



# Documentary & NYFF

Roadmap for Production Awards  
at the Documentary Section,  
New York Film Festival

# 01

## GOAL



# QUESTION

In NYFF Production Awards Section,  
once a documentary film  
is awarded the Production Fund:

What are some **factors**  
on **distribution** side  
that affect **viewership**?

# GOAL

WHAT:

Develop model analyzing  
documentary viewership

TARGET:

Viewership as measured  
by box office numbers

02

## METHODOLOGY & STRATEGY



DATA

## Data on 1969 movies

Box Office Mojo  
by IMDbPro

Search for Titles

IMDbPro

f

Twitter

Domestic

International

Worldwide

Calendar

All Time

Showdowns

Indices

Brands

Franchises

Genres

Genre Keyword: Documentary

Titles

← Previous page

1-100 of 2,055

Next page →

Rank	Title	Lifetime Gross	Max Theaters	Opening	Open Th	Release Date	Distributor
1	Fahrenheit 9/11	\$119,194,771	2,011	\$23,920,637	868	Jun 23, 2004	Lionsgate
2	March of the Penguins	\$77,437,223	2,506	\$137,492	4	Jun 24, 2005	Warner Independent Pictures (WIP)
3	Justin Bieber: Never Say Never	\$73,013,910	3,118	\$29,514,054	3,105	Feb 11, 2011	Paramount Pictures
4	This Is It	\$72,091,016	3,481	\$23,234,394	3,481	Oct 28, 2009	Sony Pictures Entertainment (SPE)
5	Under the Sea 3D	\$36,262,926	108	\$661,185	49	Feb 13, 2009	Warner Bros.
6	2016: Obama's America	\$33,449,086	2,017	\$31,610	1	Jul 13, 2012	Rocky Mountain Pictures
7	Earth	\$32,011,576	1,804	\$8,825,760	1,804	Apr 22, 2009	Walt Disney Studios Motion Pictures

# MODELING

## Data / Category

Box Office Mojo scraping  
Documentary films (by ranking)

## Modeling

Looking at relationships between features  
through linear regression (simple)

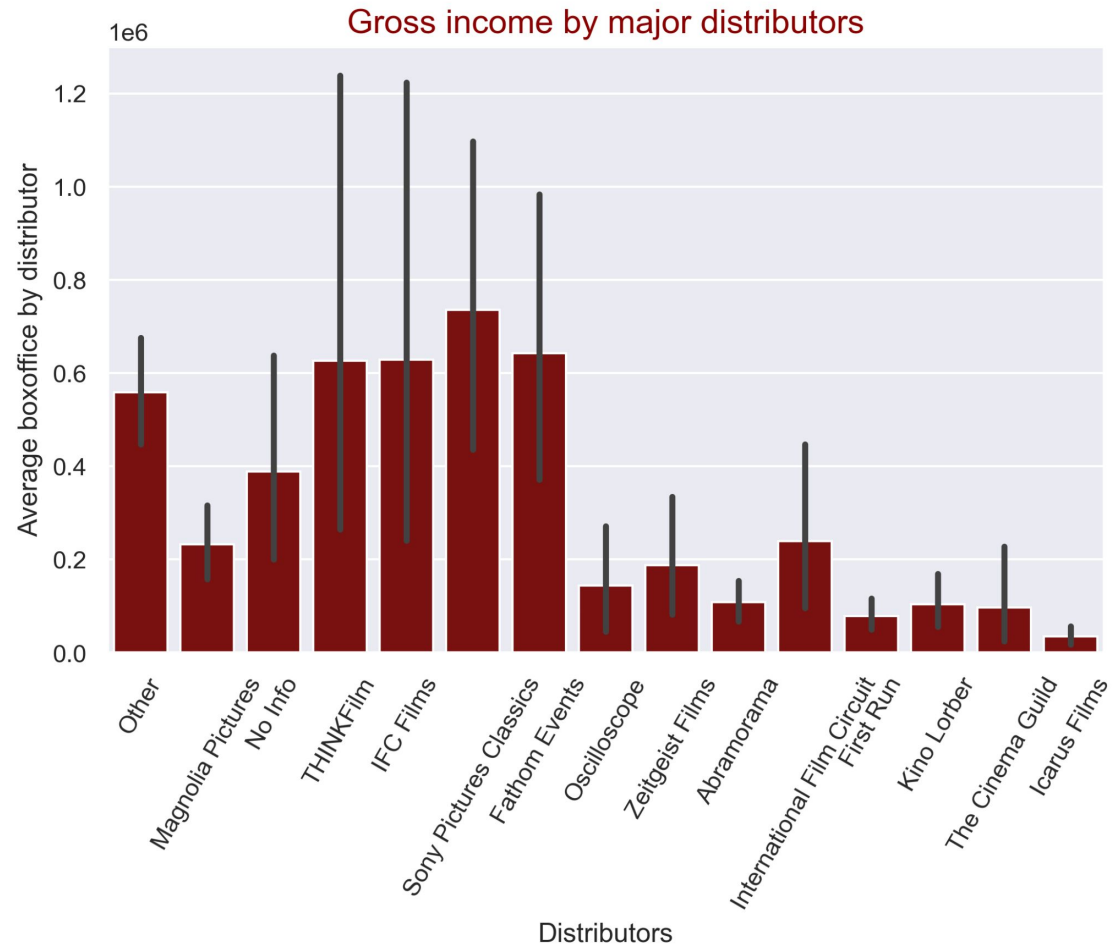
## Target: Intl Box Office

Measuring viewership

## Features: Distribution factors

- Theaters the film is shown in
- Distributor
- Runtime
- Release Year

# STRATEGY





# 03

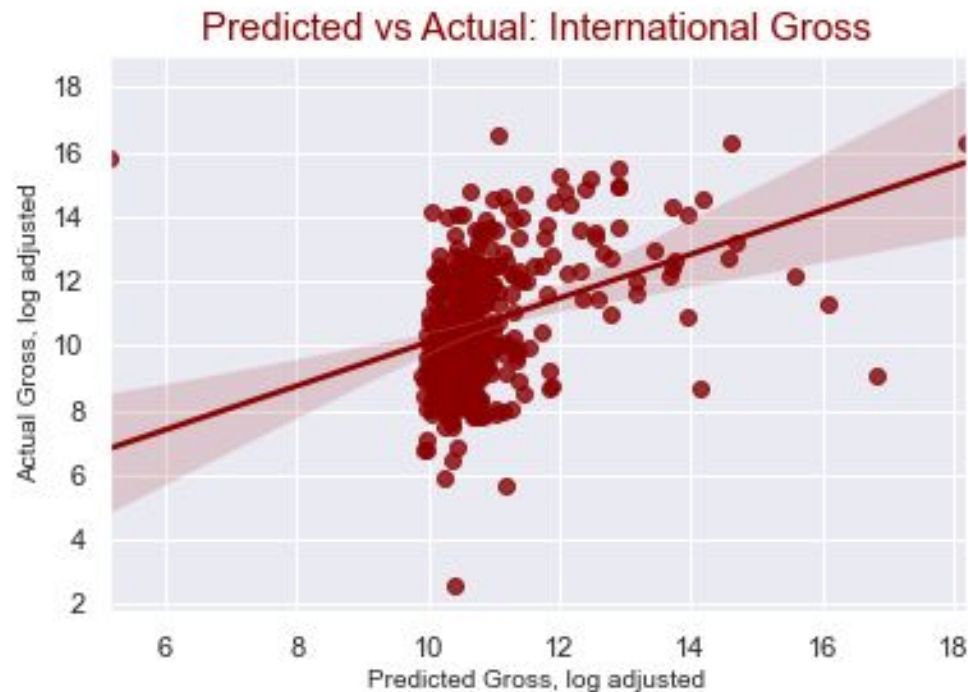
## FINDINGS



# RESULTS

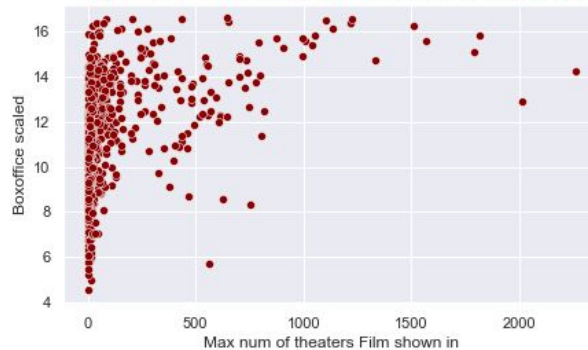
$R^2$  : 0.0974

MAE: 1.41

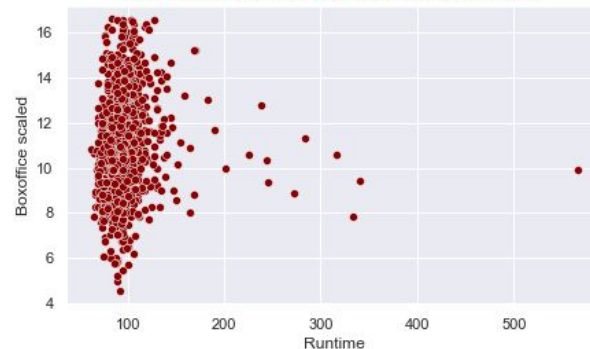


# POSSIBLE FEATURE CHARACTERISTICS THAT RESULT IN LOW CORRELATION

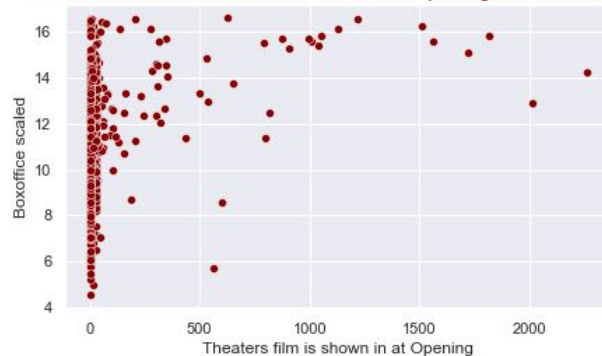
Correlation between Num of Theaters and Total Gross



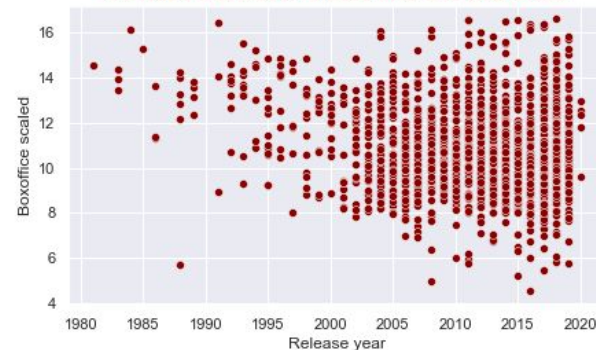
Correlation between Runtime and Total Gross



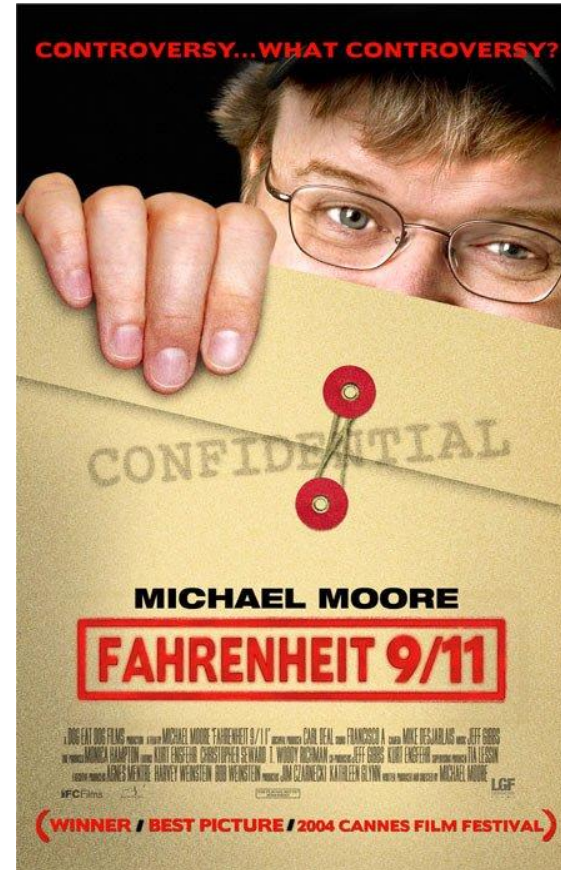
Correlation between Num of Theaters at Opening and Total Gross



Correlation between Release Year and Total Gross



# MAJOR BLOCKBUSTER



# CONCLUSIONS

**01**

Runtime Corr

**03**

# theaters, all

**05**

Distributors

**02**

Release Year Corr

**04**

# theaters, opening

**Rule out**

**Most features**

Some correlation,  
but model is underfit;  
Need more data

# 04

## FUTURE WORK



# FEATURES TO INCLUDE & OTHER CONSIDERATIONS

FEATURE **Genre:** Box Office Mojo  
How it affects viewership

FEATURE **Country of Cast:**  
Scraped from Wikipedia  
Whether some countries have a wider reach

FEATURE **Budget:**  
THINK OF : How to collect this data

METRIC **Distribution, intl**  
THINK OF : What target this would require  
LOOK AT: Better ways of combining intl & domestic gross

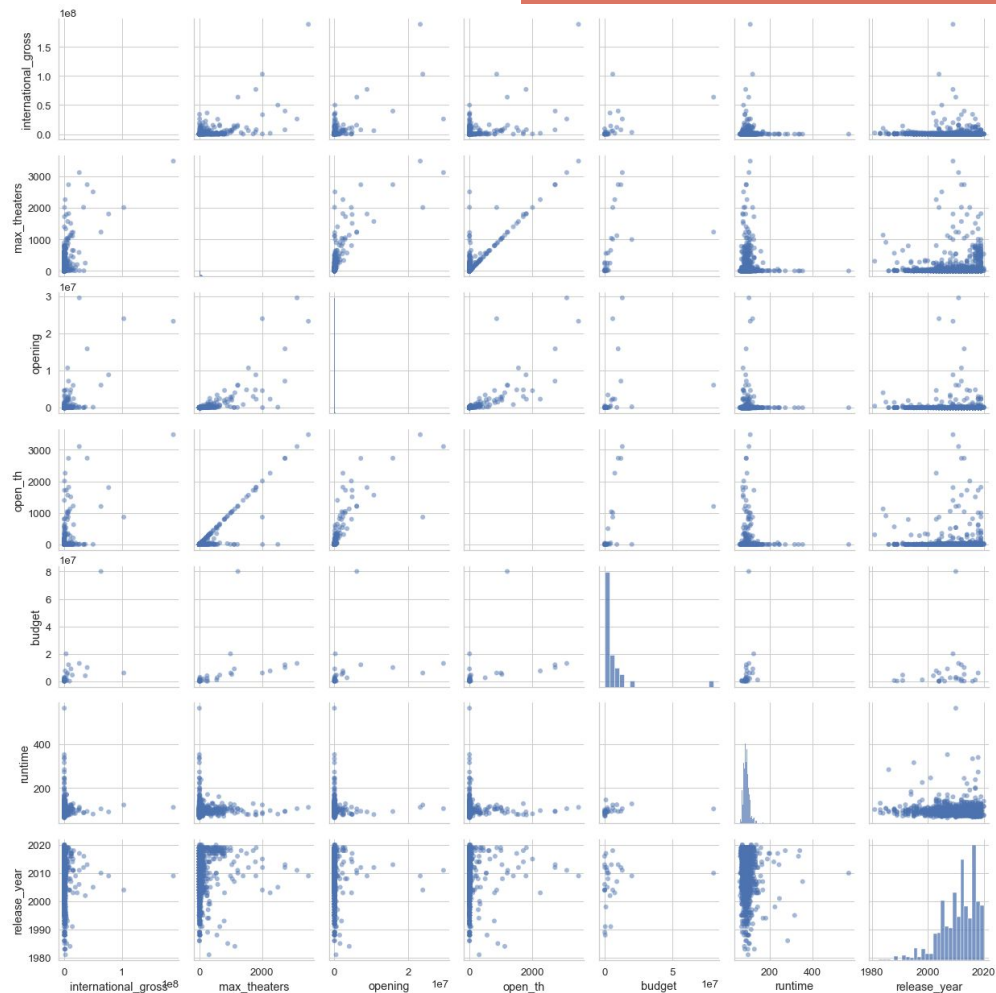


# APPENDIX

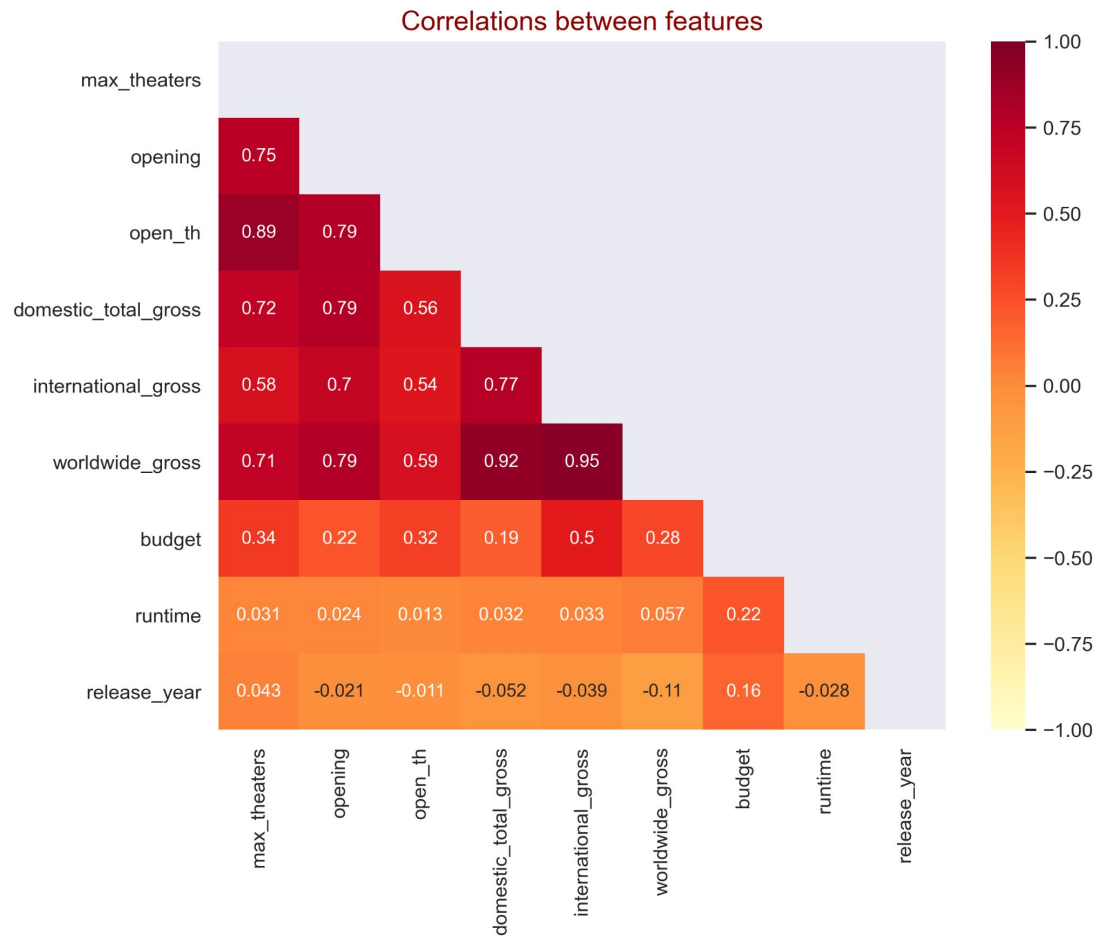




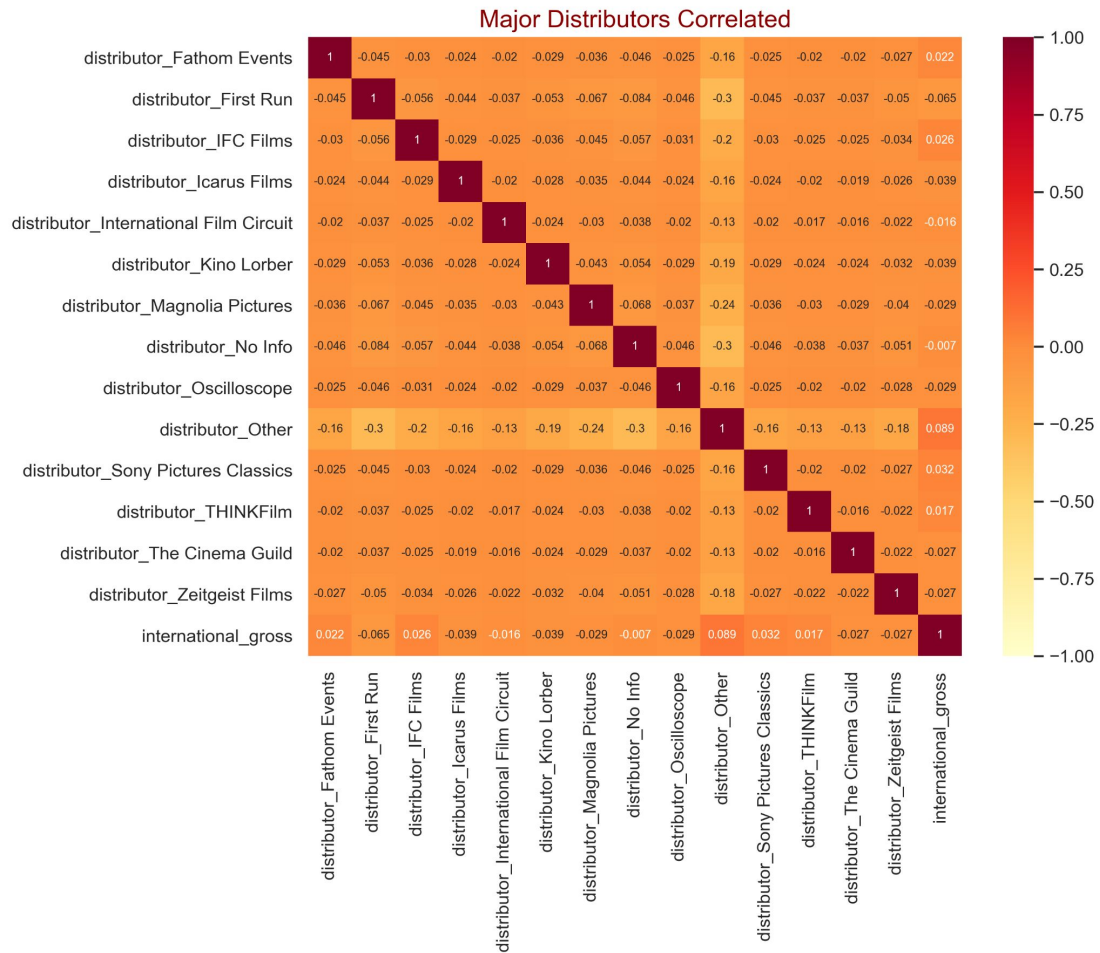
# Initial Pairplot (w/out having reduced any features)



# Numeric Feature Selection



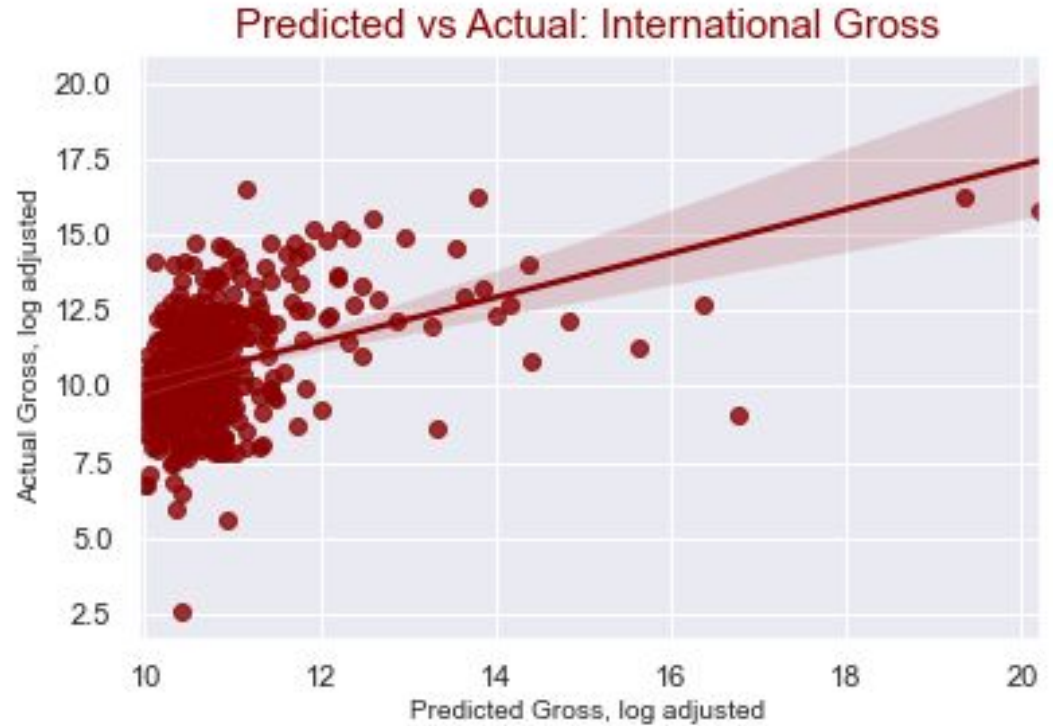
# Distributor Selection



# EARLIER RESULTS

After CV

$R^2 : 0.211$



## Equation

INTERCEPT : 123.16235734485772

COEF :

max\_theaters: 1.15708169e-02

open\_th: -1.43197170e-03

runtime: 3.35741423e-03

release\_year: -5.62119081e-02

distributor\_Fathom Events: -2.49627476e+00

distributor\_IFC Films: 6.21991486e-01

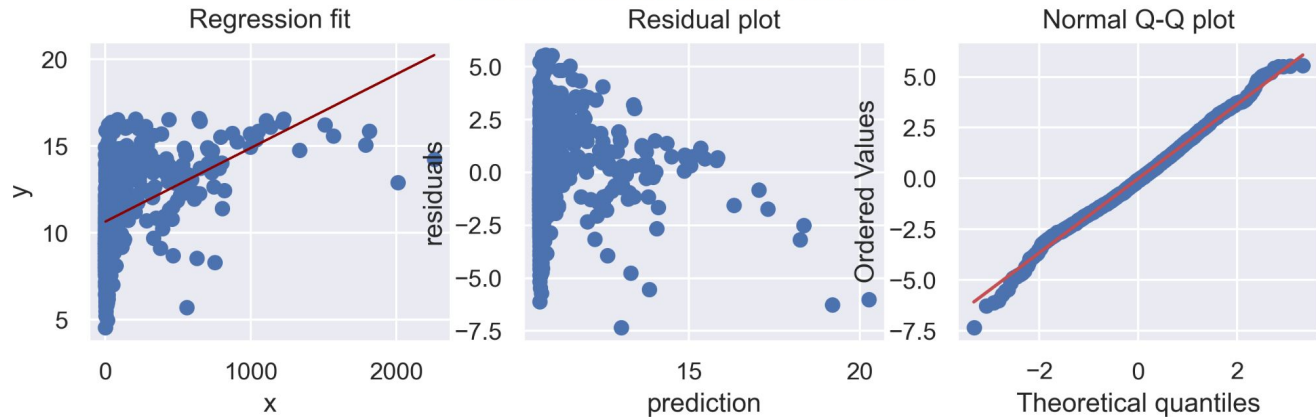
distributor\_THINKFilm: 4.92010400e-01

distributor\_Sony Pictures Classics: 8.43896374e-01

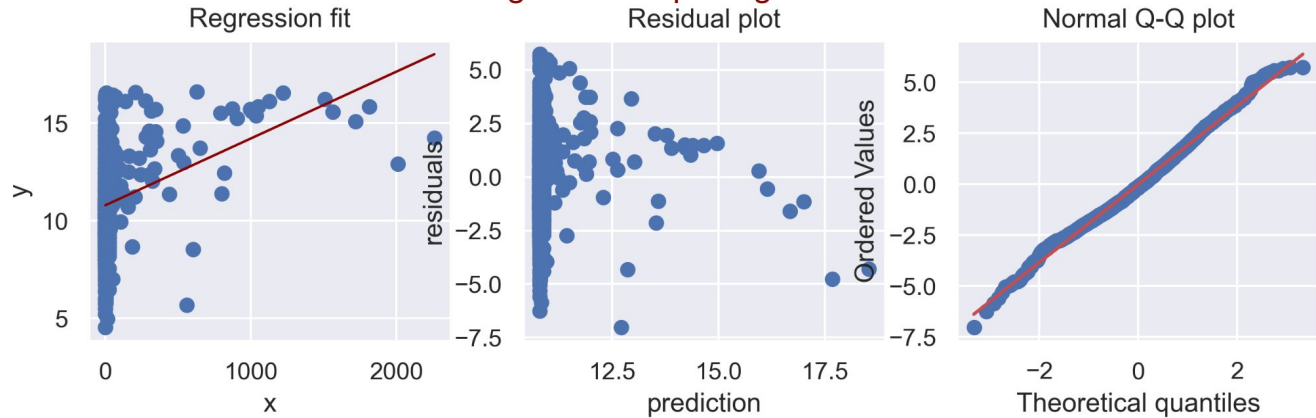
max\_th<sup>2</sup>: -4.41098067e-06

# Diagnostics 1

Diagnostics: Maximum Theaters

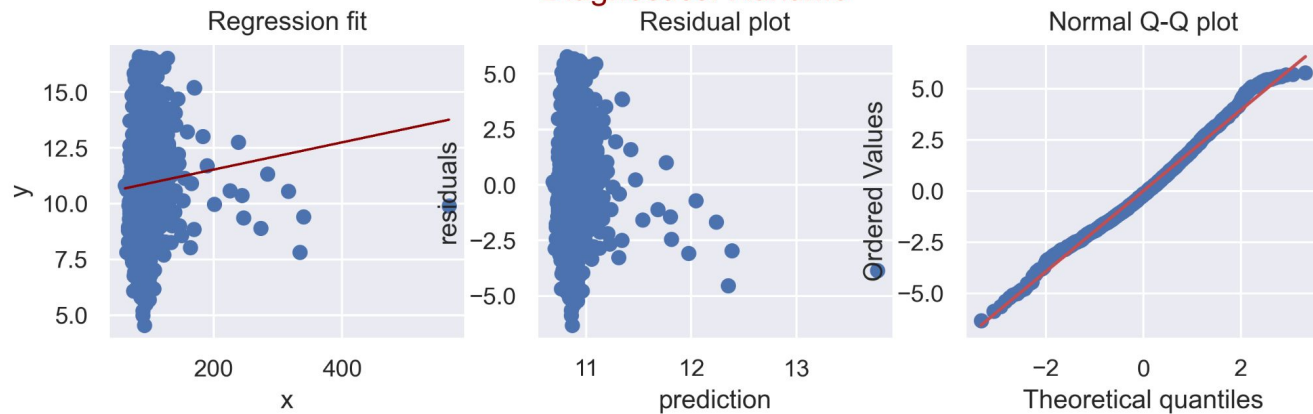


Diagnostics: Opening Theaters



# Diagnostics 2

## Diagnostics: Runtime



## Diagnostics: Release Year

