Movie Ratings: Do They Matter?

August 3, 2021

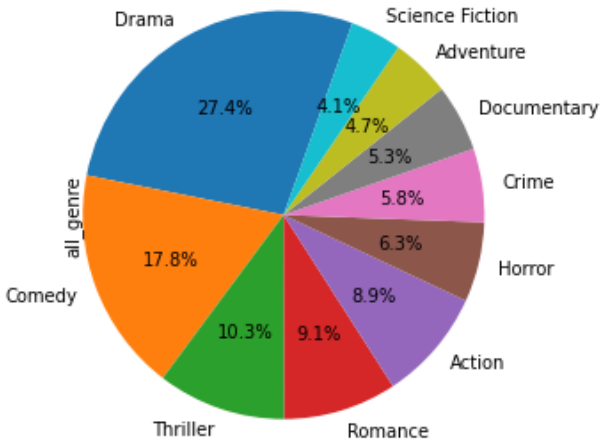
Greg Rosen & Aruna Bisht

I. Introduction

Movies are an integral part of our life for entertainment. People from all walks of life pick and watch movies based upon various parameters including genre, imdb ratings, marketing, popularity, production company, and actors. For context, global movie box office sales totaled $101B in 2019 (Escandon, The Film Industry Made A Record-Breaking $100 Billion Last Year).

While picking movies, consumers must face the decision of whether the movie is worth purchasing and watching before actually seeing it. To do so, many rely on external data to help inform this decision, such as exposure to previews via marketing, media gossip, and critic reviews and ratings or genre available.

As an example of the options available to consumers, below is a breakdown of percentages of movie genres in the dataset we will be discussing below. This breakdown is representative of the true breakdown of global movies by genre. Dramas are most popular, contributing 27.4% of all movies, followed by comedies, thrillers, romance, action, and others.



When it comes to marketing versus reviews, it may seem an easy assumption that movies with the largest budgets, best actors, and most experienced production companies generally receive higher ratings than lower-budget movies with less experienced actors and production companies. But nearly all moviegoers have experienced the emotional rollercoaster of anticipating an exciting movie only to walk out of the movie theater disappointed at the overall caliber of the movie.

II. Dataset

Seeking to answer our research question, we found the MovieLens dataset from Kaggle. It's a nearly comprehensive dataset of 40,000+ rows of US and international movies back from 1899 to until date. It includes key columns of genre, imdb rating, budget, popularity score, release date, production companies, corresponding countries etc. Some data like production companies, genre, and cast are in string json format inside the CSV, so we had to parse the data as part of data cleansing.

Key columns include: Budget, Vote Average (rating), Popularity, Actor, Production Company, Actor/Production Company Experience

III. Hypothesis / Research Question

We set out to answer the question of whether popular movies with higher budgets and more experienced starring actors and production companies actually receive higher reviews than lower budget movies with less experienced stars and production companies. Does the hype of a highly anticipated blockbuster really stack up against the critics' votes?

In our project, we analyze these assumptions and attempt to find patterns in our data if this actually holds true or false.

We plan to analyze actors and production companies' relationship with budget, popularity, and ratings. Our analytical method will include the use of boxplots, bar charts, and linear regression trendlines for trend analysis.

IV. Cleaning and Assumptions:

The original dataset had 45,466 rows and 24 columns. We manipulated the dataset by missing values, messy data, duplicate data, unclear zero values, by time, and by json formatted structure. Missing values, messy data, duplicate data, and unclear zero values in key columns were dropped to ensure we are only analyzing valid data points. Our strategy by column is listed below:

Dropped missing values and messy data for:

- Budget
- Popularity
- Rating
- Genre

Dropped unclear zero values for:

- Popularity
- Rating

Dropped Specific Years:

> The dataset contains movies with budgets dating back to 1890 onwards. We adjusted the *budget* column for inflation using the python CPI library to adjust the budget. However, we couldn't find the CPI indexes of the year previous to 1910, so we removed all movie records with release datees prior to 1910. Because less movies were made in these early years, this did not exclude a large portion of the dataset.

JSON Structures:

> The actor, production company, and genre columns were formatted as json strings within the CSV. We parsed out the specific starring actor, production company, and genre from each json into new columns.

Blockbuster movie definition:

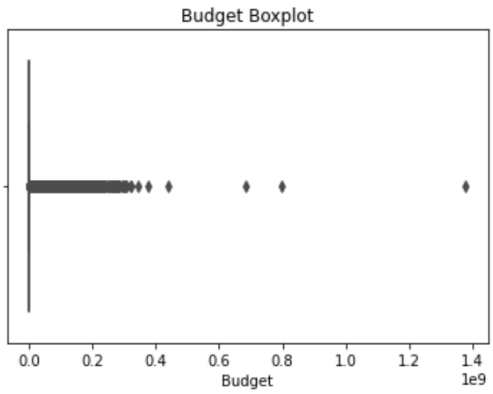
> We made the assumption that blockbuster movies are defined as movies that have highly experienced actors and/or highly experienced production companies. Experience in this sense means higher counts of movies affiliated with that actor or production company.

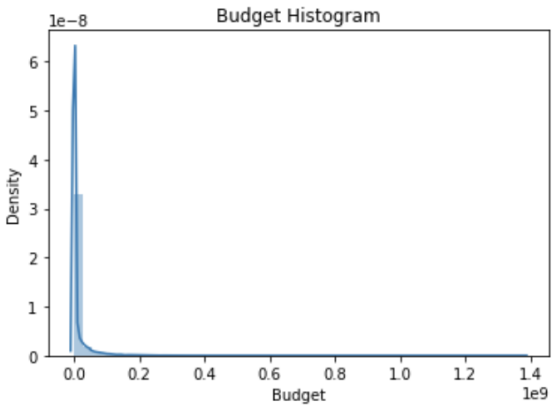
Our final dataset for analysis after manipulations contains ~32,000 rows.

V. Exploratory Questions:

We set out to analyze questions related to budget, ratings, and popularity. These questions will help to identify trends that can inform our hypotheses later on.

1. **What does the overall Budget look like in the dataset?**



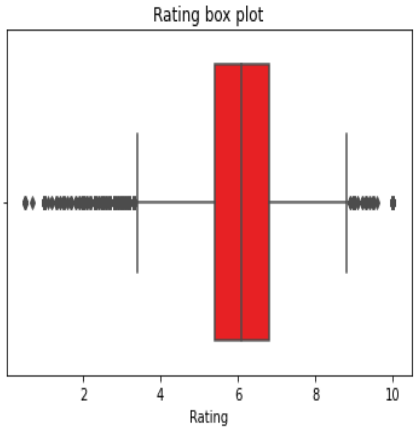


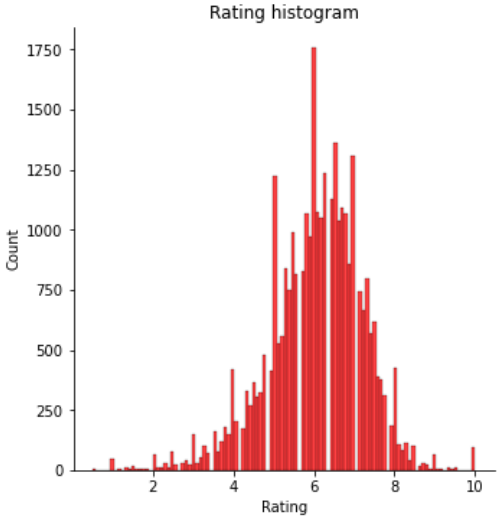
Budget boxplot: This graph shows that the majority of movies have a low budget and only a few movies have a high budget. There are ~24k budgets with 0 value (considering these low-budget movies). However, due to the high amount of 0s we will likely focus on actor and production company experience levels as a more reliable blockbuster proxy.

Budget histogram: This graph tells the count of movies, showing again that most have a low budget. This is mostly in the range of 500-1000 (to be explicit since the graph is in natural log scale).

Overall, budgets skew heavily right with most movies having very low budgets, and only a select few having high budgets.

2. **How do the overall ratings of those budget movies look in the dataset ?**



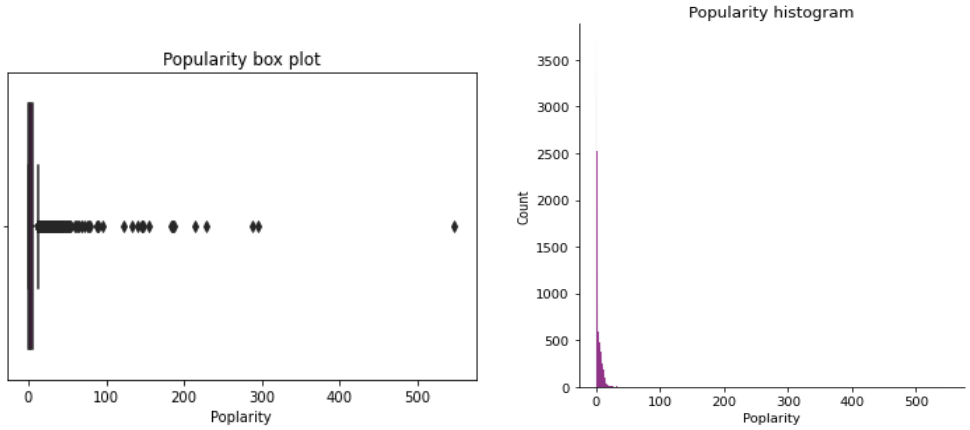


Rating box plot: This graph shows that the majority of movies have an average of 5.5 rating.

Rating Histogram: This is essentially the same thing but a different view of how the ratings are spread out across the dataset in the form of histogram. The count of movies is between 1250 to 1750 for movies ratings from 5.5 to 6.5

We can see from the above that ratings are rather normally distributed throughout the movies dataset, unlike budget which is skewed heavily to the right.

3. **What does the overall popularity look like in the dataset ?**



Popularity box: This graph shows that the majority of the movies are not popular and a very few movies are popular. There could be many reasons for this, maybe due to the difference in publicity movies got.

Popularity histogram: This graph tells the same story. Majority of unpopular movies count from 1000 - 2500.

Popularity skews heavily to the right quite similarly to budget. This may be the first semblance of possibility that budgets and popularity have more in common than budgets and ratings, but more analysis is needed.

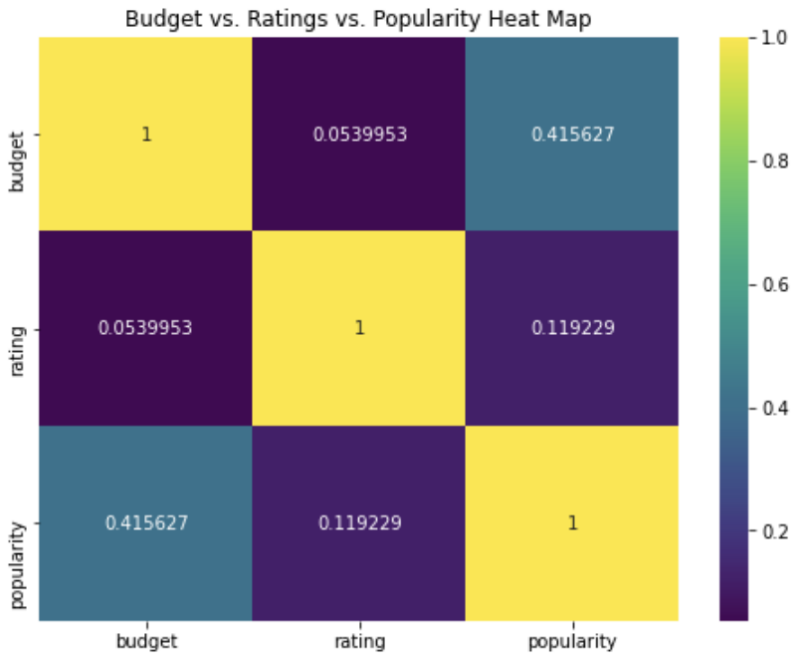
In our next analysis, we will explore the correlations between the variables explored above.

.

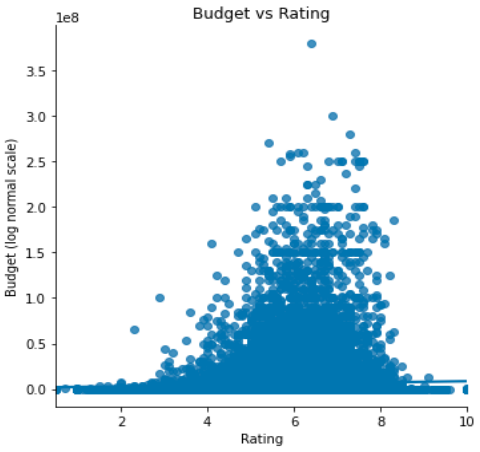
VI. Analysis

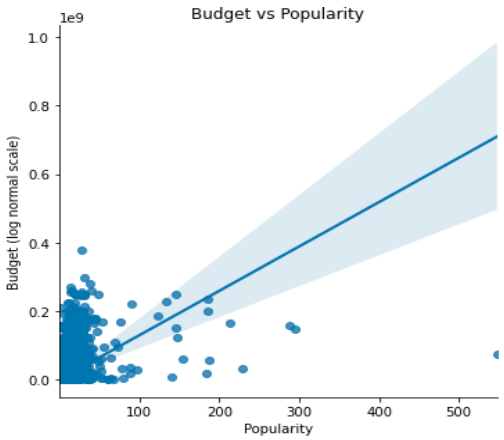
Our analysis focuses on the relationships between key variables that could have some effect on the relationship between blockbusters, ratings, popularity, experience levels, actors, and production companies.

1. **Budgets vs. ratings vs. popularity correlations**



Budget heat map: This heap map quantifies the correlations between the variables we explored before. We see popularity vs budget has a correlation of 0.31 , which is higher than in comparison to budget vs rating which is (0.12) and similarly popularity vs rating (0.18).



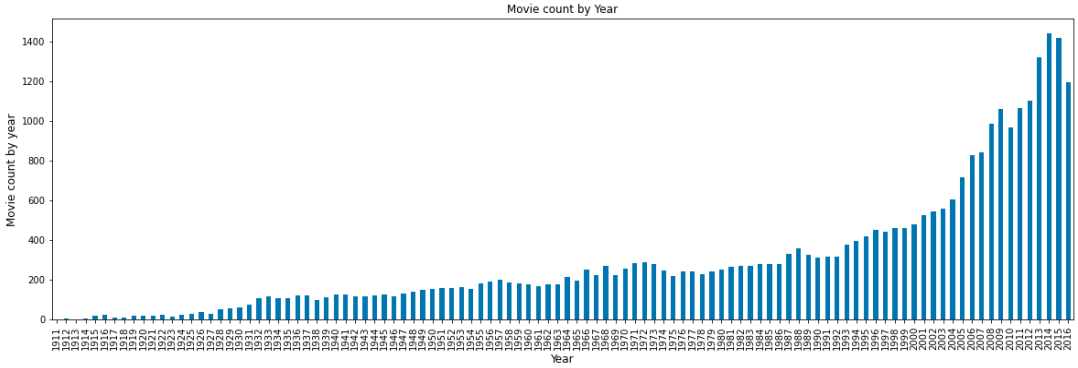


Budget vs Rating: This graph is on a natural log scale. It shows that ratings don't have a strong correlation with Budget and they are spread throughout. Most of the ratings are clustered in the range of 5.5 to 6.5 ratings.

Budget vs Popularity: This graph is also on a natural log scale. It tells that popularity has a stronger correlation with the budget. As popularity increases, the budget also increases, although most data points cluster into the 100 popularity or less with a few strong outliers.

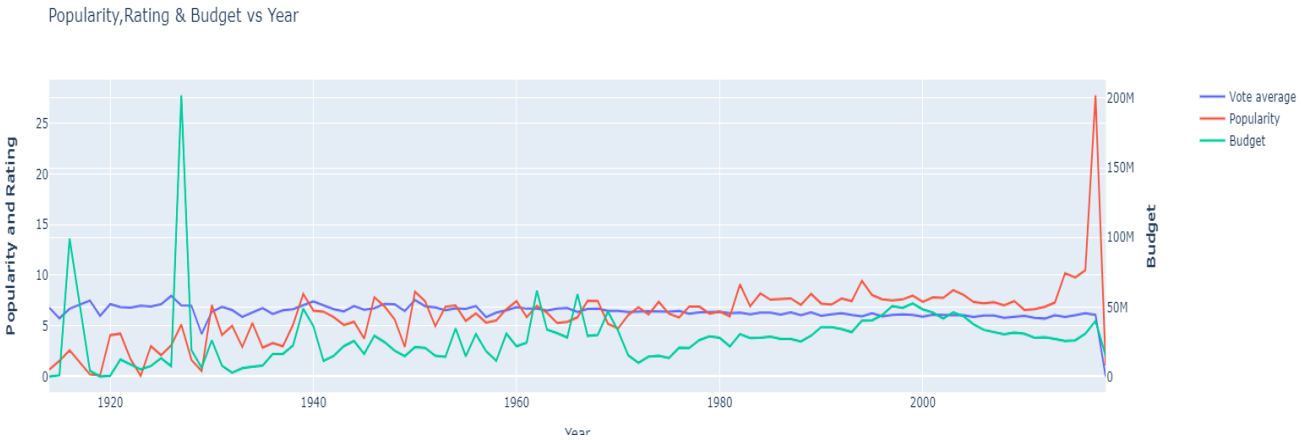
Again, there seems to be a stronger trend between budget and popularity than budget and ratings.

2. What is the frequency of those movies over the period of time?



Movie count by year: This graph shows the dynamics of these variables over time. We observed an increasing movie production trend with time -- as time moves on, more movies are made per year.

3. How do popularity, rating and budget change over time ?

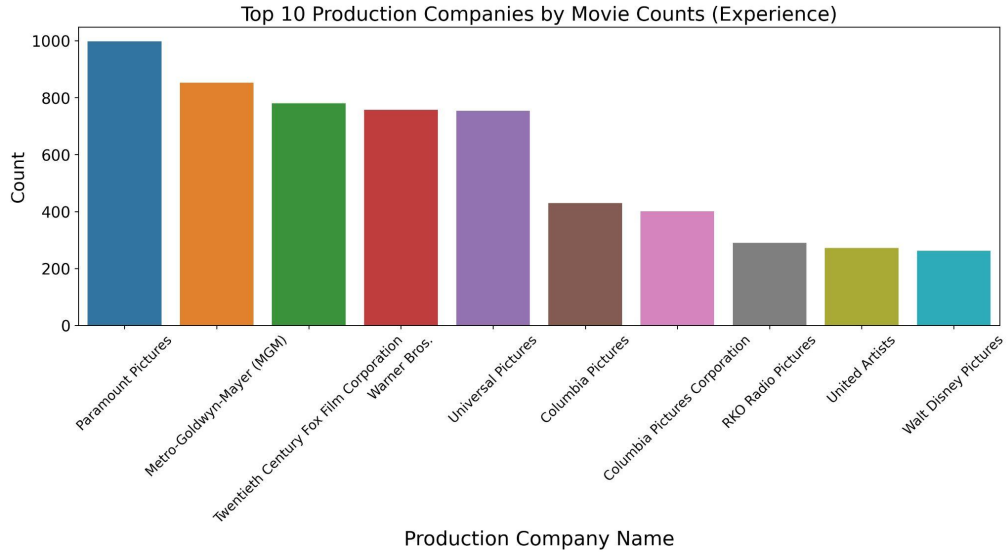


Popularity, rating & budget year: Over time, ratings stay relatively constant around an average of about 6. Yet, both popularity and budget have increased over time alongside

one another. This finding confirms our initial thoughts that budgets and popularity are interrelated while ratings are not as connected to a movie's budget or popularity.

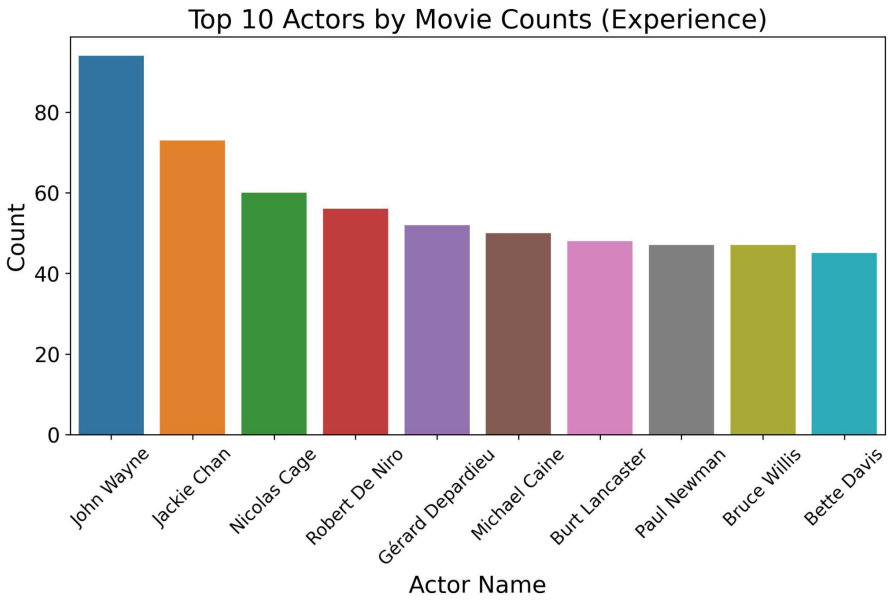
Since our time analyses are showing similar trends to overall analysis above in terms of relationships between budget, ratings, and popularity, we will likely not focus on the time variable and instead focus on the overall trends in the dataset across time.

4. What are the top production companies making up blockbuster movies?



See Number 5 below.

5. What are the top actors making up blockbuster movies?



Now we jump into the concept of blockbusters: movies with highly experienced actors andor highly experienced production companies. We can see from both questions 4 and 5 that the top actors and production companies listed are household names spanning various time periods (e.g. John Wayne and Robert De Niro). The most experienced actor has been in over 90 movies while the most experienced production company has produced nearly 1000 movies. This makes sense since production companies can produce multiple movies at once, while an actor only has the time allotted to one person.

These actors often star in the most famous of blockbusters, and these production companies produce them -- both of which we will be analyzing further in the hypothesis section.

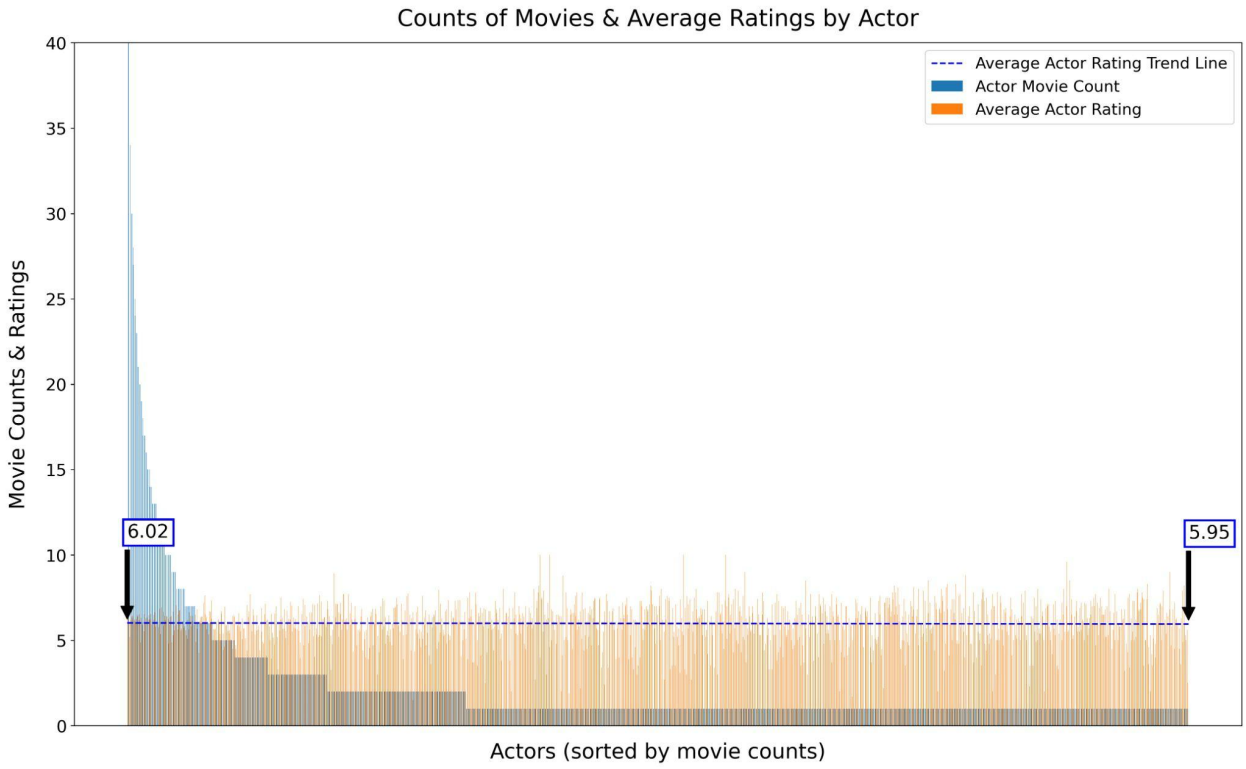
VII. Hypotheses

Based on the above findings and relationships between budgets, ratings and popularity (namely, the lack of relation between popularity and ratings with respect to blockbusters and non-blockbusters), we have four overarching hypotheses to test based on the above with the goal of comparing ratings versus popularity:

1. Experienced actors generate higher movie ratings on average
2. Experienced actors generate higher movie popularity on average
3. Experienced production companies generate higher ratings
4. Experienced production companies generate higher popularity on average

Because of the much lower frequency of movies in earlier years, we decided to ignore trends across time in favor of general movie trends regardless of time.

1. Experienced actors generate higher movie ratings on average



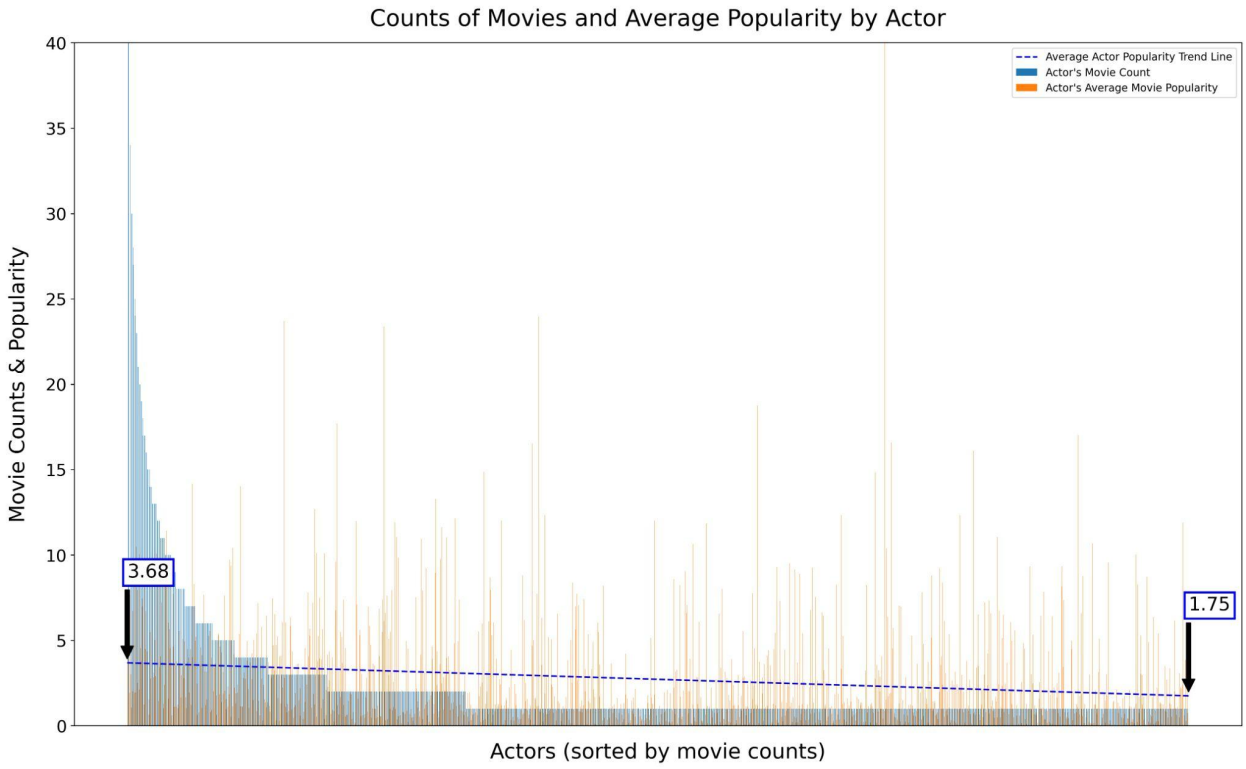
In the chart above, each blue line represents an actor, sorted by the amount of movies they have appeared in (proxy for actor's experience level). We can see that counts of movies diminish from the most experienced actors on the left to least experienced on the right for the entire sample of a few thousand actors. Alongside these blue lines are the orange lines, representing each actor's average movie rating for all movies they've appeared in.

Our hypothesis is that experienced actors generate higher ratings. If this holds true, we would expect to see similarly downward-trending movie ratings from left to right as the sorted actors diminish in experience level.

The bright blue line stretching across the graph represents a linear regression trend line of the orange average movie ratings for each actor. The minimum and maximum movie ratings of the trend line are explicitly labeled.

One can see that movie ratings do have a downward trend as actors' experience levels diminish. However, the trend is quite small -- it is nearly a straight line. The minimum to maximum average rating only has a difference of 0.07. This leads us to conclude that actor experience level may have some factor in a movie's rating but it is not necessarily a strong effect.

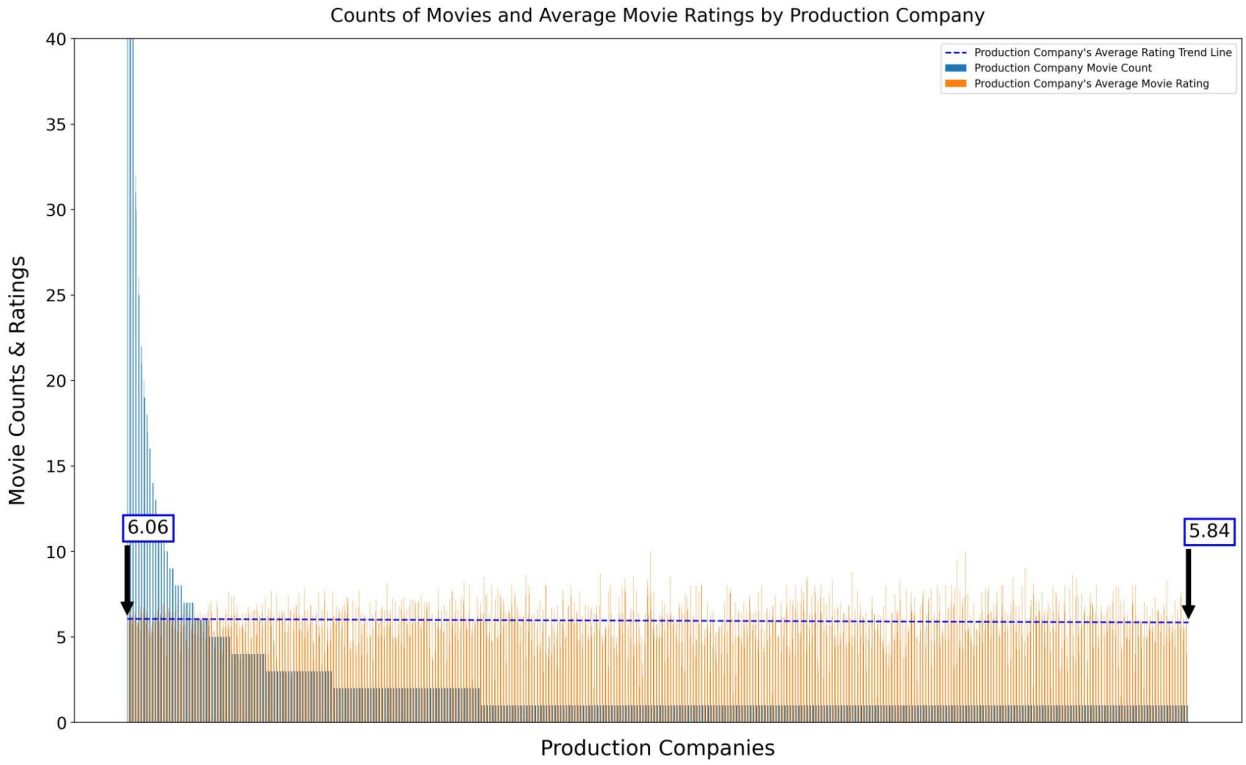
2. Experienced actors generate higher movie popularity on average



We then performed a similar test to the above on an actors' average movie popularity. Contrary to the small change in ratings, we see here a more recognizably downward-sloping trend line that has a max-min difference of 1.93 popularity points.

Though popularity has notably more variance based on the high fluctuations in orange lines, the slope is still clearly trending downwards.

3. Experienced production companies generate higher ratings

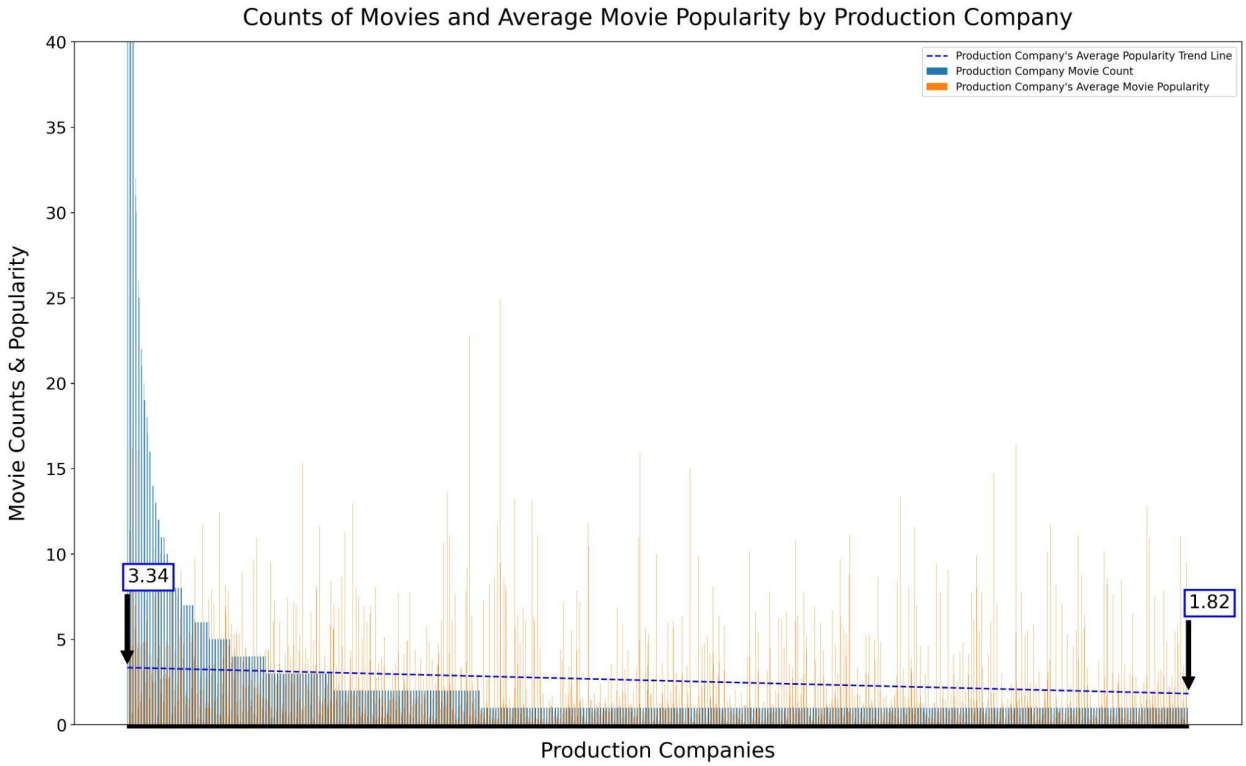


After comparing the differences in ratings and popularity by actor experience, we decided to compare the same dynamics of popularity and ratings but this time overlaid with production company experience instead of actor experience. The additional analysis will provide insight into

whether this trend of ratings versus popularity holds for more than just the starring actor, and can be expanded to other major factors affecting a movie's production.

In the above chart comparing production company experience with ratings, we see a similarly static average movie rating trend line, dropping by only 0.22 points. Again, the movie ratings do drop as production companies lessen in experience, but only slightly.

4. Experienced production companies generate higher popularity on average



The above represents the trend of popularity as production company experience decreases. Again, similar to the trends with sorted actor experience, the trend line for popularity decreases at a higher slope than for ratings in the previous chart.

IX. Limitations

There were a few limitations with respect to our analysis, both for bandwidth and data availability.

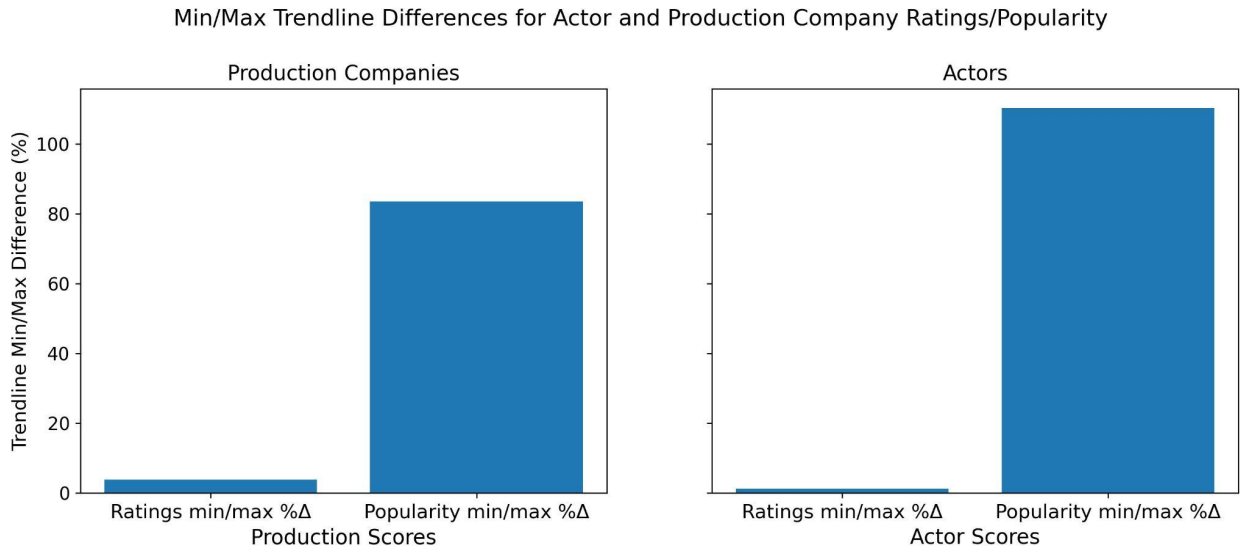
1. We did not utilize the box office revenues of movies. We instead focused on budgets because the budgets are pre-determined by the producers of the movie. Revenues focus more on a consumer's decision to see the movie, but budget gets down to how confident the producer was in the quality of the movie itself. Further analysis into revenues would be of interest for how budgets compare to consumers' behavior to actually purchase the movies.

2. We did not segment the data by sub-categories such as genre and country of origin. This was primarily due to bandwidth constraints -- we would be interested in analyzing the dynamics of ratings and popularity trends segmented by genre and country of origin in the future.

3. We did not pull movie crew data such as director, producers, writers, etc. This was due to bandwidth constraints. We would be interested in studying experience-levels by directors, producers, writers, etc. and add that to our "blockbuster" definition.

4. We did not have the ability to run a full statistical analysis with statistically significant results. In the future, we would be interested in drawing actual conclusions based on a full regression analysis and hypothesis testing.

Results:



In the charts above, we compare the percentage change in the popularity and ratings trendlines of experienced production companies and actors. We use percentage change (absolute value of the slope of the trendline) to give more context to the trends themselves, rather than their absolute point values since popularity and ratings are not scored on the same scale.

One can immediately see that the changes in ratings versus popularity with respect to experience level are markedly different, in similar proportions for both actor and production company experience levels. Popularity's trendline sees a change of over 80% for production companies (left) and over 100% for actors (right) while ratings for both change at less than 5% each.

With more experienced actors and production companies generally wielding higher budgets from our scatterplot and correlation analysis, the above results convey that marketing and budget do in fact boost the popularity of movies for both high-calibur actors and production companies. Yet, the ratings of these movies do not proportionally match the popularity of said movies. There is a disconnect between the actual ratings of a movie and its level of popularity from viewers.

X. Conclusion

Experienced actors and production companies have more credibility and budget behind them to expose more of an audience to their film. This is likely a key factor in the increase in popularity with respect to experienced actors and production companies. Reviews, on the other hand, are based only on the quality and value of the movie itself. There are many possible reasons for this.

One possibility is that movie reviews are quite subjective, so review scores may not follow a linear trend directly correlating to any one quantitative attribute.

Another possibility is that movie quality and value are not necessarily requirements for a high-budget movie with a more experienced cast and crew. Since a large production company is focused primarily on revenue, it may not be necessary to strive for a high-rated movie if the popularity and box office sales are there regardless of marketing budget alone. This would explain why popularity follows more of a trend with better actors and production companies, while ratings trend much less.

The true reason for this discrepancy is unknown but based on our analysis and findings, we believe it is likely a mix of our two theories, along with other factors we may not have accounted for.