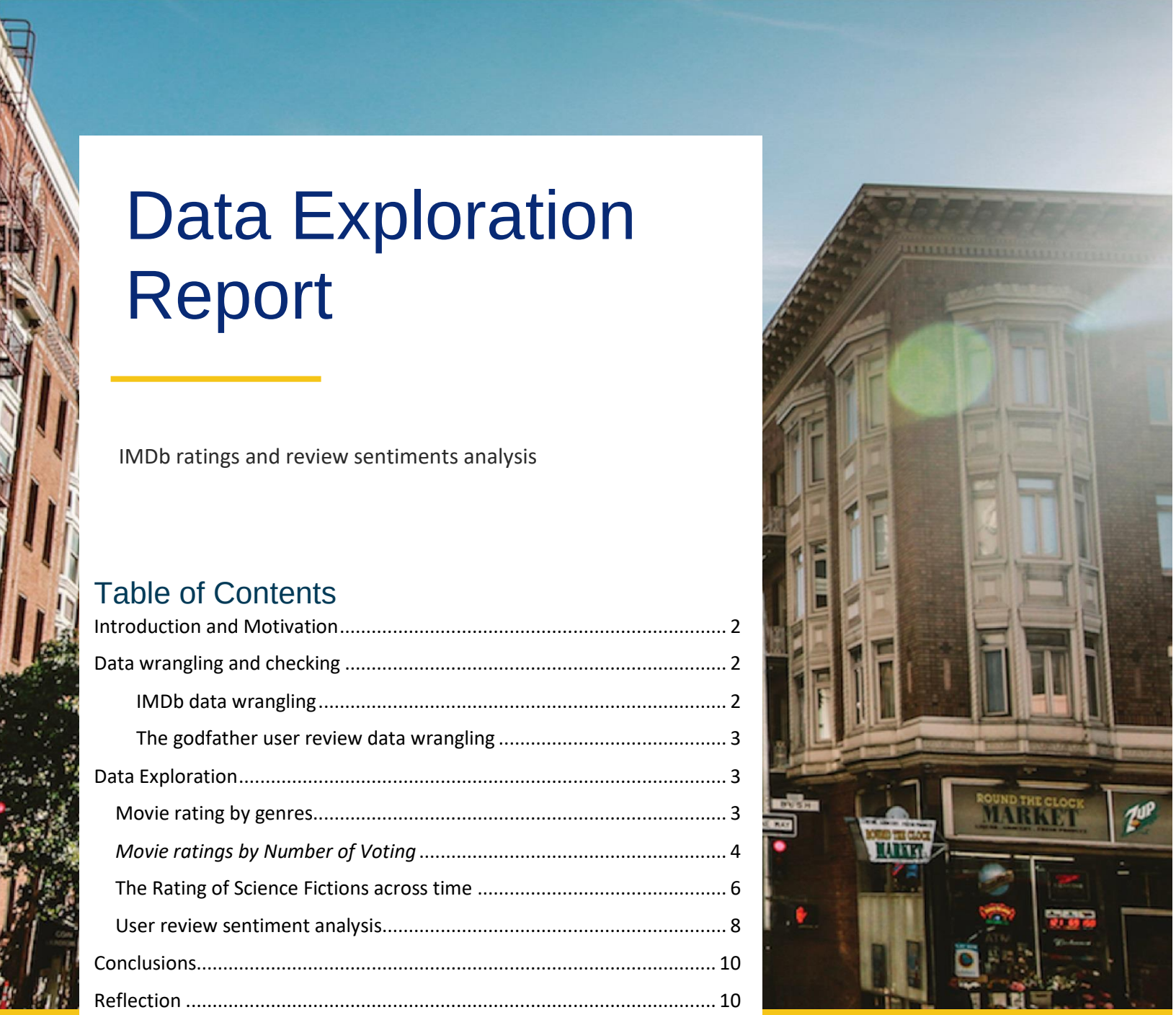


Data Exploration Report

IMDb ratings and review sentiments analysis

Table of Contents

SEPTEMBER 9

By: Xiaoyu Tian 28540964
Tutor: Angel Das and Mohit Gupta

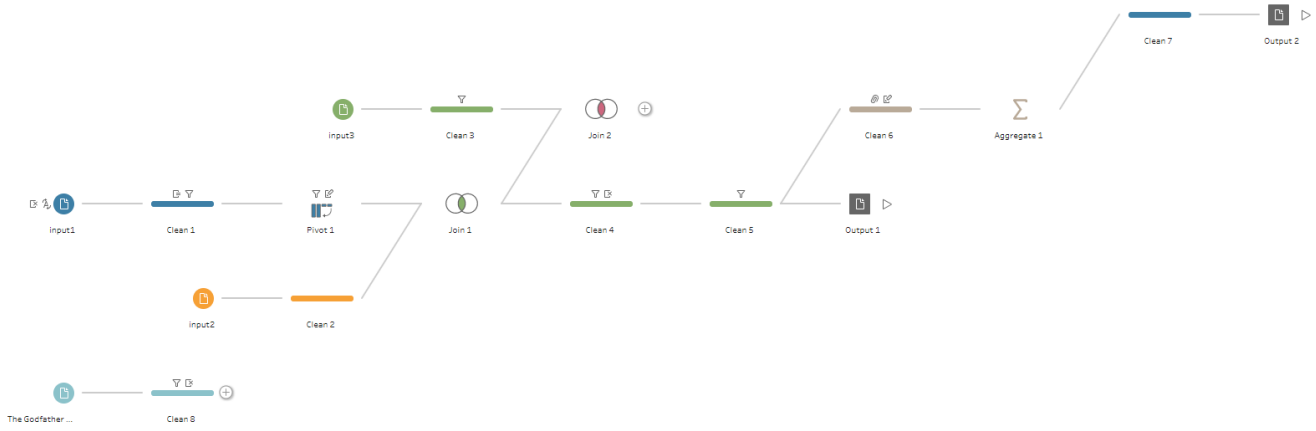
Introduction and Motivation

This report is to explore the IMDb dataset. As I am a film lover, I am always interested in exploring every aspect of a good movie. Recently, I have browsed the IMDb top rated 250 movies. Although it's not surprising that those famous movies are on that list, I am still wondering what kind of topic or attributes of a movie that the viewers mostly like. Also, since I'm a fan of science fiction movies, I would like to explore whether the recent year science fiction movies made are more favorable by viewers than the classic science movies in the past, given the fact that the film technology has been growing rapidly. What's more, as The God Father is one of my favorite movies, thus I want to find out what other viewers on the IMDb thinks of this movie. Therefore, this report will focus on the following three questions:

1. What are the attributes of a movie that would contribute to a higher user rating?
2. How is the rating of science fiction movies changing throughout the time?
3. What are the viewer's sentiments of The God Father?

Data wrangling and checking

Before the data exploration, data wrangling is first processed in Tableau Prep. The following flow chart is the wrangling process:



IMDb data wrangling

- Input1: "title.basics.tsv"; input 2: "title.ratings.tsv", input 3: "name.basics.tsv"
- Clean 1,2,3
  Remove the null from input 1,2,3. For input 1. Filter **titleType** and keep only "movie", as other productions such as tv series are not the main interests in this report. Split the column **genres** into three columns, as each movie has at most 3 genres combined with comma in the cell.
- Pivot 1
  Pivot the three split genres from columns into rows and make the data as longer format where all grouped in one column called "genres split".
- Join 1, 2

Three datasets are inner joined based on “tconst”, the unique identifier for each movie.

- Clean 4,5
  Remove the duplicated fields, change the startYear column to date variable. Exclude genres type “short”, “Reality-TV” and “News” due to low occurrence of data (less than 10).
- Output 1
  Output the dataset as “IMDb_movie.csv”
- Clean 6
  Group the years in decades.
- Aggregate 1 and output2
  Data are aggregated based on **Year in Decades, primaryTitle, genres split** respectively. The aggregated data are created as output2.

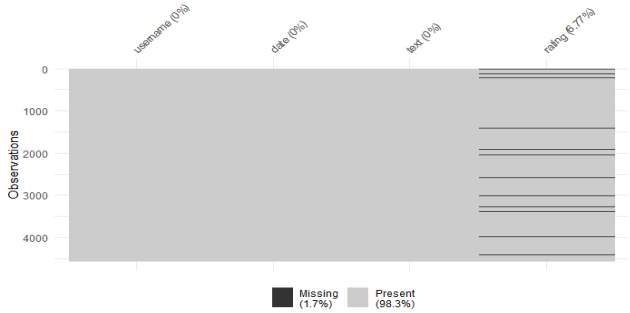
The godfather user review data wrangling

“The Godfather (1972) – IMDb.csv” is web scrapped from the IMDb review website. It contains 4.5k reviews.

Clean 8

Remove unrelated field and urls and keep the field of **username, date, text and ratings**. Since this part is text sentiment analysis, the further wrangling is conducted in Rstudio and read.csv as “review”.

Check and clean the null values



Parse the text column as long reviews are compressed from the scraping.

Unnest the text data into words by using **unnest_tokens**. Received the output as “review_words”.

Data Exploration

Movie rating by genres

As we know, there are many different topics of movies, and for each individual movie, there could also be multiple genres included. According to IMDb data, there are 26 genre types. In this report, we will focus on 23 genres, where adult, short and reality-tv genres are filtered out due to their low occurrence in the dataset (less than 10).

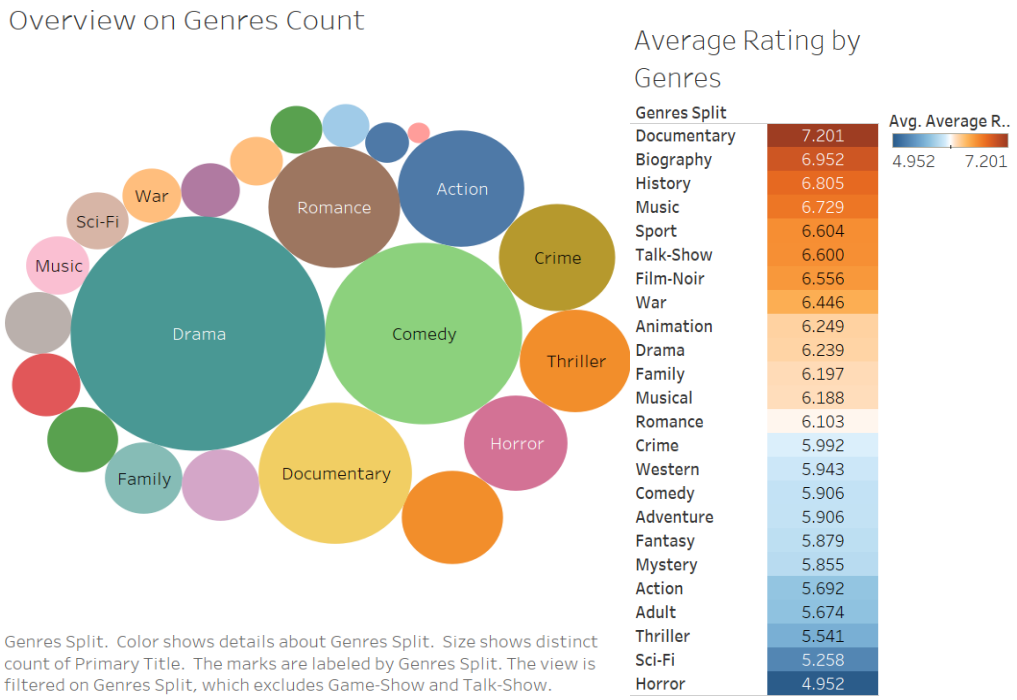


Figure 1 Overview on genres count their contribution to rating

The bubble chart shows the overview of the genres count, where the size of them shows the distinct count of movie titles. The color shows the different types of genres, sorting in descending order. Drama has the most count, followed by comedy and documentary. However, drama only ranked in the middle of all genres in terms of average rating, while documentary has the highest average rating, followed by biography and history. It is interesting to discover that most popular genres such as action, thriller, horror, etc. appear to score a low average rating which is ranked in the bottom of the table shown as blue color. We will further explore the other attributes that might affect the rating.

Movie ratings by Number of Voting

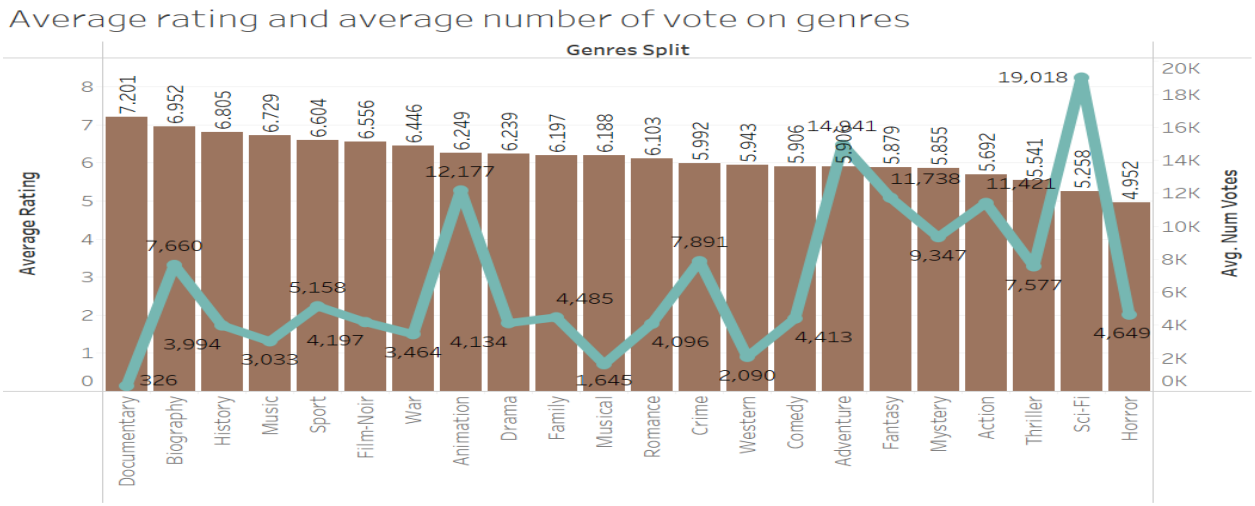
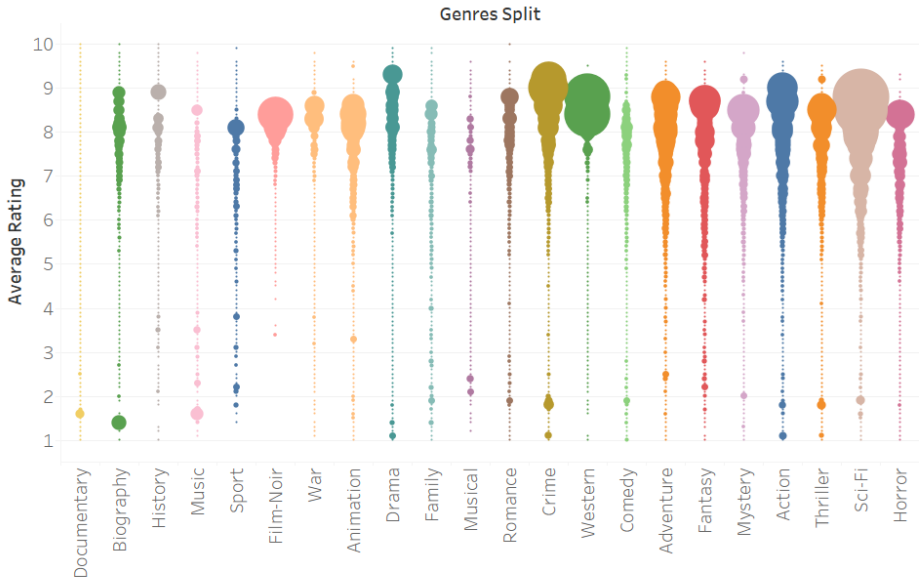


Figure 2 Average rating and number of voting on genres

By conducting a dual axis with the number of votes in each genre, an increasing trend in the number of votes is shown as blue line, when the genres moving from the higher average rating to low. This suggests movie with a smaller number of votes tends to have a higher average rating score. For example, the documentary has the highest rating but only 326 voting involved, while for science fiction, there are 19018 votes which is nearly 60 times of documentary.



Average Rating for each Genres Split. Color shows details about Genres Split. Size shows average of Num Votes. The view is filtered on Genres Split, which excludes Adult, Game-Show and Talk-Show.

Figure 3 Rating distribution by genres

Figure 3 shows the distribution of average rating by each genre. The size of the circle represents the count of votes, hence, the larger the circle is, the more votes it has. Clearly, when moving to higher scores, the counts of votes getting more and more for each genre type. This suggests the rating is somehow associated with the number of votes for that movie. Thus, the relationship between these two is displayed as follows.

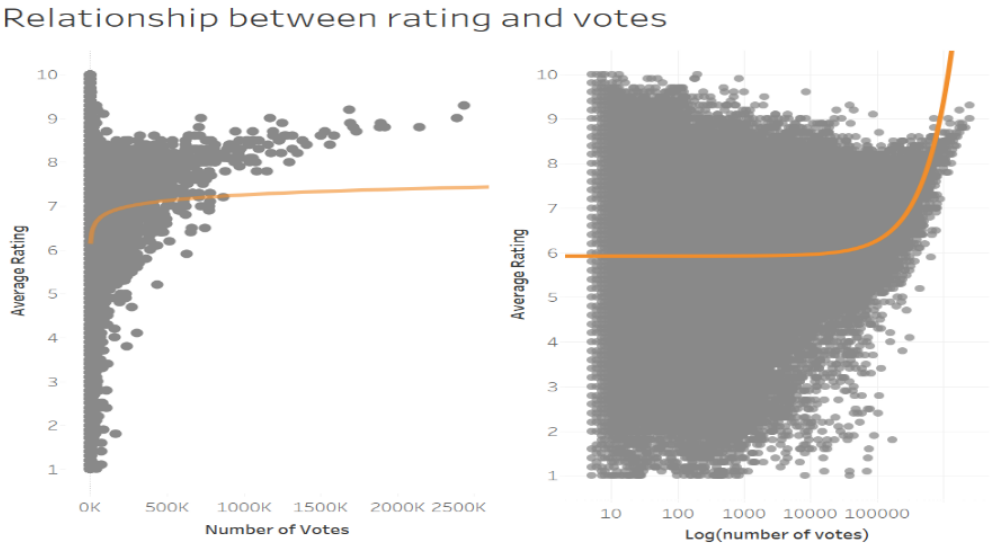
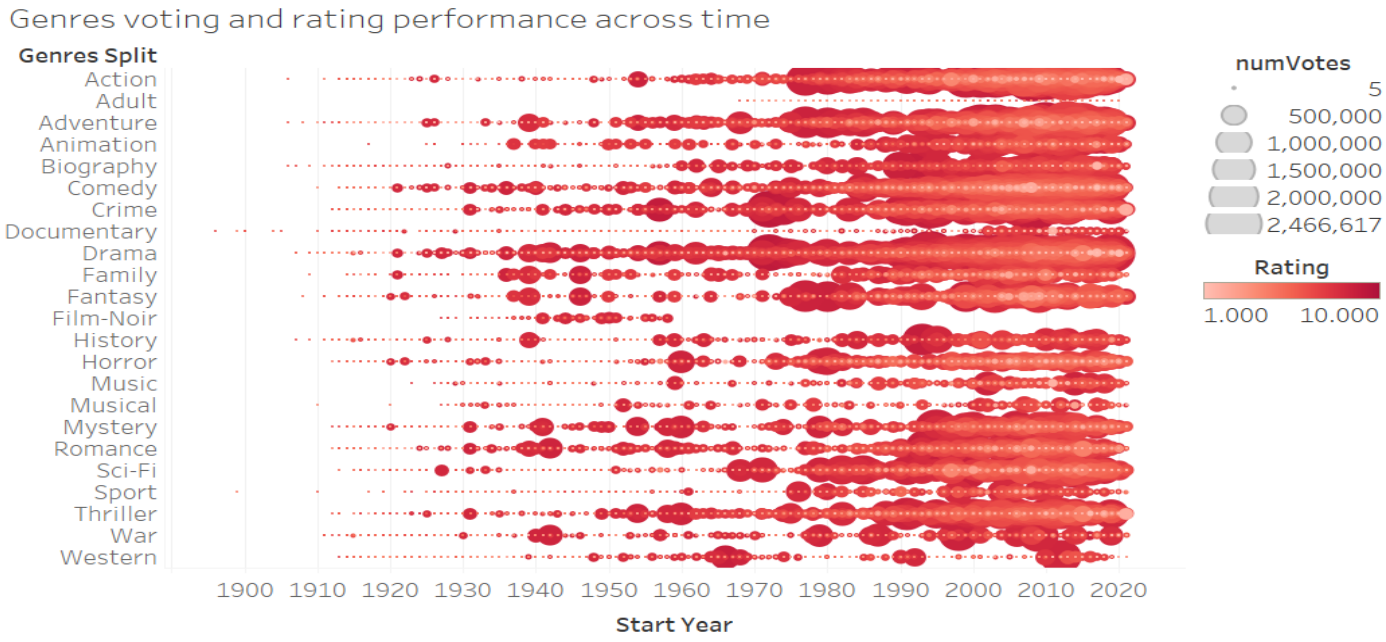


Figure 4 Relationship between rating and number of votes

The dot plot on the left-hand side shows the data is heteroscedastic and has barriers. The first barrier appears around rating 9-10. We can see that movies with high ratings (above 9) are those almost no voting involved. On the other hand, there is also a barrier in high-voting-low-rating area. One possible explanation is, the more popular the movies are (more voting), the more likely the movie to achieve a reasonably high rating score due to their popularity.
As the data is very skewed, a log transformation is conducted shown on the right-hand side of the dot plot. It is interesting to discover that some positive association only happens when large number of votes occur.

The Rating of Science Fictions across time



Start Year for each Genres Split. Color shows details about Average Rating. Size shows sum of numVotes. The view is filtered on Genres Split, which excludes Game-Show, News, Reality-TV, Short and Talk-Show.

Figure 5 Time series on genres rating

When considering the time series, we can see that the number of votes has significantly increased in the recent decades. According to the previous discussion, when the number of votes is getting bigger, the movies are more likely to have a better rating. Also, as the growth of technology in the movie-making field, it is interesting to discover whether the science fiction movies in the past 2 decades are rated higher than the movies before. In this section, we take science fictions as our focus and will determine whether the rating is getting better over recent years.

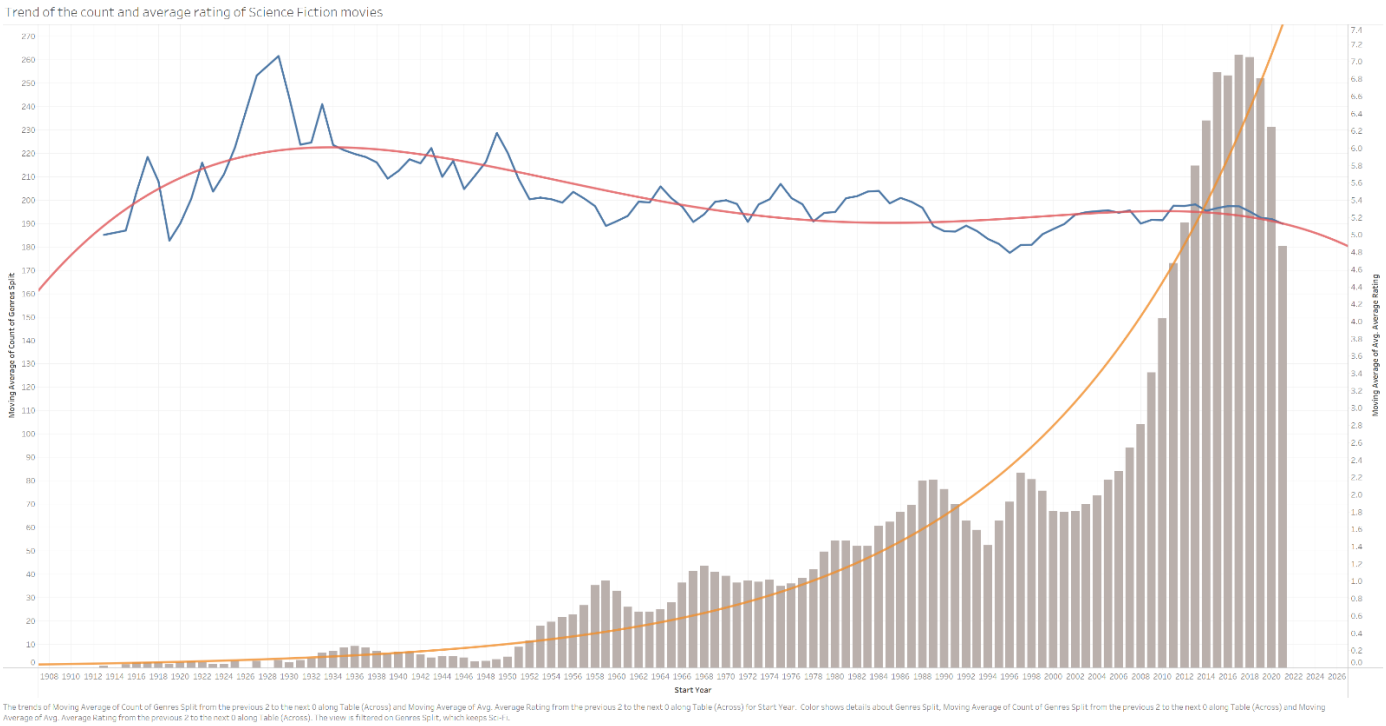


Figure 6 time series trend on science fictions

However, as the time series suggests, though the voting numbers has exponentially increased over time, the rating of them does not show an increasing trend but decreasing. This could be due to the barriers previously discussed. Before 2000, the total number of the votes from the start years is equal to the number of votes after 2020 for only one and half years' time. Low voting might cause the rating to achieve the extreme high ratings. We further conducted the hypothesis testing.

A two-sample hypothesis testing is conducted as follows"

H0: $\mu_1 - \mu_2 = 0$

HA: $\mu_1 - \mu_2 > 0$

μ_1: $The\ mean\ of\ the\ Science\ fictions\ ratings\ after\ 2000$

μ_2: $The\ mean\ of\ the\ Science\ fictions\ ratings\ before\ 2000$

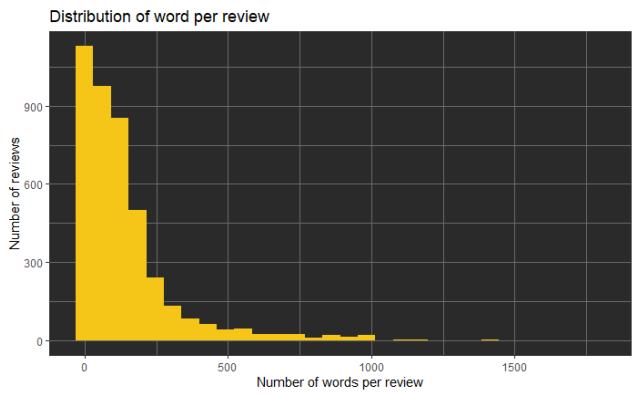
By conducting the testing in R, we retrieve the following result:

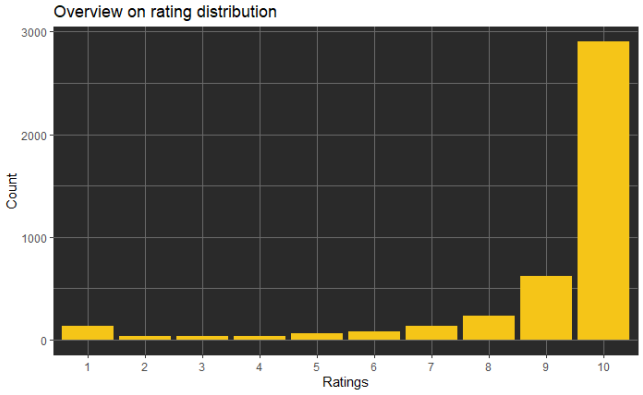
statistic	p.value	conf.high	method	alternative
-3.925764346	4.37E-05	-0.089702736	Two Sample t-test	less

Based on the low p value in the t.test, we reject the null hypothesis. This means that there is no statistical evidence to support that the science fictions after 2000 is filmed better than before 2000.

User review sentiment analysis

In this section, the text analysis is conducted on the user reviews for one of the top-rated movies “the godfather”. The movie is ranked the second in the IMDb movie list with an average rating of 9.2 and has over 4.5k reviews with ratings are uploaded.





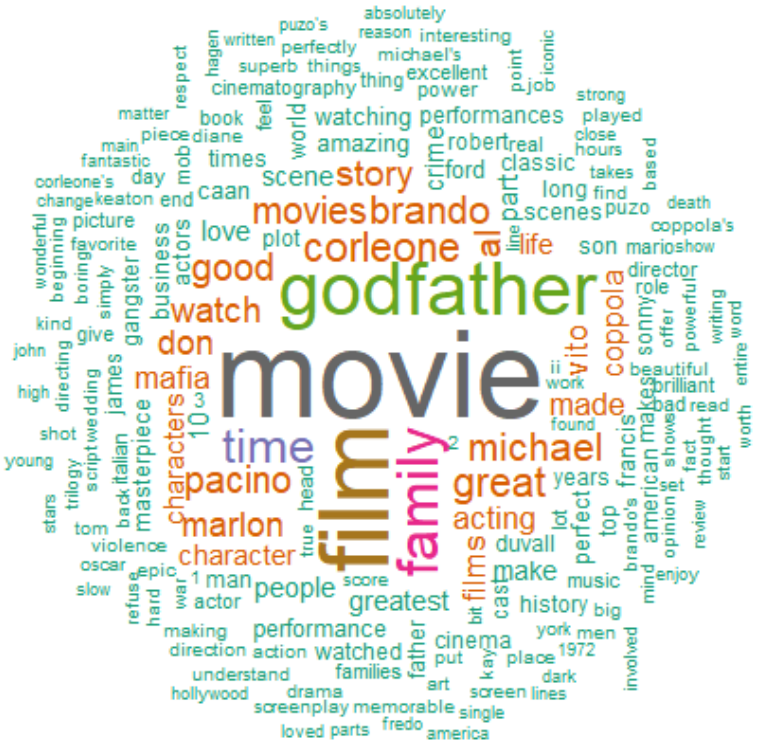
By checking the distribution of the word count per review, we can see it’s right skewed and most of the reviews are less than 250 words. For the rating distribution on these reviews, there is a completely different distribution, where it has the left skewness, and most people gave the highest rating of 10. It is interesting to discover that more people rated the lowest score of 1 than the medium ratings.

Description: df [6 x 2]

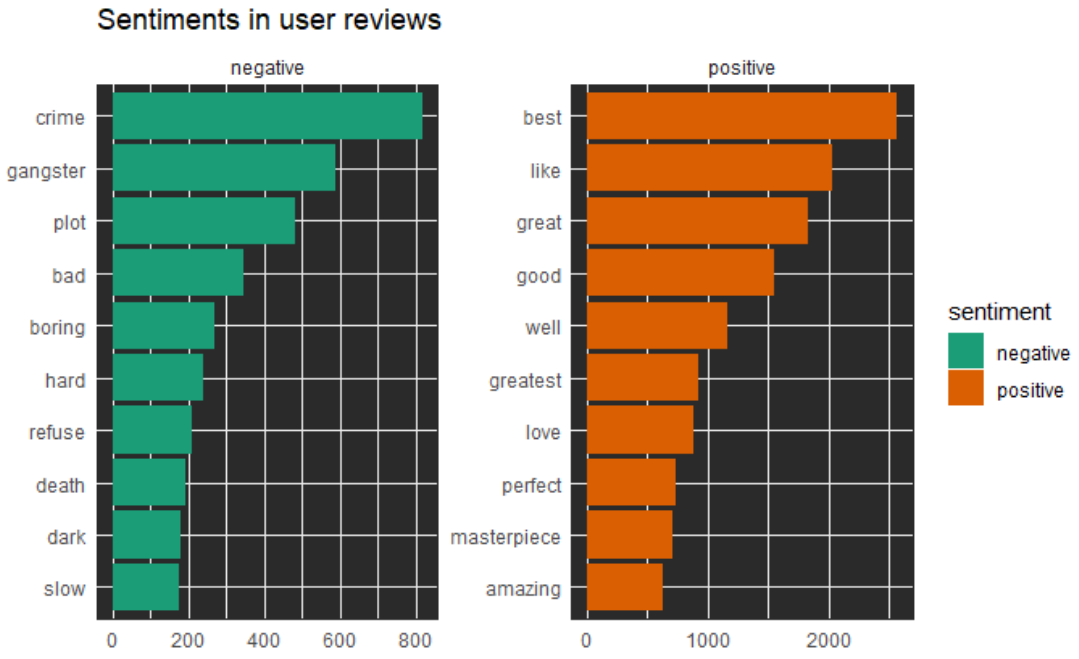
	word <chr>	n <int>
1	the	39653
2	and	17522
3	of	16985
4	is	14379
5	a	14165
6	to	12398

6 rows

By unnesting the text into words, word clouds are computed to see what the most used words in the reviews are, and hence to see the sentiment of the user reviews. However, we see that the words used most are all stop words which are the most common words in the language. Therefore, we need to filter out the stop word and obtain the word cloud as follows:



Clearly, the most mentioned words are related to the movie, such as godfather, story and acting, the character and actor names: Corleone, Brando, Michael, etc., and the topic of the movie: family, mafia. Overall, we see very positive feedback from the reviews, where the most words mentioned are closely related with the story of the film. This might suggest that people are very engaged with the story, and this somehow explains why people gave the high rating scores.



Furthermore, by using the sentiment lexicons of “bing”, the analysis is conducted based on the sentiment of users and the top 10 sentiment words are listed for most positive and negative sides. For the positive side, words such as “best”, “great”, “masterpieces” are frequently used. These are all very high regarded words. For the negative side, words such as “crime”, “gangster” and “death” are mostly mentioned. However, these words are all characteristics that relates to the story of the film which does not necessarily reflect the sentiment of the reviews, but we do see some words such as “boring”, “dark”, “slow” as negative sentiments. This shows why some viewers of the film gave the low rating. Overall, we see that the count of positive words is more than double of the negative words and that to some extend reveal the reason why the godfather is highly rated by viewers.

Conclusions

In general, the average rating of a movie varies across different genres of the movie, where the documentary and biography have the highest average rating, while the more popular genres such as actions, thriller have lower average ratings. However, when exploring the relationship between voting and average rating, we discovered that there are barriers in the data when the rating is in extreme high or extreme low ratings when few voting is involved. When having a large number of votes, the movie is more likely to achieve a reasonably high score. Secondly, the trend of the average rating for science fictions is decreasing, given the growth in movie technology in the recent years. The hypothesis testing also support to reject the null that the science fictions in recent decades are better. At last, by conducting the sentiment analysis for the godfather movie, we discovered that people are fascinated by the story and hence give very high rating scores.

Reflection

The first part of the analysis is conducted using Tableau. The data wrangling process is cumbersome due to large size of the dataset, but it is a good practice in using Tableau. In this part, it is shamed that the dataset does not contains any information regarding the box office details and the country-of-origin details. Further analysis on the world distribution and profit analysis could be done based on these data. The second part of the analysis is about the text analysis. It is a good learning as this is the first time conducting a text analysis, while the traditional analysis is all based on quantitative factors.

Bibliography

OpenMind. 2021. *7 Advances in Technology that have Revolutionized the Film Industry | OpenMind*. [online] Available at: <https://www.bbvaopenmind.com/en/technology/innovation/7-advances-in-technology-that-have-revolutionized-the-film-industry/> [Accessed 9 September 2021].

Dataset

IMDb. 2021. The Godfather (1972) - IMDb. [online] Available at: [Accessed 10 August 2021].

IMDb. 2021. IMDb Top Rated Movies. [online] Available at: [Accessed 10 August 2021]

R packages

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2021). naniar: Data Structures, Summaries, and Visualisations for Missing
Data. R package version 0.6.1. https://CRAN.R-project.org/package=naniar

Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." _JOSS_, *1*(3). doi:
10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL:
http://dx.doi.org/10.21105/joss.00037>.

Tierney N (2017). "visdat: Visualising Whole Data Frames." _JOSS_, *2*(16), 355. doi:
10.21105/joss.00355 (URL:
https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.

David Robinson (2021). gutenbergr: Download and Process Public Domain Works from Project Gutenberg. R package version 0.2.1.
https://CRAN.R-project.org/package=gutenbergr

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL
https://www.jstatsoft.org/v40/i03/.

Ian Fellows (2018). wordcloud: Word Clouds. R package version 2.6. https://CRAN.R-project.org/package=wordcloud