The Impact of film's features on IMDB Ratings: A logistic Regression approach

Yujie Tang, Jialu FU, Weiqing GUO, Bashiru Mukaila, Wanding Wang

2023-03-14

1 Introduction

The entertainment industry happens to be a competitive market and the film industry is a subset of it. The stakeholders in the film industry are interested in features that make their films successful. One key aspect of a film's success is its rating on platforms such as IMDB, which can influence audience perception and drive revenue.

The aim of this project is to investigate the relationship between the properties of films and their IMDB ratings. Specifically, we want to use a logistic regression model to know which properties of a film influence whether a film is rated by IMDB as greater than 7 or not, using variables such as year of release, length, budget, number of votes, and genre.

Section 2 speaks to the exploratory data analysis of IMDP ratings and explores the relationship between the rating and the properties of the films. Section 3 contains the results from the logistic regression model. While Section 4 gives the final remark to the research.

2 Exploratory Data Analysis

The data contains 2847 observations and 7 variables. The variables year, length, budget, votes and rating are numerical variables, and genre is a categorical variable. Table 2 shows the first 6 rows of the table. It is also worthy of note that the length of some of the films are not recorded.

Table 1: Highlight of the IMDB data

film_id	year	length	budget	votes	genre	rating
5993	1943	65	15.5	42	Action	7.6
37190	1961	87	12.3	6	Drama	6.0
43646	1987	79	16.4	161	Action	7.5
28476	1976	NA	12.2	5	Documentary	8.0
23975	1982	88	12.5	97	Action	3.5
...	...	...	...	...	...	...
50170	1936	NA	7.0	146	Drama	4.4

From the summary statistics in table 2, we can see that the year of release of the films are between 1898 to 2005. The minimum budget for film in the data set is 2.5 million while the maximum 22.3 million.

Table 2: Summary statistics of the numerical variables

year	length	budget	votes	rating
Min. :1898	Min. : 1.00	Min. : 2.50	Min. : 5	Min. :0.800
1st Qu.:1957	1st Qu.: 73.00	1st Qu.: 9.90	1st Qu.: 11	1st Qu.:3.700
Median :1982	Median : 90.00	Median :11.90	Median : 29	Median :4.600
Mean :1976	Mean : 82.22	Mean :11.85	Mean : 657	Mean :5.342
3rd Qu.:1997	3rd Qu.:101.00	3rd Qu.:13.70	3rd Qu.: 114	3rd Qu.:7.700
Max. :2005	Max. :480.00	Max. :22.30	Max. :149494	Max. :9.200
NA	NA's :131	NA	NA	NA

from the scatterplot in figure 1, we can see that there is a negative correlation between the length of the film and its rating. This implies that films with longer length tends to be rated low. There is a positive correlation between the budget of films production and it rating, however, the correlation is weak.

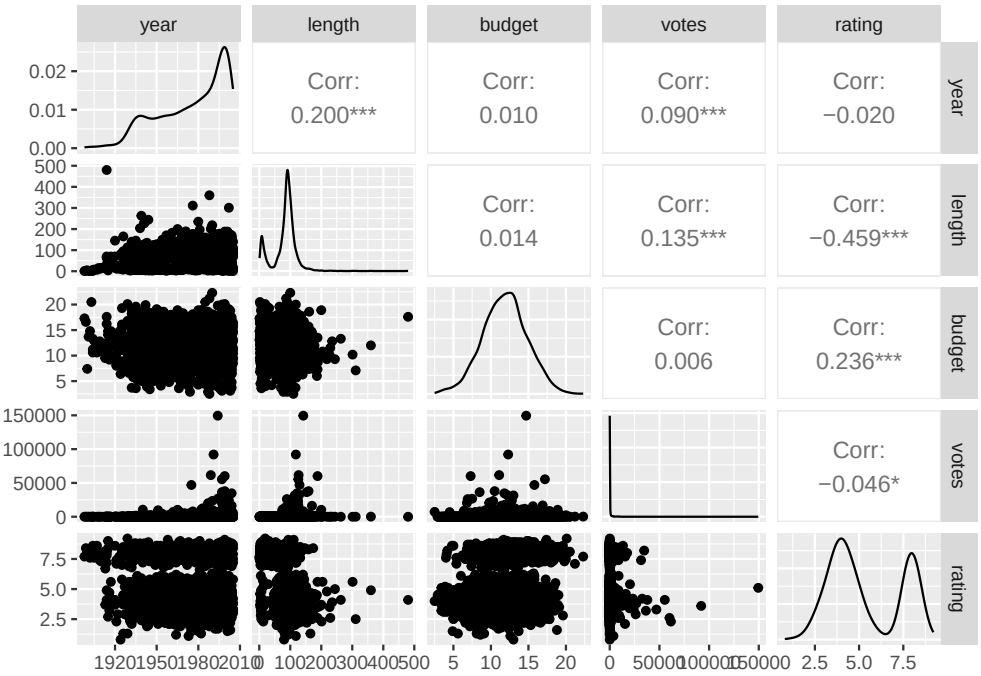

Figure 1: Scatterplot matrix of the numerical variables .

From figure 2, we can see that some genres tend to have higher ratings than others. For example, comedy tend to have higher ratings than romance films. While majority of the short films are rated higher than 7.

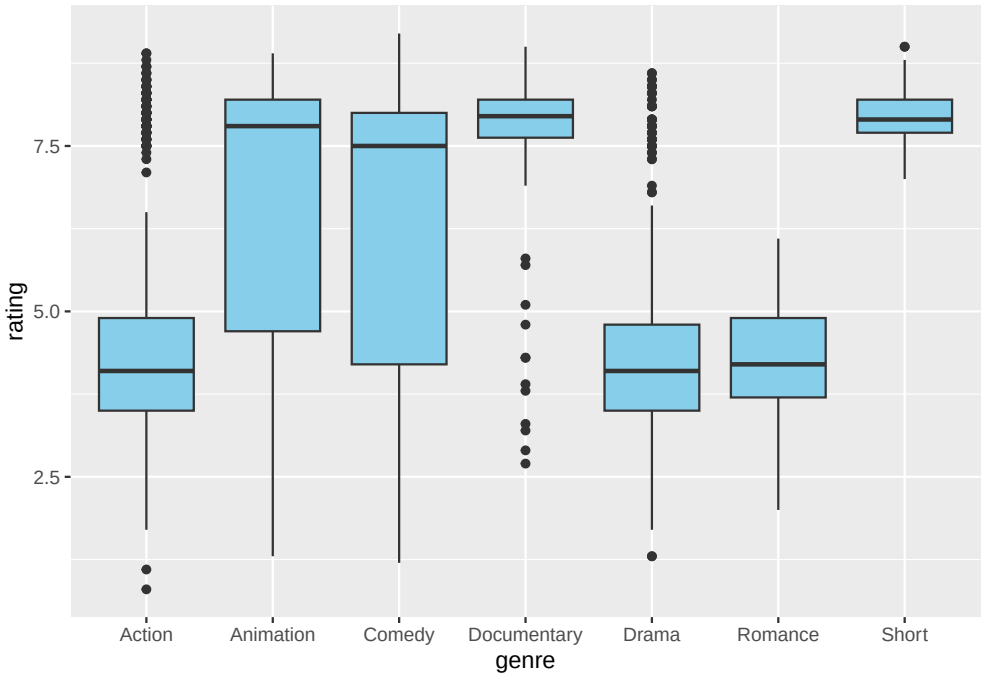

Figure 2: Boxplot of rating by genre.

3 Formal Data Analysis

4 Conclusion