MOVIE INDUSTRY ANALYSIS

About Dataset

The dataset provides a comprehensive view of the movie industry, containing various attributes that can reveal key insights about factors contributing to a movie's success or failure. It includes columns such as budget, representing the amount spent on producing the movie, and gross, which indicates its revenue, allowing for an assessment of profitability and financial performance. Other columns, such as rating, genre, and release year, offer context on the type and timing of the film, making it possible to identify trends over time and across demographics. Additionally, columns like director and writer allow us to analyze successful creative collaborations, Overall, this dataset enables a detailed exploration of how various production and release factors influence a movie's impact and reception, offering a robust foundation for analysis in the field of data science and entertainment industry studies.

Objective

The objective of analyzing this movie dataset is to uncover the key factors that contribute to a movie's success or failure, with a specific focus on identifying elements that increase the likelihood of a movie becoming a blockbuster versus a flop. By examining attributes such as budget, revenue (gross), genre, rating, and release year, as well as creative partnerships (e.g., director-writer combinations), this analysis aims to determine patterns and trends that drive high performance in the film industry. Additionally, the analysis seeks to identify the most successful genres, rating categories, and director-writer collaborations, providing actionable insights that could inform future movie production and marketing strategies.

Overview of Dataset

- There are 7408 movies in the dataset
- Total Rows : 7408
- Total Columns : 15

```
desc movies_cleaned1;
```

Field	Type	Null	Key	Default
name	text	YES		NULL
rating	text	YES		NULL
genre	text	YES		NULL
year	int	YES		NULL
released	text	YES		NULL
score	double	YES		NULL
votes	int	YES		NULL
director	text	YES		NULL
writer	text	YES		NULL
star	text	YES		NULL
country	text	YES		NULL
budget	int	YES		NULL
gross	int	YES		NULL
company	text	YES		NULL
runtime	int	YES		NULL

Data Cleaning

There are 2 null values in released and 189 null values in gross.

Removed all Missing values

No Duplicate values are found

Feature Extraction

As part of the data preparation process, new columns, final_status and movie era, were generated to enhance the analysis:

Final Status: This column categorizes movies into labels such as “blockbuster”,”Superhit”, "hit" or "flop" based on a comparison between budget and gross revenue. By grouping movies into these performance categories, we can more easily analyze what contributes to high or low success rates across different genres, ratings, and director-writer combinations.

Movie Era: To better understand historical trends, the release years were grouped into distinct eras, such as “classic”,”90’s Hits”,”Early 20’s”and”modern”. This column allows us to explore shifts in movie production styles, audience preferences, and genre popularity over time.

Analysis

Determine Factors that Lead to Blockbuster or Flop Movies:

# Compare average gross for blockbuster, Superhit,hit, and flop movies

select final_status,avg(gross) as avg_gross,avg(budget) as avg_budget from movies_cleaned1 group by final_status;

final_status	avg_gross	avg_budget
Hit	89679038.98	39682387.8
Blockbuster	180341932.9	14254423.35
Superhit	223062139.6	41970760.08
Flop	9620066.557	25332411.92

```
# Find rating categories that are common among blockbusters

select rating,count(*) as blockbuster_count from movies_cleaned1 where
final_status='blockbuster' group by rating order by blockbuster_count desc;
```

rating	blockbuster_count
R	274
PG-13	141
PG	111
Not Rated	19
G	14
NC-17	2
Unrated	1
TV-MA	1
TV-14	1

```
# Analyze genres with higher blockbuster rates

select genre,count(*) as genre_count,sum(case when final_status='blockbuster'
then 1 else 0 end) as blockbuster_count from movies_cleaned1 group by genre
order by blockbuster_count desc;
```

genre	genre_count	blockbuster_count
Comedy	2182	181
Drama	1438	106
Action	1663	95
Horror	304	63
Animation	331	33
Adventure	419	29
Crime	536	26
Biography	429	25
Fantasy	42	2
Family	10	2
Thriller	12	1
Mystery	20	1
Romance	8	0
Music	1	0
Western	3	0
Sci-Fi	8	0
Sport	1	0
Musical	1	0

```
# Find the most successful director-writer combinations

select director,writer,count(*) as movie_count,sum(case when
final_status='blockbuster' then 1 else 0 end) as blockbuster_count from
movies_cleaned1 group by director,writer order by blockbuster_count
desc,movie_count desc limit 25;
```

director	writer	movie_count	blockbuster_count
Woody Allen	Woody Allen	37	4
M. Night Shyamalan	M. Night Shyamalan	11	4
Hayao Miyazaki	Hayao Miyazaki	6	3
James Wan	Leigh Whannell	4	3
Alex Kendrick	Alex Kendrick	4	3
Peter Jackson	J.R.R. Tolkien	3	3
James DeMonaco	James DeMonaco	3	3
Pedro Almodóvar	Pedro Almodóvar	13	2
Richard Linklater	Richard Linklater	10	2
Kevin Smith	Kevin Smith	10	2
Luc Besson	Luc Besson	9	2
John Hughes	John Hughes	8	2
Robert Zemeckis	Robert Zemeckis	5	2
Eli Roth	Eli Roth	5	2
Alan Alda	Alan Alda	4	2
David Yates	Steve Kloves	3	2
Adam Wingard	Simon Barrett	3	2
Leigh Whannell	Leigh Whannell	3	2
Kenneth Lonergan	Kenneth Lonergan	2	2
Chris Columbus	J.K. Rowling	2	2
Shane Carruth	Shane Carruth	2	2
Darren Lynn Bousman	Leigh Whannell	2	2
John Carney	John Carney	2	2
John Madden	Ol Parker	2	2
James Wan	Chad Hayes	2	2

```
# Overall count of movies made by each director-writer combo

select director,writer,count(*) as total_movies from movies_cleaned1 group by
director,writer order by total_movies desc limit 25;
```

director	writer	total_movies
Woody Allen	Woody Allen	37
Pedro Almodóvar	Pedro Almodóvar	13
Jim Jarmusch	Jim Jarmusch	11
M. Night Shyamalan	M. Night Shyamalan	11
Lars von Trier	Lars von Trier	10
Kevin Smith	Kevin Smith	10
Richard Linklater	Richard Linklater	10
Tyler Perry	Tyler Perry	10
Quentin Tarantino	Quentin Tarantino	9
Luc Besson	Luc Besson	9
Robert Rodriguez	Robert Rodriguez	9
Mike Leigh	Mike Leigh	8
David Mamet	David Mamet	8
Blake Edwards	Blake Edwards	8
John Hughes	John Hughes	8
Paul Thomas Anderson	Paul Thomas Anderson	8
Spike Lee	Spike Lee	8
Gregg Araki	Gregg Araki	7
Guillermo del Toro	Guillermo del Toro	7
Noah Baumbach	Noah Baumbach	7
Peter Jackson	Fran Walsh	7
James Gray	James Gray	7
Brian De Palma	Brian De Palma	7
Ethan Coen	Joel Coen	7
Michael Haneke	Michael Haneke	7

```
# Calculate average revenue by rating category

select rating,avg(gross) as avg_gross from movies_cleaned1 group by rating
order by avg_gross desc;
```

rating	avg_gross
G	142043334.8
PG-13	127410847.3
TV-PG	120249753.3
PG	106612932.5
TV-MA	79170782.33
R	42735263.78
Approved	36565280
Not	17849586.1

Rated	
NC-17	10763242.83
X	8485984.333
TV-14	5756185
Unrated	1674056.689

# Count movies by final status and release year

```
select year,final_status,count(*) as movie_count from movies_cleaned1 group by year,final_status order by year;
```

year	final_status	movie_count
1980	Blockbuster	17
1980	Flop	18
1980	Hit	32
1980	Superhit	13
1981	Blockbuster	17
1981	Flop	37
1981	Hit	37
1981	Superhit	12
1982	Blockbuster	10
1982	Flop	49
1982	Hit	50
1982	Superhit	9
1983	Blockbuster	13
1983	Flop	51
1983	Hit	46
1983	Superhit	17
1984	Blockbuster	22
1984	Flop	77
1984	Hit	49
1984	Superhit	7
1985	Blockbuster	11
1985	Flop	92
1985	Hit	55
1985	Superhit	19
1986	Blockbuster	12

# Average gross revenue by year

```
select year,avg(gross) as avg_gross from movies_cleaned1 group by year order by avg_gross;
```

year	avg_gross
1986	20019111.67
1985	20855984.65
1983	21609771.36
1987	21672549.56
1984	22720424.6
1981	24355118.57
1988	25342333.15
1982	27013362.86
1980	31044465.19
1991	32061709.81
1989	32691169.22
1990	35993231.75
1992	38796625.82
1993	41220356.24
1994	45772352.29
1995	48200764.16
1996	49591913.03
1997	55943485.13
1998	57608366.8
1999	69148341.41
2000	69630845.96
2001	79033122.74
2002	85017182.6
2005	89333049.49
2003	89484636.06
2006	93739047.05
2004	94060219.06
2007	102427354.5
2009	102972032.2
2008	109056090.5
2010	117272336.6
2011	124812496.7
2012	127829678.6
2013	129979243.4
2014	132457412.8
2015	136257062.8
2018	141127270.4
2017	143135938.4
2016	145417057.2
2019	148886144.1
2020	198601289.3

# Average gross and budget by genre

```
select genre,avg(gross) as avg_gross,avg(budget) as avg_budget from
movies_cleaned1 group by genre order by avg_gross desc;
```

genre	avg_gross	avg_budget
Animation	241356722.4	67482386.71
Family	215787647.6	26415000
Action	142703026.8	52762798
Adventure	109558732.7	41714675.39
Mystery	101183527.7	30012500.05
Biography	48311948.67	24464659.13
Horror	47836758.47	14158460.53
Comedy	44526755.91	21016372.49
Crime	39766271.11	23555268.81
Fantasy	39251573.02	16885714.29
Drama	37667156.96	22985250.56
Sci-Fi	32561233.25	39437500
Thriller	26935259.42	13058333.33
Romance	23549374.88	24837500
Western	10675295.33	7666666.667
Musical	2217255	350000
Sport	1067629	6000000
Music	110014	8600000

# Calculate ROI for each director

```
select director,avg((gross-budget)/budget) as avg_roi from movies_cleaned1
group by director order by avg_roi desc limit 25;
```

director	avg_roi
Oren Peli	12889.3867
Daniel Myrick	4142.985
Frank Zuniga	2221.2417
Tony Chan	699.513
Lionel C. Martin	559.3747
Travis Cluff	428.6441
Jonathan Prince	386.9891
Jerry Paris	198.96045
Wolfgang Becker	157.6339
Yu Yang	120.044
Niall Johnson	108.9813
Ali Abbas Zafar	99.7465
John Pogue	88.1758

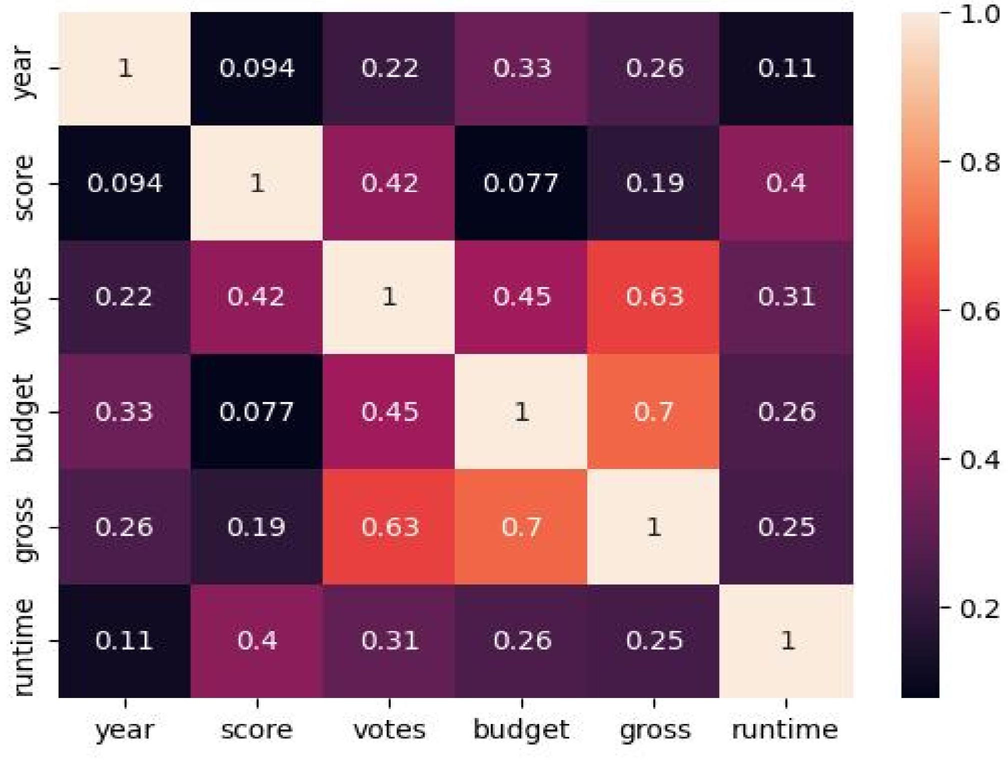
Norman René	87.06715
Aneesh Chaganty	84.7523
Andy Wolk	83.3872
Jazz Boon	81.9017
Richard Pryor	80.428
Levan Gabriadze	61.8821
Chris Kentis	57.81545
Robert J. Rosenthal	55.3259
John Carney	48.3751
Lawrence Bassoff	46.75195
Robert C. Ramirez	45.5061
Shane Carruth	43.83145

# Calculate what is the average gross of each Movie Era

```
select Movie_Era,avg(gross) from movies_cleaned1 group by Movie_Era;
```

Movie_Era	avg(gross)
Classic	24407204.62
90s Hits	47477244.31
Early 2000s	91551711.44
Modern	135005539.5

HEAT MAP



Overall Conclusion

- **Blockbusters:**Achive high revenue with relatively low budgets
- **Superhits:**Generate the highest revenue but often require larger budgets.
- **Hits:**Deliver reasonable returns on mid-level budgets.
- **Flops**:Exhibit low revenue in comparison to high budgets, leading to financial losses.
- The most successful rating categories for blockbuster movies are R, PG-13, and PG, highlighting that both mature and family-friendly movies have high revenue potential. These findings suggest that aiming for broad appeal (PG-13 and PG) or mature audiences (R) can increase the likelihood of a movie becoming a blockbuster.
- Lower frequency ratings such as Not Rated, G, NC-17, Unrated, TV-MA, and TV-14 have relatively few blockbuster films. This suggests that films with more restrictive ratings (NC-17) are less likely to achieve blockbuster status.
- Comedy movies have the highest count of 2182 movies and from that 181 are blockbusters.also Drama(106 blockbusters from 1438 movies) and Action(95 blockbusters from 1438 movies)movies are highly represented.
- Genres such as Romance and Sci-Fi have limited representation in blockbuster status, suggesting that they may cater to more specific audiences and typically do not achieve high commercial success
- By analyzing the dataset, M. Night Shyamalan's moviesandand Woody Alle's got more bluckbusters movies(4 blockbuster from 11 Movies).But we can say that the most successful director-writer combo was Peter Jackson - J.R.R Tolkien, they take 3 movies together and all 3 movies was blockbusters
- 1984 has get more number of blockbusters and Years like 1985 show the highest count of flops, suggesting industry challenges or audience shifts during those times.
- Oren Peli: Average ROI of approximately 12889.39%—a remarkably high return, indicating that his films were extremely profitable compared to their budgets.