Box Office Success: Analyzing the Influence of Top Actors and Genres on IMDb's Highest-Grossing Films

Alexis Adzich, Annie Cen, Jasmine Chu, Kevin Hamakawa, Cassia Ramelb, Rohan Saklani

STATS 140XP

Dr. Vivian Lew

13 December 2024

1 Abstract

This report examines the impact of top actors and film genres on the financial success of movies. By analyzing IMDb data and employing ANOVA tests, the study explores whether the presence of high-profile actors and specific genres correlates with increased gross revenue. The results show a strong link between popular actors and higher box office earnings, with certain actor pairs proving particularly influential, specifically from movie franchises. Additionally, certain genres, such as action, consistently outperform others in terms of financial success. These insights provide valuable guidance for industry professionals, enabling more informed decisions on casting and genre selection to optimize box office potential.

2 Introduction

2.1 Problem Statement

The film industry is a multi-billion-dollar global enterprise, where a movie's success relies on factors such as storytelling, budgets, marketing, and casting decisions. Among these, the choice of top actors and genre are key drivers of box office performance. High-profile actors can attract significant audience attention and boost a film's marketability, while certain genres tend to attract loyal followings, often leading to higher earnings. When combined, popular actors and a compelling genre can create a powerful formula for financial success.

This report investigates whether well-known actors, actor pairs, and specific genres consistently correlate with greater revenue. By analyzing historical box office data, the study aims to identify patterns that drive a film's financial success. The insights gained will help filmmakers, producers, casting directors, and investors make more informed decisions about talent selection and genre targeting in an increasingly competitive industry.

2.2 Research Questions

1. How do actor combinations (e.g., top pairings) and individual actors impact box office success?
2. Are certain genres consistently associated with higher revenue?
3. What role do franchises and recurring actor partnerships play in driving consistent box office success?

3 Methodology

3.1 Dataset Description

The dataset for this analysis was obtained from IMDb and includes the top 1,000 movies and TV shows. It contains comprehensive information about movie characteristics and performance metrics, including:

- **Series_Title:** Name of the movie.
- **Released_Year:** Year of release.
- **Certificate:** Certification earned (e.g., PG, R).
- **Runtime:** Total runtime in minutes.
- **Genre:** The primary and secondary genres of the movie.
- **IMDB_Rating:** Average user rating on a scale of 1 to 10.
- **Overview:** A short textual summary of the movie.
- **Meta_score:** Critical rating provided by Metacritic.
- **Director:** Name of the director.
- **Star1, Star2, Star3, Star4:** Names of the top-billed actors.
- **No_of_votes:** Total number of user votes.
- **Gross:** Domestic box office gross earnings (in USD).

Key descriptive statistics:

- Average domestic gross revenue: $60,513,599
- Maximum domestic gross revenue: $936,662,225 (Star Wars Episode VII: The Force Awakens)
- Most featured actor: Robert De Niro
- Most frequent genre: Drama

3.2 Data Cleaning and Preparation

The dataset was preprocessed to ensure it was clean, consistent, and ready for analysis. Key steps included handling missing values, restructuring data, and removing duplicates or irrelevant entries.

Column names were standardized and simplified for clarity, and sentiment scores were generated for the "Overview" column using a sentiment analysis tool in R. These scores provided a measure of positivity or negativity in movie descriptions, enabling trend analysis across genres and time periods.

Missing values in revenue were imputed with the mean, while other columns with missing data (e.g., runtime, votes) were either imputed with averages or excluded. Genres were simplified by grouping sub-genres (e.g., "Sci-Fi Adventure") into broader categories (e.g., "Adventure") to enable clearer trend analysis. Duplicate entries, such as alternate versions of the same movie, were removed, retaining only the highest-revenue version for analysis.

3.3 Exploratory Data Analysis

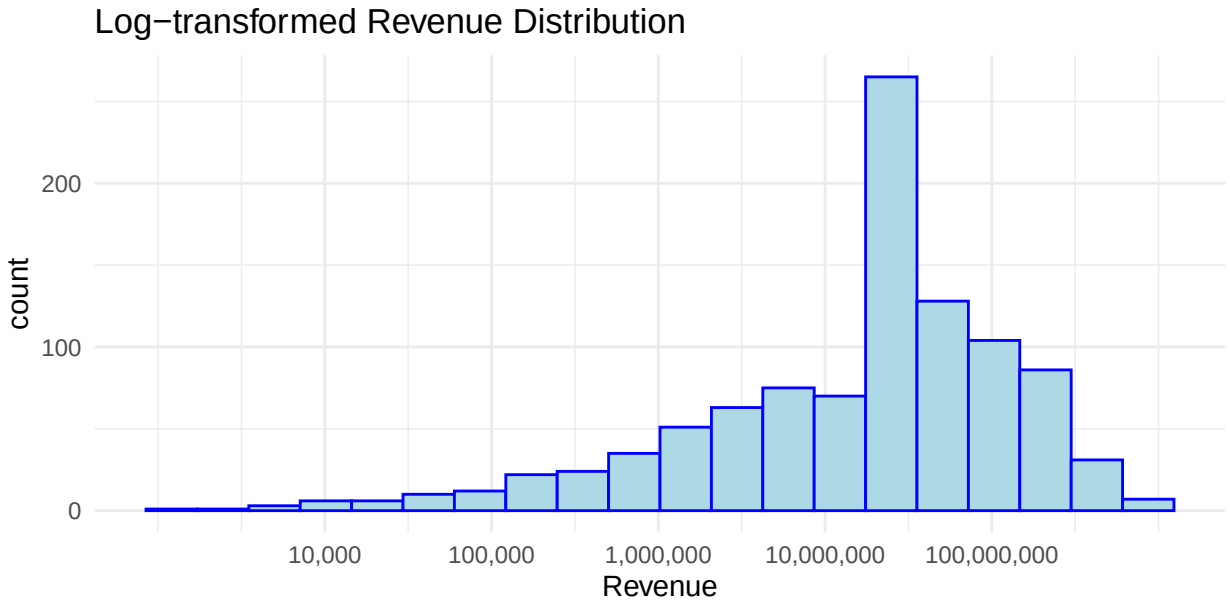

Figure 3.1: Distribution of Top 1000 Movie Revenues (Log-Transformed) from 1920 to 2020

Figure 3.1 reveals several key insights about movie revenue patterns. A prominent feature is the clear peak between $10 million and $100 million, where the majority of films fall. This indicates that while the film industry produces many movies with moderate success, high earnings are far less common. As revenue increases beyond the $100 million mark, the number of films in higher revenue ranges significantly decreases, illustrating the rarity of blockbuster-level success. Specifically, only a small proportion of films reach the $500 million to $1 billion range, emphasizing the exceptional nature of these top-grossing films.

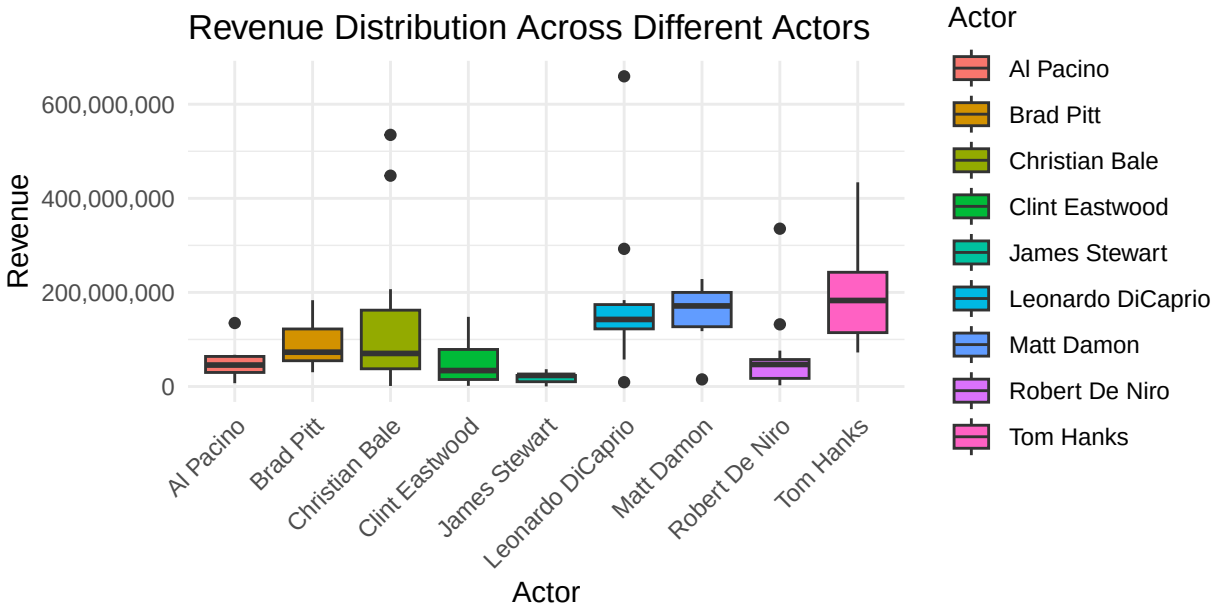

Figure 3.2: Boxplots Showing the Distribution of Revenue by Actor

Figure 3.2 uncovers critical details of revenue distribution depending on the specific actor in that film. A component of this chart is that Tom Hanks and Leonardo Dicaprio have higher median revenues compared to the other actors, which indicate that they are likely to increase revenue of a movie based on their addition to the cast. Actors like Christian Bale and Clint Eastwood have a wider gross range, indicating variability. Also, outliers reflect when specific movies achieved significantly higher/lower gross than their typical performances.

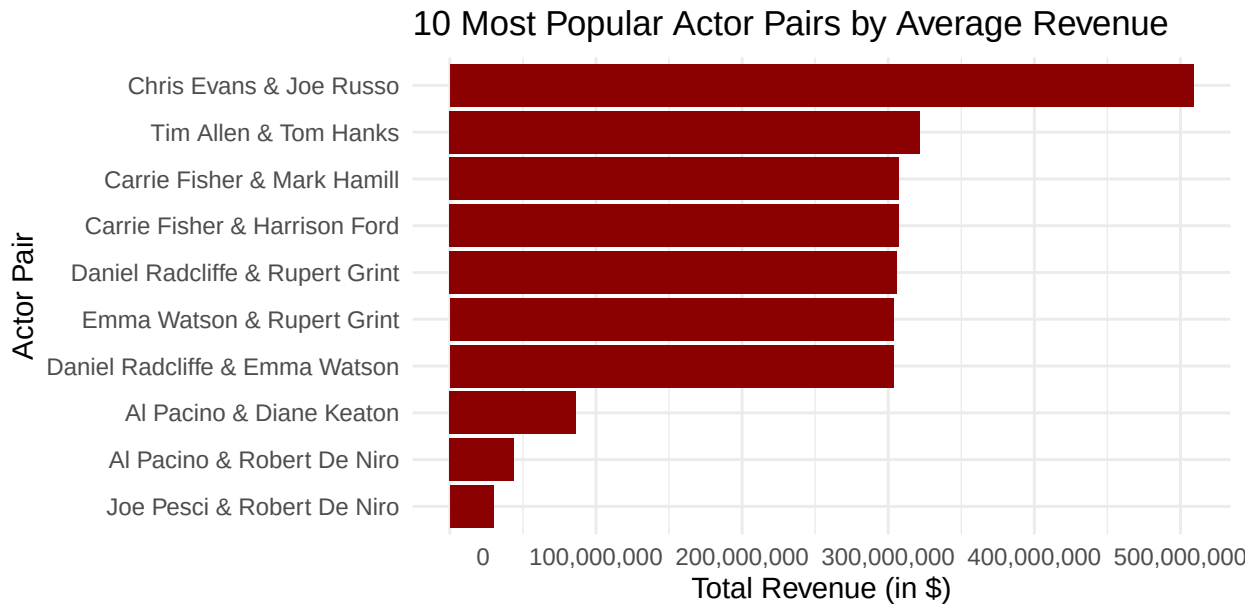

Figure 3.3: Top Actor Pairs by Avg Revenue

Figure 3.3 illustrates which actor pairs to hire in order to get more revenue. We looked at the top 10 actor pairs by average revenue made in all their films. Chris Evans and Joe Russo were the highest earning duo due to their films in Marvel. Tom Hanks and Tim Allen were second due to their involvements in the Toy Story series. Third is taken by other duos from franchises like Harry Potter and Star Wars. The dominance of these duos from these renowned franchises highlight the importance of recurring partnerships in films with high gross revenues.

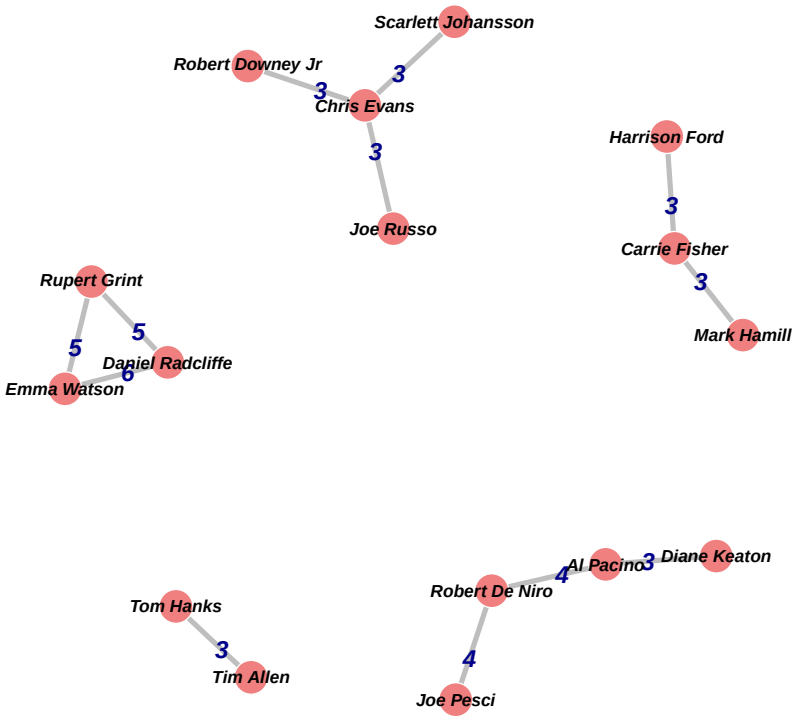

Figure 3.4: Network Graph of Actor Partnership

Figure 3.4 is a network analysis of top 5 actor partnerships in films and exemplifies how consistent actor pairing assists in financial success. The actors are grouped into clusters to display which collaboration they are a part of. The edge labels or number listed between their names is justifying the strength of the collab (# of movies together in a dataset with any actor labeled as Star1 or Star2). Harrison Ford, Carrie Fisher, and Mark Hamill are part of the Star Wars series with 3 total movies together. Daniel Radcliffe, Emma Watson, and Rupert Grint have been in 8 movies together all from the Harry Potter series (5 in this dataset) and Tom Hanks and Tim Allen have voiced roles in the 4 Toy story films (3 in this visualization because in Toy Story 2, they are listed as Star3 and Star4. Scarlett Johansson, Chris Evans, Robert Downey Jr., and Joe Russo are all from the Marvel franchise. Diane Keaton, Al Pacino, Robert De Niro, Joe Pesci have associations due to their involvements in Mafia movies like The Godfather or Goodfellas. Another collaboration not shown here but in the data is the Lord of the Rings Trilogy with Elijah Wood, Ian Mckellan, and Orlando Bloom. The Top 3 in revenue of these franchises is the Marvel Cinematic Universe, Harry Potter, and Star Wars. To sum it up, actor partnerships give substantial grossing, underscoring importance of big blockbuster franchises to increase sales perhaps due to large fan loyalty and known chemistry between actors.

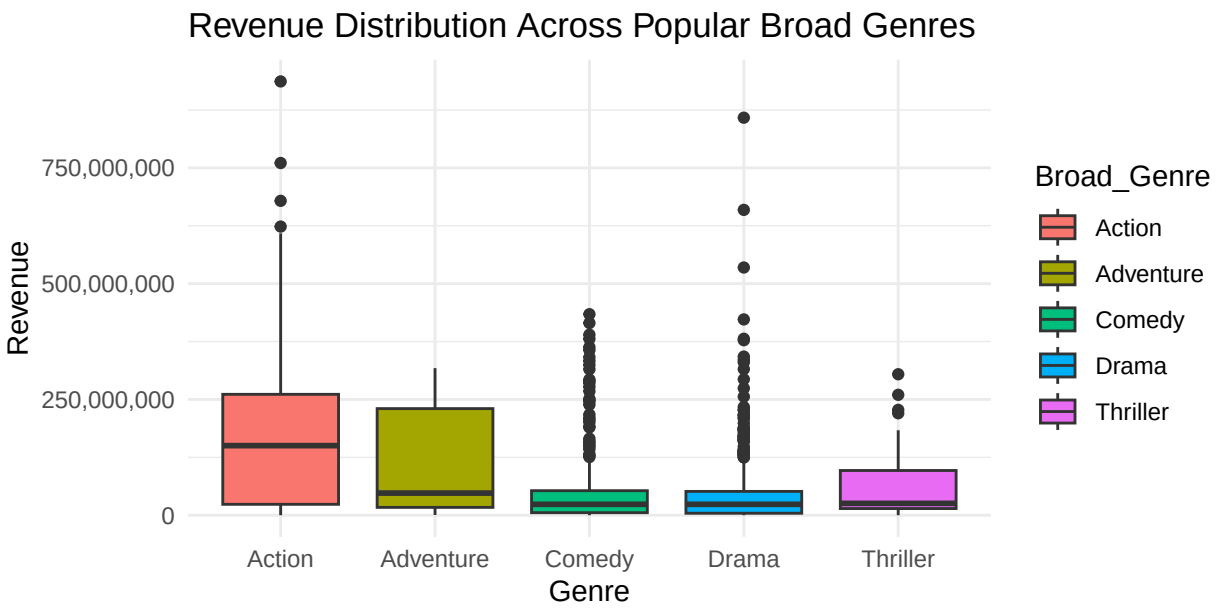

Figure 3.5: Boxplots Showing Revenue Distribution Across Popular Broad Genres

Figure 3.5 shows that Action films generate the highest revenues, as indicated by the median, with substantial variability. Several outliers with significantly higher revenues suggest that a few very successful action films may be driving up the overall revenue. Adventure films, while still successful on average, have a more modest range and a more compressed distribution, with fewer extreme outliers. In contrast, Comedy, Drama, and Thriller films generally underperform in terms of box office revenue compared to the other genres, though there are occasional standout hits.

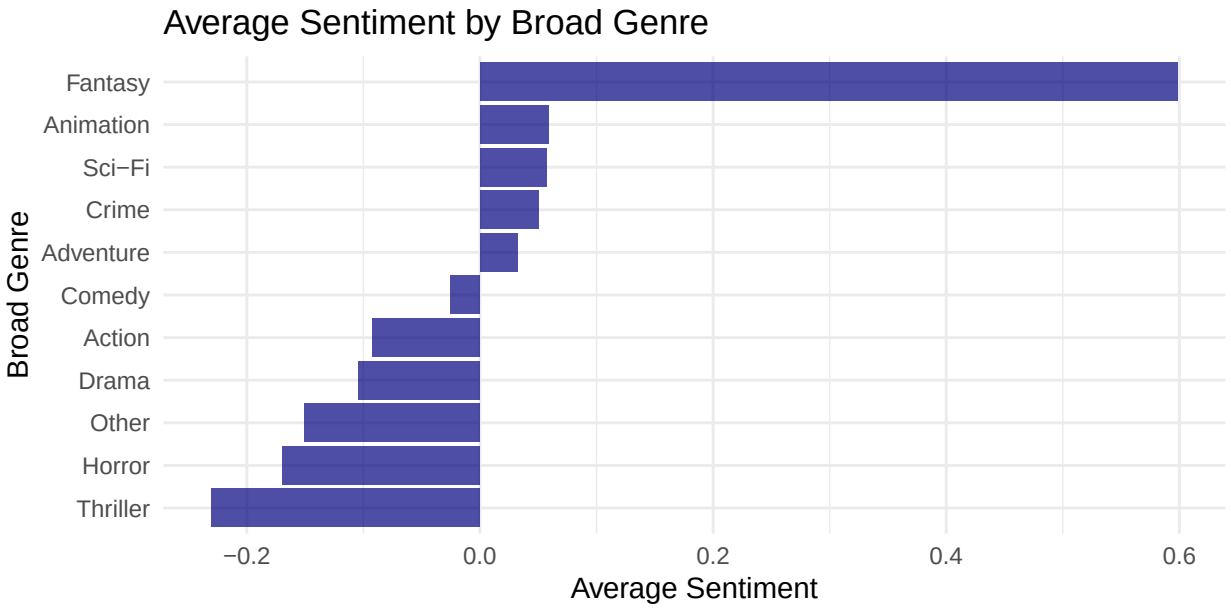

Figure 3.6: Bar Chart Depicting Average Sentiment Scores by Broad Genres

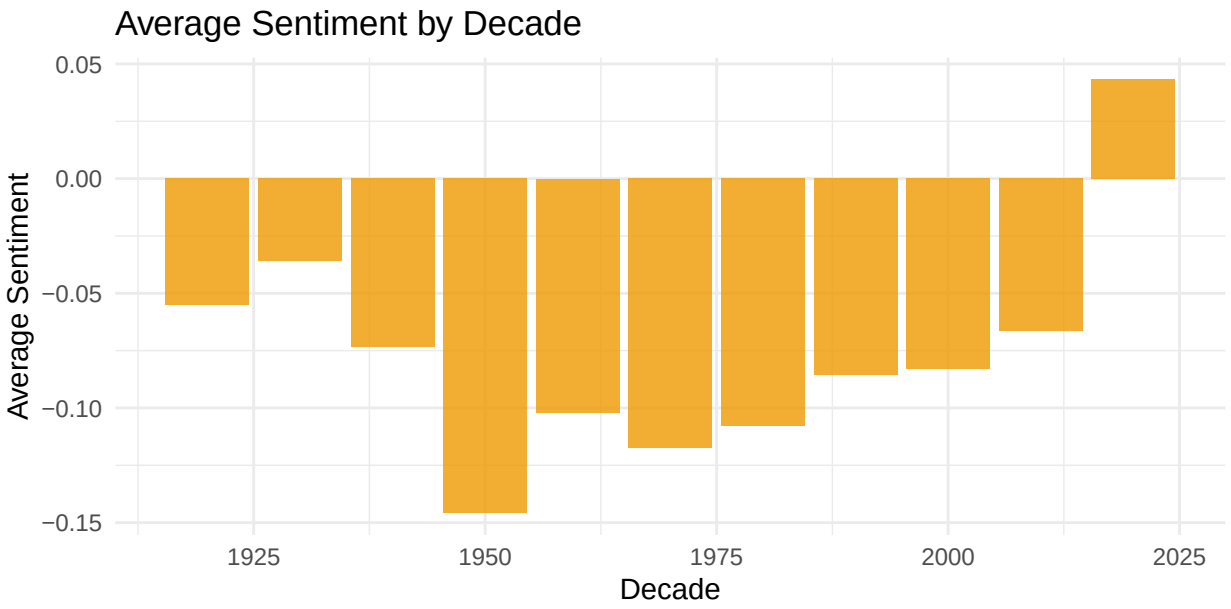

Figure 3.7: Bar Chart Depicting Average Sentiment Scores by Decade

For a closer look at genres, Figure 3.6 shows significant variation in the average sentiment by broad genres. Fantasy has the highest average sentiment, despite having the lowest sample size, possibly due to the optimistic nature of fantasy films. Animation, Sci-Fi, and Crime follow behind with a similar positive trend. On the other hand, Thriller, Horror, Drama, and Action exhibit the most negative sentiment and possibly darker themes. Comedy and Adventure fall closer to neutral, with neutral sentiment quantitatively defined as sentiment scores ranging between -0.05 and 0.05.

Additionally, Figure 3.7 reveals the sentiment trends in movie overviews appear mostly negative between the 1930s and 2010, with the lowest average sentiment around the 1950s. Historical events during these decades, such as wars and failing economies, may have influenced the more negative themes of movies during these time periods. The upward shift in the 2020s, where sentiment becomes positive for the first time, suggests a potential evolution in more optimistic movie themes.

3.4 Statistical Methods

We tested the following hypotheses to evaluate differences in mean movie revenue:

1. Actors:
H_0: There is no significant difference in mean revenue across top actors (appeared in 10+ films).
H_a: At least one top actor has a significantly different mean revenue.

2. Actor Pairings:
H_0: There is no significant difference in mean revenue across top actor pairings (appeared in 3+ films together).
H_a: At least one top actor pairing has a significantly different mean revenue.

3. Broad Genres:
H_0: There is no significant difference in mean revenue across top broad genres (genres with 10+ movies).
H_a: At least one top broad genre has a significantly different mean revenue.

For all of these hypotheses, we performed one-way Analysis of Variance (ANOVA) tests to assess overall differences. Tukey post hoc tests were then applied to identify specific actors, pairings, or genres contributing to the variations while controlling for Type I error. All analyses were conducted in R, and standard ANOVA assumptions were verified.

4 Results and Interpretation

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## Actor         8 4.617e+17 5.771e+16   5.901 3.35e-06 ***
## Residuals   102 9.974e+17 9.779e+15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.1: Summary of ANOVA Test on Revenue Across Top Actors

Figure 4.1 shows a summary of the one-way ANOVA test we performed to evaluate whether significant differences exist in the mean movie revenue generated by different actors. The test was conducted on a subset of popular actors who appeared in 10 or more films. Our analysis reveals a statistically significant difference (p-value = 0.0000335) in mean revenue across these actors, allowing us to reject the null hypothesis, which states that there are no differences in mean revenue between actors.

	diff	lwr	upr	p adj
Leonardo DiCaprio-Al Pacino	136975734	8577018	265374450	0.0273126
Tom Hanks-Al Pacino	158073469	37356237	278790701	0.0021722
Leonardo DiCaprio-Clint Eastwood	136927189	6099391	267754987	0.0329395
Tom Hanks-Clint Eastwood	158024924	34727185	281322663	0.0029727
Leonardo DiCaprio-James Stewart	168367556	31425645	305309467	0.0052741
Matt Damon-James Stewart	139207938	2266027	276149849	0.0432248
Tom Hanks-James Stewart	189465292	59698167	319232416	0.0003626
Robert De Niro-Leonardo DiCaprio	-127378222	-248656017	-6100427	0.0318426
Tom Hanks-Robert De Niro	148475957	35362195	261589719	0.0020951

Figure 4.2: Significant Tukey Test Pairwise Comparisons of Revenue Among Top Actors

Next, to identify which actors contributed most to these differences, we conducted a Tukey post hoc test. In figure 4.2, the results indicate that actors such as Tom Hanks and Leonardo DiCaprio significantly outperformed others, including Al Pacino, Clint Eastwood, Stewart, and Robert De Niro, in terms of mean movie revenue. This highlights the notion that casting certain high-performing actors could be a crucial factor in predicting box office success.

```
##             Df    Sum Sq   Mean Sq F value  Pr(>F)
## ActorPair    9 6.882e+17 7.646e+16     8.1 6.8e-06 ***
## Residuals   29 2.738e+17 9.440e+15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.3: Summary of ANOVA Test on Revenue Across Top Actor Pairs

We also conducted a one-way ANOVA test to assess what differences are generated in mean movie revenue from different popular actor pairings. The test focused on actor pairings that have been in 3 or more films together. Figure 4.3 discloses a statistically significant difference in mean revenue among these pairings (p-value = 0.000068), allowing us to reject the null hypothesis. This suggests that some specific actor pairings have higher average revenues.

	diff	lwr	upr	p adj
Chris Evans & Joe Russo-Al Pacino & Diane Keaton	422430483	151115863	693745102	0.0003812
Chris Evans & Joe Russo-Al Pacino & Robert De Niro	464829034	193514415	736143654	0.0000902
Daniel Radcliffe & Emma Watson-Al Pacino & Robert De Niro	259752824	17081651	502423997	0.0285013
Daniel Radcliffe & Rupert Grint-Al Pacino & Robert De Niro	262071233	27105880	497036586	0.0196520
Emma Watson & Rupert Grint-Al Pacino & Robert De Niro	259752824	17081651	502423997	0.0285013
Tim Allen & Tom Hanks-Al Pacino & Robert De Niro	277760552	23968965	531552140	0.0233879
Joe Pesci & Robert De Niro-Carrie Fisher & Harrison Ford	-277532202	-531323790	-23740614	0.0235626
Joe Pesci & Robert De Niro-Carrie Fisher & Mark Hamill	-277532202	-531323790	-23740614	0.0235626
Joe Pesci & Robert De Niro-Chris Evans & Joe Russo	-478826637	-732618225	-225035049	0.0000184
Joe Pesci & Robert De Niro-Daniel Radcliffe & Emma Watson	-273750427	-496658133	-50842720	0.0074791
Joe Pesci & Robert De Niro-Daniel Radcliffe & Rupert Grint	-276068835	-490561876	-61575795	0.0044435
Joe Pesci & Robert De Niro-Emma Watson & Rupert Grint	-273750427	-496658133	-50842720	0.0074791
Tim Allen & Tom Hanks-Joe Pesci & Robert De Niro	291758155	56792802	526723508	0.0066372

Figure 4.4: Significant Tukey Test Pairwise Comparisons of Revenue Among Top Actor Pairs

After finding significant results with the one-way ANOVA test, we used a Tukey post hoc test to identify actor pairings with higher revenue. Figure 4.4 shows that the top pairings come from major ensemble franchises, such as the Marvel Cinematic Universe with Chris Evans and Joe Russo, and Star Wars with Carrie Fisher and Mark Hamill. We also see that collaborations driven by two main characters with Tom Hanks and Tim Allen in Toy Story, and from multiple duos in the same trio with Daniel Radcliffe, Emma Watson, and Rupert Grint in Harry Potter bolster greater mean revenue. These findings reveal the significance of duos from franchises when it comes to predicting box office success.

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Broad_Genre   4 1.319e+18 3.298e+17   36.78 <2e-16 ***
## Residuals   972 8.716e+18 8.967e+15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.5: Summary of ANOVA Test on Revenue Across Popular Broad Genres

For the final hypothesis, Figure 4.5 demonstrates that the one-way ANOVA test reveals statistically significant differences in average movie revenue among the top broad genres. The extremely small p-value leads us to reject the null hypothesis, indicating that at least one broad genre has a significantly different mean revenue compared to the others.

	diff	lwr	upr	p adj
Adventure-Action	-74255944	-135942883	-12569006	0.0091690
Comedy-Action	-132928626	-169014333	-96842919	0.0000000
Drama-Action	-144523627	-178082183	-110965071	0.0000000
Thriller-Action	-125307869	-173040253	-77575484	0.0000000
Comedy-Adventure	-58672682	-114152181	-3193183	0.0320843
Drama-Adventure	-70267683	-124137649	-16397716	0.0035005

Figure 4.6: Significant Tukey Test Pairwise Comparisons of Revenue Among Popular Broad Genres

The Tukey test results in Figure 4.6 shows that Action films tend to outperform all other genres in revenue comparisons. Overall, Action and Adventure films emerge as the top revenue-generating genres, as evidenced by the negative differences in the table.

Interpretation of Results

The findings suggest that certain actors and actor pairings are consistently linked to higher financial returns, likely due to their strong audience appeal, alignment with high-budget productions, or association with successful franchises. Additionally, the analysis of film genres indicates that some genres outperform others in terms of gross revenue. This highlights the importance of leveraging "star power" during casting decisions and strategically selecting genres to maximize revenue potential.

5 Discussion

5.1 Summary

In conclusion, the analysis of IMDb's top 1000 movies dataset has provided valuable insights into the key factors driving box office success, with a particular focus on actor and genre influence. Our findings highlight the substantial role that individual actors, actor pairings, and genres play in generating revenue.

1. Actor Influence

Through ANOVA analysis, we found that individual actors and actor pairings significantly impact a film's financial performance.

- *Top Individual Actors*: Tom Hanks and Leonardo DiCaprio consistently outperform others like Robert De Niro and Al Pacino in mean movie revenue.
- *Top Actor Pairings*: Pairings, especially within larger cast franchises like the Marvel Cinematic Universe and Star Wars, generated the highest revenues. Actor pairings such as Chris Evans and Joe Russo stand out as top performers.

2. Genre Influence

The genre analysis revealed that Action and Adventure films take the lead in terms of revenue generation, outperforming other broad genres. This reinforces the idea that genre-specific appeal, coupled with star power, is a key driver in box office success.

Broader Insights & Key Takeaways

Investing in renowned actors and successful pairings, especially within the framework of franchise-driven films, is a highly effective strategy for maximizing revenue. These star-driven projects attract large audiences and create long-term financial value through franchise potential. When it comes to genres, Action and Adventure films consistently demonstrate broad audience appeal and have strong potential for franchise development, making them prime candidates for high-revenue success.

Ultimately, the findings underscore the critical importance of aligning film production strategies with audience preferences and emerging industry trends. By prioritizing star power and selecting the right genres, filmmakers can significantly increase their chances of achieving financial success in an increasingly competitive market.

5.2 Limitations & Future Studies

The dominance of larger franchises underscores the importance of continuity and established storylines in attracting audiences and maintaining high revenue streams. We are interested in further exploring the influence of these notable series, such as the success of a trilogy versus saga.

Further analyses should focus on non-actor factors such as budget allocation, release timing, and director influence. Additionally, examining diversity in casting (e.g., gender, race) and its impact on financial performance could provide a more inclusive perspective on box office success. In terms of genres, we can compare stand-alone genres versus hybrid genres like Action-Comedy or Sci-Fi-Thriller.

Some dataset constraints are not including variables like budget and international earnings. These additional variables to this dataset would likely make an effect on a movie's box office success and these factors should be checked out in future analyses. The gross revenue was also not adjusted for inflation, which may lead to misleading results in favor of recent movies. Incorporating information with budget allocation could also provide a relationship with revenue and determine investment levels on certain actor pairings or different genres.

Lastly, some industry trends to later explore are the rise of streaming platforms and the pandemic's global impact on theater visits which have altered box office revenue and overall revenue success.

6 References

Harshit Shankhdhar. (2020). IMDb Dataset of Top 1000 Movies and TV Shows. Kaggle. Retrieved from https://www.kaggle.com/harshitshankhdhar/eda-on-imdb-movies-dataset

IMDb. (2020). IMDb: Movies, TV and Celebrities. Retrieved from https://www.imdb.com