Data Wrangling Report

Fangnong Dai

5/7/2021

Abstract

Movies, as people's entertainment, have been influencing people in all walks of life. Through sentiment analysis on IMDb movies, it is explored that current movies, whether they are popular or highly praised, tend to write more negative information in the movie description. Perhaps this is just a means to attract the audience, because humans enjoy the process of overcoming difficulties, but it is undeniable that negative things in the film account for a large proportion.

Data collection

The data acquisition is divided into two stages. First, an api key is obtained on the website:https://imdb-api.com/api, and four lists are crawled. Bsically, they are top 250 movies, popular movies, top 250 tvs and popular tvs. In each dataset, ID, tittle, year, image, crew, IMDb rating and IMDb rating count are displayed. Apparently, there is not much information, for example, no genres, no description, no reviews. Hence, finding a data source in kaggle to make a supplement: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+ratings.csv. In the attached kaggle data source, only selected genre, duration, country, description in the IMDb movies.csv file into our original data sets, and thereby created two new dataset. E.g. the following is attributes for top 250 movies.

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'purrr' was built under R version 4.0.4
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
## Warning: package 'forcats' was built under R version 4.0.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## # A tibble: 6 x 12
##   id     rank title fullTitle  year crew  imDbRating imDbRatingCount genre
##   <chr> <int> <chr> <chr>     <int> <chr>      <dbl>           <int> <chr>
## 1 tt01~     1 The ~ The Shaw~  1994 Fran~        9.2         2379439 Drama
## 2 tt00~     2 The ~ The Godf~  1972 Fran~        9.1         1648431 Crim~
## 3 tt00~     3 The ~ The Godf~  1974 Fran~        9           1147592 Crim~
## 4 tt04~     4 The ~ The Dark~  2008 Chri~        9           2343172 Acti~
## 5 tt00~     5 12 A~ 12 Angry~  1957 Sidn~        8.9          702528 Crim~
## 6 tt01~     6 Schi~ Schindle~  1993 Stev~        8.9         1230452 Biog~
## # ... with 3 more variables: duration <int>, country <chr>, description <chr>
```

Because this data in kaggle was collected in 2020. There is a time gap between this and the one collected from IMDb API, causing some missing values. As the figure indicates, no records for Zack Snyder's Justice League and Soul. As for popular movies, there are around half of the movies have no records in genre, duration, country, description.

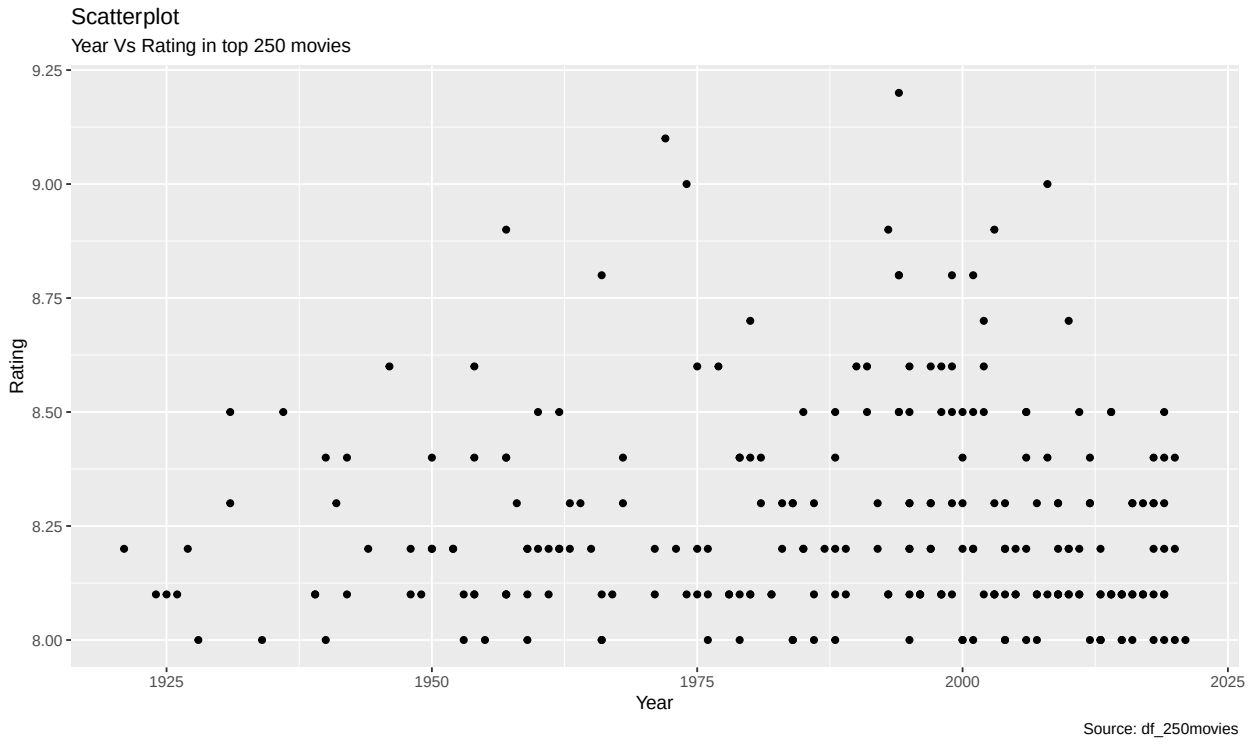
```
## # A tibble: 2 x 8
##   id         rank title                   year genre duration country description
##   <chr>     <int> <chr>                 <int> <chr>    <int> <chr>   <chr>
## 1 tt2948372   233 Soul                   2020 0            0 0       0
## 2 tt123619~   245 Zack Snyder's Justic~  2021 0            0 0       0
```

```
## # A tibble: 4 x 4
##   id          rank title                        year
##   <chr>      <int> <chr>                       <int>
## 1 tt8503618     60 Hamilton                     2020
## 2 tt10272386   132 The Father                   2020
## 3 tt2948372    233 Soul                         2020
## 4 tt12361974   245 Zack Snyder's Justice League  2021
```

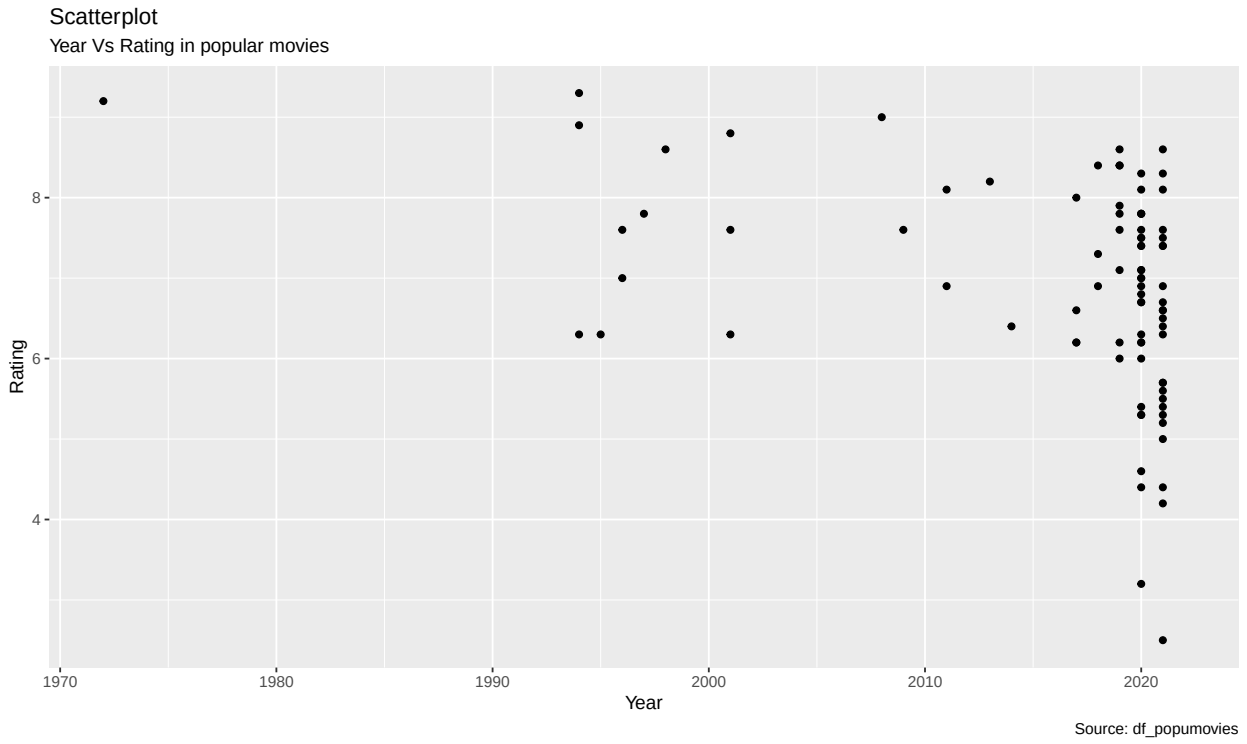
In this process, perhaps due to that, in the direction of TV series, there are many self-operated platforms such as Netflix, which can support watches, evaluates and discussions, there is not much information about IMDB's TV series data. Therefore, the TV series data will not be discussed in the this project.

Data Visualization

When considering what distinguishes the two series, the first thing that comes to mind is time. There may be some recent movies in the popular series. Therefore, first look at the relationship between time and ratings in the two lists.



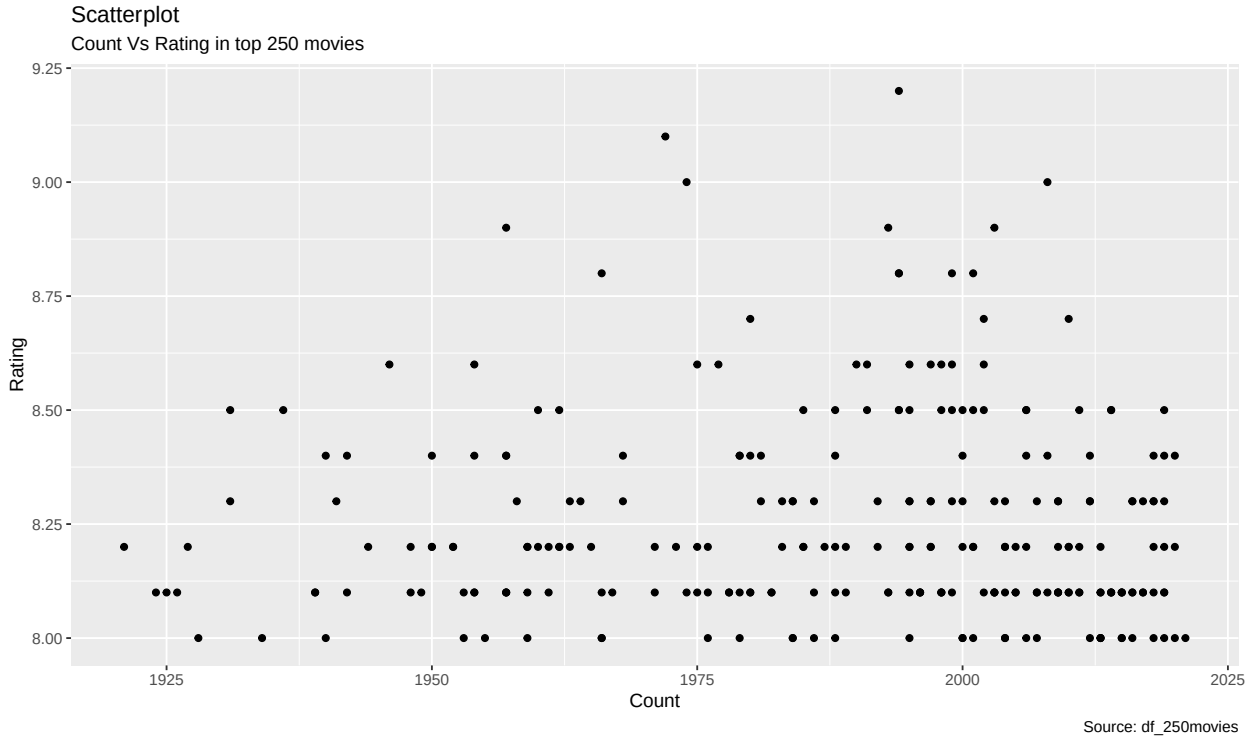

```
## Warning: Removed 13 rows containing missing values (geom_point).
```



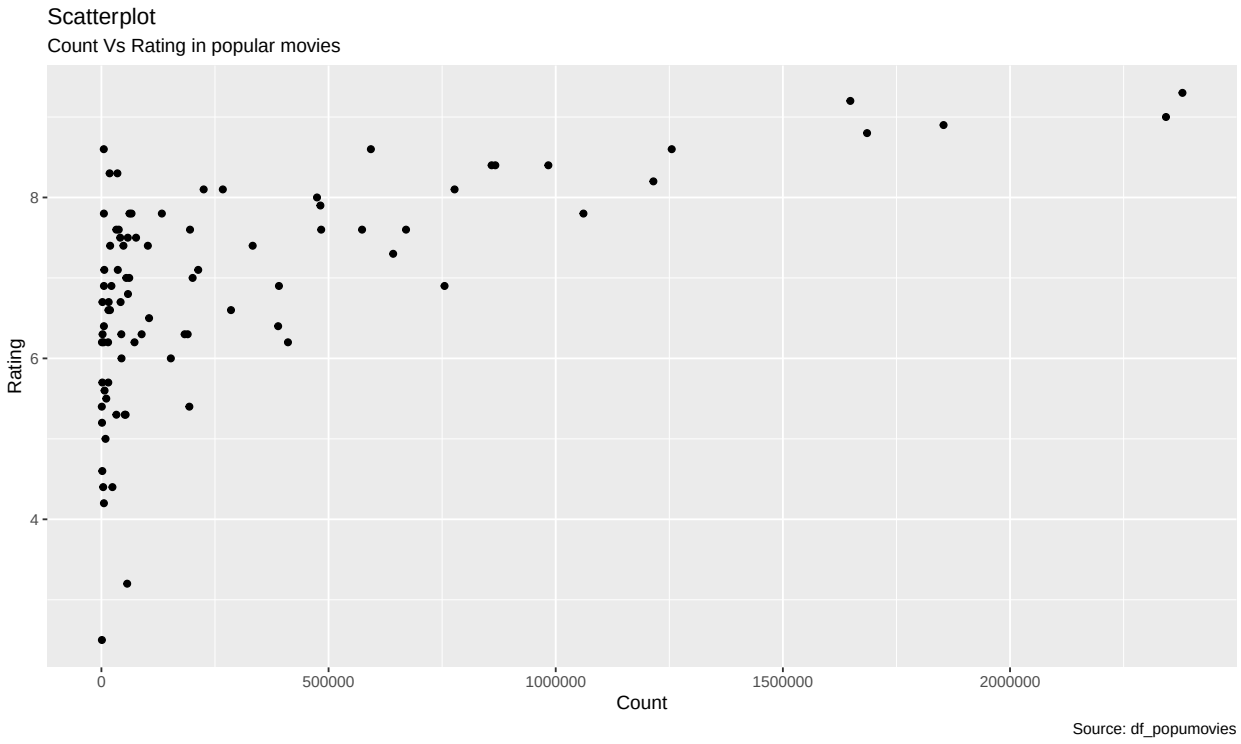
It can be concluded that time has no much affect on appreaciation of movies. Although this website is constructed in 1990, many people came to rate movies before that. The concentrated points at the right hand side imply that excellent movies entering the top 250 consequently, especially in

2020, which is a important suggestion that coronavirus did not devestate our file industry.

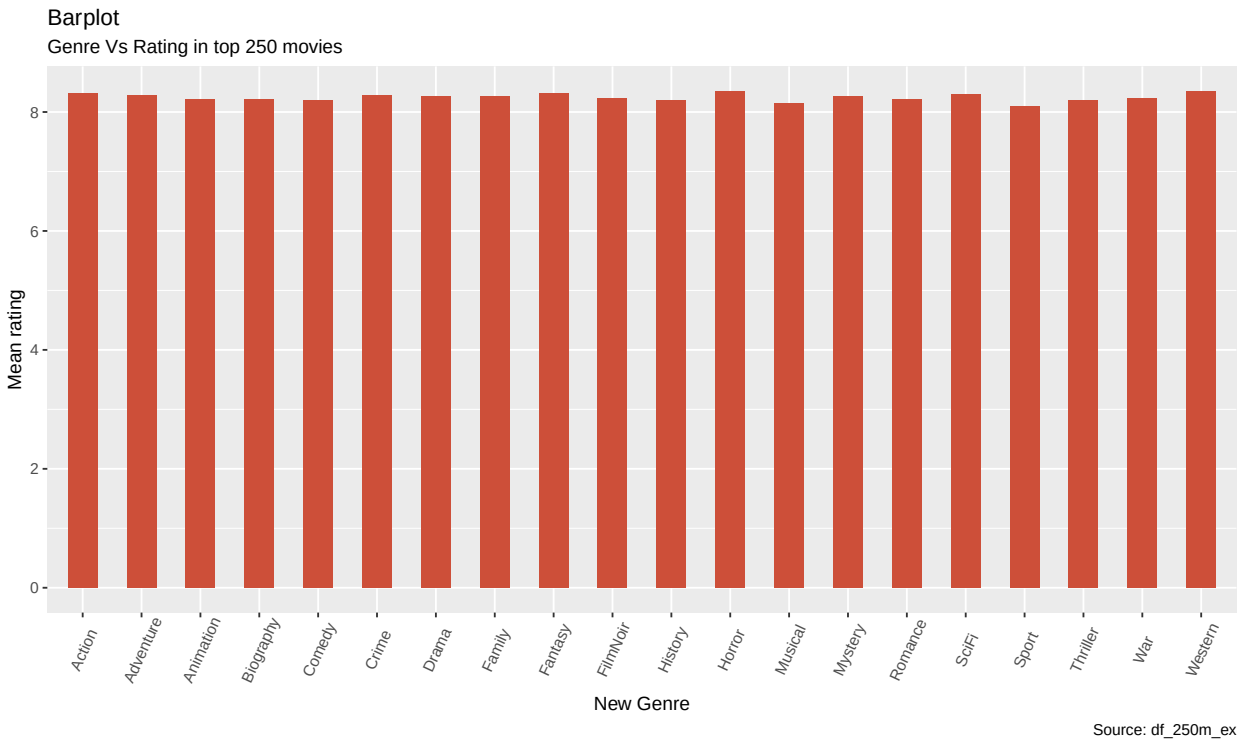
As for ratings, there are some classic movies that are hard to surpass, whether in terms of evaluation or popularity. The Shawshank Redemption (1994) and The Godfather (1972) are both highly evaluated movies, also occupy certain positions in the list of popular movies.

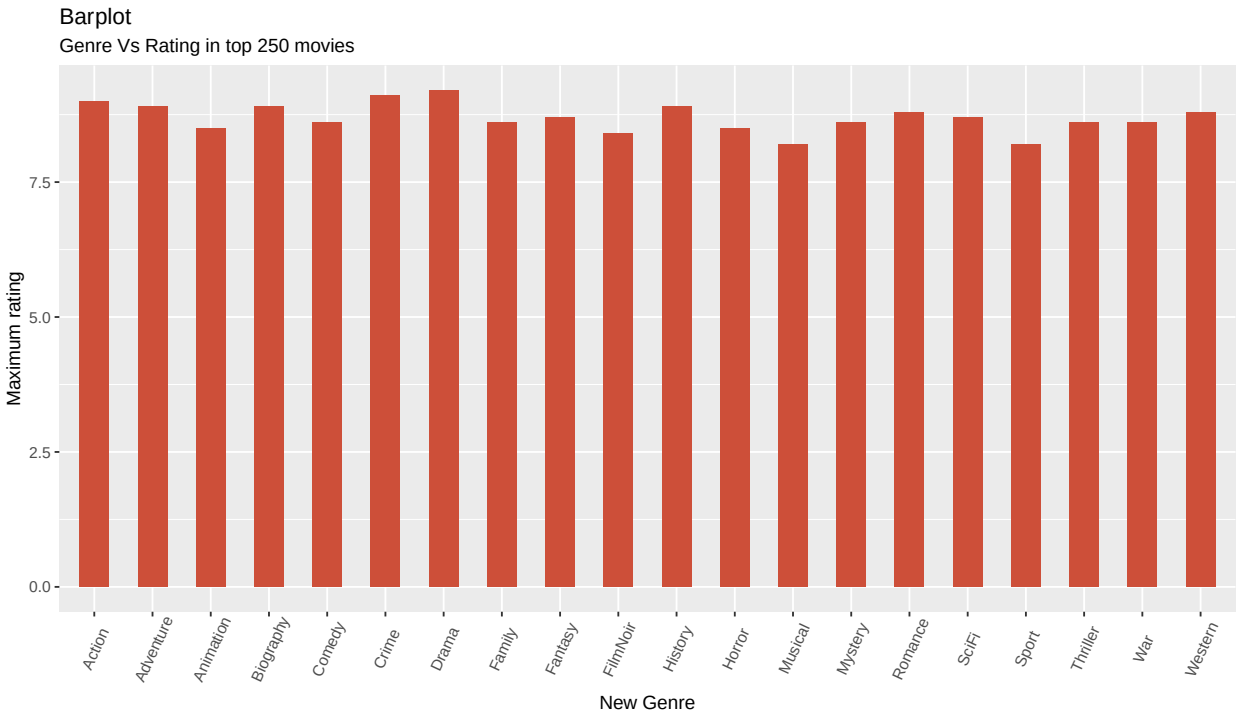


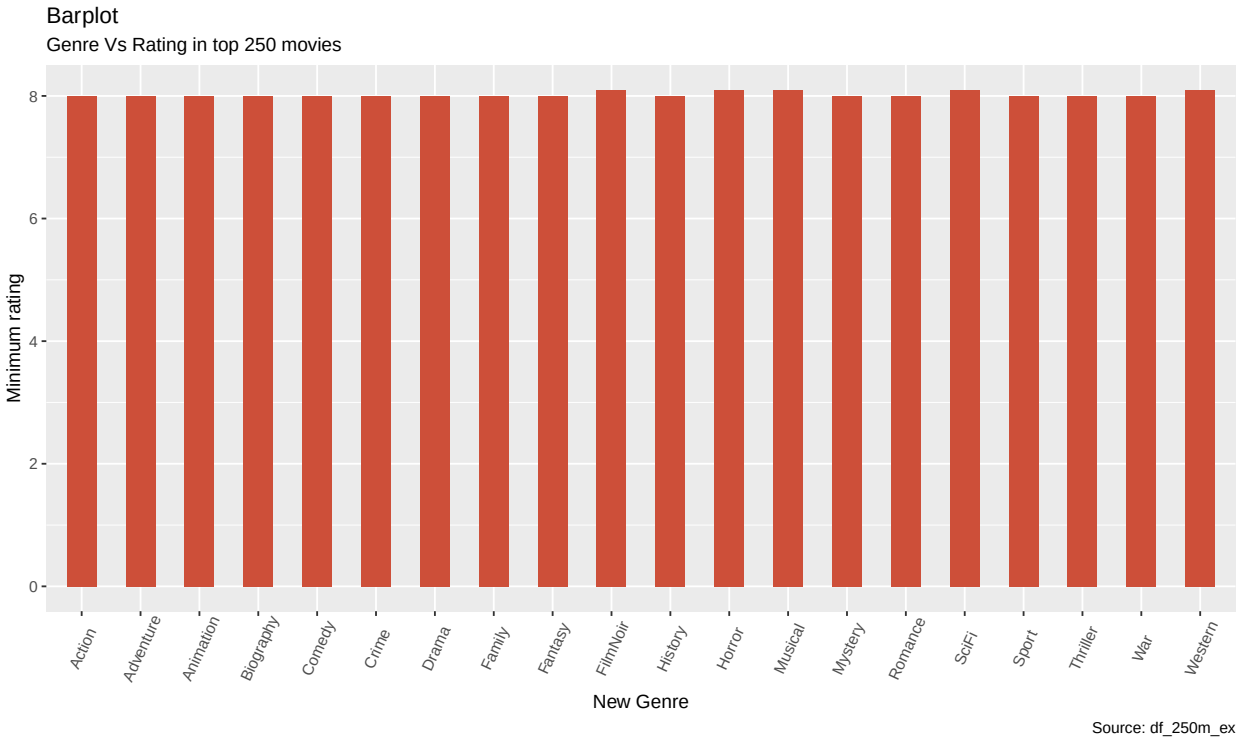

```
## Warning: Removed 13 rows containing missing values (geom_point).
```

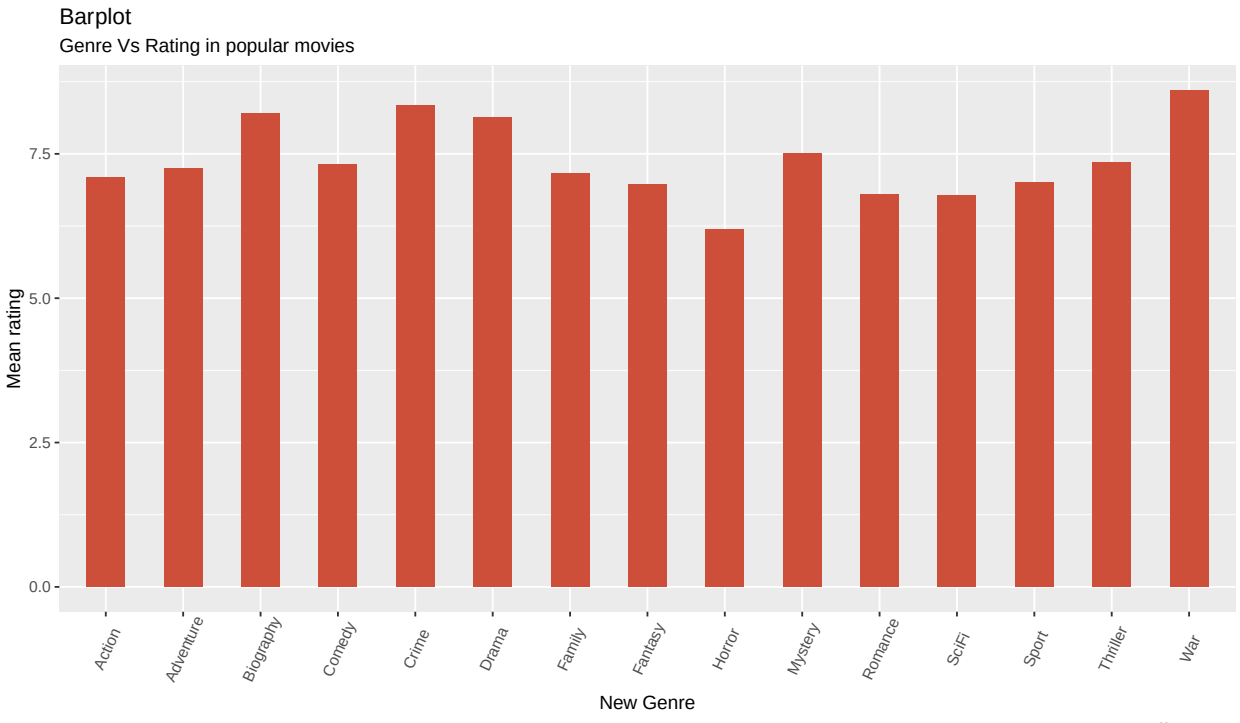


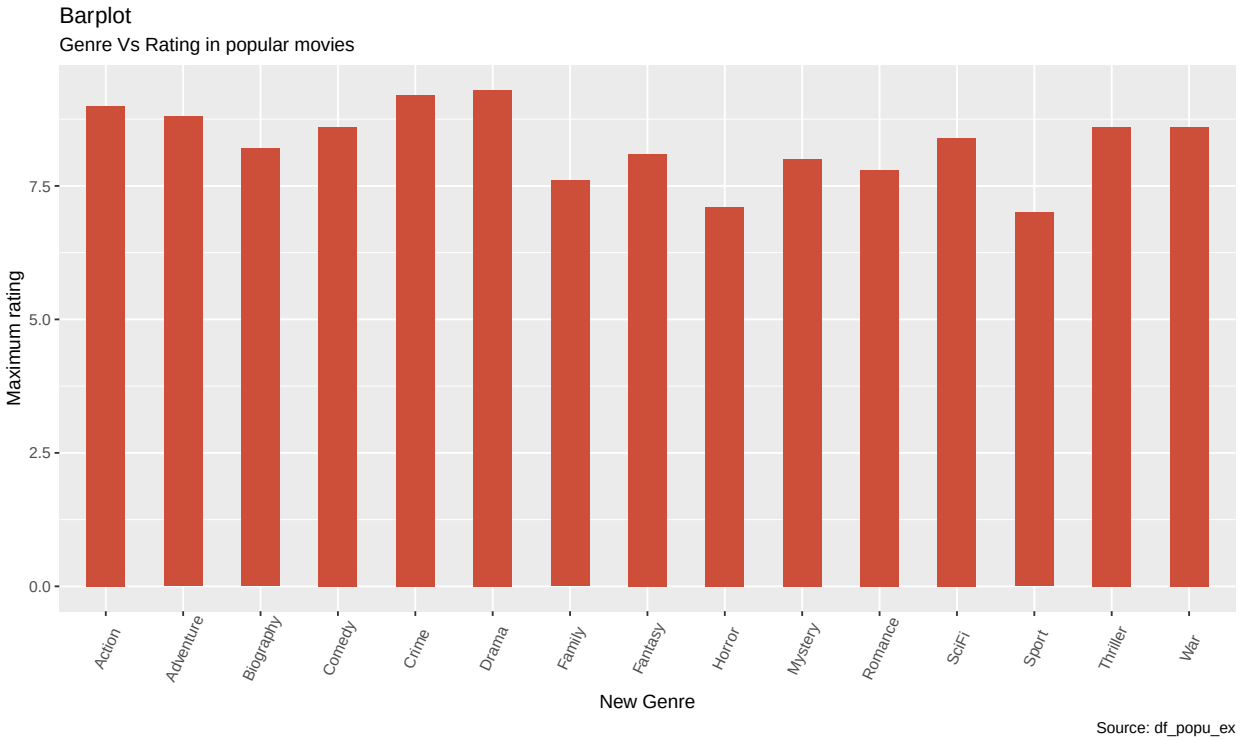
Seperate genres into individual topic, and plot the mean, maximum and minimum ratings in two lists.

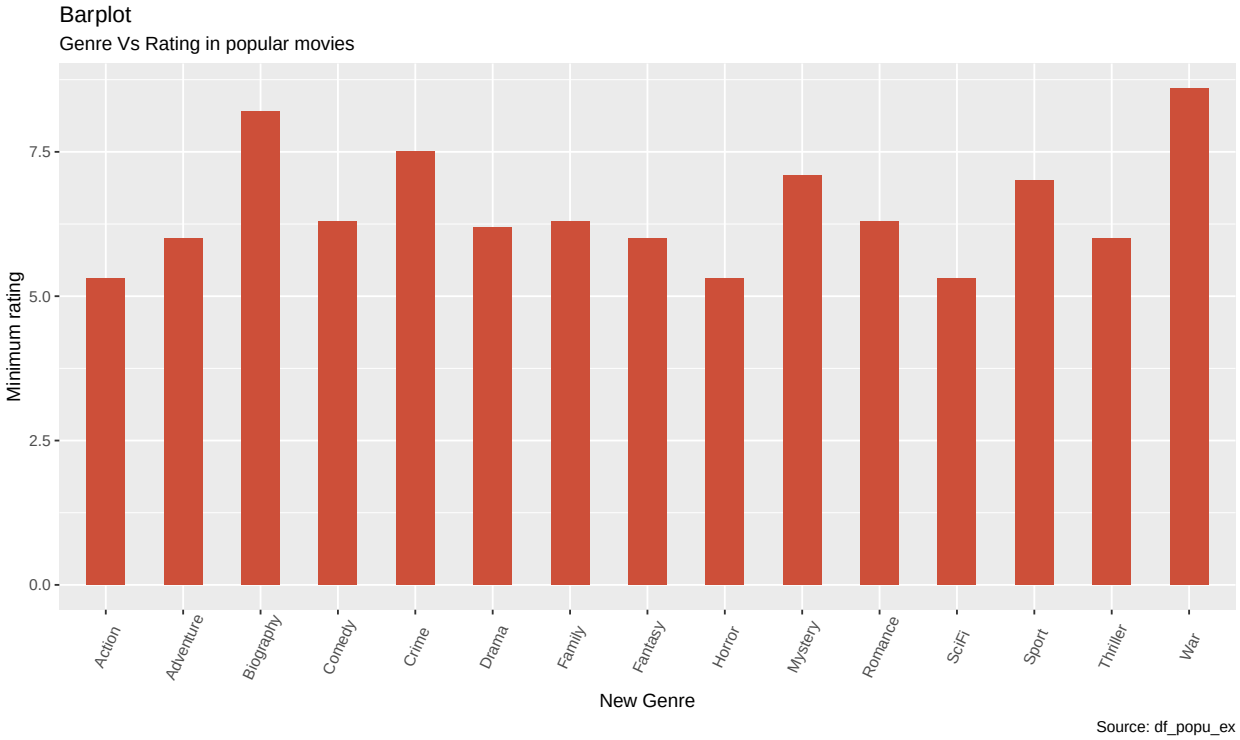



Barplot
Genre Vs Rating in top 250 movies
Maximum rating
7.5
5.0
2.5
0.0
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
FilmNoir
History
Horror
Musical
Mystery
Romance
SciFi
Sport
Thriller
War
Western
New Genre
Source: df_250m_ex


Barplot
Genre Vs Rating in top 250 movies
Minimum rating
8
6
4
2
0
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
FilmNoir
History
Horror
Musical
Mystery
Romance
SciFi
Sport
Thriller
War
Western
New Genre
Source: df_250m_ex


Barplot
Genre Vs Rating in popular movies
Mean rating
7.5
5.0
2.5
0.0
Action
Adventure
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror
Mystery
Romance
SciFi
Sport
Thriller
War
New Genre
Source: df_popu_ex


Barplot
Genre Vs Rating in popular movies
Maximum rating
7.5
5.0
2.5
0.0
Action
Adventure
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror
Mystery
Romance
SciFi
Sport
Thriller
War
New Genre
Source: df_popu_ex



There is not much difference in ratings for different genres in top 250 list, indicating top 250 is a very fair list. Each genre shares similar average rating. However, list of popular movies is in totally distinct situation. In other words, the preferences of audience for different genres are obvious. The genres in popular movies are even less than the genres in top 250 list, for example, no western, no musical in popular list.

Sentiment analysis

It seems to be difficult to distinguish in ratings of movies by only looking as the genre, especially in the top 250 movies. As a result, a sentiment analysis was conducted on the description of movies. Some common features are presented in the following figure: world, family, war and crime. There is a correlation coefficient, 0.3414 though not large.

```
## Warning: package 'tidytext' was built under R version 4.0.4
```

```
## Joining, by = "word"
```

```
## # A tibble: 6 x 2
##   word       n
##   <chr>  <int>
## 1 life      25
## 2 war       20
## 3 world     15
## 4 woman     14
## 5 family    13
## 6 son       13
```

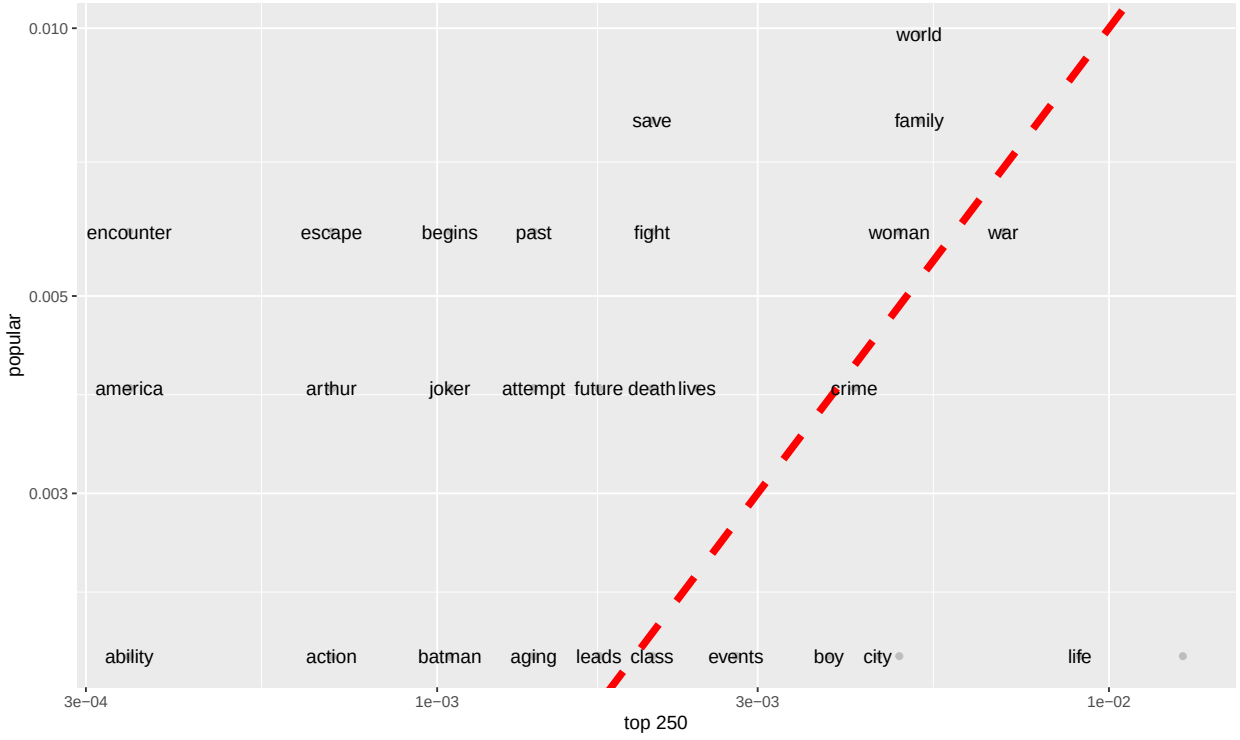
```
## Joining, by = "word"
```

```
## # A tibble: 6 x 2
##   word          n
##   <chr>     <int>
## 1 world         5
## 2 family        4
## 3 save          4
## 4 begins        3
## 5 encounter     3
## 6 escape        3
```

```
## # A tibble: 2,019 x 3
##    word      popular `top 250`
##    <chr>       <dbl>     <dbl>
##  1 abilities 0.00197 NA
##  2 ability   0.00197  0.000348
##  3 aboard    0.00197 NA
##  4 accept    0.00197  0.000348
##  5 achieve   0.00197 NA
##  6 act       0.00197  0.000348
##  7 action    0.00197  0.000696
##  8 actor     0.00197 NA
##  9 acts      0.00197  0.000348
## 10 advanced  0.00197  0.000348
## # ... with 2,009 more rows
```

```
## Warning: Removed 1747 rows containing missing values (geom_point).
```

```
## Warning: Removed 1748 rows containing missing values (geom_text).
```



```
##           popular   top 250
## popular 1.0000000 0.3356749
## top 250 0.3356749 1.0000000
```

Then check the 10 most common bigrams. Perhaps people are care about war, the top three bigrams in the top 250 movies are related to war, whether in reality or in fictional world.

```
## # A tibble: 1,125 x 2
##    bigram            n
##    <chr>         <int>
##  1 war ii            7
##  2 world war         7
##  3 darth vader       3
##  4 gotham city       3
##  5 serial killer     3
##  6 york city         3
##  7 army officer      2
##  8 bounty hunter     2
##  9 dark lord         2
## 10 enemy lines       2
## # ... with 1,115 more rows
```

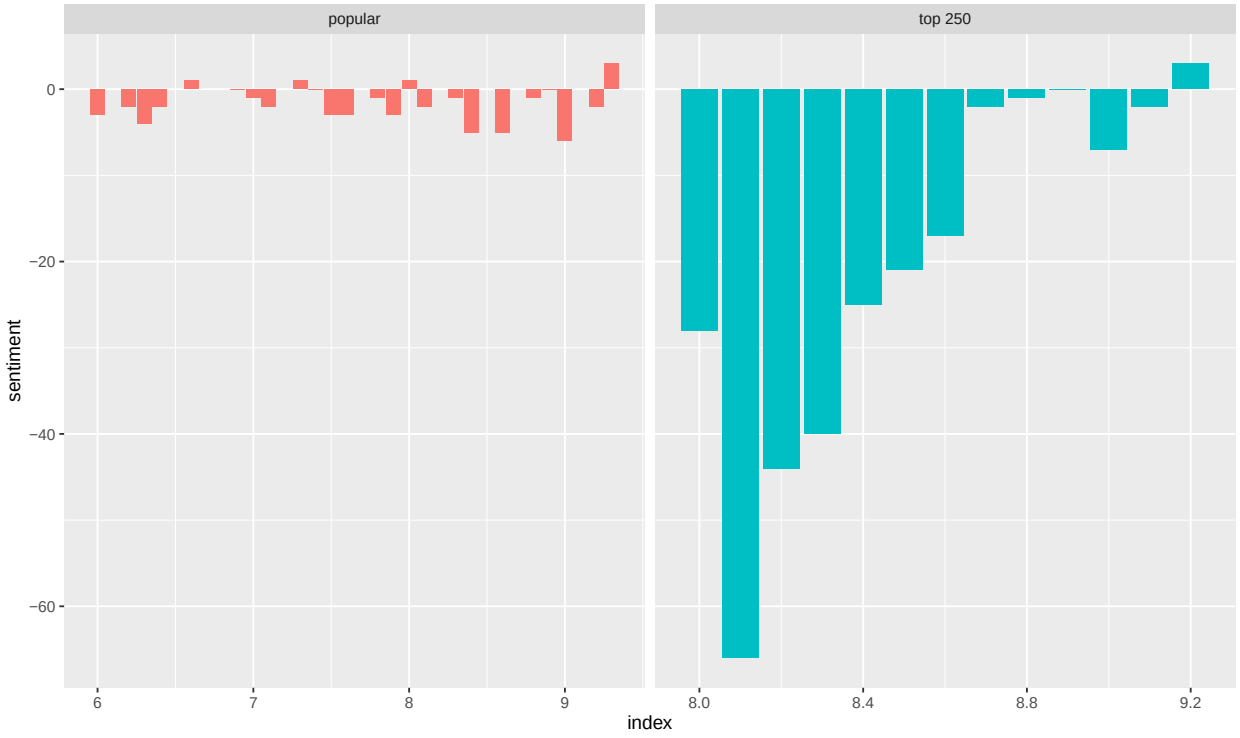
```
## # A tibble: 202 x 2
##    bigram              n
##    <chr>           <int>
##  1 blade runner        2
##  2 captain america     2
##  3 dark lord           2
##  4 1969 los            1
##  5 achieve fame        1
##  6 act bruce           1
##  7 advanced kingdom    1
##  8 agency monarch      1
##  9 aging patriarch     1
## 10 ally diana          1
## # ... with 192 more rows
```

Compute the attitude for movies in the two list. No matter which list the description is in, there are more negative information than positive. In the list of the top 250 movies, it is surprised to see that a trend of decreasing in negative information as the scores rising.

```
## # A tibble: 3,380 x 4
## # Groups:   movie [2]
##    word       movie   title                     imDbRating
##    <chr>      <chr>   <chr>                          <dbl>
##  1 imprisoned top 250 The Shawshank Redemption        9.2
##  2 bond       top 250 The Shawshank Redemption        9.2
##  3 finding    top 250 The Shawshank Redemption        9.2
##  4 solace     top 250 The Shawshank Redemption        9.2
##  5 eventual   top 250 The Shawshank Redemption        9.2
```

```
##  6 redemption top 250 The Shawshank Redemption        9.2
##  7 acts       top 250 The Shawshank Redemption        9.2
##  8 common     top 250 The Shawshank Redemption        9.2
##  9 decency    top 250 The Shawshank Redemption        9.2
## 10 aging      top 250 The Godfather                   9.1
## # ... with 3,370 more rows
```

```
## Joining, by = "word"
```



Discussion

Among the current mainstream genres, high-scoring movies can be produced under any genre. However, war movies, perhaps due to their high production costs, generally get a not-low rating. In the descriptions of these movies, more negative words appear. Does it reflect that the current movies convey too much negative information? Or maybe negative adjectives are more attractive to the audience. Do these movies make people feel positive or negative? Maybe in the future it is worthwhile to collect these movie-related reviews to study this point. Also, topic models by LDA was tried but failed. It is possible due to insufficient variables, the topic model did not fit a result.