TRƯỜNG ĐẠI HỌC KINH TẾ – ĐẠI HỌC ĐÀ NẴNG
KHOA THƯƠNG MẠI ĐIỆN TỬ

****

BÁO CÁO GIỮA KỲ

DỰ ĐOÁN THÀNH CÔNG CỦA BỘ PHIM

Học phần: *Kho và khai phá dữ liệu*

Giảng viên hướng dẫn: *ThS. Nguyễn Văn Chức*

Lớp học phần: *MIS3008_47K29.1 | Nhóm: 5*

Thành viên nhóm:

1. *Nguyễn Thị Thu Huyền*
2. *Lê Thị Phương Thảo*
3. *Nguyễn Thị Hồng Nhung*

Đà Nẵng, ngày 30 tháng 10 năm 2023

MỤC LỤC

PHẦN TRĂM ĐÓNG GÓP

Họ và tên	Nhiệm vụ	Phần trăm đóng góp
Nguyễn Thị Thu Huyền	Crawl dữ liệu, tiền xử lý, xây dựng mô hình	33.34%
Lê Thị Phương Thảo	Tiền xử lý, phân tích dữ liệu	33.33%
Nguyễn Thị Hồng Nhung	Tiền xử lý, phân tích dữ liệu	33.33%

CHƯƠNG 1: TỔNG QUAN

1.1. Giới thiệu

Trong thời đại công nghệ thông tin và xu hướng xã hội hóa ngày càng gia tăng, ngành công nghiệp giải trí đang trở thành một trong những ngành phát triển nhanh nhất thế giới. Đặc biệt, phim ảnh là một yếu tố thiết yếu của giải trí quốc gia và phát triển kinh tế ở bất kỳ quốc gia nào.

Tuy nhiên, để sản xuất được một bộ phim thành công không hề dễ dàng. Ngoài yếu tố sáng tạo, kỹ năng diễn xuất của diễn viên, kỹ năng làm phim của đạo diễn, chất lượng kịch bản và các yếu tố kinh tế, xã hội, văn hóa cũng đóng vai trò quan trọng. Vì vậy, việc tạo ra các mô hình dự đoán thành công của phim có thể giúp nhà sản xuất, đạo diễn, nhà đầu tư đưa ra quyết định đúng đắn, tránh rủi ro và tối ưu hóa lợi nhuận. Các mô hình dự đoán thành công của một bộ phim có thể dựa trên các yếu tố như: thể loại phim, diễn viên, đạo diễn, chi phí sản xuất, doanh thu, ... Khi công nghệ phát triển, việc sử dụng kỹ thuật học máy và khai thác dữ liệu sẽ cải thiện độ chính xác và độ tin cậy của các mô hình dự đoán.

Vì vậy, xây dựng mô hình ***dự đoán thành công của bộ phim*** là một đề tài nghiên cứu khoa học đầy tiềm năng, có ý nghĩa thực tiễn to lớn trong việc phát triển ngành giải trí, đồng thời cũng có vai trò trong sự phát triển của ngành giải trí. Đóng góp vào sự phát triển của khoa học và công nghệ.

1.2. Đặt vấn đề

Trước đây, phim ảnh hoàn toàn dựa trên ý tưởng, khoa học dữ liệu duy nhất xuất hiện trong ngành này chỉ đơn giản là ghi lại số lượng phòng vé hoặc doanh thu. Không thể đoán trước được một bộ phim thành công hay thất bại.

Nhưng hiện nay, ngành công nghiệp điện ảnh đã thay đổi đáng kể khi chuyển sang sử dụng khoa học dữ liệu để nâng cao hiệu quả của phim. Dự đoán thành công của một bộ phim được xếp từ thất bại đến bom tấn. Sử dụng mô hình Machine Learning kết hợp với kỹ thuật phân tích mạng xã hội và khai thác văn bản, hệ thống sẽ trích xuất một số nhóm đặc điểm như diễn viên, bao gồm: ai, phim thuộc thể loại gì, phim ra mắt khi nào,…

Sự thành công của một bộ phim dựa trên 3 yếu tố:

- Đặc điểm phim: Đặc điểm phim có các yếu tố như thể loại hấp dẫn, đạo diễn tài ba, diễn viên xuất sắc, … có xu hướng thu hút sự quan tâm của khán giả và có khả năng trở thành một bộ phim thành công.

- Doanh thu và ngân sách: Doanh thu là một chỉ số quan trọng để đánh giá thành công về mặt kinh tế của một bộ phim. Một bộ phim thành công thường có doanh thu cao, vượt qua ngân sách sản xuất và mang lại lợi nhuận. Ngân sách sản xuất cũng quan trọng, vì một ngân sách lớn có thể cung cấp tài nguyên và cơ hội để tạo ra một bộ phim chất lượng cao, nhưng không phải lúc nào cũng đảm bảo thành công.
- Điểm đánh giá TMDb: Điểm đánh giá TMDb phản ánh xếp hạng và đánh giá của khán giả đối với một bộ phim cụ thể. Điểm cao cho thấy số lượng đánh giá tích cực và tích cực từ người xem cao. Xếp hạng TMDb có thể ảnh hưởng đến khả năng thu hút và quan tâm của khán giả cũng như góp phần vào thành công của bộ phim.

1.3. Mục tiêu nghiên cứu

- Nghiên cứu các yếu tố có liên quan đến thành công của một bộ phim, bao gồm các yếu tố như rating, thể loại, đạo diễn, biên kịch, diễn viên, thời lượng, đơn vị sản xuất, tỷ số rating,...
- Xây dựng mô hình dự đoán thành công của một bộ phim dựa trên dữ liệu có sẵn.
- Đánh giá hiệu suất của mô hình dự đoán và đưa ra kết quả dự kiến.

1.4. Hướng giải quyết

- Thu thập dữ liệu về các bộ phim điện ảnh từ các nguồn đáng tin cậy.
- Tiền xử lý dữ liệu bằng cách làm sạch, chuẩn hóa và tách các trường thông tin cần thiết.
- Trực quan hóa dữ liệu: Sử dụng biểu đồ và trực quan hóa dữ liệu để hiểu các mẫu và quan hệ giữa các yếu tố.
- Phân tích đặc trưng: Xác định các đặc trưng có thể ảnh hưởng đến thành công của một bộ phim.
- Xây dựng mô hình: Sử dụng mô hình machine learning để xây dựng mô hình dự đoán thành công phim dựa trên các đặc trưng đã xác định.
- Đánh giá mô hình: Đánh giá hiệu suất của mô hình bằng cách sử dụng tập dữ liệu kiểm tra hoặc phương pháp cross-validation. Điều này giúp đánh giá khả năng dự đoán của mô hình và xác định xem liệu nó có thể áp dụng cho các bộ phim mới không.
- Dự đoán và đánh giá kết quả: Sử dụng mô hình đã huấn luyện để dự đoán thành công của các bộ phim mới. Đánh giá kết quả dự đoán bằng cách so sánh kết quả dự đoán với thực tế và tính toán các chỉ số đánh giá như độ chính xác, độ phân loại chính xác, hay độ lỗi bình phương trung bình (mean squared error).

1.5. Kết quả dự kiến

Sau khi xây dựng mô hình, chúng ta có thể sử dụng nó để dự đoán thành công của các bộ phim mới dựa trên thông tin về các yếu tố liên quan. Kết quả dự đoán sẽ cung cấp một ước lượng về khả năng thành công của các bộ phim dựa trên mô hình đã được huấn luyện.

Đánh giá hiệu suất của mô hình sẽ được thực hiện bằng cách so sánh kết quả dự đoán với thực tế. Các chỉ số đánh giá như độ chính xác, độ phân loại chính xác hoặc độ lỗi bình phương trung bình (mean squared error) có thể được tính toán để đánh giá độ chính xác và độ tin cậy của mô hình.

Kết quả dự kiến của dự án này là có thể xây dựng một mô hình dự đoán thành công của một bộ phim với mức độ chính xác và tin cậy tương đối cao. Điều này có thể hỗ trợ quyết định trong việc đầu tư và phát triển các bộ phim mới, giúp tối ưu hóa khả năng thành công và đạt được hiệu quả kinh doanh tốt hơn trong ngành công nghiệp điện ảnh.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Giới thiệu về khoa học dữ liệu

Khoa học dữ liệu là lĩnh vực sử dụng các công nghệ như trí tuệ nhân tạo và học máy để khai thác và phân tích dữ liệu lớn từ nhiều nguồn khác nhau. Dữ liệu lớn đề cập đến lượng lớn dữ liệu được thu thập từ nhiều nguồn và được xử lý bằng các công cụ và kỹ thuật AI/ML để tạo ra dự đoán và hiểu biết. Khoa học dữ liệu đã cách mạng hóa nhiều lĩnh vực công nghiệp, từ CNTT đến chăm sóc sức khỏe và chính phủ, bằng cách cung cấp quy trình kinh doanh đổi mới và trực quan.

Trong ngành điện ảnh, khoa học dữ liệu được áp dụng để sử dụng dữ liệu lớn và phân tích dự đoán để cung cấp trải nghiệm người dùng cá nhân hóa, dự đoán sở thích của khách hàng, tối ưu hóa nội dung và giảm thiểu chi phí. Bằng cách sử dụng dữ liệu và phân tích, ngành điện ảnh có thể tạo ra nhiều dự đoán khác nhau nhằm tối ưu hóa lợi nhuận.

2.2. Cách khoa học dữ liệu ứng dụng trong lĩnh vực điện ảnh

2.2.1. Hỗ trợ quá trình sản xuất phim

Trong ngành điện ảnh, các hãng phim lớn đã sử dụng mô hình học máy để xác định xu hướng và thể loại phim mà khán giả muốn xem, từ đó định hướng phát triển kịch bản. Họ cũng sử dụng phân tích dữ liệu để so sánh với các bộ phim bom tấn trước đó, nhằm xác định điểm cốt truyện và điều chỉnh nhịp độ cốt truyện để đảm bảo thành công.

Phân tích dữ liệu cũng đóng vai trò quan trọng trong quá trình quay phim để tối ưu hóa chi phí. Nhà sản xuất có thể sử dụng phân tích dự đoán để quyết định diễn viên nào phù hợp với vai diễn nào và phần nào của cốt truyện cần được nhấn mạnh. Những quyết định này được đưa ra dựa trên mối tương quan giữa doanh thu phòng vé, tiền lương và phản ứng trên mạng xã hội đối với các diễn viên cụ thể. Ngoài ra, phân tích dữ liệu cũng giúp xác định địa điểm quay phim tốt nhất và các yếu tố khác như giờ ban ngày và khí hậu để tiết kiệm chi phí.

Trong quá trình phân phối và tiếp thị, phân tích dữ liệu là cơ hội để xác định khán giả lý tưởng cho phim. Các hãng phim sử dụng mô hình phân cụm dựa trên nhân khẩu học, nội dung người xem quan tâm trên mạng xã hội và lịch sử xem phim của họ để hiểu rõ hơn về khán giả và tìm cách tiếp cận họ một cách hiệu quả.

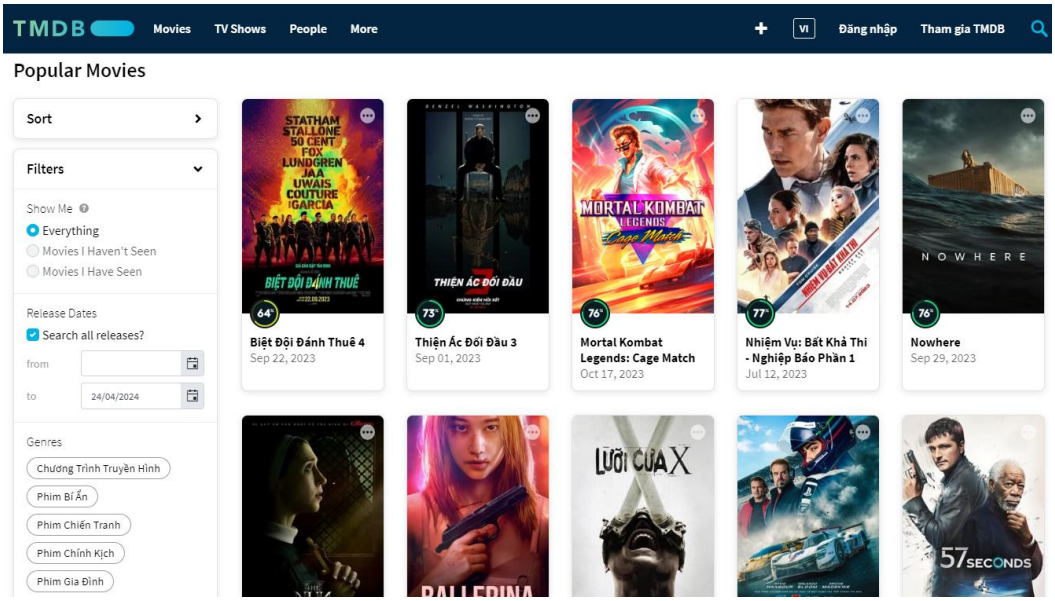
2.2.2. Đề xuất phim trên các nền tảng phát trực tuyến

Đây là một trong những trường hợp điển hình của Netflix khi nó phát triển từ dịch vụ DVD thành nền tảng phát trực tuyến lớn trên toàn cầu. Chính việc ứng dụng khoa học dữ liệu đã khiến Netflix trở nên nổi tiếng với hệ thống đề xuất phim.

Thu thập dữ liệu người dùng dựa trên thể loại, diễn viên, năm phát hành hoặc thời gian trong ngày khách hàng sử dụng. Tất cả dữ liệu này được xử lý bằng thuật toán để luôn gợi ý cho khách hàng bộ phim phù hợp ngay sau khi xem bộ phim trước.

2.2.3. Trang Web nhóm sử dụng để crawl

TMDB (The Movie Database) là một cơ sở dữ liệu trực tuyến về phim và chương trình truyền hình. Cung cấp thông tin chi tiết về các bộ phim, bao gồm thông tin về diễn viên, đạo diễn, thể loại, năm sản xuất, đánh giá, hình ảnh và nhiều nội dung khác.



2.3. Thư viện hỗ trợ

Sử dụng các thư viện có sẵn trong Python bao gồm Requests, beautifulsoup, selenium, scrapy and pandas

- Requests: bước đầu tiên trong bất kỳ quy trình cào dữ liệu nào cũng là để yêu cầu HTTP đến máy chủ web để hiển thị dữ liệu được hiển. Đây là thư viện python dùng để đơn giản hóa quá trình gửi yêu cầu từ HTTP tới một URL được chỉ định.
- bs4: beautifulsoup cho phép người dùng phân tích cú pháp HTML một cách thuận tiện. Vì beautifulsoup chỉ có thể phân tích cú pháp dữ liệu và không thể truy xuất các trang web nên nó thường được sử dụng với thư viện requests.
- pandas: được xây dựng trên thư viện NumPy cung cấp các cấu trúc dữ liệu và toán tử khác nhau để thao tác dữ liệu số.
- lxml: một thư viện python để xử lý/ phân tích cú pháp XML và HTML.
- Thư viện Selenium là một trong những công cụ kiểm thử phần mềm tự động mã nguồn mở (open source test automation tool) mạnh nhất hiện nay cho việc kiểm thử ứng dụng Web.

Selenium script có thể chạy được trên hầu hết các trình duyệt như IE, Mozilla FireFox, Chrome, Safari, Opera; và hầu hết các hệ điều hành như Windows, Mac, Linux. Selenium Python bindings cung cấp một API đơn giảm để viết functional/ acceptance test sử dụng selenium webdriver. Thông qua Selenium Python API bạn có thể truy cập tất cả các chức năng của selenium webdriver một cách trực quan. Selenium Python bindings cung cấp một API thuận tiện để truy cập Webdrivers như Firefox, IE, Chrome, Remote. Hiện tại hỗ trợ Python version 2.7,3.2.

2.4. Mô hình

2.4.1. XGB regressor

Mô hình XGBRegressor là một mô hình học máy dựa trên cây quyết định được sử dụng cho các bài toán hồi quy (regression). XGBRegressor là viết tắt của "Extreme Gradient Boosting Regressor" và nó là một thành phần của thư viện XGBoost, một thư viện học máy phổ biến và mạnh mẽ.

XGBoost được phát triển dựa trên kỹ thuật boosting, trong đó các cây quyết định yếu (weak decision trees) được xây dựng và kết hợp để tạo ra một mô hình dự đoán mạnh mẽ. XGBRegressor sử dụng thuật toán gradient boosting để tối ưu hóa hàm mất mát và xây dựng cây quyết định theo cách tuần tự. Nó kết hợp các cây quyết định để tạo ra một mô hình ensemble (tổ hợp) có khả năng dự đoán mạnh mẽ hơn so với một cây quyết định đơn lẻ.

Một số đặc điểm và lợi ích của XGBRegressor bao gồm:

- Hiệu suất cao: XGBRegressor được thiết kế để có hiệu suất cao với tốc độ và khả năng mở rộng tốt. Nó sử dụng các kỹ thuật tối ưu hóa và khai thác song song để tăng tốc quá trình huấn luyện và dự đoán.
- Xử lý dữ liệu thiếu: XGBRegressor có khả năng xử lý dữ liệu thiếu tự động. Bạn không cần phải lo lắng về việc điền giá trị thiếu trong dữ liệu vì XGBRegressor có khả năng tự động tìm cách xử lý và sử dụng các giá trị thiếu đó trong quá trình học.
- Điều chỉnh tham số linh hoạt: XGBRegressor cung cấp nhiều tham số để điều chỉnh và tinh chỉnh mô hình. Việc điều chỉnh các tham số này có thể giúp cải thiện hiệu suất và ổn định của mô hình.
- Tích hợp với Python và các thư viện học máy phổ biến: XGBRegressor là một phần của thư viện XGBoost và nó tương thích tốt với Python. Nó cũng tương thích với các thư viện học máy phổ biến như scikit-learn, cho phép bạn sử dụng và kết hợp với các công cụ và chức năng khác của Python và scikit-learn.

2.4.2. Linear Regression

Mô hình Linear Regression là một mô hình học máy sử dụng trong bài toán hồi quy (regression) để tìm mối quan hệ tuyến tính giữa các biến đầu vào và biến đầu ra.

Các đặc điểm và ưu điểm của mô hình Linear Regression:

- Đơn giản và dễ hiểu: Linear Regression là một mô hình đơn giản và dễ hiểu, với giả định về mối quan hệ tuyến tính giữa biến đầu ra và biến đầu vào. Điều này làm cho nó dễ áp dụng và giải thích kết quả dự đoán.
- Tính tường minh: Linear Regression cho phép chúng ta hiểu rõ các hệ số của mô hình, tức là mức độ ảnh hưởng của từng biến đầu vào đến biến đầu ra. Điều này giúp chúng ta đưa ra những phán đoán và giải thích về mối quan hệ giữa các biến.
- Tốc độ huấn luyện nhanh: Vì Linear Regression có một công thức toán học đơn giản để tính toán các hệ số tối ưu, quá trình huấn luyện mô hình nhanh chóng và hiệu quả, đặc biệt là đối với tập dữ liệu có kích thước lớn.
- Tính ổn định: Mô hình Linear Regression ít nhạy cảm với nhiễu và dữ liệu ngoại lai, đồng thời cũng ít bị overfitting (quá khớp) so với các mô hình phức tạp hơn. Điều này làm cho Linear Regression trở thành một lựa chọn tốt cho các tập dữ liệu có độ phức tạp thấp.

2.4.3. Random Forest Regressor

Mô hình Random Forest Regressor là một mô hình học máy được sử dụng trong bài toán hồi quy (regression) để dự đoán giá trị của một biến đầu ra dựa trên các biến đầu vào. Nó là một biến thể của mô hình Random Forest được áp dụng cho bài toán hồi quy.

Random Forest Regressor là một mô hình dự đoán dựa trên việc kết hợp nhiều cây quyết định (decision trees). Mô hình này tạo ra một tập hợp các cây quyết định ngẫu nhiên, và kết quả dự đoán cuối cùng được tính bằng cách lấy trung bình (đối với hồi quy) hoặc trung vị (đối với hồi quy loại hạng mục) của các dự đoán từ các cây quyết định thành viên. Quá trình này giúp cải thiện độ chính xác và ổn định của mô hình.

Các đặc điểm và ưu điểm chính của mô hình Random Forest Regressor:

- Độ chính xác cao: Random Forest Regressor thường cho kết quả dự đoán chính xác và ổn định do sự kết hợp của nhiều cây quyết định. Việc lấy trung bình hoặc trung vị của nhiều dự đoán từ các cây thành viên giúp giảm thiểu sai số và nhiễu có thể xuất hiện trong một cây quyết định đơn lẻ.
- Xử lý tốt dữ liệu nhiễu: Random Forest Regressor có khả năng xử lý tốt dữ liệu nhiễu, vì

nó sử dụng một tập hợp các cây quyết định và kết quả dự đoán cuối cùng được tính dựa trên đa số phiếu bầu từ các cây thành viên. Điều này giúp làm giảm tác động của các điểm dữ liệu nhiễu đến kết quả dự đoán.

- Khả năng xử lý các biến đầu vào phức tạp: Random Forest Regressor có khả năng xử lý các biến đầu vào phức tạp, bao gồm cả biến hạng mục và biến liên tục. Nó có thể tự động xử lý các biến hạng mục bằng cách thực hiện chia nhánh dựa trên các giá trị của biến hạng mục.
- Khả năng ứng dụng trong các bài toán lớn: Mô hình Random Forest Regressor có thể được áp dụng cho các bài toán với tập dữ liệu lớn. Các cây quyết định trong mô hình có thể được xây dựng song song, giúp tăng tốc quá trình huấn luyện và dự đoán trên các hệ thống tính toán song song.

CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Thu thập dữ liệu

Dữ liệu sẽ được lấy về từ các trang web:

– Crawl các url từ đường link https://www.themoviedb.org/movie. Thu được 39176 link phim và lưu vào file *link_movie.txt*.

– Từ file *link_movie.txt* có được các link bộ phim và crawl thông tin của từng bộ phim. Các thông tin của phim sẽ được crawl như là: tiêu đề, rating, thể loại, thời lượng, đạo diễn, link thông tin đạo diễn, nội dung phim, diễn viên tham gia phim, link thông tin diễn viên, ngân sách làm phim, doanh thu phim, … Thu được 39176 dòng x 28 cột và lưu vào file *Movie.csv*

– Từ các cột link thông tin đạo diễn và diễn viên, gộp và lọc các hàng trùng thành list link thông tin và crawl thông tin của đạo diễn và diễn viên từ các link đó. Các thông tin sẽ được crawl như là: tên, tiểu sử, số phim tiêu biểu, các phim tiêu biểu, năm ra mắt, giới tính, số điểm đánh giá người biết đến, … Thu được 5672 dòng x 14 cột và lưu vào file *Cast.csv*

3.2. Tiền xử lý dữ liệu

3.2.1. Tiền xử lý dữ liệu

– Vì dự đoán thành công phim dựa vào doanh thu và rating của phim nên xóa các dòng không có giá trị ở cột revenue, budget.

```
df=movie[(movie['Budget']!='-') & (movie['Revenue']!='-')]
df['tagline'] = df['tagline'].fillna('')
df['content'] = df['tagline'] + '\n' + df['overview']
df=df[['url', 'title', 'on_streaming_date', 'genre', 'certification', 'runtime', 'rating', 'content', 'Director',
       'cast1', 'cast1_link', 'cast2', 'cast2_link', 'cast3', 'cast3_link', 'Status', 'Original Language', 'Budget', 'Revenue', 'keyword']]
df.info()
```

– Lấy các cột cần thiết 'url', 'title', 'on_streaming_date', 'genre', 'certification', 'runtime', 'rating', 'content', 'director', 'cast1', 'cast1_link', 'cast2', 'cast2_link', 'cast3', 'cast3_link', 'status', 'original language', 'budget', 'revenue', 'keyword'

– Xóa các giá trị null (vì đây là những biến định tính nên không thể điền giá trị vào) và các dòng trùng.

```
df=df.dropna(axis=0)
df=df.drop_duplicates()
df = df.reset_index(drop=True)
df
```

– Xóa dấu '$' trong cột budget và revenue.

Budget	Revenue
$27,000,000.00	$7,164,778.00

```
for i in range(len(df)):
    df['Budget'][i] = df['Budget'][i][1:]
    df['Revenue'][i] = df['Revenue'][i][1:]
```

– Chuyển đổi giá trị cột budget và revenue từ object sang float và chuyển cột on_streaming_date sang kiểu date.

```
from datetime import datetime
df['on_streaming_date'] = pd.to_datetime(df['on_streaming_date'], format='%m/%d/%Y')
df['Budget'] = df['Budget'].str.replace(',', '').astype('float')
df['Revenue'] = df['Revenue'].str.replace(',', '').astype('float')
```

– Chuyển giá trị cột runtime từ object sang int bằng theo phút.

```
def convert_to_minutes(time_string):
    if len(time_string) < 5:
        hours = '0'
        minutes = time_string
    else:
        hours, minutes = time_string.split('h ')
    minutes = minutes[:-1]  # Loại bỏ ký tự 'm' cuối cùng
    total_minutes = int(hours) * 60 + int(minutes)
    return total_minutes
df['runtime'] = df['runtime'].apply(convert_to_minutes)
```

– Xóa các giá trị ngoại lai ở các cột runtime.

```
from scipy import stats
z_scores = stats.zscore(df['runtime'])
df = df[(z_scores < 3) & (z_scores > -3)]
```

```
df.describe()
```

	runtime	rating	Budget	Revenue
count	3519.000000	3519.000000	3.519000e+03	3.519000e+03
mean	105.034953	63.982097	3.974342e+07	1.106224e+08
std	22.653410	8.432083	4.871533e+07	2.080399e+08
min	2.000000	19.000000	1.000000e+00	1.000000e+00
25%	94.000000	59.000000	9.000000e+06	6.991232e+06
50%	103.000000	64.000000	2.200000e+07	3.664284e+07
75%	116.000000	69.000000	5.000000e+07	1.154525e+08
max	339.000000	100.000000	4.600000e+08	2.923706e+09

```
df.describe()
```

	runtime	rating	Budget	Revenue
count	3438.000000	3438.000000	3.438000e+03	3.438000e+03
mean	106.558464	63.767307	3.955899e+07	1.092581e+08
std	16.863447	8.140458	4.799978e+07	2.032384e+08
min	39.000000	19.000000	4.000000e+00	1.000000e+00
25%	94.000000	59.000000	9.000000e+06	7.124196e+06
50%	103.000000	64.000000	2.200000e+07	3.711138e+07
75%	116.000000	69.000000	5.000000e+07	1.153248e+08
max	172.000000	100.000000	3.790000e+08	2.923706e+09

– Thay đổi các giá trị cột ‘director_link’, ‘cast1_link’, ‘cast2_link’, ‘cast3_link’ thành các giá trị số và đổi tên cột thành ‘director_id’, ‘cast1_id’, ‘cast2_id’, ‘cast3_id’ để làm khóa chung của 2 bảng movie và cast:

```
movies['Director_link'] = movies['Director_link'].str.extractall('(\d+)').groupby(level=0)[0].apply(''.join)
movies['cast1_link'] = movies['cast1_link'].str.extractall('(\d+)').groupby(level=0)[0].apply(''.join)
movies['cast2_link'] = movies['cast2_link'].str.extractall('(\d+)').groupby(level=0)[0].apply(''.join)
movies['cast3_link'] = movies['cast3_link'].str.extractall('(\d+)').groupby(level=0)[0].apply(''.join)
movies
```

```
movies = movies.rename(columns={'Director_link': 'Director_id','cast1_link': 'cast1_id','cast2_link': 'cast2_id','cast3_link': 'cast3_id'})
```

– Sau khi xử lý lưu vào file “movie_model.csv”:

```
movies.to_csv('Movie_model.csv', index=False)
```

3.2.2. Mô tả dữ liệu

Bảng **Movie** có 3438 dòng x 21 cột

STT	Tên cột	Mô tả	Giá trị
1	url	Đường link bộ phim	
2	title	Tên bộ phim	Rush, Aline, Ma,…
3	on_streaming_date	Ngày phát hành bộ phim	từ năm 2000 - 2023
4	genre	Thể loại của bộ phim	Drama, Thriller, Crime, Mystery,...
5	certification	Ký hiệu phân loại phim	PG-13, 15, R, U, TP,...
6	runtime	Thời lượng của bộ phim	39’ - 172’
7	rating	Đánh giá của bộ phim	19 - 100
8	content	Nội dung	Get Home Safe, Your mind will not accept a game this big, …
9	Director	Đạo diễn	Jon Turteltaub, Michael Winterbottom, …
10	Director_id	Mã thông tin đạo diễn	
11	cast1	Diễn viên 1	Charlyne Yi, Virginie Ledoyen, …
12	cast1_id	Mã thông tin diễn viên 1	
13	cast2	Diễn viên 2	Lou Doillon, Demetri Martin, …
14	cast2_id	Mã thông tin diễn viên 2	
15	cast3	Diễn viên 3	Rob Riggle, Patricia Clarkson, …
16	cast3_id	Mã thông tin diễn viên 3	
17	Status	Trạng thái	Released
18	Original Language	Ngôn ngữ gốc	English, French, …
19	Budget	Ngân sách của bộ phim	$4 - $379000000
20	Revenue	Doanh thu của bộ phim	$1 - $2923706026
21	keyword	Từ khóa	“baseball”, “money”,..

Bảng **Cast** có 5672 dòng x 13 cột

STT	Tên cột	Mô tả
1	url	Đường link
2	title	Tên
3	biography	Tiểu sử
4	domestic_movie	Các bộ phim ấn tượng
5	number_domestic	Số bộ phim ấn tượng
6	launch_year	Năm ra mắt
7	Known For	Được biết đến với vai trò
8	Known Credits	Điểm đánh giá
9	Gender	Giới tính
10	Birthday	Ngày sinh
11	Place of Birth	Nơi sinh
12	Day of Death	Ngày mất
13	id	Mã số id

3.3. Xây dựng mô hình

3.3.1. Xử lý dữ liệu trước khi đưa vào mô hình

– Đọc dữ liệu 2 bảng movies và casts bằng pd.read_csv

```
movies = pd.read_csv('Movie_model.csv')
casts = pd.read_csv('Cast.csv')
```

– Tạo cột ‘id’ là mã phim bằng cách lấy id từ các đường link phim

```
movies['id'] = movies['url'].str.extractall('(\d+)').groupby(level=0)[0].apply(''.join)
```

– Chuyển giá trị ‘on_streaming_date’ sang kiểu dữ liệu datatime và tách thành 3 cột ngày, tháng, năm lần lượt có tên cột là ‘on_streaming_day’, ‘on_streaming_month’, ‘on_streaming_year’

```
movies['on_streaming_date'] = pd.to_datetime(movies['on_streaming_date'])
movies['on_streaming_month'] =movies['on_streaming_date'].dt.month
movies['on_streaming_year'] =movies['on_streaming_date'].dt.year
movies['on_streaming_day'] =movies['on_streaming_date'].dt.day
movies.head(5)
```

– Nối các thông tin cần để dự đoán trong bảng Cast vào bảng Movie dựa theo các id của Director, cast1, cast2, cast3

```
def cast(df,df1,col):
    df = df.merge(df1[['number_domestic','Known Credits','id']], left_on=col, right_on='id')
    df = df.drop('id_y', axis=1)
    df = df.rename(columns={'number_domestic': col[:-2] + 'number_domestic','Known Credits': col[:-2]+'known_credits','id_x': 'id'})
    return df
movies = cast(movies,casts,'Director_id')
movies = cast(movies,casts,'cast1_id')
movies = cast(movies,casts,'cast2_id')
movies = cast(movies,casts,'cast3_id')
movies.info()
```

– Tách các giá trị trong cột ‘genre’ và lưu vào cột ‘TL’ sau đó tạo dummy cho cột ‘TL’ mới tạo ra và cột ‘Original Language’

```
movies['TL'] = movies['genre'].str.split(', ')
dummy_df = pd.get_dummies(movies['TL'].apply(pd.Series).stack()).sum(level=0)
movies = pd.concat([movies, dummy_df], axis=1)
movies = pd.get_dummies(movies, columns=['Original Language'])
movies.head(5)
```

– Chuyển giá trị của Budget và Revenue về cùng tỷ giá hối đoái của năm 2023 và lưu vào 2 cột ‘Adjusted_Budget’ và ‘Adjusted_Revenue’

```
movies['Adjusted_Budget'] = ((((2023-movies['on_streaming_year'])*0.0322)+1)*movies['Budget'])
movies['Adjusted_Revenue'] = ((((2023-movies['on_streaming_year'])*0.0322)+1)*movies['Revenue'])
```

– Lấy các cột cần dùng trong mô hình

```
model = movies[['id', 'on_streaming_year', 'on_streaming_month', 'on_streaming_day', 'Action', 'Adventure', 'Animation', 'Comedy', 'Crime',
       'Documentary', 'Drama', 'Family', 'Fantasy', 'History', 'Horror', 'Music', 'Mystery', 'Romance', 'Science Fiction', 'TV Movie',
       'Thriller', 'War', 'Western', 'runtime', 'rating', 'Director_number_domestic', 'Director_known_credits', 'cast1_number_domestic',
        'cast1_known_credits', 'cast2_number_domestic', 'cast2_known_credits', 'cast3_number_domestic', 'cast3_known_credits',
        'Original Language_English', 'Original Language_French', 'Original Language_Spanish; Castilian','Adjusted_Budget','Adjusted_Revenue']]
model.head(5)
```

	id	on_streaming_year	on_streaming_month	on_streaming_day	Action	Adventure	Animation	Comedy	Crime	Documentary	...	cast1_known_credits	cast2_number_domestic	cast2_known_credits	cast3_number_dom
0	10851	2005	9	22	0	0	0	0	1	0	...	66	8	140	
1	12771	2002	8	23	0	0	0	1	0	0	...	73	8	59	
2	2312	2008	1	11	1	1	0	0	0	0	...	66	8	140	
3	10866	2001	10	5	0	0	0	0	0	0	...	52	8	94	
4	3489	2007	12	24	0	0	0	0	1	0	...	125	8	84	

5 rows × 38 columns

3.3.2. Xây dựng mô hình

- Tạo danh sách list_results rỗng để lưu kết quả và tách dữ liệu đầu vào của mô hình gán vào X:

```
list_results = []
```

```
X = model.drop(columns=['Adjusted_Revenue','id','rating'])
X
```

	on_streaming_year	on_streaming_month	on_streaming_day	Action	Adventure	Animation	Comedy	Crime	Documentary	Drama	...
0	2005	9	22	0	0	0	0	1	0	1	...
1	2002	8	23	0	0	0	1	0	0	0	...
2	2008	1	11	1	1	0	0	0	0	1	...
3	2001	10	5	0	0	0	0	0	0	1	...
4	2007	12	24	0	0	0	0	1	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
3433	2009	10	16	1	0	0	1	0	0	0	...
3434	2007	1	31	0	0	0	0	0	0	1	...
3435	2018	10	26	0	0	0	0	0	0	0	...
3436	2003	9	3	0	0	0	0	1	0	1	...
3437	2017	10	6	0	0	1	0	0	0	1	...

3438 rows × 35 columns

3.3.2.1. Xây dựng mô hình dự đoán doanh thu

- Đầu ra của mô hình là 'Adjusted_Revenue' gán vào y. Sau đó chuyển đổi giá trị X theo phân phối chuẩn.
- Chia X và y thành 2 tập dữ liệu train và test với tỉ lệ test_size là 0.2

```
y = model['Adjusted_Revenue']
sc = StandardScaler()
X= sc.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

- Tạo hàm funs() để fit dữ liệu train, test vào mô hình và tính các chỉ số đánh giá R2, MAE, MSE, RMSE

```
def models(funs,name):
    funs.fit(X_train,y_train)
    y_pred = funs.predict(X_test)
    r2 = r2_score(y_test, y_pred)
    mae = mean_absolute_error(y_test, y_pred)
    mse = mean_squared_error(y_test, y_pred)
    rmse = np.sqrt(mse)
    print(f'R Squared Score of {name}: {r2 * 100:.2f}%')
    print(f'Mean Absolute Error of {name}: {mae :.2f}')
    print(f'Mean Squared Error of {name}: {mse:.2f}')
    print(f'Root Mean Squared Error of {name}: {rmse:.2f}')
    dicts = {'Outcome': 'Revenue', 'Model': name,'R Squared Score':r2,'Mean Absolute Error': mae,
             'Mean Squared Error':mse,'Root Mean Squared Error': rmse}
    list_results.append(dicts)
    return dicts
```

- Sử dụng 3 mô hình dự đoán XGB Regressor, Linear Regression, Random Forest Regressor

```
xgb_r = XGBRegressor()
dict_xgb = models(xgb_r,'XGB Regressor')
```

```
R Squared Score of XGB Regressor: 60.46%
Mean Absolute Error of XGB Regressor: 92369947.94
Mean Squared Error of XGB Regressor: 26051711982136256.00
Root Mean Squared Error of XGB Regressor: 161405427.36
```

```
lr = LinearRegression()
dict_lr = models(lr,'Linear Regression')
```

```
R Squared Score of Linear Regression: 57.02%
Mean Absolute Error of Linear Regression: 99752650.39
Mean Squared Error of Linear Regression: 28321437976000796.00
Root Mean Squared Error of Linear Regression: 168289744.12
```

```
rfr = RandomForestRegressor()
dict_rfr = models(rfr,'Random Forest Regressor')
```

```
R Squared Score of Random Forest Regressor: 63.57%
Mean Absolute Error of Random Forest Regressor: 89884000.09
Mean Squared Error of Random Forest Regressor: 24003008710837276.00
Root Mean Squared Error of Random Forest Regressor: 154929044.12
```

- Kết quả được lưu về dataframe:

	Outcome	Model	R Squared Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
0	Revenue	XGB Regressor	0.604627	9.236995e+07	2.605171e+16	1.614054e+08
1	Revenue	Linear Regression	0.570181	9.975265e+07	2.832144e+16	1.682897e+08
2	Revenue	Random Forest Regressor	0.635719	8.988400e+07	2.400301e+16	1.549290e+08

- Từ đó thấy được R2 của Random Forest Regressor cao nhất và các chỉ số MAE, MSE, RMSE thấp nhất (tuy nhiên nó quá cao do giá trị dự đoán lớn).

3.3.2.2. Xây dựng mô hình dự đoán điểm đánh giá

– Đầu ra của mô hình là 'rating' gán vào y_r. Sau đó chuyển đổi giá trị X theo phân phối chuẩn.

– Chia X và y_r thành 2 tập dữ liệu train và test với tỉ lệ test_size là 0.2

```
y_r = model['rating']
sc = StandardScaler()
X= sc.fit_transform(X)
X_train, X_test, y_r_train, y_r_test = train_test_split(X, y_r, test_size=0.2, random_state=0)
```

– Tạo hàm funs() để fit dữ liệu train, test vào mô hình và tính các chỉ số đánh giá R2, MAE, MSE, RMSE

```
def models(funs,name):
    funs.fit(X_train,y_r_train)
    y_r_pred = funs.predict(X_test)
    r2 = r2_score(y_r_test, y_r_pred)
    mae = mean_absolute_error(y_r_test, y_r_pred)
    mse = mean_squared_error(y_r_test, y_r_pred)
    rmse = np.sqrt(mse)
    print(f'R Squared Score of {name}: {r2 * 100:.2f}%')
    print(f'Mean Absolute Error of {name}: {mae :.4f}')
    print(f'Mean Squared Error of {name}: {mse:.4f}')
    print(f'Root Mean Squared Error of {name}: {rmse:.4f}')
    dicts = {'Outcome': 'Rating', 'Model': name,'R Squared Score':r2,'Mean Absolute Error': mae,
             'Mean Squared Error':mse,'Root Mean Squared Error': rmse}
    list_results.append(dicts)
    return dicts
```

– Sử dụng 3 mô hình dự đoán XGB Regressor, Linear Regression, Random Forest Regressor

```
xgb_r = XGBRegressor()
dict_xgb = models(xgb_r,'XGB Regressor')
```

```
R Squared Score of XGB Regressor: 19.96%
Mean Absolute Error of XGB Regressor: 5.6178
Mean Squared Error of XGB Regressor: 53.7768
Root Mean Squared Error of XGB Regressor: 7.3333
```

```
lr = LinearRegression()
dict_lr = models(lr,'Linear Regression')
```

```
R Squared Score of Linear Regression: 31.78%
Mean Absolute Error of Linear Regression: 5.2386
Mean Squared Error of Linear Regression: 45.8363
Root Mean Squared Error of Linear Regression: 6.7703
```

```
rfr = RandomForestRegressor()
dict_rfr = models(rfr,'Random Forest Regressor')
```

```
R Squared Score of Random Forest Regressor: 29.46%
Mean Absolute Error of Random Forest Regressor: 5.3059
Mean Squared Error of Random Forest Regressor: 47.3942
Root Mean Squared Error of Random Forest Regressor: 6.8843
```

– Kết quả được lưu về dataframe:

	Outcome	Model	R Squared Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
3	Rating	XGB Regressor	0.199596	5.617850	53.776769	7.333265
4	Rating	Linear Regression	0.317781	5.238582	45.836314	6.770252
5	Rating	Random Forest Regressor	0.294594	5.305872	47.394174	6.884343

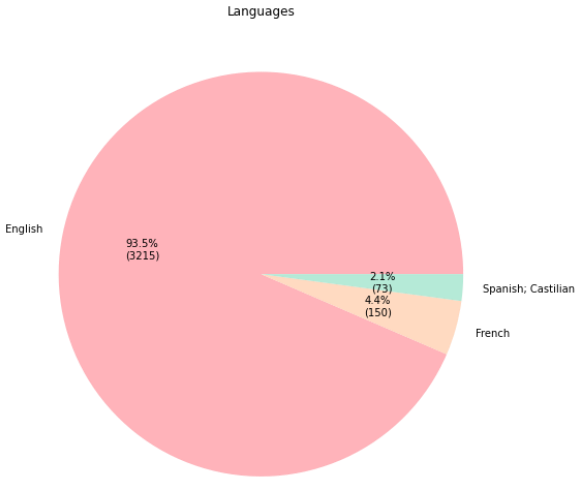
– Từ đó thấy được R2 của Linear Regression cao nhất và các chỉ số MAE, MSE, RMSE thấp nhất.

CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU

4.1. Trực quan hóa dữ liệu

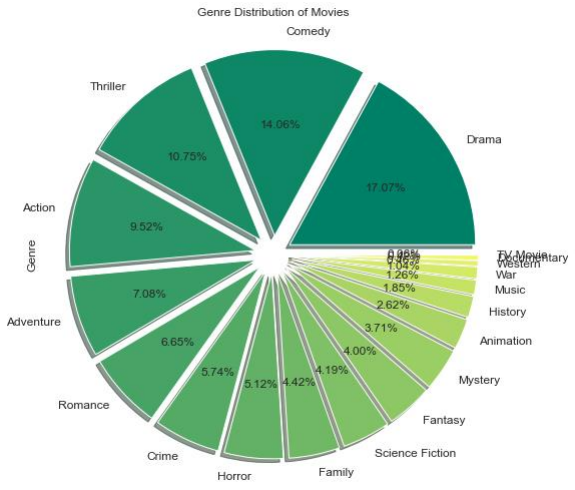
4.1.1. Tổng quan của phim:

4.1.1.1. Ngôn ngữ gốc của phim:



Biểu đồ trên cho thấy tiếng Anh (chiếm phần lớn với tỷ lệ 95.3%) là ngôn ngữ phổ biến trong ngành công nghiệp điện ảnh, trong khi tiếng Pháp (chiếm 4.4%) và tiếng Spanish, Castilian (chỉ chiếm 2.1%) đóng góp một phần nhỏ. Lý do có hiện tượng này là do số lượng phim có ngôn ngữ gốc là Tiếng Anh trong themoviedb.org chiếm số lượng lớn và nhóm chỉ chọn 3 ngôn ngữ trên để crawl dữ liệu.

4.1.1.2. Thể loại của phim:

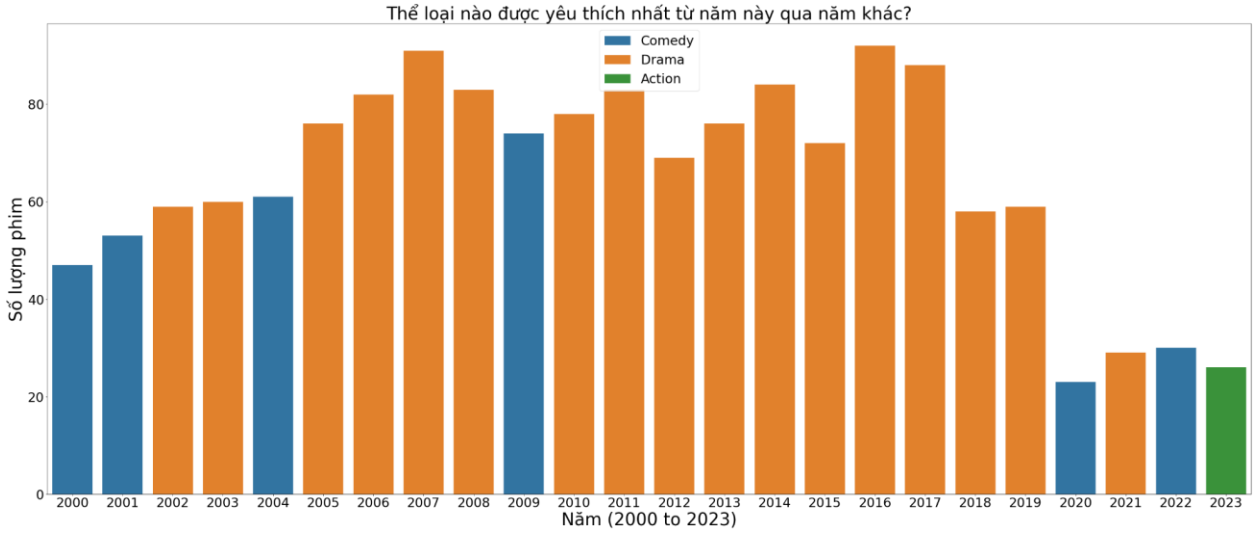


Dựa vào biểu đồ cho thấy:

- Drama: Với tỷ lệ 17.07% tần suất xuất hiện, thể loại Drama được coi là phổ biến và có sự quan tâm lớn từ khán giả. Đây là một lựa chọn tốt để tiếp tục phát triển và đầu tư, vì khả năng cao khán giả sẽ quan tâm đến các bộ phim thuộc thể loại này.

- Comedy: Với tỷ lệ 14.06% tần suất xuất hiện, thể loại Comedy cũng được coi là phổ biến. Với tính chất giải trí và khả năng mang lại tiếng cười, phim hài có thể thu hút đông đảo khán giả. Đây cũng là một thể loại đáng xem xét để phát triển và đầu tư.
- Thriller: Với tỷ lệ 10.75% tần suất xuất hiện, thể loại Thriller có thể mang đến sự hồi hộp và căng thẳng cho khán giả. Với sự pha trộn giữa yếu tố kịch tính và mạo hiểm, phim thuộc thể loại này có thể thu hút một phần khán giả đam mê những câu chuyện gây cấn.
- Documentary: Mặc dù chỉ chiếm 0.4% tần suất xuất hiện, thể loại Documentary có thể hấp dẫn một phần khán giả đặc biệt quan tâm đến việc khám phá và hiểu biết về thế giới thực. Nếu có khả năng tạo ra các bộ phim tài liệu chất lượng, đây có thể là một thị trường tiềm năng để khai thác.
- TV Show: Với tỷ lệ 0.1% tần suất xuất hiện thấp nhất, thể loại TV Show có thể không phổ biến trong tập dữ liệu này. Tuy nhiên, nếu bạn có ý định phát triển các dự án TV Show, hãy đảm bảo nghiên cứu thị trường và đáp ứng nhu cầu của khán giả.

4.1.1.3. Thể loại của phim qua các năm:

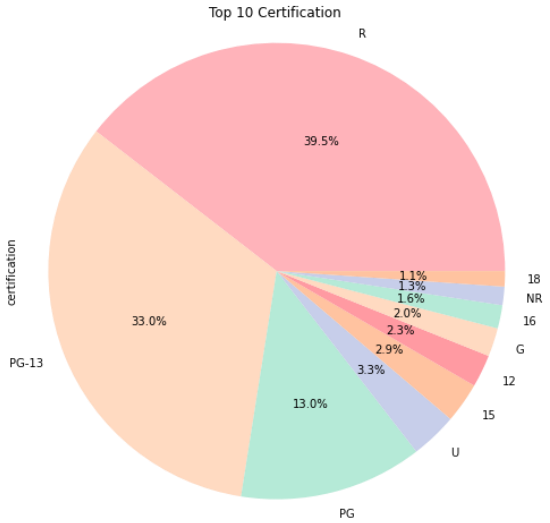


Dựa trên phân tích thể loại được ưa thích trong khoảng thời gian từ năm 2000 đến năm 2023, rút ra được kết luận:

- Drama: Thể loại Drama chiếm ưu thế trong suốt 17 năm, cho thấy sự ổn định và sự ưa thích từ khán giả trong giai đoạn đó. Drama thường tập trung vào việc xây dựng câu chuyện, phát triển nhân vật và khám phá các cảm xúc, điều này có thể giải thích sự phổ biến lâu dài của thể loại này.
- Comedy: Comedy là thể loại được yêu thích trong 6 năm, cho thấy sự ưa chuộng của khán giả đối với những bộ phim mang tính giải trí và mang lại tiếng cười. Comedy thường tạo ra một không gian vui nhộn và thoải mái cho khán giả.

– Action: Trong năm 2023, thể loại Action được yêu thích hơn so với các năm trước đó. Thể loại này thường tập trung vào các màn hành động, đánh nhau và phiêu lưu, và có thể mang lại cảm giác kịch tính và hứng khởi cho khán giả. Sự tăng trưởng ưa thích của Action có thể phản ánh xu hướng thị trường hiện tại và sự thay đổi trong sở thích của khán giả.

4.1.1.4. Phân loại của phim:



– Nhãn"R" là loại phổ biến nhất, với số lượng là 1270. Điều này cho thấy rằng trong dữ liệu, có nhiều phim được phân loại là chỉ dành cho người trên 17 tuổi hoặc có sự giám sát từ người lớn. Điều này có thể cho thấy xu hướng ưa thích và tiêu chuẩn nội dung của khán giả mục tiêu.

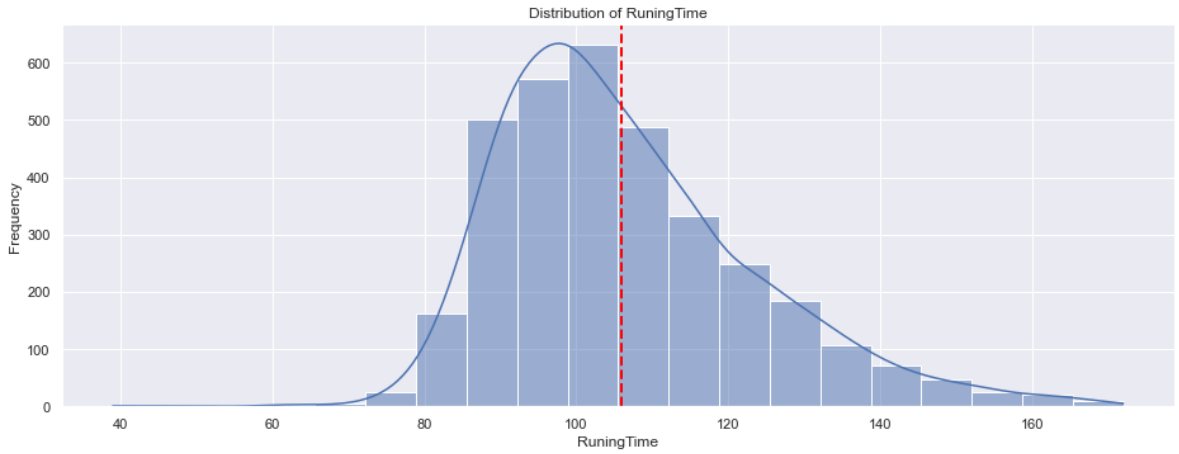
– Nhãn "PG-13" đứng thứ hai với số lượng 1061. Nhãn này chỉ ra rằng nhiều phim trong dữ liệu được phân loại cho khán giả từ 13 tuổi trở lên. Điều này cho thấy sự quan tâm đến nội dung phù hợp cho khán giả trẻ tuổi, nhưng có thể có một số nội dung không phù hợp cho trẻ em nhỏ.

– Nhãn "PG" có số lượng 417, đứng thứ ba trong danh sách. Nhãn này chỉ ra rằng một số phim trong dữ liệu được phân loại cho khán giả mọi lứa tuổi. Tuy nhiên, cũng có thể có một số nội dung không thích hợp cho trẻ em nhỏ. Điều này cho thấy sự cân nhắc về việc tạo ra nội dung phù hợp cho gia đình và khán giả trẻ em.

– Các nhãn như U15, 12, G16, NR và 18 đều chiếm tỷ lệ nhỏ hơn trong biểu đồ.

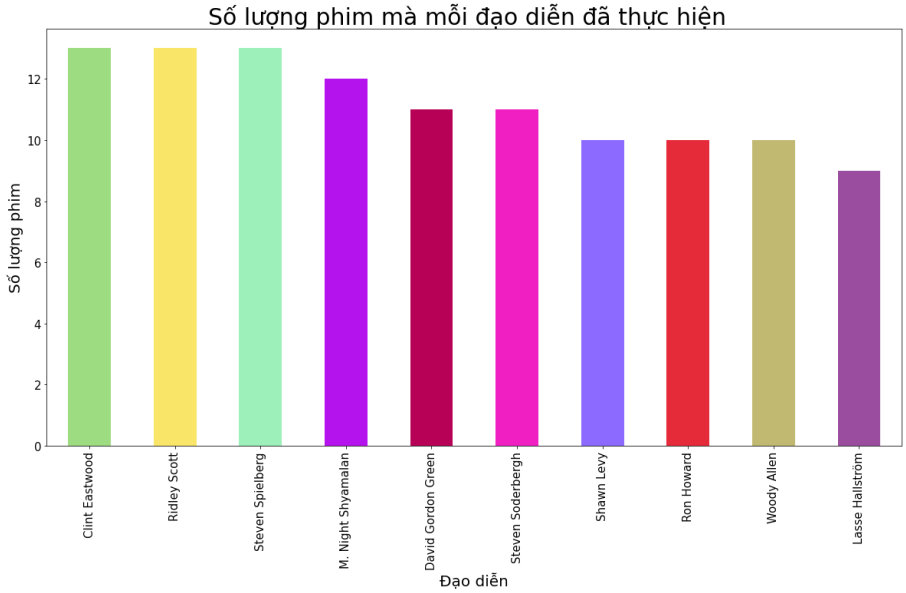
– Biểu đồ trên cho thấy rằng nhãn R và PG-13 là hai nhãn phổ biến nhất trong top 10 nhãn phân loại của các bộ phim. Điều này có thể phản ánh sự đa dạng về nội dung và mục tiêu khán giả của ngành công nghiệp điện ảnh.

4.1.1.5. Thời lượng của phim:



- Đa phần các bộ phim có thời lượng từ 95 phút đến 110 phút. Điều này cho thấy rằng thời lượng phát thông thường của các bộ phim nằm trong khoảng thời gian này. Thời lượng này có thể được coi là một mức tiêu chuẩn cho các bộ phim thông thường.
- Có một phần nhỏ các bộ phim có thời lượng dưới 80 phút. Các bộ phim ngắn hoặc tập trung vào câu chuyện ngắn gọn có thể có thời lượng ngắn hơn so với các bộ phim thông thường.
- Cũng có một phần nhỏ các bộ phim có thời lượng trên 120 phút. Các bộ phim dài này có thể là những tác phẩm nghệ thuật hoặc có nội dung phức tạp, yêu cầu thời gian để phát triển câu chuyện và nhân vật.

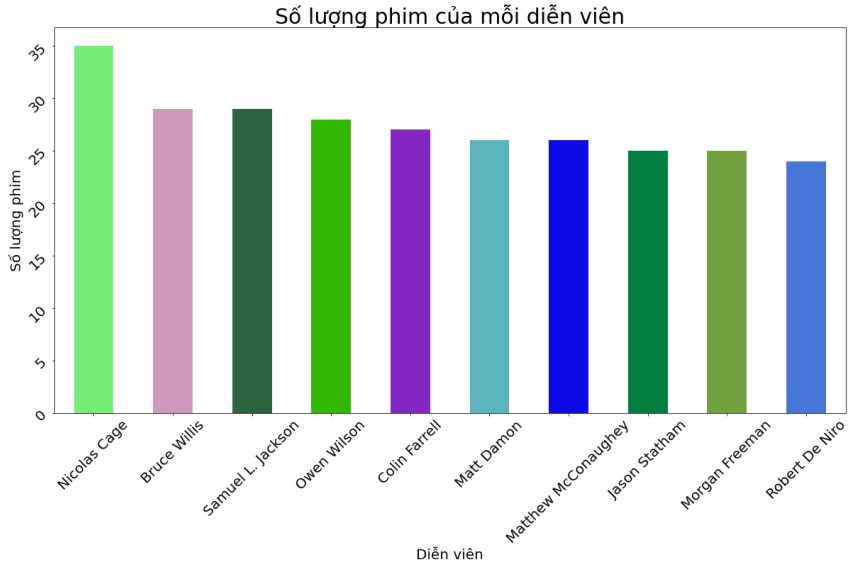
4.1.1.6. Đạo diễn:



- Clint Eastwood, Ridley Scott và Steven Spielberg đều có số lượng phim thực hiện là 13 bộ, đứng đầu trong danh sách các đạo diễn với số lượng phim nhiều nhất. Điều này cho thấy sự nổi tiếng và đóng góp lớn của họ trong ngành công nghiệp điện ảnh.

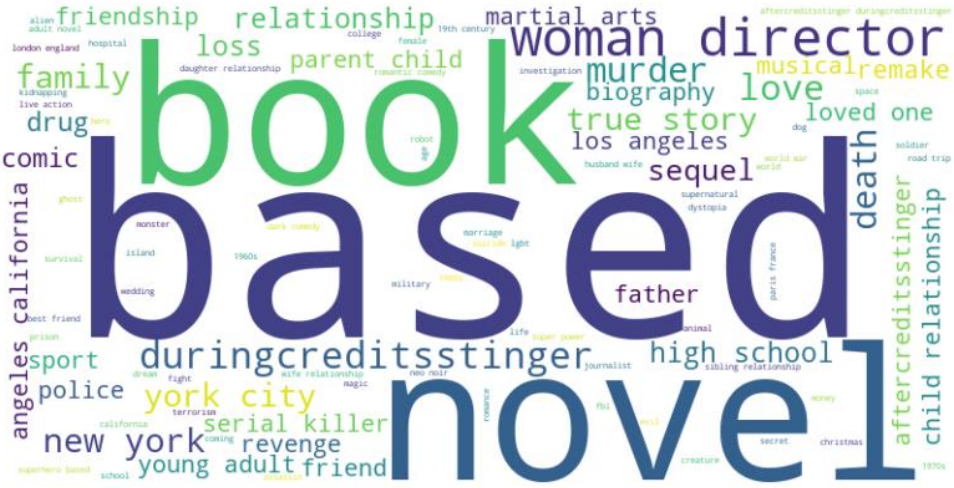
- M. Night Shyamalan cũng có số lượng phim khá đáng kể, với 12 bộ phim. Anh ta cũng nằm trong top những đạo diễn có sự ảnh hưởng đáng kể trong ngành.
- David Gordon Green và Steven Soderbergh thực hiện 11 bộ phim, trong khi Shawn Levy, Ron Howard và Woody Allen thực hiện 10 bộ phim. Các đạo diễn này cũng có sự đóng góp đáng kể trong lĩnh vực điện ảnh.
- Lasse Hall có số lượng phim thấp hơn so với các đạo diễn khác, với 9 bộ phim, nhưng vẫn có một đóng góp đáng kể trong ngành công nghiệp điện ảnh.

4.1.1.7. Diễn viên:



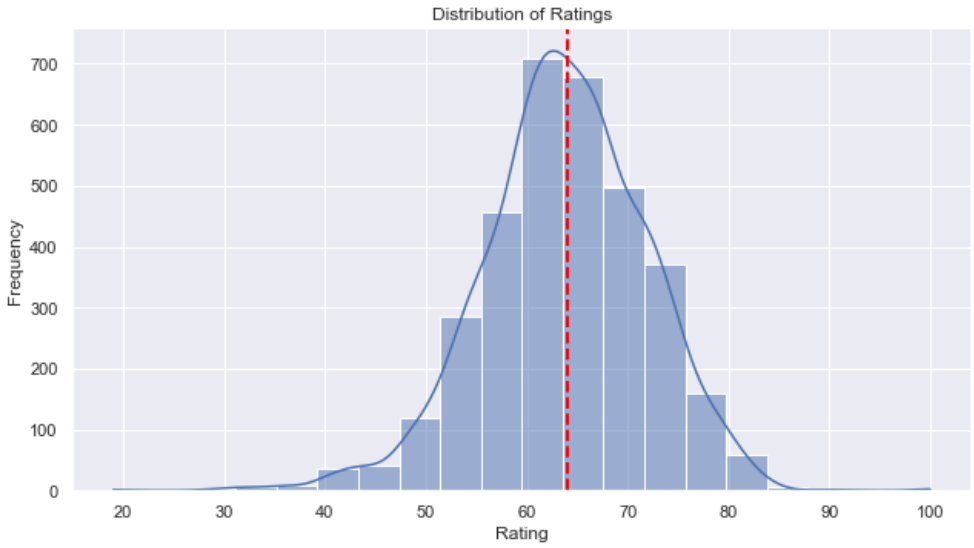
- Nicolas Cage được liệt kê với số lượng phim nhiều nhất, đạt khoảng 35 bộ phim. Điều này cho thấy anh ta có sự nổi tiếng và tham gia nhiều dự án điện ảnh.
- Bruce Willis và Jackson cũng có số lượng phim đáng kể, khoảng 30 bộ. Đây là các diễn viên có tầm ảnh hưởng và thường xuyên tham gia vào các dự án điện ảnh.
- Các diễn viên khác trong biểu đồ có số lượng phim trong khoảng từ 25 đến 30 bộ, thấp hơn so với Nicolas Cage và Bruce Willis/Jackson, tuy không cao bằng nhưng vẫn có một đóng góp khá trong ngành công nghiệp điện ảnh.

4.1.1.8. Từ khóa nổi bật của phim:



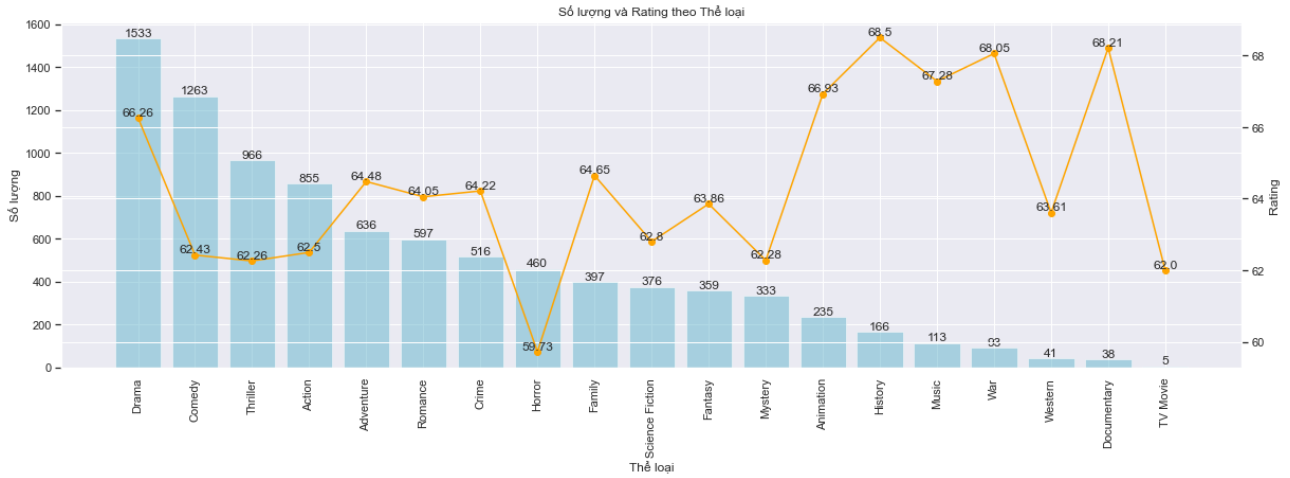
Biểu đồ WordCloud này cho thấy các từ khóa phổ biến nhất và độ phổ biến của chúng trong ngành công nghiệp điện ảnh. Từ khóa "book", "based", "novel" thường xuyên được sử dụng để chỉ ra rằng nhiều bộ phim được dựa trên các tác phẩm văn học. Các từ khóa khác như "women director", "duringcreditsstinger", "high school", "love", "family"... có độ phổ biến thấp hơn, có thể chỉ ra rằng chúng vẫn được sử dụng khá phổ biến, còn lại có nhiều keyword không được sử dụng rộng rãi trong từ khóa chính của các bộ phim.

4.1.1.9. Điểm đánh giá của phim:



Trong biểu đồ dữ liệu, ta nhận thấy rằng phân phối rating không đồng đều. Các rating chủ yếu tập trung trong khoảng từ 58 đến 68. Điều này cho thấy rằng có một sự ưu tiên đánh giá tương đối cao cho các phim thuộc khoảng rating này. Trong khi đó, các rating từ 80 đến 100 xuất hiện với tần số thấp hơn so với khoảng rating 50 đến 70, tuy nhiên vẫn có một số phim được đánh giá cao. Điều này cho thấy rằng có một số tác phẩm đặc biệt được công nhận và đánh giá cao bởi người dùng.

4.1.1.10. Điểm đánh giá theo từng thể loại của phim:



Phân tích về rating trung bình theo từng thể loại phim:

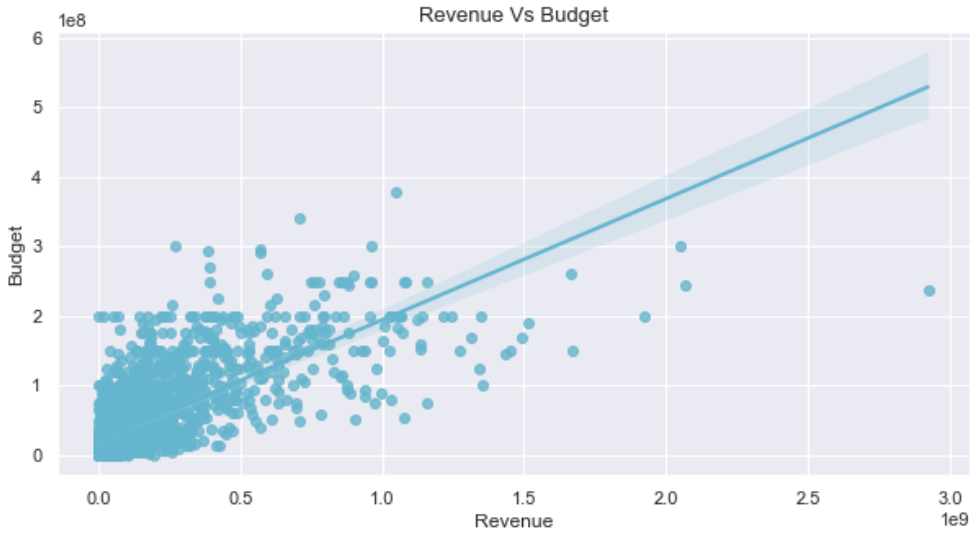
- Thể loại Drama có số lượng phim cao nhưng rating trung bình chỉ là 66.26 điểm. Điều này có thể cho thấy mặc dù có nhiều phim thuộc thể loại này, nhưng chất lượng và đánh giá của các bộ phim Drama không cao bằng các thể loại khác.
- Thể loại History có rating trung bình cao nhất là 68.5 điểm trong 166 bộ phim chiếu. Điều này cho thấy các bộ phim thuộc thể loại History được đánh giá cao và có chất lượng tốt.
- Thể loại Horror có rating trung bình thấp nhất là 59.73 điểm. Điều này cho thấy mặc dù có nhiều phim thuộc thể loại Horror, nhưng chất lượng và đánh giá của các bộ phim này không cao, có thể do yếu tố kinh dị và đánh đồng trong nội dung.
- Thể loại Documentary có rating trung bình cao 68.21 điểm, số lượng phim thuộc thể loại Documentary khá ít 38 phim. Do số lượng phim ít, việc so sánh với các thể loại có tần suất chiếu cao hơn có thể không chính xác.

Kết luận:

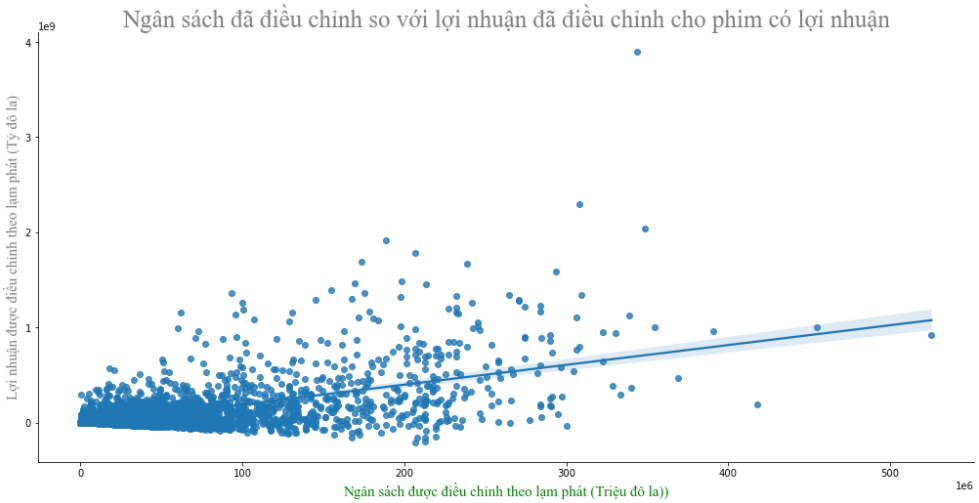
- Tận dụng sự phổ biến của tiếng Anh trong ngành công nghiệp điện ảnh, chiếm tỷ lệ lớn (95.3%). Điều này có thể bao gồm sản xuất phim chủ yếu bằng tiếng Anh hoặc bao gồm phụ đề tiếng Anh để mở rộng sự tiếp cận.
- Sử dụng nhãn "R" để tạo ra những bộ phim hấp dẫn cho khán giả từ 17 tuổi trở lên hoặc dưới 17 tuổi nhưng phải có sự giám sát từ người lớn. Phát triển nội dung phù hợp cho nhãn "PG-13" để đáp ứng nhu cầu của khán giả từ 13 tuổi trở lên.
- Tạo ra nội dung phù hợp cho gia đình với nhãn "PG" kết hợp giáo dục và giải trí.
- Khám phá tiềm năng của các nhãn khác như U15, 12, G16, NR và 18 để đáp ứng yêu cầu đa dạng của khán giả.
- Thời lượng phim thông thường nằm trong khoảng từ 95 đến 110 phút, với một số phim ngắn hơn khoảng 80 phút và một số phim dài hơn 120 phút. Cân nhắc thời lượng phù hợp dựa trên câu chuyện và khán giả mục tiêu.
- Đầu tư và phát triển các bộ phim trong thể loại Drama và Comedy để đáp ứng sự quan tâm lớn từ khán giả.
- Nắm bắt xu hướng tăng trưởng của thể loại Action và sản xuất những bộ phim hành động mạo hiểm và kịch tính.
- Sản xuất các bộ phim thuộc thể loại tài liệu chất lượng để phục vụ khán giả quan tâm đến việc khám phá và hiểu biết về thế giới thực.
- Nghiên cứu và phát triển các chương trình truyền hình độc đáo và hấp dẫn trong thể loại TV Show để thu hút sự quan tâm của khán giả.
- Đầu tư vào thể loại History và Documentary để tăng cường nội dung và thu hút khán giả.
- Nâng cao chất lượng phim Action bằng cách đảm bảo kịch bản, đạo diễn, diễn xuất, hiệu ứng hình ảnh và âm thanh chất lượng cao.
- Đánh giá lại thể loại Horror để cải thiện chất lượng và đánh giá của thể loại này.
- Tiếp tục khám phá thể loại War để tạo ra những bộ phim độc đáo và hấp dẫn.

4.1.2. Tác động đến sự thành công của phim:

4.1.2.1. Tương quan giữa ngân sách và doanh thu của phim:



Mối tương quan giữa Ngân sách và Doanh thu: Cả ngân sách và doanh thu đều có mối tương quan dương (0,695) giữa chúng. Có nghĩa là có nhiều khả năng những bộ phim có mức đầu tư cao hơn sẽ mang lại doanh thu tốt hơn.

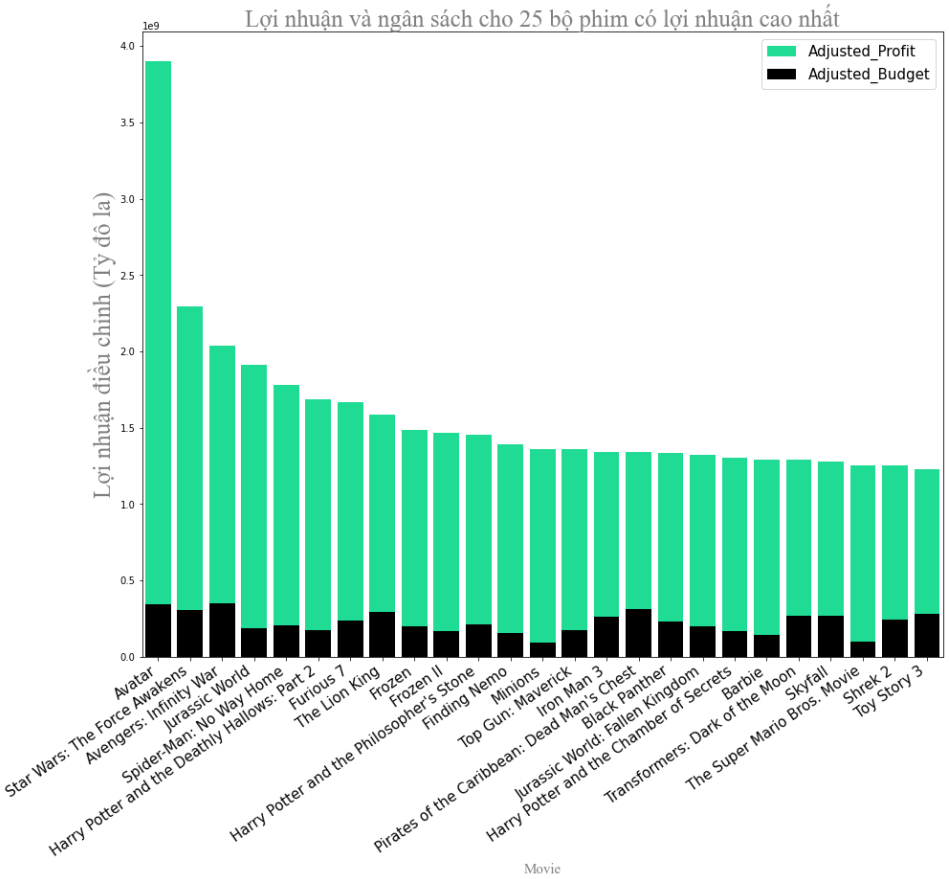


Dựa trên đồ thị phân tán giữa ngân sách (Adjusted_Budget) và lợi nhuận (Adjusted_Profit), chúng ta có thể rút ra một số nhận xét:

- Có một mối tương quan dương mạnh giữa ngân sách và lợi nhuận. Điều này có nghĩa là khi ngân sách tăng, lợi nhuận cũng tăng và ngược lại. Mối quan hệ này có giá trị hệ số tương quan Pearson (correlation coefficient) là 0.59, cho thấy mối liên hệ này khá mạnh.
- Mối tương quan dương mạnh này cho thấy rằng việc đầu tư một số lượng tài nguyên và kinh phí lớn vào một bộ phim có thể có tiềm năng tạo ra lợi nhuận cao hơn. Tuy nhiên, việc quản lý và sử dụng ngân sách hiệu quả là rất quan trọng để đảm bảo rằng các dự án phim có thể đạt được lợi nhuận mong đợi.

– Bên cạnh ngân sách, còn có nhiều yếu tố khác cũng có thể ảnh hưởng đến lợi nhuận của một bộ phim như chất lượng nội dung, kỹ năng của đội ngũ sản xuất, chiến lược tiếp thị và các yếu tố thị trường. Do đó, việc xem xét những yếu tố này cũng là rất quan trọng trong việc đạt được lợi nhuận cao trong ngành công nghiệp điện ảnh.

4.1.2.2. Ngân sách và lợi nhuận của top 25 phim có lợi nhuận cao nhất:



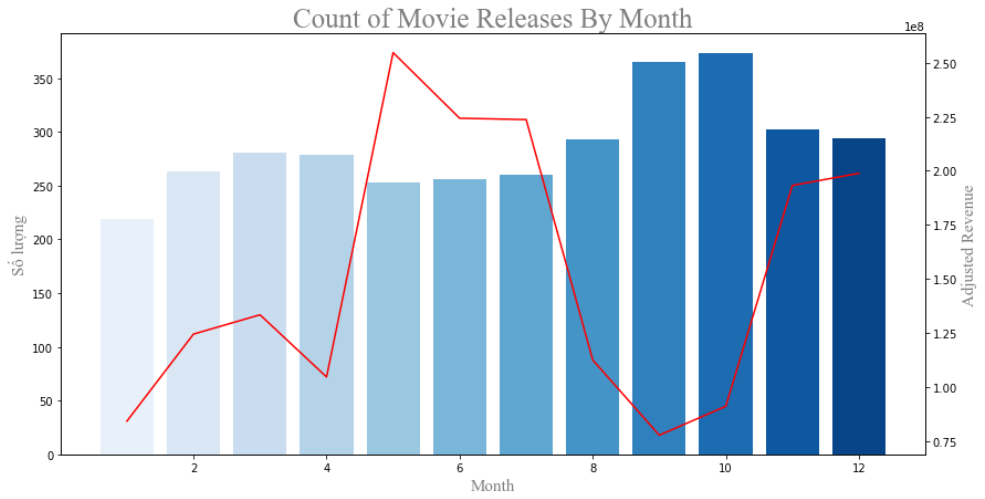
Dựa trên phân tích về ngân sách và lợi nhuận của 25 bộ phim có lợi nhuận cao nhất, rút ra được kết luận:

– Avatar là bộ phim có lợi nhuận cao nhất, đạt khoảng 3,9 tỷ đô la. Đây là một con số ấn tượng cho thấy sự thành công về mặt kinh doanh của bộ phim này. Ngân sách đầu tư cho Avatar cũng khá cao, khoảng 0,35 tỷ đô la, cho thấy việc đầu tư lớn vào sản xuất phim đã mang lại lợi nhuận đáng kể.

– Star Wars và Avengers cũng có lợi nhuận cao, lần lượt là khoảng 2,3 tỷ đô la và 2,1 tỷ đô la. Ngân sách đầu tư cho cả hai bộ phim này cũng khá ổn định, trong khoảng từ 0,3 tỷ đô la đến 0,35 tỷ đô la.

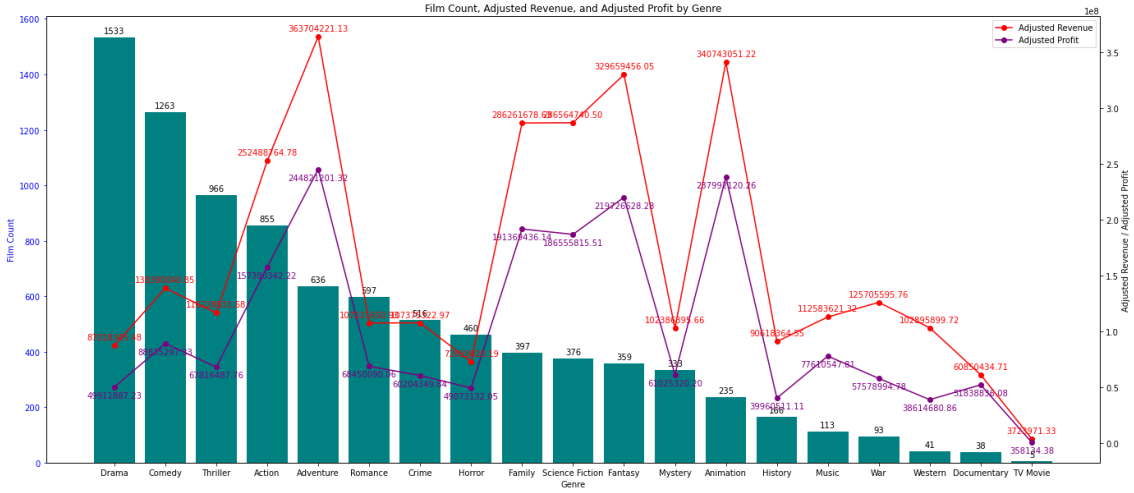
– Các bộ phim còn lại trong danh sách có lợi nhuận từ khoảng 1,3 tỷ đô la đến 2 tỷ đô la, và ngân sách đầu tư từ trên dưới 0,3 tỷ đô la. Điều này cho thấy sự đa dạng về mức lợi nhuận và ngân sách đầu tư trong danh sách các bộ phim có lợi nhuận cao nhất.

4.1.2.3. Doanh thu trung bình theo tháng của phim:



Dựa trên thông tin về số lượng phim và lợi nhuận ròng trung bình theo tháng, có thể rút ra được kết luận: tháng 9 và tháng 10 là hai tháng có số lượng phim phát sóng nhiều nhất, tuy nhiên lại đem lại doanh thu thấp nhất trong các tháng. Trong khi đó, tháng 5, tháng 6 và tháng 7 không chỉ dẫn đầu về số lượng phim phát sóng mà lại dẫn đầu về doanh thu trung bình. Điều này cho thấy mùa hè có thể tạo ra thành công lớn hơn cho ngành công nghiệp điện ảnh, có thể do sự tăng cường của đối tượng khán giả trẻ em và cha mẹ của chúng trong kỳ nghỉ hè.

4.1.2.4. Doanh thu và lợi nhuận theo từng thể loại của phim:



Dựa vào biểu đồ trên, ta thấy rằng:

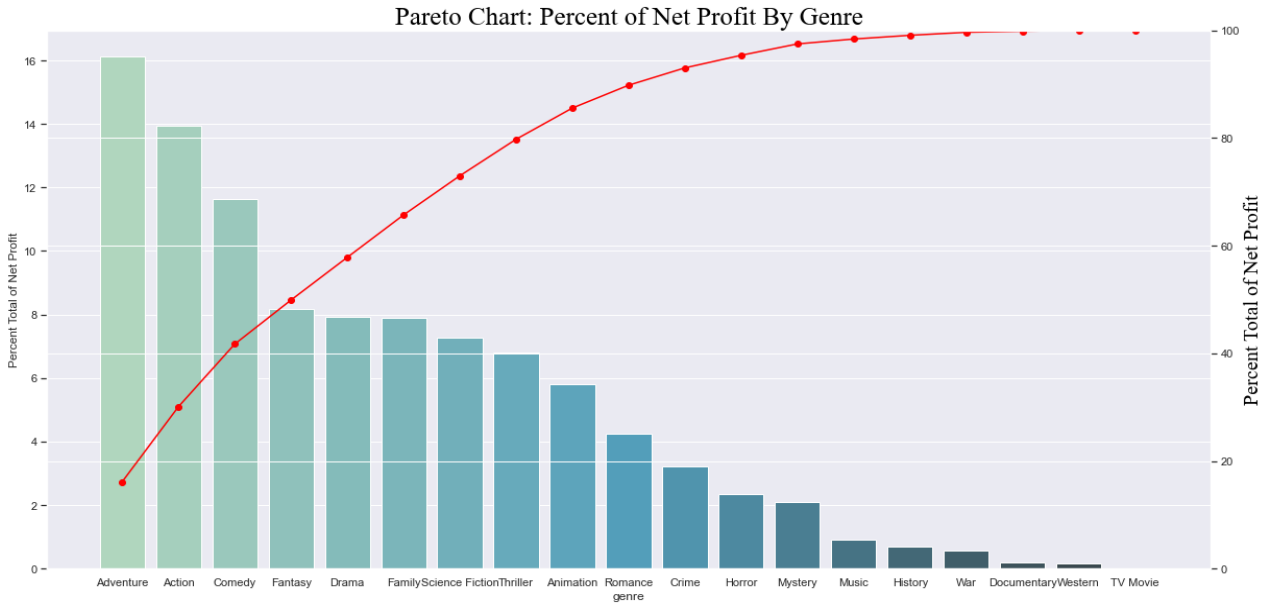
- Mặc dù thể loại Drama dẫn đầu về số lượng phim với 636 bộ phim, lợi nhuận trung bình và doanh thu trung bình của thể loại này lại thấp. Điều này có thể cho thấy mặc dù có nhiều phim, không phải tất cả các bộ phim Drama đều tạo ra lợi nhuận và doanh thu cao.
- Thể loại Adventure đứng đầu về lợi nhuận trung bình và doanh thu, mặc dù có số lượng phim không nhiều như Drama. Điều này cho thấy sự ưa chuộng của khán giả đối với những cuộc phiêu lưu mạo hiểm và câu chuyện thú vị trong thể loại này.

– Thể loại Animation, mặc dù chỉ có 235 bộ phim (xếp thứ 14 về số lượng), nhưng lại đứng thứ 2 về lợi nhuận trung bình và doanh thu trung bình sau thể loại Adventure. Điều này cho thấy sự hấp dẫn của phim hoạt hình với khán giả và khả năng tạo ra lợi nhuận cao.

– Các thể loại Fantasy, Family và Science Fiction cũng có lợi nhuận trung bình và doanh thu trung bình cao, mặc dù không phải là các thể loại phổ biến với số lượng phim phát hành nhiều. Điều này cho thấy khán giả vẫn đặt niềm tin và quan tâm đến những câu chuyện trong các thể loại này.

– Các thể loại War, Western, Documentary và TV Movie có số lượng phim dưới 100 và lợi nhuận trung bình và doanh thu trung bình thấp nhất trong bảng dữ liệu. Điều này có thể cho thấy sự hạn chế về sự quan tâm và tiếp cận của khán giả đối với những thể loại này.

4.1.2.5. Lợi nhuận theo từng thể loại của phim:



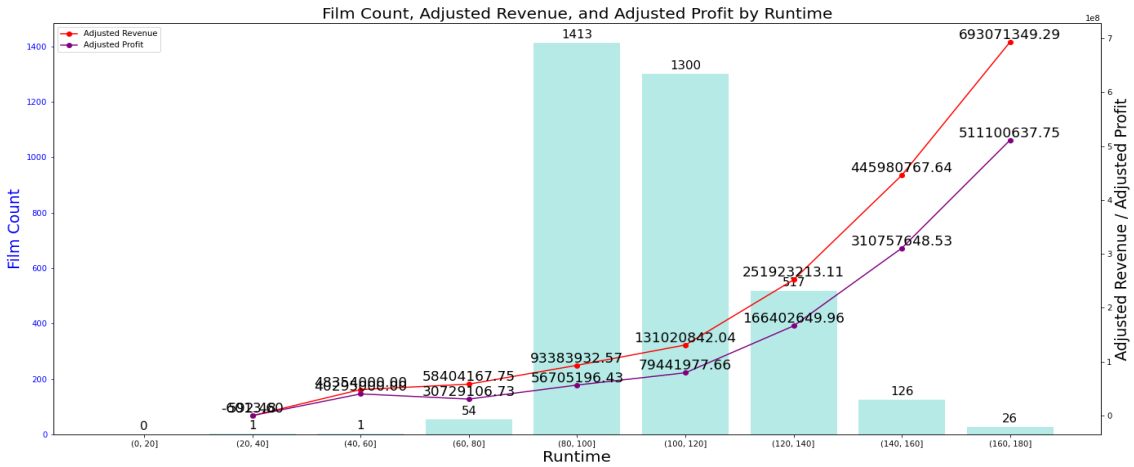
Dựa vào biểu đồ Pareto "Percent Total of Net Profit" được thể hiện ở trên, chúng ta có thể kết luận:

– Các thể loại phim chiếm 80% lợi nhuận: Adventure, Action, Comedy, Fantasy, Drama, Family, Science Fiction, Thriller và Animation là những thể loại phim tạo ra 80% tổng lợi nhuận. Điều này cho thấy sự quan trọng của các thể loại này trong việc đạt được lợi nhuận cao.

– Adventure và Action là hai thể loại chiếm tỷ trọng lợi nhuận cao nhất: Adventure chiếm 16.14% và Action chiếm 13.95% tổng lợi nhuận. Đây là hai thể loại có sự ưa chuộng và tiềm năng tạo ra lợi nhuận lớn, vì vậy nên tập trung vào sản xuất và quảng bá các bộ phim thuộc hai thể loại này.

- Comedy, Fantasy và Drama cũng có lợi nhuận đáng kể: Comedy chiếm 11.63%, Fantasy chiếm 8.18% và Drama chiếm 7.93% tổng lợi nhuận. Đầu tư vào việc sản xuất các bộ phim thuộc ba thể loại này có thể mang lại lợi nhuận ổn định.
- Family, Science Fiction, Thriller và Animation có tiềm năng tạo ra lợi nhuận: Các thể loại này có tỷ trọng lợi nhuận từ 7.27% đến 7.88%. Nên xem xét phát triển các bộ phim thuộc các thể loại này để tận dụng tiềm năng từ đối tượng khán giả của chúng.
- Các thể loại phim khác như Romance, Crime, Horror, Mystery, Music, History, War, Documentary, Western và TV Movie có tỷ trọng lợi nhuận thấp hơn. Tuy nhiên, không nên hoàn toàn loại bỏ chúng, mà có thể tiếp tục nghiên cứu và đánh giá tiềm năng của từng thể loại để xem xét khả năng tạo ra lợi nhuận trong điều kiện thị trường cụ thể.

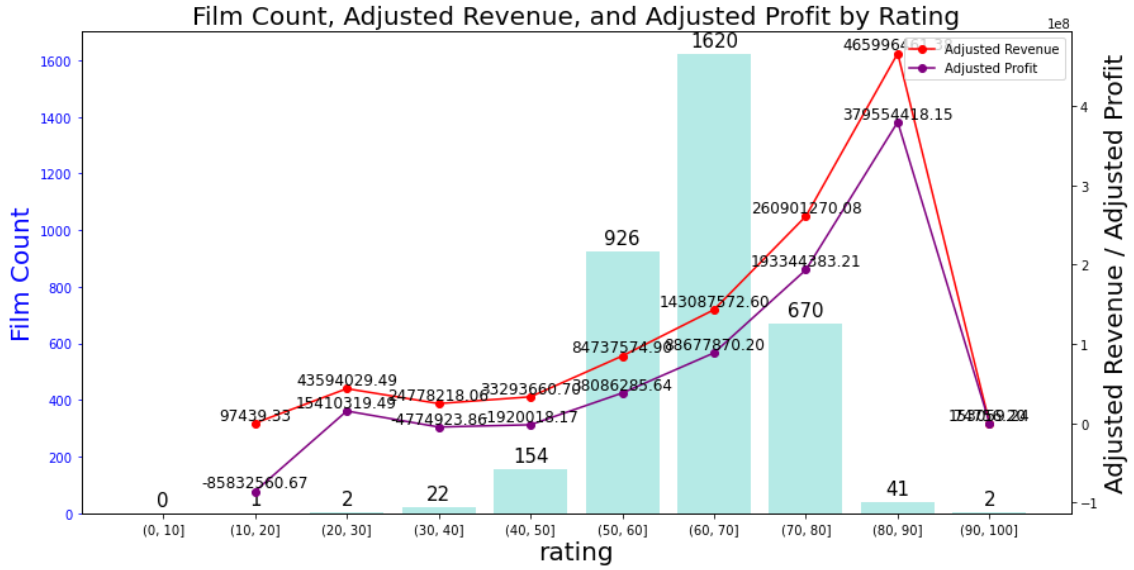
4.1.2.6. Lợi nhuận theo thời lượng chiếu phim:



Dựa vào biểu đồ trên, ta có các kết luận sau:

- Thời lượng phát sóng ngắn (20-40 phút): Chỉ có 1 bộ phim thuộc thời lượng này, và nó không đạt được lợi nhuận tốt. Điều này cho thấy rằng đối với thời lượng ngắn, khả năng thu hút khán giả và tạo ra lợi nhuận cao có thể bị hạn chế.
- Thời lượng phát sóng từ 40-60 phút: Mặc dù chỉ có 1 bộ phim trong khoảng thời gian này, nó đã mang lại doanh thu và lợi nhuận đáng kể. Điều này cho thấy rằng một bộ phim có thời lượng trung bình có thể vẫn đạt được thành công kinh doanh nếu nội dung và yếu tố khác thu hút khán giả.
- Thời lượng phát sóng từ 60-180 phút: Các bộ phim với thời lượng này có xu hướng đạt được doanh thu và lợi nhuận cao hơn. Đặc biệt, các bộ phim có thời lượng từ 140-180 phút cho thấy sự tăng trưởng lớn về doanh thu và lợi nhuận. Điều này có thể cho thấy rằng khán giả có xu hướng ưa thích các bộ phim có thời lượng dài hơn và sẵn sàng chi tiêu cho những trải nghiệm điện ảnh dài hơn.

4.1.2.7. *Lợi nhuận theo rating của phim:*

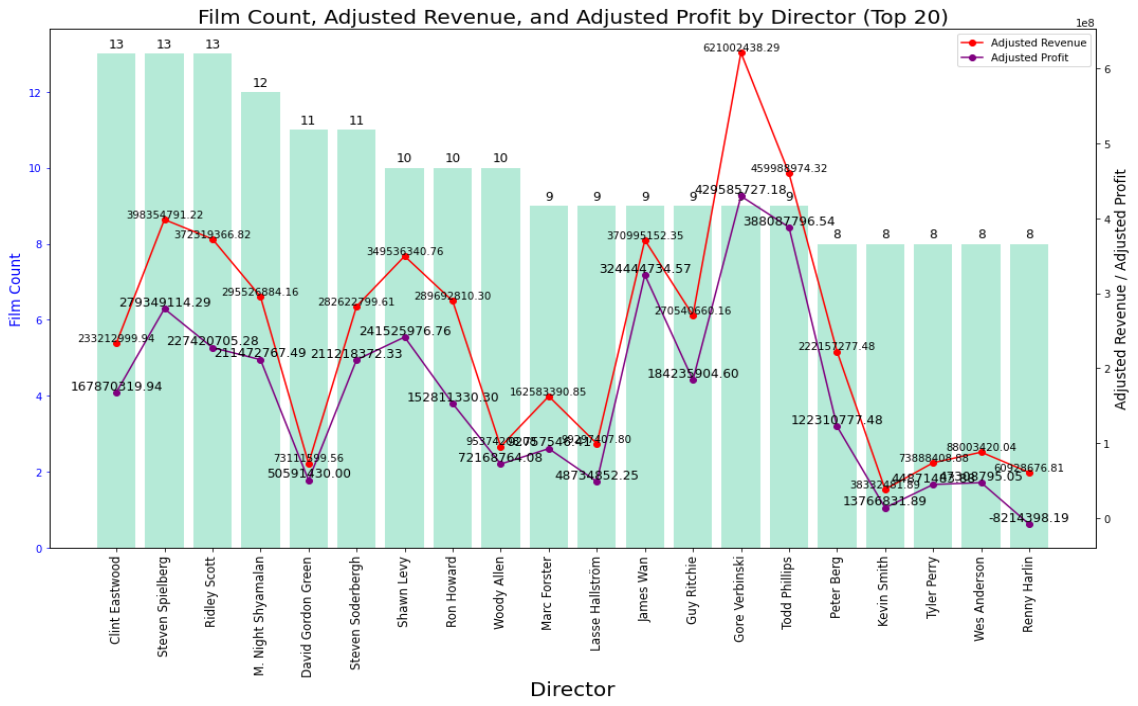


Dựa trên biểu đồ trên, chúng ta có thể rút ra một số nhận xét:

- Số lượng phim trong các khoảng rating khác nhau không đồng đều. Cụ thể, có nhiều phim nằm trong khoảng rating từ 60 đến 80, trong khi số lượng phim trong khoảng rating từ 0 đến 10 là rất ít.
- Doanh thu trung bình (Adjusted Revenue) của các phim có rating trong khoảng từ 80 đến 90 là cao nhất, với giá trị khoảng 465,996,461.38$. Điều này cho thấy rằng các phim có rating cao hơn thường có xu hướng thu được doanh thu cao hơn.
- Lợi nhuận trung bình (Adjusted Profit) của các phim có rating trong khoảng từ 70 đến 80 là cao nhất, với giá trị khoảng 193,344,383.21$. Điều này cho thấy rằng các phim có rating trong khoảng này có khả năng tạo ra lợi nhuận cao hơn so với các khoảng rating khác.

Tuy nhiên, cần lưu ý rằng có một số khoảng rating có số lượng phim rất ít, ví dụ như khoảng từ 0 đến 10 chỉ có 1 phim. Do đó, việc rút ra kết luận chính xác từ các khoảng này có thể không đại diện cho toàn bộ tập dữ liệu.

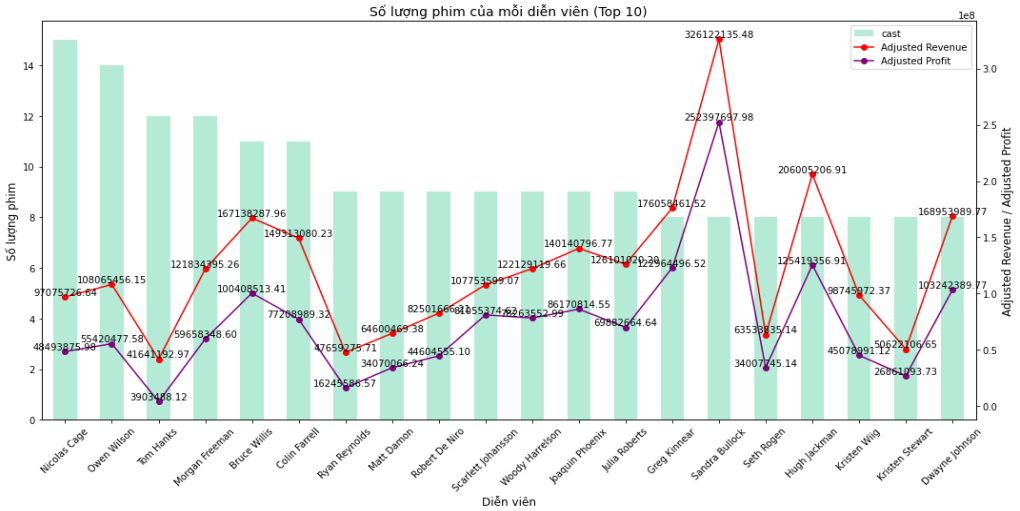
4.1.2.8. *Lợi nhuận theo 20 đạo diễn có số lượng phim cao nhất:*



Dựa trên bảng dữ liệu cho top 10 đạo diễn với số liệu về số lượng phim, trung bình doanh thu và trung bình lợi nhuận cho mỗi đạo diễn, ta có thể rút ra một số kết luận:

- Clint Eastwood, Steven Spielberg và Ridley Scott là ba đạo diễn dẫn đầu về số lượng phim được sản xuất, với mỗi người đạo diễn đã thực hiện 13 bộ phim trong dữ liệu đã cho.
- Steven Spielberg về trung bình doanh thu và lợi nhuận, với giá trị trung bình của doanh thu là khoảng 398 triệu đô la và lợi nhuận là khoảng 279 triệu đô la.
- Gore Verbinski mặc dù chỉ tham gia đạo diễn 9 bộ nhưng trung bình doanh thu và lợi nhuận cao nhất, với giá trị trung bình của doanh thu là khoảng 621 triệu đô la và lợi nhuận là khoảng 430 triệu đô la.
- Các đạo diễn M. Night Shyamalan, David Gordon Green và Steven Soderbergh cũng có hiệu suất tốt về trung bình doanh thu và lợi nhuận.
- Woody Allen có số lượng phim ít hơn so với các đạo diễn khác trong top 10, nhưng vẫn có mức trung bình doanh thu và lợi nhuận tương đối đáng kể.
- Có một số đạo diễn trong top 10 như Renny Harlin có lợi nhuận âm, cho thấy không phải tất cả các bộ phim của ông đều có hiệu quả kinh doanh tích cực.

4.1.2.9. *Lợi nhuận theo top 20 diễn viên có số lượng phim cao nhất:*



Dựa vào biểu đồ trên, chúng ta có một bảng danh sách top 20 diễn viên dựa trên số lượng phim, trung bình doanh thu và trung bình lợi nhuận. Dưới đây là một số nhận xét và kết luận từ bảng kết quả:

- Nicolas Cage dẫn đầu danh sách với 15 phim và trung bình doanh thu khoảng 97 triệu USD. Tuy nhiên, trung bình lợi nhuận của anh ta chỉ khoảng 48 triệu USD, thể hiện rằng lợi nhuận trung bình từ các phim của anh không cao.
- Owen Wilson xếp thứ hai với 14 phim và trung bình doanh thu gần 108 triệu USD. Anh ta cũng có trung bình lợi nhuận khoảng 55 triệu USD, cho thấy anh ta có khả năng mang lại lợi nhuận tốt.
- Tom Hanks, một diễn viên nổi tiếng, có 12 phim trong danh sách. Tuy nhiên, trung bình doanh thu của anh ta chỉ khoảng 41 triệu USD và trung bình lợi nhuận chỉ khoảng 3,9 triệu USD, cho thấy lợi nhuận từ các phim của anh không cao.
- Morgan Freeman cũng có 12 phim và có trung bình doanh thu khoảng 122 triệu USD. Trung bình lợi nhuận của anh ta cũng cao, khoảng 59 triệu USD, thể hiện khả năng mang lại lợi nhuận tốt.
- Bruce Willis và Colin Farrell cũng nổi bật với 11 phim mỗi người và trung bình doanh thu lần lượt là khoảng 167 triệu USD và 149 triệu USD. Cả hai đều có trung bình lợi nhuận cao, cho thấy khả năng tạo ra lợi nhuận tốt.
- Trong số các diễn viên có mặt trong danh sách, Emily Blunt, Jennifer Lawrence, và Kate Winslet có trung bình doanh thu và trung bình lợi nhuận cao hơn so với các diễn viên khác. Điều này cho thấy họ có khả năng đóng vai chính trong các bộ phim có doanh thu và lợi nhuận cao.

Kết luận:

Tận dụng mùa hè với tháng 5, 6 và 7 để sản xuất và quảng bá các bộ phim hấp dẫn dành cho khán giả trẻ em và gia đình trong mùa hè để đem lại doanh thu cao.

Thể loại phiêu lưu, hành động, hài kịch và chính kịch chiếm phần lớn lợi nhuận ròng tổng thể từ tất cả các bộ phim. Tuy nhiên, từ những các biểu, có những cơ hội lớn trong thị trường phim kinh dị và khoa học viễn tưởng do độ bão hòa thấp hơn nhưng lợi nhuận ròng trung bình cao. Nên tập trung nỗ lực vào top 6 thể loại phim có lợi nhuận cao nhất: Phiêu lưu, Hành động, Hài, Chính kịch, Khoa học viễn tưởng và phim giật gân (Thriller).

Một khuyến nghị khác để tập trung vào gia đình và Hoạt hình do ít cạnh tranh hơn và cơ hội kiếm lợi nhuận cao hơn. Bên cạnh đó thể loại phiêu lưu và hành động, được yêu thích và đem lại lợi nhuận trong những tháng mùa hè và phim phiêu lưu,hành động, phim kinh dị và hài kịch sẽ đạt được thành công tương tự nếu phát hành vào tháng 11, nhưng khuyến nghị vẫn tập trung vào mùa hè.

Cân nhắc về việc lựa chọn đạo diễn và diễn viên cho bộ phim, qua biểu đồ doanh thu và lợi nhuận trung bình ta cũng có thể thấy việc doanh thu và lợi nhuận trung đem lại cao không phụ thuộc quá nhiều vào số lượng bộ phim tham gia.

Thời lượng chiếu từ 60 đến 180 phút đóng một vai trò quan trọng trong thành công thương mại của một bộ phim. Phim có độ dài này có xu hướng tạo ra doanh thu và lợi nhuận cao hơn phim ngắn. Đặc biệt, những bộ phim có thời lượng từ 140 đến 180 phút có doanh thu và lợi nhuận tăng đáng kể. Điều này cho thấy khán giả có xu hướng thích những bộ phim dài hơn và sẵn sàng chi tiền cho một trải nghiệm phim kéo dài. Thời lượng phim dài này giúp các nhà làm phim có không gian để phát triển câu chuyện, phát triển nhân vật và tạo ra trải nghiệm sâu sắc cho khán giả.

Tập trung vào phát triển các bộ phim với rating từ 80 đến 90 để thu hút khán giả và đạt được doanh thu cao. Đầu tư vào nội dung, kịch bản và chất lượng sản xuất là quan trọng để tạo ra những bộ phim có rating cao và đem lại lợi nhuận đáng kể.Tận dụng tiềm năng của các bộ phim có rating từ 70 đến 80 bằng cách quảng bá một cách hiệu quả để thu hút khán giả và tạo ra lợi nhuận cao hơn.

Nghiên cứu chi tiết về các khoảng rating có số lượng phim ít để hiểu rõ hơn về tiềm năng và đặc điểm của những bộ phim này. Điều này giúp tìm ra cơ hội mới và phát triển nội dung đa dạng để đáp ứng nhu cầu của khán giả tiềm năng.

Theo dõi và đánh giá các dữ liệu mới liên quan đến danh sách rating phim để hiểu rõ hơn về xu hướng và thay đổi trong thị trường điện ảnh. Điều này giúp các nhà sản xuất và nhà làm phim điều chỉnh chiến lược của họ và tận dụng tối đa tiềm năng kinh doanh trong ngành công nghiệp điện ảnh.

4.2. Kết quả mô hình

– Từ bảng kết quả có thể chọn 2 mô hình để dự đoán đó là mô hình Random Forest Regressor cho dự đoán doanh thu và mô hình Linear Regression cho dự đoán điểm đánh giá.

	Outcome	Model	R Squared Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
0	Revenue	XGB Regressor	0.604627	9.236995e+07	2.605171e+16	1.614054e+08
1	Revenue	Linear Regression	0.570181	9.975265e+07	2.832144e+16	1.682897e+08
2	Revenue	Random Forest Regressor	0.635719	8.988400e+07	2.400301e+16	1.549290e+08
3	Rating	XGB Regressor	0.199596	5.617850e+00	5.377677e+01	7.333265e+00
4	Rating	Linear Regression	0.317781	5.238582e+00	4.583631e+01	6.770252e+00
5	Rating	Random Forest Regressor	0.294594	5.305872e+00	4.739417e+01	6.884343e+00

– Lưu 2 mô hình dự đoán vào file revenue.pkl và rating.pkl

```
rfr = RandomForestRegressor()
rfr.fit(X_train,y_train)
pickle.dump(rfr, open('revenue.pkl', 'wb'))
```

```
lr = LinearRegression()
lr.fit(X_train,y_r_train)
pickle.dump(rfr, open('rating.pkl', 'wb'))
```

– *Dự đoán thành công của phim dựa theo cơ sở của doanh thu và điểm đánh giá với điều kiện cần là doanh thu không bé hơn ngân sách và điều kiện đủ là phải có trên 50% số lượng điểm đánh giá của phim không bé hơn trung bình điểm đánh giá của thể loại mà phim đã chọn.*

– Xây dựng hàm để dự đoán các giá trị mới cụ thể:

+ Tính các chỉ số thống kê gắn vào biến z

```
X = model.drop(columns=['Adjusted_Revenue','id','rating'])
z=X.describe()
z
```

	on_streaming_year	on_streaming_month	on_streaming_day	Action	Adventure	Animation	Comedy	Crime	Documentary	Drama	...
count	3438.000000	3438.000000	3438.000000	3438.000000	3438.000000	3438.000000	3438.000000	3438.000000	3438.000000	3438.000000	...
mean	2010.897324	6.845259	15.893543	0.248691	0.184991	0.068354	0.367365	0.150087	0.011053	0.445899	...
std	6.189961	3.409057	8.580059	0.432317	0.388347	0.252389	0.482157	0.357209	0.104566	0.497137	...
min	2000.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	2006.000000	4.000000	9.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
50%	2011.000000	7.000000	16.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
75%	2016.000000	10.000000	23.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	...
max	2023.000000	12.000000	31.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...

8 rows × 35 columns

+ Tính các giá trị mean của từng thể loại

```
list_means=[]
for col in dummy_df.columns:
    m = movies[movies[col]==1].rating.mean()
    dicts = {'Genre':col,'Mean':m}
    list_means.append(dicts)
mean_genre = pd.DataFrame(list_means)
mean_genre = mean_genre.set_index('Genre').T
mean_genre
```

Genre	Action	Adventure	Animation	Comedy	Crime	Documentary	Drama	Family	Fantasy	History	Horror	Music	Mystery	Romance	Science Fiction	TV Movie	Thriller	War	Western
Mean	62.499415	64.47956	66.92766	62.425178	64.217054	68.210526	66.26484	64.647355	63.855153	68.5	59.726087	67.283186	62.279279	64.053601	62.803191	62.0	62.26087	68.053763	63.609756

+ Xây dựng hàm TEST() để chuyển đổi các giá trị nhập vào thành các giá trị chuẩn hóa theo phân phối chuẩn của X

```
def TEST(test,z):
    listtest=[]
    dict_test={}
    for k in test:
        if k == 'On_streaming':
            test[k] = pd.to_datetime(test[k])
            dict_test['on_streaming_day'] = (test[k].day-z['on_streaming_day'][1])/z['on_streaming_day'][2]
            dict_test['on_streaming_month'] = (test[k].month-z['on_streaming_month'][1])/z['on_streaming_month'][2]
            dict_test['on_streaming_year'] = (test[k].year-z['on_streaming_year'][1])/z['on_streaming_year'][2]
        elif k == 'Runtime':
            if 'h' not in test[k]:
                hours = '0'
                minutes = test[k]
            else:
                hours, minutes = test[k].split('h')
                minutes = minutes[:-1].strip()
            total_minutes = int(hours) * 60 + int(minutes)
            dict_test['runtime'] = (total_minutes-z['runtime'][1])/z['runtime'][2]
        elif k == 'Budget':
            year = pd.to_datetime(test['On_streaming']).year
            Adjusted = ((((2023-year)*0.0322)+1)*test[k])
            dict_test['Adjusted_Budget'] = (Adjusted-z['Adjusted_Budget'][1])/z['Adjusted_Budget'][2]
        elif k == 'Genre':
            genres = test[k].split(',')
            genres = [genre.strip() for genre in genres]
            for i in genres:
                if i in TL:
                    dict_test[i] = (1-z[i][1])/z[i][2]
            diff_genres = list(set(TL) - set(genres))
            for i in diff_genres:
                dict_test[i] = (0-z[i][1])/z[i][2]
        elif k == 'Original Language':
            dict_test['Original Language_' + str(test[k])] = (1-z['Original Language_' + str(test[k])][1])/z['Original Language_' + str(test[k])][2]
            diff_lang = list(set(language) - set([test[k]]))
            for i in diff_lang:
                dict_test['Original Language_' + i] = (1-z['Original Language_' + i][1])/z['Original Language_' + i][2]
        elif k in ['Director','cast1','cast2','cast3']:
            num = casts[casts['title']==test[k]]['number_domestic']
            dict_test[str(k) + '_number_domestic'] = (1-z[str(k) + '_number_domestic'][1])/z[str(k) + '_number_domestic'][2]
            credit = casts[casts['title']==test[k]]['Known Credits']
            dict_test[str(k) + '_known_credits'] = (1-z[str(k) + '_known_credits'][1])/z[str(k) + '_known_credits'][2]
    column_names = X.columns.tolist()
    sorted_keys = sorted(dict_test, key=lambda x: column_names.index(x))
    for key in sorted_keys:
        listtest.append(dict_test[key])
    user_input=np.array([listtest])
    return user_input, Adjusted
```

+ Tạo hàm results() để đưa ra màn hình các kết quả

```
def result(dl,Adjusted):
    revenue_model = pickle.load(open('revenue.pkl', 'rb'))
    DT = revenue_model.predict(dl)[0]
    rating_model = pickle.load(open('rating.pkl', 'rb'))
    RA = rating_model.predict(dl)[0]
    genres = test['Genre'].split(',')
    genres = [genre.strip() for genre in genres]
    kq=[]
    if DT<Adjusted:
        kq1=0
    else:
        kq1=1
    for g in genres:
        if mean_genre[g][0]>RA:
            kq.append(0)
        else:
            kq.append(1)
    name = test['Title']
    print(f'Ngân sách đã điều chỉnh của phim {name} là {Adjusted:0.2f}$')
    print(f'Doanh thu dự đoán của phim {name} là {DT:0.2f}$')
    print(f'Điểm đánh giá dự đoán của phim {name} là {RA:0.2f}%')
    if (kq1==1) & (statistics.mode(kq)==1):
        print('Dự đoán phim "'+name+'" sẽ thành công')
    else:
        print('Dự đoán phim "'+name+'" sẽ không thành công')
    print(kq1,kq)
```

- Dự đoán giá trị cụ thể:

```
test={'Title':'Revolver',
      'On_streaming':'22/08/2024',
      'Genre':'Documentary,Drama, Family',
      'Runtime':'1h 40m',
      'Director':'Guy Ritchie',
      'cast1':'Jason Statham',
      'cast2':'Ray Liotta',
      'cast3':'Vincent Pastore',
      'Original Language':'English',
      'Budget':426448245694}
dl,Adjusted = TEST(test,z)
result(dl,Adjusted)
```

```
Ngân sách đã điều chỉnh của phim Revolver là 412716612182.65$
Doanh thu dự đoán của phim Revolver là 1339682913.60$
Điểm đánh giá dự đoán của phim Revolver là 73.68%
Dự đoán phim "Revolver" sẽ không thành công
```

```
test={'Title':'VietNam airline',
      'On_streaming':'20/10/2024',
      'Genre':'Action, Adventure, Animation, Comedy',
      'Runtime':'1h 25m',
      'Director':'Ken Kwapis',
      'cast1':'Ray Liotta',
      'cast2':'Vincent Pastore',
      'cast3':'Catherine Keener',
      'Original Language':'English',
      'Budget':80898000}
dl,Adjusted = TEST(test,z)
result(dl,Adjusted)
```

```
Ngân sách đã điều chỉnh của phim VietNam airline là 78293084.40$
Doanh thu dự đoán của phim VietNam airline là 263508849.73$
Điểm đánh giá dự đoán của phim VietNam airline là 72.55%
Dự đoán phim "VietNam airline" sẽ thành công
```

CHƯƠNG 4: KẾT LUẬN

Nghiên cứu này đã thành công trong việc xây dựng một mô hình dự đoán thành công của một bộ phim dựa trên các yếu tố quan trọng như rating, thể loại, đạo diễn, diễn viên và nhiều yếu tố khác. Kết quả dự đoán từ mô hình đã cung cấp thông tin giá trị để đánh giá tiềm năng thành công của các bộ phim mới.

Việc sử dụng mô hình dự đoán trong quá trình ra quyết định đầu tư và phát triển các dự án điện ảnh có thể mang lại lợi ích to lớn. Thay vì dựa vào cảm quan và trực giác, các quyết định có thể được đưa ra dựa trên dữ liệu và dự đoán chính xác về thành công tiềm năng của một bộ phim.

Tóm lại, nghiên cứu này đã tạo ra một công cụ hữu ích giúp đánh giá và dự đoán thành công của các bộ phim mới, góp phần hỗ trợ trong quyết định đầu tư và phát triển trong lĩnh vực điện ảnh.