TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
KHOA KHOA HỌC CƠ BẢN

BÁO CÁO 2
HỌC PHẦN: DỮ LIỆU LỚN

ĐỀ TÀI: PHÂN TÍCH HÀNH VI KHÁCH HÀNG TRONG THƯƠNG MẠI ĐIỆN TỬ

Giảng viên phụ trách: PGS.TS Trần Văn Long

Sinh viên thực hiện: Vũ Thị Minh Ngọc

Mã sinh viên: 213012716

Lớp: Toán ứng dụng K63

Hà Nội, 2025

Trường ĐH GTVT
Khoa: KHCB
Bộ môn: Đại số và XSTK

Cộng hoà xã hội chủ nghĩa Việt Nam
Độc lập – Tự do – Hạnh phúc
*****

PHIẾU CHẤM ĐIỂM

MÔN HỌC: **DỮ LIỆU LỚN**

HỌC PHẦN: **BS1.114.3**

HÌNH THỨC: **BÁO CÁO 2**

Họ và tên: Vũ Thị Minh Ngọc

Mã sinh viên: 213012716

Lớp: Toán ứng dụng K63

Tiêu chí đánh giá	Thang điểm	Số điểm	Ghi chú
Báo cáo	2		
Mục đích dự án	1		
Mô tả dữ liệu	1		
Xử lý dữ liệu	2		
Phân tích dữ liệu	2		
Mô hình	2		
Tổng điểm	10		

Giảng viên chấm thi 1

Giảng viên chấm thi 2

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đến các thầy cô giáo của trường Đại học Giao Thông Vận Tải, đặc biệt là các thầy cô thuộc khoa Khoa học Cơ bản. Em rất biết ơn sự tận tâm trong công tác quản lý, tổ chức môn học và sự quan tâm, tạo điều kiện thuận lợi để em có đủ thời gian và môi trường học tập để hoàn thiện môn học một cách tốt nhất.

Em xin gửi lời cảm ơn đặc biệt tới thầy **Trần Văn Long**, người đã tận tình hướng dẫn và chỉ bảo em trong suốt quá trình học tập và thực hiện báo cáo môn **Dữ liệu lớn**. Những kiến thức quý báu và những lời khuyên của thầy đã giúp em rất nhiều trong việc hoàn thành môn học này. Em rất trân trọng sự hỗ trợ và sự tận tâm của thầy trong suốt thời gian qua.

Trong suốt quá trình thực hiện bài báo cáo, dù em đã cố gắng hết sức nhưng không tránh khỏi những thiếu sót và sai sót. Em rất mong thầy thông cảm và bỏ qua những lỗi nhỏ này. Với kinh nghiệm và kiến thức còn hạn chế, em hy vọng sẽ nhận được những ý kiến, đóng góp từ thầy để bài báo cáo có thể hoàn thiện hơn nữa, từ đó giúp em học hỏi và phát triển thêm trong các môn học sau.

Một lần nữa, em xin chân thành cảm ơn thầy và các thầy cô đã đồng hành, giúp đỡ em trong suốt quá trình học tập. Những đóng góp và sự quan tâm của thầy cô sẽ là nguồn động lực lớn lao để em không ngừng nỗ lực và phấn đấu.

Em xin chân thành cảm ơn!

MỤC LỤC

LỜI MỞ ĐẦU

Trong thời đại số, thương mại điện tử phát triển mạnh mẽ, trở thành kênh mua sắm chính của người tiêu dùng. Theo báo cáo **e-Conomy SEA 2024**, giá trị thị trường thương mại điện tử Việt Nam đạt khoảng **22 tỷ USD**, với hơn **70 triệu người dùng Internet thường xuyên**. Mỗi ngày, các website bán hàng trực tuyến ghi nhận hàng **triệu lượt truy cập, hàng trăm nghìn giao dịch**, tạo ra lượng lớn dữ liệu hành vi khách hàng.

Việc khai thác dữ liệu này một cách hiệu quả giúp doanh nghiệp:

- Hiểu rõ hành vi và nhu cầu của khách hàng.
- Phân loại khách hàng theo mức độ tương tác và tiềm năng chi tiêu.
- Triển khai các chiến lược marketing cá nhân hóa, tối ưu hóa doanh thu.

Tuy nhiên, lượng dữ liệu lớn, đa dạng và liên tục thay đổi đòi hỏi doanh nghiệp áp dụng các phương pháp **khoa học dữ liệu và Big Data** để khai thác hiệu quả thông tin. Việc này bao gồm phân tích hành vi truy cập, đánh giá mức độ tương tác, dự đoán khả năng mua hàng, từ đó đề xuất các chiến lược phù hợp cho từng nhóm khách hàng.

Mục tiêu của đề tài:

- Trình bày quá trình phân tích hành vi khách hàng trong thương mại điện tử.
- Phân cụm khách hàng và dự đoán khả năng mua hàng.
- Đề xuất chiến lược marketing phù hợp dựa trên kết quả phân tích, nhằm nâng cao hiệu quả kinh doanh.

CHƯƠNG 1: DỰ ÁN KHOA HỌC DỮ LIỆU

1.1. Bối cảnh nghiên cứu.

- Khách hàng trên các nền tảng thương mại điện tử có hành vi rất đa dạng: một số thường xuyên truy cập nhưng ít mua, một số trung thành và chi tiêu nhiều, một số mới lần đầu truy cập. Lượng dữ liệu hành vi này ngày càng lớn và phức tạp, với hàng **triệu lượt truy cập và hàng trăm nghìn giao dịch mỗi ngày**, do đó việc **hiểu rõ hành vi khách hàng** trở nên cần thiết để doanh nghiệp đưa ra quyết định kinh doanh chính xác và kịp thời.
- Phân tích hành vi khách hàng giúp:
    - Nhận diện nhóm khách hàng tiềm năng.
    - Cá nhân hóa trải nghiệm người dùng.
    - Triển khai các chiến lược marketing hiệu quả, tối ưu nguồn lực.

1.2. Đề tài nghiên cứu.

- Tên đề tài: “ Phân tích hành vi khách hàng trong thương mại điện tử”.
- Các vấn đề đặt ra trong dự án:
    - Khai phá dữ liệu lớn (Big data): Xử lý lượng dữ liệu khổng lồ từ BigQuery để phân tích hành vi và dự đoán nhu cầu khách hàng.
    - Phân loại khách hàng theo hành vi: Khách hàng có hành vi truy cập và mua hàng khác nhau, cần xác định các nhóm khách hàng với đặc điểm tương đồng để dễ triển khai chiến lược.
    - Dự đoán khả năng mua hàng của khách mới: Giúp doanh nghiệp tập trung chăm sóc khách hàng tiềm năng và tối ưu marketing.
- Những vấn đề này không chỉ phục vụ mục tiêu kinh doanh mà còn là cơ sở khoa học cho việc xây dựng **hệ thống phân tích hành vi khách hàng tự động và hiệu quả**.

1.3. Mục tiêu dự án

- Mục tiêu của dự án được chia thành:
    - **Mục tiêu tổng quát:** Phân tích hành vi khách hàng trong thương mại điện tử để tối ưu chiến lược marketing dựa trên dữ liệu lớn.
    - **Mục tiêu cụ thể:**
        1. Phân loại khách hàng theo hành vi (phân cụm).
        2. Dự đoán khả năng mua hàng của khách hàng mới.
        3. Đề xuất các chiến lược marketing phù hợp cho từng nhóm khách hàng, từ remarketing, ưu đãi cá nhân hóa đến loyalty program.

CHƯƠNG 2: MÔ TẢ DỮ LIỆU

2.1. Giới thiệu về tập dữ liệu.

- Tập dữ liệu được sử dụng trong đề tài lấy từ **Google Analytics Sample** – một tập dữ liệu công khai trên nền tảng **Google BigQuery.**
- Đây là dữ liệu thực tế về hành vi truy cập của người dùng trên một website thương mại điện tử, bao gồm thông tin về lượt truy cập, nguồn truy cập, thiết bị, vị trí địa lý và thời gian truy cập.

2.2. Cấu trúc dữ liệu.

- Tập dữ liệu bao gồm **368 bảng dữ liệu**, trong đó nổi bật là các bảng tên dạng *"ga_sessions_YYYYMMDD"* - mỗi bảng tương ứng với dữ liệu của một ngày, thể hiện các phiên truy cập của người dùng vào trang wed. Có 366 bảng dữ liệu dạng này, chúng cùng cấu trúc và kiểu dữ liệu, chỉ khác nhau về thời gian ghi nhận, do đó có thể xem đây là các phân vùng dữ liệu theo ngày.
- Ngoài ra , tập dữ liệu còn chứa 2 bảng tổng hợp sau:
  - *"daily_total_visits":* thống kê lượt truy cập theo ngày
  - *"daily_visits":* thống kê cho tiết hơn về lượt truy cập từng ngày.

2.3. Bảng dữ liệu được sử dụng.

2.3.1. Bảng "daily_total_visits".

- Bảng dữ liệu này cung cấp số lượng tổng lượt truy cập website theo từng ngày. Mỗi dòng tương ứng với một ngày cụ thể trong khoảng thời gian từ 2016 đến 2017. Đây là bảng dữ liệu tổng hợp mức cao nhất trong bộ dữ liệu.
- Bảng gồm 2 trường dữ liệu và 366 hàng dữ liệu:

Tên trường dữ liệu	Kiểu dữ liệu	Ý nghĩa	Ghi chú
Visit_date	DATE	Ngày ghi nhận lượt truy cập	Cột định danh theo thời gian
Total_visits	INTEGER	Tổng số lượt truy cập trong ngày	Không có giá trị NULL

- Đặc điểm dữ liệu:
  - Mỗi dòng dữ liệu tương ứng với một ngày.
  - Không có giá trị bị thiếu.
  - Có thể dùng để quan sát xu hướng thay đổi lượt truy cập theo thời gian (Ví dụ: theo tháng, quý, năm).

2.3.2. Bảng "daily_visits".

- Bảng này cũng thể hiện thông tin về lượt truy cập theo ngày, tương tự bảng "*daily_total_visits*", nhưng ở mức độ chi tiết hơn – có thể được sử dụng để kiểm tra, đối chiếu hoặc làm cơ sở cho các phân tích khác về tần suất truy cập.
- Bảng gồm 2 trường dữ liệu và 367 dòng dữ liệu:

Tên trường dữ liệu	Kiểu dữ liệu	Ý nghĩa	Ghi chú
Visit_date	DATE	Ngày ghi nhận lượt truy cập	Trùng với ngày trong bảng "daily_total_visits"
Total_visits	INTEGER	Số lượt truy cập trong ngày	Dữ liệu tương đồng, có thể dùng để đối chiếu

- Đặc điểm dữ liệu:
  - Không có giá trị bị thiếu.
  - Một số giá trị có thể chênh lệch nhỏ do cách tổng hợp dữ liệu khác nhau.

2.3.3. Bảng "ga_sessions_YYYYMMDD"

- Có 366 bảng cùng cấu trúc và kiểu dữ liệu, chỉ khác ghi nhận. Nên chọn bảng "*ga_sessions_20160801*" (ngày 01/08/2016) đại diện để tiến hành mô tả và phân tích.
- Bảng "*ga_sessions_20160801*" chứa dữ liệu chi tiết về từng phiên truy cập, với 23 trường dữ liệu và 2.556 dòng dữ liệu:

Tên trường dữ liệu	Kiểu dữ liệu	Ý nghĩa	Ghi chú
fullVisitorId	STRING	Mã định danh duy nhất của người dùng	Dùng để theo dõi và nhóm các phiên truy cập thuộc cùng một người dùng
visitId	INTEGER	Mã định danh duy nhất cho từng phiên truy cập	Mỗi người dùng có thể có nhiều ***visitId*** khác nhau.
visitNumber	INTEGER	Số thứ tự của phiên truy cập của người dùng	Dùng để phân biệt phiên đầu tiên, thứ hai,… của cùng một user
Date	STRING / DATE	Ngày diễn ra phiên truy cập (định dạng yyyymmdd)	Có thể chuyển sang định dạng ngày chuẩn để phân tích theo thời gian
visitStartTime	INTEGER	Thời điểm bắt đầu phiên truy cập (unix timestamp)	Có thể chuyển sang định dạng thời gian thực để trực quan hóa.
channelGrouping	STRING	Nhóm kênh tiếp thị chính	Ví dụ: Organic Search, Direct, Referral, Social, Paid Search.
socialEngagementType	STRING	Loại tương tác xã hội của phiên truy cập.	Giá trị thường gặp: "Not Socially Engaged" hoặc "Socially Engaged".
Device.browser	STRING	Trình duyệt người dùng sử dụng khi truy cập.	Ví dụ: Chrome, Safari, Firefox.
device.operatingSystem	STRING	Hệ điều hành của thiết bị.	Ví dụ: Windows, Android, iOS, macOS.
device.deviceCategory	STRING	Loại thiết bị người dùng.	Gồm mobile, desktop, hoặc tablet.
geoNetwork.continent	STRING	Châu lục của người dùng.	Ví dụ: Asia, Europe, Americas.
geoNetwork.country	STRING	Quốc gia của người dùng.	Dữ liệu được ẩn danh, nên có thể có giá trị "(not set)".
geoNetwork.city	STRING	Thành phố người dùng truy cập.	Nhiều giá trị bị thiếu hoặc "not available in demo dataset".

trafficSource.source	STRING	Nguồn truy cập cụ thể.	Ví dụ: google, bing, youtube, (direct).
trafficSource.medium	STRING	Loại nguồn truy cập.	Ví dụ: organic (tự nhiên), referral (giới thiệu), cpc (trả phí).
totals.visits	INTEGER	Số lượt truy cập được ghi nhận.	Thường là 1 cho mỗi dòng (tương ứng 1 session).
totals.hits	INTEGER	Số lượt tương tác (hits) trong một phiên.	Bao gồm pageviews, events, transactions, v.v.
totals.pageviews	INTEGER	Số trang người dùng đã xem trong phiên.	Dùng để đánh giá mức độ tương tác.
totals.timeOnSite	FLOAT / INTEGER	Tổng thời gian người dùng ở lại site trong phiên (giây).	Dùng để tính thời gian trung bình trên site.
totals.bounces	INTEGER	Phiên thoát – người dùng chỉ xem một trang rồi rời đi.	1 = thoát, 0 = không thoát.
totals.transactions	INTEGER	Số giao dịch (mua hàng) trong phiên.	Có thể bằng 0 nếu người dùng chỉ xem không mua.
totals.transactionRevenue	FLOAT / INTEGER	Doanh thu phát sinh trong phiên (micro units).	Cần chia cho 1,000,000 để đổi sang đơn vị tiền tệ.
totals.newVisits	INTEGER	Đánh dấu người dùng mới (1 nếu là lần đầu truy cập).	Giúp xác định tỷ lệ người dùng mới vs quay lại.

- Đặc điểm dữ liệu:
  - Một số cột có tỷ lệ giá trị NULL cao, ví dụ như campaign, adContent, socialEngagementType, do không áp dụng cho tất cả phiên truy cập.
  - Dữ liệu có thể khai thác được nhiều góc độ: hành vi, vị trí, thiết bị, nguồn traffic, thời lượng truy cập,…

2.4. Tổng hợp so sánh.

Bảng dữ liệu	Mức độ chi tiết	Dữ liệu đại diện cho	Ứng dụng chính
Daily_total_visits	Tổng hợp	Xu hướng truy cập theo thời gian	Vẽ biểu đồ xu hướng, phân tích biến động truy cập
Daily_visits	Tổng hợp	Kiểm tra tần suất truy cập	Đối chiếu hoặc xác minh dữ liệu
Ga_sessions_yyyymmdd	Chi tiết	Phiên truy cập từng người dùng	Phân tích hành vi người dùng, nguồn truy cập, thiết bị, vị trí

➢ Bộ dữ liệu có cấu trúc phong , cung cấp đầy đủ thông tin cho việc phân tích hành vi người dùng trên website thương mại điện tử. Tuy nhiên, do dữ liệu có nhiều trường lồng nhau và chứa giá trị trống, cần thực hiện bước tiền xử lý và làm sạch dữ liệu trước khi tiến hành phân tích sâu hơn.

CHƯƠNG 3: KHAI PHÁ DỮ LIỆU

3.1. Mục tiêu khai phá.

Mục tiêu của chương này là khám phá và phân tích sơ bộ tập dữ liệu Google Analytics Sample nhằm hiểu rõ đặc điểm, chất lượng và mối quan hệ giữa các biến.
Các nội dung thực hiện bao gồm:

- Kiểm tra dữ liệu thiếu và giá trị không hợp lệ.
- Phân tích phân phối dữ liệu.
- Đánh giá mối tương quan giữa các biến.
- Rút ra nhận xét ban đầu cho các bước xử lý tiếp theo.

3.2. Khai phá bảng dữ liệu.

3.2.1. Bảng "daily_total_visits".

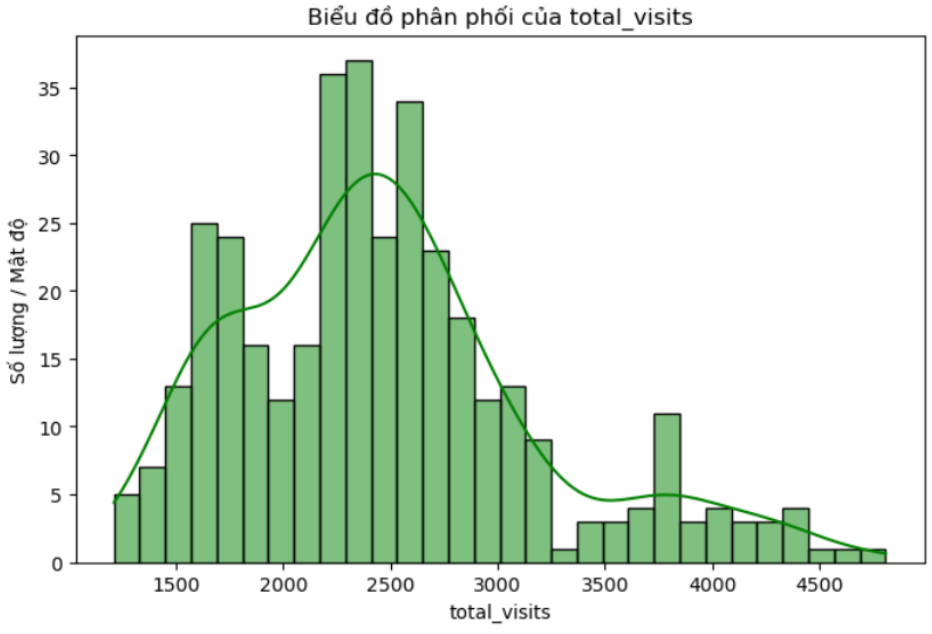
3.2.1.1. Kiểm tra dữ liệu.

- Mỗi dòng dữ liệu tương ứng với một ngày, bao gồm các cột chính:
  - **visit_date**: ngày truy cập (kiểu DATE)
  - **total_visits**: tổng số lượt truy cập trong ngày (kiểu INT64)
- Dữ liệu có 366 bản ghi, dữ liệu không có giá trị bị thiếu.
- Thống kê mô tả:

Thông số	Gía trị	Ý nghĩa
Trung bình lượt truy cập mỗi ngày: mean()	2468.997268	Mức độ truy cập trung bình
Trung vị: median()	2391	Mức truy cập điển hình
Độ lệch chuẩn: std()	707.720622	Mức độ biến động của lượt truy cập
Gía trị nhỏ nhất: min()	1211	Biên độ dao động
Gía trị lớn nhất: max()	4807	Biên độ dao động

3.2.1.2. Phân phối dữ liệu.

- Biểu đồ phân phối số lượng / mật độ của bảng dữ liệu:



➢ Biểu đồ trên minh họa **phân phối tổng số lượt truy cập hằng ngày (total_visits)** trong dữ liệu Google Analytics.
- Trục hoành (x) biểu diễn **tổng lượt truy cập mỗi ngày**.
- Trục tung (y) thể hiện **số lượng ngày** (cột histogram) và **mật độ phân phối** (đường cong KDE).
- Biểu đồ có dạng **lệch phải (right-skewed)**: phần lớn các ngày có lượng truy cập trong khoảng **1.500–2.800 lượt/ngày**, trong khi chỉ có một số ít ngày có lượt truy cập cao vượt mức **3.500**.
- Đường KDE cho thấy dữ liệu tập trung mạnh ở vùng giá trị trung bình, sau đó giảm dần về phía phải, biểu hiện sự **không đồng đều** trong hoạt động truy cập — có thể do các chiến dịch quảng bá hoặc sự kiện đặc biệt làm tăng đột biến lượt truy cập.

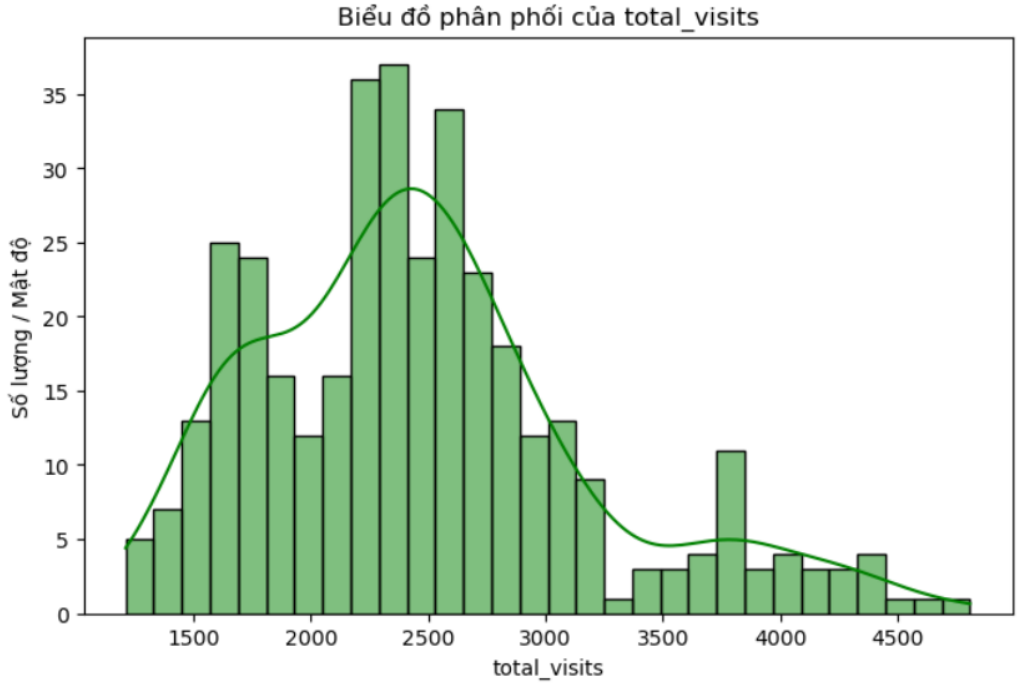
3.2.2. Bảng dữ liệu "total_visits".

3.2.2.1. Kiểm tra dữ liệu.

- Mỗi dòng dữ liệu tương ứng với một ngày, bao gồm các cột chính:
  - **visit_date**: ngày truy cập (kiểu DATE)
  - **total_visits**: số lượt truy cập trong ngày (kiểu INT64)
- Dữ liệu có 367 bản ghi, dữ liệu không có giá trị bị thiếu.
- Thống kê mô tả:

Thông số	Gía trị	Ý nghĩa
Trung bình lượt truy cập mỗi ngày: mean()	2462.269755	Mức độ truy cập trung bình
Trung vị: median()	2396	Mức truy cập điển hình
Độ lệch chuẩn: std()	712.279712	Mức độ biến động của lượt truy cập
Gía trị nhỏ nhất: min()	638	Biên độ dao động
Gía trị lớn nhất: max()	4698	Biên độ dao động

3.2.2.2. Phân phối dữ liệu.



➢ Biểu đồ trên minh họa **phân phối số lượt truy cập theo ngày (total_visits)** trong dữ liệu Google Analytics.
- Trục hoành (x) biểu diễn **số lượt truy cập mỗi ngày**.
- Trục tung (y) biểu diễn **tần suất xuất hiện (histogram)** và **mật độ xác suất ước lượng (đường KDE)**.
- Phân phối **không đối xứng**, có **độ lệch phải** (right-skewed), tức là phần lớn các ngày có lượng truy cập trung bình, trong khi một số ít ngày có lượt truy cập rất cao (tạo nên phần đuôi bên phải).
- Đỉnh của phân phối (mode) nằm khoảng **2000–2600 lượt truy cập**, nghĩa là đây là mức truy cập phổ biến nhất.
- Một số giá trị cao hơn 4000 là **ngày đặc biệt có lưu lượng truy cập cao**, có thể do chiến dịch quảng cáo hoặc sự kiện.

3.2.3. Bảng "ga_sessions_*".

3.2.3.1. Kiểm tra dữ liệu.

- Dữ liệu có tổng cộng 903653 bản ghi mỗi cột, tuy nhiên còn có nhiều dữ liệu bị thiếu, cụ thể:

Tên cột	geo_continent	geo_country	geo_city	traffic_source
Số dữ liệu bị thiếu	1468	1468	34262	69
Tên cột	traffic_medium	Totals_pageviews	totals_timeOnSite	totals_bounces
Số dữ liệu bị thiếu	117	100	451759	453023
Tên cột	totals_transactions	totals_transaction Revenue	totals_newVisits	
Số dữ liệu bị thiếu	892101	892138	200593	

- Thống kê mô tả:

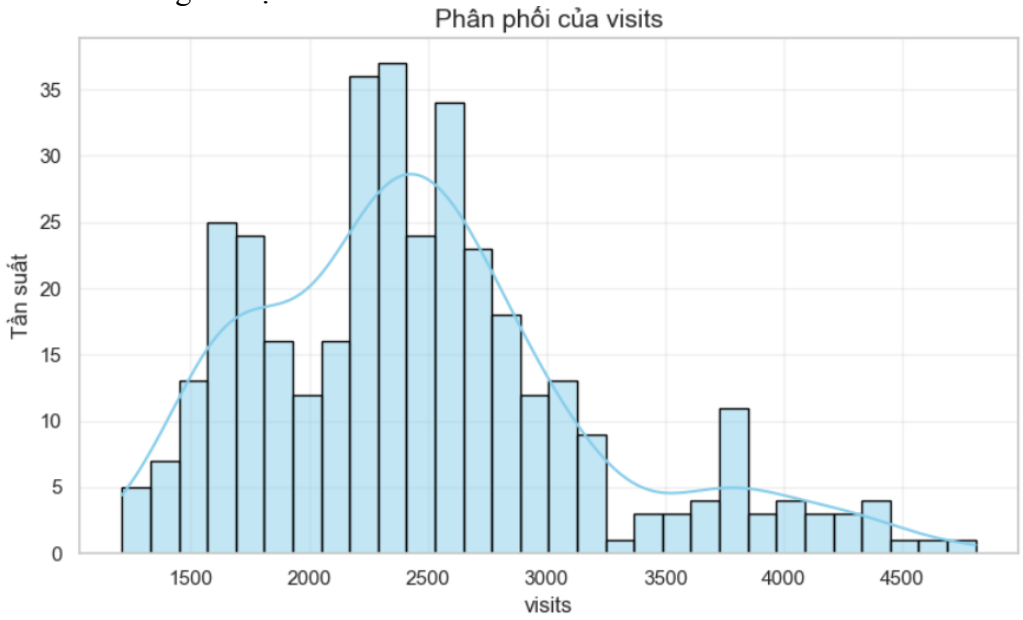
Pageviews - số trang người dùng			
Trung bình	Gía trị nhỏ nhất	Trung vị	Gía trị lớn nhất
3.849764	1	1	5
timeOnSite - thời gian người dùng			
Trung bình	Gía trị nhỏ nhất	Trung vị	Gía trị lớn nhất
262.624641	1	56	341
transactionRevenue - doanh thu phát sinh			
Trung bình	Gía trị nhỏ nhất	Trung vị	Gía trị lớn nhất
133744788.536691	10000	36980000	133980000

3.2.3.2. Phân phối dữ liệu.

3.2.3.2.1. Trường dữ liệu định lượng.

❖ Các trường dữ liệu định lượng: totals.visits, totals.hits, totals.pageviews, totals.timeOnSite, totals.bounces, totals.transactions, totals.transactionRevenue, totals.newVisits

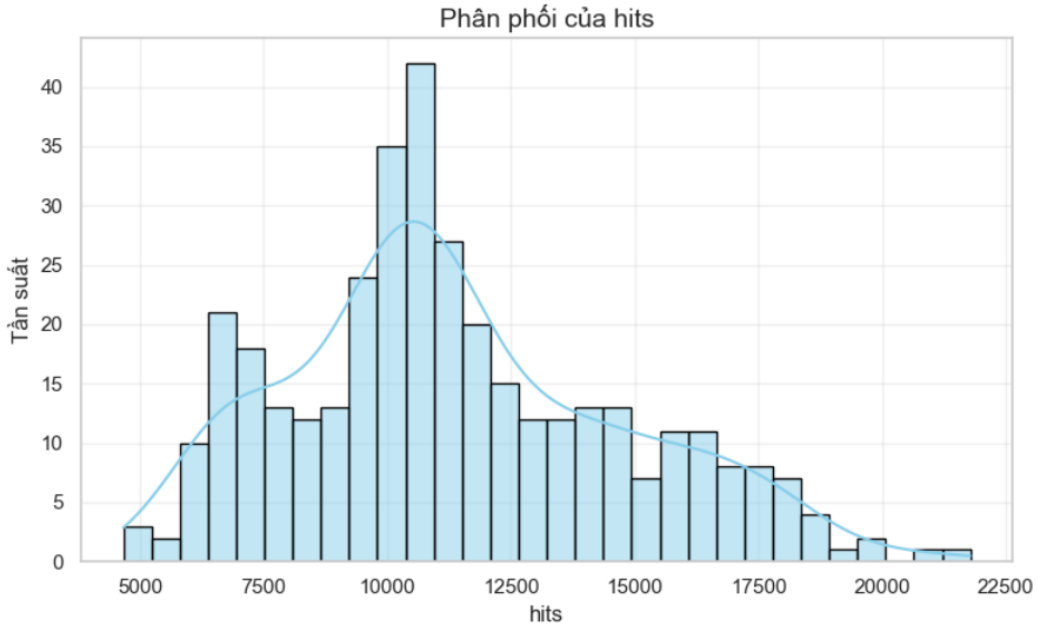
❖ Trường dữ liệu *totals.visits:*



➢ Biểu đồ trên minh họa phân phối số lượt truy cập (**visits**) trong dữ liệu:

- Trục hoành (x) biểu diễn **tổng số lượt truy cập.**
- Trục tung (y) biểu diễn **tần suất xuất hiện** (histogram) và **mật độ xác suất ước lượng** (đường KDE).
- **Hình dạng phân phối:** Phân phối này có hình dạng đa đỉnh (bimodal), với đỉnh chính rõ ràng và một cụm dữ liệu nhỏ hơn, riêng biệt, tạo thành một đỉnh phụ ở phía bên phải.
- **Đỉnh (Mode):** Đỉnh chính của phân phối nằm trong khoảng 2250–2600 lượt truy cập, cho thấy đây là mức truy cập phổ biến nhất.
- **Quan sát:** Cụm đỉnh phụ ở **khoảng 3500–4500** cho thấy có một số ngày hoặc nhóm dữ liệu với lượng truy cập cao hơn đáng kể so với mức trung bình, có thể là do các sự kiện hoặc chiến dịch đặc biệt.

❖ Trường dữ liệu *totals.hits*:

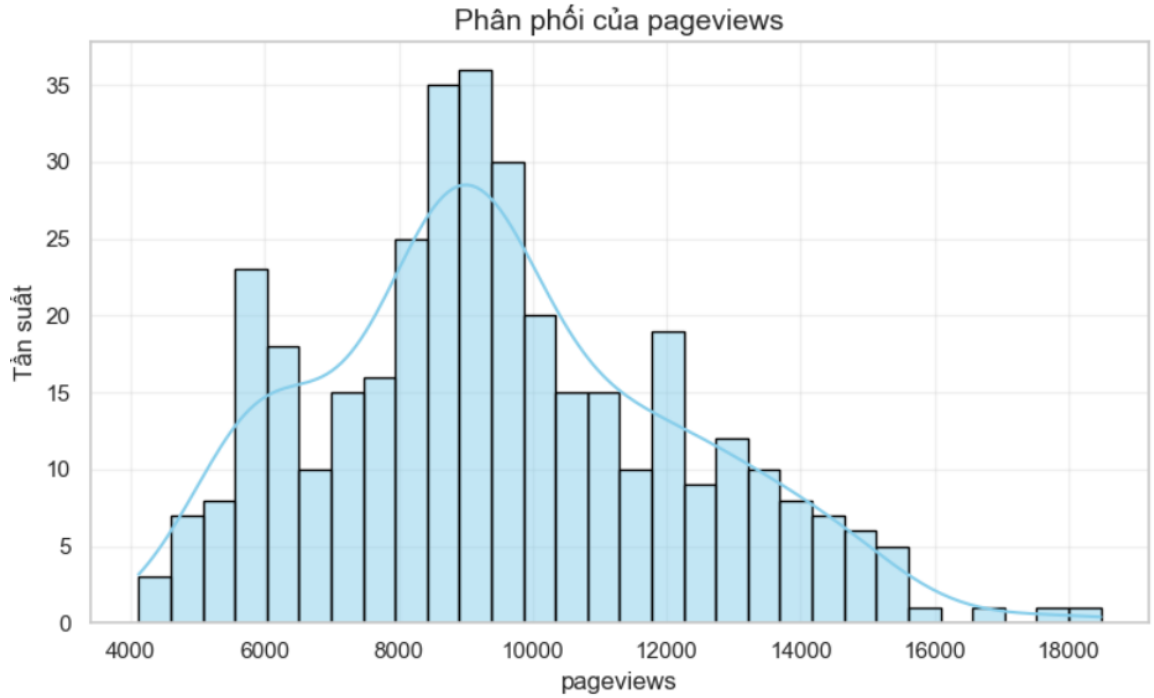


➢ Biểu đồ trên minh họa phân phối tổng số lần nhấn/tương tác (**hits**).

- Trục hoành (x) biểu diễn tổng số lần nhấn.

- Trục tung (y) biểu diễn tần suất xuất hiện và mật độ xác suất ước lượng.
- Hình dạng phân phối: Phân phối hơi lệch phải (right-skewed). Dữ liệu tập trung mạnh mẽ vào mức trung bình và giảm dần với một đuôi kéo dài về phía giá trị cao.
- Đỉnh (Mode): Đỉnh của phân phối nằm trong khoảng 10,000–11,250 lần nhấn, là mức tương tác phổ biến nhất.
- Quan sát: Phân phối có vẻ rộng (độ biến động cao), với dữ liệu trải dài từ 5,000 đến 22,500 hits, cho thấy sự khác biệt lớn về mức độ tương tác tổng thể.

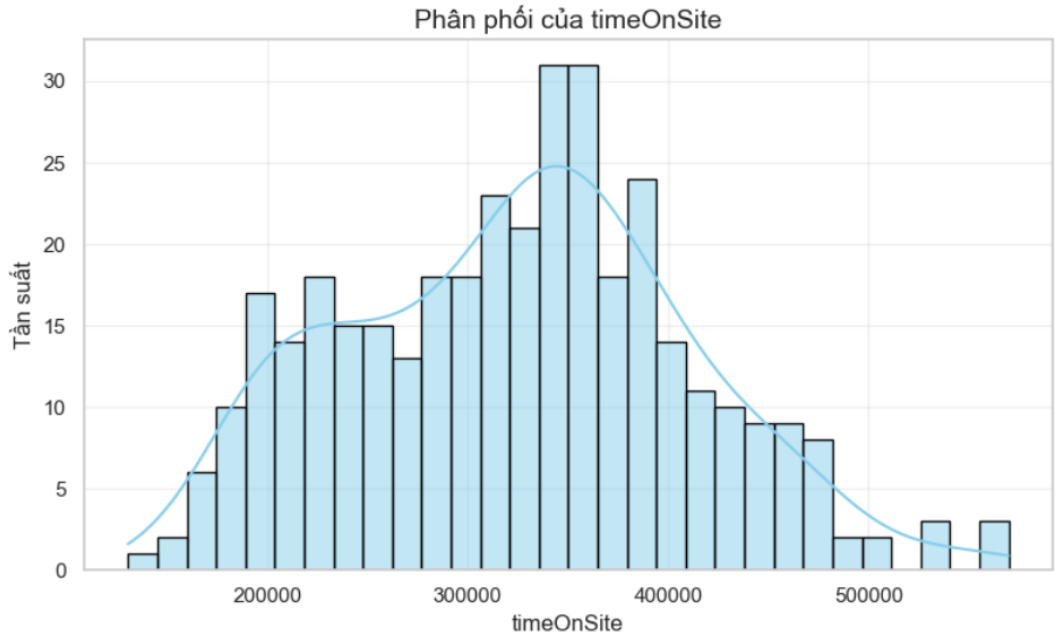
❖ Trường dữ liệu *totals.pageviews*:



➢ Biểu đồ trên minh họa phân phối tổng số lượt xem trang (**pageviews**).

- Trục hoành (x) biểu diễn tổng số lượt xem trang.
- Trục tung (y) biểu diễn tần suất xuất hiện và mật độ xác suất ước lượng.
- Hình dạng phân phối: Phân phối có độ lệch phải (right-skewed) nhẹ, tức là phần lớn dữ liệu tập trung xung quanh mức trung bình-thấp và có một đuôi dài kéo về phía các giá trị xem trang rất cao.
- Đỉnh (Mode): Đỉnh của phân phối nằm trong khoảng 9000–10000 lượt xem trang, là mức xem trang phổ biến nhất.
- Quan sát: Tần suất giảm dần sau đỉnh, nhưng vẫn có dữ liệu trải dài đến khoảng 18000 lượt xem trang, cho thấy có những ngày người dùng xem nhiều trang hơn đáng kể.

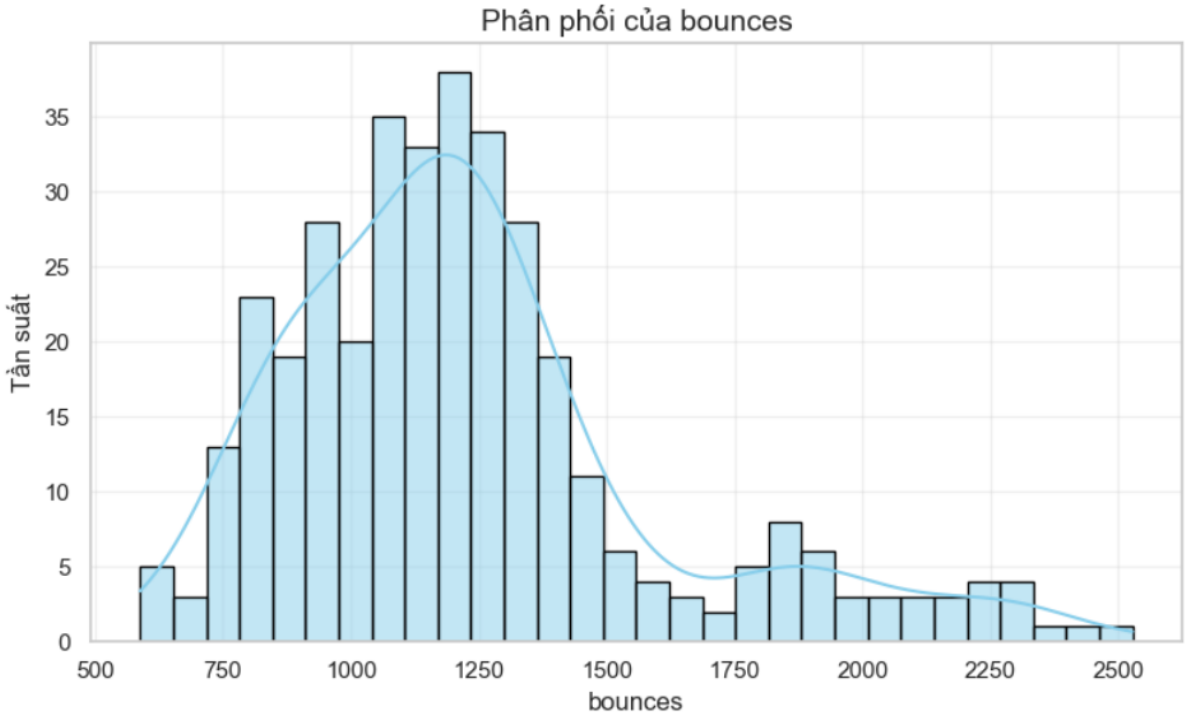
❖ Trường dữ liệu *totals.timeOnSite*:



➢ Biểu đồ trên minh họa phân phối tổng thời gian người dùng ở lại trên trang (**timeOnSite**).

- Trục hoành (x) biểu diễn tổng thời gian trên trang.
- Trục tung (y) biểu diễn tần suất xuất hiện và mật độ xác suất ước lượng.
- Hình dạng phân phối: Phân phối có vẻ gần đối xứng nhưng hơi lệch phải (right-skewed). Phân phối có một đỉnh duy nhất và phần đuôi bên phải dài hơn một chút.
- Đỉnh (Mode): Đỉnh của phân phối nằm trong khoảng 350,000–400,000 đơn vị thời gian (thường là giây hoặc mili giây), là mức thời gian ở lại trang phổ biến nhất.
- Quan sát: Dữ liệu trải dài một cách khá rộng, từ khoảng 150,000 đến 550,000, cho thấy sự đa dạng trong tổng thời gian người dùng dành cho trang web.

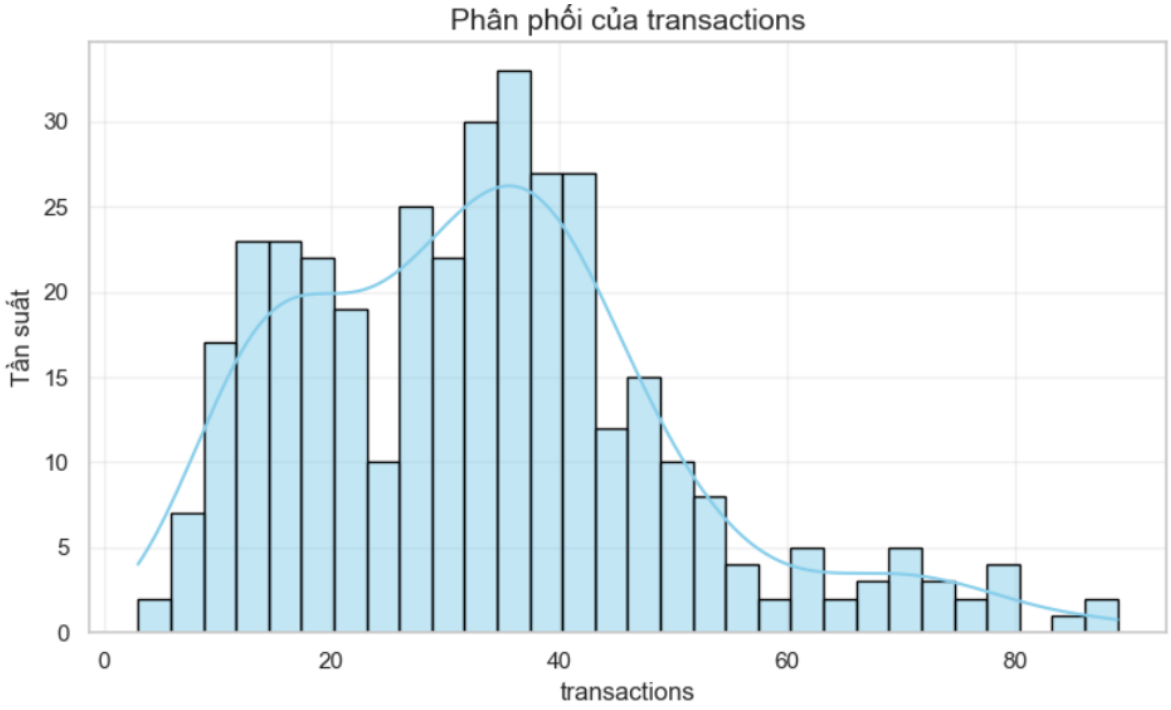
❖ Trường dữ liệu *totals.bounces*:



➢ Biểu đồ trên minh họa phân phối số lượt thoát trang (**bounces**).

- Trục hoành (x) biểu diễn số lượt thoát trang.
- Trục tung (y) biểu diễn tần suất xuất hiện và mật độ xác suất ước lượng.
- Hình dạng phân phối: Phân phối đa đỉnh (bimodal) rõ rệt, với đỉnh chính và một đỉnh phụ tách biệt hơn nhiều so với các biểu đồ khác.
- Đỉnh (Mode): Đỉnh chính nằm trong khoảng 1125–1375 lượt thoát, là mức phổ biến nhất.
- Quan sát: Đỉnh phụ nằm ở khoảng 1750–2500 lượt thoát. Sự xuất hiện của hai cụm này có thể chỉ ra hai nhóm hành vi người dùng hoặc hai loại hình truy cập khác nhau dẫn đến tỷ lệ thoát khác nhau.

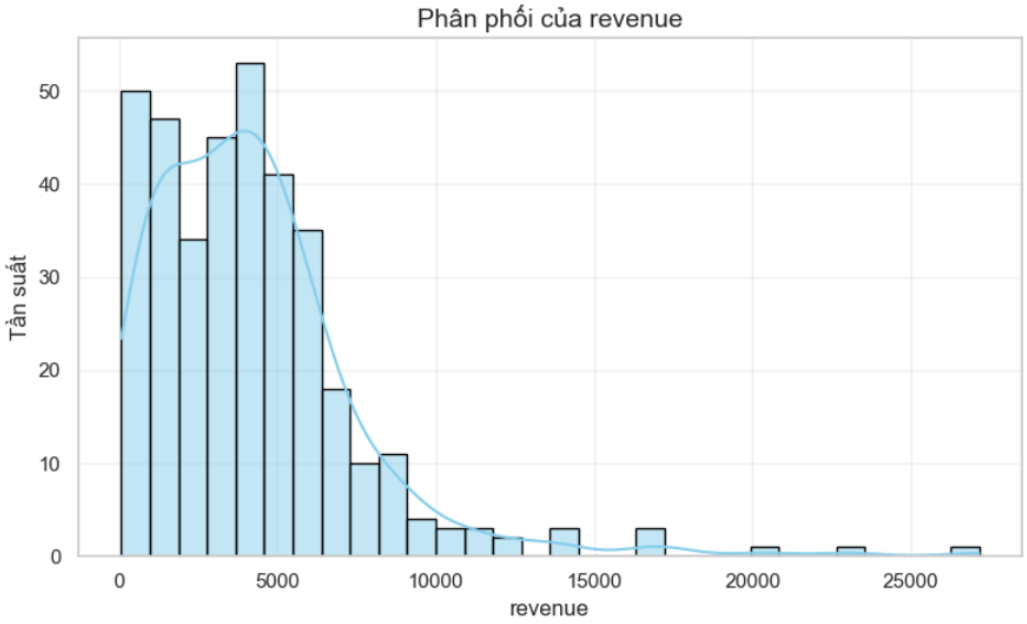
❖ Trường dữ liệu *totals.transactions*:



➢ Biểu đồ trên minh họa phân phối số lượng giao dịch (**transactions**).

- Trục hoành (x) biểu diễn số lượng giao dịch.
- Trục tung (y) biểu diễn tần suất xuất hiện và mật độ xác suất ước lượng.
- Hình dạng phân phối: Phân phối lệch phải (right-skewed), với phần lớn số ngày có lượng giao dịch thấp và một đuôi dài về phía các ngày có lượng giao dịch cao.
- Đỉnh (Mode): Đỉnh của phân phối nằm trong khoảng 35–45 giao dịch, là mức phổ biến nhất.
- Quan sát: Phân phối cho thấy sự tập trung cao của các ngày có dưới 50 giao dịch, nhưng vẫn có các ngày có số lượng giao dịch lên tới khoảng 90, cho thấy các ngày cao điểm bán hàng.

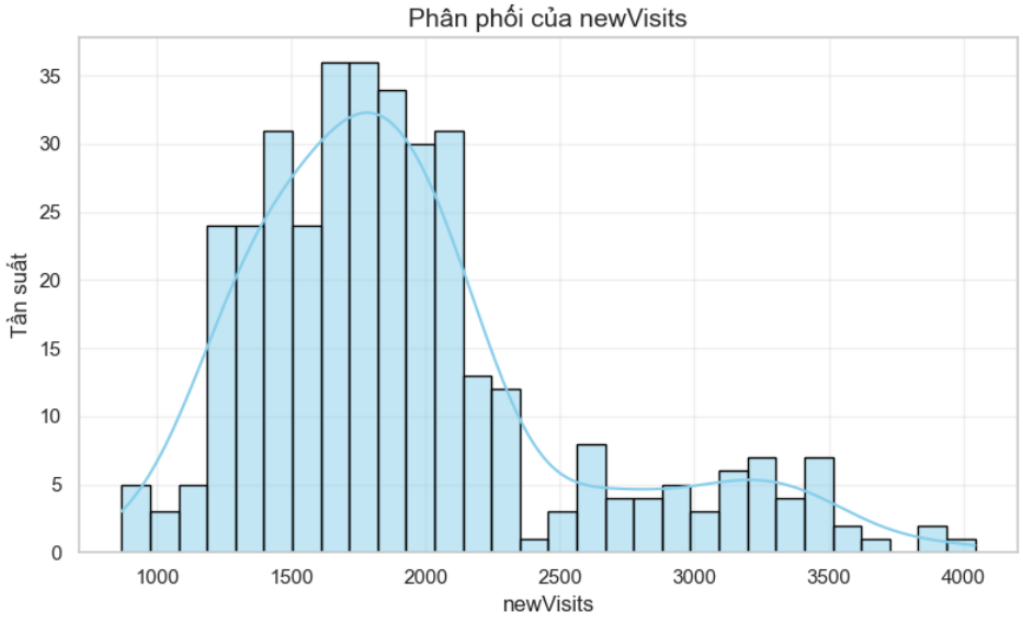
❖ Trường dữ liệu *totals.revenue*:



➢ Biểu đồ trên minh họa phân phối doanh thu (**revenue**).

- Trục hoành (x) biểu diễn giá trị doanh thu.
- Trục tung (y) biểu diễn tần suất xuất hiện và mật độ xác suất ước lượng.
- Hình dạng phân phối: Phân phối lệch phải (right-skewed) rất mạnh, cho thấy đây là một phân phối điển hình của dữ liệu tài chính. Phần lớn các giao dịch hoặc ngày có doanh thu thấp, trong khi có một số ít giao dịch/ngày tạo ra doanh thu rất cao (tạo nên phần đuôi dài).
- Đỉnh (Mode): Đỉnh của phân phối nằm trong khoảng 2500–5000 đơn vị doanh thu, là mức doanh thu phổ biến nhất.
- Quan sát: Tần suất giảm mạnh sau khoảng 7500, nhưng vẫn có các điểm dữ liệu cá biệt (outliers) kéo dài đến 27,500, minh họa sự chênh lệch lớn về giá trị giao dịch/doanh thu.

❖ Trường dữ liệu *totals.newVisits*:

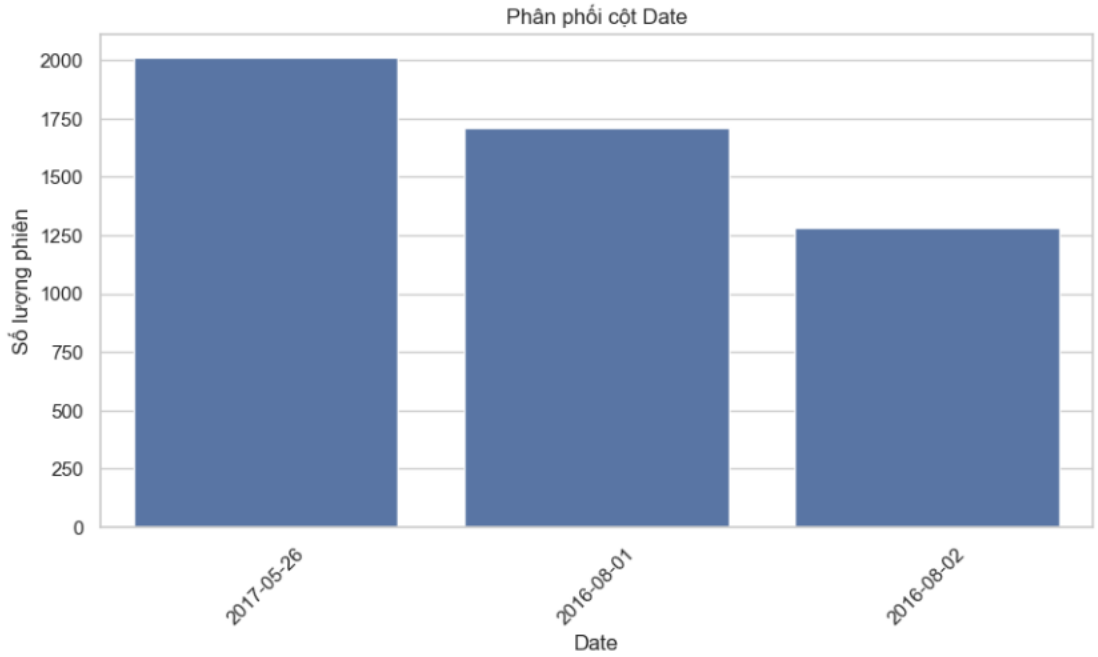


➢ Biểu đồ trên minh họa phân phối tổng thời gian người dùng ở lại trên trang (**timeOnSite**).
- Trục hoành (x) biểu diễn tổng thời gian trên trang.
- Trục tung (y) biểu diễn tần suất xuất hiện và mật độ xác suất ước lượng.
- Hình dạng phân phối: Phân phối có vẻ gần đối xứng nhưng hơi lệch phải (right-skewed). Phân phối có một đỉnh duy nhất và phần đuôi bên phải dài hơn một chút.
- Đỉnh (Mode): Đỉnh của phân phối nằm trong khoảng 350,000–400,000 đơn vị thời gian (thường là giây hoặc mili giây), là mức thời gian ở lại trang phổ biến nhất.
- Quan sát: Dữ liệu trải dài một cách khá rộng, từ khoảng 150,000 đến 550,000, cho thấy sự đa dạng trong tổng thời gian người dùng dành cho trang web.

3.2.3.2.2. Trường dữ liệu định tính.

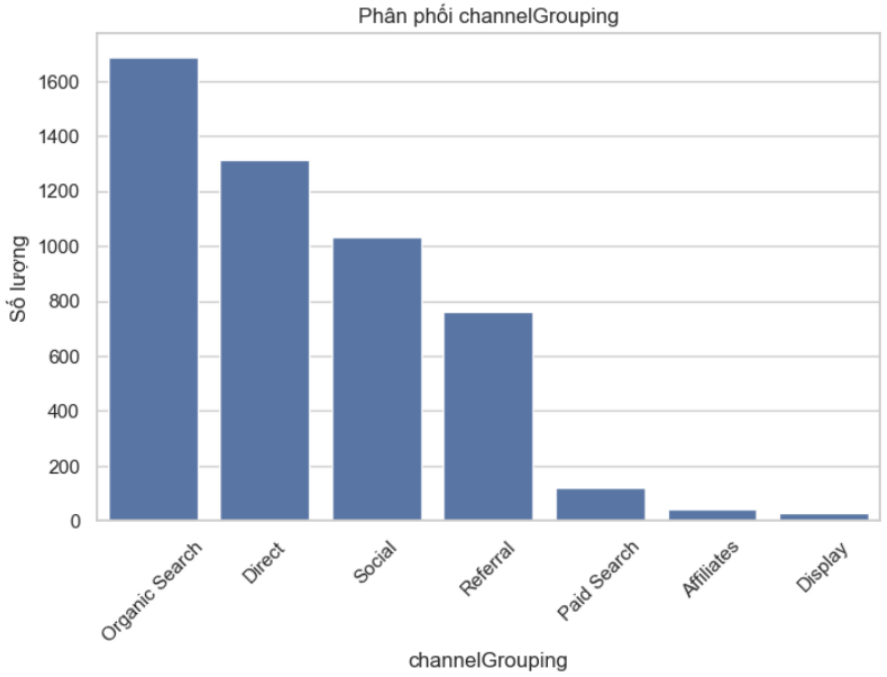
❖ Các trường dữ liệu định tính: fullVisitorId, visitId, Date, channelGrouping, socialEngagementType, Device Browser, Device Operating System, Device Category, continent, country, city, source, medium.

❖ Trường dữ liệu *Date*:



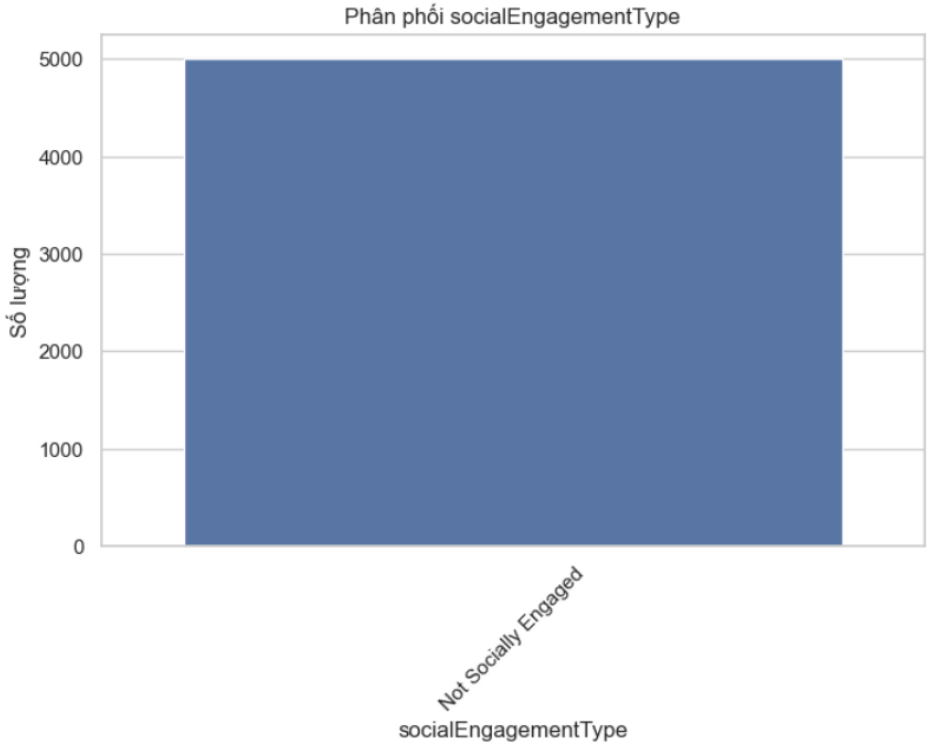
➢ Biểu đồ trên minh họa phân phối số lượng phiên trong Top 3 ngày có lưu lượng truy cập cao nhất.
- Trục hoành (x) biểu diễn ngày.
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Dữ liệu cho thấy sự khác biệt về lưu lượng truy cập giữa các ngày.
- Giá trị nổi bật:
  - Ngày 2017-05-26 có số lượng phiên cao nhất (2000).
  - Ngày 2016-08-01 đứng thứ hai (gần 1750).
  - Ngày 2016-08-02 đứng thứ ba (khoảng 1250).

❖ Trường dữ liệu *channelGrouping*:



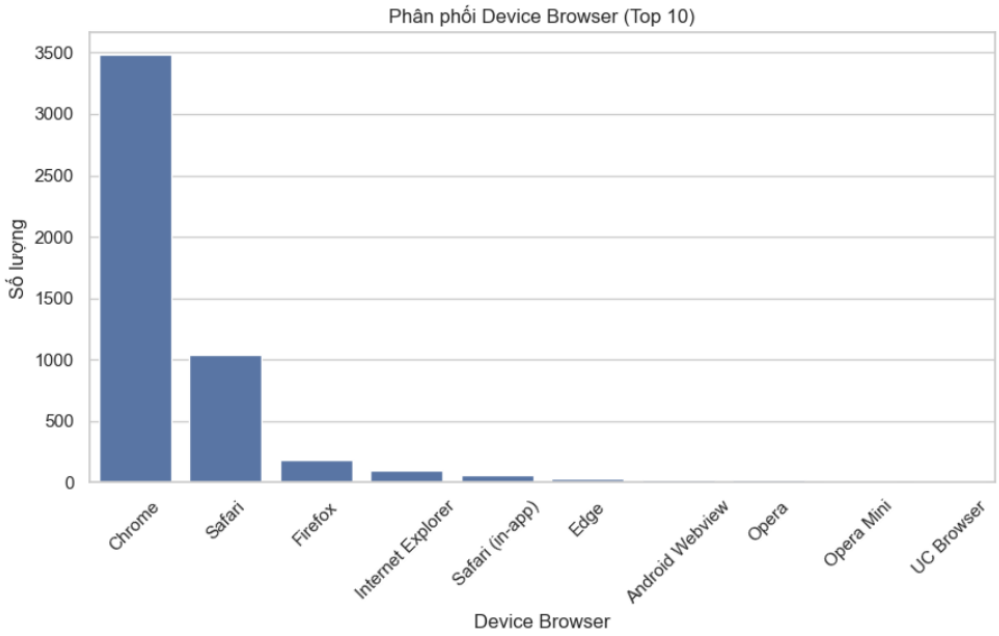
➢ Biểu đồ trên minh họa phân phối số lượng phiên theo nhóm kênh marketing.
- Trục hoành (x) biểu diễn tên nhóm kênh (Organic Search, Direct, Social, v.v.).
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Lưu lượng truy cập tập trung vào các kênh tự nhiên và trực tiếp.
- Giá trị nổi bật:
  - Organic Search (Tìm kiếm tự nhiên) dẫn đầu với số lượng phiên trên 1600.
  - Direct (Trực tiếp) đứng thứ hai với số lượng phiên khoảng 1300.
  - Social (Mạng xã hội) và Referral (Giới thiệu) theo sau với khoảng 1000 và 750 phiên.

❖ Trường dữ liệu *socialEngagementType*:



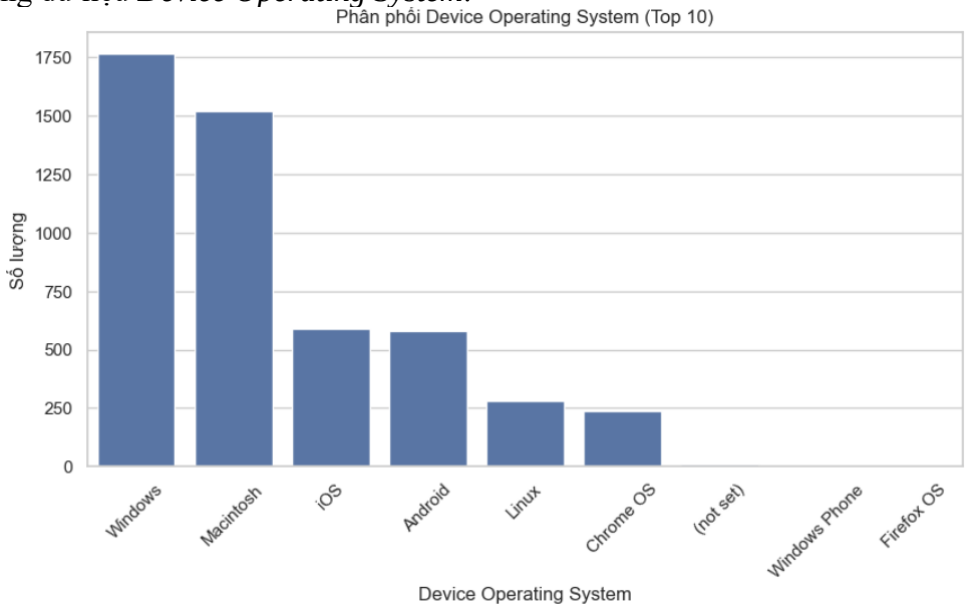
➢ Biểu đồ trên minh họa phân phối số lượng phiên theo loại tương tác xã hội.
- Trục hoành (x) biểu diễn loại tương tác.
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Toàn bộ dữ liệu nằm trong một hạng mục duy nhất.
- Giá trị nổi bật: Tất cả các phiên (khoảng 5000) đều thuộc hạng mục "Not Socially Engaged"

❖ Trường dữ liệu *Device Browser*:



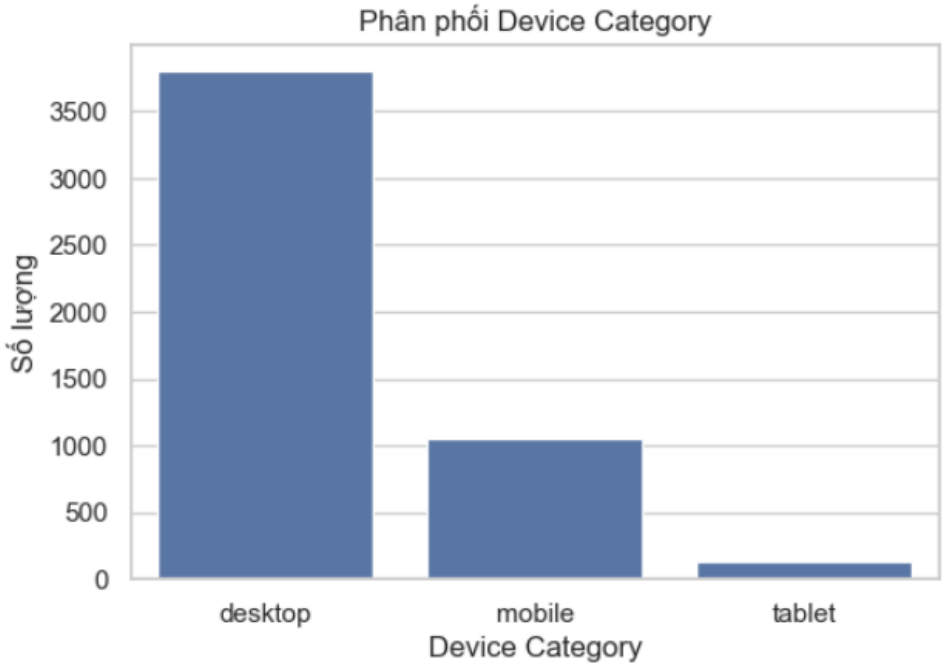
➢ Biểu đồ trên minh họa phân phối Top 10 trình duyệt được sử dụng.
- Trục hoành (x) biểu diễn tên trình duyệt.
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Có sự chiếm lĩnh thị trường của một trình duyệt.
- Giá trị nổi bật: Chrome chiếm ưu thế tuyệt đối với số lượng phiên gần 3500, cao gấp hơn 3 lần so với trình duyệt đứng thứ hai là Safari (khoảng 1000).

❖ Trường dữ liệu *Device Operating System*:



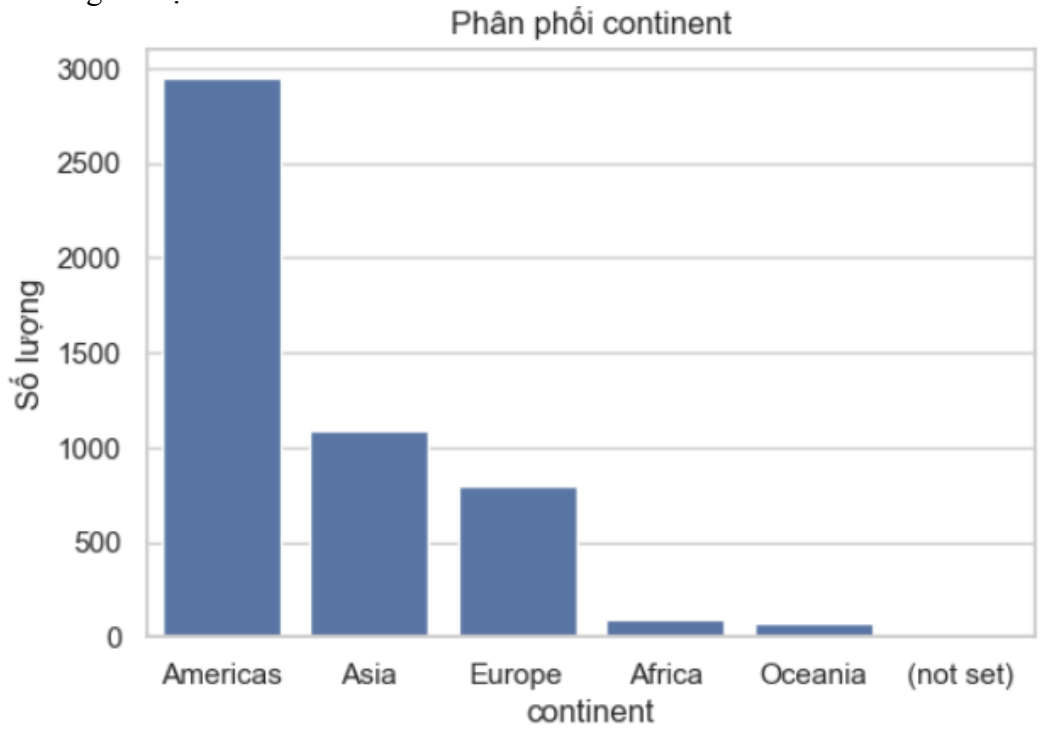
➢ Biểu đồ trên minh họa phân phối Top 10 hệ điều hành được sử dụng.
- Trục hoành (x) biểu diễn tên hệ điều hành.
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Phân phối cho thấy hai hệ điều hành dẫn đầu.
- Giá trị nổi bật: Windows và Macintosh (macOS) là hai hệ điều hành thống trị, với số lượng phiên lần lượt là trên 1750 và trên 1500.
- Quan sát: Sự ưu thế của Windows và Macintosh phù hợp với sự thống trị của loại thiết bị desktop được ghi nhận ở trên. Các hệ điều hành di động như iOS và Android đứng ở vị trí thứ ba và thứ tư (khoảng 600 phiên).

❖ Trường dữ liệu *Device Category*:



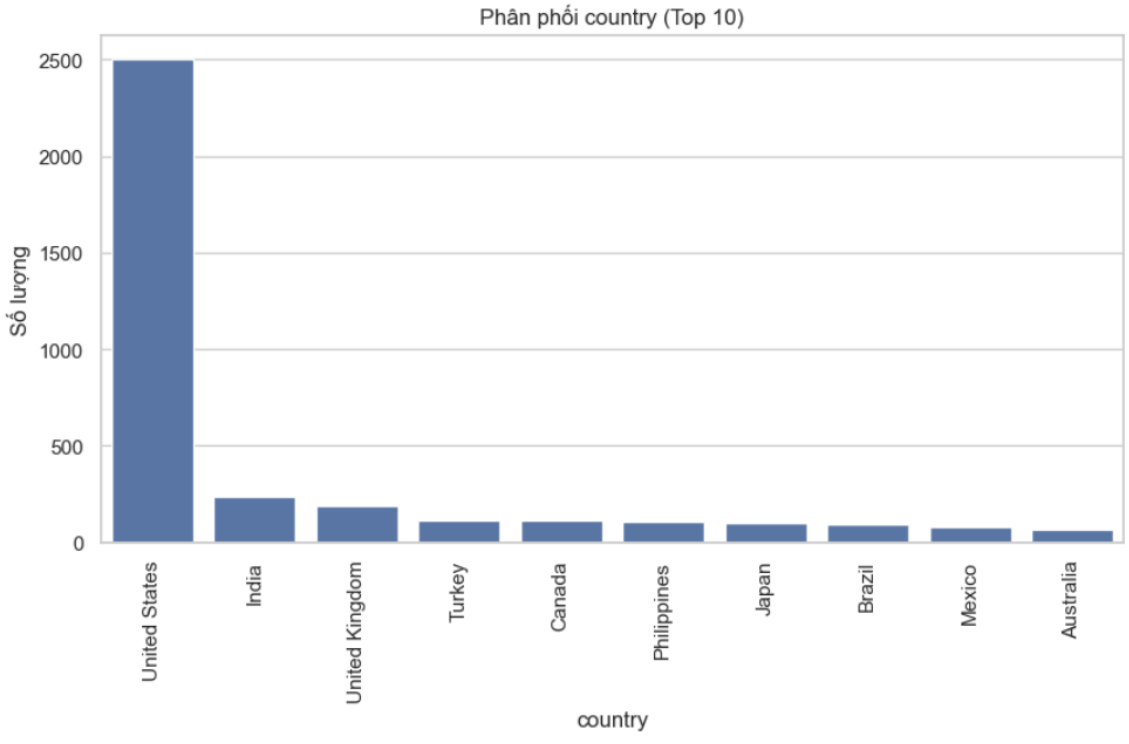
➢ Biểu đồ trên minh họa phân phối số lượng phiên theo loại thiết bị.
- Trục hoành (x) biểu diễn loại thiết bị (desktop, mobile, tablet).
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Có sự chênh lệch lớn giữa các loại thiết bị.
- Giá trị nổi bật: Thiết bị desktop (máy tính để bàn) chiếm ưu thế tuyệt đối với số lượng phiên gần 3750, gấp hơn 3 lần so với thiết bị mobile (di động) (khoảng 1000). Thiết bị tablet (máy tính bảng) có số lượng rất thấp (dưới 500).

❖ Trường dữ liệu *continent*:



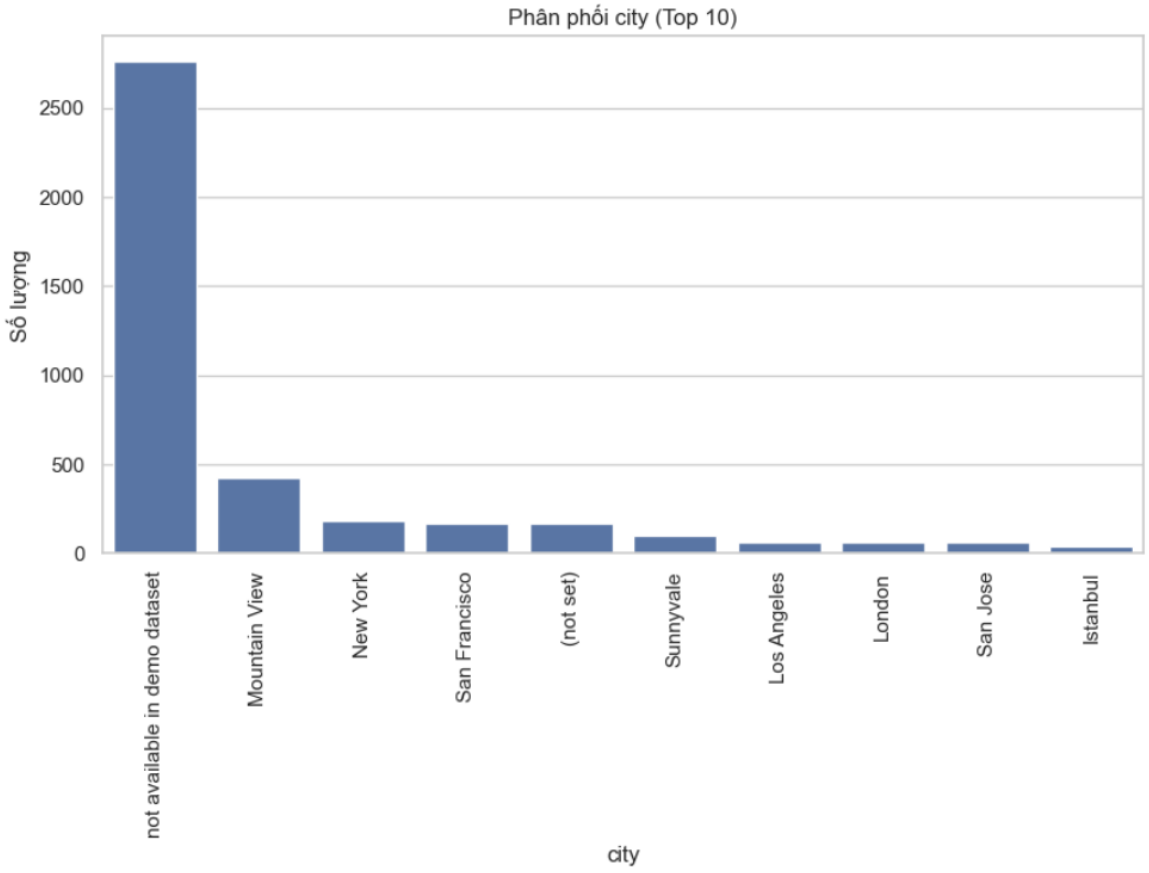
➢ Biểu đồ trên minh họa phân phối số lượng phiên theo châu lục.
- Trục hoành (x) biểu diễn tên châu lục.
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Phân phối lệch rất mạnh về một châu lục.
- Giá trị nổi bật: Châu lục Americas (Châu Mỹ) chiếm phần lớn với số lượng phiên gần 3000, cao hơn khoảng 3 lần so với châu lục đứng thứ hai là Asia (Châu Á) (khoảng 1000).

❖ Trường dữ liệu *country*:



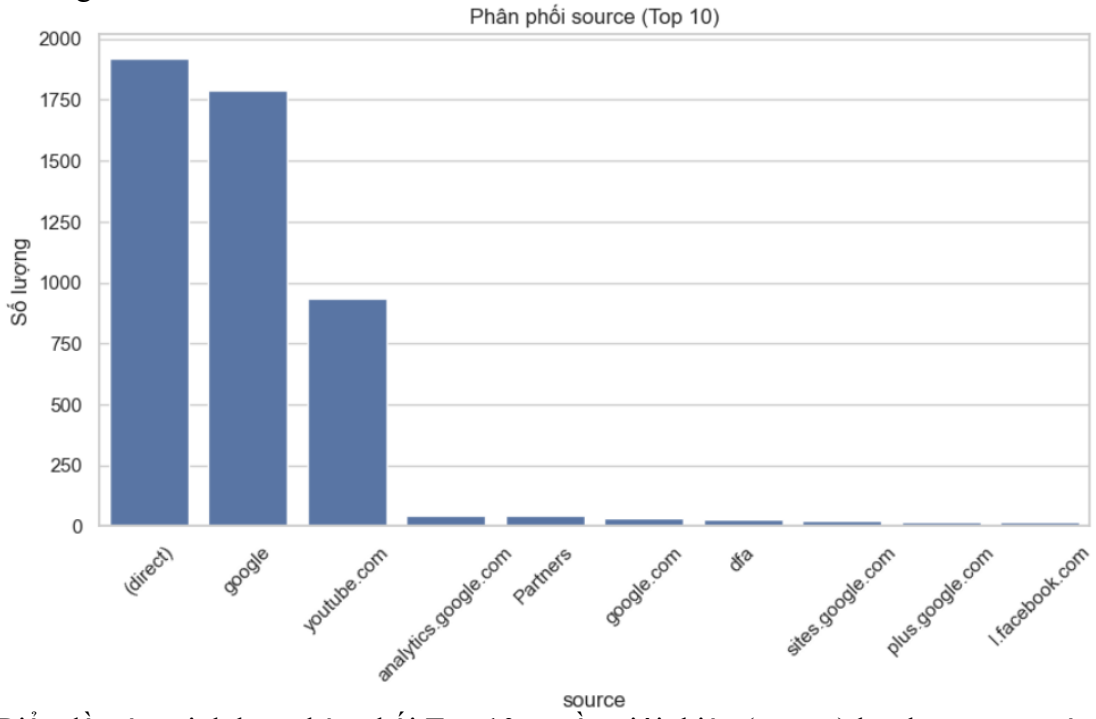
➢ Biểu đồ trên minh họa phân phối Top 10 quốc gia có số lượng phiên truy cập cao nhất.
- Trục hoành (x) biểu diễn tên quốc gia.
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Phân phối cho thấy sự tập trung cực kỳ cao vào một quốc gia duy nhất.
- Giá trị nổi bật: United States (Mỹ) chiếm ưu thế áp đảo với số lượng phiên gần 2500, cao hơn gấp 10 lần so với quốc gia đứng thứ hai là India (khoảng 250).

❖ Trường dữ liệu *city*:



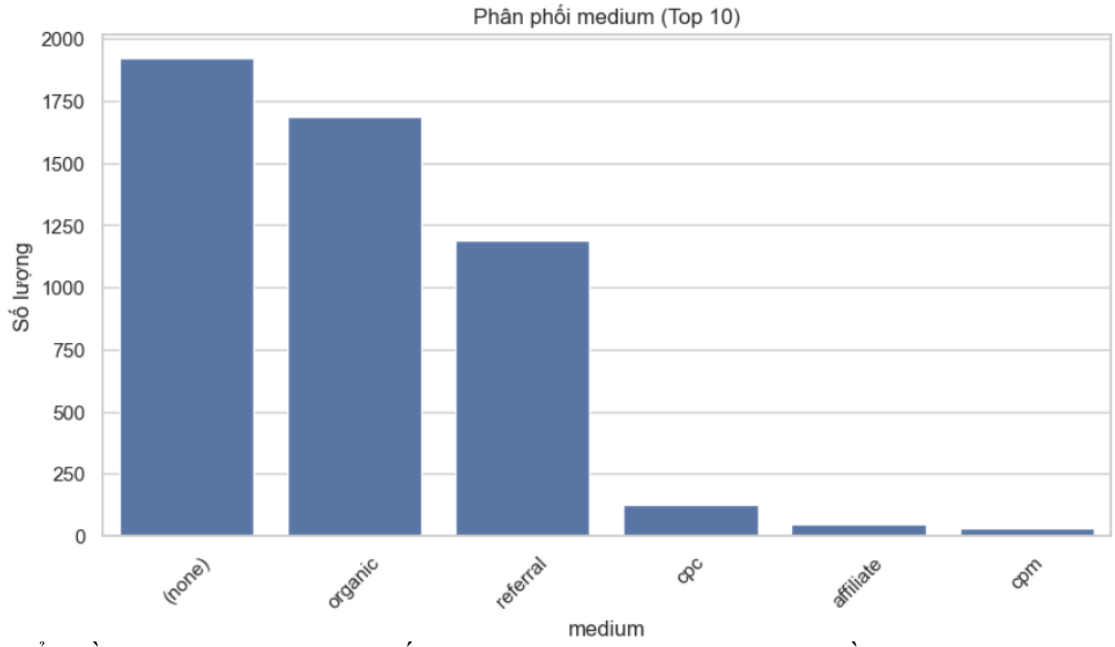
➢ Biểu đồ trên minh họa phân phối Top 10 thành phố có số lượng phiên cao nhất.
- Trục hoành (x) biểu diễn tên thành phố.
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Giá trị nổi bật nhất không phải là một thành phố cụ thể mà là một giá trị khuyết.
- Giá trị nổi bật: Hạng mục "not available in demo dataset" (không có sẵn trong bộ dữ liệu demo) chiếm vị trí cao nhất với số lượng phiên trên 2500.

❖ Trường dữ liệu *source*:



➢ Biểu đồ trên minh họa phân phối Top 10 nguồn giới thiệu (source) lưu lượng truy cập.
- Trục hoành (x) biểu diễn tên nguồn.
- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Lưu lượng truy cập tập trung vào ba nguồn chính.
- Giá trị nổi bật:
  - (direct) (Truy cập trực tiếp) đứng đầu với số lượng phiên gần 1900.
  - google đứng thứ hai với số lượng phiên trên 1750.
  - youtube.com đứng thứ ba với số lượng phiên gần 1000.

❖ Trường dữ liệu *medium*:



➢ Biểu đồ trên minh họa phân phối Top 6 phương tiện (medium) truyền tải lưu lượng truy cập.
- Trục hoành (x) biểu diễn tên phương tiện (ví dụ: (none), organic, referral, cpc).

- Trục tung (y) biểu diễn số lượng phiên.
- Quan sát chính: Lưu lượng truy cập tập trung chủ yếu vào ba phương tiện hàng đầu.
- Giá trị nổi bật:
  - Phương tiện (none) (không xác định/trực tiếp) dẫn đầu với số lượng phiên gần 1900. Giá trị này thường đi kèm với nguồn (direct).
  - Phương tiện organic (tìm kiếm tự nhiên) đứng thứ hai với số lượng phiên khoảng 1700. Phương tiện này đi kèm với nguồn google.
  - Phương tiện referral (giới thiệu) đứng thứ ba với số lượng phiên khoảng 1200.

❖ Trường dữ liệu *fullVisitorId*:

- Vì trường dữ liệu *fullVisitorId* có **hàng nghìn giá trị duy nhất**, việc vẽ biểu đồ cột hoặc phân phối sẽ **không có ý nghĩa trực quan** (biểu đồ sẽ có quá nhiều cột và không thể đọc được).
- Thay vào đó, ta chỉ **thống kê được** website có **4.605 người dùng riêng biệt**, thể hiện quy mô lượng người truy cập trong năm. Đây là cơ sở để đánh giá **mức độ tiếp cận khách hàng** và **tính trung thành** (qua tần suất truy cập lặp lại).

❖ Trường dữ liệu *visitId*:

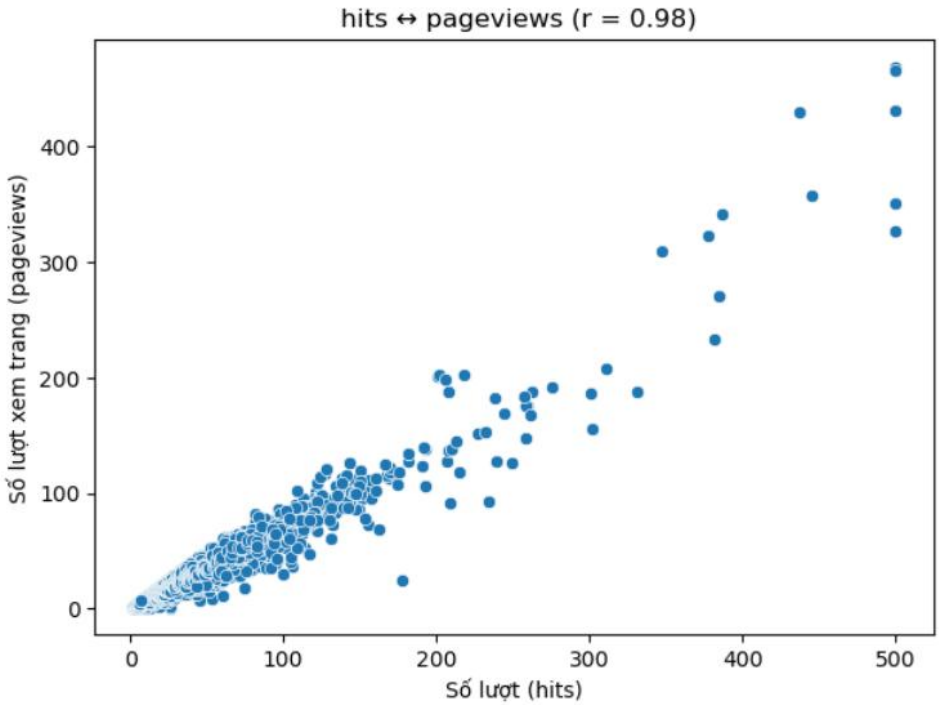
- Vì trường dữ liệu *visitId* có **rất nhiều giá trị duy nhất (gần 5.000)**, việc vẽ biểu đồ cột hoặc histogram sẽ không mang lại ý nghĩa trực quan (cột dày đặc, khó quan sát).
- Thay vào đó, ta thống kê dữ liệu có **4.942 phiên truy cập riêng biệt**, phản ánh mức độ tương tác của người dùng với website.

➢ So sánh với 4.605 người dùng duy nhất, có thể thấy trung bình **mỗi người dùng có hơn 1 phiên truy cập**, cho thấy có **sự quay lại của khách hàng** hoặc **hành vi truy cập lặp lại**.

3.2.3.3. Sự tương quan.

3.2.3.3.1. Tương quan 2 chiều.

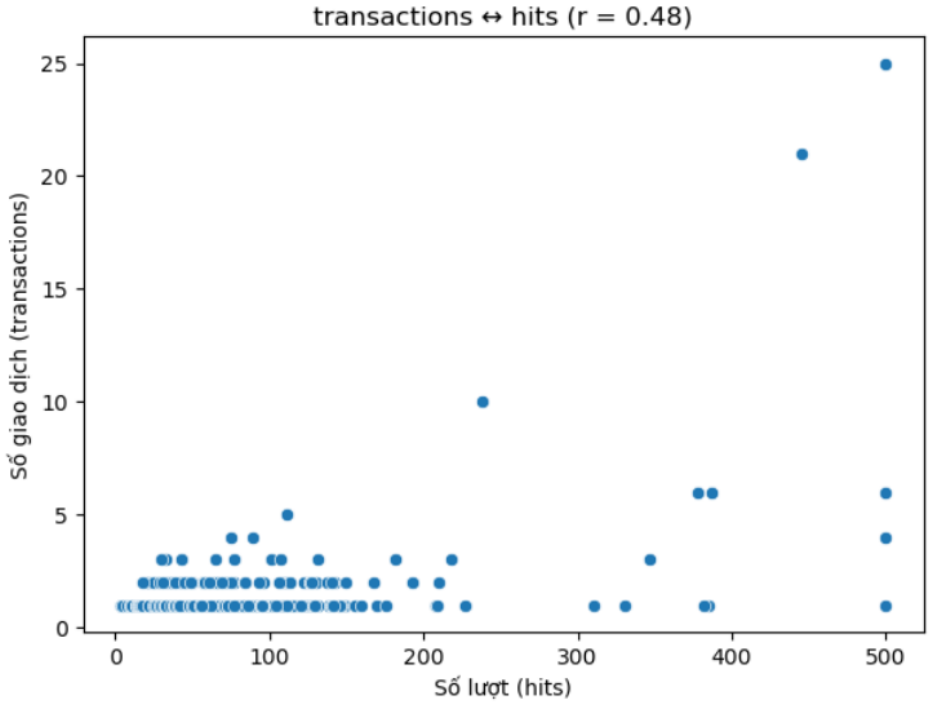
❖ **hits ↔ pageviews:**



- Hai biến có **tương quan dương rất mạnh** ($r \approx 0.98$), thể hiện rằng khi số lượt "hits" tăng thì số lượt xem trang cũng tăng gần như tương ứng.

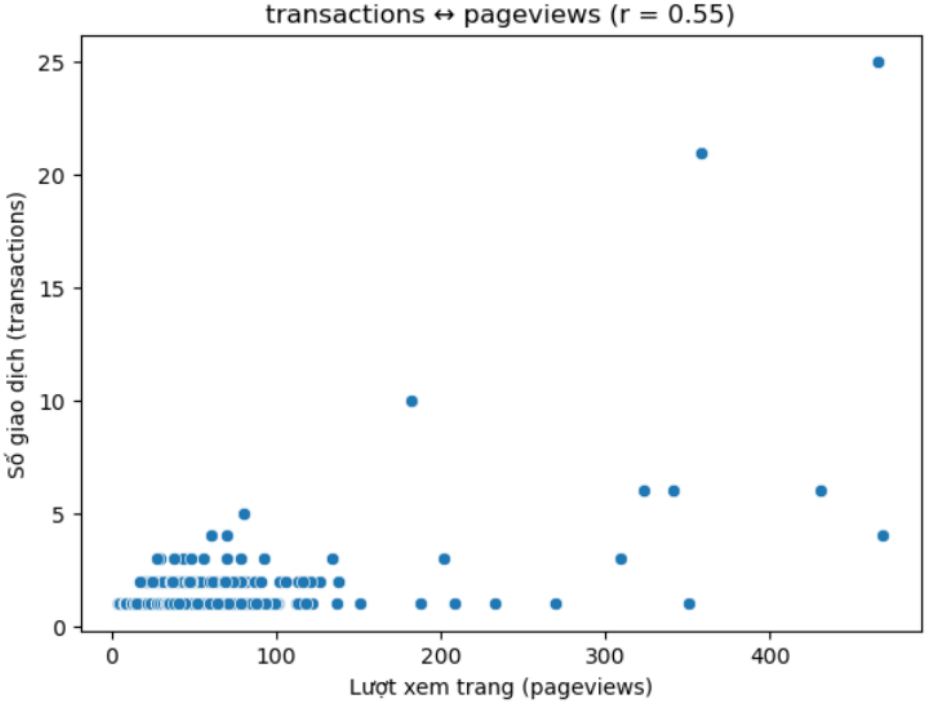
- **Ý nghĩa thực tế:** Phần lớn các hành động (hits) đến từ việc người dùng tải và xem trang, chứng tỏ các trang web có mức độ tương tác cao khi người dùng duyệt nội dung.

❖ **transactions ↔ hits**:



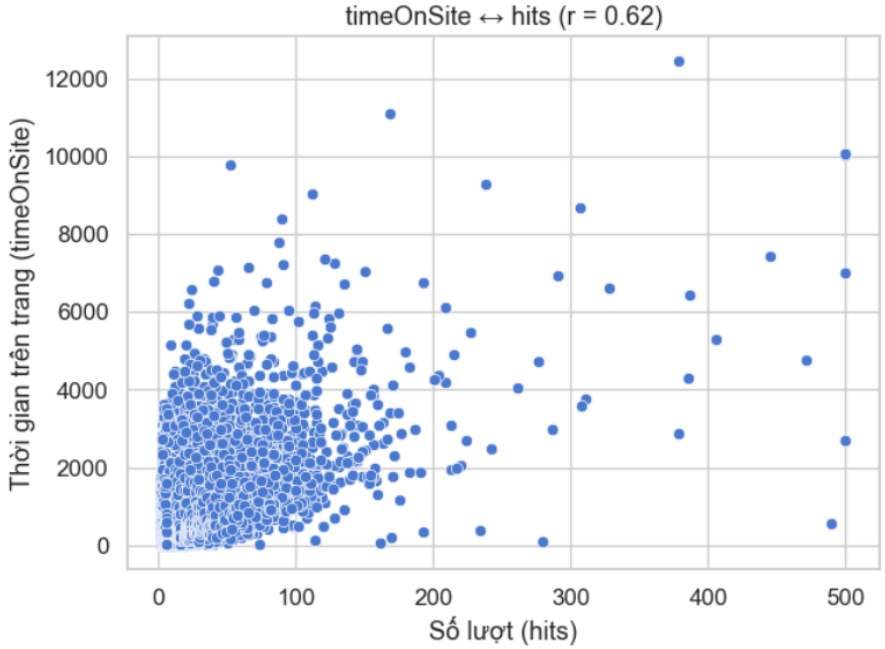
- Có **mối tương quan dương trung bình** ($r \approx 0.48$). Khi số lượt tương tác tăng, số giao dịch cũng có xu hướng tăng.
- **Ý nghĩa thực tế:** Lượng người dùng tương tác cao trên website góp phần tăng tỷ lệ chuyển đổi, nhưng chưa thật mạnh — có thể do không phải mọi lượt truy cập đều dẫn đến giao dịch.

❖ **transactions ↔ pageviews**:



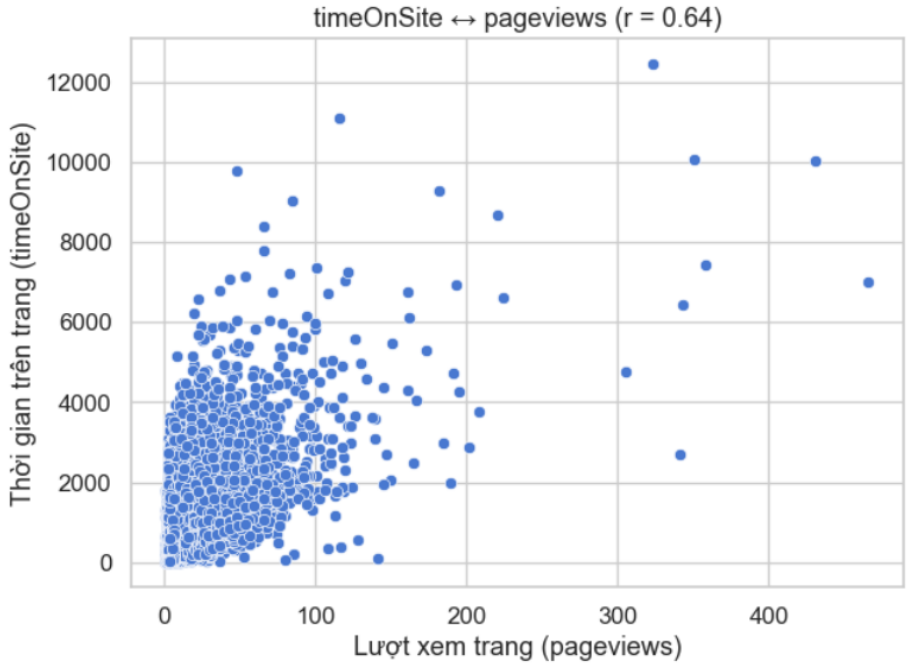
- Tương quan dương mức **trung bình khá** ($r \approx 0.55$).
- **Ý nghĩa thực tế:** Người dùng xem càng nhiều trang thì khả năng hoàn tất giao dịch càng cao — phản ánh hành vi tìm hiểu sản phẩm trước khi mua.

❖ **timeOnSite ↔ hits**:



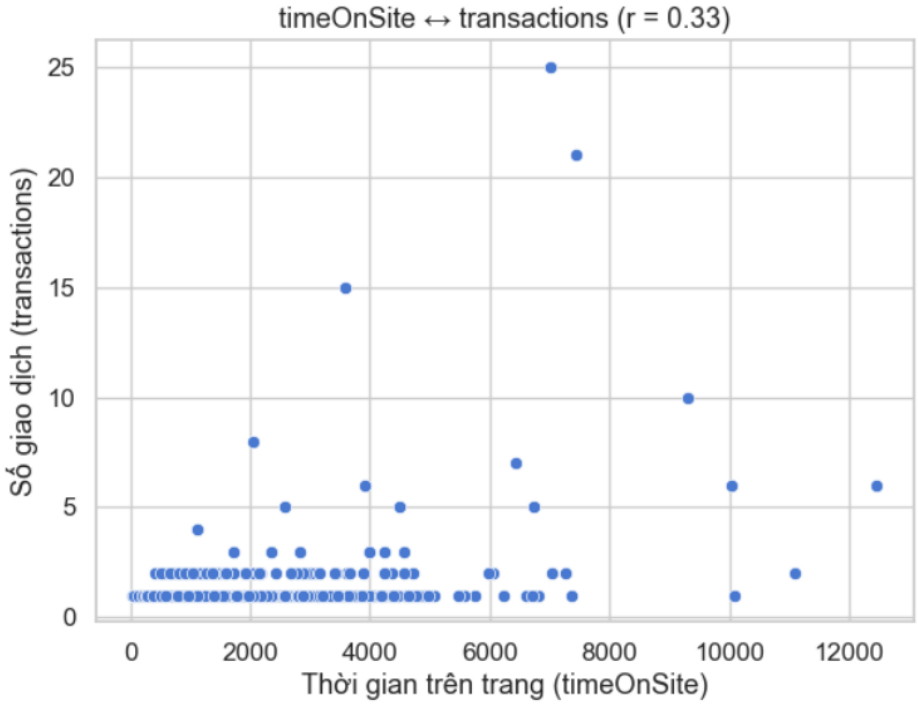
- Mối tương quan dương khá mạnh ($r \approx 0.62$), cho thấy người dùng có nhiều hoạt động (hits) hơn thường ở lại trang lâu hơn.
- **Ý nghĩa thực tế:** Điều này thể hiện trải nghiệm người dùng tích cực — các trang thu hút, giữ chân người xem.

❖ **timeOnSite ↔ pageviews**:



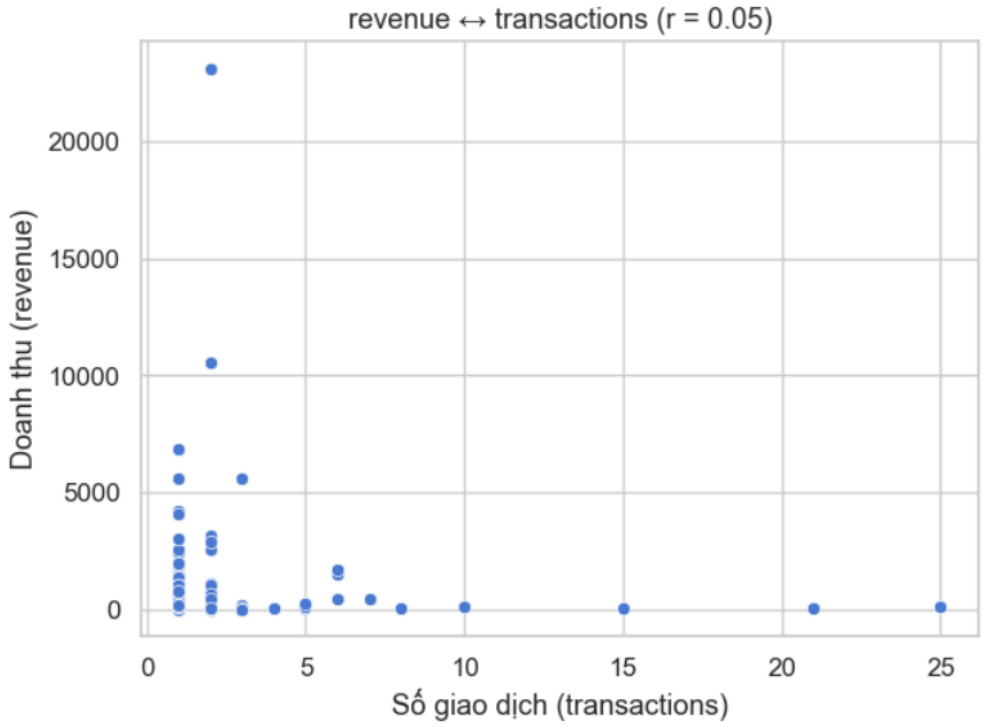
- Có **mối tương quan dương rõ rệt** ($r \approx 0.64$).
- **Ý nghĩa thực tế:** Khi người dùng xem nhiều trang hơn, thời gian họ ở lại trang web cũng tăng — phản ánh khả năng giữ chân người dùng của website.

❖ **timeOnSite ↔ transactions**:



- Mối tương quan dương yếu ($r \approx 0.33$).
- **Ý nghĩa thực tế:** Người ở lại trang lâu **có xu hướng** mua hàng nhiều hơn, nhưng không phải tất cả. Có thể một phần người dùng chỉ tìm kiếm thông tin mà không thực hiện giao dịch.

❖ **transactions ↔ revenue**:



- Mối tương quan dương rất yếu ($r \approx 0.05$).
- **Ý nghĩa thực tế:** Doanh thu không tăng đáng kể theo số giao dịch, có thể do **chênh lệch giá trị đơn hàng** — một số giao dịch có giá trị cao, trong khi nhiều giao dịch nhỏ ảnh hưởng ít đến tổng doanh thu.

➢ Tổng kết sự tương quan giữa các trường dữ liệu:

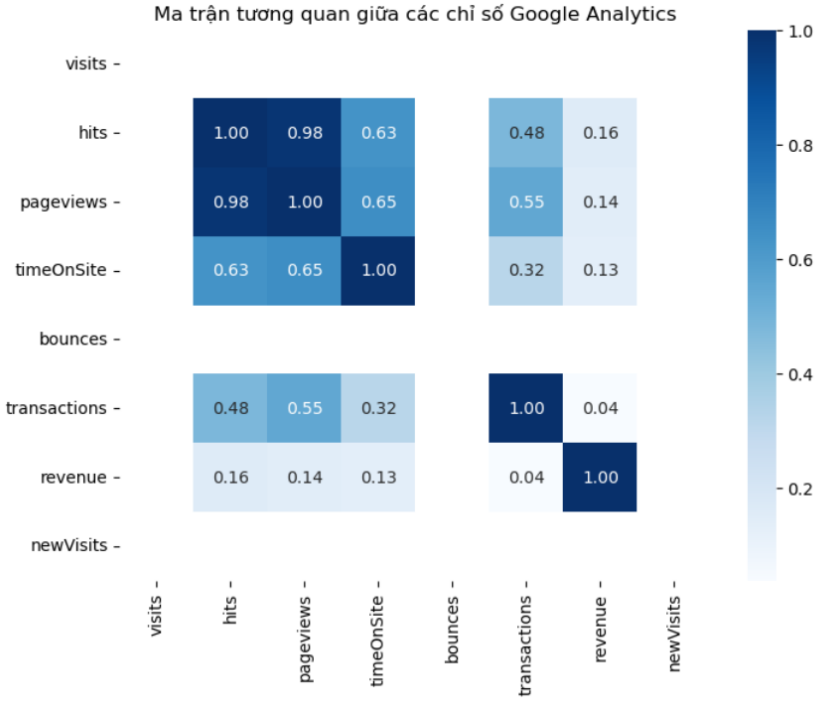
Cặp tương quan	Hệ số tương quan (r)	Mức độ	Nhân xét
hits ↔ pageviews	0.98	Rất mạnh	Hai chỉ số phản ánh cùng một hành vi người dùng
transactions ↔ hits	0.48	Trung bình	Tăng tương tác dẫn đến tăng giao dịch
transactions ↔ pageviews	0.55	Trung bình	Người xem nhiều trang hơn dễ mua hàng hơn
timeOnSite ↔ hits	0.62	Khá mạnh	Nhiều hoạt động hơn → ở lại lâu hơn
timeOnSite ↔ pageviews	0.64	Khá mạnh	Xem nhiều trang → ở lại lâu hơn
timeOnSite ↔ transactions	0.33	Yếu	Ở lâu có xu hướng mua nhiều hơn
transactions ↔ revenue	0.05	Rất yếu	Giao dịch không tỷ lệ chặt với doanh thu

3.2.3.3.2. Tương quan đa chiều.

❖ Ma trận tương quan:

	visits	hits	pageviews	timeOnSite	Bounces	transactions	revenue	newVisits
visits	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
hits	NaN	1.00	0.98	0.63	NaN	0.48	0.16	NaN
pageviews	NaN	0.98	1.00	0.65	NaN	0.55	0.14	NaN
timeOnSite	NaN	0.63	0.65	1.00	NaN	0.32	0.13	NaN
bounces	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
transactions	NaN	0.48	0.55	0.32	NaN	1.00	0.04	NaN
revenue	NaN	0.16	0.14	0.13	NaN	0.04	1.00	NaN
newVisits	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

❖ Đồ thị heatmap tương quan:



CHƯƠNG 4: XỬ LÝ DỮ LIỆU

Tệp dữ liệu gồm 3 bảng dữ liệu: *daily_visits*, *daily_total_visits* và *ga_sessions_**, với dung lượng hơn 1gb; trong đó có 2 bảng *daily_visits*, và *daily_total_visits* dữ liệu đầy đủ không bị khuyết dữ liệu nào, tuy nhiên bảng *ga_sessions_** còn khuyết khá nhiều dữ liệu vàn cần phải xử lý.

4.1. Xử lý giá trị khuyết thiếu.

- Chia bảng dữ liệu thành các trường dữ liệu định tính và định lượng (vì 2 trường dữ liệu này sẽ có cách xử lý làm sạch khác nhau):

```
cols_numeric = [
    "visits", "hits", "pageviews", "timeOnSite",
    "bounces", "transactions", "transactionRevenue", "newVisits"
]

cols_categorical = [
    "channelGrouping", "socialEngagementType", "browser",
    "operatingSystem", "deviceCategory", "continent",
    "country", "city", "source", "medium"
]
```

4.1.1. Xử lý trường dữ liệu định lượng.

- Kiểm tra giá trị khuyết thiếu: đếm các giá trị NaN của từng trường dữ liệu

```
print("\nSố lượng giá trị thiếu mỗi cột:")
print(df_num.isna().sum())
```

```
Số lượng giá trị thiếu mỗi cột:
pageviews               0
hits                    0
time_on_site         1246
bounces              1318
newVisits             684
transactions         2513
transactionRevenue   2513
sessionQualityDim       0
dtype: int64
```

➢ Nhận xét: các trường dữ liệu *"transactionRevenue", "transactions", "time_on_site"* và *"bounces"* còn thiếu nhiều dữ liệu, tuy nhiên các trường dữ liệu *"pageviews", "hits"* và *"sessionQualityDim"* lại không thiếu dữ liệu nào.

- Xử lý giá trị khuyết thiếu:

```
# Xử lý dữ liệu bị thiếu
from sklearn.impute import SimpleImputer
import numpy as np

# Cột dạng tỷ lệ/đếm => điền 0
cols_zero = ["bounces", "newVisits", "transactions", "transactionRevenue"]
df_num[cols_zero] = df_num[cols_zero].fillna(0)

# Các cột còn lại => điền median
cols_median = [col for col in df_num.columns if col not in cols_zero]
imputer = SimpleImputer(strategy='median')
df_num[cols_median] = imputer.fit_transform(df_num[cols_median])
```

➢ Nhận xét:

- Các trường dữ liệu thể hiện số lần hoặc giá trị đếm /tỷ lệ (bounces: số lần thoát trang; newVisits: số lượt truy cập mới; transactions: số giao dịch; transactionRevenue: doanh thu từ giao dịch) thay các giá trị NaN bằng 0 (tức là không có hoạt động nào xảy ra).
- Trường dữ liệu time_on_site xử lý bằng cách điền giá trị trung vị, bởi: trung vị ít bị ảnh hưởng bởi giá trị ngoại lai, và giữ được phân phối dữ liệu ổn định hơn.

➢ Kết quả:

Trước khi xử lý

	pageviews	hits	time_on_site	bounces	newVisits	transactions	transactionRevenue	sessionQualityDim
0	1	1	NaN	1.0	1.0	NaN	NaN	1
1	1	1	NaN	1.0	NaN	NaN	NaN	1
2	1	1	NaN	1.0	1.0	NaN	NaN	1
3	1	1	NaN	1.0	1.0	NaN	NaN	1
4	1	1	NaN	1.0	1.0	NaN	NaN	1

Sau khi xử lý

	pageviews	hits	time_on_site	bounces	newVisits	transactions	transactionRevenue	sessionQualityDim
0	1.0	1.0	96.5	1.0	1.0	0.0	0.0	1.0
1	1.0	1.0	96.5	1.0	0.0	0.0	0.0	1.0
2	1.0	1.0	96.5	1.0	1.0	0.0	0.0	1.0
3	1.0	1.0	96.5	1.0	1.0	0.0	0.0	1.0
4	1.0	1.0	96.5	1.0	1.0	0.0	0.0	1.0

4.1.2. Xử lý trường dữ liệu định tính.

- Sử dụng phương pháp mã hóa Label Encoding để chuyển đổi dữ liệu định tính sang dạng số, phương pháp này phù hợp khi mỗi cột là các giá trị phân loại rời rạc (không có thứ tự).
- Kết quả:

Trước khi xử lý

	deviceCategory	browser	operatingSystem	country	city	source	medium	campaign
0	desktop	chrome	windows	greece	unknown	(direct)	unknown	unknown
1	desktop	chrome	windows	india	mumbai	analytics.google.com	referral	unknown
2	desktop	chrome	windows	united kingdom	unknown	analytics.google.com	referral	unknown
3	desktop	firefox	windows	united states	dallas	analytics.google.com	referral	unknown
4	desktop	chrome	windows	united states	unknown	adwords.google.com	referral	unknown

Sau khi xử lý

	deviceCategory	browser	operatingSystem	country	city	source	medium	campaign
0	0	2	7	30	52	0	4	1
1	0	2	7	36	32	2	3	1
2	0	2	7	89	52	2	3	1
3	0	5	7	90	13	2	3	1
4	0	2	7	90	52	1	3	1

4.1.3. Xử lý bảng dữ liệu.

- Sau khi xử lý riêng các trường dữ liệu định lượng và định tính:

```
# Loại bỏ các cột định danh không hữu ích cho mô hình
df = df.drop(columns=["fullVisitorId", "visitId"])

# Chuyển cột date sang dạng datetime
df["date"] = pd.to_datetime(df["date"], format="%Y%m%d")

# Tạo thêm các đặc trưng thời gian
df["day_of_week"] = df["date"].dt.dayofweek   # thứ trong tuần (0=thứ2)
df["month"] = df["date"].dt.month
df["is_weekend"] = df["day_of_week"].isin([5,6]).astype(int)

# Xử lý giá trị thiếu (nếu có)
df = df.fillna(df.median(numeric_only=True))
df = df.drop(columns=["date"])

df.head()
```

➢ Nhận xét:

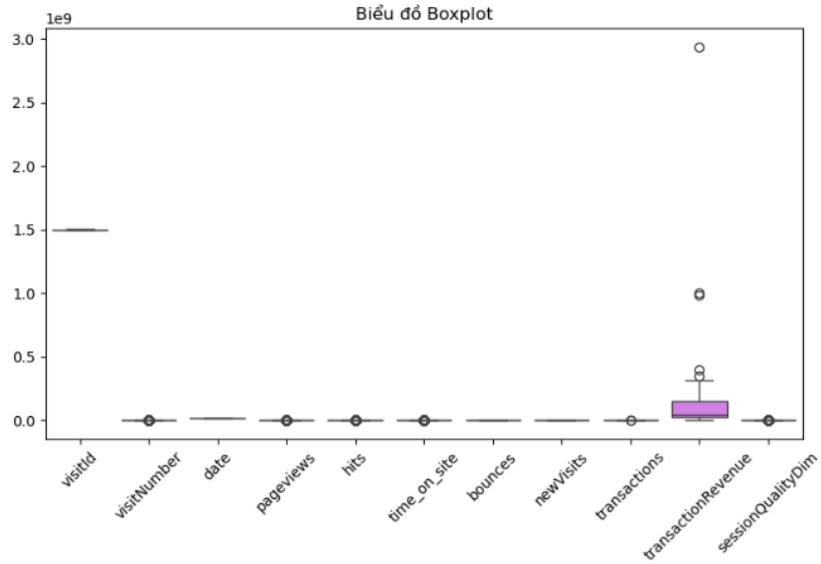
o Xóa các trường định danh duy nhất cho từng phiên truy cập hoặc người dùng: fullVisitorId và visitId. Bởi 2 trường này không mang thông tin hành vi mà chỉ đóng vai trò định danh, nên không có giá trị thống kê cho mô hình.

o Trường Date ở dữ liệu gốc ở dạng chuỗi theo định dạng "yyyymmdd" nên chuyển thành kiểu dữ liệu ngày tháng (datetime) để dễ trích xuất các đặc trưng thời gian.

o Thêm các đặc trưng thời gian giúp mô hình hiểu rõ hơn về xu hướng theo thời gian.

4.2. Xử lý giá trị ngoại lai.

- Biểu đồ Boxplot của tệp dữ liệu:



➢ Nhận xét: Xuất hiện ngoại lai ở trường dữ liệu "transactionRevenue".
- Các biến như *visitNumber*, *pageviews*, *hits*, *time_on_site*, *bounces*, *newVisits*, *transactions* có hộp boxplot rất nhỏ và không xuất hiện nhiều điểm ngoại lai -> phân bố tương đối ổn định.
- Riêng các biến *transactionRevenue* có hộp boxplot cao hơn hẳn, xuất hiện nhiều điểm ngoại lai nằm xa phần thân hộp -> Điều này cho thấy một số phiên truy cập tạo ra doanh thu hoặc số giao dịch cao bất thường, có thể ảnh hưởng mạnh đến trung bình.
- Biến *visitId* và *date* mang tính định danh (mã phiên, ngày truy cập), nên không có ý nghĩa thống kê, chỉ được hiển thị do vẫn thuộc dạng số.

- Sử dụng phương pháp Winsorization để xử lý ngoại lai:

```
# xử lý ngoại lai
import numpy as np

# Xác định các cột định lượng (numeric)
numeric_cols = df.select_dtypes(include=["int64", "float64"]).columns

# Áp dụng winsorization cho từng cột
for col in numeric_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    df[col] = np.clip(df[col], lower_bound, upper_bound)
```

➢ Nhận xét:
- Đối với các trường dữ liệu định lượng, tính các giá trị phân vị Q1 (25%) và Q3 (75%), từ đó xác định khoảng hợp lệ theo công thức:

$$[Q1 - 1{,}5IQR,\ Q3 + 1{,}5IQR]$$

- Các giá trị nằm ngoài khoảng này không bị loại bỏ mà được giới hạn lại (cắt chặn) ở giá trị biên gần nhất.
- Mục đích để giúp giảm ảnh hưởng của các giá trị cực đoan mà không làm mất dữ liệu, đảm bảo tính toàn vẹn của tập dữ liệu cho huấn luyện mô hình.

4.3. Chuẩn hóa dữ liệu.

- Sử dụng phương pháp chuẩn hóa – StandardScaler để chuẩn hóa dữ liệu:

```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import pandas as pd

features = [
    "visitNumber", "pageviews", "hits", "time_on_site", "bounces",
    "transactions", "transactionRevenue", "sessionQualityDim"
]
X = df[features].copy()
# --- Chuẩn hóa dữ liệu ---
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

➢ Nhận xét:
- Các biến đầu vào phục vụ cho mô hình phân cụm (như pageviews, hits, time_on_site, transactionRevenue, …) có đơn vị đo khác nhau và mức độ biến thiên không đồng

nhất. Nếu không chuẩn hóa, các biến có giá trị lớn sẽ chi phối thuật toán K-means, làm cho kết quả phân cụm bị lệch.

- Sử dụng StandardScaler để đưa các biến về cùng thang đo:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

Trong đó: - μ: là giá trị trung bình của biến.

- σ: là độ lệch chuẩn của biến.

- Sau khi chuẩn hóa, tất cả các biến có: Trung bình = 0 và Độ lệch chuẩn = 1, điều này giúp mô hình tính toán khoảng cách chính xác và công bằng hơn.

➢ Kết luận: Sau khi xử lý dữ liệu thu được 1 tệp dữ liệu mới "dataset" hoàn chỉnh để bắt đầu cho giai đoạn huấn luyện mô hình:

	visitNumber	pageviews	hits	time_on_site	bounces	newVisits	transactions	transactionRevenue	sessionQualityDim	deviceCategory	browser	operatingSyste
0	1.0	-0.411779	-0.353582	-0.266785	1.031804	0.604471	-0.128196	-0.049082	-0.293608	0	2	
1	2.0	-0.411779	-0.353582	-0.266785	1.031804	-1.654340	-0.128196	-0.049082	-0.293608	0	2	
2	1.0	-0.411779	-0.353582	-0.266785	1.031804	0.604471	-0.128196	-0.049082	-0.293608	0	2	
3	1.0	-0.411779	-0.353582	-0.266785	1.031804	0.604471	-0.128196	-0.049082	-0.293608	0	2	
4	1.0	-0.411779	-0.353582	-0.266785	1.031804	0.604471	-0.128196	-0.049082	-0.293608	0	2	

ctions	transactionRevenue	sessionQualityDim	deviceCategory	browser	operatingSystem	country	city	source	medium	campaign	day_of_week	month	is_weekend
28196	-0.049082	-0.293608	0	2	7	30	52	0	4	1	1	8	0
28196	-0.049082	-0.293608	0	2	7	36	32	0	4	1	1	8	0
28196	-0.049082	-0.293608	0	2	7	89	52	0	4	1	1	8	0
28196	-0.049082	-0.293608	0	2	7	90	13	0	4	1	1	8	0
28196	-0.049082	-0.293608	0	2	7	90	52	0	4	1	1	8	0

CHƯƠNG 5: HUẤN LUYỆN MÔ HÌNH

5.1. Giới thiệu mô hình học máy K-means.

- **K-Means** là một trong những mô hình học máy không giám sát (unsupervised learning) được sử dụng rộng rãi trong bài toán phân cụm dữ liệu (clustering). Mục tiêu của K-Means là chia tập dữ liệu gồm *n* điểm thành *K* cụm sao cho các điểm dữ liệu trong cùng một cụm có độ tương đồng cao, còn các điểm ở những cụm khác thì khác biệt rõ rệt.
- Thuật toán hoạt động dựa trên khoảng cách (thường là khoảng cách Euclidean) giữa các điểm dữ liệu và các tâm cụm (centroid). Cụ thể, quy trình K-Means gồm các bước sau:
  - Bước 1: Chọn ngẫu nhiên k điểm làm tâm $m_1, m_2, \ldots, m_k$ (các điểm $m_1, m_2, \ldots, m_k$ là các điểm của dữ liệu).
  - Bước 2: Tính khoảng cách từ điểm x_i đến tâm.
  - Bước 3: Với mỗi x_i gán vào nhóm C_k, nếu khoảng cách đến tâm là nhỏ nhất.
  - Bước 4: Cập nhật lại tâm $m_k = \frac{1}{|C_k|} . \sum_{x \in C_k} x.$
  - Bước 5: Lặp lại bước 2 đối với tâm mới.
  - Bước 6: Lặp đến khi không có điểm nào gán lại nhãn -> Dừng thuật toán.
- Hàm mục tiêu: là tối thiểu hóa tổng bình phương khoảng cách giữa các điểm dữ liệu và tâm cụm tương ứng. Hàm mục tiêu được biểu diễn như sau:

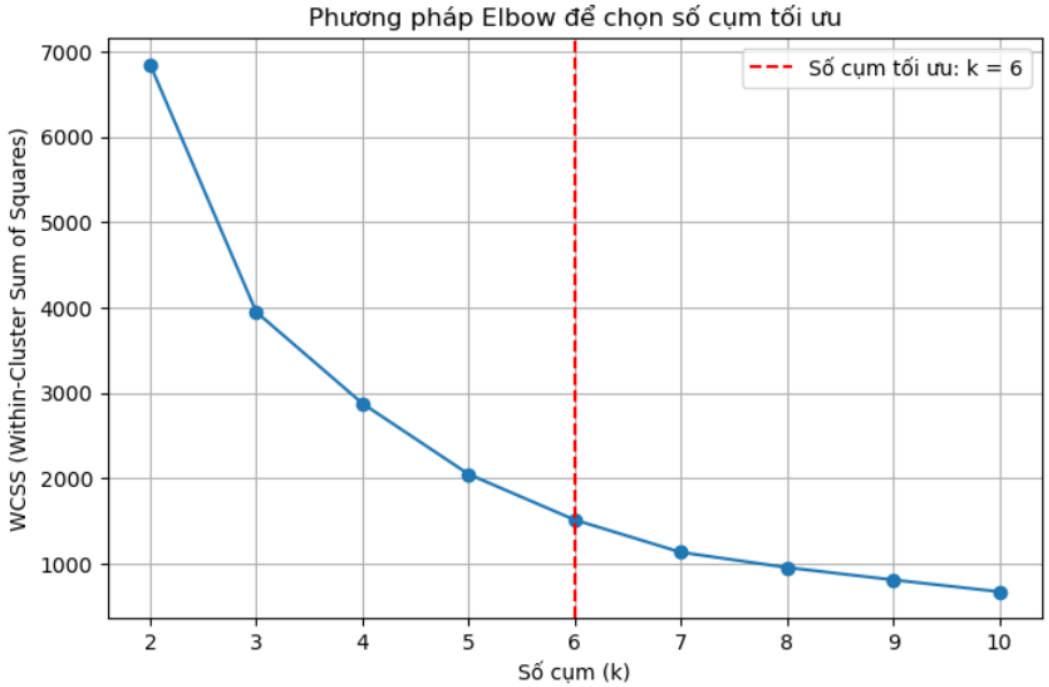
$$E = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2$$

Trong đó:

- K : số cụm
- C_k: tập các điểm thuộc cụm thứ k.
- m_k: tâm cụm thứ k.
- $\|x - m_k\|^2$: bình phương khoảng cách giữa điểm dữ liệu x đến tâm cụm m_k

5.2. Xác định số cụm dữ liệu.

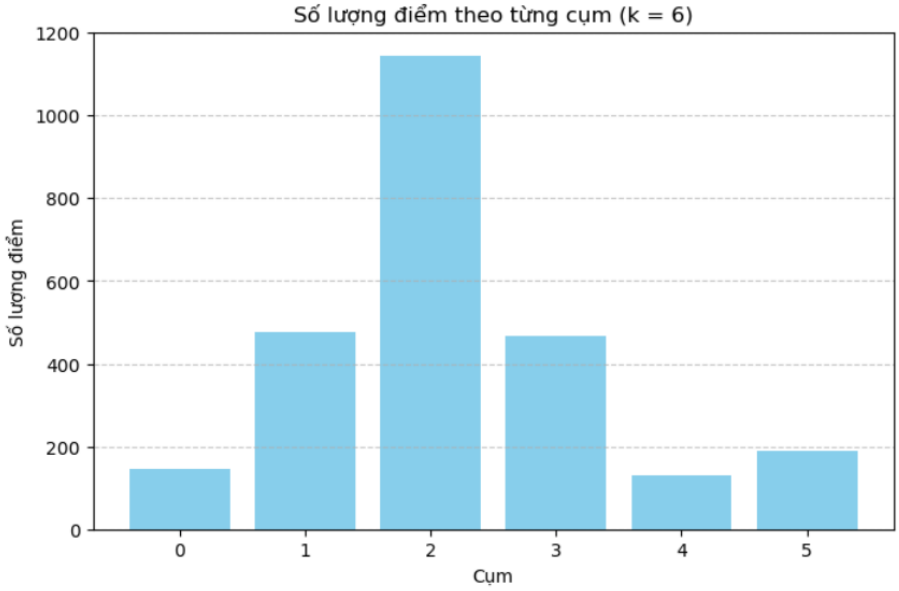
- Sử dụng phương pháp Elbow để xác định số cụm của dữ liệu:



➢ Nhận xét: Điểm "khuỷu tay" là điểm ở **k = 6 cụm** - nơi mà đường cong đột ngột thay đổi độ dốc, nơi mà việc tăng K không còn giúp giảm WCSS một cách đáng kể nữa. Gía trị K tương ứng với điểm khuỷu tay này được xem là số lượng cụm tối ưu.

5.3. Phân cụm dữ liệu.

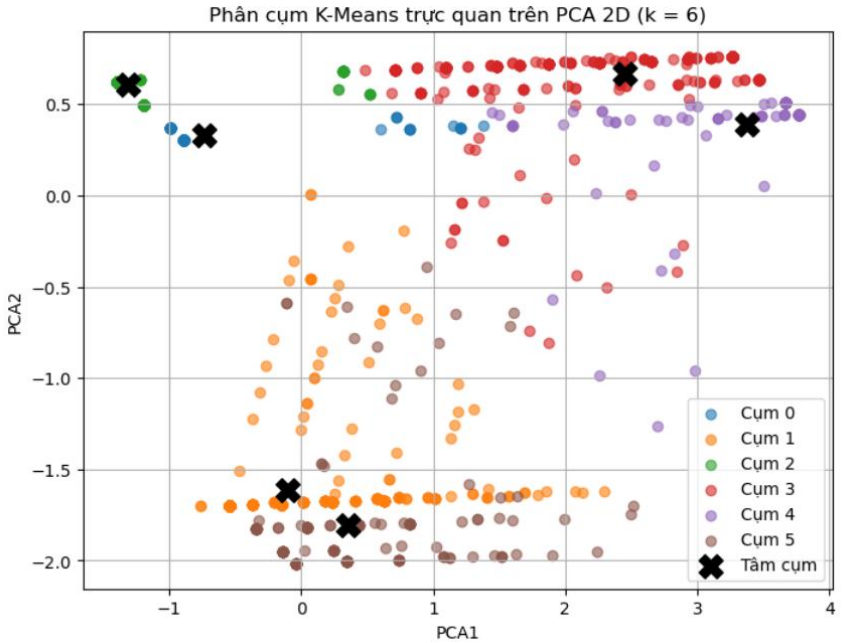
- Chia tệp dữ liệu thành k = 6 cụm, ta được:



➢ Số lượng từng cụm:

Cụm 0: 147 dữ liệu (5.75%); Cụm 3: 468 dữ liệu (18.31%);
Cụm 1: 478 dữ liệu (18.7%); Cụm 4: 131 dữ liệu (5.13%);
Cụm 2: 1144 dữ liệu (44.76%); Cụm 5: 188 dữ liệu (7.36%).

- Biểu đồ phân tán các cụm dữ liệu 2D:



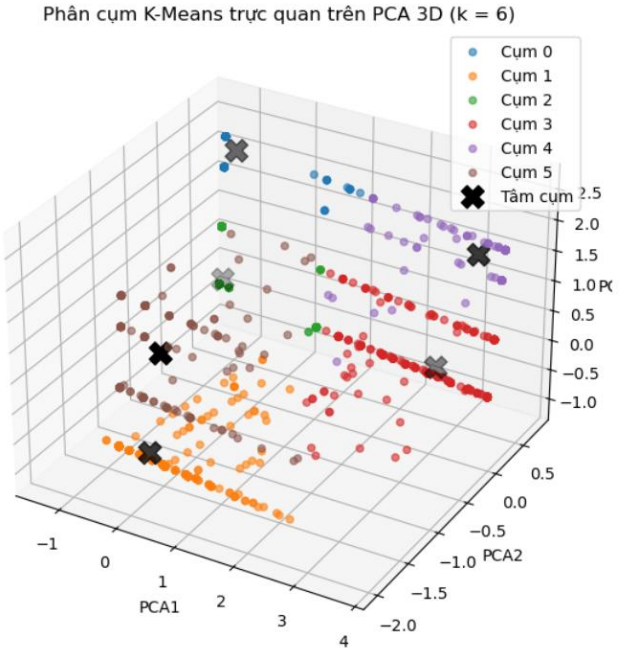
➢ Nhận xét:

- Kết quả trực quan hóa bằng PCA 2D cho thấy 6 cụm được thuật toán K-Means tạo thành những vùng phân tách tương đối rõ ràng trên không gian hai chiều. Các điểm dữ liệu trong cùng một cụm có xu hướng tập trung gần nhau phản ánh sự tương đồng về hành vi giữa các phiên truy cập. Ngược lại, các cụm khác nhau thường nằm ở các khu vực cách biệt nhau, thể hiện sự khác biệt rõ ràng về đặc trưng người dùng giữa các nhóm.
- Các tâm cụm - được biểu diễn bằng ký hiệu **X màu đen** - đều nằm gần trung tâm của từng vùng điểm dữ liệu, cho thấy mô hình K-Means đã hội tụ tốt và xác định được vị

trí trung bình đại diện cho từng cụm. Việc các tâm cụm nằm đúng vào vùng mật độ cao của dữ liệu là dấu hiệu mô hình đang hoạt động chính xác.

- Biểu đồ PCA cũng giúp giảm chiều dữ liệu từ nhiều đặc trưng xuống còn hai thành phần chính, nhưng vẫn giữ lại phần lớn phương sai quan trọng. Nhờ đó, việc quan sát trực quan giúp kiểm chứng rằng cấu trúc phân cụm là hợp lý: các cụm tương đối tách biệt nhau, không bị chồng lấn quá nhiều và phù hợp với số cụm đã chọn từ phương pháp Elbow.

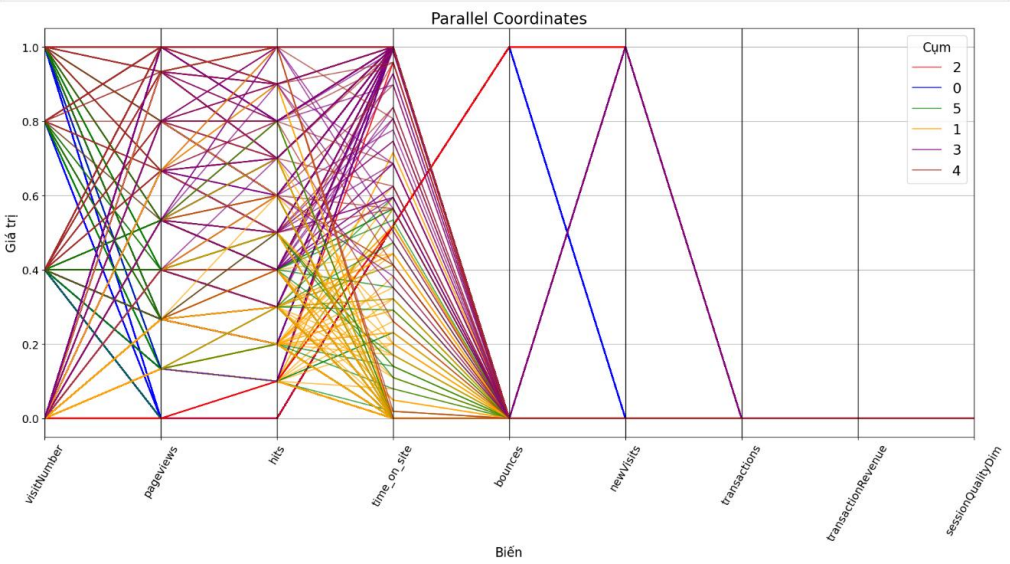
- Biểu đồ phân cụm K-Means trên không gian PCA 3D:



➢ Nhận xét:

- PCA giúp loại bỏ nhiễu và các biến ít ảnh hưởng. Biểu diễn 3D cho cái nhìn đầy đủ hơn 2D trong các trường hợp cụm có sự phân tách phức tạp.
- Biểu đồ 3D cho thấy ranh giới giữa các cụm (cluster) rõ ràng hơn, do có thêm một chiều không gian để phân tán dữ liệu. Một số cụm như cụm 3, cụm 4 và cụm 5 thể hiện sự phân tách rất tốt, nằm ở các khu vực riêng biệt mà khó thấy đầy đủ trong PCA 2D.
- Các điểm dữ liệu của mỗi cụm tạo thành những mảng màu riêng biệt, phản ánh sự tương đồng giữa các phiên truy cập thuộc cùng một nhóm hành vi.

- Biểu đồ Tọa độ Song song (Parallel Coordinates):



- Nhận xét:
  - Biểu đồ Tọa độ Song song (Parallel Coordinates) được sử dụng để phân tích sâu các đặc trưng của từng cụm sau khi phân cụm K-Means, cho thấy mỗi cụm có xu hướng giá trị riêng biệt trên các biến hành vi của người dùng.
  - Một số cụm có giá trị rất cao ở các biến *transactions* và *transactionRevenue*, thể hiện nhóm khách hàng mang lại doanh thu lớn. Một số cụm khác có *pageviews*, *hits* và *time_on_site* cao nhưng doanh thu bằng 0, thể hiện nhóm khách hàng quan tâm nhưng chưa chuyển đổi. Tuy nhiên, có những cụm có *bounces* cao và *time_on_site* thấp, cho thấy đây là nhóm truy cập không tiềm năng.
  - Tổng thể, biểu đồ xác nhận rằng mô hình K-Means đã phân chia dữ liệu thành các nhóm khác biệt rõ ràng theo nhiều chiều thông tin, và hỗ trợ rất tốt cho việc hiểu và mô tả hành vi từng nhóm khách hàng.

5.4. Phân tích các cụm dữ liệu.

Sau khi áp dụng K-Means với k = 6, mô hình đã tạo ra 6 nhóm khách hàng khác nhau dựa trên hành vi truy cập và hành vi mua hàng trên website. Dưới đây là đặc điểm chi tiết của từng cụm:

Cụm	Đặc điểm hành vi	Chiến lược đề xuất
Cụm 0 Tương tác cao, ít mua hàng	- Truy cập website nhiều, xem nhiều trang, thời gian ở lại lâu, nhưng gần như không mua hàng. Nhóm "đang cân nhắc". - Chiếm ~5.75%.	- Remarketing, ưu đãi lần đầu, nhắc giỏ hàng, cung cấp thêm thông tin sản phẩm để thúc đẩy chuyển đổi.
Cụm 1 Khách hàng trung bình	- Tương tác và mua hàng ở mức trung bình, ổn định nhưng không nổi bật. Hành vi mua thỉnh thoảng. - Chiếm ~18.7%.	- Giữ chân bằng ưu đãi định kỳ, cải thiện trải nghiệm người dùng, cá nhân hóa đề xuất sản phẩm.
Cụm 2 Khách hàng phổ thông	- Nhóm lớn nhất, tương tác và mua hàng ở mức trung tính. Đóng vai trò quan trọng trong tổng lượng truy cập và doanh thu. - Chiếm tỷ lệ lớn nhất (~44.76%).	- Xây dựng chương trình khách hàng thân thiết cơ bản, thực hiện upsell để tăng giá trị giao dịch.
Cụm 3 Mua hàng tích cực, mức chi tiêu khá	- Mua hàng tích cực, mức chi tiêu khá, số lượng giao dịch/doanh thu cao hơn trung bình, tương tác vừa phải. - Nhóm khách hàng "chốt đơn" tốt (~18.31%).	- Ưu tiên giữ chân nhóm khách này, giới thiệu sản phẩm mới/gói combo, cải thiện chăm sóc khách hàng.
Cụm 4 Ít tương tác, ít mua, có nguy cơ rời bỏ	- Ít tương tác, ít mua, dễ rời bỏ website, số lần truy cập và thời gian trên site thấp. - Chiếm ~5.13%.	- Chiến dịch kích hoạt lại, gửi email/quảng cáo nhắc nhở, ưu đãi mạnh, nội dung ngắn và hấp dẫn.
Cụm 5 VIP / Trung thành	- Giá trị cao, trung thành, doanh thu và số giao dịch cao, tương tác tốt, nhóm VIP. - Nhóm giá trị cao (~7.36%).	- Chăm sóc đặc biệt, hỗ trợ nhanh, chương trình thành viên VIP, ưu đãi cá nhân hóa, quà tri ân/sự kiện riêng.

5.5. Đánh giá mô hình.

5.5.1. Đánh giá theo Inertia.

- Tính tổng bình phương khoảng cách từ các điểm đến tâm cụm của chúng với công thức:

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2$$

Trong đó:

- K = 6 là số cụm
- C_k: tập các điểm thuộc cụm thứ k.
- m_k: tâm cụm thứ k.
- $\|x - m_k\|^2$: bình phương khoảng cách giữa điểm dữ liệu x đến tâm cụm m_k

➢ Kết quả: **WCSS ≈ 669.451**, cho thấy các cụm tách biệt rõ ràng và kết dính tốt, tức là phân cụm đạt chất lượng khá tốt.

5.5.2. Đánh giá theo Silhouette Score.

- Đo độ tách biệt giữa các cụm và kết dính bên trong cụm với công thức cho 1 điểm i:

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$$

Trong đó:

- a(i): khoảng cách trung bình từ điểm i đến tất cả các điểm khác trong cùng cụm (kết dính).
- b(i): khoảng cách trung bình từ điểm i đến tất cả các điểm trong cụm gần nhất khác (tách biệt).

➢ Kết quả: **Silhouette Score = 0.671,** cho thấy các cụm tách biệt rõ ràng và các điểm gần tâm cụm của mình, phân cụm đạt chất lượng khá tốt.

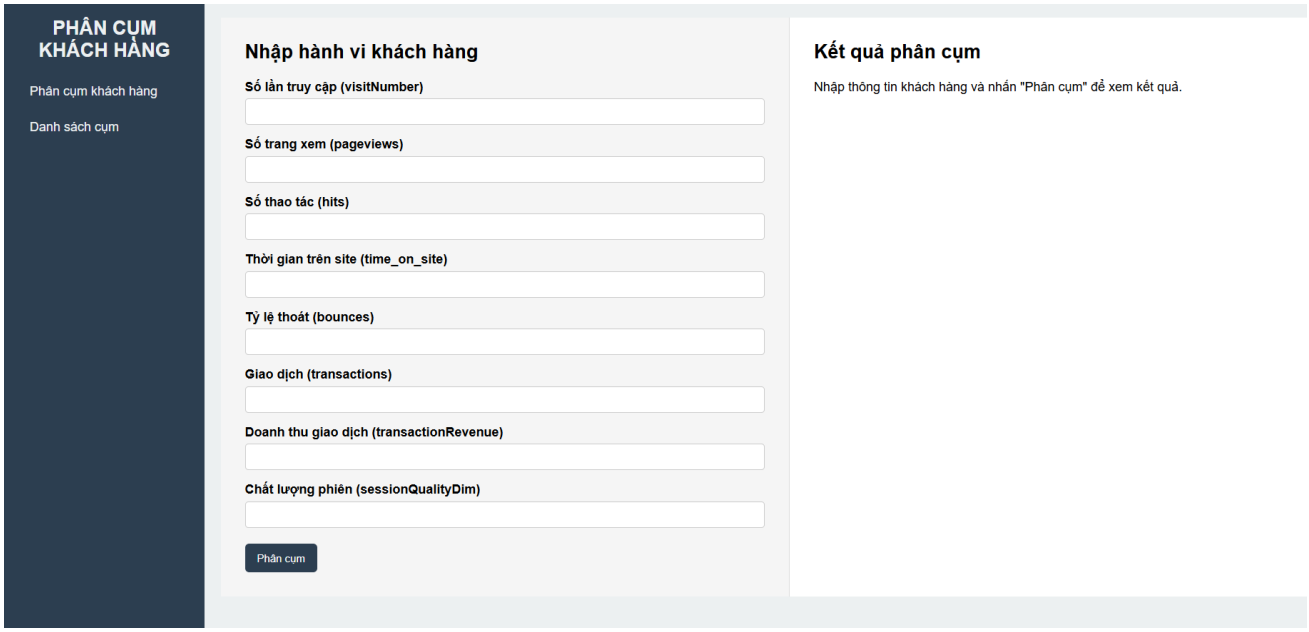
CHƯƠNG 6: TRIỂN KHAI MÔ HÌNH

6.1. Mục tiêu triển khai.

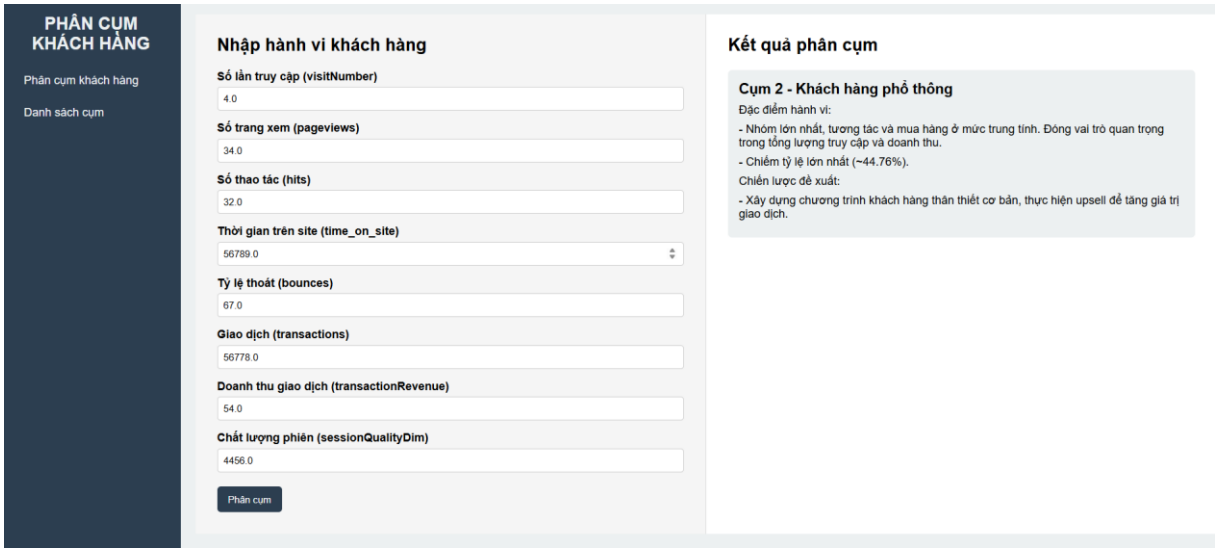
- Mục tiêu:
  - Sử dụng mô hình K-Means đã huấn luyện để phân cụm khách hàng mới dựa trên hành vi truy cập website.
  - Xác định cụm khách hàng giúp hiểu rõ hành vi, nhu cầu và mức độ tương tác của khách hàng.
- Ứng dụng:
  - Phân tích hành vi người dùng trên website.
  - Đề xuất chiến lược marketing phù hợp cho từng nhóm khách hàng.

6.2. Triển khai mô hình.

- Tải mô hình đã huấn luyện: **kmeans.pkl**
- Load file lên 1 trang localhost để nhập dữ liệu dự đoán cụm của khách hàng mới:
  - **Input:** Nhập các hành vi của khách hàng mới:



  - **Output:** khách hàng mới đó thuộc cụm nào.



KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận.

Qua quá trình thực hiện đề tài **"Phân tích hành vi khách hàng trong thương mại điện tử"**, có thể rút ra những kết luận chính sau:

- **Hiểu rõ hành vi khách hàng là yếu tố quan trọng:**
  - Khách hàng có hành vi rất đa dạng: từ việc truy cập, xem sản phẩm, thêm vào giỏ hàng, đến thực hiện giao dịch.
  - Việc phân loại khách hàng giúp doanh nghiệp cá nhân hóa trải nghiệm, tối ưu chiến lược marketing và nâng cao doanh thu.
- **Phân cụm khách hàng hiệu quả:**
  - Áp dụng mô hình K-Means đã giúp nhóm khách hàng thành các cụm có đặc điểm hành vi tương đồng.
  - Kết quả phân cụm cung cấp cơ sở khoa học để đề xuất chiến lược marketing phù hợp cho từng nhóm khách hàng (remarketing, ưu đãi, loyalty program).
- **Dự đoán hành vi mua hàng:**
  - Mô hình giúp đánh giá khả năng khách hàng mới thực hiện giao dịch.
  - Kết quả dự đoán hỗ trợ doanh nghiệp tập trung chăm sóc khách hàng tiềm năng, nâng cao hiệu quả các chiến dịch marketing.
- **Ứng dụng Big Data và khoa học dữ liệu:**
  - Dữ liệu lớn từ website, app, mạng xã hội có thể được xử lý để phân tích hành vi, dự đoán nhu cầu và triển khai chiến lược thời gian thực.
  - Việc kết hợp phân tích dữ liệu lớn với mô hình học máy tạo ra hệ thống phân tích khách hàng toàn diện, có khả năng mở rộng và áp dụng thực tế.

2. Hướng phát triển.

Dự án không chỉ dừng lại ở phân tích hành vi, phân cụm và dự đoán khả năng mua hàng, mà còn mở rộng sang:

- Thu thập và tích hợp dữ liệu đa nguồn, theo thời gian thực.
- Nâng cao mô hình học máy, áp dụng Big Data.
- Tích hợp trực tiếp vào chiến lược marketing thông minh và hệ thống quản lý khách hàng toàn diện.

➢ Với hướng phát triển này, dự án có thể trở thành một **công cụ phân tích khách hàng mạnh mẽ, hiệu quả và có khả năng ứng dụng rộng rãi** trong thương mại điện tử và các ngành khác.

TÀI LIỆU THAM KHẢO

[1] e-Conomy SEA Report 2024. *Google, Temasek, Bain & Company*.
Link: https://economysea.withgoogle.com

[2] "bigquery-public-data.imdb," *Google Cloud*.

[4] PGS.TS.Trần Văn Long, "Bài giảng Dữ liệu lớn".

[5] phamdinhkhanh, "Đánh giá mô hình phân loại trong ML," *https://phamdinhkhanh.github.io/*, 2020.