ITA5007
Digital Assignment 1

Visualizing the 'IMDb Top 250 Movies' data

Lhingnunching L
22MBS0007

WHY THIS DATASET

Analyzing and visualizing the IMDb Top 250 Movies dataset is of significant interest within the domains of data mining and business intelligence. This dataset provides a wealth of insights into highly acclaimed films, rendering it an engaging subject for exploration. The dataset offers the opportunity to uncover hidden patterns and trends in factors contributing to a film's success, such as genre preferences and directorial styles. Through effective data visualization, we can obtain a comprehensive view of the components of what makes a Blockbuster movie as well as gain insights about the evolution of the movie industry.

The "IMDb Top 250 Movies" dataset was retrieved from IMDb Top 250 Movies

Python Notebook uploaded on GitHub: GitHub

ABOUT THE DATASET

The dataset consists of the following columns:

ranking: The position of the movie in the IMDb Top 250 list.

movieTitle: The title of the movie.

movieYear: The year in which the movie was released.

rating: The IMDb user rating for the movie.

voteCount: The number of user votes received by the movie.

censorRating: The rating assigned to the movie by the relevant censorship authority.

movieLength: The length of the movie in minutes.

runtime: The duration of the movie in hours and minutes.

genre: The genre(s) to which the movie belongs.

movieMonth: The month of the movie's release.

releaseDate: The specific date of the movie's release.

summary: A brief summary or plot synopsis of the movie.

starList: A list of the movie's main cast members.

writerList: A list of the movie's writers.

director: The director(s) of the movie.

country: The country or countries where the movie was produced.

language: The primary language(s) spoken in the movie.

budget: The estimated budget for producing the movie.

gross_worldwide: The worldwide gross revenue generated by the movie.

production: The production company or companies involved in making the movie.

url: The IMDb URL of the movie for reference and additional details.

Snippet of the dataset:

```
df.head()
```

	ranking	movieTitle	movieYear	rating	voteCount	censorRating	movieLength	runtime	genre	movieMonth	...	summary	starList	
0	1	The Shawshank Redemption	1994	9.3	24,08,140	R	2h 22min	142.0	Drama	10	...	Two imprisoned men bond over a number of years...	Tim Robbins,Morgan Freeman,Bob Gunton	
1	2	The Godfather	1972	9.2	16,66,445	R	2h 55min	175.0	Crime,Drama	3	...	An organized crime dynasty's aging patriarch t...	Marlon Brando,Al Pacino,James Caan	
2	3	The Godfather: Part II	1974	9.0	11,57,841	R	3h 22min	202.0	Crime,Drama	12	...	The early life and career of Vito Corleone in ...	Al Pacino,Robert De Niro,Robert Duvall	
3	4	The Dark Knight	2008	9.0	23,67,334	PG-13	2h 32min	152.0	Action,Crime,Drama	7	...	When the menace known as the Joker wreaks havo...	Christian Bale,Heath Ledger,Aaron Eckhart	Nol
4	5	12 Angry Men	1957	9.0	7,09,744	Approved	1h 36min	96.0	Crime,Drama	4	...	A jury holdout attempts to prevent a miscarria...	Henry Fonda,Lee J. Cobb,Martin Balsam	

Figure 1: Snippet of the 'IMDb Top 250 Movies' dataset

VISUALIZING THE DATA

1. Movies released each year

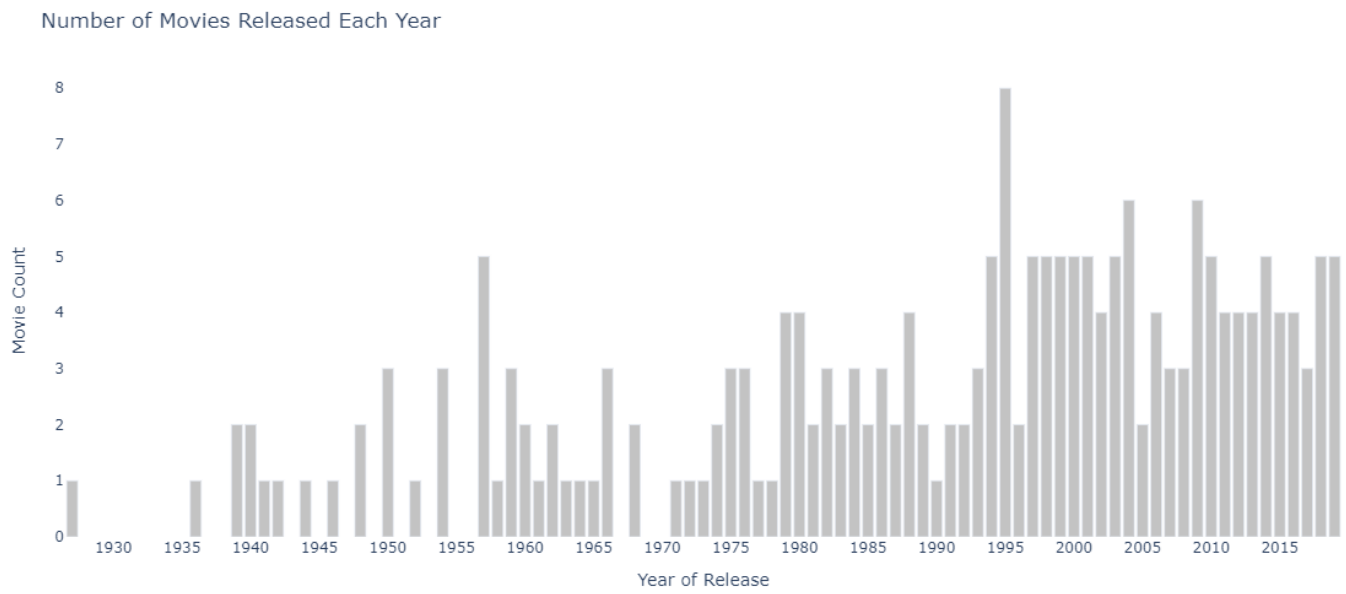


Figure 2: Bar graph of the Year of release of the movies

The year 1995 stands out as particularly significant, with the highest movie count of 8. Another intriguing observation from the bar graph is the overall trend in the number of movies in the IMDb Top 250 as the years progress. The data reveals a clear upward trajectory, suggesting a consistent increase in the number of highly-rated movies over time.

It can also be observed that there have been years in which no movies from that particular year made it to the IMDb Top 250 list. Such gaps in certain years serve as a reminder of the high standards and competitive nature of the IMDb Top 250.

2. Correlation between the numerical variables

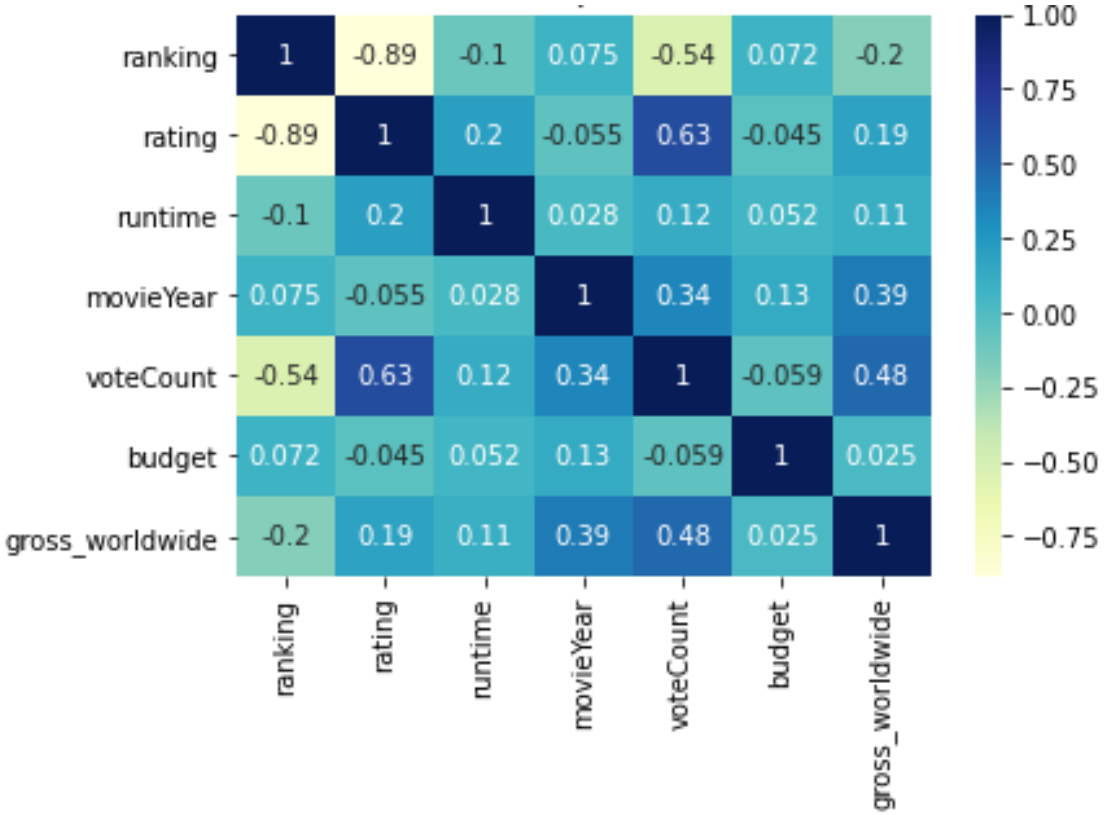


Figure 3: Correlation Heatmap of the numerical variables

The most notable correlation is the strong negative correlation between "rating" and "ranking." This indicates that as a movie's IMDb rating increases, its ranking tends to become smaller (towards 1). This finding suggests that movies rated highly by users often occupy top positions in the IMDb Top 250 list.

Another strong positive correlation was identified between "rating" and "voteCount." This implies that movies with higher vote counts tend to be higher in rating, indicating that popular and well-received films tend to garner more attention and engagement from the IMDb user community.

A strong negative correlation was observed between "ranking" and "voteCount." This implies that movies with lower ranking (towards 1) tend to have more votes.

Additionally, positive correlations were found between "movieYear" and both "grossWorldwide" and "voteCount,". These correlations suggest that more recent movies tend to have higher worldwide gross earnings and receive more

votes, indicating that contemporary films are generally more commercially successful and attract greater audience participation. It was also found that "voteCount" and "grossWorldwide" are positively related.

However, the correlations between the remaining variables were not very significant, suggesting that factors such as budget, production company, and movie length may have more complex relationships with other variables.

3. Genre: Distribution, Ratings and Profits

Table 1: Table of Genre-wise Ratings and Profits

	genre	count	Median Rating	Mean Rating	Median Profit	Mean Profit
0	Action	38	8.4	8.376316	307,710,151	334,597,726
1	Adventure	51	8.3	8.323529	341,311,890	387,801,353
2	Animation	19	8.2	8.278947	336,475,245	263,094,367
3	Biography	25	8.2	8.244	78,681,574	104,136,123
4	Comedy	33	8.2	8.251515	145,192,267	177,341,366
5	Crime	45	8.3	8.362222	53,611,975	115,455,241
6	Drama	152	8.3	8.321711	39,547,202	129,212,483
7	Family	9	8.3	8.333333	26,850,727	107,246,724
8	Fantasy	13	8.2	8.330769	26,850,727	252,219,183
9	Film-Noir	2	8.35	8.35	-1,182,500	-1,182,500
10	History	13	8.1	8.215385	59,427,638	101,635,489
11	Horror	4	8.4	8.35	29,616,704	39,787,746
12	Music	4	8.35	8.35	52,036,044	46,613,976
13	Musical	1	8.3	8.3	-675,744	-675,744
14	Mystery	25	8.3	8.296	31,047,078	80,372,850
15	Romance	18	8.3	8.288889	10,627,785	-20,357,452
16	Sci-Fi	19	8.3	8.331579	112,560,248	334,024,951
17	Sport	6	8.1	8.15	18,575,237	56,544,508
18	Thriller	28	8.25	8.278571	33,634,574	63,496,000
19	War	17	8.3	8.294118	16,357,676	74,973,256
20	Western	6	8.3	8.383333	19,226,876	85,289,066

The genre-wise analysis table presents a comprehensive view of movies categorized by genre. The mean and median ratings within each genre shed light on audience preferences, highlighting genres with higher mean and median ratings as particularly well-received by IMDb users. In addition, the mean and median profit figures offer insights into the financial performance of movies within each genre, guiding decisions on content creation and investment allocation.

3.1. Distribution of Genre

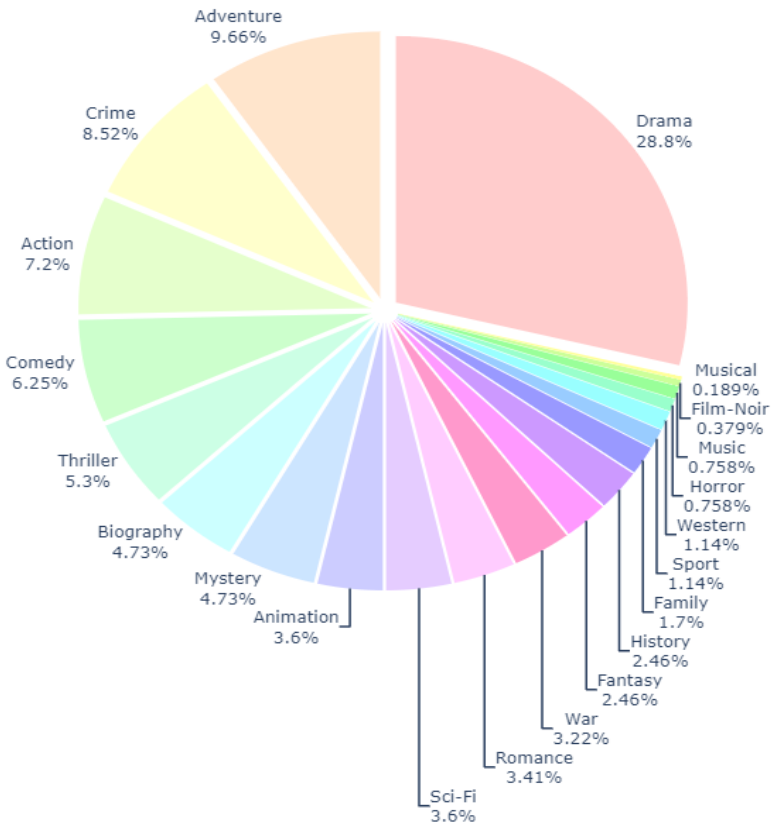


Figure 4: Pie-chart depicting the distribution of the genres

Notably, drama emerges as the dominant genre, constituting approximately 29% of the total distribution. Adventure and Crime genres follow closely behind, each making up approximately 10% and 9% of the dataset, respectively. Conversely, the Musical genre occupies the smallest portion, at around 0.2%, reflecting its relatively limited representation within the IMDb Top 250.

3.2. Profits Trends by Genre

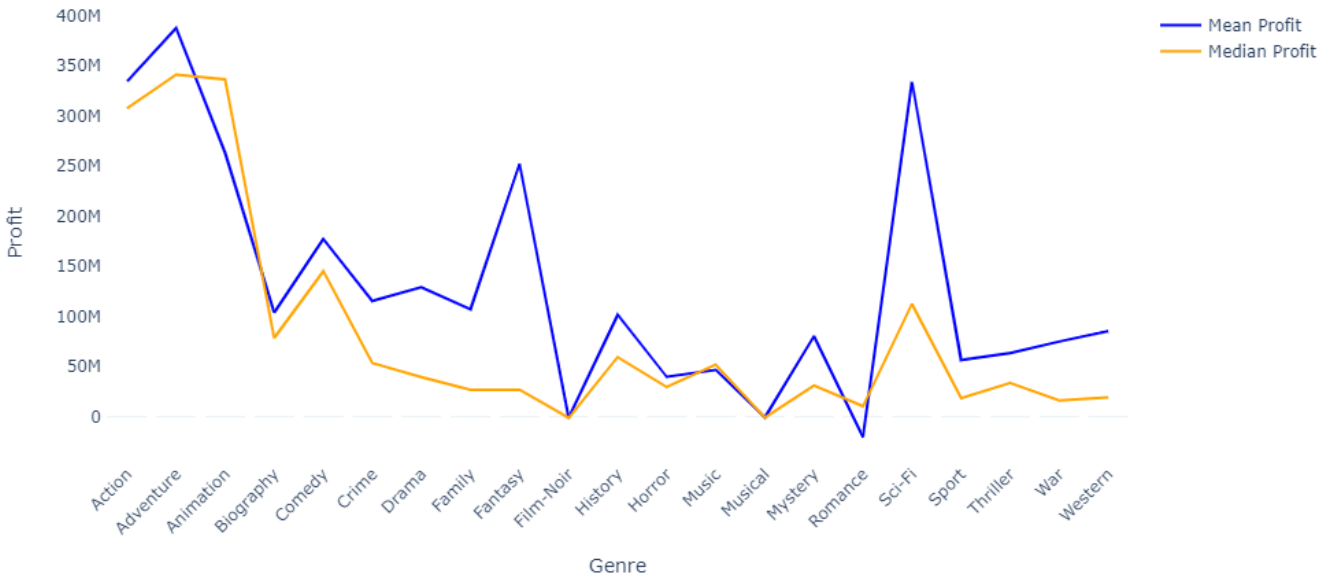


Figure 5: Genre-wise profit trends

The genre 'Adventure' emerges as the most financially lucrative, showcasing the highest peak for mean profits. Following closely behind are the genres 'Sci-Fi' and 'Fantasy,' which also exhibit notable peaks in mean profits, indicating their commercial appeal.

Conversely, genres with negative mean profits, such as 'Romance,' 'Film-Noir,' and 'Musical,' occupy the lower end of the chart. This suggests that while these genres may produce critically acclaimed films, they face challenges in achieving consistent financial success.

3.3. Genre-wise IMDb Ratings

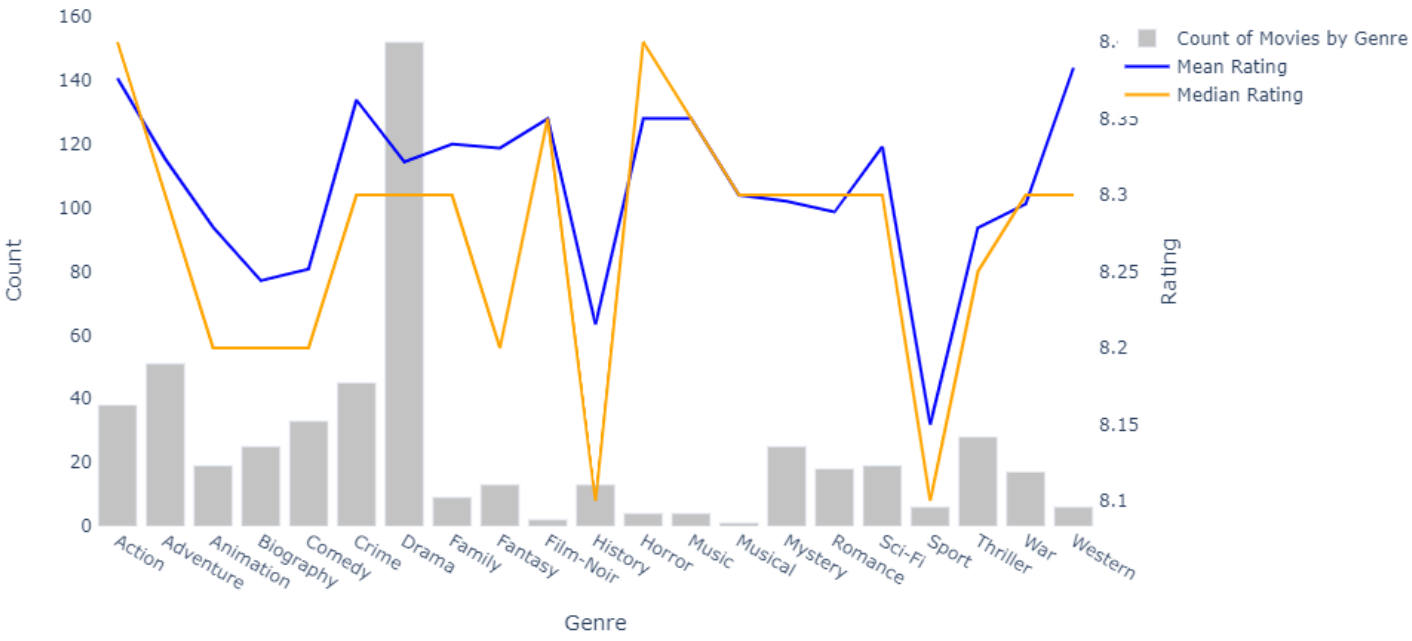
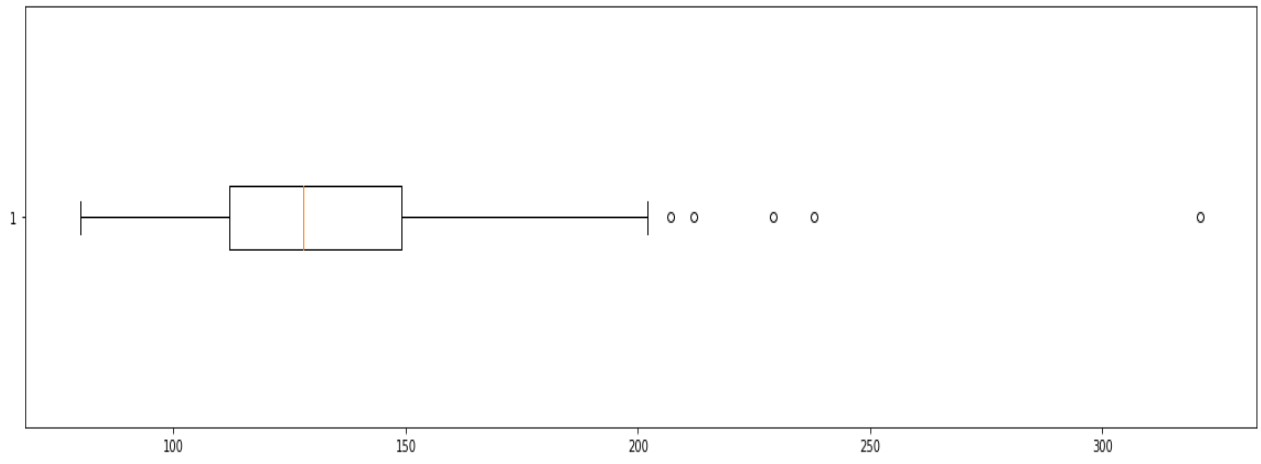


Figure 6: Count of movies by genre with trendline for IMDb ratings

The line chart, featuring dual trend lines showcasing both mean and median ratings across various genres, provides a compelling perspective on audience reception within the IMDb Top 250 movies. The genre 'Western' stands out with the highest peak for mean ratings. Following closely behind are the genres 'Action' and 'Crime,' both of which exhibit substantial peaks in mean ratings.

We can also observe that the mean and median ratings for all genres fall within a relatively narrow range of 8.1 to 8.4. This indicates a consistent level of audience appreciation across diverse genres represented in the IMDb Top 250 list. While the trendlines for the mean rating and median rating tread closely for most genres, there is a stark difference between the peaks of the trendlines for 'History' and 'Fantasy' genres.

4. 5-Number Summary of Runtime



count	mean	std	min	25%	50%	75%	max
209.0	132.14832535885168	32.00709070432944	80.0	112.0	128.0	149.0	321.0

Figure 7: Boxplot with 5 Number Summary of Runtime

The central line inside the box represents the median, which is 128 minutes, indicating that half of the movies fall below this duration, while the other half extend beyond it. The lines on the sides of the box represent the interquartile range (IQR), encapsulating the middle 50% of movie runtimes.

The whiskers extend to a maximum of 1.5 times the IQR, covering the bulk of the data points. Notably, the mean runtime of 132.15 minutes serves as a central reference point, with the data distribution slightly skewed to the right. While the majority of movies cluster around this mean, the presence of outliers, extending to a maximum of 321 minutes, indicates the existence of longer-duration films in the dataset.

5. Actor who appeared in the most number of movies

```
The actor who appeared in the most movies in the dataset is: Robert De Niro
```

Among the extensive list of actors, Robert De Niro emerges as the standout artist who has appeared in the highest number of movies. With a wide-ranging portfolio encompassing various genres and characters, Robert De Niro's enduring presence in cinema is a testament to his remarkable talent and enduring appeal to both audiences and filmmakers alike.

6. Writer who worked on the most number of movies

```
The writer who worked on the most movies in the dataset is: Quentin Tarantino
```

Quentin Tarantino emerges as the preeminent writer with the most extensive body of work among all writers in the film industry. His prolific career and distinctive storytelling style have left an indelible. With a diverse range of screenplays spanning various genres, Quentin Tarantino's creative genius and unique narratives have captivated audiences worldwide.

7. Directors: Most movies, Top-rated

7.1. Director with most movies

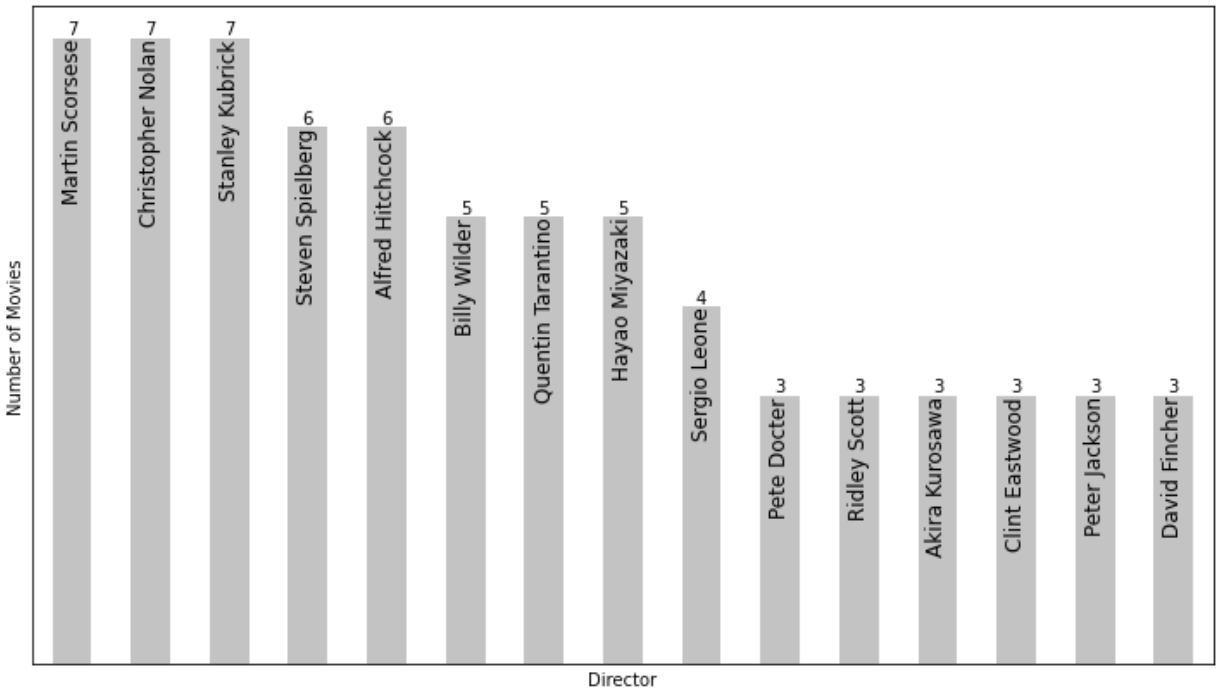


Figure 8: Top 15 directors with most number of movies

Scorsese, Christopher Nolan, and Stanley Kubrick each boast an impressive seven movies to their names, showcasing their enduring contributions to the world of cinema. Following closely behind, with six movies apiece, are the legendary Steven Spielberg and Alfred Hitchcock, further emphasizing their prolific careers and enduring influence on the film industry. This chart not only celebrates the remarkable achievements of these directors but also underscores their lasting impact on the cinematic landscape. In total there are 15 directors with more than 3 movies featuring in the list.

7.2. Top rated directors

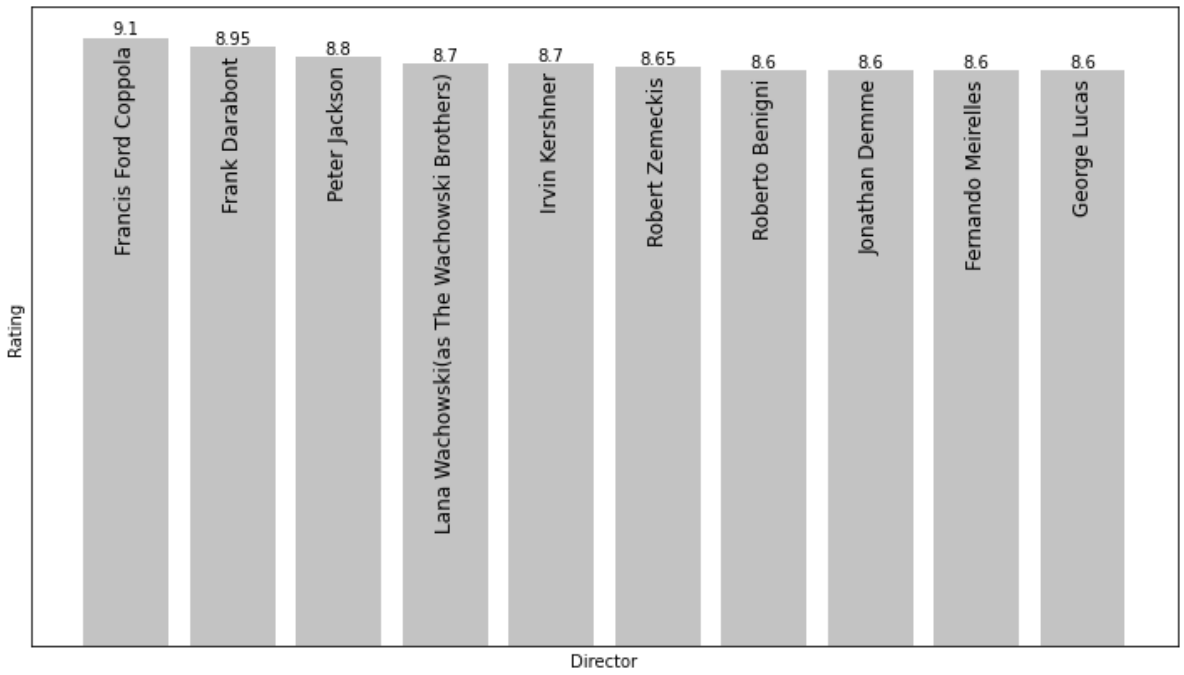


Figure 9: Top rated Directors

Topping the list with a stellar rating of 9.1 is Francis Ford Coppola, a director celebrated for his iconic contributions to cinema. Hot on his heels, Frank Darabont closely follows with an impressive rating of 8.95, showcasing the enduring appeal of his directorial prowess. The chart further underscores the exceptional talents of the remaining directors, all of whom have garnered IMDb ratings ranging from 8.6 to 8.8. Their work continues to be celebrated by audiences and critics alike, and their presence on this chart solidifies their status as true luminaries in the art of filmmaking.

8. Censor Rating: Count, IMDb Ratings

Figure 10: Count of movies by Censor rating with IMDb rating trendline

The 'PG-13' category stands out with the highest peak, suggesting that movies with this rating tend to have both strong mean and median IMDb ratings, reflecting their popularity and critical acclaim. Following closely behind are movies with the 'Approved' rating, indicating that this category also consistently garners favourable ratings. The 'R' rating, while not far behind, displays a slightly lower peak in ratings compared to the aforementioned categories.

It is also noteworthy that the 'R' rating category has the highest number of movies, totalling 105, highlighting its prevalence in the dataset. The 'PG' rating, with 31 occurrences, is the second most frequent censor rating category. This dual perspective, combining censor rating frequency with IMDb rating, suggests that while 'R' rated movies are the most common, 'PG-13' rated movies tend to have the highest mean and median IMDb ratings, making them a prominent category in terms of both quantity and quality within the IMDb Top 250 movies.

9. Top Languages: Number of Movies, Mean IMDb Rating

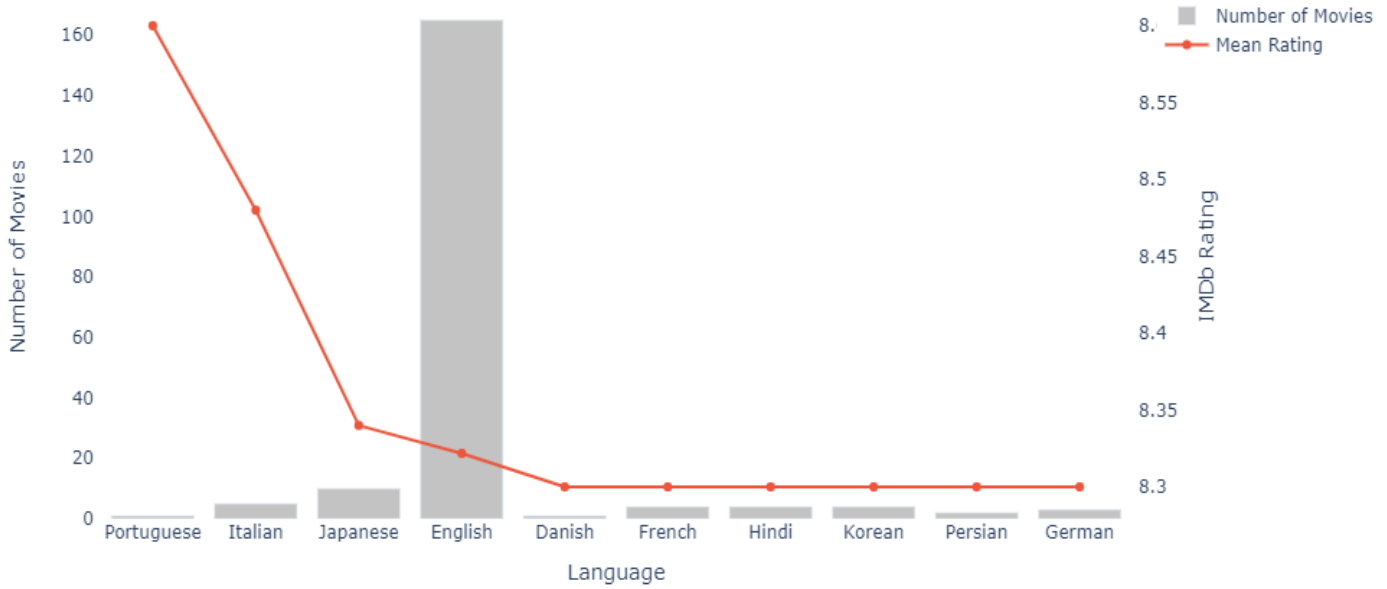


Figure 11: Top Languages by Movie count and IMDb rating

Unsurprisingly, 'English' dominates with the highest count of 165 movies, reflecting its prominence in global cinema. However, it is intriguing to note that numerous other languages, such as Portuguese, Italian, Japanese, Hindi, Persian, and more, contribute to this rich cinematic tapestry.

Despite the relatively low count of only one movie, 'Portuguese' secures the top spot in terms of mean IMDb rating with an impressive 8.6. Following closely, 'Italian' garners a commendable 8.48 mean rating. Notably, these languages, despite their limited representation, manage to achieve high IMDb ratings. The rest of the languages fall within a narrow range of IMDb ratings, from 8.3 to 8.34. This analysis emphasizes that language is not a barrier to achieving cinematic excellence and audience appreciation.

10. Top Countries: Number of Movies, Mean IMDb Rating

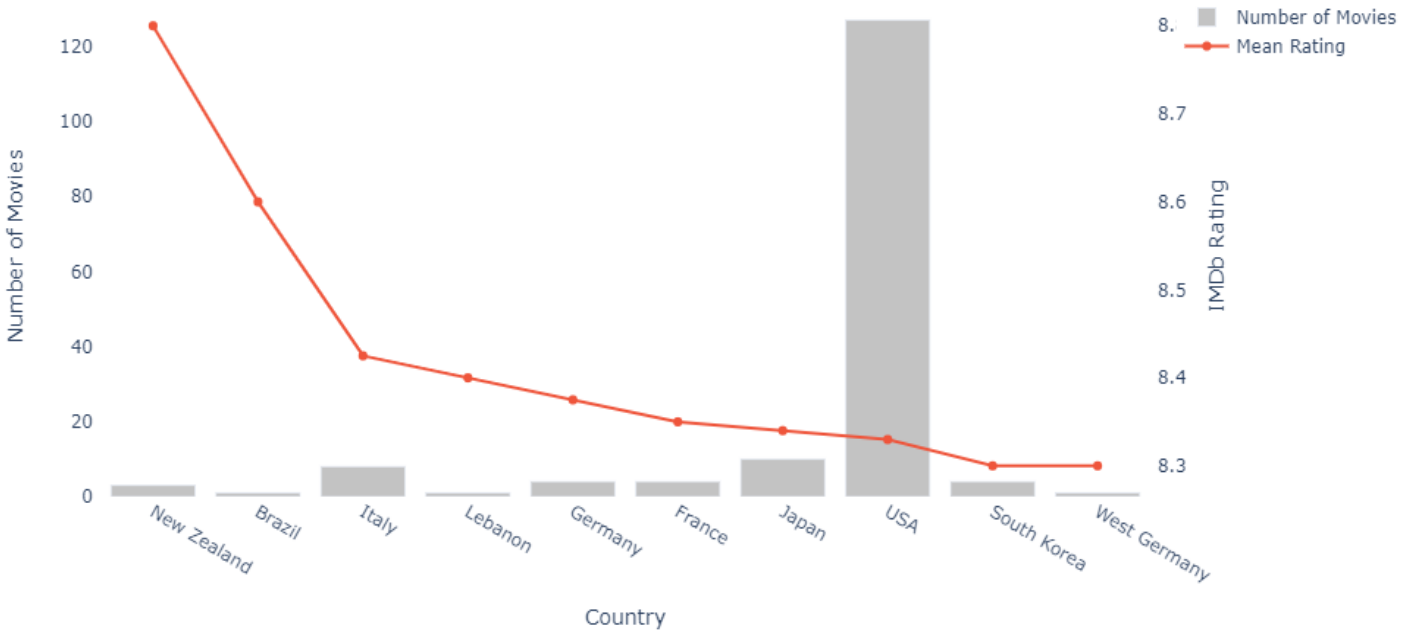


Figure 12: Top Countries by Movie count and IMDb rating

'USA' dominates with the highest count of 127 movies, reaffirming its position as a cinematic powerhouse. However, what truly stands out are the notable ratings achieved by countries with lower movie counts. Despite having only two movies, 'New Zealand' secures the top spot with an outstanding mean IMDb rating of 8.8. 'Brazil,' with just one movie, follows closely with an impressive rating of 8.6.

The remaining countries, ranging from 'Lebanon' and 'West Germany' with one movie each to 'Japan' with ten movies, consistently fall within a relatively tight range of IMDb ratings, from 8.3 to 8.425. This suggests that quality cinema knows no geographical boundaries, and these countries have managed to create movies that resonate with global audiences.

11. 20 Most Profitable Movies

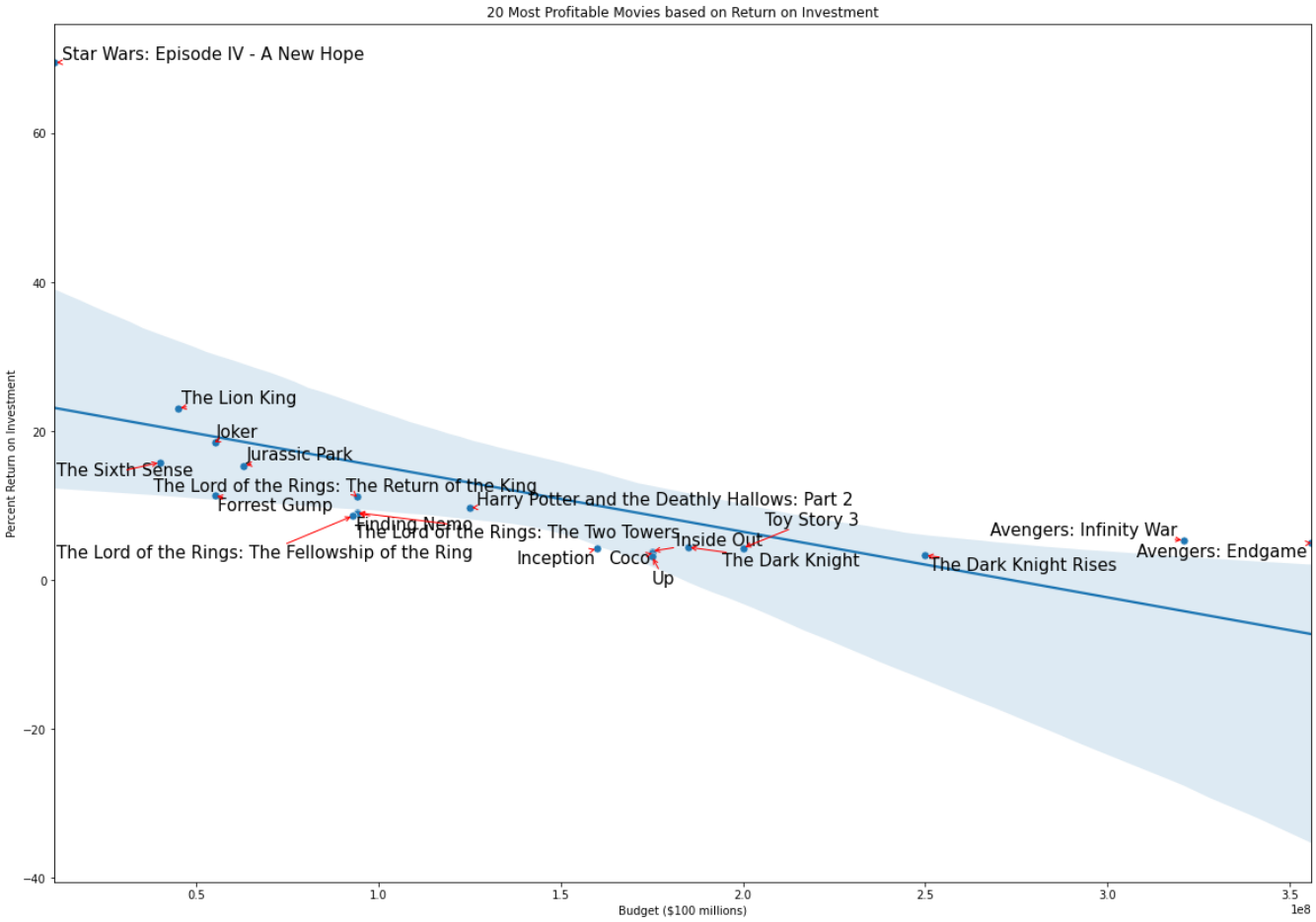


Figure 13: 20 most profitable movies based on Return on Investment

In the analysis of the "20 Most Profitable Movies based on Return on Investment, one standout observation is the significant disparity between the movie "Star Wars: Episode IV - A New Hope" and the other films within the top 20 list. "Star Wars: Episode IV - A New Hope" exhibits an astonishingly high Return on Investment (ROI) of approximately ~7000%. The mean ROI was found to be ~767% with standard deviation ~1419%.

Furthermore, the visualization highlights that higher budget does not necessarily mean higher ROI. This insight suggests that while a significant budget can contribute to a film's success, it is not the sole determinant.

INFERENCES

Year Trends: The year 1995 stands out with the highest movie count. An upward trend in the number of IMDb Top 250 movies over time is evident. Some years have no movies in the list, underscoring IMDb's high standards.

Correlation Insights: Strong negative correlation between rating and ranking (-0.89) indicates highly-rated movies secure top positions. Other correlations are less significant.

Genre Diversity: Drama dominates at approximately 29%, while Musical has the smallest share at 0.2%.

Financial Success: Adventure, Sci-Fi, and Fantasy genres lead in mean profits. Some genres like Romance and Musical face negative mean profits.

Genres with High Ratings: Western, Action, and Crime genres exhibit the highest mean ratings. All genres fall within a narrow rating range of 8.1 to 8.4.

Movie Runtime: Median runtime is 128 minutes. Runtime varies across movies.

Prolific Actor: Robert De Niro has the most movie appearances.

Prolific Writer: Quentin Tarantino is the most prolific writer.

Prolific Directors: Directors like Scorsese, Nolan, and Kubrick have seven movies each. Spielberg and Hitchcock follow with six each.

Top Rated Directors: Francis Ford Coppola and Frank Darabont lead with exceptional ratings of 9.1 and 8.95 respectively.

Censor Rating Impact: 'PG-13' has the highest peak, 'R' has the most movies. 'PG-13' tends to have the highest mean and median IMDb ratings.

Language Influence: 'English' dominates with 165 movies. 'Portuguese' and 'Italian' stand out with high mean ratings despite low counts.

Global Cinema: 'New Zealand' impresses with an outstanding mean rating of 8.8. 'Brazil' also excels (8.6) despite low representation. Other countries maintain ratings between 8.3 to 8.425.

Profit Disparity: "Star Wars: Episode IV - A New Hope" exhibits an exceptionally high ROI of ~7000% (mean ROI is just ~769%), emphasizing that budget doesn't always correlate with ROI.

These observations showcase the IMDb Top 250's dynamic nature, highlighting trends in movies, genres, ratings, and the diverse contributions of actors, writers, and directors.

Credits:
Lhingnunching L
22MBS0007

Fin.