**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

A blue and white logo

AI-generated content may be incorrect.**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU**

**PHÂN TÍCH KHÁM PHÁ DỮ LIỆU**

**Giảng viên: Đỗ Như Tài**

**Sinh viên thực hiện:**

1. **Trần Thao-3123410339**
2. **Võ Lập Đỉnh-3123410079**
3. **Võ Hoàng Thanh An - 3121410001**
4. **Hạ Thị Nhi - 3120410371**

**TP.HCM, ngày 10 tháng 10 năm 2025**

**Bảng phân công**

|  |  |
| --- | --- |
| **Thành viên** | **Công việc** |
| Hạ Thị Nhi | Làm lý thuyết + báo cáo + slide |
| Võ Hoàng Thanh An | Làm phần 1.1 |
| Võ Lập Đỉnh | Làm phần 1.2 |
| Trần Thao | Làm phần 1.3 |

**MỤC LỤC**

[I. THỐNG KÊ MÔ TẢ 1](#_Toc22058)

[1.1. Ôn tập lý thuyết 1](#_Toc229)

[1.1.1. Thống kê mô tả là gì? Nó khác gì với thống kê suy luận (inferential statistics)? 1](#_Toc8651)

[a. Thống kê mô tả là gì (Descriptive Statistics)? 1](#_Toc1924)

[b. Nó khác gì với thống kê suy luận (inferential statistics)? 1](#_Toc14857)

[c. Sự khác biệt chính 2](#_Toc153)

[1.1.2. Các thước đo thống kê mô tả chính (ví dụ: trung bình, trung vị, phương sai, độ lệch chuẩn) được sử dụng để làm gì? Trong trường hợp nào thì nên dùng trung vị thay vì trung bình? 4](#_Toc14079)

[a. Nhóm Đo lường Xu hướng Trung tâm (Measures of Central Tendency) 4](#_Toc18495)

[b. Nhóm Đo lường Độ phân tán (Measures of Dispersion) 5](#_Toc16154)

[1.1.3. Làm thế nào để xác định phân bố của một tập dữ liệu? Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)? 7](#_Toc17125)

[a. Làm thế nào để xác định phân bố của một tập dữ liệu? 7](#_Toc1507)

[b. Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)? 8](#_Toc31612)

[Điểm thi của một lớp học giỏi: Một số ít học sinh lười bị điểm thấp kéo đuôi về bên trái. 9](#_Toc8676)

[1.1.4. Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu? 9](#_Toc4160)

[a. Phạm vi (Range) 9](#_Toc31458)

[b. Độ lệch chuẩn (Standard Deviation) 10](#_Toc19084)

[1.1.5. Sự khác biệt giữa các thước đo như Q1, Q2, Q3 trong biểu đồ hộp (boxplot) là gì? 11](#_Toc23238)

[1.1.6. Làm thế nào để xử lý giá trị thiếu (missing values) trước khi tính toán các chỉ số thống kê mô tả? 13](#_Toc25249)

[1.1.7. Bạn có thể giải thích cách đọc và diễn giải một biểu đồ histogram hoặc boxplot từ dữ liệu thực tế không? 15](#_Toc16103)

[a. Biểu đồ Histogram - "Bức tranh toàn cảnh về Phân phối" 15](#_Toc13891)

[b. Biểu đồ Boxplot (Box-and-Whisker Plot) - "Bức tranh tóm tắt 5 điểm" 17](#_Toc13463)

[1.1.8. Khi gặp một tập dữ liệu có giá trị ngoại lai (outliers), bạn sẽ xử lý chúng như thế nào trước khi thực hiện thống kê mô tả? 18](#_Toc12693)

[1.2. Bài tập thực hành 1 21](#_Toc8989)

[1.3. Bài tập thực hành 2 21](#_Toc25513)

[II. XỬ LÝ VÀ TRỰC QUAN HÓA DỮ LIỆU 21](#_Toc13961)

[2.1. Ôn tập lý thuyết 21](#_Toc17552)

[2.1.1. Trực quan hóa dữ liệu có vai trò gì trong phân tích dữ liệu? Tại sao nó quan trọng trong khám phá dữ liệu (EDA)? 21](#_Toc31226)

[a. Vai trò của Trực quan hóa dữ liệu trong Phân tích Dữ liệu 21](#_Toc8434)

[b. Tại sao Trực quan hóa lại ĐẶC BIỆT QUAN TRỌNG trong Khám phá Dữ liệu (EDA)? 22](#_Toc25939)

[2.1.2. Các loại biểu đồ phổ biến (như histogram, scatter plot, boxplot, bar chart) được sử dụng trong các trường hợp nào? 24](#_Toc32640)

[a. Histogram (Biểu đồ Tần suất) 24](#_Toc379)

[b. Boxplot (Biểu đồ Hộp) 24](#_Toc14001)

[c. Scatter Plot (Biểu đồ Phân tán) 24](#_Toc8079)

[d. Bar Chart (Biểu đồ Cột) 25](#_Toc23976)

[2.1.3. Làm thế nào để chọn loại biểu đồ phù hợp với đặc điểm của dữ liệu (ví dụ: dữ liệu phân loại, dữ liệu số, dữ liệu thời gian)? 25](#_Toc17073)

[a. Dữ liệu phân loại (Biến Định Tính) 25](#_Toc22856)

[b. Dữ liệu số (Biến Định Lượng) 25](#_Toc6496)

[2.1.4. Sự khác biệt giữa các thư viện trực quan hóa trong Python như Matplotlib, Seaborn và Plotly là gì? 26](#_Toc3756)

[a. Matplotlib - "Người khổng lồ nền tảng" 27](#_Toc2684)

[b. Seaborn - "Nhà thiết kế thống kê" 28](#_Toc20131)

[c. Plotly - "Nghệ sĩ tương tác" 28](#_Toc15193)

[2.1.5. Những nguyên tắc thiết kế nào cần tuân thủ để tạo ra một biểu đồ trực quan hóa dễ hiểu và hiệu quả? 29](#_Toc27112)

[2.1.6. Làm thế nào để tạo một biểu đồ đơn giản như histogram hoặc bar chart bằng Matplotlib? Bạn có thể chia sẻ đoạn code mẫu không? 31](#_Toc24471)

[2.1.7. Làm thế nào để xuất biểu đồ từ Python ra các định dạng như PNG, PDF hoặc HTML để sử dụng trong báo cáo? 34](#_Toc7206)

[2.2. Bài tập thực hành 1 35](#_Toc26778)

[2.3. Bài tập thực hành 2 35](#_Toc11270)

[III. PHÂN TÍCH ĐƠN BIẾN VÀ HAI BIẾN 35](#_Toc6612)

[3.1. Ôn tập lý thuyết 35](#_Toc20707)

[3.1.1. Phân tích đơn biến (univariate analysis) là gì? Nó khác gì với phân tích hai biến (bivariate analysis) trong khám phá dữ liệu? 35](#_Toc28732)

[a. Phân tích đơn biến (univariate analysis) là gì? 35](#_Toc13008)

[b. Phân tích Hai biến (Bivariate Analysis) 36](#_Toc32425)

[c. Sự khác biệt cốt lỗi 37](#_Toc16401)

[3.1.2. 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)? 37](#_Toc11760)

[a. Đo lường Xu hướng Trung tâm (Measures of Central Tendency) 37](#_Toc11074)

[b. Đo lường Độ phân tán (Measures of Dispersion) 38](#_Toc1772)

[c. Đo lường Hình dạng Phân phối (Measures of Shape) 39](#_Toc16313)

[3.1.3. Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến (ví dụ: tương quan, nhân quả)? 40](#_Toc7650)

[3.1.4. Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance) trong phân tích hai biến là gì? 42](#_Toc26084)

[a. Hiệp biến (covariance) 42](#_Toc12763)

[b. Tương quan (Correlation) 43](#_Toc16307)

[3.1.5. Khi nào nên sử dụng biểu đồ trực quan hóa trong phân tích đơn biến so với phân tích hai biến? 44](#_Toc26596)

[a. PHÂN TÍCH ĐƠN BIẾN - Khi nào sử dụng? 44](#_Toc17751)

[b. PHÂN TÍCH HAI BIẾN - Khi nào sử dụng? 44](#_Toc2399)

[3.1.6. Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến? 45](#_Toc19665)

[a. Scatter Plot 45](#_Toc9672)

[b. Heatmap 45](#_Toc31533)

[c. Diễn giải 46](#_Toc15250)

[3.1.7. 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? 47](#_Toc21594)

[a. Boxplot Cơ Bản & Nâng Cao 47](#_Toc26893)

[b. Violin Plot - Hiển thị Phân phối Chi tiết 48](#_Toc1790)

[c. Phân Tích Nâng Cao với Swarm Plot & Kết Hợp 49](#_Toc28030)

[3.2. Bài tập thực hành 1 52](#_Toc15290)

[3.3. Bài tập thực hành 2 52](#_Toc17343)

**PHÂN TÍCH KHÁM PHÁ DỮ LIỆU**

1. **THỐNG KÊ MÔ TẢ**
   1. **Ôn tập lý thuyết**
      1. ***Thống kê mô tả là gì? Nó khác gì với thống kê suy luận (inferential statistics)?***
2. *Thống kê mô tả là gì (Descriptive Statistics)?*

*\* Định nghĩa:*

Thống kê mô tả là phương pháp dùng để tóm tắt và mô tả các đặc điểm chính của một tập dữ liệu, sử dụng các số liệu như trung bình, trung vị, mốt, độ lệch chuẩn, hay các biểu đồ để trình bày thông tin một cách trực quan, giúp người phân tích hiểu rõ hơn về dữ liệu mà không cần suy luận về một quần thể lớn hơn.

*\* Mục đích:*

• Mô tả: Cho biết dữ liệu trông như thế nào.

• Tóm tắt: Rút gọn một lượng lớn dữ liệu thô thành các thông số có ý nghĩa.

• Trình bày: Sử dụng các bảng biểu, đồ thị để hiển thị dữ liệu.

*\* Phạm vi:*

• Thống kê mô tả chỉ quan tâm đến chính tập dữ liệu bạn đang có (mẫu dữ liệu hoặc toàn bộ tổng thể). Nó không đưa ra kết luận hay dự đoán nào vượt ra ngoài phạm vi dữ liệu đó.

1. *Nó khác gì với thống kê suy luận (inferential statistics)?*

*\* Định nghĩa:*

Thống kê suy luận là các phương pháp sử dụng một mẫu dữ liệu nhỏ (sample) để đưa ra kết luận, ước tính hoặc dự đoán về một tập hợp dữ liệu lớn hơn (tổng thể - population). Nó liên quan đến việc suy ra các đặc tính của tổng thể dựa trên thông tin từ mẫu.

*\* Mục đích:*

• Suy luận: Đưa ra kết luận về tổng thể dựa trên mẫu.

• Ước tính: Dự đoán các tham số của tổng thể (như trung bình tổng thể, tỷ lệ tổng thể).

• Kiểm định giả thuyết: Kiểm tra xem một giả thuyết về tổng thể có đáng tin cậy hay không.

• Dự báo: Dự đoán các giá trị trong tương lai.

*\* Phạm vi:*

Thống kê suy luận vượt ra ngoài phạm vi dữ liệu hiện có. Nó sử dụng dữ liệu mẫu để khái quát hóa cho một nhóm lớn hơn.

\* Các kỹ thuật chính:

• Ước lượng (Estimation):

- Ước lượng điểm (Point Estimation): Dùng một giá trị duy nhất từ mẫu để ước tính tham số tổng thể (ví dụ: dùng trung bình mẫu để ước tính trung bình tổng thể).

- Ước lượng khoảng (Interval Estimation): Dùng một khoảng giá trị để ước tính tham số tổng thể (ví dụ: Khoảng tin cậy 95%).

• Kiểm định giả thuyết (Hypothesis Testing): Kiểm tra tính hợp lệ của một tuyên bố về tổng thể (ví dụ: kiểm tra xem một loại thuốc mới có thực sự hiệu quả hơn giả dược hay không).

• Hồi quy (Regression Analysis): Phân tích mối quan hệ giữa các biến để dự đoán.

• Phân tích phương sai (ANOVA): So sánh sự khác biệt giữa nhiều nhóm.

1. *Sự khác biệt chính*

|  |  |  |
| --- | --- | --- |
| **Đặc điểm** | **Thống kê mô tả** | **Thống kê suy luận** |
| **Mục đích** | Tóm tắt, mô tả và trình bày dữ liệu. | Suy luận, dự đoán và khái quát hóa từ mẫu ra tổng thể. |
| **Phạm vi** | Chỉ xử lý dữ liệu đang có. | Mở rộng kết luận ra ngoài dữ liệu hiện có. |
| **Kết quả** | Trả lời câu hỏi: "*Dữ liệu này trông như thế nào?*" | Trả lời câu hỏi: "*Điều gì có thể xảy ra dựa trên dữ liệu này?*" |
| **Đầu vào** | Toàn bộ tập dữ liệu (có thể là mẫu hoặc tổng thể). | Một mẫu dữ liệu được thu thập từ tổng thể. |
| **Đầu ra** | Các tham số (số trung bình, độ lệch chuẩn), bảng biểu, đồ thị. | Các ước lượng, khoảng tin cậy, kết luận kiểm định giả thuyết. |
| **Mức độ không chắc chắn** | Không có sự không chắc chắn vì chỉ mô tả dữ liệu hiện có. | Luôn tồn tại một mức độ không chắc chắn hoặc sai số (sai số mẫu). |

**Ví dụ minh họa để phân biệt**

Giả sử chúng ta muốn nghiên cứu chiều cao trung bình của sinh viên toàn trường Đại học X (tổng thể có 20,000 sinh viên).

• Bước 1 - Thống kê mô tả:

- Bạn không thể đo hết 20,000 người, nên bạn chọn ngẫu nhiên 200 sinh viên (đây là một mẫu).

- Bạn tính toán cho 200 sinh viên này: Chiều cao trung bình là 1.65m, độ lệch chuẩn là 0.1m. Bạn vẽ một biểu đồ histogram cho thấy phân phối chiều cao.

- Kết luận từ Thống kê mô tả: "Trong mẫu 200 sinh viên này, chiều cao trung bình là 1.65m."

-> Hành động này là mô tả dữ liệu của mẫu.

• Bước 2 - Thống kê suy luận:

- Từ dữ liệu mẫu (1.65m), bạn muốn suy ra chiều cao trung bình của toàn bộ 20,000 sinh viên.

- Bạn sử dụng kỹ thuật ước lượng khoảng tin cậy và kết luận: "Với độ tin cậy 95%, chúng tôi ước tính rằng chiều cao trung bình của tất cả sinh viên trường Đại học X nằm trong khoảng từ 1.63m đến 1.67m."

- Kết luận từ Thống kê suy luận: Đây là một suy luận về tổng thể, dựa trên thông tin từ mẫu, và nó thừa nhận có một mức độ không chắc chắn (khoảng 1.63m - 1.67m).

* + 1. ***Các thước đo thống kê mô tả chính (ví dụ: trung bình, trung vị, phương sai, độ lệch chuẩn) được sử dụng để làm gì? Trong trường hợp nào thì nên dùng trung vị thay vì trung bình?***

1. *Nhóm Đo lường Xu hướng Trung tâm (Measures of Central Tendency)*

• Trung bình (Mean - Average):

- Là gì: Tổng của tất cả các giá trị chia cho số lượng giá trị.

- Dùng để làm gì: Cung cấp một cái nhìn tổng quan về "mức độ trung bình" của toàn bộ tập dữ liệu. Nó rất nhạy cảm với mọi giá trị trong dữ liệu.

- Ví dụ: Điểm trung bình môn học (GPA) cho biết học lực chung của sinh viên.

• Trung vị (Median):

- Là gì: Giá trị nằm chính giữa khi sắp xếp tất cả các giá trị theo thứ tự tăng dần.

- Dùng để làm gì: Chỉ ra điểm chia đôi tập dữ liệu (50% số liệu nằm dưới, 50% nằm trên trung vị). Nó không bị ảnh hưởng bởi các giá trị ngoại lai (outliers) cực lớn hoặc cực nhỏ.

- Ví dụ: Thu nhập trung vị của một khu vực cho biết một người "ở giữa" kiếm được bao nhiêu, giúp hình dung tốt hơn về mức sống phổ biến so với thu nhập trung bình (vốn có thể bị kéo lên rất cao bởi một vài tỷ phú).

1. *Nhóm Đo lường Độ phân tán (Measures of Dispersion)*

• Phương sai (Variance):

- Là gì: Trung bình của bình phương các độ lệch so với giá trị trung bình. Nó đo lường mức độ biến thiên tổng thể của dữ liệu.

- Dùng để làm gì: Cung cấp một con số về sự "lan truyền" của dữ liệu. Tuy nhiên, đơn vị của phương sai là bình phương của đơn vị gốc (ví dụ: nếu dữ liệu là mét (m), phương sai là m²), khiến nó khó diễn giải trực tiếp.

• Độ lệch chuẩn (Standard Deviation):

- Là gì: Căn bậc hai của phương sai.

- Dùng để làm gì: Đây là thước đo quan trọng và phổ biến nhất về độ phân tán. Nó trả về cùng đơn vị với dữ liệu gốc, giúp chúng ta dễ dàng hiểu được dữ liệu biến động bao nhiêu xung quanh giá trị trung bình.

- Ví dụ: Trong đầu tư, độ lệch chuẩn của lợi nhuận cổ phiếu được dùng để đo lường rủi ro. Độ lệch chuẩn cao đồng nghĩa với sự biến động mạnh và rủi ro cao.

**\* Khi nào nên dùng Trung vị thay vì Trung bình?**

Việc lựa chọn giữa Trung vị và Trung bình phụ thuộc vào tính chất của dữ liệu. Nguyên tắc chung là: Sử dụng Trung vị khi tập dữ liệu có sự hiện diện của các giá trị ngoại lai (outliers) hoặc phân phối lệch.

Dưới đây là các trường hợp cụ thể:

*+ Khi dữ liệu có chứa giá trị ngoại lai (Outliers):*

Đây là lý do phổ biến và quan trọng nhất. Trung bình rất nhạy cảm với các giá trị cực đoan, trong khi Trung vị thì "miễn nhiễm".

• Ví dụ điển hình: Thu nhập hộ gia đình trong một khu phố.

- Giả sử chúng ta có thu nhập của 5 hộ (triệu đồng/tháng): 10, 15, 20, 25, 100 (Hộ cuối cùng là một doanh nhân giàu có).

- Trung bình = (10+15+20+25+100)/5 = 34 triệu.

- Trung vị = Giá trị ở giữa (thứ 3) = 20 triệu.

- Rõ ràng, số 34 triệu không đại diện tốt cho phần đông cư dân trong khu phố (4/5 hộ có thu nhập dưới 26 triệu). Trong khi đó, 20 triệu cho chúng ta hình ảnh chính xác hơn về mức thu nhập phổ biến. Trong trường hợp này, Trung vị là thước đo tốt hơn.

*+ Khi phân phối của dữ liệu bị lệch (Skewed Distribution):*

• Phân phối lệch phải (Positive Skew): Đuôi của phân phối kéo dài về bên phải (nhiều giá trị nhỏ, một vài giá trị rất lớn). Ví dụ: Thu nhập, giá nhà đất. => Dùng Trung vị.

• Phân phối lệch trái (Negative Skew): Đuôi của phân phối kéo dài về bên trái. => Dùng Trung vị.

• Phân phối đối xứng (Symmetrical): Dữ liệu phân bố cân đối. Ví dụ: Chiều cao của người trưởng thành. => Có thể dùng cả Trung bình và Trung vị, vì lúc này chúng gần như trùng khớp.

+ Khi thang đo là thứ tự (Ordinal Scale):

Với dữ liệu thứ tự (ví dụ: xếp hạng mức độ hài lòng: 1. Rất tệ, 2. Tệ, 3. Bình thường, 4. Tốt, 5. Rất tốt), việc tính trung bình có thể không có ý nghĩa (ví dụ: trung bình là 3.5 thì không tương ứng với hạng mục nào). Trung vị là lựa chọn phù hợp vì nó chỉ dựa trên thứ tự sắp xếp.

**\* Tóm tắt lựa chọn**

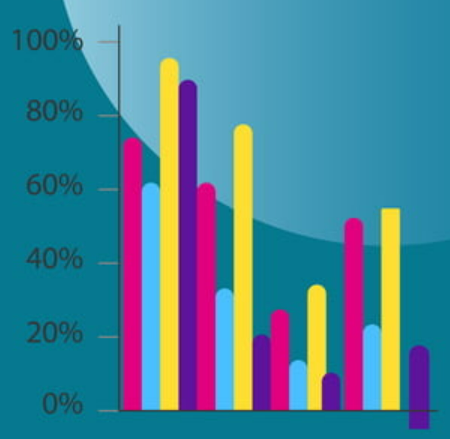
|  |  |  |
| --- | --- | --- |
| **Đặc điểm dữ liệu** | **Nên dùng** | **Lý do** |
| Phân phối đối xứng, không có ngoại lai | Trung bình | Tận dụng được tất cả thông tin từ dữ liệu, cho kết quả chính xác. |
| Có giá trị ngoại lai, phân phối lệch | Trung vị | Robust (vững vàng), không bị ảnh hưởng bởi các giá trị cực đoan, đại diện tốt hơn cho số đông. |
| Dữ liệu thứ tự (ordinal) | Trung vị | Phù hợp với bản chất của thang đo, trong khi trung bình có thể vô nghĩa. |

* + 1. ***Làm thế nào để xác định phân bố của một tập dữ liệu? Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)?***

1. *Làm thế nào để xác định phân bố của một tập dữ liệu?*

+ Trực quan hóa bằng biểu đồ (Visualization) - Phương pháp trực quan nhất: Đây là bước đầu tiên và quan trọng nhất.

• Biểu đồ Histogram (Biểu đồ tần suất):



- Cách dùng: Chia dữ liệu thành các "bin" (khoảng giá trị) và đếm xem có bao nhiêu điểm dữ liệu rơi vào mỗi bin.

- Tác dụng: Cho ta thấy ngay hình dạng của phân bố: đỉnh ở đâu, đuôi kéo dài về bên nào, có đối xứng không.

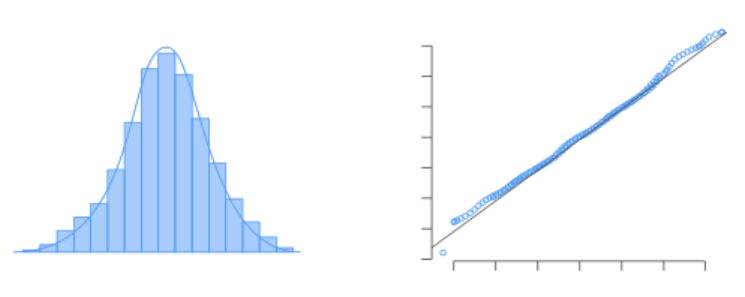
- Ví dụ: Một histogram có hình "quả chuông" gợi ý phân bố chuẩn.

• Biểu đồ Q-Q (Quantile-Quantile Plot):

- Cách dùng: So sánh các phân vị của dữ liệu thực tế với các phân vị của một phân bố lý thuyết (ví dụ: phân bố chuẩn).

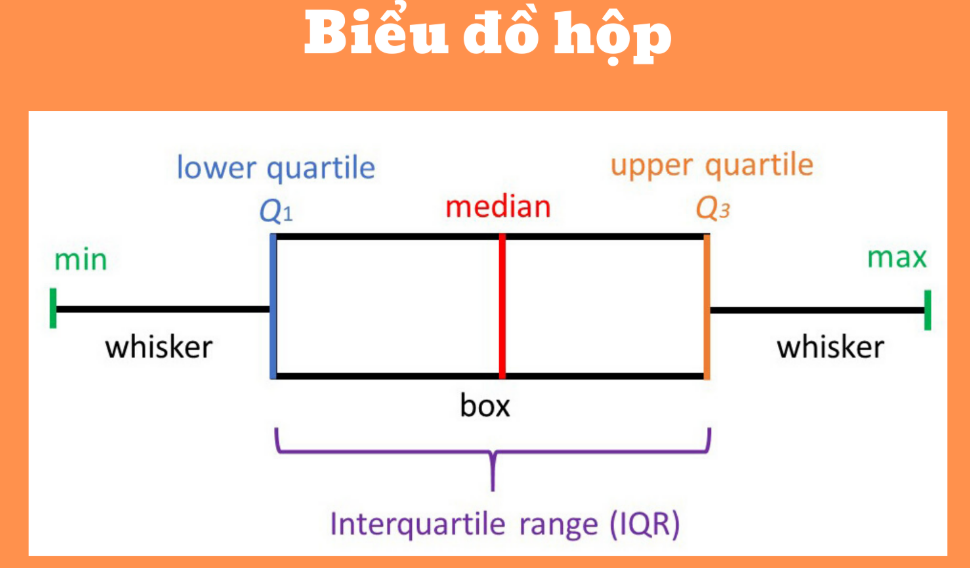
- Tác dụng: Rất tốt để kiểm tra xem dữ liệu có tuân theo một phân bố cụ thể nào không. Nếu các điểm dữ liệu nằm gần đường chéo, thì dữ liệu phù hợp với phân bố đó.

- Ví dụ: Dùng để kiểm tra tính chuẩn của dữ liệu.



*Biểu đồ dữ liệu phân phối chuẩn*

• Biểu đồ Box Plot (Biểu đồ hộp):

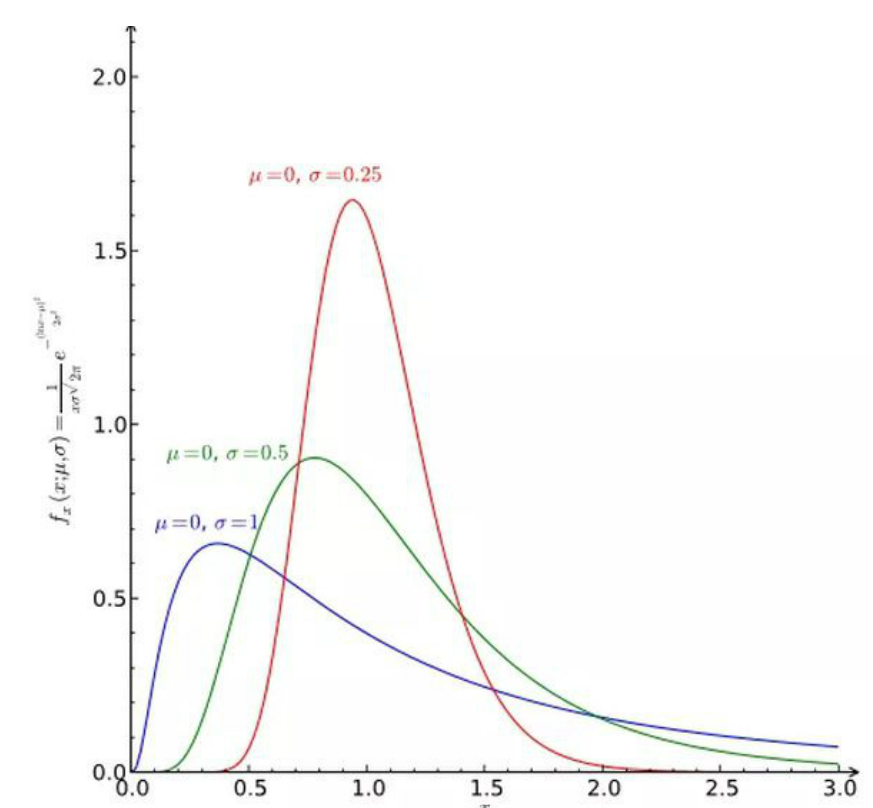


- Cách dùng: Hiển thị trung vị, tứ phân vị, và các điểm ngoại lai.

- Tác dụng: Giúp phát hiện sự lệch phải/lệch trái một cách dễ dàng. Nếu hộp bị "kéo" về một phía và râu bên kia dài hơn, đó là dấu hiệu của phân bố lệch.

+ Sử dụng các thước đo thống kê (Statistical Measures)

• Độ lệch (Skewness):



- Là gì: Thước đo mức độ đối xứng của phân bố.

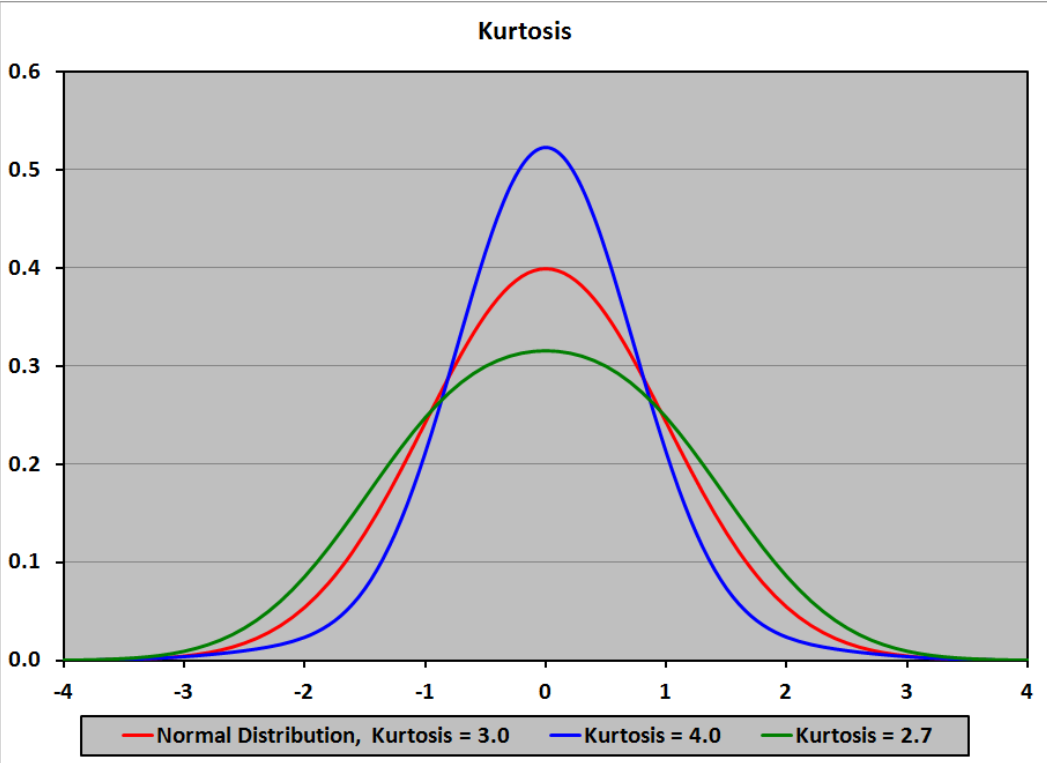
- Cách diễn giải:

Skewness ≈ 0: Phân bố đối xứng (như phân bố chuẩn).

Skewness > 0: Phân bố lệch phải (đuôi kéo dài về bên phải).

Skewness < 0: Phân bố lệch trái (đuôi kéo dài về bên trái).

• Độ nhọn (Kurtosis):



- Là gì: Thước đo độ "béo" hay "nhọn" của đuôi phân bố so với phân bố chuẩn.

- Cách diễn giải:

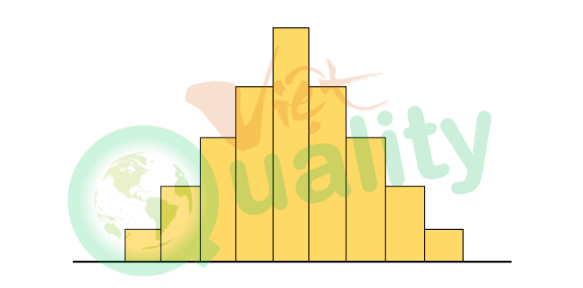
Kurtosis ≈ 0 (hoặc 3 tùy vào công thức): Độ nhọn tương tự phân bố chuẩn.

Kurtosis > 0: Phân bố "nhọn" hơn phân bố chuẩn, nhiều dữ liệu tập trung ở đuôi.

Kurtosis < 0: Phân bố "phẳng" hơn phân bố chuẩn, ít dữ liệu ở đuôi hơn.

1. *Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)?*

*+ Phân bố Chuẩn (Normal Distribution - Phân bố Gauss)*



• Hình dạng: Hình "quả chuông" đối xứng hoàn hảo.

• Đặc điểm:

- Trung bình = Trung vị = Yếu vị.

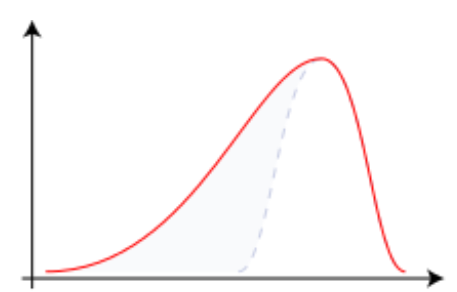
- Có dạng đối xứng qua trung bình.

- Được mô tả đầy đủ chỉ bằng hai tham số: Trung bình (μ - vị trí) và Độ lệch chuẩn (σ - độ trải rộng).

• Ứng dụng: Rất phổ biến trong tự nhiên và xã hội. Ví dụ: Chiều cao, điểm số IQ, sai số đo lường.

+ Phân bố Lệch (Skewed Distribution)

• Phân bố Lệch Phải (Positive Skew):



- Hình dạng: Đuôi kéo dài về phía bên phải (phía giá trị lớn hơn).

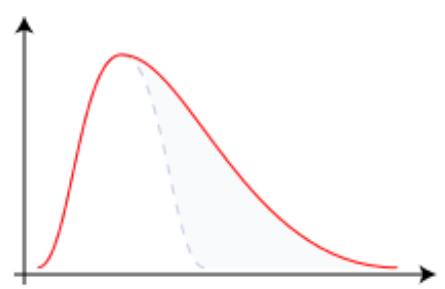
- Mối quan hệ: Trung bình > Trung vị > Yếu vị. Giá trị trung bình bị kéo lên cao bởi các giá trị ngoại lai ở đuôi phải.

- Ví dụ:

Thu nhập: Đa số người có thu nhập tập trung ở mức thấp và trung bình, một số ít tỷ phú kéo đuôi phân bố về bên phải.

Thời gian chờ đợi: Hầu hết khách hàng được phục vụ nhanh, một vài trường hợp phức tạp làm thời gian chờ kéo dài.

• Phân bố Lệch Trái (Negative Skew):



- Hình dạng: Đuôi kéo dài về phía bên trái (phía giá trị nhỏ hơn).

- Mối quan hệ: Trung bình < Trung vị < Yếu vị. Giá trị trung bình bị kéo xuống thấp bởi các giá trị ngoại lai ở đuôi trái.

- Ví dụ:

Tuổi thọ ở các nước phát triển: Số ít người mất sớm kéo đuôi về bên trái, trong khi đa số sống đến tuổi già.

Điểm thi của một lớp học giỏi: Một số ít học sinh lười bị điểm thấp kéo đuôi về bên trái.

* + 1. ***Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu?***

1. *Phạm vi (Range)*

+ Định nghĩa: Phạm vi là sự chênh lệch giữa giá trị lớn nhất và giá trị nhỏ nhất trong tập dữ liệu.

+ Công thức: Phạm vi = Giá trị lớn nhất (Max) - Giá trị nhỏ nhất (Min)

\* Ý nghĩa trong đánh giá độ phân tán:

• Ưu điểm:

- Rất dễ hiểu và dễ tính toán. Nó cho bạn một cái nhìn tức thì về "khoảng bao phủ" tổng thể của dữ liệu.

- Hữu ích để nhanh chóng nắm bắt quy mô của tập dữ liệu.

• Hạn chế lớn:

- Cực kỳ nhạy cảm với giá trị ngoại lai (outliers). Chỉ cần một giá trị cực đoan duy nhất cũng có thể làm biến dạng hoàn toàn phạm vi, khiến nó không còn phản ánh chính xác sự phân tán của đa số dữ liệu.

• Ví dụ minh họa:

Xem xét điểm thi của hai lớp học (thang điểm 10):

Lớp A: 4, 5, 5, 6, 6, 7, 7 → Phạm vi = 7 - 4 = 3

Lớp B: 4, 5, 5, 6, 6, 7, 10 → Phạm vi = 10 - 4 = 6

Nếu chỉ nhìn vào phạm vi, ta có thể kết luận điểm số của Lớp B "phân tán" gấp đôi Lớp A. Tuy nhiên, sự thật là 6/7 học sinh của Lớp B có điểm số rất tập trung, giống hệt Lớp A. Sự khác biệt lớn về phạm vi hoàn toàn đến từ một học sinh duy nhất đạt điểm 10. Phạm vi ở đây đã bị "lừa" bởi giá trị ngoại lai.

1. *Độ lệch chuẩn (Standard Deviation)*

+ Định nghĩa: Độ lệch chuẩn đo lường mức độ dao động trung bình của các điểm dữ liệu xung quanh giá trị trung bình (mean) của tập dữ liệu.

+ Công thức: (Căn bậc hai của phương sai) - nhưng quan trọng là hiểu ý nghĩa.

\* Ý nghĩa trong đánh giá độ phân tán:

• Ưu điểm:

- Sử dụng tất cả các giá trị trong tập dữ liệu, không chỉ giá trị lớn nhất và nhỏ nhất. Điều này làm cho nó là một thước đo "toàn diện" hơn nhiều.

- Ít nhạy cảm với giá trị ngoại lai hơn so với phạm vi. Nó cho biết mức độ biến động "điển hình" của dữ liệu.

- Có cùng đơn vị với dữ liệu gốc, giúp việc diễn giải trở nên dễ dàng.

• Cách diễn giải:

- Độ lệch chuẩn THẤP: Các điểm dữ liệu có xu hướng nằm rất gần với giá trị trung bình. Đồ thị phân bố sẽ cao và hẹp. → Độ phân tán THẤP, dữ liệu đồng đều.

- Độ lệch chuẩn CAO: Các điểm dữ liệu trải rộng trên một phạm vi giá trị lớn xung quanh giá trị trung bình. Đồ thị phân bố sẽ thấp và rộng. → Độ phân tán CAO, dữ liệu biến động mạnh.

• Ví dụ minh họa (tiếp nối ví dụ trên):

Giả sử chúng ta tính độ lệch chuẩn (ký hiệu là \*s\*) cho hai lớp:

Lớp A: \*s\* ≈ 1.0

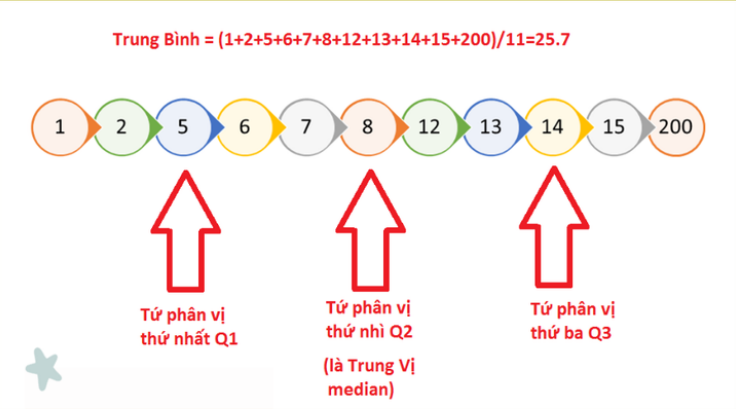
Lớp B: \*s\* ≈ 1.8

Bây giờ, chúng ta có một bức tranh chính xác hơn nhiều:

Lớp A có độ lệch chuẩn thấp (1.0), xác nhận rằng điểm số của họ rất tập trung.

Lớp B có độ lệch chuẩn cao hơn (1.8), phản ánh sự hiện diện của điểm 10 làm tăng độ biến động, nhưng mức tăng này không quá khủng khiếp như phạm vi (6) đã gợi ý. Con số 1.8 cho thấy sự phân tán tuy có cao hơn nhưng không đến mức gấp đôi.

* + 1. ***Sự khác biệt giữa các thước đo như Q1, Q2, Q3 trong biểu đồ hộp (boxplot) là gì?***



\* Định nghĩa từng Thước đo

Để hiểu chúng, trước tiên hãy sắp xếp dữ liệu theo thứ tự tăng dần. Các tứ phân vị sau đó sẽ chia tập dữ liệu này thành 4 phần bằng nhau về số lượng quan sát.

• Q2 (Tứ phân vị thứ hai):

- Đây chính là Trung vị (Median).

- Nó là gì? Giá trị nằm ở chính giữa tập dữ liệu. 50% số liệu nằm dưới Q2 và 50% nằm trên Q2.

- Trong boxplot: Đây là đường vạch ngang bên trong cái hộp. Nó cho biết "trung tâm" của dữ liệu nằm ở đâu.

• Q1 (Tứ phân vị thứ nhất):

- Nó là gì? Trung vị của nửa dưới của tập dữ liệu (nửa nằm dưới Q2).

- Vị trí: 25% số liệu nằm dưới Q1. Nó đánh dấu ranh giới dưới của "khoảng giữa" của dữ liệu.

- Trong boxplot: Đây là cạnh dưới của cái hộp.

• Q3 (Tứ phân vị thứ ba):

- Nó là gì? Trung vị của nửa trên của tập dữ liệu (nửa nằm trên Q2).

- Vị trí: 75% số liệu nằm dưới Q3. Nó đánh dấu ranh giới trên của "khoảng giữa" của dữ liệu.

- Trong boxplot: Đây là cạnh trên của cái hộp.

\* Ý Nghĩa của Sự Khác Biệt Giữa Chúng

Sự khác biệt giữa các Q không chỉ là vị trí mà còn cho chúng ta biết rất nhiều về đặc tính của dữ liệu.

+ Khoảng Liên Phân Vị (IQR - Interquartile Range)

• Là gì: IQR = Q3 - Q1

• Ý nghĩa: Đây là thước đo độ phân tán của 50% dữ liệu trung tâm. Nó cho biết khoảng cách giữa nhóm 25% dữ liệu cao nhất trong nửa trung tâm và nhóm 25% thấp nhất trong nửa trung tâm.

• Ứng dụng quan trọng nhất: Dùng để xác định các điểm ngoại lai (outliers). Một điểm dữ liệu thường được coi là ngoại lai nếu nó nằm ngoài khoảng:

- Dưới: Q1 - 1.5 \* IQR

- Trên: Q3 + 1.5 \* IQR

+ Độ Dài của "Cái Hộp" (The Box)

• Chính cái hộp trong boxplot được tạo nên bởi Q1 và Q3.

• Hộp ngắn/dẹt: Có nghĩa là IQR nhỏ. 50% dữ liệu trung tâm tập trung rất gần nhau (ít biến động).

• Hộp dài/cao: Có nghĩa là IQR lớn. 50% dữ liệu trung tâm trải rộng trên một khoảng giá trị lớn (nhiều biến động).

+ Vị trí của Q2 (Trung vị) trong Hộp

Vị trí của đường Q2 so với Q1 và Q3 tiết lộ hình dạng phân phối của dữ liệu:

• Q2 nằm ở giữa hộp (khoảng cách từ Q1 đến Q2 ≈ từ Q2 đến Q3):

- Phân phối gần như đối xứng. Dữ liệu phân bố đều ở cả hai bên trung vị.

-> Phân phối đối xứng

• Q2 nằm gần đáy hộp (khoảng cách từ Q1 đến Q2 < từ Q2 đến Q3):

- Phân phối lệch phải (Positive Skew). Phần dữ liệu bên phải trung vị (các giá trị lớn) trải dài hơn so với phần bên trái. "Đuôi" bên phải dài hơn.

-> Phân phối lệch phải

• Q2 nằm gần đỉnh hộp (khoảng cách từ Q1 đến Q2 > từ Q2 đến Q3):

- Phân phối lệch trái (Negative Skew). Phần dữ liệu bên trái trung vị (các giá trị nhỏ) trải dài hơn. "Đuôi" bên trái dài hơn.

-> Phân phối lệch trái

* + 1. ***Làm thế nào để xử lý giá trị thiếu (missing values) trước khi tính toán các chỉ số thống kê mô tả?***

Các Phương Pháp Xử Lý Giá Trị Thiếu

\* Nhóm 1: Các Phương Pháp Đơn Giản & "Ngây thơ"

+ Xóa (Deletion):

• Xóa toàn bộ dòng (Listwise Deletion): Loại bỏ bất kỳ dòng nào có chứa ít nhất một giá trị thiếu.

- Ưu điểm: Đơn giản, dễ làm.

- Nhược điểm: Lãng phí dữ liệu khổng lồ, có thể làm giảm sức mạnh thống kê và dẫn đến kết quả chệch nếu dữ liệu không thiếu ngẫu nhiên (MCAR).

- Khi nào dùng: Khi tỷ lệ giá trị thiếu rất nhỏ (<5%) và dữ liệu có vẻ như MCAR.

+ Thay thế bằng Giá trị Cố định (Imputation with Fixed Value):

• Thay thế bằng 0, "Unknown", hoặc một hằng số đặc biệt.

- Ưu điểm: Đơn giản, giữ nguyên kích thước mẫu.

- Nhược điểm: Làm biến dạng phân phối của dữ liệu, có thể tạo ra "đỉnh" giả tạo tại giá trị thay thế.

- Khi nào dùng: Khi có lý do chính đáng (ví dụ: thay thế doanh thu thiếu bằng 0 nếu chắc chắn không có giao dịch).

\* Nhóm 2: Các Phương Pháp Thay thế (Imputation) Thông minh hơn

Đây là nhóm phương pháp được khuyến nghị vì chúng tận dụng thông tin từ chính dữ liệu.

+ Thay thế bằng Giá trị Trung tâm (Central Tendency Imputation):

• Thay thế bằng Trung bình (Mean) hoặc Trung vị (Median).

- Ưu điểm: Đơn giản hơn các phương pháp phức tạp, giữ nguyên trung bình của biến.

- Nhược điểm: Làm giảm phương sai (độ phân tán) của dữ liệu. Các mối quan hệ với biến khác có thể bị yếu đi.

- Khi nào dùng:

Dùng Trung bình: Cho dữ liệu số, phân phối đối xứng, không có ngoại lai.

Dùng Trung vị: Cho dữ liệu số, phân phối lệch hoặc có ngoại lai.

Dùng Yếu vị (Mode): Cho dữ liệu phân loại (categorical).

+ Thay thế bằng Hồi quy/Phân loại (Regression/Classification Imputation):

• Cách làm: Sử dụng các biến khác không bị thiếu để dự đoán giá trị còn thiếu.

- Với biến số: Dùng Hồi quy tuyến tính.

- Với biến phân loại: Dùng Mô hình phân loại (ví dụ: Logistic Regression, Decision Tree).

• Ưu điểm: Giữ được phương sai và các mối quan hệ giữa các biến tốt hơn so với thay thế bằng trung bình.

• Nhược điểm: Phức tạp, có thể tạo ra sự không chắc chắn vì giá trị được thay thế chỉ là ước tính.

+ Phương pháp K-NN (K-Nearest Neighbors Imputation):

• Cách làm: Tìm K dòng dữ liệu "giống nhất" (gần nhất) với dòng có chứa giá trị thiếu, rồi thay thế giá trị thiếu bằng trung bình (hoặc yếu vị) của K dòng đó.

• Ưu điểm: Rất linh hoạt, không cần giả định về phân phối của dữ liệu.

• Nhược điểm: Cần chọn K phù hợp, tốn tài nguyên tính toán với tập dữ liệu lớn.

+ Phương pháp MICE (Multiple Imputation by Chained Equations):

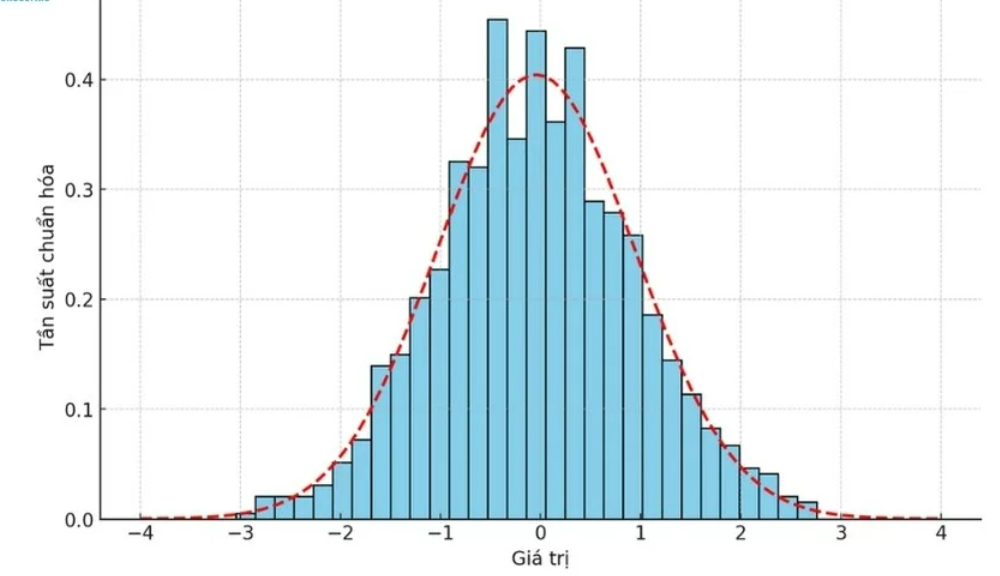
• Cách làm: Đây là phương pháp tinh vi và mạnh mẽ nhất. Nó tạo ra nhiều bản sao của tập dữ liệu, mỗi bản sao có các giá trị được thay thế khác nhau một chút. Sau đó, phân tích trên từng bản sao và kết hợp kết quả lại.

• Ưu điểm: Tính đến sự không chắc chắn của việc thay thế, cho kết quả đáng tin cậy nhất.

• Nhược điểm: Rất phức tạp để triển khai và giải thích.

* + 1. ***Bạn có thể giải thích cách đọc và diễn giải một biểu đồ histogram hoặc boxplot từ dữ liệu thực tế không?***

1. *Biểu đồ Histogram - "Bức tranh toàn cảnh về Phân phối"*



Cách đọc:

+ Trục ngang (X): Thể hiện các khoảng giá trị của dữ liệu (ví dụ: khoảng điểm số, khoảng thu nhập).

+ Trục dọc (Y): Thể hiện tần suất (số lượng) các quan sát rơi vào mỗi khoảng giá trị.

+ Các cột (Bars): Mỗi cột đại diện cho một khoảng giá trị. Cột càng cao có nghĩa là càng nhiều dữ liệu nằm trong khoảng đó.

**\* Ví dụ thực tế:** Điểm thi cuối kỳ của 100 sinh viên

Giả sử bạn có biểu đồ histogram sau:

(Hình ảnh minh họa trong tưởng tượng: Một histogram với trục X là "Điểm số" từ 0-10, trục Y là "Số sinh viên". Các cột cao dần từ 0-7 điểm, đỉnh cao nhất ở khoảng 7-8 điểm, sau đó thấp dần từ 8-10 điểm).

Cách diễn giải:

+ Hình dạng phân phối:

• Phân phối có dạng "chuông" (Unimodal): Bạn thấy một "đỉnh" duy nhất ở khoảng điểm 7-8. Điều này cho thấy đa số sinh viên có điểm số tập trung quanh mức này.

• Phân phối lệch trái (Negative Skew): Nếu đuôi của phân phối kéo dài về phía điểm thấp (0-4), điều đó có nghĩa là có một nhóm nhỏ sinh viên bị điểm rất kém, trong khi hầu hết đều đạt điểm khá cao.

+ Xu hướng trung tâm:

• Bằng mắt thường, bạn có thể ước lượng số trung bình (mean) và trung vị (median) nằm ở đâu. Trong ví dụ này, chúng có thể nằm quanh khoảng 7-8 điểm.

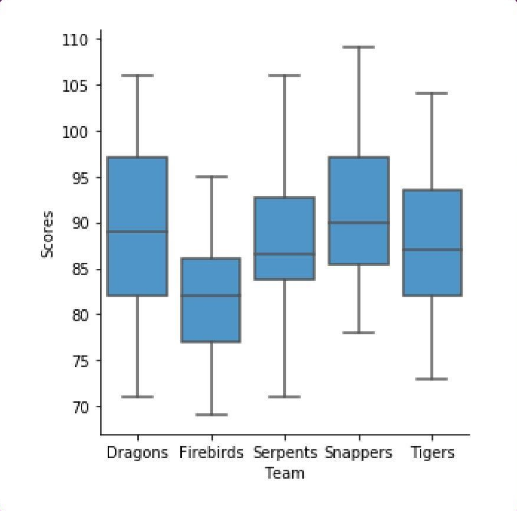
+ Độ phân tán:

• Biểu đồ trải dài từ 0 đến 10 điểm, cho thấy điểm số có sự chênh lệch đáng kể.

• Nếu các cột rất "nhọn" và tập trung quanh một khoảng hẹp, độ phân tán thấp (sinh viên đều có điểm gần như nhau). Nếu các cột "trải rộng", độ phân tán cao (điểm số chênh lệch nhiều).

-> **Kết luận từ ví dụ:** "Bài thi này có kết quả khá tốt, với phần đông sinh viên đạt điểm trong khoảng 7-8. Tuy nhiên, vẫn tồn tại một số ít sinh viên có điểm số rất thấp, kéo đuôi phân phối về bên trái."

1. *Biểu đồ Boxplot (Box-and-Whisker Plot) - "Bức tranh tóm tắt 5 điểm"*



Cách đọc:

Hãy tưởng tượng một boxplot cho cùng tập dữ liệu điểm thi trên:

(Hình ảnh minh họa trong tưởng tượng: Một boxplot nằm ngang. Từ trái sang phải: Một dấu chấm ở 1.5 điểm (outlier), "râu" trái kéo từ 4 đến 5, cạnh hộp bên trái ở 5.5, đường vạch trong hộp ở 7.5, cạnh hộp bên phải ở 8.5, "râu" phải kéo từ 8.5 đến 9.5).

Các thành phần và ý nghĩa:

+ Hộp (The Box): Đại diện cho 50% dữ liệu trung tâm (từ Q1 đến Q3).

• Cạnh dưới của hộp (Q1 - Tứ phân vị thứ nhất): 25% sinh viên có điểm dưới 5.5. (Q1 = 5.5)

• Đường vạch trong hộp (Q2 - Trung vị): 50% sinh viên có điểm dưới 7.5 và 50% có điểm trên 7.5. (Median = 7.5)

• Cạnh trên của hộp (Q3 - Tứ phân vị thứ ba): 75% sinh viên có điểm dưới 8.5. (Q3 = 8.5)

• Khoảng liên phân vị (IQR = Q3 - Q1): 8.5 - 5.5 = 3. Điều này cho thấy 50% sinh viên ở giữa có điểm số trải rộng trong khoảng 3 điểm.

+ Râu (The Whiskers): Thường kéo dài đến giá trị lớn nhất/nhỏ nhất trong phạm vi 1.5\*IQR từ Q1 và Q3.

• Râu dưới: Kéo dài đến giá trị 4. Điều này có nghĩa là sinh viên có điểm thấp nhất (không phải ngoại lai) là 4.

• Râu trên: Kéo dài đến giá trị 9.5. Điều này có nghĩa là sinh viên có điểm cao nhất (không phải ngoại lai) là 9.5.

+ Điểm ngoại lai (Outliers): Các điểm riêng lẻ nằm ngoài râu.

• Ở đây có một điểm ngoại lai ở 1.5. Đây có thể là một sinh viên có điểm số cực kỳ thấp, có thể do bỏ thi, không học, hoặc một lý do đặc biệt nào đó.

Cách diễn giải:

• Vị trí trung tâm: Trung vị là 7.5, cho thấy điểm số "điển hình" của một sinh viên là 7.5.

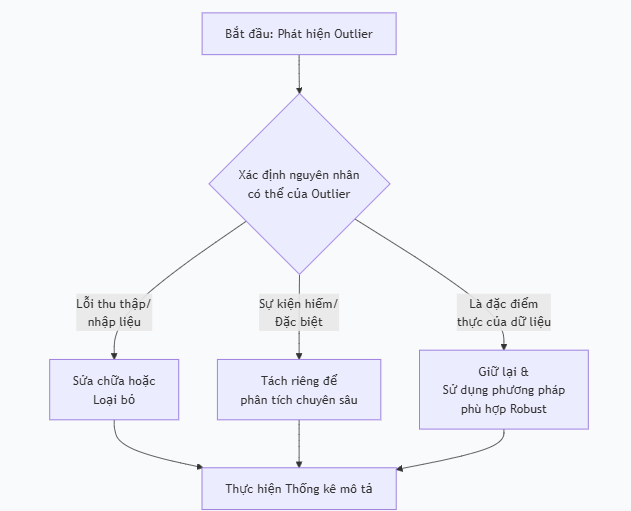
• Độ phân tán: IQR = 3 điểm. Đây là một độ phân tán trung bình.

• Hình dạng: Đường trung vị (7.5) nằm hơi lệch về phía dưới của hộp (gần Q1 hơn Q3). Điều này cho thấy phân phối hơi lệch phải (Positive Skew) - tức là phần dữ liệu bên phải (điểm cao) trải dài hơn một chút so với bên trái.

• Điểm ngoại lai: Có một giá trị ngoại lai cần được điều tra thêm.

-> **Kết luận từ ví dụ:** "Kết quả điểm thi tập trung khá tốt quanh mức 7.5. 50% sinh viên trung tâm có điểm số từ 5.5 đến 8.5. Phổ điểm có xu hướng nhẹ về phía điểm cao, và có một trường hợp điểm rất thấp (1.5) cần xem xét."

* + 1. ***Khi gặp một tập dữ liệu có giá trị ngoại lai (outliers), bạn sẽ xử lý chúng như thế nào trước khi thực hiện thống kê mô tả?***



**Bước 1: PHÁT HIỆN & ĐỊNH LƯỢNG Outlier**

+ Trực quan hóa (Visualization):

• Boxplot: Phương pháp tốt nhất để nhanh chóng nhận diện outlier. Các điểm nằm ngoài râu (whisker) được coi là nghi ngờ.

• Histogram: Quan sát các thanh tách biệt khỏi "thân chính" của phân phối.

• Scatter Plot: Với dữ liệu nhiều chiều, giúp thấy các điểm nằm tách biệt khỏi cụm chính.

+ Định lượng bằng các phương pháp thống kê:

• Quy tắc IQR (Phổ biến & Hiệu quả):

- IQR = Q3 - Q1

- Giới hạn dưới: Q1 - 1.5 \* IQR

- Giới hạn trên: Q3 + 1.5 \* IQR

- Bất kỳ điểm nào nằm ngoài hai giới hạn này đều được xem là outlier.

• Z-score (Với dữ liệu phân phối chuẩn):

- Z = (x - mean) / standard deviation

-Thông thường, nếu |Z| > 2 hoặc |Z| > 3, điểm đó được xem là outlier.

**Bước 2: HIỂU NGUYÊN NHÂN (Bước Quan trọng Nhất!)**

+ Lỗi (Errors):

• Ví dụ: Nhập sai dữ liệu (chiều cao 17m thay vì 1.7m), lỗi cảm biến, lỗi trong quá trình thu thập.

• Cách xử lý: Sửa chữa nếu có thể, nếu không thì LOẠI BỎ.

+ Sự kiện hiếm hoặc đặc biệt (Natural Variance):

• Ví dụ: Một vận động viên chạy 100m trong 9.58s (Usain Bolt) so với người bình thường; một giao dịch gian lận có giá trị cực lớn.

• Cách xử lý: CẨN TRỌNG! Đây có thể là thông tin quý giá. Cân nhắc việc phân tích riêng.

+ Là một phần của quần thể (Genuine Part of Population):

• Ví dụ: Thu nhập của Elon Musk trong tập dữ liệu thu nhập người Mỹ.

• Cách xử lý: GIỮ LẠI. Chúng phản ánh sự thật của dữ liệu.

**Bước 3: LỰA CHỌN PHƯƠNG PHÁP XỬ LÝ**

+ Nhóm 1: Loại bỏ (Deletion)

• Khi nào dùng: Khi outlier là do lỗi và không thể sửa được, hoặc khi chúng không liên quan đến mục tiêu phân tích (ví dụ: phân tích hành vi người dùng bình thường thì loại bỏ các giao dịch hack).

• Ưu điểm: Đơn giản, loại bỏ sự ảnh hưởng của dữ liệu nhiễu.

• Nhược điểm: Làm mất thông tin, có thể làm chệch (bias) mẫu nếu outlier thực sự thuộc về quần thể.

• Lưu ý: LUÔN báo cáo trong nghiên cứu rằng bạn đã loại bỏ bao nhiêu outlier và lý do.

+ Nhóm 2: Chuyển đổi (Transformation)

• Khi nào dùng: Khi dữ liệu bị lệch nặng (skewed) và bạn muốn giảm bớt ảnh hưởng của outlier mà không muốn loại bỏ chúng.

• Phương pháp:

- Log transformation: log(x)

- Square root transformation: sqrt(x)

- Box-Cox transformation: Phương pháp tổng quát hơn.

• Ưu điểm: Giữ lại tất cả các điểm dữ liệu, giúp dữ liệu có phân phối "đẹp" hơn.

• Nhược điểm: Khó giải thích kết quả vì đơn vị đã thay đổi.

+ Nhóm 3: Thay thế (Imputation)

• Khi nào dùng: Khi bạn tin rằng outlier là có thật nhưng không muốn nó ảnh hưởng quá lớn đến các chỉ số tổng hợp.

• Phương pháp:

- Thay bằng giới hạn: Thay thế các outlier bằng giá trị Q1 - 1.5\*IQR (với outlier nhỏ) hoặc Q3 + 1.5\*IQR (với outlier lớn).

- Thay bằng trung vị hoặc trung bình cắt (trimmed mean): Vì các giá trị này ít bị ảnh hưởng bởi outlier hơn số trung bình.

• Ưu điểm: Giữ được cỡ mẫu.

• Nhược điểm: Làm thay đổi phân phối thực của dữ liệu.

+ Nhóm 4: Phân tích riêng biệt (Separate Analysis)

• Khi nào dùng: Với các outlier quan trọng, đại diện cho một nhóm đặc biệt.

• Phương pháp: Thực hiện thống kê mô tả 2 lần: một lần có outlier và một lần không có outlier. Trình bày và so sánh cả hai kết quả.

• Ưu điểm: Không mất thông tin quý giá, cung cấp cái nhìn toàn diện.

• Nhược điểm: Phức tạp hơn.

* 1. **Bài tập thực hành 1**

Thực hiện thống kê mô tả trên tập dữ liệu về phân loại chất lượng rượu đỏ. Dữ liệu lấy tại https://www.kaggle.com/code/eisgandar/red-wine-quality-eda-classification.

* 1. **Bài tập thực hành 2**

Thực hiện thống kê mô tả trên tập dữ liệu về bệnh tiểu đường. Dữ liệu lấy tại https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906.

1. **XỬ LÝ VÀ TRỰC QUAN HÓA DỮ LIỆU**

**2.1. Ôn tập lý thuyết**

***2.1.1. Trực quan hóa dữ liệu có vai trò gì trong phân tích dữ liệu? Tại sao nó quan trọng trong khám phá dữ liệu (EDA)?***

1. *Vai trò của Trực quan hóa dữ liệu trong Phân tích Dữ liệu*

+ Giúp Hiểu dữ liệu Nhanh chóng và Trực quan:

• Bộ não con người xử lý thông tin hình ảnh nhanh hơn ngôn ngữ hoặc chữ số hàng nghìn lần. Một biểu đồ có thể truyền tải thông điệp trong vài giây mà một bảng số liệu dày đặc không thể làm được.

• Ví dụ: Một đường xu hướng (line chart) tăng vọt cho thấy sự tăng trưởng đột biến ngay lập tức, trong khi bạn có thể mất vài phút để nhận ra điều tương tự từ một bảng số.

+ Khám phá các Mẫu hình, Xu hướng và Mối quan hệ:

• Trực quan hóa giúp chúng ta nhìn thấy những thứ mà các phép tính thống kê đơn thuần có thể bỏ sót.

• Ví dụ:

- Xu hướng theo thời gian (Doanh thu tăng vào mùa lễ hội).

- Mối tương quan giữa hai biến (Chiều cao và Cân nặng thường tăng cùng nhau).

- Cụm (Clusters) trong dữ liệu (Phân khúc khách hàng có hành vi mua sắm tương tự nhau).

+ Phát hiện Điểm bất thường và Giá trị Ngoại lai (Outliers):

• Đây là một trong những vai trò quan trọng nhất. Một điểm nằm tách biệt hoàn toàn khỏi xu hướng chính trên biểu đồ scatter plot hoặc boxplot sẽ lập tức lộ diện.

• Ví dụ: Một giao dịch chuyển tiền khổng lồ, bất thường trên biểu đồ có thể chỉ ra gian lận.

+ Truyền đạt Thông tin và Kể chuyện bằng Dữ liệu (Data Storytelling):

• Trực quan hóa biến dữ liệu phức tạp thành một câu chuyện dễ hiểu, thuyết phục cho các bên liên quan, cấp trên hoặc công chúng.

• Ví dụ: Một dashboard với các biểu đồ then chốt giúp CEO đưa ra quyết định kinh doanh nhanh chóng.

+ Kiểm tra Giả thuyết và Xác nhận Kết quả:

• Sau khi thực hiện các phân tích thống kê phức tạp, việc vẽ biểu đồ là cách tốt nhất để kiểm tra xem kết quả có ý nghĩa và phù hợp hay không.

1. *Tại sao Trực quan hóa lại ĐẶC BIỆT QUAN TRỌNG trong Khám phá Dữ liệu (EDA)?*

Trong EDA, trực quan hóa giúp bạn trả lời những câu hỏi then chốt mà không phụ thuộc quá nhiều vào các giả định thống kê phức tạp:

+ Dữ liệu của tôi trông như thế nào? (Hình dạng phân phối)

• Biểu đồ Histogram & Boxplot: Ngay lập tức cho bạn biết dữ liệu phân bố ra sao (chuẩn, lệch, có nhiều đỉnh không?), giá trị điển hình là gì (trung vị), và có outlier không.

• Quyết định tiếp theo: Tôi nên dùng Trung bình hay Trung vị? Có cần xử lý outlier không?

+ Các biến có mối quan hệ với nhau không?

• Biểu đồ Scatter Plot & Heatmap tương quan: Giúp bạn nhìn thấy mối liên hệ (tuyến tính, phi tuyến) giữa hai hoặc nhiều biến.

• Quyết định tiếp theo: Có nên đưa các biến này vào cùng một mô hình hồi quy không?

+ Có sự khác biệt giữa các nhóm không?

• Boxplot nhóm, Biểu đồ cột nhóm: Cho phép so sánh trực quan phân phối của một biến giữa các danh mục khác nhau (ví dụ: thu nhập giữa Nam và Nữ).

• Quyết định tiếp theo: Sự khác biệt này có đáng kể không? Có nên chạy kiểm định thống kê để xác nhận?

+ Dữ liệu có thay đổi theo thời gian?

• Biểu đồ đường (Line chart): Tiết lộ xu hướng, tính thời vụ một cách rõ ràng.

• Quyết định tiếp theo: Có thể dự báo được không? Mô hình nào phù hợp?

+ Có cụm (clusters) tự nhiên nào trong dữ liệu không?

• Scatter Plot với màu sắc: Giúp phát hiện các nhóm đối tượng có đặc điểm tương đồng mà bạn có thể chưa biết trước.

• Quyết định tiếp theo: Có nên áp dụng các thuật toán phân cụm (Clustering) không?

***2.1.2. Các loại biểu đồ phổ biến (như histogram, scatter plot, boxplot, bar chart) được sử dụng trong các trường hợp nào?***

*a. Histogram (Biểu đồ Tần suất)*

+ Dùng để: Hiển thị phân phối của một biến số liệu liên tục.

+ Câu hỏi nó trả lời: "Dữ liệu của tôi phân bố như thế nào? Giá trị nào phổ biến nhất?"

+ Trường hợp sử dụng:

• Khám phá hình dạng phân phối (đối xứng, lệch trái, lệch phải, nhiều đỉnh).

• Xác định khoảng giá trị phổ biến nhất.

• Kiểm tra xem dữ liệu có tuân theo phân phối chuẩn không.

• Phát hiện các khoảng trống trong dữ liệu.

*b. Boxplot (Biểu đồ Hộp)*

+ Dùng để: Tóm tắt phân phối của một biến số liệu và so sánh giữa các nhóm.

+ Câu hỏi nó trả lời: "Phân phối của dữ liệu tập trung ở đâu? Có điểm nào bất thường không? Các nhóm khác nhau thế nào?"

+ Trường hợp sử dụng:

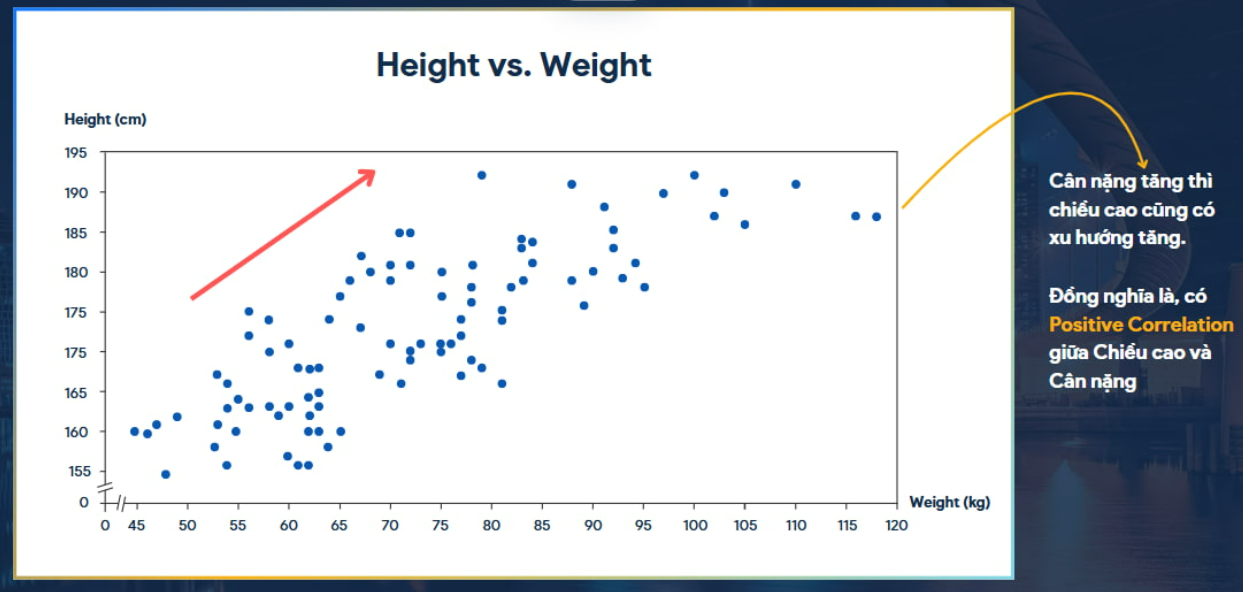
• So sánh phân phối giữa nhiều nhóm (rất hiệu quả).

• Nhận diện nhanh giá trị ngoại lai (outliers).

• Hiểu độ phân tán và xu hướng trung tâm thông qua 5 số tổng hợp.

• Khi không cần chi tiết đầy đủ về hình dạng phân phối như histogram.

1. *Scatter Plot (Biểu đồ Phân tán)*



*Biểu đồ phân tán thể hiện mối liên hệ giữa chiều cao và cân nặng*

+ Dùng để: Hiển thị mối quan hệ giữa hai biến số liệu.

+ Câu hỏi nó trả lời: "Có mối liên hệ nào giữa Biến X và Biến Y không? Mối quan hệ đó mạnh yếu ra sao?"

+ Trường hợp sử dụng:

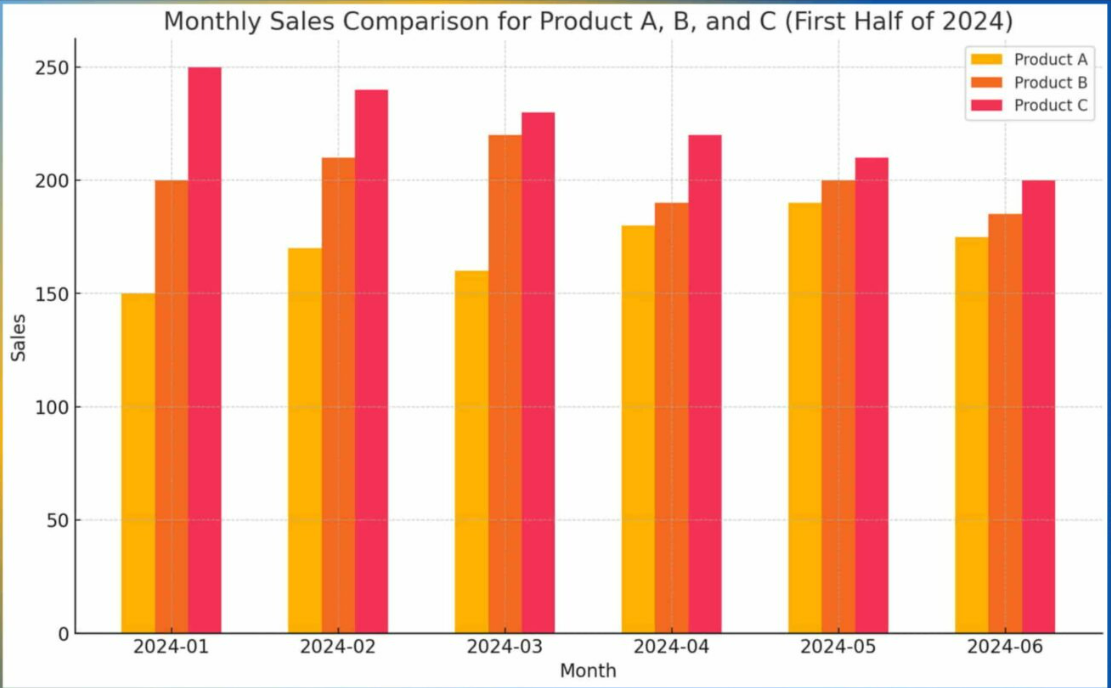
• Phát hiện mối tương quan (tuyến tính, phi tuyến) giữa hai biến.

• Nhận diện cụm (clusters) trong dữ liệu.

• Phát hiện các điểm dữ liệu bất thường (outliers).

• Kiểm tra giả thuyết về mối quan hệ nhân quả.

1. *Bar Chart (Biểu đồ Cột)*



*Biểu đồ cột minh họa doanh số bán hàng của ba loại sản phẩm trong sáu tháng đầu năm*

+ Dùng để: So sánh các danh mục (categories) với nhau.

+ Câu hỏi nó trả lời: "Danh mục nào có giá trị lớn nhất? Sự khác biệt giữa các danh mục là gì?"

+ Trường hợp sử dụng:

• So sánh giá trị giữa các nhóm rời rạc.

• Hiển thị tần suất của các danh mục.

• Thể hiện thứ hạng (xếp hạng sản phẩm bán chạy).

• Khi trục ngang là dữ liệu định tính (tên, danh mục, nhãn).

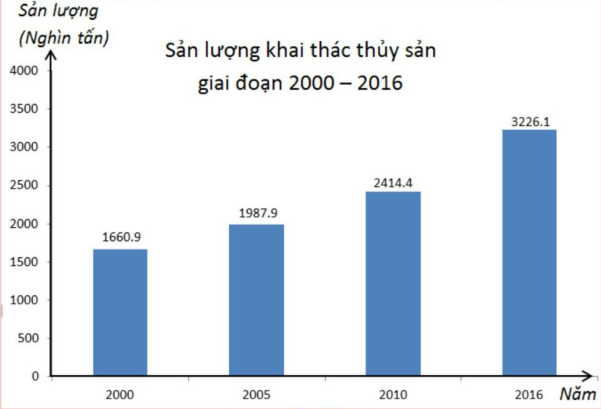
***2.1.3. Làm thế nào để chọn loại biểu đồ phù hợp với đặc điểm của dữ liệu (ví dụ: dữ liệu phân loại, dữ liệu số, dữ liệu thời gian)?***

*a. Dữ liệu phân loại (Biến Định Tính)*

Dữ liệu là các danh mục, nhãn (Ví dụ: Tên sản phẩm, Khu vực, Giới tính)

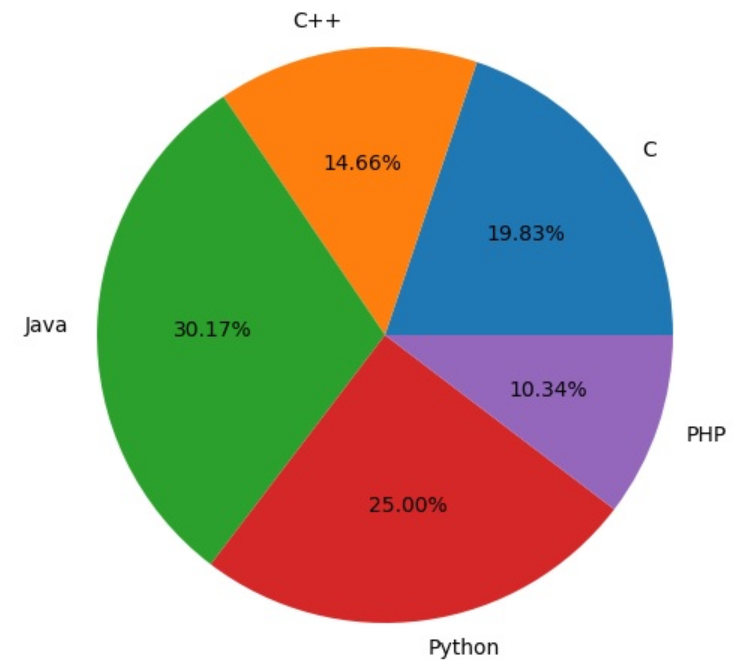
+ Mục tiêu: So sánh các danh mục

• Bar Chart (Biểu đồ cột): tốt nhất để so sánh giá trị giữa các danh mục khác nhau.

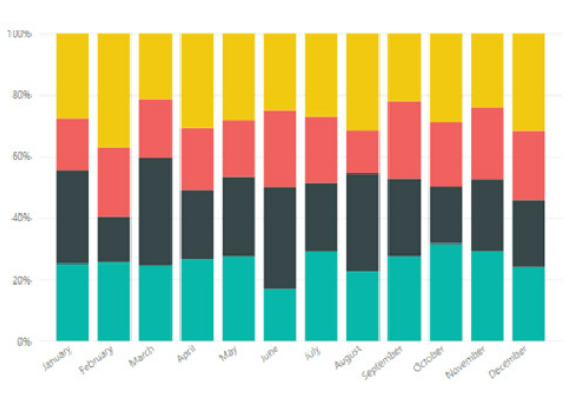


+ Mục tiêu: Hiển thị Thành phần

• Pie Chart (Biểu đồ tròn): Chỉ dùng khi bạn có ít danh mục (≤6) và muốn show phần trăm của mỗi danh mục trong một tổng thể.



• Stacked Bar Chart (Biểu đồ cột chồng): TỐT HƠN biểu đồ tròn khi có nhiều danh mục, cho phép so sánh dễ dàng hơn.



*b. Dữ liệu số (Biến Định Lượng)*

Dữ liệu là con số, có thể đo lường (Ví dụ: Thu nhập, Chiều cao, Điểm số)

+ Mục tiêu: Xem Phân bố

• Histogram: Hoàn hảo để xem hình dạng phân phối của một biến số. Xem dữ liệu tập trung ở đâu, có đối xứng không.

- Ví dụ: Phân bố điểm thi của sinh viên.

• Boxplot: Tuyệt vời để tóm tắt phân phối và nhanh chóng phát hiện điểm ngoại lai (outliers).

- Ví dụ: Phân bố lương giữa các phòng ban, xem có ai lương quá cao/thấp bất thường không.

+ Mục tiêu: Thấy Mối quan hệ giữa 2 biến số

• Scatter Plot (Biểu đồ phân tán): số một cho việc này. Mỗi chấm là một quan sát.

- Ví dụ: Mối quan hệ giữa số giờ học và điểm số.

1. *Dữ liệu thời gian*

Dữ liệu có yếu tố ngày/tháng/năm (Ví dụ: Doanh thu theo tháng, Nhiệt độ theo ngày)

+ Mục tiêu: Xem Xu hướng

• Line Chart (Biểu đồ đường): Lựa chọn tối ưu. Đường kẻ nối các điểm dữ liệu giúp theo dõi sự thay đổi một cách trực quan.

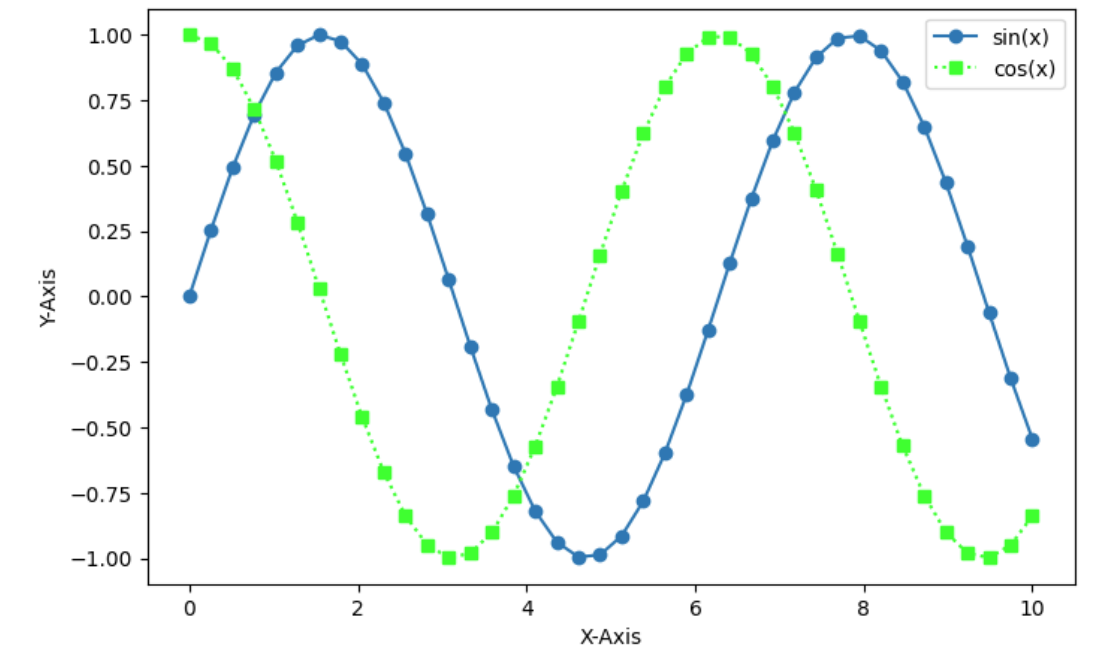


- Ví dụ: Doanh thu hàng tháng trong năm, thể hiện xu hướng tăng/giảm.

***2.1.4. Sự khác biệt giữa các thư viện trực quan hóa trong Python như Matplotlib, Seaborn và Plotly là gì?***

|  |  |  |  |
| --- | --- | --- | --- |
| **Đặc điểm** | **Matplotlib** | **Seaborn** | **Plotly** |
| **Cấp độ** | Cấp thấp (Low-level) | Cấp cao (High-level) | Cấp cao & Tương tác |
| **Độ khó** | Khó hơn | Dễ hơn | Trung bình |
| **Tính năng** | Tùy chỉnh cao | Thống kê đẹp mặc định | Tương tác & động |
| **Output** | Tĩnh | Tĩnh | Động, web-based |

1. *Matplotlib - "Người khổng lồ nền tảng"*



+ Vai trò: Thư viện vẽ đồ thị cơ bản nhất, hầu như mọi thư viện khác đều xây dựng trên nó.

+ Ưu điểm:

• Toàn quyền kiểm soát: Bạn có thể tùy chỉnh mọi chi tiết nhỏ nhất của biểu đồ.

• Ổn định và linh hoạt: Có thể vẽ được hầu hết mọi loại biểu đồ.

• Tiêu chuẩn công nghiệp: Được sử dụng rộng rãi, nhiều tài liệu hướng dẫn.

+ Nhược điểm:

• Cú pháp phức tạp: Cần nhiều dòng code để có biểu đồ đẹp.

• Mặc định không đẹp: Biểu đồ mặc định trông khá cơ bản.

• Không tương tác: Biểu đồ tĩnh.

+ Khi nào dùng:

*import matplotlib.pyplot as plt*

*# Cần nhiều code để tùy chỉnh*

*fig, ax = plt.subplots(figsize=(10, 6))*

*ax.plot([1, 2, 3, 4], [1, 4, 2, 3])*

*ax.set\_xlabel('X Label')*

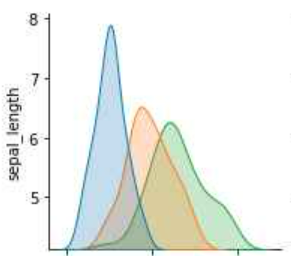
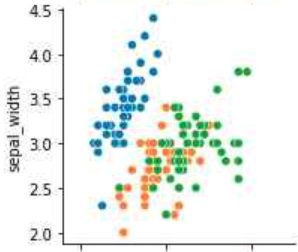
*ax.set\_ylabel('Y Label')*

*ax.set\_title('Basic Matplotlib Plot')*

*plt.show()*

+ Phù hợp cho: Biểu đồ khoa học, báo cáo học thuật, khi cần tùy chỉnh cực kỳ chi tiết.

1. *Seaborn - "Nhà thiết kế thống kê"*

+ Vai trò: Thư viện cấp cao xây dựng trên Matplotlib, tập trung vào trực quan hóa dữ liệu thống kê.

+ Ưu điểm:

• Đẹp mặc định: Giao diện hiện đại, màu sắc hài hòa.

• Cú pháp đơn giản: Rất ít code để có biểu đồ chất lượng.

• Tích hợp với Pandas: Làm việc tốt với DataFrames.

• Hỗ trợ thống kê: Nhiều biểu đồ thống kê chuyên biệt.

+ Nhược điểm:

• Ít tùy chỉnh hơn Matplotlib: Khó thay đổi những chi tiết rất nhỏ.

• Vẫn là biểu đồ tĩnh.

+ Khi nào dùng:

*import seaborn as sns*

*import pandas as pd*

*# Rất ít code cho biểu đồ phức tạp*

*tips = sns.load\_dataset("tips")*

*sns.boxplot(x='day', y='total\_bill', hue='sex', data=tips)*

*plt.show()*

+ Phù hợp cho: Phân tích khám phá (EDA), báo cáo kinh doanh, khi muốn biểu đồ đẹp nhanh chóng.

1. *Plotly - "Nghệ sĩ tương tác"*

+ Vai trò: Thư viện tương tác, hoàn hảo cho web và dashboard.

+ Ưu điểm:

• Tương tác mạnh mẽ: Zoom, pan, hover để xem giá trị, click để ẩn/hiện chuỗi.

• Biểu đồ động: Có thể cập nhật dữ liệu real-time.

• Output web-based: Dễ dàng nhúng vào webpage hoặc dashboard.

• Thư viện phong phú: Nhiều biểu đồ hiện đại như sunburst, treemap.

+ Nhược điểm:

• Nặng: File output lớn hơn, có thể chậm với dataset rất lớn.

• Khó tùy chỉnh sâu: Cú pháp tùy chỉnh khác với Matplotlib.

+ Khi nào dùng:

*import plotly.express as px*

*# Biểu đồ tương tác với vài dòng code*

*df = px.data.iris()*

*fig = px.scatter(df, x='sepal\_width', y='sepal\_length',*

*color='species', size='petal\_length',*

*hover\_data=['petal\_width'])*

*fig.show()*

+ Phù hợp cho: Dashboard, báo cáo tương tác, ứng dụng web, presentation.

***2.1.5. Những nguyên tắc thiết kế nào cần tuân thủ để tạo ra một biểu đồ trực quan hóa dễ hiểu và hiệu quả?***

+ Nguyên Tắc "Giảm Thiểu Gánh Nặng Nhận Thức" (Reduce Cognitive Load)

Bộ não con người có giới hạn trong xử lý thông tin. Hãy giúp nó làm việc hiệu quả.

• Loại bỏ "Chartjunk" (Rác đồ thị):

- TRÁNH: Nền màu sặc sỡ, đường kẻ grid dày đặc, hiệu ứng 3D không cần thiết, hình nền phức tạp.

- NÊN LÀM: Nền trắng hoặc trung tính, đường grid mờ (hoặc bỏ nếu không cần), tập trung vào dữ liệu.

- Ví dụ: Một biểu đồ cột 3D với bóng đổ thường khó đọc hơn biểu đồ 2D đơn giản.

• Sử dụng Màu Sắc Có Chủ Đích:

- Dùng màu sắc để truyền tải thông tin, không chỉ để trang trí.

- Chọn palette màu dễ phân biệt, thân thiện với người mù màu.

- Với dữ liệu định tính (categories), dùng các màu sắc khác biệt.

- Với dữ liệu định lượng (sequential), dùng các sắc thái của cùng một màu (càng đậm = giá trị càng cao).

+ Nguyên Tắc "Tập Trung Vào Thông Điệp" (Focus on the Insight)

Mọi yếu tố trên biểu đồ phải phục vụ cho việc kể câu chuyện từ dữ liệu.

• Tiêu đề Rõ Ràng và Giàu Thông Tin:

- Thay vì "Doanh số theo tháng", hãy dùng "Doanh số Tăng 25% trong Tháng 12 nhờ Chiến dịch Lễ Hội".

- Tiêu đề nên trả lời được câu hỏi "So what?" (Vậy thì sao?).

• Làm nổi bật Điểm Quan Trọng:

- Sử dụng chú thích (annotation) để giải thích các điểm dữ liệu quan trọng, đỉnh, hoặc điểm bất thường.

- Dùng màu sắc tương phản để làm nổi bật một nhóm hoặc một xu hướng cụ thể.

- Ví dụ: Trong biểu đồ đường về doanh số, tô đậm đường của năm hiện tại và làm mờ các năm trước để dễ so sánh.

+ Nguyên Tắc "Trung Thực và Rõ Ràng" (Honesty and Clarity)

Đừng vô tình (hoặc cố ý) đánh lừa người xem.

• Luôn bắt đầu trục tung (Y-axis) từ 0 với biểu đồ cột (bar chart). Việc cắt ngắn trục Y có thể phóng đại sự khác biệt một cách giả tạo.

Biểu đồ A (trục Y bắt đầu từ 0) cho thấy sự khác biệt nhỏ. Biểu đồ B (trục Y bắt đầu từ 40) phóng đại sự khác biệt, gây hiểu lầm.

• Tỷ lệ khung hình phù hợp: Một biểu đồ quá cao hoặc quá dẹt có thể làm biến dạng nhận thức về xu hướng. Tỷ lệ "golden ratio" (~1.618) thường được khuyến khích.

+ Nguyên Tắc "Tổ Chức và Cấu trúc Hợp Lý" (Organization and Structure)

Giúp người xem dễ dàng định hướng và hiểu biểu đồ.

• Sắp xếp dữ liệu một cách logic:

- Với dữ liệu định tính, sắp xếp theo thứ tự bảng chữ cái, thứ tự quy ước, hoặc theo giá trị (từ cao xuống thấp/ngược lại) để dễ so sánh.

- Ví dụ: Sắp xếp các cột trong bar chart theo giá trị giảm dần (trừ khi thứ tự thời gian hoặc danh mục cố định là quan trọng).

• Nhãn rõ ràng và dễ đọc:

- Ghi rõ tên trục, đơn vị tính.

- Xoay nhãn trục ngang nếu chúng quá dài.

- Sử dụng chú thích (legend) khi cần, đặt ở vị trí dễ thấy.

+ Nguyên Tắc "Chọn Đúng Biểu Đồ Cho Đúng Dữ Liệu" (Right Chart for Right Data)

Đây là nền tảng của mọi sự trực quan hóa hiệu quả.

• So sánh giữa các danh mục: Dùng Bar Chart.

• Thể hiện thành phần của một tổng thể: Dùng Pie Chart (chỉ với ít phần) hoặc Stacked Bar Chart.

• Hiển thị xu hướng theo thời gian: Dùng Line Chart.

• Cho thấy mối quan hệ giữa 2 biến: Dùng Scatter Plot.

• Hiển thị phân phối của dữ liệu: Dùng Histogram hoặc Box Plot.

***2.1.6. Làm thế nào để tạo một biểu đồ đơn giản như histogram hoặc bar chart bằng Matplotlib? Bạn có thể chia sẻ đoạn code mẫu không?***

+ Biểu đồ Histogram (Phân phối dữ liệu)

*import matplotlib.pyplot as plt*

*import numpy as np*

*# Tạo dữ liệu mẫu - điểm thi của sinh viên*

*np.random.seed(42) # Để kết quả có thể tái lập*

*diem\_thi = np.random.normal(7.0, 1.5, 200) # 200 điểm, trung bình=7.0, độ lệch=1.5*

*diem\_thi = np.clip(diem\_thi, 0, 10) # Giới hạn điểm từ 0-10*

*# Tạo histogram*

*plt.figure(figsize=(10, 6)) # Kích thước biểu đồ*

*# Vẽ histogram*

*plt.hist(diem\_thi, bins=15, color='skyblue', edgecolor='black', alpha=0.7)*

*# Thiết lập tiêu đề và nhãn*

*plt.title('PHÂN BỐ ĐIỂM THI CỦA SINH VIÊN', fontsize=14, fontweight='bold', pad=20)*

*plt.xlabel('Điểm Số', fontsize=12, fontweight='bold')*

*plt.ylabel('Số Sinh Viên', fontsize=12, fontweight='bold')*

*# Thêm grid và customizations*

*plt.grid(axis='y', alpha=0.3, linestyle='--')*

*plt.xticks(np.arange(0, 11, 1)) # Hiển thị tất cả các điểm từ 0-10*

*# Hiển thị biểu đồ*

*plt.tight\_layout()*

*plt.show()*

\* Giải thích các tham số quan trọng:

• bins=15: Số lượng cột (khoảng giá trị)

• color='skyblue': Màu của các cột

• edgecolor='black': Màu viền của các cột

• alpha=0.7: Độ trong suốt (0-1)

+ Biểu đồ Bar Chart (So sánh danh mục)

*import matplotlib.pyplot as plt*

*# Dữ liệu mẫu - doanh số theo sản phẩm*

*san\_pham = ['iPhone', 'Samsung', 'Xiaomi', 'Oppo', 'Nokia']*

*doanh\_so = [150, 120, 80, 60, 25] # Đơn vị: nghìn sản phẩm*

*# Tạo bar chart*

*plt.figure(figsize=(10, 6))*

*# Vẽ bar chart với màu sắc khác nhau*

*colors = ['#FF6B6B', '#4ECDC4', '#45B7D1', '#96CEB4', '#FFEAA7']*

*bars = plt.bar(san\_pham, doanh\_so, color=colors, edgecolor='black', linewidth=0.5)*

*# Thiết lập tiêu đề và nhãn*

*plt.title('DOANH SỐ BÁN HÀNG THEO SẢN PHẨM', fontsize=14, fontweight='bold', pad=20)*

*plt.xlabel('Tên Sản Phẩm', fontsize=12, fontweight='bold')*

*plt.ylabel('Doanh Số (nghìn sản phẩm)', fontsize=12, fontweight='bold')*

*# Thêm giá trị trên mỗi cột*

*for bar, value in zip(bars, doanh\_so):*

*plt.text(bar.get\_x() + bar.get\_width()/2, bar.get\_height() + 1,*

*f'{value}', ha='center', va='bottom', fontweight='bold')*

*# Tùy chỉnh trục y và thêm grid*

*plt.ylim(0, max(doanh\_so) \* 1.15) # Tạo khoảng trống phía trên cho text*

*plt.grid(axis='y', alpha=0.3, linestyle='--')*

*# Xoay nhãn trục x nếu cần*

*plt.xticks(rotation=45)*

*# Hiển thị biểu đồ*

*plt.tight\_layout()*

*plt.show()*

\* Giải thích các tham số quan trọng:

• edgecolor='black': Viền của các cột

• linewidth=0.5: Độ dày viền

• rotation=45: Xoay nhãn trục x 45 độ để dễ đọc

***2.1.7. Làm thế nào để xuất biểu đồ từ Python ra các định dạng như PNG, PDF hoặc HTML để sử dụng trong báo cáo?***

+ Phương Pháp Cơ Bản: plt.savefig()

Đây là phương pháp phổ biến nhất, hoạt động với mọi biểu đồ Matplotlib.

+ Các Tham Số Quan Trọng Khi Xuất File

bbox\_inches='tight' - THAM SỐ QUAN TRỌNG NHẤT

plt.savefig('my\_plot.png', bbox\_inches='tight')

Tác dụng: Tự động cắt bỏ khoảng trắng thừa xung quanh biểu đồ, làm cho hình ảnh chuyên nghiệp hơn.

dpi - Độ phân giải (Dots Per Inch)

plt.savefig('my\_plot.png', dpi=300, bbox\_inches='tight')

• dpi=72: Chất lượng thấp, phù hợp cho web

• dpi=150: Chất lượng trung bình, đủ cho in ấn thông thường

• dpi=300: Chất lượng cao, cho in ấn chuyên nghiệp, báo cáo

• dpi=600: Chất lượng rất cao, cho ấn phẩm khoa học

transparent=True - Nền trong suốt

plt.savefig('my\_plot.png', transparent=True, bbox\_inches='tight')

Rất hữu ích khi muốn đặt biểu đồ lên nền màu khác trong báo cáo.

+ Xuất Biểu đồ Tương Tác Plotly ra HTML

+ Xuất Nhiều Biểu đồ Cùng Lúc

+ Xuất Biểu đồ Từ Pandas DataFrame

+ Tự Động Hóa Việc Xuất Biểu đồ

+ Xuất Biểu đồ Chất Lượng Cao Cho Báo Cáo Khoa Học

Bảng So Sánh Định Dạng File

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Định Dạng** | **Chất Lượng** | **Kích Thước** | **Ưu Điểm** | **Nhược Điểm** |
| **PNG** | Cao | Trung bình | Hỗ trợ trong suốt, tốt cho web | Không vector |
| **PDF** | Rất cao | Nhỏ (vector) | Vector, in ấn sắc nét | Khó chỉnh sửa |
| **SVG** | Rất cao | Nhỏ (vector) | Vector, có thể chỉnh sửa | Cần hỗ trợ SVG |
| **JPG** | Thấp-Trung | Nhỏ | Phù hợp ảnh, nhỏ | Nén lossy, không trong suốt |
| **HTML** | - | Lớn | Tương tác (Plotly) |  |

**2.2. Bài tập thực hành 1**

Thực hiện trực quan hóa dữ liệu trên tập dữ liệu về phân loại chất lượng rượu đỏ. Dữ liệu lấy tại https://www.kaggle.com/code/eisgandar/red-wine-quality-eda-classification.

**2.3. Bài tập thực hành 2**

Thực hiện trực quan hóa dữ liệu trên tập dữ liệu về bệnh tiểu đường. Dữ liệu lấy tại https://www.kaggle.com/code/vincentlugat/pima-indians-diabetes-eda-prediction-0-906.

Thực hiện EDA trên tập dữ liệu mua sắm tại siêu thị. Tập dữ liệu lấy từ

https://www.kaggle.com/code/rajatkumar30/eda-online-retail.

1. **PHÂN TÍCH ĐƠN BIẾN VÀ HAI BIẾN**

**3.1. Ôn tập lý thuyết**

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

*a. Phân tích đơn biến (univariate analysis) là gì?*

\* Định nghĩa: Phân tích từng biến một cách riêng lẻ. Nó tập trung vào việc mô tả và tóm tắt đặc điểm của một biến duy nhất.

\* Mục tiêu: Trả lời câu hỏi "Mỗi biến riêng lẻ trông như thế nào?"

\* Các kỹ thuật chính:

+ Với Biến Định lượng (Numerical)

• Thống kê mô tả:

- Xu hướng trung tâm: Mean, Median, Mode

- Độ phân tán: Range, Variance, Standard Deviation, IQR

• Trực quan hóa:

- Histogram: Xem phân phối, hình dạng (đối xứng, lệch)

- Boxplot: Tóm tắt phân phối, phát hiện outliers

- KDE Plot: Ước lượng mật độ phân phối

+ Với Biến Phân loại (Categorical)

• Thống kê mô tả:

- Tần số (Frequency)

- Tỷ lệ phần trăm (Percentage)

• Trực quan hóa:

- Bar Chart: So sánh tần số các hạng mục

- Pie Chart: Hiển thị tỷ lệ phần trăm

*b. Phân tích Hai biến (Bivariate Analysis)*

\* Định nghĩa: Phân tích mối quan hệ giữa hai biến. Nó tìm kiếm các mẫu hình, xu hướng hoặc sự liên kết giữa hai biến.

\* Mục tiêu: Trả lời câu hỏi "Hai biến có mối quan hệ với nhau không?"

\* Các kỹ thuật chính:

+ Hai biến Định lượng (Numerical vs Numerical)

• Thống kê:

- Hệ số tương quan (Correlation): Pearson, Spearman

- Hồi quy đơn giản (Simple Regression)

• Trực quan hóa:

- Scatter Plot: Hiển thị mối quan hệ trực quan

- Heatmap Correlation: Ma trận tương quan

+ Một Định lượng & Một Phân loại (Numerical vs Categorical)

• Thống kê:

- So sánh mean giữa các nhóm (t-test, ANOVA)

- Phân phối theo nhóm

• Trực quan hóa:

- Boxplot nhóm: So sánh phân phối giữa các nhóm

- Bar Chart nhóm: So sánh giá trị trung bình

- Violin Plot: Kết hợp phân phối và thống kê

+ Hai biến Phân loại (Categorical vs Categorical)

• Thống kê:

- Bảng chéo (Crosstab)

- Kiểm định Chi-square

• Trực quan hóa:

- Stacked Bar Chart

- Heatmap của bảng chéo

- Mosaic Plot

*c. Sự khác biệt cốt lỗi*

|  |  |  |
| --- | --- | --- |
| **Đặc Điểm** | **Phân tích Đơn biến** | **Phân tích Hai biến** |
| **Số lượng biến** | 1 biến | 2 biến |
| **Mục tiêu** | Mô tả, tóm tắt từng biến | Tìm mối quan hệ giữa 2 biến |
| **Câu hỏi** | "Biến này trông như thế nào?" | "Hai biến có liên quan không?" |
| **Kỹ thuật** | Mean, Median, Histogram, Bar chart | Correlation, T-test, Scatter plot, Boxplot nhóm |
| **Kết quả** | Phân phối, Xu hướng trung tâm, Outliers | Mối quan hệ, Tương quan, Sự khác biệt nhóm |
| **Ví dụ** | Phân phối tuổi của bệnh nhân | Mối quan hệ giữa tuổi và huyết áp |

***3.1.2. 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)?***

*a. Đo lường Xu hướng Trung tâm (Measures of Central Tendency)*

+ Số Trung bình (Mean)

• Là gì: Trung bình cộng của tất cả các giá trị

• Công thức: Mean = (Tổng tất cả giá trị) / (Số lượng giá trị)

• Khi nào dùng: Dữ liệu phân phối đối xứng, không có outlier

• Ưu điểm: Sử dụng tất cả thông tin trong dữ liệu

• Nhược điểm: Rất nhạy cảm với giá trị ngoại lai (outliers)

• Ví dụ: Điểm trung bình môn học: (8 + 7 + 9 + 6 + 10) / 5 = 8.0

+ Trung vị (Median)

• Là gì: Giá trị nằm chính giữa khi sắp xếp dữ liệu theo thứ tự

• Cách tính: Sắp xếp dữ liệu → chọn giá trị ở vị trí chính giữa

• Khi nào dùng: Dữ liệu có outlier hoặc phân phối lệch

• Ưu điểm: Không bị ảnh hưởng bởi outliers

• Nhược điểm: Bỏ qua thông tin từ các giá trị khác

• Ví dụ: [3, 5, 7, 9, 11] → Trung vị = 7

+ Yếu vị (Mode)

• Là gì: Giá trị xuất hiện nhiều lần nhất trong tập dữ liệu

• Khi nào dùng: Dữ liệu phân loại (categorical) hoặc khi cần biết giá trị phổ biến nhất

• Ưu điểm: Có thể dùng cho cả dữ liệu định tính và định lượng

• Nhược điểm: Có thể có nhiều mode hoặc không có mode

• Ví dụ: [Xanh, Đỏ, Xanh, Vàng, Xanh] → Mode = Xanh

*b. Đo lường Độ phân tán (Measures of Dispersion)*

+ Khoảng biến thiên (Range)

• Là gì: Chênh lệch giữa giá trị lớn nhất và nhỏ nhất

• Công thức: Range = Max - Min

• Ưu điểm: Dễ tính, dễ hiểu

• Nhược điểm: Rất nhạy cảm với outliers

• Ví dụ: [10, 15, 20, 25, 30] → Range = 30 - 10 = 20

+ Phương sai (Variance)

• Là gì: Trung bình của bình phương các độ lệch so với giá trị trung bình

• Công thức: Variance = Σ(xi - mean)² / (n-1) (mẫu)

• Ý nghĩa: Đo lường mức độ biến thiên tổng thể

• Nhược điểm: Đơn vị là bình phương của đơn vị gốc, khó diễn giải

+ Độ lệch chuẩn (Standard Deviation)

• Là gì: Căn bậc hai của phương sai

• Công thức: SD = √Variance

• Ưu điểm: Cùng đơn vị với dữ liệu gốc, dễ diễn giải

• Ý nghĩa thực tế:

- SD nhỏ: Dữ liệu tập trung quanh mean

- SD lớn: Dữ liệu phân tán rộng

• Ví dụ: Trong tài chính, SD của lợi nhuận cổ phiếu chính là độ rủi ro

+ Khoảng tứ phân vị (IQR - Interquartile Range)

• Là gì: Chênh lệch giữa tứ phân vị thứ ba (Q3) và thứ nhất (Q1)

• Công thức: IQR = Q3 - Q1

• Ý nghĩa: Độ phân tán của 50% dữ liệu trung tâm

• Ưu điểm: Không bị ảnh hưởng bởi outliers

• Ứng dụng: Xác định outliers (giá trị < Q1 - 1.5×IQR hoặc > Q3 + 1.5×IQR)

*c. Đo lường Hình dạng Phân phối (Measures of Shape)*

+ Độ lệch (Skewness)

• Là gì: Đo lường mức độ đối xứng của phân phối

• Giá trị và ý nghĩa:

- Skewness ≈ 0: Phân phối đối xứng

- Skewness > 0: Lệch phải (đuôi dài bên phải)

- Skewness < 0: Lệch trái (đuôi dài bên trái)

• Ví dụ thực tế:

- Thu nhập: Thường lệch phải (số ít người có thu nhập rất cao)

- Điểm thi dễ: Có thể lệch trái (nhiều điểm cao)

+ Độ nhọn (Kurtosis)

• Là gì: Đo lường "độ béo" của đuôi phân phối so với phân phối chuẩn

• Giá trị và ý nghĩa:

- Kurtosis ≈ 0: Độ nhọn tương tự phân phối chuẩn

- Kurtosis > 0: Đuôi "béo" hơn, nhiều outliers hơn

- Kurtosis < 0: Đuôi "mỏng" hơn, ít outliers hơn

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

PHẦN 1: XÁC ĐỊNH MỐI QUAN HỆ - Các Bước Thực Tiễn

Bước 1: Trực Quan Hóa Trước Tiên

Bước 2: Định Lượng Mối Quan Hệ

PHẦN 2: HIỆU SỐ TƯƠNG QUAN với NHÂN QUẢ

Đây là sự phân biệt QUAN TRỌNG NHẤT:

+ Tương Quan (Correlation)

• Là gì: Hai biến thay đổi cùng nhau

• Câu hỏi: "X và Y có di chuyển cùng nhau không?"

• Ví dụ:

- Số lượng kem bán ↔ Số người chết đuối

- Doanh số áo mưa ↔ Lượng mưa

+ Nhân Quả (Causation)

• Là gì: Thay đổi trong X GÂY RA thay đổi trong Y

• Câu hỏi: "X có làm thay đổi Y không?"

• Ví dụ:

- Hút thuốc → Ung thư phổi

- Tập thể dục → Giảm cân

PHẦN 3: CÁC TÌNH HUỐNG QUAN HỆ PHỔ BIẾN

+ Quan hệ Nhân quả Trực tiếp

+ Quan hệ Ngược chiều Nhân quả

+ Quan hệ Giả tạo (Spurious)

+ Quan hệ Tình cờ

PHẦN 4: TIÊU CHUẨN ĐỂ XÁC ĐỊNH NHÂN QUẢ

Để chứng minh X gây ra Y, cần thỏa mãn:

+ Mối liên hệ Mạnh mẽ

• Tương quan cao và nhất quán

• Hiệu ứng lớn, có ý nghĩa thống kê

+ Tính Nhất quán

• Quan hệ được quan sát trong nhiều nghiên cứu

• Tái lập được trong các bối cảnh khác nhau

+ Tính Đặc hiệu

• X chỉ ảnh hưởng đến Y, không ảnh hưởng đến biến khác

• Hiệu ứng cụ thể, rõ ràng

+ Tính Thời gian

• Nguyên nhân phải xảy ra TRƯỚC kết quả

• Đây là tiêu chí QUAN TRỌNG NHẤT

+ Tính Liên tục

• Liều lượng cao hơn → Hiệu ứng mạnh hơn

• Quan hệ liều lượng - đáp ứng

+ Tính Hợp lý

• Có cơ chế sinh học/khoa học giải thích

• Phù hợp với kiến thức hiện có

+ Tính Thử nghiệm

• Có thể kiểm chứng bằng thí nghiệm

• Thay đổi X → Quan sát thay đổi trong Y

PHẦN 5: VÍ DỤ THỰC TẾ - Phân tích Mối quan hệ

Tình huống: "Số giờ học có làm tăng điểm thi không?"

# Bước 1: Trực quan hóa

*plt.scatter(df['giờ\_học'], df['điểm\_thi'])*

*plt.show()*

# Bước 2: Định lượng tương quan

*r, p = stats.pearsonr(df['giờ\_học'], df['điểm\_thi'])*

*print(f"Tương quan: r = {r:.3f}, p = {p:.4f}")*

# Bước 3: Phân tích sâu

# Kiểm soát yếu tố nhiễu (trình độ, động lực, chất lượng học)

\* Đánh giá theo tiêu chí nhân quả:

+ Liên hệ mạnh: r = 0.65 (tương quan trung bình-mạnh)

+ Nhất quán: Nhiều nghiên cứu khác cũng cho kết quả tương tự

+ Đặc hiệu: Giờ học cũng ảnh hưởng đến nhiều yếu tố khác

+ Thời gian: Giờ học xảy ra trước khi thi

+ Liên tục: Học càng nhiều → Điểm càng cao (quan hệ liều-đáp ứng)

+ Hợp lý: Có cơ chế rõ ràng (học nhiều → hiểu bài → điểm cao)

+ Thử nghiệm: Khó thực hiện thí nghiệm ngẫu nhiên

-> Kết luận: Có bằng chứng mạnh về quan hệ nhân quả

PHẦN 6: CẢNH BÁO THƯỜNG GẶP

+ Biến nhiễu (Confounding Variables)

+ Thiên lệch Chọn mẫu

• Mẫu không đại diện cho tổng thể

• Giải pháp: Kiểm tra tính đại diện của mẫu

+ Hiệu ứng Ngưỡng

• Quan hệ chỉ xảy ra trong một khoảng giá trị nhất định

• Giải pháp: Phân tích phi tuyến tính

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

*a. Hiệp biến (covariance)*

Định nghĩa: Covariance đo lường hướng của mối quan hệ tuyến tính giữa hai biến.

Công thức:

Cov(X,Y) = Σ[(Xᵢ - X̄)(Yᵢ - Ȳ)] / (n-1)

Ý nghĩa giá trị:

• Cov(X,Y) > 0: Hai biến có xu hướng thay đổi cùng chiều

• Cov(X,Y) < 0: Hai biến có xu hướng thay đổi ngược chiều

• Cov(X,Y) = 0: Không có mối quan hệ tuyến tính

*b. Tương quan (Correlation)*

Định nghĩa: Correlation đo lường cả hướng và độ mạnh của mối quan hệ tuyến tính giữa hai biến.

Công thức (Pearson):

Corr(X,Y) = Cov(X,Y) / (σₓ × σᵧ)

Ý nghĩa giá trị:

• -1 ≤ r ≤ 1 (luôn nằm trong khoảng này)

• r > 0: Tương quan thuận (cùng chiều)

• r < 0: Tương quan nghịch (ngược chiều)

• |r| ≈ 1: Tương quan mạnh

• |r| ≈ 0: Tương quan yếu hoặc không có

**SO SÁNH CHI TIẾT**

|  |  |  |
| --- | --- | --- |
| **Đặc điểm** | **Hiệp phương sai (Covariance)** | **Tương quan (Correlation)** |
| **Định nghĩa** | Đo lường hướng của quan hệ | Đo lường cả hướng và độ mạnh |
| **Phạm vi giá trị** | -∞ đến +∞ | -1 đến +1 |
| **Đơn vị** | Phụ thuộc vào đơn vị của X và Y | Không có đơn vị (chuẩn hóa) |
| **Ảnh hưởng bởi scale** | CÓ - thay đổi khi đổi đơn vị | KHÔNG - không đổi khi đổi đơn vị |
| **Khả năng so sánh** | Khó so sánh giữa các cặp biến khác nhau | Dễ dàng so sánh giữa các cặp biến |
| **Ứng dụng** | Tính toán trong các mô hình thống kê | Diễn giải và so sánh mối quan hệ |

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

Tổng quan - Khi nào dùng cái gì?

|  |  |  |  |
| --- | --- | --- | --- |
| **Phân loại** | **Mục tiêu** | **Biểu đồ chính** | **Khi nào sử dụng** |
| **Đơn biến** | Hiểu MỘT biến | Histogram, Boxplot, Bar Chart | Khi cần mô tả đặc điểm của từng biến riêng lẻ |
| **Hai biến** | Hiểu MỐI QUAN HỆ | Scatter plot, Grouped Boxplot, Heatmap | Khi cần phân tích tương tác giữa 2 biến |

1. *PHÂN TÍCH ĐƠN BIẾN - Khi nào sử dụng?*

Trường hợp cụ thể:

+ Khám phá phân phối dữ liệu

+ Phát hiện giá trị ngoại lai

+ Mô tả đặc điểm cơ bản

+ Kiểm tra giả định phân phối

1. *PHÂN TÍCH HAI BIẾN - Khi nào sử dụng?*

Trường hợp cụ thể:

+ Tìm mối tương quan giữa 2 biến số

+ So sánh các nhóm

+ Phân tích xu hướng theo thời gian

+ Kiểm tra sự khác biệt giữa các danh mục

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

*a. Scatter Plot*

*import matplotlib.pyplot as plt*

*import numpy as np*

*import pandas as pd*

*# Tạo dữ liệu mẫu*

*np.random.seed(42)*

*n = 100*

*hours\_studied = np.random.normal(20, 5, n) # Số giờ học*

*exam\_scores = 50 + 2 \* hours\_studied + np.random.normal(0, 10, n) # Điểm thi*

*# Scatter plot cơ bản*

*plt.figure(figsize=(10, 6))*

*plt.scatter(hours\_studied, exam\_scores, alpha=0.6, color='blue')*

*plt.xlabel('Số Giờ Học', fontsize=12, fontweight='bold')*

*plt.ylabel('Điểm Thi', fontsize=12, fontweight='bold')*

*plt.title('Mối Quan Hệ Giữa Số Giờ Học và Điểm Thi', fontsize=14, fontweight='bold')*

*plt.grid(True, alpha=0.3)*

*plt.show()*

*b. Heatmap*

*# Tạo dữ liệu nhiều biến*

*np.random.seed(42)*

*data = pd.DataFrame({*

*'hours\_studied': hours\_studied,*

*'exam\_scores': exam\_scores,*

*'attendance': np.random.normal(85, 10, n), # Tỷ lệ đi học*

*'homework\_score': np.random.normal(75, 15, n), # Điểm bài tập*

*'previous\_gpa': np.random.normal(7.5, 1, n) # Điểm trung bình trước*

*})*

*# Tính ma trận tương quan*

*correlation\_matrix = data.corr()*

*# Vẽ heatmap*

*plt.figure(figsize=(10, 8))*

*sns.heatmap(correlation\_matrix,*

*annot=True, # Hiển thị giá trị*

*cmap='coolwarm', # Màu sắc*

*center=0, # Tâm màu ở 0*

*square=True, Ô vuông*

*fmt='.3f', # Định dạng số*

*cbar\_kws={'shrink': 0.8})*

*plt.title('MA TRẬN TƯƠNG QUAN GIỮA CÁC BIẾN',*

*fontsize=16, fontweight='bold', pad=20)*

*plt.xticks(rotation=45)*

*plt.yticks(rotation=0)*

*plt.tight\_layout()*

*plt.show()*

1. *Diễn giải*

*+ Scatter Plot: Nhận diện mối quan hệ trực quan*

• *Điểm phân tán thành cụm → Quan hệ mạnh*

• *Điểm phân tán ngẫu nhiên → Không có quan hệ*

• *Xu hướng tăng/giảm rõ ràng → Tương quan dương/âm*

*+ Heatmap: So sánh nhiều mối quan hệ cùng lúc*

• *Màu đỏ đậm → Tương quan dương mạnh*

• *Màu xanh đậm → Tương quan âm mạnh*

• *Màu nhạt → Tương quan yếu*

***3.1.7. 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?***

*a. Boxplot Cơ Bản & Nâng Cao*

+ Boxplot Cơ Bản với Matplotlib

*import matplotlib.pyplot as plt*

*import pandas as pd*

*import numpy as np*

*# Tạo dữ liệu mẫu*

*np.random.seed(42)*

*data = {*

*'Nhóm A': np.random.normal(50, 15, 100),*

*'Nhóm B': np.random.normal(60, 12, 100),*

*'Nhóm C': np.random.normal(45, 20, 100),*

*'Nhóm D': np.random.normal(70, 10, 100)*

*}*

*# Boxplot cơ bản*

*plt.figure(figsize=(10, 6))*

*box\_plot = plt.boxplot([data['Nhóm A'], data['Nhóm B'], data['Nhóm C'], data['Nhóm D']],*

*labels=['Nhóm A', 'Nhóm B', 'Nhóm C', 'Nhóm D'],*

*patch\_artist=True) # Cho phép tô màu*

*# Tùy chỉnh màu sắc*

*colors = ['#FF6B6B', '#4ECDC4', '#45B7D1', '#96CEB4']*

*for patch, color in zip(box\_plot['boxes'], colors):*

*patch.set\_facecolor(color)*

*patch.set\_alpha(0.7)*

*plt.title('SO SÁNH ĐIỂM SỐ THEO NHÓM', fontsize=14, fontweight='bold', pad=20)*

*plt.xlabel('Nhóm', fontsize=12, fontweight='bold')*

*plt.ylabel('Điểm Số', fontsize=12, fontweight='bold')*

*plt.grid(True, alpha=0.3, axis='y')*

*plt.show()*

+ Boxplot với Seaborn (Dễ dàng & Đẹp hơn)

*import seaborn as sns*

*# Tạo DataFrame*

*df = pd.DataFrame({*

*'Điểm số': np.concatenate([data['Nhóm A'], data['Nhóm B'], data['Nhóm C'], data['Nhóm D']]),*

*'Nhóm': ['Nhóm A'] \* 100 + ['Nhóm B'] \* 100 + ['Nhóm C'] \* 100 + ['Nhóm D'] \* 100*

*})*

*plt.figure(figsize=(12, 8))*

*sns.boxplot(data=df, x='Nhóm', y='Điểm số', palette='Set2')*

*plt.title('PHÂN BỐ ĐIỂM SỐ THEO NHÓM', fontsize=16, fontweight='bold', pad=20)*

*plt.xlabel('Nhóm', fontsize=12, fontweight='bold')*

*plt.ylabel('Điểm Số', fontsize=12, fontweight='bold')*

*plt.xticks(rotation=45)*

*plt.grid(True, alpha=0.3, axis='y')*

*plt.show()*

*b. Violin Plot - Hiển thị Phân phối Chi tiết*

+ Violin Plot Cơ Bản

*plt.figure(figsize=(12, 8))*

*sns.violinplot(data=df, x='Nhóm', y='Điểm số', palette='pastel')*

*plt.title('PHÂN PHỐI ĐIỂM SỐ THEO NHÓM - VIOLIN PLOT', fontsize=16, fontweight='bold', pad=20)*

*plt.xlabel('Nhóm', fontsize=12, fontweight='bold')*

*plt.ylabel('Điểm Số', fontsize=12, fontweight='bold')*

*plt.xticks(rotation=45)*

*plt.grid(True, alpha=0.3, axis='y')*

*plt.show()*

+ Violin Plot Kết hợp với Boxplot

*plt.figure(figsize=(12, 8))*

*ax = sns.violinplot(data=df, x='Nhóm', y='Điểm số',*

*palette='light:gray', inner=None)*

*# Thêm boxplot lên trên violin plot*

*sns.boxplot(data=df, x='Nhóm', y='Điểm số',*

*width=0.3, palette='dark:red',*

*boxprops={'zorder': 2},*

*ax=ax)*

*plt.title('VIOLIN PLOT KẾT HỢP BOXPLOT', fontsize=16, fontweight='bold', pad=20)*

*plt.xlabel('Nhóm', fontsize=12, fontweight='bold')*

*plt.ylabel('Điểm Số', fontsize=12, fontweight='bold')*

*plt.xticks(rotation=45)*

*plt.grid(True, alpha=0.3, axis='y')*

*plt.show()*

1. *Phân Tích Nâng Cao với Swarm Plot & Kết Hợp*

*# Tạo visualization kết hợp*

*plt.figure(figsize=(14, 8))*

*# Violin plot làm nền*

*sns.violinplot(data=data\_real, x='Trình\_độ', y='Lương',*

*inner=None, palette='light:gray', alpha=0.7)*

*# Boxplot để hiển thị thống kê*

*sns.boxplot(data=data\_real, x='Trình\_độ', y='Lương',*

*width=0.2, palette='dark:red',*

*boxprops={'zorder': 2})*

*# Swarm plot để hiển thị từng điểm dữ liệu*

*sns.swarmplot(data=data\_real, x='Trình\_độ', y='Lương',*

*size=3, color='black', alpha=0.5)*

*plt.title('PHÂN TÍCH LƯƠNG TOÀN DIỆN THEO TRÌNH ĐỘ', fontsize=16, fontweight='bold', pad=20)*

*plt.xlabel('Trình độ', fontsize=12, fontweight='bold')*

*plt.ylabel('Lương ($)', fontsize=12, fontweight='bold')*

*plt.xticks(rotation=45)*

*plt.grid(True, alpha=0.3, axis='y')*

*# Thêm giá trị trung bình*

*means = data\_real.groupby('Trình\_độ')['Lương'].mean()*

*for i, (level, mean\_val) in enumerate(means.items()):*

*plt.text(i, mean\_val + 5000, f'Mean: {mean\_val:.0f}',*

*ha='center', va='bottom', fontweight='bold', fontsize=10)*

*plt.tight\_layout()*

*plt.show()*

1. *Phân Tích Theo Hai Biến Phân loại*

*# Thêm biến phân loại thứ hai - Giới tính*

*np.random.seed(42)*

*data\_real['Giới\_tính'] = np.random.choice(['Nam', 'Nữ'], len(data\_real))*

*# Boxplot với hue (màu sắc theo biến thứ hai)*

*plt.figure(figsize=(14, 8))*

*sns.boxplot(data=data\_real, x='Trình\_độ', y='Lương', hue='Giới\_tính', palette='Set2')*

*plt.title('LƯƠNG THEO TRÌNH ĐỘ VÀ GIỚI TÍNH', fontsize=16, fontweight='bold', pad=20)*

*plt.xlabel('Trình độ', fontsize=12, fontweight='bold')*

*plt.ylabel('Lương ($)', fontsize=12, fontweight='bold')*

*plt.legend(title='Giới tính', title\_fontsize=12)*

*plt.xticks(rotation=45)*

*plt.grid(True, alpha=0.3, axis='y')*

*plt.show()*

*# Violin plot với hue*

*plt.figure(figsize=(14, 8))*

*sns.violinplot(data=data\_real, x='Trình\_độ', y='Lương', hue='Giới\_tính', palette='Set2', split=True) # split=True để hiển thị cùng 1 violin*

*plt.title('PHÂN PHỐI LƯƠNG THEO TRÌNH ĐỘ VÀ GIỚI TÍNH', fontsize=16, fontweight='bold', pad=20)*

*plt.xlabel('Trình độ', fontsize=12, fontweight='bold')*

*plt.ylabel('Lương ($)', fontsize=12, fontweight='bold')*

*plt.legend(title='Giới tính', title\_fontsize=12)*

*plt.xticks(rotation=45)*

*plt.grid(True, alpha=0.3, axis='y')*

*plt.show()*

1. *Khi nào dùng Boxplot vs Violin plot:*

| **Loại biểu đồ** | **Ưu điểm** | **Nhược điểm** | **Khi nào dùng** |
| --- | --- | --- | --- |
| **Boxplot** | - Hiển thị rõ outliers - Thống kê tóm tắt (Q1, Q3, median) - Dễ so sánh trung vị | - Không show hình dạng phân phối - Mất thông tin về mật độ | Khi cần so sánh nhanh, tìm outliers |
| **Violin plot** | - Hiển thị hình dạng phân phối - Thấy được mật độ dữ liệu - Đa mode distribution | - Khó đọc với người mới - Không hiển thị rõ outliers | Khi cần hiểu sâu về phân phối |

**3.2. Bài tập thực hành 1**

Tìm hiểu các tính năng và cách sử dụng sản phẩm SweetViz (https://pypi.org/project/sweetviz) áp dụng trên tập dữ liệu Marketing Campaign.

**3.3. Bài tập thực hành 2**

Tìm hiểu các tính năng và cách sử dụng sản phẩm AutoViz (https://pypi.org/project/autoviz) áp dụng trên tập dữ liệu Marketing Campaign.

\_\_\_\_HẾT\_\_\_\_