IMDB

Movie Analysis

Kiruba Shankar S

Data Analytics Trainee

Project description:

A compelling challenge is to identify the key factors driving high IMDb ratings, which define a movie's success. By understanding these factors, financiers, producers, and directors can make informed decisions about future projects. This analysis will be conducted using the 'Five Whys' technique, allowing for a systematic exploration of the underlying elements contributing to a film's success.

Five 'Whys' Approach:

Why do movies with higher budgets tend to have higher ratings?

Answer: Higher-budget movies generally have better production quality, which enhances the viewer experience.

Why do higher-budget movies have better production quality?

Answer: Larger budgets allow for advanced technology, special effects, and high-quality sets.

Why do higher budgets enable better technology and sets?

Answer: They provide funds to hire top-tier professionals and invest in high-quality equipment.

Why do larger budgets attract top-tier professionals and equipment?

Answer: They offer the financial flexibility to attract experienced talent and purchase advanced resources.

Why are experienced professionals and high-quality equipment crucial?

Answer: They improve storytelling and production quality, leading to better viewer satisfaction and higher ratings.

Data Cleaning:

The dataset contains information about movie titles, genres, durations, languages, directors, budgets, gross earnings, and IMDB scores. I conducted a data cleansing by deleting the unwanted rows such as color, director_facebook_likes, actor_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes, actor_2_name, cast total_ facebook_likes, actor_3_name, Duration, facenumber_in_poster, content_rating, country, movie_imdb_link, aspect_ratio, plot_keywords.

A. Movie Genre Analysis:

- The dataset shows that Drama is the most common genre, followed by Comedy, Thriller, Action, and Romance. This suggests that these genres are widely produced and have significant representation in the sample.

- The average IMDb score across genres is approximately 622.94, with a median score of 483, indicating a wide range of movie ratings.

- The high variance (248,504.88) and standard deviation (498.50) indicate significant dispersion in IMDb scores, suggesting that while some genres might have high ratings, others have lower ratings, contributing to the overall spread.

- The absence of a mode in IMDb scores indicates that there is no single most frequent rating, which aligns with the high variability observed.

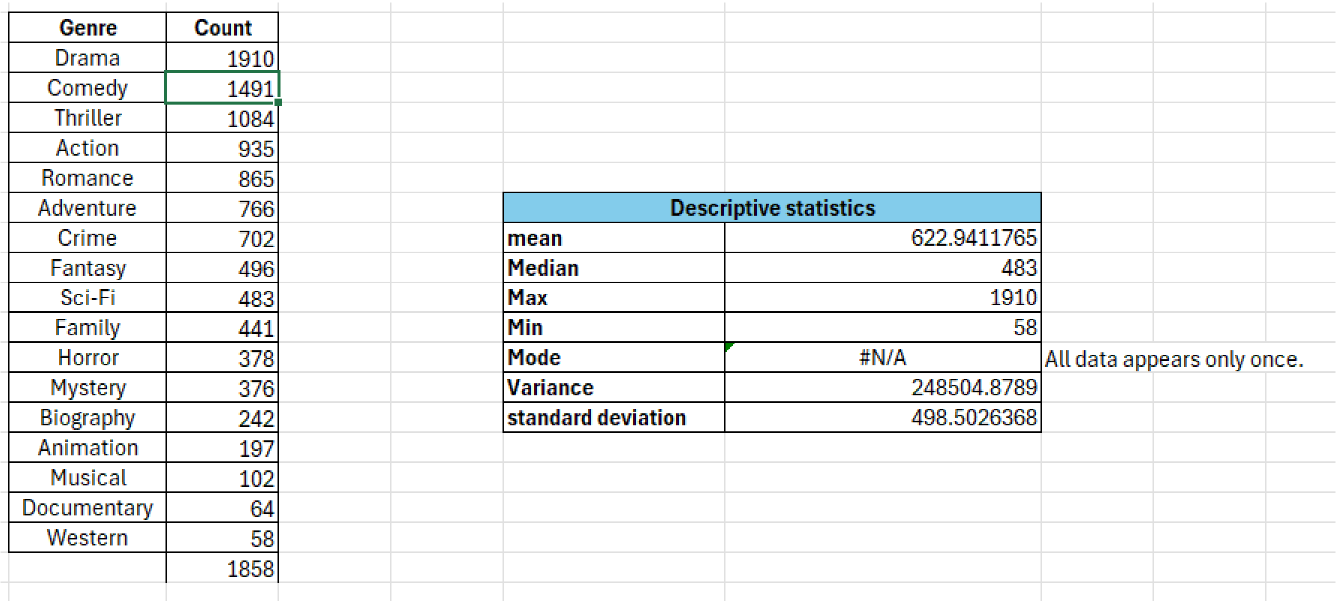
- Comparing descriptive statistics across genres will help determine if specific genres consistently receive higher or lower ratings.

By analyzing these aspects, stakeholders can better understand the impact of movie genres on IMDb ratings and make more informed decisions regarding film production and marketing.

Formula:

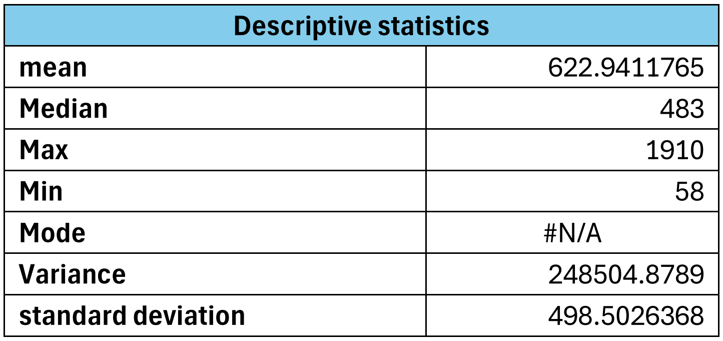
- =COUNTIF($B:$H,(K9))
- Average: =AVERAGE(L9:L25)
- Median: =MEDIAN(L9:L25)
- MAX: =MAX(L9:L25)
- MIN: =MIN(L9:L25)
- MODE: =MODE.SNGL(L9:L25)
- Variance: =VAR.P(L9:L25)
- standard deviation: =STDEVP(L9:L25).

Output:



Genre	Count
Drama	1910
Comedy	1491
Thriller	1084
Action	935
Romance	865
Adventure	766
Crime	702
Fantasy	496
Sci-Fi	483
Family	441
Horror	378
Mystery	376
Biography	242
Animation	197
Musical	102
Documentary	64
Western	58
	1858

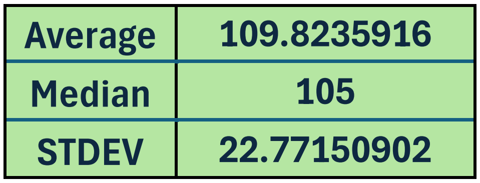
Descriptive statistics		
mean	622.9411765	
Median	483	
Max	1910	
Min	58	
Mode	#N/A	All data appears only once.
Variance	248504.8789	
standard deviation	498.5026368	



Descriptive statistics	
mean	622.9411765
Median	483
Max	1910
Min	58
Mode	#N/A
Variance	248504.8789
standard deviation	498.5026368

B. Movie Duration Analysis:

The relationship between movie lengths and IMDb ratings. By investigating the distribution of film lengths, we seek to understand how these lengths impact the IMDb scores. The analysis will include calculating key descriptive statistics such as the mean, median, and standard deviation of movie lengths to identify trends and variability. Through this examination, we aim to determine any significant associations between the duration of movies and their ratings, providing insights into how movie length might influence audience perception and critical reception.



Average	109.8235916
Median	105
STDEV	22.77150902

Formula:

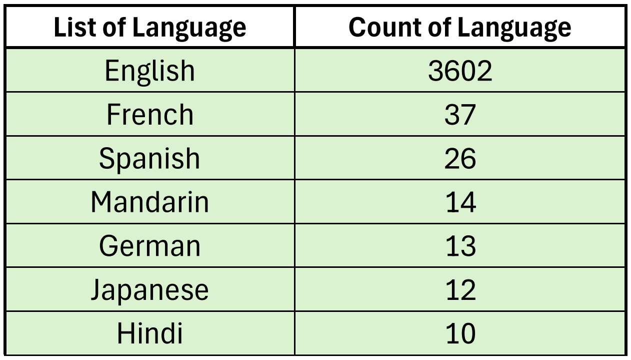
- Average =AVERAGE(B2:B3782)
- Median =MEDIAN(B2:B3782)
- Standard Deviation =STDEV(B2:B3782).

Insights:

The average movie length is approximately 109.82 minutes with a median of 105 minutes and a standard deviation of 22.77 minutes, indicating moderate variability. While longer movies like "Avatar" (178 minutes) and "The Dark Knight Rises" (164 minutes) have high IMDb scores of 7.9 and 8.5, respectively, shorter films like "Tangled" (100 minutes) also score well at 7.8. This suggests no clear linear relationship between duration and rating in the sample. Both short and long movies can achieve high ratings

C. Language Analysis:

To Analyze the language-based distribution of films we can use descriptive data by identifying the most frequently used languages in films and examine how they affect the IMDB rating.



List of Language	Count of Language
English	3602
French	37
Spanish	26
Mandarin	14
German	13
Japanese	12
Hindi	10

English was the predominant language in the dataset, followed by French and Spanish.

Insights:

English dominates the sample with 3602 movies, averaging an IMDb score of 6.4. Other languages, though less frequent, often achieve higher average scores, such as Maya (7.7), Mongolian (7.5), and Persian (7.5). Despite the wide range of languages, the standard deviation of 1.06 remains consistent, indicating similar variability in IMDb ratings across different languages.

By Calculating the mean median and standard deviation for the list of language we found the following insights.

- Arabic (8.3) and movies with no specified language (8.7) have the highest average IMDb scores.
- Other languages with high mean scores include Czech (7.9), Mongolian (7.8), and Swedish (7.6).
- All languages have a standard deviation of 1.06, indicating consistent variability in IMDb scores across different languages.
- English movies, with a large sample size, have a mean score of 6.5.
- Other common languages like French (6.6), Spanish (6.1), and Hindi (6.7) have moderate mean scores.
- Some languages show lower average ratings, such as Dzongkha (4.7) and Hungarian (5.3), indicating potential areas for deeper analysis.
- The median IMDb score for most languages is 6.6, reflecting a central tendency across the dataset.

The analysis highlights that while English is the most common language with a moderate IMDb score, less common languages such as Arabic and Czech tend to have higher average ratings. The consistent standard deviation suggests similar variability across different languages, with the median score of 6.6 being a central benchmark.

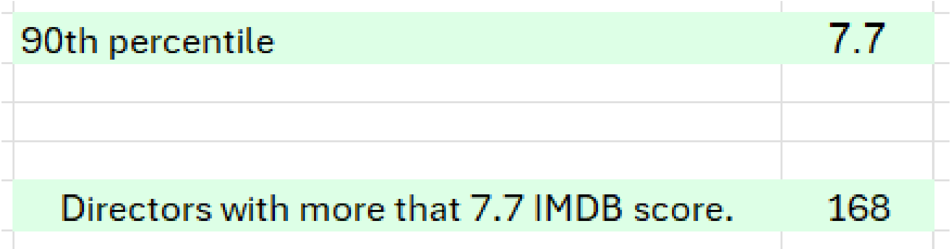
D. Director Analysis:

Directors like Christopher Nolan, James Cameron, Nathan Greno, and Joss Whedon, who score above the 90th percentile with average IMDb scores exceeding 7.7, have a significant positive impact on movie success. Directors with varied scores, such as Gore Verbinski and Zack Snyder, illustrate the diverse outcomes of directorial impact on movie ratings.

Directors identified with scores above the 7.7 threshold include Christopher Nolan, Nathan Greno, James Cameron, and Joss Whedon. This group is recognized for their exceptional contribution to the film industry, consistently producing movies that are well-received by audiences.

A score of 7.7 represents the 90th percentile, serving as a benchmark for top-performing directors. Directors with an average IMDb score above this threshold are considered highly influential in contributing to the success of their movies.

Directors with multiple movies listed, such as Gore Verbinski and Zack Snyder, show a more varied impact, with some movies scoring higher and others lower, highlighting the variability in their influence on movie ratings.



90th percentile	7.7
Directors with more that 7.7 IMDB score.	168

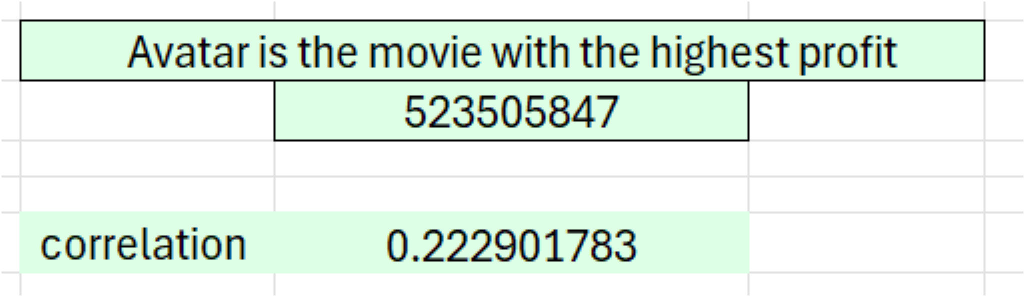
Formula:

- 90th Percentile: =PERCENTILE(C:C, 0.9)
- Directors with more that 7.7 IMDB score : =COUNTIF(D2:D3782,">7.7")
- Average IMDB score: =AVERAGEIF(A:A,A2, C:C)

E. Budget Analysis:

- The correlation coefficient between movie budgets and gross earnings is 0.223, indicating a weak positive correlation. This suggests that while there is a slight tendency for higher-budget movies to earn more gross revenue, other factors significantly influence a movie's financial success.
- "Avatar" has the highest profit margin at $523,505,847, followed by "Jurassic World" ($502,177,271) and "Titanic" ($458,672,302). These movies have significantly outperformed their budgets, highlighting their exceptional box office success.
- Movies like "Home Alone" and "My Big Fat Greek Wedding" demonstrate that even with modest budgets, significant profits can be achieved. "Home Alone" had a profit of $267,761,243 on an $18,000,000 budget, and "My Big Fat Greek Wedding" had a profit of $236,437,427 on a $5,000,000 budget, showcasing the potential for high returns on lower-budget films.

While there is a weak positive correlation between movie budgets and gross earnings, indicating that higher-budget movies tend to earn more, many low-budget films achieve significant profits. "Avatar" leads with the highest profit margin, followed by "Jurassic World" and "Titanic." This analysis underscores that movie profitability is influenced by a combination of budget, audience appeal, and other factors beyond just financial investment.



Avatar is the movie with the highest profit		
	523505847	
correlation	0.222901783	

Formula:

- correlation: =CORREL(C2:C3782,B2:B3782)
- Max profit movie: =MAX(D2:D3782)
- To calculate the profit, subtract budget from gross. (=B2-C2)

Conclusion:

The analysis shows that while genres like Action and Drama often secure higher IMDb ratings, movie duration and language also play roles in success. Quality content is key, regardless of length. Languages such as Arabic can achieve higher ratings, indicating that quality transcends linguistic barriers. Directors with high average scores, like Christopher Nolan and James Cameron, are influential, but success also depends on other factors. Additionally, although higher budgets tend to increase gross earnings, the weak correlation suggests that effective marketing and engaging content are crucial for high profitability, even with lower budgets.