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 databese, 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

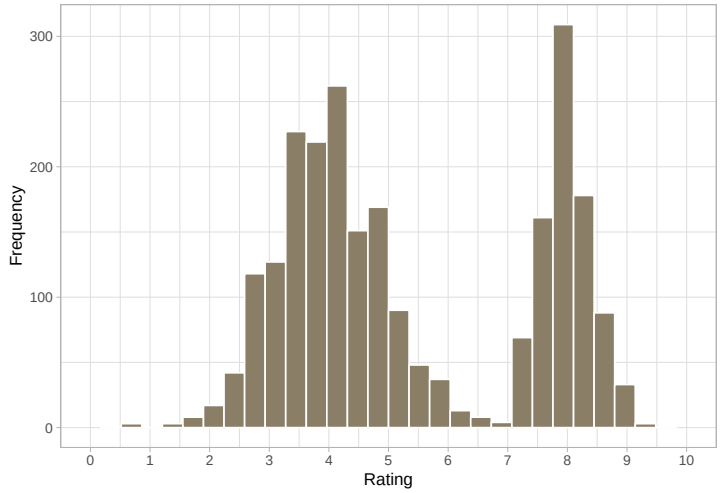

Figure 1: Density Of Ratings

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. Finaly Romance films are comparatively very rare among our list of films (only 16 films where Romantic among all 2387)

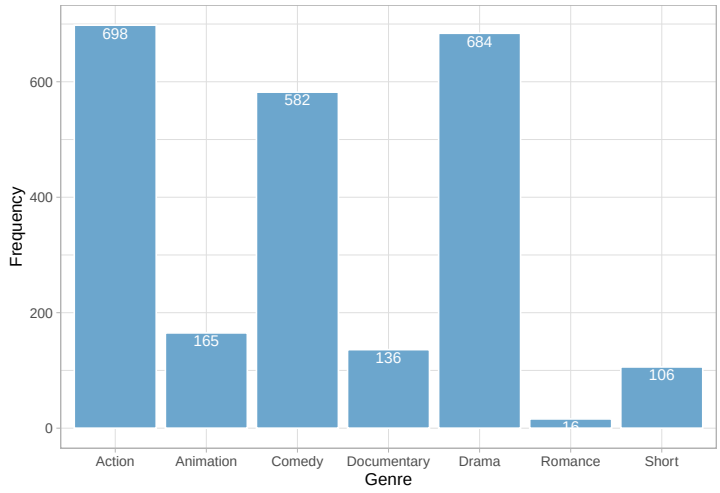

Figure 2: Frequency of each genre.

Table 2: Summary statistics of variables in the data set.

Rate	Action	Animation	Comedy	Documentary	Drama	Romance	Short
<=7	37.4% (578)	3.3% (51)	15.5% (239)	0.7% (11)	42.0% (650)	1.0% (16)	0.1% (1)
>7	14.3% (120)	13.6% (114)	40.8% (343)	14.9% (125)	4.0% (34)	0.0% (0)	12.5% (105)

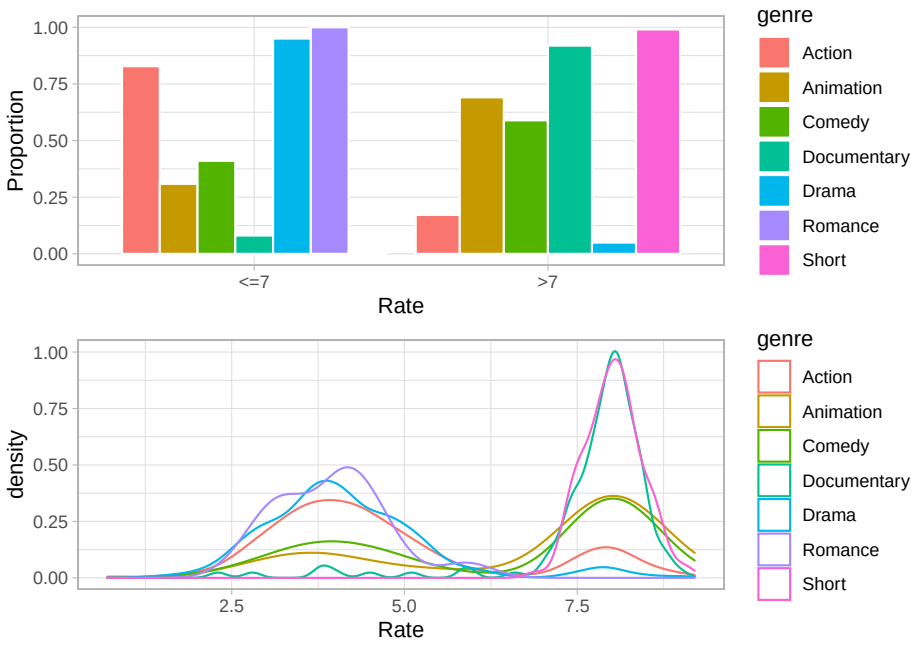

Figure 3: Distribution of genre.

From both the density graph and the barplot, we can see that Short,Documentary,Comedy and Animation films are more likely to get rates over 7.5 compared to rates given to Romance, Drama and Action films,lower than 7 (5 to be exact) Given the contingency Table 2 we can make four important notes. First, all Romance films have ratings below 7. Secondly, Animation and Short movies have a very little proportion that are rated below 7, only 3.3% and 0.1% respectively. Lastly Drama films are most liket to be rated below 7, with only 4% being rated over 7. These observations lead eliminating those genres since they do not provide useful information for the rating scale.

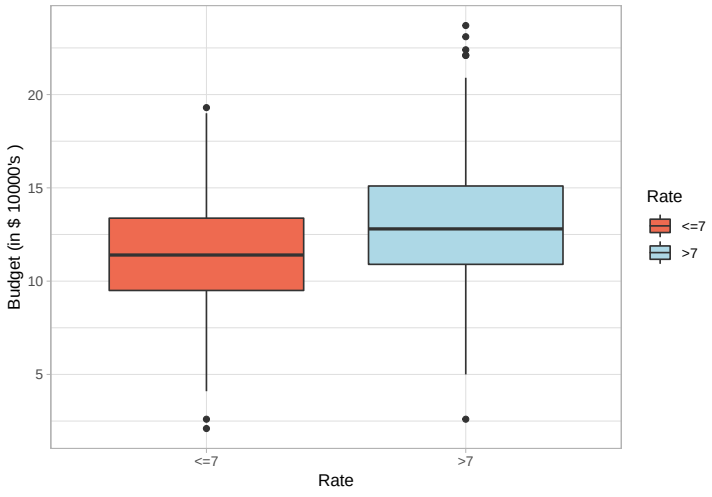

Figure 4: Budget and Rate.

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

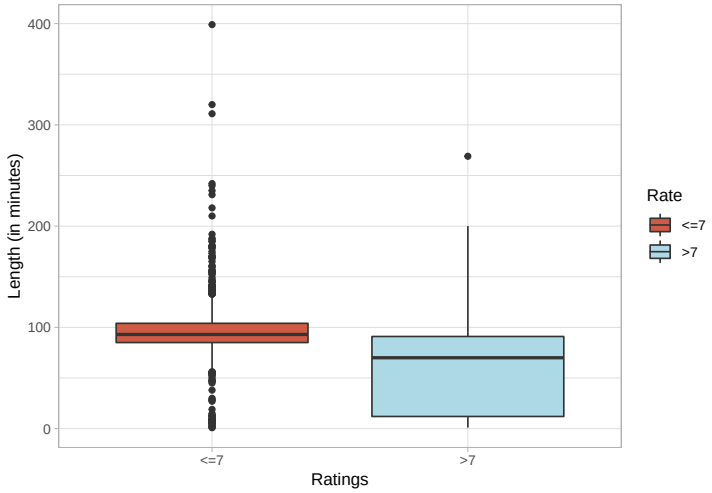

Figure 5: Rating based on length of the film.

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 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 group of rate the range of length is significantly different.

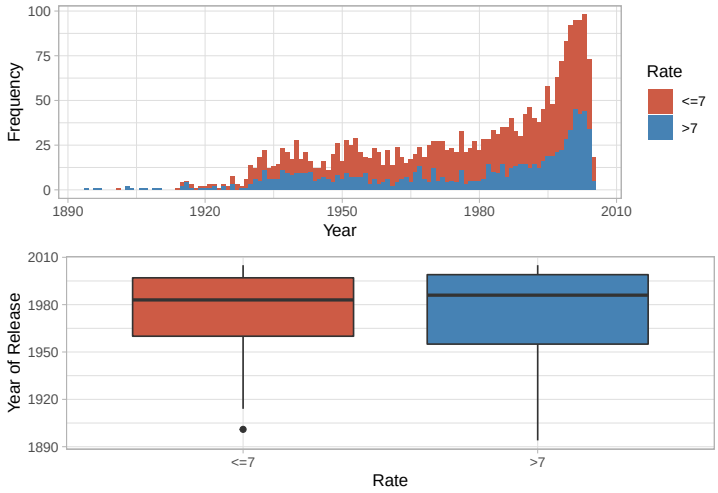

Figure 6: Rating based on Year Of Release of the film

From the second plot in Figure6, we have the rate against the year(year of release), although it seems that people tend to rate more films in general, 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 neary the same range.

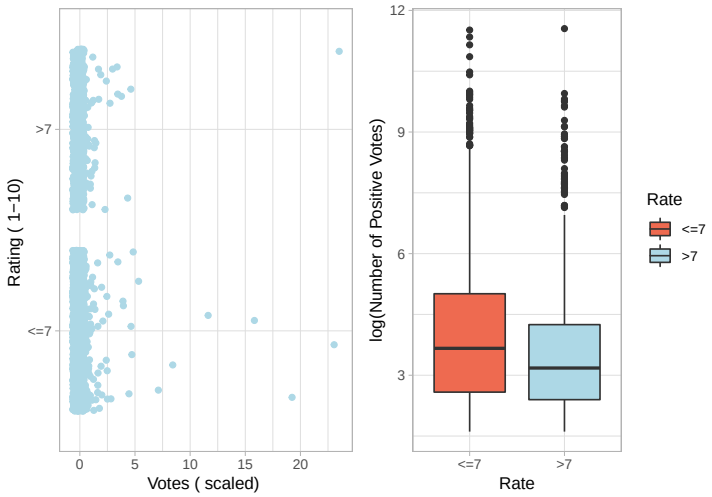

Figure 7: Rating based on Number of positive votes

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 want to check if different genres have different duraition.

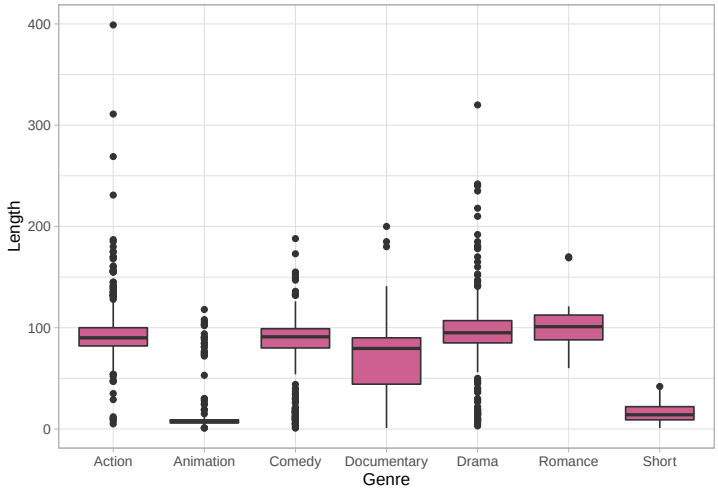

Figure 8: Genre and Length

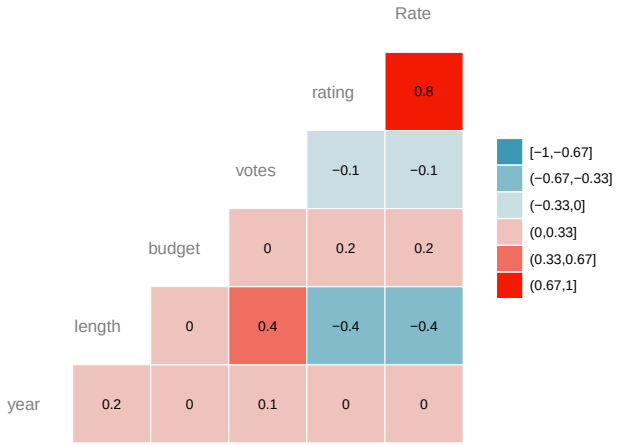

Figure 9: Genre and Length

We see that in general films' duration are roughly in same range except Documentary which has a wider range. 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 the genres Romance,Drama,Animation and Short they all do not have significant explanation thus we remove observations related to those genres. Also since we notices 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

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

where

- p is the probability a film getting a rate lower than 7.
- $\mathbb{I}_{comedy}$ indicator factor that get value 1 if the i^{th} film is comedy.
- $\mathbb{I}_{Documentary}$ indicator factor that get value 1 if the i^{th} film is documanetary.
- β_0 intercept term of the model
- β_3, β_4 are coefficients for Length and Budget of i^{th} film.

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

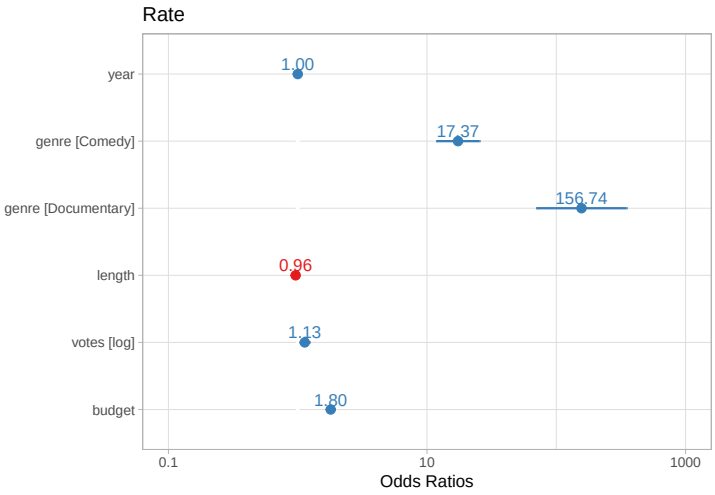

Figure 10: 95% CI for Odds ratio

The odds ratio is estimated as follows

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

$$= \exp(\hat{\beta}_0 + \hat{\beta}_1 \mathbb{I}_{comedy} + \hat{\beta}_2 \mathbb{I}_{documentary} + \hat{\beta}_3 \cdot Length_i + \hat{\beta}_4 \cdot Budget_i)$$

Figure 10 shows that the coefficient for year has an estimate of1.00065 with 95% confidence interval 0.9934032, 1.0079837, which contains 1 and that indicates that year is not a significant predictor for the odds ratio, 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
1356	982.4004	NA	NA
1355	974.4300	1	7.970443

Table 3 gives the difference in the deviance between the two models. We use Chi-squared assymptotic theorem to conduct the hypothesis test . Now $\chi^2(1; 0.95) = 3.841$ and compared to the difference of deviance, 3.841 <7.97, this indicates that there is no significant evidence to reject the null hypothesis, which states that our model is a better fit that model2.

4 Conclusion and further task