Female Protagonists: A study on how the film industry and audience treat them

Sarah Pratt

Data 211 Final

05/10/22

Introduction

When I was young one of my favorite movies was *Mulan*, then as a young adult I loved *10 Things I Hate About You*, and as I became an adult I fixated on *Jane Eyre*. Movies like these with a female protagonist were incredibly important in shaping who I am, and shaping countless other young girls like me. Which is why I am using the movies dataset from Kaggle [1] to analyze differences in revenue, budget, and popularity between movies with a female protagonist and movies with a male protagonist. The purpose of this study is to see if there are systemic differences in the way both the movie industry and audience treat female-led movies. I believe it is extremely important for children and young adults to be able to connect with and look up to characters they see in their favorite movies. Therefore, if there are significant differences in the budget, revenue, or popularity of movies with a female protagonist versus a male protagonist, work needs to be done in the industry to correct this discrepancy for women.

Methods

To distinguish which movies have a female protagonist versus a male one, I used the "overview" column of the movies dataset, which is a short plot summary. Using this column, I made a sub-set of the data frame by searching for instances of certain keywords such as "he", "her", "she", "himself", etc. And to ensure that no movies overlap I removed movies with overviews that contain male keywords for the female protagonist data frame, and vice versa. After this process, the dataset contained a larger amount of movies with a male protagonist. To further preprocess the data needed for the visualizations I also scaled the budget variable to be in millions of dollars and added the scaled version as a new column in the data frame. For all three variables, budget, revenue, and popularity I dropped any rows containing zeroes. And for the revenue and popularity visualizations I used a logarithmic transformation to show the spread of data more easily since the data was largely clustered on low values with some large outliers.

Results

For each visualization I have a color scheme with pink and blue denoting female and male, as well written labels for the female and male data. The number of movies with female protagonists vs. male protagonists are also provided, and for every figure more male-led movies are contained in the dataset.

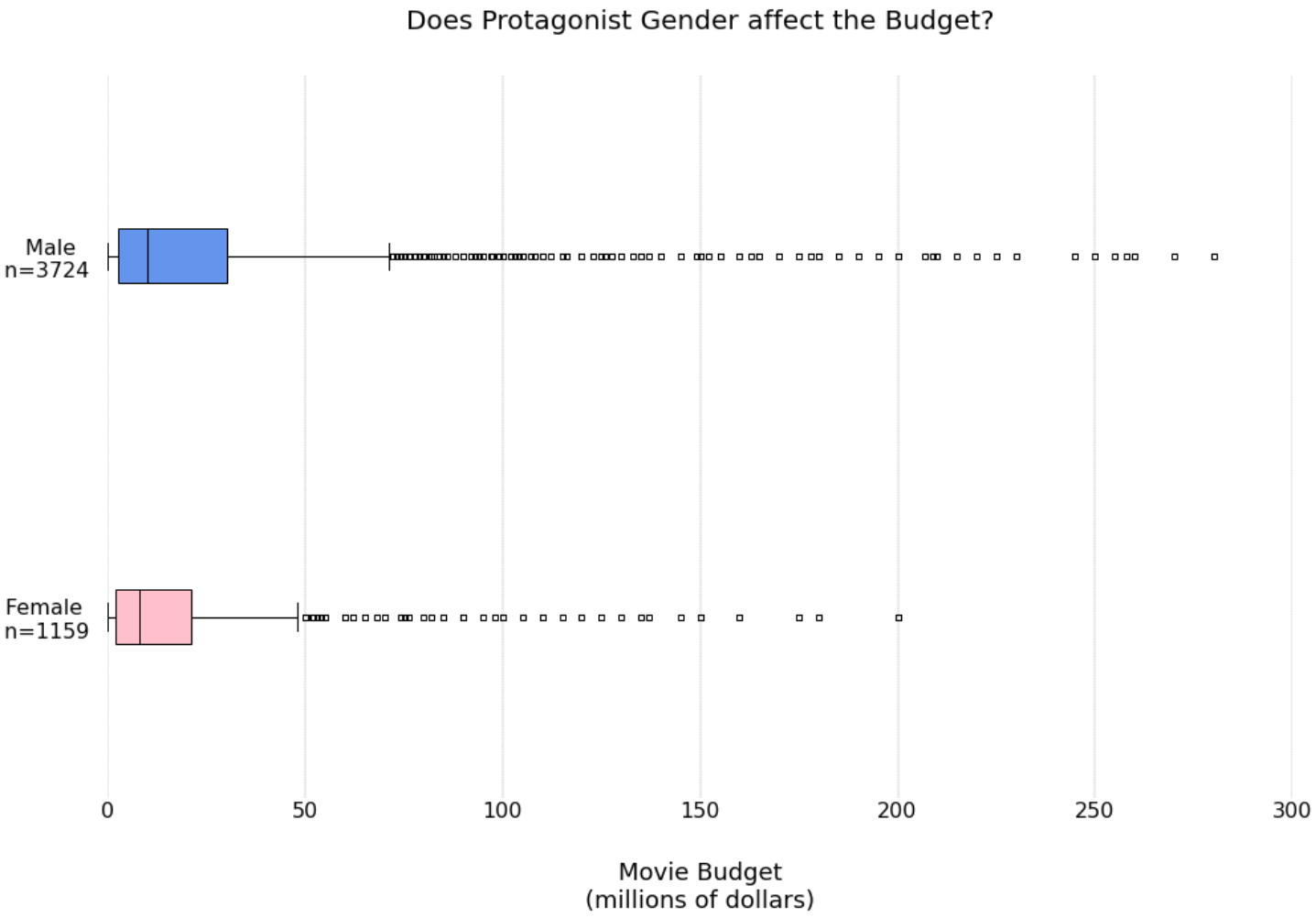

Figure 1: Two boxplots showing the budget for movies with female protagonists and male protagonists. Budget is scaled to be in millions of dollars.

Figure 1 shows the difference in budget for movies with female protagonists and male protagonists. The discrepancy is clear to see with the Interquartile Range

(IQR) of the boxplots displaying a larger spread for male-led movies, as well as the outliers which are much larger for the movies with male protagonists. The IQR for movies with female protagonists is around 0-50 million dollars, whereas the IQR for male-led movies seems to be 0-75 million dollars. The largest outlier for female-led movies is about 200 million dollars, whereas the highest budget for a male-led movie is around 280 million dollars.

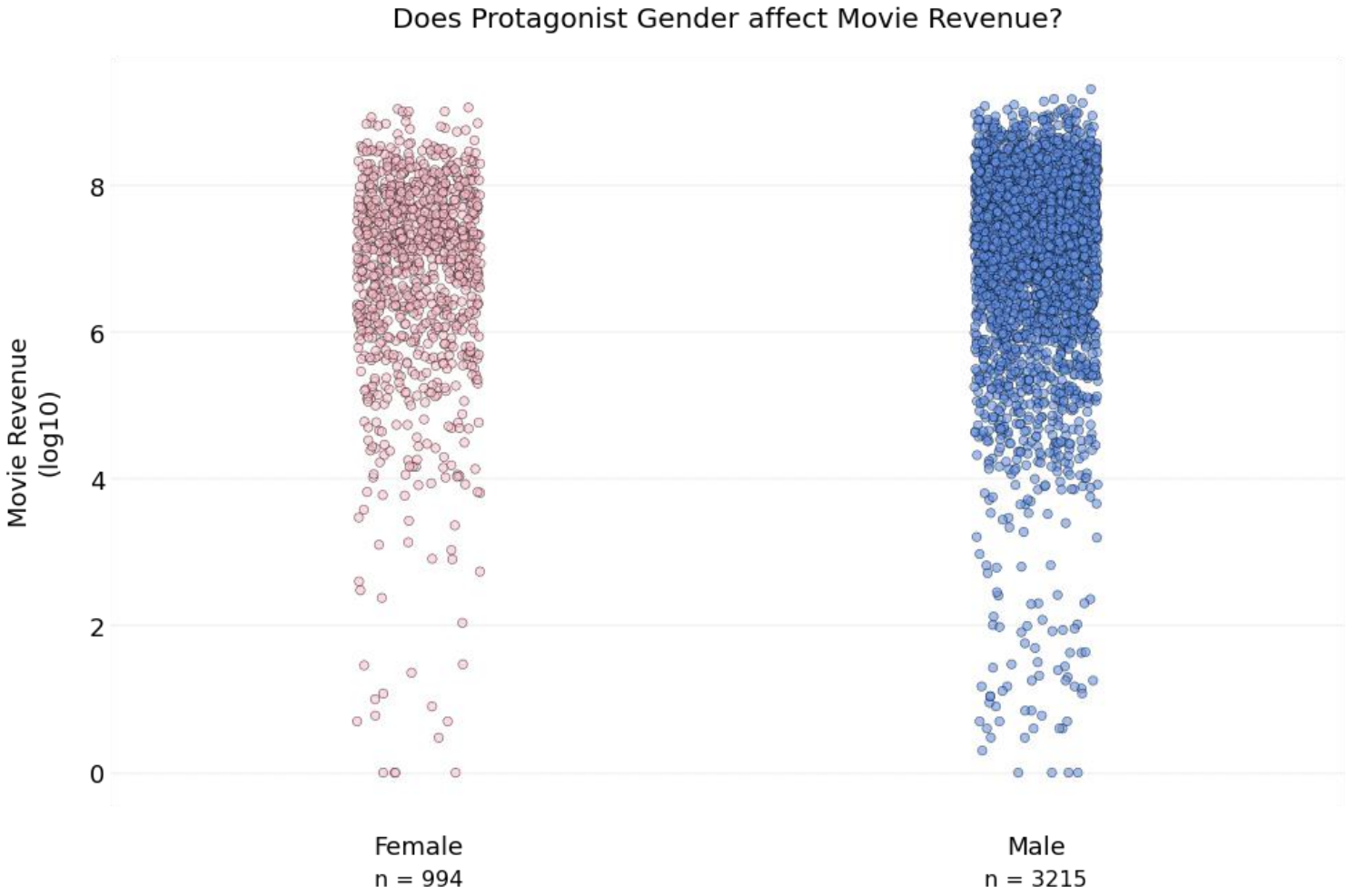

Figure 2: Two strip charts showing the revenue for movies with female protagonists and male protagonists. A logarithmic transformation is used for the revenue variable.

Figure 2 shows little difference in revenue for movies with a female protagonist versus movies with a male protagonist. The spread of data in these strip charts is

about the same between female and male-led movies. Generally, for both female and male-led movies the data spans about 8 degrees of 10, with most of the data being clustered in the range 10^6–10^8. The only noticeable difference between the two charts is the number of points, caused by the amount of movies with female protagonists in the data frame being smaller after data preprocessing.

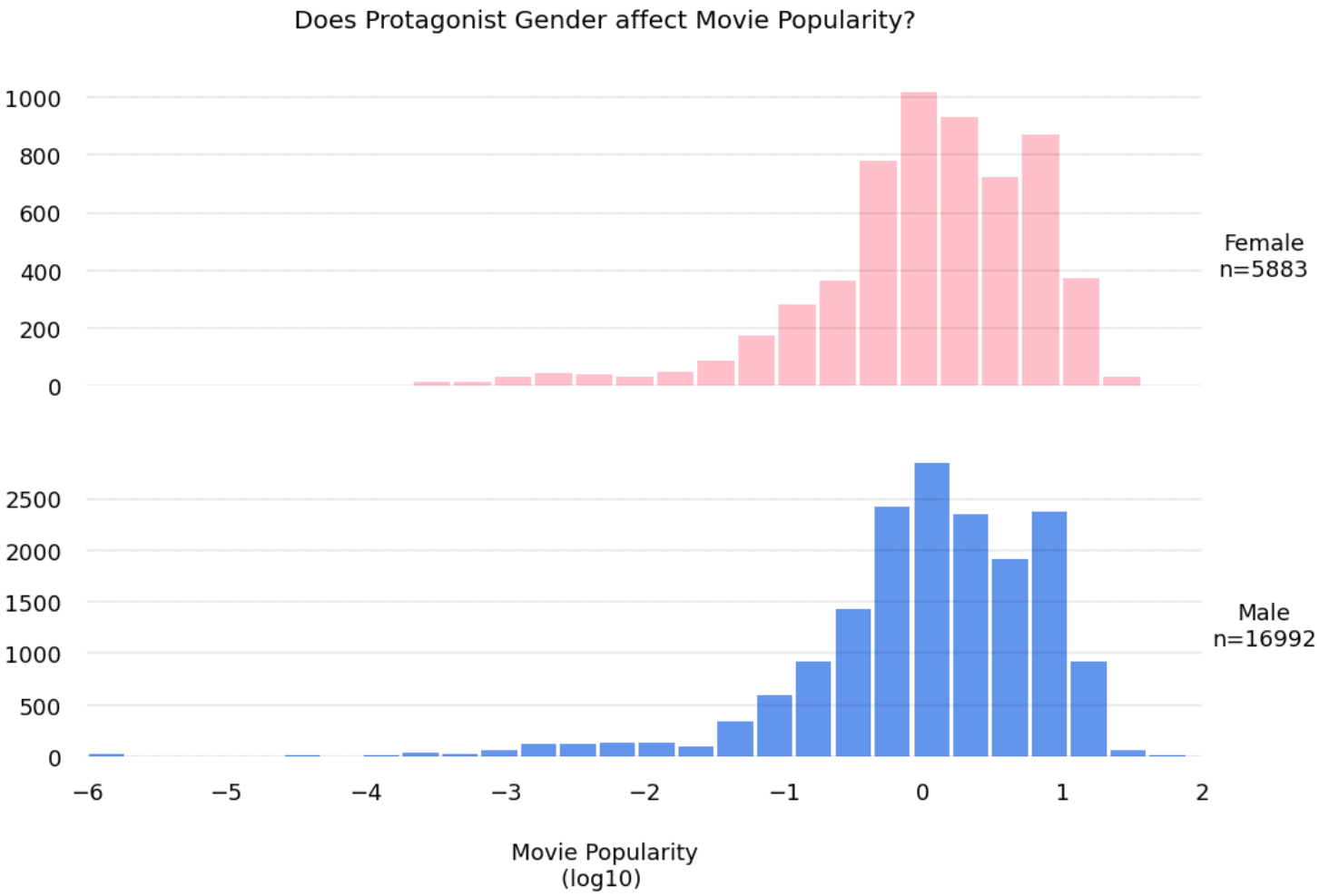

Figure 3: Two histograms showing the popularity for female and male-led movies. A logarithmic transformation is applied to the popularity variable. The count scales for the histograms are different because there are more movies with a male protagonist.

Figure 3 shows little difference between female-led and male-led movies, although it is perhaps hard to discern because of the discrepancy between the

number of male and female-led movies that contain popularity data. The overall shape of the histograms are very similar, however the movies with a male protagonist have more outliers in both directions.

Discussion

The overall objective of this study was to see if there are systemic differences in the way both the movie industry and audience treat female-led movies by way of three variables: budget, revenue, and popularity.

From Figure 1 I can conclude that the film industry does treat movies with female protagonists differently. The film industry assigns less value to a movie with a female lead than a movie with a male lead. However, Figure 2 shows that revenue is largely unaffected by protagonist gender. Although the industry might assign lower value to female protagonists, movie audiences do not. And finally Figure 3 shows little difference in the range of popularity for movies with either gender, however the difference in the amount of movies in this dataset with female vs. male protagonists shows the much higher prevalence for male-led movies in the industry. Ultimately these figures show that there is a systemic difference in the way the movie industry treats films with female protagonists, however based on both revenue and popularity it seems that movie audiences treat them similarly.

References

1. Rounak Banik. The Movies Dataset, 2017. https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset