Group_06

Analysis of IMDB data set

1 Introduction

The study aims to investigate the relationship between various film attributes and IMDB ratings, drawing data from the IMDB film database allocated. The data set comprises of the factors such as film ID, release year, duration, budget, votes, genre, and IMDB rating. The research question focuses on examining the factors that impact IMDB ratings, particularly whether specific film properties contribute to ratings greater than seven. A Generalized Linear Model (GLM) analysis is conducted to derive the relationships between these properties and IMDB ratings.

2 Data Wrangling Methods

Before we begin the analysis of our data, let's transform the data using various tools. The process below describes the detailed data wrangling techniques that are used to get the desired data set. After having a glimpse of the data set, the 'genre' column is converted to a type factor type.

A check for missing values is conducted and it is found that 103 observations are missing from the column 'length'. Missing values are imputed with the median since median is a robust measure, less impacted by outliers as much as mean. The function *median( )* reveals the median to be 90 minutes. However, it is observed in Table 1 that the median lengths vary across the different genres. With this information, the missing lengths of films are replaced by median length of the respective genre.

Table 1: Median length by genre.

genre	median_length
Action	90.0
Animation	7.0
Comedy	91.0
Documentary	73.0
Drama	96.0
Romance	92.5
Short	13.5

As per the research question, a new column 'high_rating' containing binary variables coressponding to 'rating' values is created. This column takes a value of 1 for IMDB ratings greater than or equal to seven and 0 for IMDB ratings less than seven. Additionally, another categorical variable 'rate' conveying the same is also added.

3 Exploratory Data Analysis

3.1 view the data

Check on the size of a data set

```
[1] 1937    9
```

Sample size is 1937. And it have 9 variables,7 of which are in the original data.

Let's have a look at the first five rows of the data frame.

Table 2: Glimpse of the first five rows in the IMDB data set.

film_id	year	length	budget	votes	genre	rating	high_rating	rate
31804	2002	18	9.6	15	Drama	8.0	1	Rating greater than 7
25453	2000	98	13.8	23	Action	3.3	0	Rating less than 7
5479	1989	81	11.5	57	Documentary	7.9	1	Rating greater than 7
44235	1995	100	7.5	32	Action	3.4	0	Rating less than 7
14580	2003	80	10.8	30	Action	2.6	0	Rating less than 7

The variables in Table 2

- **film.id** : The unique identifier for the film
- **year** : Year of release of the film in cinemas
- **length** : Duration (in minutes)
- **budget** : Budget for the films production (in $1000000s)
- **votes** : Number of positive votes received by viewers
- **genre** : Genre of the film
- **rating** : IMDB rating from 0 to10
- **high_rating** : 1 for IMDB ratings greater than or equal to seven and 0 for IMDB ratings less than 7
- **rate** : 'Rating greater than 7' got high_rating = 1 and 'Rating less than 7' for high_rating = 0

3.2 Summary Statistics

Since variables year, length,budget,votes,rating are continuous, we need get their summary contains mean,median,standard deviation,minimum maximum.

Table 3: Summary statistics on the IMDB data by variables.

Variables	Mean	Median	Std. Dev	Minimum	Maximum	Interquartile Range	Sample Size
year	1,976.21	1,982.00	23.44	1,896.00	2,005.00	39.00	1,937.00
length	82.88	90.00	35.62	1.00	316.00	24.00	1,937.00
budget	12.03	12.00	2.92	3.20	21.20	3.90	1,937.00
votes	590.47	31.00	3,894.33	5.00	103,854.00	104.00	1,937.00

The Table 3 shows that the summary for the columns year, length, budget and votes.

- The number of films in sample is 1937.
- For the variable year, the years of the films ranged from 1896 to 2005.
- For the variable length, the films runs from 1 minute to 316 minutes.The average length of a movie is 83.22 minutes.
- For variable budget, the budget of films is from 3.2 ($1000000s) to 21.2 ($1000000s).The median budget of a movie is 12($1000000s).
- For variable votes, the votes of films is from 5 to 103,854.The range of variation is very large, and the IQR is relatively large. The data is not stable.

3.3 Correlation

Check correlations (as scatterplots), distribution and print correlation coefficient.

Figure 1 shows rating and budget show a significant linear positive correlation. And rating is significant linear negative correlation with length. Between rating and year,votes have weak correlation.

3.4 Visualization

3.4.1 Histograms for continuous variable(rating,year,budget,length,votes)

Histograms to understand the data structures of different variables

The variable votes has a large data difference and deviation, so we can log transformation for votes.

3.4.2 Visualise the distributions of categorical variable genre

3.5 The relationship between rating and explanatory variables

Scatterplots to understand the relationship between rating and four variables(year,length,budget,log_votes)

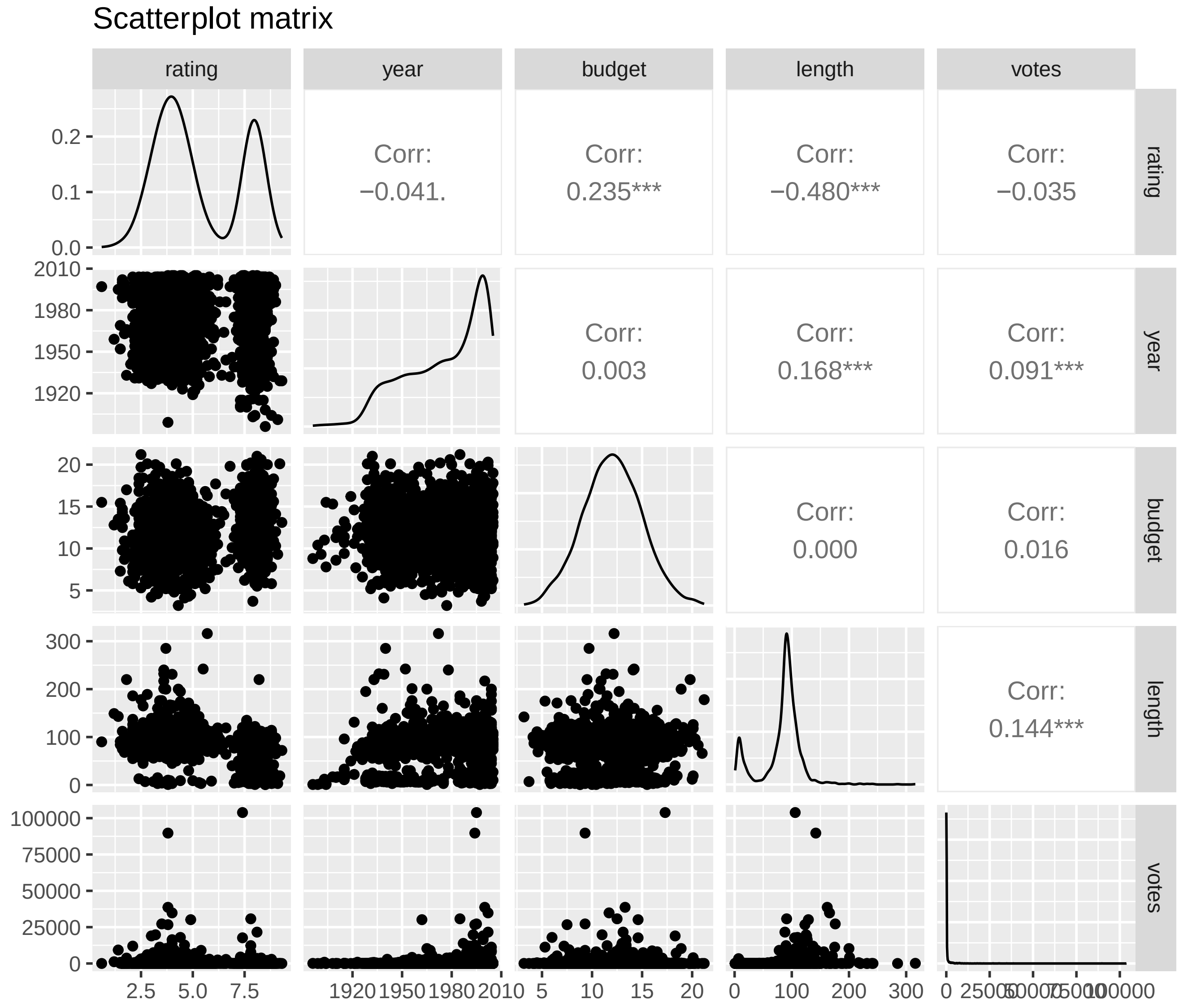

Figure 1: Scatterplot matrix between rating and explanatory variables.

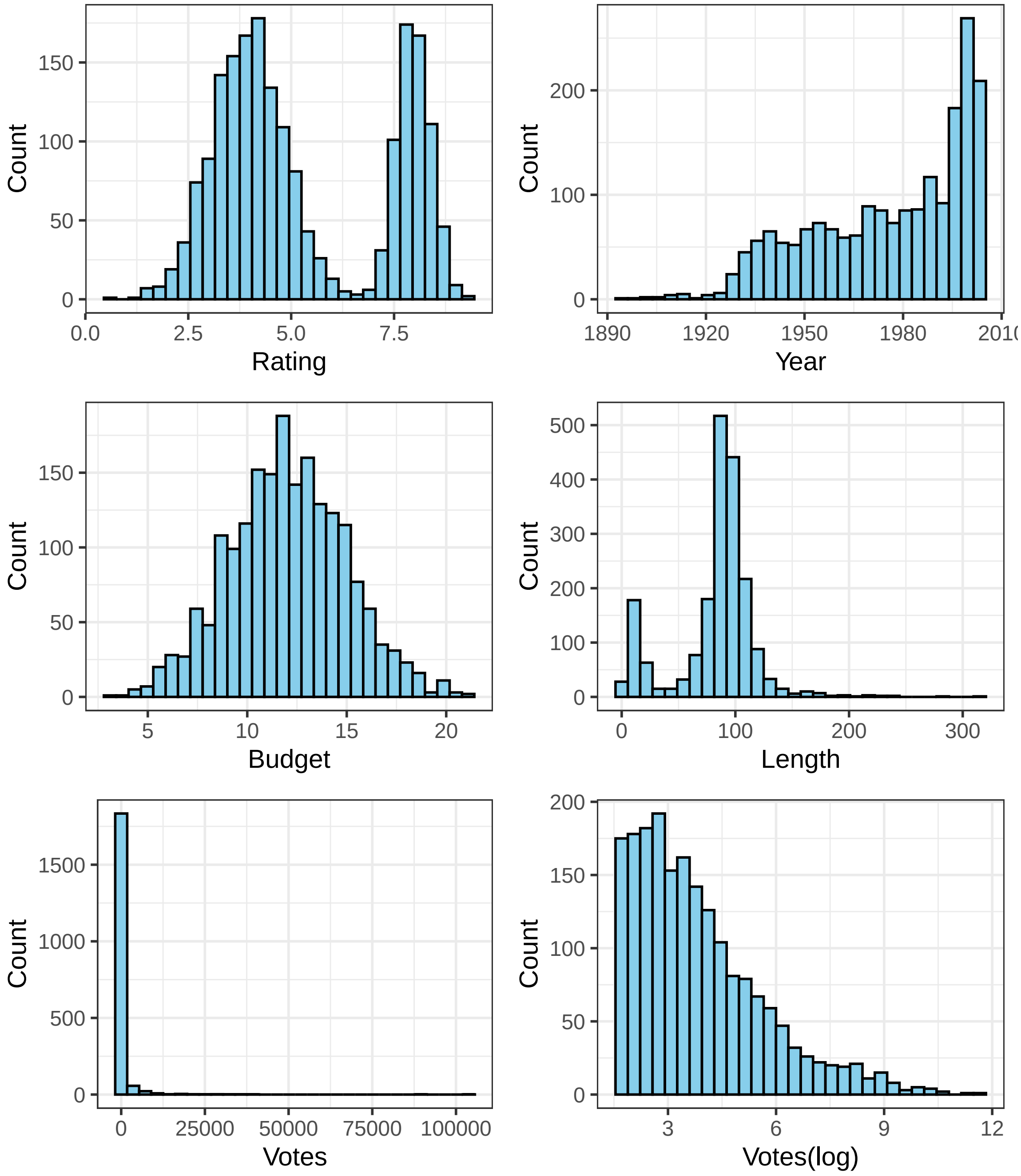

Figure 2: Histograms of statistical distribution for varibles

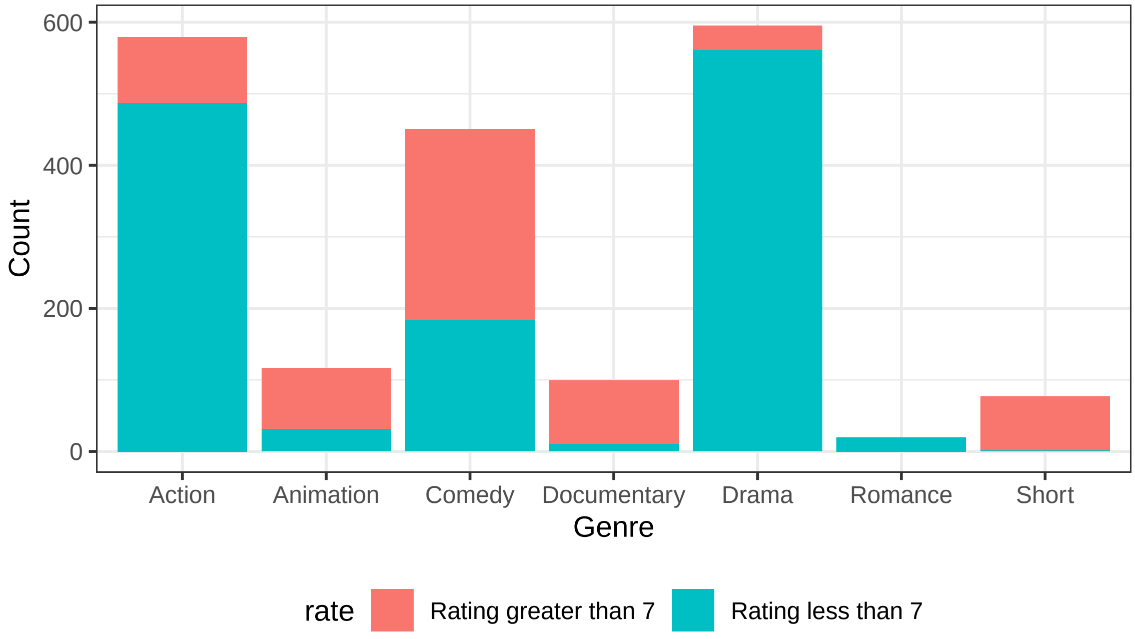

Figure 3: Stacked barplot of statistical distribution for ratings by genre.

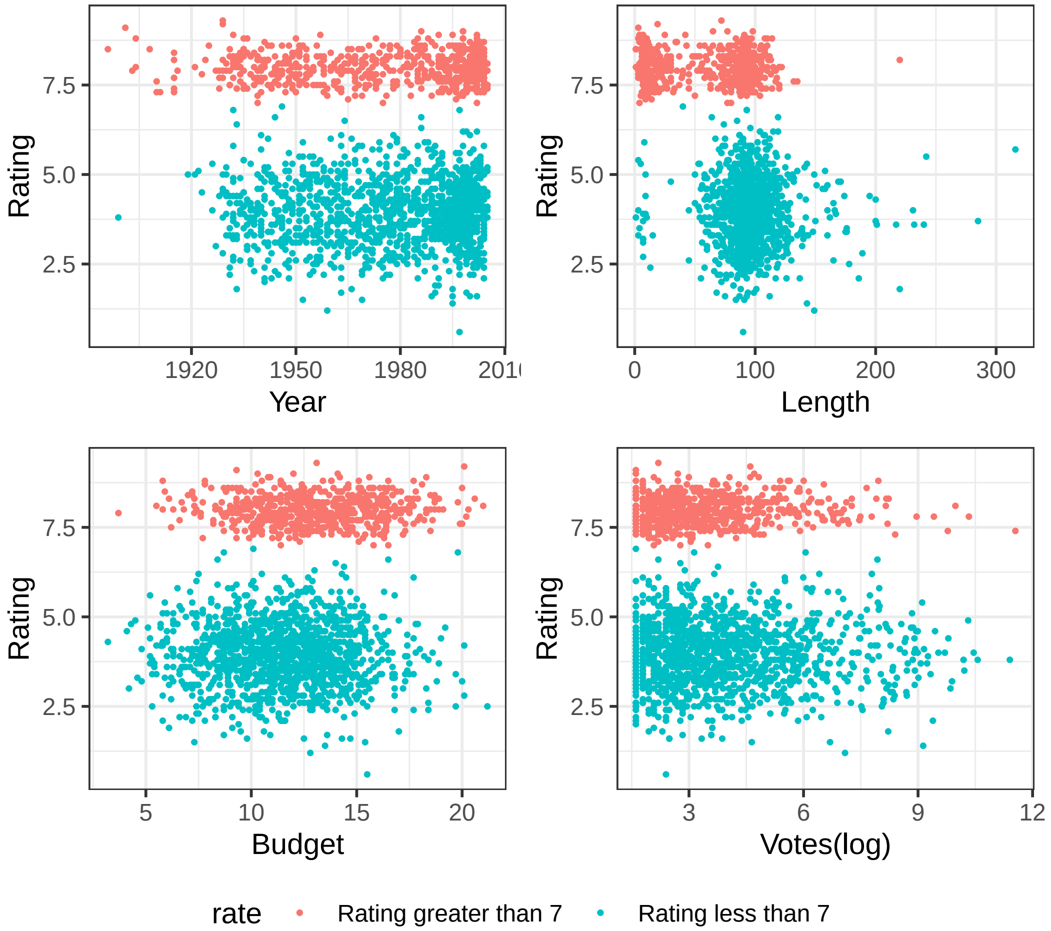

Figure 4: Scatterplots between rating and four explanatory variables.

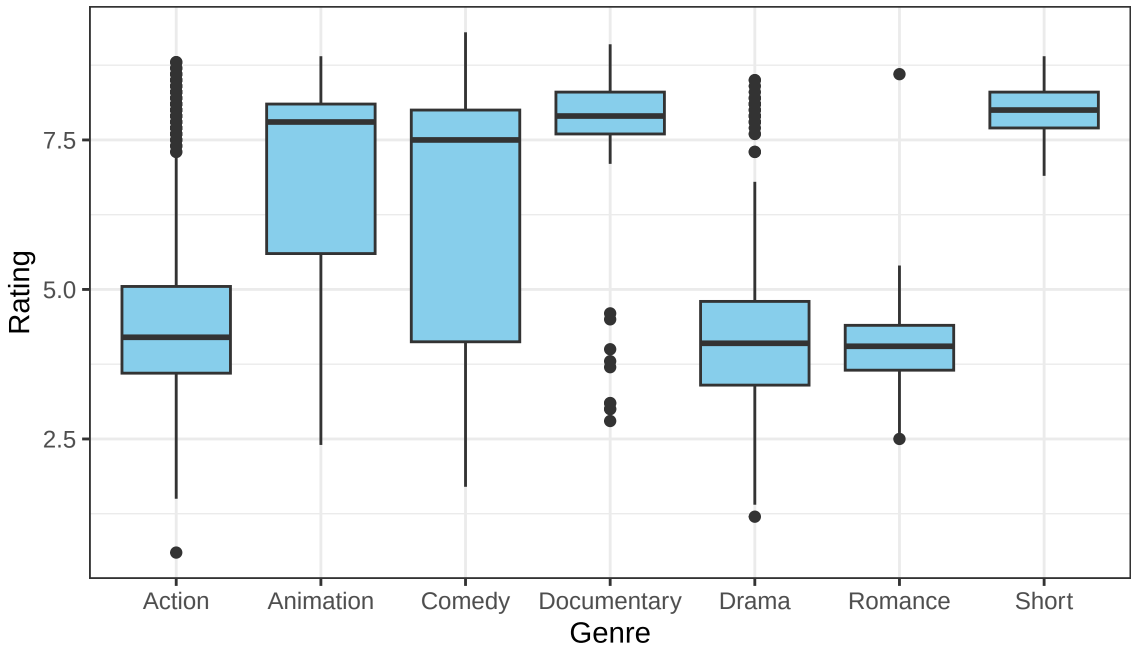

Figure 5: Boxplot of ratings by genre.

3.5.1 Boxplot to understand the relationship between rating and Categorical variable genre

3.6 The relationship response and explanatory variable

3.6.1 variable1:Length

Figure 6 shows that the median film length of high rating films is less than that of low rating films. It can Interquartile Range of low rating is smaller than that of high rating, which shows that the data about low rating is more concentrated.

Table 4: Summary statistics on length by rating.

rate	Mean	Median	Std. Dev	Minimum	Maximum	Interquartile Range	Sample Size
Rating greater than 7	57.05	73.00	39.49	1.00	220.00	78.25	644.00
Rating less than 7	95.74	94.00	25.03	1.00	316.00	18.00	1, 293.00

The Table 4 shows that the size of the film with low rating is more than twice as many as that with high rating. The mean length film with high rating(57.95) is lower than that with low rating(95.80).On the whole, highly rated films have a low parity of length phrases. But a low rated films is more stable.

3.6.2 variable2:Budget

Figure 7 shows that the median budget film of high rating films is higher than that of low rating films.

Table 5: Summary statistics on budget by rating.

rate	Mean	Median	Std. Dev	Minimum	Maximum	Interquartile Range	Sample Size
Rating greater than 7	13.09	13.00	2.84	3.70	21.00	4.10	644.00
Rating less than 7	11.51	11.50	2.82	3.20	21.20	3.90	1, 293.00

The Table 5 shows that the mean budget film with high rating(13.09 in $1000000s) is higher than that with low rating(11.51 in $1000000s).And the median budget film with high rating(13.00 in $1000000s)

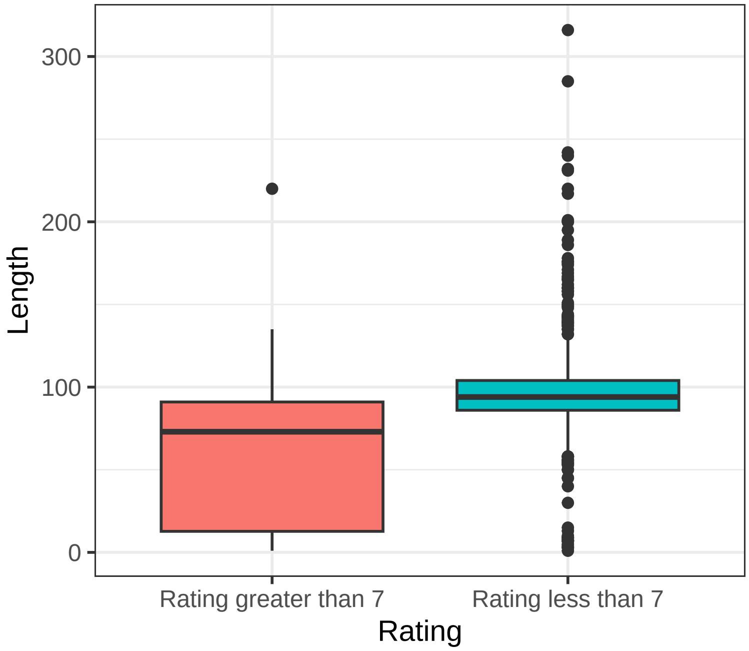

Figure 6: Boxplot of length by rating.

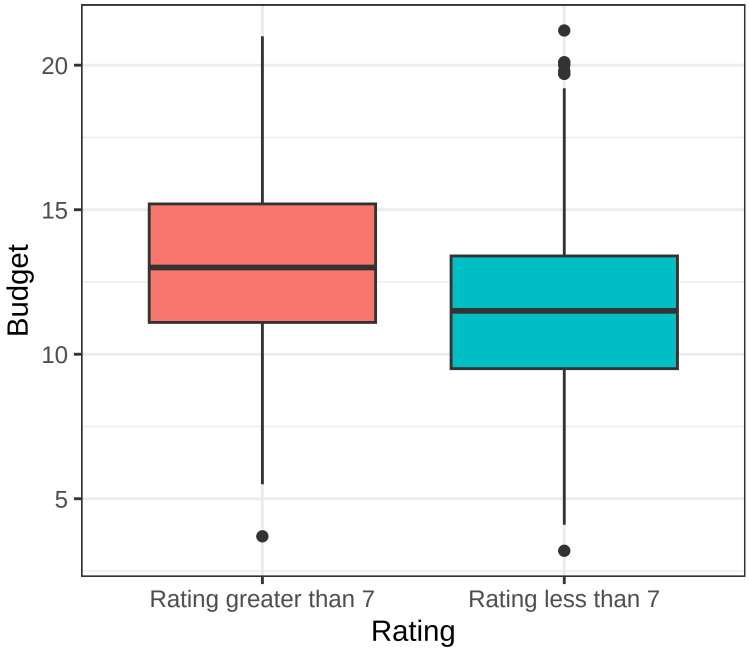

Figure 7: Boxplot of budget by rating.

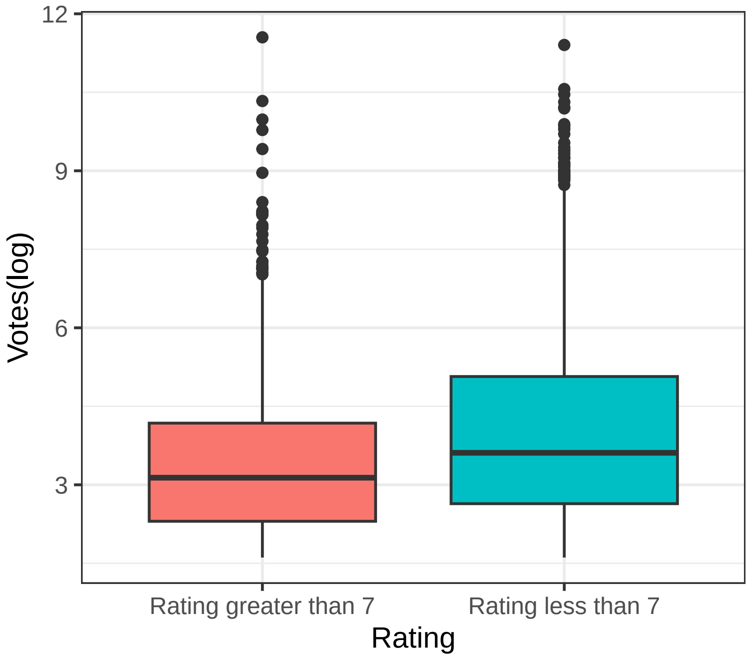

Figure 8: Boxplot of log_votes by rating.

3.6.4 variable4:genre

The ratio of ratings above 7 to ratings below 7 and sample sizes for each type

```
      genre Rating greater than 7 Rating less than 7 genre_sum_count
     Action            15.9%  (92)        84.1% (487)             579
  Animation            73.5%  (86)        26.5%  (31)             117
     Comedy            59.1% (266)        40.9% (184)             450
Documentary            89.9%  (89)        10.1%  (10)              99
      Drama             5.7%  (34)        94.3% (561)             595
    Romance             5.0%   (1)        95.0%  (19)              20
      Short            98.7%  (76)         1.3%   (1)              77
```

We can see the size sample about Romance films is only 20,which is too small to fit model.And we also can predict the positive effect of action,Drama on rate score.

Figure 9 shows that the number of low rating films is bigger than the number of high rating films in the genre films about Action,Drama,Romance.Besides,the number of low rating films is smaller than the number of high rating films in the genre films about Animation,Documentary,Short. In genre Comedy, the number of films with high ratings is about the same as the number of films with low scores.

4 Formal Data Analysis

4.1 Generalised Linear Model fit

```
Call:
glm(formula = high_rating ~ year + length + budget + log_votes +
```

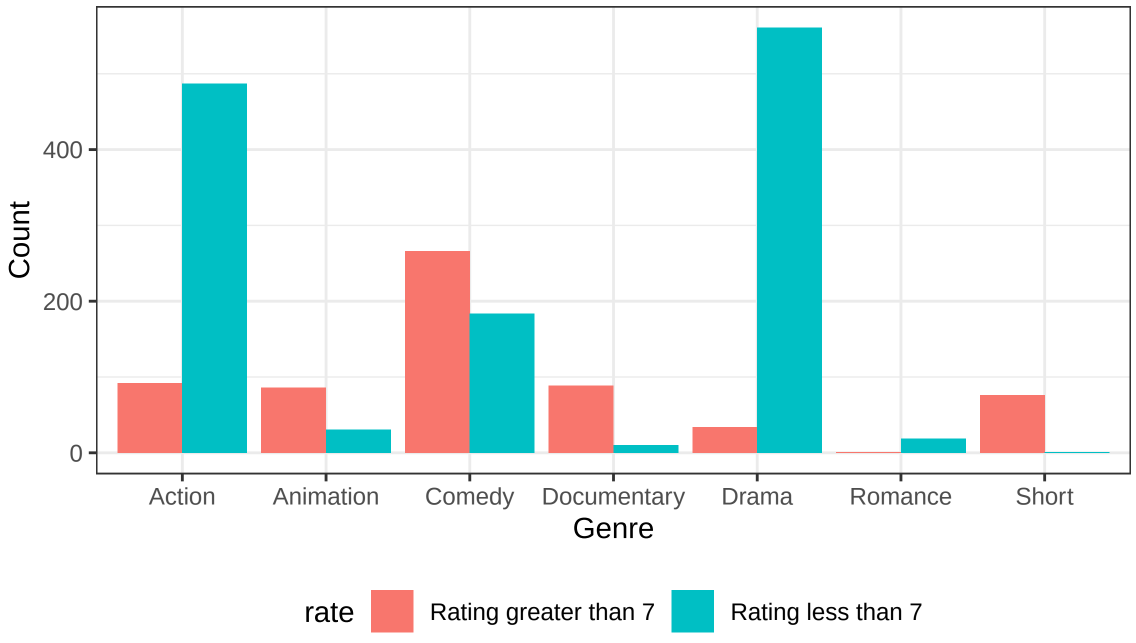

Figure 9: Dodged barplot of genre by reating.

```
    genre, family = binomial(link = "logit"), data = imdb_data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -20.999835   7.309525  -2.873  0.00407 **
year               0.009098   0.003721   2.445  0.01448 *
length            -0.063815   0.004678 -13.642  < 2e-16 ***
budget             0.512148   0.035449  14.448  < 2e-16 ***
log_votes          0.003142   0.050031   0.063  0.94992
genreAnimation    -0.547691   0.423392  -1.294  0.19581
genreComedy        3.062131   0.215119  14.235  < 2e-16 ***
genreDocumentary   5.054809   0.468057  10.800  < 2e-16 ***
genreDrama        -1.612361   0.275068  -5.862 4.58e-09 ***
genreRomance      -2.161840   1.677264  -1.289  0.19743
genreShort         3.547377   1.079457   3.286  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2463.5  on 1936  degrees of freedom
Residual deviance: 1016.8  on 1926  degrees of freedom
AIC: 1038.8

Number of Fisher Scoring iterations: 7
```

Intercept: The intercept term of the model, the log odds when all explanatory variables are zero. year: the log odds increase by about 0.009098 for each additional year, meaning that more recent films have a slightly higher chance of being rated higher than 7 compared to older films. length: For every additional minute, the log odds decrease by about 0.063815, meaning that longer films are less likely to be rated higher than 7. budget: for every one million budget increase, the log odds increase by about 0.512148, indicating that films with higher budgets are more likely to rate higher than 7. log_votes: The natural logarithm of the number of votes increases by about 0.003142 per unit, but this is not significant. genreAnimation: Animated films are less likely to receive high ratings than action films, but the difference is not statistically significant. genreComedy: Comedy films are significantly more likely to receive high ratings compared to action films. genreDocumentary: Documentaries are significantly more likely to be rated highly than action films. genreDrama: Dramas are significantly less likely to be rated highly than action films, and the difference is statistically significant. genreRomance: Romance films are significantly less likely to be rated highly than action films, but the difference is not statistically significant. genreShort: Short films are significantly more likely to receive high ratings compared to action films. ## Model Selection

```

Call:
glm(formula = high_rating ~ year + length + budget + genre, family = binomial(link = "logit"),
    data = imdb_data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -21.081669   7.192087  -2.931  0.00338 ** 
year               0.009143   0.003652   2.503  0.01230 *  
length            -0.063754   0.004573 -13.943  < 2e-16 ***
budget             0.512107   0.035442  14.449  < 2e-16 ***
genreAnimation    -0.543849   0.418900  -1.298  0.19419    
genreComedy        3.064416   0.212064  14.450  < 2e-16 ***
genreDocumentary   5.052696   0.466841  10.823  < 2e-16 ***
genreDrama        -1.612372   0.275060  -5.862 4.58e-09 ***
genreRomance      -2.158153   1.675364  -1.288  0.19769    
genreShort         3.547375   1.079435   3.286  0.00102 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2463.5  on 1936  degrees of freedom
Residual deviance: 1016.8  on 1927  degrees of freedom
AIC: 1036.8

Number of Fisher Scoring iterations: 7
```

5 Conclusions