Group_07_Analysis

Group 7

1 Introduction

As the industry of film and movie production is growing every day, understanding how viewers react towards past released films might give some approach for future content. Reactions can be measured through the overall rating a film gets from viewers. A useful online source for tackling this task is the IMDb platform, "Internet Movie Database",which as its name states it is a database on information, ratings and reviews on films, movies and many other streaming content. The dataset assigned to our group was taken from this databese which consists of records of 2387 films released from 1894 to 2005. Each film has a unique identification number (id) along with some other measurements related to it. Those are :

- $Year_i$: the year in which the film was released at cinemas
- $Length_i$: duration of film in minutes
- $Budget_i$: Budget for film production in$\$1,000,000$'s
- $Vote_i$: Total number of positive votes
- $Genre_i$: Genre of film
- $Rating_i$: IMDb rating from 1 to 10

Through this analysis we aim to find which of these properties,if any, are significant in predicting whether a film will get an IMBd rate over 7 or below 7. For this reason we first create a new binary variable that takes the value of 1 if the rating is over 7 and 0 otherwise. We begin with an exploratory analysis on our data set to detect any anomalies and to determine an appropriate class of model to fit. Next we proceed, fit the model and interpret the results given by the model and finally we evaluate our model.

2 Exploratory Data Analysis

Taking a glance at the summaries of the variables shown in Table 1 we notice that the length variable has many outliers, and that is explained 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 standard deviation, 4370 to be exact and that might be caused by fact that a very small number of films have huge number of votings compared to the rest of the films. As for the rating variable we do not observe any irregularities which is also confirmed by its density plotted in Figure 1.

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

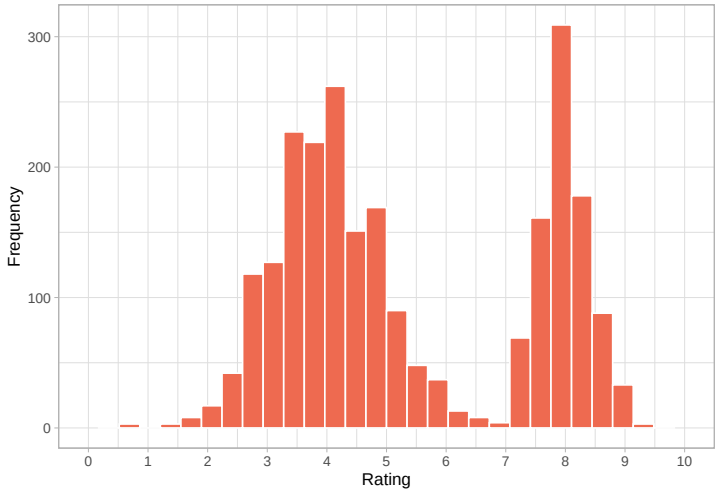

Figure 1: Density Of Ratings

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. In contrast, animation documentary and short films have relatively lower frequencies. Finally the proportion of the population of romance films within the data set seems very small. Only 16 films are romantic among the total of 2387 films.

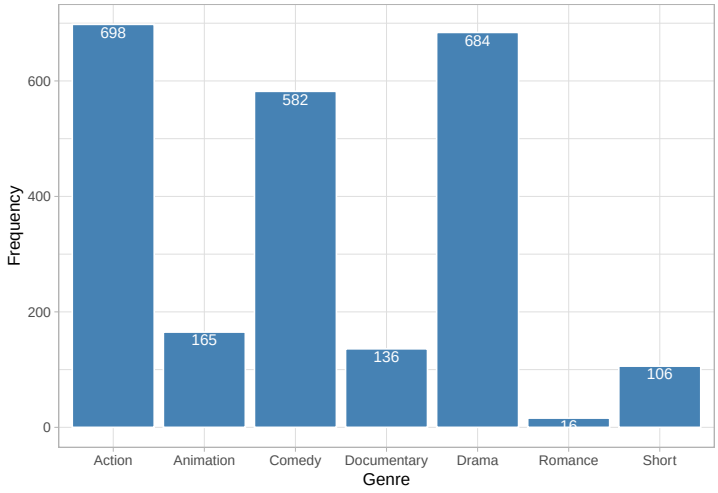

Figure 2: Frequency of each genre.

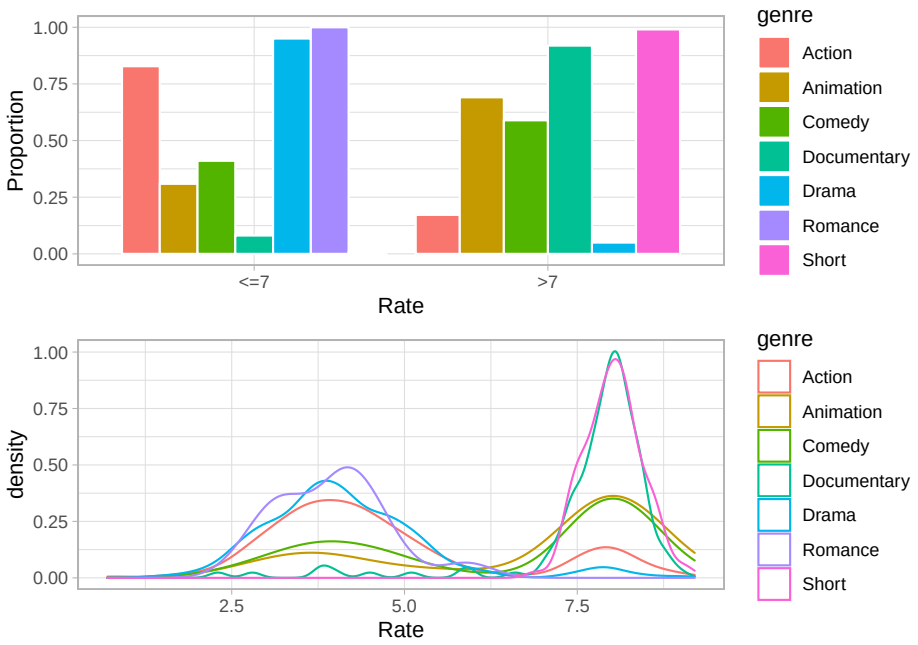

Figure 3: Distribution of genre.

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 that 7. Moving on to the next density plot of the rate for each genre we notice that some genres have very similar densities. In particular we see that documentary and short films have very similar distribution of rate. The same applies to the pair animation and comedy. As for the remaining genres, romance, action and drama there seem to be some differences in their lines but the overall shape and characteristics are very similar. These observations suggest that in fact we have 3 categories since some genres gather same reactions from the viewers and this leads 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 Figure4 although we notice that the box of rate over 7 is slightly higher than the box of rate below 7 however there is a substantial overlap between the two.

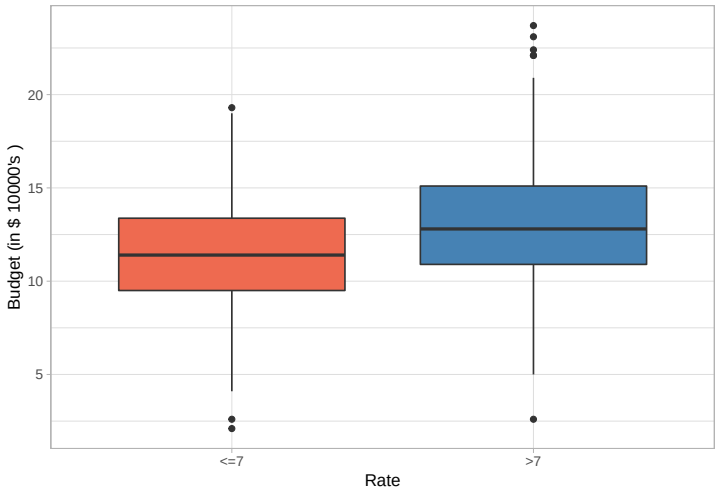

Figure 4: Budget and Rate.

Table 2: Five number summary of length variable.

Min	Q1	Mean	Q3	Max	NAs
1	72	81.41	100	399	92

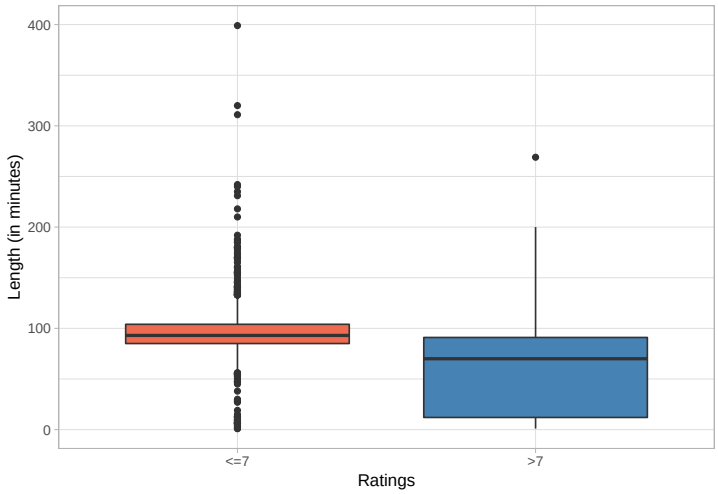

Figure 5: Rating based on length of the film.

Figure 5 shows that the length feature has many outliers for films with rate below 7. 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 3^{rd} 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.

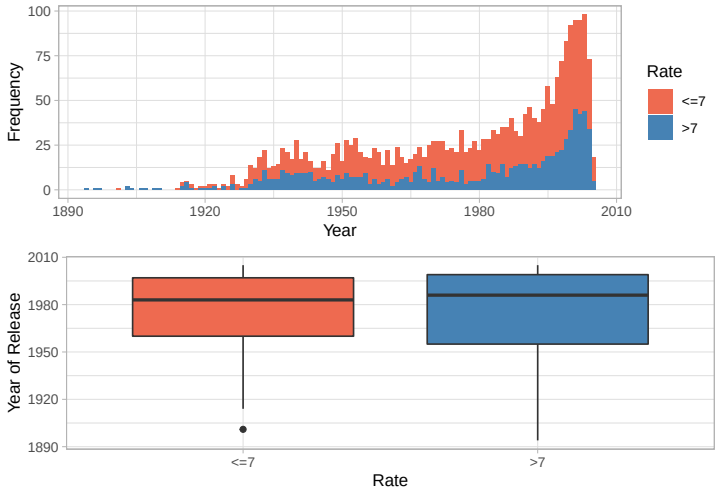

Figure 6: Rating based on Year Of Release of the film

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 confirmed in the next boxplot, where both boxes have nearly the same range and same level of the median line.

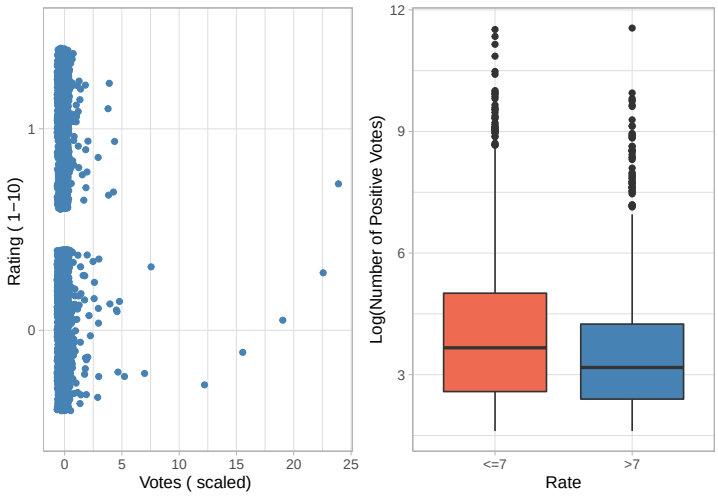

Figure 7: Rating based on Number of positive votes

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

We now check if there is any potential structure among our explanatory variables. We might want to check if genres and duration have any potential correlation.

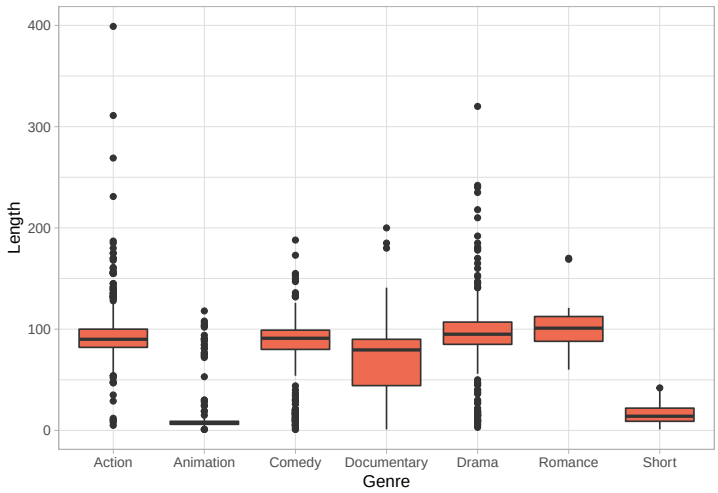

Figure 8: Genre and Length

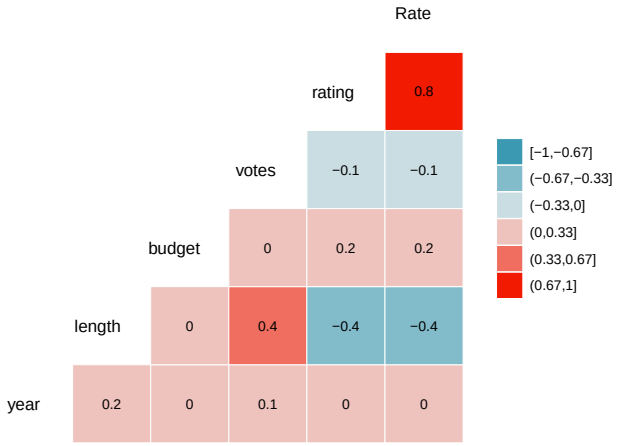

Figure 9: Genre and Length

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 of rate. We regroup our genre types 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 thus we proceed to remove observations that have missing values for length. Since the outcome variable,Rate, is binary a proper model would be a logistic model. The model we are going to fit to our data is the following.

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

where

- p is the probability of 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 coefficient of Length of i^{th} film.

Now to assess model fit we compare it to other models. First we fit the full model which has all features as explanatory variables. We check the significance of coefficients in the output and remove any non-significance variables. We then fit a reduced model with only the significance variables and we compare it to our model.

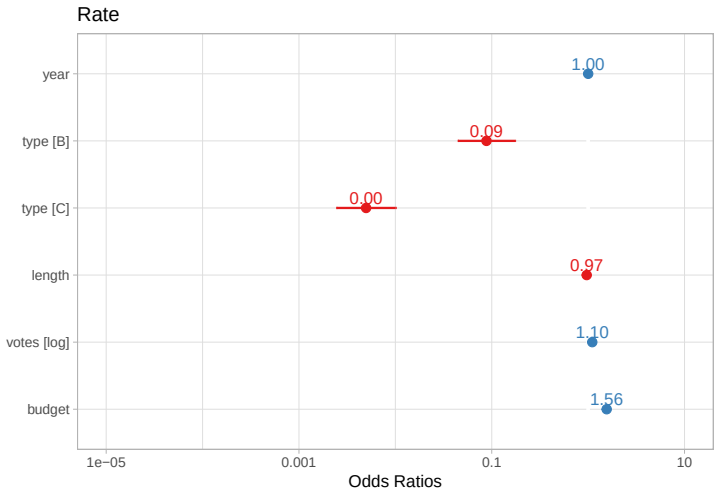

Figure 10: 95% CI for Odds ratio

The odds ratio is estimated as follows

$$\frac{\hat{p}}{1-\hat{p}} = \exp(x_i^T \cdot \hat{\beta})$$

$$= \exp(\ \hat{\beta}_0 + \hat{\beta}_1 \mathbb{I}_A + \hat{\beta}_2 \mathbb{I}_B + \hat{\beta}_3 \mathbb{I}_C + \hat{\beta}_4 \cdot Length_i\ )$$

Given the output of the full model and plotting the estimates as shown in Figure 10 we see that the coefficient for year has an estimate of 1.0008402 with 95% confidence interval 0.9951371, 1.0066013, which contains 1 and that is evidence of lack of significance as a 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 3: Defference of Deviance

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

Table 3 gives the difference in the deviance between the two models. To conduct a hypothesis test we use Chi-squared asymptotic approximation for the difference of deviances. 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

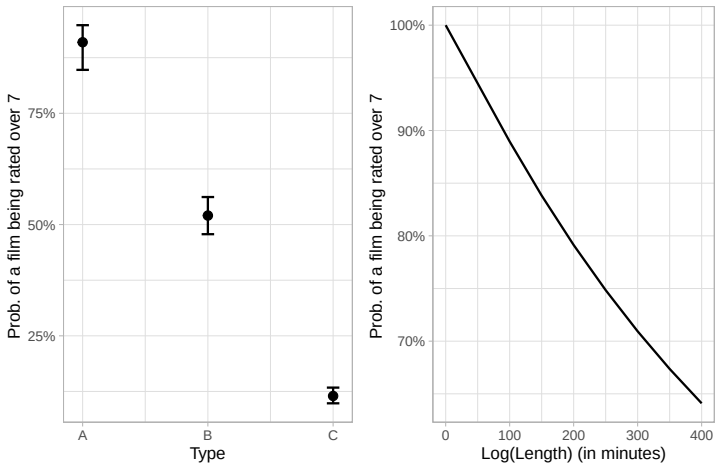

Figure 11: Predicted Probability of getting rate over 7 by Type and Length of film

Table 4: Estimates

	x
(Intercept)	7.089191
typeB	-2.228282
typeC	-4.351641
log(length)	-1.086464

Given the estimates by fitting our model we see that all estimates are negative with the magnitude of type C larger than type B suggesting that 2 films of type B have worse chance of getting rate higher than 7 and type C films have even worse chance when both compared to type A. In particular a film of type B will have log-odds lower than type A films by -2.23 while type C films will have even lower log-odds by -4.35. Moreover the log(length) coefficient also suggests that for 2 films that differ in one minute in length, the longer film's odds ratio will be -1.086 times lower than the shorter film.

We can also get estimates of the odds ratio as follows

$$\frac{\hat{p}}{1-\hat{p}} = \exp(x_i^T \cdot \hat{\beta})$$

$$= \exp(\ \ \hat{\beta}_0 + \hat{\beta}_1 \mathbb{I}_A + \hat{\beta}_2 \mathbb{I}_B + \hat{\beta}_3 \mathbb{I}_C + \hat{\beta}_4 \cdot Length_i \ \ )$$

4 Conclusion and further task

In summary we conclude that the properties which are more useful in determining whether a film will be more likely to get an overall rate over 7 or below 7, are type of film and length. By fitting our model it is suggested that films that are of type A, those are documentary and short films , have much higher probability of getting rate over 7 in comparison with the rest types of films. For future work we would consider removing some observations representing outliers such as films in the Romance category due to the analogously very

small population. A different approach could be to fit a Generalized Linear Mixed Model as the films of each genre can be thought of a sample of the population of each genre of film, in that case we would be interested in the effect of genre on rating.