**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

Icon

Description automatically generated with low confidence

**HỌC PHẦN: MÁY HỌC**

**BÁO CÁO BÀI TẬP**

**TRẢ LỜI CÂU HỎI LAB 05**

**Giảng viên hướng dẫn:**

TS. Đỗ Như Tài

**Sinh viên thực hiện:**

Nguyễn Toàn Thắng – 3123410343

Dương Tùng Thiện - 3123410350

Hồ Minh Tiến – 3123410373

Đặng Thái Tú – 3123410399

**TP.HCM, Tháng 10 năm 2025**

**PHÂN CÔNG VÀ ĐÁNH GIÁ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Thứ tự** | **Họ tên** | **Mã số sinh viên** | **Nội dung công việc** | **Mức độ đóng góp** | **Giảng viên đánh giá** |
| 1 | Nguyễn Toàn Thắng | 3123410343 | - Bộ câu hỏi lý thuyết số 3 (2 câu đầu)  - Toàn bộ bài tập thực hành của phần Phân tích đơn biến và hai biến | 25% |  |
| 2 | Dương Tùng Thiện | 3123410350 | - Bộ câu hỏi lý thuyết số 3 (5 câu sau)  - Soạn nội dung báo cáo Word  - Toàn bộ bài tập thực hành của phần Thống kê mô tả | 25% |  |
| 3 | Hồ Minh Tiến | 3123410373 | - Bộ câu hỏi lý thuyết số 1 (4 câu)  - Bài tập thực hành 2 của phần Xử lý và trực quan hóa dữ liệu | 25% |  |
| 4 | Đặng Thái Tú | 3123410399 | - Bộ câu hỏi lý thuyết số 1 (4 câu)  - Bộ câu hỏi lý thuyết số 2  - Bài tập thực hành 1 của phần Xử lý và trực quan hóa dữ liệu | 25% |  |

**NHẬN XÉT CỦA GIẢNG VIÊN**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….  
……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**……………………………………………………………………………………….**

**LỜI CẢM ƠN**

Trước tiên, chúng em xin bày tỏ lòng biết ơn sâu sắc đến thầy Đỗ Như Tài – giảng viên bộ môn Máy học, Khoa Công nghệ Thông tin, Chuyên ngành Khoa học máy tính, Trường Đại học Sài Gòn – người đã luôn tận tâm giảng dạy, truyền đạt kiến thức và đồng hành cùng chúng em trong suốt quá trình học tập cũng như thực hiện đồ án này.

Nhờ sự hướng dẫn tận tình, tâm huyết cùng những kiến thức nền tảng và kỹ năng quý báu mà thầy đã truyền đạt, chúng em mới có thể hoàn thành đề tài một cách trọn vẹn. Đây không chỉ là một sản phẩm học thuật mà còn là kết quả của một quá trình học hỏi, cố gắng và trưởng thành.

Tuy đã nỗ lực hết mình trong quá trình thực hiện, song do hạn chế về kiến thức thực tế và kinh nghiệm còn non kém, bài báo cáo này chắc chắn không thể tránh khỏi những thiếu sót. Chúng em rất mong nhận được những nhận xét, góp ý quý báu từ thầy để có thể hoàn thiện hơn trong tương lai.

Một lần nữa, chúng em xin trân trọng gửi lời cảm ơn chân thành đến thầy.

Kính chúc thầy luôn mạnh khỏe, công tác tốt và tiếp tục truyền cảm hứng cho thế hệ sinh viên chúng em.

**Chúng em xin chân thành cảm ơn!**

Mục lục

[TRẢ LỜI CÂU HỎI 5](#_Toc210408323)

[1. Thống kê mô tả 5](#_Toc210408324)

[1.1. Thống kê mô tả là gì? Nó khác gì với thống kê suy luận (inferential statistics)? 5](#_Toc210408325)

[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? 5](#_Toc210408326)

[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)? 6](#_Toc210408327)

[1.4. Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu? 7](#_Toc210408328)

[1.5. Sự khác biệt giữa các thước đo như Q1, Q2, Q3 trong biểu đồ hộp (boxplot) là gì? 7](#_Toc210408329)

[1.6. Làm thế nào để xử lý giá trị thiếu (missing values) trước khi tính toán các chỉ số thống kê mô tả? 8](#_Toc210408330)

[1.7. Bạn có thể giải thích cách đọc và diễn giải một biểu đồ histogram hoặc boxplot từ dữ liệu thực tế không? 9](#_Toc210408331)

[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? 10](#_Toc210408332)

[2. Xử lý và trực quan hóa dữ liệu 12](#_Toc210408333)

[2.1. 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)? 12](#_Toc210408334)

[2.2. Các loại biểu đồ phổ biến (như histogram, scatter plot, boxplot, bar chart) được sử dụng trong các trường hợp nào? 12](#_Toc210408335)

[2.3. 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)? 13](#_Toc210408336)

[2.4. Sự khác biệt giữa các thư viện trực quan hóa trong Python như Matplotlib, Seaborn và Plotly là gì? 15](#_Toc210408337)

[2.5. Những nguyên tắc thiết kế nào cần tuân thủ để tạo ra một biểu đồ trực quan hóa dễ hiểu và hiệu quả? 16](#_Toc210408338)

[2.6. 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? 16](#_Toc210408339)

[2.7. Làm thế nào để xuất biểu đồ từ Python ra các định dạng như PNG, PDF hoặc HTML để sử dụng trong báo cáo? 20](#_Toc210408340)

[3. Phân tích đơn biến và hai biến 21](#_Toc210408341)

[3.1. Phân tích đơn biến (univariate analysis) là gì? Nó khác gì với phân tích hai biến (bivariate analysis) trong khám phá dữ liệu? 21](#_Toc210408342)

[3.2. 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)? 21](#_Toc210408343)

[3.3. Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến (ví dụ: tương quan, nhân quả)? 22](#_Toc210408344)

[3.4. Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance) trong phân tích hai biến là gì? 24](#_Toc210408345)

[3.5. Khi nào nên sử dụng biểu đồ trực quan hóa trong phân tích đơn biến so với phân tích hai biến? 24](#_Toc210408346)

[3.6. Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến? 25](#_Toc210408347)

[3.7. 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? 26](#_Toc210408348)

# TRẢ LỜI CÂU HỎI

## 1. Thống kê mô tả

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

Thống kê mô tả (Descriptive Statistics) là các phương pháp được sử dụng để tóm tắt và mô tả chính của một tập dữ liệu. Mục tiêu là để hiểu rõ hơn về dữ liệu trước khi thực hiện các tác vụ học máy phực tạp hơn.

Sự khác nhau của thống kê mô tả (Descriptive Statistics) và thống kê suy luận (Inferential Statistics):

|  |  |  |
| --- | --- | --- |
| Tiêu chí | Thống kê mô tả (Descriptive Statistics) | Thống kê suy luận (Inferential Statistics) |
| Mục đích | Tóm tắt, trình bày dữ liệu đã có | Suy luận từ mẫu ra tổng thể |
| Công cụ | Mean, median, SD, bảng, biểu đồ | Kiểm định, khoảng tin cậy, hồi quy |
| Giới hạn | Chỉ áp dụng cho dữ liệu thu thập | Dùng xác suất để khái quát rộng hơn |
| Ví dụ | Giá trị trung bình của thuộc tính 'age' trong bộ dữ liệu Pima Indians là 33.241". Đây là một mô tả về chính bộ dữ liệu đó. | Dựa trên mẫu dữ liệu Pima Indians, sử dụng các kiểm định giả thuyết để suy luận rằng có một mối liên hệ có ý nghĩa thống kê giữa nồng độ glucose trong huyết tương và khả năng mắc bệnh tiểu đường trong toàn bộ dân số người Pima Indians, chứ không chỉ riêng trong mẫu 768 người |

### 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?

1. Trung Bình (Mean): Trung bình được tính bằng tổng các giá trị trong tập dữ liệu chia cho số lượng quan sát.

Mục đích: Phản ánh giá trị trung tâm hoặc giá trị điển hình của tập dữ liệu, cung cấp cái nhìn tổng quát về đặc điểm chung của dữ liệu.

2. Trung Vị (Median): Là giá trị nằm ở vị trí chính giữa của tập dữ liệu khi các giá trị được sắp xếp theo thứ tự tăng dần hoặc giảm dần.

Mục đích: Trung vị biểu thị điểm trung tâm của dữ liệu, ít chịu ảnh hưởng từ các giá trị ngoại lai hơn so với trung bình, do đó phù hợp hơn để mô tả dữ liệu có phân phối không đồng đều hoặc chứa các giá trị bất thường.

3. Phương Sai (Variance): Là giá trị trung bình của bình phương các độ lệch giữa mỗi giá trị trong tập dữ liệu và giá trị trung bình.

Mục đích: Thước đo này đánh giá mức độ phân tán của dữ liệu so với giá trị trung bình, phản ánh sự biến động hoặc tính không đồng đều trong tập dữ liệu.

4. Độ Lệch Chuẩn (Standard Deviation): Là căn bậc hai của phương sai.

Mục đích: Tương tự phương sai, độ lệch chuẩn đo lường mức độ phân tán của dữ liệu, nhưng được biểu diễn cùng đơn vị với dữ liệu gốc, giúp dễ dàng diễn giải và áp dụng trong thực tiễn.

1. Các Trường Hợp Nên Sử Dụng Trung Vị

a. Dữ Liệu Có Phân Bố Lệch (Skewed Distribution): Khi dữ liệu có phân bố lệch, trung bình có thể bị kéo lệch theo sự lệch của phân bố, không phản ánh chính xác xu hướng trung tâm. Trung vị chỉ phụ thuộc vào vị trí trung tâm của dữ liệu, do đó ít bị ảnh hưởng bởi độ lệch hơn.

b. Dữ Liệu Chứa Giá Trị Ngoại Lai (Outliers): Giá trị ngoại lai (quá lớn hoặc quá nhỏ so với phần lớn dữ liệu) làm trung bình thay đổi đáng kể, dẫn đến kết quả không đại diện. Trung vị không bị ảnh hưởng nhiều bởi các giá trị cực biên so với trung bình, do đó cung cấp thước đo trung tâm đáng tin cậy hơn.

### 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)?

Để xác định phân bố của một tập dữ liệu, có thể sử dụng hai phương pháp chính:

* Thống kê mô tả: tính các chỉ số như trung bình, trung vị, độ lệch (skewness), độ nhọn (kurtosis) để mô tả đặc trưng của dữ liệu và so sánh với các phân bố chuẩn.
* Trực quan hóa dữ liệu: vẽ histogram, boxplot, density plot hoặc Q-Q plot để quan sát trực tiếp hình dạng phân bố.

Các loại phân bố phổ biến:

* Phân bố Chuẩn (Normal Distribution): loại phân bố có hình dạng đặc trưng của một đường cong hình chuông. Giá trị trung bình và độ lệch chuẩn khác nhau có thể được biển đổi theo để tuân theo phân bố Guassian tiêu chuẩn (có giá trị trung bình là 0 và độ lệch chuẩn là 1) thông qua một kỹ thuật gọi là chuẩn hóa (Standardization).
* Phân bố Lệch (Skewed Distribution): loại phân bố bị dịch chuyển hoặc bị nén về một hướng (trái hoặc phải). Giá trị độ lệch tính được cho biết hướng và mức độ lệch:
  + Phân bố lệch phải (Positive Skew): khi đuôi của phân bố kéo dài về phía bên phải. Giá trị độ lệch sẽ là một số dương.
  + Phân bố lệch trái (Negative Skew): khi đuôi của phân bố kéo dài về phía bên phải. Giá trị độ lệch sẽ là một số âm.
  + Các giá trị độ lệch càng gần 0 thì phân bố càng ít bị lệch.
* Phân bố mũ (Exponential Distribution): một dạng phân bố có thể được nhận biết từ hình dạng của các biểu đồ tần suất.

Phân bố hai đỉnh (Bimodal Distribution): một dạng phân bố có đặc trưng là sự xuất hiện của hai đỉnh trên biểu đồ tần suất, cho thấy dữ liệu có thể được nhóm quanh hai giá trị trung tâm khác nhau.

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

**Phạm vi (Range):**

Định nghĩa: Range = giá trị lớn nhất – giá trị nhỏ nhất của tập dữ liệu (max – min).

Ý nghĩa:

* Cho biết khoảng biến động tối đa mà dữ liệu trải ra.
* Do chỉ dựa vào hai giá trị biên, range rất nhạy cảm với các giá trị ngoại lệ (outliers). Nếu có một giá trị cực lớn hoặc cực nhỏ, range có thể bị “kéo dài” rất nhiều mà không phản ánh đúng mức độ biến thiên chung của dữ liệu.

Ví dụ minh hoạ: Giả sử bạn có các giá trị: 5, 7, 9, 10, 100 → Range = 100 – 5 = 95. Mặc dù hầu hết dữ liệu tập trung trong khoảng 5–10, nhưng do có “100” làm ngoại lệ, phạm vi bị kéo rất rộng.

**Độ lệch chuẩn (Standard Deviation):**

Định nghĩa:

* Độ lệch chuẩn đo mức độ mà các điểm dữ liệu lệch khỏi giá trị trung bình (mean).
* Nó là căn bậc hai của phương sai (variance), giúp đưa đơn vị trở lại giống đơn vị dữ liệu gốc.

Ý nghĩa trong đánh giá sự phân tán:

* Nếu độ lệch chuẩn nhỏ, các giá trị dữ liệu gần trung bình — tức là dữ liệu “tập trung”.
* Nếu độ lệch chuẩn lớn, các giá trị phân tán xa hơn trung bình — tức là dữ liệu “rải rộng”.
* Trong phân bố chuẩn (normal distribution), khoảng 68% dữ liệu nằm trong ±1 độ lệch chuẩn quanh trung bình, khoảng 95% nằm trong ±2 độ lệch chuẩn, và khoảng 99.7% nằm trong ±3 độ lệch chuẩn (quy tắc 68–95–99.7).

Ví dụ minh hoạ: Giả sử hai tập dữ liệu:

* Tập A: {8, 9, 10, 11, 12} → độ lệch chuẩn nhỏ vì các giá trị gần nhau.
* Tập B: {5, 10, 15, 20, 25} → cùng trung bình (15) nhưng độ lệch chuẩn lớn hơn vì các giá trị cách xa trung bình hơn.

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

Sự khác biệt giữa các thước đo Q1, Q2, Q3 trong biểu đồ hộp(boxplot):

* Q1 (Phân vị thứ 25 – 25th Percentile): giá trị mà tại đó 25% dữ liệu có giá trị thấp hơn. Trong boxplot, Q1 chính là cạnh dưới của hình chữ nhật.
* Q2 (Phân vị thứ 50 – 50th Percentile hay Trung vị – Median): giá trị nằm chính giữa của tập dữ liệu khi đã được sắp xếp. Q2 chia tập dữ liệu thành 2 nửa bằng nhau. Trong boxplot, Q2 được biểu thị bằng được một đường kẻ ngang bên trong hộp.
* Q3 (Phân vị thứ 75 – 75th Percentile): giá trị mà tại đó 75% dữ liệu có giá trị thấp hơn nó (hoặc 25% dữ liệu cao hơn nó). Trong boxplot, Q3 là cạnh trên của hình chữ nhật.

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

**- Xác định và đánh giá giá trị thiếu**

Việc xác định được các giá trị bị thiếu (missing values) giúp hiểu rõ nguồn gốc và bản chất của giá trị thiếu để lựa chọn phương pháp xử lý phù hợp. Giá trị thiếu có thể xuất hiện do lỗi nhập liệu, dữ liệu không được thu thập, hoặc mang ý nghĩa cụ thể.

Phân loại giá trị thiếu:

* MCAR (Missing Completely at Random): Giá trị thiếu xảy ra hoàn toàn ngẫu nhiên, không liên quan đến các biến khác hoặc chính giá trị thiếu.
* MAR (Missing at Random): Giá trị thiếu phụ thuộc vào các biến quan sát được trong tập dữ liệu.
* MNAR (Missing Not at Random): Giá trị thiếu phụ thuộc vào chính giá trị bị thiếu, ví dụ: người tham gia khảo sát không trả lời vì câu hỏi nhạy cảm.

Lưu ý: Việc xác định loại giá trị thiếu (MCAR, MAR, hoặc MNAR) là yếu tố quan trọng để lựa chọn phương pháp xử lý phù hợp, vì nó ảnh hưởng đến tính chính xác của các chỉ số thống kê mô tả.

**- Các phương pháp xử lý giá trị thiếu**

a. Loại bỏ giá trị thiếu: Thường được áp dụng khi tỷ lệ giá trị thiếu nhỏ (thường dưới 5%) và việc loại bỏ không làm mất tính đại diện của tập dữ liệu.

Ưu điểm: Đơn giản, dễ thực hiện và không cần giả định phức tạp.

Nhược điểm: Có thể làm mất thông tin quan trọng nếu giá trị thiếu mang ý nghĩa hoặc tỷ lệ giá trị thiếu lớn.

b. Thay thế giá trị thiếu bằng các giá trị:

* Thay bằng mean, median, mode,... của trường dữ liệu đó: Mean phù hợp với dữ liệu có phân phối chuẩn, trong khi median phù hợp với dữ liệu có phân phối lệch. Mode (giá trị xuất hiện nhiều nhất) thường được dùng cho dữ liệu danh mục.
* Thay bằng giá trị cố định (dành cho dữ liệu danh mục): Gán các giá trị như "Unknown", "Other", hoặc giá trị phổ biến nhất trong tập dữ liệu.
* Thay bằng giá trị từ mô hình dự đoán: Sử dụng khi dữ liệu có mối quan hệ phức tạp giữa các biến, sử dụng các thuật toán như k-Nearest Neighbors (KNN) hoặc hồi quy để dự đoán giá trị thiếu.

c. Giữ giá trị thiếu như một danh mục riêng: Phù hợp với dữ liệu danh mục khi giá trị thiếu mang ý nghĩa cụ thể, ví dụ: "Không trả lời" trong khảo sát.

d. Sử dụng các phương pháp nâng cao

* Nội suy (Interpolation): Dành cho dữ liệu chuỗi thời gian, sử dụng các phương pháp như nội suy tuyến tính để ước lượng giá trị thiếu dựa trên các giá trị lân cận.
  + Ví dụ: Trong dữ liệu nhiệt độ hàng ngày, giá trị thiếu có thể được nội suy dựa trên các giá trị trước và sau.
* Multiple Imputation (Đa thế): Sử dụng các phương pháp như MICE (Multiple Imputation by Chained Equations) để tạo ra nhiều tập dữ liệu hoàn chỉnh, sau đó tổng hợp kết quả để giảm sai số.
  + Ví dụ: Trong nghiên cứu y học, MICE có thể được sử dụng để xử lý giá trị thiếu trong dữ liệu huyết áp, cân nặng, và tuổi của bệnh nhân.

**- Kiểm tra sau khi xử lý**

* Đảm bảo không còn giá trị thiếu: Kiểm tra tập dữ liệu để xác nhận rằng tất cả giá trị thiếu đã được xử lý.
* Đánh giá tác động: So sánh phân phối dữ liệu trước và sau khi xử lý để đảm bảo rằng phương pháp được chọn không làm sai lệch đặc tính của dữ liệu.

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

Cách đọc và diễn giải biểu đồ Histogram:

* Histogram hoạt động bằng cách nhóm dữ liệu vào các bin (khoảng giá trị) và cung cấp số lượng quan sát trong mỗi bin đó.
* Cách đọc:
  + Trục hoành (x-axis): biểu thị giá trị của thuộc tính.
  + Trục tung (y-axis): biểu thị tần suất hoặc số lượng quan sát rơi vào mỗi khoảng giá trị(bin).
  + Chiều cao cột: mức độ tập trung dữ liệu trong khoảng đó.
* Diễn giải:
  + Hình dạng phân bố:
    - Đường cong hình chuông: dữ liệu có thể theo phân bố chuẩn.
    - Lệch trái: nhiều giá trị cao.
    - Lệch phải: nhiều giá trị thấp.
  + Đỉnh (mode): khoảng có nhiều dữ liệu nhất (giá trị phổ biến).
  + Độ rộng phân bố: nếu cột trải rộng thì dữ liệu phân tán lớn và ngược lại.
  + Outlier: cột lẻ loi xa vùng trung tâm thì có ngoại lai.

Cách đọc và diễn giải biểu đồ Boxplot:

* Boxplot sẽ tóm tắt sự phân bố dữ liệu bằng cách hiển thị các giá trị thống kê quan trọng.
* Cách đọc:
  + Đường kẻ ở giữa hộp (Median): đây là giá trị trung vị (phân vị thứ 50), chia dữ liệu thành hai nửa bằng nhau.
  + Hộp: bao quanh 50% dữ liệu ở giữa, từ phân vị thứ 25 (Q1) đến phân vị thứ 75 (Q3). Khoảng cách giữa Q1 và Q3 được gọi là khoảng tứ phân vị (Interquartile Range - IQR).
  + Râu (Whiskers): các đường kẻ kéo dài từ hộp cho thấy sự lan rộng (Spread) của dữ liệu.
  + Các dấu chấm (Outliers): các điểm nằm ngoài râu được gọi là giá trị ngoại lai tiềm năng. Đây là các giá trị lớn hơn 1.5 lần IQR so với Q3 hoặc nhỏ hơn 1.5 lần IQR so với Q1.
* Diễn giải:
  + Vị trí Q2 (Median): Nếu nằm gần Q1 thì dữ liệu lệch phải hay nằm gần Q3 thì dữ liệu lệch trái.
  + Độ rộng hộp (IQR = Q3 – Q1): cho thấy mức phân tán dữ liệu.
  + Chiều dài râu: râu dài hơn về bên nào thì dữ liệu phân tán hơn về phía đó.
  + Outlier: điểm bất thường cần chú ý.

### 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?

**- Xác định giá trị ngoại lai:** Mục đích nhằm phát hiện các giá trị bất thường có khả năng làm sai lệch kết quả thống kê, từ đó quyết định cách xử lý phù hợp.

Phương pháp xác định:

* Quy tắc IQR (Interquartile Range):
  + Tính khoảng tứ phân vị: IQR = Q3 - Q1 (trong đó Q1 là tứ phân vị thứ nhất, Q3 là tứ phân vị thứ ba).
  + Xác định ngưỡng: Giá trị ngoại lai là các điểm nằm ngoài khoảng [Q1 - 1.5IQR, Q3 + 1.5IQR].
* Z-score (Điểm chuẩn hóa):
  + Tính Z-score cho mỗi giá trị: Z = (x - mean) / std (trong đó mean là trung bình, std là độ lệch chuẩn).
  + Giá trị có |Z| > 3 thường được coi là ngoại lai.
* Trực quan hóa:
  + Sử dụng biểu đồ hộp (boxplot) hoặc biểu đồ phân tán (scatter plot) để phát hiện trực quan các giá trị ngoại lai.

Lưu ý: Cần kiểm tra xem giá trị ngoại lai có mang ý nghĩa thực tế (ví dụ: doanh thu bất thường do sự kiện đặc biệt) hay chỉ là lỗi dữ liệu (ví dụ: lỗi nhập liệu).

**- Các phương pháp xử lý giá trị ngoại lai**

* Loại bỏ giá trị ngoại lai: Khi giá trị ngoại lai được xác định là lỗi dữ liệu hoặc không đại diện cho xu hướng chung của tập dữ liệu.
  + Ưu điểm: Đơn giản, giúp loại bỏ nhiễu và cải thiện tính chính xác của các chỉ số thống kê.
  + Nhược điểm: Có nguy cơ làm mất thông tin quan trọng nếu giá trị ngoại lai mang ý nghĩa thực tế.
* Thay thế giá trị ngoại lai (Imputation): Khi muốn giữ lại các hàng dữ liệu nhưng cần giảm tác động của giá trị ngoại lai.
  + Phương pháp: Thay thế giá trị ngoại lai bằng trung bình, trung vị của ngưỡng IQR (Q1 - 1.5IQR hoặc Q3 + 1.5IQR).
  + Ưu điểm: Duy trì số lượng dữ liệu và giảm ảnh hưởng của giá trị ngoại lai.
* Biến đổi dữ liệu (Transformation): Áp dụng khi dữ liệu có phân phối lệch (skewed) và giá trị ngoại lai làm sai lệch các chỉ số thống kê.
  + Phương pháp: Áp dụng các phép biến đổi như logarit, căn bậc hai, hoặc chuẩn hóa để giảm tác động của giá trị ngoại lai.
  + Ưu điểm: Giảm ảnh hưởng của giá trị ngoại lai mà không cần loại bỏ dữ liệu.
  + Nhược điểm: Các chỉ số thống kê sau biến đổi có thể khó diễn giải hơn.
* Sử dụng các phương pháp mạnh mẽ (Robust Methods): Áp dụng khi muốn tính toán các chỉ số thống kê mô tả mà không cần loại bỏ hoặc thay thế giá trị ngoại lai.
  + Phương pháp: Sử dụng các thước đo thống kê mạnh mẽ như trung vị, khoảng tứ phân vị (IQR), hoặc các phương pháp thống kê không nhạy cảm với ngoại lai (robust statistics).
  + Ưu điểm: Đảm bảo kết quả thống kê đáng tin cậy mà không cần thay đổi dữ liệu gốc.

Lưu ý: Cần báo cáo rõ ràng rằng dữ liệu chứa giá trị ngoại lai và giải thích lý do giữ nguyên/biến đổi chúng.

**- Kiểm tra sau khi xử lý**

* So sánh phân phối: Sử dụng biểu đồ hộp (boxplot) hoặc biểu đồ histogram để so sánh phân phối dữ liệu trước và sau khi xử lý giá trị ngoại lai.
* Tính toán thống kê mô tả: Tính các chỉ số như trung bình, trung vị, độ lệch chuẩn, và tứ phân vị để đánh giá dữ liệu sau xử lý.
* Đánh giá tác động: So sánh các chỉ số thống kê trước và sau khi xử lý để đảm bảo rằng phương pháp được chọn không làm mất thông tin quan trọng hoặc gây sai lệch.

## 2. Xử lý và trực quan hóa dữ liệu

### 2.1. 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)?

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

* Hỗ trợ hiểu và diễn giải dữ liệu: Việc trực quan hóa dữ liệu giúp người dùng nhanh chóng nắm bắt xu hướng, mẫu hình và mối quan hệ trong dữ liệu mà không cần phân tích số liệu thô phức tạp.
* Phát hiện xu hướng và bất thường: Giúp dễ dàng phát hiện các điểm bất thường (outliers) hoặc xu hướng ẩn trong dữ liệu, từ đó hỗ trợ ra quyết định nhanh chóng.
* Truyền đạt thông tin hiệu quả: Giúp trình bày dữ liệu một cách trực giác, dễ hiểu cho những người không chuyên về dữ liệu.
* Hỗ trợ ra quyết định: Giúp các nhà phân tích và nhà quản lý đưa ra quyết định dựa trên dữ liệu.
* Tăng cường khám phá dữ liệu: Trực quan hóa tương tác (interactive visualization) cho phép người dùng khám phá dữ liệu theo cách linh hoạt, chẳng hạn như lọc, phóng to hoặc xem chi tiết từng phần dữ liệu, từ đó phát hiện những hiểu biết mới.
* Tăng tính thuyết phục và hấp dẫn: Hình ảnh trực quan sẽ thu hút sự chú ý và thuyết phục hơn so với văn bản hoặc bảng số liệu.

Vai trò quan trọng của trực quan hóa dữ liệu trong phân tích dữ liệu khám phá (EDA)

* Hiểu cấu trúc và đặc điểm của dữ liệu: Giúp nhận diện phân phối dữ liệu, xác định tính đối xứng, độ lệch (skewness), hoặc phạm vi giá trị.
* Phát hiện bất thường và giá trị ngoại lai (outliers): Các biểu đồ như scatter plot hoặc box plot giúp dễ dàng nhận ra các điểm dữ liệu bất thường, từ đó hỗ trợ làm sạch dữ liệu hoặc điều tra nguyên nhân của các giá trị này.
* Khám phá mối quan hệ giữa các biến: Sử dụng các biểu đồ phân tán (scatter plot), ma trận tương quan (correlation matrix) hoặc bản đồ nhiệt (heatmap) giúp xác định mối quan hệ giữa các biến.
* Tăng tính tương tác và khám phá sâu hơn: Các công cụ trực quan hóa tương tác (như Tableau, Power BI) cho phép người dùng lọc, khoan sâu (drill-down), hoặc thay đổi góc nhìn dữ liệu, giúp khám phá dữ liệu linh hoạt và hiệu quả hơn.

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

1. Histogram:

Mô tả: Biểu đồ histogram hiển thị phân phối tần suất của một biến liên tục bằng cách chia dữ liệu thành các khoảng (bins) và đếm số lượng quan sát trong mỗi khoảng.

Trường hợp sử dụng:

* Phân tích phân phối của một biến: Giúp hiểu rõ đặc điểm phân phối của một biến liên tục (ví dụ: độ lệch, hình dạng phân phối như chuẩn, lệch trái/phải).
* Xác định phạm vi giá trị, đỉnh (mode), hoặc các khoảng dữ liệu tập trung.

2. Scatter Plot (Biểu đồ phân tán):

Mô tả: Biểu đồ phân tán hiển thị mối quan hệ giữa hai biến liên tục bằng cách biểu diễn các điểm dữ liệu trên mặt phẳng tọa độ.

Trường hợp sử dụng:

* Khám phá mối quan hệ (tương quan) giữa hai biến.
* Phát hiện các điểm bất thường (outliers) hoặc cụm dữ liệu (clusters).

3. Box Plot (Biểu đồ hộp):

Mô tả: Biểu đồ hộp hiển thị phân phối của dữ liệu thông qua các giá trị như trung vị (median), tứ phân vị (quartiles), và các giá trị ngoại lai (outliers).

Trường hợp sử dụng:

* So sánh phân phối của một biến giữa các nhóm.
* Phát hiện giá trị ngoại lai hoặc sự biến động của dữ liệu.

4. Bar Chart (Biểu đồ cột):

Mô tả: Biểu đồ cột hiển thị giá trị của một biến phân loại (categorical) hoặc so sánh các nhóm thông qua chiều cao của các cột.

Trường hợp sử dụng:

* So sánh giá trị giữa các danh mục.
* Hiển thị tần suất hoặc tỷ lệ của các danh mục.

### 2.3. 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)?

1. Dữ liệu phân loại (Categorical Data): Dữ liệu được chia thành các danh mục hoặc nhóm không có thứ tự số học (ví dụ: loại sản phẩm, khu vực, giới tính).

Mục tiêu phân tích:

* So sánh giá trị giữa các danh mục.
* Hiển thị tần suất hoặc tỷ lệ của từng danh mục.

Biểu đồ phù hợp:

* Bar Chart (Biểu đồ cột): So sánh giá trị định lượng giữa các danh mục.
  + Dùng bar chart khi cần so sánh trực quan giữa các danh mục.
* Pie Chart (Biểu đồ tròn): Hiển thị tỷ lệ phần trăm hoặc phân phối của các danh mục trong tổng thể.
  + Dùng pie chart khi tổng các danh mục tạo thành 100% và số danh mục ít (thường dưới 6).
  + Tránh dùng pie chart nếu có quá nhiều danh mục hoặc giá trị gần nhau, vì khó so sánh.
* Stacked Bar Chart (Biểu đồ cột chồng): So sánh nhiều biến định lượng trong cùng một danh mục.

Lưu ý: Đảm bảo các danh mục rõ ràng và không quá nhiều để tránh làm biểu đồ rối mắt.

2. Dữ liệu số (Numerical Data): Dữ liệu dạng số, có thể là liên tục (ví dụ: chiều cao, doanh thu) hoặc rời rạc (ví dụ: số lượng đơn hàng).

Mục tiêu phân tích:

* Hiểu phân phối của dữ liệu.
* Phát hiện bất thường, xu hướng, hoặc mối quan hệ giữa các biến.

Biểu đồ phù hợp:

* Histogram: Hiển thị phân phối tần suất của một biến liên tục.
  + Dùng histogram để kiểm tra hình dạng phân phối (chuẩn, lệch, đa đỉnh).
* Box Plot (Biểu đồ hộp): Tóm tắt phân phối dữ liệu (trung vị, tứ phân vị, giá trị ngoại lai).
  + Dùng box plot để so sánh phân phối giữa các nhóm hoặc phát hiện giá trị ngoại lai.
* Scatter Plot (Biểu đồ phân tán): Xem xét mối quan hệ giữa hai biến số liên tục.
  + Dùng scatter plot để khám phá tương quan hoặc cụm dữ liệu.
* Line Chart (Biểu đồ đường): Có thể dùng cho dữ liệu số liên quan đến thời gian (xem chi tiết ở mục dưới).

3. Dữ liệu thời gian (Time-Series Data): Dữ liệu được thu thập theo thời gian, thường là các giá trị số đo lường tại các mốc thời gian (ví dụ: doanh thu hàng tháng, nhiệt độ hàng ngày).

Mục tiêu phân tích:

* Xác định xu hướng, chu kỳ, hoặc biến động theo thời gian.
* So sánh dữ liệu tại các khoảng thời gian khác nhau.

Biểu đồ phù hợp:

* Line Chart (Biểu đồ đường): Hiển thị xu hướng hoặc sự thay đổi của dữ liệu qua thời gian.
  + Dùng line chart khi muốn nhấn mạnh xu hướng liên tục hoặc dài hạn.
* Area Chart (Biểu đồ vùng): Tương tự biểu đồ đường, nhưng nhấn mạnh khối lượng hoặc tổng tích lũy.
  + Dùng area chart khi cần hiển thị tổng tích lũy hoặc so sánh nhiều chuỗi dữ liệu thời gian.
* Bar Chart: So sánh giá trị tại các thời điểm cụ thể (thường dùng khi thời gian rời rạc).
  + Dùng bar chart khi các mốc thời gian cách nhau không đều hoặc muốn nhấn mạnh giá trị tại từng điểm.

Lưu ý: Đảm bảo trục thời gian được sắp xếp theo thứ tự và có khoảng cách phù hợp.

4. Nguyên tắc chung để chọn biểu đồ

* Dựa vào mục tiêu phân tích:
  + So sánh → Bar Chart, Box Plot.
  + Phân phối → Histogram, Box Plot.
  + Tương quan → Scatter Plot.
  + Xu hướng thời gian → Line Chart.
* Dựa vào số lượng biến:
  + Một biến: Histogram, Box Plot.
  + Hai biến: Scatter Plot, Line Chart.
  + Nhiều biến hoặc danh mục: Bar Chart, Stacked Bar Chart, Heatmap.

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

1. Matplotlib

* Hỗ trợ hầu hết các loại biểu đồ cơ bản (line, bar, scatter, histogram, box plot, pie chart, v.v.) và cả biểu đồ 3D.
* Cung cấp khả năng tùy chỉnh rất cao, cho phép kiểm soát chi tiết (màu sắc, nhãn, trục, lưới, v.v.).
* Chủ yếu tạo biểu đồ tĩnh, nhưng có hỗ trợ tương tác cơ bản qua các công cụ như mpl\_toolkits hoặc kết hợp với Jupyter.
* Hiệu suất: Phù hợp với các biểu đồ đơn giản hoặc phức tạp, nhưng cần viết nhiều mã hơn để tạo giao diện đẹp.
* Tính dễ sử dụng: Giao diện cấp thấp (API) yêu cầu nhiều dòng mã để tùy chỉnh, có thể phức tạp cho người mới.
* Ưu điểm:
  + Linh hoạt, phù hợp với các nhu cầu tùy chỉnh cao.
  + Tích hợp tốt với NumPy và Pandas.
  + Phù hợp để xuất bản biểu đồ chất lượng cao (ví dụ: cho báo cáo khoa học).
* Nhược điểm:
  + Giao diện mặc định không bắt mắt, cần tùy chỉnh nhiều để đẹp.
  + Ít hỗ trợ tương tác nâng cao so với Plotly.
  + Cần nhiều mã hơn cho các biểu đồ phức tạp.

2. Seaborn

* Loại biểu đồ: Tập trung vào các biểu đồ thống kê như heatmap, violin plot, pair plot, box plot, và scatter plot với các tính năng thống kê tích hợp.
* Ít linh hoạt hơn Matplotlib, nhưng cung cấp các chủ đề (themes) và bảng màu mặc định đẹp mắt.
* Chủ yếu tạo biểu đồ tĩnh, tương tự Matplotlib, nhưng dễ sử dụng hơn.
* Hiệu suất: Tối ưu cho việc trực quan hóa dữ liệu thống kê với Pandas DataFrame.
* Tính dễ sử dụng: Giao diện cấp cao, ít mã hơn Matplotlib, phù hợp cho người mới.
* Ưu điểm:
  + Tạo biểu đồ đẹp mắt với ít mã hơn Matplotlib.
  + Tích hợp tốt với Pandas, lý tưởng cho phân tích dữ liệu thống kê.
  + Hỗ trợ các biểu đồ phức tạp như ma trận tương quan (heatmap) hoặc pair plot dễ dàng.
* Nhược điểm:
  + Ít linh hoạt hơn Matplotlib khi cần tùy chỉnh chi tiết.
  + Không hỗ trợ tốt các biểu đồ tương tác.
  + Phụ thuộc vào Matplotlib, nên đôi khi cần kết hợp với Matplotlib để tùy chỉnh sâu.

3. Plotly

* Hỗ trợ nhiều loại biểu đồ (line, bar, scatter, pie, 3D, bản đồ địa lý, v.v.) và đặc biệt mạnh về biểu đồ tương tác.
* Cung cấp khả năng tùy chỉnh cao, nhưng dễ sử dụng hơn Matplotlib nhờ giao diện cấp cao.
* Tạo biểu đồ tương tác (phóng to, di chuột, lọc dữ liệu) phù hợp cho web và dashboard.
* Hiệu suất: Tốt cho dữ liệu lớn và ứng dụng web, nhưng có thể chậm hơn khi xuất biểu đồ tĩnh.
* Tính dễ sử dụng: Giao diện thân thiện, ít mã hơn Matplotlib, nhưng cần hiểu về cú pháp Plotly (Plotly Express hoặc Graph Objects).
* Ưu điểm:
  + Biểu đồ tương tác, lý tưởng cho dashboard và ứng dụng web.
  + Hỗ trợ xuất ra HTML, dễ dàng chia sẻ hoặc nhúng vào website.
  + Hỗ trợ biểu đồ phức tạp như bản đồ địa lý, biểu đồ 3D.
* Nhược điểm:
  + Phụ thuộc vào kết nối internet cho một số tính năng (trừ khi dùng chế độ offline).
  + Có thể phức tạp hơn khi làm việc với dữ liệu rất lớn hoặc cần xuất bản tĩnh.
  + Yêu cầu cài đặt thêm (Plotly, Orca cho xuất ảnh).

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

1. Tập trung vào sự rõ ràng và đơn giản: Loại bỏ các yếu tố không cần thiết, tập trung vào dữ liệu chính để tránh gây nhầm lẫn hoặc phân tâm.

2. Chọn loại biểu đồ phù hợp với dữ liệu và mục tiêu: Loại biểu đồ phải phù hợp với đặc điểm dữ liệu (phân loại, số, thời gian) và mục đích phân tích (so sánh, phân phối, xu hướng, mối quan hệ).

3. Sử dụng màu sắc hợp lý và có mục đích: Màu sắc nên tăng cường khả năng đọc hiểu, không gây rối mắt hoặc làm lu mờ thông tin.

4. Đảm bảo tính chính xác và trung thực của dữ liệu: Biểu đồ không được làm méo mó hoặc bóp méo dữ liệu, phải phản ánh đúng thông tin.

5. Cung cấp ngữ cảnh và nhãn rõ ràng: Biểu đồ cần có tiêu đề, nhãn trục, chú thích và ngữ cảnh để người xem hiểu rõ ý nghĩa.

### 2.6. 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?

1. Tạo Histogram bằng Matplotlib

- Mục đích: Hiển thị phân phối số lượng của một biến số liên tục.

- Dữ liệu mẫu: Tệp dữ liệu của những người đi khám bệnh tiểu đường (Được lấy từ bài Kaggle: pima indians diabetes - <https://www.kaggle.com/datasets/nancyalaswad90/review>)



Hình : Phần code hiển thị biểu đồ histogram phân phối độ tuổi

A graph of a number of people

AI-generated content may be incorrect.

Hình : Biểu đồ Histogram hiển thị phân phối của độ tuổi

2. Tạo Bar chart bằng Matplotlib

- Mục đích: So sánh số lượng người mắc bệnh tiểu đường và người không mắc bệnh.

- Dữ liệu mẫu: Tệp dữ liệu của những người đi khám bệnh tiểu đường (Được lấy từ bài Kaggle: pima indians diabetes - <https://www.kaggle.com/datasets/nancyalaswad90/review>)

A computer screen shot of a program code

AI-generated content may be incorrect.

Hình : Phần code vẽ biểu đồ cột hiển thị số người mắc hoặc không mắc bệnh tiểu đường

A graph showing the distribution of diabetes outcome

AI-generated content may be incorrect.

Hình : Biểu đồ cột hiển thị số người mắc hoặc không mắc bệnh tiểu đường

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

1. Xuất biểu đồ ra định dạng PNG

* Matplotlib và Seaborn: Sử dụng plt.savefig() để lưu biểu đồ.
* Plotly: Sử dụng fig.write\_image() để xuất ra PNG.

2. Xuất biểu đồ ra định dạng PDF

* Matplotlib và Seaborn: Sử dụng plt.savefig() với định dạng pdf.
* Plotly: Sử dụng fig.write\_image() hoặc fig.write\_html() (chuyển sang PDF gián tiếp).

3. Xuất biểu đồ ra định dạng HTML

* Matplotlib và Seaborn: Không hỗ trợ trực tiếp HTML tương tác, nhưng có thể nhúng vào HTML bằng cách lưu PNG và tích hợp.
* Plotly: Sử dụng fig.write\_html() để tạo file HTML tương tác.

## 3. Phân tích đơn biến và hai biến

### 3.1. 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 là quá trình khám phá và tìm hiểu từng thuộc tính (biến) trong tập dữ liệu một cách độc lập. Mục tiêu chính là để hiểu rõ hình dạng và đặc điểm phân phối của mỗi biến mà không xét đến mối quan hệ của nó với các biến khác.

Sự khác biệt với Phân tích Hai biến (Bivariate Analysis): Phân tích hai biến tập trung vào mối quan hệ giữa hai biến và cách chúng thay đổi cùng nhau. Mục tiêu là để xem liệu có sự tương tác hay liên hệ nào giữa một cặp thuộc tính hay không.

Bảng tóm tắt sự khác biệt chính:

|  |  |  |
| --- | --- | --- |
| Tiêu chí | Phân tích đơn biến | Phân tích hai biến |
| Số lượng biến | Phân tích một biến tại một thời điểm | Phân tích hai biến tại một thời điểm |
| Mục tiêu chính | Mô tả đặc điểm của một biến duy nhất: phân phối, xu hướng trung tâm, độ phân tán, độ lệch, giá trị ngoại lai… | Khám phá mối liên hệ giữa hai biến: tương quan thuận, tương quan nghịch, hoặc không có tương quan |
| Ví dụ | Thuộc tính tuổi được phân bổ như thế nào? hoặc Giá trị trung bình của thuộc tính khối lượng là bao nhiêu? | Thuộc tính bệnh tiểu đường và số lần mang thai có liên quan đến nhau không? |

### 3.2. 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)?

Phân tích đơn biến tập trung vào việc tìm hiểu từng thuộc tính một cách độc lập. Các thước đo thống kê mô tả là công cụ chính để có được cái nhìn sâu sắc về hình dạng của mỗi thuộc tính. Có 2 nhóm thước đo chính:

1. Thống kê mô tả:
   1. Count (Số lượng): Tổng số lượng quan sát (hàng) không bị thiếu dữ liệu.
   2. Mean (Giá trị trung bình): Giá trị trung bình cộng của tất cả các quan sát. Đây là một trong những thước đo bạn đã đề cập.
   3. Standard Deviation (Độ lệch chuẩn): Một thước đo cho thấy mức độ phân tán hoặc lan rộng của dữ liệu so với giá trị trung bình. Đây cũng là một thước đo bạn đã đề cập.
   4. Minimum Value (Giá trị nhỏ nhất): Giá trị nhỏ nhất trong tập dữ liệu của thuộc tính đó.
   5. 25th Percentile (Phân vị thứ 25): Giá trị mà tại đó 25% dữ liệu có giá trị nhỏ hơn hoặc bằng nó.
   6. 50th Percentile (Median - Trung vị): Đây chính là trung vị, là giá trị nằm ở giữa của tập dữ liệu đã được sắp xếp. 50% dữ liệu có giá trị nhỏ hơn hoặc bằng giá trị này. Đây là thước đo thứ ba bạn đã đề cập.
   7. 75th Percentile (Phân vị thứ 75): Giá trị mà tại đó 75% dữ liệu có giá trị nhỏ hơn hoặc bằng nó.
   8. Maximum Value (Giá trị lớn nhất): Giá trị lớn nhất trong tập dữ liệu của thuộc tính đó.
2. Độ lệch của Phân phối (Skewness): Đây là một thước đo thống kê quan trọng khác được sử dụng trong phân tích đơn biến để hiểu về hình dạng phân phối của một thuộc tính.
   1. Độ xiên đề cập đến việc một phân phối bị dịch chuyển hoặc bị nén về một hướng nào đó.
   2. Việc biết một thuộc tính có độ xiên có thể cho phép bạn thực hiện các bước chuẩn bị dữ liệu để điều chỉnh độ xiên, từ đó cải thiện độ chính xác của mô hình.
   3. Kết quả độ xiên cho thấy độ xiên dương (bên phải) hoặc âm (bên trái), trong đó các giá trị gần bằng 0 cho thấy độ xiên ít hơn.

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

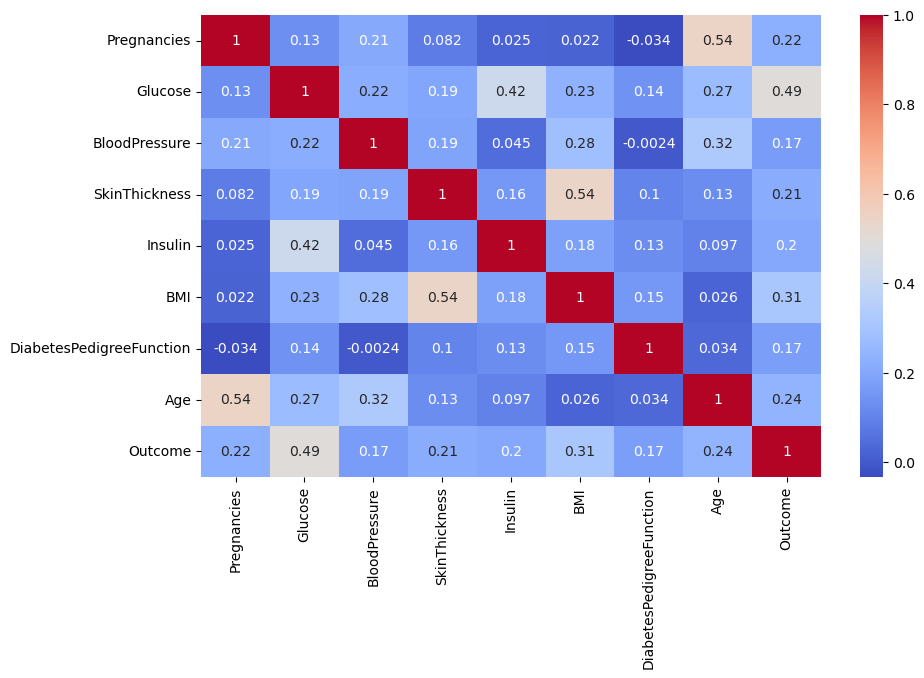
Tương quan (Correlation) đề cập đến mối quan hệ giữa hai biến và cách chúng có thể thay đổi cùng nhau hay không. Việc hiểu các mối quan hệ này là một trong các bước quan trọng để đạt được kết quả tốt nhất từ dữ liệu học máy.

1. Hệ số Tương quan Pearson (Pearson's Correlation Coefficient): Là dữ liệu cho biết được độ lớn và hướng tuyến tính giữa hai biến dữ liệu liên tục.

Đánh giá Hệ số: Hệ số tương quan nằm trong khoảng từ -1 hoặc 1

* + Giá trị gần 1 cho thấy mối tương quan thuận mạnh => Khi một giá trị của biến này tăng, giá trị của biến kia cũng tăng theo
  + Giá tri gần -1 chỉ tương quan nghịch mạnh => Khi một giá trị của biến này tăng, giá trị biến kia giảm theo
  + Giá trị gần 0 cho thấy mối quan hệ tuyến tính yếu hoặc không có => Khi một giá trị của biến này tăng, giá trị biến kia không thay đổi

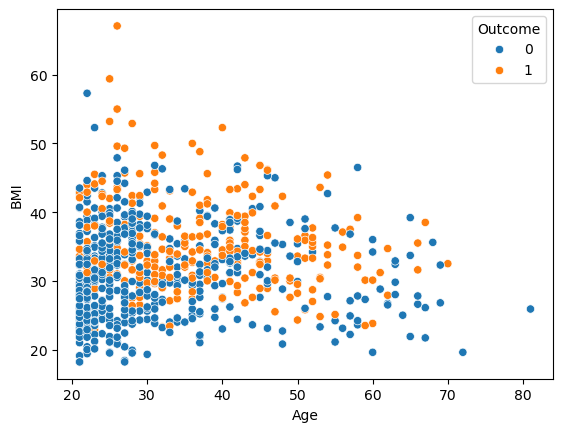
1. Ma trận Tương quan (Correlation Matrix): Ma trận tương quan liệt kê tất cả các thuộc tính dọc theo hàng ngang và hàng dọc, cung cấp tương quan giữa tất cả các cặp thuộc tính. Ma trận này có tính đối xứng. Đường chéo chạy từ góc trên bên trái xuống góc dưới bên phải của ma trận cho thấy mối tương quan hoàn hảo của mỗi thuộc tính với chính nó.



Hình : Ma trận tương quan của tệp dữ liệu tổng hợp các bệnh nhân/người đi khám bệnh tiểu đường

Ở đây ta có thể thấy, hệ số tương quan giữa độ tuổi (Age) và chỉ số BMI chỉ ở mức 0.026, rất thấp, cho thấy gần như hai kiểu dữ liệu này không có mối quan hệ tuyến tính.

1. Biểu đồ Phân tán (Scatter Plot): là một loại biểu đồ thể hiện mối quan hệ giữa hai biến định lượng bằng các điểm dữ liệu trên hệ tọa độ Descartes, trong đó trục hoành (Ox) biểu thị biến độc lập và trục tung (Oy) biểu thị biến phụ thuộc. Biểu đồ này giúp phát hiện xu hướng, tương quan (dương, âm, hoặc không có) và các điểm bất thường trong dữ liệu, từ đó hỗ trợ việc phân tích và đưa ra quyết định.



Hình : Scatter plot của tập dữ liệu của bệnh nhân/người đi khám bệnh tiểu đường thể hiện mối tương quan giữa độ tuổi (Age) và chỉ số BMI

Ta có thể thấy biểu đồ không thể hiện được sự tuyến tính giữa hai giá trị này hết, phản ánh được chỉ số tuyến tính đã được tính toán ở trên.

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

Covariance cho biết hướng mối quan hệ giữa hai biến (có xu hướng tăng/giảm cùng nhau hay ngược chiều nhau). Nếu giá trị dương, hai biến tăng hoặc giảm cùng chiều. Nếu âm, khi một biến tăng thì biến còn lại giảm. Tuy nhiên, giá trị của covariance phụ thuộc vào đơn vị đo của hai biến nên khó so sánh trực tiếp giữa các cặp biến khác nhau.

Correlation là hệ số tương quan, đo lường cả hướng và mức độ mạnh yếu của mối quan hệ, và đã được chuẩn hóa nên luôn nằm trong khoảng từ -1 đến 1. Giá trị này giúp đánh giá chính xác hơn và so sánh được giữa các cặp biến bất kể đơn vị đo lường gốc. Do đã được chuẩn hóa, nên ta thường dùng correlation làm hệ số để so sánh thay vì dùng covariance.

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

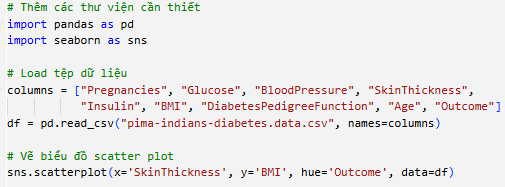
Dùng biểu đồ cho phân tích đơn biến khi:

* Sử dụng để thể hiện phân phối, tần suất hoặc tỷ lệ của một biến duy nhất.
* Thường dùng khi cần mô tả tổng quan, phát hiện xu hướng hoặc bất thường (outlier) của biến, hoặc kiểm tra giả định phân phối.
* Các loại biểu đồ thường dùng: histogram (biến liên tục), bar chart (biến phân loại), pie chart (tỷ trọng).

Dùng biểu đồ cho phân tích hai biến khi:

* Sử dụng để thể hiện mối quan hệ, ảnh hưởng của hai biến lên nhau.
* Áp dụng khi muốn kiểm tra tương quan, quan hệ nhân quả, hoặc so sánh sự thay đổi giữa hai biến.
* Các biểu đồ phổ biến: scatter plot (liên tục – liên tục), boxplot (loại – liên tục), stacked bar/clustered bar (loại – loại).

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

- Dữ liệu mẫu: Tệp dữ liệu của những người đi khám bệnh tiểu đường (Được lấy từ bài Kaggle: pima indians diabetes - <https://www.kaggle.com/datasets/nancyalaswad90/review>)

Hình : Code tạo scatter plot giữa chỉ số BMI và SkinThickness lấy từ tập dữ liệu tiểu đường

A graph of blue and orange dots

AI-generated content may be incorrect.

Hình : Scatter plot giữa chỉ số BMI và SkinThickness của tập dữ liệu tiểu đường

### 3.7. 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?

- Dữ liệu mẫu: Tệp dữ liệu của những người đi khám bệnh tiểu đường (Được lấy từ bài Kaggle: pima indians diabetes - <https://www.kaggle.com/datasets/nancyalaswad90/review>)

A computer screen shot of a computer code

AI-generated content may be incorrect.

Hình : Code tạo boxplots biểu thị mức độ tương quan giữa Glucose và Outcome

A diagram of a group of blue squares

AI-generated content may be incorrect.

Hình : Boxplots biểu thị mức độ tương quan giữa Glucose và Outcome

**TÀI LIỆU THAM KHẢO**

[1]. MLM - 01 - EDA.pdf

[2]. ML04c\_MissingValues.pdf

[3]. ML03\_EDA.pdf

[4]. MLM - 02 - Data Preprocessing.pdf

[5]. MLM - 03 - Sample Projects.pdf

[6]. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow 3rd - Aurélien Géron - O'Reilly - 2022