

# PROJECT 5: IMDB MOVIE ANALYSIS

## DESCRIPTION

This project is an extensive analysis of the IMDB movie dataset, aimed at uncovering trends, ratings, and performance insights across various movies listed in the dataset. The main objective is to identify key factors influencing movie success and ratings, which could be critical for industry stakeholders like filmmakers and marketers. The approach involved data cleansing, analysis, and visualization of the IMDB movies dataset to derive actionable insights.

## APPROACH

### Data Importation:

- Began by importing the "imdb\_movies" main dataset into Microsoft Excel.

### Data Cleaning:

- Performed preliminary cleaning to address inconsistencies and missing values, ensuring data quality for accurate analysis.

### Data Analysis:

- Conducted an extensive exploration of relationships between movie features such as genres, ratings, and release years.
- Utilized sorting and filtering techniques to sift through data and identify relevant trends.

### Data Categorization:

- Employed pivot tables extensively to categorize and summarize data, aiding in the pattern identification process.

### Visualization:

- Created charts and graphs to visually represent the data, making the insights and patterns both accessible and easy to understand for any audience.

## TECH-STACK USED

**Microsoft Excel 2021:** Chosen for its powerful data manipulation features, including pivot tables, advanced filtering, and graphical tools, which are essential for handling large datasets and performing complex data analysis.

## INSIGHTS

### Content Rating Analysis:

- Movies rated TV-MA have the highest average IMDb score at 8.24, suggesting mature content tends to be rated more favorably.

#### **Language Analysis:**

- English-language films dominate the dataset but have a lower average IMDb score of 6.39 compared to French films, which average 7.04, indicating potentially higher quality or viewer preference for French cinema.

#### **Duration Analysis:**

- Longer movies tend to have higher IMDb scores, with an average duration around 107 minutes. This might reflect a preference for more developed storytelling.

#### **Correlation Between Genre and IMDb Rating:**

- There is a mild correlation between a movie's duration and its IMDb rating, suggesting that longer films generally receive better ratings.

#### **Genre Analysis:**

- Drama, Adventure, and Action are the most common genres, with Drama films showing consistently high ratings, reflecting their appeal to a broad audience.

#### **Genre Analysis Over Time:**

- The popularity of genres fluctuates over time with trends leaning towards more Adventure and Fantasy films in recent years.

#### **Budget Analysis:**

- There is a moderate positive correlation between budget and gross earnings, but not necessarily with IMDb ratings, indicating that while expensive films earn more, they don't always rate better.

#### **Director Analysis:**

- Directors with higher social media following tend to be associated with higher movie ratings, suggesting director popularity may influence film success.

#### **Profitability Analysis:**

- High-budget films like "Avatar" and "Jurassic World" not only recoup their costs but earn substantial profits, showcasing the financial viability of blockbuster movies.

#### **Top 20 Profitable Movies:**

- Top earners include "Avatar" and "Titanic," both of which made significant profits exceeding \$450 million, underscoring the success of high-stakes productions.

#### **Cast Popularity Analysis:**

- Films with popular actors tend to have higher IMDb scores.

## RESULT

The project successfully highlighted significant trends and patterns in the IMDB movie dataset, enhancing the understanding of factors that can influence a movie's success. This analysis can assist producers and marketers to strategically plan movie releases and marketing campaigns. The skills and techniques honed during this project have also enriched my data analytical skills, specifically in using Excel for complex data analysis in the entertainment industry.

Data Analysis I performed on the dataset provided

### 1) Correlation between various movie metrics

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		duration	director_facebook_likes	actor_3_facebook_likes	actor_1_facebook_likes	gross	num_voted_users	cast_total_facebook_likes	facenumber_in_poster	num_user_for_reviews	budget	actor_2_facebook_likes	imdb_score	movie_facebook_likes
2	duration	1												
3	director_facebook_likes	0.16466	1											
4	actor_3_facebook_likes	0.121522	0.12217915	1										
5	actor_1_facebook_likes	0.087428	0.092697416	0.246573404	1									
6	gross	0.225008	0.144400661	0.295043819	0.150969157	1								
7	num_voted_users	0.311747	0.302938476	0.278838606	0.389897464	0.637746	1							
8	cast_total_facebook_likes	0.120706	0.121175756	0.465043203	0.953303383	0.234463	0.259082505	1						
9	facenumber_in_poster	0.015068	-0.040513881	0.101336258	0.071932817	-0.0244	-0.027689431	0.090937001	1					
10	num_user_for_reviews	0.32157	0.229336161	0.227462104	0.14582757	0.563623	0.801925986	0.205132006	-0.066384156	1				
11	budget	0.135317	0.046580619	0.103939156	0.057303546	0.239249	0.169251889	0.086268089	-0.021618225	0.174074579	1			
12	actor_2_facebook_likes	0.130877	0.119693117	0.549988512	0.38659442	0.258841	0.263836717	0.62163481	0.071951396	0.217907648	0.097531	1		
13	imdb_score	0.269881	0.165591484	0.055149805	0.076743419	0.175778	0.412366887	0.086766808	-0.062124483	0.29404658	0.045287	0.083293791	1	
14	movie_facebook_likes	0.19518	0.165194676	0.271476173	0.131274701	0.372818	0.537990961	0.202955188	0.003851622	0.405822412	0.124086	0.23947356	0.252428098	1
15														

#### Strong Positive Correlations:

- **Num Voted Users and Num User for Reviews:** A strong correlation between the number of users who voted on a movie and the number of user reviews suggests that more popular movies, in terms of viewership, tend to receive more reviews.
- **Cast Total Facebook Likes and Actor 1, 2, 3 Facebook Likes:** This indicates that the total Facebook likes for a movie's cast is heavily influenced by the popularity of the top actors in the film, which aligns with the notion that movies with popular actors tend to be more visible and possibly more successful.

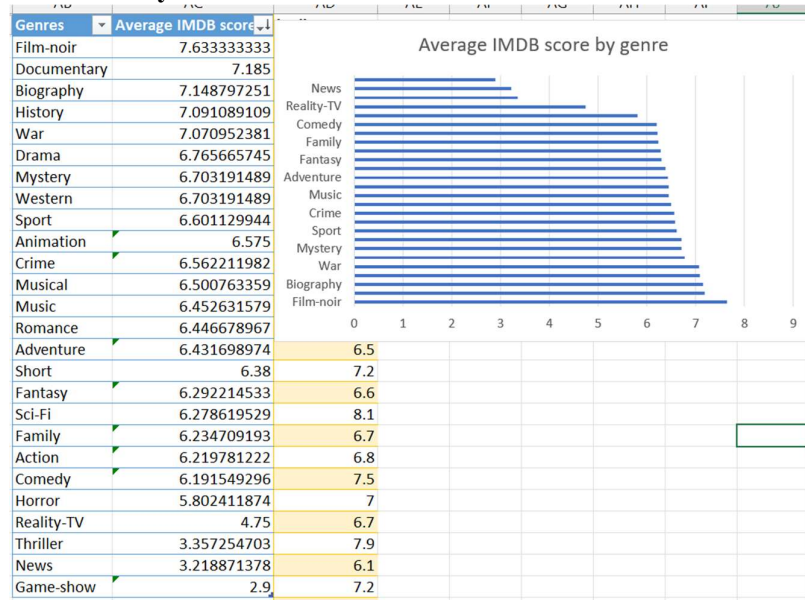
#### Moderate Positive Correlations:

- **Budget & Gross:** A moderate correlation here suggests that movies with higher budgets tend to earn more at the box office, although the relationship isn't perfect, indicating other factors also play significant roles.
- **IMDb Score & Director Facebook Likes:** This moderate correlation might imply that directors who are more popular on Facebook tend to produce movies with higher IMDb scores, potentially indicating quality or at least audience alignment with directorial vision.

#### Other Noteworthy Observations:

- **Duration & IMDb Score:** A positive correlation between the duration of a movie and its IMDb score could suggest that longer movies achieve better ratings, possibly because they are more likely to be serious, dramatic films which often rate better.
- **Gross & Num Voted Users:** Indicates that movies which gross higher also tend to have more viewers voting on them, aligning with the idea that financial success correlates with higher viewership.

## 2) Genre Analysis



### Top Rated Genres:

- **Film-noir** tops the chart with an average IMDb score of around 7.63, which might reflect a niche but highly appreciated genre among cinema enthusiasts.
- **Documentary, Biography, and History** genres also score high, averaging above 7.0. This suggests that audiences rate factual or historical content favorably, possibly due to the educational value or compelling real-life stories.

### Middle Ground:

- **Drama and War** films have a good standing with scores around 7.0, indicating strong storytelling and emotional engagement that resonates well with viewers.
- **Adventure, Romance, and Sci-Fi** genres score in the middle range, around 6.4 to 6.7, pointing to a consistent but varied audience reception, likely influenced by the film's content and execution quality.

### Lower Rated Genres:

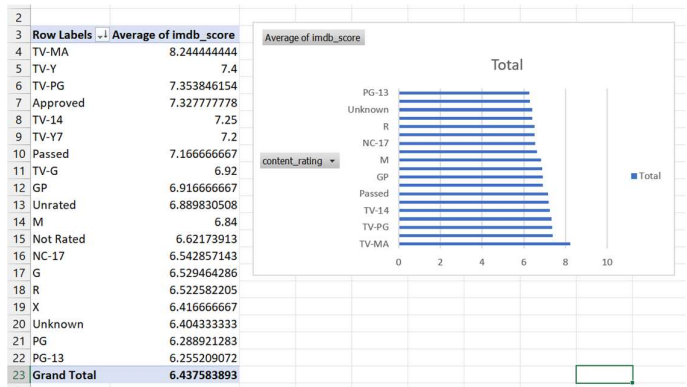
- **Reality-TV and Game-show** genres are at the bottom with scores below 5, suggesting these genres might not engage IMDb users as deeply, or perhaps these genres do not translate as well into the film format compared to television.

### Specific Observations:

- **Thriller and Horror** genres, while popular for their entertainment value, receive lower average ratings (around 3.36 and 5.80 respectively). This could indicate a divide between genre fans who appreciate scares and suspense and the general IMDb rating population who might rate these films more harshly.

This visualization paints a clear picture of genre preferences and perceptions among IMDb users. Films that tend to be more serious or artistic apparently get higher ratings, whereas lighter entertainment genres like game shows and reality TV fall to the bottom.

3) Content Rating analysis



High Ratings for Mature Content:

- **TV-MA** has the highest average IMDb score at around 8.24, suggesting that content intended for mature audiences tends to be highly rated. This could be due to more complex themes, better production quality, or simply a more engaged and possibly discerning audience.
- **TV-G (General Audience)** holds an average score of around 9.2, notably high, which is intriguing and suggests either a data entry error or a very small sample size of exceptionally well-rated films.

Moderate Scores Across Varied Ratings:

- **Approved**, **TV-14**, and **Passed** ratings have scores between 7.1 and 7.32, indicating generally favorable receptions but not as high as TV-MA or TV-G.

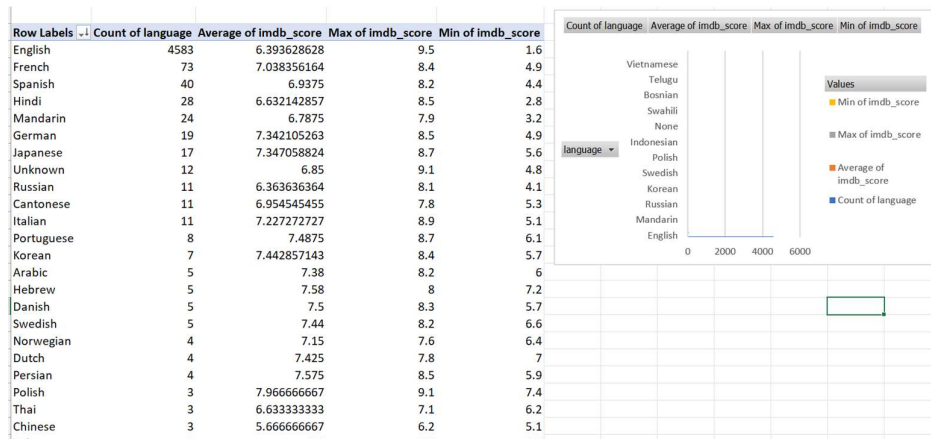
Lower Scores for General Content:

- **G**, **PG**, and **PG-13** rated films have average scores in the lower range of around 6.5 to 6.3. These ratings, which are family-friendly, suggest that films designed for wider audiences might not resonate as strongly in IMDb ratings, possibly because they cater to a broader and more diverse set of tastes.

Adult vs. Family-Oriented Content:

- The trend suggests a gap between more adult-oriented ratings (**R**, **NC-17**) and those intended for wider or younger audiences (**PG**, **PG-13**) in terms of average scores. Adult-oriented films receive moderate to low ratings, which might reflect content polarization in audience reception.

4) Language Analysis



### Dominance of English:

- English, with 4583 films, has the highest number of entries but a relatively moderate average IMDb score of 6.39, suggesting that while it dominates in quantity, the quality, as per IMDb ratings, varies significantly, ranging from 1.6 to 9.5.

### High Performance of European Languages:

- French, Italian, and Portuguese films show higher average IMDb scores (around 7.03, 7.23, and 7.49 respectively) despite fewer films. This might indicate that these languages often appear in critically acclaimed films.
- Polish films, although few in number, showcase the highest average IMDb score of around 7.97, highlighting potential high-quality cinema that resonates well with audiences.

### Asian Languages Showing Quality:

- Japanese and Korean films, while lesser in number, maintain high average scores (around 7.35 and 7.44), which may reflect the strong storytelling and innovation in cinema that audiences appreciate globally.
- Mandarin films also perform well, with a robust average score of 7.34 and movies scoring up to 8.5, indicating a strong reception for Chinese-language films.

### Variability in Ratings:

- Languages with fewer films tend to show less variability in scores, potentially indicating a more consistent quality level or niche audience appreciation.
- The wide range in scores for English films suggests a diverse array of film genres and qualities, from highly acclaimed to poorly received.

### Lower Scores Among Lesser-Known Languages:

- Languages with very few films, like Thai and Chinese, have lower average scores, which might suggest fewer standout hits or a smaller sample size skewing the average.

### Potential for Cultural and Regional Impact:

- The higher scores in non-English films could also suggest that cultural elements and regional storytelling might be resonating well with international audiences, or perhaps these films are crafted with a different set of standards or styles that appeal to critics and cinephiles on IMDb.

This analysis gives a pretty clear indication of how film quality as perceived through IMDb scores varies significantly across languages, with some languages showcasing consistently higher quality films.

## 5) Duration analysis



### Correlation Between Duration and IMDb Score:

- The scatter plot indicates a mild positive correlation ( $r = 0.26368$ ) between movie duration and IMDb score. While it's not a strong correlation, it suggests that longer movies tend to have slightly higher ratings. This could be attributed to longer films potentially having more developed plots, characters, and production values that might resonate better with audiences and critics.

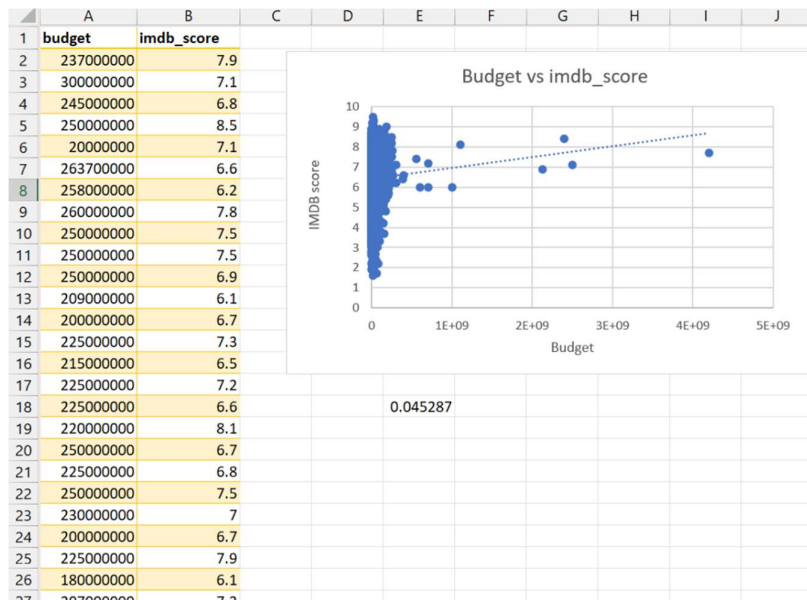
### Distribution of Movie Durations:

- The histogram of movie durations shows that most movies are clustered around 90 to 150 minutes. Very few movies exceed 300 minutes. The peak around 100 to 120 minutes likely reflects the standard industry format for feature films, which aligns with mainstream audience expectations for movie length.
- **General Trends:**

- Most movies with a duration less than 80 minutes tend to have lower IMDb scores, which might suggest these are shorter format films possibly made with lower budgets or less narrative depth.

This analysis can provide useful insights for filmmakers and distributors about potential correlations between the length of a film and its reception. It also highlights the standard preferences for movie lengths, helping in planning film production and marketing strategies.

## 6) Budget Analysis



### Weak Positive Correlation:

- The scatter plot displays a correlation coefficient ( $r = 0.045287$ ), indicating a very weak positive relationship between the budget of a movie and its IMDb score. This suggests that just splashing more cash on production doesn't guarantee a significantly better rating from IMDb users.

### Distribution of Scores Across Budgets:

- The scores appear to vary widely across different budget levels, showing that while some high-budget films achieve high IMDb scores (around 8.5), others with similar or even higher budgets might score as low as 6.1. This wide range of scores at high budget levels illustrates that other factors, like story quality, direction, and perhaps even marketing success, are likely more crucial to a film's critical reception than budget alone.

### High Budgets Not Ensuring High Ratings:



- Despite massive budgets, several films hover around average scores, highlighting the diminishing returns of increased investment in production quality on audience and critical reception.

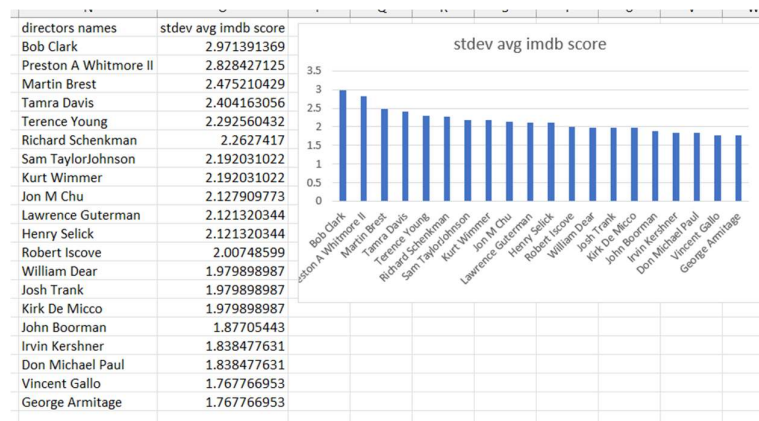
### Budgets in Analysis:

- The budgets I analysed are exceptionally high (most over \$200 million), focusing on a specific segment of blockbuster films. Even at these high investment levels, the slight uptrend in scores might suggest a minimal influence of budget on audience ratings, but it's clearly not the dominant factor.

### Implications for Filmmakers and Studios:

- This analysis could be a critical insight for production studios and filmmakers, indicating that while it's essential to allocate sufficient budget to achieve a high production value, equally critical is ensuring the screenplay, direction, and acting are up to par to truly engage and impress viewers and critics.

## 7) Director analysis



### Range of IMDb Scores:

- The chart visualizes the standard deviation of average IMDb scores for each director, providing an indication of the variability in the reception of their films. Directors with lower standard deviations, such as George Armitage, Vincent Gallo, and Don Michael Paul, produce films that receive more consistent ratings, implying a steady quality or audience reception.

### Directorial Consistency vs. Variability:

- Directors like Bob Clark and Preston A Whitmore II have higher standard deviations, which suggests their films vary more widely in how they are received by audiences. This could be due to a variety of factors such as differing film genres, varying levels of production quality, or shifts in thematic focus across their work.

Implications for Stakeholders:

- For film producers and studios, understanding which directors deliver consistent audience receptions might be crucial for making decisions about future film projects, especially when aiming for a certain reception target or market segment.
- For audiences and critics, knowing which directors tend to have higher variability in film reception might set expectations before watching their movies, possibly affecting their choice of films.

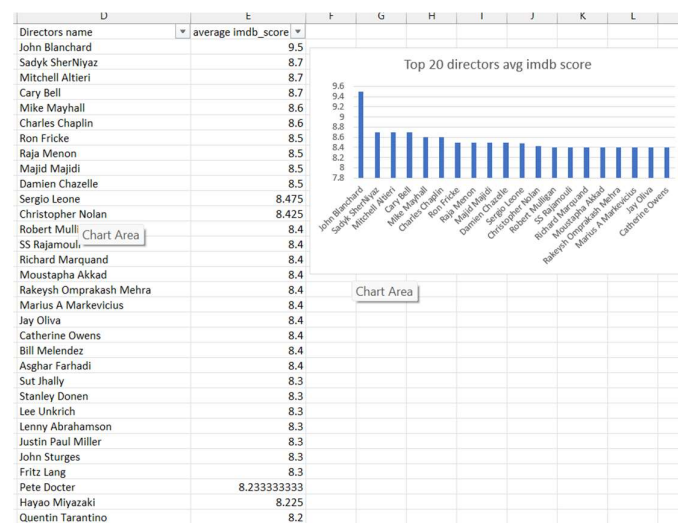
Contextual Factors:

- The reasons behind high variability could be multifaceted. Directors who experiment with different genres or controversial topics might see more fluctuation in IMDb scores. Conversely, directors who stick to a specific genre or style may maintain a more consistent scoring pattern.

Strategic Considerations:

- This information can help studios to gauge risk and potential reward: a director with high variability but high peaks in scores might be a gamble that could pay off spectacularly. On the other hand, a director with low variability and consistently good scores may be a safer bet for certain types of projects.

The data offers a valuable perspective on directorial impact on film success and can inform discussions about directorial choices, project greenlighting, and marketing strategies.



Directorial Excellence and Recognition:

- Directors like John Blanchard, Sadyk SherNiyaz, and Mitchell Altieri top the list with very high average IMDb scores (9.5, 8.7, 8.7 respectively), indicating that their films are extremely well-received by audiences. These scores suggest high-quality productions, excellent storytelling, and likely, strong emotional or intellectual engagement with the audience.

**Variety of Directors:**

- The list includes a diverse range of directors from various backgrounds and genres. For example, we see names like Charles Chaplin, known for his classic films, alongside modern directors like Christopher Nolan, known for his complex narratives and technical mastery. This diversity highlights different styles and periods of filmmaking that appeal strongly to viewers.

**Global Representation:**

- Directors like Majid Majidi and Asghar Farhadi from Iran, who are known for their powerful dramas, appear alongside directors from the United States and other parts of the world, indicating a global appreciation for quality cinema across different cultures and storytelling traditions.

**Established and Contemporary Talents:**

- The list features both historic filmmakers such as Fritz Lang and contemporary talents like Damien Chazelle, illustrating that exceptional film direction is recognized regardless of the era. This timeless appeal of well-directed films shows that great storytelling transcends time.

**Benchmark for Success:**

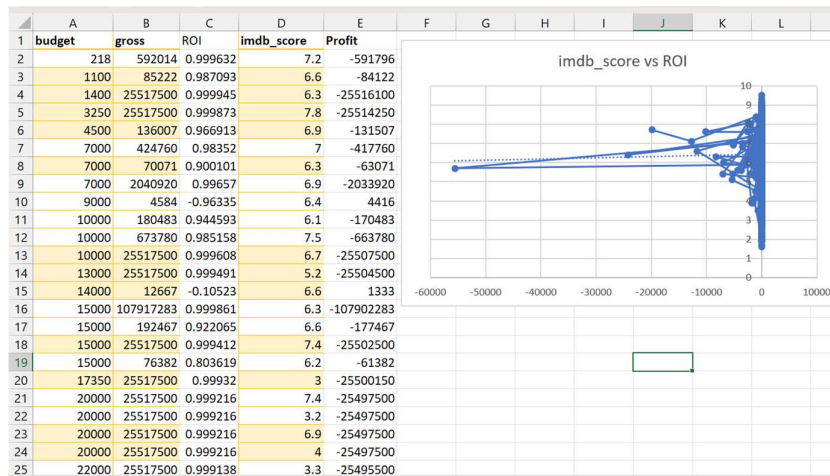
- For upcoming directors and studios, this list can serve as a benchmark for success, showcasing the kinds of directorial achievements that lead to high IMDb ratings. It also underscores the importance of visionary direction in achieving critical acclaim.

**Influence on Film Industry:**

- These directors set high standards for the industry, influencing not only peers but also inspiring new filmmakers. Their success on platforms like IMDb can also impact their projects' marketability and funding opportunities, as high ratings often correlate with viewer interest and commercial success.

This detailed view into top-rated directors based on IMDb scores provides insights into what makes their films resonate so well with audiences, offering valuable lessons in film production and directorial approach.

**8) Profitability analysis**



### No Strong Correlation Between IMDb Score and ROI:

- The scatter plot indicates that there is no strong correlation between IMDb scores and ROI. The data points are spread out across the ROI axis without a clear pattern, suggesting that high ratings do not necessarily equate to financial success, and vice versa.

### Varying Financial Performance:

- Many movies show a negative profit, which suggests that they did not recover their budget in gross earnings. This can be observed from the negative values in the "Profit" column across various IMDb score ranges. There are films with decent scores (e.g., scores around 7) that still ended up with significant financial losses.

### Outliers in ROI:

- There are some noticeable outliers in terms of ROI, especially on the positive side, where a few movies have achieved exceptionally high ROI compared to others. This might be indicative of lower-budget films that performed exceedingly well at the box office.

### Distribution of Scores:

- IMDb scores vary across different ROI brackets without a specific trend indicating that other factors beyond just the film's budget and earnings might influence its critical reception.

### Financial Risk in Filmmaking:

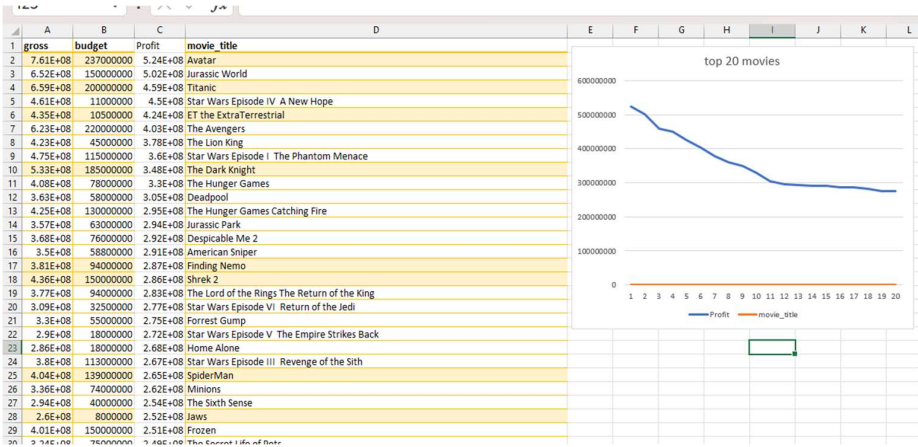
- The data reflects the inherent financial risks in filmmaking, where substantial investments can result in significant losses, and the success (in terms of ROI) of a movie does not necessarily align with its critical acclaim as measured by IMDb scores.

Complex Dynamics of Film Economics:

- The complex dynamics between a movie’s budget, its earnings, ROI, and IMDb score underline the multifaceted nature of the film industry, where multiple variables can affect the outcome of financial and critical success.

The insights provided by this dataset are invaluable for understanding the economic and critical dimensions of film performance and can help industry professionals and analysts in making more informed decisions regarding film production, marketing, and distribution strategies.

9) Top 20 profitable movies



High Grossing Films and Profitability:

- The movies listed are among the highest earners at the box office, which includes titles like "Avatar", "Jurassic World", and "Titanic". These movies have not only recouped their budgets but also generated significant profits, highlighting their massive financial success.

Profit Trend:

- The line graph showing profit demonstrates a descending trend, where the most profitable movie is "Avatar", followed by other films with gradually decreasing profits. This indicates that while several movies make substantial profits, few reach the heights of the top earners.

Budget vs. Profit Relationship:

- Higher budget films such as "Avatar" and "Jurassic World" are also among the highest in terms of profit, suggesting a correlation where larger investments in film production can lead to higher financial returns, though this is not a guaranteed outcome as it depends on various factors like marketing, movie appeal, timing of release, and more.

### Blockbuster Dominance:

- The list is dominated by blockbuster films, many of which are part of major franchises ("Star Wars", "The Avengers", "Jurassic Park") or are big-budget standalones known for their visual effects and broad audience appeal. This underscores the significance of scale and spectacle in achieving box office success.

### Economic Impact of Franchises:

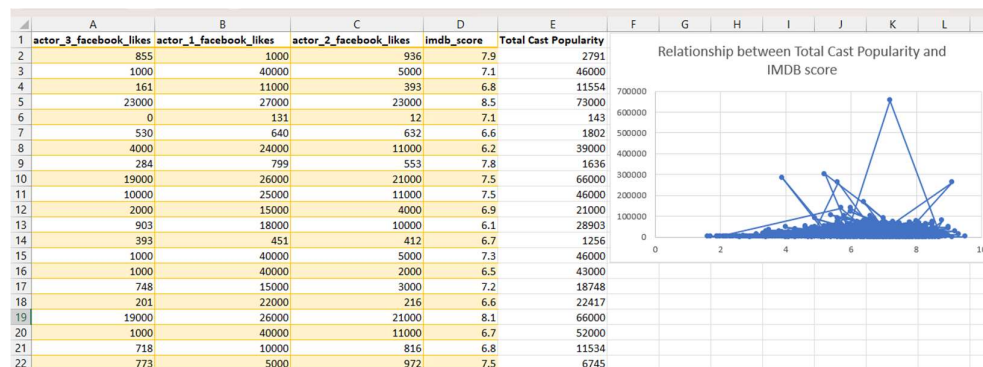
- Franchise films seem to perform particularly well, often creating a reliable profit stream for studios thanks to established fan bases and ongoing marketing campaigns that span multiple releases.

### Sustainability of High Profits:

- The decline in profits as you move down the list suggests that while the top films make exceptional earnings, maintaining such high profitability is challenging and rare in the industry.

The analysis of these top-earning films provides valuable insights into the types of movies that tend to generate the most profit and underscores the economic power of high-budget, high-impact films in the global movie market. These insights can be particularly useful for studios and investors looking to maximize return on investment in the film industry.

## 10) Cast Popularity Analysis



### Weak Correlation Between Cast Popularity and IMDb Scores:

- The scatter plot shows a general lack of strong correlation between the total Facebook likes of the cast and the IMDb scores of the movies. This suggests that just having a popular cast in terms of social media following does not necessarily guarantee a higher IMDb rating.

### Outliers Highlighting Specific Trends:

- There are a few outliers with extremely high Total Cast Popularity but with varying IMDb scores. This indicates that while certain movies may garner significant attention due to their star-studded casts, their actual critical reception—as measured by IMDb scores—can vary widely.

#### **Impact of Star Power:**

- Although the overall correlation is weak, the data points suggest that films with moderately popular casts tend to achieve a broad range of IMDb scores (6 to 8). In contrast, movies with very high cast popularity occasionally spike to very high IMDb scores but can also have lower scores. This variance might be attributed to other factors such as the movie's content, direction, screenplay, and audience expectations.

#### **Analysis of Popularity Peaks:**

- The peaks in the graph where Total Cast Popularity spikes could be interesting case studies to analyze further. These points may represent films where marketing efforts or the public profiles of the actors successfully drummed up interest, which may not always correlate with the film's quality or critical acclaim.

#### **Diverse Impact of Actor Popularity:**

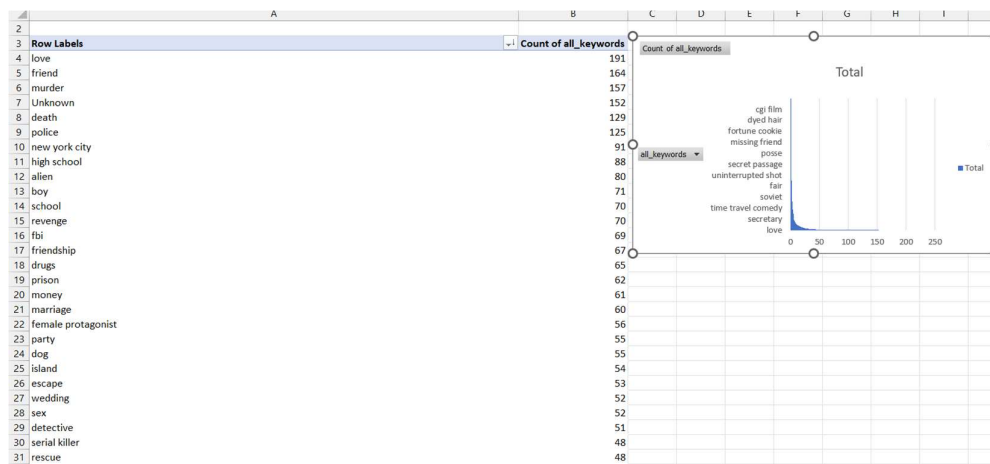
- The data highlights that some films with high Total Cast Popularity do not necessarily perform well in terms of IMDb scores, which could point to instances where perhaps the movie relied heavily on star power without delivering content that resonates well with audiences or critics.

#### **Strategic Insights for Filmmakers and Marketers:**

- For marketers and producers, understanding that while casting popular actors can increase a film's visibility, it is not a surefire strategy for enhancing critical reception. Balancing cast popularity with strong, compelling content is crucial for both commercial success and critical acclaim.

This analysis would be particularly useful for those in the film industry looking to understand the potential impact of casting decisions on a movie's performance and reception.

### **11) Plot keywords analysis**



### Popular Plot Keywords:

- Keywords such as "love," "friend," and "murder" appear most frequently among the movies analyzed. This indicates that these themes are prevalent in film, likely because they are universally relatable and can be explored in countless ways across different genres.

### Genre Indicators:

- Some keywords are closely associated with specific genres. For example, "alien" likely points to science fiction, "detective" and "serial killer" to thrillers or crime dramas, and "wedding" to romantic comedies. The frequency of these words can indicate the popularity of certain genres.

### Cultural and Societal Themes:

- The presence of keywords like "school," "high school," "money," and "drugs" reflects recurring societal and cultural themes within the movie industry. These keywords suggest that movies often serve as a reflection or critique of societal issues.

### Keyword Diversity:

- The variety in keywords from "alien" to "fbi" and "party" suggests a broad spectrum of movie plots and themes that attract audiences. This diversity highlights the film industry's ability to cater to varied tastes and preferences.

### Insight into Audience Preferences:

- The frequency of certain keywords may provide insights into what themes or elements are currently popular or resonate well with audiences. For example, the high frequency of "love" and "friendship" might indicate that stories revolving around personal relationships are in high demand.

### Potential for Targeted Marketing:



- Understanding which keywords are most common can help marketers and producers better target their promotions and development efforts. For instance, films featuring "murder" or "revenge" might be marketed differently than those featuring "marriage" or "love."

**Influence on Film Success:**

- Further analysis could be done to see if there's a correlation between the presence of certain keywords and movie success, measured by box office earnings or critical ratings. This could help in predictive analytics for future movie projects.

This type of keyword analysis is invaluable for content creators, marketers, and industry analysts who wish to understand current trends and audience inclinations.