Topic

1. Please visualize main features of this dataset using ggplot2 package
2. Create a chart with a few panels characterising 3 most important features of this dataset.

A. Data Understanding

The ‘movies’ dataset contains 58788 observations with 24 variables, with one interesting feature is the category values of the film - binary variables representing if the movie was classified as belonging to that genre.

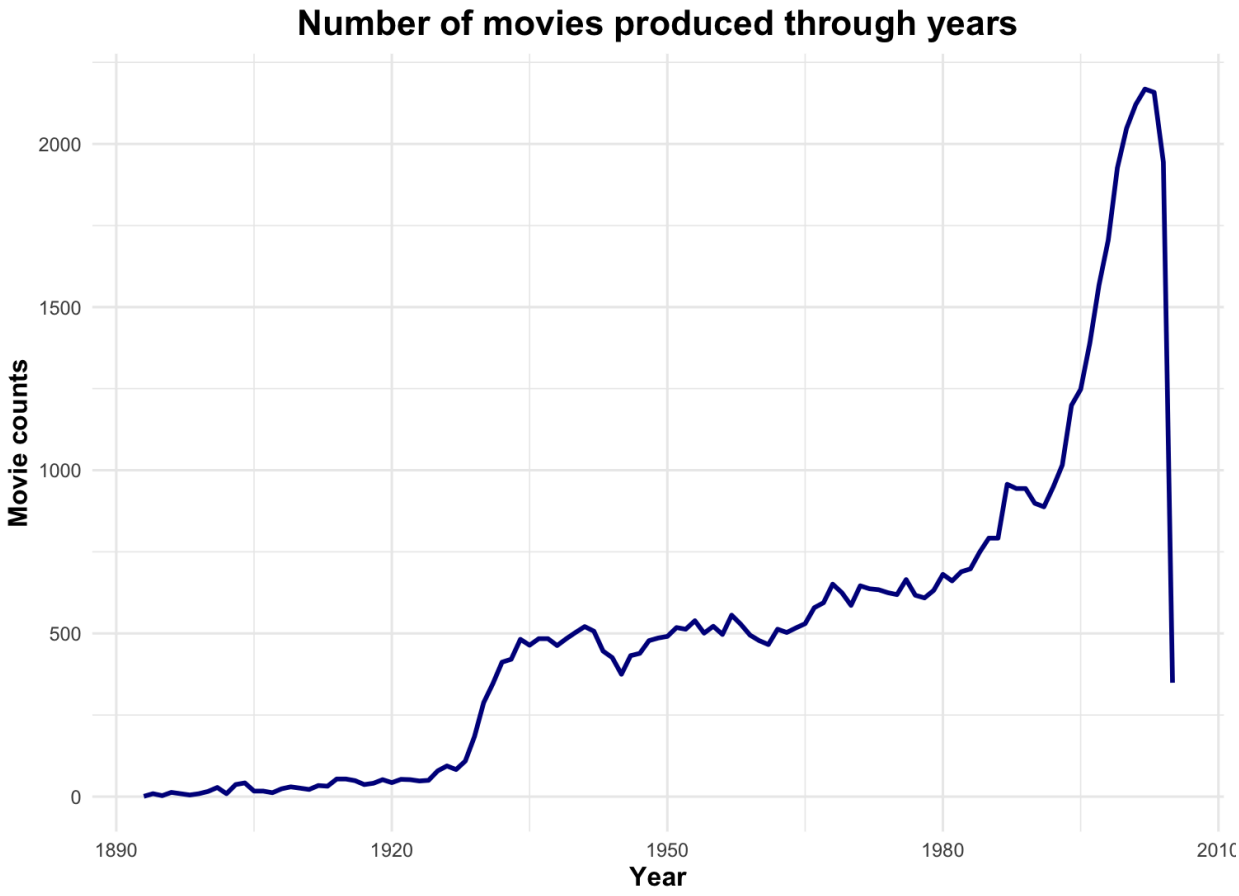
As there are some missing values in `budget` and `mpaa` , I decided not to input those variables into the data visualization.

```
> head(movies)
                     title year length budget rating votes   r1   r2  r3   r4   r5   r6
1                        $ 1971    121     NA    6.4   348  4.5  4.5 4.5  4.5 14.5 24.5
2         $1000 a Touchdown 1939     71     NA    6.0    20  0.0 14.5 4.5 24.5 14.5 14.5
3   $21 a Day Once a Month 1941      7     NA    8.2     5  0.0  0.0 0.0  0.0  0.0 24.5
4                  $40,000 1996     70     NA    8.2     6 14.5  0.0 0.0  0.0  0.0  0.0
5 $50,000 Climax Show, The 1975     71     NA    3.4    17 24.5  4.5 0.0 14.5 14.5  4.5
6                    $pent 2000     91     NA    4.3    45  4.5  4.5 4.5 14.5 14.5 14.5
    r7   r8   r9  r10 mpaa Action Animation Comedy Drama Documentary Romance Short
1 24.5 14.5  4.5  4.5           0         0      1     1           0       0     0
2 14.5  4.5  4.5 14.5           0         0      1     0           0       0     0
3  0.0 44.5 24.5 24.5           0         1      0     0           0       0     1
4  0.0  0.0 34.5 45.5           0         0      1     0           0       0     0
5  0.0  0.0  0.0 24.5           0         0      0     0           0       0     0
6  4.5  4.5 14.5 14.5           0         0      0     1           0       0     0
```

```
> str(movies)
'data.frame':   58788 obs. of  24 variables:
 $ title      : chr  "$" "$1000 a Touchdown" "$21 a Day Once a Month" "$40,000" ...
 $ year       : int  1971 1939 1941 1996 1975 2000 2002 2002 1987 1917 ...
 $ length     : int  121 71 7 70 71 91 93 25 97 61 ...
 $ budget     : int  NA NA NA NA NA NA NA NA NA NA ...
 $ rating     : num  6.4 6 8.2 8.2 3.4 4.3 5.3 6.7 6.6 6 ...
 $ votes      : int  348 20 5 6 17 45 200 24 18 51 ...
 $ r1         : num  4.5 0 0 14.5 24.5 4.5 4.5 4.5 4.5 4.5 ...
 $ r2         : num  4.5 14.5 0 0 4.5 4.5 0 4.5 4.5 0 ...
 $ r3         : num  4.5 4.5 0 0 0 4.5 4.5 4.5 4.5 4.5 ...
 $ r4         : num  4.5 24.5 0 0 14.5 14.5 4.5 4.5 0 4.5 ...
 $ r5         : num  14.5 14.5 0 0 14.5 14.5 24.5 4.5 0 4.5 ...
 $ r6         : num  24.5 14.5 24.5 0 4.5 14.5 24.5 14.5 0 44.5 ...
 $ r7         : num  24.5 14.5 0 0 0 4.5 14.5 14.5 34.5 14.5 ...
 $ r8         : num  14.5 4.5 44.5 0 0 4.5 4.5 14.5 14.5 4.5 ...
 $ r9         : num  4.5 4.5 24.5 34.5 0 14.5 4.5 4.5 4.5 4.5 ...
 $ r10        : num  4.5 14.5 24.5 45.5 24.5 14.5 14.5 14.5 24.5 4.5 ...
 $ mpaa       : chr  "" "" "" "" ...
 $ Action     : int  0 0 0 0 0 0 1 0 0 0 ...
 $ Animation  : int  0 0 1 0 0 0 0 0 0 0 ...
 $ Comedy     : int  1 1 0 1 0 0 0 0 0 0 ...
 $ Drama      : int  1 0 0 0 0 1 1 0 1 0 ...
 $ Documentary: int  0 0 0 0 0 0 0 1 0 0 ...
 $ Romance    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Short      : int  0 0 1 0 0 0 0 1 0 0 ...
```

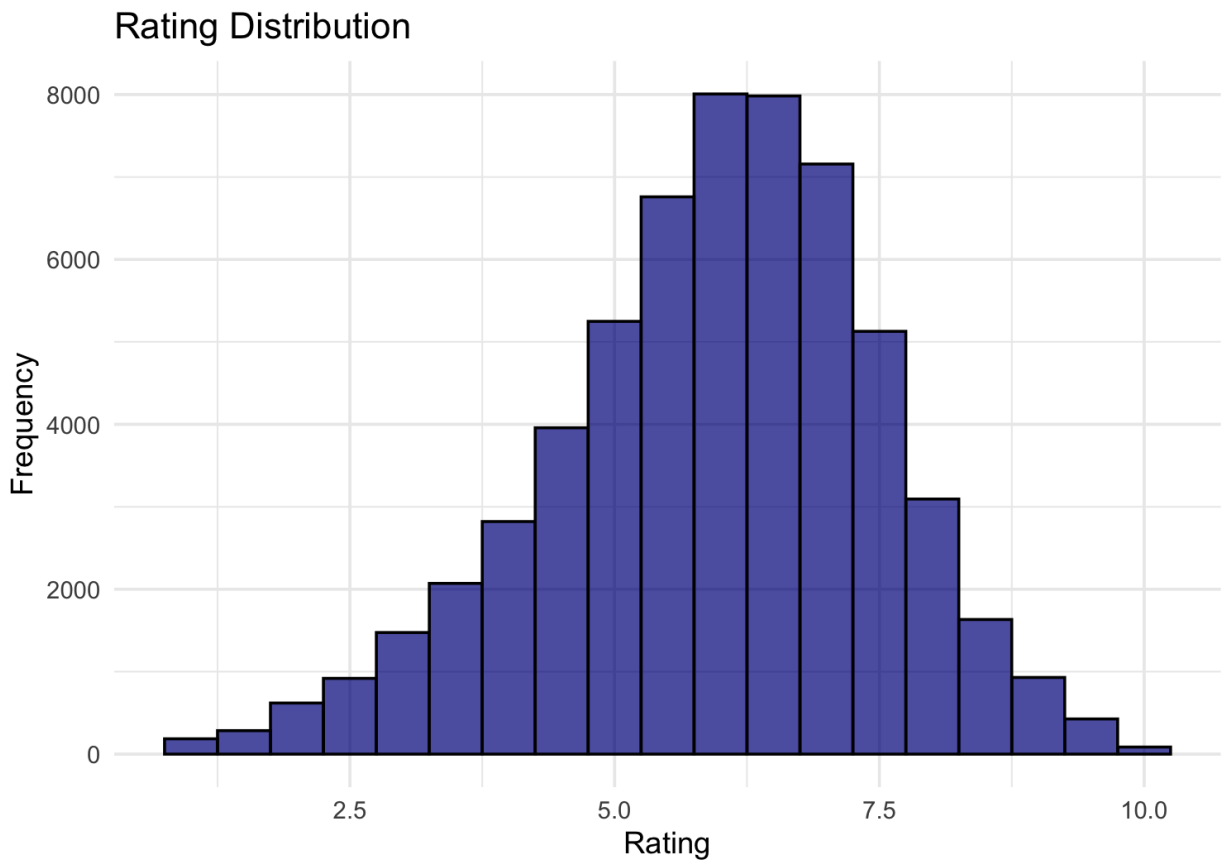
B. Data Visualization

1. Overall data



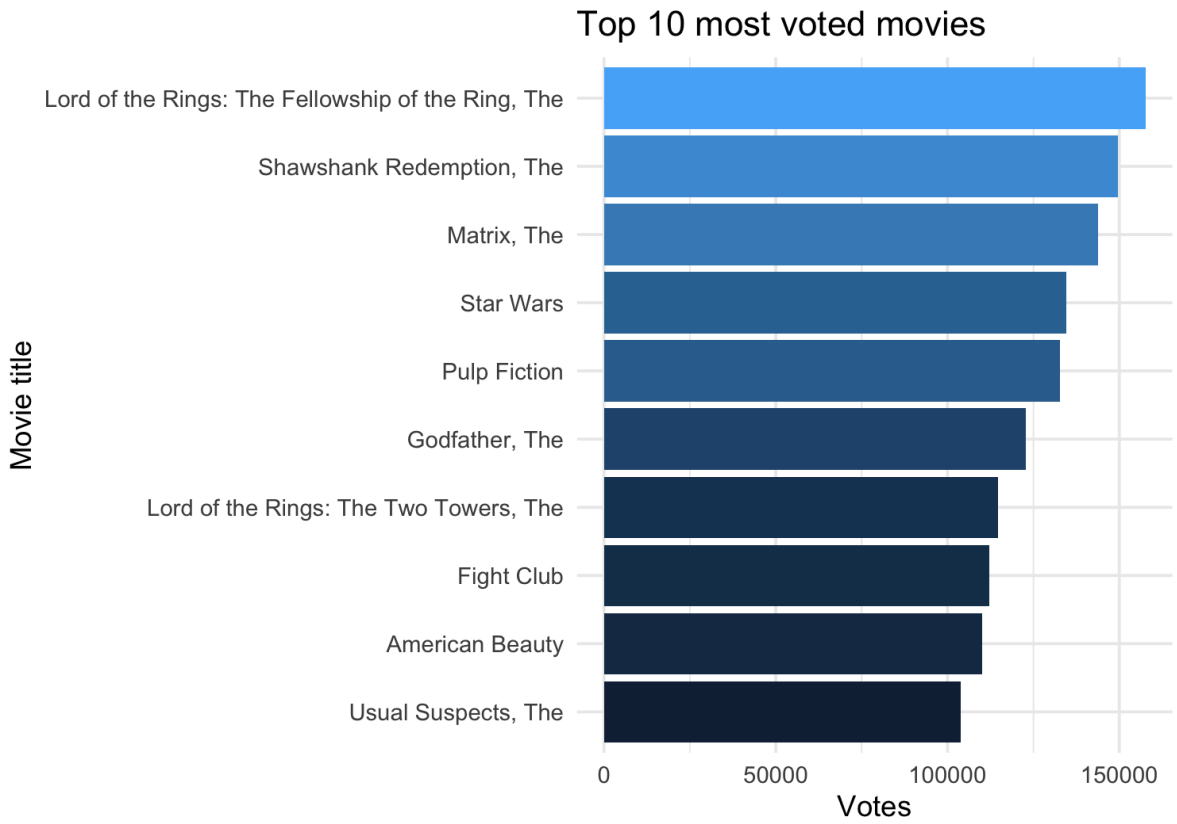
Graph 1.

The number of films produced increased steadily until around 2000, reaching nearly 2,500 films a year, then it dropped sharply to below 400.

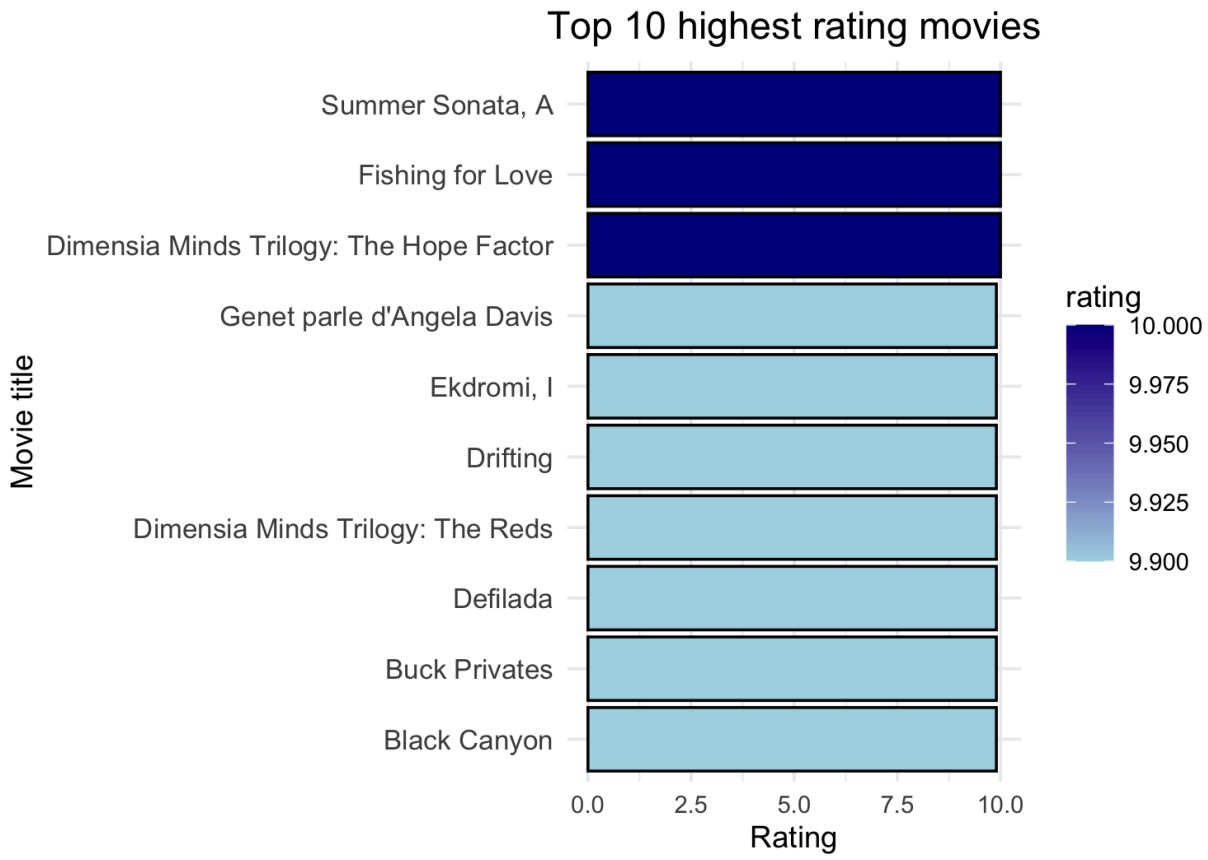


Graph 2.

The most common rating is 6-6.5, a slightly above average score, but the ratings are fairly evenly distributed.


Top 10 most voted movies
Lord of the Rings: The Fellowship of the Ring, The
Shawshank Redemption, The
Matrix, The
Star Wars
Pulp Fiction
Godfather, The
Lord of the Rings: The Two Towers, The
Fight Club
American Beauty
Usual Suspects, The
Movie title
0
50000
100000
150000
Votes

Graph 3.

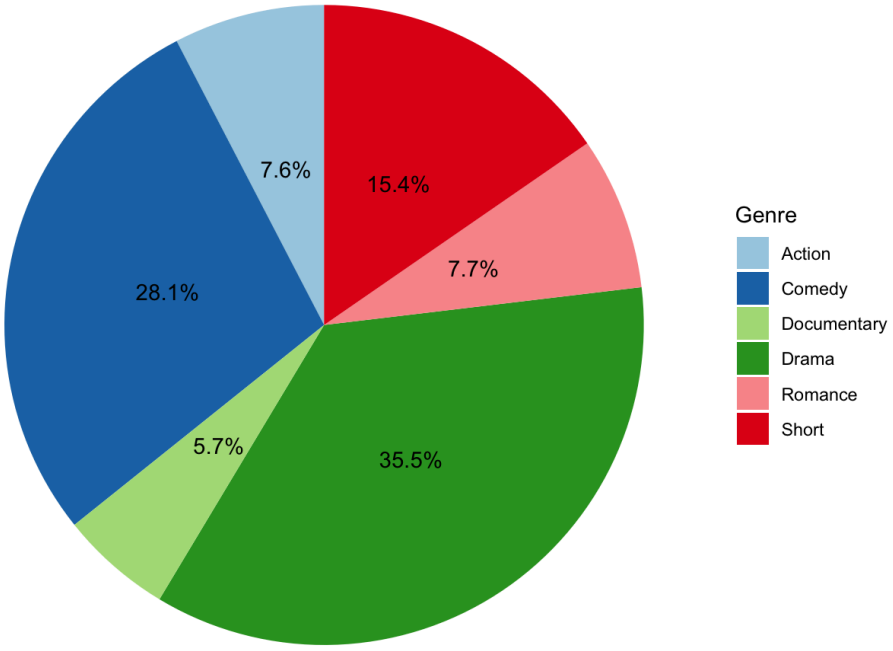


Graph 4.

Interestingly, the top 10 highest rated movies are not in the top 10 most voted movies. This may help me hypothesize that the movies that receive the most votes are famous movies, known by many people, but not everyone loves them, so the ratings are not high.

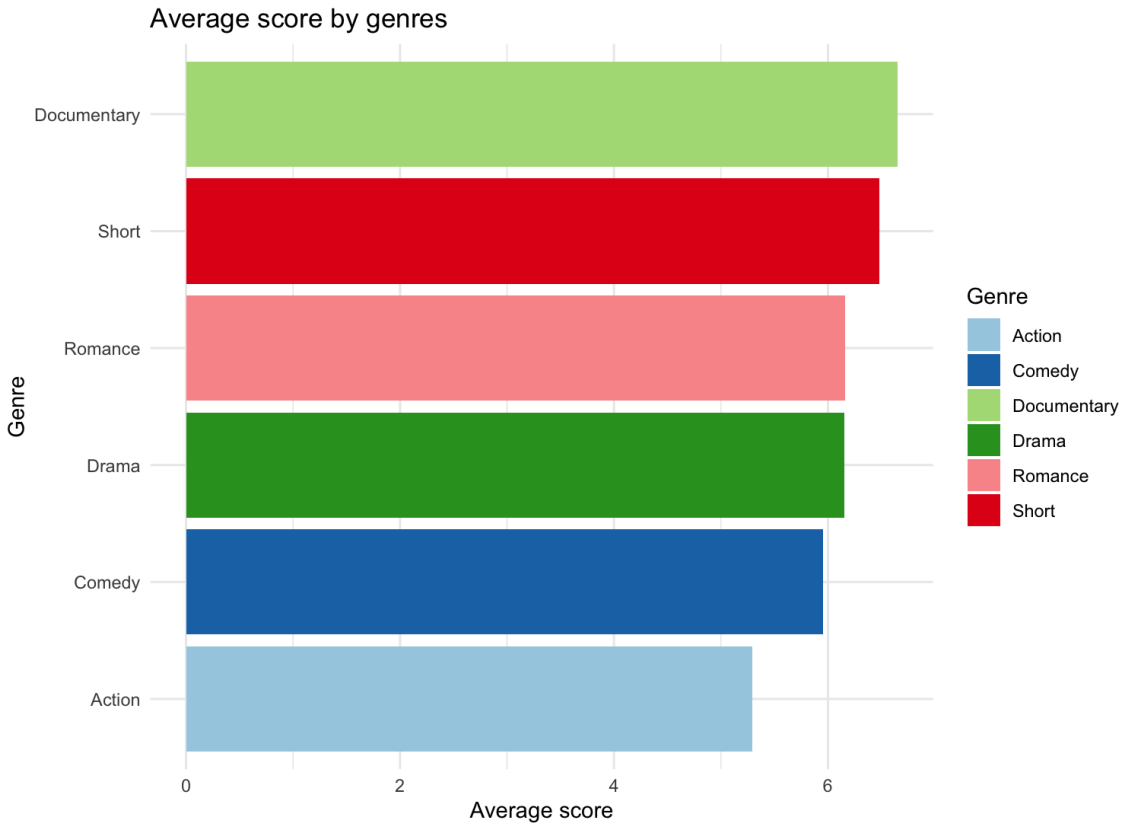
2. By genres

Movie distribution by genres



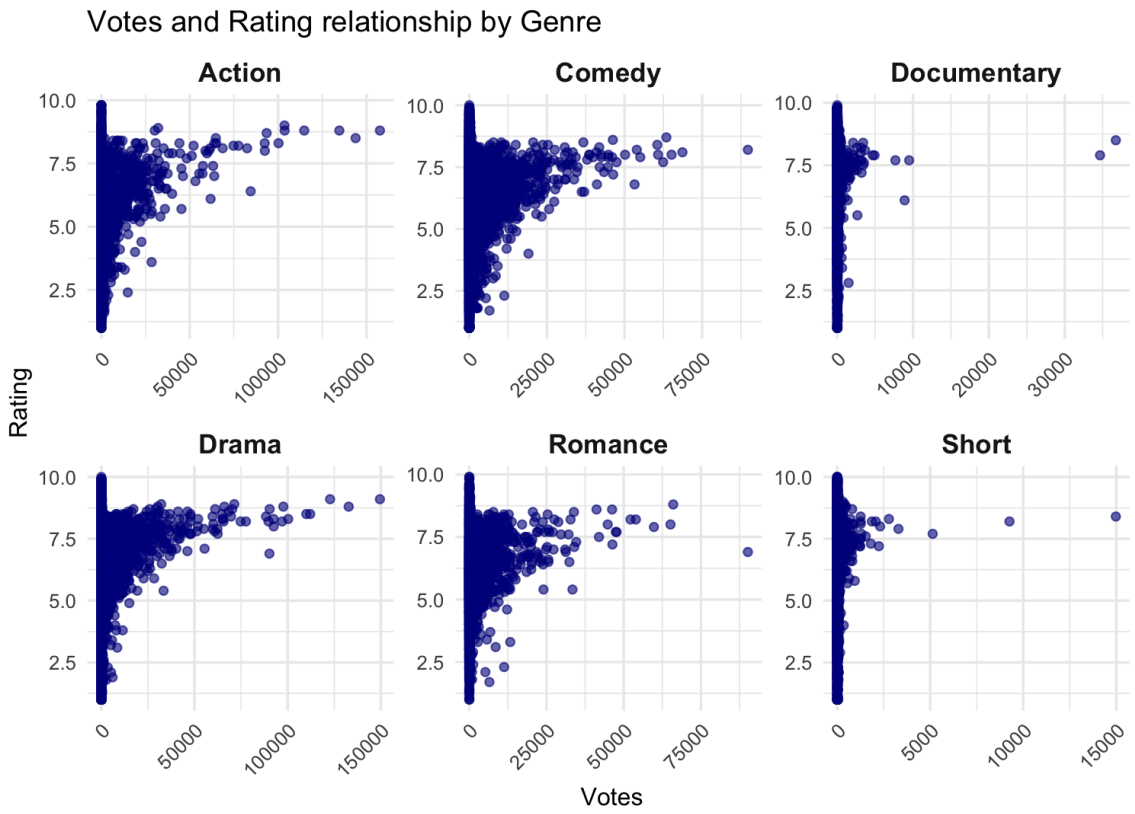
Graph 5.

Drama genre accounts for more than ⅓ of the total number of films produced, which shows that the audience has been and is very fond of this genre, so film producers have grasped the taste and continued to produce more.



Graph 6.

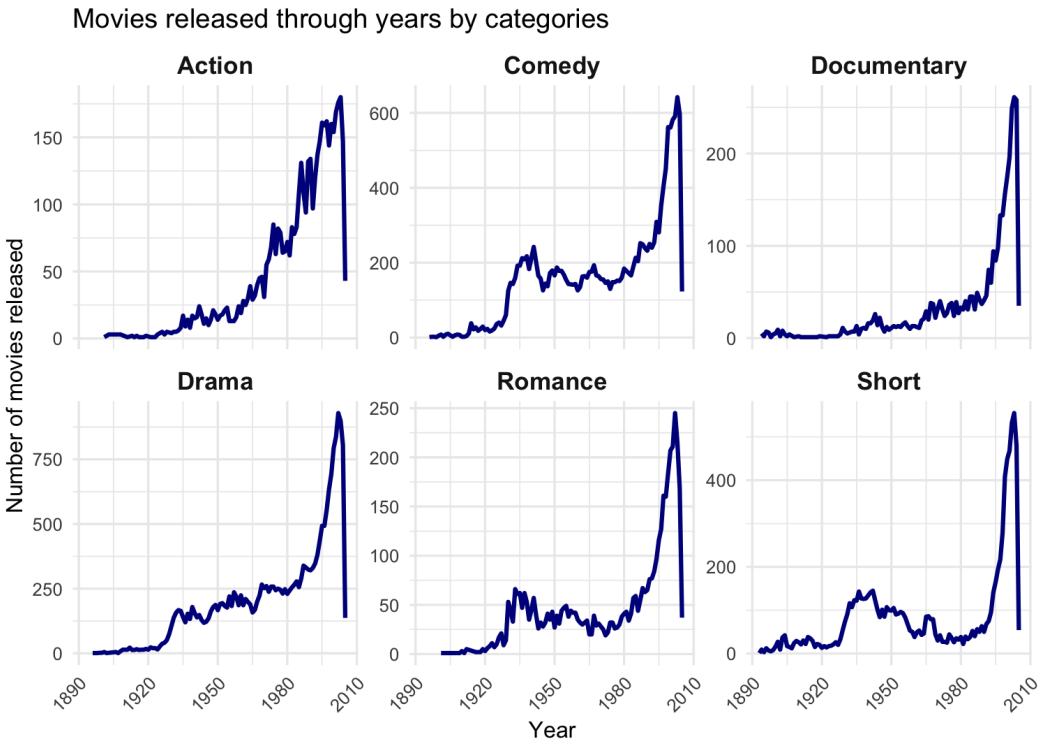
Although drama is the most produced genre, action is the most popular film based on average score, followed by short. Drama genre only ranked 4th in the average score ranking of 6 genres.



Graph 7.

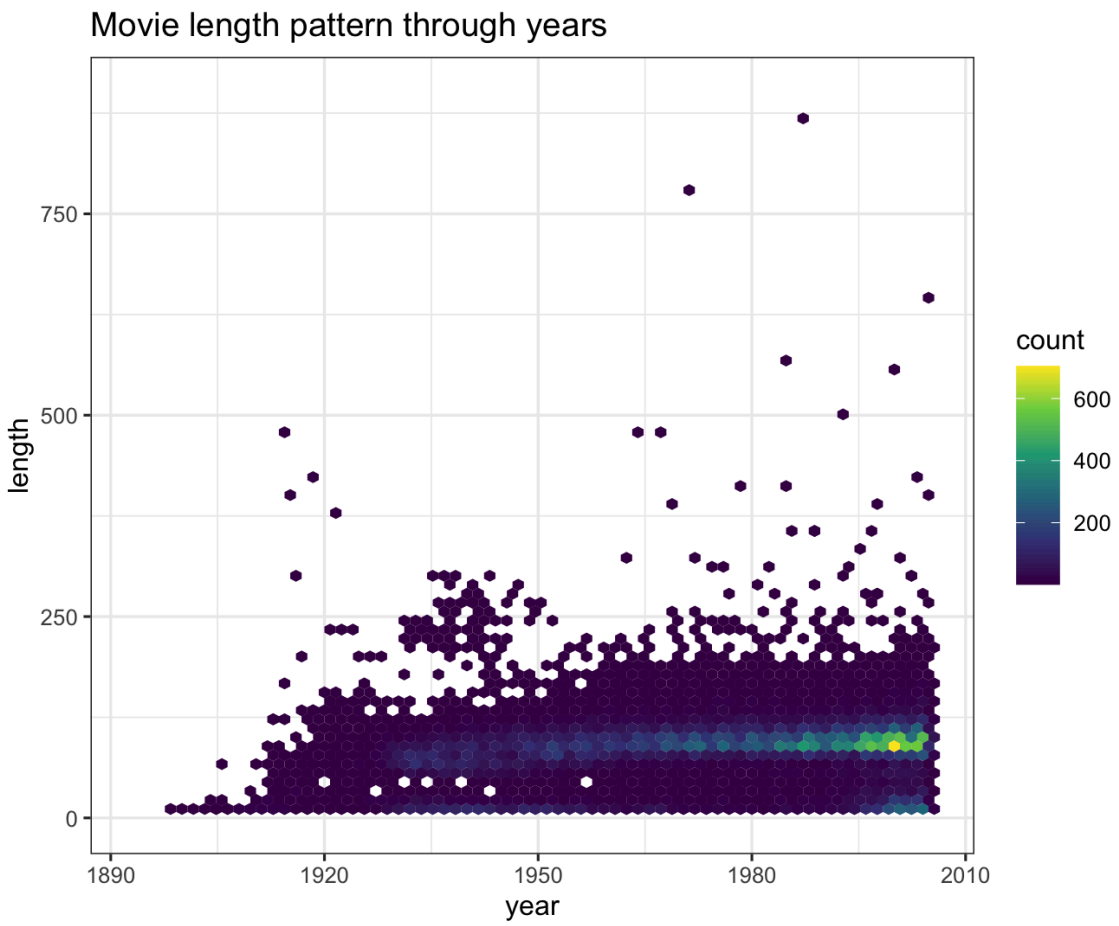
Drama had the highest number of votes among the genres, followed closely by Action; and Short was the genre with the lowest votes, possibly due to its short length, which did not create a strong enough impression on the audience. Action and Drama had a fairly similar data dispersion, both focusing on ratings from 5.0-8.5.

3. Over time

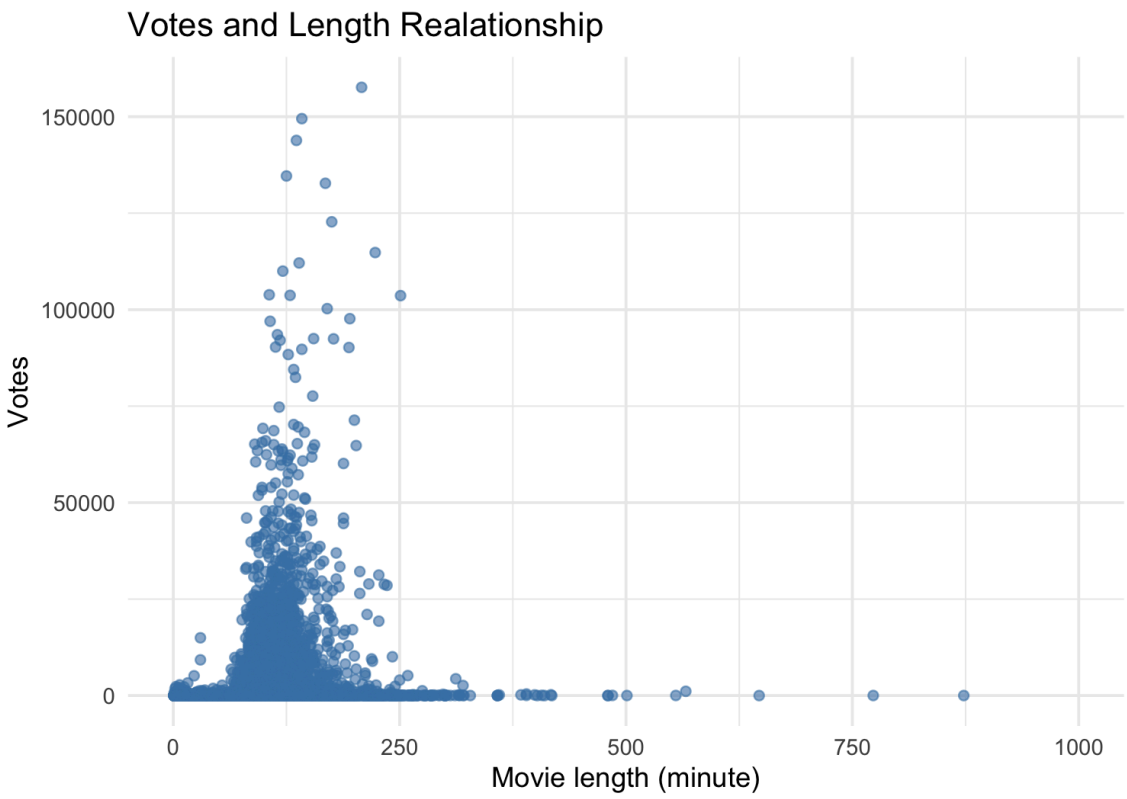


Graph 8.

All categories have a common trend of increasing network around 2000 and then decreasing sharply.



Graph 9.
Most films produced are under 250 minutes long, from 1980 onwards, film producers almost always prefer a length of about 125 minutes for their films.



Graph 10.

The highest number of votes was recorded from films with a length of 75-180 minutes, which shows that this film length is the ideal range to receive the attention of the audience.