Identifying Historical Trends in Movies

Charvi Agarwal and Julie Hong

April 3, 2023

1 Overview

IMDb, as an Internet Movie database, is a website that provides overall information of films and television programs. Development in digital visual and other technology over time, watching moving features on the screen has established itself as one of the important economic factors in people everyday life. This report is analysing IMDb data for films release from 2006 to 2016 about the impact of different factors on films' evaluation and how accuracy of a prediction on box office success affected by them. The data set is demonstrated on different type of statistical figure such as table, bar graph, histogram, and scatter-plot, which depicted trends and scale of impact of different columns on the ratings. For analysis on prediction of box office success of a film, a machine learning model was used to predict the revenue of a movie with non-commercial factors.

2 Introduction

Context and motivation The art of cinema has been around since the late nineteenth century. However, the Golden Age of cinema only came about in the 1930s. With more and more films being produced today due to the evolution of digital technology, it is evident that not every film can make it big. We are interested in finding out what exactly makes a movie a ”big-hit” and a box-office success. Particularly, we will look at the ratings on the IMDb website paired with things like revenue, 'star power', most popular genre's etc. Through this analysis we can attempt to find out the secret to film success.

Previous work

- Debates and Assumptions about Motion Picture Performance A Meta-Analysis: a paper conducted by researchers at UTS, HEC Montreal and the University of Cambridge[3]where they compared various factors to find the formula to box office success. It was found that even with great actors and stellar reviews a movie will not perform well unless it was distributed well.
- Early Predictions of Movie Success: the Who, What, and When of Profitability: a study conducted by researchers at the University of Iowa[5] to determine the profitability of a movie given a multitude of factors. It concluded that these factors can in fact measure whether a film is worth investing in or not.
- Predicting Movie Success at the Box Office - Chapter 2 <Are the Most Expensive Movies the Most Successful?>: a paper done by researchers at University of Leicester[4] It concluded that there are many different factors significantly affect predicting box office 'hit' or 'flop', which makes it difficult to predict the box office performance of a film, whereas production expense is directly influential to the factors.
- Pre-production box-office success quotient forecasting: a study done by researchers at the Capital University of Science and Technology, Islamabad, Pakistan [2] They found that despite of large size of investment, majority of investment hardly cover production budget, significantly affected by post-release and post production forecasting - such as movie critics

Objectives We are looking to find the answer to what factors determine box office success for a film. Particularly, we are planning to ask the following:

- Do bigger names(in terms of directors) have better rated movies?
- Is there a relationship between number of votes and rating?
- Is there a company that is outperforming the others in terms of revenue on their films?
- Can we predict votes and ratings will lead to box-office success?

3 Data

Data provenance The main dataset used for this project is owned by Iván González and was made available on the Kaggle website [1], which was downloaded as a CSV file format. The data set is usable under the CC0: Public Domain license, meaning that we can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. We also supplied a data set called "Movie Industry"[6] created by Daniel Grijalva who used the data on the IMDb website and made the data available on the Kaggle website. It was downloaded as a CSV file and it also has the CC0: Public Domain license.

Data description The main data set came in the form of 1 CSV file with 12 columns: Title(Name of the film), Genre(Categories that define the film based on narrative or stylistic elements), Description(Brief synopsis of the film), Director, Actors(List of main stars in the film), Year(Year the film was released), Runtime(The duration of the film in minutes), Rating(User rating given between 0-10), Votes(The number of users who left a vote on the film), Revenue(The amount of money the film made), Metascore(An aggregated average of critic scores. Values are between 0 and 100. Higher scores represent positive reviews). The second data set also can in the form of 1 CSV file with 15 columns. 12 of the 15 columns were the same as the above data set except Revenue(Millions) was called gross instead. The 3 other columns were: Rating(a classification of a film according to the age of viewers though suitable to see it), Country(the origin place of the film) and Company(the body that produced the film).

Data processing There was little cleaning to be done in the main data set. There were 128 null values in the Revenue(Millions) column and 64 null values in the Metascore column. These rows were filled with the median data from the other columns. For the other data set, we only used the company and the revenue column to find the total revenue the top 15 companies make. There were 17 null values for company and 189 null values for revenue. The null values for column were dropped and the null values for the revenue column were filled by finding the median of the revenue.

4 Exploration and analysis

We began by trying to analyse exactly which features had an impact on one another (Figure 2(a)). This gave us an idea on which factors to explore. We were particularly interested in identifying the effect of Revenue(Millions) and Votes on Rating. The heatmap (Figure 1) illustrates that there is not a strong correlation between Revenue(Millions) and Rating as it is only 0.22. However, we were curious to take a further look into this. To analyse the relationship between revenue and rating of movies, we have set a hypothesis(H_0) that the revenue has an impact on the popularity, measured by the rating in this analysis. The line regression in Figure 2(a) was used for a predictive exploration of revenue on rating of films. Statistically analysed, and it depicted that the correlation between the two variables are scattered loosely in along the regression line. On top of this, the P-value of the relationship was found to be extremely low, $p < 0.0001$, which proves that the revenue does not have sufficient level of coefficient with rating, or the null hypothesis is not likely to be the right observation of our data set.

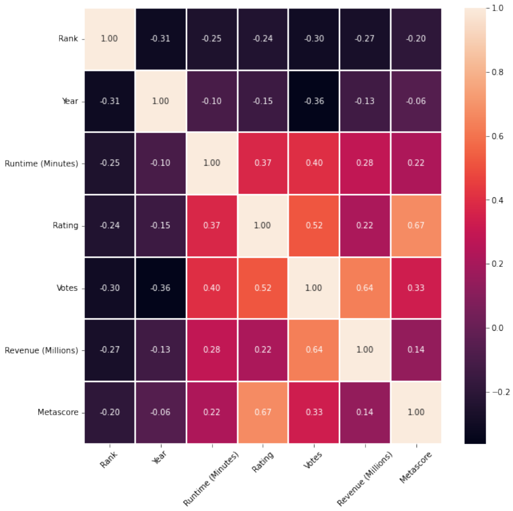
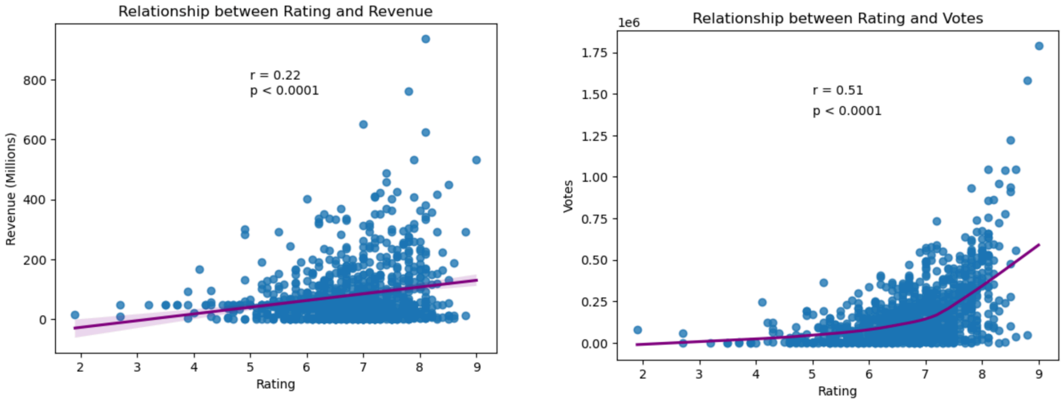


Figure 1: This is a heatmap illustrating the correlation between all of the factors in the data set.



(a) a line regression of Revenue (Millions) and Rating columns of the dataset. A clear relation is observed that the movie with larger revenue is likely to be highly rated.

(b) a line regression for votes and ratings. A clear trend of the correlation between votes and ratings is depicted by the regression line

Figure 2: Illustrates bar graphs showing the top 10 directors.

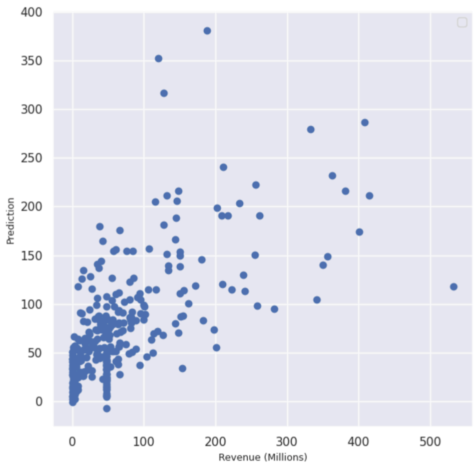


Figure 3: The Machine Learning prediction model trained with data of Rating and Votes columns. The scatter plots demonstrates accuracy of the prediction made by the two factors.

To analyse if votes are influential on the ratings of the movie, another line regression was performed for observation on votes making prediction on ratings based on our data set (Figure 2(b)). While the regression line depicts correlation of votes and rating. It shows steady and gradual increasing trend and comparatively steep rise for the ratings bigger than 7. While the p-value is extremely low as told by $p < 0.0001$, the scatter points plotted significantly intensively throughout the line and outstanding trend observed from the regression line, it was found that the votes are highly influential on rating, or the popularity of the films.

Following the correlation between vote and ratings, we decided to explore further on the factors that could predict the box office success (Figure 3) . This phase of data analysis started with us setting an basis assumption that Revenue (Millions) column of our data set corresponds to ’hit’ or ’flop’ of a movie in a box office. Exploring correlations among revenue, rating, and votes columns’ data, we wanted to see if the same set of data of rating and votes columns would have impact on revenue of the movies, while all factors are influential on one another according to the heatmap (Figure 1). First, we created a machine learning model and trained it with random 70% data of the two columns, ’Rating’ and ’Votes’, in order to make it predict the revenue based on rating and votes on the films. By plotting the prediction as a scatter plot graph, we attempted to check whether the two factors are appropriate predictor for the film’s success. Observing the figure, it was found that the number of case where the prediction is accurate, $x\ \ value = y\ \ value$, is proportionally small. This analysis brought us the result that bigger data frame is needed for more case of prediction to be observed as well as that the two factors can predict revenue.

As another possible predictor, we came up with production companies. The idea of this was that more famous and popular production companies will produce movies with higher revenues. We have demonstrated the data as a bar graph (Figure 4) and it was observed that the plotted names of companies in the data set are well-known production companies to the public. To create the figure, the data set of companies was listed in order of the one making largest revenue to the lowest revenue. To make it easier to compare and observe the trend among the companies, top 10 companies were illustrated on the bar chart. Assuming that the familiar and famous companies have more trustworthy and positive images to the public, the name of the film production company affects revenue of the films.

As an extension from the previous findings of the analysis, directors and the number of movies were examined as we were curious about the influence of famous directors in the movie industry. Organising the data for directors’ names and the number of movies in the dataset they have produced in a table helped the analysis (Table 1). From the table, it was observed that the director Ridley Scott created more movies than any other directors in the IMDb dataset. We wondered if this affects the revenue, and we have listed the name of directors of the movie with the most revenue(Table 2). It turned out that none of the directors

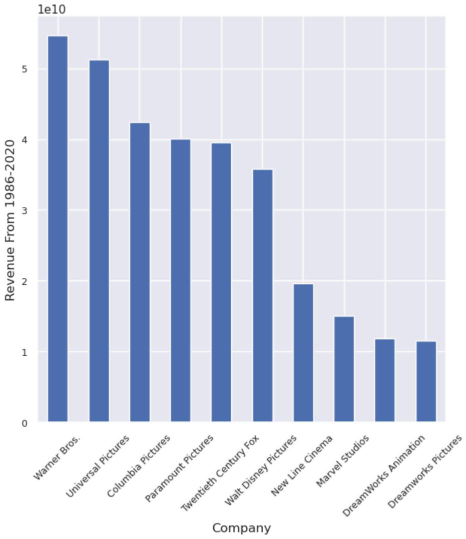


Figure 4: A bar graph of Top 10 companies and their revenues

Director	Count
Ridley Scott	8
David Yates	6
M. Night Shyamalan	6
Paul W.S. Anderson	6
Michael Bay	6

Table 1: Top 5 Directors who produced the most movies

Director	Revenue (Millions)
James Cameron	760.51
Colin Trevorrow	652.18
Joss Whedon	541.135
Lee Unkrich	414.98
Chris Buck	400.74

Table 2: Top 5 directors who produced Largest Revenues (Millions)

in Table 1 can be found in Table 2. This shows that comparatively large share of movies of a director does not affect on the success of the film in the box office.

5 Discussion and conclusions

Summary of findings This study has evaluated the success of a film in terms of its commercial profit and popularity. Rating was used as a measure of popularity of a film, and it was observed that votes has clear correlation with rating. For further analysis using the same part of data, revenue, rating, and votes, it was found that the box office success of a film can be predicted by ratings and votes, but with a limitation on the accuracy of the prediction. To find more appropriate predictor, an external data set of film production companies was used and resulted that the famous the companies are, the higher the profits they make. Last but not least, our curiosity on whether the names of directors affect revenue of a movie was proved to be not likely to occur with a clear answer that there is no remarkable correlation between the two factors, directors and Revenue (millions).

Evaluation of own work: strengths and limitations While analysing the data set of IMDb from 2006 to 2016, the main limitation that we encountered was the lack of data. The data set includes different factors that comprehensively determining the rank of the movie which made it easy to decide which part of the data we should focus on to answer the questions, or the objectives, we made for the analysis. To

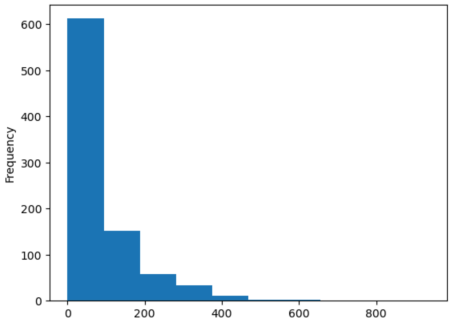
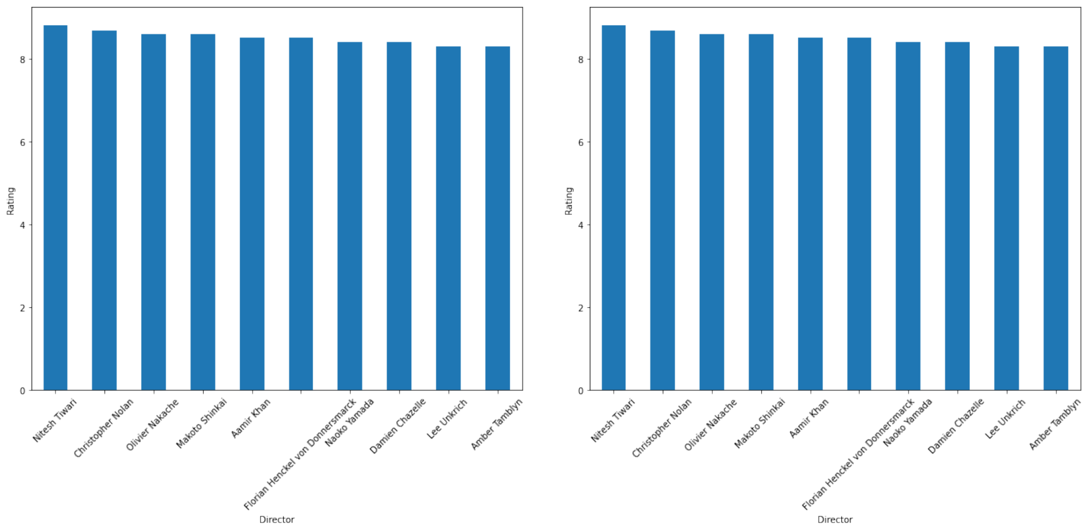


Figure 5: Histogram of distribution of null values of Revenue (Millions) column. Median = 47.985 and Mean $\approx$ 82.96. As Median < Mean, the histogram is right skewed and the null values are replaced with Median of the column.

engage the accuracy of prediction, we carefully decided with which method we will clean. There were many null values in Revenue (Millions) and Metascore columns, which was another limitation. However, as Revenue (Millions) column was the most important part of the data set for our project, we replaced them with median by checking the skew of the data Figure 5) on a histogram.

Comparison with any other related work As mentioned previously, there have been many papers that have analysed the factors that lead to box-office success. Indeed, our own findings aligned with the ones found in the papers. For example, the Meta-Analysis conducted by researchers at UTS,HEC Montreal and the University of Cambridge found that "star-power" seems to be the most telling aspect for box-office success. We went a step further by analysing directorial influence on this. However, they do mention that if the movie has been out for a while, the pull of a popular star tends to wane. Furthermore, the study conducted by the researchers at the University of Iowa also mention the fact that popularity of a film is based on big names concerning actors and/or directors. Our research suggests that factors such as rating and votes also effect the revenue but this fact is not bi-conditional.

Improvements and extensions For our extension we will explore the influence of directors in the movie industry. We want to find out if big directors have an effect on Ratings and Revenue(Millions). Starting out, we examined the top 10 rated directors. This was done by grouping the Directors and Ratings column and finding the mean rating for each director for the films they have made. Then, we took the top 10 from those calculations. As seen in figure(Figure 6) we can notice that the top rated directors have similar ratings.However, when we take a look at figure (Figure 6) we can see there is a big variation between the mean revenue each top rated director makes.However, it is also evident that the top rated directors are also the ones who make the most revenue. Therefore, it is evident from these graphs and analysis that big named directors have an influence on the movie industry as they are the ones who get rated the best and make the most amount of money.



(a) Top Rated

(b) Top Revenue

Figure 6: Illustrates bar graphs showing the top 10 directors.

References

[1] *1000 IMDB movies from 2006 to 2016. (n.d.).* Retrieved March 21, 2023. URL: https://www.kaggle.com/datasets/gan2gan/1000-imdb-movies-20062016.

[2] Usman Ahmed, Humaira Waqas, and Afzal T. Muhammad. “Pre-production box-office success quotient forecasting”. In: *Soft Computing* (2020).

[3] François A. Carrillat, Renaud Legoux, and Allègre L. Hadida. *Debates and assumptions about motion picture performance: a meta-analysis*. 2018. URL: https://doi.org/10.1007/s11747-017-0561-6.

[4] Barrie Gunter. “Predicting movie success at the box office.” In: Springer, 2018. Chap. 2.

[5] Michael T. Lash and Kang Zhao. *Debates and assumptions about motion picture performance: a meta-analysis*. 2016. URL: https://doi.org/10.1080/07421222.2016.1243969.

[6] *Movie Industry. (n.d.).* URL: https://www.kaggle.com/datasets/danielgrijalvas/movies.