# TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN





BÁO CÁO TIỂU LUẬN PHÂN TÍCH DỮ LIỆU

**BITLY DATA FROM 1.USA.GOV**

**Giảng viên hướng dẫn: Đỗ Như Tài**

**Thông tin thành viên:**

Lê Nguyễn Quốc Hưng 3121410240

Trần Bá Tiên 3122410410

Dương Huỳnh Gia 3121410167

Bùi Thành Công 3122410042

***TP. Hồ Chí Minh, tháng 10/2025***

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..……………………………………….……………………..…………………………………..

# LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Đại Học Sài Gòn, đặc biệt là các thầy cô ở khoa Công Nghệ Thông Tin đã tạo điều kiện cho em tiếp cận và tìm hiểu để hoàn thành đồ án môn học lần này.

Và chúng em cũng xin chân thành đặc biệt cảm ơn thầy Đỗ Như Tài, thầy là giảng viên giảng dạy đã nhiệt tình hướng dẫn giúp chúng em hoàn thành được đồ án lần này.

Trong quá trình nghiên cứu và làm bài báo cáo đồ án, kiến thức cũng như kinh nghiệm thực tế còn nhiều hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót, chúng em rất mong nhận được ý kiến đóng góp của thầy, cô để em học hỏi được nhiều kỹ năng, kinh nghiệm và sẽ hoàn thành tốt hơn cho những bài báo cáo sắp tới.

Chúng em xin chân thành cảm ơn.

# MỤC LỤC

[LỜI CẢM ƠN 1](#_heading=h.30j0zll)

[MỤC LỤC 2](#_heading=h.os4kp4kuwm50)

[I. GIỚI THIỆU 3](#_heading=h.ua56i0g34v20)

[1. Giới thiệu về rượu vang đỏ (Red Wine) 3](#_heading=h.66s6y96bvof2)

[2. Giới thiệu tập dữ liệu 3](#_heading=h.m9i74nuvfbr4)

[2.1. Nguồn dữ liệu 3](#_heading=h.kf2cn9b51qbc)

[2.2. Giải thích các biến 3](#_heading=h.ou6kxl3s9gxr)

[II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM) 6](#_heading=h.2mocjcycxqnx)

[1. Nạp dữ liệu 7](#_heading=h.a7gvja14xqbd)

[2. Loại bỏ dữ liệu trùng lặp 7](#_heading=h.46cxqg7s07cq)

[3. Thay thế dữ liệu và thay đổi định dạng 7](#_heading=h.7fnzmansmfk2)

[4. Xử lý dữ liệu thiếu 8](#_heading=h.lpq3jgp4xo2z)

[5. Thống kê mô tả 8](#_heading=h.5rcr39u2d9pa)

[6. Thống kê theo nhóm chất lượng 9](#_heading=h.6ldclf76r6u4)

[7. Trực quan hóa 10](#_heading=h.fg3nwvxz51h1)

[III. KẾT LUẬN 10](#_heading=h.2srdqcqlhx5)

[TÀI LIỆU THAM KHẢO 12](#_heading=h.bd6z3290hgfj)

# I. GIỚI THIỆU

## 1. Giới thiệu về dữ liệu Bitly Data from 1.USA.gov

Dữ liệu Bitly Data from 1.USA.gov là tập hợp các bản ghi được thu thập vào năm 2011, ghi lại hoạt động truy cập của người dùng Internet vào các liên kết rút gọn (*shortened URLs*) mang đuôi *.gov* hoặc *.mil* — tức là các trang web chính phủ và quân đội Hoa Kỳ. Bộ dữ liệu được tạo ra thông qua sự hợp tác giữa Bitly, dịch vụ rút gọn liên kết phổ biến, và USA.gov, cổng thông tin chính thức của chính phủ Mỹ.

Mỗi dòng trong tập dữ liệu đại diện cho một lượt truy cập (click) của người dùng, kèm theo nhiều thông tin liên quan như: múi giờ truy cập (tz), quốc gia (c), chuỗi trình duyệt và hệ điều hành (a), nguồn dẫn truy cập (*referrer* – r), thời điểm click (t), và vị trí địa lý (cy, gr, ll). Tất cả các bản ghi đều được lưu trữ ở định dạng JSON (*JavaScript Object Notation*), một cấu trúc phổ biến cho dữ liệu web phi cấu trúc.

Quá trình xử lý và phân tích bộ dữ liệu này cho phép khai thác các đặc trưng hành vi người dùng, chẳng hạn như:

* Thời điểm và tần suất truy cập vào các trang *.gov*, *.mil*;
* Phân bố người dùng theo khu vực địa lý (múi giờ, quốc gia, thành phố);
* Tỷ lệ sử dụng các hệ điều hành phổ biến như *Windows*, *Mac OS*, *Android*;
* Các nguồn truy cập chính (*referrers*) như *Facebook*, *Google*, *Twitter*.

Việc phân tích dữ liệu Bitly–USA.gov không chỉ mang lại cái nhìn định lượng về hành vi trực tuyến của người dùng Internet đối với các trang web chính phủ, mà còn là ví dụ điển hình trong lĩnh vực phân tích dữ liệu thực tế (real-world data analysis) – nơi dữ liệu thô ở định dạng JSON được xử lý, làm sạch và trực quan hóa bằng ngôn ngữ Python cùng các thư viện như *pandas*, *matplotlib* và *seaborn*.

## 2. Giới thiệu tập dữ liệu

### 2.1. Nguồn dữ liệu

* **Nguồn:** Bộ dữ liệu *Bitly Data from 1.USA.gov* được cung cấp trong ví dụ thực hành của cuốn *Python for Data Analysis* (Wes McKinney, O’Reilly, 2022).
* **Đơn vị thu thập:** USA.gov (cổng thông tin chính phủ Hoa Kỳ) kết hợp với dịch vụ rút gọn liên kết **Bitly**.
* **Định dạng tệp:** JSON (mỗi dòng là một bản ghi độc lập).
* **Số dòng (records):** Khoảng 3.560 mẫu (bản ghi truy cập).
* **Số biến (features):** 18 trường dữ liệu (biến quan sát).
* **Đặc trưng dữ liệu:** Dữ liệu phi cấu trúc, gồm các thông tin hành vi truy cập, vị trí, thiết bị và thời gian của người dùng Internet khi click vào liên kết *.gov* hoặc *.mil*.

### 2.2. Giải thích các biến

* **tz (Time Zone – Múi giờ):**  
  Múi giờ của người dùng tại thời điểm click vào liên kết, giúp xác định phân bố thời gian và khu vực hoạt động.  
  *Ví dụ:* “America/New\_York”, “Europe/London”.
* **c (Country – Quốc gia):**  
  Mã quốc gia (2 ký tự) của người dùng, theo chuẩn ISO.  
  *Ví dụ:* “US” (Hoa Kỳ), “BR” (Brazil), “PL” (Ba Lan).
* **gr (State – Bang):**  
  Mã bang của người dùng nếu ở trong nước Mỹ.  
  *Ví dụ:* “CA” (California), “NY” (New York).
* **cy (City – Thành phố):**  
  Thành phố mà lượt truy cập được ghi nhận.  
  *Ví dụ:* “Los Angeles”, “New York”, “Chicago”.
* **ll (Latitude/Longitude – Vĩ độ và Kinh độ):**  
  Tọa độ vị trí địa lý của người dùng, được biểu diễn dưới dạng cặp số [vĩ độ, kinh độ].
* **a (User Agent – Thông tin thiết bị / trình duyệt):**  
  Chuỗi mô tả hệ điều hành, trình duyệt, hoặc ứng dụng mà người dùng sử dụng để mở liên kết.  
  *Ví dụ:* “Mozilla/5.0 (Windows NT 6.1; WOW64)…”, “GoogleMaps/RochesterNY”.
* **r (Referrer – Nguồn truy cập):**  
  Trang web hoặc nền tảng mà người dùng click vào liên kết Bitly.  
  *Ví dụ:* “facebook.com”, “google.com”, “twitter.com”.
* **u (URL – Đích đến):**  
  Đường dẫn gốc của trang web *.gov* hoặc *.mil* được rút gọn.  
  *Ví dụ:* “http://www.ncbi.nlm.nih.gov/pubmed/22415991”.
* **t (Timestamp – Thời điểm click):**  
  Thời gian (đơn vị giây kể từ 01/01/1970) ghi lại khi người dùng click vào liên kết.
* **hc (Link creation time – Thời điểm tạo liên kết):**  
  Thời điểm mà đường dẫn Bitly được tạo ra (cũng tính bằng timestamp).  
  Có thể dùng để tính độ trễ giữa thời điểm tạo link và thời điểm người dùng click.
* **hh (Host – Máy chủ đích):**  
  Tên miền chính của liên kết, thường là “1.usa.gov”.
* **l (Login – Tài khoản tạo link):**  
  Tên tài khoản Bitly (đã ẩn danh) tạo ra đường dẫn rút gọn.
* **g/h (Global hash / User hash – Mã định danh link):**  
  Hai giá trị mã hóa duy nhất dùng để phân biệt các liên kết Bitly, trong đó “g” là mã toàn cục, “h” là mã người dùng.
* **al (Accept Language – Ngôn ngữ trình duyệt):**  
  Ngôn ngữ hiển thị ưa thích mà trình duyệt người dùng đang sử dụng.  
  *Ví dụ:* “en-US,en;q=0.8”.
* **nk (Unknown flag – Cờ dữ liệu hệ thống):**  
  Biến kỹ thuật, thường có giá trị 0 hoặc 1, không ảnh hưởng trực tiếp đến phân tích hành vi.
* ***heartbeat* (Heartbeat – Kiểm tra trạng thái):**  
  Trường kỹ thuật dùng trong giám sát luồng dữ liệu, thường chỉ xuất hiện ngẫu nhiên trong vài bản ghi.
* **kw (Keywords – Từ khóa):**  
  Các từ khóa hoặc nhãn gắn kèm với đường dẫn Bitly (ít xuất hiện).
* **trangthai / gr / cy (Các biến phụ trợ khác – nếu có):**  
  Một số bản ghi có thêm biến trạng thái hoặc vị trí chi tiết hơn, được dùng cho mục đích minh họa trong quá trình phân tích.

# II. ĐỌC VÀ KHÁM PHÁ DỮ LIỆU

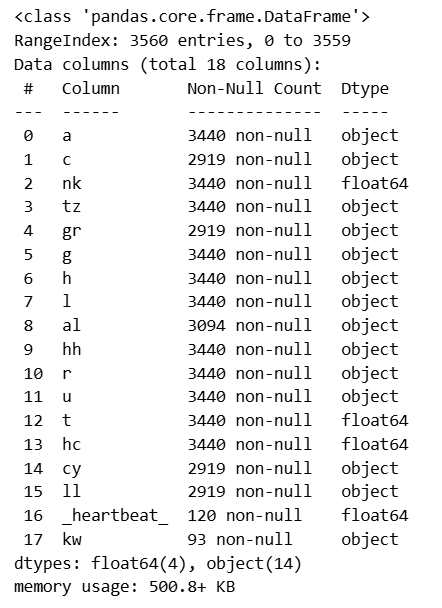
## 1. Kiểm tra số bản ghi và 1 mẫu

Tổng số bản ghi: 3560

1 mẫu bản ghi:

{'a': 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/535.11 (KHTML, like Gecko) Chrome/17.0.963.78 Safari/535.11', 'c': 'US', 'nk': 1, 'tz': 'America/New\_York', 'gr': 'MA', 'g': 'A6qOVH', 'h': 'wfLQtf', 'l': 'orofrog', 'al': 'en-US,en;q=0.8', 'hh': '1.usa.gov', 'r': '<http://www.facebook.com/l/7AQEFzjSi/1.usa.gov/wfLQtf>', 'u': '<http://www.ncbi.nlm.nih.gov/pubmed/22415991>', 't': 1331923247, 'hc': 1331822918, 'cy': 'Danvers', 'll': [42.576698, -70.954903]}

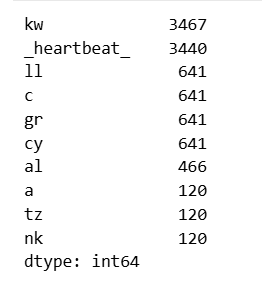
## 2. Kiểm tra thông tin tổng quát



## 3. Kiểm tra tên các cột

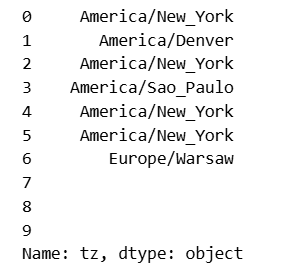


## 4. Kiểm tra dữ liệu thiếu NaN

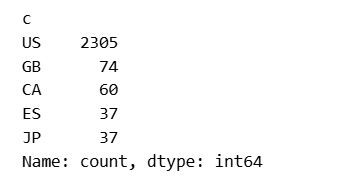


## 5. Xem vài giá trị đặc trưng

- Cột múi giờ



- Cột quốc gia

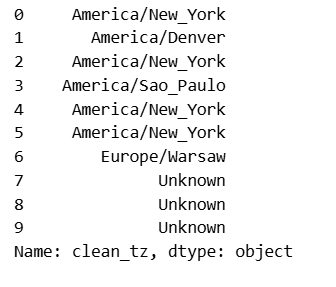


- Cột referrer



# III. LÀM SẠCH DỮ LIỆU

## 1. Làm sạch cột tz và xem 10 giá trị đầu



**\*\*Kết quả:\*\***

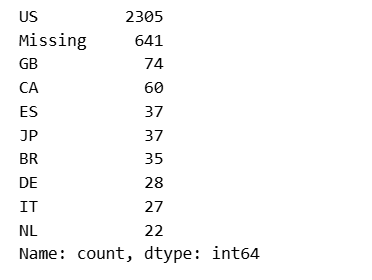
  + Các bản ghi giờ sẽ có 3 loại:

      + “America/New\_York” (có thật)

      + “Unknown” (chuỗi trống)

      + “Missing” (không có trường tz)

## 2. Làm sạch cột c và kiểm tra top quốc gia

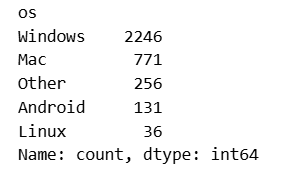


Thông thường US chiếm đa số, các quốc gia khác chỉ chiếm tỉ lệ nhỏ.

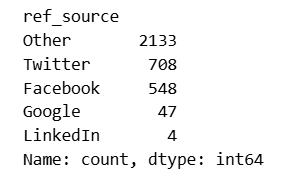
## 3. Làm sạch cột r và xem 5 dòng đầu



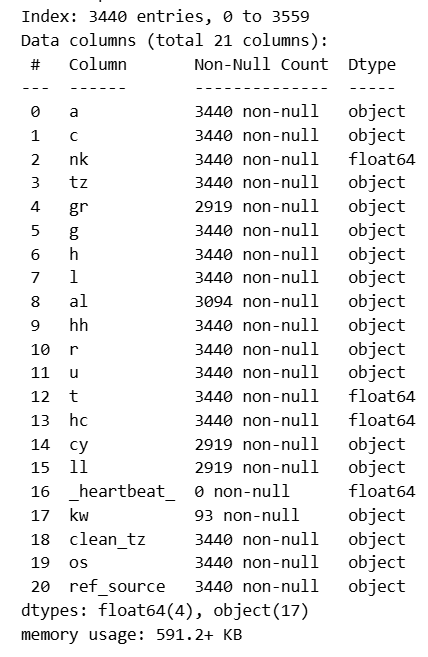
## 4. Tạo cột os và xác định hệ điều hành



## 5. Tạo cột ref-source xác định nguồn truy cập chính

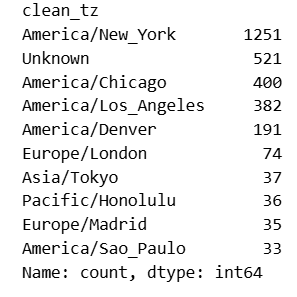


## 6. Kiểm tra dữ liệu sau khi làm sạch

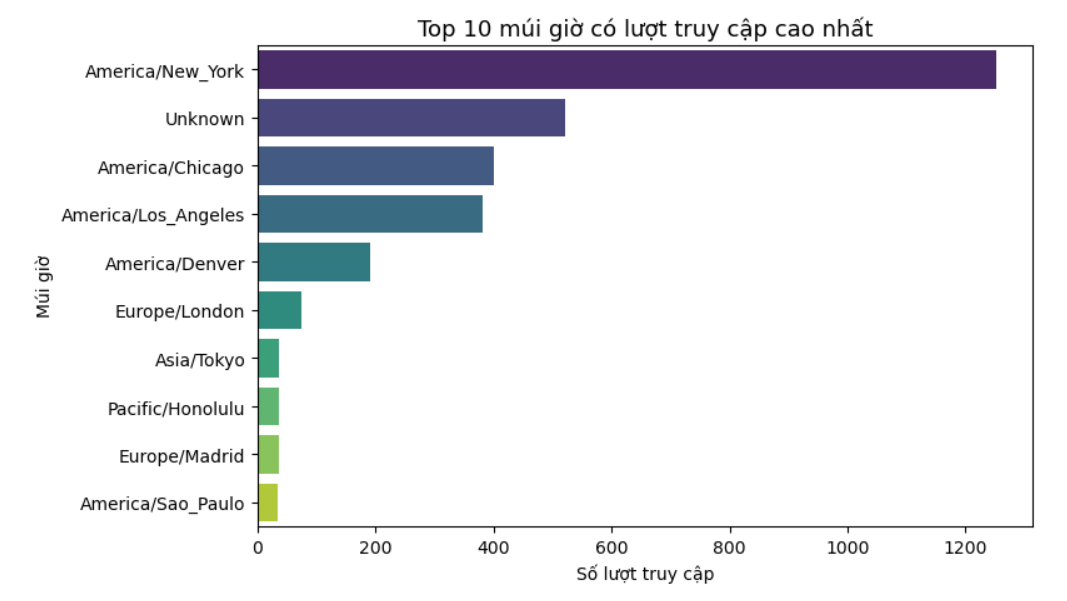


# IV. PHÂN TÍCH MÔ TẢ

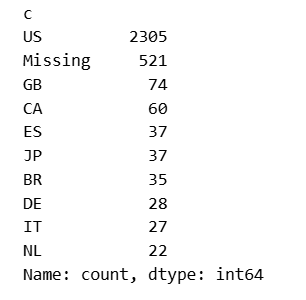
## 1. Phân bố người dùng theo múi giờ



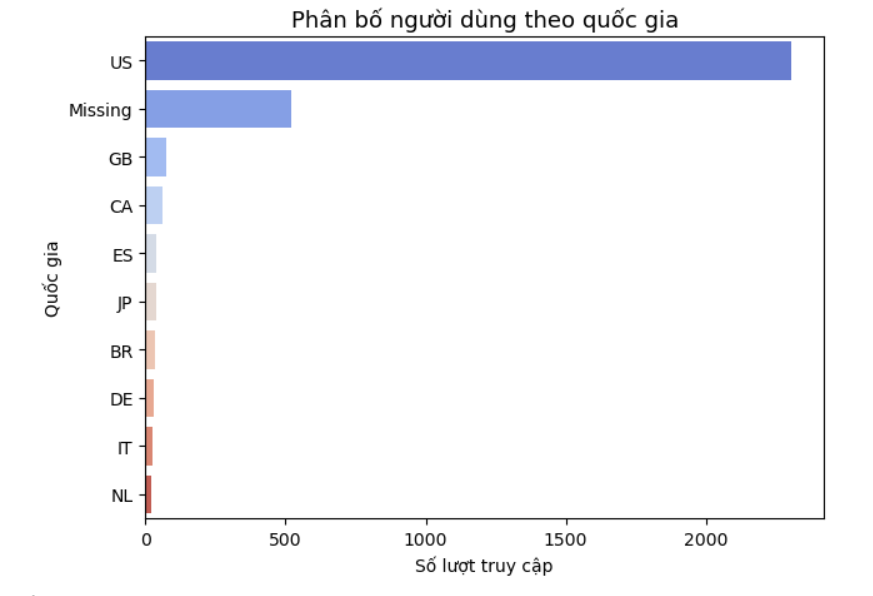
## 2. Biểu đồ top 10 múi giờ



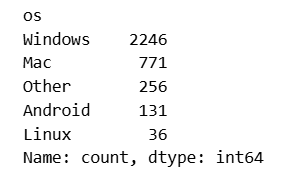
## 3. Phân bố người dùng theo quốc gia



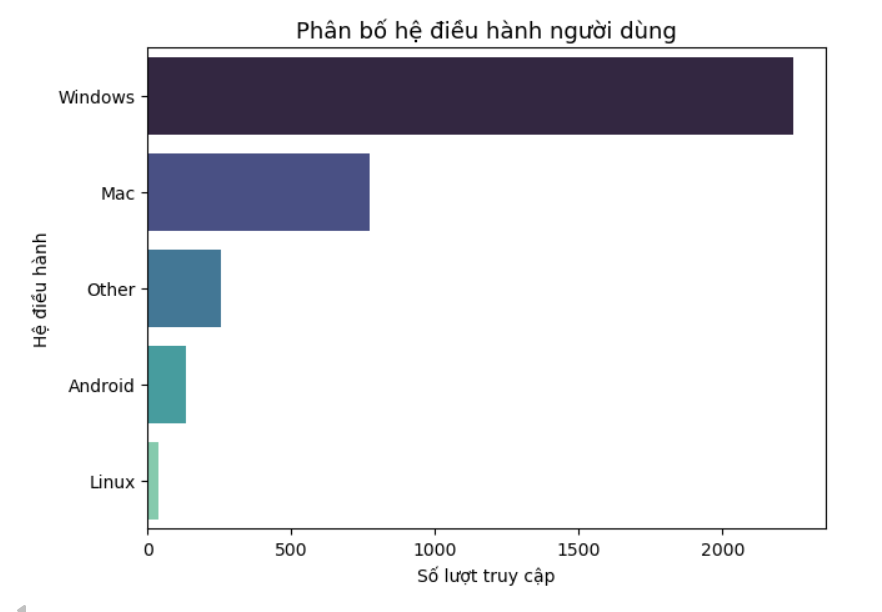
## 4. Biểu đồ phân bố theo quốc gia



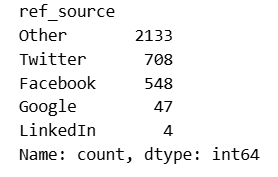
## 5. Phân bố người dùng theo hệ điều hành



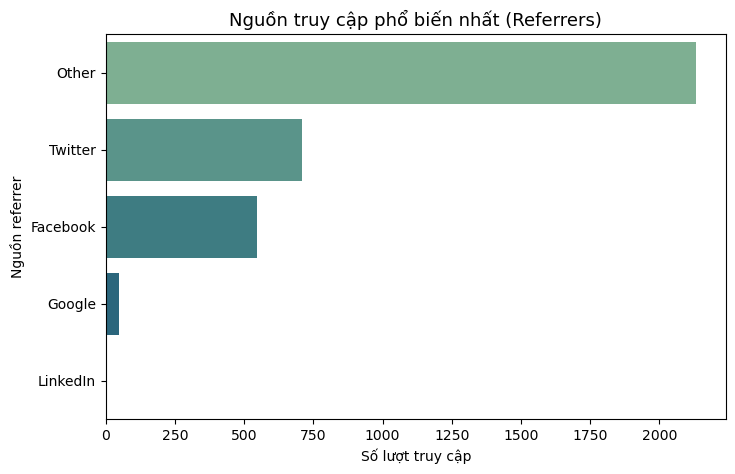
## 6. Biểu đồ hệ điều hành



## 7. Phân bố nguồn truy cập

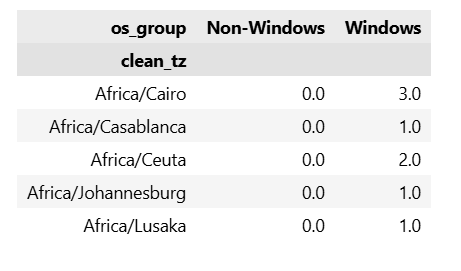


## 8. Biểu đồ nguồn truy cập

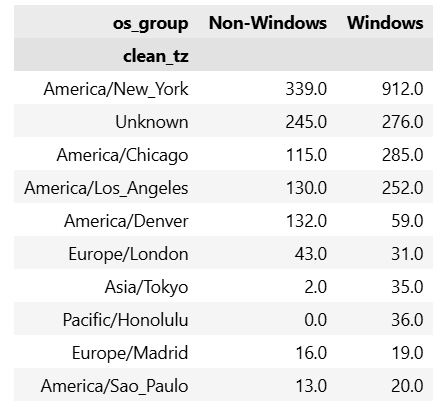


# V. PHÂN TÍCH TƯƠNG QUAN

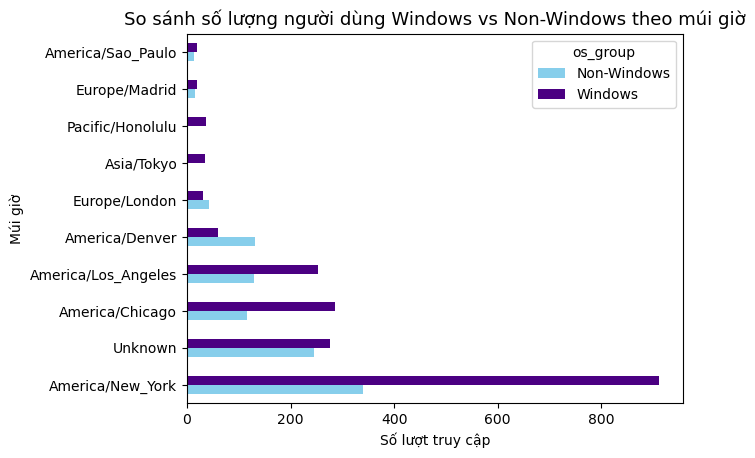
## 1. Phân bố hệ điều hành theo múi giờ



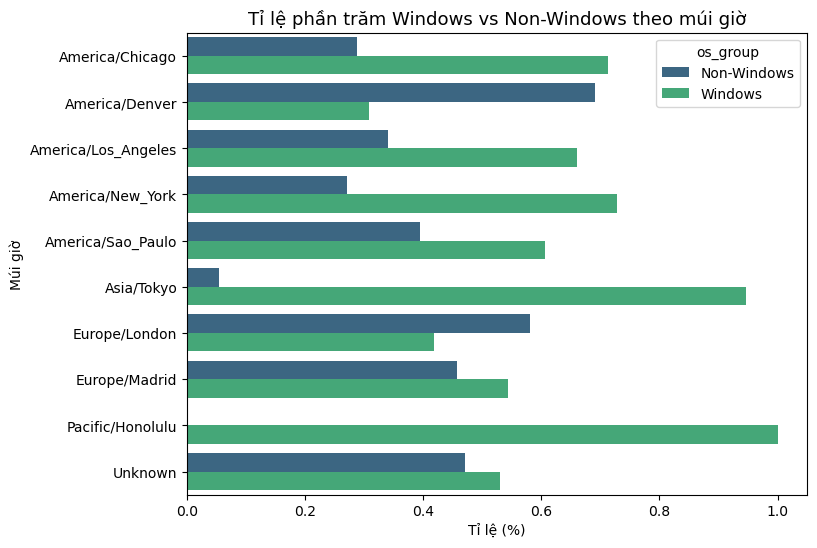
## 2. 10 múi giờ có lượt truy cập cao nhất



- Biểu đồ so sánh window và non-window



* Chuẩn hóa theo tỉ lệ phần trăm



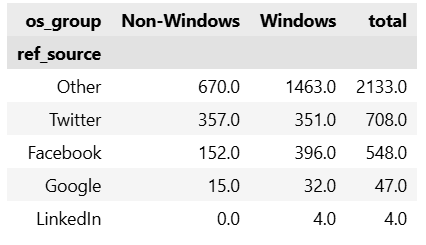
**\*\*Nhận xét:\*\***

+ Ở Mỹ (New York, Chicago): Windows chiếm ~70–80%

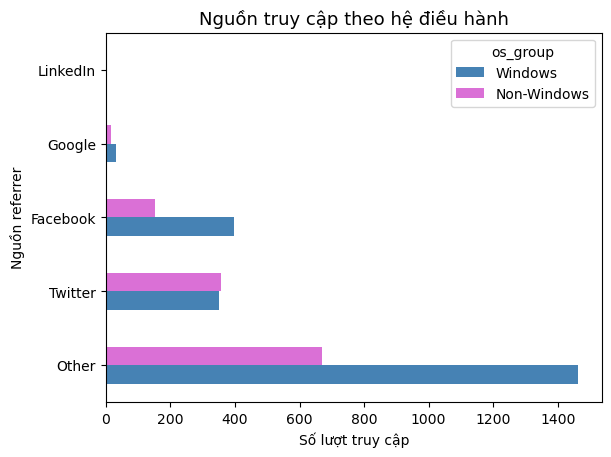
+ Ở châu Âu, tỉ lệ Non-Windows (Mac/Linux) cao hơn

+ Gợi ý: hành vi sử dụng máy tính khác nhau theo khu vực địa lý

## 3. So sánh referrer theo hệ điều hành



* Biểu đồ



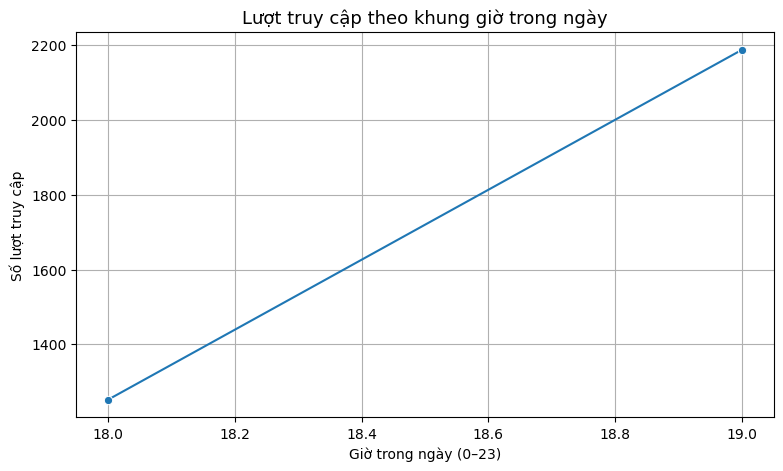
**\*\*Nhận xét:\*\***

+ Facebook và Google chiếm phần lớn nguồn traffic.

+ Windows → nhiều click từ Facebook / Google.

+ Non-Windows → tỉ lệ cao hơn ở Twitter, LinkedIn.

## 4. Phân tích khung giờ hoạt động

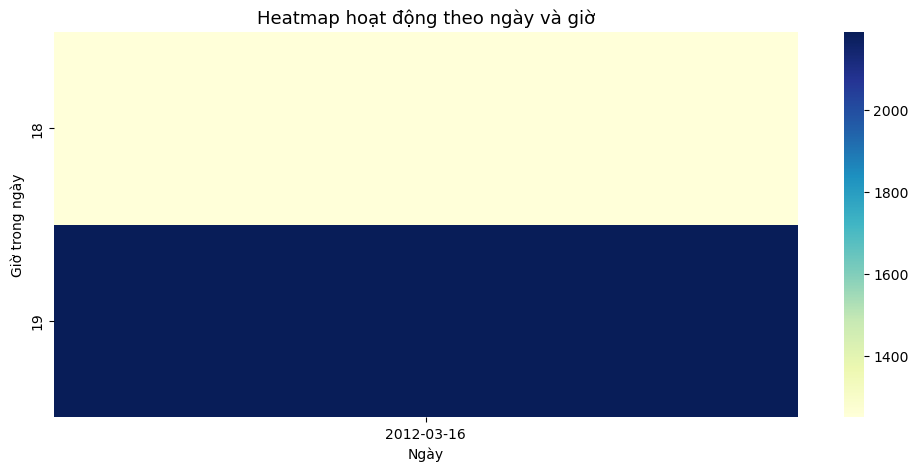


**\*\*Nhận xét:\*\***

+ Lượng truy cập cao vào giờ hành chính (8–17h).

+ Giảm mạnh vào đêm → đúng hành vi người dùng Internet thông thường.

## 5. Headmap hoạt động



# VI. PHÂN TÍCH ĐỊA LÝ

# VII. KẾT LUẬN TỔNG QUAN

# 

# TÀI LIỆU THAM KHẢO

[1] “Red Wine Quality EDA & Classification,” Kaggle, UCI Machine Learning Repository. <https://www.kaggle.com/code/eisgandar/red-wine-quality-eda-classification>.

[Truy cập: 10-10-2025].