**Lý thuyết Phân tích Đơn biến và Hai biến**

*- Phân tích đơn biến (univariate analysis) là gì? Nó khác gì với phân tích hai biến (bivariate analysis) trong khám phá dữ liệu?*

**1. Khái niệm cơ bản**

**1.1 Phân tích đơn biến (Univariate Analysis)**

**Định nghĩa**: là kỹ thuật phân tích dữ liệu đơn giản nhất, tập trung vào việc nghiên cứu và diễn giải một biến duy nhất để mô tả đặc điểm, tìm kiếm mô hình và đo lường xu hướng trung tâm, độ phân tán của dữ liệu mà không xem xét mối quan hệ với các biến khác. Các phương pháp phổ biến để thực hiện phân tích này bao gồm thống kê tóm tắt (như trung bình, trung vị), bảng phân phối tần suất và các loại biểu đồ (như biểu đồ cột, biểu đồ tần suất).

**Mục đích**:

* Hiểu rõ đặt điểm cơ bản của biến đó
* Phát hiện các xu hướng hoặc mẫu hình trong dữ liệu của biến đó

**Các kỹ thuật phổ biến**

* **Thống kê tóm tắt**:

Sử dụng các chỉ số như giá trị trung bình (mean), trung vị (median), giá trị lớn nhất (max), giá trị nhỏ nhất (min), độ lệch chuẩn (standard deviation) để tóm tắt dữ liệu.

* **Phân phối tần suất**:

Biểu diễn tần suất xuất hiện của các giá trị khác nhau trong dữ liệu, thường thông qua các bảng tần suất, biểu đồ tần suất (histogram), hoặc biểu đồ tròn (pie chart).

* **Biểu đồ**:

Sử dụng biểu đồ cột, biểu đồ thanh, hoặc biểu đồ tần suất để trực quan hóa dữ liệu, giúp dễ dàng nhận thấy xu hướng.

**1.2 Phân tích hai biến (Bivariate Analysis)**

**Định nghĩa**: Phân tích hai biến (hay phân tích song biến) là một phương pháp thống kê nhằm kiểm tra và xác định mối quan hệ giữa hai biến riêng biệt, cho phép nhà nghiên cứu hiểu được liệu chúng có liên quan với nhau hay không, mức độ mạnh yếu của mối liên hệ, và liệu một biến có thể được sử dụng để dự đoán biến kia hay không.

**Mục đích**:

* **Kiểm tra giả thuyết:**

Giúp xác định các giả thuyết đơn giản về mối liên hệ giữa hai yếu tố.

* **Xác định mối quan hệ:**

Tìm ra liệu hai biến có tương quan (liên quan) với nhau hay không.

* **Đánh giá cường độ và hướng mối quan hệ:**

Nếu có mối quan hệ, phân tích này cho biết mối quan hệ đó mạnh hay yếu, và nó là mối quan hệ thuận hay nghịch chiều.

- **Tóm lại:** Phân tích đơn biết mục đích để kiếm tra chính biến đó thay đổi như thế nào, Phân tích 2 biến nhằm mục đích kiểm tra mối quan hệ coi chúng có liên quan, ảnh hưởng với nhau không.

*+ Các thước đo thống kê nào thường được sử dụng trong phân tích đơn biến (ví dụ: trung bình, trung vị, mode, độ lệch chuẩn)?*

**2.1 Thống kê mô tả trung tâm**

* **Trung bình (Mean)**: μ = Σxi/n: Tổng tất cả các giá trị chia cho số lượng quan sát. Đây là thước đo phổ biến nhất, nhưng nhạy cảm với các giá trị ngoại lai (outliers).
* **Trung vị (Median)**: Giá trị nằm chính giữa của tập dữ liệu đã được sắp xếp. Đây là thước đo tốt hơn khi dữ liệu bị lệch (skewed) hoặc có các giá trị ngoại lai.
* **Mode**: Giá trị xuất hiện thường xuyên nhất trong tập dữ liệu. Thích hợp cho cả dữ liệu định danh (nominal) và định lượng (quantitative).

**2.2 Thống kê mô tả độ phân tán**

* **Phương sai (Variance)**: σ² = Σ(xi - μ)²/n: Trung bình của bình phương các độ lệch giữa mỗi giá trị và giá trị trung bình. Cho biết mức độ **phân tán** của dữ liệu. Phương sai càng lớn, dữ liệu càng rải rác.
* **Độ lệch chuẩn (Standard Deviation)**: σ = √σ². Cho biết **khoảng cách trung bình** mà mỗi điểm dữ liệu lệch khỏi giá trị trung bình. Độ lệch chuẩn lớn → dữ liệu rất biến động (rủi ro cao).
* **Khoảng tứ phân vị (IQR)**: Q3 - Q1. Cho biết mức độ phân tán của **50% dữ liệu ở giữa** (phần lõi). Giống như Trung vị, **ít bị ảnh hưởng bởi giá trị ngoại lai** hơn Phương sai và Độ lệch chuẩn. Nó thường được dùng để xác định các giá trị ngoại lai.
* **Min, Max, Range**

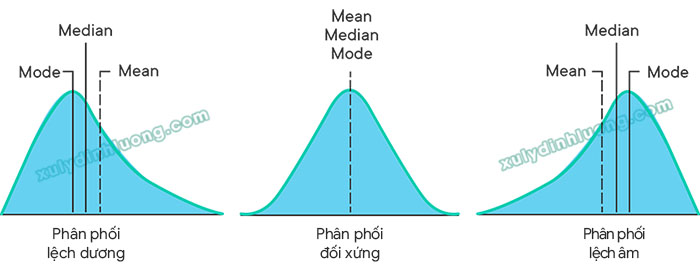
**2.3 Thống kê mô tả hình dạng phân phối**

* **Độ lệch (Skewness)**: Đo độ bất đối xứng của phân phối

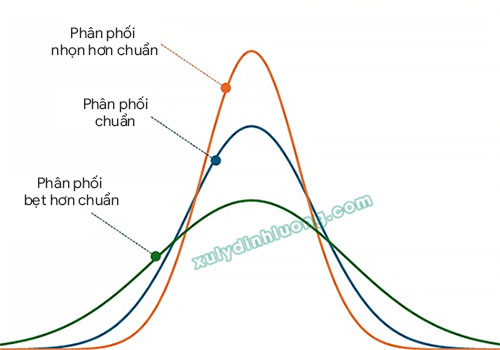
+ Skewness = 0: Phân phối đối xứng (Trung bình ≈ Trung vị).

+ Skewness > 0 (Positive Skew): Đuôi dài hơn kéo về phía bên phải. Phần lớn dữ liệu tập trung ở bên trái (các giá trị nhỏ).

+ Skewness < 0 (Negative Skew): Đuôi dài hơn kéo về phía bên trái. Phần lớn dữ liệu tập trung ở bên phải (các giá trị lớn).



* **Độ nhọn (Kurtosis)**: là một đại lượng thống kê đo lường hình dạng của phân phối dữ liệu, đặc biệt là độ nhọn của đỉnh và độ dày của đuôi so với phân phối chuẩn (phân phối có độ nhọn trung bình).



*+ Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến (ví dụ: tương quan, nhân quả)?*

Phân tích hai biến là quá trình nghiên cứu mối quan hệ giữa hai biến (hay còn gọi là hai tập dữ liệu) để xem chúng có liên quan, ảnh hưởng, hoặc đồng biến với nhau hay không.

* + **Tương quan (Correlation)**

**Định nghĩa**: **Tương quan** chỉ đo lường mức độ hai biến **thay đổi cùng nhau** (đồng biến) như thế nào. Nó không chỉ ra biến nào gây ra biến nào.

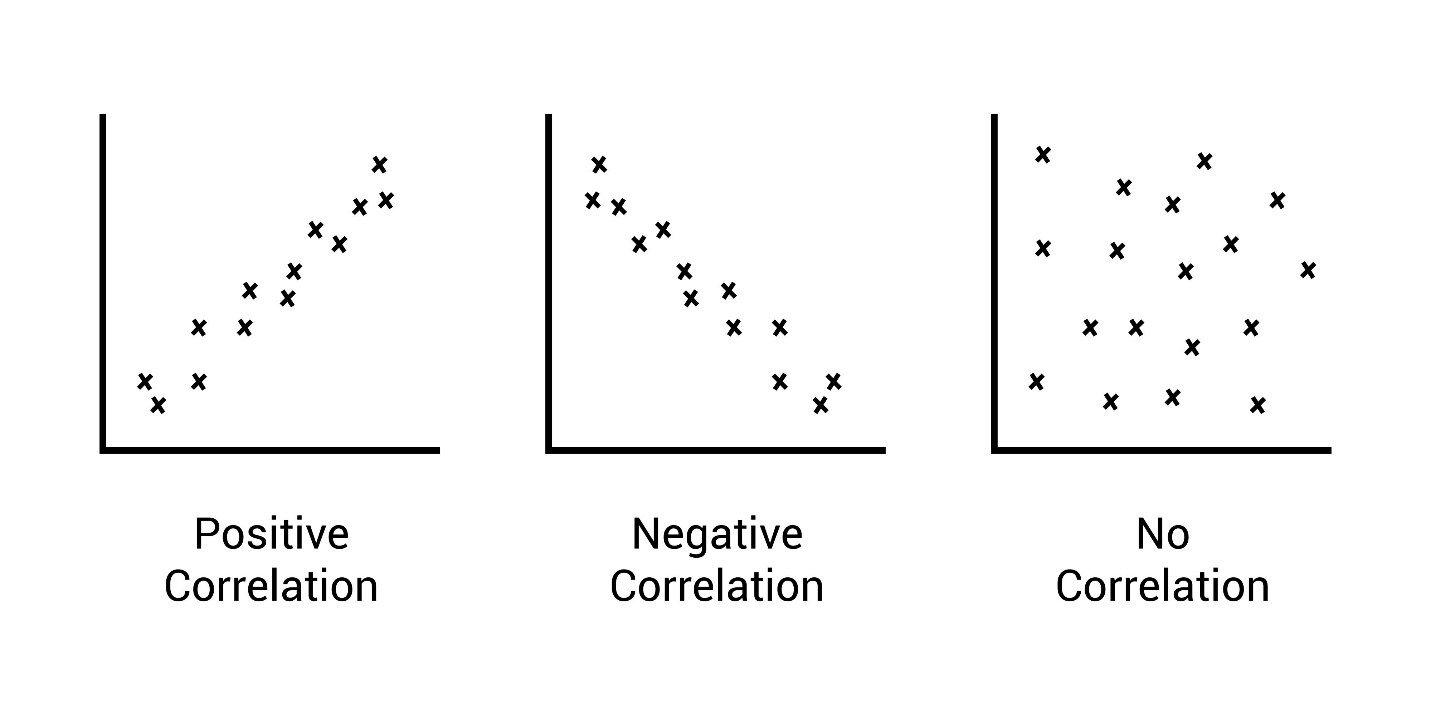
**Hệ số tương quan Pearson**:

r = Σ[(xi - x̄)(yi - ȳ)] / √[Σ(xi - x̄)² × Σ(yi - ȳ)²]

**Đặc điểm**:

* Giá trị từ -1 đến 1
* r = 1: Tương quan dương hoàn hảo. Hai biến tăng/giảm hoàn toàn cùng nhau. **Ví dụ:** Lượng mưa tăng → Mực nước sông tăng.
* r = -1: Tương quan âm hoàn hảo. Một biến tăng thì biến kia giảm. **Ví dụ:** Giá sản phẩm tăng → Lượng cầu giảm.
* r = 0: Không có tương quan tuyến tính

Note: Biểu đồ Phân tán (Scatter Plot) - Là công cụ trực quan hóa tốt nhất. Bạn vẽ một biến trên trục X và biến kia trên trục Y. Hình dạng của các điểm sẽ cho thấy mối quan hệ.



Lưu ý quan trọng: Tương quan không có nghĩa là Nhân quả!

* + **Mối quan hệ Nhân quả (Causation)**

**Nhân quả** chỉ ra rằng sự thay đổi ở **Biến Độc lập** (Nguyên nhân) trực tiếp gây ra sự thay đổi ở **Biến Phụ thuộc** (Kết quả).

**Làm thế nào để xác định Nhân quả?**

Trong Thống kê/Khoa học dữ liệu, để kết luận Nhân quả, cần phải:

1. **Có Tương quan:** Phải tồn tại mối quan hệ mạnh mẽ.
2. **Tính Thời gian:** Nguyên nhân phải xảy ra trước kết quả (Ví dụ: Thời gian học phải xảy ra trước Điểm thi).
3. **Cô lập:** Phải loại bỏ (hoặc kiểm soát bằng mô hình Hồi quy) ảnh hưởng của Biến Giả mạo (Ví dụ: Kiểm soát yếu tố "Chỉ số IQ của học sinh").
4. **Thử nghiệm:** Cách mạnh nhất là thông qua Thử nghiệm Ngẫu nhiên Có kiểm soát (A/B Testing), nơi bạn kiểm soát tất cả các yếu tố khác và chỉ thay đổi Biến Độc lập để xem kết quả.

*+ Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance) trong phân tích hai biến là gì?*

Hiệp biến cho biết **hướng** của mối quan hệ, còn Tương quan cho biết cả **hướng** và **độ mạnh** của mối quan hệ, đồng thời loại bỏ ảnh hưởng của đơn vị đo lường.

Cả hai đều được sử dụng trong phân tích hai biến để đo mức độ hai biến thay đổi cùng nhau.

**Hiệp biến (Covariance)**

**Định nghĩa**: Chỉ ra **hướng** của mối quan hệ tuyến tính giữa hai biến X và Y..

**Công thức**: Cov(X,Y) = Σ[(xi - x̄)(yi - ȳ)] / (n-1)

**Giá trị:**

* + **Dương (+):** Khi X tăng, Y cũng có xu hướng tăng (Mối quan hệ đồng biến).
  + **Âm (-):** Khi X tăng, Y có xu hướng giảm (Mối quan hệ nghịch biến). **Gần 0:** Không có mối quan hệ tuyến tính rõ ràng.

**Hạn chế:** Giá trị của Hiệp biến (ví dụ: 1000 hay 0.001) **không cho biết độ mạnh** của mối quan hệ, vì nó bị ảnh hưởng bởi thang đo và đơn vị.

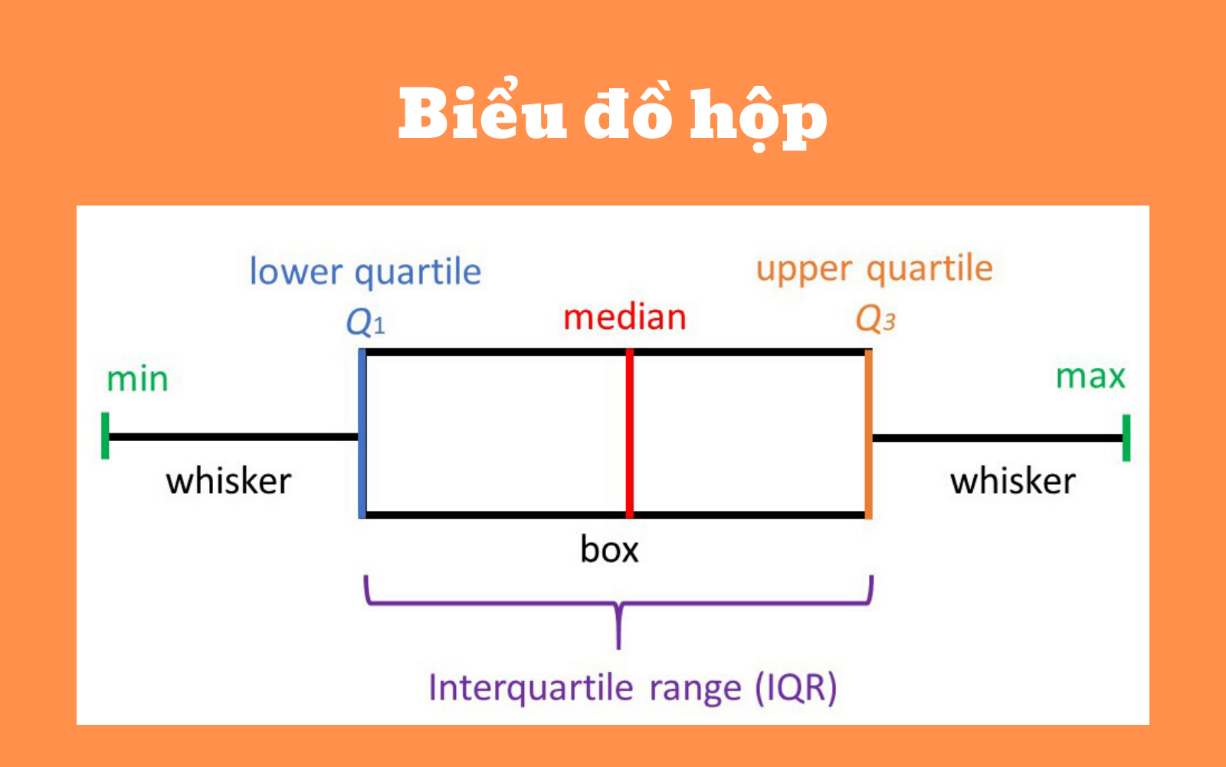
*+ Khi nào nên sử dụng biểu đồ trực quan hóa trong phân tích đơn biến so với phân tích hai biến?*

**Phân tích đơn biến:** Sử dụng khi muốn có một cái nhìn sơ bộ về nó như sự biến động, hình dạng, trung tâm, phân tán.

**Sử dụng khi**:

* Muốn hiểu phân phối của một biến
* Tìm kiếm outliers
* Kiểm tra tính chuẩn của dữ liệu
* So sánh các nhóm khác nhau của cùng một biến

**Các loại biểu đồ**:

* Histogram: Cho thấy cách dữ liệu được phân phối (ví dụ: phân phối chuẩn, lệch trái, lệch phải) và xác định các **yếu vị (Mode)**.
* Box plot: Tóm tắt nhanh **năm giá trị quan trọng** (min,Q1,Median,Q3,max) và xác định rõ ràng các **giá trị ngoại lai (Outliers)**.
* Bar chart: Dữ liệu phân loại
* Pie chart: Tỷ lệ các danh mục

-> Hiển thị tần suất của các nhóm hoặc danh mục rời rạc (ví dụ: số lượng người thuộc từng giới tính, thị phần sản phẩm).

* Biểu đồ Chuỗi thời gian (Time Series Plot): Nếu biến là dữ liệu thời gian, biểu đồ này giúp xác định xu hướng hoặc tính chu kỳ.

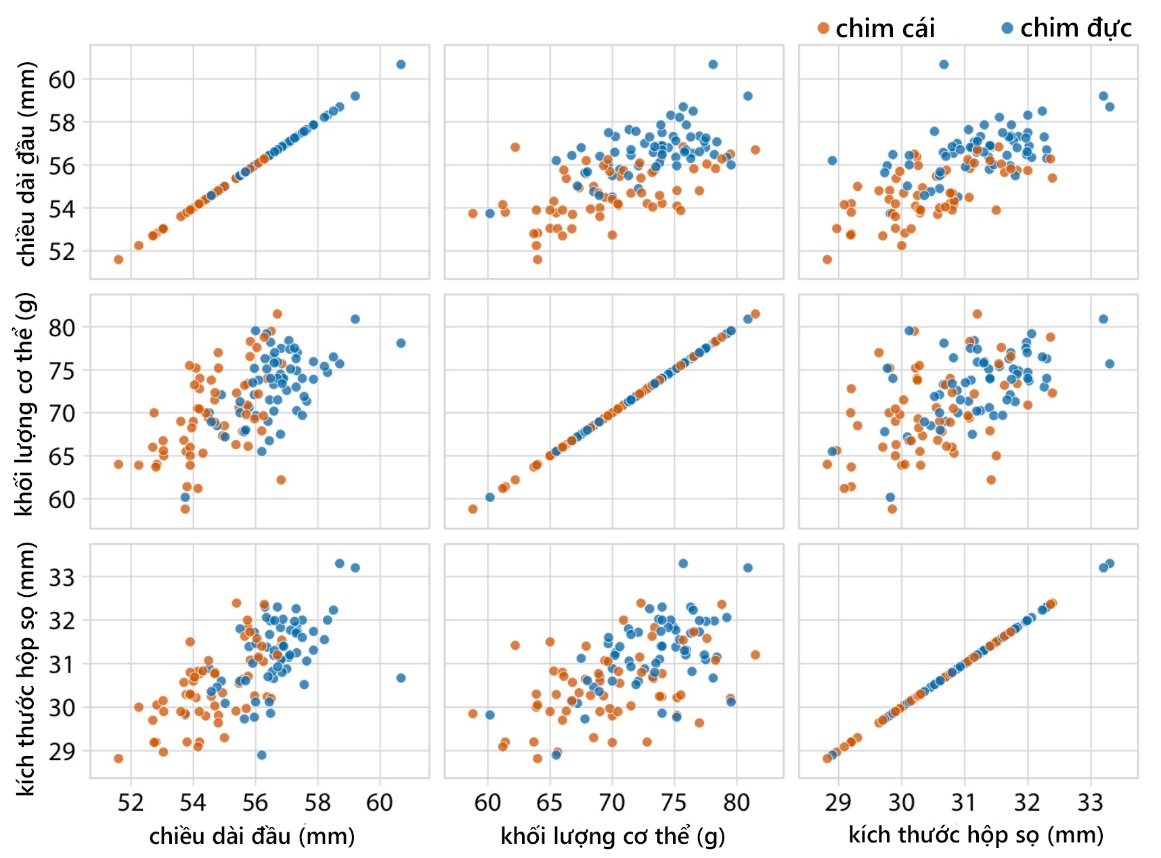
**Phân tích hai biến:** Sử dụng trực quan hóa trong phân tích hai biến khi bạn muốn hiểu mối quan hệ, tương quan, và cách một biến thay đổi khi biến kia thay đổi.

**Sử dụng khi**:

* Muốn khám phá mối quan hệ giữa hai biến
* So sánh các nhóm khác nhau
* Tìm kiếm patterns hoặc clusters
* Chuẩn bị cho modeling

**Các loại biểu đồ**:

* Scatter plot: Quan hệ giữa 2 biến liên tục



* Heatmap: Thường dùng cho các biến phân loại hoặc khi có quá nhiều điểm dữ liệu. Cho thấy sự **tập trung** của các cặp giá trị
* Grouped Box/Bar Plot: So sánh phân phối của một biến định lượng giữa các danh mục của một biến định tính
* Dual Line Plot: Theo dõi cách hai biến thay đổi cùng nhau theo thời gian (ví dụ: Doanh thu và Chi phí quảng cáo hàng tháng).

*+ Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến?*

**1. Biểu đồ Phân tán (Scatter Plot)**

Biểu đồ này là công cụ chính để phân tích mối quan hệ giữa **hai biến định lượng** (Quantitative Variables), giúp thấy rõ **tương quan** (correlation).

A graph with blue dots

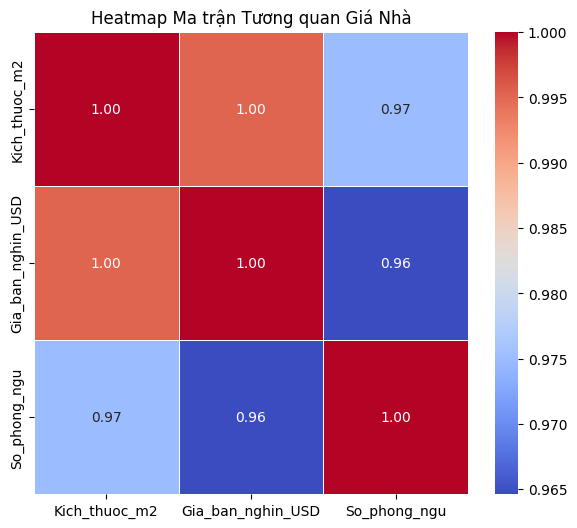
AI-generated content may be incorrect.

+ Biểu đồ sẽ hiển thị các điểm gần như nằm trên một đường thẳng có **độ dốc dương**.

+ Điều này cho thấy có một mối **tương quan dương mạnh** giữa hai biến: **Kích thước nhà càng lớn, giá bán có xu hướng càng cao** (đồng biến).

**2. Biểu đồ Nhiệt (Heatmap)**

Biểu đồ này thường được dùng để trực quan hóa **ma trận tương quan** (cho phân tích đa biến) hoặc để thể hiện mối quan hệ giữa **hai biến phân loại (categorical)** hoặc **hai biến định lượng đã được chia thành các khoảng**.



+ **Màu sắc:** Màu sắc sẽ biểu thị độ mạnh và hướng của tương quan. **Màu đỏ/cam nóng** cho thấy **tương quan dương mạnh** (gần +1), và **màu xanh dương lạnh** cho thấy **tương quan âm mạnh** (gần -1).

+ **Các ô:**

* Ô **Kích thước** và **Giá bán** sẽ có giá trị ≈+0.95 (Màu đỏ nóng), xác nhận mối quan hệ dương rất mạnh.
* Ô **Số phòng ngủ** và **Giá bán** cũng sẽ có giá trị dương mạnh, cho thấy chúng là những yếu tố quan trọng quyết định giá nhà.

*+ Làm thế nào để trực quan hóa mối quan hệ giữa một biến số và một biến phân loại bằng biểu đồ boxplot hoặc violin plot trong Python?*

Biến định lượng: là biến giá trị, thể hiện bằng những con số

Biến phân loại: ví dụ: giới tính, khu vực, loại sản phẩm

**Biểu đồ Hộp (Box Plot):**

A chart with a blue and orange rectangular box

AI-generated content may be incorrect.

**- Đường kẻ giữa Hộp (Trung vị):** Đường kẻ ngang bên trong hộp cho biết **điểm trung vị (Median)** của nhóm đó. Việc so sánh đường kẻ này giúp biết nhóm nào có xu hướng điểm cao hơn.

- **Hộp (IQR):** Chiều dài của hộp biểu thị **Khoảng tứ phân vị (IQR - Q3 - Q1)**, tức là 50% dữ liệu nằm ở đâu. Hộp càng dài, dữ liệu càng phân tán.

**- Râu (Whiskers):** Chiều dài của râu cho thấy mức độ phân tán của dữ liệu bên ngoài IQR.

- **Dấu chấm (Outliers):** Bất kỳ điểm nào nằm ngoài râu thường là **giá trị ngoại lai (Outliers)**.

-> Biểu đồ thể hiện:

- Khối xanh – Trường công lập

+ **Điểm Trung vị (Median)**Khoảng **73** 50% học sinh trường Công lập có điểm thi dưới 73.

+ Q1 ~ 67 và Q3 ~ 77: 50% học sinh tập trung trong khoảng **10 điểm** (67-77).

+ Min 62 và Max 80: Ít học sinh đạt điểm quá cao hoặc quá thấp

- Khối cam – Trường tư

+ Hơn 50% học sinh trường Tư thục có điểm thi dưới 89. Điểm trung vị **cao hơn đáng kể** so với trường Công lập.

+ Từ Q1≈84 đến Q3≈93: 50% học sinh tập trung trong khoảng điểm cao hơn (84-93).

+ Từ min≈82 đến max≈97: Phạm vi điểm cũng khá hẹp và nằm hoàn toàn ở trên so với phạm vi điểm của trường Công lập.

**Biểu đồ Vĩ cầm (Violin Plot):**

A graph of different colored shapes

AI-generated content may be incorrect. **- Hình dạng:** Chiều rộng của "vĩ cầm" tại bất kỳ điểm nào trên trục Y biểu thị **mật độ** dữ liệu tại điểm đó.

- **Phân phối:** Biểu đồ này giúp thấy rõ liệu phân phối có **nhiều đỉnh (multimodal)** hay không, hoặc phân phối có **rất tập trung** hay **rải rác** hơn so với Box Plot.

-> Ý nghĩa

- Trường Công lập (Màu Xanh Đậm)

    + Hình dạng: Hình dạng "vĩ cầm" của trường Công lập khá phẳng và rộng ở giữa, nhưng hơi thắt lại ở hai đầu.

    + Ý nghĩa: Phân phối này có xu hướng phân tán rộng hơn một chút (nhiều học sinh đạt điểm trải dài từ 60 đến 85) nhưng tập trung mạnh quanh vùng điểm trung bình của họ (70-75). Biểu đồ có vẻ hơi hai đỉnh (bimodal) nhẹ, với hai vùng mật độ cao ở khoảng 65-70 và 75-80.

- Trường Tư thục (Màu Cam Đậm)

    + Hình dạng: Hình dạng "vĩ cầm" của trường Tư thục có xu hướng gọn gàng hơn, tập trung hơn và có đỉnh rõ rệt hơn so với trường Công lập.

    + Ý nghĩa: Phân phối này cho thấy điểm số rất tập trung vào vùng cao (85-95). Điểm thi của họ ít bị kéo về phía dưới (ít học sinh đạt điểm thấp hơn 80), minh chứng cho một phân phối chất lượng cao và đồng đều hơn.