**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue circle with text

Description automatically generated****

**BÀI TẬP LỚN LAB 2**

**Danh sách sinh viên thực hiện:**

**3122410138 – Bùi Minh Huy**

**3122410128 – Phan Huy Hoàng**

**3112410336 – Hoàng Minh Quang**

**3122410106 – Nguyễn Trần Công Hậu**

**Giảng viên hướng dẫn: Đỗ Như Tài**

**Thành phố Hồ Chí Minh – Tháng 10/2025**

**MỤC LỤC**

[**LỜI MỞ ĐẦU 4**](#_Toc211513896)

[**Bối cảnh và tính cấp thiết của phân tích dữ liệu 4**](#_Toc211513897)

[**Mục tiêu và phạm vi báo cáo 4**](#_Toc211513898)

[**Tổng quan về các chuyên đề phân tích 5**](#_Toc211513899)

[**Cấu trúc báo cáo 5**](#_Toc211513900)

[**KẾ HOẠCH THỰC HIỆN 6**](#_Toc211513901)

[**CHƯƠNG 1. PHÂN TÍCH DỮ LIỆU THÔNG TIN TRUY CẬP URL RÚT GỌN 7**](#_Toc211513902)

[**1. Định nghĩa vấn đề và mục tiêu phân tích 7**](#_Toc211513903)

[**1.1. Bối cảnh vấn đề 7**](#_Toc211513904)

[**1.2. Mục tiêu phân tích 7**](#_Toc211513905)

[**1.3. Giới thiệu tổng quan về Dataset 7**](#_Toc211513906)

[**2. Hiển thị thông tin tổng quan về dữ liệu 9**](#_Toc211513907)

[**3. Phân tích khám phá và trực quan hóa dữ liệu 10**](#_Toc211513908)

[**4. Kết luận 14**](#_Toc211513909)

[**CHƯƠNG 2. PHÂN TÍCH BỘ DỮ LIỆU MOVIELENS 16**](#_Toc211513910)

[**1. Định nghĩa vấn đề và mục tiêu phân tích 16**](#_Toc211513911)

[**1.1. Bối cảnh vấn đề 16**](#_Toc211513912)

[**1.2. Mục tiêu phân tích 16**](#_Toc211513913)

[**1.3. Giới thiệu tổng quan về Dataset 16**](#_Toc211513914)

[**2. Hiển thị thông tin tổng quan về dữ liệu 17**](#_Toc211513915)

[**3. Phân tích khám phá và trực quan hóa dữ liệu 17**](#_Toc211513916)

[**4. Kết luận 21**](#_Toc211513917)

[**CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU THỰC PHẨM USDA 22**](#_Toc211513918)

[**1. Định nghĩa vấn đề và mục tiêu phân tích 22**](#_Toc211513919)

[**1.1. Bối cảnh vấn đề 22**](#_Toc211513920)

[**1.2. Mục tiêu phân tích 22**](#_Toc211513921)

[**1.3. Giới thiệu tổng quan về Dataset 23**](#_Toc211513922)

[**2. Mô tả thống kê và đánh giá dữ liệu 24**](#_Toc211513923)

[**3. Phân tích khám phá và trực quan hóa dữ liệu 24**](#_Toc211513924)

[**3.1. Phân tích đơn biến (Univariate Analysis) 24**](#_Toc211513925)

[**3.2. Phân tích đa biến (Multivariate Analysis) 26**](#_Toc211513926)

[**4. Kết luận 28**](#_Toc211513927)

[**CHƯƠNG 4. PHÂN TÍCH DỮ LIỆU VỀ TÊN BÉ TRAI VÀ BÉ GÁI Ở HOA KÌ (1880-2010) 30**](#_Toc211513928)

[**1. Định nghĩa vấn đề và mục tiêu phân tích 30**](#_Toc211513929)

[**1.1. Bối cảnh vấn đề 30**](#_Toc211513930)

[**1.2. Mục tiêu phân tích 30**](#_Toc211513931)

[**1.3. Giới thiệu tổng quan về Dataset 31**](#_Toc211513932)

[**2. Hiển thị thông tin tổng quan về dữ liệu 31**](#_Toc211513933)

[**3. Phân tích khám phá và trực quan hóa dữ liệu 32**](#_Toc211513934)

[**4. Kết luận 39**](#_Toc211513935)

[**CHƯƠNG 5. PHÂN TÍCH CƠ SỞ DỮ LIỆU ỦY BAN BẦU CỬ LIÊN BANG (FEC) 2012 41**](#_Toc211513936)

[**1. Bối cảnh vấn đề và mục tiêu phân tích 41**](#_Toc211513937)

[**1.1. Bối cảnh vấn đề 41**](#_Toc211513938)

[**1.2. Mục tiêu phân tích 41**](#_Toc211513939)

[**1.3. Giới thiệu tổng quan về Dataset 42**](#_Toc211513940)

[**2. Mô tả thống kê và đánh giá dữ liệu 43**](#_Toc211513941)

[**2.1. Hiển thị thông tin tổng quan về dữ liệu 43**](#_Toc211513942)

[**2.2. Hiển thị một dòng dữ liệu cụ thể 44**](#_Toc211513943)

[**2.3. Hiển thị danh sách các ứng viên 45**](#_Toc211513944)

[**2.4. Chuẩn bị biến phân loại (Đảng phái) 46**](#_Toc211513945)

[**3. Phân tích khám phá và trực quan hóa dữ liệu 46**](#_Toc211513946)

[**3.1. Phân tích đơn biến (Univariate Analysis) 46**](#_Toc211513947)

[**3.2. Phân tích đa biến (Multivariate Analysis) 49**](#_Toc211513948)

[**4. Kết luận 57**](#_Toc211513949)

[**KẾT LUẬN CHUNG 59**](#_Toc211513950)

# LỜI MỞ ĐẦU

## Bối cảnh và tính cấp thiết của phân tích dữ liệu

Trong kỷ nguyên số hiện nay, dữ liệu không chỉ đơn thuần là thông tin mà đã trở thành một nguồn tài nguyên chiến lược, đóng vai trò then chốt trong việc ra quyết định và định hướng chiến lược. Khả năng khai thác, phân tích và chuyển hóa dữ liệu thô thành những hiểu biết có giá trị là yếu tố then chốt giúp các tổ chức duy trì lợi thế cạnh tranh. Sự chuyển dịch từ các quyết định dựa trên kinh nghiệm sang các quyết định dựa trên bằng chứng dữ liệu đang định hình lại mọi mặt của hoạt động kinh doanh, từ phát triển sản phẩm, tối ưu hóa trải nghiệm người dùng, đến việc đánh giá hiệu quả của các chiến dịch tiếp thị.

Báo cáo này tiếp nối tinh thần đó, trình bày kết quả của việc áp dụng các kỹ thuật phân tích dữ liệu vào nhiều bộ dữ liệu thực tế khác nhau, nhằm minh họa cho sức mạnh và tính linh hoạt của các công cụ phân tích trong việc giải quyết các bài toán đa dạng.

## Mục tiêu và phạm vi báo cáo

Báo cáo này không tập trung vào một câu hỏi nghiên cứu duy nhất, mà thay vào đó, nhằm mục đích:

* **Minh họa thực tiễn:** Trình bày các bước xử lý và phân tích dữ liệu cụ thể trên những tập dữ liệu có cấu trúc và nguồn gốc khác nhau.
* **Đa dạng hóa kỹ thuật:** Áp dụng nhiều kỹ thuật phân tích, từ các thao tác cơ bản với thư viện chuẩn của Python đến các phân tích nâng cao với pandas, NumPy, và trực quan hóa với matplotlib và seaborn.
* **Khám phá thông tin chuyên sâu:** Mỗi tập dữ liệu được phân tích để trả lời những câu hỏi riêng biệt, từ phân bổ địa lý, hành vi người dùng, đến xu hướng xã hội và dinh dưỡng.

Mỗi phân tích trong báo cáo là một chuyên đề độc lập, với bối cảnh, dữ liệu đầu vào, phương pháp luận và kết luận riêng.

## Tổng quan về các chuyên đề phân tích

Báo cáo Lab 02 sẽ lần lượt đi qua các chuyên đề phân tích chính sau:

1. Phân tích Dữ liệu Bitly từ 1.USA.gov: Khám phá hành vi người dùng internet thông qua dữ liệu ẩn danh từ dịch vụ rút gọn liên kết, tập trung vào phân bổ múi giờ và thông tin trình duyệt.
2. Phân tích Bộ dữ liệu MovieLens 1M: Phân tích xu hướng đánh giá phim, sự khác biệt trong đánh giá theo nhân khẩu học và khám phá mối quan hệ giữa thể loại phim và độ tuổi người xem.
3. Phân tích Dữ liệu Tên Bé trai/Bé gái ở Mỹ (1880–2010): Nghiên cứu xu hướng đặt tên theo thời gian, sự đa dạng trong lựa chọn tên, và sự biến đổi của các đặc điểm tên (như chữ cái cuối cùng).
4. Phân tích Cơ sở dữ liệu Thực phẩm USDA: Khám phá thành phần dinh dưỡng của các nhóm thực phẩm khác nhau và xác định các thực phẩm giàu dinh dưỡng nhất cho từng nhóm chất.
5. Phân tích Cơ sở dữ liệu Ủy ban Bầu cử Liên bang (FEC) 2012: Phân tích mô hình đóng góp tài chính cho chiến dịch bầu cử, bao gồm sự khác biệt theo nghề nghiệp, nhà tuyển dụng, bang và quy mô đóng góp.

## Cấu trúc báo cáo

Để đảm bảo tính rõ ràng và mạch lạc, báo cáo được tổ chức thành các phần riêng biệt tương ứng với từng tập dữ liệu. Sau phần mở đầu này, mỗi chuyên đề phân tích sẽ được trình bày tuần tự, mỗi chuyên đề đóng vai trò như một báo cáo con hoàn chỉnh với:

* Giới thiệu về bối cảnh và tập dữ liệu.
* Quy trình xử lý và phân tích dữ liệu với các đoạn mã và giải thích cụ thể.
* Trình bày kết quả kèm theo biểu đồ trực quan.
* Thảo luận và kết luận cho từng phân tích.

# KẾ HOẠCH THỰC HIỆN

Bảng phân công công việc nhóm:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** | **Công việc** |
| 1 | Bùi Minh Huy | 3122410138 | - Phân tích Cơ sở dữ liệu Ủy ban Bầu cử Liên bang (FEC) 2012.  - Làm Word và PPT trên Dataset tương ứng.  - Tổng hợp Word và điều chỉnh |
| 2 | Phan Huy Hoàng | 3122410128 | - Phân tích Dữ liệu Bitly từ 1.USA.gov và Bộ dữ liệu MovieLens 1M  - Làm Word và PPT trên 2 Dataset tương ứng. |
| 3 | Hoàng Minh Quang | 3112410336 | - Phân tích Cơ sở dữ liệu Thực phẩm USDA  - Làm Word và PPT trên Dataset tương ứng. |
| 4 | Nguyễn Trần Công Hậu | 3122410106 | - Phân tích Dữ liệu Tên Bé trai/Bé gái ở Mỹ (1880–2010).  - Làm Word và PPT trên Dataset tương ứng. |

# CHƯƠNG 1. PHÂN TÍCH DỮ LIỆU THÔNG TIN TRUY CẬP URL RÚT GỌN

## 1. Định nghĩa vấn đề và mục tiêu phân tích

### 1.1. Bối cảnh vấn đề

Trong thời đại công nghệ số, việc chia sẻ thông tin qua Internet diễn ra với tốc độ và quy mô ngày càng lớn. Các liên kết rút gọn (shortened URLs) như *bit.ly*, *tinyurl.com*, hay *1.usa.gov* ra đời nhằm giúp người dùng chia sẻ các đường dẫn dài một cách tiện lợi hơn, đặc biệt trên mạng xã hội, email, và các nền tảng truyền thông trực tuyến Với nguồn dữ liệu công khai về tên trẻ em do **Cơ quan An sinh Xã hội Hoa Kỳ (SSA)** cung cấp trong giai đoạn **1880–2010,** việc phân tích xu hướng đặt tên trở thành một hướng nghiên cứu thú vị giúp khám phá sâu hơn về biến động xã hội và tâm lý cộng đồng người Mỹ.

Trong nghiên cứu này, việc phân tích dữ liệu thông tin truy cập các URL rút gọn từ dịch vụ 1.usa.gov được sử dụng như một ví dụ thực tế để khám phá hành vi người dùng trên quy mô toàn cầu. Dữ liệu phản ánh các lượt truy cập đến những trang web của chính phủ Hoa Kỳ, cho phép ta hiểu rõ hơn về mức độ quan tâm của công chúng, cũng như hiệu quả của hoạt động truyền thông số trong lĩnh vực công.

### 1.2. Mục tiêu phân tích

* Phân tích số lượng người dùng truy cập ở những múi giờ khác nhau.
* Phân tích số lượng app/website truy cập vào bit.ly.
* Phân tích số lượng truy cập theo từng múi giờ, được phân loại theo hệ điều hành.

### 1.3. Giới thiệu tổng quan về Dataset

Dữ liệu vào:

* a: Chuỗi User-Agent – thông tin về trình duyệt, hệ điều hành hoặc ứng dụng người dùng. (Object)
* al: Ngôn ngữ người dùng (accept-language). (Object)
* c: Mã quốc gia (ISO). (Object)
* cy: Thành phố người truy cập (Object)
* gr: Bang hoặc tỉnh. (Object)
* ll: Tọa độ địa lý . (Object)
* tz: Múi giờ của người dùng (Object)
* nk: Số lượng click (thường là 0 hoặc 1). (float64)
* g: mã rút gọn toàn cục của liên kết Bitly. (Object)
* h: mã rút gọn riêng do người dùng cụ thể tạo ra. (Object)
* l: tên tài khoản (username) của người dùng Bitly đã tạo link. (Object)
* hh: domain rút gọn mà liên kết thuộc về. (Object)
* r: Referrer – nguồn truy cập (Facebook, Twitter, direct,...). (Object)
* u: URL gốc (đích đến thật). (Object)
* t: Thời gian click (Unix timestamp). (float64)
* hc: Thời gian tạo link rút gọn (float64)
* kw: Từ khóa (nếu có).(Object)
* heartbeat: Bản ghi kỹ thuật của hệ thống (dùng theo dõi trạng thái).(float64)

## 2. Hiển thị thông tin tổng quan về dữ liệu

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer code

AI-generated content may be incorrect.

Hình 2.1. Hiển thị số dòng, số cột, kiểu dữ liệu của từng cột

Thông tin cơ bản của bộ dữ liệu:

* Dữ liệu có 18 tính chất (cột) để phân tích, bao gồm: a, c, nk, tz, gr, g, h, l, al, hh, r, u, t, hc, cy, ll, heartbeat, kw
* Các giá trị thuộc đa số thuộc kiểu object (chuỗi/text), 4 kiểu (float64).
* Tổng số dòng dữ liệu là 3560 bản ghi (tổng số lần click).

## 3. Phân tích khám phá và trực quan hóa dữ liệu

***Câu hỏi 1:*** Múi giờ nào có số lượt truy cập cao nhất trong dữ liệu? Các múi giờ khác có mức độ truy cập ra sao?

A blue and white graph

AI-generated content may be incorrect.

Hình 3.1. Biểu đồ thống kê số lượt truy cập

Nhìn vào biểu đồ ta có nhận xét:

* Người dùng ở miền Đông Hoa Kỳ (EST) chiếm phần lớn lưu lượng truy cập. Điều này hợp lý vì đây là khu vực dân cư đông, bao gồm New York, Washington D.C., Philadelphia…
* Dữ liệu thể hiện rõ mức độ tập trung người dùng ở Bắc Mỹ, đặc biệt là Hoa Kỳ.
* Rất ít người truy cập từ múi giờ ngoài nước Mỹ.

***Câu hỏi 2:*** Trình duyệt / ứng dụng nào được sử dụng nhiều nhất để truy cập các link Bitly? Điều này giúp ích gì cho nhà phân tích / doanh nghiệp?

A screen shot of a graph

AI-generated content may be incorrect.

Hình 3.2. Số lượng trình duyệt/ứng dụng truy cập

Khi quan sát biểu đồ số ta nhận thấy:

* Biểu đồ trên thể hiện 10 trình duyệt và ứng dụng được sử dụng nhiều nhất để truy cập các đường dẫn rút gọn của Bitly. Kết quả cho thấy Mozilla/5.0 chiếm tỷ lệ vượt trội, với số lượt truy cập cao gấp nhiều lần so với các nhóm còn lại. Điều này phản ánh rằng phần lớn người dùng Bitly truy cập thông qua các trình duyệt web hiện đại (như Chrome, Firefox, Safari, Edge…), bởi hầu hết các trình duyệt này đều khai báo chuỗi nhận dạng bắt đầu bằng "Mozilla/5.0" trong user-agent.
* Dựa vào biểu đồ và số liệu ở trên các nhà phân tích / doanh nghiệp biết được nền tảng truy cập phổ biến nhất từ đó giúp tối ưu trải nghiệm người dùng (UX). Nếu phần lớn người dùng đến từ “Mozilla/5.0” (trình duyệt hiện đại), ta có thể ưu tiên hỗ trợ các trình duyệt đó.

Tiếp theo chúng ta sẽ đi sâu vào việc người dùng truy cập vào bằng hệ điều hành windows hay hệ điều hành khác. Câu hỏi đặt ra: ***Múi giờ nào có lượng truy cập (click) cao nhất? Người dùng hệ điều hành nào chiếm ưu thế trong từng múi giờ? Có sự khác biệt rõ rệt giữa người dùng Windows và Not Windows không?***

A graph of a bar graph

AI-generated content may be incorrect.Hình 3.3. Hình trên sắp xếp top 10 múi giờ có tổng lượt truy cập cao nhất

(Windows + Non-Windows).

Nhận xét:

* Biểu đồ cho thấy America/New\_York là múi giờ có tổng số lượt truy cập cao nhất, tiếp theo là America/Chicago và America/Los\_Angeles.
* Trong hầu hết các múi giờ, người dùng Not Windows (Mac, Linux, iPhone, Android, v.v.)có xu hướng chiếm tỷ lệ cao hơn so với người dùng Windows.
* Điều này cho thấy Bitly được sử dụng phổ biến trên thiết bị di động hoặc hệ điều hành khác ngoài Windows.
* Các múi giờ thuộc Bắc Mỹ có hoạt động mạnh mẽ nhất, phản ánh Bitly là nền tảng phổ biến tại Hoa Kỳ.
* Các múi giờ khác (như châu Âu, châu Á) xuất hiện ít hơn, cho thấy phạm vi người dùng toàn cầu nhưng không đồng đều.

Biểu đồ cho thấy phần lớn lượt truy cập Bitly tập trung tại các múi giờ của Hoa Kỳ, đặc biệt là America/New\_York, Chicago và Los\_Angeles. Trong đó, nhóm người dùng Not Windows (tức người dùng Mac, iPhone, Android, Linux,...) chiếm ưu thế rõ rệt so với Windows. Kết quả này phản ánh xu hướng người dùng Internet hiện đại sử dụng nhiều thiết bị di động và hệ điều hành đa dạng, đồng thời khẳng định Bitly là nền tảng phổ biến mạnh nhất tại khu vực Bắc Mỹ.

Tiếp theo, ta sẽ chuẩn hóa dữ liệu (normalize) để so sánh tỷ lệ tương đối. Dựa vào dữ liệu sau khi được chuẩn hóa ta có câu hỏi sau: ***Múi giờ nào có tỷ lệ người dùng cao nhất trong toàn bộ hệ thống Bitly? Sau khi chuẩn hóa thì liệu tỉ lệ người dùng trên windows và not windows liệu có thay đổi không? Xu hướng phân bố hệ điều hành theo khu vực địa lý có gì đặc biệt?***

A graph of a bar chart

AI-generated content may be incorrect.

Hình 3.4. Số liệu chuẩn hóa top 10 múi giờ truy cập

Nhận xét:

* Biểu đồ cho thấy America/New\_York có tỷ lệ người dùng cao nhất trong toàn bộ dữ liệu — đây là múi giờ hoạt động mạnh nhất.
* Các múi giờ tiếp theo như America/Chicago và America/Los\_Angeles cũng có tỷ lệ đáng kể, phản ánh mức độ phổ biến của Bitly tại Bắc Mỹ.
* Trong phần lớn các múi giờ, người dùng Windows chiếm tỷ lệ cao hơn người không dùng Windows. Thể hiện rõ ràng nhất ở Asia/Tokyo và Pacific/Honolulu.
* Điều này cho thấy Bitly được sử dụng phổ biến trên thiết bị Windows.
* Các múi giờ thuộc Bắc Mỹ và châu Á có sự chênh lệch rõ rệt về tỷ lệ Not Windows cao.
* Trong khi đó, một số múi giờ khác có xu hướng cân bằng hơn giữa hai nhóm người dùng.

Kết luận: Biểu đồ thể hiện rằng phần lớn hoạt động của người dùng Bitly tập trung tại các múi giờ thuộc Hoa Kỳ, đặc biệt là America/New\_York. Sau khi chuẩn hóa, ta nhận thấy nhóm người dùng Not Windows chiếm ưu thế trong hầu hết các múi giờ, phản ánh xu hướng người dùng truy cập từ các thiết bị di động và hệ điều hành khác ngoài Windows. Kết quả này cho thấy Bitly có lượng người dùng đa dạng, trải rộng theo địa lý và nền tảng công nghệ.

## 4. Kết luận

Qua quá trình phân tích tập dữ liệu log truy cập từ hệ thống rút gọn liên kết (1.usa.gov), ta rút ra được một số kết luận chính như sau:

* **Đặc điểm người dùng và khu vực truy cập**: hần lớn người truy cập đến từ Hoa Kỳ, chiếm tỷ lệ cao nhất trong toàn bộ dữ liệu. Bên cạnh đó còn có người dùng từ nhiều quốc gia khác nhau cho thấy tính toàn cầu của hệ thống chia sẻ liên kết này**.**
* **Thiết bị và trình duyệt sử dụng:** Người dùng chủ yếu sử dụng trình duyệt Chrome và Firefox trên máy tính, ngoài ra có một lượng đáng kể từ thiết bị di động như iPhone, iPad, Android, BlackBerry. Xu hướng truy cập qua thiết bị di động bắt đầu tăng lên rõ rệt
* **Số lượng truy cập theo từng múi giờ, được phân loại theo hệ điều hành:** Số lượng truy cập tập trung chủ yếu ở múi giờ Hoa Kỳ, đặc biệt là “America/New\_York”. Hệ điều hành phổ biến nhất là Windows, tiếp theo là Mac OS và iOS. Người dùng tại châu Mỹ truy cập chủ yếu bằng máy tính, trong khi người dùng châu Á và châu Âu có xu hướng truy cập bằng thiết bị di động.
* **Tương quan giữa múi giờ và hệ điều hành:** Ở các múi giờ Mỹ (America/New\_York, America/Los\_Angeles), Windows và Mac OS chiếm ưu thế. Ở các múi giờ Châu Á và Châu Âu, tỷ lệ iOS và Android cao hơn, do người dùng ở đây thường truy cập qua thiết bị di động.Kết quả này gợi ý rằng người dùng ở khu vực khác nhau có thói quen và thiết bị truy cập khác biệt rõ rệt.

# CHƯƠNG 2. PHÂN TÍCH BỘ DỮ LIỆU MOVIELENS

## 1. Định nghĩa vấn đề và mục tiêu phân tích

### 1.1. Bối cảnh vấn đề

Trong kỷ nguyên số, khi khối lượng nội dung giải trí trực tuyến ngày càng phong phú, người dùng thường gặp khó khăn trong việc lựa chọn bộ phim phù hợp với sở thích của mình. Các nền tảng xem phim và dịch vụ trực tuyến như Netflix, Amazon Prime, hoặc Disney+ đều dựa vào các hệ thống gợi ý (Recommendation Systems) để cá nhân hóa trải nghiệm người dùng, giúp họ tìm thấy những nội dung yêu thích một cách nhanh chóng và hiệu quả.

Việc phân tích dữ liệu MovieLens không chỉ giúp hiểu rõ hơn về hành vi và thị hiếu người xem phim, mà còn là bước khởi đầu quan trọng trong việc phát triển các hệ thống gợi ý thông minh – một ứng dụng tiêu biểu của khoa học dữ liệu và học máy (machine learning) trong đời sống hiện đại.

### 1.2. Mục tiêu phân tích

Chuyên đề phân tích này được thực hiện nhằm thực hiện các mục tiêu sau:

* Sự phân hóa sở thích theo Giới tính
* Sự thay đổi thị hiếu theo Độ tuổi
* Xu hướng Đánh giá Phim Nghệ thuật/Kinh điển

### 1.3. Giới thiệu tổng quan về Dataset

Dữ liệu vào:

* UserID: mã người dùng (int64)
* Gender: giới tính (Object)
* Age: độ tuổi (được mã hóa theo nhóm tuổi) (int64)
* Occupation: mã nghề nghiệp (int64)
* Zip-code: mã vùng sinh sống (Object)
* MovieID: mã phim (int64)
* Title: tên phim (bao gồm năm phát hành) (Object)
* Genres: thể loại (có thể chứa nhiều thể loại, ngăn cách bằng ký tự |) (Object)
* Rating: điểm đánh giá (từ 1–5) (float64)
* Timestamp : thời điểm đánh giá (dạng UNIX time) (float64)

## 2. Hiển thị thông tin tổng quan về dữ liệu

A screenshot of a computer code

AI-generated content may be incorrect.

Hình 2.1. Hiển thị số dòng, số cột, kiểu dữ liệu của từng cột

Thông tin cơ bản của bộ dữ liệu:

* Dữ liệu có 10 tính chất (cột) để phân tích, bao gồm: user\_id, movie\_id, timestamp, gender, age, occupation, zip, title, genres.
* Các giá trị thuộc đa số thuộc kiểu object (chuỗi/text), kiểu float64, kiểu float64
* Tổng số dòng dữ liệu là 260831 bản ghi (tổng số lần đánh giá).

## 3. Phân tích khám phá và trực quan hóa dữ liệu

***Câu hỏi 1:*** Những phim nào được nữ giới đánh giá cao nhất? Thể loại phim yêu thích của phụ nữ?

A graph with red and white stripes

AI-generated content may be incorrect.

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

Nhận xét:

* Biểu đồ cho thấy 10 phim được nữ giới đánh giá cao nhất có điểm trung bình rất cao (thường trên 4.0).
* Các phim này chủ yếu thuộc thể loại tâm lý – tình cảm – lãng mạn, phù hợp với xu hướng thị hiếu của nữ giới trong bộ dữ liệu MovieLens.

Thế còn liệu điểm của nam giới có khác biệt gì so với nữ giới hay không. Chúng ta sẽ cùng đi vào phân tích Câu hỏi: ***Những phim nào được nam giới đánh giá cao vượt trội so với nữ giới? Có thể lý giải sự khác biệt này dựa trên sở thích hoặc đặc điểm nhân khẩu học không?***

A graph with blue and white stripes

AI-generated content may be incorrect.

Hình 3.2. Số lượng trẻ sinh theo năm của các tên phổ biến

Khi quan sát biểu đồ ta nhận thấy:

* Biểu đồ cho thấy top 10 phim nam chấm điểm cao hơn nữ thường có độ chênh rõ rệt (từ 0.4 đến 0.8 điểm).
* Các phim nam yêu thích hơn nữ thường thuộc thể loại hành động, chiến tranh, khoa học viễn tưởng.Những thể loại này thường có nội dung mạnh mẽ, kỹ xảo, xung đột, chiến đấu,nên thu hút đối tượng người xem nam giới nhiều hơn.

Bây giờ chúng ta sẽ đi sâu vào đánh giá của từng thể loại phim,Câu hỏi: ***Có mối liên hệ nào giữa thể loại phim và độ tranh cãi không?? Vì sao các phim này lại gây ra sự khác biệt trong cảm nhận người xem? Điều này có ảnh hưởng gì đến độ lệch chuẩn không***

A graph with orange and white lines

AI-generated content may be incorrect.

Hình 3.3 Top 10 bộ phim có độ lệch chuẩn cao

Nhận xét:

* Phần lớn các phim trong danh sách thuộc thể loại tâm lý, nghệ thuật, hoặc siêu thực. Những phim này thường có nội dung phức tạp, nhịp chậm hoặc mang tính ẩn dụ, nên người xem đại chúng có thể thấy nhàm chán, còn người yêu điện ảnh lại đánh giá cao.
* Điều này khiến cho điểm đánh giá trải rộng làm độ lệch chuẩn tăng.

Tiếp theo, chúng ta sẽ phân tích điểm đánh giá của từng thể loại phim theo độ tuổi, câu hỏi: ***Điểm đánh giá có sự thay đổi như thế nào giữa các nhóm tuổi?Có xu hướng gì trong điểm đánh giá theo nhóm tuổi không?***

A table with numbers and symbols

AI-generated content may be incorrect.

Hình 3.4. Điểm trung bình của từng thể loại phim theo nhóm tuổi

Nhận xét:

* Phần lớn các phim trong danh sách thuộc thể loại tâm lý, nghệ thuật, hoặc siêu thực. Những phim này thường có nội dung phức tạp, nhịp chậm hoặc mang tính ẩn dụ, nên người xem đại chúng có thể thấy nhàm chán, còn người yêu điện ảnh lại đánh giá cao.
* Người trẻ tuổi thường thích các thể loại mang tính giải trí, hoạt hình hoặc đơn giản như Animation, Children's, Musical.
* Người lớn tuổi thường đánh giá cao các thể loại có chiều sâu nội dung hoặc mang tính hiện thực cao như Documentary, Drama, War.
* Một số thể loại như Horror có điểm thấp nhất ở mọi độ tuổi, nhưng đặc biệt thấp ở người trẻ (có thể do yếu tố sợ hãi hoặc thiếu chiều sâu nội dung).
* Điểm trung bình nhìn chung có xu hướng cao hơn ở nhóm tuổi từ 45 trở lên, cho thấy nhóm này có xu hướng đánh giá phim cao hơn hoặc chọn lọc kỹ hơn khi xem.

## 4. Kết luận

Qua quá trình phân tích bộ dữ liệu MovieLens , ta rút ra được một số kết luận chính như sau:

* Sự phân hóa sở thích theo Giới tính: Nữ giới có xu hướng đánh giá cao các phim tập trung vào cảm xúc, mối quan hệ và sự tinh tế trong kịch bản (ví dụ: Lãng mạn, Bi kịch, Hài kịch Tinh tế). Nam giới có xu hướng đánh giá cao các phim mang tính hành động, quy mô lớn, và thể loại nặng về kỹ xảo hoặc lịch sử (ví dụ: Hành động, Chiến tranh, Khoa học Viễn tưởng).
* Sự thay đổi thị hiếu theo Độ tuổi: Người dùng trẻ tuổi thiên về các thể loại giải trí đơn giản, sống động và giàu trí tưởng tượng (ví dụ: Hoạt hình, Thiếu nhi, Âm nhạc). Người dùng lớn tuổi có xu hướng đánh giá cao các thể loại có chiều sâu nội dung, mang tính hiện thực hoặc có giá trị nghệ thuật cao (ví dụ: Phim Tài liệu, Chính kịch, Chiến tranh).
* Xu hướng Đánh giá Phim Nghệ thuật/Kinh điển: Các phim được xếp hạng cao nhất thường là những tác phẩm kinh điển, siêu thực hoặc mang tính nghệ thuật phức tạp. Những phim này không được lòng đại đa số khán giả (có thể bị xem là nhàm chán hoặc khó hiểu), nhưng lại được nhóm người yêu điện ảnh đánh giá rất cao, làm tăng điểm trung bình tổng thể.

# CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU THỰC PHẨM USDA

## 1. Định nghĩa vấn đề và mục tiêu phân tích

### 1.1. Bối cảnh vấn đề

Dinh dưỡng là yếu tố cốt lõi quyết định sức khỏe và chất lượng cuộc sống của con người. Ở Hoa Kỳ, việc hiểu rõ thành phần dinh dưỡng của thực phẩm không chỉ giúp người tiêu dùng đưa ra lựa chọn ăn uống lành mạnh mà còn hỗ trợ các chuyên gia y tế, nhà nghiên cứu và nhà hoạch định chính sách trong việc cải thiện sức khỏe cộng đồng. Cơ sở dữ liệu dinh dưỡng thực phẩm của Bộ Nông nghiệp Hoa Kỳ (USDA) được công nhận là một trong những nguồn thông tin dinh dưỡng toàn diện và đáng tin cậy nhất.

Với nguồn dữ liệu đa chiều từ USDA, bao gồm 6,636 loại thực phẩm với 375,176 bản ghi chất dinh dưỡng sau khi xử lý, việc phân tích xu hướng dinh dưỡng trở thành một hướng nghiên cứu quan trọng giúp khám phá sâu hơn về mối quan hệ giữa thực phẩm và sức khỏe. Nghiên cứu này tập trung vào việc phân tích sự phân bổ dinh dưỡng across các nhóm thực phẩm, so sánh giá trị dinh dưỡng giữa các thực phẩm, khám phá mối quan hệ giữa các chất dinh dưỡng, và xu hướng dinh dưỡng theo nhóm thực phẩm - từ đó rút ra những hiểu biết về ảnh hưởng của chế độ ăn uống đến sức khỏe cộng đồng.

### 1.2. Mục tiêu phân tích

Chuyên đề phân tích này được thực hiện nhằm đạt được các mục tiêu chính sau:

* Khám phá và trực quan hóa sự phân bổ dinh dưỡng across các nhóm thực phẩm chính.
* Đánh giá và so sánh giá trị dinh dưỡng giữa các loại thực phẩm khác nhau.
* Phân tích mối quan hệ tương quan giữa các chất dinh dưỡng trong thực phẩm.
* Nghiên cứu xu hướng dinh dưỡng theo từng nhóm thực phẩm và nhận diện các thực phẩm có giá trị dinh dưỡng đặc biệt.

### 1.3. Giới thiệu tổng quan về Dataset

a) Mô tả

* Phạm vi: Cơ sở dữ liệu dinh dưỡng thực phẩm toàn diện của Bộ Nông nghiệp Hoa Kỳ (USDA)
* Kích thước: 6,636 loại thực phẩm với 375,176 bản ghi chất dinh dưỡng sau khi xử lý
* Đặc điểm: Dữ liệu đa chiều về thành phần dinh dưỡng, định lượng phần ăn, và phân loại thực phẩm

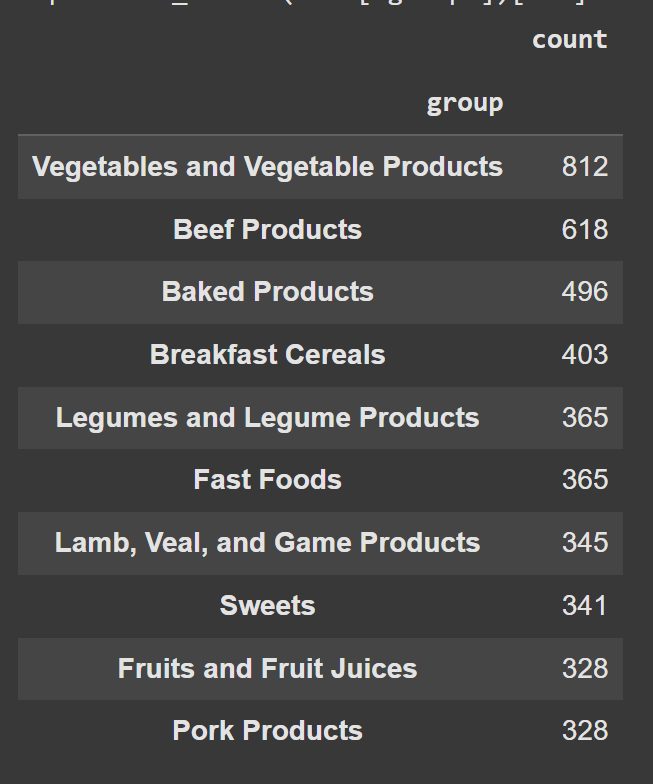
b) Dữ liệu vào

* Thông tin cơ bản (Info DataFrame):
  + id: ID duy nhất của thực phẩm (Int64)
  + description: Tên đầy đủ của thực phẩm (Object)
  + group: Nhóm thực phẩm chính (Object)
  + manufacturer: Nhà sản xuất (Object)
  + tags: Thẻ phân loại (Object)
* Thông tin dinh dưỡng (Nutrients DataFrame):
  + id: ID thực phẩm liên kết (Int64)
  + value: Giá trị dinh dưỡng (Float64)
  + units: Đơn vị đo lường (Object)
  + description: Tên chất dinh dưỡng (Object)
  + group: Nhóm chất dinh dưỡng (Object)
* Thông tin định lượng (Portions):
  + amount: Số lượng phần ăn (Float64)
  + unit: Đơn vị phần ăn (Object)
  + grams: Trọng lượng tương đương theo gram (Float64)

c) Dữ liệu ra (Kết quả phân tích)

* Các insight về dinh dưỡng và thực phẩm
  + Phân bổ dinh dưỡng across các nhóm thực phẩm
  + So sánh giá trị dinh dưỡng giữa các thực phẩm
  + Mối quan hệ giữa các chất dinh dưỡng Xu hướng dinh dưỡng theo nhóm thực phẩm

## 2. Mô tả thống kê và đánh giá dữ liệu

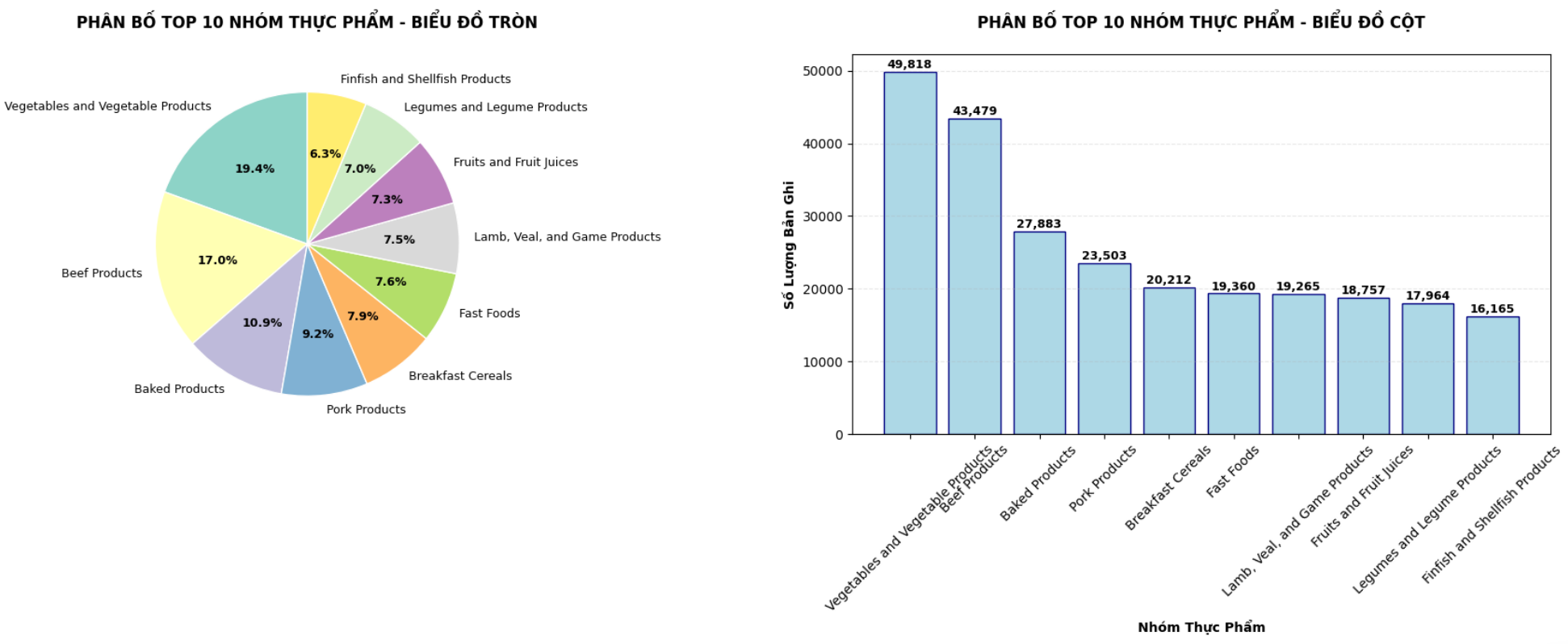


Hình 2.1. Hiển thị nhóm thực phẩm phổ biến

## 3. Phân tích khám phá và trực quan hóa dữ liệu

### 3.1. Phân tích đơn biến (Univariate Analysis)

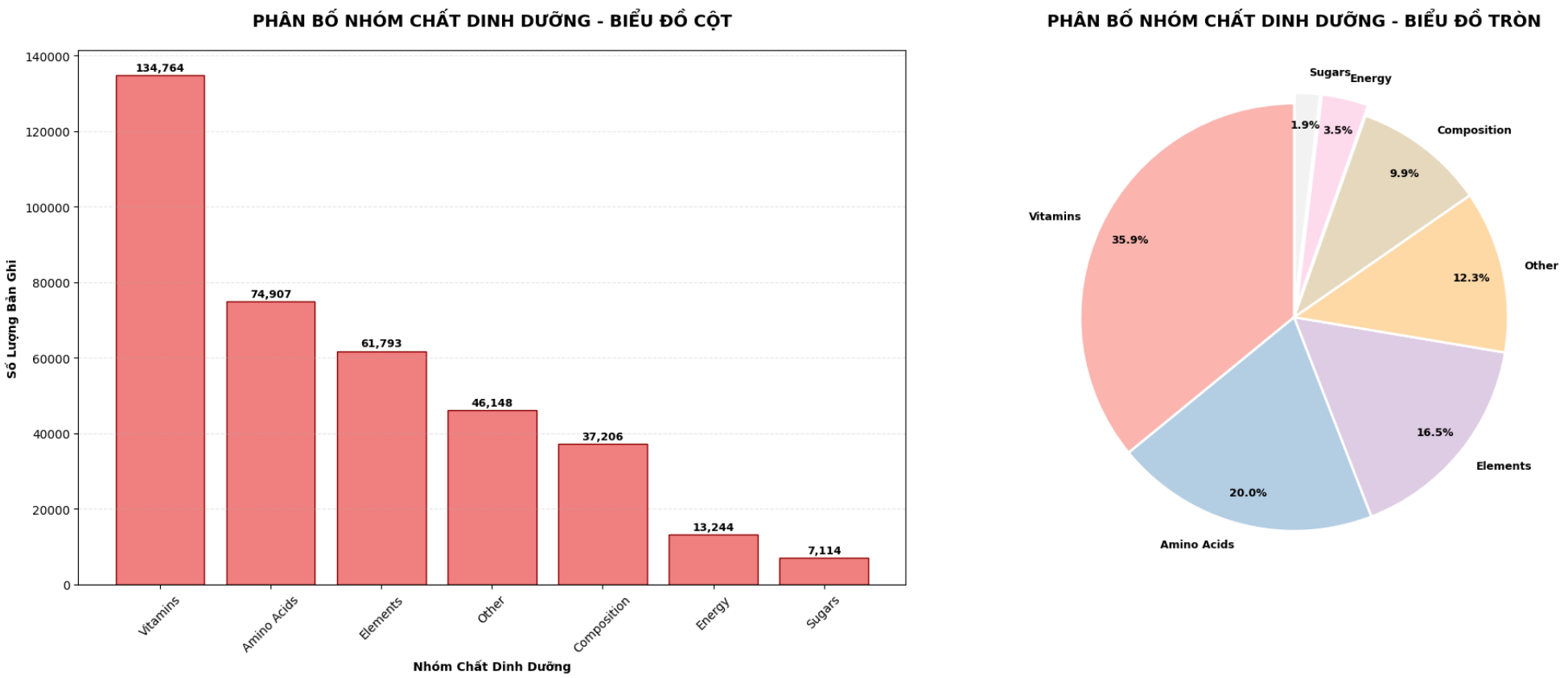
***Câu hỏi 1:*** Phân bố của các nhóm thực phẩm (fgroup) như thế nào? Điều này sẽ giúp chúng ta hiểu hơn về cấu trúc dữ liệu chính và tỷ lệ các loại thực phẩm.



Hình 3.1.1. Hiển thị top 10 nhóm thực phẩm

* **Kết quả phân tích:** Nhóm chiếm ưu thế:
* Rau củ & Sản phẩm từ rau dẫn đầu (19.4%) - phản ánh tính đa dạng và quan trọng của nhóm này
* Sản phẩm thịt bò (17.0%) - nhóm protein động vật phổ biến nhất
* Sản phẩm nướng (10.9%) - thực phẩm chế biến sẵn chiếm tỷ trọng đáng kể

***Câu hỏi 2:*** Nhóm chất dinh dưỡng nào chiếm tỷ lệ lớn trong cơ sở dữ liệu USDA?

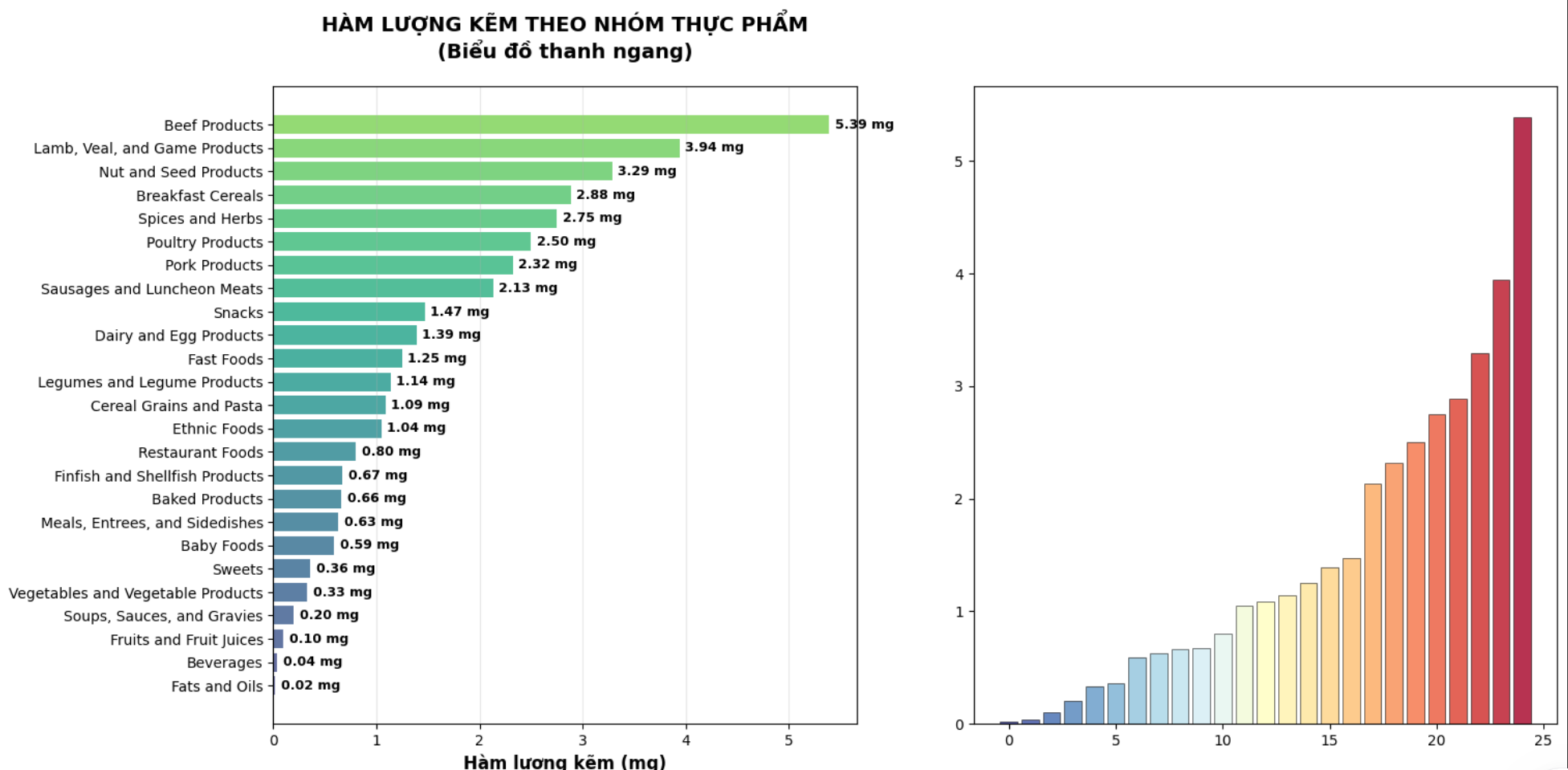


Hình 3.1.2. Hiển thị phân bổ nhóm chất dinh dưỡng

* **Kết quả phân tích:** Nhóm chiếm ưu thế:
* Vitamins dẫn đầu rõ rệt (35.9%) - Cho thấy cơ sở dữ liệu USDA rất chi tiết về vi chất dinh dưỡng
* Amino Acids (20.0%) - Quan tâm đến chất lượng protein
* Elements (16.5%) - Khoáng chất đa lượng và vi lượng
* Other (12.3%) - Các hợp chất đặc biệt khác

### 3.2. Phân tích đa biến (Multivariate Analysis)

***Câu hỏi 1:*** Nhóm thực phẩm nào có hàm lượng kẽm cao nhất và thấp nhất? Sự chênh lệch này lớn tới mức nào?

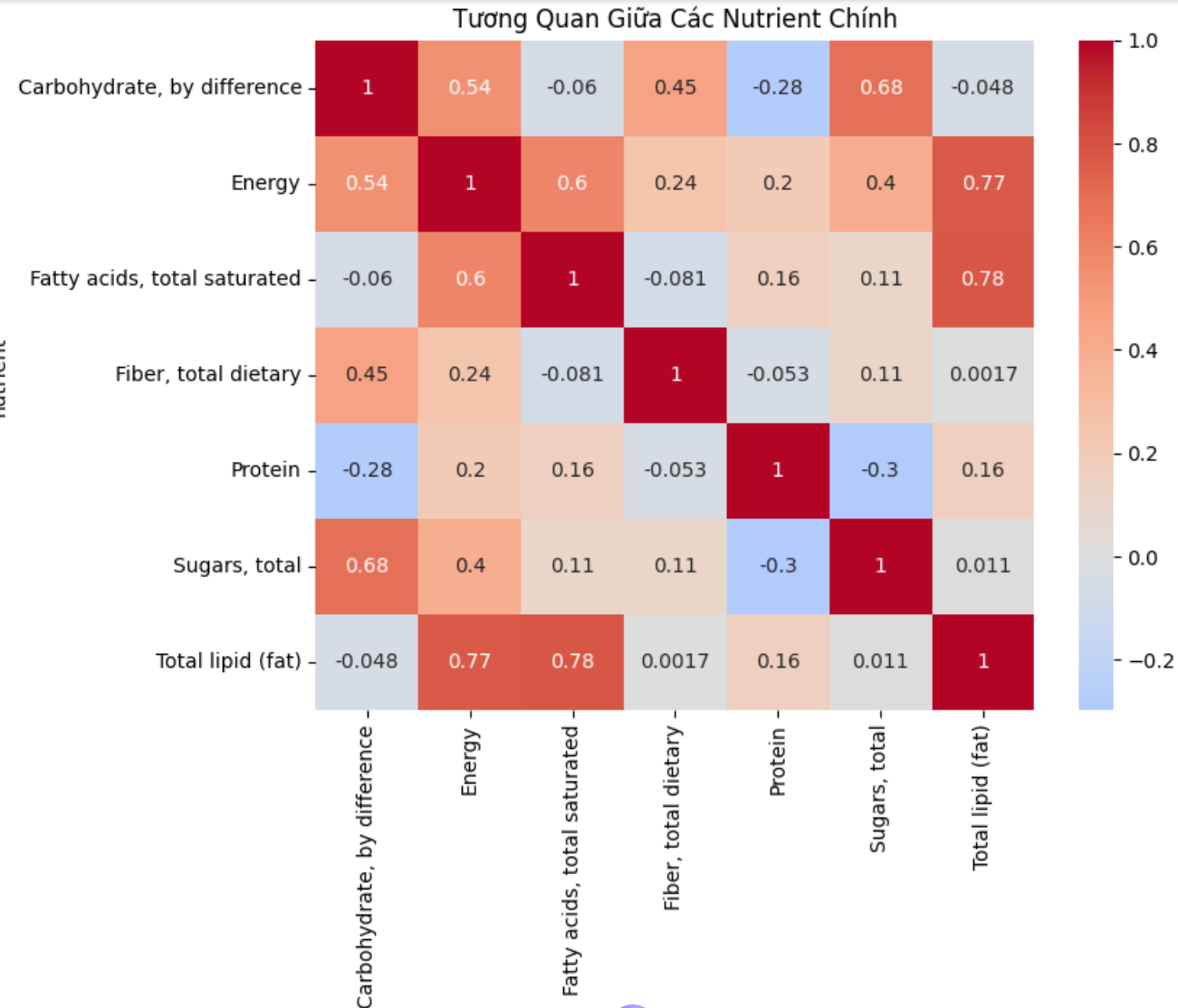


Hình 3.2.1. Hiển thị hàm lượng kẽm theo nhóm thực phẩm

* **Kết quả phân tích:** Từ biểu đồ hàm lượng kẽm, ta có khuyến nghị cụ thể:
* CHO NGƯỜI THIẾU KẼM Beef Products (5.39mg)
* Thịt bò các loại Lamb, Veal (3.94mg)
* Thịt cừu, thịt bê Nut and Seed Products (3.29mg)
* Các loại hạt Breakfast Cereals (2.88mg)
* Ngũ cốc ăn sáng Poultry Products (2.50mg)
* Thịt gia cầm

***Câu hỏi bổ sung:*** Dựa trên dữ liệu này, người ăn chay trường họ có thể bổ sung kẽm từ những nguồn thực vật nào có hàm lượng cao nhất trong biểu đồ?

* NGUỒN KẼM THAY THẾ CHO NGƯỜI ĂN CHAY:
* Nut and Seed Products (3.29mg) - HẠT CÁC LOẠI
* Breakfast Cereals (2.88mg) - NGŨ CỐC NGUYÊN HẠT
* Legumes (1.14mg) - ĐẬU CÁC LOẠI



Hình 3.2.2. Hiển thị tương quan của các biến

Chỉ số tương quan giữa "Total lipid (fat)" và "Energy" là 0.77. Tại sao lại như vậy?

* Về mặt dinh dưỡng, điều này là hoàn to hợp lý vì chất béo là nguồn cung cấp năng lượng đậm đặc nhất. Cụ thể:
* 1g Chất béo cung cấp 9 kcal.
* Trong khi đó, 1g Carbohydrate hoặc 1g Protein chỉ cung cấp 4 kcal.
* Vì vậy, khi một thực phẩm có hàm lượng chất béo (Total lipid) cao, nó chắc chắn sẽ có hàm lượng năng lượng (Energy) cao, và con số 0.77 trong bảng đã xác nhận mối quan hệ chặt chẽ này

Chỉ số tương quan giữa 'Fatty acids, total saturated' và 'Total lipid (fat)' là 0.78.

* Giải thích: Giá trị 0.78 cho thấy một mối tương quan thuận rất chặt chẽ. Điều này có nghĩa là:
* Khi hàm lượng Tổng chất béo (Total lipid) trong một loại thực phẩm tăng lên, thì hàm lượng Chất béo bão hòa (Fatty acids, total saturated) trong thực phẩm đó cũng có xu hướng tăng theo.
* Và ngược lại, thực phẩm ít chất béo nói chung thì cũng thường ít chất béo bão hòa.

Chỉ số tương quan giữa 'Carbohydrate, by difference' và 'Sugars, total' là 0.68.

* Giải thích từ góc độ dinh dưỡng: Mối tương quan "khá mạnh" (0.68) thay vì "rất mạnh" (trên 0.8) cho thấy:
* Có những thực phẩm carbohydrate cao chủ yếu là tinh bột (như cơm, bánh mì, mì ít đường) thì có thể có hàm lượng đường không cao.
* Có những thực phẩm carbohydrate cao chủ yếu là đường (như kẹo, nước ngọt, trái cây sấy) nên làm cho mối tương quan này trở nên chặt chẽ hơn.

## 4. Kết luận

**Dữ liệu cho thấy rõ ràng về mối quan hệ giữa các nhóm chất:**

* Năng lượng (Calo) chủ yếu đến từ Chất béo, và phần lớn lượng chất béo đó lại là Chất béo bão hòa - loại chất béo không có lợi cho sức khỏe.
* Đường là một thành phần chính của Carbohydrate, nhưng không phải là tất cả. Vẫn có những nguồn carb tốt (ngũ cốc nguyên hạt, đậu) ít đường.
* Điều này giải thích tại sao các chế độ ăn nhiều thực phẩm chế biến sẵn, đồ chiên rán và thịt mỡ (vừa nhiều chất béo bão hòa vừa nhiều năng lượng) lại dễ dẫn đến thừa cân, béo phì và các bệnh tim mạch.

**Giải pháp cụ thể**: **Để có một chế độ dinh dưỡng lành mạnh, cân bằng, bạn nên áp dụng các nguyên tắc sau:**

* **Giải pháp về CHẤT BÉO & NĂNG LƯỢNG**
* Ưu tiên chất béo tốt: Thay thế nguồn chất béo từ mỡ động vật, đồ chiên rán bằng các nguồn chất béo không bão hòa có lợi.
* Ví dụ: Dùng dầu ô liu, dầu hạt cải để nấu ăn.
* Ăn các loại hạt (hạnh nhân, óc chó) và quả bơ.
* Ăn cá béo (cá hồi, cá thu) ít nhất 2 lần/tuần.
* Kiểm soát tổng lượng: Chọn phương pháp chế biến ít béo (hấp, luộc, nướng) thay vì chiên xào. Đọc nhãn thực phẩm để chọn sản phẩm ít chất béo bão hòa.
* **Giải pháp về CARBOHYDRATE & ĐƯỜNG**
* Chọn Carb thông minh: Hãy phân biệt rõ giữa carb tốt và carb xấu.
* **Nên ăn:** Ngũ cốc nguyên hạt (gạo lứt, yến mạch), đậu các loại, rau củ. Chúng cung cấp năng lượng ổn định và nhiều chất xơ.
* **Hạn chế:** Thực phẩm chứa đường tinh luyện (bánh kẹo, nước ngọt, siro). Chúng cung cấp "calo rỗng" và dễ gây tăng cân.
* **Giải pháp BỔ SUNG KẼM**
* **Cho người ăn mặn:** Thịt bò, thịt cừu, thịt gia cầm và các loại hạt là nguồn kẽm tuyệt vời.
* **Cho người ăn chay:** Đảm bảo đủ kẽm bằng cách ăn đa dạng các loại hạt (hạt bí, hạt điều), ngũ cốc nguyên hạt, và đậu (đậu lăng, đậu xanh).

# CHƯƠNG 4. PHÂN TÍCH DỮ LIỆU VỀ TÊN BÉ TRAI VÀ BÉ GÁI Ở HOA KÌ (1880-2010)

## 1. Định nghĩa vấn đề và mục tiêu phân tích

### 1.1. Bối cảnh vấn đề

Tên gọi là một yếu tố văn hóa – xã hội đặc trưng, phản ánh xu hướng, giá trị và bản sắc của từng thời kỳ lịch sử. Ở Mỹ, việc đặt tên cho trẻ em không chỉ thể hiện sở thích cá nhân của cha mẹ mà còn chịu ảnh hưởng mạnh mẽ từ các yếu tố xã hội, kinh tế và truyền thông đại chúng. Qua thời gian, sự thay đổi trong cách đặt tên có thể cho thấy những biến chuyển trong quan niệm về giới tính, phong cách sống, cũng như mức độ hội nhập văn hóa.

Với nguồn dữ liệu công khai về tên trẻ em do **Cơ quan An sinh Xã hội Hoa Kỳ (SSA)** cung cấp trong giai đoạn **1880–2010,** việc phân tích xu hướng đặt tên trở thành một hướng nghiên cứu thú vị giúp khám phá sâu hơn về biến động xã hội và tâm lý cộng đồng người Mỹ. Nghiên cứu này tập trung vào việc **phân tích sự thay đổi trong tên gọi theo giới tính, mức độ phổ biến của tên qua từng thời kỳ, đặc điểm ngôn ngữ trong cấu trúc tên, và xu hướng sử dụng các tên trung tính** - từ đó rút ra những hiểu biết về ảnh hưởng của bối cảnh văn hóa – xã hội đến hành vi đặt tên của người dân Mỹ.

### 1.2. Mục tiêu phân tích

Chuyên đề phân tích này được thực hiện nhằm thực hiện các mục tiêu sau:

* Chuyên đề này được thực hiện nhằm đạt được các mục tiêu chính sau:
* Khám phá và trực quan hóa xu hướng đặt tên trẻ em theo giới tính qua các thời kỳ.
* Đánh giá mức độ phổ biến và sự đa dạng của tên gọi trong xã hội Mỹ theo thời gian.
* Phân tích đặc điểm ngôn ngữ trong cách đặt tên, tập trung vào chữ cái cuối và xu hướng phát âm.
* Nghiên cứu sự thay đổi trong việc sử dụng các tên trung tính giữa hai giới.

### 1.3. Giới thiệu tổng quan về Dataset

* Dữ liệu vào:
  + name: Tên trẻ em được đặt
  + sex: Giới tính của trẻ (M: Nam, F: Nữ)
  + births: Số lượng trẻ sinh có tên đó trong năm
  + year: Năm sinh tương ứng
* Dữ liệu ra (Kết quả phân tích): Các phát hiện về xu hướng đặt tên trẻ em tại Hoa Kỳ giai đoạn 1880–2010, dựa trên dữ liệu của Cơ quan An sinh Xã hội Mỹ (SSA).

## 2. Hiển thị thông tin tổng quan về dữ liệu

A screenshot of a computer code

AI-generated content may be incorrect.

Hình 2.1. Hiển thị số dòng, số cột, kiểu dữ liệu của từng cột

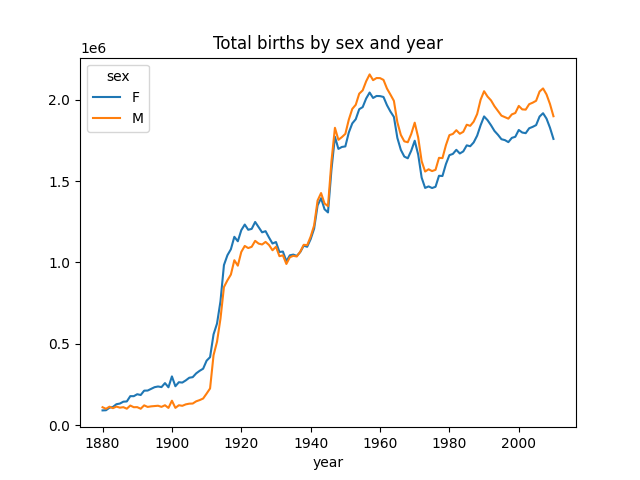
Thông tin cơ bản của bộ dữ liệu:

* Dữ liệu có 4 tính chất (cột) để phân tích, bao gồm: name, sex, births, và year.
* Các giá trị gồm 2 cột kiểu object (name, sex) và 2 cột kiểu int64 (births, year).
* Tổng số dòng dữ liệu là 1,690,784 bản ghi, tương ứng với số lượng tên được thống kê từ năm 1880 đến năm 2010.

Bộ dữ liệu này không có cột phân lớp, mà mang tính thống kê mô tả. Trong đó, biến chính để phân tích là births - biểu thị số trẻ sinh mang tên cụ thể trong từng năm và giới tính.

## 3. Phân tích khám phá và trực quan hóa dữ liệu

***Câu hỏi:*** sự thay đổi về tổng số trẻ sinh theo giới tính qua các giai đoạn (1880–2010) là như thế nào?



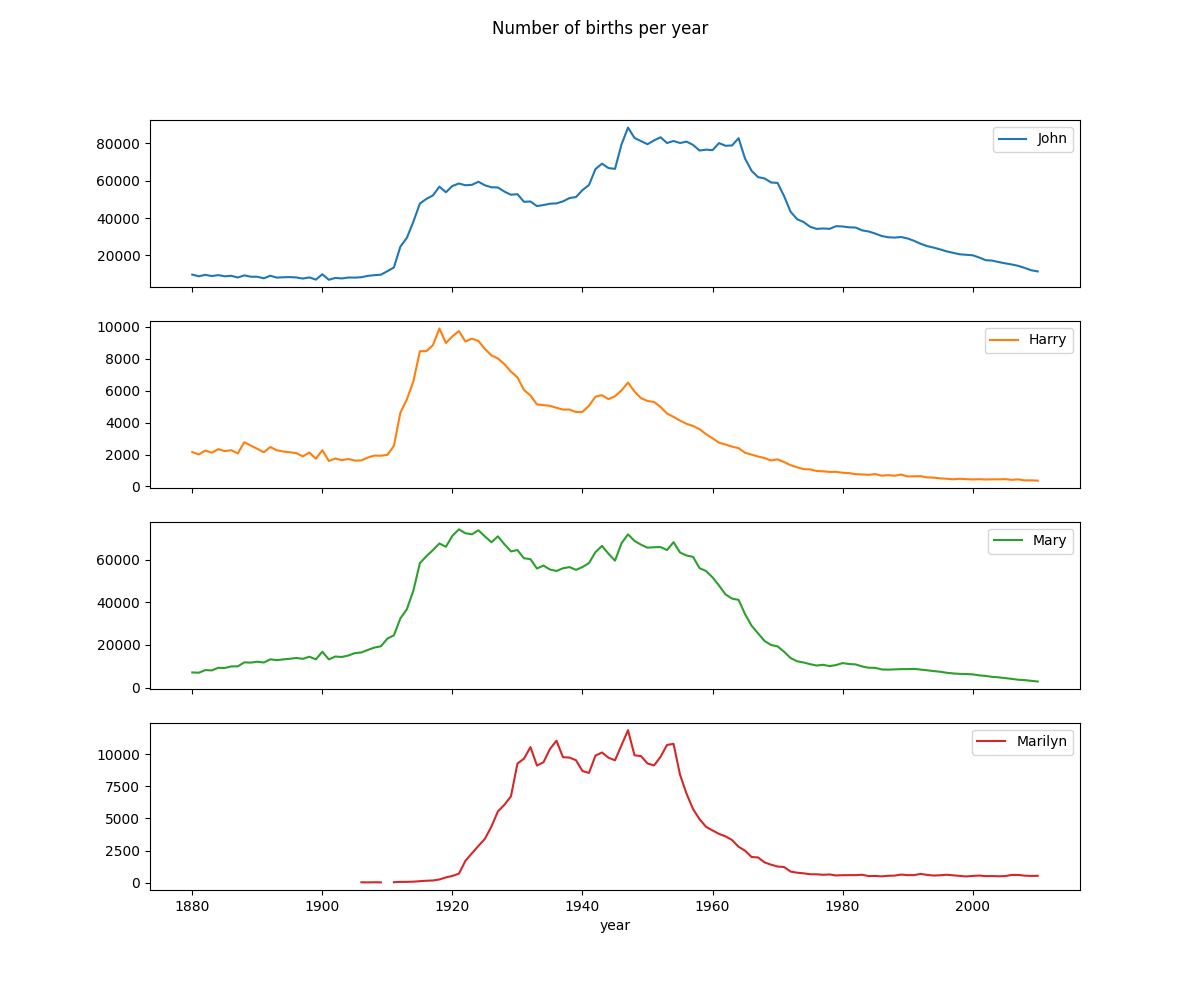
Hình 3.1. Biểu đồ line chart tổng số ca sinh theo giới tính và năm

Nhận xét:

* Giai đoạn 1880–1930:
  + Số lượng trẻ nữ sinh ra cao hơn trẻ nam, thể hiện qua đường biểu diễn của giới nữ nằm trên giới nam.
  + Điều này có thể phản ánh tỷ lệ khai sinh hoặc ghi nhận tên nữ được thống kê nhiều hơn trong giai đoạn đầu.
* Giai đoạn 1930–1950:
  + Hai đường biểu diễn gần như trùng nhau, cho thấy số lượng trẻ nam và nữ sinh ra gần bằng nhau.
  + Đây có thể là giai đoạn dữ liệu trở nên ổn định và đồng đều hơn, phản ánh sự chuẩn hóa trong thống kê hoặc tỷ lệ giới tính tự nhiên.
* Giai đoạn 1950–2010:
  + Số lượng trẻ nam sinh ra nhiều hơn nữ, thể hiện rõ qua việc đường biểu diễn của nam vượt lên trên.

Sau khi đã quan sát sự thay đổi về tổng số trẻ sinh theo giới tính qua các giai đoạn (1880–2010), ta tiếp tục đi sâu vào xu hướng đặt tên cụ thể của người dân Mỹ.

Mục tiêu: Xác định xem mức độ phổ biến của một số tên tiêu biểu (như John, Harry, Mary, Marilyn) có thay đổi theo thời gian hay không. Liệu những cái tên từng “thống trị” bảng xếp hạng đặt tên ở thế kỷ 19 có còn được ưa chuộng ở thế kỷ 21?



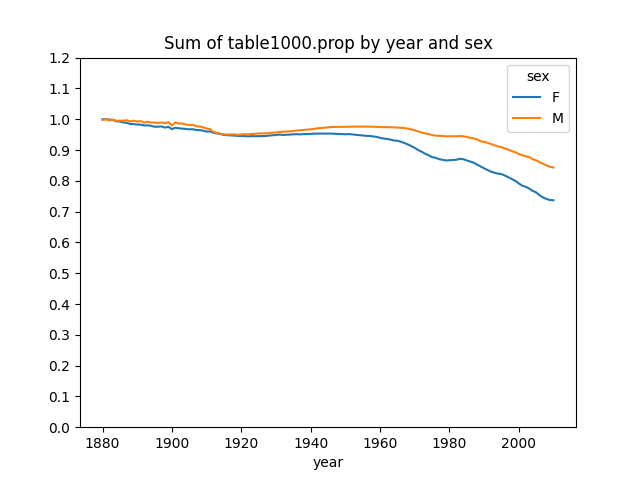
Hình 3.2. Số lượng trẻ sinh theo năm của các tên phổ biến

Khi quan sát biểu đồ số lượng trẻ được đặt tên John, Harry, Mary, Marilyn theo thời gian, ta nhận thấy:

* Tất cả các tên này đều có mức độ phổ biến cao trong khoảng từ 1900 đến 1980, nhưng đều giảm mạnh sau năm 1980.
* Một số tên như Mary từng là tên phổ biến nhất trong nhiều thập kỷ, nhưng dần mất đi vị thế trong những năm gần đây. Kết luận (vì sao lại như vậy):
* Xu hướng giảm dần này phản ánh sự thay đổi trong thị hiếu và văn hóa đặt tên của người Mỹ:
* Những cái tên này không còn được ưa chuộng nữa, ít cha mẹ chọn tên phổ biến cho con cái.
* Sự thay đổi này phản ánh thị hiếu đặt tên chuyển dịch theo thời gian, chịu ảnh hưởng bởi xu hướng văn hóa, xã hội và truyền thông (ví dụ: tên diễn viên, nhân vật nổi tiếng, hoặc trào lưu từng thời kỳ).
* Nói cách khác, việc đặt tên đã bớt truyền thống hơn và trở nên linh hoạt, cá nhân hóa hơn.

Tuy nhiên, việc giảm phổ biến của một số tên riêng lẻ chưa đủ để phản ánh toàn bộ xu hướng xã hội.Do đó, ta tiếp tục đi sâu hơn để phân tích mức độ đa dạng hóa trong đặt tên qua thời gian.

***Câu hỏi đặt ra:*** Người Mỹ có đang đặt tên cho con ngày càng đa dạng hơn không? Hay vẫn tập trung quanh một số tên phổ biến nhất?



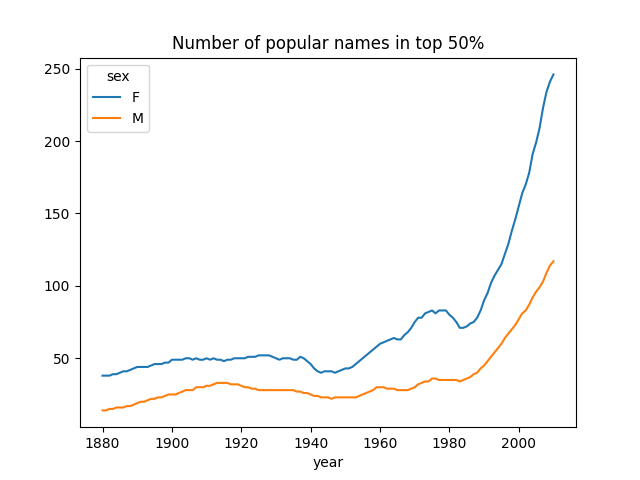
Hình 3.3. Tỷ lệ sinh được đại diện bởi 1000 tên phổ biến

nhất theo năm và giới tính

Nhận xét:

Kết quả cho thấy cả hai giới tính đều có xu hướng giảm dần rõ rệt từ năm 1880 đến 2010. Mức độ tập trung vào các tên phổ biến ngày càng giảm, hay nói cách khác, người dân có xu hướng đặt tên đa dạng hơn cho con cái — phản ánh sự thay đổi trong văn hóa và cá tính hóa lựa chọn tên.

Tiếp theo, ta đo lường xu hướng đa dạng theo giới tính và thời gian - Liệu xu hướng đa dạng hóa tên gọi có khác nhau giữa con trai và con gái không?



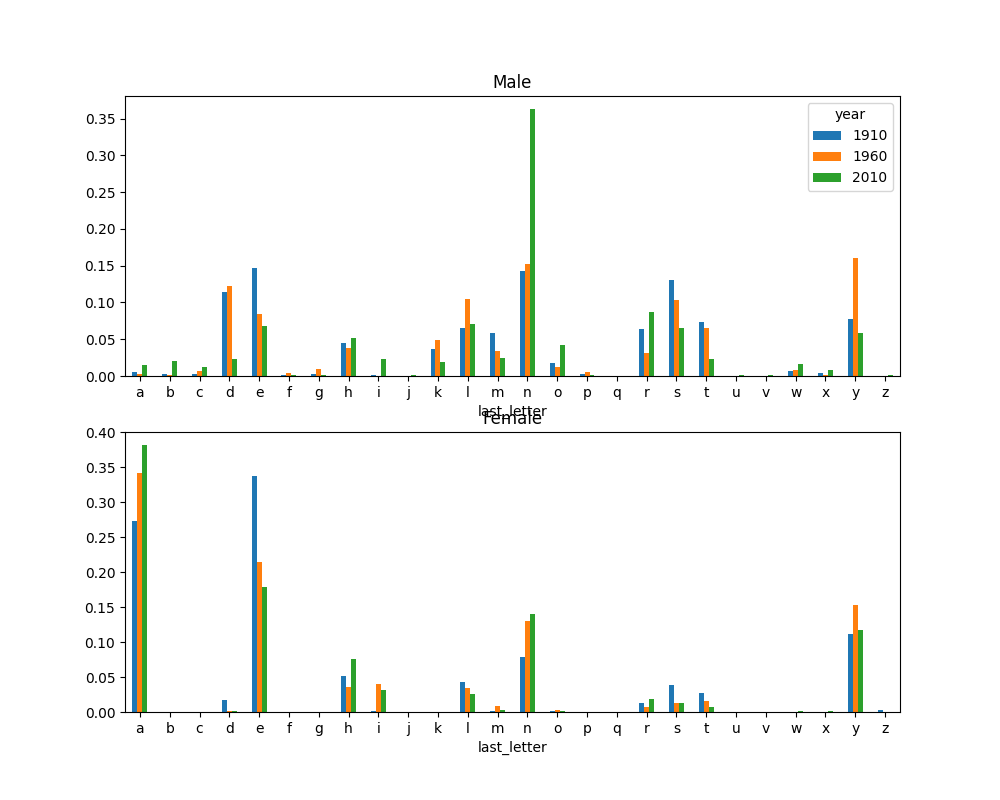
Hình 3.4. Số lượng tên chiếm 50% tổng số ca sinh theo năm và giới tính

Nhận xét: Biểu đồ cho thấy tên của con gái luôn đa dạng hơn tên con trai, và xu hướng này ngày càng rõ rệt theo thời gian.

Kết luận: Văn hoá đặt tên cho con gái thường chịu ảnh hưởng của thời trang, truyền thông và các trào lưu xã hội, dẫn đến sự thay đổi tên nhanh chóng (ví dụ: tên diễn viên, ca sĩ, nhân vật nổi tiếng). Ngược lại, tên con trai có xu hướng ổn định hơn, do truyền thống gia đình hoặc các giá trị bảo thủ về giới tính. Xu hướng gia tăng mức độ đa dạng tên phản ánh sự cá nhân hóa và tự do lựa chọn trong xã hội Mỹ hiện đại, khi cha mẹ muốn con mình có tên “độc đáo” hơn thay vì theo khuôn mẫu truyền thống.

Sau khi đã quan sát sự thay đổi về tổng số trẻ sinh theo giới tính qua các giai đoạn (1880–2010) và sự đa dạng trong việc đặt tên, ta tiếp tục đi sâu hơn vào một khía cạnh thú vị: sự thay đổi trong chữ cái cuối cùng của tên qua thời gian.

Liệu chữ cái cuối của tên trẻ có phản ánh sự thay đổi trong xu hướng đặt tên qua các thời kỳ hay không?



Hình 3.5. Phân bố tỷ lệ chữ cái cuối trong tên trẻ sinh theo giới tính và năm

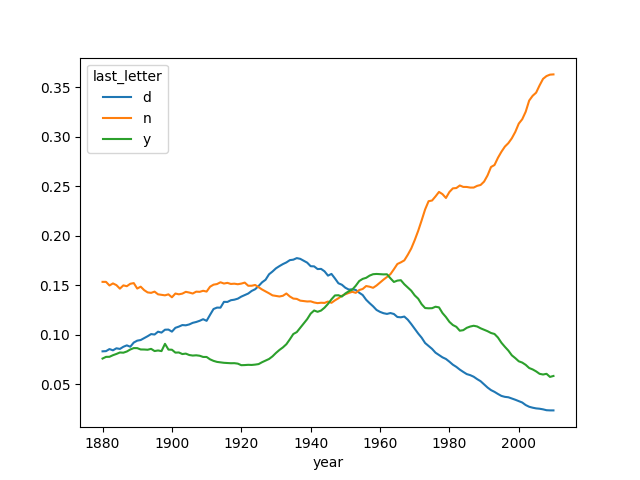
Nhận xét :

* Ở con trai, tên kết thúc bằng chữ “n” có sự tăng trưởng rõ rệt qua các năm.
* Ở con gái, sự phân bố chữ cái cuối đồng đều hơn, tuy nhiên có xu hướng thiên về các chữ “a”.

Kết luận:

* Ở con trai, sự gia tăng mạnh của các tên kết thúc bằng chữ **“**n**”** (như *Aiden, Ethan, Benjamin, Mason*) thể hiện xu hướng đặt tên hiện đại, mạnh mẽ và có âm cuối dứt khoát.
* Ở con gái, các tên kết thúc bằng chữ **“**a**”** ngày càng phổ biến, phản ánh thị hiếu ưa chuộng sự mềm mại, nữ tính trong cách đặt tên (*Emma, Olivia, Sophia, Isabella*).

Để quan sát rõ xu hướng thay đổi của một số chữ cái cụ thể, ta tính tỷ lệ (%) của các chữ cái cuối(d,n,y) trong từng năm ở tên con trai:

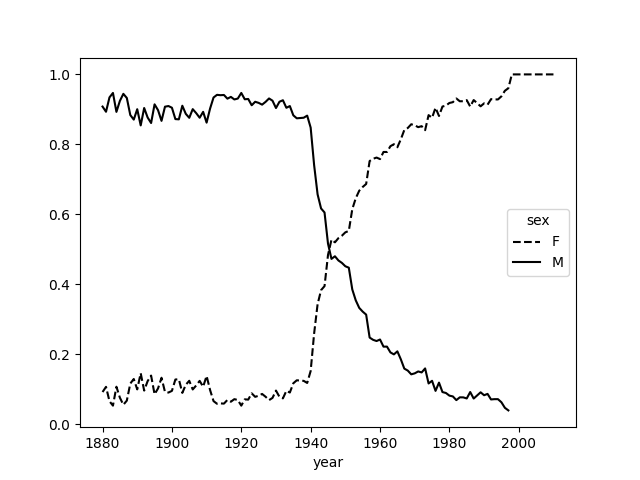


Hình 3.6. Tỷ lệ tên con trai kết thúc bằng các chữ cái ‘d’, ‘n’, ‘y’ qua thời gian (1880–2010)

Biểu đồ cho thấy: Tên kết thúc bằng “n” ngày càng phổ biến mạnh mẽ kể từ thập niên 1960, đồng thời tên kết thúc bằng “d” và “y” lại giảm rõ rệt. Điều này phản ánh sự chuyển dịch trong phong cách đặt tên:

* Các tên hiện đại như Aiden, Ethan, Benjamin (kết thúc bằng “n”) dần chiếm ưu thế.
* Trong khi các tên truyền thống như David, Larry, Harry (kết thúc bằng “d”, “y”) ngày càng ít được sử dụng.

Sau khi đã quan sát sự thay đổi trong xu hướng đặt tên theo chữ cái cuối, một hiện tượng thú vị khác được phát hiện trong dữ liệu:Liệu có tồn tại những tên vốn phổ biến cho con trai nhưng sau này lại được dùng chủ yếu cho con gái (hoặc ngược lại)? Nếu có, sự chuyển dịch này diễn ra vào giai đoạn nào và với tốc độ ra sao?



Hình 3.7. Tỷ lệ giới tính của các tên bắt đầu bằng ‘Lesl’ qua các năm

Biểu đồ cho thấy: Trong những năm đầu thế kỷ 20, các biến thể của “Lesl” (như Leslie, Lesley) được dùng chủ yếu cho con trai. Tuy nhiên, kể từ thập niên 1940 trở đi, tỷ lệ sử dụng cho con gái bắt đầu tăng nhanh, và đến giai đoạn 1960–1980, tên này trở thành tên nữ phổ biến hơn. Sự chuyển dịch này thể hiện một điểm ngoặt rõ rệt trong quan niệm về tên gọi và giới tính tại Hoa Kỳ.

Kết luận: Tên “Leslie” và các biến thể tương tự là ví dụ điển hình cho quá trình trung tính hóa giới tính trong tên gọi. Hiện tượng này có thể xuất phát từ:

* Ảnh hưởng của văn hóa đại chúng, nơi tên “Leslie” xuất hiện nhiều trong phim ảnh và âm nhạc gắn liền với hình ảnh nữ tính.
* Xu hướng xã hội cởi mở hơn với các tên phi giới tính, phản ánh thay đổi trong nhận thức về bình đẳng giới.
* Từ đó có thể thấy, việc đặt tên không chỉ là lựa chọn ngôn ngữ, mà còn là biểu hiện của sự thay đổi trong tư duy văn hóa và xã hội theo thời gian.

## 4. Kết luận

Qua phân tích từ dữ liệu lịch sử về tên trẻ em tại Hoa Kỳ giai đoạn 1880–2010, có thể rút ra các kết luận tổng hợp sau:

* Thay đổi về số lượng sinh theo giới tính: Từ 1880 đến đầu thế kỷ 20, số trẻ nữ sinh ra cao hơn trẻ nam; tuy nhiên sau giai đoạn 1950, trẻ nam lại chiếm ưu thế. Điều này phản ánh sự ổn định và chuẩn hóa trong thống kê dân số, cũng như quy mô dân số Mỹ ngày càng tăng.
* Sự thay đổi trong xu hướng đặt tên: Những cái tên từng cực kỳ phổ biến như John, Mary, Harry, Marilyn đã giảm mạnh về mức độ phổ biến theo thời gian. Điều đó cho thấy người Mỹ không còn tập trung vào một số tên truyền thống, mà đa dạng hóa lựa chọn tên cho con cái — phản ánh sự thay đổi văn hóa và cá tính hóa trong xã hội hiện đại.
* Mức độ đa dạng trong đặt tên ngày càng tăng:
  + Phân tích tỉ lệ sinh đại diện bởi 1000 tên phổ biến nhất cho thấy xu hướng giảm mạnh qua các thập niên, tức là nhiều tên mới xuất hiện hơn. Chứng tỏ mức độ đa dạng trong việc đặt tên tăng lên rõ rệt, thể hiện sự cá nhân hóa và sáng tạo trong văn hóa đặt tên.
* Cách kết thúc tên phản ánh xu hướng thời đại
  + Phân tích chữ cái cuối của tên cho thấy:
* Tên con trai kết thúc bằng “n” (như Ethan, Aiden, Benjamin) tăng mạnh sau thập niên 1960.
* Trong khi đó, các tên kết thúc bằng “d” hoặc “y” (David, Larry, Harry) giảm dần. Điều này thể hiện sự chuyển dịch về phong cách ngôn ngữ, hướng tới các âm tiết mềm mại, hiện đại hơn — phù hợp với thị hiếu đương đại.
* Sự chuyển dịch giới tính trong tên gọi: Một số tên vốn dành cho con trai (như Leslie, Lesley) dần trở thành tên con gái kể từ sau năm 1940. Hiện tượng này minh chứng cho sự thay đổi trong quan niệm giới tính, khi xã hội Mỹ dần cởi mở và bình đẳng hơn trong việc đặt tên, cho phép tên mang tính trung tính về giới.

# CHƯƠNG 5. PHÂN TÍCH CƠ SỞ DỮ LIỆU ỦY BAN BẦU CỬ LIÊN BANG (FEC) 2012

# 1. Bối cảnh vấn đề và mục tiêu phân tích

### 1.1. Bối cảnh vấn đề

Tài chính chính trị đóng một vai trò then chốt trong các cuộc bầu cử tại Hoa Kỳ, nơi mà khả năng gây quỹ thường được xem như một thước đo quan trọng cho sức mạnh và mức độ ủng hộ của một chiến dịch. Việc hiểu rõ nguồn tiền đến từ đâu, ai là những người ủng hộ chính, và xu hướng đóng góp diễn ra như thế nào là vô cùng quan trọng đối với các ứng viên, nhà chiến lược, các phương tiện truyền thông, và toàn bộ cử tri. Phân tích dữ liệu đóng góp không chỉ giúp các chiến dịch tối ưu hóa chiến lược gây quỹ mà còn mang lại sự minh bạch, giúp công chúng hiểu rõ hơn về các lực lượng kinh tế-xã hội ẩn sau mỗi cuộc chạy đua vào Nhà Trắng.

Chuyên đề phân tích này tập trung vào việc khám phá dữ liệu đóng góp cá nhân cho các chiến dịch bầu cử liên bang Hoa Kỳ trong chu kỳ bầu cử năm 2012, nhằm vẽ nên một bức tranh toàn cảnh và chi tiết về bối cảnh tài chính của cuộc đua tổng thống năm đó.

### 1.2. Mục tiêu phân tích

Chuyên đề được thực hiện nhằm trả lời các câu hỏi then chốt sau:

* Phân Tích Đối Tượng Đóng Góp: Ai là những người ủng hộ tài chính chính? Những nghề nghiệp và nhà tuyển dụng nào đóng góp nhiều nhất cho các chiến dịch?
* Phân Tích Địa Lý: Sự ủng hộ tài chính phân bố như thế nào trên bản đồ nước Mỹ? Bang nào và khu vực nào đóng góp nhiều nhất?
* So Sánh Ứng Viên: Có sự khác biệt nào rõ rệt về cơ cấu và nguồn lực tài chính giữa hai ứng viên chính là Barack Obama (Đảng Dân chủ) và Mitt Romney (Đảng Cộng hòa)?
* Phân Tích Quy Mô Đóng Góp: Các chiến dịch được hỗ trợ chủ yếu bởi những khoản đóng góp nhỏ từ quần chúng hay bởi các khoản đóng góp lớn từ giới tinh hoa?

### 1.3. Giới thiệu tổng quan về Dataset

Để thực hiện các mục tiêu trên, nghiên cứu sử dụng bộ dữ liệu từ Ủy ban Bầu cử Liên bang (FEC) về các khoản đóng góp cá nhân trong chu kỳ 2012.

* **Phạm vi:** Dữ liệu về đóng góp cá nhân cho các chiến dịch bầu cử liên bang Hoa Kỳ (Tổng thống, Thượng viện, Hạ viện) trong chu kỳ 2012.
* **Kích thước:** Bộ dữ liệu bao gồm 1,001,731 bản ghi (mẫu) về các giao dịch quyên góp.
* **Đặc điểm:** Gồm 16 trường (cột) dữ liệu, chủ yếu là thông tin định danh, thông tin người đóng góp và chi tiết giao dịch.

**Các Biến Dữ Liệu Chính (Đầu Vào):**

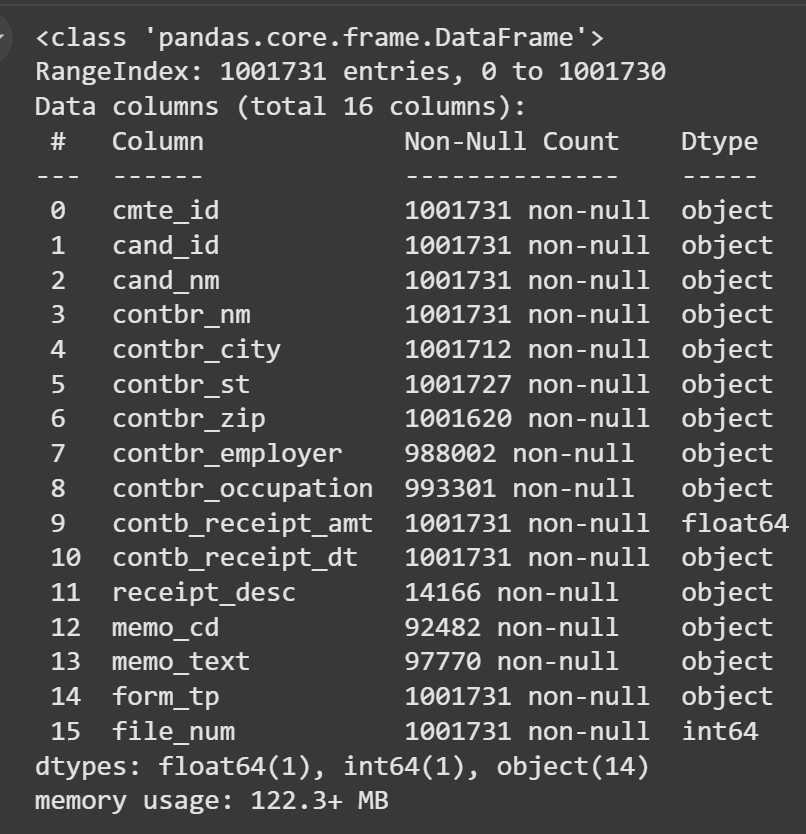
* cmte\_id: ID của Ủy ban Chiến dịch
* cand\_id: ID của ứng cử viên
* cand\_id: Tên đầy đủ của Ứng cử viên nhận quyên góp
* contbr\_nm: Tên đầy đủ của Người đóng góp(Cá nhân)
* contbr\_city: Thành phố của Người đóng góp
* contbr\_st: Bang của Người đóng góp
* contbr\_zip: Mã Zip của Người đóng góp
* contbr\_employer: Tên Công ty/Cơ quan nơi Người đóng góp làm việc
* contbr\_occupation: Nghề nghiệp của người đóng góp
* contb\_receipt\_amt: Số tiền giao dịch (Có thể là số dương: quyên góp, hoặc số âm: hoàn tiền).
* contb\_receipt\_dt: Ngày giao dịch
* receipt\_desc: Mô tả biên lai
* memo\_cd: Mã ghi nhớ để chú thích giao dịch
* memo\_text: Chi tiết nội dung ghi nhớ
* form\_tp: Loại biểu mẫu dùng để xác định loại giao dịch
* file\_num: Số hồ sơ dùng để định dạng hồ sơ

**Dữ Liệu Đầu Ra (Kết Quả Phân Tích Dự Kiến):**

* Các phát hiện (insights) định lượng và định tính về hành vi quyên góp và tài trợ chính trị, được trình bày thông qua:
* Các bảng thống kê (top nghề nghiệp, top nhà tuyển dụng, tổng số tiền theo bang).
* Biểu đồ trực quan (biểu đồ cột, biểu đồ tròn, bản đồ nhiệt).
* Nhận định và kết luận về cấu trúc tài chính của cuộc bầu cử tổng thống 2012.

## 2. Mô tả thống kê và đánh giá dữ liệu

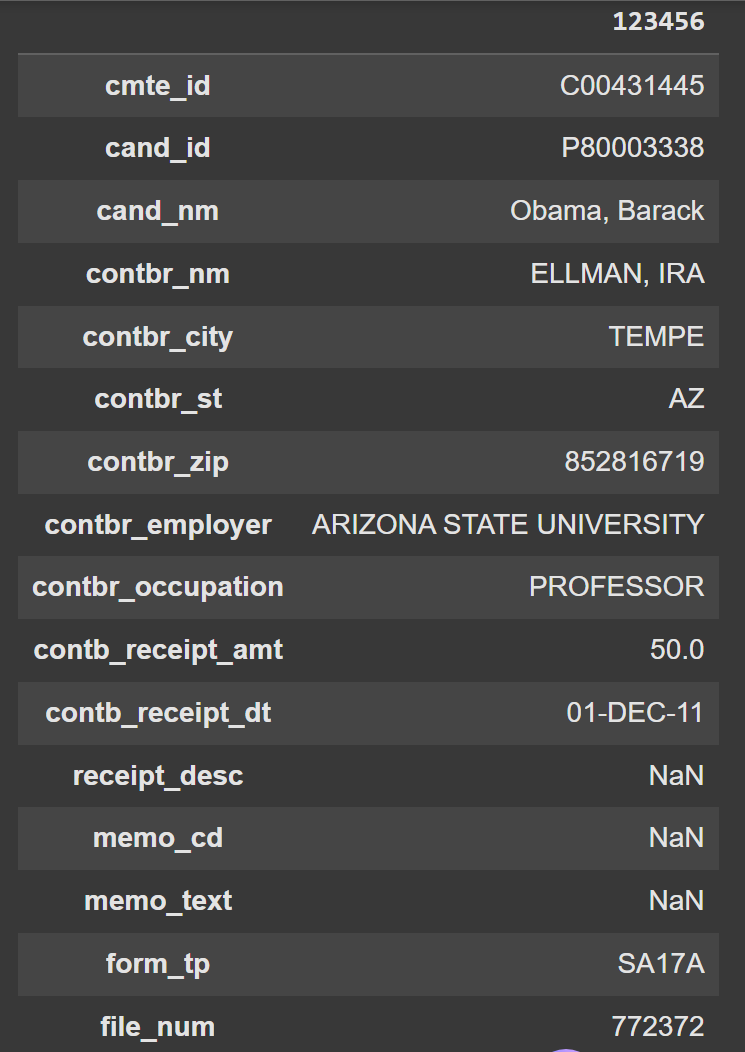
### 2.1. Hiển thị thông tin tổng quan về dữ liệu



Hình 2.1.1. Hiển thị số dòng, số cột và kiểu dữ liệu của từng cột

* **Nhận xét:**
* Dữ liệu có 16 tính chất (cột) để phân tích.
* Các giá trị thuộc đa số thuộc kiểu object (chuỗi/text), 1 kiểu float64 (cột contb\_receipt\_amt), và 1 kiểu int64 (cột file\_num).
* Tổng số dòng dữ liệu là 1,001,731 bản ghi (giao dịch quyên góp). Bộ dữ liệu này không có cột phân lớp.
* Đây là dữ liệu giao dịch tài chính/phân bổ nguồn lực, trong đó cột contb\_receipt\_amt (số tiền quyên góp) là biến số lượng chính để phân tích.

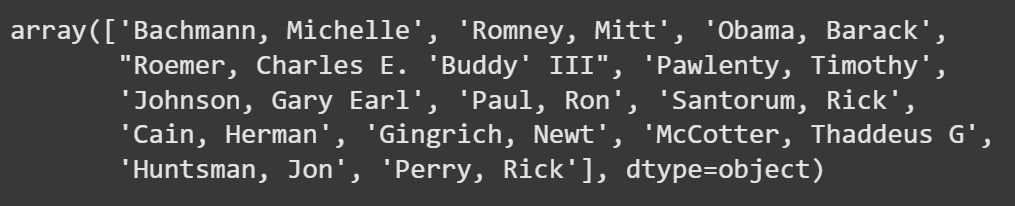
### 2.2. Hiển thị một dòng dữ liệu cụ thể



Hình 2.2.1. Hiển thị một dòng dữ liệu cụ thể

* **Nhận xét:**
* Thông tin người quyên góp:
  + Tên Người quyên góp là: ELLMAN, IRA.
  + Nghề nghiệp là: Professor - Ngành giáo dục (Professor/Giáo sư) là một nhóm nghề nghiệp có thể coi là quan trọng để phân tích xu hướng ủng hộ chính trị.
  + Công ty/Cơ quan: ARIZONA STATE UNIVERSITY - Xác định nguồn tiền đến từ một cơ sở giáo dục lớn.
  + Địa điểm( Thành phố và Bang): TEMPE, AZ - Giao dịch đến từ bang Arizona (AZ). Điều này hữu ích cho việc phân tích địa lý nguồn vốn.
* Thông tin ứng cử viên: Tên Ứng cử viên là: Obama, Barac - Xác nhận giao dịch quyên góp này dành cho chiến dịch tái tranh cử Tổng thống của Barack Obama.
* Thông tin Giao dịch (Chi tiết về số tiền và thời gian):
  + Số tiền Quyên góp: 50.0 - Đây là một khoản quyên góp nhỏ, cho thấy đây là một ủng hộ viên cá nhân, bình thường.
  + Ngày Quyên góp: 01-DEC-11 - Giao dịch này được thực hiện vào ngày 1 tháng 12 năm 2011, tức là trước năm bầu cử chính thức 2012, thuộc chu kỳ gây quỹ sớm.
  + Loại Biểu mẫu: SA17A - Đây là một loại biểu mẫu báo cáo giao dịch quyên góp cá nhân.
* Ngoài ra, 3 cột phụ trợ là receipt\_desc, memo\_cd, và memo\_text thiếu dữ liệu(NaN) là hoàn toàn bình thường và không ảnh hưởng đáng kể. Vì Ba cột dữ liệu này chỉ đóng vai trò cung cấp chi tiết giao dịch bổ sung hoặc ghi chú nội bộ.

### 2.3. Hiển thị danh sách các ứng viên



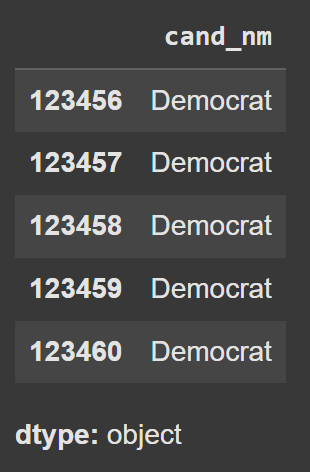
Hình 2.3.1. Hiển thị danh sách các ứng viên

### 2.4. Chuẩn bị biến phân loại (Đảng phái)

Vấn đề Dữ liệu Thô: Biến **cand\_nm** (Tên ứng cử viên) là một biến định danh dạng chuỗi thô, vốn không chứa thông tin về đảng phái, do đó không cho phép phân tích theo nhóm chính trị.

Giải pháp: Cần thực hiện thao tác Ánh xạ Dữ liệu (Data Mapping) bằng cách tạo từ điển parties. Phương pháp này chuyển đổi dữ liệu chuỗi thô thành một biến phân loại có ý nghĩa (Democrat hoặc Republican).

Tạo biến phân loại Party: Sử dụng phương thức .map() để ánh xạ từng tên ứng cử viên (từ cột cand\_nm) sang tên đảng phái tương ứng (dựa trên từ điển parties). Kết quả là tạo ra cột party mới, một biến phân loại cần thiết, chuyển đổi dữ liệu thô thành thông tin có ý nghĩa, sẵn sàng cho việc phân tích theo nhóm chính trị.

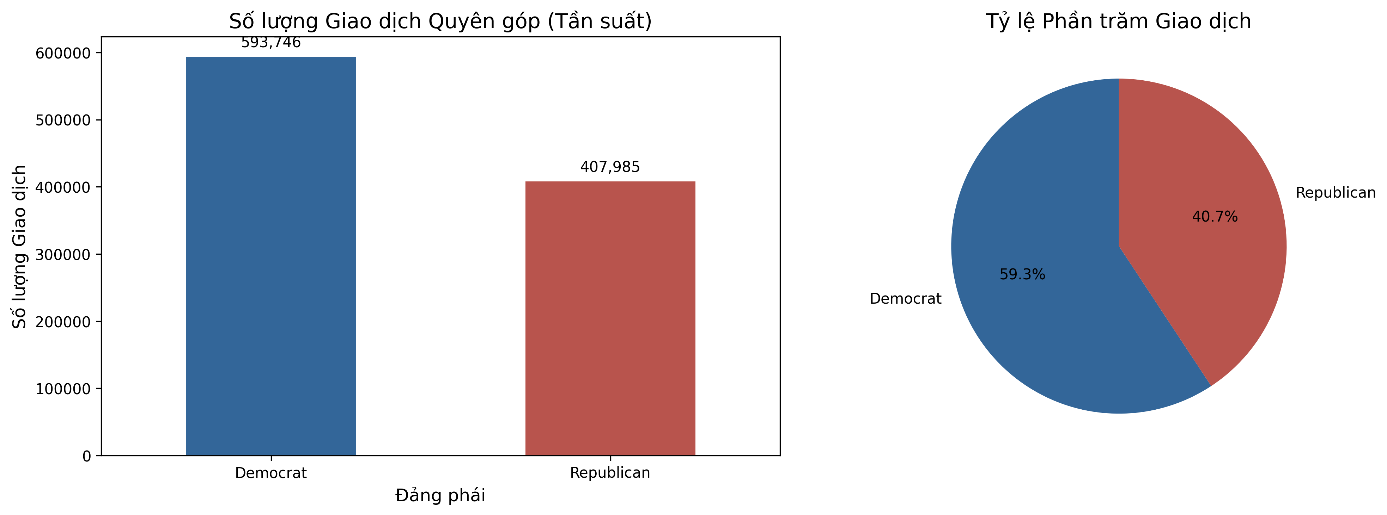
 

Hình 2.4.1. Hiển thị 5 dòng dữ liệu thô (trái) và kết quả ánh xạ (phải)

## 3. Phân tích khám phá và trực quan hóa dữ liệu

### 3.1. Phân tích đơn biến (Univariate Analysis)

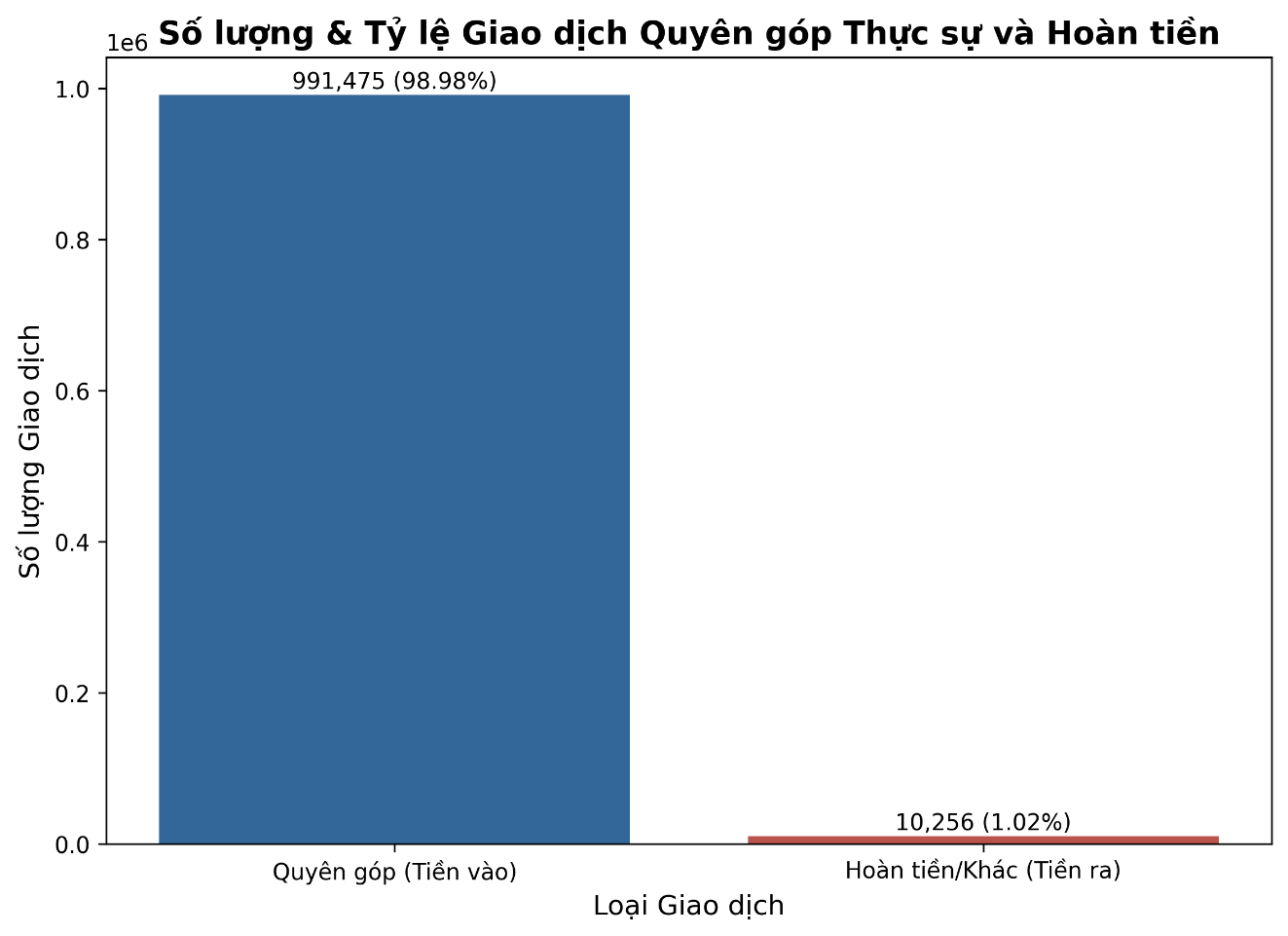
***Câu hỏi 1****:* Sự phân bổ về mặt tần suất giao dịch quyên góp giữa hai đảng chính là gì? Liệu một bên có huy động được một cơ sở ủng hộ rộng rãi hơn không?



Hình 3.1.1. Hiển thị tần số giao dịch quyên góp giữa 2 đảng

* **Kết quả phân tích:** Dựa trên kết quả trực quan hóa tần suất giao dịch quyên góp, có thể khẳng định Đảng Dân chủ(Democrat) đã huy động được một cơ sở ủng hộ rộng rãi hơn. Cụ thể, Đảng Dân chủ nhận được 593,746 giao dịch, chiếm tới 59.3% tổng số, áp đảo Đảng Cộng hòa(Republican) với 407,985 giao dịch, chỉ chiếm 40.7%. Sự chênh lệch này, lên tới hơn 45% về số lượng giao dịch, cho thấy khả năng vận động và thu hút một lượng lớn người ủng hộ cá nhân thực hiện quyên góp của Đảng Dân chủ là vượt trội. Tuy nhiên, nhận định này chỉ dựa trên tần suất, không phải tổng số tiền quyên góp thực tế.

***Câu hỏi 2:*** Tỷ lệ giữa các giao dịch quyên góp thực sự (tiền vào) và các giao dịch hoàn tiền/khác (tiền ra) là bao nhiêu, và chúng ta có nên loại bỏ các giao dịch âm khi tính tổng nguồn vốn hay không?



Hình 3.1.2. Hiển thị tỷ lệ giao dịch tiền vào và tiền ra

* **Kết quả phân tích:** Dựa trên biểu đồ cột, tỷ lệ giao dịch Hoàn tiền/Khác là rất nhỏ, chỉ 1.02% (10,256 giao dịch), trong khi 98.98% (991,475 giao dịch) là các khoản quyên góp thực sự (Tiền vào). Mặc dù tỷ lệ này không làm thay đổi đáng kể tổng số lượng giao dịch, nhưng để đảm bảo độ chính xác tuyệt đối khi tính toán tổng nguồn vốn, chúng ta phải loại bỏ các giao dịch âm (hoàn tiền hoặc có giá trị ≤0) trước khi chuyển sang Phân tích Đa Biến. Đây là một bước Làm sạch Dữ liệu cần thiết để đảm bảo tính toán tổng số tiền thực tế nhận được là đại diện và chính xác.

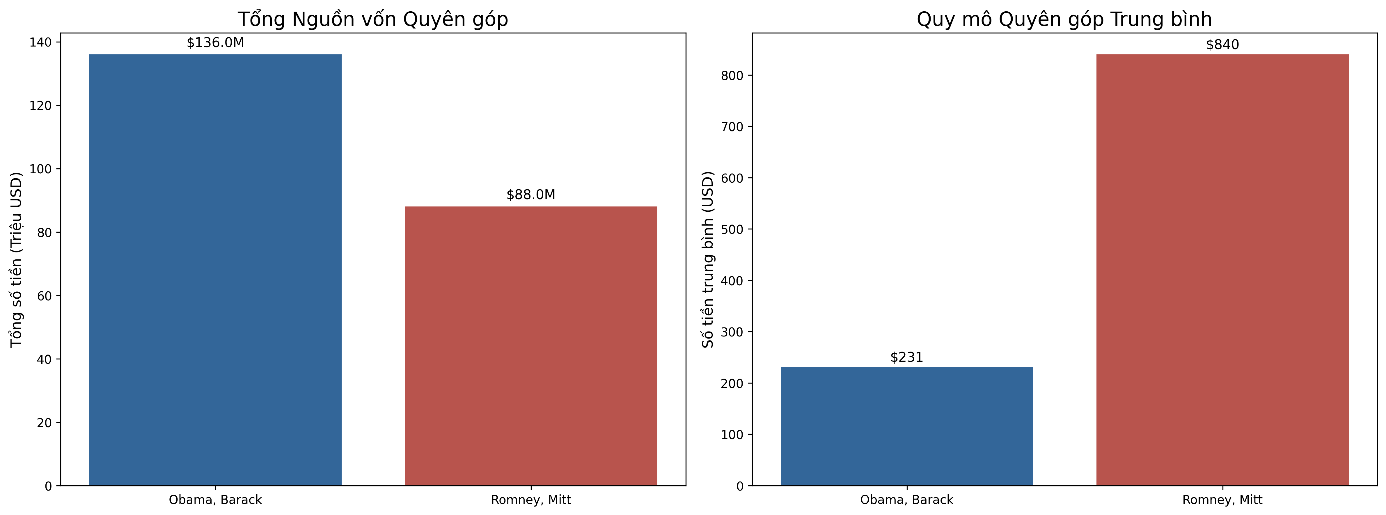
**Việc cần làm:** Thực hiện làm sạch Dữ liệu chỉ giữ lại các giao dịch có số tiền quyên góp lớn hơn 0. Việc này đảm bảo các phép tính tổng số tiền sau này sẽ chỉ phản ánh nguồn vốn ròng thực sự nhận được, loại bỏ ảnh hưởng của các khoản hoàn tiền.

### 3.2. Phân tích đa biến (Multivariate Analysis)

Vì Barack Obama và Mitt Romney là hai ứng cử viên chính, nên để đơn giản hóa việc phân tích, sẽ chuẩn bị một tập hợp con chỉ có những đóng góp cho chiến dịch của họ

***Vấn đề đặt ra:*** Chúng ta đã thấy lợi thế về tần suất người ủng hộ của Đảng Dân chủ. Bây giờ, liệu lợi thế số lượng này có thực sự chuyển thành nguồn vốn tài chính lớn hơn cho Barack Obama hay Mitt Romney không, và quy mô quyên góp trung bình của họ khác nhau như thế nào?

**Trước tiên,** thực hiện lọc DataFrame để chỉ giữ lại các giao dịch quyên góp dành cho hai ứng cử viên Tổng thống chính là Barack Obama và Mitt Romney.

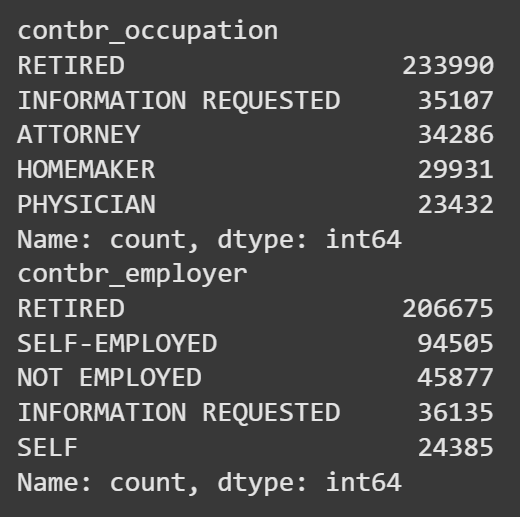


Hình 3.2.1. Hiển thị Tổng số vốn quyên góp và quy mô quyên góp trung bình

* **Kết quả phân tích:**
* Barack Obama dẫn đầu với 136 Triệu USD, cao hơn Mitt Romney (88 Triệu USD). Điều này nói lên rằng, chiến dịch của Obama dựa trên mô hình "Tần suất cao, quy mô nhỏ", huy động vốn từ một lượng lớn người ủng hộ cá nhân (phản ánh kết quả Phân tích Đơn Biến), giúp ông đạt được tổng nguồn vốn cao hơn.
* Mitt Romney có quy mô quyên góp trung bình là 840 Triệu USD, cao gấp gần 3.6 lần so với Barack Obama (231 Triệu USD). Điều này cho thấy chiến dịch của Romney dựa trên mô hình "Tần suất thấp, quy mô lớn", cho thấy nguồn vốn của ông phụ thuộc chủ yếu vào các nhà tài trợ có khả năng đóng góp những khoản tiền lớn hơn, dẫn đến số tiền trung bình cao gấp gần 3.6 lần so với Obama.
* **Insight rút ra được:** Lợi thế về tần suất của Obama đã thành công trong việc vượt qua lợi thế về quy mô đóng góp của Romney, mang lại cho ông tổng nguồn vốn lớn hơn trong cuộc đua.

Chúng ta đã thấy Obama có tổng nguồn vốn lớn hơn nhờ các khoản quyên góp nhỏ. Bây giờ, để hiểu rõ hơn về hệ sinh thái tài trợ. "***Câu hỏi đặt ra:*** Các nghề nghiệp và cơ quan/công ty nào là nguồn quyên góp tài chính lớn nhất cho hai ứng cử viên, và sự phân bổ nguồn vốn trên 2 triệu USD khác biệt như thế nào giữa Đảng Dân chủ và Đảng Cộng hòa?"

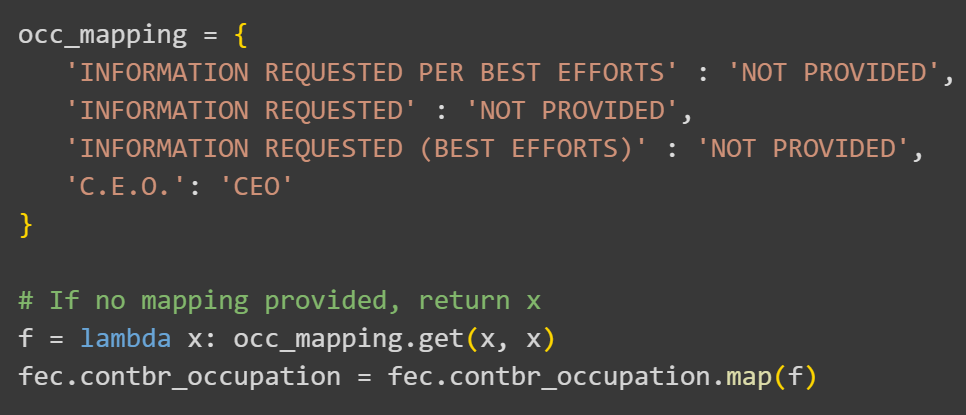
Để trả lời chính xác câu hỏi về nguồn gốc tài trợ. Trước tiên, chúng ta phải giải quyết vấn đề dữ liệu không nhất quán trong các biến phân loại. Cụ thể, các cột Nghề nghiệp (contbr\_occupation) và Cơ quan/Công ty (contbr\_employer) chứa nhiều giá trị không rõ ràng (INFORMATION REQUESTED) hoặc các biến thể của cùng một ý nghĩa (SELF, SELF EMPLOYED). Mục đích là để đảm bảo khi tính toán tổng số tiền (SUM) theo nghề nghiệp, tất cả các khoản quyên góp thuộc cùng một nhóm sẽ được tổng hợp chính xác, tránh làm sai lệch kết quả phân tích nguồn vốn lớn (trên $2 triệu).



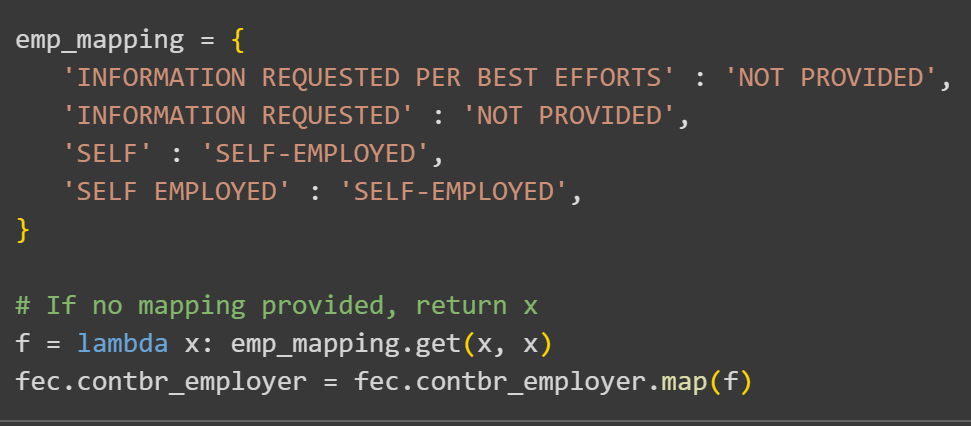
Hình 3.2.2. Hiển thị top 5 tổng số lượng của các cột Nghề nghiệp và Cơ quan

Cách thực hiện nhất quán các biến phân loại:

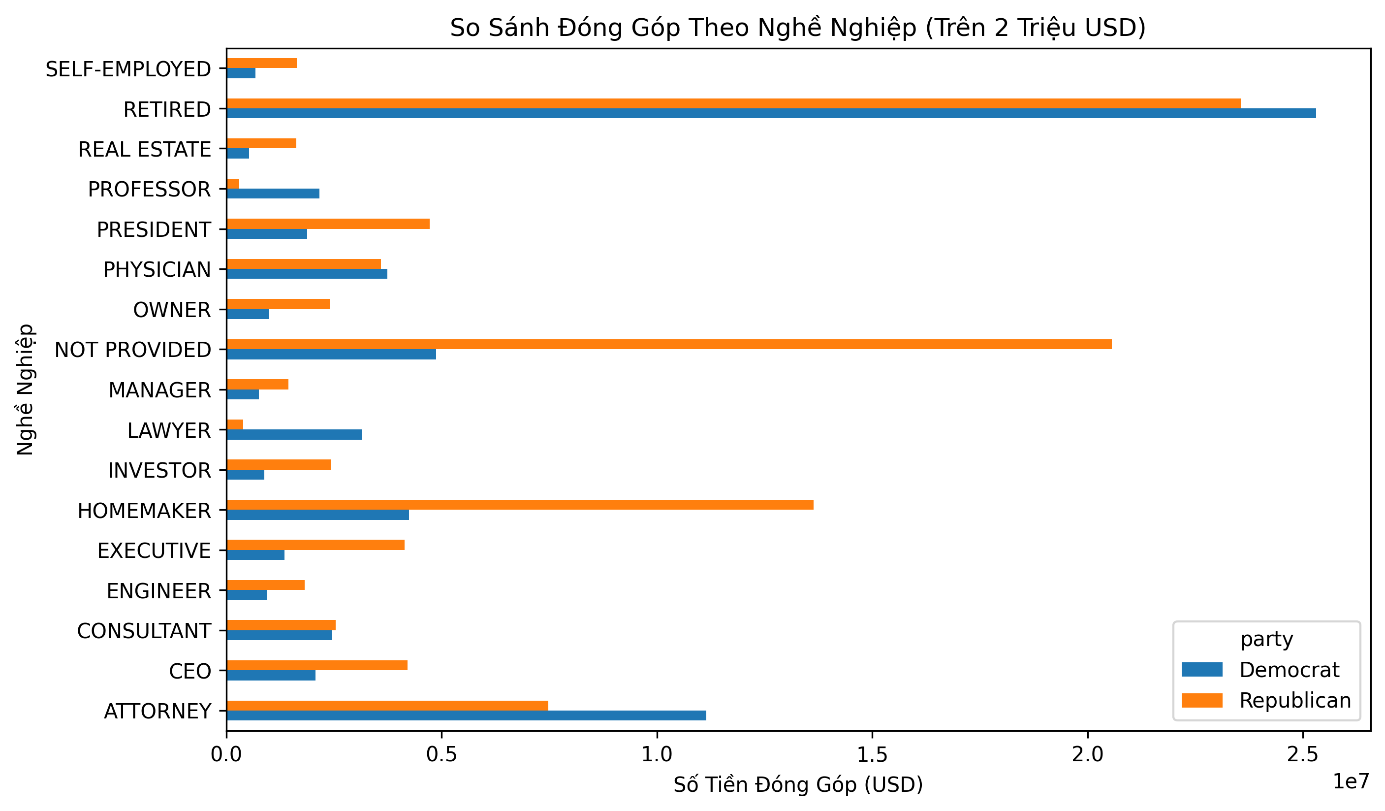
* Bước 1: Chuẩn hóa Nghề nghiệp (contbr\_occupation): Hợp nhất các biến thể của cùng một nghề nghiệp và nhóm tất cả các giá trị thiếu thông tin (INFORMATION REQUESTED) thành các danh mục chuẩn 'CEO' và 'NOT PROVIDED'. Việc này là cần thiết để tổng hợp dữ liệu chính xác và tránh sai lệch khi tính tổng nguồn vốn.



* Bước 2: Chuẩn hóa Cơ quan/Công ty (contbr\_employer): Hợp nhất các biến thể của 'Tự kinh doanh' (SELF, SELF EMPLOYED), nhóm các giá trị thiếu thông tin thành 'NOT PROVIDED', và xác định rõ 'UNEMPLOYED' (NOT EMPLOYED). Việc này loại bỏ lỗi phân loại và đảm bảo tổng hợp dữ liệu chính xác.



Sau khi đã chuẩn hóa các cột Nghề nghiệp và Cơ quan/Công ty, dữ liệu của chúng ta đã sẵn sàng để trả lời cho câu hỏi được đặt ra ở bên trên.

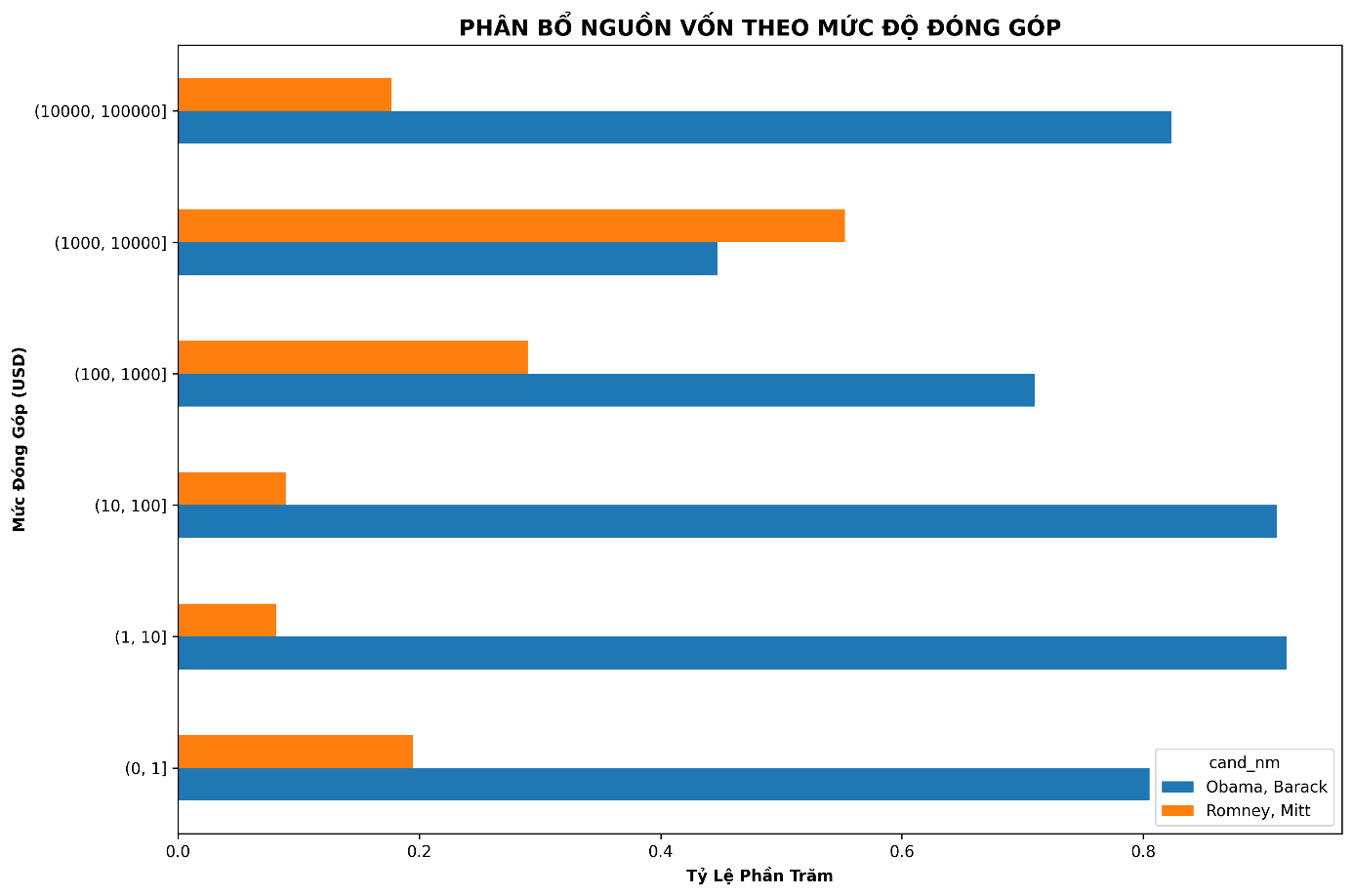


Hình 3.2.3. Hiển thị so sánh đóng góp theo nghề nghiệp

(Trên 2 Triệu USD) của 2 bang

* **Kết quả phân tích:** Dựa vào biểu đồ bên trên: cho thấy sự phân bổ nguồn vốn lớn (trên $2 triệu USD) giữa hai đảng có sự khác biệt rõ rệt, làm nổi bật hai hệ sinh thái tài trợ đối lập:
* Nguồn vốn Chiếm ưu thế của Cộng hòa (Cam): Đảng Cộng hòa nhận được nguồn vốn lớn hơn đáng kể từ các nhóm nghề nghiệp phi chuyên môn và tài chính/đầu tư. Cụ thể, các nhóm INVESTOR(Nhà đầu tư), REAL ESTATE(Bất động sản), PRESIDENT(Chủ tịch), OWNER(Chủ sở hữu), CEO, MANAGER, và HOMEMAKER (Nội trợ) là những nguồn đóng góp lớn nhất, cho thấy sự ủng hộ mạnh mẽ từ tầng lớp gia đình giàu có.
* Nguồn vốn Chiếm ưu thế của Dân chủ (Xanh): Đảng Dân chủ nhận được nguồn vốn mạnh mẽ nhất từ giới chuyên môn và pháp lý. Nguồn tiền từ ATTORNEY/LAWYER (Luật sư) chiếm ưu thế rõ rệt cho Đảng Dân chủ, cùng với sự ủng hộ từ các nghề nghiệp chuyên môn khác như PROFESSOR (Giáo sư) và PHYSICIAN (Bác sĩ).
* **Insight Rút Ra:** Phân tích nguồn vốn trên $2 triệu USD xác nhận mô hình tài trợ của Obama (Dân chủ) dựa trên tầng lớp chuyên môn và trí thức (đặc biệt là Luật sư), trong khi Romney (Cộng hòa) dựa vào tầng lớp gia đình có tài sản và tài chính/đầu tư. Cả hai đảng đều nhận được nguồn tiền lớn nhất từ nhóm RETIRED. Nhưng điểm phân biệt cốt lõi nằm ở các nhóm còn lại: Democrat(Dân chủ) thắng ở giới Luật và Republican(Cộng hòa) thắng ở giới Tài chính/Nội trợ.

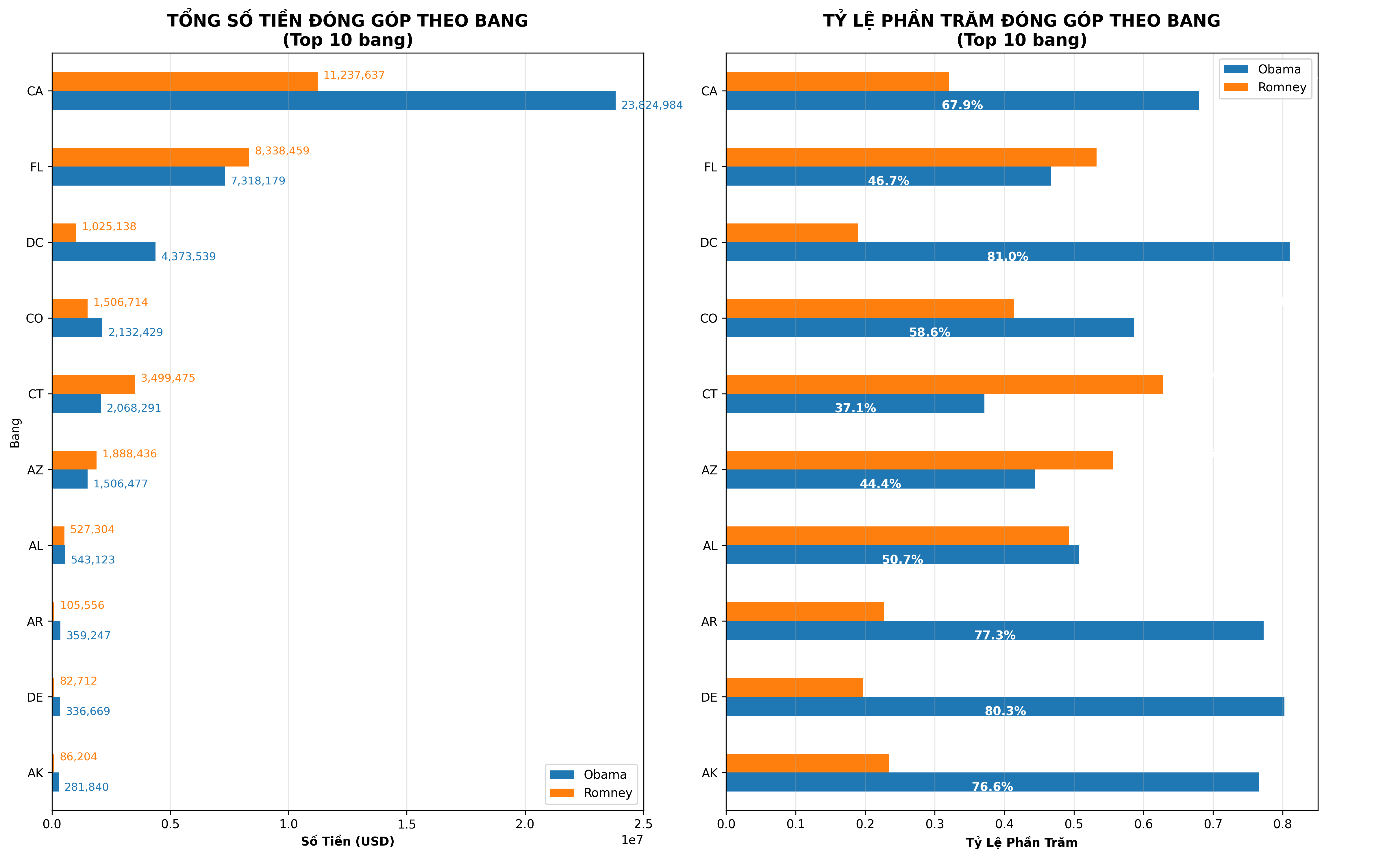
Sau khi phân tích các nguồn quyên góp lớn từ các nghề nghiệp và công ty, chúng ta đã thấy rõ sự khác biệt trong hệ sinh thái tài trợ giữa hai đảng. Tuy nhiên, một câu hỏi quan trọng khác cần được giải đáp: ***Chiến lược gây quỹ của từng ứng cử viên dựa vào các khoản đóng góp nhỏ từ quần chúng hay các khoản đóng góp lớn từ giới tài phiệt? Để trả lời câu hỏi này, chúng ta sẽ phân tích sự phân bổ các khoản đóng góp theo các mức độ khác nhau, từ những khoản chỉ 1 USD cho đến những khoản hàng triệu USD.***



Hình 3.2.4. Hiển thị sự phân bổ nguồn gốn theo mức độ đóng góp

* **Kết quả phân tích:** Dựa vào biểu đồ phân bổ nguồn vốn theo mức độ đóng góp cho thấy sự khác biệt rõ rệt trong chiến lược gây quỹ của hai ứng cử viên:
* Obama (Đảng Dân chủ) – Chiến lược “từ dưới lên” kết hợp sự đa dạng: Biểu đồ cho thấy ông Barack Obama chiếm ưu thế rõ rệt ở các mức đóng góp nhỏ và trung bình, từ 0 đến 1,000 USD, với tỷ trọng dao động từ khoảng 80% đến hơn 90% ở các nhóm này. Điều này phản ánh chiến lược gây quỹ dựa trên số lượng lớn người ủng hộ, đặc biệt từ tầng lớp trung lưu và bình dân. Đáng chú ý, ở nhóm đóng góp lớn từ 10,000 đến 100,000 USD, Obama vẫn chiếm tỷ trọng cao hơn Romney. Điều này cho thấy bên cạnh chiến lược "quần chúng", ông cũng thu hút được sự ủng hộ đáng kể từ các nhà tài trợ lớn.
* Romney (Đảng Cộng hòa) – Tập trung vào nhóm đóng góp trung bình và vừa: Ông Mitt Romney chỉ thực sự chiếm ưu thế ở nhóm 1,000 đến 10,000 USD, và có tỷ trọng đáng kể trong nhóm 100 đến 1,000 USD. Ở các mức thấp hơn (dưới 100 USD), tỷ lệ đóng góp cho ông rất thấp. Và đặc biệt, ở nhóm cao nhất (10,000–100,000 USD), ông không chiếm ưu thế so với Obama – trái với nhận định phổ biến rằng ông dựa chủ yếu vào giới siêu giàu. Chiến lược của Romney phản ánh sự tập trung vào các khoản tài trợ từ tầng lớp trung lưu khá giả, giới kinh doanh vừa, thay vì dựa vào quy mô quần chúng rộng lớn như Obama.
* **Insight Rút Ra:** Obama thành công với mô hình “từ dưới lên”, dựa vào sự đóng góp của đông đảo người dân - vừa đảm bảo nguồn lực tài chính mạnh mẽ, vừa tạo tính lan tỏa và kết nối với cử tri. Đồng thời, ông cũng thu hút được các khoản đóng góp lớn, cho thấy khả năng cân bằng chiến lược gây quỹ. Romney theo đuổi chiến lược tập trung vào nhóm đóng góp vừa và trung bình – hiệu quả về mặt tài chính nhưng ít thể hiện tính đại chúng, và không chiếm ưu thế ở cả hai đầu cực (nhỏ nhất và lớn nhất).

Sau khi đã làm rõ sự khác biệt trong chiến lược gây quỹ theo mức độ đóng góp, một khía cạnh quan trọng không kém cần được khám phá là: Sự ủng hộ tài chính không chỉ phản ánh chiến lược vận động mà còn cho thấy 'sức nóng' của cuộc đua tại từng địa phương. ***Vậy, bang nào đang dẫn đầu về mức độ hậu thuẫn tài chính cho mỗi ứng cử viên, và liệu có sự tương quan giữa nguồn tiền và kết quả bầu cử?***



Hình 3.2.5. Hiển thị sự phân bổ nguồn gốn theo mức độ đóng góp

* **Kết quả phân tích:**
* Từ biểu đồ "Tổng số tiền đóng góp theo bang" cho thấy California (CA) là "cỗ máy gây quỹ" áp đảo với tổng đóng góp lên tới 34.6 triệu USD, trong đó Romney nhận được 23.4 triệu USD, gấp đôi so với Obama (11.2 triệu USD). Florida (FL) thể hiện rõ tính chất bang chiến trường khi có sự cạnh tranh sít sao giữa hai ứng cử viên (8.3 triệu vs 7.3 triệu USD). Đáng chú ý, Romney chiếm ưu thế tại thủ đô Washington DC với 4.4 triệu USD, gấp hơn 4 lần so với Obama.
* Từ biểu đồ "Tỷ lệ phần trăm đóng góp theo bang" cho thấy các bang Colorado (CO), Connecticut (CT) và Arizona (AZ) có sự cân bằng tương đối trong đóng góp, với Obama có phần nhỉnh hơn. Trong khi đó, những bang có quy mô nhỏ hơn như Alaska (AK), Delaware (DE) và Arkansas (AR) vẫn có đóng góp đáng kể, chứng tỏ khả năng gây quỹ diện rộng của cả hai chiến dịch. Xu hướng chung cho thấy Romney có lợi thế ở các trung tâm tài chính lớn, còn Obama duy trì được sự ủng hộ rộng rãi tại nhiều địa phương với tỷ lệ phân bổ đồng đều hơn.
* **Insight quan trọng:** Romney chiếm ưu thế tại các trung tâm quyền lực và tài chính, trong khi Obama duy trì được sự ủng hộ rộng rãi ở cấp độ địa phương. Điều này không chỉ cho thấy một cuộc bầu cử cạnh tranh khốc liệt mà còn dự báo về sự chia rẽ sâu sắc trong cử tri Mỹ, nơi mỗi ứng cử viên đều có những "pháo đài" tài chính vững chắc nhưng ở các khu vực địa lý khác biệt.

## 4. Kết luận

Quá trình phân tích khám phá dữ liệu (EDA) đã tiết lộ bức tranh toàn diện về chiến lược gây quỹ và cơ cấu ủng hộ trong cuộc bầu cử tổng thống 2012:

* **Về hiệu quả gây quỹ:** Barack Obama xuất sắc vượt trội với chiến lược "tần suất cao - quy mô nhỏ", huy động thành công 59.3% tổng số giao dịch từ đông đảo quần chúng, giúp ông đạt tổng số tiền 136 triệu USD, vượt xa Mitt Romney (88 triệu USD) dù quy mô đóng góp trung bình chỉ bằng 1/3.6 so với đối thủ. Điều này cho thấy một mô hình gây quỹ bền vững và có tính lan tỏa cao.
* **Về cơ cấu ủng hộ:** Hai ứng cử viên thể hiện rõ sự phân cực trong cơ sở tài trợ. Obama được hậu thuẫn mạnh mẽ bởi giới chuyên môn (luật sư, giáo sư, bác sĩ) và cử tri bình dân với các khoản đóng góp nhỏ dưới 1,000 USD. Ngược lại, Romney tập trung vào tầng lớp doanh nhân, nhà đầu tư và gia đình giàu có với các khoản đóng góp trung bình từ 1,000-10,000 USD.
* **Về phân bổ địa lý:** Chiến lược gây quỹ cũng thể hiện rõ theo không gian địa lý. Romney thống trị tại các trung tâm tài chính lớn như California và Washington DC, trong khi Obama duy trì được sự ủng hộ rộng rãi và cân bằng across nhiều bang, từ các chiến trường then chốt đến các địa phương nhỏ.

**Thành công then chốt:** Thành công của Obama không chỉ nằm ở số tiền huy động được, mà ở cách ông xây dựng được một phong trào quần chúng. Sự kết hợp giữa sức mạnh tài chính từ các khoản đóng góp nhỏ và sự ủng hộ rộng khắp chính là nền tảng vững chắc, tạo ra lợi thế toàn diện và là chỉ báo quan trọng cho một chiến thắng trong cuộc bầu cử.

# KẾT LUẬN CHUNG

Thông qua việc phân tích năm bộ dữ liệu đa dạng, báo cáo này đã minh họa một cách sinh động sức mạnh và tính linh hoạt của việc áp dụng các kỹ thuật phân tích dữ liệu vào những lĩnh vực hoàn toàn khác biệt. Mỗi chuyên đề đều khai thác một góc nhìn riêng, từ hành vi cá nhân đến xu hướng xã hội và động lực kinh tế - chính trị, nhưng tất cả đều cùng chung một mục đích: chuyển hóa dữ liệu thô thành những hiểu biết có giá trị và có thể hành động được.

* Từ Góc Nhìn Vi Mô: Hành vi và Sở thích Cá nhân: Phân tích dữ liệu Bitly và MovieLens cho thấy sự phong phú trong thông tin ẩn chứa sau những hành vi hàng ngày của người dùng. Dữ liệu duyệt web (time zone, OS) và đánh giá phim (theo giới tính, độ tuổi) không chỉ giúp hiểu rõ đối tượng mà còn là chìa khóa để cá nhân hóa trải nghiệm và cải thiện dịch vụ.
* Đến Góc Nhìn Vĩ Mô: Xu hướng Văn hóa và Xã hội Bộ dữ liệu US Baby Names đã vẽ nên một bức tranh sống động về sự biến đổi văn hóa và xã hội Hoa Kỳ trong hơn một thế kỷ. Sự gia tăng tính đa dạng trong tên gọi và sự biến động về "thời thượng" của các chữ cái cuối phản ánh sự thay đổi trong quan điểm của cha mẹ, sự hội nhập văn hóa và cả sự năng động của xã hội.
* Ứng dụng trong Đời sống và Sức khỏe: Phân tích USDA Food Database vượt ra ngoài khuôn khổ kinh doanh, đi sâu vào lĩnh vực sức khỏe và dinh dưỡng. Bằng cách xác định các thực phẩm giàu dinh dưỡng nhất, phân tích này cung cấp một công cụ hữu ích cho các chuyên gia dinh dưỡng, nhà sản xuất thực phẩm và người tiêu dùng quan tâm đến sức khỏe.
* Khám phá Động lực Kinh tế - Chính trị: Bộ dữ liệu FEC 2012 đã "giải phẫu" nguồn lực tài chính đằng sau một cuộc bầu cử tổng thống, tiết lộ sự khác biệt rõ rệt trong chiến lược gây quỹ và cơ sở ủng hộ giữa các đảng phái. Điều này không chỉ có ý nghĩa với các nhà chiến lược chính trị mà còn rất quan trọng đối với công chúng trong việc hiểu rõ về sự minh bạch và ảnh hưởng của tiền bạc trong chính trị.

Kết luận cuối cùng, mặc dù các chuyên đề phân tích này độc lập về mặt chủ đề, chúng đều được kết nối bởi một chu trình phân tích dữ liệu thống nhất: Từ việc làm sạch và chuẩn bị dữ liệu, thông qua việc áp dụng các kỹ thuật thống kê và trực quan hóa, để cuối cùng đi đến diễn giải và đưa ra kết luận. Báo cáo đã chứng minh rằng, bất kể lĩnh vực nào, tư duy phân tích dữ liệu và sự thành thạo các công cụ như Python, Pandas, và NumPy chính là chìa khóa để mở ra những kho tàng thông tin quý giá, góp phần vào việc ra quyết định sáng suốt hơn trong mọi mặt của đời sống kinh tế - xã hội.