Breaking Stats

35 Br Bromine

16 S Sulfur

Statistics 199

3/26/2026

Jerry Barajas-Nunez, Isabel Rozes, Ben Mei-Dan, Jackson Cooke, Hala Mohammed

Topic and Motivation

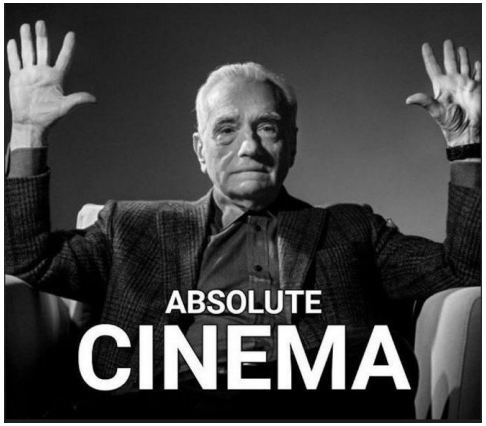
Internet Movie Database (IMDb) is an online platform where users can upload ratings and reviews on films

Ratings are weighted, from registered and frequented users

Cinema represented a $100 Billion industry worldwide

What characteristics of a movie “make it good?”

All of us have passion for movies



Research Question

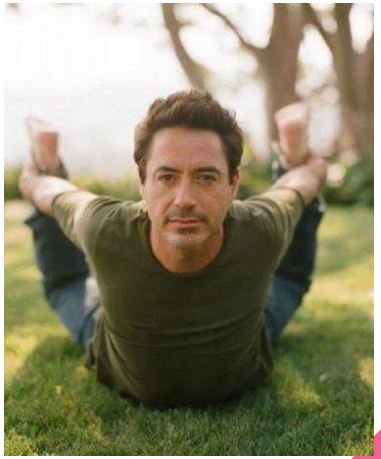
How do the numerical variables of runtime and gross revenue correlate with IMDb rating and how do the categorical variables of primary genre and leading actors influence the IMDb rating of the top 1000 movies?

Introduce the Data

The IMDB dataset: 1000 rows // 16 columns

Variables: poster link, series title, released year, certificate, runtime, genre, IMDB rating, overview, metascore, director, star 1-4, # of votes, and gross revenue.

Cleaning: changing format of all columns (clean_names, janitor package), made runtime numerical, primary genre extracted

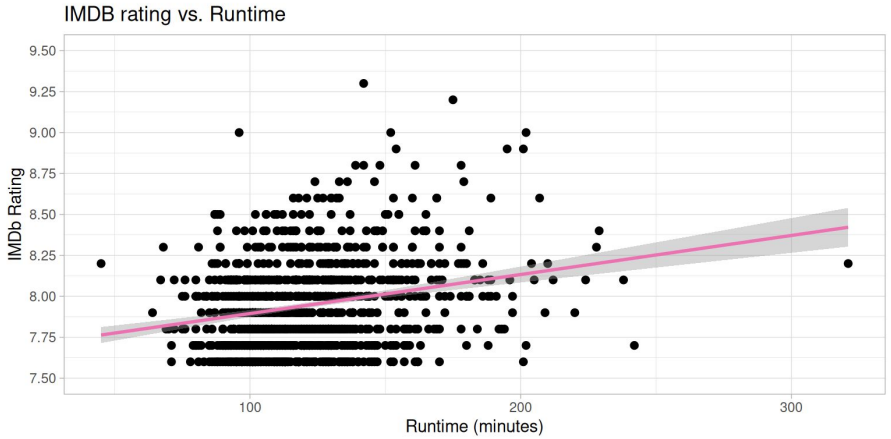


Analysis

We examined four variables to answer what might shape audience evaluations of films.

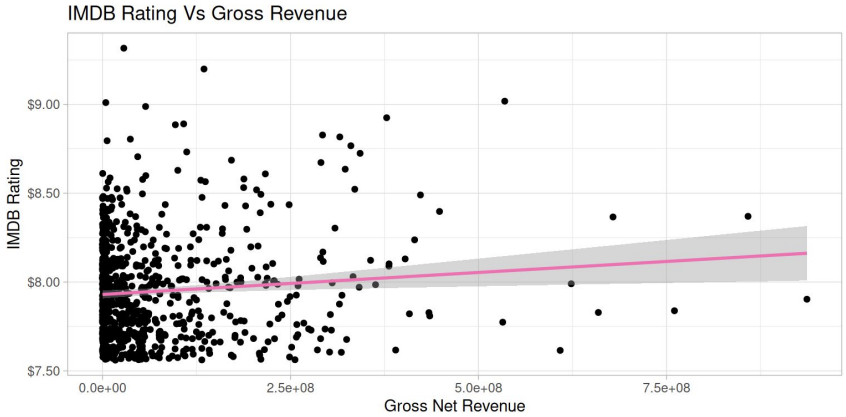
1. Runtime (numerical)
2. Gross Revenue (numerical)
3. Primary Genre (categorical)
4. Actor (categorical)

Runtime



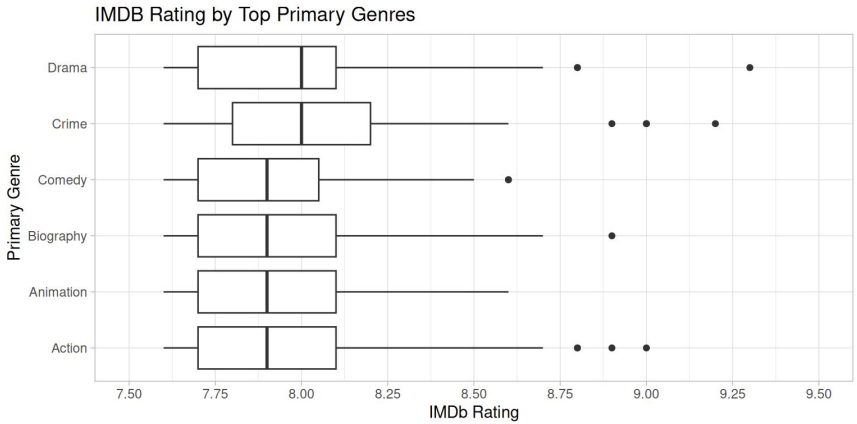
- **Slight positive relationship** r = 0.24
- The fitted line slopes upward suggesting that longer movies received somewhat higher ratings on average.
- Points are widely scattered along the line, with common runtime range of 80 - 180 minutes
- Overall, runtime alone explains little of the variation in ratings.

Gross Revenue



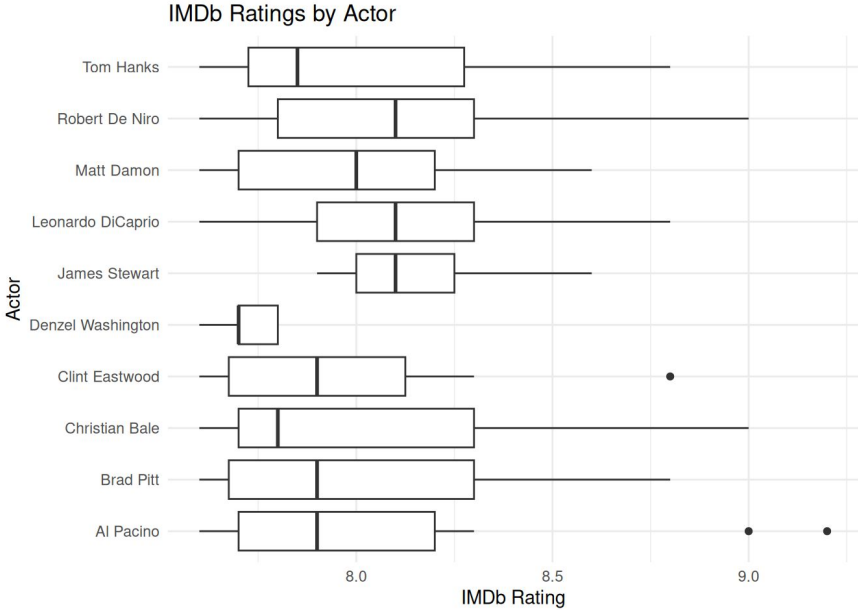
- **Slight positive relationship** r = 0.096
- The smooth line slopes upward, suggesting that higher IMDb ratings are associated with somewhat higher gross revenue on average.
- Points are widely scattered across the plot, with large variation in revenue at nearly every rating level.
- Overall, the relationship between rating and revenue appears weak.

Primary Genre



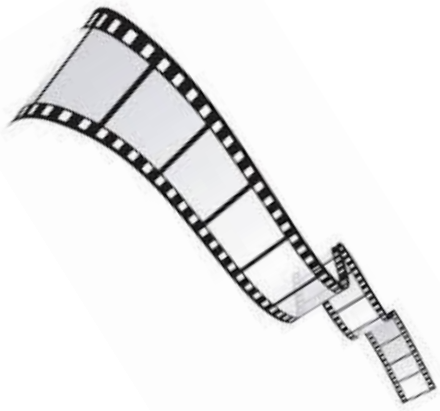
- **Negligible differences**
- Similar distributions
- Medians similar: generally around 7.9 - 8
- Crime and drama are slightly higher than the others

Leading Actor



- **Negligible differences**
- Some actors (James Stewart, Robert De Niro) appear to have slightly higher typical ratings, while others (Al Pacino, Christian Bale, and Brad Pitt) have show more variability in ratings
- Denzel Washington's ratings are the most concentrated
- Leading actor does not seem to strongly determine IMDb rating.

Discussion: Key Takeaways



- **Runtime**
  - Slight positive correlation
  - Longer films score slightly higher, but effect is minimal
- **Gross Revenue**
  - Slight positive correlation
- **Primary Genre**
  - Differences are negligible
- **Leading Actor**
  - Differences are negligible
- **Overall conclusion**
  - Within the IMDb Top 1000 dataset, no single factor strongly predicts a film's rating, suggesting that film quality is multifaceted and cannot be explained by any single variable alone

Limitations and Improvements

- Dataset is restricted to IMDb's top 1000 films,
  - Highly rated: compresses the sample size
- Genre was simplified to only the primary genre, losing information for multi genre films
- Use regression modeling to analyze the combined effect of multiple predictors rather than examining each in isolation
- Use a dataset with a broader range of ratings so that we can better detect relationships

90
Th
Thorium

ank

39
Y
Yttrium

ou!