**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**TÊN ĐỀ TÀI (CHỮ IN HOA)**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện: | | |
| STT | Họ tên | MSSV |
| 1 | Bùi Anh Khôi | 19520649 |
| 2 | Nguyễn Đình Quốc Bảo | 19520402 |
| 3 | Nguyễn Thị Thu Hà | 19521456 |

**TP. HỒ CHÍ MINH – 12/2020**

# GIỚI THIỆU

Với nhu cầu giải trí của mỗi cá nhân ngày càng cao, điện ảnh đã, đang và vẫn tiếp tục trở thành một mảnh đất màu mỡ đem lại doanh thu khổng lồ cho các nhà đầu tư. Một bộ phim được đánh giá thành công hay không chịu ảnh hưởng từ rất nhiều yếu tố như diễn viên, đạo diễn, thời điểm ra mắt, thể loại phim, thời lượng phim. Báo cáo này trình bày những tìm hiểu của chúng tôi về những ảnh hưởng của các yếu tố lên doanh thu của phim bằng các phương pháp thống kê và trực quan dữ liệu. Sau phần phân tích, chúng tôi thực hiện xây dựng dashboard trực quan các chỉ số, các yếu tố ảnh hưởng đến doanh thu của một bộ phim.

Trong đồ án này, chúng tôi đã sử dụng các thư viện hỗ trợ sau: Dash, Plotly, Pandas, Matplotlib, Numpy, Beautiful Soup…

A picture containing chart

Description automatically generated

Hình . Quy trình thực hiện

# GIỚI THIỆU BỘ DỮ LIỆU

Bộ dữ liệu được sử dụng trong đồ án này được nhóm thu thập tự động từ trang kết quả của công cụ tìm kiếm nâng cao được tích hợp sẵn trên trang IMDB [1]. Bộ dữ liệu chứa thông tin của các phim chiếu rạp ra mắt từ năm 2010 đến năm 2020 và có tên là IMDB Feature films 2010 – 2020.

Bộ dữ liệu thô khi thu thập được có 114.432 bộ phim và 21 thuộc tính. Mô tả các thuộc tính của bộ dữ liệu thô Sau khi được xử lý và làm sạch bộ dữ liệu còn lại 22.245 bộ phim và 21 thuộc tính. Chi tiết về quá trình xử lý, làm sạch sẽ được trình bày ở phần sau.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** | **Miền giá trị** | **Missing**  **(%)** |
| 1 | title | Tiêu đề phim | String | “Halloween”, “Venom”, ... | 0% |
| 2 | casts | Dàn diễn viên chính | String | “Daniel Craig,Christoph Waltz,Léa Seydoux,” | 7.8% |
| 3 | directors | Đạo diễn | String | “Sam Mendes,”, “Bong Joon Ho,”, ... | 7.7% |
| 4 | writers | Biên kịch | String | “John Logan,Neal Purvis,Robert Wade,”, ... | 11.6% |
| 5 | genres | Các tổ hợp thể loại | String | “Action,Adventure,Thriller,”, “Comedy,”,... | 8% |
| 6 | certificate | Chứng chỉ phân loại độ tuổi | String | “C13”, “C16”, “P”, “C18”, “Unrated” | 60% |
| 7 | release\_date | Thời điểm ra mắt (năm) | Numeric | 2010-2021 | 0.4% |
| 8 | countries\_of\_origin | Quốc gia xuất xứ | String | “France,United Kingdom,”, “United States,”, ... | 5.3% |
| 9 | languages | Ngôn ngữ hỗ trợ | String | “English,French,German,Icelandic,”, “English,”, ... | 6.2% |
| 10 | production\_companies | Hãng sản xuất | String | “Big View Productions,Bigview Media,”, “Talented 10th Contributions,”, ... | 15% |
| 11 | gross\_worldwide | Doanh thu toàn cầu | Numeric | 2.000000e+00 - 2.797501e+09 | 0% |
| 12 | budget | Kinh phí sản xuất ước tính | Numeric | 2.591644e-03 -  2.591644e-03 | 71.7% |
| 13 | runtime | Thời lượng phim | Numeric | 45 - 808 | 10.2% |
| 14 | color | Màu phim | String | “Color”, “Black and White” | 9.3% |
| 15 | sound\_mix | Công nghệ âm thanh sử dụng trong phim | String | “Dolby Digital,”, “Dolby Digital,DTS,”, ... | 71.5% |
| 16 | aspect\_ratio | Tỷ lệ khung hình | String | “2.35:1”, “1.78:1”, ... | 48.5% |
| 17 | score | Điểm đánh giá (trên thang 10) | Numeric | 1 - 9.4 | 8.8% |
| 18 | votes | Số lượng người tham gia đánh giá | Numeric | 0 - 2200 | 0% |
| 19 | user\_reviews | Số lượng bình luận của người dùng | Numeric | 0 - 11100 | 0% |
| 20 | critic\_reviews | Số lượng bình luận của nhà phê bình | Numeric | 0 - 1000 | 0% |
| 21 | metascore | Điểm đánh giá của nhà phê bình | Numeric | 1 - 100 | 76.8% |

Bảng . Mô tả các thuộc tính của bộ dữ liệu

# TIỀN XỬ LÝ DỮ LIỆU

## Làm sạch dữ liệu

Các thuộc tính trong bộ dữ liệu thô được nhóm xử lý, làm sạch cụ thể như sau:

* Các thuộc tính như “**casts”** và **“writers”** thường có tiền tố **“Stars,”** hoặc **“Writers,”** ở đầu chuỗi nên sẽ được loại bỏ.
* Thuộc tính **“gross\_worldwide”** và **“budget”** mang nhiều đơn vị tiền tệ khác nhau nên sẽ được chuyển đổi về đơn vị đô la Mỹ [2].
* Thuộc tính **“release\_date”** bao gồm 2 phần là thời điểm phát hành và quốc gia phát hành, nhưng phần quốc gia lại phụ thuộc vào vị trí của địa chỉ IP dùng để truy cập nên sẽ được loại bỏ. Vì vậy nhóm chỉ giữ lại phần năm trong thời điểm phát hành.
* Tương tự như **“release\_date”**, thuộc tính **“certificate”** cũng sẽ trả về thang đo tương ứng với quốc gia phụ thuộc vào vị trí của địa chỉ IP dùng để truy cập nên nhóm quyết định chuyển về thang đo chuyển của Việt Nam (“C13”, “C16”, “P”, “C18”, “Unrated”) [3][4][5].
* Thuộc tính **“runtime”** được quy đổi về dạng số tương ứng với số phút.
* Các thuộc tính như **“user\_reviews”** và **“critic\_reviews”** được quy đổi về đơn vị người.
* Thuộc tính **“votes”** được quy đổi thành đơn vị nghìn người.
* Thuộc tính **“aspect\_ratio”** được quy đổi về dạng **[x]:1** ngoại trừ một số trường hợp đặc biệt (tondoscope, circular, cinemascope).
* Thuộc tính **“color”** được chuyển về 2 dạng là **“Color”** và **“Black and White”**.

## Điền dữ liệu khuyết

* **Bước 1:** Đầu tiên, bộ dữ liệu được xây dựng với biến mục tiêu là **gross\_worldwide** nên với các điểm dữ liệu bị thiếu giá trị thuộc tính này, chúng tôi quyết định drop ra khỏi bộ dữ liệu.
* **Bước 2:** Các thuộc tính **votes**, **user\_reviews**, **critic\_reviews** được điền khuyết với giá trị 0 - tương ứng với phim đó không có ai tham gia đánh giá hoặc không có ai tham gia bình luận nhận xét hoặc không có nhà phê bình nào tham gia nhận xét về chất lượng phim.
* Các thuộc tính còn lại, chúng tôi quyết định giữ nguyên giá trị khuyết để phân tích.

# PHÂN TÍCH THĂM DÒ

## Diễn viên (casts), đạo diễn (directors), biên kịch (writers)

Diễn viên, đạo diễn và biên kịch là ba yếu tố quan trọng và không thể thiếu trong một bộ phim. Để xem chúng có ảnh hưởng đến doanh thu của bộ phim hay không, chúng tôi đã tiến hành EDA từng thuộc tính trên.

Chart, waterfall chart

Description automatically generated

Hình . Biểu đồ phân bố doanh thu theo đạo diễn

Quan sát hình 2, dễ dàng nhận thấy doanh thu phim có sự phân bố khác nhau giữa các đạo diễn. Đạo diễn Tyler Perry có doanh thu phim cao vượt trội so với các đạo diễn khác [5]. Tương tự với các yếu tố casts và writes, doanh thu phim cũng có sự phân bố khác nhau giữa các diễn viên, biên kịch. Những bộ phim có của những nhà biên kịch giỏi (Naresh Kathooria, Tyler Perry,...) hay đạo diễn tài năng (Tyler Perry, Takashi Yamazaki,...) hay các diễn viên nổi tiếng và có năng lực thì phim sẽ có doanh thu cao. Có thể kết luận rằng các yếu tố **casts**, **directors**, **writes** có sự ảnh hưởng đến doanh thu (gross\_worldwide) của bộ phim.

## Kinh phí sản xuất ước tính (budget)

Chart, scatter chart

Description automatically generated

Hình . Biểu đồ tương quan giữa kinh phí sản xuất và doanh thu

Biểu đồ hình 2, chỉ ra sự tương quan mạnh giữa hai yếu tố **budget** và **gross\_worldwide**. Điều này là hợp lí vì một bộ phim có kinh phí lớn giúp nhà sản xuất phim dám thực hiện các ý tưởng, đầu tư nhiều cho cảnh quay, trang phục, hậu kì và đẩy mạnh chiến lược marketing. Khi đó chất lượng bộ phim được nâng cao và thu hút nhiều khán giả hơn ra rạp để xem phim.

## Doanh thu theo năm

Chart, bar chart

Description automatically generated

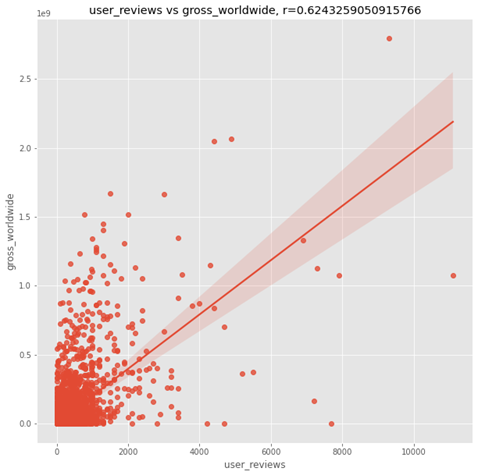
Hình . Biểu đồ phân bố doanh thu và số lượng phim ra mắt giai đoạn 2010 - nay.

Quan sát biểu đồ trên hình 3, ta thấy doanh thu phim tính theo năm trong giai đoạn 2010 đến 2019 tăng trưởng khá đều đặn. Tuy nhiên đến năm 2020 thì doanh thu phim giảm mạnh gần một nửa so với 2019 khi chỉ còn gần 1 triệu đô-la. Lý giải cho điều này là vì nửa sau năm 2020 là thời điểm bùng phát đại dịch Covid, toàn bộ mọi hoạt động từ kinh tế, chính trị, giáo dục đến giải trí đều bị trì trệ, thậm chí là tạm dừng, và doanh thu phim cũng không phải ngoại lệ.

Mặc dù doanh thu phim có sự tăng trưởng tương đối nhanh, nhưng phân bố doanh thu phim trong giai đoạn từ 2016 đến 2020 lại thấp hơn so với khoảng thời gian trước đó (2010-2015). Cụ thể vào năm 2016 chỉ còn một nửa so với 2014 và đến 2020 thì chỉ còn khoảng ⅕ so với 2014. Nguyên nhân của sự mâu thuẫn xuất phát từ việc số lượng phim tăng, nhưng doanh thu của mỗi phim trong giai đoạn từ 2016 đến 2020 lại không bằng so với thời kì từ 2010 đến 2015.

## Các số lượng đánh giá.

### Số lượng bình luận người dùng (user\_reviews)



Hình . Biểu đồ tương quan giữa số lượng bình luận (người dùng) và doanh thu

Theo quan sát hình 4, lượng userreview tập trung đa số dưới 2000 người và có sự tương quan tuyến tính với gross\_worldwide (doanh thu). Chúng tôi đã tính được hệ số tương quan của hai thuộc tính này là 0.62, con số này thể hiện có sự tương quan khá tốt giữa hai thuộc tính được nêu trên. Khi các bộ phim được nhiều lượt user\_reviews (bình luận người dùng) có nghĩa là bộ phim được nhiều người biết đến, quan tâm nên doanh thu bộ phim cao theo lượt votes.

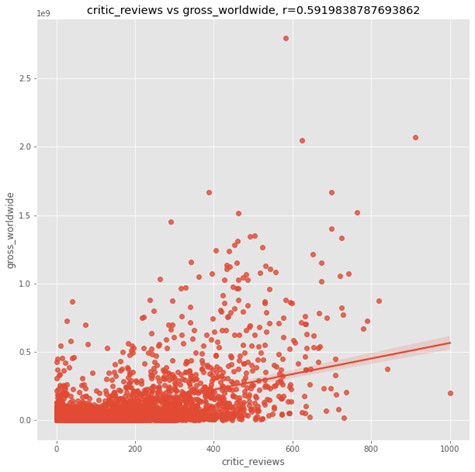
### Số lượng người tham gia đánh giá (votes)



Hình . Biểu đồ tương quan giữa số lượng đánh giá và doanh thu

Lượng votes tập trung đa số dưới 500 nghìn người và có sự tương quan tuyến tính với gross\_worldwide (doanh thu). Thường thì khi coi phim xong, thì người coi mới đánh giá phim nên khi số lượng đánh giá phim lớn cũng có nghĩa là nhiều người đã coi bộ phim này nên doanh thu bộ phim cao. Chúng tôi đã tính hệ số tượng quan giữa hai thuộc tính trên và có kết quả là 0.72 thể hiện sự tương quan tốt giữa hai thuộc tính.

### Số lượng bình luận của nhà phê bình (critic\_reviews)



Hình . Biểu đồ tương quan giữa số lượng bình luận (nhà phê bình) và doanh thu

Các bộ phim có sức ảnh hưởng thường thu hút sự quan tâm của các nhà phê bình. Các bộ phim có nhiều lượt bình luận của các phê bình (critic\_reviews) thì gross\_worldwide (doanh thu) cao.

## Doanh thu và điểm đánh giá theo thể loại phim

Chart

Description automatically generated

Hình . Biểu đồ thể hiện sự phân bố doanh thu và điểm đánh giá theo thể loại phim

Trong hai năm đầu của thập kỉ, tổ hợp thể loại phim “Comedy, Romance” và “Drama, Thriller” có doanh thu trung bình mỗi phim cao nhất trong các thị phần phim. Giai đoạn 2013-2016 doanh thu trung bình theo thể loại khá đồng đều khi những thể loại nổi bật trước đó cũng không còn quá được ưa chuộng. Nửa cuối thập kỉ từ 2017 đến 2020 chứng kiến sự bùng nổ doanh thu của thể loại “Horror” và “Horror, Thriller” [7].

# MÔ HÌNH

Trong các phần trên chúng tôi đã tìm ra các yếu tố ảnh hưởng đến doanh thu phim như đạo diễn, biên kịch, dàn diễn viên, kinh phí, điểm đánh giá cũng như là số lượt bình luận của người dùng lẫn nhà phê bình để tiến hành xây dựng mô hình dự đoán doanh thu. Dưới góc độ của một nhà sản xuất phim thì ta chỉ có thể cung cấp những thông tin có trước khi bộ phim ra mắt nên những thuộc tính có sau khi phim ra mắt như điểm đánh giá và số lượt bình luận sẽ không phù hợp. Kinh phí là một yếu tương quan cao với doanh thu tuy nhiên thuộc tính này lại bị khuyết đến 71% nên không thể sử dụng làm biến độc lập. Các thuộc tính mang yếu tố con người như đạo diễn, biên kịch, dàn diễn viên là mang rất nhiều giá trị và hầu hết đều có vai trò như nhau nên không thể áp dụng các phương pháp phân cụm hay chuyển thành dạng số vì sẽ làm tăng số chiều (one hot encode) hoặc gây ra những quan hệ không đáng có (label encode). Vì vậy nhóm quyết định không xây dựng mô hình.

Thay vì đó chúng tôi xây dựng một dashboard chứa các thông tin như total gross, total budget, … cùng với các biểu đồ so sánh doanh thu và top thể loại[[1]](#footnote-1).

# KẾT LUẬN

Phân tích và nghiên cứu bộ dữ liệu IMDB Feature Films 2010-2020 đem lại một cái nhìn tổng quan về các nhân tố ảnh hưởng đến doanh thu cuối cùng của phim. Trong báo cáo chúng tôi đã chỉ ra các yếu tố như đạo diễn, biên kịch, diễn viên có tác động đến doanh thu nhờ vào sự ảnh hưởng của họ đến khán giả. Doanh thu của các thể loại phim cũng có sự biến động theo mốc thời gian và thị hiếu của khán giả. Rõ ràng để tạo nên thành công của một bộ phim đòi hỏi nhà sản xuất phải đảm bảo và cân đối được nhiều yếu tố. Chúng tôi cũng thực hiện trả lời một số câu hỏi để tìm ra xu hướng phim trong giai đoạn gần đây, những ngôn ngữ chiếm thị phần doanh thu lớn để gợi ý cho các nhà sản xuất phim. Sau đó chúng tôi xây dựng một dashboard chứa các thông tin cần thiết để bất cứ ai nhìn vào cũng có thể nắm bắt được những yếu tố cơ bản có ảnh hưởng đến doanh thu của một bộ phim.

Mặc dù việc xây dựng mô hình để dự đoán doanh thu theo các yếu tố là bất khả thi do bộ dữ liệu không đáp ứng được, nhưng đề tài này vẫn có thể xem như một nguồn tham khảo cho các nhà sản xuất phim nhằm nâng cao chất lượng sản phẩm của mình. Trong tương lai, chúng tôi sẽ tiếp tục tìm hiểu, nghiên cứu, mở rộng bộ dữ liệu để thực hiện một số mô hình hỗ trợ việc dự đoán để hướng tới trải nghiệm giải trí tốt cho người dùng và doanh thu ổn định cho nhà làm phim.

**TÀI LIỆU THAM KHẢO**

IMDB. Link: <https://www.imdb.com/> (Ngày truy cập: 30/11/2021).

Xe. Link: <https://www.xe.com> (Ngày truy cập: 5/12/2021).

Motion Picture Association. Link: <https://www.motionpictures.org/film-ratings> (Ngày truy cập: 30/11/2021)

Thông tư số 12/2015/TT-BVHTTDL của Bộ Văn hóa, Thể thao và Du lịch.

Film Censorship Guidelines for Censors. Link: <https://www.ofnaa.gov.hk/filemanager/ofnaa/en/content_1398/filmcensorship.pdf> (Ngày truy cập: 30/11/2021)

Tyler Perry. Link: <https://tylerperry.com/tyler/story/> (Ngày truy cập 5/12/2021)

The new work time. Link: <https://www.nytimes.com/2017/10/26/movies/top-horror-movies-box-office-it-get-out.html> (Ngày truy cập: 6/12/2021)

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Bùi Anh Khôi (C) | Lên ý tưởng và kế hoạch  Thu thập dữ liệu  Xây dựng dashboard  Làm slide |
| 2 | Nguyễn Đình Quốc Bảo | Thu thập dữ liệu  Làm EDA  Thiết kế dashboard  Viết báo cáo + slide |
| 3 | Nguyễn Thị Thu Hà | Lên kế hoạch  Làm EDA  Thiết kế dashboard  Viết báo cáo + slide |

**PHỤ LỤC NỘI DUNG**

**IMDB Movie Dashboard:** <https://imdb-dashboard-ds105m11.herokuapp.com/>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu thô** | **Miền giá trị** |
| 1 | title | Tiêu đề phim | String | “Halloween”, “Venom”, ... |
| 2 | casts | Diễn viên | String | “Stars,Daniel Craig,Christoph Waltz,Léa Seydoux,” |
| 3 | directors | Đạo diễn | String | “Sam Mendes,”, “Bong Joon Ho,”, ... |
| 4 | writers | Biên kịch | String | “Writers,John Logan,Neal Purvis,Robert Wade,”, ... |
| 5 | genres | Các tổ hợp thể loại | String | “Action,Adventure,Thriller,” |
| 6 | certificate | Chứng chỉ phân loại độ tuổi | String | “TV-MA”, “IIB”, “C16”, ... |
| 7 | release\_date | Thời điểm ra mắt (năm) | String | “2018 (United States)”, “2018 (India)”, ... |
| 8 | countries\_of\_origin | Quốc gia xuất xứ | String | “France,United Kingdom,”, “United States,”, ... |
| 9 | languages | Ngôn ngữ hỗ trợ | String | “English,French,German,Icelandic,”, “English,”, ... |
| 10 | production\_companies | Hãng sản xuất | String | “Big View Productions,Bigview Media,”, “Talented 10th Contributions,”, ... |
| 11 | gross\_worldwide | Doanh thu toàn cầu | String | “$82,995”, “$1,870,392”, “₹200,000,000”, ... |
| 12 | budget | Kinh phí sản xuất ước tính | String | “€1,200,000 (estimated)”, “ HK$8,000,000 (estimated)”,... |
| 13 | runtime | Thời lượng phim | String | “1h 50min”, “1h 29min”, ... |
| 14 | color | Màu phim | String | “Color”, “Black and White”, ... |
| 15 | sound\_mix | Công nghệ âm thanh sử dụng trong phim | String | “Dolby Digital,”, “Dolby Digital,DTS,”, ... |
| 16 | aspect\_ratio | Tỷ lệ khung hình | String | “2.35 : 1”, “1.78 : 1 / (high definition)”, ... |
| 17 | score | Điểm đánh giá (trên thang 10) | String | “7.2/10”, “3.9/10”, ... |
| 18 | votes | Số lượng người tham gia đánh giá | String | “338”, “14K”, ... |
| 19 | user\_reviews | Số lượng bình luận của người dùng | String | “11.1K”, “55.0”, ... |
| 20 | critic\_reviews | Số lượng bình luận của nhà phê bình | String | “221”, “1K”, ... |
| 21 | metascore | Điểm đánh giá của nhà phê bình | Numeric | 1 - 100 |

Bảng . Mô tả các thuộc tính của bộ dữ liệu thô

Chart, waterfall chart

Description automatically generated

Hình . Biểu đồ phân bố doanh thu theo biên kịch

Chart

Description automatically generated

Hình . Biểu đồ phân bố doanh thu theo dàn diễn viên

1. IMDB Movie Dashboard: <https://imdb-dashboard-ds105m11.herokuapp.com/> [↑](#footnote-ref-1)