ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

BÀI TẬP LỚN MÔN KHAI PHÁ DỮ LIỆU (CO3029)

Phân tích và dự báo xu hướng của ngành công nghiệp điện ảnh bằng kỹ thuật Long Short-Term Memory

---

Giảng viên hướng dẫn: ThS. Bùi Tiến Đức
HK242

STT	Họ và tên	MSSV
1	Huỳnh Huy Mân	2433003
2	Đoàn Nhật Tiến	2213449
3	Bùi Thanh Tùng	2213860
4	Trần Nguyễn Anh Khoa	2211651

Thành phố Hồ Chí Minh, tháng 5 năm 2025

Mục lục

Bảng phân công nhiệm vụ

STT	MSSV	Họ và tên	Đóng góp	Nhiệm vụ
1	2433003	Huỳnh Huy Mân	100%	Tìm kiếm và tiền xử lý dữ liệu, viết báo cáo, làm slide
2	2213449	Đoàn Nhật Tiến	100%	Phân tích, xây dựng mô hình dự đoán, viết báo cáo tổng kết
3	2213860	Bùi Thanh Tùng	100%	Khai phá thông tin từ dữ liệu, viết báo cáo, làm slide
4	2211651	Trần Nguyễn Anh Khoa	100%	Phân tích các xu hướng, viết báo cáo, làm slide

Lời nói đầu

Lời đầu tiên, nhóm chúng tôi xin gửi lời cảm ơn chân thành đến ThS. Bùi Tiến Đức đã tận tình giảng dạy và tạo điều kiện thuận lợi cho nhóm trong suốt quá trình thực hiện bài tập này. Qua bài tập lớn, mỗi thành viên không chỉ củng cố được kiến thức lý thuyết mà còn có cơ hội áp dụng các phương pháp phân tích và xử lý dữ liệu vào một bài toán thực tế, từ đó giúp hiểu sâu hơn về các kỹ thuật trong Data Mining.

Trong quá trình thực hiện đề tài, các thành viên trong nhóm đã học hỏi được rất nhiều kiến thức bổ ích, từ tiền xử lý dữ liệu, phân tích, khai phá thông tin tiềm ẩn cho đến trực quan hóa kết quả. Đồng thời, nhóm cũng cải thiện được kỹ năng lập trình Python, xử lý dữ liệu, và tư duy giải quyết vấn đề. Bài tập này không chỉ giúp nhóm rèn luyện kỹ năng phân tích dữ liệu mà còn nâng cao khả năng viết báo cáo khoa học và hiểu sâu về các khái niệm như khai phá dữ liệu, mạng nơ ron, ... những kỹ năng rất hữu ích cho học tập và công việc trong tương lai.

Một lần nữa, nhóm xin chân thành cảm ơn ThS. Bùi Tiến Đức đã giao đề tài này, giúp nhóm không chỉ học hỏi thêm mà còn có cơ hội gắn kết và làm việc nghiêm túc, chủ động để hoàn thành bài báo cáo đúng tiến độ.

Nhóm sinh viên thực hiện

1 Giới thiệu

Nền công nghiệp điện ảnh đã không ngừng chuyển mình trong suốt thế kỷ 20 – từ thời kỳ phim câm, phim đen trắng đến kỷ nguyên kỹ thuật số và nền tảng phát trực tuyến. Song hành với sự phát triển công nghệ, dữ liệu phim ảnh trở thành một tài sản quan trọng, giúp các nhà làm phim và nhà đầu tư hiểu rõ hơn về thị hiếu khán giả, xu hướng nội dung và hiệu quả thương mại.

Trong nghiên cứu này, chúng tôi khai thác bộ dữ liệu từ *The Movie Database (TMDB)* nhằm phân tích đặc điểm phát triển của ngành điện ảnh và xây dựng một mô hình học sâu để dự đoán xu hướng trong tương lai. Cụ thể, chúng tôi sử dụng mạng nơ-ron hồi tiếp dài-ngắn hạn (LSTM – *Long Short-Term Memory*), một phương pháp hiệu quả trong việc học từ dữ liệu tuần tự như chuỗi doanh thu theo năm, thể loại và mức độ phổ biến.

Nghiên cứu hướng đến việc rút ra các đặc trưng nội dung, xu hướng thể loại, và các yếu tố ảnh hưởng đến thành công của phim ảnh, đồng thời chứng minh khả năng áp dụng của LSTM trong lĩnh vực giải trí.

2 Cơ sở lý thuyết

Trong phần này, chúng tôi trình bày các khái niệm và thuật toán nền tảng được sử dụng trong quá trình phân tích và dự đoán, với trọng tâm là mạng LSTM. Ngoài ra, một số khái niệm liên quan đến xử lý dữ liệu thời gian và mô hình học sâu cũng được đề cập để hỗ trợ việc hiểu rõ hơn về phương pháp tiếp cận.

2.1 Mạng nơ-ron hồi tiếp (RNN)

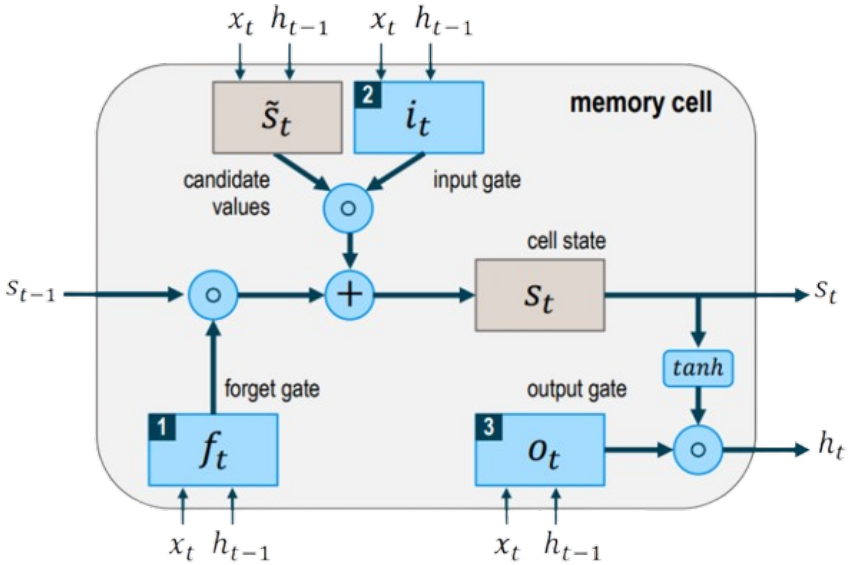
Recurrent Neural Network (RNN) là một kiến trúc mạng chuyên xử lý dữ liệu tuần tự như văn bản, chuỗi tín hiệu hoặc chuỗi thời gian. Khác với mạng nơ-ron truyền thống, RNN giữ lại thông tin từ các bước thời gian trước thông qua trạng thái ẩn.

Mặc dù hiệu quả với chuỗi ngắn, RNN truyền thống gặp khó khăn với chuỗi dài do hiện tượng *vanishing gradient*, làm giảm khả năng học các phụ thuộc dài hạn trong dữ liệu.

2.2 Mạng LSTM (Long Short-Term Memory)

LSTM là một biến thể cải tiến của RNN, được thiết kế để khắc phục nhược điểm về việc ghi nhớ các phụ thuộc dài hạn trong chuỗi dữ liệu. Khối LSTM bao gồm ba cổng chính:

- **Forget gate**: quyết định thông tin nào nên bị loại khỏi trạng thái bộ nhớ.
- **Input gate**: xác định thông tin mới nào sẽ được ghi nhớ.
- **Output gate**: quyết định thông tin nào sẽ được truyền ra ngoài.



Hình 1: Kiến trúc bên trong một khối LSTM [?]

Hình 1 mô tả cấu trúc nội tại của một khối LSTM, với các luồng thông tin qua các cổng và trạng thái bộ nhớ. Nhờ cơ chế này, LSTM có khả năng học và ghi nhớ các phụ thuộc dài hạn trong chuỗi thời gian, chẳng hạn như xu hướng doanh thu phim hoặc sự dịch chuyển thể loại qua các thập kỷ.

2.3 LSTM trong phân tích dữ liệu điện ảnh

Trong bài toán này, dữ liệu phim được sắp xếp theo trục thời gian (theo năm phát hành), với các đặc trưng như doanh thu, độ phổ biến, số lượng phim mỗi thể loại, v.v.

Đây là dạng dữ liệu chuỗi điển hình, rất phù hợp với mô hình LSTM.

Bằng cách huấn luyện mạng LSTM trên các chuỗi dữ liệu quá khứ, chúng tôi có thể dự đoán xu hướng biến động của thị trường điện ảnh, ví dụ như thể loại nào đang dần trở nên phổ biến, hay doanh thu trung bình của phim theo năm.

3 Mô tả về quy trình xử lý và phân tích dữ liệu

Dưới đây là quy trình làm sạch, xử lý và phân tích dữ liệu từ bộ dữ liệu các bộ phim.

3.1 Tải Dữ Liệu

Đầu tiên, chúng ta tải dữ liệu từ file CSV bằng hàm `load_data`:

Hàm `load_data` sẽ tải dữ liệu và cung cấp thông tin cơ bản như số lượng giá trị thiếu, số dòng trùng lặp theo cột `title`, cũng như số dòng riêng biệt.

3.2 Làm Sạch Dữ Liệu

Tiếp theo, chúng ta thực hiện làm sạch dữ liệu với hàm `clean_data`:

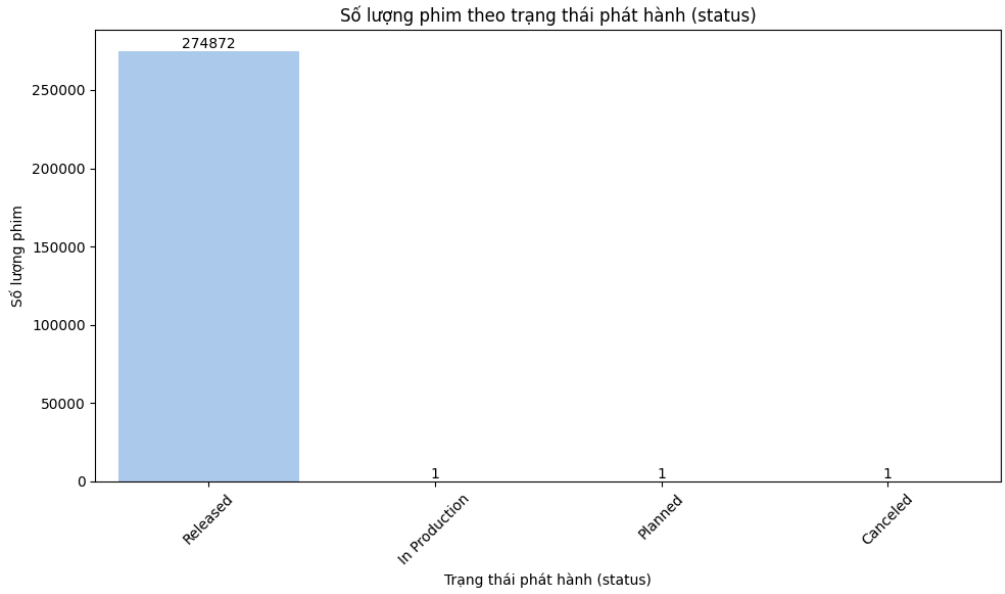
Hàm `clean_data` thực hiện các bước sau:

1. Loại bỏ các dòng trùng lặp theo cột `title`.
2. Xử lý giá trị thiếu trong các cột như `status` và `genres`.
3. Chuyển đổi cột `release_date` thành kiểu `datetime`.
4. Tách cột `genres` thành danh sách.
5. Thêm các cột `year` và `month` từ `release_date`.
6. Chỉ giữ lại các bộ phim có năm phát hành trong thế kỷ 20 (1900-1999).

Sau khi làm sạch, bộ dữ liệu được giảm thiểu với các cột và dòng không cần thiết.

3.3 Phân Tích Phân Phối Trạng Thái Phim

Để hiểu rõ hơn về trạng thái phát hành của các bộ phim, chúng ta sử dụng hàm `plot_status_distribution`:

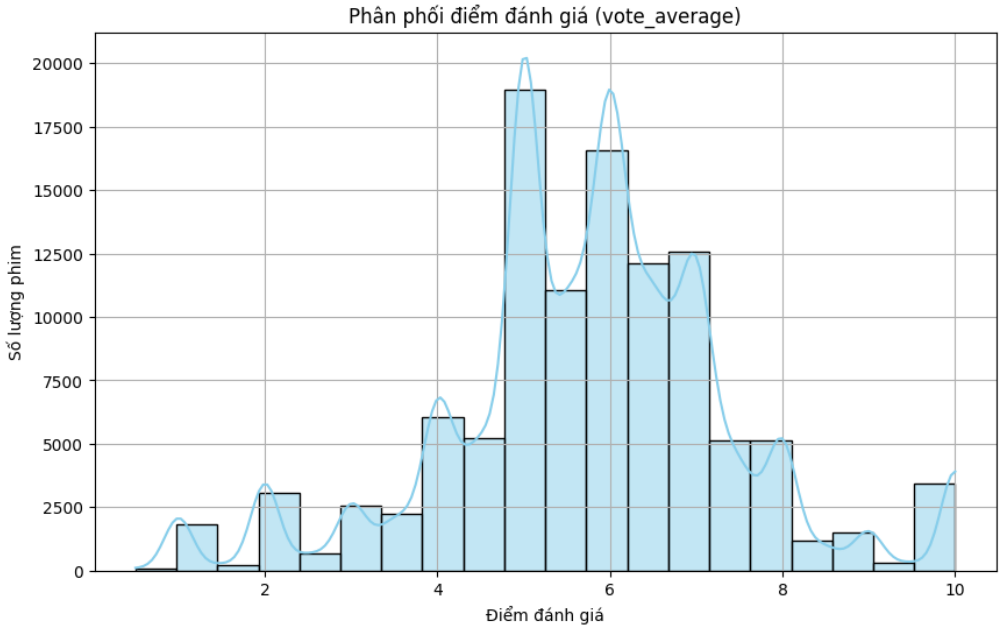


Hình 2: Phân phối số lượng phim theo trạng thái phát hành

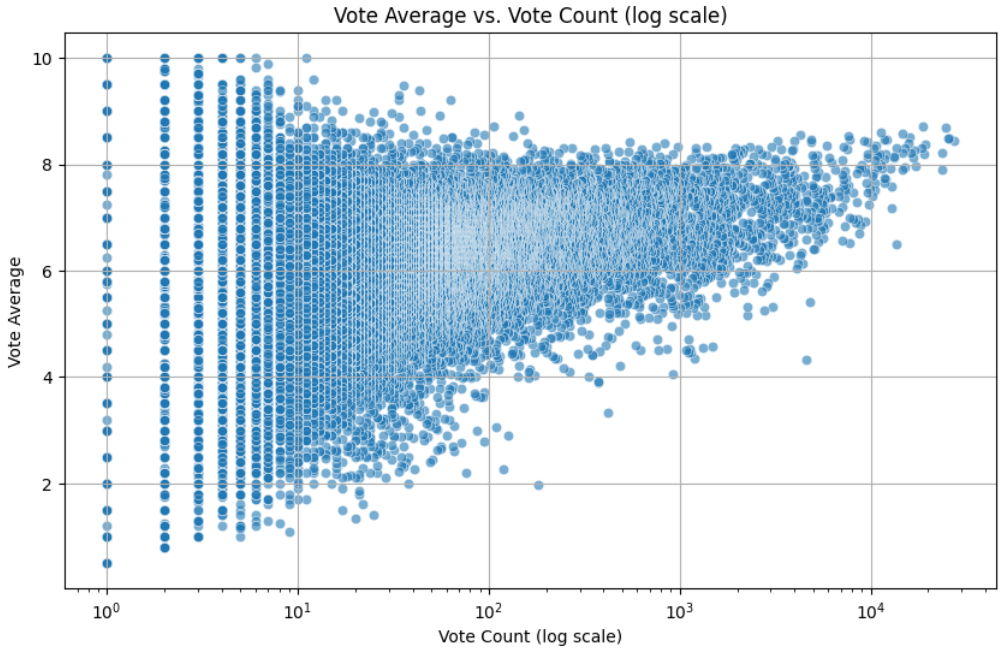
Biểu đồ trên thể hiện số lượng phim theo từng trạng thái phát hành, giúp chúng ta thấy được phần lớn phim trong bộ dữ liệu đã được phát hành (Released), trong khi một số nhỏ đang trong các trạng thái khác.

3.4 Phân Tích Đánh Giá Phim

Chúng ta tiếp tục phân tích đánh giá phim bằng hàm `analyze_votes`:



Hình 3: Phân phối điểm đánh giá của phim



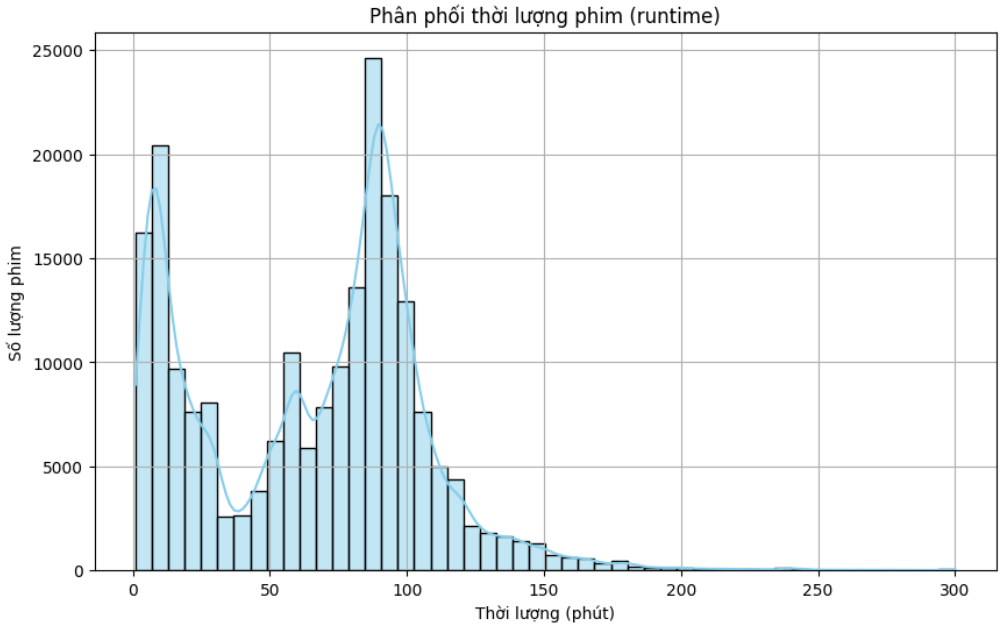
Hình 4: Mối quan hệ giữa số lượng đánh giá và điểm đánh giá trung bình

Hai biểu đồ trên cho thấy phân phối điểm đánh giá và mối quan hệ giữa số lượng đánh giá (vote_count) và điểm đánh giá trung bình (vote_average). Phân phối điểm đánh giá tập trung trong khoảng 6-7 điểm, trong khi biểu đồ thứ hai cho thấy phim

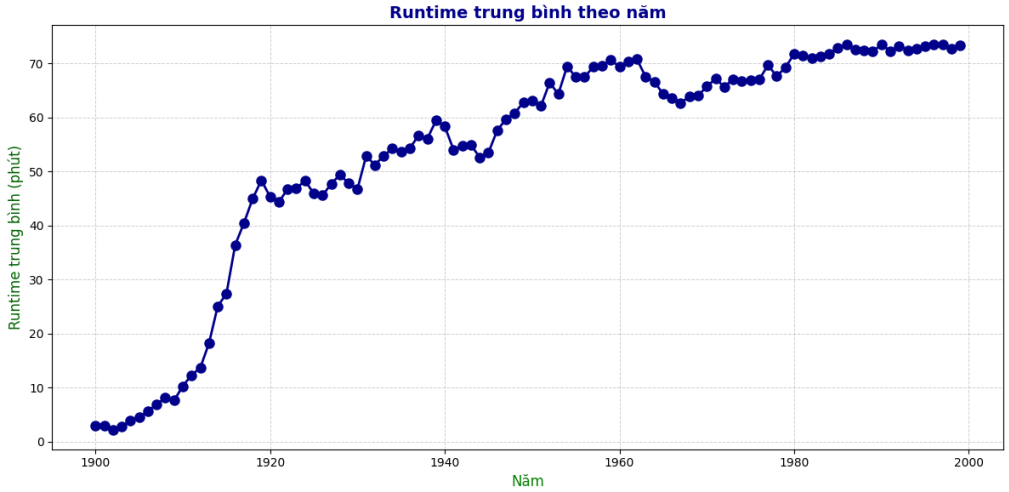
có nhiều lượt đánh giá thường có điểm trung bình tốt hơn.

3.5 Phân Tích Thời Lượng Phim

Chúng ta phân tích thời lượng phim bằng hàm `analyze_runtime`:



Hình 5: Phân phối thời lượng phim



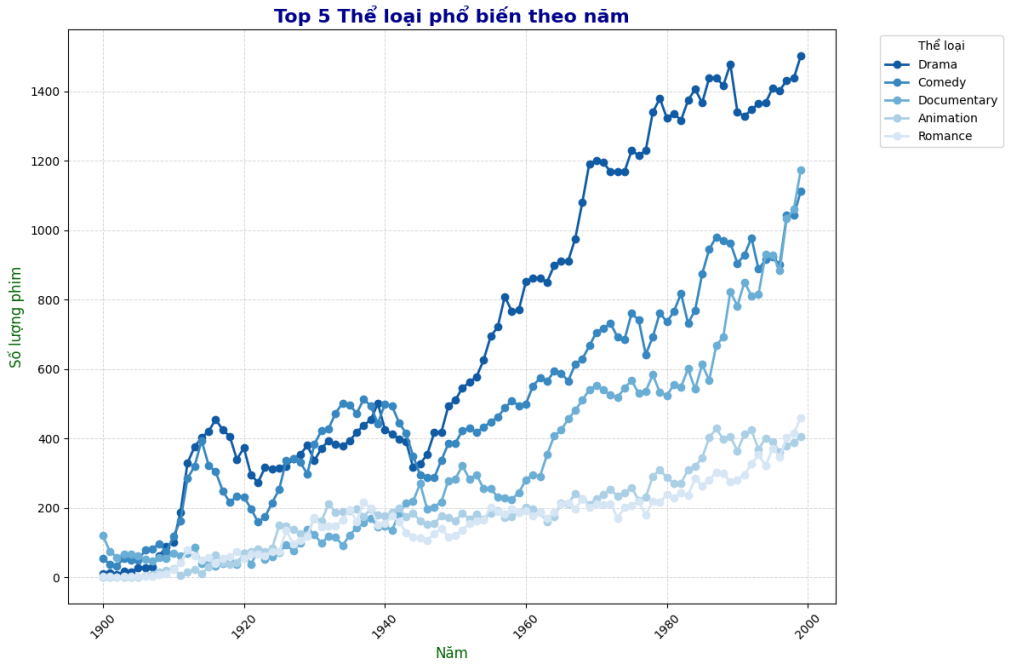
Hình 6: Thời lượng phim trung bình theo năm

Biểu đồ phân phối thời lượng phim cho thấy phần lớn phim có thời lượng khoảng

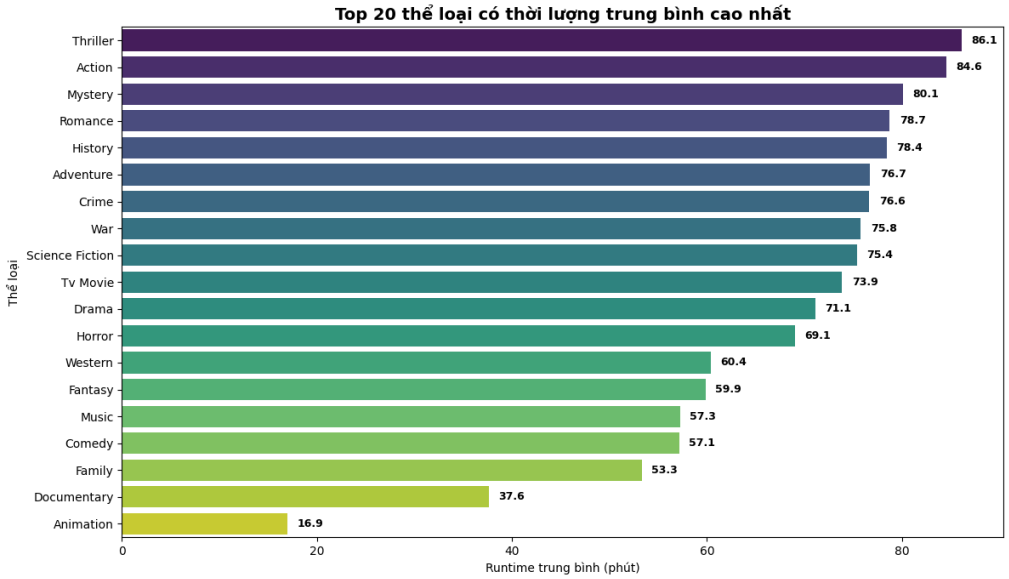
90-120 phút. Biểu đồ thứ hai thể hiện xu hướng thời lượng phim qua các năm, với sự dao động đáng kể nhưng có xu hướng tăng dần qua thời gian.

3.6 Phân Tích Thể Loại Phim

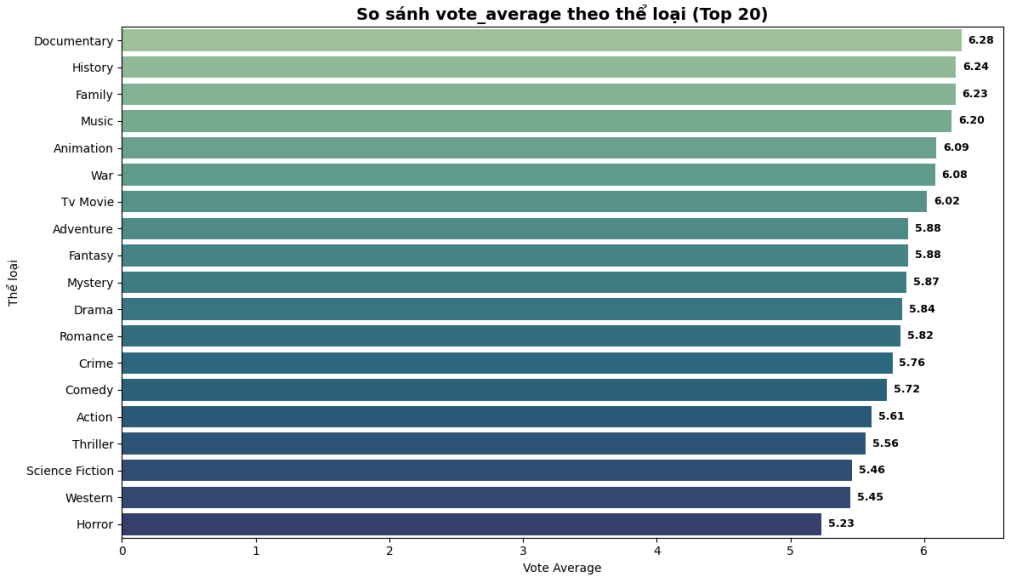
Tiếp theo, chúng ta phân tích thể loại phim bằng hàm `analyze_genres`:



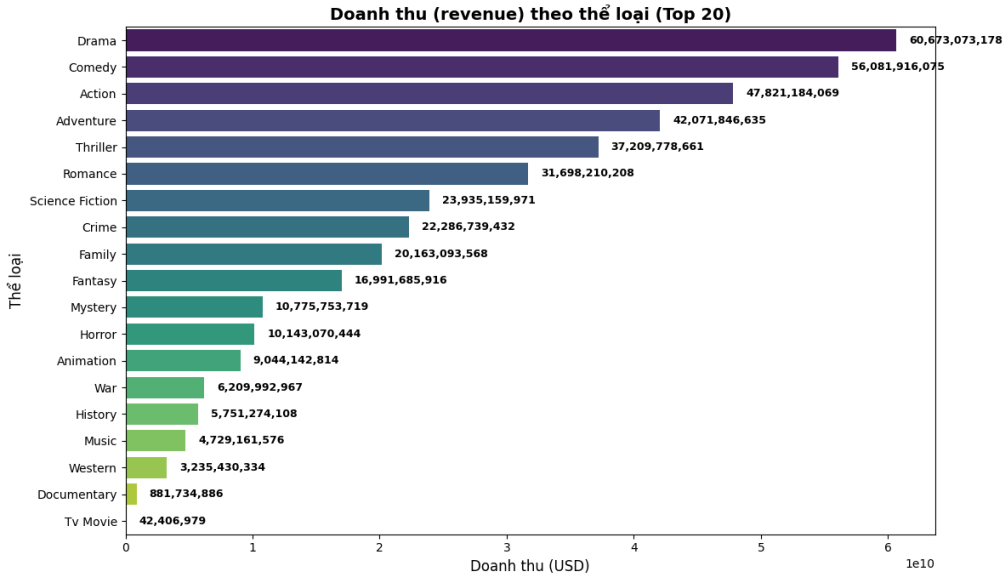
Hình 7: Top 5 thể loại phim phổ biến theo năm



Hình 8: Thời lượng trung bình theo thể loại phim



Hình 9: Điểm đánh giá trung bình theo thể loại phim

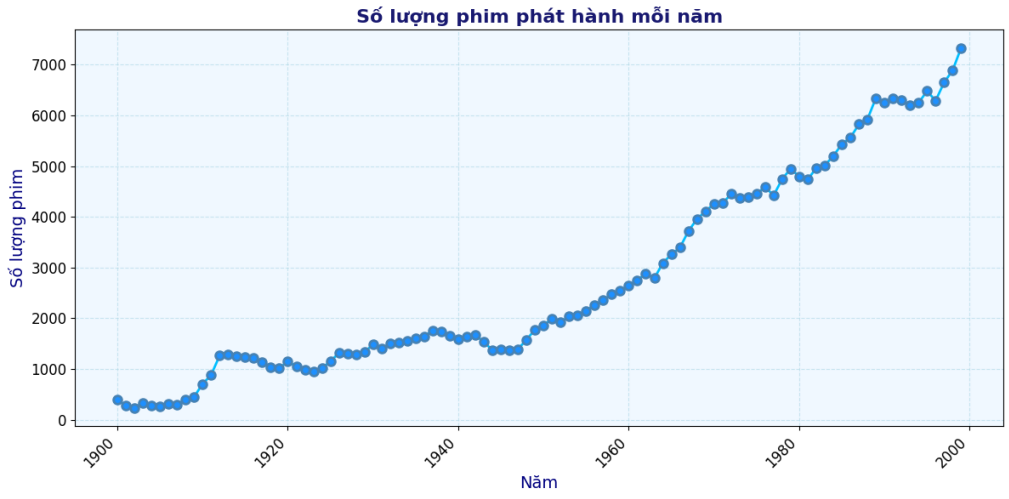


Hình 10: Doanh thu theo thể loại phim

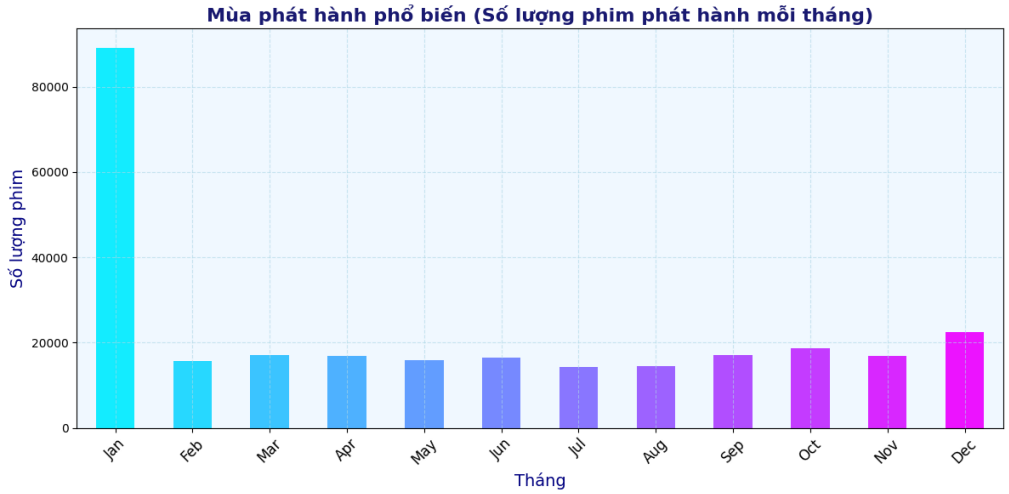
Các biểu đồ trên cho thấy xu hướng của các thể loại phim phổ biến qua các năm, thời lượng trung bình theo thể loại, điểm đánh giá trung bình và doanh thu theo từng thể loại. Có thể thấy một số thể loại như Drama và Comedy luôn dẫn đầu về số lượng, trong khi các thể loại như History và War có thời lượng trung bình cao hơn.

3.7 Phân Tích Phim Theo Năm và Tháng

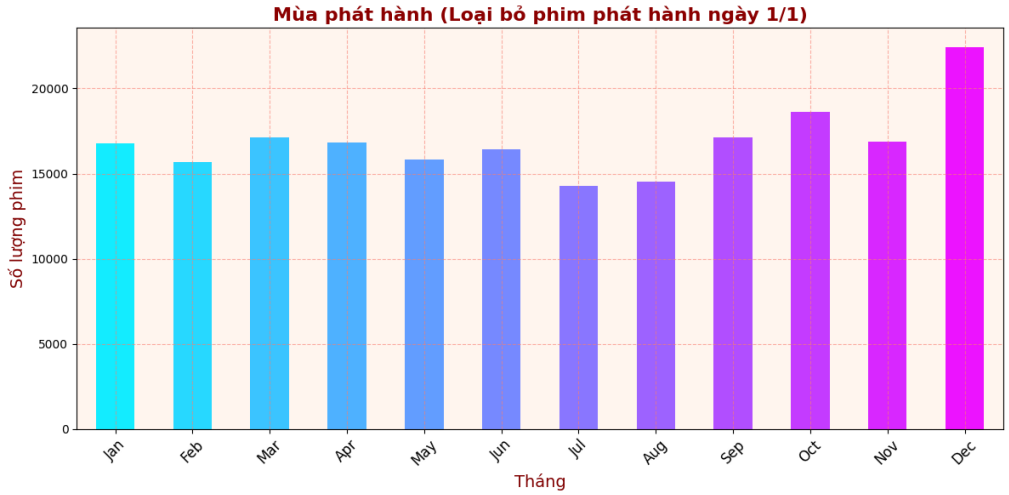
Chúng ta tiếp tục phân tích số lượng phim theo năm và tháng bằng các hàm `plot_movies_per_year` và `plot_movies_per_month`:



Hình 11: Số lượng phim phát hành mỗi năm



Hình 12: Số lượng phim phát hành mỗi tháng

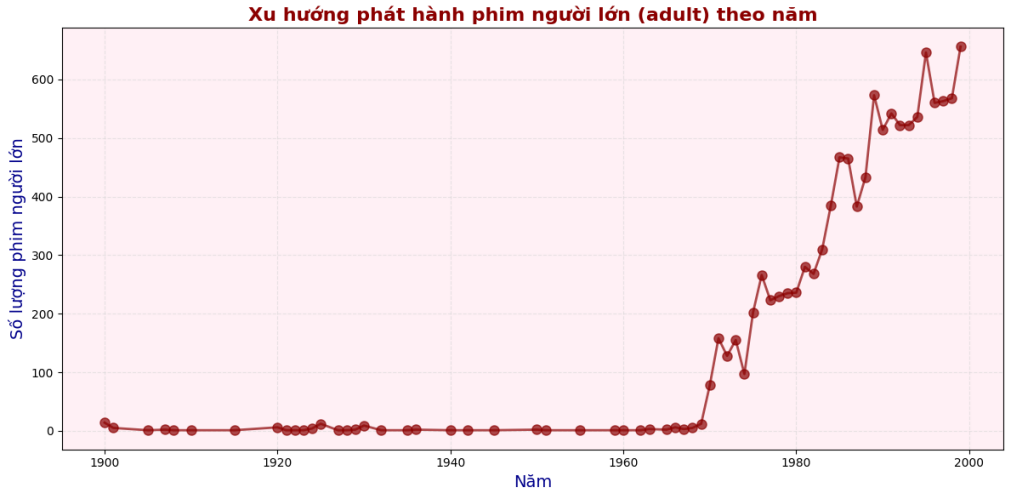


Hình 13: Số lượng phim phát hành mỗi tháng (loại bỏ ngày 1/1)

Biểu đồ số lượng phim theo năm cho thấy sự tăng trưởng mạnh về số lượng phim sản xuất theo thời gian. Biểu đồ theo tháng cho thấy tháng 1 có số lượng phim phát hành cao đột biến, nhưng khi loại bỏ các phim phát hành ngày 1/1 (thường là ngày mặc định), phân phối theo tháng trở nên cân đối hơn.

3.8 Phân Tích Phim Người Lớn

Chúng ta phân tích xu hướng phát hành phim người lớn theo năm:

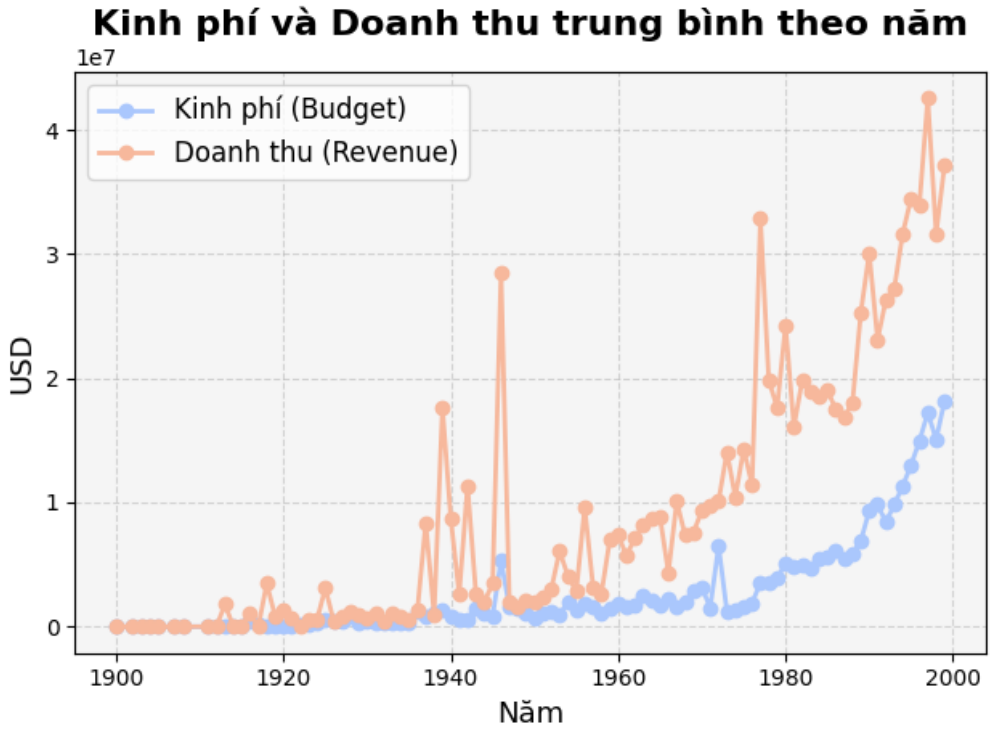


Hình 14: Số lượng phim người lớn phát hành theo năm

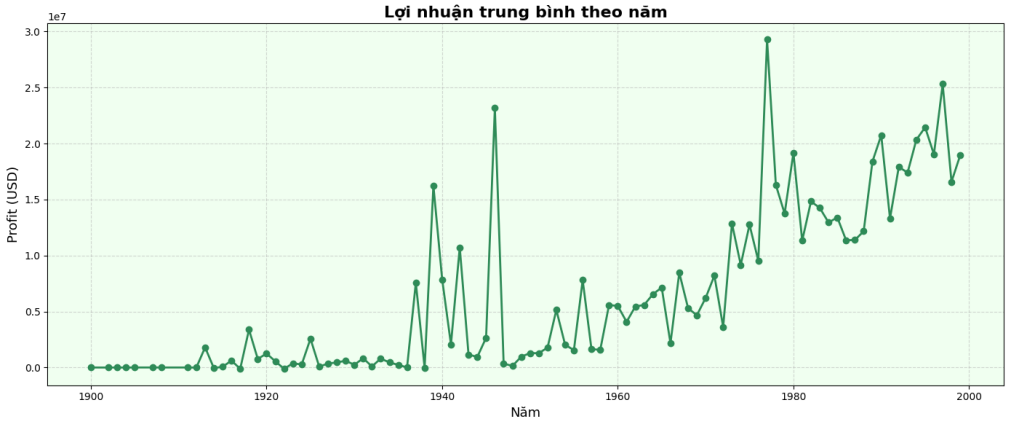
Biểu đồ trên thể hiện xu hướng phát hành phim người lớn theo các năm, cho thấy sự biến động theo thời gian với một số đỉnh điểm đáng chú ý.

3.9 Phân Tích Xu Hướng Tài Chính

Chúng ta phân tích xu hướng tài chính của các bộ phim:



Hình 15: Kinh phí và Doanh thu trung bình theo năm



Hình 16: Lợi nhuận trung bình theo năm

Hai biểu đồ trên thể hiện xu hướng kinh phí, doanh thu và lợi nhuận trung bình của các bộ phim theo năm. Có thể thấy cả kinh phí và doanh thu đều có xu hướng tăng theo thời gian, đặc biệt từ những năm 1960 trở đi. Lợi nhuận trung bình cũng có xu hướng tăng tương tự, mặc dù có một số biến động đáng chú ý vào những năm nhất định.

4 Chuẩn Bị Dữ Liệu Cho Mô Hình Dự Đoán

Để chuẩn bị cho việc xây dựng mô hình dự đoán, chúng ta thực hiện các bước làm sạch và trích xuất dữ liệu cần thiết:

4.1 Trích Xuất và Làm Sạch Dữ Liệu

Hàm `clean_and_extract_data` thực hiện các thao tác sau:

Hàm này trích xuất các cột cần thiết, chuyển đổi `release_date` thành kiểu datetime, lọc phim trong khoảng năm 1900-2025, loại bỏ các giá trị âm, xử lý giá trị thiếu, và thêm cột `movies_count` để đếm số lượng phim mỗi năm.

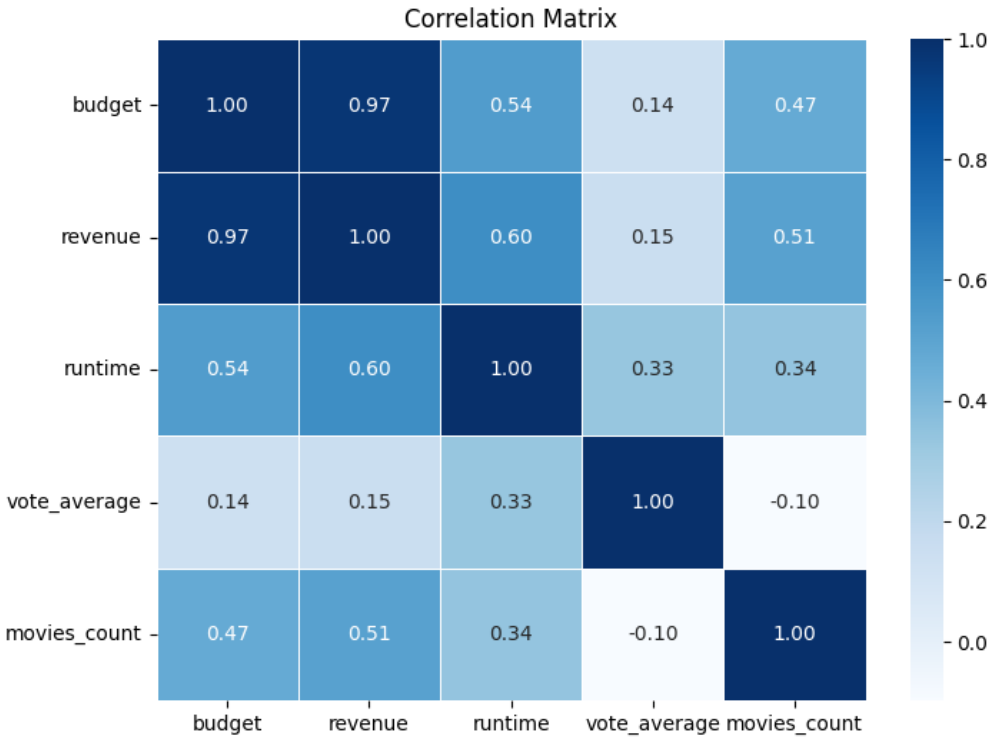
4.2 Nhóm Dữ Liệu Theo Năm

Sau khi làm sạch dữ liệu, chúng ta nhóm dữ liệu theo năm để chuẩn bị cho việc xây dựng mô hình:

Hàm này nhóm dữ liệu theo năm và tính giá trị trung bình cho các biến như kinh phí, doanh thu, thời lượng, điểm đánh giá và số lượng phim.

4.3 Phân Tích Tương Quan

Trước khi xây dựng mô hình, chúng ta phân tích mối tương quan giữa các biến:

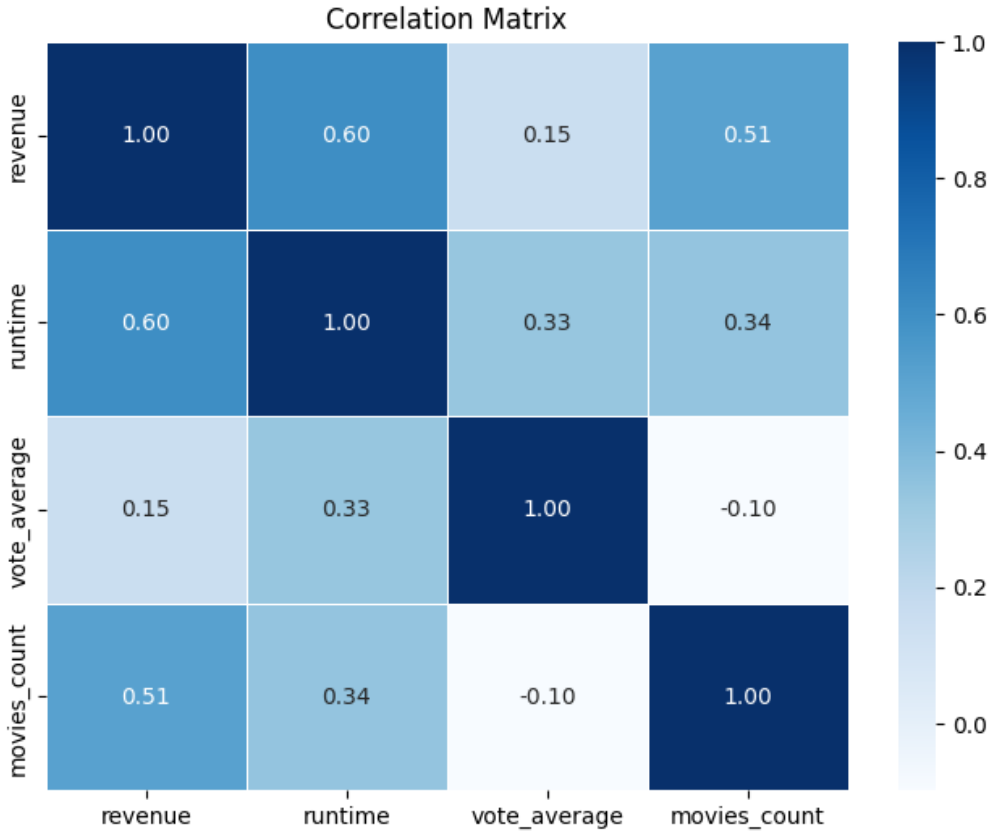


Hình 17: Ma trận tương quan giữa các biến

Biểu đồ ma trận tương quan cho thấy mối quan hệ giữa các biến. Có thể thấy có một số cặp biến có tương quan khá cao, điều này có thể gây ra vấn đề đa cộng tuyến khi xây dựng mô hình.

4.4 Loại Bỏ Biến Có Tương Quan Cao

Để giảm thiểu vấn đề đa cộng tuyến, chúng ta loại bỏ các biến có tương quan quá cao:

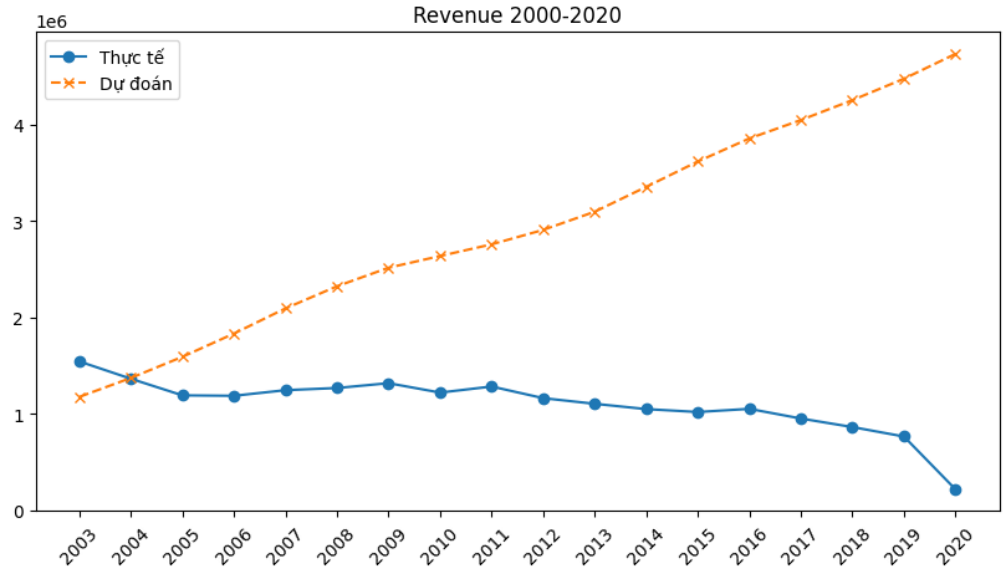


Hình 18: Ma trận tương quan sau khi loại bỏ biến có tương quan cao

Sau khi loại bỏ các biến có tương quan cao (>0.9), ma trận tương quan mới có cấu trúc hợp lý hơn cho việc xây dựng mô hình.

5 Xây Dựng Mô Hình Dự Đoán

Chúng ta xây dựng mô hình LSTM để dự đoán doanh thu phim dựa trên dữ liệu lịch sử:

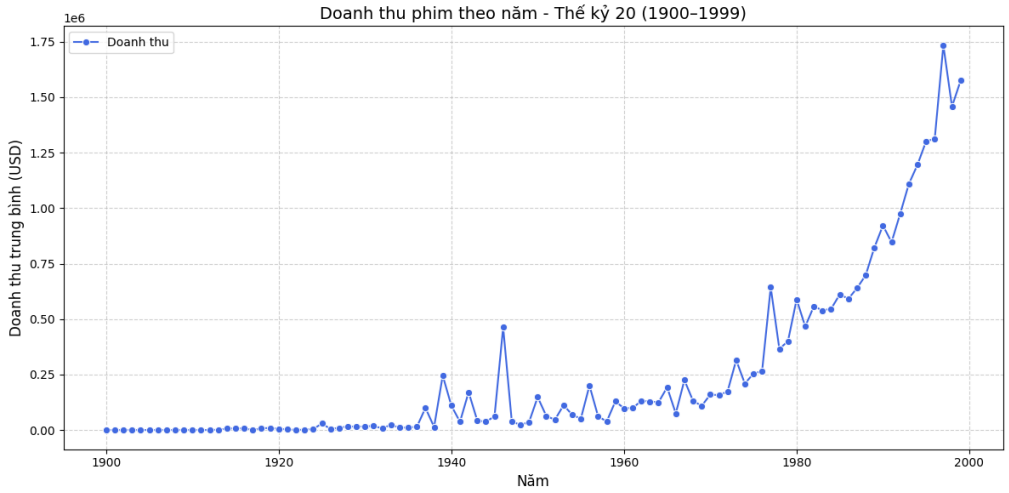


Hình 19: Kết quả dự đoán doanh thu phim (2000-2020)

Mô hình LSTM được huấn luyện trên dữ liệu thế kỷ 20 (1900-1999) và dự đoán doanh thu phim trong giai đoạn 2000-2020. Biểu đồ trên so sánh giá trị doanh thu thực tế với giá trị dự đoán. Mặc dù có một số sai lệch, mô hình đã bắt được xu hướng chung của doanh thu phim trong giai đoạn này.

6 Phân Tích Doanh Thu Thế Kỷ 20

Cuối cùng, chúng ta phân tích doanh thu phim trong thế kỷ 20:



Hình 20: Doanh thu phim theo năm trong thế kỷ 20 (1900-1999)

Biểu đồ doanh thu phim trong thế kỷ 20 cho thấy xu hướng tăng đáng kể từ những năm 1970, với một số biến động mạnh vào các giai đoạn cuối thế kỷ.

7 Kết luận và phương hướng phát triển

7.1 Kết luận

Báo cáo đã phân tích chuyên sâu dữ liệu phim từ đầu thế kỷ 20, mang lại nhiều phát hiện đáng chú ý về sự phát triển của ngành điện ảnh:

- **Tăng trưởng sản xuất:** Số lượng phim tăng mạnh, đặc biệt từ thập niên 1970, phản ánh sự mở rộng quy mô ngành công nghiệp.
- **Thể loại phong phú:** Drama và Comedy phổ biến nhất, trong khi các thể loại như Adventure và Science Fiction lại dẫn đầu về doanh thu, cho thấy khác biệt giữa thị hiếu và tiềm năng thương mại.
- **Xu hướng tài chính:** Kinh phí và doanh thu trung bình ngày càng cao, đặc biệt từ những năm 1980, thể hiện mức đầu tư lớn hơn vào các dự án điện ảnh.
- **Thời điểm phát hành chiến lược:** Sau khi loại bỏ dữ liệu bất thường, phim thường ra mắt vào mùa hè và cuối năm—giai đoạn vàng để tiếp cận khán giả.
- **Thời lượng tối ưu:** Phim dài khoảng 90–120 phút là phổ biến, cân bằng giữa nội dung và trải nghiệm người xem.
- **Dự đoán xu hướng khả thi:** Mô hình LSTM cho thấy tiềm năng trong việc dự đoán doanh thu dựa trên dữ liệu lịch sử.

Những phát hiện này cung cấp nền tảng hữu ích cho việc đưa ra quyết định trong sản xuất và phân phối phim.

7.2 Phương hướng phát triển

Dựa trên kết quả hiện tại, một số hướng nghiên cứu và ứng dụng tiếp theo được đề xuất:

1. **Mở rộng dữ liệu:** Kết hợp thêm dữ liệu từ IMDb, Rotten Tomatoes và mạng xã hội để phân tích đa chiều hơn.

2. **Phân tích theo khu vực:** Nghiên cứu xu hướng từng quốc gia nhằm tối ưu chiến lược phát hành toàn cầu.

3. **Phân tích yếu tố nhân sự:** Đánh giá tác động của đạo diễn, diễn viên, nhà sản xuất đến thành công phim qua phân tích mạng lưới.

4. **Nâng cao mô hình dự đoán:** Áp dụng mô hình học sâu và mô hình kết hợp để dự đoán doanh thu, đánh giá và thời lượng hiệu quả.

5. **Phân tích nội dung:** Sử dụng NLP và thị giác máy để khai thác nội dung phim, từ đó đánh giá ảnh hưởng đến hiệu quả thương mại.

6. **Hệ thống khuyến nghị thông minh:** Hỗ trợ nhà sản xuất xác định thời điểm phát hành, ngân sách và hướng nội dung dựa trên dữ liệu.

Những định hướng này không chỉ mở rộng phạm vi nghiên cứu mà còn mang tính ứng dụng cao, góp phần nâng cao hiệu quả ra quyết định trong ngành điện ảnh.

8 Phụ lục

8.1 Nguồn dữ liệu

- **Bộ dữ liệu phim (TMDB)**: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
- **Mã nguồn**: https://github.com/khoahotran/Movie-CO3029

8.2 Tài liệu tham khảo

- **Mô hình LSTM cho dự đoán doanh thu**
  - TensorFlow LSTM: https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM
  - Keras Time Series: https://keras.io/examples/timeseries/
  - Mạng LSTM: https://nguyentruonglong.net/giai-thich-chi-tiet-ve-mang-long-short-term-memory-lstm.html
- **Phân tích chuỗi thời gian**
  - TensorFlow Time Series: https://www.tensorflow.org/tutorials/structured_data/time_series
  - Scikit-learn Time Series: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html