Data Analysis of IMDb Movie Ratings Using Python

(COMP3125 Individual Project)

Hayden Saunders   
*Data Science Fundamentals*

*Abstract*—This project analyzes IMDb movie data to explore patterns in film ratings, genres, and runtimes. The dataset, sources from Kaggles IMDb collection, was processed in Python using pandas and visualized with matplotlib. Key findings include th distribution of IMDb ratings, genre popularity, genre- based rating differences, and runtime-rating relationships. These insights demonstrate how open-source data can be used to better understand audience preferences and industry trends.

Keywords—IMDb, Data analysis, Python, visualization, movie ratings

# Introduction (*Heading 1*)

The film industry generates massive amounts of data through audience ratings, box office performance, and genre trends. Understanding these patterns can benefit filmmakers, distributors, and researchers in predicting movie success and tailoring creative strategies. This project focuses on analyzing IMDb’s movie dataset to identify relationships between movie genres, audience ratings and runtimes. Previous studies have leverages IMDb data to explore factors influence success, revealing that genre and runtime often play a role in how audiences perceive a film. By using Python’s data analysis libraries, we create visual representations of these relationships to present clear, interpretable insights.

# Datasets

## Source of dataset

The dataset was obtained from Kaggle (ashirwadsangwan/imdb-dataset), which hosts IMDb’s official TSV files: title.basics.tsv and title,ratings.tsv. IMSB is a credible and widely recognized source of film-related information, making the dataset reliable for analysis.

## Character of the datasets

The dataset contains metadata for thousands of titles, including films, TV shows, and documentaries. After merging the basics and ratings datasets on the tconst identifier, the final dataset contained the following areas:

| Column Name | Description | Unit / Format |
| --- | --- | --- |
| tconst | Unique IMDb ID | String |
| primaryTitle | Main release title | Text |
| originalTitle | Original language title | Text |
| titleType | Type of title (e.g., movie, short, TV) | Categorical |
| startYear | Release year | Year |
| runtimeMinutes | Film duration | Minutes |
| genres | One or more genres | Comma-separated text |
| averageRating | IMDb audience rating | Scale: 0–10 |
| numVotes | Number of ratings received | Count |

# Methodology

The analysis was conducted entirely in Python, using the following packages:

* **pandas**: for dataset loading, merging, cleaning, and manipulation
* **matplotlib**: for generating visualizations
* **kagglehub**: for automated dataset download from Kaggle

The key steps included:

1. **Dataset Acquisition** – Automated download from Kaggle, merging title.basics.tsv and title.ratings.tsv.
2. **Data Cleaning** – Handling \N placeholders for missing values, converting runtimes and ratings to numeric formats, and splitting genre strings into multiple rows.
3. **Visualization** – Generating four key plots:
   * IMDb rating distribution
   * Top 15 genres by number of titles
   * Average rating per genre
   * Runtime vs. rating scatter plot

# Results

##### A. Rating Distribution

##### The histogram of IMDb ratings revealed a peak between 6.0 and 7.5, suggesting most films receive moderately positive ratings. A graph of a distribution of imdb rating AI-generated content may be incorrect.

##### B. Genre Popularity

##### The top three most common genres were Drama, Comedy, and Thriller, indicating these dominate the dataset in terms of sheer volume.

A graph with red bars

AI-generated content may be incorrect.

##### C. Average Ratings by Genre

##### Documentary and History titles tended to receive higher average ratings, while genres like Horror and Action showed more variability.

A green and white bar graph

AI-generated content may be incorrect.

##### D. Runtime vs. Rating

##### The scatter plot revealed no strict correlation between runtime and rating, though extremely short and extremely long runtimes were less common and tended to have more polarizing ratings.

A graph with numbers and symbols

AI-generated content may be incorrect.

# Discussion

# While IMDb provides extensive data, several limitations affect the analysis:

# Sampling Bias: Popular genres and high-budget films tend to receive more votes, skewing averages.

# Incomplete Data: Missing runtimes and genres limit the sample size for certain visualizations.

# No Financial Metrics: Without revenue data, the analysis focuses only on audience perception, not profitability.

# Future work could include integrating box office revenue data to assess commercial success alongside ratings and applying machine learning models to predict ratings based on genre, runtime, and other factors.

# Conclusion

This project demonstrated how open datasets from IMDb can be used to analyze film trends and audience preferences. The results confirmed that Drama and Comedy dominate the industry by volume, but critically acclaimed films often belong to niche genres such as Documentary or History. These insights could help inform production and marketing strategies in the film industry.

##### Acknowledgment *(Heading 5)*

##### I would like to thank the IMDb and Kaggle teams for providing open access to structured film data, as well as the developers of pandas, matplotlib, and kagglehub for enabling efficient data analysis and visualization.

##### References

1. [1] IMDb, "IMDb datasets," [Online]. Available: <https://www.imdb.com/interfaces/>. [Accessed: Aug. 10, 2025].
2. [2] Kaggle, "IMDb dataset," [Online]. Available: <https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset>. [Accessed: Aug. 10, 2025].
3. [3] W. McKinney, *Python for Data Analysis*, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2017.