THỰC HÀNH VỀ EDA

**1.1.1**

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

-Thống kê mô tả là phương pháp sử dụng các công cụ như biểu đồ và số liệu (trung bình, phương sai) để tóm tắt, sắp xếp và trình bày dữ liệu một cách có ý nghĩa, chỉ mô tả đặc điểm của tập dữ liệu đã thu thập. Ngược lại, thống kê suy luận sử dụng dữ liệu từ một mẫu đại diện để rút ra kết luận, đưa ra dự đoán và khái quát hóa về một quần thể lớn hơn.

- So sánh:

\*Thống kê mô tả:

+Mục đích: Mô tả các đặc điểm chính của một tập dữ liệu hiện có.

+Cách thực hiện: Sử dụng các phép đo xu hướng trung tâm (như trung bình, trung vị, yếu vị) và độ phân tán (như phương sai, độ lệch chuẩn).

+Phạm vi: Chỉ áp dụng cho chính tập dữ liệu mà bạn đang phân tích, không đưa ra kết luận về tổng thể lớn hơn.

+Ví dụ: Tính điểm trung bình của học sinh trong một lớp học hoặc biểu diễn doanh số bán hàng hàng ngày bằng biểu đồ.

\*Thống kê suy luận:

+Mục đích: Rút ra kết luận, đưa ra suy đoán hoặc dự đoán về một quần thể lớn hơn dựa trên dữ liệu thu thập từ một mẫu.

+Cách thực hiện: Sử dụng các phương pháp như kiểm định giả thuyết, phân tích hồi quy, và ước lượng tham số.

+Phạm vi: Khái quát hóa kết quả từ mẫu cho toàn bộ quần thể.

+Ví dụ: Sử dụng kết quả khảo sát ý kiến của một nhóm khách hàng để dự đoán xu hướng tiêu dùng của tất cả khách hàng.

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

-Các thước đo thống kê mô tả chính như trung bình, trung vị, phương sai và độ lệch chuẩn giúp tóm tắt và mô tả các đặc điểm của tập dữ liệu. Trung bình và trung vị cho biết xu hướng trung tâm của dữ liệu, trong khi phương sai và độ lệch chuẩn cho biết mức độ phân tán của dữ liệu xung quanh giá trị trung bình.

\*Mục đích của các thước đo thống kê mô tả chính

+Trung bình (Mean): Là tổng của tất cả các giá trị chia cho số lượng giá trị, cho biết "trung tâm" điển hình của dữ liệu.

+Trung vị (Median): Là giá trị ở giữa của tập dữ liệu khi được sắp xếp theo thứ tự, biểu thị một điểm mà một nửa dữ liệu nằm dưới nó và một nửa nằm trên nó.

+Phương sai (Variance): Đo lường mức độ biến thiên của từng số trong tập dữ liệu so với giá trị trung bình, tính toán giá trị trung bình của các bình phương sai số.

+Độ lệch chuẩn (Standard Deviation): Là căn bậc hai của phương sai, cho biết mức độ các giá trị của tập dữ liệu tập trung xung quanh giá trị trung bình. Giá trị độ lệch chuẩn thấp cho thấy dữ liệu gần trung bình, giá trị lớn cho thấy dữ liệu phân tán hơn.

-Nên dùng trung vị thay vì trung bình khi dữ liệu có các giá trị ngoại lệ hoặc [phân phối bị lệch](https://www.google.com/search?sca_esv=13677aa2df667dbd&cs=0&sxsrf=AE3TifOTtphpmygXVnWlbgZ9goUr1bT8GQ:1759234371841&q=ph%C3%A2n+ph%E1%BB%91i+b%E1%BB%8B+l%E1%BB%87ch&sa=X&ved=2ahUKEwigkZneuoCQAxU51DQHHQ3BOzUQxccNegQIDRAB&mstk=AUtExfBQbMu2NiKc6qOJ0_eQu66fR45iWonIOUADVIHQ-Wzs5d5mRgId7QmeSxXjGwJBUfZVXoenw3y_udN6L5TFyacMbJ5iz-nCPJLItgJG5YzGqkoWgF-epKx_ha4K61ewFLaD4Q6b9iNqGppV8NvQjCTn5N32cMIlrlkv82OwgdRx_haezBhDLpeYHszVSu1_63bpTVhTF1ronLEu7MXUF7ReSl4z7EikVKKtT-KgoVJBfO0D6M_frcXsSaSbz79CoVbPsVpBeqU_o5slCrLttMpP&csui=3), vì trung vị ít bị ảnh hưởng bởi các giá trị cực đoan này hơn.

\*Khi nào nên dùng trung vị thay vì trung bình

Trung vị nên được sử dụng khi tập dữ liệu có các giá trị ngoại lệ hoặc có phân phối bị lệch, vì các giá trị này ảnh hưởng mạnh đến giá trị trung bình.

+Phân phối lệch: Khi dữ liệu bị lệch (không đối xứng), một số giá trị cao hoặc thấp bất thường sẽ kéo giá trị trung bình. Trong trường hợp này, trung vị vẫn sẽ đại diện tốt hơn cho giá trị trung tâm của dữ liệu, do nó ít bị ảnh hưởng bởi các giá trị cực đoan.

+Sự hiện diện của các giá trị ngoại lệ: Một giá trị ngoại lệ là một điểm dữ liệu khác biệt đáng kể so với các quan sát khác. Nếu có các giá trị ngoại lệ, chúng có thể làm sai lệch giá trị trung bình, trong khi trung vị sẽ cho một cái nhìn chính xác hơn về xu hướng trung tâm của phần lớn dữ liệu.

**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)?**

-Để xác định phân bố của một tập dữ liệu, bạn có thể sử dụng biểu đồ histogram, biểu đồ hộp, hệ số độ lệch, và các kiểm định phân phối để kiểm tra xem dữ liệu có phù hợp với phân phối chuẩn hoặc các loại phân phối lệch khác nhau hay không.

\*Cách xác định phân bố của tập dữ liệu

1.Trực quan hóa dữ liệu:

+Biểu đồ Histogram: Vẽ biểu đồ histogram để xem hình dạng của phân bố. +Dữ liệu có hình chuông đối xứng có thể là phân phối chuẩn, trong khi đuôi dài về bên nào cho thấy sự lệch tương ứng.

+Biểu đồ hộp (Box Plot): Biểu đồ này giúp bạn hình dung sự phân tán của dữ liệu và xác định các ngoại lệ. Nó cũng có thể cho thấy sự lệch.

2.Sử dụng các thống kê mô tả:

+Hệ số độ lệch (Skewness): Đo lường tính bất đối xứng của phân phối.

+Độ lệch bằng 0: Phân phối đối xứng (tương tự phân phối chuẩn).

+Độ lệch âm: Phân phối lệch trái, đuôi dài về bên trái.

+Độ lệch dương: Phân phối lệch phải, đuôi dài về bên phải.

+Hệ số độ nhọn (Kurtosis): Đo độ dày của đuôi phân phối so với phân phối chuẩn.

3.Sử dụng các kiểm định phân phối:

+Kiểm định độ phù hợp phân phối: Đây là các kiểm định giả thuyết để xác định xem dữ liệu mẫu có tuân theo một phân phối xác suất cụ thể (ví dụ: phân phối chuẩn) hay không.

-Các loại phân phối phổ biến bao gồm phân phối chuẩn (đối xứng, hình chuông), phân phối lệch phải (đuôi dài về bên phải, trung vị < trung bình) và phân phối lệch trái (đuôi dài về bên trái, trung vị > trung bình).

\*Các loại phân bố phổ biến

1.Phân phối chuẩn (Normal Distribution):

+Có hình dạng đối xứng giống quả chuông.

+Trung bình, trung vị và mốt có giá trị gần bằng nhau.

+Dữ liệu được phân phối đều xung quanh giá trị trung bình.

+Khoảng 68% dữ liệu nằm trong phạm vi 1 độ lệch chuẩn so với trung bình, 95% trong 2 độ lệch chuẩn, và 99.7% trong 3 độ lệch chuẩn (Quy tắc thực nghiệm).

2.Phân phối lệch trái (Left-Skewed Distribution):

+Đuôi của phân phối dài hơn về phía bên trái (phía các giá trị âm).

+Trung vị thường lớn hơn giá trị trung bình.

+Phần lớn dữ liệu tập trung ở phía bên phải của biểu đồ.

3.Phân phối lệch phải (Right-Skewed Distribution):

+Đuôi của phân phối dài hơn về phía bên phải (phía các giá trị dương).

+Giá trị trung bình thường lớn hơn trung vị.

+Phần lớn dữ liệu tập trung ở phía bên trái của biểu đồ.

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

-Độ lệch chuẩn và phạm vi đều đánh giá sự phân tán dữ liệu, nhưng độ lệch chuẩn đo lường mức độ tập trung của dữ liệu xung quanh giá trị trung bình (càng thấp càng tập trung) và có đơn vị đo lường tương tự dữ liệu gốc, trong khi phạm vi (range) chỉ đơn giản là hiệu số giữa giá trị lớn nhất và nhỏ nhất, cho thấy khoảng cách bao quát của toàn bộ dữ liệu nhưng ít nhạy cảm với sự biến thiên ở giữa.

-Ý nghĩa:

1.Độ lệch chuẩn (Standard Deviation)

+Ý nghĩa: Đo lường mức độ phân tán trung bình của các điểm dữ liệu so với giá trị trung bình của tập dữ liệu.

+Độ lệch chuẩn cao: Các giá trị dữ liệu phân tán rộng rãi, xa giá trị trung bình.

+Độ lệch chuẩn thấp: Các giá trị dữ liệu tập trung chặt chẽ xung quanh giá trị trung bình.

+Ưu điểm: Có cùng đơn vị đo lường với dữ liệu gốc, giúp dễ hình dung và so sánh các tập dữ liệu khác nhau một cách trực quan hơn.

+Ứng dụng: Được sử dụng để đo lường rủi ro trong tài chính, kiểm soát chất lượng, và đánh giá tính ổn định hoặc biến động của dữ liệu theo thời gian.

2.Phạm vi (Range)

+Ý nghĩa: Là sự khác biệt giữa giá trị lớn nhất và nhỏ nhất trong một tập dữ liệu.

+Cách tính: Giá trị lớn nhất - Giá trị nhỏ nhất.

+Ý nghĩa của sự phân tán: Chỉ ra toàn bộ khoảng cách bao quát của dữ liệu, giúp xác định các giá trị ngoại lai (outliers) ở hai đầu.

+Hạn chế: Ít nhạy cảm với sự biến thiên của các giá trị nằm giữa phạm vi và có thể bị ảnh hưởng bởi các giá trị ngoại lai cực đoan.

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

-Trong biểu đồ hộp (boxplot), Q1 là phần tư thứ nhất (25% dữ liệu nhỏ hơn), Q2 (trung vị) là phần tư thứ hai (50% dữ liệu nhỏ hơn), và Q3 là phần tư thứ ba (75% dữ liệu nhỏ hơn). Ba thước đo này đánh dấu các vị trí quan trọng trong tập dữ liệu, được biểu thị bằng các đường thẳng trong hộp của biểu đồ, giúp ta hiểu được sự phân bố và phân tán của dữ liệu.

\*Chi tiết về các thước đo:

+Q1 (Tứ phân vị thứ nhất): Là giá trị mà 25% các điểm dữ liệu trong tập hợp nhỏ hơn nó.

+Q2 (Trung vị - Median): Là giá trị ở giữa của tập dữ liệu, chia tập dữ liệu thành hai nửa bằng nhau. 50% các điểm dữ liệu nhỏ hơn Q2.

+Q3 (Tứ phân vị thứ ba): Là giá trị mà 75% các điểm dữ liệu trong tập hợp nhỏ hơn nó.

\*Cách chúng được thể hiện trong biểu đồ hộp:

+Hộp: Phần thân của biểu đồ, được giới hạn bởi Q1 và Q3.

+Đường trong hộp: Kẻ một đường thẳng trong hộp để biểu thị Q2 (trung vị).

+Khoảng tứ phân vị (IQR): Khoảng cách giữa Q3 và Q1, cho biết sự phân tán của 50% dữ liệu ở giữa.

+Râu (whiskers): Các đường kẻ kéo dài từ hộp đến các giá trị ngoại lai hoặc các giá trị cực đại/cực tiểu không có ngoại lai, tùy thuộc vào cách biểu đồ được thiết kế.

**6. 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

1. Loại Bỏ Dữ Liệu (Deletion)

-Loại bỏ hàng (Listwise Deletion): Phương pháp này xóa toàn bộ hàng (quan sát) nếu nó có ít nhất một giá trị thiếu, bất kể biến nào. Đây là cách đơn giản nhất.

-Loại bỏ theo cặp (Pairwise Deletion): Phương pháp này chỉ sử dụng các quan sát đầy đủ cho cặp biến cụ thể đang được tính toán. Ví dụ, khi tính toán trung bình của biến A, chỉ cần loại bỏ các hàng thiếu A, nhưng vẫn giữ lại các hàng này khi tính toán thống kê cho biến B.

2. Điền Khuyết (Imputation)

A. Phương Pháp Đơn Giản

-Điền bằng Trung bình (Mean Imputation): Thay thế giá trị thiếu bằng giá trị trung bình của cột đó. Thường dùng cho biến định lượng.

-Điền bằng Trung vị (Median Imputation): Thay thế giá trị thiếu bằng giá trị trung vị của cột đó. Đây là lựa chọn tốt hơn khi dữ liệu có nhiều ngoại lai.

-Điền bằng Mode (Mode Imputation): Thay thế giá trị thiếu bằng giá trị xuất hiện nhiều nhất (mode) của cột đó. Thường dùng cho biến định danh (categorical).

-Điền bằng Hằng số (Constant Imputation): Thay thế bằng một hằng số cố định, chẳng hạn như 0 hoặc một mã đặc biệt để biểu thị dữ liệu bị thiếu.

B. Phương Pháp Nâng Cao

-Hồi quy (Regression Imputation): Sử dụng các biến khác trong tập dữ liệu để xây dựng mô hình hồi quy và dự đoán giá trị thiếu, qua đó tính đến mối quan hệ giữa các biến.

-Hàng xóm gần nhất (K-Nearest Neighbors - KNN Imputation): Điền giá trị thiếu bằng trung bình (hoặc mode) của K quan sát tương đồng nhất với quan sát bị thiếu.

-Đa điền khuyết (Multiple Imputation - MI): Tạo ra nhiều bộ dữ liệu đã được điền khuyết, phân tích từng bộ dữ liệu, và sau đó kết hợp các kết quả thống kê lại. Đây là phương pháp mạnh mẽ giúp ước tính cả độ không chắc chắn.

\* Lựa Chọn Phương Pháp

-Việc lựa chọn phương pháp phù hợp phụ thuộc vào mô hình của dữ liệu thiếu và tần suất thiếu.

-Nếu dữ liệu thiếu rất nhỏ (ví dụ: dưới 5%) và có mô hình MCAR (thiếu hoàn toàn ngẫu nhiên), bạn có thể cân nhắc Loại bỏ hàng hoặc Điền Trung bình/Trung vị cho thống kê mô tả đơn giản.

-Nếu lượng thiếu lớn hơn (ví dụ: 5% đến 20%), hoặc mục tiêu là phân tích sâu hơn (như mô hình hồi quy), nên ưu tiên các kỹ thuật Điền khuyết nâng cao như KNN hoặc Đa điền khuyết (MI) để có kết quả chính xác hơn.

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

-Để diễn giải biểu đồ histogram, hãy xem xét hình dạng phân phối (dạng chuông, lệch, hai đỉnh), trung tâm (giá trị trung bình, trung vị) và độ phân tán (sự biến động), cùng với các điểm ngoại lai. Với boxplot, bạn cần chú ý đến giá trị trung vị, phần tư (Q1, Q3), phạm vi (từ điểm cuối "râu" đến điểm cuối khác), và các điểm ngoại lai được biểu thị bằng các dấu chấm riêng biệt.

-Đọc và diễn giải Histogram:

1.Quan sát hình dạng phân phối:

+Phân phối đối xứng (dạng chuông): Dữ liệu tập trung ở giữa và đối xứng đều hai bên.

+Phân phối lệch trái: Đuôi đồ thị kéo dài về bên trái, có nghĩa là có nhiều giá trị lớn và ít giá trị nhỏ.

+Phân phối lệch phải: Đuôi đồ thị kéo dài về bên phải, có nghĩa là có nhiều giá trị nhỏ và ít giá trị lớn.

+Hai đỉnh: Dữ liệu có hai nhóm tập trung chính.

2.Xác định trung tâm:

Ước tính trung bình và trung vị: Nhìn vào vị trí trung tâm của các cột để ước tính giá trị trung bình hoặc trung vị của dữ liệu.

3.Đánh giá độ phân tán:

Độ rộng của biểu đồ: Một biểu đồ rộng cho thấy dữ liệu phân tán nhiều, trong khi biểu đồ hẹp cho thấy dữ liệu tập trung hơn.

4.Tìm điểm ngoại lai:

Kiểm tra các cột nhỏ, nằm riêng biệt ở hai đầu biểu đồ, đó là các điểm dữ liệu bất thường hoặc ngoại lai.

-Đọc và diễn giải Boxplot:

1.Xem xét đường trung vị (Median):

Đường kẻ ngang bên trong hộp cho biết giá trị trung vị (phân vị thứ 50) của dữ liệu.

2.Xác định các phần tư:

+Hộp: Biểu đồ biểu thị phân vị thứ 25 (Q1 - cạnh dưới hộp) và phân vị thứ 75 (Q3 - cạnh trên hộp), khoảng cách giữa Q1 và Q3 là khoảng cách liên tứ phân vị (IQR).

+Cạnh "râu": Các đường kẻ từ hộp đến điểm cuối cùng không phải là điểm ngoại lai.

3.Phát hiện điểm ngoại lai:

Các dấu chấm nằm ngoài hai cạnh "râu" của hộp là các điểm ngoại lai, cho thấy các giá trị bất thường.

4.Đánh giá độ lệch:

Nếu đường trung vị không nằm giữa hộp, điều đó cho thấy dữ liệu bị lệch.

**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?**

1. Phát hiện và Phân tích Ngoại lai

-Trực quan hóa: Sử dụng biểu đồ hộp (Box Plot) để trực quan hóa phạm vi và xác định các điểm nằm ngoài ranh giới (thường là 1.5×IQR).

-Các phương pháp thống kê:

+Sử dụng Điểm Z (Z-Score): Xác định các giá trị vượt quá một ngưỡng nhất định (ví dụ: Z>3 hoặc Z<−3).

+Sử dụng Phạm vi liên tứ phân vị (Interquartile Range - IQR): Xác định các giá trị nằm ngoài [Q1−1.5×IQR;Q3+1.5×IQR].

-Phân tích nguyên nhân: Tìm hiểu xem ngoại lai là do lỗi nhập liệu (sai số) hay là hiện tượng tự nhiên (ví dụ: thu nhập của một tỷ phú trong mẫu dân số).

2. Phương pháp Xử lý Ngoại lai

A. Loại bỏ (Deletion)

-Mô tả: Xóa các hàng (quan sát) chứa giá trị ngoại lai.

-Khi sử dụng: Chỉ nên áp dụng khi ngoại lai được xác định rõ là lỗi nhập liệu hoặc lỗi thu thập dữ liệu, hoặc khi mẫu dữ liệu rất lớn và việc loại bỏ sẽ không làm sai lệch đáng kể phân phối chung.

-Hạn chế: Dẫn đến mất thông tin và làm giảm kích thước mẫu.

B. Chuyển đổi (Transformation)

-Mô tả: Áp dụng các hàm toán học để giảm nhẹ tác động của các giá trị cực đoan, làm cho phân phối dữ liệu gần với phân phối chuẩn hơn.

-Các kỹ thuật phổ biến:

+Logarit (log): Thường được dùng cho các dữ liệu lệch phải (right-skewed).

+Căn bậc hai: Ít mạnh mẽ hơn logarit, nhưng hiệu quả.

-Lợi ích: Giữ lại tất cả dữ liệu, làm giảm ảnh hưởng của ngoại lai lên phương sai.

C. Giới hạn/Cắt xén (Capping/Winsorizing)

-Mô tả: Thay thế các giá trị ngoại lai bằng giá trị nằm ở ngưỡng tối đa hoặc tối thiểu được chấp nhận.

-Cách làm:

+Thay thế tất cả giá trị lớn hơn Q3+1.5×IQR bằng chính giá trị Q3+1.5×IQR.

+Thay thế tất cả giá trị nhỏ hơn Q1−1.5×IQR bằng chính giá trị Q1−1.5×IQR.

-Lợi ích: Giảm thiểu tác động của ngoại lai lên các thống kê, đồng thời giữ nguyên kích thước mẫu.

D. Sử dụng Thống kê bền vững (Robust Statistics)

Phương pháp này tập trung vào việc sử dụng các chỉ số thống kê mô tả ít bị ảnh hưởng bởi các giá trị cực đoan. Thay vì xử lý hoặc loại bỏ các ngoại lai trong tập dữ liệu, chúng ta chỉ cần lựa chọn các công cụ đo lường có khả năng "chịu đựng" tốt hơn sự hiện diện của chúng.

**1.2.1**

**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)?**

Trực quan hóa dữ liệu giúp hiểu, khám phá mẫu hình, xu hướng, và mối quan hệ phức tạp trong dữ liệu, đồng thời truyền đạt thông tin một cách rõ ràng, nhanh chóng cho nhiều đối tượng. Nó quan trọng trong khám phá dữ liệu (EDA) vì biến các con số khô khan thành biểu đồ sinh động, giúp các nhà phân tích dễ dàng nhận diện những thông tin quan trọng bị bỏ sót, phát hiện các yếu tố dự đoán tiềm năng, và hỗ trợ đưa ra quyết định sáng suốt dựa trên dữ liệu.

\*Vai trò của trực quan hóa dữ liệu trong phân tích dữ liệu:

-Hiểu sâu sắc dữ liệu: Biến các con số và chữ cái khô khan thành biểu đồ, hình ảnh trực quan, giúp não bộ dễ dàng nhận biết và xử lý thông tin.

-Phát hiện thông tin chuyên sâu (insight): Giúp tìm ra các mẫu hình, xu hướng, và mối quan hệ ẩn sau hàng triệu dòng dữ liệu mà có thể bị bỏ qua khi chỉ nhìn vào dữ liệu số.

-Thúc đẩy việc ra quyết định: Dữ liệu được trình bày rõ ràng, trực quan giúp người xem dễ dàng nắm bắt thông tin, hỗ trợ đưa ra các quyết định nhanh chóng và chính xác hơn.

-Truyền đạt thông tin hiệu quả: Cung cấp một cách trình bày thông tin chi tiết, rõ ràng và hấp dẫn, đặc biệt hữu ích khi truyền đạt kết quả phân tích cho những người không chuyên về dữ liệu.

-Khám phá ý tưởng mới: Là một công cụ mạnh mẽ để thúc đẩy việc tạo nên ý tưởng mới hoặc giải quyết các vấn đề phức tạp bằng cách khám phá dữ liệu từ nhiều góc độ khác nhau.

\*Tại sao trực quan hóa dữ liệu quan trọng trong khám phá dữ liệu (EDA):

-Giúp phát hiện xu hướng và mẫu hình: Trong giai đoạn EDA, trực quan hóa dữ liệu giúp các nhà phân tích khám phá xu hướng, sự biến đổi của dữ liệu theo thời gian và các mẫu hình bất thường một cách trực quan.

-Làm nổi bật các thông tin quan trọng: Các biểu đồ và đồ thị giúp làm nổi bật những điểm quan trọng và những thông tin có thể bị bỏ sót trong khối lượng dữ liệu lớn, từ đó định hướng các bước phân tích tiếp theo.

-Hiểu rõ các mối quan hệ phức tạp: Trực quan hóa dữ liệu giúp làm rõ mối quan hệ giữa các biến khác nhau, bao gồm cả những yếu tố dự đoán tiềm năng, trong quá trình mô hình dự đoán.

-Hỗ trợ giao tiếp: Cải thiện khả năng giao tiếp và truyền đạt các thông tin chuyên sâu mà nhà phân tích tìm thấy trong dữ liệu cho các bên liên quan, ngay cả khi họ không có chuyên môn sâu về phân tích.

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

Biểu đồ histogram dùng để xem phân phối tần suất của một biến liên tục. Biểu đồ phân tán (scatter plot) được dùng để quan sát mối quan hệ giữa hai biến. Biểu đồ hộp (boxplot) hữu ích khi so sánh sự phân phối của nhiều nhóm dữ liệu, còn biểu đồ thanh (bar chart) phù hợp với dữ liệu định danh để so sánh các giá trị.

1. Biểu đồ Histogram

-Khi nào sử dụng: Khi bạn muốn hiểu cách dữ liệu của một biến liên tục được phân bố, ví dụ như tần suất xuất hiện trong các khoảng giá trị khác nhau.

-Mục đích: Để xem xét các điểm tập trung của dữ liệu, khoảng trống, điểm cực đại, và các giá trị bất thường trong một biến số.

2. Biểu đồ Phân tán (Scatter Plot)

-Khi nào sử dụng: Khi bạn cần tìm hiểu mối quan hệ tương quan giữa hai biến số.

-Mục đích: Để xác định xem hai biến có tương quan dương (cùng tăng), tương quan âm (một tăng, một giảm), hoặc không tương quan hay không. Biểu đồ này cũng giúp phát hiện các mẫu hình, khoảng trống hoặc các điểm ngoại lệ trong dữ liệu.

3. Biểu đồ Hộp (Box Plot)

-Khi nào sử dụng: Khi bạn muốn so sánh sự phân phối của một biến liên tục qua nhiều nhóm dữ liệu khác nhau.

-Mục đích: Biểu đồ hộp cung cấp cái nhìn tổng quan về phân vị, độ phân tán và các giá trị ngoại lệ, giúp so sánh các nhóm dữ liệu một cách dễ dàng.

4. Biểu đồ Thanh (Bar Chart)

-Khi nào sử dụng: Khi bạn có dữ liệu định danh (categorical data) và muốn so sánh các giá trị hoặc tần suất giữa các nhóm khác nhau.

-Mục đích: Biểu đồ này thể hiện tần suất của các biến định danh dưới dạng các cột, giúp so sánh trực quan các mức độ khác nhau của các danh mục.

**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)?**

Để chọn biểu đồ phù hợp, bạn cần xác định mục đích hiển thị dữ liệu của mình: dùng để so sánh các hạng mục (biểu đồ cột, thanh), thể hiện sự phân bố (biểu đồ tròn, histogram, biểu đồ hộp), truyền tải sự thay đổi theo thời gian (biểu đồ đường), hoặc khám phá mối quan hệ giữa các biến số (biểu đồ phân tán).

1. Dữ liệu Phân loại (Categorical Data)

-Mục đích: So sánh các hạng mục riêng lẻ.

-Loại biểu đồ phù hợp:

+Biểu đồ cột (Column Chart): Phù hợp khi so sánh các giá trị khác nhau giữa các danh mục.

+Biểu đồ thanh (Bar Chart): Tốt cho các nhãn trục dài hoặc khi so sánh các giá trị trong một khoảng thời gian.

+Biểu đồ tròn (Pie Chart): Dùng để thể hiện tỷ trọng, thị phần, các phần trăm của một tổng thể chung.

2. Dữ liệu Số (Numerical Data)

-Mục đích: Hiển thị sự phân bố, xu hướng hoặc mối quan hệ giữa các biến số.

-Loại biểu đồ phù hợp:

+Biểu đồ phân tán (Scatter Plot): Hiển thị mối quan hệ giữa hai biến số, giúp phát hiện xu hướng và điểm ngoại lệ.

+Biểu đồ histogram (Histogram): Cho biết phân bố tần suất của một tập dữ liệu số.

+Biểu đồ hộp (Box Plot): Trực quan hóa sự biến động và phân bố của dữ liệu.

3. Dữ liệu Thời gian (Time-Series Data)

-Mục đích: Theo dõi và hiển thị sự thay đổi của dữ liệu qua các khoảng thời gian.

-Loại biểu đồ phù hợp:

+Biểu đồ đường (Line Chart): Là lựa chọn tiêu chuẩn để theo dõi sự thay đổi của một giá trị theo thời gian.

+Biểu đồ khu vực (Area Chart): Tương tự biểu đồ đường, nhưng làm nổi bật khối lượng dữ liệu theo thời gian.

\*Các Yếu Tố Cần Cân Nhắc Khác

-Mục tiêu của bạn: Bạn muốn người xem hiểu điều gì từ biểu đồ? (So sánh, phân bố, xu hướng, mối quan hệ?).

-Độ phức tạp của dữ liệu: Dữ liệu đơn giản có thể dùng biểu đồ cơ bản, còn dữ liệu phức tạp có thể cần biểu đồ mạng, treemap, hoặc biểu đồ phân cấp.

-Số lượng biến: Nếu có nhiều hơn hai biến, bạn có thể cần biểu đồ bong bóng hoặc thêm màu sắc/kích thước cho các điểm trong biểu đồ phân tán.

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

Matplotlib phù hợp cho việc tùy chỉnh đồ thị chuyên sâu, Seaborn được tối ưu cho biểu đồ thống kê đẹp mắt và dễ sử dụng, còn Plotly tạo ra các biểu đồ tương tác, động, thích hợp cho ứng dụng web và bảng điều khiển.

1.Matplotlib

-Mục đích: Là thư viện cốt lõi cho phép bạn tạo ra các biểu đồ cơ bản như đường, cột, phân tán, v.v.

-Ưu điểm: Cung cấp khả năng kiểm soát chi tiết và tùy chỉnh cao độ cho giao diện và bố cục của biểu đồ.

-Nhược điểm: Đòi hỏi nhiều mã hơn để tạo các biểu đồ phức tạp, đặc biệt là biểu đồ thống kê.

2.Seaborn

-Mục đích: Xây dựng trên nền tảng của Matplotlib, Seaborn cung cấp một giao diện cấp cao để tạo các biểu đồ thống kê đẹp mắt và dễ nhìn hơn với mã nguồn tối thiểu.

-Ưu điểm: Tích hợp mạnh mẽ với thư viện Pandas và cung cấp các tùy chỉnh mẫu sẵn cho các biểu đồ thống kê như biểu đồ dải, biểu đồ hộp.

-Nhược điểm: Ít linh hoạt hơn Matplotlib trong việc tùy chỉnh sâu.

3.Plotly

-Mục đích: Tạo ra các biểu đồ tương tác, động, nơi người dùng có thể zoom, pan, và xem chú giải chi tiết khi di chuột qua dữ liệu.

-Ưu điểm: Mạnh mẽ trong việc phát triển các bảng điều khiển (dashboards) và các ứng dụng web do tính tương tác và động của biểu đồ.

-Nhược điểm: Một số tùy chỉnh nâng cao có thể yêu cầu các giải pháp phức tạp.

4.Chọn thư viện nào?

-Chọn Matplotlib khi bạn cần kiểm soát tối đa giao diện và có khả năng viết nhiều mã.

-Chọn Seaborn khi bạn muốn tạo nhanh các biểu đồ thống kê đẹp mắt và thông tin mà không cần quá nhiều mã.

-Chọn Plotly khi bạn cần các biểu đồ tương tác, động để sử dụng trong các bảng điều khiển hoặc ứng dụng web.

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

Để tạo ra biểu đồ trực quan hóa dễ hiểu và hiệu quả, cần tuân thủ các nguyên tắc: ưu tiên sự rõ ràng và đơn giản, chọn loại biểu đồ phù hợp, sử dụng màu sắc và phông chữ hợp lý, loại bỏ các yếu tố thừa, nhấn mạnh thông tin quan trọng, và hiểu rõ khán giả mục tiêu.

\*Nguyên tắc thiết kế chính

-Ưu tiên sự rõ ràng và đơn giản: Loại bỏ những yếu tố không cần thiết làm lộn xộn biểu đồ, sử dụng nhãn rõ ràng, và chỉ sử dụng những gì thực sự nâng cao sự hiểu biết của người xem.

-Chọn loại biểu đồ phù hợp: Chọn biểu đồ (như biểu đồ đường, biểu đồ cột) phù hợp với loại dữ liệu bạn muốn trình bày, ví dụ: sử dụng biểu đồ đường để theo dõi thay đổi theo thời gian.

-Sử dụng màu sắc hợp lý: Chọn bảng màu nhất quán và có ý nghĩa để phân biệt các nhóm dữ liệu hoặc mức độ quan trọng, tránh lạm dụng màu sắc.

-Sử dụng phông chữ phù hợp: Giữ kiểu chữ đơn giản, dễ đọc, sử dụng một vài kích thước khác nhau để tạo hệ thống phân cấp và làm nổi bật thông tin chính.

-Giảm sự lộn xộn (làm sạch dữ liệu): Loại bỏ các thành phần không cần thiết để tránh tạo ra hình ảnh trực quan khó hiểu và tập trung vào thông tin chính.

-Nhấn mạnh thông tin quan trọng: Sử dụng hệ thống phân cấp trực quan để hướng dẫn người xem, làm nổi bật các yếu tố quan trọng nhất trước tiên.

-Hiểu rõ khán giả mục tiêu: Đồng cảm với khán giả bằng cách hiểu trình độ kiến thức và sở thích của họ để thiết kế trực quan hóa dữ liệu phù hợp với họ.

-Bắt đầu biểu đồ ở mức 0: Đối với biểu đồ thanh và cột, luôn bắt đầu từ mức 0 để tránh làm sai lệch nhận thức về dữ liệu.

-Bao gồm chú thích và nhãn rõ ràng: Cung cấp chú thích và nhãn để giúp người xem diễn giải biểu đồ một cách chính xác.

-Tránh hiệu ứng 3D không cần thiết: Hiệu ứng 3D có thể gây hiểu lầm và làm giảm khả năng đọc biểu đồ.

**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?**

Để tạo biểu đồ histogram hoặc biểu đồ thanh (bar chart) bằng Matplotlib, bạn cần nhập matplotlib.pyplot, dùng plt.hist() cho histogram hoặc plt.bar() cho biểu đồ thanh với dữ liệu của bạn, sau đó sử dụng plt.show() để hiển thị biểu đồ. Dưới đây là các đoạn mã mẫu cho từng loại biểu đồ:

1. Histogram được sử dụng để trực quan hóa phân phối của dữ liệu bằng cách nhóm dữ liệu vào các "bin" (thùng) và đếm tần suất xuất hiện trong mỗi bin.

Code:

import matplotlib.pyplot as plt

import numpy as np # Thường dùng để tạo dữ liệu mẫu

# Dữ liệu mẫu

danh\_sach\_gia\_tri = [22, 55, 62, 45, 21, 22, 34, 42, 42, 4, 2, 102, 95, 85, 55, 110, 120, 70, 65, 55, 111, 115, 80, 75, 65, 54, 44, 43, 42, 48]

so\_luong\_bin = 10 # Số lượng bin cho biểu đồ histogram

# Tạo biểu đồ histogram

plt.hist(danh\_sach\_gia\_tri, bins=so\_luong\_bin, color='skyblue', edgecolor='black')

# Thêm tiêu đề và nhãn trục

plt.xlabel('Giá trị')

plt.ylabel('Tần suất')

plt.title('Biểu đồ Histogram của một tập dữ liệu')

# Hiển thị biểu đồ

plt.show()

2. Biểu đồ thanh sử dụng các thanh hình chữ nhật để hiển thị các giá trị của các danh mục khác nhau.

Code:  
import matplotlib.pyplot as plt

# Dữ liệu mẫu

ten\_thu\_vien = ['Matplotlib', 'Seaborn', 'Plotly', 'Pandas']

so\_luong\_nguoi\_dung = [100, 120, 80, 90]

# Tạo biểu đồ thanh

plt.bar(ten\_thu\_vien, so\_luong\_nguoi\_dung, color='lightgreen')

# Thêm tiêu đề và nhãn trục

plt.xlabel('Thư viện')

plt.ylabel('Số lượng người dùng')

plt.title('Số lượng người dùng cho các thư viện trực quan hóa')

# Hiển thị biểu đồ

plt.show()

**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?**

Để xuất biểu đồ từ Python ra các định dạng PNG, PDF, hoặc HTML, bạn có thể sử dụng phương thức savefig() của Matplotlib với các định dạng tệp khác nhau, ví dụ: plt.savefig('my\_plot.png') cho PNG, plt.savefig('my\_plot.pdf') cho PDF. Đối với định dạng HTML, bạn có thể lưu biểu đồ thành tệp HTML để nhúng vào các báo cáo web bằng cách sử dụng các thư viện như [Plotly](https://www.google.com/search?sca_esv=db06df8ffc1e69f4&cs=0&sxsrf=AE3TifNi29Jke4zf71kCj7Dl1RuZMir7jw:1759312133551&q=Plotly&sa=X&ved=2ahUKEwjEzYW23IKQAxX1M1kFHU-XKCgQxccNegQIBBAB&mstk=AUtExfAysSyezOJqpTLMf-leY3xXULgt-lU5I1Z935qWlULMBy-HFzErlVMJDKQld7fhZoTBzZtBs9mlzh8kHnBb0mYXRNVKl3z4TiD2ubf3XJ4aO4CEFvKRl9coOMXg4piZA0gWy2hwXojuCZpSGpGfJfALovIu2u3gg4m4WKZ1OtvQ12vp8jtYv6RmaNK4RWRi6JiRxUoRcVeb81bhhv5BOaVROUtIvVpuEuUhhFmKiJnGbrXm0EF7eWEBpbNVzynhk2ASVdDLzUSDynTtMCELghpOPFG58yTspsVihsIeueFtV50gPo7JJaIS6vsOKkp3BJF_-s8M97AcWfS-1GrBBBxqxR6t0S6WOBOh-VV8hhzkfH4dWkdMb5e_EnGhQkngZwU4rfVUIO9c3vJM1VRKNA&csui=3) hoặc [Altair](https://www.google.com/search?sca_esv=db06df8ffc1e69f4&cs=0&sxsrf=AE3TifNi29Jke4zf71kCj7Dl1RuZMir7jw:1759312133551&q=Altair&sa=X&ved=2ahUKEwjEzYW23IKQAxX1M1kFHU-XKCgQxccNegQIBBAC&mstk=AUtExfAysSyezOJqpTLMf-leY3xXULgt-lU5I1Z935qWlULMBy-HFzErlVMJDKQld7fhZoTBzZtBs9mlzh8kHnBb0mYXRNVKl3z4TiD2ubf3XJ4aO4CEFvKRl9coOMXg4piZA0gWy2hwXojuCZpSGpGfJfALovIu2u3gg4m4WKZ1OtvQ12vp8jtYv6RmaNK4RWRi6JiRxUoRcVeb81bhhv5BOaVROUtIvVpuEuUhhFmKiJnGbrXm0EF7eWEBpbNVzynhk2ASVdDLzUSDynTtMCELghpOPFG58yTspsVihsIeueFtV50gPo7JJaIS6vsOKkp3BJF_-s8M97AcWfS-1GrBBBxqxR6t0S6WOBOh-VV8hhzkfH4dWkdMb5e_EnGhQkngZwU4rfVUIO9c3vJM1VRKNA&csui=3).

\*Xuất biểu đồ (PNG, PDF) bằng Matplotlib

-Thư viện Matplotlib cung cấp một cách đơn giản để lưu biểu đồ ở nhiều định dạng khác nhau.

+Tạo biểu đồ: Đầu tiên, bạn tạo biểu đồ bằng Matplotlib theo cách thông thường.

+Lưu biểu đồ: Sử dụng phương thức savefig() để lưu biểu đồ vào một tệp.

Định dạng PNG:

import matplotlib.pyplot as plt

# ... mã tạo biểu đồ của bạn ...

plt.savefig('ten\_bieu\_do.png')

Định dạng PDF:

import matplotlib.pyplot as plt

# ... mã tạo biểu đồ của bạn ...

plt.savefig('ten\_bieu\_do.pdf')

\*Xuất biểu đồ định dạng HTML

-Để tạo biểu đồ tương tác ở định dạng HTML, bạn có thể sử dụng các thư viện chuyên dụng như Plotly hoặc Altair.

-Sử dụng Plotly:

+Cài đặt thư viện: Nếu chưa có, hãy cài đặt Plotly: pip install plotly.

+Tạo biểu đồ: Sử dụng Plotly để tạo biểu đồ tương tác.

+Xuất sang HTML: Sử dụng phương thức write\_html() để lưu biểu đồ thành một tệp HTML

import plotly.express as px

# ... mã tạo biểu đồ với Plotly ...

fig = px.bar(x=["A", "B", "C"], y=[1, 3, 2])

fig.write\_html("ten\_bieu\_do\_tuong\_tac.html")

-Sử dụng Altair:

+Cài đặt thư viện: pip install altair.

+Tạo biểu đồ: Sử dụng Altair để tạo biểu đồ.

+Xuất sang HTML: Sử dụng phương thức save() để lưu biểu đồ thành tệp HTML

import altair as alt

import pandas as pd

# ... mã tạo biểu đồ với Altair ...

df = pd.DataFrame({'x': [1, 2, 3], 'y': [2, 1, 3]})

chart = alt.Chart(df).mark\_bar().encode(x='x', y='y')

chart.save('ten\_bieu\_do\_altair.html')

\*Sử dụng trong báo cáo

-Tệp PNG hoặc PDF: Bạn có thể chèn tệp hình ảnh PNG hoặc PDF trực tiếp vào các ứng dụng xử lý văn bản như Microsoft Word, Google Docs hoặc các công cụ báo cáo khác.

-Tệp HTML: Tệp HTML có thể được nhúng vào các báo cáo dựa trên web hoặc mở trực tiếp trong trình duyệt để hiển thị biểu đồ tương tác, cho phép người dùng khám phá dữ liệu một cách chi tiết hơn.

**1.3.1:**

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

* **Phân tích đơn biến (univariate analysis)** là phân tích một biến duy nhất trong dữ liệu để hiểu đặc điểm, phân phối và xu hướng của nó.
* **Mục đích:**
* Xem biến đó có dạng dữ liệu gì (định tính, định lượng)
* Đo lường trung tâm (mean, median, mode)
* Đo lường độ phân tán (variance, standard deviation, range, IQR)
* Quan sát phân phối (histogram, boxplot)
* **Ví dụ:**
* Phân tích chiều cao của học sinh trong một lớp: trung bình bao nhiêu? có phân phối chuẩn hay không?
* Vẽ histogram số lượng khách theo độ tuổi

→ **Đặc điểm:** chỉ tập trung một biến, không xét mối quan hệ với biến khác

* **Phân tích hai biến (bivariate analysis)** là phân tích hai biến cùng lúc để tìm hiểu mối quan hệ hoặc sự phụ thuộc giữa chúng
* **Mục đích:**
* Kiểm tra xem biến này ảnh hưởng như thế nào đến biến kia
* Phát hiện mối tương quan, xu hướng hoặc sự khác biệt
* **Ví dụ:**
* Xem mối quan hệ giữa chiều cao và cân nặng (scatter plot, hệ số tương quan Pearson)
* So sánh thu nhập trung bình theo giới tính (boxplot hoặc bar chart)
* Kiểm định thống kê: t-test, chi-square test, …

→ **Đặc điểm:** phân tích cặp biến, thường có một biến độc lập (x) và một biến phụ thuộc (y)

* Sự khác nhau giữa Univariate and Bivariate

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | Univariate | Bivariate |
| **Số biến** | 1 biến | 2 biến |
| **Mục tiêu** | Hiểu đặc điểm phân phối của biến | Tìm mối quan hệ giữa hai biến |
| **Công cụ** | Histogram, Boxplot, thống kê mô tả | Scatter plot, Bar chart, Correlation, t-test |
| **Kết quả trả lời** | Biến này có đặc điểm gì? | Hai biến có liên quan/thay đổi theo nhau không? |

**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)?**

Trong phân tích đơn biến, ta thường dùng các thống kê mô tả để tóm tắt và hiểu rõ đặc điểm của dữ liệu. Có thể chia làm 3 nhóm chính:

1. **Các thước đo vị trí (Central Tendency)** dùng để mô tả giá trị điển hình của dữ liệu:

* Trung bình (mean): Giá trị trung bình cộng
* Trung vị (median): Giá trị nằm ở giữa khi sắp xếp dữ liệu
* Mốt (mode): Giá trị xuất hiện nhiều nhất
* Ngoài ra: Percentile/Quartile: Vị trí dữ liệu theo phần trăm (ví dụ Q1,Q2,Q3)

1. **Các thước đo phân tán (Dispersion/Variability)** dùng để xem dữ liệu phân tán nhiều hay ít quanh giá trị trung tâm:

* Range (khoảng biến thiên) = Max - Min
* Phương sai (variance)
* Độ lệch chuẩn (standard deviation, SD): căn bậc hai của phương sai
* IQR (Interquartile Range) = Q3 - Q1 (dùng trong boxplot để phát hiện outlier)
* MAD (Mean Absolute Deviation): trung bình của khoảng cách tuyệt đối tới trung bình

1. **Các thước đo hình dạng phân phối (Shape of Distribution)** giúp hiểu dữ liệu có lệch hay nhọn không:

* Skewness (Độ lệch):
* Skew > 0 → lệch phải (đuôi dài bên phải)
* Skew < 0 → lệch trái
* Kurtosis (Độ nhọn):
* Kurtosis cao → phân phối nhọn, tập trung
* Kurtosis thấp → phân phối bẹt, phân tán

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

1. Xác định mối quan hệ (Association/ Correlation) đây là bước khám phá mối liên hệ:

* Nếu cả hai là biến định lượng (số):
* Dùng hệ số tương quan Pearson (linear correlation)
* Giá trị từ -1 → +1
* +1 = tương quan thuận hoàn hảo, -1 = tương quan nghịch hoàn hảo, 0 = không có quan hệ tuyến.
* Nếu dữ liệu không tuyến tính hoặc không phân phối chuẩn → dùng Spearman’s rank correlation hoặc Kendall’s tau
* Biểu đồ thường dùng: scatter plot, regression line
* Nếu một biến định tính, một biến định lượng:
* Dùng so sánh trung bình (t-test, ANOVA) để xem biến định tính có ảnh hưởng đến biến định lượng không
* Ví dụ: thu nhập khác nhau giữa nam và nữ
* Nếu cả hai là biến định tính:
* Dùng bảng chéo (cross-tabulation)
* Kiểm định Chi-square test of independence để xem có sự phụ thuộc không

1. Xác định mối quan hệ nhân quả (Causality) đây là bước khó hơn, vì tương quan khác với nhân quả

* Tương quan (Correlation): chỉ cho thấy hai biến thay đổi cùng nhau, nhưng không khẳng định biến này gây ra biến kia
* Ví dụ: Doanh số kem tăng cùng với số ca cuối nước → thật ra biến ẩn (confounder) là nhiệt độ mùa hè
* Nhân quả (Causation): để chứng minh biến A gây ảnh hưởng đến biến B, cần:
* Thực nghiệm (Controlled Experiment): thay đổi A có dẫn đến thay đổi B không (ví dụ: thử nghiệm A/B)
* Phân tích thống kê nâng cao: hồi quy đa biến (multivariate regression), kiểm soát biến gây nhiễu
* Mô hình chuỗi thời gian (Time Series Analysis): kiểm định Granger causality để xem biến nào dự đoán biến nào
* Nguyên tắc Bradford Hill trong y học: tính nhất quán, hợp lý sinh học, mối quan hệ liều-đáp ứng, trình tự thời gian…

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

1. Hiệp biến (Covariance)

* Định nghĩa: Đo mức độ hai biến thay đổi cùng nhau
* Công thức:



xi = một giá trị x đã cho trong tập dữ liệu

xm = giá trị trung bình hoặc trung bình của các giá trị x

yi = giá trị y trong tập dữ liệu tương ứng với xi

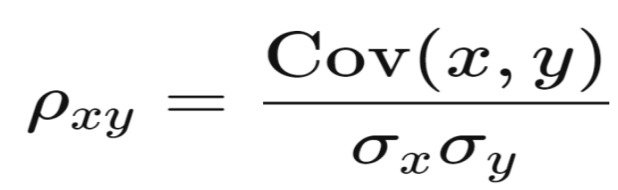
ym = giá trị trung bình hoặc trung bình của các giá trị y

n = số điểm dữ liệu

* Ý nghĩa:
* Cov > 0 → hai biến tăng/giảm cùng chiều
* Cov < 0 → một biến tăng thì biến kia giảm (ngược chiều)
* Cov ≃0 → gần như không có mối liên hệ tuyến tính
* Đơn vị: Phụ thuộc vào đơn vị đo của X và Y → khó so sánh giữa các cặp biến khác nhau

1. Tương quan (Correlation)

* Định nghĩa: là hiệp biến được chuẩn hóa, cho biết mức độ chặt chẽ của mối quan hệ tuyến tính
* Công thức:



pxy = Hệ số tương quan Pearson

cov = hiệp biến của x và y

ox = độ lệch chuẩn của x

oy = độ lệch chuẩn của y

* Ý nghĩa:
* Luôn nằm trong khoảng [-1;1]
* +1 → tương quan thuận hoàn hảo
* -1 → tương quan nghịch hoàn hảo
* 0 → không có mối quan hệ tuyến tính

1. So sánh

|  |  |  |
| --- | --- | --- |
| Tiêu chí | Hiệp biến | Tương quan |
| Thang đo | Có thể rất lớn hoặc nhỏ, phụ thuộc vào đơn vị | Luôn trong [-1;1] |
| Đơn vị | Phụ thuộc vào đơn vị của biến | Không có đơn vị ( đã chuẩn hóa) |
| Ý nghĩa | Chỉ cho biết chiều quay (cùng/ngược) | Cho biết cả chiều và mức độ chặt chẽ |
| Ứng dụng | Trung gian trong tính toán  thống kê | Phổ biến hơn trong phân tích dữ liệu |

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

1. Trong phân tích đơn biến

Biểu đồ được dùng khi muốn nhìn rõ phân phối và đặc điểm của một biến duy nhất

* Histogram → phân phối của biến liên tục (tuổi, điểm số)
* Boxplot → phát hiện giá trị ngoại lai (outlier), xem độ phân tán
* Bar chart → tần suất/ tỷ lệ của biến phân loại (giới tính, ngành học)
* Density plot → so sánh dạng phân phối mượt hơn histogram

→ Dùng khi: muốn mô tả “biến này có hình dạng phân phối ra sao?”

1. Trong phân tích hai biến

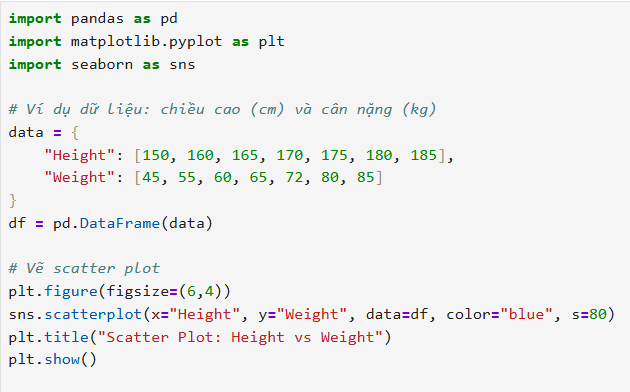
Biểu đồ được dùng khi muốn xem mối quan hệ hoặc sự phụ thuộc giữa hai biến

* Scatter plot → mối quan hệ giữa hai biến số (chiều cao <-> cân nặng)
* Box plot/ Violin plot → so sánh phân phối biến số giữa các nhóm phân loại (thu nhập theo giới tính)
* Grouped bar chart/ Mosaic plot → quan hệ giữa hai biến phân loại (giới tính <-> nghề nghiệp)
* Line chart → mối quan hệ theo thời gian (doanh thu theo tháng)

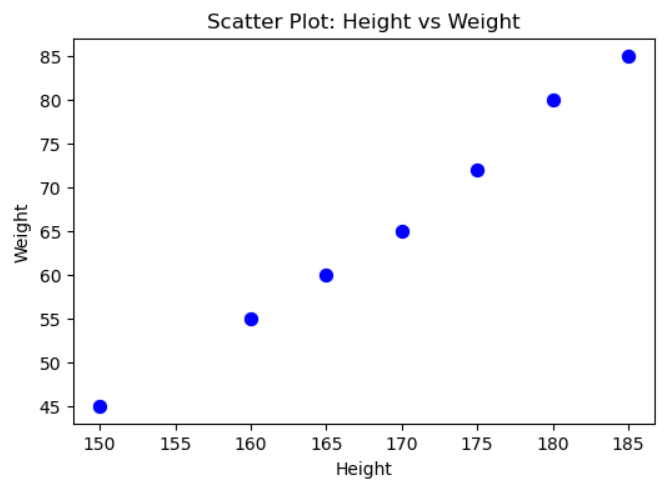
→ Dùng khi: muốn trả lời “hai biến này có liên hệ với nhau không, và liên hệ như thế nào?”

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

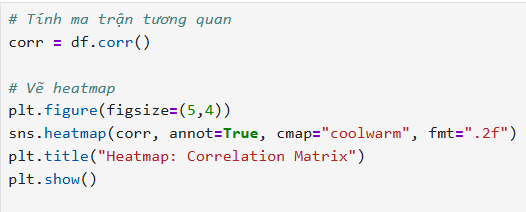
1. Scatter plot (quan hệ giữa hai biến số)



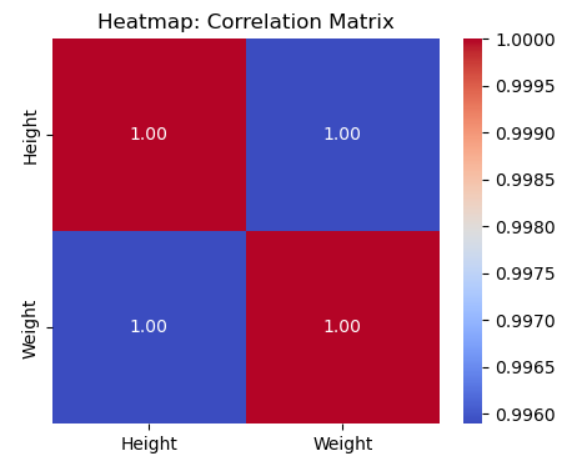
→ Kết quả: mỗi điểm biểu diễn một người, dễ thấy quan hệ thuận giữa chiều cao và cân nặng



1. Heatmap (tương quan giữa nhiều biến)



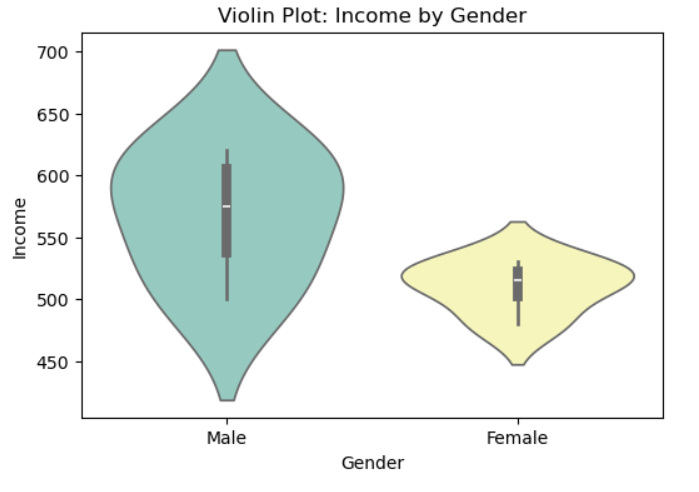
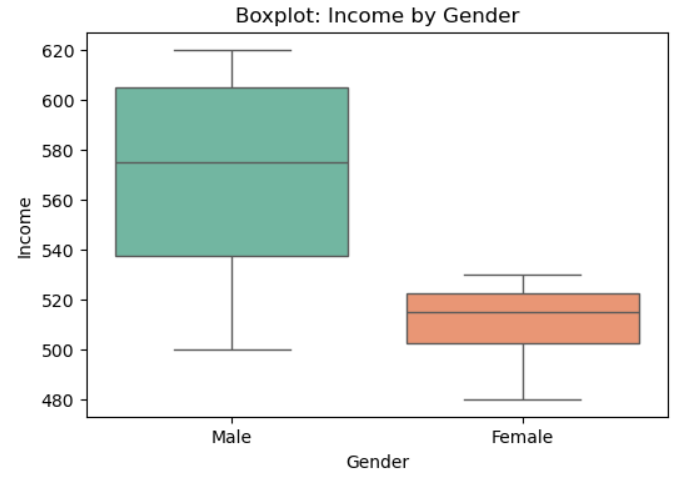
→ Kết quả: ma trận hiển thị hệ số tương quan giữa các biến



**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?**

Ví dụ bằng Python : giả sử có dữ liệu thu nhập theo giới tính:





* Ý nghĩa:
* Boxplot:
* Thể hiện median (trung vị), Q1 - Q3 (IQR), và các outlier
* Dùng khi muốn tập trung vào thống kê tóm tắt
* Violin plot:
* Kết hợp boxplot với dạng phân phối xác suất (density)
* Dùng khi muốn thấy hình dạng phân phối (có lệch/trải rộng không)