Exercise:

- Analyse dataset movies
- Visualize main features of this dataset using ggplot2 package
- Create a chart with a few panels characterising 3 most important features of this dataset.

I. Data overview

```
   title               year          length            budget
Length:58788       Min.   :1893   Min.   :   1.00   Min.   :        0
Class :character   1st Qu.:1958   1st Qu.:  74.00   1st Qu.:   250000
Mode  :character   Median :1983   Median :  90.00   Median :  3000000
                   Mean   :1976   Mean   :  82.34   Mean   : 13412513
                   3rd Qu.:1997   3rd Qu.: 100.00   3rd Qu.: 15000000
                   Max.   :2005   Max.   :5220.00   Max.   :200000000
                                                    NA's   :53573
    rating           votes                r1                r2
Min.   : 1.000   Min.   :     5.0   Min.   :  0.000   Min.   : 0.000
1st Qu.: 5.000   1st Qu.:    11.0   1st Qu.:  0.000   1st Qu.: 0.000
Median : 6.100   Median :    30.0   Median :  4.500   Median : 4.500
Mean   : 5.933   Mean   :   632.1   Mean   :  7.014   Mean   : 4.022
3rd Qu.: 7.000   3rd Qu.:   112.0   3rd Qu.:  4.500   3rd Qu.: 4.500
Max.   :10.000   Max.   :157608.0   Max.   :100.000   Max.   :84.500

      r3               r4                r5               r6
Min.   : 0.000   Min.   :  0.000   Min.   :  0.000   Min.   : 0.00
1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:  4.500   1st Qu.: 4.50
Median : 4.500   Median :  4.500   Median :  4.500   Median :14.50
Mean   : 4.721   Mean   :  6.375   Mean   :  9.797   Mean   :13.04
3rd Qu.: 4.500   3rd Qu.:  4.500   3rd Qu.: 14.500   3rd Qu.:14.50
Max.   :84.500   Max.   :100.000   Max.   :100.000   Max.   :84.50

      r7               r8               r9               r10
Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
1st Qu.:  4.50   1st Qu.:  4.50   1st Qu.:  4.500   1st Qu.:  4.50
Median : 14.50   Median : 14.50   Median :  4.500   Median : 14.50
Mean   : 15.55   Mean   : 13.88   Mean   :  8.954   Mean   : 16.85
3rd Qu.: 24.50   3rd Qu.: 24.50   3rd Qu.: 14.500   3rd Qu.: 24.50
Max.   :100.00   Max.   :100.00   Max.   :100.000   Max.   :100.00
```

```
    mpaa               Action          Animation           Comedy
Length:58788       Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
Class :character   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
Mode  :character   Median :0.00000   Median :0.00000   Median :0.0000
                   Mean   :0.07974   Mean   :0.06277   Mean   :0.2938
                   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
                   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000

    Drama        Documentary         Romance           Short
Min.   :0.000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.000   Median :0.00000   Median :0.0000   Median :0.0000
Mean   :0.371   Mean   :0.05906   Mean   :0.0807   Mean   :0.1609
3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
```

The internet movie database imdb.com is a website devoted to collecting movie data supplied by studios and fans. The data frame contains 28819 rows and 24 variables:

- title: Title of the movie
- year: Year of release
- budget: Total budget (if known) in US dollars
- length: Duration of the movies in minutes
- rate: Average IMDB user rating
- votes: Number of IMDB user who rated this movie
- r1-10: Multiplying by ten gives percentile (to nearest 10%) of users who rated this movie a 1.
- mpaa: MPAA rating
- action, animation, comedy, drama, documentary, romance, short: Binary variables representing if a movie was classified as belonging to that genre.

Notably, the “budget” column contains missing data as there are movies’ budgets that are unknown. In addition, “mpaa” column also has a lot of blank cells.

II. Data processing

1. **Create “category” column**
2. **Drop binary category columns**
3. **Edit data type of the remaining variables**
4. **Create “unknown” value for missing value in the “budget” column**

After these steps, we have a data frame as follow:

df

Filter

	title	year	length	budget	rating	votes	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	mpaa	category
1	$	1971	121	Unknown	6.4	348	4.5	4.5	4.5	4.5	14.5	24.5	24.5	14.5	4.5	4.5		Comedy
2	$1000 a Touchdown	1939	71	Unknown	6.0	20	0.0	14.5	4.5	24.5	14.5	14.5	14.5	4.5	4.5	14.5		Comedy
3	$21 a Day Once a Month	1941	7	Unknown	8.2	5	0.0	0.0	0.0	0.0	0.0	24.5	0.0	44.5	24.5	24.5		Animation
4	$40,000	1996	70	Unknown	8.2	6	14.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	34.5	45.5		Comedy
5	$50,000 Climax Show, The	1975	71	Unknown	3.4	17	24.5	4.5	0.0	14.5	14.5	4.5	0.0	0.0	0.0	24.5		Other
6	$pent	2000	91	Unknown	4.3	45	4.5	4.5	4.5	14.5	14.5	14.5	4.5	4.5	14.5	14.5		Drama
7	$windle	2002	93	Unknown	5.3	200	4.5	0.0	4.5	4.5	24.5	24.5	14.5	4.5	4.5	14.5	R	Action
8	'15'	2002	25	Unknown	6.7	24	4.5	4.5	4.5	4.5	4.5	14.5	14.5	14.5	4.5	14.5		Documentary
9	'38	1987	97	Unknown	6.6	18	4.5	4.5	4.5	0.0	0.0	0.0	34.5	14.5	4.5	24.5		Drama
10	'49-'17	1917	61	Unknown	6.0	51	4.5	0.0	4.5	4.5	4.5	44.5	14.5	4.5	4.5	4.5		Other
11	'68	1988	99	Unknown	5.4	23	4.5	0.0	4.5	14.5	24.5	4.5	24.5	4.5	14.5	4.5		Drama
12	'94 du bi dao zhi qing	1994	96	Unknown	5.9	53	4.5	0.0	4.5	4.5	4.5	14.5	24.5	14.5	4.5	24.5		Other
13	'7' Motorist, The	1906	10	Unknown	7.0	44	4.5	0.0	0.0	0.0	4.5	14.5	34.5	14.5	4.5	14.5		Comedy
14	'A'	1965	10	Unknown	6.7	11	14.5	0.0	0.0	0.0	0.0	14.5	4.5	4.5	14.5	24.5		Short
15	'A' gai waak	1983	106	Unknown	7.1	1259	4.5	4.5	4.5	4.5	4.5	14.5	14.5	14.5	14.5	24.5	PG-13	Action
16	'A' gai waak juk jaap	1987	101	Unknown	7.2	614	4.5	4.5	4.5	4.5	4.5	14.5	24.5	24.5	14.5	14.5	PG-13	Action
17	'Breaker' Morant	1980	107	Unknown	7.9	2718	4.5	4.5	4.5	4.5	4.5	4.5	14.5	24.5	24.5	24.5		Drama
18	'Bullitt': Steve McQueen's Commitment to Reality	1968	10	Unknown	6.6	37	0.0	0.0	4.5	4.5	4.5	14.5	44.5	14.5	4.5	4.5		Documentary
19	'Crocodile' Dundee II	1988	110	Unknown	5.0	7252	4.5	4.5	4.5	14.5	24.5	24.5	14.5	4.5	4.5	4.5		Action
20	'E'	1981	7	Unknown	8.6	15	0.0	0.0	0.0	0.0	4.5	4.5	0.0	14.5	24.5	45.5		Animation
21	'El Chicko' - der Verdacht	1995	90	Unknown	3.9	10	24.5	14.5	24.5	14.5	0.0	14.5	14.5	0.0	0.0	24.5		Comedy
22	'G' Men	1935	85	450000	7.2	281	0.0	4.5	4.5	4.5	4.5	14.5	34.5	34.5	4.5	4.5		Drama
23	'Gator Bait	1974	88	Unknown	3.5	100	14.5	14.5	24.5	14.5	14.5	4.5	4.5	4.5	4.5	4.5		Action
24	'Gator Bait II: Cajun Justice	1988	95	Unknown	3.1	54	24.5	14.5	4.5	14.5	4.5	14.5	0.0	4.5	0.0	4.5		Other
25	'High Sign', The	1921	18	Unknown	7.8	145	0.0	0.0	4.5	0.0	4.5	14.5	24.5	24.5	24.5	14.5		Comedy
26	'Hukkunud Alpinisti' hotell	1979	80	Unknown	7.7	45	4.5	0.0	4.5	4.5	4.5	4.5	24.5	14.5	14.5	34.5		Other
27	'Hyp-Nut-Tist', The	1935	6	Unknown	6.2	18	4.5	0.0	0.0	24.5	0.0	24.5	14.5	14.5	14.5	0.0		Animation
28	'I Do...'	1989	86	Unknown	6.2	10	0.0	0.0	0.0	24.5	24.5	14.5	24.5	14.5	14.5	14.5		Other
29	'I Know Where I'm Going!'	1945	92	Unknown	7.7	825	4.5	4.5	4.5	4.5	4.5	4.5	14.5	14.5	14.5	34.5		Drama
30	'Java Madness' formerly titled 'Coffee Madness'	1995	7	Unknown	6.8	10	0.0	14.5	0.0	0.0	14.5	14.5	34.5	0.0	14.5	34.5		Comedy
31	'Je vous salue, Marie'	1985	105	Unknown	6.4	322	4.5	4.5	4.5	4.5	4.5	14.5	14.5	14.5	14.5	14.5		Drama
32	'Kaash'	1987	140	Unknown	5.9	13	4.5	4.5	0.0	0.0	24.5	24.5	0.0	24.5	0.0	14.5		Drama

Showing 1 to 33 of 58,788 entries, 18 total columns

5. Create sub table for rating and votes of each movie (rate_votes)

title	year	rating	votes
$	1971	6.4	348
$1000 a Touchdown	1939	6.0	20
$21 a Day Once a Month	1941	8.2	5
$40,000	1996	8.2	6
$50,000 Climax Show, The	1975	3.4	17
$pent	2000	4.3	45
$windle	2002	5.3	200
'15'	2002	6.7	24
'38	1987	6.6	18
'49-'17	1917	6.0	51

6. Create sub table for number and rating of category (genre_summary)

category	count	avg_rate	avg_votes
Action	4688	5.292022	2086.72440
Animation	3606	6.590627	311.25901
Comedy	14269	5.872612	730.96979
Documentary	3183	6.656645	85.81464
Drama	16952	6.189352	689.87394
Other	12786	5.448045	285.18622
Romance	580	6.058276	264.96379
Short	2724	6.287372	21.49339

7. **Create sub table about rates and number of movies each year (yearly_trends)**

year	avg_rate	movie_count
1893	7.000000	1
1894	4.888889	9
1895	5.500000	3
1896	5.269231	13
1897	4.677778	9
1898	5.040000	5

8. **Create sub table about movies duration and movies rating (length_analysis)**
    - Short: Duration under 60 mins
    - Medium: Duration between 60 mins and 120 mins

- Long: Duration more than 120 mins

length_category	avg_rate	movie_count
Long	6.792058	3752
Medium	5.738166	44220
Short	6.430741	10816

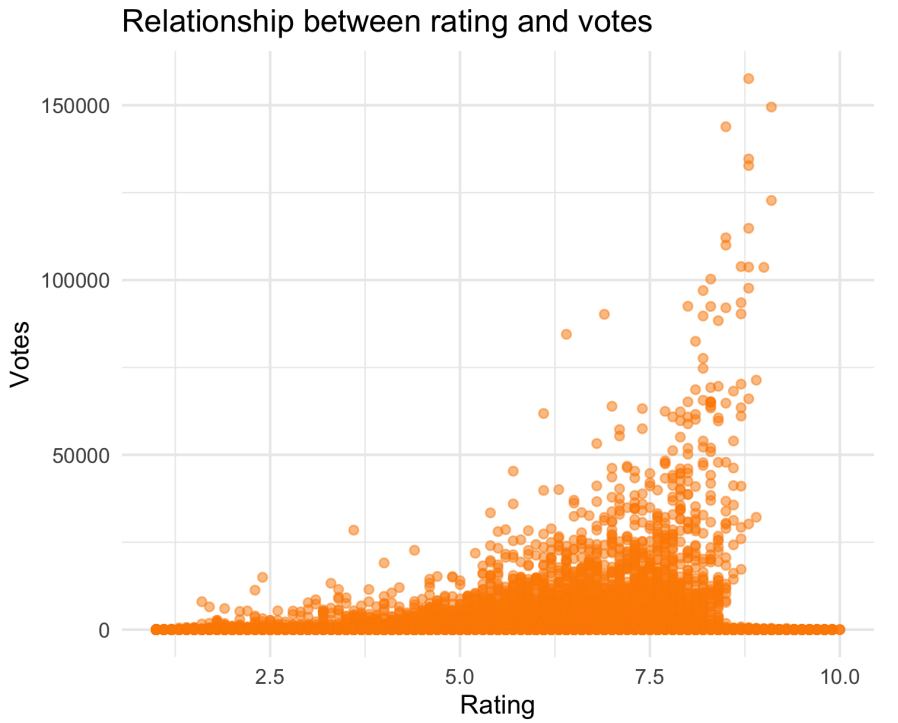
9. **Create sub table for top movies in the dataset (top_movies)**

title	year	rating	votes
Dimensia Minds Trilogy: The Hope Factor	2004	10.0	5
Fishing for Love	2001	10.0	5
Summer Sonata, A	2004	10.0	5
Drifting	2004	9.9	41
Himala	1982	9.9	26
Ivan Groznyy III	1988	9.9	26
New World, The	1982	9.9	24
Dimensia Minds Trilogy: The Reds	2004	9.9	15
Titanic vals	1964	9.9	13
Plight of Clownana, The	2004	9.9	11

10. **Create sub table for top movies in the dataset with votes more than 30 (top_movies_filtered)**

title	year	rating	votes
Drifting	2004	9.9	41
Cashback	2004	9.8	51
Glissando	1985	9.8	37
Goodnite Charlie	2005	9.8	34
Tis kakomoiras	1963	9.6	86
Classic Queen	1992	9.6	39
Da no tien gu	1965	9.5	94
Punto y raya	2004	9.5	62

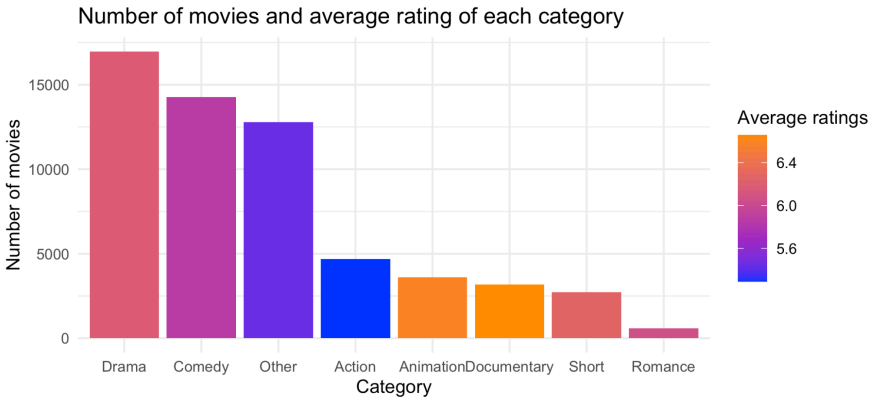
III. Data analysis



Movies with ratings between 6-8 usually receive the most votes

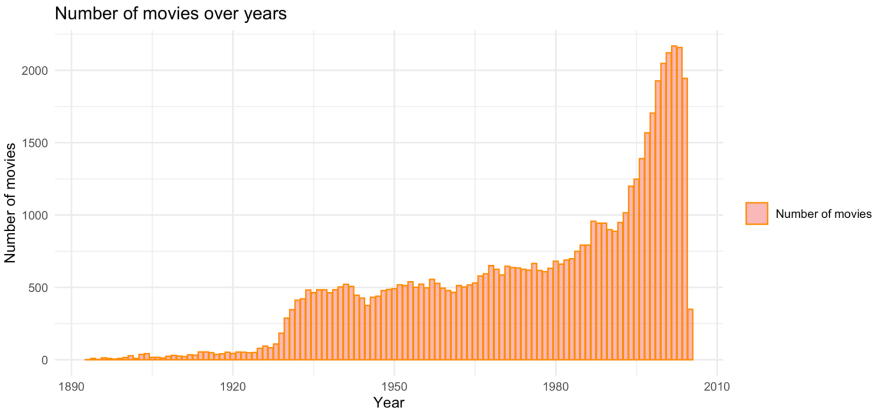
Movies with low ratings have few views, which is understandable because if the movie is not attractive, it will have few viewers

Interestingly, movies with very high ratings (>9) also do not receive many votes, maybe because these movies are selective of viewers

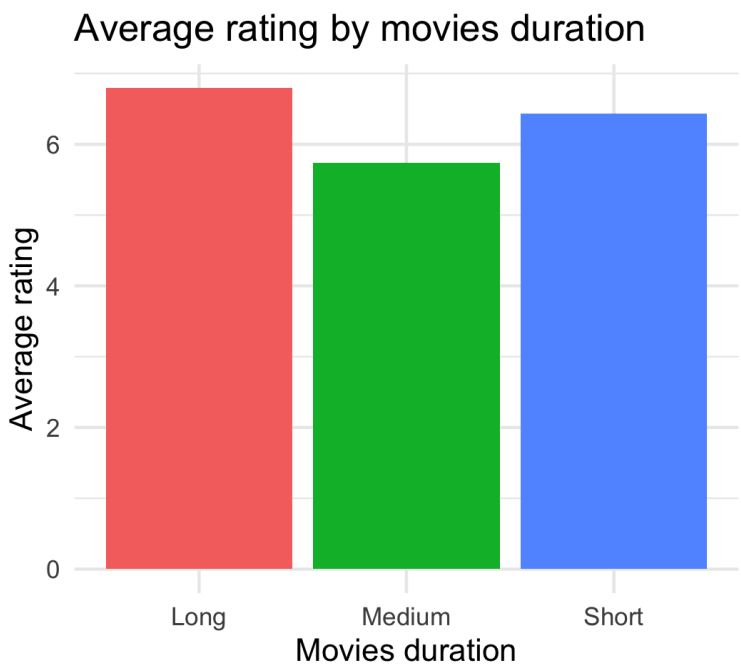


It can be seen that drama and comedy are the two most popular genres, receiving the most ratings

The genres that receive the highest ratings are animation and documentary, short

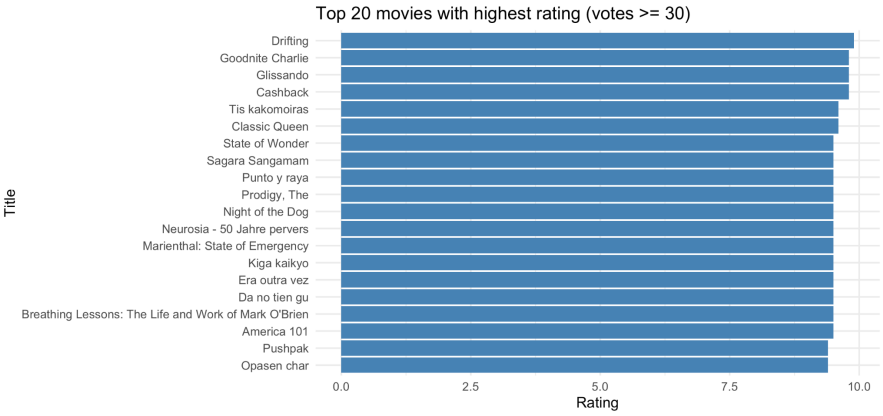


The number of films increased steadily over the years and increased significantly after the 2000s, reflecting the development of the film industry during this period.

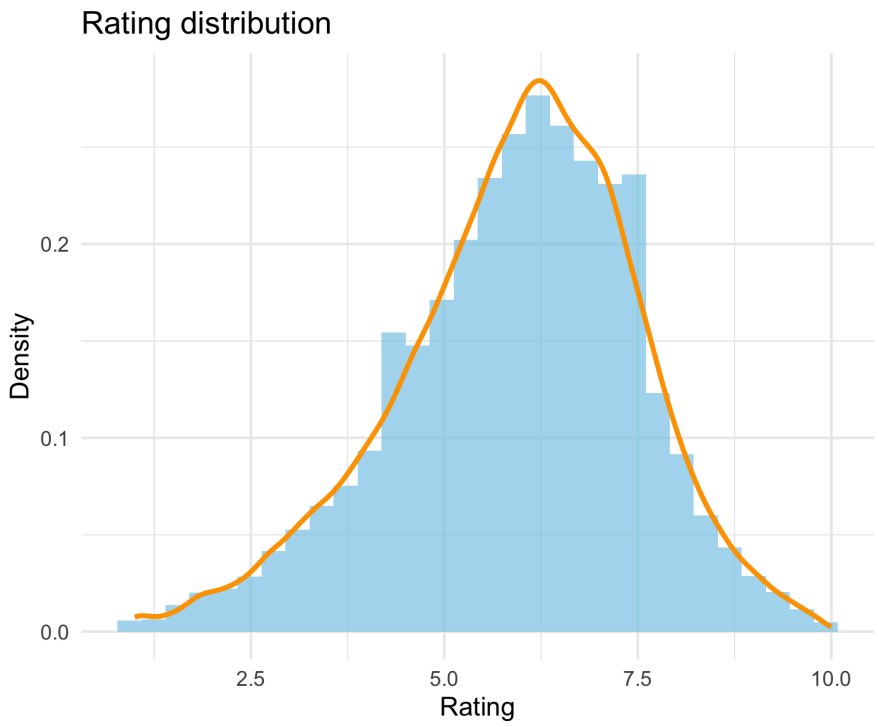


Medium length films (60-120 minutes) are generally rated highest.

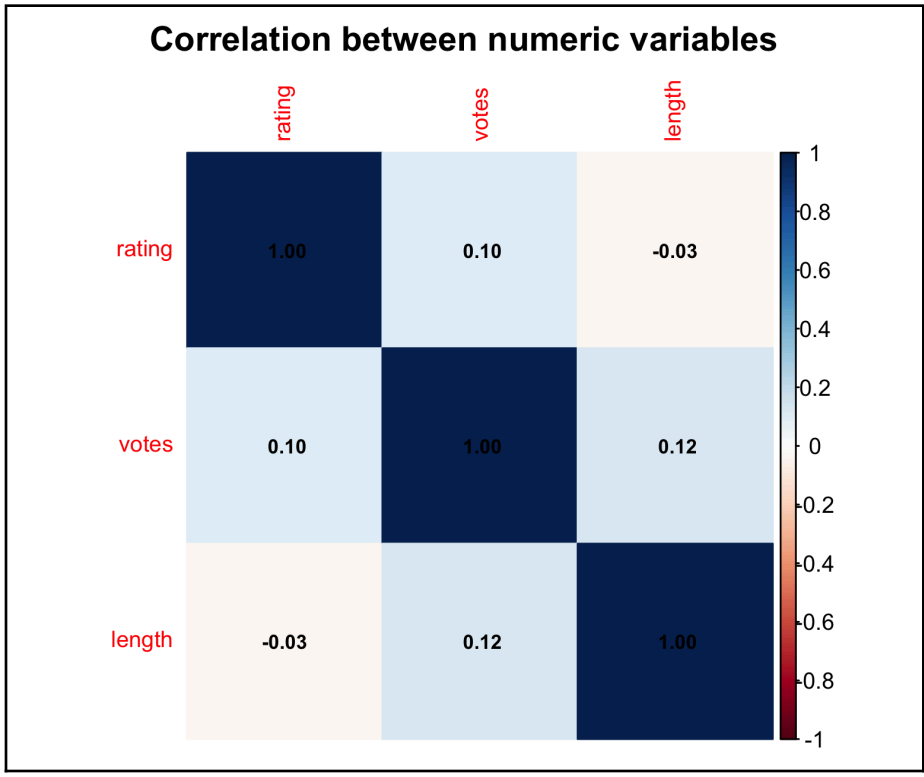
Short films (<60 minutes) are rated lower, possibly due to less engaging content.



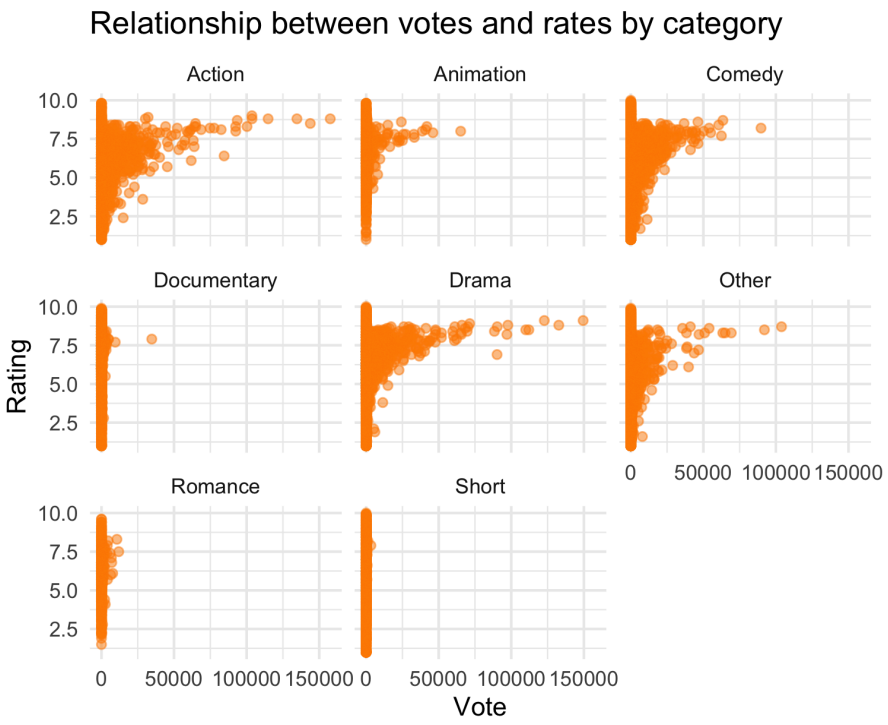
The chart shows the top 20 highest rated movies (with at least 30 votes)



It can be seen that the most popular movies are rated at 6-7.

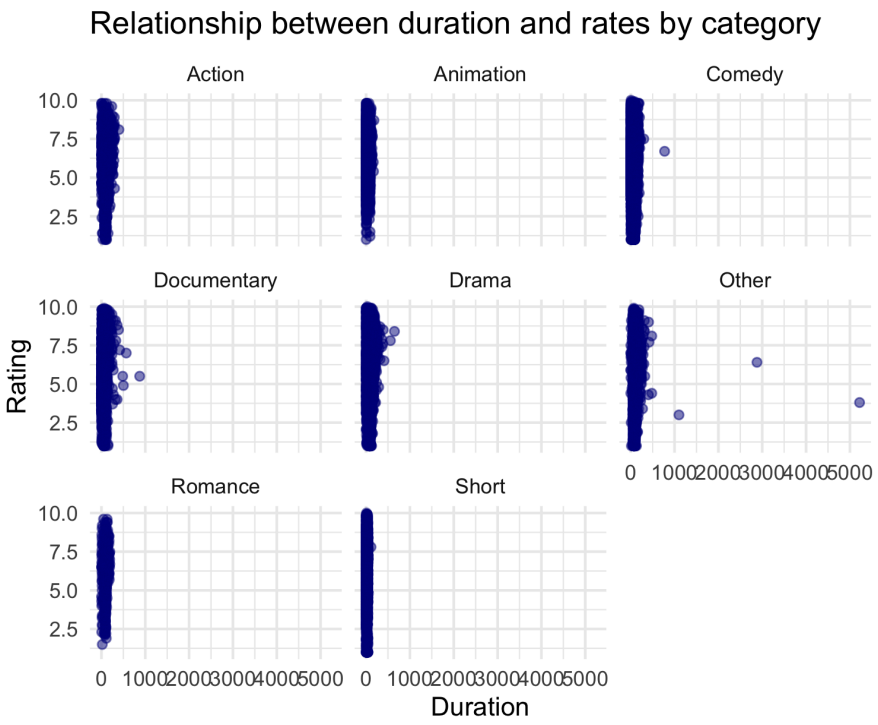


It can be seen that there is no significant negative or positive correlation between the numerical variables.



As can be seen, the genres with the strongest correlation between rating and votes are action, drama, followed by comedy and unspecified genres respectively.

It can be seen that short films often receive very few votes.



There isn't really a special correlation between film length and film rating. It's understandable since films are often rated on content, not length.