Nhóm 17

Danh sách thành viên:

Nguyễn Nhật Hoàng – 20520516

Code: Link

HOMEWORK 2:

Yêu cầu 1: Xác định mối quan hệ giữa số lượt thích của một bài đăng và độ tuổi của người đăng.

Answer:

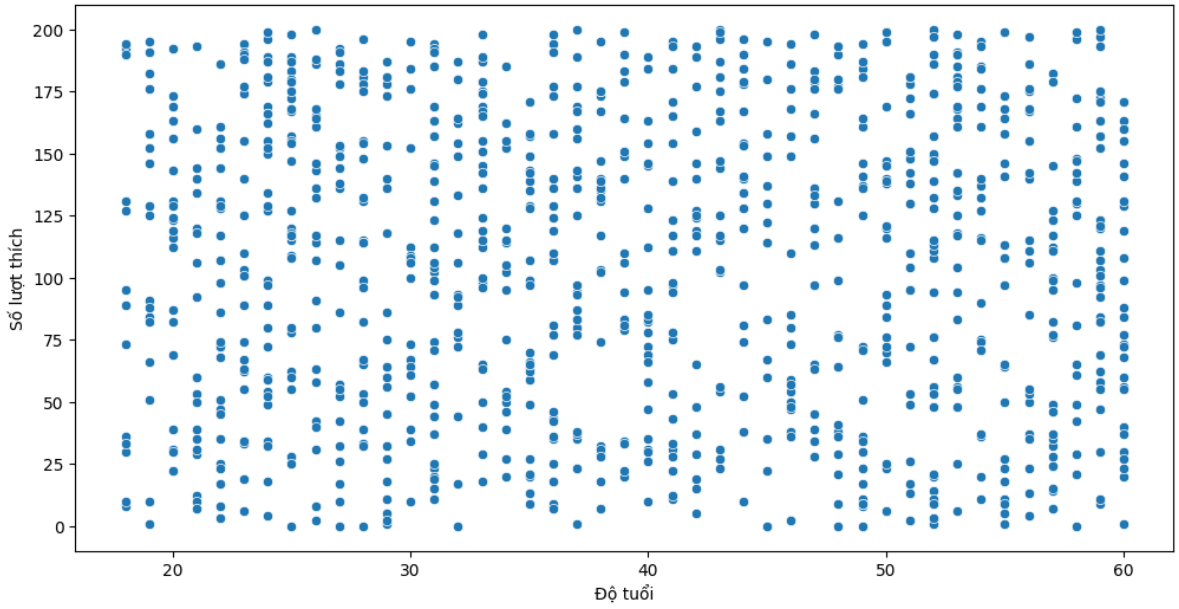
Đầu tiên, để đơn giản hóa, ta có thể xem xét mối quan hệ trực tiếp giữa số lượt thích và độ tuổi sử dụng biểu đồ cột, biểu đồ hộp hay biểu đồ phân tán.

Giải thích lý do sử dụng các biểu đồ trên:

- Biểu đồ cột: đây là loại biểu đồ phù hợp khi hai thuộc tính đang xét có 1 loại là category (Age - vì có thể chia thành các nhóm tuổi) và 1 loại là numeric (Likes).

- Biểu đồ hộp: biểu đồ dùng cho việc quan sát phân bố của các điểm dữ liệu.

- Biểu đồ phân tán: một trong những biểu đồ cơ bản để quan sát mối quan hệ của 2 thuộc tính.

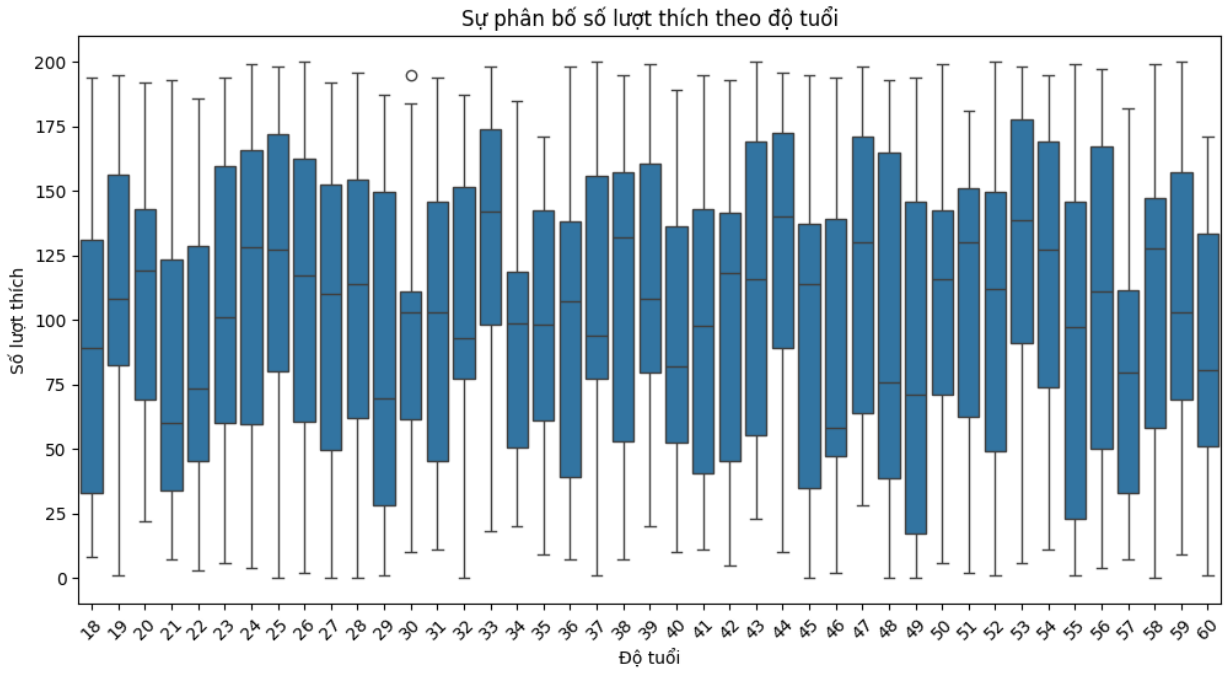


1. Biểu đồ phân tán số lượt theo độ tuổi

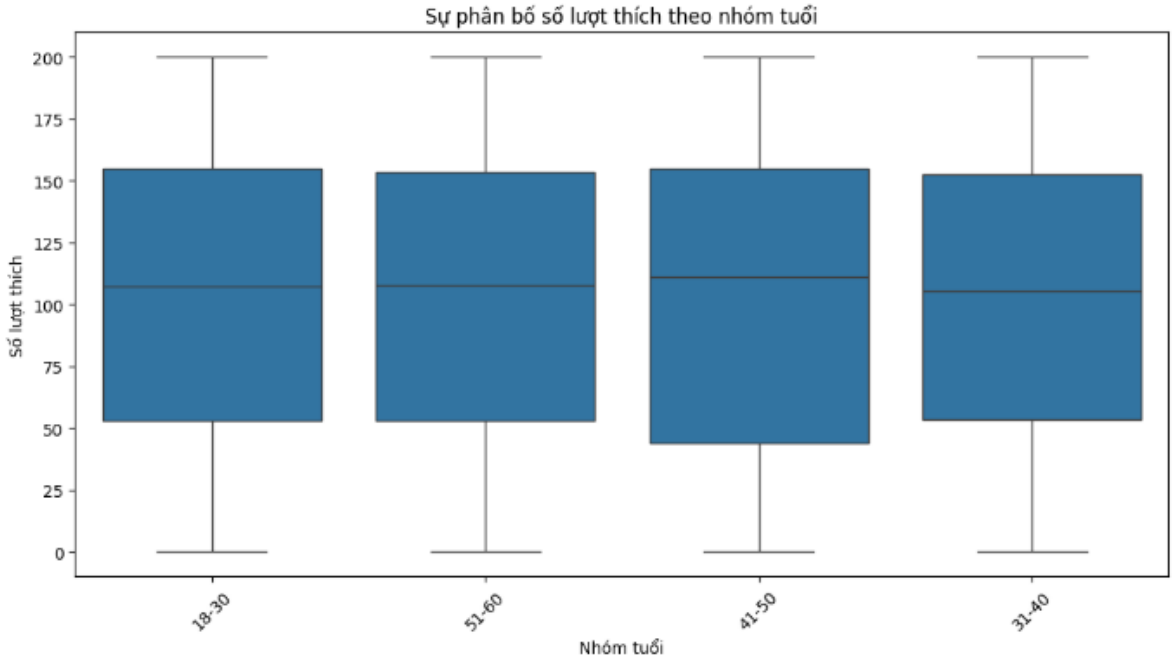
> *Insight (1):* dựa vào biểu đồ scatter thu được, dễ dàng nhận ra với bộ dữ liệu này thì giữa số lượt thích và độ tuổi không thể hiện bất kì một mối quan hệ trực tiếp nào cả vì sự phân bố của các điểm dữ liệu khá ngẫu nhiên và rời rạc.

Tuy nhiên, cần có thêm những phân tích sâu hơn trước khi kết luận về mối quan hệ giữa hai thuộc tính này.

Đầu tiên chúng ta sẽ quan sát về sự phân bổ của các giá trị lượt thích theo từng độ tuổi bằng cách sử dụng box plot.

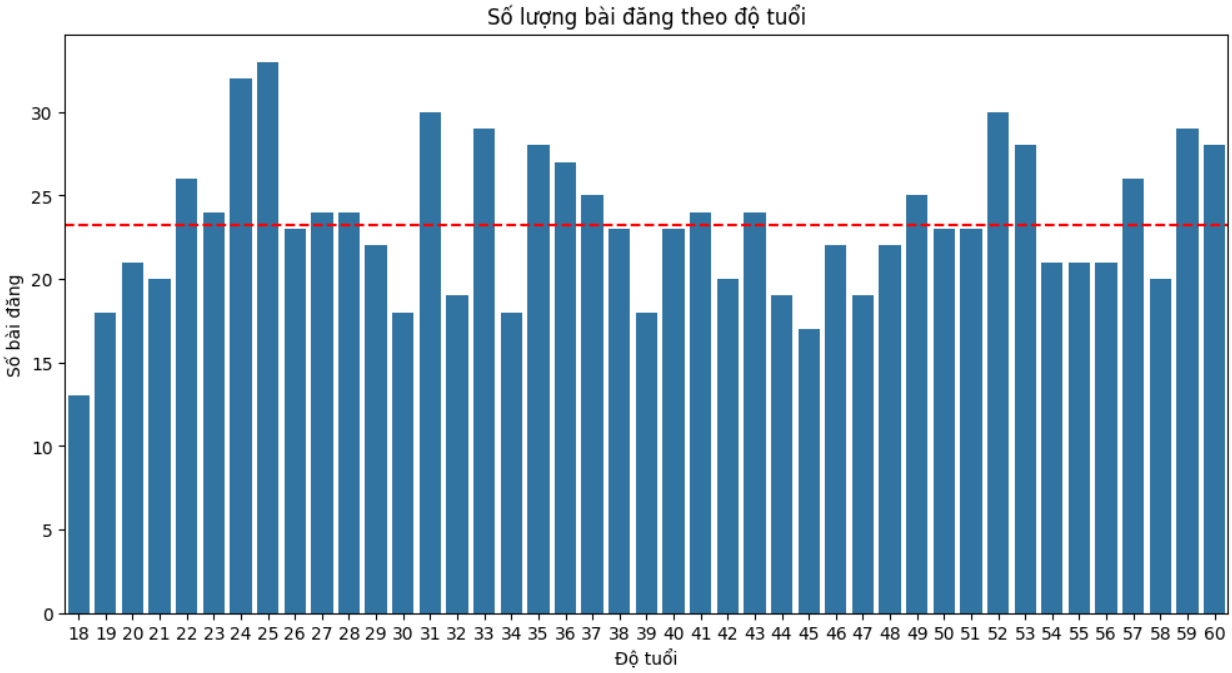


2. Phân bổ lượt thích theo độ tuổi

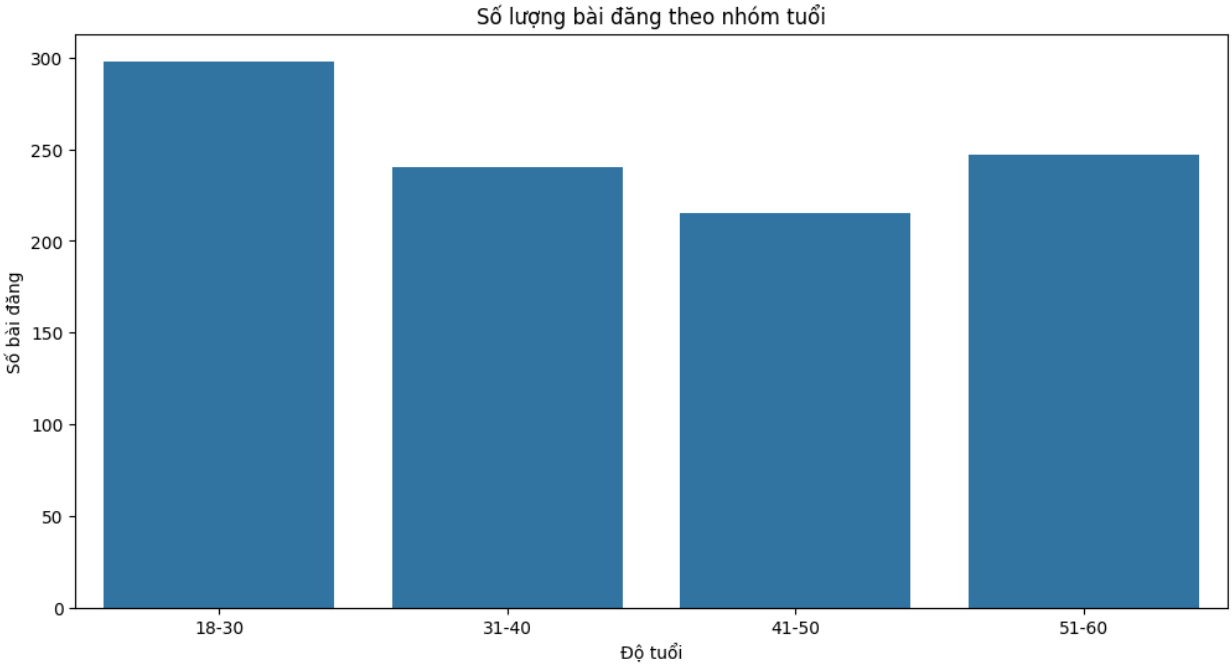


3. Phân bố số lượt thích theo nhóm tuổi

Ta có thể thấy được là biên độ số lượt thích của mỗi độ tuổi khá là rộng và gần như tương tự nhau, phần lớn tập trung ở số lượt thích từ 50 đến 150 và gần như không có giá trị ngoại lai.

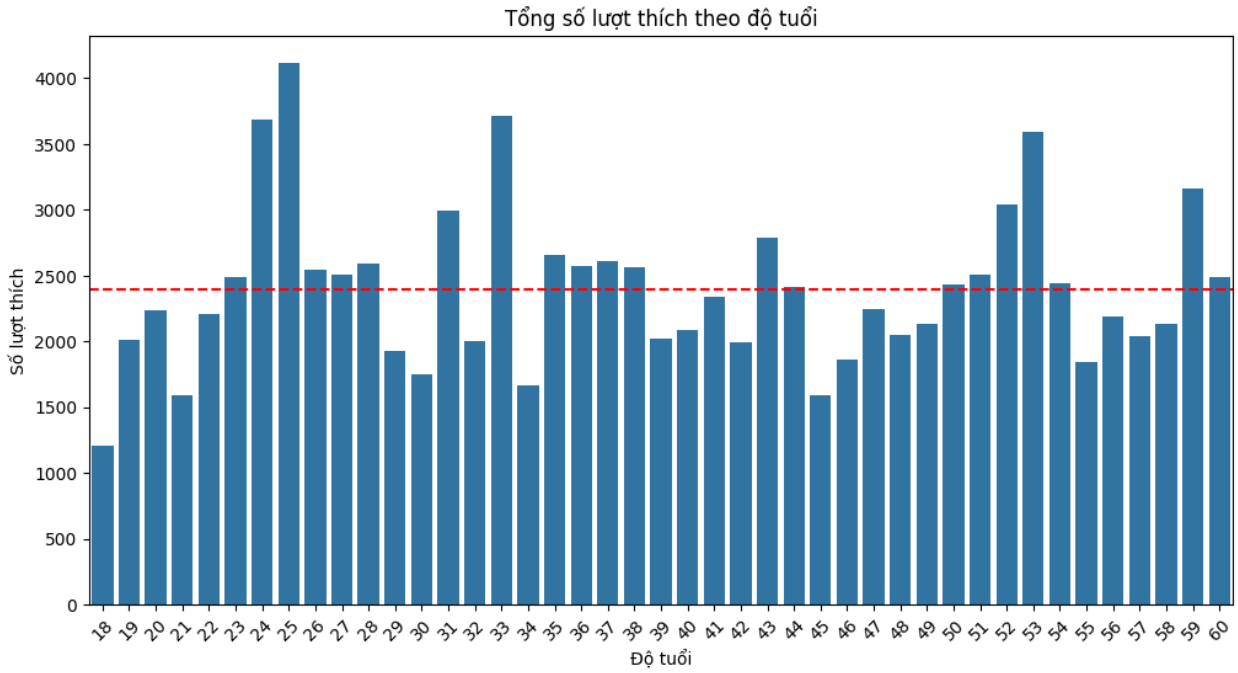


4. Số bài đăng theo độ tuổi

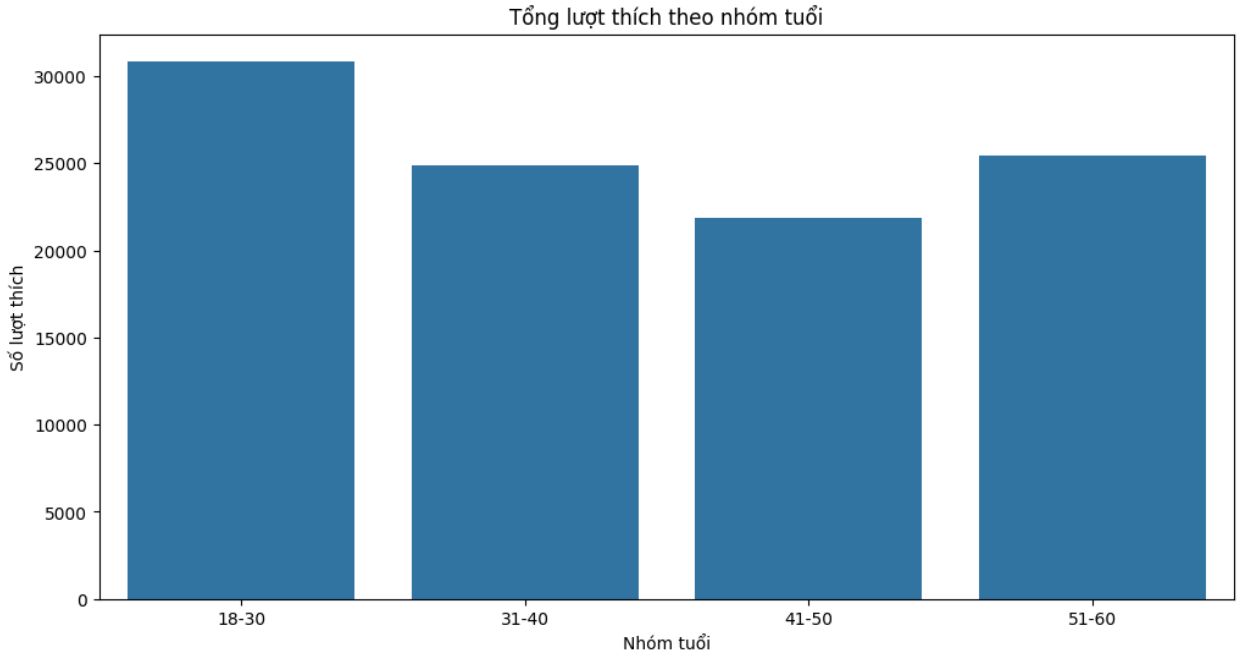


5. Số bài đăng theo nhóm tuổi

Số bài đăng ở từng độ tuổi cũng là khác nhau. Tuy nhiên có xu hướng giảm dần về nhóm tuổi từ 41-50.



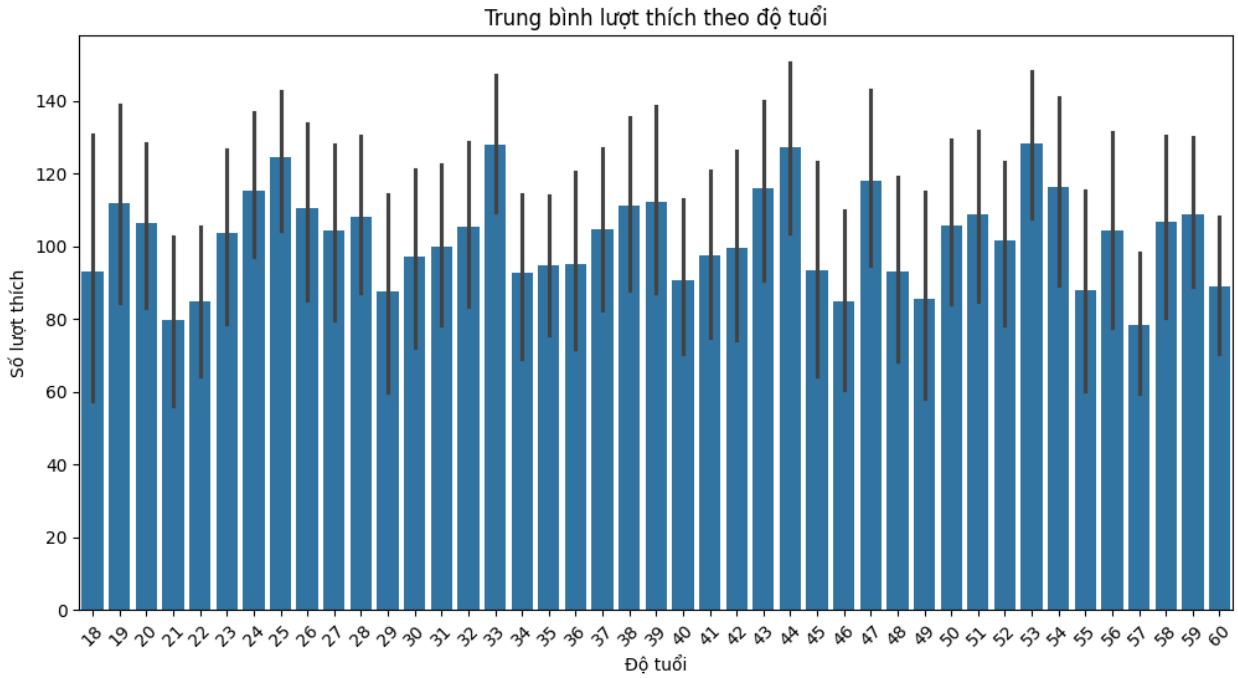
6. Tổng số lượt thích theo độ tuổi



7. Tổng số lượt thích theo nhóm tuổi

Nếu xét theo độ tuổi, ta không thể thấy được xu hướng rõ ràng của biểu đồ, nhưng có thể thấy được các xu hướng cục bộ là sự thấp hơn trung bình sau đó tăng mạnh vượt trội ở một vài cụm độ tuổi.

Biểu đồ nhóm tuổi thì thể hiện rõ nét hơn xu hướng, khi tổng số lượt thích đạt cao nhất ở độ tuổi từ trước 30, về sau thì giảm dần về mức thấp nhất ở độ tuổi từ 41-50 sau tăng lên ở độ tuổi ngoài 50.



8. Trung bình lượt thích theo độ tuổi

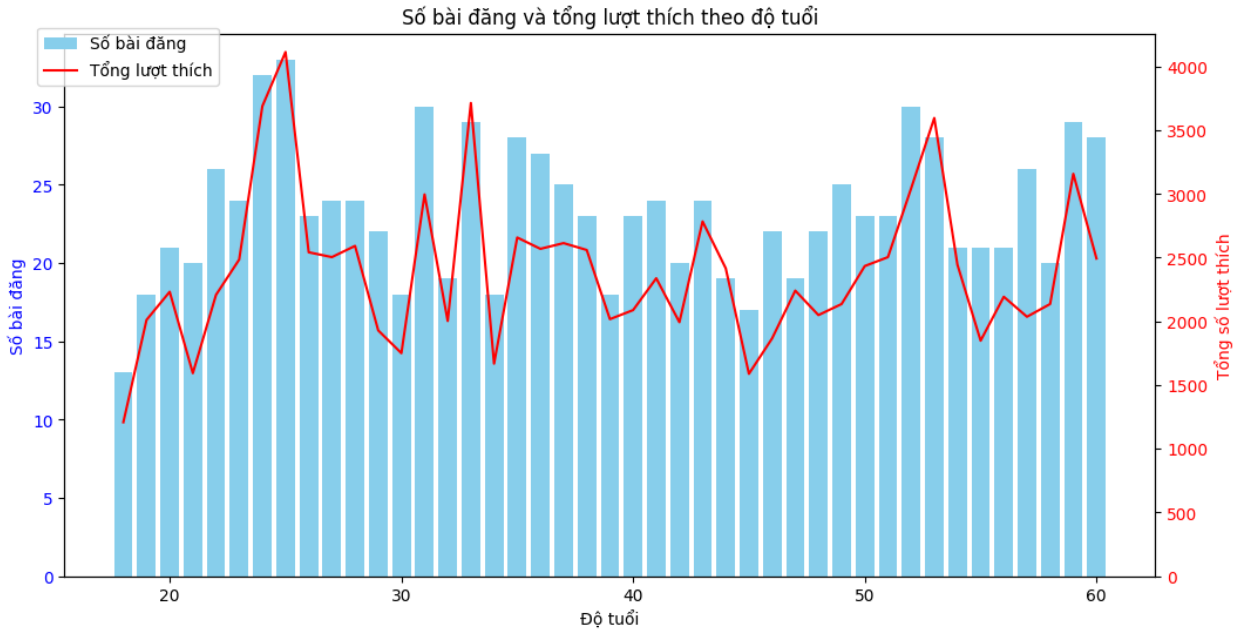
Biểu đồ về trung bình lượt thích theo độ tuổi cũng cho thấy xu hướng cục bộ đã nêu ở trên.

> *Insight (2):* tổng số lượt thích có sự khác nhau ở mỗi độ tuổi, không có xu hướng chung cho toàn bộ nhưng có thể thấy một xu hướng cục bộ đó là tổng số lượt thích sẽ thấp hơn trung bình ở một vài độ tuổi và sau đó tăng vọt ở một hoặc một vài độ tuổi.

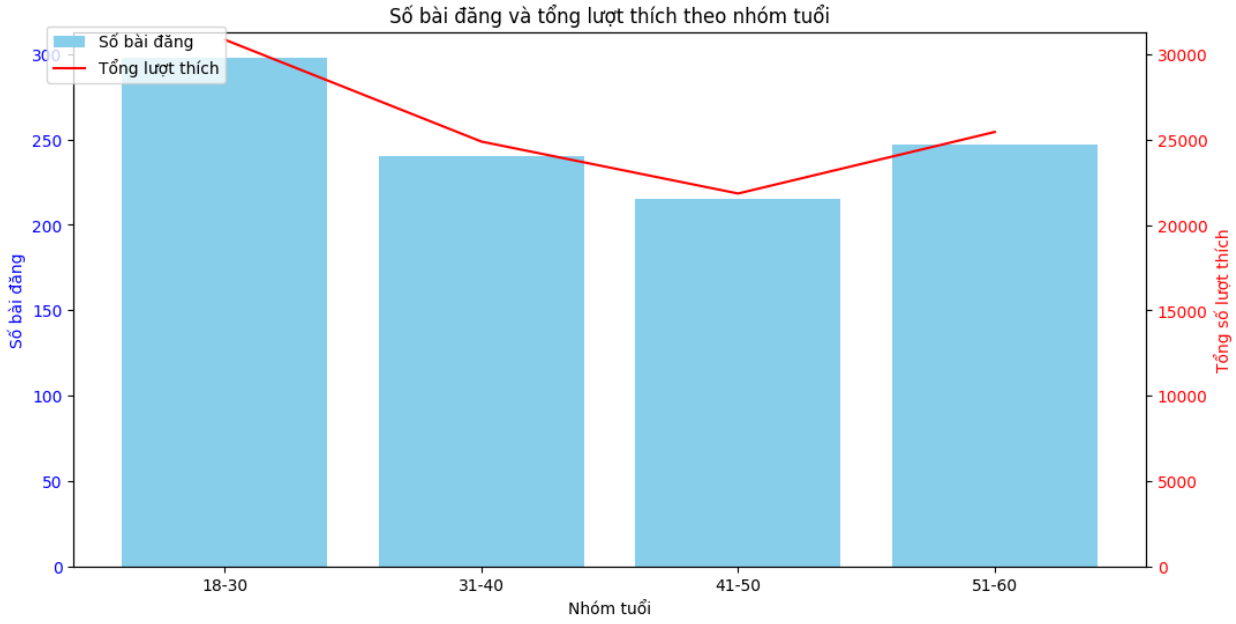
Qua hai phân tích trên, ta có thể thấy hình dạng của biểu đồ "tổng lượt thích theo độ tuổi" và "số bài đăng theo độ tuổi" có hình dạng khá tương đồng, điều này càng rõ hơn khi ta quan sát hai biểu đồ theo nhóm tuổi. Đây có thể là "manh mối" cho ta thấy được có mối liên hệ nào đó giữa tổng số lượt thích và số bài đăng theo từng độ tuổi.

Để làm rõ suy đoán trên, ta sẽ tiến hành xây dựng một biểu đồ hỗn hợp gồm 2 biểu đồ:

- Biểu đồ cột thể hiện số bài đăng
- Biểu đồ đường thể hiện sự thay đổi của số lượt thích



9. Sự thay đổi của số bài đăng và số lượt thích theo độ tuổi



10. Sự thay đổi về số bài đăng và số lượt thích theo nhóm tuổi

> *Insight (3):* ta có thể thấy được là tổng số lượt thích sẽ thay đổi theo đố lượng bài đăng ở từng độ tuổi, với những độ tuổi có nhiều bài đăng sẽ cho tổng số lượt thích cao hơn, ngược lại nếu số bài đăng ít hơn thì số lượt thích cũng ít hơn.

Kết luận: *dựa theo những gì đã phân tích ở trên, ta có thể kết luận một cách cơ bản là số lượt thích không phụ thuộc vào độ tuổi của người đăng, tuy nhiên tổng số lượt thích lại tỉ lệ thuận vào việc người dùng trong một độ tuổi đăng bài ít hay nhiều. Dữ liệu trong bộ dữ liệu khá ổn định và gần như không có giá trị ngoại lai. Và nếu xét theo từng độ tuổi thì số lượng bài đăng còn khá ít, điều này làm giảm thách thức và hạn chế cho khác phân tích chuyên sâu hơn.*

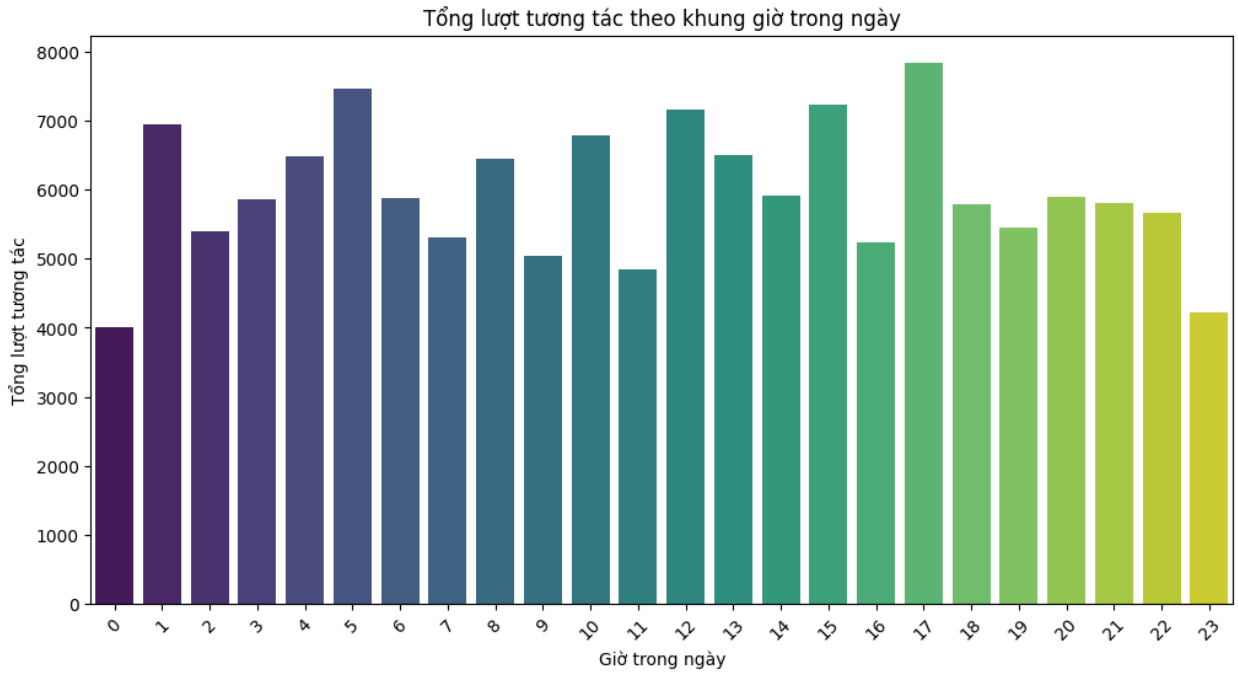
<u>Yêu cầu 2:</u> Bài đăng được đăng vào các khung giờ nào trong ngày có xu hướng nhận được nhiều/ít lượt tương tác nhất?

Answer:

Để giải quyết bài toán này, ta sẽ cần tính tổng số tương tác của mỗi bài đăng theo công thức:

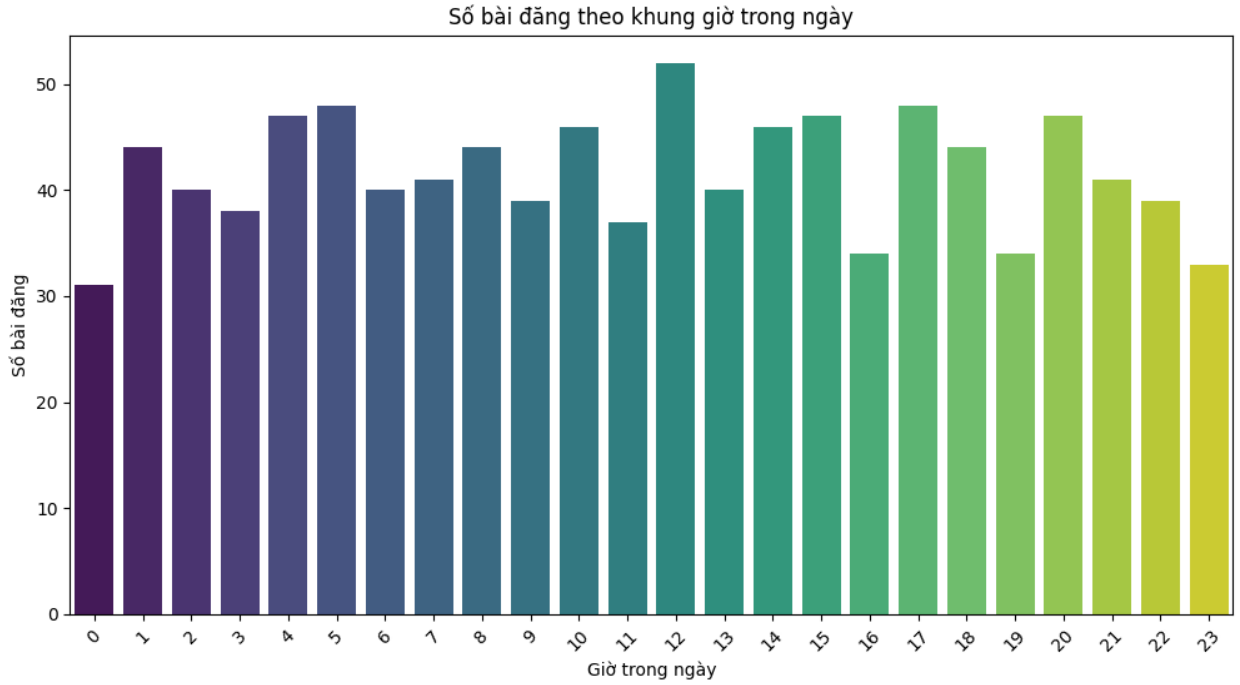
Tương tác = số likes + số shares + số comments

Sau đó tổng hợp lại theo mỗi độ tuổi, ta thu được những biểu đồ sau:



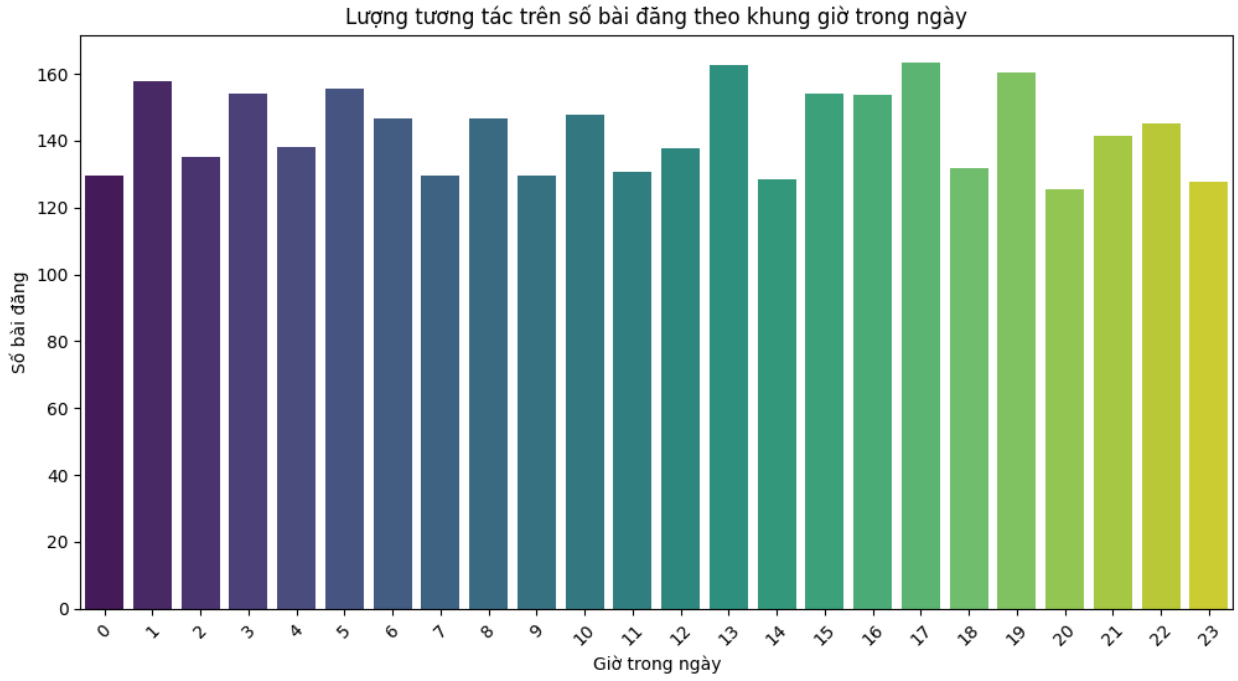
11. Lượng tương tác theo giờ trong ngày

> *Insight (4):* các khung giờ đăng bài có được nhiều lượng tương tác là 1h, 5h, 12h, 15h và cao nhất là 17h; các khung giờ đăng bài có được ít lượng tương tác ít nhất là 0h và 23h.



12. Lượng bài đăng theo giờ trong ngày

Tuy nhiên, mỗi khung giờ có số lượng bài đăng khác nhau, điều này có thể gây ra mất cân bằng dữ liệu (vd khung giờ có nhiều bài đăng hơn sẽ nhận nhiều lượt thích hơn và ngược lại) làm giảm tính khách quan khi xác định xu hướng. Để khắc phục điều này, ta cần một giá trị có tính khách quan hơn là hiệu suất tương tác trong mỗi khung giờ (số tương tác/số bài đăng).



13. Lượng tương tác mỗi bài đăng theo khung giờ trong ngày

Kết luận: *các khung giờ có tổng lượng tương tác cao thường là các khung giờ cao điểm như 1h, 5h, 10h, 12h, 15h và 17h; tổng tương tác thấp thường là các khung giờ "nghỉ" 11h, 23h và 0h. Tuy nhiên, nếu xét về hiệu suất thì các khung giờ có hiệu suất tương tác cao là 1h, 13h, 17h và 19h; hiệu suất tương tác thấp là 0h, 7h, 9h, 11h, 14, 20h và 23h.*

Mặc dù vậy, bộ dữ liệu không cho biết rõ là lượng tương tác được thu thập trong thời gian bao lâu (vd lượt thích được thu thập trong 1h sau khi đăng bài), vì thế không thể đảm bảo được sự cân bằng về thời gian tương tác của các bài đăng.