Possible Bias in the Film Industry

By Aria Fisher

Abstract:

In this project, I explore two datasets featured in articles about the film industry. They contain data on films and were sourced from GitHub. Each set includes, among other variables, the title, year of release, and the amount of money that the films grossed. The datasets are especially unique in that one accounts for whether a film has passed the Bechdel test, wherein two female characters are named and converse about a topic other than a man, and the other focuses in on biopics specifically and accounts for whether the lead actor is a person of color. The purpose of this project was to discover how, if at all, representation in films correlates with their performance in terms of profits as well as to simply observe general trends concerning representation in the film industry.

Acknowledgement:

This project would not have been possible without the teaching of Professor Herath and the availability of datasets on GitHub. I would also like to thank the authors of the articles, Hannah Fingerhut and Walt Hickey, that used this data, and which inspired my own exploration in this project.

Introduction:

As aforementioned, the data includes basic information about a copious number of movies. However, each dataset also includes information that is superfluous to my goals in the project and has occasional missing data for some films. Thus, my approach to analyzing this data first centers around cleaning it by removing rows with missing data, dropping columns with unnecessary information, and formatting the data for clarity and readability. The basic questions that I am pursuing in my analysis are broadly about correlation between variables (e.g., time and representation) so the next element of my approach is to make sure the variables are compatible in terms of finding their relationship to each other. This may entail changing the type of some variables from character to numeric. Then, to further filter down the amount of information I am dealing with, subsets are created with certain conditions, like a subset where the only movies included are the ones that pass the Bechdel test. Now that everything is sufficiently broken down, the final part of my approach is visualization, and in some cases the calculation of relevant mathematical figures.

My report does not reflect this entire approach, because it is focused on presenting any findings I have, so it is sectioned into different visualizations and occasionally numerical findings. For each result, I will explain the combination of data involved, but the focus will be on what the graphic representation or numerical finding demonstrates about the relationship between certain variables. All that can safely be concluded is correlation, but I find that full analysis requires postulation of cause-effect relationships that could reasonably exist within these datasets, so a bit of sociological involvement will take place as well, more prevalently in the conclusion of this report.

Results:

The following section of results have been extrapolated from the biopics dataset and analyzed with SAS.

Table

Description automatically generated

This table visualizes the frequency distribution of where the biopics in the dataset have been made, with 58.85% being made in the US and about 18% being made in the UK. Table

Description automatically generated

Chart

Description automatically generated

The above table (and the accompanying visualization of its results) uses the categories “Inclusive” and “Exclusive” to represent whether a biopic has been categorized as having a person of color as the leading actor or not, and it then presents the frequency of each instance within the dataset. This shows that 82.68% of the biopics have a white person as the leading actor, while 17.32% have a person of color as the leading actor. It must be considered that two of the movies did not have data on the race of the leading actor, but the figures remain significantly different despite that small confounding element. The frequency plot visualizes how steep that difference is.

Chart, bar chart, histogram

Description automatically generated Table

Description automatically generated

This table and frequency plot represent the distribution of subject matter for biopics that are led by actors of color. Besides the “other” category, which is vague and thus largely unhelpful for drawing conclusions, the most frequently covered subject matter for these biopics appears to be about figures who are activists, athletes, and musicians. The table confirms that, as the percent frequency of those subjects, in order, is 14.67%, 14.67%, and 13.33% of all the POC-led biopics in the dataset.

Chart, bar chart

Description automatically generated Table

Description automatically generated

Finally, this frequency table and its graphic visualization represent the distribution of subject matter for biopics wherein the leading actor is a white person. There is a larger number of subjects covered than the previous subject distribution for POC-led films, and the visualization demonstrates that, besides the ever-popular “other” category, the most frequent subjects for these biopics are criminals, athletes, and authors. The frequency table lends values to those visual observations, as criminal biopics make up 11.45% of the white-led movies in the dataset, athletes make up 9.78%, and authors make up 9.22%.

The following results have been extrapolated from the Bechdel dataset and analyzed with R.



Chart, scatter chart

Description automatically generated

The calculation of the correlation coefficient for domestic and international profits within the dataset was found to be about 0.94. That, coupled with the scatterplot and regression line relating the two variables, demonstrates a strong positive correlation between domestic and international profits. Thus, for the following calculations and visualizations, international profits were used where most figures of profit are concerned because it can be understood that the value will be closely alike to that of domestic profits and will not be a confounding factor in analyzing the general relationship between profit and other variables.

Chart, histogram

Description automatically generated

This line graph compares international profits with the year of release, but only for the movies within the data set that were categorized as passing the Bechdel test. There appears to be a steep increase beginning around 1990, culminating with a large spike in profits around the year 1996, and then a stagnation of that increase with profits leveling off at around 1.0e+09 dollars.

Chart, histogram

Description automatically generated

This line graph compares international profits with the year of release, but only for movies that were categorized as failing the Bechdel test. There appears to be a steady, small-growing increase in profits over time, with one sharp increase in profits around the year 2009.

A picture containing text

Description automatically generated

The figures above represent the average values of the release year, domestic profits, and international profits for different subsets of the Bechdel data. First, for the entire dataset, the average year of release is about 2003 and the average international profits are about 1.5e+08. For movies that specifically fail the Bechdel test, their average release year is around 2002 and their average profits are about 1.6e+08. Finally, for the movies that pass the Bechdel test, the average release year is about 2003 and their mean international profits are around 1.3e+08.

Text

Description automatically generated

Lastly, the figures above represent the calculation of the percent of movies, both for the subset of the data with Bechdel test failures and the subset with Bechdel test passes, that had profits higher than the overall average profits for the entire dataset. Because 350 out of the 1004 movies that failed the test profited more than the overall average, about 35% of Bechdel test failures had above average profits. For those movies that passed, only 216 of the 825 profited above the overall average, so about 26% of the movies that passed the Bechdel test had profits that high.

Conclusion:

Through analysis of the biopics dataset, it was demonstrated that a significantly small percentage of those films feature people of color as the lead actors, 17.32% to be exact. This means that there is a gap of about 65 percentage points between the biopics that do represent people of color and those that do not (at least not as leading actors). This disparity points to a well-known practice of exclusion in the entertainment industry, but also, because these movies are biopics, it points to a general understanding of history and the world around us as one that only occasionally has a person of color important enough to be the subject of a film. Furthermore, the examination of subject matter distribution for POC-led versus white-led movies demonstrated that there is a disproportionate focus on certain stories for people of color. Not only were there simply less biopics, those that were categorized as inclusive were largely about activists, athletes, or musicians. While athletes were still a significant portion of white-led biopics, the subject was only the focus for 9.78% of those films, in comparison to the 14.67% of POC-led films that focused on athletes. Musicians were also prevalent in both, but still were more frequent in POC-led films. One staggering difference in subject matter was found in biopics about activists: while 14.67% of the inclusive films covered the story of activists, only 3.91% of non-inclusive films centered on that subject. Part of these differences may be due to the smaller breadth of subjects covered by biopics led by people of color, because with a smaller total, it takes fewer films to get high percentages in certain categories. However, the difference is still stark, and the lack of breadth in subject matter also speaks in its own way to the exclusion and reduction of people of color to certain acceptable stereotypes. Perhaps only certain athletes, activists, or musicians are seen ss people of color worthy of commemoration, because there is a historical misconception that POC have not been exceptional in any other field. Either way, it points to a need for deep inclusivity and more mindful film-making that seeks to remedy biases in the way people of color are perceived.

Examination and analysis of the Bechdel test dataset revealed that, at first glance, movies that pass the Bechdel test do a bit worse at the box office than those that do not. On the scale of billion-dollar profits, the amount is relatively small, but it is still a difference. The average international profit for movies that fail the test is about 1.6e+08 dollars, while it is about 1.3e+08 dollars for those that pass. The overall average international profit for every movie in the dataset is around 1.5e+08, and while 35% of movies that fail the Bechdel test manage to go beyond that profit mark, only around 26% of those that passed had above-average profits. However, two confounding factors must be noted before one draws the conclusion that movies are simply worse if they include women who have names and talk to each other about subjects other than men. Firstly, the graphic representations of international profits over time demonstrate that this dataset spans from 1970 to 2013 and, keeping in mind that the Bechdel test wasn’t even invented until 1985, there is a strong leftward skew of the graph showing movies that pass the test. Until around 1994, coinciding with the rise of third-wave feminism, movies that passed the Bechdel test were far from reaching the benchmark of 1e+09 dollars that the graph shows them consistently achieving after the ‘90s. That period of pre-feminism likely skews the average profit data, causing movies that pass to appear less profitable. Of course, there is a leftward skew of the graph showing profits over time for the movies that fail, but even those smaller, earlier values were closer to 1e+09 dollars than those of the Bechdel movies, thus demonstrating that the average of the test-failing movies is not skewed to the extent of the test-passing ones. Secondly, bias in audiences, rather than quality of film, may be what affects the profits negatively. In some cases, if a film features prominent female characters and appears to likely pass the Bechdel test, the bias of moviegoers can prevent them from paying to see the film. This has nothing to do with the quality of the movie, but simply with the state of society. Even if the pre-feminism skew is ignored and it is accepted that movies featuring well-written female character do worse at the box office, one must recognize the flawed social attitudes that can play a part in that. Just as profits for test-passing movies rose with third-wave feminism, they may rise again as social attitudes towards women shift positively in the future.

References:

Fingerhut, Hannah. “'Straight Outta Compton' Is the Rare Biopic Not About White Dudes.” *FiveThirtyEight*, FiveThirtyEight, 14 Aug. 2015, https://fivethirtyeight.com/features/straight-outta-compton-is-the-rare-biopic-not-about-white-dudes/.

Herath, Wiranthe. “Class Notes, Weeks 3-12.” 20 November 2022, STAT 040, College of Business & Public Administration, Drake University. Microsoft PowerPoint presentation

Hickey, Walt. “The Dollar-and-Cents Case Against Hollywood's Exclusion of Women.” *FiveThirtyEight*, FiveThirtyEight, 1 Apr. 2014, https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/.