# TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN





BÁO CÁO TIỂU LUẬN PHÂN TÍCH DỮ LIỆU

**PHÂN TÍCH KHÁM PHÁ DỮ LIỆU VỀ MOVIELENS 1M**

**Giảng viên hướng dẫn: Đỗ Như Tài**

**Thông tin thành viên:**

Lê Nguyễn Quốc Hưng 3121410240

Trần Bá Tiên 3122410410

Dương Huỳnh Gia 3121410167

Bùi Thành Công 3122410042

***TP. Hồ Chí Minh, tháng 9/2025***

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..……………………………………….……………………..…………………………………........................................................................................................................

# LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Đại Học Sài Gòn, đặc biệt là các thầy cô ở khoa Công Nghệ Thông Tin đã tạo điều kiện cho em tiếp cận và tìm hiểu để hoàn thành đồ án môn học lần này.

Và chúng em cũng xin chân thành đặc biệt cảm ơn thầy Đỗ Như Tài, thầy là giảng viên giảng dạy đã nhiệt tình hướng dẫn giúp chúng em hoàn thành được đồ án lần này.

Trong quá trình nghiên cứu và làm bài báo cáo đồ án, kiến thức cũng như kinh nghiệm thực tế còn nhiều hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót, chúng em rất mong nhận được ý kiến đóng góp của thầy, cô để em học hỏi được nhiều kỹ năng, kinh nghiệm và sẽ hoàn thành tốt hơn cho những bài báo cáo sắp tới.

Chúng em xin chân thành cảm ơn!

# PHÂN CÔNG CÔNG VIỆC

|  |  |
| --- | --- |
| Dương Huỳnh Gia | Word |

# MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc30318)

[PHÂN CÔNG CÔNG VIỆC 2](#_Toc22899)

[MỤC LỤC 3](#_Toc1019)

[DANH MỤC HÌNH ẢNH 4](#_Toc12969)

[I. GIỚI THIỆU 6](#_Toc21053)

[1. Nguồn gốc 6](#_Toc16087)

[2. Phạm vi dữ liệu 6](#_Toc27766)

[3. Cấu trúc dữ liệu 6](#_Toc18692)

[II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM) 7](#_Toc1031)

[1. Thông tin người dùng (users.dat) 7](#_Toc7719)

[2. Thông tin phim (movie.dat) 8](#_Toc30401)

[3. Thông tin đánh giá (rating.dat) 8](#_Toc7859)

[III. CHUẨN BỊ VẤN ĐỀ (PREPARE PROBLEM) 9](#_Toc23942)

[IV. PHÂN TÍCH 9](#_Toc22850)

[1. EDA không dùng biểu đồ (Non-graphie EDA) 9](#_Toc26851)

[1.1 Xử lý giá trị bị thiếu 9](#_Toc28468)

[2. EDA trực quan bằng sơ đồ (Graphical EDA) 10](#_Toc17299)

[2.1 Biểu đồ Phân bổ giới tính 10](#_Toc15283)

[2.2 Biểu đồ Phân bổ độ tuổi 11](#_Toc12813)

[2.3 Biểu đồ phân bố nghề nghiệp theo tên 12](#_Toc14209)

[2.4 Phân bố Điểm đánh giá (Rating) 13](#_Toc8073)

[2.5 Các thể loại phim phổ biến nhất 14](#_Toc10356)

[2.6 Các thể loại phim có >200 lượt đánh giá 15](#_Toc3104)

[3. Độ lệch chuẩn 16](#_Toc24970)

[3.1 Độ lệch chuẩn Rating 16](#_Toc11503)

[3.2 Độ lệch chuẩn trung bình theo thể loại phim 17](#_Toc22126)

[3.3 Độ lệch chuẩn của rating theo nhóm tuổi 18](#_Toc19031)

[3.4. Biểu đồ trực quan hóa ma trận tương quan (Correlogram) 19](#_Toc3525)

[3.5 So sánh độ lệch chuẩn giữa các thể loại 20](#_Toc20869)

[3.6 Các yếu tố ảnh hưởng đến độ lệch chuẩn 21](#_Toc1450)

[3.7 Độ lệch chuẩn theo thập niên 22](#_Toc15060)

[4. Mối tương quan 23](#_Toc18146)

[4.1 Mối tương quan giữa độ phổ biến và độ phân hóa 23](#_Toc15637)

[4.2 Các loại phim tranh cãi nhất giữa nam và nữ 24](#_Toc27826)

[4.3 Tỷ lệ đánh giá cực đoan theo nhóm tuổi 25](#_Toc14135)

[III. KẾT LUẬN 26](#_Toc28987)

[TÀI LIỆU THAM KHẢO 27](#_Toc21957)

# DANH MỤC HÌNH ẢNH

[Hình 1 Bộ dữ liệu 8](#_Toc30166)

[Hình 2 : Xử lý giá trị bị thiếu 8](#_Toc27120)

[Hình 3 : Hình phân bố giới tính 9](#_Toc20051)

[Hình 4 : Biểu đồ phân bố độ tuổi 10](#_Toc3988)

[Hình 5 : Biểu đồ phân bố nghề nghiệp theo tên 11](#_Toc17701)

[Hình 6 : Biểu đồ phân bố điểm đánh giá 12](#_Toc23165)

[Hình 7 : Biểu đồ thể hiện các thể loại phim phổ biến nhất 13](#_Toc4687)

[Hình 8 : Biểu đồ thể hiện các phim có trên 200 lượt đánh giá 14](#_Toc15483)

[Hình 9 : Biểu đồ thể hiện độ lệch chuẩn rating 15](#_Toc31913)

[Hình 10 : Biểu đồ độ lệch chuẩn trung bình theo thể loại phim 16](#_Toc11955)

[Hình 11 : Biểu đồ độ lệch chuẩn rating theo nhóm tuổi 17](#_Toc7998)

[Hình 12 : Biểu đồ trực quan hóa ma trận tương quan 18](#_Toc16943)

[Hình 13 : Biểu đồ so sánh độ lệch chuẩn giữa các thể loại 19](#_Toc7666)

[Hình 14 : Biểu đồ thể hiện các yếu tố ảnh hưởng độ lệch chuẩn 20](#_Toc6554)

[Hình 15 : Biểu đồ thể hiện độ lệch chuẩn theo thập niên 21](#_Toc3432)

[Hình 16 : Mối tương quan giữa độ phổ biến và độ phân hóa 22](#_Toc12356)

[Hình 17 : Các loại phim tranh cãi nhất giữa nam và nữ 23](#_Toc4642)

[Hình 18 : Tý lệ đánh giá cực đoan theo nhóm tuổi 24](#_Toc4717)

# I. GIỚI THIỆU

## 1. Nguồn gốc

Dữ liệu được thu thập và công bố bởi phòng thí nghiệm nghiên cứu GroupLens tại Đại học Minnesota (University of Minnesota). Dữ liệu này được thu thập thông qua trang web MovieLens, một hệ thống đề xuất phim.

## Phạm vi dữ liệu

Bộ dữ liệu này (phiên bản MovieLens 1M) chứa 1.000.209 lượt xếp hạng ẩn danh từ 6.040 người dùng cho khoảng 3.900 bộ phim. Ngoài ra, bộ dữ liệu còn bao gồm:

- Thông tin nhân khẩu học của người dùng (Giới tính, Tuổi, Nghề nghiệp, Mã Zip).

- Thông tin chi tiết của phim (Tiêu đề, Năm sản xuất, và Thể loại).

## 3. Cấu trúc dữ liệu

Bộ dữ liệu gốc được cung cấp dưới dạng ba tệp văn bản (.dat), bao gồm: users.dat, ratings.dat, và movies.dat.

Bên trong mỗi tệp, mỗi dòng dữ liệu có cấu trúc được phân tách bằng dấu hai chấm kép (::)

- users.dat: Mỗi dòng chứa thông tin về một người dùng.

Cấu trúc: UserID::Gender::Age::Occupation::Zip-code

- ratings.dat: Mỗi dòng chứa thông tin về một lượt xếp hạng của người dùng cho phim.

Cấu trúc: UserID::MovieID::Rating::Timestamp

- movies.dat: Mỗi dòng chứa thông tin về một bộ phim.

Cấu trúc: MovieID::Title::Genres

Lưu ý: Trong trường Thể loại (Genres), các thể loại khác nhau được phân tách bằng dấu gạch đứng (|).

# II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM)

Bạn đã bao giờ tự hỏi **những yếu tố nào ảnh hưởng đến cách khán giả đánh giá phim**? Liệu **thể loại, năm phát hành, giới tính, độ tuổi** hay **mức độ phổ biến của phim** đóng vai trò quan trọng nhất trong việc **tạo ra sự khác biệt trong điểm đánh giá**? Hay các **phim kinh dị, tâm lý, chính trị** có xu hướng **gây tranh cãi hơn** so với các **phim hài, tình cảm**?  
Bộ dữ liệu **MovieLens 1M** có thể giúp chúng ta khám phá những câu hỏi đó!

Bộ dữ liệu này chứa **hơn 1 triệu lượt đánh giá phim** do người dùng thực hiện, bao gồm thông tin chi tiết về **người xem (tuổi, giới tính, nghề nghiệp)**, **phim (thể loại, năm phát hành)** và **điểm đánh giá (rating)**. Tất cả dữ liệu cá nhân đã được ẩn danh để đảm bảo quyền riêng tư.

Hãy cùng phân tích và khám phá bộ dữ liệu này để tìm ra **những yếu tố quan trọng ảnh hưởng đến sự phân tán trong đánh giá phim**, từ đó hiểu rõ hơn về **hành vi và sở thích khán giả**, cũng như **hỗ trợ mô hình dự đoán xu hướng đánh giá phim trong tương lai**.

- Mô tả:

Bộ dữ liệu bao gồm thông tin về **người dùng, phim và điểm đánh giá** được thu thập từ hệ thống MovieLens.

Các đặc trưng mô tả **đặc điểm của người xem** và **đặc tính của phim** giúp phục vụ cho việc **phân tích hành vi đánh giá và mức độ đồng thuận** của khán giả.

- Dữ liệu vào:

1. Thông tin người dùng (users.dat)

**UserID**: mã định danh người dùng

**Gender**: giới tính (Nam/Nữ)

**Age**: độ tuổi của người xem

**Occupation**: nghề nghiệp của người xem

2. Thông tin phim (movie.dat)

**MovieID**: mã định danh phim

**Title**: tên phim

**Genres**: thể loại phim (có thể gồm nhiều thể loại)

**Year**: năm phát hành phim (trích xuất từ tên phim)

3. Thông tin đánh giá (rating.dat)

**UserID**: mã người dùng

**MovieID**: mã phim

**Rating**: điểm đánh giá (từ 1 đến 5)

**Timestamp**: thời điểm đánh giá

- Từ bộ dữ liệu MovieLens 1M, chúng ta sẽ:

Hiểu rõ hơn **hành vi đánh giá và sở thích của khán giả**.

Xác định **những yếu tố quan trọng ảnh hưởng đến sự tranh cãi trong đánh giá phim**.

Làm nền tảng cho **mô hình dự đoán điểm đánh giá hoặc đề xuất phim cá nhân hóa** trong tương lai.

# III. CHUẨN BỊ VẤN ĐỀ (PREPARE PROBLEM)

* Bộ dữ liệu được lấy từ Kaggle .

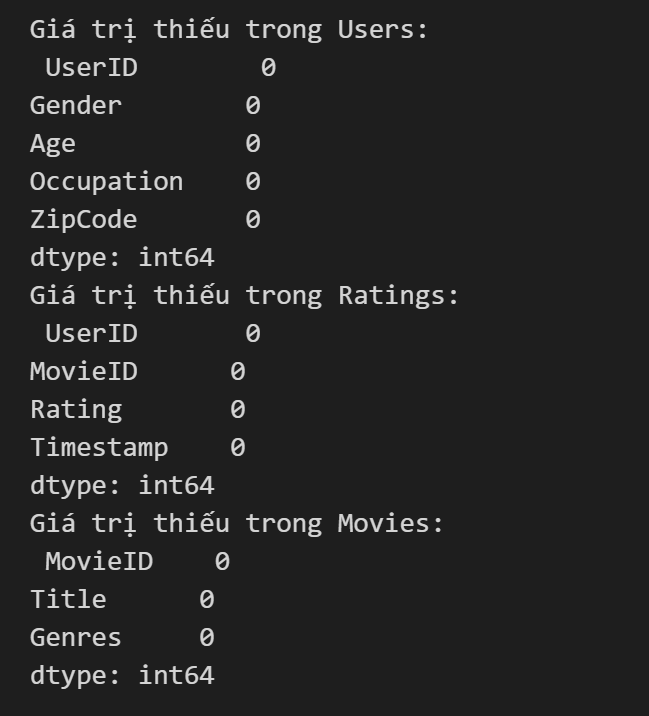


Hình 1 Bộ dữ liệu

# IV. PHÂN TÍCH

## 1. EDA không dùng biểu đồ (Non-graphie EDA)

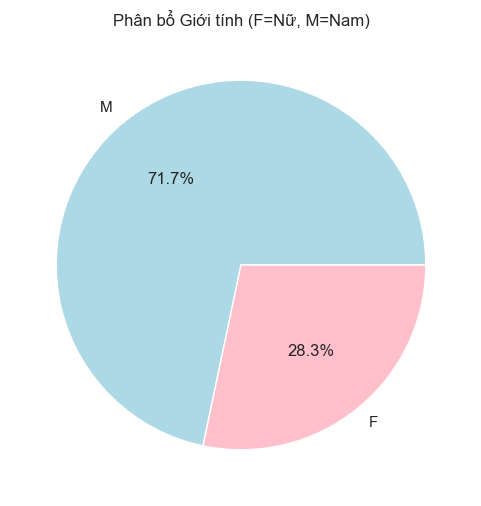
### 1.1 Xử lý giá trị bị thiếu



Hình 2: Xử lý giá trị bị thiếu

## 2. EDA trực quan bằng sơ đồ (Graphical EDA)

### 2.1 Biểu đồ Phân bổ giới tính



Hình 3: Hình phân bố giới tính

**Sự Áp đảo của Nam giới:** Dữ liệu cho thấy một sự chênh lệch rất lớn về giới tính.

**Nam giới (M)** chiếm tỷ lệ áp đảo với **71.7%**.

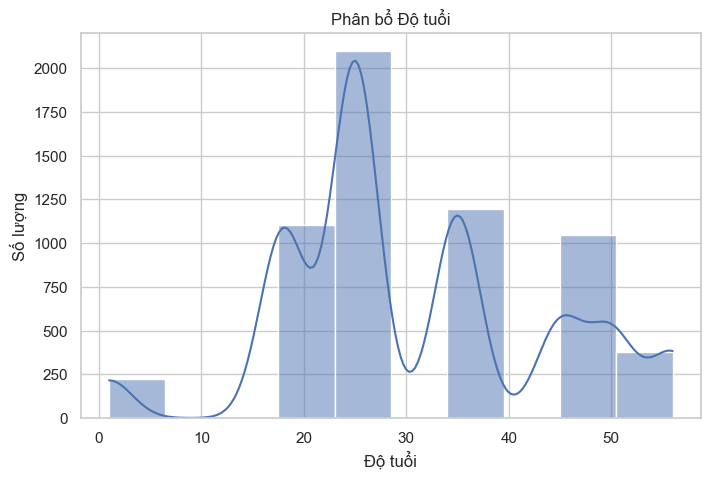
**Nữ giới (F)** chỉ chiếm **28.3%** còn lại.

**Kết luận:**

Tỷ lệ nam giới trong tập dữ liệu này cao gần gấp 2.5 lần so với nữ giới.

Điều này có nghĩa là các phân tích và xu hướng mà chúng ta đã thấy trong các biểu đồ trước (về thể loại, độ tranh cãi, v.v.) chủ yếu bị chi phối bởi quan điểm và hành vi đánh giá của nam giới.

### 2.2 Biểu đồ Phân bổ độ tuổi



Hình 4: Biểu đồ phân bố độ tuổi

- Các đỉnh nổi bật:

+ Đỉnh cao nhất (Mode): Nhóm tuổi 25-30 chiếm số lượng đông đảo nhất, với hơn 2000 người. Đây rõ ràng là nhóm người dùng cốt lõi.

+ Đỉnh thứ hai: Nhóm tuổi 35-40 cũng rất đáng kể, với khoảng 1200 người.

+ Đỉnh thứ ba: Nhóm tuổi 18-22 (khoảng 1100 người) và nhóm 45-50 (khoảng 1050 người) cũng là các nhóm quan trọng.

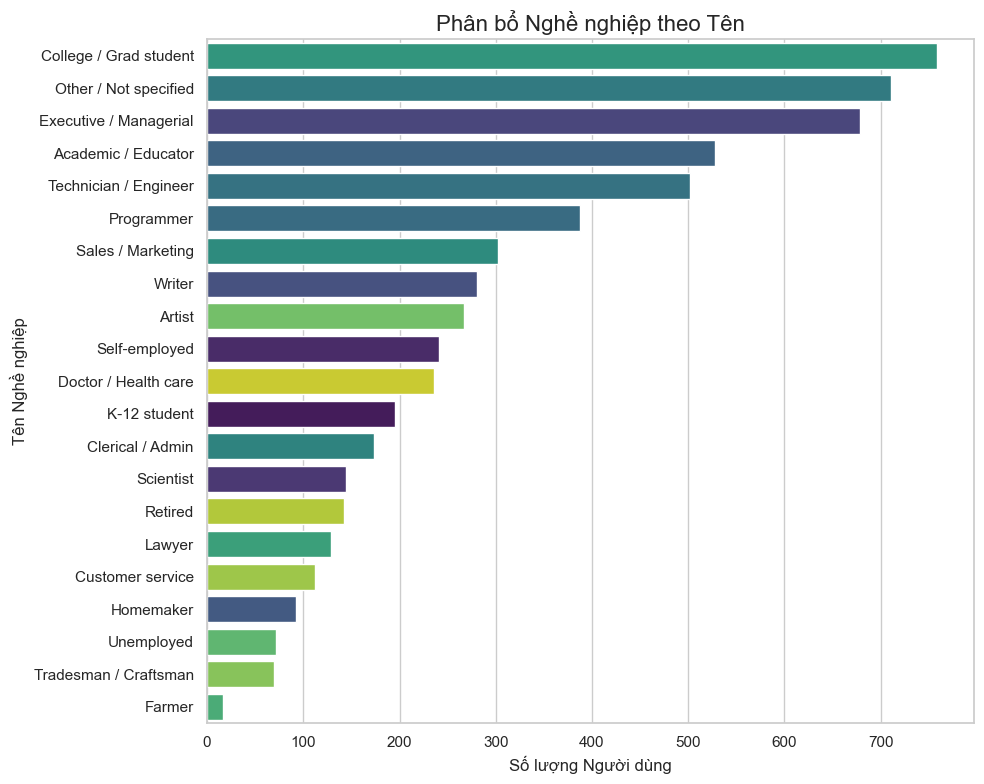
- Điểm đáng chú ý:

+ Có một nhóm nhỏ trẻ em (khoảng 0-10 tuổi).

+ Có một "khoảng trống" dữ liệu đáng kể ở độ tuổi thiếu niên (khoảng 10-17 tuổi), cho thấy đây không phải là đối tượng của dịch vụ này.

**Nhận xét:** Dữ liệu người dùng này bao gồm nhiều nhóm nhân khẩu học riêng biệt, không đồng nhất. Các nhóm chính có thể tương ứng với: Sinh viên/Mới ra trường (18-22), Chuyên gia trẻ (25-30), Chuyên gia có kinh nghiệm (35-40) và Người trung niên (45-50).

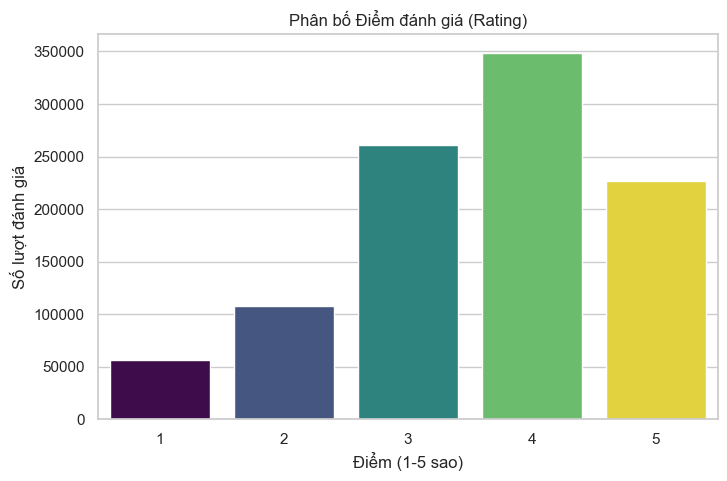
### 2.3 Biểu đồ phân bố nghề nghiệp theo tên



Hình 5: Biểu đồ phân bố nghề nghiệp theo tên

**Nhận xét:** Tập dữ liệu này cho thấy rõ ràng sự thiên lệch về cơ cấu nghề nghiệp. Phần lớn người dùng là những người có trình độ học vấn cao (sinh viên, học giả) hoặc làm trong các lĩnh vực chuyên môn (quản lý, kỹ sư, lập trình viên). Ngược lại, các ngành nghề lao động chân tay hoặc truyền thống (như nông dân) gần như không được đại diện. Điều này gợi ý mạnh mẽ rằng dữ liệu được thu thập từ một cộng đồng trực tuyến, có học thức và có thể thuộc về môi trường đô thị.

### 2.4 Phân bố Điểm đánh giá (Rating)



Hình 6: Biểu đồ phân bố điểm đánh giá

- Xu hướng Chung: Lệch về phía tích cực: Biểu đồ phân bổ bị lệch rõ rệt về bên phải (lệch âm), cho thấy người dùng trong tập dữ liệu này có xu hướng đánh giá phim một cách tích cực.

- Đỉnh cao nhất: 4 sao

+ Mức điểm phổ biến nhất, được trao nhiều nhất là 4 sao, với gần 350,000 lượt đánh giá.

+ Điều này cho thấy khán giả có xu hướng "thích" phim, nhưng có thể khá "tiết kiệm" điểm 5 sao, khiến 4 sao trở thành mức điểm "tốt" phổ biến nhất.

- Thứ tự Phổ biến:

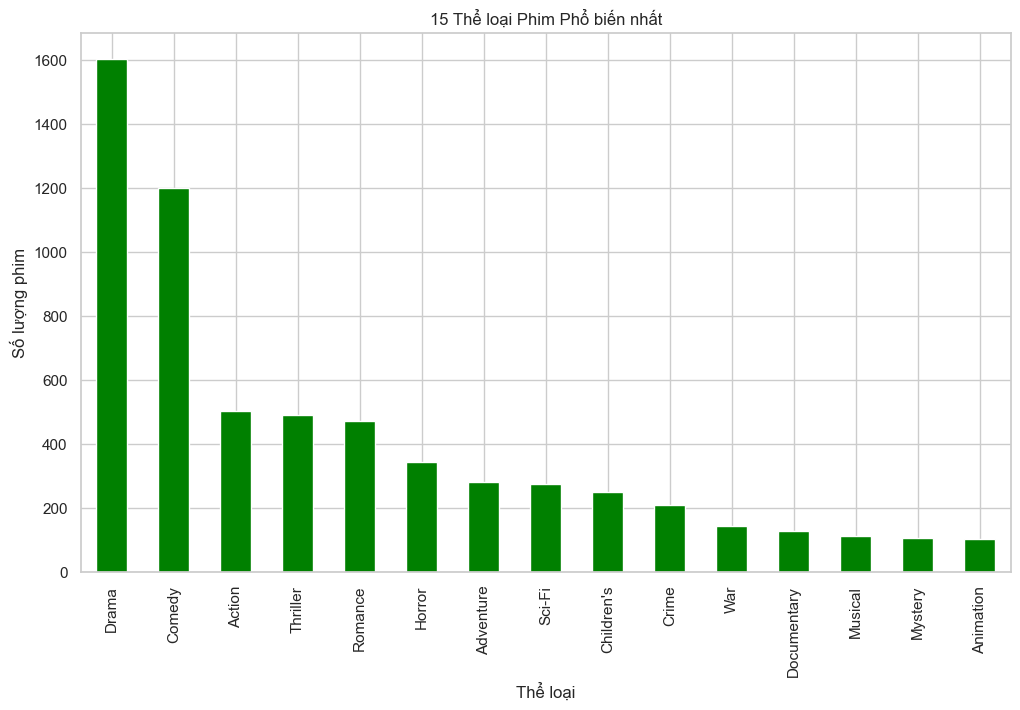
+ Sau 4 sao, mức điểm phổ biến thứ hai là 3 sao (khoảng 260,000).

+ 5 sao đứng ở vị trí thứ ba (khoảng 225,000).

+ Các đánh giá tiêu cực 1 sao (~55,000) và 2 sao (~105,000) là ít phổ biến nhất.

Kết luận: Nhìn chung, khán giả trong mẫu dữ liệu này có thiên hướng đánh giá tốt. Họ trao điểm 4 sao nhiều hơn đáng kể so với điểm 5 sao, và số lượng đánh giá tích cực (3, 4, 5 sao) hoàn toàn áp đảo các đánh giá tiêu cực (1, 2 sao).

### 2.5 Các thể loại phim phổ biến nhất



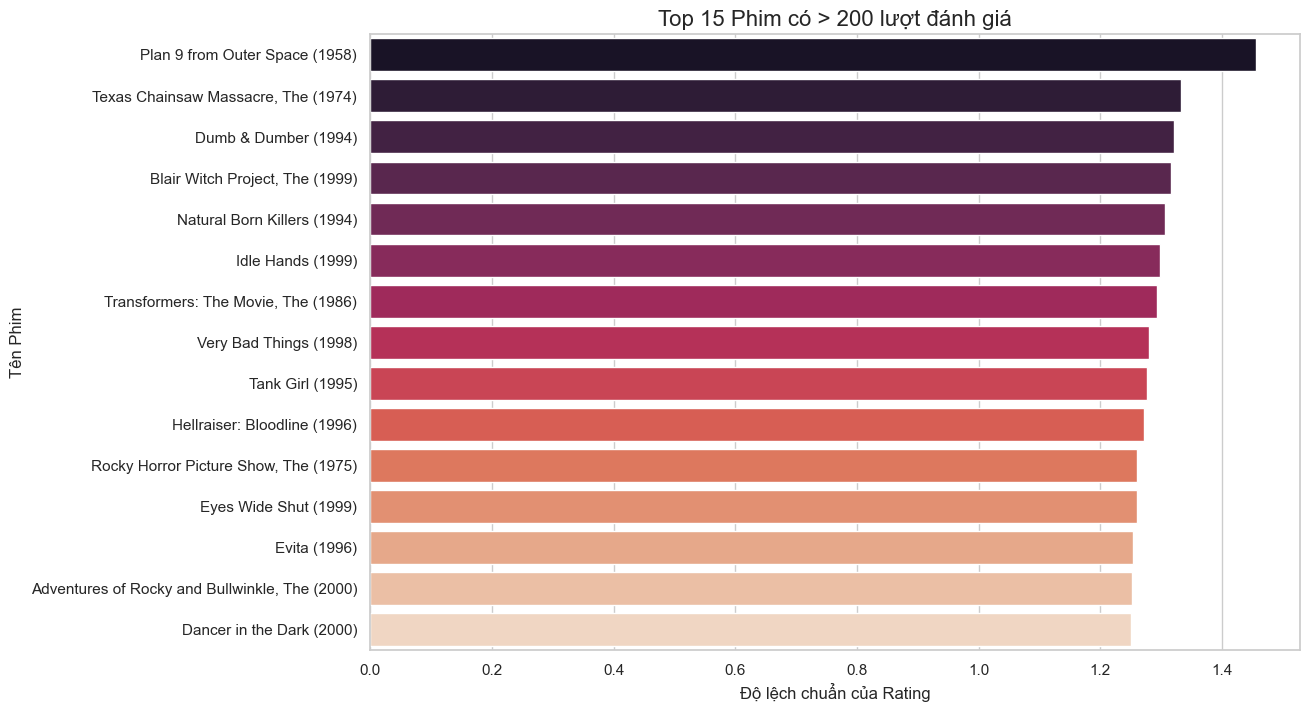
Hình 7: Biểu đồ thể hiện các thể loại phim phổ biến nhất

- Sự thống trị của Hài kịch (Comedy): Thể loại Comedy đứng đầu một cách vượt trội, với tổng số lượt xem cao hơn đáng kể so với tất cả các thể loại còn lại. Điều này cho thấy phim hài là thể loại có tính giải trí cao, dễ tiếp cận và thu hút được lượng khán giả đại chúng lớn nhất.

- Nhóm dẫn đầu: Cùng với Comedy, thể loại Drama (Kịch) cũng là thể loại phổ biến nhất. Đây là các thể loại "chủ lực" (mainstream) luôn có sức hút lớn trên thị trường và nhận được nhiều sự quan tâm.

- Nhóm bám đuổi: Có một khoảng cách rõ rệt giữa top 2 và các thể loại còn lại. Nhóm bám đuổi bao gồm Action (Hành động), Thriller (Giật gân), Sci-Fi (Khoa học viễn tưởng), và Adventure (Phiêu lưu). Đây đều là những thể loại có lượng fan trung thành và phổ biến.

### 2.6 Các thể loại phim có >200 lượt đánh giá



Hình 8: Biểu đồ thể hiện các phim có trên 200 lượt đánh giá

**Nhận xét:**

- Quán quân gây Tranh cãi: "Plan 9 from Outer Space (1958)". Lý do: Bộ phim này nổi tiếng được mệnh danh là "bộ phim dở nhất mọi thời đại". Chính vì điều này, nó tạo ra sự phân cực rất lớn.

- Các Phim Gây Tranh cãi Cao Khác:

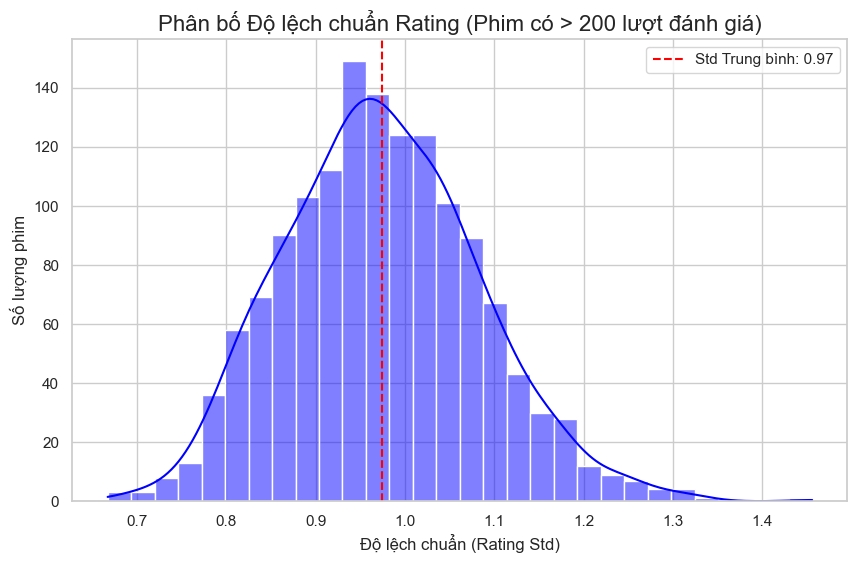
"Texas Chainsaw Massacre, The (1974)" (hạng 2): Một phim kinh dị có tính bạo lực đột phá vào thời điểm ra mắt, gây sốc cho nhiều người nhưng cũng được tôn sùng bởi fan hâm mộ thể loại này.

"Dumb & Dumber (1994)" (hạng 3): Một phim hài "nhảm" (silly comedy). Thể loại hài này rất chủ quan: người xem hoặc thấy nó cực kỳ hài hước, hoặc thấy nó cực kỳ ngớ ngẩn.

- Những bộ phim này là những phim tạo ra phản ứng cảm xúc mạnh mẽ và đối nghịch. Chúng hiếm khi nhận được đánh giá "trung bình" (2, 3, 4 sao) mà thay vào đó, chúng nhận được rất nhiều điểm tuyệt đối (1 sao hoặc 5 sao), như đã được minh họa trong biểu đồ 3 về đánh giá cực đoan.

## Độ lệch chuẩn

### 3.1 Độ lệch chuẩn Rating



Hình 9: Biểu đồ thể hiện độ lệch chuẩn rating

**Nhận xét :**

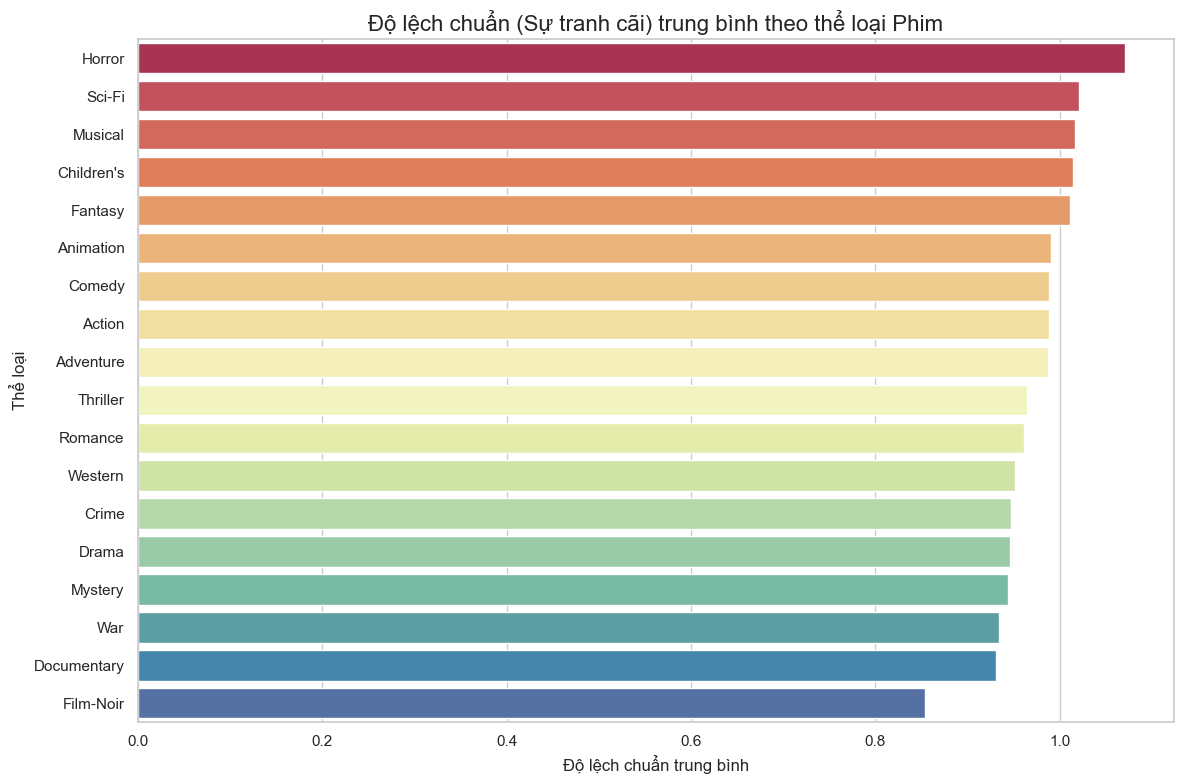
- Điều này cho thấy hầu hết các bộ phim đều có mức độ tranh cãi ở mức trung bình, tập trung dày đặc quanh giá trị 0.97 (đường gạch đỏ).

- Có rất ít phim mà mọi người đều đồng thuận tuyệt đối (độ lệch chuẩn rất thấp, ở phía bên trái).

- Cũng có rất ít phim cực kỳ phân cực (độ lệch chuẩn rất cao, ở phía bên phải).

Kết nối với Biểu đồ "Top 15": 15 bộ phim gây tranh cãi nhất mà chúng ta thấy ở biểu đồ trước (như "Plan 9" với độ lệch chuẩn ~1.45) chính là những trường hợp cực kỳ hiếm và đặc biệt, nằm ở phần đuôi ngoài cùng bên phải của biểu đồ hình chuông này.

### 3.2 Độ lệch chuẩn trung bình theo thể loại phim



Hình 10: Biểu đồ độ lệch chuẩn trung bình theo thể loại phim

**Nhận xét:**

- Thể loại Gây Tranh cãi Nhất:

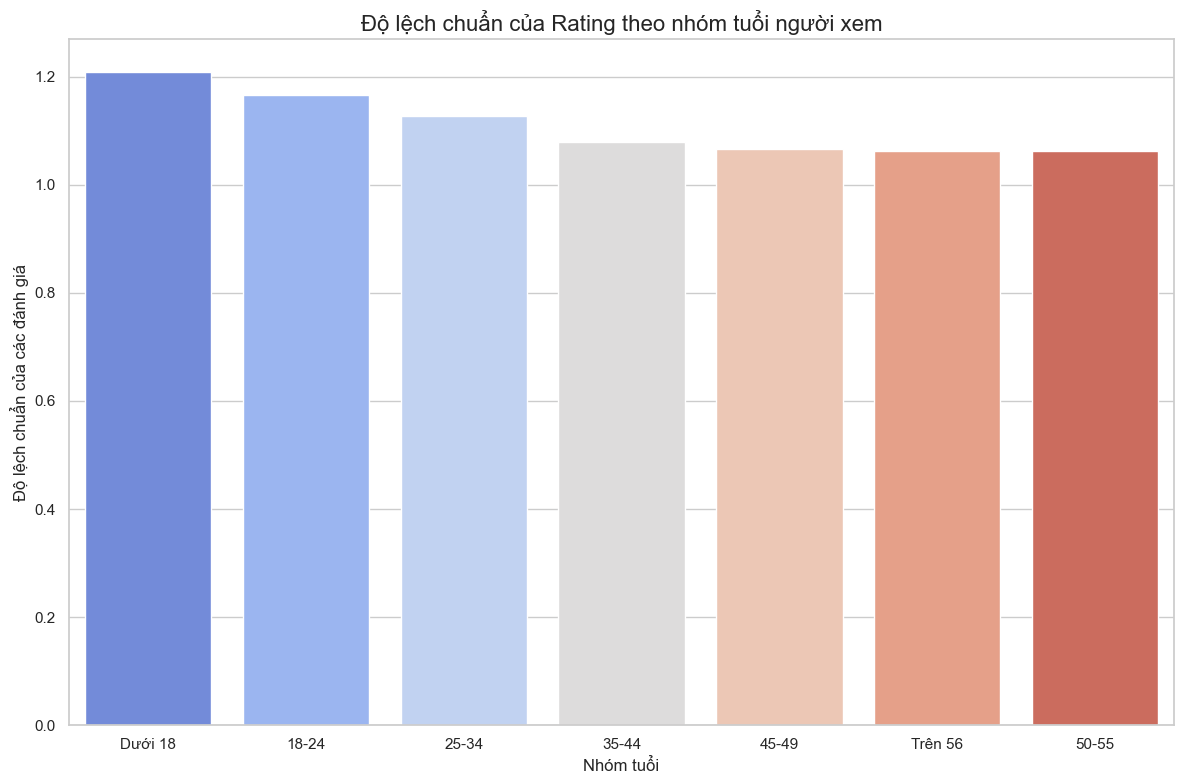
Horror (Kinh dị) đứng đầu một cách rõ rệt. Đây là thể loại "yêu hoặc ghét" điển hình; yếu tố sợ hãi, bạo lực... mang tính chủ quan cao, dẫn đến sự phân cực mạnh mẽ.

Các thể loại theo sau là Sci-Fi (Khoa học viễn tưởng), Musical (Ca nhạc), Children's (Thiếu nhi) và Fantasy (Kỳ ảo). Đây cũng là những thể loại kén khán giả, đòi hỏi người xem phải chấp nhận những yếu tố phi thực tế, dẫn đến nhiều ý kiến trái chiều.

- Thể loại Ít Tranh cãi Nhất (Đồng thuận cao): Film-Noir là thể loại ít gây tranh cãi nhất

- Theo sau là Documentary (Tài liệu), War (Chiến tranh), và Mystery (Huyền bí). Những thể loại này thường có lượng khán giả hâm mộ riêng (niche) hoặc dựa trên sự thật, dẫn đến các đánh giá có xu hướng đồng nhất hơn.

### 3.3 Độ lệch chuẩn của rating theo nhóm tuổi



Hình 11: Biểu đồ độ lệch chuẩn rating theo nhóm tuổi

**Nhận xét :**

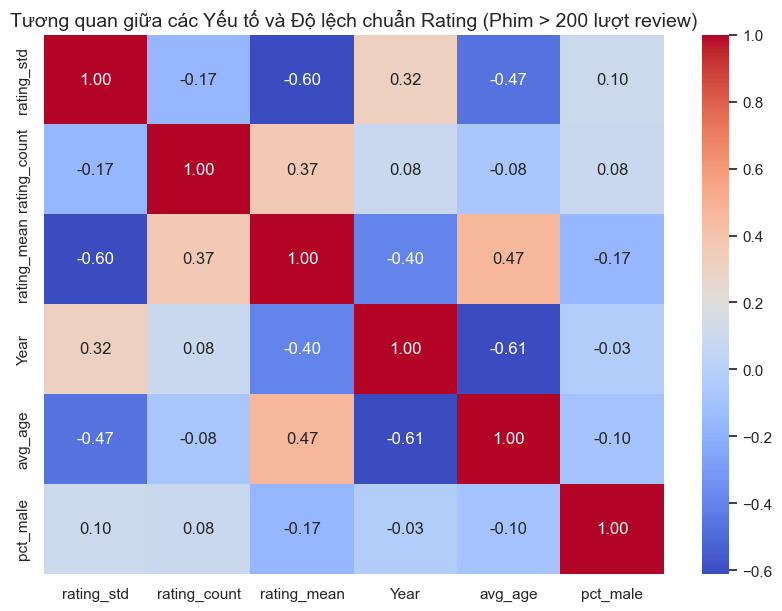
Xu hướng chính: Mức độ tranh cãi cao nhất ở nhóm tuổi trẻ nhất ("Dưới 18") và giảm dần một cách ổn định khi độ tuổi tăng lên.

Ý nghĩa:

- Người xem trẻ tuổi có xu hướng đánh giá phân cực nhất. Các đánh giá của họ có độ phân tán cao, cho thấy sự chia rẽ "yêu-ghét" rõ rệt.

- Người xem lớn tuổi (đặc biệt là các nhóm từ 45 tuổi trở lên) có xu hướng đồng thuận cao nhất. Các đánh giá của họ ít phân tán hơn, cho thấy họ có xu hướng đồng ý với nhau nhiều hơn về chất lượng của một bộ phim.

### 3.4. **Biểu** đồ trực quan hóa ma trận tương quan (Correlogram)



Hình 12: Biểu đồ trực quan hóa ma trận tương quan

**Nhận xét:**

- rating\_mean (Điểm trung bình): -0.6. Đây là mối tương quan nghịch mạnh nhất trong cột.

Ý nghĩa: Phim có điểm trung bình càng cao (tức là phim càng được đánh giá là hay) thì độ tranh cãi càng thấp (tức là mọi người càng đồng thuận). Ngược lại, những phim có điểm trung bình thấp có xu hướng gây tranh cãi nhiều hơn.

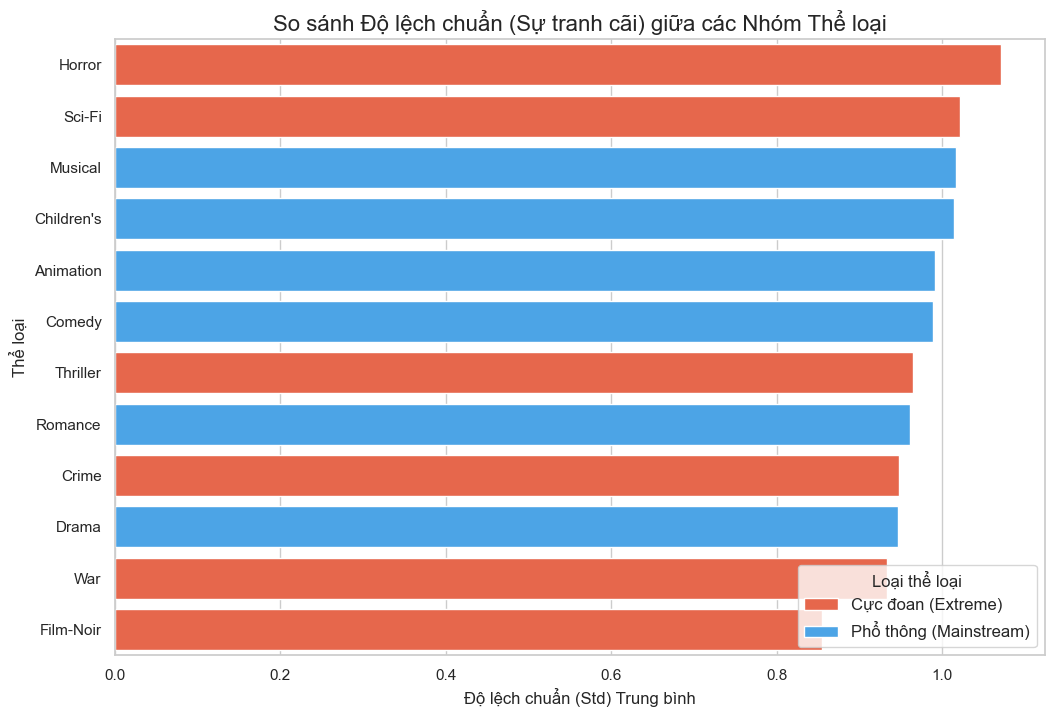
- avg\_age (Tuổi trung bình): -0.47. Đây là mối tương quan nghịch vừa phải.

Ý nghĩa: Khi độ tuổi trung bình của khán giả xem phim càng cao, mức độ tranh cãi càng giảm. Điều này cho thấy người xem lớn tuổi có xu hướng đồng thuận với nhau hơn.

- Year (Năm sản xuất): 0.32. Đây là mối tương quan thuận vừa phải.

Ý nghĩa: Phim càng mới (năm sản xuất càng lớn) thì mức độ tranh cãi càng cao.

### 3.5 So sánh độ lệch chuẩn giữa các thể loại



Hình 13: Biểu đồ so sánh độ lệch chuẩn giữa các thể loại

**Nhận xét về biểu đồ:**

Nhóm "Phổ thông (Mainstream)"

Bao gồm: Musical, Children's, Animation, Comedy, Romance, Drama.

- Phân tích: Điểm đáng chú ý nhất là các thể loại này có mức độ tranh cãi rất đồng nhất. Tất cả chúng đều tụ lại thành một cụm ở mức trung bình đến cao (khoảng 0.95 đến 1.01).

- Ý nghĩa: Điều này cho thấy các thể loại phim đại chúng, dù khác nhau về nội dung, nhưng lại có chung một mức độ gây chia rẽ trong khán giả tương đối ổn định.

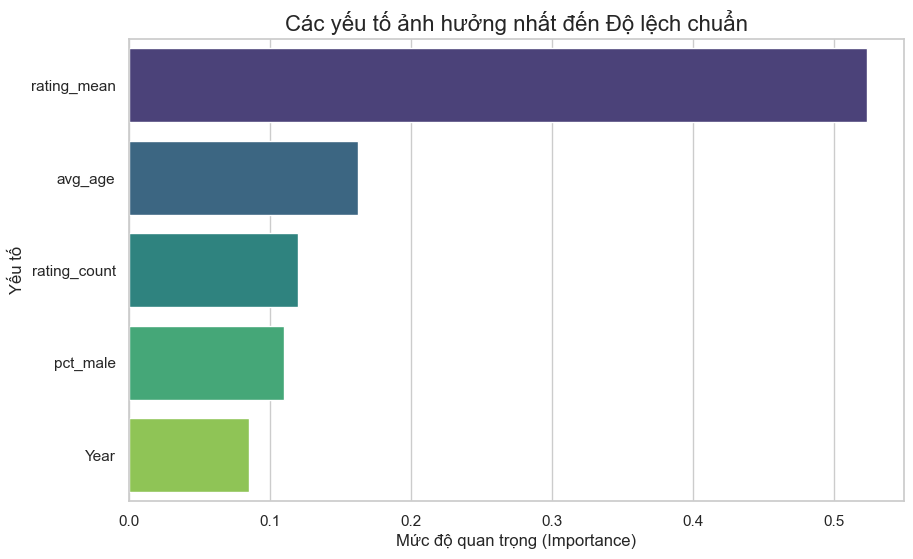
Nhóm "Cực đoan (Extreme)"

Bao gồm: Horror, Sci-Fi, Thriller, Crime, War, Film-Noir.

- Phân tích: Ngược lại với nhóm "Phổ thông", nhóm "Cực đoan" có biên độ tranh cãi rất rộng.

- Ý nghĩa: Nhóm này chứa 2 thể loại gây tranh cãi nhất (Horror và Sci-Fi). Đồng thời, nó cũng chứa các thể loại ít gây tranh cãi nhất (War và Film-Noir). Điều này ngụ ý rằng các thể loại "Extreme" (có thể hiểu là "kén khán giả" hoặc "niche") có tính chất rất khác nhau: một số (như Horror) được làm ra để gây phân cực, trong khi một số khác (như Film-Noir) có thể thu hút một lượng fan rất trung thành và đồng điệu, dẫn đến sự đồng thuận cao.

### 3.6 Các yếu tố ảnh hưởng đến độ lệch chuẩn



Hình 14: Biểu đồ thể hiện các yếu tố ảnh hưởng độ lệch chuẩn

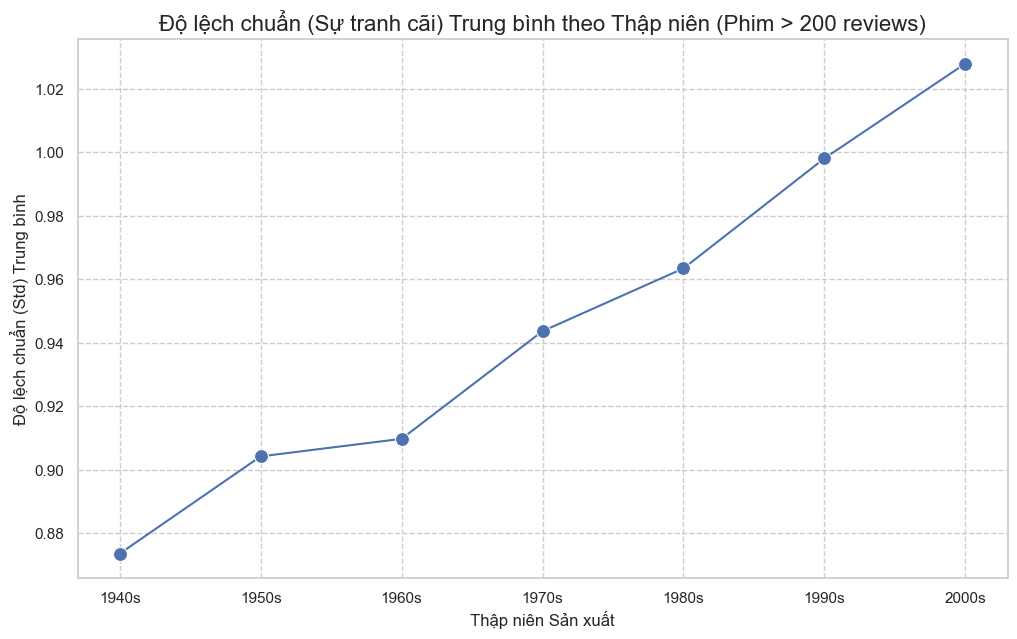
**Nhận xét về biểu đồ:**

Để biết một bộ phim có gây tranh cãi hay không, yếu tố quan trọng nhất cần xem xét chính là điểm đánh giá trung bình của nó. Mức độ quan trọng của nó (hơn 50%) còn lớn hơn tổng của cả 4 yếu tố còn lại cộng lại.

Nhân khẩu học (Tuổi tác) quan trọng: Yếu tố quan trọng thứ hai là "gu" của các thế hệ. Sự khác biệt về tuổi tác của khán giả là một yếu tố lớn thứ hai tạo ra sự phân cực.

Năm sản xuất ít ảnh hưởng nhất (trong top 5): Mặc dù biểu đồ trước cho thấy sự tranh cãi tăng theo thập niên (Year), nhưng khi phân tích cùng các yếu tố khác, bản thân "Năm sản xuất" lại là yếu tố có sức ảnh hưởng trực tiếp thấp nhất. Điều này có thể ngụ ý rằng sự gia tăng tranh cãi theo thời gian (biểu đồ 1) có thể bị ảnh hưởng bởi các yếu tố khác thay đổi theo thời gian (như rating\_count tăng lên, hoặc avg\_age của người dùng thay đổi).

### 3.7 Độ lệch chuẩn theo thập niên



Hình 15: Biểu đồ thể hiện độ lệch chuẩn theo thập niên

**Nhận xét về biểu đồ:**

Phim ảnh ngày càng phân cực: Càng về sau, các bộ phim càng có xu hướng tạo ra nhiều ý kiến trái chiều. Khán giả ít có "tiếng nói chung" hơn so với trước đây.

Sự đồng thuận giảm dần: Trong những năm 1940, dường như dễ dàng hơn để tìm thấy những bộ phim mà đa số mọi người đều đồng ý là "hay" hoặc "dở". Ngày nay (ít nhất là đến những năm 2000), việc đó khó khăn hơn nhiều.

Các giả thuyết có thể giải thích cho xu hướng này (ngoài phạm vi biểu đồ):

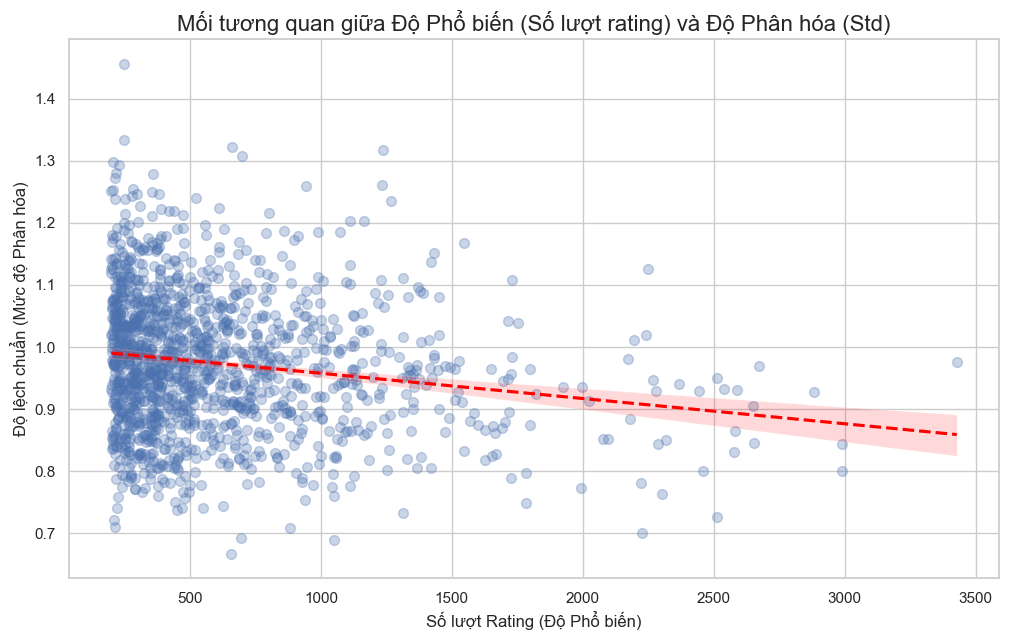
Sự đa dạng hóa về thể loại và chủ đề (nhiều phim "kén" khán giả hơn).

Sự phát triển của internet và văn hóa review (mọi người có xu hướng đưa ra các đánh giá cực đoan hơn).

Sự thay đổi trong cách làm phim, tập trung vào việc tạo ra các "bom tấn" gây tiếng vang lớn (cả tích cực và tiêu cực).

## Mối tương quan

### 4.1 Mối tương quan giữa độ phổ biến và độ phân hóa



Hình 16: Mối tương quan giữa độ phổ biến và độ phân hóa

**Nhận xét về biểu đồ:**

-Phân tích sự Phân tán Dữ liệu (Các chấm xanh)

Phim ít phổ biến (Phía bên trái, < 1000 rating): Đây là khu vực có sự biến động lớn nhất. Các chấm xanh phân tán rất rộng theo chiều dọc (từ 0.7 đến 1.4+).

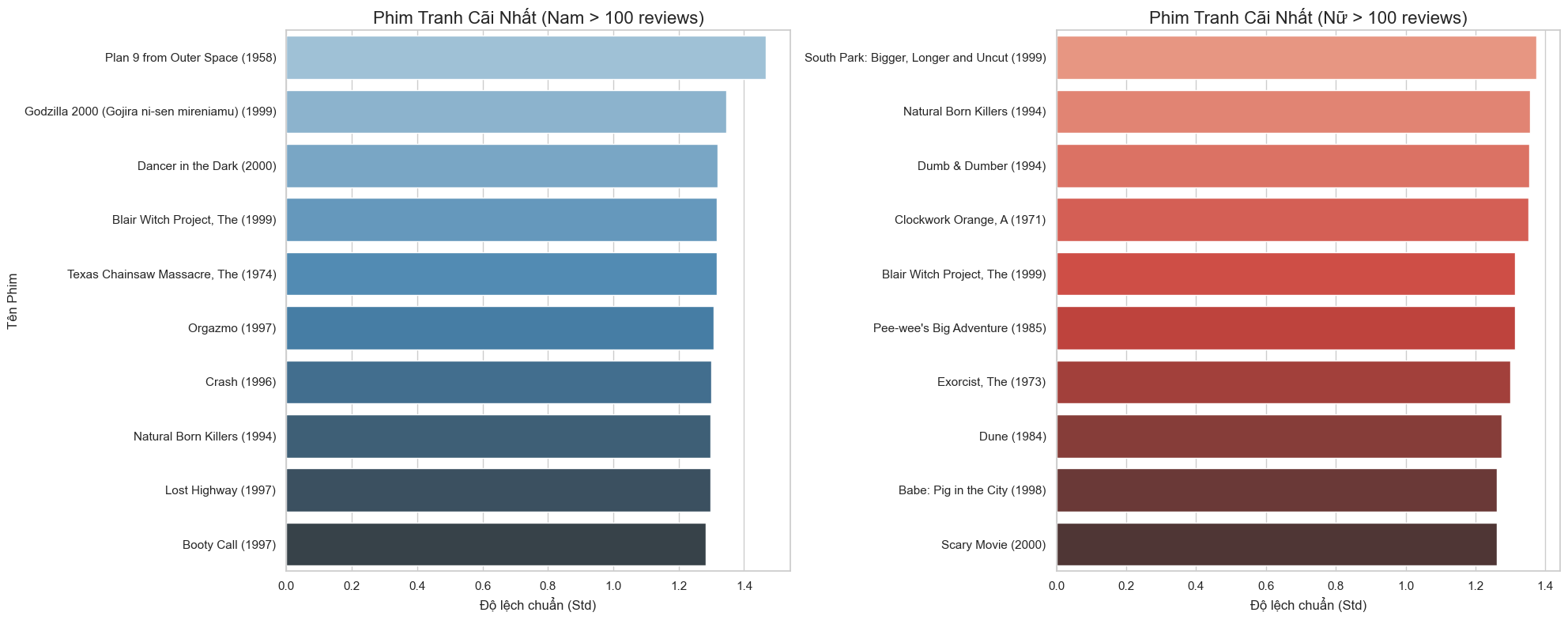
Điều này có nghĩa là trong nhóm phim "niche" (kén khán giả), tồn tại cả những phim cực kỳ gây tranh cãi (độ lệch chuẩn cao, ở phía trên) và những phim có độ đồng thuận rất cao (độ lệch chuẩn thấp, ở phía dưới).

Phim rất phổ biến (Phía bên phải, > 1500 rating): Các chấm xanh trở nên thưa thớt hơn và có xu hướng tụ lại gần nhau, chủ yếu ở mức độ tranh cãi từ trung bình đến thấp (hầu hết dưới 1.1).

Rất hiếm có bộ phim nào vừa rất nổi tiếng lại vừa rất gây tranh cãi.

Sự tranh cãi và phân cực mạnh mẽ chủ yếu xảy ra ở các bộ phim "niche" hoặc ít phổ biến. Các bộ phim muốn trở thành "bom tấn" (thu hút được nhiều người xem) thường phải có nội dung dung hòa và ít gây chia rẽ hơn để đạt được sự đồng thuận của số đông.

### 4.2 Các loại phim tranh cãi nhất giữa nam và nữ



Hình 17: Các loại phim tranh cãi nhất giữa nam và nữ

Danh sách Gây tranh cãi Rất Khác biệt: Điểm nổi bật nhất là danh sách top 10 của hai giới gần như hoàn toàn khác nhau. Điều này cho thấy yếu tố tạo nên sự "yêu-ghét" (phân cực) cho một bộ phim phụ thuộc mạnh mẽ vào giới tính của người xem.

Với Nam: Phim tranh cãi nhất là "Plan 9 from Outer Space (1958)". Đây là một phim "cult classic" nổi tiếng vì bị xem là "phim dở nhất mọi thời đại", do đó tạo ra sự chia rẽ giữa người xem vì mục đích giải trí ("so bad it's good") và người xem nghiêm túc.

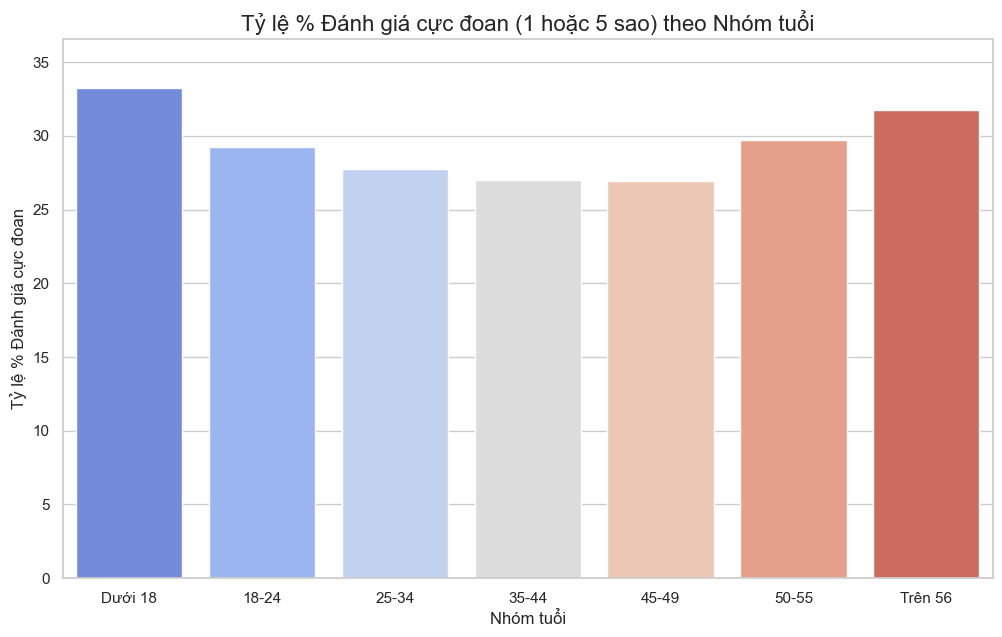
Với Nữ: Phim tranh cãi nhất là "South Park: Bigger, Longer and Uncut (1999)". Đây là một phim hài hoạt hình tục tĩu, gây chia rẽ rõ rệt giữa những người thấy nó hài hước và những người thấy nó phản cảm.

- Điểm chung Hiếm hoi: Chỉ có hai bộ phim xuất hiện trong top 10 của cả hai giới:

"Natural Born Killers (1994)", "Blair Witch Project, The (1999)". Điều này cho thấy đây là những bộ phim có phong cách độc đáo (bạo lực cách điệu hoặc giả tài liệu) có khả năng gây chia rẽ sâu sắc cho bất kỳ khán giả nào, bất kể giới tính.

Kết luận: Nhìn chung, "gu" phim gây tranh cãi của nam và nữ là rất khác biệt. Nam giới bị phân cực bởi các phim "cult" hoặc khoa học viễn tưởng ("Godzilla 2000"), trong khi nữ giới bị phân cực bởi các phim hài ("Dumb & Dumber") hoặc phim có chủ đề gây sốc ("A Clockwork Orange", "The Exorcist").

### 4.3 Tỷ lệ đánh giá cực đoan theo nhóm tuổi



Hình 18: Tý lệ đánh giá cực đoan theo nhóm tuổi

**Nhận xét về biểu đồ:**

Xu hướng "Hình chữ U": Khán giả ở hai đầu của phổ tuổi (trẻ nhất và lớn tuổi nhất) là những người có xu hướng đưa ra các đánh giá phân cực nhất.

Sự "Ôn hòa" của Tuổi trung niên: Các nhóm tuổi ở giữa (đặc biệt là từ 25 đến 49 tuổi) có xu hướng "ôn hòa" hơn trong các đánh giá của họ. Họ ít đưa ra các mức điểm tuyệt đối (1 hoặc 5 sao) và có thể có xu hướng cho điểm ở mức giữa (2, 3, 4 sao) nhiều hơn.

# III. KẾT LUẬN

- Phân tích tập dữ liệu đánh giá phim đã chứng minh một xu hướng rõ rệt: sự phân cực (tranh cãi) trong thị hiếu khán giả đã tăng lên đáng kể qua các thập niên, với các yếu tố như thể loại, tuổi tác và giới tính ảnh hưởng mạnh mẽ đến sự chia rẽ này.

- Tập dữ liệu MovieLens (được phân tích) đã và đang trở thành một benchmark (tiêu chuẩn) kinh điển trong nghiên cứu và phát triển Hệ thống đề xuất (Recommender Systems).

- Hướng phát triển: Ứng dụng các mô hình Machine Learning và Deep Learning (như Collaborative Filtering, Matrix Factorization, Neural Networks) để xây dựng hệ thống dự đoán và gợi ý phim chính xác. Tích hợp thêm dữ liệu phi cấu trúc (như nội dung bình luận) để hiểu sâu hơn về lý do gây ra tranh cãi.

- Ý nghĩa thực tế:

+ Hỗ trợ các nền tảng streaming (như Netflix) và rạp chiếu phim tối ưu hóa việc cá nhân hóa nội dung và chiến lược marketing.

+ Giúp người dùng cuối nhận được các gợi ý phim phù hợp, nâng cao trải nghiệm giải trí và giảm thời gian tìm kiếm.

+ Cung cấp cho các nhà làm phim thông tin chi tiết về các phân khúc khán giả để định hướng sản xuất và quảng bá.

# TÀI LIỆU THAM KHẢO

[1] https://grouplens.org/datasets/movielens/