Analysis of the Bechdel test and awards on IMBD scores

William Bevidas, Olachi Mbakwe, Claudia Silverstein

2023-12-13

Abstract

Introduction and Background

The movie industry has been around for over 100 years and along with it has come rating systems. There are many different types of ratings that can help define how good a movie is such as Rotten Tomatoes, Letterboxd, IMBD, etc. One category that does a good job at combining ratings from around the world are IMBD scores. Since these are scores from people around the world, we can point out what movies do better or worse and might even be able to explain some bias in these votes. One of the ways to measure a movies diversity is through the Bechdel test. There is also potential for bias in movies that have won awards at award shows.

We are investing a way to determine bias in IMBD scores through the use of the Bechdel test and award winning movies. We have multiple research questions that we would like to answer:

- Does a films score on the Bechdel test make a statistical difference in IMBD scores?
- Does the number of movie awards won by a film make a statistical difference in IMBD scores?
- Does the interaction of Bechdel test scores and number of awards make a statistical difference in IMBD score?

We will first discuss some additional background information and describe the methods we used to collect our data. We will then describe the collected data and share the results of analysis. Finally, we will close with discussion of results and some post-hoc analysis questions if there is a significant impact such as:

- Given the number of awards, is failing the Bechdel test associated with a higher IMBD score?
- Given the Bechdel test score, is winning more awards associated with a higher IMBD score?

IMBD Scores

The way IMBD scores are calculated is by taking the votes from registered IMBD users who voted on specific titles in the IMBD database using their account. Each person gets one vote per movie. Instead of taking the arithmetic mean, they use a weighted average based on the votes given and the rating can sometimes be influenced on unusual voting activity. Their method for calculating the final score is not disclosed but they claim that it works the best in keeping the scores reliable. These scores are helpful in helping people to decide what to watch, meaning it is a good measure of how good a movie is.

Award Ceremonies

Another way of understanding how well a movie can be is through award ceremonies. Three common award ceremonies that take place yearly are the Oscars, the Golden Globes, and the Bafta Awards. Getting an award from these ceremonies shows that you must have done something that was impressive in film, however this might not always be the case. Award ceremonies and Hollywood have been known to show bias against minority groups. One of the groups involved has been women and a big issue is the lack of actresses in the cast.

In a 2012 study, it was proven that from 1950 to 2006, the number of male main characters has outnumbered female characters by more than two to one in the most popular films during that time period. So, more popular movies who will probably be getting Oscars typically has less females than males. Along with that, a majority of Oscar voters are males, giving into even more potential bias as a man might vote for a movie that has more men in it. While there are awards for best lead actress, we are only going to be interested in awards given to the entire film, to try and find the underlying bias.

The Bechdel Test

The Bechdel test was created by cartoonist Alison Bechdel from one of their 1985 cartoons and is a measure of the representation of women in film and other fiction. It was meant to call attention to gender inequality in fiction (not just movies). The rules to pass a test are simple:

- The movie has to have at least two women in it
- These women must talk to each other
- These women must talk about something other than a man

It doesn't seem too hard to pass the film, yet there are many movies out there that fail. This test is able to break open movies and point out where some underlying bias is that might not have been noticed before. There can however, also be skewed results within test. For instance, *The Lord of the Rings: The Two Towers* actually passes the test, but only because there is a 5 second clip of a mother talking to her daughter. Yes there were two women talking and there is some debate on whether or not it was about a man (the daughter asks where her father is), but it doesn't mean the film deserved to pass as these two characters are not relevant to the story. Since this test might not be extremely accurate, it might be hard to justify any trends but we think it is an easy and useful way to recognzie where the underlying bias might be in the movies.

Previous Studies

While there are no studies to our knowledge doing exactly what we are doing, some people have used the Bechdel test in studies. In a 2021 paper written by Nelli Abdullaeva titled *The issues of an oversimplified analysis in assessing the representation of women in cinema*, they discuss how passing the Bechdel test doesn't necessarily mean it is a good movie or it gives an adequate representation of women in film. They compared Bechdel test scores with their own personal feelings on how the representation of women and compared the two. They found that 15 of the 30 movies they looked at passed both the Bechdel test, and the personal evaluation. Two movies failed the test but passed the evaluation, three passed but failed the personal evaluation, and seven failed both. There were a few controversial calls, but overall this simple test showed that 22 out of the 30 movies were correctly representative based on the author's personal feelings.

Data Collection

The overall goal of collecting data is to get a data set that contains the primary response (IMBD scores) and our two factors (Bechdel test scores and number of awards). We are not using a block or any other response attributes for this study. We were unable to find a data set that had all of these items, so we decided to create our own. We will make use of 5 different data sets that have been collected. The first data set consists of a list of 5,043 IMBD scores and their equivalent movies. There was other unnecessary data for this report included here as well. The next three data sets included 10,765 different Oscar awards, 7,991 different Golden Globe awards, and 4,176 Bafta awards that are either a person or movie and that either won that specific award or did not. The last data set has 1,794 movies each with a binary variable on whether or not it passes the Bechdel test. Depending on what movies match the Bechdel test, we should have a maximum of 1,794 movies to work with in this study.

The first step in making a data set to use is finding out how many awards each movie won. For each row of data, if the award is True then it got 1 award and if is False it got 0. This counts for all types of awards, as long as it the entire film getting the award and not an actor or actress. Then for each awards data set, we aggregated the rows and added up the awards to get how many awards were given to each specific movie at that award ceremony. From there, the data was joined into one table, where all cases were kept from each of the three tables. Any cases of missing data between the three were given a value of 0. This is because if a movie was not on another list, it was assumed that it wasn't apart of that award show since each of the original lists contained all known information.

Next, the awards are added up from each of the award ceremonies and this final list consists of 11,003 different movies/people who received awards along with how many awards they have won. In order to include all movies that have Bechdel test scores that we have access to, we join these two data sets together, along with the IMBD score data set. This leaves us with a data set consisting of 578 observations that include the movie, the number of awards won, the year it came out, its Bechdel test score, and its IMBD score.

While the data is a subset of the population (all movies) and we also did not pick this data set directly, we never used randomization to pick the movies or the treatments and because of this our study will be an observational study. This means that we cannot make any statements about causation, but we can depict an association between each of the factors and the response. Although we have minimum control over the measurement units, we are able to account for some bias unlike if we were running an experiment. As a note, our measurement units are movies, and since these movies have the treatments that are already given, the experimental units are the same as the measurement units.

The final item to use for analysis is dealing with the amount of awards. Since there are between 0 and 15 awards, we are going to use categories to simplify the levels of this factors. We decided to use none, low, medium, and high to determine the amount of awards given to the movies. None is defined as 0 awards, low is between 1 and 5 awards, medium is between 6 and 10 awards, and high is between 11 and 15 awards.

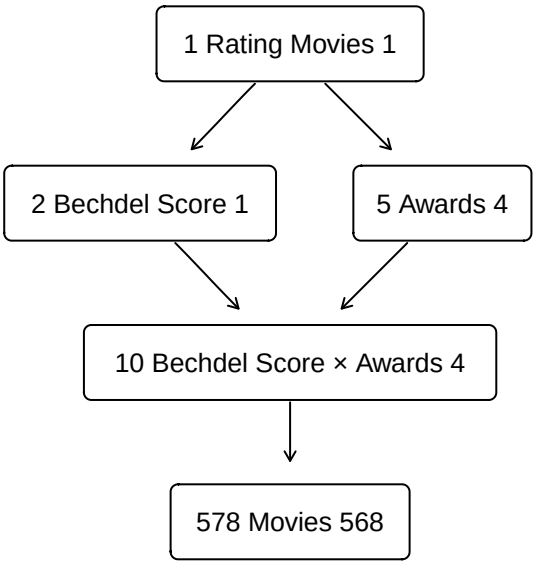
Analytical Methods

To analyze our data and answer our research questions we will use R (version 4.3.0) and make use of ANOVA methods, in particular a full factorial (two-way) ANOVA model.

Appropriateness of ANOVA

Our response variable is the arithmetic mean score given to a movie according to the IMBD website, so this would be a continuous response. We also have two factors: the Bechdel test score and the amount of awards won. The Bechdel test score is categorical with two levels (True or False). The amount of awards won is also categorical and we decided to simplify the model to make it have 5 levels instead of 15 (no awards, a few awards, some awards, many awards, and a lot of awards).

From the Hasse diagram, we can see the two factors (both fixed) and their interaction. Our sample size from the population of movies is 578, so there are many degrees of freedom to estimate all main effects, interaction, and error terms. All of this points to using ANOVA methods to answer our research questions. For this study, we will be making use of a two-way ANOVA model. This makes sense to use because we have two factors of interest and we want to explore the potential interaction between them. This is more appropriate than a one-way model as we can distinguish between the main effects and interaction terms. Also, since the two factors are not attributes of the movies themselves and just outcomes caused by the movie, we cannot use either of these as a block, so a block design would also not be of interest here. If we were to use a block design, we could have chosen a variable such as budget that would be intended to block out the effect of more expensive movies winning more awards.



Hypothesis

The null hypothesis for our study is that there is no statistically significant impact on IMBD scores of movies. The alternative hypothesis is that there is a statistically significant impact on the IMBD of movies. We express these hypotheses as:

$$H_0 : y_{ijk} = \mu_{..} + \epsilon_{ijk}$$
$$H_A : y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$$

This equation along with the Hasse diagram both represent the two-way ANOVA model for our study. The $\mu_{..}$ representes the baseline effect of obtaining a score on IMBD, α_i represents the main effect of the Bechdel test score, β_j represents the main effect of the amount of awards, and $\alpha\beta_{ij}$ is the interaction between Bechdel score and awards. The error term given as ϵ_{ijk}, represents the variability within each of the individual movies as well as any outside sources that are not accounted for in the model.

Addressing the Multiple Comparisons Problem

Within each hypothesis test, we will use an Unusualness Thresheold equivalent to 0.1.

Exploratory Data Analysis

Exploration 1: Awards Distribution

When we watch movies, awards are a way to tell us which ones are exceptional. But how often do movies get these awards?

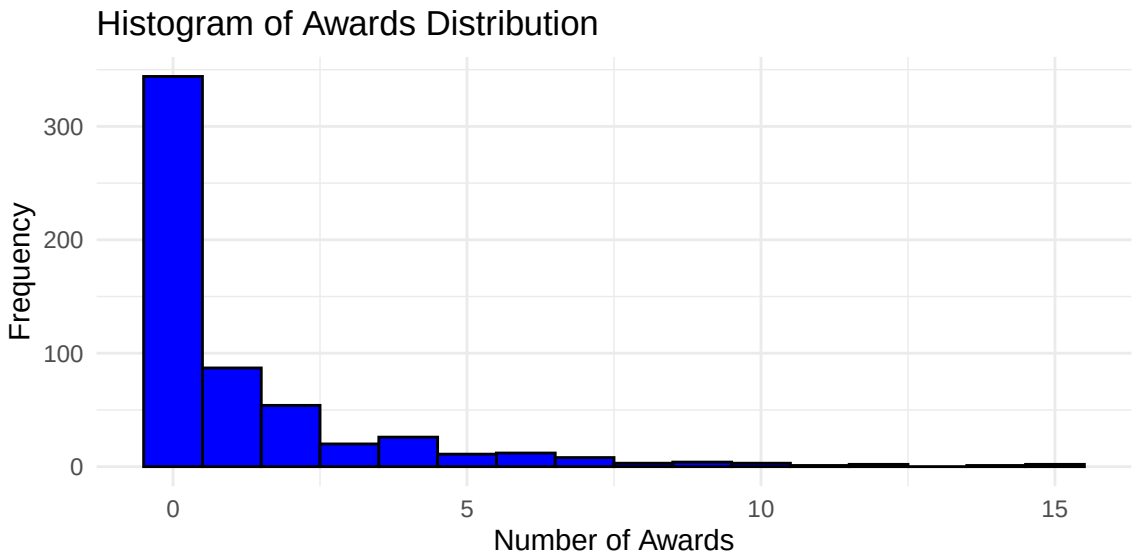


Figure 1: Histogram of Awards Distribution

The histogram above provides a clear picture of how awards are distributed among movies. We can immediately see that a significant number of movies don't receive any awards. As we look for movies with more awards, they become less and less common. In fact, it's quite rare for movies to receive a large number of awards. To make our analysis easier to understand, we've grouped the movies based on the number of awards they've received. This helps us see patterns and compare different sets of movies.

Given the observed distribution, the awards have been categorized into the following levels for analysis: - **No Award**: Movies that have not received any awards. - **Few Awards**: Movies that have received 1 to 2 awards. - **Some Awards**: Movies that have received 3 to 6 awards. - **Many Awards**: Movies that have received 7 to 10 awards. - **Numerous Awards**: Movies that have received 11 or more awards.

Exploration 2: IMBD Scores of Movies

The histogram above shows us the frequency of movies across different IMBD score ranges. From it we can see that most movies in the data set have scores around the 7 to 8. There are fewer movies with very high scores (close to 9) or lower scores (around 5 or 6) this means most movies have average scores, and fewer movies are rated as very good or bad.

We'll keep this in mind as we analyze how scores relate to awards and the representation of women in these films.

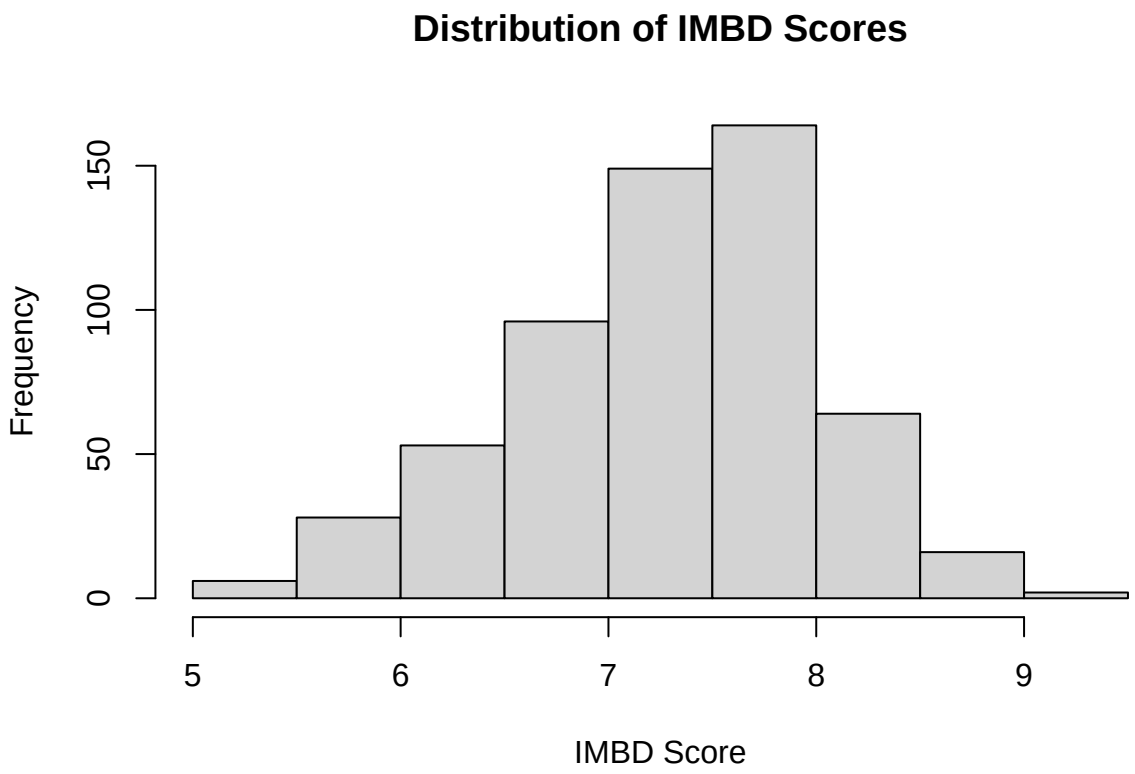


Figure 2: Histogram showing the distribution of IMBD scores for movies

Exploration 3: IMBD Scores and the Bechdel Test

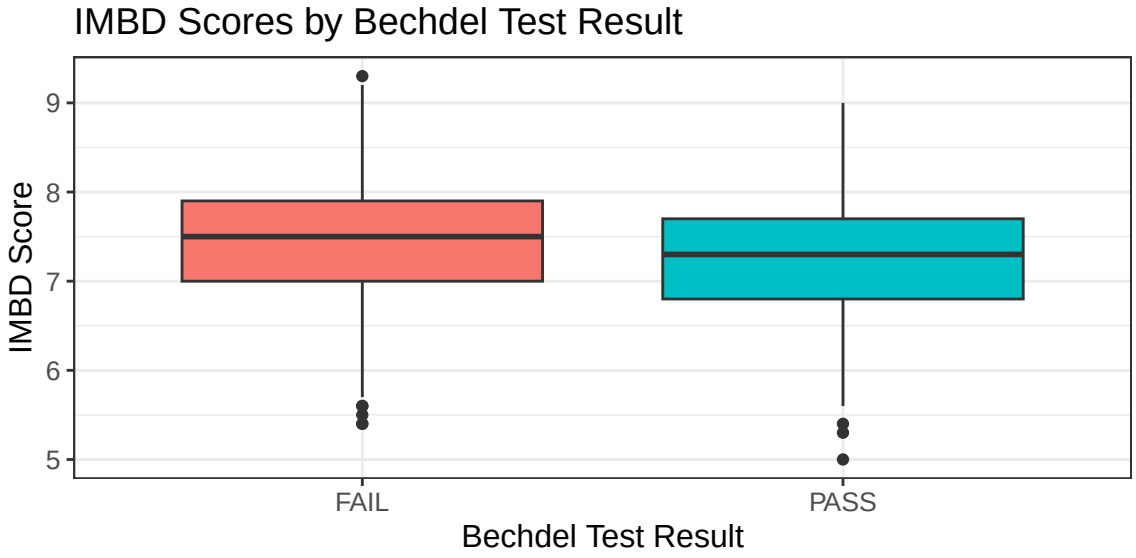


Figure 3: Box Plots of IMBD Scores by Bechdel Test Result

From the box plot above The median IMBD score for movies that pass the Bechdel test seems to be slightly higher than for those that fail. There is a wide range of scores for both categories, but the 'Pass' category has a slightly tighter interquartile range, indicating that scores are more concentrated around the median. Lastly there are a few outliers in both categories, indicating some movies scored much lower or higher than

the typical range.

This visual suggests that, on average, movies that pass the Bechdel test might be more favorably reviewed, but the difference is not stark. It's a reminder that while the Bechdel test can signal something about the content of a movie, it's not a definitive measure of quality or viewer satisfaction.

Exploration 4: The Impact of Awards on IMBD Scores

The box plot below segments movies according to the number of awards they've received and displays their range of IMBD scores within each category.

we turn our attention to the relationship between the number of awards a movie has won and its IMBD score. IMBD scores, which range from 1 to 10, reflect the average viewer's rating of a film, and we're curious to see if winning awards correlates with higher ratings.

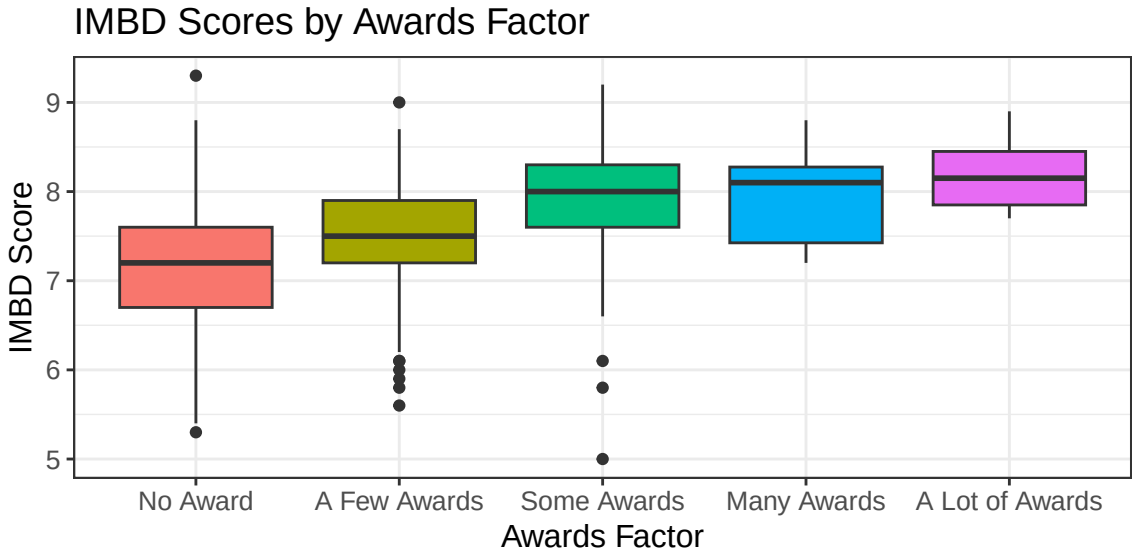


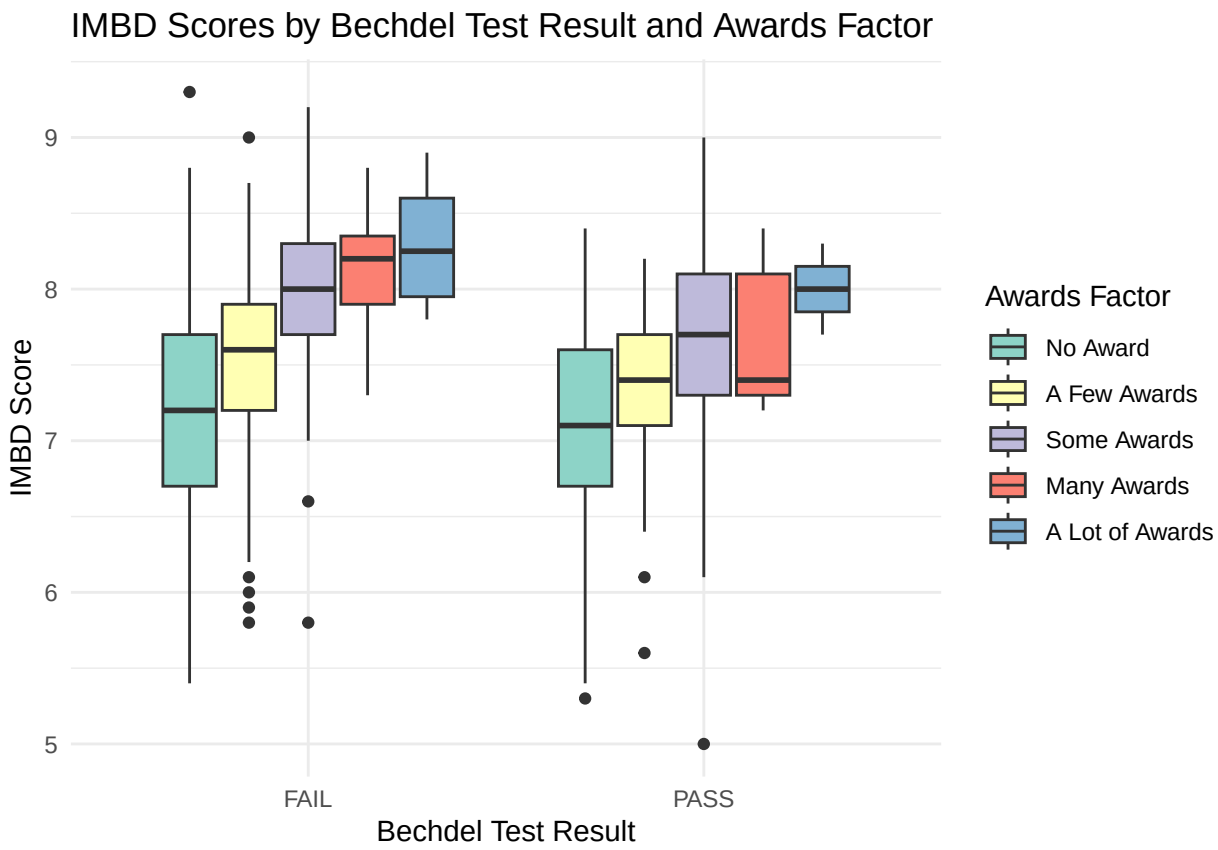
Figure 4: Box plot showing IMBD Scores by Awards Factor

From the box plot, we see that Movies that have not won any awards tend to have a wider range of IMBD scores, with a number of outliers indicating variability in viewer ratings. As the number of awards increases, the median IMBD score appears to increase slightly, suggesting that movies with more awards might be more favorably viewed by the audience. Interestingly, the highest award category (11+ awards) does not necessarily have the highest median score, indicating that a large number of awards does not always correspond to the highest viewer ratings.

This visualization suggests that while there may be a positive relationship between the number of awards and IMBD scores, it is not a strict one-to-one correlation. Other factors not captured by award counts may also play a significant role in determining a movie's IMBD score.

Exploration 5: IMBD Scores by Bechdel Test Result and Awards Factor

We have created a box plot that lays out IMBD scores based on whether movies pass or fail the Bechdel test, as well as how many awards they have won. This dual comparison provides a more nuanced view of the data.

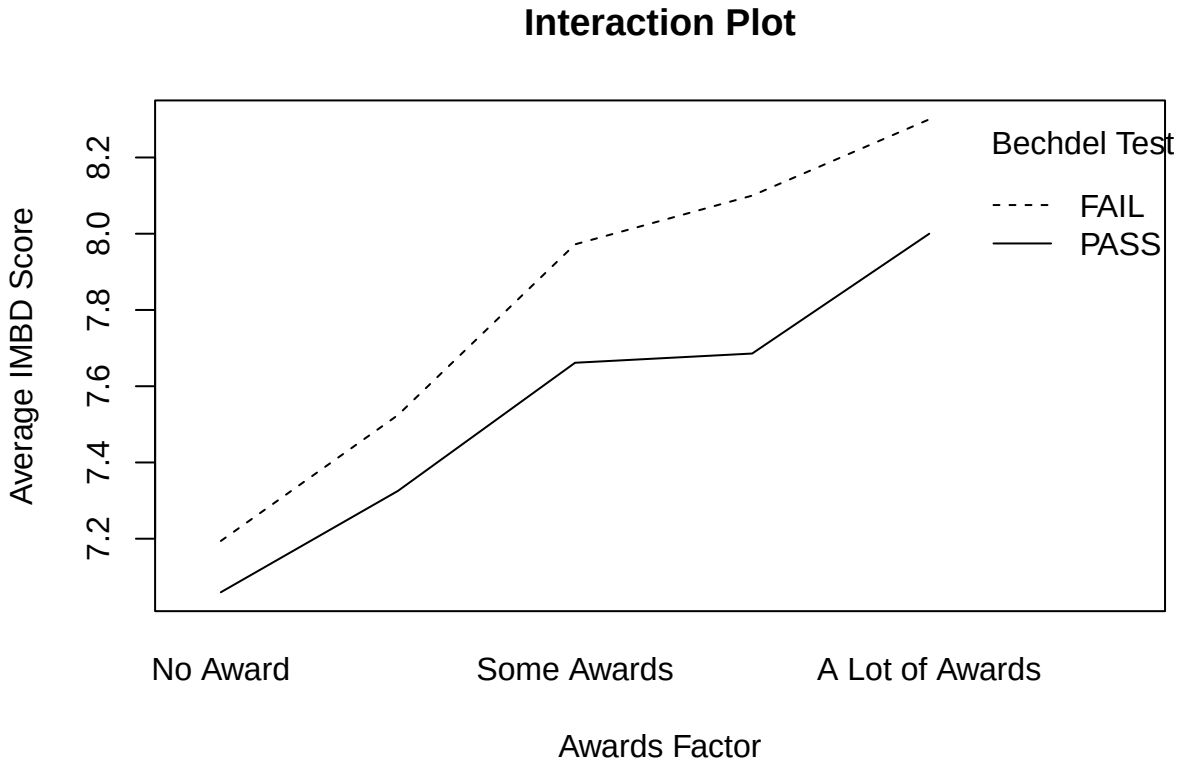


Across both Bechdel test results, movies with no awards have a wider range of IMBD scores, indicating variability in how viewers rate them. For movies that pass the Bechdel test, as the number of awards increases, there seems to be a slight trend towards higher IMBD scores. Interestingly, movies that fail the Bechdel test show a less clear pattern between awards and IMBD scores, suggesting that other factors might influence the viewer ratings of these films.

This visual analysis allows us to observe that while awards seem to be associated with higher IMBD scores for movies passing the Bechdel test, the story isn't as straightforward for movies that fail it. It hints at the complex relationship between movie ratings, gender representation, and critical acclaim.

Exploration 6: Interaction Between Awards and Bechdel Test Results

The interaction plot below helps us visualize whether the relationship between the number of awards and IMBD scores is different for movies that pass the Bechdel test compared to those that fail.



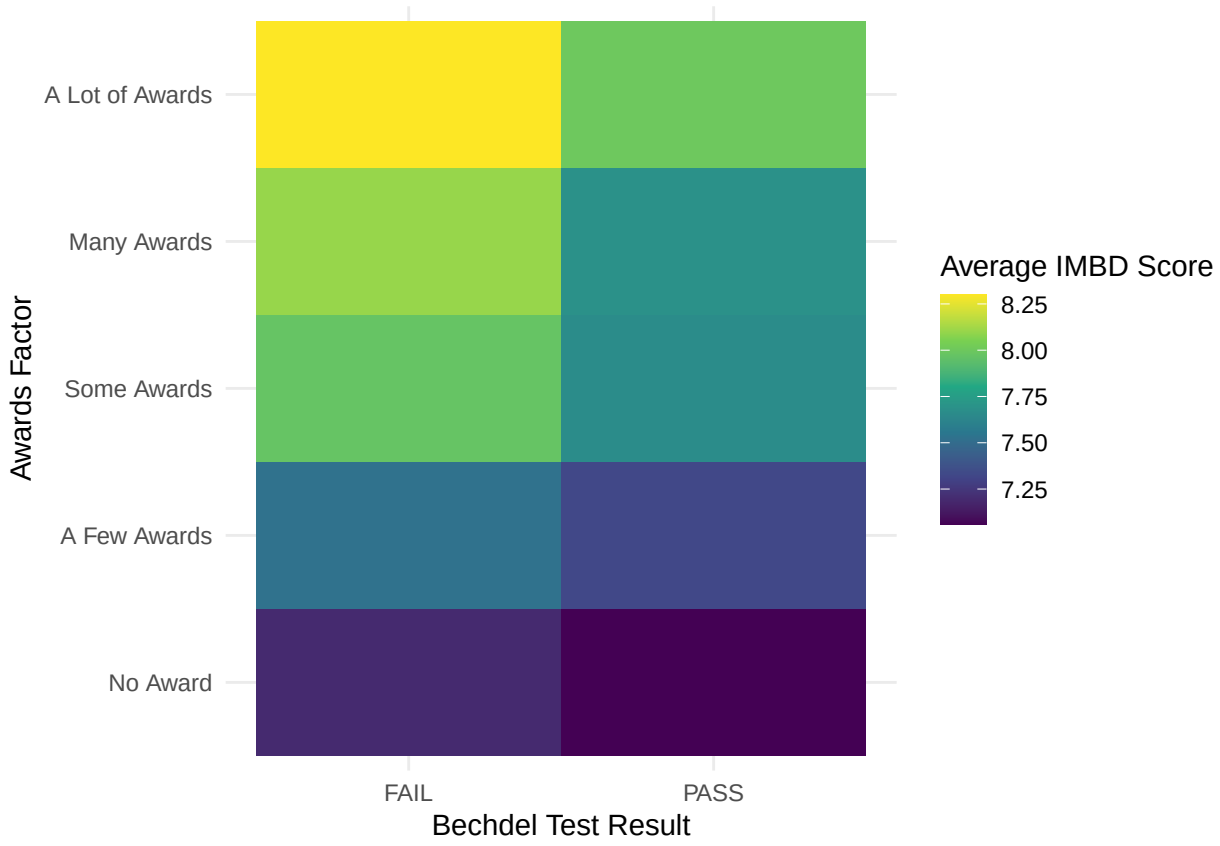
Movies that pass the Bechdel test (solid line) generally have higher average IMBD scores than those that fail (dashed line) across all awards categories. The influence of awards on IMBD scores appears to be more pronounced for movies that pass the Bechdel test, particularly in the '3-6 Awards' and '11+ Awards' categories.There seems to be a notable increase in average IMBD scores for movies that pass the Bechdel test as the number of awards goes from '7-10' to '11+' awards, which is less evident for movies that fail the test.

This pattern indicates a potential synergistic effect, where both awards and positive Bechdel test results may be associated with higher viewer ratings. However, it also suggests that the Bechdel test alone does not explain all the variability in IMBD scores, and winning awards is an important factor regardless of Bechdel test outcome.

This interaction suggests that awards and positive representation of women (as measured by the Bechdel test) might collectively influence audience appreciation of movies. The interaction effect, particularly pronounced in films with a high number of awards, indicates that these two factors may not be independent in their impact on IMBD scores.

Exploration 7: Visualizing the Influence of Awards and Bechdel Test Results

The following heat map allows us to see at a glance how the average IMBD scores differ across movies that pass or fail the Bechdel test and how many awards they've won.



This visualization helps us understand that there is a positive association between movies that fail the Bechdel test, receive a higher number of awards, and achieve higher IMBD scores. This pattern highlights the potential impact of gender representation and critical acclaim on audience ratings.

Exploration 8: Distribution of Movies by Bechdel Test and Awards

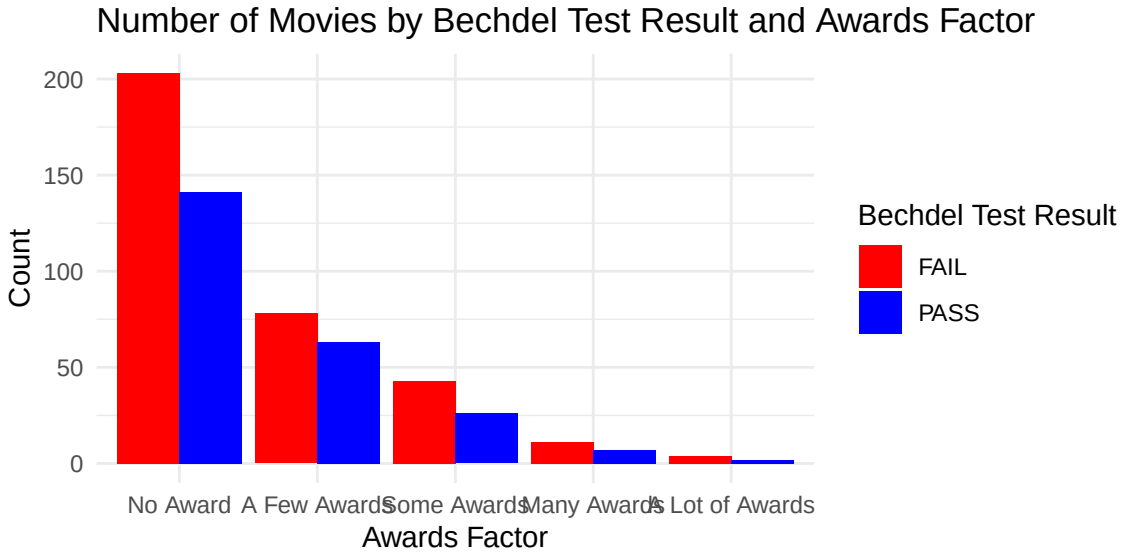


Figure 5: Bar chart showing the number of movies by Bechdel Test Result and Awards Factor

It suggests that movies which fail the Bechdel test might be more likely to receive awards, or conversely, that movies with more awards tend to fail the Bechdel test more often.

Exploration 9: Plot of IMBD Scores by Awards Factor and Bechdel Test

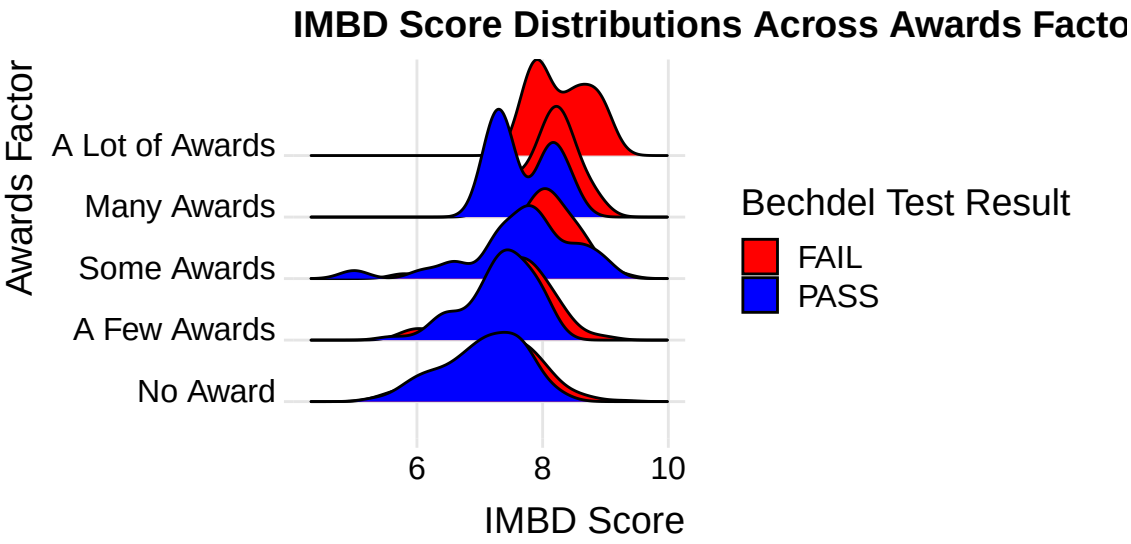


Figure 6: Ridge plot showing IMBD Score Distributions Across Awards Factor

The comparison between movies that fail and pass the Bechdel test shows different distribution shapes, with failing movies showing a tighter concentration of higher scores, especially in the higher awards categories.For movies with no awards, the distribution of scores is wider, suggesting a greater diversity in audience ratings regardless of the Bechdel test result.

Exploration 10: Awards Over Time by Bechdel Test Result

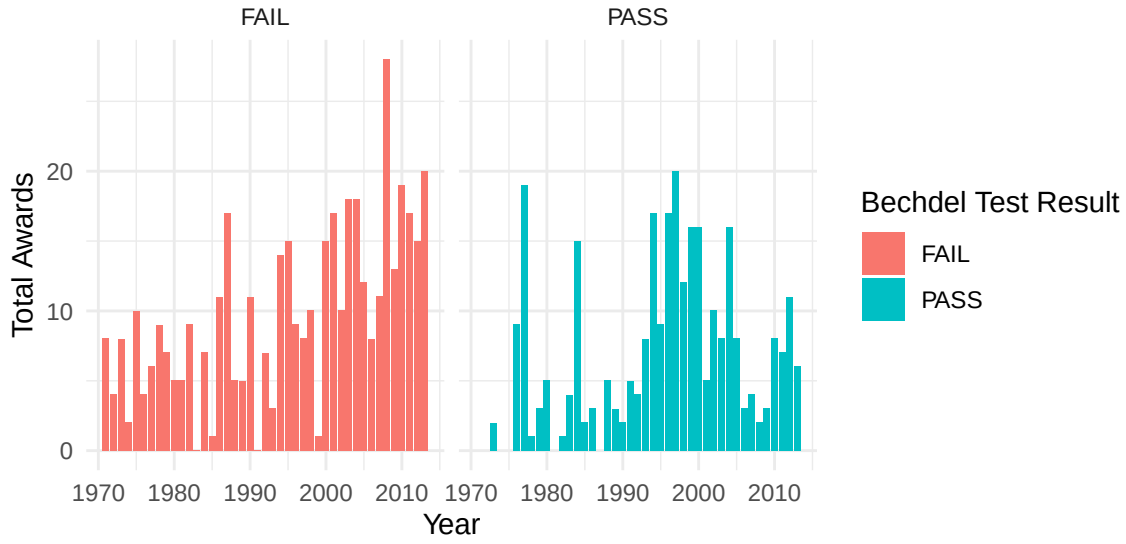


Figure 7: Stacked bar chart showing the total number of awards by year and Bechdel Test Result

Movies that fail the Bechdel test have, in several years, received a higher total number of awards compared to those that pass. This could be indicative of historical trends in the film industry's recognition practices.

This visualization helps to contextualize the relationship between the film industry's award patterns and the representation of women over time. It suggests that while there has been a historical gap in recognition, there may be a trend towards more equitable recognition for movies that provide better representation of women, as indicated by their passing of the Bechdel test.

Exploration 11: IMBD Score Distributions by Awards Factor

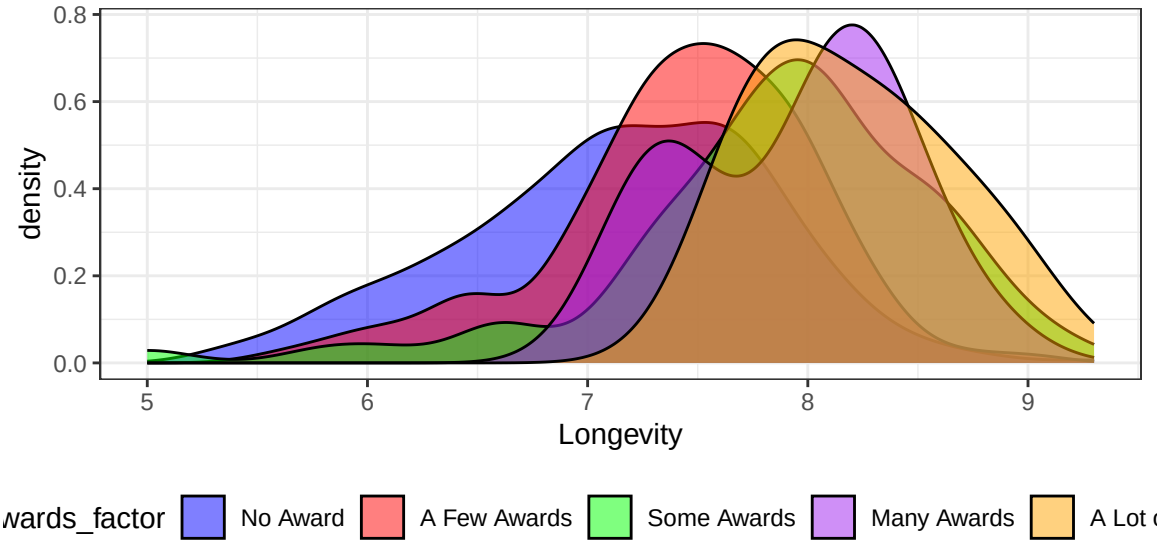


Figure 8: Density plot of IMBD scores across awards categories

The distributions show how the IMBD scores are spread within each awards category. We can see that some awards categories, such as '7+ Awards', tend to have a higher density around higher IMBD scores, suggesting that movies with more awards may be rated more favorably. The overlap between the distributions of different awards categories indicates that while there may be a trend, the number of awards is not the sole determinant of IMBD scores.

Exploration 12: Shadowgram Visualization of IMBD Scores

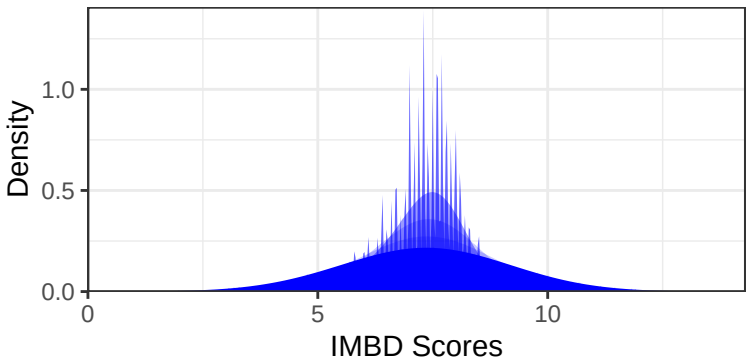


Figure 9: Shadowgram of IMBD Scores

The visualization shows a multi-layered perspective of the IMBD scores, with peaks at certain scores suggesting common ratings given by viewers.The spread of the scores across different layers indicates the diversity in the viewers' ratings, with some scores being much more common than others.

Exploration 13: Descriptive Statistics of IMBD Scores by Bechdel Test Result

Table 1: Summary Statistics for IMBD Scores by Bechdel Test Result

	n	Min	Q1	Median	Q3	Max	MAD	SAM	SASD	Sample Skew	Sample Ex. Kurtosis
FAIL	339	5.4	7.0	7.5	7.9	9.3	0.741	7.411	0.740	-0.373	-0.097
PASS	239	5.0	6.8	7.3	7.7	9.0	0.593	7.221	0.685	-0.453	0.172

From Table 13, we discern that movies failing the Bechdel test have a marginally higher median IMBD score, a subtle yet intriguing finding. The median score, an indicator less sensitive to outliers, suggests that movies with less female presence or those that do not emphasize female-centric conversations are not necessarily disadvantaged in terms of viewer ratings. However, the range of IMBD scores (from minimum to maximum) for Bechdel-failing movies is broader, indicating a more varied reception amongst viewers.

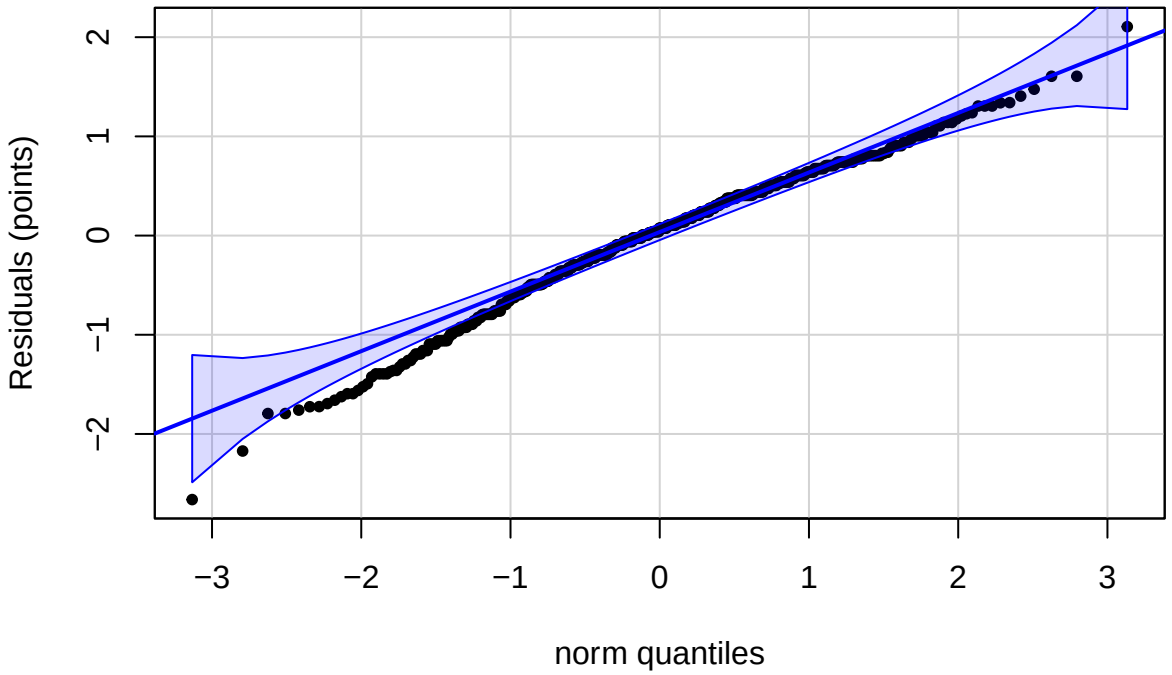
Conversely, movies that pass the Bechdel test exhibit a slightly lower median, suggesting that a focus on female representation does not always correlate with higher ratings. Nonetheless, the more concentrated interquartile range (Q1 to Q3) for Bechdel-passing films implies a consistency in scoring, perhaps indicating a more uniform viewer appreciation.

The skewness of both distributions leans negatively, hinting at a longer tail of lower scores. Yet, it is the excess kurtosis where we notice a stark contrast: Bechdel-failing films have a negative excess kurtosis, implying fewer outliers than a normal distribution, whereas Bechdel-passing films show a positive excess kurtosis, indicating a presence of outlier scores that are unusually high

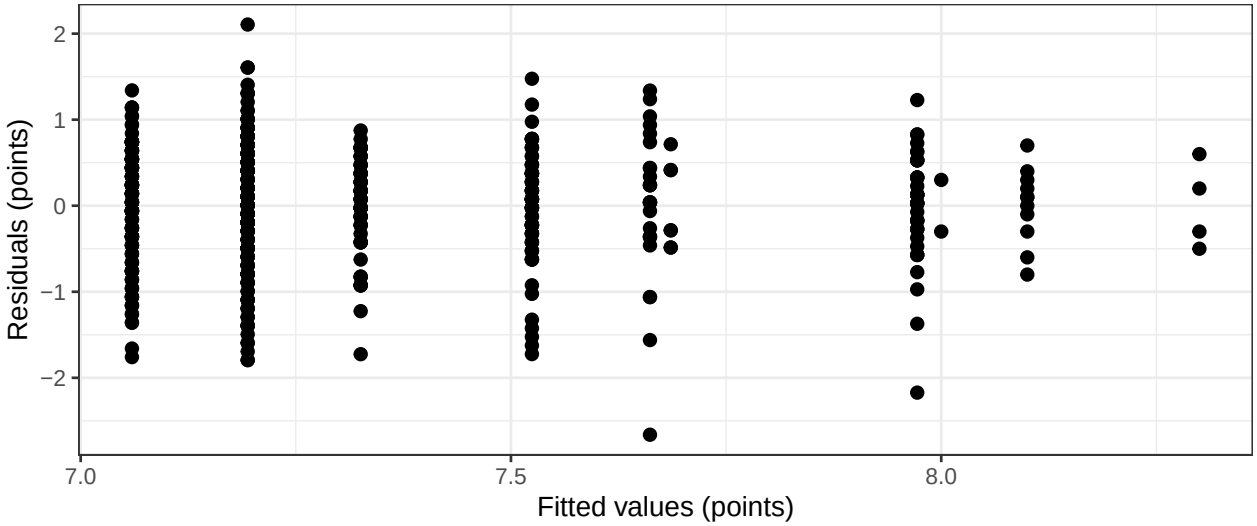
Results

Assumptions

In order to test the assumptions for using a parametric test, we are going to look at the normality and the homoscedasticity. Since the scores given on IMBD's website are completely random of when they were given, we do not need to look at the measurement order. We could look at the measurement order with respect to the year that it came out, but that wouldn't make sense when looking at IMBD scores, since a lot of these movies came out long before people began giving scores.



Looking at this first assumption of normality, it appears that there are many points over the 90% envelope. This happens mainly on the lower end, but it is still enough to consider this assumption not met. This will have to be looked into, but first we need to assess the equal variance assumption.

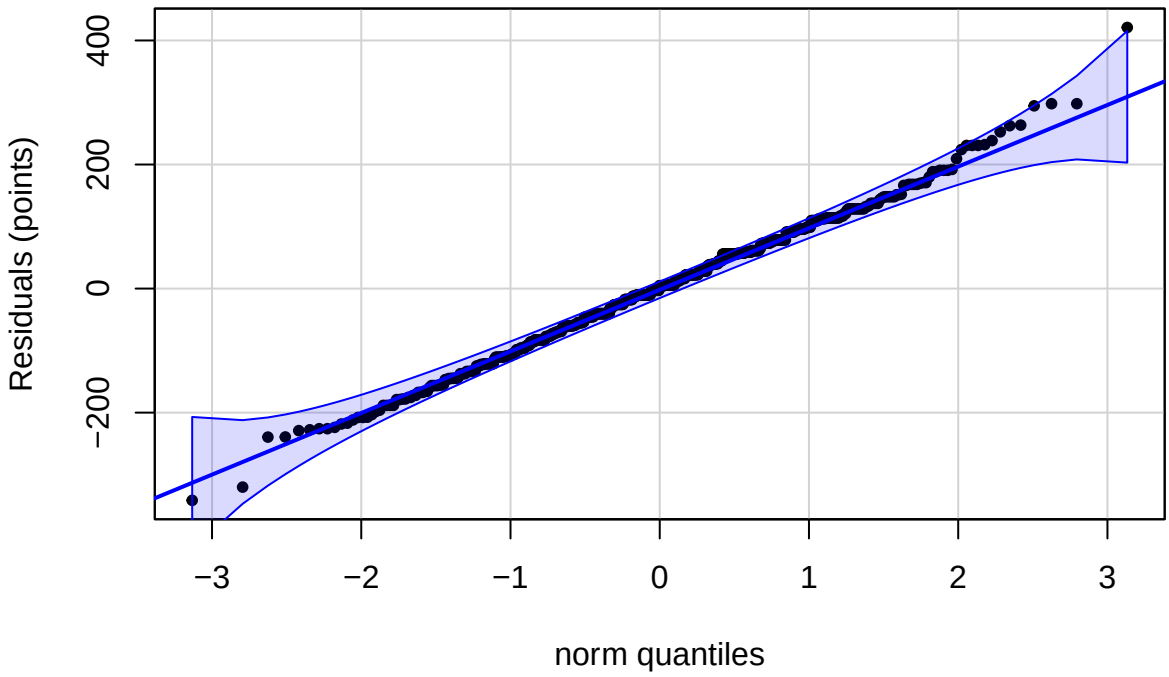


This assumption would be on the edge of being met. The main concern here are the three strips that have less than 4 points. Since these are here, it is hard to asses whether or not the assumption is met, but since there is no fanning or funneling throughout the entire strip chart, we think it is safe to assume this assumption is met and we can continue, keeping in mind the possiblility of this being incorrect.

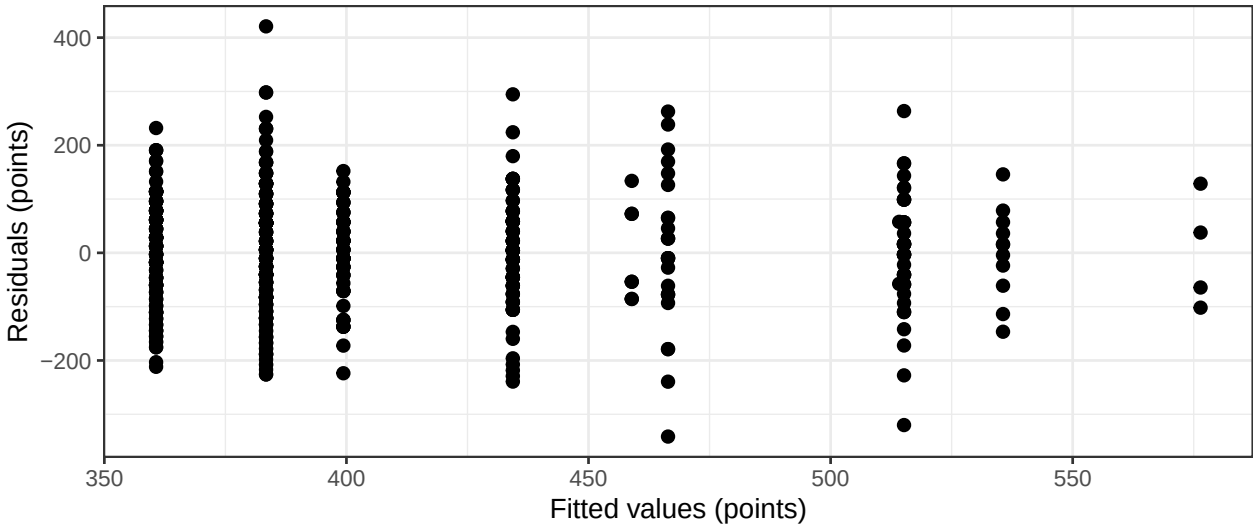
Transformations

Some of the possible transformations that we looked into did not help. These included squaring, square rooting, and logging the response variable. While searching for other possible ways to help, we found that

cubing the response was the best. Here, we are going to reconsider both assumptions made and use this transformation for the rest of the paper.



After applying the transformation, it is easy to see that it worked. Now, mostly all the points are within the envelope, and since we have so many there is no need to worry about these going over. With this transformation, the normality assumption is met.



Looking at the homoscedasticity of the residuals, it looks almost the same as before. We still show caution with the three strips of points that are smaller, but we believe it is safe to continue on with the assumption being met points swap places Since the two assumptions are now met, we can use this transformed that to now look at the Omnibus report and begin to answer our research questions.

Omnibus

Table 2: Modern ANOVA Table for Study

Source	SS	df	MS	F	p-value	Eta Sq.	Omega Sq.	Epsilon Sq.
movies.code	148165.8	1	148165.813	13.6334	0.0002	0.0234	0.0214	0.0217
awards_factor	1240276.1	4	310069.027	28.5308	< 0.0001	0.1673	0.1600	0.1614
movies.code:awards_factor	21597.6	4	5399.401	0.4968	0.7381	0.0035	0.0000	0.0000
Residuals	6172958.3	568	10867.884					

Note. Computer rounding has made the p-value look like zero.

Table 3: Post Hoc Tukey HSD Comparisons

	Difference	Lower Bound	Upper Bound	Adj. p-Value
PASS-FAIL	-32.512	-55.269	-9.755	0

Post Hoc Analysis

Table 4: Post Hoc Tukey HSD Comparisons Across All Factors

	Difference	Lower Bound	Upper Bound	Adj. p-Value	Factor
PASS-FAIL	-32.512	-55.269	-9.755	0.000	movies.co
A Few Awards-No Award	45.843	11.752	79.934	0.000	awards_fac
Some Awards-No Award	121.635	76.665	166.606	0.000	awards_fac
Many Awards-No Award	131.020	48.589	213.452	0.000	awards_fac
A Lot of Awards-No Award	179.084	38.695	319.474	0.000	awards_fac
Some Awards-A Few Awards	75.792	25.704	125.880	0.000	awards_fac
Many Awards-A Few Awards	85.177	-0.154	170.508	0.010	awards_fac
A Lot of Awards-A Few Awards	133.241	-8.870	275.352	0.019	awards_fac
Many Awards-Some Awards	9.385	-80.845	99.616	0.997	awards_fac
A Lot of Awards-Some Awards	57.449	-87.657	202.555	0.695	awards_fac
A Lot of Awards-Many Awards	48.064	-112.649	208.776	0.865	awards_fac

Comparing Bechdel Scores Given the Amount of Awards

Comparing Awards Given the Bechdel Score

Discussion

Limitations

Future Work

There is potential here for this project to handle future work. First, every year new award shows are presented.

References

https://en.wikipedia.org/wiki/Bechdel_test#cite_note-7

https://www.jahonline.org/article/S1054-139X(12)00069-9/fulltext

https://help.imdb.com/article/imdb/track-movies-tv/weighted-average-ratings/GWT2DSBYVT2F25SK?ref_=ttrt_wtavg#

https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#

https://blogs.fu-berlin.de/abv-gender-diversity/2021/12/13/the-bechdel-test-and-gender-equality-in-the-film-industry/

Author Contributions