Data Analysis of IMDB Top 1000 Movies and TV Shows

27 July 2025

Zaio Institute of Technology
Authored by: Sesethu M. Bango

Analysis Overview Report

Background

The film and television/streaming industry is highly revered. Many movies are produced on an annual basis. They entertain, inform and often leave us inspired and in awe. For production companies to remain in business and successful they need to consistently churn out quality movies and shows. Achieving high IMDB Ratings and Meta Scores are vitally important, so is the bottom line, revenue and profits.

By analyzing the relevant data, trends can be found that paint a vivid picture of how to potentially achieve continuously successful ratings and revenue numbers.

Summary of Findings

- The number of missing values in the dataset is a little troubling. 250 rows had to be dropped because the Gross and IMDB_Rating columns had missing values for those rows. Perhaps even more insight could've been gained if this data wasn't missing.
- High IMDB rated movies generally also have high Meta scores.
- Dramas are greatly represented in the overall scope of genres. They are at least 3 times as likely to make up a film or show's genre classification as their nearest competition.
- Films and shows with a high Number of votes do not in any meaningful way correspond to high gross revenue figures.
- There is no one certificate that clearly has higher ratings than the others.

Reflection

- Index resetting. After dropping the null rows, errors were incurred. It took some time to figure out that the indices needed to be reset.
- A 'KeyError' kept arising after the implementation of the 'groupby' method. To resolve this error the relevant dataframe was saved to a CSV file. The file was then re-read into python and analysis continued.

Data Collection

The IMDB dataset was downloaded from Kaggle using a line of code within a Python script. Provided is a link which when clicked will open a webpage where the raw dataset was downloaded;
harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

Table 1 shows the number of missing and None values in the dataset/dataframe. The table shows data before the cleaning of null values and after cleaning. Rows with missing critical data were removed, i.e.; rows with missing Gross and IMDB_Rating data.

Table 1

Check missing values of raw dataset	
Number of missing values:	427
Total number of values in the table:	16000
Proportion of missing data to the total:	2.67% of data is missing
Check missing values after dataset has been cleaned	
Number of missing values:	36
Total number of values in the table:	12000
Proportion of missing data to the total:	0.3% of data is missing

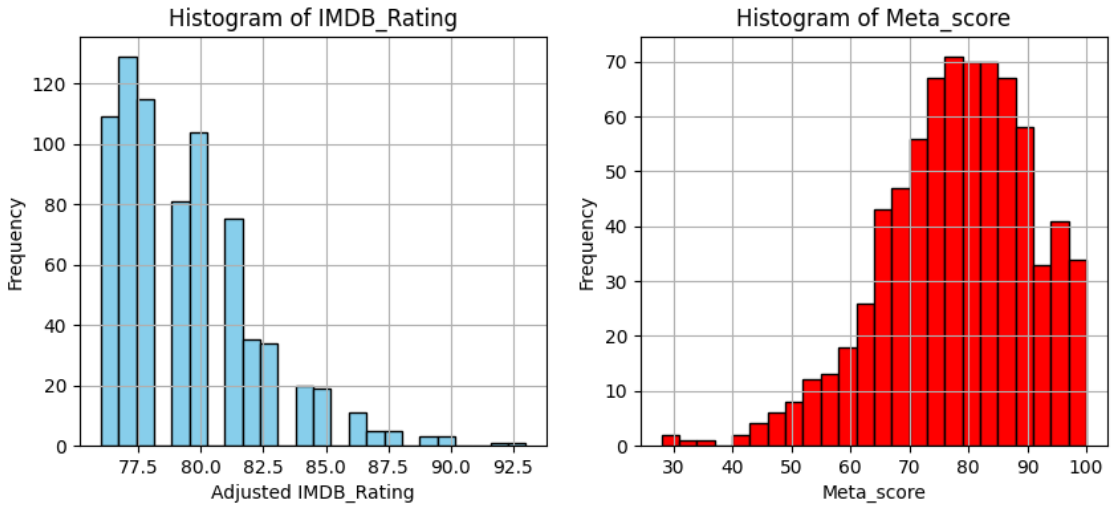
Data Preparation

A program was run to check for and remove duplicate rows. No duplicate rows were found in the dataset.
The column Runtime's data type was changed from object to numerical.
From the Released_Year column, the Decade for each row value was extracted.
The columns Star1, Star2, Star3 and Star4 were combined under a new Lead_Actors column.
These results can be found in the document named Markdown.ipynb.

Data Visualisation

Distribution Analysis:
Figure 1 shows the distribution of movie rating scores in the form of histograms of IMDB_Rating and Meta_score ratings.

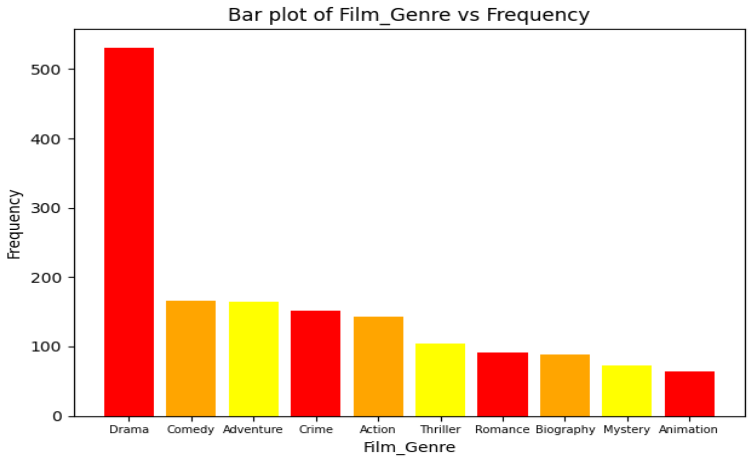
Figure1



On the x-axis is the rating value and on the y-axis the frequency, the number of times a specific score is recorded. The IMDB_rating needed to be adjusted, previously it had a rating out of 10. In order to compare like to like, this number was multiplied by 10, hence each rating value is now a rating out of 100 points.
Comparing these two histograms, they are both very much skewed towards the higher half of the rating score range. With IMDB ratings all being at 75 points or higher, which makes sense as we are dealing with the top 1000 IMDB rated movies.
High IMDB ratings do generally correspond to high Meta scores. The Meta score distribution is more bell shaped, albeit skewed to the right. They both point to consistently high ratings, each having median and mean values of around 80.

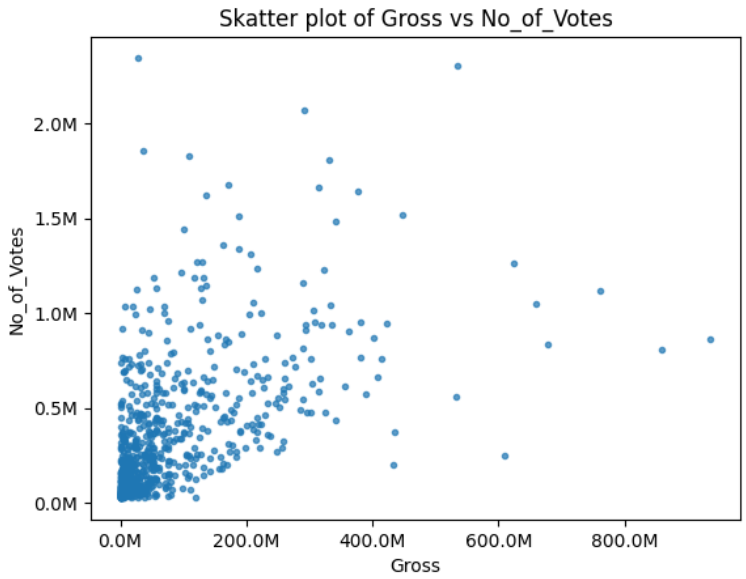
Categorical Analysis:
Figure 2 shows a bar plot of Genre vs Frequency. Very few films are single genre only films, most are categorized as having two or more genres. From the bar plot one can see that dramas lead the other genres by a great margin. Dramas are at least 3 times more prevalent than their nearest competitors. It is often dramas paired up with one or more of the other genres that make up a film's genre classification.

Figure 2



Relationship Analysis:
Figure 3 shows a scatter plot of Gross profit vs Number of votes.
From figure 3 we see that most of the data is compressed into the range of 0 – 300M Gross, and 0 – 1M No. of votes. What's also very clear is the lack of data points in the high Gross, high No. of votes section of the graph. This clearly means that a high vote count does not at all correlate to a high gross revenue. The films that have grossed the highest have a No of votes count of around 1M, which is above average.

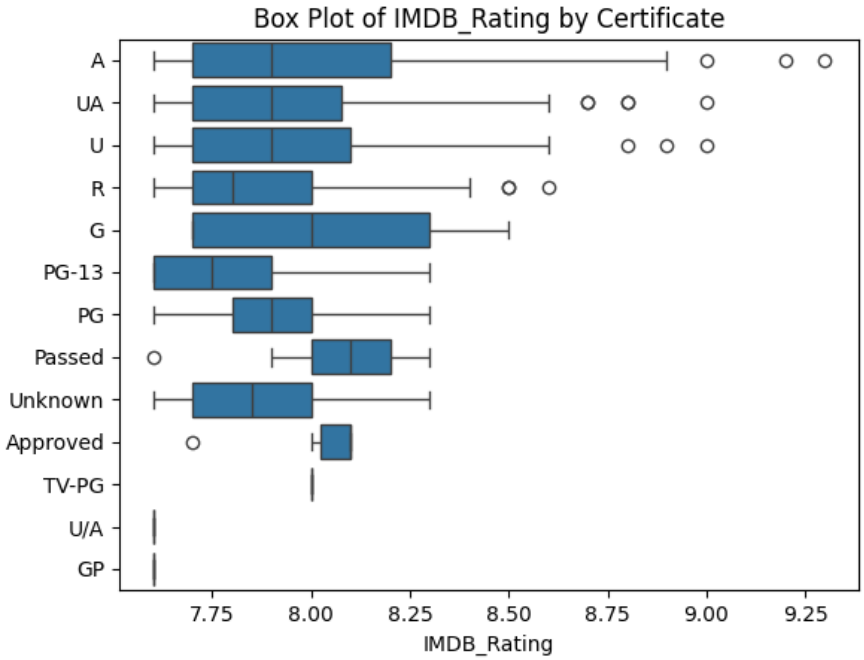
Figure 3



Comparative Analysis:

Figure 4 shows box plots of IMDB rating by Certificate. The plot corresponds to the general trend in the IMDB rating histogram chart in Figure 1. Most of the box plots have very similar minimum, Q1 and Q2 values. The first three certificates have the widest range in terms of rating. They also account for most of the outlier values. There is no one certificate that clearly has higher ratings than the rest.

Figure 4



Statistical Analysis

Table 2 shows the mean, median and standard deviation for Gross, No. of votes and IMDB rating. The distributions of Gross and No. of votes were checked and are exponential in shape. According to (ScienceDirect, 2025) in an exponential distribution the standard deviation is estimated to be equal to the mean.

Table 2

Values for mean, median and standard deviation for the following fields:			
	Mean	Median	Std dev
Gross:	74,952,069.0	31,900,000.0	74,952,069.0
No_of_Votes:	342133.0	219734.0	342133.0
IMDB_Rating:	7,9	7,9	0,3

The **Pearson correlation** between Gross revenue and No. of votes is **0.55546**. A value of +1 would've meant a perfect positive linear relationship. This value is above just the halfway mark between 0 and +1, indicating a moderate correlation. Hence as No. of votes increases, there's a modest increase in revenue as well, and visa versa.

The Inter-Quartile Range was used to identify outliers in Gross revenue. The **upper boundary** being **237 700 000.**

The outliers are:

[534858444, 377845905, 292576195, 315544750, 330252182, 342551365, 290475067, 322740140, 335451311, 422783777, 858373000, 678815482, 448139099, 248159971, 309125409, 293004164, 415004880, 356461711, 381011219, 380843261, 289916256, 293506292, 402453882, 341268248, 333176600, 363070709, 623279547, 305413918, 261441092, 260000000, 315058289, 936662225, 257730019, 318412101, 249358727, 245852179, 532177324, 408084349, 434038008, 258366855, 303003568, 760507625, 267665011, 659325379, 238632124, 435110554, 324591735, 259766572, 257760692, 274092705, 277322503, 304360277, 295983305, 290013036, 389813101, 608581744, 248757044, 251513985, 255959475, 301959197, 317575550, 285761243]

Conclusion

To ensure the continued success of the film and television series industries it is important to understand trends. With the insights gained in this analysis, one can have a better understanding of the factors that do and don't contribute to high IMDB ratings and high Gross revenue, among other things. With more data and a deeper, more thorough analysis, further insights can be gained.

References

Cover image 1
QuickBooksEMEA (2024). "*The zetigest-defining movie trend of the 2020's*". Reddit. Available at: https://www.reddit.com/r/decadeology/comments/1b5zips/the_zetigestdefining_movie_trend_of_the_2020s_the/ Accessed: 25 July 2025.

Cover image 2
N. Semlyen, J. Dye at al (2025). "*The 100 Best Movies Of All Time*". Empire. Available at: https://www.empireonline.com/movies/features/best-movies-2/ Accessed: 25 July 2025.

Image 3
A. Laskowski (2025). "*From Superman to F1, Expect a Summer of Blockbusters*". Boston University. Available at: https://www.bu.edu/articles/2025/summer-of-movie-blockbusters/ Accessed: 26 July 2025.

(ScienceDirect, 2025) "*Exponential Distribution*". ScienceDirect. Available at: https://www.sciencedirect.com/topics/mathematics/exponential-distribution#:~:text=always%20very%20skewed.-,For%20an%20Exponential%20Distribution,the%20next%20event%20is%20exponential. Accessed: 26 July 2025.