Bechdel Test Final Report

Samantha Chai, Samantha Candelo-Ortegon, Chi Le, Grace Lessig

1. Introduction

1.1. Background

The Bechdel test, also known as the Bechdel-Wallace test, is one of the better-known “tests” to gauge a film’s female representation. Originally invented in 1985 by Allison Bechdel as a joke in her comic strip “Dykes to Watch Out For” and inspired by the writings of Virgina Woolf and Liz Wallace, the test asks the viewer three simple questions. First, whether there are two named women, second, whether these women are having an on-screen discussion, and third, whether that conversation is about something other than a man. By “passing” this test, a film is thought to be “more inclusive” than those that fail it. Since its inception in 1985, it has become famous not just for the test itself but for the cultural phenomenon behind it. People might not know where it originated or even that it started as a joke, but they know of it. Is the test really that accurate, however? Are these three “requirements” enough to test the entire scope of the film's female representation?

1.2. Motivation

The motivation behind our research stems from both quantitative and qualitative observations around the test. If the test is a seamless and perfect detector of female representation, what can we learn about the films that pass it? We want to know what aspects dictate whether a film will pass the test and what percentage of films have those features. When it comes to those features, which are the most important? While there is cultural discussion about the importance of having a female director or crew, does the math support this? Are there actual shifts in passing the test when implementing these quotas for females in the crew and cast? From a qualitative standpoint,

we are interested in what this data means in a contextual sense. Is this test an accurate representation, or are there other tests or measures of a film that do a better job of judging the female representation?

2. Literature Review

The intersectionality of the test offers some incredibly interesting and useful information to note. Much research has been done, in a similar fashion to ours, trying to find a relationship between passing the test and other quantitative factors. There has been research that shows the films that pass the test achieve higher box office earnings[1] while other research has shown to show the opposite, and that box office earnings will decrease[2]. This clear disagreement with the facts leads many to question whether the test should continue to be used. It was a tool used scientifically, yet not made by a scientist for that purpose. A film could still pass the test even with incredibly harsh gender stereotypes (a mother and daughter discussing the importance of cleaning over schoolwork could pass the test despite the conversation being about reinforcing negative gender stereotypes)[3]. The Bechdel test ignores the racial and cultural background of the "named female character," leading to many "subtests" that account for these other factors. Some examples of these alternative tests include the Duvernay test and the Waithe test, which were made to test the representation of African-American women and film, and the Landau test, which tests not what a

[1] Johann Valentowitsch, "Does Female Screen Presence Pay off at the Box Office? An Exploratory Analysis with Special Emphasis on the Bechdel Test," *Journal of International Business and Economics* 22, no. 1 (March 1, 2022): 40–69, https://doi.org/10.18374/jibe-22-1.4.
[2] Andrew M. Lindner, Melissa Lindquist, and Julie Arnold, "Million Dollar Maybe? The Effect of Female Presence in Movies on Box Office Returns," *Sociological Inquiry* 85, no. 3 (March 14, 2015): 407–28, https://doi.org/10.1111/soin.12081.
[3] Johann Valentowitsch, "Does Female Screen Presence Pay off at the Box Office? An Exploratory Analysis with Special Emphasis on the Bechdel Test," *Journal of International Business and Economics* 22, no. 1 (March 1, 2022): 40–69, https://doi.org/10.18374/jibe-22-1.4.

film does have but what it doesn't have[4]. These tests would find a similar fate to the Bechdel test in failing to account for all aspects of intersectionality. The Duvernay test, for example, focuses on African American women but ignores the presence of Latina or Asian women, while the Villalobos test accounts for Latina women but ignores the presence of African-American women. Nearly 60% of films fail the Bechdel test, despite it being so open and broad, so what would the failure rate look like for these other tests? What then is the right test?[5]

3. Dataset Construction and Preparation

We drew from two separate data sources to construct our primary dataset. The first was a Kaggle "9000+ Movies: IMDb and Bechdel" dataset (n = 9478), which provided a wide-ranging collection of films alongside their Bechdel Test scores and other film feature information (release year, IMDb average rating, runtime, and genres). This dataset was our main source for capturing the gender representation variable and general film classification data. The second key source was the FiveThirtyEight GitHub "Bechdel Test" dataset (n =1795), which offered a list of films with additional detailed production attributes (budget, domestic gross revenue, and international gross revenue). Though smaller, this dataset was essential in its inclusion of financial information absent in the Kaggle dataset, offering another layer of analysis and feature variety to look at.

---

[4] Terri Waters, "7 Tests (That Aren't the Bechdel Test) That Measure Movies for Gender Equality and Representation," The Unedit, January 23, 2019, https://www.the-unedit.com/posts/2018/8/20/7-tests-that-arent-the-bechdel-test-that-measure-movies-for-gender-equality-and-representation.

[5] Johann Valentowitsch, "Does Female Screen Presence Pay off at the Box Office? An Exploratory Analysis with Special Emphasis on the Bechdel Test," *Journal of International Business and Economics* 22, no. 1 (March 1, 2022): 40–69, https://doi.org/10.18374/jibe-22-1.4.

Our target variable was the Bechdel Test rating, a score ranging from 0 to 3 that indicates the level of female representation in a film. A score of 3 means the film fully passes the test: it includes at least two named women who talk to each other about something other than a man. This rating served as the classification label for our predictive modeling task. We were able to indicate a clear pass/fail metric by transforming the Bechdel Test rating into a binary indicator: a value of 1 if a movie received a Bechdel Test score of 3 (i.e., it fully passed the test), and 0 otherwise. This allowed us to shift from a 4-point ordinal scale to a clear classification task for prediction.

4. Methodology

4.1. Data Preparation

To prepare the data for analysis and modeling, we cleaned and merged the two datasets by using the movie title as our merge key. This is essential in resolving inconsistencies in formatting, such as punctuation differences, alternate title spellings, and capitalization, as well as handling missing values across both sources.

Our first step was to select a set of relevant features: imdb_average_rating, bechdel_rating, budget, domestic_gross, international_gross, duration, genre, and release_year. We selected these features as they were held in common amongst the datasets (something essential we had to account for since we were combining them into one) and included a range of both categorical and quantitative data.

Secondly, we engineered new features, total_gross (domestic_gross + international_gross), profit (total_gross - budget), release decade (release year floor divided by 10 then multiplied by 10 again to round to the nearest tenth), period (a binary variable categorizing films as either "pre-Bechdel" (released before 1985) or "post-Bechdel" (released from 1985 onwards)), and bechdel_pass (1 if a film scored 1 or higher on the Bechdel Test, 0 otherwise). The new features, release decade, period, and bechdel_pass, were primarily used to group the data together to easily identify relationships with a movie's pass result and release time. Other new features, such as profit and total_gross, were only used for our training models as an accurate measure for financial success. Engineering these features was essential, as it allows us to consider more complex relationships between a movie's past result and other significant attributes.

Thirdly, we dropped rows with missing values in selected predictors to ensure model training quality. The datasets were from openly sourced data sites, meaning that not every row was filled in.

Lastly, to prepare for the hot encoding of the movie "genre" category, the genres with sample sizes greater than 50 were kept (such as Action, Comedy and Drama), while the movie genres that had low counts (such as Fantasy, Animation and Mystery) were grouped into the new genre category "Other." This was done because of their small sample sizes.

4.2. Feature Selection

We chose five features to focus on as potential predictors for passing the Bechdel test. These features were the film's duration, release year, budget, IMDb rating, and profit. Categorical features like genre were initially explored but not encoded for the first round of model training. We wanted to focus on these five features as they were unbiased and public. Features like rating or genre are subjective to the movie-goers, and their opinions rely not just on the biases of their opinions, but also on the biases of their decade. The ways a movie is reviewed and the genre it is put into are not a clear factual statement, but more of a consensus opinion. In terms of feature selection, we wanted to look for more factual numerical pieces of data. The duration, release year, and budget are all decided upon by teams of people over months of work and discussion. These levels of discussions a movie production must go through strip it of the effect of a singular person's bias, making it more trustworthy for testing.

4.3. Data Visualization

To better understand how movie characteristics relate to Bechdel Test outcomes, we implemented three major visualization strategies: trend analysis using categorical comparison using bar charts, line plots, and structural analysis using network graphs. These approaches allowed us to examine both temporal trends and relational patterns within the dataset.

4.3.1. Bar Chart Visualization of Movie Features by Bechdel Pass status

To investigate how movie characteristics differ based on Bechdel Test outcomes, we created bar charts that compare the average values of several features across two groups: movies that pass

the Bechdel Test and those that do not. Rather than using the full 0–3 rating scale, we simplified the variable into a binary Bechdel_pass column, where a value of 1 indicates a movie scored 3 and thus passed the test, and 0 otherwise (because scores below 3 indicate failing the test).

4.3.2 Line Plot Visualization of Movie Attributes by Decade and Bechdel Rating

To explore trends in movie characteristics and how they relate to Bechdel Test performance, we grouped our dataset by release decade and Bechdel Test rating. We computed the average of key numeric attributes within each (decade, bechdel_rating) group, allowing us to examine how factors such as budget, IMDb average rating, and domestic gross revenue varied over time and across levels of gender representation.

Using the grouped data, we created a series of line plots with seaborn to visualize these trends. Each line represents a specific Bechdel rating (0–3), and the x-axis denotes the decade of release. Movies that scored a 3 fully passed the Bechdel Test, while lower scores indicate decreasing levels of female representation. The y-axis varied by attribute, capturing the average production budget, IMDb rating, or domestic gross for each category over time.

This visualization approach provided a comparative view of industry patterns and revealed how movies that pass or fail the Bechdel Test differ in terms of production investment, critical reception, and financial performance across decades. It also allowed us to assess whether progress in gender representation correlates with broader trends in the film industry.

4.3.3 Complex Networks

To explore the relationships between a movie's characteristics and its Bechdel Test outcome, we constructed network graphs using NetworkX, treating each movie as a node and drawing edges between movies that share a common attribute. This structure allowed us to visually examine patterns in gender representation across different groupings.

These graph-based visualizations and metrics offered an intuitive way to explore the influence of genre and era on gender representation, revealing not just individual outliers but also broader structural patterns in the dataset.

4.4. Modeling Approach

We trained four types of classification models throughout our analysis: a decision tree classifier, a random forest classifier, a logistic regression, and K-nearest neighbors. Each model was trained on an 80/20 train-test split. Scaling (StandardScaler) was applied for Logistic Regression and KNN models. Firstly, four models were trained using the combined movies dataset, and then tested on a filtered dataset (only including movies released in the years 1970-2010) to examine model behavior when limiting outliers. Once the movie-genre categorical variable onto the combined dataset through hot encoding, the four models were retrained, and then again, but with only the genres as predictors. This was repeated, but using the original IMDb dataset to investigate the results if the dataset size was larger.

Model performance was evaluated using the following metrics: accuracy (percentage of correct predictions), precision/recall / F1-score (to understand class-specific performance), and confusion matrices (to visualize prediction errors between passing and failing films).

5. Results & Analysis

5.1. Data Visualization

5.1.1 Bar Chart Visualization of Movie Features by Bechdel Pass status

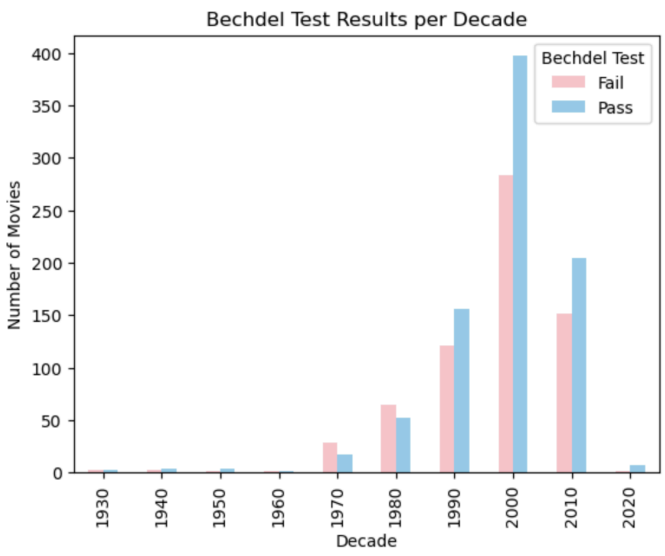

Figure 1.1 Bar chart displaying the proportion of movies that failed and passed the Bechdel test by decade (1930-2020) and the number of movies per decade. Data source: combined_movies_df (n = 1500)

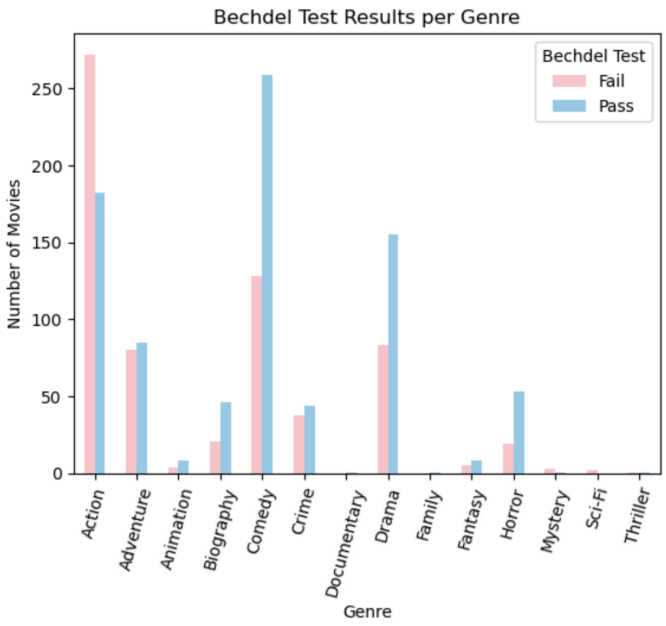

Figure 1.2 Bar chart displaying the proportion of movies that failed and passed the Bechdel test, by genre and number of movies per genre. Data source: combined_movies_df (n=1500)

The exploratory analysis displays that the dataset is imbalanced in terms of movie release date; movies released before 1980 and after 2010 have low representation, while movies released in 2000 are the most represented (Figure 1.1). Notably, in the decades 1970 and 1980, there is a higher proportion of movies failing the Bechdel test than passing (Figure 1.1). However, after 1980 (and it is important to note that the Bechdel test was created in 1985), the proportion of movies passing is greater, with it being the greatest in the decade 2000 (Figure 1.1). The dataset is also imbalanced in terms of genre representation; a wide majority of movies represented are of the action or comedy genre (Figure 1.2). Most action movies in this sample have failed the test, while most of the other genres ' movies have passed (Figure 1.2). However, it is notable that the proportion of failing to pass is almost equal for genres such as adventure or crime.

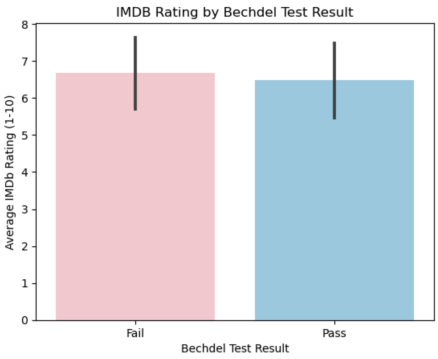

Figure 1.3 Bar chart displaying the proportion of movies failing or passing by their average IMDB rating (1-10). Data source: combined_movies_df (n = 1500)

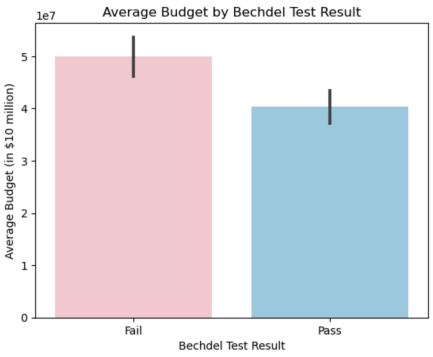

Figure 1.4 Bar chart displaying the proportion of movies failing or passing by their average budget (in $10 million). Data source: combined_movies_df (n = 1500)

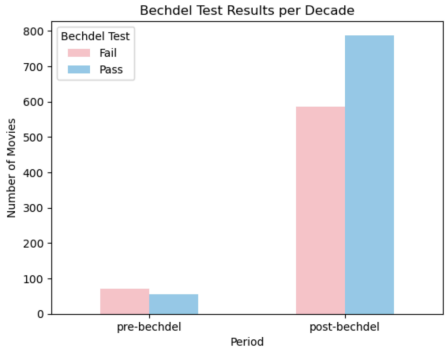

Figure 1.5 Bar chart displaying the proportion of movies released before and after the creation of the Bechdel test, by the movie count represented in the dataset. Data source: combined_movies_df (n = 1500)

There seems to be little difference in the average IMDB rating for movies from the dataset that have failed and passed the Bechdel test (Figure 1.3). However, there is a notable difference in terms of budget; the average budget for movies that have failed is greater than for movies that

have passed (Figure 1.4). There is an overrepresentation of movies that were released post-Bechdel (after 1985) in the dataset (Figure 1.5). Before the test's creation, the proportion of movies that failed was slightly greater than those that passed; while post-Bechdel, the proportion of movies passing is much greater (Figure 1.5).

5.1.2 Line Plot Visualization of Movie Attributes by Decade and Bechdel Rating

5.1.2.1 With full dataset

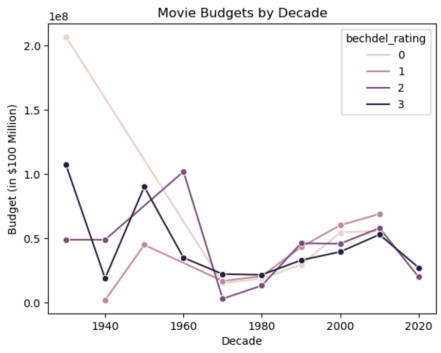

Figure 2.1 Line plot of average movie budget (in $100 million) by release decade and Bechdel rating (0-3). Data source: combined_movies_df (n = 1500)

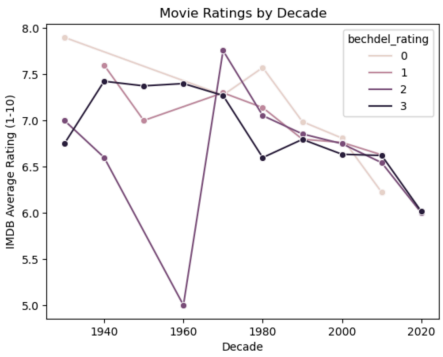

Figure 2.2 Line plot of average IMBd movie rating (1-10) by release decade and Bechdel rating (0-3). Data source: combined_movies_df (n = 1500)

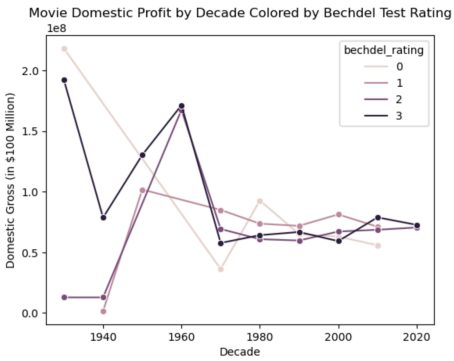

Figure 2.3 Line plot of movie average domestic gross (in $100 million) by release decade and Bechdel rating (0-3). Data source: combined_movies_df (n = 1500)

To analyze how movie characteristics vary by gender representation over time, we created line plots showing average budget, IMDb rating, and domestic gross profit grouped by release decade and Bechdel Test rating. These plots allow us to visualize how films that fully pass the Bechdel Test (score of 3) compare to those that do not (scores 0–2) across different periods.

Across all three plots, data from decades before the 1970s shows noticeable volatility, particularly in the 1930s–1950s. This instability is likely due to a limited number of movies with complete data from those early decades, making the averages more sensitive to outliers. As such, trends from the 1970s onward provide a more reliable basis for interpretation.

In terms of movie budgets, films that fully pass the Bechdel Test (score 3) generally trend similarly to other rating categories in later decades, with some indication that they began receiving comparable financial investment by the 1990s and 2000s. Earlier decades show no consistent pattern, again likely due to sparse data.

When looking at IMDb average ratings, films with higher Bechdel scores (particularly 2s and 3s) tend to maintain more stable and slightly higher ratings from the 1970s through the 2000s. This suggests that gender-inclusive films were critically well-received, contradicting any assumptions that such films perform poorly with audiences or critics (Figure 2.1).

For domestic gross revenue, the lines converge more tightly in recent decades, but earlier patterns are less conclusive due to noise in the data. Films that pass the Bechdel Test do not appear to perform significantly worse or better financially, suggesting that representation does not hinder box office success (Figure 2.2).

Together, these visualizations suggest that while earlier decades suffer from data sparsity, from the 1970s onward, films that meet Bechdel Test criteria show comparable—if not

stronger—performance across budget, ratings, and profit metrics, reinforcing the idea that gender representation and production success can coexist (Figure 2.3).

5.1.2.2 With filtered dataset (1970–2010)

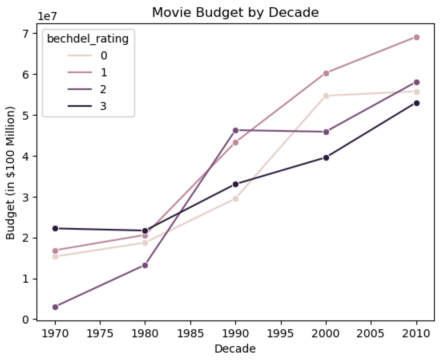

Figure 2.4 Line plot of average movie budget (in $100 million) by release decade and Bechdel rating (0-3). Data source: filtered_df (n=1500)

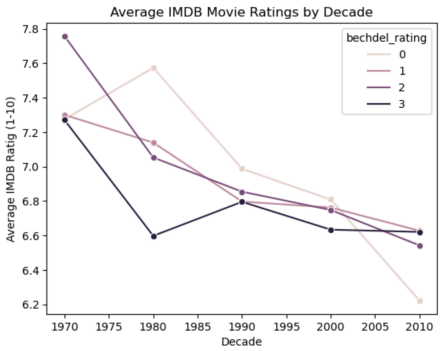

Figure 2.5 Line plot of average IMBd movie rating (1-10) by release decade and Bechdel rating (0-3). Data source: filtered_df (n=1500)

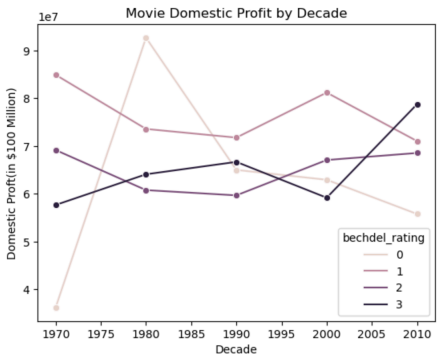

Figure 2.6 Line plot of movie average domestic gross (in $100 million) by release decade and Bechdel rating (0-3). Data source: filtered_df (n=1500)

After the outliers have been filtered out, the data appears less volatile across the four Bechdel ratings, and trends are more easily able to be assessed.

In terms of movie budgets, there is a steady increase in budget over time. Up until 1980, movies that passed the Bechdel test had the largest average budget, while movies with a Bechdel score of 2 (almost passing) trailed behind. Then from 1980-1990, Bechdel score 2 leads with the biggest budget, while score 0 (failing) trails. Post 1990, movies with scores of 1 (barely failing) have the highest average budget, and movies with scores of 3 (passing) have the lowest average budget. This is an interesting disparity that warrants further investigation (Figure 2.4).

In terms of IMDb ratings, there is a notable dip in ratings for movies passing the test (as well as those almost passing) from 1970-1980, while there is a rise in ratings for movies with the lowest Bechdel score over the same period. As time progresses, ratings for movies with each score all steadily decline to about the same average rating, except for movies with the lowest Bechdel score, which do much worse (Figure 2.5).

In terms of domestic profit, movies with the lowest Bechdel scores profited the most, while those that passed or nearly passed profited the least from 1970-1980. From 1980-1990, profits steadily increased for movies passing, and faced a deep decline for movies failing. Post 1990, there are a couple of changing profit trends for movies of all scores, with movies of score 1 (barely failing) leading most years, until in the very recent decade we see movies passing leading, and movies with the lowest score trailing (Figure 2.6).

5.1.3 Complex Networks

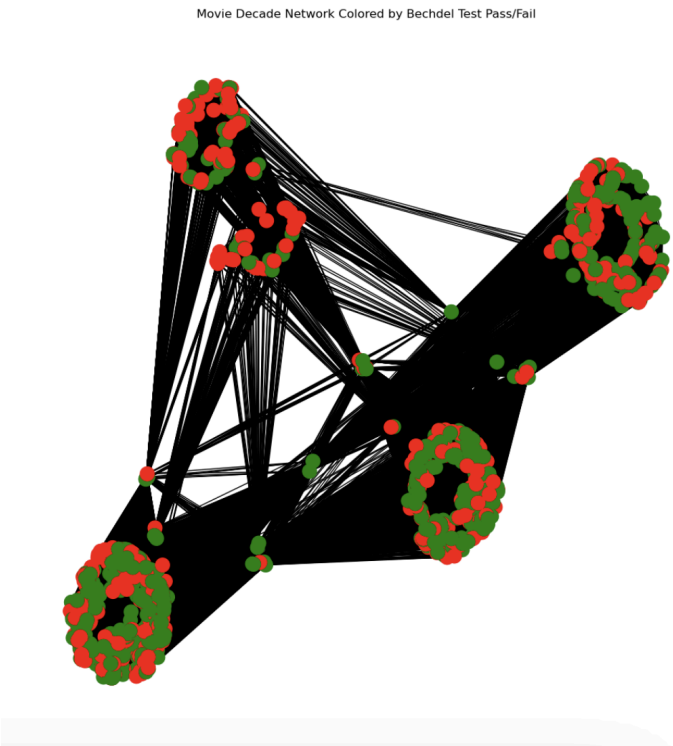

Figure 3.1

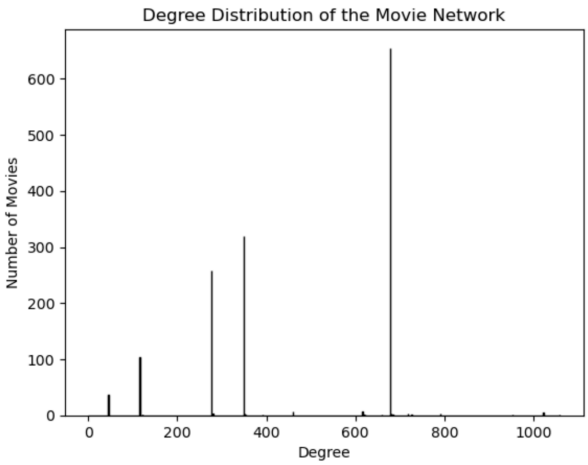

Figure 3.2

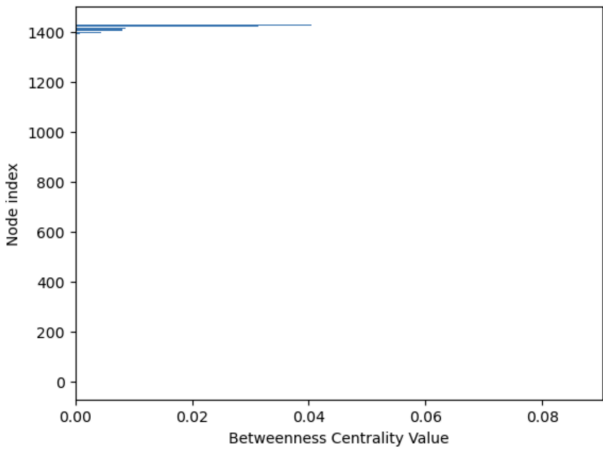

Figure 3.3

For the network, we investigated how shared release decade relate to Bechdel Test results. We grouped the movies by their assigned decade and added edges between every unique pair of

movies within each group. Movies were grouped by decade, and edges were added between all movies released in the same decade.

For the graph, nodes were colored according to Bechdel Test pass/fail status—green for movies that fully pass (score of 3) and red for those that do not (scores 0–2). This visual distinction helped identify clusters or communities where Bechdel-passing movies were more or less prevalent.

Additionally, we calculated two key network metrics: degree and betweenness centrality. Degree measures how many connections each movie has—i.e., how many other movies it shares a decade with. We identified the movie with the highest degree and visualized the overall degree distribution to examine connectivity patterns. We also computed betweenness centrality, which captures how often a movie lies on the shortest paths between others, indicating its potential role as a bridge within the network.

5.2. Model Training

5.2.1 Numerical Categories

We trained and evaluated four classification models (Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN)) to predict whether a movie would pass the Bechdel Test, using numeric features including duration, budget, profit, and release year. The goal was to assess how well these production-related attributes alone could serve as predictors of

gender representation. Their performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices.

Table 1. Results from supervised learning models, where predictors = duration, budget, profit, release_year, and dependent variable = failing (0) / passing (1) the Bechdel test. Data source: combined_movie_df (n=1500) and filtered_df (n=1500)

Model	Dataset	Accuracy	Precision (0 / 1)	Recall (0 / 1)	F1-Score (0 / 1)
Decision Tree	Unfiltered Combined	0.597	0.52 / 0.65	0.50 / 0.67	0.51 / 0.66
	Filtered Combined	0.597	0.49 / 0.64	0.34 / 0.77	0.40 / 0.70
Random Forest	Unfiltered Combined	0.587	0.51 / 0.64	0.48 / 0.67	0.49 / 0.65
	Filtered Combined	0.569	0.46 / 0.65	0.49 / 0.62	0.48 / 0.63
Logistic Regression	Unfiltered Combined	0.613	0.60 / 0.62	0.23 / 0.89	0.33 / 0.73
	Filtered Combined	0.597	0.49 / 0.63	0.28 / 0.81	0.36 / 0.71

KNN (no scaling)	Unfiltered Combined	0.587	0.51 / 0.64	0.47 / 0.67	0.49 / 0.65
	Filtered Combined	0.576	0.47 / 0.66	0.52 / 0.62	0.49 / 0.64
KNN (with scaling)	Unfiltered Combined	0.577	0.50 / 0.62	0.42 / 0.69	0.45 / 0.65
	Filtered Combined	0.522	0.41 / 0.61	0.44 / 0.58	0.42 / 0.59

5.2.1.1 With full dataset

On the full dataset, Logistic Regression achieved the highest accuracy (0.613), with strong performance on class 1 (precision: 0.62, recall: 0.89, F1-score: 0.73), though it struggled to correctly identify class 0 (recall: 0.23). This indicates a significant bias toward predicting class 1. Decision Tree and Random Forest followed closely with accuracies of 0.597 and 0.587, respectively, and showed more balanced performance between the two classes compared to Logistic Regression. The Decision Tree had slightly better recall on class 0 (0.50 vs. Random Forest's 0.48), but the Random Forest had marginally better precision on both classes.

KNN with scaling and KNN without scaling performed similarly (accuracy ~0.577–0.587). Both models favored class 1, with better recall and F1-scores for class 1 than class 0. However, scaling the features slightly improved recall for class 0 (0.47 vs. 0.42) and gave a slightly more balanced

outcome between the classes. Overall, performance across models on the full dataset was modest, with a general trend of better predictions for class 1 than class 0.

5.2.1.2. With filtered dataset (1970–2010)

In the filtered dataset, Decision Tree maintained the same accuracy as it did with the full dataset (0.597), but its class-wise performance shifted: recall for class 1 improved to 0.77, while recall for class 0 dropped to 0.34. This suggests the model became more confident in predicting class 1 at the expense of correctly identifying class 0.

Random Forest and Logistic Regression both saw slight decreases in overall accuracy (to 0.569 and 0.597, respectively). However, both continued to favor class 1, maintaining decent recall and F1-scores for that class while still underperforming on class 0. Logistic Regression again showed particularly strong recall for class 1 (0.81) but low performance for class 0 (recall: 0.28), reinforcing its tendency to predict the majority class.

KNN with scaling performed slightly worse after filtering (accuracy dropped from 0.587 to 0.576), though class-wise recall became more balanced (0.52 for class 0 and 0.62 for class 1), suggesting some improvement in identifying class 0. KNN without scaling had the lowest performance of all models on the filtered data (accuracy: 0.522), with both precision and recall lower than their full dataset counterparts, highlighting the detrimental effect of not scaling the features in distance-based models.

a) Comparison between the full dataset and the filtered dataset:

Overall, filtering the dataset to include only movies from 1970 to 2010 did not lead to a significant improvement in accuracy for any of the models. Most models experienced either a plateau or a slight drop in performance.

However, the effect of filtering varied depending on the model. For Decision Tree and Logistic Regression, filtering shifted model behavior to more confidently predict class 1, improving recall for that class but worsening performance on class 0. This indicates that the filtered data might have strengthened patterns associated with class 1 but made class 0 harder to distinguish.

KNN (with scaling) showed one of the more balanced results post-filtering, suggesting that filtering could slightly improve the model's ability to handle both classes when scaling is applied. In contrast, KNN without scaling showed worse results after filtering, further emphasizing the importance of preprocessing for distance-based algorithms.

In conclusion, while filtering helped certain class-level metrics, particularly for class 1, it did not consistently improve model performance.

5.2.2 Categorical Categories

A few models were retrained, accounting for the “genre” variable. These models now contain five predictors: genre (hot encoded into 8 categories), duration, budget, profit, and release year.

Table 2. Results from supervised learning models retrained to include movie genre, where predictors = genre, duration, budget, profit, release_year, and dependent variable = failing (0) / passing (1) the Bechdel test. Data source: combined_movied_df (n=1500)

Model	Dataset	Accuracy	Precision (0 / 1)	Recall (0 / 1)	F1-Score (0 / 1)
Decision Tree	Unfiltered Combined	0.633	0.57/0.67	0.49/0.74	0.53/0.70
Random Forest	Unfiltered Combined	0.637	0.57/0.68	0.52/0.72	0.55/0.70
Logistic Regression	Unfiltered Combined	0.633	0.58/0.66	0.48/0.74	0.53/0.70

The three models resulted in very similar accuracy, with Random Forest leading (0.637). As each has an accuracy of ~0.7, we conclude that they are moderate to good predictors, since they can predict a movie passing the Bechdel test ~70% of the time. The Decision Tree performed moderately well on class 1 (precision: 0.67, recall: 0.74, F1-score: 0.70), but not well when identifying class 0 (precision: 0.57, recall: 0.49, F1-score: 0.53). This indicates a significant bias toward predicting class 1. The Random Forest and Logistic Regression models had very similar metrics for identifying class 0 (hovering in the 0.5-0.6 range), also illustrating moderate performance (Table 2).

We also wanted to explore results with genre as the only predictor variable, and with a larger sample size of movies. Thus, we retrained a few of the models using the combined dataset (n=1500), and then on the original "9000+ Movies: IMDB and Bechdel" dataset (n=9478).

Table 3. Results from supervised learning models, where the only predictor is movie genre and the dependent variable = failing (0) / passing (1) the Bechdel test. Data source: combined_movied_df (n=1500) and original_bechdel_df (n=9478)

Model	**Dataset**	**Accuracy**	**Precision (0 / 1)**	**Recall (0 / 1)**	**F1-Score (0 / 1)**
Decision Tree	n/a	n/a	n/a	n/a	n/a
	Original Bechdel	0.609	0.57/0.62	0.34/0.81	0.43/0.70
Random Forest	n/a	n/a	n/a	n/a	n/a
	Original Bechdel	0.609	0.57/0.62	0.34/0.81	0.43/0.70
Logistic Regression	Combined (Unfiltered)	0.633	0.58/0.66	0.44/0.76	0.50/0.71
	Original Bechdel	0.609	0.57/0.62	0.34/0.81	0.43/0.70

KNN (no scaling)	Combined (Unfiltered)	0.603	0.53/0.65	0.48/0.70	0.50/0.67
	Original Bechdel	0.595	0.54/0.62	0.37/0.77	0.44/0.68
KNN (with scaling)	Combined (Unfiltered)	0.533	0.46/0.63	0.63/0.45	0.53/0.53
	Original Bechdel	0.478	0.36/0.54	0.29/0.62	0.33/0.57

The three models retrained with the same combined dataset as all the previous models here are Logistic Regression, and KNN with and without scaling. All three produced similar accuracies (in the 0.5-0.6 range), with Logistic Regression leading (0.633), performing moderately well, and KNN with scaling trailing (0.533), performing moderately. The Logistic regression (precision: 0.66, recall: 0.76, F1-score:0.71) and KNN (without scaling) ((precision: 0.65, recall: 0.70, F1-score:0.67) performed moderately well toward predicting class 1 with metrics in the 0.6-0.7 value range. However, their performance towards predicting class 0 was poorer (LR - precision: 0.58, recall: 0.44, F1-score:0.50) and (KNN - precision: 0.53, recall: 0.48, F1-score:0.50). The KNN with scaling's precision was noticeably greater for predicting class 1 (0.63) than 0 (0.46). However, this switches in terms of recall (class 0: 0.63 and class 1: 0.45). This model's F-1 score (0.53) indicates poor to moderate performance.

All five models were retrained on the larger dataset (n=9478). The Decision Tree (precision: 0.62, recall: 0.81, F1-score:0.70), Random Forest (precision: 0.62, recall: 0.81, F1-score:0.70), Logistic Regression (precision: 0.62, recall: 0.81, F1-score:0.70), and KNN without scaling (precision: 0.62, recall: 0.77, F1-score:0.68) faired very similarly when predicting class 1. As these metrics hover in the 0.6-0.7 range, and recall scores ~0.8, we find that the models are moderately good predictors. However, when predicting class 0, the Decision Tree (precision: 0.57, recall: 0.34, F1-score:0.43), Random Forest (0.57, recall: 0.34, F1-score:0.43), Logistic Regression (0.57, recall: 0.34, F1-score:0.43), and KNN without scaling (precision: 0.54, recall: 0.37, F1-score:0.44) falter. Precision remains poor to moderate (~0.5), but the recall metrics are low (~0.3). As for the KNN with scaling model, it performed moderately for predicting class 1 (precision: 0.54, recall: 0.62, F1-score: 0.57). However, poor performance is seen for predicting class 0 (precision: 0.36, recall: 0.29, F1-score: 0.33).

6. Conclusion and Limitations

In this study, we used four classification models (Decision Trees, Random Forest, Logistic Regression and KNN with/out scaling) to predict whether a movie passes (1) or fails (0) the Bechdel Test based on five predictors: genre (hot encoded into 8 categories), movie duration, budget, profit, and release year. They all yielded accuracies ranging between 0.5-0.6, indicating poor to moderate performance. Thus, we conclude that, according to this analysis, our five predictors were not the best at predicting whether a movie passes or fails the Bechdel test. Through our exploratory analysis, interesting trends surfaced, such as how the average budget for movies that failed the test is much higher than for movies that passed in the most recent decades. These findings warrant further investigation in future studies.

There were many limitations present in this study. In terms of the dataset, movie count by decade and genre differed significantly. This imbalance and skewness in sample sizes have an effect on our figures and the conclusions we have deduced from them. Moreover, in the process of combining both original datasets ("9000+ Movies: IMBd and Bechdel" (n = 9478) and "movies.csv" (n = 1795)) into one, our total sample size of movies had decreased to n =1500. In the presence of both more and a larger variety of movies, our results could have been different. Moreover, there may be inaccuracies present in the data, such as for the reported budget and domestic gross.

For future similar studies, it would be interesting to include more predictors in the models, such as the gender of the director, director "prestige" (measured through award count), character dialogue, and more. These variables could not be included in our analysis due to the data being unavailable or difficult to incorporate (e.g., creating a "gender" column and filling it in manually). It would also be recommended to use other classification models that are better suited for dealing with imbalanced data and a binary dependent variable.

Ultimately, the motivations of this study were to determine what variables could predict whether a movie passes or fails the Bechdel Test, and to continue the conversation of the representation of women in recent media. What is important to note is that the Bechdel Test is not the end-all be-all metric for gender equity in media; there are other ways to assess this, as well as other tools out there that would be interesting to further look into. Our work shows this instability of the

Bechdel test, as none of our results provided strong or concrete results to warrant a direction correlation between passing the Bechdel test and any of its production features.

7. Works Cited

Lindner, Andrew M., Melissa Lindquist, and Julie Arnold. "Million Dollar Maybe? The Effect of Female Presence in Movies on Box Office Returns." Sociological Inquiry 85, no. 3 (March 14, 2015): 407–28. https://doi.org/10.1111/soin.12081.

Valentowitsch, Johann. "Does Female Screen Presence Pay off at the Box Office? An Exploratory Analysis with Special Emphasis on the Bechdel Test." Journal of International Business and Economics 22, no. 1 (March 1, 2022): 40–69. https://doi.org/10.18374/jibe-22-1.4.

Waters, Terri. "7 Tests (That Aren't the Bechdel Test) That Measure Movies for Gender Equality and Representation." The Unedit, January 23, 2019. https://www.the-unedit.com/posts/2018/8/20/7-tests-that-arent-the-bechdel-test-that-measure-movies-for-gender-equality-and-representation.

Datasets:

Movies Dataset (movies.csv)

9000+ Movies : IMDb and Bechdel (Bechdel_IMDB_Merge0524.csv)