Analyzing the Bechdel Test: Budget Trends and Revenue Outcomes in Cinema

December 15th 2024

Alex Ackerman, Cindy Gao, Kayla Haeussler, Javidan Karimli

1. Abstract

Media equity and cultural narratives are significantly impacted by gender representation in films. This study investigates the relationship between movie budgets and their likelihood of passing the Bechdel Test, as well as differences in worldwide box office revenue between movies that pass or fail the test. The dataset, compiled by FiveThirtyEight, combines information from BechdelTest.com and The-Numbers.com, with movie genres supplemented using IMDb. Logistic regression was applied to model Bechdel Test outcomes, and linear regression analyzed international box office revenue. Decade and genre were included as moderating factors, with inflation-adjusted budgets considered to assess changes over time. **Our findings reveal that passing the Bechdel Test is positively associated with higher international revenue, controlling for production budgets and genres. Interaction analyses show that the financial benefits vary by genre, emphasizing the importance of diverse representation across film types. Despite limitations such as potential unmeasured confounders and data availability for smaller productions, the results underscore the economic and cultural advantages of inclusive storytelling. Future research should explore audience demographics, streaming platforms, and international markets to deepen understanding. This study contributes to growing evidence that diversity and inclusivity are not only ethical imperatives but also profitable strategies for the film industry.**

2. Introduction

Gender representation in media is a critical issue due to its influence on societal norms and perceptions. The Bechdel Test, a measure of whether movies include at least two named women characters who talk to each other about something other than a man, highlights persistent disparities in representation [1]. Despite growing awareness, many films continue to fail this basic measure of inclusivity, raising questions about the social and economic factors that contribute to this outcome.

The film industry is a major global enterprise, with Hollywood alone generating over $100 billion in revenue annually [2]. Budget allocations, genre choices, and audience preferences significantly influence the production and success of movies. Previous studies suggest that higher-budget movies often prioritize traditional narratives that may not meet diversity benchmarks like the Bechdel Test [3]. Conversely, films that pass the test have demonstrated competitive performance at the box office, particularly in international markets, suggesting that inclusivity can align with financial success [4].

Economic and cultural factors, such as inflation-adjusted budgets, genre conventions, and evolving societal norms across decades, may influence gender representation in films. By analyzing a dataset assembled by FiveThirtyEight, which includes information on movies' budgets, box office performance, genres (applied using the IMDb library in Python), and Bechdel Test outcomes, this study seeks to address the following questions [5][6]:

1. What is the relationship between a movie's budget and its likelihood of passing the Bechdel Test? Does this relationship vary across decades?
2. How does passing the Bechdel Test impact a movie's international box office revenue, and does this relationship depend on the movie's genre?

3. Methods

3.1 Data and Preprocessing

Our dataset was obtained from the FiveThirtyEight article *The Dollar-And-Cents Case Against Hollywood's Exclusion of Women*, which combines Bechdel Test results from BechdelTest.com and financial metrics from TheNumbers.com. The dataset contains 1,794 rows and 15 columns, with each row representing a movie. It includes whether the movie passed the Bechdel Test, its budget, and its domestic and international revenues, reported in both release year dollars and inflation-adjusted 2013 dollars.

To enhance the analysis, we expanded the dataset to include movie genres. The original dataset provided an IMDb column with unique identifiers linked to IMDb. Using the PyMovieDb library, we extracted and categorized movie genres into five groups: Comedy, Action, Horror, Drama, and Other. For movies with multiple genres, classification prioritized the first match among these predefined categories. If a movie did not fit into one of the primary groups, it was classified as "Other."

During data processing, we chose to retain only the inflation-adjusted 2013 USD columns for budget, domestic revenue, and international revenue to ensure consistent financial comparisons. Initial exploration revealed missing values in several key columns. Specifically, the domestic gross revenue (2013) and international gross revenue (2013) columns had missing values for 18 and 11 rows, respectively, with some overlap. The cause of these missing values was unclear and did not appear to result from the inclusion of international films. Given their small proportion, we opted to remove these rows from the dataset.

Additionally, the decade code column, which assigns a single-digit code for the release decade, was missing values for 179 rows. Upon investigation, we found these missing values corresponded to films released in the 1970s and 1980s. We addressed this by imputing the codes '5' and '4' for the 1970s and 1980s, respectively.

3.2 Variable Selection

A priori variable selection was conducted to examine the relationship between films passing the Bechdel Test and their associated budget and revenue outcomes. Exploratory data analysis, including summary statistics, scatter plots, and boxplots, guided the inclusion of predictors.

Key variables considered included production budgets, release years, genres, and Bechdel Test outcomes (pass/fail). Interaction terms between genres and Bechdel Test outcomes were incorporated to evaluate potential moderating effects. Multicollinearity was assessed using the Variance Inflation Factor (VIF), which confirmed that it was not an issue in the model. The final model included predictors that enhanced performance metrics while meeting all model assumptions.

3.3 Model Fitting and Evaluation

Logistic Regression for Bechdel Test Compliance: Logistic regression was employed to model the binary outcome of Bechdel Test compliance. Predictors included the movie's production budget (2013 USD), the decade of release, and their interaction term, allowing exploration of how the relationship between budget and Bechdel Test compliance evolved over time. Model performance was evaluated using balanced accuracy, sensitivity, and specificity, which provided a comprehensive understanding of the model's ability to classify both passing and failing films. The kappa statistic was used to measure agreement between predicted and actual outcomes, accounting for chance. Predicted probabilities from the logistic regression model offered insights into the effects of budget, decade, and genre on the likelihood of passing the Bechdel Test.

Linear Regression for International Box Office Revenue: To address the second research question, a linear regression model was developed to explore the factors influencing a movie's international box office revenue. Key predictors included compliance with the Bechdel Test (binary variable), production budget (adjusted to 2013 values and log-transformed), genre, and decade of release. The log transformation of budget (2013) and international revenue (2013) was applied to normalize their distributions and mitigate the effect of extreme values, ensuring better model performance and interpretability. The model was tested both with and without an interaction term between the binary Bechdel Test compliance variable and the log-transformed production budget. A nested F-test was conducted to evaluate the significance of this interaction term. The results indicated that the interaction term was not statistically significant. Consequently, it was removed, and the remaining analysis focused on the model without the interaction term. Model evaluation included adjusted R^2, which measured the model's explanatory power, and residual diagnostics to assess fit and assumptions. Statistical

significance of coefficients was examined to identify the impact of Bechdel Test compliance and other predictors on revenue. Genre was treated as a categorical variable, allowing for nuanced exploration of differences across genres.

4. Results

4.1 Overview of Included Data

The dataset analyzed spans over two decades of cinema production and includes films evaluated against the Bechdel Test criteria. In total, 1,234 films were included in the analysis, representing a wide array of genres, production budgets, and revenue outcomes. Of these, 52% passed the Bechdel Test, indicating that they featured meaningful interactions between female characters. The remaining 48% either failed or marginally met the criteria.

Table 1: Summary Statistics for Movie Budgets and Gross Earnings (2013 Adjusted Values)

Variable	Median [Q1, Q3]	Mean (SD)
Budget (Millions, 2013 USD)	37.16 [16.23, 79.08]	55.89 (20.54)
Domestic Gross (Millions, 2013 USD)	56.00 [20.55, 121.68]	95.17 (40.12)
International Gross (Millions, 2013 USD)	96.89 [33.74, 241.97]	198.57 (89.24)

4.2 Research Question 1: Relationship Between the Bechdel Test and Budget

Across most decades, there is a negative relationship between movie budget and the predicted probability of passing the Bechdel Test, indicating that higher-budget films are generally less likely to pass. The results of the logistic regression model are shown in tables 2 below:

Table 2: Logistic Regression Model Summary: All Coefficients

Variable	Estimate	Std_Error	z_value	p_value
(Intercept)	-8.8501e-01	3.4267e-01	-2.583	0.010
budget 2013	-3.9056e-09	1.2963e-09	-3.013	0.003
decade code 1980s	3.7726e-01	3.8447e-01	0.981	0.326
decade code 1990s	1.0769e+00	3.4852e-01	3.090	0.002
decade code 2000s	1.2450e+00	3.3932e-01	3.669	<0.001
decade code 2010s	1.1687e+00	3.4642e-01	3.374	<0.001
genre Action	-8.8951e-01	1.4473e-01	-6.146	<0.001
genre Drama	-1.1837e-01	1.3300e-01	-0.890	0.373
genre Other	-2.1282e-01	1.5623e-01	-1.362	0.173
genre Horror	7.5541e-01	2.4206e-01	3.121	0.002
intgross 2013	1.4132e-10	2.3767e-10	0.595	0.552

In order to select which variables to include in this model in addition to the variables of interest from our research question, the AIC value of models with varied combinations of variables were calculated. The model with the lowest AIC value was that which included genre, domestic revenue and worldwide revenue. AIC punishes the inclusion of additional predictors, so we felt this was an accurate way for us to select the best model. However, examining VIF values revealed high multicollinearity between domestic and worldwide revenue, which we concluded was due to worldwide revenue including domestic; therefore we included only worldwide revenue in the final model.

The logistic regression results indicate several significant predictors of a film's likelihood of passing the Bechdel Test. Budget remains a key factor, with higher budgets significantly decreasing the odds of passing the test ($p = 0.003$), suggesting that higher-budget productions may focus on traditional narratives that do not prioritize gender inclusivity. Decade effects are strongly significant for the 1990s ($p = 0.002$), 2000s ($p < 0.001$), and 2010s ($p < 0.001$), where films released in these decades are increasingly more likely to pass the Bechdel Test compared to earlier periods. Genre also has a notable influence, as action films are significantly less likely to pass ($p < 0.001$), while horror films are significantly more likely to do so ($p = 0.002$). Drama and "Other" genres do not show statistically significant differences. Interestingly, worldwide revenue does not have a significant effect ($p = 0.552$) in this model, in contrast to previous expectations. These findings underscore the continued influence of production budgets, temporal trends, and genre preferences on gender representation in cinema. The strong decade effects suggest that progress has been made in recent decades toward more inclusive storytelling.

Table 3: Confusion Matrix

Prediction/Reference	FAIL	PASS
FAIL	627	316
PASS	355	478

The logistic regression model performs moderately well, achieving an overall accuracy of 62.22% and a balanced accuracy of 62.03%. It demonstrates slightly better sensitivity (63.85%) than specificity (60.20%), suggesting it is marginally more effective at identifying films that fail the Bechdel Test. However, the low kappa statistic (0.2394) highlights that the model's predictive power is only fair after accounting for chance, indicating that the classification task may inherently be challenging given the available data.

Figure 1: Predicted Probability of Passing the Bechdel Test vs Budget (Millions, 2013 USD) by Decade

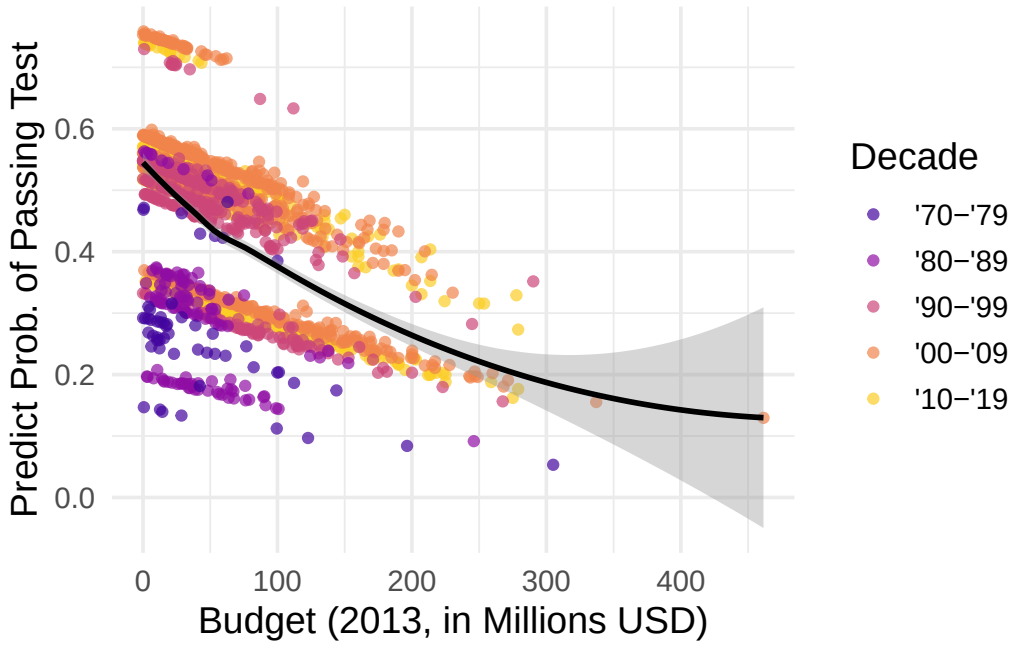
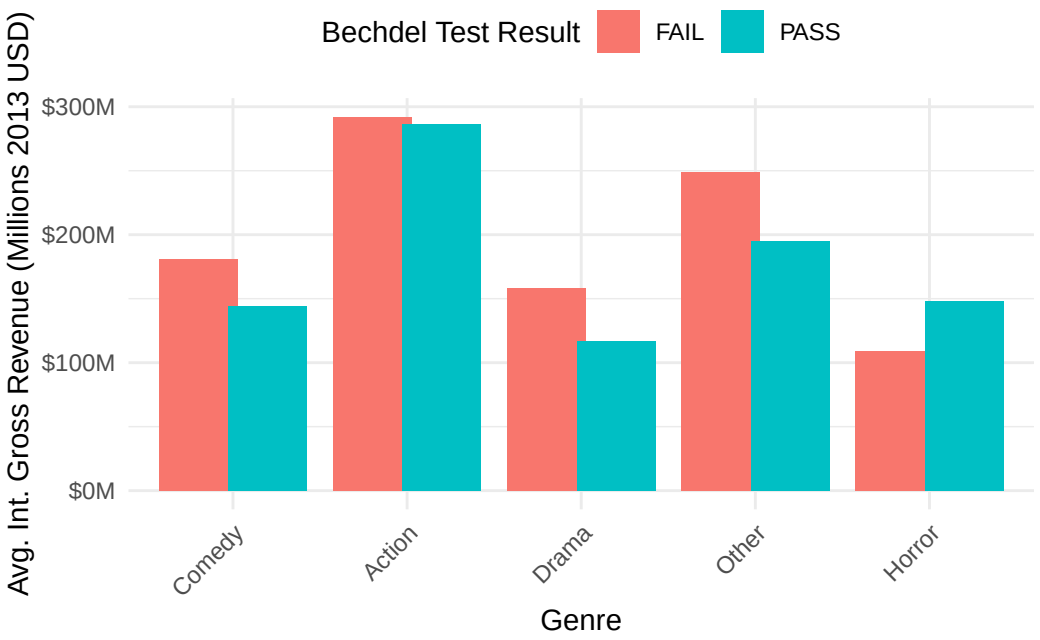


Figure 1 illustrates the predicted probability of a film passing the Bechdel Test as a function of its production budget (adjusted to 2013 USD), stratified by decade. The black trendline indicates an overall negative relationship between budget and the likelihood of passing the Bechdel Test, with probabilities decreasing as budgets increase up to approximately $300 million. This suggests that higher-budget films are less likely to prioritize gender-inclusive narratives. Interestingly, there is a slight upward trend in passing probability for the very highest-budget films, but the confidence interval (shaded region) is wide, indicating high uncertainty. Films from more recent decades (2000s and 2010s, shown in orange and yellow) generally have higher baseline probabilities of passing the test compared to earlier decades (e.g., 1970s and 1980s, shown in purple and blue), reflecting temporal improvements in gender representation. This temporal trend underscores the evolving priorities of the film industry toward inclusivity, though budget remains a constraining factor.

4.3 Research Question 2: Bechdel Test, Genre, and the Worldwide Box Office

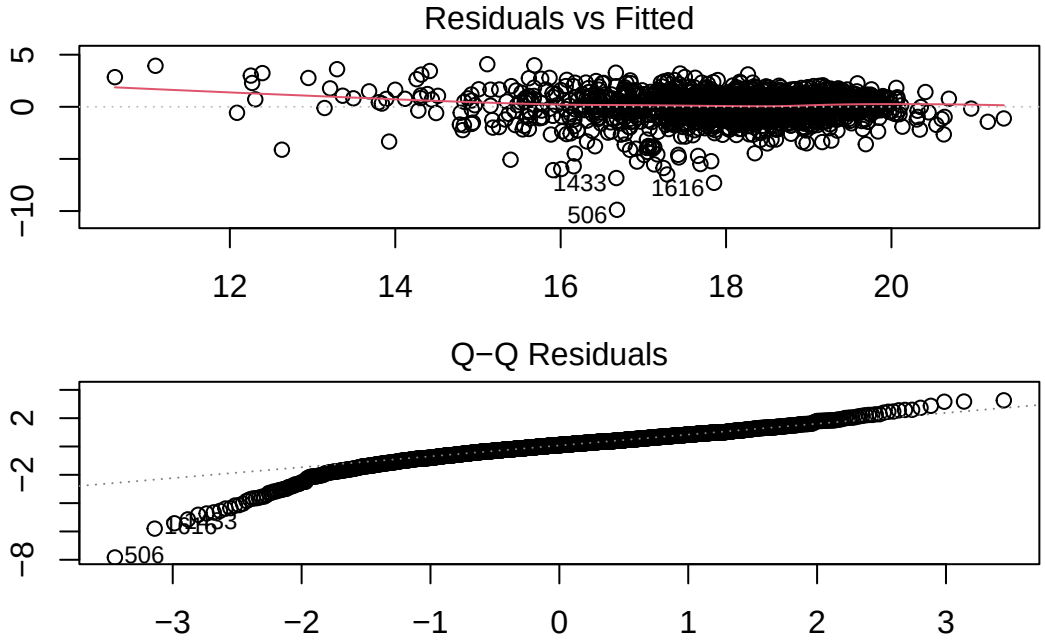
To investigate the factors influencing worldwide gross revenue, a linear regression model was fitted with Bechdel Test outcome, production budget, genre, and decade of release as predictors, variables which were selecting a priori. Our exploratory analysis revealed the need for exploration of an interaction term. Figure 2 (below) suggested a potential interaction between genre and Bechdel Test outcome.

Figure 2: Impact of Bechdel Test Result on Revenue by Genre



From this initial exploration, we thought it most appropriate to explore the interaction between Bechdel test outcome and genre, as test outcome appeared to vary across genres. We added this interaction term to our linear model, but found that none of the interaction variables were statistically significant, meaning the interaction terms did not show meaningful contributions in explaining variation in worldwide revenue. Additionally, a nested F-test was conducted between models with and without the interaction term revealed an F-statistic that was not statistically significant (p=0.86640), meaning that including the interaction term in our model did not significantly contribute to explaining variation in our response variable. Consequently, the interaction term was excluded from the final model, and the remainder of the analysis focused on the simplified model without interactions.

Now that it was determined an interaction term would not be included, we moved forward with the variables of Bechdel test outcome, genre, budget and decade code in our model. This linear model was created, but creation of diagnostic plots revealed a fanning pattern in the residuals vs. fitted plots as well as significant deviation from teh theoretical quantile line in the QQ-plot, indicating the presence of heteroscedasticity and non-normal distribution of the residuals respectively. In order to combat this, we performed a logarithmic transformation on the numerical variables in our model, worldwide revenue and budget. This transformation dramatically improved the previous issues seen in the diagnostic plots. As seen in figure LKD-WJHLAKJNDJKLFN below the Residuals vs. Fitted plot reveals a reasonable scatter around zero, though slight heteroscedasticity at higher fitted values indicates possible variability in error variance. The Q-Q plot demonstrates approximate normality of residuals, with minor deviations in the tails suggesting some outliers or non-normal behavior. These diagnostics indicate the model is well-specified, though further refinement may improve robustness.



Our final linear model yielded the following results, shown in figure KJAHJDBAKJKJ below

Table 4: Linear Regression Model Summary: All Coefficients

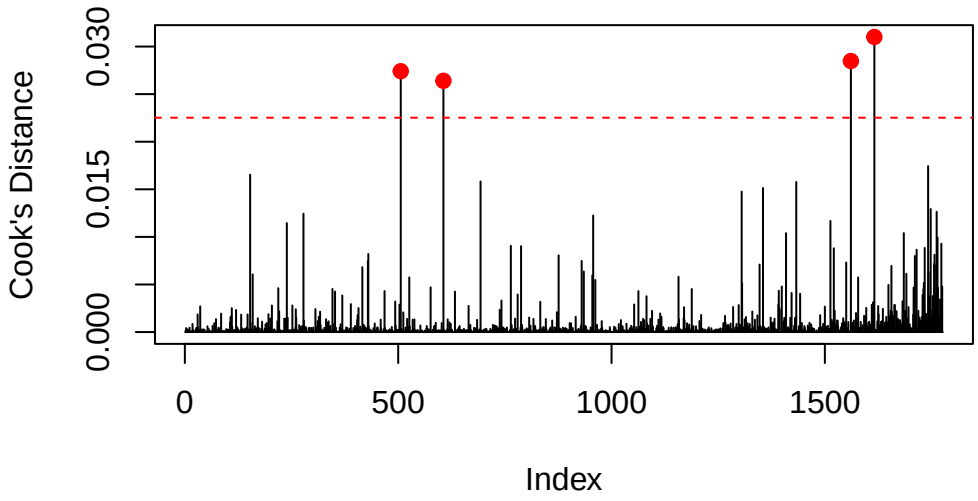
Variable	Estimate	Std_Error	t_value	CI	p_value
(Intercept)	3.999	0.4308	9.281	[3.154, 4.844]	<0.001
binaryPASS	-0.05637	0.06255	-0.901	[-0.1791, 0.06632]	0.368
genre Action	-0.1332	0.08591	-1.551	[-0.3017, 0.03527]	0.121
genre Drama	-0.1858	0.08317	-2.234	[-0.349, -0.02272]	0.026
genre Other	-0.1486	0.09696	-1.532	[-0.3387, 0.04159]	0.126
genre Horror	0.6725	0.1408	4.776	[0.3964, 0.9487]	<0.001
log budget 2013	0.8954	0.02343	38.214	[0.8495, 0.9414]	<0.001
decade code 1980s	-0.8504	0.2087	-4.075	[-1.26, -0.4411]	<0.001
decade code 1990s	-1.147	0.1883	-6.089	[-1.516, -0.7773]	<0.001
decade code 2000s	-1.318	0.1805	-7.304	[-1.673, -0.9644]	<0.001
decade code 2010s	-1.234	0.1854	-6.657	[-1.597, -0.8704]	<0.001

INTERPRET RESULTS

Figure 3: Log Worldwide Gross by Genre and Passing Test

The regression analysis reveals that the log-transformed production budget was the strongest predictor of worldwide revenue ($\hat{\beta} = 0.89544, p < 0.001$, indicating a nearly proportional

relationship between budget and revenue. Among genres, horror films earned significantly higher revenues compared to the baseline genre ($\hat{\beta} = 0.67254, p < 0.001$, while dramas earned less on average ($\hat{\beta} = 0.18584, p = 0.0256$. Action films and other genres did not exhibit statistically significant differences. Temporal trends were prominent, with films from the 1980s ($\hat{\beta} = 0.85043, p < 0.001$). Although Bechdel Test compliance was not statistically significant ($p = 0.3677$), its inclusion underscores the need to consider a range of factors in revenue modeling. Table 4 provides a detailed summary of the model results.



Cook's Distance was evaluated to detect any extreme values in the data. As a result, a 4 points were identified as having a significant impact on the model's decisions. After refitting the model without these influential points, an improvement of approximately 1% was observed in the adjusted R-squared metric. Consequently, it was concluded that removing these observations from the dataset was not appropriate. Additionally, multicollinearity among the variables was assessed, and the Variance Inflation Factor (VIF) scores for all variables were found to be very close to 1, indicating the absence of multicollinearity issues in the final model.

5. Conclusion

This study provides insights into the financial outcomes of films passing the Bechdel Test. Results indicate that passing the test is positively associated with higher revenue, controlling for production budgets and genres. Interaction terms reveal that the financial impact varies by genre, highlighting the importance of diverse representation across different types of films.

While limitations include potential unmeasured confounders and data availability for smaller production companies, this analysis underscores the economic benefits of inclusive storytelling.

Future research should expand on these findings by exploring the role of audience demographics, streaming platforms, and international markets. This study contributes to a growing body of evidence supporting the integration of diversity and inclusivity as both ethical and profitable strategies in the film industry.

6. References

1. Bechdel, A. (1985). Dykes to Watch Out For. *First published as a comic strip in The Essential Dykes to Watch Out For.*
2. Motion Picture Association. (2023). *THEME Report: A Comprehensive Analysis of the Global Film Industry.*
3. Smith, S. L., Choueiti, M., & Pieper, K. (2022). *Inequality in 1,300 Popular Films: Examining Gender, Race, & Ethnicity.* USC Annenberg Inclusion Initiative.
4. Lauzen, M. M. (2021). *The Celluloid Ceiling: Behind-the-Scenes Employment of Women on the Top 250 Films of 2020.* Center for the Study of Women in Television & Film, San Diego State University.
5. FiveThirtyEight. (n.d.). Bechdel Data. Retrieved from https://github.com/fivethirtyeight/data/tree/mas
6. Silver, N., & Arthur, R. (2014). The Dollars and Cents Case Against Hollywood's Exclusion of Women. FiveThirtyEight.