Xi'an Jiaotong-Liverpool University
西交利物浦大学

School of Advanced Technology

Project 1 Report

Project Title: WEB DATA SCRAPING

Student Name: TANIA NATASYA TANUWIJAYA

Student ID: 1825453

Project field: INT303

Supervisor:

Co-supervisor (if applicable):

Top 100 Movies from IMDB Analysis: Does the IMDB rating reliable?

In recent years, box office movies have gotten more attention from the public. Mcclintock (2021) predicts that the revenues of the global box office movies in 2021 will be 80% ahead of 2020 but still 49% behind 2019. Therefore, it is essential to analyse all-time favourite movies to give viewers insight into what is measured as a good movie. The top 100 movies from Internet Movie Database (IMDB) have been scraped into a single CSV file using BeautifulSoup as the Python library to make this analysis. Below is provided the analysis result for the top movies based on the IMDB users' preference.

1. All-time favourite movies based on IMDB: The top 10 movies

   According to the IMDB website, "The Shawshank Redemption" movie released in 1994 has the highest rating in IMDB, 9.3, with nearly 2.5 million people votes. In fact, 1994 is also the mode and median for the all-time favourite movies, which means films in 1994 are mostly liked by the viewers. For instance, there are five titles mentioned in the list produced in 1994, and 2 of those movies are in the top 10 ratings (Table 1). Cykiert (2014) also argued that 1994 was the best year for the film industry. In addition, from the first top ten movies, only three movies were produced after 2000 but before 2010 (Table 1). The old movies are preferred rather than the new movies. In addition, drama is considered as the best genre for a movie. 90% of the movies in the top 10 list were categorized as a drama, and people also like the combination of drama and crime genre in a movie.

   1.1. All-time favourite movies based on IMDB: The old movies VS the new movies

   Furthermore, the range of movies' release years is 89 years, from 1931 to 2020. Interestingly, 65 movies were produced before 2000, whereas only 35 were produced after 2000 (Figure 1). Do those old movies still good enough and worth watching now? One of the reasons movies before 2000 still successfully hit the top 100 list is that the old movies receive more reviews than the new movies (Bischoff, 2016). It is possible that since the IMDB started its service in the 1990s and within a few years, this site became popular, and it affected the movies review (Bischoff, 2016). Nevertheless, reviewing a film in IMDB is quite rare activity nowadays, especially for Generation Z. That is why new movies have less rating rather than the old movies.

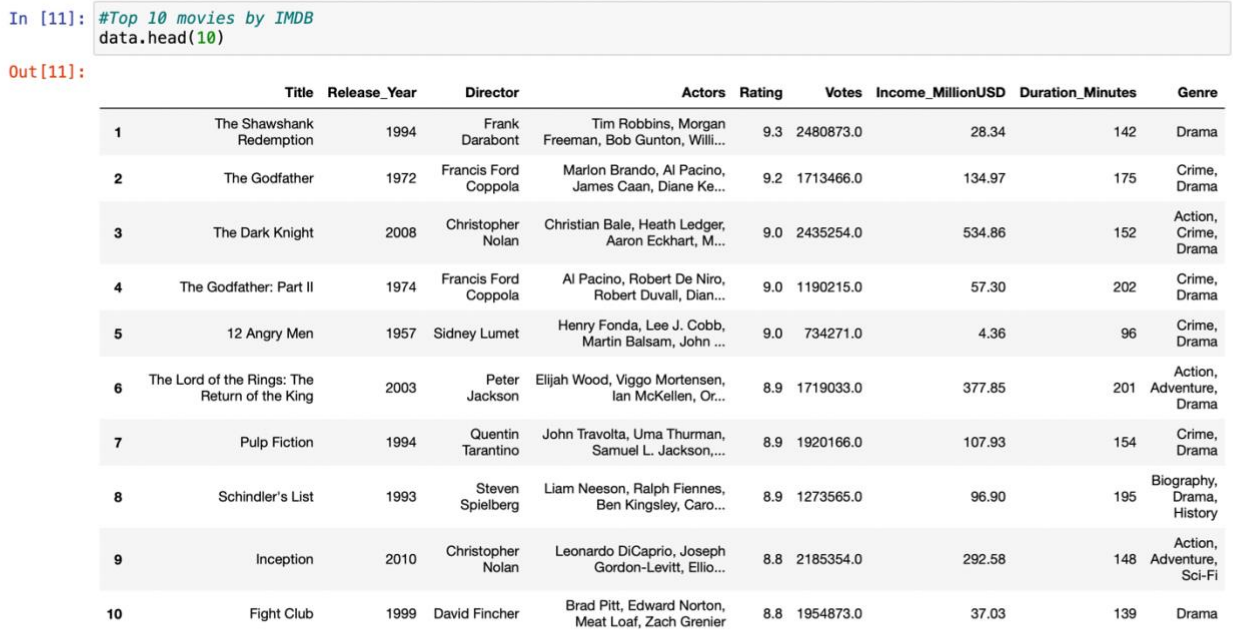
2. Revenue: What factors do affect the movie's revenue?

   Based on the analysis, movies that were produced in 2019 had the highest revenue, accounting for a total of 1247.19 million USD, while the average is 415.7 million USD (Figure 2). In that year, one of the Avengers sequels, Endgame, reached the highest cumulative income, 858 million USD (Table 2). In addition, another Avengers sequel, Infinite War, produced in 2018, also placed in a second position with more than 678 million US dollars. If we analyze it further, it is interesting that most of the movies in the highest revenue position mostly are the continuation, such as Avengers, The Dark Knight, The Lord of the Rings, and Star Wars. In other words, the success of these movies might depend on the content, which is the director and the actors that they were known before. For instance, the Avengers and the Star Wars' main casts remain the same for their sequel. On the contrary, Christopher Nolan has directed six movies chosen as the all-time best movie. Also, Tom Hanks played 4 times in the top 100 movies, which is the most compared to other actors in the top 100 movies. Although they did not contribute to the sequel movies, they still successfully hit the top 100. However, the mean rating from 2019 shows no relationship between rating with either income or votes (Figure 3 and 4).

3. Duration: Does the movie's runtime affect the rating of the movie?
   Dahlgreen (2015) stated that the ideal duration for a movie is between 90 to 120 minutes. However, the mean duration of the all-time favorite movies is 134.62 minutes or approximately equal to 2 hours and 15 minutes (Figure 5). In fact, more than half of the movies exceed the ideal time and even the average time (Figure 6). The longest duration in the 100 best movies is "Once Upon a Time in America" at 229 minutes, while the shortest is "Toy Story" at 81 minutes. Rosser (2019) reported that duration does not affect the movie's revenue and rating. In the report, he took "Avengers: Endgame" from 2019 as an example. It ran for 181 minutes yet still reached a high rating and cumulative income.

In conclusion, people do not care about the rating and the duration of a movie, but they are more considerate about the director and the actors, which is the content they have recognized before. It is proven from the top movies' revenue list, mostly it has the same directors and actors. When the revenue is high, it means more people had spent their money to watch the movie that they consider worth watching. Jasper (2018) mentioned that we could use the movie review website as a reference, but we cannot fully believe in IMDB reviews since they are biased towards men's preferences (Reynolds, 2017). Therefore, IMDB rating is not reliable for people to judge whether it is a good movie or not.

Appendix



```
In [11]: #Top 10 movies by IMDB
         data.head(10)
```

Out[11]:

	Title	Release_Year	Director	Actors	Rating	Votes	Income_MillionUSD	Duration_Minutes	Genre
1	The Shawshank Redemption	1994	Frank Darabont	Tim Robbins, Morgan Freeman, Bob Gunton, Willi...	9.3	2480873.0	28.34	142	Drama
2	The Godfather	1972	Francis Ford Coppola	Marlon Brando, Al Pacino, James Caan, Diane Ke...	9.2	1713466.0	134.97	175	Crime, Drama
3	The Dark Knight	2008	Christopher Nolan	Christian Bale, Heath Ledger, Aaron Eckhart, M...	9.0	2435254.0	534.86	152	Action, Crime, Drama
4	The Godfather: Part II	1974	Francis Ford Coppola	Al Pacino, Robert De Niro, Robert Duvall, Dian...	9.0	1190215.0	57.30	202	Crime, Drama
5	12 Angry Men	1957	Sidney Lumet	Henry Fonda, Lee J. Cobb, Martin Balsam, John ...	9.0	734271.0	4.36	96	Crime, Drama
6	The Lord of the Rings: The Return of the King	2003	Peter Jackson	Elijah Wood, Viggo Mortensen, Ian McKellen, Or...	8.9	1719033.0	377.85	201	Action, Adventure, Drama
7	Pulp Fiction	1994	Quentin Tarantino	John Travolta, Uma Thurman, Samuel L. Jackson,...	8.9	1920166.0	107.93	154	Crime, Drama
8	Schindler's List	1993	Steven Spielberg	Liam Neeson, Ralph Fiennes, Ben Kingsley, Caro...	8.9	1273565.0	96.90	195	Biography, Drama, History
9	Inception	2010	Christopher Nolan	Leonardo DiCaprio, Joseph Gordon-Levitt, Ellio...	8.8	2185354.0	292.58	148	Action, Adventure, Sci-Fi
10	Fight Club	1999	David Fincher	Brad Pitt, Edward Norton, Meat Loaf, Zach Grenier	8.8	1954873.0	37.03	139	Drama

Table 1

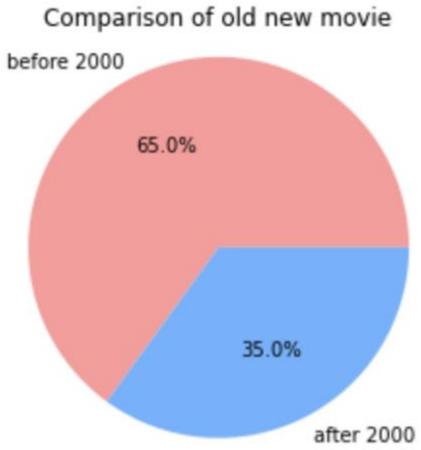


Figure 1

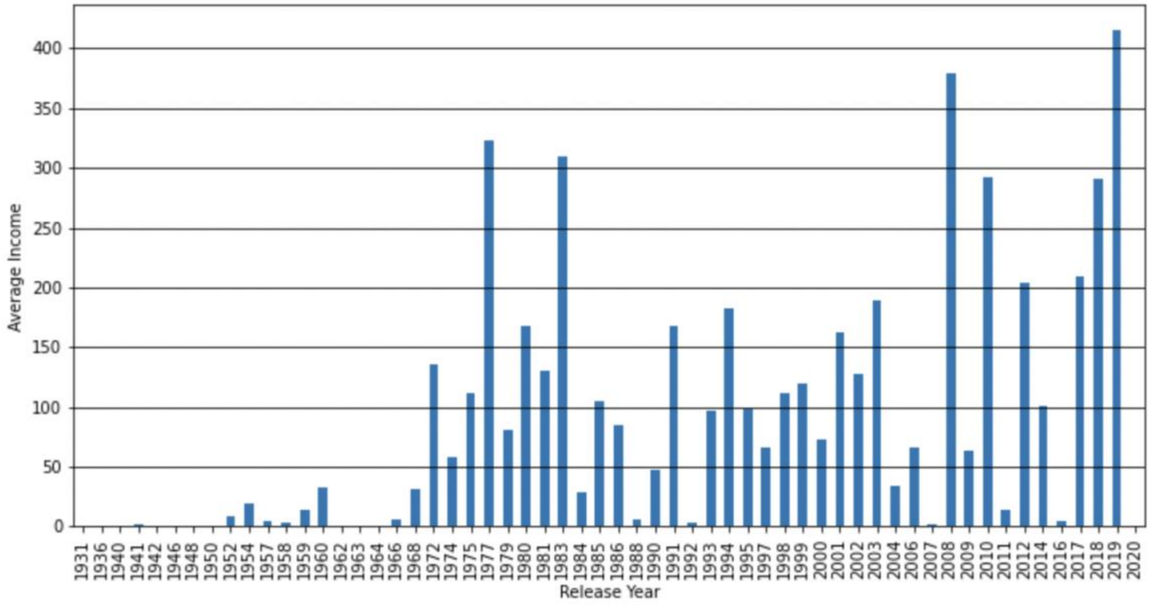
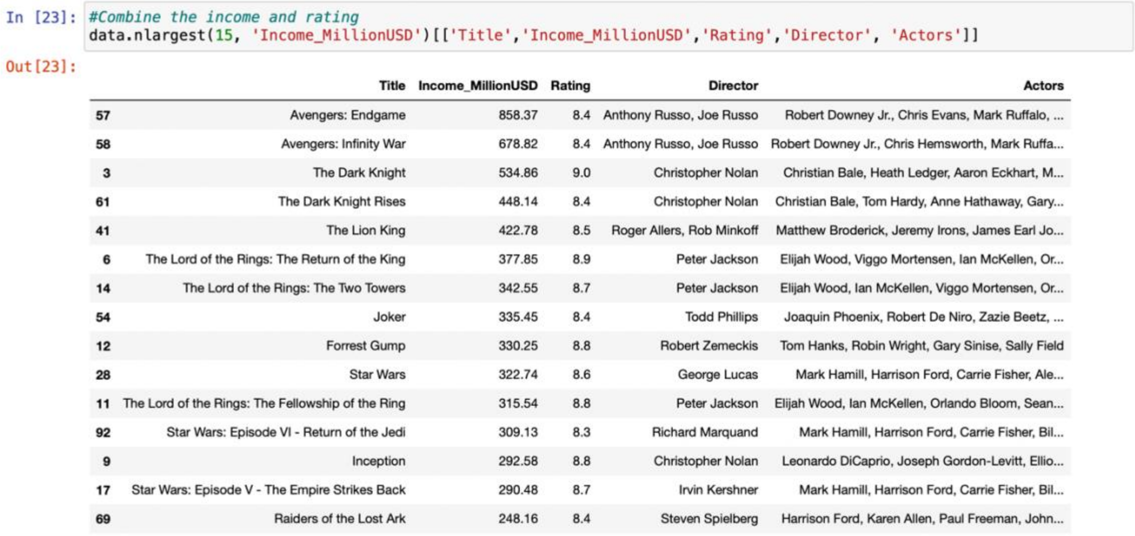


Figure 2



In [23]:
```
#Combine the income and rating
data.nlargest(15, 'Income_MillionUSD')[['Title','Income_MillionUSD','Rating','Director', 'Actors']]
```

Out[23]:

	Title	Income_MillionUSD	Rating	Director	Actors
57	Avengers: Endgame	858.37	8.4	Anthony Russo, Joe Russo	Robert Downey Jr., Chris Evans, Mark Ruffalo, ...
58	Avengers: Infinity War	678.82	8.4	Anthony Russo, Joe Russo	Robert Downey Jr., Chris Hemsworth, Mark Ruffa...
3	The Dark Knight	534.86	9.0	Christopher Nolan	Christian Bale, Heath Ledger, Aaron Eckhart, M...
61	The Dark Knight Rises	448.14	8.4	Christopher Nolan	Christian Bale, Tom Hardy, Anne Hathaway, Gary...
41	The Lion King	422.78	8.5	Roger Allers, Rob Minkoff	Matthew Broderick, Jeremy Irons, James Earl Jo...
6	The Lord of the Rings: The Return of the King	377.85	8.9	Peter Jackson	Elijah Wood, Viggo Mortensen, Ian McKellen, Or...
14	The Lord of the Rings: The Two Towers	342.55	8.7	Peter Jackson	Elijah Wood, Ian McKellen, Viggo Mortensen, Or...
54	Joker	335.45	8.4	Todd Phillips	Joaquin Phoenix, Robert De Niro, Zazie Beetz, ...
12	Forrest Gump	330.25	8.8	Robert Zemeckis	Tom Hanks, Robin Wright, Gary Sinise, Sally Field
28	Star Wars	322.74	8.6	George Lucas	Mark Hamill, Harrison Ford, Carrie Fisher, Ale...
11	The Lord of the Rings: The Fellowship of the Ring	315.54	8.8	Peter Jackson	Elijah Wood, Ian McKellen, Orlando Bloom, Sean...
92	Star Wars: Episode VI - Return of the Jedi	309.13	8.3	Richard Marquand	Mark Hamill, Harrison Ford, Carrie Fisher, Bil...
9	Inception	292.58	8.8	Christopher Nolan	Leonardo DiCaprio, Joseph Gordon-Levitt, Ellio...
17	Star Wars: Episode V - The Empire Strikes Back	290.48	8.7	Irvin Kershner	Mark Hamill, Harrison Ford, Carrie Fisher, Bil...
69	Raiders of the Lost Ark	248.16	8.4	Steven Spielberg	Harrison Ford, Karen Allen, Paul Freeman, John...

Table 2

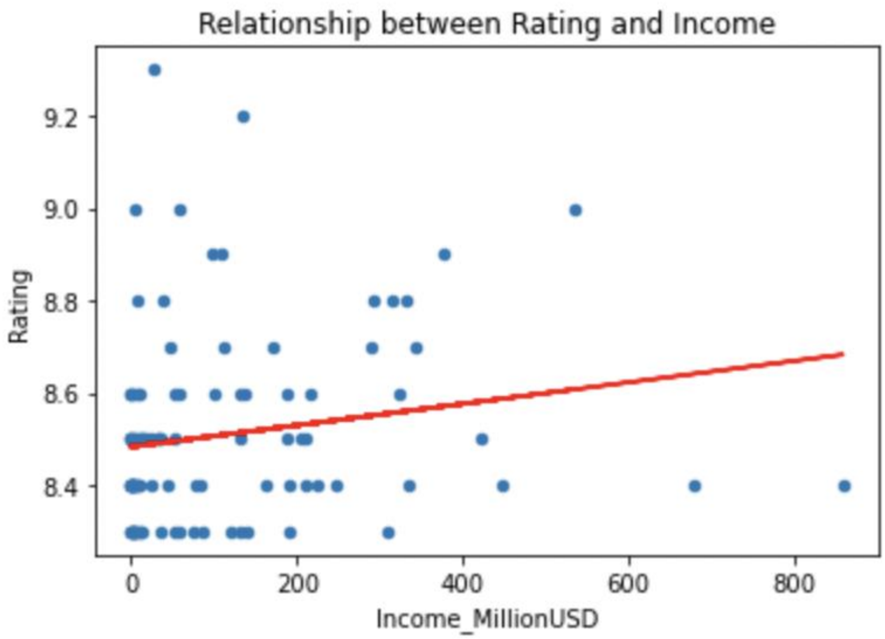


Figure 3

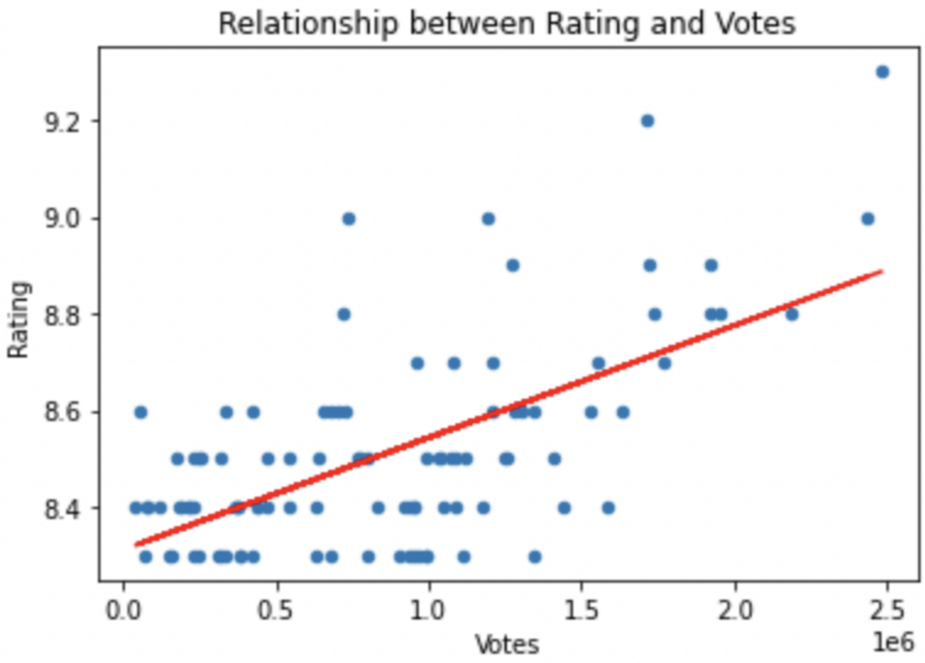


Figure 4

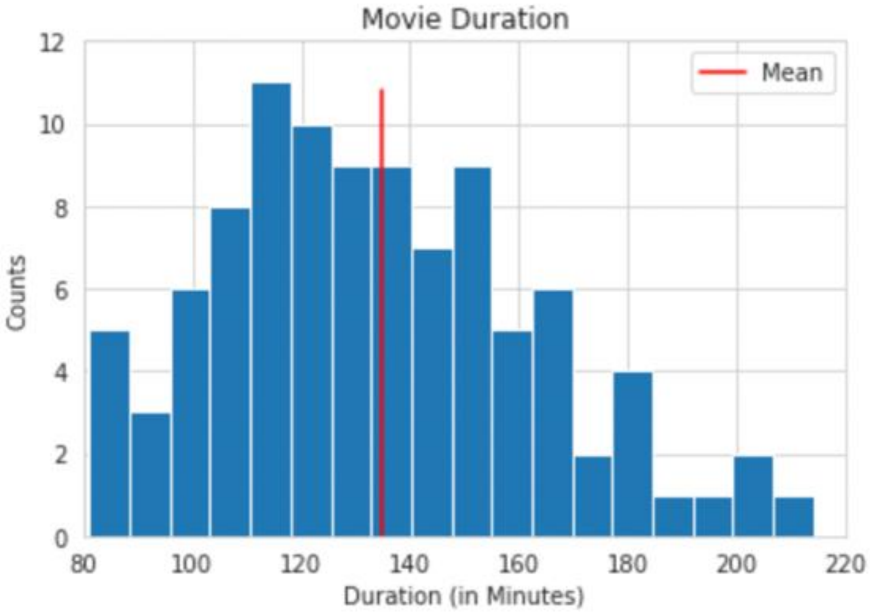


Figure 5

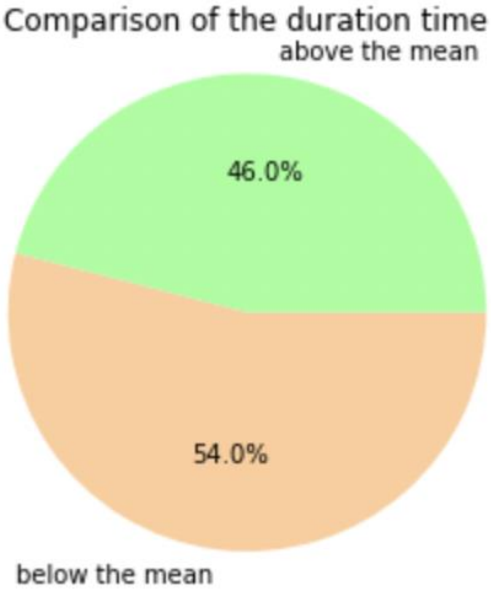


Figure 6

Reference List

Bischoff, P. (2016) *Why old movies get better ratings on Rotten Tomatoes, Metacritic, and IMDB*. Available at: https://medium.com/@pabischoff/why-old-movies-get-better-ratings-on-rotten-tomatoes-metacritic-and-imdb-a5f030031834 (Accessed: 2 November 2021).

Cykiert, N. (2014) *The greatest film debate of our generation: 1993 or 1994?*. Available at: https://www.theverge.com/2014/8/22/6056729/1993-vs-1994-what-was-the-best-year-in-film#1994 (Accessed: 2 November 2021).

Dahlgreen, W. (2015) *Note to producers: the ideal movie length is under 2 hours*. Available at: https://yougov.co.uk/topics/lifestyle/articles-reports/2015/09/26/note-producers-perfect-movie-length-under-2-hours (Accessed: 2 November 2021).

Jasper, M. G. (2018) *10 Essential Elements for Movie Reviews*. Available at: https://medium.com/@m.g.jasper/10-essential-elements-for-movie-reviews-921230d7fb1e (Accessed: 2 November 2021).

Mcclintock, P. (2021) *Global box office: Forecast for 2021 Revenue Improves to $21.6B*. Available at: https://www.hollywoodreporter.com/movies/movie-news/global-box-office-forecast-improves-1235036692/ (Accessed: 2 November 2021).

Reynolds, M. (2017) *You should ignore film ratings on IMDB and Rotten Tomatoes*. Available at: https://www.wired.co.uk/article/which-film-ranking-site-should-i-trust-rotten-tomatoes-imdb-metacritic (Accessed: 2 November 2021).

Rosser, M. (2019) *Does a long running time help or hurt a film's box office performance?*. Available at: https://www.screendaily.com/features/does-a-long-running-time-help-or-hurt-a-films-box-office-performance/5144271.article (Accessed: 2 November 2021).