Parallel Computing with Letterboxd Data to Analyze Movies

Group 11: Jonathan Dietrich, Matthew Giadla, Kaelin Silas, Julie Walsh, Sijia Wu

Introduction

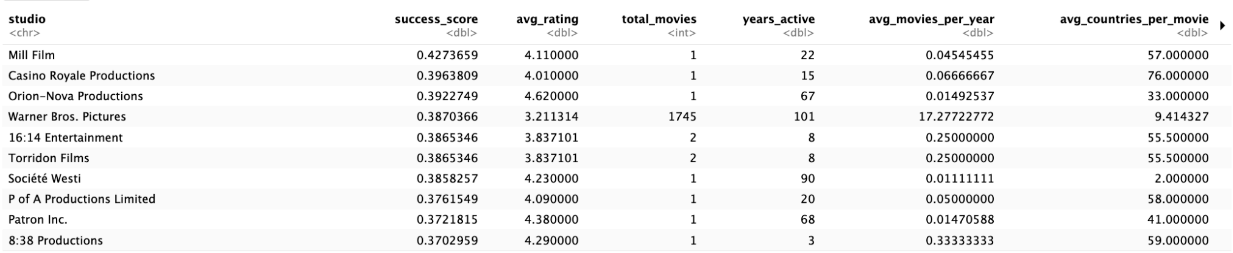
In this project, we investigated what makes a movie studio "successful" and what types of movies tend to emerge from different measures of success. Our goal was to investigate whether a successful studio is better measured in two different ways: through cumulative studio metrics or through user popularity alone over time. In **Method 1**, we built a cumulative success metric by parallelly integrating data across multiple CSV files, combining factors such as movie ratings, global reach, studio longevity, and total movie production. Average ratings were only weighted at 25% of the total success score. In **Method 2**, we shifted focus to ratings alone, analyzing how a studio's ratings evolved over the past decade without considering production scale. This allowed us to compare the two definitions of success. Through Method 1, we identified successful studios producing long-standing, high-budget films, such as Warner Bros. and Mill Film (*Gladiator*), reflecting strong commercial endurance. In contrast, Method 2 resulted in studios such as BBC, ARTE, and France 2/3 Cinéma, which lean towards educational, documentary, and socially driven storytelling. Our project revealed that different success metrics via Letterboxd yield varying studios and movies, whether cumulative factors or user ratings are prioritized.

Body

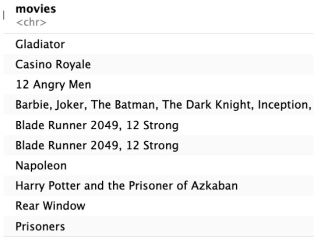
We used a 24GB Letterboxd dataset from Kaggle with information on 950,000+ movies. We analyzed CSV files with information on actors, countries, crew, genres, languages, movies, release dates, studios, and themes, including a posters database linking movie posters.

Method 1: Cumulative metrics

We calculated multiple studio-level metrics to capture dimensions of a successful studio. We submitted an R job focused on cleaning, joining, and summarizing, requesting 1 CPU, 4GB of memory, and 2GB of disk space.



studio <chr>	success_score <dbl>	avg_rating <dbl>	total_movies <int>	years_active <dbl>	avg_movies_per_year <dbl>	avg_countries_per_movie <dbl>
Mill Film	0.4273659	4.110000	1	22	0.04545455	57.000000
Casino Royale Productions	0.3963809	4.010000	1	15	0.06666667	76.000000
Orion–Nova Productions	0.3922749	4.620000	1	67	0.01492537	33.000000
Warner Bros. Pictures	0.3870366	3.211314	1745	101	17.27722772	9.414327
16:14 Entertainment	0.3865346	3.837101	2	8	0.25000000	55.500000
Torridon Films	0.3865346	3.837101	2	8	0.25000000	55.500000
Société Westi	0.3858257	4.230000	1	90	0.01111111	2.000000
P of A Productions Limited	0.3761549	4.090000	1	20	0.05000000	58.000000
Patron Inc.	0.3721815	4.380000	1	68	0.01470588	41.000000
8:38 Productions	0.3702959	4.290000	1	3	0.33333333	59.000000

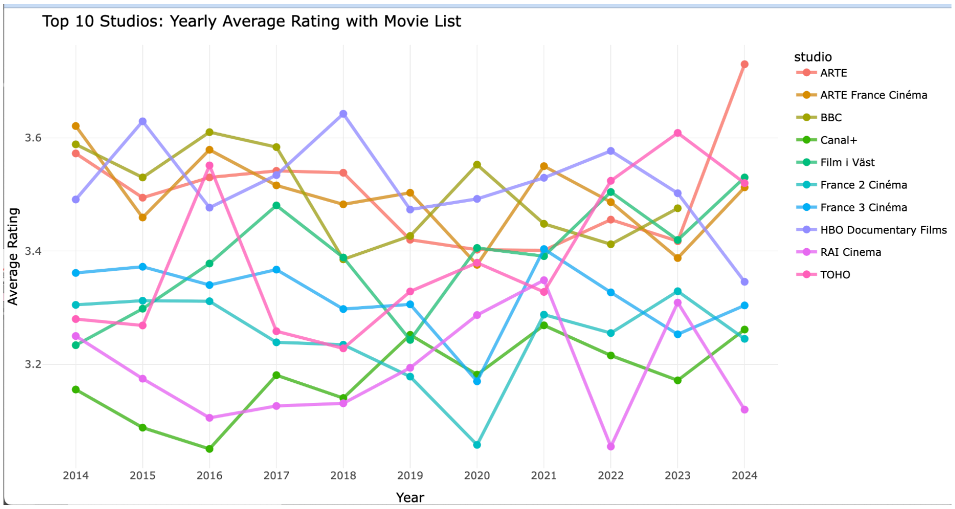


movies <chr>
Gladiator
Casino Royale
12 Angry Men
Barbie, Joker, The Batman, The Dark Knight, Inception,
Blade Runner 2049, 12 Strong
Blade Runner 2049, 12 Strong
Napoleon
Harry Potter and the Prisoner of Azkaban
Rear Window
Prisoners

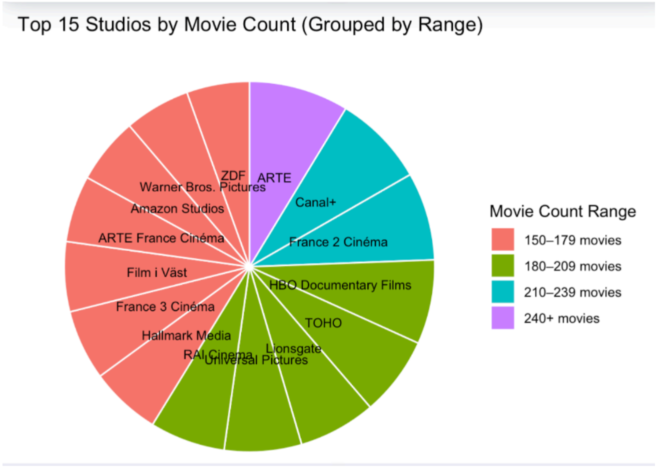
Method 1 Results: Our results, validated through outside research, identified studios like Mill Film (*Gladiator*, five Academy Awards), Casino Royale Productions (17 awards), Orion-Nova Productions (*12 Angry Men*, three Oscars), and Warner Bros. The studios that emerged from this cumulative metric approach tended to reflect those with greater longevity, critical acclaim, and box office success, capturing widespread audience reach over time.

Method 2: User rating alone

We created *top10_movies.csv,* which retained only movies with complete information. To manage the volume of over 950,000 movie records, we used parallel computing to extract and partition four groups for merging and processing. We submitted 10 separate jobs, each focused on analyzing one studio's data, with each job requiring less than one minute to run while requesting 512 MB of memory and 4096 MB of disk space on the CHTC system. We grouped by studio and year, with the average rating calculated for each studio-year combination.



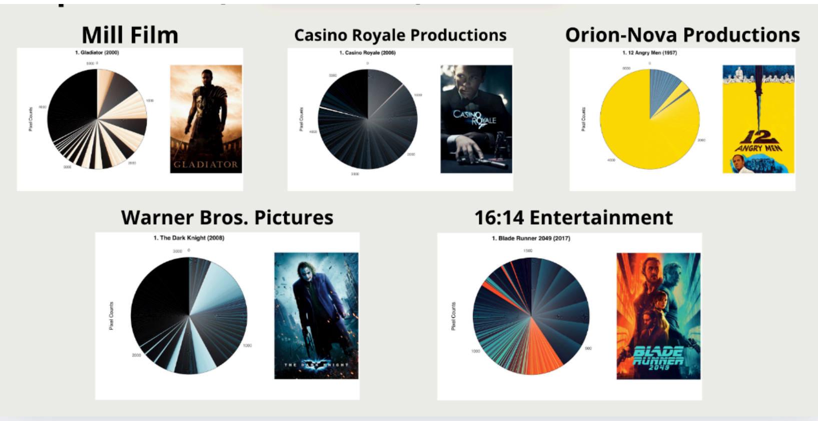
Method 2 Results: We noticed that a majority of studios, such as BBC, ARTE, and France 2/3 Cinéma are linked to public broadcasting or government-supported cinema. This suggests that studios focused on cultural and artistic content receive higher user ratings, even if they are not the most commercially dominant. From ratings alone, we see that Letterboxd users value critical storytelling over mainstream popularity.



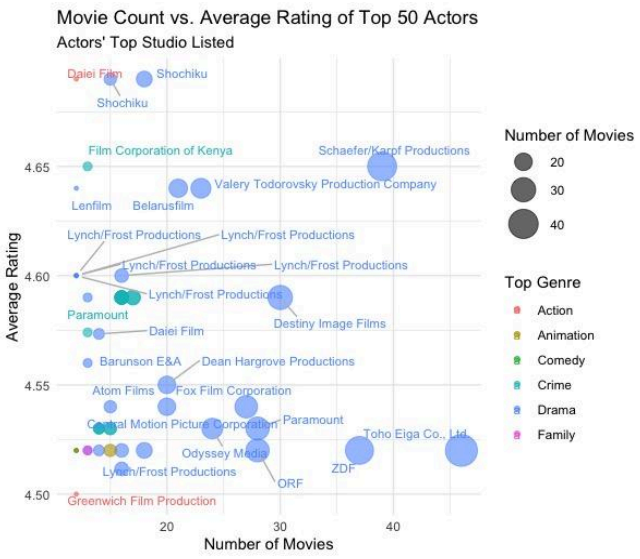
Conclusion:

In this project, we explored what makes a movie studio successful, using two approaches: cumulative studio metrics and user ratings over time. The cumulative method emphasized studios with longevity, large productions, and global reach, reflecting blockbuster-style success. In contrast, user ratings highlighted artistic, independent studios valued by cinephiles. Limitations include Letterboxd user bias, such as users only representing a younger age range. Future work could involve statistically determining optimal metric weights to refine how success is measured across different types of studios.

Supplemental Visualizations



Top studio’s movie posters: Color study



Studios and Average Actor Ratings

Contributions

Name	Proposal	Code	Report	Presentation
Scarlett Wu	**1**	**1**	**0.6**	**1**
Kaelin Silas	**1**	**.5**	**.5**	**1**
Julie Walsh	**.7**	**.8**	**1**	**1**
Jonathan Dietrich	**0.7**	**1**	**0.5**	**1**
Matthew Giadla	**.7**	**.5**	**.7**	**1**