Wikipedia Data Analysis

Ⅰ Introduction and Literature Review

The film industry is a vast market worth tens of billions, attracting significant attention from both aesthetic and commercial perspectives. Despite this, there's limited understanding of how online content, typically online encyclopedia, shapes audience decision-making. Box office performance plays a crucial role in the industry, with audience choices now heavily influenced by online platforms. As such, we propose a hypothesis: the content within Wikipedia entries about a movie has a substantial impact on its box office success.

To delve deeper into this topic, we searched for relevant literature. In one paper, researchers noted a linear correlation between the views of film entries on Wikipedia and their box office performance[1]. Building upon this article, other researchers offered a preliminary explanation for this phenomenon: many people decide what movie to watch by consulting Wikipedia, making Wikipedia a reflection of a movie's search popularity[2]. In our paper, we take a different approach. We believe that Wikipedia is not just a counter of search numbers but also a crucial means for audiences to choose movies. We will focus investigating on the quality of Wikipedia content of the film and its impact on box office performance. This will help us examine the influence of Wikipedia, a platform often considered educational and informative, on the reputation of films.

Ⅱ Methodology

To evaluate the quality of the content of a specific film, we make the assumption that on average, a Wikipedia entry about a film with a higher number of edits indicates higher quality. We can substantiate this assumption with the following points:

1: All Wikipedia entries related to films can be created by any audience member. Therefore, the textual quality of these film entries is entirely random when they are first created.

2: According to IBM's report, Wikipedia exhibits a rapid removal rate for destructive behavior. Consequently, the majority of retained edits contribute positively to the textual quality.[3]

Based on these points, it can be argued that as the number of edits on a film's Wikipedia entry increases, the quality of the entry is more likely to be higher.

Therefore, to study the relationship between a film's box office performance and the quality of its Wikipedia entry, we only need to examine the editing activity on the Wikipedia entry during the film's release period along with its total box office earnings. We chose the top 40 films in terms of box office earnings for the years 2022 and 2023 as the data reference. We retrieved the total number of edits to these films' entries on Wikipedia within one month of their release using the Wikipedia API and conducted regression analysis with their box office earnings.

Ⅲ Result Analysis

As depicted in the scatter plot below, each point represents a film. The x-axis

represents the number of edits the film's Wikipedia entry received within one month of its release, while the y-axis represents the total box office earnings of the film, measured in US dollars.

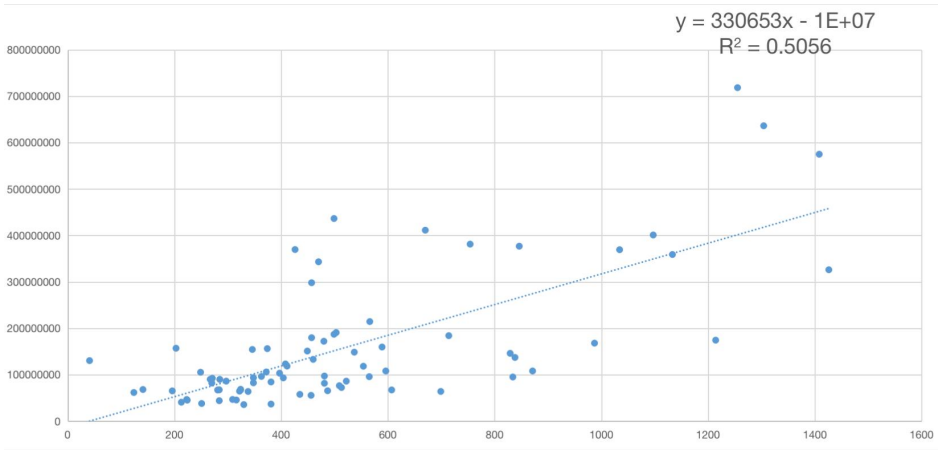

Figure 1

After a simple linear analysis, it was found that the number of edits a film receives is positively correlated with its box office earnings. On average, for each additional edit, the total box office earnings increase by $330,653. Due to the limited sample size, the R-squared value for this linear regression is 0.51. With a larger sample size, more precise values can be obtained.

However, this linear regression only reflects the relationship between the number of edits and box office earnings but does not establish causality. It remains unclear whether the increase in edits leads to increased audience willingness to watch the film, thereby boosting box office earnings, or if the rising box office earnings generate increased interest among audiences, leading to more edits. To understand if there is a causal relationship between these two factors, we have selected three typical examples for detailed analysis.

Typical Data Analysis:

We have chosen three typical examples:

Don't Worry Darling; 224 edits; total box office $45,309,403

Uncharted; 537 edits; total box office $1,486,449,929

Guardians of the Galaxy Vol. 3; 1133 edits; total box office $358,995,815

These three data points have low intercepts from the regression line, representing low, medium, and high values, respectively. Therefore, they are suitable as typical examples for reference. The following three graphs respectively show the daily number of article edits and daily box office revenue for these three movies within one month of release. The red line represents article edits, and the blue line represents box office revenue.

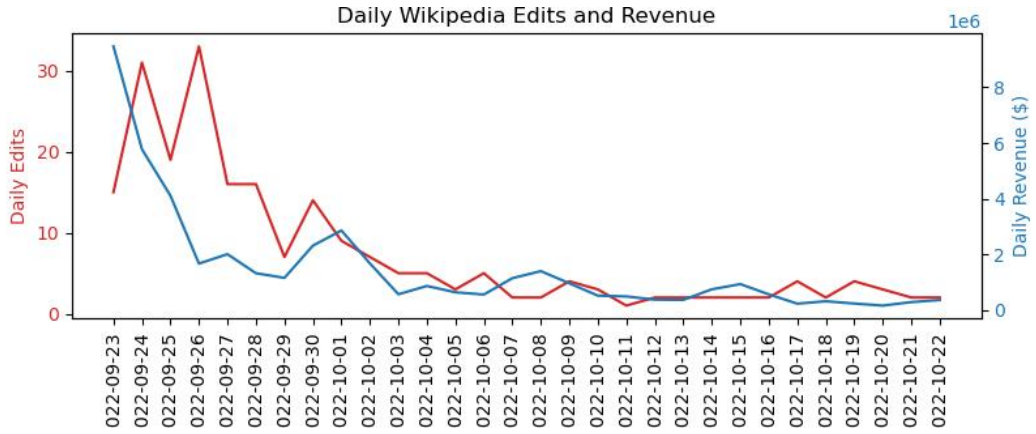

Figure 2: Don't Worry Darling

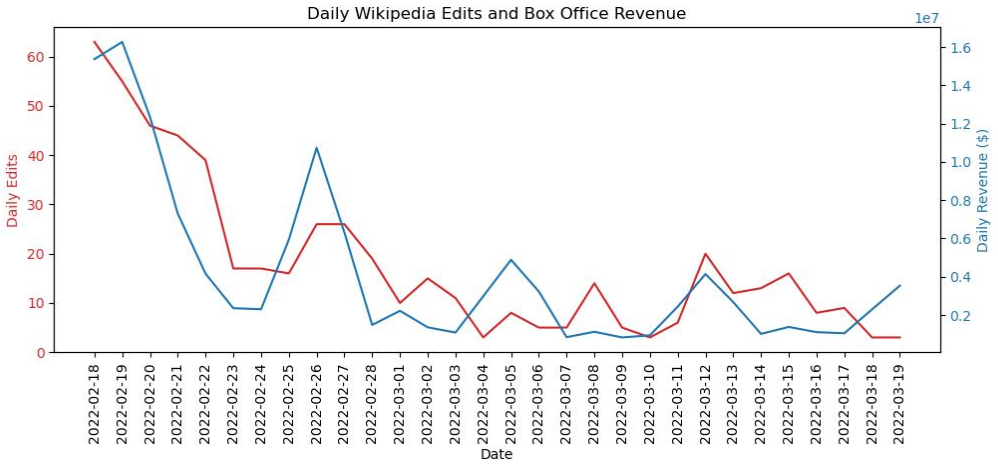

Figure 3: Uncharted

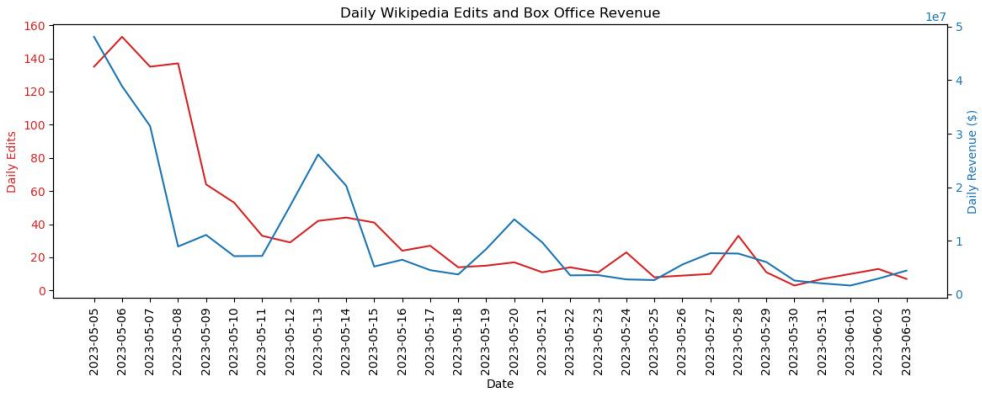

Figure 4: Guardians of the Galaxy Vol. 3

There is a strong correlation between the daily article edits and box office revenue even for different dates of the same movie. This is understandable as more people finish watching the movie in a day, leading to more edits. Interestingly, the magnitude of changes in article edits also indicates changes in box office revenue. Taking figures 2 and 3 as examples, the movie represented in Figure 2 no longer has significant changes in the Wikipedia article edits after a week of release, thus it led to the low box office sales in the following weeks. In contrast, in Figure 3, despite being released for two weeks, after the weekend of 03-05 screening, there is still active editing on 03-08, which also hints at a surge in box office sales a week later.

IV Conclusion

In terms of results, the conclusion of the study aligns with our hypothesis: higher article edits imply better text quality, which in turn implies higher box office revenue. However, the underlying principles may not necessarily be as we assumed. The rise and fall of a movie's reputation have many aspects, and Wiki is just one part of many, not a decisive factor. But due to Wiki's openness, inclusiveness, and transparent data, it serves as a microcosm for studying various online platforms. If more people are willing to share information about the movie on Wiki, then naturally more people are willing to share about the movie on platforms like TikTok, Instagram, increasing the movie's reputation and box office revenue.

In conclusion, although there is no evidence that improving the text quality of a movie's Wiki page can increase its box office revenue, the amount of Wiki article edits can serve as a good indicator, showing in real-time the audience's reputation and engagement with the movie. This may be useful for future movie market analysis.

[1] Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. PloS one, 8(8), e71226.

[2] de Silva, B., & Compton, R. (2014). Prediction of foreign box office revenues based on wikipedia page activity. arXiv preprint arXiv:1405.5924.

[3] https://en.wikipedia.org/wiki/Reliability_of_Wikipedia