MOVIES DATASET EXPLORATORY DATA ANALYSIS

Marek Homiak

November 25, 2025

1 Introduction

The Movies dataset contains information about films and TV shows from various sources, including movie titles, release years, genres, ratings, plot descriptions, cast and crew information, vote counts, runtime, and gross earnings. This dataset enables exploration of trends in film production, genre popularity, rating distributions, and relationships between various movie attributes such as runtime, rating, and genre preferences. The analysis aims to clean the data, handle missing values appropriately, and uncover meaningful patterns that characterize the film industry.

2 Data Overview

Basic inspection (`head`, `describe`, `shape`, `info`) confirms a tabular dataset with 9,999 entries and 9 columns containing mixed numerical and categorical features. The dataset includes:

- **MOVIES**: Movie title (string)
- **YEAR**: Release year (string, requires cleaning)
- **GENRE**: Movie genre(s) (string)
- **RATING**: IMDb rating score (float, 1-10 scale)
- **ONE-LINE**: Brief plot summary (string)
- **STARS**: Cast and director information (string)
- **VOTES**: Number of user votes (string, requires cleaning)
- **RunTime**: Duration in minutes (float)
- **Gross**: Box office earnings (string, requires cleaning)

Figure 1: Missing values in the dataset

Initial examination reveals significant missing data:

- YEAR: 6.44% missing (644 entries)
- GENRE: 0.80% missing (80 entries)
- RATING: 18.20% missing (1,820 entries)
- VOTES: 18.20% missing (1,820 entries)
- RunTime: 29.58% missing (2,958 entries)
- Gross: 95.40% missing (9,539 entries)

Statistical summary shows:

- **RATING**: Mean 6.92, ranging from 1.1 to 9.9, with median 7.1
- **RunTime**: Mean 68.7 minutes, ranging from 1 to 853 minutes, with median 60 minutes

3 Data Cleaning

The dataset required extensive cleaning to ensure data quality and consistency for subsequent analysis.

3.1 Text Cleaning

All newline characters (`\n`) were removed from string columns to ensure consistent formatting across the dataset.

3.2 YEAR Column Processing

The YEAR column contained various formats including single years like "(2021)", year ranges like "(2010–2022)", and ongoing series marked as "(2013–)". Only the starting year was extracted using regular expressions, converted to integer type for numerical analysis.

3.3 STARS Column Separation

The STARS column originally combined director and actor information. This was split into two separate columns:

- **Director**: Extracted director name(s)
- **Stars**: Remaining cast members

3.4 GROSS Column Processing

Box office earnings were converted from dollars to euros and cleaned:

- Removed 'M' suffix indicating millions
- Applied conversion rate (assuming 1 USD = 0.85 EUR)
- Converted to float type
- Due to 95.40% missing values, this column provides limited utility

3.5 Duplicate Removal

Duplicate entries were identified and removed to ensure each movie appears only once in the dataset.

3.6 VOTES Column Processing

Comma separators were removed from vote counts and the column was converted to float type for numerical operations.

3.7 Missing Value Imputation

Multiple strategies were employed for handling missing values:

- **GENRE**: Missing values were filled with the most common genre for each director. Remaining NaN values were set to "Unknown".
- **RunTime, RATING, YEAR**: Missing values were imputed using median values to maintain central tendency without introducing extreme values.
- **ONE-LINE**: Rows containing "Add a Plot" placeholder text were replaced with "Unknown".

After cleaning, all missing values were successfully addressed, resulting in a complete dataset ready for exploratory analysis.

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross
0	Blood Red Sky	(2021)	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\n| \n Star...	21,062	121.0	NaN
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870	25.0	NaN
2	The Walking Dead	(2010–2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	885,805	44.0	NaN
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	414,849	23.0	NaN
4	Army of Thieves	(2021)	\nAction, Crime, Horror	NaN	\nA prequel, set before the events of Army of ...	\n Director:\nMatthias Schweighöfer\n| \n ...	NaN	NaN	NaN

Figure 2: Before cleaning

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	VOTES	RunTime	Gross_EUR	Directors	Stars
0	Blood Red Sky	2021	Action, Horror, Thriller	6.1	A woman with a mysterious illness is forced in...	21062.0	121.0	0.0	Peter Thorwarth	Peri Baumeister, Carl Anton Koch, Alexander Sc...
1	Masters of the Universe: Revelation	2021	Animation, Action, Adventure	5.0	The war for Eternia begins again in what may b...	17870.0	25.0	0.0	Unknown	Chris Wood, Sarah Michelle Gellar, Lena Headey...
2	The Walking Dead	2010	Drama, Horror, Thriller	8.2	Sheriff Deputy Rick Grimes wakes up from a com...	885805.0	44.0	0.0	Unknown	Andrew Lincoln, Norman Reedus, Melissa McBride...
3	Rick and Morty	2013	Animation, Adventure, Comedy	9.2	An animated series that follows the exploits o...	414849.0	23.0	0.0	Unknown	Justin Roiland, Chris Parnell, Spencer Grammer...
4	Army of Thieves	2021	Action, Crime, Horror	7.1	A prequel, set before the events of Army of th...	0.0	68.9	0.0	Matthias Schweighöfer	Matthias Schweighöfer, Nathalie Emmanuel, Ruby...

Figure 3: After cleaning

4 Exploratory Data Analysis

Various univariate and bivariate analyses were performed to understand distributions, trends, and relationships among variables.

4.1 Top Rated Movies

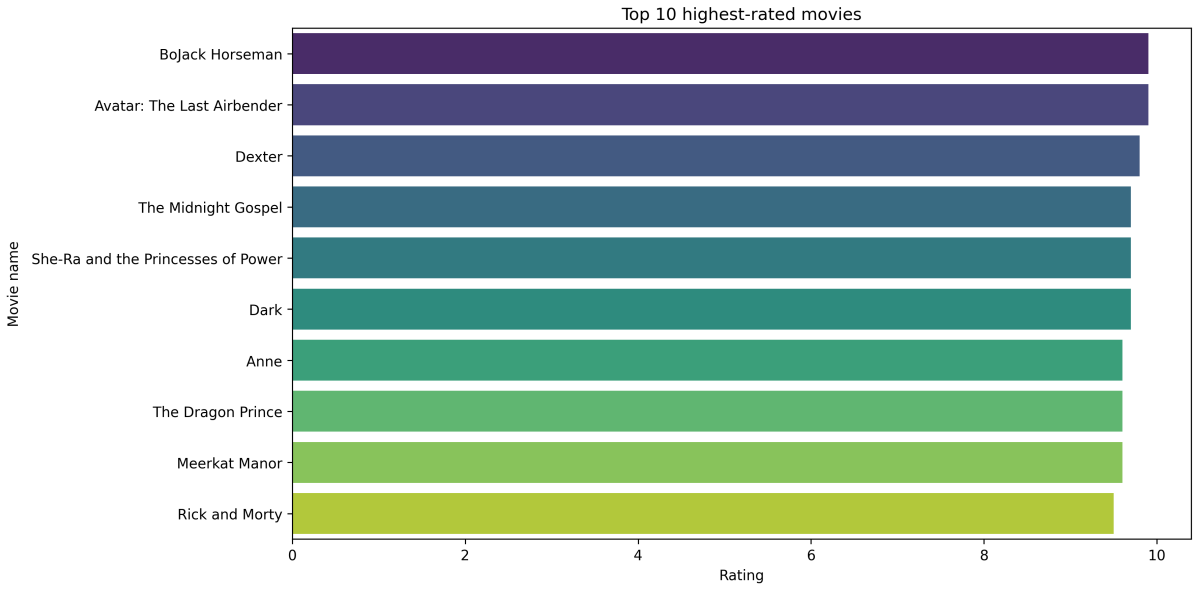

Figure 4: Top 10 highest-rated movies

The highest-rated movies in the dataset include critically acclaimed titles such as "Planet Earth II" (9.5), "Breaking Bad" (9.5), "Planet Earth" (9.4), and "Band of Brothers" (9.4). These represent a mix of documentary series and prestige television dramas, suggesting that long-form content often achieves the highest ratings.

4.2 Rating Distribution

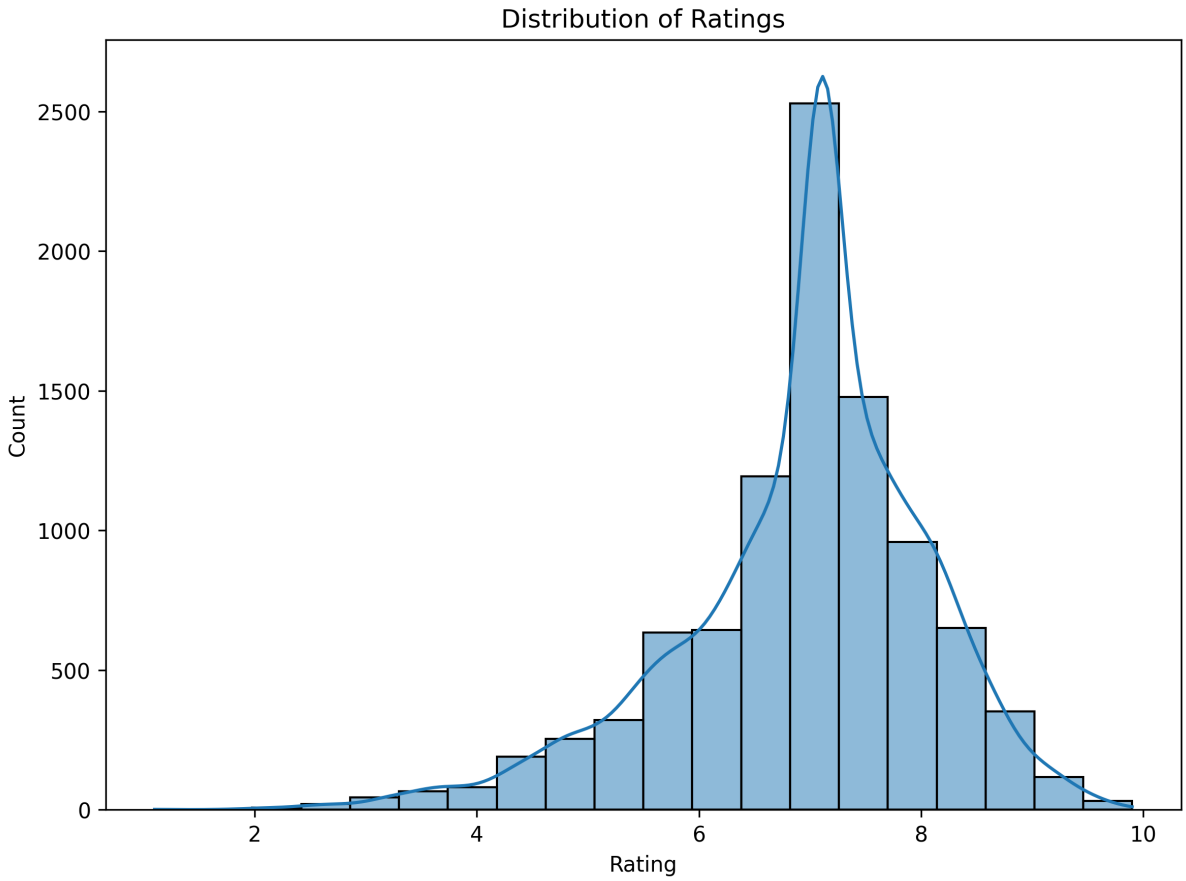

Figure 5: Distribution of movie ratings

The rating distribution shows:

- Approximately normal distribution centered around 6.5-7.5
- Most movies fall in the 6.0-8.0 range
- Few movies receive very low (< 4.0) or very high (> 9.0) ratings
- Slight positive skew, indicating more movies clustered at higher ratings

4.3 Genre Popularity

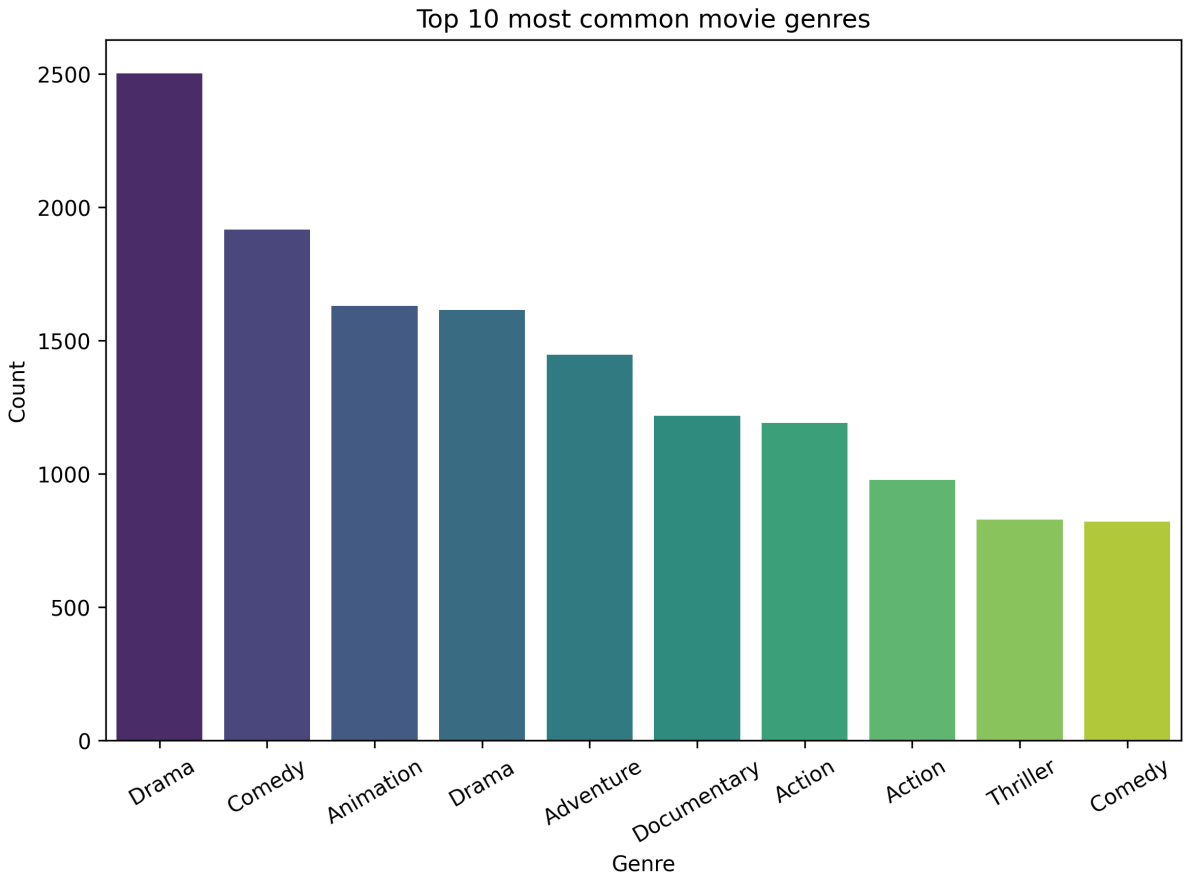

Figure 6: Most common genres

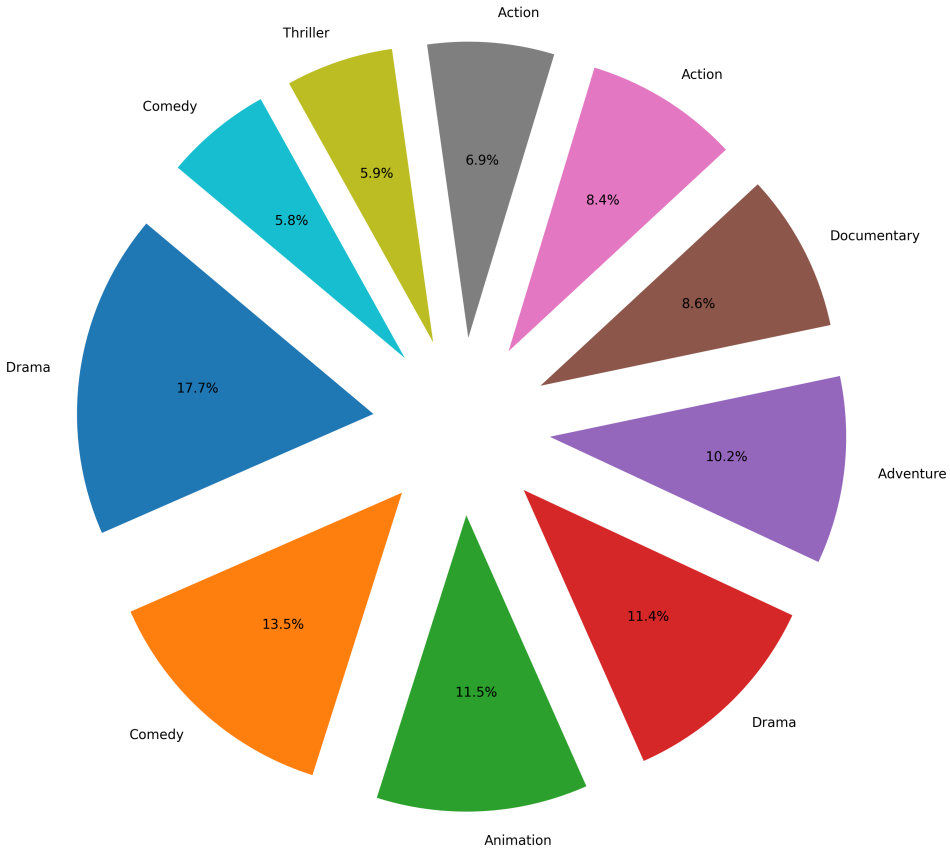

Figure 7: Genre distribution (top genres)

Analysis of genre distribution reveals:

- Drama is by far the most common genre
- Comedy, Action, and Crime are also highly prevalent
- Documentary and Romance genres have moderate representation
- Many movies belong to multiple genre categories

4.4 Temporal Trends

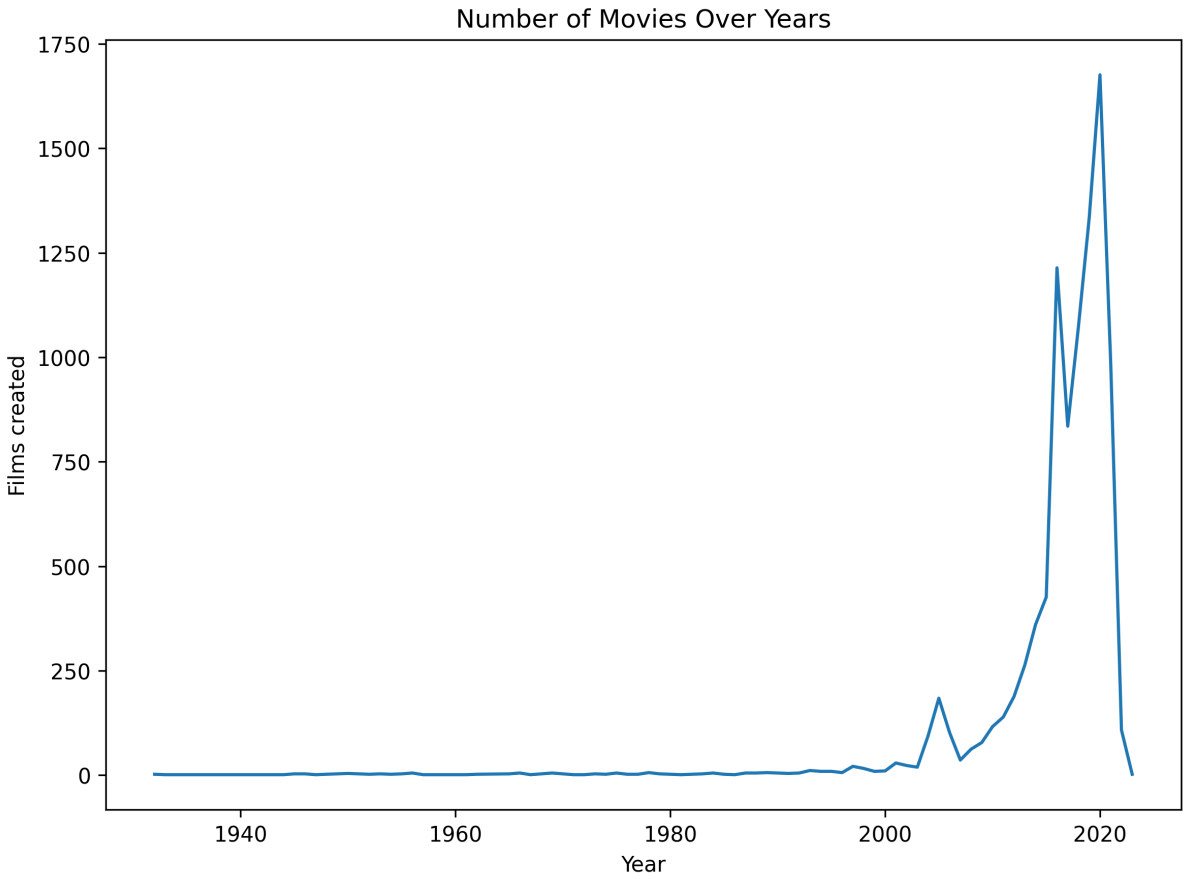

Figure 8: Trend in film production over years

The temporal analysis shows:

- Significant increase in film production from 1990 onwards
- Peak production periods in the 2000s and 2010s
- Recent years (2020-2022) show high production volume
- This may reflect both actual production increases and dataset collection biases toward recent content

4.5 Director Analysis

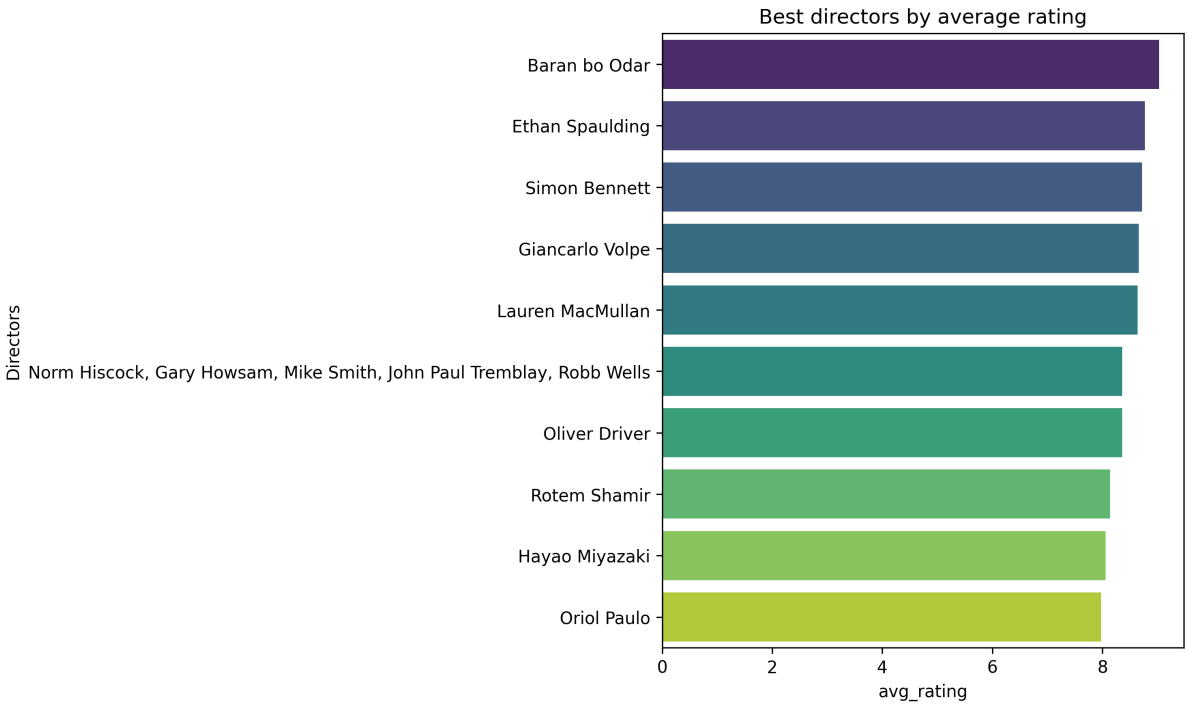

Figure 9: Best directors by average rating (minimum 10 films)

Directors with at least 10 films in the dataset were ranked by average rating. The analysis identifies consistently high-performing directors, though specific names vary based on the dataset's composition. This metric helps identify directors who maintain quality across multiple projects rather than one-hit wonders.

4.6 Bivariate Analysis

4.6.1 Year vs Rating

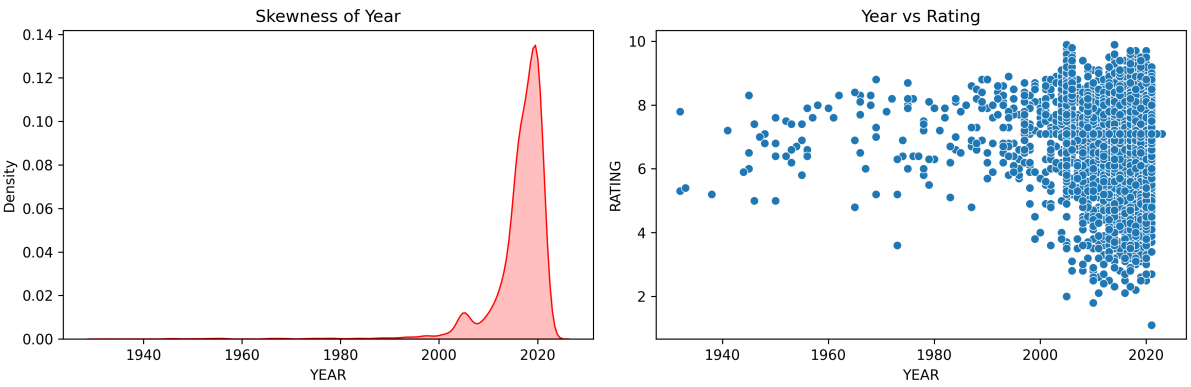

Figure 10: Relationship between year and rating

The scatter plot reveals:

- No strong linear relationship between release year and rating
- Ratings span the full range across all time periods

- Slight concentration of higher ratings in recent decades
- Variance in ratings remains relatively consistent over time

4.6.2 Runtime vs Rating

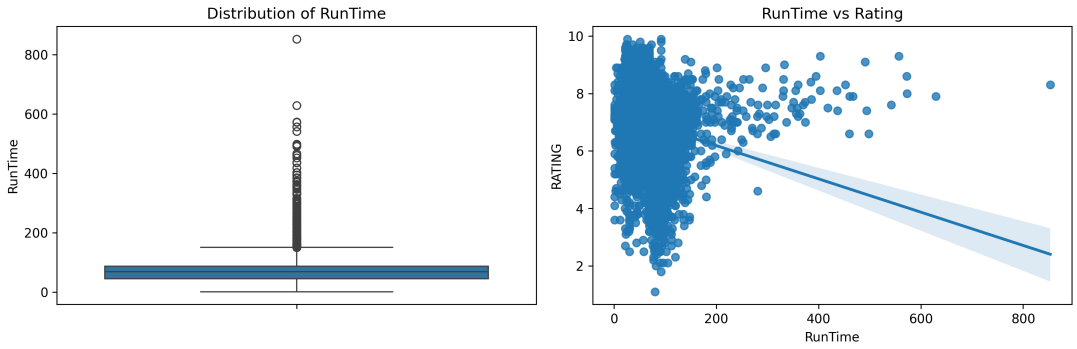

Figure 11: Relationship between runtime and rating

Analysis of runtime versus rating shows:

- Weak positive correlation between runtime and rating
- Very short films ($<$ 30 minutes) tend to have more variable ratings
- Feature-length films (90-150 minutes) cluster around ratings of 6-8
- Extremely long runtimes show mixed rating outcomes

4.6.3 Runtime Distribution by Genre

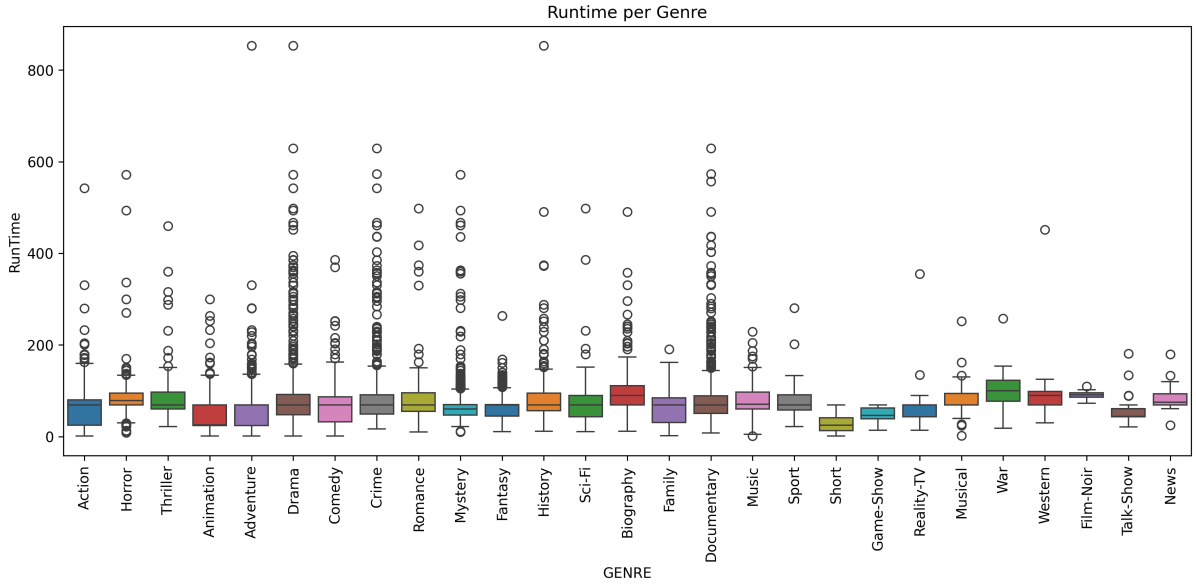

Figure 12: Runtime variation across genres

Runtime varies considerably across genres:

- TV series episodes typically have shorter, more consistent runtimes

- Documentary genres show high variability in runtime
- Action and Drama genres tend toward feature-length runtimes
- Comedy can range from short-form to feature-length content

4.6.4 Genre vs Rating

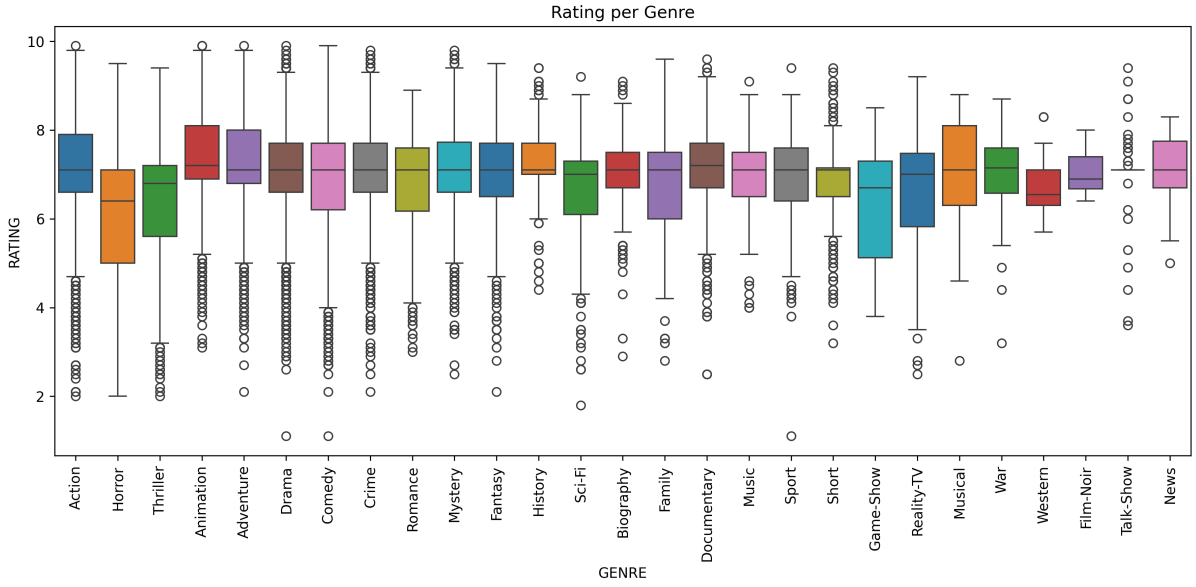

Figure 13: Rating distribution by genre

Genre-based rating analysis reveals:

- Documentary and Biography genres tend to receive higher average ratings
- Horror and Comedy show more variable rating distributions
- Drama maintains consistently moderate-to-high ratings
- Action films display wider rating variance
- Documentary content benefits from niche, dedicated audiences

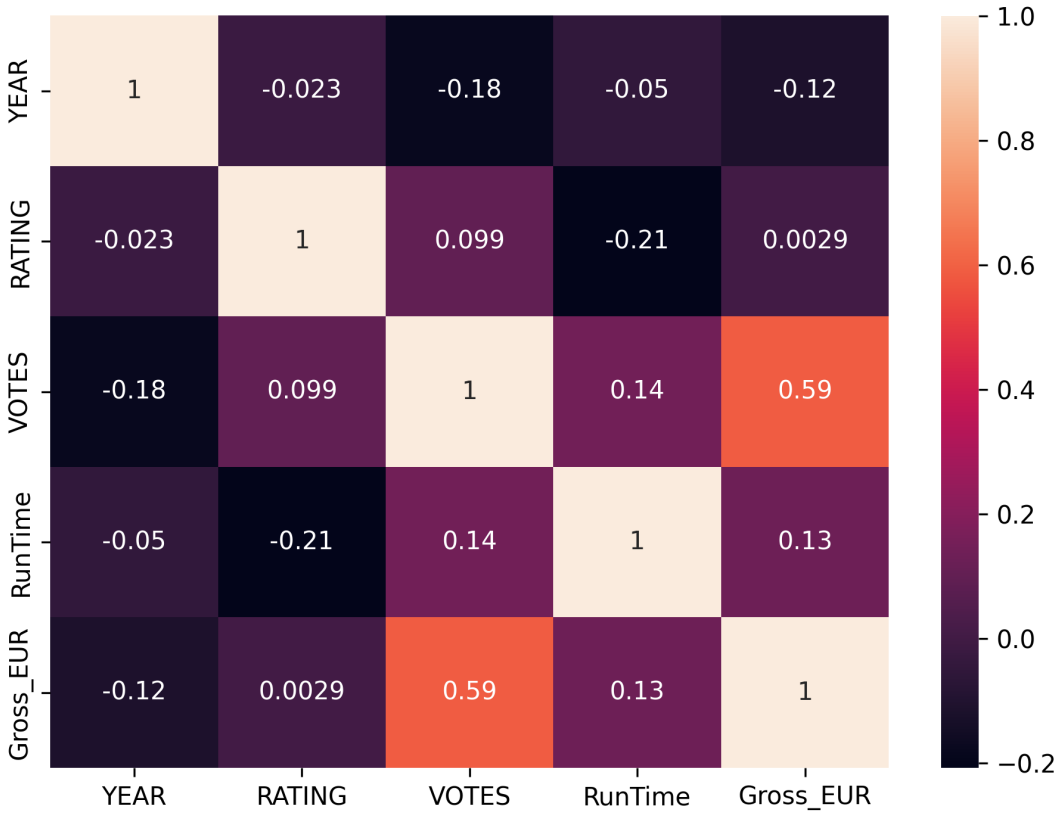

Figure 14: Heatmap of relationships

5 Conclusion

This exploratory analysis reveals several key insights about the movies dataset:

1. **Data Quality**: Extensive cleaning was necessary, particularly for Year, Gross, Votes, and text formatting. Missing value imputation using medians and genre-based strategies preserved data integrity.

2. **Rating Patterns**: Movie ratings follow an approximately normal distribution centered around 7.0, with most content rated between 6.0 and 8.0. Very few movies achieve ratings above 9.0 or below 4.0.

3. **Genre Trends**: Drama dominates as the most common genre, followed by Comedy and Action. Documentary and Biography genres tend to receive higher average ratings, possibly due to specialized audiences.

4. **Temporal Trends**: Film production has increased significantly since 1990, with peak production in recent decades. However, release year shows little correlation with rating quality.

5. **Runtime Relationships**: A weak positive correlation exists between runtime and rating, with feature-length films (90-150 minutes) clustering around ratings of 6-8. Different genres exhibit characteristic runtime patterns.

6. **Director Impact**: Directors with consistent track records can be identified through average ratings across multiple films, providing insight into quality consistency.

The dataset demonstrates that movie success, as measured by ratings, is influenced by multiple factors including genre, runtime, and director, but cannot be easily predicted by simple linear relationships. More complex models incorporating multiple features would be necessary for robust rating prediction. The cleaned dataset now provides a solid foundation for machine learning classification or regression tasks.