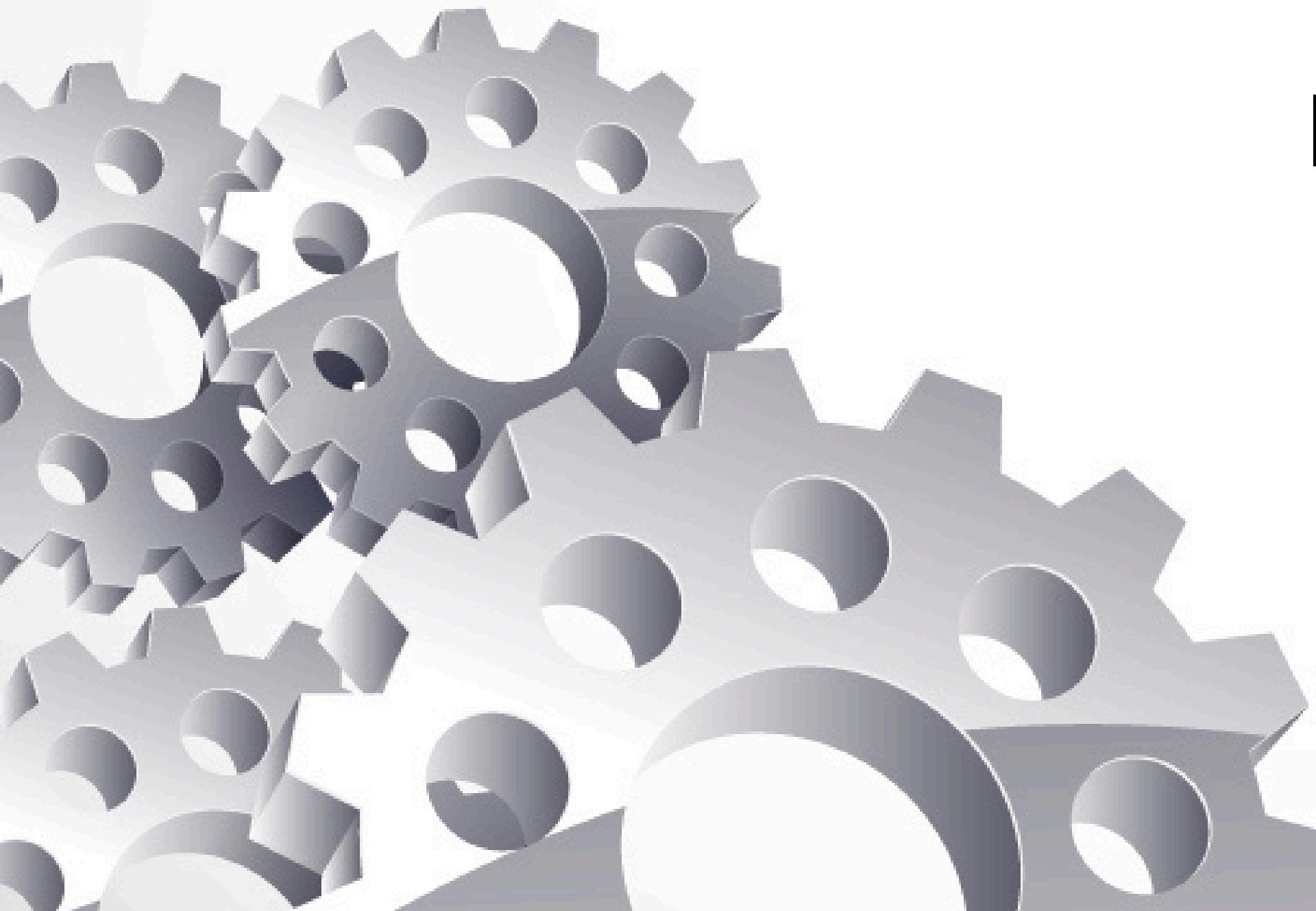


Exploratory Data Analysis

Pima Indians Diabetes



Bảng phân công

Họ tên	Công việc
Nguyễn Đăng Khoa	Tìm hiểu thông tin, code, ppt
Lê Thị Trúc Ly	Tìm thông tin, code, làm slide thuyết trình
Lê Đoàn Kim Ngân	Tìm hiểu, tổng hợp thông tin, code
Lâm Tú Nhi	Tìm thông tin, code, làm slide ppt

Giới thiệu

- Tập dữ liệu tiểu đường Pima Indian dùng để dự đoán sự khởi phát của bệnh tiểu đường dựa trên các biện pháp chẩn đoán.
- Nguồn gốc từ Viện Quốc gia về Bệnh tiểu đường, Tiêu hóa và Bệnh thận.
- Mục tiêu: Dự đoán xem bệnh nhân có bị tiểu đường hay không dựa trên các phép đo chẩn đoán.

Các điểm cần thảo luận:

- Tóm tắt dữ liệu
- Phát hiện giá trị bất thường & ngoại lệ
- Phân tích đơn biến
- Phân tích theo kết quả
- Phân tích theo độ tuổi
- Phân tích theo số lần mang thai
- Phân tích theo BMI
- Phân tích theo Glucose
- Một số câu hỏi quan trọng
- Bản đồ nhiệt tương quan
- Kết luận

Tóm tắt dữ liệu

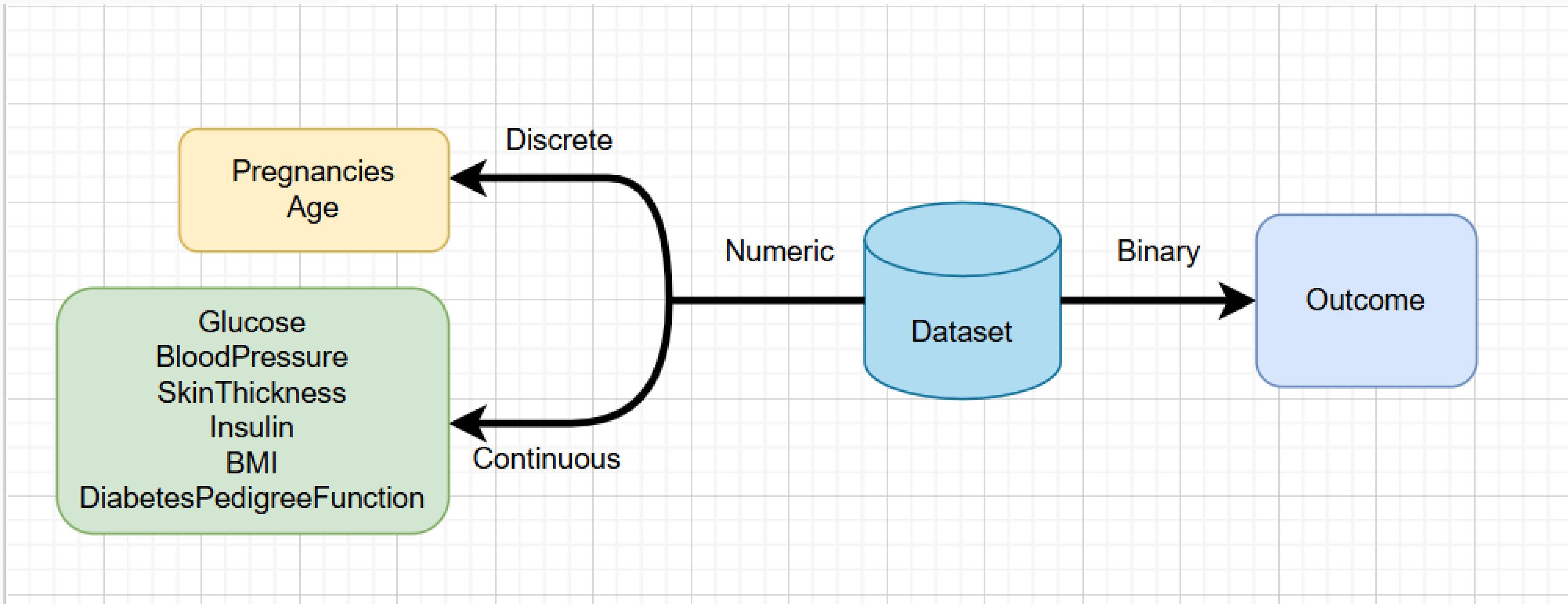
Tập dữ liệu cho trước gồm hồ sơ cá nhân của các bệnh nhân nữ, các hồ sơ đều có 9 thuộc tính như sau:

- **Pregnancies**: Số lần mang thai của bệnh nhân.
- **Plasma glucose concentration (Glucose)** : Nồng độ Glucose trong huyết tương sau 2 giờ trong xét nghiệm dung nạp glucose đường uống.
- **BloodPressure** : Huyết áp tâm trương (mức áp lực của máu tác động lên thành động mạch khi tim đang giãn ra giữa hai nhịp đập).

Tóm tắt dữ liệu

- **SkinThickness** : Độ dày nếp gấp da (đây là phép đo độ dày lớp mỡ dưới da ở vùng cơ phía sau cánh tay).
- **Insulin**: Nồng độ insulin huyết thanh đo sau 2 giờ.
- **BMI (Body Mass Index)** : Chỉ số khối cơ thể.
- **DiabetesPedigreeFunction** : Chỉ số “di truyền tiểu đường” dựa trên tiền sử gia đình.
- **Age** : Tuổi của bệnh nhân.
- **Outcome** : Kết quả chuẩn đoán, 1 nếu mắc bệnh tiểu đường, 0 nếu không mắc bệnh.

Tóm tắt dữ liệu



Kiểm tra missing & zero values

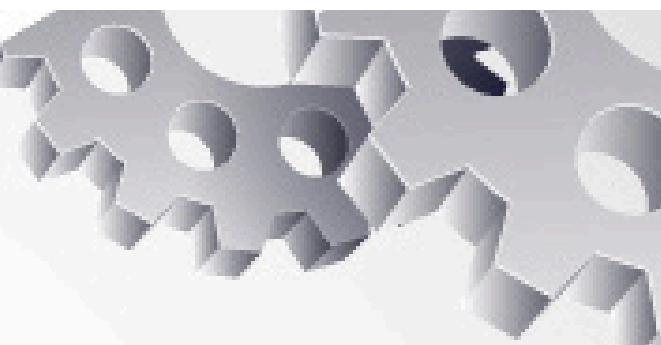


```
# Kiểm tra missing & zero values
cols_with_zero = ["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]
print("Missing values trước khi xử lý:")
print(df.isnull().sum())

df[cols_with_zero] = df[cols_with_zero].replace(0, np.nan)

print("zero values trước khi xử lý:")
print(df.isnull().sum())
```

Kiểm tra missing & zero values



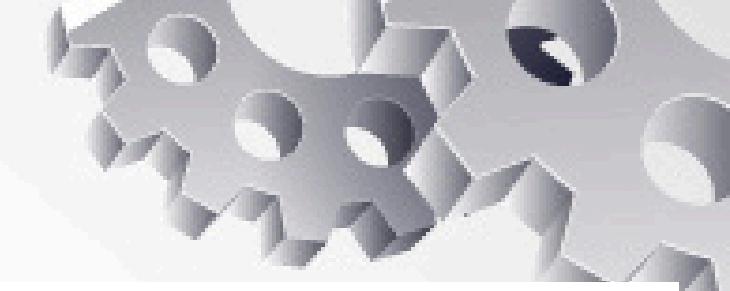
Missing values trước khi xử lý:

```
Pregnancies          0  
Glucose             0  
BloodPressure       0  
SkinThickness       0  
Insulin             0  
BMI                 0  
DiabetesPedigreeFunction 0  
Age                 0  
Outcome             0  
dtype: int64
```

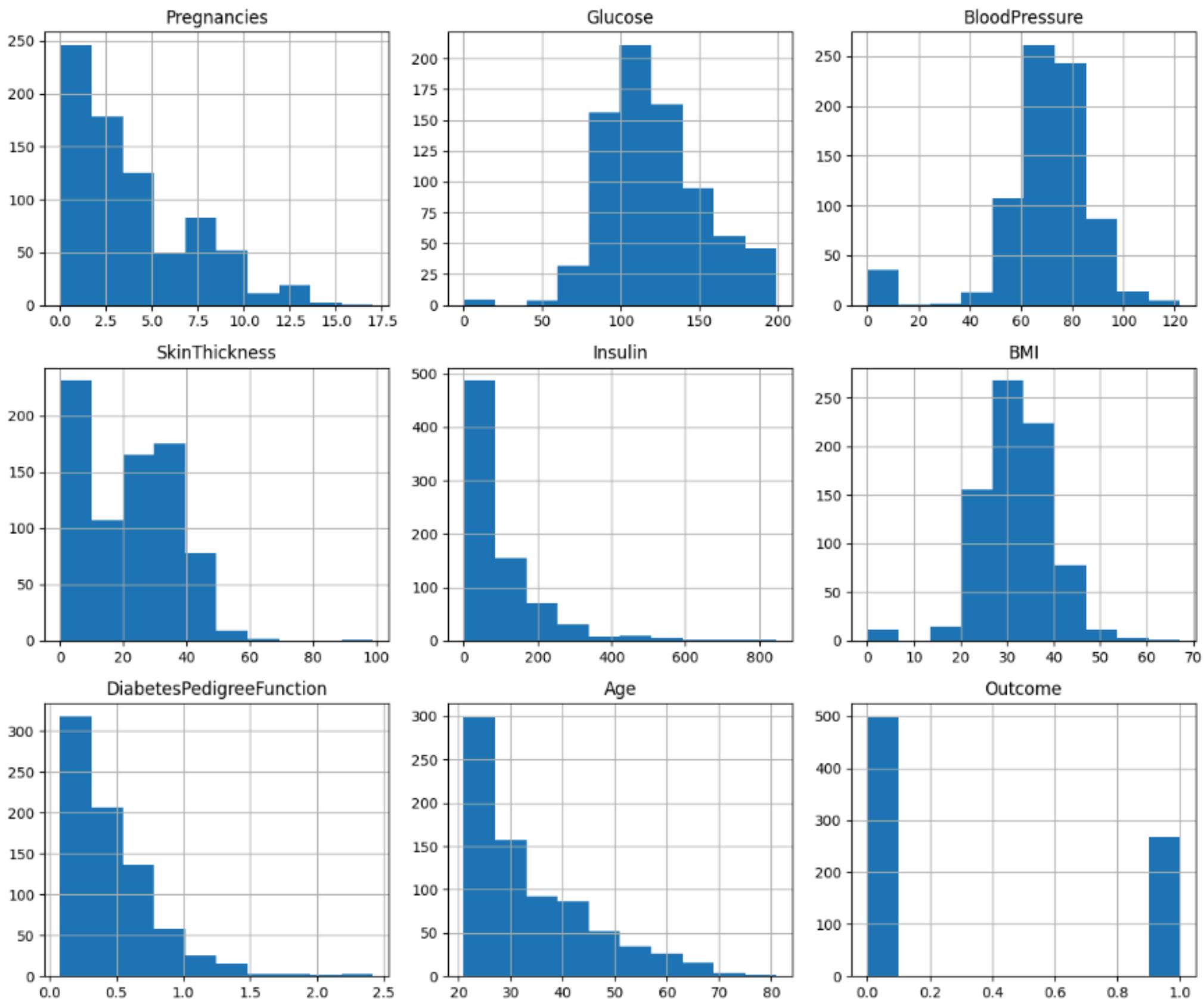
Zero values trước khi xử lý:

```
Pregnancies          0  
Glucose              5  
BloodPressure        35  
SkinThickness        227  
Insulin              374  
BMI                 11  
DiabetesPedigreeFunction 0  
Age                 0  
Outcome              0  
dtype: int64
```

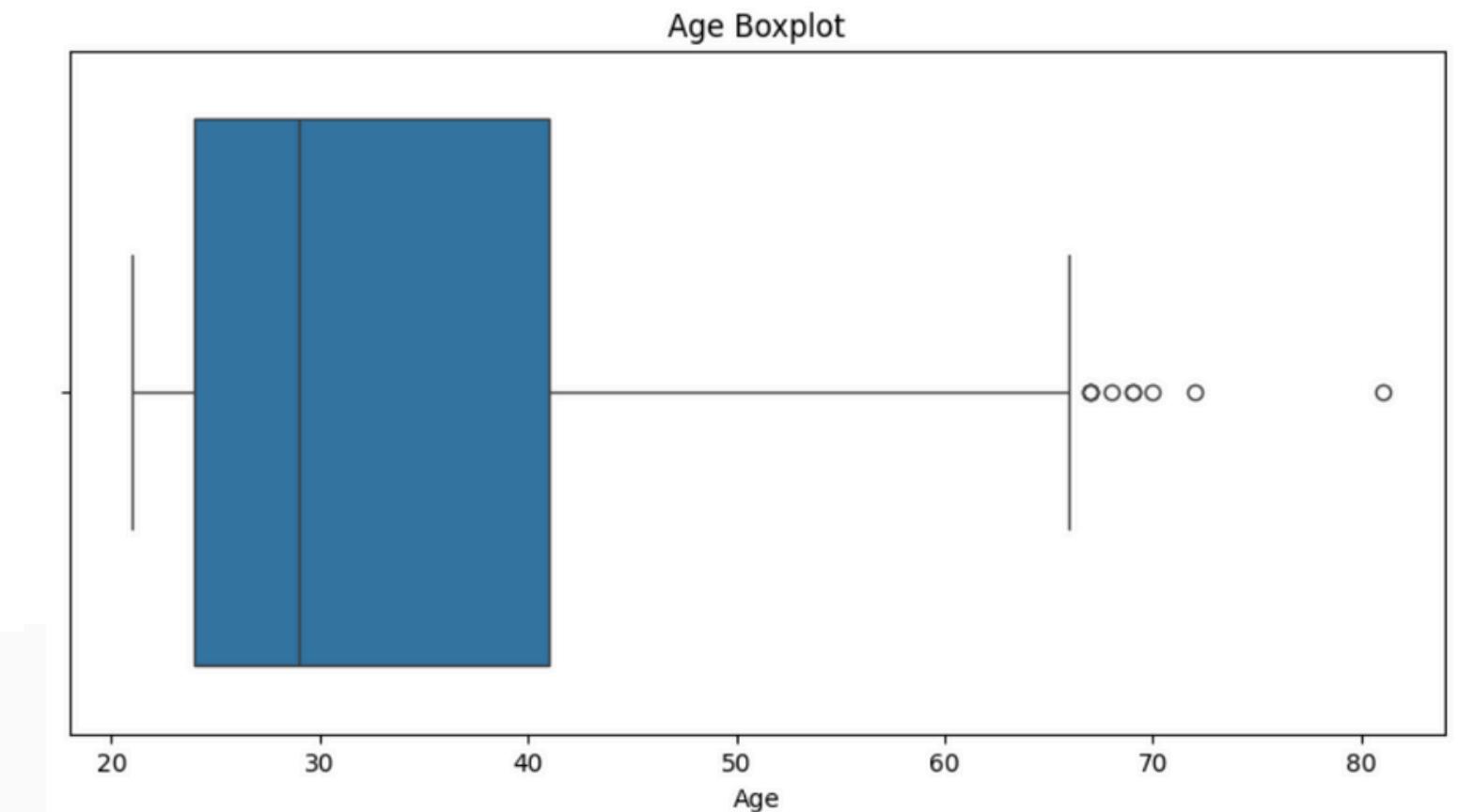
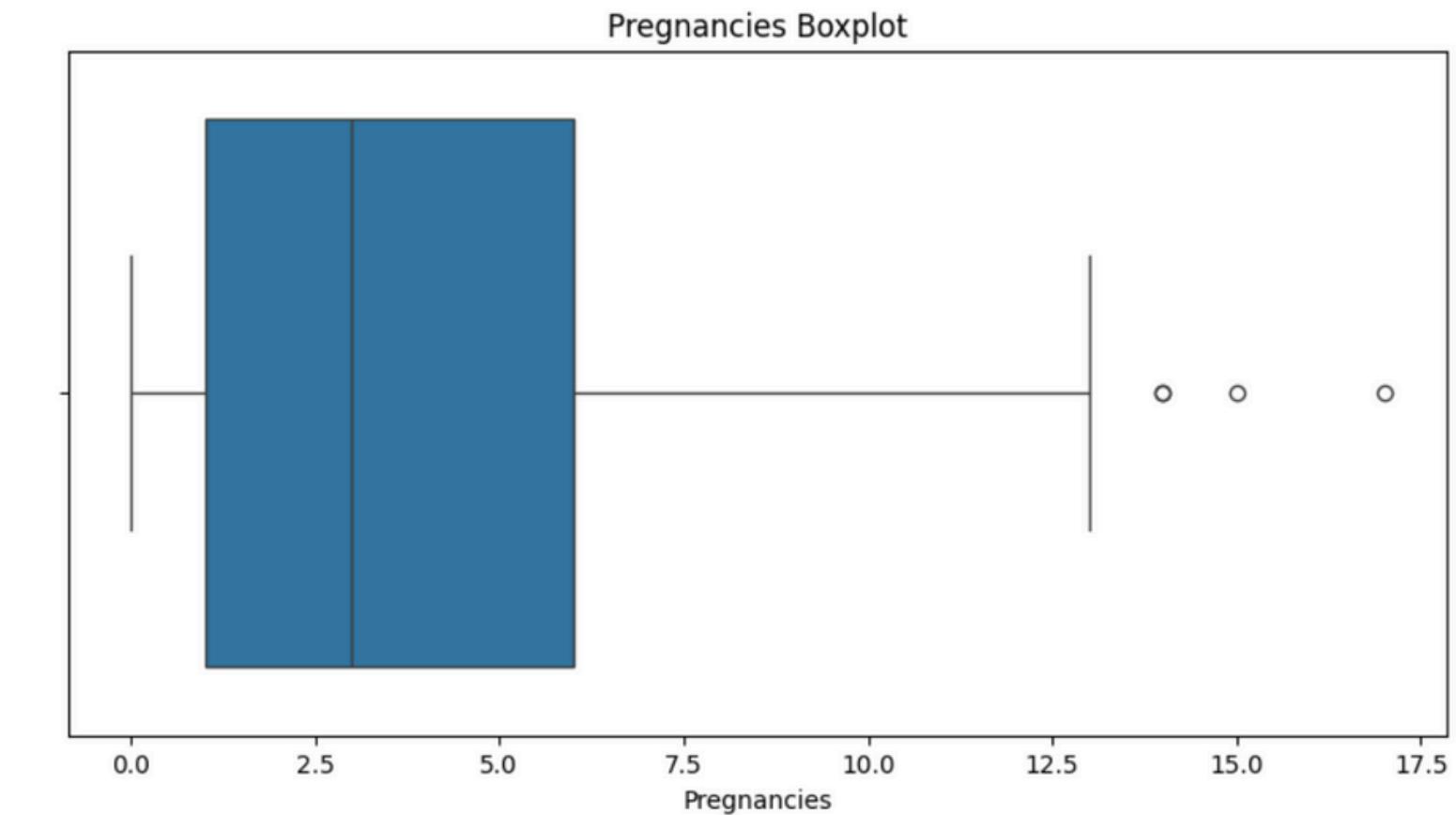
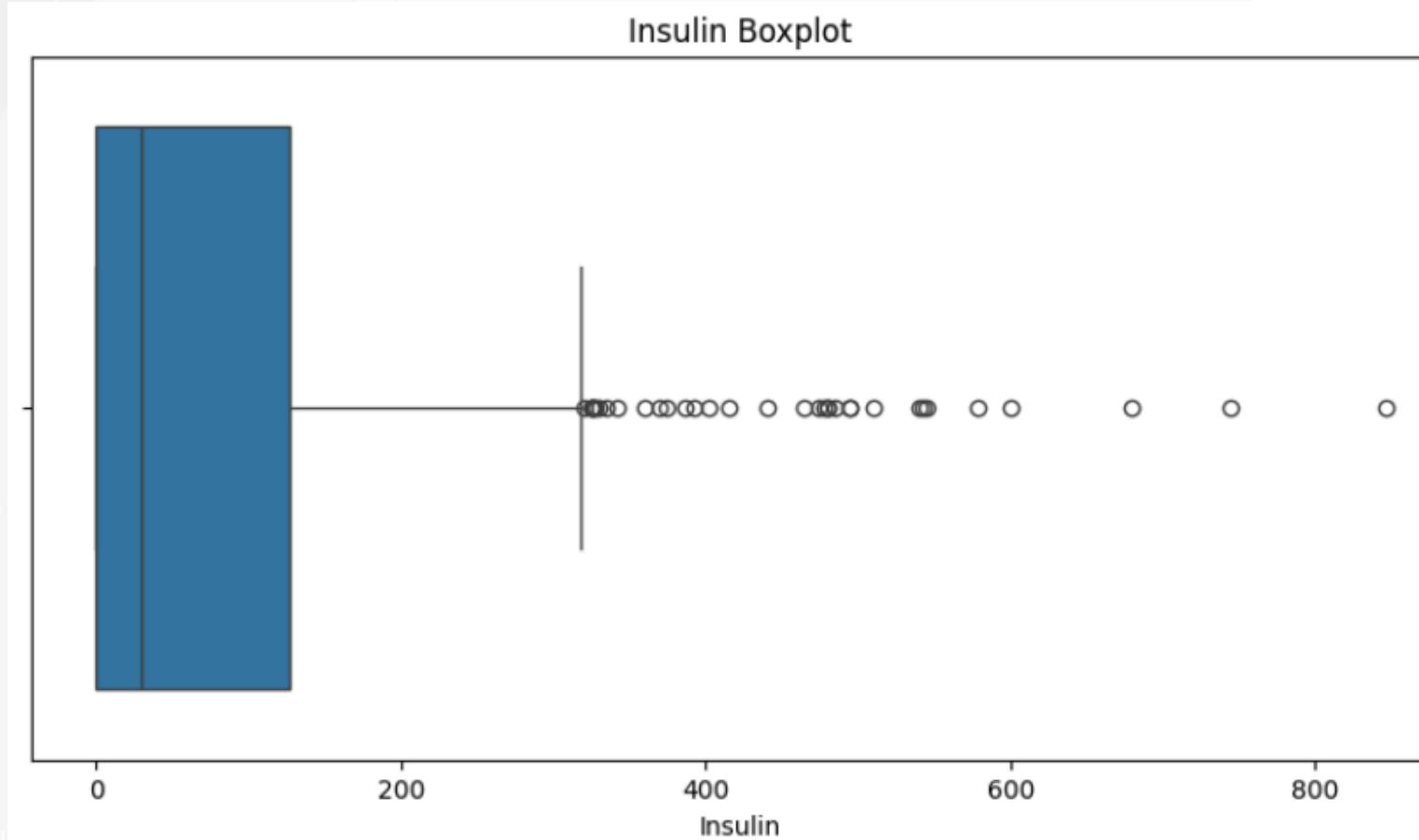
Phát hiện giá trị bất thường & ngoại lệ



- Các thuộc tính Glucose, BloodPressure, SkinThickness, Insulin, BMI có nhiều giá trị bằng 0, các giá trị này có thể được xem là **missing values**.
- Nhiều thuộc tính có phân bố **lệch phải** như: Pregnancies, Insulin, DiabetesPedigreeFunction, Age.
- Dữ liệu **imbalanced** ở biến mục tiêu Outcome (tỉ lệ mắc bệnh khoảng 35%).



Phát hiện giá trị bất thường & ngoại lệ



Phát hiện 1 số **outliers** như:

- *Insulin* có rất nhiều giá trị ngoại lệ lớn (> 600)
- *Pregnancies* và *Age* cũng có những giá trị cực đoan như: mang thai 17 lần, hơn 80 tuổi.

Phân tích đơn biến

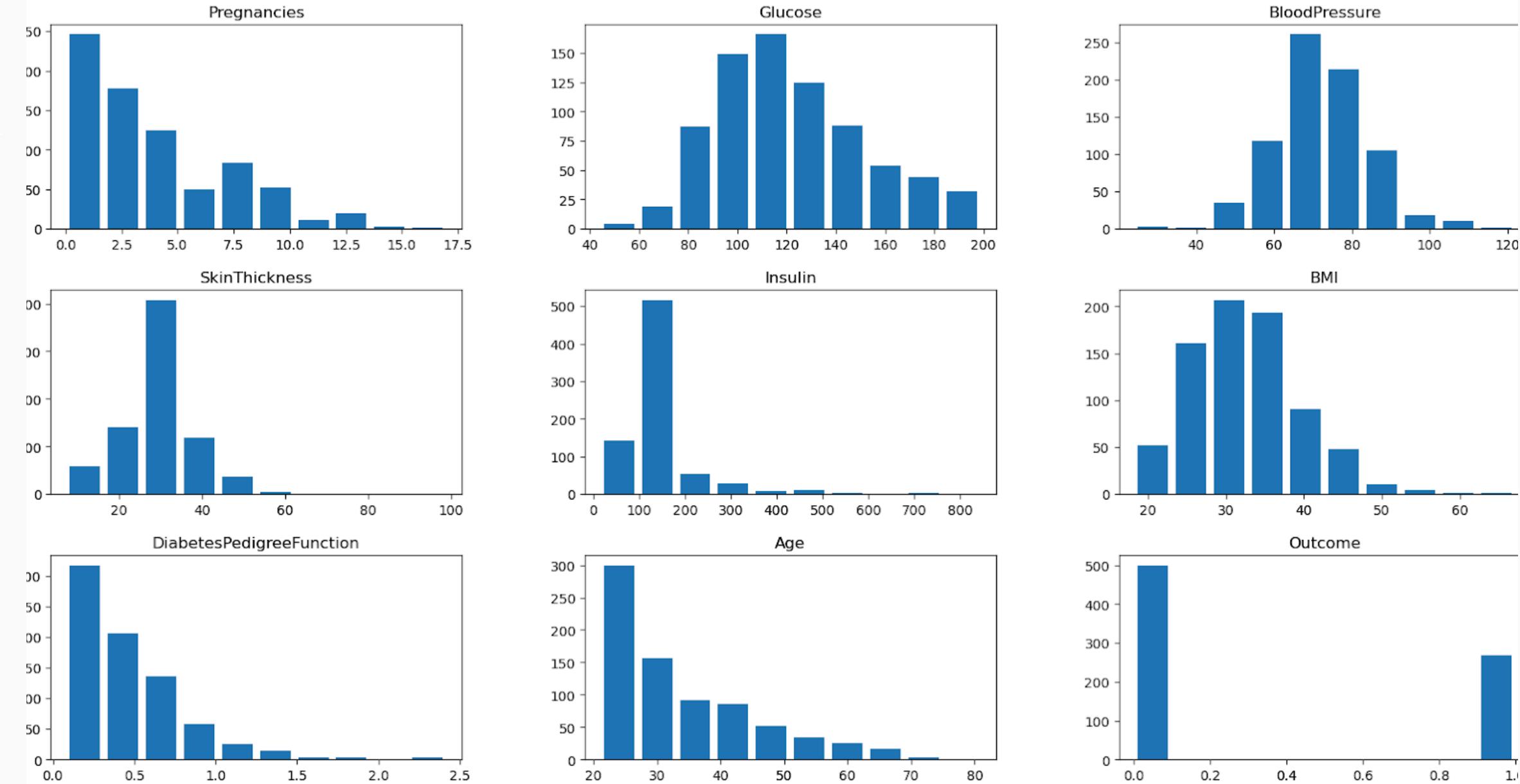
Khi phân tích đơn biến sẽ biết được phân phối của từng biến độc lập và có thể trả lời các câu hỏi sau:

- Nhóm tuổi nào chiếm nhiều nhất?
- Glucose, BMI phân bố ra sao?
- Bao nhiêu người mắc tiểu đường?
- Số lần mang thai thường rơi vào mức bao nhiêu?

Phân tích đơn biến

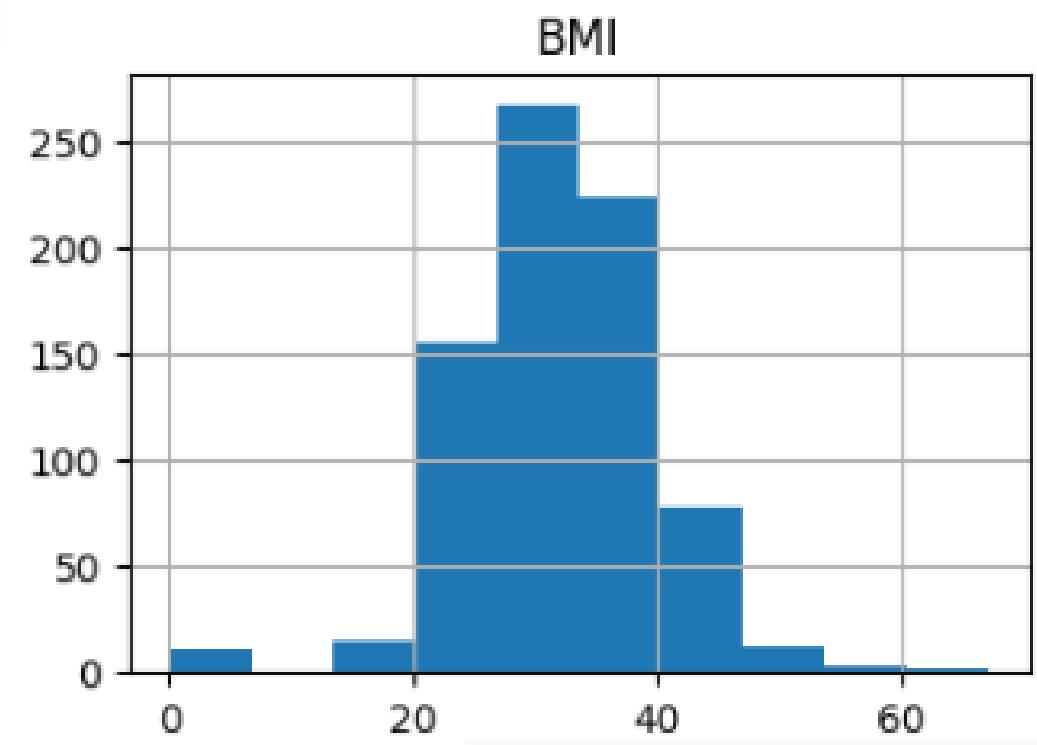
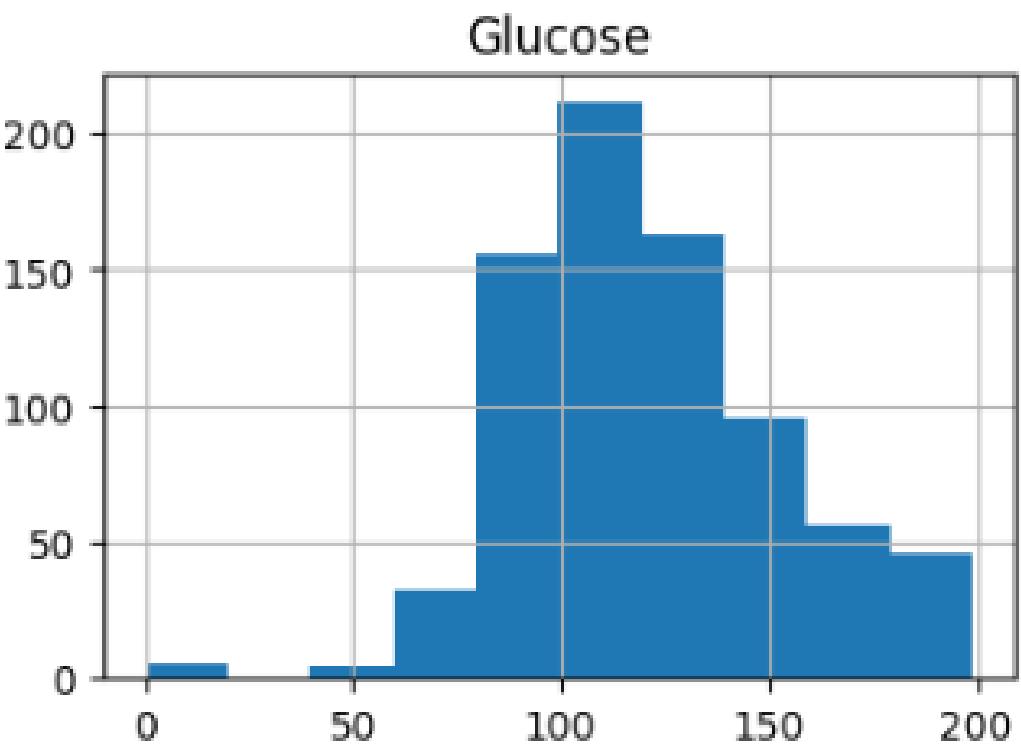
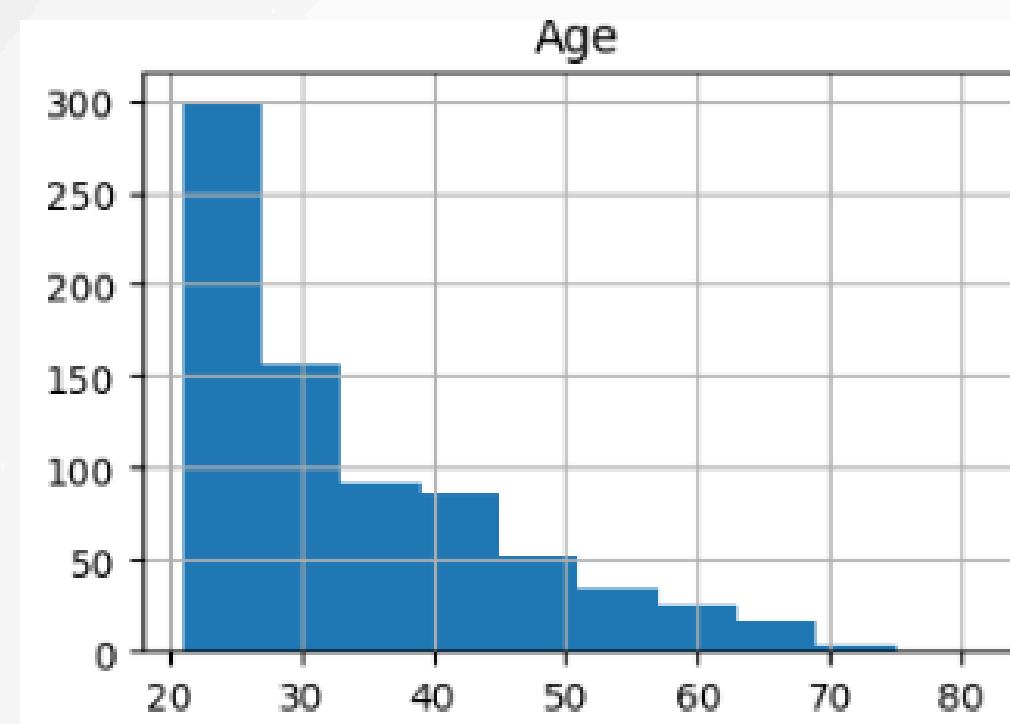


Histogram of Features



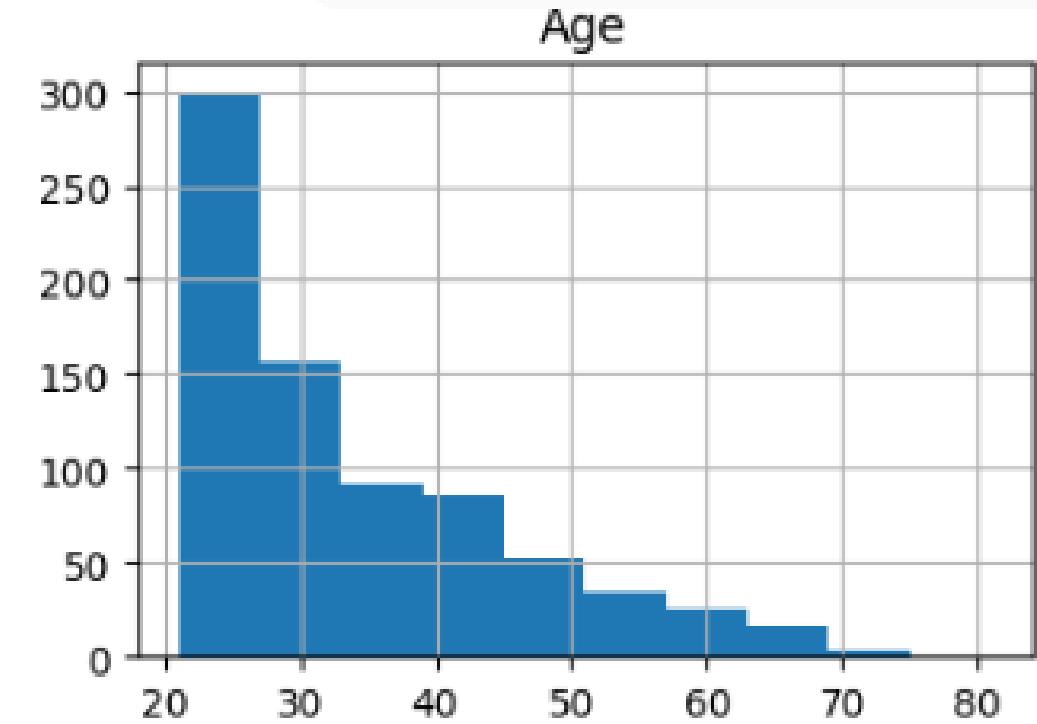
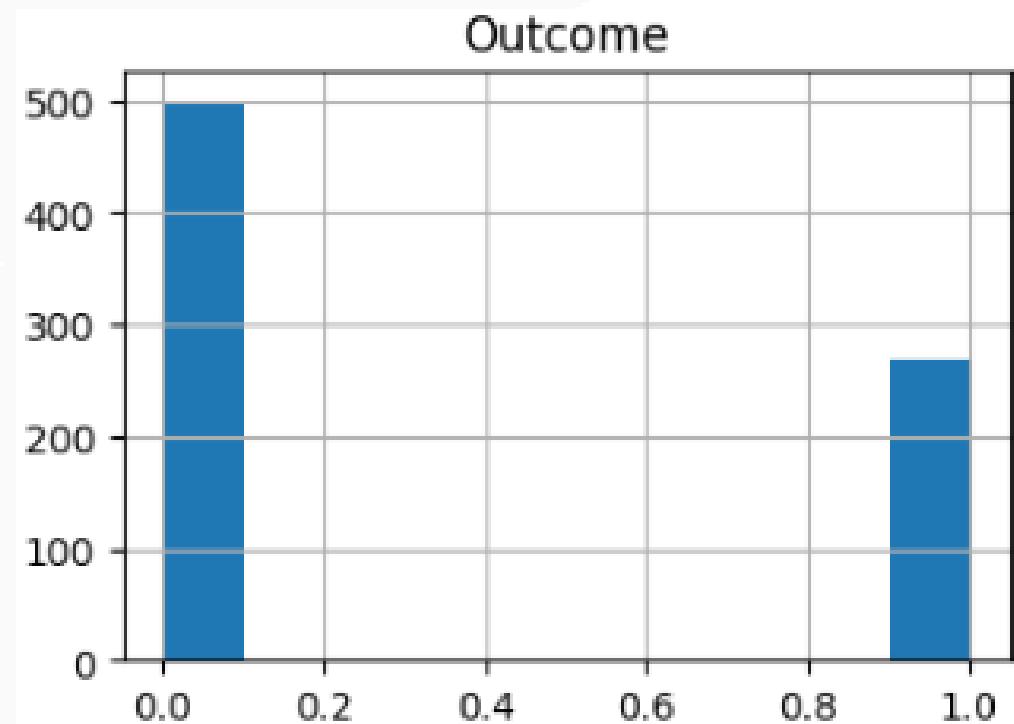
Mục đích: Để biết các biến nào phân bố như thế nào. phát hiện lệch hay mất cân bằng, xác định vấn đề cần xử lý trước khi modeling.

Phân tích đơn biến



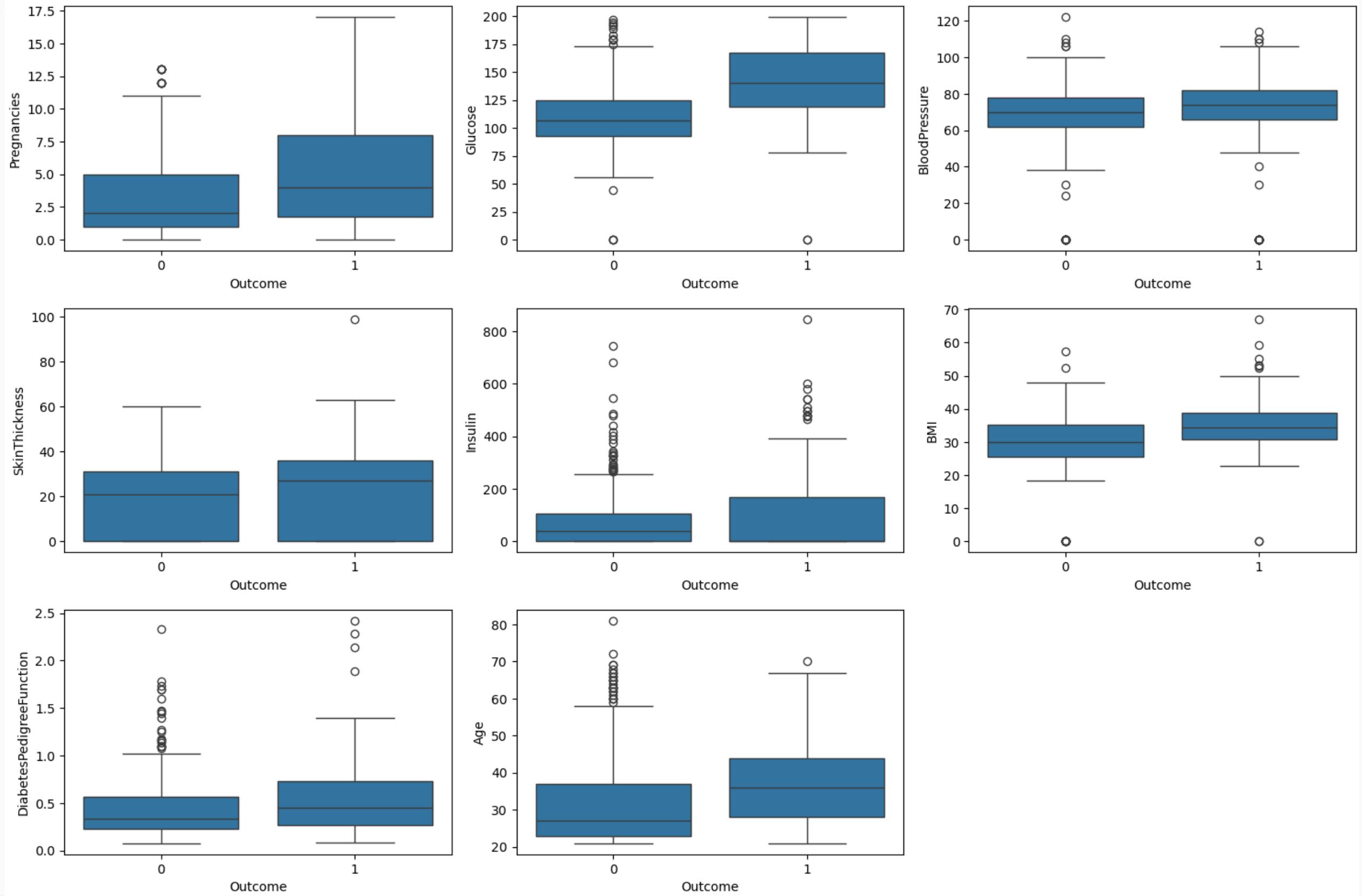
- Nhóm tuổi chiếm nhiều nhất là từ 20 – 30 tuổi.
- Glucose: phân bố nghiêng phải, tập trung nhiều nhất trong khoảng 90-140.
- BMI phân bố quanh 25-35, có đỉnh ở khoảng 30 (nhiều trường hợp thừa cân).

Phân tích đơn biến



- Có khoảng 500 người không mắc tiểu đường và khoảng 260 mắc bệnh tiểu đường (khoảng 34%), mất cân bằng dữ liệu cần oversampling, undersampling khi xây dựng mô hình phân loại
- Số lần mang thai phổ biến nhất là 0-2 lần mang thai.

Phân tích theo kết quả

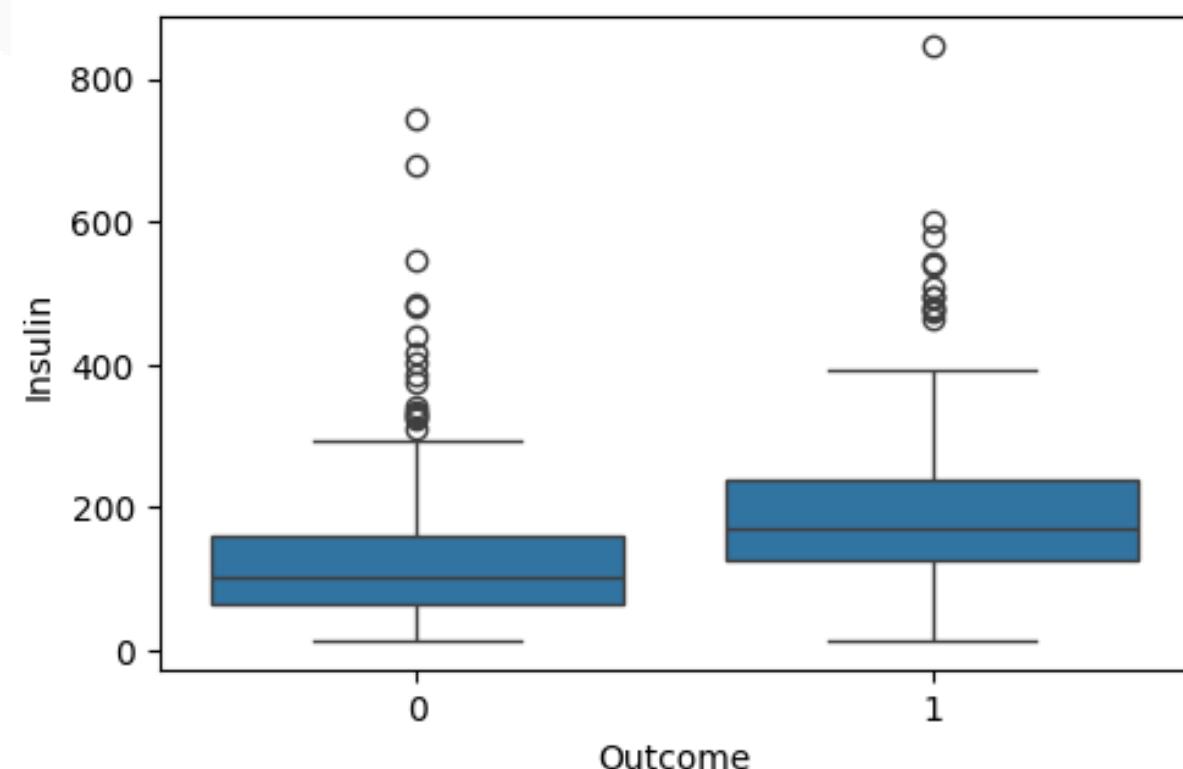
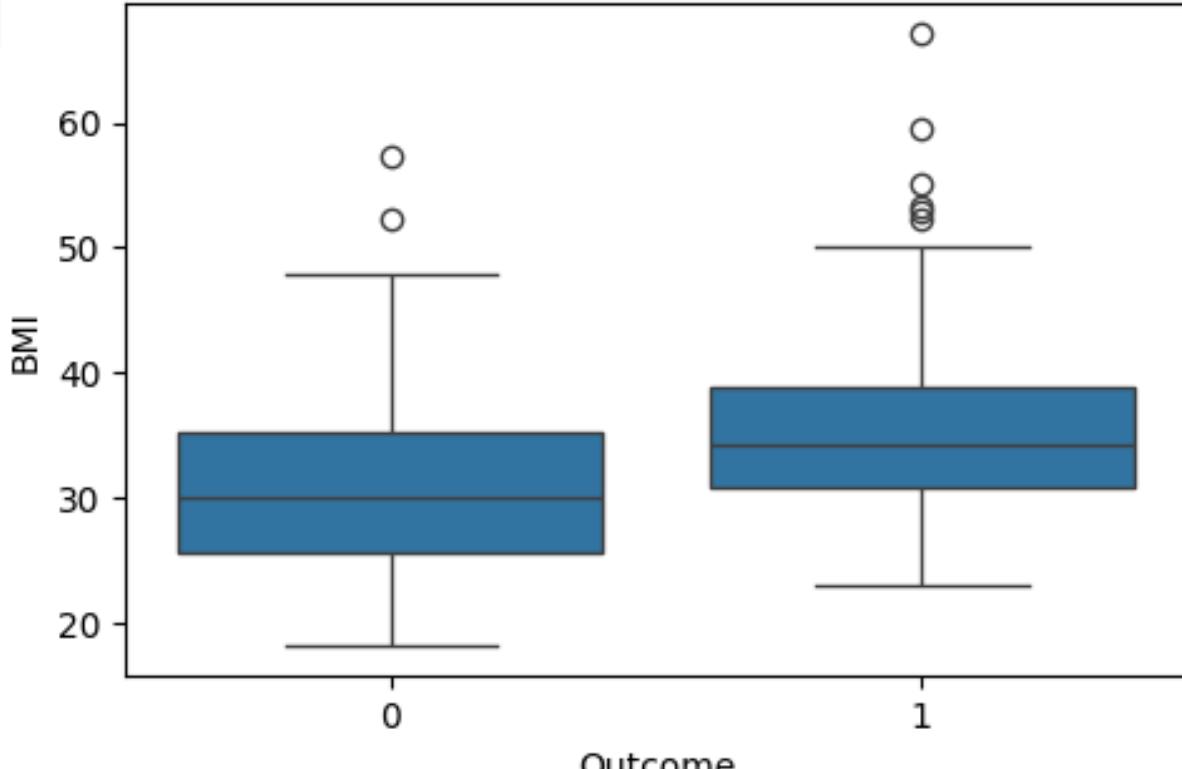
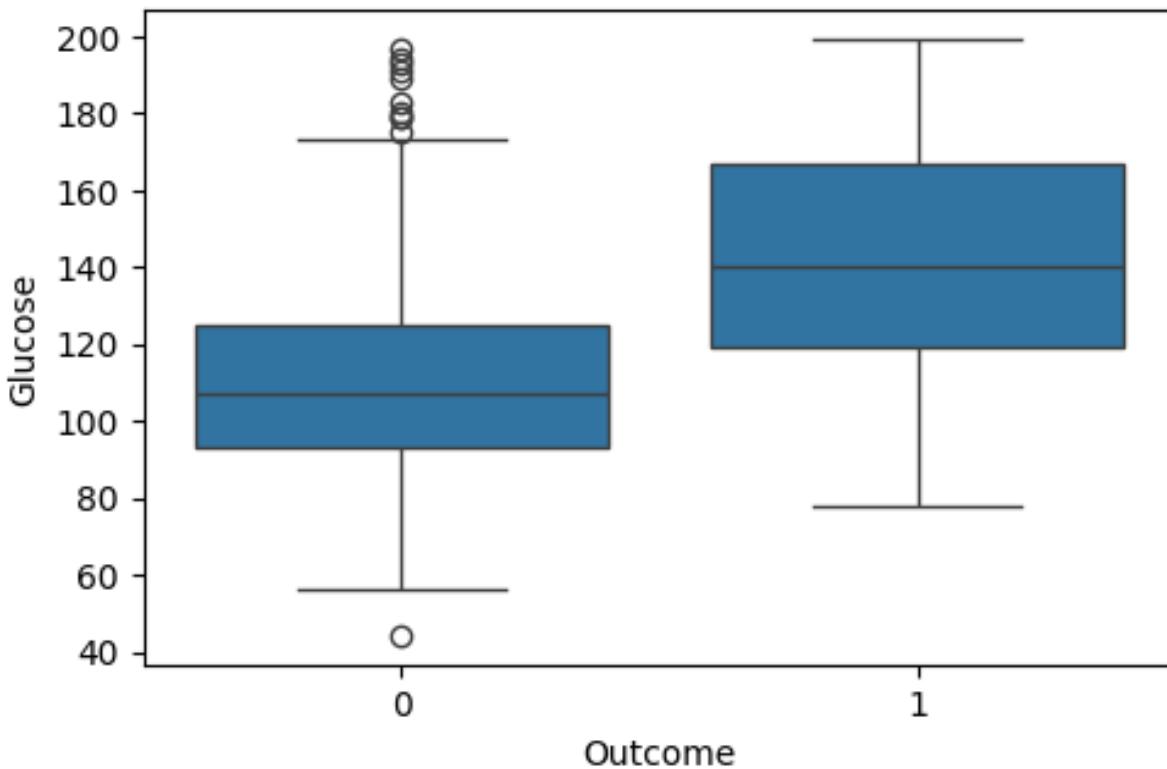


Mục đích: Xác định biến quan trọng để dùng cho mô hình

Phân tích theo kết quả

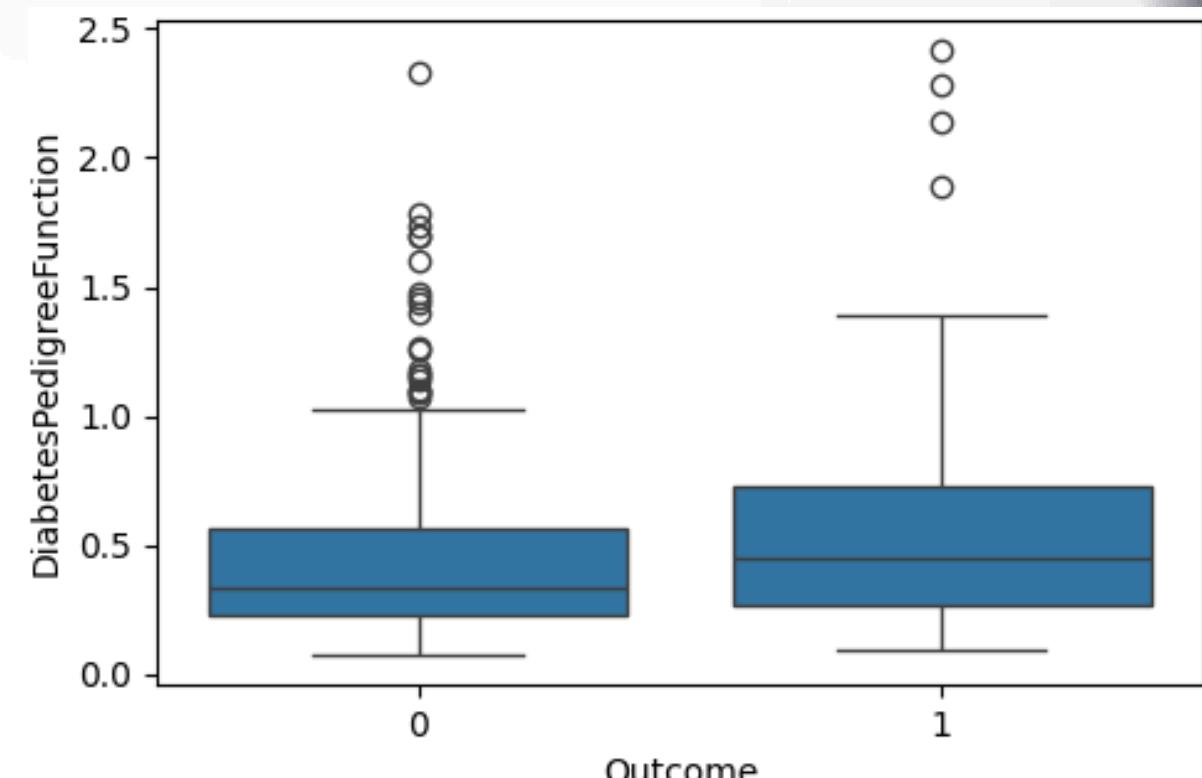
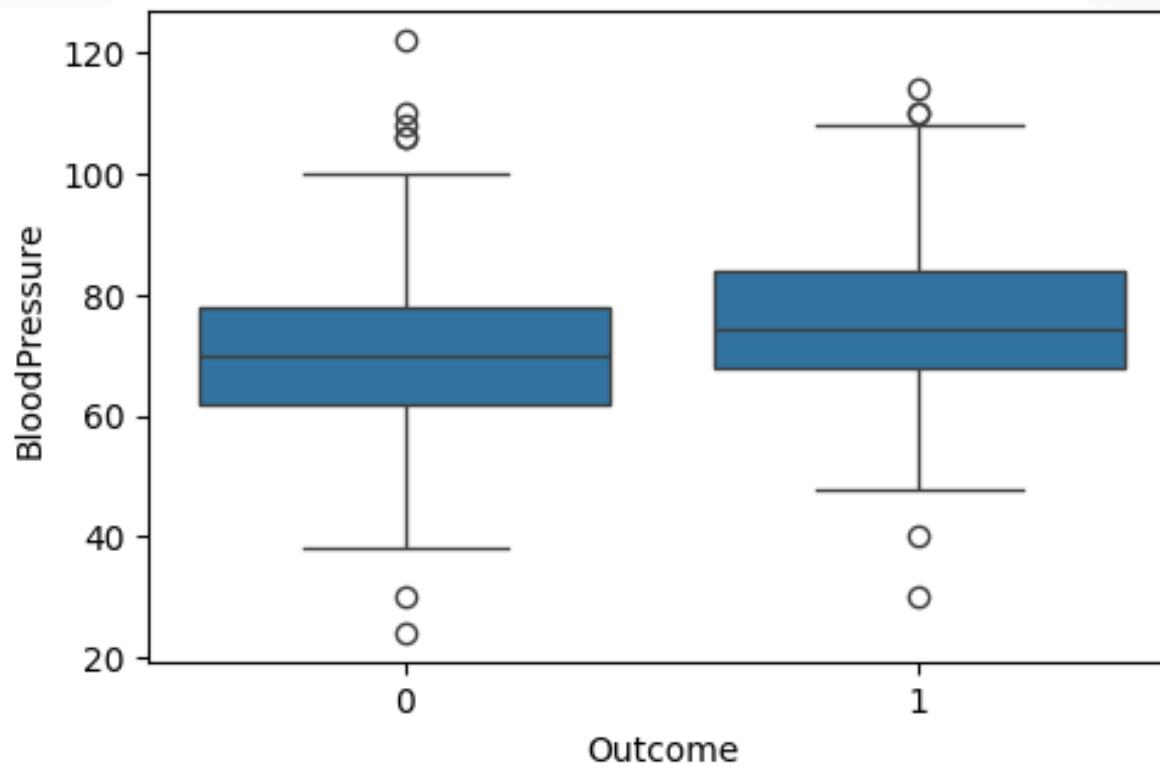
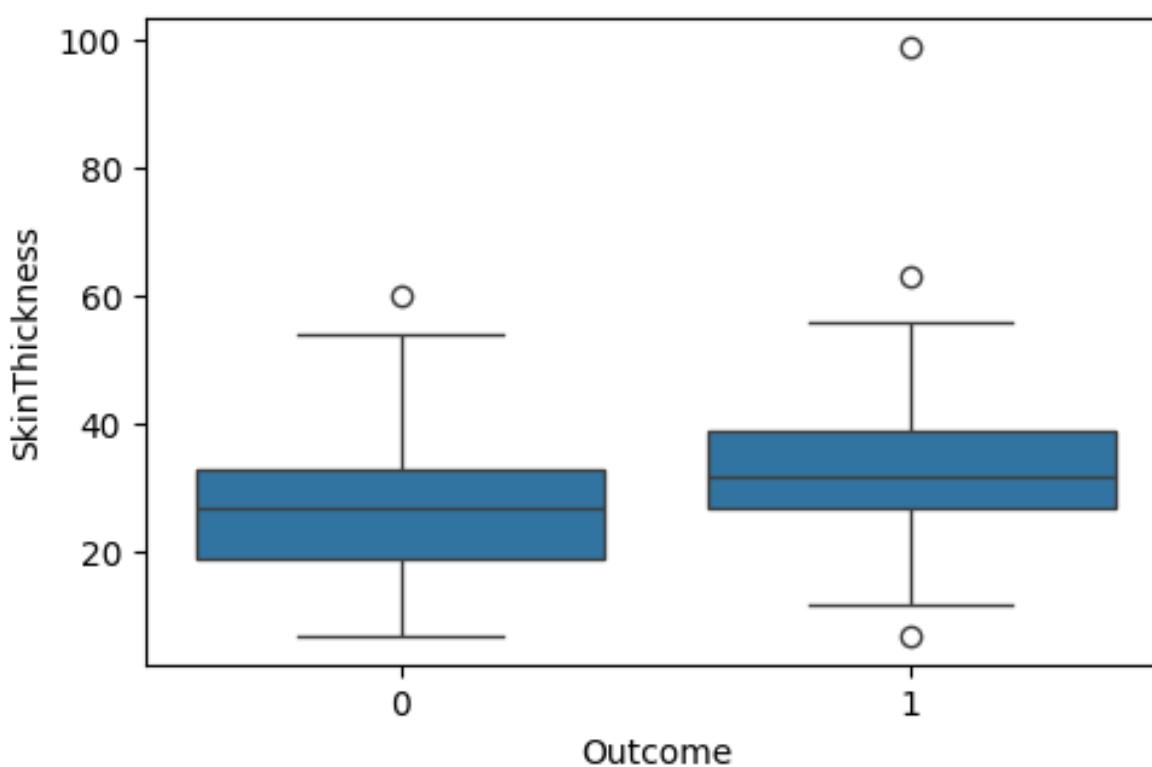
- Glucose cao thì nhóm nào mắc tiểu đường nhiều hơn?
- Mắc tiểu đường thì BMI thế nào?
- Insulin thấp/cao ảnh hưởng thế nào đến Outcome?
- SkinThickness có thực sự tách biệt 2 nhóm Outcome không?
- BloodPressure trung bình có khác biệt giữa nhóm bệnh và không bệnh không?
- Diabetes Pedigree Function cao có đồng nghĩa nguy cơ cao hơn không?

Phân tích theo kết quả



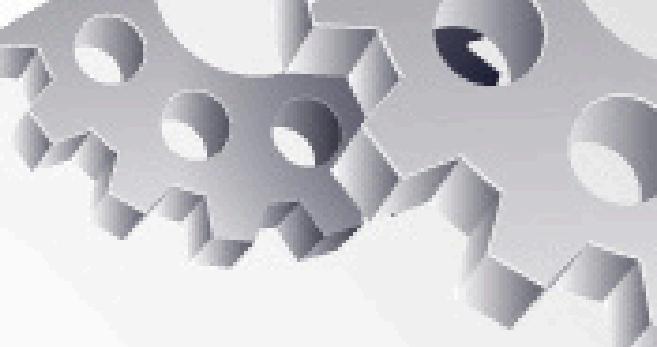
- Người có Glucose cao dễ mắc tiểu đường hơn.
- Người mắc tiểu đường thường có BMI cao hơn so với nhóm không bệnh.
- Insulin có nhiều giá trị ngoại lai (outlier), phân bố rộng và chồng lấn giữa hai nhóm. Nhìn chung, Insulin không cho thấy sự khác biệt rõ ràng giữa Outcome 0 và 1, nên không phải yếu tố tách biệt tốt

Phân tích theo kết quả



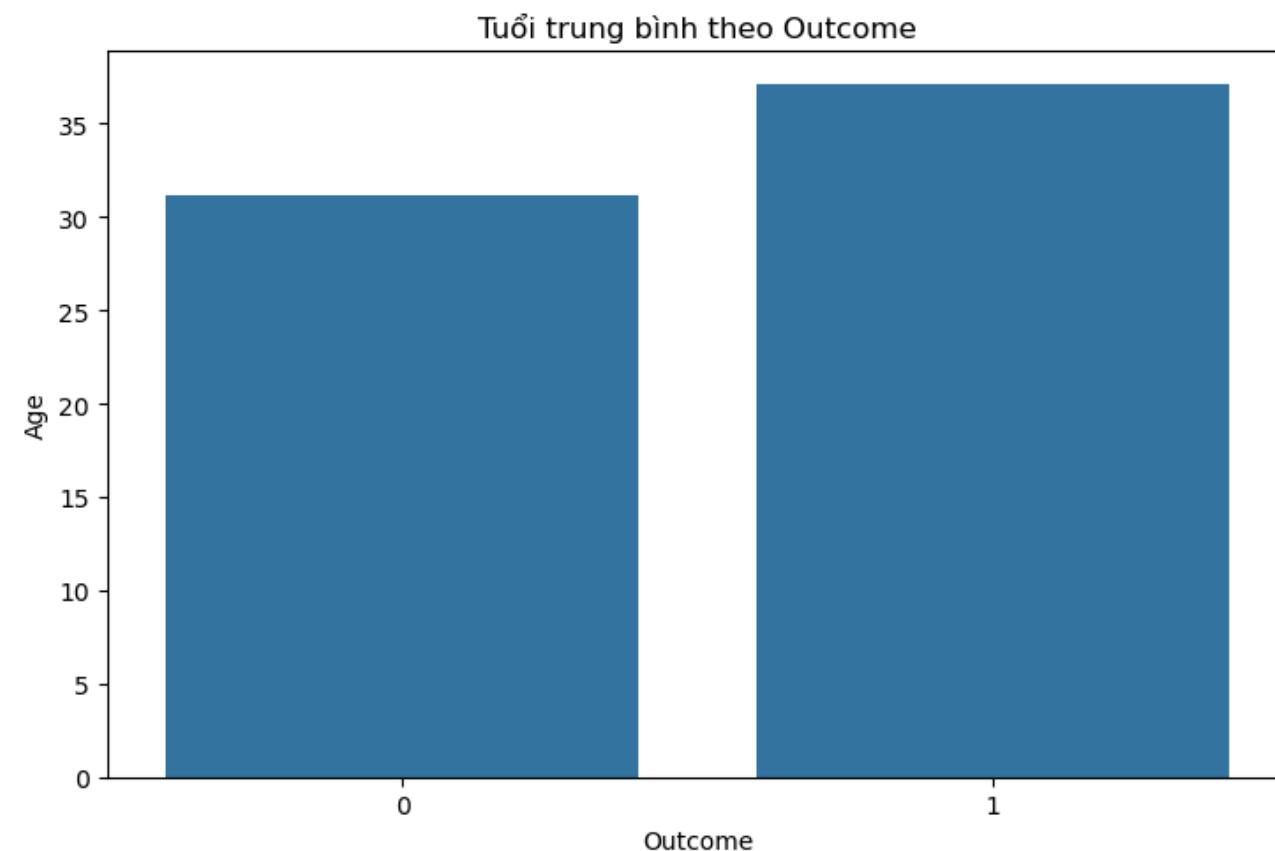
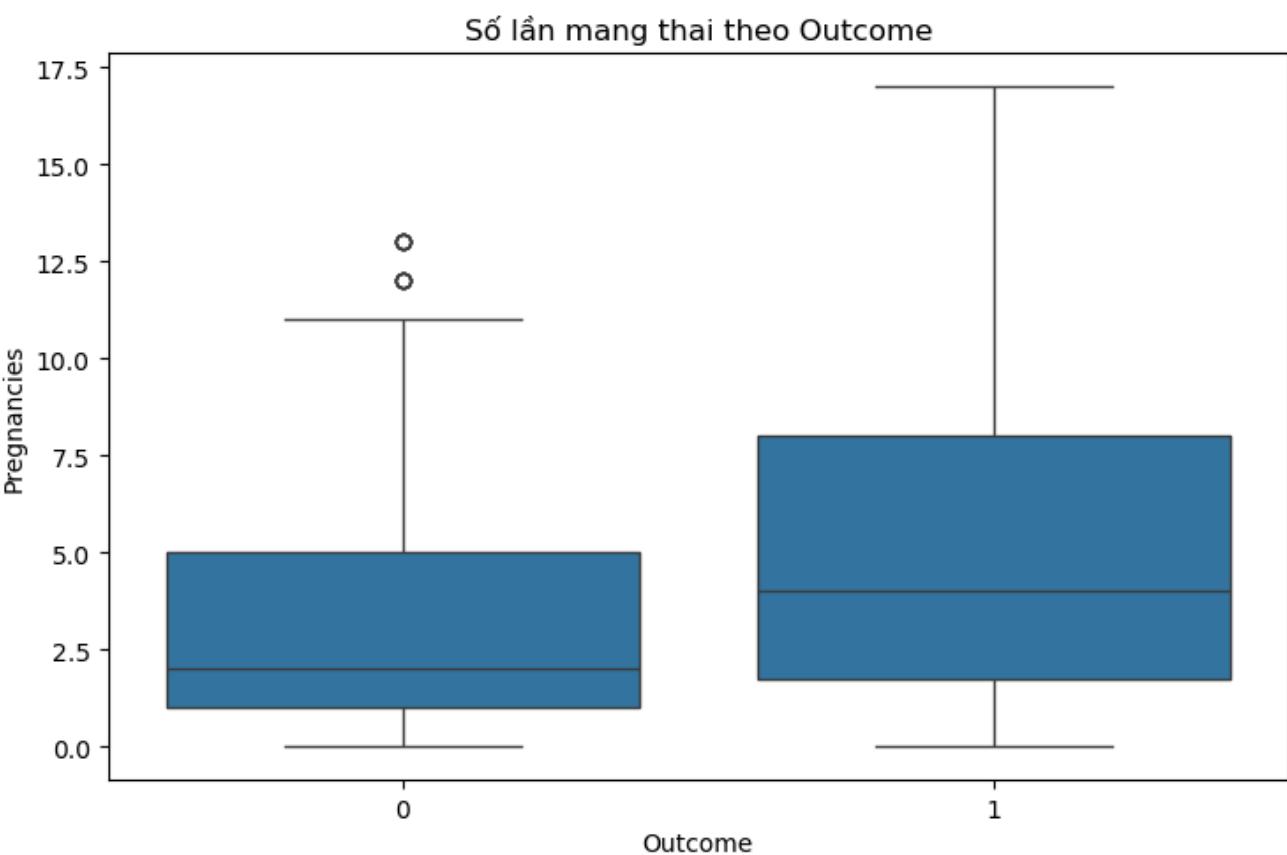
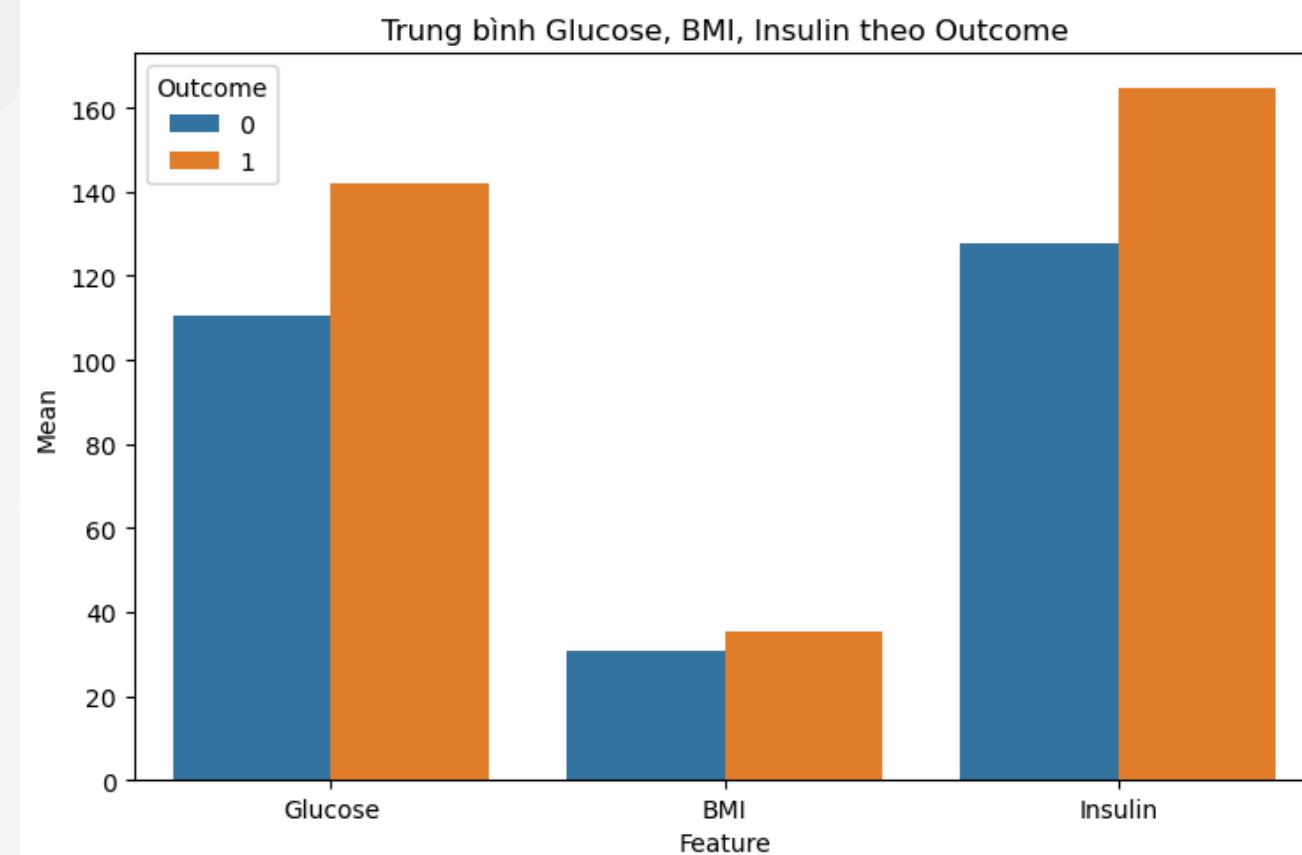
- Hai median chênh không nhiều, chồng lấn mạnh. SkinThickness không tách biệt rõ rệt giữa người có và không có tiểu - đường.
- Huyết áp trung bình không có sự khác biệt rõ rệt giữa người tiểu đường và không tiểu đường
- Nhóm mắc bệnh có giá trị DiabetesPedigreeFunction nhỉnh hơn một chút, nhưng phân bố vẫn khá chồng lấn. Như vậy, chỉ số này có ảnh hưởng nhưng không mạnh bằng Glucose hay BMI.

Phân tích theo kết quả



- Các biến quan trọng nhất khi so sánh với Outcome: **Glucose, BMI, Age, Pregnancies.**
- Các biến yếu hơn hoặc ít khác biệt: **BloodPressure, SkinThickness, Insulin, DiabetesPedigreeFunction.**
- Điều này phù hợp với y học: đường huyết cao, thừa cân, tuổi cao là yếu tố nguy cơ chính gây tiểu đường type 2.

Phân tích theo kết quả



- Trung bình Glucose, BMI, Insulin khác nhau thế nào giữa Outcome=0 và 1?
- Người mắc tiểu đường có số lần mang thai cao hơn không?
- Tuổi trung bình của nhóm mắc bệnh và không mắc bệnh khác nhau bao nhiêu?

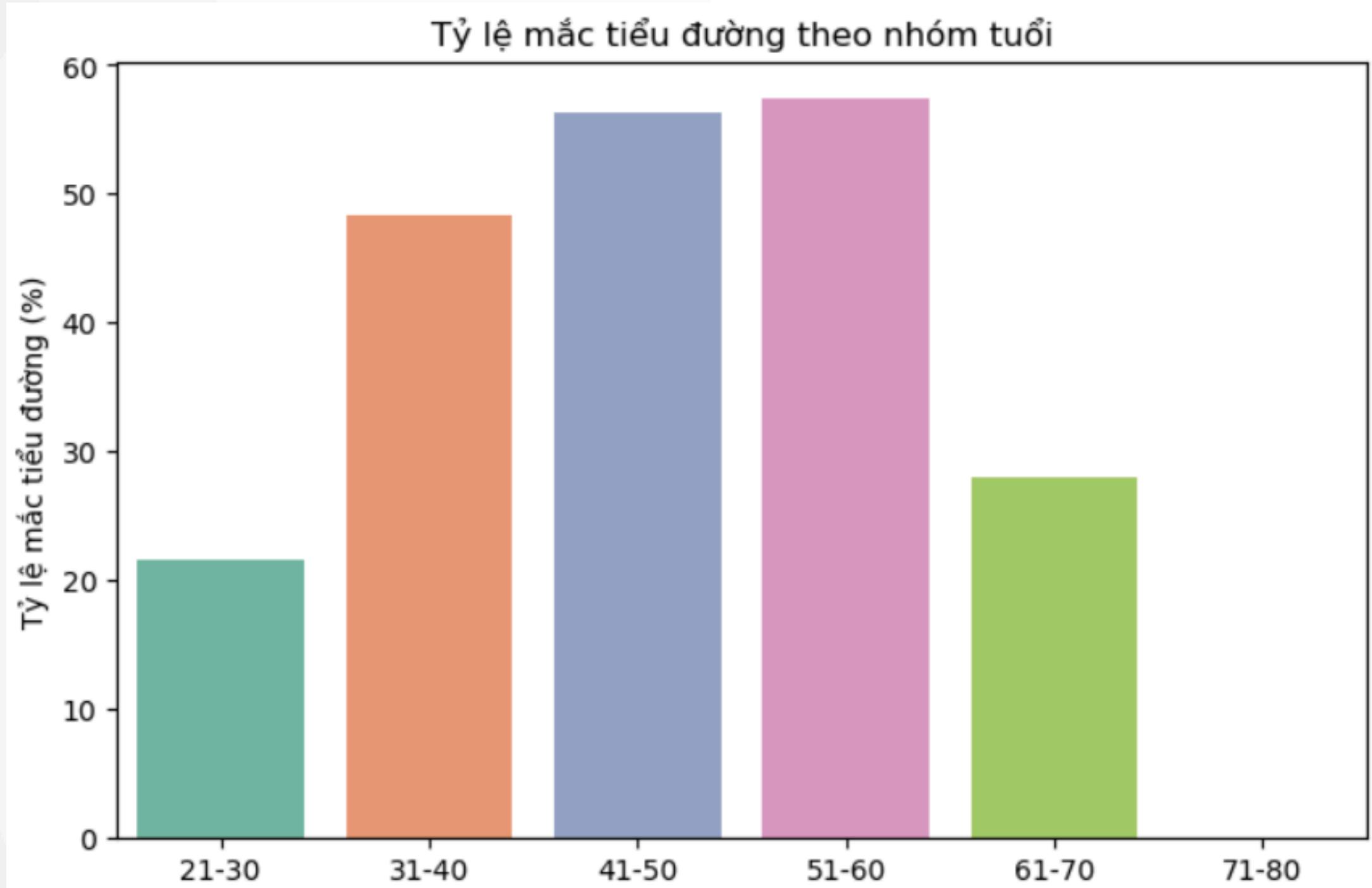
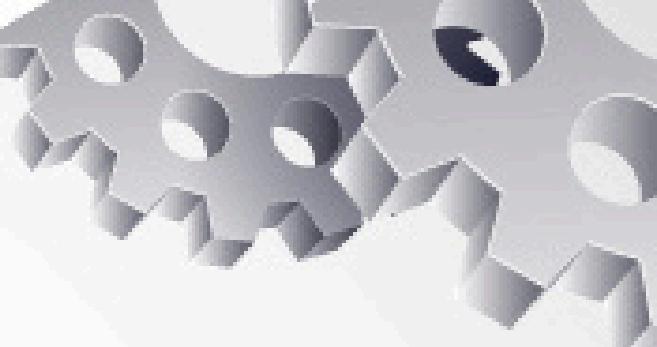
Phân tích theo kết quả

- Người mắc tiểu đường (Outcome=1) có Glucose trung bình cao hơn hẳn so với người không mắc. BMI trung bình của nhóm Outcome=1 cũng cao hơn một chút. Insulin trung bình ở nhóm Outcome=1 cao hơn so với nhóm Outcome=0.
- Boxplot cho thấy nhóm Outcome=1 có median số lần mang thai cao hơn so với nhóm Outcome=0. Có xu hướng rằng người mắc tiểu đường thường có nhiều lần mang thai hơn.
- Người mắc tiểu đường lớn tuổi hơn trung bình 6 năm so với người không mắc.

Phân tích theo độ tuổi

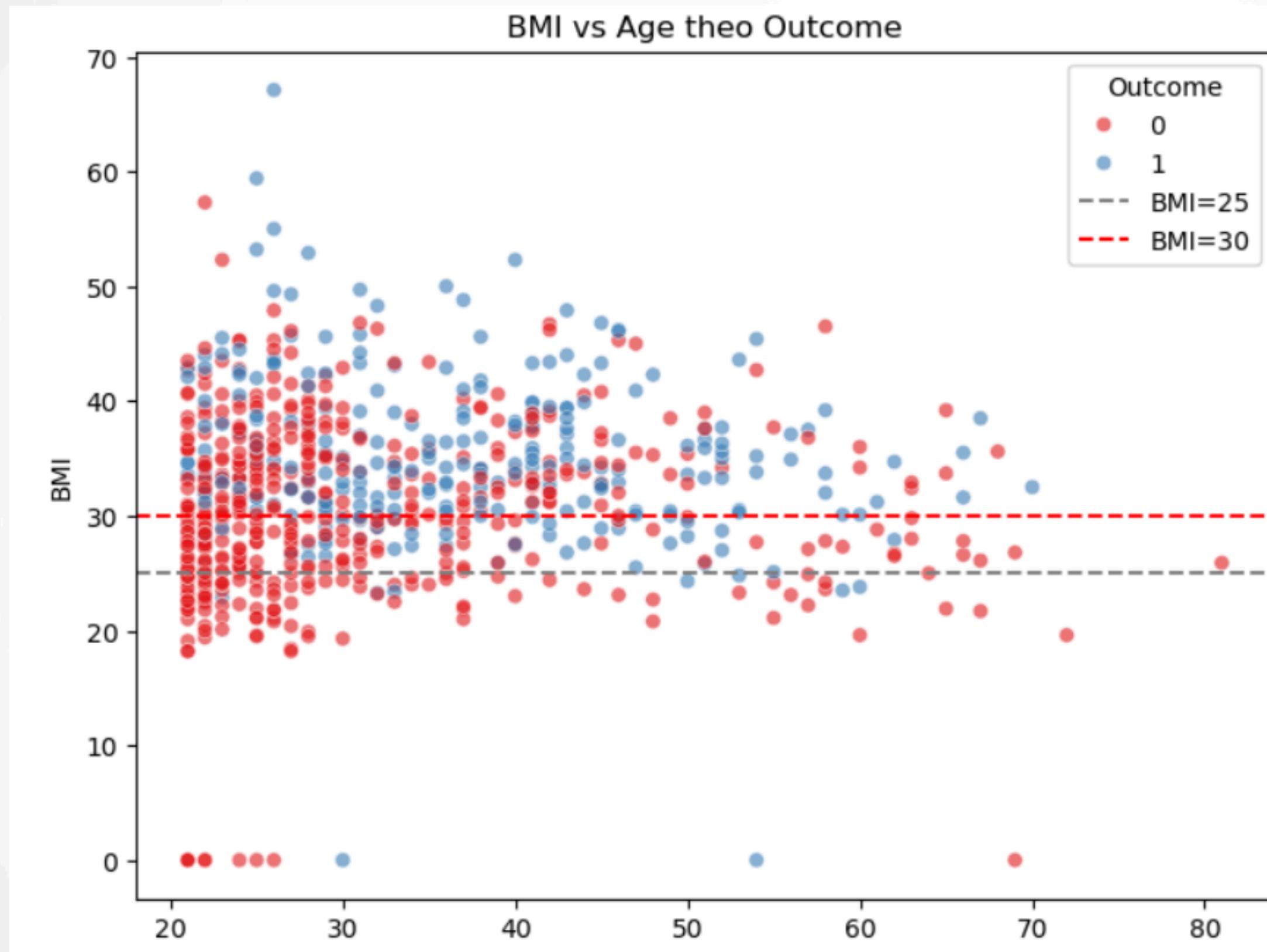
- Nhóm tuổi nào có tỷ lệ tiểu đường cao nhất?
- Tỷ lệ mắc tiểu đường tăng dần theo tuổi như thế nào?
- Trong nhóm tuổi 40–50, BMI > 30 thì nguy cơ cao hơn BMI < 25 bao nhiêu?
- Có mối quan hệ nào giữa tuổi và số lần mang thai không?

Phân tích theo độ tuổi



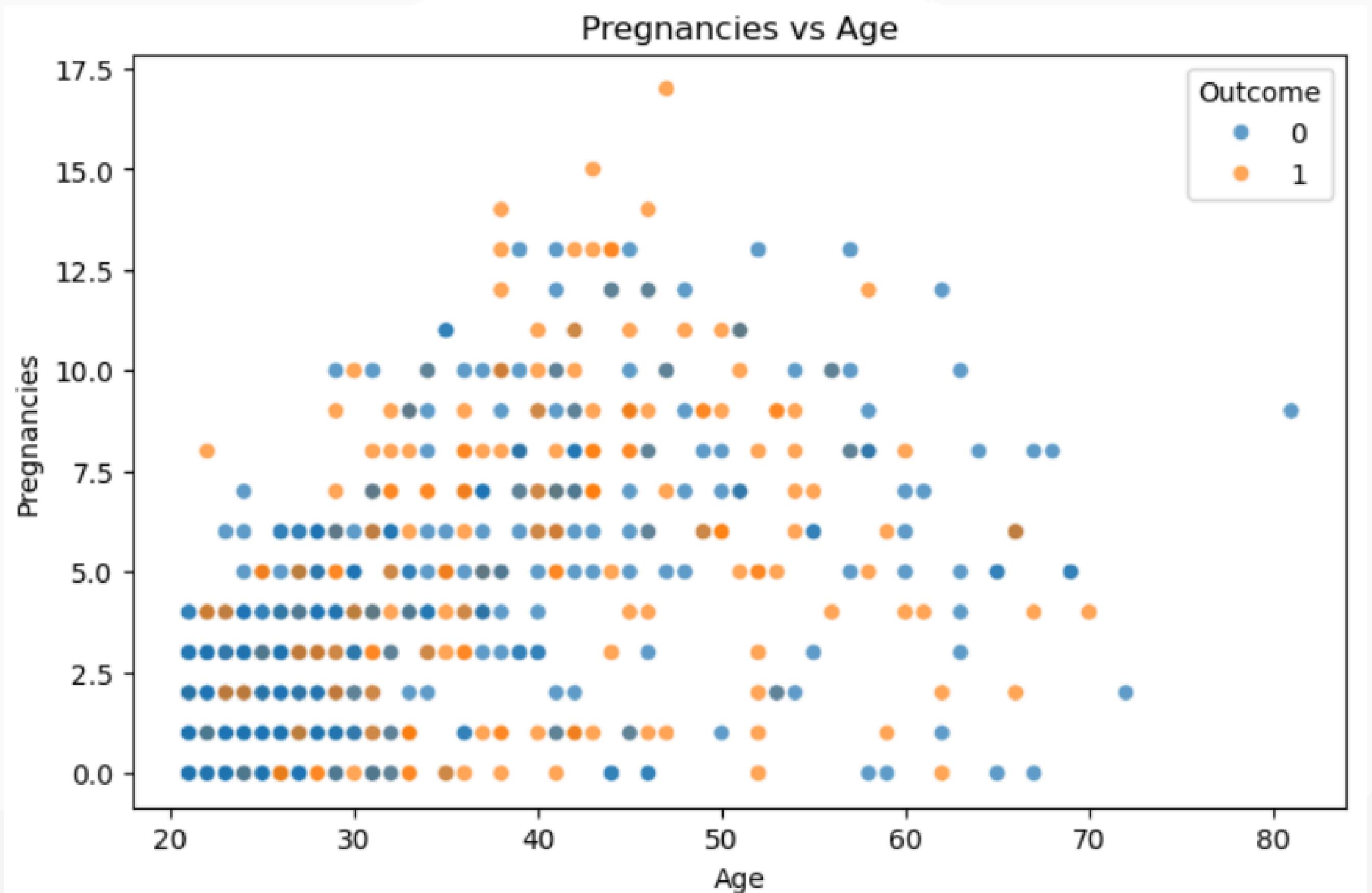
- Nhóm tuổi có tỷ lệ tiểu đường cao nhất là: 51- 60 tuổi
- Tỷ lệ mắc tiểu đường tăng dần theo tuổi cho đến khoản 60 tuổi trở đi thì bắt đầu giảm

Phân tích theo độ tuổi



- Trong nhóm tuổi 40–50, người có $\text{BMI} > 30$ có nguy cơ tiểu đường cao hơn hẳn so với $\text{BMI} < 25$.
- Kết luận:** BMI cao là yếu tố mạnh quyết định nguy cơ tiểu đường, ngay cả khi tuổi giống nhau.

Phân tích theo tuổi



Phân tích theo tuổi

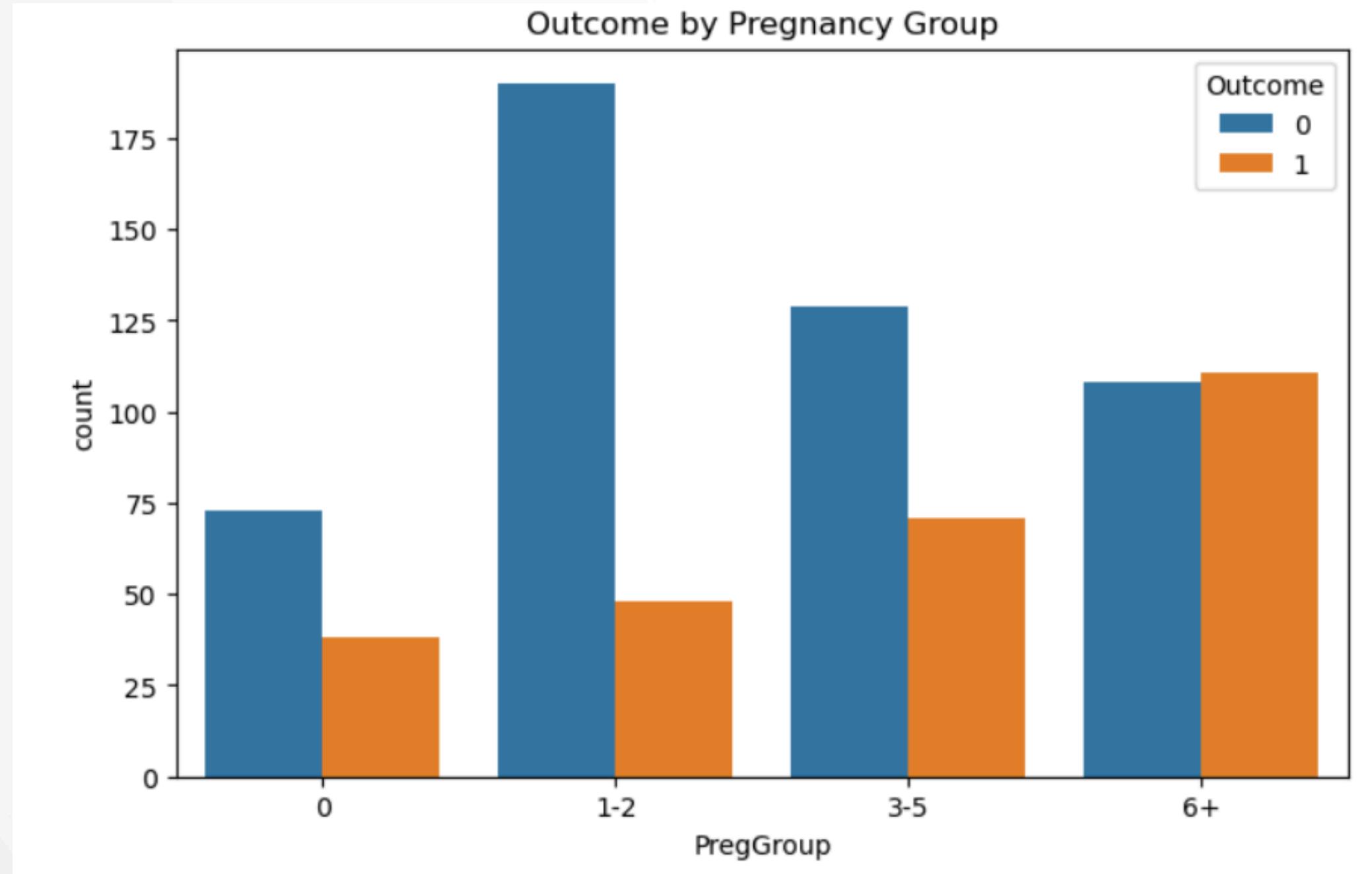
- Một số phụ nữ trẻ (~20–30 tuổi) đã có nhiều lần mang thai (5–10 lần), nhưng phần lớn nằm ở 0–5 lần.
- Phụ nữ lớn tuổi (~50–60 tuổi) có số lần mang thai khá đa dạng, nhưng trung bình vẫn cao hơn nhóm trẻ.
- Ảnh hưởng của Outcome: Không có sự phân tách rõ ràng giữa Outcome=0 và Outcome=1 theo số lần mang thai

Phân tích theo số lần mang thai



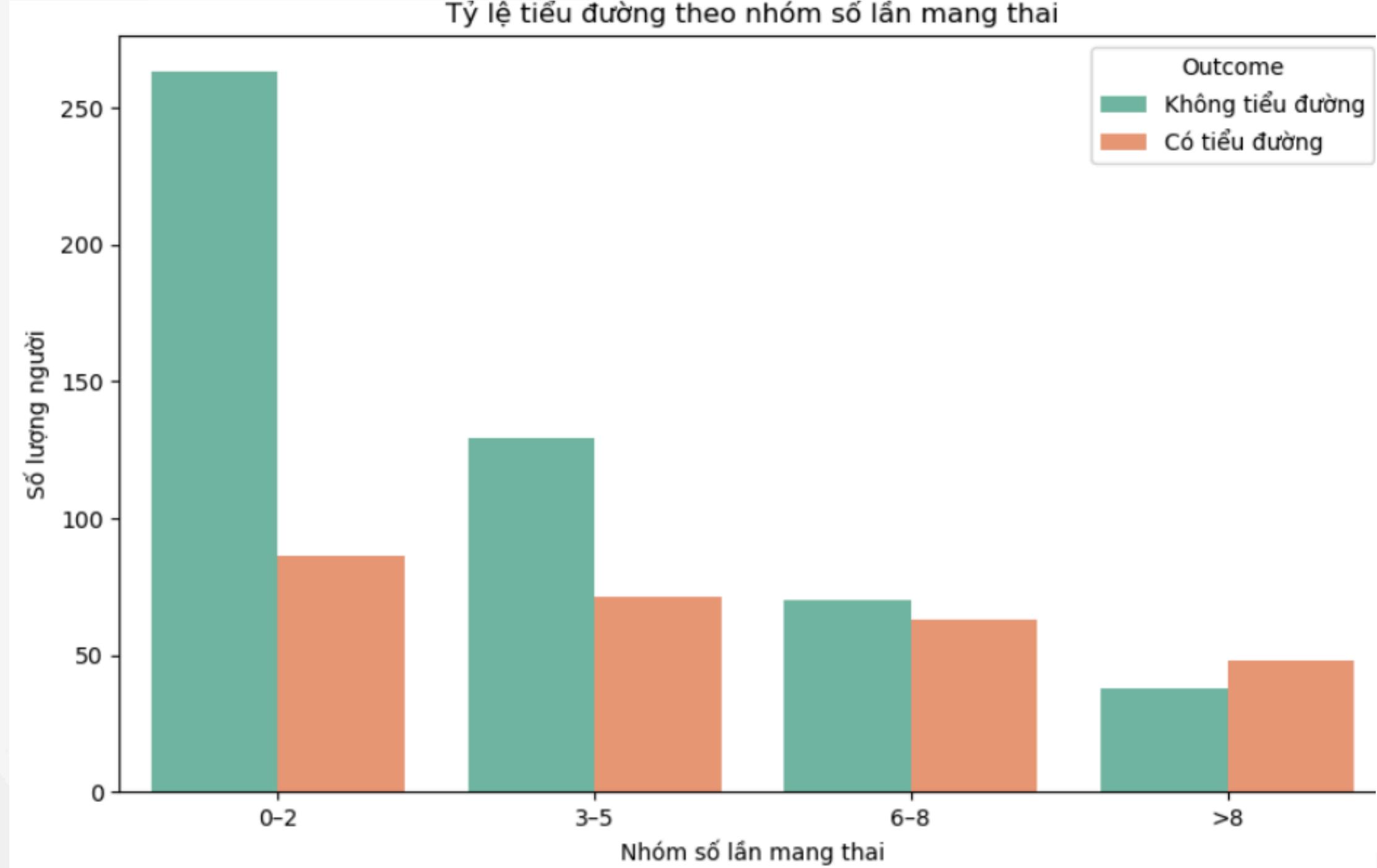
- Phân bố số lần mang thai (Pregnancies) trong dataset như thế nào?
- Người có nhiều lần mang thai hơn có nguy cơ tiểu đường cao hơn không?
- Có ngưỡng số lần mang thai nào làm tăng nguy cơ tiểu đường?
- So sánh số lần mang thai trung bình giữa nhóm có tiểu đường (Outcome=1) và nhóm không có tiểu đường (Outcome=0).

Phân tích theo số lần mang thai



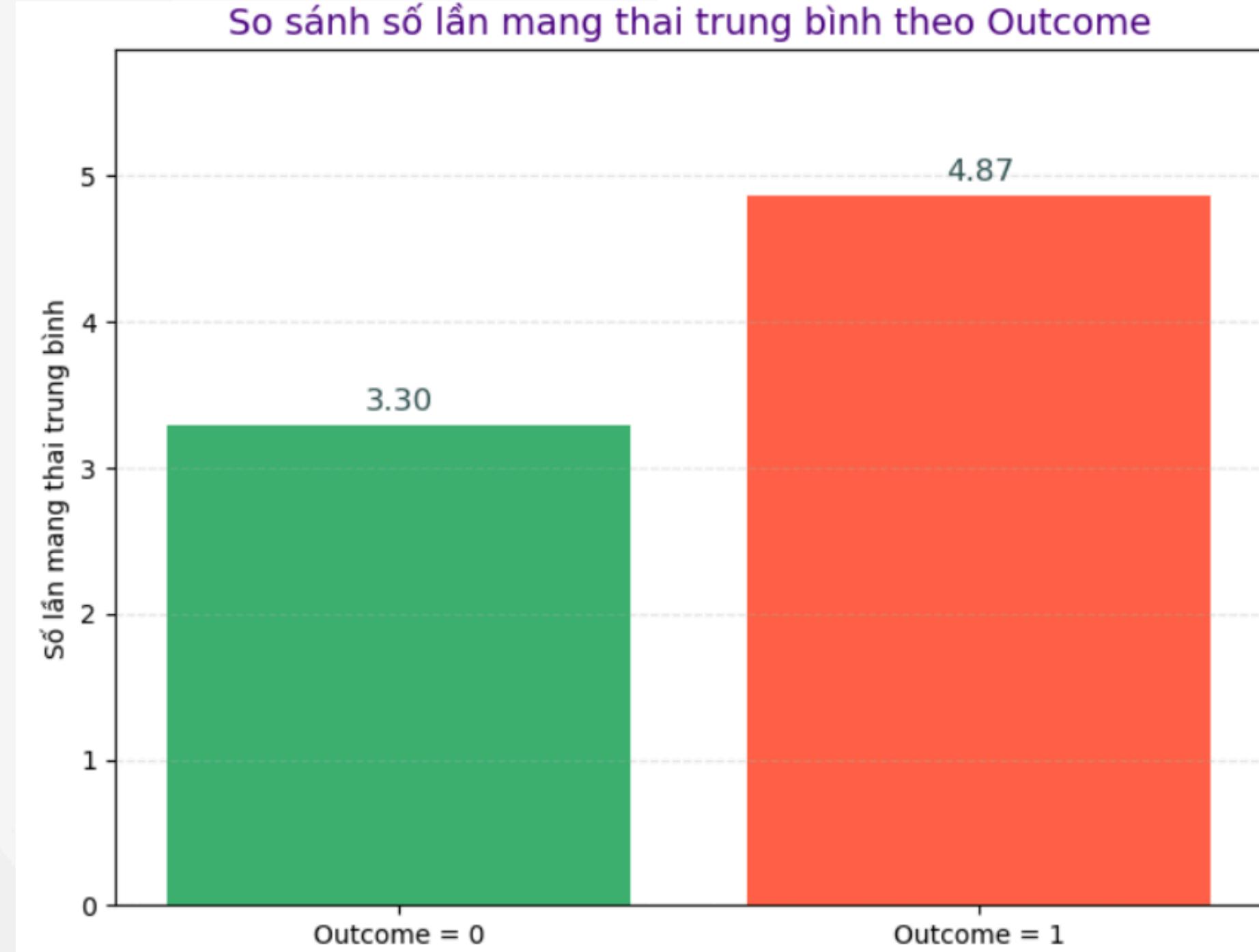
- Nguy cơ tiểu đường tăng theo số lần mang thai:
- Nhóm chưa từng mang thai có nguy cơ thấp
- Nhóm mang thai nhiều lần (6+) có nguy cơ cao nhất
- Biểu đồ cho thấy mối quan hệ rõ ràng giữa số lần mang thai và tỷ lệ mắc tiểu đường.

Phân tích theo số lần mang thai



- Nhóm 0-2 là nhóm có nguy cơ thấp nhất.
- Từ khoảng 6 lần mang thai trở lên, nguy cơ mắc tiểu đường bắt đầu tăng rõ.

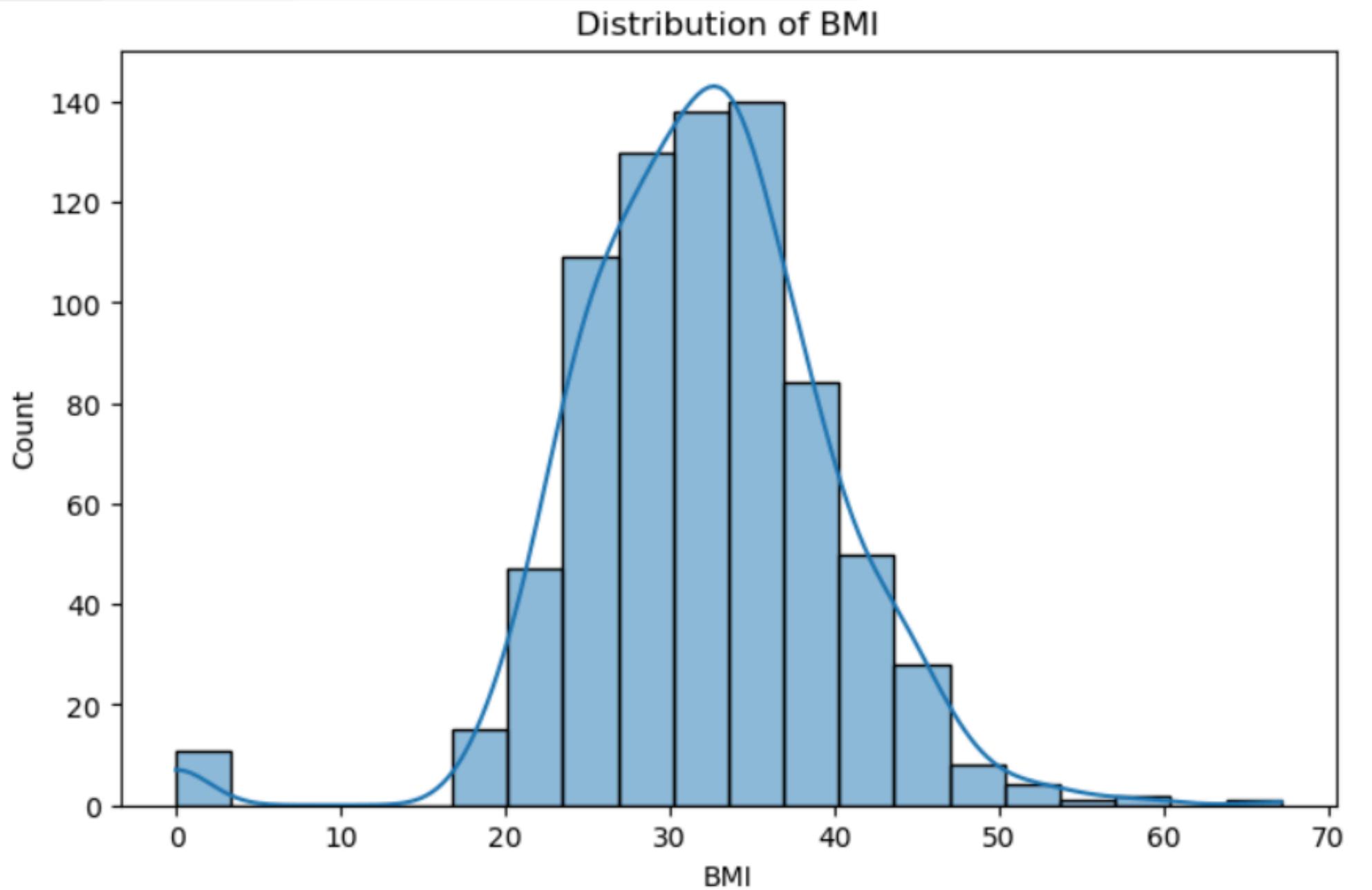
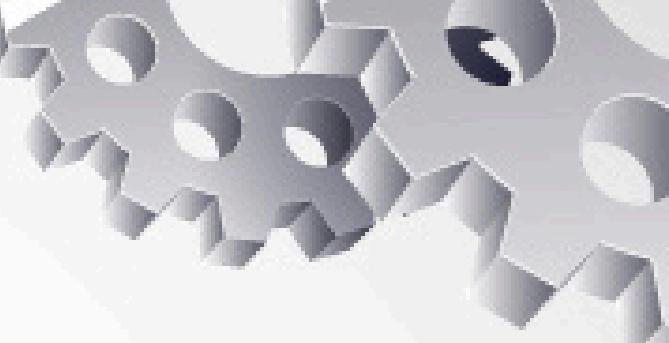
Phân tích theo số lần mang thai



- Không tiểu đường: trung bình 3.30 lần mang thai
 - Có tiểu đường: 4.87 trung bình lần mang thai
- => Pregnancies có thể là yếu tố hỗ trợ phân loại nguy cơ, dù không phải là biến mạnh nhất.

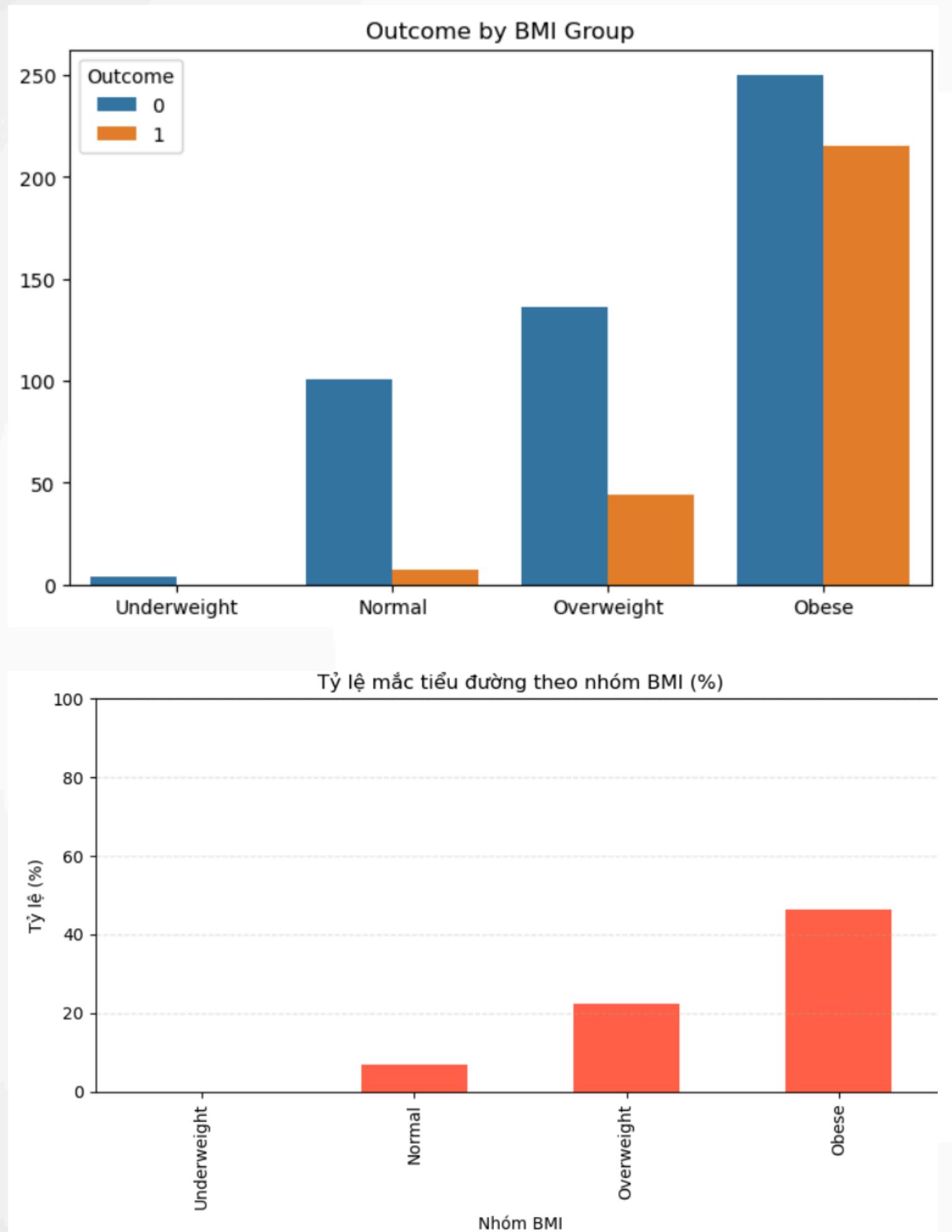
Phân tích theo BMI

- Phân bố chỉ số BMI trong dataset như thế nào?
- Nếu phân nhóm (Underweight <18.5, Normal 18.5-24.9, Overweight 25-29.9, Obese >=30), Outcome=1 xuất hiện nhiều ở nhóm nào?
- BMI cao có đi kèm với Glucose cao không?



Phân bố chỉ số BMI trong dataset như thế nào?

- Phân bố gần chuẩn, tập trung nhiều ở khoảng 25-35
- Một số giá trị $>50 \rightarrow$ có thể là outlier
- BMI trung bình quanh mức 30



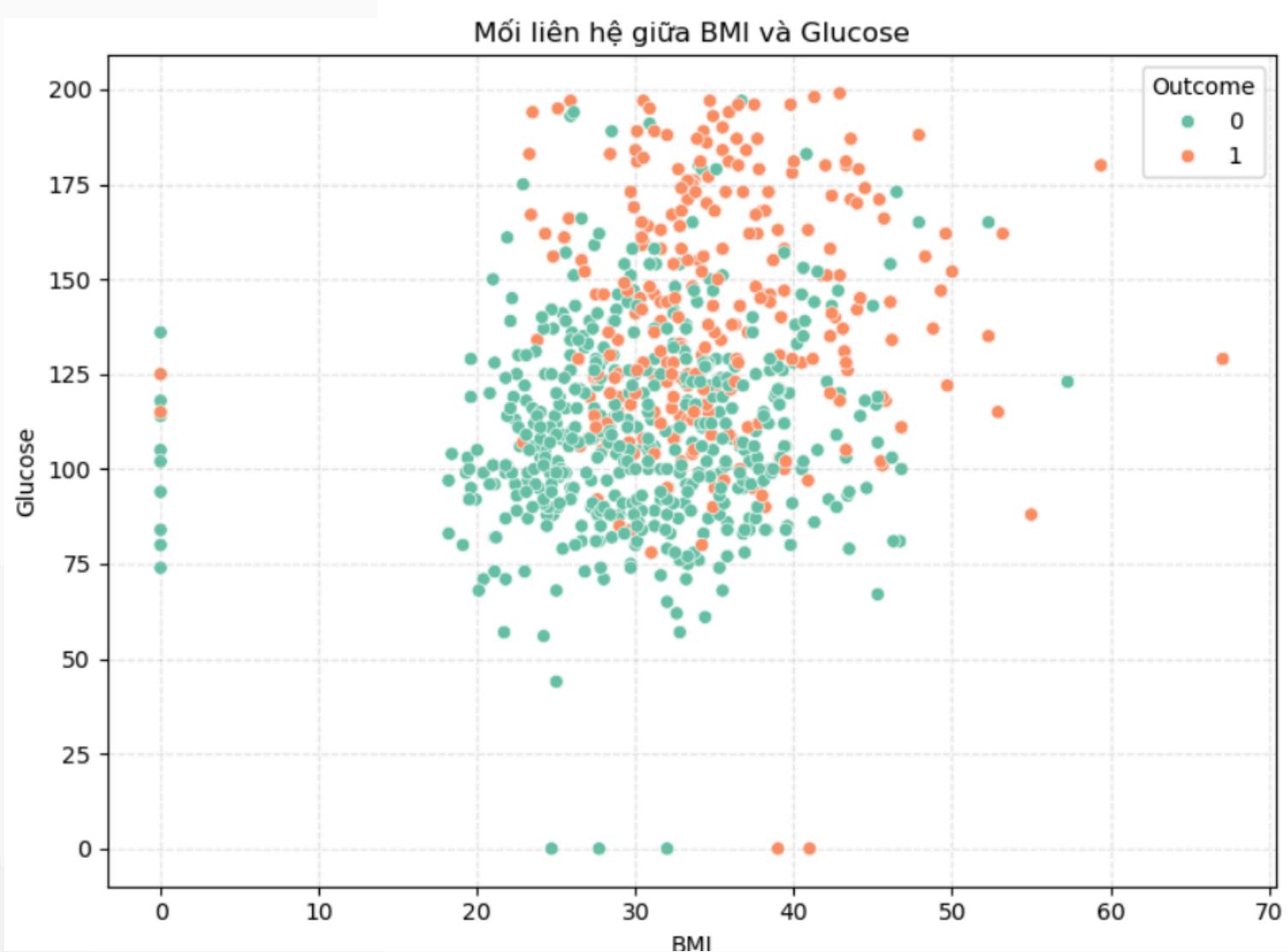
Nguy cơ tiểu đường xuất hiện nhiều ở nhóm BMI nào?

- Nhóm Obese ($BMI \geq 30$) có số lượng Outcome = 1 cao nhất
- Nhóm Underweight ít người, không ảnh hưởng nhiều đến phân tích
- BMI cao là yếu tố nguy cơ rõ rệt → cần theo dõi trong mô hình

Mối liên hệ giữa BMI và Glucose

```
# Tính hệ số tương quan Pearson  
correlation = df['BMI'].corr(df['Glucose'])  
print(f' H t  tương quan gi a BMI v  Glucose: {correlation:.2f} ')
```

Hệ số tương quan gi a BMI v  Glucose: 0.22

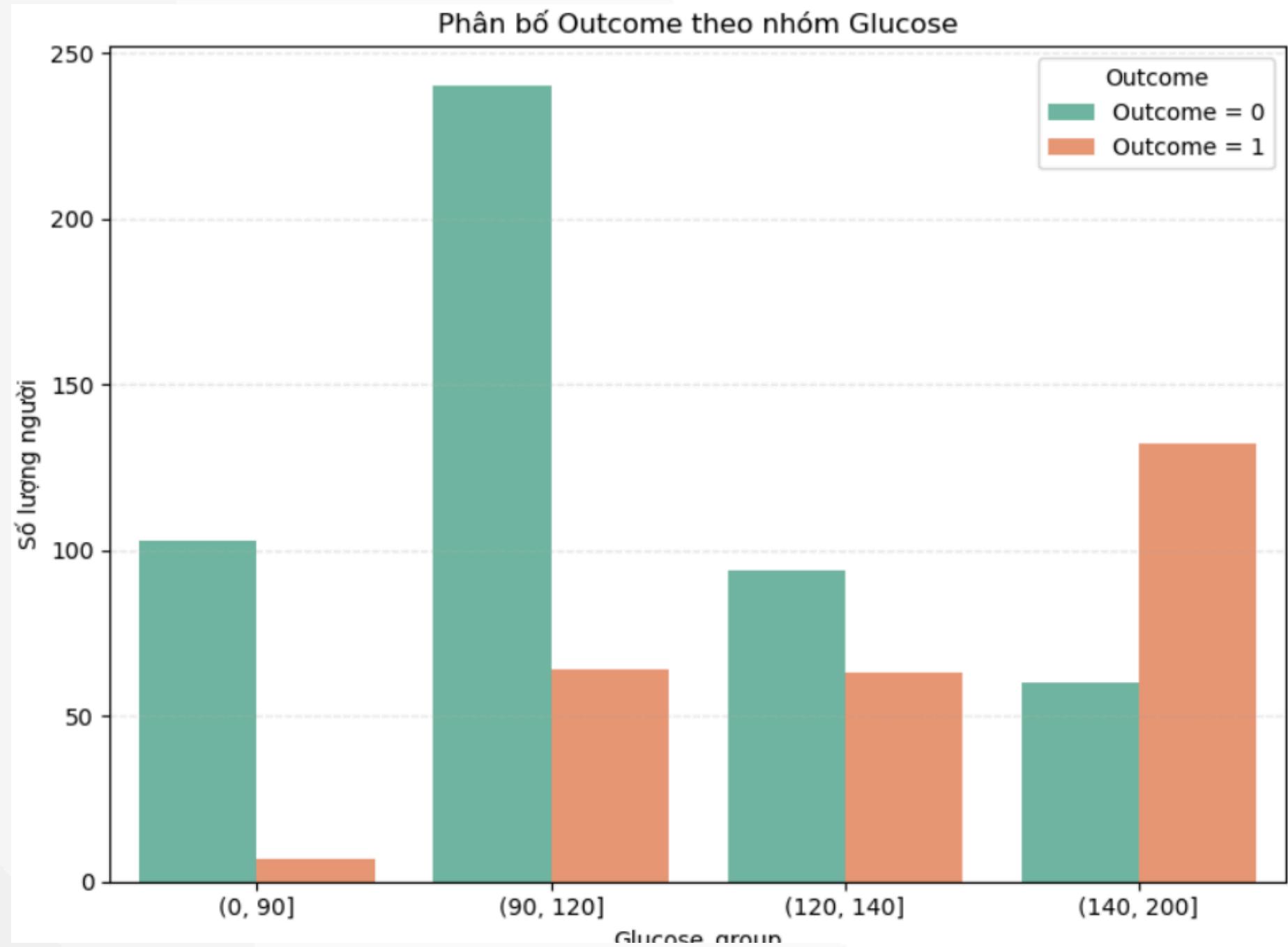


- Các điểm dữ liệu rải rác, không tạo thành đường xu hướng rõ rệt
- Có một số cụm người BMI cao – Glucose cao, nhưng không phổ biến
- Nhóm Outcome = 1 (cam) có xu hướng tập trung ở vùng Glucose cao, nhưng không nhất thiết BMI cao

Phân tích theo Glucose

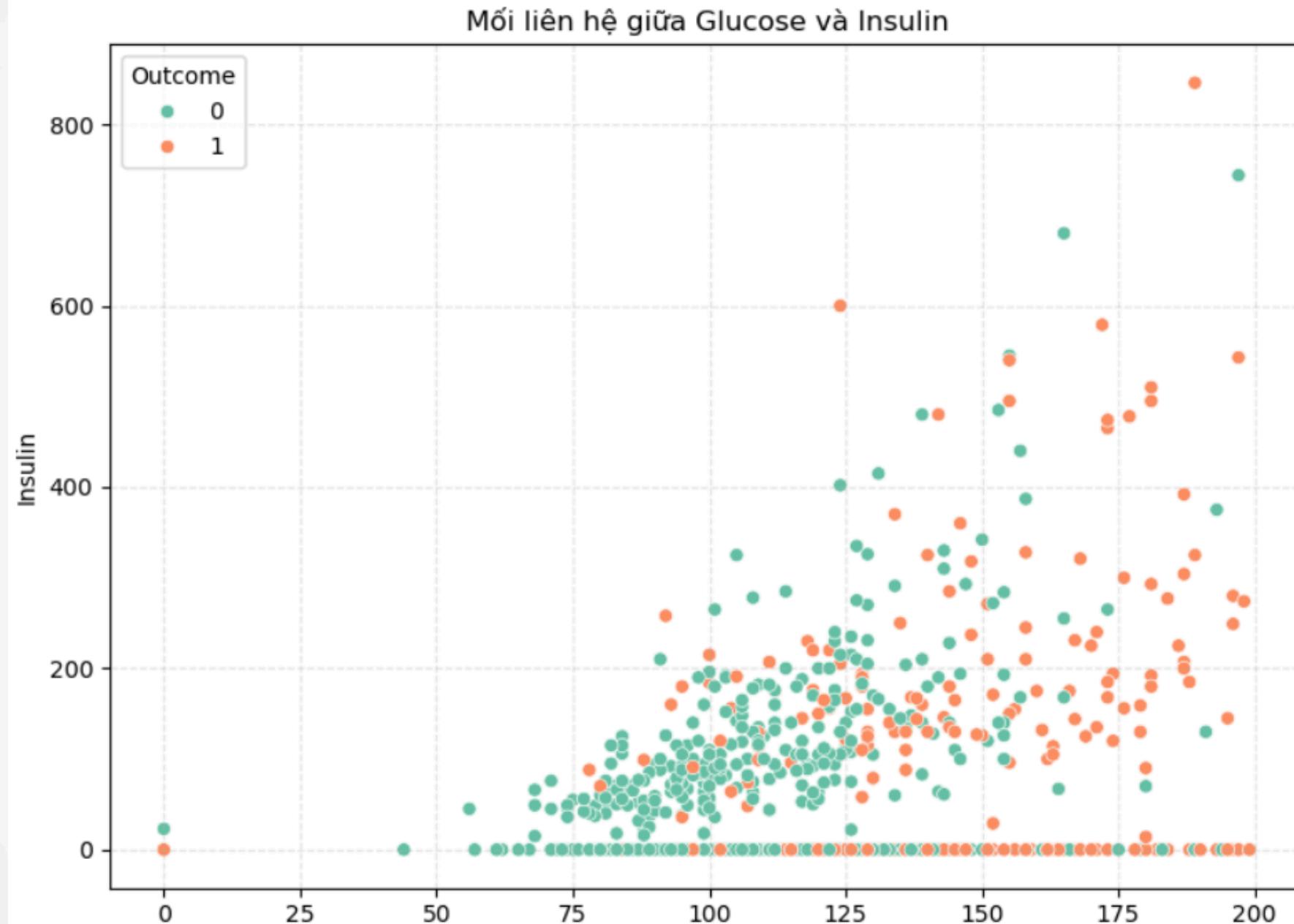
- Phân bố chỉ số Glucose trong dataset như thế nào?
- Insulin có mối liên hệ như thế nào với Glucose?
- Glucose và Age có mối quan hệ tuyến tính không?

Phân tích theo Glucose



- Tỷ lệ Outcome = 1 tăng rõ rệt ở nhóm Glucose cao (140–200)
- Nhóm (90–120] có tỷ lệ Outcome = 0 cao nhất → vùng “an toàn”
- Glucose là biến phân loại mạnh, có thể được chọn làm feature chính

Insulin có mối liên hệ như thế nào với Glucose?



```
# Tính hệ số tương quan giữa Insulin và Glucose  
corr = df['Insulin'].corr(df['Glucose'])  
print(f"Hệ số tương quan giữa Insulin và Glucose: {corr:.2f}")
```

Hệ số tương quan giữa Insulin và Glucose: 0.33

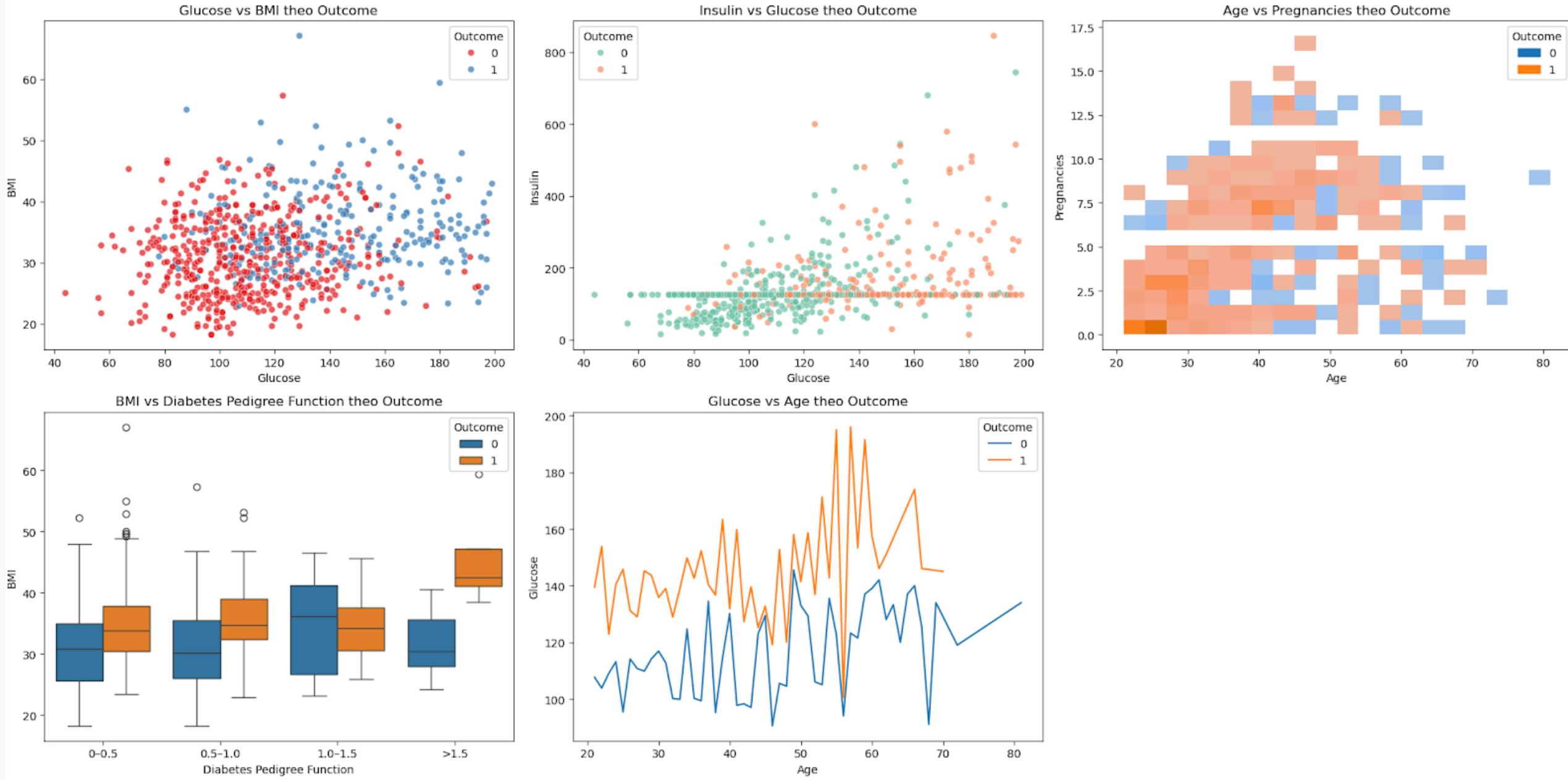
- Một số cụm Glucose cao – Insulin cao xuất hiện ở nhóm Outcome = 1
- Mối liên hệ giữa Glucose và Insulin có xu hướng cùng chiều, mức độ trung bình

Một số câu hỏi quan trọng



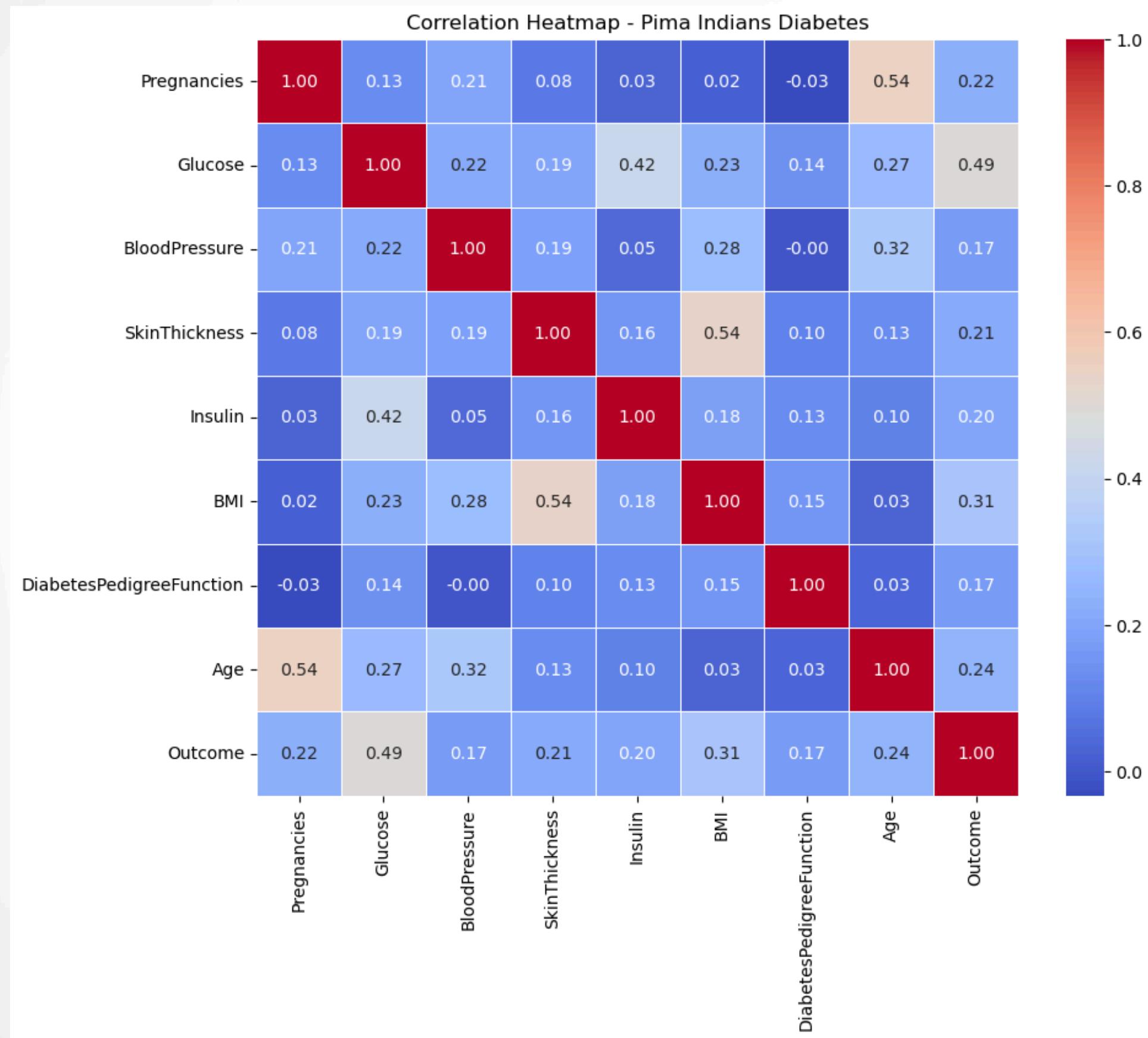
- Glucose + BMI: Kết hợp 2 biến này thì nhóm mắc tiểu đường phân bố ở đâu?
- Insulin + Glucose: Hai biến này có quan hệ tuyến tính không?
Outcome=1 nằm tập trung ở vùng nào?
- Pregnancies + Age: Phụ nữ nhiều lần mang thai và lớn tuổi có nguy cơ cao hơn không?
- BMI + DPF: Người có BMI cao và DPF cao thì tỉ lệ mắc bệnh ra sao?
- Glucose + Age: Người lớn tuổi với Glucose cao có nguy cơ cao nhất không?

Một số câu hỏi quan trọng



- Glucose & BMI: Nhóm Outcome=1 tập trung ở Glucose ≥ 120 và BMI ≥ 30 .
- Insulin: Outcome=1 có Insulin rải rác nhưng thường cao hơn nhóm 0.
- Pregnancies & Age: Nguy cơ tăng rõ ở phụ nữ >35 tuổi và mang thai >5 lần.
- BMI & DPF: Cùng cao (BMI >30 & DPF >1.0) \rightarrow nguy cơ mắc bệnh cao nhất.
- Age & Glucose: Người ≥ 40 tuổi với Glucose ≥ 140 có nguy cơ cao nhất.

Ma trận tương quan



- Glucose (0.49): mạnh nhất → người có Glucose cao dễ mắc tiểu đường.
- BMI (0.31): tương quan khá rõ → BMI cao tăng khả năng mắc bệnh.
- Age (0.24): người lớn tuổi có nguy cơ cao hơn.
- SkinThickness và BMI (0.54): khá cao → có thể gây multicollinearity.
- Pregnancies và Age (0.54): tương quan cao (tuổi lớn → số lần mang thai nhiều).
- BloodPressure (0.17), SkinThickness (0.21), Insulin (0.20): liên quan yếu với Outcome.

Kết luận

- Glucose là yếu tố phân biệt mạnh nhất: người có Glucose cao dễ mắc tiểu đường.
- BMI và Age cũng có ảnh hưởng rõ: BMI cao và tuổi lớn làm tăng nguy cơ mắc bệnh.
- Pregnancies có vai trò nhất định, đặc biệt khi kết hợp với tuổi (phụ nữ lớn tuổi, nhiều lần mang thai → nguy cơ cao).
- Diabetes Pedigree Function (DPF) góp phần bổ sung, nguy cơ tăng mạnh khi kết hợp với BMI cao.
- BloodPressure, SkinThickness, Insulin có liên quan yếu đến Outcome, ít giá trị phân biệt.
- Một số cặp biến có tương quan cao (multicollinearity) như SkinThickness-BMI, Pregnancies-Age.
- Bộ dữ liệu mất cân bằng lớp (Outcome=0 chiếm đa số), cần xử lý khi huấn luyện mô hình.
-> Nhìn chung: Glucose, BMI, Age và Pregnancies là những đặc trưng quan trọng nhất để dự báo tiểu đường trong cộng đồng Pima Indians.