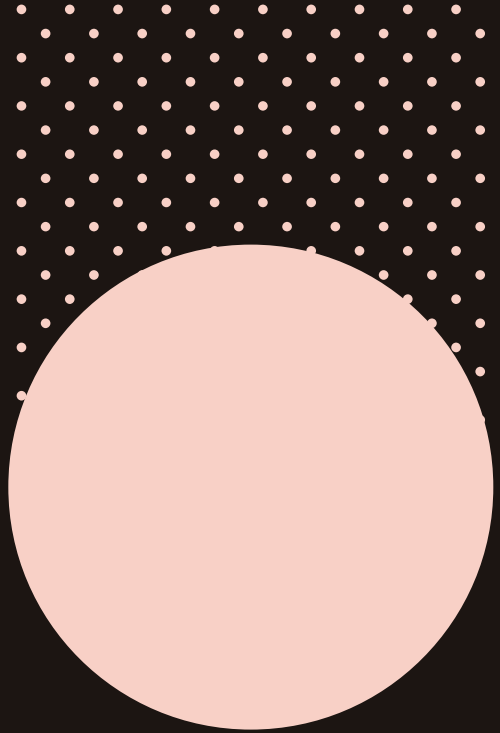
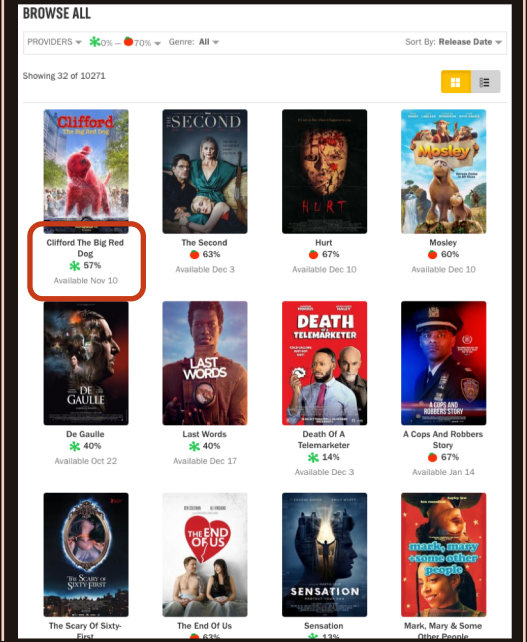


Movie Profit Analysis

A dive into the indicators of a successful movie

Web Scraping Pipeline

Rotten Tomatoes scrape of all movies with a Tomatometer score below 70%



Merge with IMDB datasets on movie title, year, and IMDB's unique ID tconst

Tomatoes URL	Title	Tomatometer	Year	N_Tomatometer
https://www.rottentomatoes.com/m/the_second	The Second	63%	2018	8
https://www.rottentomatoes.com/m/needle_in_a_timestack	Needle in a Timestack	34%	2021	48
https://www.rottentomatoes.com/m/adventures_of_a_mathematician	Adventures of a Mathematician	50%	2020	10
https://www.rottentomatoes.com/m/de_gaulle	De Gaulle	40%	2020	10
https://www.rottentomatoes.com/m/the_estate_2021	The Estate	33%	2020	9

+

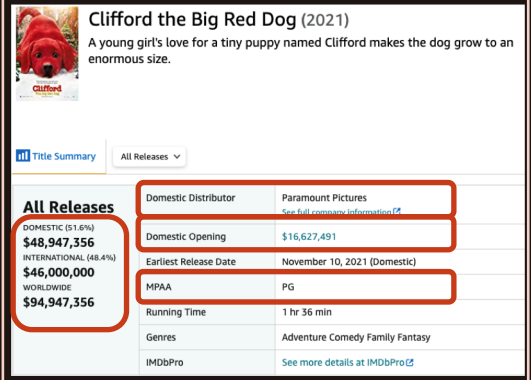
tconst	Title	isAdult	Year	runtimeMinutes	genres
tt0214985	Nellu	0	1974	\N	Drama,Romance
tt1814917	Trippers	0	2010	74	Drama,Romance
tt0446994	Devenir	0	2005	78	Documentary

+

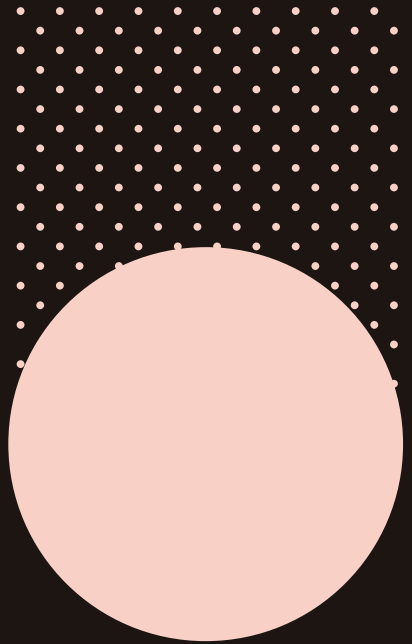
tconst	averageRating	numVotes
tt0398438	8.0	138
tt2988442	8.0	187
tt1079781	6.0	5
tt2319530	5.2	21

Scrape Box Office Mojo for data on revenue, MPAA rating, etc.

https://www.boxofficemojo.com/title/tt2397461



Characteristics of this dataset

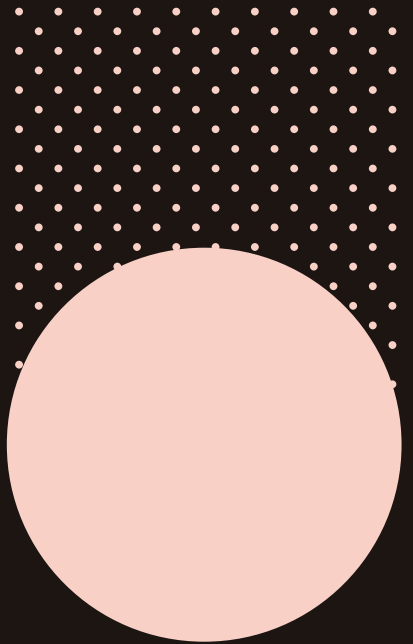


Contains only movies with a Rotten Tomatoes score of 70% or lower

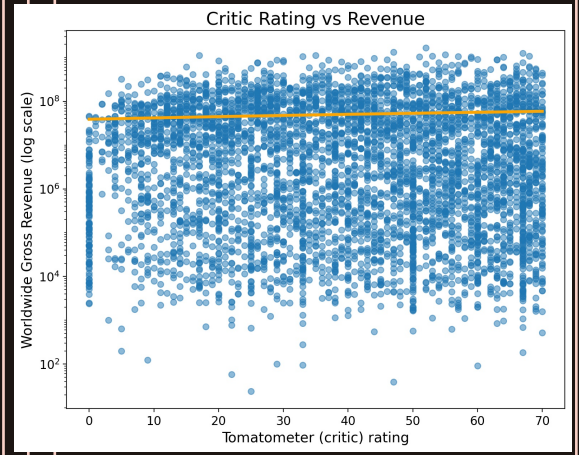
8000 rows and 18 columns of data, including 2 categorical variables

Only movies released in the last 20 years were used in data analysis

Worldwide Gross Revenue was chosen as the target variable / indicator of movie success



Critic and Audience Movie Ratings



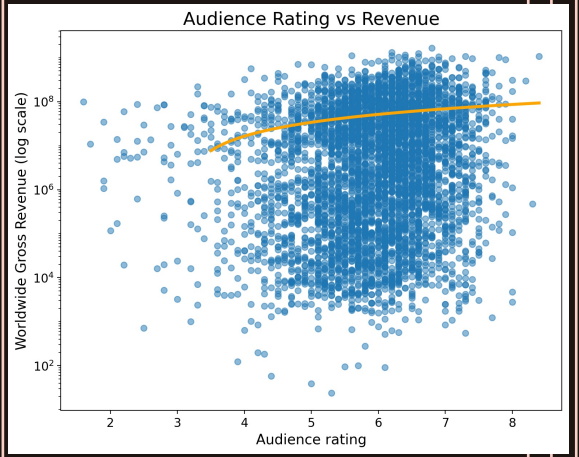
R2: 0.002
slope: 286,838.63
y-intercept: 38778532.79

HYPOTHESIS

Higher rated movies will correlate with more success at the box office

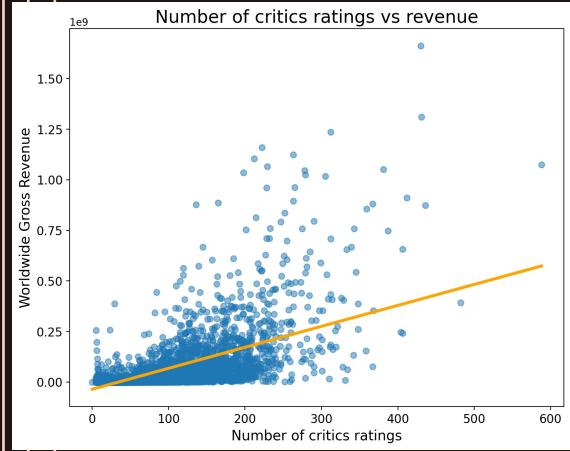
OBSERVATION / CONCLUSION

Movie ratings (both critic and audience) are positively correlated with worldwide gross revenue



R2: 0.02
slope: 17,486,441.96
y-intercept: -53508880.75

Number of Movie Ratings



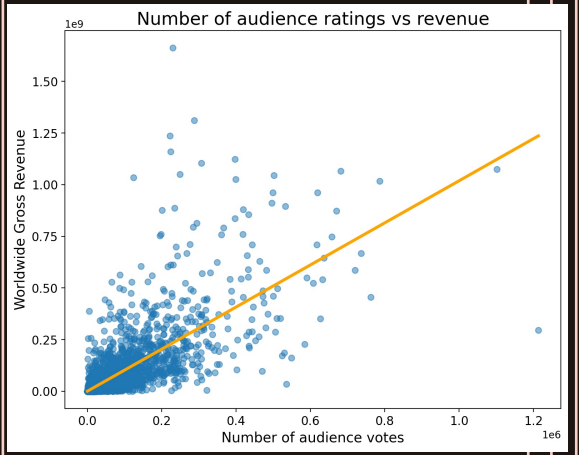
R2: 0.39
slope: 1,036,550.07
y-intercept: -35496902.90

HYPOTHESIS

The number of ratings for each film will be positively correlated to box office success

OBSERVATION / CONCLUSION

The number of ratings (both critic and audience) do correlate positively with box office success



R2: 0.56
slope: 1018.62
y-intercept: 527721.74

MPAA Rating

HYPOTHESIS

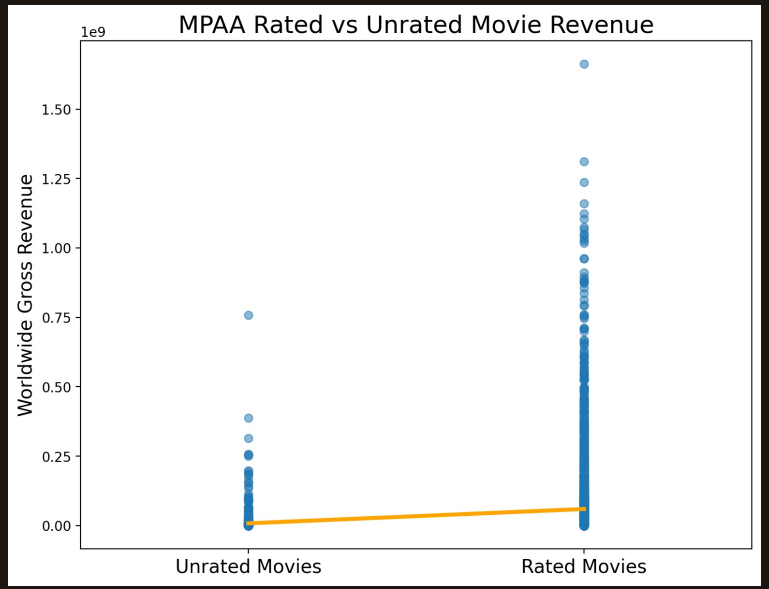
Unrated movies do less well at the box office

OBSERVATIONS

Movies that are unrated
average a WGR of $59 million
The average Worldwide Gross is $51 million higher for movies that have an MPAA rating

CONCLUSION

Movies with an MPAA rating do much better at the box office



R2 = 0.03
slope: 51,560,644.23
y-intercept: 59,777,843.73

Movie Genre

HYPOTHESIS

Movies in the top 5 genres do better at the box office

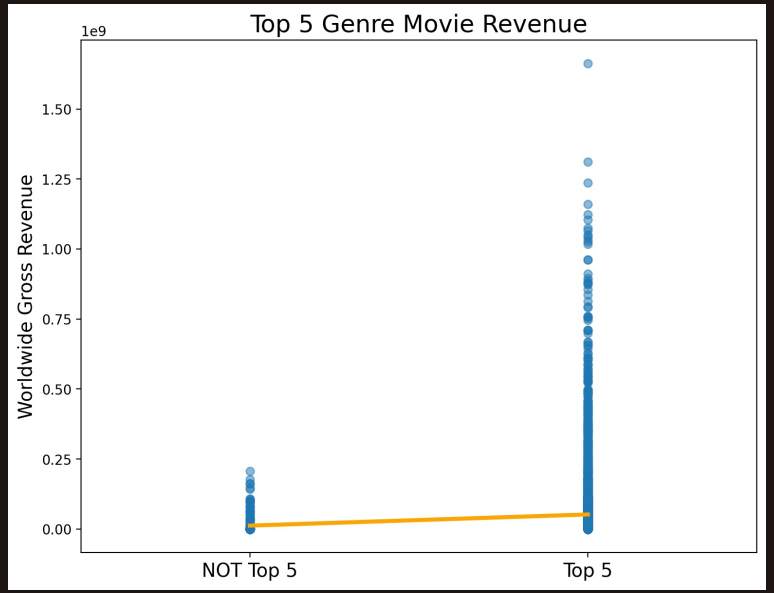
Action
Adventure
Drama
Comedy
Thriller

OBSERVATIONS

Movies that are not in the Top 5 genres average a WGR of $12 million
The average Worldwide Gross is $40 million higher for movies that are in one of the top 5 genres

CONCLUSION

Movies in the top 5 genres do much better at the box office



R2: 0.006
slope: 39,817,265.21
y-intercept: 12,382,206.25

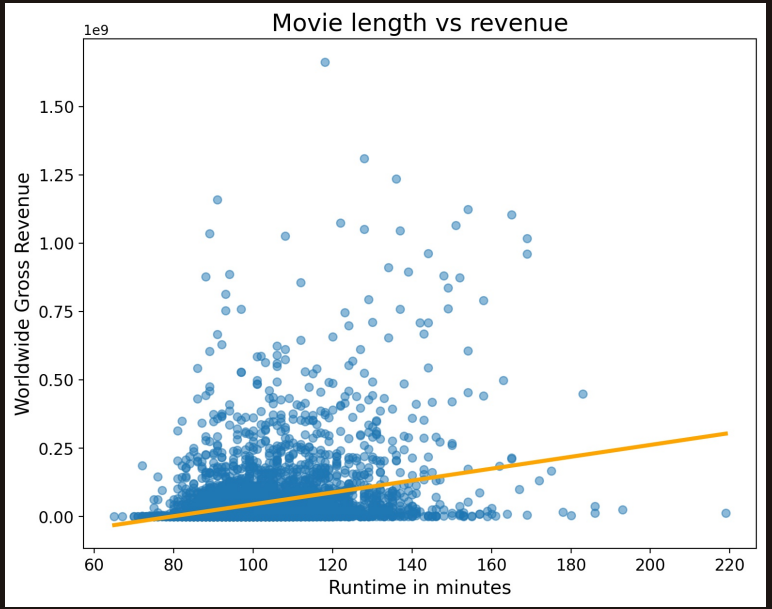
Movie Length

HYPOTHESIS

Movie length does not correlated with how well a movie does at the box office

OBSERVATIONS / CONCLUSION

Movie length has a positive correlation with worldwide gross revenue



R2: 0.076
slope: 2,172,733.13
y-intercept: -17,2713,176.22

Regularization – multiple feature linear regression

TARGET:
- Worldwide Gross Revenue

8 FEATURES:
- Tomatometer Score
- Year Released
- Number of Tomatometer Ratings
- Run Time
- Average Audience Rating
- Number of Audience Votes
- Domestic Opening
- Is movie rated?
- Is movie in the Top 5 Genres?

#	Column	Non-Null Count	Dtype
0	Tomatometer	4580 non-null	int64
1	Year	4580 non-null	int64
2	N_Tomatometer	4580 non-null	int64
3	runtimeMinutes	4580 non-null	float64
4	Average_Audience_Rating	4580 non-null	float64
5	Number_Audience_Votes	4580 non-null	int64
6	Worldwide_Gross	4580 non-null	float64
7	Domestic_Opening	4580 non-null	float64
8	Unrated	4580 non-null	int64
9	Top_5_Genres	4580 non-null	float64

R-squared:	0.837
Adj. R-squared:	0.837

	coef	std err	t	P>|t|
const	-2.107e+09	2.73e+08	-7.726	0.000
Tomatometer	6.874e+04	4.49e+04	1.533	0.125
Year	1.022e+06	1.35e+05	7.550	0.000
N_Tomatometer	495.3461	1.58e+04	0.031	0.975
runtimeMinutes	2.189e+05	5.35e+04	4.094	0.000
Average_Audience_Rating	2.124e+06	1.04e+06	2.033	0.042
Number_Audience_Votes	195.9089	13.947	14.046	0.000
Domestic_Opening	6.6582	0.080	83.010	0.000
Unrated	6.134e+06	2.06e+06	2.977	0.003
Top_5_Genres	5.8e+06	3.25e+06	1.784	0.074

Omnibus:	3484.068	Durbin-Watson:	1.940
Prob(Omnibus):	0.000	Jarque-Bera (JB):	216107.342
Skew:	3.079	Prob(JB):	0.00
Kurtosis:	36.084	Cond. No.	6.13e+09

> Drop features with Pvalue>0.05 : Tomatometer, N_Tomatometer, Top_5_Genres

R-squared:	0.837
Adj. R-squared:	0.837

	coef	std err	t	P>|t|
const	-2.133e+09	2.71e+08	-7.864	0.000
Year	1.037e+06	1.35e+05	7.693	0.000
runtimeMinutes	2.317e+05	5.17e+04	4.477	0.000
Average_Audience_Rating	2.807e+06	8.78e+05	3.199	0.001
Number_Audience_Votes	197.2809	13.199	14.947	0.000
Domestic_Opening	6.6529	0.077	86.816	0.000
Unrated	6.025e+06	1.88e+06	3.198	0.001

Omnibus:	3486.675	Durbin-Watson:	1.939
Prob(Omnibus):	0.000	Jarque-Bera (JB):	216545.149
Skew:	3.082	Prob(JB):	0.00
Kurtosis:	36.117	Cond. No.	6.10e+09

> R2 and adj R2 remain the same, but this is a better model since it's less complex, low p-values
> high Condition No. still indicates strong multicollinearity

Fitting an ElasticNet Regression

TARGET:
- Worldwide Gross Revenue

8 FEATURES:
- Tomatometer Score
- Year Released
- Number of Tomatometer Ratings
- Run Time
- Average Audience Rating
- Number of Audience Votes
- Domestic Opening
- Is movie rated?
- Is movie in the Top 5 Genres?

#	Column	Non-Null Count	Dtype
0	Tomatometer	4580 non-null	int64
1	Year	4580 non-null	int64
2	N_Tomatometer	4580 non-null	int64
3	runtimeMinutes	4580 non-null	float64
4	Average_Audience_Rating	4580 non-null	float64
5	Number_Audience_Votes	4580 non-null	int64
6	Worldwide_Gross	4580 non-null	float64
7	Domestic_Opening	4580 non-null	float64
8	Unrated	4580 non-null	int64
9	Top_5_Genres	4580 non-null	float64

> Split data into 80% training / 20% testing data
> Scale the values using standard scalar
> test_r_squared = 0.767 (with standard alpha = 1, l1_ratio=.5)
> ElasticNet coefficients:
- ('Tomatometer', -293299.27326428954),
- ('Year', 3542483.6379300943),
- ('N_Tomatometer', 13844947.69541261),
- ('runtimeMinutes', 3686288.199194253),
- ('Average_Audience_Rating', -381037.31166927726),
  ('Number_Audience_Votes', 26288007.541837066),
- ('Domestic_Opening', 51607661.30830691),
- ('Unrated', 945657.9293973312),
- ('Top_5_Genres', 330971.45738288405)

r_squared_train: 0.833
r_squared_test: 0.852
alpha: 100.0
l1_ratio: 1.0

Thanks for watching!