Film Earnings vs Ratings

Name: Gareth McKane

GitHub URL: https://github.com/UCDPAGarethMcKane/UCDPA_GarethMcKane/upload/main

Abstract

The global film industry was valued at $90 billion in 2021 (www.grandviewresearch.com). The highest grossing films of all time have been released in the last 25 years. It would be logical to assume that where these films placed in the earnings rankings, they would have the same positions in the critics' rankings.

This analysis aims to investigate comparisons between 20 of the top grossing films of all time and their critically acclaimed ratings and audience ratings. Charts in the form of scatterplot will be used to demonstrate this.

With the data that is available, it will also investigate the comparative differences in reviews of these films and illustrate this using a rating pyramid chart.

5 insights will be deduced from the analysis.

Introduction

The vast majority of people are aware of the global film industry. It originally started in the early 20th century and has gradually snowballed into the size it is today. However, the phenomenon of blockbuster films with billion dollar ticket sales is relatively new – the first being Titanic in 1998. Since then there have been over 50. Hence, the growth of these extremely high grossing films has occurred within my own lifetime and I have viewed the majority of them in a cinema.

"The nature of film criticism is to enlighten and enrich one's experience with the art of film, not to interpret film for them." (www.universityobserver.ie). Despite this, it is difficult to argue against critic reviews not having an impact on a films earnings potential. However, ultimately the earnings for a film come from the numbers of sales regardless of reviews. Higher ticket sales should go hand in hand with high audience reviews.

As I have personally contributed to these films by 'paying for a ticket', regardless of critical reviews, I decided to carry out an analysis as to whether audience reviews actually do outweigh critical reviews with regards to earnings.

Datasets

There were 2 datasets used in the production of this analysis.

- Dataset 1 – List of Highest Grossing Films
  - Source: Wikipedia - https://en.wikipedia.org/wiki/List_of_highest-grossing_films
  - An up to date referenced dataset with all relevant data for the analysis.
- Dataset 2 - rotten_tomatoes_movies.csv
  - Source: Kaggle - https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset
  - The data in this dataset was last updated 2 years ago.
  - Contains 21 columns of data.
- Both datasets contained several similar columns of data.

Implementation Process

The initial steps were to import the modules that would be required for the analysis. These were Pandas, Numpy, BeautifulSoup, Seaborn and Matplotlib.

It was my intention to use 2 sources of data that could be imported in different methods. Prior to settling on this analysis, I investigated several other options that required the web scraping of multiple tables from various Wikipedia pages. After repeating the web scraping process several times, I defined the function 'wikipage' that enabled me to quickly enter a Wikipedia URL and obtain the relevant table from the page in question. I used this function code in this analysis.

The Wikipedia page that I web scraped from was titled 'List of highest grossing films'. This was then converted into a dataframe from which I removed several columns using the '.drop()' function and renamed columns. Several of the films in the dataframe had letters placed in front of the $values. I was aware that I would need to convert these values into a float value to use later on. However, I was not able to produce code that enabled this - hence I removed these rows unfortunately. However, after a dtypes() check, the World Gross column was still an object and not float values. Hence, I purposely sourced code that enabled the separation of the $ symbol from the values. Subsequently, I reduced the values by 1,000,000 to produce easier values to work with.

The Rotten tomatoes dataset was obtained from Kaggle.com. This was imported using the '.read_csv()' function. Once again, I removed the unwanted columns from this dataset and renamed relevant columns. During a scan of the csv dataset, I noticed that there were multiple films with the same name from different years i.e. remakes. Hence, it would be necessary to merge the 2 dataframes using 2 subset columns – 'Title' and 'Year'. However, the Year column values in the Rotten Tomatoes dataframe was not in the same format as those in the Films dataset. This required code inputted that created a new column 'Year', and extracted the year only from the 'original release date' column. The 'original release date' column was deleted after this.

The 2 dataframes, 'Films' and 'Red_tomatoes' were then merged using the 'Title' and 'Year' columns from each of them. Unfortunately, this revealed that several of the rows contained NaN values. On checking the original csv dataset of 17,712 rows, these entries did not have any data other than the basic title and release date. Hence, I removed all rows that contained NaN values and reduced the dataframe 16,546 rows. This resulted in a merged dataframe with 37 rows. I reduced this down to the top 20 by removing rows in the range between 21 – 37. Hence, I was left with my final dataframe – 'movies_tomatoes'.

Results

Chart 1: Pairplot of all data columns

This pairplot essentially shows all scatter plot combinations. This is useful to get a general visualisation of all the data comparisons that can be configured.

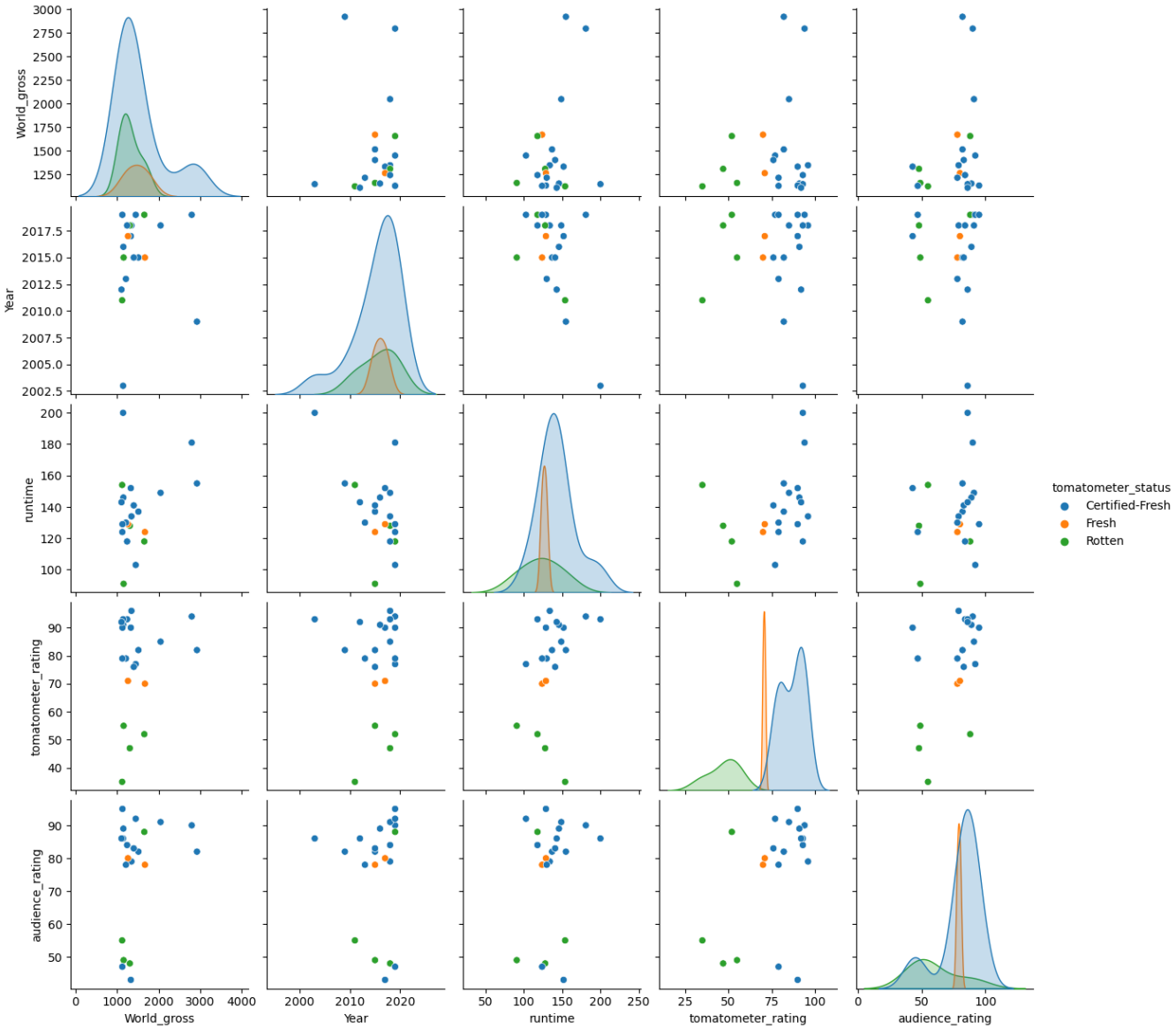


Chart 2 - Gross Earnings vs Tomatometer Ratings

"The Tomatometer score represents the percentage of professional critic reviews that are positive for a given film or television show. A Tomatometer score is calculated for a movie or TV show after it receives at least five reviews." (www.rottentomatoes.com)

The results here are not as spread out in the y-axis as in Chart 3. There are also only 4 films in the lower range of the ratings in comparison to Chart 3s' 5. There is also a cluster of 7 films of similar earnings in the upper left with ratings of over 90. Only the second highest earning film is above 90. This indicates that the highest critics' ratings do not necessarily transfer over to the highest grossing films.

In first viewing, there could be a strong correlation made for the Tomatometer Status as these are clearly grouped together.

However, rottentomatoes.com states that 'If the positive reviews make up 60% or more, the film is considered "fresh". If the positive reviews are less than 60%, the film is considered "rotten".' Hence, the chart simply highlights that this statement is true.

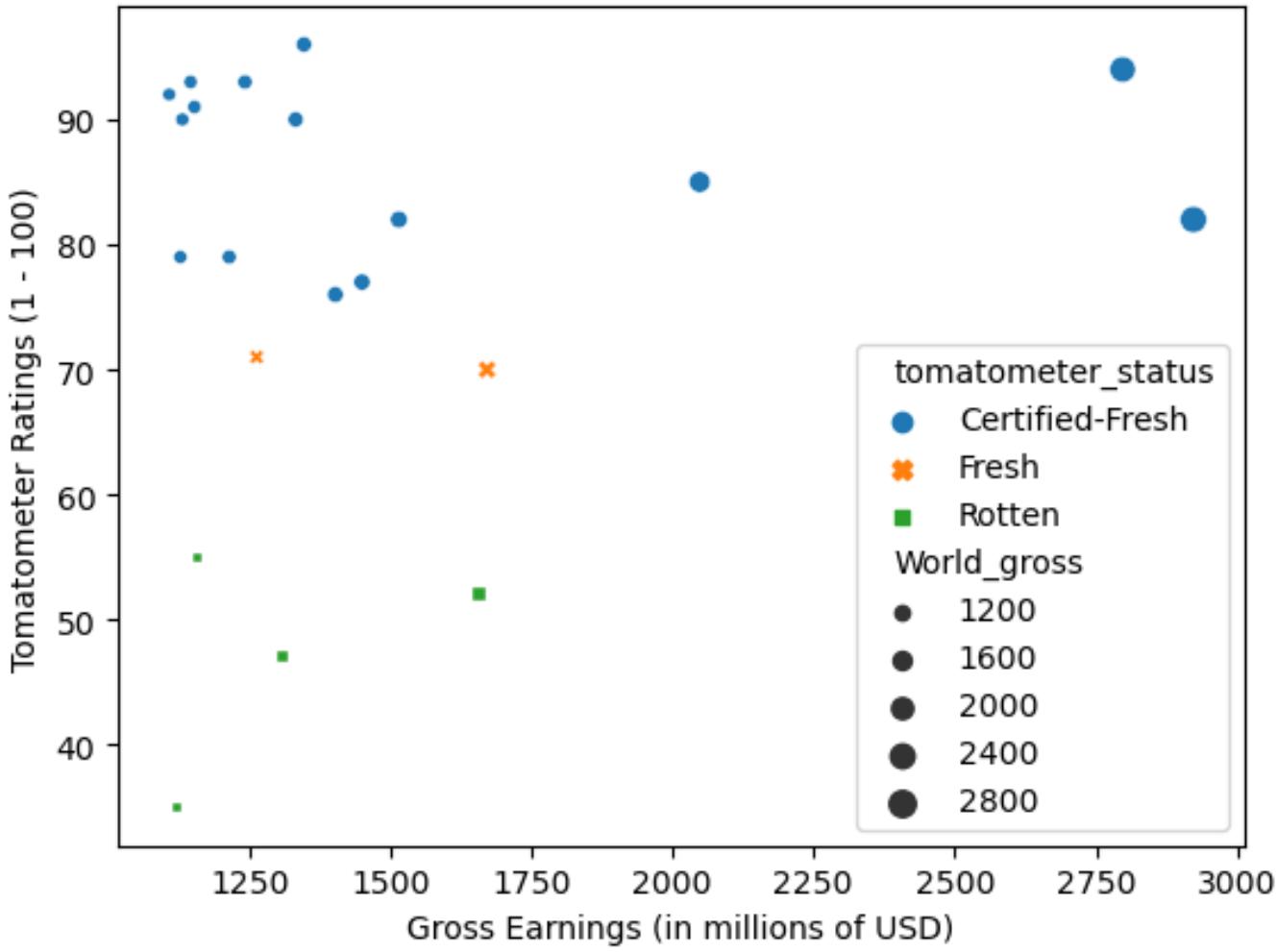


Chart 3: Gross Earnings vs Audience Ratings

There is a large gap in the audience ratings between the 5 lowest earning films and the other 15 – circa 55 – 78. This cannot necessarily be explained by the data. However, 3 of the highest-ranking films are also the highest earning.

In addition, 2 of the lowest 5 carry 'Certified Fresh' status. At the same time, 1 of the 'Rotten' films is highly ranked at near 90. This highlights the divergence between audience ratings and critics' ratings i.e. the Tomatometer.

Generally, though, there is a weak positive correlation between the audience rating and the gross earnings.

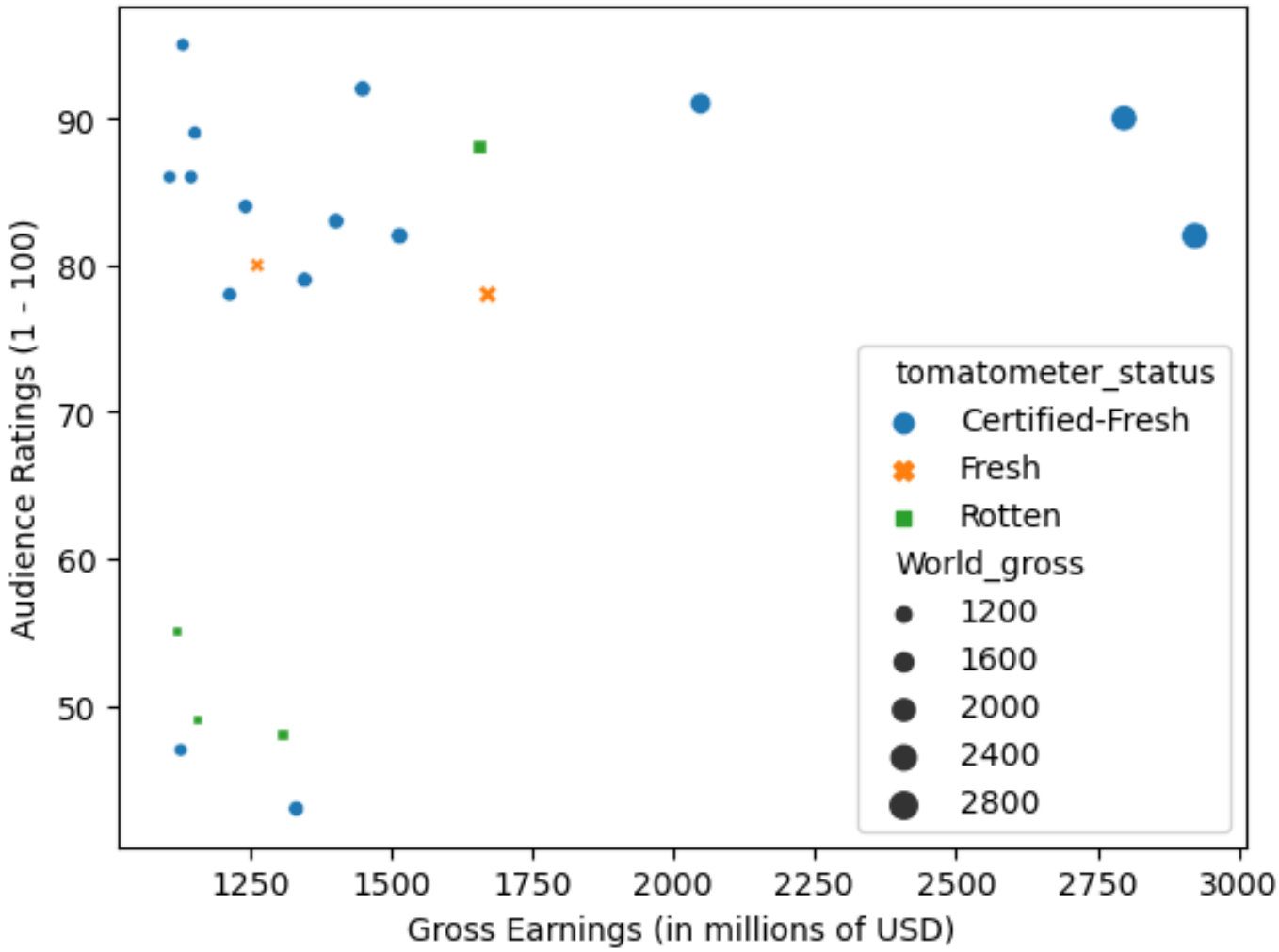


Chart 4: Audience Ratings vs Tomatometer Ratings

Generally there is a strong consensus between both audience and critic ratings concerning the highest-ranking films. This is evidenced by the large clustering of films in above circa 75 in the upper right of the chart.

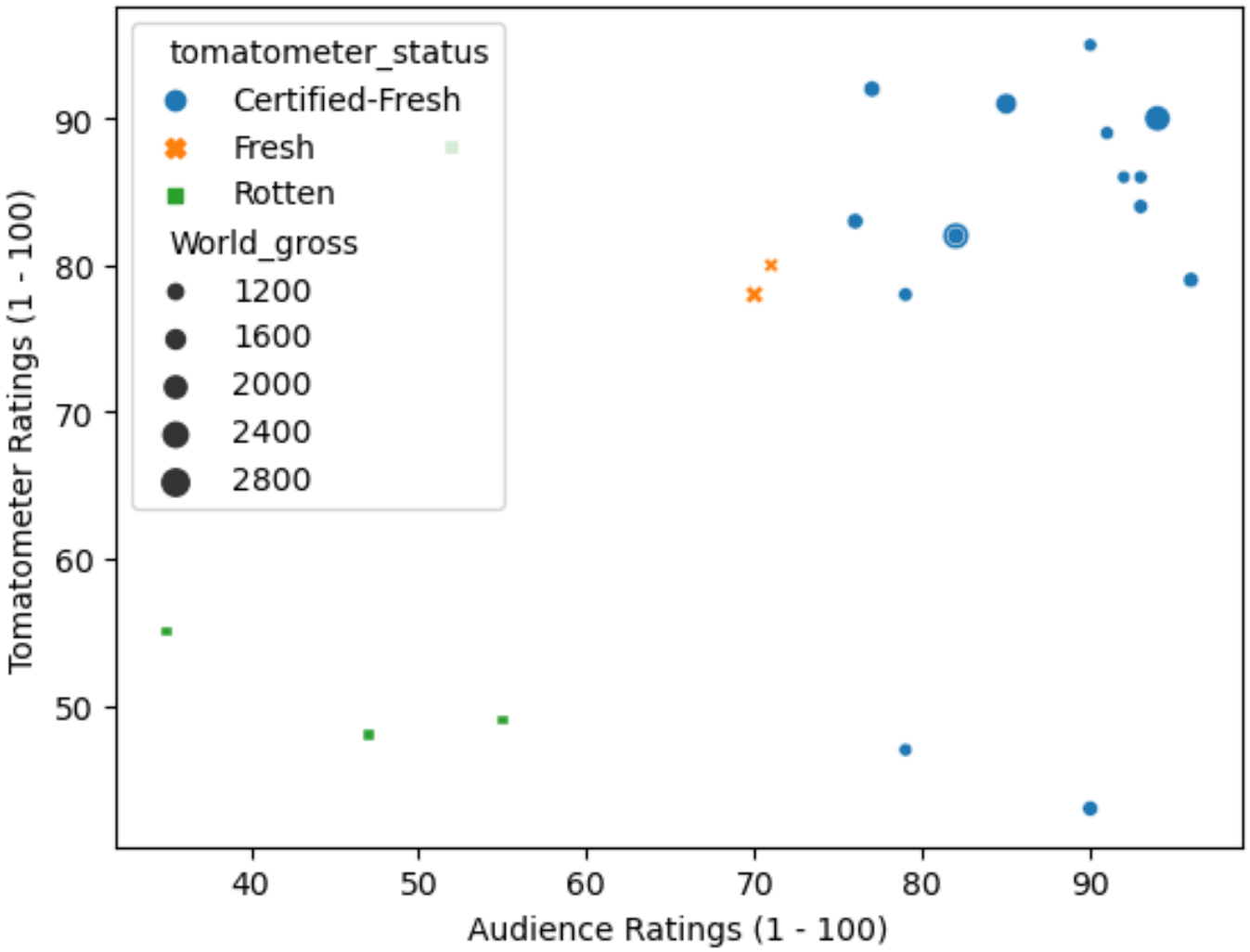


Chart 5: Gross Earnings vs Year

The chart illustrates that the vast majority of the highest earning films were released after 2010. The largest grouping of these between 1 and 1.75 billion dollars – as grouped in the upper left of the chart.

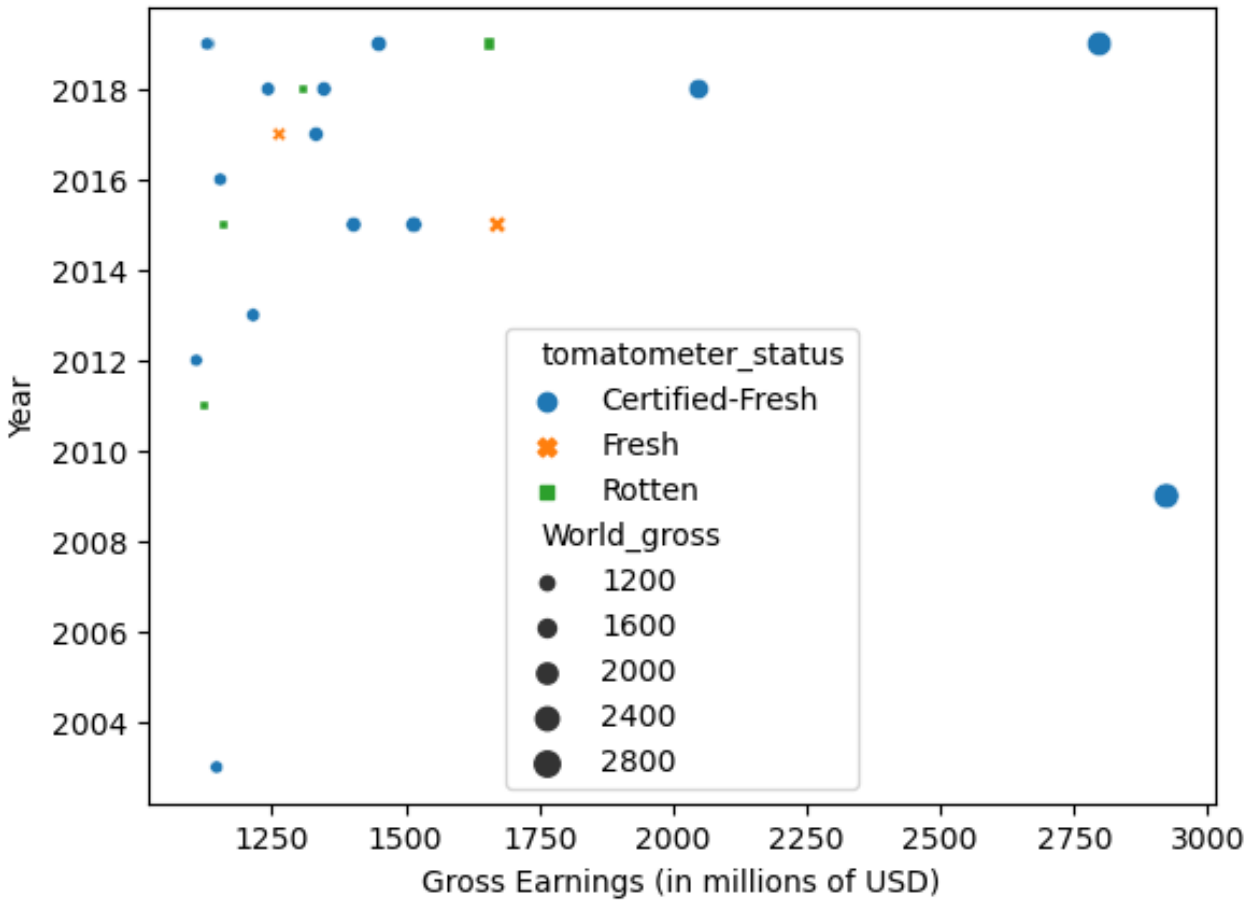
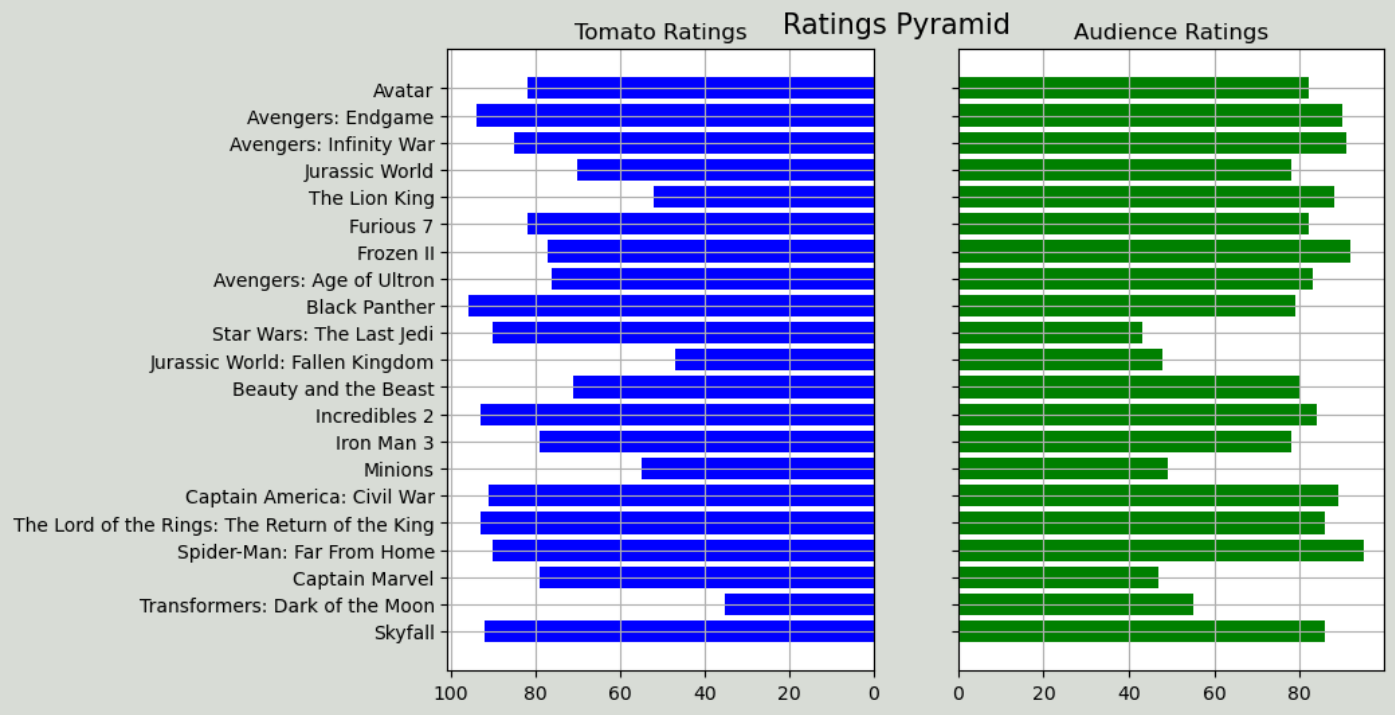


Chart 6: Ratings pyramid

The pyramid indicates that the films that had the highest gross takings tend to have higher audience ratings. This is most evident from 'Black Panther' up. This would make sense however, as film audiences are where earnings are generated. Inverse to this, tomato ratings i.e. critics ratings tend not have as great an impact on film earnings.

The films below 'Black Panther' have a very similar pyramid shapes.



Insights

1. There is a weak positive correlation between the audience rating and the gross earnings of the highest earning films.
2. The highest critical ratings do not necessarily transfer over to the highest grossing films.
3. There is a strong positive correlation between audience and critic ratings i.e. they are similar.
4. The vast bulk of highest earning films were released after 2010.
5. Audience ratings have a greater impact on gross earnings than critic ratings.

Machine Learning

It may be possible to use machine-learning tools, combined with larger amounts of data from sources such as IMDb, Meta Critic and of course, Rotten Tomatoes, to determine potential earnings based on prior critic reviews and early audience reviews of films before their run in theatres comes to end. However, extra data such as marketing budgets, cast, film release dates and type of film could also be incorporated into this.

References

www.wikipedia.com

www.kaggle.com

https://universityobserver.ie/the-importance-of-film-criticism/