# "CineScope: Crafting Custom Recommendations & Insights from IMDb Data"

Team: **Data Dazzlers**

A Project By

Jayneet Ashish Jain
Hari Dave
Naitik Shah
Mihika Amit Kubadia
Vaishnavi Kulkarni
Usha Gourigari

Date: 12/10/2024

# Table Of Contents

# 1. Introduction

The entertainment sector has undergone substantial changes over the previous decade, owing to the rapid expansion of Over-The-Top (OTT) platforms. These platforms, which provide a varied selection of movies and television series, have transformed how material is consumed, disrupting old watching habits and bringing new kinds of engagement. With millions of titles available to consumers, knowing the patterns of content consumption is critical for forecasting future trends and enhancing user experiences. This project uses IMDb's extensive dataset to examine key characteristics of content consumption, such as the number of movies versus series released and their average ratings across genres. Additionally, a recommendation system is incorporated to analyze user preferences and suggest personalized content. By leveraging features like genres, ratings, and release years, the system enhances viewer satisfaction and supports OTT platforms in refining their strategies to maximize engagement.

.

# 2. Literature Review

### 2.1 Motivation

The rapid rise of Over-The-Top (OTT) platforms has transformed how people consume entertainment material. With the increasing popularity of these platforms, analyzing content consumption habits is critical for content creators, distributors, and marketers to remain competitive. Using IMDb's extensive dataset, this analysis seeks to find trends in viewer preferences, highlight regional and genre variations, and investigate issues such as runtime and audience involvement. These insights are critical for OTT platforms to optimise their strategy, provide personalised recommendations, and make data-driven decisions that improve user pleasure and engagement.

## 2.2 Research Question

1. **What are the trends in the number of series vs. movies released over the last decade?**
   This question focuses on understanding whether series are overtaking movies in popularity, potentially indicating shifts in production priorities and viewer consumption habits.
2. **How are the top genres distributed in ratings?**
   This explores whether certain genres are more suited to one type of content over another, helping producers tailor content to match audience preferences.

### 2.3 Summary of Literature

OTT platforms have reshaped content consumption habits, with Vinay and Rani (2020) highlighting their impact on traditional film. They discovered that region-specific and genre-based products have a considerable impact on customer decisions, highlighting the necessity of adapting material to localized preferences. This fundamental understanding reinforces the importance of doing a broad analysis of trends across genres and geographies [1]. Saha and Srivastava (2023) broadened this perspective by investigating the elements that influence user adoption of OTT platforms. Their findings indicate that personalization, affordability, and cultural congruence are important drivers of audience engagement, especially in emerging markets. This lends weight to the premise that regional and genre-based analysis might provide practical insights into viewing behavior [2].Pandey et al. (2016) emphasized the need of strong research procedures when dealing with huge datasets. Their emphasis on organized data analysis and ethical procedures establishes a framework for properly utilizing IMDb data. This assures that the conclusions drawn from this study are valid, reproducible, and impactful [3].

Together, these studies lay the groundwork for investigating how geographical preferences, audience engagement, and genre-based patterns influence content consumption on OTT platforms.

# 3. Methods

### 3.1 Data Description

The dataset for this project was sourced from IMDb, a large portal for film and television information. The dataset contains several tables, the main ones being title.basics.tsv.gz, title.ratings.tsv.gz, and title.akas.tsv.gz. These tables contain significant information about numerous aspects of movies and television shows.

**Key Variables**:

tconst: A unique identifier for each title (film or television series).

titleType: Indicates whether the entry is a movie, a series, or another sort of material (for example, documentary, short film).

startYear: The year that the title was released or first aired.

runtimeMinutes: The title's runtime in minutes, which applies to movies and series.

genres: The genre(s) connected with the title (for example, comedy, drama, action).

averageRating: The title's average user rating, as determined by IMDb user reviews.

numVotes: The number of user votes cast for the title.

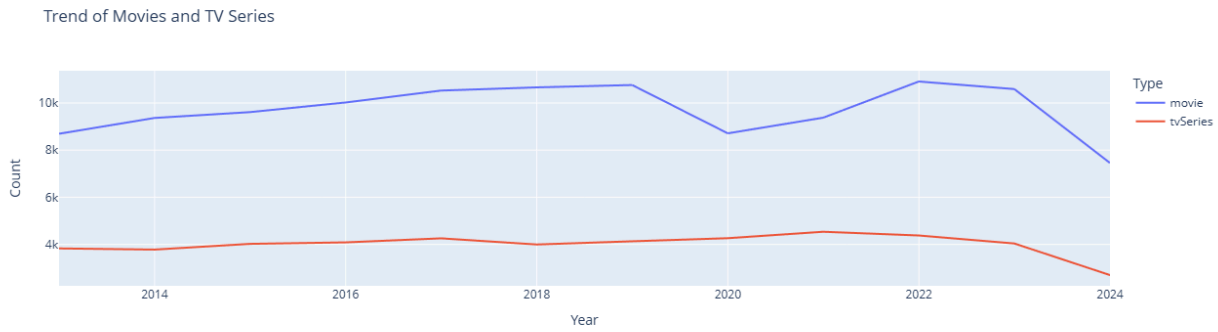
primaryTitle: The title's name (for example, movie or series name).

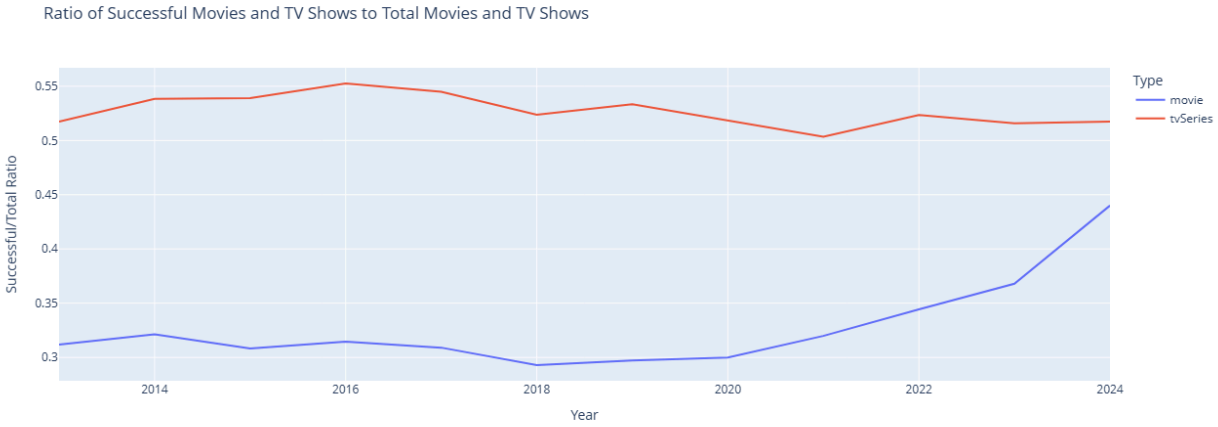
Data source: https://developer.imdb.com/non-commercial-datasets/

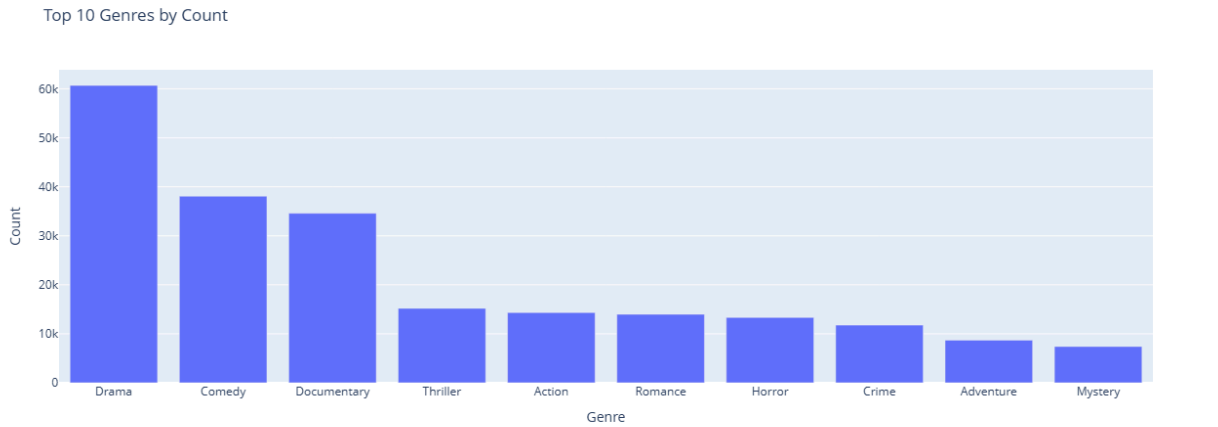
### 3.2 Data Analysis

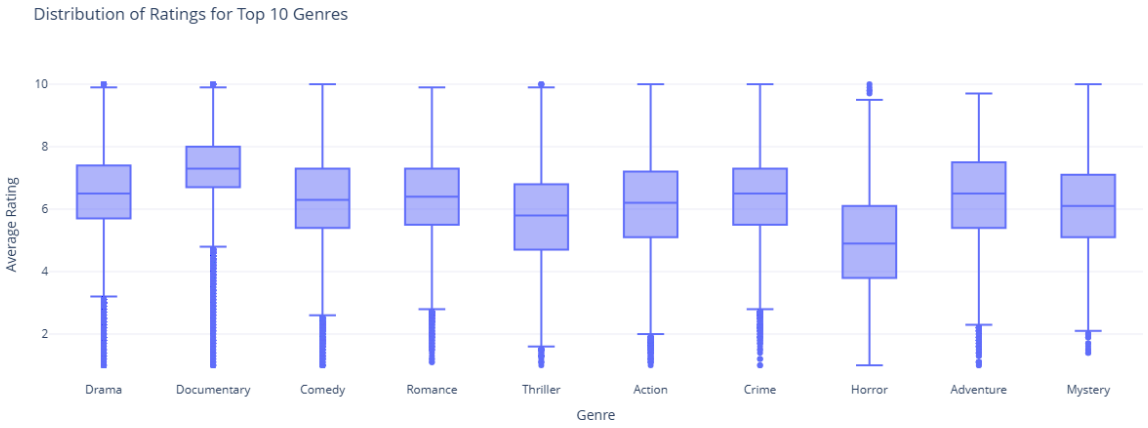
The project employs a content-based filtering approach for the recommendation system, using cosine similarity to identify titles closely aligned with user preferences. Movie features, including genres, average ratings, and release years, are encoded as tensors to facilitate efficient similarity calculations using PyTorch. The methodology aggregates feature vectors of user-selected titles to compute a mean vector, which is then compared against all other titles to derive similarity scores. Recommendations are refined by applying constraints such as release year and a minimum average rating threshold to ensure relevance. Advanced regression or predictive modeling techniques are not incorporated in this implementation. The system's effectiveness is validated through precision-recall metrics and user feedback, with post-hoc filtering applied to exclude irrelevant titles and improve recommendation quality.

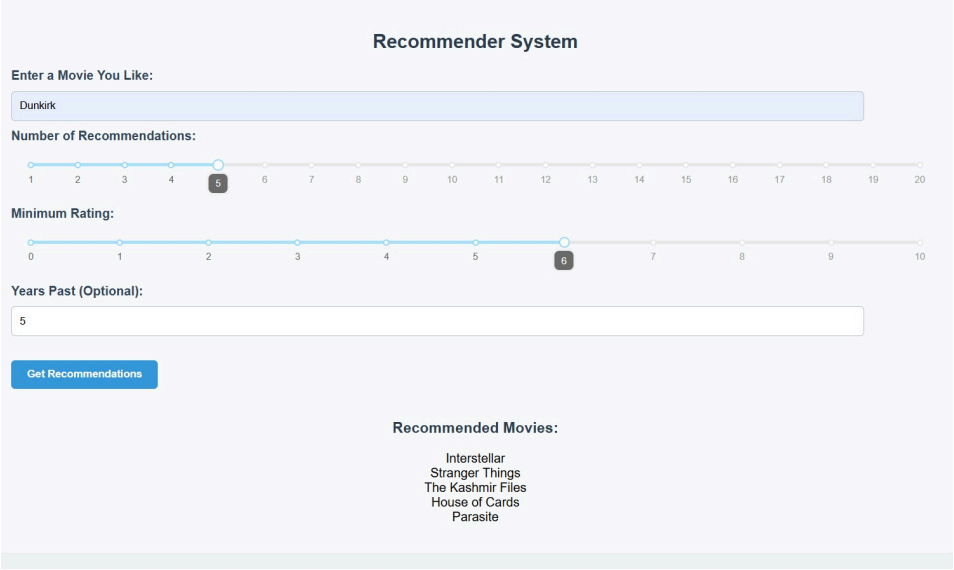
# 4. Results


Trend of Movies and TV Series
Type
movie
tvSeries
Count
10k
8k
6k
4k
2014
2016
2018
2020
2022
2024
Year


Ratio of Successful Movies and TV Shows to Total Movies and TV Shows
Type
movie
tvSeries
Successful/Total Ratio
0.55
0.5
0.45
0.4
0.35
0.3
2014
2016
2018
2020
2022
2024
Year


Top 10 Genres by Count
Count
60k
50k
40k
30k
20k
10k
0
Drama
Comedy
Documentary
Thriller
Action
Romance
Horror
Crime
Adventure
Mystery
Genre


Distribution of Ratings for Top 10 Genres
Average Rating
10
8
6
4
2
Drama
Documentary
Comedy
Romance
Thriller
Action
Crime
Horror
Adventure
Mystery
Genre


Recommender System
Enter a Movie You Like:
Dunkirk
Number of Recommendations:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Minimum Rating:
0 1 2 3 4 5 6 7 8 9 10
Years Past (Optional):
5
Get Recommendations
Recommended Movies:
Interstellar
Stranger Things
The Kashmir Files
House of Cards
Parasite

# 5. Discussion

The research questions examined in this analysis focused on:

1. **Trends in the number of series vs. movies released over the last decade**

The analysis reveals distinct trends in the release patterns of series and movies over the past decade. While series historically dominate in terms of success ratio, movies show a steady increase, particularly post-COVID-19, indicating a shift in viewer preferences and production strategies**.**

2. **Comparison of average ratings for series vs. movies across genres.**

Genres also exhibit notable variations in ratings, with certain genres like drama and thriller receiving higher average ratings in series, while action and adventure remain dominant for movies. These insights suggest that series are gaining ground in audience engagement, particularly in genres that emphasize character depth and long-form storytelling, helping producers align content strategies with evolving viewer expectations.

### Limitations:

- **Missing Data**: Incomplete data, especially on votes and ratings, could introduce bias.
- **Geographical Data**: Lack of direct geographical identifiers limits understanding of regional preferences.
- **Genre Classification**: Broad genre categorization may not capture the complexity of hybrid genres.

### Suggestions for Improvement:

- **Incorporate Time-Based Analysis**: Analyzing trends over multiple years could uncover shifting viewer preferences.
- **Refine Genre Classification**: Implementing a more detailed genre classification system would enhance the understanding of genre influences.
- **Geographic Analysis**: More granular regional data would improve insights into cultural preferences and viewing habits.

### Future work:

Future research may explore different interesting methods to improve our understanding of content consumption habits. One such approach is to use sentiment analysis from social media platforms to acquire a better understanding of viewer perceptions and reactions to movies and series. Furthermore, researching worldwide distribution and local content strategies could help

OTT platforms adjust their services to certain locations, increasing user engagement. Exploring language-based content trends, such as those in Bollywood, Hollywood, and other regional cinema industries, is another possible research field. Platforms might provide individualized suggestions by analyzing language preferences and content popularity across different linguistic markets, increasing user happiness and engagement in a variety of geographies.

### Conclusion

The visualizations show us that while tv shows are overall more successful, but in the recent years (particularly post covid), success ratios of movies have increased significantly. Additionally, some genres have a wide range of ratings (e.g. Drama, Action) indicating diverse content within the genre. Genres like Documentary and Mystery have a higher level of consistency in their ratings. Overall, series and movies have both declined in numbers but overall the success rates of movies increasingly matching tv shows shows a greater improvement in the overall industry.

## 6. References and Citation

[1] Vinay, M., & Rani, A. (2020). A study on factors influencing OTT video streaming and its impact on cinema. Journal of Content, Community, and Communication, 12(6), 65-73. https://doi.org/10.31620/JCCC.12.20/08

[2] Pandey, A., Aggarwal, A., Maulik, M., & Sharma, M. (2016). Research methodology and publication ethics. Indian Journal of Anaesthesia. https://doi.org/10.4103/0019-5049.190621

[3] Saha, S., & Srivastava, A. (2023, February). Digitized Entertainment–Factors influencing Consumer Adoption of OTT Platforms. In *2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON)* (pp. 1-6). IEEE.