Deliverable 2: TMDB Movie Dataset

4/24/2019

Group 3: Junyi Chen, Zhe Gu, Leo Li, Xu Ning, Yiyu Yang

Background

When we see a trailer, we could decide if we want to watch the movie. How could movie companies foresee success before the release? Is there a formula for "successful movies"? Since movie production cost is quite expensive, if the industry could figure out "the formula for success," there would be better predictions on how to strategically produce all-time favorite movies. Then how to define the success of a movie, by its online ratings or box office? For this project, our team uses the data from The Movie Database (TMDb) website to explore "the formula for success" in filmdom and provide useful research support for filmmakers.

Narrative of the Dataset

The sample size of this database is 4803, involving 20 variables. The variables we are going to dig into include necessary information about the movie, investment budget, producer's information, revenues, and so on. The types of variable are numeric, textual, categorical and date. All the data included is up to date at the end of 2018.

Statement of the problem

Our group examines how different variables influence the revenue and the voting score of the film. This paper aims to analyze the movies data with various graphics and gives an interpretation of these data. After the analysis, we will conclude with the final findings on which specific variables contribute the most to the result of the film (revenue, voting score).

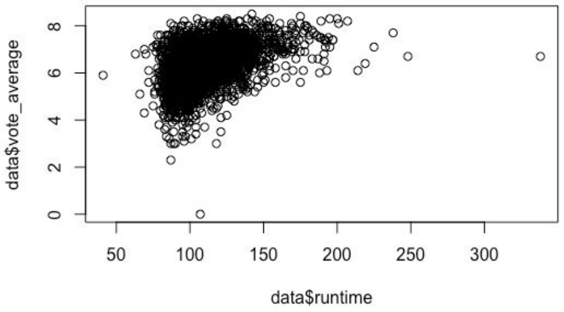
Reasons behind the choice of analytical techniques.

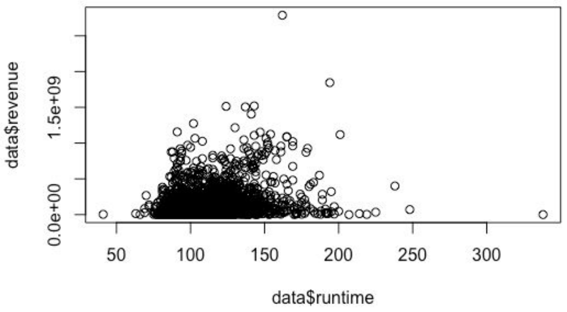
In this analysis, we used four primary analytical techniques: clustering, text mining, sentiment analysis, and data visualization. Clustering is very helpful in an initial exploratory analysis, as it groups movies with similar features and provides insights. Text mining is a must for our dataset because our dataset contains many text data in variables keywords, overview, and taglines. Text mining will be able to extract information from the text and explore possible relationships between the variables and movie revenue & ratings. Sentiment analysis will be built upon the result from text mining and further used to examine the positivity and emotions involved in those text variables. We use data visualization not only to summarize and demonstrate findings but also makes it easier for future review and examination.

Variable Analysis

1）Runtime

Runtime is an essential feature of a movie, and we would like to explore the correlation between the runtime and the movie's vote_average and revenue. Surprisingly, the correlation between runtime and vote average reaches 0.407 (graph on the left), which means there is a moderate correlation between the length of the movie and the rating of the movie. However, the correlation between runtime and revenue is much weaker at 0.245 (graph on the right).





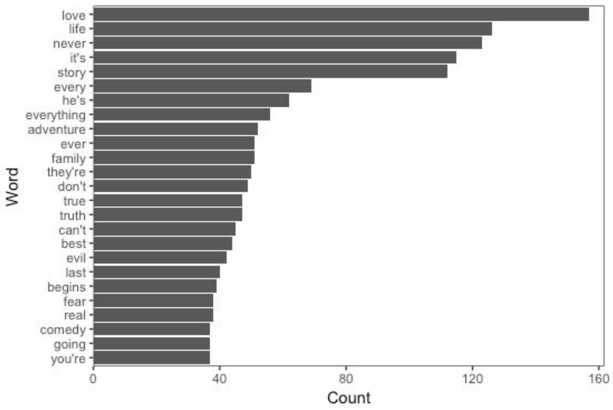
In conclusion, the longer the movie is, it is more likely to have a higher rating, but it is hard to infer the revenue of the movie from the runtime.

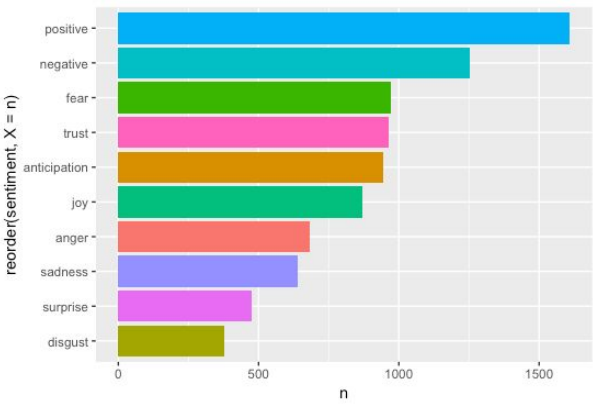
2）Tagline

A tagline is a short text that movies use to promote and market themselves. It is usually a summarization of the movie. We use text mining to look into different taglines and see if the word usage in taglines could affect movie ratings and revenues.

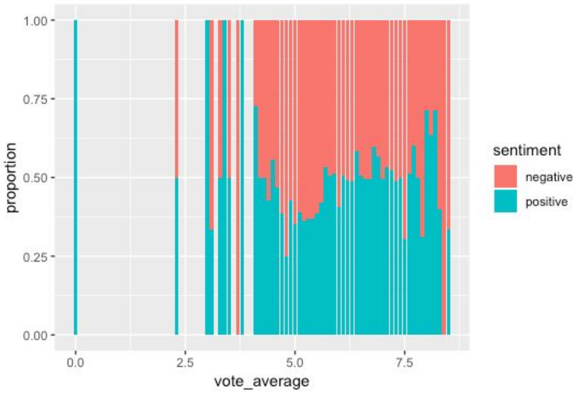
First, we examined the length of the tagline. The average characters of a tagline are 41.82, and the average number of words is 7.76. The shortest tagline contains only one word: "Ka-ciao!", also, the longest tagline has as many as 53 words (how can they fit that into a poster?). The correlations between the length of the tagline and the movie's rating and profitability are at 0.11 and -0.10. Both are very weak.

We also looked into the most common words used in taglines. From the word count graph, we can tell that positive words such as love, life, family and true are widely used, while some negative words including evil and fear also find their place.





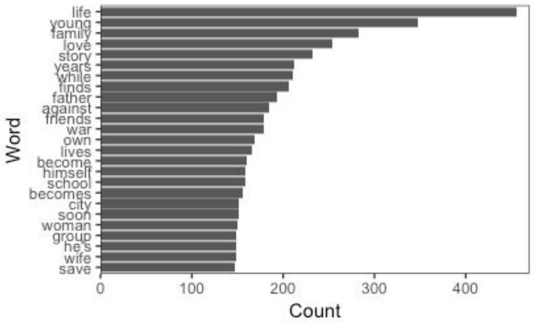
We conducted sentiments analysis to examine emotions within the taglines. We can tell that although we have more positive emotions than negative ones, the most common emotion type appeared in taglines is fear. After fear, there comes to trust, anticipation, and joy.



Do positive and negative emotions of taglines affect movie rating or revenue? We draw a graph between the sentiments and movie scores, and from the figure, there isn't a clear correlation. It's also confirmed by the correlation between positivity and vote average -0.0263. It's the same case for emotions on movie profits. The correlation between positivity and movie profit is also -0.0263, meaning that there is no correlation between the two.

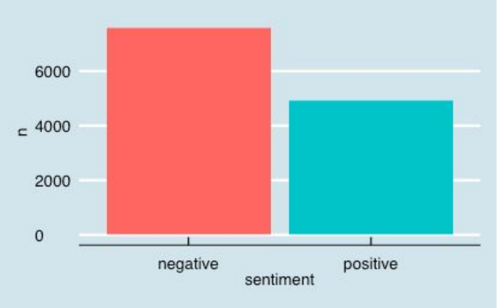
For movie taglines, we can conclude that movies use words with negative emotions (especially fear) more frequently than we expected, and there is no clear correlation between the positivity in taglines with movie ratings and movie profit.

3）Overview

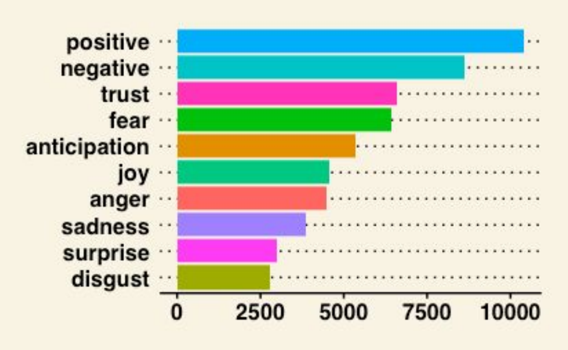


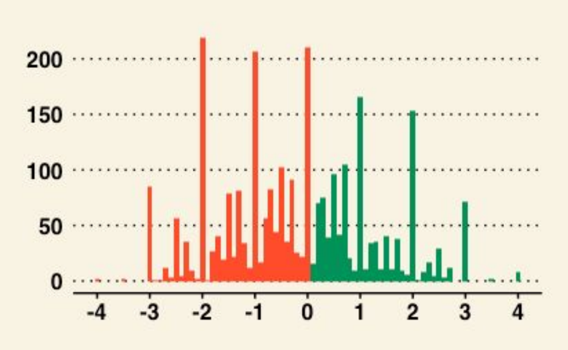
The overview describes the plot of the film in several short sentences. An attractive overview will probably raise the public interest to watch the movie. Hence, it is a crucial feature to explore. By doing a word count, we found that the longest overview contains 175 words, the shortest one has 9 words, and the average is around 52 words. Then, we explored the correlation between overview length and the factors of vote average and revenue. The results show a weak relation between them, with a correlation coefficient of -0.02 and 0.05 respectively. We also found the most common words in all overviews; the plot above shows the top 25 common words, in which the most common word is "life."

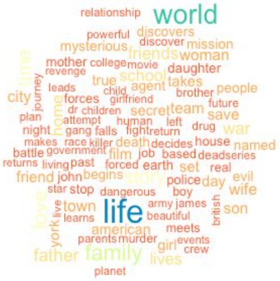
Then we conducted a sentiment analysis. We first use the "bing" lexicon to classify all the words into "positive" and "negative" and compare the total number of each. The number of negative words is 7,564, much more than positive words at 4,895. We also examined the relationship between the proportion of positive words and the vote average and revenue, and it turned out to be a weak correlation coefficient(both around 0.03) between the percentage of positive words and those two factors.

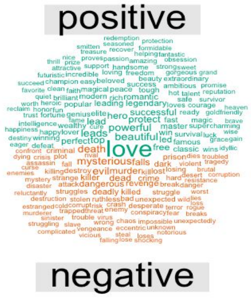


To further analyze the sentiments among overviews, we categorized the words based on emotion conveyed. The first chart below shows that instead of generally saying there are more negative words, the words represent specific emotions such as trust and fear appear most commonly in overviews. Then we scored words based on the extent to which they are positive or negative. The findings are in the second chart below, showing that the overall sentiment tends to the side of negative. We can also see that the words scored 0, -1, -2 and 1 are the most common words.







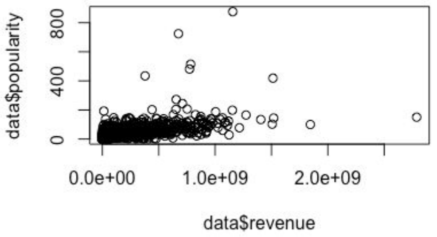


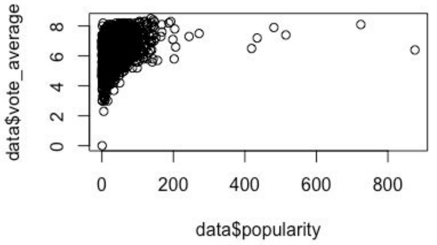
To gain a clear view of the word count ranking and the sentiments they represent, we constructed the word clouds. From the results, we conclude that the most common words after excluding the stop words such as "the" and "with," are "life" and "world."
We build a comparison cloud, showing that the most common word used in the overview with positive emotion is "love."

4）Popularity

Popularity is an index, describing how popular the film is among the audience, and we would like to explore the correlation between the popularity and the movie success. The correlation between the popularity and the revenue reaches 0.58, indicating a moderate to a strong correlation; while the correlation between popularity and vote_average is 0.297, slightly weaker relation than the previous pair.

It is inferred that the more popular the movie is, generally, it is going to earn higher revenue, but the vote_average score does not have so robust correlation with the movie's popularity.

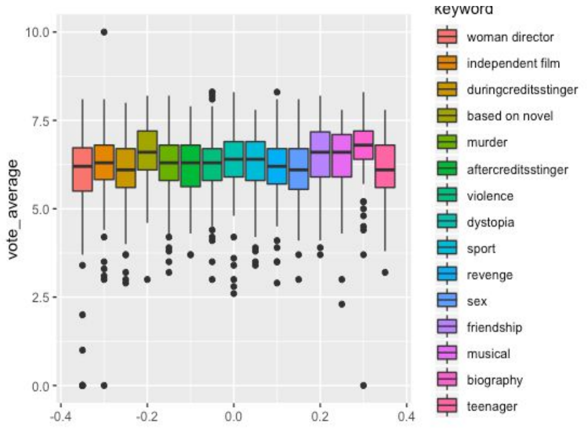


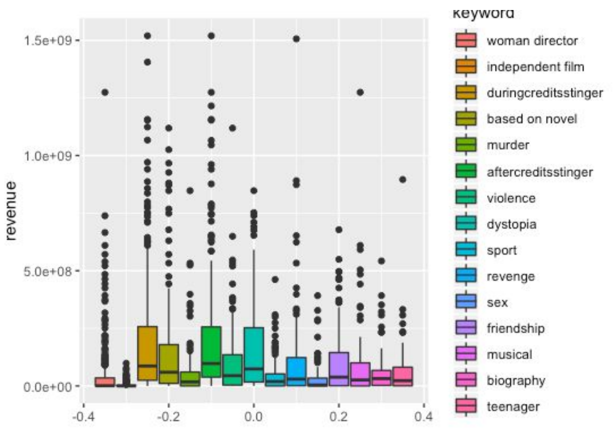


5）Keywords

The keyword is a list of string data, and it describes the theme, background, scene and other key characteristics of the movie. We reorganized data based on keywords to explore how they will affect movie rating and revenue.

Firstly, we filtered out the top 15 most frequently used keywords. "woman director," "independent film," and "duringcreditsstinger" are the top 3 most used keywords. After that, we plot the keywords along with movie ratings and revenue.



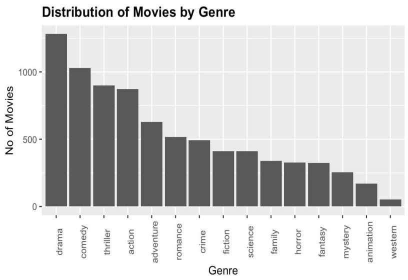


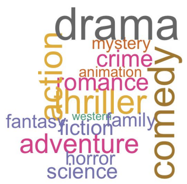
The findings are very intuitive. As for movie ratings, we can tell that biography (median rating 6.8) and movies based on a novel (median rating 6.6) usually have a higher score than all others. We assume that it's because those movies are based on better stories and plots. On the other hand, movies with teenager tags (6.1) usually have much lower ratings.

In terms of movie revenue, it's an entirely different story. This time movies with keywords aftercreditsstinger (median revenue $97 million), duringcreditsstinger ($86 million) and dystopia ($74 million) have the best performances. In contrast, more than half of the independent films have extremely low revenue due to their niche market among professional film critics or independent theatergoers.

6）Genre

We look at the number of movies by genre, sorted by count. Drama is the most prevalent genre, followed by Comedy, Thriller, Action, and Adventure. Some of the categories have many fewer films (less than 100) and so due to sampling error may not be as reliable for assessing overall trends.

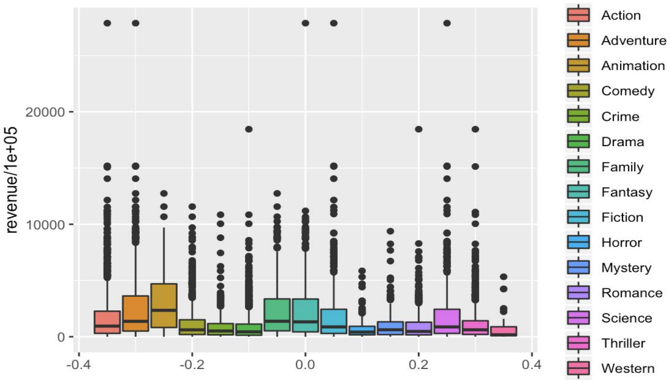


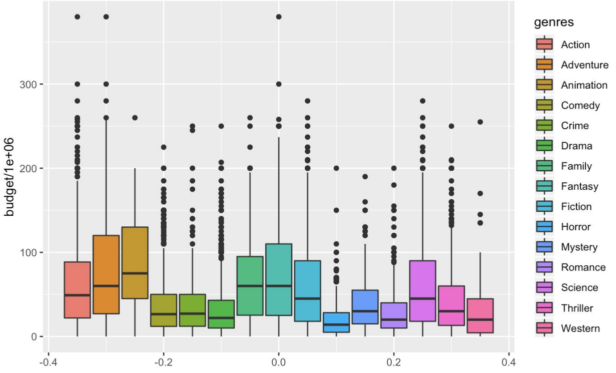


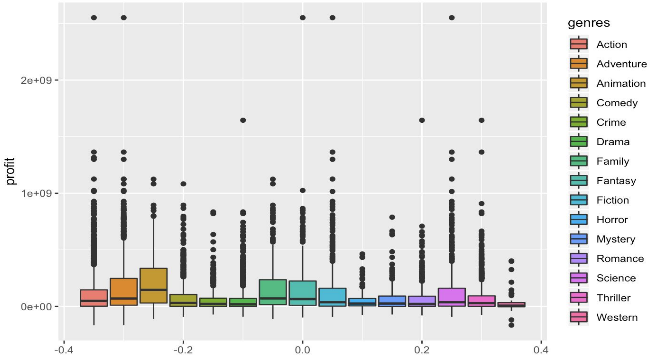
These box-plots show the distribution of budget, revenue, vote_average, profit, run_time, popularity, vote_count of each genre. It shows that, at the median, Animation has the highest revenue, followed by Adventure films. Next, Action, Drama and Family films show similar central tendencies. Movies from the same genre tend to cluster within a similar revenue range, while for each genre, outliers are mostly on the high end, outperforming others.

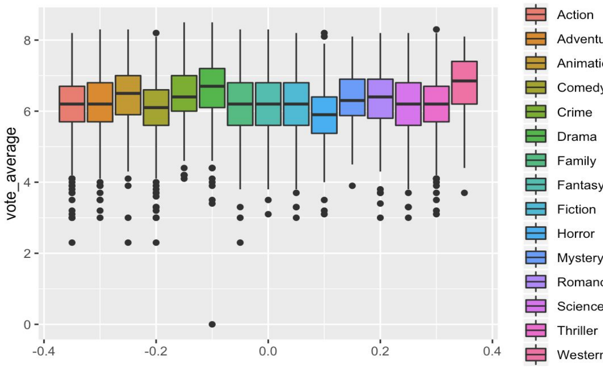
We also note differences in the tails and variance of the distribution, as some genres have a more substantial variance in terms of budget, revenue and thus profit, while movies from other genres share more similarity in these financial aspects.

In terms of the distribution of vote_average across genres, movies of the same genre tend to cluster within a certain range, while for each genre, outliers are mostly on the low end, and get extremely low scores.



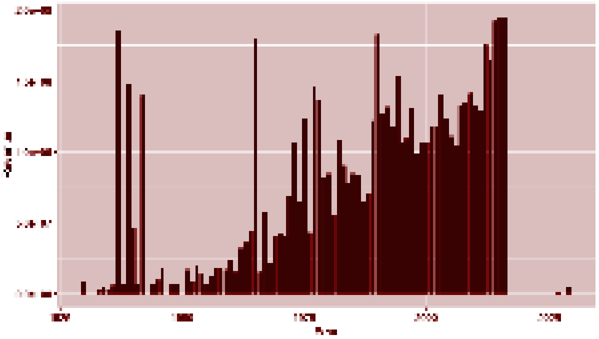


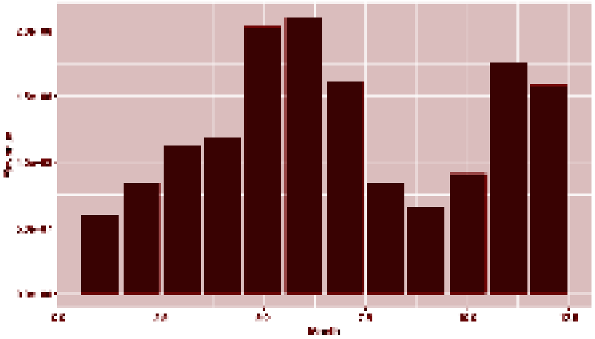




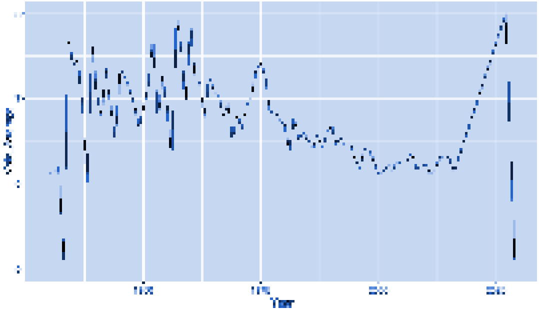
7）Released Date

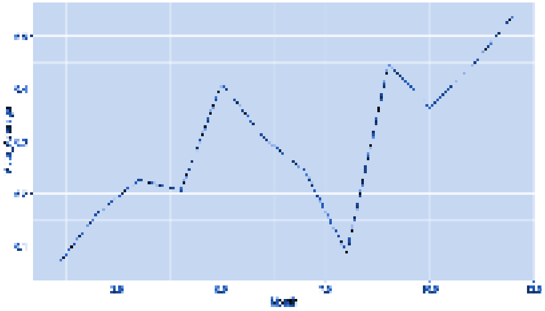
A release date is a date that indicates the premiere of artistic production and its presentation and marketing to the public. Therefore, we are curious about that is there any underlying rules between the released data and the success of a movie.





To more accurately observe the influence of time, we explored the released year and month respectively. Observing revenue from the perspective of the time dimension, we find that the box office increases with the increase of years in trend. Specifically, there are four-time nodes of the outbreak in 1937, 1965, 1990 and 2015. However, the box office shows regular periodic fluctuations in a year. Mid-year and end-year have the highest box office, while other months are relatively depressed.

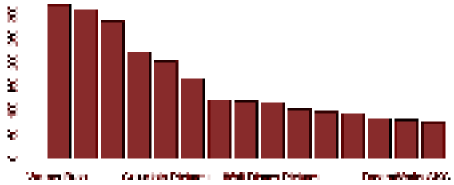


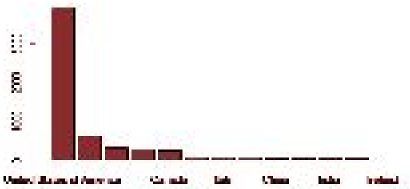


As for the movie scoring market, from the released time analysis we can see that TMDb scoring has no apparent positive correlation with the year as the revenue has, but the fluctuation of the scoring is getting steadier as years increase (the data after 2018 is the forecasted, which is not of high reference value because of the small amount of data.). Moreover, among the month, the ratings of the audience present the same periodic fluctuation with the box office.

8）Production Companies and Countries

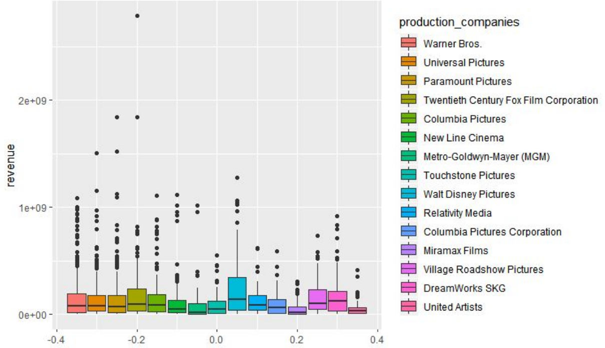
The production companies and countries of movies are likely to mark the quality of a movie, such as Disney, Warner Brothers, which we know well, have released many classic films works. We can discover that Warner Brothers are the most significant production company in the world now and

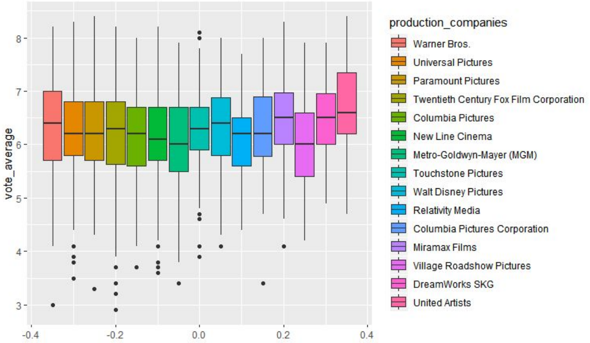




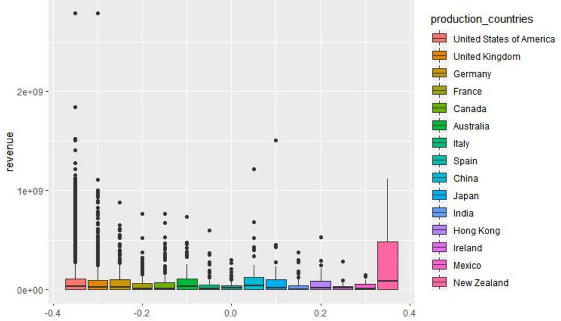
the U.S. has far more movie production than other countries, accounting for 90% of the global movie productions.

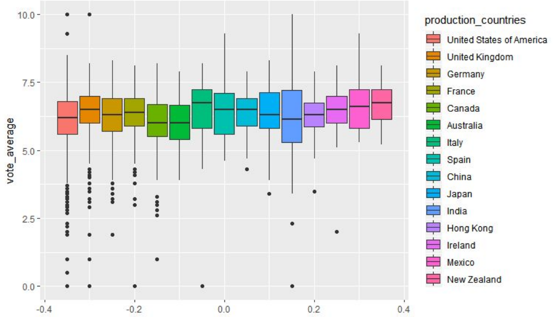
However, a higher quantity does not equal higher quality. From the revenue side, Walt Disney and DreamWorks occupy the top positions. Also, more than half of Disney's movies have a box office that far exceeds the industry average, and it set the highest box office record in movie history. In terms of movie rating, United Artists ranks the top and once set the most elevated rating record. However, movies released by big companies like Warner and Disney have high scores on average. It is worth noting that movies released by Village Roadshow have achieved remarkable results in the box office, while the audience rating has become the lowest.



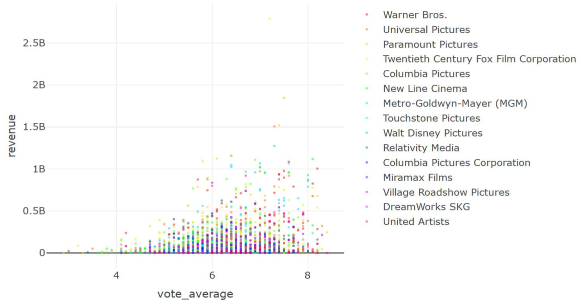


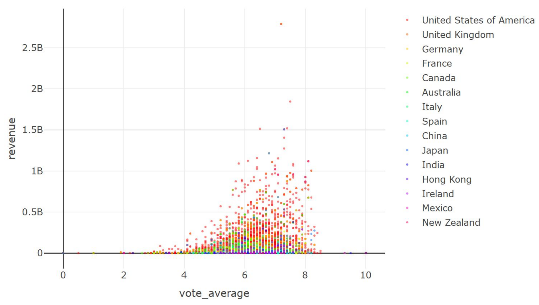
Then, focusing on the releasing countries, we discover that Australia's movie industry is small but with outstanding performance. Movies released by New Zealand has the highest average box office, while Australian movies have the highest ratings. Besides, although the average box office sales of Indian movies are relatively low, they have the most dispersed movie ratings and see both the highest and lowest ratings among all countries.





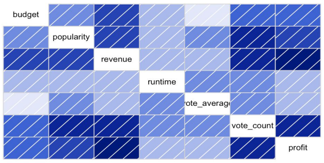
Furthermore, we go further to analyze the relationships between revenue and average scoring depending on different production companies and countries. Walt Disney and the U.S. are the most outstanding company and country that have the highest level of revenue and rating at the same time. Beyond that, one surprising finding is that the highest-grossing movies do not locate on the highest-scoring areas. The whole distribution presents a shape of normal distribution and movie rated as about 7 generates the highest revenue.



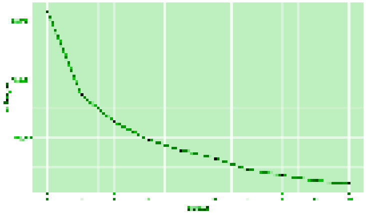


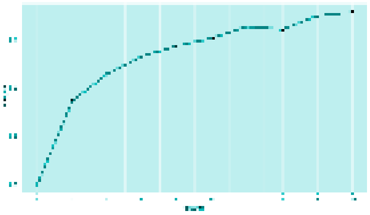
Clustering

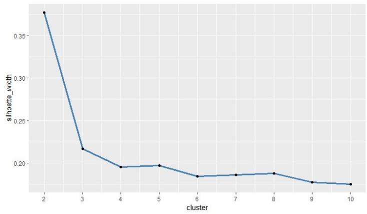
Next, we consider conducting a k-mean clustering for movies using all numerical variables from the dataset. Take a look at the correlation among these variables first.



We used total within the sum of squares plot, ratio plot and silhouette method to determine the number of clusters.

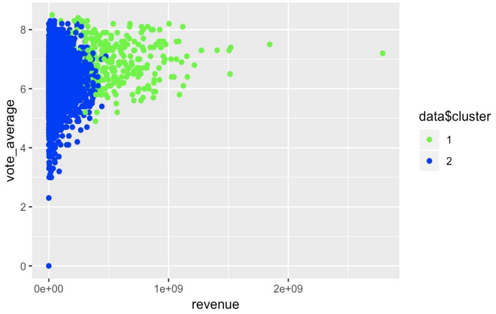


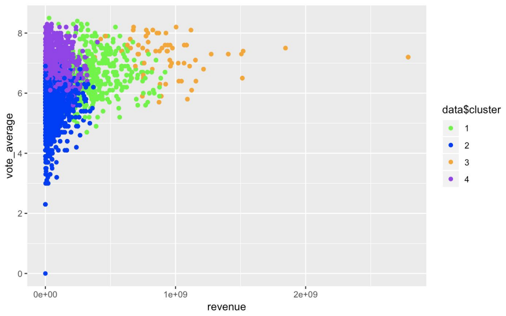




The first two tests indicate that the ideal number of clusters is two, while the silhouette method recommends a four-cluster solution. We chose two clusters as dividing data into four groups would not leave enough data to fit a model.

The clustering result is as below. We obtain the conclusion that the majority of movies are in Cluster 2, which generated low box offices and with high variations in IMDb scores; whereas movies in Cluster 1 have relatively higher gross and ratings compared to other clusters.





Conclusions & Recommendations

This report is mainly an exploratory analysis of the impact of various factors on the movie's box office and ratings. After the analysis, we came up with 3 Ps, three principles for filmmakers to achieve success in today's movie industry.

First, powerful product. In terms of genre, Animation and Action generate the most revenue, while Western and Drama movies have more decent ratings. Through text mining, we find biographical and novel-based movies usually receive higher ratings, while teenager movies have much lower ratings. A movie is a form of dramatic and artistic storytelling, so the content and format both matter a lot. A compelling plot combined with heart-shaking visual & sound effects and groundbreaking film technology are the key elements of success in the industry.

Second, proper perspective. Filmmakers need to have an analytical perspective that some factors may affect the movie success unnoticeably. As we discovered, the runtime has a moderate correlation with movie ratings and box office. Maybe long movies tend to be good, but more likely it is because good movies are usually long. Also, filmmakers should meticulously choose the release date, knowing that movies gain more attention in summers and winters.

Last but not least, persistent innovation. Major film studios such as Hollywood's "Big Five" have developed themselves along with history. For example, Walt Disney and DreamWorks are the most successful producers. Walt Disney, in particular, has an average box office far above the industry average. Movies produced by United Artists Picture on average obtain the highest ratings and once set the highest rating record. The industry has constantly reinvented itself in its over 100-year History. However, Streaming platforms like Netflix, Amazon Studio, and HBO are the latest paradigm shift, moving the industry away from theatres, which perceived as a threat to the "Big Five" major film studios. The market competition will become more robust as more players join in, stimulating this industry which centers around creativity.

The movie business is a blend of art and commerce that evolves with technological advancement, as well as the spirit and the determination of entrepreneurs. Looking back at the data, we discovered factors influencing success in the past. However, as new technology trends emerge and the new movie business models materialize, creative artists and entrepreneurs will pioneer opportunities to expand the industry and redefine the success in the future.

References

Movie metadata from the https://www.themoviedb.org, and https://developers.themoviedb.org/3/getting-started.

This dataset was generated from The Movie Database API. This product uses the TMDb API but is not endorsed or certified by TMDb. Their API also provides access to data on many additional movies, actors and actresses, crew members, and TV shows.