A Cinematic Journey through Data

Movie Data Wrangling & Analysis
Sai Lohith Chimbili & Sahil Parab
Course: Data Wrangling | Instructor: Stevenson Bolivar Atuesta
Date: May 6, 2025

Executive Summary

This project unravels the evolution of global cinema through data—spanning decades, genres, runtimes, and directorial styles. As movie lovers and data explorers, we fused our passion for storytelling with analytical rigour, scraping and merging datasets from IMDb to discover how filmmaking trends have changed over the years.

Main objective:
Our goal was to clean, integrate, and analyse movie data to answer key questions:

1. Which decades defined the golden age of cinema in terms of both volume and quality?
2. Who are the most consistent directors and actors based on IMDb ratings?
3. How have genres and runtimes shifted over time?
4. Does longer screen time correlate with better ratings?

Key insights:
Key findings include a dramatic rise and fall in movie production, genre booms, and runtime inflation in the streaming era. Directors like Masaki Kobayashi and Nolan showed exceptional consistency, directors who consistently produced high-quality films, such as Masaki Kobayashi and Christopher Nolan, built strong legacies defined by critical acclaim. Conversely, actors with prolific filmographies did not always maintain high average ratings, while genres like Adventure and Drama evolved in compelling ways. The results offer valuable insights for industry professionals and movie enthusiasts alike.

Brief Methodology:
We began by web-scraping a large list of movies' genres, directors, cast, and runtimes and enriched it with additional information such as IMDb ratings using the OMDb APIs. The raw data contained inconsistencies like multi-valued cells, missing entries, and irregular formats. We cleaned and standardised the dataset—separating nested fields (e.g., genre, cast), handling missing values, and converting durations into numeric formats. We then created multiple versions of the dataset for analysis by cast, director, and genre. Exploratory Data Analysis (EDA) was followed, using visualisations to uncover trends in movie counts, ratings, runtimes, and career consistency across actors and directors.

Introduction: Context & Relevance

In an age of streaming wars and franchise fatigue, we set out to analyse film history with data. This project doesn't just ask which movies are "good"—it examines how and why certain eras, directors, genres, and formats capture audience appreciation. Our dataset—scraped from Flickchart, enhanced via API, and refined through multiple transformation layers—allowed us to observe long-term trends in production, genre preferences, runtimes, and critical acclaim. These insights are valuable for filmmakers, streaming platforms, and fans of cinema.

Data Wrangling Process *[Appendix]*

We began with raw data scraped from Flickchart and enhanced it using OMDb APIs. The datasets contained messy entries: multiple genres and cast members packed into single cells, inconsistent runtimes, and missing genres and ratings. Assigned sequel numbers and title group IDs to each movie within a series or reboot cluster for structured comparison. Our cleaning process involved:

- Splitting composite columns (e.g., separating cast and genre entries)
- Diagnosed and corrected a data misalignment on page 30 of the scraped genre data by manually inserting a missing entry ("Comedy • Comedy Drama") and realigning subsequent rows to ensure accurate sequencing.
- Removing missing values (NAs) from ratings
- Standardising formatting (e.g., trimming whitespace, correcting year entries)
- Creating per-entity datasets (by cast, director, genre) for granular analysis

The cleaned dataset (cleaned_dataset_master.csv) enabled us to analyse trends across various dimensions.

Columns Retained & Created:

Column Name	Description
Title	Movie title (cleaned and standardized)
Year	Release year (numeric)
Director	List of directors
Duration	Runtime in minutes
Cast	List of lead actors/actresses
Genre	Cleaned list-column of genres
IMBD Rating	IMDb rating from OMDb

	title	year	director	duration	cast	genre	imdb_rating
1	Rashômon	1950	Akira Kurosawa	88	Toshirô Mifune, Machiko Kyô, Minoru Chiaki	Based-on-20th-Century-Literature • Crime Drama ...	8.2
2	Sunset Blvd.	1950	Billy Wilder	110	William Holden, Gloria Swanson, Nancy Olson	Film Noir • Mystery • Political Thriller	5.2
3	In a Lonely Place	1950	Nicholas Ray	94	Humphrey Bogart, Gloria Grahame, Frank Lovejoy	Drama • Film Noir • Media Satire	7.9
4	All About Eve	1950	Joseph L. Mankiewicz	138	Bette Davis, Anne Baxter, George Sanders	Based-on-Theatre • Comedy • Romance	8.2
5	Harvey	1950	Henry Koster	104	James Stewart, Josephine Hull, Peggy Dow	Drama • Foreign Language Film • Psychological Dr...	7.9
6	The Asphalt Jungle	1950	John Huston	112	Louis Calhern, James Whitmore, Sam Jaffe	Coming-of-Age • Drama • Foreign Language Film	7.8
7	Winchester '73	1950	Anthony Mann	92	Dan Duryea, James Stewart, Shelley Winters	Western	7.6
8	Night and the City	1950	Jules Dassin	96	Richard Widmark, Gene Tierney, Googie Withers	Absurd Comedy • Comedy • Farce	7.8
9	Gun Crazy	1950	Joseph H. Lewis	86	Peggy Cummins, John Dall, Berry Kroeger	Based-on-20th-Century-Literature • Crime Drama ...	7.6
10	Los olvidados	1950	Luis Buñuel	85	Estela Inda, Miguel Inclán, Alfonso Mejía	Action Thriller • Action • Based-on-20th-Century-...	6.0
11	Born Yesterday	1950	George Cukor	103	Judy Holliday, Broderick Crawford, William Holden	Based-on-20th-Century-Literature • Comedy Thrill...	7.5
12	Orpheus	1950	Jean Cocteau	95	Jean Marais, François Perier, Maria Casarès	Drama • Paranoid Thriller • Psychological Thriller	7.9
13	The Gunfighter	1950	Henry King	85	Gregory Peck, Helen Westcott, Karl Malden	Comedy Drama • Coming-of-Age • Cross-Dressin...	7.7
14	Rio Grande	1950	John Ford	105	John Wayne, Maureen O'Hara, Ben Johnson	Horror • Monster Film • Survival Film	7.0
15	Cinderella	1950	Clyde Geronimi	74	Ilene Woods, Verna Felton, Don Barclay	Comedy Western • Foreign Language Film • Outlaw	7.3
16	D.O.A.	1950	Rudolph Maté	83	Edmond O'Brien, Pamela Britton, Luther Adler	Action Thriller • Action • Adventure	6.7
17	Where the Sidewalk Ends	1950	Otto Preminger	95	Bert Freed, Gary Merrill, Dana Andrews	Biography • Documentary • Gay and Lesbian Film	7.6
18	Rabbit of Seville	1950	Chuck Jones	7	Mel Blanc, Arthur Q. Bryan	Drama • Ensemble Film • Family Drama	8.3

Key Insights

1. Movie Production Over Time

Production peaked in the 2000s with over 1,600 films(fig1). A sharp decline follows in the 2010s and 2020s, likely due to the COVID-19 pandemic and changing media consumption.

2. Top Directors by IMDb Rating

Directors like Masaki Kobayashi, Sergio Leone, and Christopher Nolan maintain averages above 8.0(fig2), showcasing consistent quality across their works.

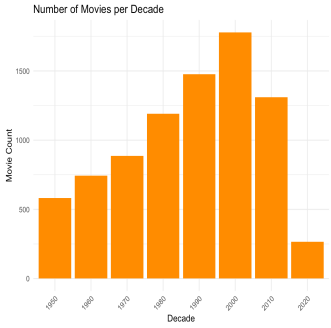

fig1

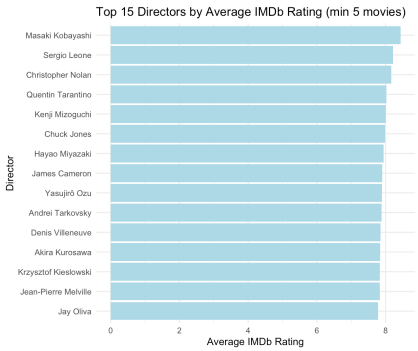

fig2

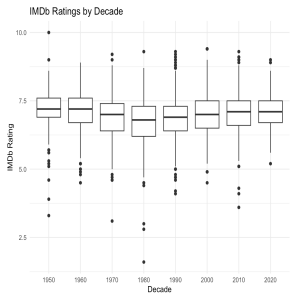

fig3

3. IMDb Ratings by Decade

IMDb ratings have remained stable from the 1950s to 2020s, with most films falling between 6.75 and 7.5. The best-rated (10/10) movie came from the 1950s, while the lowest-rated (2/10) emerged in the 1980s.(fig3)

4. Genre Evolution

- Drama dominates across decades.
- Action and Comedy surged after the 1980s.
- Adventure and True-Story genres rose post-2000.
- The 2020s show a decline due to incomplete data and fewer releases.(fig4)

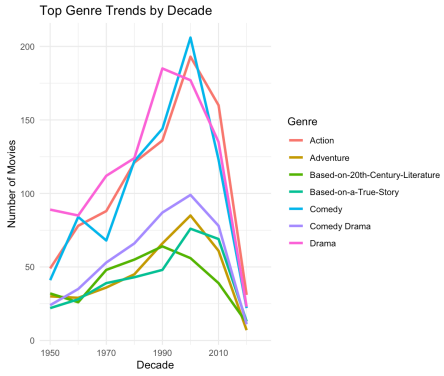

fig4

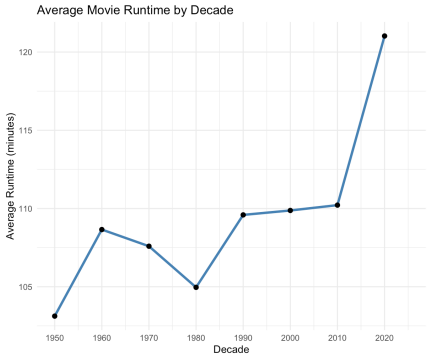

fig5

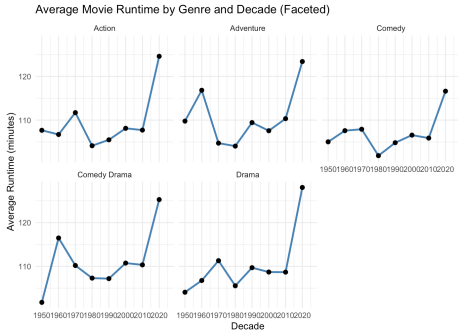

fig6

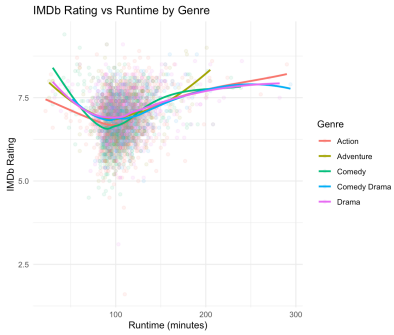

fig7

5. Runtime Trends

Movie runtimes remained stable (~105–110 mins) from 1950–2010,(fig5) then jumped past 120 mins in the 2020s, likely due to streaming freedom and longer narratives in franchises.

6. Runtime by Genre

All genres saw an uptick in runtime in the 2020s(fig6). Drama and Comedy Drama showed the highest runtime increases, aligning with deeper emotional storytelling.

7. IMDb Rating vs Runtime

- Adventure films perform better with longer runtimes (>150 mins).
- Comedy does worse beyond 90 mins—audiences may experience humour fatigue.
- Drama peaks around 180 mins but flattens after.(fig7)
- Comedy Drama blends both emotional depth and humour, improving with duration.

8. Actors: Frequency vs Quality

Actors like Robert De Niro and Nicolas Cage are prolific(fig9). However, when ranked by IMDb average, international actors shine, likely due to fewer but high-quality roles.(fig11)

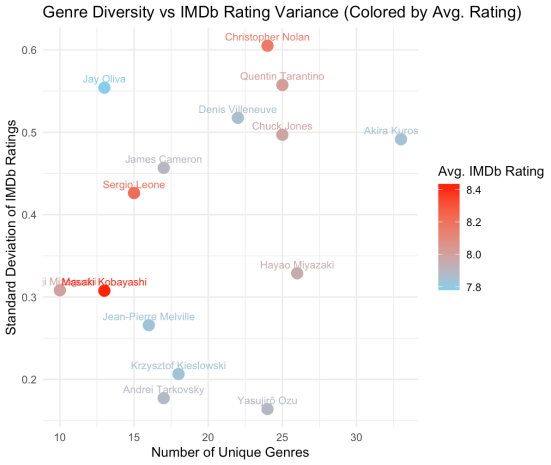

fig8

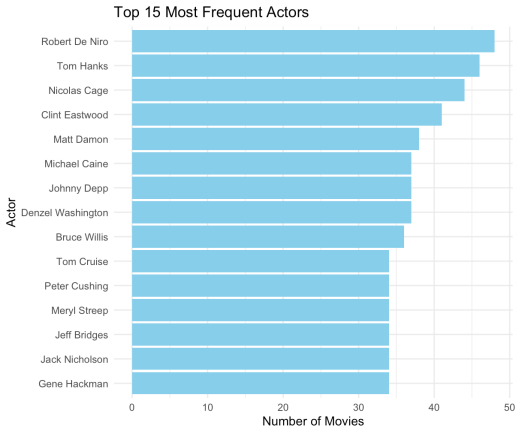

fig9

9. Director-Genre Strength

Top directors have specific genres where they consistently perform well(fig8). For example:

- Nolan in thrillers
- Miyazaki in animation
- Tarantino in crime

10. Genre Diversity of Top 15 Directors

Akira Kurosawa leads with the widest genre diversity(fig10), followed by Hayao Miyazaki and Quentin Tarantino. This shows the adaptability and broad storytelling range of top directors.

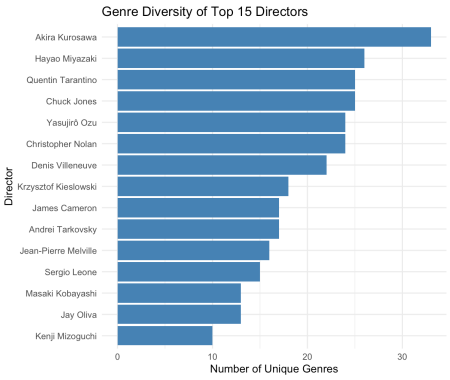

fig10

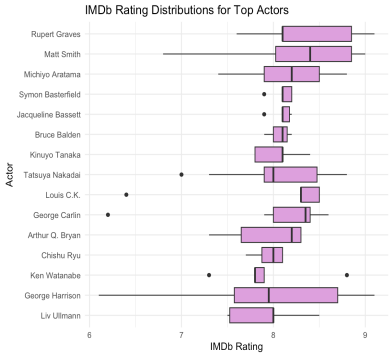

fig11

11. Genre Diversity vs IMDb Rating Variance

This scatter plot(fig8) reveals that directors with wider genre experience don't always have high rating variance. For instance, Masaki Kobayashi maintains both high ratings and low variance—suggesting consistent excellence.

fig12

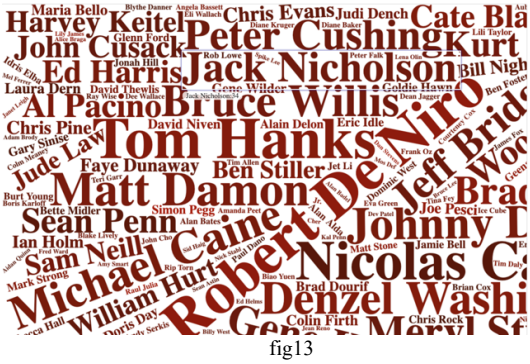

fig13

Interpretation of Director vs. Actor Trends

An intriguing distinction emerges when comparing the trends in directorial versus acting careers using visualizations like word clouds and IMDb rating distributions.

Directors:
We observe that director names predominantly appear in red, signifying that many of the most prolific directors not only directed a large number of films but also maintained consistently high IMDb ratings. This suggests a strong correlation between directorial experience and critical success. Sustained success in directing appears to require a high level of artistic vision, control over production, and storytelling acumen—factors that are directly reflected in viewer ratings.(fig12)

Actors:
In contrast, the actor word cloud reveals a different trend. Some actors with very high movie counts appear in blue, indicating lower average IMDb ratings despite their extensive filmography. This suggests that while many actors remain consistently in demand, their films do not always achieve high critical acclaim. Factors such as popularity, charisma, or typecasting may influence casting decisions more heavily than their track record of critical success.(fig13)

These contrasting trends imply that in the film industry, sustained critical success is more central to a director's legacy, whereas actors may sustain prolific careers through popularity or broad appeal, even without consistent critical acclaim.

Conclusion

This project began as a quest to clean messy movie data. But it evolved into a cinematic voyage through decades of creativity, innovation, and storytelling. From vintage classics to modern masterpieces, the data tells a story of changing tastes, rising directors, and evolving formats.

We leave this project not just with insights, but a deeper appreciation for the craft of filmmaking—and the power of data to illuminate art.

Business or Practical Implications –

Content Acquisition & Curation:

- Streaming platforms can prioritize high-performing genres, directors and actors when making licensing or production decisions.

Recommendation Engines:

- The enriched dataset can serve as a base for personalized movie recommendations or clustering based on user preferences.

Production Insights:

- Studios might consider the identified optimal duration and genre trends when planning new releases for better audience reception.

References –

Data sources:

- Flickchart - https://www.flickchart.com
- OMDB API - https://www.omdbapi.com
- IMDB ratings were accessed via OMDB API but originally sourced from https://www.imdb.com.

R Tools and Packages Used:

- rvest - For web scraping HTML content from Flickchart.
- httr - For making HTTP requests to the OMDB API.
- jsonlite - To parse JSON responses from API calls.
- tidyverse - For data manipulation and cleaning.
- stringr - For string normalization and parsing.
- ggplot2 - For data visualization (histograms, boxplots, etc.).

GitHub:

- Used for version control, project organization, and code sharing.
- Repository platform: https://github.com

Appendix (Technical Details & Code):

- Data Cleaning Process – Full details on transformations, missing value handling, etc.
- Code Snippets – Well-commented Python or R code used in the analysis.
- Additional Visualizations – Any extra plots that support the findings.