Group Project 2

Group 10: Linxuan Guo, PiuSheung TANG, Yiqing Wen, Chenghao YANG, Yanqi Zhu

1 Introduction

IMDB stands for Internet Movie Database. It is an online database which allows users to rate and review movies and TV shows. In this project, we are going to investigate the following question of interest related to IMDB: which properties of films influence whether they are rated by IMDB as greater than 7 or not. We used a generalised linear model(GLM) to fit the data in the formal analysis.

2 Data processing

Table 1 below shows the first five rows of the row data.

Table 1: The first 5 rows of the raw data

film_id	year	length	budget	votes	genre	rating
49834	1963	107	11.4	225	Romance	3.1
53923	1984	NA	9.7	59	Comedy	2.3
30020	1992	32	15.4	6	Documentary	7.7
46364	2000	NA	11.5	69	Action	2.3
19967	1964	87	9.4	34	Comedy	5.5

As there are some NA values in this dataset which will influence our result, we first omit those NA value. Then, we use the column rating to create a binary factor. And we omit the invalid column film_id and rating.

According to the dataset, we need to transform the characters type values into factor type values. Finally, we set the level of the response variable to make the graph of the data visualization more understandable.

Table 2 below displays the first five rows of the data after processed. The last column "greater_than_7" was added to each row to present whether the rating of the film is greater than 7.

Table 2: The first 5 rows of the data

	year	length	budget	votes	genre	greater_than_7
1	1963	107	11.4	225	Romance	IMDB rate less than 7
3	1992	32	15.4	6	Documentary	IMDB rate greater than 7
5	1964	87	9.4	34	Comedy	IMDB rate less than 7
6	2003	95	13.3	22	Action	IMDB rate less than 7
7	1983	96	9.6	10	Drama	IMDB rate less than 7

3 Exploratory Data Analysis

The Scatterplot matrix **figure 1** shows that most correlations between explanatory variables are really insignificant, therefore the initial model was built without censoring the variables.

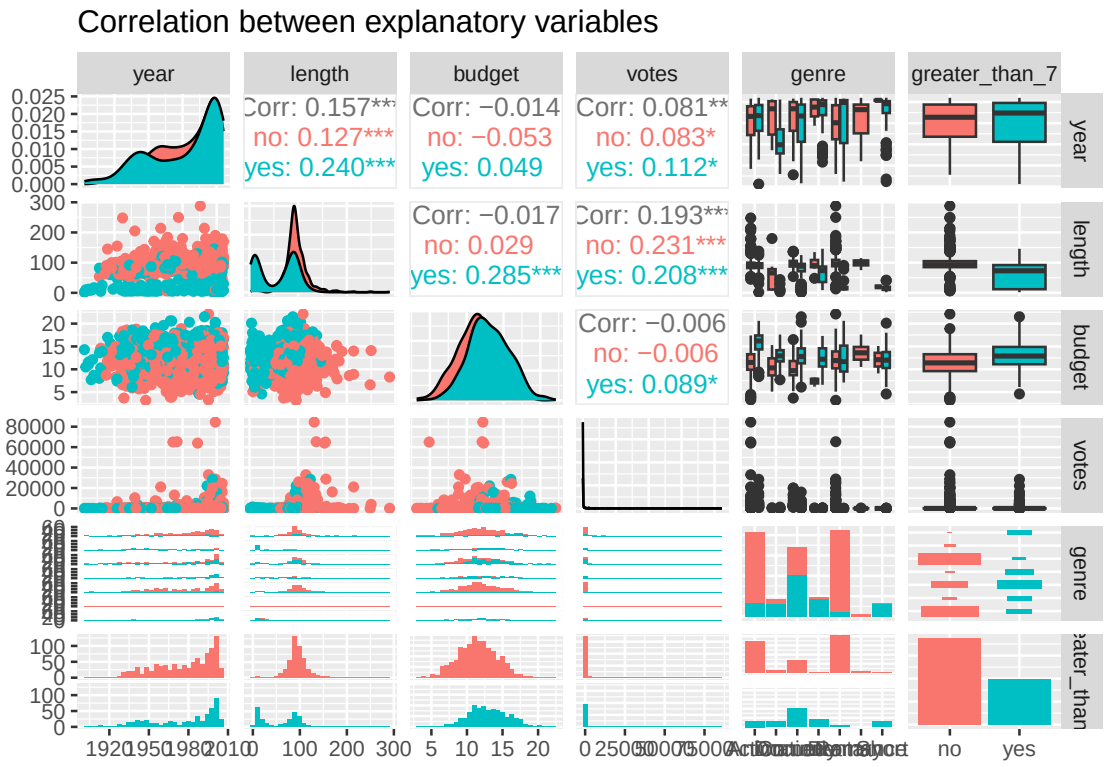

Figure 1: Correlation between explanatory variables.

Figure 2 demonstrates the relationships between the four numerical variables and the response variable Y(IMDB rate greater than 7) through multiple boxplots. These plots can be used to identify associations, trends, outliers, and non-linear relationships between the variables, thus providing valuable insights for further analysis and model development.

-There are zero potential outliers in the top left corner plot.

-There are many potential outliers in the top right corner plot.

-There are eight potential outliers in the bottom left corner plot.

-There are all potential outliers in the bottom right corner plot.

```
#Plot of year against greater_than_7
p1 = ggplot(data = DataSet, aes(x = greater_than_7, y = year, fill = greater_than_7)) +
  geom_boxplot() +
  labs(x = "IMDB rate greater than 7 or not", y = "Year",
       title = "Year and IMDB rate greater than 7 or not") +
  theme(legend.position = "none", plot.title = element_text(size = 10),
        axis.text = element_text(size = 7))
#Notice that the factor year seems not to influence rates

#Plot of length against greater_than_7
```

```
p2 = ggplot(data = DataSet, aes(x = greater_than_7, y = length, fill = greater_than_7)) +
  geom_boxplot() +
  labs(x = "IMDB rate greater than 7 or not", y = "Length",
       title = "Length and IMDB rate greater than 7 or not") +
  theme(legend.position = "none", plot.title = element_text(size = 10),
        axis.text = element_text(size = 7))

#boxplot of budget against greater_than_7
p3 = ggplot(data = DataSet, aes(x = greater_than_7, y = budget, fill = greater_than_7)) +
  geom_boxplot() +
  labs(x = "IMDB rate greater than 7 or not", y = "Budget",
       title = "Budget and IMDB rate greater than 7 or not") +
  theme(legend.position = "none", plot.title = element_text(size = 10),
        axis.text = element_text(size = 7))

#boxplot of votes against greater_than_7
p4 = ggplot(data = DataSet, aes(x = greater_than_7, y = votes, fill = greater_than_7)) +
  geom_boxplot() +
  labs(x = "IMDB rate greater than 7 or not", y = "Votes",
       title = "Votes and IMDB rate greater than 7 or not") +
  theme(legend.position = "none", plot.title = element_text(size = 10),
        axis.text = element_text(size = 7))

grid.arrange(p1,p2,p3,p4,
             top = "The Boxplot of IMDB rating greater than 7 versus each variable",
             ncol = 2,nrow=2)
```

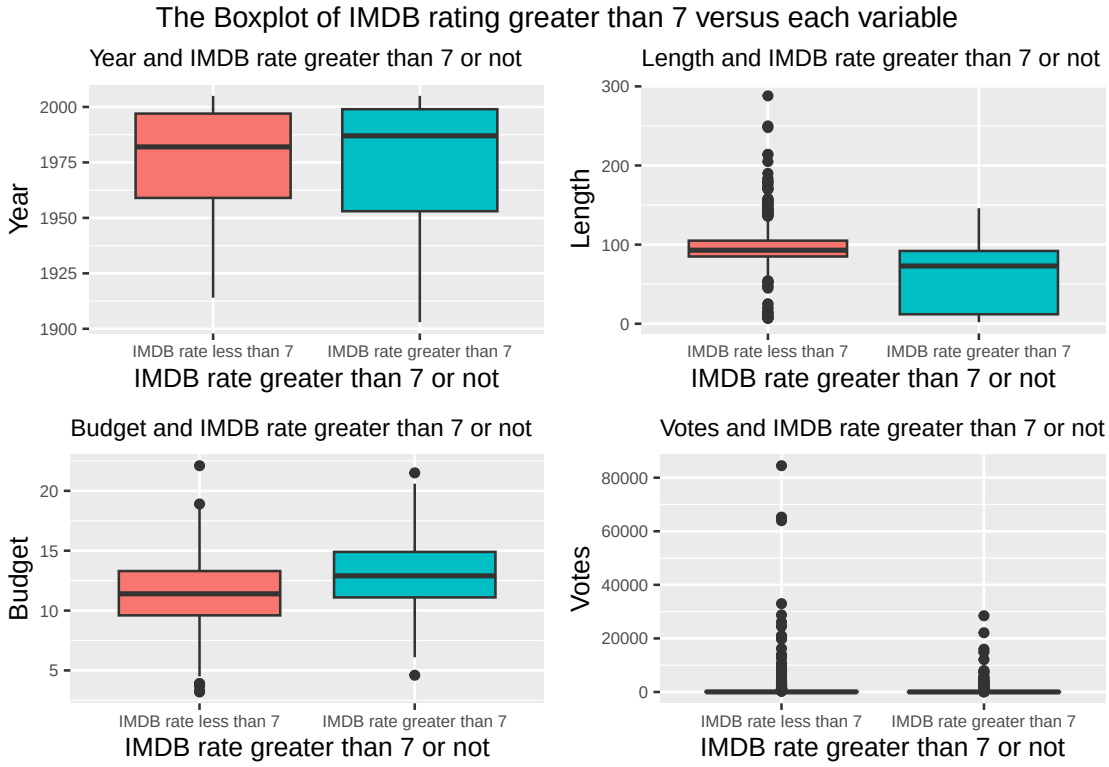

Figure 2: The Boxplot of IMDB rating greater than 7 versus each variable

The histogram(**figure 3**) shows how many IMDB rates are less than 7 and more than 7 in different movie genre, respectively.

```
#Plot of genre against greater_than_7
ggplot(data = DataSet, aes(x= greater_than_7,  y = ..prop.., group=genre, fill=genre)) +
  geom_bar(position="dodge", stat="count") +
  labs(x = "IMDB rate greater than 7 or not", y = "Proportion",
       title = "Barplot of the proportion of whether IMDB rate greater than 7 by genre")
```

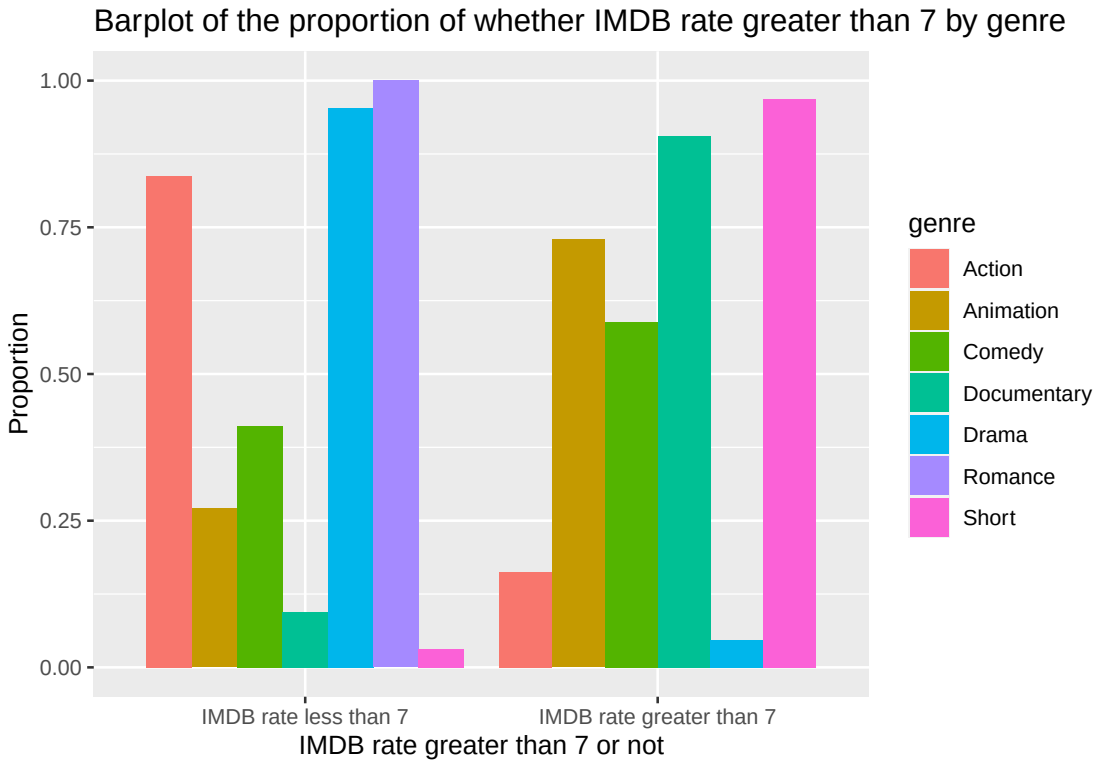

Figure 3: The Histograms of the proportion of movie categories and IMDB rate greater than 7 or not.

4 Modeling

4.1 Initial modelling

Initial modelling was carried out using the GLM function and the model results were as follows.

```
model = glm(greater_than_7 ~ year + length + budget + votes + genre, data = DataSet,
            family = binomial(link = "logit"))
model %>%
  summary()
```

```
Call:
glm(formula = greater_than_7 ~ year + length + budget + votes +
```

```
    genre, family = binomial(link = "logit"), data = DataSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6363  -0.3475  -0.0829   0.2049   3.2741

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.150e+01  8.442e+00  -1.362  0.17326
year              4.014e-03  4.306e-03   0.932  0.35123
length           -6.783e-02  5.984e-03 -11.336  < 2e-16 ***
budget            5.810e-01  4.478e-02  12.975  < 2e-16 ***
votes             5.676e-05  2.046e-05   2.774  0.00553 **
genreAnimation   -8.599e-01  5.371e-01  -1.601  0.10940
genreComedy       3.213e+00  2.599e-01  12.360  < 2e-16 ***
genreDocumentary  5.491e+00  5.468e-01  10.042  < 2e-16 ***
genreDrama       -2.141e+00  3.505e-01  -6.107 1.02e-09 ***
genreRomance     -1.541e+01  5.649e+02  -0.027  0.97824
genreShort        2.648e+00  8.840e-01   2.995  0.00274 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1854.97  on 1435  degrees of freedom
Residual deviance:  720.94  on 1425  degrees of freedom
AIC: 742.94

Number of Fisher Scoring iterations: 15
```

Hence, the fitted model for "the probability of IMDB rate greater than 7" is given by:

$$\ln\left(\frac{p}{1-p}\right) = -11.5 + 0.004 \cdot \text{year} - 0.07 \cdot \text{length} + 0.58 \cdot \text{budget} + 0.00005 \cdot \text{votes} - 0.86 \cdot \mathbf{I}_{(Animation)} + 3.21 \cdot \mathbf{I}_{(Comedy)} + 5.49 \cdot \mathbf{I}_{(Documentary)} - 2.14 \cdot \mathbf{I}_{Drama} - 15.41 \cdot \mathbf{I}_{(Romance)} + 2.65 \cdot \mathbf{I}_{(Short)}$$

where $\mathbf{p}$ is the probability of IMDB rate greater than 7, and $\mathbf{1-p}$ is the probability of IMDB rate less than 7.

From the results it is clear that:

- The slope of the Year variable is estimated to be positive, which means that, all else being equal, films released later are more likely to receive a higher rating. The parametric test, it has a high p-value, so it has a high probability of being a non-significant variable.
- The point estimate of odds for length is negative, suggesting that, all else being equal, the longer the film the more likely it is to receive a poor rating.
- A positive odds estimate for budget means that, all else being equal, the higher the budget, the more likely it is that the film will have a high rating.
- A positive odds point estimate for the votes variable suggests that, all else being equal, the higher the number of votes, the more likely the film is to receive a high rating.
- Different film categories all have different effects on film ratings; for example, when all else is equal, films in the Animation/ Drama/ Romance category are less likely to be well-received; films in the Comedy/ Documentary/ Short category are more likely to be well-recognised.

4.2 Model optimisation

Based on the initial modelling results, the YEAR variable looks more likely to be a non-significant variable, so this variable will be removed and the model will be reconstructed using the remaining variables, resulting in the following.

```
Call:
glm(formula = greater_than_7 ~ length + budget + votes + genre,
    family = binomial(link = "logit"), data = DataSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6107  -0.3395  -0.0820   0.2024   3.2722

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.653e+00  6.256e-01  -5.839 5.24e-09 ***
length           -6.661e-02  5.814e-03 -11.456  < 2e-16 ***
budget            5.796e-01  4.469e-02  12.970  < 2e-16 ***
votes             5.784e-05  2.026e-05   2.855  0.00431 **
genreAnimation   -8.336e-01  5.359e-01  -1.556  0.11979
genreComedy       3.221e+00  2.602e-01  12.379  < 2e-16 ***
genreDocumentary  5.550e+00  5.418e-01  10.243  < 2e-16 ***
genreDrama       -2.137e+00  3.486e-01  -6.131 8.76e-10 ***
genreRomance     -1.549e+01  5.587e+02  -0.028  0.97788
genreShort        2.750e+00  8.782e-01   3.132  0.00174 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1854.97  on 1435  degrees of freedom
Residual deviance:  721.81  on 1426  degrees of freedom
AIC: 741.81

Number of Fisher Scoring iterations: 15
```

Therefore, the fitted generalised linear model is:

$$\ln\left(\frac{p}{1-p}\right) = -3.65 + -0.07 \cdot \text{length} + 0.58 \cdot \text{budget} + 0.00005 \cdot \text{votes} - 0.83 \cdot \mathbf{I}_{(Animation)} + 3.22 \cdot \mathbf{I}_{(Comedy)} + 5.55 \cdot \mathbf{I}_{(Documentary)} - 2.14 \cdot \mathbf{I}_{Drama} - 15.49 \cdot \mathbf{I}_{(Romance)} + 2.75 \cdot \mathbf{I}_{(Short)}$$

5 Conclusions

After adjustment and re-modelling, we found that most of the variables were significant, but some variables in genre were still non-significant. Predicting the results of the model allows us to calculate the probability of an IMDB score greater than 7.

Assuming a probability greater than 0.5, the predicted outcome is classified as a rating greater than 7 and vice versa. This gives a prediction accuracy of 90%, which is relatively close to 1. Therefore, we consider that this model performs well in terms of explanatory power and level of prediction.

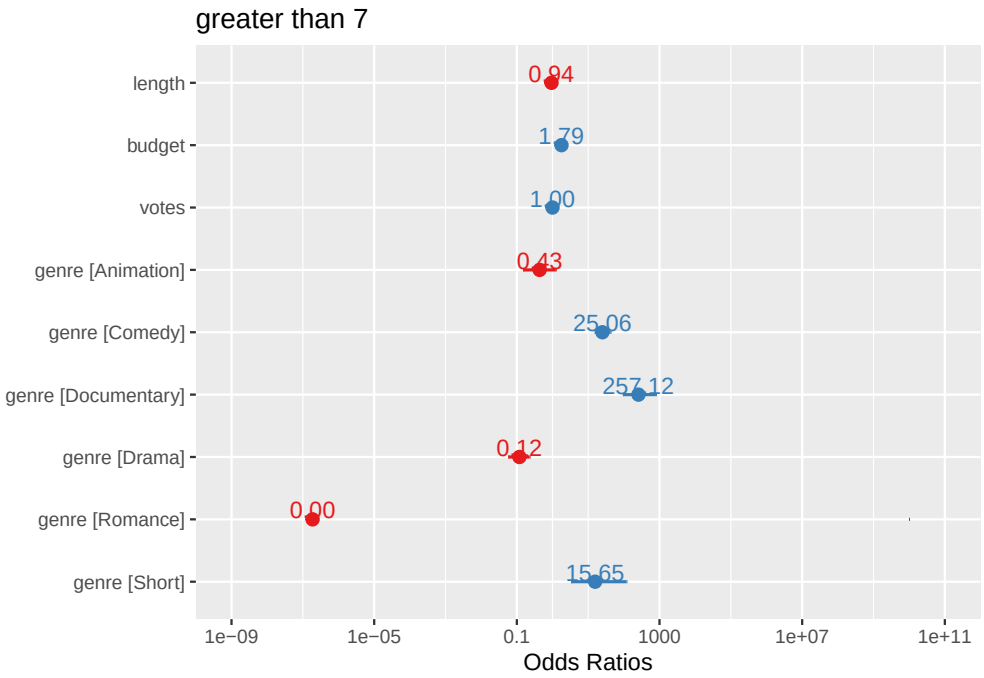

Figure 4: Odds of IMDB rate greater than 7

figure 4 above shows the odds ratios of each variable.A red number indicates that the variable it corresponds to has a negative impact on the movie rating, and conversely, the variable corresponding to a blue number has a positive impact on the movie.

The value of odds can be interpreted as follows:

- For two films which differ by one unit in length, the Longer film is 0.97 times more likely to be rated more than 7 than the shorter film.
- For two films which differ by one unit in the budget, the film on a lager budget is 1.79 times more likely to be rated more than 7 than the film on less budget.

According to all the analysis presented above and the model, we are able to know that the factors that influence whether a film rating is greater than 7 are: length, budget, votes, and genre.