Behind the Curtain: Statistical Insights into Movie Success

Code

AUTHOR
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, in the age of data-driven decision-making, this project leverages a Generalized Linear Model (GLM) to systematically evaluate these factors, offering a more empirical basis for understanding cinematic success.

Our dataset 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 logistic 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, our 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.

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

Overview

```
Rows: 2,387
Columns: 8
$ film_id <int> 39891, 33810, 20282, 33131, 50633, 37020, 55337, 28037, 13291,…
$ year    <int> 2003, 2004, 1941, 1959, 1917, 1934, 2003, 1988, 1981, 1935, 19…
$ length  <int> 75, 120, 78, 106, 70, 64, 91, 101, 78, 7, 21, 90, 99, 101, 110…
$ budget  <dbl> 10.9, 19.6, 11.7, 12.0, 14.8, 11.6, 12.6, 10.1, 14.2, 6.6, 10.…
$ votes   <int> 17, 21, 14, 14, 9, 8, 182, 274, 61, 10, 5, 8, 349, 24, 20168, …
$ genre   <fct> Action, Documentary, Action, Drama, Drama, Drama, Comedy, Acti…
$ rating  <dbl> 4.4, 7.3, 2.7, 4.9, 5.6, 4.7, 4.4, 4.3, 4.3, 8.8, 7.3, 8.3, 7.…
$ above_7 <fct> 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1,…
```

Statistical Summary

```
    film_id           year          length           budget     
 Min.   :   33   Min.   :1894   Min.   :  1.00   Min.   : 2.10  
 1st Qu.:14799   1st Qu.:1958   1st Qu.: 74.00   1st Qu.:10.00  
 Median :30259   Median :1984   Median : 90.00   Median :12.00  
 Mean   :29942   Mean   :1977   Mean   : 81.75   Mean   :11.95  
 3rd Qu.:44670   3rd Qu.:1998   3rd Qu.:100.00   3rd Qu.:13.90  
 Max.   :58780   Max.   :2005   Max.   :399.00   Max.   :23.70  

     votes                genre         rating      above_7    
 Min.   :     5   Action     :698   Min.   :0.700   0:1546  
 1st Qu.:    12   Animation  :165   1st Qu.:3.700   1: 841  
 Median :    32   Comedy     :582   Median :4.700           
 Mean   :   659   Documentary:136   Mean   :5.414           
 3rd Qu.:   118   Drama      :684   3rd Qu.:7.800           
 Max.   :103854   Romance    : 16   Max.   :9.200           
                  Short      :106                           
```

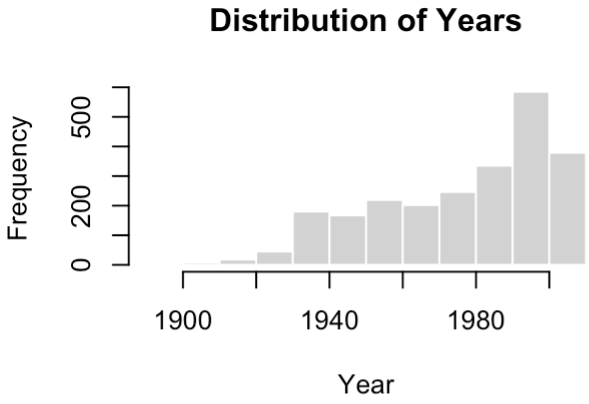
Outliers

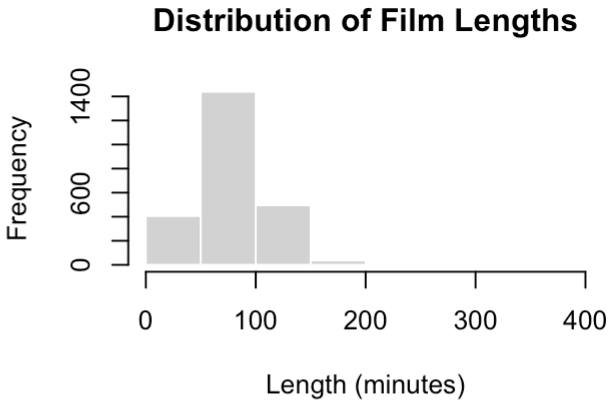
	length	budget	votes
Proportion of Outliers	0.1805614	0.004608295	0.1625471

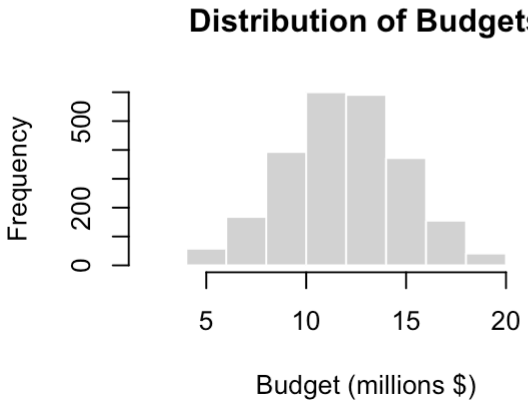
Outliers after log transformation

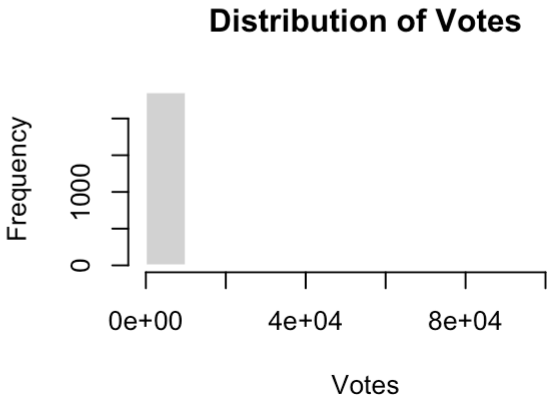
	length_log	votes_log
Proportion of Outliers	0.1818182	0.03812317

Numerical variable distribution (Histogram)

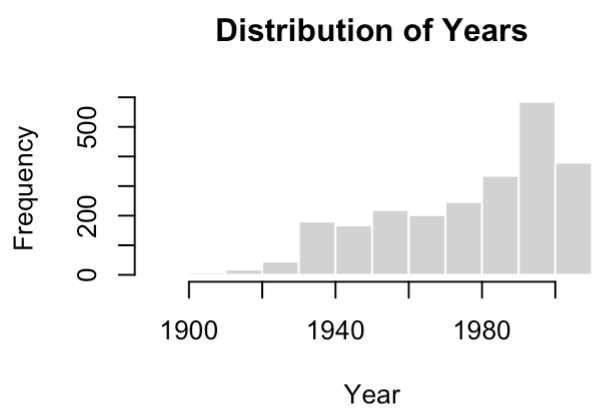


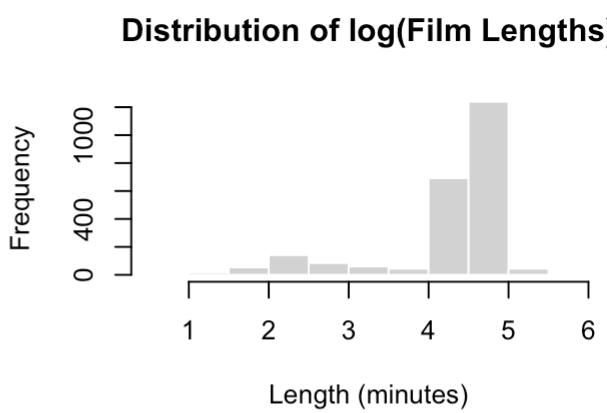


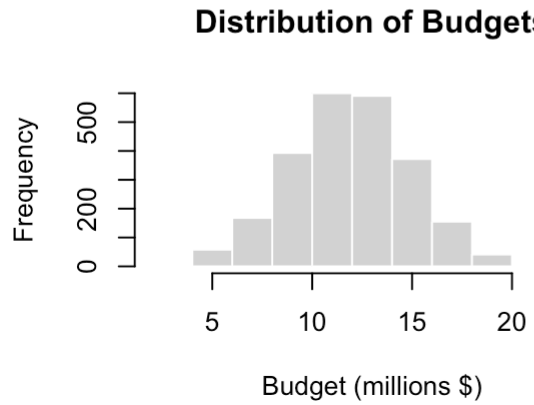


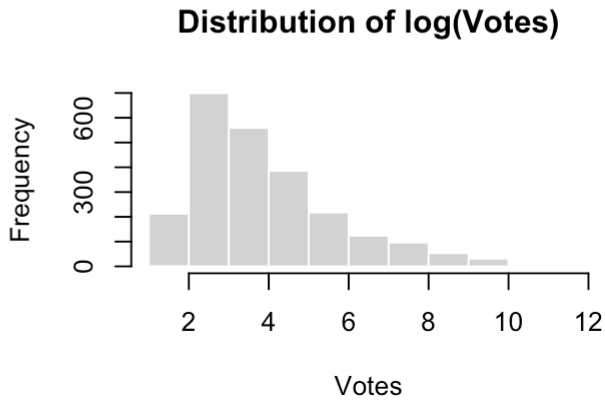


Numerical variable distribution after log (Histogram)

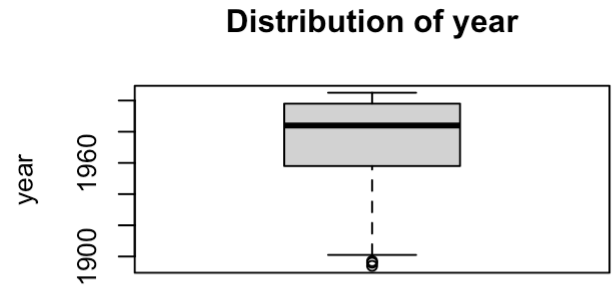


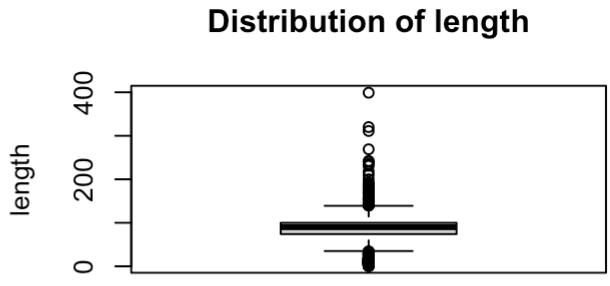


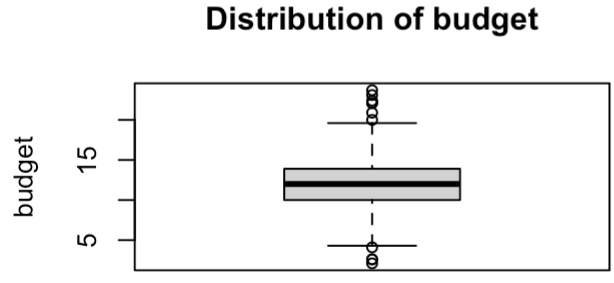


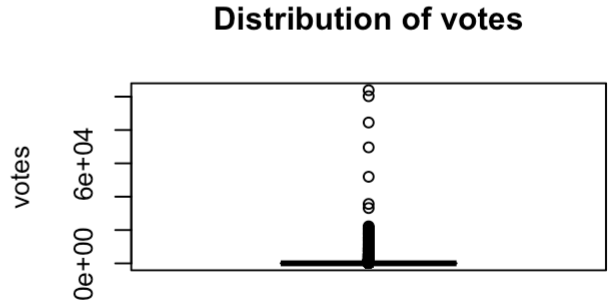


Numerical variable distribution (Boxplot)

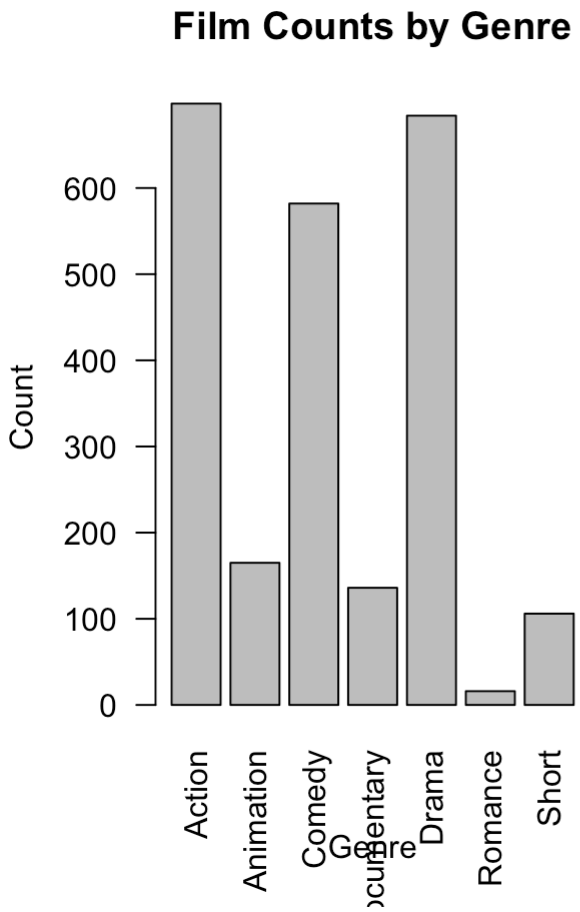


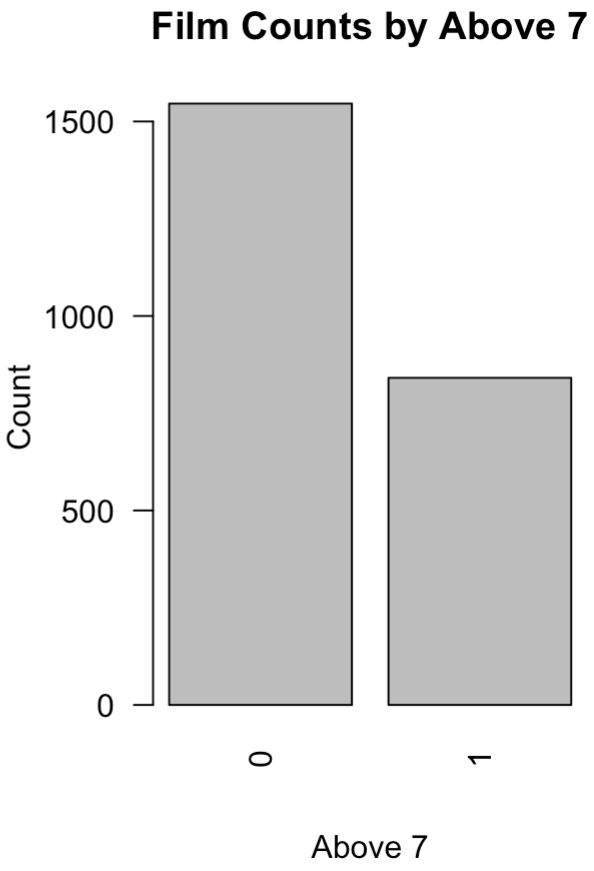




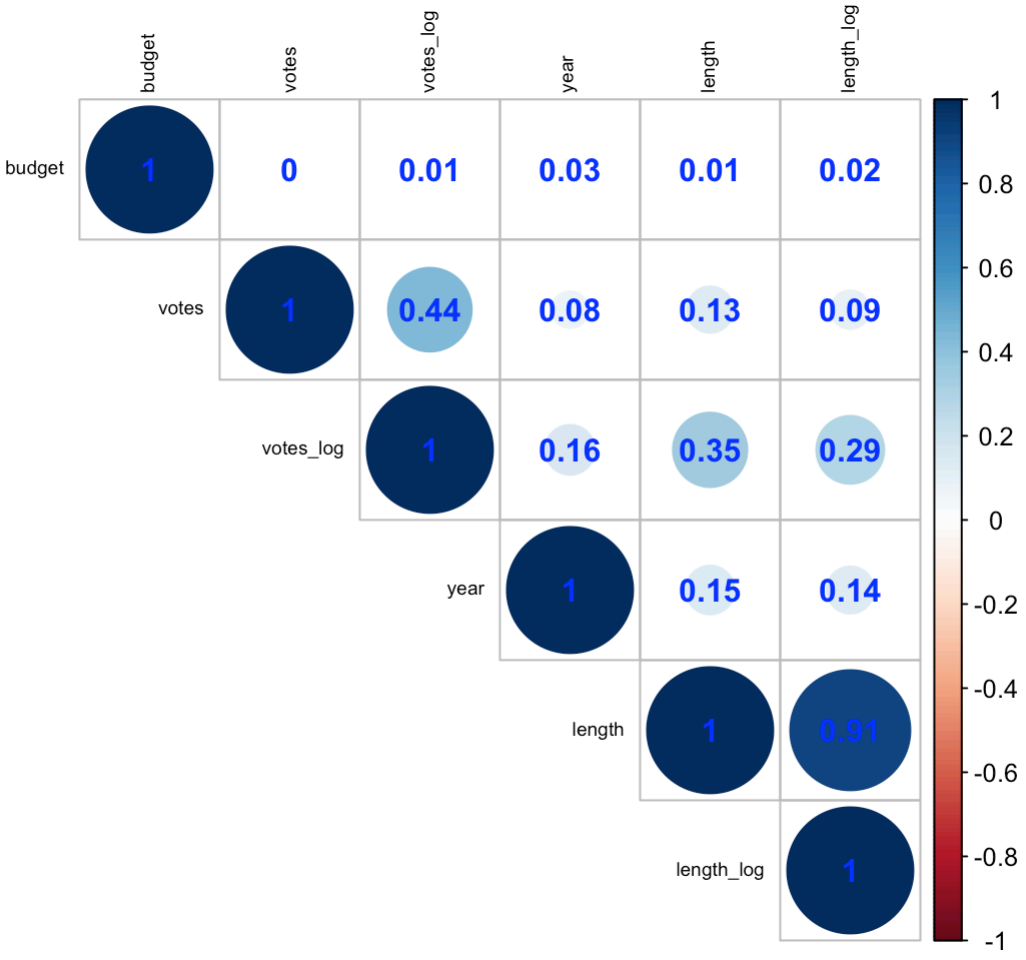


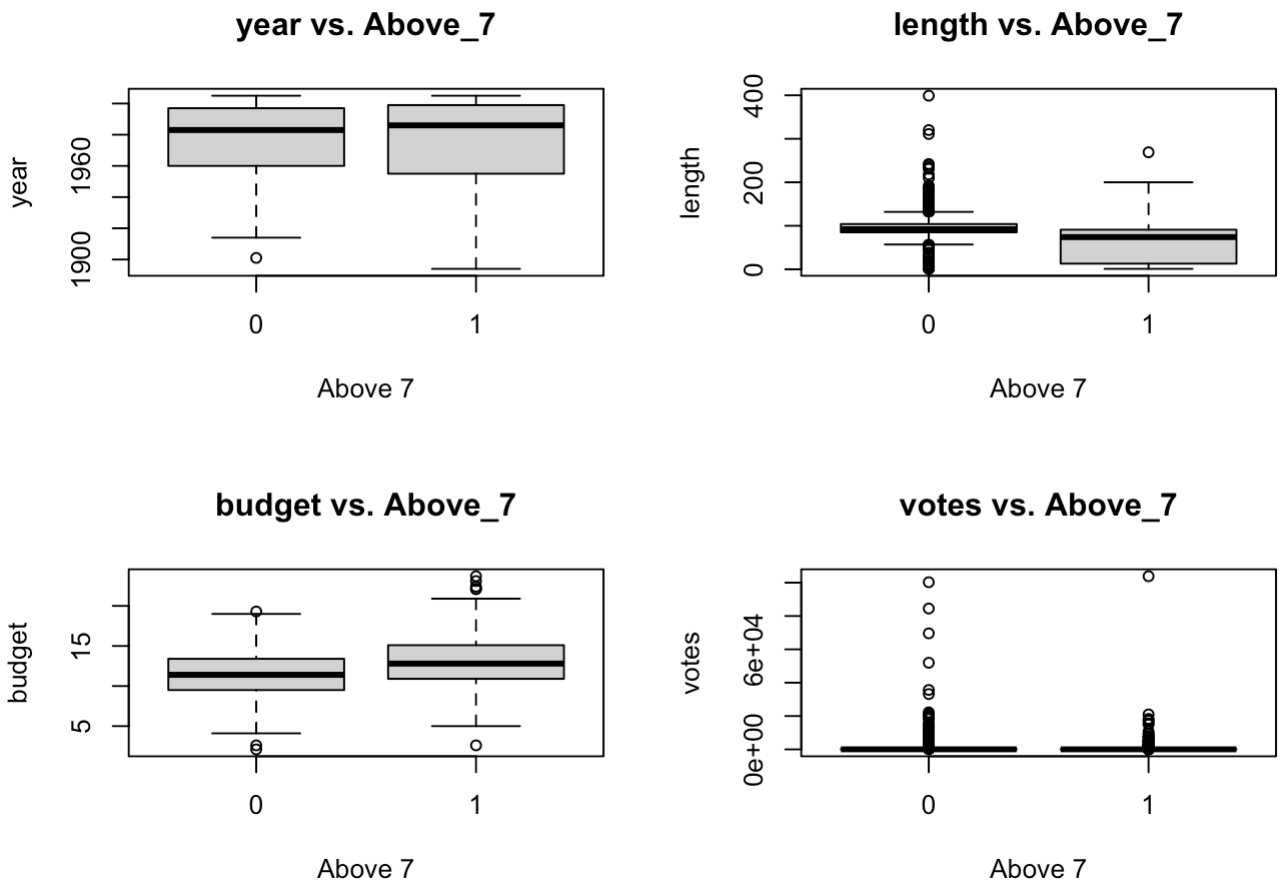
Count plot of categorical variables


Film Counts by Genre
Count
0
100
200
300
400
500
600
Action
Animation
Comedy
Drama
Romance
Short
Genre

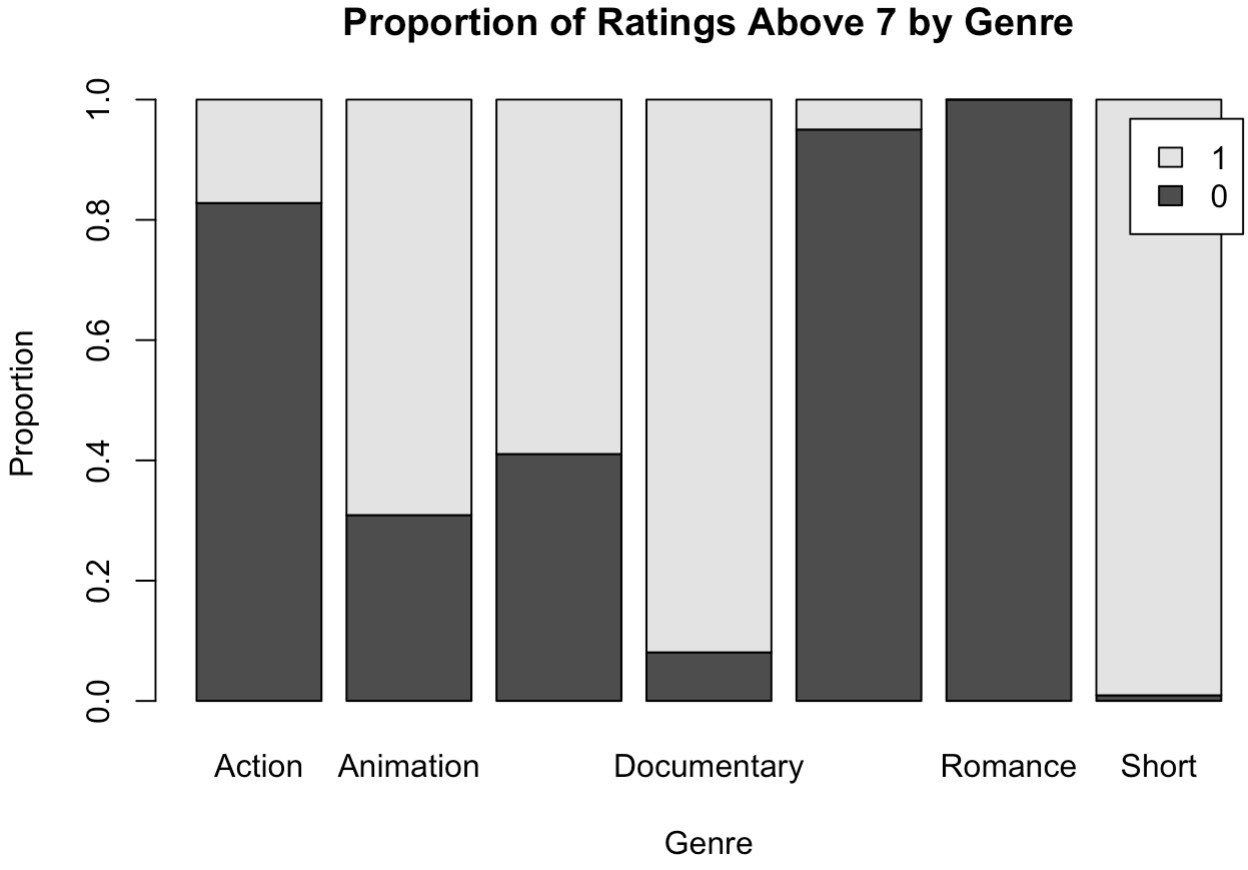

Film Counts by Above 7
Count
0
500
1000
1500
0
1
Above 7

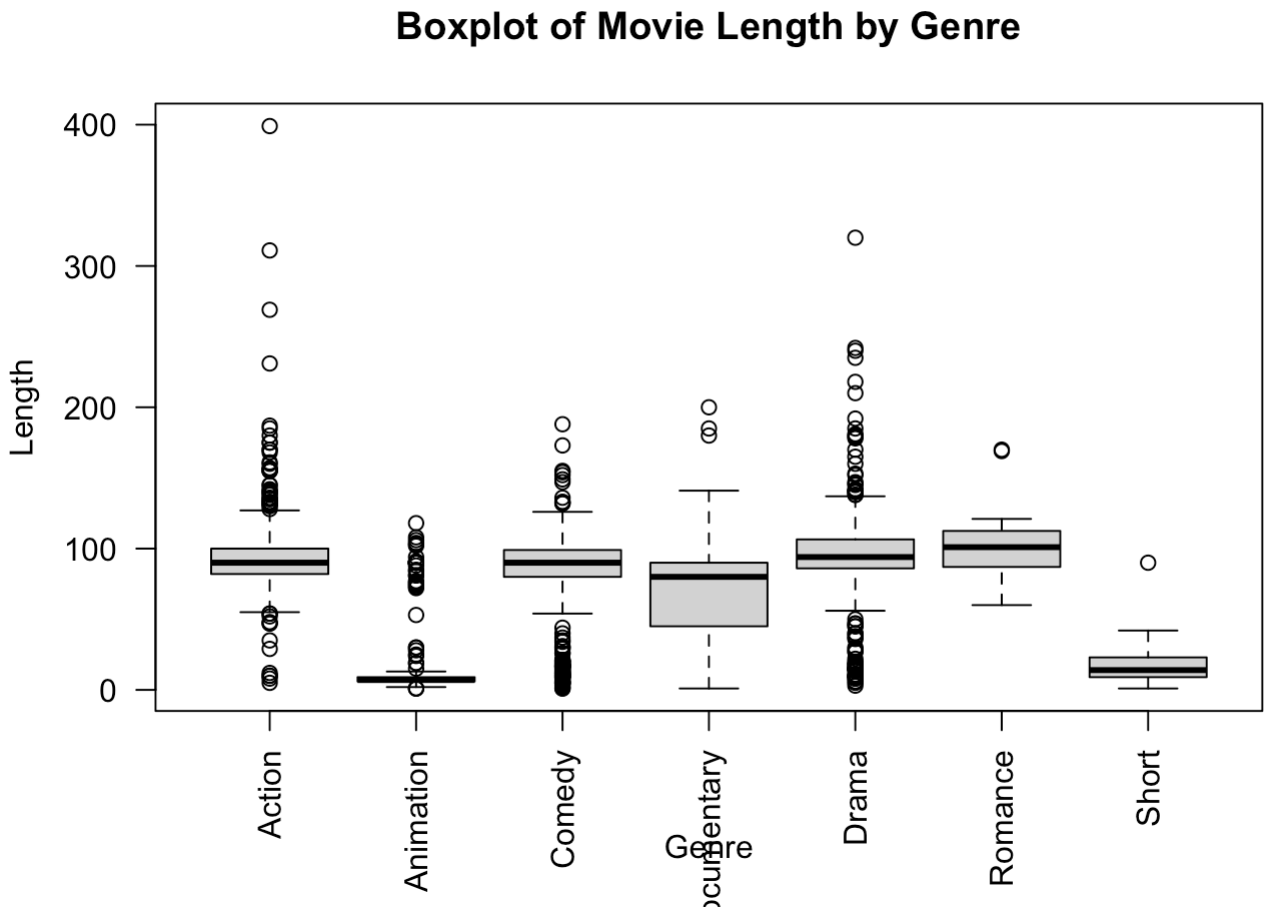
Pairplot

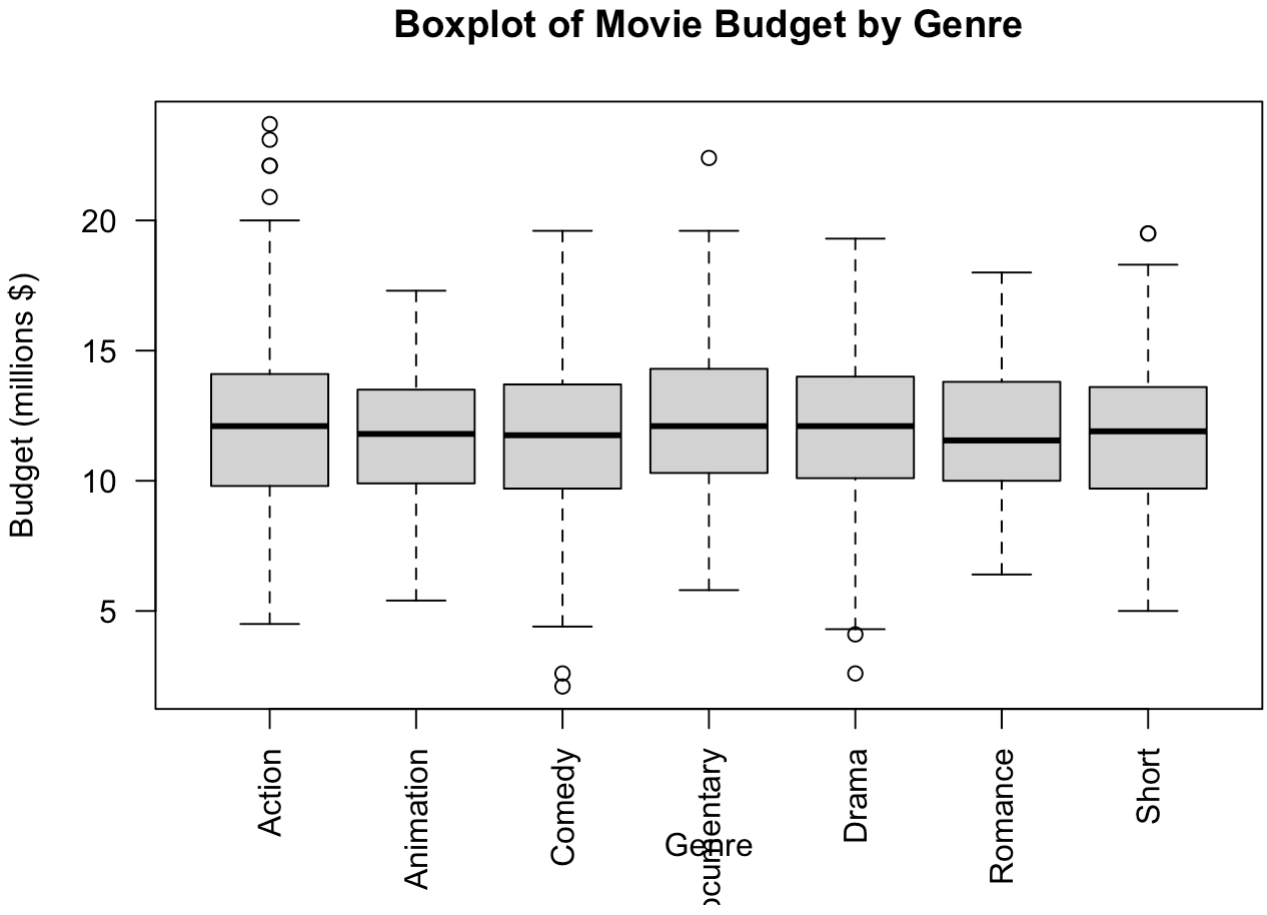

budget
votes
votes_log
year
length
length_log
1
0
0.01
0.03
0.01
0.02
1
0.44
0.08
0.13
0.09
1
0.16
0.35
0.29
1
0.15
0.14
1
0.91
1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

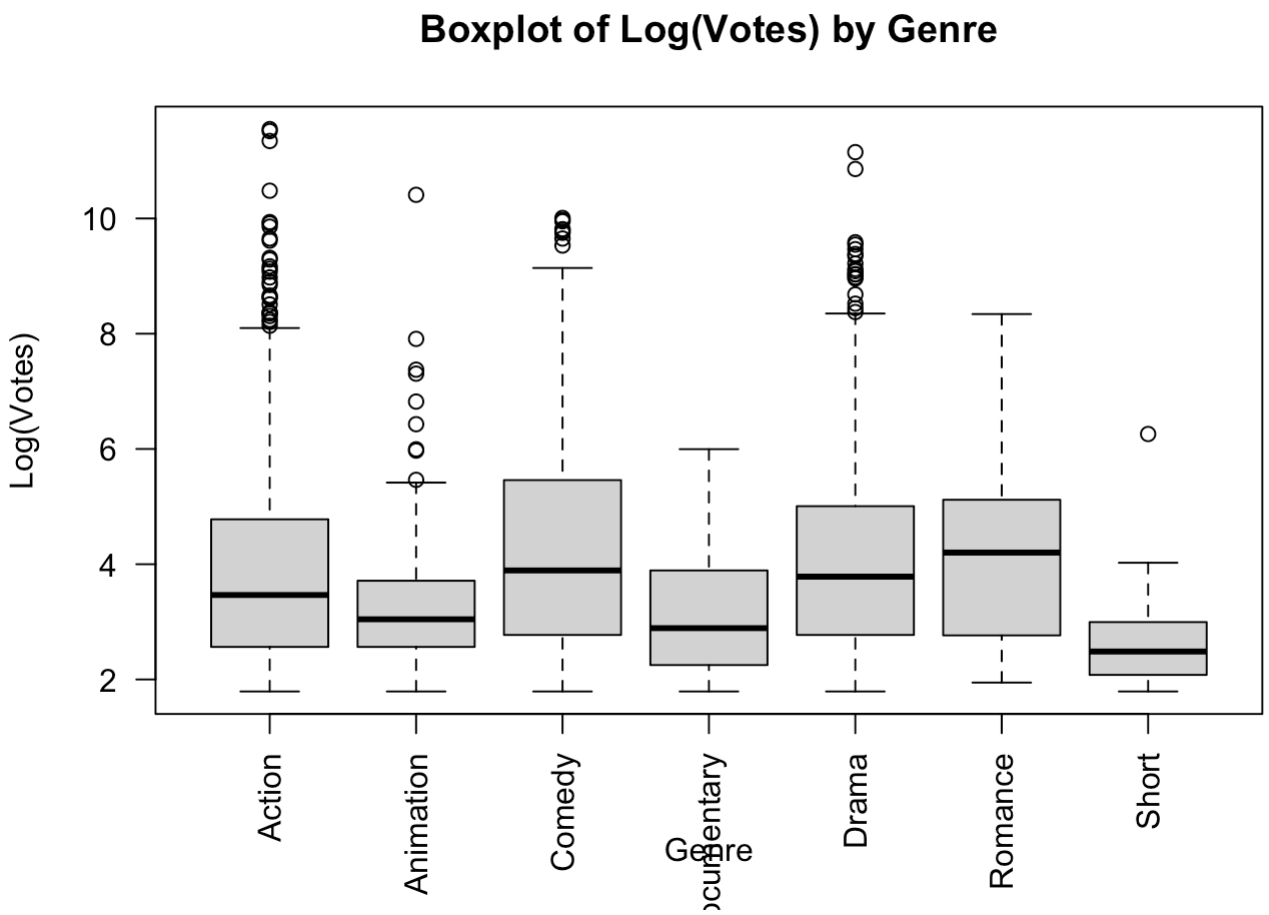


Proportion of ratings above 7 by genre




Boxplot of Movie Length by Genre
Length
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)
10
8
6
4
2
Action
Animation
Comedy
Drama
Romance
Short
Genre


Parallel Coordinates Plot for Movie Data without Genre
1.00
0.75
0.50
0.25
0.00
value
year
budget
length_log
votes_log
variable
Rating Above 7
Below 7
Above 7

3.1 EDA findings

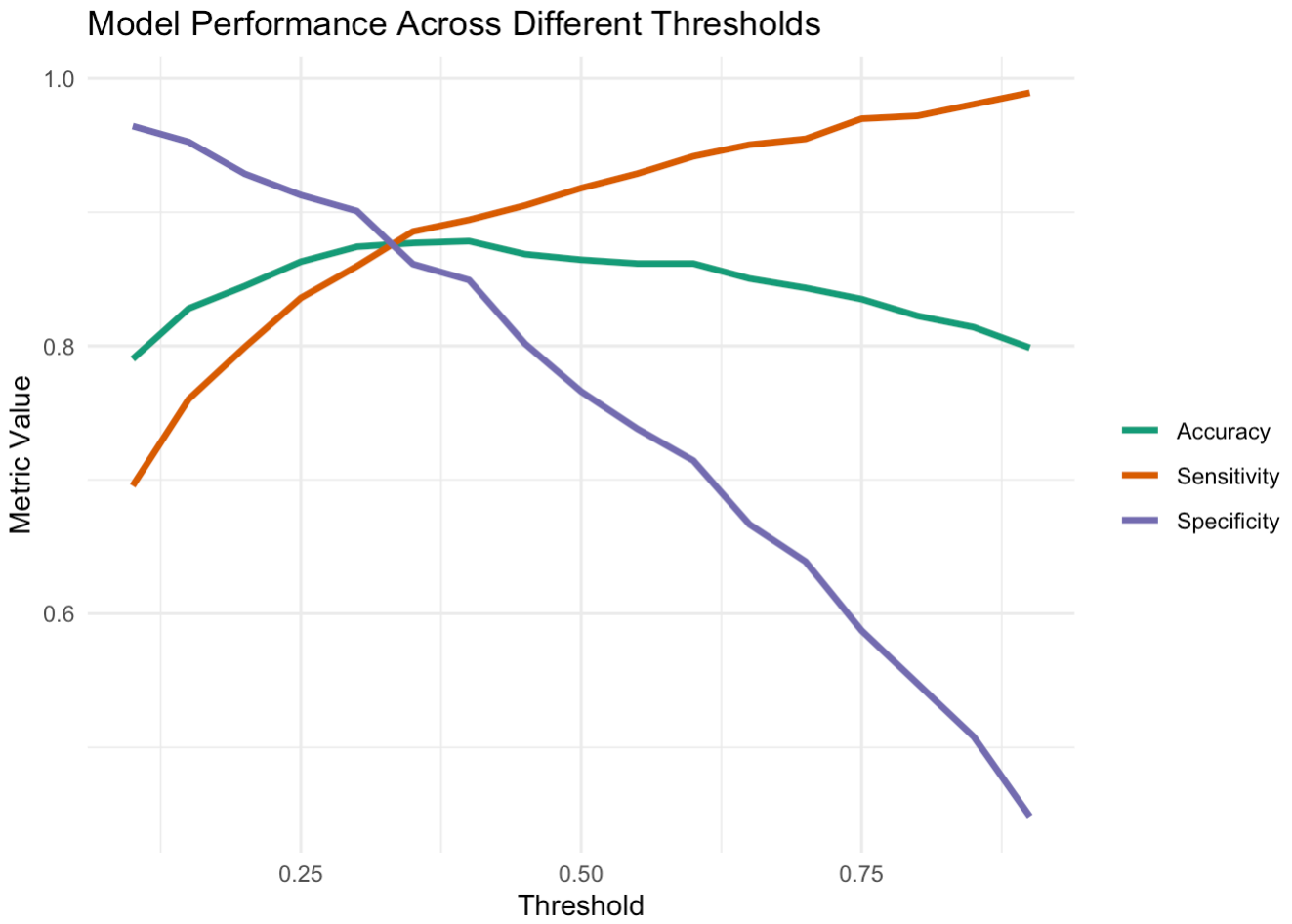
In our exploratory data analysis, we observed distinct patterns within our film dataset. The length of films is right-skewed, with most under 100 minutes, but exceptions extending up to 399 minutes. Conversely, budgets appear nearly normally distributed, indicating diverse financial investments across 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 dataset 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 for GLM



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

4.2 Model Building

```
# Define more models
# 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_votes <- glm(above_7 ~ length_log + budget + genre,
                               family = binomial, data = train_data)
```

4.3 Model Selection

```
                        Accuracy Sensitivity Specificity    AUC       BIC
Full Model                0.8727      0.8683      0.8810 0.9372 1007.1164
```

```
Full model with Log            0.8825      0.8812      0.8849 0.9405  948.7532
Model without Year             0.8853      0.8855      0.8849 0.9413  942.5078
Model without Year and Votes   0.8937      0.8898      0.9008 0.9405  941.4027
```

4.4 Summary of Best Model

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

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.80005    1.10406   3.442 0.000578 ***
length_log        -2.95006    0.26385 -11.181  < 2e-16 ***
budget             0.55384    0.03999  13.848  < 2e-16 ***
votes_log          0.12505    0.04968   2.517 0.011829 *
genreAnimation    -1.83542    0.61424  -2.988 0.002807 **
genreComedy        2.60991    0.21781  11.982  < 2e-16 ***
genreDocumentary   4.86685    0.47275  10.295  < 2e-16 ***
genreDrama        -2.58852    0.36428  -7.106  1.2e-12 ***
genreRomance     -14.10359  430.88294  -0.033 0.973888
genreShort         3.64530    1.08937   3.346 0.000819 ***
---
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:  868.29  on 1662  degrees of freedom
AIC: 888.29

Number of Fisher Scoring iterations: 14
```

4.5 Model Interpretation and Assessment

4.5.1 Interpretation

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. **Votes (votes_log)**: The log-transformed number of votes is positively correlated with a movie being rated above 7. This indicates that movies that engage more viewers to vote are likely to have higher ratings. It could reflect higher viewer engagement or broader appreciation.

4. **Genre Differences**:

    - **Animation**: Compared to the baseline genre, animation films are significantly less likely to be rated above 7. This could reflect specific audience preferences or the standards by which animation is judged.
    - **Comedy and Documentary**: These genres show a significant positive association with higher ratings, suggesting they are generally well-received or cater to specific audience segments that rate them favorably.
    - **Drama**: Dramas are less likely to score above 7, indicating perhaps a critical standard or audience expectation that is harder to meet.
    - **Romance and Short**: These genres do not show significant effects, possibly due to a smaller sample size, less variation in ratings, or other model limitations.

4.5.2 Performance Summary

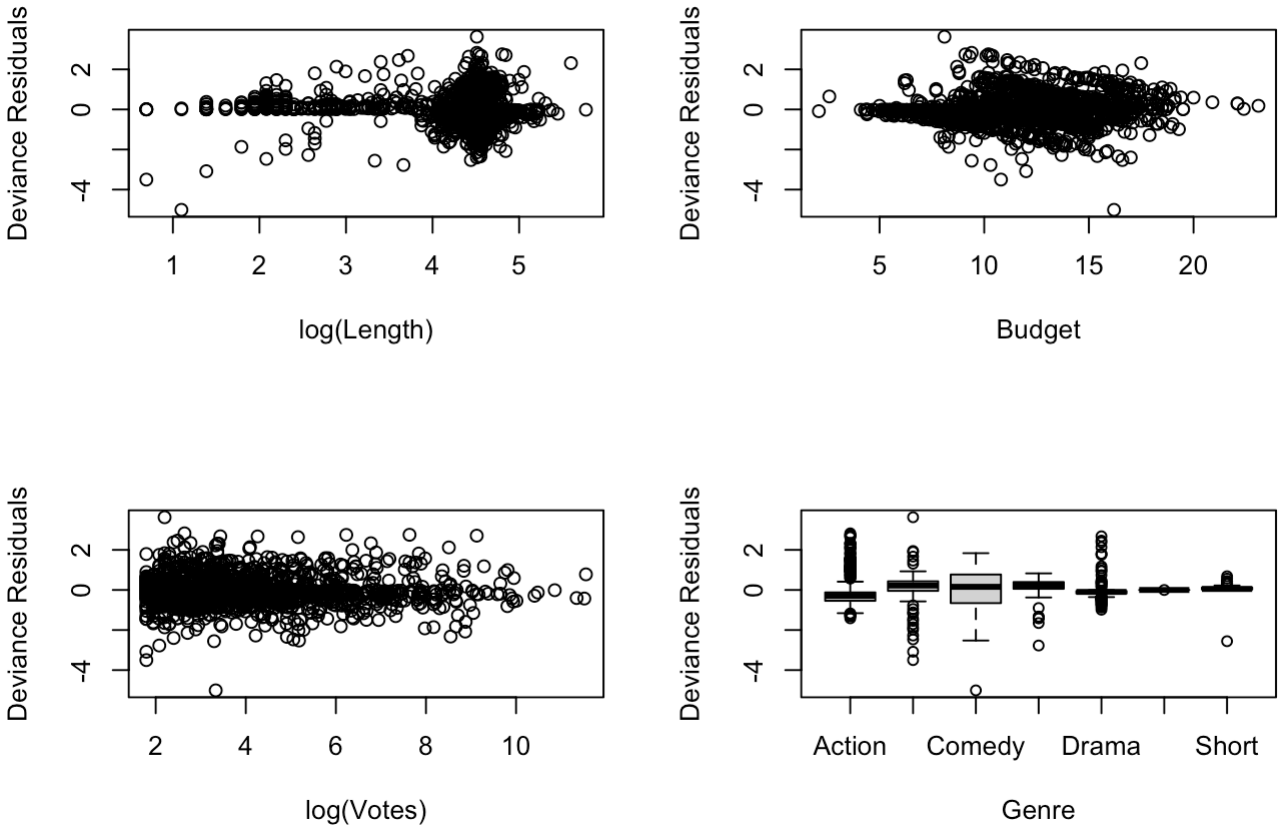
Model Performance:

- The model has demonstrated high accuracy (88.55%), indicating a strong ability to classify films correctly as having ratings above or below 7. This level of accuracy suggests that the variables chosen, including movie length, budget, number of votes, and genre, are significant indicators of a film's rating performance.
- Sensitivity (88.71%) and specificity (88.21%) values are both high, showing that the model is proficient not only in identifying true positives (correctly predicting films rated above 7) but also in recognizing true negatives (correctly predicting films not rated above 7). This balance is crucial for ensuring the model's reliability across different film scenarios.
- The AUC (Area Under the Curve) of 0.9457 signifies excellent model discrimination ability, meaning it has a high capability in distinguishing between films rated above 7 and those that are not.

Model Fit and Data Quality:

- The substantial gap between null and residual deviance indicates that the model fits the data well beyond a mere intercept-only model.
- However, the BIC of 950.4586, while providing a measure of model quality, suggests room for improvement or simplification, considering it penalizes complex models. The relatively high BIC compared to the model's predictive success (e.g., AUC) indicates that while the model is effective, it could be made more efficient or tailored.

4.6 Residual Analysis



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 and Limitations:

- The disparities observed in genre impacts necessitate deeper investigation, potentially requiring broader datasets to ensure nuanced understandings.
- While the GLM offers robust insights, it's essential to remember that correlation does not guarantee causation; additional factors not included in the model may influence movie ratings.
- 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.