

# TRỰC QUAN HÓA DỮ LIỆU

GIẢNG VIÊN HƯỚNG DẪN: TS. BÙI TIẾN LÊN

## PHIM CHIẾU RẠP Ở VIỆT NAM

Trình bày bởi: Nhóm 07



# THÀNH VIÊN NHÓM

21127104 - Đoàn Ngọc Mai

21127115 - Trần Thanh Ngân

21127129 - Lê Nguyễn Kiều Oanh

21127229 - Dương Trường Bình

21127616 - Lê Phước Quang Huy



# NỘI DUNG TRÌNH BÀY

1

Giới thiệu tập dữ liệu

4

Trực quan hóa dữ liệu

2

Khám phá và phân tích thống kê

5

Các tiêu chí đánh giá

3

Tương quan stat model

6

Đề xuất cải tiến

NỘI DUNG 1

# GIỚI THIỆU TẬP DỮ LIỆU

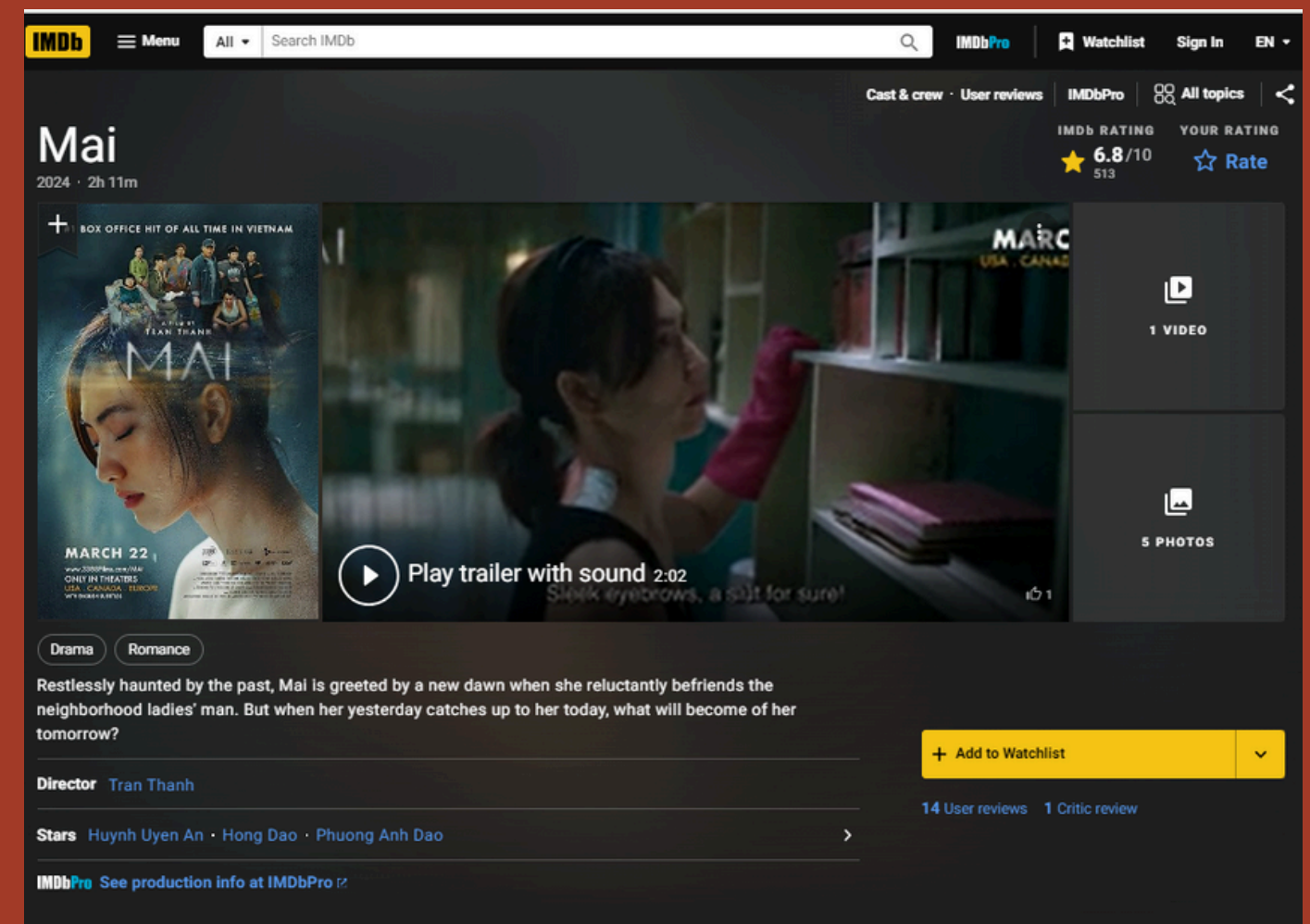
# 1.1 THÔNG TIN CƠ BẢN

Tên: Movies Shown In Vietnam

Nguồn gốc: Nhóm đã thu thập dữ liệu từ nhiều nguồn khác nhau

- Cào dữ liệu từ website moveek và imdb
- Sử dụng API của TMDB

Chủ đề: Thông tin chi tiết của các bộ phim đã từng chiếu rạp tại Việt Nam



# 1.2 MÔ TẢ CHI TIẾT

Mỗi dòng trong bộ dữ liệu mô tả các thông tin về các bộ phim (tên, doanh thu, ngân sách thể loại, năm phát hành ...) ngoài ra thông tin như đạo diễn, biên kịch, diễn viên chỉ có ở những phim điện ảnh Việt Nam.

STT	Tên cột	Ý nghĩa
1	Id	Mã định danh cho mỗi bộ phim
2	Title	Tên của bộ phim
3	Original Title	Tên gốc của phim (ngôn ngữ của quốc gia sản xuất phim đó)
4	Original Language	Ngôn ngữ gốc của bộ phim
5	Overview	Tóm tắt và mô tả nội dung của bộ phim
6	Revenue	Doanh thu của bộ phim
7	Budget	Ngân sách sản xuất của bộ phim
8	Runtime	Thời lượng của bộ phim (đơn vị: phút)
9	Release Date	Ngày phát hành của bộ phim
10	Vote Average	Điểm đánh giá trung bình của bộ phim
11	Vote Count	Số lượt đánh giá mà bộ phim nhận được
12	Genres	Thể loại của bộ phim, có thể bao gồm nhiều thể loại khác nhau
13	Production Companies	Các công ty sản xuất bộ phim
14	Production Countries	Các quốc gia tham gia sản xuất bộ phim
15	Spoken Languages	Các ngôn ngữ được sử dụng trong bộ phim
16	Director	Đạo diễn của bộ phim
17	Stars	Các diễn viên chính tham gia trong bộ phim

Bảng 1: Ý nghĩa của các cột trong bộ dữ liệu

## NỘI DUNG 2

# KHÁM PHÁ VÀ PHÂN TÍCH THỐNG KÊ

# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Cột	Số lượng giá trị thiếu	Tỉ lệ giá trị thiếu (%)
Stars	1737	79.2
Director	1730	78.9
Overview	110	5.01

Bảng 2: Các cột bị thiếu dữ liệu ở tất cả các bộ phim

Cột	Số lượng giá trị thiếu	Tỉ lệ giá trị thiếu (%)
Overview	108	23.5
Stars	12	2.6
Director	4	0.87

Bảng 3: Các cột bị thiếu dữ liệu ở các bộ phim Việt Nam



# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

- Revenue

Bảng 4: Thống kê và khoảng tin cậy 95%

Thống kê	Giá trị
Trung bình (mean)	1,637,583
Độ lệch chuẩn (std)	2,796,444
Giá trị nhỏ nhất (min)	975
Phân vị 25%	124,606
Phân vị 50% (median)	665,224
Phân vị 75%	2,136,495
Giá trị lớn nhất (max)	22,119,910
Khoảng tin cậy 95%	(1,275,393; 1,999,773)

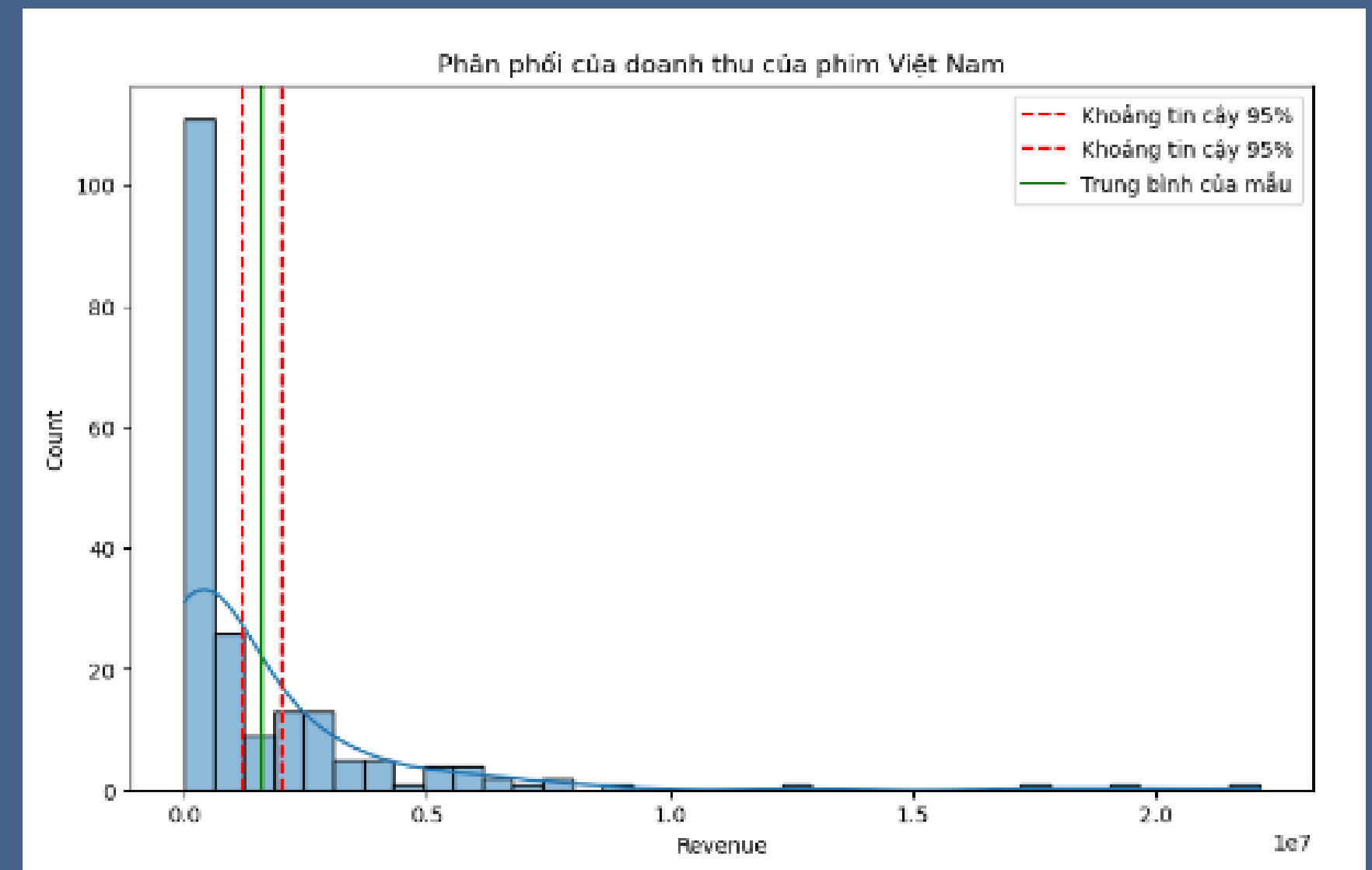
# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

- Revenue



# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

- Revenue
- Budget

Bảng 5: Thống kê và khoảng tin cậy 95%

Thống kê	Giá trị
Trung bình (mean)	906,335
Độ lệch chuẩn (std)	925,063
Giá trị nhỏ nhất (min)	8000
Phân vị 25%	300,000
Phân vị 50% (median)	669,159
Phân vị 75%	1,023,219
Giá trị lớn nhất (max)	6,000,000
Khoảng tin cậy 95%	(725,926; 1,086,744)

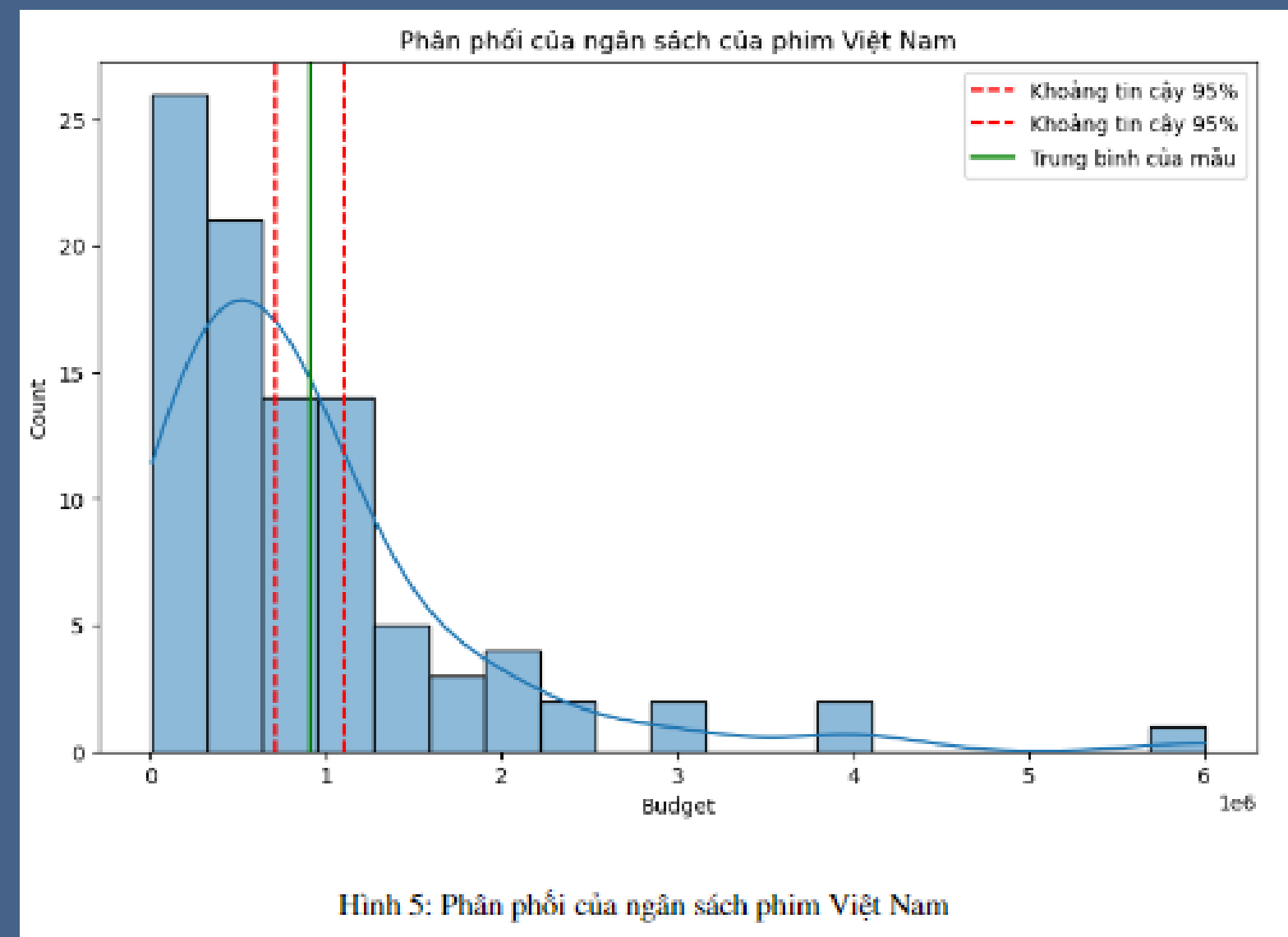
# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

- Revenue
- Budget



# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

- Revenue
- Budget
- Runtime

Bảng 6: Thống kê và khoảng tin cậy 95%

Thống kê	Giá trị
Trung bình (mean)	97.88
Độ lệch chuẩn (std)	16.04
Giá trị nhỏ nhất (min)	11
Phân vị 25%	90
Phân vị 50% (median)	98
Phân vị 75%	107
Giá trị lớn nhất (max)	179
Khoảng tin cậy 95%	(96.3; 99.4)

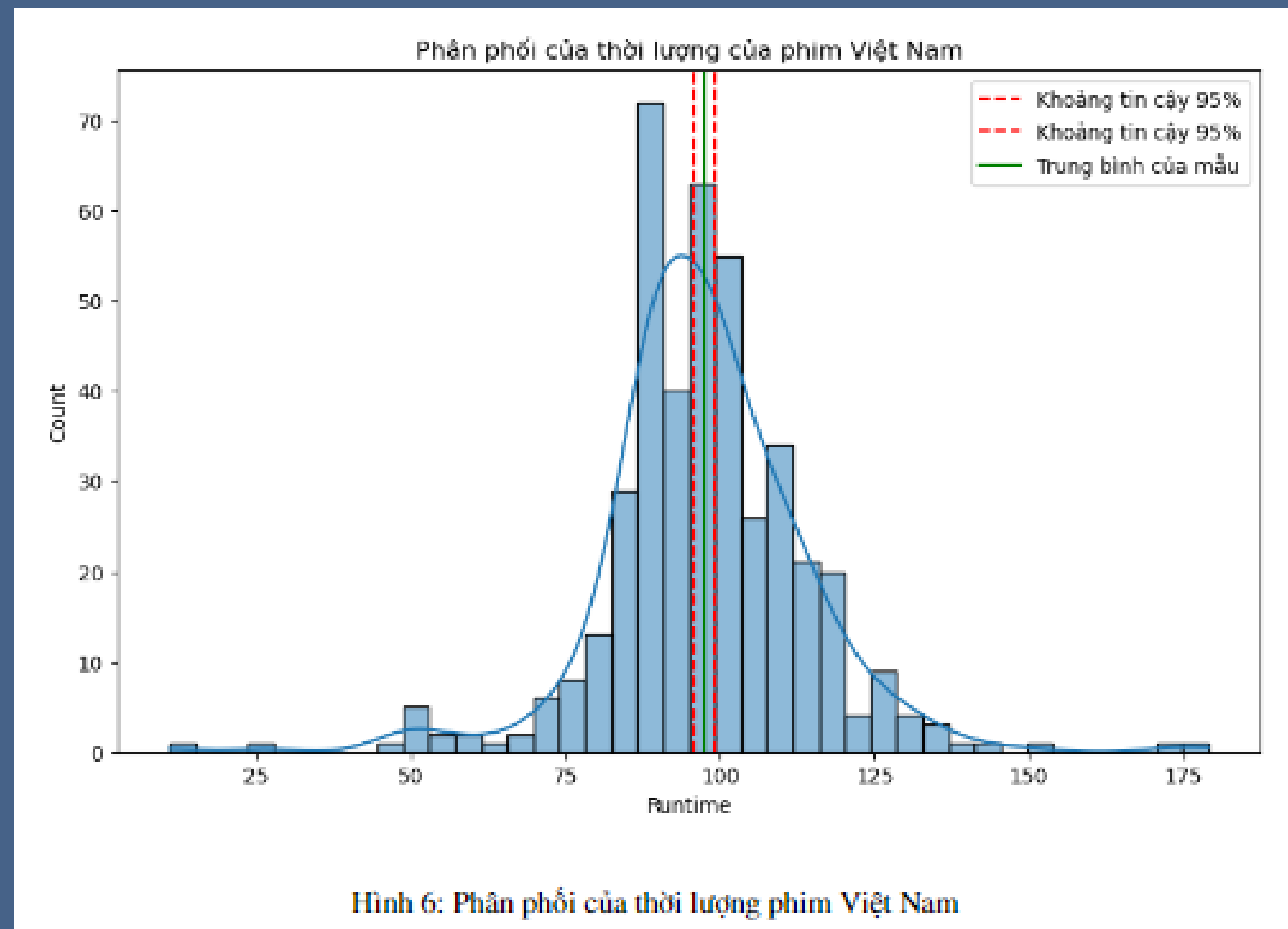
# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

- Revenue
- Budget
- Runtime



# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

- Revenue
- Budget
- Runtime
- Vote Average

Bảng 7: Thống kê và khoảng tin cậy 95%

Thống kê	Giá trị
Trung bình (mean)	6.15
Độ lệch chuẩn (std)	1.21
Giá trị nhỏ nhất (min)	1.7
Phân vị 25%	5.4
Phân vị 50% (median)	6.2
Phân vị 75%	7.00
Giá trị lớn nhất (max)	9.4
Khoảng tin cậy 95%	(6.03 ; 6.26)

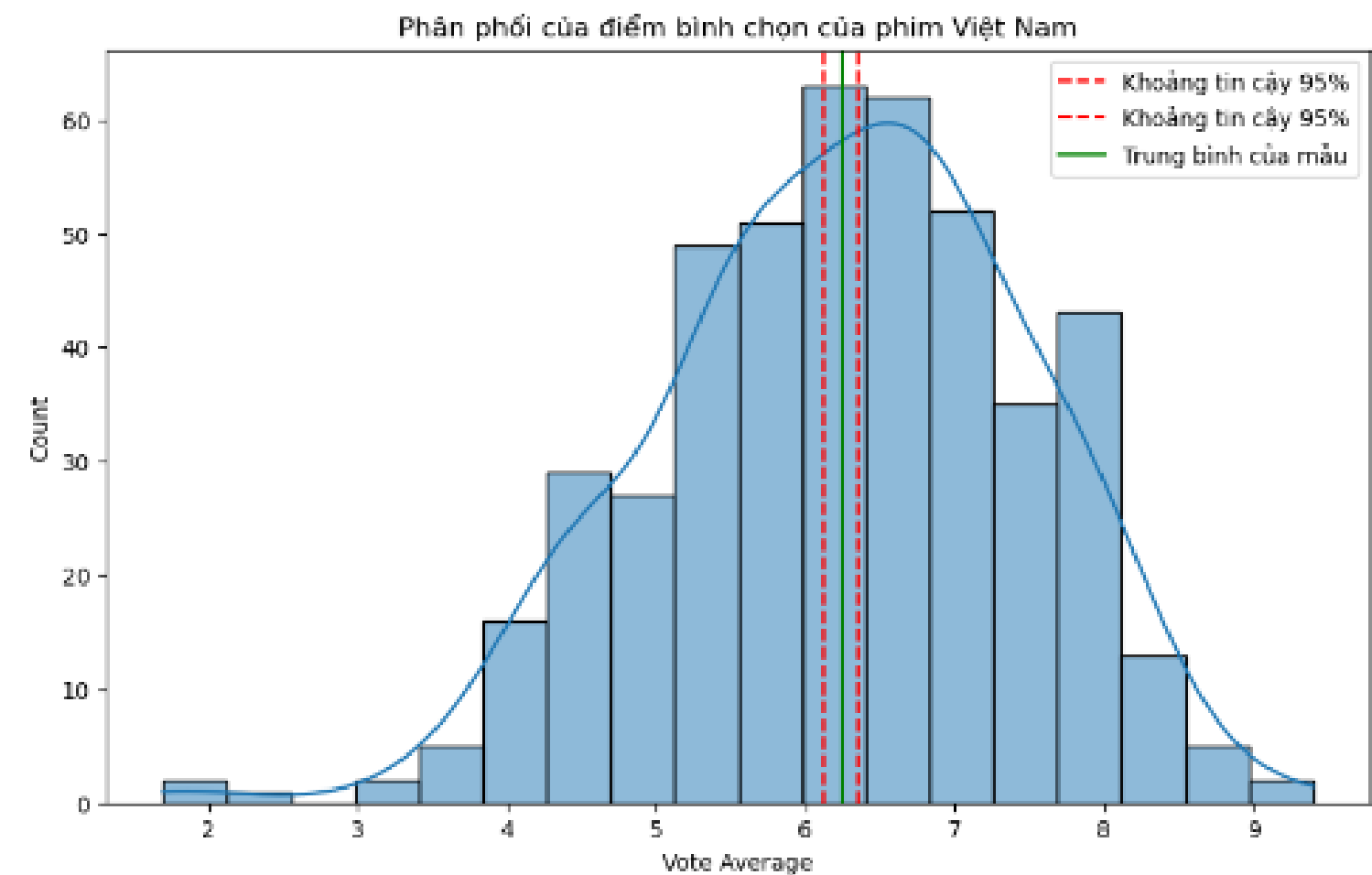
# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

- Revenue
- Budget
- Runtime
- Vote Average



Hình 7: Phân phối của điểm bình chọn phim Việt Nam



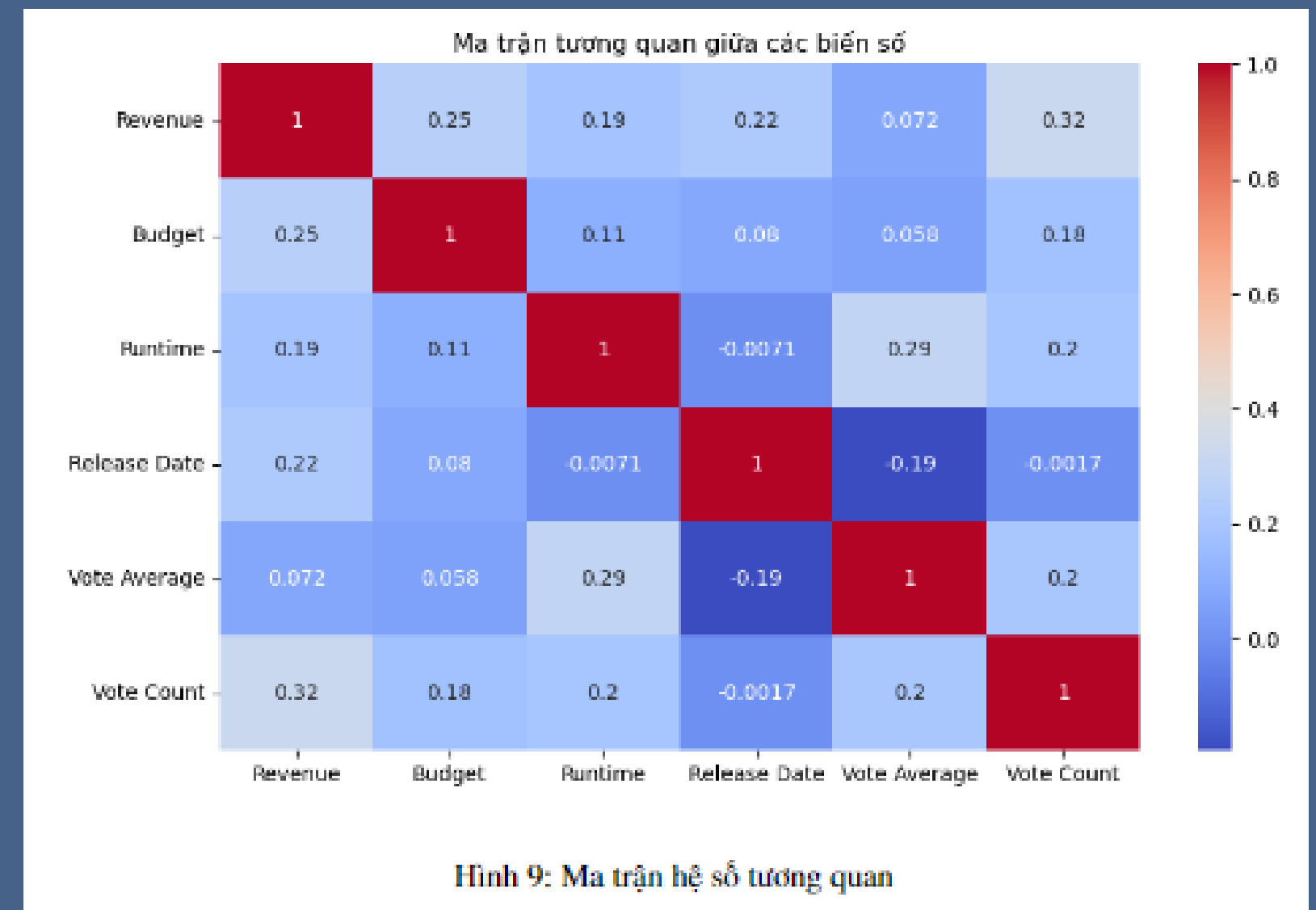
# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

Tương quan giữa các biến



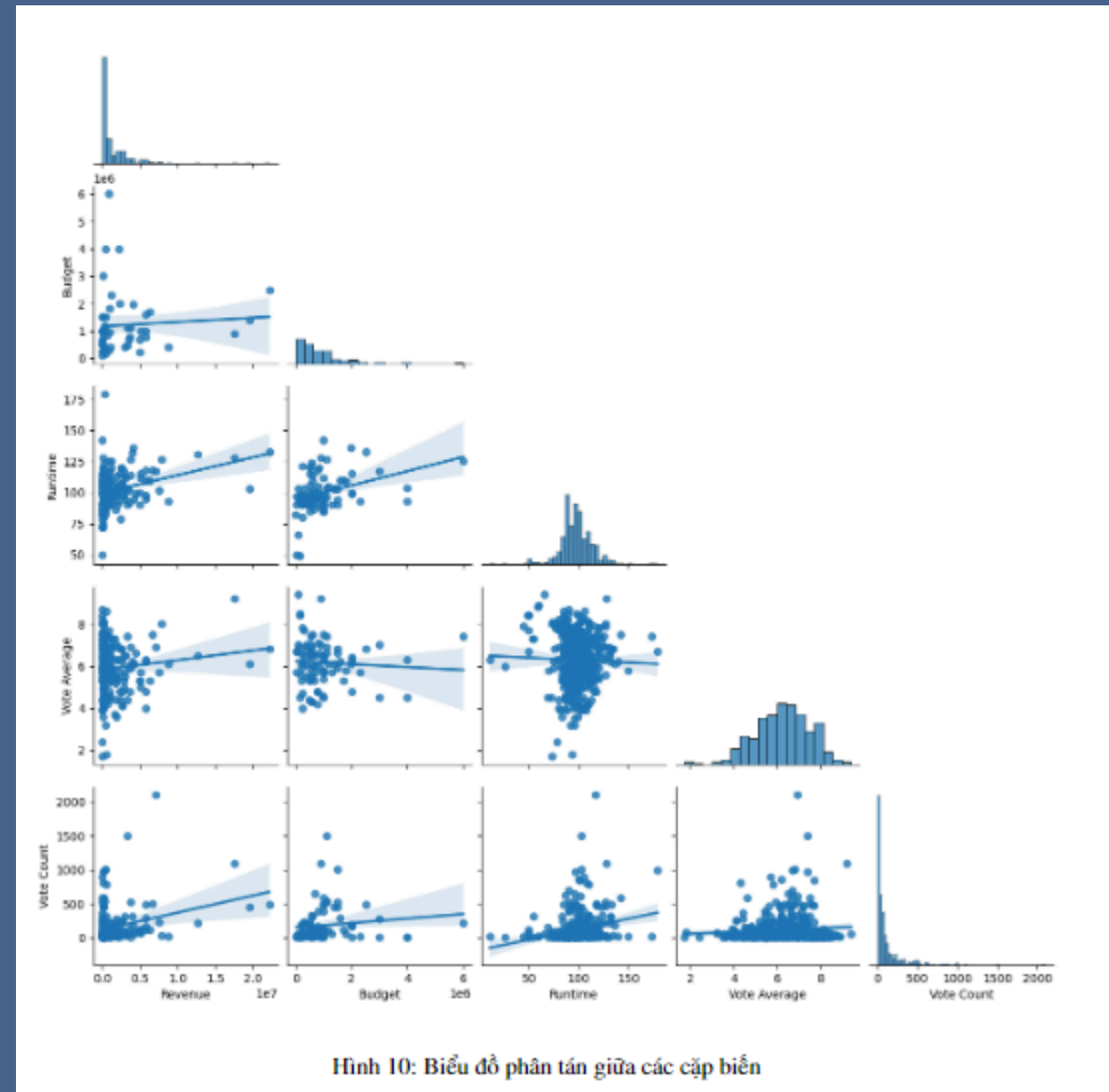
# KIỂM TRA DỮ LIỆU

Trùng lặp

Dữ liệu thiếu

Phân tích thống kê

Tương quan giữa các biến



**NỘI DUNG 3**

# **TƯƠNG QUAN STAT MODEL**

# 3.1 PHƯƠNG PHÁP

Sử dụng mô hình hồi quy tuyến tính OLS để ước lượng mối quan hệ.

$$\text{Biến Mục Tiêu} = \beta_0 + \beta_1 \times [\text{Biến Độc Lập}] + \epsilon \quad (1)$$

Trong đó,  $(\beta_0)$  là hệ số chặn,  $(\beta_1)$  là hệ số hồi quy cho [biến độc lập], và  $(\epsilon)$  là sai số ngẫu nhiên.

## 3.2 KẾT QUẢ

# Mô hình dự đoán Revenue

OLS Regression Results						
=====						
Dep. Variable:	Revenue	R-squared (uncentered):	0.773			
Model:	OLS	Adj. R-squared (uncentered):	0.773			
Method:	Least Squares	F-statistic:	1880.			
Date:	Wed, 08 May 2024	Prob (F-statistic):	0.00			
Time:	21:41:10	Log-Likelihood:	-44295.			
No. Observations:	2211	AIC:	8.860e+04			
Df Residuals:	2207	BIC:	8.862e+04			
Df Model:	4					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
Vote Average	-1.696e+05	1.56e+06	-0.109	0.913	-3.23e+06	2.89e+06
Vote Count	2.538e+04	860.065	29.507	0.000	2.37e+04	2.71e+04
Budget	2.0404	0.065	31.296	0.000	1.913	2.168
Runtime	-1.43e+05	9.72e+04	-1.470	0.142	-3.34e+05	4.77e+04
=====						
Omnibus:	1869.764	Durbin-Watson:	2.028			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	114483.019			
Skew:	3.603	Prob(JB):	0.00			
Kurtosis:	37.508	Cond. No.	3.83e+07			
=====						

## 3.2 KẾT QUẢ

# Mô hình dự đoán Vote Average

```

=====
OLS Regression Results
=====
Dep. Variable:          Vote Average    R-squared (uncentered):          0.826
Model:                  OLS             Adj. R-squared (uncentered):      0.825
Method:                 Least Squares   F-statistic:                     539.3
Date:                   Wed, 08 May 2024 Prob (F-statistic):               6.44e-171
Time:                   21:41:10        Log-Likelihood:                   -1071.9
No. Observations:      458             AIC:                             2152.
Df Residuals:          454             BIC:                             2168.
Df Model:               4
Covariance Type:        nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
Vote Count	0.0019	0.001	3.022	0.003	0.001	0.003
Revenue	-6.179e-08	6e-08	-1.031	0.303	-1.8e-07	5.6e-08
Budget	1.464e-07	2.14e-07	0.684	0.494	-2.74e-07	5.67e-07
Runtime	0.0569	0.002	36.627	0.000	0.054	0.060

```

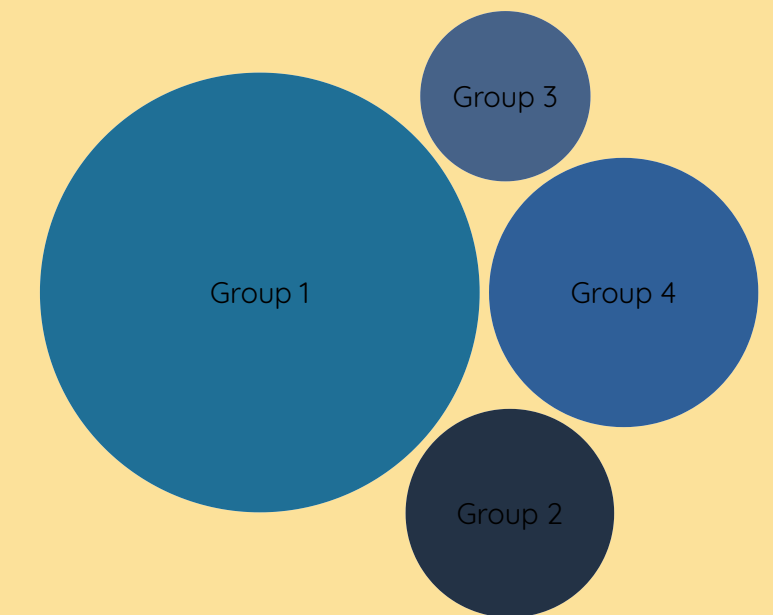
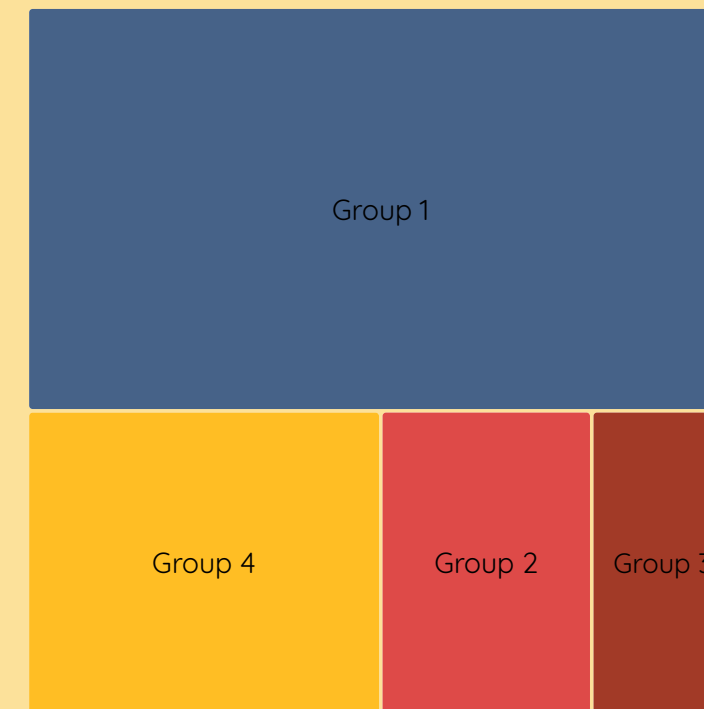
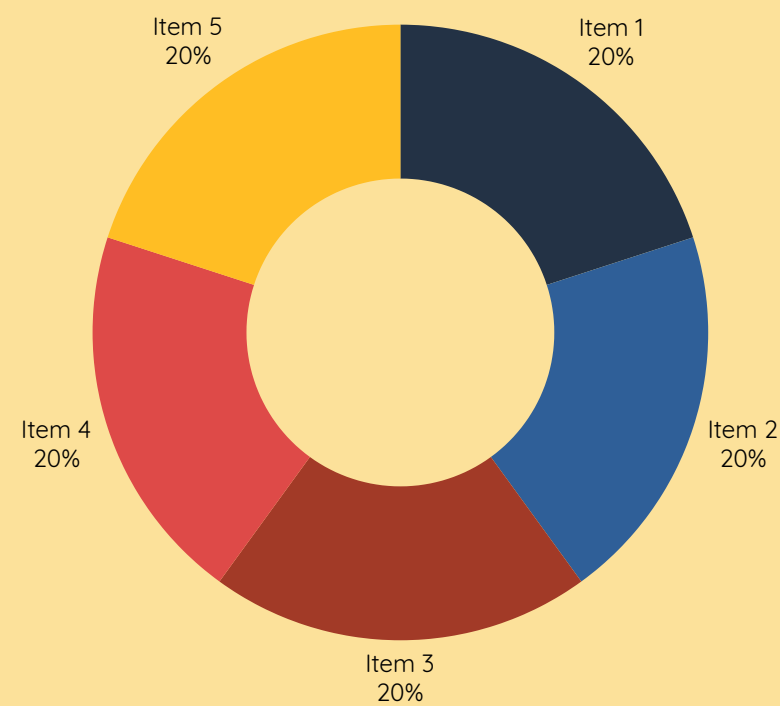
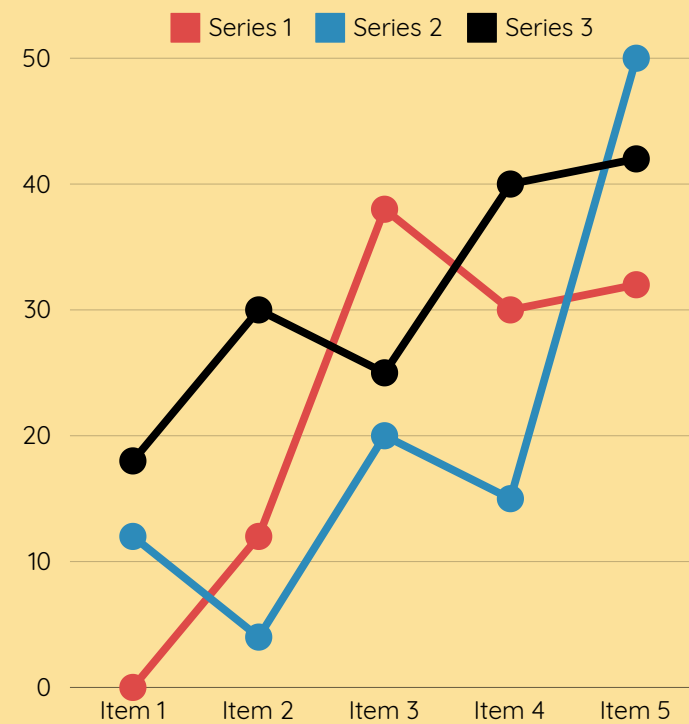
=====
Omnibus:                29.278    Durbin-Watson:                1.679
Prob(Omnibus):           0.000    Jarque-Bera (JB):            40.606
Skew:                    0.504    Prob(JB):                     1.52e-09
Kurtosis:                4.055    Cond. No.                     3.07e+04
=====

```

**NỘI DUNG 3**

# **TRỰC QUAN HÓA**

# 3.1 CHỌN BIỂU ĐỒ TRỰC QUAN



Biểu đồ đường

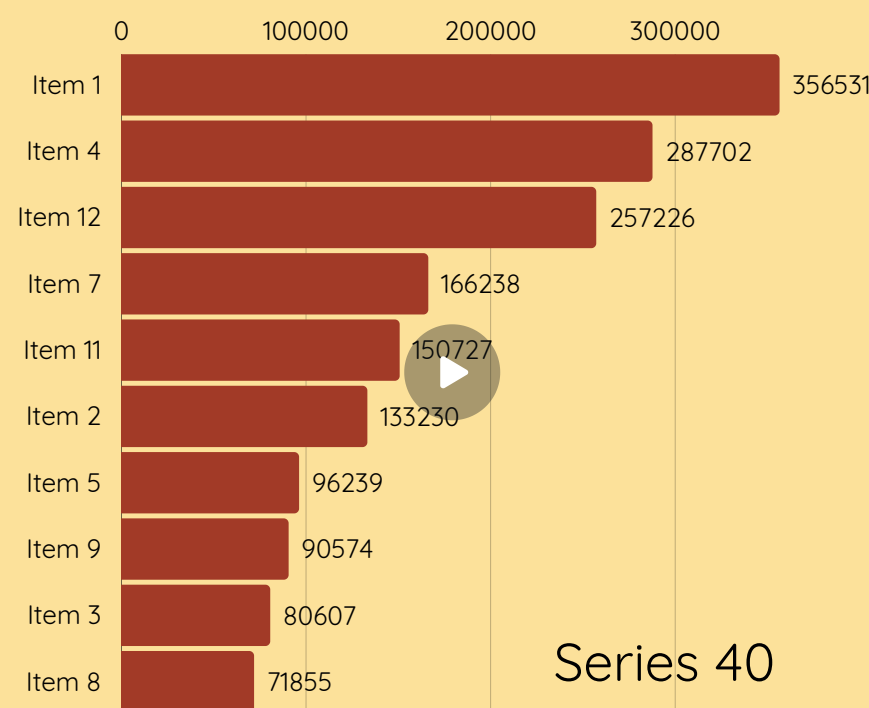
Biểu đồ donut

Biểu đồ treemaps

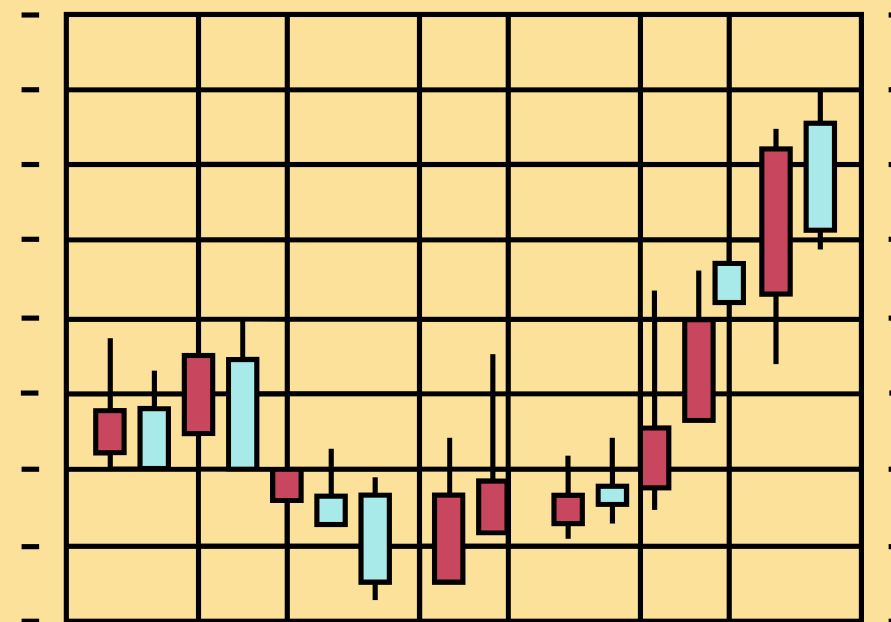
Biểu đồ bong bóng



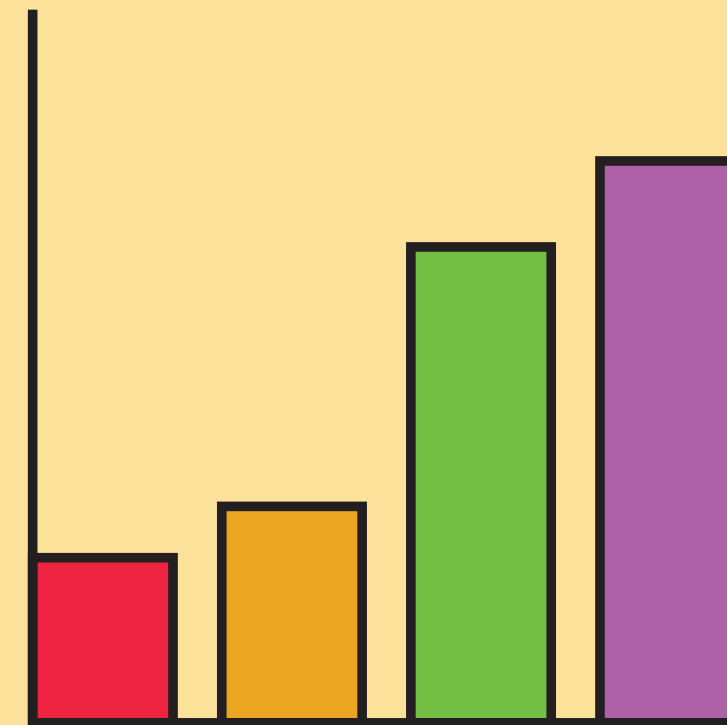
# 3.1 CHỌN BIỂU ĐỒ TRỰC QUAN



Biểu đồ ranking



Biểu đồ boxplot



Biểu đồ cột

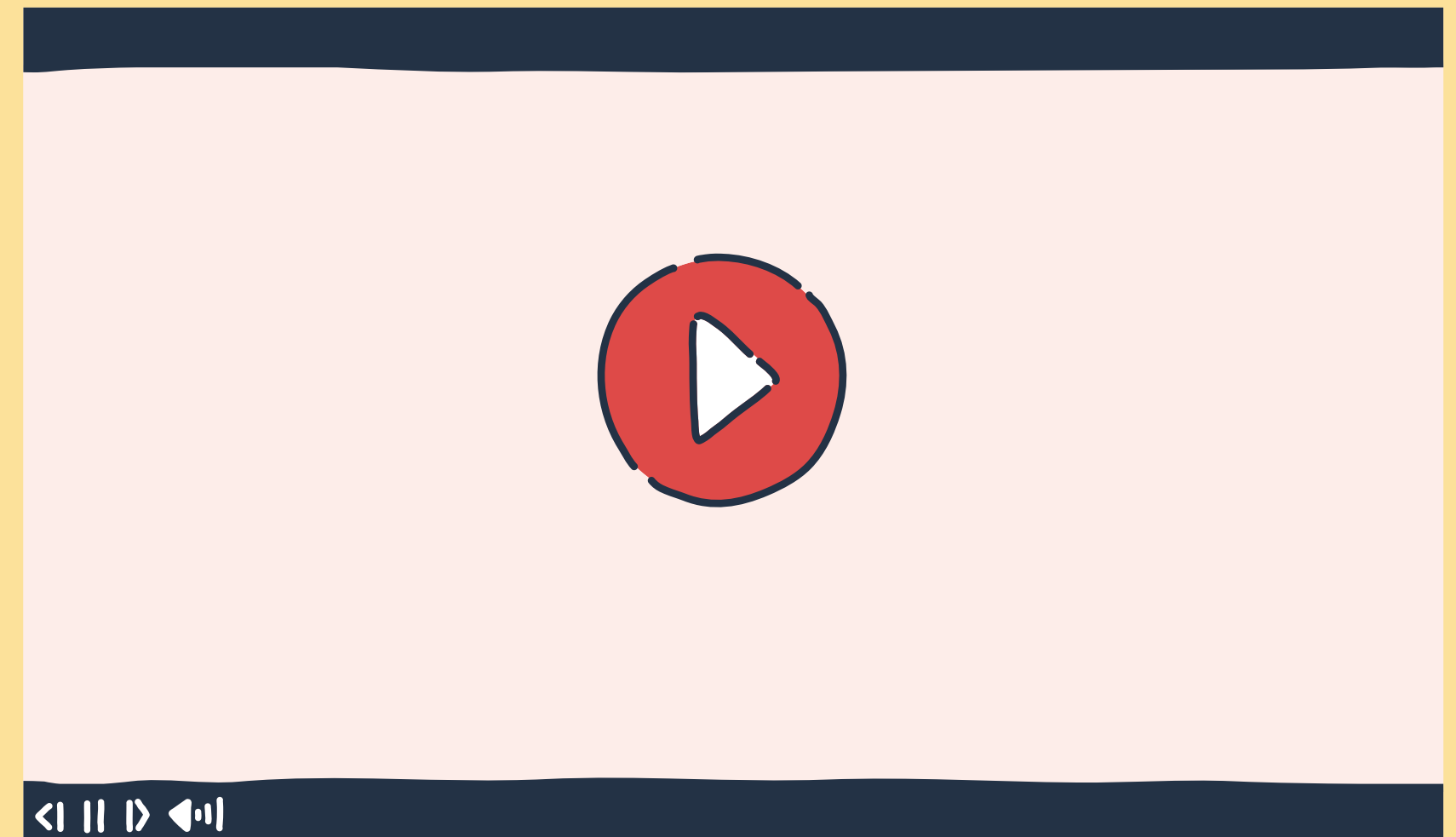


Biểu đồ maps

## 3.2 TRỰC QUAN HÓA

Gồm có 6 dashboard chính:

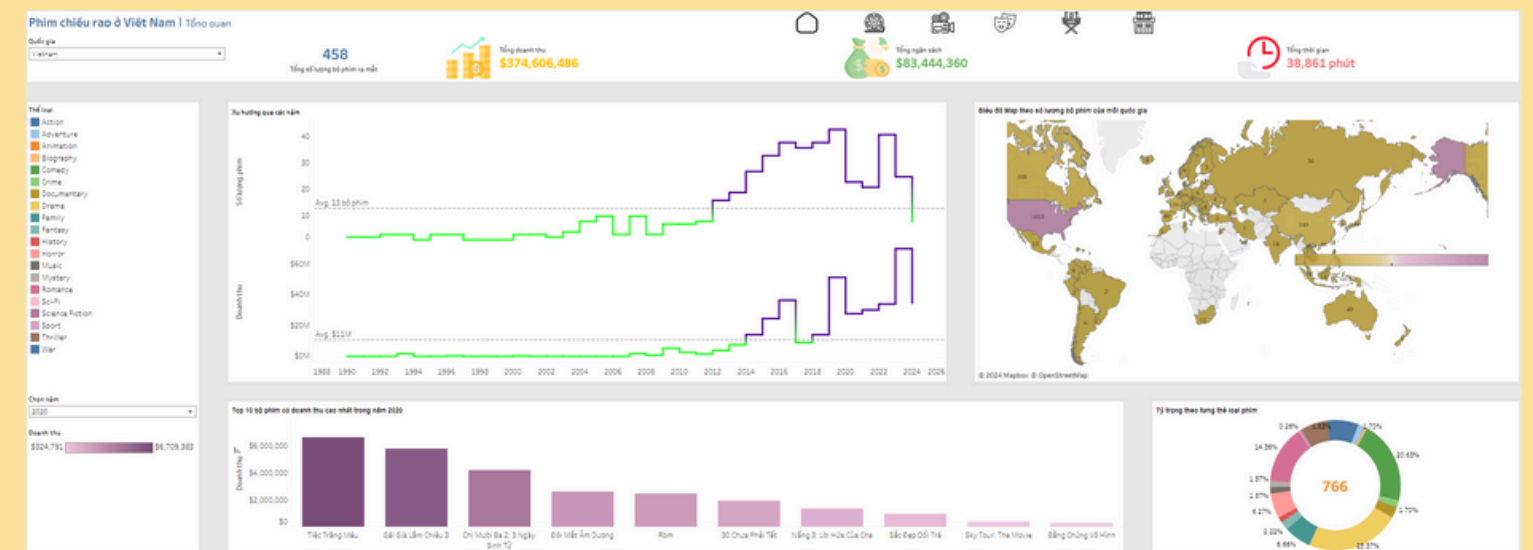
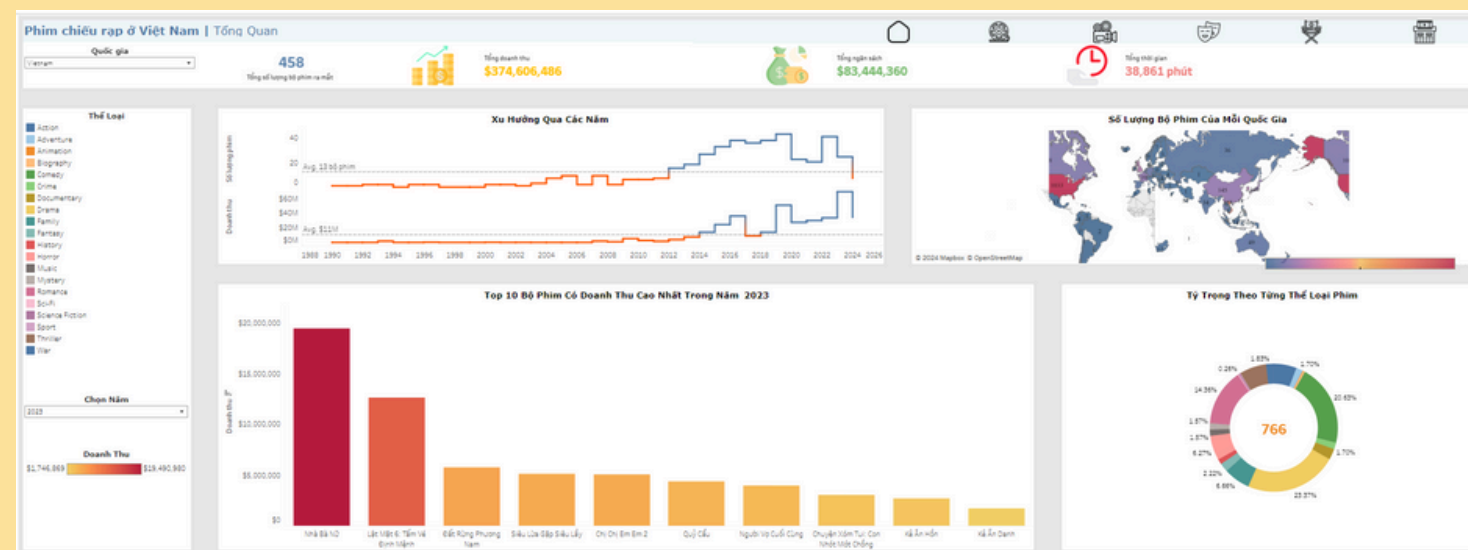
- Dashboard tổng quan
- Movie Dashboard
- Genres Dashboard
- Actors Dashboard
- Directors Dashboard
- Company Dashboard



# 3.2 TRỰC QUAN HÓA











Phù hợp với mọi người

Đối với những người mù màu



## NỘI DUNG 5

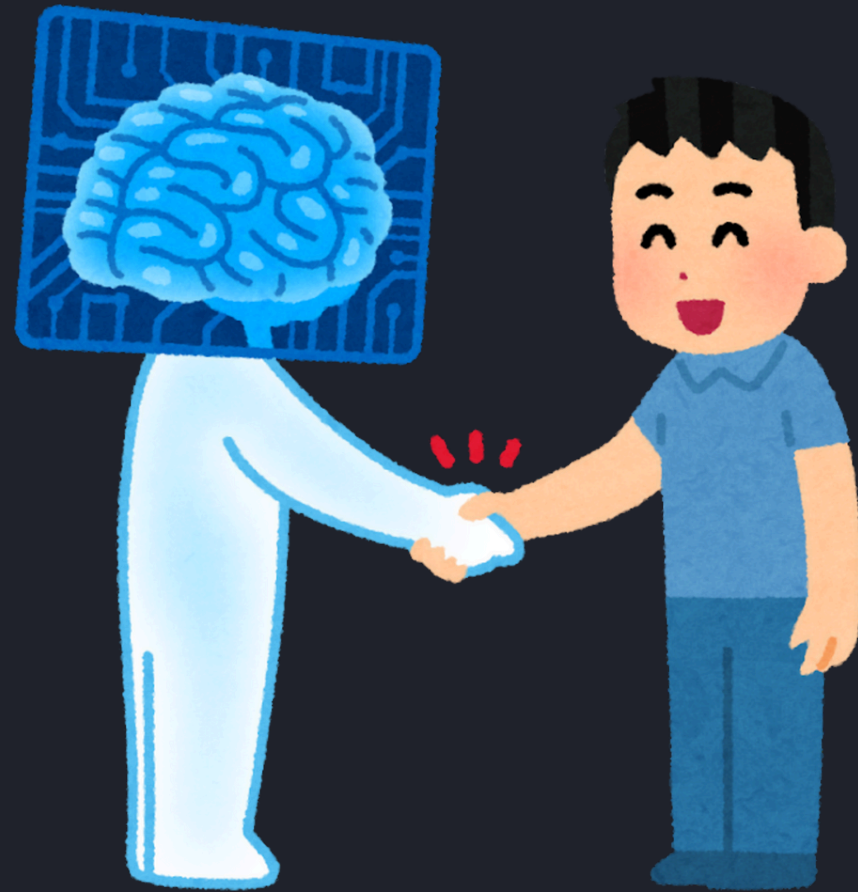
# CÁC TIÊU CHÍ ĐÁNH GIÁ

STT	 <b>TIÊU CHÍ</b>	<b>MÔ TẢ</b> 	<b>ĐÁNH GIÁ</b> 
1	Kết hợp nguồn dữ liệu đáng tin cậy	Nhóm thu thập dữ liệu từ các nguồn uy tín là: Moveek, IMDB và TMDB.	
2	Phù hợp với mục đích	Các biểu đồ tròn, cột, đường, bong bóng, treemaps,... được sử dụng đúng công dụng	
3	Rõ ràng và dễ hiểu	Biểu đồ và đồ thị rõ ràng, dễ hiểu, có chú thích và đơn vị đo lường rõ ràng, giúp người xem nhanh chóng hiểu thông tin.	
4	Sự tích hợp và liên kết	Các biểu đồ được liên kết trực quan, tạo cái nhìn toàn diện về chủ đề được đề cập.	
5	Phân tích được sự thay đổi và xu hướng	Trực quan hóa rõ sự thay đổi doanh thu, kinh phí theo thời gian của các bộ phim chiếu rạp và mối quan hệ giữa thể loại, diễn viên, đạo diễn và nhà sản xuất phim.	
6	Tương tác và điều hướng	Cho phép người xem tùy chỉnh và lọc để xem dữ liệu từ nhiều hướng và góc nhìn. Từ đó cung cấp cho người xem nhiều thông tin hữu ích về các bộ phim chiếu rạp tại VN.	
7	Thiết kế hấp dẫn	Dashboard được thiết kế dựa vào nền là màu xám nhạt và trắng. Chọn những màu thiên nhạt để làm dịu mắt người xem và thu hút người xem. Có lựa chọn thích hợp cho những người mù màu cũng có thể xem và phân biệt.	

## NỘI DUNG 6

# ĐỀ XUẤT CẢI TIẾN

# 6. ĐỀ XUẤT CẢI TIẾN



- Tích hợp hệ thống gợi ý và phản hồi của người dùng
- Giúp cải thiện tương tác với người dùng



- Phân tích sâu hơn về mô hình dự đoán và xu hướng
- Đề xuất các xu hướng trong tương lai



- Thiết kế luồng dữ liệu tự động cập nhập
- Dữ liệu có thể cập nhật và đồng bộ với dashboard

**THE END**

**THANK  
YOU!**

**Trình bày bởi: Nhóm 07**