Behind the Curtain: Statistical Insights into Movie Success

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

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. The selection of variables such as 'length', 'budget', and 'votes' is predicated on the hypothesis that these factors collectively encapsulate elements of narrative compactness, production quality, and audience engagement—each a potential predictor of a film's rating.

As we navigate 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 pre-processed, 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—namely, length, budget, viewer engagement (votes), and genre—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 residual analysis to evaluate 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

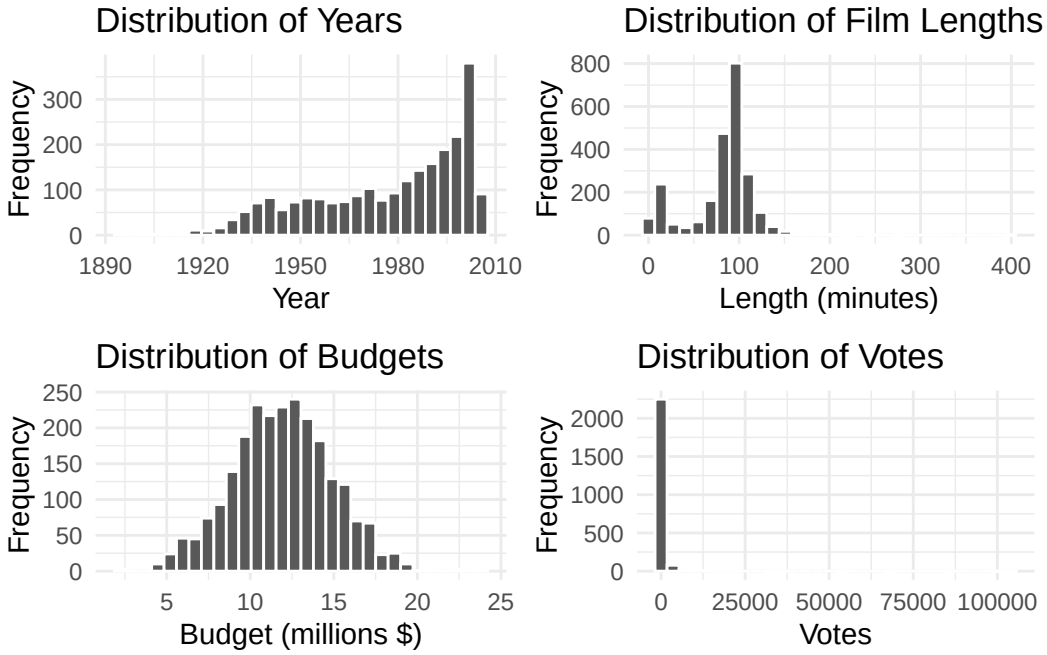
Statistical Summary of Categorical Variables

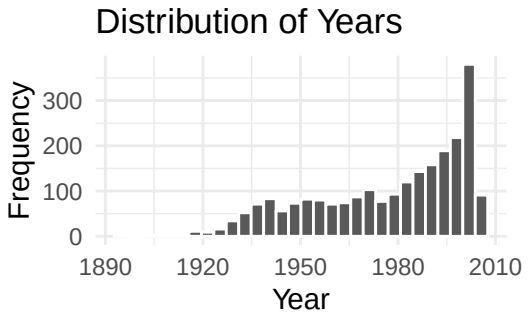
variable	n_levels	levels	
genre	7	Action, Animation, Comedy, Documentary, Drama, Romance, Short	698, 165, 582, 13
above_7	2	0, 1	

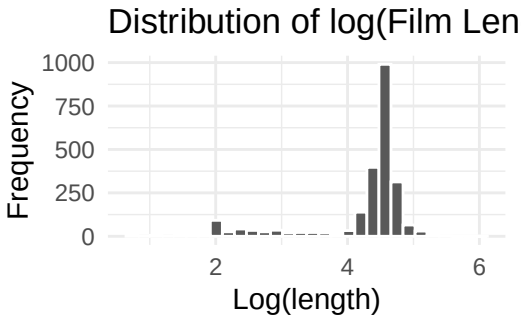
3.2 Outliers

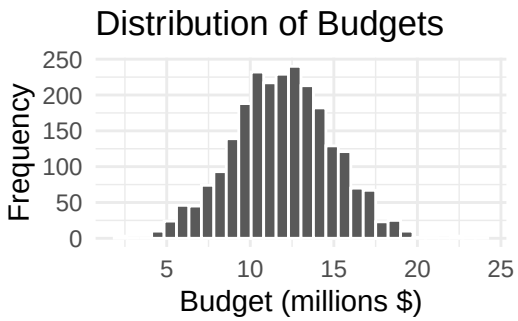
```
                          length      budget     votes
Proportion of Outliers 0.1805614 0.004608295 0.1625471
```

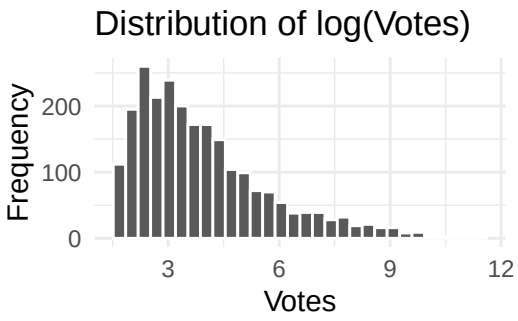
3.3 Visualisation

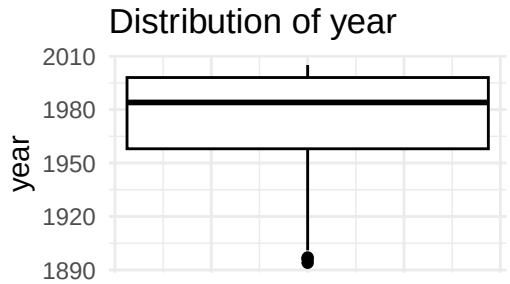


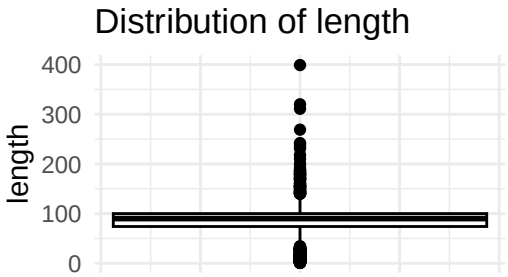

Distribution of Years
Frequency
300
200
100
0
1890
1920
1950
1980
2010
Year

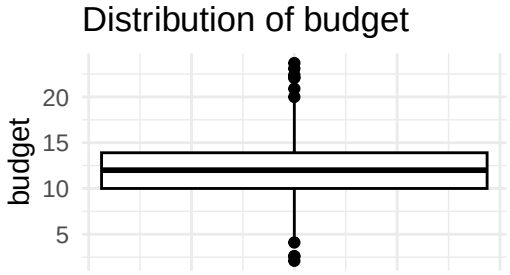

Distribution of log(Film Len
Frequency
1000
750
500
250
0
2
4
6
Log(length)

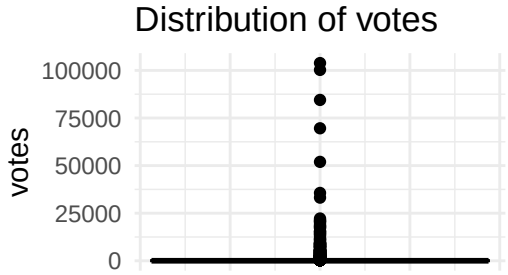

Distribution of Budgets
Frequency
250
200
150
100
50
0
5
10
15
20
25
Budget (millions $)

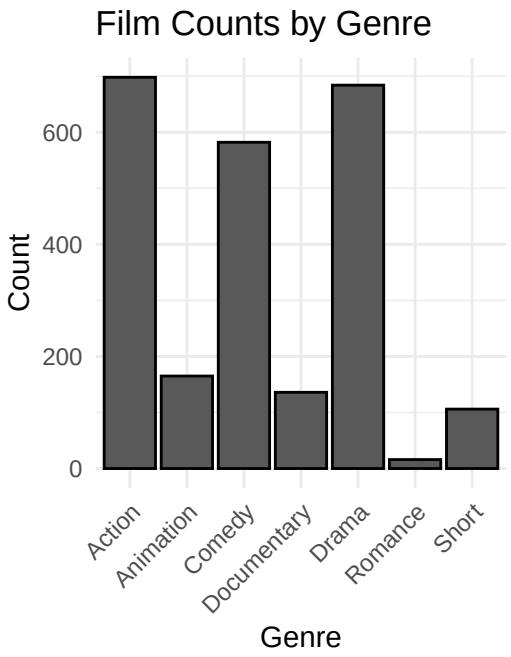

Distribution of log(Votes)
Frequency
200
100
0
3
6
9
12
Votes

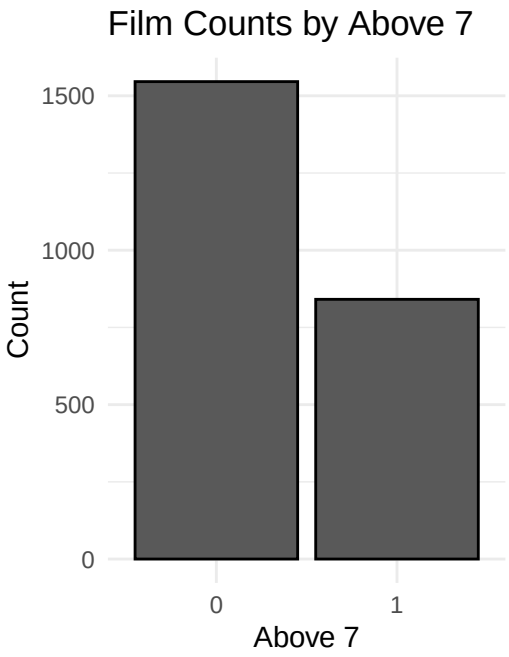

Distribution of year
year
2010
1980
1950
1920
1890

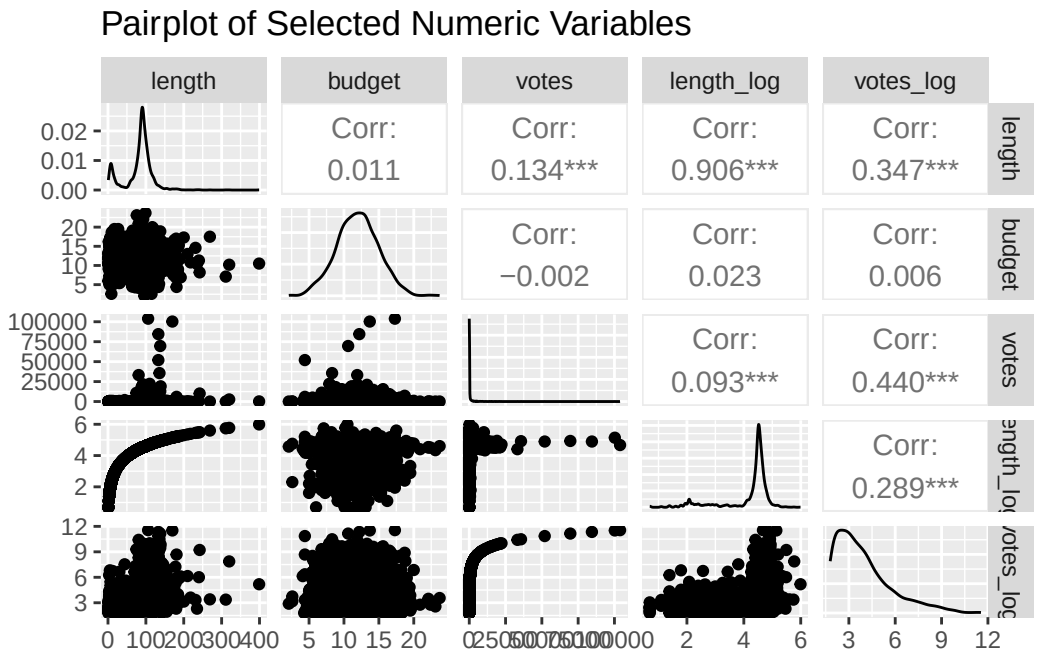

Distribution of length
length
400
300
200
100
0

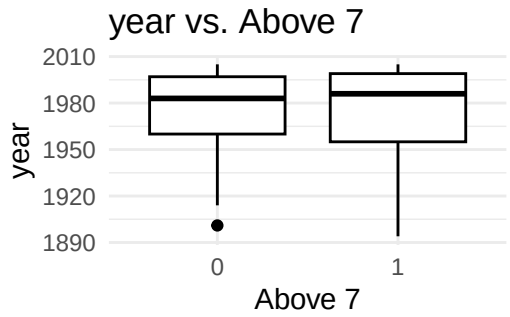

Distribution of budget
budget
20
15
10
5

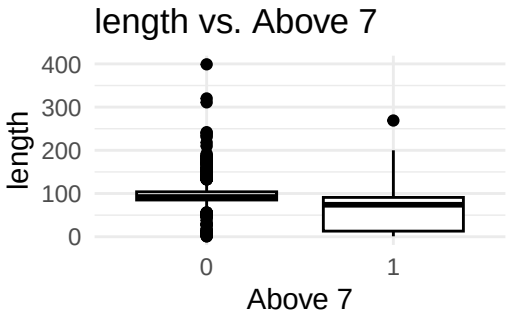

Distribution of votes
votes
100000
75000
50000
25000
0

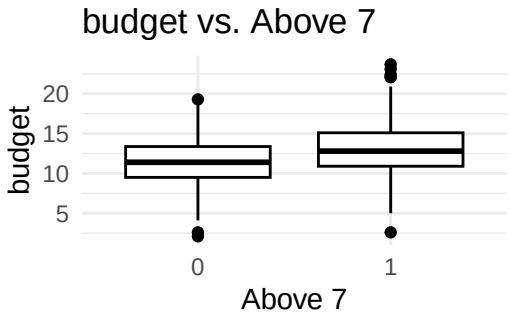

Film Counts by Genre
Count
600
400
200
0
Action
Animation
Comedy
Documentary
Drama
Romance
Short
Genre

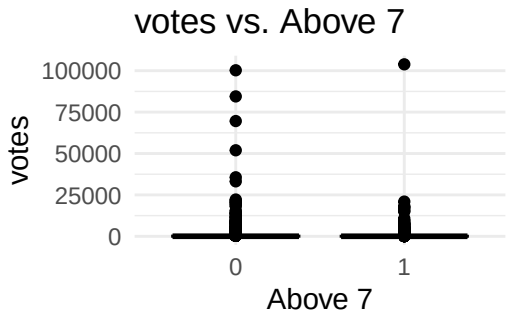

Film Counts by Above 7
Count
1500
1000
500
0
0
1
Above 7

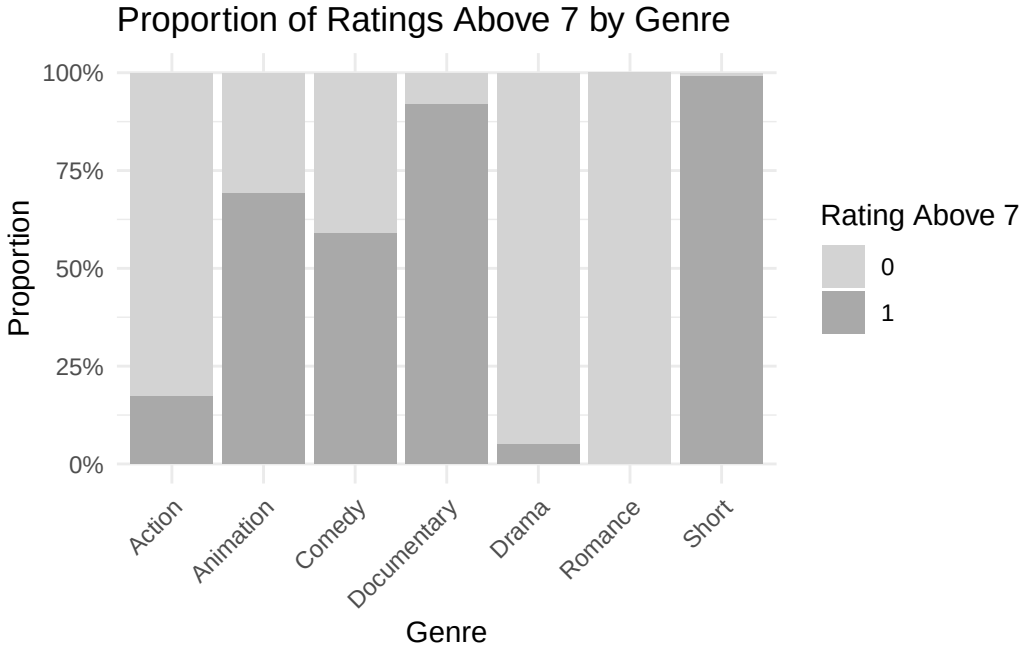

Pairplot of Selected Numeric Variables
length
budget
votes
length_log
votes_log
Corr:
0.011
Corr:
0.134***
Corr:
0.906***
Corr:
0.347***
Corr:
−0.002
Corr:
0.023
Corr:
0.006
Corr:
0.093***
Corr:
0.440***
Corr:
0.289***

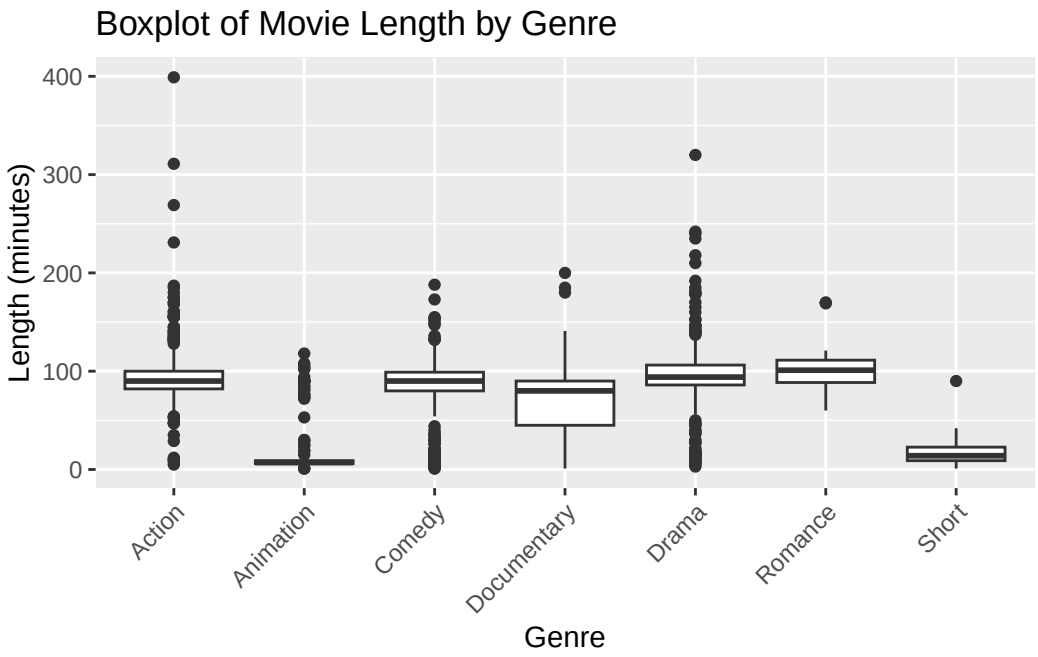

year vs. Above 7
year
2010
1980
1950
1920
1890
0
1
Above 7

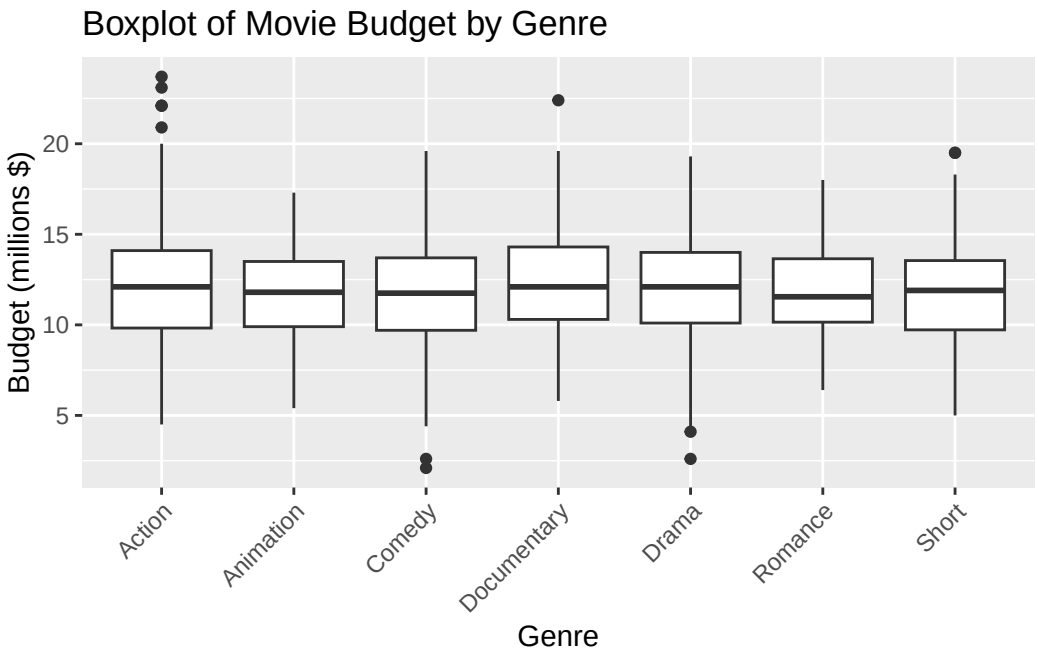

length vs. Above 7
length
400
300
200
100
0
0
1
Above 7

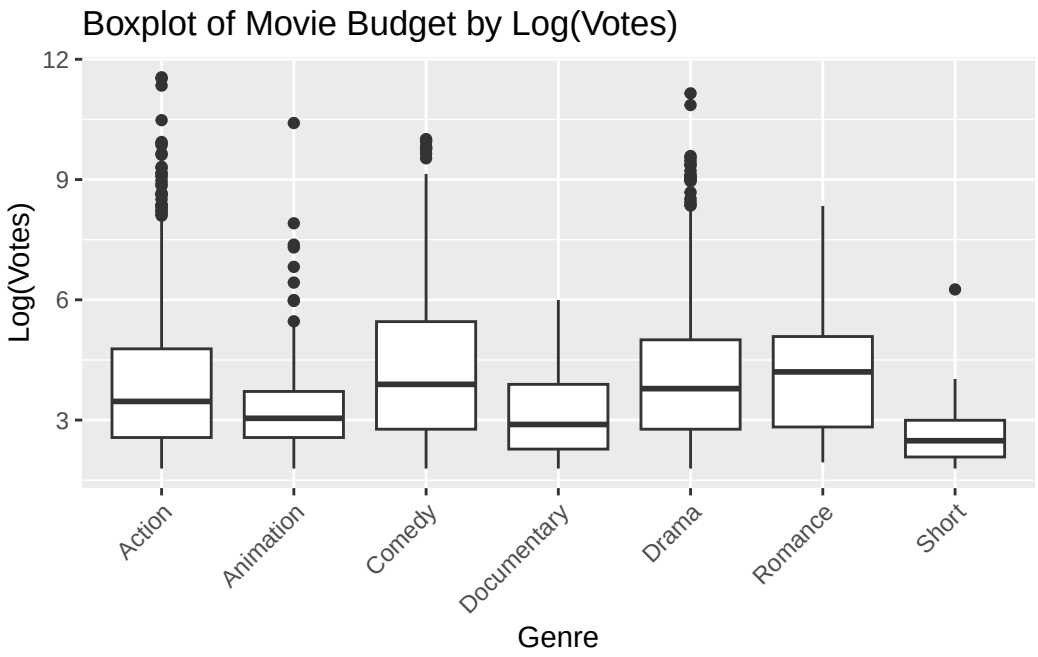

budget vs. Above 7
budget
20
15
10
5
0
1
Above 7

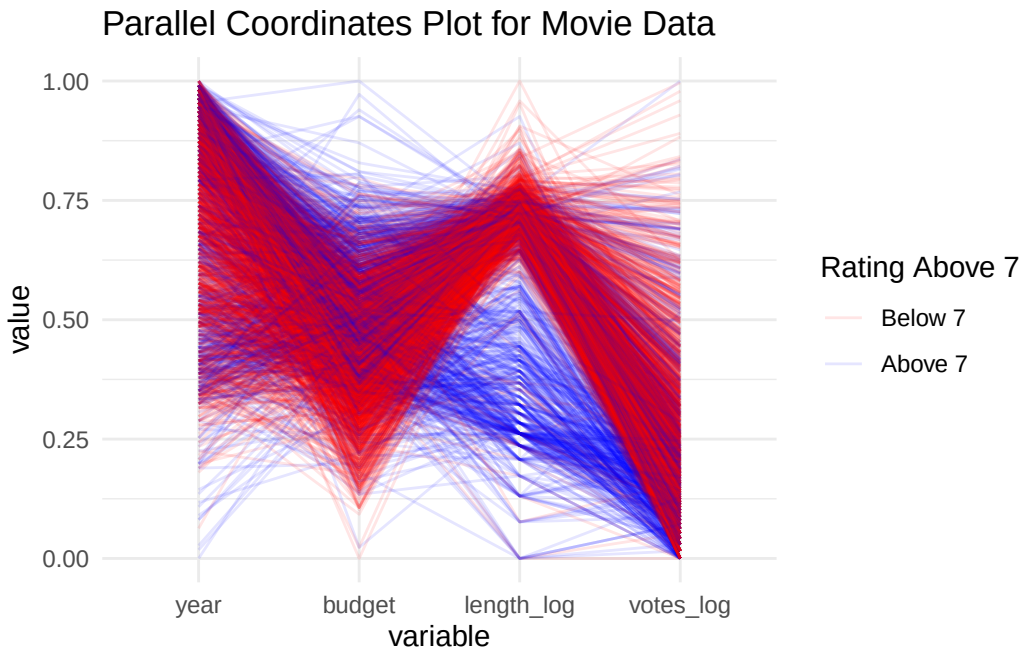

votes vs. Above 7
votes
100000
75000
50000
25000
0
0
1
Above 7


Proportion of Ratings Above 7 by Genre
Proportion
100%
75%
50%
25%
0%
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 Movie Budget by Log(Votes)
Log(Votes)
12
9
6
3
Action
Animation
Comedy
Documentary
Drama
Romance
Short
Genre


Parallel Coordinates Plot for Movie Data
value
1.00
0.75
0.50
0.25
0.00
year
budget
length_log
votes_log
variable
Rating Above 7
Below 7
Above 7

3.4 EDA Findings

In our 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. Conversely, budgets appear nearly normally distributed, indicating diverse financial investments across films. The 'votes' distribution is significantly right-skewed, highlighting a disparity in viewer engagement.

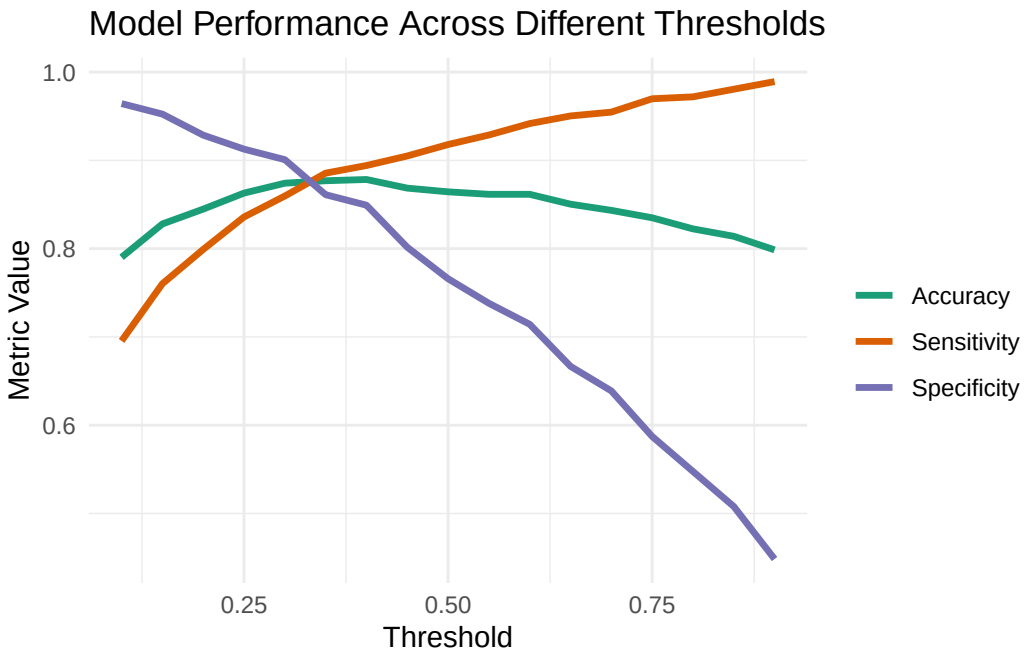
After log transformations, the distributions of 'length' and 'votes' approached closer to normality but still exhibited skewness. The data set predominantly features action, drama, and comedy genres, with fewer romantic and short films. Notably, only 35% of movies are rated above 7.

There is a medium positive correlation between log-transformed votes and length, suggesting films of longer duration may engage viewers more. Budget analyses indicate movies rated above 7 typically have higher budgets. Genre-wise, documentaries stand out with a highest proportion 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. Despite uniform budget distribution across genres, 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 Find Optimal Threshold

In determining the optimal classification threshold for our logistic regression model, especially when faced with an imbalanced target variable, a systematic approach is adopted. A range of potential thresholds is evaluated to assess their impact on key performance metrics: Accuracy, Sensitivity, and Specificity. This process allows for the identification of a threshold that strikes the best balance between correctly identifying true positives and true negatives. The objective is to enhance the model's predictive power and ensure a more informed, context-specific application, particularly important in scenarios where accurate classification holds significant consequences. This methodology ensures that the chosen threshold aligns with the specific needs and goals of the analysis, addressing the challenges posed by an imbalanced dataset.



Based on the graph, we decide to set threshold to 0.32 for best balance of accuracy, sensitivity, and specificity.

4.2 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
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 the Area Under the Curve (AUC).

4.3 Model Selection

```
                                                Variables Accuracy Sensitivity
Model 1         year + length + budget + votes + genre   0.8727      0.8683
Model 2 year + length_log + budget + votes_log + genre   0.8825      0.8812
Model 3        length_log + budget + votes_log + genre   0.8853      0.8855
Model 4                     length_log + budget + genre   0.8937      0.8898
Model 5                      votes_log + budget + genre   0.8503      0.8531
        Specificity    AUC       BIC
Model 1      0.8810 0.9372 1007.1164
Model 2      0.8849 0.9405  948.7532
Model 3      0.8849 0.9413  942.5078
Model 4      0.9008 0.9405  941.4027
Model 5      0.8452 0.9009 1192.5274
```

Model 4, featuring log-transformed film length, budget, and genre, is selected as the optimal model due to its superior balance of performance and simplicity. Exhibiting the highest specificity (0.9008) among the evaluated models, it effectively identifies films not surpassing the rating threshold. Its accuracy (0.8937) and sensitivity (0.8898) are commendable, with an AUC value of 0.9405 indicating strong discriminative power. The reduced Bayesian Information Criterion (BIC) of 941.4027 suggests efficient modeling with fewer predictors, underlining its effectiveness without undue complexity.

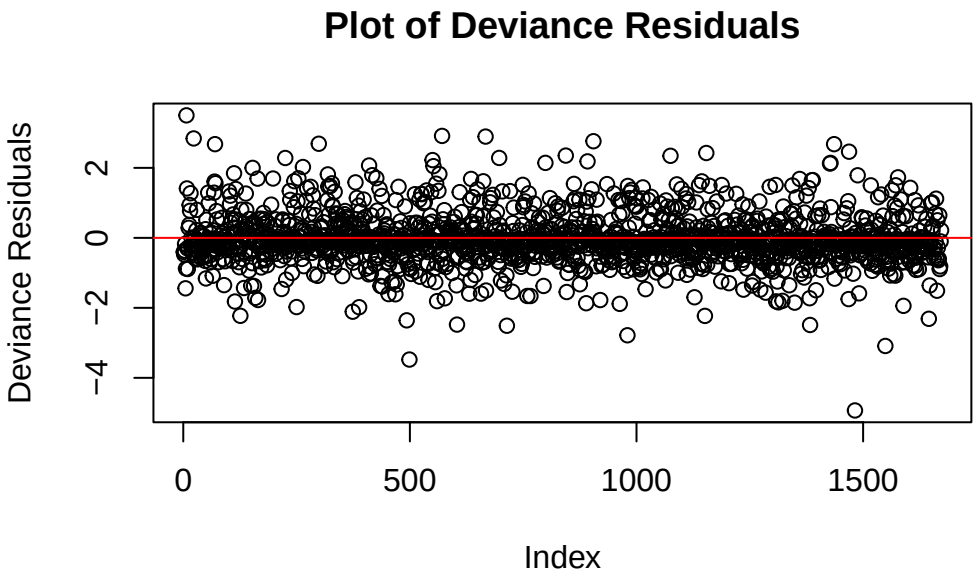
4.4 Model Interpretation

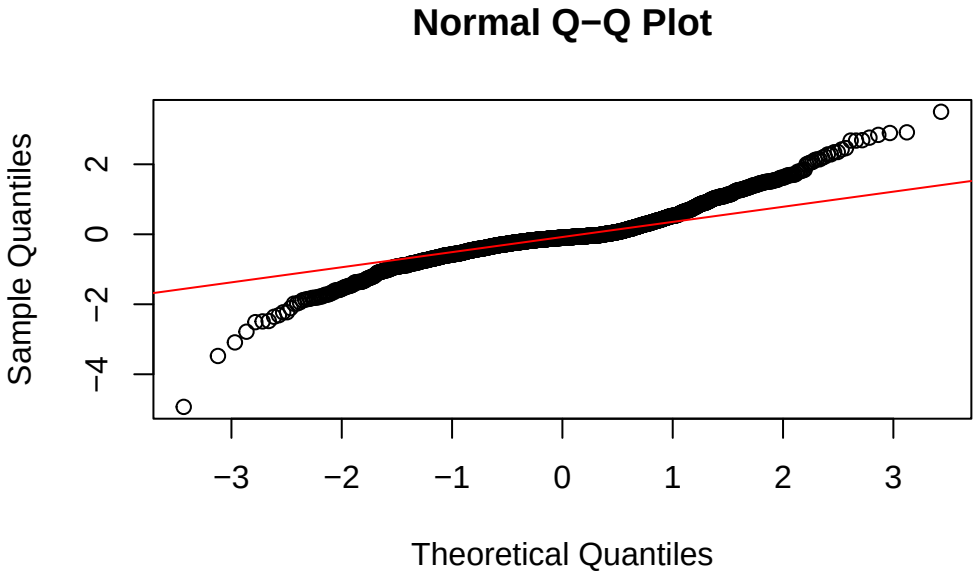
```
Call:
glm(formula = above_7 ~ length_log + budget + genre, family = binomial,
```

```
    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 Differences**:
    - **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 Residual Analysis





The residual analysis indicates that the model is reasonably well-fitted, with residuals displaying no systematic bias and consistent spread. However, the presence of outliers and deviations from normality in the tails, as shown in the Q-Q plot, suggest that the data may have more extreme values than a standard normal distribution would predict. This implies that while the model generally captures the data's central tendency, it may need refinement to better accommodate the extreme values or outliers observed.

5 Conclusion

Our analysis reveals that film length, budget, viewer votes, and genre significantly impact movie ratings. Specifically, shorter films, higher budgets, and increased viewer engagement (as measured by votes) are positively correlated with ratings above 7, underscoring the importance of narrative conciseness, financial investment, and audience interaction in cinematic success. Among genres, documentaries stand out for their high proportion of well-rated films, while action, drama, and romance show varying levels of success. These insights underscore a multifaceted approach to predicting film success, suggesting that filmmakers can enhance audience reception by strategically balancing these key factors within the creative and production processes.

6 Discussion

6.1 Practical Implications

- Filmmakers and producers can leverage insights from this model, particularly around film length, budget, and targeted genre, to optimize their projects for higher audience ratings.
- The significant predictors offer a blueprint for aligning movie projects with characteristics correlated with success, though considerations of artistic intent and narrative integrity remain paramount.

6.2 Further Research

- The disparities observed in genre impacts necessitate deeper investigation, potentially requiring broader datasets to ensure nuanced understandings.
- Future research should address the data limitations, particularly for underrepresented genres, and explore external factors beyond the scope of the current model to provide a more comprehensive understanding.