Data Final Report - Jessie Reynolds - 12/13/24

Introduction / Problem Statement

For my final project, I chose to use the dataset provided on the Bechdel Test. The Bechdel Test, created in 1985 by Allison Bechdel, was a way to measure female representation, or the lack of, in film and fiction. Movies are scored based on the following criteria; includes at least two named women, the women have at least one conversation with one another, and the conversation is about something other than a man. Based on this criteria, movies can score from 0-3 on the Bechdel Test. The test was created to highlight the lack of female presence in film and call action to gender inequalities. Though the Bechdel Test is not an accurate representation of females in film, it is at least a step in the right direction. A film may include two named women who have a conversation with one another about something other than a man, but that does not mean the film is used for female empowerment or has accurate female representation and equality.

For my analysis, I was interested in focusing on various movie characteristics and how these related to scores on the test. Specifically, I wanted to look at factors including genre, year released, movie length, sex of the director, sex of the writer and sex of the top stars from each movie. In the film industry, there are genres or eras that have higher female representation than others. Along with sex of the stars and directors/writers, society has preconceived notions about how these factors play into the role of female representation. I was interested in which characteristics specifically led to movies passing the Bechdel Test with a score of 3.

For my project, I enriched the Bechdel Test data set provided by Dr. Bray with two different datasets from Kaggle. I used one data set including the Top 250 IMDb movies and their characteristics (Kumar) along with an additional dataset that had various movie characteristics for a wide variety of movies (Mhatre). I chose to use the Top 250 IMDb movies dataset because IMDb is the world's most popular and credible source for film content. This dataset provided information on the top ranked movies that are relevant in society today. Since this dataset did not provide movie genres or writers, I found a second dataset that could provide this information.

Methods

To begin my cleaning process, I started by joining the Bechdel Test dataset and the Movies Initial dataset with a left join. I joined these two dataframes by the film’s IMDb ID, which is just an identifier for movies from IMDb’s platform. For my analysis I would be comparing various variables to scores on the Bechdel Test, so I wanted to have all observations included from the original Bechdel Test dataset. From here, I then joined this combined dataset with the Top 250

Movies dataset with a right join. I joined these datasets by movie title, since their IMDb IDs were not provided in the top 250 movies dataset. My dataset now went from over 46,000 observations to 292 observations. Once all of my datasets were joined, I could see there were duplicates from the different datasets. To drop duplicates, I filtered observations by movie title, year and director because there are many movies that do share the same title, but by filtering on year and director I was removing duplicates of the same movies. I kept only distinct movies on title, year and director and this took my dataset down to 282 observations. After dropping duplicates I cleaned variable names because there were a lot of repeated variables in my combined dataset. Next, I removed all NA values for scores on the Bechdel test because this is crucial for my analysis. After dropping NA values for Bechdel score, my final dataframe had 215 observations.

Not all movies fall into only one genre category, so many of these observations had multiple genres listed out. For my analysis, I thought it would be easier to only look at the first genre listed as this was the most prominent one for each film. I separated the genre variable and kept only the first one provided. I repeated this same process for the director variable because there were many people listed who worked on each movie. Again, I thought the most prominent director was listed first. From here, I faced some challenges in cleaning. I had one director for each movie, but I was interested in analyzing sex of the director so I had to figure out what would be the easiest way to place these directors into male or female categories. To tackle this obstacle, I separated directors into first names and last names. I began looking through the first names and placing them into the male category and researching names I was unfamiliar with. After looking at all of the director's names, I found that all of the directors in my final dataframe were male, so I chose to stop with any cleaning and analysis for this variable.

Each observation had the top 3 stars provided, so I separated these into their own variables Star 1, Star 2 and Star 3. From here, I repeated a similar process when tackling the directors by separating these each into first names and last names. I went through the first names of each star and put these into male and female categories. I did face some obstacles at this point because there were some female and male stars that had the same first name, so I couldn't just assign these names to one category. I had to problem solve on how to assign these to the correct category with the help of chatGPT. I then wanted to create a new variable representing the proportion of top stars that were female. I converted "female" to 1 and "male" to 0 and totaled the number of females for each observation and divided them by 3. These proportions were either 0/3, 1/3, 2/3 or 3/3.

The final stage in my cleaning process, and arguably the most difficult, was tackling the writer variable. Each movie had multiple people listed in this variable and some provided the writer of the novel/book adaptation of the movie, some had the story writer and some had the screenplay writer provided. With the assistance of ChatGPT and Dr. Bray, I used code to pull out the screenplay writer, if it was listed, and if not pull out the first name listed from that variable. From here, I repeated the same process as before with separating this variable into first and last name to then place into male or female categories. Finally, I selected only the variables I was

interested in for analysis, including; title, year, length, genre, IMDb rating, Bechdel score, proportion of female stars and sex of the writer.

For my multivariate analysis, I chose ordinal regression because my response variable, Bechdel score, had multiple categories. For my analysis, I used genre, proportion of top female stars and sex of the writer as my predictor variables when running my model.

Results

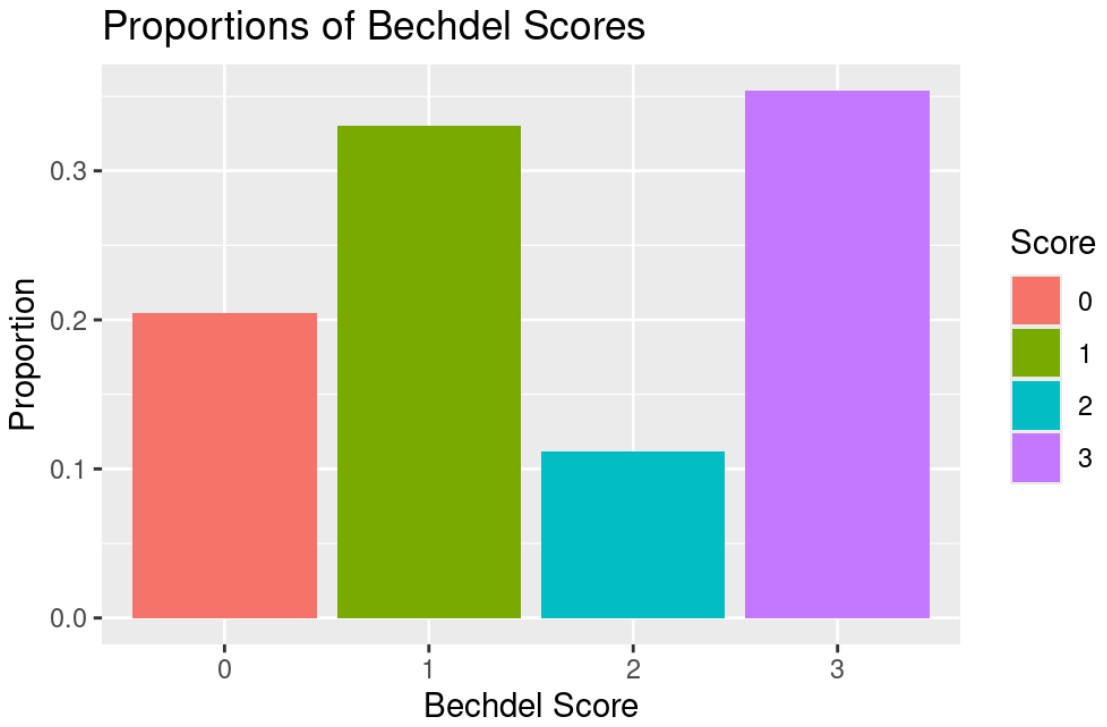

Figure 1: *Proportions of movies by various Bechdel Scores*

Before analyzing relationships between Bechdel Scores and movie characteristics, I wanted to see how the movies in my final dataset scored on the test. Out of my 215 films in my final dataframe, 44 scored 0, 71 scored 1, 24 scored 2 and 76 scored 3. This means around 35% of my final dataframe passed the Bechdel test, which did not come as a shock, but I was still hoping this would be higher. My data frame also had a similar proportion of movies with a score of 1 with around 33% of movies falling into this category. From this graph we can see that there was a very small proportion of movies that scored 2 on the test, this was around 11%.

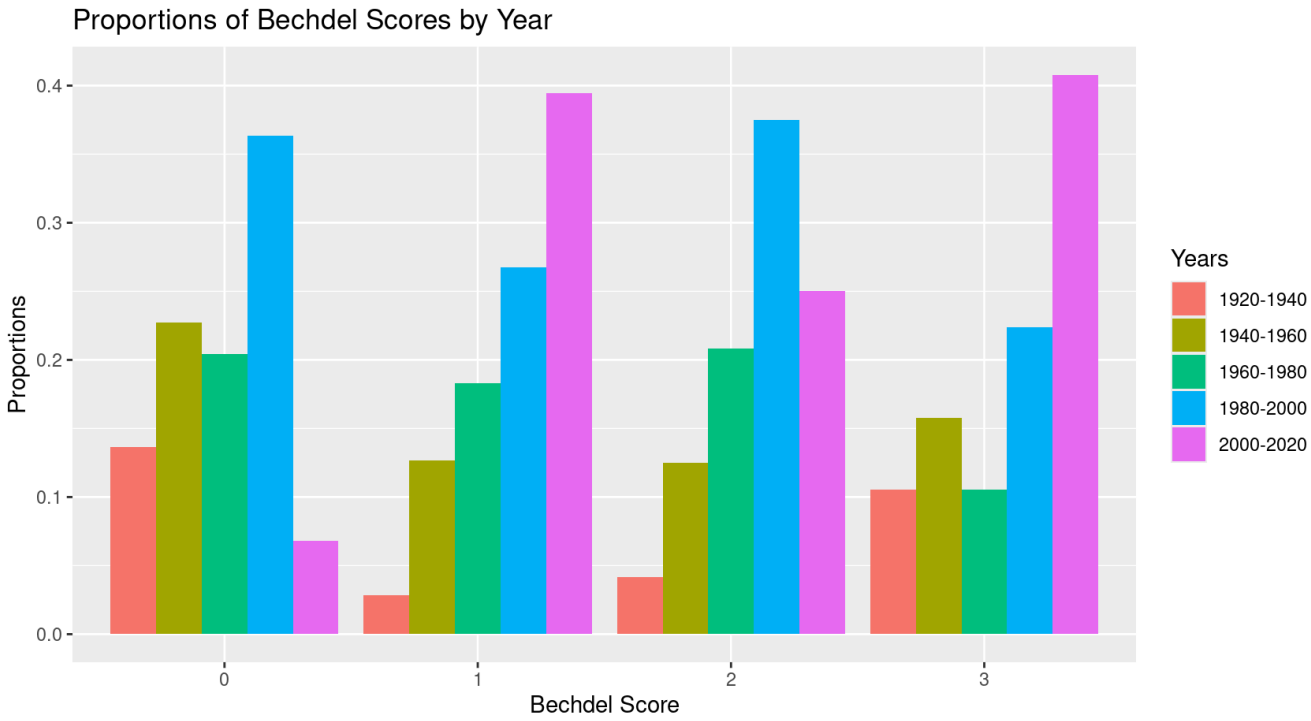

Figure 2: *Proportions of various Bechdel Scores by Year*

The films in my final dataset ranged from 1921-2017, so I decided to look at these in 20 year increments to make analysis easier. When analyzing Bechdel scores over the years, I predicted that the proportion of movies scoring 0 and 1 would decrease over time as the proportion of movies scoring 2 or 3 would go up over time. Based on ***Figure 2***, we can see that the largest proportion of movies with a score of 0 occurred in 1980-2000. This proportion was much higher than 1920-1980, which was quite shocking. For movies that scored 1 on the Bechdel test, we can see that the highest proportion of these came from 2000-2020, which again was quite shocking. We would expect as society progresses that modernity would lead to more female representation in film now than in the past. For movies that scored 2 and 3 we see an overall increase in proportion as time progresses, as we had predicted.

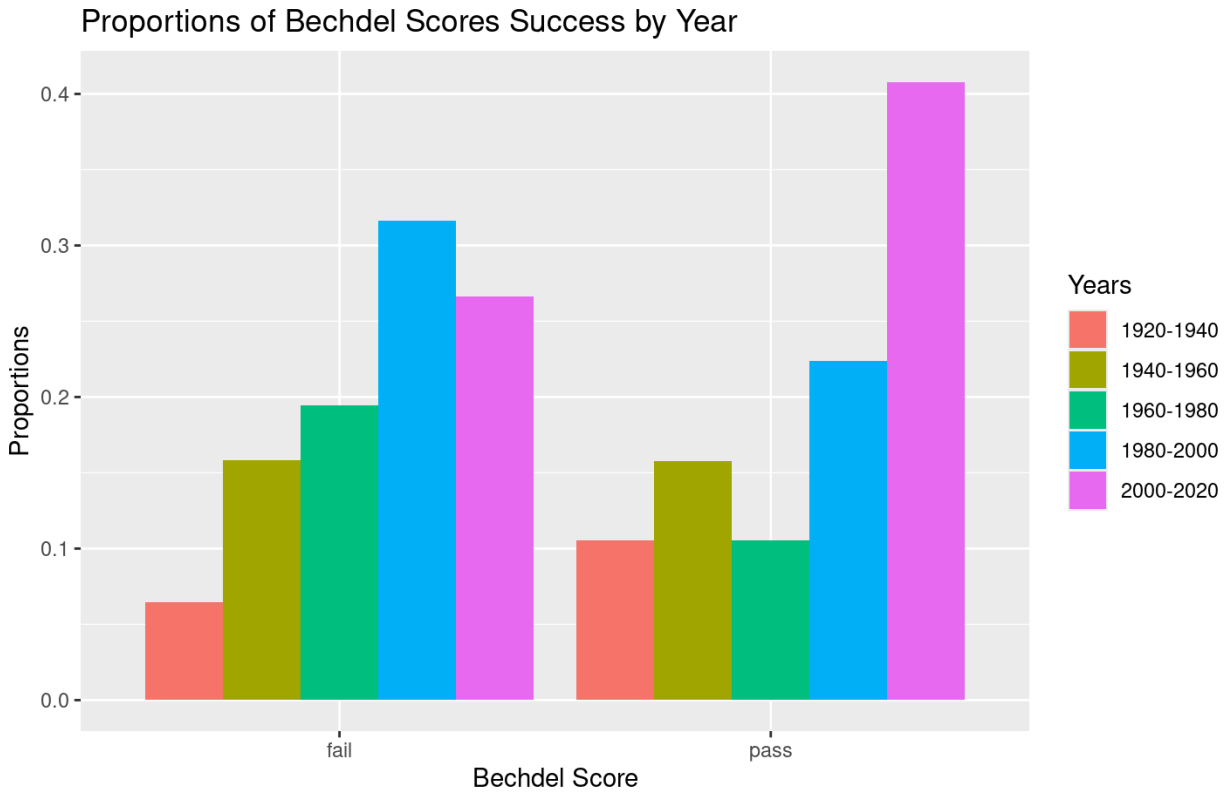

Figure 3: *Proportions of films that pass/fail the Bechdel Test based on year*

I then categorized scores into passing, films that scored 3, or failing, films that scored 0 - 2. From ***Figure 3***, we can see that the proportion of failing films was lowest in 1920-1940 and highest in 1980-2000 which was quite surprising, we would expect that the proportion of failing films to be lower in more recent years. We can also see that the proportion of passing films was lowest in 1960-1980 and highest in 2000-2020. This makes sense as we expect the highest proportion of passing films to come from more recent years.

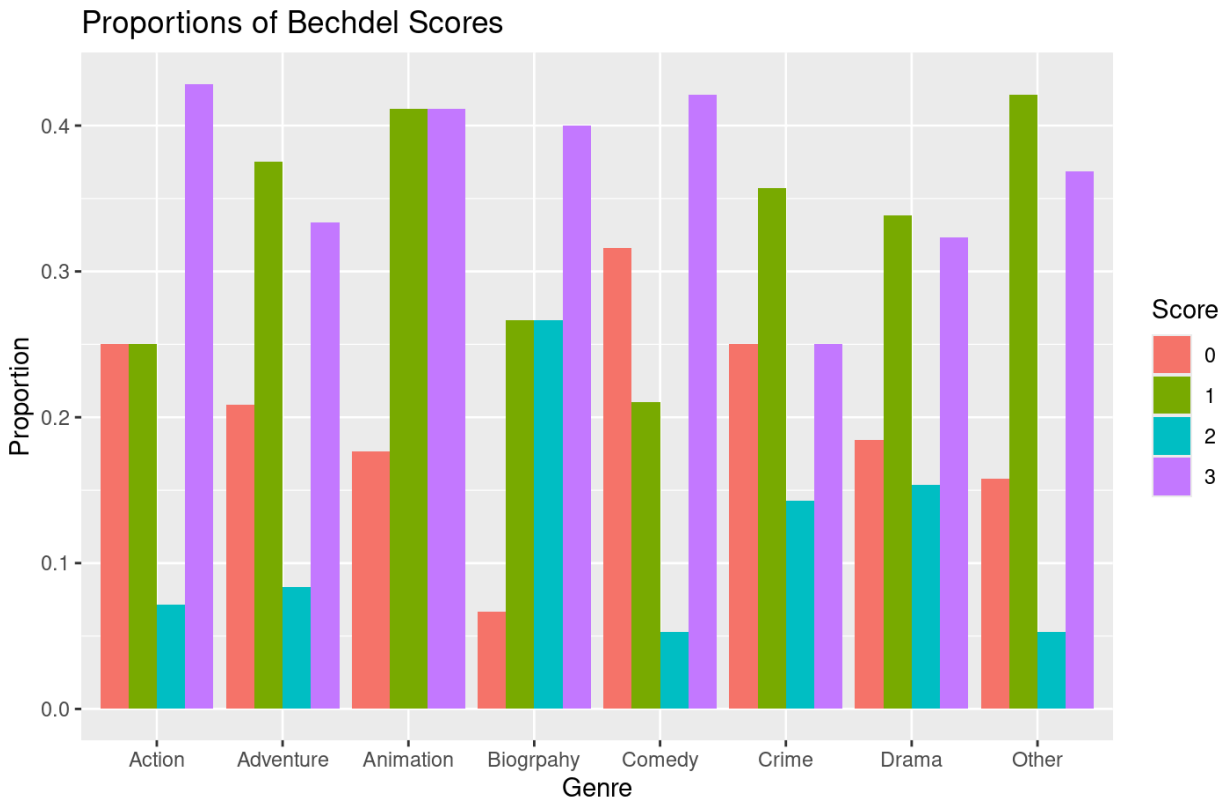

Figure 4: *Proportion of movies by genres over various Bechdel scores*

Figure 4 represents each genre of film and proportions for various Bechdel scores. From this figure we can see that in every genre, the score 3 had the highest proportion. Comedy, action, animation and biographies were the genres with the proportions of passing scores. This was surprising as we would expect drama movies to have the highest proportion scoring 3 because these films are often based on female characters or aimed toward female audiences.

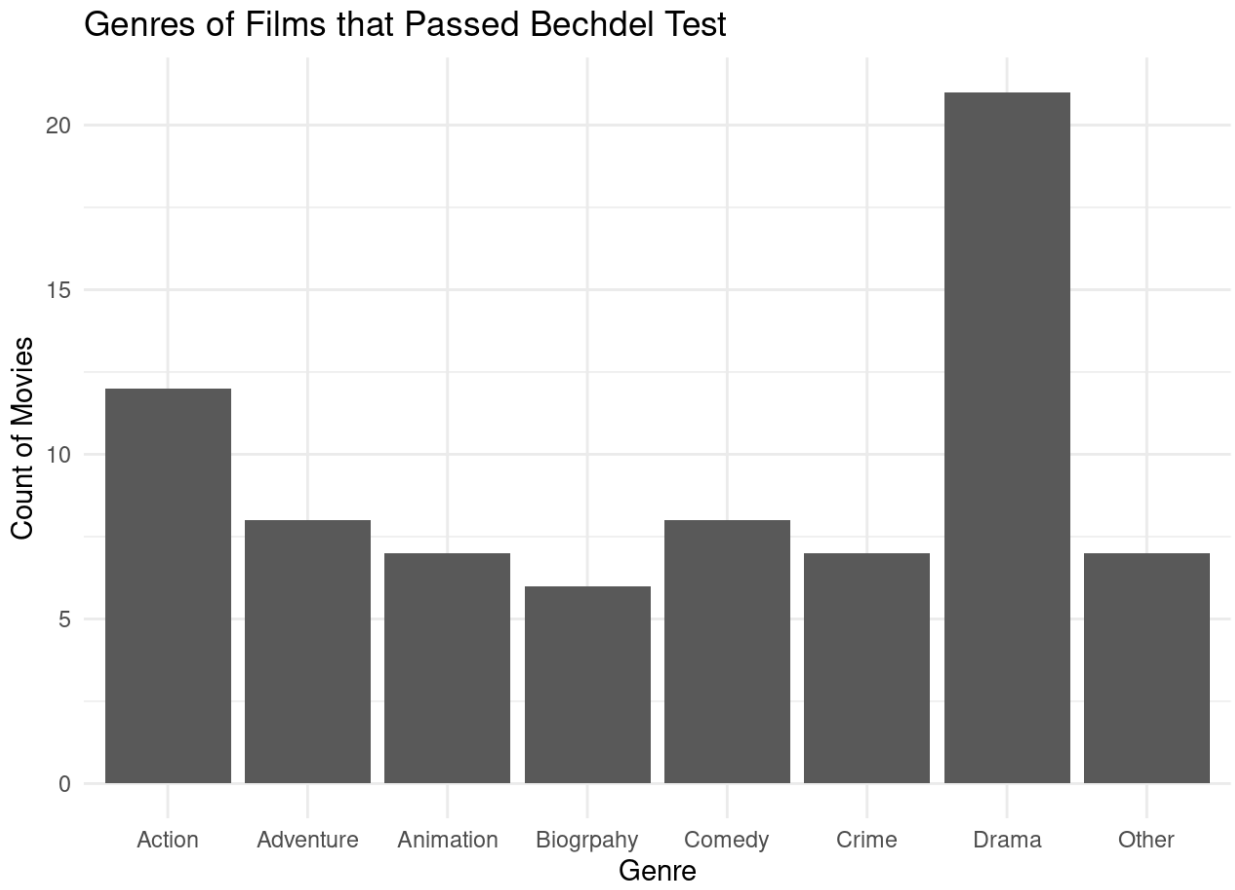

Figure 5: *Count of films by genre for films that pass the Bechdel Test*

To further my analysis of genres, I analyzed genres from only films that passed the Bechdel test. From Figure 5, we can see that the genre with the most passing films was drama. This was not surprising as female characters often appear in drama movies. We can also see that biographies and other, which includes western, crime, sci-fi, etc. all have a low count of movies that pass the test. Surprisingly, the genre with the second highest number of passing films was action. Action movies often include aspects like car chases, violence or explosions and are often focused on male characters.

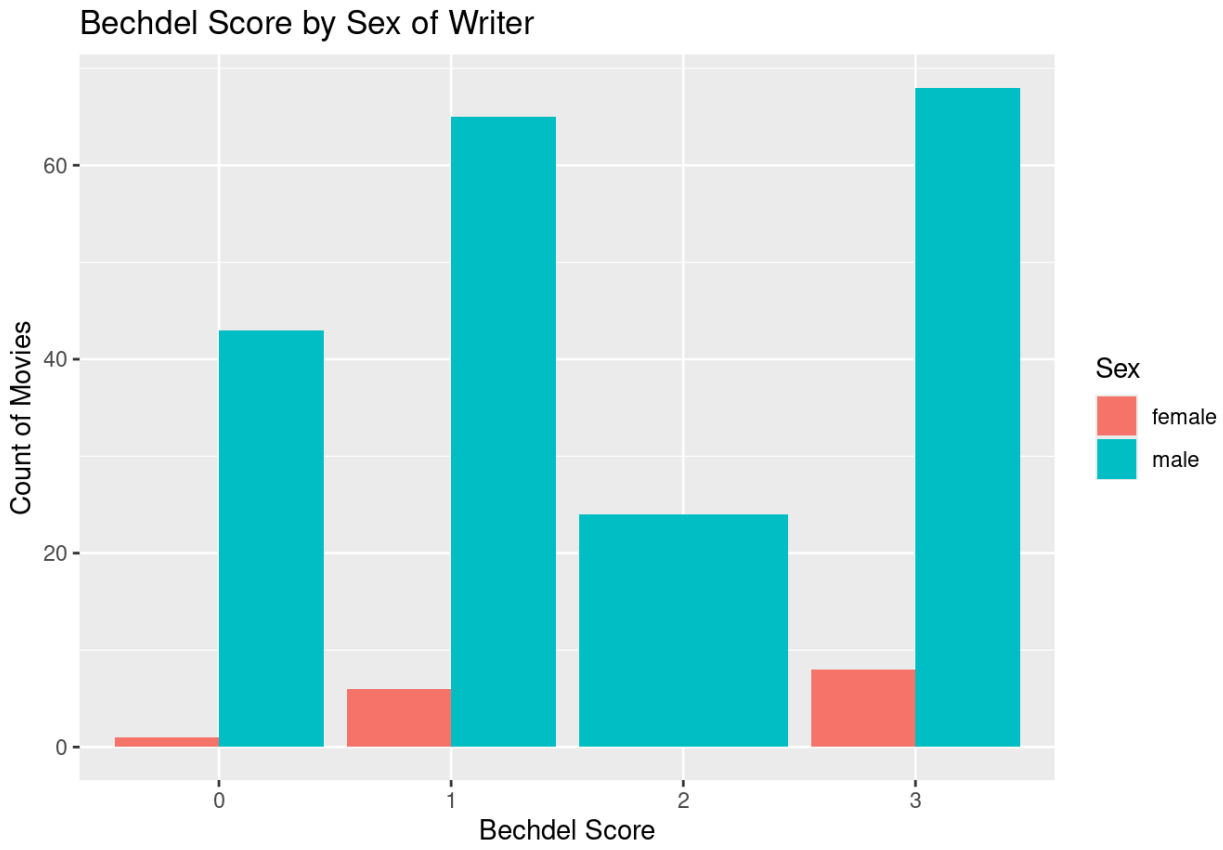

Figure 6: *Count of films with male or female writers over various Bechdel Scores.*

In my final dataframe, there were 200 films written by males and only 15 films written by females. It is difficult to draw strong conclusions because there are very few films written by a female, but based on Figure 6, we can see that most of the movies written by a female did pass the Bechdel Test with a score of 3. This makes sense as we would predict movies written by females would be more likely to pass the test. There were still a large number of movies with a passing score written by males.

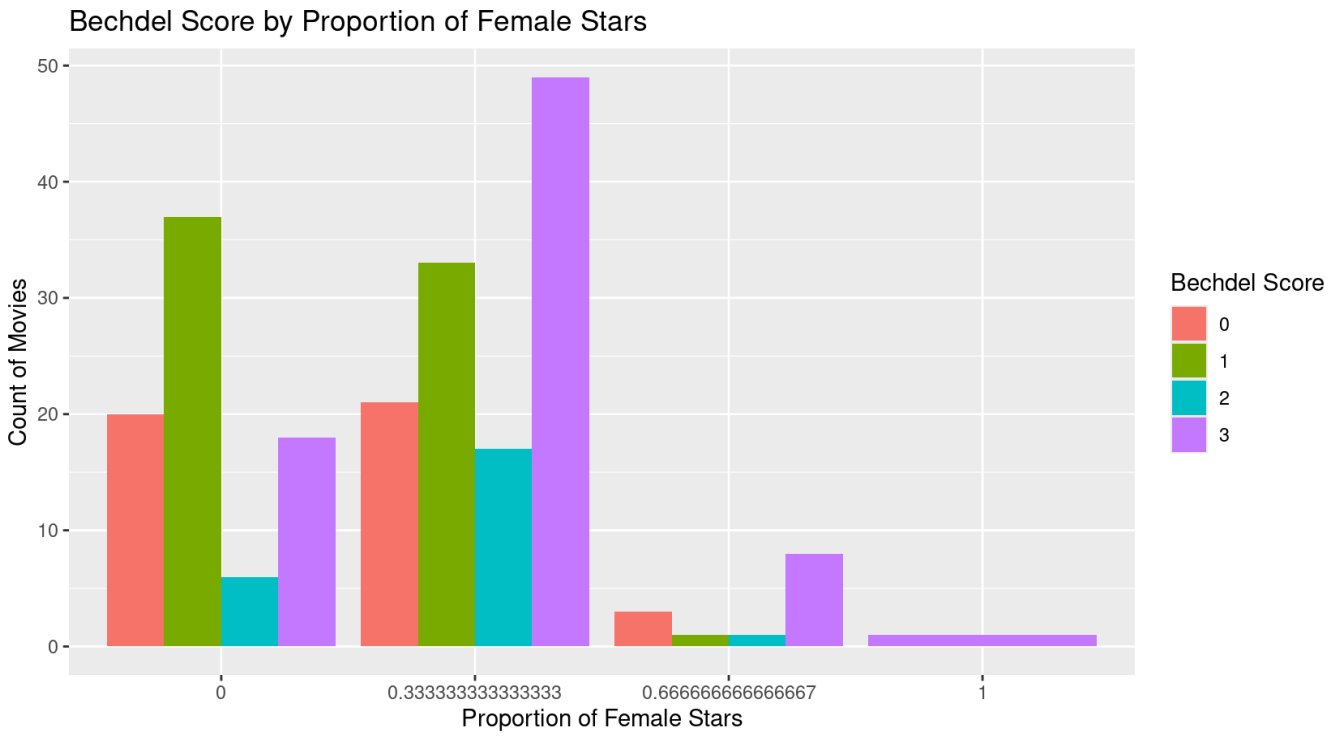

Figure 7: *Count of films based on proportion of female stars over various Bechdel Scores. All films were based on what proportion of their top 3 stars were female.*

Most of the films in my dataset had either 0 or 1 top female stars and there were very few with 2 or 3 top female stars. From ***Figure 7***, we can see that there was a very small count of movies with all 3 of the top stars as female, but these films did pass the Bechdel test. Surprisingly, there was a very high number of movies with 1 female star that passed the test. This means that there was another female character in the movie to meet any of the criteria, but they were not one of the top stars. Most movies with no top female stars scored a 1 on the test, which again means there were women in this film but they were not the top stars. From these results we can see that a majority of the passing films had at least 1 of the top stars as female.

Multivariate Analysis

For my multivariate analysis, I used ordinal regression to see how variables would predict the various scores on the Bechdel Test. In my analysis I included predictor variables including genre, proportion of female stars and sex of the writer. For my ordinal regression model, I used the genre action, 0/3 top female stars and a female writer as my base levels for each variable. Based on my model, I found that all scores on the test would decrease for every other genre in my final dataset compared to action, other than biographies. This may be because all biography movies in my dataset with a score of 3 are most likely biographies about a woman so they would be more likely to meet the criteria than a biography about a man. I can also see that as the proportion of top female stars increases, so do scores on the test. Finally, films scored lower

on the test if they had a male writer compared to female writers. As female presence increases, whether that is number actresses or writers, scores also increase.

Discussion

From my final data set, we can see that the majority of films either scored 1 or 3 on the Bechdel test and there are very few that scored 2. Based on my analysis of genres, we can see that Drama movies had the highest number of movies that passed over other genres, with a surprising amount from action movies. Based on the year a film was released, we can see that the time frame with the highest proportion of failing films was 1980-2000 and the time frame with the highest proportion of passing films was 2000-2020. We did not see a clear correlation that as time progressed scores were increasing. Based on my analysis of sex of the writer and the fact that all of my films had male directors, it is difficult to draw conclusions about how sex is affected by Bechdel test scores. We were able to see, even though there were a small number of films in this category, that those movies with higher proportions of top female stars did pass the Bechdel test. Based on these results, we can conclude that there is still a lack of female representation and gender inequality needs to be addressed in the film industry.

References

Kumar, Yarana. *IMDb Top 250 Movies*. 2023, Kaggle, https://www.kaggle.com/datasets/yaranathakur/imdb-top-250-movies.

Mhatre, Samruddhi. *IMDB Movies Analysis.* 2021, Kaggle, https://www.kaggle.com/datasets/samruddhim/imdb-movies-analysis