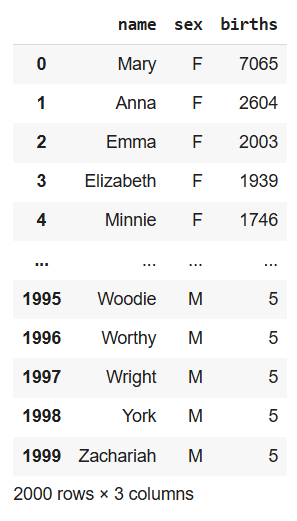
**(Heading 1)**

**13.3 US Baby Names 1880–2010**

***Import thư viện và nạp dữ liệu***

Bảng 1: Dữ liệu sơ bộ năm 1880



**Nhận xét**

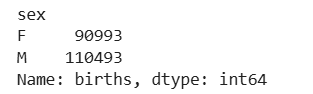
Tải và xem sơ bộ dữ liệu gốc năm 1880

Dữ liệu có 3 cột chính:

name: Tên của trẻ em được ghi nhận.  
sex: Giới tính của đứa trẻ là "M" (nam) hoặc "F" (nữ).  
births: Số lượng trẻ sinh ra có tên đó.

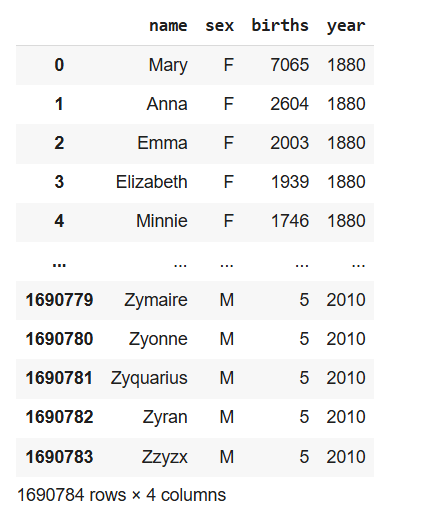
***Tổng hợp số lượng sinh theo giới tính***

Bảng 2: Tổng số lượng sinh năm 1880



***Hợp nhất dữ liệu từ 1880 đến 2010***

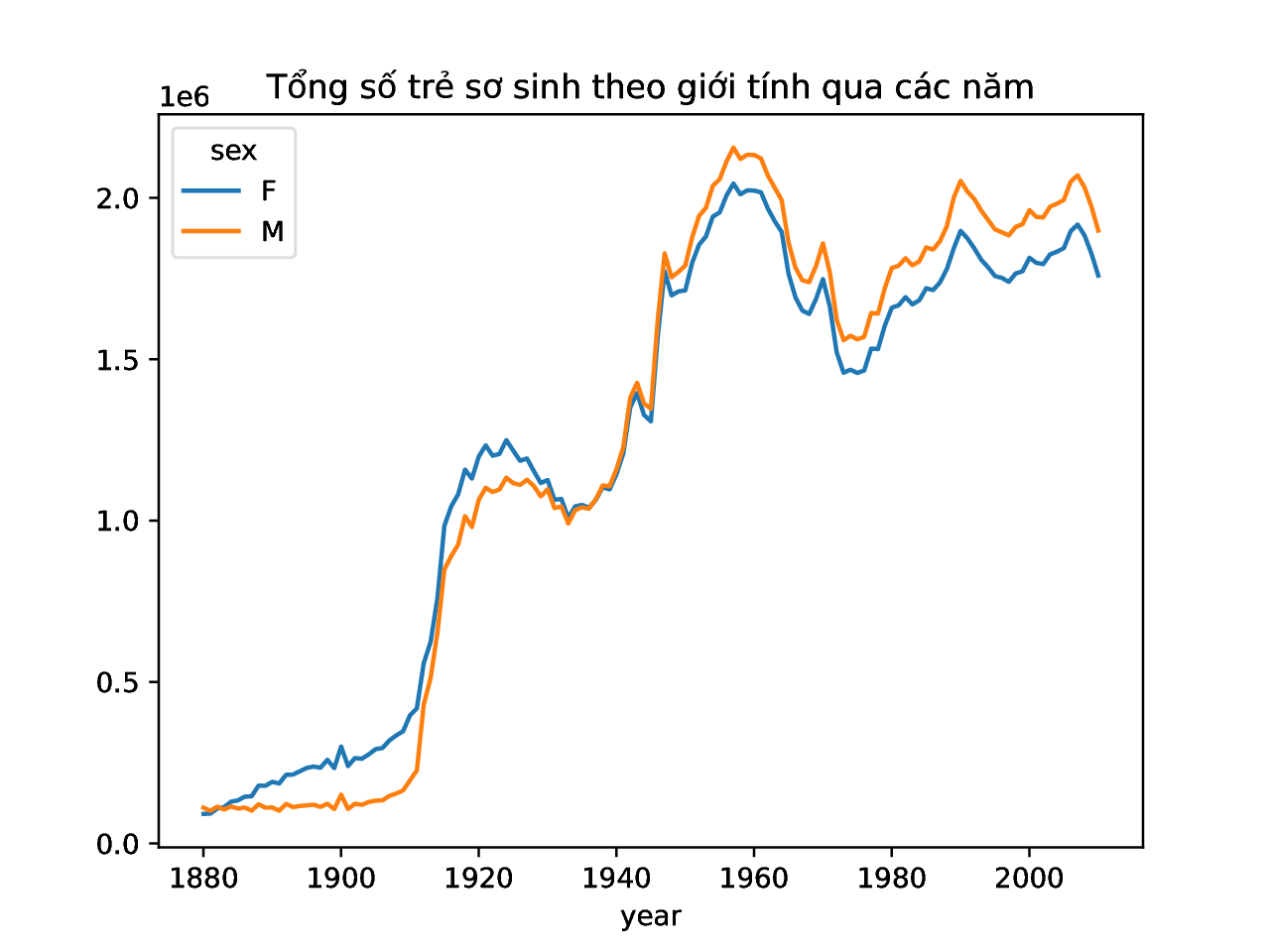
Bảng 3: Dữ liệu đã gộp từ năm 1880 - 2010



**Nhận xét:**

Đọc dữ liệu tên trẻ em từ năm 1880 đến 2010, thêm cột “year” để ghi lại năm tương ứng

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



Hình 1: Biểu đồ tổng số lượng trẻ theo giới tính qua các năm

**Nhận xét:**

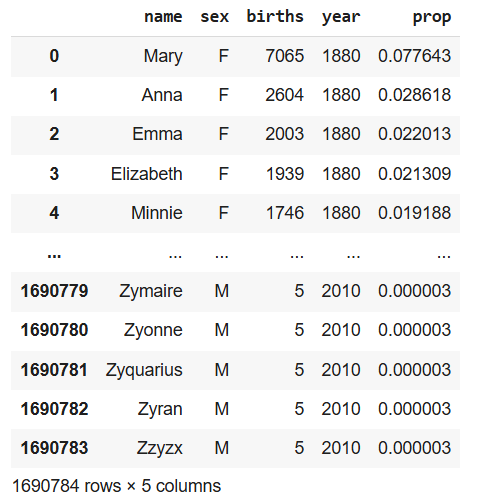
Biểu đồ cho thấy số trẻ sinh tăng mạnh dần qua các năm

Trẻ nam (M) luôn nhiều hơn trẻ nữ (F) trong hầu hết các năm.

Khoảng cách giữa hai đường F và M thường khá ổn định, cho thấy xu hướng sinh con trai cao hơn con gái là phổ biến.

***Tính tỉ lệ phổ biến của từng tên***

Bảng 4: Tỉ lệ phổ biến của từng tên



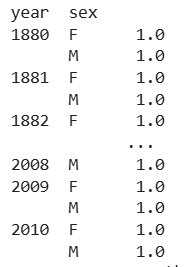
**Nhận xét:**

Thêm cột prop để biểu thị tỉ lệ trẻ mang mỗi tên trong tổng số trẻ sinh cùng năm và giới tính.

Điều này giúp so sánh mức độ phổ biến của các tên theo thời gian, thay vì chỉ dựa vào số lượng tuyệt đối.

***Kiểm tra tổng tỉ lệ (prop) theo năm và giới tính***

Bảng 5:  Tỉ lệ prop theo năm và giới tính

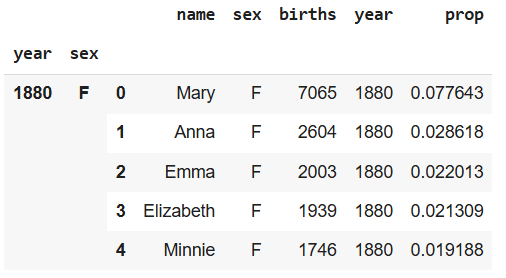


**Nhận xét:**

Kiểm tra xem tổng tỉ lệ prop trong mỗi nhóm (năm + giới tính) có bằng 1 hay không.

***Lấy 1000 tên phổ biến nhất theo năm và giới tính***

Bảng 6: 1000 tên phổ biến nhất theo năm và giới tính

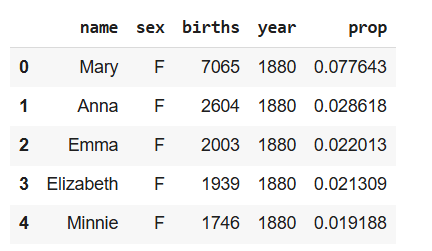


**Nhận xét:**

Lấy danh sách nhằm phục vụ cho việc phân tích xu hướng tên qua thời gian.

***Đặt lại chỉ mục cho bảng Top 1000 tên***

Bảng 7: Chỉ mục cho bảng Top 1000 tên sau khi đặt lại



**Nhận xét:**

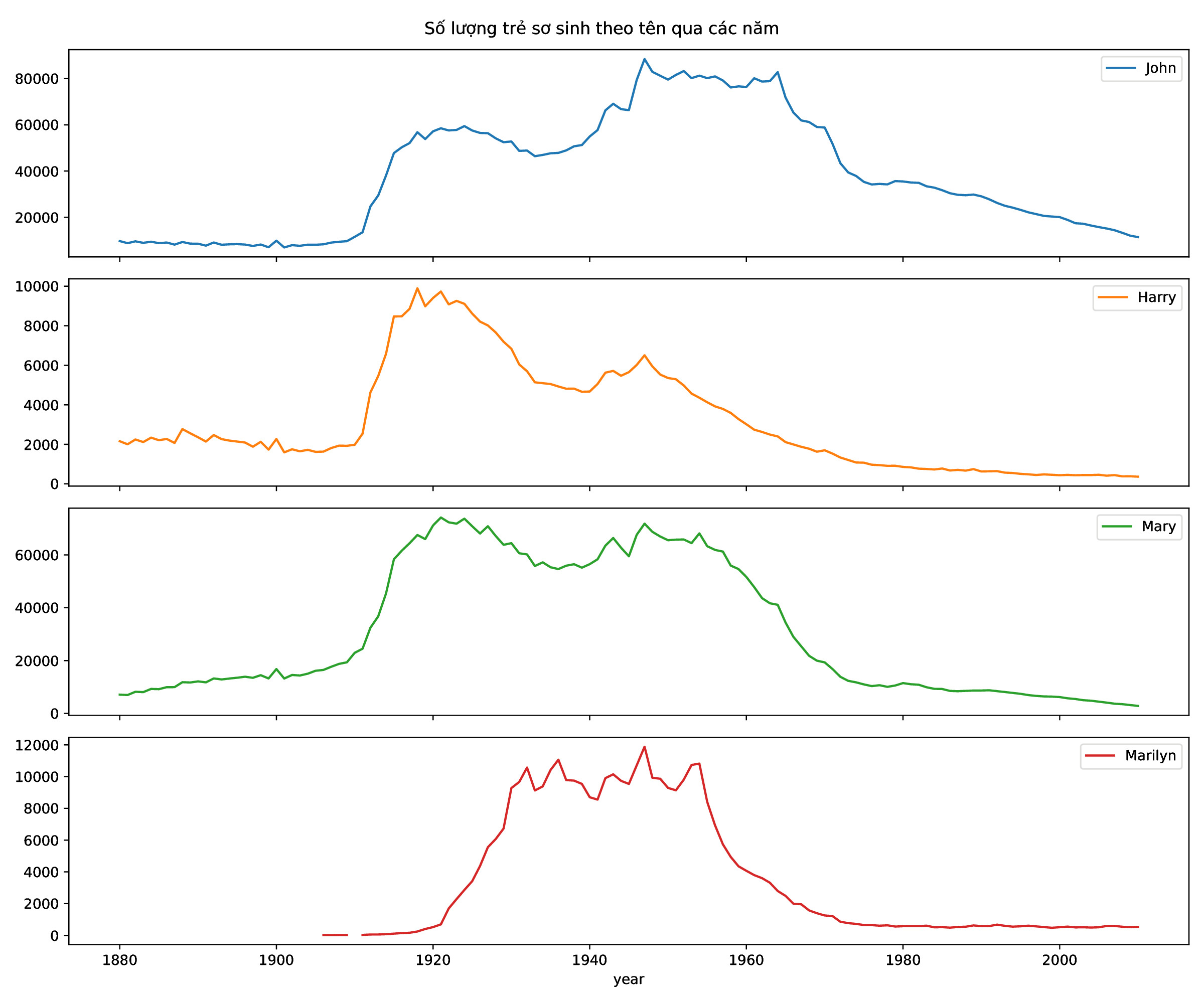
Nhằm xóa các chỉ mục cũ và tạo lại chỉ mục tuần tự mới giúp dữ liệu gọn gàng, dễ truy cập và trực quan hơn khi hiển thị.

**(Heading 2)**

**Analyzing Naming Trends**

*Phân tích xu hướng đặt tên*

***Vẽ biểu đồ xu hướng của một số tên tiêu biểu***



Hình 2: Biểu đồ số lượng trẻ của 1 số tên tiêu biểu của trẻ qua các năm

**Nhận xét:**

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.

**(Heading 3)**

**Measuring the increase in naming diversity**

*Đo lường mức độ đa dạng của tên theo thời gian*

***Tổng tỉ lệ phần trăm trong top 1000 tên theo giới tính qua từng năm***



Hình 3: Biểu đồ tỉ lệ tổng top 1000 tên theo giới tính qua từng năm

**Nhận xét:**

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.

**Kết luận:**

Tên của bé gái đa dạng hơn bé trai, vì tỉ lệ của nhóm nữ giảm nhanh hơn.

Xã hội hiện đại có xu hướng cá nhân hóa tên gọi và ít trùng lặp hơn so với thế kỷ trước.

***Lọc dữ liệu tên bé trai năm 2010***

Bảng 8: Dữ liệu tên bé trai năm 2010

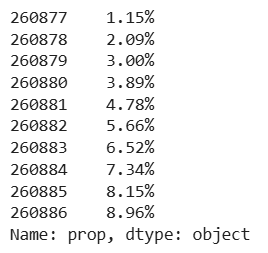


**Nhận xét:**

Dùng để phân tích top 10 tên bé trai năm 2010.

***Tính tích lũy tỉ lệ phần trăm của các tên bé trai năm 2010***

Bảng 9: Tí lệ phần trăm của top 10 tên bé trai năm 2010



**Nhận xét:**

Tên phổ biến nhất chiếm 1.15% tổng số bé trai năm 2010.

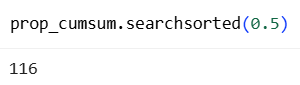
Top 2 tên phổ biến nhất cộng lại chiếm 2.09% tổng số.

Top 3 tên chiếm gần 3.0%.

Top 10 tên phổ biến nhất năm 2010 chỉ chiếm khoảng 9% tổng số trẻ trai, cho thấy các tên khá đa dạng, không tập trung vào vài tên như những năm xưa.

***Xác định số lượng tên chiếm 50% tổng số trẻ sinh trong năm 2010***

Bảng 10: Số lượng tên bé trai năm 2010 chiếm 50% tổng số trẻ sinh



**Nhận xét:**

Cần 117 tên phổ biến nhất để chiếm 50% tổng số trẻ sinh năm 2010. (Vì mảng bắt đầu từ 0)

Điều này cho thấy độ đa dạng tên rất cao, vì phải tới 117 tên mới chiếm nửa dân số bé trai.

***Xác định số lượng tên bé trai năm 1900 chiếm 50% tổng số trẻ sinh***

Bảng 11: Số lượng tên bé trai năm 1900 chiếm 50% tổng số trẻ sinh



**Nhận xét:**

Cần 26 tên phổ biến nhất để chiếm 50% tổng số trẻ sinh năm 1990.

Chứng tỏ nhiều trẻ được đặt cùng vài tên phổ biến. Mức đa dạng tên thấp hơn so với các năm sau.

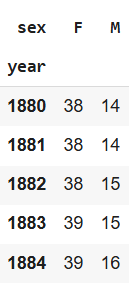
***Tính độ đa dạng tên theo thời gian (điểm phân vị 50%)***

Giúp đo lường mức độ đa dạng tên qua các năm — cụ thể là bao nhiêu tên phổ biến nhất chiếm 50% tổng số trẻ sinh.

Giá trị càng cao → tên càng đa dạng.

Giá trị thấp → tập trung vào ít tên phổ biến hơn.

Bảng 12: Độ đa dạng tên



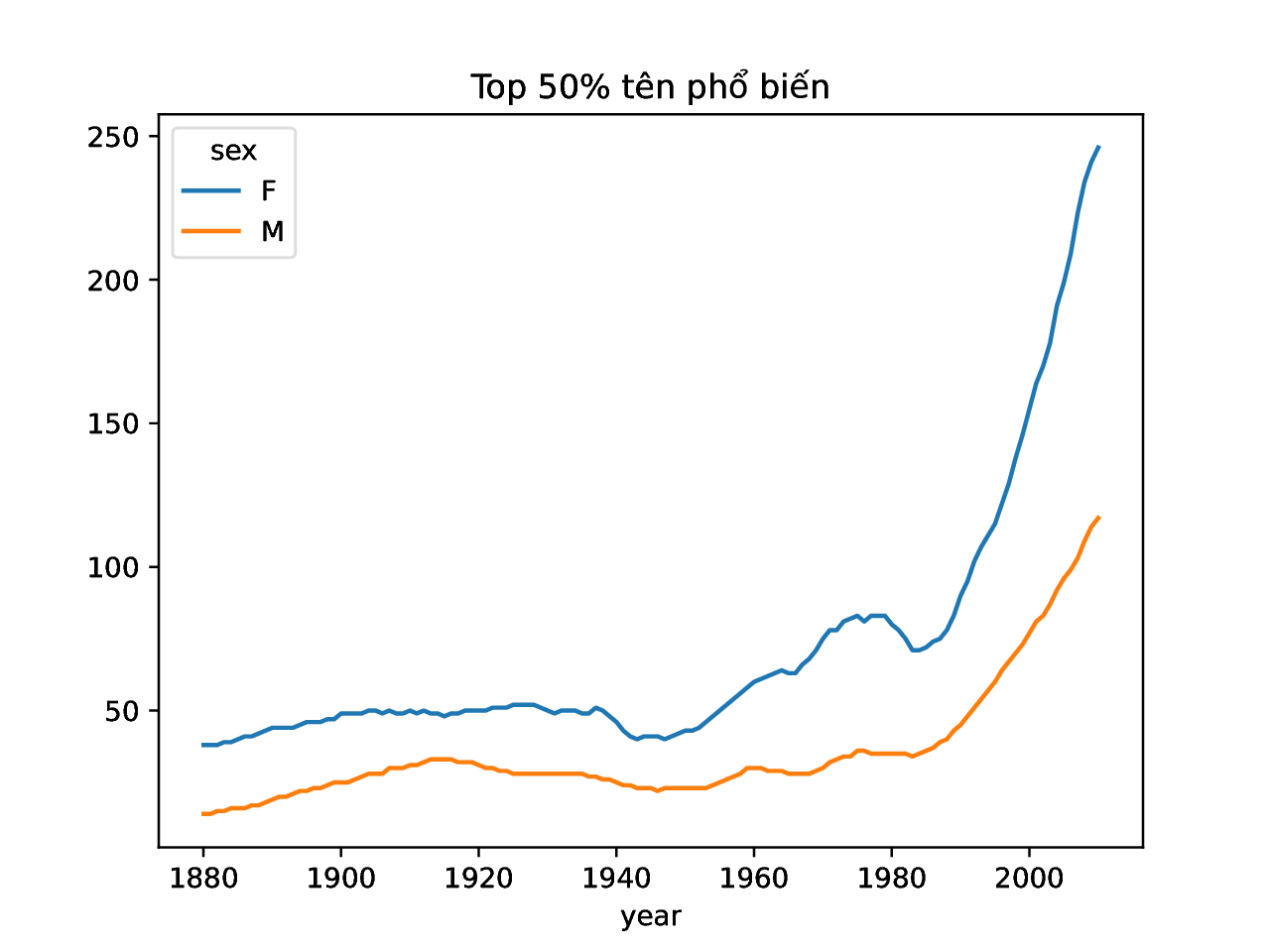
**Nhận xét:**

Năm 1880 cần 38 tên nữ và 14 tên nam để đạt 50% tổng số trẻ sinh

Năm 1884 cần 39 tên nữ và 16 tên nam.

Nhìn chung, số lượng tên nữ luôn lớn hơn số tên nam, cho thấy tên nữ đa dạng hơn so với tên nam.

***Biểu đồ số lượng tên phổ biến chiếm 50% tổng trẻ sinh***



Hình 4: Biểu đồ thể hiện Top 50% tên phổ biến

**Nhận xét:**

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.

**(Heading 3)**

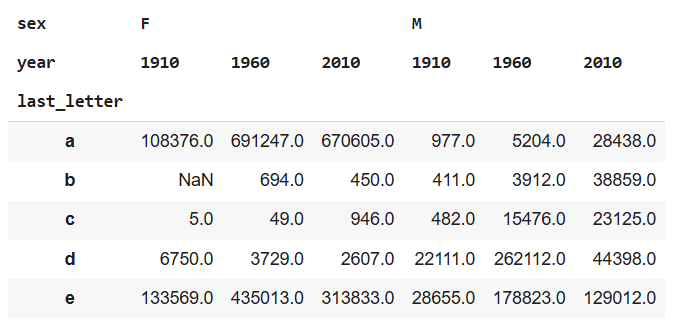
**The “last letter” revolution**

*Cuộc cách mạng chữ cái cuối trong tên*

Ví dụ: “Anna” → “a”, “David” → “d”

***Phân tích chữ cái cuối của tên***

Bảng 13: Dữ liệu chữ cái cuối cho các năm tiêu biểu

******

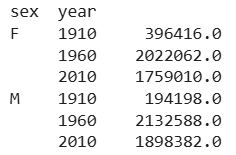
**Nhận xét:**

Chữ “a”: cực phổ biến ở nữ (F), số lượng tăng mạnh từ 108k (1910) → 691k (1960) → 670k (2010). Ở nam (M) gần như không phổ biến.

Chữ "d" ở nam (M) giảm từ 262k (1960) → 44k (2010), cho thấy thay đổi xu hướng tên nam. Ở nữ ít dùng.

***Tổng số trẻ sinh theo chữ cái cuối, giới tính và năm***

Bảng 14: Tổng số trẻ sinh theo chữ cái cuối, giới tính và năm

******

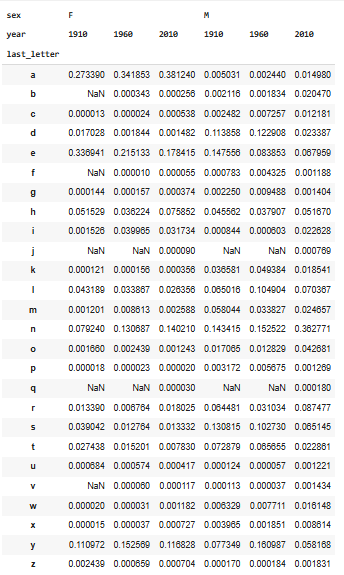
**Nhận xét:**

Nữ (F): số trẻ tăng mạnh từ 1910 (≈ 0.4 triệu) → 1960 (≈ 2 triệu), sau đó giảm nhẹ 2010 (≈ 1.76 triệu).

Nam (M): số trẻ tăng từ 1910 (≈ 0.19 triệu) → 1960 (≈ 2.13 triệu), rồi giảm nhẹ 2010 (≈ 1.9 triệu).

***Chuẩn hóa chữ cái cuối thành tỷ lệ***

Bảng 15: Tỉ lệ số trẻ sinh theo chữ cái cuối, giới tính và năm

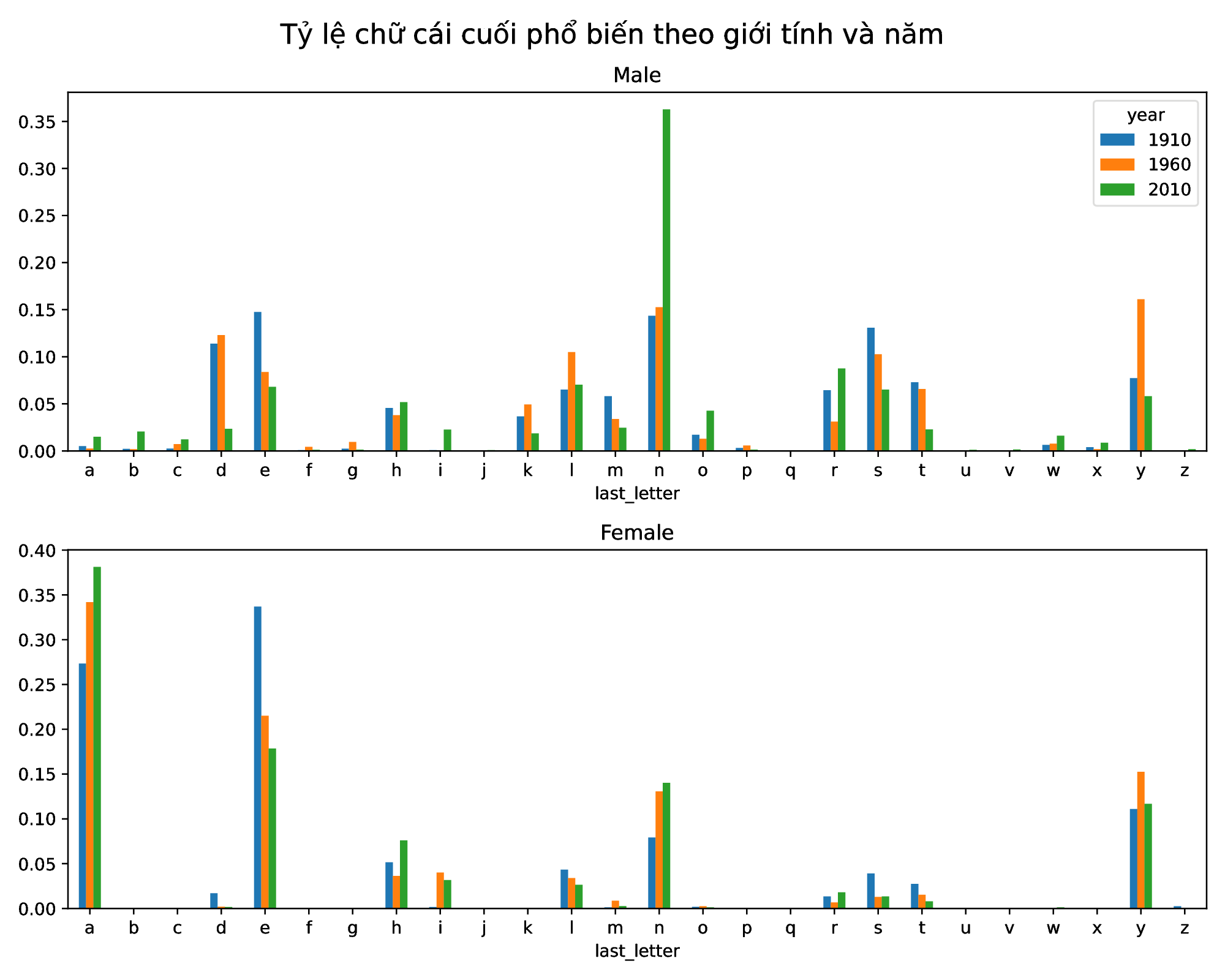


**Nhận xét:**

Nữ (F): chữ cái cuối “a”, “e”, “y” chiếm tỷ lệ cao nhất, với “a” tăng dần từ 1910 → 2010, phản ánh xu hướng tên nữ phổ biến.

Nam (M): chữ cái cuối “d”, “n”, “y” phổ biến nhất, với “n” tăng mạnh năm 2010, cho thấy thay đổi xu hướng tên nam theo thời gian.

***Vẽ biểu đồ cột tỷ lệ chữ cái cuối theo giới tính***



Hình 5: Biểu đồ phân bố chữ cái cuối của tên theo giới tính và năm

**Nhận xét:**

***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.

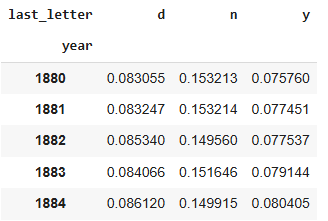
Chữ “e” giảm từ 33.7% (1910) → 17.8% (2010), trong khi “n” tăng từ 7.9% (1910) → 14.0% (2010).

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.

***Lấy tỷ lệ các chữ cái cuối của tên nam theo từng năm, tập trung vào ba chữ cái phổ biến nhất: “d”, “n”, và “y”.***

Bảng 16: Tỷ lệ các chữ cái cuối của tên nam theo từng năm, tập trung vào ba chữ cái phổ biến nhất: “d”, “n”, và “y”

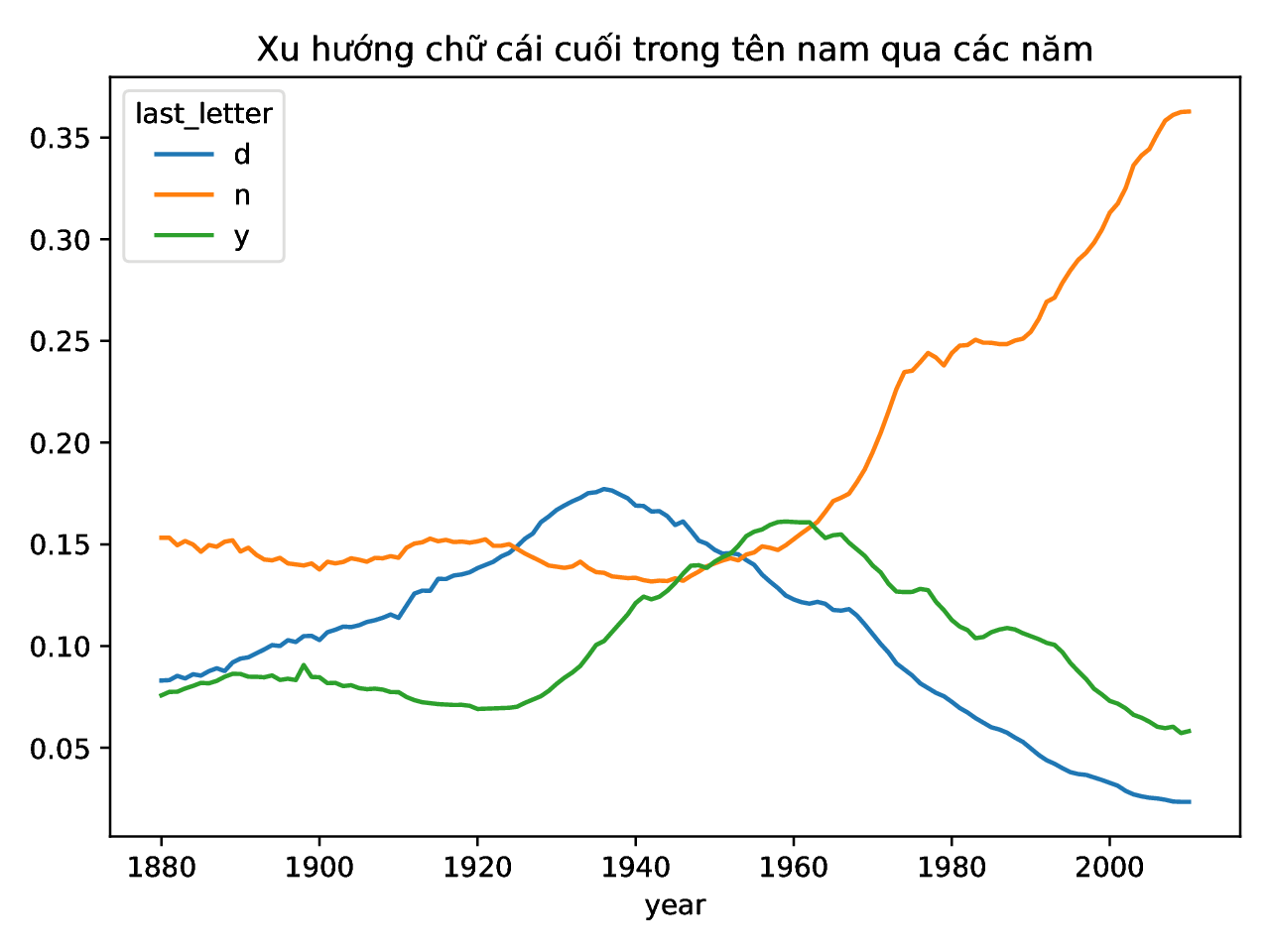
******

**Nhận xét:**

Trong giai đoạn 1880 – 1884, tỷ lệ các tên nam kết thúc bằng chữ “n” luôn cao nhất (khoảng 15%), thể hiện đây là dạng kết thúc phổ biến nhất.

Chữ “d” duy trì quanh mức 8 – 8.6%, ổn định nhẹ qua các năm.

Chữ “y” có xu hướng tăng nhẹ từ 7.6% → 8.0%, cho thấy ngày càng được ưa chuộng hơn trong cách đặt tên nam



Hình 6:Biểu đồ xu hướng tên nam kết thúc bằng d, n, y theo năm

**Nhận xét:**

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.

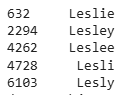
**(Heading 3)**

**Boy names that became girl names (and vice versa)**

*Phân tích các tên từng phổ biến ở giới nam nhưng sau đó trở nên phổ biến ở giới nữ (và ngược lại)*

***Lọc các tên có chứa chuỗi "Lesl" trong danh sách các tên phổ biến (từ top 1000).***

Bảng 17: Các tên có chứa chuỗi "Lesl" trong danh sách các tên phổ biến (từ top 1000).

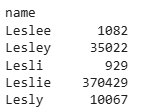
******

**Nhận xét:**

Nhằm xác định các biến thể của tên “Lesl” để phân tích xu hướng giới tính ở bước tiếp theo.

***Chọn các tên liên quan tới chuỗi* "Lesl" *và tính tổng số trẻ sinh theo từng tên***

Bảng 18: Tổng các tên liên quan có chứa chuỗi "Lesl"



**Nhận xét:**

Tên “Leslie” chiếm đa số tuyệt đối với hơn 370 nghìn lượt đặt, cho thấy đây là biến thể phổ biến nhất trong nhóm.

Các biến thể khác như “Lesley” và “Lesly” có tần suất thấp hơn nhiều, còn “Leslee” và “Lesli” rất hiếm gặp.

***Tỷ lệ phần trăm trẻ sinh theo giới tính (F/M) cho các biến thể tên***

Bảng 19: Tỷ lệ phần trăm trẻ sinh theo giới tính có chứa chuỗi "Lesl"

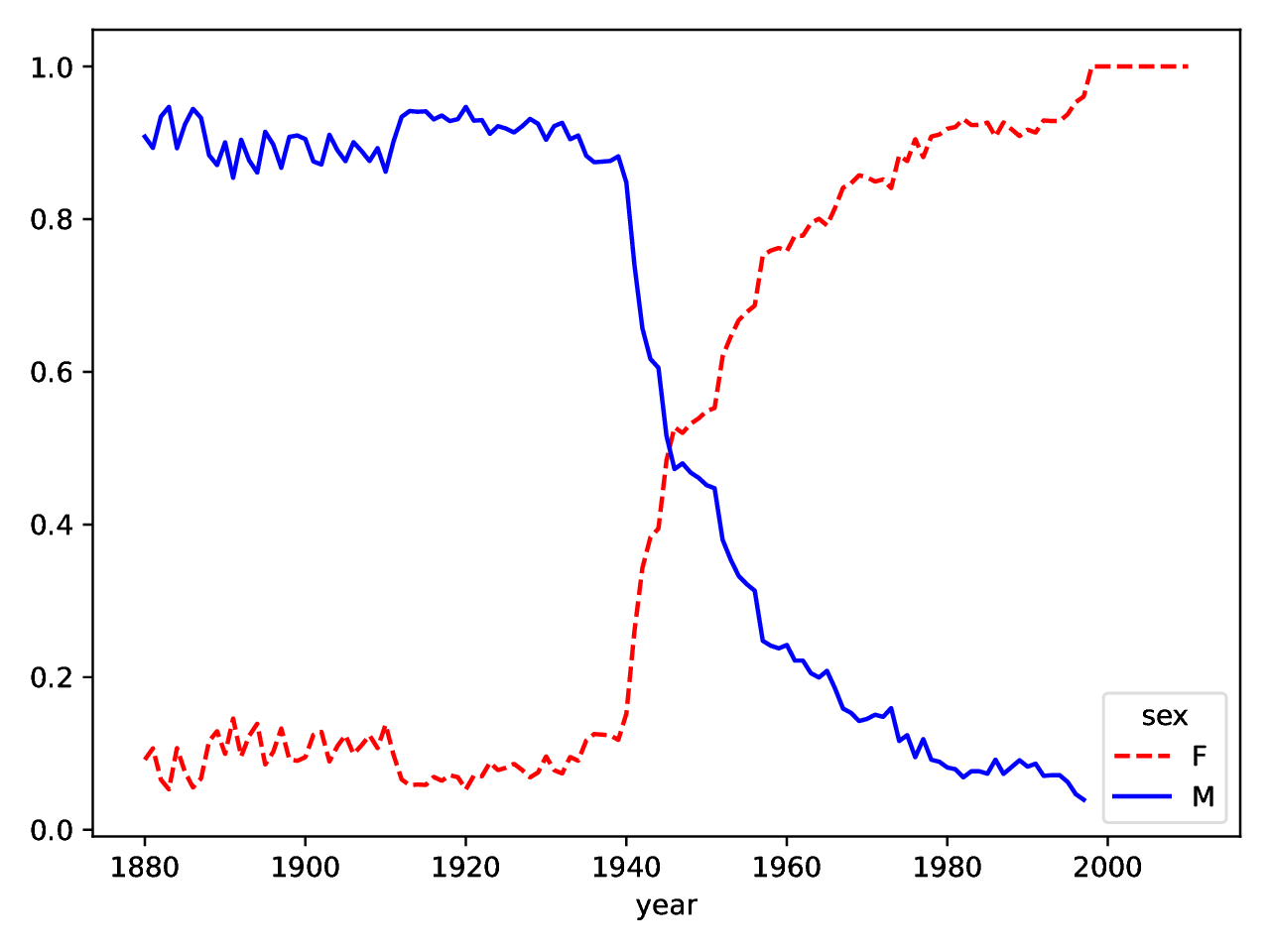


**Nhận xét:**

Trong các năm 2006 – 2010, tất cả trẻ mang tên trong danh sách đều là nữ. Các tên này gần đây chỉ dùng cho nữ, không còn xuất hiện ở nam.

Cột nam (M) = NaN vì không có trẻ nam nào mang các tên này trong các năm đó.

***Vẽ biểu đồ tỷ lệ trẻ sinh theo giới tính***



Hình 7: Biểu đồ thể hiện sự chuyển đổi giới tính của tên 'Lesl' qua thời gian (1880–2010)

**Nhận xét:**

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.

**Kết luận:**

“Lesl” là ví dụ điển hình của một tên chuyển giới tính theo thời gian — từ nam sang nữ.

Biểu đồ minh họa rõ sự “đảo chiều giới tính” của cái tên này: Từ ~1.0 (nam chiếm gần hết) → ~0.0 (nữ chiếm gần hết).

**(Heading1)**

**13.4 USDA Food Database**

***Nạp dữ liệu và kiểm tra độ dài file***

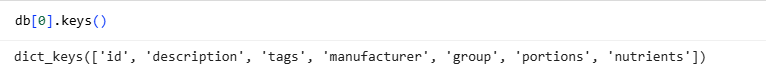


Hình 8: Nạp dữ liệu và kiểm tra độ dài file

**Nhận xét:**

Dữ liệu từ file JSON chứa thông tin thực phẩm của USDA, với tổng cộng 6,636 dòng (mỗi dòng tương ứng với một mục thực phẩm).

***Kiểm tra các trường dữ liệu trong bản ghi đầu tiên của cơ sở dữ liệu USDA***



Hình 9: liệt kê tất cả các khóa trong phần tử đầu tiên của danh sách

**Nhận xét:**

id: Mã định danh duy nhất của thực phẩm trong cơ sở dữ liệu.

description: Tên hoặc mô tả của thực phẩm.

tags: Các thẻ mô tả thêm.

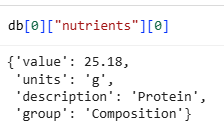
manufacturer: Tên nhà sản xuất hoặc thương hiệu.

group: Nhóm thực phẩm.

portions: Thông tin về khẩu phần.

nutrients: Danh sách các chất dinh dưỡng.

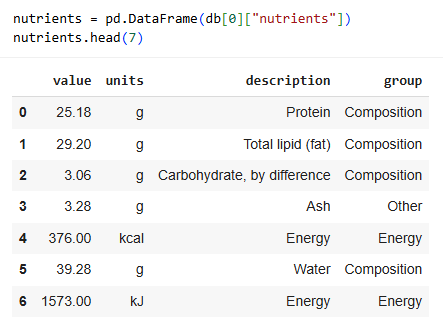
***Truy cập thông tin chất dinh dưỡng đầu tiên của thực phẩm đầu tiên trong cơ sở dữ liệu USDA***



Hình 10: Mô tả phần tử đầu tiên trong danh sách chất dinh dưỡng

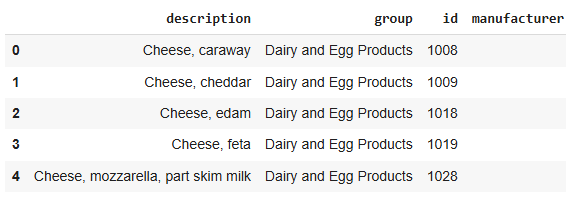
***Chuyển danh sách các chất dinh dưỡng của thực phẩm đầu tiên thành bảng dữ liệu***

Bảng 20: Danh sách các chất dinh dưỡng

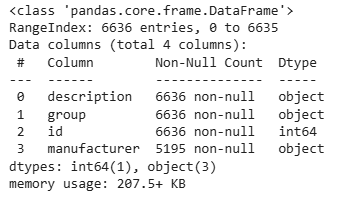


***Tạo bảng thông tin tổng quát từ dữ liệu USDA.***

Bảng 21: bảng thông tin tổng quát từ dữ liệu USDA

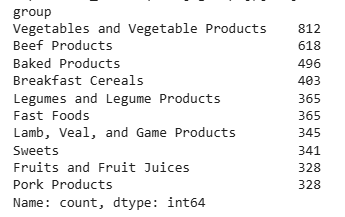


***Kiểm tra thông tin tổng quát về cấu trúc và kiểu dữ liệu của bảng***



Hình 11: Thông tin tổng quan về dữ liệu

***Thống kê 10 nhóm thực phẩm phổ biến nhất trong cơ sở dữ liệu USDA***



Hình 12: 10 nhóm thực phẩm phổ biến nhất trong cơ sở dữ liệu

**Nhận xét:**

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.).

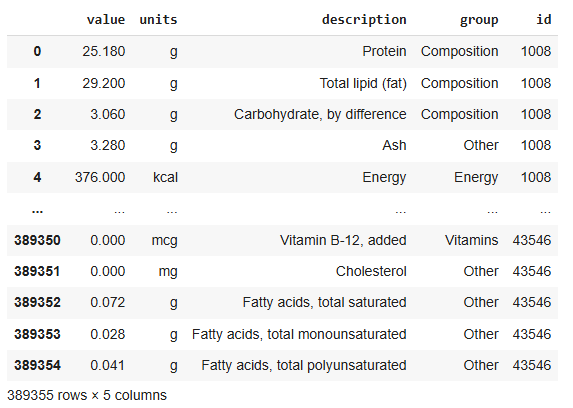
Các nhóm “Baked Products” và “Breakfast Cereals” → Xuất hiện khá nhiều (400–500 mục), cho thấy cơ sở dữ liệu cũng tập trung vào nguồn thực phẩm chế biến phổ biến trong khẩu phần ăn hiện đại.

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.

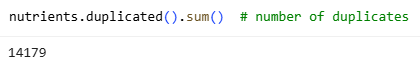
Nhóm “Sweets”, “Fruits and Fruit Juices”, “Pork Products” → Chiếm tỷ lệ nhỏ hơn, nhưng vẫn thể hiện sự đa dạng nguồn thực phẩm từ nhiều nhóm dinh dưỡng khác nhau.

***Tổng hợp toàn bộ dữ liệu chất dinh dưỡng của tất cả thực phẩm trong cơ sở dữ liệu USDA thành một bảng duy nhất.***

Bảng 22: Tổng hợp toàn bộ dữ liệu chất dinh dưỡng của tất cả thực phẩm trong cơ sở dữ liệu USDA



***Kiểm tra số lượng bản ghi bị trùng lặp trong bảng nutrients***

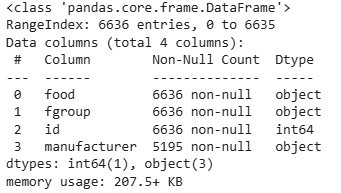


Hình 13: Số lượng dòng bị trùng lặp

**Nhận xét:**

Trong bảng nutrients có 14.179 bản ghi bị trùng lặp hoàn toàn

***Làm sạch dữ liệu chất dinh dưỡng(nutrients) và chuẩn hóa tên cột trong bảng thực phẩm***



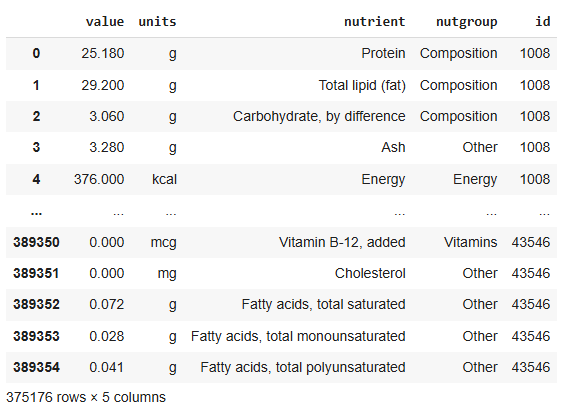
Hình 14:

**Nhận xét:**

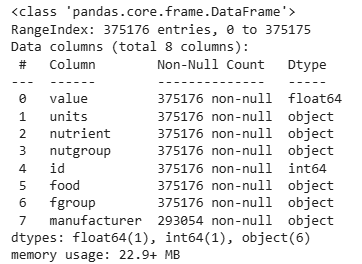
Cột manufacturer có giá trị thiếu 1.441 dòng, khoảng 22% dữ liệu bị thiếu.

***Chuẩn hóa tên cột trong bảng nutrients để dễ phân biệt và tránh trùng tên khi phân tích dữ liệu.***

Bảng 23: Chuẩn hóa tên cột trong bảng nutrients

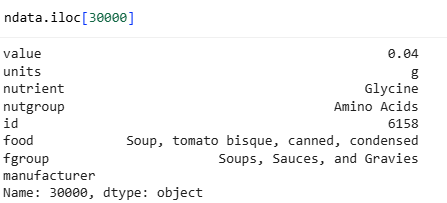


***Gộp hai bảng dữ liệu nutrients và info***

******

Hình 15: Merge dữ liệu dinh dưỡng và thông tin thực phẩm

***Truy cập và xem chi tiết một hàng cụ thể trong bảng ndata sau khi merge dữ liệu***

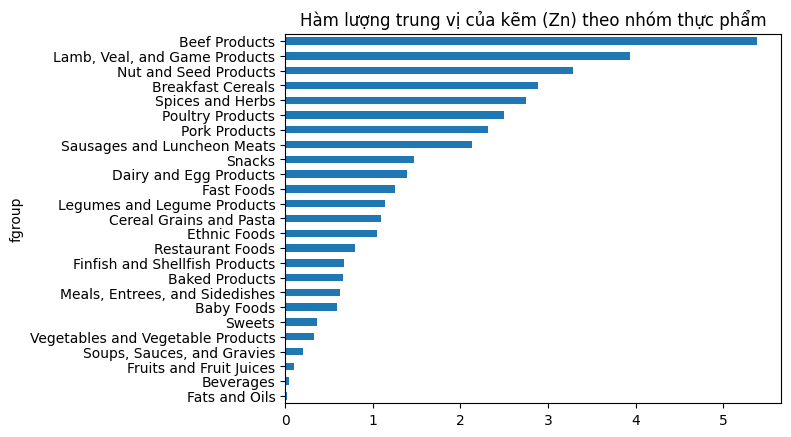


Hình 16: Tổng quan dòng thứ 30,001 sau khi gộp dữ liệu

**Nhận xét:**

Phân tích hàm lượng Glycine trong món súp cà chua đóng hộp, thuộc nhóm thực phẩm "Soups, Sauces, and Gravies".

***So sánh giá trị trung vị (median) của hàm lượng kẽm (Zinc, Zn) trong các nhóm thực phẩm (fgroup).***



Hình 17: Hàm lượng Kẽm (Zn) Trung vị theo Nhóm Thực phẩm

**Nhận xét:**

Beef Products và Lamb/Veal/Game Products thường có hàm lượng kẽm cao nhất.

Cereals, Legumes, và Nuts/Seeds cũng chứa lượng kẽm đáng kể.

Fruits, Vegetables, và Beverages có hàm lượng kẽm thấp.

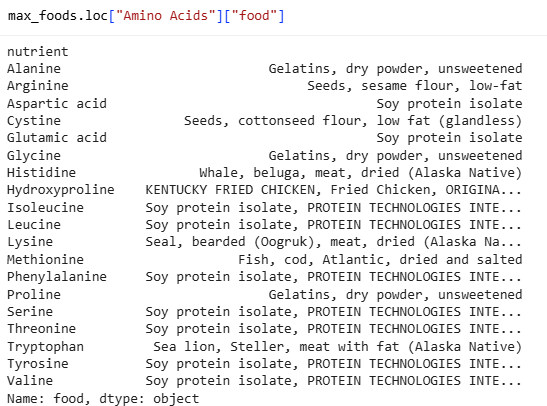
→ Biểu đồ giúp nhận ra những nhóm thực phẩm giàu kẽm nhất, hữu ích cho: Việc lập chế độ dinh dưỡng cân bằng và phân tích nguồn cung cấp vi chất trong thực phẩm.

***Tìm loại thực phẩm có hàm lượng cao nhất cho từng chất dinh dưỡng trong từng nhóm dinh dưỡng:***

Tạo bảng tóm tắt thực phẩm giàu nhất về từng chất dinh dưỡng.

Mỗi dòng thể hiện loại thực phẩm đứng đầu cho một chất dinh dưỡng cụ thể.

*Ví dụ: Đây là danh sách các thực phẩm chứa hàm lượng cao nhất cho mỗi loại axit amin trong nhóm “Amino Acids” của cơ sở dữ liệu USDA.*



Hình 18: liệt kê các loại thực phẩm chứa hàm lượng cao nhất của từng loại Axit Amin

**Nhận xét:**

Thực phẩm từ đậu nành (Soy protein isolate):

Xuất hiện nhiều nhất trong danh sách (Aspartic acid, Glutamic acid, Leucine, Phenylalanine, Serine, Threonine, Tyrosine, Valine...). → Đậu nành và các sản phẩm cô lập protein đậu nành là nguồn cung cấp axit amin hoàn chỉnh, đặc biệt phù hợp cho người ăn chay hoặc luyện tập thể hình.

Gelatin (Gelatins, dry powder, unsweetened):

Nổi bật ở các axit amin như Alanine, Glycine, Proline. → Gelatin giàu collagen, góp phần vào sức khỏe da, tóc, móng và khớp.

Thực phẩm nguồn gốc động vật:

Histidine → Whale (thịt cá voi khô)

Lysine → Seal (thịt hải cẩu khô)

Tryptophan → Sea lion (hải sư) → Đây là các dữ liệu thuộc nhóm thực phẩm bản địa (Alaska Native), cho thấy cơ sở dữ liệu USDA có phạm vi khá đa dạng.

Các loại hạt và cá:

Arginine → Hạt mè (sesame flour)

Cystine → Hạt bông (cottonseed flour)

Methionine → Cá tuyết khô (cod, dried and salted) → Nguồn cung cấp axit amin chứa lưu huỳnh (Methionine, Cystine) thường đến từ hạt và cá biển.