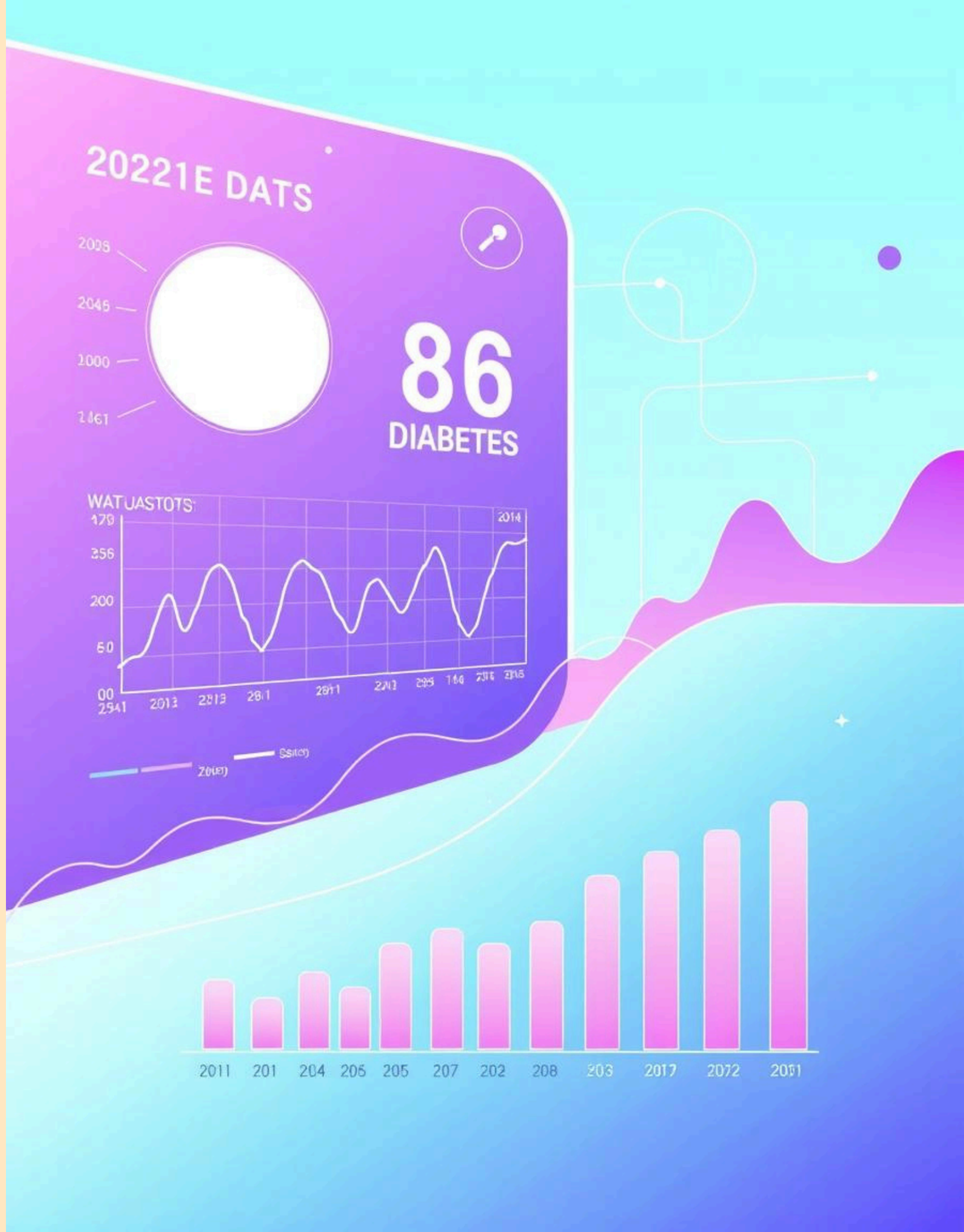
Báo cáo Lab 4

Pima-indians-diabetes Analyst



Nhóm ()

Phân công

Nguyễn Thái Vinh

Tổng hợp và tóm tắt tài liệu

Phạm Thanh Tuấn

Soạn thảo và chỉnh sửa code

Đặng Nhật Đức

Võ Văn Tuấn

Thiết kế Slide bạn đang xem

Định nghĩa vấn đề:

Phân tích dữ liệu về việc mắc bệnh tiểu đường đối với những phụ nữ thuộc bộ tộc Pima ở Arizona Mỹ

Đơn vị thu thập dữ liệu

Viện Quốc gia về Tiểu đường và Bệnh tiêu hóa, Thận (National Institute of Diabetes and Digestive and Kidney Diseases – NIDDK).

Tổng quan dữ liệu

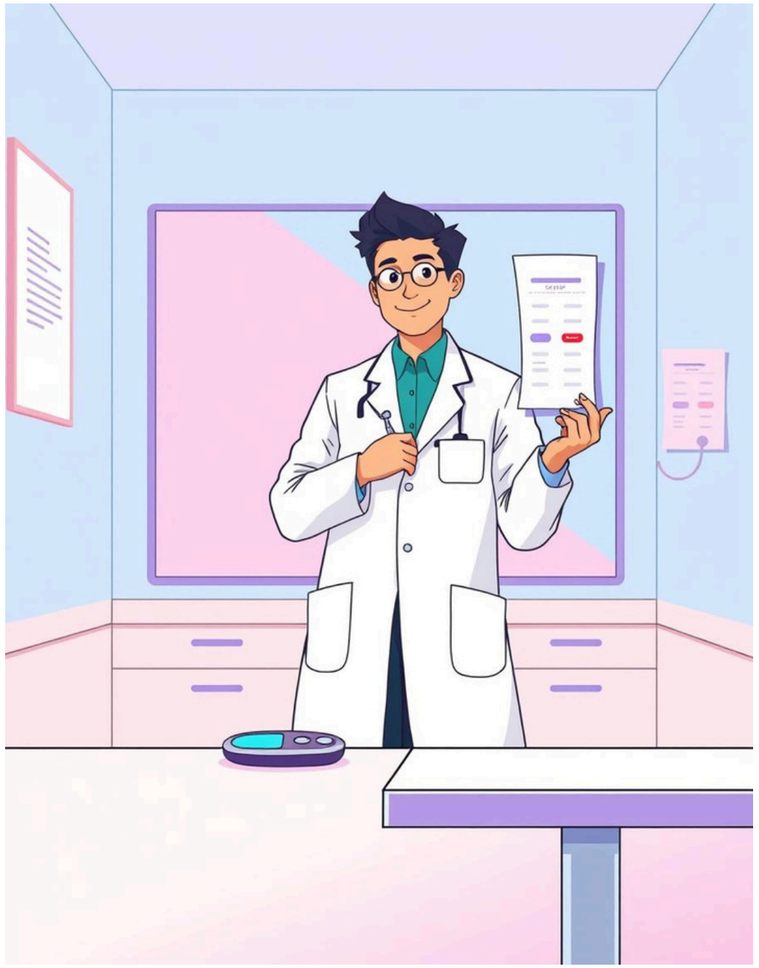
768 phụ nữ da đỏ thuộc bộ lạc Pima, sống ở Arizona, Hoa Kỳ.

9 thuộc tính

1 trường Kết quả dạng binary

Chi tiết các trường dữ liệu

1. **pregnancies**: Số lần mang thai
2. **glucose**: Nồng độ glucose huyết tương
3. **blood_pressure**: Huyết áp tâm trương
4. **skin_thickness**: Độ dày nếp gấp da
5. **insulin**: Insulin huyết thanh 2 giờ
6. **bmi**: chỉ số khối cơ thể
7. **diabetes_pedigree_function**: Hàm di truyền tiểu đường
8. **age:** độ tuổi
9. **outcome:** kết quả co bị tiểu đường hay không



Tính toàn vẹn của dữ liệu

0

dữ liệu bị trống

dữ liệu bị lỗi

mẫu trùng

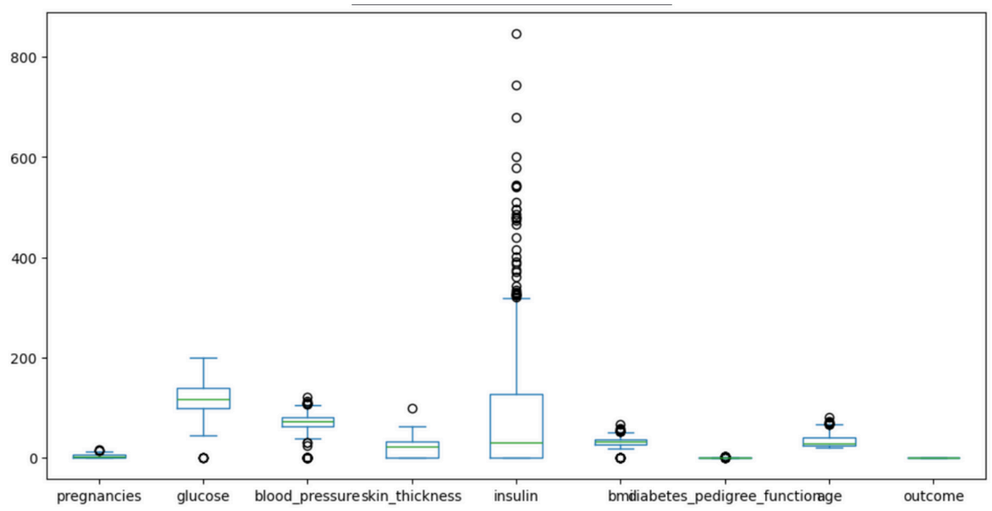
Xác Định Tính Chất và Thách Thức của Dữ Liệu

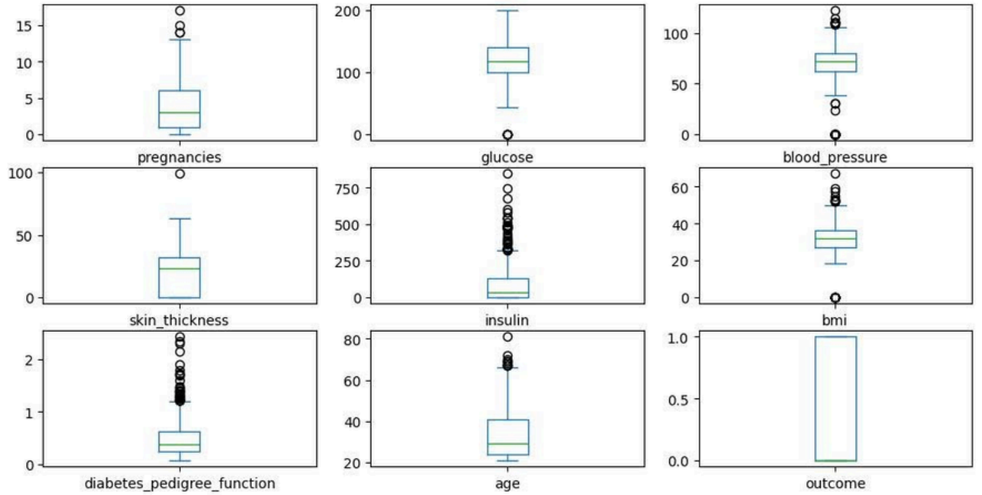
- **Giá Trị Insulin:** Nhiều giá trị 0 không hợp lý trong trường hợp insulin, cho thấy khả năng thiếu dữ liệu.
- **Phân Loại Biến**
  Biến định lượng: Glucose, BMI, Tuổi, Insulin.
  Biến định tính: Tình trạng mắc bệnh tiểu đường (Có/Không).

- **Độ Dày Da (Skin Thickness):** Tương tự insulin, giá trị 0 trong độ dày da cũng chỉ ra dữ liệu bị thiếu hoặc ghi nhận sai.

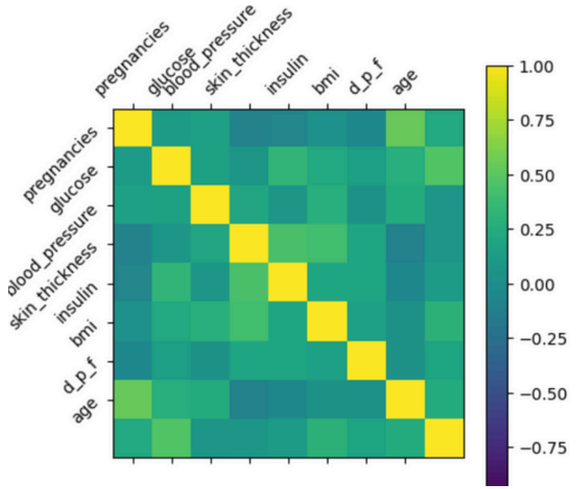
Việc nhận diện và xử lí các giá trị 0 bất hợp lý (missing value) là bước quan trọng để đảm bảo tính toàn vẹn và chính xác của phân tích dữ liệu, đặc biệt đối với các biến như Insulin và độ dày da, nơi giá trị 0 là không thể có trong thực tế

Phân tích dữ liệu





	pregnancies	glucose	blood_pressure	skin_thickness	insulin	bmi	d_p_f	age	outcome
pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
blood_pressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
skin_thickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
bmi	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
d_p_f	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000



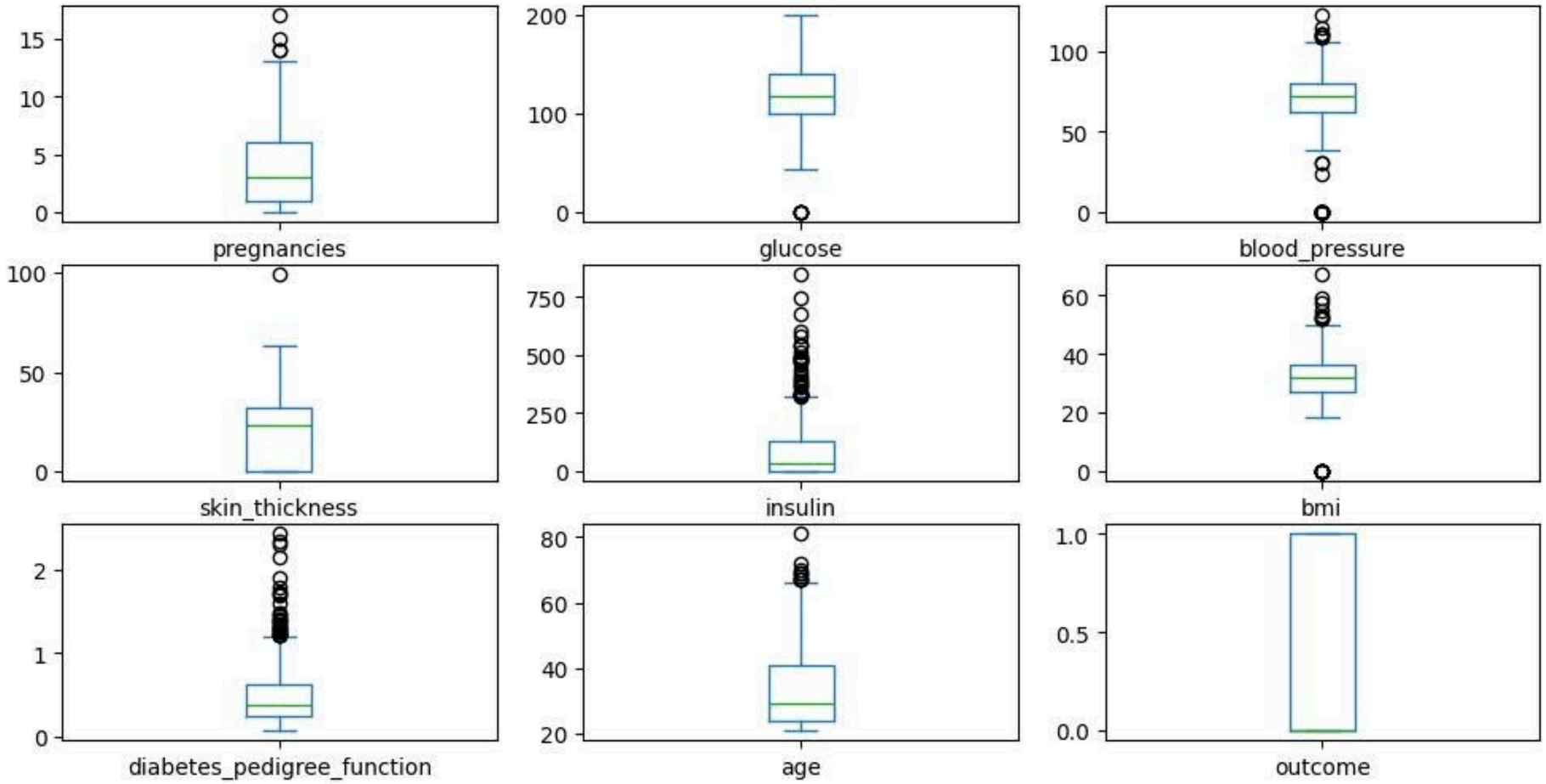
Mối tương quan giữa các tính chất

	pregnancies	glucose	blood_pressure	skin_thickness	insulin	bmi	d_p_f	age	outcome
pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
blood_pressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
skin_thickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
bmi	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
d_p_f	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Nhận xét sự tương quan giữa các đặc trưng

- Glucose và Outcome: có hệ số tương quan khá cao (0.4666), cho thấy mức glucose là yếu tố quan trọng trong việc dự đoán khả năng mắc tiểu đường.
- BMI và Outcome: có mức tương quan dương (0.2927) – những người có BMI cao có xu hướng dễ mắc tiểu đường hơn.
- Pregnancies và Age: có tương quan cao (0.5444), hợp lý vì tuổi càng cao thì số lần mang thai (ở nữ) thường nhiều hơn.
- Skin Thickness và Insulin: tương quan khá mạnh (0.4369).
- Skin Thickness và BMI: cũng có tương quan đáng kể (0.3926).

Phân tích đơn biến



Box and whisker plots

Nhận xét:

Pregnancies(Số lần mang thai):

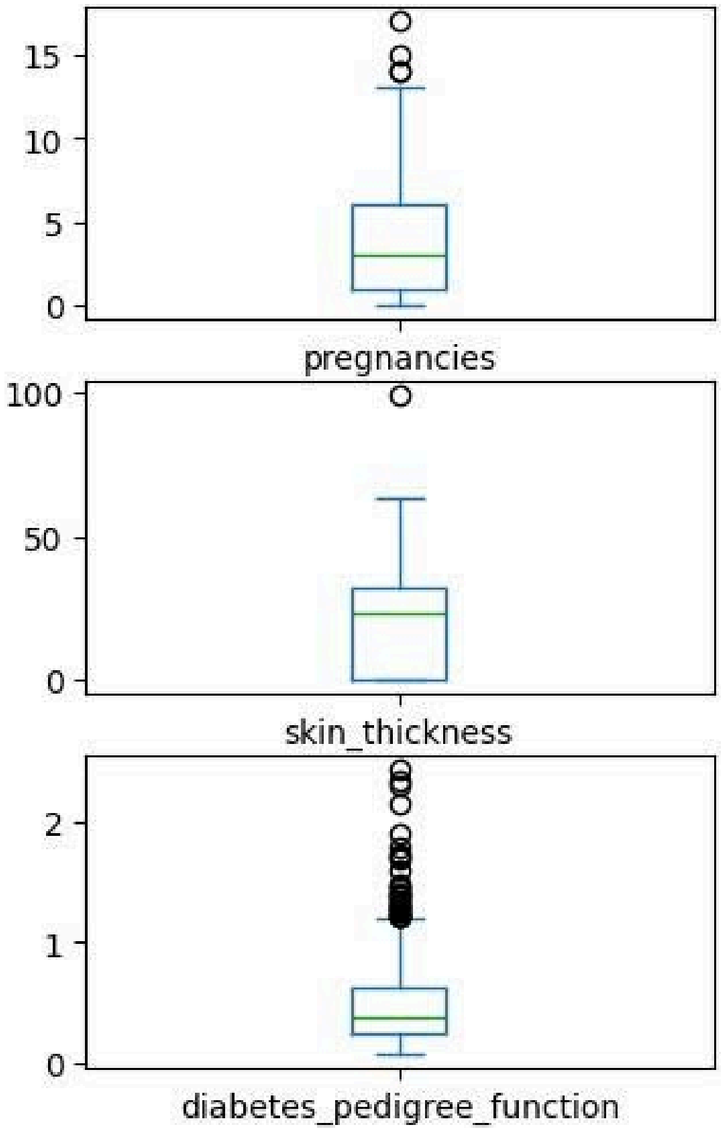
- Phần lớn giá trị nằm trong khoảng 0-6
- Có một số giá trị ngoại lai (outliers) lớn hơn 12

Skin Thickness (độ dày da):

- Nhiều giá trị bằng 0 (có thể dữ liệu thiếu hoặc nhập sai).
- Trung vị khoảng 25.

Diabetes Pedigree Function:

- Phân bố lệch phải (skewed), phần lớn giá trị nhỏ hơn 1.



Glucose (lượng đường huyết):

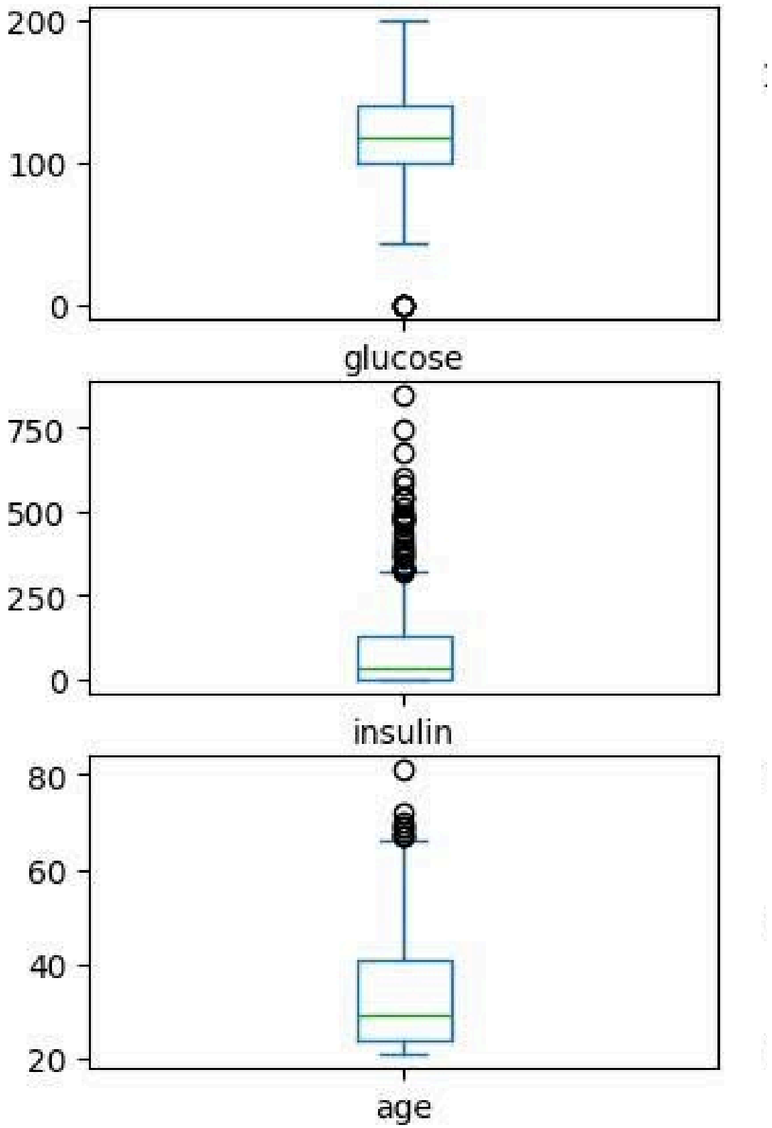
- Giá trị trung vị khoảng 110-120.
- Nhiều giá trị bằng 0 (không hợp lệ) và cao (>200)

Insulin:

- Có rất nhiều giá trị bằng 0 (không thực tế, cho thấy dữ liệu bị thiếu).

Age (tuổi):

- Trung vị khoảng 30.
- Nhiều giá trị ngoại lai ở tuổi trên 70.



Blood Pressure (huyết áp):

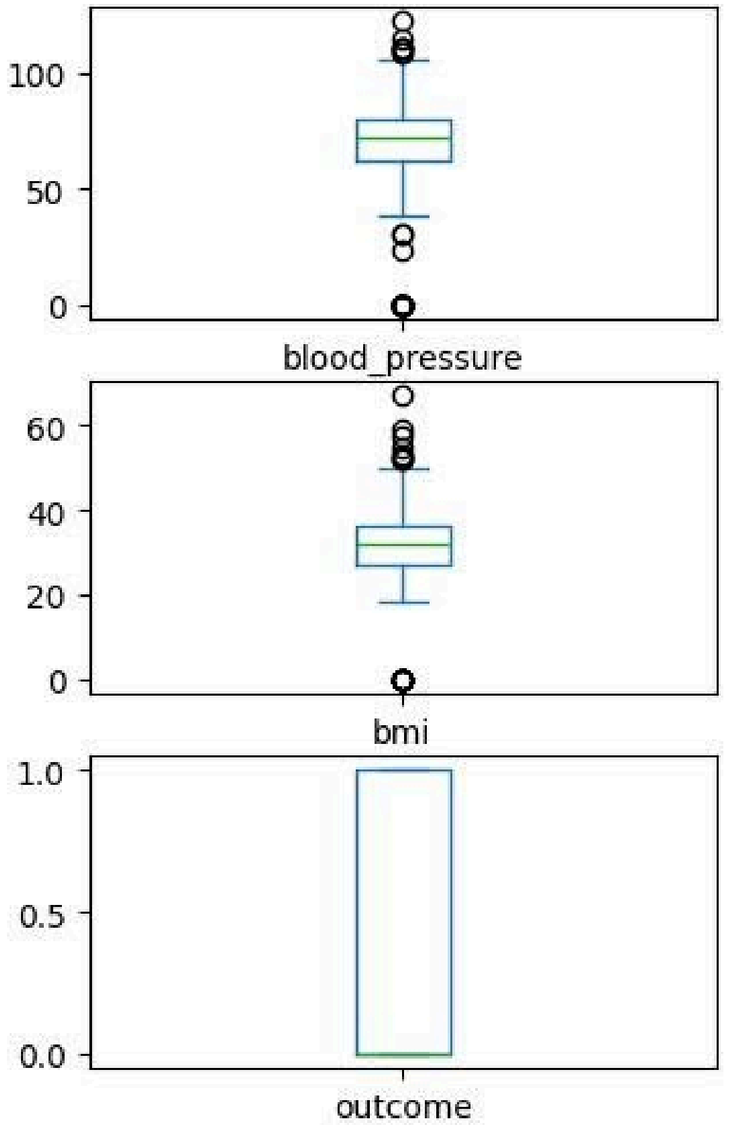
- Trung vị khoảng 70
- Có một số giá trị bằng 0 (không thực tế với dữ liệu huyết áp).
- Một số outliers cao trên 120.

BMI (chỉ số khối cơ thể):

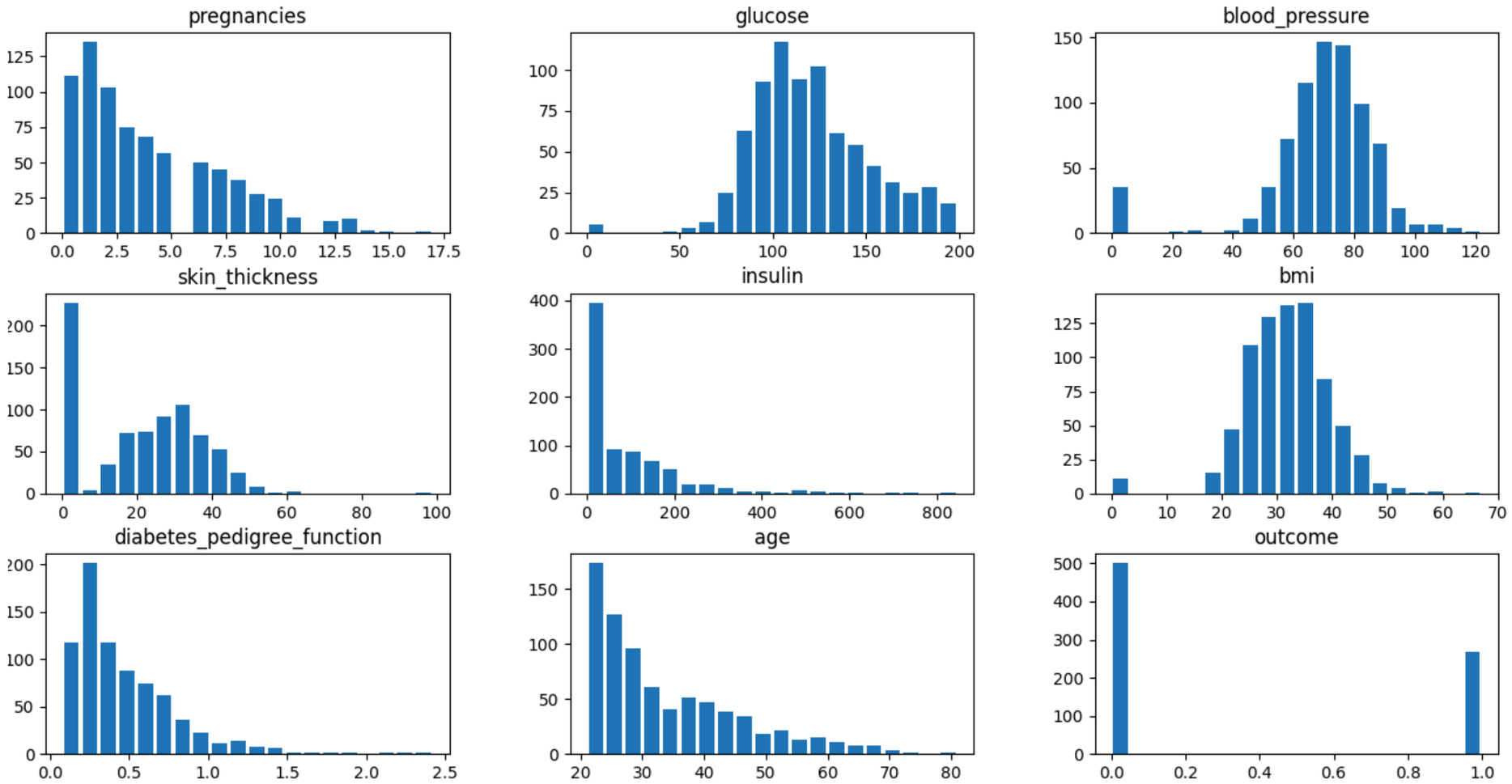
- Trung vị khoảng 30.
- Có một số giá trị bằng 0 (không thực tế).
- Một số outliers trên 60.

Outcome (nhãn bệnh)

- Dữ liệu mất cân bằng: số người không mắc bệnh (0) nhiều hơn đáng kể so với số người mắc bệnh (1).



Phân tích đơn biến

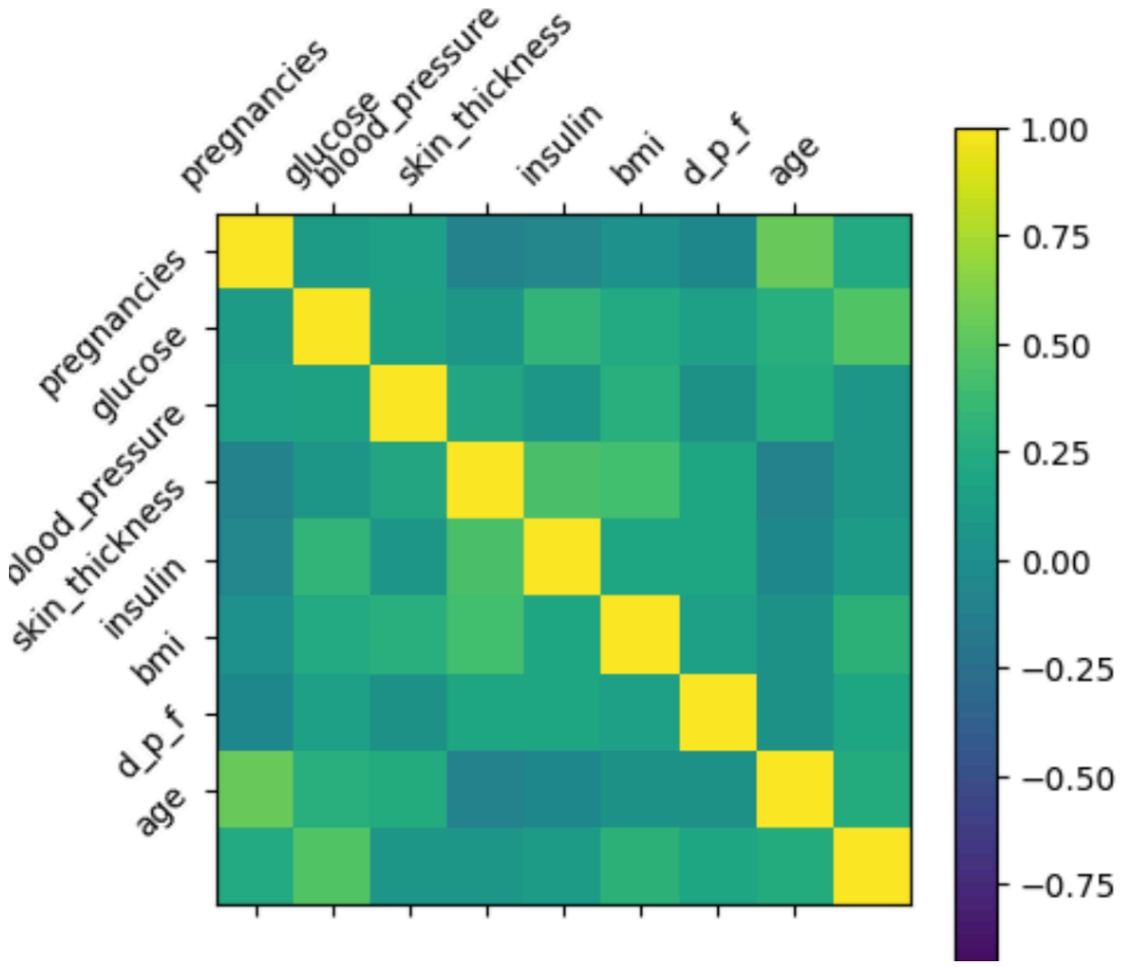


Biểu đồ histogram

Dựa vào biểu đồ ta có thể thấy:

- Nhóm tuổi chiếm nhiều nhất là từ 21 – 30 tuổi.
- Glucose cao nhất ở mức 100 – 125.
- BMI cao nhất là từ khoảng 20 – 40.
- Có 268/768 người được chẩn đoán mắc tiểu đường, 500 người không mắc.
- Số lần mang thai phổ biến nhất là 1 lần mang thai.

Phân tích đa biến



Heatmap Tương quan

Nhận xét

- Đa số cặp biến: Mức tương quan thấp đến trung bình (màu xanh nhạt → xanh dương).
- Glucose – Outcome: Tương quan khá rõ (màu sáng hơn), phù hợp vì glucose ảnh hưởng mạnh đến tiểu đường.
- BMI – Skin Thickness: Tương quan cao (màu sáng gần vàng), hợp lý vì cả hai đều phản ánh mỡ cơ thể.
- Insulin – Glucose: Tương quan trung bình–khá (màu sáng hơn), hợp lý do insulin thường tăng khi glucose cao.
- Pregnancies – Age: Tương quan cao (màu sáng), phù hợp vì phụ nữ lớn tuổi thường có nhiều lần mang thai hơn.
- Các biến khác (Blood Pressure, DPF, Age, ...): Chủ yếu tương quan yếu, màu xanh đậm hơn → ít ảnh hưởng trực tiếp.

Nhận xét:

1. Mối quan hệ giữa các biến độc lập với outcome

- Glucose (0.466): có tương quan mạnh nhất với outcome → mức đường huyết là yếu tố quan trọng nhất trong việc dự đoán tiểu đường.
- BMI (0.293): có tương quan dương, người có BMI cao dễ mắc tiểu đường hơn.
- Age (0.238) và Pregnancies (0.222): có mối tương quan vừa phải, tuổi cao và số lần mang thai nhiều thì nguy cơ mắc tiểu đường tăng.
- Diabetes Pedigree Function (0.174): có mức tương quan thấp nhưng vẫn có ý nghĩa → yếu tố di truyền cũng ảnh hưởng.
- Insulin (0.131), Skin Thickness (0.074), Blood Pressure (0.065): tương quan rất thấp → ít ảnh hưởng trực tiếp đến outcome trong dataset này.

2. Mối quan hệ giữa các biến độc lập với nhau

- Skin Thickness và Insulin (0.437): có tương quan cao → hợp lý vì độ dày da thường liên quan đến mức insulin.
- BMI và Skin Thickness (0.393): có tương quan vừa, phản ánh rằng lớp mỡ dưới da ảnh hưởng đến BMI.
- Pregnancies và Age (0.544): tương quan khá cao, dễ hiểu vì tuổi càng cao thì số lần mang thai có xu hướng nhiều hơn.

Nhận xét chung

- Glucose là biến dự đoán quan trọng nhất.
- BMI, Age, Pregnancies, Diabetes Pedigree Function cũng có đóng góp nhưng nhỏ hơn.
- Một số biến (Blood Pressure, Skin Thickness, Insulin) có tương quan thấp với outcome → cần kiểm tra kỹ, có thể kết hợp theo dạng phi tuyến hoặc tương tác mới có tác dụng.
- Không có cặp biến nào có tương quan quá cao (gần ±1), nên không lo vấn đề đa cộng tuyến nghiêm trọng.

Giá trị Thiếu và Mẫu Bất thường trong Dữ liệu

Trong bộ dữ liệu y tế, việc phát hiện và xử lý các giá trị thiếu (missing values) và mẫu bất thường (anomalous samples) là rất quan trọng để đảm bảo tính chính xác của phân tích.

Giá trị 0 Bất hợp lý

Đối với các biến như Insulin và Độ dày da (Skin Thickness), việc có nhiều giá trị bằng 0 thường được xem là các giá trị thiếu, vì trên thực tế các chỉ số này không thể bằng 0.
Tuy nhiên, các giá trị 0 cho Huyết áp hoặc Glucose lại cực kỳ hiếm và có thể coi là dữ liệu bất thường (outliers) hoặc lỗi nhập liệu nghiêm trọng.

Chiến Lược Xử Lý

Để duy trì tính toàn vẹn của dữ liệu, có thể thay thế các giá trị bị thiếu bằng giá trị trung vị (median) hoặc giá trị trung bình (mean), hoặc loại bỏ các mẫu dữ liệu có giá trị bất thường nếu chúng chiếm tỷ lệ nhỏ.

Xử lí các giá trị 0 (zero values)

Trước hết thay thế các giá trị 0 bằng NaN

```
df_clean['glucose'].replace(0, np.nan, inplace=True)
df_clean['blood_pressure'].replace(0, np.nan, inplace=True)
df_clean['skin_thickness'].replace(0, np.nan, inplace=True)
df_clean['insulin'].replace(0, np.nan, inplace=True)
df_clean['bmi'].replace(0, np.nan, inplace=True)
```

Sau khi kiểm tra ta có được số lượng zero values của từng thuộc tính

```
df_clean.isna().sum()
```

	0
pregnancies	0
glucose	5
blood_pressure	35
skin_thickness	227
insulin	374
bmi	11
diabetes_pedigree_function	0
age	0
outcome	0

Với các thuộc tính bị thiếu ít giá trị, ta điền các giá trị bị thiếu bằng median

```
df_clean['glucose'].fillna(df_clean['glucose'].median(), inplace=True)
df_clean['blood_pressure'].fillna(df_clean['blood_pressure'].median(), inplace=True)
df_clean['bmi'].fillna(df_clean['bmi'].median(), inplace=True)
```

Skin thickness có độ tương quan cao với BMI cả về mặt dữ liệu và sinh học thực tế, ta phân theo nhóm BMI và điền median tương ứng với từng nhóm

```
df_clean['bmi_group'] = pd.cut(df_clean['bmi'], bins=[0, 18.5, 25, 30, 35, 100],
                               labels=['Underweight', 'Normal', 'Overweight', 'Obese I', 'Obese II+'])
df_clean['skin_thickness'].fillna(df_clean.groupby('bmi_group')['skin_thickness'].transform('median'), inplace=True)
```

Insulin tương quan cao với skin thickness và glucose, ta điền median theo nhóm

```
df_clean['insulin'] = df_clean.groupby(['glucose', 'skin_thickness'])['insulin']\
                  .transform(lambda x: x.fillna(x.median()))

# Nếu vẫn còn NaN (vì nhóm nào đó toàn bộ insulin đều NaN), có thể điền thêm median chung
df_clean['insulin'] = df_clean['insulin'].fillna(df_clean['insulin'].median())
```

Kết quả

	pregnancies	glucose	blood_pressure	skin_thickness	insulin	bmi	diabetes_pedigree_function	age	outcome
pregnancies	1.000000	0.128213	0.208615	0.078829	0.023145	0.021559	-0.033523	0.544341	0.221898
glucose	0.128213	1.000000	0.218937	0.213698	0.423182	0.231049	0.137327	0.266909	0.492782
blood_pressure	0.208615	0.218937	1.000000	0.210935	0.047135	0.281257	-0.002378	0.324915	0.165723
skin_thickness	0.078829	0.213698	0.210935	1.000000	0.150196	0.703276	0.125048	0.097404	0.266268
insulin	0.023145	0.423182	0.047135	0.150196	1.000000	0.178974	0.121377	0.103604	0.211731
bmi	0.021559	0.231049	0.281257	0.703276	0.178974	1.000000	0.153438	0.025597	0.312038
diabetes_pedigree_function	-0.033523	0.137327	-0.002378	0.125048	0.121377	0.153438	1.000000	0.033561	0.173844
age	0.544341	0.266909	0.324915	0.097404	0.103604	0.025597	0.033561	1.000000	0.238356
outcome	0.221898	0.492782	0.165723	0.266268	0.211731	0.312038	0.173844	0.238356	1.000000

So sánh ma trận tương quan trước & sau khi điền giá trị

- Glucose – Outcome: Giữ nguyên (~0.49), mối liên hệ chính không thay đổi.
- Insulin – Glucose: Tăng (0.42 so với ~0.25), hợp lý hơn về mặt sinh lý.
- BMI – SkinThickness: Tăng mạnh (0.70 so với ~0.58), do liên quan trực tiếp khi điền.
- Blood Pressure – BMI: Tăng nhẹ (0.28 so với ~0.22), do điền theo nhóm.
- Các biến khác (Age, Pregnancies, DPF): Ít thay đổi.

Kết luận: Sau khi điền, mối quan hệ sinh lý hợp lý hơn

Ngoại lệ trong dữ liệu nghiên cứu

Những trường hợp được chẩn đoán đái tháo đường trong vòng 1 năm sau lần khám (index examination) sẽ bị loại bỏ.

- Lý do: đây là những ca **"quá dễ dự đoán"** vì phần lớn (75%) sẽ phát bệnh trong vòng 6 tháng.
- Nếu giữ lại sẽ khiến mô hình học nhanh nhưng không phản ánh đúng khả năng dự báo lâu dài (5 năm).paper 2

Ý nghĩa

- Đây là cách xử lý ngoại lệ (outlier handling) trong dữ liệu y sinh học:
- Loại bỏ những trường hợp có nguy cơ **làm sai lệch** mô hình học máy.
- Đảm bảo dữ liệu huấn luyện phản ánh đúng **mục tiêu dự báo** (phát bệnh trong 5 năm, không phải chỉ vài tháng).

Thank you