Analysis of the Bechdel test and awards on IMBD scores

William Bevidas, Olachi Mbakwe, Claudia Silverstein

2023-12-15

Executive Summary

To understand the relationship of the factors of a movies Bechdel test score and the amount of respective awards given to how they perform on IMBD scoring websites, we have conducted an observational study to assist in understanding these relationships. Using the 2 given levels of Bechdel test scores (Pass/Fail) and 5 different levels of amount of awards given (None, A Few, Some, Many, A Lot), we found that the main effect of the Bechdel test score as well as the amount of awards had a statistically significant impact on the score on IMBD's website. Additional post hoc analysis revealed that when a movie passes the Bechdel test, they tend to get a lower IMBD score. We were also able to find out that when a movie recieves less than 2 awards, they also tend to get a lower IMBD score.

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 investigating 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:

- Is failing the Bechdel test associated with a higher IMBD score?
- Is winning a lot of 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 mean 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 it is described as 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 woman talking to a little girl. 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 recognize 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 that it gives an adequate representation of women in film. They compared Bechdel test scores with their own personal feelings of the films representation of women. 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.

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.

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 is 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, each list 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 not apart of that award and therefore receieved 0 awards.-

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.

Appropriateness of ANOVA

Our response variable is the weighted 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 in Figure 1, 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.

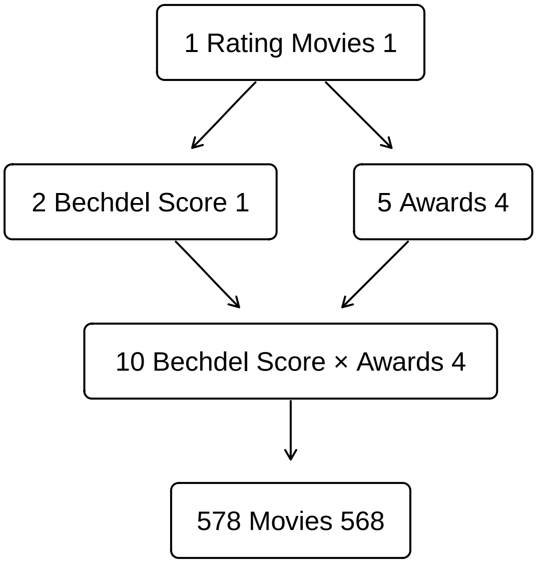

Figure 1: Hasse Diagram

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.

Type I Error Control

We have decided to set our overall Type I error rate at 10%. We decided this for statistical testing without being overly conservative in order to not mask genuine effects.For multiple comparisons, we will maintain the Simultaneous Confidence Interval error rate at the same level using Tukey's HSD. Our Unusualness threshold for declaring results as statistically unusual will be set at 0.01 (1%).

Exploratory Data Analysis

In our data it is important to note that they 339 movies Fail the bechdel test while 239 Pass.

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?

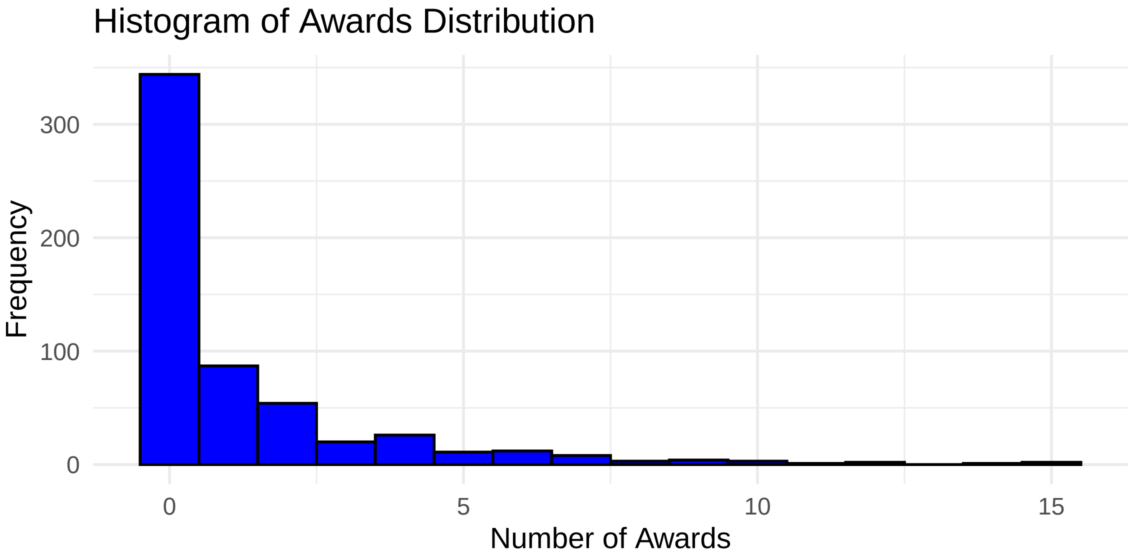

Figure 2: Histogram of Awards Distribution

The histogram in Figure 2 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. And 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.
- **A lot of 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

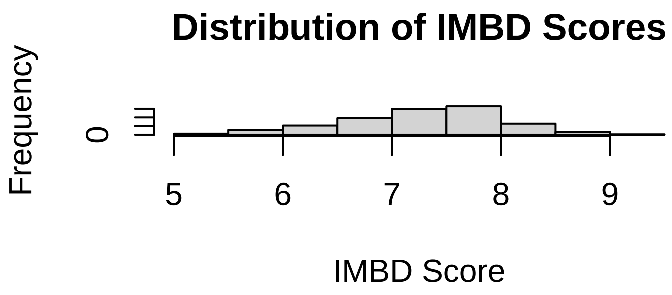

Figure 3: Imbd Score Distribution

The histogram in Figure 3 shows us the distribution of IMBD scores across a selection of movies. From it we can see that most movies in the data set have scores around 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 median scores. This could imply that while many films are deemed satisfactory, only a select few are rated as exceptional or poor

Exploration 3: IMBD Scores of Statistics

Table 1: Summary Statistics for Imdb Scores

	n	Min	Q1	Median	Q3	Max	MAD	SAM	SASD	Sample Skew	Sample Ex. Kurtosis
No Award	344	5.3	6.700	7.20	7.600	9.3	0.741	7.139	0.696	-0.255	-0.172
Some Awards	141	5.6	7.200	7.50	7.900	9.0	0.593	7.435	0.593	-0.612	0.599
A Few Awards	69	5.0	7.600	8.00	8.300	9.2	0.593	7.855	0.737	-1.238	2.570
A Lot of Awards	18	7.2	7.425	8.10	8.275	8.8	0.445	7.939	0.492	-0.168	-1.360
Numerous Awards	6	7.7	7.850	8.15	8.450	8.9	0.519	8.200	0.456	0.306	-1.698

Note: Descriptive statistics calculated for Imdb score rating

This table in figure above is the statistical summary of IMDB scores categorized by the number of awards received by movies. Movies that receive no awards have the most number of movies (n=344) with the lowest median IMDB score observed (7.2). But noticeably movies that receive Numerous awards have the least number of movies and median score of 8.15

Exploration 4: IMBD Scores and the Bechdel Test

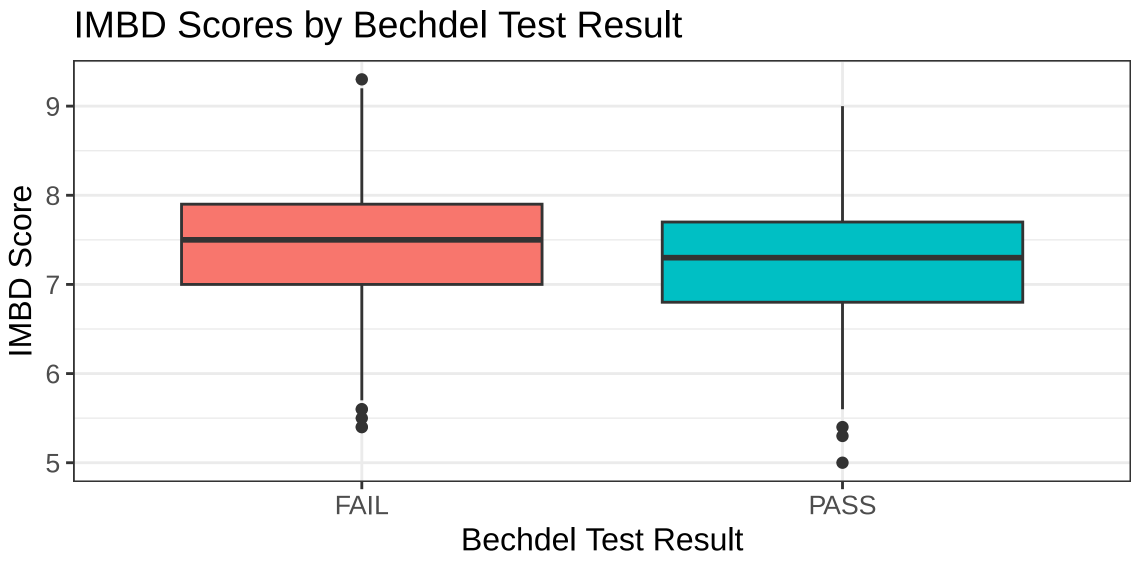

Figure 4: Box Plots of IMBD Scores by Bechdel Test Result

From the box plot in Figure 4, the median IMBD score for movies that pass the Bechdel test seems to be slightly lower 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 movies that fail 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 5: The Impact of Awards on IMBD Scores

The box plot in Figure 5 segments movies according to the number of awards they've received and displays their range of IMBD scores within each category.

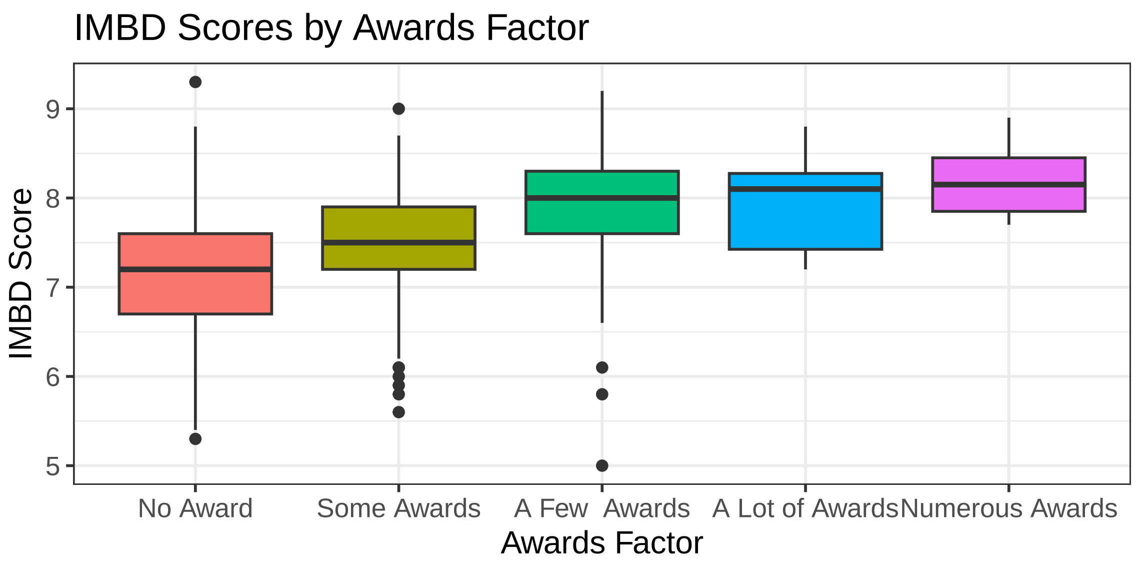

Figure 5: Box plot showing IMBD Scores by Awards Factor

From the box plot, we see that Movies that have not won any awards tend to have a wider range of IMBD scores. 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. This visualization suggests that there may be a positive relationship between the number of awards and IMBD scores.

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.

Exploration 6: IMBD Scores by Bechdel Test Result and Awards Factor

We have created a box plot in Figure 6 that lays out IMBD scores based on whether movies pass or fail the Bechdel test, as well as how many awards they have won.

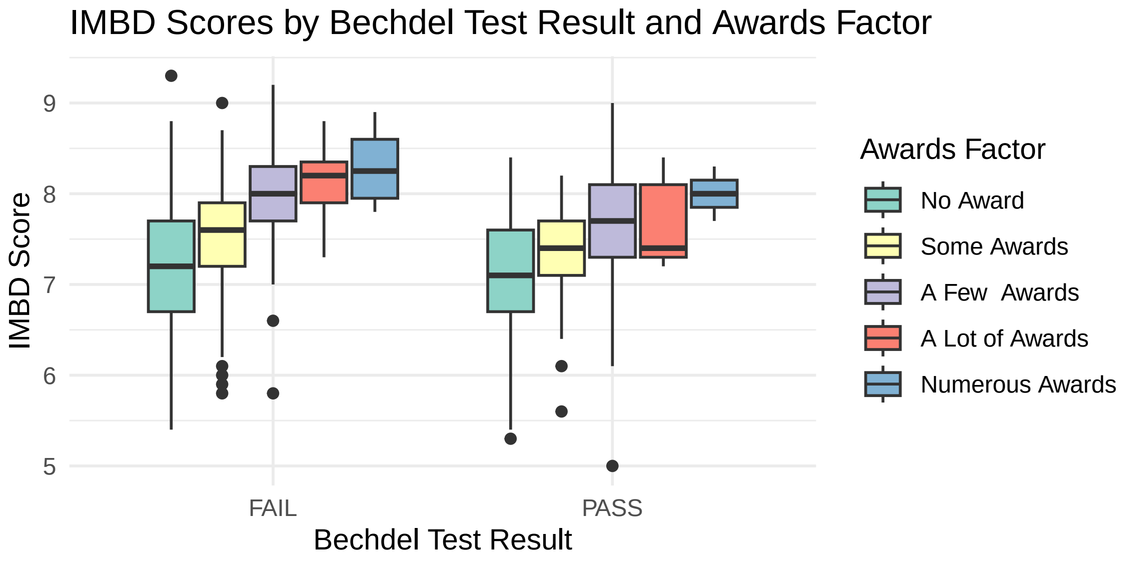

Figure 6: Box plot showing IMBD Scores by Bechdel Test Result and Awards Factor

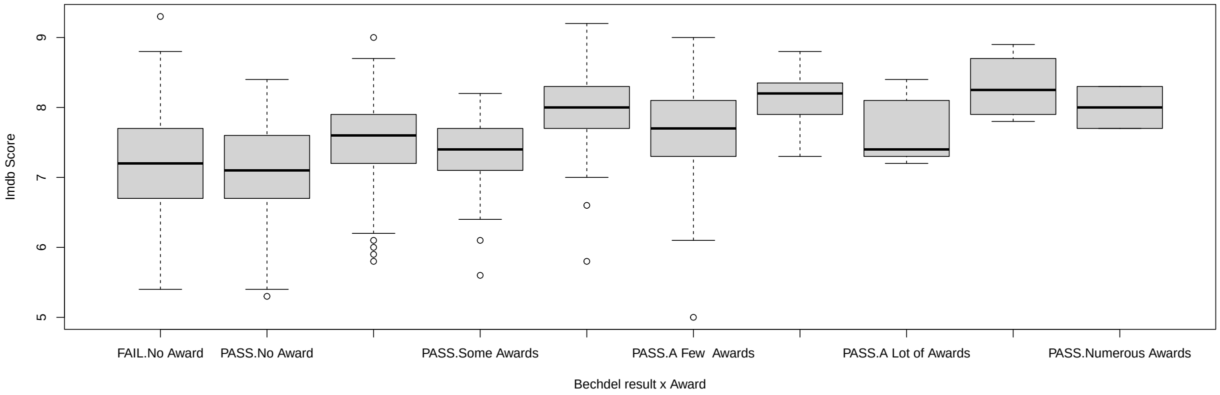

Figure 7: Box plot of IMBD Scores by Bechdel Test Result and Awards Factor

Across both Bechdel test results, Movies show a general trend where an increase in the number of awards correlates with higher median IMBD scores. Also movies that passed the Bechdel test exhibit a less pronounced difference in median scores across award categories.

Exploration 7: Interaction Between Awards and Bechdel Test Results

The interaction plot in Figure 8 & 9 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.

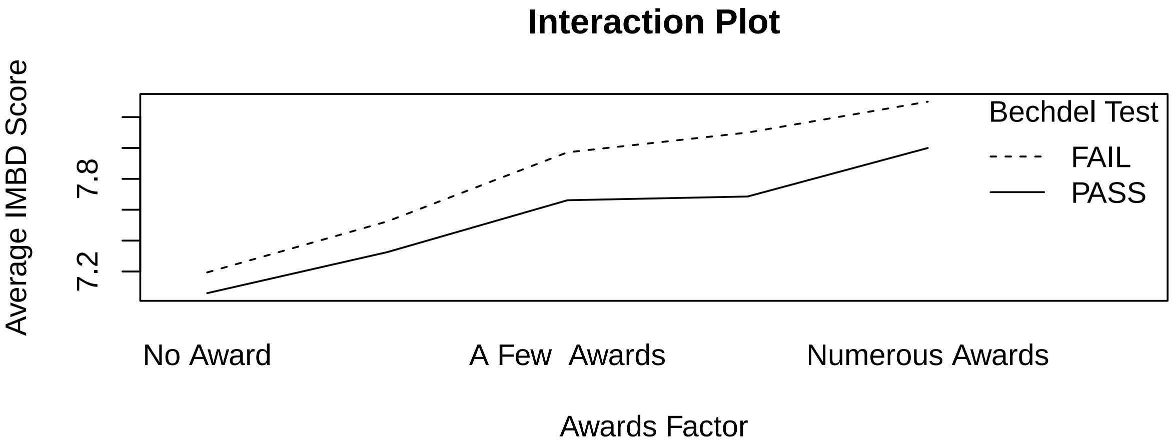

Figure 8: Interaction plot between IMBD Scores, Bechdel Test Result, and Awards Factor

Movies that fail the Bechdel test (dashed line) suggests that for movies failing the Bechdel test, the average IMBD score tends to increase with the number of awards received. Movies passing the Bechdel test shows a similar trend, with average scores increasing as movies receive more awards. However, the slope of the line is less steep compared to movies that fail the Bechdel test, especially between ‘A Few Awards’ and ‘A Lot of Awards’ categories.

The interaction between the two factors is evident in the differing slopes of the lines. While both lines trend upward, indicating that more awards generally correspond to higher IMBD scores, the rate of increase differs based on the Bechdel test result. This suggests that the Bechdel test result may modulate the effect of awards on IMBD scores.

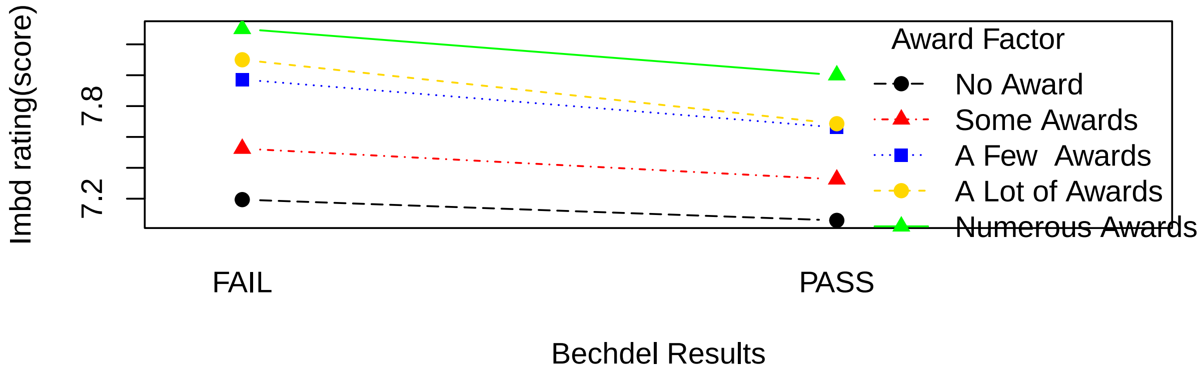

Figure 9: Interaction plot between IMBD Scores, Bechdel Test Result, and Awards Factor

Looking at another view of the interaction plot the figure 9 presents a comparison of average IMBD scores for movies, segmented by Bechdel Test results ('FAIL' and 'PASS') and across varying levels of award recognition. For movies that fail the Bechdel test, the plot reveals a general upward trend in IMBD scores with an increase in awards received. However, the slope of the increase varies between award levels. For movies that pass the Bechdel test, a similar trend is noticeable, but the 'Numerous Awards' category shows a particularly steep increase, suggesting a strong correlation between receiving numerous awards and higher IMBD scores for these movies.

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 8: Visualizing the Influence of Awards and Bechdel Test Results

The heat map in Figure 10 allows us to see at a glance how the IMBD scores differ across movies that pass or fail the Bechdel test and how many awards they've won.

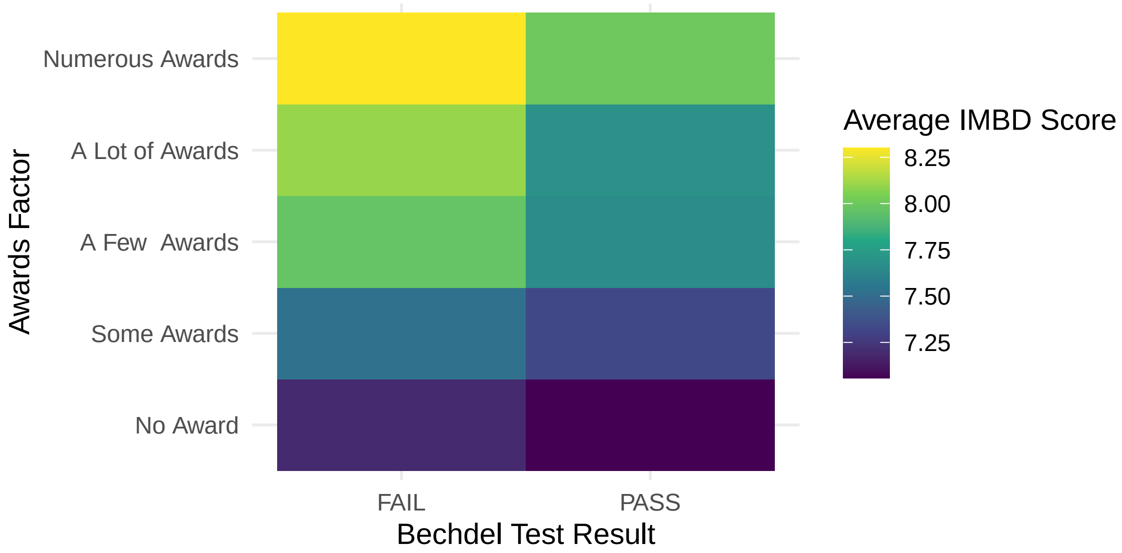

Figure 10: Heat map showing the average IMBD Scores by Bechdel Test Result and Awards Factor

This visualization shows that tiles under the 'PASS' column generally exhibit warmer tones compared to the 'FAIL' column, suggesting that movies passing the Bechdel test tend to have higher IMBD scores. Moreover, a gradient shift is observable as one moves from 'No Award' to 'Numerous Awards', indicating a potential trend where more awarded movies receive higher ratings which is consitent with previous exploration

Exploration 9: Distribution of Movies by Bechdel Test and Awards

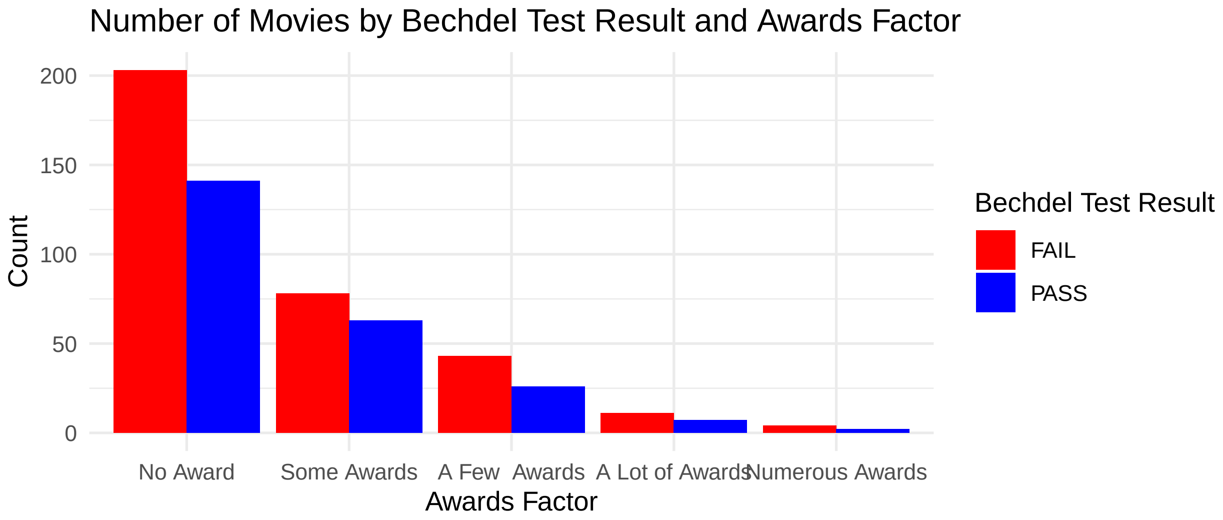

Figure 11: Bar chart showing the number of movies by Bechdel Test Result and Awards Factor

The histogram in Figure 11 suggests that movies which fail the Bechdel test are more likely to receive awards, or conversely, that movies with more awards tend to fail the Bechdel test more often.

Exploration 10: Plot of IMBD Scores by Awards Factor and Bechdel Test

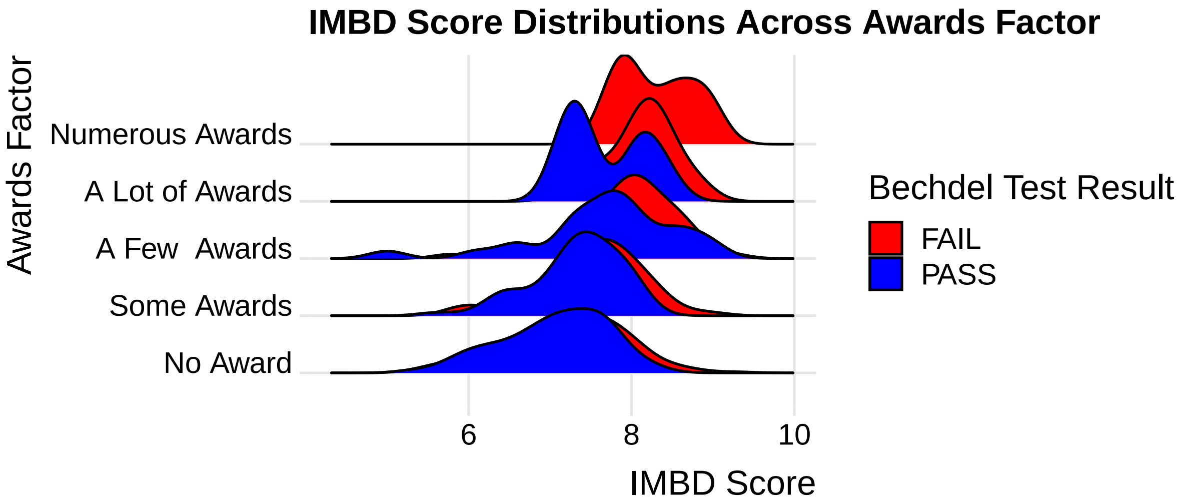

Figure 12: Ridge plot showing IMBD Score Distributions Across Awards Factor

The comparison between movies that fail and pass the Bechdel test in Figure 12 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 11: Awards Over Time by Bechdel Test Result

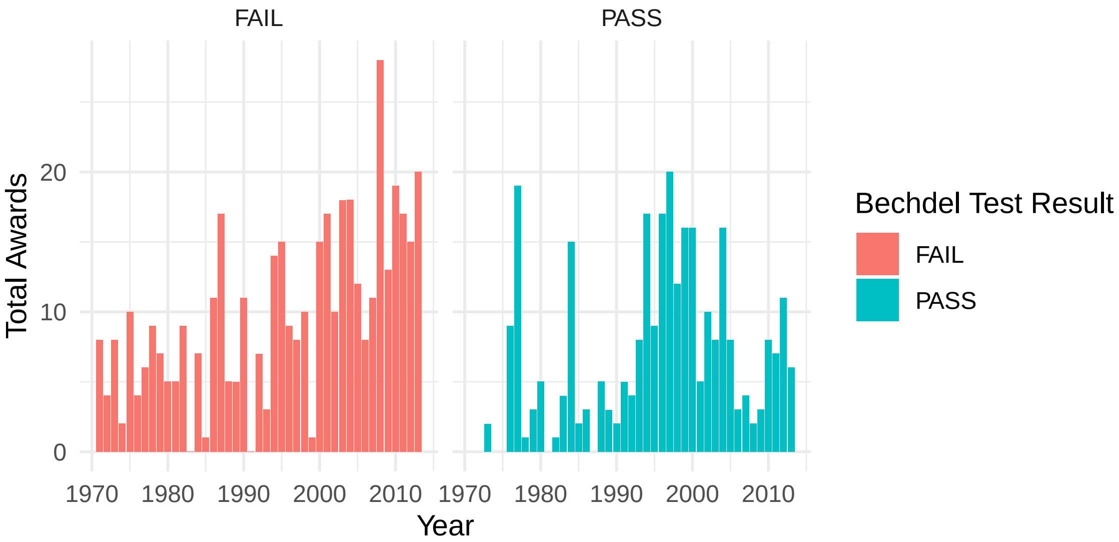

Figure 13: Stacked bar chart showing the total number of awards by year and Bechdel Test Result

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 12: IMBD Score Distributions by Awards Factor

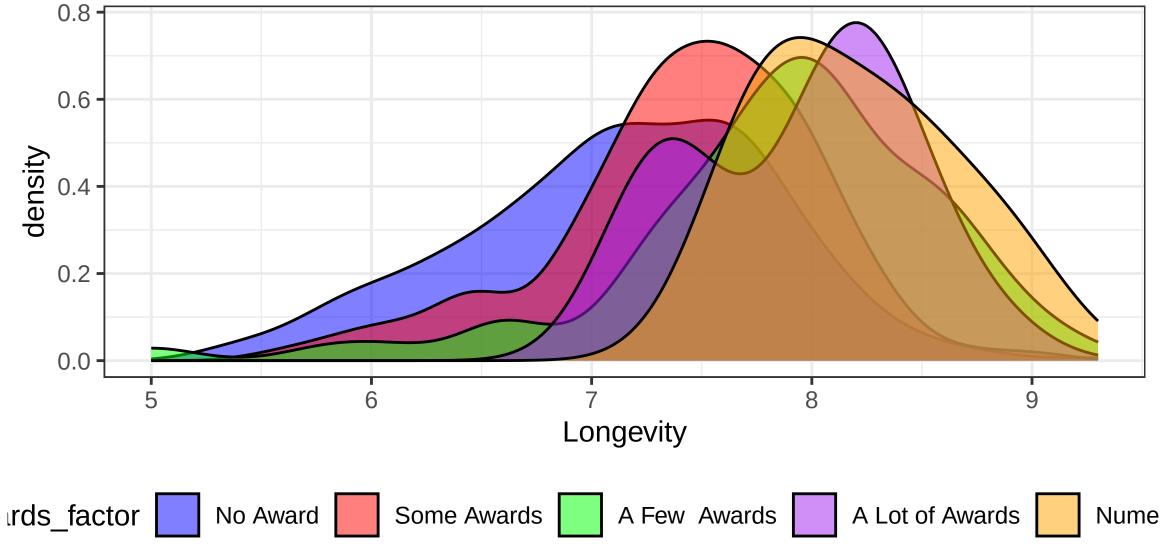

Figure 14: Density plot of IMBD scores across awards categories

The distributions in Figure 14 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 13: Shadowgram Visualization of IMBD Scores

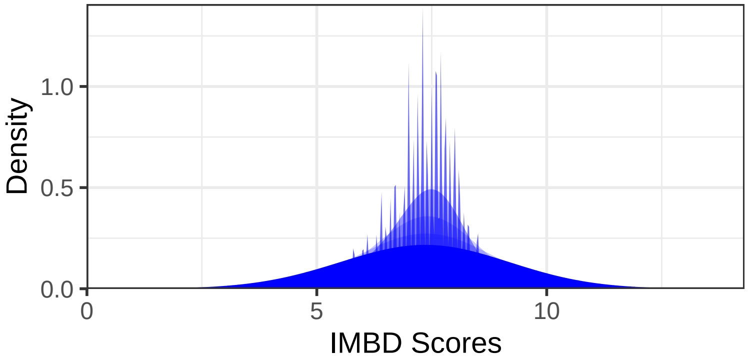

Figure 15: Shadowgram of IMBD Scores

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 14: Descriptive Statistics of IMBD Scores by Bechdel Test Result

Table 2: 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 2, 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

Model

Results

We present our results in three sections. First, we will discuss the assumptions of the parametric shortcut. Then we will move on to answering our omnibus questions before ending this section with post hoc analysis, if appropriate

Assessing 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.

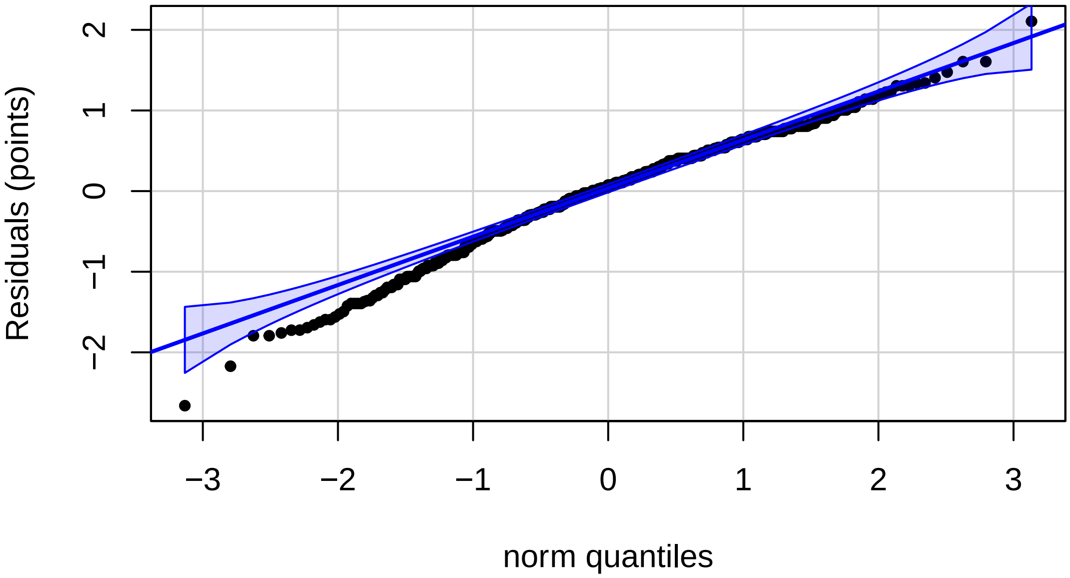

Figure 16: QQ Plot for Gaussian Residuals Assumption

Looking at this first assumption in Figure 16 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.

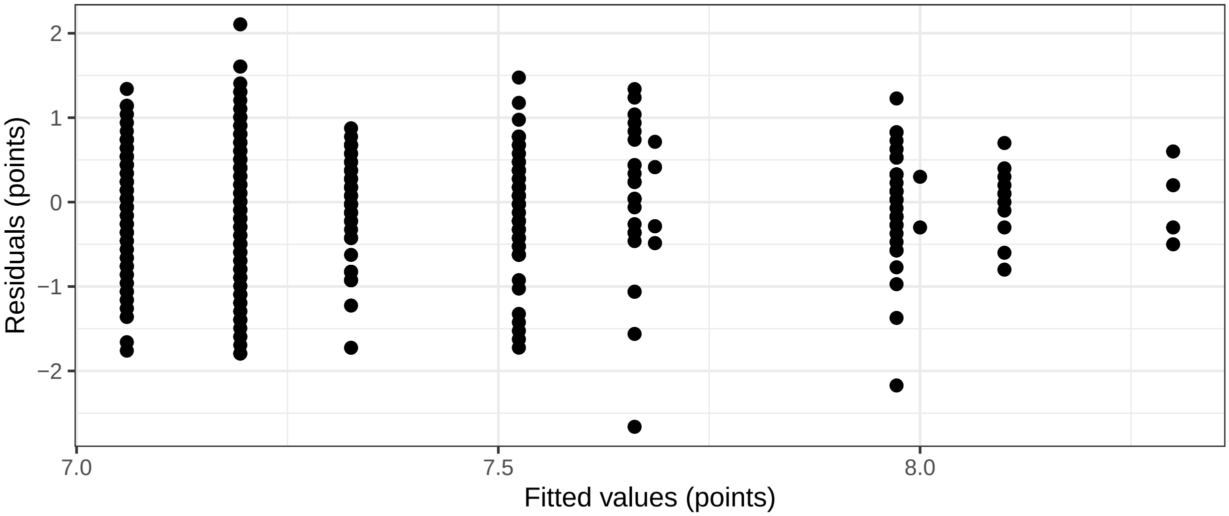

Figure 17: Tukey-Anscombe Plot for Equal Variance Assumption

The equal variance assumption in Figure 17 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 assess 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

Since the normality assumption was not met, we needed to look into other possible solutions. 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.

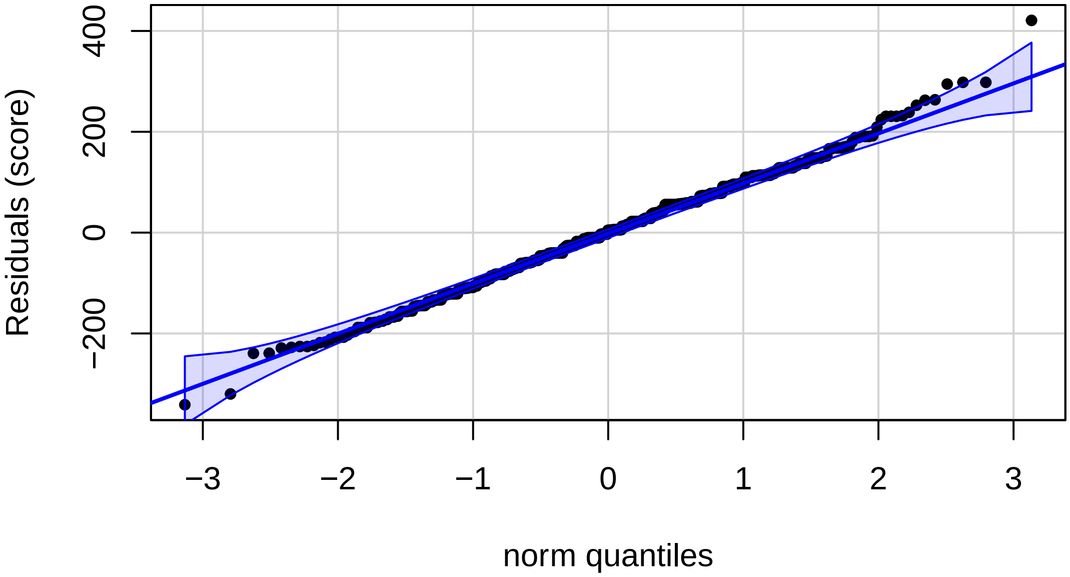

Figure 18: QQ Plot for Transformed Gaussian Residuals

After applying the transformation, it is hard to tell but it worked. We did not want to dedicate an entire page to the qq plot in Figure 18 to see, but after expanding the plot outside of the paper, we easily saw that only some of the points were outside of the envelope. This is a much bigger improvement from before, and we can now say that the normality assumption is met.

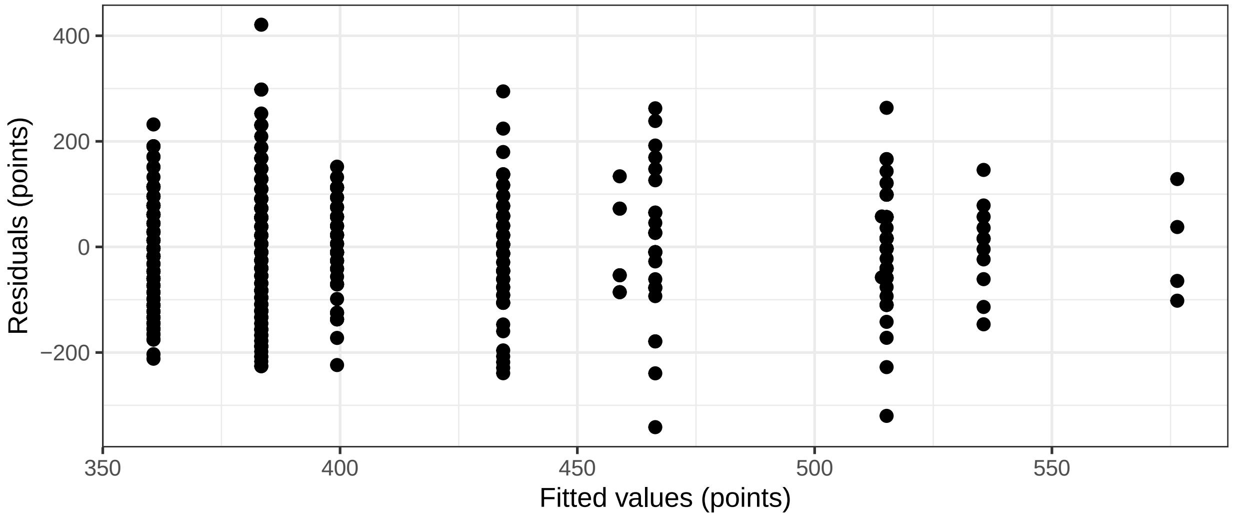

Figure 19: Tukey-Anscombe Plot for Transformed Data

Looking at the homoscedasticity of the residuals in Figure 19, 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 data to now look at the Omnibus report and begin to answer our research questions.

Index Plots

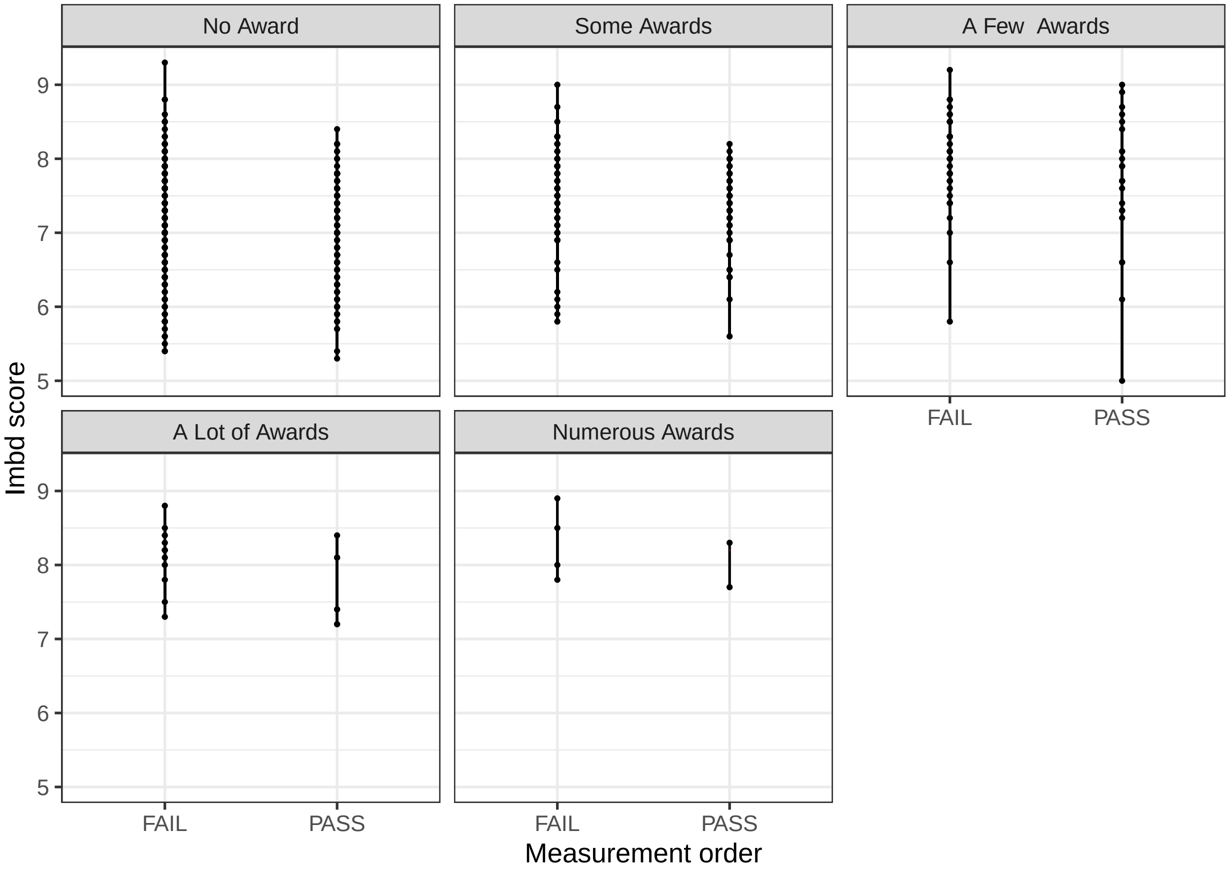

Figure 20: Index Plots Within Groups for Bechdel Test Results and Awards

The index plots in figure 20 provides a visual exploration of the relationship between the Bechdel test results, the number of awards received, and the corresponding IMBD scores. The plots are organized into separate panels for each awards category, allowing for a comparison across different levels of critical acclaim.

Based on the plot we can see variations in IMBD scores across different award categories. Differences in score distributions between movies that pass and fail the Bechdel test. More specifically The 'PASS' category across different awards levels does not show a significant variation in scores, implying a more consistent rating for movies that pass the Bechdel test. Conversely, the 'FAIL' category exhibits more variability, especially in the 'Some Awards' and 'A Few Awards' groups. Across the various groups, there are outliers present, indicating that there are movies with significantly higher or lower IMBD scores than the average within their respective awards categories.

Omnibus

Table 3: 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.

As we examine the ANOVA table in Figure 3, it becomes evident that the Bechdel test results, labeled here as movies.code, are associated with a notable amount of the variability in IMBD scores. Specifically, the Bechdel test results explain a greater degree of variation in IMBD scores than what we would expect by chance alone. This is indicated by the F-statistic, which suggests that such a high value would occur by chance less than 0.2% of the time ($p = 0.0002$). Since our p-value is less than our predetermined threshold for statistical significance, we reject the null hypothesis and conclude that the Bechdel test result does indeed impact IMBD scores. Furthermore, the awards_factor has an even more pronounced effect on IMBD scores, accounting for an F-statistic that would be observed approximately once in over 10,000 repetitions of this study under the null hypothesis ($p < 0.0001$). This strongly suggests that the number of awards a movie receives is significantly associated with its IMBD score.

However, the interaction between the Bechdel test results and the number of awards (movies.code:awards_factor) does not seem to contribute meaningfully to the variability in IMBD scores. This interaction effect is not statistically significant ($p = 0.7381$), suggesting that the Bechdel test's impact on IMBD scores does not differ across various levels of award recognition.

Looking at the effect sizes, η^2, ω^2, and ϵ^2, for movies.code are relatively small, indicating a modest effect. In contrast, the effect sizes for awards_factor are quite substantial. According to Cohen's benchmarks, effect sizes are considered large when ^{2} and ^{2} are above 0.14. With our ^{2} and ^{2} values of 0.1673 and 0.1600, respectively, for awards_factor, we are in the realm of a large effect size, underscoring the significant impact awards recognition has on IMBD scores.

Post Hoc Analysis

Given the significant results observed in our ANOVA, we will proceed with post hoc analysis to further dissect the impact of the Bechdel test results (movies.code) and the number of awards received (awards_factor) on IMBD scores. The significant F-statistics for these factors suggest that there are differences to explore between the levels of awards and potentially between movies that pass or fail the Bechdel test.

For the post-hoc analysis, we will be looking at the comparisons between the individual main effects, and answer our two questions from the beginning:

- Is failing the Bechdel test associated with a higher IMBD score?
- Is winning a lot of awards associated with a higher IMBD score?

Individual Main effect Comparisons

Table 4: Post Hoc Comparisons for Main Effect Bechdel Score

Pair	Difference	SE	DF	t	p-value	Cohen's d	Prob. of Superiority
FAIL - PASS	49.079	21.73	568	2.259	0.024	0.471	0.63

To begin, we want to first look at the p-value given in Table 4 and compare it directly with the unusualness threshold we decided from before of 0.1. The p-value came out to 0.024, which is below the threshold level. Since it is below, this means that we must reject the null hypothesis that there is no difference in the Bechdel test scores and act as if there is a difference in IMBD scores when looking at the Bechdel test score. As mentioned before, we can not make any causal statements about failing or passing the Bechdel test leading to better or worse scores, but we can say that there is a significant statistical difference between IMBD scores depending on if you fail or pass the Bechdel test. We can also take a look at Cohen's d and claim that there are 0.471 standard deviations between these two groups. This is not a lot of distance between the two, but with a p-value being so low, there must be a small variance for both tests. The probability of superiority explains what percent of the time a randomly selected number of the first group will have a higher numeric value of the response than a randomly selescted member of the second. In our case, this value is 0.63, meaning that 63% of the time a random value is chosen from movies that fail the Bechdel test, their IMBD score will be higher than a randomly selected value from movies that pass. This backs up our question because it shows that a large percent of the time a movie fails (shows less representation of women), they have a higher IMBD score.

Table 5: Post Hoc Comparisons for Main Effect Awards

Pair	Difference	SE	DF	t	p-value	Cohen's d	Prob. of Superiority
No Award - Some Awards	-44.828	10.517	568	-4.262	0.000	-0.430	0.381
No Award - A Few Awards	-118.753	14.154	568	-8.390	0.000	-1.139	0.210
No Award - A Lot of Awards	-125.230	25.842	568	-4.846	0.000	-1.201	0.198
No Award - Numerous Awards	-173.245	45.501	568	-3.807	0.000	-1.662	0.120
Some Awards - A Few Awards	-73.925	15.673	568	-4.717	0.000	-0.709	0.308
Some Awards - A Lot of Awards	-80.403	26.704	568	-3.011	0.005	-0.771	0.293
Some Awards - Numerous Awards	-128.417	45.997	568	-2.792	0.008	-1.232	0.192
A Few Awards - A Lot of Awards	-6.478	28.334	568	-0.229	0.819	-0.062	0.482
A Few Awards - Numerous Awards	-54.492	46.962	568	-1.160	0.308	-0.523	0.356
A Lot of Awards - Numerous Awards	-48.014	51.700	568	-0.929	0.393	-0.461	0.372

A similar breakdown can be used to look at the number of awards received in Table 5. While looking at every p-value for movies that received no awards, they are all 0. This would indicate that we reject the null hypothesis that there is no statistical difference between getting no awards and getting any of the other types, and act as if there is a difference between them. We can tell that movies that get no awards tend to have less IMBD scores because each of the cohen's d values are less than 0 (meaning no awards are that many standard deviation below its respective counterpart) and the probability of superiority is less than 50%, meaning for every random movie that has no awards is far less likely to have a higher IMBD score than those that have any awards at all.

There are only 3 cases of failing to reject the null hypotheses and having no statistical difference between the two, and these are for some and many awards, some and a lot of awards, and many and a lot of awards. This could be due to the lack of data points in each of these categories, but it is quite interesting to see that movies who get over 3 awards tend to have similar IMBD scores.

Discussion

There does appear to be a correlation of passing the Bechdel test and getting a higher score on IMBD. There also appears to be a correlation between getting no awards and having a lower IMBD score. As a reinstatement, since we are working with an observational study, we cannot make any claims about the Bechdel test or the number of awards causing IMBD scores to be what they are. We can however state that there was a statistical difference between the Bechdel test scores and also the number of awards received. It was interesting to see that the Omnibus results stated that the interaction term had no statistical evidence of the IMBD scores being different, yet there was a difference in the extremes of the factors (passing with no awards and failing with a lot of awards).

Limitations and Future Work

In this section, we will discuss specific limitations to our project and how this could be incorporated into future work.

To start, our data set only consisted of movies between the years 1971 and 2013 due to how the combined data sets just happened to line up. There are probably movies not on this list that would be able to get included, especially from recent times. It would be an interesting question to see how over time, the difference in Bechdel test scores could matter more or less. This type of future work could be making a data set that is able to have all movies from 1900 all the way to present day and see the impacts that different years make, and how additional years impact over time.

A second limitation consists of including two factors or potential blocks. First, we did not take into consideration any type of genre throughout. It is possible that the type of movie also impacts the IMBD score, and this would be interesting to see what genres tend to do better. The second is the year it came out. There is potential here for this to be used as a factor, but also as a block. Obviously, movies from the 1960s will probably have less representation of women than those in the 2010s, so using this as a block could have potential to block out the impacts that the year might have on the Bechdel test score.

The final limitation has to do with our response. We used IMBD score because we thought it would be able to represent the populations interests the best. However, there are many other ways to score a movie. Rotten tomatoes is probably the biggest out there, but this is dependent mainly on critics ratings, and the percentage you see is typically what percentage of critics who rated the movie rated it above 3.5. This seems biased mainly towards critics and not a good representation of the score, but nonetheless it would be interesting to use this as a response, or even as a factor. Another potential score could be Letterboxd scores, as they represent a much larger population of people.

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

The authors of this report would like to acknowledge their individual contributions to the report. Both authors contributed to ongoing discussions about study design and analysis.

- Olachi Mbakwe contributed to the Exploration of the Data, analysis of data(Omnibus Results), coding, and writing of the report.
- William Bevidas contributed to the Introduction and Background, Study Design and Methods, Exploration of the Data, analysis of data(Assumptions), coding, and writing of the report.
- Claudia Silverstein contributed to the analysis of data(Omnibus Results) and writing of the report.

Code Appendix

```
knitr::opts_chunk$set(
    echo = FALSE,
    fig.align = "center",
    message = FALSE,
    warning = FALSE,
    dpi = 300,
    cache = TRUE
)


# Add additional packages by name to the following list ----
packages <- c(
  "tidyverse", "knitr", "kableExtra", "hasseDiagram",
  "psych", "car", "parameters",'emmeans'
  )
lapply(X = packages, FUN = library, character.only = TRUE, quietly = TRUE)

# Loading Helper Files and Setting Global Options ----
options(knitr.kable.NA = "")
options("contrasts" = c("contr.sum", "contr.poly"))
source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/ANOVATools.R")

source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/shadowgram.R")


bechdel_awards <- read_csv("bechdel_awards.csv")

bechdel_awards$movies.code <- factor(
x = bechdel_awards$movies.code,
levels = c("FAIL", "PASS")
)

modelLabels <- c("1 Rating Movies 1", "2 Bechdel Score 1", "5 Awards 4", "10 Bechdel Score × Awards 4",
modelMatrix <- matrix(
  data = c(FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALS
  nrow = 5,
  ncol = 5,
  byrow = FALSE
)
hasseDiagram::hasse(
 data = modelMatrix,
 labels = modelLabels
)
# Create a histogram of the awards distribution
ggplot(bechdel_awards, aes(x = awards)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(title = "Histogram of Awards Distribution",
       x = "Number of Awards",
       y = "Frequency") +
  theme_minimal()
bechdel_awards$awards_factor <- cut(
  x = bechdel_awards$awards,
  breaks = c(-Inf, 0, 2, 6, 10, Inf),
  labels = c("No Award", "Some Awards", "A Few  Awards","A Lot of Awards","Numerous Awards"),
  include.lowest = TRUE
)

# Distribution of IMBD scores
hist(bechdel_awards$imbd_score, main = "Distribution of IMBD Scores", xlab = "IMBD Score")
# Descriptive statistics on longevity by condition ----
BechdelStats <- psych::describeBy(
x = bechdel_awards$imbd_score,
group = bechdel_awards$awards_factor,
na.rm = TRUE,
skew = TRUE,
ranges = TRUE,
quant = c(0.25, 0.75),
IQR = FALSE,
mat = TRUE
)

BechdelStats %>%
tibble::remove_rownames() %>%
tibble::column_to_rownames(
var = "group1"
) %>%
dplyr::select(
n, min, Q0.25, median, Q0.75, max, mad, mean, sd, skew, kurtosis
) %>%
knitr::kable(
caption = "Summary Statistics for Imdb Scores",
digits = 3,
format.args = list(big.mark = ","),
align = rep('c', 11),
col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "SAM", "SASD",
"Sample Skew", "Sample Ex. Kurtosis"),
booktabs = TRUE
) %>%
kableExtra::kable_classic_2(
font_size = 12,
latex_options = c("scale_down", "HOLD_position")
)%>%
  add_header_above() %>%
  footnote(
    general = "Descriptive statistics calculated for Imdb score rating",
    footnote_as_chunk = TRUE
  )

ggplot(
  data = bechdel_awards,
  mapping = aes(x = movies.code, y = imbd_score, fill = movies.code)
) +
  geom_boxplot() +
  theme_bw() +
```

```r
   labs(title = "IMBD Scores by Bechdel Test Result")+
  xlab("Bechdel Test Result") +
  ylab("IMBD Score") +
  theme(
    legend.position = "none",
    text = element_text(size = 12)
  )
ggplot(
  data = bechdel_awards,
  mapping = aes(x = awards_factor, y = imbd_score, fill = awards_factor)
) +
  geom_boxplot() +
  theme_bw() +
   labs(title = "IMBD Scores by Awards Factor")+
  xlab("Awards Factor") +
  ylab("IMBD Score") +
  theme(
    legend.position = "none",
    text = element_text(size = 12)
  )
# Boxplot of IMBD scores by Bechdel test result with awards factor
ggplot(bechdel_awards, aes(x = movies.code, y = imbd_score, fill = awards_factor)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "IMBD Scores by Bechdel Test Result and Awards Factor",
       x = "Bechdel Test Result",
       y = "IMBD Score",
       fill = "Awards Factor") +
  scale_fill_brewer(palette = "Set3")
boxplot(
formula = imbd_score ~ movies.code:awards_factor,
data = bechdel_awards,
ylab = "Imdb Score",
xlab = "Bechdel result x Award"
)

# Interaction plot for Bechdel test result and awards factor
interaction.plot(bechdel_awards$awards_factor, bechdel_awards$movies.code, bechdel_awards$imbd_score,
                 main = "Interaction Plot", xlab = "Awards Factor", ylab = "Average IMBD Score",
                 trace.label = "Bechdel Test", fixed = TRUE)
interaction.plot(
x.factor = bechdel_awards$movies.code, # First Factor
trace.factor = bechdel_awards$awards_factor, # Second Factor
response = bechdel_awards$imbd_score, # Response
fun = mean,
type = "b", # Both points and lines
col = c("black","red","blue","gold","green","purple"), # Set colors for trace
pch = c(19, 17, 15), # Set symbols for trace
fixed = TRUE,
legend = TRUE,
xlab = "Bechdel Results",
ylab = "Imbd rating(score)",
trace.label = "Award Factor")
```

```r
bechdel_awards %>%
  group_by(movies.code, awards_factor) %>%
  summarise(average_score = mean(imbd_score, na.rm = TRUE), .groups = 'drop') %>%
  ggplot(aes(x = movies.code, y = awards_factor, fill = average_score)) +
  geom_tile() +
  scale_fill_viridis_c() +
  labs(x = "Bechdel Test Result", y = "Awards Factor", fill = "Average IMBD Score") +
  theme_minimal()


ggplot(bechdel_awards, aes(x = awards_factor, fill = movies.code)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = c("red", "blue")) +
  labs(title = "Number of Movies by Bechdel Test Result and Awards Factor", x = "Awards Factor", y = "C
  theme_minimal()

library(ggridges)
ggplot(bechdel_awards, aes(x = imbd_score, y = awards_factor, fill = movies.code)) +
  geom_density_ridges() +
  scale_fill_manual(values = c("red", "blue")) +
  labs(title = "IMBD Score Distributions Across Awards Factor", x = "IMBD Score", y = "Awards Factor", 
  theme_ridges()


ggplot(bechdel_awards, aes(x = movies.year, y = awards, fill = movies.code)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~movies.code) +
  theme_minimal() +
  labs(x = "Year", y = "Total Awards", fill = "Bechdel Test Result")

bechdel_awards %>%
  drop_na(awards_factor) %>%
  ggplot(mapping = aes(x = imbd_score, fill = awards_factor)) +
  geom_density(na.rm = TRUE, alpha = 0.5) +
  theme_bw() +
  scale_fill_manual(values = c("blue", "red", "green", "purple", "orange")) + # Make sure to adjust the
  xlab("Longevity") +
  theme(legend.position = "bottom")
# Assuming the 'shadowgram' function is available in your R environment
# and 'bechdel_awards' is your dataset with an 'imbd_score' column
shadowgram(
  dataVec = bechdel_awards$imbd_score,
  label = "IMBD Scores",
  layers = 5,
  color = "blue",
  aStep = 4
)

# Descriptive statistics on IMBD scores by Bechdel test result
bechdel_stats <- psych::describeBy(
  x = bechdel_awards$imbd_score,
  group = bechdel_awards$movies.code,
  na.rm = TRUE,
  skew = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = FALSE,
  mat = TRUE
)

bechdel_stats %>%
  tibble::remove_rownames() %>%
  tibble::column_to_rownames(var = "group1") %>%  # Use the actual grouping column name here
  dplyr::select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd, skew, kurtosis
  ) %>%
  knitr::kable(
    caption = "Summary Statistics for IMBD Scores by Bechdel Test Result",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "SAM", "SASD", "Sample Skew", "Sample
    booktabs = TRUE
  ) %>%
  kableExtra::kable_classic_2(
    font_size = 12,
    latex_options = c("scale_down", "HOLD_position")
  )


# Fit the model for ANOVA ----
imbd_model <- aov(
  formula = imbd_score ~ movies.code * awards_factor,
  data = bechdel_awards,
  na.action = "na.omit"
)
# Assumption Assessment Visualizations ----
## Gaussian Residuals Assumption
car::qqPlot(
  x = imbd_model$residuals,
  distribution = "norm",
  envelope = 0.9,
  id = FALSE,
  pch = 20,
  ylab = "Residuals (points)"
)
## Strip Chart for Homoscedasticity ----
ggplot(
  data = data.frame(
    residuals = imbd_model$residuals,
    fitted = imbd_model$fitted.values
  ),
  mapping = aes(x = fitted, y = residuals)
) +
  geom_point(size = 2) +
  theme_bw() +
```

```r
  xlab("Fitted values (points)") +
  ylab("Residuals (points)")
# Fit the model for ANOVA ----
bechdel_awards$imbd_score3 <- (bechdel_awards$imbd_score)^3
imbd_model <- aov(
  formula = imbd_score3 ~ movies.code * awards_factor,
  data = bechdel_awards,
  na.action = "na.omit"
)
# Assumption Assessment Visualizations ----
## Gaussian Residuals Assumption
car::qqPlot(
  x = imbd_model$residuals,
  distribution = "norm",
  envelope = 0.9,
  id = FALSE,
  pch = 20,
  ylab = "Residuals (score)"
)
## Strip Chart for Homoscedasticity ----
ggplot(
  data = data.frame(
    residuals = imbd_model$residuals,
    fitted = imbd_model$fitted.values
  ),
  mapping = aes(x = fitted, y = residuals)
) +
  geom_point(size = 2) +
  theme_bw() +
  xlab("Fitted values (points)") +
  ylab("Residuals (points)")
# Create new data frame ---
means <- bechdel_awards %>%
group_by(awards_factor) %>%
summarize(
sam = mean(imbd_score, na.rm = TRUE)
)

bechdel_awardsData <- left_join(
x = bechdel_awards,
y = means,
by = c("awards_factor" = "awards_factor")
)

 # Create the Index Plot by Group/Team ----
ggplot(
data = bechdel_awardsData,
mapping = aes(
x = movies.code,
y = imbd_score
)
) +
geom_point(size = 0.5) +
geom_line() +
geom_smooth(
formula = y ~ x,
method = "lm",
se = TRUE,
level = 0.9,
linetype = 1
) +
geom_line(
data = bechdel_awardsData,
mapping = aes(y = sam, x = movies.code),
linetype = 2,
color = "red"
) +
theme_bw() +
xlab("Measurement order") +
ylab("Imbd score ") +
facet_wrap(
facets = vars(awards_factor),
scales = "fixed"
)
# Modern ANOVA Table ----

parameters::model_parameters(
model = imbd_model,
effectsize_type = c("eta", "omega", "epsilon")
) %>%
  dplyr::mutate(
p = ifelse(
test = is.na(p),
yes = NA,
no = pvalRound(p, digits = 4)
)
) %>%
knitr::kable(
digits = 4,
col.names = c(
"Source", "SS", "df", "MS", "F", "p-value",
"Eta Sq.", "Omega Sq.", "Epsilon Sq."),
caption = "Modern ANOVA Table for Study",
booktabs = TRUE,
align = c("l", rep("c", 8))
) %>%
kableExtra::kable_classic(
font_size = 10,
latex_options = c("HOLD_position")) %>%
  kableExtra::footnote(
general = "Computer rounding has made the p-value look like zero.",
general_title = "Note. ",
footnote_as_chunk = TRUE
)

IMBDPostHoc <- emmeans::emmeans(
  object = imbd_model,
  specs = pairwise ~ movies.code, # Notice the use of the |
  adjust = "BH",
  level = 0.9
)

IMBDEffects <- as.data.frame(
    eff_size(
    object = IMBDPostHoc,
    sigma = sigma(imbd_model),
    edf = df.residual(imbd_model)
    )
) %>%
dplyr::mutate( # The eff_size command places the pairs inside parentheses
contrast = gsub(pattern = "[()]", replacement = "", x = contrast),
ps = probSup(effect.size),
.after = effect.size
) %>%
dplyr::select(contrast, effect.size, ps)

### Build table
as.data.frame(IMBDPostHoc$contrasts) %>%
left_join(
  y = IMBDEffects,
  by = join_by(contrast == contrast)
) %>%
knitr::kable(
digits = 3,
caption = "Post Hoc Comparisons for Main Effect Bechdel Score",
col.names = c("Pair", "Difference", "SE", "DF", "t", "p-value", "Cohen's d",
"Prob. of Superiority"),
align = "lcccccc",
booktabs = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("condensed", "boardered"),
font_size = 10,
latex_options = c("HOLD_position")
)
fig.align='center'
IMBDPostHoc <- emmeans::emmeans(
  object = imbd_model,
  specs = pairwise ~ awards_factor,
  adjust = "BH",
  level = 0.9
)

IMBDEffects <- as.data.frame(
    eff_size(
    object = IMBDPostHoc,
    sigma = sigma(imbd_model),
    edf = df.residual(imbd_model)
    )
```

```
) %>%
dplyr::mutate( # The eff_size command places the pairs inside parentheses
contrast = gsub(pattern = "[()]", replacement = "", x = contrast),
ps = probSup(effect.size),
.after = effect.size
) %>%
dplyr::select(contrast, effect.size, ps)

### Build table
as.data.frame(IMBDPostHoc$contrasts) %>%
left_join(
  y = IMBDEffects,
  by = join_by(contrast == contrast)
) %>%
knitr::kable(
digits = 3,
caption = "Post Hoc Comparisons for Main Effect Awards",
col.names = c("Pair", "Difference", "SE", "DF", "t", "p-value", "Cohen's d",
"Prob. of Superiority"),
align = "lccccccc",
booktabs = TRUE
) %>%
kableExtra::kable_styling(
bootstrap_options = c("condensed", "boardered"),
font_size = 10,
latex_options = c("HOLD_position")
)
```