Data Wrangling on Movie dataset: Web Scraping, IMDB API Integration, and Visualization with Key Insights

Group 5 –

Sahil Parab (sp2627)
Sai Lohith Chimbili (sc2975)

Course Name –

16:954:597:01 - Data Wrangling and Husbandry with R

Professor Name –

Stevenson Bolivar-Atuesta

Submission Date –

9th May 2025

Executive Summary –

Context of the Project –

With the increasing consumption of online content, movie recommendation systems and analysis tools rely heavily on clean and integrated movie datasets. However, such datasets are often either incomplete or behind paywalls. This project aims to build a comprehensive, enriched movie dataset by combining publicly available data sources.

Data Sources –

Flickchart:

Used as the primary source for metadata of 10,000 movies including title, release year, runtime, genre, director, and cast, due to its broad coverage and unique ranking-based listing.

OMDB API:

Chosen to supplement the dataset with IMDb ratings for ~8,200 movies, as IMDb is a widely trusted benchmark for movie popularity and quality.

Main Objective –

To create a cleaned and enriched dataset of movies released between 1950 and 2024 that consolidates key metadata and IMDb ratings, enabling future analysis or machine learning tasks such as recommendation systems, popularity trend analysis, or genre-based clustering.

Key Insights –

- Successfully matched IMDb ratings for over 80% of the scraped movies, indicating a strong overlap and compatibility between Flickchart and OMDB data.
- Significant variability in movie durations across genres, which could be used to model genre-specific patterns.
- Certain directors and cast members consistently appear in top-rated movies, showing potential for predictive modelling based on contributors.

Brief Methodology –

Web Scraping:

Used R packages to scrape movie details (title, year, duration, genre, director, cast) from Flickchart for 10,000 movies.

Data Cleaning:

Performed data wrangling to normalize column formats, handle missing or malformed entries, and remove duplicates.

API Integration:

Fetched IMDb ratings for ~8,200 movies using the OMDb API. Implemented title-based matching to ensure alignment.

Dataset Consolidation:

Merged API results with scraped metadata to produce a clean, enriched dataset ready for exploratory analysis or modelling.

Introduction: Context & Project Relevance –

Problem Statement –

The project addresses the challenge of integrating fragmented movie information across different sources into a unified, high-quality dataset which provides both detailed metadata and standardized ratings in a clean, analysis-ready format.

Project Relevance –

- Clean and enriched movie datasets are essential for data scientists, media analysts, and developers building recommendation engines, sentiment models, or trend analyses.
- This project streamlines data acquisition and preparation, making it easier for such stakeholders to derive insights without spending excessive time on preprocessing.

Dataset Overview –

Primary Source – Flickchart:

Scraped metadata for 10,000 movies, including title, release year, duration, genre, director, and cast. Chosen for its extensive and user-ranked database.
[Appendix A.1]

Secondary Source – OMDb API:

Retrieved IMDb ratings for ~8,200 of these movies using API calls based on title and year. IMDb is a widely trusted rating source.
[Appendix A.2]

Initial Observations:

The scraped data varied in structure and completeness; cleaning was required to handle missing fields and standardize formats. A significant number of movies had valid and retrievable IMDb ratings, enabling meaningful integration.

Goal of the Analysis –

- To construct a comprehensive, cleaned, and enriched movie dataset for the movies released between 1950 and 2024 by combining Flickchart metadata with IMDb ratings.
- The expected outcome is a structured dataset that can support exploratory data analysis, trend identification, and machine learning tasks in the movie domain.

Data Wrangling & Cleaning –

Initial State of the Data –

Flickchart Scraped Dataset (Raw):

- Size: 10,000 movies
- Issues Identified:
    - Genres and cast listed as unstructured strings or with nested tags.
    - One of movie was missing genre.
    - Cast data had unwanted control characters and white spaces.
    - Cast data has string ‘Starring:’.
    - Duration column is not in numeric format.

OMDb API Dataset (Raw):

- Size: ~8,200 successful matches
- Issues Identified:
    - Missing or null values for some IMDb ratings.

Cleaning Process –

- Manually inserted missing genre and shifted subsequent rows down.
- Removed unwanted control characters: carriage returns, new lines, tabs.
- Removed extra white spaces from start, end, and collapsed multiple spaces in between.
- Extracted everything that comes after the phrase "Starring: ".
- Cleaned the 'duration' column to make it numeric-ready.

[Appendix A.1.6, A.1.8, A.1.9]

Data Merging and Integration –

- Each metadata column from Flikchart was web scraped separately, cleaned and merged directly using column bind.
- Joined the cleaned Flickchart and OMDbAPI datasets using title as primary keys.

[Appendix A.1, A.2]

Final Cleaned Dataset –

Size: ~8,200 fully enriched movie records

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

[Appendix A.4]

Exploratory Data Analysis (EDA) –

Overview of the Data After Cleaning –

After cleaning and integration:

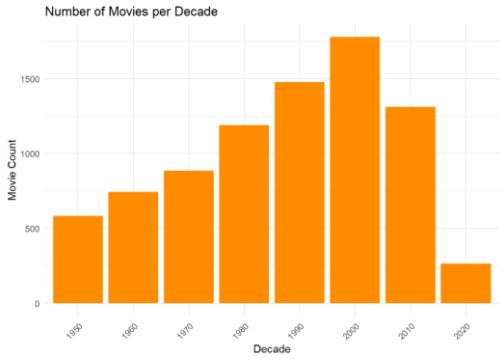
- The dataset used is a cleaned and integrated version of the original data, with all missing values (NAs) removed to ensure completeness.
- It consists of seven columns, among which the 'cast', 'directors', and 'genre' columns may contain multiple values. For accurate analysis—particularly when evaluating individual actors, directors, or genres—these multi-valued entries will be properly separated.

[Appendix A.4]

Visualizations and Insights –

[Appendix B]

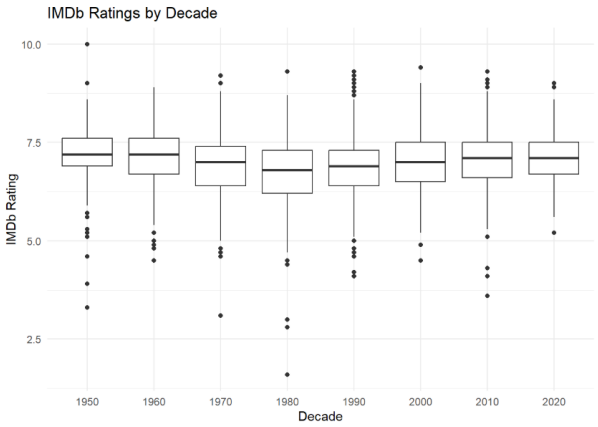
Number of Movies per decade:



- The number of movies steadily increased from the 1950s to the 2000s, peaking around 2000, before declining sharply in the 2010s and further in the 2020s.
- The lower count for the 2020s may be due to the decade still being in progress.

[Appendix B.1]

IMDB ratings per decade:

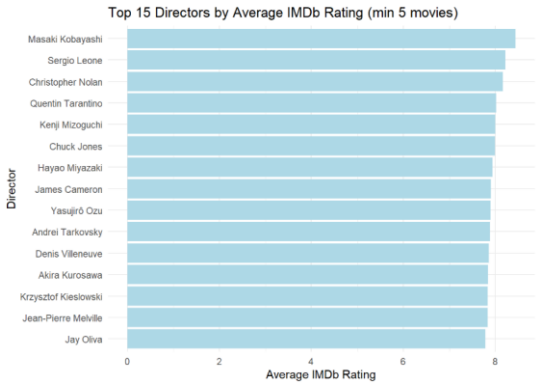


- IMDB ratings have remained relatively steady from the 1950s to the 2020s, with 50% of the ratings consistently falling between 6.75 and 7.5 across decades, though each decade has its share of outliers.
- This stable range suggests that, on average, there has been neither significant improvement nor decline in filmmaking quality over the past 70 years. The highest-rated movie observed—a perfect 10—appeared in the 1950s, while the lowest-rated, at 2 out of 10, came from the 1980s.

- Notably, movies from the 1980s appear to have received a weaker reception overall, as reflected by the generally lower placement of their boxplot compared to other decades.

[Appendix B.2]

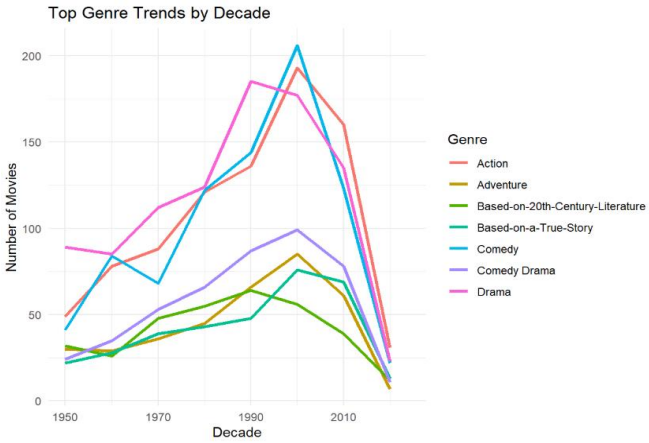
Top Directors by Average IMDB Rating:



- Masaki Kobayashi leads the list with the highest average IMDb rating — just above 8.0 — suggesting strong consistency and audience acclaim across his works.
- Sergio Leone, and Christopher Nolan follow closely, reinforcing their reputation for producing critically acclaimed films with wide audience appeal.
- All directors in this list maintain average IMDb ratings above ~7.8, indicating that their filmographies are not just one-hit wonders but consistently well-received.

[Appendix B.3]

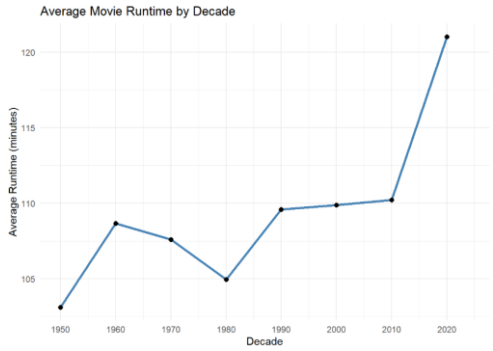
Top Genre trends over the Decades:



- Drama dominates: Drama has been the most consistently produced genre, especially peaking in the 1990s and 2000s. It shows a steady upward trend from 1960s to 1990s, followed by a sharp decline in the 2020s (likely due to incomplete data or limited releases post-2020).
- Action & Comedy surged post-1970: Both genres experience explosive growth starting in the 1970s, peaking around the 2000s–2010s. Reflects the rise of blockbusters, franchise films, and global streaming appeal.
- Based-on-a-True-Story and Adventure grows steadily: Real-life inspired films and Adventure gained popularity in the post-1980s era, likely due to more biopics and historical dramas and rise of enthusiasm for adventure.
- Based-on-20th-Century-Literature remains steady but modest: Remains a niche genre with stable, lower production volume across decades.
- Comedy Drama: Comedy had a stable and a visible rise in the 1950s–2000s, indicating a trend toward hybrid emotional storytelling.
- Sharp drop after 2010: The decline across all genres in the 2020s is likely due to:
    - Fewer releases post-2020 (pandemic impact).
    - Limited dataset coverage for recent years.

[Appendix B.4]

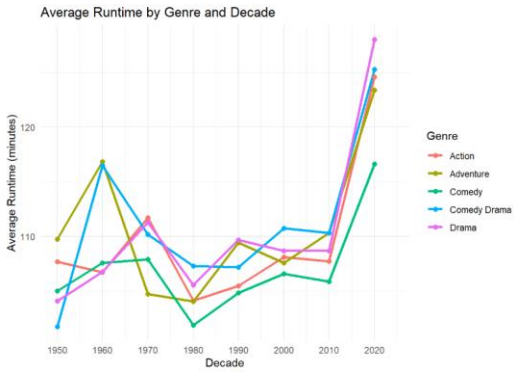
Average Movie Runtime by Decade:



- Stabilization from 1950s–2000s (~100–110 mins): For about 6 decades, average runtime stays relatively stable. Hollywood and global films found an optimal storytelling length here.
- Modern Increase in the 2020s (~120 mins): Likely be attributed to several factors: streaming platforms enabling longer formats, the rise of superhero franchises, epic dramas, and multi-part stories, as well as reduced theatrical constraints, which offer directors greater creative freedom.

[Appendix B.5]

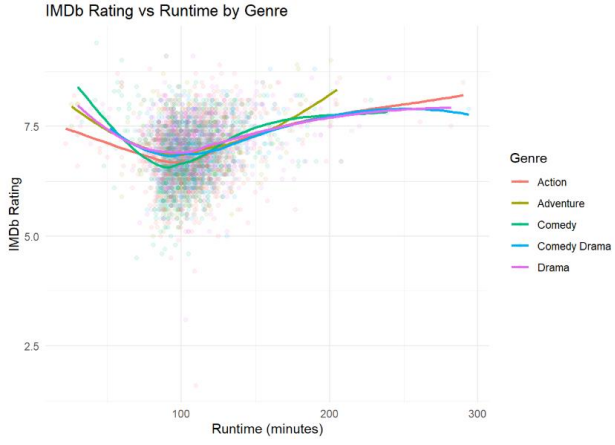
Average Runtime by Genre and Decade:



- All genres show stable runtime over the decades except for 1960s where we see a sudden spike and finally a notable spike in the 2020s.
- The rise may be due to:
  - Streaming platforms enabling more flexible formats.
  - Increasing viewer tolerance for long-form storytelling.
  - Bigger budgets, multi-part sagas, and extended cut.

[Appendix B.6]

IMDB Rating vs Average Runtime by Genre:

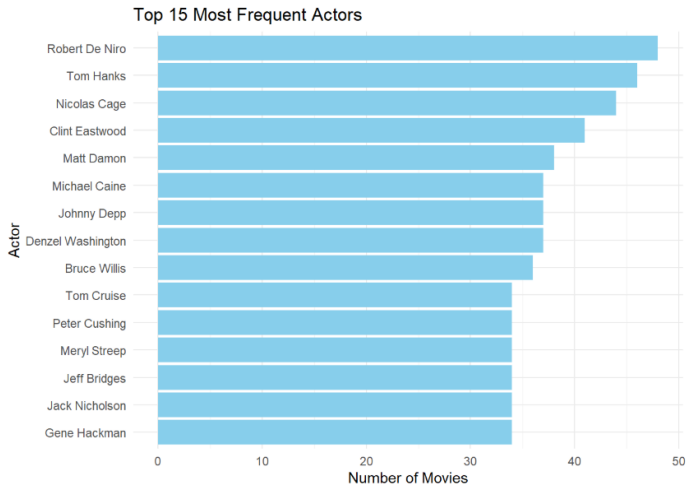


- Adventure: Longer Adventure movies tend to receive significantly higher ratings, especially beyond the 150-minute mark. Suggests that audiences reward deeper, expansive storytelling typical of epic adventures.

- Comedy: Shorter comedies (<90 mins) tend to get higher ratings. There's a sudden drop in ratings if movie drags on beyond 90 mins mark which might be because of sturation of humour. A subtle improvement in ratings with increasing length, but still flats around 7.5—implying that comedy success is more about content than length.
- Comedy Drama: Shows a similar pattern to Comedy, but overall has higher ratings across the board. Ratings increase with length, especially beyond 200 mins, showing the value of combining emotional weight with humor.
- Action: Longer Action movies tend to receive significantly higher ratings as the duration increases.
- Drama: Fairly stable trend, with moderate increases until ~180 mins. Beyond that, ratings stabilize or dip, possibly reflecting diminishing returns for overly long dramas.

[Appendix B.7]

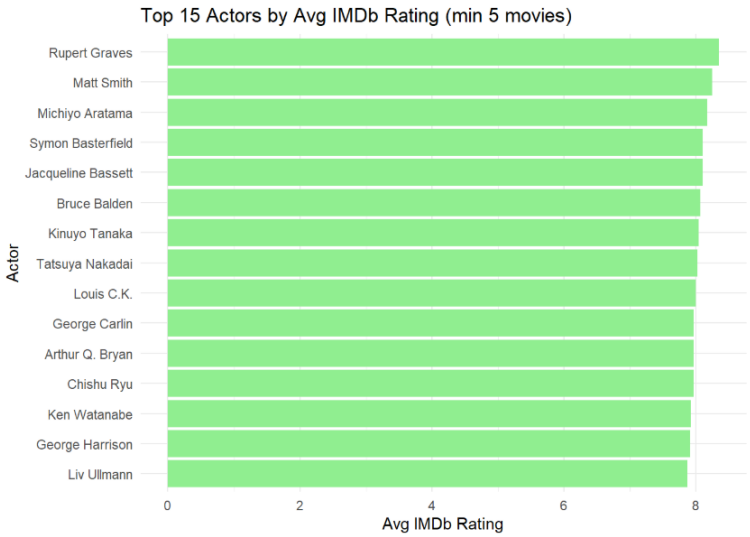
Top 15 most Frequent Actors:



- Robert De Niro, Tom Hanks, Nicolas Cage, and Clint Eastwood dominate the charts, each appearing in over 40 movies between the 1950s and 2020s.
- Notably, the list is dominated by Hollywood actors—which is not surprising given Hollywood's historically high volume of film production compared to international film industries.

[Appendix B.8]

Top 15 Actors by Average IMDB Rating:



- An important observation is that when IMDB ratings are used to identify top actors, the charts are largely dominated by actors from international film industries rather than Hollywood.
- A possible explanation is that these international actors appeared in fewer films compared to their Hollywood counterparts, and the limited number of roles they took on were rated more highly—resulting in a higher average IMDB rating.

[Appendix B.9]

Key Insights & Discussion –

Summary of Findings –

- Film ratings have become more varied but stabilized in quality since the mid-20th century. Early cinema is underrepresented and may skew low due to lack of data or retrospective evaluation.
- The consistent median ratings from the 1950s onward show that audience expectations and critical standards have levelled out, despite increasing film production.
- A mix of international directors (e.g., Jean-Pierre Melville, Hayao Miyazaki, Tarkovsky) and modern mainstream directors (e.g., Nolan, Villeneuve, Cameron) suggests that excellence in movie directions spans both global cinema and Hollywood.
- Highlights how social tastes, media formats, and production trends have shifted over time. Drama remains the most produced, while Action, Adventure, and true-story adaptations have risen with modern cinema's commercial and emotional trends.
- Longer runtimes tend to correlate with higher ratings for genres like Adventure, Drama, and Comedy Drama. However, extreme runtimes do not necessarily lead to higher ratings and may hurt pacing and engagement.

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.

Challenges Faced –

- Directors, Cast and Genres were combined.
- OMDB API calls had long runtime.
- Unwanted control characters and white spaces.

Limitations & Next Steps –

- IMDB ratings were not available for ~1,800 movies, limiting full dataset analysis. Future work could explore alternative rating sources (e.g., TMDb, Rotten Tomatoes).
- The dataset captures a fixed point in time; integrating time-series data (e.g., ratings over years) could enhance trend analysis.
- Current dataset includes numeric ratings but lacks textual review sentiment, which could offer deeper insights into audience perception.
- Future Work:
    - Perform sentiment analysis using user reviews (if available).
    - Cluster movies using NLP on synopsis or tags.
    - Build a simple recommender system using this enriched dataset.

Conclusions –

Project Workflow Summary –

Web Scraping:

Collected metadata for 10,000 movies from Flickchart using R (rvest, httr), including title, year, genre, director, cast, and duration.

Data Cleaning:

Standardized formats, removed duplicates, handled missing values, and cleaned text fields.

API Integration:

Queried OMDB API to fetch IMDB ratings for ~8,200 movies; handled matching issues and merged with the scraped dataset.

Exploratory Data Analysis:

Conducted EDA to explore distributions, genre-wise trends, and relationships between variables like duration and rating.

Recommendations & Future Work –

- A well-integrated movie dataset can be useful for a variety of stakeholders — data analysts studying media trends, streaming services curating content, or even marketers understanding viewer behaviour.
- By consolidating raw data into a clean format, this project can enable deeper analysis of the film industry, such as identifying genre-based rating trends, comparing movie performance, or examining how features like runtime or release year relate to viewer ratings.

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