Hacking the Box Office

Ngoc Hoang, Maryam Khalili, Alem Shaimardanov, Sashank Silwal

Department of Computer Science, New York University Abu Dhabi

The film industry is a significant contributor to the global economy. In the United States alone, 2.2 million people work in jobs supported by the industry (Pangarker and Smit 2013) (*1*). Thus, it's important to understand the variables that influence the box office revenue of the movies. Using the IMDB database, we explore whether the demographic data of the cast, such as ethnicity, gender, age, and star power, has an effect on the global box office of the movies. Overall, actors have a higher presence in movies than actresses with longer careers and more leading roles. The number of white actors and actresses remains much higher than actors and actresses of color, and movies with more diverse cast were found to make lower profit than those with majority white cast.

Introduction

The film industry is a significant contributor to the global economy. A crucial indicator of the commercial success of a movie is undoubtedly its box office revenue. It is commonly believed that factors such as budget and the popularity of the director/cast influence the box office revenue of a movie. Understanding the significance of such factors is useful in predicting

the financial success of films. In this research, we studied the demographic data of cast and its impact on the financial success of movies.

Literature Review

Previous studies have addressed some of the factors that influence the box office revenue of movies. An empirical investigation of direct links to box office success and their measurement reveals that the star power factor was statistically significant. Other significant factors were critics' reviews, screen coverage, and top distributor. Contrary to popular belief, the number of Academy awards was found to be insignificant. Seasonality, being a sequel, and being a popular genre was also not statistically significant. Star power (*2*) is defined using two approaches: (1) in terms of the number of nominees and winners of Academy awards for all key players in each film before a sample year. Best Actor/Actress, Best Supporting Actor/Actress, Best Director. [Star power defined by this approach was not found to be statistically significant], and (2) in terms of stars' earning power, which is defined as the average value of box office revenue generated throughout all key players' entire acting careers before a sample year. Dataset: Box Office Mojo website. [Star power defined by this approach was found to be statistically significant]. In our research, star power is a metric to measure the popularity of a movie's cast prior to that movie.

Most significant expenses are associated with production factors such as the ones mentioned above. Another study (*3*) took a more interesting turn and looked at an inexpensive, usually not talked about factor: the title of the movies. The analysis shows that an informative movie title, that is, a movie title that contains information about its genre or storyline, affects a movie's box office revenue. While it cannot be denied that the budget is one of the major factors that determine box office success, the authors point out that a high production budget has not been a necessary condition for box office success in the history of filmmaking. Lee and Choeh

(2020) (*4*) look at the impact of online review and electronic word of mouth in order to forecast the box office revenue and their findings suggest that movies with more helpful reviews or those that are reviewed by more helpful reviewers show better performances at predicting box office revenues. Other studies have looked at the relationship between genre and revenue. Anast (1967) has found a negative correlation between the action/adventure genre and revenue and a positive correlation between adventure/erotic genre and box office performance.

The literature reviewed by the authors above, serve as the cornerstone for future studies that aim to identify what factors influence box office revenue. One particular variable that remains mostly unexplored when it comes to studying and determining the financial success of movies is the diversity of the cast. A Hollywood Diversity Report by UCLA (2016) (*5*) concludes that audience favor diversity on screen. The findings of the study suggests that a more diverse cast (inclusive of women and minority) leads to higher ticket sales.

We can see from the various and sometimes contradictory results that there is not a single, consistent factor in determining the success of a movie in terms of its generated revenue. Our literature review leads to the following potential factors in regards to their impact on box office: ethnicity, gender, and star power (of the cast). We label these variables under the umbrella term 'demographic data'.

Results

We first look into the gender dimension of the film industry by comparing the average career lengths of actors and actresses. An actor/actress' career length is defined as the time between his or her first movie (i.e. first movie in the IMDb dataset where he or she is listed among the principal cast) and his or her latest movie. Figure 1a compares, for each genre, the average career length of actors (x-axis) and the average career length of actresses (y-axis). On average, actors have longer career than actresses across all genres. Some genres observe a small gap,

such as romance (626 movies) where actors' and actresses' average career lengths are 21.6 years and 20.9 years respectively, and comedy (1,252 movies) with 21.4 years for actors and 20.1 years for actresses. Some other genres observe a bigger gap, such as action (993 movies) with 23.7 years for actors and 17.6 years for actresses, a gap of 6 years. Documentary and news also observe big gaps, but these two genres include a low count of movies (25 and 2, respectively). Figure 1b visualizes the numbers of actors (x-axis) and actresses (y-axis) by genre. The genres with the highest counts of actors and actresses also generally observe the biggest gaps in the numbers of actors and actresses, for example, drama (1,884 movies) has about 1,700 more actors than actresses. Romance, similar to the average career lengths, observes a more balanced ratio of actors and actresses, as does family. Most genres lie close to the $y = x$ line and to the origin since these genres include fewer movies (and consequently, fewer actors and actresses) and they observe a smaller gap in actors and actresses.

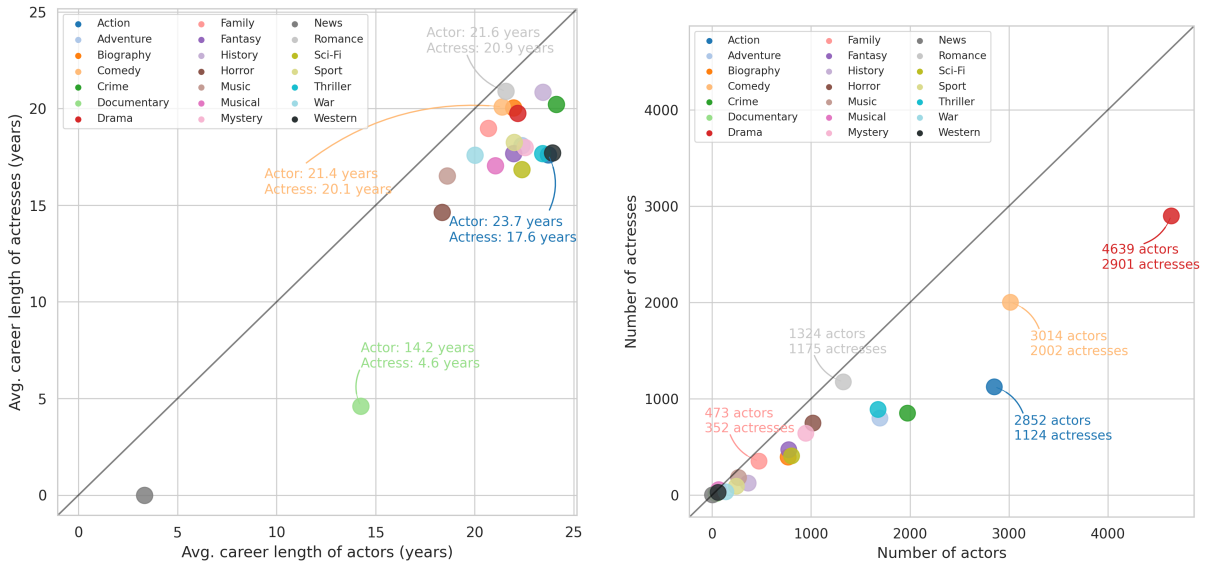

Figure 1: **Gender differences in the film industry.**

To further study the gender gap in career in the film industry, we used a simple regression

		Coefficient	95% CI	p
Actors	Intercept (α_{actor})	-1.6581	[-2.157, -1.159]	<0.001
	Career length (β_{actor})	0.8176	[0.792, 0.843]	<0.001
	Observations	3244		
	Adjusted R^2	0.556		
Actresses	Intercept ($\alpha_{actress}$)	-0.0827	[-0.526, 0.360]	0.714
	Career length ($\beta_{actress}$)	0.6657	[0.641, 0.691]	<0.001
	Observations	2028		
	Adjusted R^2	0.573		

Table 1: **Regression results on the relationship between career length and total number of movies.**

model to analyze the differences in the relationship between career length and the number of movies accumulated for actors and actresses. The regression model is specified as follows: $totalMovies = \alpha + \beta \times careerLength$. We ran this model on two groups of cast members with one including only the actors and the other one including only the actresses. The regression results shown in Table 1 can be interpreted that for every increase of one year in a cast member's career, an actor will accumulate an additional 0.82 of a movie while an actress will accumulate an additional 0.67 of a movie. This regression model serves as a quick way to study the two aspects of a cast member's career and does not control for other variables. See Supplemental Figure 2 for the detailed regression plots.

Next, we investigated the gender composition of movie casts. Figure 2 shows, for each year, the proportion of movies released with a higher proportion of actors, a higher proportion of actresses, or an equal proportion of both. It is evident that in all years, more than half of the movies have a higher proportion of actors (i.e., more actors than actresses). Movies with a higher proportion of actresses accounted for a small portion of all movies (less than 20% in all years). This trend remained consistent for the entire time period from 2000 to 2019.

Figure 3 shows the average revenue of movies in each year, when they are grouped into three categories based on the proportion of actors versus actresses in the principal cast. In 17

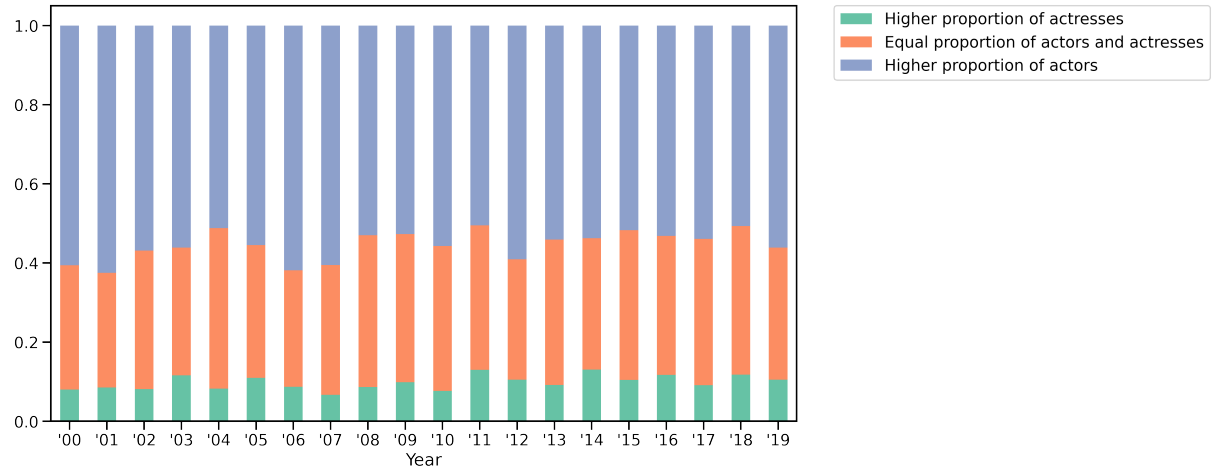

Figure 2: **Proportions of movies with regards to gender composition of the cast over the years.**

out of the 20 years, movies with a higher proportion of actors earned more revenue than movies with a higher proportion of actresses or an equal proportion of genders. In 2009, movies with a higher proportion of actresses had the highest average revenue (and also a high standard error). This may be due to the impact of the movie Avatar, which was released in the same year and is among the highest grossing movies of all time, and was included in the movies with a higher proportion of actresses. In 12 out of the 20 years, movies with a higher proportion of actresses earned less revenue than movies with an equal or higher proportion of actors. See Supplemental Figure 3 for the median revenue of movies in each group.

We further analyzed the data based on the gender homogeneity of the cast. Generally, movies with a homogeneous gender composition (i.e. include all actors or all actresses, but not a mix of both) generated the same average revenue as movies with a mix of actors and actresses. In 2012, however, movies with all actors or actresses started to generate higher average revenue. For detailed graphs on this analysis, see Supplemental Figure 4. To further study movies with a homogeneous gender composition, we looked into actor and actress homogeneity (Supplemental Figure 5). In studying actor homogeneity, we compared average revenue of movies with only actors versus the rest of the dataset (i.e. movies with either only actresses, or a

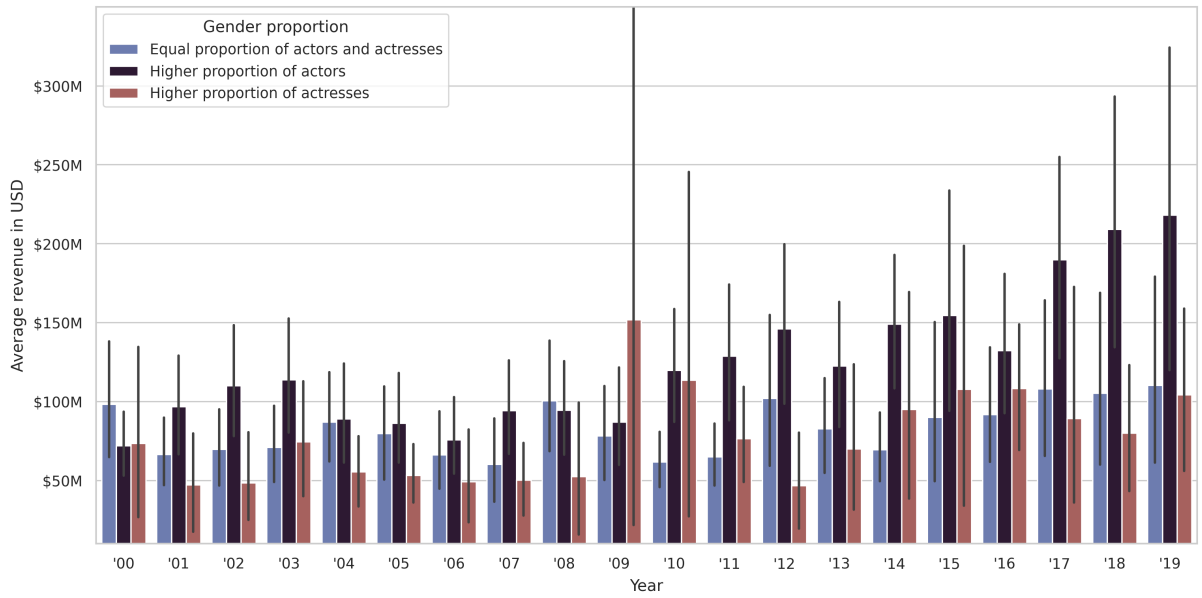

Figure 3: **Average revenue of movies with regards to gender composition of the cast.**

mix) and found that in 14 out of the 20 years, movies with only actors generated higher average revenue. In studying actress homogeneity, we followed the same process for movies with only actresses versus the rest, where the results showed that in 16 out of the 20 years, movies with only actresses generated lower average revenue.

We now move on to the ethnicity dimension of the film industry. Figure 4 visualizes the number of principal cast members (actors and actresses) by ethnicity. Throughout the years, the number of actors and actresses in all groups except for white remain low: for all years, no group (other than white) has more than about 120 actors and actresses. Thus, the gap remains wide between the number of white actors and actresses compared to their colleagues who are people of color. There is a decline towards the later years (from 2016) in the number of principal cast members across most groups, including white cast. This is due to a gap in our dataset, as newer movies tend to lack financial information (budget and revenue) and thus are not included in our analysis (see **Methods** for more details).

For each movie, we used Gini impurity index as a measure of the diversity of the principal

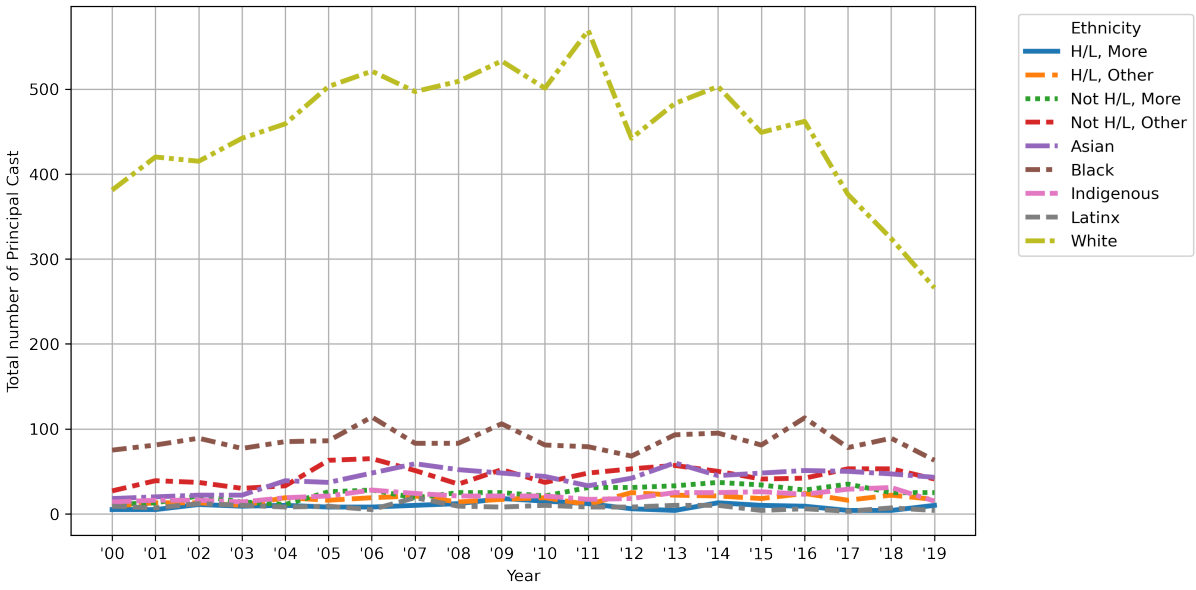

Figure 4: **Counts of cast members by ethnicity over the years.**

cast, where the higher the index, the more diverse the cast is (see **Methods** for details on classifying ethnicity and calculating diversity index). Figure 5 visualizes the average diversity index throughout the years of movies within 5 genres with the highest movie counts. Generally, there was an upward trend in terms of diversity index across the genres. As of now, we do not draw any definite conclusion based on this observation, as it either reflects an actual increase in cast diversity in the film industry or was caused by a drop in the number of white cast towards the later years, as previously discussed.

We divided the movies into three groups with regards to cast ethnicity to study its impact on the financial success of movies. The first group includes all the movies where the majority of the cast are white actors and actresses, while the second group includes the movies where people of color account for a majority of the cast, and the third group includes movies where there is a balance between white versus non-white cast. Figure 6 summarizes the results from this analysis. On average, movies with majority white cast accumulated higher revenue in 16 out of the 20 years. Additionally, in 17 out of the 20 years, movies with majority non-white cast

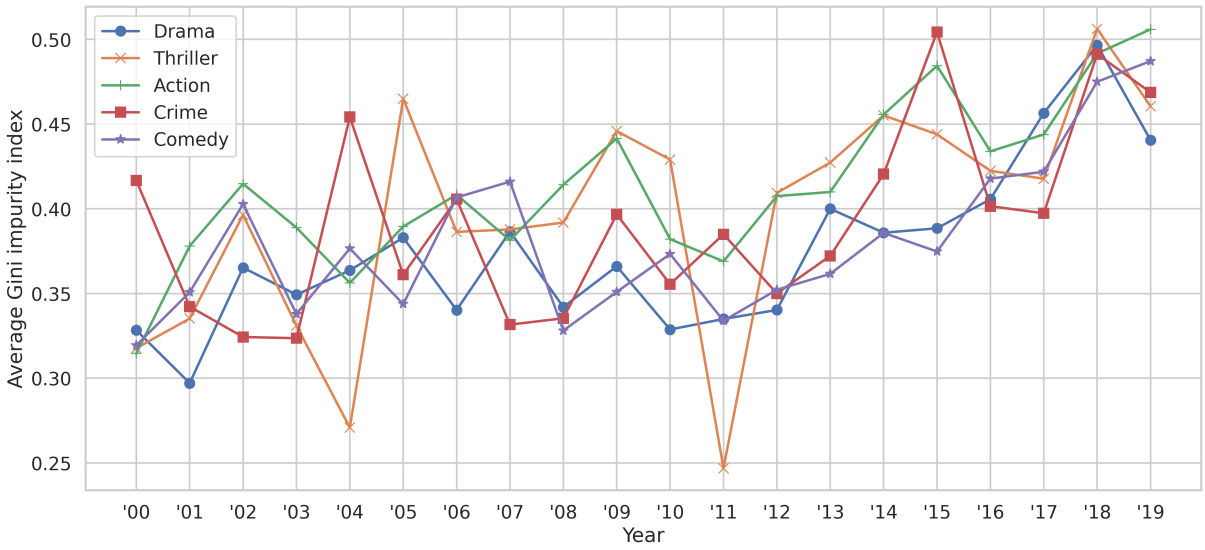

Figure 5: **Ethnic diversity of cast (Gini impurity index) over the years for top 5 genres with the highest movie counts.**

accumulated the lowest average revenue out of the three groups. In 2007 and 2018, the group of movies with majority non-white cast generated higher average revenue than movies with equal white and non-white cast, and our speculation is that this has to do with a number of outlier majority non-white movies in these two years that generated very high revenue, such as Black Panther (2018), Crazy Rich Asians (2018), Rush Hour 3 (2007), and I am Legend (2007).

We further studied the ethnicity dimension of the problem by considering the ethnic homogeneity of the cast (see Supplemental Figure 6). We first divided the movies into movies with ethnically homogeneous cast (e.g. fully white, fully black, fully Asian) and movies with ethnically diverse cast. The results from this analysis showed that movies with ethnically homogeneous cast, on average, generated higher revenue. We further isolated movies with fully white cast and found that such movies, on average, generated higher revenue than the remaining movies.

Next, we looked into the intersection of gender and ethnicity with regards to movie career. For this, we considered the ordering of principal cast, an aspect that has not been discussed so

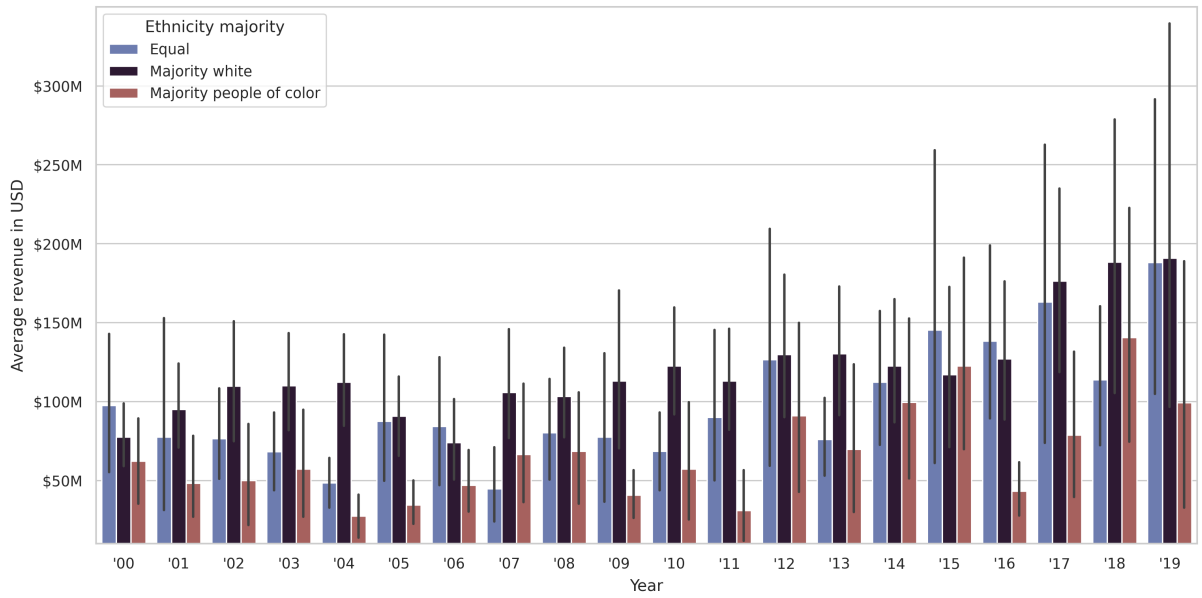

Figure 6: **Average revenue of movies with regards to ethnicity majority.**

far. For a certain movie, the ordering of the actors and actresses of its principal cast as shown in the IDMb dataset can reflect the prominence of their movie roles e.g. the actor/actress listed first plays the main character. Figure 7 summarizes the results of the analysis, where each cell $c_{e,r,g}$ at row e, column r of gender g (actors or actresses) reflects the number of movies that has an actor or actress of ethnicity e playing role number r. For example, $c_{white,1,actors}$ = 1,825 (top left cell in dark blue color) reflects that there are 1,825 movies in our dataset that list a white actor as the first role. All cells in the figure, except for the row of white actors and actresses, are in pale colors, effectively showing the big gap in opportunities available to actors and actresses of color compared to their white colleagues. Looking at the total numbers, it can be observed across all ethnic groups that actors most frequently play the first roles, while actresses most frequently play the second or third roles. In many groups (black, Asian, indigenous, etc.), actresses show the highest concentration in the third or later roles.

To dive deeper into the relationship between our demographic variables and their influence on the financial success of movies, we ran two separate regression models. The first model used

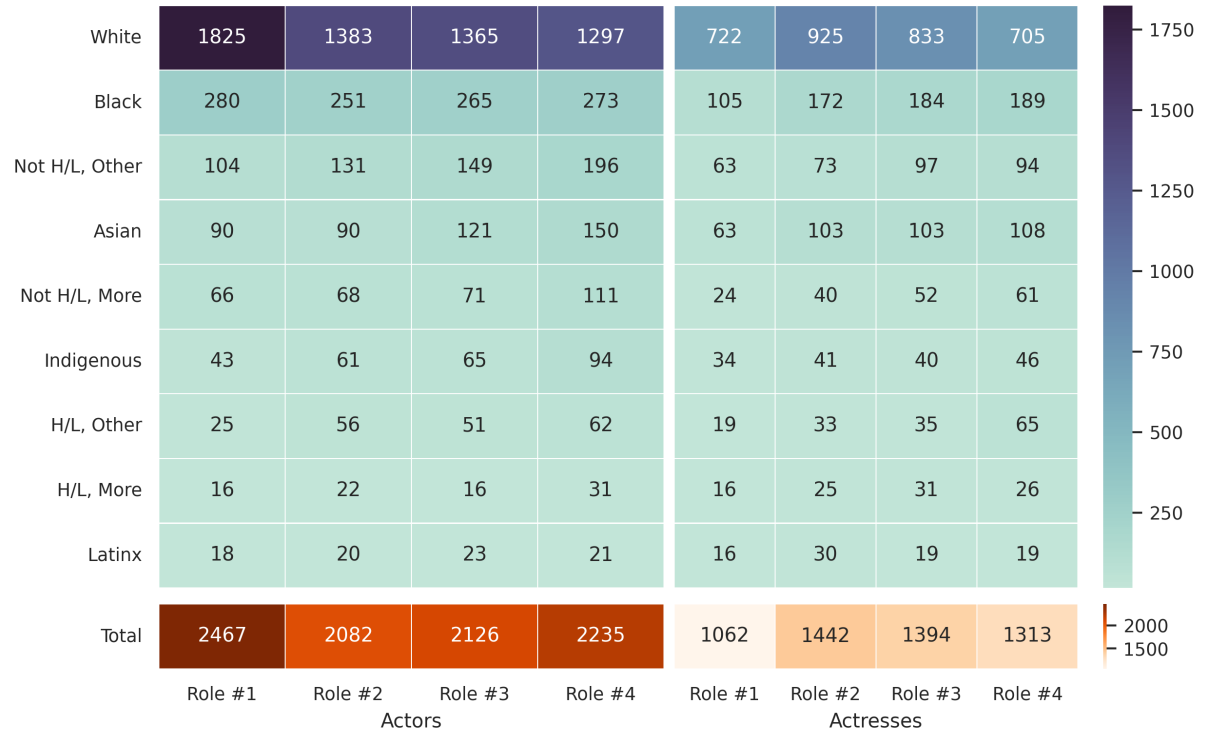

Figure 7: **Distribution of role orderings by gender and ethnicity.** The majority of movies ($N = 3491$) list four cast members in the principal cast. For movies with more than five cast members, ($N = 10$), all roles after the fourth roles are grouped into Roles #4.

revenue as the dependent variable while incorporating budget[1], ethnic diversity of cast (Gini impurity index), star power (average and maximum), and proportion of (male) actors as the independent variables, in addition to a number of dummy variables to account for the genres of the movies. Results of this regression model are summarized in Table 2. The result that stood out the most from this regression model is the disproportionate influence of budget on revenue (p-value < 0.001), to the point that only a handful of genre variables in the model were

[1]Our initial design was to include budget as a control variable by discretizing budget values into appropriate budget bins. The problem with this approach was that we would need to take into account movie genre in the discretization, as movies from different genres can fall into very different ranges of budgets (e.g. action, adventure, and science fiction movies tend to cost much more in production compared to other genres), which meant we would need to create unique binning methods for the genres separately. This would lead to a high number of unique genre-budget groups having few movies. Furthermore, one movie could fall into multiple genres, further complicating the discretization. Thus, we decided to use the budget values as is.

statistically significant, and even then, all but one of such variables had negative coefficients (the only statistically significant genre that positively impacts the revenue is adventure). While this result made intuitive and practical sense – high-budget movies tend to accumulate high revenue, and movies will strive to earn, at a minimum, as much as the production costs – it limited our ability to study the inter-relationships of the different variables in our model.

We wanted to build a model that accounts for budget since it is an integral part of movie production while, at the same time, ensuring that budget alone does not overshadow the effects of other attributes of a movie. Thus, we ran a second regression model using profit as the dependent variable. Profit, computed by subtracting budget from revenue, was devised as an alternative metric to measure the financial success of movies. Using profit, we were able to both include budget in our model and restrain its disproportionate, overarching influence on the revenue. This second model incorporated ethnic diversity of cast, star power, and proportion of actors as the independent variables, in addition to genres as a control variables. Regression results of this model are summarized in Table 3. For this model, out of all the independent variables, diversity of cast and average star power were computed to be statistically significant, along with a number of genre variables. Specifically, diversity of cast had a negative coefficient, which can be interpreted that the more ethnically diverse a cast ensemble is, the lower the profit the movie might make. On the contrary, average star power had a positive coefficient, indicating that the popularity of the cast prior to the movie can positively impact the movie profit. A number of movie genres seem to have statistically significant impact in improving profit, including science fiction, musical, fantasy, action, and adventure, while a number of other genres seem to have the opposite effect, including comedy, horror, crime, and drama. While the regression results did not control for all possible confounding variables, such results could point towards potential research in studying the influence of the cultural and ethnic background of actors and actresses on audience reception.

Y = revenue	Coefficients	Standard errors	p
is_Family	-23,773,578.47**	11,354,862.71	0.036360
is_Music	1,715,074.23	14,222,444.29	0.904020
is_Romance	-5,857,063.92	7,652,465.03	0.444100
is_War	-54,177,651.79**	22,436,357.21	0.015800
is_Comedy	-16881947.58**	7,336,231.17	0.021440
is_Horror	7,854,587.64	9,317,970.54	0.399310
is_Sci-Fi	-919,713.91	10,018,789.89	0.926860
is_History	-44,504,402.36***	14,174,137.02	0.001700
is_Crime	-16,680,272.25**	7,145,256.68	0.019630
is_Western	-54,446,662.01*	31,226,632.45	0.081320
is_Mystery	-5,025,272.55	8,732,901.28	0.565030
is_News	13,808,445.84	106,483,674.67	0.896830
is_Musical	8,858,364.67	26,724,105.87	0.740310
is_Drama	-14,542,952.13**	6,813,365.09	0.032870
is_Biography	-6,127,566.91	9,853,773.10	0.534080
is_Sport	-37,316,552.76**	16,491,463.45	0.023710
is_Fantasy	-11,579,177.88	9,532,632.95	0.224570
is_Action	-34,049,730.87***	7,409,595.89	0.000000
is_Adventure	21,223,974.96**	8,379,624.69	0.011360
is_Documentary	11,325,660.47	32,302,751.80	0.725900
is_Thriller	-6,882,312.19	7,935,872.42	0.385870
Budget	3.24***	0.06	0.000000
Ethnic diversity of cast	7,312,188.85	10,643,960.64	0.492140
Average star power	-2,461,008.02	2,197,619.02	0.262850
Maximum star power	1,499,248.63	2,863,215.36	0.600570
Proportion of (male) actors	-13,986,226.94	11,818,420.59	0.236720
Intercept	17,665,054.83	19,937,909.17	0.375680
Observations	3,540		
Adjusted R^2	0.578		

Table 2: **Regression table of the model using revenue as the dependent variable and budget as one of the independent variables.** All genre variables are binary; budget and revenue are in dollars (thus the small scale of budget's coefficient compared to all other variables); ethnic diversity (Gini index) ranges from 0 to 1; star power (average and maximum) ranges from 0 to 10; proportion of (male) actors ranges from 0 to 1. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Discussion

Y = profit	Coefficients	Standard errors	p
is_Family	19,576,140.15	13,210,401.27	0.138460
is_Music	8,141,377.04	16,639,011.61	0.624660
is_Romance	1,108,545.20	8,950,557.51	0.901440
is_War	-41,523,861.56	26,247,416.19	0.113740
is_Comedy	-35,324,854.69***	8,562,489.77	0.000040
is_Horror	-34,319,674.07***	10,815,648.71	0.001520
is_Sci-Fi	65,442,128.24***	11,522,114.21	0.000000
is_History	-27,423,379.15*	16,574,516.88	0.098100
is_Crime	-27,140,746.74***	8,353,072.83	0.001170
is_Western	-55,918,757.99	36,535,262.95	0.125970
is_Mystery	2,207,965.23	10,214,830.18	0.828880
is_News	16,217,791.72	124,586,332.64	0.896440
is_Musical	59,137,174.05*	31,224,634.91	0.058320
is_Drama	-42,946,840.06***	7,918,092.21	0.000000
is_Biography	-10,061,340.51	11,528,248.74	0.382860
is_Sport	-22,953,352.22	19,289,441.49	0.234150
is_Fantasy	42,043,812.80***	11,016,366.05	0.000140
is_Action	18,031,918.56**	8,502,584.11	0.034010
is_Adventure	109,934,560.58***	9,371,159.71	0.000000
is_Documentary	-37,475,314.36	37,761,099.52	0.321060
is_Thriller	-8,845,279.33	9,284,785.91	0.340830
Ethnic diversity of cast	-35,533,424.62***	12,375,465.77	0.004110
Average star power	12,092,893.09***	2,527,388.33	0.000000
Maximum star power	2,945,535.41	3,349,644.85	0.379270
Proportion of (male) actors	11,182,159.81	13,803,413.13	0.417940
Intercept	18,488,951.83	23,327,424.10	0.428070
Observations	3,540		
Adjusted R^2	0.174		

Table 3: **Regression table of the model using profit as the dependent variable.** All genre variables are binary; profit is in dollars; ethnic diversity (Gini index) ranges from 0 to 1; star power (average and maximum) ranges from 0 to 10; proportion of (male) actors ranges from 0 to 1. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

The key findings of this study indicate that there is a gender imbalance in the film industry, with more actors than actresses across all genres. On average, actors also have longer careers than actresses, and a longer career length leads to an increase in the total number of movies for actors but not for actresses. The findings also suggest that over the years, there have been

more movies with a higher proportion of actors. These movies have generally generated higher revenues than movies with a higher proportion of actresses.

When it comes to ethnic diversity, the number of actors and actresses from groups other than white remains low over the years. There has been an upward trend in the Gini impurity index for ethnicity, indicating that movies are becoming more diverse. However, movies with a majority non-white cast have on average generated lower revenues than movies with a majority white cast and equally white and non-white cast. Actors are also more likely to be cast in the top roles, across all ethnicity groups, while actresses are more often cast in the second or third roles.

Regression analysis shows that budget has a disproportionately strong correlation with revenue. When profit is used as the dependent variable, the results indicate that ethnic diversity has a significant negative impact on profit, while average star power has a significant positive impact. The proportion of male actors has a positive but statistically insignificant impact on profit. Certain genres, such as fantasy, action, adventure, and science fiction, have a significant positive impact on profit, while others, such as comedy, horror, crime, and drama, have a significant negative impact.

The findings of this study have several implications for the film industry. The gender imbalance in the industry, with more actors than actresses and longer careers for actors, suggests that there may be gender discrimination in casting and career advancement. The lower revenues generated by movies with a majority non-white cast and the under-representation of actors and actresses from non-white groups suggest that there may be discrimination based on ethnicity as well.

These findings also have implications for the audience and for society at large. The lack of diversity in the film industry may limit the range of stories and perspectives that are represented on screen and may perpetuate stereotypes and prejudices.

From a business perspective, the findings of this study suggest that increasing diversity in the cast and crew of a movie may have a negative impact on revenue. However, this may not necessarily be the case, as the lower revenues could be due to other factors such as discrimination or lack of access to resources and opportunities. Additionally, the positive impact of certain genres on profit suggests that there may be opportunities for the film industry to diversify its offerings and appeal to a wider audience.

While our study provides insights into the gender and ethnic diversity of the film industry, it does have some limitations. One limitation is the availability and quality of the data used in the study. The budget and revenue data used in the analysis may not be a good representation of the overall film industry. Another limitation is the accuracy of the IMDb dataset used in the study. This dataset only lists an average of four main actors/actresses per movie, which may not be representative of the full cast in some cases. This can affect the analysis of the diversity of movie casts and the range of Gini impurity index values.

Additionally, the accuracy of the tools used for classifying ethnicity is a limitation of the study. The ethnicolr tool used in the analysis has limited accuracy and may lead to misclassification of actors and actresses into different racial groups. This could affect the results of the analysis and the conclusions drawn about the diversity of movie casts. Despite these limitations, the findings of this study still provide valuable insights into the gender and ethnic diversity of the film industry.

Methods

Data sources

The bulk of our data is sourced from IMDb's public datasets[2] which contain the basic information– such as year of release, genres, principal cast and crew, etc.– for all titles available on IMDb.

[2]https://www.imdb.com/interfaces/

From all available IMDb datasets, we mainly made use of the following:

- `title.basics.tsv.gz`: containing basic information for the titles including languages, title types (movie, short, TV series, TV episode, etc.), year of release, genres. This dataset also contains `tconst`s of the titles, which are unique IDs assigned to the titles that are used to merge data across different datasets.

- `title.principals.tsv.gz`: contains the principal cast and crew for the titles. Each row in this dataset corresponds to one principal cast or crew (denoted with a unique ID `nconst`) in a specific title (denoted with its `tconst`) and includes the person's category of job (actor, actress, director, writer, cinematographer, etc.).

- `title.ratings.tsv.gz`: contains the votes and ratings that the titles have received on IMDb.

Since the data available from IMDb does not contain information on the budget and revenue of the movies, we used data available from a second source, The Movie Database (TMDB). Data from TMDB is available through its free API, which also accepts search queries that use IMDb's `tconst`.

Data cleaning

Starting from the dataset of all titles available on IMDb (over 9 million titles), we first filtered out only those of type movies. This left us with over 600,000 movies, which we then filtered by our time window of interest (2000 - 2019, inclusive on both ends) to keep over 230,000 movies. We dropped all movies of genre animation, for these movies include only voice actors who do not appear on screen and thus are presumably irrelevant to our analysis. We also kept only movies that are in English.

From this pool of remaining movies, we used the TMDB API to query their budget and revenue data. Out of over 230,000 movies, 3,721 movies returned query results that include budget and revenue data. We conducted manual inspection on these movies to check for possible data entry errors. Some movies have seemingly infeasible budget or revenue, for example, movies with budget/revenue under 500 (USD). We manually checked other sources e.g. Wikipedia to fix such issues wherever more accurate data is available. Those that we could not find reliable information for, even after manual inspection, were dropped. We also dropped movies where there are no principal cast i.e. no actors and actresses are listed. Such movies included a large number of documentaries (where the people involved usually portray themselves without acting roles), which is the reason for the discrepancy between the proportions of documentaries in the original dataset and our finalized dataset. In the last filtering step, we kept movies that are in the 2.5 percentile of the original IMDB dataset in terms of votes received, so as not to give more weights to movies with high ratings but low number of raters. Our finalized dataset included 3,540 movies. From these movies, we detected 14,122 principal roles, corresponding to 5,273 unique actors and actresses.

Our finalized dataset of movies span 21 genres, where some of the genres with highest counts of movies include drama, comedy, action, and crime, whereas some of the genres with the lowest counts are news, western, documentary, and musical, among others. Figure 8 visualizes the counts by genres of movies in our finalized dataset (the columns corresponding to the left y-axis) compared to movies in the original IMDb dataset filtered by the same time window, 2000 - 2019 (the red line corresponding to the right y-axis). In general, the distribution of movies in our dataset follows the same trend as the original data, with the exception of documentaries, the reason for which was discussed above. The columns do not sum up to the same number of movies in our dataset (3,540) since one movie can be listed under multiple genres.

In addition to budget and revenue, to study the financial success of movies while also ac-

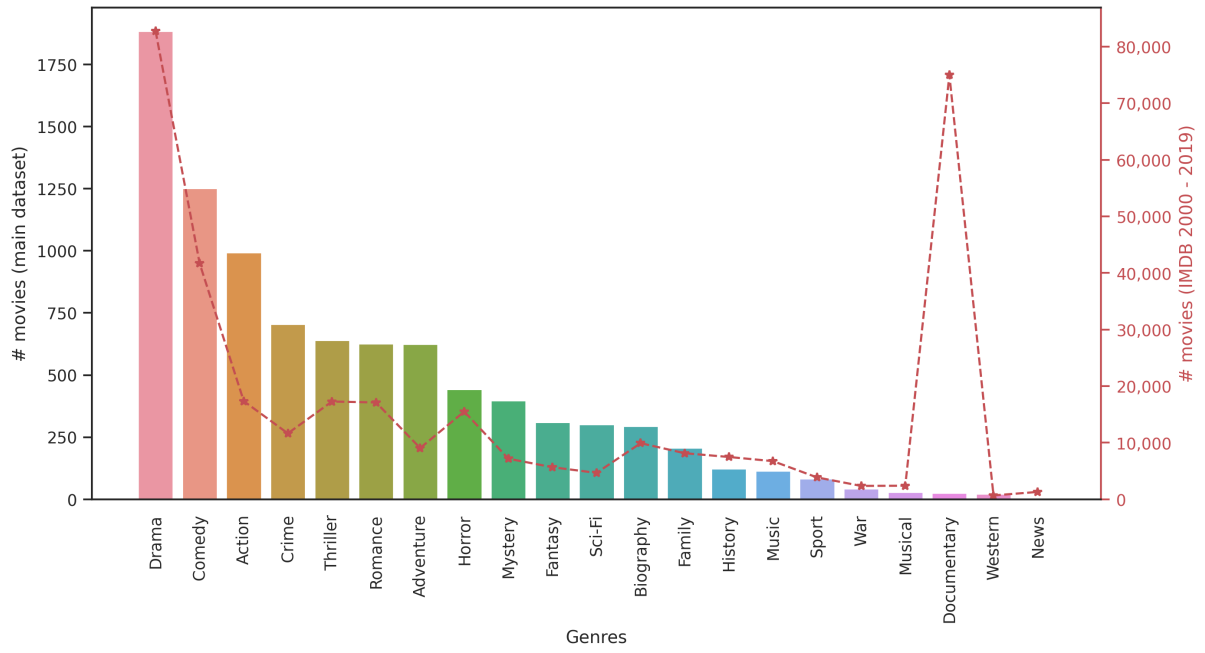

Figure 8: **Movie counts by genre.**

counting for their budgets, we included profit in our analysis, which is calculated by subtracting budget from revenue. Figure 9 includes the average revenue, budget, and profit of movies by years. It can be observed that since budget stays relatively stable through the year, the profit generally reflects the same trends as revenue. Supplemental Figure 7 visualizes the annual average budget and revenue for each genre.

Ethnicity prediction

We used the Ethnicolor API to identify the race of each principal cast member in our dataset. This API uses a combination of US census data, Florida voting registration data, and Wikipedia data to predict a person's race and ethnicity based on their first and last names. The API provides the following racial categories: White, Black, Latinx, Asian, Others, Indigenous, and more than two. However, we found that the API had some errors in predicting the race of certain cast members. To address this, we also used the Racial Lines dataset (*6*) (available at `https:`

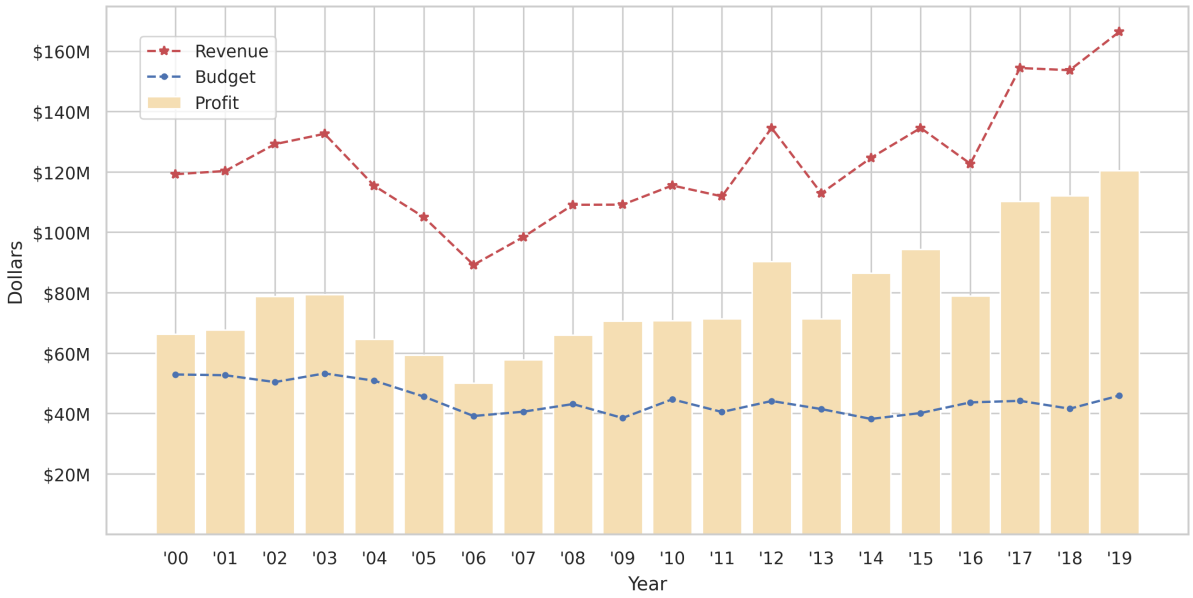

Figure 9: **Average budget, revenue, and profit by year.** For a certain movie, profit = revenue - budget. Both revenue and budget considered are adjusted for inflation.

//doi.org/10.7910/DVN/KERZQY) as a reference. We kept the classifications of entries in the Racial Lines dataset that overlapped with our dataset, resulting in a final dataset of 1267 entries from the Racial Lines dataset and remaining 2273 entries from the Ethnicolor API. In addition, we merged the East Asian and South Asian categories in the Racial Lines dataset into a single Asian category for consistency.

In cases where the Ethnicolor API returned ”Others” or ”More than two” races for a cast member, we used additional ethnic group data to further sub-categorize them as either Hispanic and Latino or not. This allowed us to more accurately determine the race of each cast member in our study. Supplemental Table 1 provides the final count of principal cast members classified into different races.

Gini Impurity Index

The Gini impurity index is used as a measure of ethnic diversity of principal cast within a movie. Let Pi denote the set of principal cast for movie M_i. The formula for the Gini impurity index for a movie M is given by:

$$G(M_i) = 1 - \sum_{i=1}^{n} p_i^2 \tag{1}$$

where n is the number of unique races of principal cast for the movie M_i and p_i is the proportion of the race in the movie.

For instance, for any given movie, M_i, we could have: [eth(P_i) : Principal Cast(p)] = [Asian; Asian; Black; White]. By using the equation above, we can calculate the Gini impurity index for the movie M_i by first finding the number of unique races in the principal cast, which in this case would be three (Asian, Black, and White). Next, we would calculate the proportion of each race in the cast by dividing the number of individuals of each race by the total number of principal cast members. For the example above, the proportions would be 0.5 for Asian, 0.25 for Black, and 0.25 for White. Finally, we can plug these values into the equation to calculate the Gini impurity index for the movie M_i. In this case, the Gini impurity index would be

$$G(M_i) = 1 - (0.5^2 + 0.25^2 + 0.25^2) = 0.625$$

Star power

Since there is no standardization for star power, we devised our own metric to measure the popularity of the cast of a movie prior to that movie based on movie ratings available in the IMDb datasets. First, let $m_{j,C}$ denote a movie released at year j and has the list of principal cast C. For an actor or actress a at a certain year y, their star power$_{a,y}$ is defined as the average ratings of movies where movies $= \{m_{j,C} | j < y, a \in C\}$. Then, for a certain movie $m_{j,C}$, we

computed the average and and maximum of the star power$_{a,j}$ for all $a \in C$. An actor or actress with no prior principal role prior to a certain year (i.e. the actor/actress was not listed in the principal cast for any movie before that year) is assigned a star power of 0. Supplemental Figure 8 visualizes the average star power by year for each genre.

Inflation adjustment

Since our analysis spans a 20-year window, we adjusted the budget and revenue of all movies to account for inflation. We converted all monetary values to the baseline of 2019. See Supplemental Figure 9 for detailed graphs comparing the average original and adjusted budget/revenue for each year. Inflation adjustment was conducted based on the consumer price indices (CPIs) of the United States from 2000 to 2019 obtained from the International Monetary Fund (*7*).

References and Notes

1. N. A. Pangarker, E. Smit, The determinants of box office performance in the film industry revisited (2013).

2. G. Selvaretnam, J. Yang, Factors Affecting the Financial Success of Motion Pictures: What is the Role of Star Power? (2015).

3. G. Bae, H. jin Kim, The impact of movie titles on box office success (2019).

4. S. Lee, J. Y. Choeh, The impact of online review helpfulness and word of mouth communication on box office performance predictions (2020).

5. D. Hunt, A.-C. Ramón, M. Tran, 2016 Hollywood Diversity Report: Business as Usual? (2016).

6. V. Svaikovsky, A. Meisner, E. Kraicer, M. Sims, Racial Lines (2018).

7. International Financial Statistics - International Monetary Fund Data, https://data.imf.org.

Author contributions

Conceptualization: Ngoc Hoang, Maryam Khalili, Alem Shaimardanov, Sashank Silwal; Methodology: Ngoc Hoang, Sashank Silwal, Alem Shaimardanov; Data collection and processing: Ngoc Hoang, Maryam Khalili, Alem Shaimardanov, Sashank Silwal; Visualization: Ngoc Hoang, Maryam Khalili, Alem Shaimardanov, Sashank Silwal; Writing: Ngoc Hoang, Maryam Khalili, Sashank Silwal