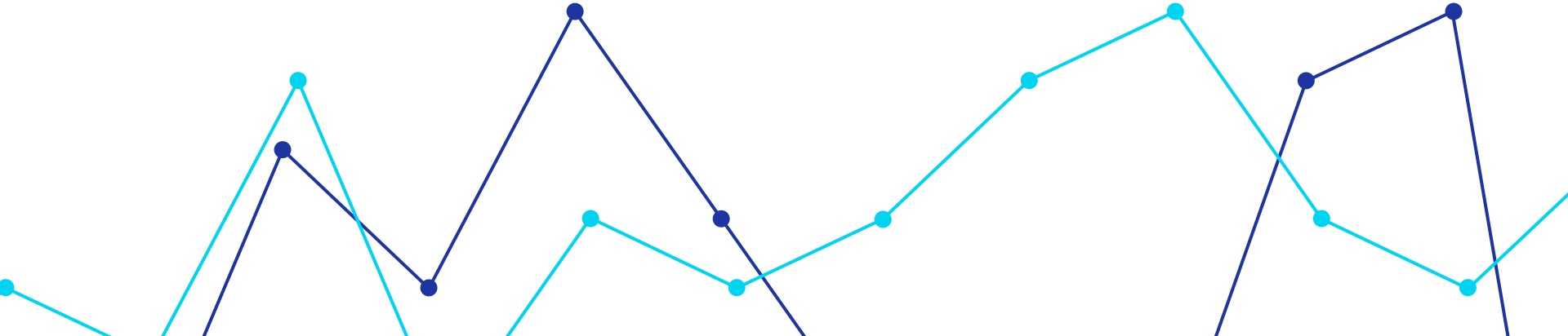
The Impact of Film's Properties on IMDB Ratings : A Logistic Regression Approach

Wanding Wang, Yujie Tang, Weiqing Guo, Bashiru Mukaila, Jialu FU

Group 08



Background

IMDb

- **IMDb is a popular platform for the competitive entertainment industry, with ratings influencing audience perception and revenue.**

Table 1: Highlight of the IMDB data

film_id	year	length	budget	votes	genre	rating
5993	1943	65	15.5	42	Action	7.6
37190	1961	87	12.3	6	Drama	6.0
43646	1987	79	16.4	161	Action	7.5
28476	1976	NA	12.2	5	Documentary	8.0
23975	1982	88	12.5	97	Action	3.5
50170	1936	NA	7.0	146	Drama	4.4

Aims

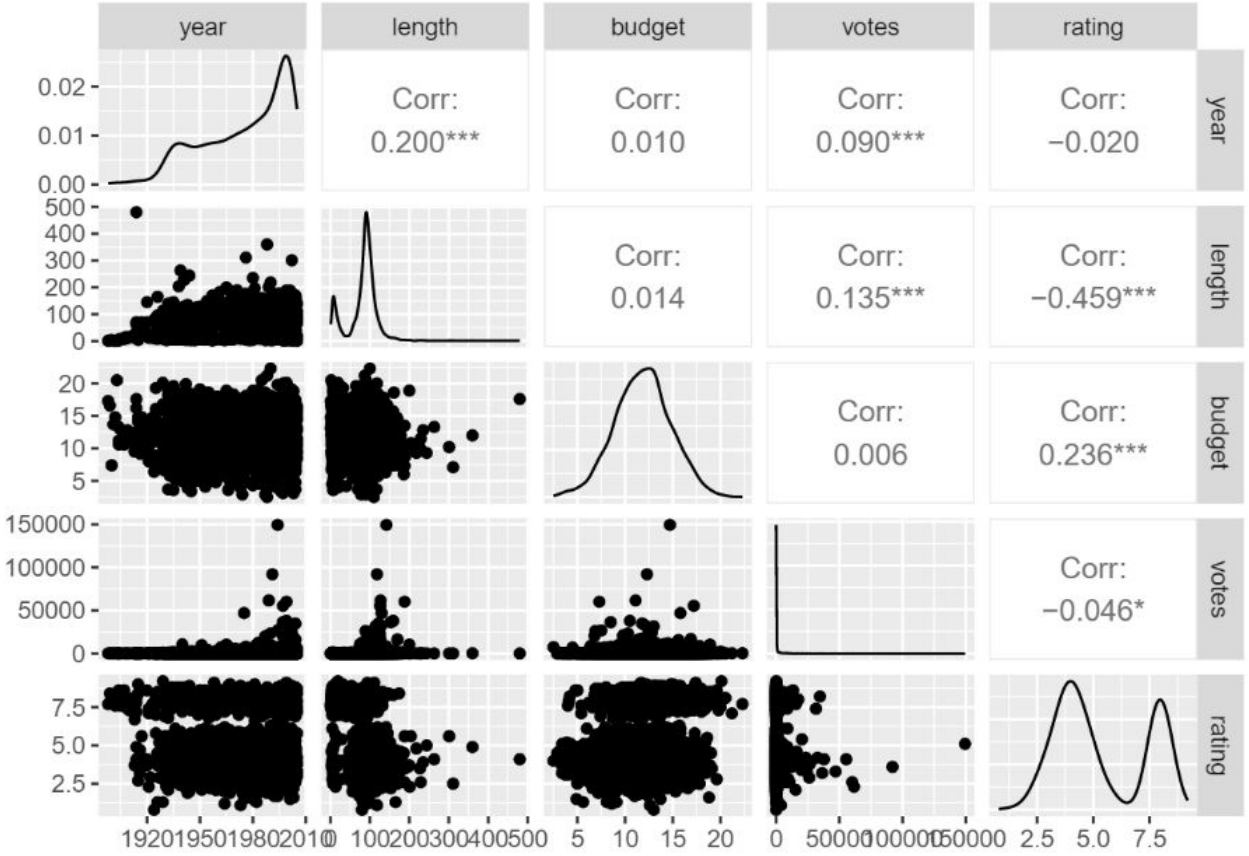
Main purpose

- **Explore the relationship between the properties of films and the IMDB ratings.**

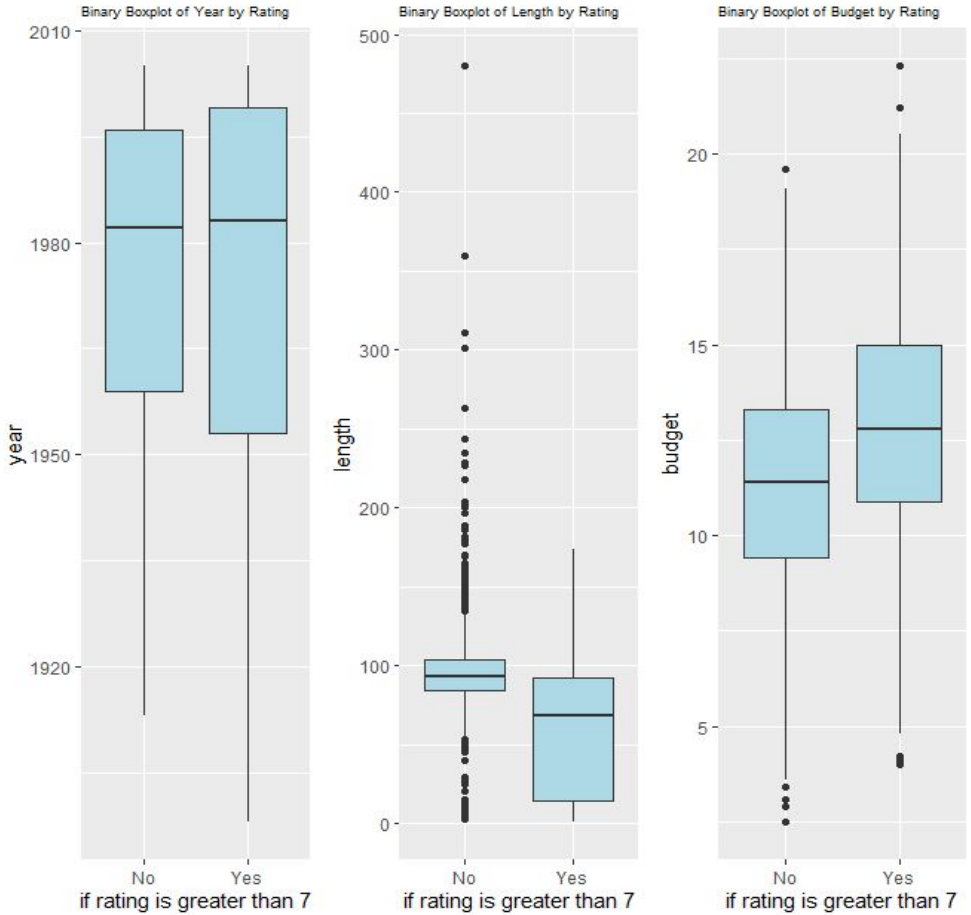
Research Question

- **Investigate the specific properties of a movie that have a significant influence on whether it receives an IMDB rating of 7 or higher.**

Exploratory Analysis – Scatterplot Matrix of the Numerical Variables

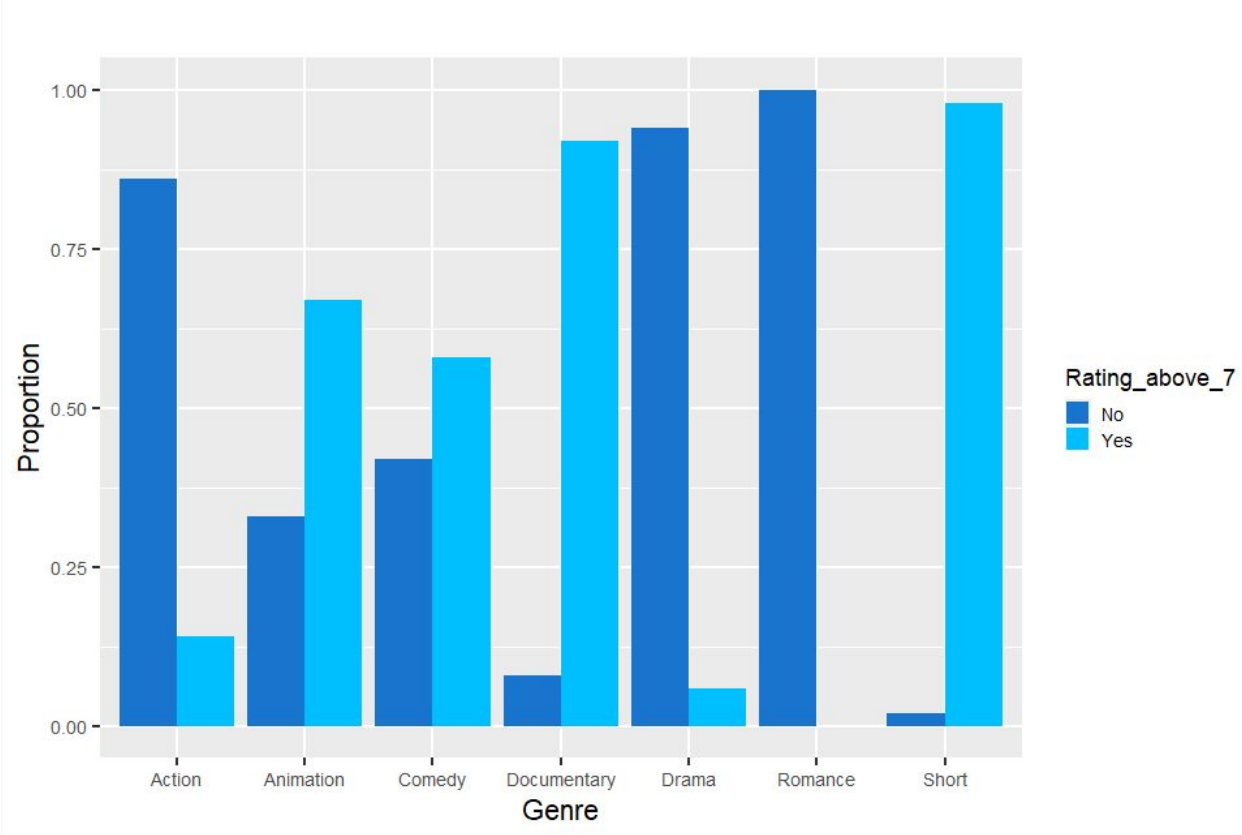


Exploratory Analysis – Binary Boxplots of Year, Length, Budget by Rating



- The middle 50% of ratings fall between 1953 and 1999, with ratings greater than 7 showing **more variability in length** and **higher budgets** compared to those under 7.

Proportions of rating greater than 7 by genre



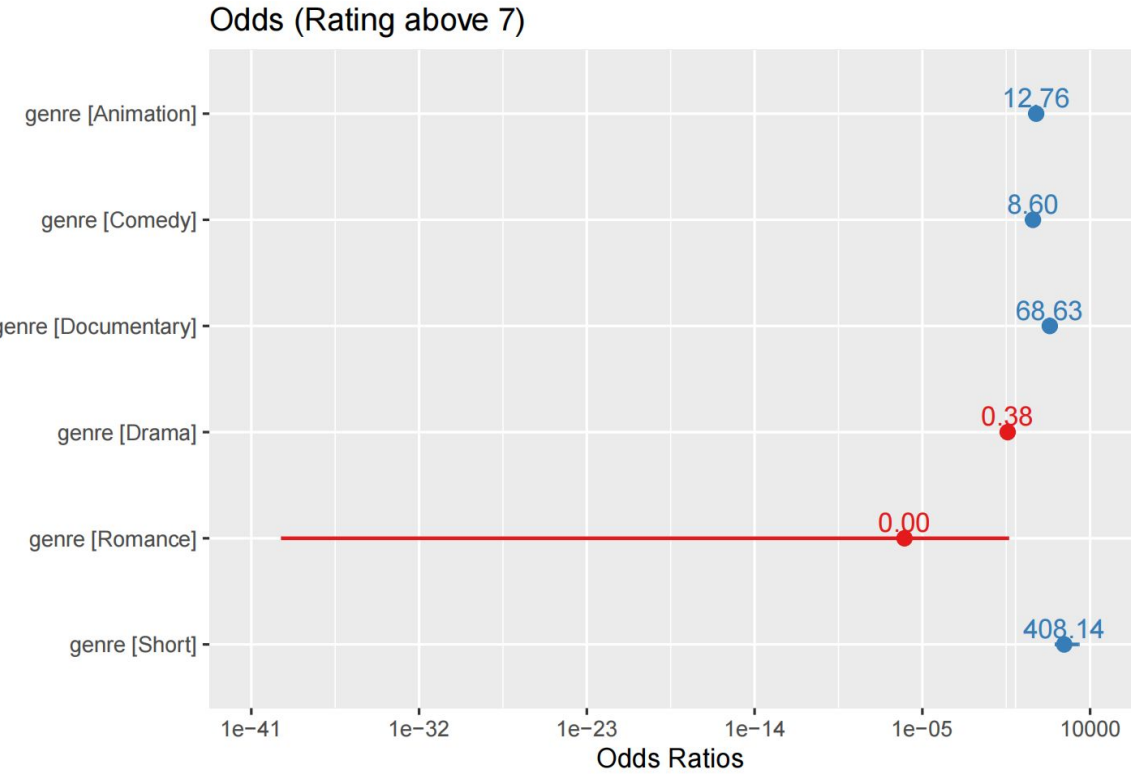
Model fitting in R

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.84494	0.09865	-18.702	< 2e-16	***
genreAnimation	2.54653	0.18731	13.595	< 2e-16	***
genreComedy	2.15141	0.12537	17.160	< 2e-16	***
genreDocumentary	4.22875	0.30618	13.811	< 2e-16	***
genreDrama	-0.97113	0.18244	-5.323	1.02e-07	***
genreRomance	-13.72113	261.39713	0.052	0.958	
genreShort	6.01161	0.71936	8.357	< 2e-16	***

$$g(p_i) = log(\frac{p_i}{1-p_i}) = \alpha + \beta \cdot genre_i$$

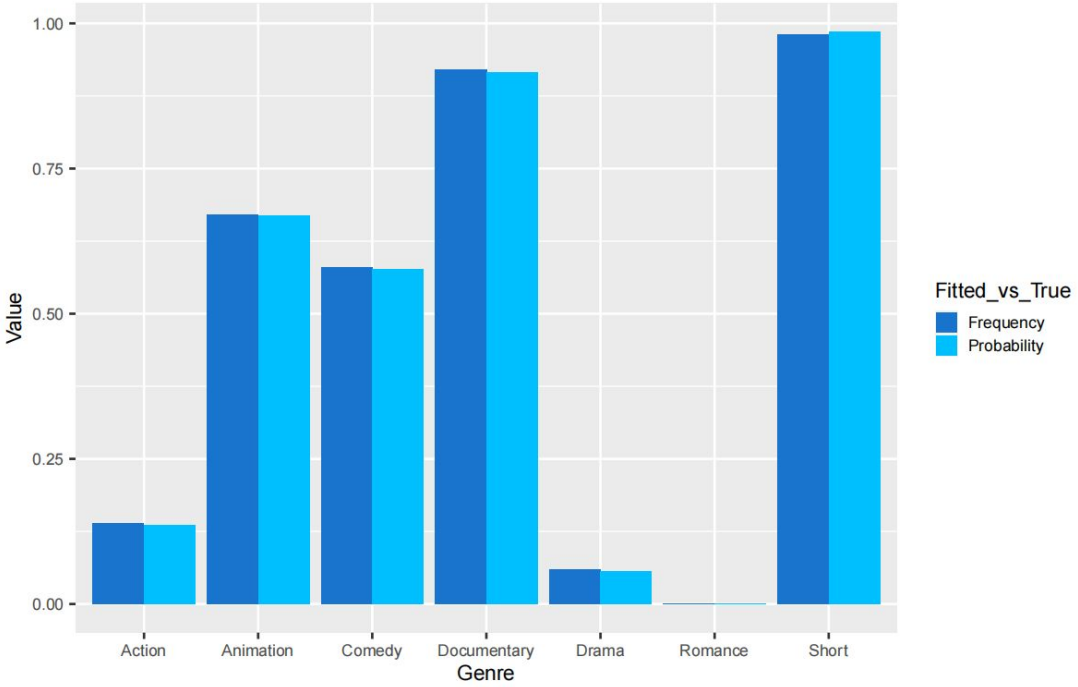
- where ***p_i*** denotes the probability of receiving a high rating, and ***β*** represents the coefficient for genres such as Action, Animation, Comedy, etc.

Odds Ratios



- For every unit increase in **Short**, the probability of rating above 7 becomes 408.14 times higher than that of **Action**.
- For every unit increase in **Drama**, the probability of rating above 7 becomes 0.38 times higher than that of **Action**.

Probability of films with rating above 7



$$\hat{p} = \frac{\exp(\hat{\alpha} + \hat{\beta} \cdot \mathrm{II}_{genre}(\cdot))}{1 + \exp(\hat{\alpha} + \hat{\beta} \cdot \mathrm{II}_{genre}(\cdot))}$$

- The **estimated probability** of all kinds of films is denoted by the dark blue column, while the light blue denotes the **actual frequency** in the dataset.

Model fitting in R

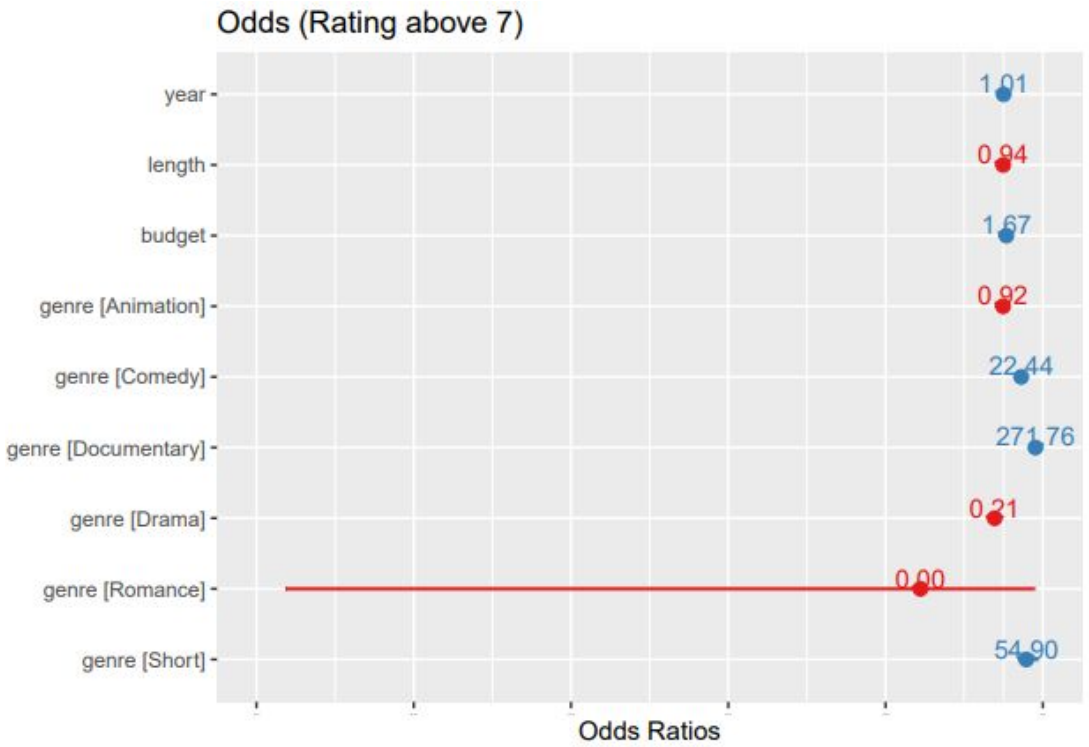
- We use a **stepwise approach** and select the best model based on **AIC** as the selection criterion.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-23.864236	5.767450	-4.138	3.51e-05
year	0.010238	0.002940	3.483	0.000497
length	-0.056869	0.003537	-16.077	< 2e-16
budget	0.509979	0.030117	16.933	< 2e-16
genreAnimation	-0.078708	0.320139	-0.246	0.805793
genreComedy	3.110781	0.179174	17.362	< 2e-16
genreDocumentary	5.604906	0.442282	12.673	< 2e-16
genreDrama	-1.556649	0.239136	-6.509	7.54e-11
genreRomance	-14.607631	391.828859	-0.037	0.970261
genreShort	4.005562	0.796669	5.028	4.96e-07

- The best model removed the **votes** variable from the full model.

$$g(p_i) = log(\frac{p_i}{1-p_i}) = \alpha + \beta_1 \cdot year_i + \beta_2 \cdot length_i + \beta_3 \cdot budget_i + \beta_4 \cdot genre_i$$

Odds Ratios



- Statistically, a film with a rating above 7 is 1.01 times **more likely than** an **older** film.
- A minute **longer** movie is 0.94 times **less** likely to be rated above 7.
- For every million dollar **increase** in **budget**, a film is 1.67 times **more likely** to receive a 7+ rating.
- The influences of the genre are similar to those discussed previously.

Conclusion

Genre preferences of the audience

- The audiences prefer movies with **logic** and **depth** (Documentary and Short) over **formulaic commercial** films (Action, Drama, and Romance).
- Animation and Comedy genres have mixed ratings with more high-rated movies, aligning with **popular audience trends**.

Impact of numerical factors on movie ratings

- **Release year** and **production budget** positively impact a movie's rating, while **movie length** negatively impacts it due to people's preference for faster-paced content.
- **Votes** do not significantly impact a movie's rating, possibly due to their subjectivity.

Future Work

Additional Movie Attributes

- Explore the impact of other movie attributes such as **box office** and **investment in art and music** on movie ratings, to further enhance the accuracy of the classifier's functionality.

More Response Variable Categories

- Consider dividing the response variable into more categories to improve **discriminability**.

Improved Sample Collection

- Obtain a more comprehensive sample to improve the model's performance, as the uneven distribution of genres in the current sample has led to errors.