

Group_07_Analysis

Group 7

1 Introduction

IMDb is a platform that provides information , ratings and reviews on films, movies and many other streaming content. Our group was given a dataset taken from this database, which has records of 2387 films released from 1894 to 2005. Each film has a unique identification number (id) and some other measurments related to it. Those are :

- $Year_i$: the year the film was released at cinemas
- $Length_i$: duration of film in minutes
- $Budget_i$: Budget for film production in 10^5 's
- $Vote_i$: Number of positive votes
- $Genre_i$: Genre of film
- $Rating_i$: IMDb rating from 1 to 10

2 Exploratory Data Analysis

Taking a glance at the summaries of the variables shown in Table 1 we notice that length variable has many outliers , and that is given by the big difference between its maximum value and its third quartile value ,that is 399 minutes and 100 minutes respectively. Budget also has a wide range of variability and an amount of outliers . It is worth mentioning that votes have a large sandard deviation , 4370 to be exact and that might be because of an outlier that presents a huge difference with the rest. Possibly a small number of films had respectively huge number of votings compared to the rest of films.

Table 1: Summary statistics of variables in the data set.

Variable	Mean	SD	Min	Q1	Median	Q3	Max	IQR
length	81.414	37.675	1.0	72.0	90.0	100.0	399.0	10.0
budget	11.948	2.968	2.1	10.0	12.0	13.9	23.7	1.9
votes	658.969	4370.038	5.0	12.0	32.0	118.0	103854.0	86.0
rating	5.414	2.069	0.7	3.7	4.7	7.8	9.2	3.1

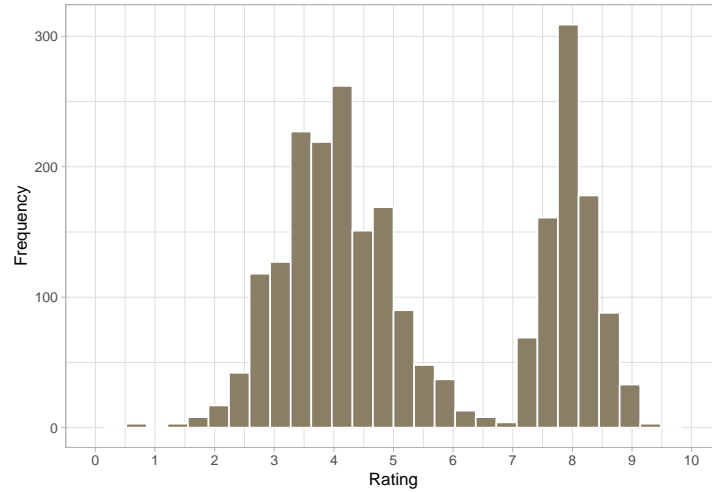


Figure 1: Density Of Ratings

Given the barplot in Figure2 we notice that the majority of films are within one of the 3 most dominant categories Action, Comedy and Drama. On the opposite Animation Documentary and Short films have relatively lower frequencies. Finally Romance films are comparatively very rare among our list of films (only 16 films where Romantic among all 2387)

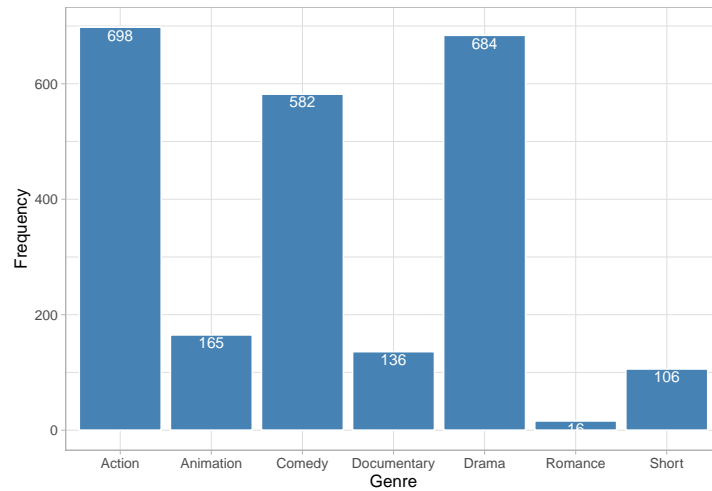


Figure 2: Frequency of each genre.

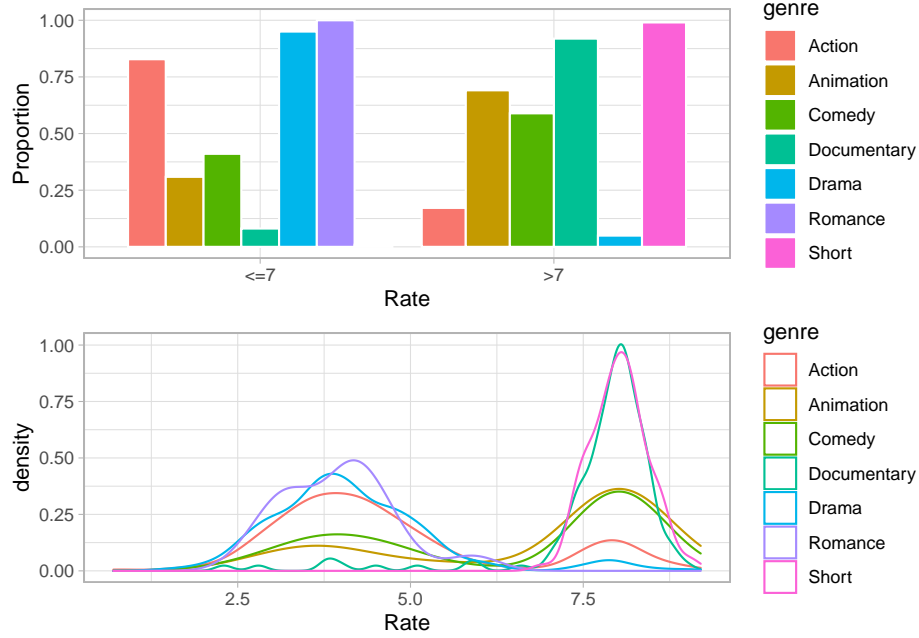


Figure 3: Distribution of genre.

Figure 3 displays two plots. The first barplot shows that Action, Romance and Drama films have higher probabilities getting rate below 7 compared to the other genres, while Documentary and Short types are more likely to get rate higher than 7. Moving on to the next density plot of the rate for each genre we notice that some genres have very similar lines. In particular we see that documentary and short films have very similar distribution of rate. The same applies for animation and comedy. As for the remaining genres, romance, action and drama we do see some more noticeable differences in their lines but the overall shape and characteristics are very similar. These observations might lead to regroup the genres to a new categorical variable of 3 levels instead of 7.

Next we explore the budget property. Given the boxplot in Figure 4 although we notice that the box of rate over 7 is slightly higher than the box of rate below 7 however the mean line is not that different.

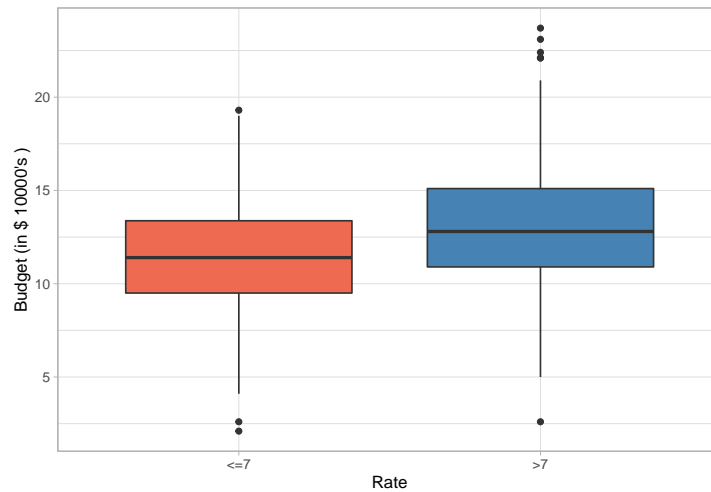


Figure 4: Budget and Rate.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1.00	72.00	90.00	81.41	100.00	399.00	92

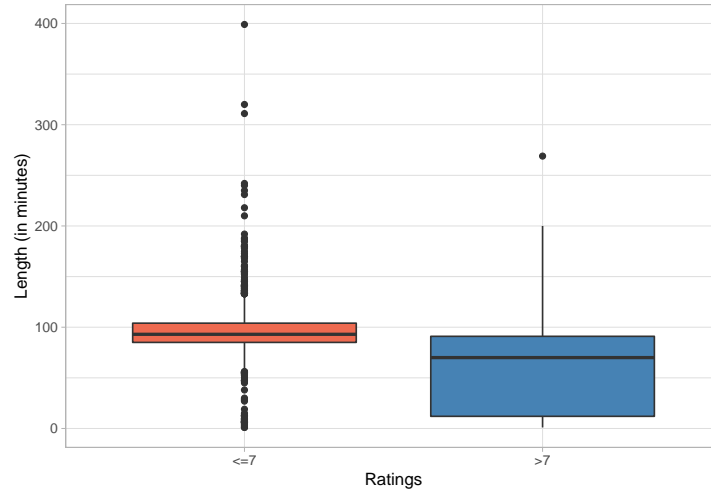


Figure 5: Rating based on length of the film.

Figure 5 shows that the length feature has many outliers. That is explained by the large number of points graphed outside the box. Those points represent films that have duration longer than 100 minutes which is our 3rd quartile. We have 47 outliers. Despite the outliers we might assume that length in fact has an effect on the rate since for each scale of rate the range of length is significantly different.

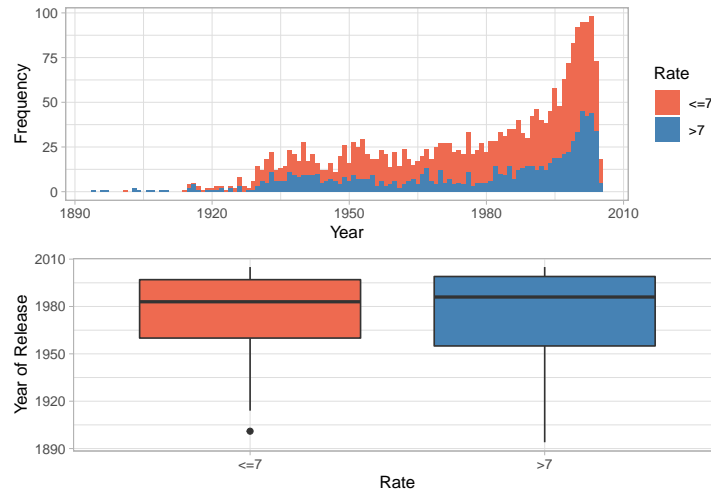


Figure 6: Rating based on Year Of Release of the film

From the second plot in Figure6, we have the rate against the year(year of release), although it seems that most recent films get rated by more people, however the difference between giving a rate over 7 and lower than 7 for different years of release did not change dramatically. That is also obvious in the next boxplot, where both boxes have nearly the same range.

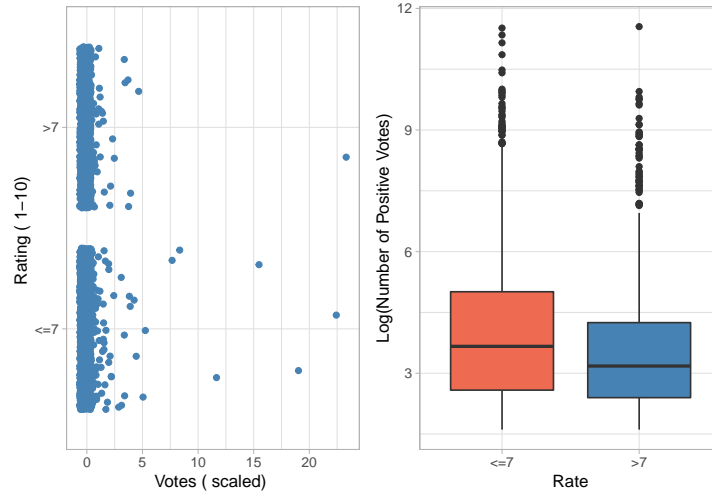


Figure 7: Rating based on Number of positive votes

Taking into account the number of positive votes again does not help identifying if a film is more probably getting a rate lower than 7 or higher than 7 .

We might now check if there is any potential structure among our explanatory variables. First we check if genres and duration have any hidden structure.

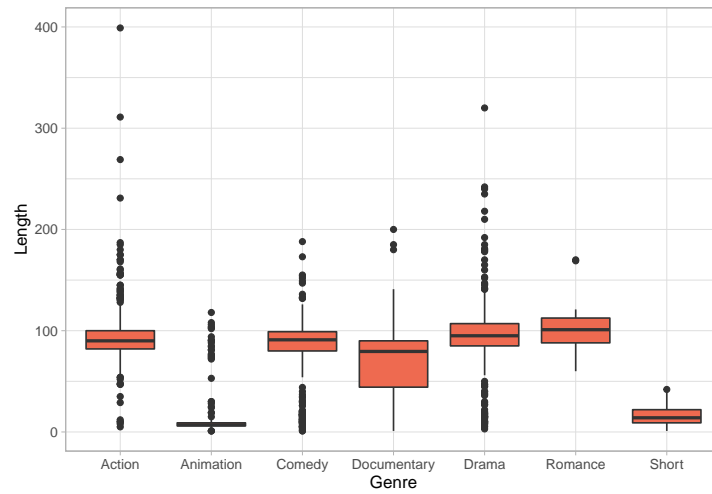


Figure 8: Genre and Length

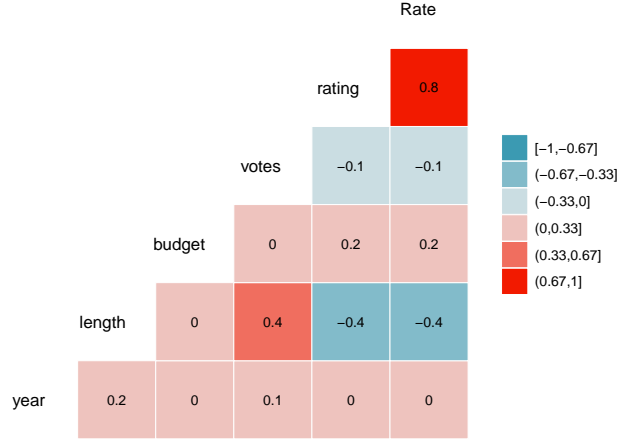


Figure 9: Genre and Length

We see that in general films' duration are roughly in same range except Documentary which has a wider range and also short and animation are genres with shorter duration. In summary we do not observe any correlation between any of our continuous explanatory variables as all have their absolute value of pairwise correlation coefficient lower than 0.4.

3 Formal Data Analysis

As we noticed from our exploratory analysis some genres can be grouped together as they have similar distributions. We regroup our genre type as follows :

- Documentary and Short films as type A
- Comedy and Animation as type B
- Romance, Drama and Action as type C

Recall we also noticed that length of the film might be a significant property that can determine the rating scale we proceed and remove observations that have missing values for length. The model we are going to fit to our data is the following.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathbb{I}_A + \beta_2 \mathbb{I}_B + \beta_3 \mathbb{I}_C + \beta_4 \cdot Length_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where

- p is the probability a film getting a rate higher than 7..
- \mathbb{I}_A indicator factor that gets value 1 if the i^{th} film is either documentary or short.
- \mathbb{I}_B indicator factor that gets value 1 if the i^{th} film is either comedy or animation.
- \mathbb{I}_C indicator factor that gets value 1 if the i^{th} film is either one of romance, drama or action.
- β_0 intercept term of the model

- β_4 is the coefficients for Length of i^{th} film.

Now to assess model fit we compare it to other models. First we compare it with the full model, which has all features as explanatory variables.

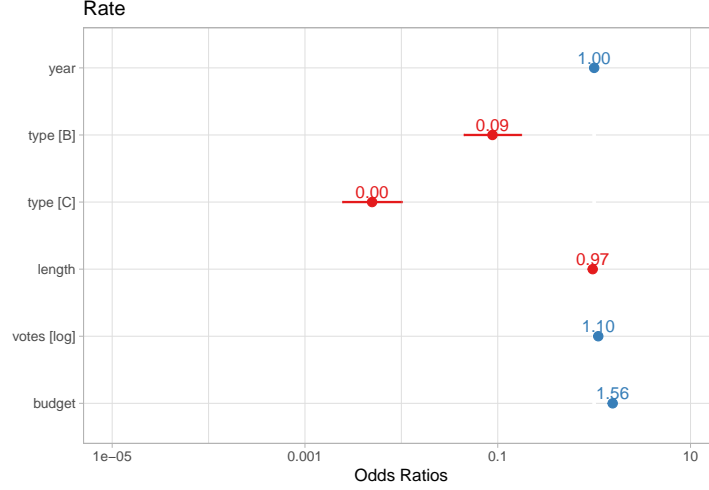


Figure 10: 95% CI for Odds ratio

The odds ratio is estimated as follows

$$\begin{aligned} \frac{\hat{p}}{1 - \hat{p}} &= \exp(x_i^T \cdot \hat{\beta}) \\ &= \exp(\hat{\beta}_0 + \hat{\beta}_1 \mathbb{I}_A + \hat{\beta}_2 \mathbb{I}_B + \hat{\beta}_3 \mathbb{I}_C + \hat{\beta}_4 \cdot \text{Length}_i) \end{aligned}$$

Given the output of the full model and plotting the estimates as shown in Figure 10 shows that the coefficient for year has an estimate of 1.0008402 with 95% confidence interval 0.9951371, 1.0066013, which contains 1 and that indicates that the variable year is not a significant predictor for the odds ratio of getting rate over 7, thus we drop it from the model. Next we apply a model with year removed. Using the deviance difference we compare it with our model.

Table 2: Defference of Deviance

Resid. Df	Resid. Dev	Df	Deviance
2291	1791.962	NA	NA
2289	1449.920	2	342.0414

Table 2 gives the difference in the deviance between the two models. We use Chi-squared asymptotic approximation to conduct a hypothesis test. Now $\chi^2(1; 0.95) = 3.841$ and compared to the difference of deviance, $3.841 < 342.041$, this indicates that there is no significant evidence to reject the null hypothesis, which states that our model is a better fit than model2.

3.1 Interpretation of model's estimates

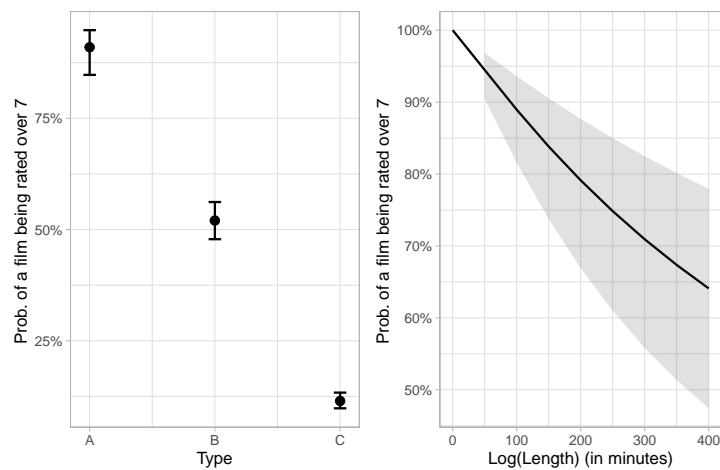


Figure 11: Predicted Probability of getting rate over 7 by Type and Length of film

3.2 Residuals

4 Conclusion and further task

5 Reference