Analysis of IMDB Rating with GLM

Group 10

1 Introduction

Studying the factors that can affect film ratings is an interesting topic to be explored. IMDB dataset containing information about film rating and their properties, such as length or duration, budget, votes, year of release, and genre. There are 1495 films (Film ID) included in the dataset.

In this analysis, the research question is to investigate which properties of films influence whether IMDB rating exceeding 7 or not. The binary rating (i.e., 1 if greater than 7 and 0 otherwise) will be the response variables, and the film properties will be the explanatory variables. The GLM (Generalized Linear Model) for binary response variables, Logistic Regression, will be used to investigate the relationship between binary rating and film properties.

```
library(tidyverse)
library(gt)

# Read CSV from data dir
df <- read_csv("../data/dataset10.csv")

# Display the first 5 rows
df |>
  slice_head(n=5) |>
  gt() |>
  cols_label(
    film_id = html("Film ID"),
    year = html("Year"),
    length = html("Length"),
    budget = html("Budget"),
    votes = html("Votes"),
    genre = html("Genre"),
    rating = html("Rating")
  )
```

Table 1: First five entries of the IMDB Dataset.

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

Description:

- Film ID: The unique identifier for the film.
- Year: Year of release of the film in cinemas.
- Length: Duration (in minutes).
- Budget: Budget for the films production (in $1000000).
- Votes: Number of positive votes received by viewers.
- Genre: Genre of the film.
- Rating: IMDB rating from 0-10.

2 Data Wrangling

```
# Import Libraries

library(skimr)
library(knitr)
library(corrplot)
library(ggplot2)
library(gridExtra)
library(dplyr)
library(stats)
library(jtools)
library(sjPlot)
library(broom)
library(huxtable)
library(lmtest)
library(zoo)
```

```
# Calculate the number of NAs for each column

# Checking NAs
na_sum <- colSums(is.na(df))

# Plotting number of NAs
bp <- barplot(na_sum,
              main = "Missing Values Count",
              ylab = "Count",
              col = "skyblue",
              names.arg = colnames(df),
              ylim = c(0, max(na_sum) + 180), # space for labels
```

```
              las=1) # rotating x-axis labels
text(x = bp,
     y = na_sum + 2,
     labels = na_sum,
     pos = 3,
     col = "red")
```

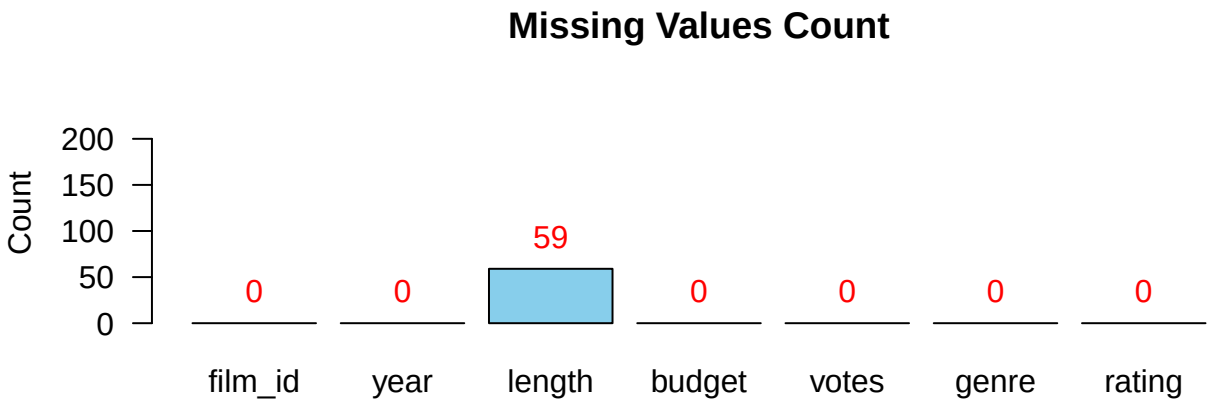

Figure 1: The number of NAs for each column in the dataset.

In the column length, there are 59 from 1945 (10.03%) rows containing NA values. Moreover, they will be removed as the proportion is pretty small. Another reason is to avoid imputing inaccurate information relative to other explanatory variables which might give impact to the statistical analysis result and conclusion.

2.1 Preprocessing Steps

```
# Create new columns: rating_higher_than_7
df <- df %>%
  mutate(rating_higher_than_7 = ifelse(rating <= 7, 0, 1))

# Remove NAs
df <- na.omit(df)
```

The data preprocessing is performed to create new columns to categorize the rating is higher than 7 or not. If yes, it will be marked as 1, and 0 otherwise. Next, the rows which have NAs are being removed from the analysis. Later on, rating_higher_than_7 is going to be the response variable for the following Logistic Regression (GLM) analysis.

Table 2: Summary statistics of the IMDB Dataset.

(a) Data summary

Name	IMDB
Number of rows	1436
Number of columns	8
Column type frequency:	
character	3
numeric	5
Group variables	None

Variable type: character

skim_variable	n	min	max	empty	n_unique	whitespace
film_id	1436	2	5	0	1436	0
genre	1436	5	11	0	7	0
rating_higher_than_7	1436	1	1	0	2	0

Variable type: numeric

skim_variable	n	mean	sd	p25	p50	p75
year	1436	1976.72	23.80	1958.00	1984.0	1998.00
length	1436	82.29	35.67	74.75	90.0	100.00
budget	1436	12.00	2.89	10.10	11.9	13.90
votes	1436	788.62	4501.33	12.00	33.0	122.25
rating	1436	5.36	2.09	3.70	4.6	7.80

3 Exploratory Data Analysis

3.1 Statistics Descriptive

```
# Creating summary statistics

# Convert film_id and rating_higher_than_7 as categorical
IMDB <- df
IMDB <- IMDB %>%
  mutate(film_id = as.character(film_id),
         rating_higher_than_7 = as.character(rating_higher_than_7))

# Summary statistics with adjusted skim()
my_skim <- skim_with(base = sfl(n = length),
                     numeric = sfl(p0 = NULL, p100 = NULL,hist = NULL))
knit_print(my_skim(IMDB ))
```

Based on the summary tables Table 2, there is no duplication for the Film ID, and it means each observation is already unique. Then, the categorical explanatory variables, genre, has seven unique values. Furthermore, the votes has a very wide values by looking at the standard deviation, mean, and median. Year and length are slightly skewed to the left, and then budget and rating are slightly skewed to the right. It can be seen by comparing mean and median position.

3.2 Correlation

```
# Calculate the correlation coefficient between numeric variables

# Filter out non-numeric columns
numeric_df <- df[sapply(df, is.numeric)]
numeric_df <- numeric_df[, !names(numeric_df) %in% c("film_id",
                                                     "rating_higher_than_7")]
# Compute correlation matrix
correlation_matrix <- cor(numeric_df, use = "complete.obs")
# Creating correlation heatmap
corrplot(cor(numeric_df), method = "color",
         type = "lower", addCoef.col = 'grey')
```

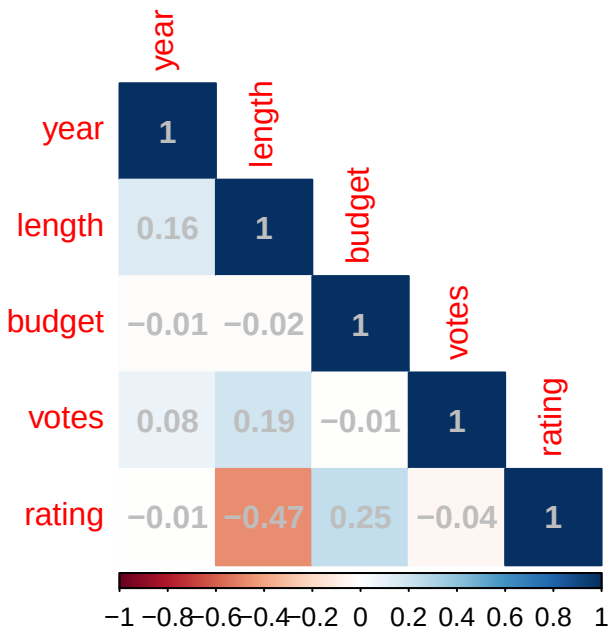

Figure 2: The correlation between numerical variables.

From the Figure 2, it reveals that rating has a weak negative correlation (-0.47) to length, and has a weak positive correlation (0.25) to budget. Moreover, year and votes show a very weak negative correlation to the rating, -0.01 and -0.04 respectively. It means there is no film properties that can give strong signal (linearly) to the rating. Further investigation will be performed visually using Figure 3.

3.3 Scatterplot (Continuous Relationship)

```
# Creating scatterplot between rating and explanatory variables

# Custom color palette
custom_colors <- c("1" = "lightskyblue", "0" = "dodgerblue4")

# Scatterplot Rating vs. Year with some adjustments
p1 <- ggplot(df, aes(x = year, y = rating,
                     color = factor(rating_higher_than_7))) +
  geom_point() +
  geom_hline(yintercept = 7, linetype = "dashed", color = "red") +
  geom_smooth(method = "lm", se = FALSE, color = "Orange") +
  scale_color_manual(values = custom_colors) +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Rating vs Year", x = "Year", y = "Rating")

# Scatterplot Rating vs. Length with some adjustments
p2 <- ggplot(df, aes(x = length, y = rating,
                     color = factor(rating_higher_than_7))) +
  geom_point() +
  geom_hline(yintercept = 7, linetype = "dashed", color = "red") +
  geom_smooth(method = "lm", se = FALSE, color = "Orange") +
  scale_color_manual(values = custom_colors) +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Rating vs Length", x = "Length", y = "Rating")

# Scatterplot Rating vs. Budget with some adjustments
p3 <- ggplot(df, aes(x = budget, y = rating,
                     color = factor(rating_higher_than_7))) +
  geom_point() +
  geom_hline(yintercept = 7, linetype = "dashed", color = "red") +
  geom_smooth(method = "lm", se = FALSE, color = "Orange") +
  scale_color_manual(values = custom_colors) +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Rating vs Budget", x = "Budget", y = "Rating")

# Scatterplot Rating vs. Votes with some adjustments
p4 <- ggplot(df, aes(x = votes, y = rating,
                     color = factor(rating_higher_than_7))) +
  geom_point() +
  geom_hline(yintercept = 7, linetype = "dashed", color = "red") +
  geom_smooth(method = "lm", se = FALSE, color = "Orange") +
```

```
  scale_color_manual(values = custom_colors) +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Rating vs Votes", x = "Votes", y = "Rating")

# Plot in Grid
grid.arrange(p1, p2, p3, p4, ncol=2)
```

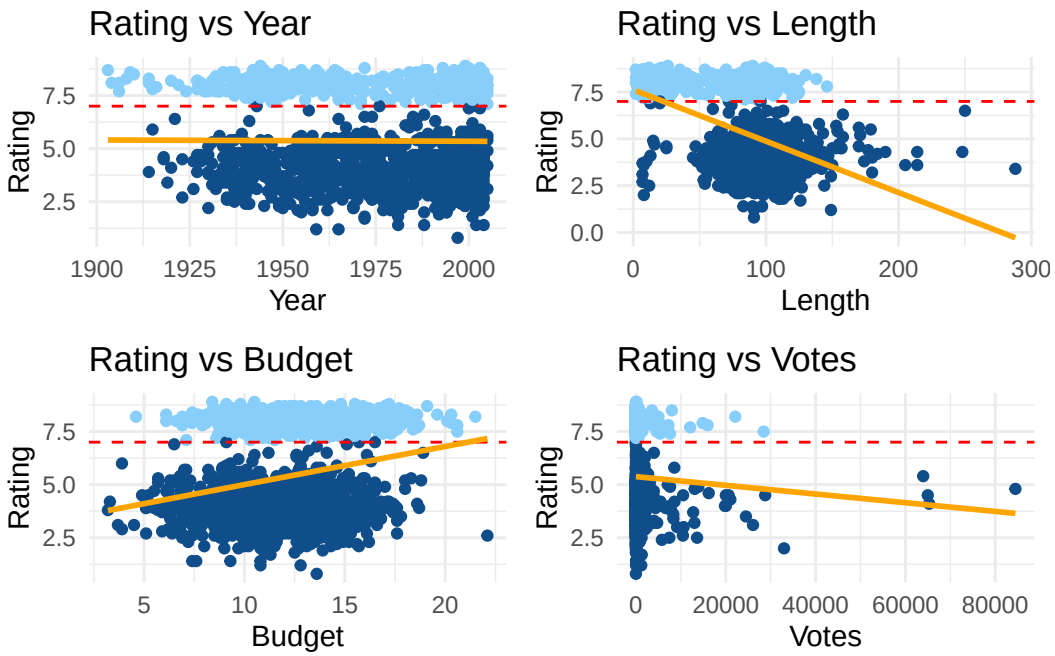

Figure 3: The relationship between rating and continuous explanatory variables.

The red line is the indicator of rating equal to 7. Based on Figure 3, most of explanatory variables' values overlap to each other between rating greater and lower than 7. Hypothetically, it might affect the logistic regression performance. However, weak linear relationship to the rating is still noticeable for Budget and Length. Then, the points for year and rating scatterplot are very scattered, indicating very weak relationship between them. Furthermore, there are some outliers for votes showing films which have very high votes (greater than 60k), and interestingly they have ratings lower than 7. These are the Film IDs:

Table 5: Films having votes greater than 60000.

film_id	votes	rating
1679	63961	5.4
48909	84488	4.8
282	64982	4.5
10414	65283	4.1

3.4 Boxplot and Barplot (Categorical Relationship)

```
# Creating boxplot and barplot between rating_higher_than_7 and explanatory variables

# Custom color palette
custom_colors <- c("1" = "lightskyblue", "0" = "dodgerblue4")

# Boxplot Rating > 7 vs. Year with some adjustments
p1 <- ggplot(data = df,
             mapping = aes(x = factor(rating_higher_than_7),
                           y = year,
                           fill = factor(rating_higher_than_7))) +
  geom_boxplot() +
  labs(y = "Year", x = "Rating > 7") +
  scale_fill_manual(values = custom_colors) +
  theme(legend.position = "none") # remove legend

# Boxplot Rating > 7 vs. Length with some adjustments
p2 <- ggplot(data = df,
             mapping = aes(x = factor(rating_higher_than_7),
                           y = length,
                           fill = factor(rating_higher_than_7))) +
  geom_boxplot() +
  labs(y = "Length", x = "Rating > 7") +
  scale_fill_manual(values = custom_colors) +
  theme(legend.position = "none") # remove legend

# Boxplot Rating > 7 vs. Budget with some adjustments
p3 <- ggplot(data = df,
             mapping = aes(x = factor(rating_higher_than_7),
                           y = budget,
                           fill = factor(rating_higher_than_7))) +
  geom_boxplot() +
  labs(y = "Budget", x = "Rating > 7") +
  scale_fill_manual(values = custom_colors) +
  theme(legend.position = "none") # remove legend

# Boxplot Rating > 7 vs. Votes with some adjustments
p4 <- ggplot(data = df,
             mapping = aes(x = factor(rating_higher_than_7),
                           y = votes,
                           fill = factor(rating_higher_than_7))) +
  geom_boxplot() +
  labs(y = "Votes", x = "Rating > 7") +
  scale_fill_manual(values = custom_colors) +
  theme(legend.position = "none") # remove legend
```

```
# Barplot Rating > 7 vs. Genre with some adjustments
p6 <- ggplot(df, aes(x = genre,  y = ..prop..,
                     group =factor(rating_higher_than_7),
                     fill = factor(rating_higher_than_7))) +
  geom_bar(position="dodge", stat="count") +
  labs(y = "Proportion", fill = "Rating > 7") +
  scale_fill_manual(values = custom_colors) +
  theme_minimal()

# Plot in Grid
grid.arrange(arrangeGrob(p1, p2, p3, p4, ncol=2),
             p6, nrow=2, heights = c(2, 1))
```

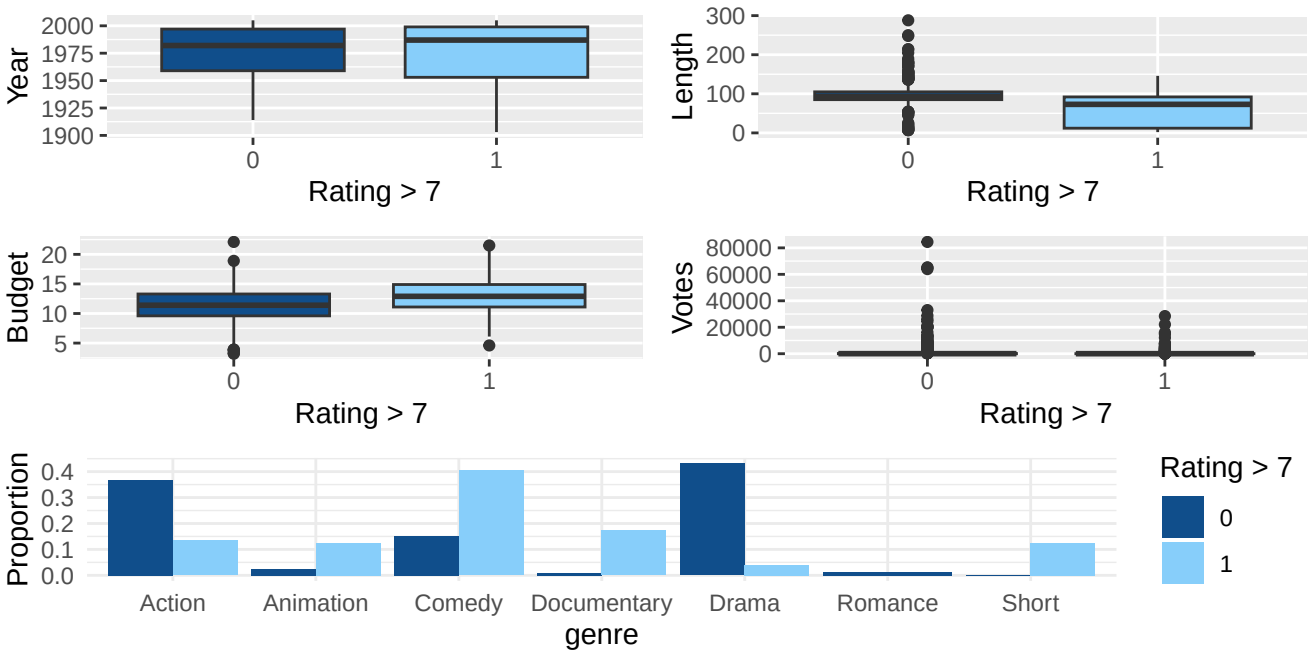

Figure 4: The relationship between binary rating and explanatory variables.

From the boxplot on the Figure 4, it is more clear to see that Budget and Length is more helpful to distinguish the rating will be higher or lower than 7. This can be seen by comparing their median lines inside the box. Moreover, there are a lot of points detected as outliers for length and votes.

3.5 Barplot (Genre vs. Rating)

```
# Calculate proportions and average of ratings for each genre categories
ratings_genre <- df %>%
  group_by(genre) %>%
  summarize(
    proportion_higher_than_7 = round(mean(rating_higher_than_7, na.rm = TRUE),3),
```

```
    average_rating = round(mean(rating, na.rm = TRUE),2)
    ) %>%
  ungroup() %>% # Grouping is removed so it can be sorted
  arrange(desc(average_rating))

# Create barplot
ggplot(ratings_genre, aes(x = genre)) +
  geom_bar(aes(y = proportion_higher_than_7),
           stat = "identity", fill = "skyblue", width = 0.5) +
  geom_point(aes(y = average_rating),
             stat="identity",color="red",size=2)+
  labs(x = "Genre", y = "Proportions of Rating > 7",
       title = "Rating vs. Genre") +
  scale_y_continuous(sec.axis=sec_axis(~.,name="Average of Rating"),
                     limits = c(0,9)) + # creating second y-axis
  geom_text(aes(y = proportion_higher_than_7,
                label = round(proportion_higher_than_7, 2)),
            vjust = -0.5, color = "black", size = 3.5) +
  geom_text(aes(y = average_rating,label = average_rating),
            vjust = -0.5, color = "black", size = 3.5)
```

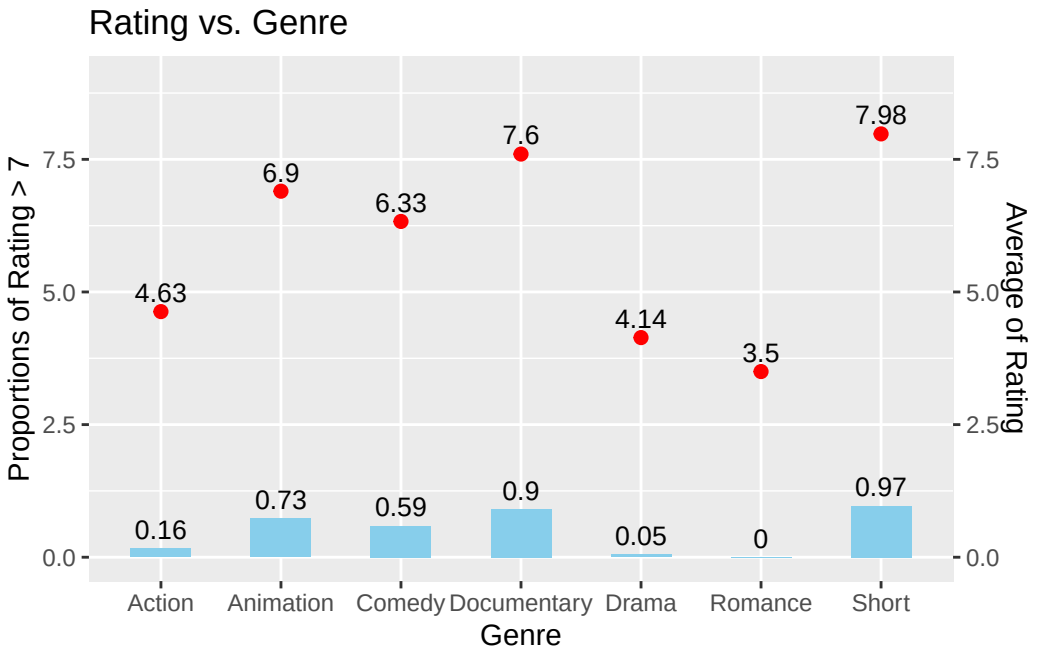

Figure 5: The proportion and average of rating for each genre.

Figure 5 shows that Action (16%) and Drama (5%) have a low proportion of having rating greater than 7, and their average of ratings are 4.63 and 4.14 respectively. The Short genre becomes the highest for both proportion (97%) and average (7.98) of rating. Visually, genre is helpful to distinguish whether the rating will be high or low.

4 Statistical Analysis (GLM)

4.1 Model Fitting and Selection

```
# Comparing models with different explanatory variables

model <- glm(rating_higher_than_7 ~ length + budget + genre,
             data = df,
             family = binomial(link = "logit"))
model_2 <- glm(rating_higher_than_7 ~ year + length + budget + votes + genre,
               data = df,
               family = binomial(link = "logit"))

suppressWarnings(export_summs(model, model_2,
                 error_format = "[{conf.low}, {conf.high}]"))
```

Following a comprehensive investigations in Table 6, it is observed that the model incorporating the variables of year and votes (Model 2) has a slightly lower Akaike Information Criterion (AIC), but a noticeable increase in the Bayesian Information Criterion (BIC). Moreover, the coefficient associated with the year variable is not statistically and practically significant by looking at P-value higher than 5% Alpha and 95% CI contains 0. Similarly, the votes coefficient is not practically significant because of having 95% CI is very close to 0. Additionally, by looking at Figure 2 reveals a very weak correlation from the year and votes to the rating variables. Hence, Model 1 is selected to be the better model. Removing year and votes resulting in a minor increment in AIC but significantly reduce the BIC. In this case, BIC favors simpler model.

Furthermore, one caveat of using this model is genreAnimation and genreRomance are not statistically significant. However, if genre is being removed, AIC will increase drastically to 1298. Please refer to the Section 6. For now, it is reasonable to keep genre as one of the explanatory variables. Goodness of fit test will be performed in the next step.

Table 6: Hypothesis Testing and Goodness of fit. The one on the left-hand side is the model without the year and votes explanatory variables.

	Model 1	Model 2
(Intercept)	-3.74 ***	-11.50
	[-4.96, -2.52]	[-28.04, 5.05]
length	-0.06 ***	-0.07 ***
	[-0.08, -0.05]	[-0.08, -0.06]
budget	0.57 ***	0.58 ***
	[0.49, 0.66]	[0.49, 0.67]
genreAnimation	-0.72	-0.86
	[-1.75, 0.32]	[-1.91, 0.19]
genreComedy	3.19 ***	3.21 ***
	[2.68, 3.69]	[2.70, 3.72]
genreDocumentary	5.46 ***	5.49 ***
	[4.41, 6.51]	[4.42, 6.56]
genreDrama	-2.11 ***	-2.14 ***
	[-2.79, -1.44]	[-2.83, -1.45]
genreRomance	-15.52	-15.41
	[-1115.47, 1084.43]	[-1122.52, 1091.70]
genreShort	2.82 **	2.65 **
	[1.11, 4.54]	[0.92, 4.38]
year		0.00
		[-0.00, 0.01]
votes		0.00 **
		[0.00, 0.00]
N	1436	1436
AIC	744.85	742.94
BIC	792.28	800.90
Pseudo R2	0.75	0.75

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

4.2 Examination of the Final Model

4.2.1 Hypothesis Testing and Goodness of fit

```
# Display the model summary (log-odds)
model %>%
  summ()
```

Table 7: Hypothesis Testing and Goodness of fit.

Observations	1436
Dependent variable	rating_higher_than_7
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(8)$	1128.12
Pseudo-R^2 (Cragg-Uhler)	0.75
Pseudo-R^2 (McFadden)	0.61
AIC	744.85
BIC	792.28

	Est.	S.E.	z val.	p
(Intercept)	-3.74	0.62	-6.01	0.00
length	-0.06	0.01	-11.40	0.00
budget	0.57	0.04	12.97	0.00
genreAnimation	-0.72	0.53	-1.36	0.17
genreComedy	3.19	0.26	12.34	0.00
genreDocumentary	5.46	0.53	10.20	0.00
genreDrama	-2.11	0.35	-6.12	0.00
genreRomance	-15.52	561.21	-0.03	0.98
genreShort	2.82	0.87	3.23	0.00

Standard errors: MLE

```r
# Display CI for the log-odds
confint(model) %>%
  kable()
```

Table 8: Confidence Intervals for Hypothesis Testing.

	2.5 %	97.5 %
(Intercept)	-4.9858324	-2.5416678
length	-0.0756187	-0.0535324
budget	0.4906868	0.6645180
genreAnimation	-1.7608457	0.3114839
genreComedy	2.6955471	3.7090221
genreDocumentary	4.4661592	6.5702495
genreDrama	-2.8217494	-1.4629401
genreRomance	NA	23.1715368
genreShort	1.2849315	4.8429135

The Logistic Regression formula can be written as:

$$log(\frac{P_i}{1-P_i}) = \alpha + \beta_{budget} \cdot budget_i + \beta_{length} \cdot length_i + \beta_{\text{Animation}} \cdot \mathbb{I}_{\text{Animation}}(i) + \beta_{\text{Comedy}} \cdot \mathbb{I}_{\text{Comedy}}(i) +$$

$$\beta_{\text{Documentary}} \cdot \mathbb{I}_{\text{Documentary}}(i) + \beta_{\text{Drama}} \cdot \mathbb{I}_{\text{Drama}}(i) + \beta_{\text{Romance}} \cdot \mathbb{I}_{\text{Romance}}(i) + \beta_{\text{Short}} \cdot \mathbb{I}_{\text{Short}}(i)$$

where $\beta_{\text{Genre}} \cdot \mathbb{I}_{\text{Genre}}$ is an indicator function such that

$$\mathbb{I}_{\text{Genre}}(x) = \begin{cases} 1 & \text{xth observation is part of the genre,} \\ 0 & \text{Otherwise.} \end{cases}$$

Based on the result Table 7, the log-odds coefficients associated with the budget variable is positive. This suggests that as the budget values increase, the likelihood of the rating greater than 7 also increases. Furthermore, the log-odds coefficients for length is negative, and it indicates that the lower the values, the rating to be higher than 7 is more likely. Table 8 shows that these two coefficients are statistically significant because the P-values are lower than 0.05 and practically significant because Confidence Intervals (CI) do not contain 0. There are two coefficients which are not statistically significant, genreAnimation and genreRomance.

```r
# Hosmer-Lemeshow goodness of fit test

source(url("http://www.chrisbilder.com/categorical/Chapter5/AllGOFTests.R"))

HLTest(model, g=6)
```

```
Warning in HLTest(model, g = 6): Some expected counts are less than 5. Use
smaller number of groups

    Hosmer and Lemeshow goodness-of-fit test with 6 bins

data:  model
X2 = 7.6182, df = 4, p-value = 0.1066
```

Next, Hosmer-Lemeshow goodness of fit test is performed. Large P-values indicates no evidence for lack of fit. However, some expected counts are less than 5, so the test might not be very reliable. As there is no evidence of lack of fit, explanatory variables are worth to be retained. Moreover, checking residuals for model diagnostic is not informative for binary response variable.

4.2.2 Odds-ratios

```r
# Plot Odds Ratios for each parameters
plot_model(model, show.values=TRUE)
```

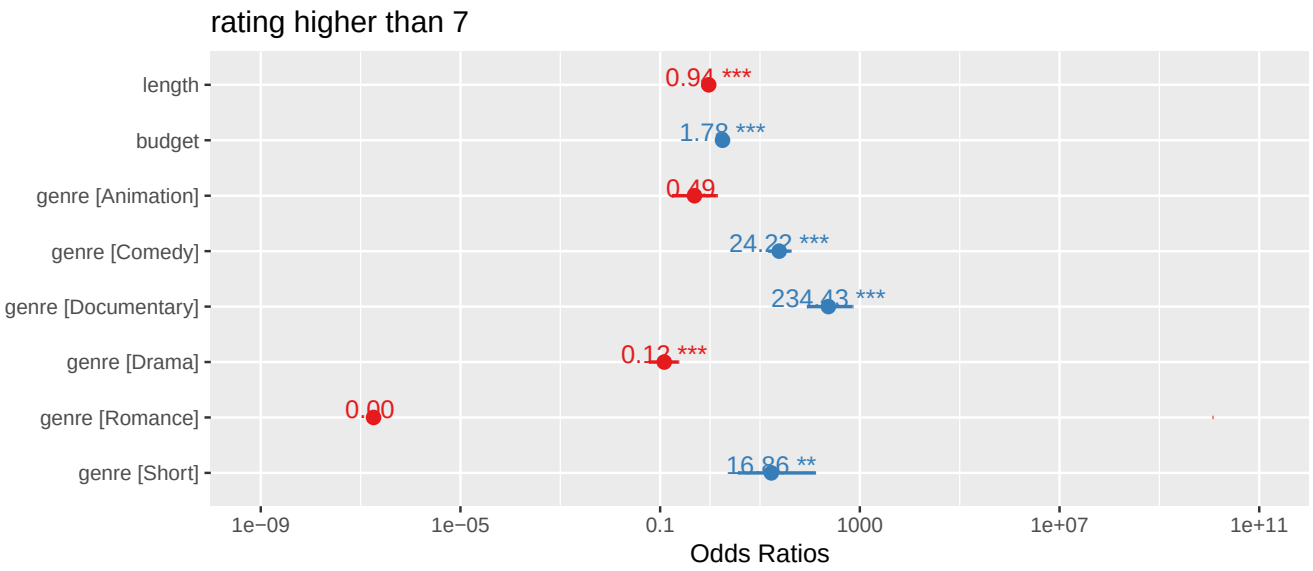

Figure 6: Red dots implying negative relationship, while blue dots suggest positive relationship. Significant coefficients are marked with stars.

According to Figure 6, as length and budget increase, the odds of having rating greater 7 will be decreasing (multiply by 0.94) and increasing (multiply by 1.78) respectively. In this analysis, the baseline for Genre is Action. For Comedy, Documentary, and Short films, the odds ratios have values significantly higher than 1. For instance, Comedy's odds of having rating higher than 7 is 24.22 times those of Action films.

Interestingly, the Animation's odds is 0.49 times to Action's odds, and this is not align with the fact that Animation's proportion (of having rating higher than 7) is bigger than Action. Moreover, we notice that Animation has a mean of length 19.92 min, which is lower than the overall mean, 82.29 min from Table 2. This is also already align with the negative correlation between length and rating which can be seen from Figure 2. One hypothesis, compared to other genres, Animation is very skewed in terms of length values, and its IQR is 15 times smaller than SD - Table 9. However, Animation's odds is not statistically significant.

```
df |>
   summarize('Mean' = mean(length),
             'Median' = median(length),
             'St.Dev' = sd(length),
             'IQR' = quantile(length,0.75)-quantile(length,0.25),
             'Sample_size' = n(),
.by = genre) |>
 gt() |>
  fmt_number(decimals=2) |>
  cols_label(
    Mean = html("Mean"),
    Median = html("Median"),
    St.Dev = html("Std. Dev"),
    IQR = html("Interquartile Range"),
    Sample_size = html("Sample Size")
  )
```

Table 9: Summary statistics of length for each genre.

genre	Mean	Median	Std. Dev	Interquartile Range	Sample Size
Romance	99.46	102.00	11.58	16.00	13.00
Documentary	67.22	75.00	33.05	48.50	95.00
Comedy	84.18	90.00	28.52	20.00	343.00
Action	92.50	91.00	23.03	18.00	412.00
Drama	96.26	95.50	29.78	23.25	424.00
Animation	19.92	7.00	30.31	2.00	85.00
Short	15.52	14.50	8.40	10.00	64.00

4.2.3 Predicted Probabilities

```
# Plot predicted probabilities of having rating higher than 7
plot_model(model,type="pred",terms=c("length","genre"))
```

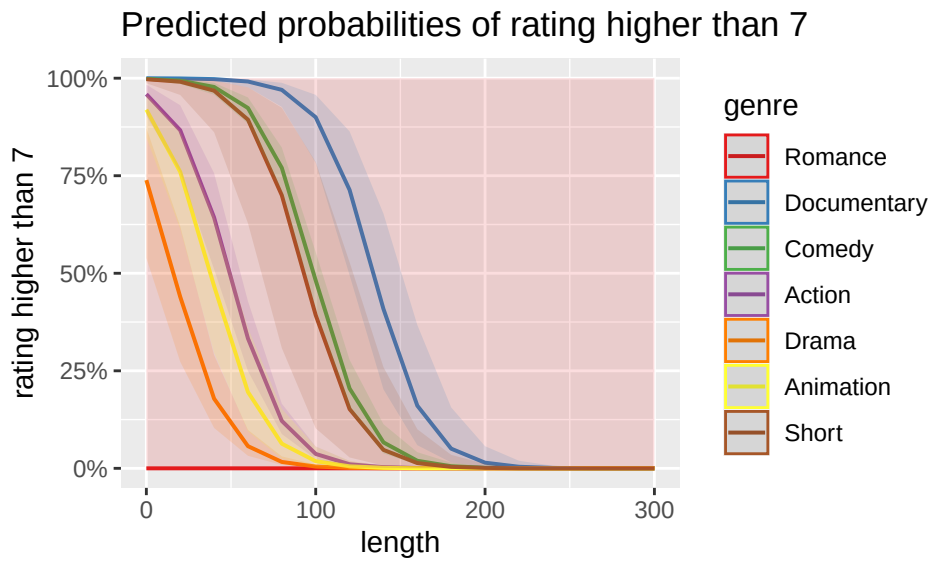

Figure 7: Predicted probabilities of rating higher than 7 for budget and genre.

```
# Plot predicted probabilities of having rating higher than 7
plot_model(model,type="pred",terms=c("budget","genre"))
```

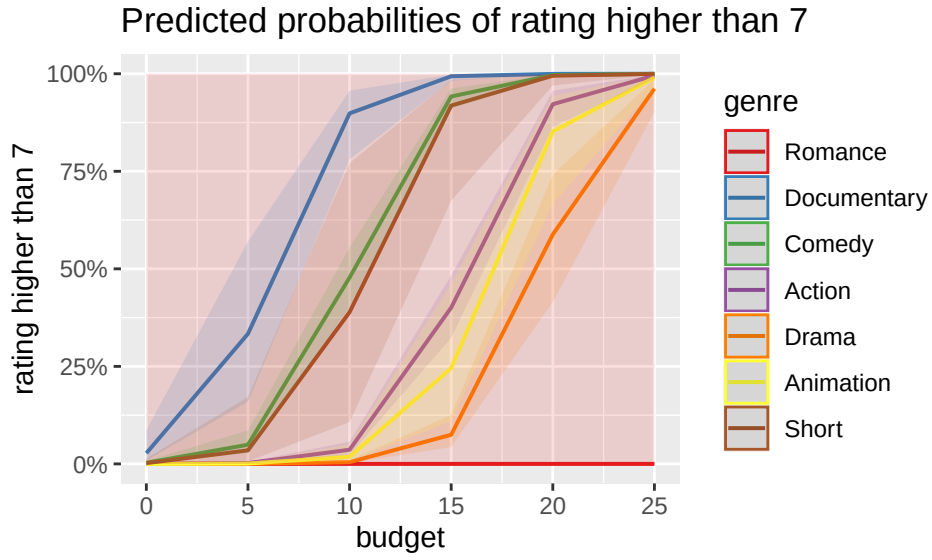

Figure 8: Predicted probabilities of rating higher than 7 for length and genre.

Other than Romance genre, the probabilities of having rating higher than 7 decreasing by length and increasing by budget. It can also be seen from odds sign as well in Figure 6. Moreover, we can also compare to Figure 5 that Romance (0%), Drama (5%), and Action (16%) are the lowest in terms of proportion of having rating greater than 7, and this is align with their predicted probabilities. However, this is not the case for Animation predicted probabilities. For Romance, it will be always predicted to have rating lower than and equal to 7.

4.3 Sanity Checking: Linear Regression

```
# Fitting the model_3 (rating > 7) and model_4 (rating <= 7)

df_higher <- df[df$rating_higher_than_7 == 1, ]
df_lower <- df[df$rating_higher_than_7 == 0, ]

model_3 <- lm(rating ~ length + budget + genre,
              data = df_lower)
model_4 <- lm(rating ~ length + budget +  genre,
              data = df_higher)

suppressWarnings(export_summs(model_3, model_4,
                error_format = "[{conf.low}, {conf.high}]"))
```

In this part, the one on the left is model which has observations with rating <= 7 and the one on the right having observation with rating > 7. They are fitted separately in order to analyze the effects of variables separately on high-scoring and low-scoring movies. In model 2, when rating is higher than 7, budget has a slightly higher impact on rating than length by looking at the effect size or the magnitude of the coefficients. However, both coefficients are very close to 0, so they are not practically significant. As the budget gets higher then rating is also getting higher, but we see the opposite for the length. For Model 1, the rating gets higher as the budget and length getting higher. However, for both models, the coefficients are not statistically and practically significant.

Table 10: Hypothesis Testing and Goodness of fit. The one on the left is model which has observations with rating <= 7 and the one on the right having observation with rating > 7.

	Model 1	Model 2
(Intercept)	3.72 ***	7.99 ***
	[3.35, 4.09]	[7.76, 8.21]
length	0.00	-0.00
	[-0.00, 0.00]	[-0.00, 0.00]
budget	0.01	0.00
	[-0.02, 0.03]	[-0.01, 0.01]
genreAnimation	-0.11	0.05
	[-0.52, 0.30]	[-0.11, 0.21]
genreComedy	-0.01	-0.01
	[-0.20, 0.18]	[-0.12, 0.10]
genreDocumentary	0.00	-0.02
	[-0.63, 0.64]	[-0.15, 0.10]
genreDrama	-0.06	0.08
	[-0.19, 0.08]	[-0.12, 0.28]
genreRomance	-0.51	
	[-1.03, 0.02]	
genreShort	3.11 ***	0.01
	[1.78, 4.44]	[-0.15, 0.17]
N	937	499
R2	0.03	0.01

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

5 Conclusion

The model which has budget, length, and genre as the predictors is chosen to be the best model. The AIC is slightly lower than model which also has year and votes as the predictors, but BIC is noticeably higher. In this case, simpler model is chosen.

The caveat from this model that it has coefficients which are not statistically significant (genreAnimation and genreRomance). However, if genre is being removed, AIC will increase significantly. It is better to incorporate genre as predictors. The hypothesis on Hosmer-Lemeshow also shows no evidence for lack of fit.

In summary, answering the research question, budget and length have significant relationship to binary rating, whether greater than 7 or not. For Genre, two of them are not significant.

6 Appendix

```
# Comparing models with different explanatory variables

model_6 <- glm(rating_higher_than_7 ~ length + budget,
               data = df,
               family = binomial(link = "logit"))

# Display the model summary (log-odds)
model_6 %>%
  summ(model.info = FALSE)
```

Table 11: Hypothesis Testing and Goodness of fit. Model without year, votes, and genre.

$\chi^2(2)$	562.99
Pseudo-R^2 (Cragg-Uhler)	0.45
Pseudo-R^2 (McFadden)	0.30
AIC	1297.98
BIC	1313.79

	Est.	S.E.	z val.	p
(Intercept)	-0.60	0.35	-1.74	0.08
length	-0.04	0.00	-16.47	0.00
budget	0.29	0.03	10.92	0.00

Standard errors: MLE