Final Project

Jin Yan

2022-12-01

Abstract

The full name of IMBD is Internet Movie Database, which has a wealth of information about movie works, including basic information such as movie actors, directors, plots, and movie reviews, as well as deeper movie revenue and other content. It is an authoritative movie data organization. Increasingly, the movie box office has become an excellent basis for judging whether a movie is good or bad. This year's movie "Top Gun 2" won the annual box office champion with a box office of 1.4 billion US dollars, which aroused my interest in exploring the factors that affect the box office: what factors will Is it the year of release that affects the trend of the movie box office? Online evaluation? Or a movie genre? In order to clarify this problem, I built a multi-layer data set with movie type and release year as the data set to explore the influencing factors of movie box office, and found that IMBD score and movie type will affect the box office trend. For example, medium-length action movies have always been The box office leader, but the metacritic score has no positive impact on the box office, this report analyzes through introduction, methods, results, and discussion.

Introducation

Usually, the box office of a movie is related to the plot and content of the movie. However, due to the mixed evaluations of movies by movie fans, the evaluation of movie fans after watching usually affects the viewing decision of fans who have not watched the movie. One of the important factors for sustainable development. Usually, movies with high ratings will get higher box office. For example, the IMBD score of "TOP GUN 2" is 8.0. The higher rating also makes it the annual box office champion, but there are still movies The score is better, but the box office is not satisfactory, so we have to consider the impact of the movie content, that is, the type of movie, on the movie box office. Generally speaking, movies that make people adrenaline soar, such as action movies and science fiction movies, have novel content and shocking pictures. , it will bring a good box office, but although some movies have better special effects, their box office is lower. Therefore, I decided to introduce a multi-level model to find out the impact of fixed effects (such as multi-site ratings, movie duration, number of votes, etc.) and random effects (movie type, release year)

Methods

Data clean

I downloaded the IMBD data from kaggle: https://www.kaggle.com/datasets/preetviradiya/imdb-movies-ratings-details. This data records the detailed data of 1000 movies from 1925 to 2018. Since the years are too long, I rearranged the release time of the movies, that is, the release time of movies during 1920-1930 is 1920s, and so on. In addition, the data Without considering the quarterly impact factors, two website scores were introduced: IMBD and Metacritic scores, and the number of votes was counted.

column names	explanation
X	Movie ID
name	Name of the movie
year	Year of release
runtime	Movie runtime
genre1	Movie Genre
rating	IMDB Ratings
metascore	Metascore on IMDB
timeline	Short storyline of the movie
votes	Total votes earned on IMDB
gross	Box-office grossings
decade	decade of release

By aforementioned part, I've got a `data` with 747 observations and 11 variables,I try to figure out whether or not to use the 11 variables.

Average data

I get the average values of data group by genre and decade

```
average_data_genre
```

```
## # A tibble: 14 x 6
##    genre1    runtime rating metascore   votes gross
##    <chr>       <dbl>  <dbl>     <dbl>   <dbl> <dbl>
##  1 Action      129      7.9      73.5 548476. 156.
##  2 Adventure   133.     7.9      77.5 398450.  93.3
##  3 Animation    99.5    7.9      81.9 321309. 132.
##  4 Biography   135.     7.9      76.8 325030.  64.6
##  5 Comedy      110.     7.9      78.2 233686.  37.3
##  6 Crime       128.     8        77.1 389467.  39.2
##  7 Drama       127.     7.9      79   272261.  42.8
##  8 Family      108.     7.8      79   279978. 220.
##  9 Fantasy     100      7.6      66   190214   14.4
## 10 Film-Noir    93      8.1      97   161382    0.4
## 11 Horror      108.     7.9      78.5 354228.  67.4
## 12 Mystery     126.     8.2      78.2 641221.  42.4
## 13 Thriller    108      7.8      81    28211   17.6
## 14 Western     152.     8.3      78.2 327916.  14.6
```

```
average_data_decade
```

```
## # A tibble: 10 x 6
##    decade runtime rating metascore   votes gross
##    <chr>    <dbl>  <dbl>     <dbl>   <dbl> <dbl>
##  1 1920s     110.    8.2      97.5 108368.   0.6
##  2 1930s     118.    8.1      91.7 149789.  23.7
##  3 1940s     115.    8.1      95.5 157126   10.3
##  4 1950s     123.    8.1      92.2 174154.  13.3
##  5 1960s     132.    8        83.3 166938.  36.3
##  6 1970s     122.    8        81.3 264995.  54.7
##  7 1980s     120.    7.9      76.7 293093.  65.6
##  8 1990s     123.    8        74.4 410219.  67.3
##  9 2000s     124.    7.9      74.3 390968.  72.8
## 10 2010s     124.    7.9      77.5 401408. 119.
```

genre data vs decade graphs

```
grid.arrange(runtime_by_genre, runtime_by_decade, ncol = 2)
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_smooth()`).
## Removed 2 rows containing non-finite values (`stat_smooth()`).
```

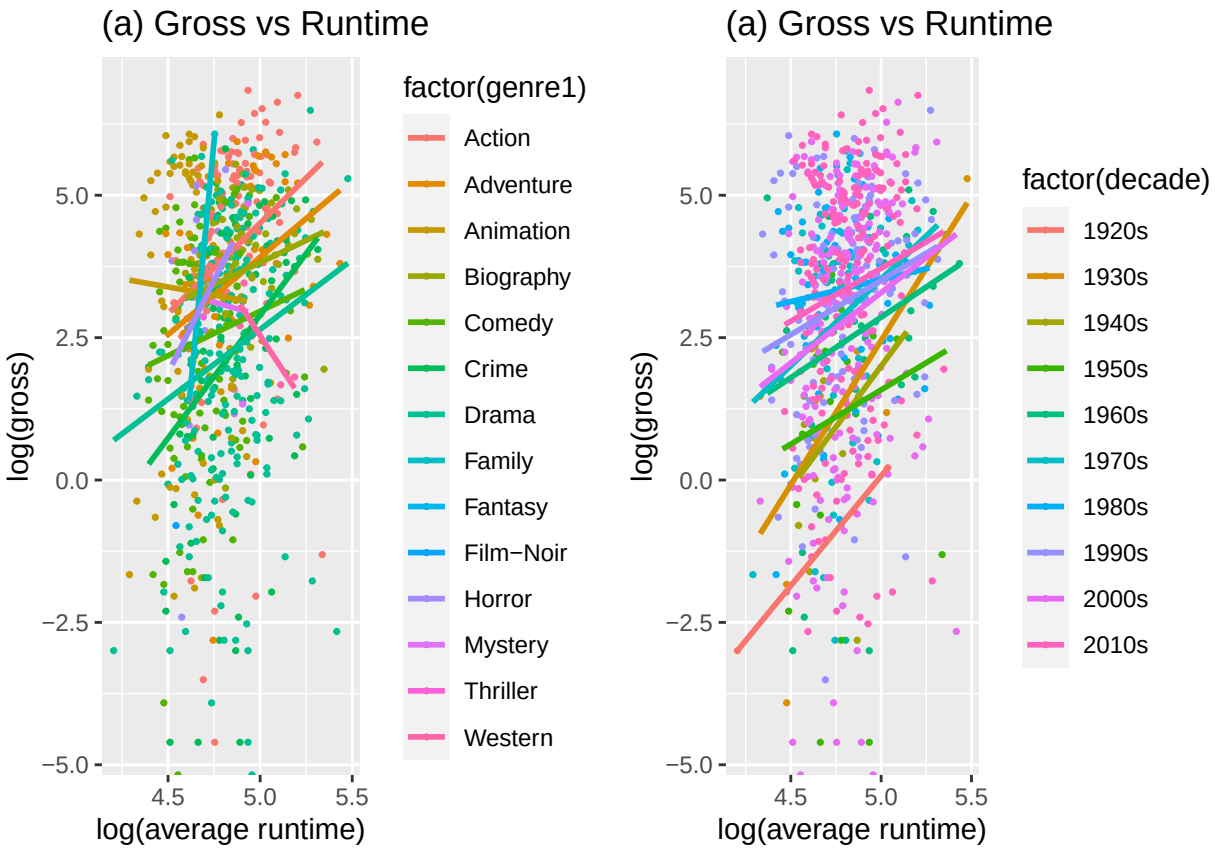


Figure 1 illustrates the relationship between gross and average runtime, while fig(a) is in genre level and fig(b) is in decade level. However, whatever the level, gross show the increasing trend as runtime going up. And in different genre and decade, the intercepts and slopes show slights differences. After I draw the graph of gross versus metascore, votes, the figures are quite similar. Thus I put them in the appendix.

```
grid.arrange(rating_by_genre, rating_by_decade, ncol = 2)
```

```
## Warning: Removed 2 rows containing non-finite values (‘stat_smooth()‘).
## Removed 2 rows containing non-finite values (‘stat_smooth()‘).
```

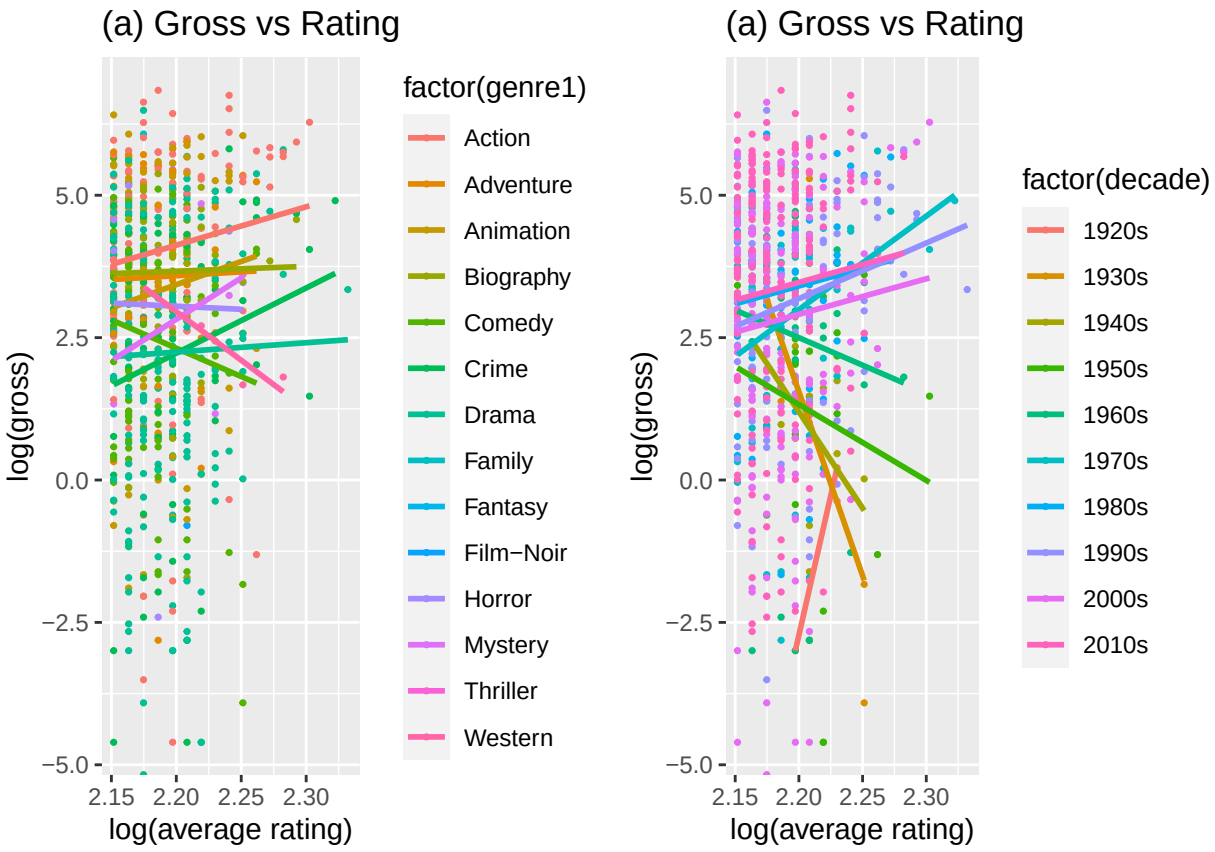


Figure 2 shows the correlation between gross and rating. Similarly, figure(a) is in genre level while figure(b) is in decade level. But the result is very strange. In the middle of the twentieth century, the ratings of movies were inversely proportional to the box office, and the ratings of some types of movies were also inversely proportional to the box office, such as western movies and crime movies.

Gross for genre/decade distribution

I found the gross are highly correlated with decade, I thought may be the rising in social incomes.

Model fitting

Since different movie genres and decades have a considerable impact on the model, I decided to use a multilevel model to fit the data. Since all variables are more or less skewed and have heavy tails, I use log(variable + 1) to create new variables. See the appendix of this report for the original distribution diagram of all variables. For the next step, I plot the Pearson correlation matrix for predictor selection.

Model fitting data summarize

correlation between

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```

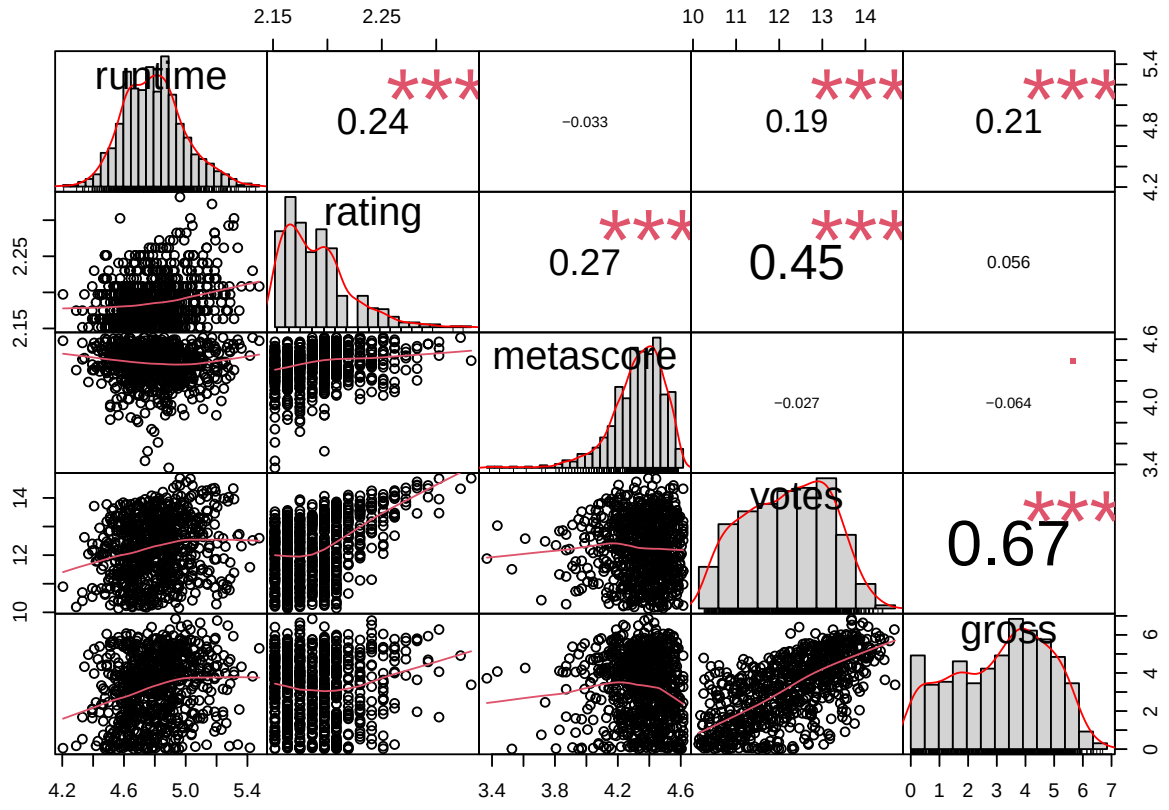


Figure 1: Correlation Matrix

From the graphs,I decides to use the all variables to fit my model

Model Fitting

```
model <- lmer(gross~runtime+rating+metascore+votes
              +(1+votes+rating+runtime|genre1)+
```

```
              (1+metascore|decade),
            data=log_data)
```

```
## boundary (singular) fit: see help('isSingular')
```

```
## Warning: Model failed to converge with 2 negative eigenvalues: -2.9e+00 -1.5e+02
```

Due to the results of the model, I founld the variable of metascore is not significant, so i remove it to fit my model.Here is the summary of model(fixed effect) and all variables here are considered as statistically significant at $\alpha = 0.5$ level. To be more clear, a fixed effect parameters are also include

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	29.24	5.12	581.38	5.71	0.00 ***
log_runtime	1.62	0.33	8.90	5.00	0.00 ***
log_rating	-21.35	2.04	65.90	-10.45	0.00 ***
log_metascore	-0.80	0.71	629.92	-1.13	0.26
log_votes	1.35	0.12	7.29	11.11	0.00 ***

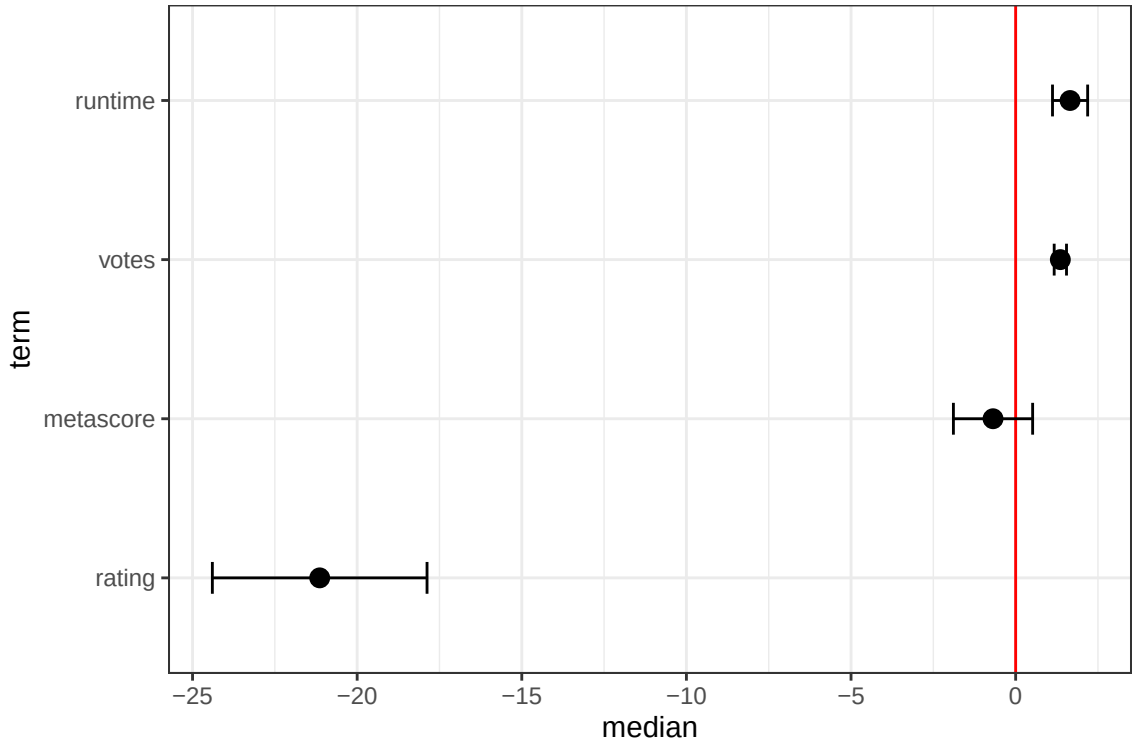


Figure 2: Fixed Effect of gross Model

And the following tables are the summary of random effects. The first one is random effect of movie genre and the second one is movie decade.

```
round(ranef(model)$genre1, digits = 2)[1:14, ]
```

```
##           (Intercept) votes rating runtime
## Action          -2.47  0.32  -2.06    0.63
## Adventure       -1.83  0.15  -0.76    0.35
## Animation       14.71  0.57  -8.71   -0.39
## Biography       12.06 -0.13  -2.10   -1.14
## Comedy          -0.62 -0.28   2.48   -0.32
## Crime           -8.94 -0.06   2.91    0.62
## Drama           -0.49 -0.19   1.71   -0.22
## Family          -1.93  0.12  -0.52    0.32
## Fantasy         -2.45 -0.04   0.99    0.14
## Film-Noir       -4.10  0.01   1.00    0.34
## Horror          -1.91  0.12  -0.54    0.32
## Mystery         -4.69 -0.11   2.19    0.22
## Thriller         3.07 -0.18   0.71   -0.49
## Western         -0.43 -0.31   2.70   -0.38
```

```
round(ranef(model)$decade, digits = 2)
```

```
##       (Intercept) metascore
## 1920s        0.50     -0.13
## 1930s       -0.02      0.00
## 1940s        1.86     -0.44
## 1950s        5.65     -1.24
## 1960s       10.28     -2.24
## 1970s        2.38     -0.45
## 1980s        0.06      0.02
## 1990s       -7.35      1.65
## 2000s       -6.58      1.37
## 2010s       -6.78      1.47
```

Additionally, a random effect plot for `genre` level are included. we can come to the conclusion that baseline of gross for each genre are quite different. This exactly verify that animation movies are willing to get a higher gross. Another parameter that differs most is `decade`, which means 1960s is a The heyday of cinema.

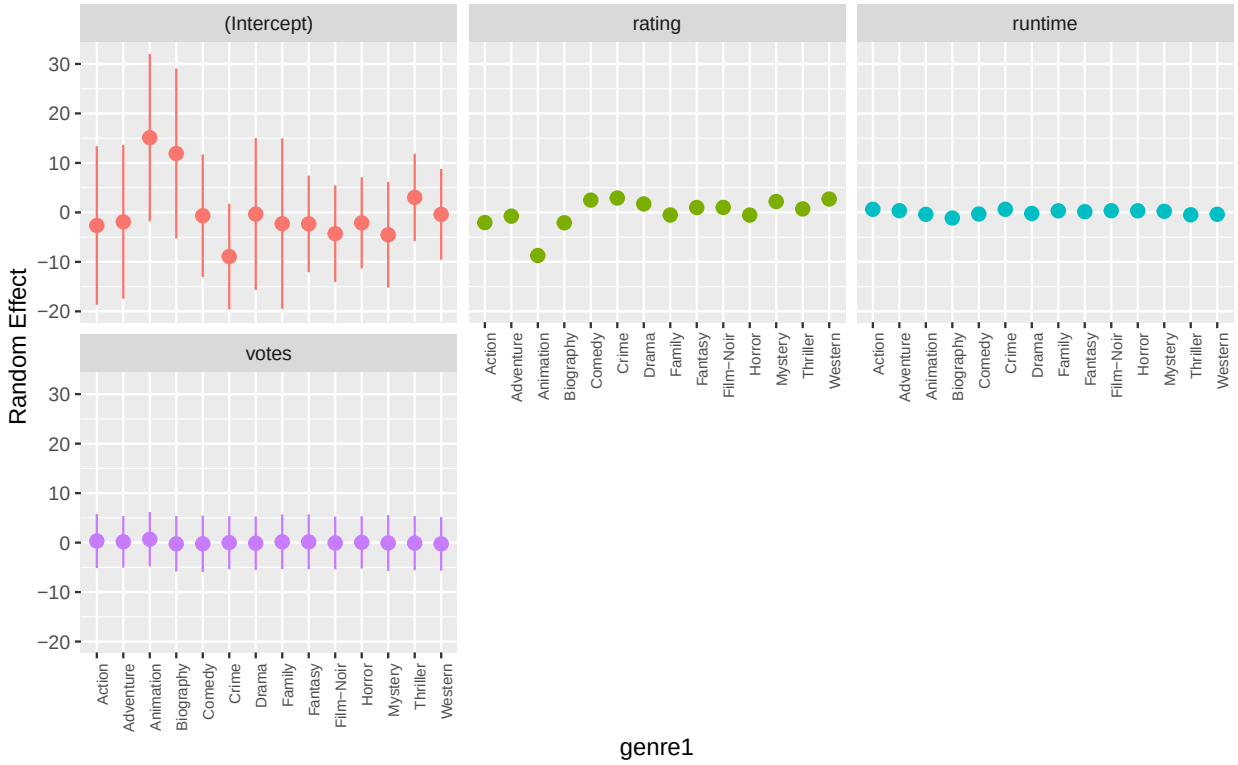


Figure 3: Random Effect of gross Model

Result

Interpretation

we are able to get the following formula of fixed effect:

$$log(gross + 1) = 29.24 + 1.62 \times log(runtime + 1) - 21.35 \times log(rating + 1) + 1.35 \times log(votes + 1)$$

Then add the random effect to the intercepts and slopes and get the estimated formula:

$$log(gross + 1) = 102.21 + 2.13 \times log(runtime + 1) - 5.46 \times log(rating + 1) + 1.46 \times log(votes + 1)$$

In the formula, all parameters except rating are positive, which means that the duration of the movie and the number of votes have a positive impact on the box office of the movie. Looking at the rating, the better the score, the lower the box office may be related to the malicious accusation of the movie. Therefore The negative impact of rating on the box office is reasonable. In the model, for every 1% increase in gross, runtime will increase by 2.13%, and for every 1% increase in gross, voting will increase by 1.46%.

Model Checking

The left plot of Figure 8 is residual plot and the right one is residual Q-Q plot. According to it, the mean value of residuals is approximately 0. Yet as the fitted value close to 0, there's no negative residuals. This phenomenon can be explained by there exists clear lower bound for actual gross and when we make prediction with multilevel model, that would not happen. As for Q-Q plot in Figure 8, majority points except tail ones are on the normal distribution line, thus the normality check is acceptable.

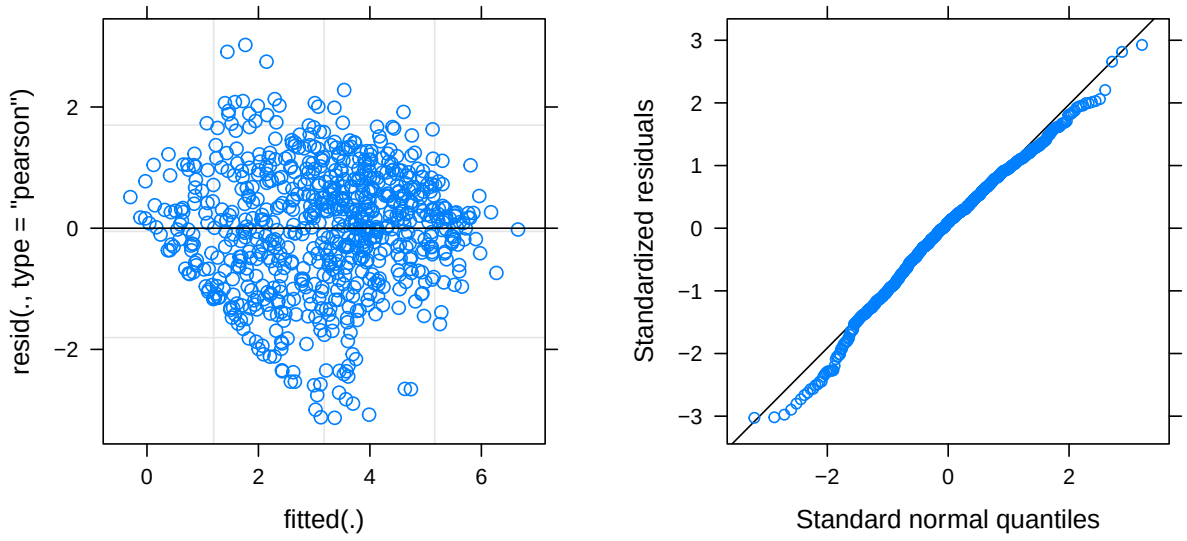


Figure 4: Residual plot and Q-Q plot.

Discussion

In this report, we use a multi-level model to calculate the relationship between the box office of a movie and several basic details of the movie. In addition, the model also considers two levels: the genre of the movie and the release period of the movie. Generally speaking, from the perspective of fixed effects, movie ratings and fans' votes have a positive impact on the movie's box office, and ratings have a negative impact. The better, no matter in terms of movie type or release year, such factors are reasonable, and the final model explanation is also very good.

However, this report also has some limitations. First of all, the data sets I constructed included World War II, post-war economic reconstruction, the Cold War, and the period of rapid economic development. These off-site factors will have a greater impact and influence on the film industry, especially It is the epidemic in recent years, which also has a great impact on the box office of movies, so I should choose different movies in the same period for analysis and fit a multi-level model.

Appendix

Variable distributions

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
```

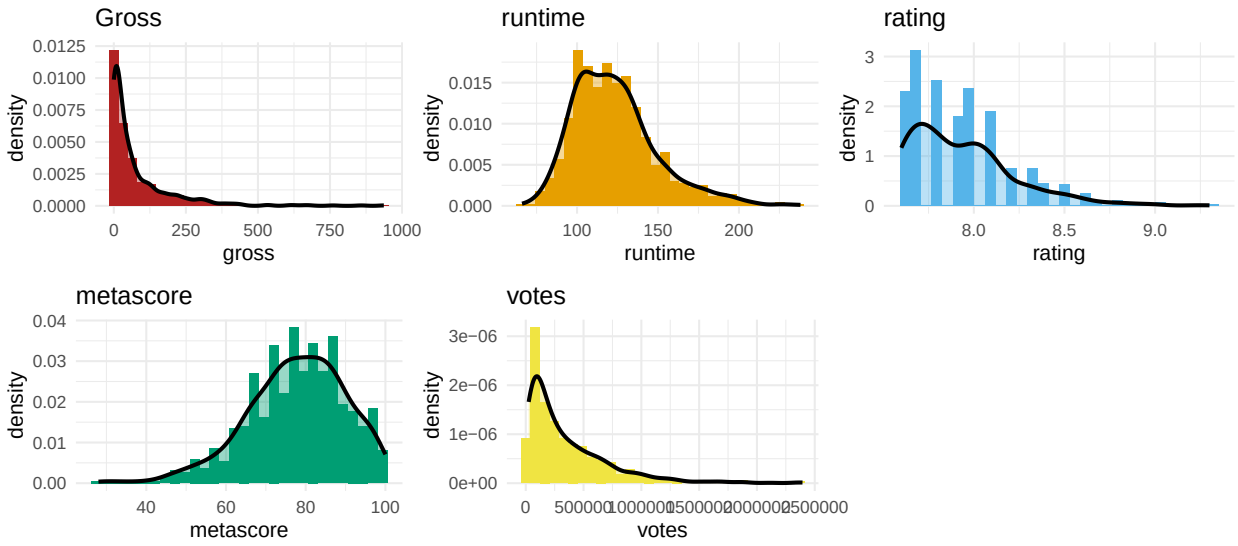
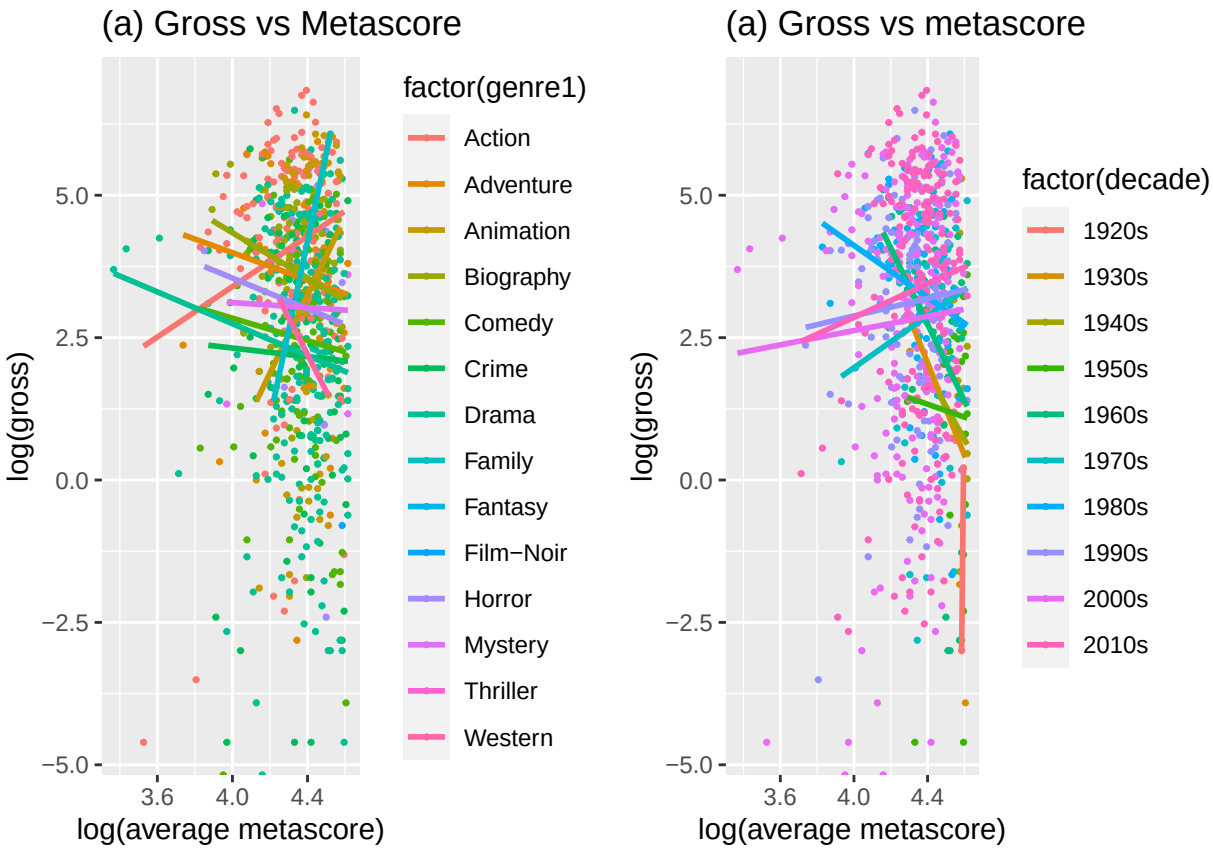


Figure 5: EDA: distribution plots

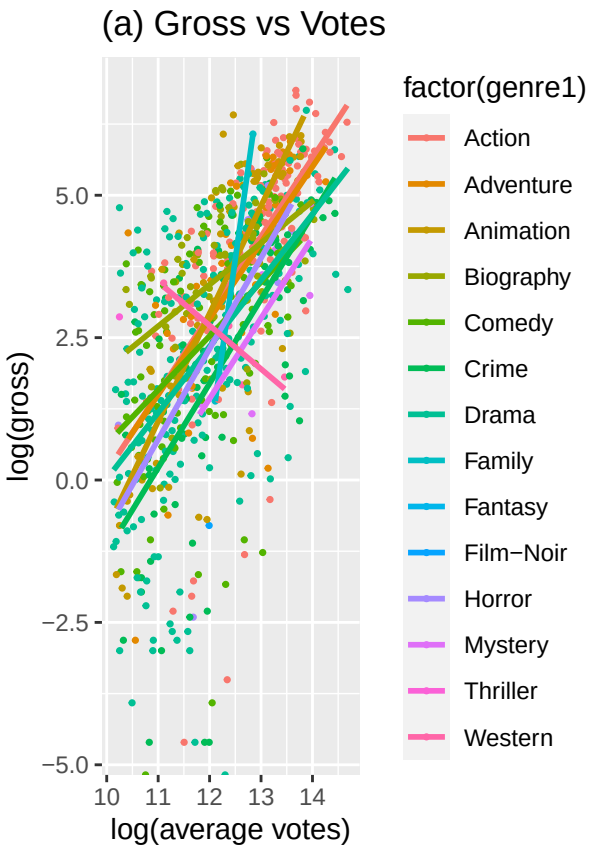
```
grid.arrange(metascore_by_genre1, metascore_by_decade, ncol = 2)
```

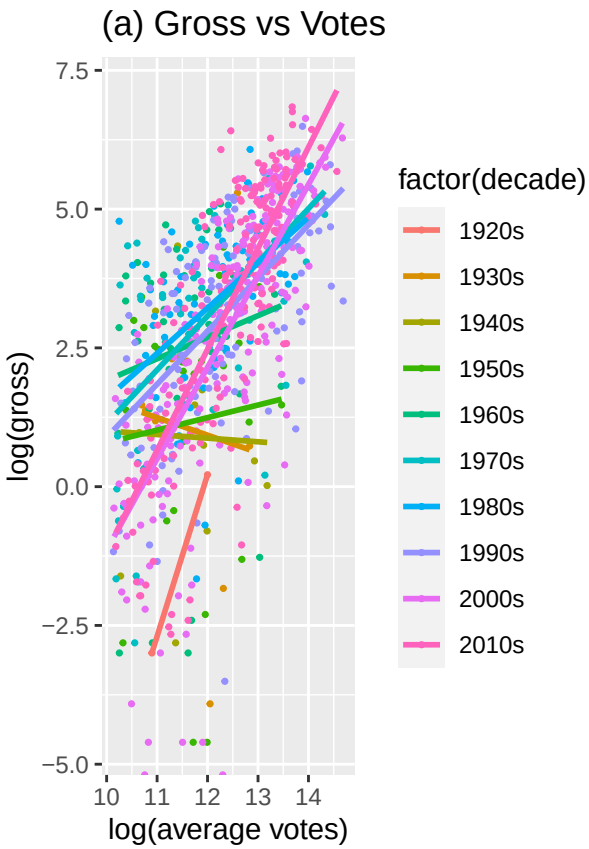
```
## Warning: Removed 2 rows containing non-finite values (‘stat_smooth()‘).
## Removed 2 rows containing non-finite values (‘stat_smooth()‘).
```




```
grid.arrange(votes_by_genre1, votes_by_decade, ncol = 2)
```

```
## Warning: Removed 2 rows containing non-finite values (‘stat_smooth()‘).
## Removed 2 rows containing non-finite values (‘stat_smooth()‘).
```


(a) Gross vs Votes
log(gross)
log(average votes)
5.0
2.5
0.0
−2.5
−5.0
10
11
12
13
14
factor(genre1)
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Film−Noir
Horror
Mystery
Thriller
Western


(a) Gross vs Votes
log(gross)
log(average votes)
7.5
5.0
2.5
0.0
−2.5
−5.0
10
11
12
13
14
factor(decade)
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s