

NGŨ VỊ HƯƠNG

Chủ đề: Phân tích những yếu tố ảnh hưởng tới thị hiếu khán giả và dự đoán rating phim

THÀNH VIÊN NHÓM

- Trần Mạnh Hùng - 22120117
- Nguyễn Thị Anh Thi - 22120339
- Dương Ngọc Kiều Trinh - 22120389
- Nguyễn Đoàn Minh Uyên - 22120421
- Nguyễn Phạm Tú Uyên - 22120422

NỘI DUNG

1. TỔNG QUAN ĐỒ ÁN
2. NỘI DUNG ĐỒ ÁN

TỔNG QUAN ĐỒ ÁN

GIỚI THIỆU VỀ ĐỀ TÀI LÝ DO CHỌN ĐỀ TÀI

- Ngành điện ảnh phát triển mạnh mẽ về số lượng và thể loại phim
- Sự thay đổi nhanh chóng trong thị hiếu khán giả

Ý NGHĨA CỦA ĐỀ TÀI

- Cung cấp cái nhìn toàn diện về thị hiếu và hành vi khán giả.
- Giúp các nhà làm phim, phát hành và đầu tư tối ưu hóa chiến lược kinh doanh.

MỤC TIÊU ĐỒ ÁN

- Phân tích thị hiếu xem phim của khán giả: Xác định các yếu tố ảnh hưởng đến sở thích phim trong 2 năm qua
- Dự đoán rating phim: Sử dụng dữ liệu đặc điểm phim và tương tác khán giả để dự đoán mức độ yêu thích.

NỘI DUNG ĐỒ ÁN

THU THẬP DỮ LIỆU

- JustWatch: Cung cấp thông tin về dịch vụ phát trực tuyến, danh sách phim, chương trình truyền hình, giá cả và nền tảng phát hành.
- OMDB: Cung cấp API truy xuất thông tin chi tiết về phim, chương trình truyền hình, đánh giá, xếp hạng, doanh thu và giải thưởng.

KHÁM PHÁ TIỀN XỬ LÝ DỮ LIỆU

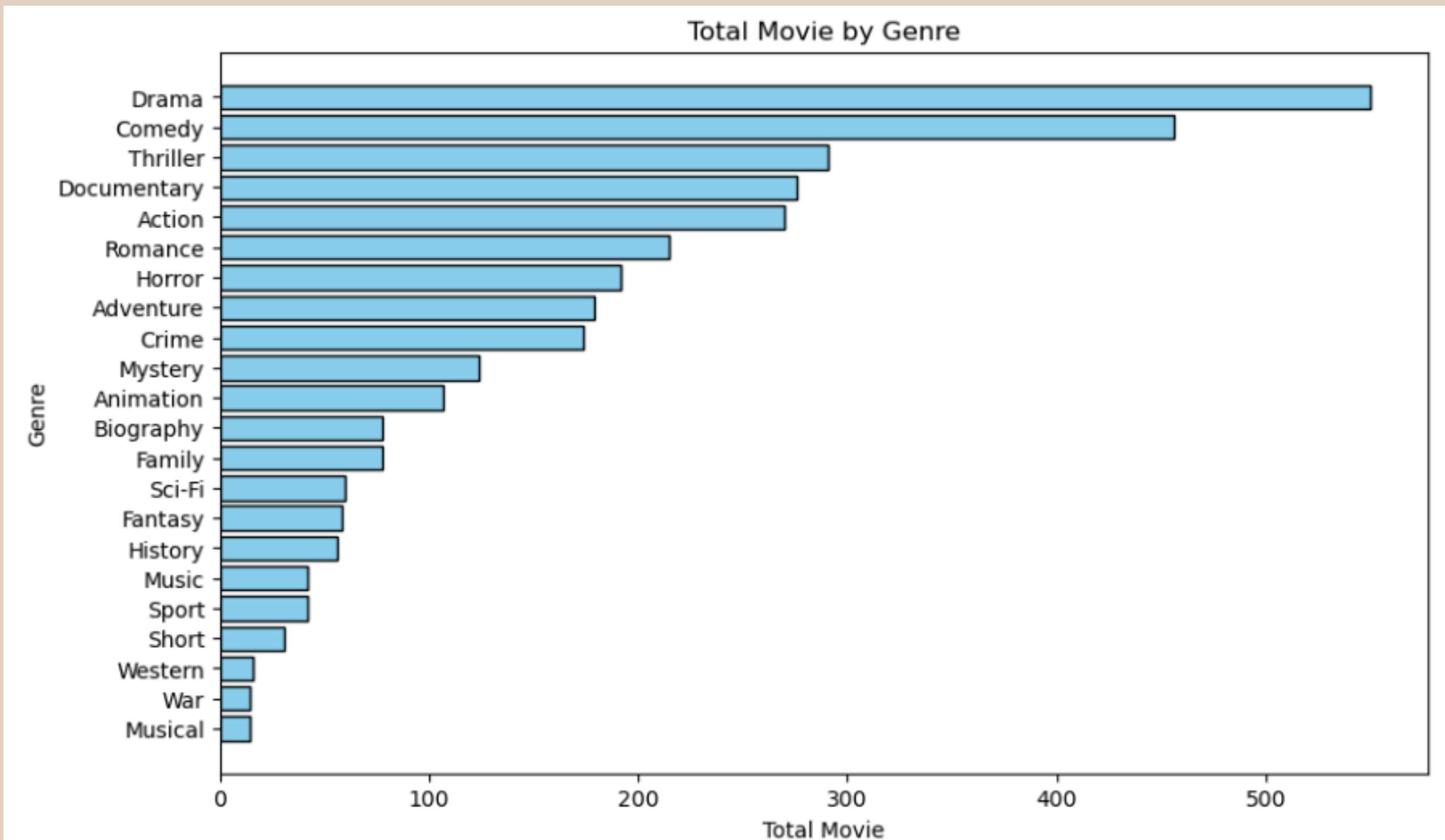
- Tổng hợp dữ liệu:** Hợp nhất các nguồn dữ liệu, kiểm tra số lượng dòng, cột và kiểu dữ liệu.
- Phân tích dữ liệu:** Xác định ý nghĩa từng cột, phân bố dữ liệu qua biểu đồ.
- Xử lý dữ liệu:**
 - Loại bỏ giá trị trùng lặp và dòng dữ liệu không cần thiết.
 - Điền giá trị thiếu (trung bình, trung vị, giá trị phổ biến) hoặc xóa cột có tỉ lệ thiếu cao.
 - Không xử lý outliers trong giai đoạn này.

KẾT QUẢ ĐẠT ĐƯỢC

- Tạo bộ dữ liệu nhất quán và đã được làm sạch, sẵn sàng cho các tác vụ tiếp theo.

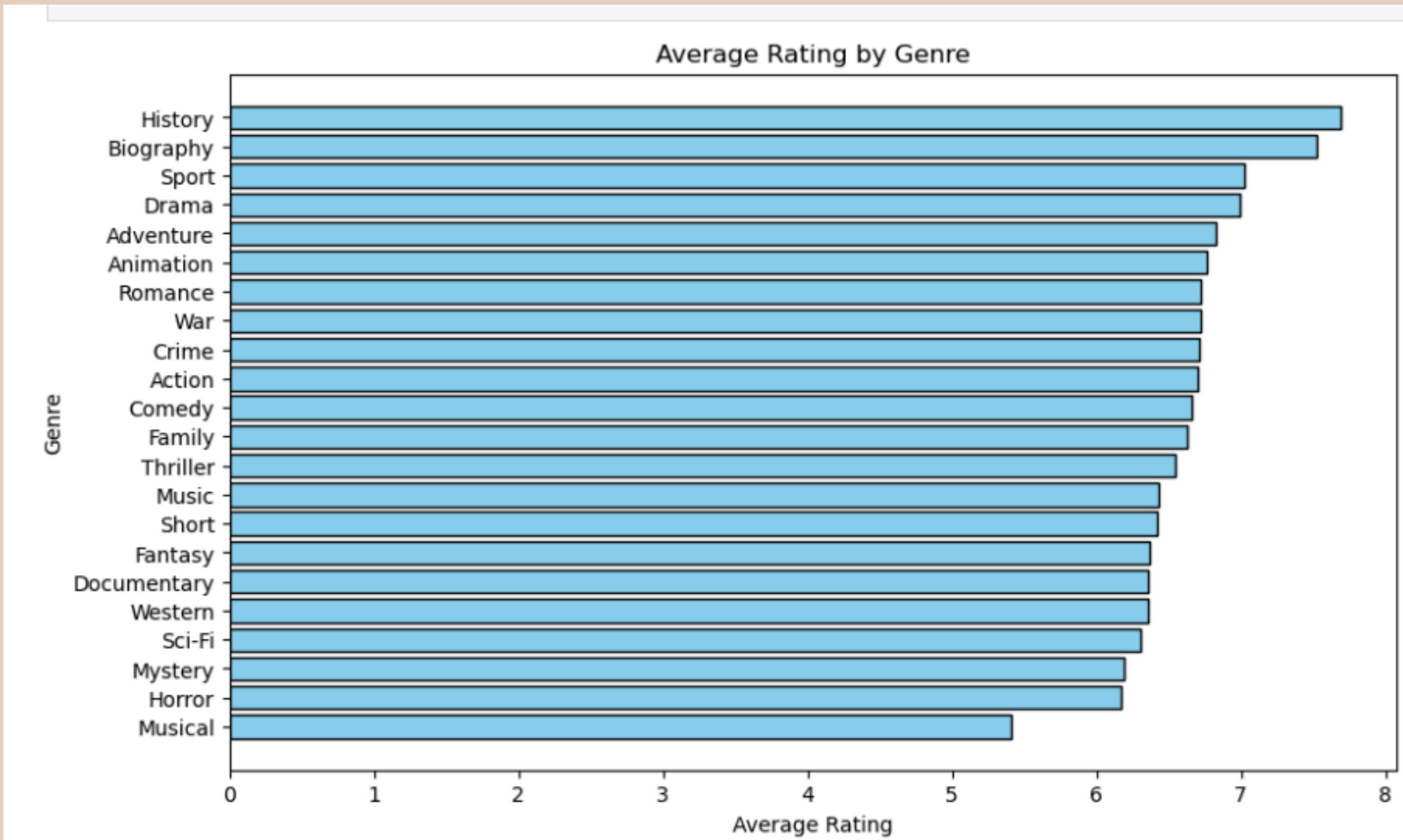
ĐẶT CÂU HỎI CÓ Ý NGHĨA

Câu hỏi 1: Với từng thể loại khác nhau thì đánh giá của khán giả khác nhau như thế nào ? Liệu số lượng bình chọn có ảnh hưởng tới đánh giá của từng thể loại hay không?



ĐẶT CÂU HỎI CÓ Ý NGHĨA

Câu hỏi 1: Với từng thể loại khác nhau thì đánh giá của khán giả khác nhau như thế nào ? Liệu số lượng bình chọn có ảnh hưởng tới đánh giá của từng thể loại hay không?



ĐẶT CÂU HỎI CÓ Ý NGHĨA

Câu hỏi 1: Với từng thể loại khác nhau thì đánh giá của khán giả khác nhau như thế nào? Liệu số lượng bình chọn có ảnh hưởng tới đánh giá của từng thể loại hay không?

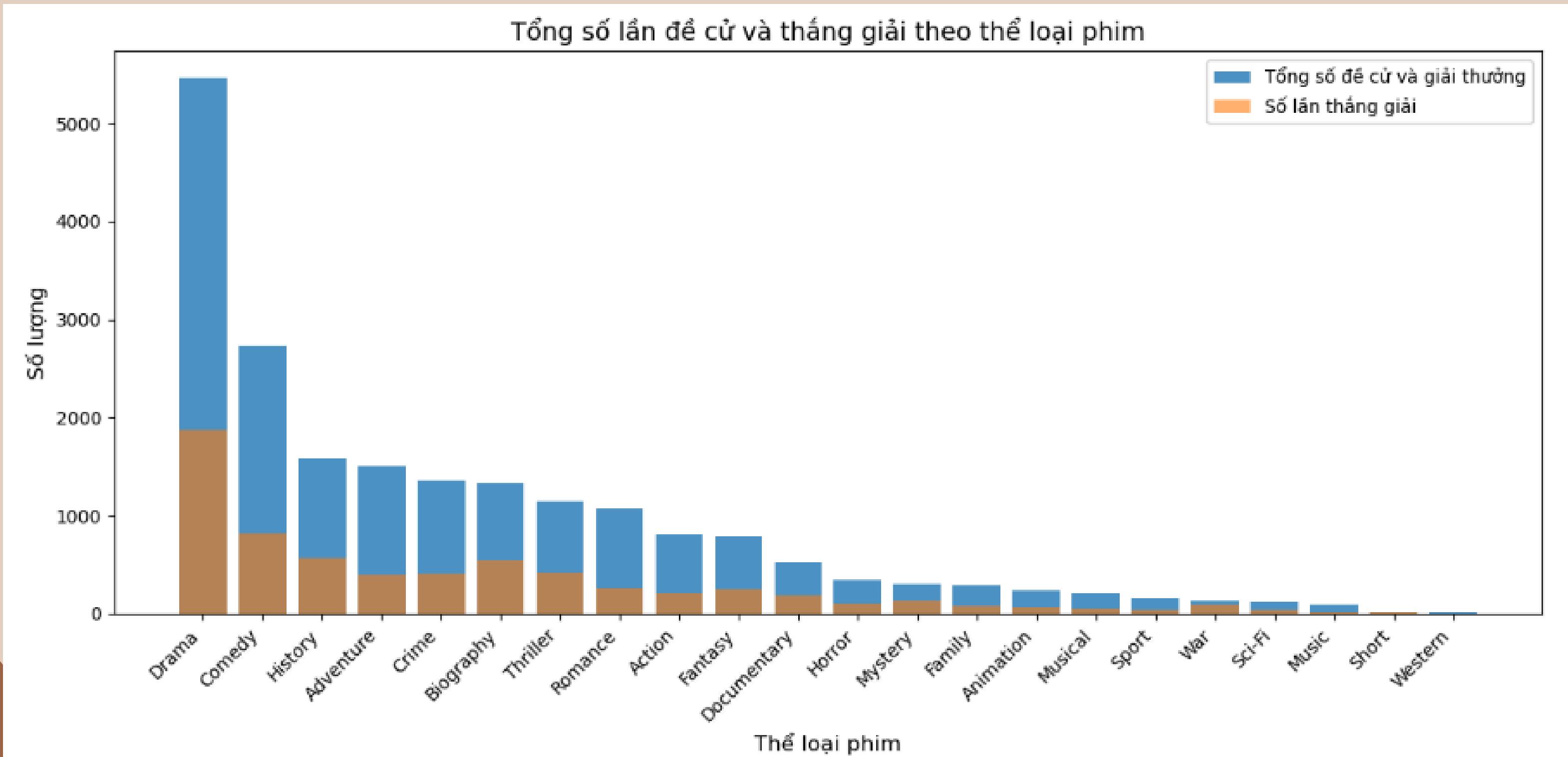
LỢI ÍCH

- Giúp hiểu sở thích và khẩu vị của khán giả với từng thể loại phim

KẾT LUẬN

- Số lượng không tỷ lệ thuận với chất lượng
- Thể loại ít nhưng chất lượng cao: History, Biography và Sport
- Thể loại phổ biến nhưng chất lượng trung bình: Drama, Comedy và Documentary
- Thể loại kém hấp dẫn: Sci-Fi, Horror và Musical

Câu hỏi 2: Các phim thuộc thể loại nào có nhiều khả năng được đề cử hoặc giành giải thưởng nhất?



Câu hỏi 2: Các phim thuộc thể loại nào có nhiều khả năng được đề cử hoặc giành giải thưởng nhất?

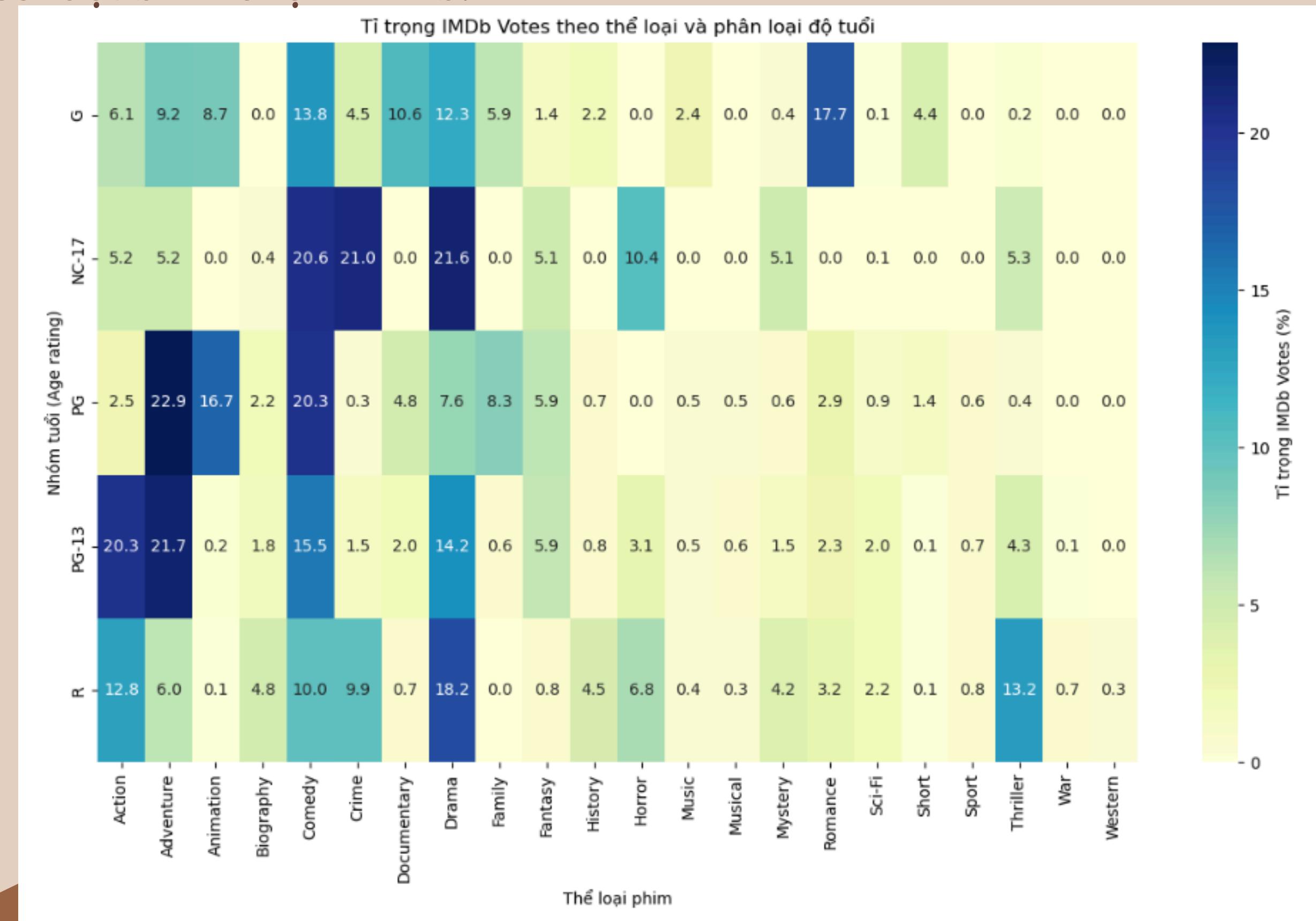
LỢI ÍCH

- Giúp nhà làm phim và nhà đầu tư tối ưu hóa chiến lược phát triển
- Hỗ trợ lựa chọn thể loại phim có khả năng được đề cử
- Hiểu rõ xu hướng thị hiếu và tiêu chí đánh giá từ các hội đồng giải thưởng

KẾT LUẬN

- Thể loại Drama dẫn đầu về số đề cử và giải thưởng
- Các thể loại Comedy, History và Adventure có tiềm năng lớn

Câu hỏi 3: Nhóm tuổi (Age rating) ảnh hưởng thế nào đến sự phổ biến của các thể loại phim qua số lượt bình chọn IMDb?



Câu hỏi 3: Nhóm tuổi (Age rating) ảnh hưởng thế nào đến sự phổ biến của các thể loại phim qua số lượt bình chọn IMDb?

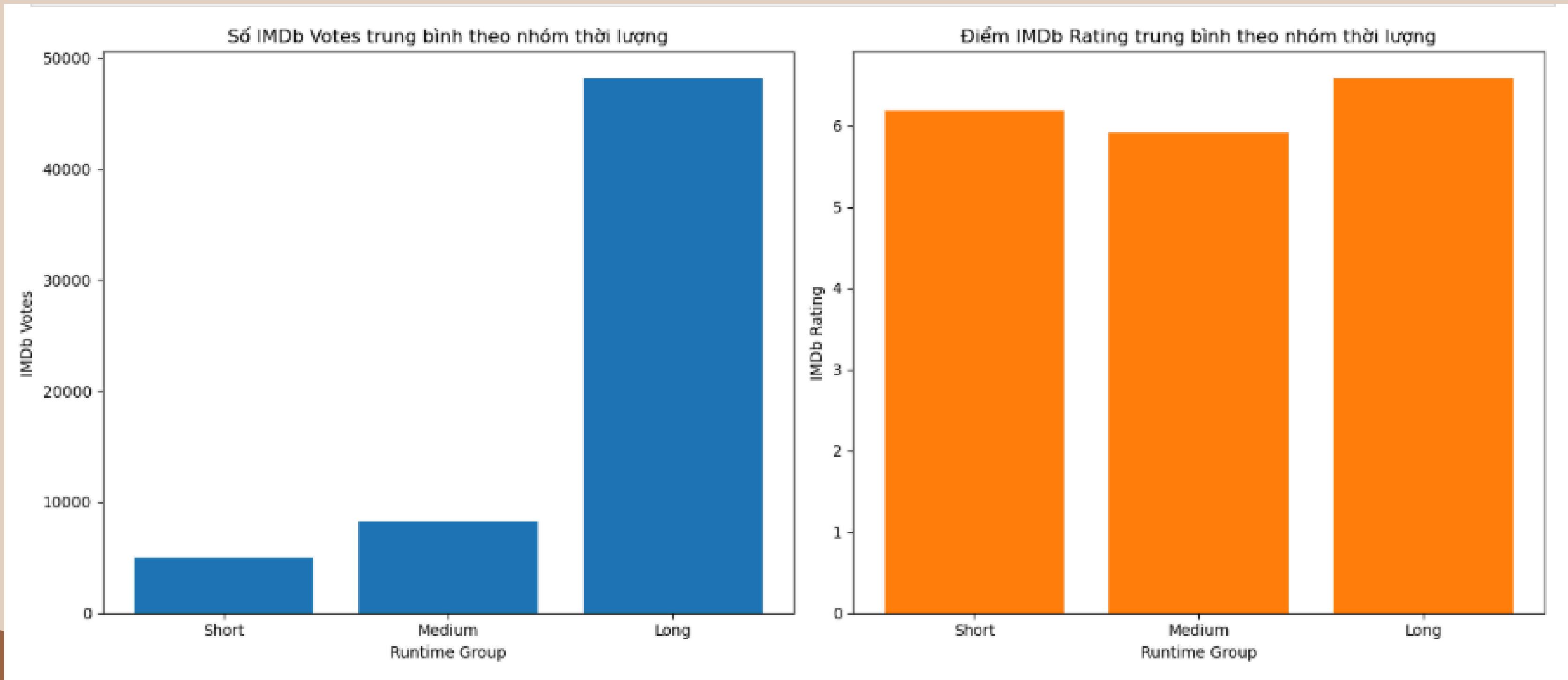
LỢI ÍCH

- Giúp hiểu rõ thị hiếu của từng nhóm tuổi.
- Hỗ trợ nhà sản xuất trong việc phát triển các thể loại phim phù hợp với từng độ tuổi, tối ưu hóa lợi nhuận.
- Giúp các nhà phát hành cập nhật các thể loại phim thu hút đối tượng.
- Cải thiện trải nghiệm người dùng trên nền tảng xem phim trực tuyến.

KẾT LUẬN

- Nhóm tuổi PG-13 và R là nhóm khán giả đóng góp lớn vào sự phổ biến của các thể loại Action, Drama và Thriller
- Nhóm tuổi nhỏ quan tâm đến Animation và Family

Câu hỏi 4: Thời lượng phim (Runtime) ảnh hưởng như thế nào đến số lượng IMDb Votes và IMDb Rating?



Câu hỏi 4: Thời lượng phim (Runtime) ảnh hưởng như thế nào đến số lượng IMDb Votes và IMDb Rating?

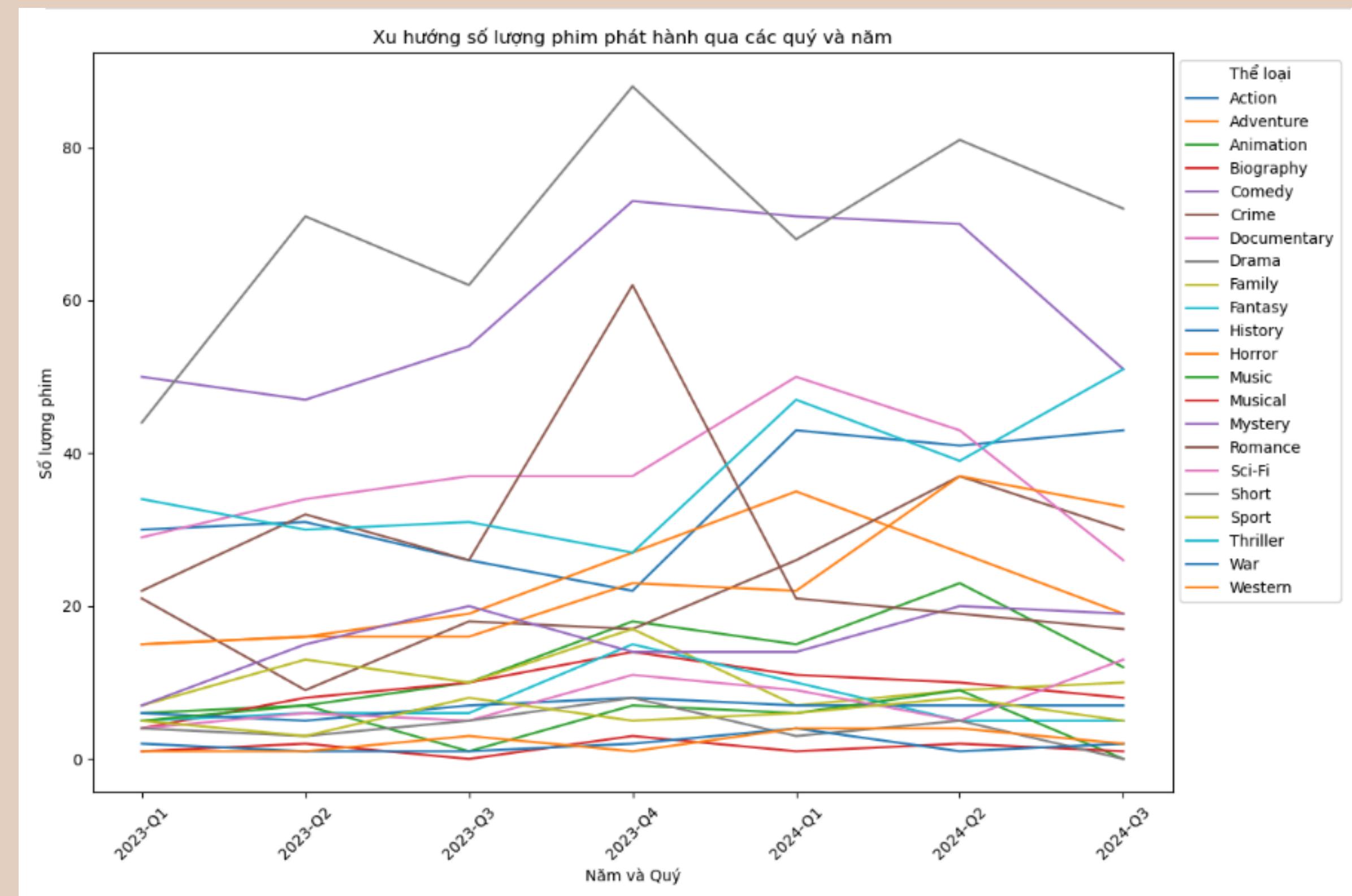
LỢI ÍCH

- Xác định thời lượng lý tưởng của một bộ phim để thu hút khán giả và đạt được điểm đánh giá cao.

KẾT LUẬN

- Phim chất lượng cao: Trên 120 phút, phù hợp với các tác phẩm chuyên sâu, mang lại trải nghiệm đa chiều.
- Phim thương mại phổ thông: 80-120 phút, cân bằng giữa nội dung hấp dẫn và độ dài vừa phải để thu hút khán giả.

Câu hỏi 5: Thể loại nào có xu hướng phát triển mạnh nhất qua các quý của năm dựa trên số lượng phim phát hành?



Câu hỏi 5: Thể loại nào có xu hướng phát triển mạnh nhất qua các quý của năm dựa trên số lượng phim phát hành?

LỢI ÍCH

- Dự đoán xu hướng phát triển về thể loại của ngành công nghiệp điện ảnh để định hướng đầu tư

KẾT LUẬN

- Thể loại ít phổ biến: Short và Western có thị trường hẹp và ít tiềm năng phát triển.
- Thể loại thống trị thị trường: Drama, Documentary và Comedy có sự tăng trưởng ổn định, phù hợp với thị hiếu khán giả

XÂY DỰNG MÔ HÌNH

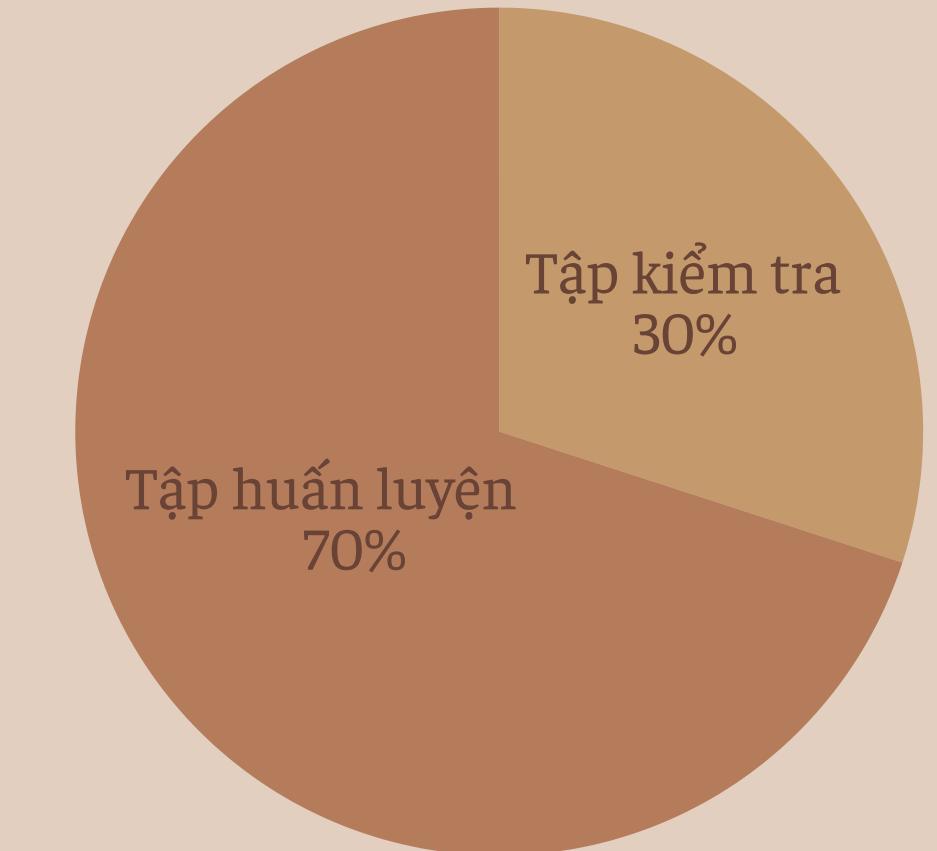
Vấn đề 1: Dự đoán phim có thành công hay không (Ratings > 7.0), phim thành công thường có những yếu tố nào (thời lượng, độ tuổi giới hạn, lượt bình chọn, thể loại).

Vấn đề 2: Dự đoán Ratings của phim dựa trên các thông tin được cho trước, bao gồm thời lượng, giới hạn độ tuổi, thể loại, số lượt đề cử, số lần thắng giải và số lượt bình chọn.

Vấn đề 1: Dự đoán phim có thành công hay không (Ratings > 7.0), phim thành công thường có những yếu tố nào (thời lượng, độ tuổi giới hạn, lượt bình chọn, thể loại).

Tập dữ liệu

Xử lý mất cân bằng dữ liệu bằng SMOTE
(Synthetic Minority Oversampling
Technique)

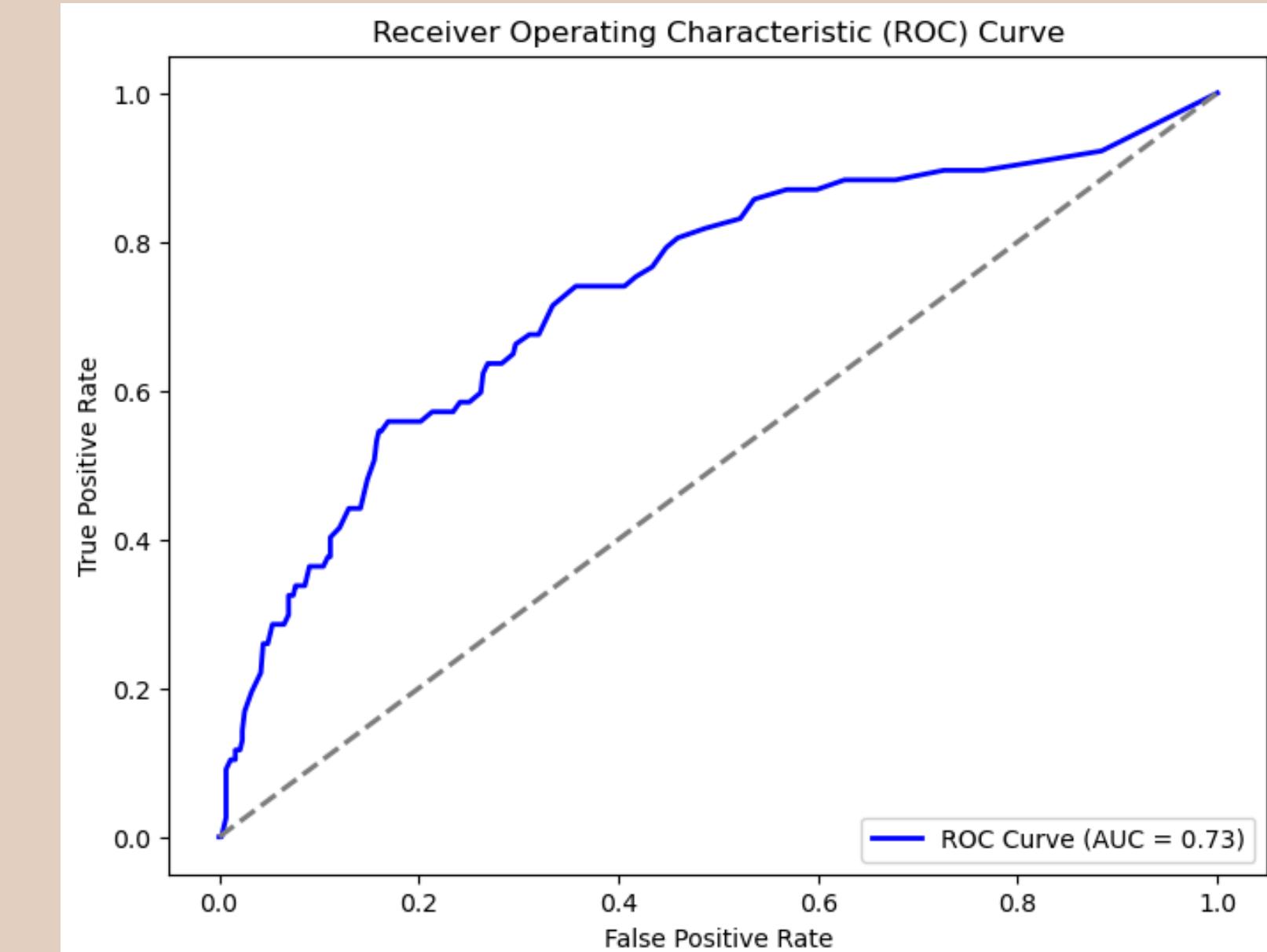
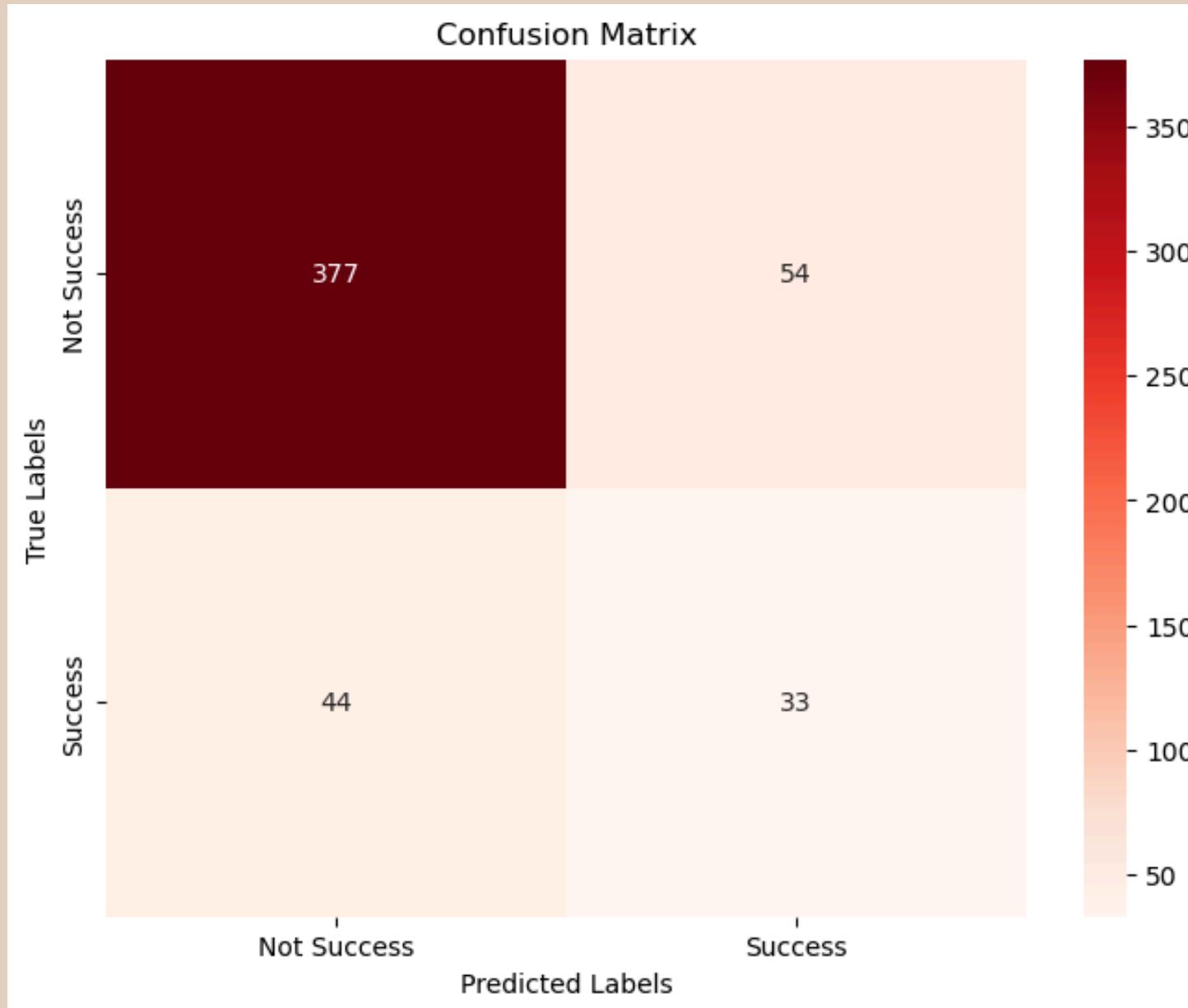


MÔ HÌNH RANDOM FOREST CLASS

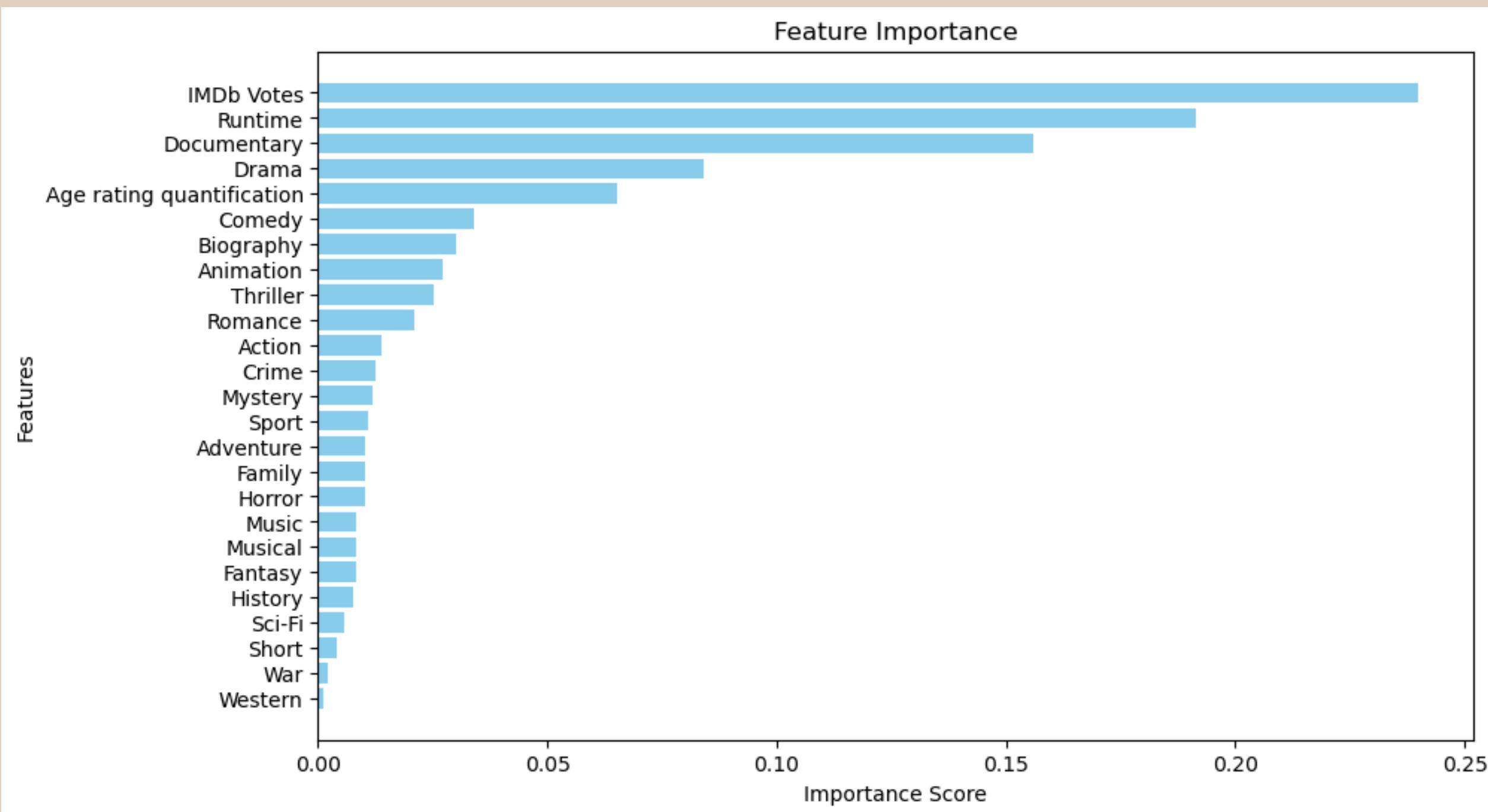
MỤC TIÊU MÔ HÌNH

- Mục tiêu là xây dựng mô hình machine learning để dự đoán phim có thành công hay không dựa trên các đặc trưng như IMDb Rating, số lượt bình chọn, thể loại phim và các thông tin liên quan.
- Hiểu rõ các yếu tố ảnh hưởng đến sự thành công của phim.
- Định hướng và cải thiện các dự án phim trong tương lai.

MÔ HÌNH RANDOM FOREST CLASS



MÔ HÌNH RANDOM FOREST CLASS



MÔ HÌNH RANDOM FOREST CLASS

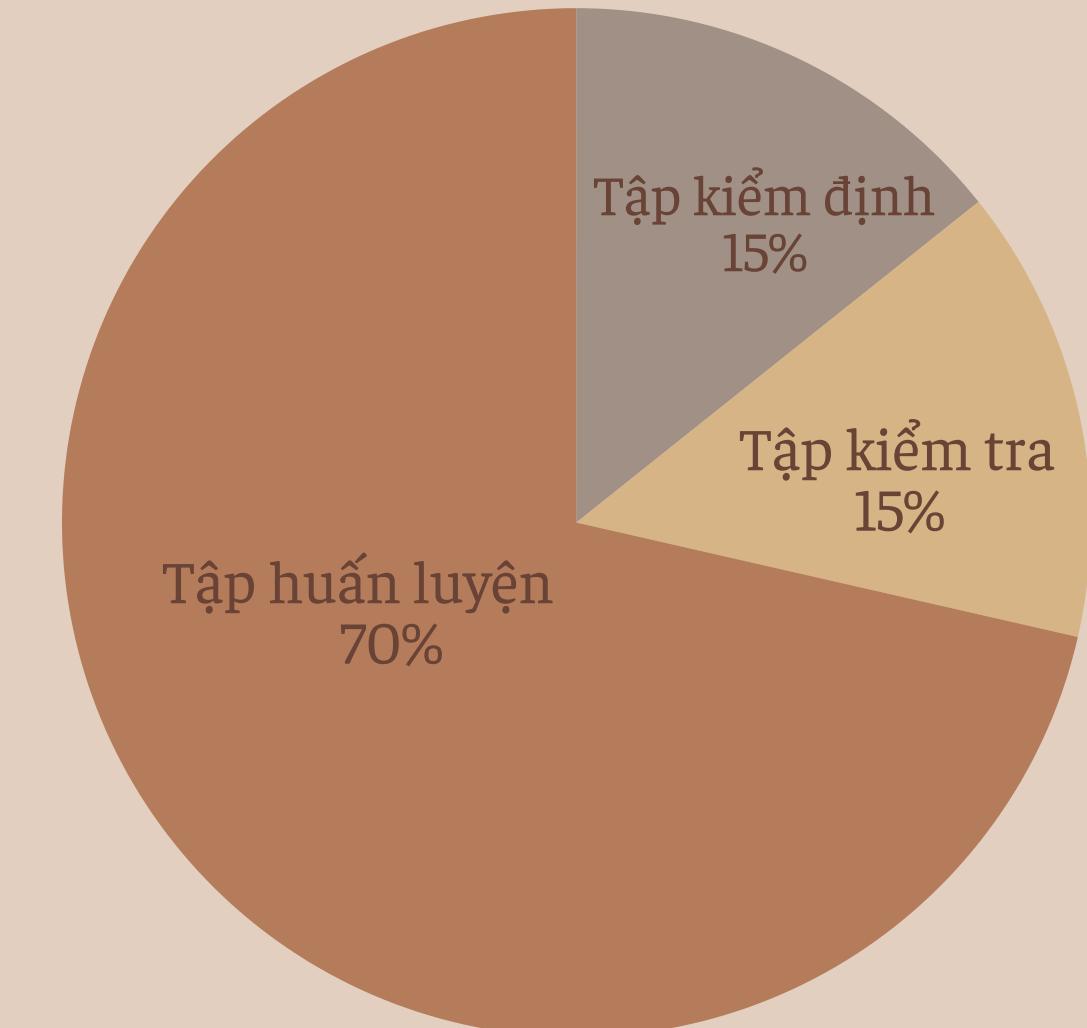
KẾT QUẢ

- Mô hình đã dự đoán thành công phim dựa trên các đặc trưng như IMDb Votes và thể loại phim.
- Xác định được các yếu tố quan trọng ảnh hưởng đến sự thành công của bộ phim.

Vấn đề 2: Dự đoán Ratings của phim dựa trên các thông tin được cho trước, bao gồm thời lượng, giới hạn độ tuổi, thể loại, số lượt đề cử, số lần thắng giải và số lượt bình chọn.

Loại bỏ outliers cho đặc trưng IMDb Rating

Tập dữ liệu



MÔ HÌNH LINEAR REGRESSION

LÝ DO CHỌN MÔ HÌNH

- Đơn giản, dễ triển khai
- Cung cấp mức hiệu suất tốt thiểu để đánh giá các mô hình phức tạp hơn

Ý NGHĨA MÔ HÌNH

Sử dụng làm mô hình cơ sở để đánh giá các mô hình phức tạp hơn

GRADIENT BOOSTING VÀ RANDOM FOREST REGRESSOR

LÝ DO CHỌN MÔ HÌNH

- Phù hợp với nhiều kiểu dữ liệu
- Xử lý tốt dữ liệu phức tạp
- Không yêu cầu chuẩn hóa đặc trưng
- Giảm overfitting bằng cách trung bình hóa nhiều cây
- Hoạt động hiệu quả trên dữ liệu nhiễu

Ý NGHĨA MÔ HÌNH

Dự đoán rating phim dựa trên các thông tin phim có sẵn, đánh giá khả năng sử dụng mô hình trong việc điền giá trị thiếu

ĐÁNH GIÁ CÁC MÔ HÌNH

Model	MAE	RMSE
Random Forest	0.516311	0.686432
Gradient Boosting	0.519526	0.684072
Linear Regression	0.562162	0.719547

MAE for test set: 0.49962969856522077
RMSE for test set: 0.655902824386394

- Mô hình Linear Regression có hiệu suất tệ nhất, trong khi đó mô hình Random Forest Regressor có hiệu suất tốt nhất dựa trên so sánh MAE và RMSE.
- Khi áp dụng mô hình Random Forest trên tập test, hiệu suất tương đương giữa tập valid và tập test cho thấy mô hình tương đối đáng tin cậy, có khả năng tổng quát hóa tốt.

KẾT QUẢ ĐẠT ĐƯỢC

- Mô hình đã dự đoán rating phim với sai số khoảng 5-7%. Đây là sai số tạm chấp nhận được với dữ liệu hiện tại.
- Đánh giá mô hình có thể được cân nhắc sử dụng vào bài toán điền dữ liệu rating thiếu nếu các feature input được xử lý tốt và có nhiều dữ liệu train hơn.

REFLECTION

1

Khó tìm nguồn dữ liệu chính thống được cào hợp pháp

BỎ QUA NHỮNG TRANG WEB LẬU, NHỮNG TRANG KHÔNG CHO PHÉP CÀO DỮ LIỆU HOẶC CÓ NHIỀU YÊU CẦU PHÁP LÝ.

2

Dữ liệu trong 1 trang không đủ feature cần thiết

KẾT HỢP DỮ LIỆU CỦA 2 TRANG ĐÃ CHỌN ĐỂ THU THẬP (TẠM) ĐỦ CÁC FEATURE NHÓM MONG MUỐN CÓ.

3

Key API có giới hạn theo ngày.

DỮ LIỆU ĐƯỢC CHIA THÀNH 3 PHẦN, MỖI THÀNH VIÊN DÙNG 1 API KEY -> TỐI ƯU HÓA THỜI GIAN VÀ TÀI NGUYÊN.

4

Không tải trang kịp dẫn tới các giá trị NULL không mong muốn

TĂNG TIME SLEEP VÀ CHỌN NƠI CÓ KẾT NỐI MẠNG TỐT HƠN

5

Mô hình chưa quá tốt do thiếu sample và feature

RÚT KINH NGHIỆM CHO CÁC LẦN TIẾP THEO

PHÂN CÔNG CÔNG VIỆC

Người thực hiện	Công việc
Hùng	<ul style="list-style-type: none">Xử lý dữ liệu kiểu định lượng, viết báo cáo tiền xử lý và đặt câu hỏi, trả lời câu hỏi 5, format code
Thi	<ul style="list-style-type: none">Cào dữ liệu trong JustWatch, trả lời câu hỏi 2,3, build mô hình và viết báo cáo mô hình 1
Trinh	<ul style="list-style-type: none">Cào link phim, trả lời câu hỏi 6, format code, viết báo cáo P2
Minh Uyên	<ul style="list-style-type: none">Cào dữ liệu trong OMDb, viết báo cáo cào và khám phá dữ liệu, trả lời câu hỏi 1, build mô hình và format code
Tú Uyên	<ul style="list-style-type: none">Kiểm tra kiểu dữ liệu, xử lý dữ liệu kiểu định tính, trả lời câu hỏi 4, format code, viết báo cáo các bước và insight câu hỏi, mô hình học máy 2

REFERENCES

- 1 **Thư viện Scikit-learn**
<HTTPS://SCIKIT-LEARN.ORG/1.5/API/INDEX.HTML>
- 2 **Thư viện Pandas**
<HTTPS://PANDAS.PYDATA.ORG/DOCS/REFERENCE/>
- 3 **Thư viện Matplotlib**
<HTTPS://MATPLOTLIB.ORG/STABLE/>
- 4 **Thư viện Seaborn**
<HTTPS://SEABORN.PYDATA.ORG/API.HTML>
- 5 **Công thức Bayesian average**
<HTTPS://WWW.ALGOLIA.COM/DOC/GUIDES/MANAGING-RESULTS/MUST-DO/CUSTOM-RANKING/HOW-TO/BAYESIAN-AVERAGE/>



**THANK
YOU**