REPORT: Project 2 - Bechdel Movie Test

https://github.com/UC-Berkeley-I-School/Project2_Chance_Fromuth_Sinani

1. Summary of Conclusions

Following the analysis of a merged dataset between the Bechdel Test and IMDB data, there are no significant relationships between features of a movie and the passing of the Bechdel test. As such, we cannot determine with any certainty that any of the features influence the outcome of the Bechdel test. Looking at the correlation analysis and features like budget, IMDB score, and gross revenue, there were minimal differences between those movies that passed the test and those that failed. When looking at the trends, there were unique findings, but none were not consistent enough to demonstrate a true relationship between the features and passing.

2. Overview of Purpose and Analytic Objectives

In an era in which the aim for female representation across all niches to hit an all time high, it is important to understand how those females are being represented. This is particularly true in pop culture and media. Some would argue that this representation is some of the most important in building the future, as it represents to the younger generation and the world how our culture views the female role in society.

When looking at movies in particular, this topic stimulated an interesting conversation about understanding certain patterns associated with such representation. In 1985 a measure known as the Bechdel Test emerged, which evaluated the representation of when in film and other fiction. This test became more widely discussed in the 2000's and continues today. It measures the representation of women in film in the three following categories. To pass, a movie must satisfy all three of these requirements.

1) at least two women are in it,
2) the women talk to each other, and
3) they are talking about something besides a man.

With the ability to merge datasets, however, there is an opportunity to look at another facet of the Bechdel test – what features of a movie are most important to that movie passing the Bechdel test**.** Using a combination of the Bechdel movie test data and IMDB movie data, one could identify patterns that would be identifiable across the film industry and look for strong correlations with Bechdel scores. As such, through the consolidated dataset, the objective of this **report is to identify what attributes of a movie contribute most to passing the Bechdel Test.** To meet this objective, we broke down our analysis into two components: analyzing general trends and assessing correlation between the Bechdel score and different features of a movie.

Within the general trend analysis, our intent is to assess if there are any significant trends that could indicate a potential correlation with a higher or a passing Bechdel score. For example, are

there specific directors or plot types that occur most often with a passing Bechdel test. Below lists the trend analysis we elected for this portion of our study. Of note, given the wide range of possible options within these features, we determined that a trend analysis versus a true correlation analysis was most appropriate.

- Changes in Bechdel Scores Pass Rate Over Time
- Genres with Highest Pass Rate; with Lowest Pass Rate
- Most Common Plot Topics with Passing Bechdel Scores; with Failing Bechdel Scores
- Movie Rating vs. Bechdel Scores

To best meet the objective of the study, however, we are also going to look at the correlation of various features of the movie to the Bechdel test. In doing so, we will be able to highlight how a change in budget or genre could impact the likelihood of a movie passing the Bechdel test. Below lists the type of correlation analysis we have elected for this portion of the study. Of note, during this portion of the analysis, we will look not just as those passing the test but also how these features impact the score (0-3) with the intent of measuring incremental correlation.

- Revenue & Bechdel Score
- Budget & Bechdel Score
- IMDB Score & Bechdel Score

In addition to the directed analytic questions above, there are several sub-questions that we will analyze to ensure we are evaluating all aspects of the study. These sub-questions are comparative in nature to illuminate if there is a directional correlation or trend between the two variables. We listed those questions below:

- What is the average IMDB score of a movie that passes the Bechdel test as compared to the average IMDB score of a movie that fails the Bechdel test?
- What is the difference in gross revenue for a movie that passes the Bechdel test as compared to a movie that fails the Bechdel test?
- Which directors have the highest percentage of Bechdel passing movies as compared to which directors have the lowest percentage?
- How many of the top 5 directors with passing Bechdel tests are male vs. female?
- How many of the top 5 actors in passing Bechtel tests are male vs. female?

3. Description of the Dataset

This study utilizes two distinct Kaggle datasets that were merged together to create a new dataset. The first dataset is Movie Bechdel Test Scores | Kaggle, which is a database of all movies evaluated and scored under the Bechdel testing rubric. The second dataset is IMDB 5000 Movie Dataset | Kaggle, which provides several features of a movie. Of note each of these datasets have unique attributes and features/columns, but both have a column for the movie's IMDB number. While the format of this number was different between the two datasets, a simple transformation technique allowed for the same format.

Description of Original Datasets

Investigating the datasets more closely, starting with the Bechdel movie test scores, it consists of 9,372 movie scores in total. The movie release dates range from as early as 1874 to as recent as 2021. The test score that we are particularly focused on contains inputs ranging from 0 to 3. Zero being that none of the three Bechdel requirements were met and as each requirement gets met the score rises until it becomes a passing score of three. The following are the variables of interest within this dataset and description of input *(variable name / data type - range)*:

- Bechdel Score (rating / integer - 0 to 3)
- IMDB ID (imdbid / float - 1 to 15943414)

This dataset is quite simplistic in comparison to the information provided by the IMDB movie dataset. Although only having 5042 entries, there is a wide breadth of variables included in the IMDB dataset with 28 unique variables. The following encompasses the most relevant variables to this study and a brief description of the inputs *(variable name / data type - range)*:

- Bechdel Score (rating / integer - 0 to 3)
- IMDB ID (movie_imdb_link / string - includes tt number that is the IMDB ID)
- Revenue (gross / integer - null to 760505847)
- Rating (content_rating / string - R, PG-13, PG, Not Rated, G)
- Genre (genre / string - Drama, Comedy, Comedy|Drama, Comedy|Drama|Romance, Comedy|Romance)
- Budget (budget / integer - 218 to 12.2Bill)
- IMDB Score (imbd_score / float - 1.6 to 9.5)
- Directors (director_name / string - multiple)
- #1 Actor (actor_1_name / string - multiple)
- Year Released (title_year / integer - 1916 to 2016)
- Plot Keywords (plot_keywords / string - multiple)

Strengths & Weaknesses of the Datasets

When investigating weaknesses associated with the datasets in question, the biggest one is the fact that the two datasets do not consist of all the same movies. The results in the overall merged data set shrinking and having less entries to conduct an analysis on. Contributing to this shrinking is the fact that both datasets have a varying range of dates of movies included. Once merged, it is evident that matched movies ranged over a 9 year period from 2007 to 2016. Additionally, the IMDB Dataset included several null values in key fields/columns of interest.

The strengths of the datasets, specifically the IMDB movie dataset, is the broad scope of variables being included allowing for several different correlations to be run and not contribute to a sense of restrictiveness.

Data Merging, Cleaning, and Transformation

In terms of formulating the final database used for analysis purposes, the Bechdel movie dataset alone did not contain the complexity of variables necessary to conduct the wide variety of correlation tests desired which is where the incorporation of the IMDB data set was introduced. By merging the Bechdel dataset with the IMDB dataset through the IMDB ID, the new dataset provided greater features for trend and correlation analysis on each of the Bechdel-rated movies.

Prior to the merge, we conducted transformation and cleaning of the individual datasets. For the IMDB dataset, we transformed the IMDB ID to match the Bechdel movie dataset, removed 'xa0' from the movie title field, and we removed the '.0' in front of the title year. We also dropped columns/features that were not relevant, creating a new dataframe 'movie_df'. On the Bechdel dataset, we created two new columns. The first was a transformation of the existing IMDB column, removing the '.0' from the end of the ID. The second, was a new column indicating if the movie passed or did not pass the Bechdel test, using the rating column as the source and identifying those with a rating of greater than 2 as an indication of pass. This offered simplicity when running correlations between variables and interpreting the analysis. Like the IMDB dataset, we then created a new dataframe, bechdel_df, including only those columns/features that were relevant to our analysis. These included variables such as director/actor facebook likes, supporting actor/actress names, language, and several more.

The final transformation we conducted on both initial datasets. This was to remove all duplicates of the IMDB ID from each of the datasets and set the index to the IMDB ID. Before removing duplicates, we analyzed the duplicate information and found that most of the inputs were exactly the same. As such, we removed the second iteration of each duplicate.

We then conducted a merge of the two datasets and continued our final phase of data cleaning. We merged the data using a left join on the Bechdel dataset using the IMDB as the key value for the merge. This merge was necessary in order to gain additional information about attributes that could be linked to a passing Bechdel movie score that was not obtainable by a single dataset. Following the merge, we conducted additional data cleaning. The first step dropped any of the rows that did not have a value for the columns/features of interest. We then reset the index but kept the IMDB ID as a column.

Of note, in our analysis of the raw datasets, we identified very few null values in the Bechdel data but several in the IMDB dataset. As such, we expected that there would be a significant drop in usable data and our expectations were founded - total movies being analyzed dropped from 9413 movies in the Bechdel dataset to 804. Analyzing the null values in the merged dataset, a lack of budget data and revenue were the most significant features in the decrease. The fact that there were 139 movies in the Bechdel dataset that premiered before 1916, which is the min year

in the IMDB dataset, also contributed to the drop. The biggest reason, however, for the drop is a limited overlap in IMDB IDs between the two datasets.

Summary of Assumptions

- Duplicate IMDB IDs in the raw Bechdel dataset were all similar; once removed the IMDB ID could act as the index for the dataset
- Duplicate IMDB movie links in the IMDB dataset were all similar in the fields that were important to the study; once removed the IMDB ID could act as the index for the dataset
- If there are any null values in the fields/columns we are analyzing, they will be dropped

4. Discussion of Findings: Trends and Correlation

Trend Analysis

Pass Rate by Year

Beginning with assessing the overall passing rates of the movies by year included in the merged dataset, as seen in Figure 1.0, it is evident that passing rates almost consistently have been above 50% despite the years 2008 and 2009. It can be seen that in 2007 it is calculated that 100% of the movies passed the test, but taking note of the n=2 total movies makes us conclude this is relevant to the dataset and not as applicable to real life trends. Similarly, Figure 1.1, demonstrates the fluctuations with passing rates across the years. The 100% passing rate can be seen as the initial spike in the graph, but the sudden drop should be interpreted with caution.

	title_year	total_movies	total_passes	pass_rate
0	2007	2	2	1.000000
1	2008	23	9	0.391304
2	2009	75	32	0.426667
3	2010	97	55	0.567010
4	2011	118	60	0.508475
5	2012	112	65	0.580357
6	2013	130	80	0.615385
7	2014	122	64	0.524590
8	2015	110	60	0.545455
9	2016	53	33	0.622642

Figure 1.0 Table displaying passing rates

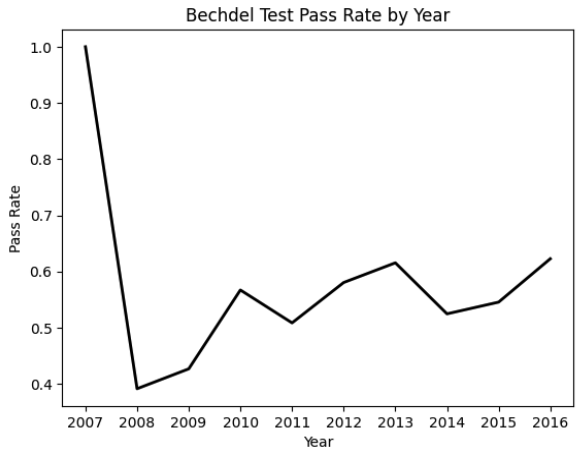
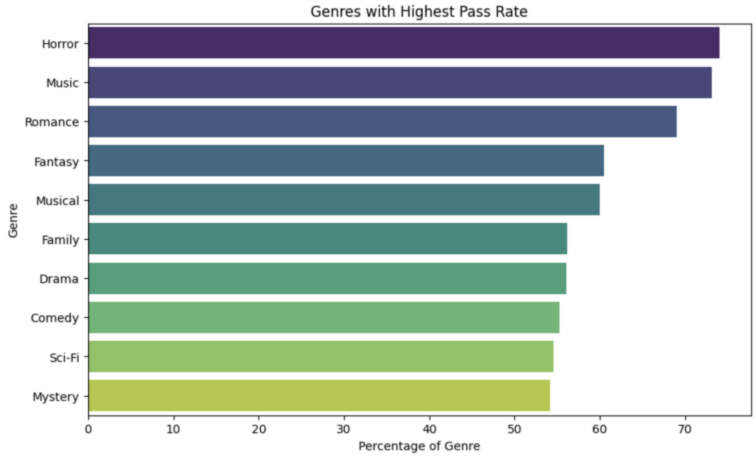


Figure 1.1 line graph demonstrating fluctuations in passing rates

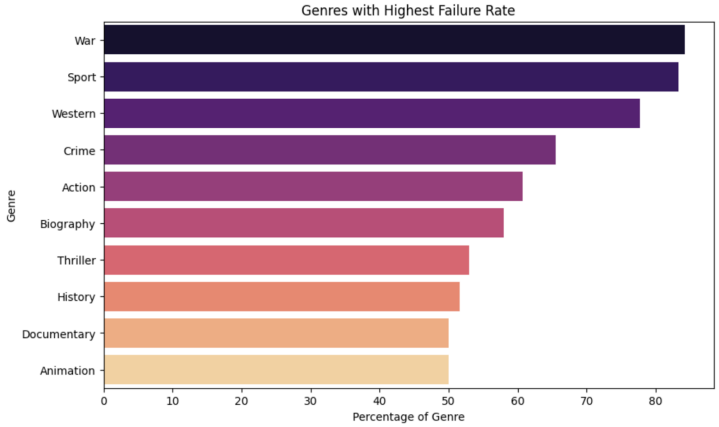
Genre & Passing Rates

Attempting to investigate the first possible attribute contributing to such passing rates, movie genres were investigated to see which were associated more closely with either passing or non passing score. As seen in figure 2.0, nearly 70% of both horror and music based movies had a passing Bechdel score. Closely following these genres are romance, fantasy, musical, and family,



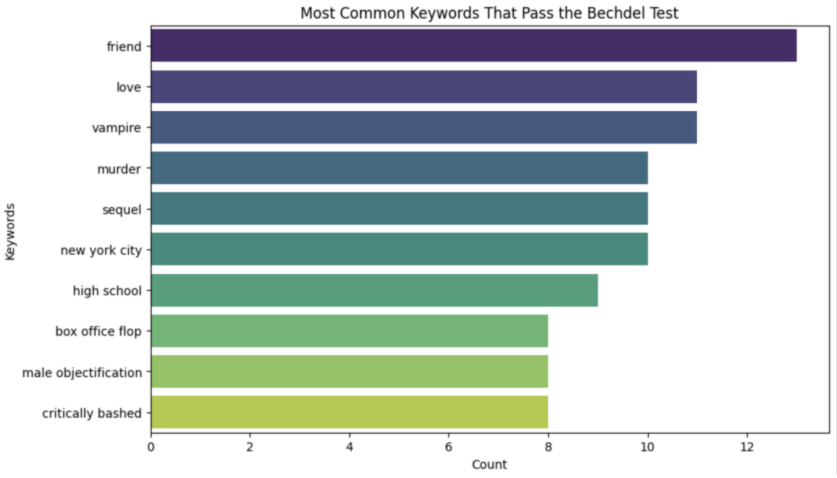
which may be explained by the tendency for such films to be targeted towards the female population. In contrast, when investigating the genres with the lowest passing rates, as seen in figure 2.1, genres like war, sport, western, crime, and action take the lead. Due to such genres typically being towards the male population, a

Figure 2.0 Genres associated with the highest Bechdel passing rates



juxtaposition from the genres mentioned before, can best explain why they experience the lowest passing rates. The Bechdel test aims to evaluate the bias against women in films and the genres with the lowest passing rates are known for not being focused on mitigating such a bias.

Figure 2.1 Genres associated with the highest Failure Rate



Keywords & Bechdel Test Pass vs Fail

Similar trends are demonstrated when observing what key plot words within a movie are linked to it receiving a passing or failing Bechdel movie score. Movies with keywords such as friend, love, vampire, male objectification, and high school tended to have higher counts of passing the Bechdel test, as seen in figure 3.0.

Figure 3.1 Plot keywords associated with a passing Bechdel score

Comparing such keywords with ones associated with a failed Bechdel test, as seen in figure 3.1, similar explanations from the section above can be applied. Words such as alien, spy, death, police, and fbi are seamlessly aligned with the genres with the lowest passing rates which once again emphasize the focus towards a male centered movie approach.

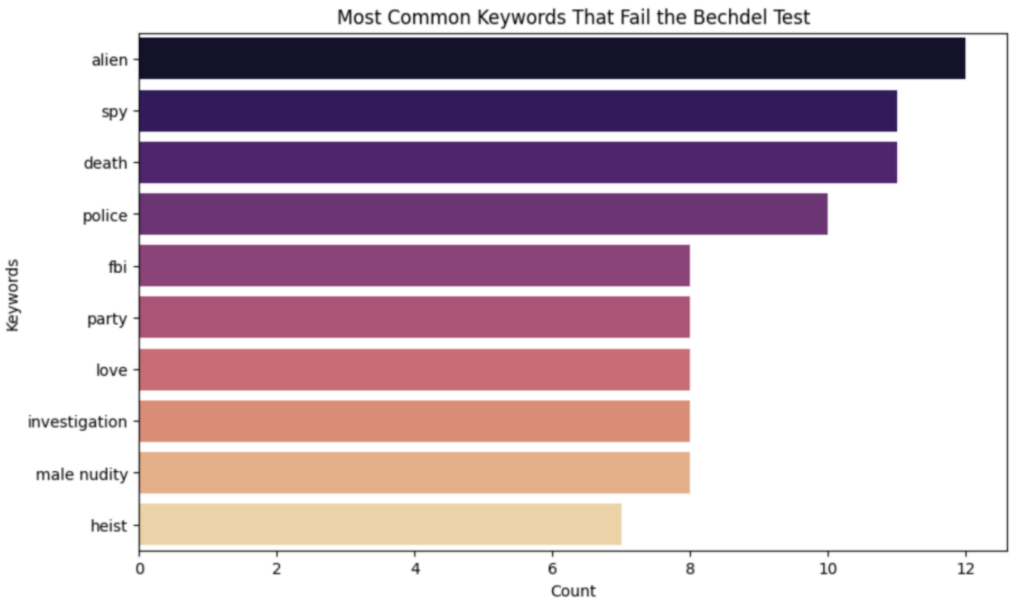


Figure 3.1 Plot keywords associated with a failed Bechdel score

Rating & Bechdel Test Pass vs Fail

Attempting to investigate if a passing score can be explained by the overall movie rating led to results not as polarizing as expected, as seen in figure 4.0. Rated R movies tended to have an equal distribution of passing and non passing Bechdel movie scores. In terms of PG and PG-13 rated movies, passing bechdel scores seem to push the equal distribution to a 60-40 representation instead.

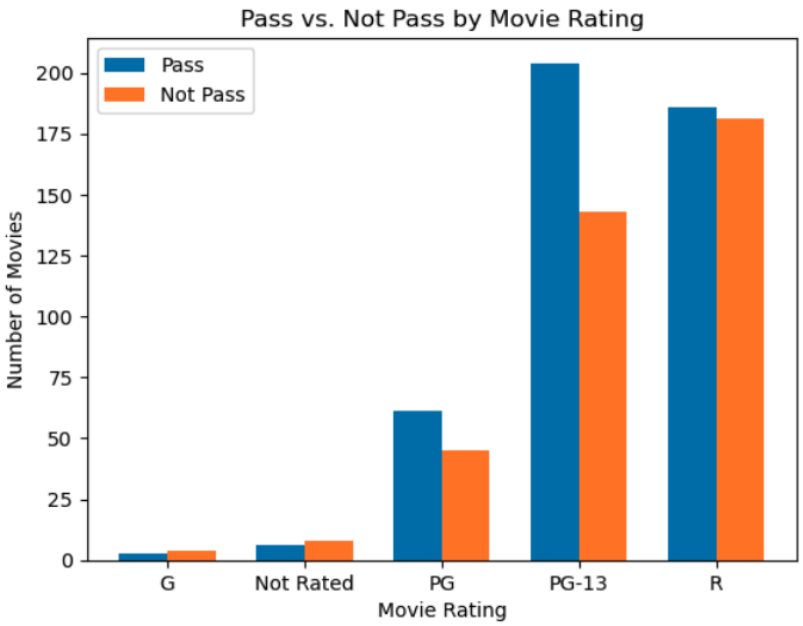


Figure 4.0 Histogram of passing count via movie ratings

Correlation Analysis

Revenue & Bechdel Score

The existing data demonstrates that the impact of the outcome of the Bechdel test on gross revenue is statistically similar in regards to spread, but those that pass the Bechdel test have slightly higher measurements. Looking at figure 5.0, both those that pass and those that fail have a relatively similar average with those that pass having a slightly higher mean. The quartiles and the max is also higher for

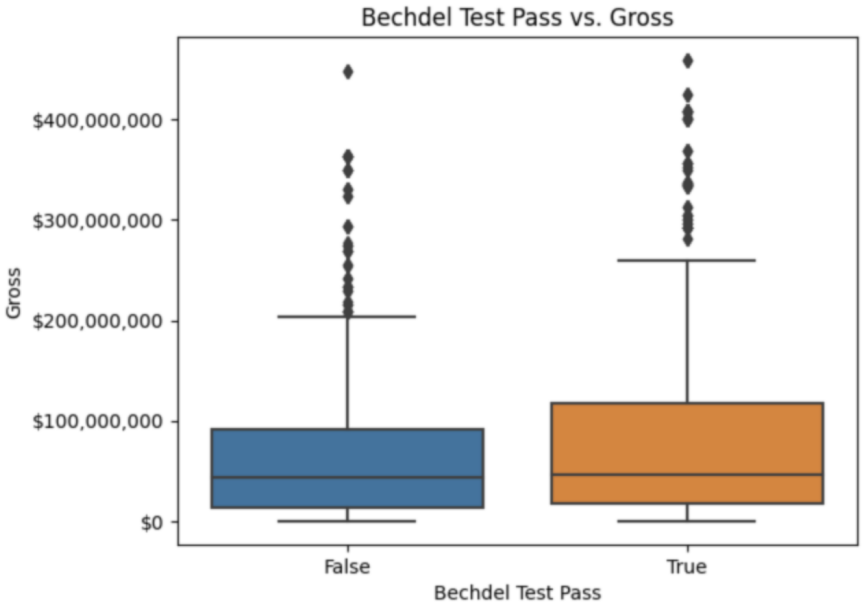


Figure 5.0 Bechdel test vs. gross revenue

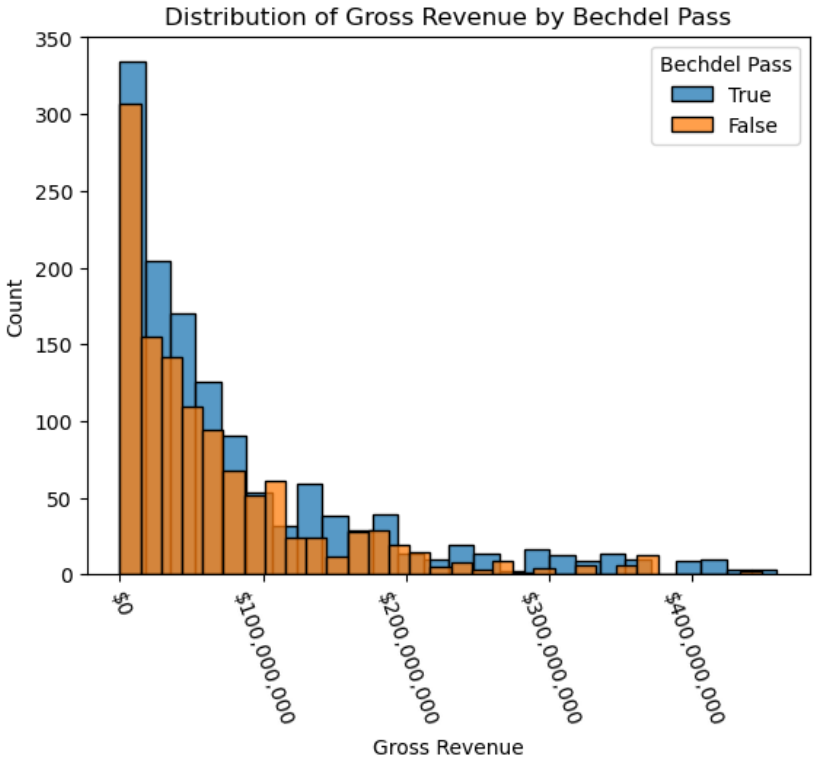


Figure 5.1 Number of Bechdel test passed over gross revenue

those that pass the Bechdel test, but the min is lower. Figure 5.1 confirms as well, that there are a greater number of passing movies with lower gross revenue, and those that pass the Bechdel test appear to have more outliers at higher revenues. Both distributions follow a skewed right model as expected with counts tailing off the higher gross revenues.

Budget & Bechdel Score

Switching from investigating gross revenue to budget, it is evident similar distributions are taking place, as seen in figure 6.0. Movies that pass or fail the Bechdel test have almost an identical interquartile range, with the maximum budget being slightly higher for movies that pass the test. Additionally, both instances contain several outliers within the spread of their data.

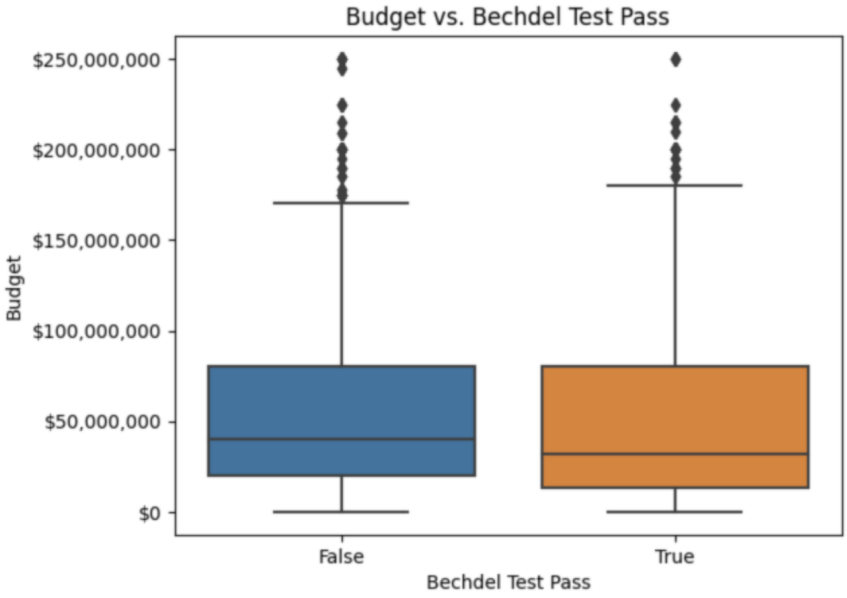


Figure 6.0 Bechdel test passing vs. budget

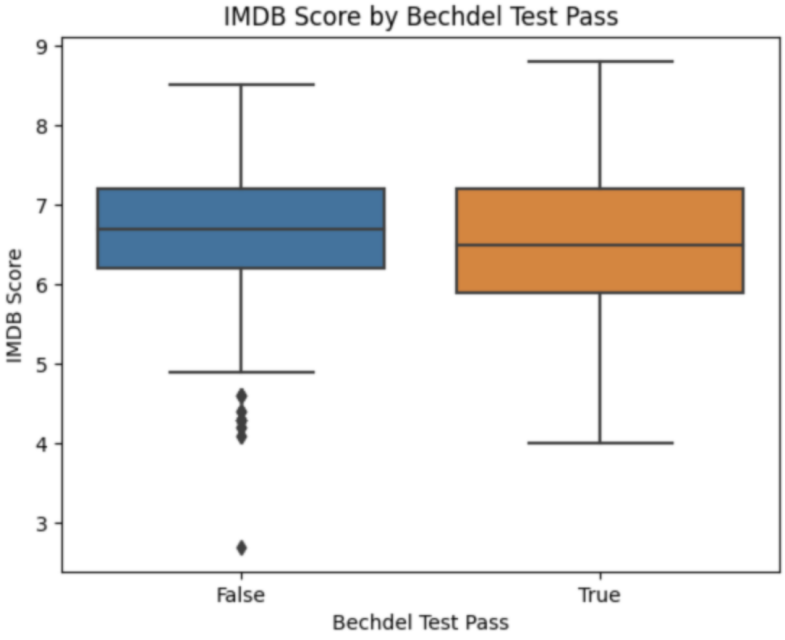


Figure 7.0 Bechdel test passing vs. IMDB scores

IMDB Score & Bechdel Score

Addressing the relationship between IMDB score and passing the Bechdel test, demonstrated in figure 7.0 , it seems as if non-passing scores have a tighter spread ranging from an IMDB score of 5-9 with several lower bound outliers. In contrast, passing tests had an IMDB score spread from 4 to 9 with no outliers. Both have an average IMDB score around 6.75. Significant differences between the two boxplots are not demonstrated hinting at a Bechdel passing movie not having an attribute of a high IMDB score.

5. Conclusions

The results of our findings do not indicate a strong relationship between the features of the movies we measured and the passing of the Bechdel test. This was true for both the trend analysis as well as the correlation analysis, although there were unique observations and differences in the features we examined in the trend analysis section. As such, we cannot conclude that any of the features influence the outcome of the Bechdel test. Of note, given we only looked at 804 of the 9361 movies that were evaluated in the Bechdel dataset, it would be hard to definitively state that none of the features influenced the likelihood that a movie would pass the Bechdel score. To confidently verify this conclusion, more data is needed.

Looking at the statistical analysis of revenue, budget, and IMDB score, there is very little correlation between these features and the Bechdel test. When looking at the quartiles for each of these features, it does appear that the results from the movies that pass the test are somewhat lower than those that pass. The spread of results may be weighted towards the lower scale but that there are higher outliers that offset this and increase the mean. While this split indicates a greater variability in the features of those movies that pass, it is relatively minimal as opposed to those movies that fail.

When looking at the trend analysis, there were some unique observations, but none were substantially different enough that we could comfortably conclude correlation between the features and the likelihood of passing the Bechdel test. For example, the horror genre demonstrated the highest passing rate, but it was only a few percentage points from music and romance. In fact each of the top rated genre's only have an approximately 20% split. Similarly, those genres most common to movies that fail include war, sports and westerns. There is a slightly larger split here at a 35% split but, again, none demonstrate a significant difference. Uniquely, keywords came close to highlighting a trend with a notable trend. Friend, for example, highlighted an advantage (3x more counts) over murder, sequel, love and New York City in most common keywords for movies that passed the Bechdel test. Conversely, death, alien and spy had the same size advantage over policy in the most common keywords for those movies that failed. Given we evaluated 804 movies, however, this trend was interesting but not significant.