Behind the Curtain: Statistical Insights into Movie Success

Cheng Tang, Mingcan Wang, Yiang Liang, Yuxuan Zhao, Zilu Wang

Table of contents

1 Introduction **1**

2 Methodology **2**

3 Exploratory Data Anlaysis **3**
3.1 Statistical Summary . . . 3
3.2 Outliers . . . 4
3.3 Visualisation . . . 4
3.4 EDA Findings . . . 10

4 Formal Analysis **11**
4.1 Model Building . . . 11
4.2 Model Selection . . . 11
4.3 Model Tuning . . . 12
4.4 Model Interpretation . . . 13
4.5 Assumption Checking . . . 15

5 Conclusion **16**

6 Discussion **16**
6.1 Implications . . . 16
6.2 Limitation . . . 17
6.3 Further Research . . . 17

1 Introduction

In the evolving landscape of cinematic entertainment, the question of what factors lead a film to be favorably received by audiences has intrigued producers, directors, and marketers alike. This project, titled "Behind the Curtain: Statistical Insights into Movie Success" embarks on a statistical journey to decipher the complex dynamics between various film attributes and their resulting viewer ratings, specifically focusing on the critical threshold of a rating above 7, often considered a benchmark for success in the industry.

The inception of this analysis is rooted in the premise that a film's length, budget, viewer engagement (measured through votes), and genre hold significant sway over its overall reception. Traditionally, the entertainment industry has relied on anecdotal evidence or isolated case studies to gauge the potential success of film projects. However, this project leverages a Generalized Linear Model (GLM) to evaluate these factors, offering a more empirical basis for understanding cinematic success.

The data set comprises diverse films spanning various years, genres, and production scales, enabling a comprehensive analysis that transcends specific market trends or cultural biases. By employing a generalized linear regression framework, we aim to predict the likelihood of a film achieving a rating above 7, transforming subjective notions of quality and appeal into quantifiable probabilities.

Through this project, the goal is to distill actionable insights that can guide filmmakers and studios in crafting content that resonates with viewers. Beyond its immediate application, this study contributes to the broader discourse on the quantification of artistic and entertainment value, marking a confluence of creativity and analytics.

2 Methodology

The methodology of the project involves a systematic approach to understanding the factors contributing to movie success, as measured by audience ratings. Initially, the data is cleansed and preprocessed, which includes handling missing values and transforming skewed distributions through log transformations for variables such as film length and votes to achieve distributions closer to normal. Subsequently, a binary variable is created to distinguish films based on whether they have achieved a rating above 7.

An extensive Exploratory Data Analysis (EDA) is conducted to gain deeper insights into underlying patterns and relationships. This includes examining the distributions of key variables, identifying outliers, and assessing correlations.

The analysis then employs a Generalized Linear Model (GLM), specifically logistic regression, to examine the influence of various film attributes on the likelihood of a film receiving a rating above 7, which is considered indicative of success. The model's predictive power and fit are assessed through accuracy, sensitivity, specificity, and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) metrics.

To fine-tune the model, a series of candidate thresholds for classification are evaluated to identify the optimal balance between sensitivity and specificity. This involves calculating performance metrics across different threshold values and selecting the one that provides the best compromise according to the project's objectives.

The methodology also encompasses evaluation of the model's assumptions and the fit to the data, ensuring the reliability and validity of the findings. Finally, based on the insights gained from the EDA and GLM analysis, strategic recommendations are formulated to guide filmmakers and producers in aligning their projects with the attributes associated with higher-rated films.

3 Exploratory Data Anlaysis

3.1 Statistical Summary

Statistical Summary of Numerical Variables

Variable	Mean	Standard Deviation	Median	Interquartile Range	Minimum	Maximum
year	1976.872225	23.739365	1984.0	40.0	1894.0	2005.0
length	81.745287	36.978082	90.0	26.0	1.0	399.0
budget	11.948136	2.967745	12.0	3.9	2.1	23.7
votes	658.969418	4370.037987	32.0	106.0	5.0	103854.0
rating	5.414328	2.069483	4.7	4.1	0.7	9.2

Frequency and Percentage Summary by Genre

Genre	Frequency	Percentage (%)
Action	698	29.24
Drama	684	28.66
Comedy	582	24.38
Animation	165	6.91
Documentary	136	5.70
Short	106	4.44
Romance	16	0.67

Frequency and Percentage Summary for Ratings Above 7

Above Rating 7	Frequency	Percentage (%)
0	1546	64.77
1	841	35.23

3.2 Outliers

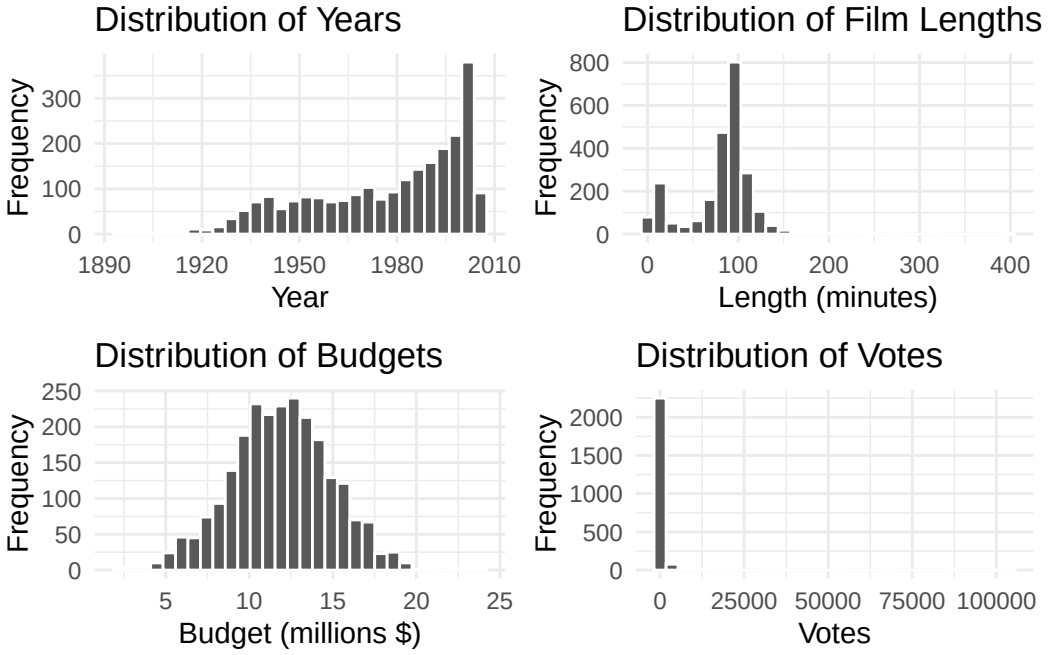
Proportions of outliers for each numeric variable

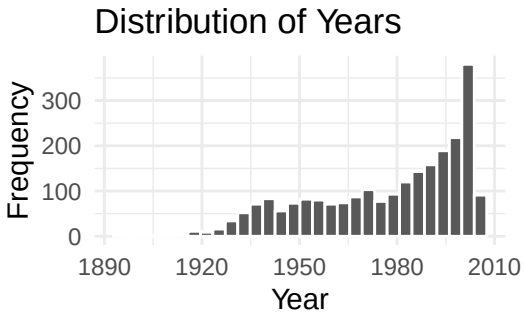
```
                            length      budget     votes
Proportion of Outliers 0.1805614 0.004608295 0.1625471
```

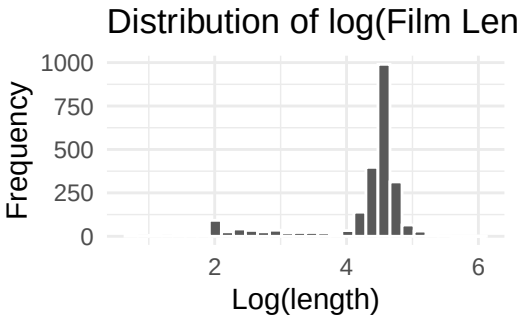
Proportions of outliers for numeric variables after log

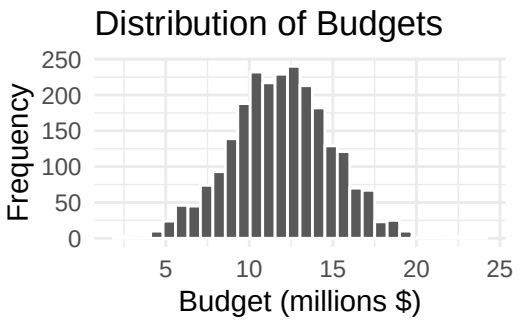
```
                       length_log  votes_log
Proportion of Outliers  0.1818182 0.03812317
```

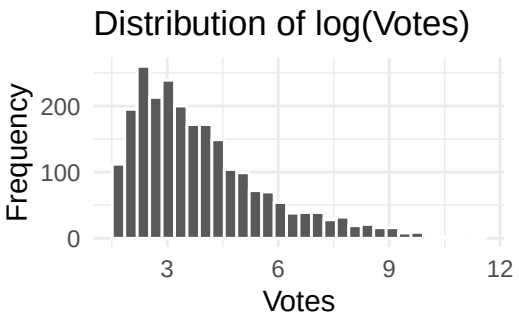
3.3 Visualisation

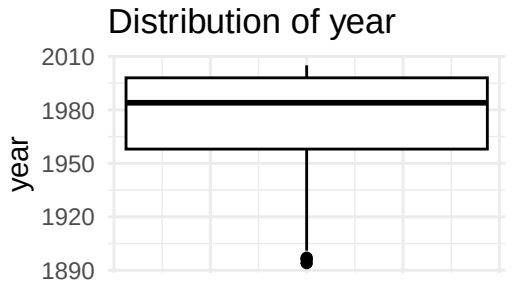


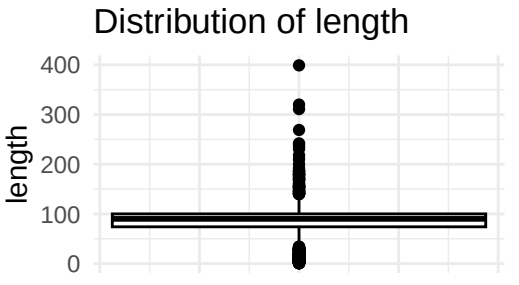

Distribution of Years
Frequency
300
200
100
0
1890
1920
1950
1980
2010
Year

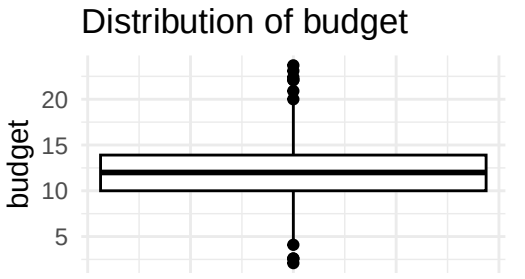

Distribution of log(Film Len
Frequency
1000
750
500
250
0
2
4
6
Log(length)

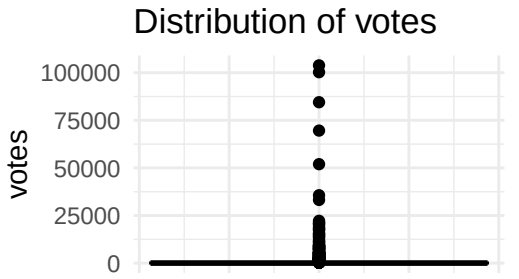

Distribution of Budgets
Frequency
250
200
150
100
50
0
5
10
15
20
25
Budget (millions $)

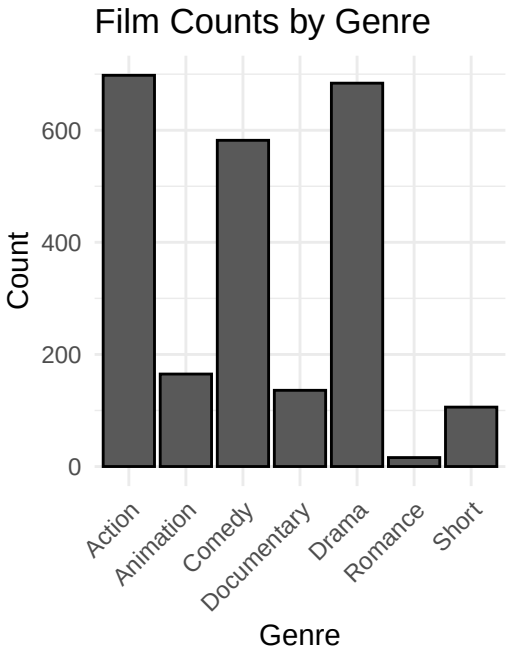

Distribution of log(Votes)
Frequency
200
100
0
3
6
9
12
Votes

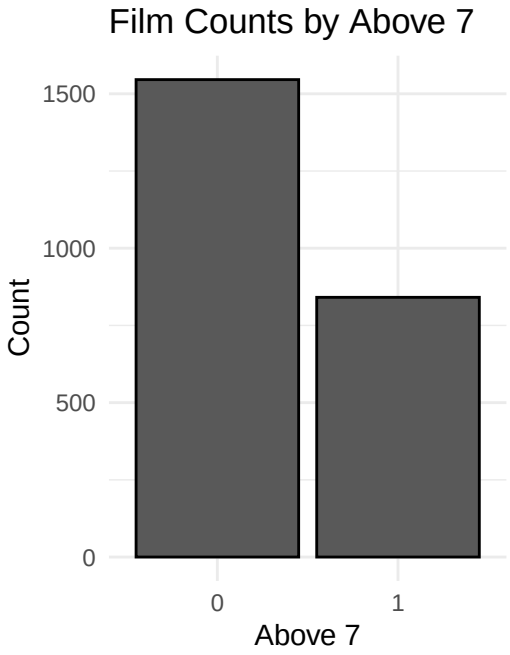

Distribution of year
year
2010
1980
1950
1920
1890

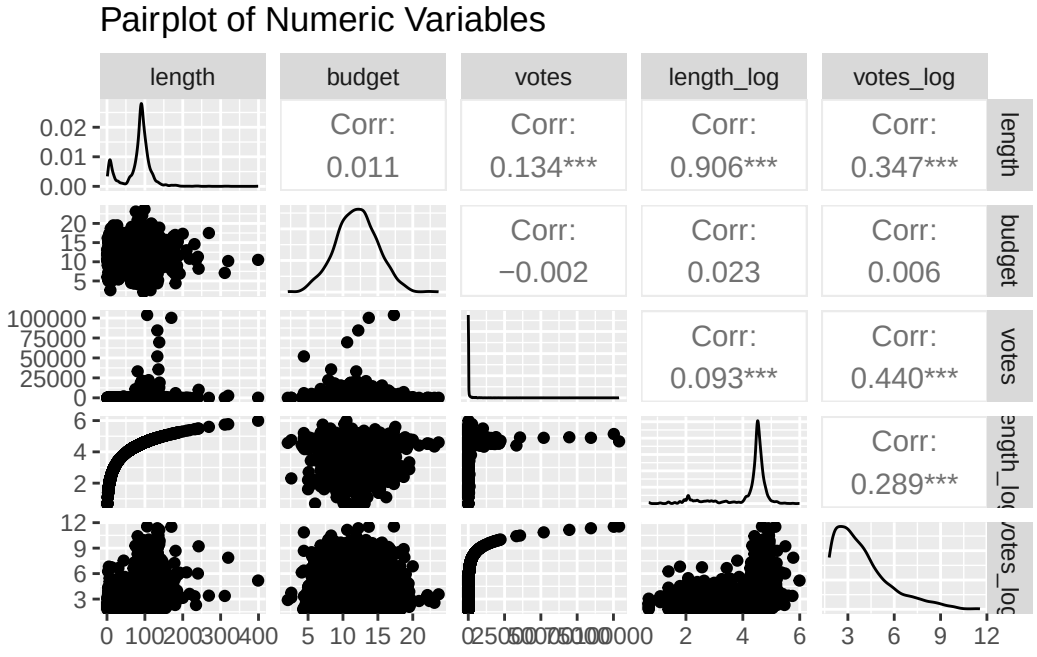

Distribution of length
length
400
300
200
100
0

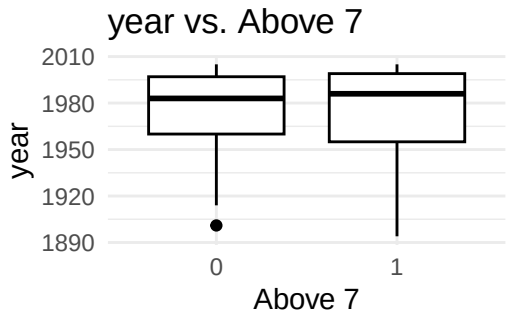

Distribution of budget
budget
20
15
10
5

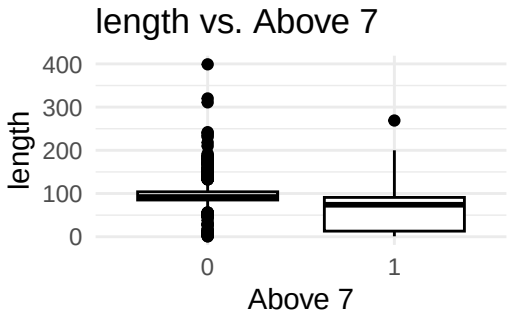

Distribution of votes
votes
100000
75000
50000
25000
0

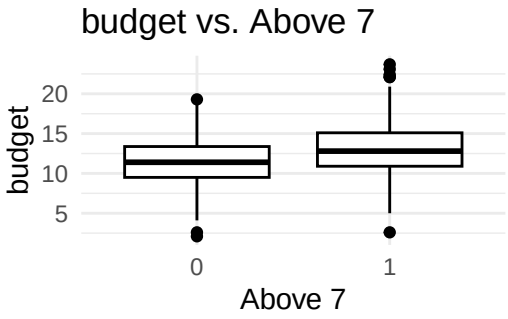

Film Counts by Genre
Count
600
400
200
0
Action
Animation
Comedy
Documentary
Drama
Romance
Short
Genre

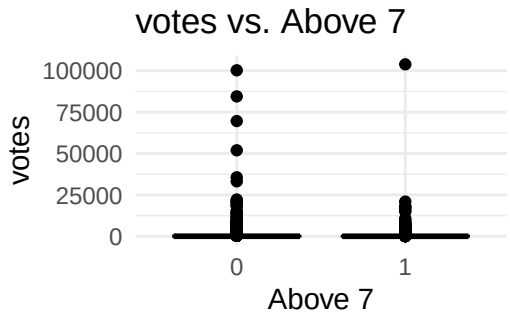

Film Counts by Above 7
Count
1500
1000
500
0
0
1
Above 7

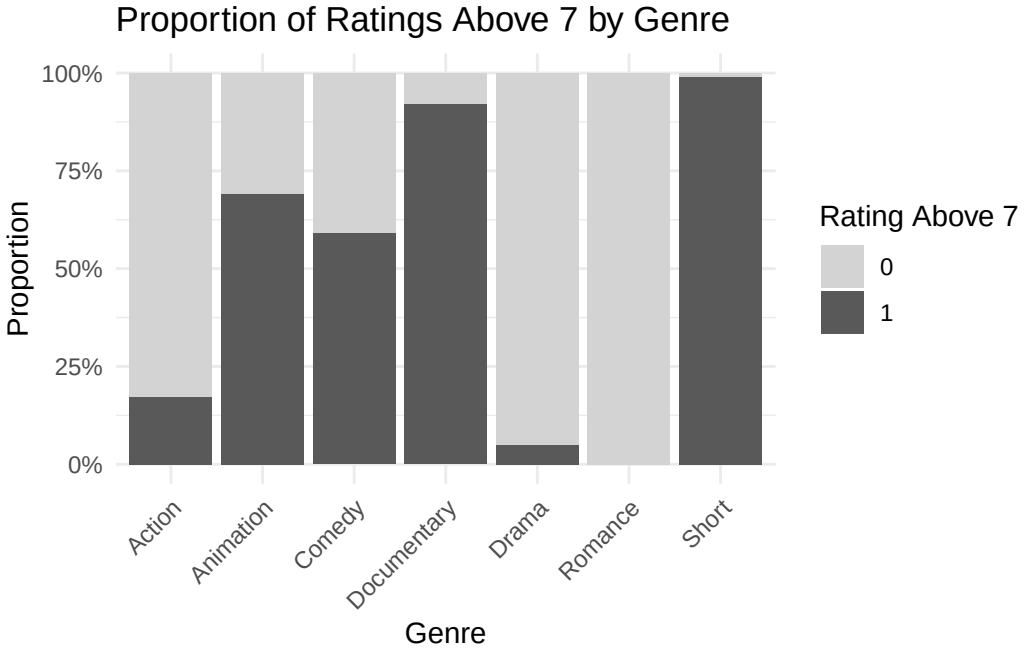

Pairplot of Numeric Variables
length
budget
votes
length_log
votes_log
Corr:
0.011
Corr:
0.134***
Corr:
0.906***
Corr:
0.347***
Corr:
−0.002
Corr:
0.023
Corr:
0.006
Corr:
0.093***
Corr:
0.440***
Corr:
0.289***

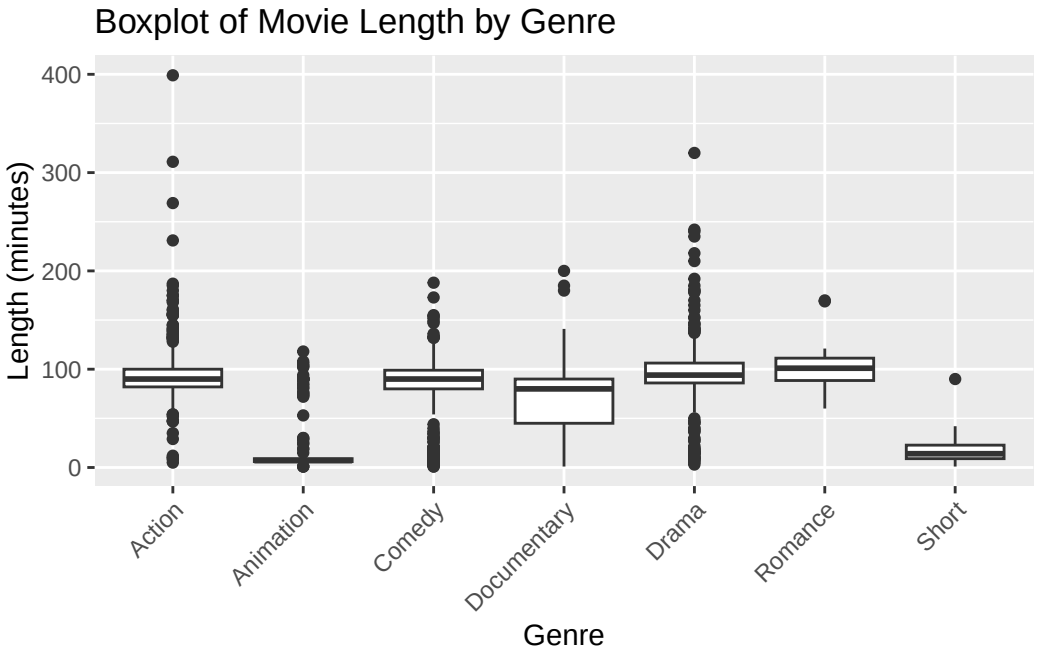

year vs. Above 7
2010
1980
1950
1920
1890
year
0
1
Above 7

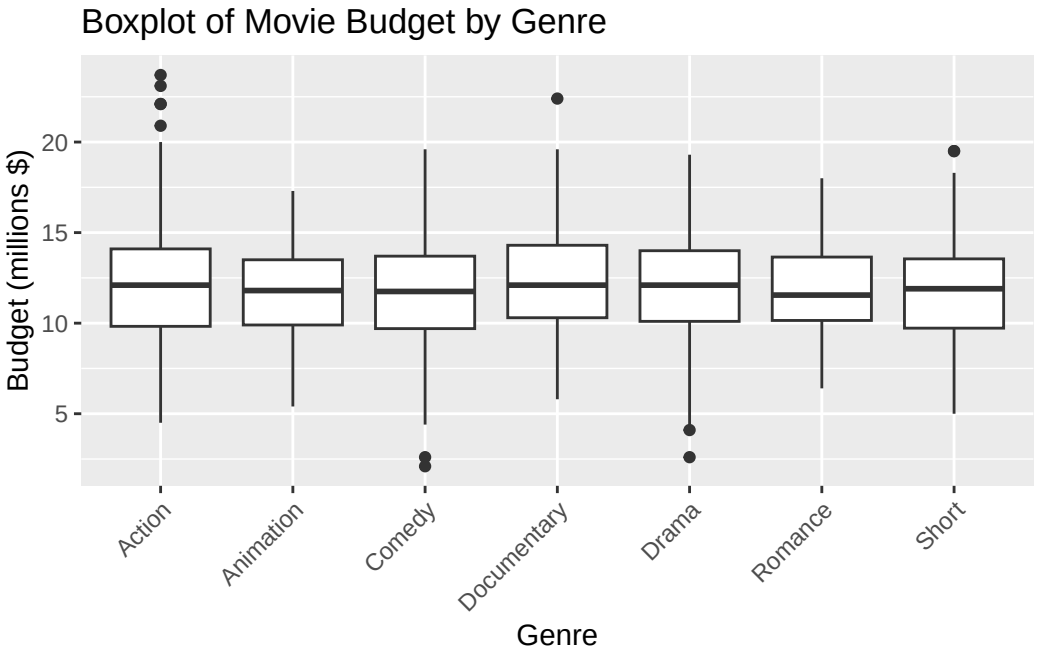

length vs. Above 7
400
300
200
100
0
length
0
1
Above 7

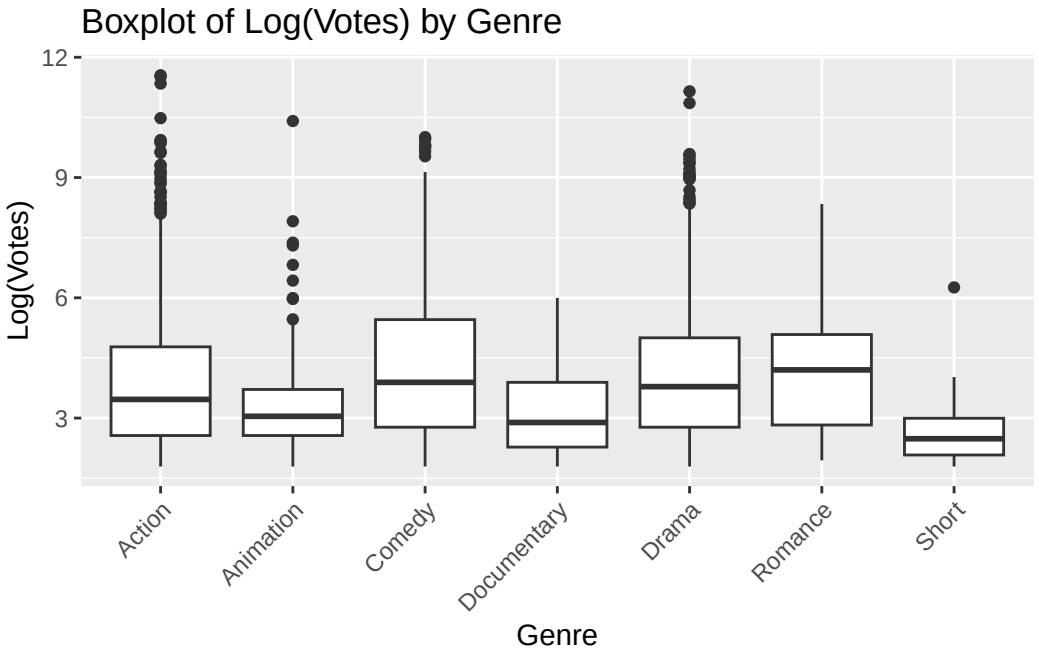

budget vs. Above 7
20
15
10
5
budget
0
1
Above 7

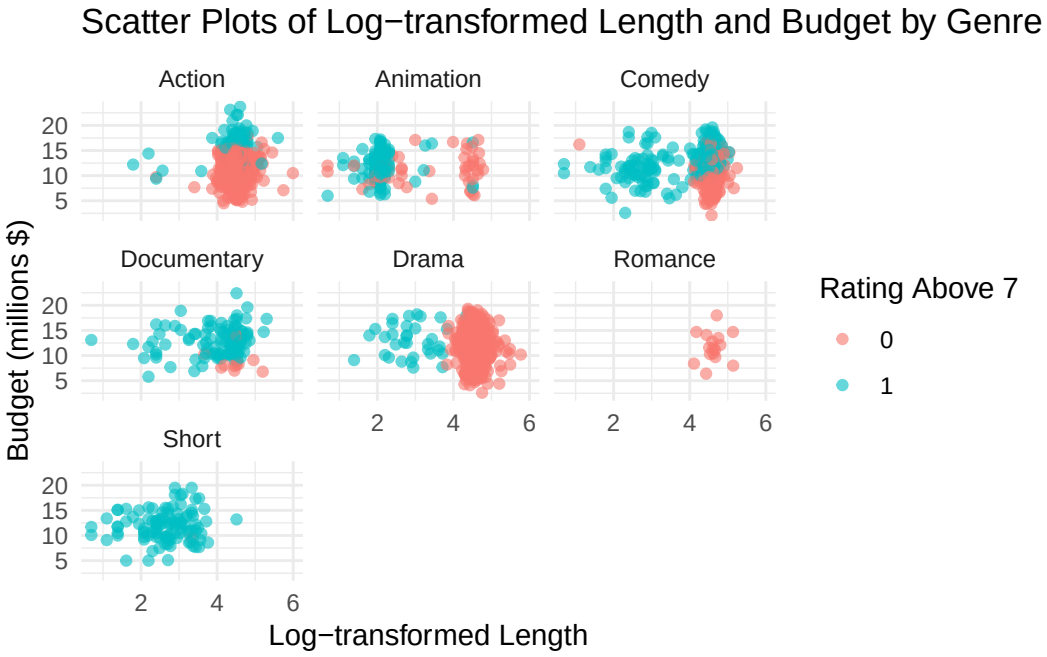

votes vs. Above 7
100000
75000
50000
25000
0
votes
0
1
Above 7

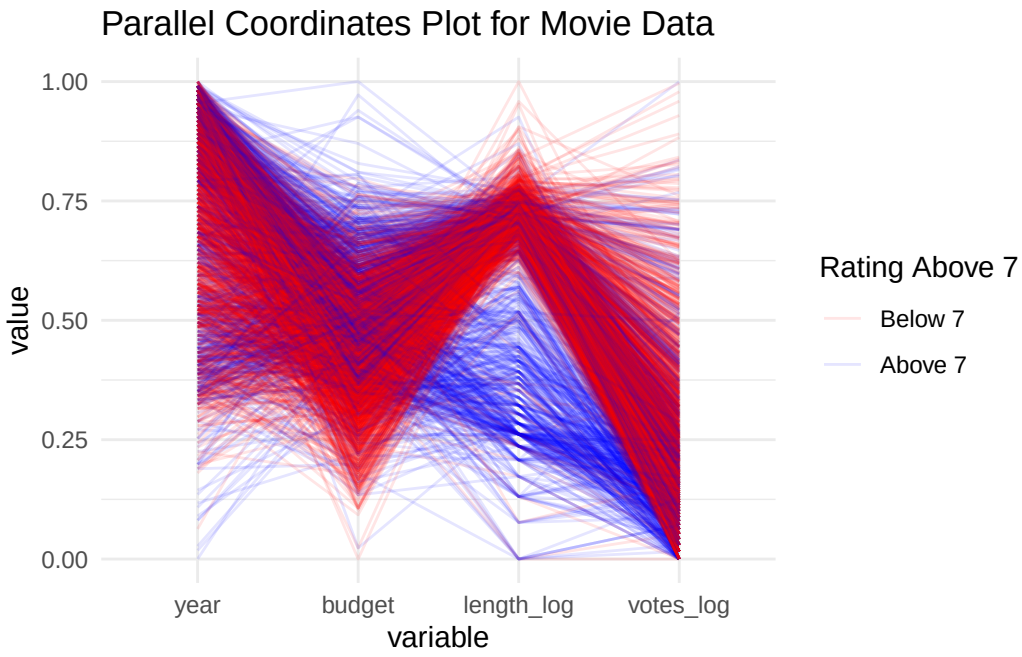

Proportion of Ratings Above 7 by Genre
100%
75%
50%
25%
0%
Proportion
Action
Animation
Comedy
Documentary
Drama
Romance
Short
Genre
Rating Above 7
0
1


Boxplot of Movie Length by Genre
Length (minutes)
400
300
200
100
0
Action
Animation
Comedy
Documentary
Drama
Romance
Short
Genre


Boxplot of Movie Budget by Genre
Budget (millions $)
20
15
10
5
Action
Animation
Comedy
Documentary
Drama
Romance
Short
Genre


Boxplot of Log(Votes) by Genre
Log(Votes)
12
9
6
3
Action
Animation
Comedy
Documentary
Drama
Romance
Short
Genre


Scatter Plots of Log−transformed Length and Budget by Genre
Action
Animation
Comedy
Documentary
Drama
Romance
Short
Budget (millions $)
20
15
10
5
2
4
6
Log−transformed Length
Rating Above 7
0
1



3.4 EDA Findings

In the exploratory data analysis, we observed distinct patterns within the film data set. The length of films is right-skewed, with most under 100 minutes, but exceptions extending up to 399 minutes. Similarly, the 'votes' distribution is significantly right-skewed, highlighting a disparity in viewer engagement. Conversely, budgets appear nearly normally distributed, indicating diverse financial investments across films.

After log transformations, the distributions of 'length' and 'votes' approached closer to normality but still exhibited skewness. The data set predominantly features action, drama, and comedy genres, with fewer romantic and short films. Notably, only 35% of movies are rated above 7.

There is a medium positive correlation between log-transformed votes and length, suggesting films of longer duration may engage viewers more. It is also shown that movies rated above 7 typically have higher budgets. Genre-wise, short and documentaries stand out with highest proportions of high-rated films, whereas romance, drama, and action genres show fewer films surpassing the rating threshold. Short films and animations are generally shorter, whereas romance tends to be longer. The budget across various genres does not vary significantly, with action and documentaries exhibit slightly higher budgets. Lastly, romance genre films receive the most votes, while short films receive the fewest, indicating varying audience engagement levels by genre.

4 Formal Analysis

4.1 Model Building

```
# Full model
glm_model_full <- glm(above_7 ~ year + length + budget + votes + genre,
                      family = binomial, data = train_data)
# Full Model with Log Transformation
glm_model_log <- glm(above_7 ~ year + length_log + budget + votes_log + genre,
                     family = binomial, data = train_data)

# Model without Year
glm_model_no_year <- glm(above_7 ~ length_log + budget + votes_log + genre,
                         family = binomial, data = train_data)

# Model without Year and Votes_log
glm_model_no_year_votes <- glm(above_7 ~ length_log + budget + genre,
                               family = binomial, data = train_data)
# Model without Year and Length_log
glm_model_no_year_length <- glm(above_7 ~ votes_log + budget + genre,
                               family = binomial, data = train_data)
```

In this project, the modeling principle involved constructing and refining a series of logistic regression models to identify key factors influencing a movie's success, defined as achieving a rating above 7. The full model included all variables (except ID), offering a comprehensive baseline for analysis.

Subsequent models were developed by applying log transformation and removing variables based on their statistical significance, assessed through p-values, and their impact on the model's overall performance. This iterative process aimed to streamline the model, removing less impactful variables while observing changes in performance metrics like accuracy, sensitivity, specificity, and AUC.

4.2 Model Selection

```
Variables Accuracy Sensitivity
```

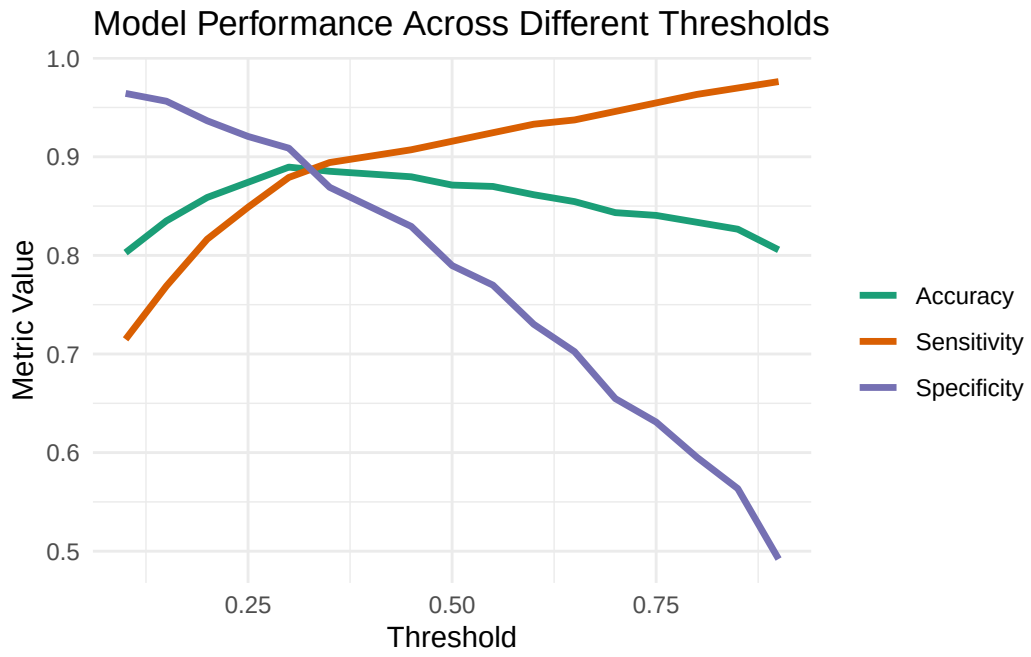
Model 1	year + length + budget + votes + genre	0.8643	0.9179
Model 2	year + length_log + budget + votes_log + genre	0.8755	0.9136
Model 3	length_log + budget + votes_log + genre	0.8755	0.9158
Model 4	length_log + budget + genre	0.8713	0.9158
Model 5	votes_log + budget + genre	0.8420	0.9071

	Specificity	AUC	BIC
Model 1	0.7659	0.9372	1007.1164
Model 2	0.8056	0.9405	948.7532
Model 3	0.8016	0.9413	942.5078
Model 4	0.7897	0.9405	941.4027
Model 5	0.7222	0.9009	1192.5274

Model 4 was chosen for further tuning due to its balance between simplicity and performance. Despite not having the absolute highest accuracy, it provides high sensitivity (0.9158) and decent specificity (0.7897), alongside a strong AUC of 0.9405, indicating good discriminative power. With a BIC of 941.4027, it suggests efficiency in balancing model fit with complexity. This makes Model 4 a strong candidate for detailed analysis and threshold optimization.

4.3 Model Tuning

In fine-tuning our logistic regression model, particularly for an imbalanced dataset, we focus on optimizing the classification threshold. This involves systematically assessing various thresholds to find an optimal balance between Accuracy, Sensitivity, and Specificity. The goal is to improve model precision by correctly balancing true positive and negative predictions. This targeted approach ensures our model is better suited to the specific challenges and objectives of our analysis, thereby enhancing its predictive reliability and relevance.



Metric	Value
1 Accuracy	0.8923
2 Sensitivity	0.8898
3 Specificity	0.8968
4 AUC	0.9405
5 BIC	941.4027

The classification threshold of 0.33, as observed from the plot, optimally balances accuracy, sensitivity, and specificity. This threshold reflects a strategic compromise, enhancing the model's ability to correctly identify films rated above and below 7, without heavily sacrificing one metric for another.

Model 4, which utilizes length_log, budget, and genre as predictors, demonstrates strong predictive performance. With an accuracy of 89.23%, it effectively distinguishes between movies rated above and below 7. The model is equally balanced in terms of sensitivity (88.98%) and specificity (89.68%), indicating it is reliable in identifying both high-rated and lower-rated films. The AUC value of 0.9405 suggests good discrimination between the positive and negative classes. Furthermore, a BIC of 941.4027 reflects the model's efficiency, balancing model complexity with fit to the data.

4.4 Model Interpretation

```
Call:
```

```
glm(formula = above_7 ~ length_log + budget + genre, family = binomial,
    data = train_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.66547    1.07287   3.416 0.000634 ***
length_log        -2.79886    0.24870 -11.254  < 2e-16 ***
budget             0.54983    0.03966  13.865  < 2e-16 ***
genreAnimation    -1.61972    0.59253  -2.734 0.006265 **
genreComedy        2.66538    0.21699  12.283  < 2e-16 ***
genreDocumentary   4.77889    0.46931  10.183  < 2e-16 ***
genreDrama        -2.54760    0.36048  -7.067 1.58e-12 ***
genreRomance     -13.95741  432.52053  -0.032 0.974257
genreShort         3.60198    1.08387   3.323 0.000890 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2169.73  on 1671  degrees of freedom
Residual deviance:  874.61  on 1663  degrees of freedom
AIC: 892.61

Number of Fisher Scoring iterations: 14
```

1. **Length of Movies (length_log)**: There is a significant negative relationship between the log-transformed length of movies and their likelihood of being rated above 7. This suggests that longer movies are less likely to receive high ratings, potentially indicating viewer preferences for shorter films or perhaps an association with certain film types or genres that are longer but less popular.

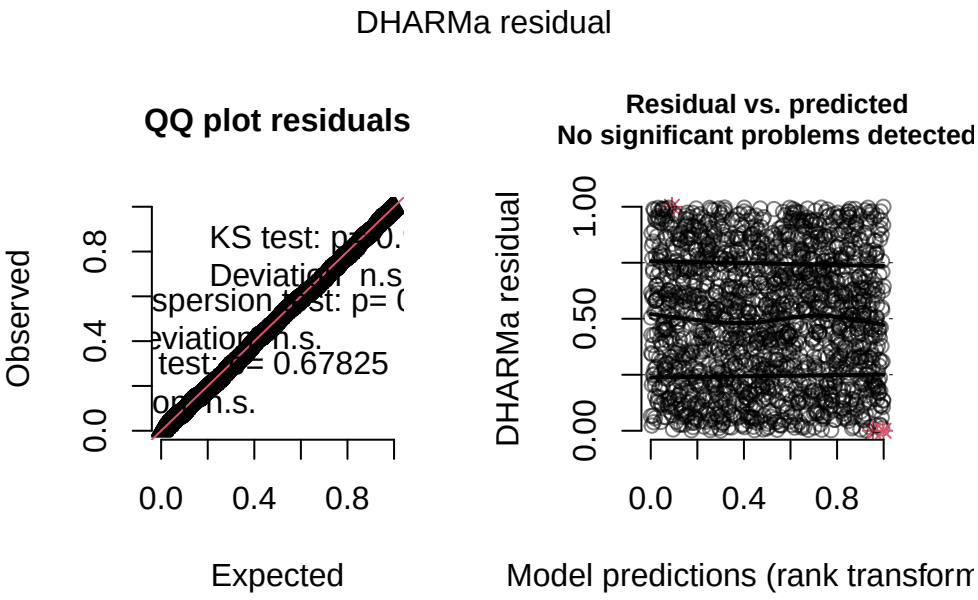
2. **Budget (budget)**: The budget of a movie shows a significant positive association with the likelihood of being rated above 7. This might imply that higher-budget movies, which can

afford better production quality, actors, and marketing, are more likely to be well-received by audiences.

3. **Genre**:

   - **Animation and Drama**: Compared to the baseline genre (action), animation and drama films are significantly less likely to be rated above 7.
   - **Short, Comedy, and Documentary**: These genres show higher probability of receiving high ratings compared to action, suggesting they are generally well-received or cater to specific audience segments that rate them favorably.
   - **Romance**: This genre do not show significant effects, possibly due to a smaller sample size, less variation in ratings, or other model limitations.

4.5 Assumption Checking



The DHARMa diagnostic plots for the generalized linear model indicate that the model assumptions are largely satisfied. The quantile-quantile plot reveals that residuals closely follow the expected uniform distribution, suggesting that the model does not exhibit significant misfit. The uniformity is further supported by a high p-value in the Kolmogorov-Smirnov test, indicating no significant deviation from the expected distribution. The residuals versus predicted values plot shows no discernible pattern, implying consistent variance across predicted values and supporting the ho-

moscedasticity assumption. Although there are a few outliers, they don't appear to systematically affect the model's validity. Overall, the analysis suggests that the model is well-fitted to the data under the constraints of GLM assumptions.

5 Conclusion

This analysis has shed light on the influential dynamics between film attributes and audience ratings. Specifically, it has been observed that film length, budget, and genre play pivotal roles in determining a movie's rating. Notably, shorter films tend to receive higher ratings, highlighting the audience's preference for narrative conciseness and the ability to maintain engagement. Similarly, films with higher budgets are generally correlated with ratings above 7, indicating that substantial financial investment in production quality, star power, and marketing can significantly influence viewer perceptions and satisfaction. Among the various genres analyzed, short films and documentaries have demonstrated particularly high success rates, suggesting that niche audiences or the unique storytelling and educational aspects of these genres resonate well with viewers. On the other hand, genres such as drama and animation appear to face greater challenges in achieving high ratings, potentially due to genre-specific expectations or saturated market conditions.

These findings provide valuable insights into the multifaceted nature of film success. They suggest that filmmakers, producers, and studios should consider a holistic approach when planning new projects. By balancing the essential elements of film length, budget allocation, and genre selection, and aligning them with target audience preferences and market trends, filmmakers can optimize their strategies to enhance audience reception and increase the likelihood of producing critically acclaimed and commercially successful films.

6 Discussion

6.1 Implications

This study highlights key factors influencing audience ratings: film length, budget, and genre. Findings suggest filmmakers should focus on concise storytelling, invest in production quality, and select genres wisely to enhance film success. This also impacts marketing strategies, advocating for targeted campaigns, especially for niche genres like documentaries.

6.2 Limitation

The study's limitations include reliance on viewer ratings as the sole success metric and a data set that may not reflect all aspects of movie success. Additionally, the dynamic nature of audience preferences and unobserved variables like directorial style or societal trends could affect outcomes.

6.3 Further Research

Future research could extend to analyzing film success through both audience ratings and box office revenue, offering a comprehensive view of what defines success in the film industry. Additionally, examining subtleties within film genres—such as specific sub-genres, themes, or narrative structures—could uncover elements that significantly impact a film's success. This detailed analysis would help filmmakers and studios better align their projects with audience preferences and market trends, potentially leading to higher ratings and greater financial success.