**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**PHÂN TÍCH CÁC THUỘC TÍNH ẢNH HƯỞNG ĐẾN DOANH THU PHIM**

|  |  |  |
| --- | --- | --- |
| 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

# 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. (Trích dẫn) 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 trước khi xử lý có 114.432 bộ phim và 21 thuộc tính. Sau khi được xử lý và làm sạch bộ dữ liệu còn lại 22245 bộ phim và 21 thuộc tính.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** | **Miền giá trị** |
| 1 | title | Tiêu đề phim | String | “Halloween”, “Venom”, ... |
| 2 | casts | Diễn viên | String | “Daniel Craig,Christoph Waltz,Léa Seydoux,” |
| 3 | directors | Đạo diễn | String | “Sam Mendes,”, “Bong Joon Ho,”, ... |
| 4 | writers | Biên kịch | String | “John Logan,Neal Purvis,Robert Wade,”, ... |
| 5 | genres | Các tổ hợp thể loại | String | “Action,Adventure,Thriller,”, “Comedy,”,... |
| 6 | certificate | Chứng chỉ phân loại độ tuổi | String | “C13”, “C16”, “P”, “C18”, “Unrated” |
| 7 | release\_date | Thời điểm ra mắt (năm) | Numeric | 2010-2021 |
| 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 | Numeric | 2.000000e+00 - 2.797501e+09 |
| 12 | budget | Kinh phí sản xuất ước tính | Numeric | 2.591644e-03 -  2.591644e-03 |
| 13 | runtime | Thời lượng phim | Numeric | 45 - 808 |
| 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”, ... |
| 17 | score | Điểm đánh giá (trên thang 10) | Numeric | 1 - 9.4 |
| 18 | votes | Số lượng người tham gia đánh giá | Numeric | 0 - 2200 |
| 19 | user\_reviews | Số lượng bình luận của người dùng | Numeric | 0 - 11100 |
| 20 | critic\_reviews | Số lượng bình luận của nhà phê bình | Numeric | 0 - 1000 |
| 21 | metascore | Điểm đánh giá của nhà phê bình | Numeric | 1 - 100 |

Bảng 1. Mô tả các thuộc tính của bộ dữ liệu

# TIỀN XỬ LÝ DỮ LIỆU

## Làm sạch dữ liệu

* Các thuộc tính như “**casts”** và **“writer”** 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. (Tài liệu tham khảo)
* 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”). (Tài liệu tham khảo)
* 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.

# Phân tích thăm dò

## Diễn viên (casts), đạo diễn (director), biên kịch (writer)

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.

(Plot)

Quan sát (plot), dễ dàng nhận thấy doanh thu phim có sự phân bố khác nhau giữa các nhà biên kịch, đạo diễn và dàn diễn viên khác nhau. Nhà biên kịch, đạo diễn Tyler Perry có doanh thu phim cao vượt trội so với các nhà biên kịch, đạo diễn khác.(Tltk) Có thể kết luận rằng các yếu tố **casts**, **director**, **writer** có sự ảnh hưởng đến doanh thu (gross\_worldwide) của bộ phim. 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.

## Kinh phí sản xuất ước tính (budget)

### Kinh phí sản xuất ước tính (budget) với doanh thu (gross\_worldwide)

(Plot)

The scatter plot of budget and gross\_worldwide, chỉ ra sự tương quan mạnh giữa hai yếu tố này với nhau. 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. Khi đó chất lượng bộ phim được nâng cao và thu hút nhiều khán giả hơn ra rạp để xem phim.

### Kinh phí sản xuất ước tính (budget) với thời lượng phim (runtime)

(plot)

Trong biểu đồ đầu “distribution of running time” cho thấy thời lượng phim trong giai đoạn này bị lệch dương khi mean là 103 phút và median là 100 phút. Có 50% phim có thời lượng ngắn hơn 103 phút, thời lượng tăng nhẹ trong khoảng từ 103 đến 200 phút, cá biệt có phim dài tới 808 phút.

Biểu đồ budget vs runtime: Như đã giải thích ở phần budget vs gross, runtime cũng tương quan mạnh với budget, bởi 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. Từ đó kéo dài thời gian phim nhằm truyền tải nhiều thông điệp nội dung hơn.

## Doanh thu theo năm

(Plot)

Quan sát biểu đồ … 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 … million dollars. 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 tăng, 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 khoảng một nửa so với 2014 và đến 2020 thì chỉ còn tầm ⅕ so với 2014tr. 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 quá lớn so với thời kì từ 2010 đến 2015.

## Các số lượng đánh giá.

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

Chart, scatter chart

Description automatically generatedTheo quan sát hình … , lượng user-review 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 tr 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

### Chart, scatter chart Description automatically generatedSố lượng người tham gia đánh giá (votes)

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.

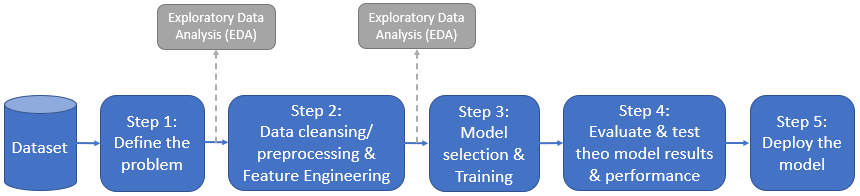
### Chart, scatter chart Description automatically generatedSố lượng bình luận của nhà phê bình (critic\_reviews)

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.

Phải có mô tả dataset trước khi tiến hành phân tích. Nếu là dataset tự thu thập thì phải trình bày phương pháp thu thập.

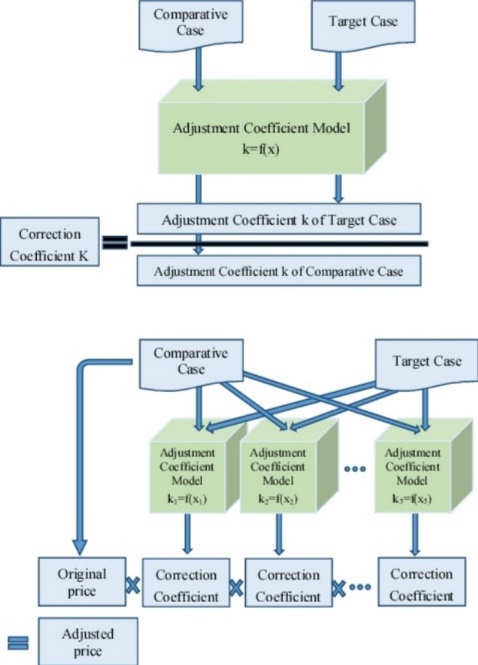
Phương pháp phân tích, có thể tiếp cận theo các các sau

* Khuyến nghị tiếp cận theo quy trình hoặc các bước thực hiện phân tích dữ liệu đã được học. Hình:



*Hình 1. Quy trình PTDL.*

* Hoặc, theo style:
  + 1. Phương pháp (Methodology)
  + 2. Thiết kế thực nghiệm (Empirical design)
  + 3. Kết quả thực nghiệm (Empirical results)
* Hoặc, theo phương pháp, hoặc style khác.
* Nên có một hình thể hiện các bước thực hiện cả quá trình. Ví dụ:



*Hình 2. Quy trình abc. Hình vừa phải không quá lớn hoặc quá nhỏ.*

## Tiêu đề Heading 2

Nội dung đoạn văn dùng Style BT

* Gạch đầu dòng này dùng **Style G1**.

## Tiêu đề Heading 2

# KẾT LUẬN

Tóm tắt lại quá trình và kết quả hiện có.

Tóm tắt kết quả đạt được?

Viết 10 dòng, hoặc nửa trang.

Viết tốt được **2 điểm**

**Sinh viên trình bày không quá 10 trang cho tổng cộng 3 phần:**

**Giới thiệu.**

**Nội dung.**

**Kết luận.**

**(Không tính số trang tài liệu tham khảo)**

**TÀI LIỆU THAM KHẢO**

Họ và tên tác giả 1, Họ và tên tác giả 2, …, Tên bài tham khảo, năm.

Tên trang web. Link: http:abc.com.vn (Ngày truy cập). Đối với tài liệu online hạn chế/ko tham khảo các blog công nghệ, wikipedia, facebook, youtube, mạng xã hội…

Trình bày không đúng style TLTK **– 1 điểm**.

Vi phạm các điều KHÔNG nên làm – 0.5 điểm

Nếu đề tài có công bố ở Hội nghị Khoa học trẻ **+ 1.5 điểm**. (Nói rõ thêm về điều cộng điểm để không thiếu nại).

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

***(Chú ý: Ghi rõ từng nhiệm vụ chi tiết của mỗi thành viên).***

Khi làm bài thu hoạch SV phải thực hiện đúng các quy định sau:

1. Sử đụng đúng Template này. Không đúng thì **– 2 điểm**.
2. Trình bày đề tài tối thiểu là 05 trang và tối đa là 10 trang A4
   * Không tính Trang Bìa, và Phụ lục phân công công việc trong template.
   * Phụ lục code
3. Copy y chan trên Internet thì gọi là đạo văn và **nhận 0 điểm** cho bài thu hoạch.
4. Nộp đúng hạn theo thông báo của GV.
5. Không trình bày code trong nội dung bài thu hoạch. Nếu muốn thì trình bày tại phụ lục.

Sản phẩm:

1. File word (thêm pdf) báo cáo, ko in ra, nộp file mềm.
2. Demo kết quả.
3. Slide thuyết trình (Nếu nhóm được chọn).