Sharif University of Technology
Machine Learning in Physics project:
Success in Movies
Phase 1: Data Collection

Ali Setareh Kokab , Reyhane Ghanbari

April 2, 2021

Contents

Abstract

In our project, we try to develop a learning model that can successfully predict movies' commercial success before they officially release. The first phase of our project is collecting data and cleaning it. In this phase, we try to find related factors to the success of a movie. We use IMDB.com and boxofficemojo.com to obtain our desired information about movies released in the US between 2010 and 2019. At first, we had 27 columns in our data set, but after processing our data, including convert non-numerical columns to numeric ones (using one-hot encoding and introduce a ranking system for producers, writers, publishers, and stars), our final data set has 290 features.

1 Introduction

The film industry is one of the most lucrative industries. In 2019, the global movie industry was worth 42.2 billion dollars. As a result, predicting movie sales prediction has always been an attractive problem for the people in industry and academia. Distributors can use these predictions to decide which movie they should invest or use them to make a better decision about the release month of a movie. With its success in various fields, machine learning is one the most promising approaches that scientists have taken during the past decade to tackle this problem. They are variety of ways to use data for movie's success prediction. from sentiment analysis of Twitter [2] to Wikipedia movies' page activity analysis [4]. To develop a machine learning model, the first step is to find the important features and making a data set. To this end, We use IMDB.com and boxofficemojo.com and use Python to make a web crawler to gather information from these sites.

2 Related works

Predicting the success of cultural goods has long been the focus of scientists. Our work's general idea is taken from the Albert-László Barabási et al. 2019 paper, "Success in books: predicting book sales before publication "[5]. They try to predict the success of books published in the US using different machine learning algorithms. They use BookScan data base to obtain information about the books. Their study uses three groups of features: author features, book features, and publisher features. Because the book sales distribution is a heavy tale distribution, they develop their own learning algorithm called L2P (learning to place). In this algorithm, unlike other algorithms, instead of predicting the sales of a book, they try to predict the place of a book in terms of its sales among other books.
In 2013 Márton Mestyán,Taha Yasseri, and János Kertész [4] proposed a model to predict box office success of movies based on Wikipedia activity. They use boxofficemojo.com to gather financial information on 535 movies released in 2010 in the US. Then out of this data, they could find 312 Wikipedia pages of

them and use Wikipedia data dumps to find their activity and use it to predict the box office of movies.
Some factors related to a movie's success on the box office are genre, director(s), producer(s), date of release, viewers' opinions, etc. Anand Bhave et al. [3] try to find these factors and classify these factors into two factors: social factors (like viewers' opinion) and classical factors like directors and producers.
Abidi, S.M.R., Xu, Y., Ni, J. et al. [1] applied different machine learning algorithms on movies data and try to find hidden factors to make a success model before publication for movies. They also suggested a ranking system to turn non-numerical data like directors' names to numerical ones using their movie scores.

3 Data collection

For making our data set, we have used two sources. The first one is Boxofficemojo.com, and the second is IMDB.com. We used the python library BeautifulSoup to read the HTML web pages and utilize regular expressions (regex) to extract specific string pieces. In our study, we focused on the US movie market between 2010 and 2019. So we only pull movies' information that has been released during this period in the United States. To gather the required information, we took the following steps for each year:

1. Go to yearly domestic box office page in boxofficemojo.com
2. Go to each movie page and extract the required information, including the IMDB ID of the movie with Beautiful Soup
3. Go to IMDB page of the movie using IMDB ID of the movie
4. Read the movie IMDB page and extract the information using Beautiful Soup

In fig 1, you can see the schematic picture of this procedure.

Figure 1: Data collection procedure. For each year we should repeat these steps.

4 Data set description

Note that we consider all the movies that went on screen in a year in the United States between 2010 and 2019. So we also have movies whose first initial release was before 2010, but they re-released in this period like Jean-Luc Godard's

movies.
Our data consists of 7096 movies that have been released in the US from 2010 to 2019. We saved our data yearly in separate CSV files. We have two types of features in our data. The features that are obtained from boxofficemojo.com and the ones that are gathered from IMDB.com. Each year data set has 27 columns which we describe below:

- **boxofficemojo features**:
    1. IMDB ID: A unique combination of numbers for each movie on the IMDB website
    2. Mojo ID: A unique combination of numbers for each movie on the box office mojo website
    3. Title: The original title of the movie
    4. Genres: The Genre(s) of the movie
    5. Year: The year which the film went on screen
    6. Domestic Gross ($): total income of the movie in the US in dollars
    7. Worldwide Gross ($): net income of the movie in the world
    8. Opening ($): Opening weekend gross of the movie
    9. Budget ($): the amount of money that is spent to make the movie
    10. Opening Theaters: number of theaters that show the movie when it is initially released
    11. MPAA: the age rating of the movie
    12. In Release (Days): number consecutive days that the movie has been on the screen
    13. Widest Release (Days): the most significant number of theaters that showed the movie
    14. Release date: the date which movie released
- **IMDB information**:
    1. Stars: The names of the first five actors of the movie according to the actors' list on the IMDB site. There are 20474 actors in our data set.
    2. Director(s): the director(s) of the movie. There are 5662 directors in our data set.
    3. Writer(s): the writer(s) of the movie. There are 10270 writers in our data set.
    4. Producer(s): the producer(s) of the movie. We have 33222 producers in our data set.
    5. Runtime: duration of the film in minutes

6. IMDB score: the IMDB score of the movie
7. IMDB votes: number of the people who vote for that movie on IMDB
8. Metascore: the score of the movie on the Metacritic website
9. Meta users: the composition of the people (regular user or critic) who wrote a review about the movie
10. Country: the country(s) that made the movie
11. Language: The original language of the movie
12. Distributor: The company which distributed the movie
13. Plot outline: a short description of the movie on IMDB

5 Data processing

Our data contains both numerical and non-numerical values. For using this data in machine learning algorithms, we should convert non-numerical values to numerical ones:

- **release date**: The release date of a movie is made of a number and a word like '18, Jan'. To convert this data to only numerical values, we first separate the number (day) and then attribute a number from 1 (January) to 12 (December) to each month. So the release date column becomes two separate columns, namely 'Day' and 'Month.'

- **Genre**: For converting the Genre column to numerical values, we first find all genres present in our data set. We have 26 categories in total: 'Action', 'Adult', 'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', 'Documentary', 'Drama', 'Family', 'Fantasy', 'Film-Noir', 'History', 'Horror', 'Music', 'Musical', 'Mystery', 'News', 'Reality-TV', 'Romance', 'Sci-Fi', 'Short', 'Sport', 'Thriller', 'War', 'Western'. We use one-hot encoding to turn these categorical values into numerical values. Therefore, we make a new column for each category in our data set. If a movie belongs to a category, the corresponding genre column's value for that movie is one. Otherwise, it is zero. You can see the number of movies in each genre in fig 4.

- **MPAA**: We also use one-hot encoding to turn age ratings into numerical values. This time we have 9 categories: 'G', 'M/PG', 'NC-17', 'Not Rated', 'PG', 'PG-13', 'R', 'TV-PG', 'Unrated'.So here again, we make a new column for each age rating. We determine which genre a movie belongs to with zeros and ones.

- **Language**: We have 84 different languages in our data set. Here again, we use one-hot encoding and add each language as a new feature in our data set and specify the movie's language with zeros and ones.

- **Country**: There are 146 countries in our data set. The country column specifies the country or countries involved in making a film. Note that a movie can be a joint product of multiple countries. We use one-hot encoding to convert country features to numerical values.

- **Meta Users**: The meta users feature indicates the composition of review writers of a movie. There are two groups of reviewers: the regular people and the professional movie critics. In the Meta Users column of yearly data sets, the number of people belonging to these groups is given. We have divided this column into two columns called "User" and "Critic" in our final data set.

- **producer(s), director(s), writer(s), stars**: To make numerical values from these features, we made a list of names for each column and made new data by multiplying IMDB votes and IMDB score for each person and add all for persons in each movie. Then, to make it more sensible, we divide these numbers into 5 groups and rank each movie from 0 to 4. The higher the rank, the better the movie.

You can find the final data set here. This data set has 293 columns and 290 features (considering Worldwide Gross $ as our Y column). Note that we keep Mojo ID and IMDB ID in this data set to retain movie titles if necessary. However, we do not use them as features in our learning algorithm.

6 Exploring the data

It is always good to see how our data looks like before we feed it to our machine learning algorithms. In fig 2 you can see the correlation between different features.

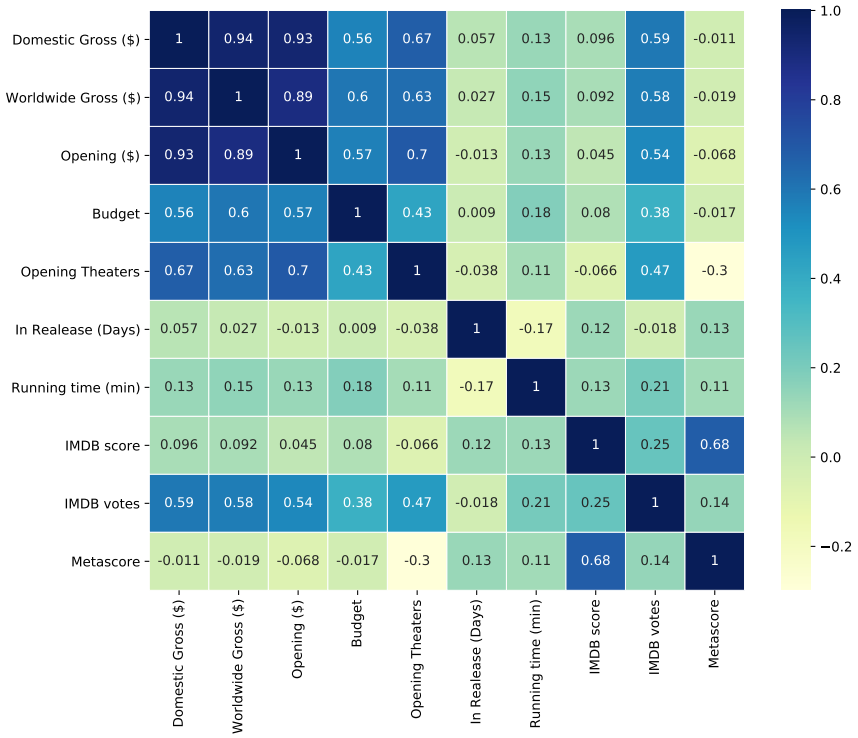

Figure 2: Correlation between different features

In fig 3 you see the scatter matrix plot of some of the features.

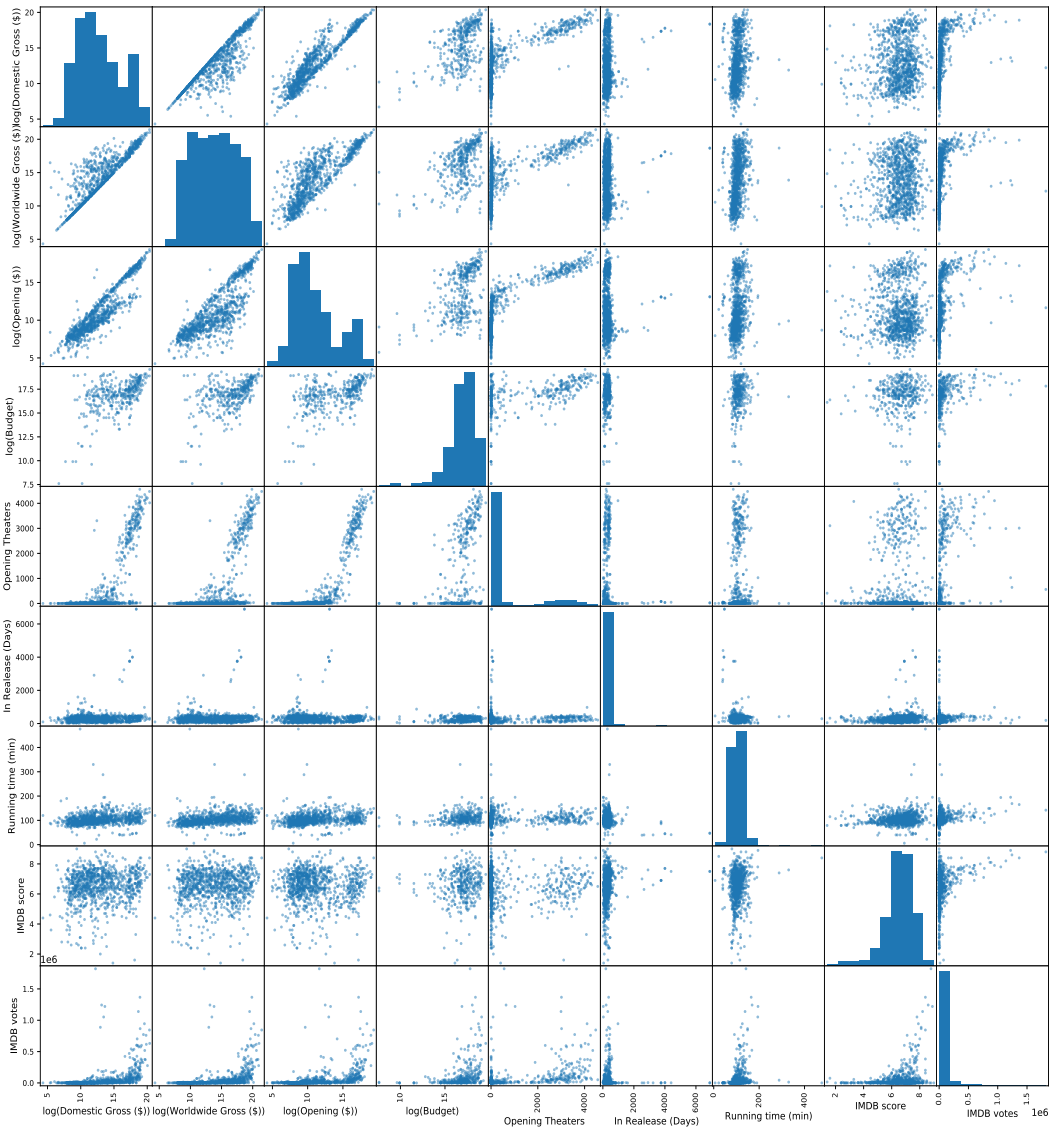

Figure 3: scatter matrix plot of some of the features

In fig 3 we plot the logarithm of some of the features to have better visualization. To see this, In fig 6 and 5 you can see the histogram of the worldwide gross in both linear and logarithmic scale. As you can see, Because this data is so skewed, the logarithmic scale gives us a better view of the data.

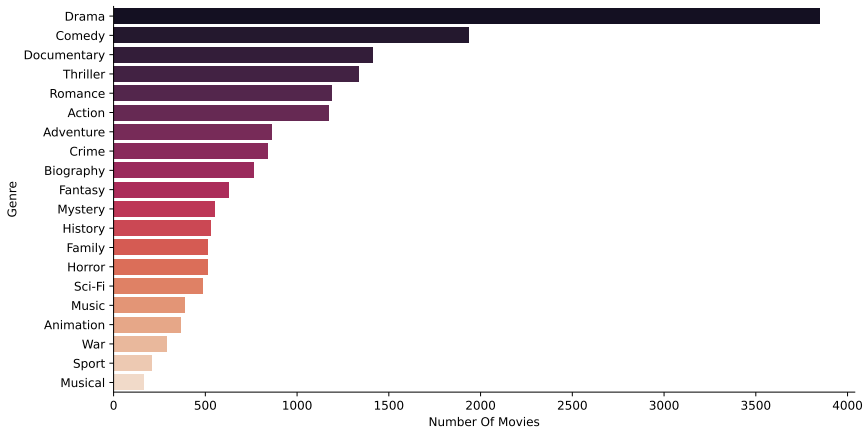

Figure 4: number of movies in each genre

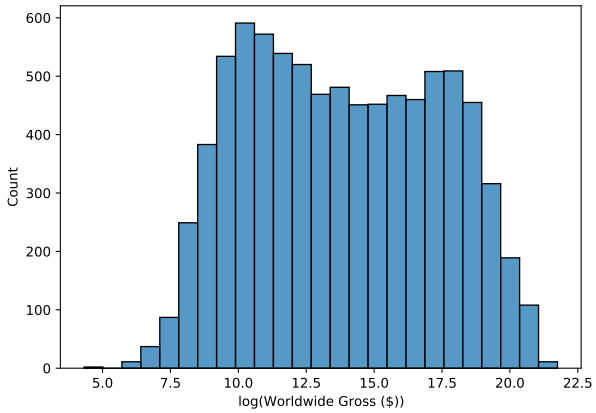

Figure 5: histogram of log (worldwide gross)

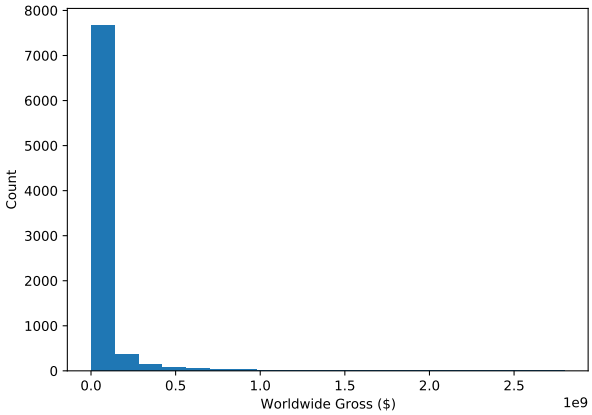

Figure 6: histogram of worldwide gross

In fig 7 you see the box plot of language vs. log(worldwide gross) box plot. In this plot, The languages are sorted from the highest mean gross to the lowest. Fig 8 shows the ten languages with the most movies. From this plot we see that English has the largest number of films by far.

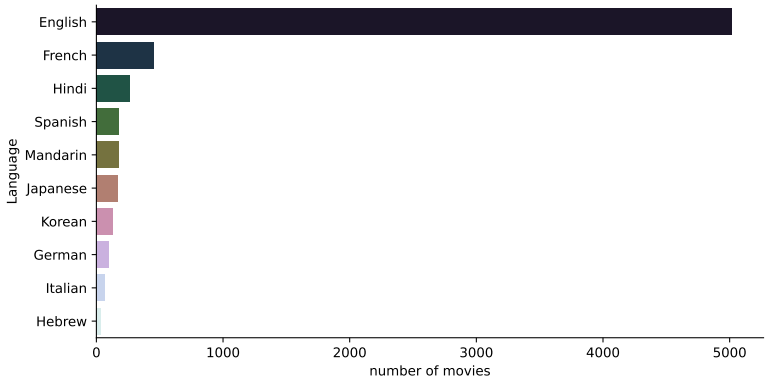

Figure 8: Number of movies in each language

Fig 9 shows the most frequent words in movie titles. The bigger the word is, the higher frequency that word appeared in the title of the movies.

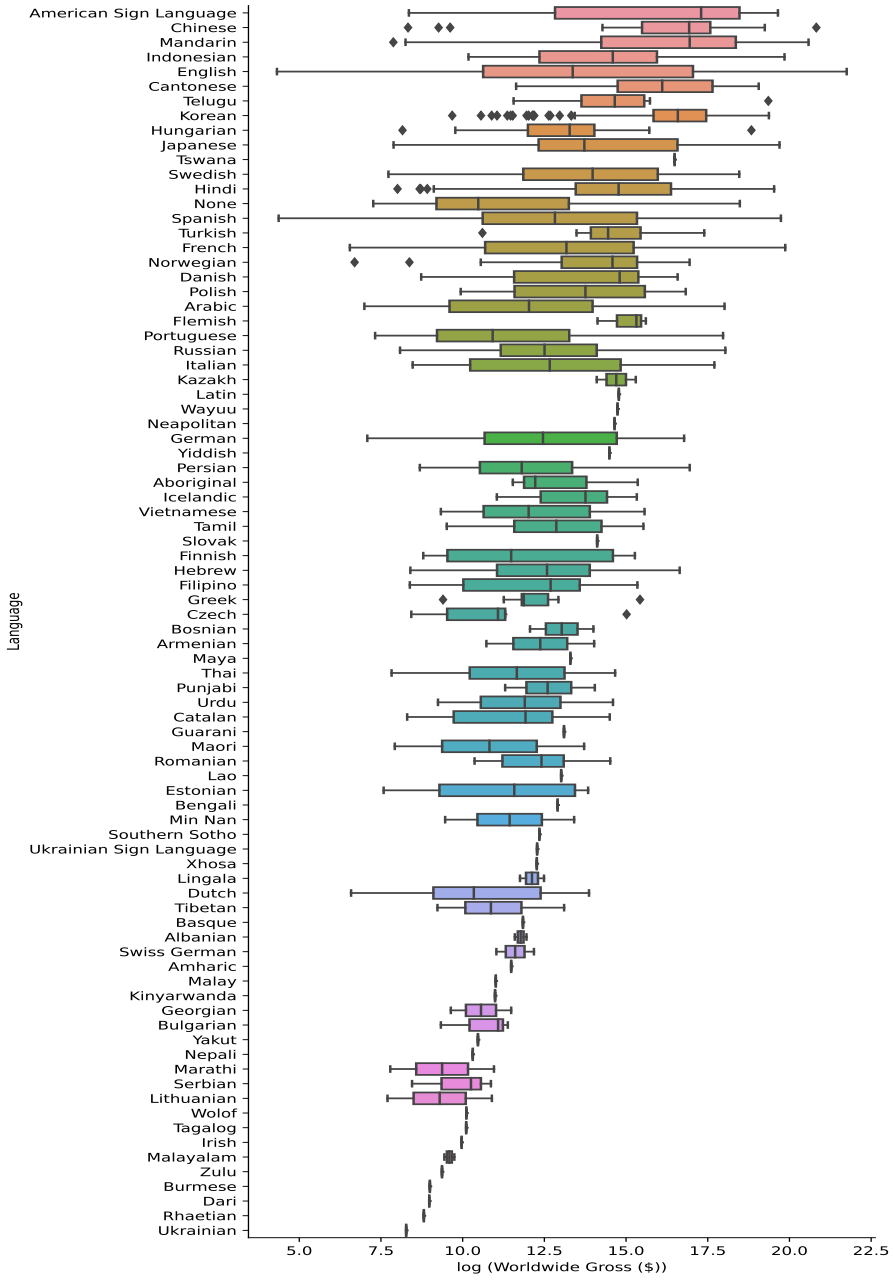

Figure 7: language vs worldwide gross

Figure 9: Word cloud of the most frequent words in title of the movies. To generate this world cloud we use python wordcloud library

Fig 10 and 11 show the histogram and worldwide gross versus the release month of the movies, respectively. Evidently, movies released in December had the highest sales on average, followed by July and June. From fig 10 we see that most movies were released in September.

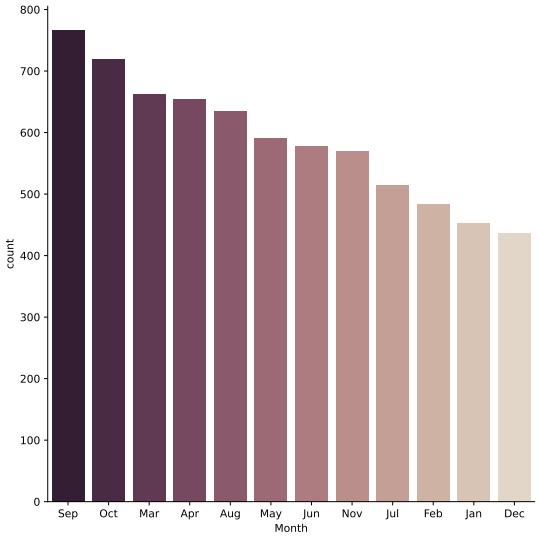

Figure 10: number of movies have been released each month

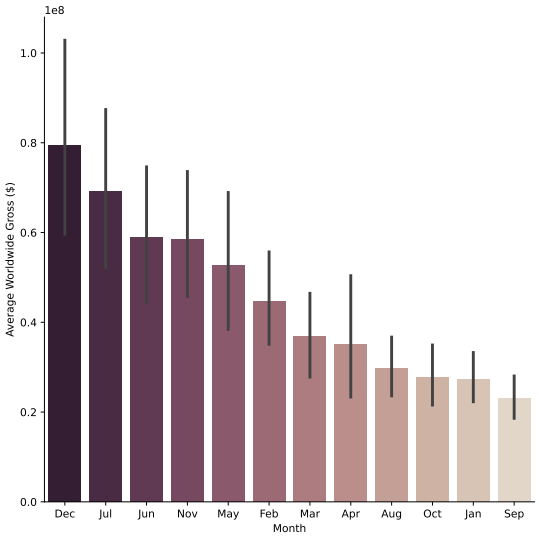

Figure 11: Worldwide gross vs month of release

Fig 12 shows the number of movies for each director. From this figure, we see that Robert Zemeckis and Hayao Miyazaki have the most films. It should be noted that from this figure, we can not necessarily conclude that which directors were busier between 2010 and 1019. Because some movies' initial release date was before this period, but they re-released during this 10 years.

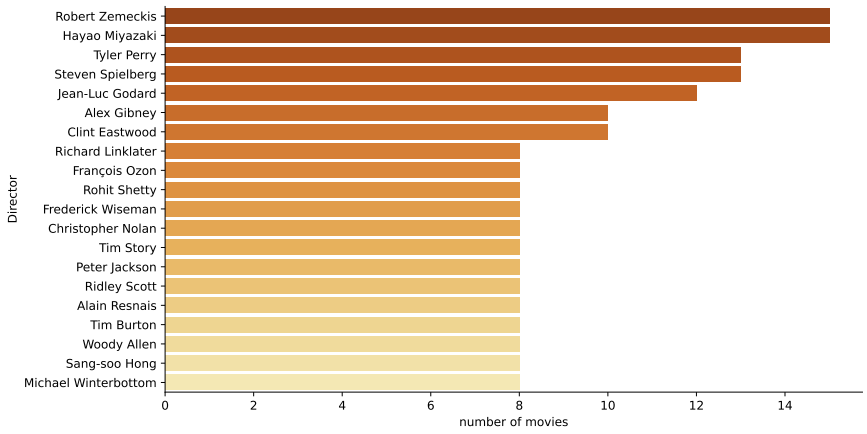

Figure 12: 20 directors with the most films between 2010 and 2019. Note that the horizontal axis shows number of movies from a director which showed in cinemas during this period although the first release date of a movie may differe. Thus we see that Jean-Luc Godard is in fifth place even though most of his movies did not release between 2010 and 2019.

Fig 13 shows the first 20 directors with the average highest IMDB score $\times$ Number of voters. It is important to consider the number of voters because some movies have high IMDB scores, but their voters are meager. From this figure, Frank Darabont, Director of acclaimed movies The Shawshank Redemption and The Green Mile, is in the first place followed by Christopher Nolan and Francis Ford Coppola.

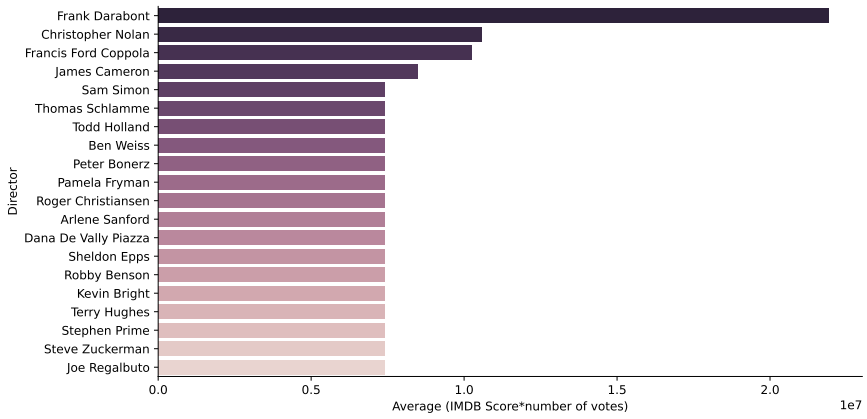

Figure 13: Twenty directors with the highest Average IMDB scores $\times$ IMDB votes

Fig 14 and 15 shows the directors with the most average worldwide gross and

worldwide gross. These two plots express different things. When we average the number of movies for a director, we measure the average performance on the director's box office. However, when we calculate the sum of the worldwide gross of movies from a director (fig 15), what is important for us is how much money that director has made besides their single movie performance. For instance, in fig 15 Micheal bay is in tenth place, but he is not among the first twenty directors in fig 14. In both of these figures Anthony Russo and Joe Russo, directors of Marvel Avengers movies, are in the first place.

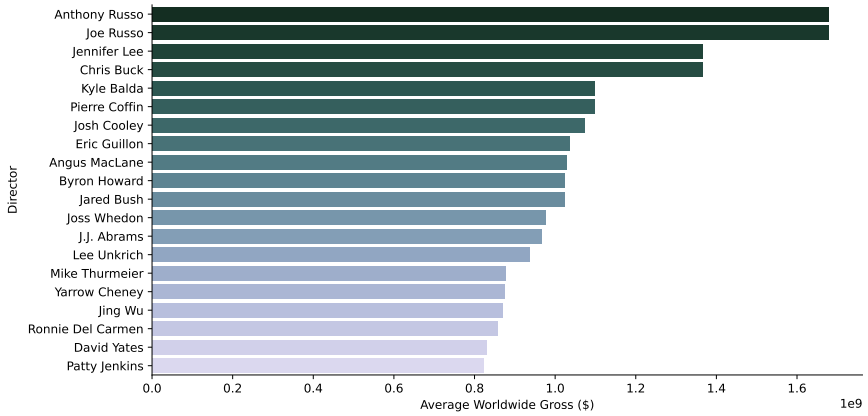

Figure 14: 20 directors with highest worldwide gross averaged over number of movies.

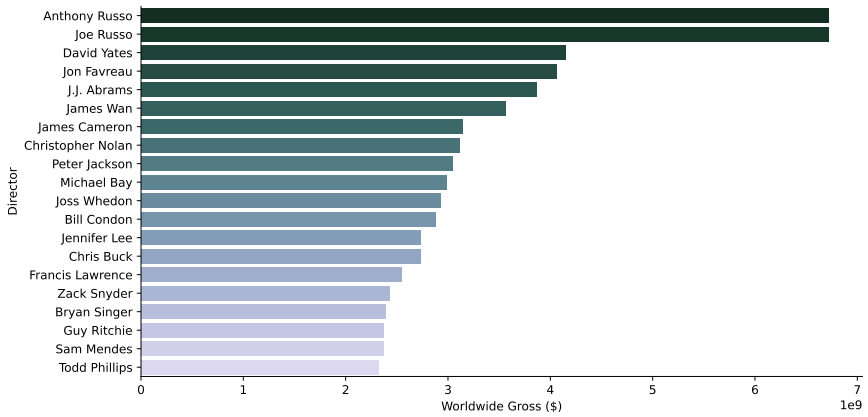

Figure 15: Twenty directors with highest worldwide gross. Note the difference between this plot and fig 14.

Fig 16 shows the average number of reviews that a movie (from each director)

got on IMDB. From this plot, we see that Nolan's movies have been talked about the most on average between 2010 and 2020.

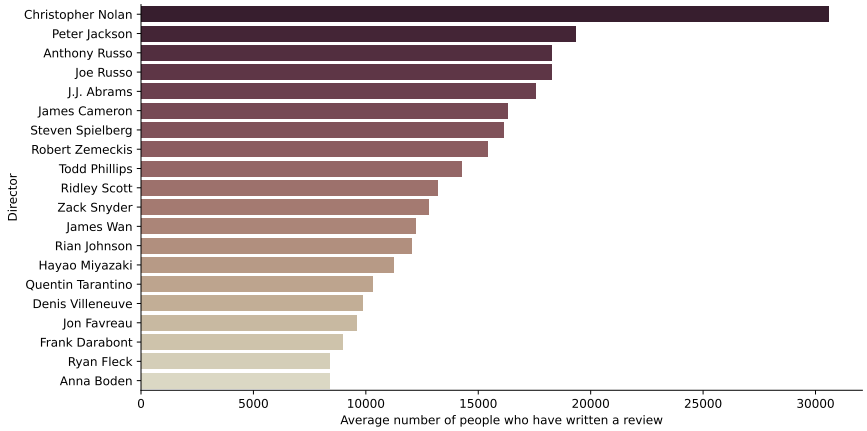

Figure 16: Average number of reviewers for each director

Distributors play a crucial role in the success of movies. In fig 17 you see the average worldwide gross for the twenty first distributors. As can be seen, The H Collective is in the first place, followed by Walt Disney and DreamWorks.

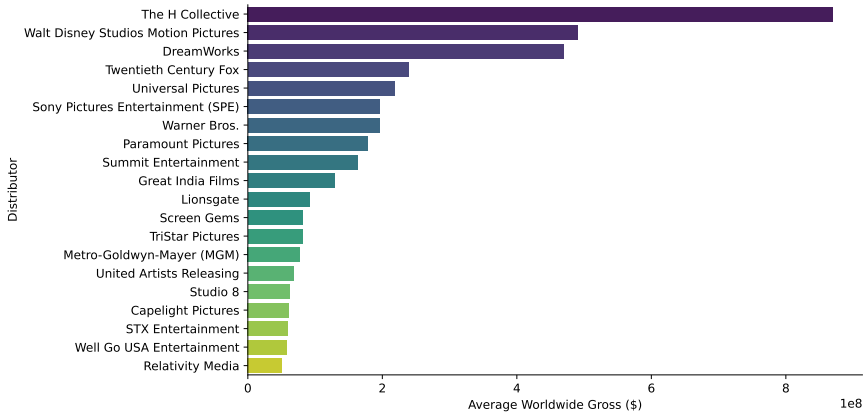

Figure 17: Twenty first distributors in terms of average WorldWide gross.

Fig 18 shows the average IMDB score $\times$ number of votes for each company. As shown, Universal Home Video is in the first place.

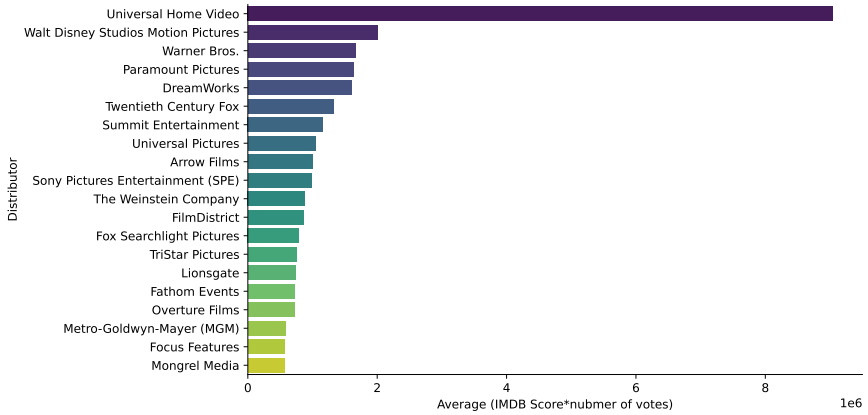

Figure 18: Twenty first distributors by average IMDB score $\times$ number of votes

Fig 19 shows the twenty first actors who had the most movies between 2010 and 2019. Liam Neeson, Robert De Niro, and Isabelle Huppert are in first to third place, respectively.

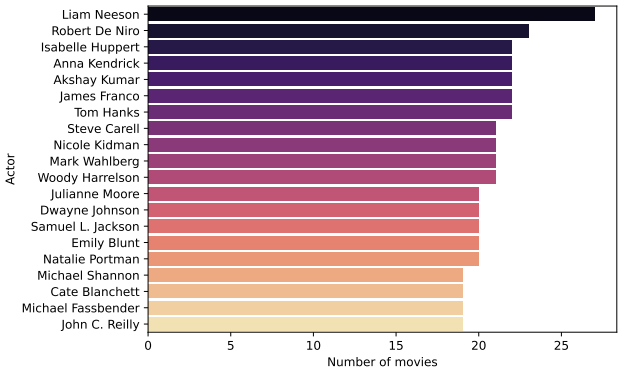

Figure 19: Actors number of movies

Fig 20 and 21 show the top twenty movies by the number of written reviews and worldwide gross. As it can be seen, Joker got the highest number of reviews, and Avengers: Endgame has the highest worldwide gross.

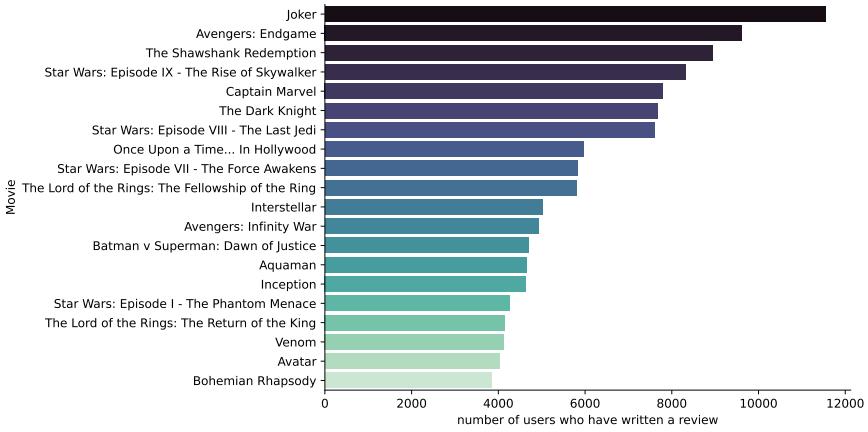

Figure 20: Movies's number of reviews

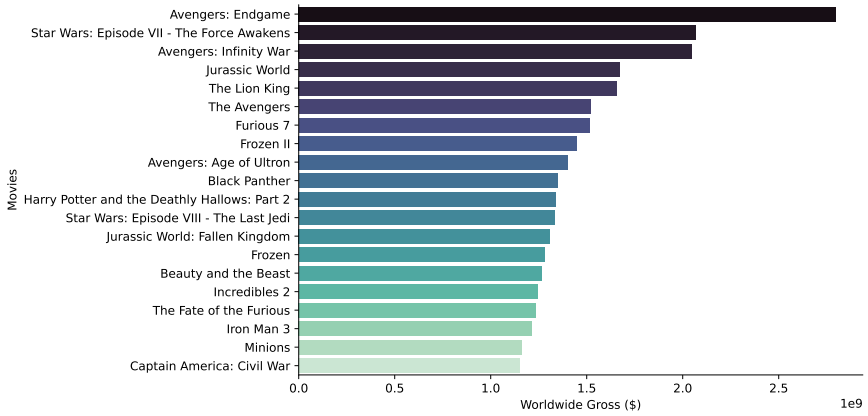

Figure 21: Movies' worldwide gross

7 supplementary information

- The project repository including all codes and figures can be found here.
- Yearly CSV files can be found here.
- Final data set can be found here.

References

[1] Syed Muhammad Raza Abidi, Yonglin Xu, Jianyue Ni, Xiangmeng Wang, and Wu Zhang. Popularity prediction of movies: from statistical modeling to machine learning techniques. *Multimedia Tools and Applications*, 79(47):35583–35617, Dec 2020.

[2] S. Asur and B. A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 492–499, 2010.

[3] A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar. Role of different factors in predicting movie success. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–4, 2015.

[4] Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PLoS ONE*, 8(8):e71226, August 2013.

[5] Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. Success in books: predicting book sales before publication. *EPJ Data Science*, 8(1):31, Oct 2019.