Analysis of Film Box Office Forecast and Word of Mouth

Kunyang Jiang
Chengdu Jiaxiang Foreign Language School, Chengdu 610000, China.

Abstract: This paper uses China's film box office and word of mouth data in the past 20 years, and uses Python for data analysis, To explore the influence of word of mouth on film box office, to analyze the relationship between film shooting budget, film score and box office,film production, film box office and film style with the change trend of time, find out the most popular film style, assist the film business company to make the film online plan, improve the revenue.

Keywords: Film Box Office; Film Industry; Box Office Data

Introduction

Since the 21st century, China's progress in the film industry is obvious to all. In terms of the number of box office figures, China's annual total box office has made a qualitative leap from 1 billion yuan initially to 60 billion yuan now today. China's domestic films occupy an important seat in the whole world film market.

Box office and word of mouth are two important indicators to evaluate the quality of films. In today's society where the Internet has become so developed, the reference value of word of mouth seems to become higher and higher for consumers. In terms of film, the score of a film is the concrete form of film reputation.But because of the water army deliberately bad comments or praise phenomenon. Scoring also loses its reference value to some extent. Compared with positive word of mouth, negative word of mouth has a more significant impact on the box office. Therefore, word of mouth is not the only factor affecting the box office.

Star actors are also an important factor affecting the film box office. Famous actors acting in the film will become the first choice of the audience. Similarly, well-known directors will also bring a certain box office revenue,the director is the core character of the film, and he has accumulated a lot of experience before making the film, and his ability size often determines the quality of a film.

The film genre will also reflect the box office situation of the film. If most women may be particularly afraid of watching horror movies, horror movies generally provide men at the box office. Comedy, science fiction-genre movies are relatively more suitable for all genders, but science fiction movies may not be suitable for older people,science fiction films may be offered by young or middle-aged people. So the movie genre also affects the box office revenue.

Release time also affects the box office. Whenever the Spring Festival, the National Day will appear "immortal fight" scene. In the context of the holidays, a lot of good movies will flock to them. And the release time is also closely related to the genre of films, like the New Year films can only be released during the Spring Festival.

The above discussion shows that: 1. Word of mouth and rating have an impact on the box office, but not significant.2. The star power of actors and directors will have an impact on the box office, but it is not the decisive factor. A dark horse shot by unknown directors big works also abound.3. Specific types of movies will broaden but will also limit the number of moviegoers, thus affecting the box office. This paper will discuss the problem of popular films in China from the perspective of film reputation, director and actor popularity, film type, film length and film budget.

1. Analyze the target

Film is a huge social and cultural cause, and the development of film cause has a positive significance to enhance China's cultural

confidence. China's film market also occupies an important seat in the world film. But different films have different box office.So, what are the factors affecting the box office and word of mouth? What is the reason for the huge box office difference?

2. Data collection； DC

This paper uses the data of box office, evaluation, budget and popularity of Chinese films in the past 20 years.

3. Data cleaning and processing

Repeat the value processing Duplicate value refers to the repeated values in the dataset, which we first need to locate and modify, otherwise it will cause a bias in the subsequent data analysis. The d r o p _ duplicates() function performs the deletion of the duplicate values.Use the ones in the dataframe. The duplicated method returns a boolean-type Series, showing whether there are duplicate rows, and False without duplicate rows, and with duplicate rows, showing True in the second duplicate value.

Missing value processing The presence of missing values can affect the grasp of the data rules, so the missing values need to be filled in or deleted. Missing values in the dataset require the data cleaning with Numpy and Pandas's tools. We can count the number of missing values, and then process them. By processing the missing values, it is possible for us to further analyze the data. The following figure (Figure 1) is the missing part of the original dataset.

```
budget                    0
genres                    0
homepage               2608
keywords                  0
original_language         0
original_title            0
overview                  2
popularity                0
production_companies      0
production_countries      0
release_date              0
revenue                   0
runtime                   2
status                    0
tagline                 502
title_x                   0
vote_average              0
vote_count                0
movie_id                  0
title_y                   0
actors                    0
director                  1
writer                 2950
producer               4172
dtype: int64
```

Figure 1. Missing values for the data set

Outliers Handling Outliers refer to individual data in a data set that obviously deviate from other samples or violate common sense. Generally, we will modify or delete outliers if there are any outliers.We can use the box diagram to view the outliers, as shown in the following figure (Figure 2), to clean and screen the dataset more perfectly.

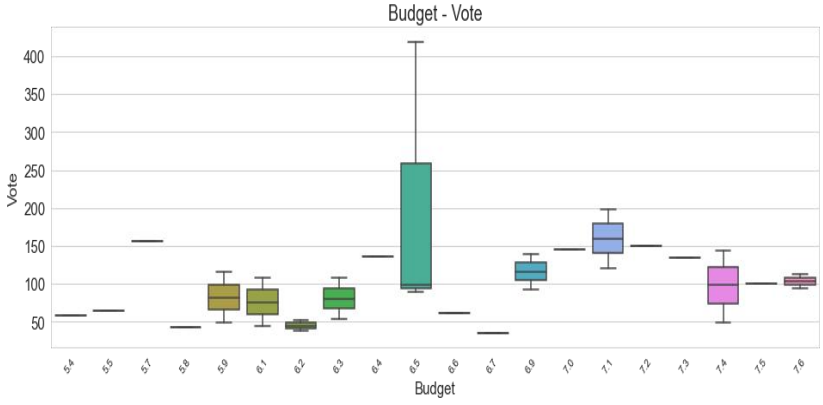

Figure 2. Box plot of the data to view the outliers

Through the analysis of outliers, we can see that some films' box office revenue is much higher than other films, but combined

with the film itself and found that the box office for these films is really high, so we still keep these outliers.

Test data By checking the data (Figure 3), the average budget is 3 million, up to 40 million; the average box office revenue is 90 million, 200 million, and the median is 3 million, which makes some films raise the average box office. So what does the budget have to do with the box office? We need to do our research with data analysis tools.

	budget	popularity	revenue	vote_average	vote_count	movie_id
count	4.172000e+03	4172.000000	4.172000e+03	4172.000000	4172.000000	4172.000000
mean	3.266745e+07	24.260428	9.385619e+07	6.224976	787.653883	50407.210211
std	4.220379e+07	33.213366	1.714749e+08	0.969845	1296.127823	82452.706664
min	0.000000e+00	0.001586	0.000000e+00	0.000000	0.000000	5.000000
25%	3.000000e+06	6.958402	4.233102e+05	5.700000	95.750000	8054.750000
50%	1.800000e+07	15.811272	3.007637e+07	6.300000	314.500000	12179.500000
75%	4.500000e+07	31.402622	1.085573e+08	6.900000	874.000000	49014.000000
max	3.800000e+08	875.581305	2.787965e+09	10.000000	13752.000000	459488.000000

Figure 3. Dataset

By ranking the ratings of the data set and the number of ratings, you can see that the top three movies are all about Inception, Batman: The Dark Knight, and Avatar.

title_x	vote_average	vote_count
Inception	8.1	13752
The Dark Knight	8.2	12002
Avatar	7.2	11800

Figure 4. Top three rated films

3.1 Data screening

According to this dataset, only factors such as "budget, genres, popularity, production_companies, production_countries," can have an effect on the film box office, while data such as "homepage" will have no effect on the box office. Here is a visual analysis of these factors.

4. Data visualization

After processing the repeated values, missing values and outliers in the data set, the data can be visualized and analyzed, and the following conclusions can be obtained:

4.1 The relationship between the budget and the box office

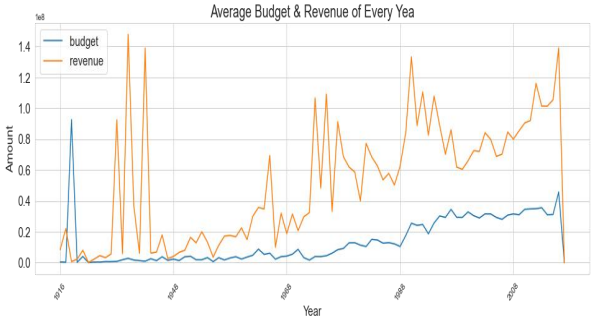

Figure 5. Relationship between budget and box office in different years (line chart)

The picture reflects a significant increase in film industry revenue during this period in 1990. And you can see that the revenue of the film industry also fluctuates periodically.But the steep increase around 1940 was puzzling, so after looking for data and searching for online sources, Pinocchio, Snow White and the Seven Dwarfs, and Gone with the Wind shone brilliantly in that era. Even a hundred years later, they are still the brightest stars in film history.

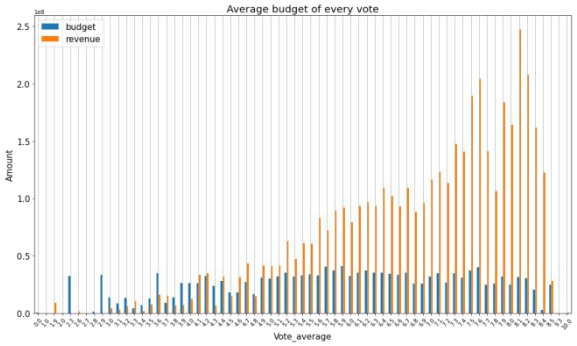

Figure 6. Relations between budget and box office (histogram)

Generally speaking, the general trend is that the higher the score, the higher the box office, and some film budget is not completely positively correlated to the box office. So these films may stand out from the extreme because of the perfection of the director or the skill of the actors.

4.2 The relationship between the film genre and the box office.

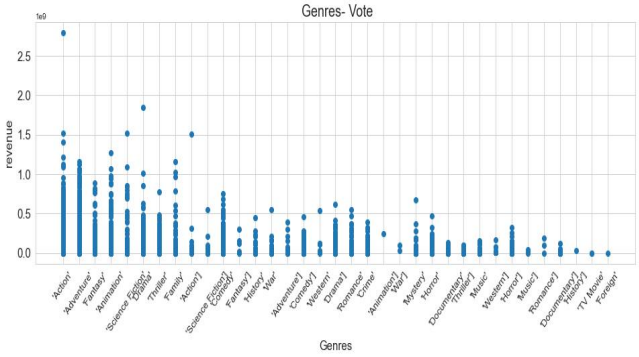

Figure 7. The Relationship between the film genre and the box office

As you can see the action, the top of all genres. It means that this kind of film is more popular with audiences.

4.3 The relationship between the actors and the box office

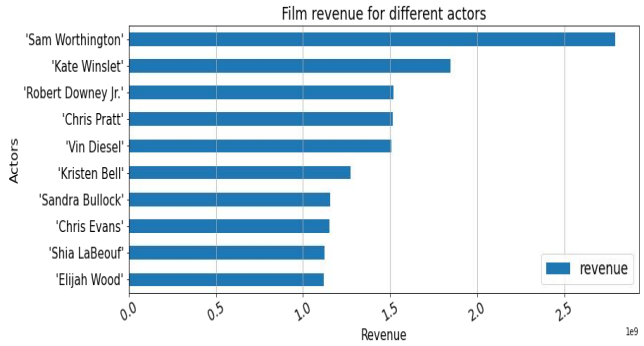

Figure 8. The relationship between the actors and the box office

In this picture, Sam Worthington (Avatar), Sam Worthington (Guardians of the Galaxy, Jurassic World), Robert Downey Jr (Iron Man and Avengers) and Kate Winslet (Titanic) had the highest-grossing actors.

4.4 The relationship between the director and the box office

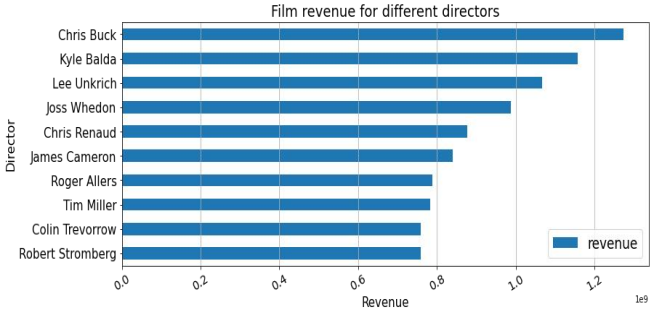

Figure 9. The relationship between the director and the box office

Therefore, directors Chris Buck (Frozen 1,2), Kyle Balda (Minions, Despicable Me) and Lee Unkrich (Coco and Toy Story) are high at the box office.

Conclusion

Film factors affecting the film box office: film type, production budget, duration, production company or country; director and actor factors: the director award and popularity, the popularity of the actors.

All of these factors will affect the box office. Therefore, for the investors and producers of films, it is best to focus on the type of films on action films, and invest more in remakes and sequels, while for directors and actors, try to choose the winning director and well-known actors. Such a movie is a movie that can eventually get a high revenue box office.

References

[1] Zhou RX. Value Evaluation of Domestic Film Copyright Based on BP Neural Network [D]. Chongqing University of Technology, 2022.

[2] Diao JY. How to improve the quality of the film industry and promote cultural consumption [N]. China City News, 2022-02-14 (A14).

[3] Xie YN, Yang CM. Film box Office data collection and visualization study [J]. Information and Computer (theoretical edition), 2021,33 (23): 176-178.0.