# 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](#_Toc212412503)

[MỤC LỤC 2](#_Toc212412504)

[I. GIỚI THIỆU 4](#_Toc212412505)

[1. Giới thiệu về dữ liệu Bitly Data from 1.USA.gov 4](#_Toc212412506)

[2. Giới thiệu tập dữ liệu 5](#_Toc212412507)

[2.1. Nguồn dữ liệu 5](#_Toc212412508)

[2.2. Giải thích các biến 5](#_Toc212412509)

[II. ĐỌC VÀ KHÁM PHÁ DỮ LIỆU 6](#_Toc212412510)

[1. Kiểm tra số bản ghi và 1 mẫu 6](#_Toc212412511)

[2. Kiểm tra thông tin tổng quát 7](#_Toc212412512)

[3. Kiểm tra tên các cột 7](#_Toc212412513)

[4. Kiểm tra dữ liệu thiếu NaN 7](#_Toc212412514)

[5. Xem vài giá trị đặc trưng 8](#_Toc212412515)

[III. LÀM SẠCH DỮ LIỆU 9](#_Toc212412516)

[1. Làm sạch cột tz và xem 10 giá trị đầu 9](#_Toc212412517)

[2. Làm sạch cột c và kiểm tra top quốc gia 9](#_Toc212412518)

[3. Làm sạch cột r và xem 5 dòng đầu 9](#_Toc212412519)

[4. Tạo cột os và xác định hệ điều hành 10](#_Toc212412520)

[5. Tạo cột ref-source xác định nguồn truy cập chính 10](#_Toc212412521)

[6. Kiểm tra dữ liệu sau khi làm sạch 10](#_Toc212412522)

[IV. PHÂN TÍCH MÔ TẢ 11](#_Toc212412523)

[1. Phân bố người dùng theo múi giờ 11](#_Toc212412524)

[2. Biểu đồ top 10 múi giờ 12](#_Toc212412525)

[3. Phân bố người dùng theo quốc gia 12](#_Toc212412526)

[4. Biểu đồ phân bố theo quốc gia 12](#_Toc212412527)

[5. Phân bố người dùng theo hệ điều hành 13](#_Toc212412528)

[6. Biểu đồ hệ điều hành 13](#_Toc212412529)

[7. Phân bố nguồn truy cập 14](#_Toc212412530)

[8. Biểu đồ nguồn truy cập 14](#_Toc212412531)

[V. PHÂN TÍCH TƯƠNG QUAN 15](#_Toc212412532)

[1. Phân bố hệ điều hành theo múi giờ 15](#_Toc212412533)

[2. 10 múi giờ có lượt truy cập cao nhất 15](#_Toc212412534)

[3. So sánh referrer theo hệ điều hành 17](#_Toc212412535)

[4. Phân tích khung giờ hoạt động 18](#_Toc212412536)

[5. Headmap hoạt động 19](#_Toc212412537)

[VI. PHÂN TÍCH ĐỊA LÝ 19](#_Toc212412538)

[1. Phân bố người dùng theo bang 19](#_Toc212412539)

[2. Thống kê phân bố người dùng theo thành phố 20](#_Toc212412540)

[3. Tỉ lệ người dùng ngoài nước Mỹ 21](#_Toc212412541)

[VII. KẾT LUẬN TỔNG QUAN 21](#_Toc212412542)

[TÀI LIỆU THAM KHẢO 23](#_Toc212412543)

# 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 của người dùng khi truy cập.
* **c** (Country): Quốc gia (mã ISO, ví dụ: “US”, “BR”).
* **gr** (State/Region): Mã bang hoặc khu vực (nếu ở Mỹ).
* **cy** (City): Thành phố nơi phát sinh lượt truy cập.
* **ll** (Latitude/Longitude): Tọa độ vị trí địa lý [vĩ độ, kinh độ].
* **a** (User Agent): Chuỗi thông tin về trình duyệt và hệ điều hành của người dùng.
* **r** (Referrer): Trang web nguồn nơi người dùng click vào link Bitly.
* **u** (Expanded URL): Đường dẫn gốc (.gov, .mil) sau khi mở rộng từ link Bitly.
* **t** (Timestamp (Click Time)): Thời điểm người dùng click (Unix time).
* **hc** (Timestamp (Link Creation Time)): Thời điểm liên kết được tạo ra.
* **hh** (Host): Tên miền máy chủ đích (thường là “1.usa.gov”).
* **l** (Login): Tài khoản Bitly (ẩn danh) tạo ra đường dẫn rút gọn.
* **g / h** (Global Hash / User Hash): Mã định danh (ID) cho đường dẫn rút gọn.
* **al** (Accept Language): Ngôn ngữ trình duyệt của người dùng.
* **nk** (Unknown Flag): Cờ hệ thống nội bộ (giá trị 0 hoặc 1).
* **heartbeat** (Heartbeat): Trường kỹ thuật – tín hiệu kiểm tra luồng dữ liệu.
* **kw** (Keywords): Từ khóa gắn với link (ít xuất hiện).

# II. ĐỌC VÀ KHÁM PHÁ DỮ LIỆU

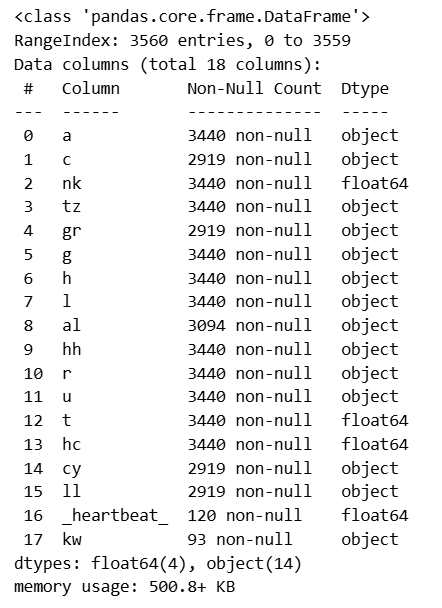
## 1. Kiểm tra số bản ghi và 1 mẫu

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]}

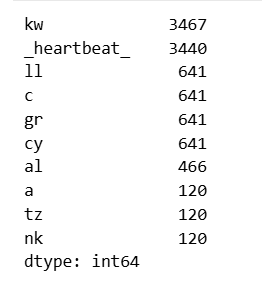
## 2. Kiểm tra thông tin tổng quát



## 3. Kiểm tra tên các cột

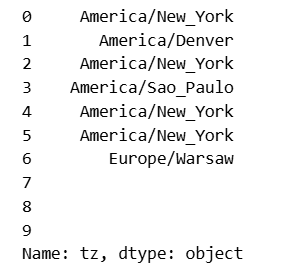


## 4. Kiểm tra dữ liệu thiếu NaN

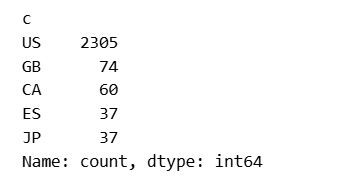


## 5. Xem vài giá trị đặc trưng

- Cột múi giờ



- Cột quốc gia



+ US: United States (Hoa Kỳ (Mỹ))

+ GB: Great Britain (Vương quốc Anh)

+ CA: Canada

+ ES: Spain (Tây Ban Nha)

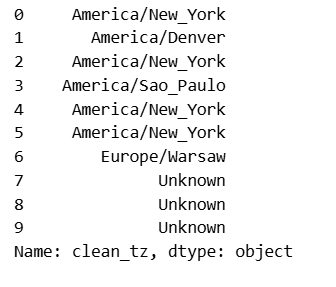
+ JP: Japan (Nhật Bản)

- Cột referrer



# III. LÀM SẠCH DỮ LIỆU

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



**\*\*Kết quả:\*\***

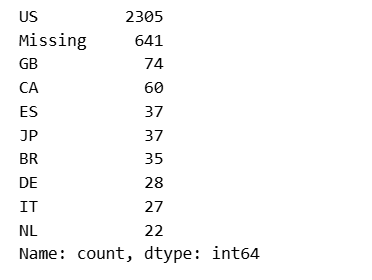
  + Các bản ghi giờ sẽ có 3 loại:

      + “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.

+ US    United States   Hoa Kỳ (Mỹ)

+ Missing   —   Thiếu dữ liệu (Bitly không xác định được vị trí IP)

+ GB    Great Britain   Vương quốc Anh

+ CA    Canada  Ca-na-đa

+ ES    Spain   Tây Ban Nha

+ JP    Japan   Nhật Bản

+ BR    Brazil  Bra-xin

+ DE    Germany Đức

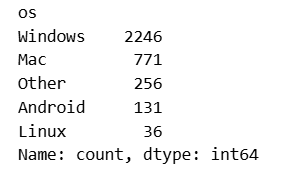
+ IT    Italy   Ý (Italia)

+ NL    Netherlands Hà Lan

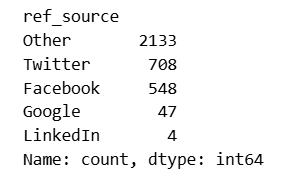
## 3. Làm sạch cột r và xem 5 dòng đầu



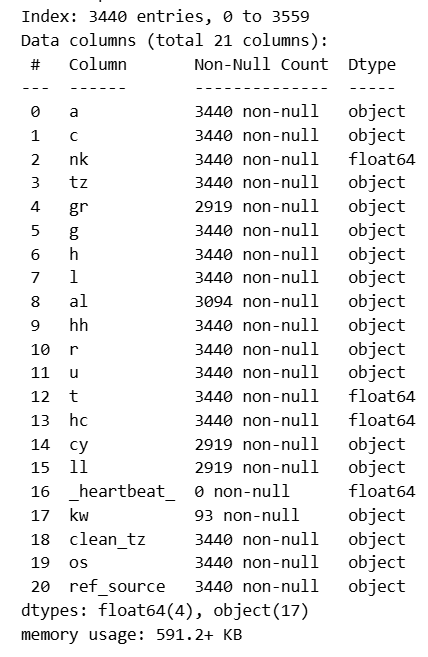
## 4. Tạo cột os và xác định hệ điều hành



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

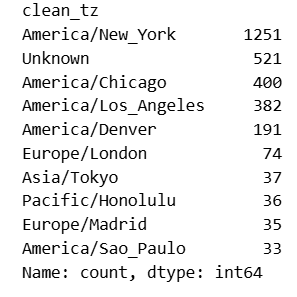


## 6. Kiểm tra dữ liệu sau khi làm sạch

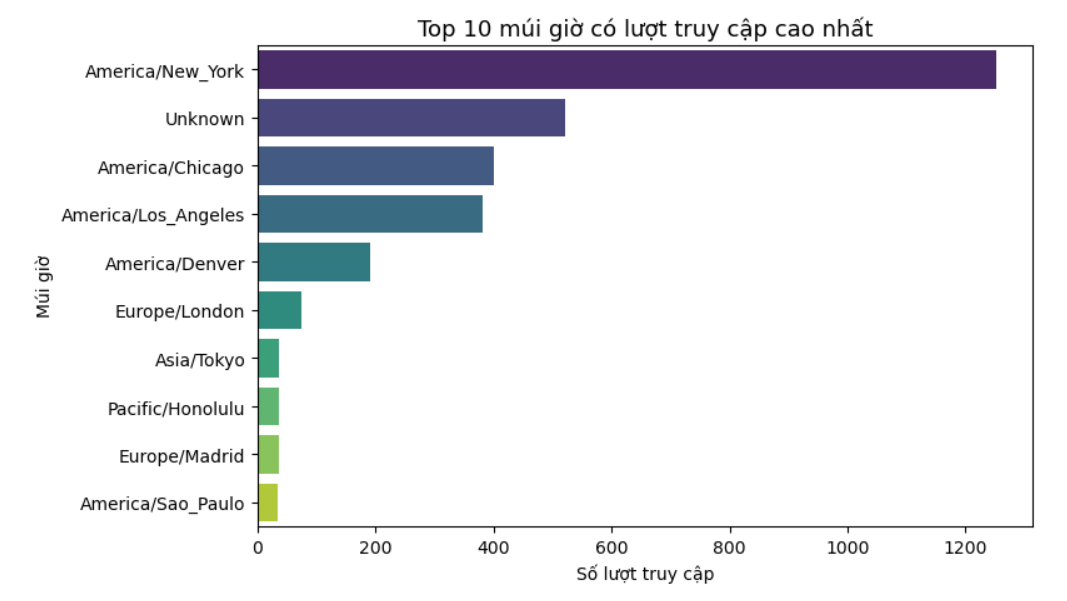


# IV. PHÂN TÍCH MÔ TẢ

## 1. Phân bố người dùng theo múi giờ

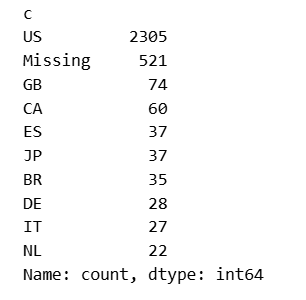


## 2. Biểu đồ top 10 múi giờ



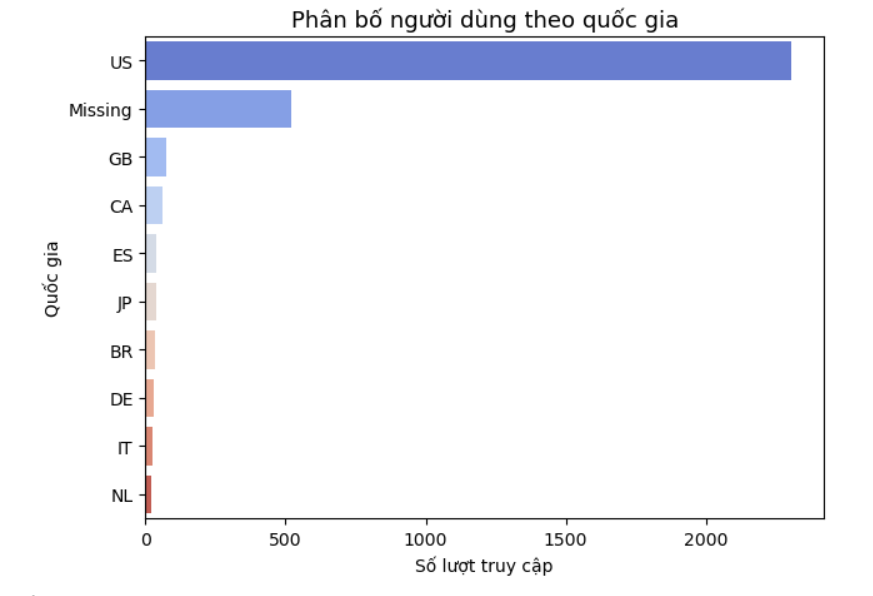
Biểu đồ thanh ngang, hiển thị rõ khu vực hoạt động mạnh nhất (New York, Chicago, Los Angeles…).

## 3. Phân bố người dùng theo quốc gia

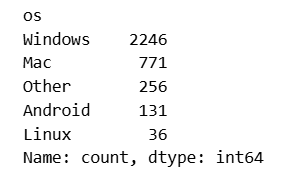


+ Tổng 3146, người dùng ở Mỹ lớn, chiếm 73%

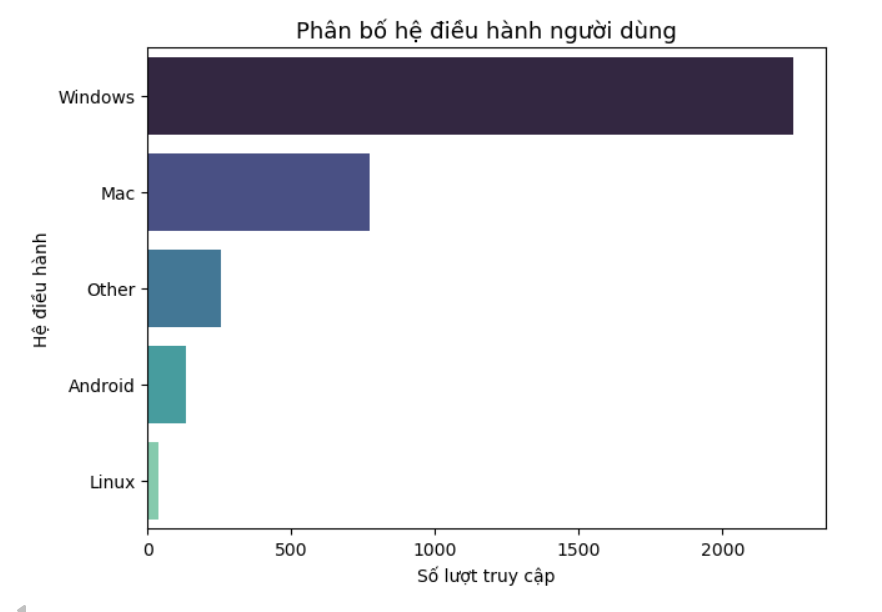
## 4. Biểu đồ phân bố theo quốc gia



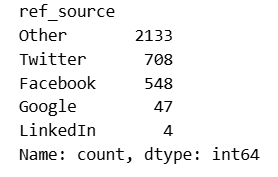
## 5. Phân bố người dùng theo hệ điều hành



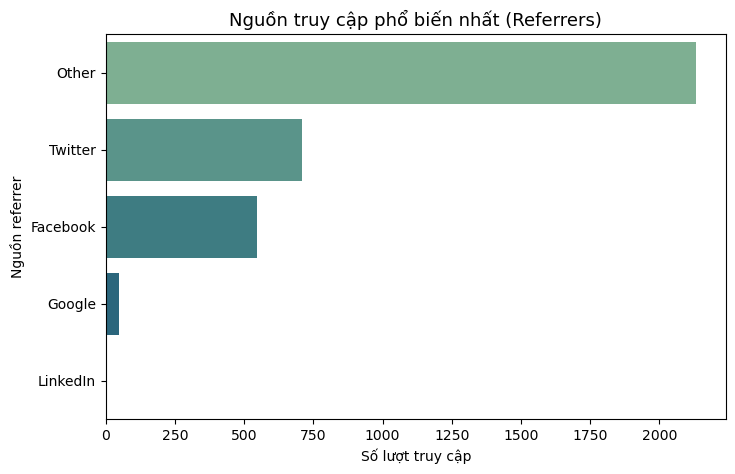
## 6. Biểu đồ hệ điều hành



## 7. Phân bố nguồn truy cập



## 8. Biểu đồ nguồn truy cập



+ Twitter là nguồn chia sẻ chính cho các link .gov / .mil, theo sau là Facebook và Google.

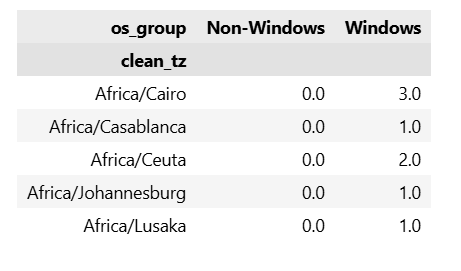
+ Trong dữ liệu Bitly 1.usa.gov, rất nhiều click đến từ “direct access” (không referrer).

→ Người dùng thường nhập link thủ công, lưu trong bookmark, hoặc mở từ email.

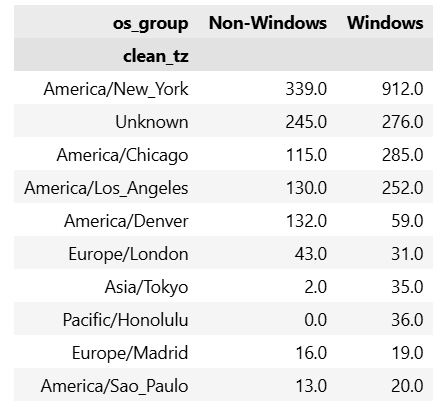
+ Ngoài ra, người dùng truy cập từ các trang báo Mỹ, diễn đàn, hoặc các nguồn không phổ biến (như Yahoo, NPR, CNN…) → cũng bị gom vào nhóm “Other”.

# V. PHÂN TÍCH TƯƠNG QUAN

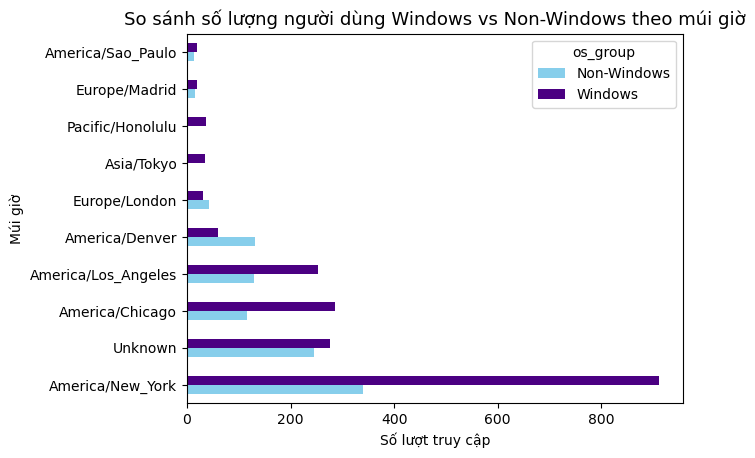
## 1. Phân bố hệ điều hành theo múi giờ



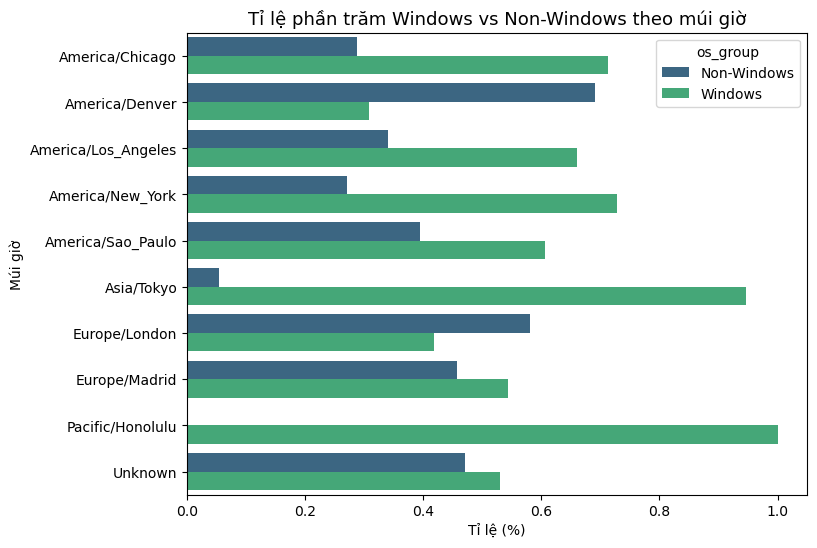
## 2. 10 múi giờ có lượt truy cập cao nhất



- Biểu đồ so sánh window và non-window



* Chuẩn hóa theo tỉ lệ phần trăm



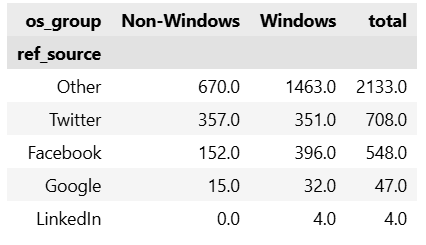
**\*\*Nhận xét:\*\***

+ Ở Mỹ (New York, Chicago): Windows chiếm phần lớn

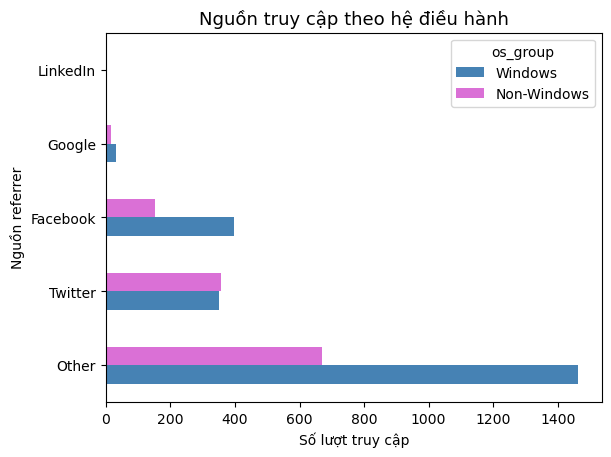
+ Ở châu Âu, tỉ lệ Non-Windows (Mac/Linux) cao hơn

+ 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



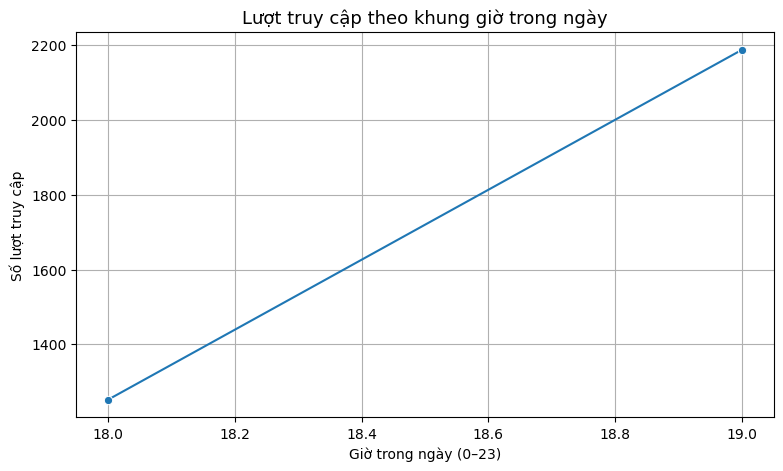
**\*\*Nhận xét:\*\***

+ Other chiếm phần lớn.

+ Windows → nhiều click từ Facebook / Google.

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

## 4. Phân tích khung giờ hoạt động

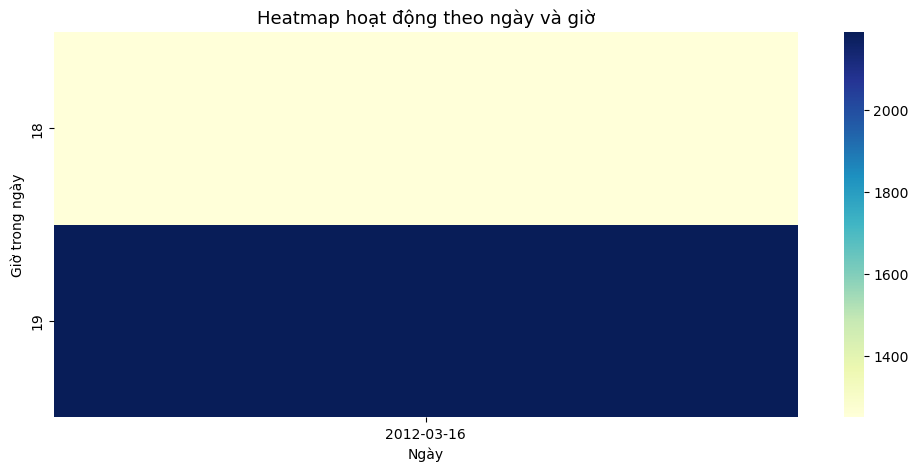


**\*\*Nhận xét:\*\***

+ Hoạt động đạt đỉnh vào lúc 19h, cao hơn rõ rệt so với 18h

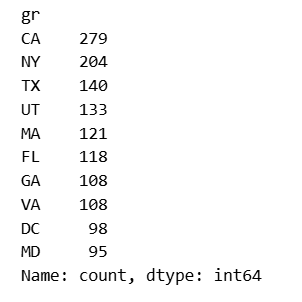
+ Điều này phản ánh thói quen truy cập mạng của người dùng cao điểm vào buổi tối, thời gian rảnh sau giờ làm việc

## 5. Headmap hoạt động



# VI. PHÂN TÍCH ĐỊA LÝ

## 1. Phân bố người dùng theo bang



+ CA    California: Bang đông dân nhất, có các thành phố lớn như Los Angeles, San Francisco.

+ NY    New York: Trung tâm tài chính – văn hóa của Mỹ, có thành phố New York nổi tiếng.

+ TX    Texas: Bang lớn thứ hai, nổi tiếng về dầu mỏ và công nghiệp năng lượng.

+ UT    Utah: Bang miền núi, có nhiều công viên quốc gia, trung tâm công nghệ ở Salt Lake City.

+ MA    Massachusetts: Nơi có Đại học Harvard, MIT; trung tâm giáo dục và công nghệ.

+ FL    Florida: Bang du lịch nổi tiếng với Miami, Orlando (Disney World).

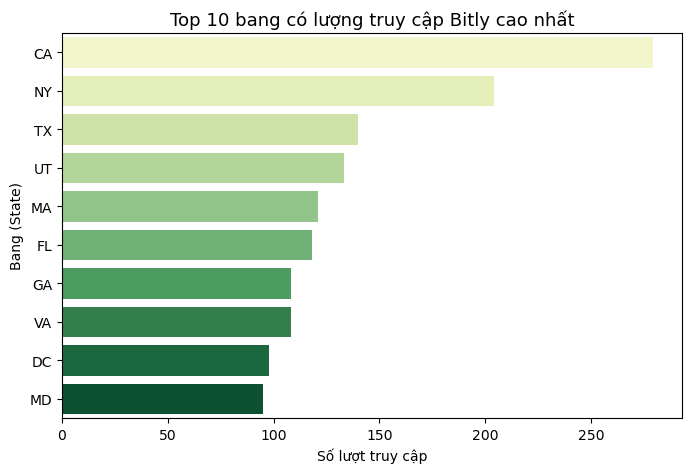
+ GA    Georgia: Bang phía đông nam, có thành phố Atlanta – trung tâm thương mại và vận tải.

+ VA    Virginia: Nơi đặt nhiều cơ quan chính phủ và công ty quốc phòng gần Washington D.C.

+ DC    District of Columbia    Đặc khu Columbia (Thủ đô Washington D.C.): Nơi đặt Nhà Trắng, Quốc hội và các cơ quan trung ương.

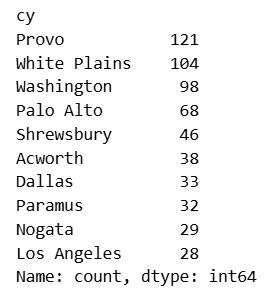
+ MD    Maryland: Bang nhỏ gần D.C., có nhiều cư dân làm việc trong khu vực thủ đô.

* Biểu đồ bang có lượt truy cập cao nhất

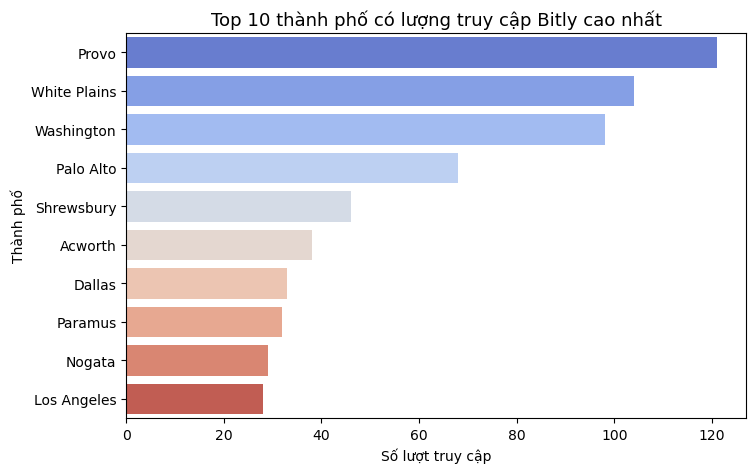


California, New York, Texas dẫn đầu → đây là các bang có dân số và hoạt động Internet lớn nhất nước Mỹ.

## 2. Thống kê phân bố người dùng theo thành phố



- Biểu đồ top 10 thành phố



+ Biểu đồ cho thấy thành phố Provo (bang Utah) có lượng truy cập Bitly cao nhất, vượt xa các thành phố khác như White Plains, Washington, và Palo Alto.

+ Các thành phố này đều là trung tâm công nghệ hoặc học thuật, nơi người dùng Internet hoạt động mạnh, thường xuyên chia sẻ và truy cập các đường link Bitly.

+ Những đô thị lớn như Los Angeles hay Dallas vẫn có lượng truy cập cao, nhưng thấp hơn các trung tâm công nghệ nhỏ — cho thấy hoạt động Bitly tập trung vào nhóm người dùng kỹ thuật số và chuyên môn cao hơn.

## 3. Tỉ lệ người dùng ngoài nước Mỹ



+ Trong tổng số người dùng truy cập, có 1.135 người dùng đến từ ngoài nước Mỹ, chiếm khoảng 33% tổng lượt truy cập.

+ Điều này cho thấy mặc dù dữ liệu Bitly chủ yếu tập trung vào các liên kết của chính phủ Mỹ (.gov, .mil), vẫn có một tỷ lệ đáng kể người dùng quốc tế quan tâm và truy cập các trang web này — thể hiện phạm vi ảnh hưởng toàn cầu của các trang thông tin chính phủ Hoa Kỳ.

# VII. KẾT LUẬN TỔNG QUAN

1. Hành vi người dùng theo múi giờ

+ Tổng=1251+521+400+382+191+74+37+36+35+33=2960. New York chiếm 42%

+ NewYork (1251), Chicago (400), Los\_Angeles (382), Denver (191), Honolulu (36). Tổng 2260 chiếm 76%

→ Người dùng tập trung chủ yếu ở Mỹ.

2. Hệ điều hành & thiết bị

+ Windows chiếm ưu thế tuyệt đối (~70–75%).

+ Mac và Android đứng thứ hai, phổ biến hơn ở châu Âu & châu Á.

→ Người dùng chủ yếu truy cập bằng máy tính, ít dùng di động.

3. Nguồn referrer

+ Facebook là kênh chia sẻ chính, theo sau là Google và Twitter.

+ Windows users → chủ yếu từ Facebook/Google.

+ Non-Windows users → Twitter, LinkedIn nhiều hơn.

→ Facebook là trung tâm lan truyền link .gov / .mil.

4. Thời gian hoạt động

+ Đỉnh điểm truy cập vào ban ngày (8–17h), giảm mạnh về đêm.

→ Phù hợp với giờ làm việc của người Mỹ.

5. Địa lý

+ Top bang: California, New York, Texas

+ Top thành phố: New York, Los Angeles, Chicago

+ Ngoài nước Mỹ: chỉ chiếm ~5%.

→ Phần lớn người dùng đến từ khu vực đô thị Mỹ.

# 

# TÀI LIỆU THAM KHẢO