**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**--------\*\*\*--------**

A blue button with white text

AI-generated content may be incorrect.

BÁO CÁO THỰC HÀNH

MÔN PHÂN TÍCH DỮ LIỆU

**TÊN ĐỀ TÀI PHÂN TÍCH DỮ LIỆU THỰC TẾ VỚI PYTHON**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | **ThS. Đỗ Như Tài** |
| **Nhóm sinh viên thực hiện:** |  |
| Hồ Thị Thanh Thảo | 3122410389 |
| Nguyễn Thị Hồng Thắm | 3122410392 |
| Phan Văn Thảo | 3122410391 |
| Nguyễn Hoàng Thiên Bảo | 3122410019 |

***Thành phố Hồ Chí Minh, tháng 10 năm 2025***

**LỜI MỞ ĐẦU**

Phân tích dữ liệu là một kỹ năng thiết yếu trong thời đại số hóa hiện nay. Với sự phát triển nhanh chóng của công nghệ, lượng dữ liệu được tạo ra mỗi ngày tăng theo cấp số nhân. Dữ liệu này tồn tại ở khắp nơi - từ các nền tảng mạng xã hội, thương mại điện tử, y tế cho đến chính phủ. Tuy nhiên, dữ liệu thô không có giá trị nếu chúng ta không biết cách khai thác và phân tích để rút ra những thông tin có ý nghĩa từ đó.

Python đã trở thành một trong những ngôn ngữ lập trình phổ biến nhất cho phân tích dữ liệu nhờ vào những thư viện mạnh mẽ như Pandas, NumPy và Matplotlib. Những công cụ này cung cấp cho chúng ta khả năng xử lý, trực quan hóa và phân tích dữ liệu một cách hiệu quả.

Báo cáo này được thực hiện dựa trên Chương 13 - "Data Analysis Examples" của cuốn sách "Python for Data Analysis" (O'Reilly 2022) của tác giả Wes McKinney. Chương này cung cấp những ví dụ thực tế về cách ứng dụng các kỹ thuật phân tích dữ liệu với Python vào các tập dữ liệu khác nhau từ các lĩnh vực khác nhau.

Trong báo cáo này, chúng em sẽ:

* Tìm hiểu về 5 tập dữ liệu thực tế: Bitly Data, MovieLens, US Baby Names, USDA Food Database, và Federal Election Commission Data
* Học cách tải, làm sạch, và xử lý dữ liệu từ các nguồn khác nhau (JSON, CSV)
* Áp dụng các kỹ thuật phân tích như groupby, pivot\_table, merge, explode
* Trực quan hóa dữ liệu bằng Matplotlib và Seaborn
* Rút ra những insights có giá trị từ dữ liệu

Mục đích của báo cáo này không chỉ là để hiểu rõ các kỹ thuật phân tích mà còn để thấy được sự ứng dụng thực tế của chúng trong các tình huống thực tế. Qua đó, chúng em hy vọng có thể nâng cao kỹ năng phân tích dữ liệu và góp phần áp dụng Python một cách hiệu quả hơn trong công việc sau này.

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn chân thành đến *Ths. Đỗ Như Tài*, người đã tận tình hướng dẫn và chia sẻ những kiến thức quý báu trong suốt quá trình thực hiện đề tài này. Sự chỉ bảo tận tình và những góp ý xây dựng từ thầy đã giúp chúng em hiểu rõ hơn về các vấn đề chuyên môn và hoàn thiện sản phẩm một cách tốt nhất.

Mặc dù đã rất cố gắng, nhưng do hạn chế về thời gian và kinh nghiệm, chắc chắn rằng không thể tránh khỏi những sai sót. Chúng em rất mong nhận được ý kiến đóng góp từ thầy để có thể cải thiện và hoàn thiện hơn nữa.

Một lần nữa, chúng em xin chân thành cảm ơn!

**PHÂN CÔNG**

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên** | **Công việc** |
| 1 | Hồ Thị Thanh Thảo | Làm báo cáo, tổng hợp code |
| 2 | Nguyễn Thị Hồng Thắm | Phân tích dataset bitly  Phân tích dataset MovieLens |
| 3 | Nguyễn Hoàng Thiên Bảo | Phân tích dataset babynames  Phân tích dataset USDA |
| 4 | Phan Văn Thảo | Phân tích dataset FEC, làm slide |

**MỤC LỤC**

[CHƯƠNG 1. GIỚI THIỆU CHƯƠNG 1](#_Toc211438162)

[1.1. Mục đích của chương 1](#_Toc211438163)

[1.2. Các dataset được sử dụng 1](#_Toc211438164)

[1.3. Các kỹ thuật chính được áp dụng 2](#_Toc211438165)

[1.3.1. Tải dữ liệu (Data loading) 2](#_Toc211438166)

[1.3.2. Làm sạch và xử lý dữ liệu (Data cleaning) 2](#_Toc211438167)

[1.3.3. Phân tích và Aggregation 3](#_Toc211438168)

[1.3.4. Trực quan hoá 3](#_Toc211438169)

[1.3.5. Thống kê và tổng hợp 3](#_Toc211438170)

[1.4. Cấu trúc học tập 3](#_Toc211438171)

[CHƯƠNG 2. PHÂN TÍCH DỮ LIỆU BITLY (1.USA.GOV) 5](#_Toc211438172)

[2.1. Giới thiệu về dữ liệu 5](#_Toc211438175)

[2.1.1. Nguồn gốc dữ liệu 5](#_Toc211438176)

[2.1.2. Định dạng dữ liệu 5](#_Toc211438177)

[2.1.3. Cấu trúc dữ liệu chính 6](#_Toc211438178)

[2.2. Quy trình thực hiện 6](#_Toc211438179)

[2.3. Kết quả 8](#_Toc211438180)

[2.3.1. Top 10 múi giờ phổ biến 8](#_Toc211438181)

[2.3.2. So sánh Windows vs Not-Windows - absolute counts 9](#_Toc211438182)

[2.3.3. So sánh Windows vs Not-Windows - tỉ lệ phần trăm 10](#_Toc211438183)

[2.4. Nhận xét 11](#_Toc211438184)

[CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU MOVIELENS 1M 13](#_Toc211438185)

[3.1. Giới thiệu về dữ liệu 13](#_Toc211438187)

[3.2. Quy trình thực hiện 13](#_Toc211438188)

[3.3. Kết quả 14](#_Toc211438189)

[3.3.1. Phân tích đánh giá trung bình theo giới tính 14](#_Toc211438190)

[3.3.2. Phân tích sự khác biệt giới tính trong điểm đánh giá 16](#_Toc211438191)

[3.3.3. Độ biến thiên của điểm số và mức độ đồng thuận giữa khán giả 17](#_Toc211438192)

[3.3.4. Xu hướng đánh giá theo thể loại và nhóm tuổi 19](#_Toc211438193)

[3.4. Nhận xét 20](#_Toc211438194)

[CHƯƠNG 4. PHÂN TÍCH DỮ LIỆU US BABY NAMES 23](#_Toc211438195)

[4.1. Giới thiệu về dữ liệu 23](#_Toc211438197)

[4.2. Quy trình thực hiện 23](#_Toc211438198)

[4.3. Kết quả 25](#_Toc211438199)

[4.3.1. Phân tích tổng số trẻ sơ sinh theo năm và giới tính 25](#_Toc211438200)

[4.3.2. Tỷ lệ phổ biến của từng tên 27](#_Toc211438201)

[4.3.3. Phân tích xu hướng đặt tên 28](#_Toc211438202)

[4.4. Nhận xét 33](#_Toc211438203)

[CHƯƠNG 5. PHÂN TÍCH DỮ LIỆU USDA FOOD DATABASE 36](#_Toc211438204)

[5.1. Giới thiệu về dữ liệu 36](#_Toc211438205)

[5.2. Quy trình thực hiện 37](#_Toc211438206)

[5.3. Kết quả 38](#_Toc211438207)

[5.3.1. Thực phẩm phổ biến nhất trong dữ liệu 38](#_Toc211438208)

[5.3.2. Phân tích hàm lượng trung vị của kẽm theo nhóm thực phẩm 39](#_Toc211438209)

[5.3.3. Thực phẩm có hàm lượng axit amin cao nhất 40](#_Toc211438210)

[5.4. Nhận xét 41](#_Toc211438211)

[CHƯƠNG 6. PHÂN TÍCH DỮ LIỆU FEC 2012 43](#_Toc211438212)

[6.1. Giới thiệu về dữ liệu 43](#_Toc211438213)

[6.2. Quy trình thực hiện 43](#_Toc211438214)

[6.3. Kết quả 44](#_Toc211438215)

[6.3.1. Phân tích quyên góp theo nghề nghiệp 44](#_Toc211438216)

[6.3.2. Phân tích quyên góp theo nhà tuyển dụng 46](#_Toc211438217)

[6.3.3. Phân tích theo mức đóng góp 47](#_Toc211438218)

[6.3.4. Phân tích theo bang 48](#_Toc211438219)

[6.4. Nhận xét 49](#_Toc211438220)

[CHƯƠNG 7. KẾT LUẬN 51](#_Toc211438221)

[7.1. Tổng kết quá trình và mục tiêu 51](#_Toc211438222)

[7.2. Ý nghĩa tổng hợp từ các phân tích 51](#_Toc211438223)

[7.3. Đánh giá về công cụ và kỹ thuật 52](#_Toc211438224)

[7.4. Hạn chế và hướng phát triển 52](#_Toc211438225)

[7.5. Kết luận chung 52](#_Toc211438226)

[TÀI LIỆU THAM KHẢO 54](#_Toc211438227)

**DANH MỤC HÌNH ẢNH**

[Hình 2.1. Top 10 múi giờ phổ biến nhất trong dữ liệu Bitly 8](#_Toc211437820)

[Hình 2.2. Phân bố hệ điều hành Windows và Not-Windows theo số lượng tuyệt đối tại 10 múi giờ phổ biến nhất 9](#_Toc211437821)

[Hình 2.3. So sánh hệ điều hành Windows và Not-Windows theo tỉ lệ phần trăm trong 10 múi giờ phổ biến nhất 10](#_Toc211437822)

[Hình 3.1. Top 10 phim được nữ giới đánh giá cao nhất 14](#_Toc211437823)

[Hình 3.2. Top 10 phim được nam giới đánh giá cao nhất 15](#_Toc211437824)

[Hình 3.3. Biểu đồ thể hiện 20 bộ phim có sự khác biệt điểm trung bình lớn nhất giữa nam và nữ 16](#_Toc211437825)

[Hình 3.4. Top 10 bộ phim có độ lệch chuẩn đánh giá cao nhất 17](#_Toc211437826)

[Hình 3.5. Biểu đồ nhiệt thể hiện điểm đánh giá trung bình của các thể loại phim theo nhóm tuổi 19](#_Toc211437827)

[Hình 4.1. Biểu đồ tổng số lượng trẻ theo giới tính qua các năm 26](#_Toc211437828)

[Hình 4.2. Số lượng trẻ sơ sinh của một số tên tiêu biển qua các năm 28](#_Toc211437829)

[Hình 4.3. Biểu đồ tỉ lệ tổng top 1000 tên theo giới tính qua từng năm 29](#_Toc211437830)

[Hình 4.4. Biểu đồ thể hiện top 50% tên phổ biến 30](#_Toc211437831)

[Hình 4.5. Biểu đồ phân bố chữ cái cuối của tên theo giới tính và năm 31](#_Toc211437832)

[Hình 4.6. Biểu đồ xu hướng tên nam kết thúc bằng d, n, y theo năm 32](#_Toc211437833)

[Hình 4.7. Biểu đồ thể hiện sự chuyển đổi giới tính của tên 'Lesl' qua thời gian (1880–2010) 33](#_Toc211437834)

[Hình 5.1. 10 nhóm thực phẩm phổ biến nhất trong cơ sở dữ liệu 38](#_Toc211437835)

[Hình 5.2. Hàm lượng trung vị kẽm trong các nhóm thực phẩm 39](#_Toc211437836)

[Hình 5.3. Các loại thực phẩm chứa hàm lượng axit amin cao nhất theo từng chất dinh dưỡng 40](#_Toc211437837)

[Hình 6.1. Biểu đồ tổng quyên góp theo nghề nghiệp và đảng phái 45](#_Toc211437838)

[Hình 6.2. Top 10 nhà tuyển dụng đóng góp nhiều nhất theo ứng viên 46](#_Toc211437839)

[Hình 6.3. Tỷ lệ tổng số tiền quyên góp theo từng mức đóng góp 47](#_Toc211437840)

[Hình 6.4. Tỷ lệ quyên góp theo bang 48](#_Toc211437841)

**DANH MỤC BẢNG**

[Bảng 1.1. Danh sách bộ dữ liệu được phân tích 2](#_Toc211437842)

[Bảng 1.2. Các kỹ thuật đọc dữ liệu được sử dụng 2](#_Toc211437843)

[Bảng 1.3. Các kỹ thuật xử lý và phân tích dữ liệu được sử dụng 3](#_Toc211437844)

[Bảng 1.4. Tổng hợp các phần thực hành phân tích dữ liệu 3](#_Toc211437845)

[Bảng 2.1. Ý nghĩa các trường trong bộ dữ liệu Bitly Data from 1.USA.gov 6](#_Toc211437846)

[Bảng 3.1. Cấu trúc các tệp dữ liệu trong MovieLens 1M Dataset 13](#_Toc211437847)

[Bảng 4.1. Các trường dữ liệu trong dataset US Baby Names 23](#_Toc211437848)

[Bảng 4.2. Tỷ lệ phổ biến của từng tên 27](#_Toc211437849)

[Bảng 5.1. Các trường dữ liệu trong dataset USDA FOOD 36](#_Toc211437850)

[Bảng 6.1. Các trường dữ liệu trong dataset FEC 2012 43](#_Toc211437851)

# GIỚI THIỆU CHƯƠNG

## Mục đích của chương

Chương 13 – *“Data Analysis Examples”* trong sách Python for Data Analysis (O’Reilly, 2022) được biên soạn nhằm giúp người học ứng dụng các kỹ thuật phân tích dữ liệu đã học ở các chương trước vào các bài toán thực tế.  
Thông qua các ví dụ cụ thể, chương này giúp người đọc hiểu rõ toàn bộ quy trình phân tích dữ liệu - từ thu thập dữ liệu, làm sạch, biến đổi, trực quan hóa đến rút ra kết luận.

**Mục đích chính:**

* Áp dụng các kỹ thuật đã học trong các chương trước vào các tình huống phân tích dữ liệu thực tế
* Hiểu rõ quy trình hoàn chỉnh của phân tích dữ liệu: từ tải dữ liệu → làm sạch → phân tích → trực quan hóa → rút ra kết luận
* Làm quen với các loại dữ liệu và định dạng khác nhau (JSON, CSV, v.v.)
* Phát triển kỹ năng xử lý dữ liệu messier, không chuẩn hóa (real-world data)

**Mục đích phụ:**

* Cung cấp những ví dụ có thể tái sử dụng cho các dự án phân tích dữ liệu trong tương lai
* Giới thiệu những best practices trong data cleaning và data wrangling
* Khuyến khích học viên tự khám phá và thử nghiệm với nhiều nguồn dữ liệu khác nhau

## Các dataset được sử dụng

Chương này bao gồm 5 tập dữ liệu thực tế đến từ nhiều lĩnh vực khác nhau:

Bảng 1.1. Danh sách bộ dữ liệu được phân tích

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên dataset** | **Mô tả ngắn gọn** | **Định dạng dữ liệu** |
| 1 | Bitly Data from 1.USA.gov | Dữ liệu người dùng rút gọn link *.gov/.mil*, chứa thông tin trình duyệt và múi giờ. | JSON |
| 2 | MovieLens 1M Dataset | Dữ liệu đánh giá phim từ người dùng (giới tính, độ tuổi, nghề nghiệp). | Text (CSV-like) |
| 3 | US Baby Names 1880–2010 | Dữ liệu tên trẻ em ở Mỹ trong hơn 130 năm. | CSV |
| 4 | USDA Food Database | Cơ sở dữ liệu thông tin dinh dưỡng của các món ăn và sản phẩm. | JSON |
| 5 | 2012 Federal Election Commission Database | Dữ liệu quyên góp cho chiến dịch tranh cử tổng thống Mỹ 2012. | CSV |

## Các kỹ thuật chính được áp dụng

### Tải dữ liệu (Data loading)

Bảng 1.2. Các kỹ thuật đọc dữ liệu được sử dụng

|  |  |  |
| --- | --- | --- |
| **Kỹ thuật** | **Dataset** | **Mục đích** |
| json.loads() | Bitly, USDA | Đọc dữ liệu JSON từ file |
| pd.read\_csv() | MovieLens, Baby Names, Election | Đọc dữ liệu CSV |
| pd.read\_table() | MovieLens | Đọc dữ liệu delimiter khác |

### Làm sạch và xử lý dữ liệu (Data cleaning)

* Xử lý missing values: fillna(), dropna(), notna()
* Xử lý dữ liệu không hợp lệ: Lọc, kiểm tra điều kiện
* Loại bỏ duplicates: drop\_duplicates()
* Chuyển đổi kiểu dữ liệu: Mapping, casting
* Tách và ghép dữ liệu: String manipulation, explode()

### Phân tích và Aggregation

Bảng 1.3. Các kỹ thuật xử lý và phân tích dữ liệu được sử dụng

|  |  |  |
| --- | --- | --- |
| **Kỹ thuật** | **Mục đích** | **Ví dụ** |
| value\_counts() | Đếm tần suất | Đếm múi giờ, tên phim |
| groupby() | Nhóm dữ liệu | Nhóm theo giới tính, lứa tuổi |
| pivot\_table() | Tạo bảng tổng hợp | Mean ratings theo giới tính |
| merge() / join() | Kết hợp bảng | Merge users với ratings |
| explode() | Giải nén danh sách | Tách genres thành các hàng riêng |

### Trực quan hoá

* Matplotlib: Plotting cơ bản
* Seaborn: Bar plots, heatmaps
* Trực quan hóa biểu đồ cơ bản: Line charts, bar charts, scatter plots

### Thống kê và tổng hợp

* Tính mean, median, std: Để so sánh và phân tích
* Cumulative sum (cumsum()): Để tìm diversity
* Searchsorted: Tìm vị trí chèn
* Normalization: Chuẩn hóa dữ liệu để so sánh

## Cấu trúc học tập

Chương được chia thành nhiều phần, mỗi phần tập trung vào một bộ dữ liệu cụ thể và mục tiêu học tập riêng:

Bảng 1.4. Tổng hợp các phần thực hành phân tích dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| **Phần** | **Tên ví dụ** | **Nội dung chính** | **Kỹ năng đạt được** |
| 13.1 | Bitly Data from 1.USA.gov | Phân tích JSON, thống kê múi giờ, phân loại người dùng theo hệ điều hành. | Xử lý file JSON, thống kê, visualization |
| 13.2 | MovieLens 1M Dataset | Phân tích điểm đánh giá phim theo giới tính, tuổi, thể loại. | Merge, groupby, pivot\_table |
| 13.3 | US Baby Names 1880–2010 | Phân tích xu hướng đặt tên và mức độ đa dạng theo thời gian. | concat, groupby, time-series |
| 13.4 | USDA Food Database | Phân tích thành phần dinh dưỡng và nhóm thực phẩm. | Làm việc với dữ liệu JSON lồng nhau |
| 13.5 | Federal Election Commission Dataset | Phân tích quyên góp chính trị, theo nghề nghiệp, bang, và ứng viên. | Binning, phân tích định lượng, visualization |

# PHÂN TÍCH DỮ LIỆU BITLY (1.USA.GOV)



## Giới thiệu về dữ liệu

### Nguồn gốc dữ liệu

Bộ dữ liệu được lấy từ dịch vụ rút gọn liên kết 1.usa.gov – một nền tảng do Bitly phối hợp cùng chính phủ Hoa Kỳ. Mỗi bản ghi trong dữ liệu đại diện cho một lượt người dùng truy cập hoặc rút gọn đường dẫn thuộc tên miền .gov hoặc .mil. Bộ dữ liệu này được dùng để thực hành quy trình phân tích dữ liệu thực tế, gồm các bước: đọc dữ liệu JSON, làm sạch, thống kê, và trực quan hóa.

### Định dạng dữ liệu

Dữ liệu được lưu trong file JSON, mỗi dòng là một chuỗi JSON chứa thông tin về người dùng.

Một dòng dữ liệu mẫu:

{"a": "Mozilla/5.0 (Windows NT 6.1; WOW64)...",

"c": "US",

"nk": 1,

"tz": "America/New\_York",

"gr": "NY",

"g": "A5J7Td",

"h": "c1lhGj",

"l": "orofrog",

"t": 1331923247,

"hc": 1331822918,

"r": "http://www.facebook.com/l/...",

"u": "http://www.ncbi.nlm.nih.gov/pubmed/22415991"}

### Cấu trúc dữ liệu chính

Bảng 2.1. Ý nghĩa các trường trong bộ dữ liệu Bitly Data from 1.USA.gov

|  |  |  |
| --- | --- | --- |
| **Trường** | **Ý nghĩa** | **Ví dụ** |
| tz | Múi giờ của người dùng | America/New\_York |
| a | Thông tin trình duyệt và hệ điều hành (User Agent) | Mozilla/5.0 (Windows NT 6.1; WOW64) |
| r | Referrer (trang web người dùng đến từ) | <http://www.facebook.com/>... |
| u | Đường dẫn gốc được rút gọn | <http://www.ncbi.nlm.nih.gov/>... |
| c | Quốc gia | US |

## Quy trình thực hiện

**Bước 1: Đọc và xử lý dữ liệu JSON**

import json

path = "datasets/bitly\_usagov/example.txt"

# Đọc dữ liệu

with open(path) as f:

    records = [json.loads(line) for line in f]

# In 5 bản ghi đầu tiên

for i, rec in enumerate(records[:5], 1):

    print(f"Record {i}: {rec}")

Sau khi đọc, dữ liệu được chuyển thành danh sách dictionary (list of dict), dễ dàng thao tác bằng pandas.

**Bước 2: Tạo DataFrame và kiểm tra dữ liệu**

import pandas as pd

frame = pd.DataFrame(records)

frame.info()

frame.head()

* Xác định các cột có giá trị trống (NaN) hoặc thiếu (None).
* Các giá trị thiếu trong cột tz được thay bằng 'Missing', giá trị rỗng ('') được thay bằng 'Unknown'.

**Bước 3: Phân tích múi giờ (Time Zone)**

* Đếm tần suất xuất hiện của các múi giờ:

clean\_tz = frame["tz"].fillna("Missing")

clean\_tz[clean\_tz == ""] = "Unknown"

tz\_counts = clean\_tz.value\_counts()

tz\_counts.head()

* Hiển thị top múi giờ phổ biến nhất bằng biểu đồ thanh:

import seaborn as sns

subset = tz\_counts.head()

sns.barplot(y=subset.index, x=subset.to\_numpy())

**Bước 4: Phân tích thông tin trình duyệt và hệ điều hành**

* Tách thông tin hệ điều hành từ chuỗi User Agent:

cframe = frame[frame["a"].notna()].copy()

cframe["os"] = np.where(cframe["a"].str.contains("Windows"),

                        "Windows", "Not Windows")

cframe["os"].head(5)

* Đếm số lượng người dùng theo múi giờ và hệ điều hành:

by\_tz\_os = cframe.groupby(["tz", "os"])

agg\_counts = by\_tz\_os.size().unstack().fillna(0)

agg\_counts.head()

**Bước 5: Nhóm dữ liệu và trực quan hóa kết quả**

* Nhóm dữ liệu theo time zone và OS
* Sử dụng unstack() để tạo pivot table
* Vẽ biểu đồ với seaborn
* Chuẩn hóa dữ liệu (normalization)

## Kết quả

### Top 10 múi giờ phổ biến

A graph with blue bars

AI-generated content may be incorrect.

Hình 2.1. Top 10 múi giờ phổ biến nhất trong dữ liệu Bitly

Kết quả phân tích tần suất cho thấy dữ liệu tập trung chủ yếu ở các múi giờ của Hoa Kỳ (Hình 2.1). America/New\_York nổi bật với 1.251 bản ghi, chiếm hơn 1/3 tổng số dữ liệu (~35%). Nhóm “Unknown” cũng chiếm tỷ lệ đáng kể (521 bản ghi, tương đương khoảng 15%), phản ánh hiện tượng thiếu thông tin về múi giờ trong nhiều lượt truy cập. Đây là một yếu tố quan trọng cần lưu ý trong các bước phân tích sau, vì dữ liệu thiếu có thể gây sai lệch khi diễn giải kết quả.

Ngoài ra, các múi giờ lớn khác gồm America/Chicago (400 bản ghi) và America/Los\_Angeles (382 bản ghi), cho thấy sự phân bố người dùng khá tập trung tại các trung tâm dân cư lớn của Hoa Kỳ. America/Denver cũng đóng góp một lượng dữ liệu đáng kể (191 bản ghi), mặc dù thấp hơn đáng kể so với ba khu vực chính kể trên.

Bên ngoài phạm vi Hoa Kỳ, sự xuất hiện của các múi giờ quốc tế như Europe/London (74 bản ghi) hay Asia/Tokyo (37 bản ghi) cho thấy phạm vi sử dụng dịch vụ có tính toàn cầu, nhưng số lượng tương đối nhỏ so với khối lượng dữ liệu từ Mỹ. Các múi giờ khác như Pacific/Honolulu (36 bản ghi) và Europe/Madrid (35 bản ghi) chỉ chiếm tỷ trọng không đáng kể.

### So sánh Windows vs Not-Windows - absolute counts

A graph of a bar graph

AI-generated content may be incorrect.

Hình 2.2. Phân bố hệ điều hành Windows và Not-Windows theo số lượng tuyệt đối tại 10 múi giờ phổ biến nhất

Phân tích số lượng tuyệt đối theo từng múi giờ cho thấy sự khác biệt rõ rệt trong phân bố hệ điều hành. America/New\_York là khu vực chiếm ưu thế nhất với 912 lượt Windows trên tổng 1.251 quan sát (72.90%), tương ứng khoảng 35% toàn bộ dữ liệu có thông tin múi giờ. Điều này khẳng định phần lớn lưu lượng rút gọn liên kết trong tập dữ liệu xuất phát từ bờ Đông Hoa Kỳ và chủ yếu đến từ thiết bị sử dụng Windows. Tình hình ở America/Chicago và America/Los\_Angeles cũng tương tự, với tỷ lệ Windows lần lượt đạt 71.25% và 65.97%, cho thấy sự phổ biến của Windows tại các trung tâm dân cư lớn của Mỹ.

Ngược lại, một số trường hợp ngoại lệ đáng chú ý xuất hiện. America/Denver có 59 lượt Windows so với 132 Not-Windows, tức Windows chỉ chiếm 30.89%. Đây là khu vực duy nhất trong nhóm có sự vượt trội rõ rệt của Not-Windows (69.11%), phản ánh khác biệt tiềm tàng về cơ cấu thiết bị. Europe/London cũng ghi nhận tỷ lệ Not-Windows cao hơn Windows (43 so với 31, tương ứng 58.11% Not-Windows), gợi ý sự khác biệt về hành vi sử dụng thiết bị giữa Mỹ và châu Âu.

Một số múi giờ có quy mô nhỏ lại cho kết quả cực đoan. Asia/Tokyo có đến 94.59% quan sát là Windows (35/37), trong khi Pacific/Honolulu thậm chí ghi nhận 100% Windows (36/36). Tuy nhiên, do kích thước mẫu ở các khu vực này rất nhỏ, những tỷ lệ này khó mang tính khái quát. Tương tự, Europe/Madrid (54.29% Windows, n=35) và America/Sao\_Paulo (60.61% Windows, n=33) cho thấy phân bố gần cân bằng nhưng chưa đủ cơ sở để kết luận xu hướng lâu dài.

Nhóm dữ liệu Unknown (không ghi múi giờ) cũng chiếm tỷ trọng đáng kể với 521 quan sát, trong đó Windows và Not-Windows phân bố khá cân bằng (52.98% và 47.02%). Điều này cho thấy các bản ghi thiếu thông tin không thiên lệch về một hệ điều hành cụ thể và việc loại bỏ hoàn toàn nhóm này có thể gây mất mát thông tin quan trọng.

### So sánh Windows vs Not-Windows - tỉ lệ phần trăm

A graph of a bar chart

AI-generated content may be incorrect.

Hình 2.3. So sánh hệ điều hành Windows và Not-Windows theo tỉ lệ phần trăm trong 10 múi giờ phổ biến nhất

Biểu đồ hình 2.3 trình bày tỉ lệ phần trăm giữa Windows và Not-Windows trong từng múi giờ, cho phép quan sát sự phân bố hệ điều hành một cách độc lập với kích thước mẫu. Kết quả cho thấy Windows vẫn chiếm ưu thế tại phần lớn múi giờ, đặc biệt ở America/New\_York, America/Chicago và America/Los\_Angeles, nơi Windows chiếm trên 65% số quan sát. Điều này củng cố nhận định rằng ở các trung tâm lớn của Hoa Kỳ, thiết bị Windows vẫn là nền tảng phổ biến.

Ngược lại, ở America/Denver sự phân bố đảo chiều rõ rệt: chỉ khoảng 31% người dùng thuộc nhóm Windows, trong khi 69% còn lại thuộc Not-Windows. Đây là một ngoại lệ quan trọng, cho thấy cấu trúc thiết bị ở khu vực này khác biệt đáng kể so với các khu vực khác. Europe/London cũng thể hiện xu hướng tương tự với Not-Windows chiếm đa số (≈58%), phản ánh sự khác biệt địa lý và thói quen sử dụng hệ điều hành giữa Mỹ và châu Âu.

Một số múi giờ nhỏ lại thể hiện tỉ lệ cực đoan. Asia/Tokyo có gần 95% quan sát thuộc Windows, trong khi Pacific/Honolulu thậm chí đạt 100%. Tuy nhiên, do kích thước mẫu nhỏ (dưới 40 quan sát), các kết quả này có độ tin cậy thấp và cần được diễn giải thận trọng. Ngược lại, Europe/Madrid và America/Sao\_Paulo cho thấy phân bố khá cân bằng giữa hai hệ điều hành, nhưng số lượng mẫu cũng còn hạn chế.

## Nhận xét

Kết quả phân tích cho thấy hệ điều hành Windows chiếm ưu thế tại hầu hết các múi giờ, đặc biệt ở các khu vực lớn của Hoa Kỳ như America/New\_York, America/Chicago và America/Los\_Angeles. Thực tế, giai đoạn 2011 Windows vẫn là hệ điều hành phổ biến nhất trên thị trường máy tính cá nhân, được sử dụng rộng rãi trong môi trường văn phòng, trường học và hộ gia đình. Điều này giải thích tại sao phần lớn lưu lượng rút gọn liên kết đến từ thiết bị Windows.

Tuy nhiên, một số ngoại lệ đáng chú ý lại phản ánh sự đa dạng thực tế về thói quen công nghệ. Tại America/Denver, tỷ lệ người dùng Not-Windows vượt trội có thể liên quan đến đặc thù dân cư và môi trường công nghệ: khu vực này tập trung nhiều trường đại học, viện nghiên cứu và cộng đồng công nghệ, vốn có xu hướng sử dụng macOS hoặc Linux. Tại Europe/London, sự phổ biến cao của Not-Windows cũng phản ánh xu hướng người dùng châu Âu ưa chuộng thiết bị Apple trong phân khúc trung và cao cấp, cùng với tỷ lệ sử dụng smartphone sớm hơn Mỹ ở thời điểm đó.

Một số kết quả cực đoan ở các mẫu nhỏ (như Pacific/Honolulu 100% Windows, hay Asia/Tokyo gần 95% Windows) có thể xuất phát từ đặc thù thị trường địa phương hoặc sự thiên lệch ngẫu nhiên do số lượng quan sát ít. Ví dụ, tại Nhật Bản năm 2011, Windows chiếm phần lớn trong máy tính để bàn và laptop, trong khi di động (Android/iOS) chưa chiếm ưu thế hoàn toàn, nên tỷ lệ cao Windows ở Asia/Tokyo cũng phản ánh xu thế thực tế lúc đó.

Nhóm Unknown (~15% dữ liệu) cho thấy hạn chế của dữ liệu thu thập tự động: nhiều bản ghi không chứa thông tin múi giờ. Việc này thường xảy ra khi trình duyệt hoặc hệ thống không gửi thông tin đầy đủ, hoặc do người dùng sử dụng các công cụ bảo mật/ẩn danh. Điều đáng chú ý là phân bố Windows/Not-Windows ở nhóm này khá cân bằng, tức không gây thiên lệch nghiêm trọng nhưng vẫn cần được tính đến khi đánh giá toàn bộ dữ liệu.

Nhìn chung, các kết quả phản ánh đúng bối cảnh công nghệ năm 2011: Windows thống trị toàn cầu, song những khác biệt địa lý (Denver, London) gợi ý rằng nhân khẩu học và thói quen tiêu dùng địa phương có ảnh hưởng đáng kể đến cơ cấu hệ điều hành. Điều này khẳng định dữ liệu thực tế từ log truy cập có thể cung cấp insight không chỉ về hành vi trực tuyến, mà còn phản ánh cấu trúc thị trường công nghệ theo vùng địa lý và văn hóa.

# PHÂN TÍCH DỮ LIỆU MOVIELENS 1M



## Giới thiệu về dữ liệu

Bộ dữ liệu MovieLens 1M được phát hành bởi GroupLens Research (Đại học Minnesota), chứa khoảng 1 triệu lượt đánh giá phim từ người dùng.  
Mục tiêu là phân tích hành vi đánh giá phim dựa trên giới tính, độ tuổi, và thể loại phim.

Cấu trúc dữ liệu gồm 3 file:

Bảng 3.1. Cấu trúc các tệp dữ liệu trong MovieLens 1M Dataset

|  |  |  |
| --- | --- | --- |
| **File** | **Mô tả** | **Số dòng** |
| users.dat | Thông tin người dùng (UserID, Gender, Age, Occupation, Zip-code) | 6,040 |
| ratings.dat | Điểm đánh giá phim (UserID, MovieID, Rating, Timestamp) | 1,000,209 |
| movies.dat | Thông tin phim (MovieID, Title, Genres) | 3,952 |

## Quy trình thực hiện

**Bước 1: Tải và đọc dữ liệu**

import pandas as pd

unames = ["user\_id", "gender", "age", "occupation", "zip"]

users = pd.read\_table("datasets/movielens/users.dat", sep="::",

                      header=None, names=unames, engine="python")

rnames = ["user\_id", "movie\_id", "rating", "timestamp"]

ratings = pd.read\_table("datasets/movielens/ratings.dat", sep="::",

                        header=None, names=rnames, engine="python")

mnames = ["movie\_id", "title", "genres"]

movies = pd.read\_table("datasets/movielens/movies.dat", sep="::",

                       header=None, names=mnames, engine="python")

**Bước 2: Gộp các bảng dữ liệu**

a = pd.merge(pd.merge(ratings, users), movies)

Tạo ra một DataFrame duy nhất chứa thông tin người dùng, điểm số và phim.

**Bước 3: Phân tích trung bình điểm đánh giá theo giới tính**

mean\_ratings = data.pivot\_table("rating", index="title",

                                columns="gender", aggfunc="mean")

mean\_ratings.head(5)

**Bước 4: Lọc phim có ít nhất 250 lượt đánh giá**

ratings\_by\_title = data.groupby("title").size()

ratings\_by\_title.head()

active\_titles = ratings\_by\_title.index[ratings\_by\_title >= 250]

active\_titles

**Bước 5: Top 10 phim được nữ giới đánh giá cao nhất**

top\_female\_ratings = mean\_ratings.sort\_values("F", ascending=False)

top\_female\_ratings.head(10)

**Bước 6: Phân tích độ chênh lệch giữa nam và nữ**

mean\_ratings["diff"] = mean\_ratings["M"] - mean\_ratings["F"]

sorted\_by\_diff = mean\_ratings.sort\_values("diff")

**Bước 7: Phân tích độ chia rẽ (disagreement)**

rating\_std\_by\_title = data.groupby("title")["rating"].std()

rating\_std\_by\_title = rating\_std\_by\_title.loc[active\_titles]

rating\_std\_by\_title.sort\_values(ascending=False)[:10]

## Kết quả

### Phân tích đánh giá trung bình theo giới tính

A graph with red and white lines

AI-generated content may be incorrect.

Hình 3.1. Top 10 phim được nữ giới đánh giá cao nhất

A blue and white bar graph

AI-generated content may be incorrect.

Hình 3.2. Top 10 phim được nam giới đánh giá cao nhất

Kết quả phân tích điểm đánh giá trung bình theo giới tính cho thấy có sự khác biệt đáng kể trong thị hiếu điện ảnh giữa nam và nữ, mặc dù vẫn tồn tại những điểm giao thoa nhất định ở các tác phẩm được xem là kinh điển. Biểu đồ Top 10 phim được nữ giới đánh giá cao nhất (Hình 3.1) cho thấy nữ giới có xu hướng ưu ái các bộ phim hoạt hình ngắn, mang tính nhân văn hoặc chứa đựng yếu tố cảm xúc sâu sắc. Các tác phẩm như A Close Shave (1995) (điểm trung bình nữ 4.64; nam 4.47) và The Wrong Trousers (1993) (4.59; 4.46) đều thuộc thể loại hoạt hình stop-motion, nhẹ nhàng và hài hước. Bên cạnh đó, các phim chính kịch – tâm lý xã hội như Schindler’s List (1993) và The Shawshank Redemption (1994) cũng nhận được đánh giá rất cao từ khán giả nữ, thể hiện sự quan tâm của nhóm này đối với các chủ đề mang tính nhân văn và giá trị đạo đức. Ngược lại, biểu đồ Top 10 phim được nam giới đánh giá cao nhất (Hình 3.2) phản ánh thị hiếu khác biệt rõ rệt. Các bộ phim đứng đầu danh sách như The Godfather (1972), Seven Samurai (1954), Raiders of the Lost Ark (1981) và Star Wars: Episode IV – A New Hope (1977) chủ yếu thuộc thể loại hành động, sử thi hoặc tội phạm, đặc trưng bởi cốt truyện kịch tính, xung đột mạnh và nhân vật mang tính anh hùng. Điều này cho thấy nam giới có xu hướng đánh giá cao hơn các tác phẩm đề cao kỹ năng, bản lĩnh và sức mạnh nhân vật. Mặc dù vậy, vẫn có nhiều bộ phim xuất hiện trong cả hai bảng xếp hạng như Shawshank Redemption, Schindler’s List và The Usual Suspects, phản ánh mức độ đồng thuận cao giữa hai giới đối với các tác phẩm được xem là kiệt tác điện ảnh. Tuy nhiên, phân tích thống kê cho thấy điểm trung bình của nữ giới thường cao hơn khoảng 0.1–0.2 điểm so với nam giới, cho thấy xu hướng nữ giới đánh giá rộng lượng hơn, đặc biệt với các tác phẩm cảm động hoặc mang thông điệp tích cực.

### Phân tích sự khác biệt giới tính trong điểm đánh giá

A graph with a bar chart

AI-generated content may be incorrect.

Hình 3.3. Biểu đồ thể hiện 20 bộ phim có sự khác biệt điểm trung bình lớn nhất giữa nam và nữ

Hình 3.3 minh họa 20 bộ phim có sự chênh lệch điểm trung bình lớn nhất giữa hai giới tính. Trục hoành thể hiện giá trị Difference (Male – Female), trong đó các cột màu đỏ biểu thị phim được nữ giới đánh giá cao hơn, còn các cột màu xanh thể hiện phim được nam giới đánh giá cao hơn.

Giá trị càng xa khỏi mốc 0 cho thấy mức độ khác biệt trong đánh giá giữa hai giới càng lớn. Kết quả cho thấy sự phân hóa rõ rệt về thị hiếu điện ảnh giữa nam và nữ.

Nhóm phim được nữ giới yêu thích hơn chủ yếu thuộc thể loại tình cảm, nhạc kịch hoặc tâm lý – xã hội, điển hình như Dirty Dancing (1987), Grease (1978), Little Women (1994), Steel Magnolias (1989) hay The Color Purple (1985). Đây đều là những bộ phim mang thông điệp về tình yêu, gia đình và khẳng định vai trò của phụ nữ, phản ánh mối quan tâm đặc trưng của khán giả nữ đối với các giá trị cảm xúc và nhân văn. Trong đó, Dirty Dancing (1987) có mức chênh lệch cao nhất với diff ≈ -0.83, cho thấy nữ giới đánh giá phim này cao hơn nam giới gần một điểm trung bình. Ngược lại, nhóm phim được nam giới đánh giá cao hơn lại tập trung vào các thể loại hành động, chiến tranh, phiêu lưu hoặc hài, chẳng hạn The Good, The Bad and The Ugly (1966), Dumb & Dumber (1994), Evil Dead II (1987), Rocky III (1982) hay For a Few Dollars More (1965). Các tác phẩm này thường khai thác yếu tố kỹ thuật, hành động, cạnh tranh hoặc hài hước đậm chất “nam tính”, do đó dễ thu hút nhóm khán giả nam hơn.

Giá trị diff trong nhóm này dao động quanh +0.6 đến +0.7, phản ánh xu hướng nam giới đánh giá cao hơn đáng kể các phim thuộc thể loại này. Điểm đáng chú ý là mặc dù tồn tại khác biệt giới tính rõ ràng, song không có sự đối lập tuyệt đối giữa hai nhóm người xem. Một số tác phẩm như The Shawshank Redemption (1994) hay Schindler’s List (1993) được cả hai giới đánh giá rất cao và có mức chênh lệch nhỏ, thể hiện vùng đồng thuận thẩm mỹ trong cộng đồng khán giả đối với các bộ phim kinh điển có giá trị nghệ thuật và nhân văn vượt thời gian.

### Độ biến thiên của điểm số và mức độ đồng thuận giữa khán giả

A graph with different colored bars

AI-generated content may be incorrect.

Hình 3.4. Top 10 bộ phim có độ lệch chuẩn đánh giá cao nhất

Bên cạnh việc xem xét điểm trung bình, việc phân tích độ lệch chuẩn (standard deviation – std) của điểm đánh giá giúp nhận diện mức độ đồng thuận hay tranh cãi giữa các khán giả về một bộ phim. Độ lệch chuẩn cao phản ánh sự phân tán lớn trong đánh giá — tức là cùng một bộ phim nhưng người xem có quan điểm rất khác nhau, trong khi độ lệch chuẩn thấp cho thấy sự đồng thuận cao giữa các nhóm người xem. Biểu đồ trong Hình 3.4 trình bày Top 10 bộ phim có độ lệch chuẩn cao nhất, tức là những tác phẩm gây nhiều tranh cãi nhất trong cộng đồng người dùng MovieLens. Dễ nhận thấy rằng phần lớn các phim này thuộc nhóm hài, kinh dị hoặc tâm lý – nghệ thuật, vốn thường nhận được phản ứng trái chiều do khác biệt trong cảm nhận cá nhân và kỳ vọng thể loại.

Đứng đầu danh sách là Dumb & Dumber (1994) và The Blair Witch Project (1999) — hai bộ phim có độ lệch chuẩn xấp xỉ 1.25, thể hiện sự chia rẽ rõ rệt giữa khán giả. Dumb & Dumber là phim hài mang tính phi lý, gây cười bằng yếu tố ngớ ngẩn, khiến người xem hoặc rất yêu thích hoặc hoàn toàn không đồng cảm. The Blair Witch Project lại là phim kinh dị quay theo phong cách “found footage”, mới lạ với thời điểm phát hành, nên nhận được cả lời khen đột phá lẫn chỉ trích vì hình thức thô ráp. Các tác phẩm khác như Natural Born Killers (1994), Eyes Wide Shut (1999) hay Fear and Loathing in Las Vegas (1998) đều thuộc thể loại tâm lý hoặc nghệ thuật, thường có nội dung trừu tượng, gây tranh luận về thông điệp và phong cách thể hiện. Ngược lại, những bộ phim có độ lệch chuẩn thấp (không thể hiện trong hình) thường là các tác phẩm được đánh giá ổn định hơn, chẳng hạn The Shawshank Redemption hay Toy Story, cho thấy mức độ đồng thuận cao trong cộng đồng người xem về chất lượng nghệ thuật và giá trị cảm xúc.

### Xu hướng đánh giá theo thể loại và nhóm tuổi

A blue and green chart

AI-generated content may be incorrect.

Hình 3.5. Biểu đồ nhiệt thể hiện điểm đánh giá trung bình của các thể loại phim theo nhóm tuổi

Để khảo sát sự khác biệt trong thị hiếu điện ảnh giữa các nhóm tuổi, phân tích này sử dụng biểu đồ nhiệt (heatmap) thể hiện điểm đánh giá trung bình của từng thể loại phim (genre) theo nhóm tuổi. Trục tung biểu diễn thể loại, trong khi trục hoành thể hiện các nhóm tuổi được mã hóa theo dữ liệu MovieLens (1: <18 tuổi, 18: 18–24, 25: 25–34, 35: 35–44, 45: 45–49, 50: 50–55, 56: ≥56 tuổi). Màu sắc trong biểu đồ phản ánh mức độ yêu thích – điểm càng đậm biểu thị đánh giá càng cao.

Kết quả cho thấy xu hướng đa dạng về sở thích điện ảnh giữa các độ tuổi, đồng thời phản ánh quá trình thay đổi trong nhận thức và trải nghiệm của người xem theo thời gian. Trước hết, nhóm khán giả trẻ (<25 tuổi) có xu hướng đánh giá cao các thể loại mang tính hoạt hình, âm nhạc và hành động nhẹ nhàng, điển hình như Animation (điểm trung bình 3.62–3.70) và Musical (3.56–3.62). Điều này phù hợp với đặc điểm tâm lý của nhóm tuổi này: ưa thích giải trí, hình ảnh sinh động và cốt truyện dễ tiếp cận. Ngược lại, các thể loại phức tạp hoặc giàu chiều sâu như Drama và Film-Noir được nhóm này đánh giá thấp hơn so với người lớn tuổi.

Ở nhóm trung niên (35–50 tuổi), sự yêu thích trở nên cân bằng hơn giữa các thể loại. Các nhóm này thường đánh giá cao các phim Documentary, Drama và Mystery, với điểm trung bình đạt khoảng 3.9–4.0, cho thấy sự quan tâm tới các nội dung phản ánh xã hội, nhân sinh hoặc khai thác yếu tố hiện thực. Đáng chú ý, thể loại Film-Noir duy trì điểm trung bình cao nhất toàn bảng (khoảng 4.1–4.2) ở hầu hết các nhóm tuổi, cho thấy đây là thể loại được đánh giá cao ổn định, đặc biệt trong các nhóm khán giả trưởng thành.

Nhóm lớn tuổi (≥56 tuổi) thể hiện xu hướng đánh giá cao hơn toàn diện ở hầu hết các thể loại, đặc biệt với War (4.07), Film-Noir (4.13) và Documentary (3.96). Điều này có thể phản ánh sự đồng cảm lớn hơn với các chủ đề lịch sử, chiến tranh và giá trị hiện thực, thường gắn liền với trải nghiệm sống và bối cảnh thế hệ của nhóm này.

Ngược lại, thể loại Horror (kinh dị) có điểm trung bình thấp nhất trong hầu hết các nhóm tuổi (dao động 3.16–3.28), cho thấy thể loại này vẫn gây chia rẽ trong đánh giá, có thể do tính kén người xem hoặc cảm xúc tiêu cực mà nó gợi ra.

## Nhận xét

Phân tích bộ dữ liệu MovieLens 1M đã cho thấy mối liên hệ rõ rệt giữa đặc điểm nhân khẩu học (giới tính, độ tuổi) và hành vi đánh giá phim của khán giả. Các kết quả chỉ ra rằng mặc dù mức độ yêu thích phim có sự đồng thuận nhất định ở các tác phẩm kinh điển, nhưng vẫn tồn tại những khác biệt có ý nghĩa thống kê giữa các nhóm người xem.

Thứ nhất, giới tính là yếu tố ảnh hưởng mạnh mẽ đến xu hướng đánh giá.

Nữ giới có xu hướng ưu ái các phim mang tính nhân văn, cảm xúc và hoạt hình, thể hiện qua việc đánh giá cao hơn các tác phẩm như A Close Shave (1995) hay Schindler’s List (1993). Ngược lại, nam giới lại dành điểm cao hơn cho các bộ phim hành động, sử thi hoặc tội phạm như The Godfather (1972) và Seven Samurai (1954).

Sự khác biệt này phản ánh đặc điểm tâm lý và sở thích tiếp nhận của hai giới: trong khi nữ giới tập trung vào yếu tố cảm xúc và thông điệp nhân văn, nam giới bị thu hút bởi kỹ năng, hành động và cấu trúc kịch tính. Tuy nhiên, vẫn tồn tại những vùng giao thoa — điển hình là The Shawshank Redemption (1994) — thể hiện giá trị phổ quát vượt lên trên ranh giới giới tính.

Thứ hai, độ biến thiên điểm số cung cấp cái nhìn sâu hơn về mức độ đồng thuận trong cộng đồng người xem.

Các phim có độ lệch chuẩn cao như Dumb & Dumber (1994), The Blair Witch Project (1999) hay Natural Born Killers (1994) cho thấy sự chia rẽ mạnh mẽ trong cảm nhận khán giả. Phần lớn các tác phẩm này thuộc thể loại hài, kinh dị hoặc nghệ thuật thử nghiệm – nơi ranh giới giữa “hay” và “dở” mang tính chủ quan cao.

Ngược lại, các phim như The Shawshank Redemption hoặc Toy Story có độ lệch chuẩn thấp, thể hiện sự đồng thuận cao về chất lượng. Điều này khẳng định rằng độ lệch chuẩn không chỉ là chỉ số thống kê, mà còn là thước đo xã hội học phản ánh sự đa dạng trong diễn giải nghệ thuật và kỳ vọng thể loại.

Thứ ba, độ tuổi cũng cho thấy tác động đáng kể đến xu hướng đánh giá.

Người trẻ (<25 tuổi) có xu hướng đánh giá cao các thể loại mang tính giải trí và thị giác như Animation và Musical, trong khi nhóm trung niên và lớn tuổi (>35 tuổi) ưa chuộng Drama, Documentary và Film-Noir. Đặc biệt, thể loại Film-Noir duy trì điểm trung bình cao ổn định (≈4.1–4.2) ở hầu hết các nhóm tuổi, chứng tỏ sức hút bền vững của dòng phim cổ điển.

Điểm trung bình có xu hướng tăng dần theo độ tuổi, cho thấy khán giả lớn tuổi thường đánh giá hào phóng hơn, có thể do trải nghiệm sống và sự đồng cảm sâu sắc hơn với các chủ đề nhân sinh.

Tổng hợp lại, kết quả nghiên cứu khẳng định rằng các yếu tố nhân khẩu học có vai trò quyết định trong hành vi đánh giá phim. Giới tính, độ tuổi và sở thích thể loại không chỉ ảnh hưởng đến điểm số trung bình mà còn đến mức độ đồng thuận và hướng cảm xúc trong đánh giá.

Những phát hiện này mang giá trị ứng dụng thực tiễn trong việc xây dựng hệ thống gợi ý phim cá nhân hóa (personalized recommendation system) — nơi mô hình có thể tích hợp thông tin nhân khẩu học để đề xuất phim phù hợp hơn với từng nhóm người dùng. Ngoài ra, chúng cũng góp phần cung cấp bằng chứng định lượng cho các nghiên cứu xã hội học và tâm lý học về hành vi tiếp nhận văn hóa trong môi trường số.

# PHÂN TÍCH DỮ LIỆU US BABY NAMES



## Giới thiệu về dữ liệu

Dataset US Baby Names được cung cấp bởi US Social Security Administration (SSA) - Cơ quan An sinh Xã hội Hoa Kỳ, bao gồm dữ liệu về tên được đặt cho trẻ em từ năm 1880 đến năm 2010, kéo dài trong 130 năm. Dữ liệu được tổ chức thành 131 files CSV riêng biệt, mỗi file tương ứng với một năm, với tên file theo định dạng yob1880.txt, yob1881.txt cho đến yob2010.txt. Tổng cộng có khoảng 1,690,784 bản ghi trong toàn bộ dataset.

Các trường dữ liệu:

Bảng 4.1. Các trường dữ liệu trong dataset US Baby Names

|  |  |  |
| --- | --- | --- |
| **Trường** | **Ý nghĩa** | **Ví dụ** |
| name | Tên của trẻ | Emma |
| sex | Giới tính | F |
| births | Số trẻ mang tên đó trong năm | 20799 |

Một đặc điểm quan trọng của dataset này là nó chỉ bao gồm những tên xuất hiện ít nhất 5 lần trong một năm để đảm bảo tính riêng tư. Dữ liệu trong mỗi file được sắp xếp theo số lượng giảm dần, không có missing values và có cấu trúc rất tốt, khiến nó trở thành một dataset lý tưởng để học về phân tích chuỗi thời gian.

Dataset này có nhiều ứng dụng thực tế quan trọng như nghiên cứu xu hướng đặt tên theo thời gian, phân tích sự đa dạng văn hóa qua các thập kỷ, dự đoán tên phổ biến trong tương lai, cũng như nghiên cứu xã hội học về xu hướng giới tính và sự thay đổi quan niệm về tên người trong xã hội Mỹ.

## Quy trình thực hiện

**Bước 1: Đọc và kết hợp dữ liệu**

import pandas as pd

pieces = []

for year in range(1880, 2011):

    path = f"datasets/babynames/yob{year}.txt"

    frame = pd.read\_csv(path, names=["name", "sex", "births"])

    # Add a column for the year

    frame["year"] = year

    pieces.append(frame)

# Concatenate everything into a single DataFrame

names = pd.concat(pieces, ignore\_index=True)

display(names)

**Bước 2: Tính tổng số trẻ sinh theo năm và giới tính**

total\_births = names.pivot\_table("births", index="year",

                                 columns="sex", aggfunc=sum)

total\_births.tail()

**Bước 3: Tính tỷ lệ phổ biến cho mỗi tên**

def add\_prop(group):

    group["prop"] = group["births"] / group["births"].sum()

    return group

names = names.groupby(["year", "sex"], group\_keys=False).apply(add\_prop)

display(names)

**Bước 4: Tạo danh sách top 1000 tên theo năm và giới tính**

def get\_top1000(group):

    return group.sort\_values("births", ascending=False)[:1000]

grouped = names.groupby(["year", "sex"])

top1000 = grouped.apply(get\_top1000)

Lý do:

* Dataset đầy gồm 1.7 triệu records, quá lớn để phân tích
* Top 1000 đại diện cho phần lớn tên phổ biến
* Dễ dàng visualize và phân tích hơn

**Bước 5: Phân tích xu hướng tên**

* **Phân tích cụ thể theo thời gian**

subset = total\_births[["John", "Harry", "Mary", "Marilyn"]]

#subset.plot(subplots=True, figsize=(12, 10),

#           title="Number of births per year")

subset = subset.plot(subplots=True, figsize=(12, 10),

                 title="Số lượng trẻ sơ sinh theo tên qua các năm")

* **Đo lường Diversity**

Mục tiêu: Tính xem bao nhiêu tên chiếm 50% tổng số sinh

def get\_quantile\_count(group, q=0.5):

    group = group.sort\_values("prop", ascending=False)

    return group.prop.cumsum().searchsorted(q) + 1

diversity = top1000.groupby(["year", "sex"]).apply(get\_quantile\_count)

diversity = diversity.unstack()

* **Phân tích chữ cái cuối cùng trong tên**

def get\_last\_letter(x):

    return x[-1]

last\_letters = names["name"].map(get\_last\_letter)

last\_letters.name = "last\_letter"

table = names.pivot\_table("births", index=last\_letters,

                          columns=["sex", "year"], aggfunc=sum

* **Tên chuyển đổi giới tính**

Mục tiêu: Tìm tên từng là nam nhưng trở thành nữ (hoặc ngược lại)

# Tìm tất cả tên có chứa "Lesl"

all\_names = pd.Series(top1000["name"].unique())

lesley\_like = all\_names[all\_names.str.contains("Lesl")]

# Lọc dữ liệu

filtered = top1000[top1000["name"].isin(lesley\_like)]

filtered.groupby("name")["births"].sum()

# Tạo pivot table

table = filtered.pivot\_table("births", index="year",

                             columns="sex", aggfunc="sum")

table = table.div(table.sum(axis="columns"), axis="index")

fig = plt.figure()

table.plot(style={"M": "b-", "F": "r--"})

## Kết quả

### Phân tích tổng số trẻ sơ sinh theo năm và giới tính

A graph with blue and orange lines

AI-generated content may be incorrect.

Hình 4.1. Biểu đồ tổng số lượng trẻ theo giới tính qua các năm

Biểu đồ trên thể hiện số lượng trẻ sơ sinh theo giới tính (nam và nữ) qua các năm từ 1880 đến 2010. Dễ dàng nhận thấy rằng số trẻ sinh ra tăng mạnh dần qua thời gian, đặc biệt là giai đoạn từ khoảng năm 1900 đến sau Thế chiến II, khi hiện tượng “baby boom” diễn ra. Sau đó, số lượng trẻ sinh ra có xu hướng dao động nhẹ nhưng vẫn duy trì ở mức cao cho đến đầu thế kỷ XXI. Ngoài ra, có thể thấy trẻ nam (M) luôn nhiều hơn trẻ nữ (F) trong hầu hết các năm, tuy nhiên khoảng cách giữa hai đường biểu diễn tương đối ổn định. Điều này cho thấy xu hướng sinh con trai cao hơn con gái là phổ biến nhưng không quá chênh lệch, phản ánh sự cân bằng tương đối về giới tính khi sinh trong suốt giai đoạn được thống kê.

### Tỷ lệ phổ biến của từng tên

Bảng 4.2. Tỷ lệ phổ biến của từng tên

A screenshot of a table

AI-generated content may be incorrect.

Nhìn chung, dữ liệu cho thấy có sự đa dạng ngày càng tăng trong việc đặt tên cho trẻ sơ sinh theo thời gian — khi ở các năm đầu (như 1880) chỉ có vài tên phổ biến chiếm tỷ lệ lớn, còn ở các năm gần đây, số lượng tên xuất hiện nhiều hơn nhưng mỗi tên lại chiếm tỷ lệ nhỏ hơn trong tổng số trẻ sinh ra.

### Phân tích xu hướng đặt tên

A graph of different colored lines

AI-generated content may be incorrect.

Hình 4.2. Số lượng trẻ sơ sinh của một số tên tiêu biển qua các năm

Biểu đồ này minh họa mức độ phổ biến của các tên John, Harry, Mary, Marilyn qua thời gian.

* John: Xuất hiện ổn định từ 1880s, đạt đỉnh khoảng 1940–1950 với hơn 80.000 trẻ được đặt tên John mỗi năm. Sau 1960, độ phổ biến giảm dần nhưng vẫn là một trong những tên phổ biến
* Harry: Tăng nhanh vào khoảng 1910–1920, đạt đỉnh gần 10.000 trẻ/năm. Sau đó giảm dần, đặc biệt sau 1950.
* Mary: Là tên nữ phổ biến nhất trong suốt nửa đầu thế kỷ 20, đặc biệt 1910–1950 với hơn 60.000–70.000 trẻ/năm.
* Marilyn: Xuất hiện muộn hơn, nổi bật trong thập niên 1930–1950, sau đó giảm mạnh.

A graph with blue and orange lines

AI-generated content may be incorrect.

Hình 4.3. Biểu đồ tỉ lệ tổng top 1000 tên theo giới tính qua từng năm

Giai đoạn đầu (1880–1940), tổng tỉ lệ của top 1000 tên gần bằng 1.0 → Hầu như tất cả trẻ em đều mang tên nằm trong nhóm phổ biến nhất. Sau năm 1950, đặc biệt là từ 1980 trở đi, đường biểu diễn giảm dần, rõ rệt hơn ở giới nữ (F). Điều này cho thấy xu hướng đa dạng hóa tên: ngày càng nhiều tên mới được đặt cho trẻ, khiến 1000 tên phổ biến nhất chiếm tỷ lệ nhỏ hơn tổng số trẻ sinh ra.

A graph with blue and orange lines

AI-generated content may be incorrect.

Hình 4.4. Biểu đồ thể hiện top 50% tên phổ biến

Hình 4.4. cho thấy, số lượng tên phổ biến của bé gái luôn cao hơn bé trai → Tên nữ đa dạng hơn. Từ khoảng năm 1940–1960, số lượng tên giữ ổn định, sau đó tăng mạnh, đặc biệt từ 1980 trở đi.

A screenshot of a graph

AI-generated content may be incorrect.

Hình 4.5. Biểu đồ phân bố chữ cái cuối của tên theo giới tính và năm

***Tên nam (M):***

Chữ “n” tăng mạnh từ 14.3% (1910) → 36.3% (2010), trở thành kết thúc phổ biến nhất.

Chữ “d” và “s” giảm mạnh (11.3% → 2.3%, 13.0% → 6.5%).

Chữ “r” tăng nhẹ (6.4% (1910) → 8.7% (2010)), trong khi “l” và “t” giảm dần (10.4% → 7.0%, 6.5% → 2.3%).

Một số chữ hiếm như “o”, “y”, “w” tăng nhẹ trong giai đoạn 1960–2010.

***Tên nữ (F):***

Chữ “a” tăng mạnh từ 27.3% (1910) → 38.1% (2010), trở thành kết thúc phổ biến nhất.

Chữ “e” giảm từ 33.7% (1910) → 17.8% (2010), trong khi “n” tăng từ 7.9% (1910) → 14.0% (2010).

Chữ “y” tăng lên 15.2% (1960) rồi giảm còn 11.6% (2010).

Các chữ “t”, “s”, “l” đều giảm dần qua ba mốc 1910 → 1960 → 2010.

A graph with different colored lines

AI-generated content may be incorrect.

Hình 4.6. Biểu đồ xu hướng tên nam kết thúc bằng d, n, y theo năm

Tên nam kết thúc bằng “n” tăng mạnh và trở thành xu hướng phổ biến nhất, từ khoảng 15% năm 1880 lên hơn 35% sau năm 2000.

Tên kết thúc bằng “d” phổ biến trong giai đoạn 1910–1950 (đạt đỉnh gần 20%) nhưng sau đó giảm mạnh.

Tên kết thúc bằng “y” tăng nhẹ đến giữa thế kỷ 20 rồi giảm dần sau năm 1960.

→ Xu hướng chung: các tên nam hiện đại có xu hướng ưu tiên kết thúc bằng “n”, trong khi “d” và “y” xuất hiện ít thường xuyên hơn.

A graph showing the growth of the us dollar

AI-generated content may be incorrect.

Hình 4.7. Biểu đồ thể hiện sự chuyển đổi giới tính của tên 'Lesl' qua thời gian (1880–2010)

Hình 4.7 cho thấy:

Giai đoạn 1880–1940: Hầu hết người mang tên “Lesl” là nam giới. Tỷ lệ nam chiếm gần 100%. Tên này lúc đầu được xem là tên nam truyền thống.

Khoảng 1940–1950: Điểm giao nhau giữa hai đường (M và F) xuất hiện — tức là “Lesl” bắt đầu được đặt cho cả hai giới gần như ngang nhau.

Sau 1950: Tên “Lesl” nhanh chóng chuyển dịch sang phổ biến ở nữ. Tỷ lệ nữ chiếm gần như 100% sau năm 2000, trong khi tỷ lệ nam giảm xuống gần 0.

## Nhận xét

Sự đa dạng hóa tên là xu hướng rõ ràng và liên tục từ 1880 đến 2010, với tốc độ tăng nhanh hơn sau năm 1960. Có nhiều nguyên nhân giải thích cho hiện tượng này. Toàn cầu hóa đã giúp người Mỹ tiếp xúc với nhiều văn hóa và ngôn ngữ khác nhau, dẫn đến việc họ chọn tên từ nhiều nguồn gốc khác nhau hơn. Xu hướng cá nhân hóa ngày càng mạnh mẽ trong xã hội hiện đại khiến cha mẹ muốn con mình có tên độc đáo, khác biệt thay vì theo đuổi những tên phổ biến. Truyền thông đại chúng, đặc biệt là phim ảnh, TV và celebrities đã tạo ra nhiều xu hướng tên mới và nhanh chóng. Hơn nữa, áp lực truyền thống về việc đặt tên theo gia đình hoặc tôn giáo đã giảm đáng kể, cho phép cha mẹ tự do lựa chọn hơn. Điều này phản ánh sự thay đổi lớn trong văn hóa xã hội Mỹ hướng tới cá nhân hóa và tự do lựa chọn.

Một quan sát thú vị là tên nữ luôn đa dạng hơn tên nam trong mọi thời kỳ, với số lượng tên cần để đạt 50% gần gấp đôi. Có nhiều lý do giải thích cho hiện tượng này. Truyền thống gia đình thường khuyến khích truyền tên nam từ cha hoặc ông xuống con trai, dẫn đến ít thay đổi hơn trong tên nam. Kỳ vọng xã hội về tên nam thường yêu cầu chúng phải "mạnh mẽ" và "truyền thống", giới hạn phạm vi lựa chọn. Ngược lại, tên nữ được chấp nhận sáng tạo và thay đổi nhiều hơn. Ví dụ minh họa rõ điều này là các tên nam như John, William, James vẫn rất phổ biến qua nhiều thế hệ, trong khi tên nữ thay đổi nhanh chóng theo xu hướng từ Emma, Sophia đến Isabella.

Xu hướng chữ cái cuối, đặc biệt là sự gia tăng của tên nam kết thúc bằng 'n', có thể được giải thích bởi nhiều yếu tố. Từ góc độ ngữ âm, âm 'n' được nhiều người cho là dễ nghe và có vẻ hiện đại. Ảnh hưởng của celebrities và người nổi tiếng có tên kết thúc bằng 'n' cũng góp phần tạo ra xu hướng này. Hơn nữa, xu hướng có tính lan truyền - khi một tên trở nên phổ biến, nó dẫn đến sự xuất hiện của nhiều tên tương tự khác. Toàn bộ nhóm tên như Mason, Logan, Ethan, Aiden, Jayden xuất hiện nhiều trong top 100, tạo thành một "gia đình" tên có đặc điểm chung.

Hiện tượng gender fluid names (tên không phân biệt giới tính rõ ràng) phản ánh sự thay đổi sâu sắc trong quan niệm xã hội về giới tính. Ngày càng nhiều tên trở thành unisex (dùng cho cả nam và nữ) và một số tên chuyển từ nam sang nữ hoàn toàn. Điều này có ý nghĩa xã hội quan trọng: quan niệm về giới tính không còn rạch ròi như trước, phong trào bình đẳng giới cho rằng tên không nên phân biệt giới tính quá rõ, và cha mẹ có tự do chọn tên bất kể truyền thống giới tính của nó. Xu hướng này cho thấy xã hội Mỹ ngày càng cởi mở và tiến bộ trong quan điểm về giới tính và vai trò xã hội.

Từ góc độ ứng dụng thực tế, dataset này có giá trị lớn trong nhiều lĩnh vực. Trong business và marketing, hiểu xu hướng đặt tên giúp target đúng khách hàng và chọn tên sản phẩm phù hợp với thị hiếu thị trường. Trong demographic analysis, tên có thể được sử dụng để ước tính độ tuổi của khách hàng. Trong nghiên cứu khoa học, dataset này là nguồn tài liệu quý giá cho xã hội học nghiên cứu thay đổi văn hóa, cho ngôn ngữ học phân tích xu hướng ngôn ngữ, và cho nhân khẩu học dự báo xu hướng dân số. Trong cuộc sống hàng ngày, cha mẹ có thể tham khảo xu hướng này để đặt tên con và hiểu về background văn hóa của các tên khác nhau.

Tóm lại, dataset US Baby Names là một ví dụ xuất sắc về time series analysis và social data analysis. Nó cho thấy rằng dữ liệu đơn giản có thể mang lại insights sâu sắc về xã hội, visualization đóng vai trò quan trọng trong việc phát hiện patterns, context xã hội là cần thiết để giải thích dữ liệu một cách có ý nghĩa, và phân tích dữ liệu không chỉ là về số liệu mà còn là về kể câu chuyện đằng sau những con số đó.

# PHÂN TÍCH DỮ LIỆU USDA FOOD DATABASE

## Giới thiệu về dữ liệu

Dataset USDA Food Database được cung cấp bởi US Department of Agriculture (USDA) - Bộ Nông nghiệp Hoa Kỳ nhằm mục đích cung cấp thông tin dinh dưỡng chi tiết cho các loại thực phẩm. Phiên bản được sử dụng trong chương này do programmer Ashley Williams chuyển đổi sang định dạng JSON để dễ dàng xử lý hơn. Dataset chứa thông tin về 6,636 loại thực phẩm khác nhau với cấu trúc hierarchical/nested phức tạp, được lưu trong file database.json.

Các trường dữ liệu chính:

Bảng 5.1. Các trường dữ liệu trong dataset USDA FOOD

|  |  |  |  |
| --- | --- | --- | --- |
| **Trường** | **Kiểu** | **Mô tả** | **Ví dụ** |
| id | integer | ID duy nhất của thực phẩm | 21441 |
| description | string | Tên mô tả thực phẩm | "Cheese, caraway" |
| tags | array | Nhãn phân loại | ["KFC"], ["Organic"] |
| manufacturer | string | Nhà sản xuất | "Kentucky Fried Chicken" |
| group | string | Nhóm thực phẩm | "Fast Foods", "Dairy" |
| portions | array | Phần ăn tiêu chuẩn | [{amount, unit, grams}] |
| nutrients | array | Thông tin dinh dưỡng | [{value, units, description, group}] |

Dataset này có nhiều ứng dụng thực tế quan trọng trong nhiều lĩnh vực. Nó có thể được sử dụng để phát triển các ứng dụng mobile về dinh dưỡng giúp người dùng tra cứu thông tin dinh dưỡng của thực phẩm. Trong nghiên cứu sức khỏe cộng đồng, dataset này cung cấp dữ liệu để phân tích mối liên hệ giữa chế độ ăn và sức khỏe. Các chuyên gia dinh dưỡng sử dụng nó để tư vấn cho khách hàng về cách chọn thực phẩm phù hợp. Nó cũng cho phép so sánh giá trị dinh dưỡng giữa các loại thực phẩm khác nhau và hỗ trợ xây dựng thực đơn healthy và cân bằng.

## Quy trình thực hiện

**Bước 1: Đọc dữ liệu JSON**

import json

db = json.load(open("datasets/usda\_food/database.json"))

**Bước 2: Tạo DataFrame**

nutrients = pd.DataFrame(db[0]["nutrients"])

info\_keys = ["description", "group", "id", "manufacturer"]

info = pd.DataFrame(db, columns=info\_keys)

**Bước 3: Gộp thông tin thực phẩm và chất dinh dưỡng**

nutrients = []

for rec in db:

    fnuts = pd.DataFrame(rec["nutrients"])

    fnuts["id"] = rec["id"]

    nutrients.append(fnuts)

nutrients = pd.concat(nutrients, ignore\_index=True)

**Bước 4: Loại bỏ trùng lặp**

nutrients = nutrients.drop\_duplicates()

**Bước 5: Phân tích**

Tính trung vị giá trị dinh dưỡng theo nhóm:

fig = plt.figure()

result = ndata.groupby(["nutrient", "fgroup"])["value"].quantile(0.5)

result["Zinc, Zn"].sort\_values().plot(kind="barh", title="Hàm lượng trung vị của kẽm (Zn) theo nhóm thực phẩm")

## Kết quả

### Thực phẩm phổ biến nhất trong dữ liệu

A screenshot of a computer screen

AI-generated content may be incorrect.

Hình 5.1. 10 nhóm thực phẩm phổ biến nhất trong cơ sở dữ liệu

Hình 5.1 cho thấy:

Nhóm “Vegetables and Vegetable Products” (Rau củ và các sản phẩm từ rau củ) → Có 812 mục, là nhóm đa dạng nhất, chứng tỏ dữ liệu USDA chú trọng mạnh đến nguồn thực phẩm thực vật, phản ánh tầm quan trọng của rau củ trong dinh dưỡng.

Nhóm “Beef Products” (Sản phẩm từ thịt bò) đứng thứ 2 với 618 mục, → Thể hiện sự phong phú trong các loại thịt bò và cách chế biến khác nhau (tươi, đông lạnh, chế biến sẵn, v.v.).

Các nhóm “Baked Products” và “Breakfast Cereals” → Xuất hiện khá nhiều (400–500 mục), cho thấy cơ sở dữ liệu cũng tập trung vào nguồn thực phẩm chế biến phổ biến trong khẩu phần ăn hiện đại.

Nhóm “Legumes and Legume Products” và “Fast Foods” → Có số lượng tương đương (365), phản ánh cả thực phẩm lành mạnh lẫn đồ ăn nhanh đều được ghi nhận đầy đủ trong cơ sở dữ liệu.

Nhóm “Sweets”, “Fruits and Fruit Juices”, “Pork Products” → Chiếm tỷ lệ nhỏ hơn, nhưng vẫn thể hiện sự đa dạng nguồn thực phẩm từ nhiều nhóm dinh dưỡng khác nhau.

### Phân tích hàm lượng trung vị của kẽm theo nhóm thực phẩm

A graph with text and numbers

AI-generated content may be incorrect.

Hình 5.2. Hàm lượng trung vị kẽm trong các nhóm thực phẩm

Biểu đồ thể hiện hàm lượng trung vị của kẽm (Zn) trong các nhóm thực phẩm khác nhau. Kết quả cho thấy nhóm thực phẩm từ thịt bò (Beef Products) có hàm lượng kẽm cao nhất, đạt giá trị trung vị trên 5 mg, tiếp theo là thịt cừu và thịt thú rừng (Lamb, Veal, and Game Products) cùng các loại hạt và sản phẩm từ hạt (Nut and Seed Products).

Các nhóm thực phẩm như ngũ cốc ăn sáng (Breakfast Cereals), gia vị và thảo mộc (Spices and Herbs), thịt gia cầm (Poultry Products) và thịt heo (Pork Products) cũng có hàm lượng kẽm tương đối cao.

Ngược lại, nhóm thực phẩm có hàm lượng kẽm thấp nhất là đồ uống (Beverages), trái cây và nước trái cây (Fruits and Fruit Juices), chất béo và dầu (Fats and Oils) — các nhóm này gần như không chứa kẽm đáng kể.

Nhìn chung, các thực phẩm có nguồn gốc từ động vật và các loại hạt là nguồn cung cấp kẽm dồi dào nhất, trong khi thực phẩm có nguồn gốc thực vật hoặc chế biến lỏng chứa hàm lượng kẽm thấp hơn rõ rệt.

### Thực phẩm có hàm lượng axit amin cao nhất

A screenshot of a computer

AI-generated content may be incorrect.

Hình 5.3. Các loại thực phẩm chứa hàm lượng axit amin cao nhất theo từng chất dinh dưỡng

Bảng dữ liệu trên hiển thị những loại thực phẩm có hàm lượng cao nhất của từng loại axit amin trong nhóm “Amino Acids”.

Kết quả cho thấy đạm đậu nành (Soy protein isolate) là nguồn chứa nhiều loại axit amin nhất như Aspartic acid, Glutamic acid, Isoleucine, Leucine, Phenylalanine, Serine, Threonine, Tyrosine và Valine. Điều này khẳng định đậu nành là nguồn cung cấp protein thực vật chất lượng cao.

Ngoài ra, bột gelatin cũng giàu Alanine, Glycine, và Proline, còn các loại hải sản như cá tuyết (cod), hải cẩu (seal) và sư tử biển (sea lion) có hàm lượng cao các axit amin thiết yếu như Methionine, Lysine, và Tryptophan.

Một số thực phẩm đặc biệt khác như thịt cá voi khô (whale, beluga) và gà rán Kentucky Fried Chicken cũng xuất hiện trong danh sách, cho thấy nguồn axit amin phong phú không chỉ đến từ thực phẩm tự nhiên mà còn từ các món ăn chế biến.

Nhìn chung, các thực phẩm giàu protein – đặc biệt là đậu nành, gelatin và hải sản – là nguồn cung cấp axit amin chủ yếu cho cơ thể, góp phần quan trọng trong quá trình xây dựng và phục hồi mô.

## Nhận xét

Qua các biểu đồ và kết quả phân tích trong mục 5.3, có thể thấy bộ dữ liệu USDA Food Database phản ánh rõ đặc điểm đa dạng và chi tiết của nguồn thực phẩm. Biểu đồ hình 5.1 cho thấy nhóm “Vegetables and Vegetable Products” chiếm tỷ lệ lớn nhất trong cơ sở dữ liệu, thể hiện sự chú trọng của tổ chức USDA đối với các nguồn thực phẩm có nguồn gốc thực vật. Điều này phù hợp với xu hướng dinh dưỡng hiện nay, khi rau củ và trái cây ngày càng được xem là thành phần thiết yếu trong chế độ ăn lành mạnh. Sự xuất hiện đáng kể của các nhóm như “Beef Products” và “Baked Products” cũng cho thấy cơ sở dữ liệu được xây dựng toàn diện, bao phủ cả thực phẩm tươi sống lẫn thực phẩm chế biến.

Trong biểu đồ hình 5.2, phân tích hàm lượng trung vị của kẽm (Zn) cho thấy các nhóm thực phẩm có nguồn gốc động vật — đặc biệt là thịt bò, thịt cừu và các loại hạt — chứa hàm lượng kẽm cao hơn hẳn so với nhóm thực vật hoặc đồ uống. Điều này phản ánh chính xác đặc tính sinh học của kẽm, một khoáng chất chủ yếu được tìm thấy trong protein động vật. Kết quả này có thể được ứng dụng trong các nghiên cứu về dinh dưỡng và xây dựng thực đơn cân bằng khoáng chất cho từng nhóm đối tượng khác nhau.

Cuối cùng, biểu đồ hình 5.3 về hàm lượng axit amin cho thấy sự khác biệt rõ ràng giữa các loại thực phẩm. Đạm đậu nành (Soy protein isolate) nổi bật với hàm lượng cao của nhiều loại axit amin thiết yếu, khẳng định đậu nành là nguồn protein thực vật chất lượng cao, tương đương hoặc vượt trội so với protein động vật. Gelatin và hải sản cũng là nguồn cung cấp axit amin quan trọng, cho thấy vai trò của các nhóm thực phẩm giàu protein trong việc tái tạo mô và phát triển cơ bắp. Nhìn chung, phân tích này giúp người học không chỉ hiểu cách xử lý và trực quan hóa dữ liệu JSON phức tạp, mà còn thấy rõ ý nghĩa thực tế của việc phân tích dinh dưỡng, phục vụ cho các ứng dụng trong y học, sức khỏe và khoa học thực phẩm.

# PHÂN TÍCH DỮ LIỆU FEC 2012

## Giới thiệu về dữ liệu

Bộ dữ liệu do Ủy ban Bầu cử Liên bang (FEC) công bố, chứa thông tin các khoản quyên góp cho chiến dịch tranh cử tổng thống năm 2012 (Obama vs Romney). Dữ liệu được lưu trong tệp CSV có kích thước lớn, chứa hơn một triệu bản ghi, phản ánh chi tiết tên người đóng góp, nghề nghiệp, nhà tuyển dụng, bang sinh sống, ứng cử viên nhận tiền và số tiền quyên góp. Mục tiêu của tác giả là hướng dẫn người đọc cách xử lý và phân tích một tập dữ liệu thực tế, quy mô lớn bằng pandas – từ việc đọc, lọc, đến trực quan hóa và rút ra thông tin ý nghĩa.

Các trường dữ liệu chính:

Bảng 6.1. Các trường dữ liệu trong dataset FEC 2012

|  |  |
| --- | --- |
| **Trường** | **Ý nghĩa** |
| cand\_nm | Tên ứng viên |
| contbr\_nm | Người đóng góp |
| contbr\_employer | Nơi làm việc |
| contbr\_occupation | Nghề nghiệp |
| contb\_receipt\_amt | Số tiền đóng góp |
| contb\_receipt\_dt | Ngày đóng góp |
| contbr\_st | Bang |

## Quy trình thực hiện

**Bước 1: Đọc dữ liệu**

fec = pd.read\_csv("P00000001-ALL.csv", low\_memory=False)

fec.info()

**Bước 2: Lọc dữ liệu hợp lệ**

fec = fec[fec["contb\_receipt\_amt"] > 0]

**Bước 3: Gắn nhãn đảng phái**

parties = {'Bachmann, Michelle': 'Republican',

           'Cain, Herman': 'Republican',

           'Gingrich, Newt': 'Republican',

           'Huntsman, Jon': 'Republican',

           'Johnson, Gary Earl': 'Republican',

           'McCotter, Thaddeus G': 'Republican',

           'Obama, Barack': 'Democrat',

           'Paul, Ron': 'Republican',

           'Pawlenty, Timothy': 'Republican',

           'Perry, Rick': 'Republican',

           "Roemer, Charles E. 'Buddy' III": 'Republican',

           'Romney, Mitt': 'Republican',

           'Santorum, Rick': 'Republican'}

fec.cand\_nm[123456:123461]

fec["party"] = fec["cand\_nm"].map(parties)

**Bước 4: Phân tích đóng góp theo nghề nghiệp**

by\_occupation = fec.pivot\_table("contb\_receipt\_amt",

                                index="contbr\_occupation",

                                columns="party", aggfunc="sum")

over\_2mm = by\_occupation[by\_occupation.sum(axis="columns") > 2000000]

**Bước 5: Binning theo mức đóng góp**

bins = np.array([0, 1, 10, 100, 1\_000, 10\_000, 100\_000, 1\_000\_000, 10\_000\_000])

labels = pd.cut(fec\_mrbo.contb\_receipt\_amt, bins)

**Bước 6: Nhóm theo bang**

grouped = fec\_mrbo.groupby(['cand\_nm', 'contbr\_st'])

totals = grouped.contb\_receipt\_amt.sum().unstack(0).fillna(0)

totals = totals[totals.sum(1) > 100\_000]

## Kết quả

### Phân tích quyên góp theo nghề nghiệp

A graph with numbers and lines

AI-generated content may be incorrect.

Hình 6.1. Biểu đồ tổng quyên góp theo nghề nghiệp và đảng phái

Dựa trên kết quả phân tích dữ liệu đóng góp cho chiến dịch bầu cử tổng thống Hoa Kỳ năm 2012, biểu đồ cung cấp một cái nhìn sâu sắc về mối tương quan giữa nghề nghiệp và khuynh hướng chính trị. Có thể thấy một sự phân hóa rõ ràng trong xu hướng ủng hộ tài chính giữa hai đảng. Các nghề nghiệp trong lĩnh vực học thuật và chuyên môn như Giáo sư (PROFESSOR) và Luật sư (ATTORNEY, LAWYER) có xu hướng ủng hộ Đảng Dân chủ mạnh mẽ, trong khi đó, giới doanh nhân và điều hành, bao gồm các Chủ tịch (PRESIDENT), Giám đốc điều hành (CEO, EXECUTIVE) và Nhà đầu tư (INVESTOR), lại thiên về Đảng Cộng hòa. Đặc biệt, nhóm Nội trợ (HOMEMAKER) nổi bật là nhóm đóng góp số tiền lớn nhất và nghiêng hẳn về phe Cộng hòa. Một điểm đáng chú ý khác là nhóm Người đã nghỉ hưu (RETIRED), với số tiền đóng góp khổng lồ cho cả hai đảng, phản ánh đây là một lực lượng ủng hộ tài chính quan trọng và tương đối cân bằng trong cử tri. Nhìn chung, biểu đồ không chỉ minh chứng cho sự khác biệt trong quan điểm chính trị gắn liền với đặc thù nghề nghiệp mà còn phác họa bức tranh sinh động về các nhóm cử tri then chốt mà mỗi đảng chính trị hướng đến.

### Phân tích quyên góp theo nhà tuyển dụng

A screenshot of a computer

AI-generated content may be incorrect.

Hình 6.2. Top 10 nhà tuyển dụng đóng góp nhiều nhất theo ứng viên

Đối với ứng viên Obama, nhóm đóng góp lớn nhất đến từ các cá nhân đã nghỉ hưu (RETIRED), tiếp theo là nhóm tự làm chủ (SELF-EMPLOYED) và những người không có thông tin về nghề nghiệp (NOT EMPLOYED).

Trong khi đó, ứng viên Romney cũng nhận được sự ủng hộ tài chính mạnh mẽ từ nhóm nghỉ hưu và nội trợ (HOMEMAKER), nhưng đặc biệt có sự tham gia của một số tổ chức tài chính lớn như Credit Suisse và Morgan Stanley trong top 10.

### Phân tích theo mức đóng góp

A graph with blue and orange bars

AI-generated content may be incorrect.

Hình 6.3. Tỷ lệ tổng số tiền quyên góp theo từng mức đóng góp

Biểu đồ hình 6.3 cho thấy một sự khác biệt rõ rệt và rất thú vị trong chiến lược gây quỹ giữa hai ứng viên Obama và Romney. Ở những khoản đóng góp nhỏ, dưới 1,000 USD, Tổng thống Obama có được sự ủng hộ áp đảo từ quần chúng, với tỷ lệ đóng góp trong các nhóm từ dưới 1 USD đến 1,000 USD đều chiếm trên 70%, thậm chí gần 92% ở nhóm 1-10 USD. Điều này phản ánh chiến lược vận động thành công dựa vào số đông cử tri với những khoản đóng góp nhỏ lẻ. Ngược lại, chiến dịch của ông Romney thể hiện sức mạnh ở phân khúc các nhà tài trợ lớn. Tỷ lệ đóng góp của ông Romney tăng dần theo quy mô khoản tiền và vượt trội hoàn toàn ở nhóm từ 1,000 đến 10,000 USD, chiếm tới 55%. Sự phân cực này không chỉ làm nổi bật sự khác biệt trong cơ sở ủng hộ mà còn cho thấy hai triết lý gây quỹ riêng biệt: một bên dựa vào sức mạnh của đám đông và một bên tập trung vào các nhà tài trợ hạng sang.

### Phân tích theo bang

A screenshot of a black screen

AI-generated content may be incorrect.

Hình 6.4. Tỷ lệ quyên góp theo bang

Dựa trên dữ liệu tỷ lệ quyên góp theo bang, có thể rút ra một số nhận định chiến lược quan trọng sau:

**1. Lợi thế áp đảo của Obama tại các "Pháo đài Dân chủ":**

Tổng thống Obama không chỉ dẫn đầu mà còn chiếm ưu thế tuyệt đối tại các bang vốn được xem là "căn cứ địa" truyền thống của Đảng Dân chủ. Điều này được minh chứng rõ ràng qua các con số: ông nhận được sự ủng hộ tài chính vượt trội tại Đặc khu Columbia (DC - 81%), Delaware (DE - 80.3%), và Arkansas (AR - 77.3%). Tỷ lệ cao tại DC và Delaware phản ánh sự ủng hộ mạnh mẽ từ các cử tri trong các cơ quan chính phủ liên bang và giới tài chính (do Delaware là nơi đăng ký của nhiều tập đoàn lớn).

**2. Romney thống trị tại các "Pháo đài Cộng hòa" và giành lợi thế tại các Bang Chiến trường then chốt:**

Ông Romney thể hiện sức mạnh không thể phủ nhận tại các bang theo truyền thống ủng hộ Đảng Cộng hòa. Ông nhận được tỷ lệ đóng góp áp đảo ở Arizona (AZ - 55.6%), quê nhà của ứng cử viên phó tổng thống của ông, Paul Ryan, vào thời điểm đó. Quan trọng hơn, ông đã giành được lợi thế tài chính tại một số bang chiến trường (Swing States/Battleground States) cực kỳ quan trọng. Việc dẫn trước ở Florida (FL - 53.3%), bang có số phiếu cử tri đoàn lớn, là một tín hiệu chiến lược rất tích cực cho chiến dịch của Romney, cho thấy sự ủng hộ từ các cử tri cao tuổi và cộng đồng doanh nhân.

**3. Sự cạnh tranh khốc liệt tại các Bang "Vàng":**

Bảng số liệu phơi bày những điểm cạnh tranh gay gắt nhất. Tại Alabama (AL), tỷ lệ ủng hộ tài chính cho hai ứng viên gần như sít sao (50.7% so với 49.3%), biến nó thành một "bang tím" (purple state) thực sự trên bản đồ tài chính, bất chấp khuynh hướng bầu cử truyền thống. Tương tự, Colorado (CO) với tỷ lệ 58.6% cho Obama so với 41.4% cho Romney, cho thấy đây là một chiến trường mà cả hai bên đều đã đầu tư nguồn lực đáng kể.

Phân tích này cho thấy không có sự ngẫu nhiên trong chiến lược vận động. Các chiến dịch đã có sự tập trung nguồn lực rất rõ ràng: củng cố lợi thế tại các bang "pháo đài" và dồn sức tranh giành các bang chiến trường then chốt, nơi mỗi phần trăm ủng hộ đều có ý nghĩa quyết định. Dữ liệu tài chính này là một chỉ báo quan trọng, phản chiếu và dự báo cho cuộc đua giành phiếu cử tri đoàn trong ngày bầu cử.

## Nhận xét

Quá trình phân tích dữ liệu đóng góp cho chiến dịch bầu cử tổng thống Hoa Kỳ năm 2012 đã mang lại những góc nhìn sâu sắc về cơ cấu và chiến lược gây quỹ của hai ứng viên chính: Barack Obama và Mitt Romney. Kết quả phân tích theo từng khía cạnh cho thấy một bức tranh tổng thể rất rõ nét.

Về mặt nghề nghiệp, dữ liệu chỉ ra sự phân hóa rõ ràng trong cơ sở ủng hộ. Ứng viên Obama nhận được sự ủng hộ tài chính vượt trội từ các ngành nghề tri thức và chuyên môn như Giáo sư và Luật sư. Ngược lại, ứng viên Romney lại được hưởng lợi từ sự ủng hộ mạnh mẽ của giới doanh nhân, điều hành cấp cao (CEO, Chủ tịch) và đặc biệt là nhóm Nội trợ. Sự khác biệt này càng được khẳng định khi phân tích theo nhà tuyển dụng, cho thấy nguồn lực của Obama đến từ đông đảo quần chúng và những cá nhân đã nghỉ hưu, trong khi Romney có được sự hậu thuẫn từ các tổ chức tài chính lớn.

Chiến lược gây quỹ của hai ứng viên cũng thể hiện sự khác biệt rõ rệt. Chiến dịch của Tổng thống Obama thành công vang dội trong việc huy động các khoản đóng góp nhỏ lẻ, thể hiện qua việc ông chiếm gần như tuyệt đối tỷ lệ đóng góp trong các nhóm dưới 1,000 USD. Điều này cho thấy một mô hình vận động tài chính dựa vào sức mạnh của số đông. Trái lại, ông Romney thể hiện ưu thế ở phân khúc các nhà tài trợ lớn, với tỷ lệ đóng góp vượt trội trong nhóm từ 1,000 đến 10,000 USD.

Trên bình diện địa lý, sự cạnh tranh càng trở nên rõ nét. Mặc dù ông Obama chiếm ưu thế tuyệt đối tại các bang có truyền thống ủng hộ Đảng Dân chủ, thì ông Romney lại không chỉ củng cố được các "pháo đài" Cộng hòa mà còn giành được lợi thế tài chính quan trọng tại các bang chiến trường then chốt như Florida và Arizona. Điều này dự báo một cuộc đua sít sao và cho thấy cả hai chiến dịch đều có những tính toán chiến lược rất rõ ràng trong việc phân bổ nguồn lực.

Tóm lại, dữ liệu FEC 2012 không chỉ phản ánh sự khác biệt về đặc điểm nhân khẩu học và địa lý của các nhóm ủng hộ, mà quan trọng hơn, nó đã làm sáng tỏ hai triết lý vận động và gây quỹ khác biệt, góp phần định hình diện mạo của một trong những cuộc bầu cử tổng thống được theo dõi sát sao nhất trong lịch sử hiện đại Hoa Kỳ.

# KẾT LUẬN

## Tổng kết quá trình và mục tiêu

Qua quá trình thực hiện báo cáo này, chúng em đã hoàn thành việc nghiên cứu và ứng dụng các kỹ thuật phân tích dữ liệu với Python vào 5 tập dữ liệu thực tế khác nhau, bao gồm: dữ liệu duyệt web từ Bitly (phân tích múi giờ và user agents), đánh giá phim từ MovieLens (1 triệu ratings), tên trẻ em Hoa Kỳ từ 1880-2010 (130 năm dữ liệu), cơ sở dữ liệu thực phẩm USDA (6,636 loại thực phẩm), và dữ liệu đóng góp chính trị FEC năm 2012 (hơn 1 triệu records). Thông qua các case study này, báo cáo đã minh họa một cách toàn diện quy trình phân tích dữ liệu hoàn chỉnh từ tải dữ liệu (JSON, CSV), làm sạch và xử lý dữ liệu (handling missing values, duplicates, data transformation), phân tích và aggregation (groupby, pivot\_table, merge), cho đến trực quan hóa (matplotlib, seaborn), nhằm trích xuất những thông tin và tri thức có giá trị từ dữ liệu thô.

## Ý nghĩa tổng hợp từ các phân tích

Tổng hợp từ 5 tập dữ liệu, có thể rút ra một số kết luận chung về sức mạnh của phân tích dữ liệu:

* Khám phá hành vi và xu hướng: Phân tích cho thấy khả năng làm sáng tỏ các mẫu hành vi phức tạp. Từ xu hướng đặt tên trẻ em thay đổi theo thời gian, sự khác biệt trong đánh giá phim theo giới tính, cho đến các mô hình duyệt web theo múi giờ, dữ liệu đã tiết lộ những câu chuyện và xu hướng văn hóa - xã hội sâu sắc.
* Hỗ trợ ra quyết định trong các lĩnh vực chuyên môn: Dữ liệu không chỉ mô tả quá khứ mà còn định hướng cho các quyết định trong tương lai. Việc xác định các thực phẩm giàu dinh dưỡng nhất trong cơ sở dữ liệu USDA có thể hỗ trợ cho các khuyến nghị về sức khỏe. Tương tự, việc phân tích các nhóm ủng hộ chính then chốt trong dữ liệu FEC giúp các chiến dịch chính trị phân bổ nguồn lực một cách tối ưu.
* Làm nổi bật sự khác biệt và đa dạng: Các kỹ thuật phân tích, đặc biệt là việc nhóm (groupby) và so sánh, đã liên tục làm nổi bật sự đa dạng và khác biệt trong dữ liệu – sự khác biệt về sở thích giữa các nhóm người dùng, sự đa dạng trong văn hóa đặt tên, hay sự phân cực trong khuynh hướng chính trị.

## Đánh giá về công cụ và kỹ thuật

Toàn bộ báo cáo là một minh chứng cho sức mạnh và tính linh hoạt của hệ sinh thái Python cho khoa học dữ liệu. Các thư viện chính như Pandas (cho thao tác dữ liệu cấp cao), NumPy (cho tính toán số học hiệu năng cao), và Matplotlib/Seaborn (cho trực quan hóa) đã tỏ ra cực kỳ hiệu quả trong việc xử lý các tập dữ liệu đa dạng, từ vài nghìn đến hàng triệu bản ghi. Các kỹ thuật then chốt được áp dụng xuyên suốt, chẳng hạn như groupby, pivot\_table, merge và lập chỉ mục boolean, đã trở thành công cụ nền tảng để khai thác và làm sáng tỏ ý nghĩa của dữ liệu.

## Hạn chế và hướng phát triển

Mặc dù đã đạt được các mục tiêu đề ra, báo cáo vẫn còn một số hạn chế cần được lưu ý. Một số phân tích, như xu hướng đặt tên, có thể được đào sâu hơn nữa bằng cách kết hợp với các dữ liệu kinh tế - xã hội khác để lý giải nguyên nhân. Đối với dữ liệu FEC, việc phân tích mạng lưới của các nhà đóng góp hoặc kết hợp với kết quả bầu cử theo bang có thể mang lại những hiểu biết phong phú hơn. Về mặt kỹ thuật, việc khám phá các thư viện máy học như Scikit-learn để dự đoán đánh giá phim hoặc phân cụm người dùng là một hướng phát triển đầy tiềm năng trong tương lai.

## Kết luận chung

Tóm lại, thông qua 5 case study cụ thể, báo cáo này đã khẳng định giá trị then chốt của tư duy phân tích dữ liệu trong việc giải quyết các vấn đề đa dạng, từ kinh doanh, xã hội học đến chính trị và sức khỏe cộng đồng. Kỷ nguyên dữ liệu lớn không chỉ nằm ở quy mô của dữ liệu mà còn ở khả năng biến chúng thành những hiểu biết có thể hành động được. Các kỹ năng và công cụ được trình bày trong báo cáo này chính là nền tảng để khai phá sức mạnh đó, mở ra vô số khả năng cho các phân tích sâu hơn và các ứng dụng thực tiễn trong tương lai.

# TÀI LIỆU THAM KHẢO

[1] W. McKinney, *Python for Data Analysis*, 3rd ed. Sebastopol, CA: O’Reilly Media, 2022.

[2] The Pandas Development Team, *pandas: Powerful Python Data Analysis Toolkit*, 2024.

[3] Matplotlib Developers, *Matplotlib: Visualization with Python*, 2024.

[4] Seaborn Developers, *Seaborn: Statistical Data Visualization*, 2024. [Online]. Available: [*https://seaborn.pydata.org*](https://seaborn.pydata.org)

[5] Python Software Foundation, *Python Language Reference, version 3.x*, 2024. [Online]. Available: [*https://www.python.org*](https://www.python.org)

[6] U.S. Social Security Administration, *National Data on Baby Names (1880–2010)*. [Online]. Available:[*https://www.ssa.gov/oact/babynames/*](https://www.ssa.gov/oact/babynames/)

[7] GroupLens Research, University of Minnesota, *MovieLens 1M Dataset*. [Online]. Available: [*https://grouplens.org/datasets/movielens/*](https://grouplens.org/datasets/movielens/)

[8] United States Department of Agriculture (USDA), *National Nutrient Database for Standard Reference*.

[9] Federal Election Commission (FEC), *Campaign Finance Data 2012*.