Movie Rating and Revenue Analysis

by

Petra Ivanovic (pi2018), Jon Dinh (jhd9252), Nabiya Alam (na2794)

Processing Big Data for Analytics Applications,

New York University, Fall 2023

Abstract:

Our study aims to learn more about the different dynamics and nuances of the film industry, particularly focusing on how various elements such as cast, directors, genres, and studios, influence movie ratings and box-office revenues. We believe that our results can be used by various groups to decide which movies to watch and which to invest in. We used big data platforms and methods such as HDFS, Hadoop, and Scala, to analyze multiple datasets, including those from IMDb and MovieLens. Our key findings reveal significant correlations between the gender of the main actor/actress, the studio that produced the movie, the time (month) that the movie was released, and the movie's performance, both in terms of box office and rating. Our insights not only offer a deeper understanding of the determinants of movie success but also provide valuable guidance for stakeholders in the film industry and highlight potential issues that need to be resolved such as bias in casting. We concluded our research with interpretations of the data that underline the complexities of predicting movie success. Further, we offered insight to readers on how to understand what they can expect from movies based on the cast, director, studio, and time of the year it was released. Finally, we compiled recommendations to industry professionals and studios on how to recognize and reduce potential bias in their field.

Key Words: Movie Analysis, Box-Office Revenue Analysis, Ratings, Big Data, Film Industry Analysis, Predictive Analytics

1 Introduction

In this study we conducted a detailed analysis of the various factors influencing movie success, particularly focusing on box-office revenues and ratings. This research is important given the evolving nature of the film industry, where understanding which factors lead to success is crucial for stakeholders, from filmmakers to marketers and audiences. In the past, we knew that different factors influenced the movie industry, such as star power, but we did not understand the nuances of it or the extent to which they can influence success. Our goal was to conduct a holistic analysis that would take into account many different factors and their relationship to movie success. This knowledge gap motivated our comprehensive approach, leveraging big data analytics to explore datasets from IMDb and MovieLens. This study will present a more detailed approach to analyze a few different factors that influence movie performance, such as time of release, gender of lead actors, and directors, in order to provide a more intricate understanding of the film industry's success factors. This approach will offer valuable insights to various stakeholders to both make decisions on a personal level - whether they will watch the film - and on a professional level such as which actors to cast and which movies to fund.

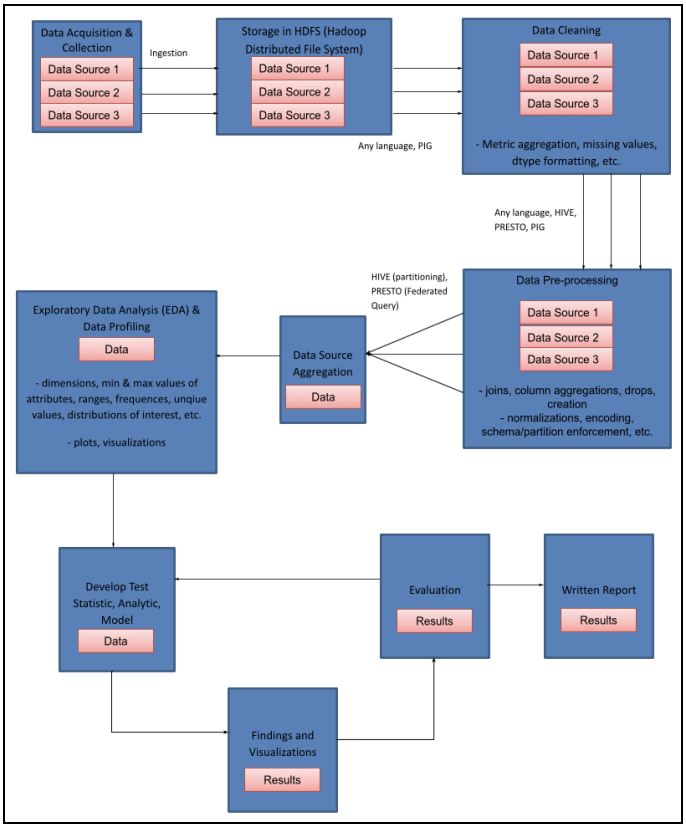


Figure 1: Data Flow Diagram

Figure 1 shows our Data Flow Diagram which details the steps we took from data collection and processing to evaluation and presentation. Starting with individual dataset collection and ingestion into HDFS, we leveraged Scala and Hive for data cleaning. After this, each dataset was pre-processed to be compatible for merging with the other 2 datasets. The consolidated dataset was cleaned again using Scala to remove unnecessary and duplicate features. Then, we conducted an Exploratory Data Analysis on the consolidated dataset. At this point we had a very good understanding of the data we were working with and potential findings we would like to look for. After we had a couple of relevant and interesting findings, such as the fact that releasing movies in certain months led to significantly higher box office, we looked at our references and different journal articles to see if they could corroborate our findings. Finally, we evaluated our findings, created valuable visualizations to present them, and repeated the process with many different interesting analytics.

2 Motivation

Regardless of one's background, one thing most people share is a love for movies. While they might not be movies of the same genre, our analysis can be used by anyone even slightly interested in either watching or creating movies. We addressed the question, "What makes a movie successful in terms of the money it earns and the ratings it receives?". We wanted to see how things like who's acting or directing, the type of story being told, and when the movie is released can make the movie successful (or unsuccessful.) Our goal was to find what matters the most in making a popular and well-rated movie, which can help many different stakeholders. More precisely, for movie audiences, it can help them decide which movie to watch when indecisive as our analysis can show them what kind of ratings they can expect based on the factors such as cast and genre. Further, the analysis can also be used by movie producers and advertisers who might want to know what kind of movies would be most popular when deciding on budgets and cast. Additionally, it can be used by smaller cinemas to decide which movies to choose to screen and when to screen them for the most profit. Finally, it can be used by streaming platforms such as Netflix or Hulu to decide which movies to include on their applications to reach larger audiences.

3 Related Works

3.1 *Is Studio Size Important to Box Office Success?* by Barrie Gunter, 2018

In this chapter of *Predicting Movie Success at the Box Office* (2018), Barrie Gunter talks about the relationship between studio size and movies' success in the box office. This strongly corroborated our findings that there are 6 studios that hold more than 70% of the lifetime gross out of all the studios. Further, in this chapter, Gunter states that "five most active companies (Disney, Sony, Warner Brothers, Paramount and MGM) that year [2002] released more than half of the total films"[1]. This not only shows similar results to ours but we can also see that the same studios are still leading the industry now as they were in 2002.

3.2 *The promise is great: the blockbuster and the Hollywood economy* by Marco Cucco, 2009

In this journal article, Marco Cucco discusses the effect of blockbusters on the Hollywood economy and talks about why these movies are released at certain times of the year. The article highlights the importance of good timing when releasing movies with the most promise. Cucco also emphasizes the importance of communication with other large studios to make sure two

[1] Barrie Gunter, 'Is Studio Size Important to Box Office Success?', in *Predicting Movie Success at the Box Office* (Springer International Publishing, 2018), 38.

blockbuster movies are not coming out at the same time as that could cause diminished earnings for both. Further, our findings showed that in certain months of the year (May, June, July, and December), movies with higher box offices come out. We assumed that this could be due to the fact that that is when blockbuster movies are released and also when people have more time to spend at the movies. The article strongly corroborates both of these findings as it claims "that the favorite period for the release of blockbusters starts on Memorial Day weekend (28 May) and ends on Labor Day (the first Monday of September). In other words, blockbusters tend to come out in summer when the network's competition is weaker, people are more willing to go out and youngsters have more spare time."[2] This shows us that the article validates both our findings and our assumptions about the reasoning behind the findings.

3.3 *Correlations between user voting data, budget and box office for films in the internet movie database* by Max Wasserman, Satyam Mukherjee, Konner Scott, Xiao Han T. Zeng, Filippo Radicchi, Luís A. N. Amaral, 2014

Wasserman et al., in the *Journal of The Association for Information Science and Technology*, surveys the underlying distribution and existing relationships between user generated data and economic standards such as box office gross[3]. In their research, features were compared with double log standards according to time period, while language and origin country bias was filtered out using a generated acyclic graph of movie references. Corroborating our feature significance and correlation estimates, they found a strong correlation between the user generated number of votes and IMDB provided economic features. Their results suggest that there is a relationship between total user votes and film prominence, specifically using the number of user votes as an indicator. The research article aligns with our estimates and findings using the same datasets in addition to two more databases.

3.4 *Early Prediction of Movie Success using Machine Learning Models* by D.M.L. Dissanayake, V.G.T.N. Vidanagama, 2021

Dissanayake and Vidanagama focus on identifying industry controllable features such as director and cast engagement with the public in order to discover factors relevant to the economic success of a movie[4]. Pertaining to our research findings, the researchers also utilized user engagement generated data found on Facebook such as the number of likes on a director's post. The feature

[2] Marco Cucco, 'The Promise Is Great: The Blockbuster and the Hollywood Economy', *Media, Culture & Society* 31, no. 2 (2009): 227.
[3] Wasserman, M., S. Mukherjee, K. Scott, X. H. T. Zeng, F. Radicchi, and L. A. N. Amaral. "Correlations between User Voting Data, Budget, and Box Office for Films in the Internet Movie Database." Journal of the Association for Information Science and Technology 66, no. 4 (2015): 858-868.
[4] Dissanayake, D. M. L., and V. G. T. N. Vidanagama. "Early Prediction of Movie Success Using Machine Learning Models." International Journal of Computer Applications 183, no. 44 (2021).

significance methods of Chi-square, F-Test, feature importance and Recursive Feature Elimination were applied. The results of those methods suggest that the ten most pertinent features in predicting early success of a film were from user and cast engagement. Nine of the ten selected features post selection were features that were related to number of user reviews and employee engagements on Facebook. This aligns with our findings that total user engagement or voting data holds a high impact on the media's success.

3.5 *Tears or Fears? Comparing Gender Stereotypes about Movie Preferences to Actual Preferences* by Peter Wühr, et al., 2017

Peter Wühr et al. investigated the accuracy of stereotypes about gender and movie preferences. As we know, romance and comedy movies are often aimed at female audiences, whereas action and horror movies are generally tailored to men's preferences. This is often assumed due to the gender stereotypes many believe are true when it comes to movie preferences. The paper analyzes the legitimacy of these claims by having participants of both genders watch the same movies and then assign them ratings. Since our research aims to determine what makes movies successful, this paper is helpful as it delves into the details of the demographics and their movie preferences and specifically how representation makes movies more attractive to a certain audience. While we are not explicitly investigating genre, this study gives us a baseline for understanding what types of movies appeal to different audiences.

4 Description of Datasets

4.1 IMDb Datasets:

This dataset consists of seven individual datasets in TSV format, six of which are pertinent to our study. We have merged these datasets using shared identifiers to create an extensive dataset for detailed analysis and predictive modeling. Key features under examination include movie titles, crew details (such as directors and writers), principal figures (including cast and crew members), as well as movie ratings and vote counts. Comprehensive schema details are accessible here. The full size of all tables combined into the master dataset is 4.63 GB. All of these datasets can be found here.

Table 1: IMDb Final Dataset Schema

Column Name	Description
nconst (String)	Unique identifier of the primaryName for a particular movie.
tconst (String)	Unique IMDb identifier for each movie.
titleType (String)	Movie category - movie or TV movie.
primaryTitle (String)	Main title that the movie goes by.
originalTitle (String)	Original title of the movie.
isAdult (Binary)	Binary variable (0 or 1) on whether a movie is appropriate for only adult audience (1) or for all audience (0).
startYear (Int)	Year the movie was released.
runtimeMinutes (Int)	Length of the movie in minutes.
genres (String)	Categorical variable describing genre of the movie.
directors (String)	String giving the identifier(s) for director(s) for the movie.
writers (String)	String giving the identifier(s) for writer(s) for the movie.
job (String)	Job in which primaryName for the movie is in (usually actor / actress).
averageRating (Float)	Average IMDb rating for the movie.
numVotes (Int)	Number of IMDb votes for the movie.
primaryName (String)	Most known person (name) for the movie.

4.2 Box Office Data:

This dataset is from the website Data World. It contains a large list of Hollywood movies. Each row includes a movie with its ranking, the studio it was produced by, the lifetime domestic gross (not accounting for inflation), and the year it was released. The movie release years date back to the 1920s.

The size of the dataset is 1.22 MB with 16543 rows. It can be downloaded here. Additional information about the data, as well as an interactive dashboard for data exploration, can be found here.

Table 2: Box Office Dataset Schema

Column Name	Description
rank (string)	The ranking of the movie in descending order based on lifetime gross.
title (string)	The title of the movie.
studio (string)	The abbreviated name of the studio that produced the movie.
lifetime_gross (int)	The total amount of money the movie has made since its release.
year (int)	The year the movie was released.

4.3 Small MovieLens Dataset from GroupLens:

This collection of data comes from the GroupLens Research team at the Research Lab in the Department of Computer Science and Engineering at the University of Minnesota. Specifically, this dataset is a subsample of a larger collected rating dataset from the MovieLens website, a movie recommendation service. It contains an approximate 100,836 ratings and 3,683 tag applications, across 9,742 movies. The ratings were created by 610 users between the years 1996-2018. For quality control, and accurate metrics, the users were randomly selected, with each user having at least 20 ratings for consistent results.

There are a total of four CSV files named *links*, *movies*, *ratings* and *tags*. The *link* file included connecting keys to the IMDb and TMDB databases for additional desired features. The *movies* file contained basic features of the film (title, year, genres, etc.). The *ratings* file contains the userId, movieId, and rating. The *tags* file contained descriptive String comments.

Additional movie information was obtained through the PyMovieDb package using the provided IMDb unique identifiers, before the datasets were merged. Aligning with the focus of this research, tags and ratings were grouped by movies and aggregated before merging onto the industry controllable movie features.

The size of the data files total 3.71 MB. The individual datasets can be downloaded here, whereas additional information can be found here.

Table 3: Small MovieLens Dataset Schema

Column Name	Description
movieId	Unique identifier of movie in MovieLens database.
imdbId	Unique identifier of movie in IMDB database.
tmdbId	Unique identifier of movie in TMDB.
mediaType	Type of media (Movie, Short Film, etc.).
imdb_ratingCount	Total number of submitted ratings on IMDB grouped on movie.
imdb_bestRating	Highest rating given to a movie.
imdb_worstRating	Lowest rating given to a movie.
imdb_averageRating	Mean rating of movie after grouping and aggregation.
contentRating	Designated abbreviation for target audience (G, PG, PG-13, R, etc.).
releaseDate	YYYY-MM-DD format.
smallLensRatingCount	Total number of submitted ratings on MovieLens grouped on movie.
smallLensRatingAvg	Mean rating of movie after grouping and aggregation.

5. Analytic Stages, process

5.1 Data Ingestion

IMDb Non-Commercial Datasets were accessed on November 4th, 2023 from the IMDb website. After that the data was moved to DataProc and placed on HDFS. Lastly, it was ingested via Spark Scala.

The Box Office dataset was accessed on November 4th, 2023 from the Data World website. The data was uploaded on to the DataProc and into the HDFS. It was ingested through Spark Scala.

The MovieLens dataset was accessed on November 4th, 2023 from the MovieLens website. Additional information was retrieved through unique movie identifiers and scraping. The data was staged onto DataProc before ingested via Spark Scala.

5.2 Data Merging and Cleaning

5.2.1 IMDb Data Merging and Cleaning

IMDb Non-Commercial Datasets consisted of 7 individual datasets, 6 of which were relevant to our study. Due to this our first step was to join these datasets into 1 main IMDb Dataset. After merging the datasets we chose to remove some columns, such as the end year of the movie (most were NA), birth and death year of main actors, character names and some others.

The IMDb Dataset consisted of different media formats such as movies, TV shows, and video games so we decided to only focus on “movies” and “TV movies” in our analysis. After filtering the dataset for these categories, we also decided to transform some column formats, for instance most columns were of “string” format so we transformed the average rating and number of votes columns to numeric format. Additionally, we transformed all the column names to lowercase for simpler use and all the title names were transformed so that if they start with “The” it was moved to the end of the movie name as most movies are known just by their name without the article. We also decided to create another binary column “aboveAverage” to see which and how many movies had an above average rating in our dataset. Finally, we removed any duplicate rows or rows that had NA for our movie title or movie ID. After this, our IMDb dataset was ready to be merged with the other 2 datasets.

5.2.2 Box Office Data Cleaning

The raw Box Office Dataset contained the following columns: rank, title, studio, lifetime_gross, and year. In our specific case, the rank of the movie along with the year that it was released were

not necessary for our research so they were removed using MapReduce code. In addition to this, we checked for missing or repeating values. As there were none we did not need to do any additional cleaning. The remaining columns relevant to our analysis were title, studio, and lifetime gross. At this point our box office data was ready to be merged with the other 2 cleaned datasets.

5.2.3 Small MovieLens Data Merging and Cleaning

The MovieLens dataset consists of 4 separate CSV files that contained pertinent features for our focus, and therefore were merged. With our goal of exploring the industry controllable features for successful economic standards, tertiary features such as ratings and tags were grouped by movie before being aggregated. The data was then merged by common keys such as movieId or title.

Cleaning the data consisted of three parts, data formatting, data manipulation and schema enforcement. The former, consisted of lowercasing all non-ordinal columns (title, genres, content rating, etc.), replacing all commas within string features for easy data handling tasks, and limiting decimal places. Data manipulation consisted of splitting the releaseDate column into separate releaseYear and releaseMonth features since intuitively, these may represent acyclic and cyclic trends, respectively. To enforce schema, missing values were either replaced or dropped depending on total column counts to reduce sparse areas. Column data types were then viewed and explicitly casted to an appropriate type for profiling and analytics. After this, our data was ready for merger with the other 2 datasets.

5.2.4 Final (Joined) Dataset Merging and Cleaning

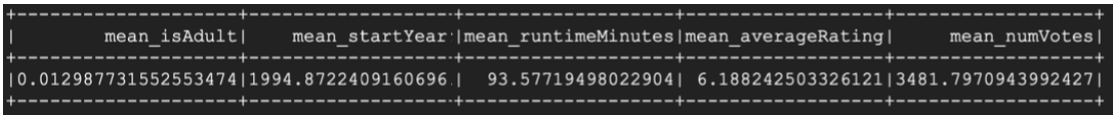
After we cleaned all 3 of our datasets we identified primary keys, and merged them together into 1 main dataset. We were lucky in the first part of the merger as 2 of our datasets, IMDb and Small MovieLens, both had IMDb identifiers for movies so they were simple to merge. However, when joining these 2 datasets with the Box Office Data we had to make sure our title columns were formatted in the same way to ensure that all the corresponding movies would be identified correctly. We did this by making sure that the movie title columns in both datasets were in their original form and in lower case. We used “inner join” so that we only retained movies present in all 3 datasets. This gave us the most information about each movie but also significantly decreased the amount of data we were working with to 6344 rows.

After the successful merger into the main Movie Dataset, we cleaned this dataset by removing all duplicate rows and all the rows that contained NA in movie title or primary key column. Further, we removed some columns that appeared in more than one of our dataset such as movie and title IDs, average ratings, and start years of our movies. After this, our dataset was ready for further analysis.

5.3 Data Profiling

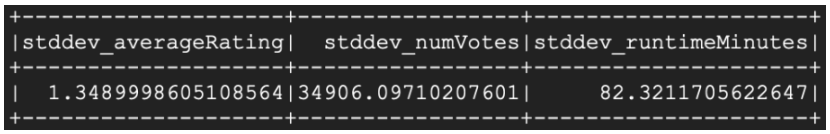
5.3.1 IMDb Data Profiling

The IMDb data features could be separated into 4 categories, unique IDs (movie title and primary name IDs), numerical columns, categorical columns, and string columns (directors and actors names). We decided to calculate relevant statistics such as mean, median, mode and standard deviation for all of our numerical columns. The Numerical columns we were working with were the binary "isAdult" column, which indicated whether a movie was appropriate for children (0) or only adults (1), start year, runtime in minutes, average rating, and number of votes. Figure 2 shows mean values for each of these columns and Figure 3 shows standard deviations only for average rating, number of votes and runtime columns.



mean_isAdult	mean_startYear	mean_runtimeMinutes	mean_averageRating	mean_numVotes
0.012987731552553474	1994.8722409160696	93.57719498022904	6.188242503326121	3481.7970943992427

Figure 2: IMDb Data Mean Values for Numerical Features



stddev_averageRating	stddev_numVotes	stddev_runtimeMinutes
1.3489998605108564	34906.09710207601	82.3211705622647

Figure 3: IMDb Data Standard Deviation Values

In addition to the profiling we did on numerical features we also looked at our categorical features such as the genre column and the job category (just "category" in schema) column. This resulted in an interesting, yet slightly complicated finding. As IMDb allowed for the same movie to be categorized as multiple genres, our genre column had 1322 unique values. This initially seemed like a big issue but in the end we decided to drop this column altogether as the Small MovieLens also had a genre column that only retained the main genre instead of many. Further, while we had 12 (not including null) possible values for our job category column, as shown in Figure 4, it is worth mentioning that we mostly had information about actors, actresses, and directors.

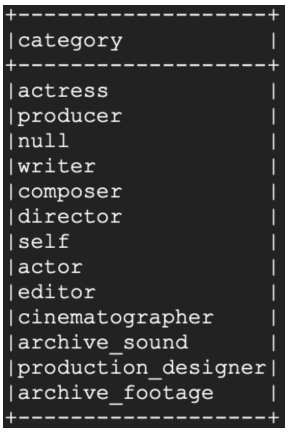


Figure 4: IMDb Data Unique Job Categories

5.3.2 Box Office Data Profiling

Profiling of the Box Office dataset was done using Spark Scala to find relevant statistics such as the mean, mode, median of the lifetime grosses of all the movies. The lifetime gross of a movie is all the revenue it made from the release date to present from any means such as theaters and streaming platforms. Due to this, the lifetime grosses for many of the more popular movies in the dataset are incredibly large, while there are some movies that have inexplicably low earnings, under $100. As a result, the range of lifetime grosses in this dataset is incredibly varied. Out of all the data profiling, the mean was the most significant statistic for our work as it serves as a baseline for understanding what the middle ground of lifetime grosses is, specifically for our dataset.

```
scala> val count=  boxoffice.count
count: Long = 16541

scala> val sumLG=  boxoffice.agg(sum("lifetime_gross")).first.get(0)
sumLG: Any = 2.93793787897E11

scala> val mean = sumLG.asInstanceOf[Double] / count
mean: Double = 1.7761549355963968E7
```

Figure 5: The Mean of the Lifetime Gross

In Figure 5, we can see the mean value for the lifetime_gross column, which was found to be $17,761,549. In other words, the movies in our dataset seem to be making over around $17,500,000 on average.

```
scala> val median = boxofficeD.stat.approxQuantile("lifetime_gross_double", Array(0.5), 0.01)(0)
median: Double = 707343.0

scala>
```

Figure 6: Median

In Figure 6, we can see that the median value for lifetime_gross is significantly lower than the mean, at $707,343. This shows us that our data is highly skewed, as our median is less than 4% of our mean, and that we most likely have many more movies with very high box offices.

```
scala> val mode = boxofficeWithDoubleType.groupBy("lifetime_gross").count().orderBy(desc("count")).selec
t("lifetime_gross").limit(1).collect()(0)(0)
mode: Any = 7500000
```

Figure 7: Mode

The mode of the lifetime gross was also found to be $7,500,000 as seen in Figure 7. The frequency of the mode was incredibly low as expected considering the fact that the likelihood of many movies having the same lifetime gross is low.

5.3.3 Small MovieLens Data Profiling

Initial EDA on the compiled MovieLens data revealed that our observations are categorized into 4 classes of media: movie, TV series, TV episode and music video. The contentRating categorical feature was dominated by those observations labeled R and PG-13.

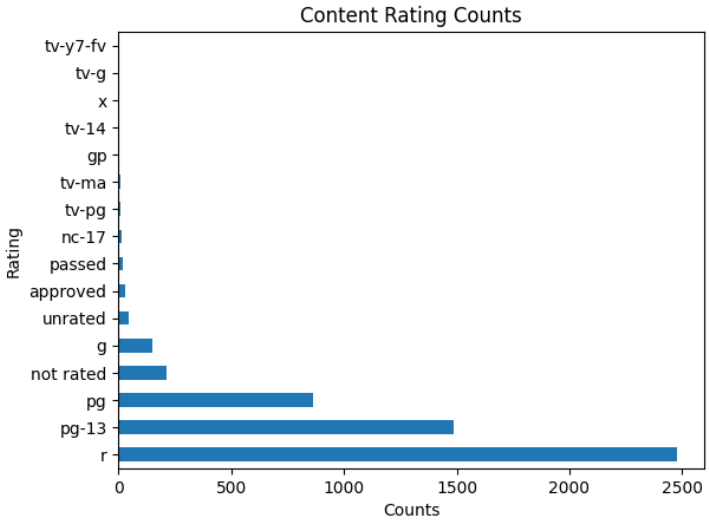


Figure 8: Content Rating Classes Ordered by Descending Number of Occurrences

We find that the mean rating counts of the MovieLens dataset is 10.36, with a median of 3.0. Likewise, we find the average rating (with a scale of 1 to 5) to be 3.26 with a median of 3.41 and a standard error of 0.86. The average number of applied tags per movie is 2.34 with a massive standard error of 5.56, suggesting this feature is not consistent and is extremely sparse.

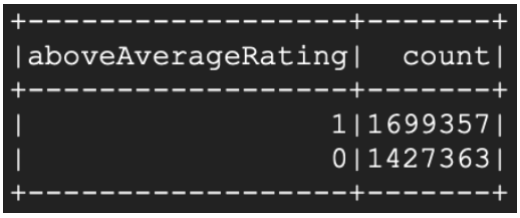
5.3.4 Final (Joined) Dataset Profiling

The profiling of the joined dataset was done almost identically to the profiling of the IMDb dataset. Mean, median, mode, and standard deviation were calculated for numeric columns, and unique values were identified for categorical columns. One thing that is worth mentioning is that while the joint dataset was significantly smaller than our original ones, when comparing it to the IMDb dataset, the mean, median, mode, and standard deviation did not change significantly. Moreover, the number of unique values in the categorical columns were the same. Therefore, we can say that the joined and cleaned dataset remained representative of the initial datasets.

5.4 Data Analysis

5.4.1 IMDb Data Analysis

Our first step in analyzing IMDb data was to understand the features we were working with and their relationships. As mentioned above we created a binary column that was 1 if the movie rating was above average (average over the entire dataset) and 0 if it was below average. We then counted the number of movies that are below and above average as shown in Figure 9.



aboveAverageRating	count
1	1699357
0	1427363

Figure 9: IMDb Data Number of Above and Below Average Ratings

We can notice that we have similar numbers of above and below average ratings which means our movie ratings are balanced. However we can also notice that there are slightly more movies with above than below average rating which could indicate that people are more likely to watch movies they expect to like.

In addition to this we also calculated the correlation between our numerical columns as shown in Figure 10.

```
Correlation between isAdult and averageRating is: -0.054899679375195756
Correlation between numVotes and averageRating is: 0.06108360676232029
Correlation between startYear and averageRating is: 0.016549967873012977
Correlation between runtimeMinutes and averageRating is: -0.0017017614438090277
```

Figure 10: IMDb Data Correlations

We can notice that the correlation is very low which indicates that there are no strong relationships between any of the numerical columns. Further, we attempted to build Regression models on this data using Spark ML, however this resulted in extremely low R^2 (below 0.1) values which further showed us a need to join our IMDb dataset with the other 2 to get more insightful findings.

5.4.2 Box Office Data Analysis

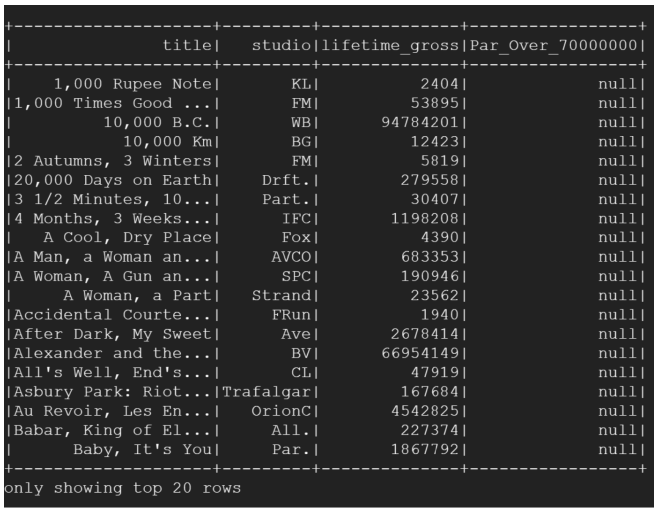
As our dataset contained a very large range of movies, with very different lifetime grosses. we wanted to see how many grossed over $70,000,000. This was chosen as the threshold as we wanted to only retain movies that were in the 90th percentile of the box office. In other words, we only wanted to look at the best 10% of all the movies, according to the box office. Only a few of the studios had any movies that fit this criteria. Specifically, Warner Brothers (referred to as "WB"), Paramount, and Fox. In Figure 11, we can see which of the Warner Brothers movies made it past our threshold (only looking at the partial dataset output). Looking at the figure, we can see that in the first 20 rows there is only 1 movie - 10,000 B.C - that had a box office larger than our threshold. Throughout the entire dataset, Warner Brothers had 169 movies that had box offices above our threshold.

title	studio	lifetime_gross	WB_Over_70000000
1,000 Rupee Note	KL	2404	null
1,000 Times Good ...	FM	53895	null
10,000 B.C.	WB	94784201	10,000 B.C.
10,000 Km	BG	12423	null
2 Autumns, 3 Winters	FM	5819	null
20,000 Days on Earth	Drft.	279558	null
3 1/2 Minutes, 10...	Part.	30407	null
4 Months, 3 Weeks...	IFC	1198208	null
A Cool, Dry Place	Fox	4390	null
A Man, a Woman an...	AVCO	683353	null
A Woman, A Gun an...	SPC	190946	null
A Woman, a Part	Strand	23562	null
Accidental Courte...	FRun	1940	null
After Dark, My Sweet	Ave	2678414	null
Alexander and the...	BV	66954149	null
All's Well, End's...	CL	47919	null
Asbury Park: Riot...	Trafalgar	167684	null
Au Revoir, Les En...	OrionC	4542825	null
Babar, King of El...	All.	227374	null
Baby, It's You	Par.	1867792	null

only showing top 20 rows

Figure 11: Warner Brothers Movies with over $70,000,000 in lifetime gross

In Figure 12, we can see the same output as figure 11, but this time focusing on Paramount movies that have box offices over our threshold. We can see that in the first 20 movies there are no such movies (as all values are null) but over our entire dataset there were 145 Paramount movies that had box offices larger than our threshold.

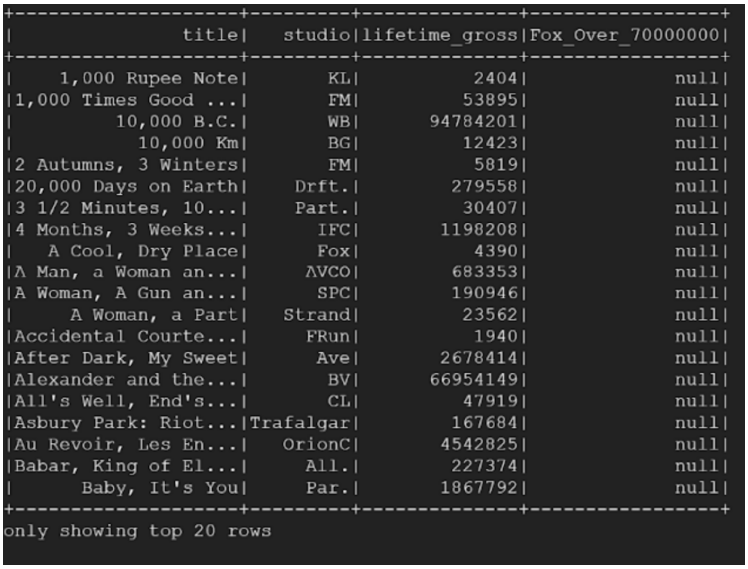


title	studio	lifetime_gross	Par_Over_70000000
1,000 Rupee Note	KL	2404	null
1,000 Times Good ...	FM	53895	null
10,000 B.C.	WB	94784201	null
10,000 Km	BG	12423	null
2 Autumns, 3 Winters	FM	5819	null
20,000 Days on Earth	Drft.	279558	null
3 1/2 Minutes, 10...	Part.	30407	null
4 Months, 3 Weeks...	IFC	1198208	null
A Cool, Dry Place	Fox	4390	null
A Man, a Woman an...	AVCO	683353	null
A Woman, A Gun an...	SPC	190946	null
A Woman, a Part	Strand	23562	null
Accidental Courte...	FRun	1940	null
After Dark, My Sweet	Ave	2678414	null
Alexander and the...	BV	66954149	null
All's Well, End's...	CL	47919	null
Asbury Park: Riot...	Trafalgar	167684	null
Au Revoir, Les En...	OrionC	4542825	null
Babar, King of El...	All.	227374	null
Baby, It's You	Par.	1867792	null

only showing top 20 rows

Figure 12:Paramount Movies with over $70,000,000 in lifetime gross

In Figure 13, we focused on Fox movies that have box offices over our threshold. We can see that in the first 20 movies there are once again no such movies (as all values are null) but over our entire dataset there were 149 Fox movies that had box offices larger than our threshold.



title	studio	lifetime_gross	Fox_Over_70000000
1,000 Rupee Note	KL	2404	null
1,000 Times Good ...	FM	53895	null
10,000 B.C.	WB	94784201	null
10,000 Km	BG	12423	null
2 Autumns, 3 Winters	FM	5819	null
20,000 Days on Earth	Drft.	279558	null
3 1/2 Minutes, 10...	Part.	30407	null
4 Months, 3 Weeks...	IFC	1198208	null
A Cool, Dry Place	Fox	4390	null
A Man, a Woman an...	AVCO	683353	null
A Woman, A Gun an...	SPC	190946	null
A Woman, a Part	Strand	23562	null
Accidental Courte...	FRun	1940	null
After Dark, My Sweet	Ave	2678414	null
Alexander and the...	BV	66954149	null
All's Well, End's...	CL	47919	null
Asbury Park: Riot...	Trafalgar	167684	null
Au Revoir, Les En...	OrionC	4542825	null
Babar, King of El...	All.	227374	null
Baby, It's You	Par.	1867792	null

only showing top 20 rows

Figure 13: Fox movies with over $70,000,000 in lifetime gross

5.4.3 Small MovieLens Data Analysis

Looking at the distribution of the economic gross feature, we see a continuous decreasing exponential distribution between the gross and frequency of occurrence. The form of the distribution is strikingly similar to the standard exponential distribution where the μ is zero and β is one.

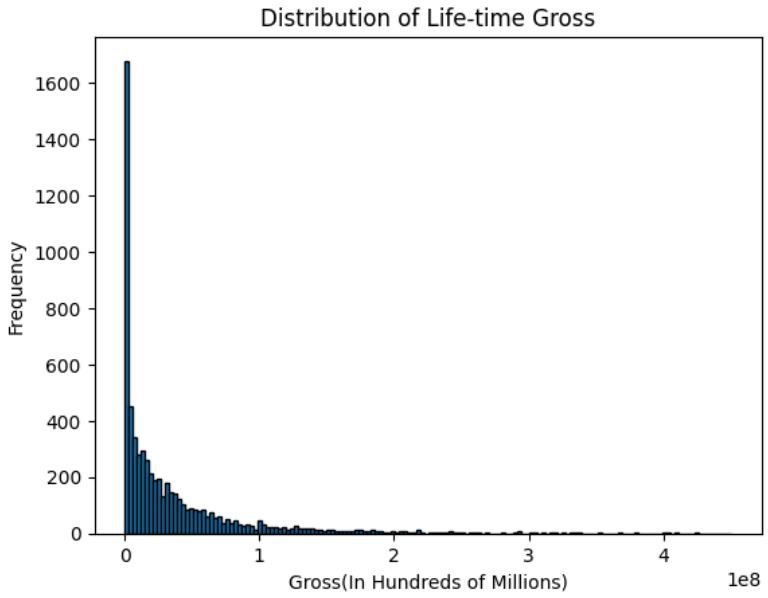


Figure 14: Distribution of Life-time Media Gross

Interestingly, comparing the runtime occurrence to gross shows a mean runtime of approximately 110 minutes of runtime length. Here we can observe that while the majority of observations have a runtime between 60 minutes and 150 minutes, a majority of the successful outliers are scattered upwards of the mean runtime, more so than to the right. Examining the ordinality of any causal relationship between runtime and gross is beyond the focus of this research, however we can take away that there exists a dense range of runtimes that hold a majority of observations labeled as successes.

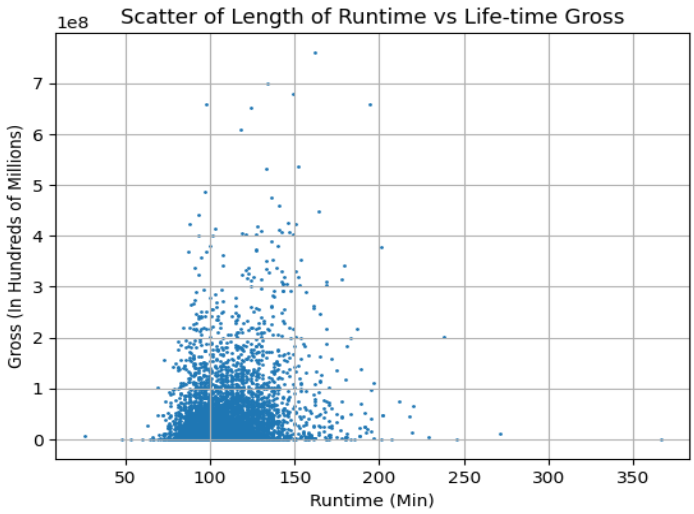


Figure 15: Scatter Plot of Runtime Against Life-time Gross

5.4.4 Final (Joined) Dataset Analysis

While performing preliminary analyses on our individual datasets we realized that having such limited information makes it difficult to come up with comprehensive findings. Due to this, analyses performed on the joined dataset, including correlation calculations and exploring relationships between new features, was instrumental in our overall project. Through our analysis we identified different trends in our data, such as relationships between box office and month of movie release, gender bias in the movie industry, and correlation between the number of votes a movie receives and its box office. After identifying relevant relationships and confirming our findings with the previously mentioned related works, we used Tableau to present our findings.

6. Results

In this section we will go more in depth into the findings we presented in the Related Works section and explore some other interesting trends we found in our dataset.

6.1 Month of Movie Release and Average Movie IMDb Rating

We analyzed the relationship between Month of Movie Release and its Average IMDb Rating score. Our findings showed that there is not a very significant difference in movie ratings depending on the month of its release. In Figure 16 we can see the trend of Average and Median IMDb Rating depending on the release month of the movie. We can see that the average and median ratings fluctuate throughout the year but remain within a narrow range of 6.4 to 7.4. We can also notice that there are slight peaks observed in June and September, suggesting a marginal increase in ratings during these months. However, the overall variation is less than 0.5 points, indicating that the month of release does not greatly affect the movie's reception (as measured by IMDb ratings). This implies that seasonal factors and specific release strategies do not significantly influence average movie ratings.

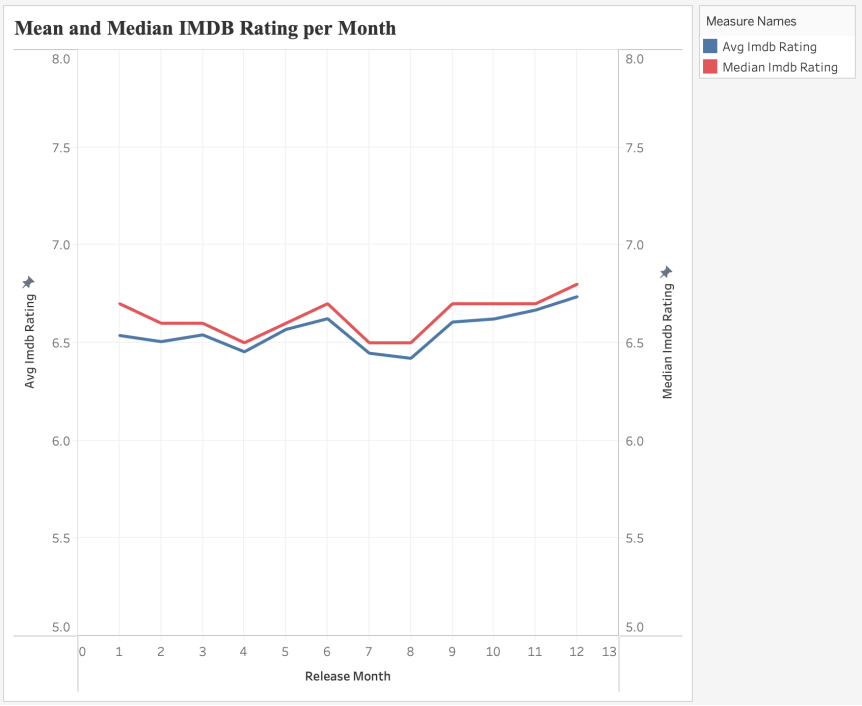


Figure 16: Month of Release and Average IMDb Rating

6.2 Month of Movie Release and Movie Box Office

After not finding a significant relationship between release month and average IMDb rating we proceeded to analyze the relationship between month of release and movies' lifetime box office. This relationship was much more significant than the previous one, with the difference in average box office between January and June being almost 40 million dollars. Looking at Figure 17 we can see that the largest peaks are in summer and winter months. More specifically, the box office is the highest for movies released in May, June, July, and December. This corresponds to the information in our Related Works (Section 3.2). Further, we can notice that there is a steady increase in box office from October to December. Some of the reasons that might be causing a higher box office in summer and winter months is the fact that those are generally the months when blockbusters are released which appeal to wide audiences. Some examples of this are 'Barbie' (the highest grossing movie of 2023), released in July 2023, and 'Spiderman No Way Home' (the highest grossing movie of 2021), released in December 2021. Additionally, these are the times of year when people usually have more free time to spend with their families and go to the movies. Addressing the steady rise in box office from October to December, one reason for this could be the fact that, at that time of the year, award-winning films are released, such as 'Killers of the Flower Moon', released in October 2023, which attract a lot of public attention and with that, movie ticket sales. Furthermore, this is also the time when prestigious movie directors are releasing their films, which also attracts the attention of the public. While there might be other factors, such as economic and social factors that influence the box office and cause the peaks in certain months of the year, it is safe to say that there is a significant relationship between Month of Movie Release and its Lifetime Gross.

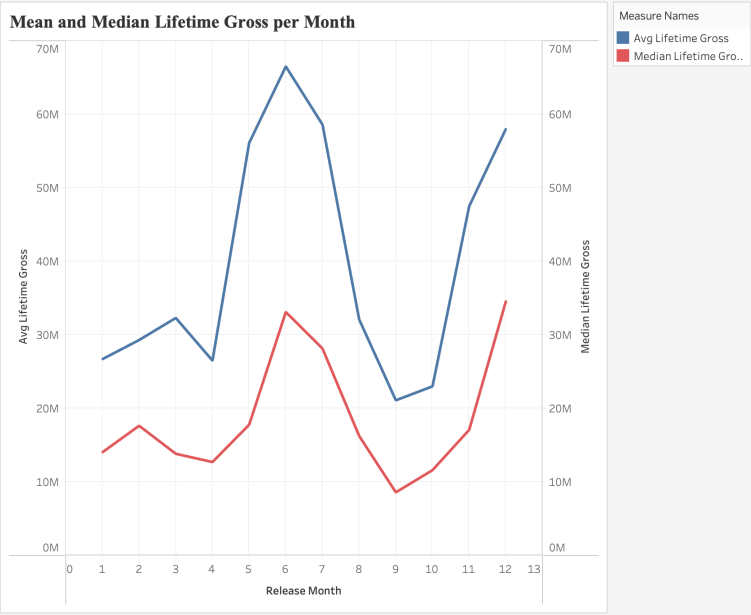


Figure 17: Month of Release and Lifetime Gross

6.3 Lifetime Gross per Studio

Next, we looked at the data about film studios and their corresponding Lifetime grosses. Our findings showed that a handful of studios held most of the lifetime gross of the industry. This was corroborated by our references in the Related Works (Section 3.1). Looking at Figure 18 we can see that 6 studios (Sony, Paramount, Fox, Universal, Warner Brothers, and Buena Vista (now Disney)) tend to produce movies with the highest lifetime gross, almost appearing as an oligarchy in Hollywood. This could result from the fact that these were some of the first studios in the film industry during the "Golden Age of Hollywood" in the 1910s and 1920s. In addition to being the first in the industry, just the fact that they have been active and successful for more than 100 years adds to their popularity which in turn results in higher public interest in the movies that are released by these studios.

It is worth noting that, as our data has a cutoff of 2018, some of these lifetime grosses will be even higher due to mergers and acquisitions that have happened since.

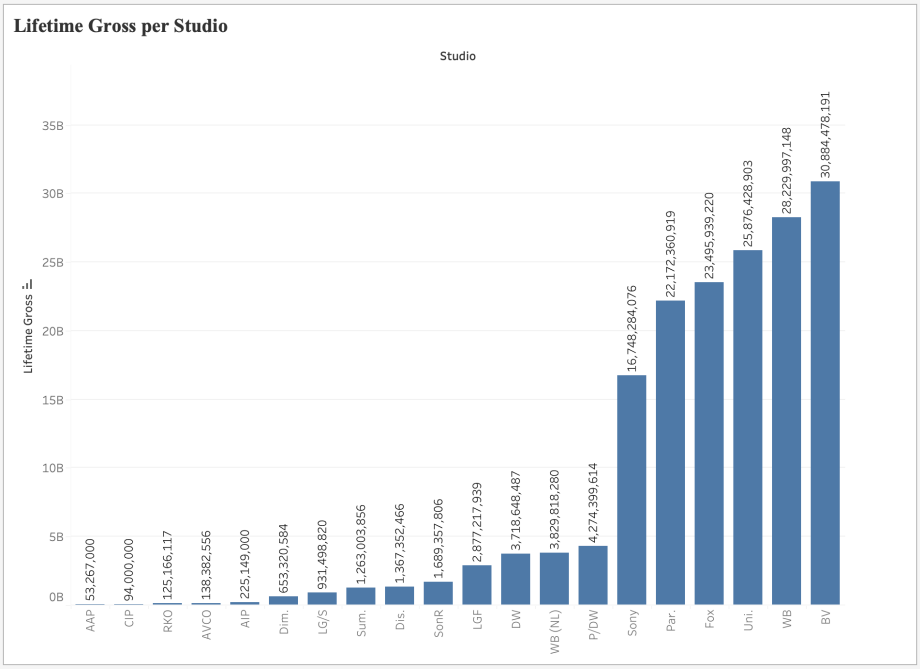


Figure 18: Lifetime Gross per Studio

6.4 Analyzing the Impact of Actors and Actresses on Lifetime Gross and Rating

In our final dataset, each movie was associated with a notable person mentioned in the "primary name" column. Further, we have a corresponding column, "category," which indicates the occupation of each of the individuals in the primary name column, most of which were actors / actresses. We wanted to further analyze how the gender of the primary person is related to a movie's lifetime gross and average rating. Due to this, using Scala, we created 2 separate tables - one with all the actors and their corresponding lifetime gross and average rating, and one for actresses with their corresponding lifetime gross and average rating.

After creating these tables and calculating average lifetime grosses and ratings for each of the actors and actresses, we ordered them in descending order and only retained the first 10 individuals.

In the following Figures, 19 - 22, we can see these relationships between actors and actresses and their corresponding average lifetime gross of movies and average ratings.

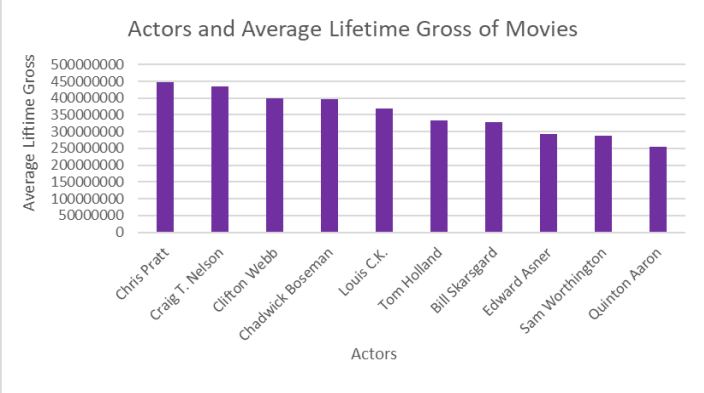


Figure 19: Actors and their Average Lifetime Grosses

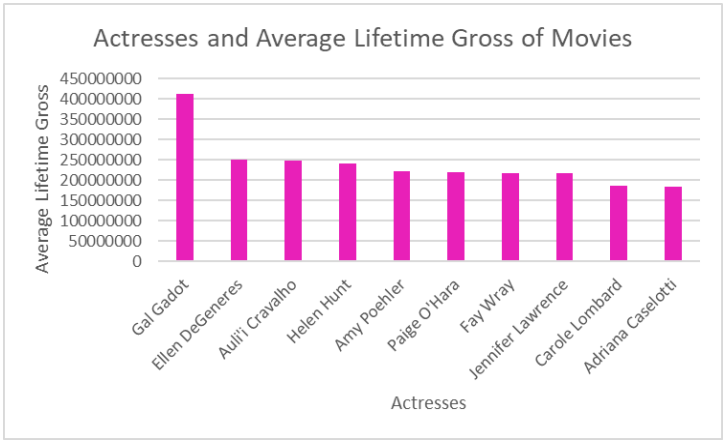


Figure 20: Actresses and their Average Lifetime Grosses

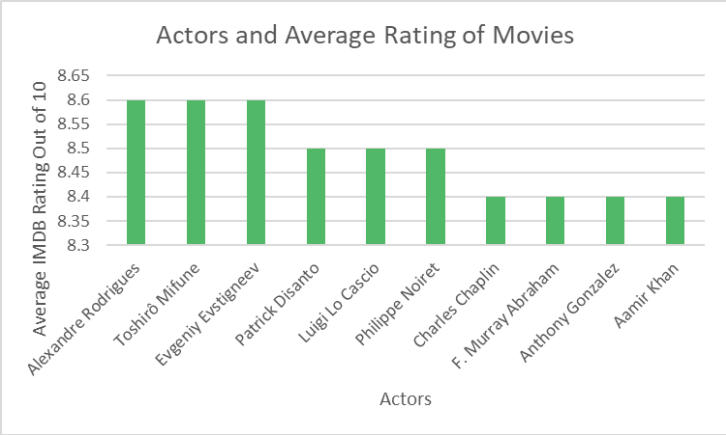


Figure 21: Actors and their Average Movie Ratings

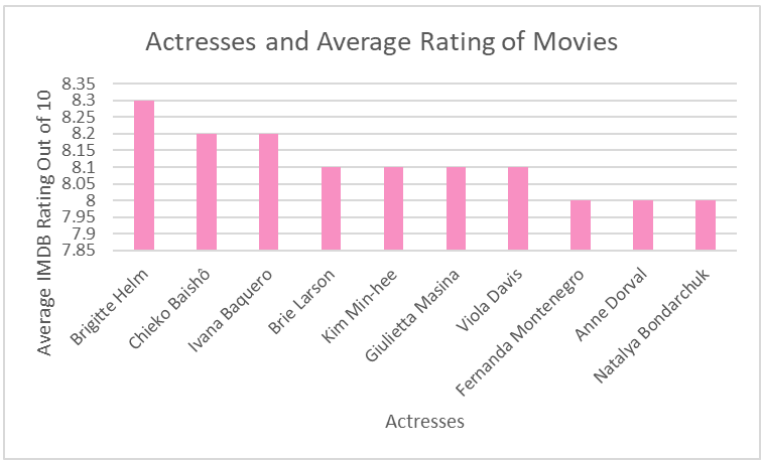


Figure 22: Actresses and their Average Movie Ratings

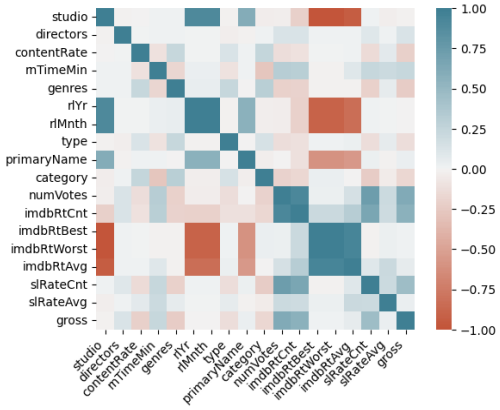
We can see that the top actor with the highest grossing movies was Chris Pratt, while the actress with the highest grossing movies was Gal Gadot. This is notable as both actors mentioned are part of large movie franchises with Chris Pratt playing the superhero Star Lord for the Marvel franchise and Gal Gadot playing Wonder Woman for DC Extended Universe (DCEU). In addition to Chris Pratt, we also see the actors Chadwick Boseman and Tom Holland, who are also part of the Marvel franchise and both starred in 'Avengers: Endgame', which happens to be the highest grossing movie of all time. In addition to previously mentioned criticism about highest grossing movies being made by the same studios, there is also a lot of controversy about superhero genre movies being "automatic blockbusters". These movies tend to make millions in the box office on their opening nights, while movies from other genres seem to fail to get the same level of recognition. According to Sarah E. Joseph, movies being a part of a franchise, such as the MCU (Marvel Cinematic Universe) or DCEU, is the most important factor in predicting

movie success[5]. Her study claimed that movies that are already a part of a huge franchise have a higher likelihood of achieving high box office revenue while other movies tend to flounder in comparison.

According to Wühr et al., content analyses showed that male protagonists dominate in film genres that are preferred by both men and women, whereas female protagonists are more likely to be found in film genres that are preferred solely by women[6]. In our data, the "primary name" referred to the most prominent person associated with the movie, which in most cases is the lead actor. Interestingly, we can see that movies with actresses in the lead role rather than actors tended to make less lifetime gross as well as resulted in lower average rating. In the book *Invisible Women*, Perez touches on this particular gender bias by mentioning how women are more willing to accept men as role models, but men won't do the same for women[7]. This means that movies fronted by men would be able to appeal to wider audiences rather than movies fronted by women which could be a possible explanation for higher lifetime grosses and ratings seen for movies associated with actors.

6.5 Significance between User Voting Data and Economic Standard of Gross

After normalizing the appropriate features, one-hot encoding non-sequential categorical columns and label encoding natural sequential columns, we conducted a pairwise correlation test using Pearson's method. Intuitively, we can observe that the features that record the number of user votes per database hold a higher relative correlation to the economic standard of gross profit. Similarly, when applying a ranking order to the correlations, we see the same result to corroborate the observation.



[5] Joseph, Sarah E. "What Makes a Movie Successful: Using Analytics to Study Box Office Hits." TRACE: Tennessee Research and Creative Exchange, 2019.
[6] Wühr, Peter, et al. "Tears or Fears? Comparing Gender Stereotypes about Movie Preferences to Actual Preferences." Frontiers in Psychology, March 24, 2017.
[7] Perez, Caroline Criado. *INVISIBLE WOMEN: Data bias in a world designed for men*. New York: Abrams Press, 2021.

Figure 23: Heatmap of Pairwise Pearson's Correlation between Features

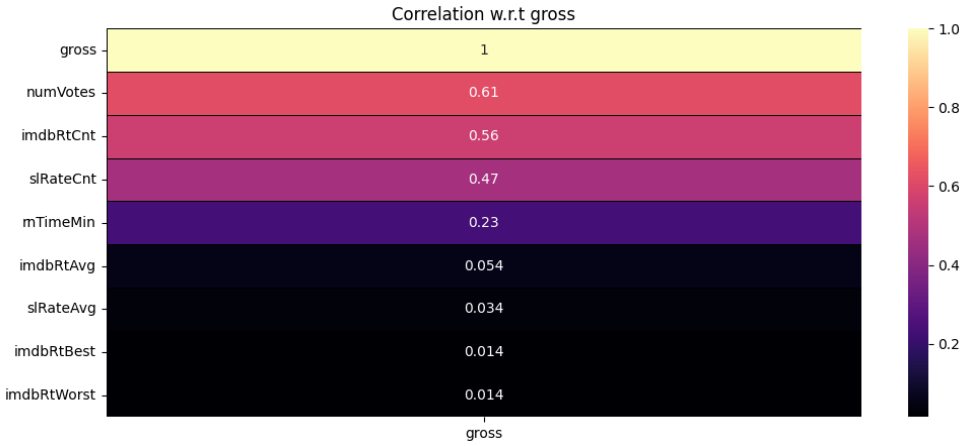


Figure 24: Lifetime Gross per Studio (Note the highest three correlations to gross)

To provide a broader affirmation of this, a coefficient ranking was conducted along with a feature importance test from a linear ridge and random forest regressor model respectively. Ranked correlations provide a view into the relationships when the data is viewed in an observational study paradigm, however we are far more interested in the industry controllable features that are relevant to media success. The coefficient ranking assumes that the underlying relationship of the data is linear.

The random forest regressor instance, using a squared error criterion, ranks feature importance by impurity. Specifically, the importance of a feature is a normalized reduction of the criterion, otherwise known as the Gini importance;

$$Imp(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t) = X_m} p(t) \Delta i(s_t, t)$$

Figure 25: Gini Importance Estimand Formula

10	numVotes	0.55958
11	imdbRtCnt	0.09536
14	imdbRtAvg	0.07820
3	rnTimeMin	0.07264
15	slRateCnt	0.05359
12	imdbRtBest	0.02809
13	imdbRtWorst	0.02809
0	studio	0.02809
16	slRateAvg	0.02210
8	primaryName	0.00113
7	type	0.00030
2	contentRate	0.00020
5	rlYr	0.00016
6	rlMnth	0.00016
1	directors	0.00004
4	genres	0.00001
9	category	0.00000

10	numVotes	0.68633
2	contentRate	0.09916
16	slRateAvg	0.07211
11	imdbRtCnt	0.02900
15	slRateCnt	0.02669
9	category	0.02640
14	imdbRtAvg	0.02123
3	rnTimeMin	0.01519
1	directors	0.01248
4	genres	0.00458
8	primaryName	0.00264
6	rlMnth	0.00193
5	rlYr	0.00189
7	type	0.00035
13	imdbRtWorst	0.00001
0	studio	0.00001
12	imdbRtBest	0.00000

Figures 26 and 27: Ranked Linear Ridge Coefficients & RFR Feature Importances

We can observe that not only did the total number of rating counts in each dataset hold the highest three correlations with the economic standard of interest, from the coefficient and feature importance rankings above, we see that the total number of rating counts for each dataset consistently appeared in the top five most important features. Each of these different estimations of feature significance rely on different assumptions, target estimands and experimental paradigms. Together, the consistent results indicate a corroborating suggestion that it is the user engagement (the action of rating) that holds the most sway over economic success of the media.

This is further supported by Wasserman et al., where the strongest correlation, post bias reduction and double log comparison, was the total number of user votes[8]. Their findings suggest an adjacent conclusion that the total number of user votes is an indicator of media notability, which in turn impacts the probability of media success.

Dissanayake and Vidanagama provide a similar conclusion through a different set of features wherein user engagement data was collected through Facebook[9]. Their feature selection methods result in similar discoveries to our own findings, with the caveat that it is publicly viewable user and cast generated engagement that have the highest significant impact to the early predictions of film success.

[8] Wasserman, M., S. Mukherjee, K. Scott, X. H. T. Zeng, F. Radicchi, and L. A. N. Amaral. "Correlations between …."
[9] Dissanayake, D. M. L., and V. G. T. N. Vidanagama. "Early Prediction of Movie Success Using Machine Learning Models."

7. Conclusion

Navigating the crossroads of big data and cinema, our research casts a new light on the age-old question of what truly makes a movie successful, both in the eyes of critics and at the box office. Our findings highlight the fact that one of the biggest determinants of movie success is what studio produces it. We found that 6 studios, Sony, Paramount, Fox, Universal, Warner Brothers, and Buena Vista (now Disney), are a de facto oligarchy in Hollywood as they hold over 70% of all the box office earnings compared to all other studios. Another factor that strongly influences box office is the month the movie is released in. More specifically, movies released in summer and winter months have significantly higher box offices on average, most likely due to the fact they are blockbusters and award contenders. We also found strong evidence that user engagement and generated data (such as reviews) hold one of the highest, if not the highest, impact towards a film's revenue. Our research indicated that the highest correlations were between gross and total number of user votes the movie had. In addition to determinants of movie success, we found a strong gender bias that showed that movies with men in lead roles tend to make more in box office and also be rated higher. From these conclusions, recommendations can be made to streaming platforms when deciding which movies to stream for their specific audiences. As well as this, they can help producers decide which actors to cast and which films to take on. These findings can also be helpful to screenwriters and authors in deciding which studios to sell their work to. Lastly, they can be used by viewers in order to decide which movies to watch based on their potential ratings and casts. Possible avenues for expanding upon our research could be to look into more specific demographics of the audience such as their age, gender, race, etc. to be able to understand what specific factors appeal to them and thus make better tailored recommendations. Concerning possible further exploration of features related to film profitability, perhaps more features could be identified and gathered that are intuitively an indicator of user engagement, such as the number of Reddit threads, advertisements, promotional budgets, scaled sentiment analysis of tweets, etc. Furthermore, as past research has supported the notion that user generated engagement is impactful, further research could be conducted on differentiating the types of effective engagement.

Acknowledgements

We would like to thank NYU High Performance Computing for their support. We would also like to thank Tableau for providing students with a free year of their platform access. We would like to thank IMDb for giving us access to their movie data information. We would like to thank the GroupLens Research Team at the University of Minnesota for having collected and made available subsets of data from MovieLens. We would like to thank Hemant Malik for creating and maintaining the PyMovieDb Module, a wrapper representing the IMDb API that allowed us to obtain additional information via scraping.

References

Cucco, Marco. “The Promise Is Great: The Blockbuster and the Hollywood Economy.” Media, Culture & Society 31, no. 2 (2009): 215-230. https://doi.org/10.1177/0163443708100315.

Dissanayake, D. M. L., and V. G. T. N. Vidanagama. "Early Prediction of Movie Success Using Machine Learning Models." International Journal of Computer Applications 183, no. 44 (2021). https://www.ijcaonline.org/archives/volume183/number44/dissanayake-2021-ijca-921847.pdf.

Gunter, Barrie. “Is Studio Size Important to Box Office Success?” In Predicting Movie Success at the Box Office, 35-49. Cham: Palgrave Macmillan, 2018. https://doi.org/10.1007/978-3-319-71803-3_3.

Joseph, Sarah E. "What Makes a Movie Successful: Using Analytics to Study Box Office Hits." TRACE: Tennessee Research and Creative Exchange, 2019. https://trace.tennessee.edu/cgi/viewcontent.cgi?article=3282&context=utk_chanhonoproj.

Perez, Caroline Criado. *INVISIBLE WOMEN: Data bias in a world designed for men*. New York: Abrams Press, 2021.

Wasserman, M., S. Mukherjee, K. Scott, X. H. T. Zeng, F. Radicchi, and L. A. N. Amaral. "Correlations between User Voting Data, Budget, and Box Office for Films in the Internet Movie Database." Journal of the Association for Information Science and Technology 66, no. 4 (2015): 858-868. https://doi.org/10.1002/asi.23213.

Wühr, Peter, et al. “Tears or Fears? Comparing Gender Stereotypes about Movie Preferences to Actual Preferences.” Frontiers in Psychology, March 24, 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5364821/.

Datasets

Elias Dabbas. “Boxofficemojo Alltime Domestic Data - Dataset by Eliasdabbas.” Data.World, August 4, 2019. https://data.world/eliasdabbas/boxofficemojo-alltime-domestic-data.

“Datasets.” IMDb Datasets. Accessed December 14, 2023. https://datasets.imdbws.com/.

“Movielens.” GroupLens. Last modified December 8, 2021. https://grouplens.org/datasets/movielens/.