

Trường Đại học Khoa học tự nhiên – Khoa Công nghệ thông tin.

Đồ án thực hành số 01

Trực quan hóa dữ liệu – Data Visualization.

Nhóm 16
Tháng 10, 2024.

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN THỰC HÀNH SỐ 01

Bộ môn: Trực quan hóa dữ liệu.

Tên đề tài: “*Data Visualization with Python*”.

STT nhóm: 16.

Thành viên:

1. 22120378 – Nguyễn Ngọc Khánh Trân.
2. 22120384 – Nguyễn Đình Trí.
3. 22120387 – Trần Đức Trí.
4. 22120412 – Nguyễn Anh Tường.

Thông tin chung:

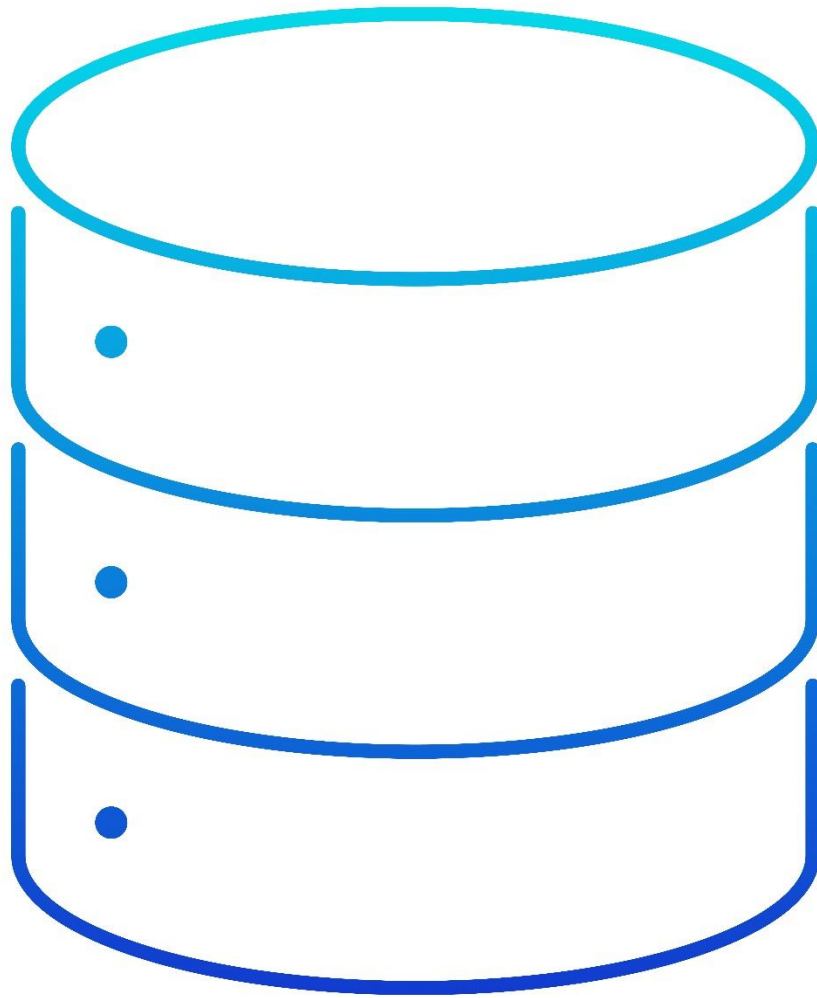
1. **Bộ môn:** Trắc quan hóa dữ liệu.
2. **Giảng viên lý thuyết:** Thầy Bùi Tiến Lên.
3. **Giảng viên thực hành:** Thầy Lê Nhựt Nam.
4. **Mã lớp:** 22_21.
5. **STT nhóm:** 16.
6. **Danh sách thành viên:**
 - a. 22120378 – Nguyễn Ngọc Khánh Trân.
 - b. 22120384 – Nguyễn Đình Trí.
 - c. 22120387 – Trần Đức Trí.
7. **Link Notion quản lý công việc:** [Click](#)

MỤC LỤC

ĐỒ ÁN THỰC HÀNH SỐ 01	2
Thông tin chung:	3
Phần 00: Đóng góp của các thành viên.	6
Phân công công việc theo yêu cầu của đồ án:	7
Phần 01: Giới thiệu tập dữ liệu được dùng.	9
I. Giới thiệu thông tin:	10
II. Giới thiệu các trường dữ liệu:	10
Section 02: Thu thập dữ liệu.	12
I. Động lực và nội dung của tập dữ liệu:	13
II. Cách tập dữ liệu được xây dựng?	13
III. Cách sử dụng tập dữ liệu? Liệu nó có phù hợp trong việc học tập không?	15
Section 03: Khai thác dữ liệu.	16
I. Từng dòng dữ liệu có nghĩa là gì? Nó có ảnh hưởng gì nếu các dòng thay đổi ý nghĩa khác không?	17
II. Những thuộc tính ở cột có nghĩa là gì?	17
III. Kiểu dữ liệu nào từng cột dữ liệu đang có? Có cột nào mà kiểu dữ liệu của nó không phù hợp cho việc khảo sát sau này không?	18
IV. Với từng cột, sự phân bố của các dữ liệu như thế nào? Có cần thiết phải tiền xử lý dữ liệu không? Nếu có thì bạn làm cách nào?	19
Section 04: Chọn lọc, phiên dịch và trực quan các trường dữ liệu và tìm hiểu về những mối quan hệ ẩn của chúng.	20
I. Phân tích tổng số lượng phim trong cuộc khảo sát:	21
1. Phân tích tỉ lệ số phim được phát hành trong từng thập kỉ:	21
II. Phân tích về thể loại phim (genres):	22
1. Phân tích mức độ phổ biến của thể loại (genres):	22
2. Nhận xét:	24
III. Phân tích về loại hình phim (types):	25
1. Phân bố sự phổ biến của loại hình phim	25
2.	26
IV. Phân tích mối tương quan giữa các thuộc tính:	27

1.	Mối quan hệ giữa loại hình phim, mức điểm đánh giá và năm phát hành:	27
2.	Mối quan hệ giữa số lượt bình chọn (votes) và điểm đánh giá trung bình (average rating) thay đổi theo năm phát hành (release year):.....	30
V.	MỘT SỐ GÓC NHÌN CẬN HON:	33
1.	Top 3 thể loại phim (genres) được sản xuất nhiều nhất:.....	33
2.	Mức điểm đánh giá trung bình của tất cả các bộ phim trong hai thập kỉ gần đây:	36
Section 5: Hệ thống gợi ý dựa trên Mức đánh giá, Thể loại, Số lượng vote, và Loại hình phim.		38
<i>Hệ thống nhận gợi ý theo Tựa phim.</i>		<i>38</i>
I.	Giới thiệu mô hình:	39
II.	Các thư viện sử dụng:.....	39
	Thư viện scikit-learn:	39
	Thư viện copy:	39
	Thư viện Numpy:	39
III.	Cài đặt thuật toán:	40
1.	Class Recommender:.....	40
2.	Class Recommender > def Prepare_feature():.....	40
3.	Class Recommender > def get_recommendation():.....	40
IV.	Recommendation System GUI:	41
1.	Cửa sổ chính.	41
2.	Cửa sổ database:	42
3.	Cửa sổ recommendation results:	42

Phần 00: *Đóng góp của các thành viên.*



Phân công công việc theo yêu cầu của đồ án:

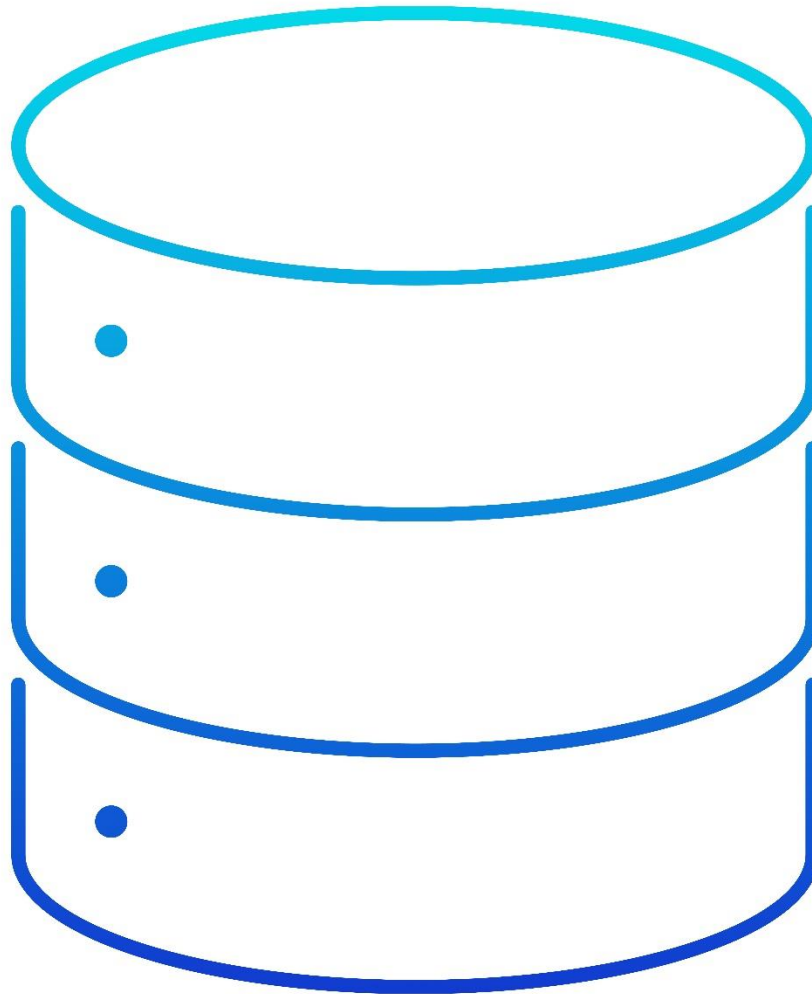
Tên công việc	Người đảm nhận	Đóng góp	% Điểm
Thu thập dữ liệu và tiền xử lý	Nguyễn Anh Tường	100%	5%
Chọn lọc, phiên dịch và trực quan các thuộc tính và tìm ra những mối quan hệ ẩn trong chúng	Nguyễn Ngọc Khánh Trân	28.33%	50%
	Nguyễn Đức Trí	28.33%	
	Trần Đức Trí	28.33%	
	Nguyễn Anh Tường	15%	
Xem xét các mối quan hệ và các quan điểm khác nhau	Nguyễn Ngọc Khánh Trân	33.33%	10%
	Trần Đức Trí	33.33%	
	Nguyễn Đình Trí	33.33%	
Hiểu được code mình nộp lên	Nguyễn Ngọc Khánh Trân	25%	5%
	Trần Đức Trí	25%	
	Nguyễn Đình Trí	25%	
	Nguyễn Anh Tường	25%	
Phân tích và trực quan dữ liệu với các biểu đồ mới mà vẫn cung cấp được thông tin hữu ích. Có sử dụng các mô hình học máy cơ bản.	Nguyễn Anh Tường	100%	5%
Cung cấp nội dung mạch lạc và dễ hiểu sau mỗi dữ liệu được trực quan.	Nguyễn Ngọc Khánh Trân	30%	20%
	Trần Đức Trí	30%	
	Nguyễn Đình Trí	30%	
	Nguyễn Anh Tường	10%	
Bản báo cáo được trình bày trực quan và dễ hiểu	Nguyễn Ngọc Khánh Trân	30%	15%
	Trần Đức Trí	15%	
	Nguyễn Đình Trí	15%	
	Nguyễn Anh Tường	40%	
Kiểm tra lại code và bài báo cáo	Nguyễn Ngọc Khánh Trân	50%	---
	Nguyễn Đình Trí	25%	
	Trần Đức Trí	25%	

Mức độ hoàn thành các chỉ tiêu:

100%



Phần 01: *Giới thiệu tập dữ liệu được dùng.*



I. Giới thiệu thông tin:

Trang web: Kaggle.

Tên tập dữ liệu: IMDb Top Rated Titles (Movies & TV Series).

Liên kết: [Click here!](#)

Bản quyền: không.

Sơ lược: Đây là tập dữ liệu bao gồm hơn 6000 những bộ phim có đánh giá tốt trên IMDb, có cả bản điện ảnh và TVSeries. Đánh giá trung bình nhỏ nhất sẽ là 7 và với hơn 10.000 phiếu bầu chọn.

Lưu ý: Đây là tập dữ liệu được **cập nhật liên tục** mỗi 10:00 AM hằng ngày theo giờ Trung Âu - CET. Chúng em sẽ chỉ sử dụng dữ liệu được cập nhật đến Thứ 7 – ngày 26/10/2024.

II. Giới thiệu các trường dữ liệu:

1. Id – IMDb ID (mã ID theo IMDb):

Kiểu dữ liệu: string.

Đặc điểm: mỗi một ID là duy nhất, dùng để phân biệt các bộ phim với nhau.

Số lượng ID: 6029.

2. Title – Title of Films (Tên phim):

Kiểu dữ liệu: string.

Đặc điểm: Tên bộ phim có thể trùng, do định dạng ở thể loại khác nhau.

Số lượng Title: 6029.

3. Type – Type of Films (Loại hình phim):

Kiểu dữ liệu: string.

Đặc điểm: Loại hình có thể trùng, gồm nhiều loại hình khác nhau,

Ví dụ như: tvSeries, tvMiniSeries, movie, ...

4. Genres – Genres of Films (Thể loại phim):

Kiểu dữ liệu: string.

Đặc điểm: Thể loại có thể trùng, gồm nhiều thể loại khác nhau.

Ví dụ như: Drama, Comedy, Family, ...

5. averageRating – The Average Rating (mức điểm đánh giá trung bình):

Kiểu dữ liệu: float.

Đặc điểm: Có giá trị từ 0.0 đến 10.0, thể hiện đánh giá của người dùng dựa trên mức độ ưa thích của tựa phim đó.

6. numVotes – The Number of Votes (số lượng vote):

Kiểu dữ liệu: int.

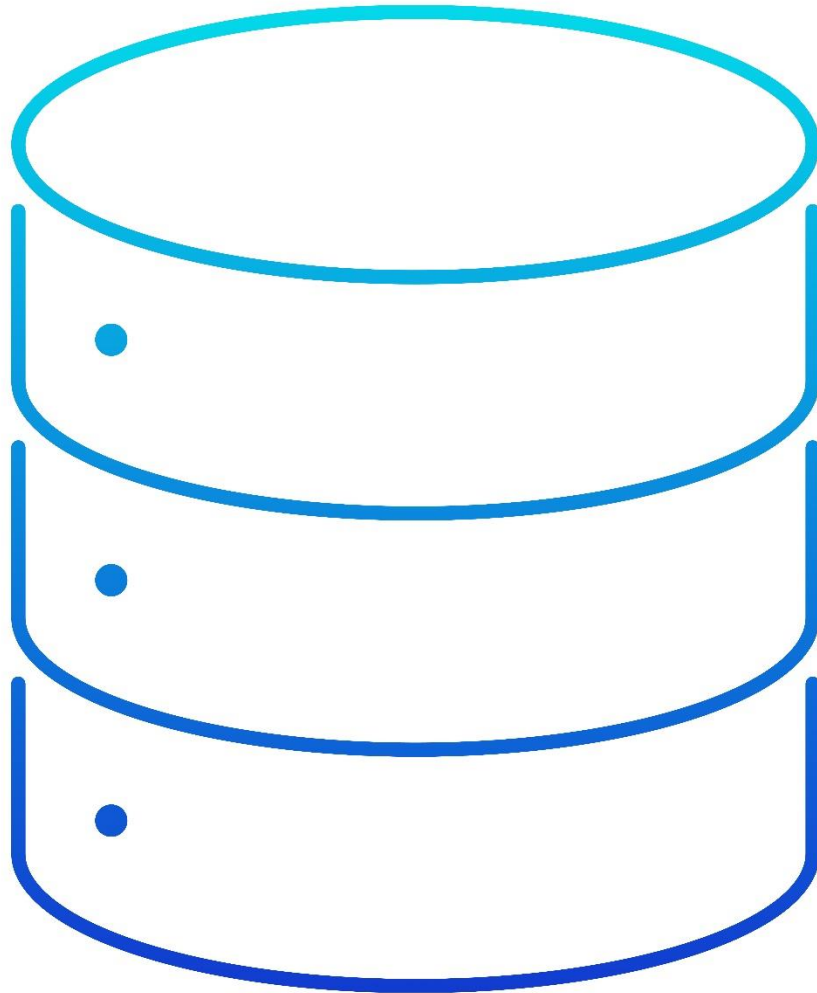
Đặc điểm: Thể hiện số lượng phiếu đánh giá của các tựa phim.

7. releaseYear – Release Year (năm phát hành phim):

Kiểu dữ liệu: int.

Đặc điểm: Thể hiện năm ra mắt tựa phim đó. Tính từ 1916 đến năm 2024.

Section 02: *Thu thập dữ liệu.*



I. Động lực và nội dung của tập dữ liệu:

“IMDb Top Rated Titles (Movies & TV Series)” là một tập dữ liệu gồm các đánh giá về Top các bộ phim hot trên trang web IMDb – một website đánh giá và review nội dung phim uy tín của thế giới.

Trong tập dữ liệu này có hơn **6000 tác phẩm** phim với nhiều thể loại, tổng cộng đến hơn **700 triệu lượt đánh giá** các film trải dài từ những năm 1916 đến 2024.

Đây là một tập dữ liệu được **cập nhật liên tục** hàng ngày vào mỗi sáng 10 giờ theo giờ Trung Âu (CET). Cho nên việc sử dụng dữ liệu sẽ có khác biệt giữa các thời điểm khác nhau.

Đây là một tập dữ liệu có tính tổng quát rộng, với cả hai loại hình phát sóng là TV Series (Phim truyền hình dài tập) và Movies (Phim điện ảnh).

Vì đây là những bộ phim từ hot cho đến rất hot nên mục đánh giá sẽ có phần khắc khe, tập dữ liệu chỉ chọn ra các tác phẩm có rating cao trên 7.0 và số lượt vote phải lớn hơn hoặc bằng 10.000 (votes).

Chúng em chọn tập dữ liệu này vì những điều sau:

1. Chúng em là những người yêu thích phim ảnh.
2. Chúng em đang muốn khám phá thêm nhiều tựa phim hay.
3. Chúng em đang kỳ vọng mình có thể tạo được một mô hình gợi ý phim cho khán giả dựa vào thể loại và thời lượng phim và mức độ đánh giá và năm phát hành.

II. Cách tập dữ liệu được xây dựng?

Đây là một tập dữ liệu nằm trong một series gồm 4 tập dữ liệu lớn:

1. [IMDb Top 1000 Movies.](#)
2. [IMDb Top 1000 TV Series.](#)
3. [IMDb Top 1000 Worst Rated Titles.](#)
4. [IMDb Top Rated Titles \(Movies & TV Series \).](#)

Dữ liệu của tập dữ liệu được lấy trực tiếp từ website [IMDb.com](https://www.imdb.com).

IMDb (viết tắt của *Internet Movie Database*) là một cơ sở dữ liệu trực tuyến về thông tin liên quan đến phim, phim truyền hình, podcast, video gia đình, trò chơi điện tử và nội dung phát trực tuyến – bao gồm dàn diễn viên, đoàn làm phim và tiểu sử cá nhân, tóm tắt cốt truyện, thông tin thú vị, xếp hạng và đánh giá của người hâm mộ và phê bình.

IMDb là trang web được truy cập nhiều thứ 52 trên Internet, theo xếp hạng của Alexa . Tính đến tháng 3 năm 2022, cơ sở dữ liệu chứa khoảng 10,1 triệu tựa phim (bao gồm cả các tập phim truyền hình), 11,5 triệu hồ sơ cá nhân và 83 triệu người dùng đã đăng ký.

Mặc dù công thức hiện tại không được tiết lộ, nhưng IMDb ban đầu đã sử dụng công thức sau để tính xếp hạng có trọng số của họ:

$$W = \frac{R \cdot v + C \cdot m}{v + m}$$

Trong đó:

- W là xếp hạng có trọng số.
- R là điểm đánh giá trung bình của bộ phim, từ 1 đến 10.
- v là số phiếu bầu cho bộ phim.
- m là số phiếu bầu tối thiểu cần thiết để được liệt kê trong Top 250 (25.000 tính đến năm 2013).
- C là rating trung bình. (7,0 tính đến năm 2013).

IMDb cũng có tính năng Bottom 100 được xây dựng thông qua một quy trình tương tự mặc dù chỉ cần nhận được 10.000 phiếu bầu để đủ điều kiện vào danh sách.

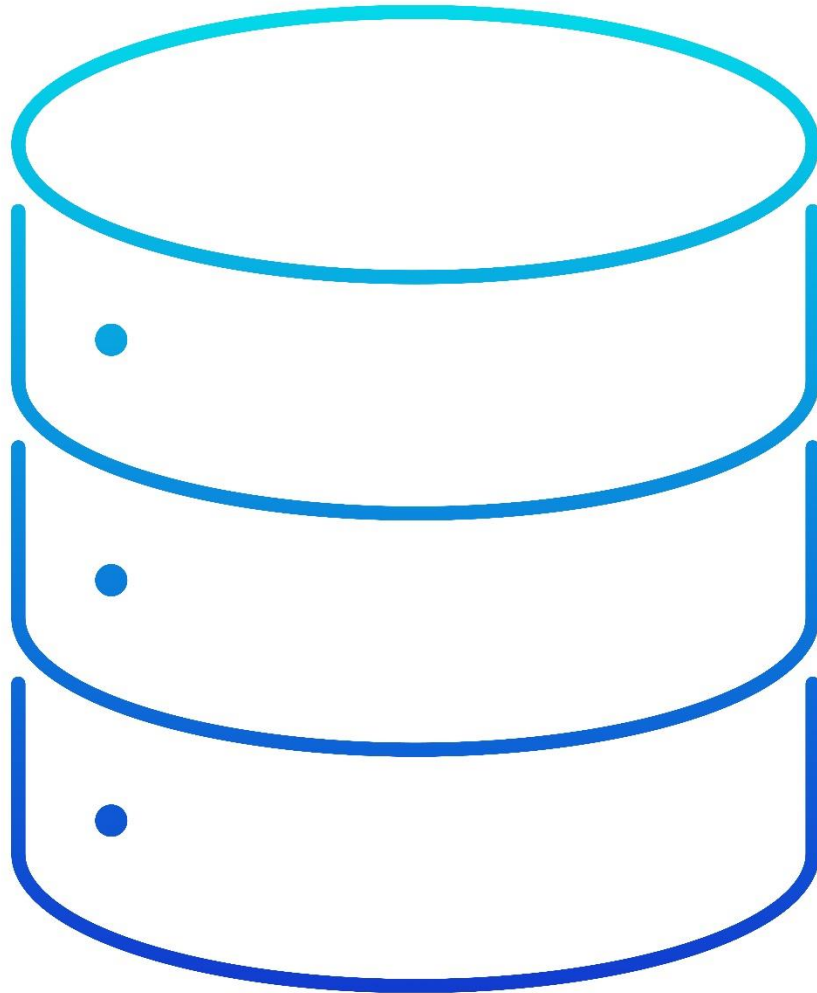
III. Cách sử dụng tập dữ liệu? Liệu nó có phù hợp trong việc học tập không?

Đây là một tập dữ liệu được sử dụng một cách hoàn toàn miễn phí và tự do. Với chế độ bản quyền thuộc **CC0 1.0 Universal**: No Copyright.

Với dạng bản quyền này thì ta có thể sử dụng tập dữ liệu cho thao tác từ việc copy, chỉnh sửa, phân phát, sử dụng trong công việc, v.v hay thậm chí là dành cho việc kinh doanh, có lợi nhuận.

Đây là liên kết bản quyền của tập dữ liệu: [Click here!](#)

Section 03: *Khai thác dữ liệu.*



I. Từng dòng dữ liệu có nghĩa là gì? Nó có ảnh hưởng gì nếu các dòng thay đổi ý nghĩa khác không?

Mỗi dòng tương ứng với một dữ liệu một bộ phim.

Mỗi dòng bao gồm các thuộc tính:

IMDb ID	Movie Title	Movie Type	Movie Genres	Average Rating	Num Votes	Release Year
---------	-------------	------------	--------------	----------------	-----------	--------------

Việc thay đổi các dòng sẽ dẫn đến các sai sót về mặt dữ liệu và các ràng buộc liên quan. Ví dụ như trùng ID, nhầm lẫn giữa Movie Type và Movie Genres,...

II. Những thuộc tính ở cột có nghĩa là gì?

IMDb ID: Mã số định danh của mỗi bộ phim thuộc IMDb, mã số này là duy nhất cho từng bộ phim. (kể cả việc trùng tên nhưng sai khác ở dạng phát sóng – **movie type**)

Movie Title: Tên của bộ phim đó.

Movie Type: Định dạng phát sóng của bộ phim. Phim điện ảnh (movies), Phim truyền hình dài tập (TVSeries), Phim truyền hình ngắn tập (MiniTV Series).

Movie Genres: Thể loại của bộ phim đó. Ví dụ: Drama (kịch tính), Family (gia đình), Horror (Kinh dị),...

Average Rating: Đánh giá trung bình của người xem về bộ phim đó. Có giá trị từ 1 đến 10. Riêng trong tập dữ liệu này thì thấp nhất là 7.0.

Num Votes: Tổng số phiếu bầu đánh giá của bộ phim đó.

Release Year: Năm phát hành bộ phim.

III. Kiểu dữ liệu nào từng cột dữ liệu đang có? Có cột nào mà kiểu dữ liệu của nó không phù hợp cho việc khảo sát sau này không?

IMDb ID: String – Kiểu chuỗi.

Movie Title: String – Kiểu chuỗi.

Movie Type: String – Kiểu chuỗi.

Movie Genres: String – Kiểu chuỗi.

Average Rating: Float – Kiểu số thực.

Num Votes: Int – Kiểu số nguyên.

Release Year: Int – Kiểu số nguyên.

Ở đây chỉ có cột Movie Title và IMDb ID là ít có khả năng sử dụng trong việc xử lý và xây dựng mô hình ở giai đoạn sau.

Vì mô hình tại em hướng đến là gợi ý phim dựa trên thể loại, thời lượng, rating, số lượng votes và năm phát hành. Với hướng phát triển này thì tại em chú trọng vào xử lý các phần sau:

1. **Movie Type:** Dùng để lấy ra thời lượng phim.
 1. Movies: Thời lượng phim ngắn – Đa phần từ 1 – 3 tiếng.
 2. MiniTV Series: Thời lượng ngắn – Kéo dài qua ít tập – Từ 1 đến 12 tập phát sóng.
 3. TV Series: Thời lượng dài – Kéo dài qua nhiều tập – Lớn hơn 12 tập phát sóng và đa phần dừng lại ở tập thứ 24 hoặc kéo sang các phần tiếp theo.
2. **Movie Genres:** Dùng để lấy ra thể loại phim. Bao gồm nhiều thể loại tùy theo nhu cầu của người dùng, ví dụ: hành động, giải trí, gia đình, kinh dị, hình sự, hoạt hình, hài hước,...
3. **Average Rating:** Dùng để lấy ra những đánh giá từ những người dùng khác đã xem qua từ đó đưa ra gợi ý người trực tiếp cho người dùng hiện tại.
4. **NumVotes:** Dùng để lấy ra số lượng vote của các phim, vì có nhiều phim điểm đánh giá tương đồng nhưng lượt vote lại chênh lệch. Ví dụ cùng lượt đánh giá thì nhiều lượt vote hơn sẽ hay hơn và ngược lại.
5. **Release Year:** Dùng để lấy ra năm phát hành, được sắp theo thứ tự ưu tiên giảm dần từ gần hiện tại nhất (ra mắt trong năm 2024) đến hết.

IV. Với từng cột, sự phân bố của các dữ liệu như thế nào? Có cần thiết phải tiền xử lý dữ liệu không? Nếu có thì bạn làm cách nào?

Đối với 2 thuộc tính là Average Rating và Release Year hiện đang có kiểu dữ liệu số (float và int), phù hợp với việc tính toán nên chỉ cần format lại theo chung một định dạng (ví dụ như làm tròn 2 chữ số cho cột Average Rating) nếu cần thiết.

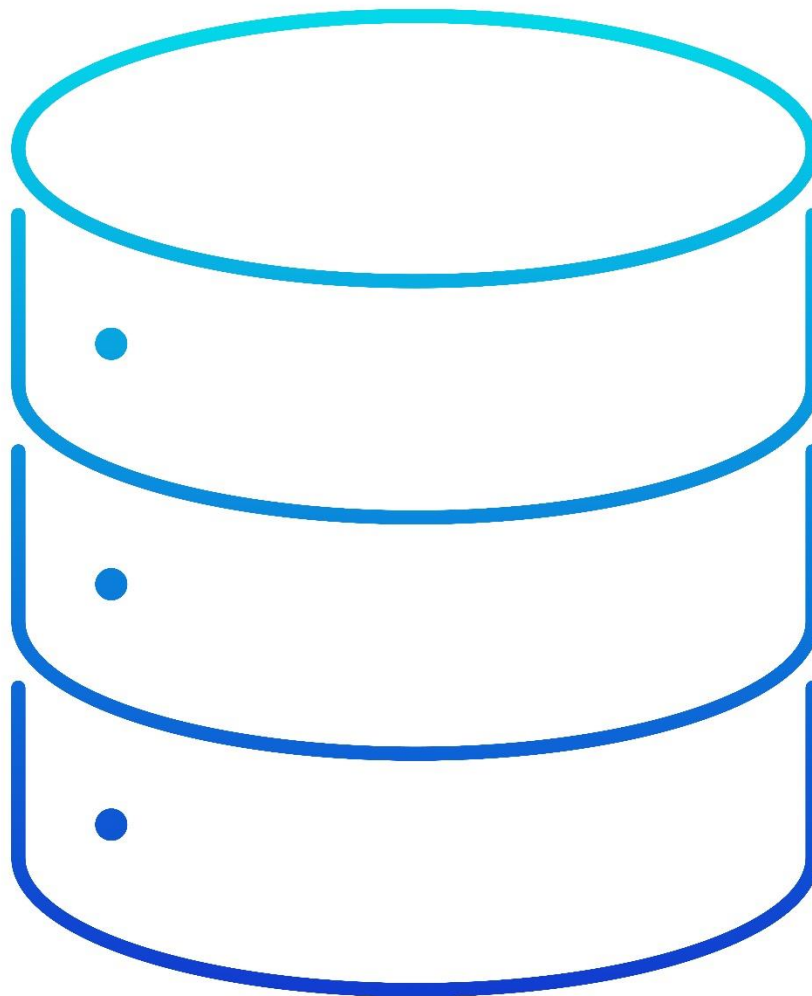
Đối với 2 cột Movie Type và Movie Genres hiện tại đang ở kiểu dữ liệu chuỗi - string, do vậy cần chuyển đổi sang định dạng số thích hợp.

1. Bước đầu tiên là xác định xem có bao nhiêu phân tử phân biệt của các cột này.
2. Bước tiếp theo là chuyển đổi định dạng chuỗi sang số bằng cách thêm định nghĩa cho các cột này.
 - a. Ví dụ Movie Type có 3 dạng là Movies, TV Series, MiniTV Series thì sẽ lần lượt đánh số là 1, 2, 3 cho 3 định dạng phân phối video này.
 - b. Vì một tựa phim có thể có nhiều thể loại cho nên cách đánh số 0, 1, 2, v.v sẽ không hiệu quả đối với cột này. Do đó ta cần phải tách cột này thành nhiều cột khác nhau với cùng kiểu dữ liệu số nguyên, giá trị là 0 hoặc 1.

Trong đó:

- Giá trị 1 đại diện cho việc tựa phim có thể loại đó.
- Giá trị 0 đại diện cho việc tựa phim không có thể loại đó.

Section 04: *Chọn lọc, phiên dịch và trực quan các trường dữ liệu và tìm hiểu về những mối quan hệ ẩn của chúng.*



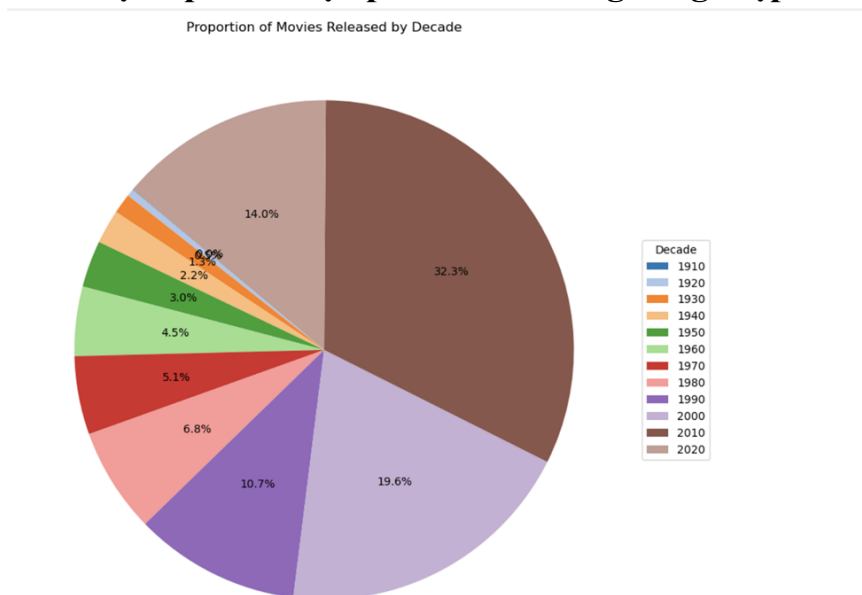
Trong section này, ta sẽ phân tích các thuộc tính phim khác nhau để tiết lộ xu hướng và mối quan hệ ẩn bên trong bộ dữ liệu IMDb. Các phân tích được tổ chức như sau:

1. Phân tích về tổng số lượng phim trong cuộc khảo sát.
2. Phân tích về thể loại phim (genres).
3. Phân tích về các loại hình phim (types).
4. Phân tích mối tương quan giữa các thuộc tính.
5. Một số góc nhìn cận hơn.

I. Phân tích tổng số lượng phim trong cuộc khảo sát:

- Tổng số lượng phim được khảo sát là: 6029
- Thời gian phát hành phim trải dài từ 1916 đến 2024

1. Phân tích tỉ lệ số phim được phát hành trong từng thập kỉ:



Hình 1. Biểu đồ tròn thể hiện tỉ lệ số lượng phim được phát hành trong từng thập kỉ.

Lý do chọn dạng biểu đồ: Biểu đồ tròn được sử dụng ở đây là phù hợp vì nó sẽ hiển thị rõ ràng sự phân bố tỷ lệ của số lượng phim được phát hành theo từng thập kỉ. Điều này giúp ta nhận biết và so sánh phần trăm của số lượng phim trong các thập kỉ khác nhau, và nhận định được đâu là thập kỉ có nhiều phim được sản xuất nhất.

Nhận xét:

- Số lượng phim tăng dần qua từng thập kỷ:
 - Ta thấy được số lượng phim phát hành tăng mạnh từ những năm 1980 trở đi, đặc biệt là trong giai đoạn 2000 và 2010.

- Thập kỷ 2010 chiếm tỷ lệ cao nhất với 32.3%, cho thấy đây là thời kỳ mà công nghiệp điện ảnh phát triển mạnh mẽ và có nhiều phim được sản xuất.
- Giai đoạn phát triển vượt bậc (từ thập kỷ 1990 đến thập kỷ 2010):
 - Các thập kỷ 1990, 2000 và 2010 chiếm phần lớn tỷ lệ, với các mức lần lượt là 10.7%, 19.6%, và 32.3%.
 - Điều này có thể liên quan đến sự bùng nổ của công nghệ, cho phép việc sản xuất phim trở nên dễ dàng hơn, cùng với sự phát triển của các nền tảng phát hành phim trực tuyến.
- Các thập kỷ trước 1980:
 - Từ năm 1910 đến 1980, số lượng phim phát hành ít hơn đáng kể, với tỷ lệ dưới 7% cho mỗi thập kỷ. Điều này phản ánh giai đoạn đầu phát triển của ngành công nghiệp điện ảnh khi công nghệ và các phương tiện sản xuất phim còn nhiều hạn chế.
- Thập kỷ 2020:
 - Thập kỷ 2020 chỉ chiếm 2.2% trên biểu đồ, có thể do dữ liệu chưa đầy đủ cho thập kỷ này hoặc ảnh hưởng của đại dịch COVID-19, khiến nhiều phim bị trì hoãn hoặc sản xuất ít hơn.

Kết luận:

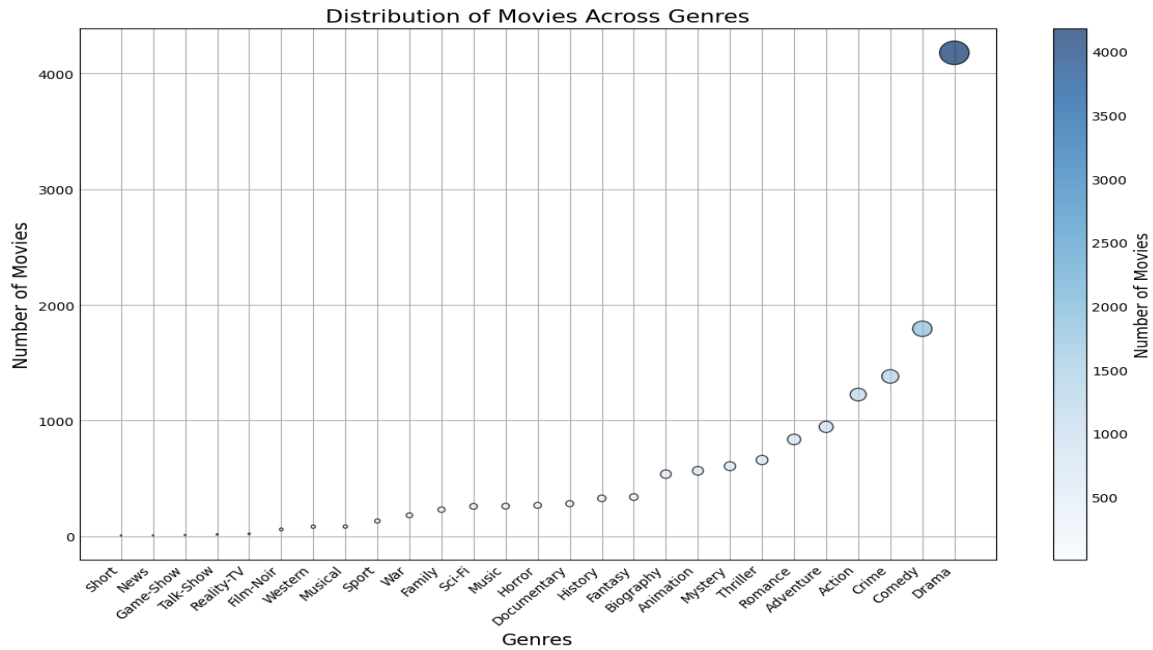
Biểu đồ cho thấy sự tăng trưởng liên tục trong số lượng phim được phát hành qua các thập kỷ, đặc biệt là từ những năm 1980 trở đi. Thập kỷ 2010 là thời kỳ có tỷ lệ phim phát hành cao nhất, trong khi các thập kỷ đầu tiên có số lượng phim rất ít, phản ánh sự phát triển của ngành công nghiệp điện ảnh qua thời gian.

II. Phân tích về thể loại phim (genres):

1. Phân tích mức độ phổ biến của thể loại (genres):

- Bước 1: Tính số lượng mỗi thể loại (genres) theo các bộ phim, lưu vào một biến dictionary.
- Bước 2: Sắp xếp các genres theo tứ tự tăng dần trước khi vẽ các biểu đồ thể hiện phân phối và phân bố.
- Bước 3: Vẽ các biểu đồ thể hiện sự phân tán, phân bố dữ liệu của các genres.

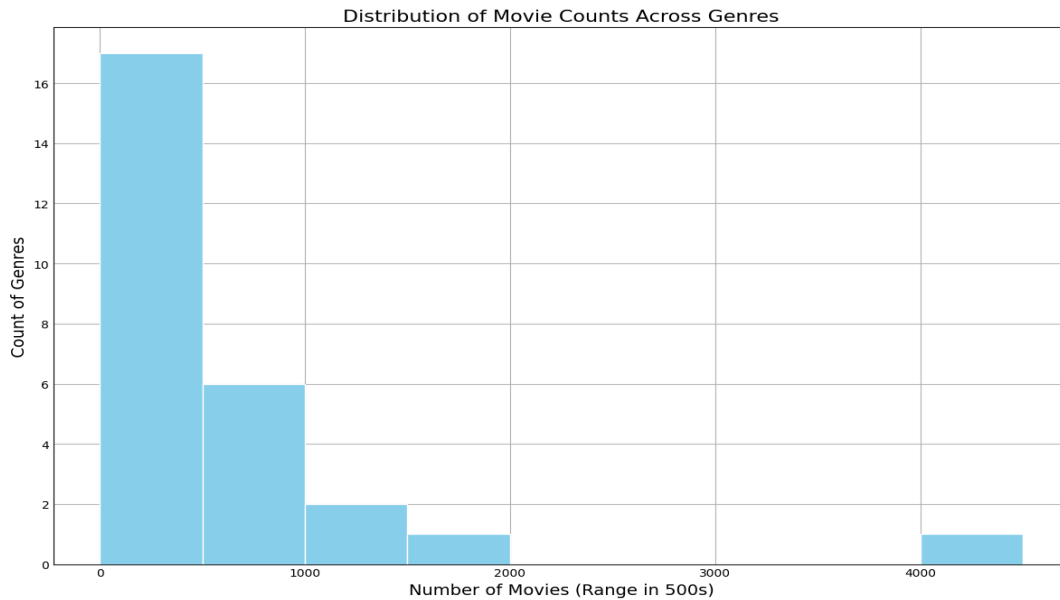
1.1. Biểu đồ phân tán phân bố dữ liệu của các thể loại phim (genres)



Hình 2. Biểu đồ phân tán thể hiện số lượng phim theo các genres.

Lý do chọn dạng biểu đồ: Biểu đồ phân tán được sử dụng trong việc phân tích mức độ phổ biến của các thể loại (genres) là phù hợp vì nó truyền tải được mối quan hệ giữa các thể loại và số lượng phim. Kích thước của các điểm cung cấp cho ta góc nhìn rõ ràng hơn về số lượng và thứ hạng phổ biến của từng thể loại, với điểm càng lớn thì càng phổ biến.

1.2. Biểu đồ thể hiện phân phối số lượng phim của các thể loại phim theo các khoảng cách nhau 500 số lượng



Hình 3. Biểu đồ thể hiện phân phối số lượng bộ phim trong mỗi genres.

Lý do chọn dạng biểu đồ: Việc sử dụng biểu đồ thể hiện phân phối số lượng bộ phim trong mỗi thể loại phim (genres) và được chia ra thành các khoảng nhỏ sẽ giúp cho người dùng dễ dàng quan sát, từ đó người dùng có thể nhận ra số lượng của các thể loại mà ở đó các thể loại này có nhiều phim nhất.

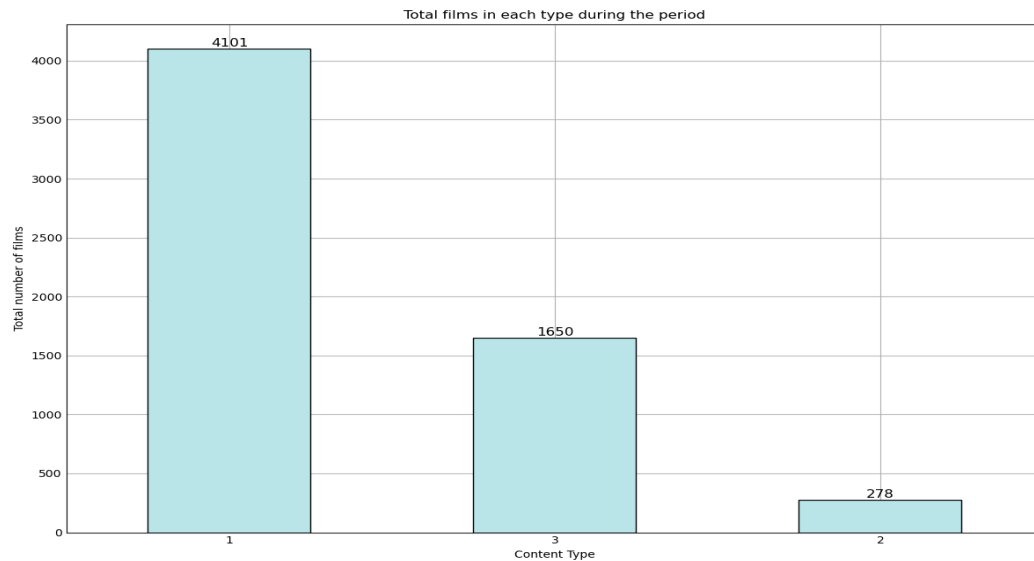
2. Nhận xét:

- Từ các biểu đồ ta thấy trên tổng số hơn 6000 tựa phim, thể loại drama – kịch tính là thể loại có nhiều bộ phim nhất – với hơn 4000 phim. Sau đó là thể loại hài kịch (comedy) với hơn 1700 phim và tội phạm (crime) với hơn 1300 phim.
- Top 3 thể loại có ít phim nhất theo thứ tự từ cao nhất đến thấp nhất là Gameshow, news và short. Với số lượng phim lần lượt là 11 – 8 – 7.
- Mỗi genres thường có khoảng 500 bộ phim lọt vào trong top list này, cá biệt trong đó có các trường hợp dramatic – kịch tính với hơn 4000 bộ phim, có thể nói đây là thể loại có ở hầu hết các tựa phim hoặc hầu hết các bộ phim trong top list này đều có chứa yếu tố kịch tính.
- Có thể thấy người xem ưa chuộng các thể loại giải trí cao, kịch tính, hài hước (comedy và crime) do vậy đánh giá của họ rất tốt đối với dạng thể loại này, dẫn đến danh sách lọt vào rất nhiều tựa phim có cùng thể loại yếu tố.

III. Phân tích về loại hình phim (types):

1. Phân bố sự phổ biến của loại hình phim

- Biểu đồ cột (bar chart) cho biết tổng số lượng của từng thể loại phim được ghi nhận trong suốt khoảng thời gian được khảo sát:



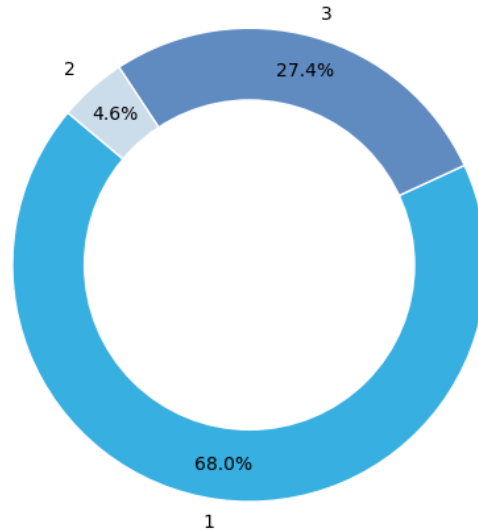
Hình 4. Biểu đồ cột thể hiện tổng số lượng phim theo từng loại hình phim (type).

Lý do chọn dạng biểu đồ: Việc sử dụng biểu đồ cột ở đây sẽ giúp cho người dùng có thể so sánh trực quan tổng số lượng phim giữa các thể loại. Mỗi cột có một chiều cao riêng, nên người dùng từ đó mà dễ dàng so sánh xem cột nào nhiều hơn. Đồng thời người dùng cũng có thể nhận thấy được thứ bậc dữ liệu của các loại hình phim.

Để cụ thể hơn, ta xét về tỉ lệ của từng loại hình phim trong khoảng thời gian được khảo sát như sau:

- Biểu đồ donut (donut chart) cho biết tỉ lệ của từng loại hình phim:

Distribution of Total Films by Content Type



Hình 5. Biểu đồ donut thể hiện tỉ lệ số lượng phim theo từng loại hình phim (type).

Lý do chọn dạng biểu đồ: Bản chất của donut chart khá giống với pie chart. Việc sử dụng donut chart giúp người dùng nhận biết được tỉ lệ các loại hình phim trong khoảng thời gian được khảo sát. Tuy nhiên donut chart có điểm nổi bật ở chỗ nó trực quan dữ liệu rõ ràng và dễ nhìn hơn, vì người dùng sẽ tập trung hơn vào vị trí có màu (chính là phần nội dung)

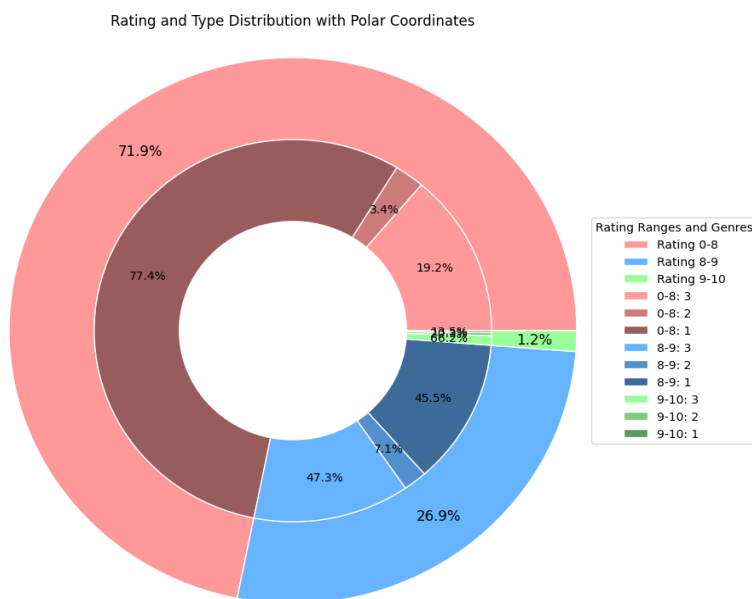
2. Kết luận:

- Ta thấy tỷ lệ của 3 loại phim, với loại 1- phim chiếu rạp (Movie) được sản xuất nhiều nhất từ trước đến nay, chiếm hơn 68% tổng số phim, theo sau đó loại 3- phim truyền hình dài tập (TVSeries) và cuối cùng là loại 2 - phim truyền hình ngắn tập (TVMiniSeries).
- Xu hướng từ trước đến nay là xem phim ở rạp thay vì xem phim đài truyền hình, một phần cũng là do thời gian không cho phép người xem dành thời gian mỗi ngày để xem phim, phần còn lại có thể do các yếu tố âm thanh, hình ảnh sắc nét hơn của phim rạp so với việc xem phim trên đài truyền hình.

IV. Phân tích mối tương quan giữa các thuộc tính:

1. Mối quan hệ giữa loại hình phim, mức điểm đánh giá và năm phát hành:

1.1. Phân bố mức đánh giá trung bình (average rating) theo loại hình phim (types):



Hình 6. Biểu đồ donut với nhiều vòng lồng nhau thể hiện tỉ lệ mức điểm đánh giá (rating) của từng loại hình phim (type).

Lý do chọn dạng biểu đồ: Việc sử dụng biểu đồ donut với nhiều vòng lồng nhau là phù hợp khi biểu diễn mối quan hệ giữa loại hình phim, mức điểm đánh giá và năm phát hành vì nó cho phép hiển thị thông tin cùng một lúc (như việc chia tỉ lệ các mức điểm khác nhau ở vòng trong, trong khi vòng ngoài là các loại hình phim). Mặt khác, việc trình bày như thế này sẽ giúp người xem nhận ra loại hình nào có mức đánh giá là cao nhất hoặc thấp nhất.

Nhận xét:

- Đầu tiên ta xét đến phân bố các mức rating, ở đây ta xét 3 mức là từ 0-8, 8-9 và 9-10.
- Ta có thể quan sát vòng tròn ngoài cùng, có thể nói rằng đa số phim được đánh giá ở mức khá (dưới 8) và không có bộ phim nào dưới 7, tỷ lệ phim ở mức đánh giá này lên đến xấp xỉ 72%. Ở vòng trong, trong mức đánh giá này, ta thấy đa số các phim thuộc loại hình 1 – phim ngắn (movie). Khả năng chất lượng các bộ phim chưa thể khiến một số lượng lớn người xem hài lòng.
- Ở phân khúc 8-9, chiếm khoảng 27%, trong đó, hầu như chỉ toàn phim loại hình 1 và 3, số lượng phim khiến cho người xem cảm thấy thích thú chiếm 1 phần tương đối, nhưng nhìn chung vẫn khó so với phân khúc trên (0-8).

- Phân khúc 9-10 chiếm một tỷ lệ khá thấp với chỉ 1,2% tổng số phim. Đây là những bộ phim xuất sắc để lại nhiều ấn tượng và lấy được lòng khán giả, trong đó loại hình 3 (phim truyền hình dài tập) chiếm đa số trong phân khúc này.

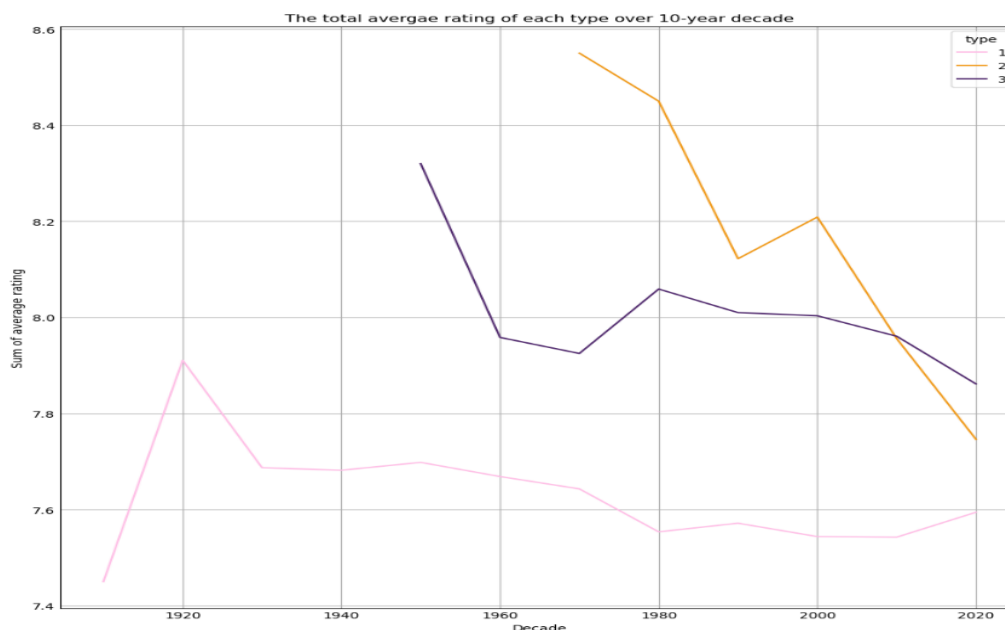
Giải pháp: Cần tập trung cải thiện chất lượng phim điện ảnh (movie) để làm hài lòng người xem. Bên cạnh đó cũng duy trì và phát huy thêm sản xuất phim truyền hình dài tập khi điểm rating của nó là cao nhất trong các loại phim.

1.2. Phân tích xu hướng của loại hình phim (types) và mức đánh giá trung bình (average rating) theo thời gian năm phát hành (release year):

- Ta chia khoảng thời gian trong khảo sát thành những cụm 10 năm và xem xét sự thay đổi về chất lượng được đánh giá của từng thể loại trong những cụm 10 năm này.

- Ta xem xét giá trị trung bình (mean values) của những mức đánh giá trung bình (average rating) trong khảo sát.

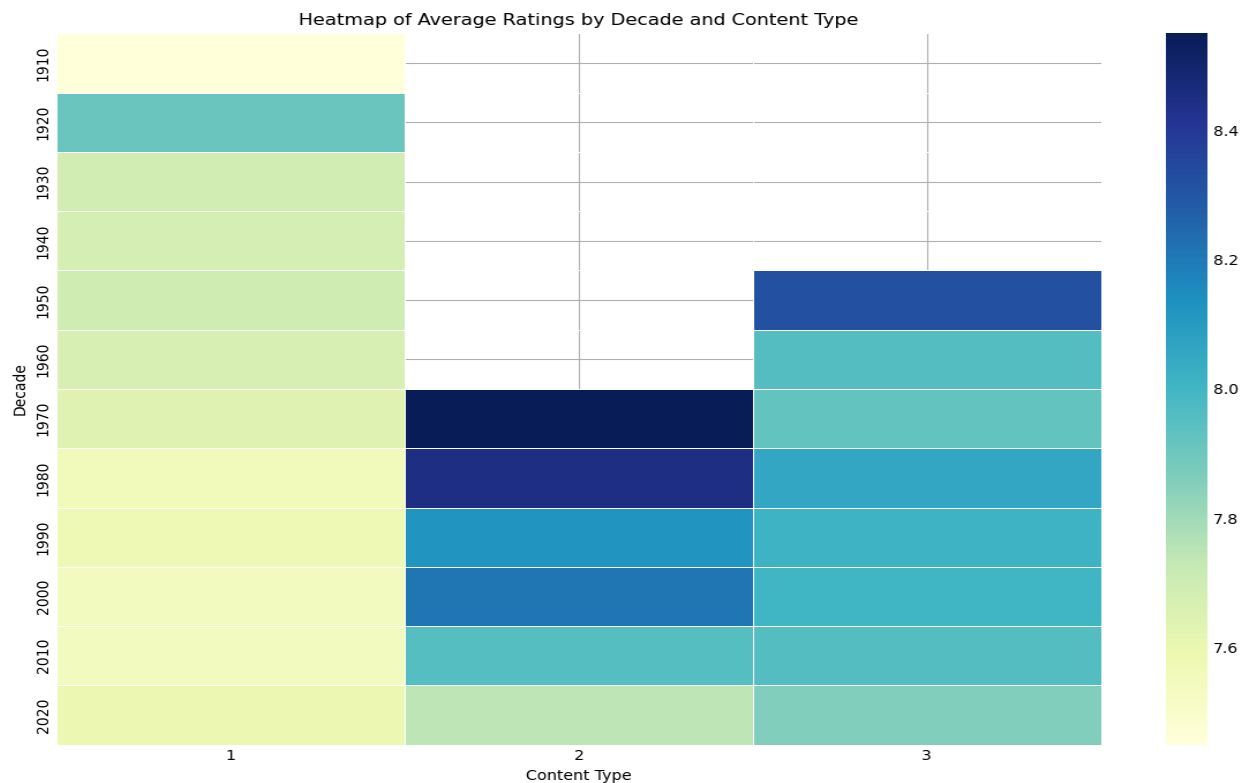
- Ta sử dụng biểu đồ đường để thể hiện xu hướng của mức đánh giá trung bình của từng loại hình phim trong từng cụm 10 năm:



Hình 7. Biểu đồ đường thể hiện xu hướng của mức đánh giá trung bình của từng loại hình phim trong mỗi thập kỉ.

Lý do chọn dạng biểu đồ: Biểu đồ đường được sử dụng với nội dung này là dùng để thể hiện sự thay đổi và xu hướng theo thời gian của mức đánh giá trung bình của các loại hình phim. Từ biểu đồ đường này, ta trực quan được việc điểm trung bình của các loại hình là cao hay thấp ở từng thập kỉ khác nhau và ta cũng biết được vào thời gian nào thì loại hình nào có điểm trung bình cao nhất.

- Kết hợp với biểu đồ phân tán, ta thấy cụ thể hơn về mức điểm rating của các loại hình phim và sự tương quan của nó đến các thập kỉ:



Hình 8. Biểu đồ phân tán thể hiện mức điểm trung bình của rating của từng loại hình theo từng thập kỉ

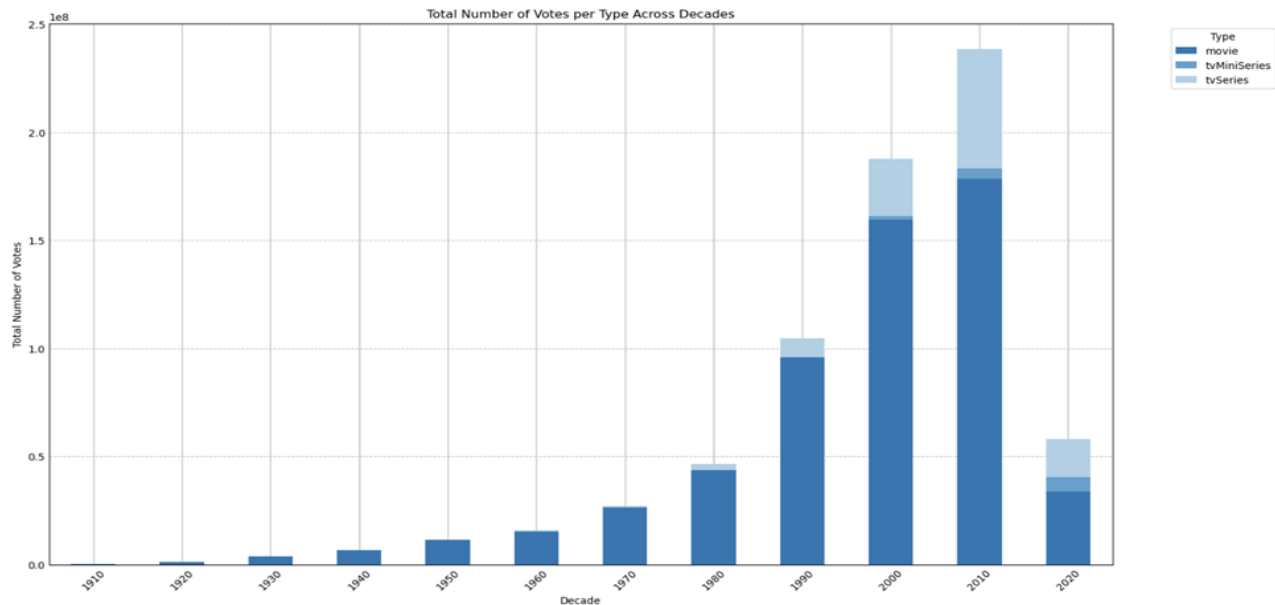
Lý do chọn dạng biểu đồ: Kết hợp với dạng biểu đồ đường thể hiện xu hướng ở trên, biểu đồ phân tán này sẽ thể hiện rõ ràng thang điểm trung bình của từng loại hình qua các thập kỉ. Độ đậm nhạt về màu sắc cũng góp phần cho ta góc nhìn rõ ràng hơn về sự phân tán của thang điểm trung bình.

Kết luận:

- Đối với loại hình 1 (movie): các đánh giá từ những thập kỉ đầu (1910 đến 1960) còn khá thấp, tuy nhiên vào thập kỉ 1980, loại hình này đã được đánh giá ở các mức cao hơn, nhưng vẫn chưa có gì nổi bật khi điểm đánh giá tầm khoảng dưới 8 điểm.
- Đối với loại hình 2 (TV Miniseries): ngay từ những thập kỉ đầu, loại hình này đã được đánh giá rất cao (có khi được đánh giá lên đến 8.6 vào thập kỉ 1970). Tuy nhiên càng về sau, loại hình này có xu hướng được đánh giá thấp dần, nhưng đây vẫn là loại hình được đánh giá cao nhất trong toàn bộ khoảng thời gian được khảo sát.
- Đối với loại hình 3 (TV Series): đây cũng là một loại hình được đánh giá khá cao, tuy nhiên có thể không cao bằng loại hình TV Miniseries trong cuộc khảo sát. Loại hình này được đánh giá cao nhất là vào những năm 1950 đến những năm 1960.

2. Mối quan hệ giữa số lượt bình chọn (votes) và điểm đánh giá trung bình (average rating) thay đổi theo năm phát hành (release year):

2.1. Số lượt đánh giá (votes) của từng loại phim (types) theo từng thập kỷ của năm phát hành (từ 1910 đến 2020)



Hình 9. Biểu đồ cột chồng thể hiện tổng số lượt vote của từng loại hình phim theo từng thập kỷ.

Lý do chọn dạng biểu đồ: Biểu đồ cột chồng được sử dụng với nội dung này là phù hợp vì nó vừa thể hiện được tổng số lượng vote qua từng thập kỷ, đồng thời nó cũng cho biết xu hướng của từng loại phim dựa vào màu sắc phân biệt. Nhờ vậy mà ta dễ dàng so sánh tỉ lệ giữa các loại phim.

Nhận xét:

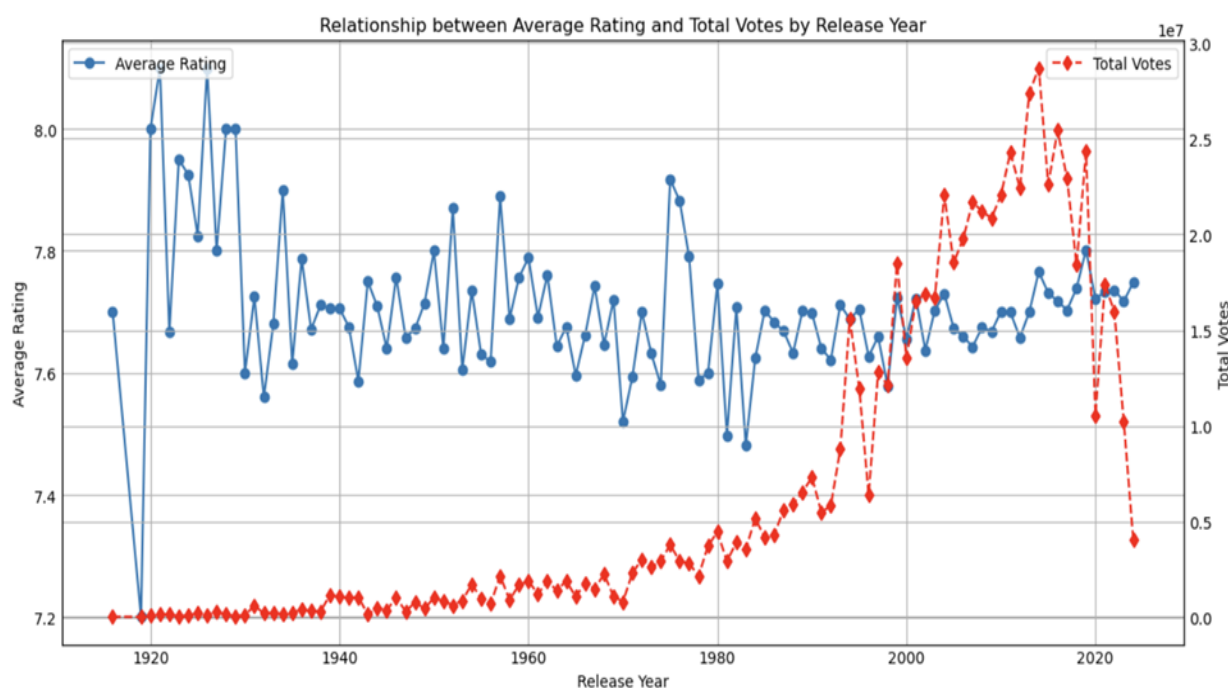
- Từ những năm 1990 trở đi, tổng số lượt bình chọn cho các loại hình giải trí tăng lên đáng kể, đặc biệt là từ thập kỷ 2000 và 2010. Phản ánh sự phát triển vượt bậc của ngành giải trí và sự quan tâm lớn của khán giả đối với các loại hình phim.
- Phim điện ảnh (movie) chiếm tỷ lệ lớn nhất trong tổng số lượt bình chọn, đặc biệt là trong các thập kỷ gần đây (2000 và 2010). Điều này cho thấy rằng dù có sự xuất hiện của các loại hình khác như phim truyền hình và miniseries, phim điện ảnh vẫn giữ được sự phổ biến và thu hút lượng bình chọn lớn từ người xem.
- Phim truyền hình (tvSeries) bắt đầu có số lượt bình chọn đáng kể từ những năm 2000, và con số này tiếp tục tăng trong thập kỷ 2010. Sự gia tăng này có thể liên quan đến sự phát triển của các nền tảng phát trực tuyến và xu hướng xem phim truyền hình dài tập.

- Miniseries có sự xuất hiện trong lượt bình chọn nhưng với số lượng nhỏ hơn nhiều so với phim điện ảnh và phim truyền hình. Tuy nhiên, chúng vẫn có sự đóng góp nhất định, đặc biệt trong thập kỷ 2010 và 2020.

- Tổng số lượt bình chọn giảm so với thập kỷ 2010. Điều này có thể do dữ liệu của thập kỷ này chưa hoàn chỉnh hoặc do sự ảnh hưởng của đại dịch COVID -19, dẫn đến giảm sản xuất phim và số lượng bình chọn từ khán giả.

Kết luận: Biểu đồ cho thấy sự tăng trưởng liên tục trong tổng số lượt bình chọn của khán giả qua các thập kỷ, đặc biệt là từ những năm 2000 trở đi. Phim điện ảnh vẫn là loại hình phổ biến nhất, trong khi phim truyền hình và miniseries cũng dần thu hút sự quan tâm của khán giả, đặc biệt trong những năm gần đây.

2.2. *Mối quan hệ giữa mức điểm rating và tổng số bình chọn theo những năm phát hành được chia theo thập kỷ:*



Hình 10. Biểu đồ đường kép thể hiện mối quan hệ giữa mức điểm rating và tổng số vote theo từng thập kỷ.

Lý do chọn dạng biểu đồ: Việc sử dụng biểu đồ đường kép (dual line chart) trong việc biểu diễn nội dung mối quan hệ giữa mức điểm rating và tổng số vote theo từng thập kỷ là phù hợp vì nó thể hiện rõ mối quan hệ giữa hai chỉ số riêng biệt theo thời gian mà vẫn thể hiện được sự tương quan của hai chỉ số này.

Nhận xét:

- Điểm đánh giá trung bình (đường màu xanh dương):

- Trong những năm đầu thế kỷ 20, điểm đánh giá trung bình có sự dao động khá lớn, với những đỉnh cao và thấp rõ rệt.
- Từ khoảng năm 1940 đến 1980, điểm đánh giá trung bình có xu hướng ổn định quanh mức 7.5 và ít dao động hơn.
- Sau năm 1980, điểm đánh giá trung bình bắt đầu có sự tăng nhẹ và duy trì ổn định ở mức cao hơn.
- Tổng số lượt bình chọn (đường màu đỏ):
- Trước năm 1990, tổng số lượt bình chọn cho mỗi năm khá thấp và ổn định.
- Từ năm 2000 trở đi, tổng số lượt bình chọn tăng mạnh, đạt đỉnh vào những năm sau 2010. Điều này có thể liên quan đến sự phát triển của Internet và các nền tảng đánh giá trực tuyến như IMDb, nơi mà người xem có thể dễ dàng tham gia đánh giá phim hơn.
- Đặc biệt, từ năm 2020 trở đi, tổng số lượt bình chọn giảm mạnh. Điều này có thể do sự ảnh hưởng của đại dịch COVID-19, khi mà nhiều bộ phim bị trì hoãn hoặc ra mắt ít hơn, dẫn đến lượng bình chọn giảm sút.

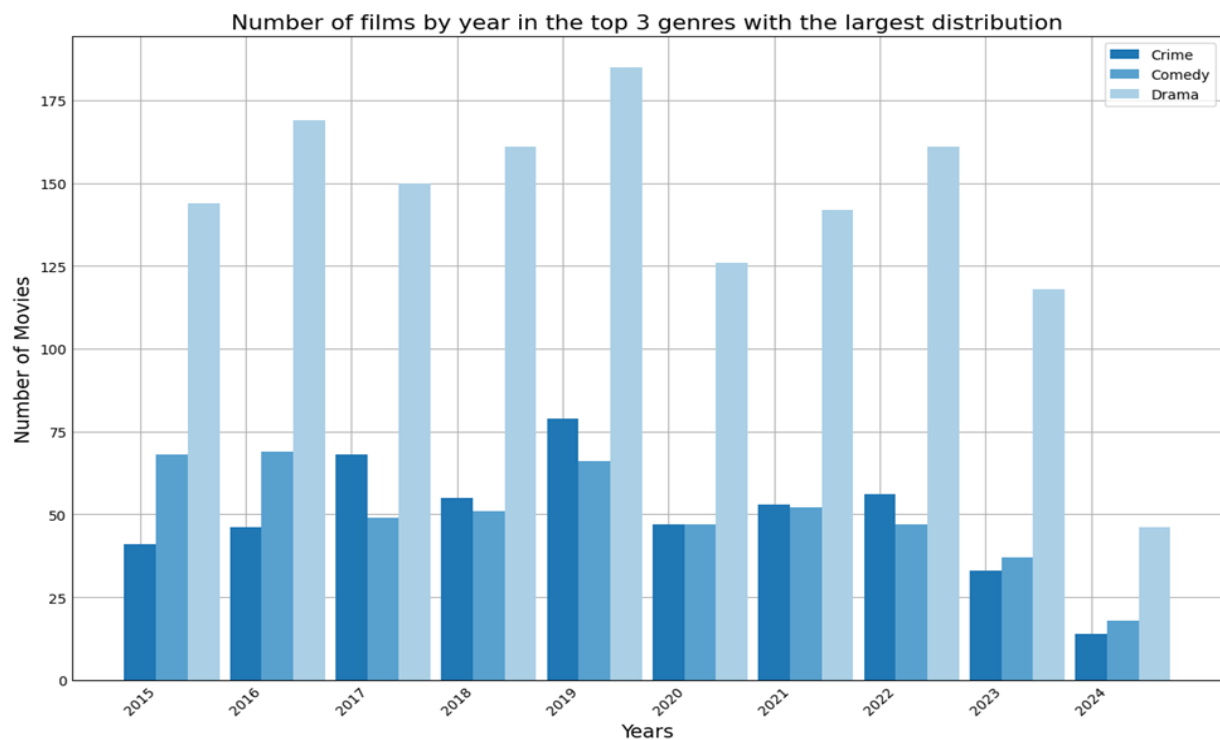
Kết luận: Biểu đồ cho thấy một xu hướng là mặc dù điểm đánh giá trung bình không thay đổi nhiều theo thời gian, tổng số lượt bình chọn đã tăng đáng kể trong những năm gần đây, đặc biệt là từ năm 2000 trở đi.

V. MỘT SỐ GÓC NHÌN CẬN HƠN:

1. Top 3 thể loại phim (genres) được sản xuất nhiều nhất:

- Top 3 thể loại phim được sản xuất nhiều nhất theo thứ tự giảm dần lần lượt là Drama, Comedy và Crime. Thể loại phim drama lên tới con số 4180 bộ, bỏ xa Comedy và Drama với chỉ 1794 và 1382 bộ (biểu đồ phân bố của các thể loại phim, phần II mục 1.1). Có thể thấy nhu cầu của khán giả đa số là thích xem những bộ phim nhiều tình tiết gay cấn, hài hước.

- Chúng ta sẽ đi tìm hiểu xem số lượng bộ phim được sản xuất trong 10 năm gần nhất (2015-2024) của 3 ông lớn này thông qua biểu đồ dưới đây:

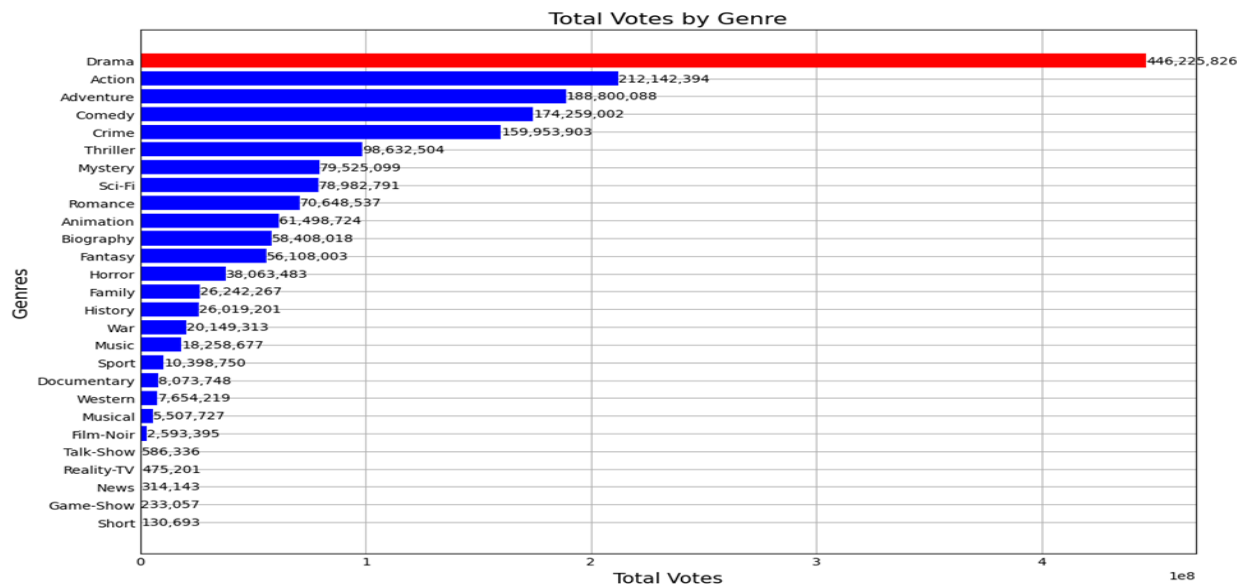


Hình 11. Biểu đồ cột gom nhóm thể hiện số lượng phim của 3 thể loại Crime, Comedy và Drama trong 10 năm gần đây nhất (2015 –2024)

Lý do chọn dạng biểu đồ: Việc sử dụng biểu đồ cột có gom nhóm (group bar chart) cho nội dung này là phù hợp vì với màu sắc riêng biệt, ta có thể so sánh số lượng phim thuộc từng thể loại (crime, comedy, drama) một cách rõ ràng ở từng thập kỉ, đồng thời vẫn có góc nhìn tổng quan đối với xu hướng tăng giảm của từng thể loại phim trong suốt thời gian phát hành được ghi nhận.

- Có thể thấy thể loại phim Drama rất được ưa chuộng phát triển tới từ các nhà sản xuất phim.

- Để biết chắc rằng số lượng có đi kèm với chất lượng không, chúng ta cùng xem biểu đồ thể hiện các thể loại phim có số lượng votes cao nhất từ trước đến nay.

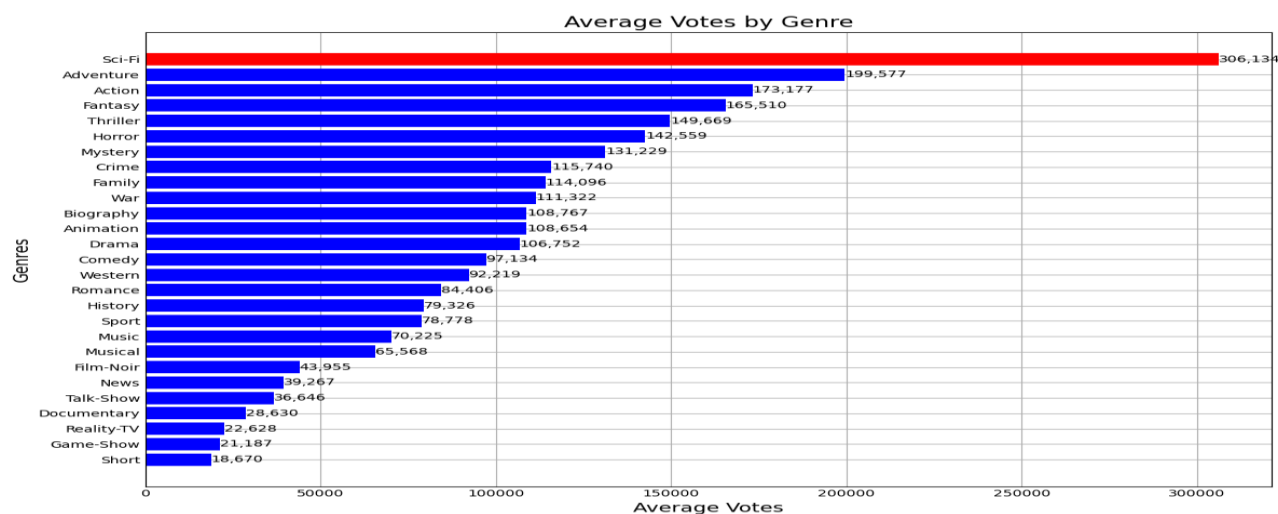


Hình 12. Biểu đồ thanh ngang thể hiện tổng số vote của các thể loại phim.

Lý do chọn dạng biểu đồ: Việc sử dụng biểu đồ thanh ngang đối với nội dung so sánh số lượng vote của từng thể loại phim là phù hợp vì biểu đồ thanh ngang biểu diễn dữ liệu được nhiều hơn (đặc biệt đối với dữ liệu có nhiều thuộc tính). Đồng thời, người dùng vẫn có thể so sánh được các thể loại phim một cách tổng quan nhất.

- Theo biểu đồ này ta thấy 3 thể loại phim có lượt vote nhiều nhất đứng đầu là Drama, tiếp theo là Action và cuối cùng là Adventure.

- Biểu đồ này thật ra chưa chắc rằng drama là bộ phim được yêu thích nhất, chúng ta cùng xem tiếp biểu đồ thể hiện số lượng vote trung bình cho 1 bộ phim của từng thể loại.



Hình 13. Biểu đồ thanh ngang thể hiện số lượng vote trung bình của các thể loại phim.

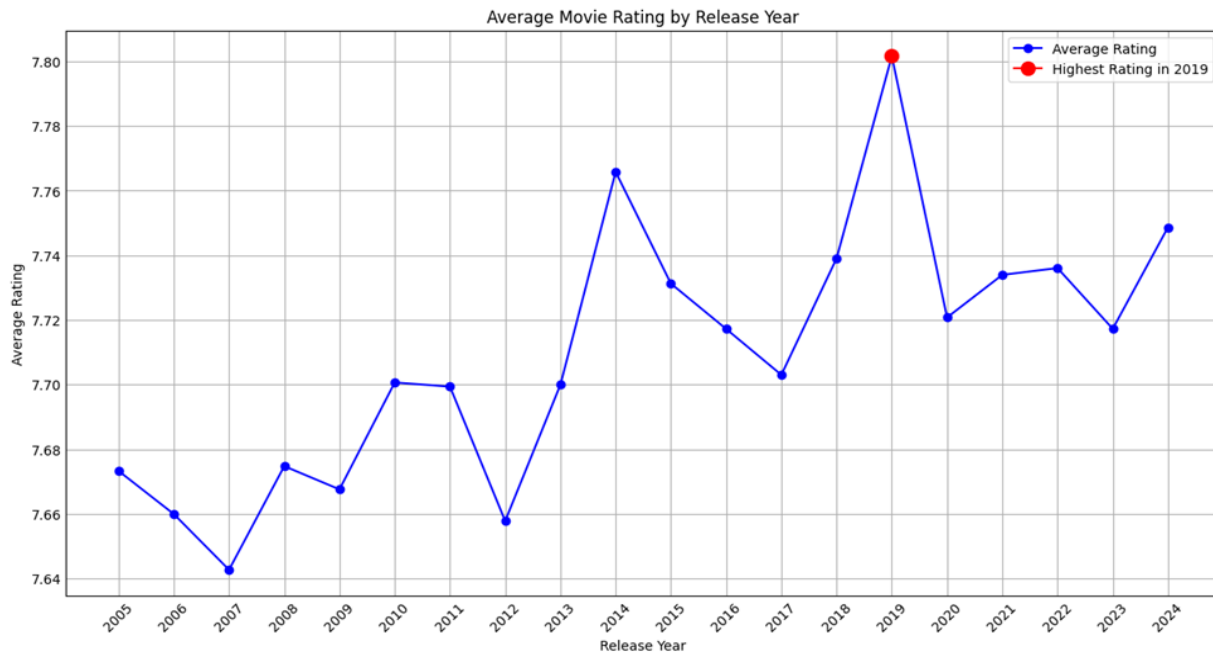
- Tới đây chúng ta có thể thấy rõ rằng, Sci-Fi mới là loại phim được yêu thích nhất với hơn 306,000 lượt vote trung bình trên 1 bộ phim thuộc thể loại này, thiết lập một khoảng cách rất xa so với các đối thủ ở vị trí số 2 và 3 là Adventure và Action.

Nhật xét:

- Mặc dù số lượng bộ phim Sci-Fi là khá thấp (258 bộ, dựa theo biểu đồ phân phối các thể loại ở mục 3.2) nhưng lại nhận về số lượng vote rất cao (gần 79,000,000 lượt vote), điều này cho ta thấy rằng chất lượng các bộ phim Sci-Fi đang đạt mức cao và thu hút người xem.

- Adventure và Action vẫn giữ được vị thế của mình, riêng Drama thì số lượng nhiều nhưng chất lượng thì chưa đáp ứng yêu cầu.

2. Mức điểm đánh giá trung bình của tất cả các bộ phim trong hai thập kỉ gần đây:



Hình 14. Biểu đồ đường thể hiện mức điểm đánh giá trung bình của tất cả các bộ phim trong hai thập kỉ gần đây.

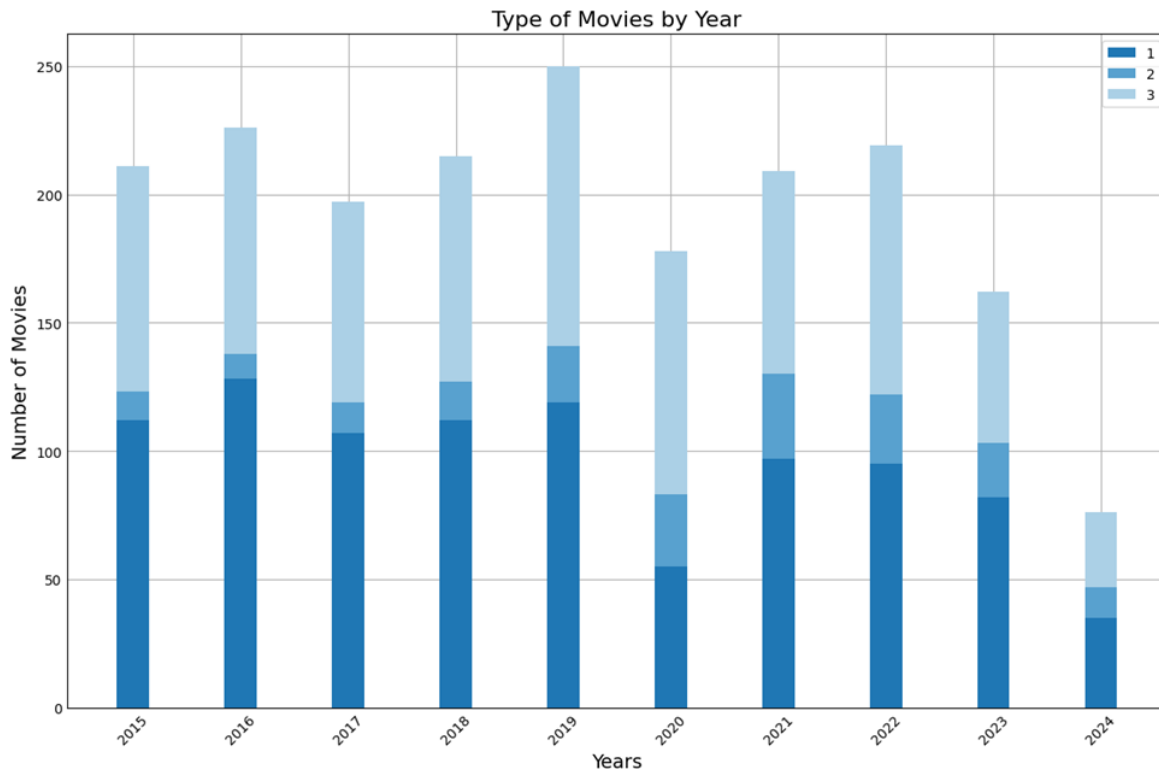
Lý do chọn dạng biểu đồ: Việc sử dụng biểu đồ đường trong nội dung về mức điểm đánh giá trung bình của tất cả các bộ phim trong hai thập kỉ gần đây là phù hợp vì biểu đồ đường thể hiện được xu hướng của mức điểm trung bình được đánh giá nhiều nhất trong từng năm. Việc này giúp người đọc có thể dễ dàng nắm được xu hướng của mức điểm đánh giá và dễ dàng so sánh được các mức điểm này qua từng năm.

Nhận xét:

- Nhìn chung, mức điểm đánh giá trung bình tăng dần qua từng năm đến năm 2019, đây là một tín hiệu tốt.
- Năm 2019 cũng là năm mà mức điểm trung bình đạt đỉnh điểm trong 10 năm qua với 7.8 điểm có thể do nhu cầu cao trong thời kỳ đại dịch.
- Kể từ năm 2019 đang có xu hướng giảm dần. Chất lượng phim khả năng đang bị mất dần đi.

Giải pháp: Cần cải thiện chất lượng các bộ phim thay vì số lượng.

Để cụ thể hơn, ta tìm hiểu về phân bố của các loại hình phim trong 10 năm gần đây nhất (2015 – 2024) thông qua biểu đồ cột chồng sau:



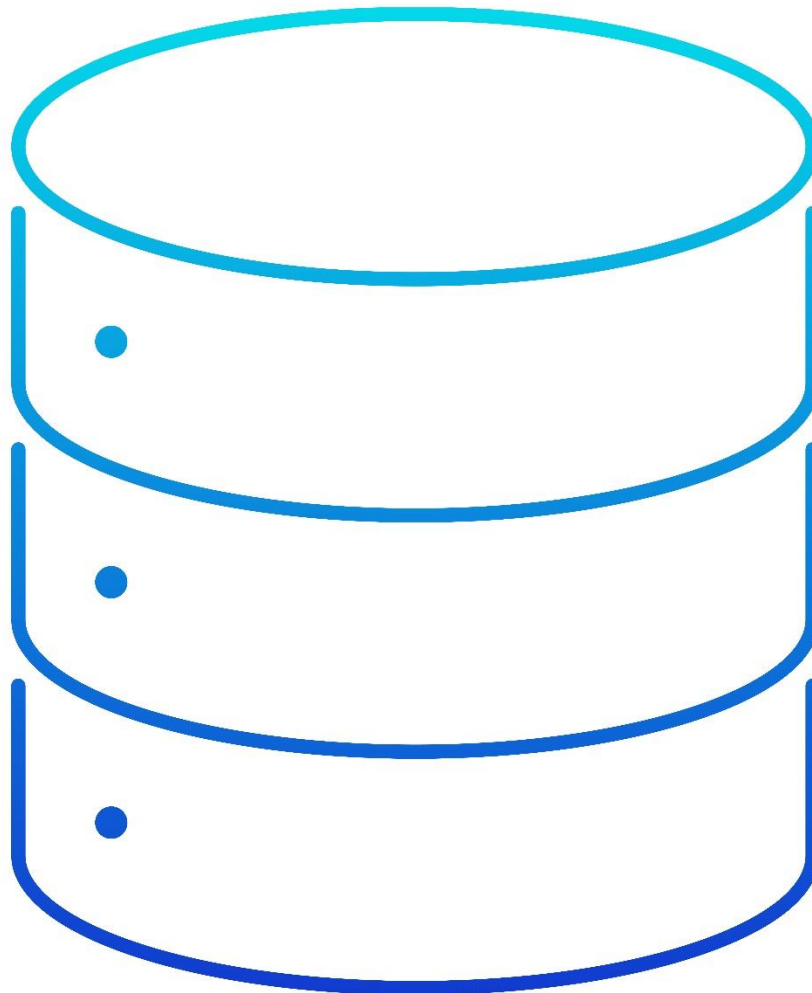
Hình 15. Biểu đồ cột chồng thể hiện sự phân bố của các loại hình phim trong 10 năm gần đây nhất (2015 – 2024)

Lý do chọn dạng biểu đồ: Biểu đồ cột chồng phù hợp cho nội dung phân bố của các loại hình trong 10 năm gần đây nhất (2015 – 2024) vì nó có thể thể hiện sự phân bố từng loại phim qua từng năm cùng một lúc mà người dùng vẫn có góc nhìn tổng quan về sự thay đổi của từng loại phim qua từng năm.

Nhận xét:

- Nhìn chung, xu hướng sản xuất phim ổn định từ 2015-2018.
- Năm 2019 số lượng phim được sản xuất đạt ngưỡng cao nhất với 250 bộ phim, trong đó phim ngắn chiếm hơn 47% theo sau là phim truyền hình dài tập với gần 44%. Sở dĩ vậy cũng là do đây là để phục vụ nhu cầu cho người dân toàn cầu khi đây là thời điểm dịch Covid và người dân ít ra đường.
- Sau đó, số lượng phim có xu hướng giảm dần. Trong năm 2024, tính đến thời điểm hiện tại thì với số lượng 76 bộ phim là khá thấp so với những năm trước, năm nay có khả năng số lượng phim đạt mức thấp nhất trong 10 năm trở lại đây.

Section 5: *Hệ thống gợi ý dựa trên Mức đánh giá, Thẻ loại, Số lượng vote, và Loại hình phim.*



Hệ thống nhận gợi ý theo Tựa phim.

I. Giới thiệu mô hình:

Mô hình hệ thống tư vấn gợi ý các bộ phim khác từ một bộ phim do người dùng nhập vào thông qua mối liên hệ của bộ phim đó và các bộ phim khác trong các yếu tố:

1. Thể loại công chiếu: movies (1), TVSeries (2), MiniTVSeries (3).
2. Thể loại phim (yếu tố phim): Kinh dị, hài hước, hành động, kịch tính,...
3. Rating: đánh giá của người dùng trên thang điểm 10.
4. Numvotes: số lượng đánh giá.
5. ReleaseYear: năm phát hành.

II. Các thư viện sử dụng:

Vì thời gian có hạn cho nên em chỉ xây dựng nên khung của mô hình, sau đó sử dụng các thư viện với nhiều hàm cài đặt sẵn cho có tác vụ này.

Trong đó gồm:

Thư viện scikit-learn:

1. sklearn.preprocessing: sử dụng MinMaxScaler.

Mục đích: dùng để chuẩn hóa các vector về trong đoạn $[0,1]$ nhằm đảm bảo rằng các đặc trưng có trọng số tương đương nhau khi tính toán độ tương đồng (hạn chế bias), ngăn chặn việc chiếm ưu thế quá lớn của một yếu tố nào đó.

2. sklearn.metrics.pairwise: sử dụng cosine_similarity.

Mục đích: dùng để tính độ tương đồng cosine cho từng cặp phim. Trong đó độ tương đồng giữa hai vector A, B được tính theo công thức sau:

$$Similarity = \frac{A.B}{||A||.||B||}$$

Thư viện copy:

1. deepcopy:

Mục đích: dùng để tạo ra một bản sao theo kiểu deepcopy cho một đối tượng.

Thư viện Numpy:

1. log1p: - quan trọng.

Mục đích: Yếu tố này vốn thường có giá trị lớn và phân phối không đồng đều –do đó log1p sẽ giúp giảm bớt ảnh hưởng của các giá trị cực đại. Điều này làm cho dữ liệu trở nên dễ quản lý hơn bằng cách "làm phẳng" các giá trị lớn, biến đổi chúng để giảm thiểu sự thiên lệch.

Ngoài ra việc này còn giúp chuẩn hóa về dạng phân phối chuẩn tốt hơn, từ đó áp dụng các tính toán cho độ tương đồng được nhanh chóng và chính xác.

Ví dụ, với giá trị `numVotes = 100,000` và `numVotes = 10`, sau khi áp dụng `log1p`, giá trị sẽ trở nên gần nhau hơn và không quá lệch:

- $\log1p(100,000) \approx 11.51$
- $\log1p(10) \approx 2.4$

2. Một số hàm tính toán liên quan đến ma trận, mảng,...

III. Cài đặt thuật toán:

1. Class Recommender:

Mô tả: class chứa các thuộc tính, cài đặt hàm dùng để tạo nên một recommendation system.

2. Class Recommender > def Prepare_feature():

Mô tả: Đây là hàm dùng để chuẩn bị các yếu tố cần thiết cho một recommendation system và tính toán ma trận tương đồng.

Bước 1: Chuẩn hóa vector 'rating', 'numVotes', 'releaseYear' về trong đoạn [0,1] bằng hàm `fit_transform`.

Bước 1.2: Lưu ý, ở vector `numVotes` ta sẽ sử dụng `numpy` để chuẩn hóa thuộc tính này bằng `log1p` trước khi chuẩn hóa bằng `fit_transform` của thư viện `scikit-learn`.

Bước 2: Tạo các thuộc tính mới kèm theo từng trọng số cho các thuộc tính (cột) đã được chuẩn hóa.

Lưu ý: Ở đây mô hình của em sẽ chú trọng vào phần `rating` và `genres` cho nên 2 thuộc tính này sẽ có trọng số cao nhất, ngược lại vì `numVotes` chênh lệch tương đối lớn giữa các phim cho nên em sẽ cho thuộc tính này trọng số thấp nhất.

Bước 3: Tính toán ma trận tương đồng thông qua thư viện cho các thuộc tính.

3. Class Recommender > def get_recommendation():

Ý tưởng: dựa trên ma trận tương đồng ta tìm ra các phim có độ tương đồng (similarity score) với phim cần tìm kiếm cao nhất.

Bước 1: Tìm ra vị trí của phim hiện tại cần tìm kiếm.

Bước 2: Lấy ra mức điểm tương đồng của phim hiện tại.

Bước 3: Sắp xếp thứ tự của ma trận tương đồng này để lấy ra các vị trí có độ tương đồng gần với mức điểm của phim hiện tại.

Bước 4: Thêm vào cột Similarity Score.

Bước 5: Trả về một Dataframe gồm các phim có similarity score cao nhất được sắp xếp theo thứ tự giảm dần. Số lượng phim này sẽ do người dùng nhập vào.

IV. Recommendation System GUI:

1. Cửa sổ chính.

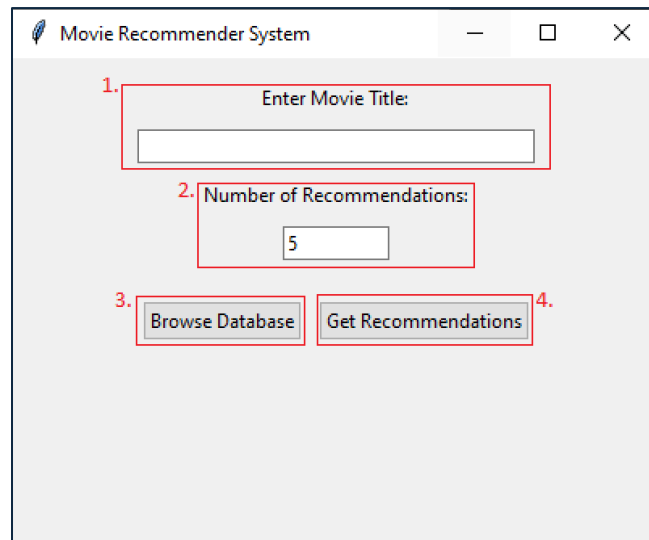
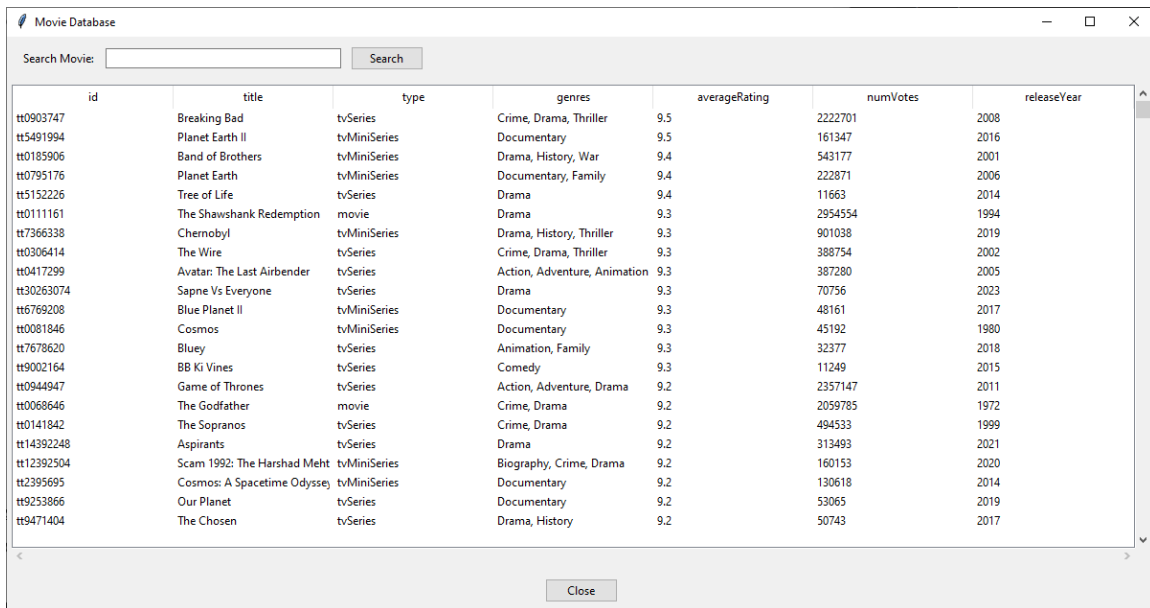


Figure 10. Giao diện chính.

- a) **Khung nhập xuất:** nhập tên của bộ phim cần được gợi ý.
- b) **Khung số lượng:** nhập số lượng tên phim sẽ được gợi ý. Ví dụ nhập 10 thì sẽ gợi ý ra 10 bộ phim.
- c) **Browse Database:** dùng để xem lại trong database có những phim gì.
- d) **Get Recommendation:** xem kết quả gợi ý.

2. Cửa sổ database:

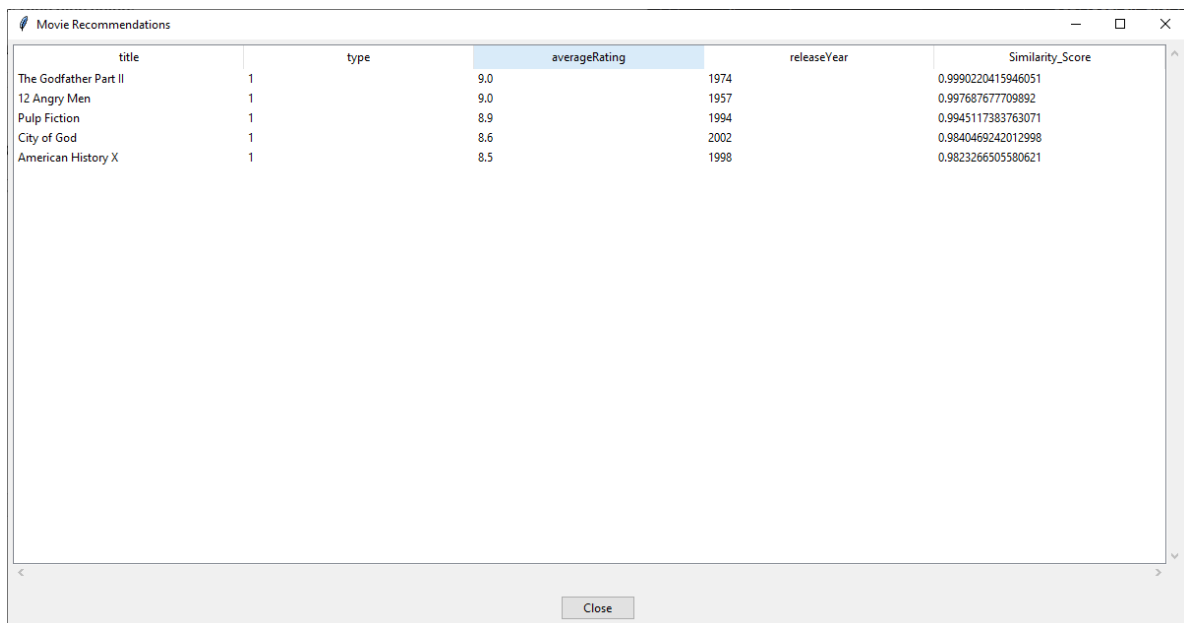


id	title	type	genres	averageRating	numVotes	releaseYear
tt0903747	Breaking Bad	tvSeries	Crime, Drama, Thriller	9.5	2222701	2008
tt5491994	Planet Earth II	tvMiniSeries	Documentary	9.5	161347	2016
tt0185906	Band of Brothers	tvMiniSeries	Drama, History, War	9.4	543177	2001
tt0795176	Planet Earth	tvMiniSeries	Documentary, Family	9.4	222871	2006
tt5152226	Tree of Life	tvSeries	Drama	9.4	11663	2014
tt0111161	The Shawshank Redemption	movie	Drama	9.3	2954554	1994
tt7366338	Chernobyl	tvMiniSeries	Drama, History, Thriller	9.3	901038	2019
tt0306414	The Wire	tvSeries	Crime, Drama, Thriller	9.3	388754	2002
tt0417299	Avatar: The Last Airbender	tvSeries	Action, Adventure, Animation	9.3	387280	2005
tt30263074	Sapne Vs Everyone	tvSeries	Drama	9.3	70756	2023
tt6769208	Blue Planet II	tvMiniSeries	Documentary	9.3	48161	2017
tt0081846	Cosmos	tvMiniSeries	Documentary	9.3	45192	1980
tt7678620	Bluey	tvSeries	Animation, Family	9.3	32377	2018
tt9002164	BB Ki Vines	tvSeries	Comedy	9.3	11249	2015
tt0944947	Game of Thrones	tvSeries	Action, Adventure, Drama	9.2	2357147	2011
tt0068646	The Godfather	movie	Crime, Drama	9.2	2059785	1972
tt0141842	The Sopranos	tvSeries	Crime, Drama	9.2	494533	1999
tt14392248	Aspirants	tvSeries	Drama	9.2	313493	2021
tt12392504	Scam 1992: The Harshad Meht	tvMiniSeries	Biography, Crime, Drama	9.2	160153	2020
tt2395695	Cosmos: A Spacetime Odyssey	tvMiniSeries	Documentary	9.2	130618	2014
tt9253866	Our Planet	tvSeries	Documentary	9.2	53065	2019
tt9471404	The Chosen	tvSeries	Drama, History	9.2	50743	2017

Figure 11. Browser Database GUI.

Chức năng: Tìm kiếm hoặc xem thông tin phim theo nhu cầu.

3. Cửa sổ recommendation results:



title	type	averageRating	releaseYear	Similarity_Score
The Godfather Part II	1	9.0	1974	0.9990220415946051
12 Angry Men	1	9.0	1957	0.997687677709892
Pulp Fiction	1	8.9	1994	0.9945117383763071
City of God	1	8.6	2002	0.9840469242012998
American History X	1	8.5	1998	0.9823266505580621

Figure 12. Top 5 Phim liên quan đến 'The Godfather'.

Chức năng: Thể hiện các kết quả gợi ý.