FINAL PROJECT – NHÓM 7

Thành viên nhóm:
20120041 - Trần Kim Bảo
20120053 - Nguyễn Thành Đạt
20120071 - Nguyễn Thị Bích Hà
20120113 - Lê Nguyên Khang

IMDB – 1000 PHIM HAY NHẤT MỌI THỜI ĐẠI

IMDb là một trang cơ sở dữ liệu trực tuyến về điện ảnh thế giới, cung cấp thông tin về phim, diễn viên, đạo diễn, nhà làm phim,.. và tất cả những người, công ty trong lĩnh vực sản xuất phim, phim truyền hình và cả trò chơi video.

THU THẬP DỮ LIỆU

- Nhóm sử dụng **Scrapy** kết hợp với **css** và **xpath** để crawl dữ liệu từ web về.
- Nhóm chỉ lấy 1000 phim đầu danh sách để thực hiện khám phá.
- Tổng cộng nhóm crawl 20 cột dữ liệu, 20 cột này là các thông tin của một phim nằm trong bảng xếp hạng.
- Vì nhóm chia nhau ra crawl dữ liệu, nên khi crawl xong thì lưu vào 2 file Json khác nhau. Do đó khi đọc dữ liệu từ file Json lên, nhóm phải thực hiện merge dữ liệu lại thành một DataFrame duy nhất.

THU THẬP DỮ LIỆU

	STT	ID	Name	Published Year	Rated	Duration	Genres	Director	Writers	Stars	IMDb RATING	Budget (estimated)	Gross	Popularity	Votes
0	1.	tt0068646	The Godfather	1972	R	[2, , hours, , 55, , minutes]	[Crime, Drama]	[Francis Ford Coppola]	[Francis Ford Coppola, Mario Puzo]	[Al Pacino, James Caan, Marlon Brando]	9.2	$6,000,000	134,966,411	92	1849023
1	2.	tt0099685	Goodfellas	1990	R	[2, , hours, , 25, , minutes]	[Biography, Crime, Drama]	[Martin Scorsese]	[Martin Scorsese, Nicholas Pileggi]	[Joe Pesci, Ray Liotta, Robert De Niro]	8.7	$25,000,000	46,836,394	194	1156616
2	3.	tt0110912	Pulp Fiction	1994	R	[2, , hours, , 34, , minutes]	[Crime, Drama]	[Quentin Tarantino]	[Roger Avary, Quentin Tarantino]	[Uma Thurman, John Travolta, Samuel L. Jackson]	8.9	$8,000,000	107,928,762	115	2042809
3	4.	tt0114814	The Usual Suspects	1995	R	[1, , hour, , 46, , minutes]	[Crime, Drama, Mystery]	[Bryan Singer]	[Christopher McQuarrie]	[Chazz Palminteri, Gabriel Byrne, Kevin Spacey]	8.5	$6,000,000	23,341,568	425	1083153
4	5.	tt0078788	Apocalypse Now	1979	R	[2, , hours, , 27, , minutes]	[Drama, Mystery, War]	[Francis Ford Coppola]	[John Milius, Michael Herr, Francis Ford Coppola]	[Martin Sheen, Robert Duvall, Marlon Brando]	8.5	$31,500,000	83,471,511	307	666748

- Đây là Dataframe **film_info_df** sau khi merge lại. Gồm 20 cột và 1000 dòng dữ liệu.
- Mỗi một dòng là các thông tin của một phim.
- Mỗi một cột là một trường thông tin riêng của phim.
- Và không có dòng dữ liệu nào bị trùng lắp.

STT	Tên cột dữ liệu	Mô tả	Đơn vị
1	Top	Thể hiện thứ hạng của phim trên bảng xếp hạng	
2	ID	Mã định danh của phim	
3	Name	Tên của phim	
4	Published Year	Năm xuất bản phim	năm
5	Rated	Loại phim dựa trên Hệ thống phân loại phim của MPAA, để phân loại phim cho các đối tượng xem trẻ em, thanh thiếu niên và người lớn	
6	Duration	Thời lượng phim	giờ
7	Genres	Thể loại phim	
8	Director	Tên các đạo diễn sản xuất phim	
9	Writers	Tên các biên kịch phim	
10	Stars	Tên các diễn viên tham gia (ở cột này chỉ lấy tên các diễn viên được hiển thị trên trang web)	
11	IMDb RATING	Điểm IMDb được tính theo thang điểm 10, nghĩa là nếu điểm số càng cao thì chất lượng càng tốt	
12	Budget (estimated)	Kinh phí thực hiện phim (ước tính)	$
13	Gross	Doanh thu của phim tính đến thời điểm hiện tại	$
14	Popularity	Độ phổ biến của phim (thứ tự trên bảng xếp hạng) được tính đến tuần hiện tại	
15	Votes	Số lượt bình chọn của người xem trên trang IMDb dành cho phim	
16	User reviews	Số lượng nhận xét của người xem	
17	Critic reviews	Số lượng nhận xét của các nhà phê bình	
18	Meta score	Điểm trung bình của các nhà phê bình lấy từ bộ điểm số của Metacritic, với thang điểm là 100	
19	Wins	Số lượng giải thưởng và đề cử mà phim đạt được trong Giải thưởng Viện Hàn lâm, thường được biết đến với tên Giải Oscar	
20	URL	Địa chỉ truy cập vào trang web mô tả chi tiết của phim	

Ý NGHĨA CỤ THỂ CỦA MỖI CỘT

TIỀN XỬ LÝ DỮ LIỆU

- Nhóm chia các cột ra thành 2 loại: **Numeric** và **Categorical** để tiền xử lý.
- Mục đích chính là đưa các cột về đúng kiểu dữ liệu. Những cột kiểu dữ liệu multi, nhóm không xử lý trực tiếp ở phần này mà tự xử lý ở phần trả lời câu hỏi.

ĐƯA RA CÂU HỎI
ĐỂ KHÁM PHÁ DỮ LIỆU

Câu 1: Phân tích trong 10 năm để tìm ra thể loại phim được yêu thích trong mỗi khoảng thời gian. Tương quan giữa các thể loại (genre) với doanh thu, độ nổi tiếng, các thành tựu đạt được.

Ý nghĩa khi trả lời câu hỏi:

- Biết được thể loại phim nào được yêu thích nhất trong mỗi thập niên.
- Thể loại được yêu thích có phải chỉ tập trung vào một vài thập niên không, hay là gần như được yêu thích với mọi thập niên.
- Hiểu rõ được thể loại được yêu thích là do gì.

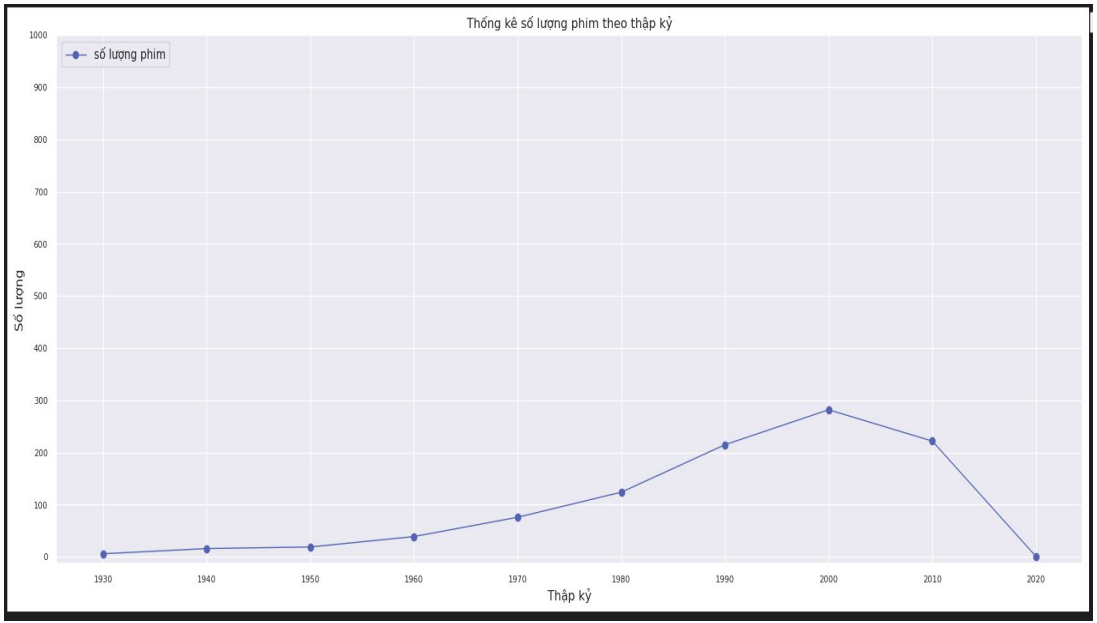
Nguồn cảm hứng đặt câu hỏi:

- Cảm hứng bắt nguồn từ câu hỏi thực tiễn rằng liệu bộ phim được ưa chuộng, dù có gây tiếng vang, nhưng đôi khi vẫn bị lỗ vốn hoặc không đạt giải thưởng vì một số lý do nào đó.Thể loại nào thường mang về những giải Oscar.
- Có phải những thể loại được yêu thích thường mang lại nhiều lợi nhuận hơn không?

Câu 1: Các bước thực hiện, mục đích thực hiện.

Bước	Nội dung	Mục đích
1	Thống kê số lượng phim theo thập niên	Biết số lượng của phim thay đổi theo từng thập niên như thế nào, có sự bất thường nào không.
2	Thống kê số lượng phim theo thể loại	Biết thể loại được yêu thích cho đến hiện tại. Là bước xử lý trước để nhận xét cho bước bước 3.
3	Tương quan giữa thể loại phim và thập niên	Khẳng định sự yêu thích của thể loại qua từng thập niên.
4	Tương quan với cột Gross và cột Wins	Trả lời câu thể loại được yêu thích thì sẽ mang nhiều doanh thu về nhất. Hay thể loại được yêu thích sẽ là những yếu tố đóng góp nhiều cho các giải thưởng của phim.

Vẽ biểu đồ thể hiện số lượng phim theo thập niên

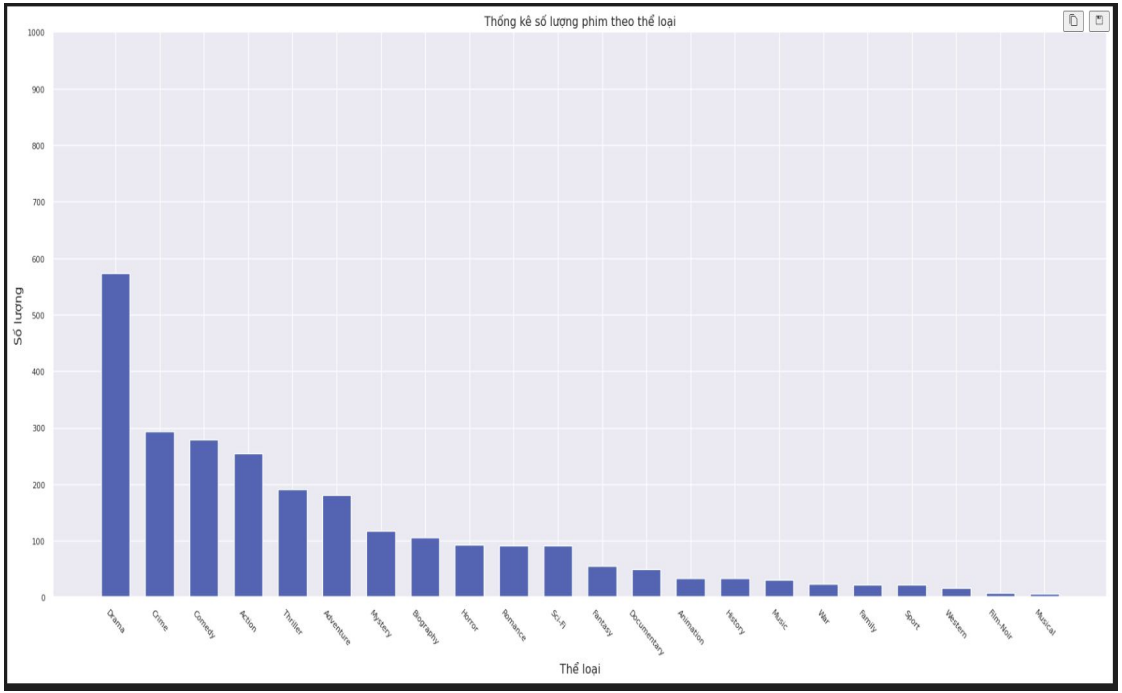


Cập nhật 12/11/2021

Nhận xét:

- Sự thay đổi bất thường của thập niên 2010 và thập niên 2020.

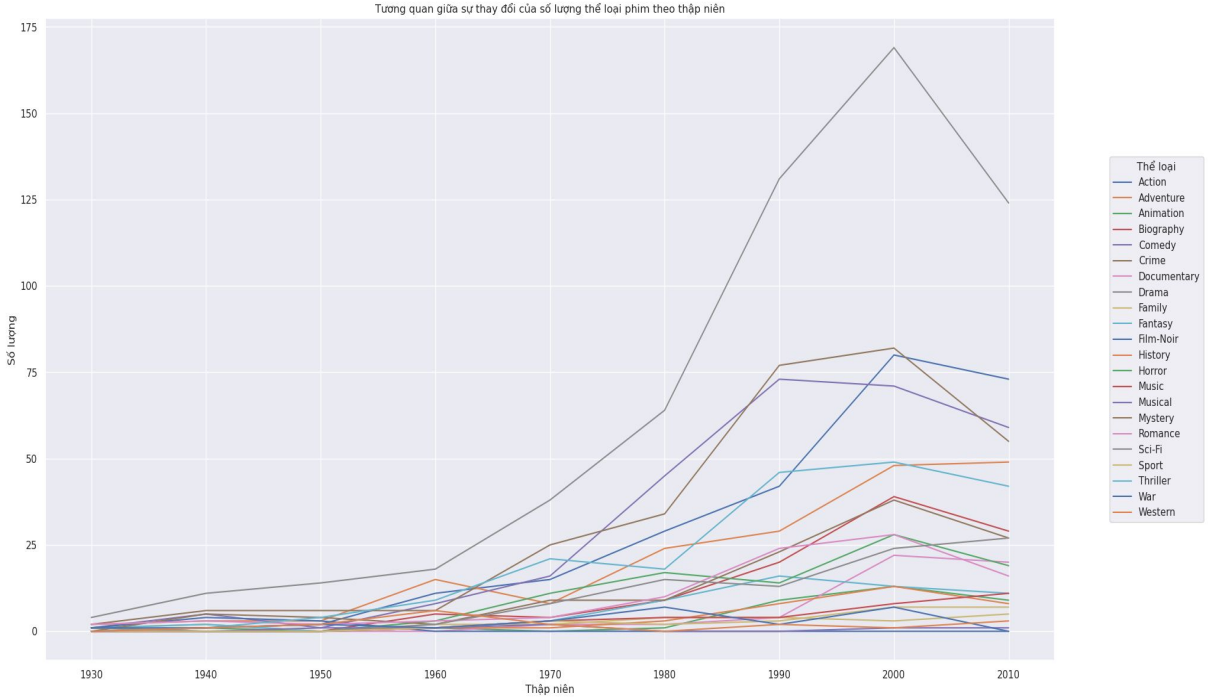
Bước 2: Thống kê số lượng phim theo thể loại



Nhận xét:

- Ta thấy được xu hướng các phim có thể loại Drama chiếm hơn 1 nửa số lượng phim và cũng là thể loại chiếm số lượng cao nhất.
- Tiếp đến là các thể loại phổ biến như: **tội phạm, hài kịch, hành động, giật gân, phiêu lưu** cũng là những thể loại chiếm đa số. Ở đây trừ hài kịch mục đích gây cười, có tính giải trí cao *(điều ai cũng cần)*, thì có thể thấy gần như hầu hết đều là những thể loại mạnh mẽ, nhịp độ phim nhanh gây kích thích, hứng thú cho người xem.
- Các thể loại còn lại chỉ chiếm từ 0 đến 10% số lượng phim trong danh sách.

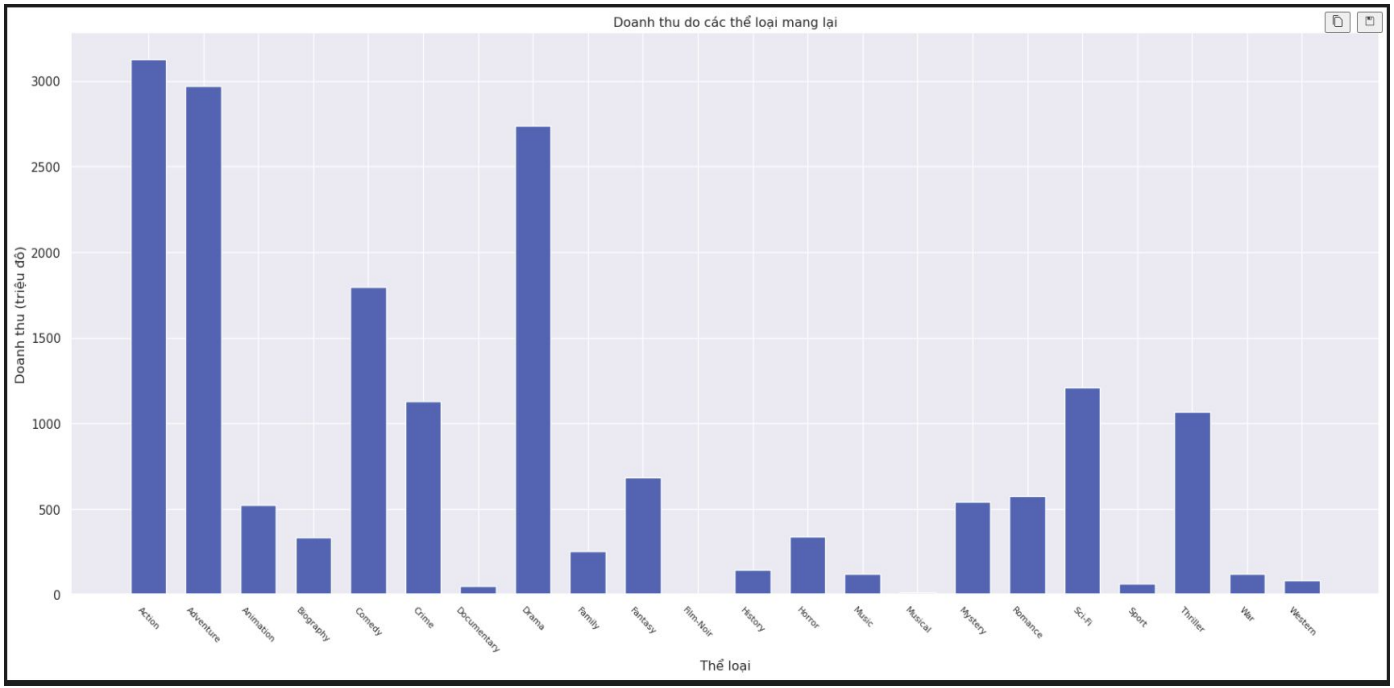
Bước 3: Tương quan giữa thể loại phim và thập niên



Nhận xét:

- Ta thấy rằng ở thể loại **Drama** ở mọi thập niên đều đứng đầu trong số lượng phim lọt top.
- Không có sự thay đổi các thể loại: **hành động, hài kịch, phiêu lưu, tội phạm.**
- Vì số lượng phim lọt top của thập niên 2010 giảm hơn 1/5 so với thập niên 2000, nên ta có thể thấy rằng chỉ vài thể loại tăng lên. Tuy rằng nhu cầu giải trí mỗi ngày càng cao, số lượng phim sản xuất cũng tăng lên và có nhiều sự khác biệt trong công nghệ làm phim.

Bước 4: Tương quan với những cột khác

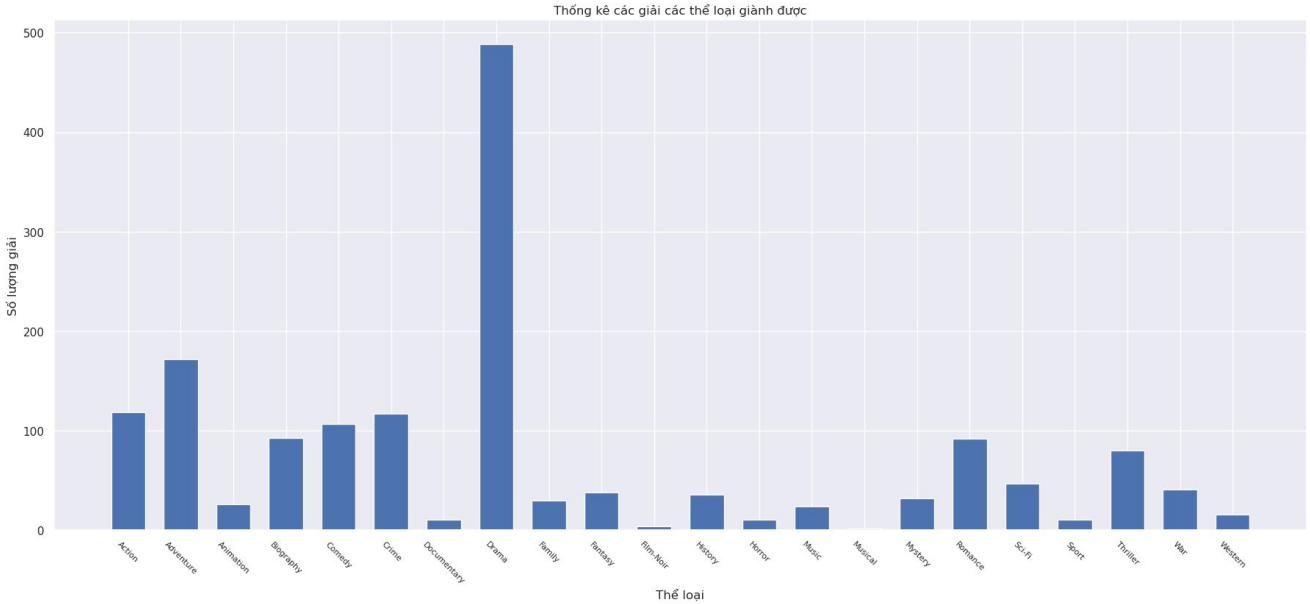


Biểu đồ giữa doanh thu và thể loại

Nhận xét:

- Ta thấy được doanh thu mang về của thể loại **Drama** vẫn không cao bằng 2 thể loại cùng đứng top là **hành động** và **phiêu lưu**.

Bước 4: Tương quan với những cột khác



Số giải thắng được của từng thể loại

Nhận xét:

- Nhìn vào biểu đồ trên ta đã hoàn toàn thấy sự khác biệt số lần thắng giải của thể loại **Chính kịch**, gấp hơn 2.5 lần so với các thể loại cùng đứng top trên.

Kết luận

- Thể loại được yêu thích nhất mọi thập niên: **Chính kịch.**
- Top các thể loại yêu thích kế tiếp *(không mang tính thứ tự - Vì mỗi thập niên có sự thay đổi qua lại vị trí giữa những thể loại đó)*: **Phiêu lưu, Tội phạm, Hài kịch, Hành động.**
- Thể loại mang về nhiều giải thưởng nhất: **Chính kịch.**

Câu 2: Tìm hiểu xem top 5 đạo diễn xuất hiện nhiều nhất trong Dataset. Xuất hiện nhiều đồng nghĩa với việc họ sở hữu số lượng phim áp đảo những đạo diễn còn lại. Liệu rằng số phim mà các đạo diễn này sản xuất ra có thực sự là một bộ phim chất lượng (đánh giá dựa trên các nhà phê bình có chuyên môn - Metascore).

Ý nghĩa khi trả lời câu hỏi:

- Ta sẽ biết được về các đạo diễn gạo cội, những người tạo ra những bộ phim để đời.
- Ngoài ra, còn biết thêm việc, liệu những đạo diễn này có thật sự sản xuất ra những bộ phim chất lượng hay không.

Nguồn cảm hứng đặt câu hỏi:

- Cảm hứng bắt nguồn từ sự tò mò của bản thân, vì em nghĩ, nếu các đạo diễn này tham gia vào nhiều phim thì sẽ có hai hướng suy nghĩ. Một là, đạo diễn này thật sự là một đạo diễn tài năng và việc sản xuất ra được những bộ phim "chất lượng" lọt top là điều hiển nhiên. Hai là, đạo diễn này có được nhiều nguồn đầu tư nên mời được những diễn viên nổi tiếng, gạo cội nhờ đó làm phim được quan tâm nhiều hơn. Do đó tần suất xuất hiện cũng nhiều hơn. Và để làm rõ điều đó. Em mới đặt ra câu hỏi này.

Tìm hiểu thêm về Metascore:

- Metacritic thu thập đánh giá của các nhà phê bình phim và các trang đánh giá phim sau đó cho điểm các đánh giá này từ 0 đến 100 điểm.

- Ở Metacritic, các tác phẩm được đánh giá hiển thị theo 3 màu: đỏ - vàng – xanh lá. Trong đó, màu đỏ dành cho các phim được cho là không hay. Màu vàng dành cho phim ở mức trung bình (có nhiều sạn, có điểm chưa hợp lý, diễn xuất chưa tốt,...). Màu xanh là phim được cho là hay.

Màu sắc	Đánh giá	Phim/Chương trình TV/Nhạc
Xanh lá	Universal acclaim (Hoan nghênh nhiệt liệt)	81-100
Xanh lá	Generally favorable (Nhìn chung là ý kiến tán thành)	61-80
Vàng	Mixed or average (Hỗn tạp hoặc trung bình)	40-60
Đỏ	Generally unfavorable (Nhìn chung là ý kiến không tán thành)	20-39
Đỏ	Overwhelming dislike (Hoàn toàn không thích)	0-19

Bước 1: Tìm ra 5 đạo diễn xuất hiện nhiều nhất trong tập dữ liệu cùng với các phim của các đạo diễn này sản xuất.

	Director	Name
0	Martin Scorsese	[Goodfellas, Taxi Driver, Raging Bull, The Departed, Casino, The Color of Money, The Wolf of Wal...
1	Steven Spielberg	[Schindler's List, Jaws, Indiana Jones and the Raiders of the Lost Ark, Jurassic Park, Saving Pr...
2	Joel Coen	[Fargo, The Big Lebowski, Miller's Crossing, No Country for Old Men, O Brother, Where Art Thou?,...
3	Ethan Coen	[Fargo, The Big Lebowski, Miller's Crossing, No Country for Old Men, O Brother, Where Art Thou?,...
4	Christopher Nolan	[The Dark Knight, The Prestige, Batman Begins, Memento, Inception, Insomnia, The Dark Knight Ris...

Bước 2: Tiến hành phân tích theo từng đạo diễn. Ta sẽ tách các phim từ dạng list ra và bổ sung thêm cột Metascore của mỗi phim.

	Name	Meta score
0	Goodfellas	90.0
1	Taxi Driver	94.0
2	Raging Bull	89.0
3	The Departed	85.0
4	Casino	73.0
5	The Color of Money	77.0
6	The Wolf of Wall Street	75.0
7	Hugo	83.0
8	The Irishman	94.0
9	The Aviator	77.0
10	Gangs of New York	72.0
11	The Last Waltz	88.0
12	Shutter Island	63.0
13	The King of Comedy	73.0
14	Bringing Out the Dead	70.0
15	After Hours	90.0
16	Cape Fear	73.0
17	Mean Streets	96.0

Martin Scorsese

	Name	Meta score
0	Schindler's List	94.0
1	Jaws	87.0
2	Indiana Jones and the Raiders of the Lost Ark	85.0
3	Jurassic Park	68.0
4	Saving Private Ryan	91.0
5	Indiana Jones and the Last Crusade	65.0
6	Minority Report	80.0
7	Close Encounters of the Third Kind	90.0
8	Catch Me If You Can	75.0
9	A.I. Artificial Intelligence	65.0
10	Munich	74.0
11	War of the Worlds	73.0
12	E.T. the Extra-Terrestrial	91.0
13	Empire of the Sun	62.0
14	Indiana Jones and the Temple of Doom	57.0
15	The Adventures of Tintin	68.0
16	The Post	83.0

Steven Spielberg

Bước 3: Ta sẽ thêm một cột **Color** vào để hỗ trợ cho phần visualize phía sau. Cột Color này mang nghĩa phân loại điểm Metascore dựa trên màu sắc mà ta đã đề cập ở **phần tìm hiểu**.

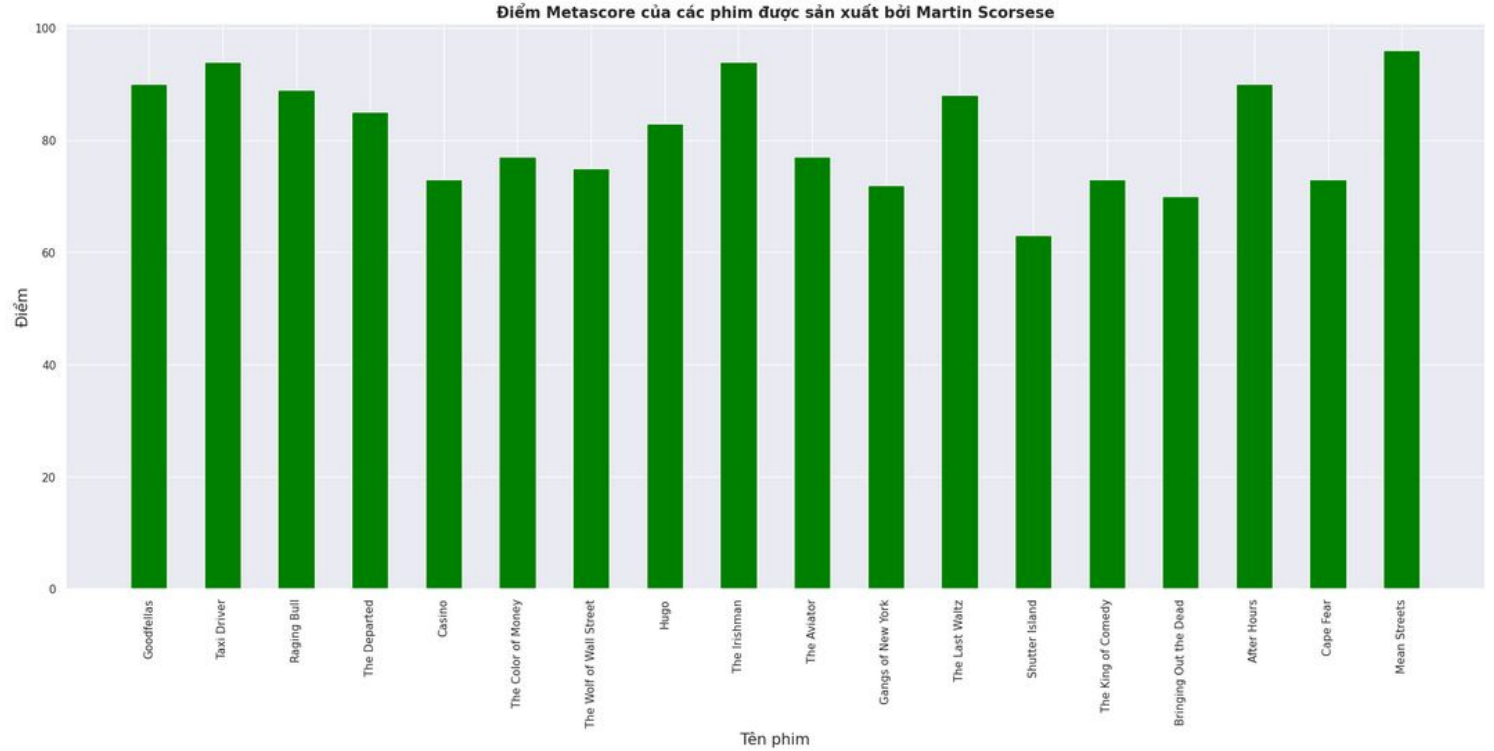
Martin Scorsese

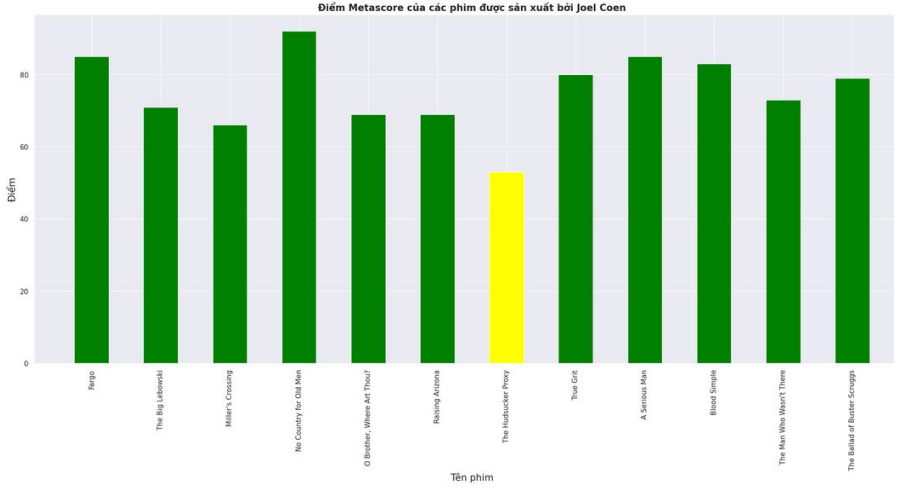
	Name	Meta score	Color
0	Goodfellas	90.0	Green
1	Taxi Driver	94.0	Green
2	Raging Bull	89.0	Green
3	The Departed	85.0	Green
4	Casino	73.0	Green
5	The Color of Money	77.0	Green
6	The Wolf of Wall Street	75.0	Green
7	Hugo	83.0	Green
8	The Irishman	94.0	Green
9	The Aviator	77.0	Green
10	Gangs of New York	72.0	Green
11	The Last Waltz	88.0	Green
12	Shutter Island	63.0	Green
13	The King of Comedy	73.0	Green
14	Bringing Out the Dead	70.0	Green
15	After Hours	90.0	Green
16	Cape Fear	73.0	Green
17	Mean Streets	96.0	Green

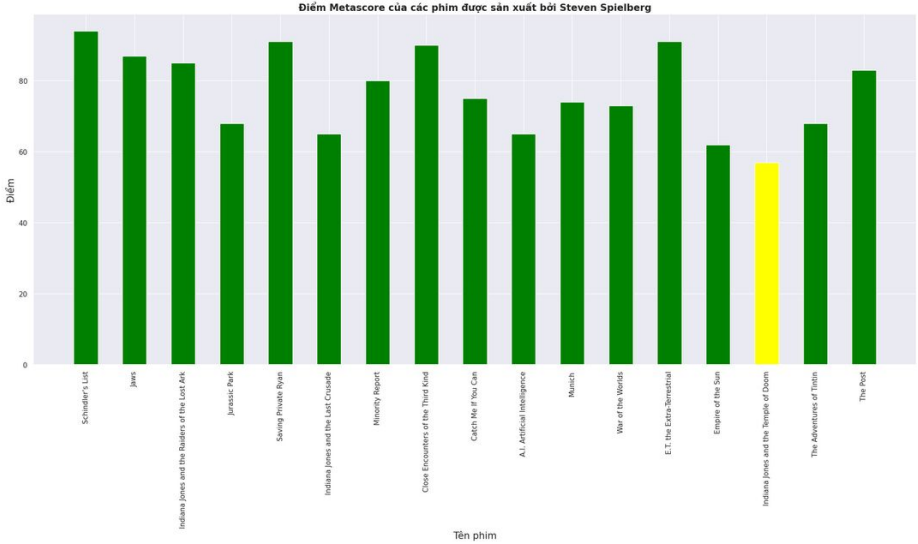
Steven Spielberg

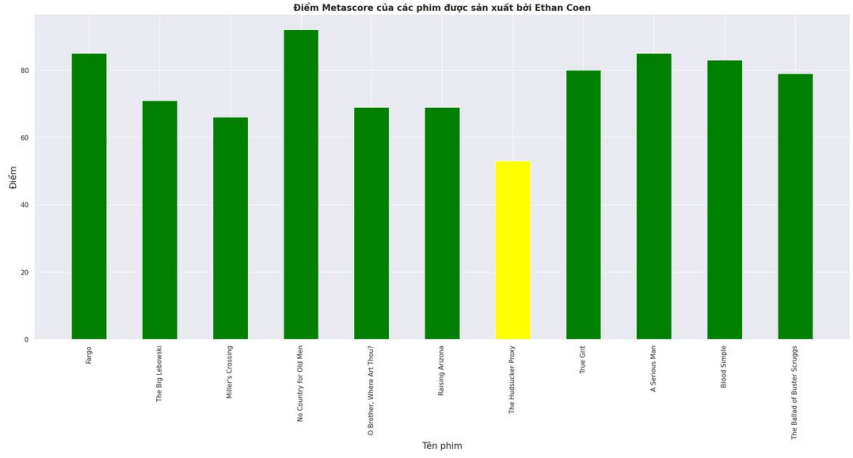
	Name	Meta score	Color
0	Schindler's List	94.0	Green
1	Jaws	87.0	Green
2	Indiana Jones and the Raiders of the Lost Ark	85.0	Green
3	Jurassic Park	68.0	Green
4	Saving Private Ryan	91.0	Green
5	Indiana Jones and the Last Crusade	65.0	Green
6	Minority Report	80.0	Green
7	Close Encounters of the Third Kind	90.0	Green
8	Catch Me If You Can	75.0	Green
9	A.I. Artificial Intelligence	65.0	Green
10	Munich	74.0	Green
11	War of the Worlds	73.0	Green
12	E.T. the Extra-Terrestrial	91.0	Green
13	Empire of the Sun	62.0	Green
14	Indiana Jones and the Temple of Doom	57.0	Yellow
15	The Adventures of Tintin	68.0	Green
16	The Post	83.0	Green

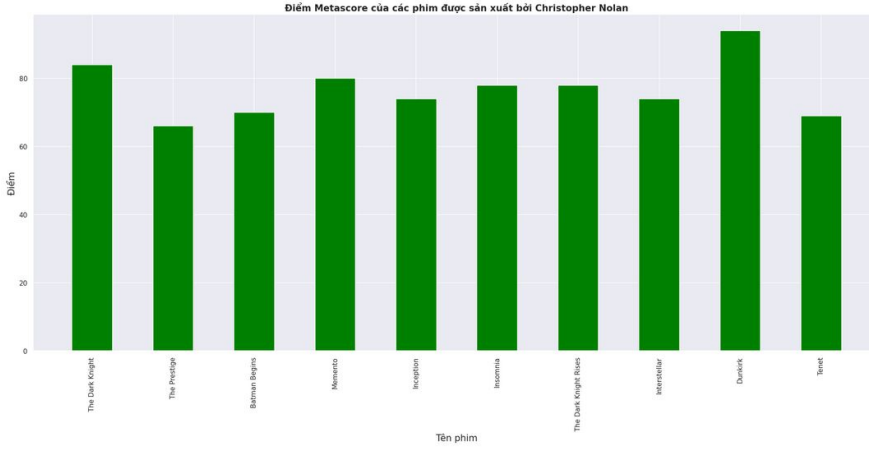
Bước 4: Ta tiến hành visualize bằng Bar chart để nhận xét.




Điểm Metascore của các phim được sản xuất bởi Joel Coen
Điểm
Tên phim
Fargo
The Big Lebowski
Miller's Crossing
No Country for Old Men
O Brother, Where Art Thou?
Raising Arizona
The Hudsucker Proxy
True Grit
A Serious Man
Blood Simple
The Man Who Wasn't There
The Ballad of Buster Scruggs


Điểm Metascore của các phim được sản xuất bởi Steven Spielberg
Điểm
Tên phim
Schindler's List
Jaws
Indiana Jones and the Raiders of the Lost Ark
Jurassic Park
Saving Private Ryan
Indiana Jones and the Last Crusade
Minority Report
Close Encounters of the Third Kind
Catch Me If You Can
A.I. Artificial Intelligence
Munich
War of the Worlds
E.T. the Extra-Terrestrial
Empire of the Sun
Indiana Jones and the Temple of Doom
The Adventures of Tintin
The Post


Điểm Metascore của các phim được sản xuất bởi Ethan Coen
Điểm
Tên phim
Fargo
The Big Lebowski
Miller's Crossing
No Country for Old Men
O Brother, Where Art Thou?
Raising Arizona
The Hudsucker Proxy
True Grit
A Serious Man
Blood Simple
The Ballad of Buster Scruggs


Điểm Metascore của các phim được sản xuất bởi Christopher Nolan
Điểm
Tên phim
The Dark Knight
The Prestige
Batman Begins
Memento
Inception
Insomnia
The Dark Knight Rises
Interstellar
Dunkirk
Tenet

Kết luận

- Hầu hết các phim đều được xếp vào hạng điểm xanh lá, nghĩa là phim hay. Chỉ có một vài phim phải nhận hạng điểm vàng (trung bình).
- Và qua những gì đã phân tích, ta thấy được rằng theo nhận xét và đánh giá của các chuyên gia, các phim của các đạo diễn trên đều đạt được những đánh giá cao về mặt nghệ thuật.
- Ngoài ra, các đạo diễn trên đều xuất hiện rất nhiều lần trong dataset và phim của họ của họ cũng nhận được những đánh giá cao.
- Ta thấy được rằng theo nhận xét và đánh giá của các chuyên gia, các phim của các đạo diễn trên đều đạt được những đánh giá cao về mặt nghệ thuật.

Câu 3: Liệu những phim được đóng bởi diễn viên xuất hiện nhiều trong top phim hay có còn trending ở thời điểm hiện tại không? Nếu không thì có mâu thuẫn gì với điểm IMDb không?

Ý nghĩa khi trả lời câu hỏi:

- Trả lời câu hỏi sẽ giúp những nhà làm phim xác định được những bộ phim do các diễn viên hạng A đóng có tạo trend và tạo ấn tượng lâu dài cho người xem không. Bởi không ai muốn xem bộ phim bị flop cả, nhà làm phim lẫn đạo diễn cũng không mong điều đó dù là thời gian nào.
- Điều này sẽ giúp đảm bảo được lợi nhuận nếu gửi gắm nhân vật đúng diễn viên hạng A có thể gánh phim.

Nguồn cảm hứng đặt câu hỏi:

- Hồi trước khi xem phim, người ta thường để ý tới những diễn viên hạng A như một hiện tượng, một sự lôi kéo người ta ra rạp xem, dù chẳng cần biết là về nội dung gì và chất lượng ra sao.
- Nhưng dạo gần đây, cái tên của những diễn viên hạng A dường như không đủ sức nặng và gây ấn tượng với người xem như trước. Người ta dần để ý hơn về đạo diễn phim và chủ đề phim.
- Vậy nên, mình mới thắc mắc những diễn viên đó dù không được như những thập niên 9X thì có còn trending ở hiện tại không.

Ý tưởng giải quyết

Sử dụng thang điểm Popularity để kiểm tra mức độ trending ở thời điểm hiện tại bằng cách tính bình quân mỗi phim do từng diễn viên gạo cội đóng có vượt trên điểm trung bình Popularity của toàn danh sách không.

Đồng thời cũng tiến hành kiểm tra với thang điểm IMDb để đưa ra nhận xét.

Popularity là gì?

- Nếu nói về IMDb thì chắc các bạn đều quen thuộc cả. Nhưng khi nói đến thang điểm Popularity thì chắc ít ai để ý.
- Thang điểm Popularity đánh giá mức độ phổ biến của phim theo nhiều phương diện.

Bước 1: Tìm những diễn viên gạo cội xuất hiện nhiều trong top phim hay

```
Top 5 diễn viên gạo cội xuất hiện nhiều nhất trong top phim hay:
```

Out[70]:

	Actor
0	Robert De Niro
1	Tom Cruise
2	Tom Hanks
3	Al Pacino
4	Brad Pitt

Bước 2: Xác định những bộ phim do diễn viên đó đóng.

Những phim do diễn viên top 1 đóng:

	Film
0	Goodfellas
1	Taxi Driver
2	The Godfather Part II
3	Heat
4	Raging Bull
5	Once Upon a Time in America
6	Casino
7	The Untouchables
8	The Deer Hunter
9	Meet the Parents
10	The Irishman
11	Brazil

Những phim do diễn viên top 2 đóng:

	Film
0	Magnolia
1	Rain Man
2	Minority Report
3	The Color of Money
4	Interview with the Vampire: The Vampire Chronicles
5	Top Gun
6	Jerry Maguire
7	Mission: Impossible - Ghost Protocol
8	Mission: Impossible
9	Collateral
10	Edge of Tomorrow
11	Mission: Impossible - Fallout

Những phim do diễn viên top 3 đóng:

	Film
0	Toy Story
1	The Green Mile
2	Forrest Gump
3	Toy Story 2
4	Saving Private Ryan
5	Road to Perdition
6	Toy Story 3
7	Catch Me If You Can
8	Big
9	Captain Phillips
10	Cast Away
11	Apollo 13
12	Philadelphia
13	The Post

Những phim do diễn viên top 4 đóng:

	Film
0	The Godfather
1	The Godfather Part II
2	Scarface
3	Heat
4	Dog Day Afternoon
5	Glengarry Glen Ross
6	Carlito's Way
7	Donnie Brasco
8	The Irishman
9	Insomnia
10	The Insider
11	Scent of a Woman
12	The Devil's Advocate
13	The Godfather Part III

Bước 3: Tính bình quân điểm IMDb và Popularity của từng diễn viên tìm được.

Out[70]:

	Top	Average Popularity	Average IMDb
0	Top 1	1650.1	7.7
1	Top 2	1560.1	7.3
2	Top 3	1048.2	8.0
3	Top 4	1315.2	8.0
4	Top 5	860.3	7.9

Bước 4: Xét xem mỗi điểm của từng diễn viên có lớn hơn điểm bình quân của toàn danh sách không.

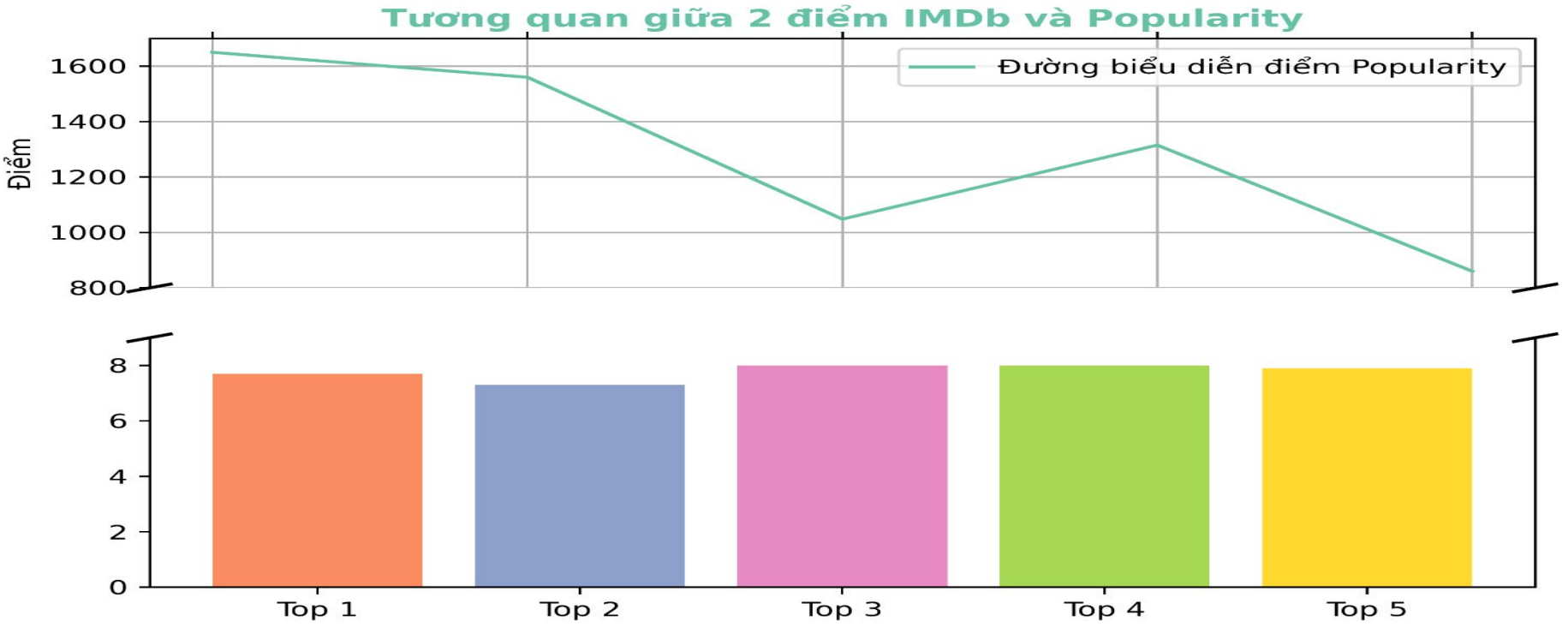
Diễn viên có bình quân môĩ phim trending cao hơn trung bình:

	Top	Average Popularity	Average IMDb
0	Top 1	1650.1	7.7
1	Top 2	1560.1	7.3

Diễn viên có bình quân môĩ phim được người xem đánh giá cao hơn trung bình:

	Top	Average Popularity	Average IMDb
0	Top 1	1650.1	7.7
2	Top 3	1048.2	8.0
3	Top 4	1315.2	8.0
4	Top 5	860.3	7.9

Vẽ biểu đồ minh hoạ



Kết luận

- Những bộ phim của những diễn viên gạo cội (xuất hiện nhiều trong top phim hay) không hoàn toàn trending ở hiện tại. Có thể lý giải là vì, theo thời gian, thị hiếu người xem thay đổi, có những phim nhân văn và ý nghĩa dần trở nên kén người xem. Dòng phim nghệ thuật cũng phải cạnh tranh với các dòng phim thương mại "kiểu mì ăn liền", đã không còn hiệu ứng trending như những năm ở thế kỷ trước.

- Đồng thời, những phim do diễn viên đó đóng hầu như sẽ cao hơn bình quân điểm IMDb. Hơi trái ngược với điểm Popularity nhưng có thể hiểu rằng, bởi nếu ai dành thời gian xem những bộ phim đó đều sẽ dành lời khen và đánh giá tích cực, chỉ là ở thời điểm hiện tại, quá nhiều bộ phim tràn lan ngoài kia được quay mỗi ngày và chúng ta ai cũng thích xem những phim mới chứ sẽ thường không đào lại những phim cũ của thế kỷ trước, hoặc mười mấy năm trước. Vậy nên điểm Popularity sẽ rất khác với IMDb và điều đó không mâu thuẫn gì.

Câu 4: Thể loại phim nào đang chiếm giữ phần trăm về số lượng trong bảng xếp hạng lớn nhất?

Ý nghĩa khi trả lời câu hỏi:

- Ta sẽ biết được về thể loại phim xuất hiện nhiều nhất trên bảng xếp hạng.
- Câu hỏi này giúp chúng ta nhìn ra thông tin cơ bản trong tập dữ liệu, thể loại phim nào chiếm phần lớn trong tập dữ liệu này.

Nguồn cảm hứng đặt câu hỏi:

- Thường thì khi tìm phim để xem, ta thường tìm với những từ khóa về thể loại nhiều hơn là tên phim hay các diễn viên trong phim. Do đó, em mới nghĩ ra câu hỏi này.

Bước 1: Ta lấy cột Genres từ tập dữ liệu và tiền xử lý. Sau đó đếm số lượng của mỗi thể loại.

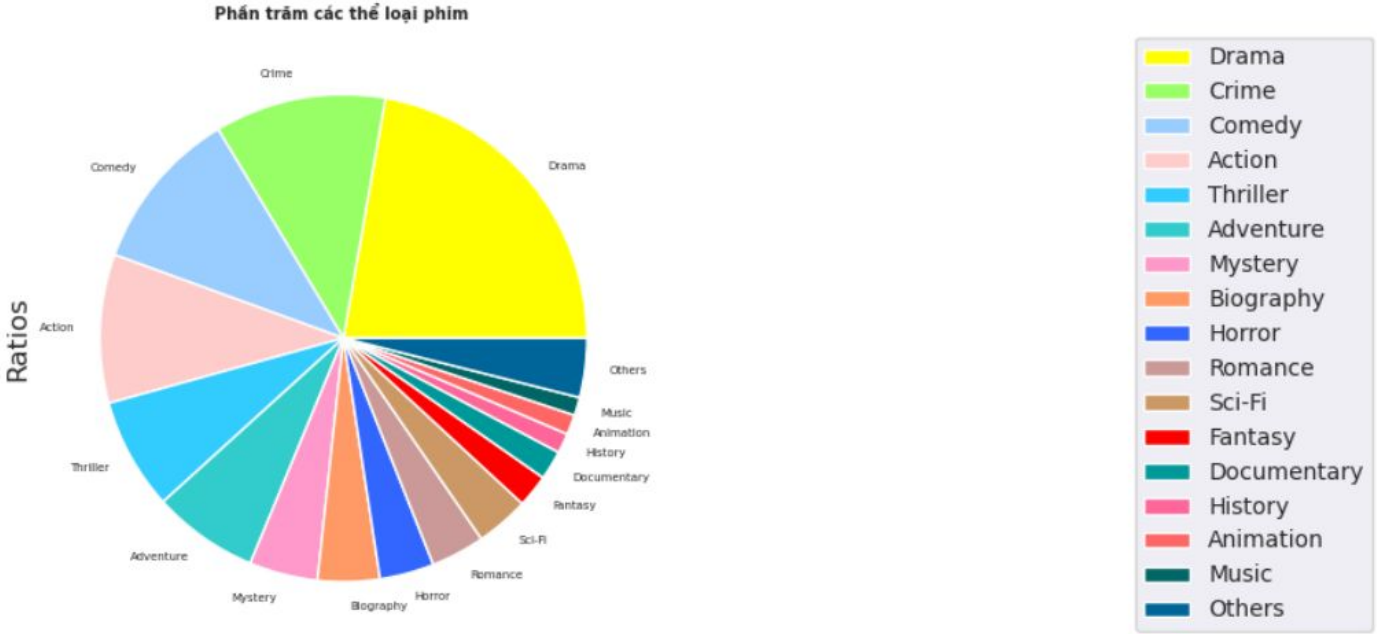
	Sum each genre
Drama	573
Crime	293
Comedy	279
Action	254
Thriller	191
Adventure	181
Mystery	117
Biography	106
Horror	93
Romance	92
Sci-Fi	91
Fantasy	56
Documentary	50
History	34
Animation	34
Music	31
War	24
Sport	23
Family	23
Western	17
Film-Noir	8
Musical	6

Bước 2: Ta tiến hành tính phần trăm mỗi thể loại.

	Sum each genre
Drama	573
Crime	293
Comedy	279
Action	254
Thriller	191
Adventure	181
Mystery	117
Biography	106
Horror	93
Romance	92
Sci-Fi	91
Fantasy	56
Documentary	50
History	34
Animation	34
Music	31
Others	101

	Ratios
Drama	22.24379
Crime	11.37422
Comedy	10.83075
Action	9.86025
Thriller	7.41460
Adventure	7.02640
Mystery	4.54193
Biography	4.11491
Horror	3.61025
Romance	3.57143
Sci-Fi	3.53261
Fantasy	2.17391
Documentary	1.94099
History	1.31988
Animation	1.31988
Music	1.20342
Others	3.92081

Bước 3: Dùng Pie chart để visualize và nhận xét



Kết luận

- Nhìn vào biểu đồ, ta có thể thấy ngay thể loại chiếm tỉ lệ cao nhất là thể loại Drama (Chính kịch).
- Ngoài ra một số thể loại Action, Comedy, Crime cũng chiếm một phần không nhỏ trong bảng xếp hạng 1000 phim này.

Câu 5: Tương quan giữa doanh thu (gross) và kinh phí (budget) với một số vấn đề liên quan về Metascore, User reviews, IMDb RATING, Votes, Popularity => để nhận xét: Liệu các phim có hoàn được vốn(lợi nhuận) hay không ?

Ý nghĩa khi trả lời câu hỏi:

- Cố gắng tìm kiếm các yếu tố nào của một bộ phim có ảnh hướng đến lợi nhuận của bộ phim nhất. Căn cứ vào các yếu tố đó nhà làm phim có thể sản xuất phim và không bị lỗ cũng như đẩy lợi nhuận cao hơn của các bộ phim trong tương lai.

Nguồn cảm hứng đặt câu hỏi:

- Xuất phát từ mục đích của bất cứ công việc làm gì trong cuộc sống nói chung và làm phim của các nhà sản xuất nói riêng là kiếm ra lợi nhuận từ các việc đầu tư. Từ đó cụ thể hóa các yếu tố tác động đến phim trở nên rõ ràng hơn giúp ích cho các nhà sản xuất giải quyết bài toán kinh tế.

Câu hỏi 5: Các bước thực hiện.

Bước 1: Tiền xử lý một số dữ liệu cần thiết (xử lý trường hợp NaN).

Bước 2: Thêm cột lợi nhuận (Profit = gross-budget), Sort data tăng dần theo lợi nhuận (Profit), gôm thành các khoảng 10 phim theo thứ tự và tính trung bình theo từng yếu tố (Metascore, User reviews, IMDb RATING, Votes, Popularity).

Bước 3: Vẽ các biểu đồ cột lợi nhuận.

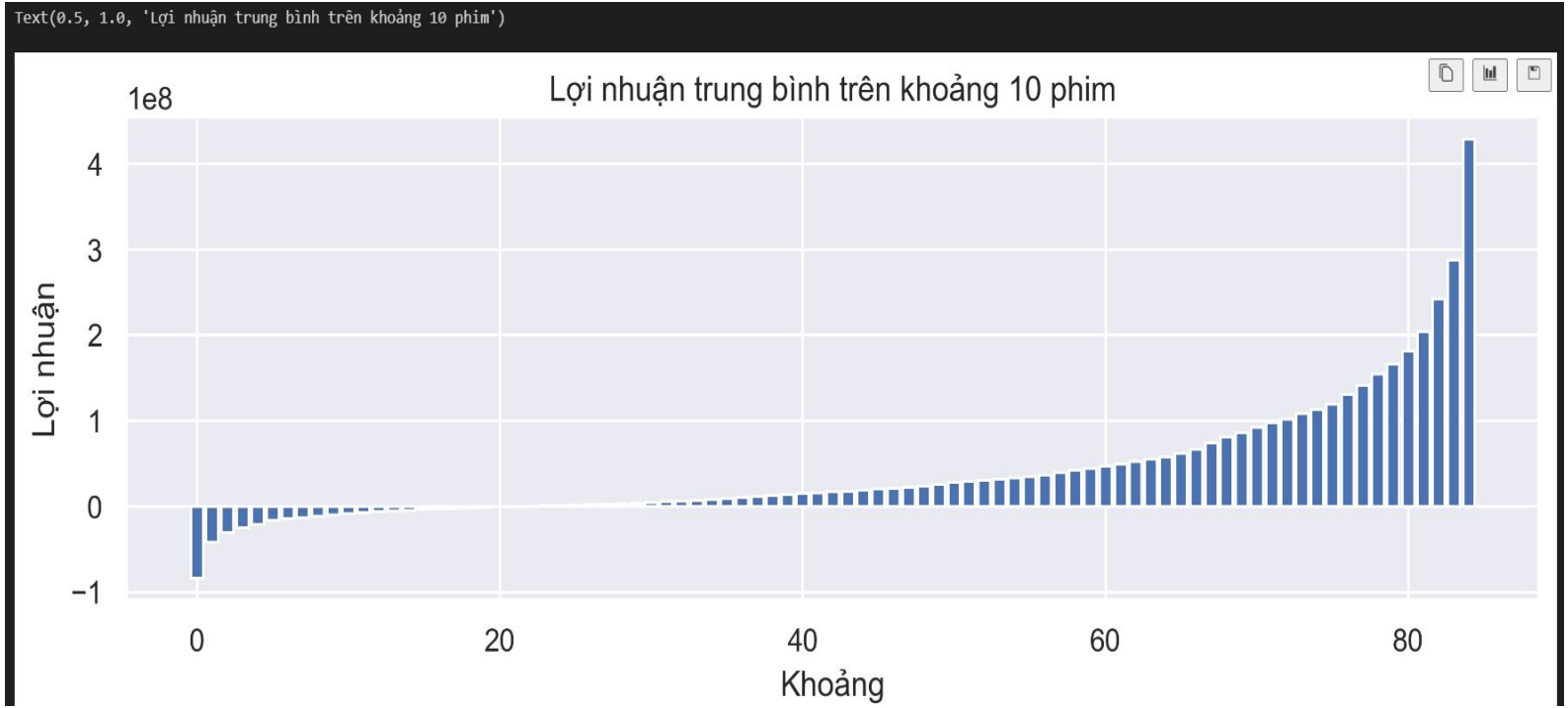
Bước 4: Vẽ các biểu đồ cột từng yếu tố từ đó tương quan có tăng theo lợi nhuận không và kết luận.

Minh họa bước 2:

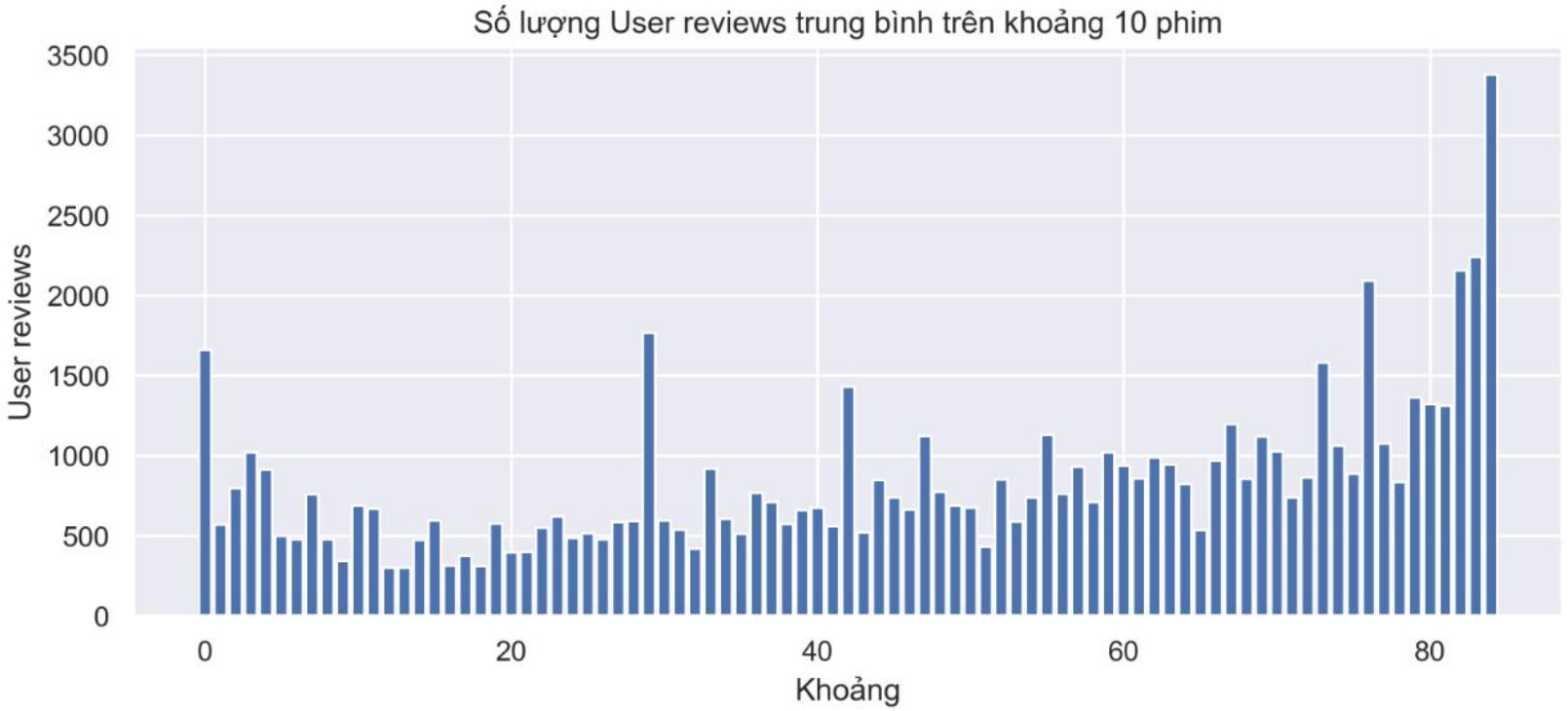
```
correlate_profit_with_x.head()
```

✓ 0.1s

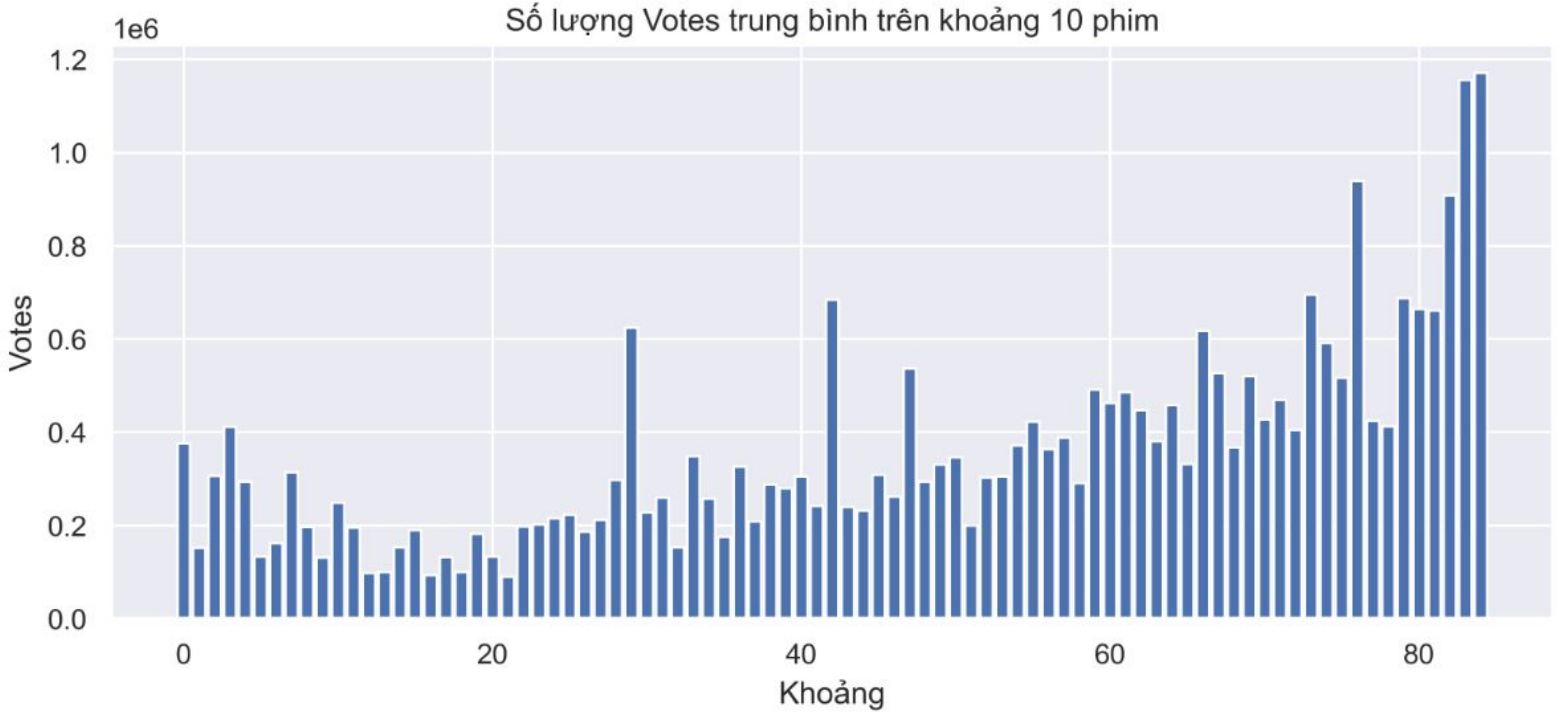
	Profit	Meta score	User reviews	IMDb RATING	Votes	Popularity
0	-84096863.1	75.5	1659.8	7.4	376482.2	1441.7
1	-41762269.5	67.1	569.8	7.4	151907.0	5783.8
2	-30564143.7	70.4	797.2	7.3	305619.1	2339.6
3	-24877939.4	64.8	1021.4	7.4	411859.2	3538.2
4	-20815091.1	71.3	912.1	7.5	293728.2	3242.2



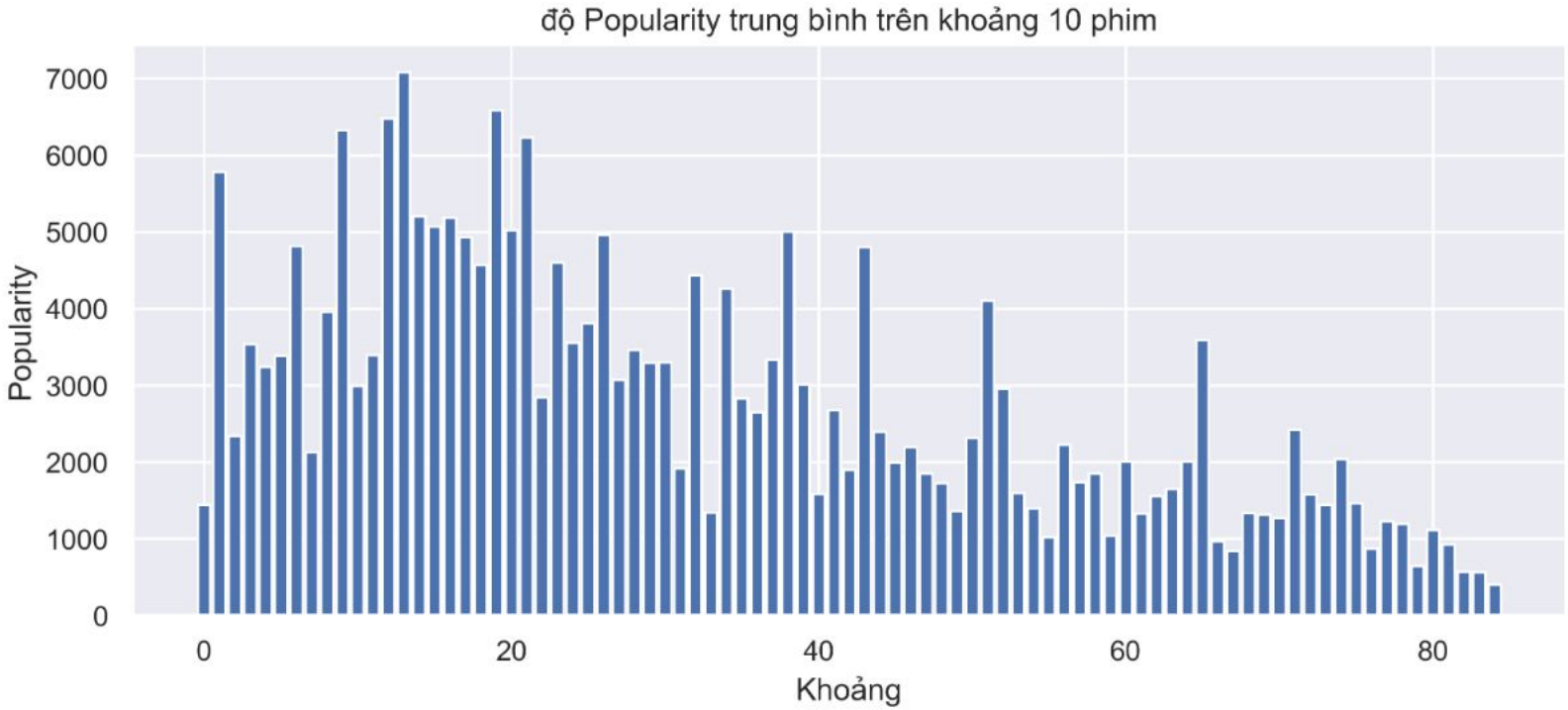
Lợi nhuận trung bình trên khoảng 10 phim được sắp xếp tăng



Số lượng user reviews phản ảnh khá rõ ràng nếu tăng số lượng user reviews thì lợi nhuận sẽ tăng.



Tương tự như user reviews, số lượng vote tăng cao thì lợi nhuận sẽ tăng.



Các phim có top càng đứng đầu (càng nhỏ) lợi nhuận sẽ cao.

Nhận xét câu hỏi 5:

- Ta có thể dựa vào các yếu tố như: User reviews, Votes, Popularity là các chỉ số phản ánh rõ ràng nhất trực tiếp từ người xem mà suy luận khả năng sinh lợi nhuận của một phim

- Song ta cũng có thể dựa vào các yếu tố như: IMDb RATING, Metascore là các điểm được chấm từ chuyên gia về chất lượng phim, hình ảnh cũng phản ánh khách quan nếu điểm cao thì lợi nhuận cao nhưng đôi khi chuyên môn cũng không phản ánh chính xác mà chỉ phần nào.

➔ Để phán đoán phim làm ra có đạt lợi nhuận cao hay không phần nào có thể đánh giá dựa trên các yếu tố như: User reviews, Votes, IMDb RATING, Metascore,... Nhưng đôi lúc cũng không hoàn toàn chính bởi các yếu tố khách quan hay chủ quan nào đó nên mới có các phim lỗ dù đầu tư cao về nhiều mặt.

Phần 2: Mô hình hoá dữ liệu

Bài toán đặt ra

- Chúng ta đã biết rằng, có rất nhiều kiểu để phân loại một bộ phim.
- Ví dụ như, phân loại theo thể loại: Hành động, Phiêu lưu, Chính kịch,... , theo độ tuổi xem, R, PG, G,... Hay cũng có xếp hạng theo điểm IMDb, theo độ nổi tiếng.
- Vậy liệu rằng chúng ta có thể dựa trên các thông tin ở trên để đáng giá một bộ phim thành công nhiều hay ít không?

Hướng giải quyết

Để giải quyết vấn đề trên nhóm quyết định dựa trên các thuật toán hỗ trợ trong Sklearn để cố phân cụm các phim, nhằm đánh giá mức thành công của bộ phim đó.

Dựa trên những giá trị ở các cột để phân cụm:

- IMDb RATING
- Popularity
- Meta score
- Budget
- Gross
- Wins

Thuật toán sử dụng

Thuật toán sử dụng	Giải thích sử dụng thuật toán
• K-Means • DBSCAN	• Những thuật toán phổ biến, được sử dụng phổ biến.

Mô hình phân cụm

K-Means

Các bước thực hiện

STT	Thực hiện	Nội dung
1	Chọn cột thuộc tính	Chọn các cột đầu vào, quyết định cho các cụm đầu ra.
2	Xử lý cột thuộc tính	• Xóa các phim thiếu 2 ô dữ liệu trở lên. • Cột Numeric: điền các giá trị mean vào những phần dữ liệu còn thiếu. Tạo cột: Balance = Gross - Budget • Cột Category: chuyển dữ liệu thành kiểu Numeric với những ước lượng mà nhóm đặt ra.
3	Chạy mô hình	Chọn siêu tham số, cho chạy mô hình
4	Đánh giá mô hình	Tinh chỉnh siêu tham số đầu vào theo biểu đồ đánh giá Elbow, Silhouette Score.
5	Phân cụm dữ liệu theo cụm của mô hình.	Xử lý, sửa đổi nội dung các cụm.

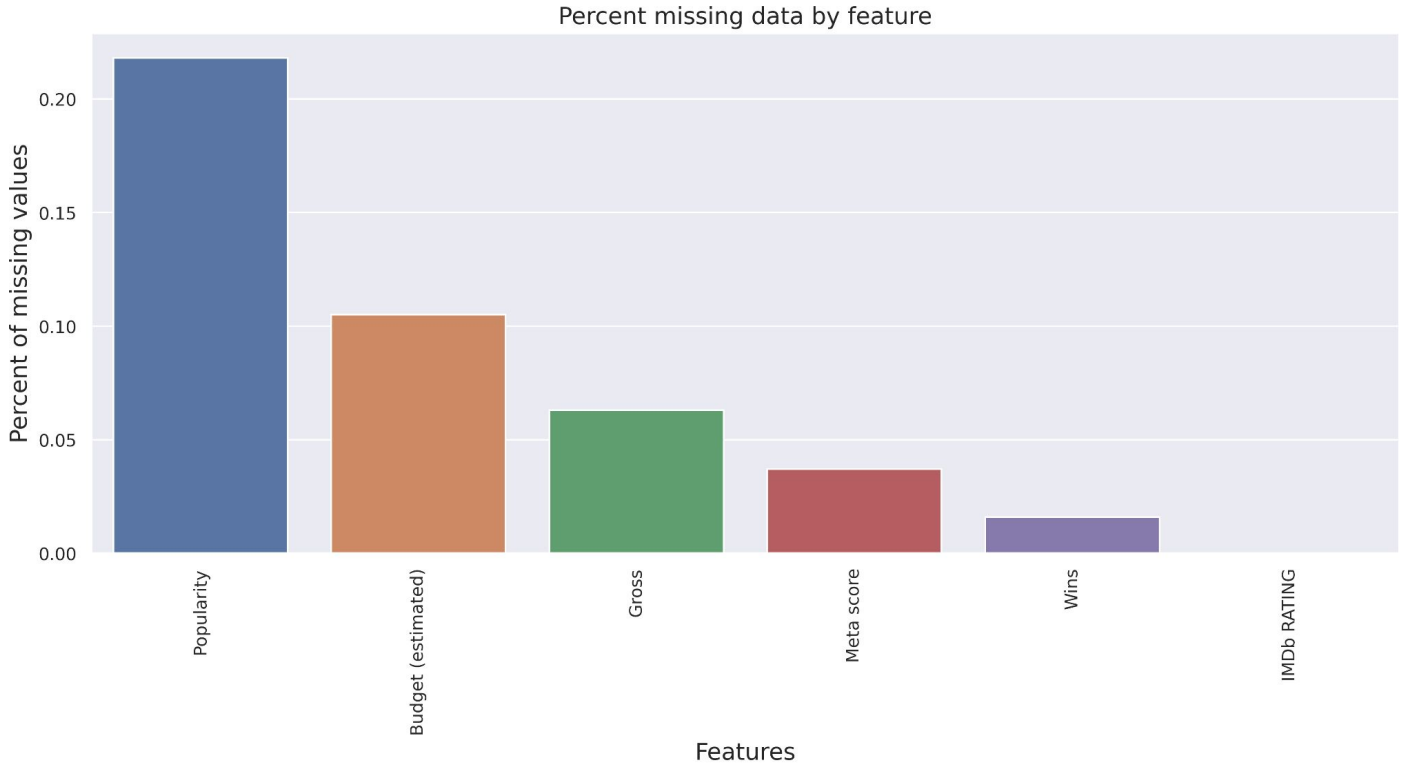
Bước 1: Chọn cột thuộc tính

```
cluster_df=cluster_full_info_df.loc[:,['IMDb RATING','Popularity','Meta score','Budget (estimated)','Gross','Wins']]
cluster_df.head(10)
```

✓ 0.1s

	IMDb RATING	Popularity	Meta score	Budget (estimated)	Gross	Wins
0	9.2	92.0	100.0	6000000.0	134966411.0	Won 3 Oscars
1	8.7	194.0	90.0	25000000.0	46836394.0	Won 1 Oscar
2	8.9	115.0	94.0	8000000.0	107928762.0	Won 1 Oscar
3	8.5	425.0	77.0	6000000.0	23341568.0	Won 2 Oscars
4	8.5	307.0	94.0	31500000.0	83471511.0	Won 2 Oscars
5	8.1	784.0	83.0	1830000.0	16501785.0	Nominated for 1 Oscar
6	8.8	139.0	66.0	63000000.0	37030102.0	Nominated for 1 Oscar
7	9.0	219.0	94.0	22000000.0	96898818.0	Won 7 Oscars
8	7.9	462.0	85.0	15000000.0	26400640.0	Nominated for 3 Oscars
9	8.3	381.0	79.0	1200000.0	2832029.0	Awards

Bước 2: Xử lý các cột



Bước 2: Xử lý các cột

```
X1=cluster_df.iloc[:,:5]
X2=cluster_df.iloc[:,:5]
X1
```

	IMDb RATING	Popularity	Meta score	Balance	Wins
0	9.2	92.0	100.0	128966411.0	5.482062e+08
1	8.7	194.0	90.0	21836394.0	1.827354e+08
2	8.9	115.0	94.0	99928762.0	1.827354e+08
3	8.5	425.0	77.0	17341568.0	3.654708e+08
4	8.5	307.0	94.0	51971511.0	3.654708e+08
...	...	...	...	...	...
995	6.7	2463.0	52.0	28348319.0	3.654708e+07
996	7.6	1362.0	56.0	15903593.0	3.654708e+07
997	7.5	206.0	68.0	-9308792.0	3.654708e+07
998	7.2	1207.0	63.0	23722567.0	3.654708e+07
999	7.7	506.0	66.0	214591735.0	3.654708e+07

964 rows × 5 columns

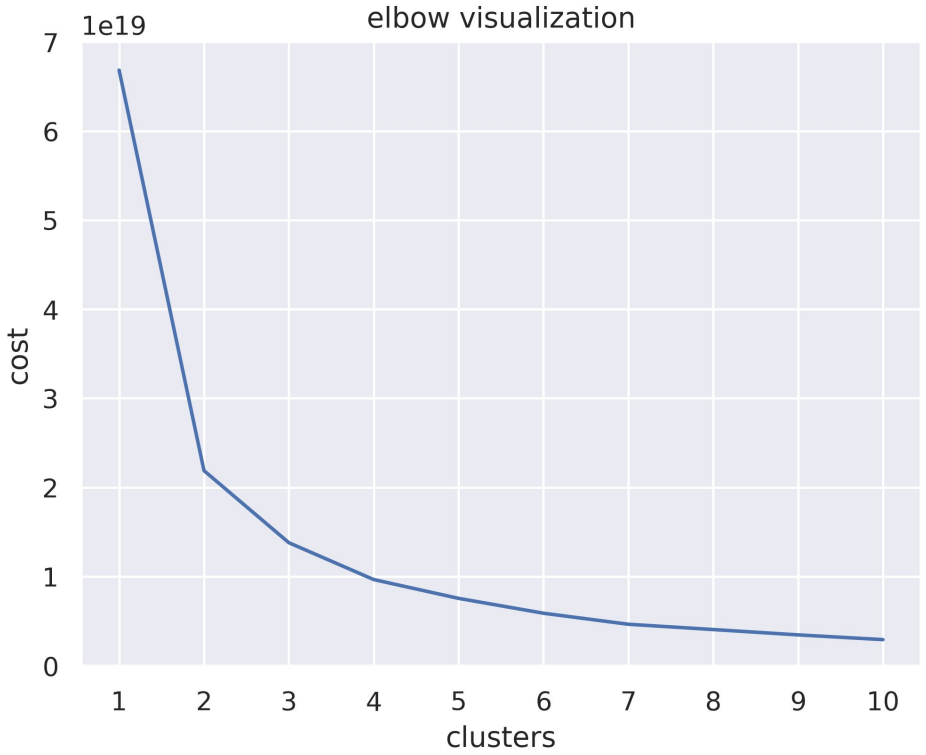
- 1 phim trung bình sẽ có thể nhận đến tối đa 40 đề cử Oscar, và trong 1 nội dung thắng, thì trung bình có 5 đề cử. Nên ở đây nhóm sẽ cố gắng quy đổi thông tin các giải thưởng mà phim nhận được thành dạng numeric.
- Xem như các giải thưởng thắng được là ngang nhau trong thứ bậc, và đề cử có đồng giá trị như nhau. Ở đây:
    - 1 Win = 5 Nominations
    - 1 Nomination = 1 Award
    - None = 0
- Mặt khác nhóm còn muốn cột Wins đóng vai trò lớn nhất trong phần quyết định rằng phim đó thành công ra sao.

Bước 3: Chạy mô hình

Những siêu tham số cần xác định trong mô hình phân cụm theo K-Means:

- **n_nit**: Số lần khởi tạo tâm và chọn giá trị hội tụ phù hợp nhất *(default=10)*, Nhóm muốn tăng số lần khởi tạo để có thể quét được nhiều nhất không gian của đối tượng đó. Chọn **n_nit=50**
- **max_iter**: Với n_init lần chạy và mỗi lần chạy lặp lại max_iter lần *(default=300)*, được hiểu trong một lần chạy, điểm sẽ được gán cho các cụm khác nhau và chi phí được tính cho max_iter lần. Với max_iter ở giá trị cao, thì ta được đảm bảo rằng sẽ khám phá toàn bộ không gian đối tượng, nhưng đồng nghĩa với chi phí sẽ tăng lên. Chọn **max_iter=500**
- **nit**: tọa độ tâm của cụm: với lượng dữ liệu lớn, khó để xác định được đúng tâm, nên ở đây nhóm để ở **nit='random'**
- **random_state**: là hạt giống sinh số ngẫu nhiên, nhằm tránh việc mỗi lần cho ra kết quả khác nhau. Ta có thể chọn bất kỳ giá trị nào, ở đây nhóm chọn **random_state=2020.**
- **n_clusters**: đóng vai trò **quan trọng nhất**, số cụm, ở đây ngoài cách thử và sai ra thì, khó có thể chọn được siêu tham số này một cách phù hợp nhất. Thử và sai **n_clusters=[1,10].**

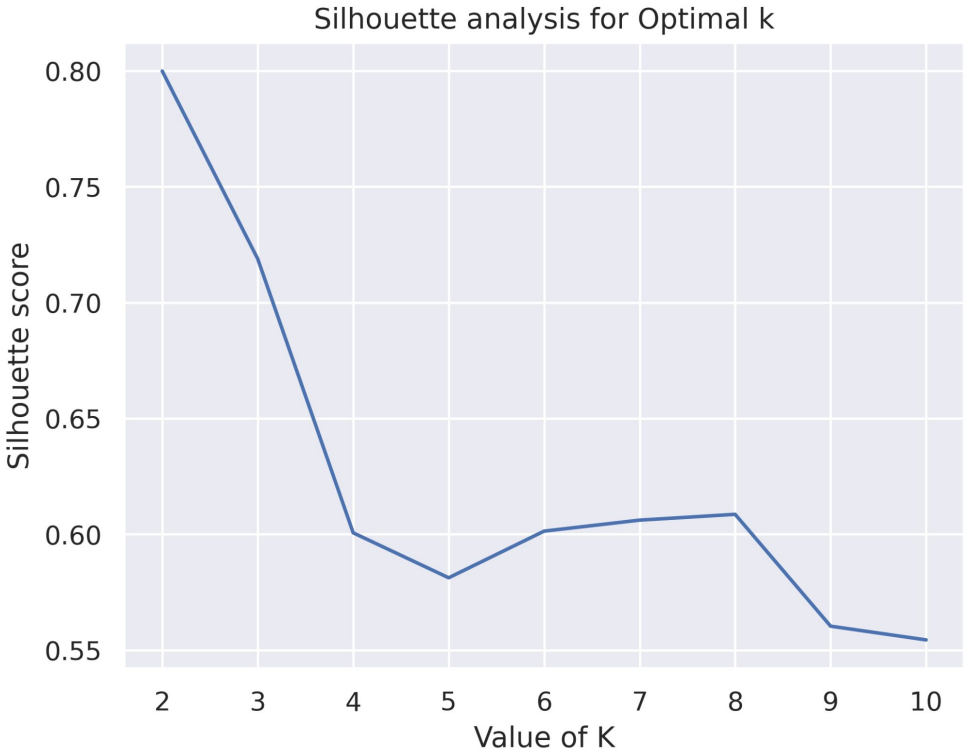
Bước 4: Đánh giá mô hình



Nhận xét:

- Ta thấy có sự thay đổi nhiều khi phân từ 1 cụm về 2 cụm. Nhưng khi phân từ 2 cụm trở lên, chi phí không thay đổi nhiều.

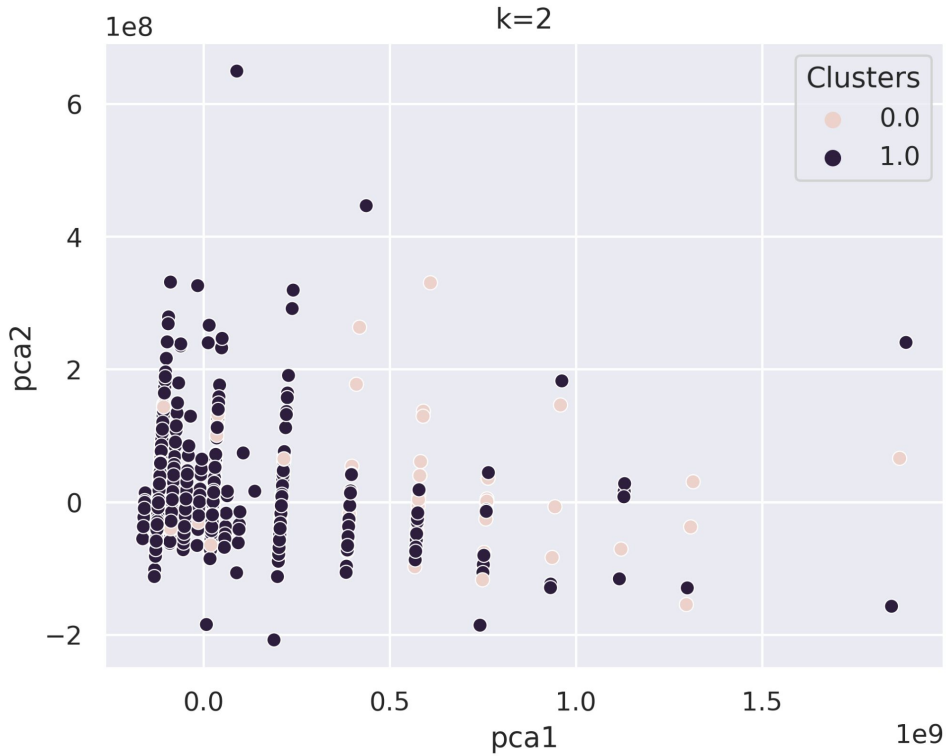
Bước 4: Đánh giá mô hình



Nhận xét:

- Ta thấy tại k=2 có giá trị Silhouette score lớn nhất.

Bước 4: Đánh giá mô hình




```
cluster_full_info_df_method1[cluster_full_info_df_method1.Clusters==1].Wins.unique()
```

```
array(['Won 3 Oscars', 'Won 7 Oscars', 'Won 6 Oscars', 'Won 4 Oscars',
       'Won 5 Oscars', 'Won 11 Oscars', 'Won 8 Oscars'], dtype=object)
```

Nhận xét:

- Các phim dựa trên mô hình được chia làm 2 cụm như bên hình.

Bước 5: Áp dụng vào phân cụm các phim

	STT	ID	Name	Published Year	Rated	Duration	Genres	Director	Writers	Stars	IMDb RATING	Budget (estimated)	Gross	Popularity	Votes	User reviews	Critic reviews	Meta score	Wins	URL	Clusters
0	1	tt0068646	The Godfather	1972	R	175.0	[Crime, Drama]	[Francis Ford Coppola]	[Francis Ford Coppola, Mario Puzo]	[Al Pacino, Marlon Brando, James Caan]	9.2	6000000.0	134966411.0	92.0	1849463	5200.0	193.0	100.0	Won 3 Oscars	/title/tt0068646/	Xuất Sắc
1	2	tt0099685	Goodfellas	1990	R	145.0	[Biography, Crime, Drama]	[Martin Scorsese]	[Nicholas Pileggi, Martin Scorsese]	[Ray Liotta, Robert De Niro, Joe Pesci]	8.7	25000000.0	46836394.0	194.0	1156922	1500.0	163.0	90.0	Won 1 Oscar	/title/tt0099685/	Thành công
2	3	tt0110912	Pulp Fiction	1994	R	154.0	[Crime, Drama]	[Quentin Tarantino]	[Roger Avary, Quentin Tarantino]	[Samuel L. Jackson, Uma Thurman, John Travolta]	8.9	8000000.0	107928762.0	115.0	2043441	3500.0	306.0	94.0	Won 1 Oscar	/title/tt0110912/	Thành công
3	4	tt0114814	The Usual Suspects	1995	R	106.0	[Crime, Drama, Mystery]	[Bryan Singer]	[Christopher McQuarrie]	[Chazz Palminteri, Kevin Spacey, Gabriel Byrne]	8.5	6000000.0	23341568.0	425.0	1083363	1400.0	154.0	77.0	Won 2 Oscars	/title/tt0114814/	Thành công
4	5	tt0078788	Apocalypse Now	1979	R	147.0	[Drama, Mystery, War]	[Francis Ford Coppola]	[Michael Herr, Francis Ford Coppola, John Milius]	[Martin Sheen, Marlon Brando, Robert Duvall]	8.5	31500000.0	83471511.0	307.0	666884	1300.0	302.0	94.0	Won 2 Oscars	/title/tt0078788/	Thành công

Dataframe sau khi phân cụm

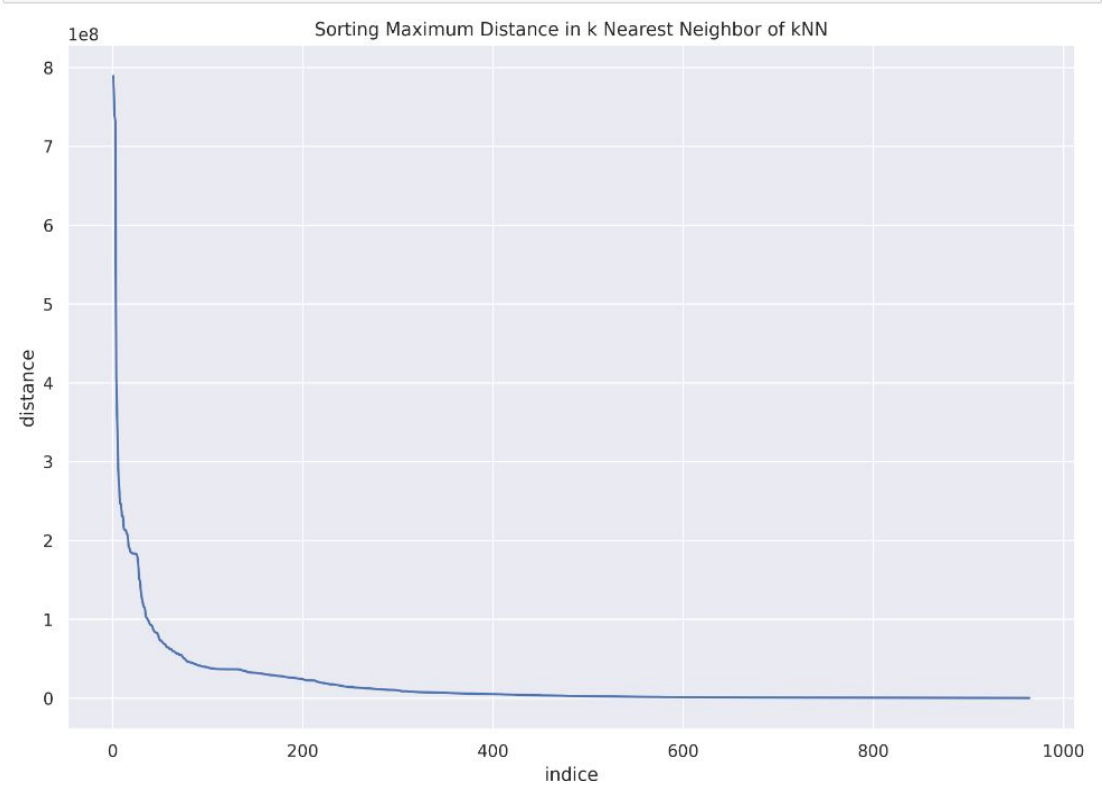
Mô hình phân cụm

Thuật toán DBSCAN

Những siêu tham số cần xác định trong mô hình phân cụm theo DBSCAN

- ***eps***: đóng vai trò **quan trọng nhất,** Khoảng cách tối đa giữa hai mẫu để một mẫu được coi là lân cận với mẫu kia. Được xác định dựa trên phương pháp K-Nearest Neighbours.
- ***min_samples***: Số lượng mẫu trong một vùng lân cận cho một điểm được coi là điểm cốt lõi. Hướng chọn tối ưu của siêu tham số này được đề xuất là chọn gấp đôi số cột. Chọn **min_samples = 10.**
- ***metric***: Khoảng cách được tính bằng phương pháp Euclean. Chọn **metric = eucliean**.

Xác định *eps* qua biểu đồ sắp xếp khoảng cách tối đa trong k hàng xóm gần nhất

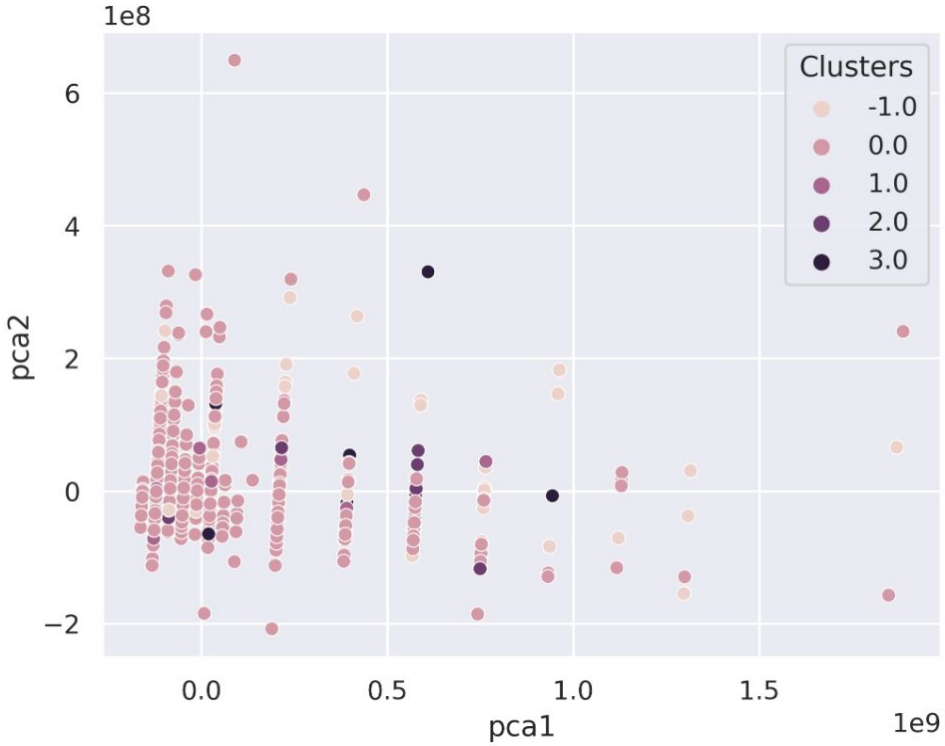


Nhận xét:

esp=0.5*10^8

là khoảng cách tối ưu giữa các cụm

Biểu đồ phân cụm



Nhận xét:

- Các phim được chia làm 5 cụm

```
cluster_full_info_df_method2[cluster_full_info_df_method2.Clusters==0].Wins.unique()
```
✓ 0.2s

```
array(['Won 1 Oscar', 'Nominated for 1 Oscar', 'Nominated for 3 Oscars',
       'Awards', 'Nominated for 7 Oscars', 'Nominated for 4 Oscars',
       'Won 1 BAFTA Award', 'Nominated for 2 Oscars',
       'Nominated for 1 BAFTA Award', 'Nominated for 5 BAFTA Awards',
       'Nominated for 5 Oscars', 'Nominated for 6 Oscars',
       'Nominated for 2 BAFTA Awards', None,
       'Nominated for 3 BAFTA Awards', 'Nominated for 2 Primetime Emmys',
       'Nominated for 3 Primetime Emmys'], dtype=object)
```

```
cluster_full_info_df_method2[cluster_full_info_df_method2.Clusters==1].Wins.unique()
```
✓ 0.3s

```
array(['Won 2 Oscars', 'Won 2 BAFTA Awards', 'Nominated for 10 Oscars'],
      dtype=object)
```

```
cluster_full_info_df_method2[cluster_full_info_df_method2.Clusters==2].Wins.unique()
```
✓ 0.3s

```
array(['Won 4 Oscars'], dtype=object)
```

```
cluster_full_info_df_method2[cluster_full_info_df_method2.Clusters==3].Wins.unique()
```
✓ 0.4s

```
array(['Won 3 Oscars'], dtype=object)
```

```
cluster_full_info_df_method2[cluster_full_info_df_method2.Clusters==-1].Wins.unique()
```
✓ 0.4s

```
array(['Won 7 Oscars', 'Won 3 Oscars', 'Won 6 Oscars', 'Won 2 Oscars',
       'Won 5 Oscars', 'Won 4 Oscars', 'Won 11 Oscars', 'Won 8 Oscars',
       'Nominated for 6 Oscars', 'Won 1 Oscar', 'Nominated for 10 Oscars',
       'Nominated for 1 Oscar', 'Nominated for 8 Oscars',
       'Nominated for 5 Oscars', 'Awards', 'Nominated for 4 Oscars',
       'Nominated for 3 Oscars', 'Nominated for 7 Oscars'], dtype=object)
```

- Dựa trên những giải xuất hiện trong các cụm có thể nhận thấy rằng, độ thành công được sắp xếp tăng dần như sau: Cụm 0 , Cụm 1, Cụm 3, Cụm 2
- Cụm **Clusters=-1**: chứa cả những bộ phim đạt nhiều giải Oscar, và cả những phim không được giải thưởng lớn nào. Và được mô hình đánh giá là những phim mang giá trị nhiễu.

Nhận xét 2 mô hình phân cụm:

- Từng thuật toán có cách chia cụm khác nhau.
- DBSCAN:
  - Chịu nhiễu tốt.
  - Chưa phù hợp để phân cụm cho bài toán đặt ra ở đầu bài.
  - Các siêu tham số khó xác định chính xác hơn.

- KMeans:
  - Dễ bị ảnh hưởng bởi những thông tin nhiễu.
  - Phù hợp để trả lời cho bài toán đã đặt ra ở đầu bài.
  - Các siêu tham số dễ xác định chính xác hơn.

Mô hình hồi quy tuyến tính

Mô hình hồi quy tuyến tính

Bài toán đặt ra:

- Phát triển tiếp câu hỏi số 5 của nhóm, bài toán sẽ là dự đoán lợi nhuận của một bộ phim sao cho gần và chính xác nhất so với thực tế
- Dựa vào việc đồng biến hoặc nghịch biến của các yếu tố câu hỏi 5 theo lợi nhuận dự đoán sử dụng mô hình hồi quy tuyến tính.

Giải quyết vấn đề trên bằng hai mô hình hồi quy :

- Hồi quy tuyến tính đơn.
- Hồi quy tuyến tính nhiều biến.

Hồi quy tuyến tính đơn

Bước 1: Chọn dữ liệu đặc trưng (features)

- Chọn những dữ liệu đặc trưng có tác động đến lợi nhuận (Profit) như câu hỏi số 5 cho vào tập XX.
- Chọn cột lợi nhuận cần dự đoán gần đúng cho vào tập YY.

```
XX = film_info_df_copy[['Meta score', 'User reviews', 'IMDb RATING', 'Votes', 'Popularity']]
YY =  film_info_df_copy['Profit']
```

✓ 0.8s

Hồi quy tuyến tính đơn

Bước 2: Splitting the dataset

- Để tránh việc overfitting trong quá trình học ta chia tập dữ liệu thành 2 phần 70% dùng để train và 30% dùng để test (XX, YY -> X_train, X_test, y_train, y_test)

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
                    XX, YY, test_size = 0.3, random_state = 0)
```

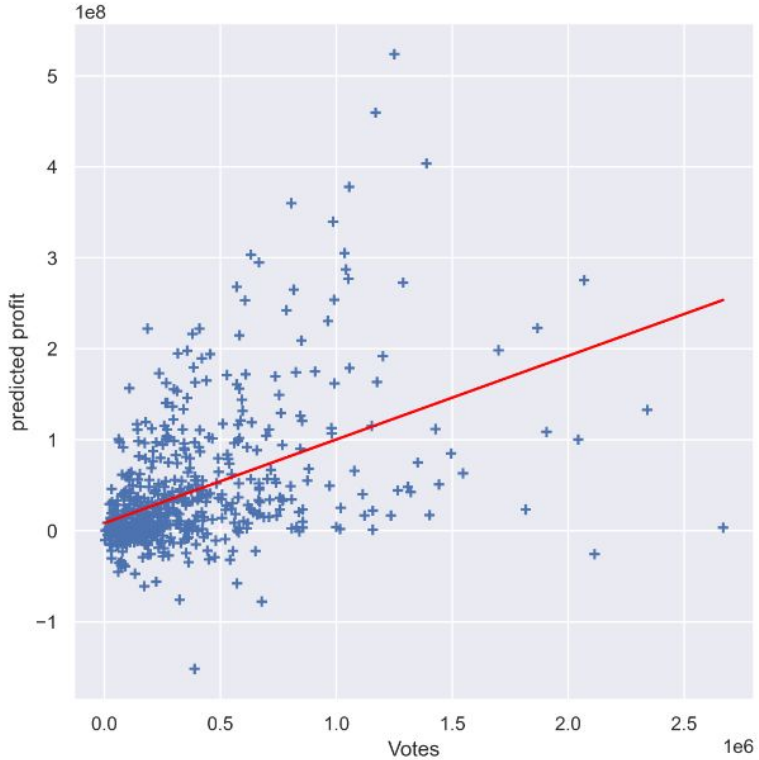
✓ 0.1s

Hồi quy tuyến tính đơn

Bước 3: Training machine learning model

- Dùng lỗi bình phương trung bình - MSE để tìm ra yếu tố tác động nhiều nhất đến lợi nhuận (MSE bé nhất)
- Xây dựng mô hình hồi quy tuyến tính đơn với yếu tố có MSE bé nhất.
- Xác định các tham số m và b trong đường hồi quy y = m*x + b.

Hồi quy tuyến tính đơn



Minh họa mô hình hồi quy tuyến tính đơn

Hồi quy tuyến tính đơn

Bước 4: Đánh giá hiệu suất model (model evaluation)

- Giá trị y và y dự đoán

	Y	y_preds
59	76801374.0	1.274082e+08
50	9858926.0	3.923550e+07
551	69043761.0	4.120593e+07
262	16847564.0	2.118997e+07
370	27014192.0	3.873148e+07

- Tính giá trị lỗi bình phương trung bình - MSE cho phần data testing

```
MSE = metrics.mean_squared_error(y_test,
                                 linear_regression.predict(X_test[[id_XX]]))
MSE
```

✓ 0.1s

```
5182736112664674.0
```

Hồi quy tuyến tính đơn

Nhận xét:

- Mô hình đã giúp ta dự đoán được các lợi nhuận của các phim dự vào yếu tố đặc trưng (Votes).
- Với tập dữ liệu lớn ta thấy giá trị lợi nhuận dự đoán y_preds khá khác với giá trị y theo giả thuyết.
- Dù các giá trị là khá lớn (10^7) nhưng MSE cho giá trị vẫn khá lớn
- Mô hình hồi quy tuyến tính đơn không quá phù hợp cho bài toán đặt ra cùng kiểm thử với mô hình hồi quy tuyến tính nhiều biến.

Hồi quy tuyến tính nhiều biến

Bước 1: Chọn dữ liệu đặc trưng (features)

- Chọn những dữ liệu đặc trưng có tác động đến lợi nhuận (Profit) như câu hỏi số 5 cho vào tập XX.
- Chọn cột lợi nhuận cần dự đoán gần đúng cho vào tập YY.

```
XX2 = film_info_df_copy[['Meta score', 'User reviews', 'IMDb RATING', 'Votes', 'Popularity']]
YY2 =  film_info_df_copy['Profit']
✓ 0.8s
```

Hồi quy tuyến tính nhiều biến

Bước 2: Splitting the dataset

- Để tránh việc overfitting trong quá trình học ta chia tập dữ liệu thành 2 phần 70% dùng để train và 30% dùng để test (XX, YY -> X_train, X_test, y_train, y_test).

```
from sklearn.model_selection import train_test_split

X_train2, X_test2, y_train2, y_test2 = train_test_split(
                                    XX2, YY2, test_size = 0.3, random_state = 0)
```

✓ 0.8s

Hồi quy tuyến tính nhiều biến

Bước 3: Training machine learning model

- Xây dựng mô hình hồi quy tuyến tính với tất cả các dữ liệu đặc trưng.
- Xác định tham số đường hồi quy y = b + w1*x1 + w2*x2 + ... + wn*xn với w = [w1, w2, ..., wn]

Nhận xét: Mô hình hồi quy nhiều hơn 3 biến không thể biểu diễn matplotlib được.

Hồi quy tuyến tính nhiều biến

Bước 4: Đánh giá hiệu suất model (model evaluation)

- Giá trị y và y dự đoán

	Y	y_preds
128	33979328.0	3.195418e+07
868	101697350.0	3.154963e+07
570	-4677400.0	5.000575e+07
301	25605492.0	6.275075e+07
988	-9245330.0	5.212527e+07

- Tính giá trị lỗi bình phương trung bình - MSE cho phần data testing

```
MSE = metrics.mean_squared_error(y_test2,
    linear_regression.predict(X_test2))
MSE
```

✓ 0.8s

```
5306289859389350.0
```

Hồi quy tuyến tính nhiều biến

Mở rộng: Ta có thể dùng lỗi bình phương trung bình - MSE để tìm ra yếu tố tác động nhiều nhất đến lợi nhuận (MSE bé nhất) để xây dựng mô hình hồi quy tuyến tính với các yếu tố có ảnh MSE bé nhất.

Nhận xét:

- Mô hình hồi quy tuyến tính nhiều biến đã đưa ra các giá trị lợi nhuận dự đoán tốt hơn với mô hình hồi quy tuyến tính đơn.
- Mô hình hồi quy không quá phù hợp với bài toán này.