**TRƯỜNG ĐẠI HỌC SÀI GÒN**



KHOA CÔNG NGHỆ THÔNG TIN

**----------**🙡🕮🙣**----------**

BÁO CÁO ĐỒ ÁN CUỐI KỲ

HỌC PHẦN: PHÂN TÍCH DỮ LIỆU

ĐỀ TÀI:

**PHÂN TÍCH DỮ LIỆU US BABYNAMES 1880 - 2010**

**GVHD: LÊ TẤN LONG**

**NHÓM: 4**

**TÊN THÀNH VIÊN:**

**LÊ NGUYỄN QUỐC HƯNG – 3121410240**

**TRẦN BÁ TIÊN – 3122410410**

**BÙI THÀNH CÔNG – 3122410042**

**BÙI HUỲNH GIA – 3121410167**

**Mục lục**

[**PHẦN I: ĐỊNH NGHĨA VẤN ĐỀ** 5](#_Toc212326125)

[**1.1** **Nguồn gốc dữ liệu** 5](#_Toc212326126)

[**1.2**  **Phạm vi dữ liệu** 6](#_Toc212326127)

[**1.3**  **Cấu trúc dữ liệu** 7](#_Toc212326128)

[**1.4** **Mục tiêu, dữ liệu đầu vào (Input), dữ liệu đầu ra (Output)** 7](#_Toc212326129)

[**2. SƠ ĐỒ EDA** 9](#_Toc212326130)

[**2.1**  **Phân tích các xu hướng đặt tên** 9](#_Toc212326131)

[**2.2**  **Đo lường sự tăng đa dạng trong đặt tên** 11](#_Toc212326132)

[**2.3**  **Tổng hợp dữ liệu theo cấp độ nămvà giới tính** 12](#_Toc212326133)

[**2.4**  **SỐ LƯỢNG TÊN PHỔ BIẾN TRONG TOP 50%(1880 - 2010)** 13](#_Toc212326134)

[**2.5**  **CUỘC CÁCH MẠNG CHỮ CÁI CUỐI CÙNG** 14](#_Toc212326135)

[**2.6**  **Tỷ lệ bé trai sinh ra có tên kết thúc bằng d/n/y** 15](#_Toc212326136)

[**2.7**  **Tỷ lệ tên giống Lesly theo giới tính theo thời gian** 16](#_Toc212326137)

[**3. Câu hỏi và ý nghĩa của phân tích dữ liệu US BABYNAMES** 18](#_Toc212326138)

[**3.1 Ý nghĩa của phân tích dữ liệu US BABYNAMES** 18](#_Toc212326139)

[**3.2 Câu hỏi phân tích dữ liệu** 18](#_Toc212326140)

**Phụ lục hình ảnh**

[**Hình 1. Tổng cục An sinh xã hội (SSA)** 4](#_Toc211931600)

[**Hình 2. Cấu trúc dữ liệu US babynames (txt)** 5](#_Toc211931601)

[**Hình 3. Biểu đồ phân tích các xu hướng đặt tên giai đoạn 1880 - 2000** 7](#_Toc211931602)

[**Hình 4. Đo lường sự tăng đa dạng trong đặt tên** 9](#_Toc211931603)

[**Hình 5. Tổng hợp dữ liệu theo cấp độ năm('year') và giới tính('sex')** 10](#_Toc211931604)

[**Hình 6. Số lượng tên phổ biến trong top 50% (50% số lượng người có tên này bao gồm nam('M') và nữ('F') ) (1880 - 2010)** 11](#_Toc211931605)

[**Hình 7. Biểu đồ phân tích cuộc cách mạng “chữ cái cuối cùng”** 12](#_Toc211931606)

[**Hình 8. Biểu đồ mô tả tỷ lệ bé trai sinh ra có tên kết thúc bằng chữ d/n/y theo thời gian** 13](#_Toc211931607)

[**Hình 9. Biểu đồ mô tả tỷ lệ tên giống Lesley theo giới tính nam/nữ theo thời gian** 14](#_Toc211931608)

# **PHẦN I: ĐỊNH NGHĨA VẤN ĐỀ**

## **Nguồn gốc dữ liệu**

**A building with a sign on the front

AI-generated content may be incorrect.**

**Hình 1. Tổng cục An sinh xã hội (SSA)**

Dữ liệu được thu thập và công bố bởi **Cục An sinh Xã hội Hoa Kỳ (Social Security Administration - SSA)**. Nó dựa trên đơn đăng ký thẻ An sinh Xã hội cho các ca sinh tại Hoa Kỳ.

## **1.2 Phạm vi dữ liệu**

Bộ dữ liệu này chứa danh sách tên trẻ em sinh ra tại Hoa Kỳ trong khoảng thời gian **từ năm 1880 đến năm 2010**. (Hiện nay đã có các phiên bản cập nhật đến những năm gần đây hơn).

## **1.3 Cấu trúc dữ liệu**

A computer screen shot of a black screen

AI-generated content may be incorrect.

**Hình 2. Cấu trúc dữ liệu US babynames (txt)**

Bộ dữ liệu gốc thường được cung cấp dưới dạng một tập hợp các tệp văn bản (.txt), **mỗi tệp đại diện cho một năm** (ví dụ: yob1880.txt, yob1881.txt,... cho đến yob2010.txt).

Bên trong mỗi tệp, mỗi dòng dữ liệu có cấu trúc rất đơn giản, bao gồm 3 trường (cột) được phân tách bằng dấu phẩy:

* **Tên (Name)**: Ví dụ: Mary, John.
* **Giới tính (Sex)**: F (cho Nữ - Female) hoặc M (cho Nam - Male).
* **Số lượng (Count)**: Tổng số trẻ sinh ra trong năm đó có tên và giới tính tương ứng.

## **1.4 Mục tiêu, dữ liệu đầu vào (Input), dữ liệu đầu ra (Output)**

|  |  |  |
| --- | --- | --- |
| **Mục tiêu** | **Đầu vào (Input)** | **Đầu ra (Output)** |
| Thống kê mô tả EDA | Dữ liệu cột ‘name’, ‘sex’, ‘births’(được chia thành các năm) | Xu hướng đặt tên  Xu hướng giới tính  Tên phổ biến theo thời gian |
| Phân tích theo giới tính | Dự liệu cột ‘name’,’sex’,’year’ | Dự đoán xu hướng sinh trong từng giai đoạn |
| Phân tích xu hướng thời gian | Dữ liệu cột ’year’,’name’ | Xu hướng đặt tên theo thời gian |
| Tổng hợp dữ liệu | Dữ liệu thô được nhóm theo từng năm(yob1880txt, …) | Hợp nhất toàn bộ dữ liệu vào dataset trong giai đoạn 1880 - 2010 |
| Trực quan hóa dữ liệu | Dữ liệu cột ‘name’, ‘sex’, ‘births’,’year’ (toàn bộ dữ liệu) | Biểu đồ line chart, barchart |

# **2. SƠ ĐỒ EDA**

## **2.1 Phân tích các xu hướng đặt tên**

A graph of different colored lines

AI-generated content may be incorrect.

**Hình 3. Biểu đồ phân tích các xu hướng đặt tên giai đoạn 1880 - 2000**

**Nhận xét về biểu đồ:**

* Xu hướng biến động: Biểu đồ cho thấy sự thay đổi rõ rệt về mức độ phổ biến (số lượng sinh tuyệt đối) của bốn tên ("John", "Harry", "Mary", "Marilyn") trong giai đoạn 1880-2010.
* Ảnh hưởng Baby Boom: Các tên "John" và "Mary" đạt đỉnh cao phổ biến rõ rệt trong thời kỳ Baby Boom (khoảng 1945-1960). "Harry" cũng có một đỉnh nhỏ hơn trong giai đoạn này.
* Ảnh hưởng văn hóa: Tên "Marilyn" gần như không xuất hiện cho đến những năm 1920, sau đó tăng vọt và đạt đỉnh vào những năm 1950 (trùng với thời kỳ nổi tiếng đỉnh cao của Marilyn Monroe), rồi giảm nhanh chóng. Điều này cho thấy ảnh hưởng mạnh mẽ của văn hóa đại chúng lên việc đặt tên.
* Suy giảm gần đây: Cả bốn tên đều cho thấy xu hướng giảm đáng kể về số lượng tuyệt đối vào cuối giai đoạn (từ khoảng 1960 trở đi), phản ánh sự đa dạng hóa ngày càng tăng trong việc lựa chọn tên cho trẻ em.

## **2.2 Đo lường sự tăng đa dạng trong đặt tên**

A graph with lines and numbers

AI-generated content may be incorrect.

**Hình 4. Đo lường sự tăng đa dạng trong đặt tên**

**Nhận xét về biểu đồ:**

* **Xu hướng chính:** Biểu đồ cho thấy rõ ràng rằng sự đa dạng trong việc đặt tên ngày càng tăng theo thời gian cho cả hai giới tính.
* **Quá khứ:** Vào cuối thế kỷ 19 (khoảng 1880), 1000 tên phổ biến nhất chiếm gần như toàn bộ (gần 100%) số lượng trẻ em được sinh ra, cho thấy sự tập trung cao độ vào các tên thông thường.
* **Hiện tại:** Tỷ lệ này giảm dần, đặc biệt là từ giữa thế kỷ 20 trở đi. Đến năm 2010, 1000 tên phổ biến nhất chỉ chiếm khoảng 75% số bé gái và khoảng 85% số bé trai.
* **Khác biệt giới tính:** Xu hướng đa dạng hóa tên gọi mạnh mẽ hơn ở bé gái so với bé trai, thể hiện qua đường biểu đồ ('F') dốc xuống nhanh hơn đường ('M') trong những thập kỷ gần đây.

## **2.3 Tổng hợp dữ liệu theo cấp độ nămvà giới tính**

A graph of birth rate

AI-generated content may be incorrect.

**Hình 5. Tổng hợp dữ liệu theo cấp độ năm('year') và giới tính('sex')**

**Nhận xét về biểu đồ:**

* Xu hướng sinh tăng dần từ 1880 đến 2010.
* Có một đợt bùng nổ sinh (baby boom) rõ rệt đạt đỉnh vào khoảng năm 1960.
* Một quan sát thú vị là đường 'M' (Nam) luôn nằm trên đường 'F' (Nữ), cho thấy số lượng bé trai sinh ra hàng năm luôn cao hơn một chút so với bé gái.

## **2.4 SỐ LƯỢNG TÊN PHỔ BIẾN TRONG TOP 50%(1880 - 2010)**

A graph of a number of babies

AI-generated content may be incorrect.

**Hình 6. Số lượng tên phổ biến trong top 50% (50% số lượng người có tên này bao gồm nam('M') và nữ('F') ) (1880 - 2010)**

**nhận xét về biểu đồ:**

* **Xác nhận xu hướng:** Biểu đồ này xác nhận mạnh mẽ rằng sự đa dạng trong việc đặt tên đang tăng lên đáng kể theo thời gian cho cả hai giới.
* **Ý nghĩa:** Cần ngày càng nhiều tên hơn để chiếm được 50% tổng số trẻ em sinh ra. Điều này có nghĩa là sự phổ biến không còn tập trung vào một số ít tên như trước đây nữa.
* **Quá khứ:** Vào năm 1880, chỉ cần khoảng 14 tên bé trai và 38 tên bé gái là đã chiếm 50% số lượng sinh.
* **Hiện tại:** Đến năm 2010, con số này đã tăng vọt lên khoảng 120 tên bé trai và gần 250 tên bé gái.
* **Khác biệt giới tính rõ rệt:** Sự gia tăng đa dạng tên gọi diễn ra mạnh mẽ hơn nhiều ở bé gái so với bé trai, đặc biệt là từ khoảng năm 1980 trở đi

## **2.5 CUỘC CÁCH MẠNG CHỮ CÁI CUỐI CÙNG**

A screenshot of a graph

AI-generated content may be incorrect.

**Hình 7. Biểu đồ phân tích cuộc cách mạng “chữ cái cuối cùng”**

Nhận xét về biểu đồ:

* Nữ (Biểu đồ dưới): Có sự thay đổi rõ rệt. Tên kết thúc bằng 'e' rất phổ biến vào năm 1910 (khoảng 34%) nhưng đã giảm mạnh. Ngược lại, tên kết thúc bằng 'a' đã tăng vọt, trở nên chiếm ưu thế tuyệt đối vào năm 2010 (gần 40%).
* Nam (Biểu đồ trên): Xu hướng nổi bật nhất là sự gia tăng đột biến của các tên kết thúc bằng 'n' vào năm 2010 (lên đến hơn 35%), vượt xa tất cả các chữ cái khác. Trong khi đó, các chữ cái từng phổ biến như 'd', 'e', 'r', 's' đã giảm tỷ lệ.

## **2.6 Tỷ lệ bé trai sinh ra có tên kết thúc bằng d/n/y**

A graph of different colored lines

AI-generated content may be incorrect.

**Hình 8. Biểu đồ mô tả tỷ lệ bé trai sinh ra có tên kết thúc bằng chữ d/n/y theo thời gian**

**nhận xét ngắn gọn về biểu đồ:**

* **Xu hướng chính:** Sự thay đổi lớn nhất là sự gia tăng mạnh mẽ của các tên kết thúc bằng chữ 'n' (màu cam), bắt đầu từ khoảng những năm 1940 và tiếp tục tăng vọt, trở thành phổ biến nhất trong ba loại này vào cuối biểu đồ.
* **Chữ 'd' (màu xanh dương):** Từng phổ biến hơn 'y', đạt đỉnh vào khoảng năm 1940, nhưng sau đó đã giảm sút liên tục và trở nên ít phổ biến nhất.
* **Chữ 'y' (màu xanh lá):** Có một giai đoạn phổ biến vào giữa thế kỷ 20 (đạt đỉnh quanh năm 1960), nhưng sau đó cũng có xu hướng giảm dần.

## **2.7 Tỷ lệ tên giống Lesly theo giới tính theo thời gian**

A graph showing the difference between male and female

AI-generated content may be incorrect.

**Hình 9. Biểu đồ mô tả tỷ lệ tên giống Lesley theo giới tính nam/nữ theo thời gian**

**Nhận xét về biểu đồ:**

* **Chuyển đổi giới tính rõ rệt: Biểu đồ minh họa một sự thay đổi ngoạn mục trong cách các tên chứa "Lesl" (như Leslie, Lesley) được sử dụng cho nam và nữ theo thời gian.**
* **Quá khứ: Ban đầu (cuối thế kỷ 19, đầu thế kỷ 20), các tên này chủ yếu được đặt cho bé trai (đường liền 'M' ở mức cao, gần 90%).**
* **Giai đoạn chuyển tiếp: Khoảng những năm 1940, có một sự giao thoa nhanh chóng, tỷ lệ đặt tên cho bé gái (đường đứt 'F') tăng vọt trong khi tỷ lệ cho bé trai giảm mạnh.**
* **Hiện tại: Từ giữa thế kỷ 20 trở đi, các tên này trở nên gần như độc quyền cho bé gái, với tỷ lệ 'F' tiến gần đến 100% vào cuối giai đoạn.**
* **Kết luận: Đây là một ví dụ điển hình về một cái tên ban đầu phổ biến cho nam giới đã dần dần trở thành một cái tên chủ yếu dành cho nữ giới.**

# **3. Câu hỏi và ý nghĩa của phân tích dữ liệu US BABYNAMES**

## **3.1 Ý nghĩa của phân tích dữ liệu US BABYNAMES**

Khám phá xu hướng đặt tên trẻ em ở Mỹ trong hơn 130 năm (1880–2010):

**2.1 Phân tích các xu hướng đặt tên**

Tìm hiểu sự phổ biến của từng tên theo thời gian:

**2.2 Đo lường sự tăng đa dạng trong đặt tên**

Đánh giá sự đa dạng văn hóa và thay đổi giới tính trong đặt tên:

**2.2 Đo lường sự tăng đa dạng trong đặt tên**

**2.4 SỐ LƯỢNG TÊN PHỔ BIẾN TRONG TOP 50%(1880 - 2010)**

Xác định các mô hình dài hạn (long-term patterns) về độ phổ biến, vòng đời của tên, và tác động xã hội:

**2.6 Tỷ lệ bé trai sinh ra có tên kết thúc bằng d/n/y**

**2.7 Tỷ lệ tên giống Lesly theo giới tính theo thời gian**

## **3.2 Câu hỏi phân tích dữ liệu**

1. Phân tích Mô tả

Câu hỏi 1: Số lượng trẻ sinh theo giới tính thay đổi như thế nào từ 1880–2010?

A graph with red and blue lines

AI-generated content may be incorrect.

Biểu đồ cho thấy tổng số trẻ em sinh ra ở Mỹ (1880–2010) đã **tăng đáng kể** qua 130 năm.

1. **So sánh Giới tính:** Số lượng bé trai (M, đường màu xanh) **luôn cao hơn** số lượng bé gái (F, đường màu đỏ) trong suốt toàn bộ thời kỳ.
2. **Xu hướng chính:** Cả hai giới đều có chung xu hướng tăng trưởng, đặc biệt là giai đoạn tăng trưởng chậm (1880-1910) và tăng nhanh (1910-1940).
3. **Sự kiện nổi bật:** Có một sự **bùng nổ dân số (Baby Boom)** rất rõ rệt sau năm 1945, đạt đỉnh vào khoảng năm 1960, sau đó sụt giảm trước khi tăng trở lại và dao động ở mức cao.

Câu 2: Tên nào phổ biến nhất mỗi thập kỷ (top 10 theo năm hoặc giới)?

A screenshot of a graph

AI-generated content may be incorrect.

Biểu đồ so sánh Top 5 tên nam phổ biến giữa thập kỷ 1880s và 2000s, cho thấy sự **thay đổi hoàn toàn** trong xu hướng đặt tên.

1. **Thay đổi tên:** **Không có tên nào** trong Top 5 của 1880s (John, William, James, George, Charles) còn xuất hiện trong Top 5 của 2000s (Jacob, Michael, Joshua, Matthew, Daniel).
2. **Quy mô (Số lượng):** Số lượng sinh của các tên hàng đầu trong thập kỷ 2000s (trên 200.000) **cao hơn đáng kể** so với thập kỷ 1880s (dưới 90.000). Điều này phản ánh sự gia tăng dân số.
3. **Độ phổ biến:** Vào những năm 1880, tên "John" phổ biến vượt trội so với các tên còn lại. Trong khi đó, vào những năm 2000, sự chênh lệch giữa các tên trong Top 5 (như "Jacob" và "Michael") không quá lớn, cho thấy xu hướng đặt tên đa dạng hơn.

Câu 3: Độ tập trung tên có thay đổi theo thời gian không (tên phổ biến chiếm bao nhiêu % tổng số sinh)?

A graph showing the value of a stock market

AI-generated content may be incorrect.

Biểu đồ cho thấy một xu hướng xã hội rõ rệt: **tên gọi ngày càng trở nên đa dạng hơn** từ năm 1880 đến 2010.

1. **Xu hướng chính:** Cả hai đường (Nam - M và Nữ - F) đều cho thấy một **sự sụt giảm nghiêm trọng** theo thời gian. Vào năm 1880, Top 50 tên phổ biến nhất chiếm tới 70% tổng số bé trai và gần 60% bé gái. Đến năm 2010, con số này giảm mạnh, chỉ còn khoảng 32% (Nam) và 22% (Nữ).
2. **So sánh Giới tính:** Đường màu xanh (Nam) luôn nằm trên đường màu đỏ (Nữ). Điều này có nghĩa là tên của nam giới **luôn có độ tập trung cao hơn** (ít đa dạng hơn) so với tên của nữ giới trong suốt 130 năm.

Câu 4: Xu hướng đa dạng hóa tên (name diversity) có tăng theo thời gian không?

A graph with a line

AI-generated content may be incorrect.

Biểu đồ "Số lượng tên cần để chiếm 50% số trẻ theo năm":

Biểu đồ này minh họa cho sự đa dạng hóa trong việc đặt tên. Số lượng tên riêng biệt cần thiết để chiếm một nửa (50%) tổng số trẻ em đã tăng lên rõ rệt, đặc biệt là từ khoảng những năm 1980. Điều này có nghĩa là ngày càng có nhiều tên độc đáo được sử dụng hơn, thay vì chỉ tập trung vào một số ít tên phổ biến như trước đây.

Câu 5: Xu hướng tên trung tính (Unisex)

A graph with a line

AI-generated content may be incorrect.

Biểu đồ "Xu hướng tên Unisex theo năm":

 Biểu đồ này cho thấy số lượng hoặc sự phổ biến của các tên unisex (tên có thể dùng cho cả nam và nữ) đã tăng lên đáng kể theo thời gian. Sau một giai đoạn tương đối ổn định, xu hướng này bắt đầu tăng mạnh từ khoảng những năm 1970 và tiếp tục tăng vọt cho đến cuối giai đoạn được hiển thị.

Câu 6: Sự phổ biến của tên có khác nhau giữa bé trai và bé gái không (độ tập trung tên nữ có cao hơn)

A graph with numbers and lines

AI-generated content may be incorrect.

Đây là nhận xét về biểu đồ:

Biểu đồ cho thấy xu hướng đặt tên cho cả nam (M) và nữ (F) đều ngày càng trở nên đa dạng hơn theo thời gian, thể hiện qua việc cần nhiều tên hơn để đạt 50% tổng số trẻ sinh.

Tuy nhiên, xu hướng đặt tên cho nữ (F) luôn luôn đa dạng hơn nam (M). Đường biểu diễn của nữ (màu xanh) luôn nằm trên đường của nam (màu cam), nghĩa là luôn cần nhiều tên nữ khác nhau hơn để đạt mốc 50% so với tên nam trong suốt giai đoạn lịch sử được ghi lại.

Câu 7: Những tên xuất hiện ngắn hạn (short-lived) khác gì so với tên phổ biến lâu dài (long-lived)?

A graph with blue and orange lines

AI-generated content may be incorrect.

nhận xét về biểu đồ:

Biểu đồ này minh họa rõ rệt sự khác biệt giữa một tên "lâu dài" và một tên "ngắn hạn":

Mary (Lâu dài - màu xanh): Tên này duy trì sự phổ biến ở mức độ cao trong một khoảng thời gian rất dài (trải dài hơn một thế kỷ), với các xu hướng tăng và giảm diễn ra từ từ.

Debra (Ngắn hạn - màu cam): Tên này là một ví dụ điển hình của "mốt" nhất thời. Nó bùng nổ rất nhanh, đạt đỉnh cao đột ngột vào khoảng những năm 1950, và sau đó cũng suy giảm nhanh chóng, gần như biến mất.

Câu 8: Có thể dự đoán tên nào sẽ biến mất hoặc quay trở lại phổ biến dựa trên xu hướng giảm/tăng trong 20 năm gần nhất?

A graph with a line

AI-generated content may be incorrect.

Biểu đồ này cho thấy tên 'Jennifer' là một ví dụ điển hình của một tên "mốt" ngắn hạn (short-lived fad).

Bùng nổ đột ngột: Tên này gần như không tồn tại trước những năm 1960.

Đạt đỉnh cao: Nó bùng nổ mạnh mẽ, đạt đỉnh cao chót vót về sự phổ biến vào khoảng năm 1980.

Suy giảm nhanh chóng: Ngay sau khi đạt đỉnh, tên 'Jennifer' đã suy giảm với tốc độ "lao dốc" (như được tô màu trong vùng màu đỏ). Trong vòng 20 năm, nó đã giảm từ đỉnh cao nhất xuống gần như bằng không, cho thấy nó đã "lỗi mốt" rất nhanh.

Câu 9: Chu kỳ tên "Vintage" (Ví dụ: Evelyn)

A green line graph with numbers

AI-generated content may be incorrect.

Nhận xét biểu đồ:

Tên này rất phổ biến vào đầu những năm 1900.

Sau đó, nó giảm dần độ phổ biến và gần như "biến mất" vào khoảng những năm 1980.

Trong 20 năm gần nhất (vùng được tô màu xanh lá), tên 'Evelyn' đã bắt đầu tăng trở lại, cho thấy nó đang quay trở lại thành "mốt" sau một chu kỳ dài.