1. **Phân tích dataset Bitly (timezone)**
2. **Giới thiệu**

Năm 2011, dịch vụ rút gọn URL Bitly đã hợp tác với cổng thông tin 1.USA.gov để cung cấp luồng dữ liệu ẩn danh liên quan đến các đường dẫn kết thúc bằng miền .gov hoặc .mil. Các dữ liệu này được ghi nhận theo từng yêu cầu rút gọn URL, chứa nhiều thông tin về người dùng như múi giờ, hệ điều hành hay trình duyệt sử dụng. Bộ dữ liệu phản ánh hành vi truy cập vào các trang web chính phủ Hoa Kỳ trong bối cảnh mạng xã hội và nền tảng chia sẻ đường dẫn đang phát triển mạnh mẽ.

Việc phân tích tập dữ liệu này giúp nhận diện đặc điểm phân bố người dùng theo khu vực địa lý và hệ điều hành, qua đó cung cấp góc nhìn về mô hình tiêu thụ thông tin điện tử trong giai đoạn đầu của thập niên 2010. Bên cạnh giá trị thực tiễn trong nghiên cứu hành vi người dùng, đây còn là ví dụ điển hình minh họa cho việc khai thác dữ liệu dạng JSON và áp dụng các công cụ phân tích trong Python.

1. **Dữ liệu & Phương pháp**

**2.1 Tiền xử lý:**

Dữ liệu thô được cung cấp dưới dạng các bản ghi JSON. Trước hết, các bản ghi này được chuyển đổi sang cấu trúc bảng (DataFrame) để thuận tiện cho phân tích. Một số biến quan trọng cần xử lý bao gồm:

* Múi giờ (tz): Các giá trị khuyết (NaN) được thay thế bằng nhãn “Missing”, trong khi các chuỗi rỗng được gán thành “Unknown”. Việc này nhằm duy trì toàn bộ dữ liệu gốc, đồng thời vẫn phản ánh rõ sự tồn tại của thông tin bị thiếu.
* Chuỗi user-agent (a): Những bản ghi thiếu giá trị trong trường này bị loại bỏ. Từ các chuỗi còn lại, một biến mới phản ánh hệ điều hành được xây dựng dựa trên quy tắc đơn giản: nếu chuỗi có chứa từ khóa “Windows” thì gán nhãn “Windows”, ngược lại gán nhãn “Not Windows”.
* Xử lý dữ liệu thiếu: Đáng chú ý, tỷ lệ bản ghi có trường múi giờ rỗng khá lớn; thay vì loại bỏ, chúng được giữ lại và gán nhãn “Unknown” để phân tích như một nhóm riêng biệt.

**2.2 Phương pháp thống kê / trực quan hóa:**

Sau khi tiền xử lý, dữ liệu được khai thác theo các bước sau:

* Thống kê tần suất múi giờ: Tính số lần xuất hiện của từng múi giờ và lựa chọn 10 múi giờ phổ biến nhất để phân tích chi tiết.
* So sánh hệ điều hành: Tạo bảng hai chiều giữa múi giờ và hệ điều hành (Windows/Not-Windows), qua đó cho phép so sánh số lượng người dùng theo từng nhóm.
* Phân tích tập trung (top 10 múi giờ): Lựa chọn 10 múi giờ có tổng số bản ghi lớn nhất để trực quan hóa phân bố người dùng theo hệ điều hành, dưới dạng biểu đồ cột nhóm (grouped bar chart).
* Chuẩn hóa tỷ lệ: Để loại bỏ ảnh hưởng của kích thước mẫu giữa các múi giờ, dữ liệu được chuẩn hóa về tỷ lệ phần trăm trong từng múi giờ, từ đó vẽ biểu đồ cột chuẩn hóa (normalized bar chart).
* Phân tích nâng cao (tùy chọn): Để kiểm định sự khác biệt có ý nghĩa thống kê giữa múi giờ và hệ điều hành, có thể áp dụng kiểm định Chi-square trên bảng tần số (contingency table) cho những múi giờ có kích thước mẫu đủ lớn.

1. **Kết quả**

**3.1. Top 10 múi giờ**

A graph with blue bars

AI-generated content may be incorrect.

Hình 3.1. Top 10 múi giờ phổ biến nhất trong dữ liệu Bitly

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

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

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

**3.2. So sánh Windows vs Not-Windows — absolute counts (top time zones)**

A graph of a bar graph

AI-generated content may be incorrect.

Hình 3.2. Phân bố hệ điều hành Windows và Not-Windows theo số lượng tuyệt đối tại 10 múi giờ phổ biến nhất

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

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

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

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

**3.3. So sánh Windows vs Not-Windows — tỉ lệ phần trăm (trong mỗi tz)**

A graph of a bar chart

AI-generated content may be incorrect.

Hình 3.3. So sánh hệ điều hành Windows và Not-Windows theo tỉ lệ phần trăm trong 10 múi giờ phổ biến nhất

Biểu đồ Hình 3 trình bày tỉ lệ phần trăm giữa Windows và Not-Windows trong từng múi giờ, cho phép quan sát sự phân bố hệ điều hành một cách độc lập với kích thước mẫu. Kết quả cho thấy Windows vẫn chiếm ưu thế tại phần lớn múi giờ, đặc biệt ở America/New\_York, America/Chicago và America/Los\_Angeles, nơi Windows chiếm trên 65% số quan sát. Điều này củng cố nhận định rằng ở các trung tâm lớn của Hoa Kỳ, thiết bị Windows vẫn là nền tảng phổ biến.

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

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

1. **Nhận xét và thảo luận:**

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

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

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

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

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

1. **Phân tích dataset MoviesLens**
2. **Giới thiệu:**

Bộ dữ liệu MovieLens 1M do GroupLens Research phát hành chứa hơn 1 triệu lượt đánh giá phim từ khoảng 6.000 người dùng trên 4.000 bộ phim, được thu thập cuối thập niên 1990 – đầu 2000. Dữ liệu bao gồm ba bảng: thông tin người dùng (giới tính, tuổi, nghề nghiệp, mã vùng), thông tin phim (tựa đề, thể loại, năm phát hành) và dữ liệu đánh giá (rating). Bộ dữ liệu này thường được sử dụng để phát triển hệ thống gợi ý phim dựa trên machine learning. Trong phân tích này, chúng tôi tập trung mô tả và so sánh hành vi đánh giá theo giới tính, độ tuổi và thể loại, thay vì xây dựng mô hình dự đoán.

1. **Phương pháp:**

Bộ dữ liệu MovieLens 1M gồm ba bảng người dùng, phim và đánh giá, được hợp nhất bằng *pandas.merge* thành một bảng duy nhất. Chúng tôi tính điểm trung bình của từng phim theo giới tính bằng pivot\_table, đồng thời chỉ giữ các phim có ≥250 lượt đánh giá để đảm bảo tính đại diện. Sự khác biệt thị hiếu được đo bằng chênh lệch điểm trung bình giữa nam và nữ. Mức độ gây tranh cãi được xác định qua độ lệch chuẩn của điểm số. Ngoài ra, dữ liệu thể loại được tách (explode) để phân tích xu hướng đánh giá theo cặp thể loại – nhóm tuổi.

1. **Kết quả:**
   1. **Phân tích đánh giá trung bình theo giới tính**

A graph with red and white lines

AI-generated content may be incorrect.

Hình 3.1.1. Top 10 phim được nữ giới đánh giá cao nhất

A blue and white bar graph

AI-generated content may be incorrect.

Hình 3.1.2. Top 10 phim được nam giới đánh giá cao nhất

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

* 1. **Phân tích sự khác biệt giới tính trong điểm đánh giá**

A graph with a bar chart

AI-generated content may be incorrect.

**Hình 3.2.** Biểu đồ thể hiện 20 bộ phim có sự khác biệt điểm trung bình lớn nhất giữa nam và nữ.

Hình 3.2 minh họa 20 bộ phim có sự chênh lệch điểm trung bình lớn nhất giữa hai giới tính. Trục hoành thể hiện giá trị Difference (Male – Female), trong đó các cột màu đỏ biểu thị phim được nữ giới đánh giá cao hơn, còn các cột màu xanh thể hiện phim được nam giới đánh giá cao hơn.

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

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

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

* 1. **Độ biến thiên của điểm số và mức độ đồng thuận giữa khán giả**

A graph with different colored bars

AI-generated content may be incorrect.

**Hình 3.3.** Top 10 bộ phim có độ lệch chuẩn điểm đánh giá cao nhất, thể hiện mức độ tranh cãi lớn trong cộng đồng người xem.

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

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

* 1. **Xu hướng đánh giá theo thể loại và nhóm tuổi**

A blue and green chart

AI-generated content may be incorrect.

**Hình 3.4.** Biểu đồ nhiệt thể hiện điểm đánh giá trung bình của các thể loại phim theo nhóm tuổi.

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

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

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

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

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

1. **Nhận xét và thảo luận:**

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

**Thứ nhất, giới tính** là yếu tố ảnh hưởng mạnh mẽ đến xu hướng đánh giá.  
Nữ giới có xu hướng ưu ái các phim mang tính nhân văn, cảm xúc và hoạt hình, thể hiện qua việc đánh giá cao hơn các tác phẩm như A Close Shave (1995) hay Schindler’s List (1993). Ngược lại, nam giới lại dành điểm cao hơn cho các bộ phim hành động, sử thi hoặc tội phạm như The Godfather (1972) và Seven Samurai (1954).  
Sự khác biệt này phản ánh đặc điểm tâm lý và sở thích tiếp nhận của hai giới: trong khi nữ giới tập trung vào yếu tố cảm xúc và thông điệp nhân văn, nam giới bị thu hút bởi kỹ năng, hành động và cấu trúc kịch tính. Tuy nhiên, vẫn tồn tại những vùng giao thoa — điển hình là The Shawshank Redemption (1994) — thể hiện giá trị phổ quát vượt lên trên ranh giới giới tính.

**Thứ hai, độ biến thiên điểm số** cung cấp cái nhìn sâu hơn về mức độ đồng thuận trong cộng đồng người xem.  
Các phim có độ lệch chuẩn cao như Dumb & Dumber (1994), The Blair Witch Project (1999) hay Natural Born Killers (1994) cho thấy sự chia rẽ mạnh mẽ trong cảm nhận khán giả. Phần lớn các tác phẩm này thuộc thể loại hài, kinh dị hoặc nghệ thuật thử nghiệm – nơi ranh giới giữa “hay” và “dở” mang tính chủ quan cao.  
Ngược lại, các phim như The Shawshank Redemption hoặc Toy Story có độ lệch chuẩn thấp, thể hiện sự đồng thuận cao về chất lượng. Điều này khẳng định rằng độ lệch chuẩn không chỉ là chỉ số thống kê, mà còn là thước đo xã hội học phản ánh sự đa dạng trong diễn giải nghệ thuật và kỳ vọng thể loại.

**Thứ ba, độ tuổi** cũng cho thấy tác động đáng kể đến xu hướng đánh giá.  
Người trẻ (<25 tuổi) có xu hướng đánh giá cao các thể loại mang tính giải trí và thị giác như Animation và Musical, trong khi nhóm trung niên và lớn tuổi (>35 tuổi) ưa chuộng Drama, Documentary và Film-Noir. Đặc biệt, thể loại Film-Noir duy trì điểm trung bình cao ổn định (≈4.1–4.2) ở hầu hết các nhóm tuổi, chứng tỏ sức hút bền vững của dòng phim cổ điển.  
Điểm trung bình có xu hướng tăng dần theo độ tuổi, cho thấy khán giả lớn tuổi thường đánh giá hào phóng hơn, có thể do trải nghiệm sống và sự đồng cảm sâu sắc hơn với các chủ đề nhân sinh.

Tổng hợp lại, kết quả nghiên cứu khẳng định rằng **các yếu tố nhân khẩu học có vai trò quyết định trong hành vi đánh giá phim**. Giới tính, độ tuổi và sở thích thể loại không chỉ ảnh hưởng đến điểm số trung bình mà còn đến mức độ đồng thuận và hướng cảm xúc trong đánh giá.  
Những phát hiện này mang giá trị ứng dụng thực tiễn trong việc **xây dựng hệ thống gợi ý phim cá nhân hóa (personalized recommendation system)** — nơi mô hình có thể tích hợp thông tin nhân khẩu học để đề xuất phim phù hợp hơn với từng nhóm người dùng. Ngoài ra, chúng cũng góp phần cung cấp bằng chứng định lượng cho các nghiên cứu xã hội học và tâm lý học về hành vi tiếp nhận văn hóa trong môi trường số.