The Bechdel Test & Film Metrics

Audrey Yang (awy2006), Ting Zhou (tz2089)

December 16, 2023

Abstract

The Bechdel test is a popular metric by which to analyze the extent of female representation in film media. In this project, we quantified and attempted to predict the relationship between females' active presence in a movie and the movies' features, including the genders of writers & directors, as well as the commercial performance metrics and viewer perceptions of the film. We aimed to do this by joining a dataset containing films' Bechdel scores with another dataset of the additional data on these films, then analyzing the relationships of these variables through categorical and machine learning analysis. By examining the correlation between a film's Bechdel performance and its writer/director gender, we found films directed or written by women have significantly higher Bechdel scores than those by men. Moreover, we compared movies' Bechdel distribution across genres, finding Short, Western, Documentary, and War films to have poorer performance. Next, we implemented machine learning algorithms to examine the predictive ability of the films' other information on their Bechdel scores. Overall, we found that the most prominent numerical variables in the IMDb dataset are not ideal predictors of the Bechdel score, and we offer potential explanations for why this may be.

1 Introduction

Alison Bechdel introduced the Bechdel test in her 1985 comic strip *Dykes to Watch Out For* featured in a feminist newspaper. The test consists of three criteria:

1. Two female characters with names are present in the media
2. The two female characters speak to one another
3. The conversation between the two female characters is about a subject that is not a man.

Although Bechdel's thought experiment was very simple, it turns out to fail on many world-famous films whose release dates range from several decades ago to today, including *Forrest Gump*, *The Hobbit: An Unexpected Journey*, *Ant-Man*, etc. This illustrates its continued usefulness, both in analyzing representation of women in film in the past as well as conducting research on how to potentially improve female representation in the future.

We are interested in the Bechdel test because we care about accurate, relevant, and meaningful representation of women in film. As discussions about the role of women in media rage on in the present day, we hope that our project and analysis will contribute to the growing understanding of the part that feminism can play in continued development of the movie industry.

We now discuss some background on both the legitimacy and limitations of the Bechdel test. First: the test, though not being a standardized metric, is positively correlated with the centrality of female characters in films, lending legitimacy to its use. Many past researchers have been automating Bechdel tests through social network analysis. Specifically, Apoorv Agarwal from Columbia University conducted an in-depth analysis on films' scripts and their Bechdel performance, discovering that women were indeed portrayed as less-central or less-important characters in films in movies that failed the test (Agarwal et al., 2015). Thus, the Bechdel test appears to be a reasonable baseline metric to evaluate female representation in media.

However, we also note that the test does have limitations. For example, the score itself can be ambiguous for certain films: viewers may argue and different raters can disagree on whether certain elements of the film pass the criteria of the test. Second, it was worth noting that not all movies that passed the test were considered unbiased against women by filmgoers. For instance, the movie *Spider-Man: No Way Home* passed the test but drew criticism from many viewers for featuring only female characters whose conversations predominantly revolved around the male protagonist, Peter. Additionally, some believe that the movie is guilty of "fridging," where underdeveloped female characters exist primarily to be harmed or killed, and thus whose sole purpose is driving the male protagonist's development.

Using this knowledge, we intend to explore how movies' production and popularity are related to their Bechdel performance.

We describe the workflow of this project:

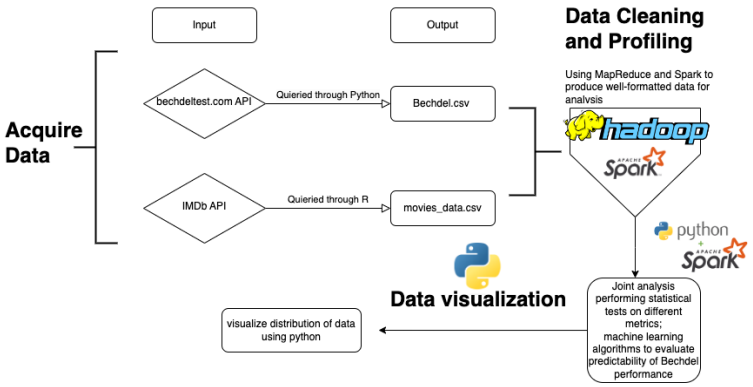

Figure 1: Work Flow Design Diagram

We first ingest the two datasets we will use in this project, one on Bechdel score ratings and one on IMDb film data (such as box office performance, the average viewer rating on IMDb, etc.).

Next, once the datasets are acquired, we go through a stage of data cleaning and data profiling, the former to ensure the data is prepared for joint analysis, and the latter to get a overall understanding of the datasets. We used both Hadoop MapReduce and Spark Scala for these processes.

After the datasets are adequately processed, we join them, and we conduct a two-part joint analysis. First, we analyze the relationships between the Bechdel scores and its categorical data for each film, such as the gender of the writers and directors as well as the genre of the movie. This was done in PySpark. We then use two machine learning algorithms to determine how well the numerical IMDb data can predict the Bechdel scores of the films. This section was also done in PySpark, using MLlib.

Data visualizations were used throughout the process to gain a preliminary understanding of the data and inter-column relationships, as well as to illustrate our conclusions.

2 Motivation

The importance of a sophisticated, independent female character in creative productions was described as early as in Virginia Woolf's essay 'A Room of One's Own.' Around the same time, film had emerged as an experimental medium, revolutionizing people's way of seeing and accessing contemporary culture. Considering how vibrant yet still male-dominant the film industry was, we decided it was worth exploring insights a Woolf-inspired test brought regarding popular films. This topic is relevant to everyone who cares about female representation in film, and indeed anyone who is concerned with women's place in society, as the film industry is powerful in shaping our perception of many ideas, including the role of women in the modern day.

While trying to research on Bechdel test's current relevance, Ting discovered bechdeltest.com, which has recorded around 10,000 movies whose release dates range from the 1870s to the 2020s. We then decided that we wanted to analyze these films through the lens of their performance metrics, such as box office data and average movie rating, and Audrey found an R library omdbapi which allowed us to query through the IMDb Open Movie Database to extract additional details for each film.

The majority of the records were submitted after the 2010s, indicating a heightened contemporary focus on gender portrayal in cinema. Therefore, we hoped that our attempt to capture movies' categorical features' influence on their Bechdel performance would reveal potential gender biases in the film industry. Moreover, our result sought to provide audiences with an comprehensive lens to critically examine nuanced inequalities prevalent in the film industry.

3 Related Work

Prior to 2014, some investigations on female and male presence in literature works were conducted through manual effort, requiring multiple researchers' inspection through thousands of materials (Smith et al., 2011). Later, more data scientists have started to automate Bechdel tests through social network analysis, a computational approach based on graph theory in mathematics (Garcia et al., 2014; Agarwal, 2015; Selisker, 2015). Specifically, researchers extracted scripts from open datasource and assess characters' presence in the network through their social network features such as mean degree, closeness, and betweenness centrality (Agarwal, 2016). Specifically, Agarwal employed natural language processing in automating the Bechdel test, revealing a significant correlation between the importance of female presence in movies with their Bechdel performance (Agarwal, 2015).

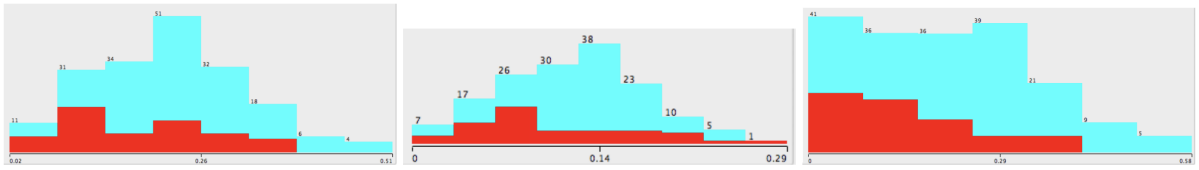

Figure 2: Data from Agarwal: Distribution of three SNA features (left to right): mean degree centrality, mean closeness centrality, and mean betweenness centrality of named women. Red histogram is for movies that fail and the Blue histogram is for movies that pass the third Bechdel Test. The histograms show that the average centralities of women are higher for movies that pass the Bechdel test.

In our experiment, we did not find strong correlation between movie's performance and their Bechdel scores. While increasing proportion of movies passed the test over years, we notice that this prevalent trend was not evident for Western, crime, and documentary films. In a project led by a group of students from Carnegie-Mellon University, similar trend of Bechdel test performance was observed, and the Bechdel test failed to show significant impact over a movie's performance as well. However, they noticed a stereotypical patterns in the plot descriptions for movies with different Bechdel ratings, corroborating our analysis on different genres' performance. (See Fig. 3.)

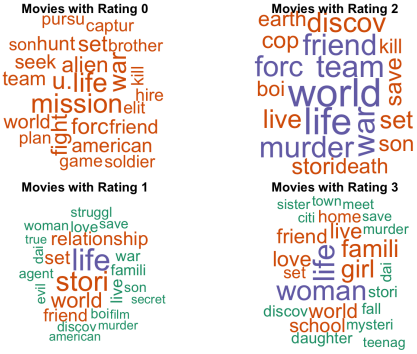

Figure 3: Graph from stat.CMU.edu: Word Cloud of Plot by Bechdel Rating

Movies with poor performance tend to have "american", "mission", "soldier" in their plot description while those that passed the test tend to have "woman", "love" "girl", "daughter".

In further discussion of related work, we will also compare our results to the 2020 paper "Passing the Bechdel Test and the Influence of Internet and Social Media Advertising on Seeing a New Movie Release" by Joshua Fogel and Kara Criscione. Although we do not have data on social media advertising of the films, the paper also discusses trends in the other metrics we do have data on, and describes some previous analysis done on the correlation between Bechdel test scores and these film metrics. We will discuss the results from Fogel & Criscione's paper when comparing them to our own results in the *Results & Visuals* section of this paper.

4 Description of Datasets

4.1 Bechdel Dataset

4.1.1 Dataset Origin & Access

This dataset was acquired from bechdeltest.com – a website that has been collecting Bechdel ratings of around 10,000 movies since its creation in 2008, and it has an average of 595 submissions per year. The website was still under active management from the website administrator at the time we obtained the data, around mid-October. The platform is an open-source system of obtaining Bechdel scores, enabling public contribution. Each accepted vision had comment sections allowing users to challenge the rating for a potential change; submitters are required to provide a comment justifying their rating, and other users can challenge and initiate further discussion around the score. For example, the movie *Dune* was initially considered failing the test, prompting other users to challenge the score by providing dialogues between character Shadout Mapes and Lady Jessica. The score was eventually raised to 3.

4.1.2 Dataset Description

We now provide a description of the columns available in the dataset. For each submitted movie, the dataset contained its visibility information, submission date, submitter ID, Bechdel rating, dubious quality (depending on users' review of the score), IMDbID, website ID, movie title, and release year:

1. visible: Has this movie been approved (currently only approved movies are returned when querying the data, so this value will always be 1).

2. date: The date this movie was added to the list

3. submitterid: The ID of the submitter. Since submitter information is currently not available through the API, this column is of no use.

4. rating: The Bechdel score. Number from 0 to 3 (0 means fewer than two women, 1 means no talking, 2 means talking about a man, 3 means it passes the test).

5. dubious: Whether the submitter considered the rating dubious.
6. imdbid: The IMDb id – a unique value IMDb uses to identify films.
7. id: The bechdeltest.com unique id.
8. title: The title of the movie. Any characters are HTML encoded.
9. year: The year this movie was released (according to IMDb).

4.2 IMDb Dataset

4.2.1 Dataset Origin & Access

The second dataset used in this project was obtained through an R library called omdbapi, the Open Movie Database. This library allows the user to make queries to the database based on the IMDb ID, which we obtain from the Bechdel dataset.

With these IMDb IDs, we obtain the data of all films included in the Bechdel dataset, and this forms the IMDb dataset we use in the analysis. These queries were made in late October, and the entire dataset we ultimately use was generated with this process in one session.

4.2.2 Dataset Description

The dataset contains many columns, a subset of which we use in our analysis. The most important columns are:

1. imdID: This allows us to join the IMDb data with the Bechdel data.
2. Year: We are given the year of release for each film.
3. Director: This contains a list of the directors of the film.
4. Writer: This contains a list of the writers of the film.
5. imdbRating: This is the average rating, on a scale from 1-10, rounded to the nearest decimal point, of ratings by users on the IMDb website.
6. imdbVotes: This describes the number of votes present for the film on the IMDb website: how many users rated the film.
7. BoxOffice: This describes the box office performance for each film, in dollars, if known.
8. Plot: A short description of the plot of the film.
9. All other columns: there are many more columns present, such as a list of actors, a link to an image of the poster of the for the movie, the language(s) the films are in, etc. These are either text data and difficult to analyze, contain too many nulls to use, or are otherwise not necessary to include.

5 Analytics Stages, Process

5.1 Data Cleaning and Profiling

In the Bechdel dataset, the titles were formatted for future operation such that its special characters were HTML encoded and its leading articles were removed and added at the end, delimited by a comma. We defined a transformation function so that the titles would appear in readable format for further analysis in PySpark. Moreover, using MapReduce, we dropped columns including "visible", "submitterid", and "id" while updating "date" to "year submitted" and "rating" to "Bechdel Score" to avoid confusion when joining with the other dataset. After profiling and dropping null records, our records decreased from 9903 to 9603 movies in the Bechdel dataset, which is still comparable to the original dataset.

Preliminary profiling was done in Spark. Peaking in 2010s, the website received comparable amount of submissions from users each year with a mean of 595.88 and all years are within one standard deviation from the mean. Moreover, we noticed most users submitted films they've recently watched by looking at the difference between film's year released and year submitted, confirming the relevance of our data.

The IMDb dataset also required substantial cleaning. In this dataset, we noticed that there were commas embedded in several of the fields, and found two ways to solve this: in Spark, adding a quote character fixed this problem when importing the data. In MapReduce, we used simple algebra to switch commas with a different delimiter within fields. The former was done when doing categorical joint analysis, and the latter produced a clean dataset that was used for the machine learning component of this project.

The R library omdbapi appears to have generated duplicate records in the first version of the dataset. Using MapReduce automatically took care of this problem for us, as the identical keys were combined into a single output. This reduced redundancy and decreased the number of rows by more than half.

With a clean dataset, we conducted some preliminary data profiling in both MapReduce and Spark. We calculated basic summary statistics such as the mean, median, mode, and standard deviation of film ratings by each decade and each year to see if there were any anomalies. There appear to be none, and the fluctuation of the average ratings appears mostly random over time.

Next, we moved on to additional data processing for the two sections of our joint analysis: the categorical analysis, and the machine learning. It was necessary to process the datasets differently for these two sections, as we needed a different subset of both rows and columns for each part.

When joining two datasets for the categorical analysis, escape commands were used to avoid comma delimiters in list value under IMDb rating. After dropping duplicate records and cleaning the data, we had 9534 lines of records. The overall schema was then consist of

"Title", "IMDb ID", "Year", "Bechdel Score", "IMDb Rating", "Genre", "Director", and "Writer." The joint dataset was used for further analysis.

In the machine learning component of the project, all text columns were removed (as binary logistic regression and random forests cannot handle non-numeric data). We thus kept only the numeric values in this version of the dataset. Datacleaning was done using MapReduce jobs on the two datasets. We retained the imdbID values in each dataset (so that a join would be possible), the Bechdel score from the first dataset, and the year, IMDb rating, IMDb votes, and box office columns in the latter dataset. To avoid nulls, it was necessary to subset the rows to only those containing data for all five of the columns we use. Next, it was necessary to create two separate datasets for the two different machine learning algorithms we implemented. The first, being a binary logistic regression, asks that the outcome variable be binary, 0 or 1. We thus mapped scores 0 through 2 to a "FAILURE" class, and assigned it a score of 0. Only a score of 3 is considered "PASSING", which gets a value of 1. The random forest classifier has no such requirement, so we keep the exact score breakdown. Thus, after the cleaning process of the datasets for machine learning, we obtain two datasets: each with four predictors and one outcome, but one where the outcome is binary, and one where the outcome is just the raw Bechdel score from the original data.

6 Results & Visuals

Out of curiosity, we first generated word clouds of the films in each Bechdel score category. Many renowned films still failed the simple test.

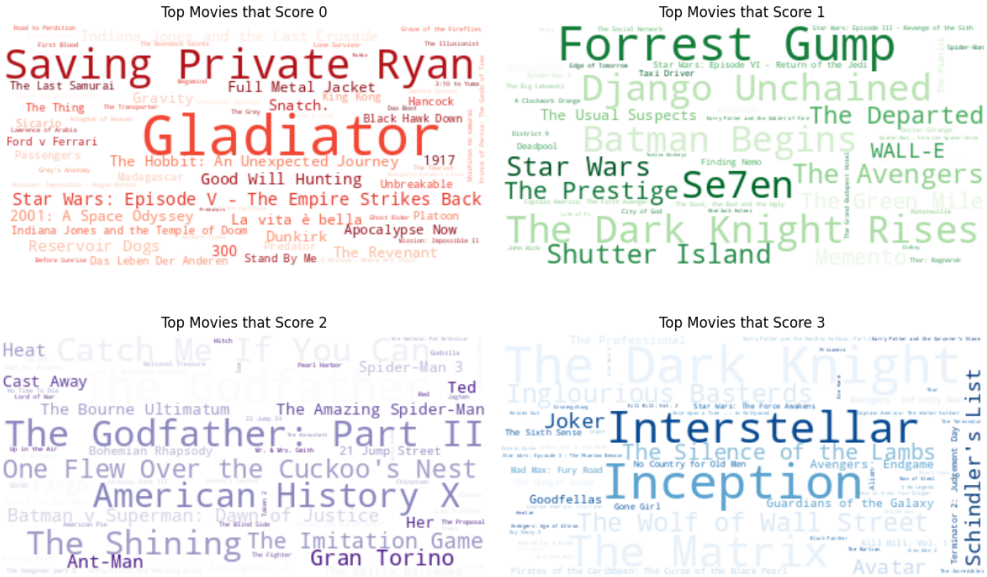

Figure 4: Word Cloud of Top Movies by Bechdel Rating

We next analyze the fluctuation of Bechdel scores throughout each year. The Bechdel scores over years showed a consistent increase in proportion of movies that passed the test.

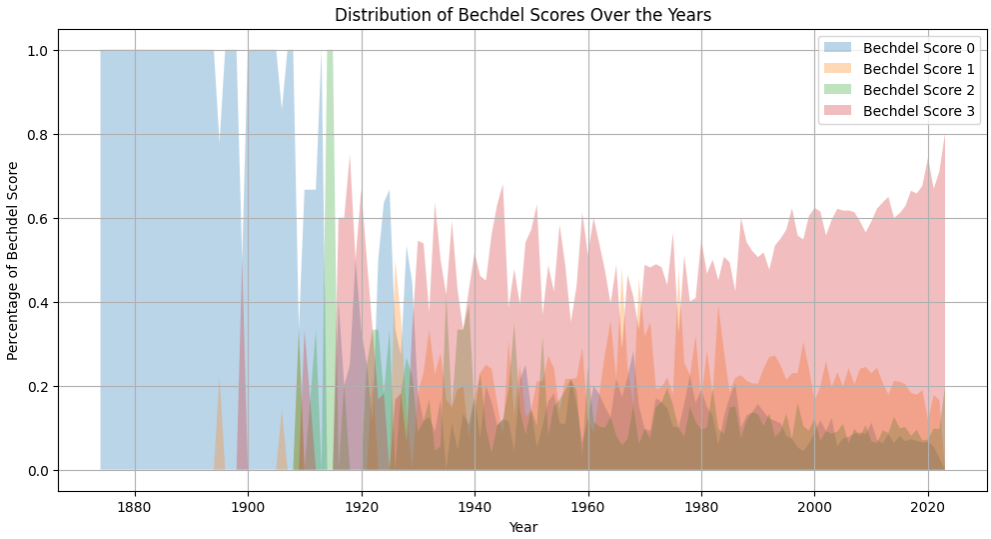

Figure 5: Increasing percentage of movies passed the test since the onset of 1970s.

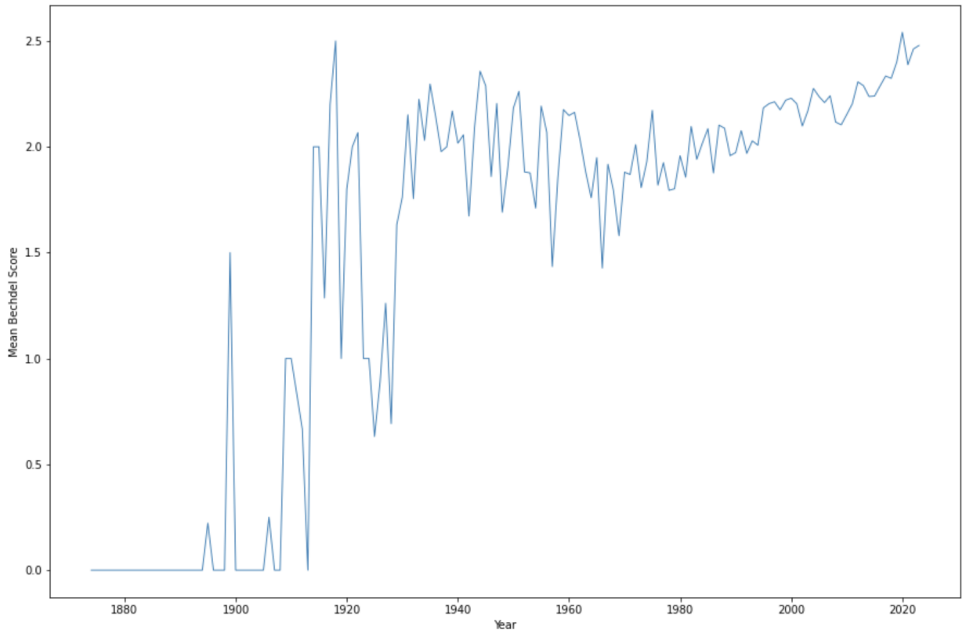

Figure 6: Average Bechdel score over the years.

6.1 Categorical Analysis

We employed PySpark in categorical analysis on our joint dataset.

With high expectations, we first evaluated the relationship between a movie's IMDb rating and its Bechdel score. Unfortunately, the results are not extremely meaningful: poor correlation with a value close to zero was found between two variables. We visualized the

relationship in two ways: with a box plot, and with a histogram. They both tell the same story as our quantitative analysis: that the distribution indicated that a movies' Bechdel performance had little impact on audiences' perception of its goodness.

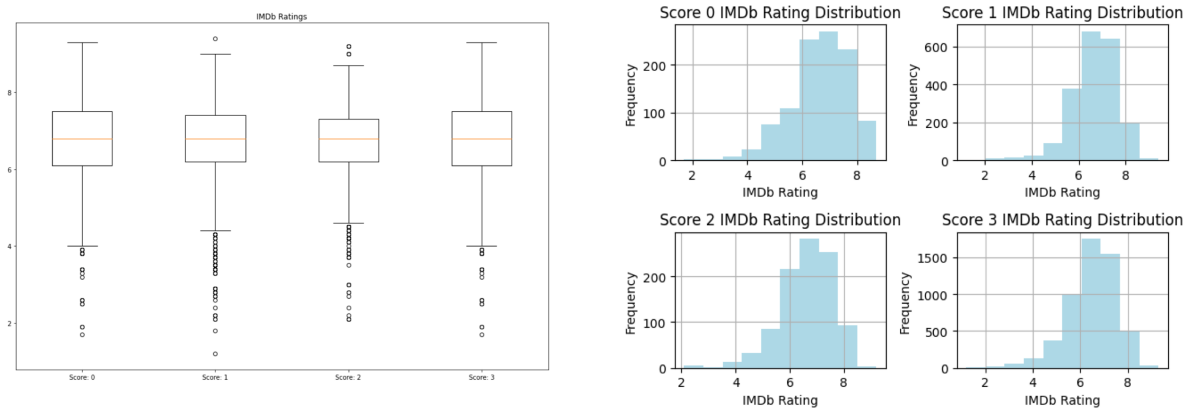

Figure 7: IMDb ratings appear to have similar distribution for different Bechdel scores

Our result indicated that a movie's Bechdel performance did not have significant influence over its performance. This conclusion seems to be in agreement with previous findings: in Fogel & Criscione's paper, they describe the relationship between the average viewer rating, and even the box office performance itself, as having no significant relationship with the Bechdel score (Fogel and Criscione, 2020).

However, we wondered why this result might be the case and conducted further analysis. To account for such a finding, we reconsidered the over IMDb rating distribution alone and discovered that, movies, especially popular ones, were likely to have a homogeneous rating distributions, meaning there would be similar percentage of people loving and hating it. Moreover, the rating distribution similarity could also be subject to submitter bias when users of bechdeltest.com were likely to watch a certain range of films.

Next, we borrowed a gender-guesser library and used it to analyze how a movie's director/writer's gender would influence its Bechdel scores. It was worth mentioning that the film industry had predominantly male directors/writers. In our recorded films, (leading) female directors/writers were consistently less than a quarter of (leading) male directors/writers over the decades. (From a set of data published on Statista, the reality had a even more biased distribution (Hunt et al., 2023).) We performed hypothesis testing on director/gender's effect on their movie's Bechdel performance and discovered that movies written by female are more likely to pass the test – the p value is essentially 0, meaning that the difference between likelihood of passing the test is *significantly* higher with female directors and writers. We thus concluded that while male directors/writers consistently outnumbered their female counterparts, female directors/writers demonstrated significantly higher rate of passing the test.

In both graphs below, from left to right, the histograms showed counts for male, female, and unknown. The gender-guesser library returned male/female based on the empirical gender distribution of the name and returned unknown when it was equally likely to be male/female. (Non-binary genders were not included in the library and might contribute to the error of the result.)

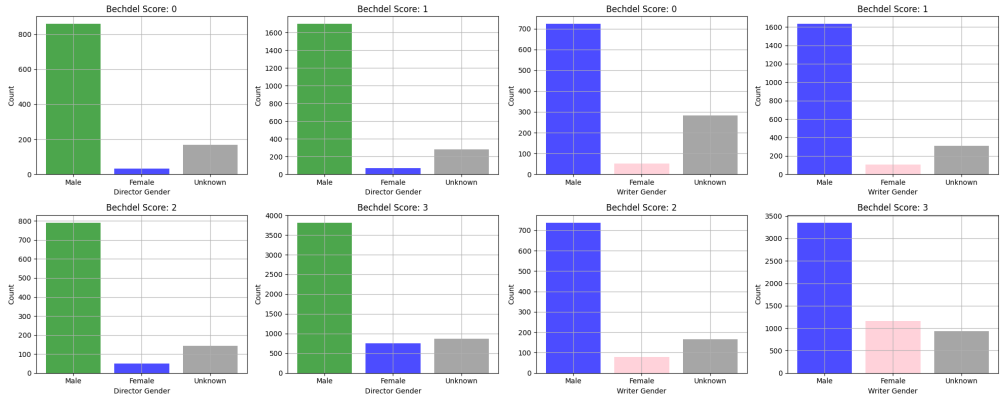

Figure 8: Bechdel Test Result for Different Director Gender

Figure 9: Bechdel Test Result for Different Director Gender

Lastly, we look at distribution of Bechdel scores across different genres. Among all 22 genres with over 100 submissions, drama has significantly more submissions (over thousands). Ranking their mean Bechdel scores, we noticed Short, Western, Documentary, and War films had a mean Bechdel scores of lower than 1.5 while more than half of other genres had a mean score greater than 2.

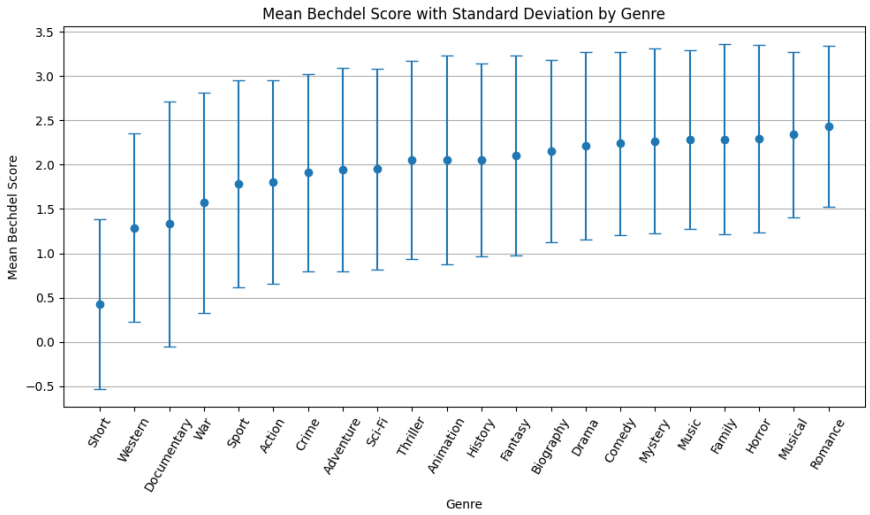

Figure 10: Mean Bechdel Score for Different Genres

We found the low Bechdel score for short films reasonable; a potential explanation for their poor Bechdel performance could be that their plots often revolved around a single

protagonist or topic, typically without explicit relevance to female characters, in conventional production. Another factor that could account for their low score was their low submission rate. For instance, the documentary films had 150 submissions in our recorded dataset, among one of the lowest submission rates.

6.2 Machine Learning

We also implemented two simple machine learning models to further examine the predictive ability the IMDb data has on the Bechdel score.

6.2.1 Data Processing for ML

6.2.2 Binary Logistic Regression

We first use a very simple 2-class classification model, a binary logistic regression. We use the following as predictors:

1. Year Released
2. IMDb Average Rating
3. IMDb Votes
4. Box Office Performance

For this analysis, as there are only 2 classes, we implement a cutoff of 2: either the film does not pass the Bechdel test with a score of 2 or below, or it does pass with a score of 3.

Using PySpark and the machine learning library, we implement a binary logistic regression model. We obtain the following table of coefficients and odds ratios:

Coefficients & Odds Ratios		
Predictor Name	**Calculated Parameter**	**Odds Ratio**
Year	1.53 e-2	1.015
IMDb Rating	1.62 e-2	1.016
IMDb Votes	-1.25 e-6	1.00
Box Office	1.79 e-9	1.00
Intercept	1.0	N/A

And we obtain an accuracy of around 58.5%, meaning around 58.5% of films were accurately predictable using our four predictors.

If the accuracy score were higher, we would be more interested in exactly how well the model performs. In this case, we could introduce a train/test split on the data and evaluate the model on the test set. For our results, we found that the predictors we chose did not

perform very well even on the training set itself, so we forgo this process.

We offer possible explanations of the relatively poor performance:

1. The first possible explanation is that these predictors are simply not very predictive of Bechdel scores. For instance, if there is no relationship (or at least, no monotone relationship) between year released/box office performance/IMDb metrics and the Bechdel score, that would explain why they make poor predictors. However, as we can see from the other visualizations, this might not be the case: as the year increases, so does the Bechdel score, in general, so there does exist a relationship.

2. A more likely explanation is that the Bechdel test itself is slightly ambiguous. "Good" and "accurate" portrayal of women in film is a complex problem, and the Bechdel test is an imperfect way to map this concept to a score (and indeed, probably any test would have flaws). Further, since the difference between a 0 and a 1 score is as simple as giving two characters names (if we decide to quickly name our characters 'Lisa' and 'Mary', does that mean our film suddenly has better female representation?), and since a difference between scoring a 2 and fully passing the test with a score of 3 could be the inclusion or exclusion of a couple of words, implementing a specific cutoff is also ambiguous. If the difference between a 2 and a 3 score is very slight, does the fault lie with the machine learning algorithm that our predictive accuracy is not ideal?

6.2.3 Random Forest Classifier

A similar story is told in the multi-class classification model with the random forest. As random forests, unlike binary logistic regression models, can easily handle multiple classes, we choose not to map the Bechdel score variable to a binary pass/fail, and instead elect to keep the exact score breakdown (0, 1, 2, and 3).

This makes for a more complex classification task, as the model must choose one class out of four instead of one out of two, but we find that the accuracy doesn't suffer too badly: we obtain a rate of approximately 56.8%.

Given that there are four possible classes to choose from, that the random forest model has only very slightly less accuracy than the logistic regression is not unimpressive.

Similarly to the logistic regression, the cutoff between each of the scores can be ambiguous, which could contribute to the poor performance of these selected predictors.

7 Conclusion

In this paper, using data on movies' Bechdel test performance and IMDb metrics, we found an increase in both the number and percentage of movies passing the test as the industry progressed. The result indicated a higher gender awareness in modern film production. By evaluating relationships between various movies' metrics and their Bechdel Performance, we

concluded that movies' Bechdel ratings had no significant impact on audiences' perception of its quality. Moreover, we discovered that among directors/writers of different genders, female directors/writers, while being outnumbered, are more likely to produce movies with greater female presence than their male counterparts. Seeing such results, we believe the film industry could continue raising their awareness around minority representation in both their produced movies and their working environment, as the advantages of having female writers and directors is clear. Across different genres, we found Short, Western, Documentary, and War films to have poorer performance limited to their time, themes, and conventional approach in production. Our result, consistent with Banerjee's finding from CMU, revealed an stereotypical associations between movies' plot and Bechdel performance. Movie that failed the test were commonly associated with conventionally masculine themes. We encourage retrospection on and avoidance of currently prevalent stereotypes in future film production, such as the usage of underdeveloped minority characters whose sole existence revolves around the protagonists.

Our study did not reveal a significant influence of Bechdel Scores on movie performance metrics. We speculated that the uniform distribution across different Bechdel Scores might stem from the existence of blockbusters with varying Bechdel Scores. As for future research, we suggested dividing genres into specific sub-categories by controlling for particular metrics. This could involve analyzing subsets like Marvel Movies (sharing the same production company), Independent Films (with similar production costs), and films within the IMDb Top 50 of 2023 (sharing comparable ratings within the same year). As the a film's production team had great impact over their Bechdel test performance, we proposed to examine Bechdel Score distribution within a set of female and male directors/writers who frequented in certain themes of movies (by tracking movies' genres and keywords in plots). The result could provide readers a more detailed narrative on gender awareness by examining who is telling what stories.

References

[1] Agarwal, A. and Zheng, J. and Kamath, S. and Balasubramanian, S. and Dey, S. A. Key female characters in film have more to talk about besides men: Automating the bechdel test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 830–840, 2015.

[2] Banerjee, D. and Fu, L. and Lai, C. and Quan, J. Gender inequality in movies over time through the bechdel test, 2022.

[3] Fogel, J. and Criscione, K. Passing the bechdel test and the influence of internet and social media advertising on seeing a new movie release. *International Journal of Arts Management*, 22(3):66–77, 2020.

[4] Garcia, D. and Weber, I. and Garimella, V. Gender asymmetries in reality and fiction: The bechdel test of social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 131–140, May 2014.

[5] Selisker, S. The bechdel test and the social form of character networks. *New Literary History*, 46(3):505–523, 2015.

[6] Statista. Movie director gender distribution in the u.s. 2011-2022, n.d.