An Analysis of Factors Influencing IMDB Movie Ratings

A Statistical Study Based on IMDB Dataset

Group 10

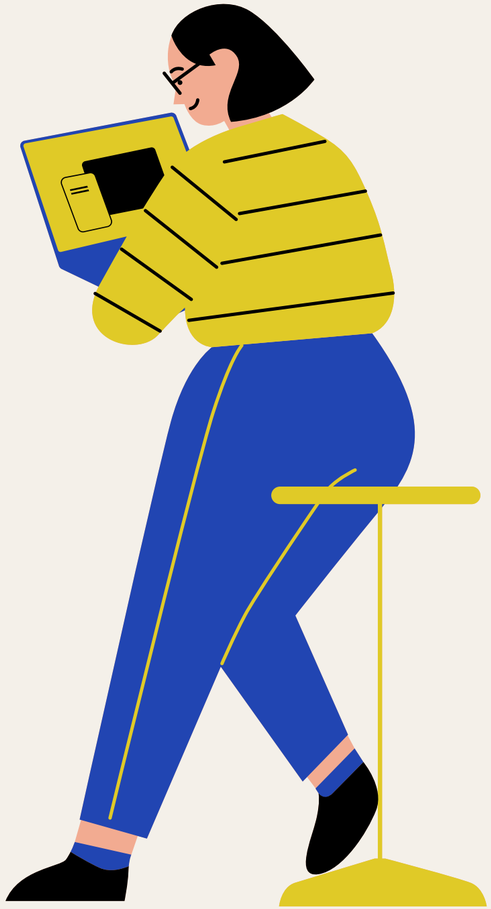
ZHUOYU ZANG、XINYI SHENG、HAOCHENG YU、YANZHEN JIA、Azel Zhuman

01 - Introduction

02 - Exploratory Data Analysis

03 - Statistical Modeling

04 - Conclusions



Data

Visualization

01 - Introduction

Data Source

- IMDB movie dataset with 1,495 films.
- Variables: film_id, year, length, budget, votes, genre, rating.

Research Question

- What factors influence whether a movie gets an IMDB rating above 7?

Research Methodology

- Exploratory Data Analysis (EDA) to understand data patterns.
- Logistic Regression (GLM) to analyze the impact of movie attributes on high ratings.
- Model selection using AIC and BIC for best fit.

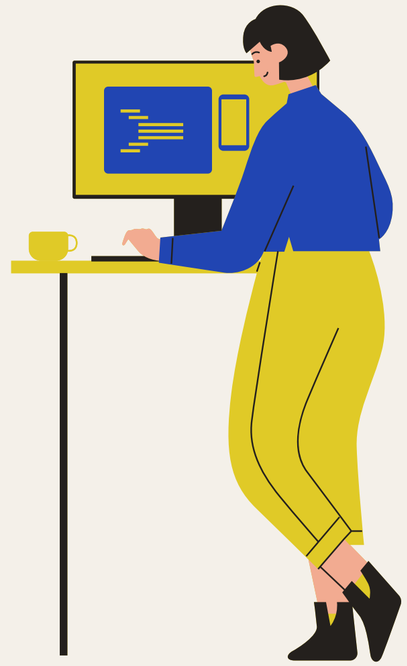
Analyzing data enables informed decision-making

Data

Visualization

02 - Exploratory Data Analysis

Overview of Dataset



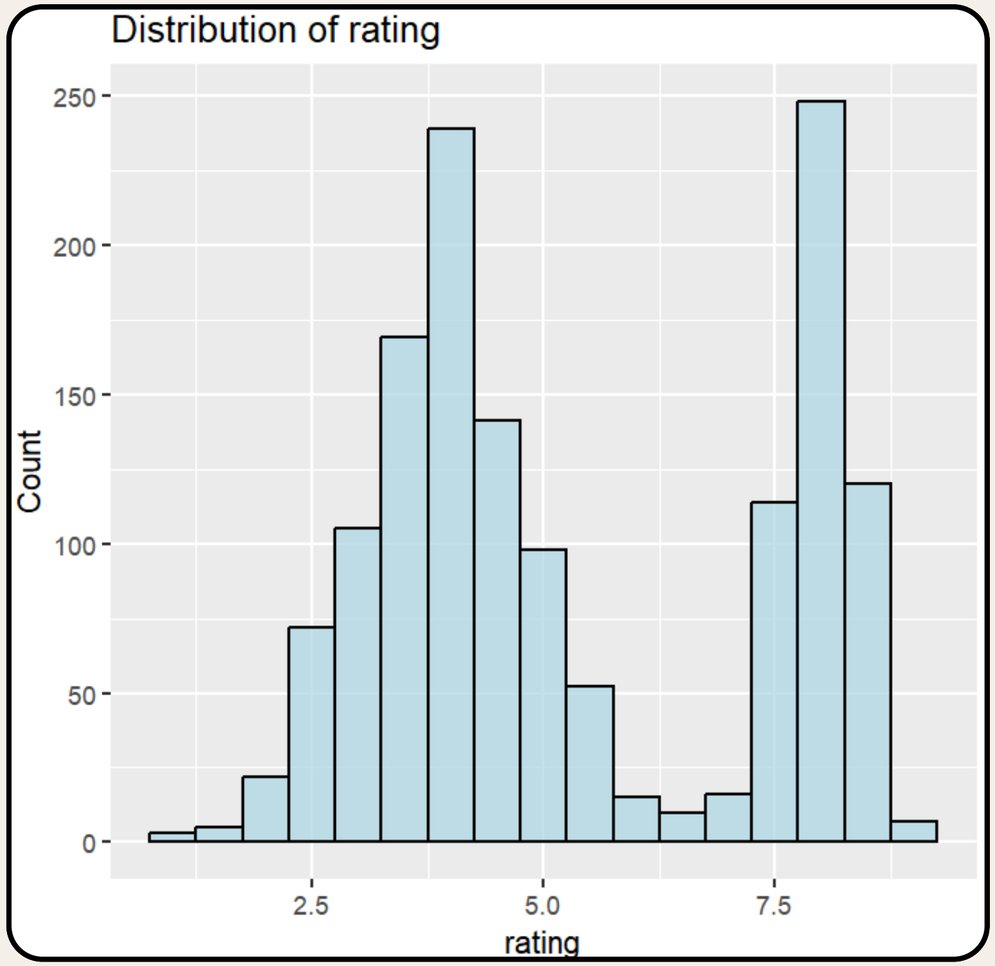
Dataset Description:

- The dataset contains information on 1,495 movies, with the following variables:
  a. film_id – Unique identifier for each film
  b. year – Year of movie release
  c. length – Duration of the film (in minutes)
  d. budget – Production budget (in millions of USD)
  e. votes – Number of positive audience votes
  f. genre – Film genre (e.g., Drama, Action, Comedy)
  g. rating – IMDB rating (from 0 to 10)

Handling of Missing Data:

- There are 59 missing values in the dataset.
- All missing data were removed using the na.omit() function to ensure clean and reliable analysis.

02 - Exploratory Data Analysis

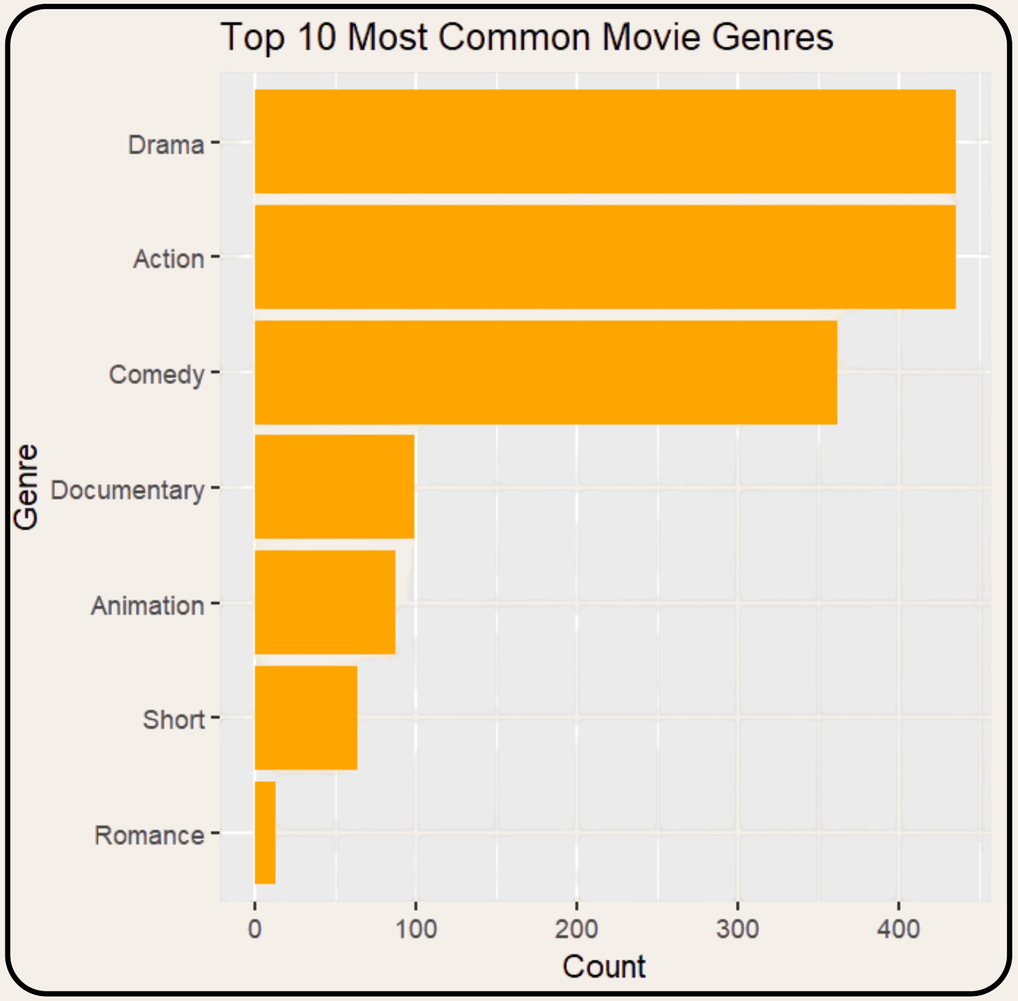


Distribution of Ratings

Key Observations:

- The distribution of ratings is bimodal, meaning there are two prominent peaks.
- Most ratings are concentrated in the ranges of 3–5 and 7–8.
- This indicates a clear polarization in movie ratings, with films tending to receive either low or high scores rather than middle-range scores.

02 - Exploratory Data Analysis



Distribution of Genres

Key Observations:

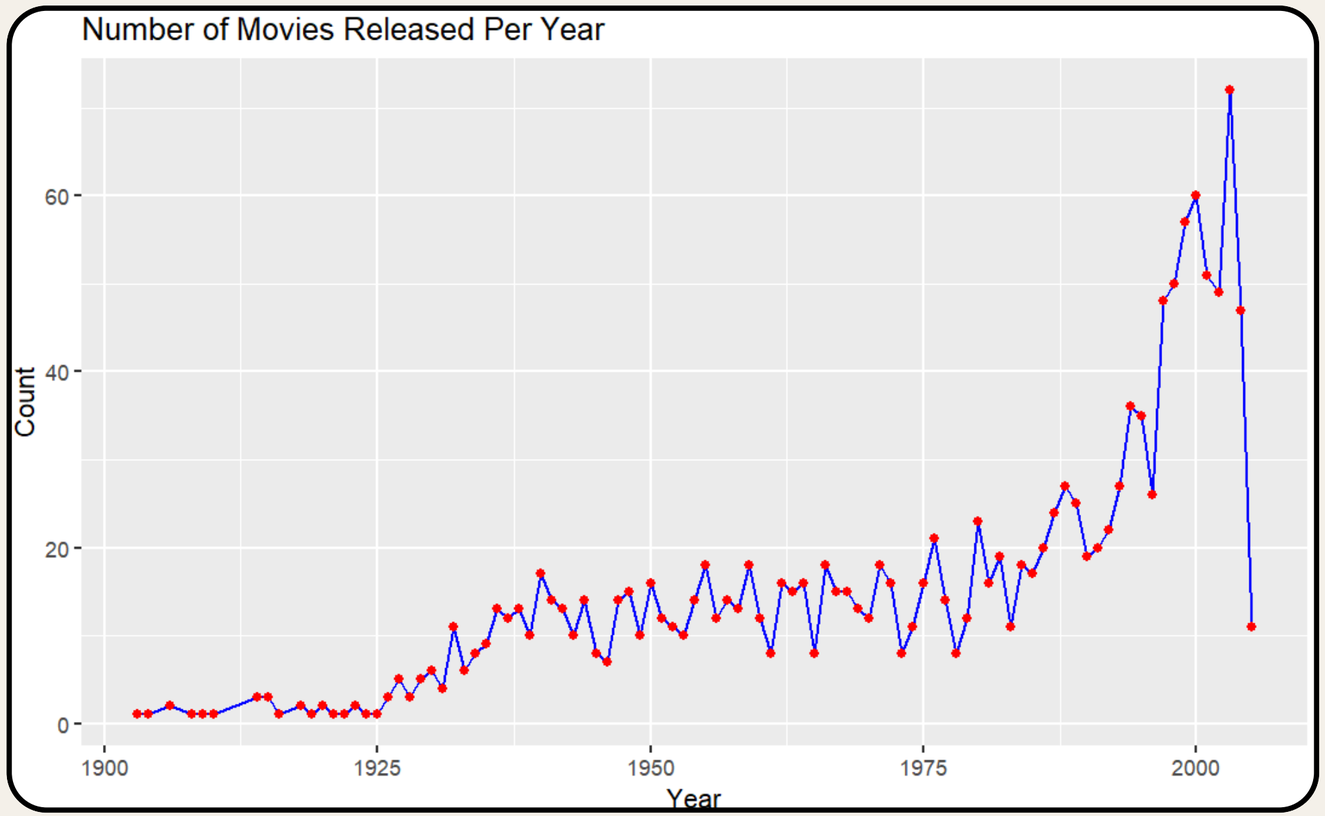
The three most frequent genres are:

- Drama
- Action
- Comedy

Drama and Action genres dominate the dataset, making them the primary focus for analyzing trends in high ratings.

02 - Exploratory Data Analysis

Distribution of Genres



Key Observations:

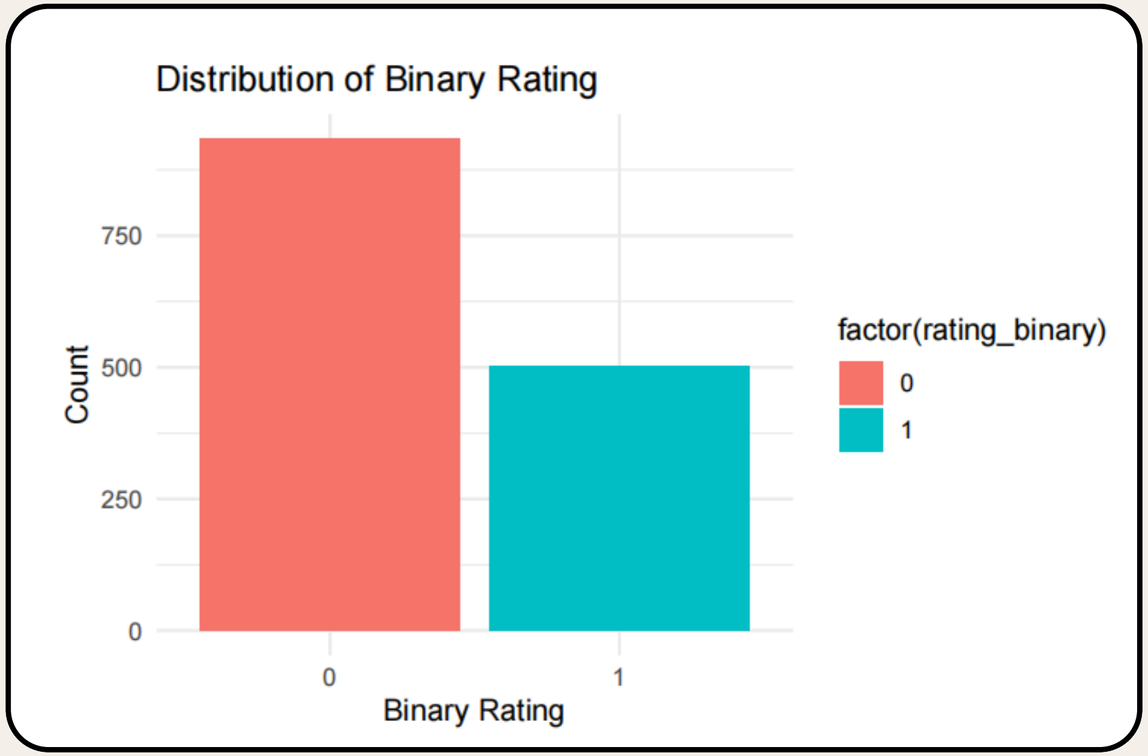
- The number of films released per year has increased significantly since the 1980s.
- The trend shows a sharp rise between 1990 and 2000, reaching a peak around the early 2000s.
- This reflects the growing global film industry and possibly the rising popularity of IMDB as a rating platform.

02 - Exploratory Data Analysis

Distribution of Binary Ratings (High vs. Low)

Key Observations:

- The majority of films in the dataset received ratings below 7, indicating a skew towards lower ratings.
- The approximate ratio of low-rated (≤7) to high-rated (>7) films is about 2:1.
- This imbalance highlights the need to understand what differentiates the smaller group of highly-rated films from the rest.



03 - Statistical Modeling

- ***Model 1 (Linear Regression Model)***

$$rating = 4.523 - 0.1382 \times \text{year} - 4.6849 \times \text{length} + 3.5523 \times \text{budget} + 1.6732 \times \text{votes} + \text{genre\_effects}$$

Regression Model Summary

```
Call:
lm(formula = rating ~ year + length + budget + votes + genre_num,
    data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7995 -0.9983  0.0002  1.0404  4.4763

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.5230     0.2224  20.336  < 2e-16 ***
year                 -0.1382     0.1780  -0.777  0.43751
length               -4.6849     0.4341 -10.793  < 2e-16 ***
budget                3.5523     0.2597  13.677  < 2e-16 ***
votes                 1.6732     0.7613   2.198  0.02814 *
genre_numAnimation    1.0747     0.2078   5.171 2.66e-07 ***
genre_numComedy       1.6255     0.1104  14.730  < 2e-16 ***
genre_numDocumentary  2.6033     0.1767  14.731  < 2e-16 ***
genre_numDrama       -0.4226     0.1042  -4.054 5.30e-05 ***
genre_numRomance     -1.2728     0.4231  -3.008  0.00267 **
genre_numShort        2.1542     0.2342   9.198  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.499 on 1425 degrees of freedom
Multiple R-squared:  0.4887,    Adjusted R-squared:  0.4851
F-statistic: 136.2 on 10 and 1425 DF,  p-value: < 2.2e-16
```

Key Observations:

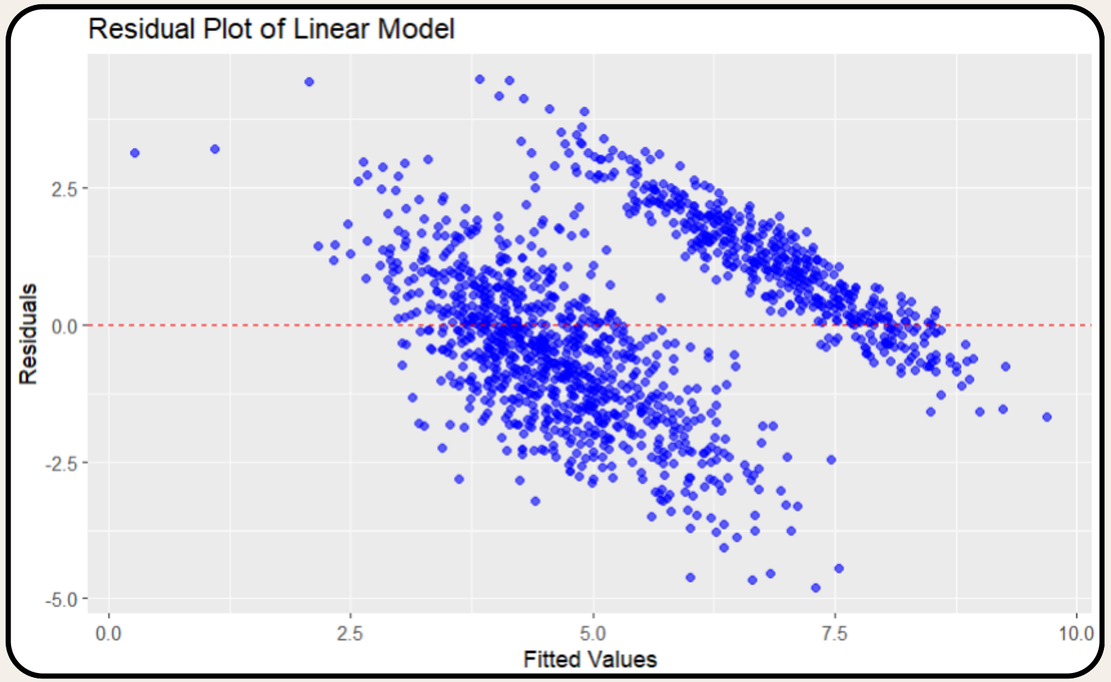
- Model 1 explains 48.87% of rating variability, but a significant portion remains unexplained.
- The F-statistic (136.2, p < 2.2e-16) indicates that the overall model is highly significant.

03 - Statistical Modeling

- ***Model 1 (Linear Regression Model)***

$$rating = 4.523 - 0.1382 \times \text{year} - 4.6849 \times \text{length} + 3.5523 \times \text{budget} + 1.6732 \times \text{votes} + \text{genre\_effects}$$

Pearson Residuals Plot



Key Observations:

- The residual plot shows non-random patterns, indicating that the model might not fully capture all relationships.
- A GLM approach could improve model fit and address these issues.

03 - Statistical Modeling

- ***Model 2 (Logistic Regression Model)***

$$\log\left(\frac{P(\text{rating\_binary}=1)}{1-P(\text{rating\_binary}=1)}\right) = -9.4789 + 0.0038 \times \text{year} - 19.5733 \times \text{length} + 11.2481 \times \text{budget} + 4.8041 \times \text{votes} + \text{genre\_effects}$$

Regression Model Summary

```
Call:
glm(formula = rating_binary ~ year + length + budget + votes +
    genre_num, family = binomial(link = "logit"), data = data1)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.478856   8.486884  -1.117  0.26404
year                0.003814   0.004335   0.880  0.37895
length            -19.573258   1.725614 -11.343  < 2e-16 ***
budget             11.248114   0.861953  13.050  < 2e-16 ***
votes               4.804115   1.740554   2.760  0.00578 **
genre_numAnimation -0.866847   0.539843  -1.606  0.10833
genre_numComedy     3.274858   0.263524  12.427  < 2e-16 ***
genre_numDocumentary 5.561898  0.552412  10.068  < 2e-16 ***
genre_numDrama     -2.077900   0.345890  -6.007 1.89e-09 ***
genre_numRomance  -15.446373 562.295627  -0.027  0.97808
genre_numShort      3.535718   1.143062   3.093  0.00198 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1858.7  on 1435  degrees of freedom
Residual deviance:  711.5  on 1425  degrees of freedom
AIC: 733.5
```

Key Observations:

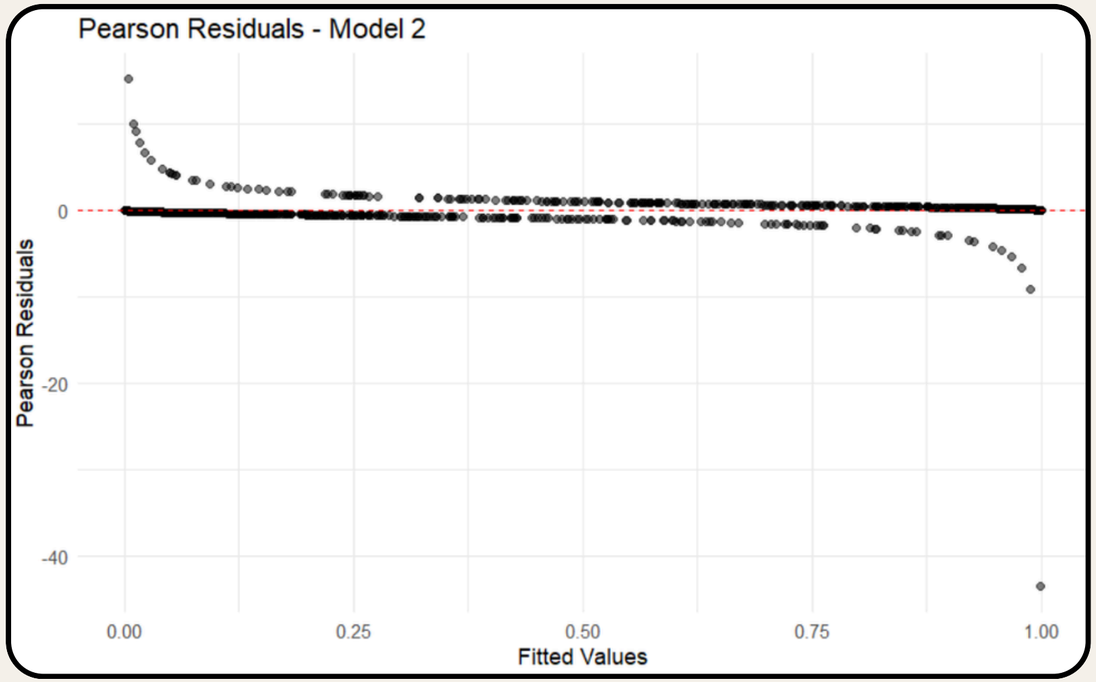
- The null deviance (1858.7) vs. residual deviance (711.5) shows a substantial reduction, indicating that the model explains a significant portion of variability.
- The AIC (733.5) suggests the model's overall quality, with lower values indicating better fit.
- Year is not statistically significant (p = 0.379) and could be removed to improve model performance.

03 - Statistical Modeling

- ***Model 2 (Logistic Regression Model)***

$$\log\left(\frac{P(\text{rating\_binary}=1)}{1-P(\text{rating\_binary}=1)}\right) = -9.4789 + 0.0038 \times \text{year} - 19.5733 \times \text{length} + 11.2481 \times \text{budget} + 4.8041 \times \text{votes} + \text{genre\_effects}$$

Pearson Residuals Plot



Key Observations:

- Residuals are mostly around zero, but large errors at extreme fitted values suggest non-constant variance.

03 - Statistical Modeling

- ***Model 3 (Logistic Regression Model (Without Year Variable))***

$$\log\left(\frac{P(\text{rating\_binary}=1)}{1-P(\text{rating\_binary}=1)}\right) = -2.0297 - 19.2431 \times \text{length} + 11.2226 \times \text{budget} + 4.8916 \times \text{votes} + \text{genre\_effects}$$

Regression Model Summary

```
Call:
glm(formula = rating_binary ~ length + budget + votes + genre_num,
    family = binomial, data = data1)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.0297     0.5379  -3.773 0.000161 ***
length            -19.2431     1.6783 -11.466  < 2e-16 ***
budget             11.2226     0.8603  13.046  < 2e-16 ***
votes               4.8916     1.7233   2.838 0.004534 **
genre_numAnimation -0.8426     0.5387  -1.564 0.117774
genre_numComedy     3.2833     0.2638  12.446  < 2e-16 ***
genre_numDocumentary 5.6177    0.5474  10.262  < 2e-16 ***
genre_numDrama     -2.0757     0.3442  -6.031 1.63e-09 ***
genre_numRomance  -15.5254   556.2689  -0.028 0.977734
genre_numShort      3.6338     1.1388   3.191 0.001418 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1858.72  on 1435  degrees of freedom
Residual deviance:  712.27  on 1426  degrees of freedom
AIC: 732.27

Number of Fisher Scoring iterations: 15
```

Key Observations:

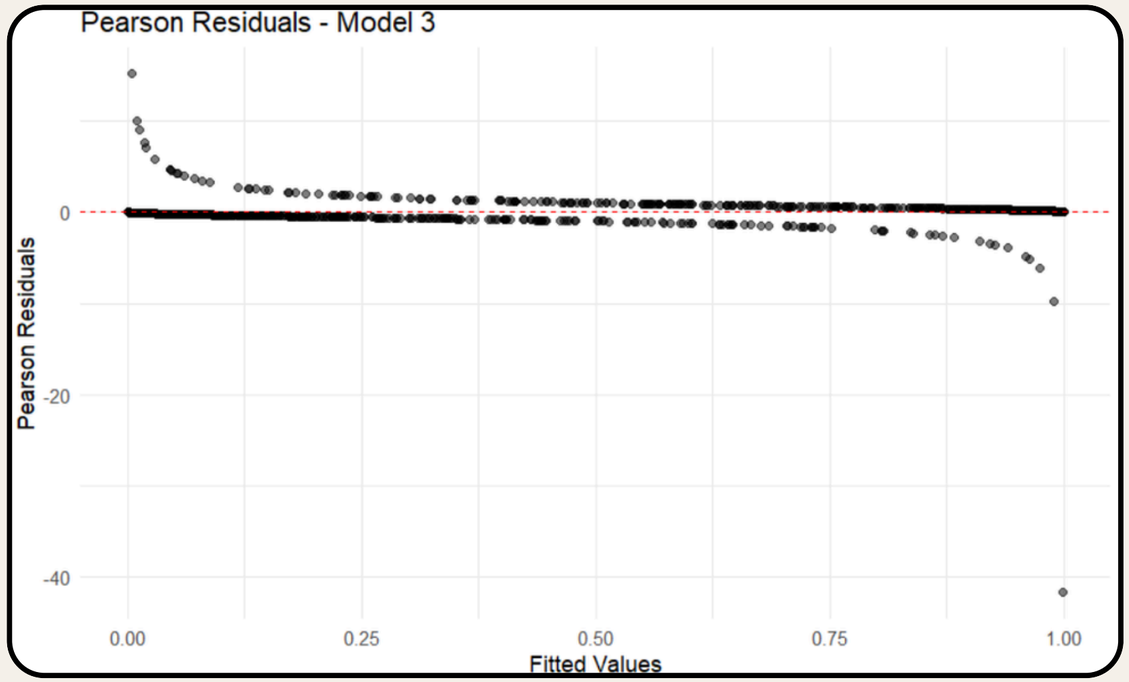
- Most variables have p-values < 0.05, indicating a significant impact on ratings.
- Compared to Model2, the lower AIC (732.27) suggests that Model3 has a better fit and is more efficient in balancing goodness-of-fit and complexity.

03 - Statistical Modeling

- ***Model 3 (Logistic Regression Model (Without Year Variable))***

$$\log\left(\frac{P(\text{rating\_binary}=1)}{1-P(\text{rating\_binary}=1)}\right) = -2.0297 - 19.2431 \times \text{length} + 11.2226 \times \text{budget} + 4.8916 \times \text{votes} + \text{genre\_effects}$$

Pearson Residuals Plot



Key Observations:

- Model 3 shows a more concentrated residual distribution, reduced non-constant variance, and fewer extreme residuals, indicating better fit and stability.

03 - Statistical Modeling

Final Model Selection

- Model 3 has mostly significant regression coefficients ($p < 0.05$), a more stable Pearson residual plot, and lower AIC value, achieving a better balance between goodness-of-fit and model complexity. Therefore, Model 3 is selected as the final model.

- ***Model 3 (Logistic Regression Model (Without Year Variable))***

$$\log\left(\frac{P(\text{rating\_binary}=1)}{1-P(\text{rating\_binary}=1)}\right) = -2.0297 - 19.2431 \times \text{length} + 11.2226 \times \text{budget}$$
$$+4.8916 \times \text{votes} - 0.8426 \times \text{genre\_Animation}$$
$$+3.2833 \times \text{genre\_Comedy} + 5.6177 \times \text{genre\_Documentary}$$
$$-2.0757 \times \text{genre\_Drama} - 15.5254 \times \text{genre\_Romance}$$
$$+3.6338 \times \text{genre\_Short}$$

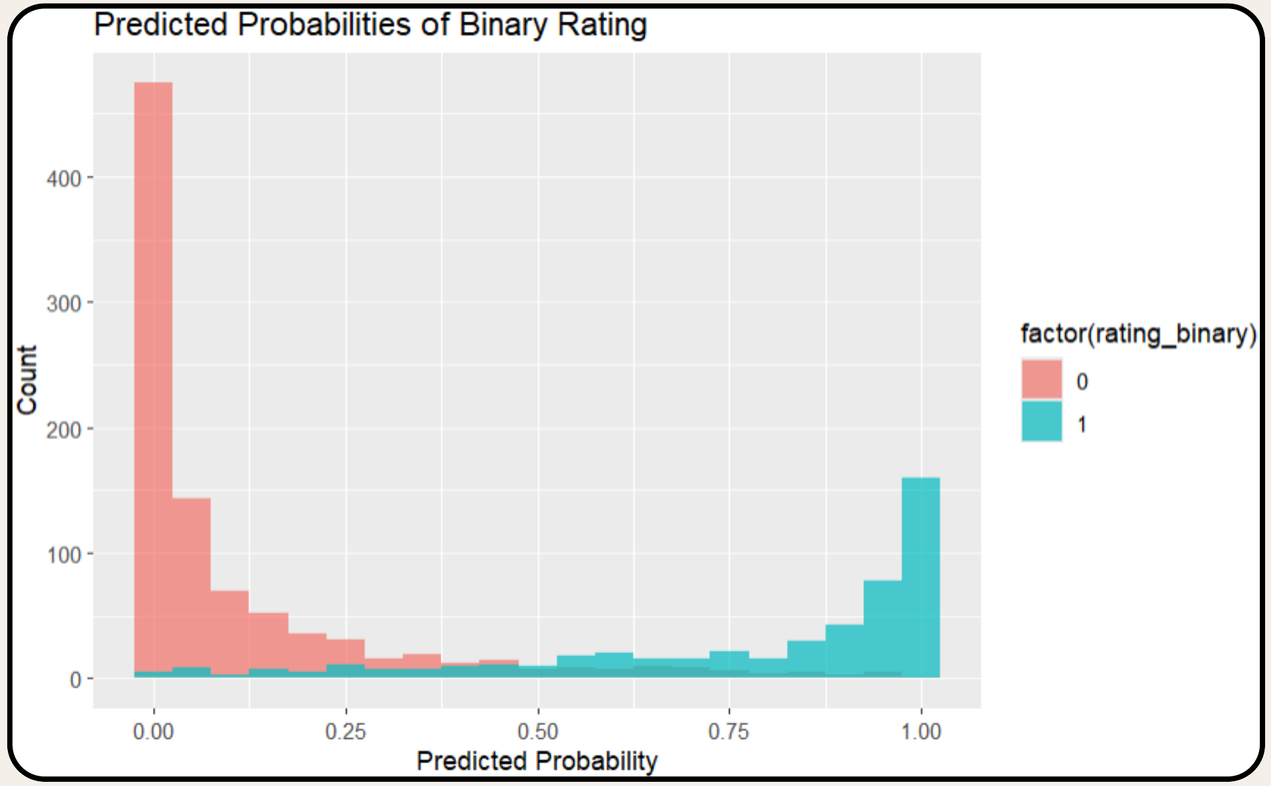
03 - Statistical Modeling

Model Interpretation

- The model indicates that length, budget, and votes significantly affect movie ratings. Longer movies are less likely to receive high ratings, while higher budgets and more votes increase the likelihood of high ratings.
- Regarding film genres, documentaries, short films, and comedies are more likely to receive high ratings, with documentaries having the strongest positive effect. In contrast, dramas tend to receive lower ratings. Animation and romance films show no significant effect, possibly due to sample distribution or other factors.

03 - Statistical Modeling

Assessing Model Performance



Key Observations:

- The histogram shows the predicted probabilities of binary movie ratings (0 or 1) from Model 3. Low-rated movies (red) are mostly near 0, while high-rated movies (blue) cluster around 1, indicating strong classification.
- However, some overlap suggests misclassification, which could be improved with interaction terms or data balancing.

04 - Conclusions

Practical Insights for the Film Industry

- The findings suggest that movie length, budget, and audience engagement (votes) significantly impact ratings, which has practical implications for film production and marketing. Filmmakers should avoid excessively long runtimes to maintain audience engagement, invest in higher budgets to improve production quality, and focus on effective promotion strategies to increase audience participation (votes).

04 - Conclusions

Practical Insights for the Film Industry

- In terms of genre selection, documentaries, short films, and comedies tend to receive higher ratings, indicating strong audience preference for these genres. Dramas are less likely to achieve high ratings, possibly due to higher viewer expectations or heavier themes. Animation and romance films show no clear rating trend, which may be influenced by market segmentation or audience preferences. Therefore, filmmakers should carefully consider target audience preferences and market demand when selecting a film genre.

Thanks

Group 10