ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

TRỰC QUAN HÓA DỮ LIỆU

< Lab 02 - NETFLIX >

Sinh viên thực hiện: 21127115 - Trần Thanh Ngân
21127229 - Dương Trường Bình
21127616 - Lê Phước Quang Huy

Giảng viên hướng dẫn: TS. Bùi Tiến Lên

Lớp: 21KHDL

Mục lục

Thông tin nhóm và phân công công việc

MSSV	Họ và tên	Công việc được phân công	Mức độ hoàn thành
21127115	Trần Thanh Ngân	• A. Mô tả dữ liệu • B. Khám phá và phân tích dữ liệu (câu hỏi 1, 2)	100%
21127229	Dương Trường Bình	• B. Khám phá và phân tích dữ liệu (câu hỏi 3, 4) • C. Trực quan dữ liệu tương tác (1)	100%
21127616	Lê Phước Quang Huy	• B. Khám phá và phân tích dữ liệu (câu hỏi 5, 6) • C. Trực quan dữ liệu tương tác (2)	100%

Tiến độ công việc

Phần	Nội dung	Mức độ hoàn thành
A. Mô tả dữ liệu	1. Viết bảng mô tả về tập dữ liệu	100%
	2. Phân tích tỷ lệ missing rate	100%
B. Khám phá và phân tích dữ liệu	Đặt câu hỏi về tập dữ liệu và trả lời (câu hỏi 1 - 6)	100%
C. Trực quan dữ liệu tương tác	Trực quan các biểu đồ có thể tương tác trực tiếp	100%
D. Insights	Chia sẻ các phát hiện thú vị.	100%

1 Tổng quan

> **Netflix** là dịch vụ giải trí trực tuyến hàng đầu thế giới với hơn 208 triệu lượt người sử dụng trên toàn cầu tại hơn 130 quốc gia, bao gồm đa dạng các loại chương trình truyền hình, phim dài tập, phim tài liệu và phim truyện thuộc nhiều thể loại và ngôn ngữ.

Netflix Movie Rating Dataset bao gồm hai tập dữ liệu là: **Netflix_Dataset_Movie** và **Netflix_Dataset_Rating**.

1.1 Netflix Dataset Movie

Tập dữ liệu `Netflix_Dataset_Movie.csv` cung cấp thông tin về các bộ phim như tên và năm, bao gồm 17,770 dòng và 3 cột, mỗi dòng tương ứng cho một bộ phim theo định dạng sau:

`<Movie_ID>` `<Year>` `<Name>`

STT	Tên thuộc tính	Mô tả	Giá trị	Kiểu dữ liệu
1	Movie_ID	Mã bộ phim	Nằm trong khoảng từ 1 đến 17,770	Integer
2	Year	Năm mà bộ phim được ra mắt	Từ năm 1915 đến năm 2005	Integer
3	Name	Tên đầy đủ của bộ phim		String

1.2 Netflix Dataset Rating

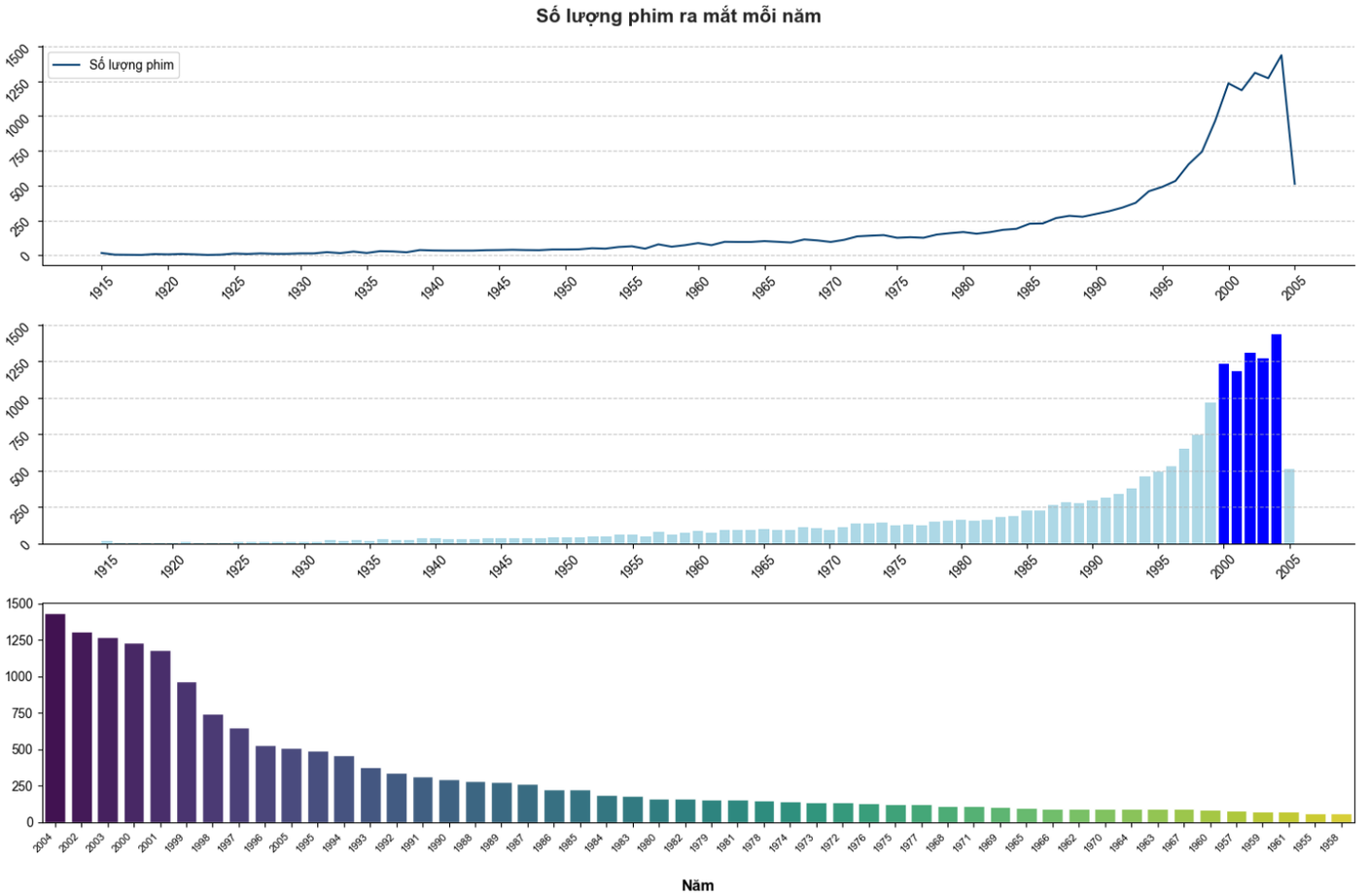
Tập dữ liệu `Netflix_Dataset_Rating.csv` cung cấp thông tin về các đánh giá của người xem về bộ phim, bao gồm 17,337,458 dòng và 3 cột, mỗi dòng tương ứng cho một đánh giá từ người xem theo định dạng sau:

`<User_ID>` `<Rating>` `<Movie_ID>`

STT	Tên thuộc tính	Mô tả	Giá trị	Kiểu dữ liệu
1	User_ID	Mã định danh người dùng	Nằm trong phạm vi từ 1 đến 2,649,429, với những khoảng trống giữa chúng. Có tổng cộng 480,189 người dùng	Integer
2	Rating	Đánh giá	Trên thang đo 5 sao, tương ứng từ 1 đến 5 sao.	Integer
3	Movie_ID	Mã bộ phim	Nằm trong khoảng từ 1 đến 17,770	Integer

2 Khám phá và phân tích dữ liệu

2.1 Câu hỏi 1: Phân tích xu hướng trong ngành công nghiệp điện ảnh qua các năm. Năm nào có số lượng phim được ra mắt nhiều nhất?

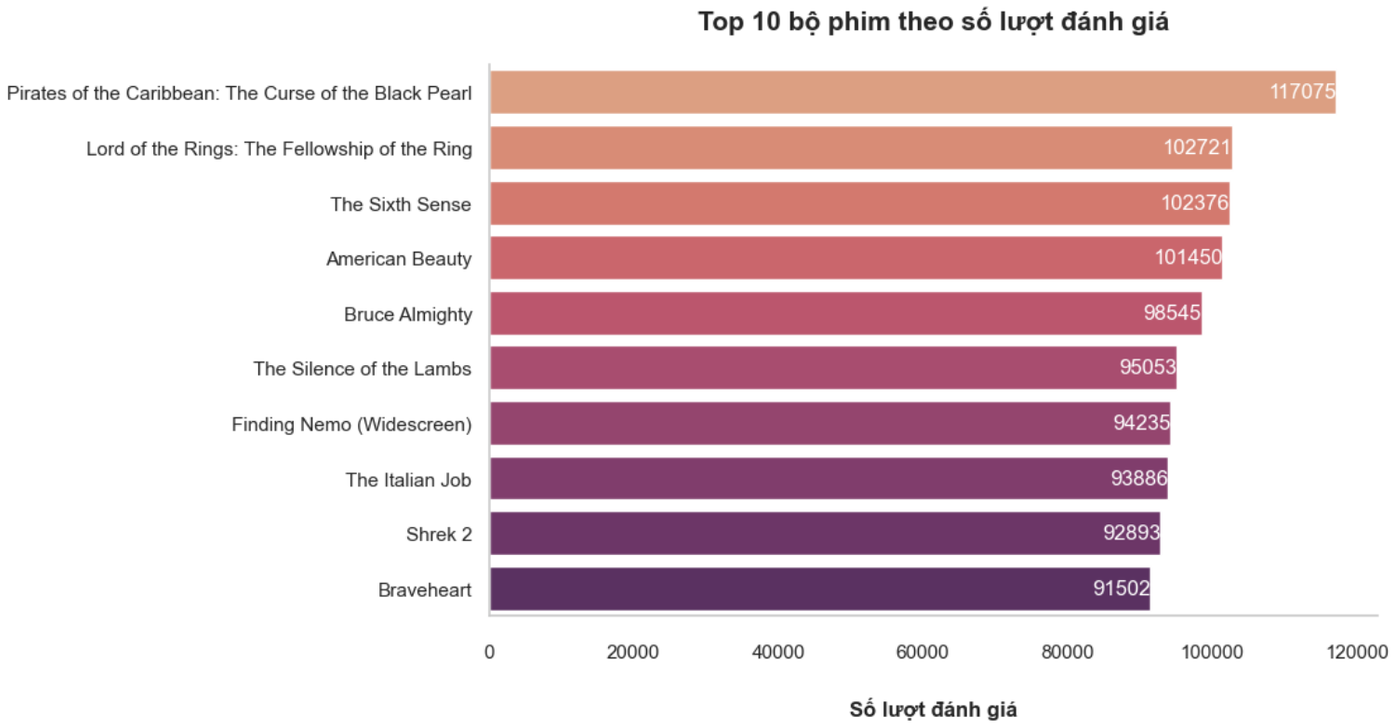


Trả lời:

- Trước hết, chúng ta có thể thấy số lượng phim ra mắt mỗi năm có xu hướng tăng dần. Điều này phản ánh sự phát triển của ngành công nghiệp điện ảnh, cũng như sự tăng trưởng của thị trường giải trí.
- Năm đạt đỉnh cao trong lịch sử sản xuất phim có thể kể đến là chuỗi 5 năm liên tiếp từ năm 2000 đến năm 2004, khi số lượng phim ra mắt vượt qua mốc 1000 bộ phim so với các năm trước, trong đó 2004 là năm có số lượng cao nhất - gần 1500 bộ phim. Điều này chỉ ra một thời kỳ sôi động và phát triển mạnh mẽ của ngành công nghiệp điện ảnh, có

thể do sự đầu tư mạnh mẽ từ các công ty sản xuất, sự hấp dẫn của thị trường và nhu cầu giải trí của công chúng.

2.2 Câu hỏi 2: Những bộ phim nào có số lượt đánh giá cao nhất ?




```
Bộ phim có thứ hạng thấp nhất:
       Movie_ID Tên bộ phim  Số lượt đánh giá
1269      4238   Inu-Yasha             1042
```

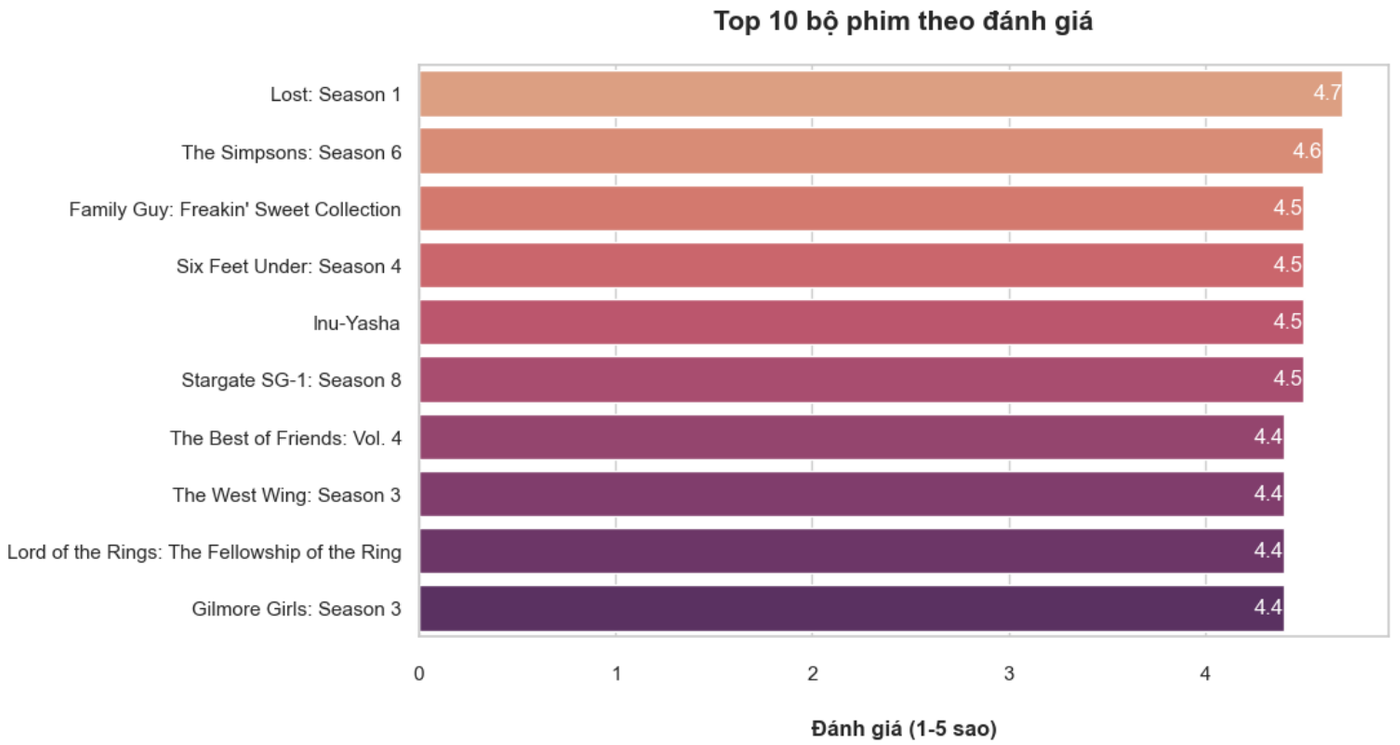
Trả lời:

- Phim **Pirates of the Caribbean: The Curse of the Black Pearl** thu hút sự quan tâm lớn nhất từ khán giả với tổng cộng **117,075**‘ lượt đánh giá, cao hơn khoảng 1,14 lần so với bộ phim xếp thứ 2 là **Lord of the Rings: The Fellowship of the Ring** với 102,721 lượt đánh giá.
- 2 bộ phim trên đã đi vào lịch sử nền điện ảnh thế giới khi bộ phim **Cướp biển vùng Caribe** quá nổi tiếng với nhân vật **Jack Sparrow** do **Johnny Depp** thủ vai và bộ phim đã đạt 5 giải thưởng trong các giải phim **Oscar**. Còn với bộ phim **Chúa tể của những chiếc nhẫn**, đây là là một trong những bộ phim đạt nhiều giải **Oscar** nhất trên thế giới.
- Từ hạng 3 đến hạng 10 lần lượt là:

3. The Sixth Sense - **102376 lượt**
4. American Beauty - **101450 lượt**
5. Bruce Almighty - **98545 lượt**
6. The Silence of the Lambs - **95053 lượt**
7. Finding Nemo (Widescreen) - **94235 lượt**
8. The Italian Job - **93886 lượt**
9. Shrek 2 - **92893 lượt**
10. Braveheart - **91502 lượt**

- Chênh lệch giữa bộ phim có thứ hạng đầu tiên là "Pirates of the Caribbean: The Curse of the Black Pearl"và bộ phim có thứ hạng cuối cùng là "Inu-Yasha"(1042 lượt đánh giá) là khá lớn, trong đó "Pirates of the Caribbean: The Curse of the Black Pearl"thu về số lượt đánh giá cao hơn, gấp 112 lần so với "Inu-Yasha".

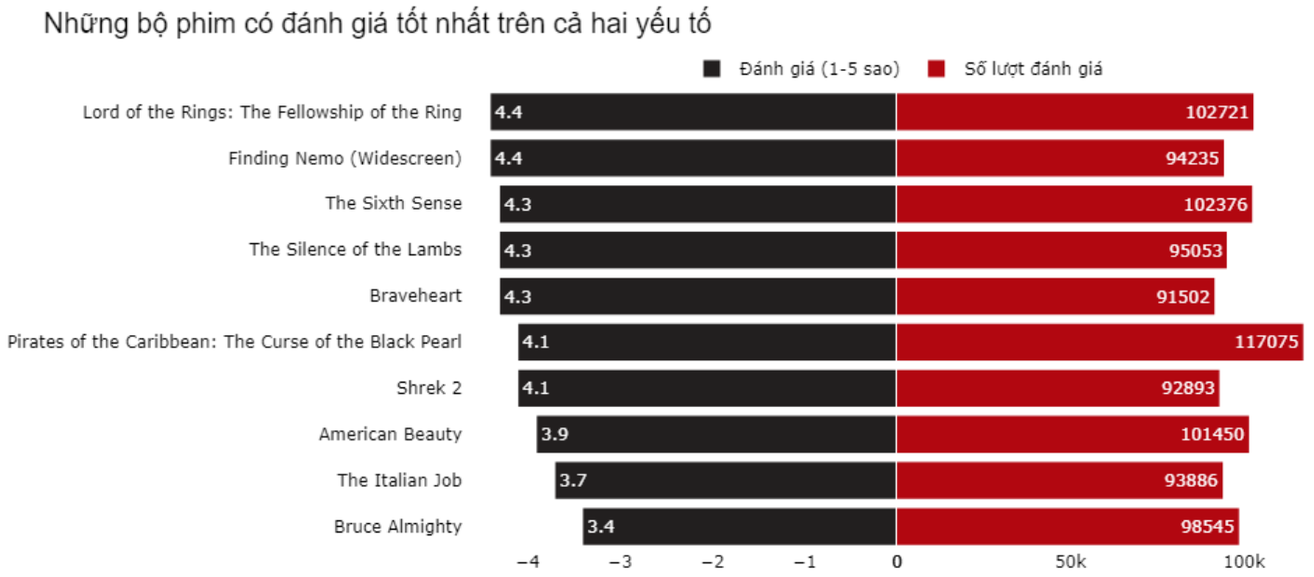
2.3 Câu hỏi 3: Những bộ phim nào có đánh giá cao nhất ? (1-5 sao)



Trả lời:

- Bộ phim được đánh giá cao nhất là **Lost: Season 1** với điểm số lên đến **4.7 sao**. Hạng 2 thuộc về bộ phim "The Simpsons: Season 6"với 4.6 sao.
- Từ hạng 3 đến hạng 10 lần lượt là:
  3. Family Guy: Freakin' Sweet Collection - **4.5 sao**
  4. Six Feet Under: Season 4 - **4.5 sao**
  5. Inu-Yasha - **4.5 sao**
  6. Stargate SG-1: Season 8 - **4.5 sao**
  7. The Best of Friends: Vol. 4 - **4.4 sao**
  8. The West Wing: Season 3 - **4.4 sao**
  9. Lord of the Rings: The Fellowship of the Ring - **4.4 sao**
  10. Gilmore Girls: Season 3 - **4.4 sao**
- Tuy nhiên, thứ hạng chỉ là một phần của bảng xếp hạng này. Mặc dù một số bộ phim có đánh giá cao, nhưng số lượng lượt đánh giá có thể không nhiều bằng những bộ phim khác. Do đó, cần phải xem xét cả hai yếu tố này để có cái nhìn toàn diện về sức hút của một bộ phim.

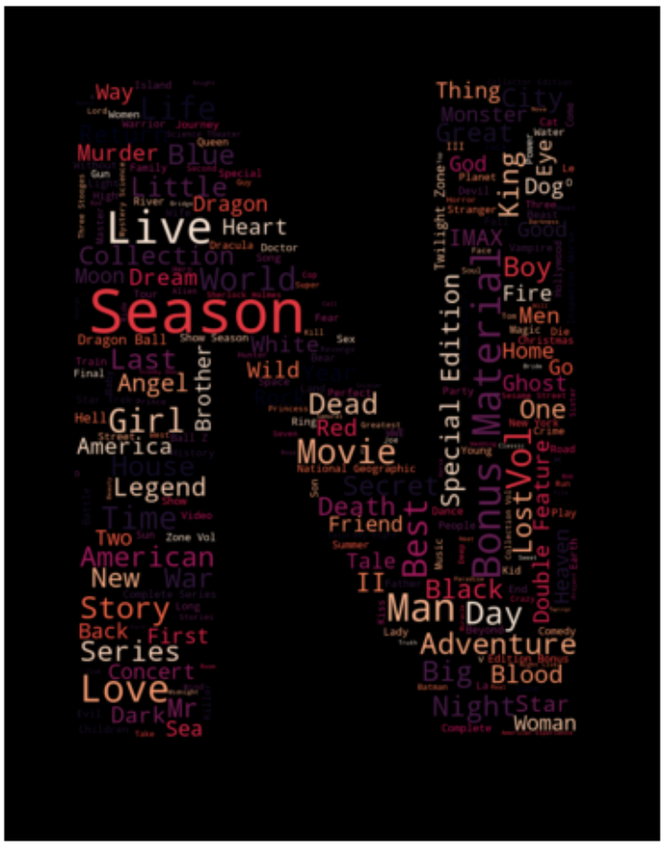
2.4 Câu hỏi 4: Dựa trên việc xem xét cả hai yếu tố về Đánh giá và Tổng số lượt đánh giá, bộ phim nào có đánh giá tốt nhất ?



Trả lời:

- Ta có thể thấy các bộ phim đã được xếp thứ hạng một cách tổng quát hơn dựa trên hai yếu tố là **đánh giá** và **số lượt đánh giá**.
- Theo đó, bộ phim có thứ hạng cao nhất theo thang đánh giá 1-5 sao thuộc về "Lord of the Rings: The Fellowship of the Ring"và "Finding Nemo (Widescreen)", với 4.4 sao và tương ứng với số lượt đánh giá lần lượt là 102,721 và 94,235 lượt.
- Tuy nhiên, nếu nhìn vào số lượng lượt đánh giá thì bộ phim "Pirates of the Caribbean: The Curse of the Black Pearl"lại chiếm vị trí hàng đầu với 117,075 lượt và đạt điểm số trung bình 4.1 sao. Điều này cho thấy sức hút lớn của bộ phim này đối với khán giả, dù điểm số trung bình có thể không cao bằng một số bộ phim khác.

2.5 Câu hỏi 5: Những từ nào xuất hiện nhiều nhất trong tên của các bộ phim ?



Insight

Các từ có kích thước lớn trong wordcloud thường là những từ xuất hiện nhiều nhất trong tên của các bộ phim. Bằng cách này, chúng ta có thể nhận ra các từ khóa phổ biến mà nhà sản xuất phim thường sử dụng để thu hút sự chú ý của khán giả.

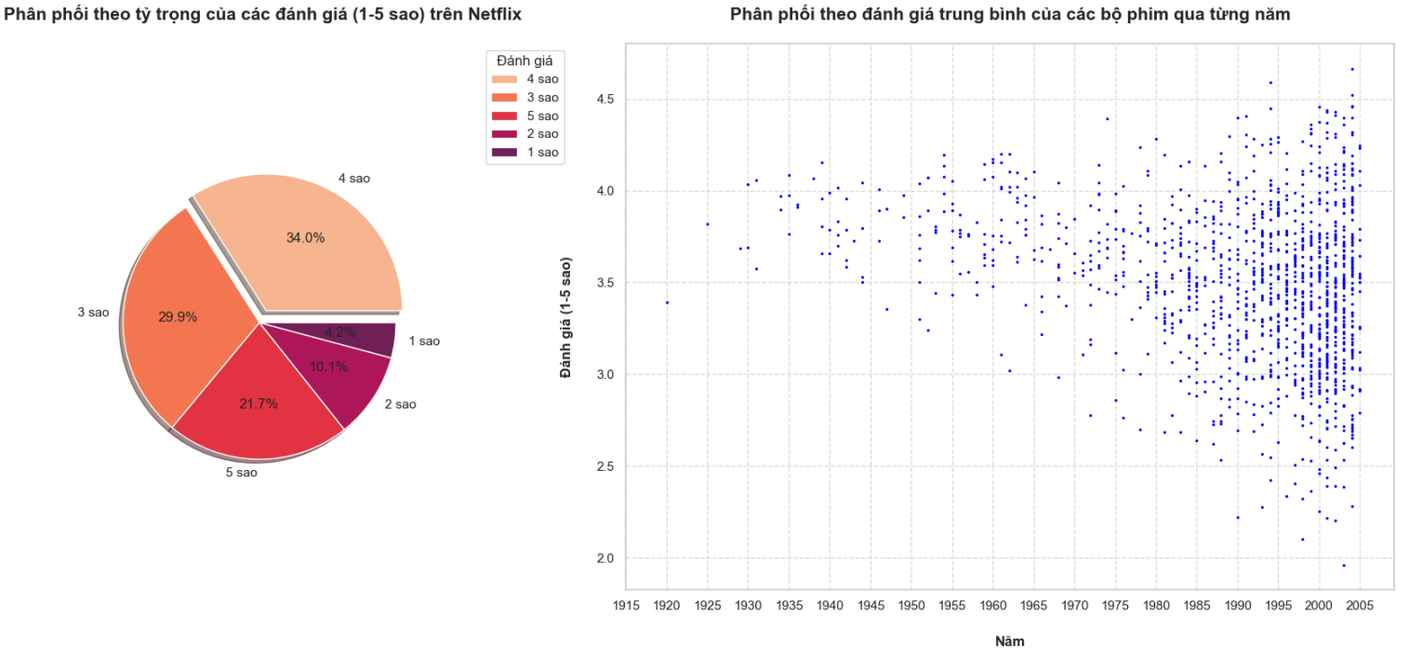
Những từ khóa phổ biến có thể kể đến như: Season, Live, Love, Vol và World,... Ta thấy những bộ phim nổi tiếng thường kéo dài qua rất nhiều mùa được đông đảo người xem ưa thích, ví dụ như phim "The Simpsons".

Nhìn chung, qua đây chúng ta có thể xác định được xu hướng, sở thích của khán giả, cũng như xu hướng mới hoặc các sự kiện nổi bật trong ngành công nghiệp điện ảnh. Bằng cách nhận diện các từ xuất hiện nhiều trong wordcloud, mọi người có thể sử dụng thông tin này để tạo nội dung tiếp thị hoặc quảng cáo phim có thể thu hút sự chú ý của các đối tượng mục tiêu.

Trả lời:

- Các từ có kích thước lớn trong wordcloud thường là những từ xuất hiện nhiều nhất trong tên của các bộ phim. Bằng cách này, chúng ta có thể nhận ra các từ khóa phổ biến mà nhà sản xuất phim thường sử dụng để thu hút sự chú ý của khán giả.
- Những từ khóa phổ biến có thể kể đến như: Season, Live, Love, Vol và World,... Ta thấy những bộ phim nổi tiếng thường kéo dài qua rất nhiều mùa được đông đảo người xem ưa thích, ví dụ như phim "The Simpsons".
- Nhìn chung, qua đây chúng ta có thể xác định được xu hướng, sở thích của khán giả, cũng như xu hướng mới hoặc các sự kiện nổi bật trong ngành công nghiệp điện ảnh. Bằng cách nhận diện các từ xuất hiện nhiều trong wordcloud, mọi người có thể sử dụng thông tin này để tạo nội dung tiếp thị hoặc quảng cáo phim có thể thu hút sự chú ý của các đối tượng mục tiêu.

2.6 Câu hỏi 6: Mỗi khoảng đánh giá (1-5 sao) lần lượt chiếm tỷ trọng bao nhiêu trong toàn bộ kết quả đánh giá từ người xem ?



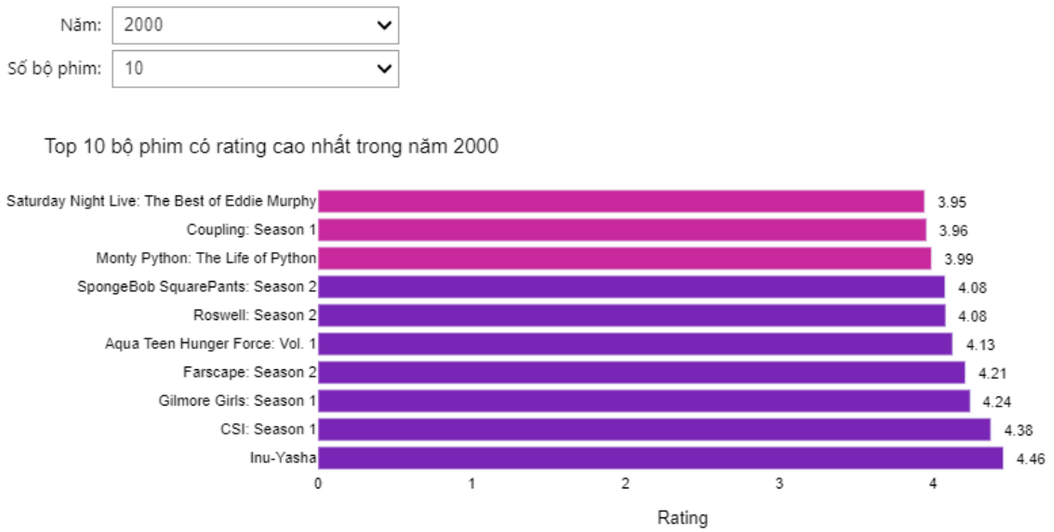
Trả lời:

- Phần lớn người xem (chiếm 34.0%) đánh giá các bộ phim với mức 4 sao, cho thấy mức độ hài lòng lớn với chất lượng và trải nghiệm xem phim.
- Trong khi đó các bộ phim được đánh giá với mức 5 - mức cao nhất, chỉ chiếm 21.7%, có thể do người xem đặt tiêu chuẩn cao hơn cho mức đánh giá cao nhất này.
- Số lượng phim được đánh giá với mức trung bình là 3 sao cũng khá phổ biến, chiếm 29.9% trên tổng số, cho thấy một phần đáng kể của người xem có ý kiến trung bình về các bộ phim này.
- Cuối cùng là các đánh giá thấp hơn - 2 sao và 1 sao - lần lượt là 10,1% và 4.2%. Tuy chỉ chiếm tỷ trọng nhỏ, nhưng qua đó cũng có thể thấy một số ít người xem không hài lòng hoặc có nhận định tiêu cực về một số bộ phim.
- Khi quan sát kĩ hơn qua từng năm *(biểu đồ bên phải)*, ta nhận thấy một tăng trưởng đáng kể trong số lượt đánh giá từ người xem trong những năm gần đây, đặc biệt là từ năm 1980 trở đi. Trước đó, đa số là những lượt đánh giá từ 3 sao đến 4.3 sao. Trong khi đó, trong những năm gần đây, phân bố này trải rộng hơn, từ các đánh giá thấp nhất là 2 sao đến những đánh giá cao nhất là hơn 4.5 sao.

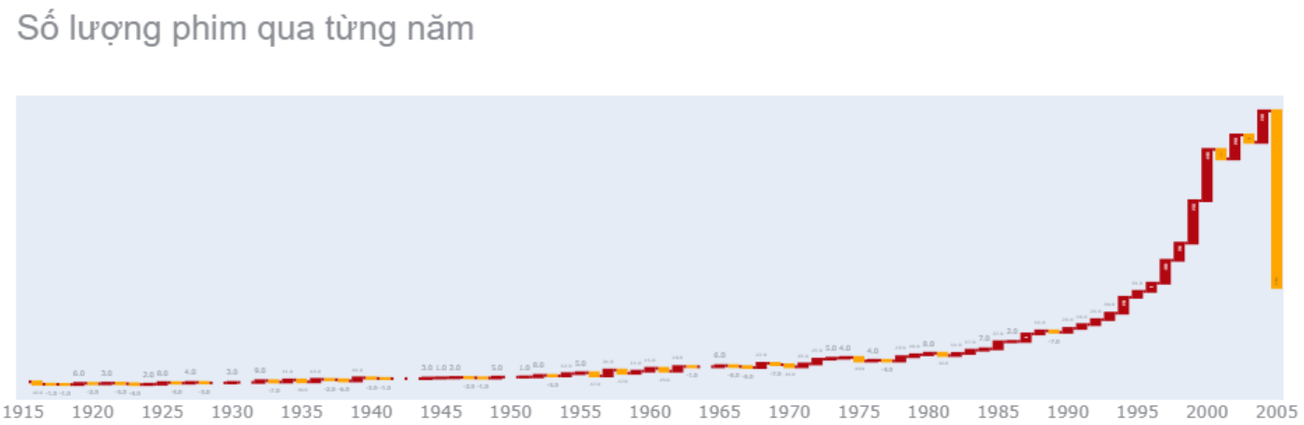
3 Trực quan dữ liệu tương tác

Để khám phá dữ liệu theo cách tự nhiên và linh hoạt hơn, từ đó tạo ra trải nghiệm tương tác thú vị và cá nhân hóa theo từng mục đích sử dụng của mỗi người, ở phần này cho phép mọi người có thể tương tác trực tiếp với các biểu đồ thông qua các tùy chọn *(dropdown options)* cũng như kéo thả thao tác trên chính biểu đồ đó.

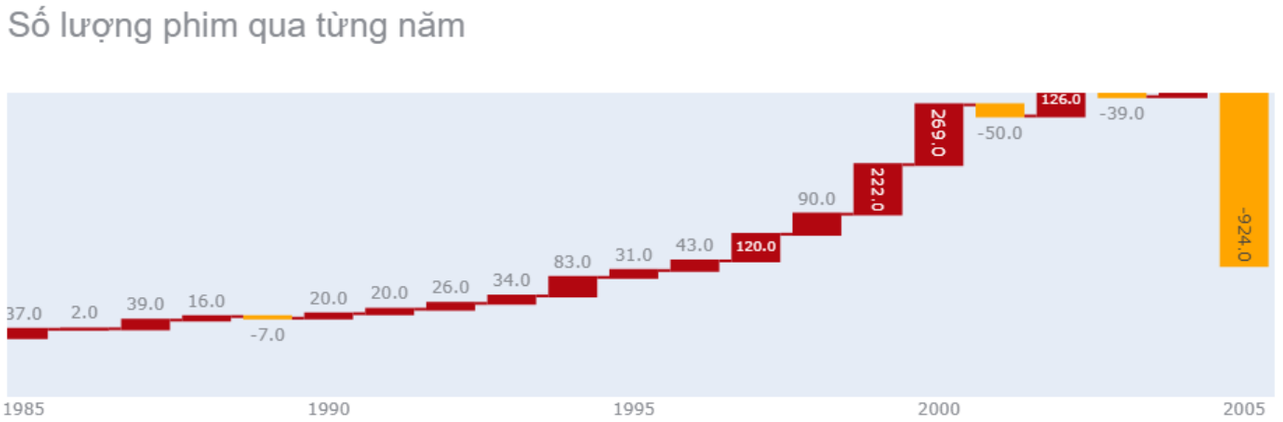
3.1 Top các bộ phim có rating cao nhất theo từng năm



3.2 Sự biến động (tăng/giảm) của số lượng phim qua từng năm



Hình 3.2.1: Từ năm 1915 đến năm 2005



Hình 3.2.2: Từ năm 1985 đến năm 2005

4 Insights

Trong quá trình khám phá toàn diện tập dữ liệu `Netflix_Movie_Rating_Dataset` bằng các công cụ trực quan phân tích như Matplotlib, Seaborn và Plotly, nhóm đã tìm hiểu sâu về vô số khía cạnh của bối cảnh nội dung trên nền tảng Netflix. Trong phần này, nhóm sẽ chia sẻ những phát hiện thú vị nhất, bao gồm các xu hướng thị trường trong ngành công nghiệp điện ảnh, sở thích và đánh giá của người xem đối với các bộ phim và chương trình truyền hình:

- **Khám phá sâu về các thị trường điện ảnh qua các thập kỷ:**
  - Phân tích tiết lộ các xu hướng của thị trường trong ngành công nghiệp điện ảnh từ những năm 1915 đến năm 2005, cũng như sở thích và đánh giá của người xem đối với các nội dung trên Netflix.
  - Qua việc xem xét xu hướng sản xuất và tiêu thụ nội dung điện ảnh qua các thập kỷ, chúng ta thấy được sự thay đổi, phát triển và ngày một tăng trưởng của ngành công nghiệp điện ảnh trong thời gian qua, nhất là liên tiếp 5 năm từ 2000-2004 số lượng phim được ra mắt luôn cao hơn so với các năm trước và vượt ngưỡng 1000 bộ phim, có năm gần chạm mức 1500 phim trong một năm.
- **Đa dạng và phân bổ của nội dung trên Netflix:**
  - Khi đi sâu hơn vào các bộ phim và xếp hạng của chúng, ta có thể thấy được sự đa dạng và phân bổ của các bộ phim và chương trình truyền hình cùng với mức độ xếp hạng khác nhau của chúng.
- **Hiểu sâu hơn về sở thích và đánh giá của người dùng:**
  - Bằng cách phân tích các đánh giá và xếp hạng từ người dùng trên Netflix, chúng ta hiểu rõ hơn về sở thích và xu hướng xem phim của họ.
  - Những bộ phim nổi tiếng vừa có điểm số đánh giá cao vừa có đông đảo lượng đánh giá từ người xem, có thể kể đến như: "Pirates of the Caribbean: The Curse of the Black Pearl"hay "Lord of the Rings: The Fellowship of the Ring", cũng như những tựa phim kéo dài nhiều mùa được ưa thích như The Simpsons mà đáng chú ý là "The Simpsons: Season 6"với 4.6 sao từ khán giả.