpler charts like pie charts or bar charts might be more effective.

- Bước 5: Chọn loại biểu đồ phù hợp → thử nghiệm nhiều biểu đồ khác nhau và so sánh để chọn cách thể hiện hiệu quả nhất, vì không có giải pháp "one-size-fits-all".

Electricity Transformer Dataset

	date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
0	2016-07-01 00:00:00	5.827	2.009	1.599	0.462	4.203	1.340	30.531000
1	2016-07-01 00:15:00	5.760	2.076	1.492	0.426	4.264	1.401	30.459999
2	2016-07-01 00:30:00	5.760	1.942	1.492	0.391	4.234	1.310	30.038000
3	2016-07-01 00:45:00	5.760	1.942	1.492	0.426	4.234	1.310	27.013000
4	2016-07-01 01:00:00	5.693	2.076	1.492	0.426	4.142	1.371	27.787001

Phần 2: Case Study ETTH Dataset

Ví dụ: ETTH Dataset (Time Series)

Với bộ dữ liệu ETTH, mục tiêu là tìm ra thời điểm mức tiêu thụ điện cao nhất. Quy trình phân tích đi theo: xác định mục tiêu → tìm mối quan hệ giữa các biến (nhiệt độ dầu, tải máy biến áp) → kiểm tra phân phối dữ liệu để phát hiện bất thường → đánh giá tỷ lệ dữ liệu lỗi hoặc thiếu. Kết quả giúp lựa chọn biểu đồ line chart cho xu hướng, box plot để xem phân phối, và correlation chart để phân tích mối quan hệ giữa biến.

2.1 Chuẩn bị dữ liệu

Tải dữ liệu và tính các thống kê cơ bản (số lượng, trung bình, độ lệch chuẩn, phần trăm vị trí) để đánh giá chất lượng và phân phối dữ liệu.

2.2 Mục tiêu trực quan hóa

- Xác định thời điểm nhiệt độ dầu đạt đỉnh.

- Hiểu mối tương quan giữa các biến.
- Phân tích phân phối dữ liệu và phát hiện dữ liệu sai lệch.
- Đánh giá phân phối tải (HUFL, MUFL, LUFL).

2.3 Các loại biểu đồ sử dụng

- **Line chart:** Trực quan hóa xu hướng và chuỗi thời gian (ví dụ: nhiệt độ dầu, loại tải).
- **Box plot:** Thể hiện phân phối dữ liệu, ngoại lai và giá trị trung tâm.
- **Bar chart và Pie chart:** Biểu diễn dữ liệu phân loại và tỷ lệ giữa các loại tải.
- **Donut chart và Correlation chart:** Thể hiện quan hệ phần-toàn bộ và mối tương quan giữa các biến.

Phần 3: Case Study Iris Dataset

Bộ dữ liệu Iris – kinh điển trong học máy – cho thấy các kỹ thuật trực quan hóa nâng cao cho dữ liệu liên tục và đa biến.

3.1 Tổng quan dữ liệu

Chứa các thông số đo của hoa iris theo loài, cho phép phân tích mối quan hệ giữa các biến số.

3.2 Các trực quan sử dụng

- **Histogram:** Thể hiện phân phối tần suất và độ lệch (ví dụ: chiều dài cánh hoa, chiều rộng đài hoa).
- **Scatter plot:** Phân tích mối quan hệ cặp, chỉ ra kiểu tương quan.
- **Bubble chart:** Thêm biến thứ ba bằng kích thước bong bóng, phục vụ phân tích đa chiều.
- **KDE plot:** Minh họa mật độ xác suất cho phân phối đơn biến và nhị biến.
- **Displot chart:** Kết hợp histogram và KDE để có cái nhìn bổ sung về phân phối.
- **Biểu đồ 3D:** Thể hiện quan hệ giữa ba biến để phân tích phụ thuộc phức tạp.

Tóm tắt: Case study này làm nổi bật cách nhiều loại biểu đồ khác nhau giúp tăng khả năng diễn giải và khám phá insight.

3.3 Kỹ thuật trực quan hóa

- **Bar chart:** Thể hiện sự thay đổi rời rạc giữa các nhóm (ví dụ: giới tính, học vấn của phụ huynh, việc chuẩn bị ôn thi).
- **Count plot:** Hiển thị tần suất xuất hiện của biến phân loại để xác định nhóm chiếm ưu thế hoặc xu hướng.

Tóm tắt: Các trực quan hóa này giúp phát hiện yếu tố ảnh hưởng đến kết quả học tập, hỗ trợ các biện pháp giáo dục có mục tiêu.

Ôn tập: Cách chọn biểu đồ phù hợp

1. Xác định loại dữ liệu: định lượng, phân loại, chuỗi thời gian, hoặc không gian.
2. Xác định quan hệ giữa các biến: so sánh, phân phối, quan hệ.
3. Xác định mục tiêu trực quan hóa: xu hướng, so sánh, phân phối.
4. Xác định đối tượng: chuyên gia (heatmap) hay công chúng (pie, bar).
5. Thử nghiệm nhiều loại biểu đồ để chọn lựa phù hợp nhất.

Lời Kết

Dữ liệu số là cho máy còn hình ảnh trực quan từ dữ liệu là cho mình. Để hiểu cách máy hoạt động như thế nào, để nhận viết được mối quan hệ trong đời sống thực tế có thể là thời gian đi học và thời gian tắc đường hoặc hành vi mua hàng đến yếu tố làm giao động cổ phiếu. DÙ cho gì đi nữa, thì team TimeSeries cũng tin rằng chỉ với trực quan hóa, con người mới có thể thực sự hiểu những con số mà mình vận dụng.