Analysing IMDB film dataset to predict IMDB ratings

1 Introduction

With the gradual improvement of materialistic lifestyles, the pursuit of human life is not only limited to having good food and clothing, but also begins to have spiritual goals. Jackson T and Marks N. believe that people's life has shifted from material pursuit to spiritual pursuit in the past 40 years according to their consumption outlook (Jackson T, Marks N.1999). In recent years, watching movies has become an indispensable form of entertainment(Kubrak T.2020). According to sociological analysis, a film is not only a good medium to spread cultural and social values, but also a kind of relaxed entertainment (Balabantaray S R.2014). In most movies, the social phenomena expressed are deeply engraved in the minds of audiences by bringing people emotions such as touching, happiness, sadness and memories (Homan R L.1997).

In order to make a better a successful movie, this analysis will explore the influencing factors of movie ratings by analysing 3001 movies from the IMDB movie database. IMDb, originally a hobby organization of international fan groups, is currently a website owned by Amazon and contains as much information as possible about every movie (Peralta v. 2007). According to the description of films in the database, it can beseen that they are related to six main factors, which are year, running time, budget, number of positive votes, genre of film, and IMDB rating. Throughout the analysis, we'll be looking at different charts and then establish a generalized linear model to analyse the relationship between IMDB rating and several important factors affecting the rating. The rating score of 7 is critical boundary to judge the film quality. The knowledge of significant influencing factors can be applied to the field of film production.

This paper mainly consists of six parts. The Section 1 is introduction. The Section 2 is the main definition and method of the factors used in the analysis. The Section 3 is the exploratory data analysis of IMDB rating and these factors, and studies the relationship between them. In Section 4, the generalized linear model (GLM) is used to analyze the data. The Section 5 is the summary of the analysis. The Section 7 is the reference.

2 Main Definitions And Method

2.1 Population and Data

2.1.1 Population

3001 movies from the IMDB movie database

2.1.2 Data Collection

The data comes from the IMDB movie database. The variables are shown in the followed table.

2.2 Objective

2.2.1 Primary Objective

Analyze the properties of films that influence whether they'll be rated greater than 7 or not.

Table 1: Variable description

Variable	Type	Description
film.id	Identification variable	Unique identifier for the film
year	Continuous independent variable	Year in which the film was released in cinemas
length	Continuous independent variable	The run-time of the film(in minutes)
budget	Continuous independent variable	Budget required for the film production(in millions)
votes	Continuous independent variable	Number of positive votes received by the viewers
genre	Independent Categorical Variable	Genre of the film
rating	Response variable	IMDB Rating from 0-10(10 being the highest)

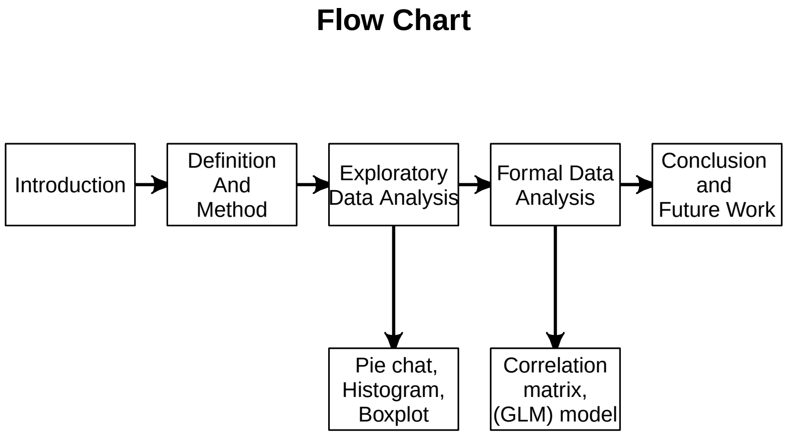
Table 2: Genre

Different_Genre
Action
Animation
Comedy
Documentary
Drama
Romance
Short

2.2.2 Secondary Objectives

Analyze the most significant features which influence the IMDB ratings of a film. Analyze the relationship between the six properties of films namely, year, running time, budget, number of positive votes, genre and IMDB rating.

2.3 Work-flow


Flow Chart
Introduction
Definition And Method
Exploratory Data Analysis
Formal Data Analysis
Conclusion and Future Work
Pie chat, Histogram, Boxplot
Correlation matrix, (GLM) model

3 Exploratory Data Analysis

3.1 Basic Feature Analysis of Films

First we'll just take a look at the dataset. First few rows of the dataset are shown below.

Table 3: Visualization of part of the data

year	length	budget	votes	genre	rating
1984	103	14.4	17	Comedy	8.0
2001	60	10.2	11	Documentary	8.1
1999	105	13.4	3216	Documentary	7.9
1970	135	11.6	73	Comedy	7.1
1939	117	17.0	1988	Action	8.0
1961	90	10.7	7	Action	2.8

Then we summarize this dataset.

Table 4: Summary statistics

name	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
Year	3001	1975.9	24.1	1895.0	1957.0	1983.0	1997.0	2005.0
Length	3001	81.6	38.7	1.0	73.0	90.0	100.0	555.0
Budget	3001	12.0	3.0	1.2	10.1	12.1	14.0	23.4
Votes	3001	655.8	3780.1	5.0	11.0	30.0	118.0	92437.0
Rating	3001	5.4	2.1	0.8	3.7	4.7	7.8	9.2

Firstly, in order to explore the data, the summary of different variables is necessary. The above table is the summary table of five numeric variables along with their descriptive statistic. The general analysis of the sample data shows that the total sample size is 3001. The oldest film in the sample was released in the year 1895, and the latest was released in 2005. The average length of the films is 81.6 minutes, with the shortest being just one minute and the longest bring 555 minutes. Budgets ranges from 1.2 million dollars to 23.4 million dollars, and according to the quartile, most films still cost more than 10 million dollars. It can be seen from the standard deviation of votes that there is a large difference between different films, with the standard deviation reaching 3780.1, while the average value is only 655.8. In terms of film rating data, most films are rated around 5 out of 10, and the best film being rated as 9.2 while the lowest is only rated 0.8.

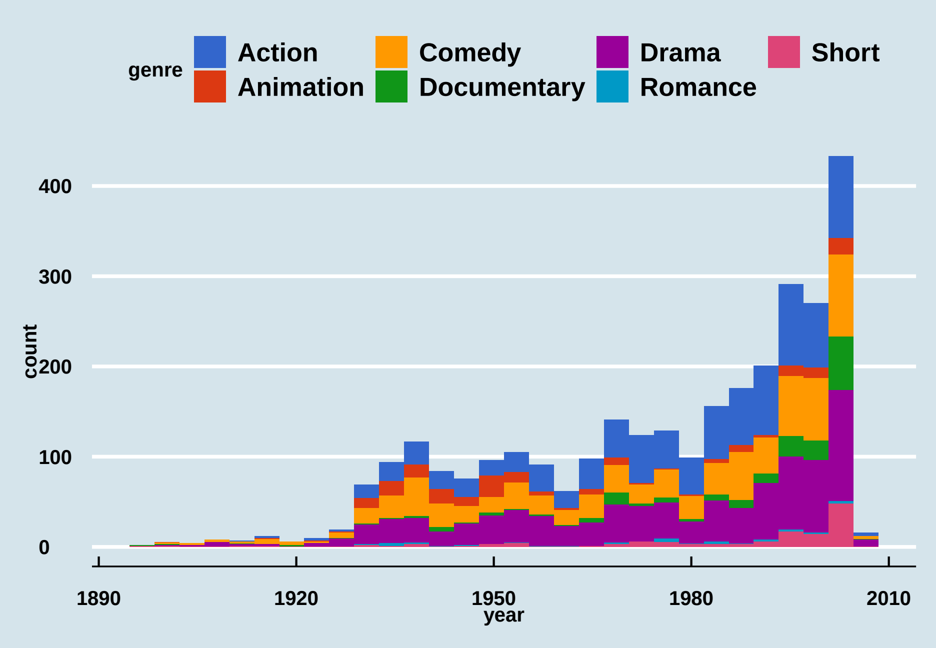


Figure 1: Year histogram

Figure 1 is a bar chart which illustrates the number of different kinds of films published in different years from 1890 to 2005. According to the trend in the bar chart, it can be seen that with each passing year, the number of the sum of all kinds of films getting released also increased. Short films and proportion of drama films can be seen ton have a gradually increasing trend. The number of documentary films also begun to increase in recent years, but the proportion it shares with other genre is still not high. Action films and comedy films have always dominated, while animated films, after their golden age from 1930 to 1956, now account for a relatively small proportion. The number of romantic films being released are always in acute numbers.

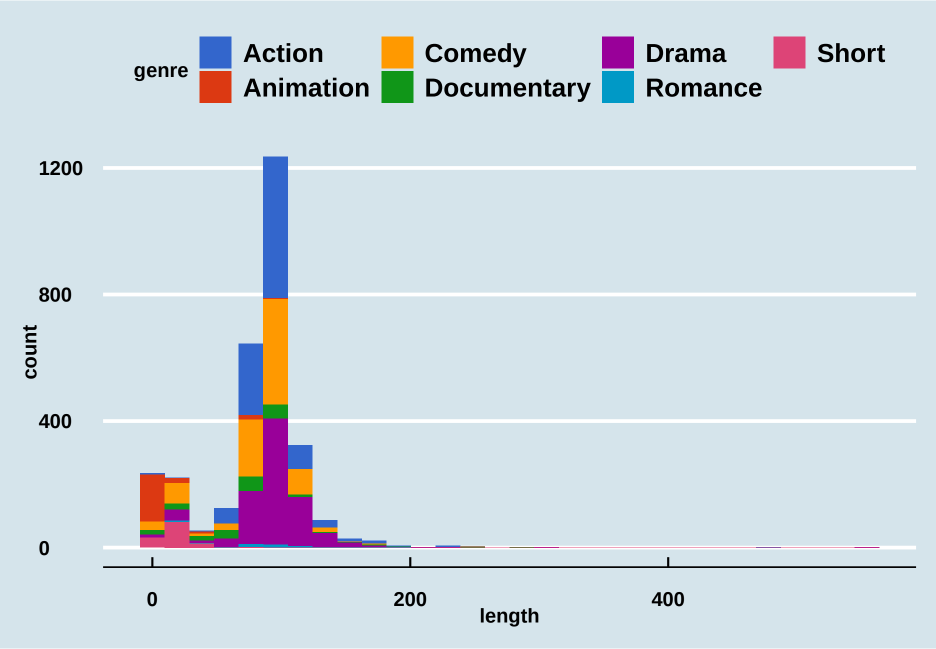


Figure 2: Length histogram

Because the data has many outliers(which is much larger) in the length variable, so we use a log transformation of this variable and then draw a histograms with the log(length).

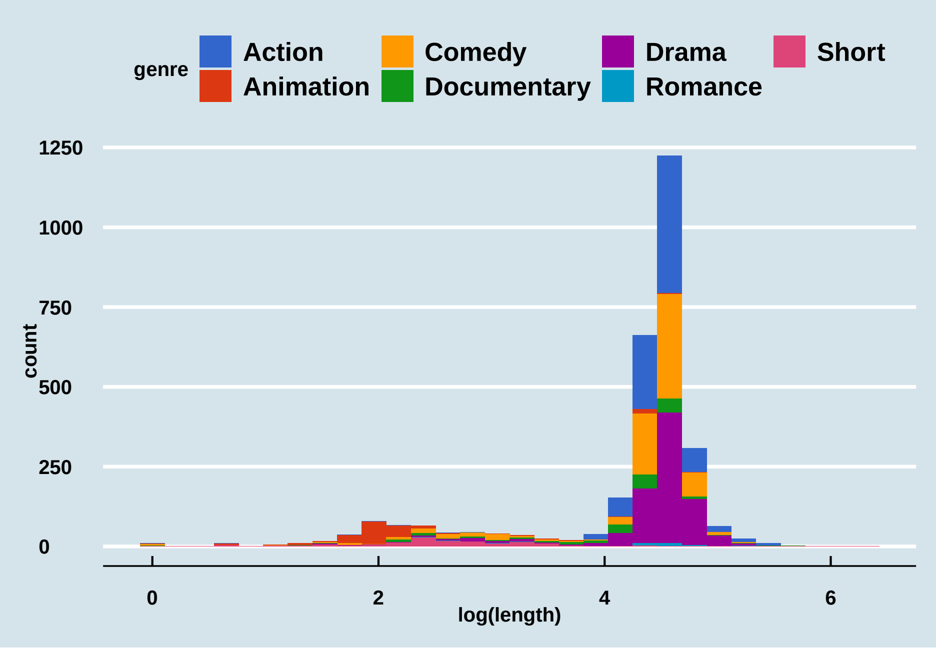


Figure 3: log(Length) histogram

After comparing Figure 2 using data without transforming into log and the Figure using data with log transformation, We can see the later one is better. We find that Figure 3 describes the number of different genres of films related to the length. According to the bar chart, the log(length) of a large number of films are located between 4 and 5(about 55.5 to 151.6 minutes). Almost all Short films and Animation films are under 55.5 minutes while Action films are mostly above 55.5 minutes. In addition, Drama and Comedies have different length interval.

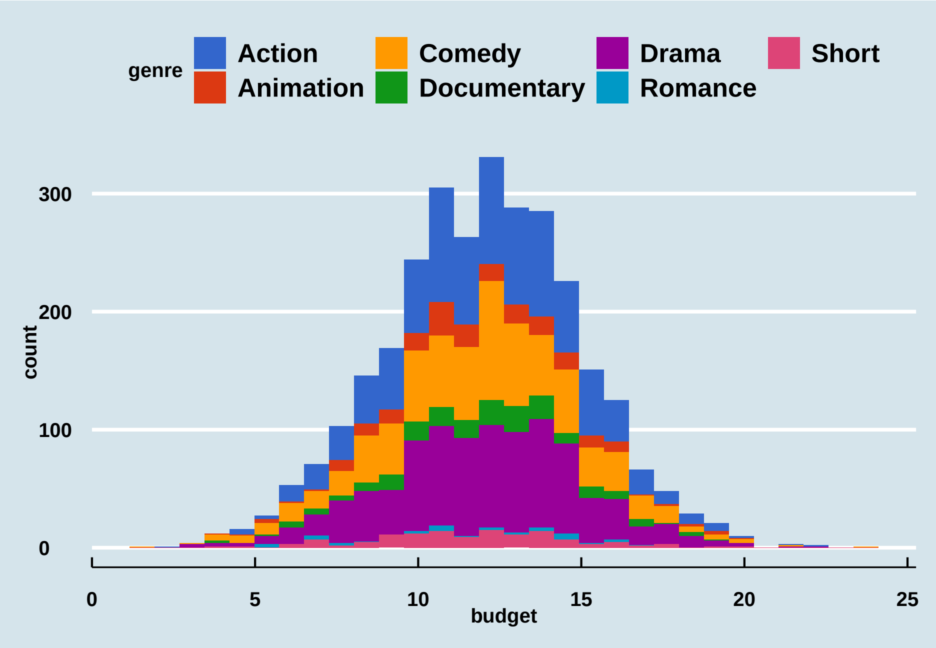


Figure 4: Budget histogram

Figure 4 is the relationship between budget and the number of different kinds of films. It's not difficult to see that this is a bell shape histogram. The film, whichever genres it belongs to, the count rises first and then falls and reachs peak at about 12 million dollars.

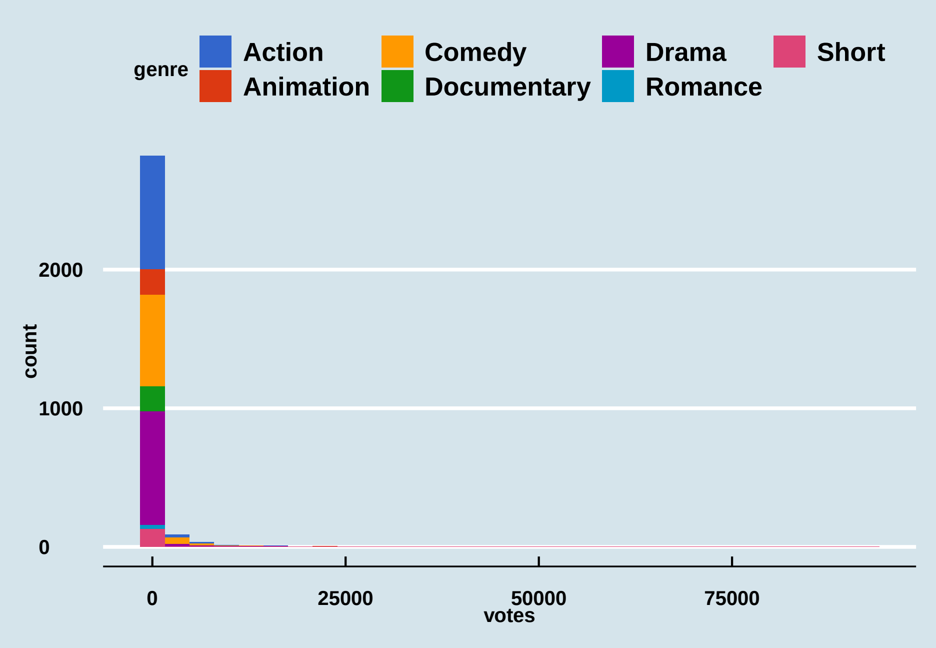


Figure 5: Votes histogram

Because most of the data is located near zero and has many outliers(which is much larger) in the votes variables, we log transform this variable and then draw a histograms with the log(votes).

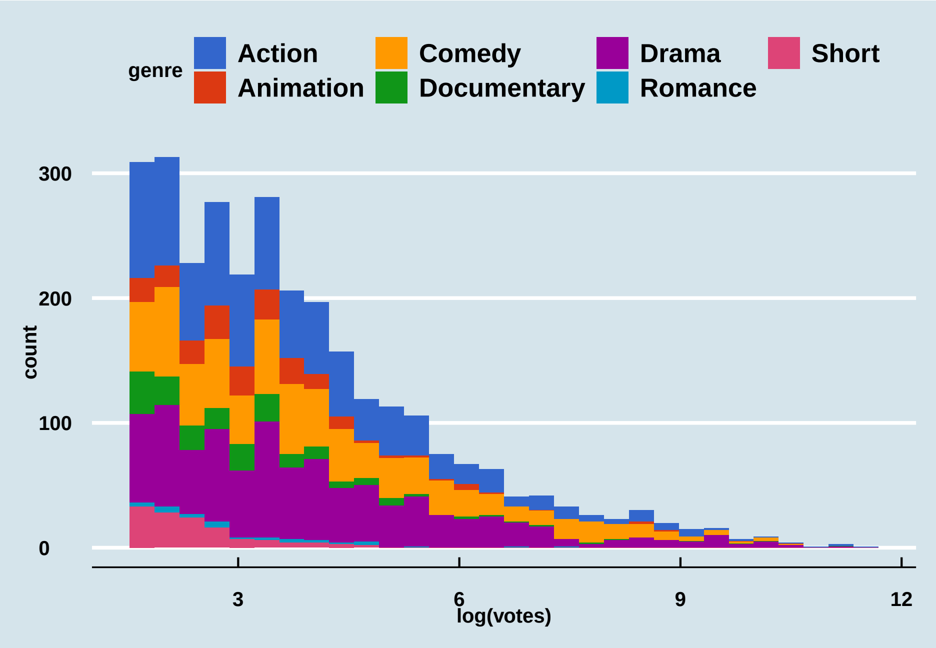


Figure 6: log(Votes) histogram

Comparing Figure 5 using data without log transformation and Figure 6 using the data with log transform, We see the later one is better. Figure 6 is the relationship between votes and the number of different kinds of films. As we can see, small number of films receive a positive evaluation from the audience, while most of the movies receive low or moderate positive votes.

3.2 Characteristic Comparative Analysis of Films with Different Rating

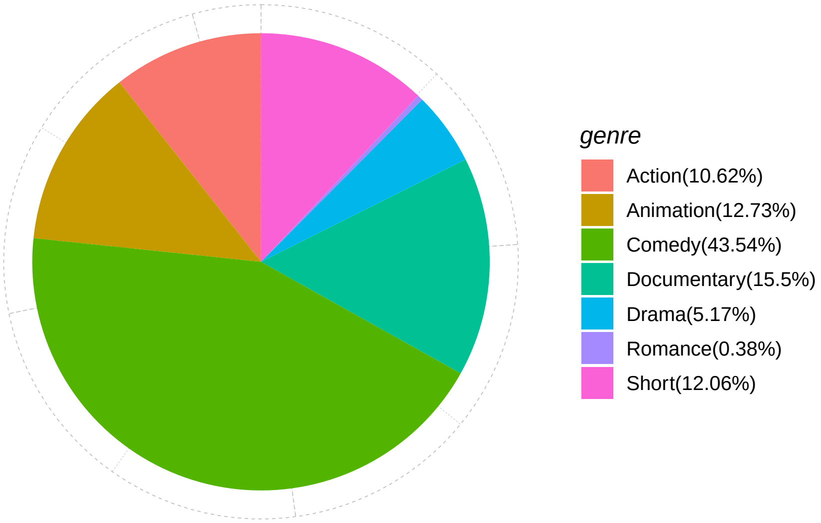


Figure 7: genre pie chat(greater than 7)

Figure 7 is a pie plot which shows the proportion of films of different genre that gets rated higher than 7. We can see that Comedy, Documentary and Animation are the top 3 genres of films which account for 43.54%, 15.5% and 12.73% respectively.

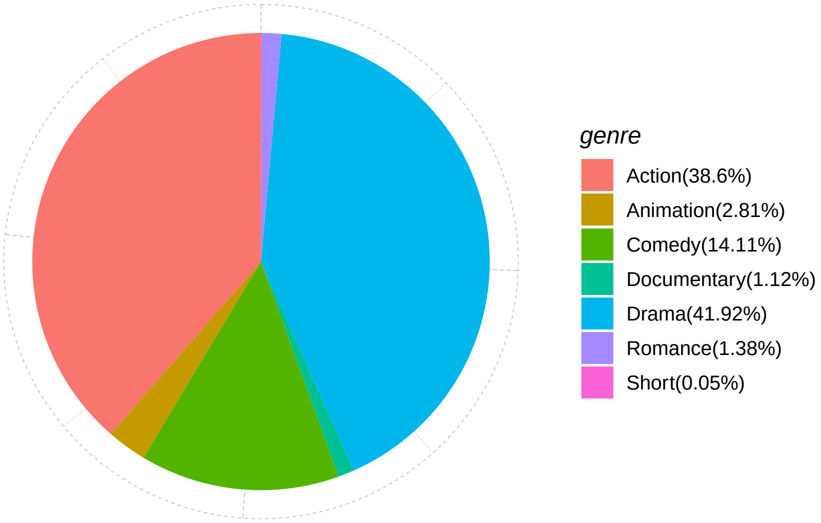


Figure 8: genre pie chat(lower than 7)

Figure 8 is a pie plot which shows the proportion of different genre of films that gets rated lower than 7. According to the plot we can see that the Drama and Action almost account for 80.52%.

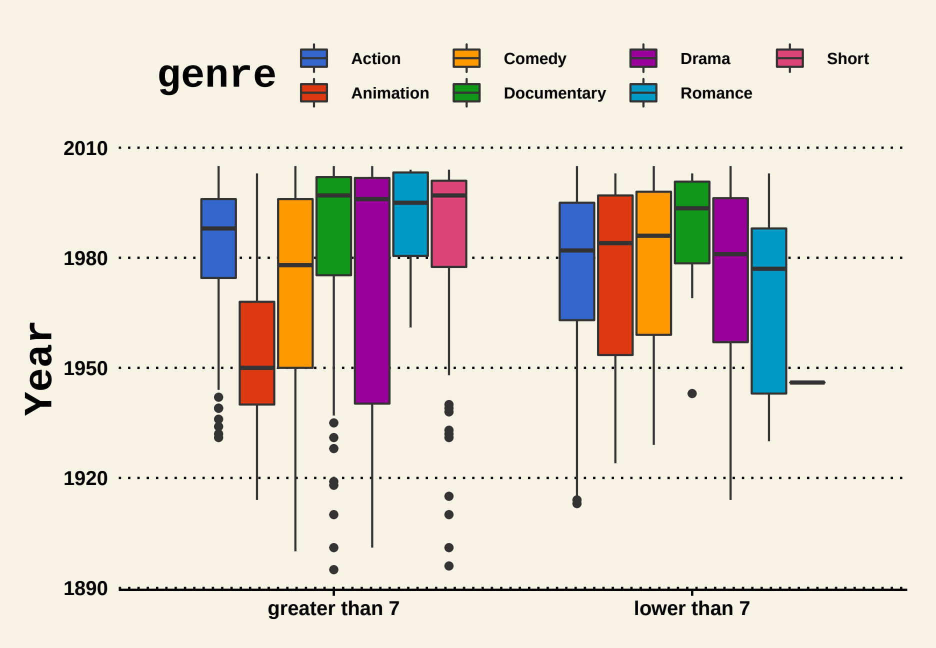


Figure 9: Year boxplot

Figure 9 is a box plot shows the distribution of different kinds genres which have scores lower and greater than 7 in different years. According to the plot we can see something special, almost all genres of films have a similar distribution irrespective of ratings(lower than 7 or upper than 7) except for Animation and short. It can be easily interpreted that in early years Animated films had higher rating. While after 1980, the ratings of Animated films reduced, large number of films are seen to get rated lower than 7. In addition, It can be seen that very few short films before 1950 got rated lower(Excluding the outliers) which mean Short films are easy to get rated higher on IMDB.

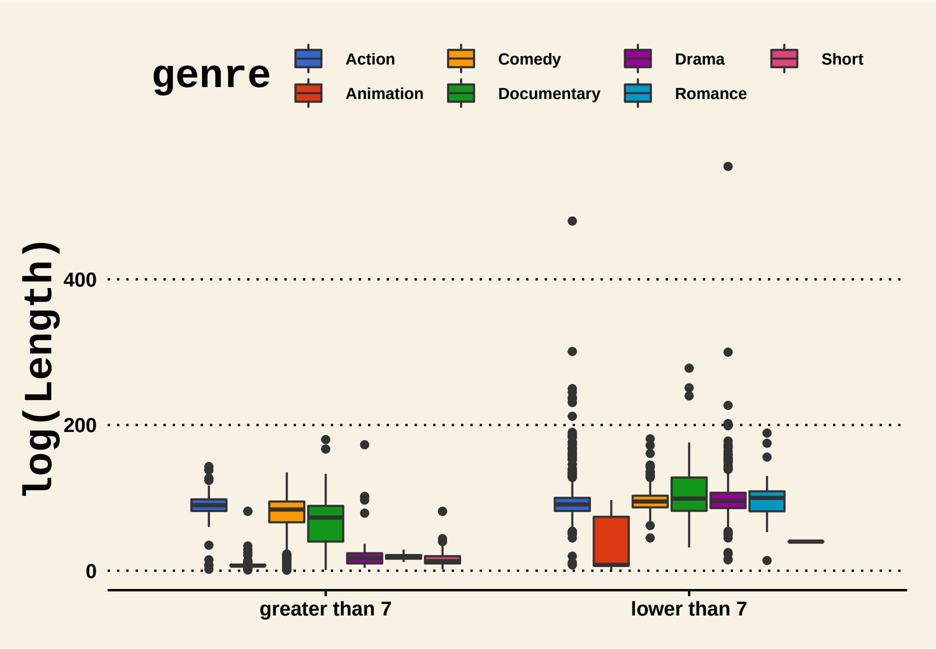


Figure 10: Length boxplot

Because the data has many outliers in the length variable, so we log transform this variable and then draw boxplots with the log(length).

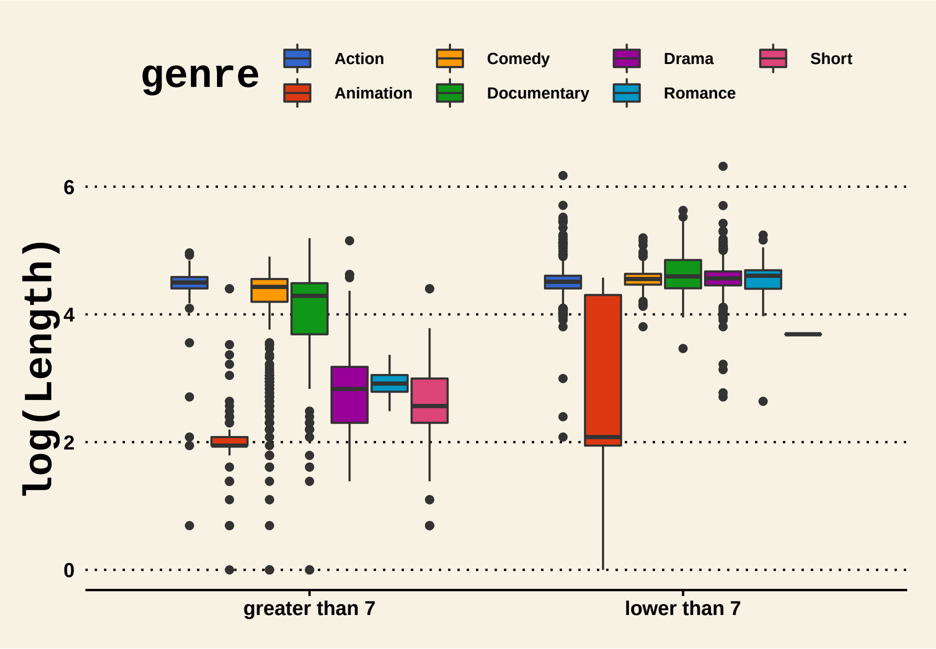


Figure 11: log(Length) boxplot

So after comparing the Figure 10 (without transforming into log) and the Figure 11 (using data with transformation), We see the later gives better visual representation. Figure 11 shows the distribution of different kinds of films which have scores lower and upper than 7 of different length. We can easily get the information that the shorter the Drama, Romance and Short films are, higher are the chances for movie of these geners to get rated greater than 7.

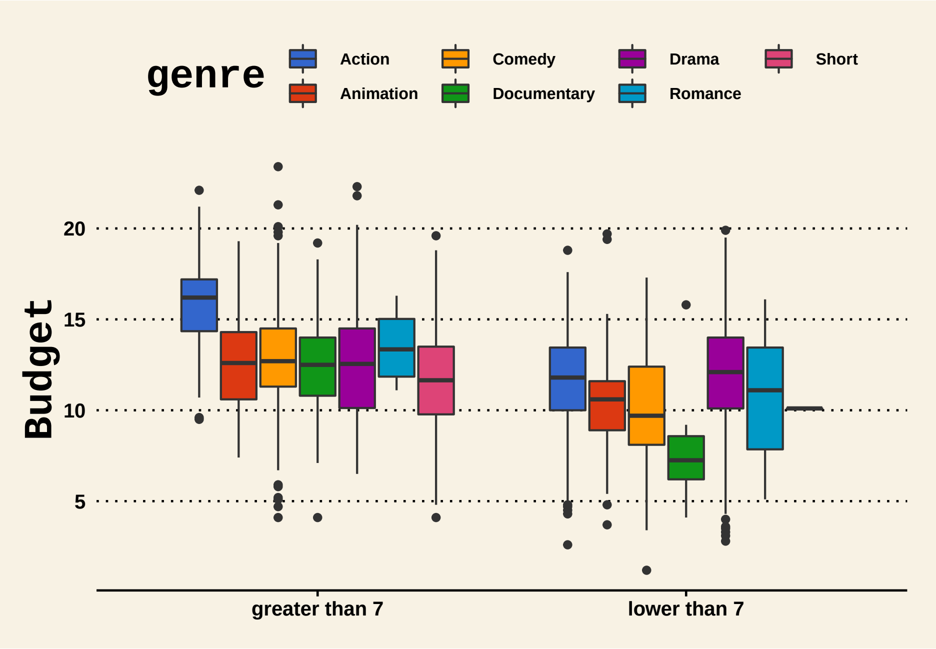


Figure 12: Budget boxplot

From figure 12, we can clearly see that on an average, films with higher budget tend to get rated greater than 7. It is reasonable that a guest appearance of a well-known actor, cost of special effects production, magnificent sets are all in the budget. And these elements can all add up to an increased budget and influence to the rating.

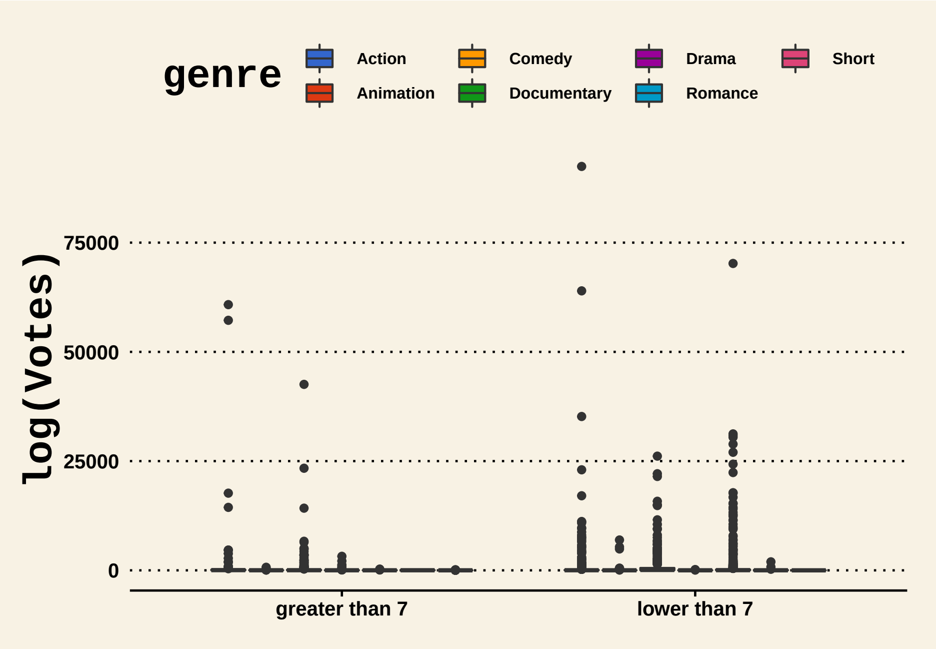


Figure 13: Votes boxplot

Since the votes data has too many outliers, we log transform this variable and then draw a plot of boxplots with the log(votes) on vertical axis.

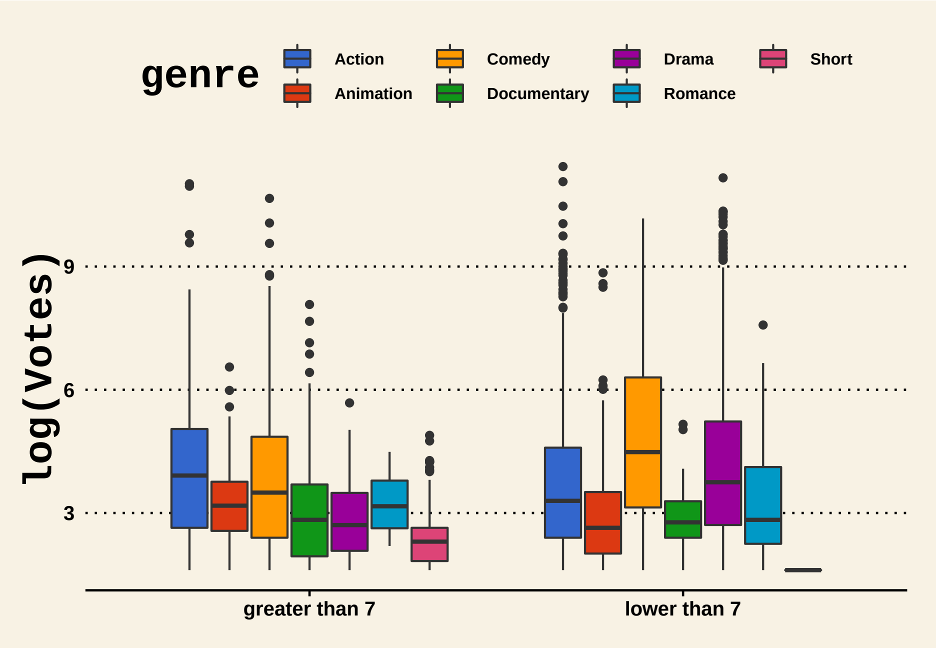


Figure 14: log(Votes) boxplot

Figure 13 and Figure 14 are boxplots of different genre plotted against votes, the later one is log transformed. Figure 14 shows the distribution of different genres of films which have scores lower and greater than 7 plotted against log(votes). From the plot we can see that on an average, films get similar amount of positive votes irrespective of the ratings. Almost all the genre have similar distribution and even when the films are grouped by ratings.

In above figures, we can see some similar phenomenons. Firstly, in all four plots, we see many outliers which may be because the variable score is a very subjective. This means no matter how good a movie is, there will always be people who dislike it. Secondly, lower-than-7 Short film's distribution is always a "line" when boxplots are seen, which means there are very few short films which got rated lower than 7. Finally, we cannot see any obvious relationship between the four variables (year, length, budget and votes) and the score through the boxplots, so the following exploration is necessary.

3.3 Correlation Analysis Between Independent Variables

According to previous studies(histograms and boxplots), we find that it is necessary to change votes and length into log(votes) and log(length) to reduce the effect of outliers. So we use the log(votes) and log(length) in the correlation analysis.

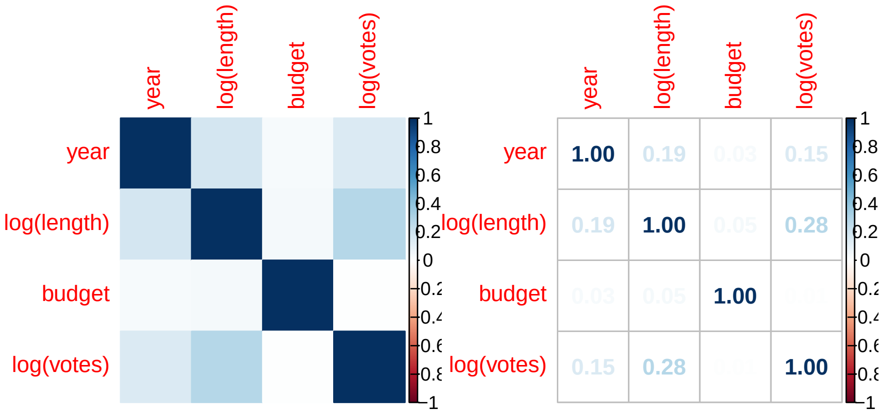


Figure 15: The relationship of factors related to independent variables

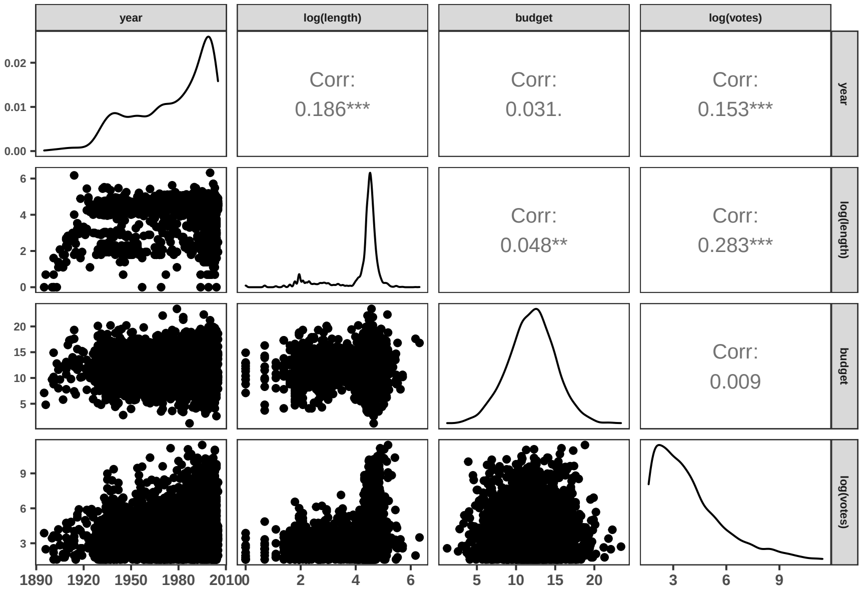


Figure 16: Correlation matrix of data

Figure 15 and Figure 16 are the correlation matrix between the variables (year, length, budget and votes). According to the correlation matrix, we can see that these variables are not much related to each other which means we should not be worried about the multicollinearity problem in the following data analysis.

4 Formal Data Analysis

4.1 The Construction of GLM Model

From the previous studies(histogram and boxplot), we find that it is necessary to change votes and length into log(votes) and log(length) to reduce the effect of outliers. So we use the log(votes) and log(length) in the Formal Data Analysis analysis.

For the pre-processing of the data, the rating variable is converted to a binary variable: $\mathbb{I}_{\text{score}}(i)$ is an indicator function such that

$$\text{score}(i) = \begin{cases} 1 & \text{if score of observations is greater than 7,} \\ 0 & \text{if score of observations is lower than 7.} \end{cases}$$

we can use a logistic regression model for the probability of success(movie getting rated greater than7). We first write it down in mathematical notation by letting $\mathrm{p_i} = \mathrm{P}(\mathrm{Y_i} = 1)$ denotes the probability of success for the film. We assume that the $\mathrm{Y_i}$ are independent random variables which follow the $\mathrm{Bin}(1, \mathrm{p_i})$(or Bernoulli $(\mathrm{p_i})$ distribution. The full GLM can be written as

$$log[(\mathrm{p_{ij}})/(1-\mathrm{p_{ij}})] = \widehat{\alpha}+\widehat{\beta}_1\cdot\mathrm{year_j}+\widehat{\beta}_2\cdot\log(\mathrm{length})_\mathrm{j}+\widehat{\beta}_3\cdot\mathrm{budget_j}+\widehat{\beta}_4\cdot\log(\mathrm{votes})_\mathrm{j}+\widehat{\beta}_5\cdot\mathbb{I}_{\mathrm{genre}}(i)+\epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where

$\mathrm{p_{ij}}$ is the probability of the $j$thth films score is greater than 7 in the ith genre;

the intercept α is the mean score for the baseline category of Action genre

$t\beta_{\text{1-4}}$ is the value of score that changes as the corresponding independent variable changes by one unit;

β_5 is the difference in the mean score of 6 other genres;

$\text{variables}_\mathrm{j}$ is the jth observations of the continuous variables;

and

$\mathbb{I}_{\text{genre}}(i)$ is an indicator function such that

$$\mathbb{I}_{\text{genre}}(i) = \begin{cases} 1 & \text{if genre of } j\text{th observations belongs to the genre}(i), \\ 0 & \text{if genre of } j\text{th observations belongs to the genre other than genre}(i). \end{cases}$$

Here is the result of the model

Table 5: Estimates of the parameters from the fitted GLM regression model1.

	Est.	S.E.	z val.	p
(Intercept)	-2.0576839	5.8118236	-0.3540513	0.7233004
year	0.0032561	0.0029827	1.0916521	0.2749860
log(length)	-3.0157109	0.1991005	-15.1466767	0.0000000
budget	0.5350513	0.0296474	18.0471651	0.0000000
log(votes)	-0.0030168	0.0383801	-0.0786030	0.9373484
genreAnimation	-2.6784479	0.4932708	-5.4299745	0.0000001
genreComedy	3.2136717	0.1745749	18.4085507	0.0000000
genreDocumentary	5.1205571	0.3542294	14.4554834	0.0000000
genreDrama	-1.7807422	0.2437356	-7.3060398	0.0000000
genreRomance	-1.0297621	0.8516224	-1.2091769	0.2265949
genreShort	4.3200190	1.0639637	4.0603066	0.0000490

According to the p-value, we can see two variables are not or partly not significant. First is year(p-value of 0.2749860) , second is log(votes)(p-value of 0.9373484) and last is genre(p-value of genreRomance is 0.2265949).

4.2 The Selection of GLM Model by Stepwise Regression Approach

We use the method of stepwise regression to choose the best model which contains all the significant independent variables and has the lowest AIC score,

```
Start:  AIC=1545.08
score ~ year + log(length) + budget + log(votes) + genre

              Df Deviance    AIC
- log(votes)   1   1523.1 1543.1
- year         1   1524.3 1544.3
<none>             1523.1 1545.1
- log(length)  1   1999.1 2019.1
- budget       1   2028.5 2048.5
- genre        6   2759.0 2769.0

Step:  AIC=1543.08
score ~ year + log(length) + budget + genre

              Df Deviance    AIC
- year         1   1524.3 1542.3
<none>             1523.1 1543.1
+ log(votes)   1   1523.1 1545.1
- budget       1   2028.6 2046.6
- log(length)  1   2030.9 2048.9
- genre        6   2759.3 2767.3

Step:  AIC=1542.29
score ~ log(length) + budget + genre

              Df Deviance    AIC
<none>             1524.3 1542.3
+ year         1   1523.1 1543.1
+ log(votes)   1   1524.3 1544.3
- budget       1   2029.7 2045.7
- log(length)  1   2045.1 2061.1
- genre        6   2771.4 2777.4

Call:
glm(formula = score ~ log(length) + budget + genre, family = binomial(link = "logit"),
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9020  -0.3453  -0.1161   0.1805   3.3849

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.22330    0.79889   5.286 1.25e-07 ***
```

```
log(length)      -2.98172    0.19104 -15.608  < 2e-16 ***
budget            0.53453    0.02959  18.062  < 2e-16 ***
genreAnimation   -2.65115    0.48396  -5.478 4.30e-08 ***
genreComedy       3.20445    0.17224  18.604  < 2e-16 ***
genreDocumentary  5.16415    0.35106  14.710  < 2e-16 ***
genreDrama       -1.77421    0.24228  -7.323 2.43e-13 ***
genreRomance     -0.99242    0.83917  -1.183    0.237
genreShort        4.37694    1.06083   4.126 3.69e-05 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3879.3  on 3000  degrees of freedom
Residual deviance: 1524.3  on 2992  degrees of freedom
AIC: 1542.3

Number of Fisher Scoring iterations: 7
```

The AIC of model3(with independent variables log(length), budget and genre) is 1542.3 and according to the stepwise regression, we know this model is the best one with a lowest AIC. So we choose this model.

4.3 Analysis of The Optimal GLM Model

we finally analyse the results from the Optimal GLM Model.

$$log[(\mathrm{p_{ij}})/(1-\mathrm{p_{ij}})] = \widehat{\alpha} + \widehat{\beta}_1 \cdot \mathrm{log(length)_i} + \widehat{\beta}_2 \cdot \mathrm{budget_i} + \widehat{\beta}_3 \cdot \mathbb{I}_{\mathrm{genre}}(i) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

Table 6: Estimates of the parameters from The Optimal GLM Model.

	Est.	S.E.	z val.	p
(Intercept)	4.2233004	0.7988877	5.286475	0.0000001
log(length)	-2.9817179	0.1910419	-15.607663	0.0000000
budget	0.5345333	0.0295940	18.062214	0.0000000
genreAnimation	-2.6511514	0.4839597	-5.478041	0.0000000
genreComedy	3.2044451	0.1722443	18.604072	0.0000000
genreDocumentary	5.1641478	0.3510643	14.709977	0.0000000
genreDrama	-1.7742132	0.2422830	-7.322896	0.0000000
genreRomance	-0.9924192	0.8391659	-1.182626	0.2369575
genreShort	4.3769420	1.0608271	4.125971	0.0000369

Table 7: Estimates of Odds In 95percentage confidence Interval.

	odd_ratio	2.5 %	97.5 %
(Intercept)	68.2583908	14.9192403	343.3447329
log(length)	0.0507057	0.0343656	0.0727095
budget	1.7066516	1.6126514	1.8111438
genreAnimation	0.0705699	0.0266336	0.1779787
genreComedy	24.6418232	17.6953605	34.7754332
genreDocumentary	174.8883611	90.2359522	358.0368420
genreDrama	0.1696169	0.1038577	0.2690718
genreRomance	0.3706789	0.0569247	1.6142459
genreShort	79.5942587	14.9206871	1482.0931919

We see from the output that the coefficient for length is negative, indicating a lower score for films with longer length. Similarly the coefficients for genreAnimation, genreDrama and genreRomance are negative.

To quantify the effect of each of these predictors, we look at odds ratios which can be computed as $exp(\widehat{\beta})$. These are shown in the plot below.

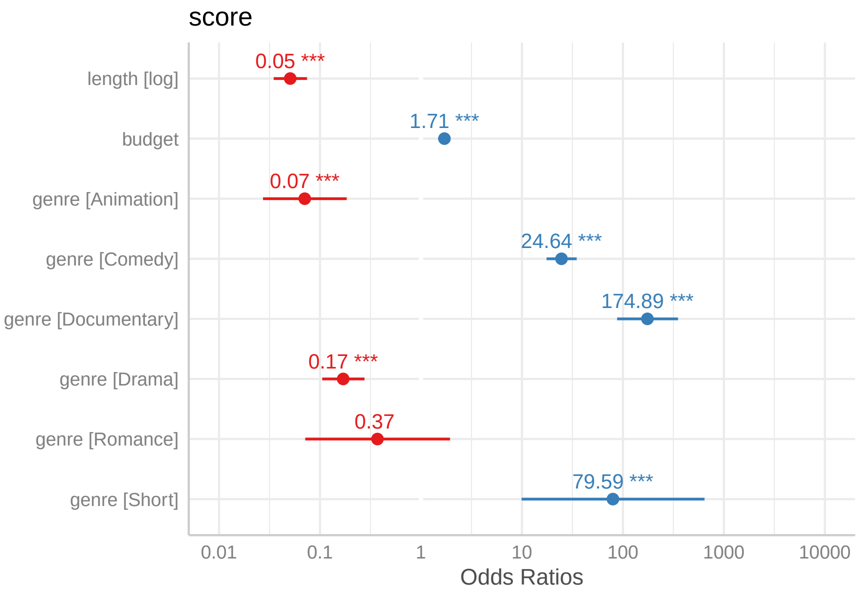


Figure 17: Odds Ratios Plot

We interpret the odds ratios as follows:

1. For each minute increase in the log(length) of film, the odds of film getting rated greater than 7 decrease (get multiplied by a factor of 0.05);
2. For each $1000000 increase in the budget of film, the odds of film getting rated greater than 7 increase (get multiplied by a factor of 1.71);
3. Animation's odds of score is 0.07 times those of Action; Comedy's odds of score is 24.64 times those of Action; Documentary's odds of score is 174.89 times those of Action; Drama's odds of score is 0.17 times those of Action; Short's odds of score is 79.59 times those of Action; Romance's odds of score is same as those of Action(not significant).

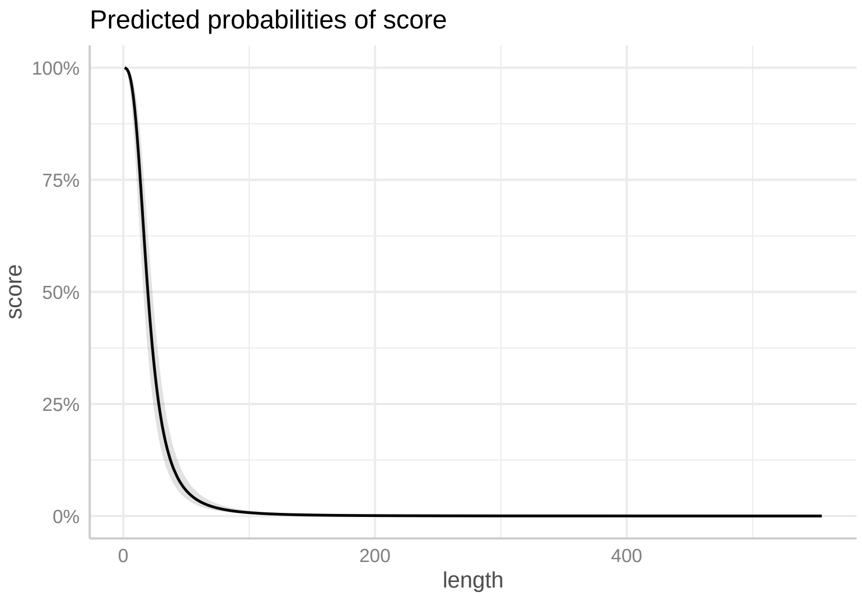


Figure 18: Predicted Probabilities Of Film Score Plot

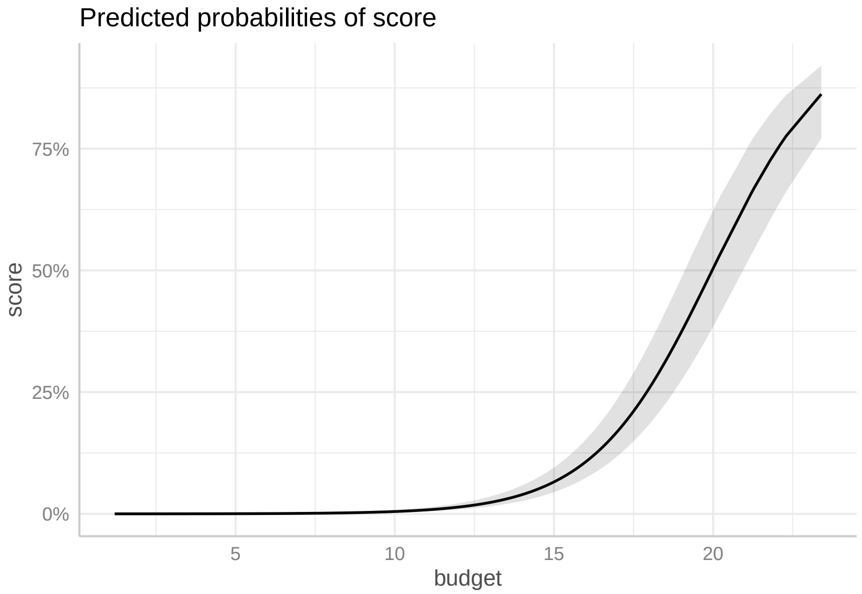


Figure 19: Predicted Probabilities Of Film Score Plot

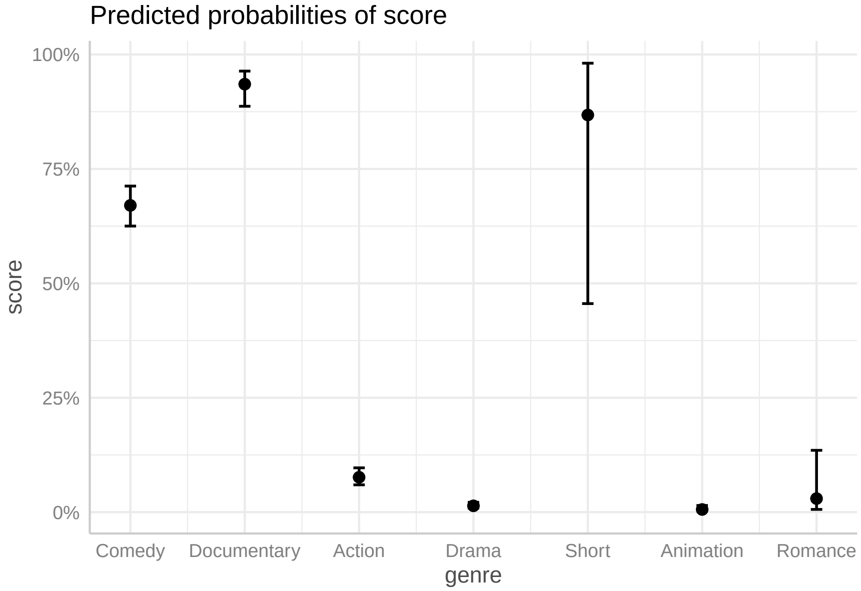


Figure 20: Predicted Probabilities Of Film Score Plot

4.4 Analysis of assessing the predictive power

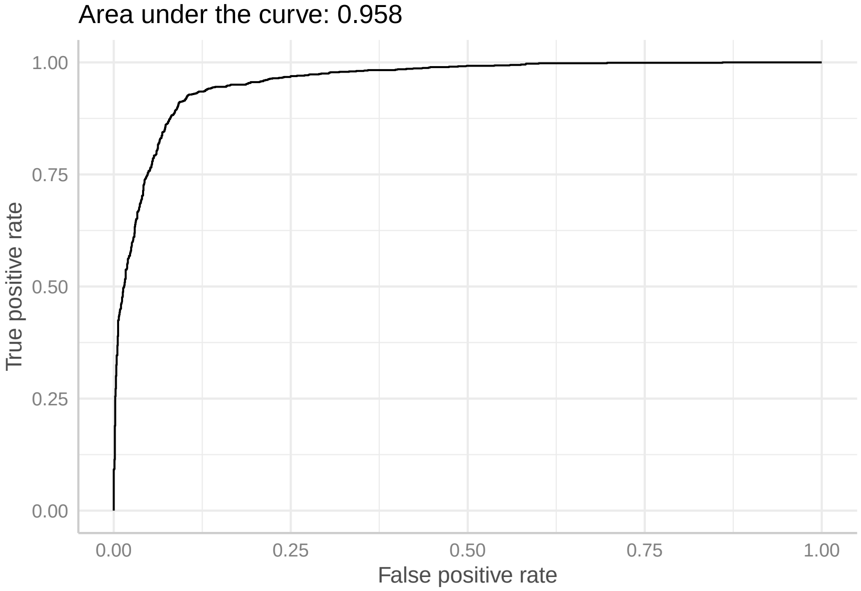


Figure 21: Predicted Probabilities Of Film Score Plot

On the horizontal axis of the ROC curve we have the false positive rate and on the vertical axis the true positive rate. The area under the ROC curve, known as AUC, is used as a measure of a diagnostic test's discriminatory power. The area under the curve is about 0.958. So the model2 performs better than a random guess and the model2 which we choose has a good predictive power.

5 Conclusions

IMDB ratings can be considered as a criterion to judge a movie by it's entertainment value. Thus, IMDB ratings might influence the viewer's decision to watch a film or not. Therefore, factors influencing the IMDB ratings may be of atmost priority to film producers.

From the analysis, It can be concluded that factors like length of the movie, budget of the movie and genre can highly influence IMDB ratings and can determine which movie might get rated above 7 or less.

Growing time constraints may be a significat reason as to why length of a film may affect its IMDB rating. Audience, nowadays, prefer watching a movie with high entertainment value which gets over within a brief period. Hence, shorter the runtime of a movie, the more likely it is for a movie to get rated greater than 7.

The majesticity of a movie can play a huge role in affecting the ratings of a movie, A visually appealing movie has a higher chance of getting rated more than 7 on IMDB. These visual effects and technicalities may lead to an increasing budget of the movies, consequently making budget as an important predictor of IMDB ratings greater than 7 or less.

People tend to favour watching a movie of a particular genre than other, though it a highly subjective choice. The analysis show that films of some genres perform well in getting rated above 7 on IMDB than others. It can be concluded from the analysis that a Comedy film, documentary style film or a short film have better odds of getting rated above 7 on IMDB compared to a action film. Hence, genre of a film may also be helpful in predicting whether or not a movie will be rated above 7 on IMDB ratings or not.

6 Future work

The dataset observed of the analysis is just a subset of IMDB movie database, and in our model we supressed outliers. In this paper, the rating level of 7 is considered as the standard of judging a film(good or bad). The analysis of plethora of movies, multiple databases can be done to reduce prediction errors of IMDB ratings in the future. This paper mainly uses a generalized linear model to analyze the data, and future work may involves deeper analysis of films and genre. In the future work, we can mainly focus on the following points:

1. Add rating weights for films from different decades to the model. Because different eras have different cultural interests, a film that is rated less than 7 in one decade could be rated higher than 7 if released in another decade.

2. The dataset chosen for the analysis is just a subset of IMDB films database. In future, we can try to study the complete database and several other movie databases at the same time and compare and analyze them, so as to provide a comprehensive result and determine all the factors that might affect the IMDB ratings

3. The way to build the model can be changed to more complex logistic regression which might give a better prediction with low error rates.

In the analysis of films getting rated greater than 7, this paper finds that comedies account for 43.54 percent which is a large a number, so it needs to be backed with several other analysis of different genre of film in order to convince the targeted audience. Also, for films of different eras, outliers mostly appeared before 1950. Whether the rating of that era is valid needs to be explored, because 70 years have passed, and cultural and social differences might affect the rating level. So we should do a weighted analysis of films from different eras.In addition, average budget required for different types of films is not consistent. For example, an action blockbuster requires a large budget for setting special effects, and action sequences, while a comedy usually doesn't. So the budget distribution Cannot be generalised.

7 Reference

1.Balabantaray S R. Impact of Indian cinema on culture and creation of world view among youth: A sociological analysis of Bollywood movies[J]. Journal of Public Affairs, 2014: e2405.

2.Homan R L. The Everyman movie, circa 1991[J]. Journal of Popular Film and Television, 1997, 25(1): 21-30.

3.Jackson T, Marks N. Consumption, sustainable welfare and human needs—with reference to UK expenditure patterns between 1954 and 1994[J]. Ecological economics, 1999, 28(3): 421-441.

4.Kubrak T. Impact of Films: Changes in Young People's Attitudes after Watching a Movie[J]. Behavioral Sciences, 2020, 10(5): 86.

5.Peralta V. Extraction and integration of movielens and imdb data[J]. Laboratoire Prisme, Université de Versailles, Versailles, France, 2007.