

# **PHÂN TÍCH KHÁM PHÁ VỀ BỆNH ĐÁI THẢO DƯỜNG**

**Thực hiện: Nhóm SKTT**

# Giới Thiệu Đề Tài

Phân tích dữ liệu bệnh tiểu đường đóng vai trò quan trọng trong chẩn đoán sớm, cá nhân hóa điều trị và phòng ngừa biến chứng. Đề tài sử dụng hai nguồn dữ liệu khác nhau để khám phá đặc điểm sinh học, hành vi điều trị và khả năng dự đoán bệnh. Việc kết hợp Pima Indians Dataset và UCI Diabetes Dataset giúp đánh giá hiệu quả các phương pháp phân tích trong nhiều ngữ cảnh lâm sàng

# THÀNH VIÊN



Thiên Sơn



Minh Khang



Thanh Tâm



Duy Tân

# Bộ Dữ Liệu

## PIMA INDIANS DIABETES

### MÔ TẢ

- 768 mẫu từ phụ nữ Pima Indian
- 8 chỉ số y tế được đo

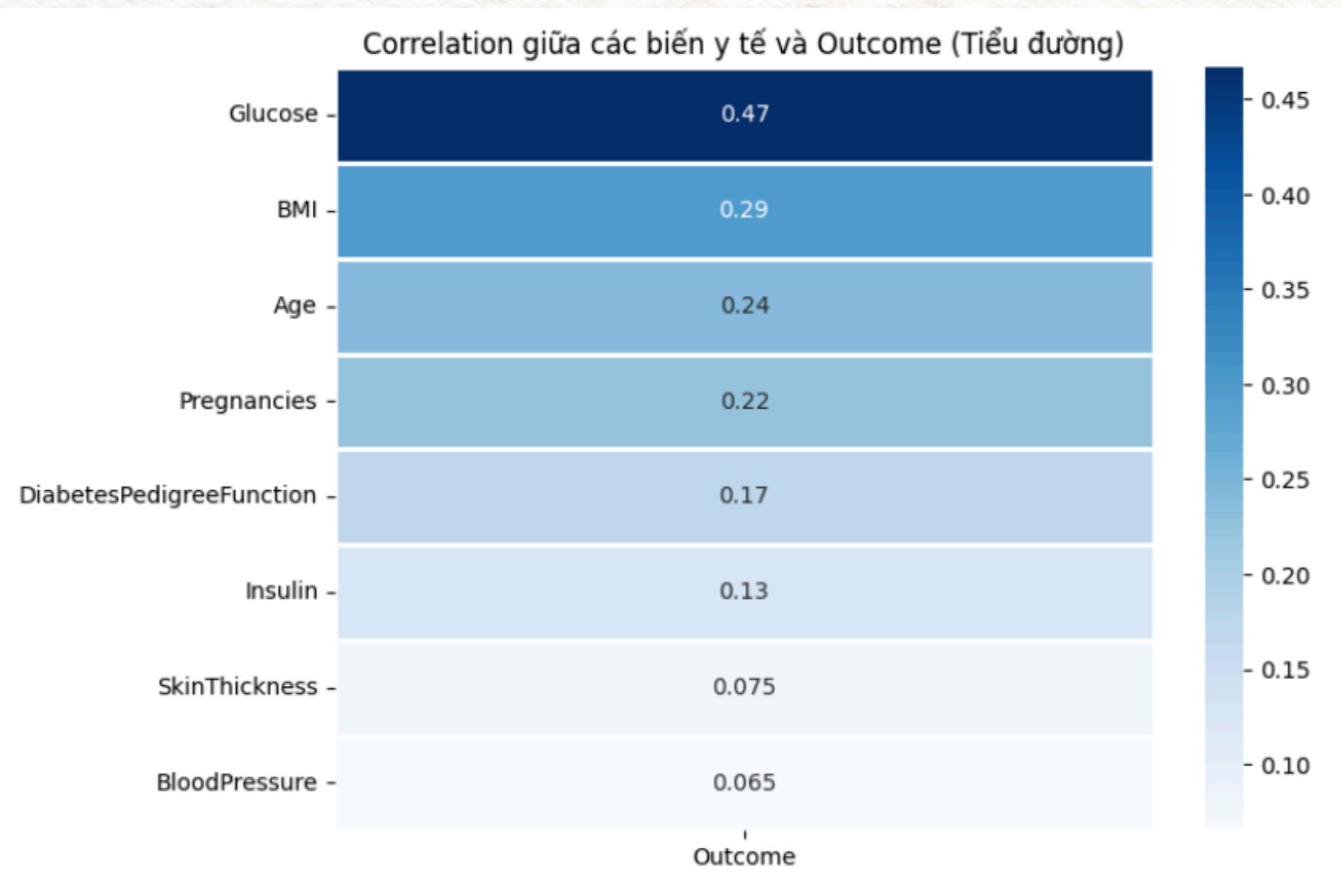
### CÁC BIỂN

- Pregnancies, Glucose, BloodPressure, SkinThickness
- Insulin, BMI, DiabetesPedigreeFunction, Age
- Outcome

# DATA SUMMARY

- **Pregnancies:** Số lần mang thai
- **Glucose:** Nồng độ glucose huyết tương sau 2 giờ OGTT
- **BloodPressure:** Huyết áp (mm Hg)
- **SkinThickness:** Độ dày nếp gấp da ở tay sau (mm)
- **Insulin:** Insulin huyết thanh sau 2 giờ, mu U/ml)
- **BMI:** Chỉ số khối cơ thể, kg/m<sup>2</sup>)
- **DiabetesPedigreeFunction:** Hàm phả hệ tiểu đường
- **Age:** Tuổi
- **Kết quả:** Outcome (0: Không mắc tiểu đường, 1: Mắc tiểu đường)

# MỐI TƯƠNG QUAN GIỮA CÁC BIẾN Y TẾ VÀ KẾT QUẢ



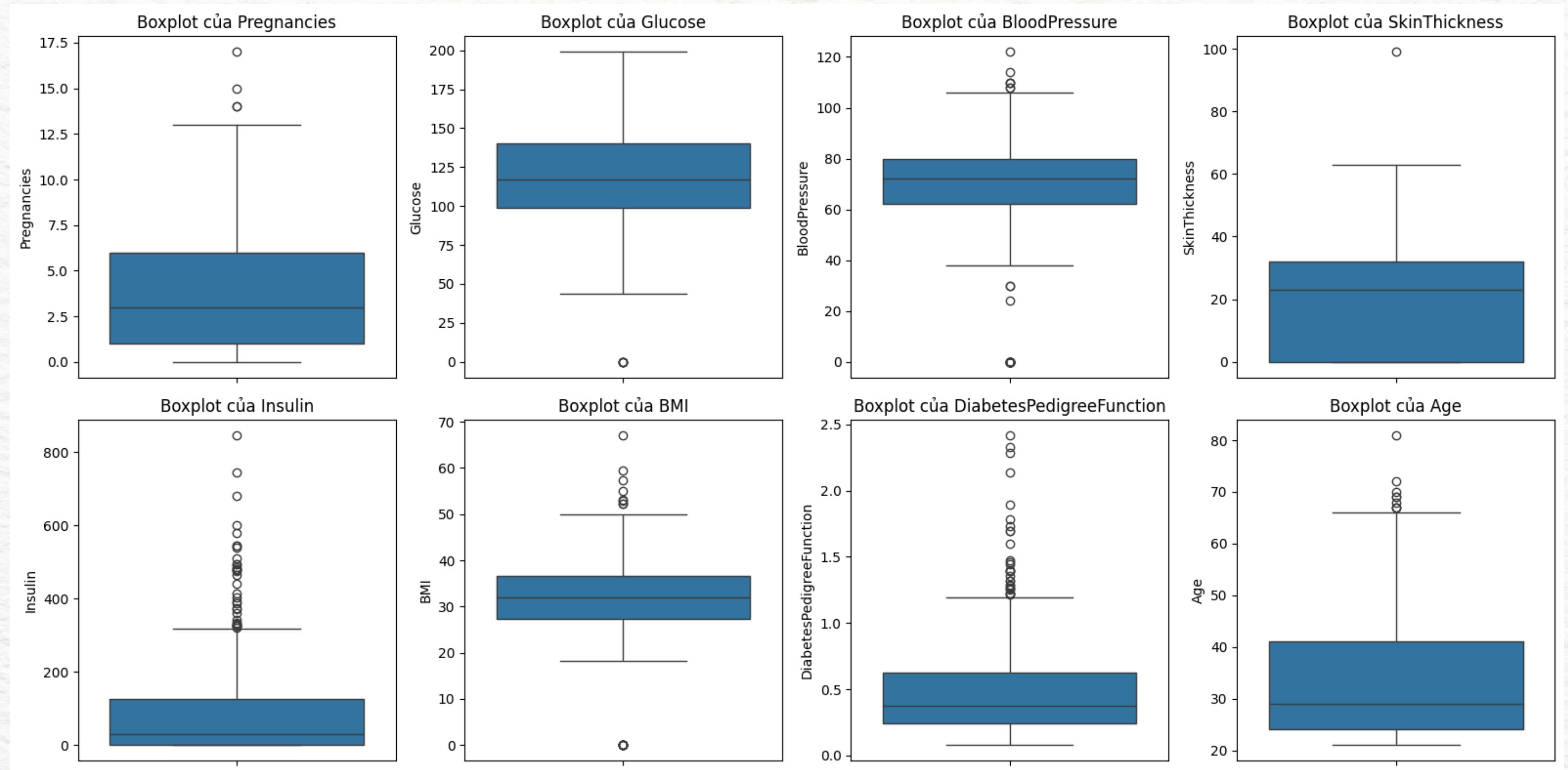
## NHẬN XÉT

- **Glucose** có tương quan trung bình ( $0.47 \rightarrow$  quan trọng nhất).
- **BMI, Age, Pregnancies, DiabetesPedigreeFunction** → tương quan yếu ( $0.17-0.29$ ).
- **Insulin, SkinThickness, BloodPressure** → gần như không liên quan ( $<0.13$ ).

# PHÂN TÍCH THEO đơn BIẾN

## CÂU HỎI

1. Phần lớn phụ nữ trong bộ dữ liệu có bao nhiêu lần mang thai?
2. Phần lớn người trong dữ liệu có Glucose và BMI ở mức bình thường hay cao?
3. Biến nào có nhiều giá trị bất thường cần lưu ý?



- Đa số phụ nữ trong bộ dữ liệu có khoảng 3 lần mang thai, tuổi trung vị 29 (9 cá nhân cao tuổi khác biệt).
- Glucose và BMI tập trung quanh mức cao (Glucose ~117, BMI ~32 → nhiều người ở mức thừa cân, béo phì so với chuẩn (Glucose < 100, BMI < 30)).
- BloodPressure, DiabetesPedigreeFunction và Insulin lại có khá nhiều giá trị bất thường (outlier), dữ liệu hơi lộn xộn.

# PHÂN TÍCH THEO KẾT QUẢ

## CÂU HỎI

1. Phân bố Glucose giữa nhóm mắc và không mắc có khác biệt thống kê đáng kể không?
2. Nguy cơ mắc tiểu đường có tăng theo tuổi không?
3. Có thể xác định ngưỡng BMI nguy cơ cao không?
4. Liệu có ngưỡng số lần mang thai mà nguy cơ tăng mạnh không?

## CHỌN CHỈ SỐ Y TẾ

Chọn chỉ số có độ tương quan ẩn tượng

- **Pregnancies**
- **Glucose**
- **Insulin**
- **BMI**

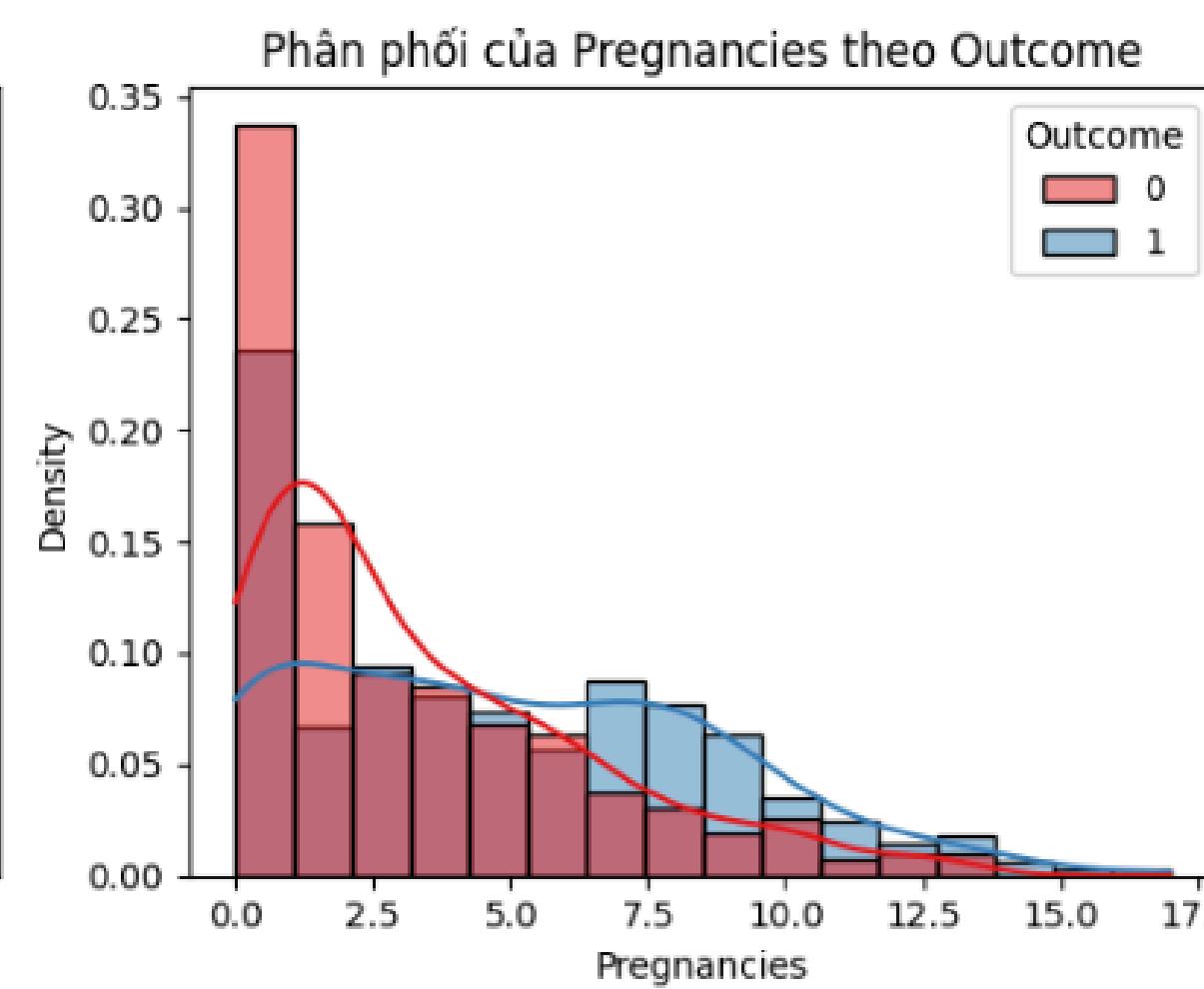
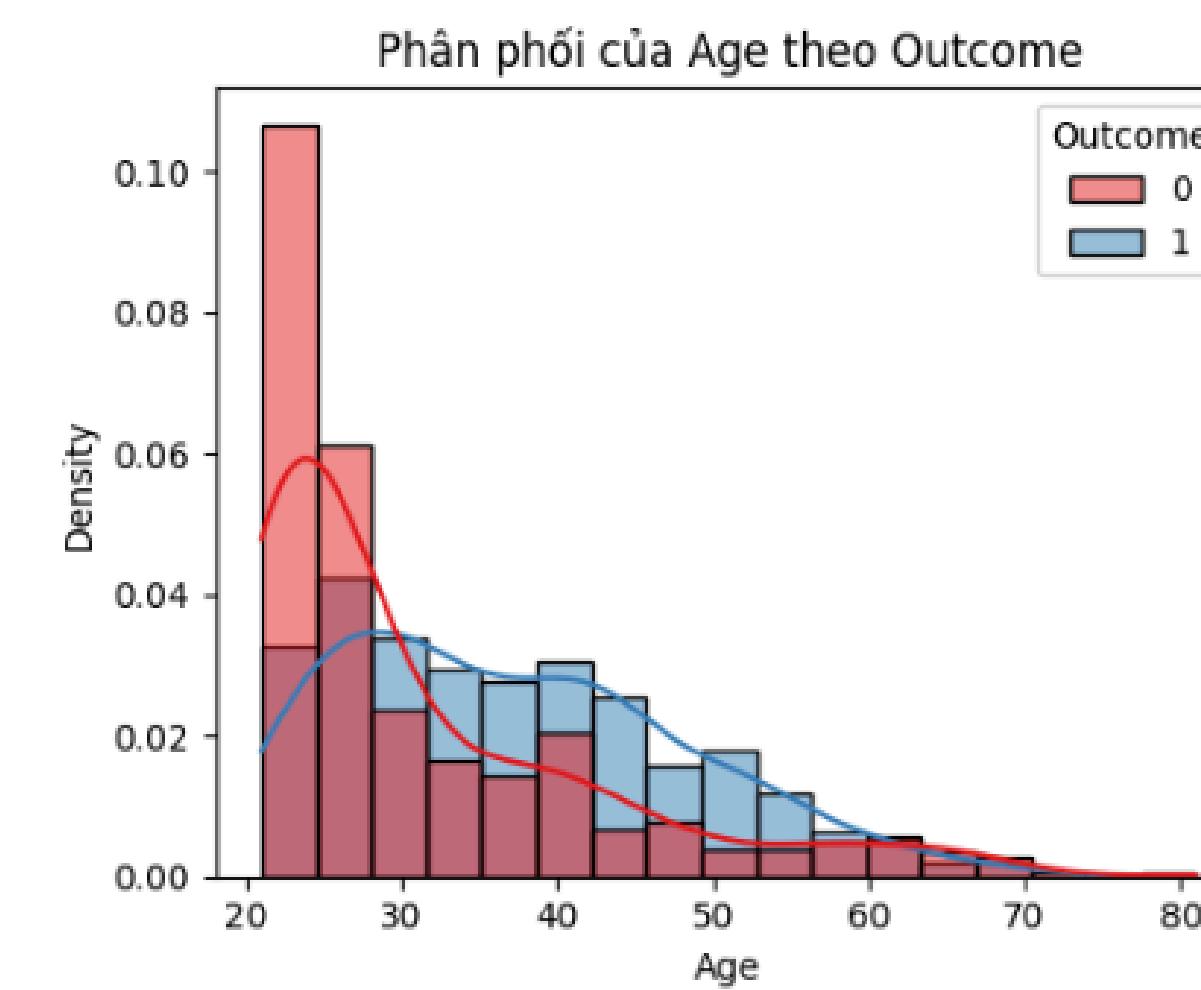
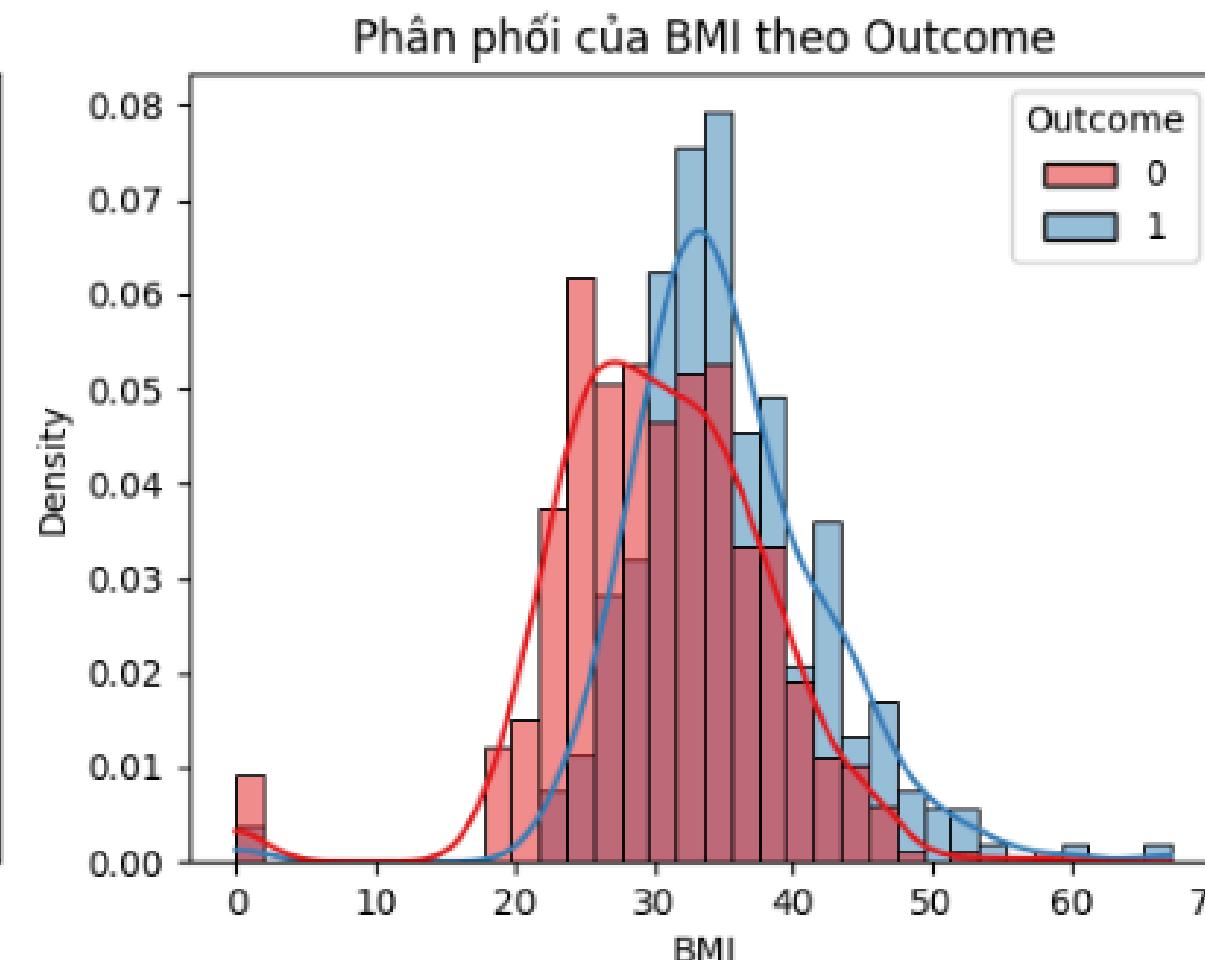
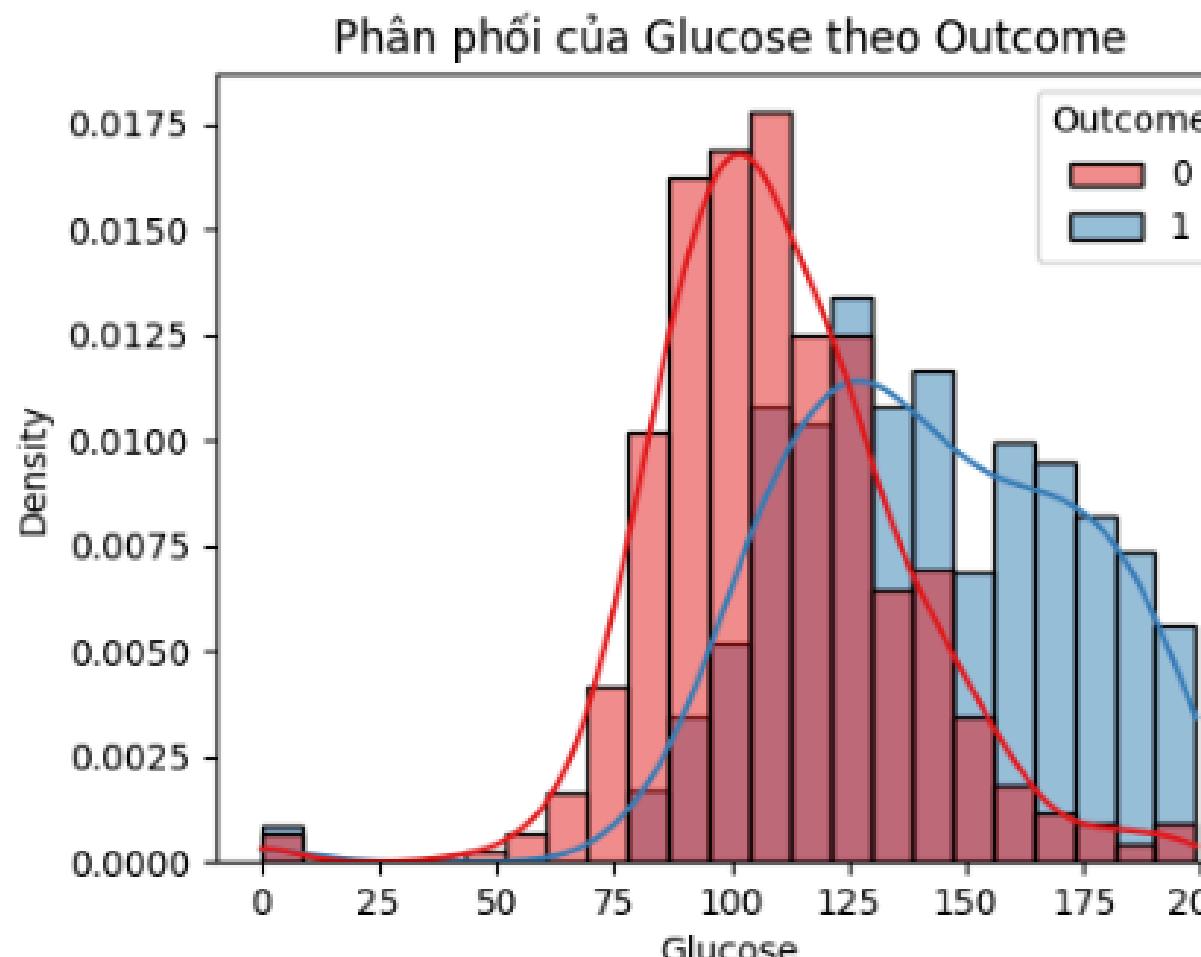
## NHẬN XÉT

### Nhóm không tiểu đường

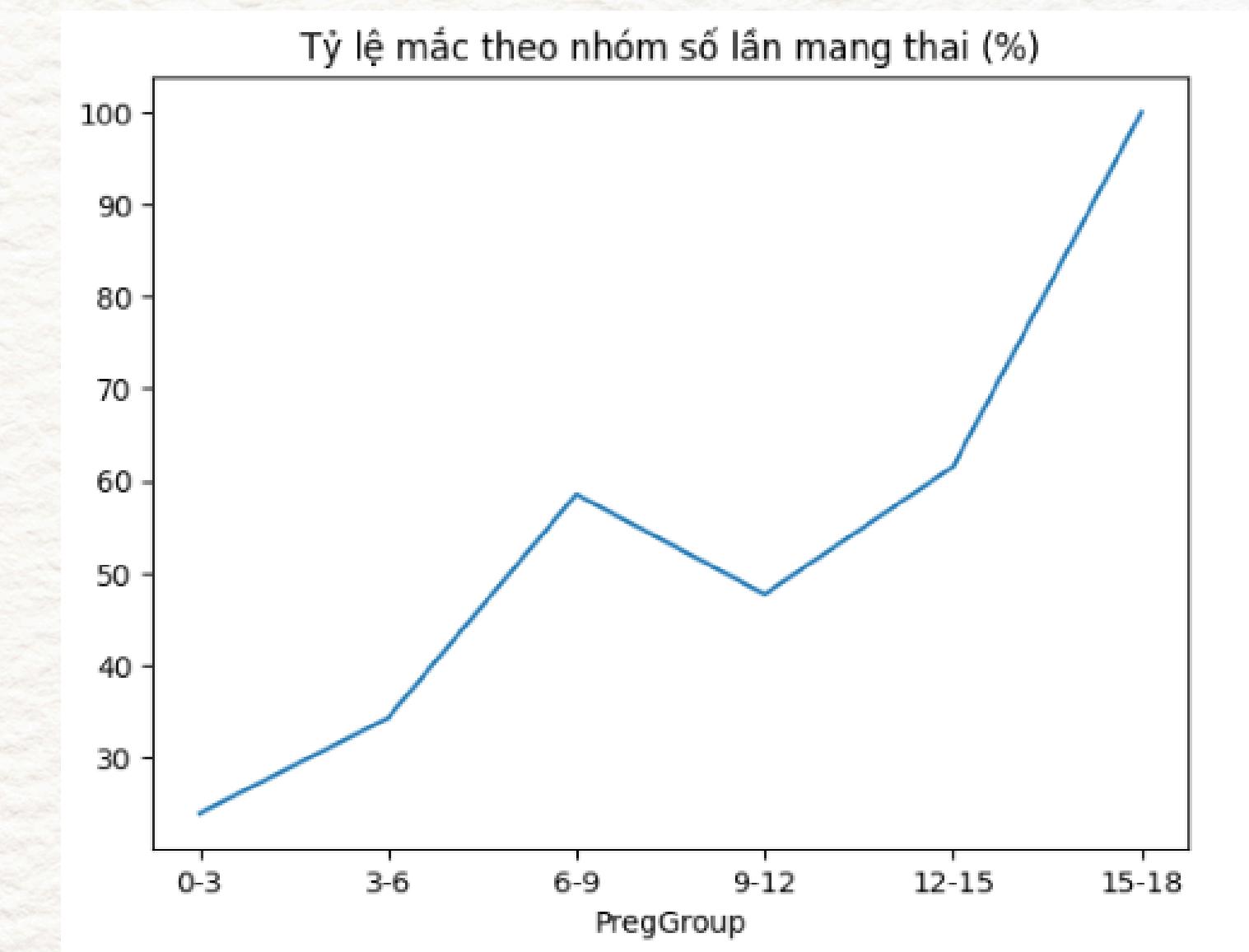
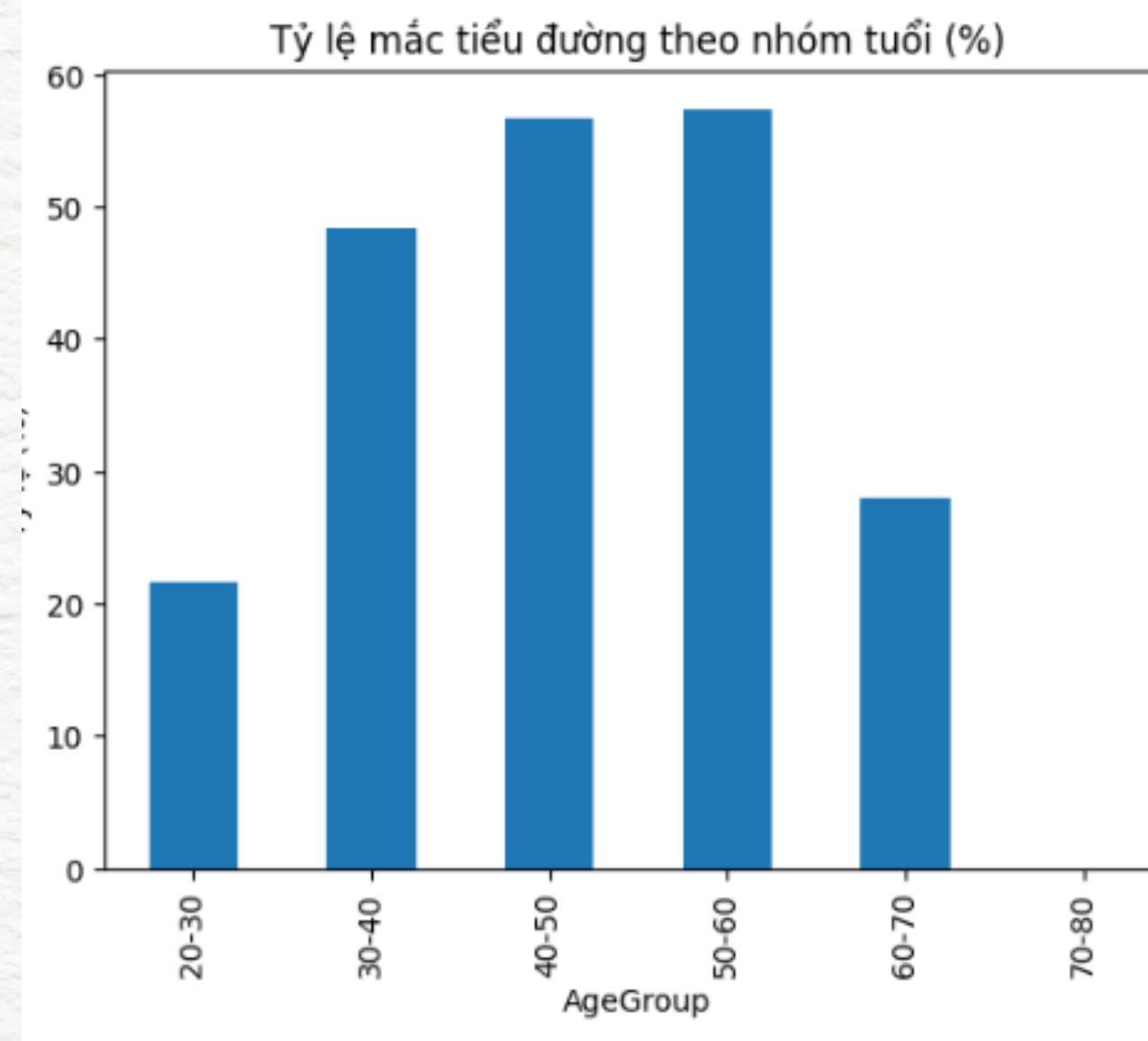
- Glucose: Tập trung chủ yếu khoảng 80 – 125 → phần lớn người khỏe mạnh có mức đường huyết bình thường.
- Age (Tuổi): Tập trung ở nhóm trẻ (20–25 tuổi), càng lớn tuổi càng ít người không mắc → nguy cơ tăng dần theo tuổi.
- BMI: Vẫn có nhiều người BMI 30–40 nhưng Outcome = 0 → ngoài béo phì còn có yếu tố khác ảnh hưởng đến nguy cơ mắc bệnh.

### Nhóm mắc tiểu đường

- Glucose: Phân bố lệch sang bên phải, nhiều người có Glucose  $\geq 100$ , thậm chí cao hơn 150–200 → đường huyết cao liên quan trực tiếp đến tiểu đường.
- Age (Tuổi): Phân bố đều ở nhiều độ tuổi.
- BMI:
  - BMI  $< 25$  (bình thường hoặc gầy): đa phần không mắc tiểu đường.
  - BMI 25–30 (thừa cân): số ca mắc bắt đầu tăng.
  - BMI 30–40 (béo phì): số ca mắc vượt trội hơn hẳn → béo phì là yếu tố nguy cơ mạnh.
- Pregnancies (Số lần mang thai): Người có từ 6–12 lần mang thai có xu hướng mắc tiểu đường cao hơn so với nhóm không mắc.

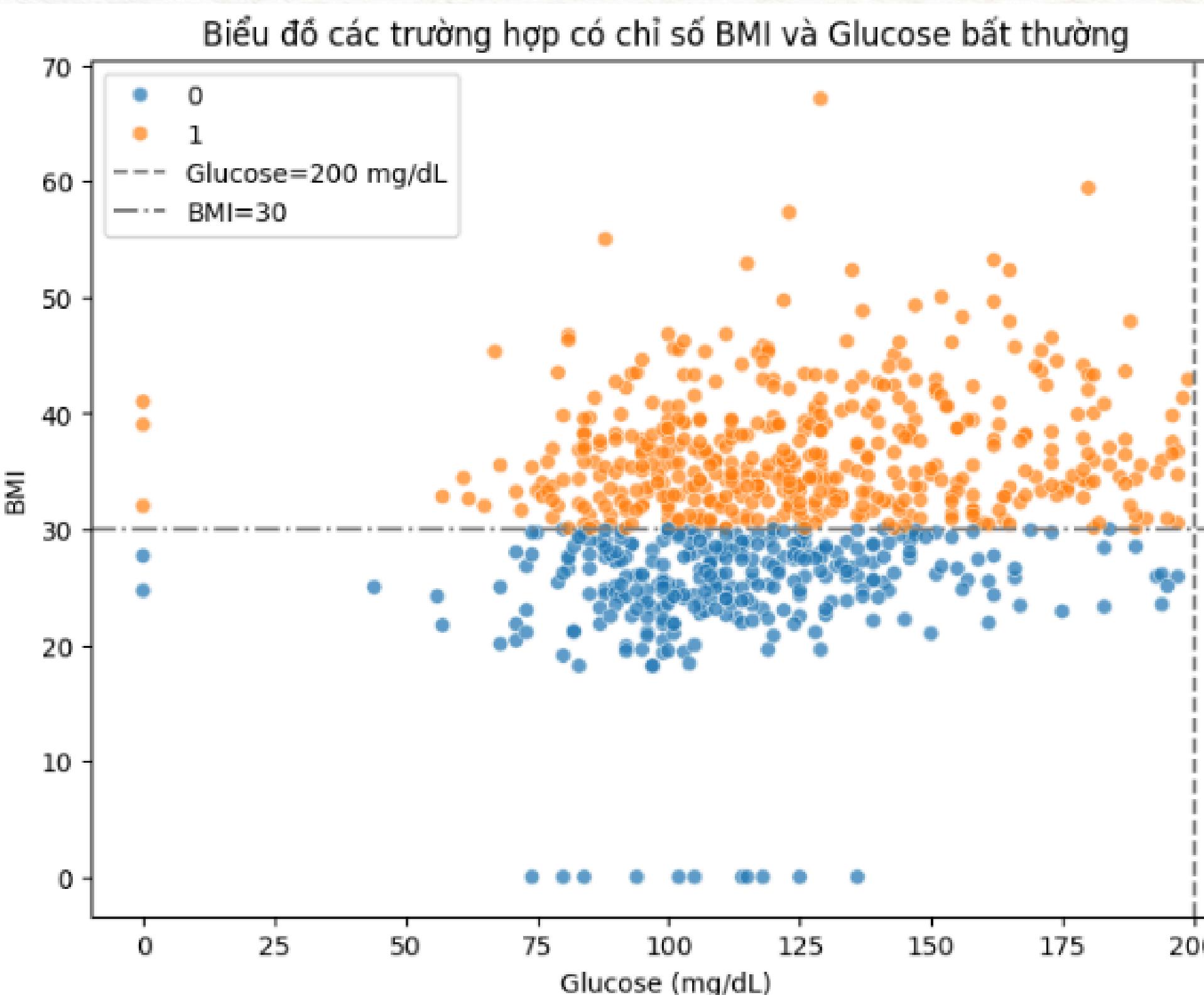


# PHÂN TÍCH Tỷ Lệ Mắc Bệnh Theo TUỔI, SỐ LẦN MANG THAI



Kết luận: Tuổi > 40 và mang thai > 6 có tỷ lệ cao hơn

# PHÂN TÍCH TRƯỜNG HỢP ĐẶC BIỆT



## ĐIỀU KIỆN XÉT

Theo bài báo tên 'PaperI' trang 33, 58

### Béo phì trung tâm:

- Nam: vòng eo / vòng hông  $> 0.90$
- Nữ: vòng eo / vòng hông  $> 0.85$
- Hoặc  $\text{BMI} > 30 \text{ kg/m}^2$

### Glucose máu:

- Nồng độ đo qua tĩnh mạch  $> 200 \text{ mg/dL}$   
→ Những yếu tố này làm tăng khả năng mắc bệnh tiểu đường.

## Nhận Xét

Mặc dù có rất nhiều phụ nữ vượt ngưỡng BMI 30, nhưng không ai có mức glucose vượt 200 mg/dL sau khi OGTT

# Bộ Dữ Liệu UCI DIABETES

## Mô Tả

- 29330 dòng
- 4 biến đầu vào

## CÁC BIẾN

- Date
- Time
- Code
- Value

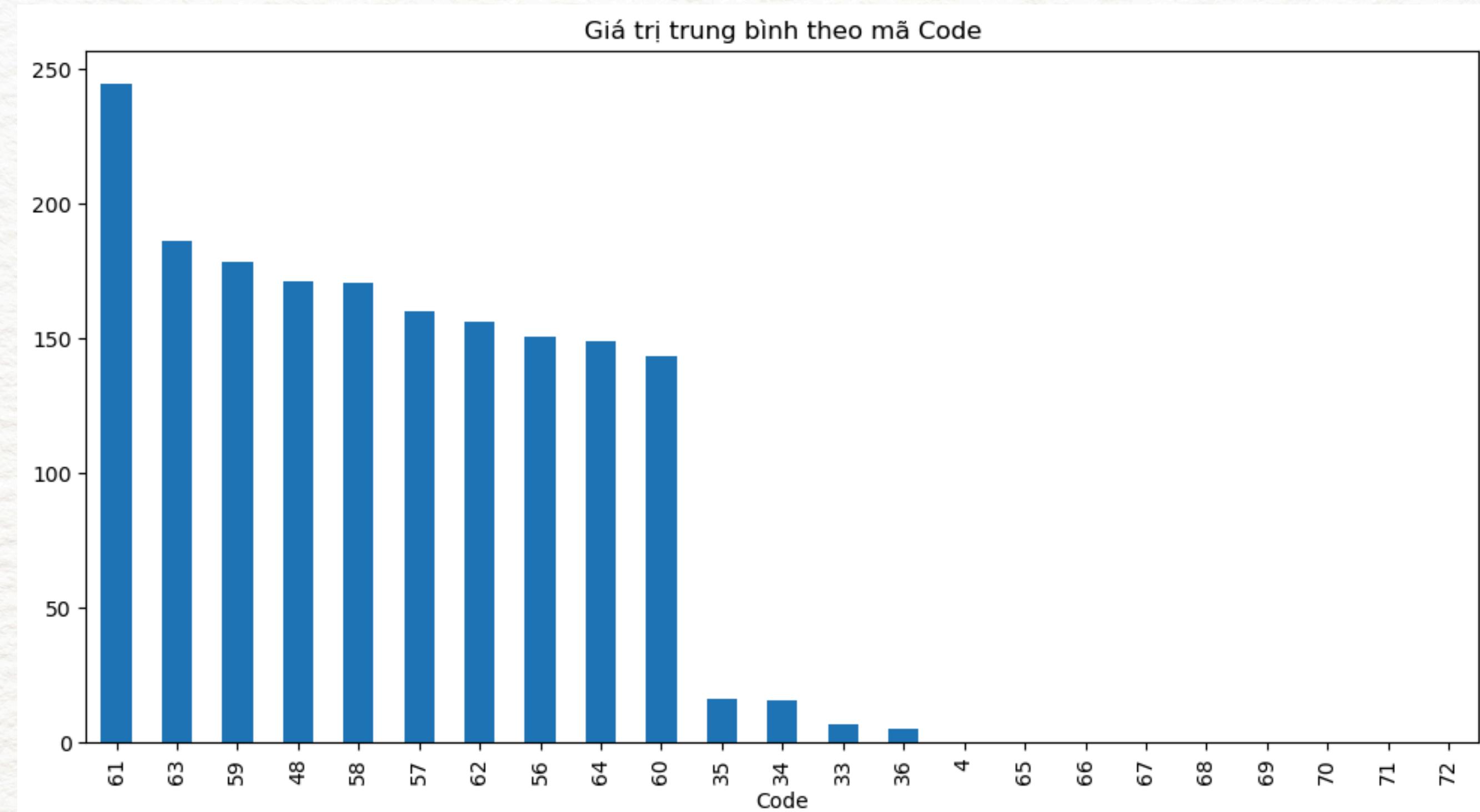
# DATA SUMMARY

- **Date:** Ngày diễn ra sự kiện y tế (định dạng MM-DD-YYYY)
- **Time:** Thời gian cụ thể của sự kiện (định dạng HH:MM)
- **Code:** Mã hóa loại sự kiện hoặc hành vi y tế (ví dụ: đo glucose, tiêm insulin, ăn uống...)
- **Value:** Giá trị đo lường tương ứng với mã Code (có thể là nồng độ glucose, liều insulin, v.v.)

# PHÂN TÍCH TỔNG QUAN: GIÁ TRỊ TRUNG BÌNH THEO CODE

## CÂU HỎI

1. Mã Code nào có giá trị trung bình cao nhất?
2. Có nhóm mã nào nổi bật về mức độ đo lường?
3. Những mã nào có giá trị thấp hoặc không có dữ liệu?



- Code 61 có giá trị trung bình cao nhất, vượt trội so với các mã còn lại → có thể phản ánh hành vi điều trị đặc biệt hoặc chỉ số sinh hóa quan trọng.
- Nhóm mã 48, 54, 57, 62, 59, 53, 58 dao động ở mức trung bình khá cao (150–190) → cho thấy mức độ hoạt động y tế đáng chú ý.
- Các mã 33–36 có giá trị rất thấp và ít biến động → phản ánh các hành vi thường xuyên, ổn định hoặc ít ảnh hưởng đến phân tích.
- Nhiều mã từ 65 trở đi không có dữ liệu → cần loại bỏ hoặc kiểm tra lại trong quá trình tiền xử lý.

# PHÂN TÍCH CHUỖI THỜI GIAN TRONG DỮ LIỆU TIỂU ĐƯỜNG

## CÂU HỎI

1. Các mã Code có xu hướng biến động như thế nào theo thời gian?
2. Có mã nào thể hiện rủi ro cao hoặc bất thường không?
3. Dữ liệu có bị gián đoạn hay thiếu hụt ở giai đoạn nào?

## CHỌN CHỈ SỐ Y TẾ

Chọn nhóm mã Code dày dữ liệu nhất:

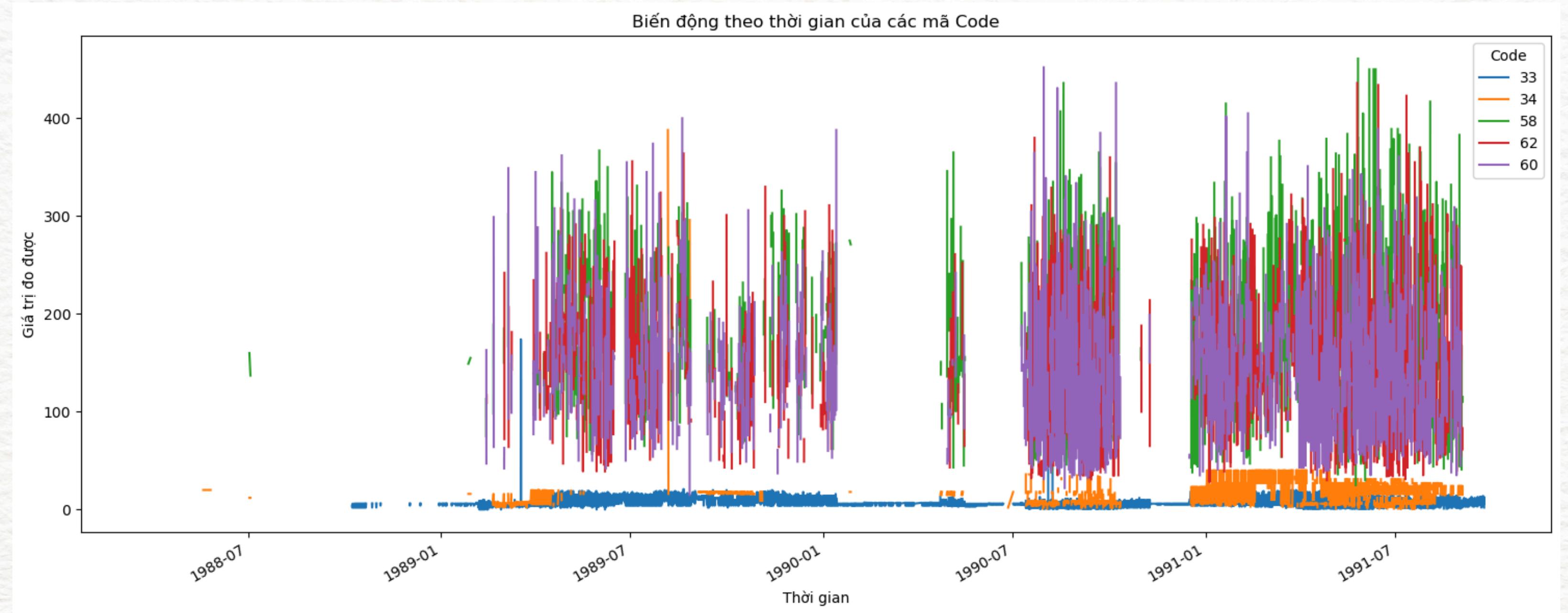
33

34

58

62

60



- Các mã 33 và 34 có giá trị thấp và ổn định theo thời gian → phản ánh các hoạt động thường xuyên như đo glucose hoặc tiêm insulin.
- Nhóm mã 58, 60, 62 có giá trị dao động mạnh, nhiều thời điểm vượt mức 400 → cho thấy nguy cơ sinh học cao hoặc hành vi điều trị không ổn định.
- Từ năm 1989 trở đi, dữ liệu có xu hướng biến động rõ rệt hơn → có thể liên quan đến thay đổi trong quy trình thu thập hoặc điều trị.
- Một số khoảng thời gian xuất hiện gián đoạn dữ liệu → cần kiểm tra lại nguồn thu thập

# PHÁT HIỆN ĐIỂM BẤT THƯỜNG QUA CHUỖI THỜI GIAN

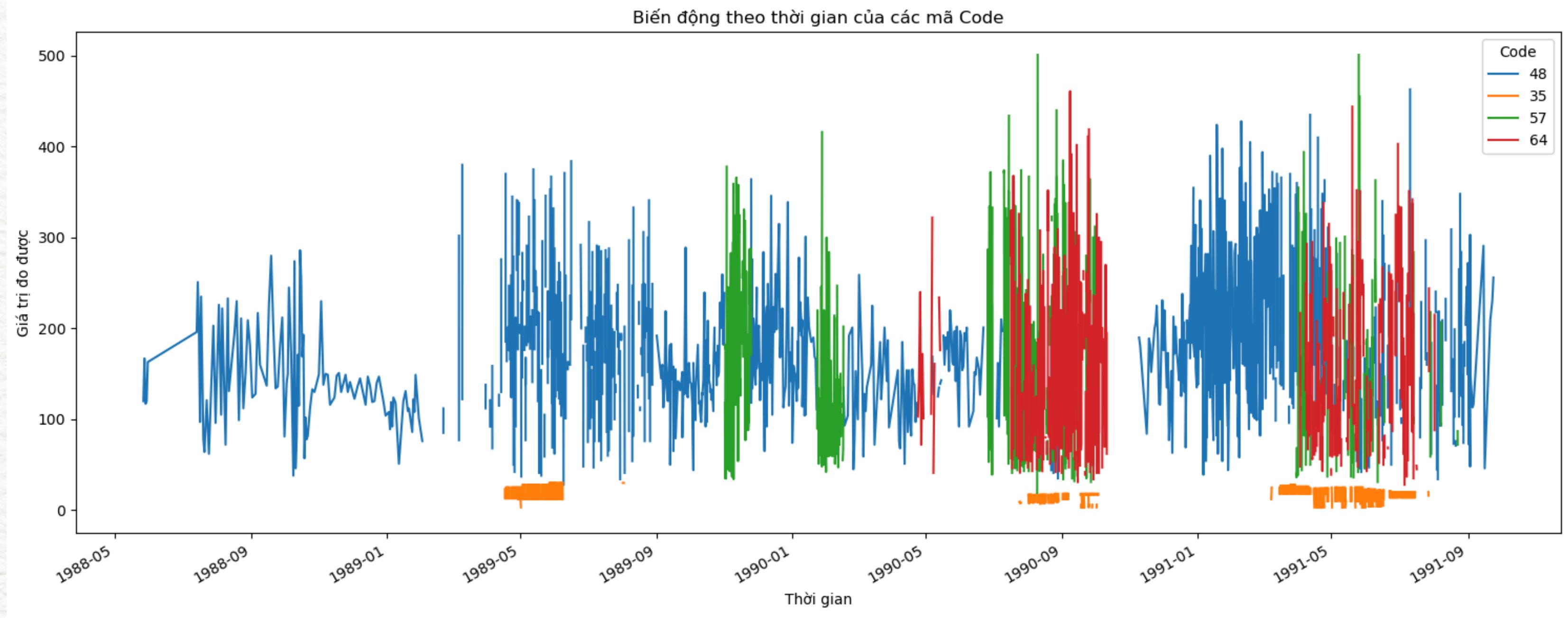
## CÂU HỎI

1. Mã Code nào có xu hướng dao động mạnh theo thời gian?
2. Có mã nào xuất hiện spikes (điểm bất thường) rõ rệt?
3. Dữ liệu có mã nào ổn định hơn so với phần còn lại?

## CHỌN CHỈ SỐ Y TẾ

Chọn nhóm mã Code dày dữ liệu vừa phải:

48  
35  
57  
64

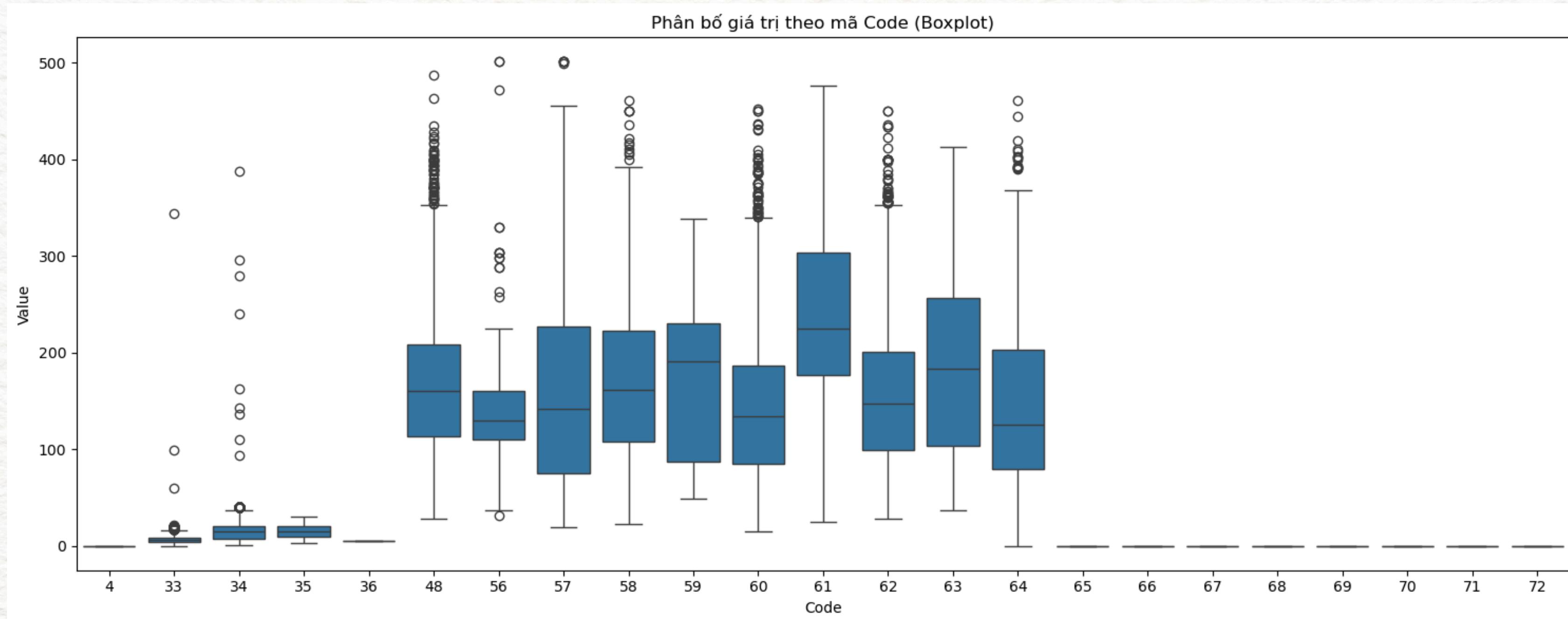


- Code 48 xuất hiện liên tục với biên độ dao động rộng → là mã nổi bật nhất trong nhóm, có thể phản ánh hành vi điều trị thường xuyên nhưng biến động.
- Code 35 dao động ở mức rất thấp, tập trung gần trực hoành → cho thấy mức độ ổn định cao hoặc ít được sử dụng.
- Code 57 và 64 có những giai đoạn bùng phát mạnh, nhiều thời điểm vượt trên 400 → phản ánh tính bất ổn cao, cần kiểm tra kỹ về ý nghĩa lâm sàng.
- Một số khoảng thời gian có dấu hiệu gián đoạn dữ liệu → cần kiểm tra lại nguồn thu thập

# PHÂN TÍCH PHÂN BỐ GIÁ TRỊ THEO MÃ CODE

## CÂU HỎI

1. Các mã Code có phân bố giá trị như thế nào?
2. PMã nào có giá trị đo lường cao nhất hoặc biến động mạnh nhất?
3. Có xuất hiện điểm ngoại lai (outlier) trong dữ liệu không?



- Các mã 33–46 có giá trị nhỏ, phân bố hẹp, ít biến động → phản ánh các hành vi thường xuyên, ổn định.
- Nhóm mã 48, 57, 58, 59, 61, 62, 63 có giá trị trung vị cao và độ trai rộng lớn → cho thấy mức độ dao động mạnh, có thể liên quan đến các chỉ số sinh hóa hoặc liều điều trị.
- Code 61 nổi bật nhất với trung vị cao và nhiều điểm ngoại lai → cần được chú ý trong phân tích chuyên sâu.
- Các mã từ 65 trở đi gần như không có dữ liệu thực tế → có thể loại bỏ trong quá trình tiền xử lý để tránh nhiễu.

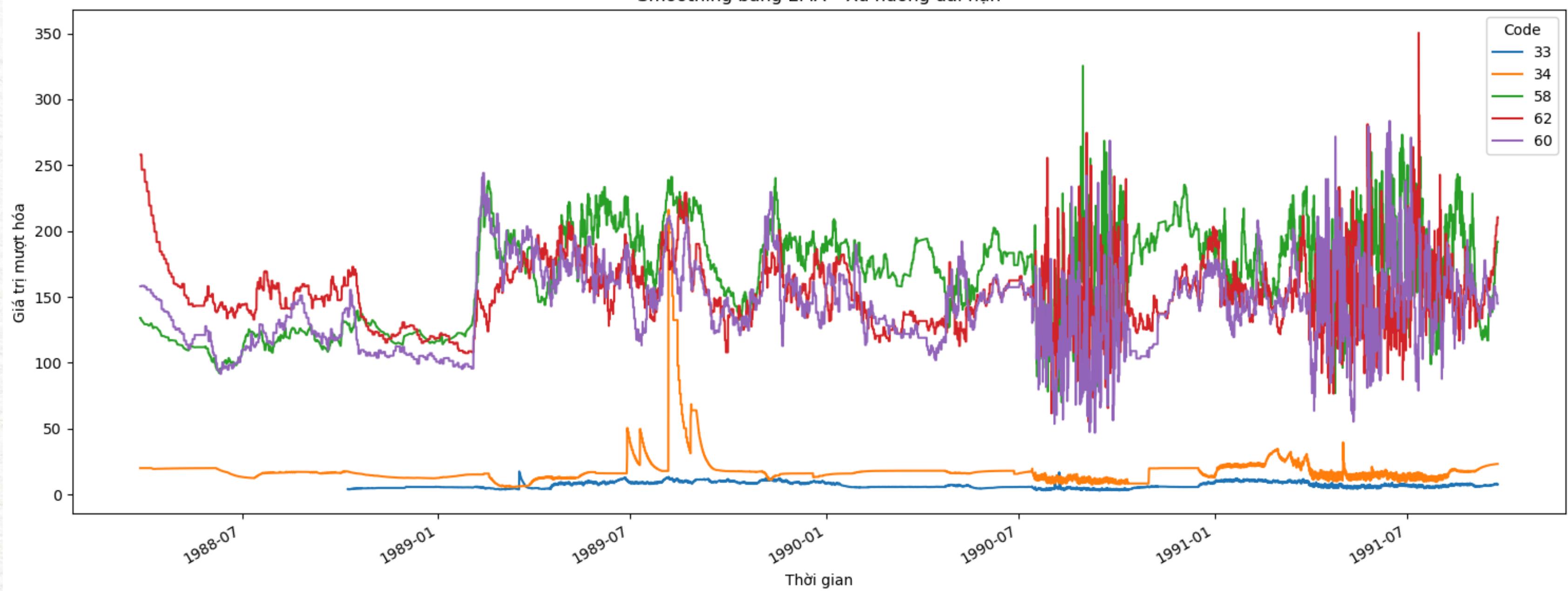
# PHÂN TÍCH XU HƯỚNG DÀI

## HẠN BẮNG EMA

CÂU HỎI

1. Các mã Code có xu hướng tăng hay giảm trong dài hạn?
2. Mã nào thể hiện sự ổn định rõ rệt?

### Smoothing bằng EMA - Xu hướng dài hạn



- Code 33 và 34 duy trì ở mức thấp, dao động nhẹ → thể hiện sự ổn định cao trong suốt thời gian quan sát.
- Code 58, 60 và 62 có giá trị cao hơn và dao động mạnh → đặc biệt sau năm 1989, xuất hiện xu hướng tăng rồi giảm rõ rệt.

# KẾT LUẬN

- Phân tích hai tập dữ liệu Pima và UCI cho thấy glucose, BMI và yếu tố nhân khẩu học là những chỉ số dự đoán quan trọng trong bệnh tiểu đường.
- Pima Dataset phù hợp cho phân tích tĩnh và xây dựng mô hình dự đoán, trong khi UCI Dataset phản ánh rõ biến động theo thời gian và hành vi điều trị.
- Đề xuất sàng lọc định kỳ cho nhóm tuổi  $>40$ , BMI  $>30$ , số lần mang thai  $>6$ , kết hợp thay đổi lối sống để giảm nguy cơ.
- Nghiên cứu cung cấp nền tảng cho việc phát triển mô hình học máy hỗ trợ chẩn đoán và phòng ngừa bệnh đái tháo đường trong tương lai.

CÀM ƠN ĐÃ

LĂNG NGHE