CineInsight

Saiteja Siddana
Courant Institute of
Mathematical Sciences
New York University
New York, USA

Praharsh Allada
Courant Institute of
Mathematical Sciences
New York University
New York, USA

Harsha Joshi
Courant Institute of
Mathematical Sciences
New York University
New York, USA

Abstract—

In this paper about the project "CineInsight" is a comprehensive project that focuses on collecting and analyzing movie data from notable sources such as TMDB, IMDb, and MovieLens. With this project we aim to provide insightful analytics on the trends in movie ratings, and popularity over the years by leveraging MapReduce for data cleanup, and profiling followed by Hive and Trino for querying.

Keywords—analytics, movies, big data, ratings, averages

I. Introduction

In the ever-evolving landscape of film and media, the analysis of movie data presents a unique opportunity to understand trends, preferences, and patterns in the global film industry. This report delves into a comprehensive analysis of movie data sourced from three major platforms: IMDb (Internet Movie Database), TMDb (The Movie Database), and MovieLens. These platforms are renowned for their extensive databases, offering a rich tapestry of information on films, including ratings, reviews, metadata, and viewer preferences.

The primary objective of this project was to conduct a thorough analysis of movie data to extract meaningful insights that could benefit stakeholders in the film industry, including filmmakers, marketers, and film enthusiasts. To achieve this, the project was structured in distinct phases. Initially, a MapReduce approach was employed for data cleaning and profiling. This step was crucial to ensure the integrity and quality of the data, which is fundamental for accurate analysis. MapReduce, known for its efficiency in processing large datasets, was instrumental in organizing and refining the data for subsequent stages.

Following the data preparation phase, a series of analyses were conducted on the resulting datasets. These analyses aimed to uncover trends in movie ratings, genre popularity, viewer preferences, and other critical metrics that could inform understanding of current market dynamics and future trends in the film industry.

This report presents the findings of this analysis, offering insights into the complexities and nuances of movie data and its implications for the film industry. Through a combination of data-driven methodologies and comprehensive analysis, the study seeks to contribute to a deeper understanding of film data and its potential applications.

II. Motivation

The motivation behind our analytics is rooted in the recognition of big data's transformative power in the film industry. By harnessing comprehensive insights into genre popularity, audience ratings, and runtime trends, we aim to equip film producers and studios with actionable intelligence that can drive the creation of resonant and successful content. This analysis is not merely academic; it has practical applications, enabling industry players to make well-informed decisions that can shape the future of entertainment. From determining which types of films receive green lights to tailoring marketing campaigns that effectively target the right audiences, our analytics serve as a compass in the complex landscape of cinematic production and distribution.

In the rapidly evolving domain of streaming services, our analysis offers a strategic edge. By decoding viewer preferences, these platforms can refine their recommendation algorithms to curate content that captivates and retains subscribers. Our analytic work thus becomes instrumental in streamlining content discovery, enhancing user satisfaction, and fostering a deeper connection between the digital screen and its audience.

Ultimately, the purpose of our analysis is to bolster the film industry's capacity to produce content that resonates with diverse audiences and to refine the movie-watching ecosystem at large. By aligning film production with audience preferences, we not only elevate the art of storytelling but also ensure its economic viability in a competitive market. This analytic is designed to be a beacon for the industry, guiding it towards a future where data-driven decisions lead to richer narratives, more engaged viewers, and a thriving cinematic culture.

III. Design and Implementation

The workflow depicted in the diagram represents our design and approach of this analytics. Initially, disparate datasets from various sources such as IMDB, TMDb, and MovieLens are collected, each offering unique pieces of information about movies and then ingested into HDFS. These datasets are then subjected to a 'Clean and Profile' phase, involving data preprocessing steps like handling missing values, removing duplicates and profiling involved calculating basic statistics, which gave us initial insights like the number of movies in each genre and the average ratings for each movie. This phase is crucial as it ensures the quality and reliability of the data, which is foundational for accurate analysis. The cleaned data sets are then merged using a MapReduce framework, which allows for efficient processing of large-scale data by distributing the task across multiple nodes and then consolidating the results.

Upon merging, we converted our resultant dataset into HIVE tables so that we can do our analysis and distributed SQL querying using Trino. In our analysis, various columns and combinations of data are examined to uncover insights such as viewer preferences, performance metrics, and trends across different dimensions of the movie data. If the initial analysis reveals gaps or prompts further questions, adjustments are made to the process, iterating until the refined data meets the requirements for final reporting. Finally, we visualized our analytic results in bar graphs, line charts, using Excel and Tableau.

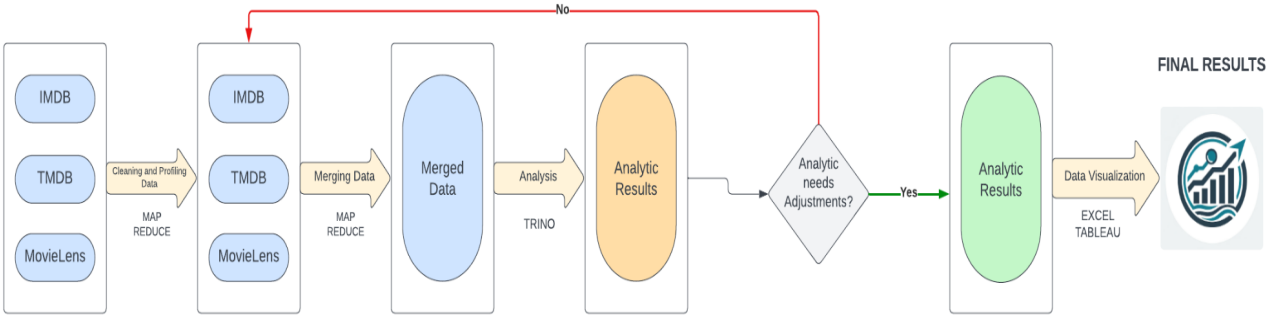


Figure: Data Design Diagram

IV. Data Sources and Profiling

A. Movie Lens 25M Dataset

The MovieLens 25M dataset, one of the primary data sources for our analysis, is a vast and detailed collection obtained from the MovieLens website, encompassing a plethora of movie ratings and associated metadata across 25 million entries. For our project, we have selectively utilized three of its CSV files, which provide us with user ratings, movie details, and essential links to two other data sources.

1. Ratings Data (ratings.csv): This file is pivotal to understanding user preferences, containing a million instances of user ratings ranging from 0.5 to 5 stars. Each record includes a unique user identifier, the movie's identifier, the rating awarded, and a timestamp marking the exact moment the rating was submitted.
2. Movies Data (movies.csv): Serving as the central reference for movie details, this file associates each film with a unique identifier, its title, and a set of genres. The genres are categorized in a pipe-separated list, allowing for the representation of multiple genres associated with a single movie.
3. Movie Links (links.csv): This file acts as a bridge to external databases, providing unique IMDB and TMDB identifiers for each movie. With over 62,000 rows, it enables the possibility of enriching our dataset with additional metadata from these external sources.

movies

movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action|Crime|Thriller
7	Sabrina (1995)	Comedy|Romance
8	Tom and Huck (1995)	Adventure|Children
9	Sudden Death (1995)	Action
10	GoldenEye (1995)	Action|Adventure|Thriller
11	American President, The (1995)	Comedy|Drama|Romance
12	Dracula: Dead and Loving It (1995)	Comedy|Horror
13	Balto (1995)	Adventure|Animation|Children
14	Nixon (1995)	Drama
15	Cutthroat Island (1995)	Action|Adventure|Romance
16	Casino (1995)	Crime|Drama
17	Sense and Sensibility (1995)	Drama|Romance
18	Four Rooms (1995)	Comedy
19	Ace Ventura: When Nature Calls (1995)	Comedy
20	Money Train (1995)	Action|Comedy|Crime|Drama|Thriller
21	Get Shorty (1995)	Comedy|Crime|Thriller
22	Copycat (1995)	Crime|Drama|Horror|Mystery|Thriller
23	Assassins (1995)	Action|Crime|Thriller
24	Powder (1995)	Drama|Sci-Fi
25	Leaving Las Vegas (1995)	Drama|Romance
26	Othello (1995)	Drama
27	Now and Then (1995)	Children|Drama
28	Persuasion (1995)	Drama|Romance
29	City of Lost Children, The (Cité des enfants perdus, La) (1995)	Adventure|Drama|Fantasy|Mystery|Sci-Fi
30	Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)	Crime|Drama
31	Dangerous Minds (1995)	Drama
32	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	Mystery|Sci-Fi|Thriller
33	Wings of Courage (1995)	Adventure|Romance|IMAX
34	Babe (1995)	Children|Drama
35	Carrington (1995)	Drama|Romance

Data cleaning just involved the process of data validation, dropping any NULL values, and dropping unwanted columns. Then, for data profiling I implemented the map-reduce job for calculating average rating of each movie given by users, and also counted the number of movies that we have in this dataset in each genre.

B. IMDB Non Commercial Dataset

The IMDb dataset serves as an extensive repository of data for film and television titles, categorized by type, title, release year, and genre. Each entry includes a unique identifier, titles in both promotional and original languages, content rating, and duration. The dataset also provides viewer ratings and vote counts, offering insights into audience reception. This information is crucial for analysis and insights in the entertainment industry.

The dataset was cleaned of noise and profiled using a map reduce program. Before the data was cleaned, it had 10,327,720 entries. After the data cleaning steps, we were left with 962,885. The resulting data looked like the following:

movieId	releaseYear	movieTitle	runtime	genres	ratings	votes
tt0000001	1894	Carmencita	1	Documentary,Short	5.7	2007
tt0000002	1892	Le clown et ses chiens	5	Animation,Short	5.8	269
tt0000003	1892	Pauvre Pierrot	4	Animation,Comedy,Romance	6.5	1912
tt0000004	1892	Un bon bock	12	Animation,Short	5.5	178
tt0000005	1893	Blacksmith Scene	1	Comedy,Short	6.2	2694
tt0000006	1894	Chinese Opium Den	1	Short	5.0	182
tt0000007	1894	Corbett and Courtney Before the Kinetograph	1	Short,Sport	5.4	842
tt0000008	1894	Edison Kinetoscopic Record of a Sneeze	1	Documentary,Short	5.4	2154
tt0000009	1894	Miss Jerry	45	Romance	5.3	207
tt0000010	1895	Leaving the Factory	1	Documentary,Short	6.9	7352
tt0000011	1895	Akrobatisches Potpourri	1	Documentary,Short	5.2	378
tt0000012	1896	The Arrival of a Train	1	Documentary,Short	7.4	12539
tt0000013	1895	The Photographical Congress Arrives in Lyon	1	Documentary,Short	5.7	1919
tt0000014	1895	The Waterer Watered	1	Comedy,Short	7.1	5666
tt0000015	1894	Autour d'une cabine	2	Animation,Short	6.2	1129

The profiling involved finding the total number of movies released for every genre for every year of the dataset.

C. TMDB Dataset

TMDB is widely recognized as a standard repository for movie data. However, the publicly accessible dataset is limited to around 5000 records, making it very incomplete. This dataset has been scraped off the internet using the TMDB API provided by TMDB, this API can be obtained in the TMDB Settings page. The dataset has an option of scraping data such as Title, Overview, Release date, Runtime, Genres, Poster path, Backdrop path, Vote average, Vote count, budget etc., of which we have selected the columns id, popularity, original_title, cast, overview, vote_count, vote_average, release_year, genres all the way from the year 1874 to 2023.

The data scraped off the TMDB website as such contains 1,65,555 records.Shown below is an image of what the data looked like . The cleanup was fairly straightforward by removing the rows which did not have one or more of the entries or those which had an empty string as an entry. However when cleaning and profiling the data using map-reduce some we had issues with the description column and hence the mapping cannot be done with each record in a new line since the description sometimes had newline character which needed special care and we had to add the condition that it is a new record if the first 2 parts of new line are numbers then its a new records and its not a new record otherwise and continued with the description of the old record.

```
id,popularity,original_title,cast,overview,vote_count,vote_average,release_year,genres
315946,4.793000,"Passage de Venus","","Photo sequence of the rare transit of Venus over the face of the Sun, one of the first cl
921940,1.960000,"Zim, Boum, Boum","","Praxinoscope animation of a green-suited boy wearing a drum and cymbal.  Series 1, number
766094,1.960000,"La Rosace Magique","","Praxinoscope strip of a shifting rosette. Series 2, number 5.",17,5.60,1878,"Animation"
751212,1.827000,"Le Singe Musicien","","A pre-cinematograph colour animation of the monkey playing his violin.",21,6.00,1878,"A
921930,1.438000,"La Danse sur la Corde","","Early Praxinoscope strip showing a girl dancing on a tightrope.  Series 1, Number 7
922511,0.768000,"Le Jongleur","","Praxinoscope animation of a juggler balancing a spinning plate on their nose.  Series 1, numb
921939,1.400000,"Le Jeu de Corde","","Praxinoscope of a girl in a blue dress skipping rope. Series 1, number 9",8,4.60,1878,"An
922010,0.600000,"Les Papillons","","Praxinoscope animation of a butterfly fluttering about a flower.  Series 2, number 6.",4,4.
921938,1.231000,"L'Aquarium","","Praxinoscope reel, Series 1, number 1",4,3.30,1878,"Animation"
922515,1.230000,"Les Bulles de Savon","","Praxinoscope animation of a girl in a blue dress blowing soap bubbles.  Series 1, numb
922184,0.728000,"Les Chiens Savants","","Praxinoscope animation of a boy holding a hoop for two trained dogs to leap through.
922081,0.600000,"Le Steeple-chase","","Praxinoscope animation of an equestrian riding.  Series 3, number 8.",4,4.30,1878,"Anima
```

As for the Profiling part of it we used 2 other map reduce functions to calculate the genre wise average rating of movies which showed the highest for documentaries and the lowest for Horror and for the other map reduce we tried to get the most popular and the highest rated movies which came out to be “How the Grinch Stole Christmas” and “The 1st 13th Annual Fancy Anvil Awards Show Program Special: Live in Stereo”.

V. Results

A. Top Movies

The top 10 films, when considering both popularity and average ratings, showcase a diverse array of cinematic achievements that have successfully captured audience attention while also receiving critical acclaim. These movies, ranging from the immersive visual experience of "IMAX" features to the storytelling prowess of "The Shawshank Redemption," highlight a blend of mass appeal and quality that resonates across a broad spectrum of viewers, combining box office success with enduring cultural impact.

Title	Average Rating
The Shawshank Redemption	9.281693674
The Nagano Tapes: Rewound, Replayed & Reviewed	9.192779783
The Godfather	9.181107855
Steven Banks: Home Entertainment Center	9.151028807
Mirror Game	8.997655102
The Private Life of Plants	8.994068387
Ashi Hi Banwa Banwi	8.993812271
The Godfather Part II	8.984858576
The Dark Knight	8.984386624
Michael Jackson Live in Bucharest: The Dangerous Tour	8.981184333

Table 1.1 Top 10 Movies based on Ratings

Title	Popularity
Dragon Kingdom	1184.555
How the Grinch Stole Christmas	468.862
The Grinch	430.867
Krampus	421.717
The Hunger Games: Mockingjay - Part 1	341.679
Casper's Haunted Christmas	335.158
Charlie and the Chocolate Factory	334.336
Home Alone 2: Lost in New York	330.406
Willy Wonka & the Chocolate Factory	317.242
What Every Frenchwoman Wants	311.493

Table 1.2 Top 10 Movies based on Popularity

B. Popularity and Rating trends

The overall Popularity of movies has been increasing however the rating has been hugely fluctuating.

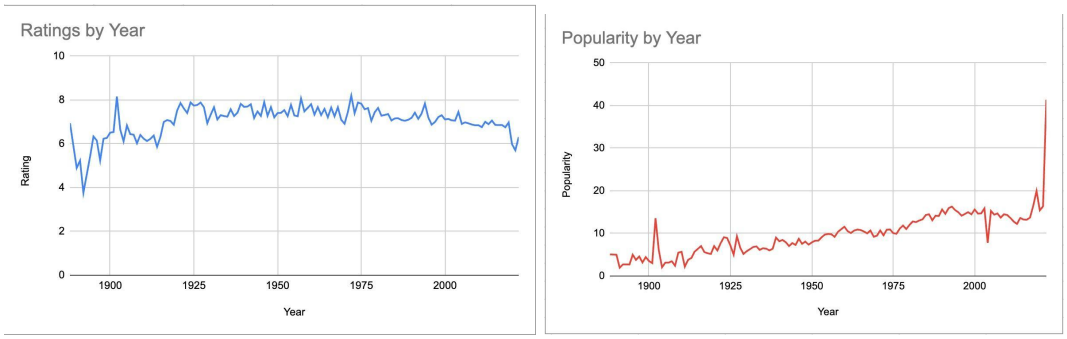


Figure 1

a) ACTION

The data on action genre films over the years reveals a consistent trend of moderate ratings and varying levels of popularity. From the early 1900s, action films have evolved in both style and reception. In recent decades, the genre has seen a significant increase in popularity, as reflected in the surge of audience engagement in the 21st century. Despite fluctuations in ratings, the enduring appeal of action films is evident, with the genre maintaining a substantial presence and capturing the attention of a wide audience, reaching its peak in popularity in 2019.

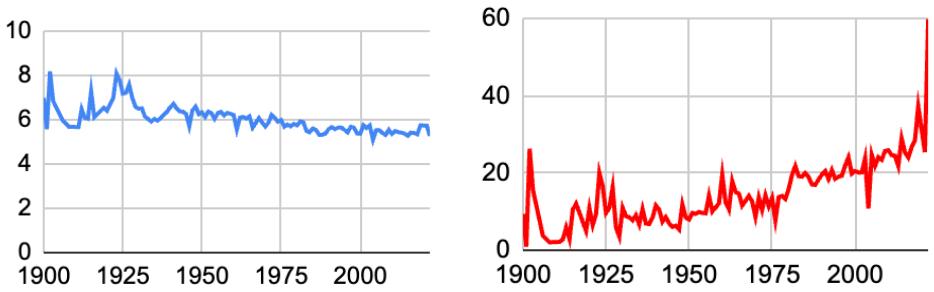


Figure 2.1

b) ADULT

In the early 1970s, the genre started with a modest rating of 2.43 and garnered relatively high popularity at 33.85. The subsequent years witnessed fluctuations in both ratings and popularity, with peaks and troughs reflecting changing tastes and cultural influences. By 1999, the adult genre had seen a significant rise in both ratings (6.78) and a moderate level of popularity (7.52), indicative of a shift in societal attitudes and perhaps advancements in production quality.

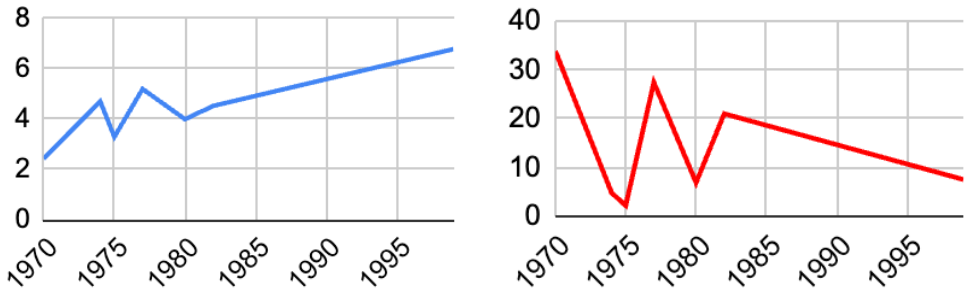


Figure 2.2

c) ADVENTURE

The Adventure genre has evolved over the years, with a steady rise in ratings and popularity from the 1920s onward. The 1940s and 1950s marked peaks, while the 1970s saw significant popularity. Despite challenges in the late 20th century, Adventure films made a strong comeback in the 21st century, culminating in a remarkable surge in ratings and popularity in 2018 and 2019, followed by a substantial peak in 2022.

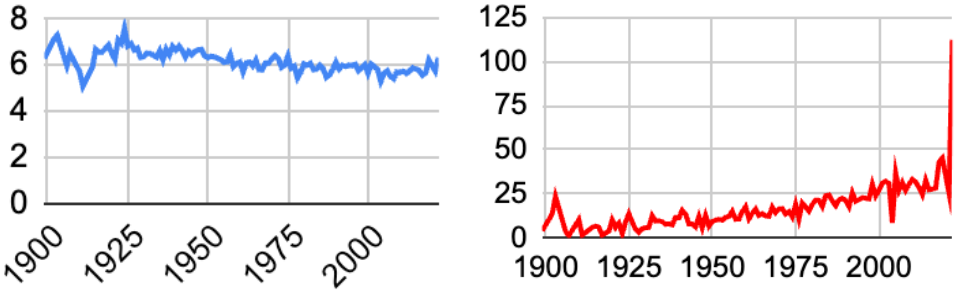


Figure 2.3

d) Animation

The Animation genre has evolved significantly over the years, with a notable rise in both ratings and popularity. Starting from the early 1900s, the genre gained traction, reaching a peak in the late 1950s and 1960s. Despite occasional fluctuations, Animation maintained a steady presence, demonstrating resilience and enduring appeal. The genre experienced a resurgence in the late 2000s, with increasing popularity and continued strong ratings, highlighting its enduring impact on audiences

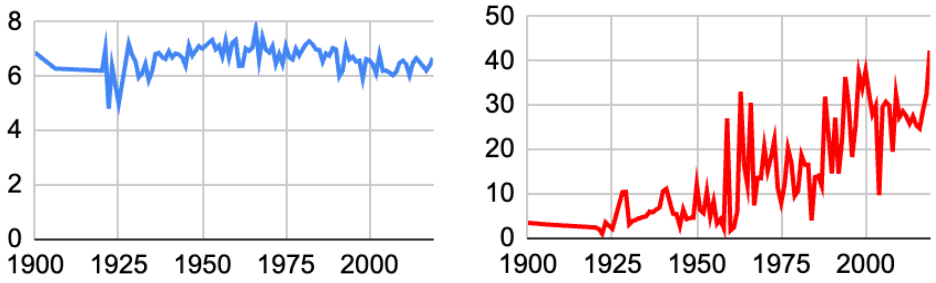


Figure 2.4

e) Biography

The Biography genre has evolved significantly over the years, with a notable increase in both ratings and popularity. Starting in the late 19th century, the genre saw a steady rise in ratings, peaking in the 1920s and 1930s. Despite some fluctuations, Biography maintained a strong presence, particularly in the mid-20th century.

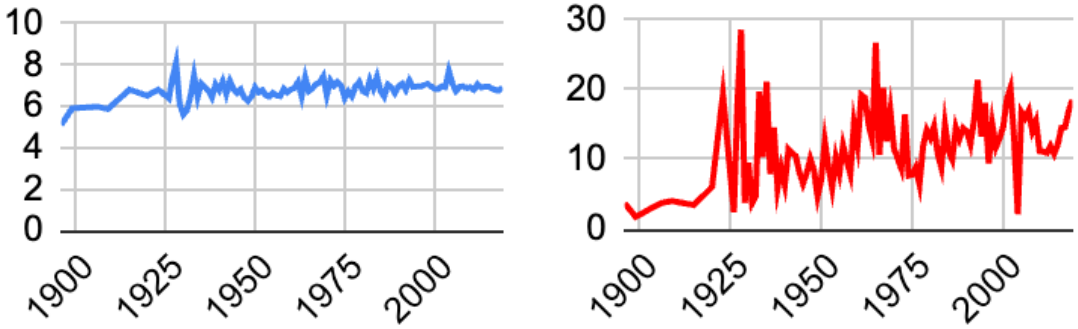


Figure 2.5

f) Children

The children's genre has witnessed a steady increase in ratings and popularity over the years. Starting from a modest 6.48 in 1899, the genre experienced fluctuations but generally maintained positive growth. Notably, the popularity soared in recent years, reaching its peak with a rating of 6.52 and an impressive popularity score of 44.80 in 2019, showcasing the enduring appeal of children's content across the decades.

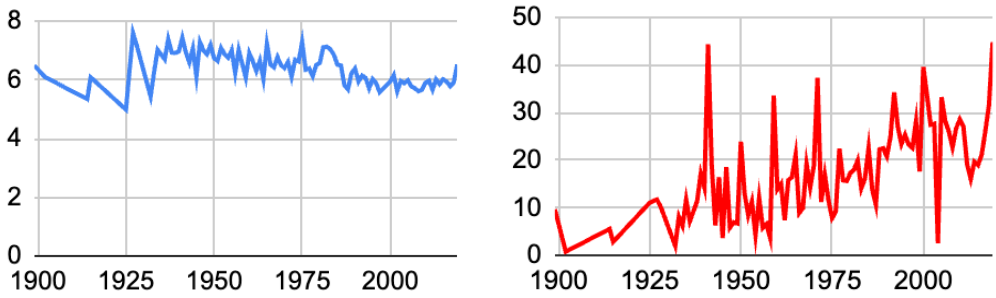


Figure 2.6

g) Comedy

The comedy genre has evolved significantly over the years, with its popularity and ratings fluctuating. In the early 1900s, comedy films like those in 1902 received high ratings and popularity. The genre experienced a resurgence in the 1950s, marked by increased ratings and audience engagement. However, in recent years, the genre's average ratings have shown a decline, while its popularity has remained relatively stable.

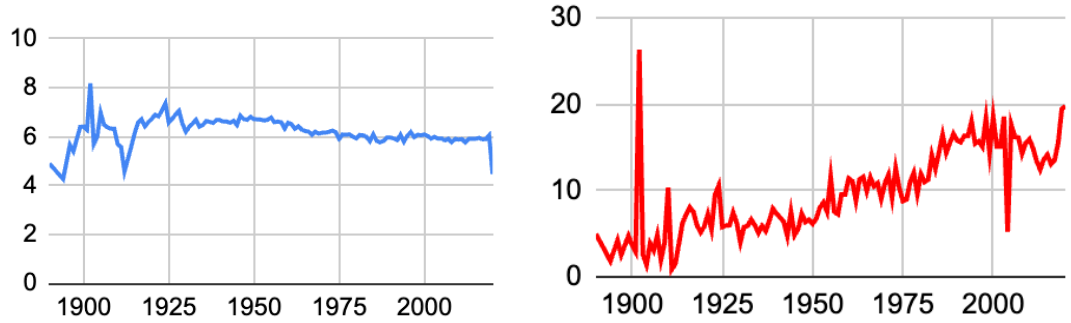


Figure 2.7

h) Crime

The Crime genre has maintained a consistent presence in the film industry, with ratings ranging from 5.7 to 7.3 over the years. Despite fluctuations, it gained popularity steadily, reaching a peak in 2019 with a rating of 6.05 and a popularity score of 22.0. The genre's resilience and enduring appeal are evident in its ability to engage audiences across different eras.

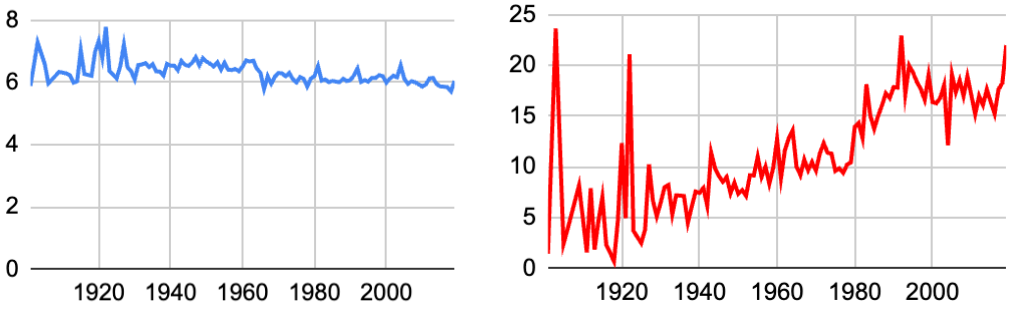


Figure 2.8

i) Documentary

The Documentary genre has evolved significantly over the years, with a notable increase in both ratings and popularity. Starting in the late 19th century with modest ratings around 5, the genre experienced a surge in the early 20th century, reaching peaks in the 1930s and 1950s, with consistently high ratings above 7. The trend continued into the 21st century, showcasing sustained popularity and quality, although a slight dip in ratings was observed in the year 2020.

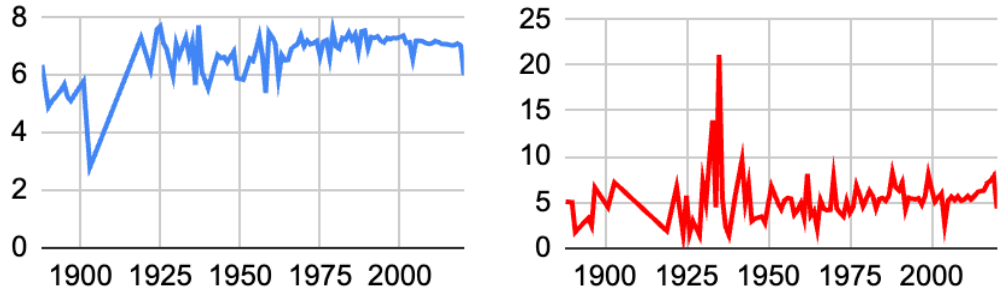


Figure 2.9

j) Drama

The Drama genre has experienced a steady evolution in ratings and popularity over the years. Starting in 1895 with a rating of 5.87 and a popularity of 3.5, the genre gradually gained traction, reaching a peak in 2019 with a rating of 6.15 and an impressive popularity score of 19.44. However, in 2020, there was a noticeable decline in both rating (5.31) and popularity (23.74), suggesting a potential shift or challenge for the genre.

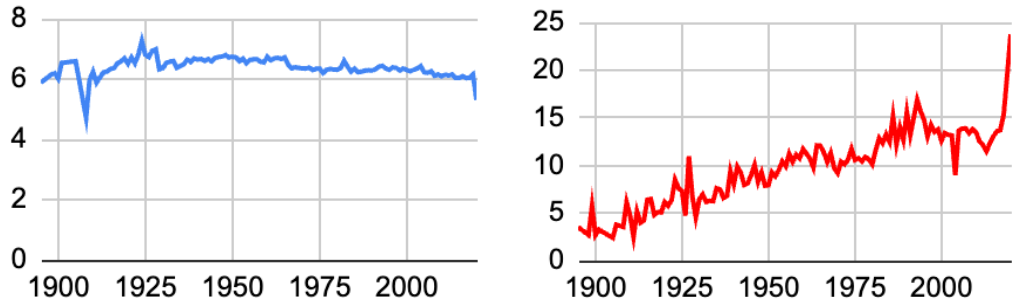


Figure 2.10

k) Family

The family genre has consistently maintained a moderate to high average rating over the years, ranging from 6.0 to 7.0. Despite fluctuations in popularity, the genre has demonstrated enduring appeal, with increasing popularity in recent years, peaking at 29.5 in 2019. The positive correlation between ratings and popularity suggests a continued audience interest in family-oriented content across different decades.

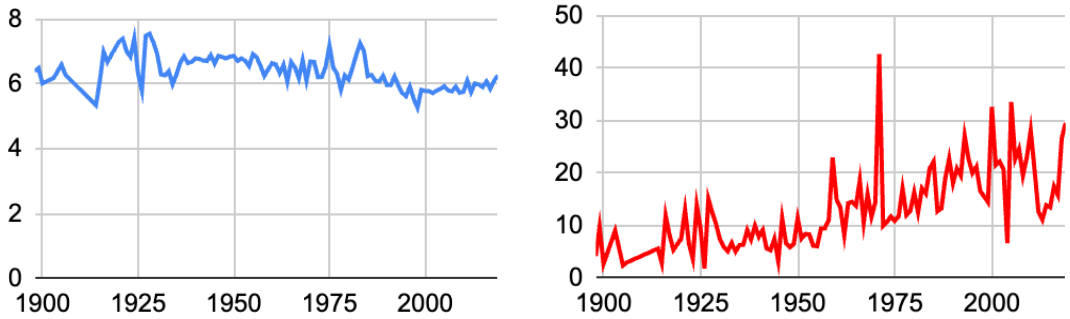


Figure 2.11

l) Fantasy

The Fantasy genre has experienced a significant evolution over the years, with a notable spike in popularity during the early 1930s, marked by an impressive rating of 7.397. The genre continued to captivate audiences, reaching its peak in 1993 with a substantial popularity score of 31.74.

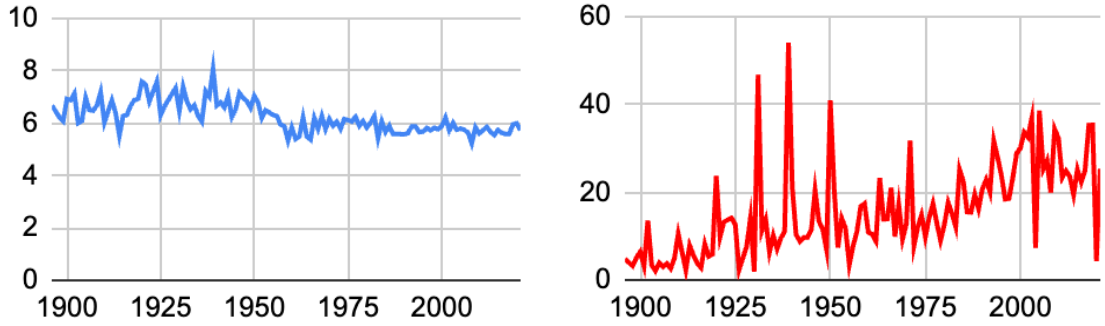


Figure 2.12

m) Film-Noir

The Film-Noir genre, characterized by its dark and atmospheric storytelling, gained prominence in the 1940s. During this period, films like "1941" (6.74 rating, 8.25 popularity) and "1944" (6.87 rating, 10.91 popularity) exemplify the genre's compelling narratives and growing popularity. Despite a decline in the genre's prominence in subsequent years, it experienced a resurgence with notable films such as "1974" (8.20 rating, 47.42 popularity) showcasing its enduring appeal.

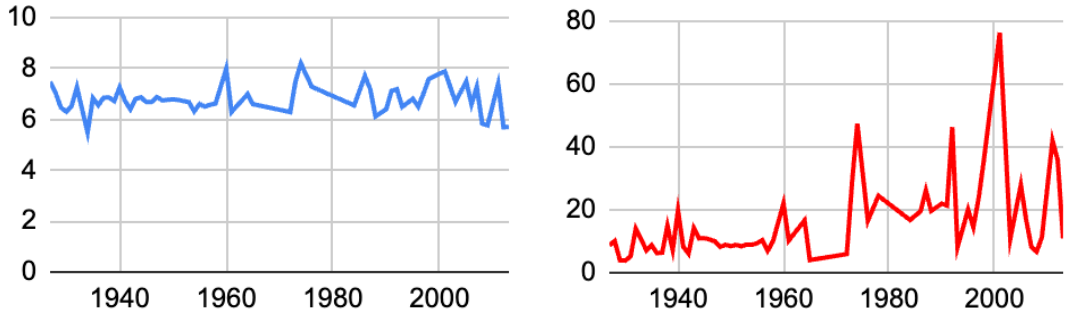


Figure 2.13

n) History

The History genre has evolved significantly over the years, witnessing a steady rise in ratings and popularity. Beginning in 1895 with a modest rating of 6.67, the genre gained momentum, reaching peaks such as 7.33 in 1984 and a remarkable popularity of 32.13 in 2019. This enduring trend suggests a consistent appeal and appreciation for historical content across different eras.

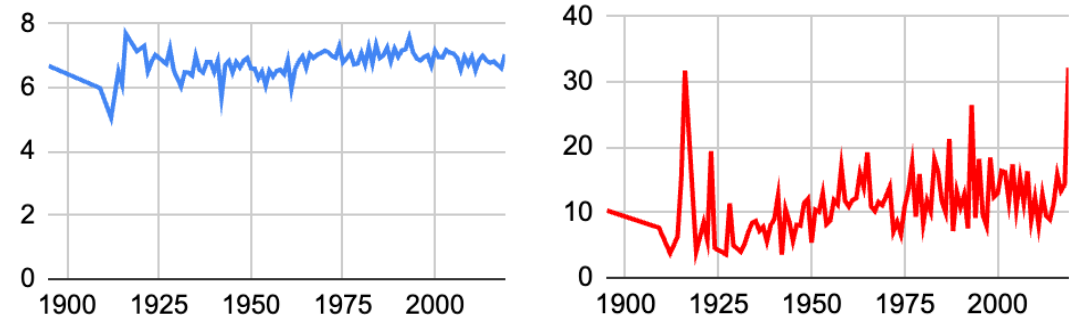


Figure 2.14

o) Horror

The horror genre has evolved over the years, experiencing fluctuations in both ratings and popularity. While the early 1900s saw a steady rise in horror film ratings, the genre gained significant popularity in the 1920s, with notably high ratings. However, the subsequent decades witnessed a fluctuating pattern, with occasional peaks and troughs in both ratings and audience engagement. Despite variations, the horror genre continues to captivate audiences, maintaining a substantial level of popularity in recent years.

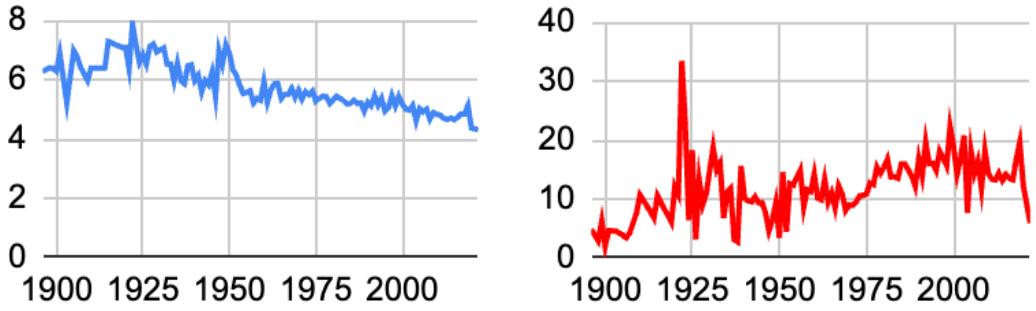


Figure 2.15

p) IMAX

Over the years, the IMAX genre has demonstrated varying trends in ratings and popularity. While the ratings have seen fluctuations, reaching a peak of 7.95 in 1991, the genre's popularity, measured in audience numbers, has notably surged in certain years, such as 2005 with 131.49 and 2010 with 95.31. The IMAX genre appears to have a dynamic relationship between critical acclaim and audience appeal.

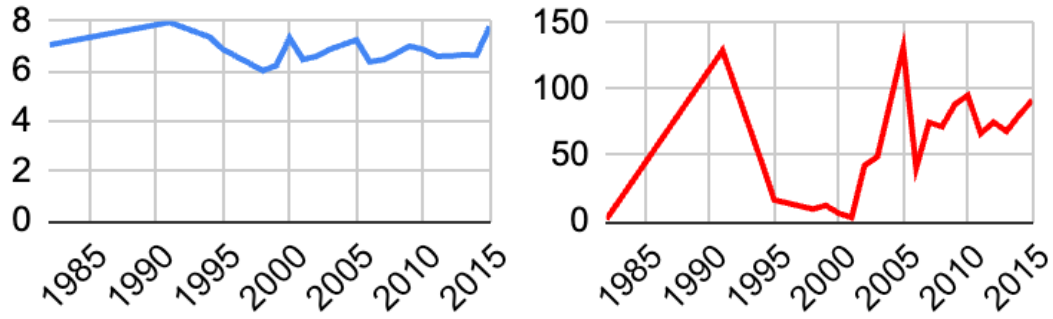


Figure 2.16

q) Music

Over the years, the Music genre has experienced fluctuations in ratings and popularity. Beginning in the early 20th century, the genre witnessed varying reception, with notable peaks in the 1950s and a resurgence in the 1970s and 1980s. Despite occasional dips, the genre has maintained a generally positive trajectory, culminating in a high average rating and popularity in recent years, emphasizing its enduring and evolving appeal.

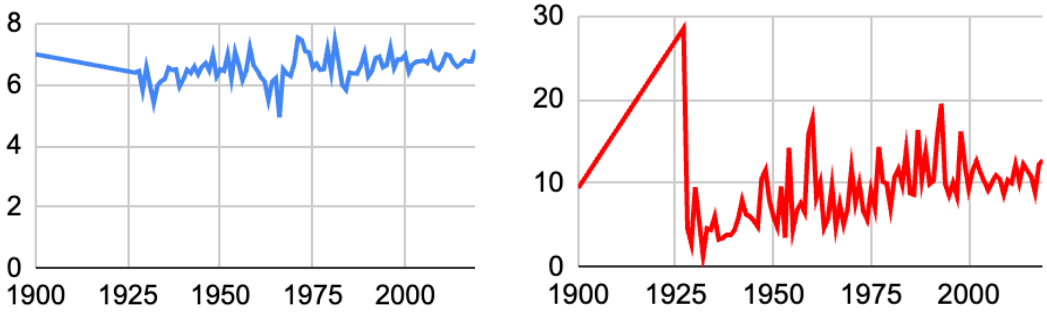


Figure 2.17

r) Musical

Over the decades, the Musical genre has experienced fluctuations in both ratings and popularity. Beginning in the late 1920s, the genre saw varying levels of acclaim and audience engagement, with notable peaks in the early 1970s. Despite some periods of lower ratings, the Musical genre has demonstrated resilience, maintaining a consistent presence in the entertainment landscape.

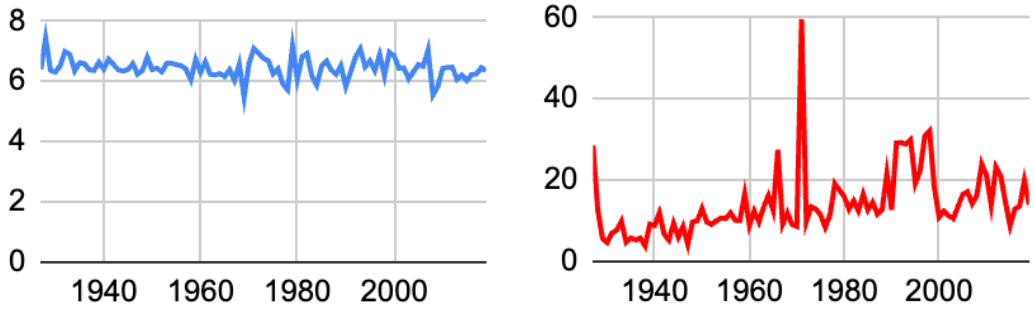


Figure 2.18

s) Mystery

Over the years, the Mystery genre has shown a fluctuating trend in both ratings and popularity. Starting in the early 20th century with modest ratings and popularity, it experienced a significant boost in the 1920s. The genre continued to evolve, with occasional peaks and troughs, reaching its pinnacle in the late 1950s. However, from the 1980s onwards, Mystery witnessed a gradual decline in both ratings and popularity, culminating in a notable decrease in recent years.

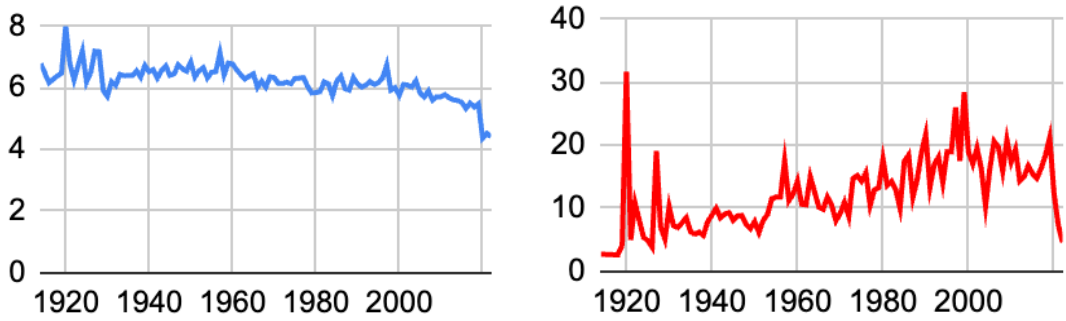


Figure 2.19

t) News

The News genre's ratings have shown fluctuations over the years, with a peak in 2014 at 7.74. Despite occasional variations, the genre has maintained a generally high standard, ranging from 6.99 to 7.77. Popularity has also varied, reaching a notable peak in 2006 at 23.2 and experiencing fluctuations in subsequent years.

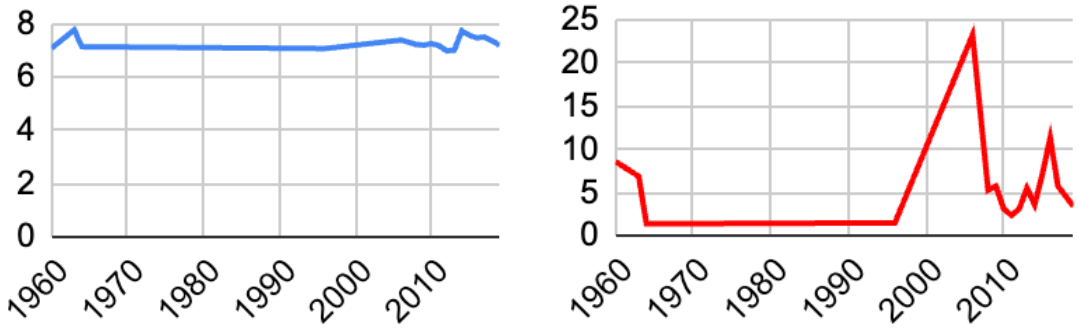


Figure 2.20

u) Reality-TV

The Reality-TV genre has undergone fluctuations in ratings and popularity over the years. In 1993, it started with a moderate rating of 6.59 and high popularity at 8.85. By 2014, ratings increased to 7.10, but popularity declined to 6.03. Subsequently, in 2018, the genre regained popularity at 7.11 but experienced a decrease in ratings, showcasing its dynamic nature.

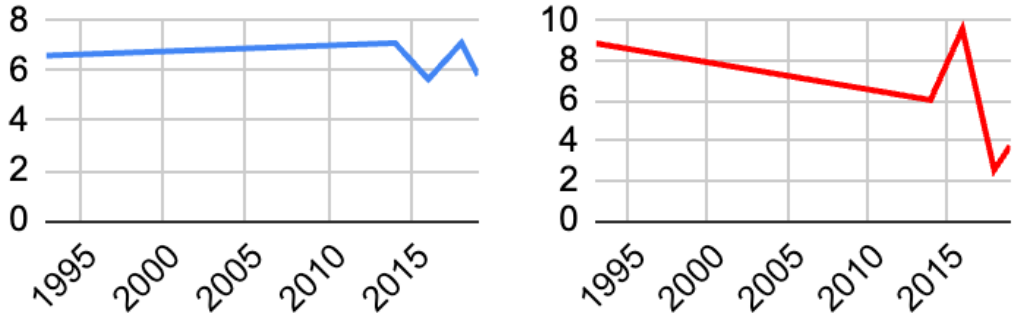


Figure 2.21

v) Romance

The Romance genre has seen a steady evolution over the years, with ratings fluctuating between 5.5 and 6.2 from 1900 to 2020. Popularity, measured by audience interest, has shown an increasing trend, reaching its peak at 24.022 in 2020. Despite variations in ratings, the genre maintains a consistent appeal and remains a popular choice among audiences.

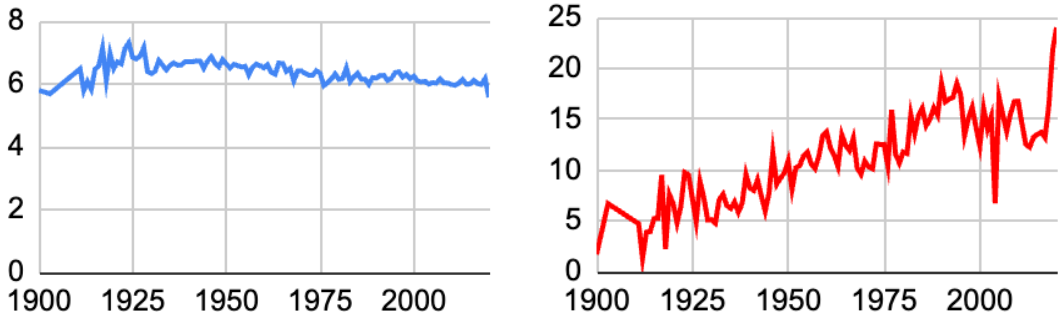


Figure 2.22

w) Sci-Fi

Over the years, the Sci-Fi genre has evolved significantly in both ratings and popularity. Starting in the early 20th century with moderate ratings, it saw a notable rise in the late 1920s, coinciding with increased popularity. The genre continued to fluctuate in the mid-20th century but experienced a resurgence

in the late 20th and early 21st centuries, reaching its peak in 2018 with high ratings and unprecedented popularity.

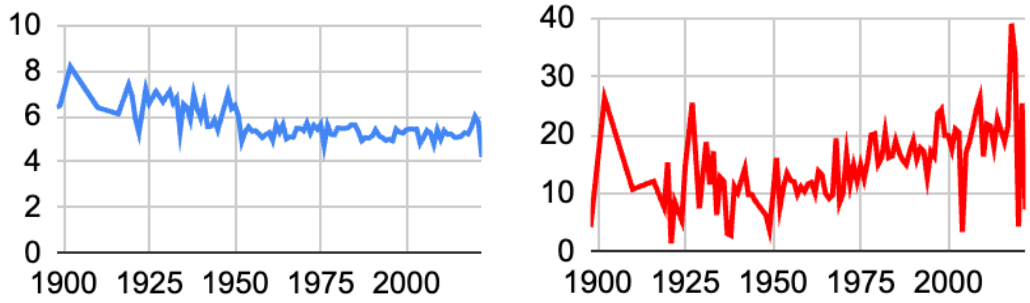


Figure 2.23

x) Sport

The sport genre has evolved over the years, with ratings fluctuating but generally maintaining a positive trend. The popularity of sport-related content has seen notable peaks, such as in 1961 and 1985, reflecting the genre's ability to capture audience interest. Despite occasional variations, the genre consistently resonates with viewers, blending competitive narratives with compelling entertainment.

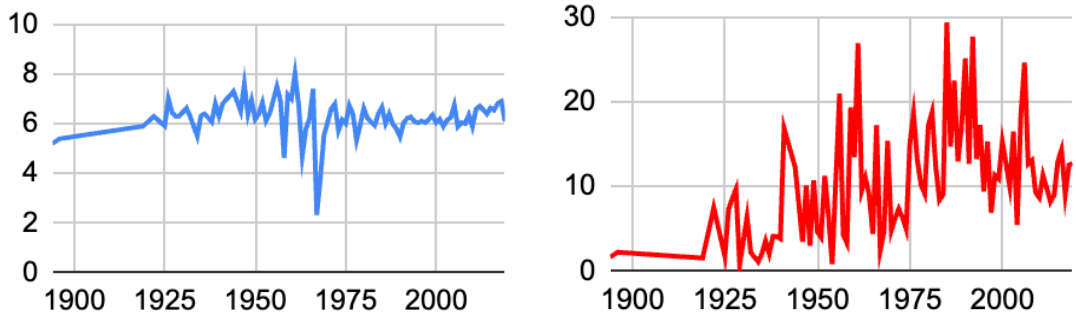


Figure 2.24

y) Short

The Short genre has seen a fluctuating trend in ratings and popularity over the years. Starting in the late 19th century, ratings varied between 3.76 and 7.28, with a notable peak in 1975 at a perfect 8. Popularity has generally increased, reaching its highest point in 2019 with a popularity score of 7.49. Despite occasional dips, the Short genre has maintained a steady presence in the film landscape.

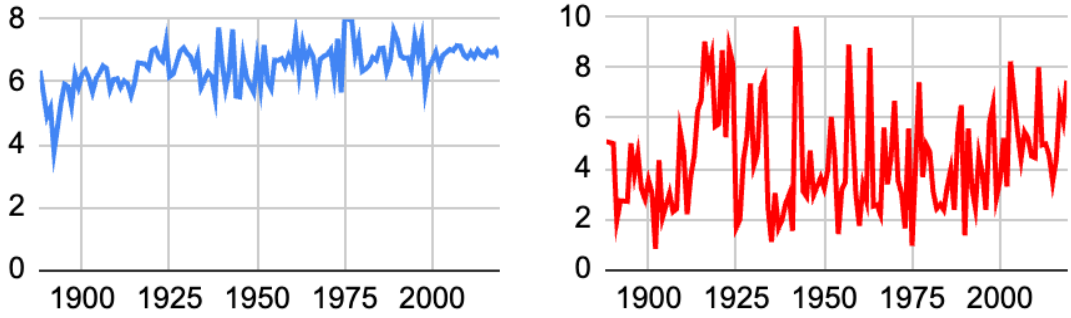


Figure 2.25

z) Thriller

The Thriller genre has experienced fluctuations in ratings and popularity over the years. In the early 1900s, it gained moderate ratings, but its popularity steadily increased, reaching a peak in the 1960s and 1970s. However, the genre saw a decline in both ratings and popularity in the 1980s and 1990s. In recent years, despite a decrease in ratings, Thriller has regained some popularity, especially in 2019, but experienced a notable drop in 2020 and 2021.

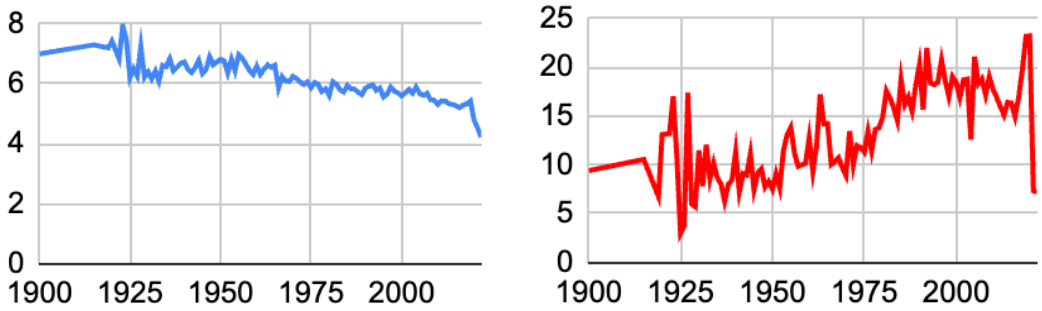


Figure 2.26

aa) War

The War genre has exhibited varying trends in both ratings and popularity over the years. In the early 20th century, from 1910 to 1940, there was a steady increase in ratings, reaching a peak during World War II. The genre continued to thrive in the post-war years, with a resurgence in popularity in the late 1970s and 1980s. The highest-rated and most popular War films in recent years have maintained a consistent level of acclaim and audience engagement.

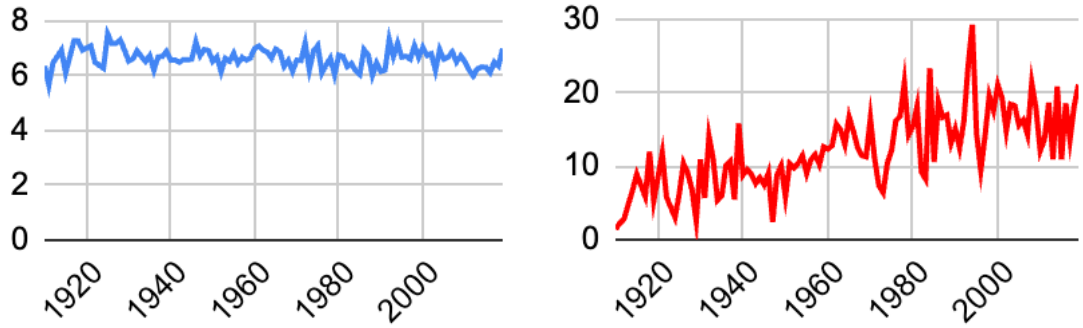


Figure 2.27

ab) Western

The Western genre has evolved significantly over the years, with its popularity peaking in the early 1960s. During this period, the genre achieved high ratings, exemplified by the peak rating of 6.880 in 1961. However, in recent years, the Western genre has seen a decline in both ratings and popularity, with the lowest rating recorded in 2016 at 5.366 and a decrease in popularity from the 1960s peak. Despite this, the genre continues to hold a lasting impact on cinematic history.

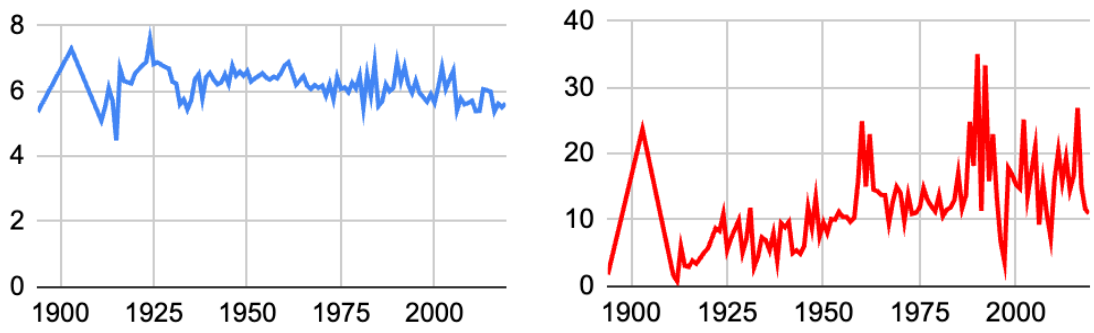


Figure 2.28

C. Optimum Rumtime Range

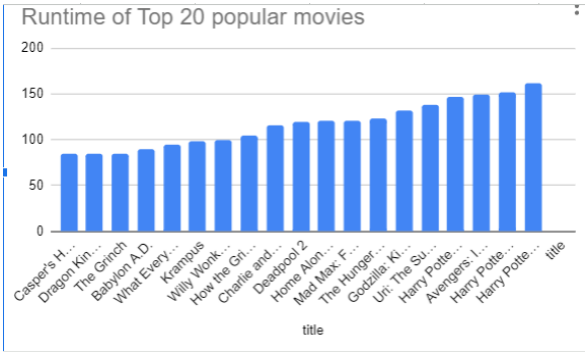
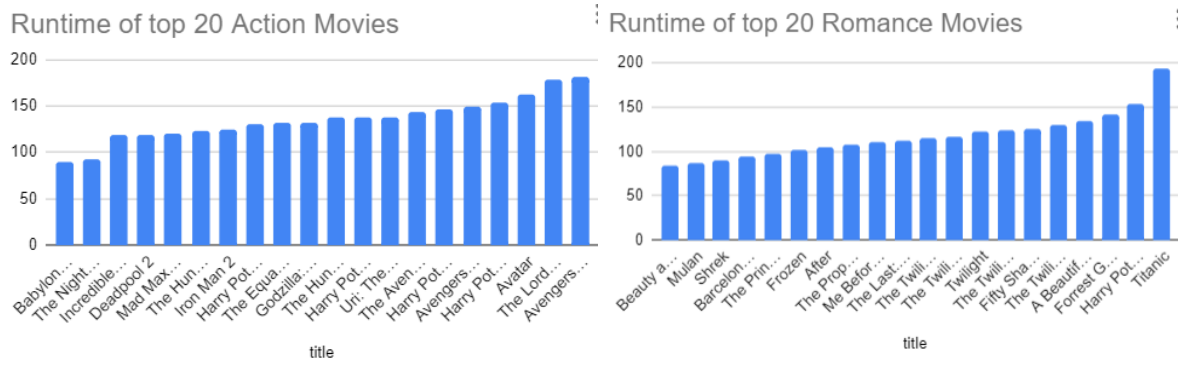


Figure 3.1: Runtime of top 20 movies

The above graph offers a compelling visualization of the runtime distribution for the top 20 popular movies across various genres. A critical observation is that all these movies share a runtime within the range of 80 to 155 minutes, suggesting a preferred movie length among audiences and filmmakers alike.

Similarly, analysis can also be made for specific genres giving us more insights. Following are the analysis for Action, Romance and Comedy genre.



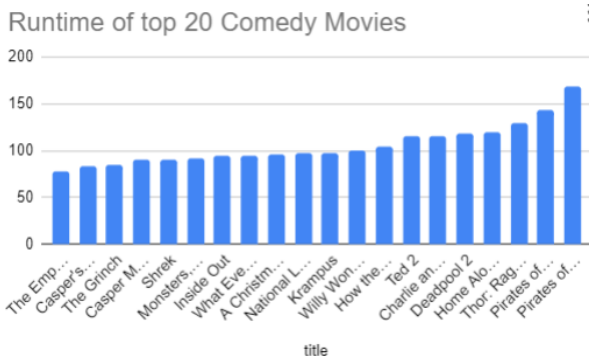


Figure 3.2: Runtime Analysis of Top 20 Movies based on specific genre

This convergence towards a specific runtime range can be attributed to several factors. Firstly, it aligns with the standard narrative structure that provides adequate time for character development, plot progression, and a satisfying resolution. Movies that are too short may not offer the depth that viewers expect, while those that are too long could lead to viewer fatigue.

Secondly, this runtime range is also logistically practical, fitting into standard programming slots for theaters and accommodating the typical audience attention span. It also facilitates the inclusion of these movies into streaming platforms that often look for content that can be consumed in a single sitting.

Furthermore, the data indicates a potential sweet spot for filmmakers when deciding the length of a movie. Staying within this range may increase the likelihood of a film achieving popularity, provided that other elements of the film, such as story, direction, and acting, are also well-executed.

The analysis of these top-performing movies' runtimes offers an invaluable reference for future film projects. By understanding the audience's preferences, filmmakers can tailor the movie length to align with market expectations, potentially increasing the film's reception and success.

The data from this graph suggests that while the content and quality of a film are paramount, the runtime is also a crucial factor that can influence a movie's popularity. This insight lays a foundation for further research into how runtime affects audience satisfaction and box office performance.

D. Genre Trends

The analysis of average ratings across various film genres reveals intriguing insights into viewer preferences and genre popularity. As shown in figure 4.1, leading the chart is Film-Noir, with a compelling average rating of 7.65, closely followed by News and War genres, suggesting a high viewer appreciation for the thematic complexity and historical significance often embodied within these categories. Biographical films and historical documentaries also receive high acclaim, averaging above 7.5, a testament to the audience's interest in narrative depth and educational value. Genres traditionally associated with broad commercial appeal such as Drama, Animation, and Crime exhibit strong performance as well, all scoring above 7.2, indicating a consistent demand for their distinctive blend of storytelling and visual engagement. At the lower end of the spectrum, Reality-TV and Adult genres face significant challenges in garnering viewer approval, with ratings of approximately 5.98 and 4.19 respectively, highlighting a critical or discerning attitude from viewers towards content within these categories.

These findings underscore the subjective nature of film appreciation and the complex interplay between genre, viewer expectation, and critical reception.

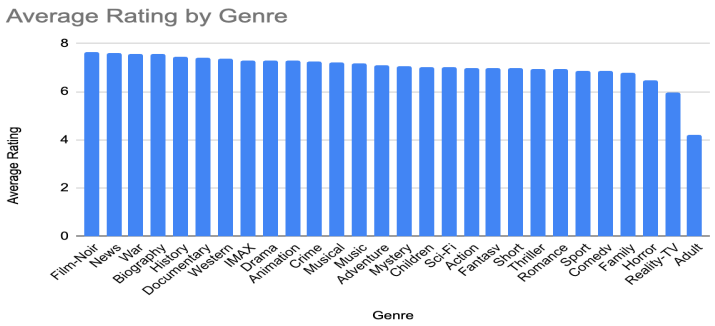


Figure 4.1: Average ratings of each Genre

Coming to the analysis of genre popularity, as shown in figure 4.2, IMAX leads significantly with a popularity score of over 70, a testament to the immersive cinematic experiences it offers, which appear to resonate profoundly with audiences. This is followed by a cluster of genres — Adventure, Children, Fantasy, Animation, and Action — each with popularity scores ranging from approximately 20 to 24. Such figures suggest a robust preference for escapism, visual spectacle, and dynamic storytelling, which these genres typically provide.

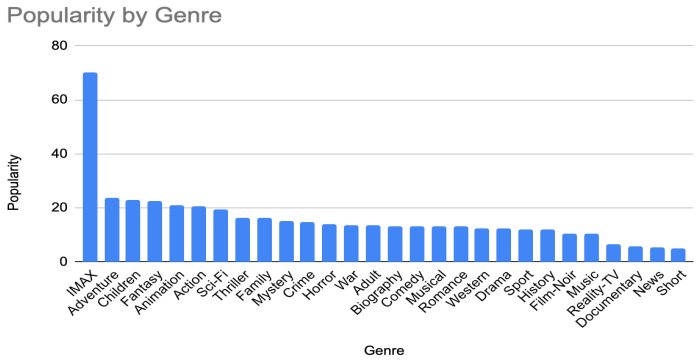


Figure 4.2: Average Popularities of each Genre

When juxtaposing these popularity scores against the average ratings data, an interesting dichotomy emerges. Genres with the highest ratings, such as Film-Noir and War, do not correspond to the most popular genres, indicating that while certain genres may achieve critical acclaim and resonate deeply with a particular segment of viewers, they may not necessarily command widespread popularity. Conversely, IMAX's towering popularity does not correlate with a leading position in average ratings, underscoring a potential divide between the experiential draw of certain genres and their perceived quality or impact.

Additionally, our analysis of genre pairings provides a quantitative lens through which we can examine the interplay between different cinematic genres within individual films. As shown in table 2, Comedy and Drama are the most frequently paired genres, with 5318 movies blending these two styles, underscoring a preference for narratives that merge humor with more serious undertones. Similarly, the Drama and Romance pair, with 5263 occurrences, highlights the enduring appeal of love stories and emotional narratives in cinema. The high incidence of Drama with Thriller, and Crime with Drama, with 4380 and 4149 movies respectively, indicates a market favorability towards films that promise both an engaging plot and a deeper, more nuanced storyline. These

combinations are likely crafted to deliver a richer and more complex viewing experience, often resulting in a gripping emotional journey that resonates well with audiences.

Genre 1	Genre 2	Number of Movies
Comedy	Drama	5318
Drama	Romance	5263
Drama	Thriller	4380
Crime	Drama	4149
Comedy	Romance	3694
Crime	Thriller	2806
Action	Drama	2683
Horror	Thriller	2632
Action	Thriller	2251
Mystery	Thriller	2162
Drama	Mystery	2101
Action	Adventure	2063
Action	Crime	1827
Adventure	Drama	1707
Adventure	Comedy	1600
Drama	Horror	1488
Comedy	Crime	1482
Biography	Drama	1464
Action	Comedy	1422
Comedy	Family	1385

Table 2: Movie Count by Genre Pairings

For filmmakers and studios, all these analytics offer a roadmap to expand their audience reach by strategically combining genres, which are not experimented till now or by even following the proven combination, that resonate with diverse viewer groups. By understanding which genre pairings are most prevalent and appealing, and which genres have wider audience content creators can craft narratives that blend elements from different genres, thereby appealing to a wider spectrum of moviegoers and potentially increasing the commercial success of their films.

VI. Future Work

The insights gleaned from this analysis of IMDb, TMDb, and MovieLens data offer fertile ground for future exploration and application. The following avenues present promising opportunities for extending the scope of this project:

1. Genre Evolution Study: With the observed shifts in genre popularity and ratings over time, a deeper investigation into the factors driving these changes could be illuminating. Future studies could incorporate social and economic indicators to understand how external variables influence film genre trends.
2. Hybrid Genre Analysis: Recognizing the high ratings of Film-Noir and the widespread popularity of IMAX, further research could focus on the potential of hybrid genres. Analyzing audience reception to films that blend elements of high-rated but less popular genres with widely popular formats could uncover new market opportunities.
3. Runtime Optimization: Given the observed optimal runtime range for top movies, predictive models could be developed to suggest the ideal length for new films based on genre, intended audience, and distribution method (theatrical release versus streaming platforms)..
4. International Market Trends: Expanding the dataset to include regional film databases could help in understanding global market trends and regional preferences, which are crucial for international releases and marketing.
5. Longitudinal Studies: Conducting a longitudinal study to monitor how the trends identified in this analysis evolve over the next decade could provide continuous insights into the changing landscape of the film industry.

VII. Conclusions

In the ever-evolving landscape of the film and media industry, this project has endeavored to unlock the transformative power of big data through a meticulous analysis of movie data from IMDb, TMDb, and MovieLens. Our journey was motivated by a deep understanding of the potential impact comprehensive insights can have on the film industry, from content creation to audience engagement.

The design and implementation of our analytics workflow, involving data collection, cleaning, profiling, and in-depth analysis, allowed us to extract meaningful insights from three major datasets. Utilizing the MovieLens 25M, IMDb Non-Commercial, and TMDB datasets, we addressed crucial aspects such as viewer preferences, genre popularity, and the dynamic relationship between ratings and popularity.

Our findings are presented in various analyses, showcasing top movies based on both ratings and popularity, discerning trends in genre-specific preferences, identifying an optimum runtime range, and understanding the nuanced evolution of genres over time. The top 10 movies, a deep dive into genre trends, and the examination of the interplay between ratings and popularity provide stakeholders in the film industry with actionable intelligence.

As a compass in the complex cinematic production and distribution landscape, our analysis offers strategic insights for filmmakers, studios, and streaming services. By aligning film production with audience preferences, our findings contribute to a deeper understanding of market dynamics and trends, providing a foundation for data-driven decisions that lead to richer narratives, more engaged viewers, and a thriving cinematic culture.

In essence, our work goes beyond the academic realm; it is a practical guide for industry players to navigate the competitive market, make informed decisions, and shape the future of entertainment. From genre-specific preferences to strategic runtime considerations, our analysis empowers the film industry to create content that not only resonates with diverse audiences but also ensures economic viability in an ever-changing landscape.

In the era of streaming services, our insights provide a strategic edge by decoding viewer preferences and refining recommendation algorithms, thereby streamlining content discovery and enhancing user satisfaction.

In conclusion, this project serves as a beacon for the film industry, illuminating a path towards a future where data-driven decisions lead to richer narratives, more engaged viewers, and a cinematic culture that thrives on the pulse of audience preferences.

Acknowledgments

Thanks to Google DataProc for providing the clusters used for MapReduce Job and for querying using Trino. We appreciate Grouplens for making the data used in the development of Movielens publicly available through their parent website. We would also like to Acknowledge IMDB for creating, and maintaining their developer website which aids developers and students access their data to be used for projects or other analysis. We are Grateful to TMDB for providing an API for everyone alike to be able to scrape data directly off the parent website and hence making it very flexible for people using it to selectively get . Appreciation to Google for providing google sheets and to trino both of which enabled us to make some quick graphs.

References

1. https://www.lafabbricadellarealta.com/open-data-entertainment/
2. https://grouplens.org/datasets/movielens/
3. https://developer.imdb.com/non-commercial-datasets/
4. https://developer.themoviedb.org/reference/intro/getting-started
5. https://www.themoviedb.org/settings/api?language=en-US