An analysis of the genre that gives the best rating for Christmas movies

Shufan Wang

1 Load libs

```
pacman::p_load(tidyverse,readxl,gt,targets)
```

2 Introduction

This analysis aims to find the genre of Christmas movies having the best ratings.

Firstly, we explored and cleaned the data by deleting the irrelevant variables and removing the observations with missing values. We found some genres have very few observations, which may lead to our result lacking statistical significance. Therefore, we considered the genres more generously in this analysis. To be specific, for observations having genres = "Action, Adventure, Animation", we split the observation into 3 observations that have genres = "Action", "Adventure" and "Animation", respectively. The new dataset is called movies_split.

After we cleaned the data, we made box plots to show the relationship between genres and average ratings. There are still genres that have few observations, so we classified the genres with a rate of occurrence less than 0.01 as "other". The box plots Figure 6 show that the majority of genres have outliers. Therefore, we chose median value to measure the average ratings, because the mean values will be significantly influenced by the outliers. By comparing the median values, we found that the genre "Documentary" had the highest median average rating.

However, there is a big difference between the number of different genres' mean "num_votes". For example, the mean of num_votes in "Documentary" is almost 30 times less than in "Drama". It suggests that selecting the best genre only dependent on average ratings might have some potential bias. For example, the high average ratings in "Documentary" were probably rated by a few people who particularly like documentaries. Therefore, we trained a random forest model and extracted the feature importance to find out which genre has the largest positive influence on average rating.

The random forest's hyper-parameters were tuned by 10-fold validation and the best model was selected by the minimum rmse. Excluding genres and num_votes, we also added the interaction terms of them in the model. According to the variable importance plot Figure 7, the genre "Horror" had the largest positive influence on the average rating. However, if we consider the interaction term, the genre "Drama" had the biggest boost to the average rating.

In summary, if we do not consider the influence of num_votes, the genre "Documentary" has the best rating. However, if we consider the interaction between num_votes and genres, then "Drama" can lead to the best rating.

3 EDA

We explored the raw data. The procedure of EDA is given: 1. Checking the basic structure of the data: number of variables, number of observations. 2. Checking if there are missing values in the data. 3. Checking the number of observations in each genre. 4. Checking the distribution of the response (average rating) 5. Checking the distribution of the num_votes

We discussed our results of EDA in detail in the following.

```
tar_load(movies_EDA)
```

The dataset has 2265 observations and 14 variables (6 categorical variables, 4 logical variables and 4 numeric variables).

```
movies_EDA$movies_summary
```

```
# A tibble: 14 x 19
   skim_type skim_variable   n_missing complete_rate character.min character.max
 * <chr>     <chr>               <int>         <dbl>         <int>         <int>
 1 character tconst                  0         1                 9            10
 2 character title_type              0         1                 5             7
 3 character primary_title           0         1                 5            97
 4 character original_title          0         1                 4            97
 5 character genres                 32         0.986             5            27
 6 character simple_title              0         1                 4            96
 7 logical   christmas               0         1                NA            NA
 8 logical   hanukkah                0         1                NA            NA
 9 logical   kwanzaa                 0         1                NA            NA
```

```
10 logical   holiday                 0          1                 NA             NA
11 numeric   year                    0          1                 NA             NA
12 numeric   runtime_minutes       189          0.917             NA             NA
13 numeric   average_rating          0          1                 NA             NA
14 numeric   num_votes               0          1                 NA             NA
# i 13 more variables: character.empty <int>, character.n_unique <int>,
#   character.whitespace <int>, logical.mean <dbl>, logical.count <chr>,
#   numeric.mean <dbl>, numeric.sd <dbl>, numeric.p0 <dbl>, numeric.p25 <dbl>,
#   numeric.p50 <dbl>, numeric.p75 <dbl>, numeric.p100 <dbl>,
#   numeric.hist <chr>
```

According to the Figure 1, variables "runtime_minutes" and "genres" have missing values. We may need to delete the observations that contain the missing values.

```
movies_EDA$missing_plot
```

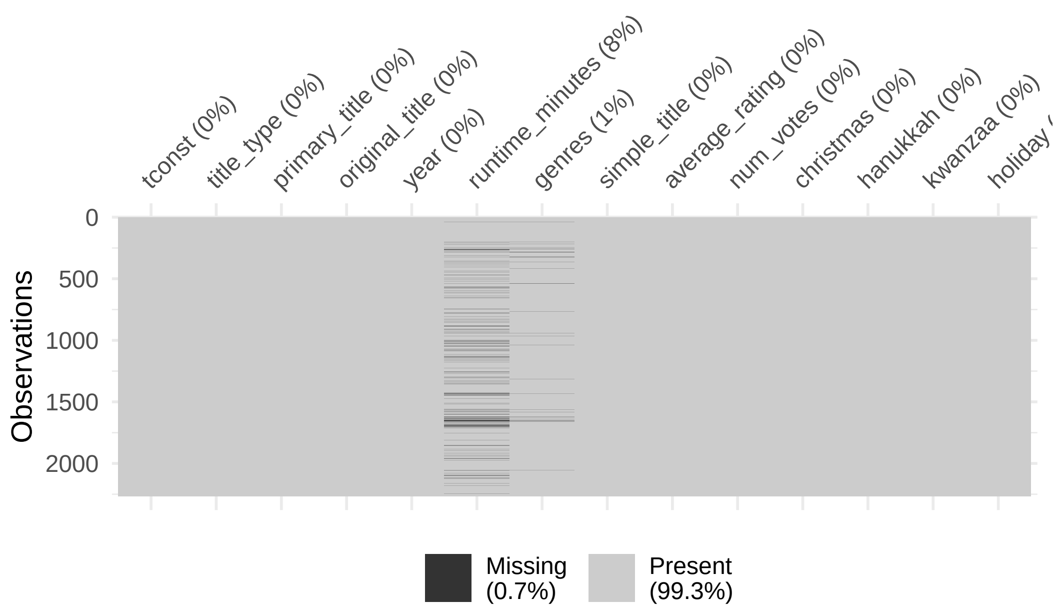

Figure 1: Missing values plot

Table 1 shows the number of observations in each unique class of variable "genre". There are numerous genres classes that only have <10 observations. We may need to divide the genres into more general classes.

```
movies_EDA$genres_count
```

Table 1: genres

	n
Action	
	1.00
Action,Adventure,Animation	
	3.00
Action,Adventure,Comedy	
	2.00
Action,Adventure,Drama	
	3.00
Action,Comedy	
	3.00
Action,Comedy,Crime	
	4.00
Action,Comedy,Drama	
	2.00
Action,Comedy,Family	
	1.00
Action,Comedy,Horror	
	4.00
Action,Comedy,Romance	
	1.00
Action,Crime,Drama	
	3.00
Action,Crime,Thriller	
	1.00
Action,Drama	
	1.00

Action,Family
1.00
Action,Horror,Thriller
1.00
Adventure
1.00
Adventure,Animation,Comedy
38.00
Adventure,Animation,Drama
2.00
Adventure,Animation,Family
14.00
Adventure,Animation,Fantasy
1.00
Adventure,Comedy
3.00
Adventure,Comedy,Crime
1.00
Adventure,Comedy,Drama
7.00
Adventure,Comedy,Family
17.00
Adventure,Comedy,Fantasy
1.00
Adventure,Comedy,Musical
1.00
Adventure,Crime,Family
1.00
Adventure,Crime,Mystery
1.00
Adventure,Crime,Romance

1.00
Adventure,Drama,Family
6.00
Adventure,Drama,Fantasy
1.00
Adventure,Drama,Mystery
1.00
Adventure,Family
5.00
Adventure,Family,Fantasy
4.00
Adventure,Family,Musical
1.00
Adventure,Fantasy,Horror
1.00
Adventure,Fantasy,Romance
1.00
Animation
46.00
Animation,Comedy
5.00
Animation,Comedy,Drama
4.00
Animation,Comedy,Family
34.00
Animation,Comedy,Fantasy
1.00
Animation,Comedy,Musical
1.00
Animation,Comedy,Romance
3.00

Animation,Comedy,Sci-Fi
1.00
Animation,Comedy,Short
2.00
Animation,Drama
1.00
Animation,Drama,Family
6.00
Animation,Drama,Fantasy
2.00
Animation,Family
49.00
Animation,Family,Fantasy
13.00
Animation,Family,Music
2.00
Animation,Family,Musical
6.00
Animation,Family,Sci-Fi
2.00
Animation,Family,Short
9.00
Animation,Fantasy
2.00
Animation,Fantasy,Horror
1.00
Animation,Fantasy,Short
2.00
Animation,Horror,Short
1.00
Animation,Musical

	2.00
Animation,Romance	
	1.00
Animation,Short	
	14.00
Biography,Comedy,Drama	
	1.00
Biography,Documentary,Sport	
	1.00
Biography,Drama	
	1.00
Biography,Drama,Family	
	1.00
Biography,Drama,Music	
	1.00
Biography,Romance,War	
	1.00
Comedy	
	182.00
Comedy,Crime	
	1.00
Comedy,Crime,Drama	
	3.00
Comedy,Crime,Family	
	3.00
Comedy,Crime,Horror	
	1.00
Comedy,Crime,Mystery	
	3.00
Comedy,Crime,Romance	
	1.00

Comedy,Crime,Thriller
2.00
Comedy,Crime,Western
1.00
Comedy,Documentary
2.00
Comedy,Documentary,Family
2.00
Comedy,Documentary,Music
1.00
Comedy,Documentary,Short
2.00
Comedy,Documentary,War
1.00
Comedy,Drama
44.00
Comedy,Drama,Family
99.00
Comedy,Drama,Fantasy
14.00
Comedy,Drama,Horror
1.00
Comedy,Drama,Music
8.00
Comedy,Drama,Musical
3.00
Comedy,Drama,Mystery
2.00
Comedy,Drama,Romance
148.00
Comedy,Drama,Sci-Fi

1.00
Comedy,Drama,Thriller
1.00
Comedy,Family
50.00
Comedy,Family,Fantasy
25.00
Comedy,Family,Horror
1.00
Comedy,Family,Music
8.00
Comedy,Family,Musical
6.00
Comedy,Family,Mystery
1.00
Comedy,Family,Romance
32.00
Comedy,Family,Sci-Fi
1.00
Comedy,Family,Short
3.00
Comedy,Fantasy
15.00
Comedy,Fantasy,Horror
1.00
Comedy,Fantasy,Romance
9.00
Comedy,Fantasy,Sci-Fi
1.00
Comedy,History
1.00

Comedy,History,Musical	
	1.00
Comedy,Horror	
	7.00
Comedy,Horror,Mystery	
	1.00
Comedy,Horror,Sci-Fi	
	1.00
Comedy,Music	
	4.00
Comedy,Music,Romance	
	3.00
Comedy,Musical	
	6.00
Comedy,Musical,Romance	
	5.00
Comedy,Musical,Short	
	1.00
Comedy,Mystery	
	2.00
Comedy,Mystery,Romance	
	2.00
Comedy,Romance	
	156.00
Comedy,Romance,Thriller	
	1.00
Comedy,Romance,War	
	1.00
Comedy,Sci-Fi	
	1.00
Comedy,Sci-Fi,Short	

1.00
Comedy,Short
14.00
Comedy,Thriller
1.00
Crime
3.00
Crime,Documentary,Drama
1.00
Crime,Drama
2.00
Crime,Drama,Film-Noir
2.00
Crime,Drama,Mystery
1.00
Crime,Drama,Romance
2.00
Crime,Drama,Thriller
4.00
Crime,Horror,Thriller
1.00
Crime,Thriller
1.00
Documentary
52.00
Documentary,Family
4.00
Documentary,History
4.00
Documentary,History,Music
2.00

Documentary,History,War
1.00
Documentary,Horror
1.00
Documentary,Horror,Short
1.00
Documentary,Music
10.00
Documentary,Short
16.00
Drama
111.00
Drama,Family
53.00
Drama,Family,Fantasy
33.00
Drama,Family,History
1.00
Drama,Family,Music
1.00
Drama,Family,Musical
3.00
Drama,Family,Mystery
2.00
Drama,Family,Romance
37.00
Drama,Fantasy
14.00
Drama,Fantasy,Music
2.00
Drama,Fantasy,Musical

1.00
Drama,Fantasy,Romance
11.00
Drama,History,Music
1.00
Drama,History,War
1.00
Drama,Horror
1.00
Drama,Horror,Mystery
1.00
Drama,Music
2.00
Drama,Music,Romance
8.00
Drama,Musical
2.00
Drama,Musical,Romance
4.00
Drama,Mystery
1.00
Drama,Mystery,Romance
5.00
Drama,Mystery,Thriller
2.00
Drama,Romance
143.00
Drama,Romance,Sport
1.00
Drama,Short
1.00

Drama,Thriller	
	2.00
Drama,War	
	3.00
Drama,Western	
	2.00
Family	
	113.00
Family,Fantasy	
	10.00
Family,Fantasy,Musical	
	2.00
Family,Fantasy,Romance	
	1.00
Family,Music	
	5.00
Family,Music,Musical	
	1.00
Family,Music,Romance	
	1.00
Family,Music,Short	
	1.00
Family,Musical	
	8.00
Family,Musical,News	
	1.00
Family,Mystery,Sport	
	1.00
Family,Romance	
	18.00
Family,Sci-Fi	

1.00
Family,Short
7.00
Fantasy
7.00
Fantasy,Horror,Mystery
2.00
Fantasy,Music
2.00
Fantasy,Romance
5.00
History,Short,War
1.00
Horror
23.00
Horror,Mystery
2.00
Horror,Mystery,Thriller
3.00
Horror,Short
2.00
Horror,Thriller
5.00
Music
19.00
Music,Romance
2.00
Music,Sci-Fi
1.00
Music,Short
6.00

Musical
23.00
Mystery,Romance
3.00
Mystery,Thriller
1.00
Reality-TV,Short
1.00
Romance
128.00
Romance,Sci-Fi
1.00
Romance,Western
1.00
Sci-Fi
2.00
Short
10.00
Short,Talk-Show
1.00
Sport
2.00
Talk-Show
1.00
Thriller
6.00
Western
2.00
NA
32.00

The histogram of "average_rating" is given Figure 2.

The distribution is unimodal and balanced. There are no obvious outliers in the plot.

```
movies_EDA$ave_rate_hist
```

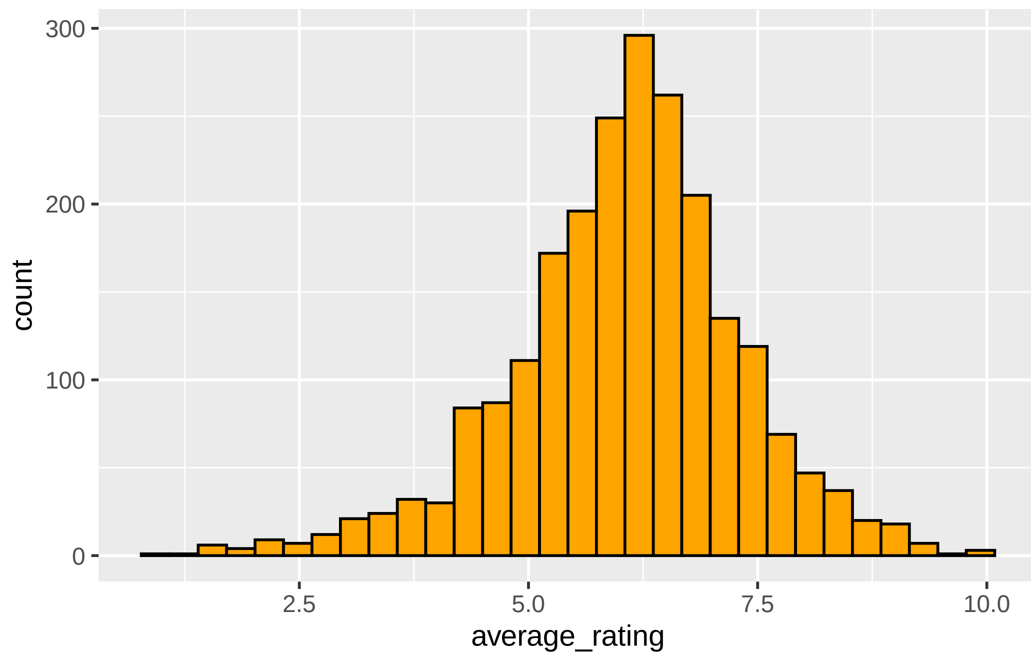

Figure 2: A histogram of average rating.

The histogram of the average rating is given Figure 3.

There are few values of "num_votes" very large but it is reasonable for some movies that have numerous audiences.

We limited the num_votes to less than 10000 and redrew the histogram, given Figure 4. The histogram plot is unimodal and strongly left-skewed. The majority of num_votes had values between 0 and 2500.

```
movies_EDA$num_votes_hist
```

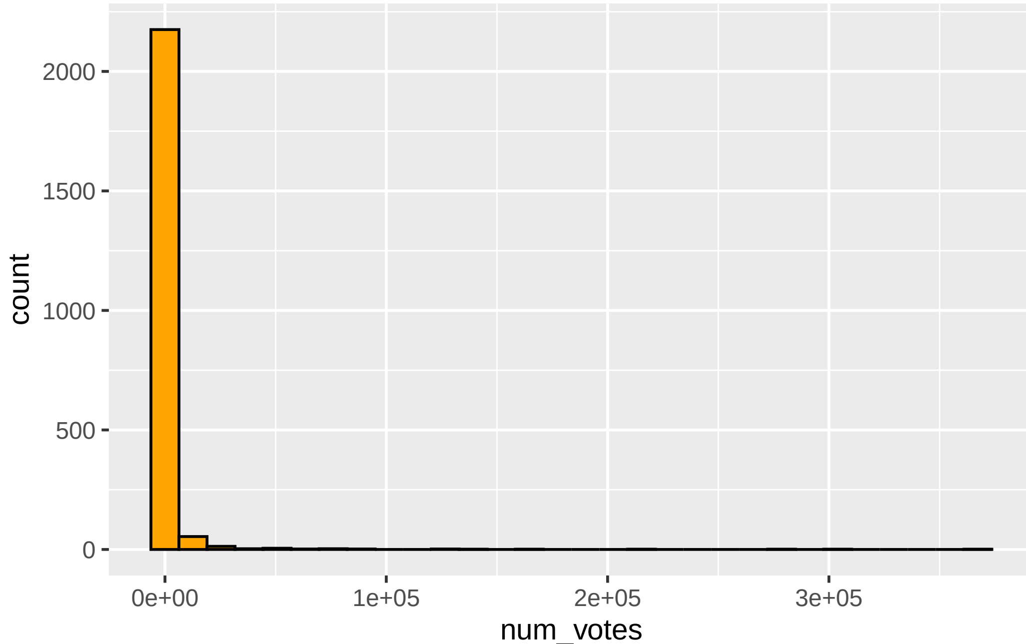

Figure 3: A histogram of the number of votes.

```
movies_EDA$num_votes_hist+xlim(0,10000)
```

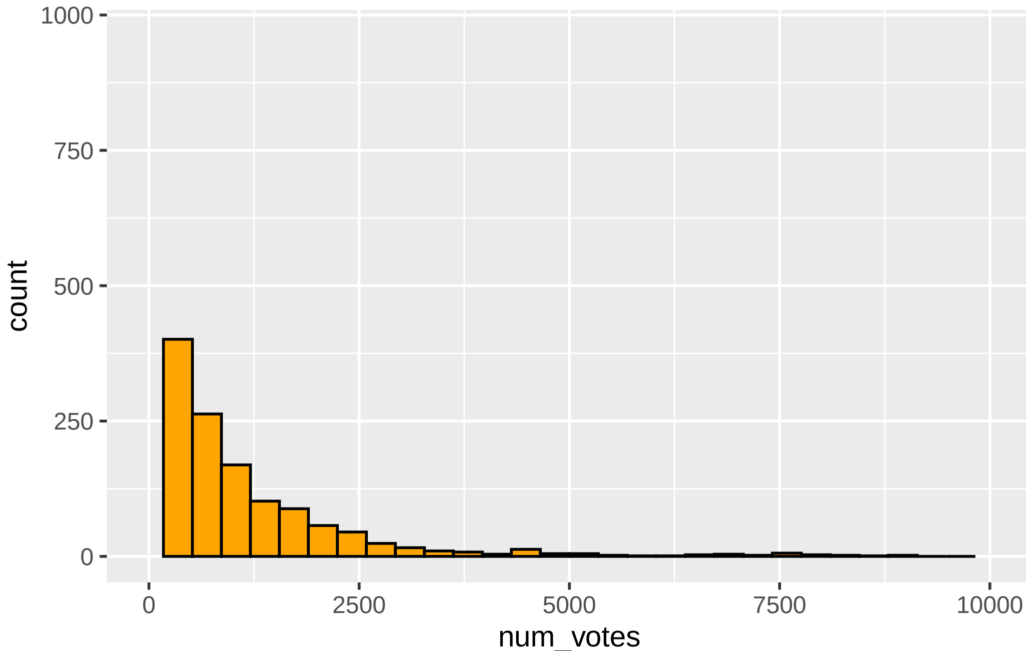

Figure 4: A histogram of the number of votes.

Figure 5 shows the relationship between log(num_votes) and log(average_rating). According to the plot, the average_rating doesn't have homogeneity. As num_votes grows larger the data points become less but more concentrated and there is a slight downward trend in the mode value of average_rating.

```
movies_EDA$scatter_plot
```

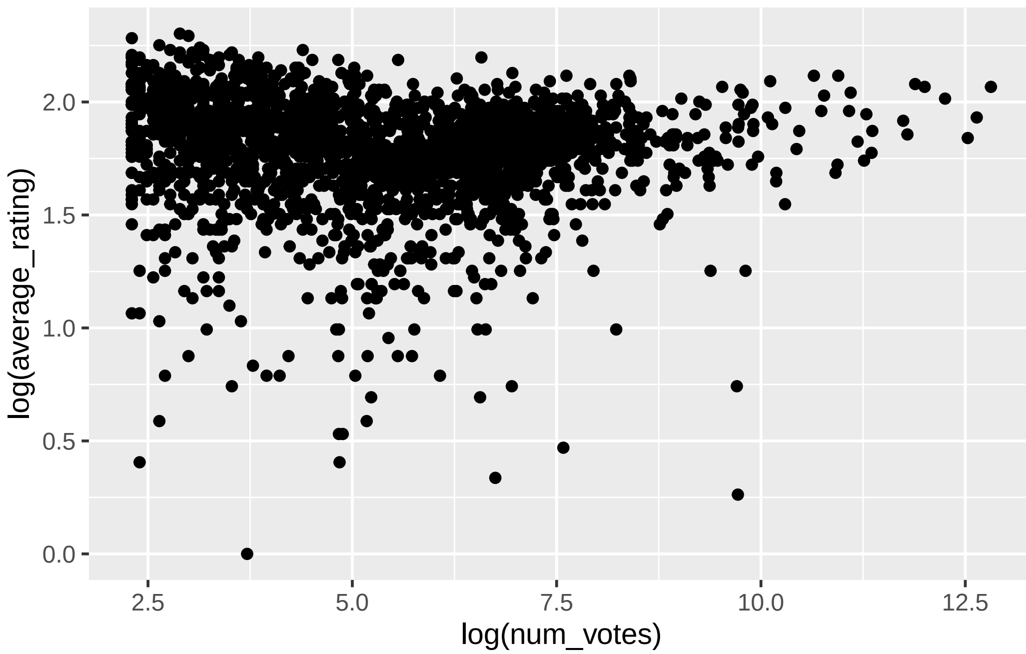

Figure 5: A scatter plot of num_votes vs average_rating

4 Data Clean

We filtered the Christmas movies, selected the variables that will be used in this analysis (genres, average_rating and num_votes), and deleted the observations that contain missing values.

Now the dataset contains 1903 observations and 3 variables. **?@tbl-mov_clean** displays the first 6 observations in the cleaned dataset.

```
tar_load(movies_clean)
head(movies_clean)
```

Then, we create a new dataset with genres split into general classes. For example, observation with genres = “Drama,Family,Fantasy” has been split into 3 observations with the same average_rating” and “num_votes”, but genres = “Drama”, “Family” and “Fantasy” respectively.

The new dataset has 3855 observations and 3 variables. The **?@tbl-mov_split** displays the first 6 observations of the new dataset.

Table 2: **?(caption)**

```
# A tibble: 6 x 3
  genres                      average_rating num_votes
  <chr>                                <dbl>     <dbl>
1 Drama,Family,Fantasy                   7.5      8312
2 Comedy,Drama,Romance                   7.4      4172
3 Crime,Drama,Film-Noir                  6.5      1583
4 Comedy,Romance                         7.3     11196
5 Action,Adventure,Comedy                6.2       515
6 Comedy,Drama                           5.7       805
```

```
#| tbl-cap: movies split
#| label: tbl-mov_split
tar_load(movies_split)
head(movies_split)
```

```
# A tibble: 6 x 3
  genres  average_rating num_votes
  <chr>            <dbl>     <dbl>
1 Drama              7.5      8312
2 Family             7.5      8312
3 Fantasy            7.5      8312
4 Comedy             7.4      4172
5 Drama              7.4      4172
6 Romance            7.4      4172
```

We classified the genres which have rates of occurrence <0.01 as "other", considering the statistical significance. Then, we plot box plots (Figure 6) of average_rating VS genres.

In one box plot, the lowest/ highest data point denotes the minimum/ maximum average rating in the related genres. The upper/ lower edge of the box denotes the 75%/ 25% quantile of the average rating, marked as Q3/ Q1 The line splitting the box denotes the median value and the single points denote the outliers, which are larger/ smaller than Q3+1.5(Q3-Q1) or Q1-1.5(Q3-Q1).

According to Figure 6, the majority of the genres had outliers. Considering the mean values will be significantly influenced by the outliers, we chose median values to compare the average ratings of the genres.

We found the "Documentary" has the highest median value. Therefore, "Documentary" had the best average rating if we only consider the relationship between genres and average ratings.

```r
tar_read(movies_boxplot)
```

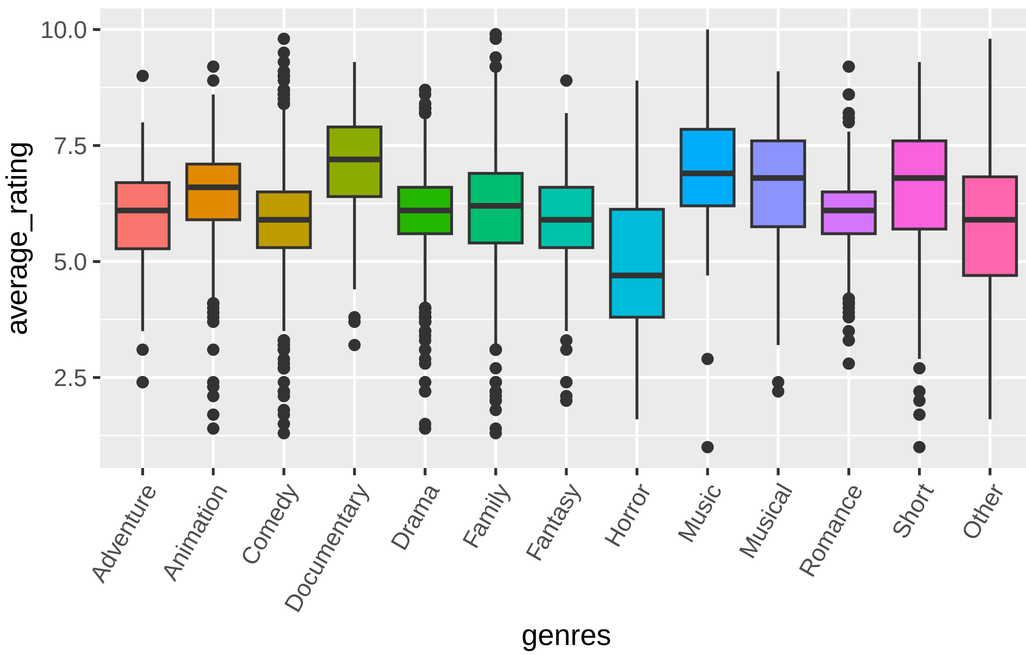

Figure 6: Box plots of average_rating and genres

We calculated the mean num_votes in each genre and found significant differences. For example, the mean of num_votes in "Documentary" is only around 56, which is almost 30 times less than in "Drama". Excessive differences in the number of votes may bias the results of the average ratings. For example, the high average ratings in the genre "Documentary" were probably rated by a few people who particularly like documentaries. Therefore, it is necessary to consider the influence of num_votes, to ensure the genre which has the best average ratings also being widely recognized.

```r
tar_load(numvote_mean)
numvote_mean
```

Table 3: Mean of number of votes

genres	mean
Adventure	5, 145.94
Animation	3, 174.52
Comedy	3, 155.38

Documentary	56.38
Drama	1,676.40
Family	2,792.69
Fantasy	5,765.16
Horror	2,423.75
Music	411.03
Musical	1,613.50
Romance	1,542.73
Short	72.72
Other	2,109.96

We built random forest with the movies_split dataset. The numerical variables (average_ratings, num_votes) are normalized, the genres whose rates of occurrence are less than 0.01 were classified as "other" and dummy variables were created with one hot encoding. We aimed to investigate the interactive influence of the num_votes and genres on average_rating. Therefore, we added the interaction term between the dummy variables and num_votes.

We used 10-fold cross-validation to tune the mtry (i.e. the number of variables randomly sampled at each split) and min_n (the minimum number of data points in a node that can be split further) and the random forest has 100 decision trees.

The model having the minimum rmse was selected. Our best result has mtry=4, min_n=40, trees=100 and the rmse = 1.135.

```
tar_load(movies_rf)
print(movies_rf$model)
```

```
$pre
$actions
$actions$recipe
$recipe

$blueprint
NULL

attr(,"class")
[1] "action_recipe" "action_pre"    "action"


$mold
NULL
```

```
$case_weights
NULL

attr(,"class")
[1] "stage_pre" "stage"

$fit
$actions
$actions$model
$spec
Random Forest Model Specification (regression)

Main Arguments:
  mtry = 4
  trees = 100
  min_n = 40

Engine-Specific Arguments:
  importance = permutation

Computational engine: ranger


$formula
NULL

attr(,"class")
[1] "action_model" "action_fit"   "action"


$fit
NULL

attr(,"class")
[1] "stage_fit" "stage"

$post
$actions
named list()

attr(,"class")
[1] "stage_post" "stage"
```

```
$trained
[1] FALSE

attr(,"class")
[1] "workflow"
```

```
sprintf('rmse=%f' ,movies_rf$rmse)
```

```
[1] "rmse=1.135212"
```

Based on the best model, we calculated the feature importance of each variable. According to Figure Figure 7, the num_votes has significant influence on the average_rating. If we do not consider the interaction terms, the genre "Horror" has the largest positive influence on the average rating. If we consider the interactive influence of num_votes and genres, then the genre "Drama" is the genre with the biggest boost to the average rating.

```
movies_rf$VIP
```

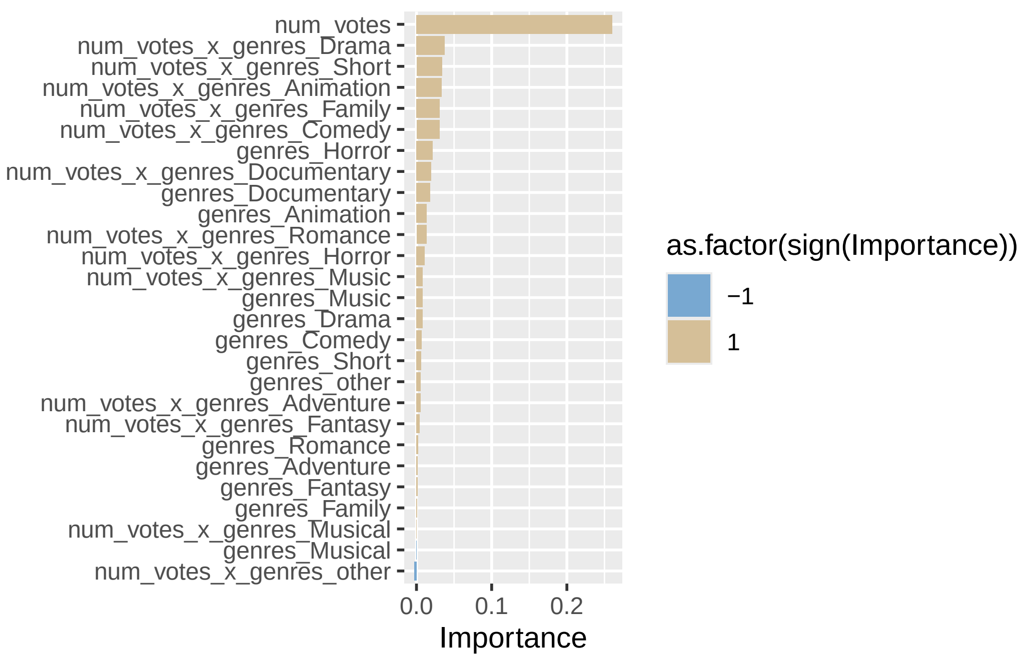

Figure 7: Variable Importance Plot