**Thành viên:**

Huỳnh Anh Nhựt

Nguyễn Tiến Minh

**Ôn tập lý thuyết**

**1. Thống kê mô tả và Thống kê suy luận**

Thống kê mô tả: Tóm tắt, trình bày dữ liệu đã có (trung bình, trung vị, histogram, boxplot). Giúp hiểu đặc điểm của tập dữ liệu hiện tại

Thống kê suy luận: Sử dụng dữ liệu mẫu để suy ra hoặc dự đoán cho toàn bộ quần thể thông qua kiểm định giả thuyết, ước lượng, hồi quy

Khác biệt: Mô tả = kể lại dữ liệu hiện tại, Suy luận = dự đoán cho tổng thể

* **Các thước đo thống kê mô tả chính**

Trung bình (Mean): Giá trị đại diện của dữ liệu

Trung vị (Median): Giá trị ở giữa khi sắp xếp dữ liệu, ít bị ảnh hưởng bởi ngoại lai

Phương sai (Variance): Đo độ biến động của dữ liệu

Độ lệch chuẩn (Standard deviation): Đo mức phân tán so với trung bình

Khi dữ liệu có ngoại lai hoặc lệch mạnh thì nên dùng trung vị thay vì trung bình

* **Xác định phân bố dữ liệu**

Có thể xác định bằng cách vẽ histogram, density plot, hoặc dùng các chỉ số skewness, kurtosis

Có thể thực hiện kiểm định chuẩn (Shapiro-Wilk, Kolmogorov-Smirnov)

Các phân bố phổ biến:

Phân bố chuẩn (Normal): Chuông cân đối

Lệch trái (Left-skewed): Đuôi dài bên trái

Lệch phải (Right-skewed): Đuôi dài bên phải

Phân bố đồng đều (Uniform): Các giá trị xuất hiện gần như ngang nhau

* **Độ lệch chuẩn và Phạm vi**

Range (Max – Min): Độ rộng của dữ liệu, nhưng chỉ dựa trên 2 giá trị biên và dễ bị ảnh hưởng bởi ngoại lai

Standard deviation: Đo mức phân tán dựa trên toàn bộ dữ liệu, phản ánh tốt hơn mức biến động

* **Q1, Q2, Q3 trong Boxplot**

Q1 (Quartile 1): Phân vị 25%, 25% dữ liệu nhỏ hơn Q1

Q2 (Quartile 2 = Median): Phân vị 50%

Q3 (Quartile 3): Phân vị 75%

IQR = Q3 – Q1, thể hiện mức độ phân tán của phần lớn dữ liệu

* **Xử lý giá trị thiếu (Missing Values)**

Loại bỏ dòng hoặc cột chứa missing nếu số lượng ít

Thay thế bằng trung bình, trung vị hoặc mode

Ước lượng bằng mô hình dự đoán (Regression, KNN imputation, MICE)

Giữ nguyên nếu mô hình phân tích có thể xử lý missing

* **Đọc biểu đồ Histogram và Boxplot**

Histogram: Cho thấy dữ liệu phân bố thế nào (chuẩn, lệch trái/phải, có nhiều hay ít đỉnh). Ví dụ: Histogram thu nhập có đuôi dài bên phải cho thấy nhiều người thu nhập thấp, ít người thu nhập rất cao

Boxplot: Hộp thể hiện khoảng IQR, đường giữa là median, râu thể hiện phạm vi dữ liệu không ngoại lai, các điểm ngoài là outliers. Giúp so sánh nhanh phân bố giữa nhiều nhóm

* **Xử lý Outliers trước khi thống kê mô tả**

Giữ nguyên nếu outlier có ý nghĩa thực (ví dụ: doanh thu cực cao)

Loại bỏ nếu chắc chắn đó là lỗi nhập liệu

Biến đổi dữ liệu bằng log transform hoặc winsorization

Phân tích riêng biệt nhóm có outlier để so sánh

1. **Thống kê mô tả**

**a. Giới thiệu Dataset**

Nguồn: UCI Machine Learning Repository.

Mục tiêu: Dự đoán bệnh tiểu đường dựa trên các chỉ số y tế.

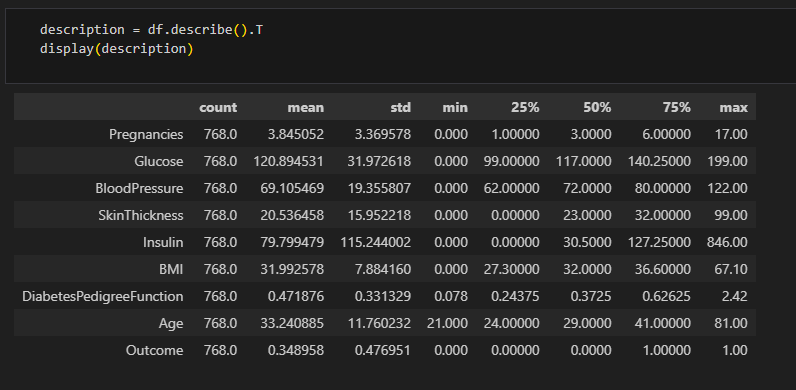
Số dòng (instances): 768

Số cột (features): 9 (8 input, 1 output - “Outcome”)

Các thuộc tính:

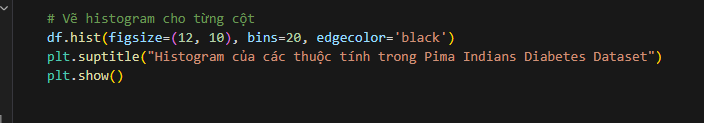
1. Pregnancies – số lần mang thai
2. Glucose – nồng độ glucose huyết tương sau 2h trong oral glucose tolerance test
3. BloodPressure – huyết áp tâm trương (mm Hg)
4. SkinThickness – độ dày nếp gấp da (mm)
5. Insulin – nồng độ insulin trong huyết thanh (mu U/ml)
6. BMI – chỉ số khối cơ thể (kg/m²)
7. DiabetesPedigreeFunction – chức năng phả hệ tiểu đường (nguy cơ di truyền)
8. Age – tuổi
9. Outcome – 0 (không tiểu đường), 1 (tiểu đường)

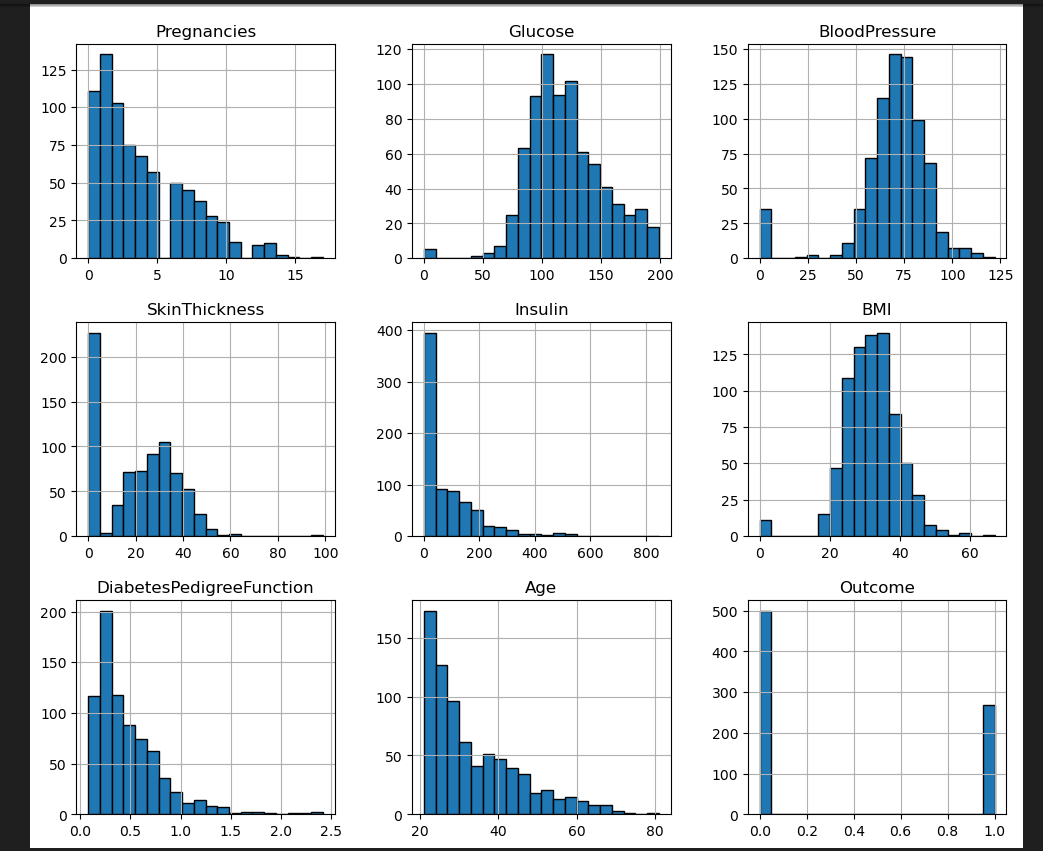
Sử dụng df.describe() sẽ ra **mean, std, min, 25%, 50%, 75%, max** cho từng biến.



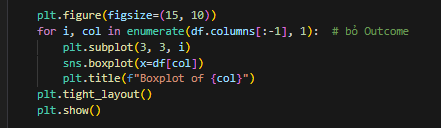
1.2.3.a

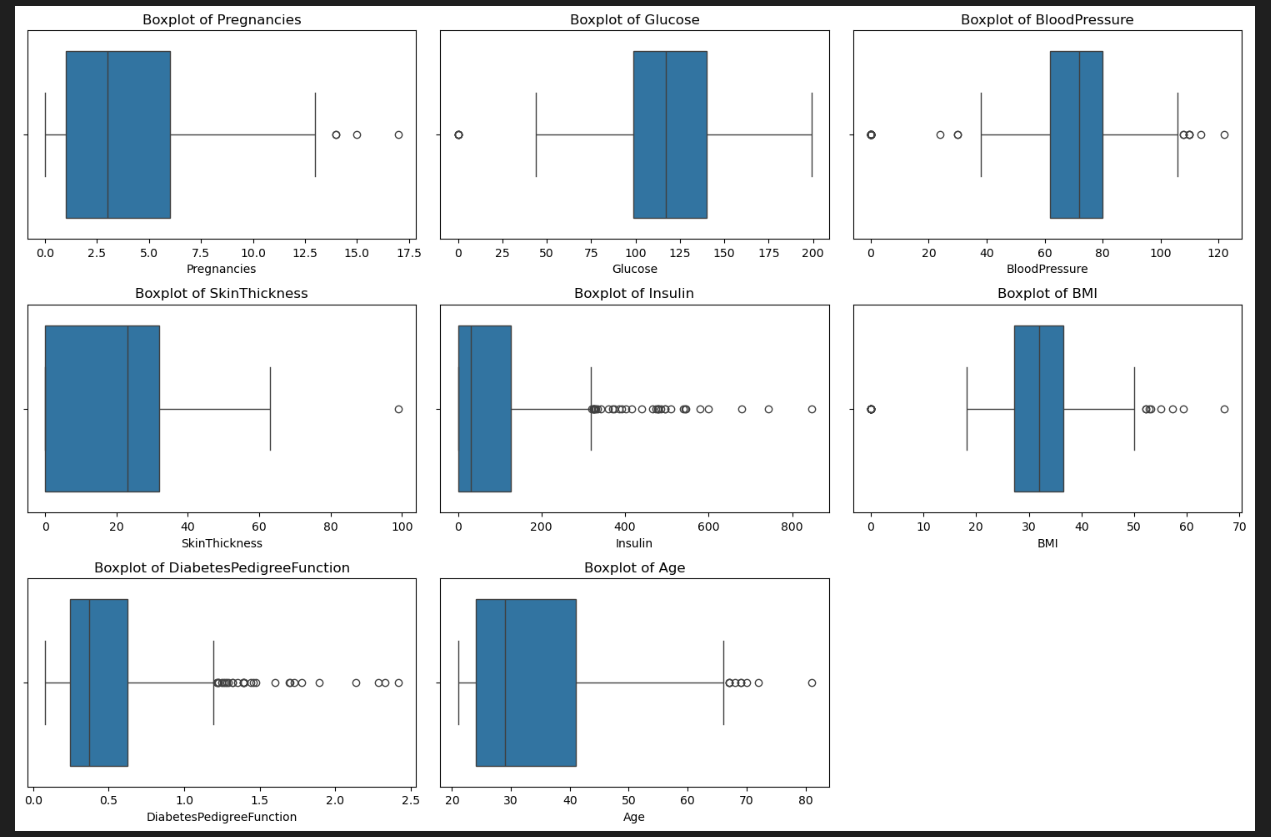
1.Phân bố dữ liệu (Histogram)



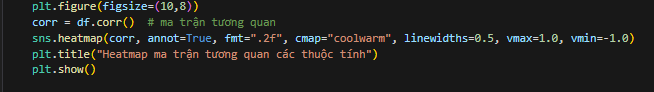


2.Kiểm tra outlier (Boxplot)



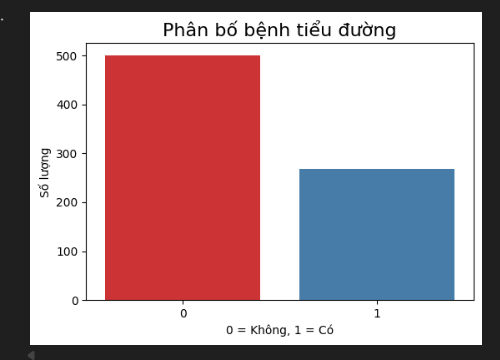
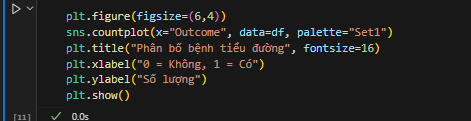


**3. Tương quan giữa các biến (Heatmap)**

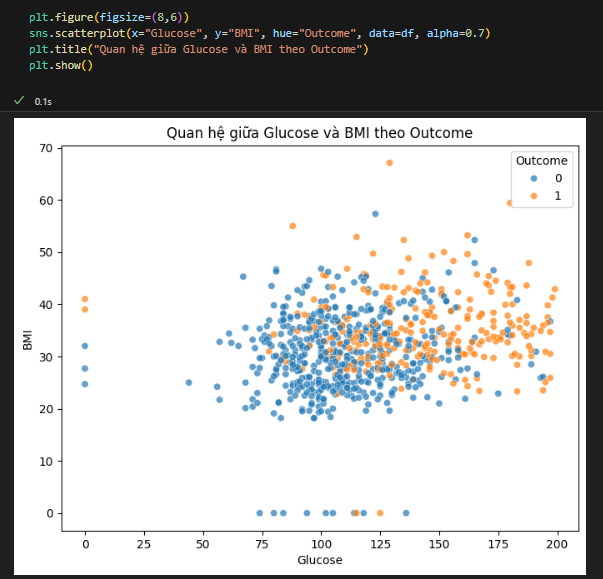




4. Phân bố biến mục tiêu (Outcome) dùng countplot

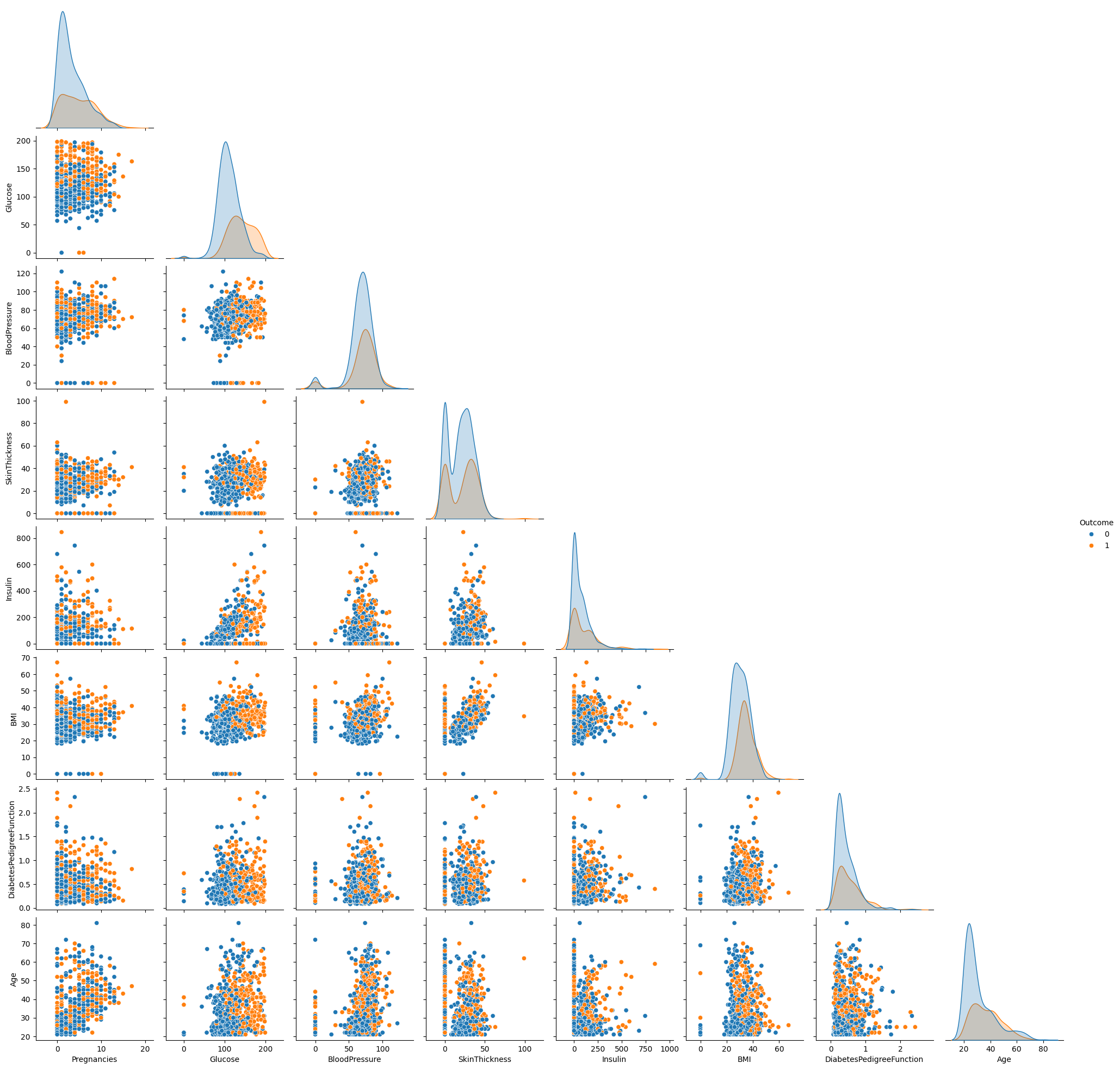


5.Dùng scatter plot tìm mối quan hệ giữa 2 biến



6. Pairplot để xem toàn cảnh





1.2.1 Ôn tập lý thuyết

**- 1. Vai trò của trực quan hóa dữ liệu trong phân tích dữ liệu**

Giúp nhìn thấy bức tranh tổng quan về dữ liệu một cách nhanh chóng

Hỗ trợ phát hiện xu hướng, mẫu, mối quan hệ và ngoại lai trong dữ liệu

Là công cụ quan trọng trong giai đoạn khám phá dữ liệu (EDA) vì con người dễ hiểu trực quan hơn là qua số liệu thô

Giúp truyền đạt kết quả phân tích dễ dàng, trực tiếp hơn cho người khác

* **2. Các loại biểu đồ phổ biến và khi nào sử dụng**

Histogram: Dùng cho dữ liệu số, để quan sát phân bố và tần suất

Scatter plot: Dùng để xem mối quan hệ giữa hai biến số, phát hiện tương quan hoặc ngoại lai

Boxplot: Dùng để xem phân bố, phát hiện outliers và so sánh giữa nhiều nhóm

Bar chart: Dùng cho dữ liệu phân loại, so sánh tần suất hoặc giá trị trung bình giữa các nhóm

* **3. Cách chọn loại biểu đồ phù hợp**

Dữ liệu phân loại: Bar chart, pie chart (ít dùng hơn)

Dữ liệu số: Histogram, boxplot, scatter plot

Dữ liệu thời gian: Line chart để theo dõi xu hướng theo thời gian

Dữ liệu nhiều biến: Heatmap hoặc scatter matrix để xem mối quan hệ giữa nhiều biến cùng lúc

* **4. Sự khác biệt giữa Matplotlib, Seaborn và Plotly**

Matplotlib: Thư viện nền tảng, mạnh mẽ nhưng code dài, nhiều tùy chỉnh thủ công

Seaborn: Xây dựng trên Matplotlib, cú pháp gọn hơn, dễ vẽ biểu đồ thống kê, có sẵn theme đẹp

Plotly: Hỗ trợ biểu đồ tương tác, zoom, hover, xuất HTML, phù hợp cho dashboard và trình bày kết quả trực quan động

* **5. Nguyên tắc thiết kế biểu đồ trực quan hiệu quả**

Rõ ràng, dễ hiểu, không nhồi nhét quá nhiều thông tin

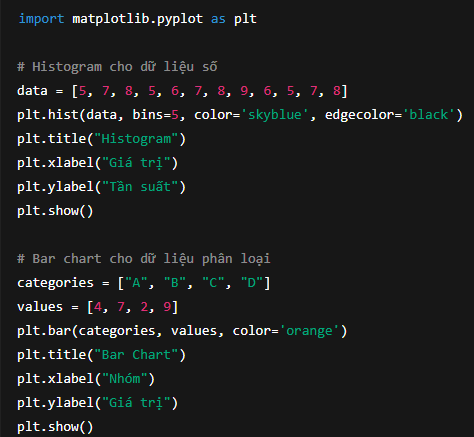
Chọn loại biểu đồ đúng với dữ liệu

Tránh dùng quá nhiều màu sắc gây rối mắt

Có tiêu đề, nhãn trục rõ ràng, chú thích khi cần thiết

Nhấn mạnh thông tin chính, hạn chế chi tiết không cần thiết

* 6. Ví dụ code tạo histogram và bar chart bằng Matplotlib



* **7. Xuất biểu đồ ra PNG, PDF, HTML**

Với Matplotlib: dùng plt.savefig("tenfile.png") hoặc plt.savefig("tenfile.pdf")

Với Plotly: dùng fig.write\_html("tenfile.html") để xuất ra HTML tương tác

Ví dụ với Matplotlib:

plt.savefig("bieu\_do.png", dpi=300) #xuat png

plt.savefig("bieu\_do.pdf") #xuat pdf

1.3.1. Ôn lý thuyết

**- Phân tích đơn biến (Univariate analysis) và phân tích hai biến (Bivariate analysis)**

Phân tích đơn biến: Xem xét, mô tả, tóm tắt **một biến duy nhất** để hiểu đặc điểm của nó (trung bình, trung vị, histogram, boxplot...)

Phân tích hai biến: Xem xét mối quan hệ giữa **hai biến** (ví dụ: chiều cao và cân nặng, điểm số và số giờ học). Có thể dùng scatter plot, hệ số tương quan, hồi quy tuyến tính...

Khác biệt: Đơn biến tập trung vào **tính chất của một biến**, hai biến tập trung vào **mối quan hệ giữa hai biến**

* **Các thước đo thống kê thường dùng trong phân tích đơn biến**

Trung bình (Mean)

Trung vị (Median)

Mode (giá trị xuất hiện nhiều nhất)

Độ lệch chuẩn (Standard deviation)

Phương sai (Variance)

Phạm vi (Range = Max – Min)

Quartiles (Q1, Q2, Q3) và IQR (Q3 – Q1)

* **Xác định mối quan hệ giữa hai biến**

Dùng biểu đồ scatter plot để quan sát xu hướng

Tính hệ số tương quan (Pearson, Spearman) để đo mức độ và chiều hướng mối quan hệ tuyến tính

Kiểm định thống kê (ví dụ kiểm định Chi-square cho biến phân loại)

Dùng hồi quy để kiểm tra quan hệ nhân quả, không chỉ dừng lại ở tương quan

* **Tương quan (Correlation) và Hiệp biến (Covariance)**

Covariance: Đo mức độ hai biến thay đổi cùng nhau. Nếu cùng tăng thì dương, một tăng một giảm thì âm. Tuy nhiên giá trị khó so sánh vì phụ thuộc đơn vị đo lường

Correlation: Chuẩn hóa từ covariance, giá trị nằm trong [-1, 1], dễ so sánh. 1 = tương quan dương hoàn hảo, -1 = tương quan âm hoàn hảo, 0 = không có mối quan hệ tuyến tính

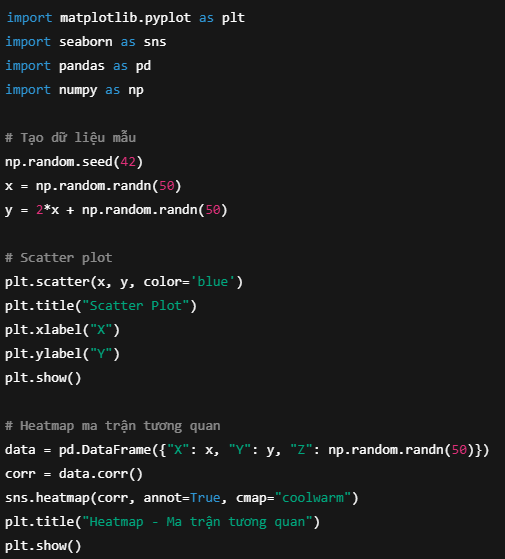
Khác biệt: Covariance cho biết hướng thay đổi, correlation cho biết cả hướng lẫn mức độ và có thang đo chuẩn hóa

* **Khi nào dùng biểu đồ trực quan trong phân tích đơn biến và hai biến**

Đơn biến: Dùng histogram, boxplot, bar chart → để hiểu phân bố, độ lệch, ngoại lai

Hai biến: Dùng scatter plot, heatmap, boxplot nhóm theo biến phân loại → để thấy mối quan hệ, xu hướng, sự khác biệt giữa các nhóm

* 6. Code mẫu Scatter plot và Heatmap trong Python



* 7. Trực quan hóa mối quan hệ giữa biến số và biến phân loại

