

Analysis of IMDB data set

Group_06

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 ?? 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 corresponding 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

The data set has 1937 rows and 9 columns, 7 of which are from 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 ?? are defined as:

- **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 to 10
- **high_rating** : 1 for IMDB ratings greater than or equal to seven and 0 for 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

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

| Variables | Mean | Median | Std. Dev | Min | Max | IQR | 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 |
| rating | 5.29 | 4.60 | 2.05 | 0.60 | 9.30 | 4.00 | 1,937.00 |

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

- For the variable year, the years of the films ranges from 1896 to 2005.
- For the variable length, the films runs from 1 minute to 316 minutes. The median for length of films is 90 minutes.
- For variable budget, the budget of films is from 3.2 (\$1000000s) to 21.2 (\$1000000s). The median budget of a film is 12(\$1000000s).
- For variable votes, the votes of films ranges from 5 to 103,854 which suggests large variation. It can be observed the IQR is relatively large as well.

3.3 Correlation

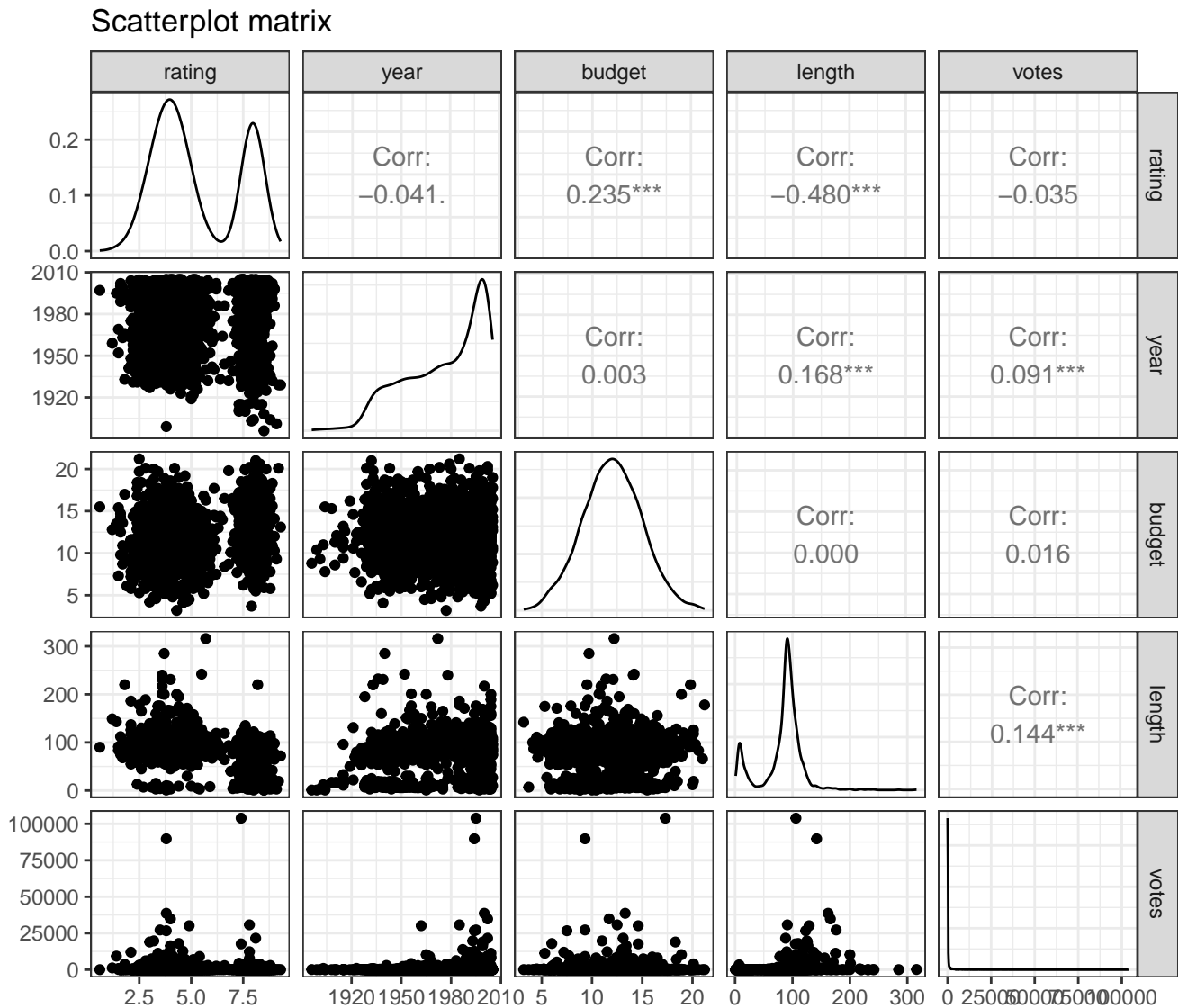


Figure 1: Scatterplot matrix between rating and explanatory variables.

The Figure ?? shows weak correlation between the variables. 'length' shows not so strong negative correlation with 'rating'.

3.4 Visualization

3.4.1 Histograms

The Figure ?? shows that the data structures follow exponential distributions. The variable 'votes' displays skewness due to large difference in values of maximum and minimum values. To reduce this skewness and facilitate more robust analysis, a logarithmic transformation is used.

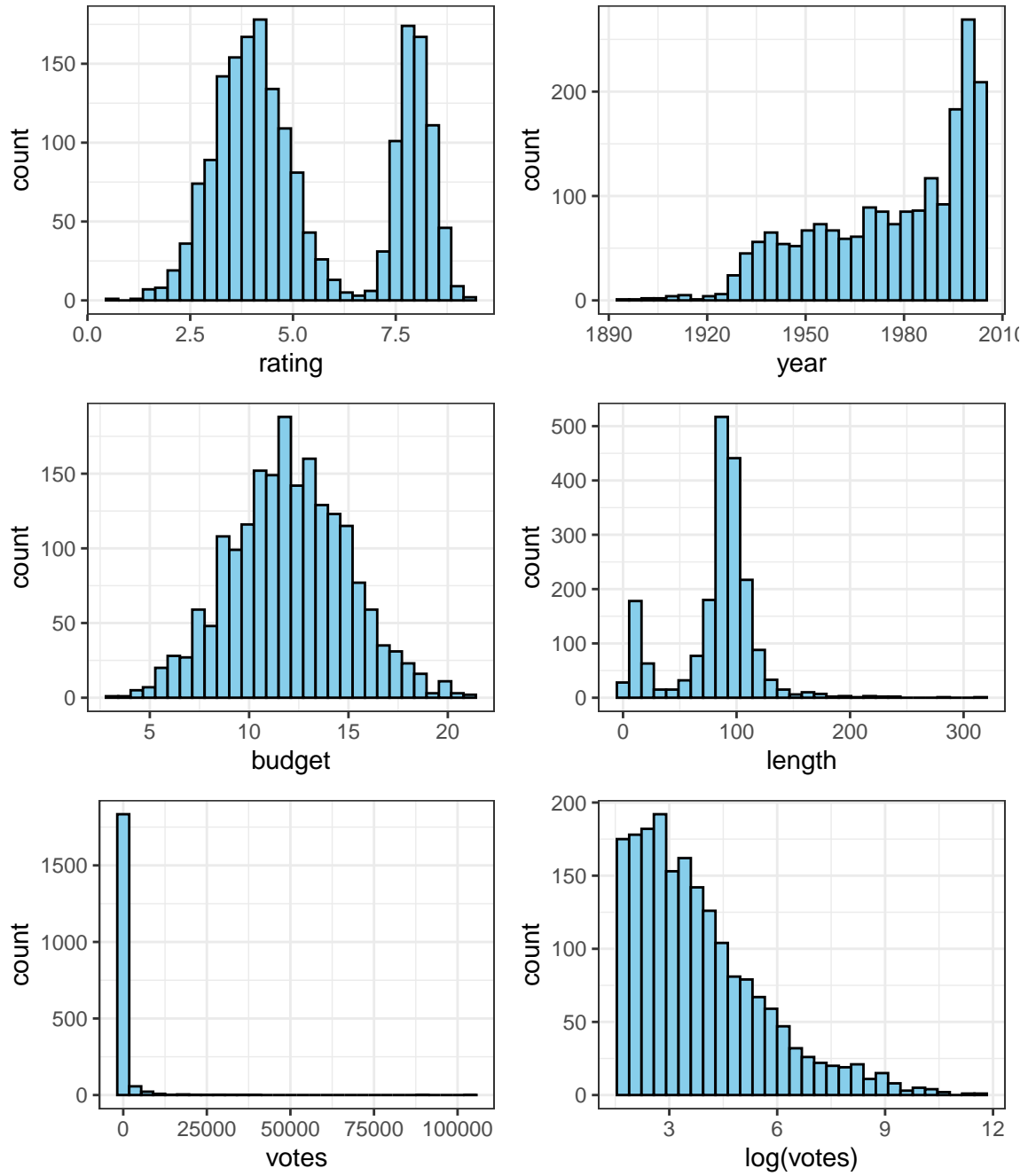


Figure 2: Histograms of statistical distribution for variables

3.4.2 Scatter plot for rating vs explanatory variables

The Figure ?? suggests that there seems to be no linear relationship between the response variable and the explanatory variables which justifies the weak correlation observed earlier.

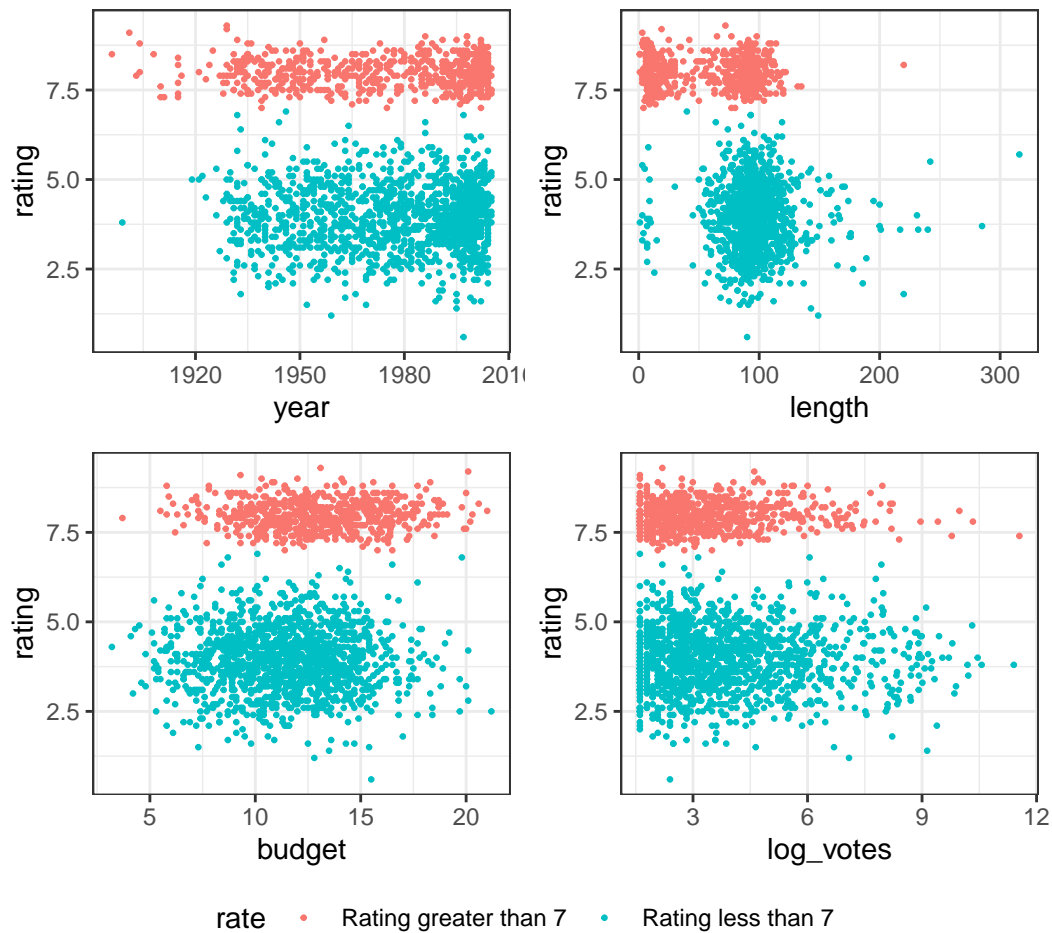


Figure 3: Scatterplots between rating and four explanatory variables.

3.4.3 Boxplot for genre

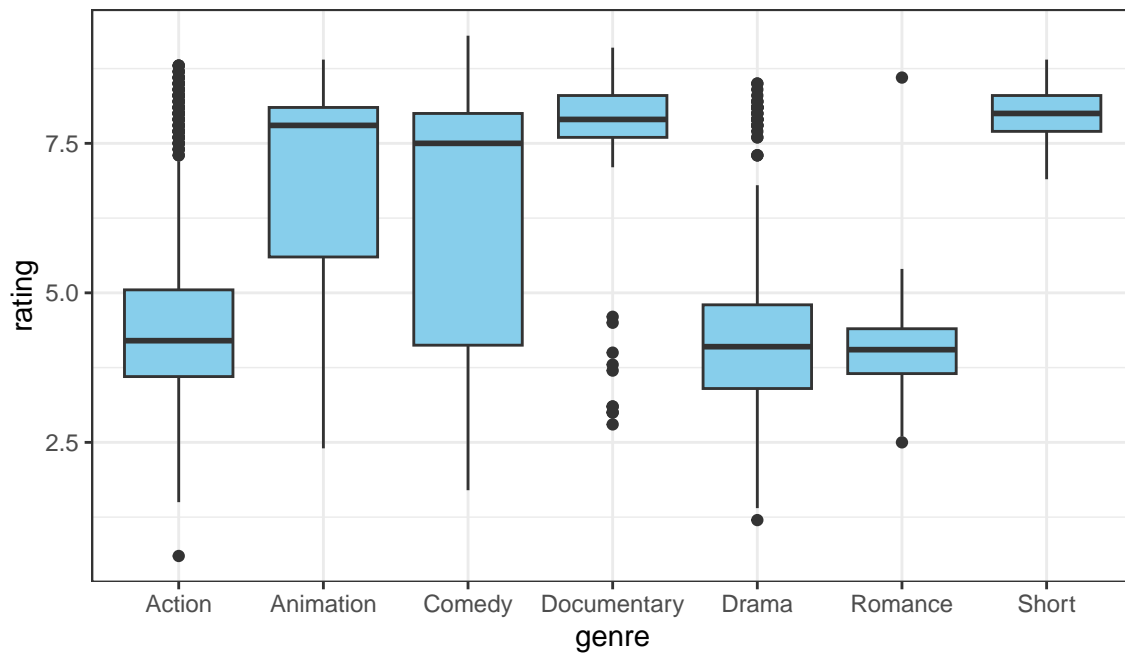


Figure 4: Boxplot of ratings by genre.

The Figure ?? distribution of rating genre wise. Outliers for ratings can be seen for the genre's Action, Documentary, Drama and Romance.

3.5 The relationship response and explanatory variable

3.5.1 Variable 1: Length

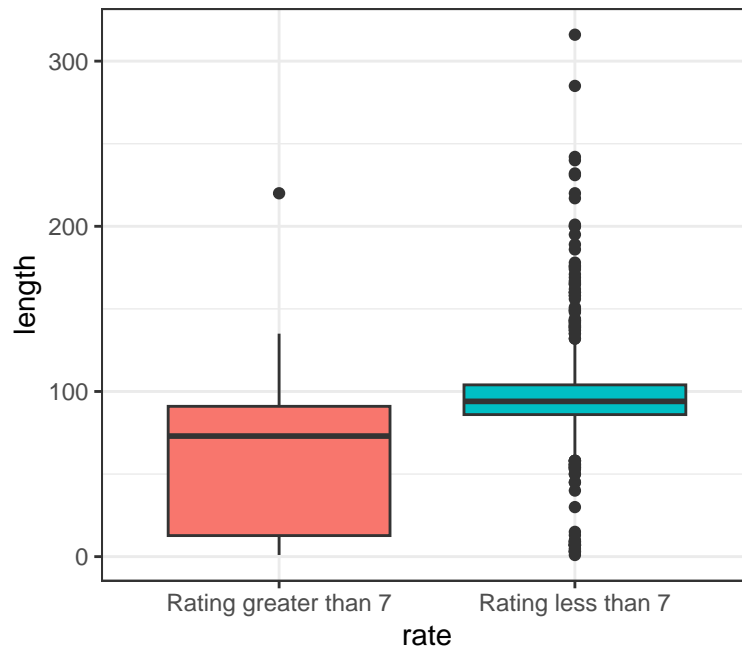


Figure 5: Boxplot of length by rating.

The Figure ?? shows that the median film length of films with ‘Rating greater than 7’ is less than that of ‘Rating less than 7’ films. It can be observed IQR of ‘Rating less than 7’ is smaller but has many outliers.

Table 4: Summary statistics on length by rating.

| rate | Mean | Median | St.Dev | Min | Max | IQR | 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 ?? The median length film with ‘Rating greater than 7’ is (73 minutes) lower than that with ‘Rating less than 7’ (95.74 minutes).

3.5.2 Variable 2 : budget

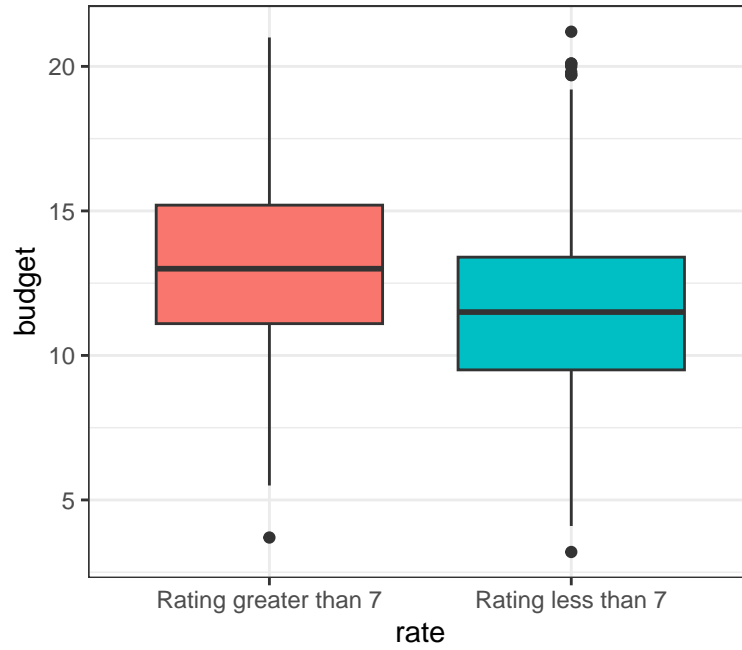


Figure 6: Boxplot of budget by rating.

The Figure ?? shows that the median budget film of ‘Rating greater than 7’ is slightly higher than that of ‘Rating less than 7’ films. There are 9 outliers.

Table 5: Summary statistics on budget by rating.

| rate | Mean | Median | Std. Dev | Min | Max | IQR | 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 ?? shows that the mean and median for ‘Rating greater than 7’ is almost equal. Similarly, it can be observed for and ‘Rating less than 7’ as well. This suggests a normal distribution. The variability is also equivalent for the 2 categories.

3.5.3 Variable 3 : log_votes

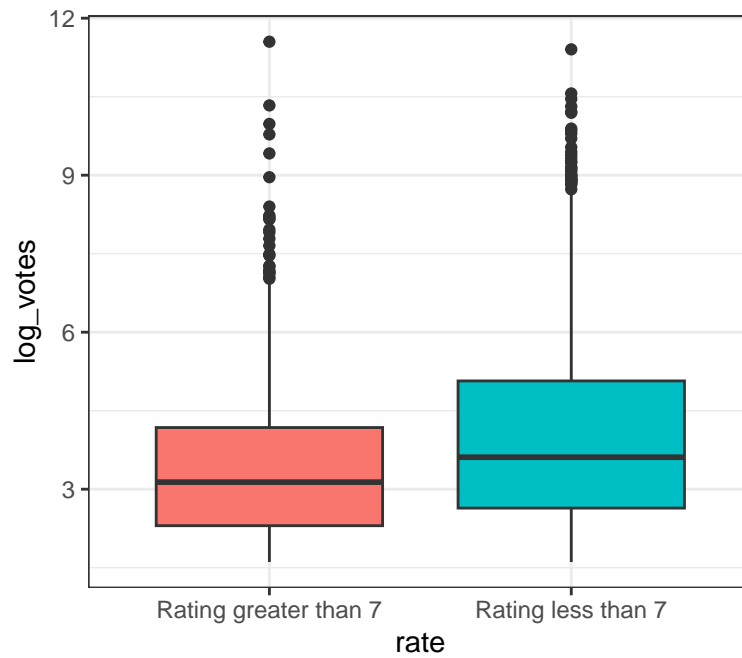


Figure 7: Boxplot of log_votes by rating.

The Figure ?? shows that the median log_votes film of ‘Rating greater than 7’ films is lower than that of ‘Rating less than 7’ films.

Table 6: Summary statistics of votes(log) by rating.

| rate | Mean | Median | Std. Dev | Min | Max | IQR | Sample Size |
|-----------------------|------|--------|----------|------|-------|------|-------------|
| Rating greater than 7 | 3.47 | 3.14 | 1.58 | 1.61 | 11.55 | 1.88 | 644.00 |
| Rating less than 7 | 4.03 | 3.61 | 1.86 | 1.61 | 11.40 | 2.43 | 1,293.00 |

The Table ?? shows that the mean and median for ‘Rating greater than 7’ is almost equal.

3.5.4 Variable 4 : genre

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

Table 7: Summary statistics of genre

| 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 |

It can be seen the size for the genre 'Romance' is only 20, which is comparatively small. We observe the following: - Animation, Documentary, Short - have more films with 'Rating greater than 7' - Action, Drama, Romance - have more films with 'Rating less than 7' - Comedy - films with 'Rating greater than 7' are moderately higher than 'Rating less than 7'

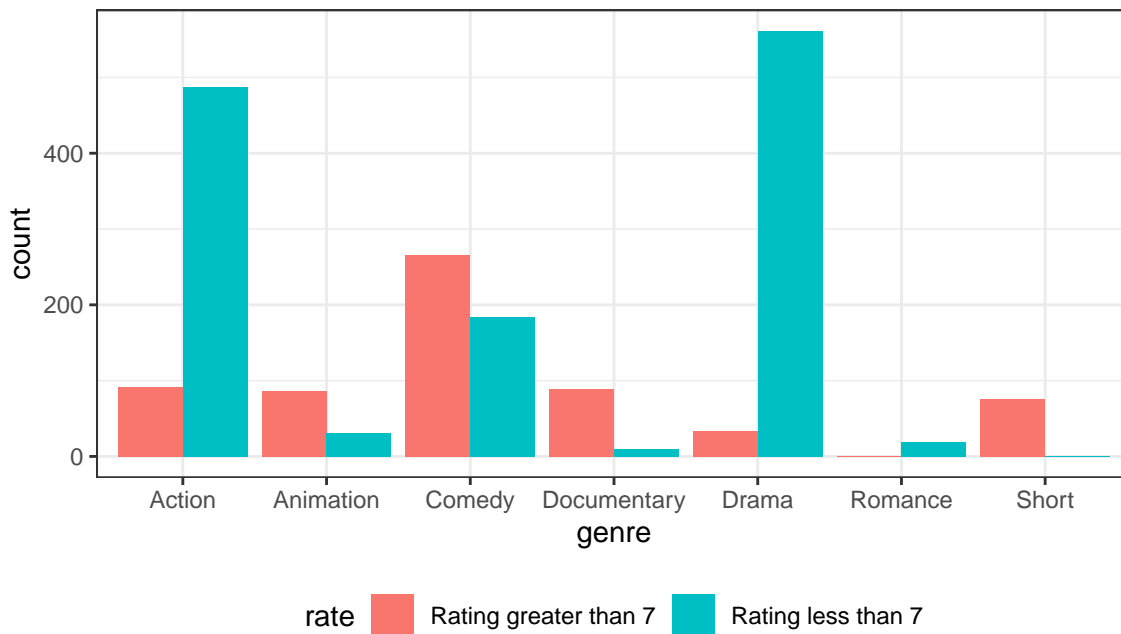


Figure 8: Dodged barplot of genre by rating.

The Figure ?? displays the information in the table Table ??