Project Proposal

Project Name:

Investigating the Relationship Between Movie Box Office Opening Revenue and Ratings on IMDb and Rotten Tomatoes.

Author: 必填!!

Motivation：

The motivation behind this project is to explore the relationship between a movie's box office opening revenue and its ratings on IMDb and Rotten Tomatoes, aiming to understand the relationship between commercial success and critical acclaim in the film industry. By analyzing correlations and discrepancies between financial performance and ratings, this study seeks to uncover insights into audience preferences, marketing strategies, and the dynamics of film reception across different platforms.

Data Source：

1. Movie Box Office Data

URL: https://www.boxofficemojo.com/year/2023/
Instructions: Movie Box Office provides revenue data of each movie. Utilize pandas.read_html to scrape 2023 movies data from Movie Box Office website. Then use BeautifulSoup and Request to request every movie's html to scrape the title, unique identity code and opening revenue of the movie. There are some advanced anti-scraping techniques on the website so we may need to use Selenium to get data (Domestic Box Office For 2023, 2023).

2. IMDb Rating Data

URL: https://developer.imdb.com/non-commercial-datasets/
Instructions: IMDb provides a series of opensource datasets including basic information, rating data, crew data, etc. The datasets can be downloaded from the URL (IMDb Non-Commercial Datasets, n.d.).

3. Rotten Tomatoes Rating Data

URL: https://www.rottentomatoes.com/
Instructions: Rotten Tomatoes have stopped issuing API keys for developers. So, we need to use Request and BeautifulSoup to scrape rating data. we will construct URLs based on titles of movies in 2023 to scrape rating data on Rotten Tomatoes. There are some advanced anti-scraping techniques such as Shadow DOM so we may need to use Selenium to get data (Rotten Tomatoes, n.d.).

For each data source, I collected 200 samples based on the number of 2023 movies data from Movie Box Office website.

During the data collection process, two main challenges were encountered. Firstly, establishing a reliable linkage between movie box office data and rating data posed difficulties. It cost some time to find out that Box Office data provided a unique identifier (UID) that could be used to match with IMDb Rating Data. But no such UID was available for the Rotten Tomatoes Rating Data, requiring reliance on movie titles which risked matching issues due to identically named movies. Secondly, when scraping the Rotten Tomatoes Rating Data, server restrictions resulted in a 403 error code. To overcome this, I employed tools to generate various user agents to update request headers, eventually successfully retrieving the data.

Integrated Data Model:

There are three tables in the model, they can be integrated by ‘Release’ or ‘uid’, where ‘Release’ is movie title and ‘uid’ is unique identifier value.

a. movies_2023_boxoffice

	Release	Gross	Theaters	Release Date	Info_url	uid	openingGross	openingTheaters
0	Barbie	$636,225,983	4,337	Jul 21	/release/rl1077904129/?ref_=bo_yld_table_1	tt1517268	$162,022,044	4,243
1	The Super Mario Bros. Movie	$574,934,330	4,371	Apr 5	/release/rl1930593025/?ref_=bo_yld_table_2	tt6718170	$146,361,865	4,343
2	Spider-Man: Across the Spider-Verse	$381,311,319	4,332	Jun 2	/release/rl2812183041/?ref_=bo_yld_table_3	tt9362722	$120,663,589	4,313
3	Guardians of the Galaxy Vol. 3	$358,995,815	4,450	May 5	/release/rl2977202945/?ref_=bo_yld_table_4	tt6791350	$118,414,021	4,450
4	Oppenheimer	$326,101,370	3,761	Jul 21	/release/rl3725886209/?ref_=bo_yld_table_5	tt15398776	$82,455,420	3,610

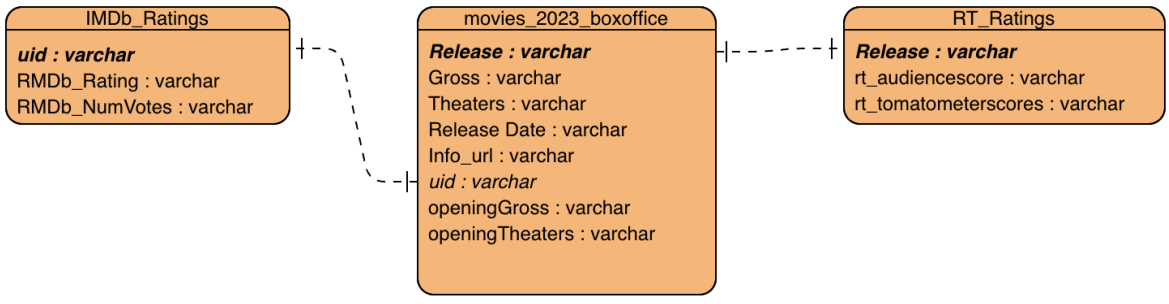
b. IMDb_Ratings

	uid	IMDb_Rating	IMDb_numVotes
0	tt0000001	5.7	2036
1	tt0000002	5.7	272
2	tt0000003	6.5	1986
3	tt0000004	5.4	178
4	tt0000005	6.2	2746

c. RT_Ratings

	Release	rt_audiencescore	rt_tomatometerscores
195	Broker	72	94
196	A Good Person	96	58
197	Maybe I Do	61	33
198	Fear	42	NaN
199	The Lost King	92	77

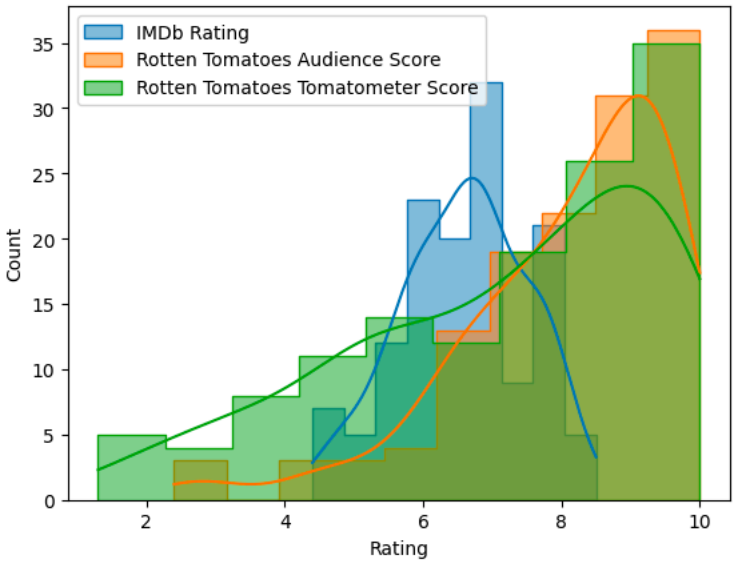
d. Diagram

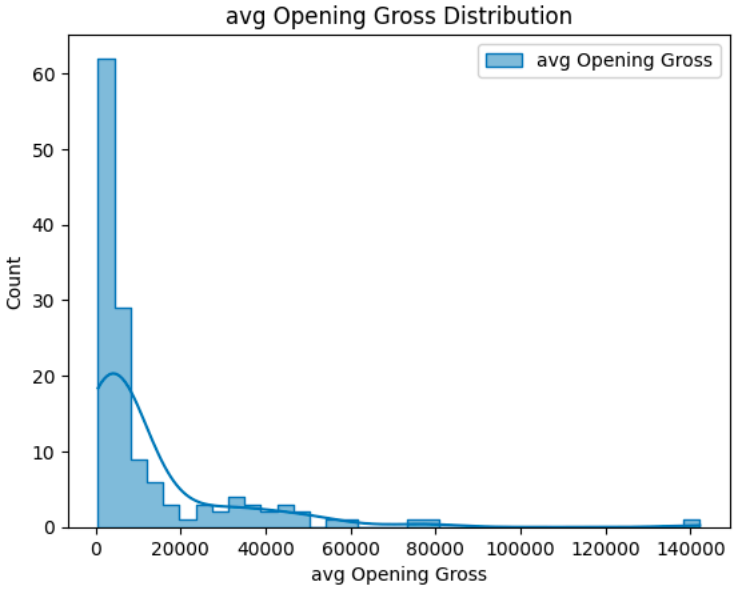


Analyses/Visualizations:

First of all, I used histogram to analyze the distribution of ratings and gross revenue. To make the ratings in same scale, I normalized them between 1 to 10. To mitigate the impact of the number of theaters on gross revenue, I utilized average gross data, calculated as gross revenue divided by the number of theaters. This approach was applied to both overall gross and opening gross to normalize the revenue figures.

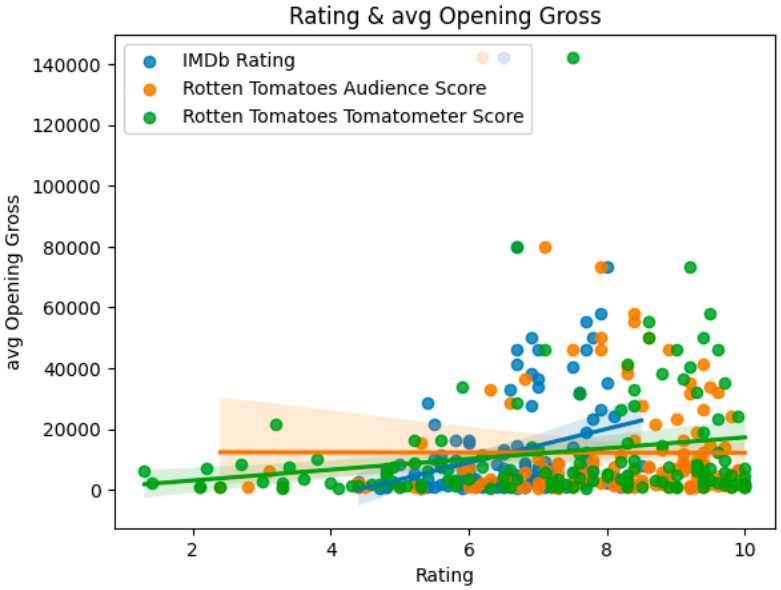
It appears that IMDb ratings range from 4 to 9, and Rotten Tomatoes Audience and Tomatometer scores range from 1 to 10 from the rating histogram. And it appears the average opening gross has a long tail on the right.



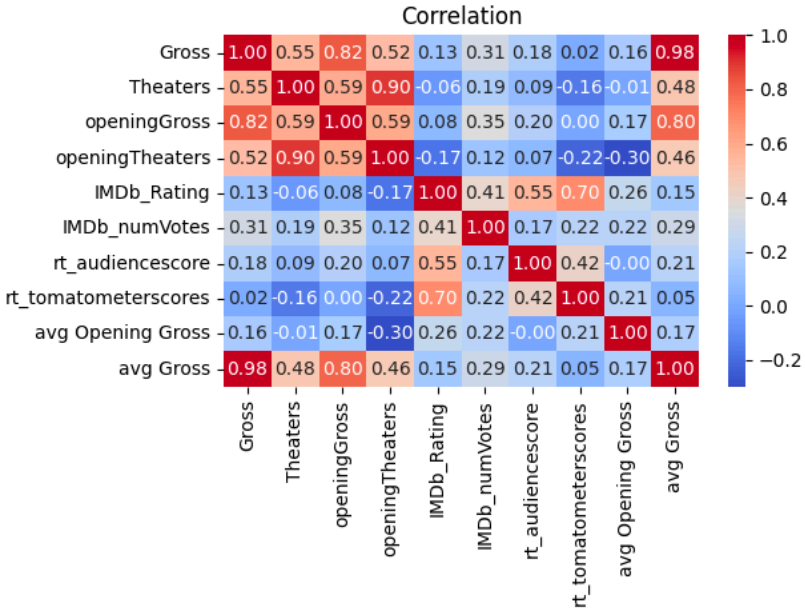


Then, I use scatter figure to analyze the relationship between ratings and gross revenue. To make it visually significant, I use linear regression to plot the trendline.

It appears that the opening gross is positively correlated to ratings of both IMDb and Rotten Tomatoes Tomatometer. The higher the rating, the higher the opening gross revenue. But in terms of Rotten Tomatoes audience scores, it appears to be not correlated to the opening gross revenue.



Finally, I utilize the correlation heat map to visualize the correlation between all factors. We can confirm that the opening gross is positively correlated to ratings of both IMDb and Rotten Tomatoes Tomatometer. And the Rotten Tomatoes audience scores are not correlated to the opening gross.



There are a lot of analyses and visualizations done during this process, more details can be found in /results/analyze_visualize.ipynb.

Conclusions:

There are some relationships between a movie's box office opening revenue and its ratings on IMDb and Rotten Tomatoes. The opening gross is positively correlated to ratings of both IMDb and Rotten Tomatoes Tomatometer. The higher the rating, the higher the opening gross revenue. But the Rotten Tomatoes audience scores is not correlated to the opening gross revenue.

We can suggest that theaters prioritize the IMDb and Rotten Tomatoes Tomatometer ratings when scheduling showtimes for new movies, rather than relying heavily on the Rotten Tomatoes audience ratings. This approach could potentially offer a more objective and critical assessment of the movie's quality, helping theaters make more informed decisions that align with audience preferences and expectations.

Future Works:

Firstly, I would conduct a more in-depth analysis of the factors influencing the discrepancies between box office revenue and ratings. Explore additional variables such as genre, cast, director, and promotional activities to identify underlying patterns and insights.

Furthermore, I would like to Implement machine learning algorithms to develop predictive models that can forecast a movie's box office performance based on its ratings, genre, promotional efforts, and other relevant features.