Pass the Popcorn and the Bechdel Test

W200, Section 4
Project 2: Data Analysis
Richard Ryu, TK Truong, Dili Wang

1. Introduction

Movies often reflect the values and state of society. Over time, female representation on the big screen has evolved as women's roles have changed in society. There are many ways to examine how women are portrayed in film. One method, the Bechdel Test, has received a lot of attention in the past few years for its simplicity. It derives from a comic strip called *Dykes to Watch Out For* by Alison Bechdel, first published in 1985.

Excerpt from *Dykes to Watch Out For* regarding the Bechdel Test

In order to pass, a movie has to satisfy three criteria: it has to have at least two named female characters, they have to talk to each other, and they have to talk to each other about something other than a man. Still, many films end up failing the Bechdel Test--even from the first criterion. The test is meant to highlight that many movies depict women as passive and without agency. Often times, their characterization is developed in relation to male characters.

However, the Bechdel Test isn't perfect. Some critics point out that some films narrowly pass all the criteria due to the female characters talking about shopping, marriage, or babies. This is just one of the ways the Bechdel Test may miss nuances in female representation (i.e. movies that pass may still have sexist content), but its simple criteria focuses on basic tenets of equal representation.

2. Research Questions

Utilizing the user-maintained Bechdel Test database with other datasets from Internet Movie Database (IMDB), Box Office Mojo and Forbes, we hope to analyze and compare the differences between the movies that pass and fail the test in relation to content, popularity, financial gross, and star power.

1. How many of the top 250 highest-rated IMDB movies pass the Bechdel Test?
2. How many movies pass the Bechdel test when grouped by IMDB ratings?
3. Does Runtime have any effects on Bechdel ratings?
4. How many of the the top earning movies pass the Bechdel Test?
5. Since women's roles have changed over time, does the number of films that pass the Bechdel Test increase over time?
6. What genres have the most movies that pass the Bechdel Test?
7. Are highest grossing actors and actresses in movies that pass the Bechdel Test?
8. Are the current highest paid actresses' movies passing the Bechdel Test?

3. Datasets

3.1 Bechdel Test Ratings

Source: https://bechdeltest.com/api/v1/doc

Variable Name	Description
bechdel_id	Unique ID for Bechdel Test website
imdb_id	Unique ID from IMDB titles table
rating	Bechdel Test rating on a scale of 0-3 (Only 3 indicates a pass)
title	Title of the movie
release_yr	Release year of movie

3.2 IMDB title.basics

Source: https://www.imdb.com/interfaces/

Variable Name	Description
imdb_id	Unique ID for media titles in IMDB
titleType	Type of the title such as movie, tvMovie, tvEpisode, etc.
primaryTitle	Most popular title
originalTitle	Original title
isAdult	0: non-adult title, 1: adult title
startYear	Release year for title
endYear	For TV Series end year. Not applicable for other types
runtimeMinutes	Runtime of the title in minutes
genres	Array of up to three genres associated with the title

3.3 IMDB title.principals

Source: https://www.imdb.com/interfaces/

Variable Name	Description
imdb_id	Unique ID for media titles in IMDB
ordering	Number to identify rows for a tconst ID
nconst	Unique ID for name/person

category	Job category (actor, director, crew member, etc.)
job	Specific job title if applicable
characters	Names of character played if applicable (mainly actors and actresses)

3.4 IMDB title.ratings

Source: https://www.imdb.com/interfaces/

Variable Name	**Description**
imdb_id	Unique ID for media titles in IMDB
averageRating	Weighted average of all individual user ratings
numVotes	Number of votes the title received

3.5 IMDB name.basics

Source: https://www.imdb.com/interfaces/

Variable Name	**Description**
nconst	Unique ID for name/person
primaryName	Name by which the person is most often credited
birthYear	Birth Year
deathYear	Death Year
primaryProfession	Array of top-3 professions of the person
knownForTitles	Array of tconsts the person is known for

3.6 Box Office Mojo Highest Gross All Time World

Source: https://www.boxofficemojo.com/alltime/world/

Variable Name	**Description**
Rank	Rank number in terms of gross
Title	Movie Title
Studio	Studio abbreviation
Worldwide	Worldwide gross in millions
Domestic	Domestic gross in millions
Domestic %	% of domestic gross out of worldwide gross

Overseas	Overseas gross in millions
Overseas %	% of overseas gross out of worldwide gross
Year	Release year

3.7 Box Office Mojo People Index (Highest Grossing Actors)

Source:
https://www.boxofficemojo.com/people/?view=Actor&pagenum=1&sort=sumgross&order=DESC&&p=.htm

Variable Name	**Description**
Rank	Rank number in terms of total gross
primaryname	Actor/Actress name
Total_Gross_Millions	Total gross in millions
Num_Movies	Number of movies acted in
Average_Millions	Average gross in terms of millions (Total gross / Number of Movies)
Num_1_Picture	Top movie for the actor/actress
Gross_Millions	Gross for top movie in millions

3.8 Forbes Highest Paid Actresses in 2017

Source:
https://www.forbes.com/pictures/598b86c731358e60d775bca6/worlds-highest-paid-actre/#6643d43e60b0

Variable Name	**Description**
Rank	Actress rank according to salary
primaryname	Actress name
Salary_Millions	2017 salary in millions

4. Data Transformation

4.1 Bechdel Test Dataset

After obtaining the Bechdel Test data from the website, we loaded the JSON file in Python to explore the data and convert it into a CSV format. We did some initial checks to look for duplicates and anomalies in the dataset.

The Bechdel Test rating is codified on a scale of 0-3, where 0 means the movie does not have at least 2 named female characters, 1 means the movie has at least 2 named female characters, 2 means these two characters talk to each other, and 3 means they talk to each other about something other than a man. Only a score of 3 indicates passing the Bechdel Test. In our analysis, we kept the nuances of this type of scale, but also occasionally grouped 0-2 as fail and 3 as pass.

4.2 Filtering and Cleaning in SQL

Due to the volume of some the IMDB tables, we did some initial filtering and cleaning in SQL. The title.principals table was over 29 million rows with extraneous information that we did not need since we're interested in analyzing actors and actresses. We imported the table into PostGreSQL and filtered the table down to just actors and actresses, which was over 17 million rows. We joined this subset of title.principals table with the Bechdel Test data to get rows from only the movies we have Bechdel Test data for, which resulted in approximately 31,000 rows. We also queried other subsets from title.principals for the Highest Grossing Actors/Actresses and Highest Paid Actresses analysis to get data for all of their films. These were exported into CSVs and analyzed using pandas.

4.3 File Merging and Joining

The main file from IMDB containing film names, runtime, and genres is called title.basics.tsv. This file was read into a Pandas dataframe and subsequently merged with the Bechdel Test data set in 4.1 using the imdb_id, contained in both tables, as the key. This dataframe was later used for a number of the analysis questions, such as

6.3 - what is the relationship between bechdel test scores and runtime of films?
6.6 - which genres have the most movies that pass the bechdel test?
6.7 - are highest grossing actors and actress in movies that pass the bechdel test?

Next we merged this dataframe containing Bechdel test scores and IMDB movies data with the IMDB file title.ratings.tsv using the imdb_id. This merged dataframe was later used to perform analysis for questions like:

6.2 - how many movies pass the bechdel test when grouped by IMDB ratings?
6.1 - how many of the top 250 highest-rated IMDB movies passed the Bechdel Test?

We then merged the dataframe for bechdel_titlebasics_ratings with the dataframe for the web scrape from Box Office Mojo for the top earning movies of all time. For this merge, we created a key using the primaryTitle and Year concatenation for both tables. This allowed us to answer question 6.4 - how many of the top earning movies pass the bechdel test.

In further analysis of questions 6.7 and 6.8, name.basics.tsv and a clean version from SQL of title.principals.tsv were further merged, using nconst (an IMDB constant for individual actors names). This resulted in a dataframe of all actors, their names, and the movies the were in. Using either the web scrape from Box Office Mojo for the top 50 highest grossing actors, or

the list of Forbes highest paid actress of 2017, we filtered this dataframe twice to create 2 additional data frames for actors of interest pertain to each question. Both dataframes were individually merged with a dataframe containing films and bechdel ratings, using imdb_id as the key.

Here is the summary of our File Joining and Merging:

Merged Tables	**Merge / Join Key**	**Filters**	**Questions**
Title.basics.tsv + Bechdel Test Table	imdb_id	Only IMDB movies with bechdel ratings were included in resulting dataframe	● Relationship between runtime and bechdel scores ● Which genres have the most Bechdel passing films ● Which highest grossing actors and actresses are in movies that pass the Bechdel test
(Title.basics.tsv + Bechdel Test Table) + title.ratings.tsv	imdb_id	Only IMDB movies with Bechdel ratings and IMDB ratings were included in resulting dataframe.	● How many movies pass the bechdel test when grouped by IMDB ratings ● How many of the top 250 highest rated IMDB movies passed the Bechdel test
Bechdel_titlebasics_ratings (result of previous 2 rows) + Box Office Mojo web scrape for top earning movies all time	Created key using concatenation of Name+Year	Only movies listed by Box Office Mojo with Bechdel ratings were included in resulting dataframe	● how many of the top earning movies pass the bechdel test?
name.basics.tsv+title.principals.tsv	nconst	Title.principles.tsv was first filtered in SQL for only Actors, then filtered again for only imdb_ids that were included in Bechdel Test Table.	● Which highest grossing actors and actresses are in movies that pass the Bechdel test ● Are the current highest paid Actresses' movies passing the Bechdel Test
(Title.basics.tsv + Bechdel Test Table) + (name.basics.tsv+title.principals.tsv)	imdb_id		
(dataframe from previous row) + (Box Office Mojo web scrape for top 50 highest grossing actors OR Forbes Highest Paid Actresses)	Actors'/Actress' Names	Only IMDB movies that the Actors/Actress starred in were included in resulting dataframes Movies were kept regardless of whether or not an Bechdel rating was retrieved.	

5. Data Integrity

5.1 Reliability of Bechdel Test data

The Bechdel Test dataset comes from user-maintained website, which comes with inherent biases. We don't have any information on the users, but it's worth noting that it's likely that a certain kind of user is interested and motivated to rate movies on the website. While the Bechdel Test has simple criteria, there may be errors as a result of "crowd-sourcing" the ratings since there are little restrictions on user input. For example, the website does allow for commenting in case a user wants to dispute a rating, but there's little moderation or standardization when it comes to ratings. There's a lack of inter-rater reliability as a result.

There could be other user input errors in the dataset as well. Initial data exploration also revealed 6 duplicate movies in the Bechdel Test dataset, which we had to investigate. It's unclear what restrictions, if any, are in place to prevent duplication. Users also manually input the movie title and release year, which can result in errors as well.

5.2 IMDB Ratings

We noticed something odd about the imdb ratings. If you look closely, movies named "The Clock" and "Aruvi" are ranked 3rd and 4th on the imdb rankings. I have never heard of such movies, while movies like "The Godfather" and "The Shawshank Redemption" make sense to be highly rated. Digging further, we noticed the number of votes associated to each movies rated on imdb under the (numVotes) variable. Based on the values from (numVotes) variable, famous movies like "The Shawshank Redemption" and "The Godfather" have more than 1 million votes, while "The Clock" and "Aruvi" have less than 10,000 votes. This led us to believe that the imdb ratings for each movie should not be taken as face value and have made the decision to filter out movies with insignificant number of votes by only counting the movies that are in the top 25% range in terms of number of votes.

5.3 IMDB Genres

IMDB provides multiple genres for many movies, and this is expected as many movies do not fall into a specific genre. In the raw data of title.basics.tsv, genres are listed in one column, separated by comma, and appear to be in alphabetical order. Therefore, we have no further data on which genre is considered the primary, nor is there data on the weight of a specific genre in the list. In this case, we considered 2 approaches to analyzing the genres -

1. Using raw count - meaning that a movie would be counted once for every genre in its list. Therefore if a movie has 7 genres, it is counted once in all 7 genres and therefore counted 7 times across the genres
2. We can split a movie into its list of genres - this would mean that if a movie has 7 genres, it would counted as 1/7 of a film in each of the 7 genres.

Ultimately, both methods are rather unfair. In the first case, movies are counted more than once and can skew the counts for some categories such as Drama and Comedy which are listed more often than others. However, the second method would then not give justice to films that are predominantly one genre, but then become diluted if its genre list is very long. Ultimately, since our analysis regarding genres pertains to a comparison of which genres have the most Bechdel Test passing films, we decided to go with method 1, although we fully realize that this would mean films are counted more than once.

6. Analysis Questions

6.1 How many of the top 250 highest-rated IMDB movies passed the Bechdel Test?

One of the initial questions that we were curious about was how many movies passed the Bechdel Test from the top 250 highest-rated IMDB movies. In order to answer this question, we converted “bechdel_titlebasics_ratings.csv” into Python Pandas dataframe and then filtered the ‘titleType’ values for movie or tvMovie. After applying the filters to the dataframe, we sorted the filtered dataframe by ‘averageRating’ in ascending order, then used the .head(250) method to obtain the top 250 highest-rated IMDB movies.

However, we noticed our filtered and sorted dataframe was inaccurate, since movies like “The Clock” and “Aruvi” appeared in the top 250 list with 165 votes and 6,343 votes respectively. These movies had higher average ratings than popular movies like “The Godfather: Part II” and “The Dark Knight” with 935,170 votes and 1,944,155 votes respectively. In order to provide an accurate representation of the top 250 IMDB movies by average rating, we decided to focus on movies or tvMovies that were in the top 25% range for number of votes by analyzing all the movies in the “bechdel_titlebasics_ratings.csv” file for number of votes using the .describe() method from the Python Pandas package. As a result of the analysis, we were able to identify the 75% cut-off range for the number of votes and removed any movies that had less than 77,045 votes from the initial dataframe, then reapplied the filters and sorting order mentioned above.

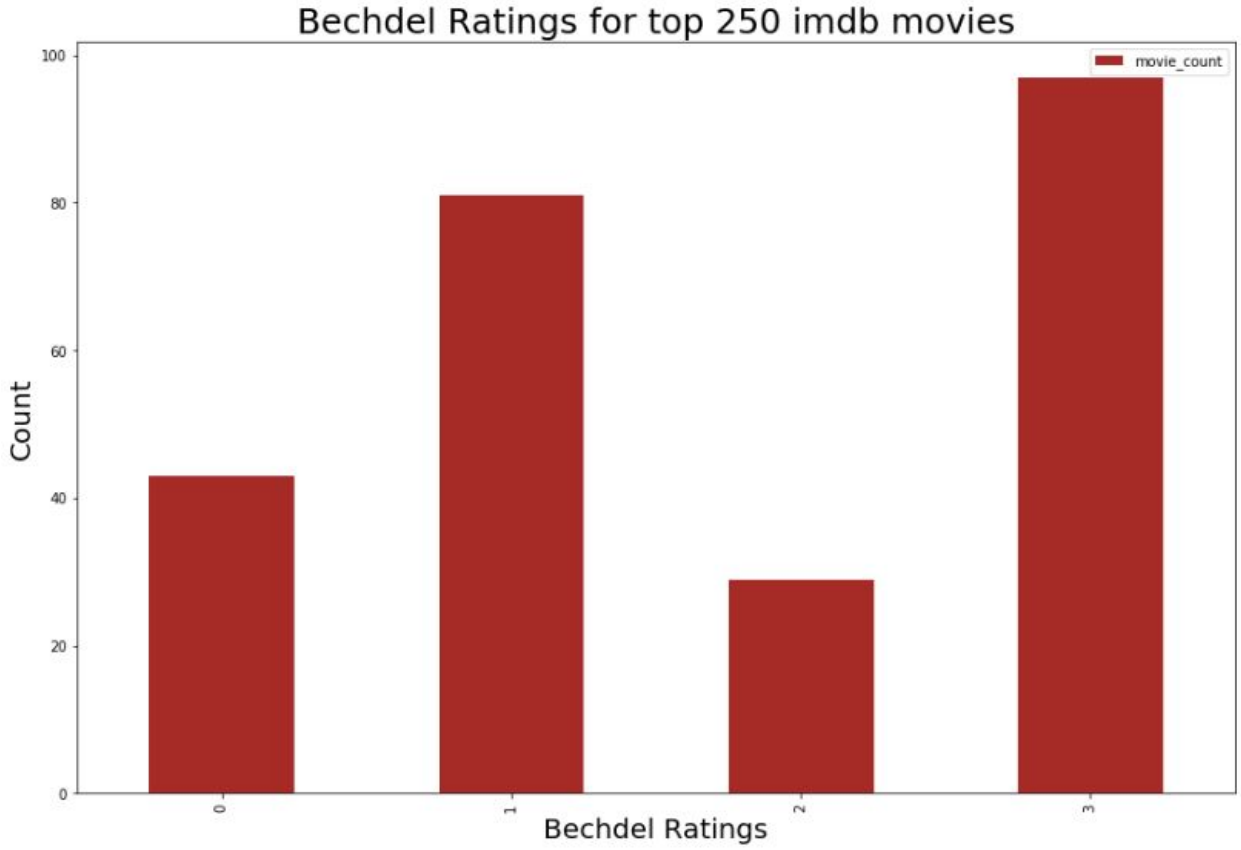
With the re-filtered top 250 highest-rated IMDB movies dataframe, we were able to learn that 97 movies passed the Bechdel Test by applying additional filter on “rating” column for value of 3 and using .describe() method. To provide some perspective, popular movies like “The Shawshank Redemption” and “Star Wars: Episode V - The Empire Strikes Back” did not pass the Bechdel test, while movies like “Star Wars: Episode VII - The Force Awakens”, “The Dark Knight”, and “Pulp Fiction” have passed the Bechdel test.

For additional analysis, we decided to examine how top 250 highest-rated IMDB movies break down in terms of Bechdel Test ratings. In order to achieve this, we grouped the re-filtered dataframe by “rating” and counted the number of movies using .groupby() method and .count() method. Then we added “percentage” column to the dataframe by dividing the number of movies in each group with the total sum of movies for all groups. Figure 1 and Figure 2 below is the result of this analysis and it was interesting to find out that 81 movies or 32.40% of the top 250 highest-rated IMDB movies did not pass the Bechdel Test despite having more than 2 female characters in the movies.

Fig. 1: Bechdel Rating Analysis for Top 250 IMDB Movies

bechdel_rating	movie_count	percentage
0	43	17.20%
1	81	32.40%
2	29	11.60%
3	97	38.80%

Fig. 2: Bar Chart for Bechdel Rating Analysis for Top 250 IMDB Movies

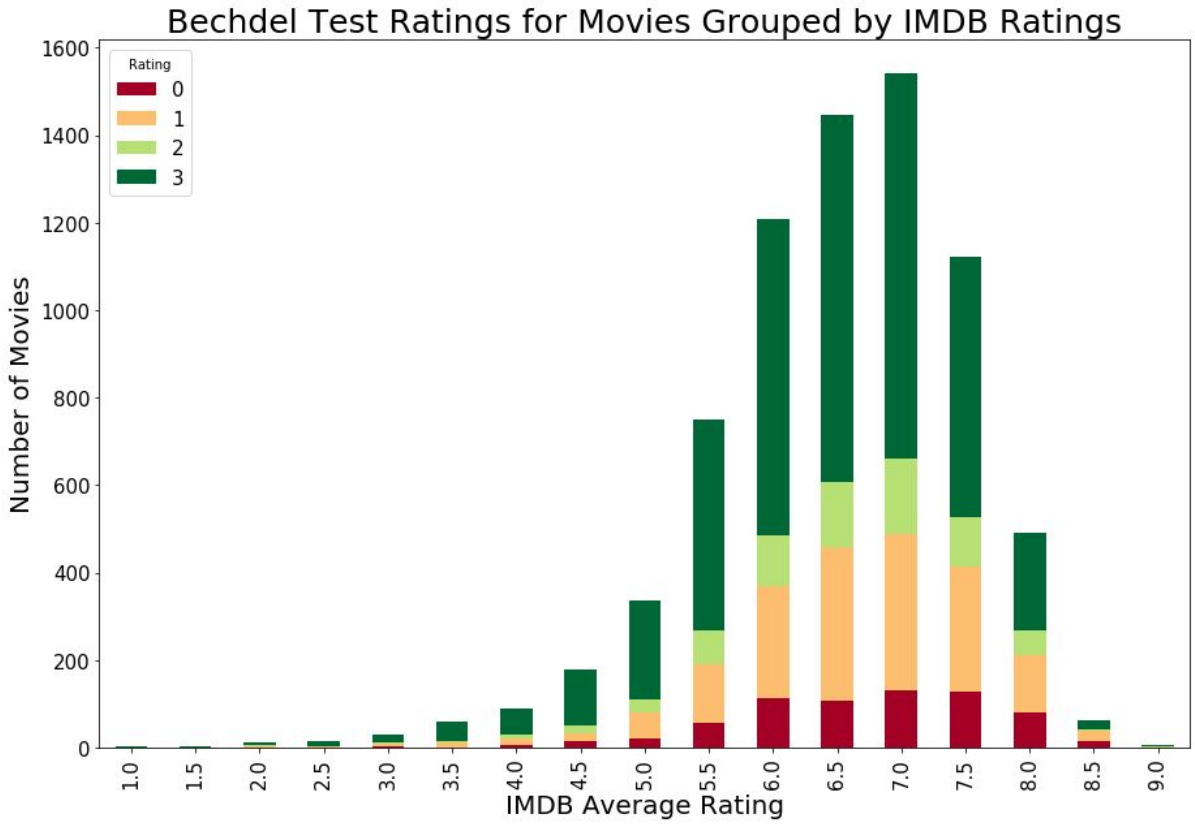


6.2 How many movies pass the Bechdel test when grouped by IMDB ratings?

In this questions, we are primarily concerned with whether or not movies with high IMDB ratings have higher Bechdel Test scores or if there is any visible correlations between IMDB ratings and Bechdel ratings. To prepare the data for analysis, pandas dataframes were created from reading the IMDB raw data files title.basics.tsv and title.ratings.tsv. A third pandas dataframe was created for Bechdel test data extracted from https://bechdeltest.com/api/v1/doc. The three dataframes were merged using imdb_id in a 2 step process, which resulted in a final dataframe where only movies with IMDB ratings and Bechdel ratings were kept. IMDB films without Bechdel ratings or without IMDB ratings were not included. Data was filtered include only title type of Movies and TV movies, and also adult films were excluded.

First, we plotted all movies that we had both IMDB ratings and Bechel ratings in a stacked bar chart, where movies were categorized by bin size of 0.5 IMDB ratings, and stacked using the Bechdel rating as the category. Please note that when reading the IMDB ratings along the x-axis, the tick marker indicates that movies within the bar have a rating of at least the value of the tick mark, but does not exceed the rating of the next tick mark. In other words, movies within the 8.5 bar means that they have a rating of greater than or equal to 8.5 but less than 9.0.

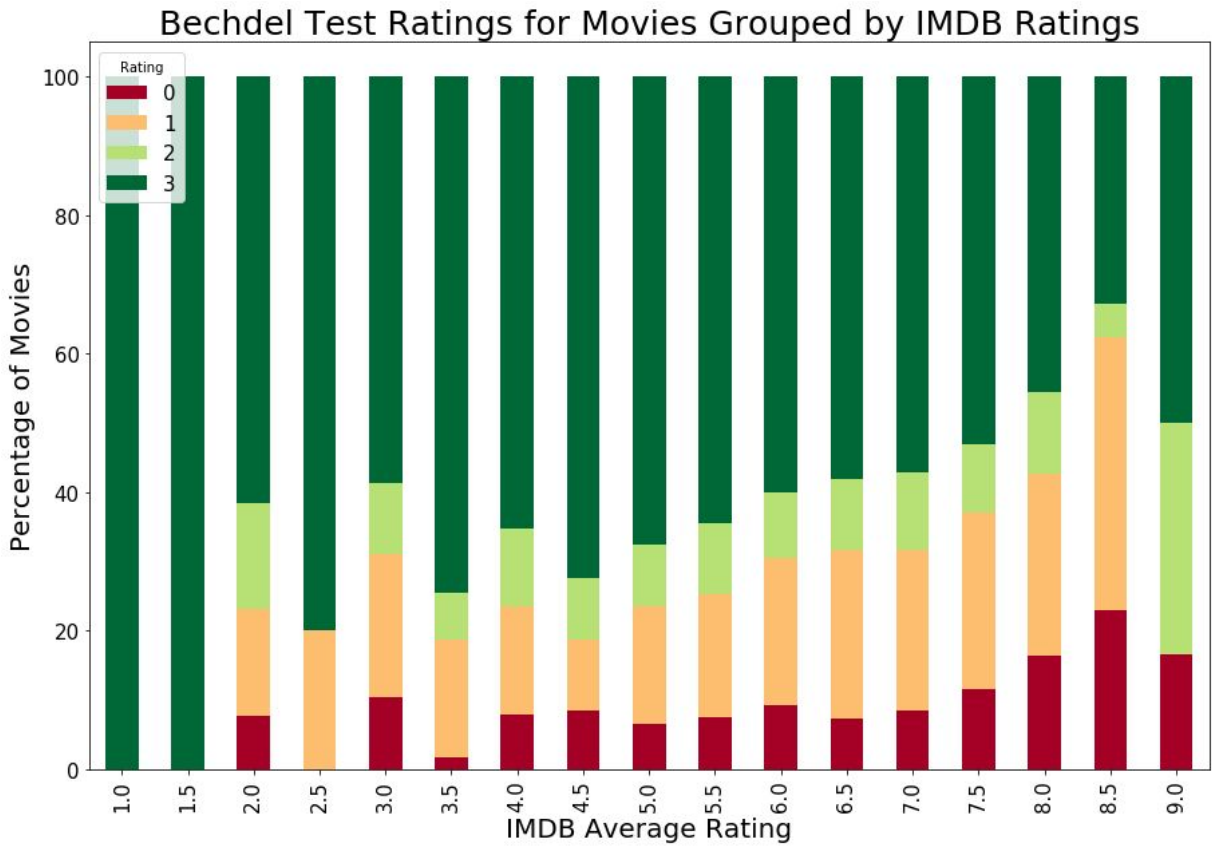
Fig. 3: Stacked Bar Chart for Bechdel Test Ratings for Movies Grouped by IMDB Ratings



Since this bar plot displays raw counts, it shows some selection bias within the Bechdel data, as Bechdel rated films are not evenly distributed amongst all IMDB ratings. The majority of movies with Bechdel ratings from the Bechdel test data source have an IMDB rating between 4.5 and <9.0, while there are very few movies with ratings less than 4.0 and greater than 9.0. We see that movies with IMDB rating from 6.0 to 7.5 have the greatest raw counts of movies that pass the Bechdel test (with a score of 3), however, these IMDB rating groups also have the greatest counts of movies compared to other IMDB ratings.

To better ascertain if there is any noticeable correlation between Bechdel pass rate and IMDB rating, we decided to further transform the data so we can plot the stacked bar by percentages within each category:

Fig. 4: Movies grouped by IMDB Average Rating of bin size 0.5. Percent stacked bars are calculated by percentage of films within each bin that have a Bechdel score



The first observation from this plot is that the percentage of movies that pass the Bechdel test with a score of 3 actually appears to decrease for increasing IMDB average rating, within the range of IMDB rating from 4.5 to <= 9. In comparison to the plot of raw counts, this is also the IMDB rating range where the majority of the Bechdel data lies. In our consideration of percentage pass rate of Bechdel films, it would seem that a high IMDB average rating does not therefore guarantee a higher pass rate for the Bechdel test.

In analyzing data for IMDB ratings, our biggest concern was how to consider the fairness of an IMDB rating given the number of votes for each rating. For instance, can we really say that a movie with 10 votes and an IMDB rating of 9.5 is better than a movie with 3,000,000 votes and an IMDB rating of 8.5? Are they even comparable? Without further knowledge of the data of voter demographics, we felt that number of votes should not be simply ignored.

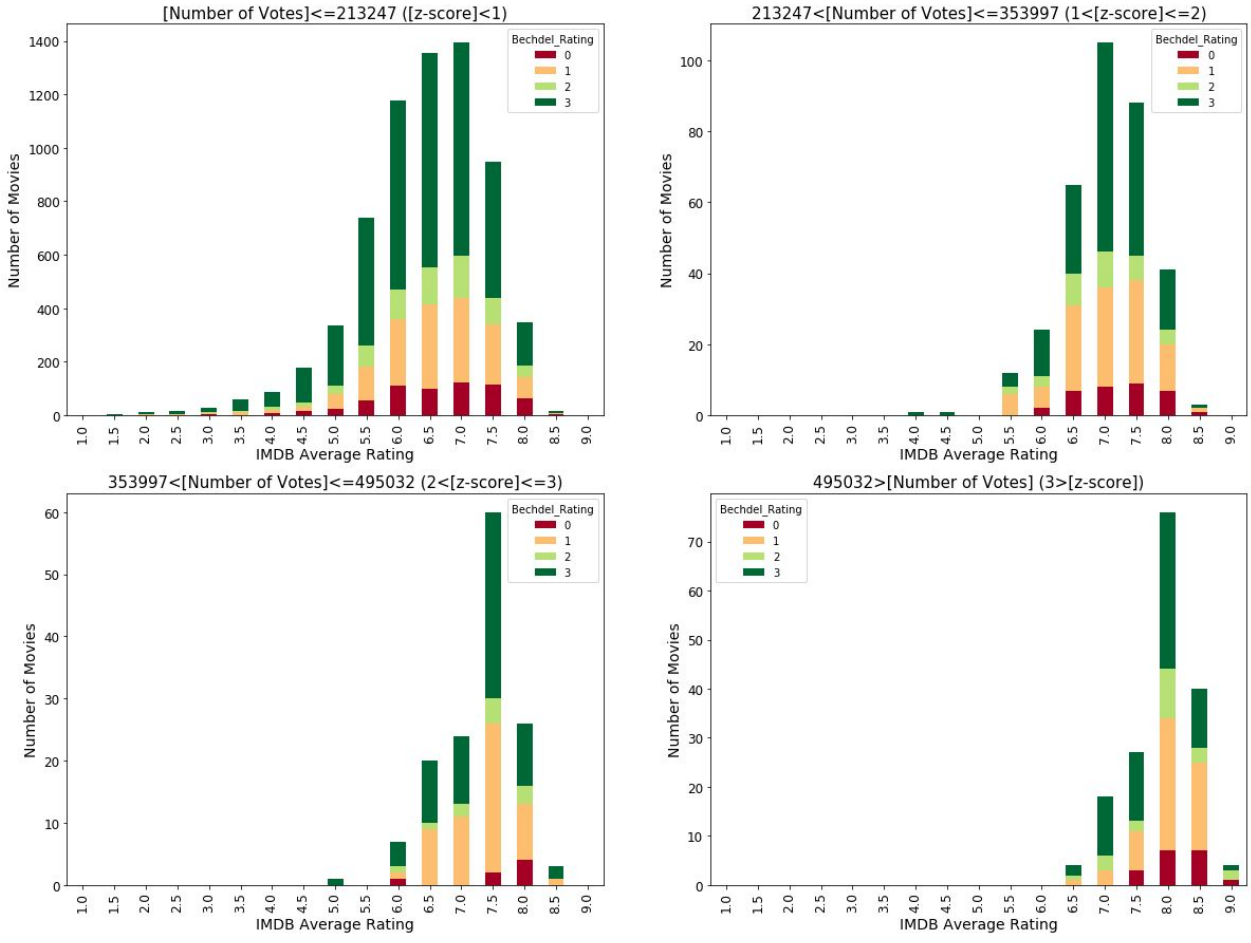
In taking the number of votes into consideration, we first applied a z-score function to the IMDB number of votes in merged dataframe. A quick exploration in Jupyter notebook showed that the majority of films have a very low number of votes with a mean of 72,524 votes and a standard deviation of 141,100 votes. However, the maximum number of votes is close to 2,000,000. This means that the majority of movies, more than 95%, have number of votes less than 354,000. In the range from 354,000 to 2,000,000, there are very few movie counts. It is important to note, however, that we did not filter out movies with large number of votes, because we use the number of votes as a proxy for popularity and thus are really interested in how Bechdel test scores compare for movies with large number of votes.

In a preliminary study, we decided to split the data into subsets, so that we can compare the number of Bechdel passing versus IMDB ratings for movies that have similar numbers of votes. The only exception would be movies with number of votes greater than 3 standard deviations from the mean, as the range for those votes is quite large (495,000 - 2,000,000). Therefore the data was split into 4 sections -

1. Z-score < =1
2. 1 < Z-score < =2
3. 2 < Z-score < =3
4. 3 < Z-score

Plots of raw counts showed the following:

Fig. 5: Movie counts grouped by IMDB Average Rating of bin size 0.5, stacked by Bechdel test scores. Each bar plot in this figure corresponds to a different Number of Votes z-score.

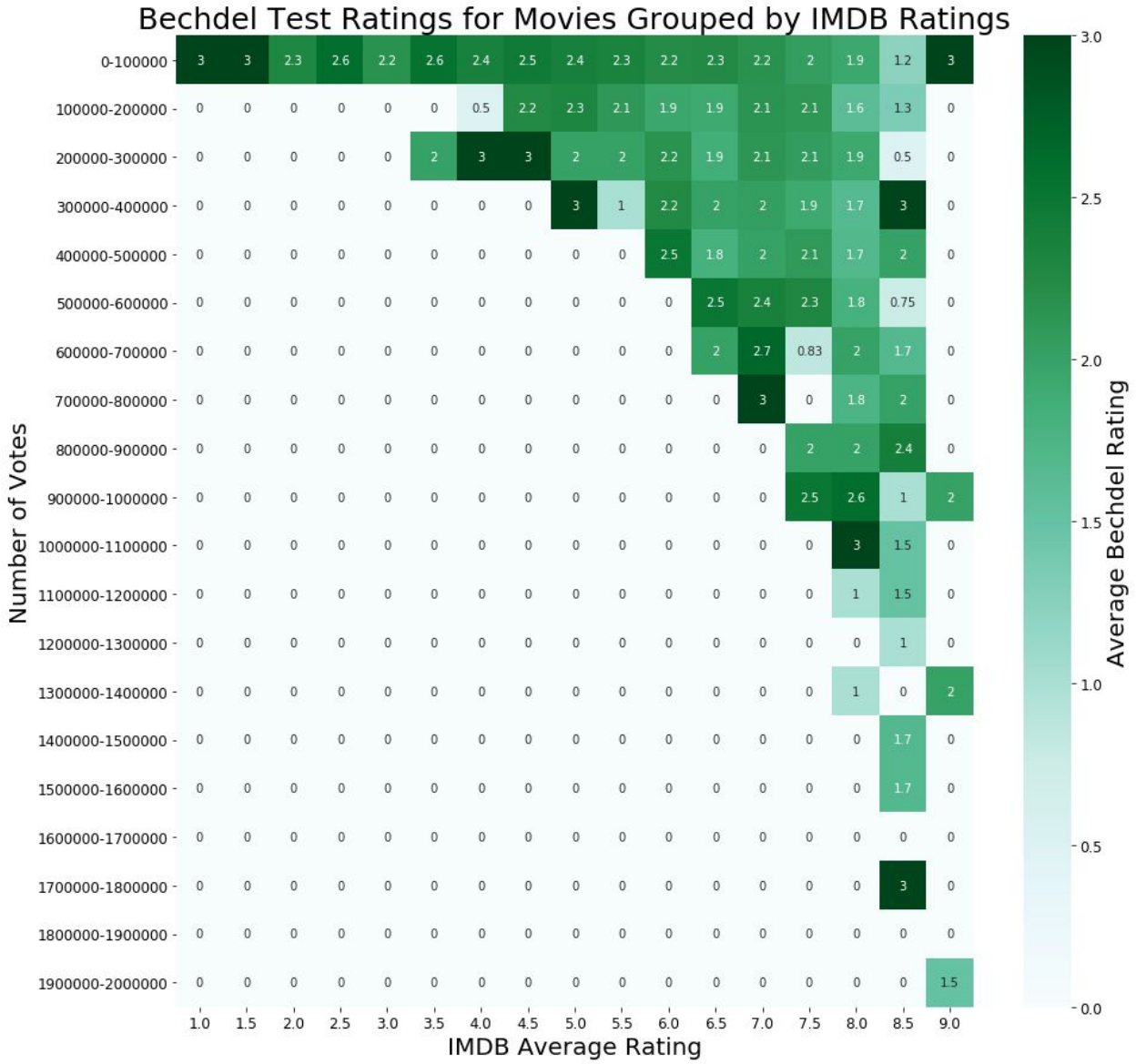


One thing that we can see from these 4 plots is that from a counts perspective, the number of movies that pass the Bechdel test has a maximum at 6.5-7.5 IMDB rating range for number of votes within 1 standard deviation, but the max Bechdel movies pass count moves to 7.0-7.5 IMDB rating range for number of votes between 1 and 2 standard deviations, and then to 7.5-8.0 IMDB rating range for number of votes between 2 and 3 standard deviations. For

very large number of votes, > 3 standard deviations from the mean, the maximum number of movies that pass the bechdel test is has an IMDB rating range of between 8.0-8.5.

To see the effect of the number of votes another way, we decided to group the movies by number of votes bin size of 100,000 and then grouping again by 0.5 IMDB rating range within each 100,000 number of votes ranges. To see if there the combination of IMDB rating and number of votes is related to Bechdel test scores, Bechdel average scores (total Bechdel score/total movies count) was plotted per number of votes range and IMDB rating range via a heat map:

Fig. 6: Heatmap of Bechdel average rating for movies grouped by IMDB Average ratings (x-axis) and number of votes (y-axis)



The number of 0s we see in this plot further highlights the selection bias in the Bechdel data that we had seen evidence of in the other plots. It seems that as the number of votes increases, the Bechdel data is more limited to a decreasing range of IMDB ratings skewed towards increasingly higher IMDB average rating values. For number of votes between 0 to

100,000, there is Bechdel data for movies of all IMDB rating increments, but as the number of votes increases, the Bechdel data is limited to increasingly smaller IMDB ratings range, until we only see data for IMDB ratings between 8.0 and 9.5 for movies with number of votes greater than 1,100,000.

Therefore, the shift that we see in the previous figure of 4 bar charts - where the maximum count of the number of of Bechdel passing movies move from IMDB rating range of 6.5-7.5 for low number of votes to 8.0-8.5 for high number of votes - may not necessary be a true correlation. In other words, more popular films that have a higher number of votes and a high IMDB rating are not guaranteed a higher pass rate for the Bechdel Test. The reason we see the maximum count shift in IMDB rating range is more likely due to the fact that Bechdel data contains fewer movies as the number of votes increases and as skew towards higher IMDB ratings occurs for increasing number of votes, so too will the maximum count, because there is simply no data for movies with lower IMDB ratings and high number of votes.

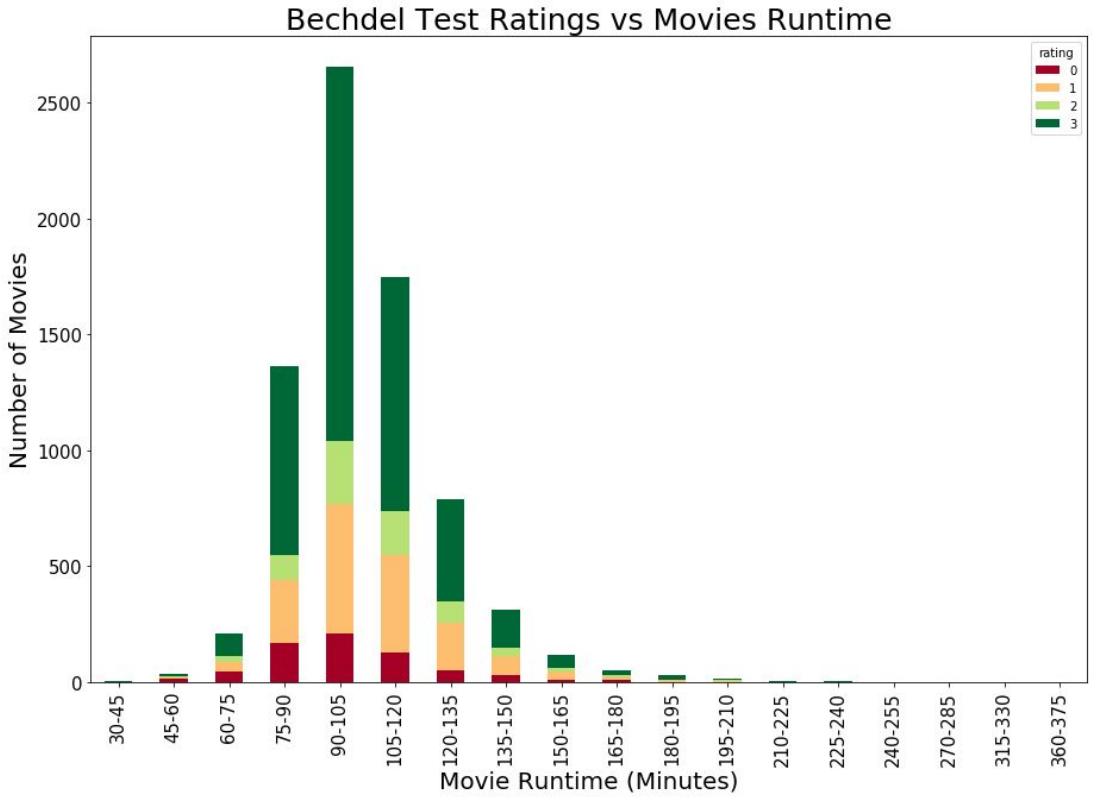
The fact that the Bechdel data becomes more limited as the number of votes increases is a bit surprising. One would think that comparatively, less popular movies would have less votes on IMDB and thus would be less likely to be scored for the Bechdel test. In addition, from the heat map, it seems Bechdel test averages are highest for movies with less than 1.1 million votes. With a few scattered exceptions, we see a little evidence that the Bechdel test averages peak seems to move to lower IMDB ratings as the number of votes decreases. However, to determine if there is true correlation, further statistical analysis will be required in a future phase.

6.3 Does Runtime have any effects on Bechdel ratings?

The basis for this question came from our initial curiosity - if a movie has a longer runtime in number of minutes, is it more likely for the film to pass the Bechdel test, since a longer film would have more opportunities for female characters to speak to each other. For the runtime data, we merged title.basics.tsv with Bechdel test data from https://bechdeltest.com/api/v1/doc using pandas merge on the imdb_id of dataframes created from the two data sets. We then filtered for title type to include movies and TV movies, and also removed adult films. Since we need to see the relationship between runtime and number of movies that pass the Bechdel test, movies with no runtime ('NA' in runtime column) were also removed from our dataframe.

Removing the outlier film the clock (with a runtime of 1440 minutes), IMDB movies with Bechdel ratings have an average runtime of 105 minutes, with a standard deviation of 21 minutes. The interquartile range (25th percentile to 75th percentile) is from 92 to 115 minutes. When we group the data by increments of 15 minute ranges in runtime and plot raw counts stacked by Bechdel rating scores, we see the following:

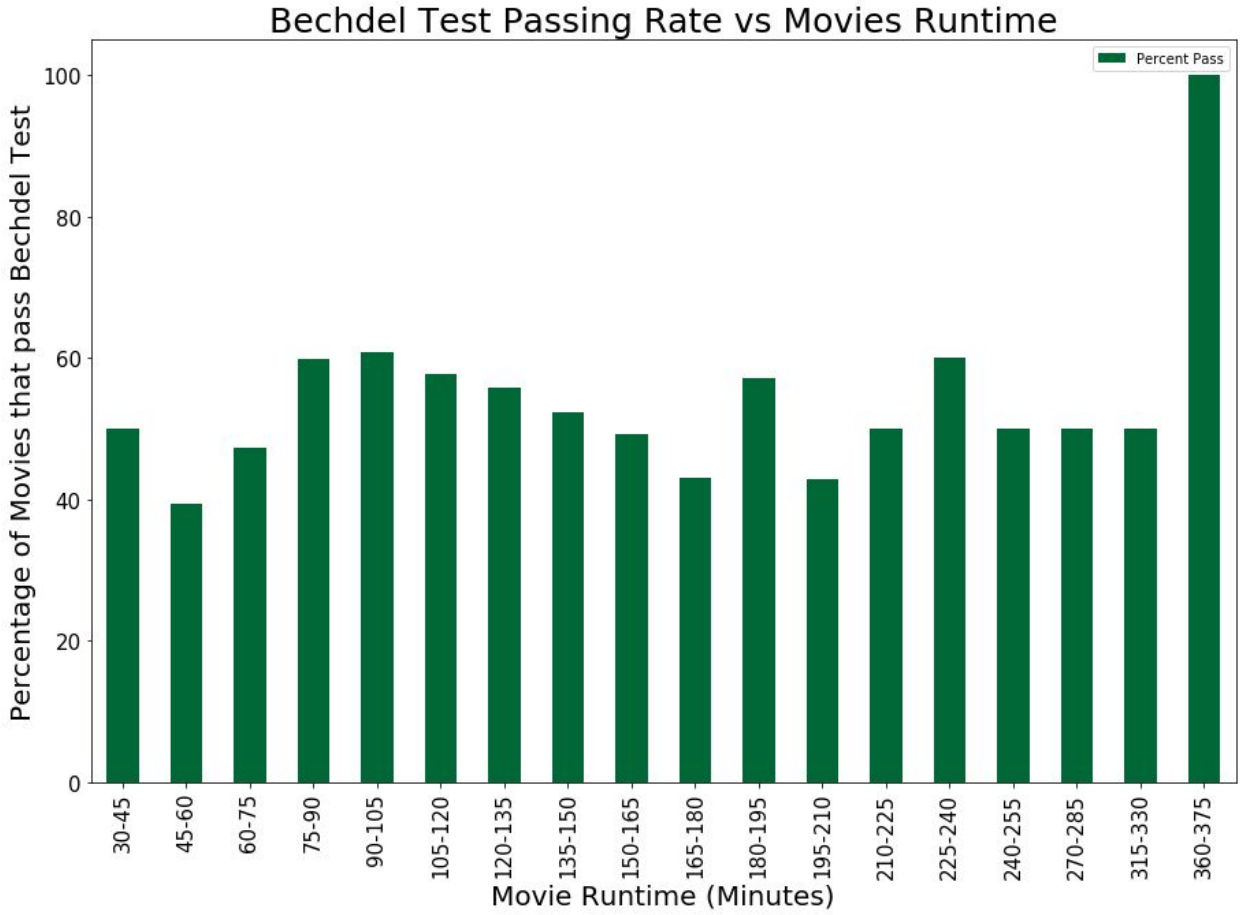
Fig. 7: Movie counts grouped by IMDB Runtime of bin size 15 minutes, stacked by Bechdel test scores.



Based on raw counts alone, the maximum number of movies that pass the Bechdel test is for films that have a runtime of 90-105 minutes. However, this runtime range has has the most number of movies with Bechdel ratings. From this graph alone, one could say that it is difficult to judge if runtime has any effect on Bechdel passing rate, since the proportion of Bechdel movies that score a 3 look roughly the same across the bars for 75 minutes to 135 min.

In fact, when we plot the percentage of movies that score a 3 for the Bechdel test within each Runtime range, we see the following:

Fig. 8: Percentage of movies that pass the Bechdel test with a score of 3, for IMDB movies grouped by runtime of bin size 15 minutes.



The 100% bar at the 360-375 range is a bit misleading as it corresponds to only 1 film that happened to score a 3. Ignoring this and focusing on the rest of the data, we see no clear correlation between the percentage of movies that pass the bechdel test and run time. For almost any runtime, the percentage of bechdel passing films like between 40% and 60%. There is no noticeable peak at the 90-105 runtime minutes range, where the raw counts peak occurred. From this bar chart, we would conclude that longer runtimes do not ensure that films have a higher rate of passing the Bechdel test.

6.4 How many of the the top earning movies pass the Bechdel Test?

In order to answer how many of the top earning movies in terms of worldwide box office gross ranking passed the Bechdel Test, we had to web scrape Box Office Mojo's worldwide gross list from https://www.boxofficemojo.com/alltime/world/ page into a dataframe called "movies_df". Then we added additional column called 'TitleYear', which was a concatenation of 'primaryTitle' column and 'Year' column on both "movies_df" dataframe and "bechdel_titlebasics_ratings" dataframe to serve as primary key for merging the two dataframes. The merged dataframe yielded 597 rows, which meant that 80.03% movies had results for the Bechdel Test from the original 746 movies listed in the Box Office Mojo's worldwide gross list. In order to analyze the 597 movies, we created a new dataframe grouped by 'rating' and counted the number of movies using .groupby() method and .count() method. Then we added "percentage" column to the new dataframe by dividing the number of movies in each group with

the total sum of movies for all groups. As a result we were able to learn that 306 movies out of 597 movies passed the Bechdel Test as shown in Figure 9 below.

Fig. 9: Bechdel Test Results on Box Office Mojo's Worldwide Gross Ranking

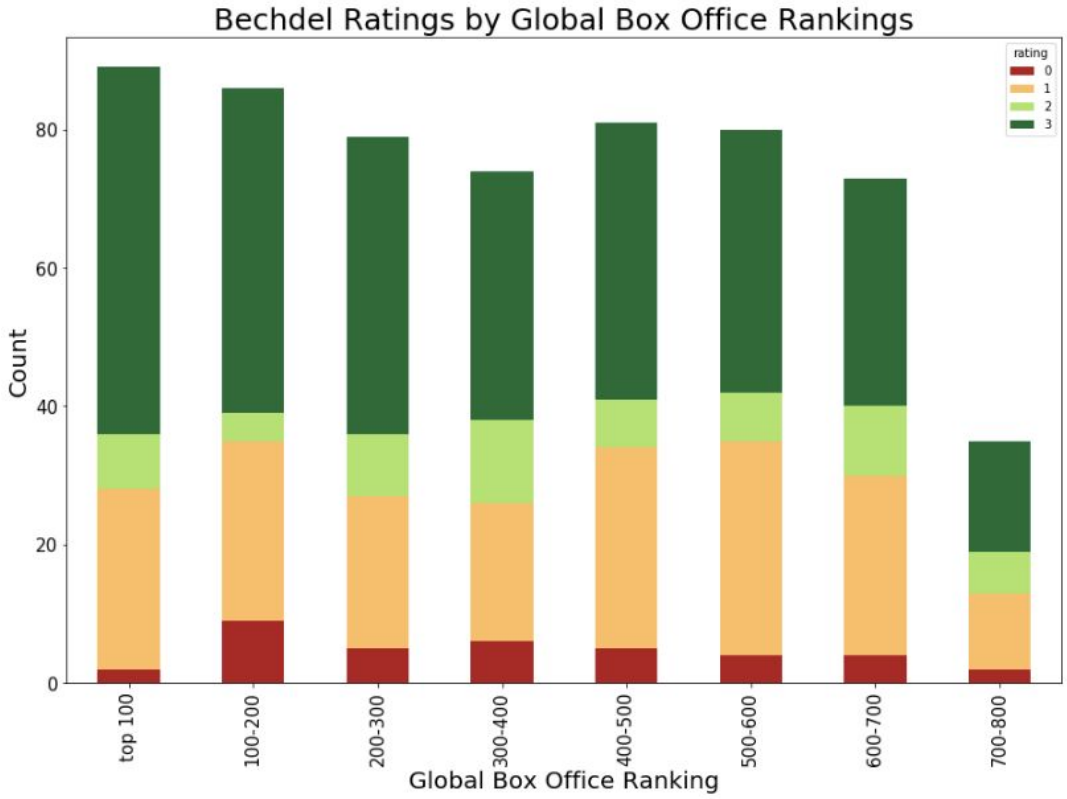
bechdel_rating	movie_count	percentage
0	37	6.20%
1	191	31.99%
2	63	10.55%
3	306	51.26%

For additional insights, we decided to bin Box Office Mojo's worldwide gross ranking by 100 using .cut() method and created a new dataframe grouped by 'bin' and 'rating' to show the number of movies in each bin for each Bechdel Test rating. To provide better perspective, we added 'percentage' column by dividing the number of movies for each bin and rating by the total number of movies in each bin. The resulting dataframe was a bit hard to analyze since the dataframe had 32 indexes as a result of multi-indexing. Therefore, we went ahead and filtered the multi-index dataframe by resetting the index and then applying a filter for the bechdel rating of 3. Figure 10 below represents the filtered dataframe and it is very interesting to note the decreasing percentage of Bechdel Test pass rate as the binned ranking index gets lower. Perhaps there is a possible correlation between the amount of money that a movie generates and the results of the Bechdel Test for the same movie.

Fig. 10: Bechdel Test Pass Rate by Binned Box Office Mojo's Worldwide Gross Ranking

gross_ranking	bechdel	movie_count	percentage
top 100	3	53	59.55%
100-200	3	47	54.65%
200-300	3	43	54.43%
300-400	3	36	48.65%
400-500	3	40	49.38%
500-600	3	38	47.50%
600-700	3	33	45.21%
700-800	3	16	45.71%

Fig. 11: Stacked Bar Chart for Bechdel Ratings by Global Box Office Rankings



6.5 Since women's roles have changed over time, does the number of films that pass the Bechdel Test increase over time?

Movies often reflect the current state of society. In the 1880s, when cinema was first established in the United States, women didn't even have the right to vote yet. Along those lines, early female representation in film was not very empowering. As women's roles changed in society, female representation on screen is likely to reflect that. We examined this in the Bechdel Test dataset by looking at the movie release years binned into release decades as a categorical variable.

Fig. 12: The raw counts of Bechdel Test ratings for movies binned by release decade show that there are far more ratings for films in the last two decades.

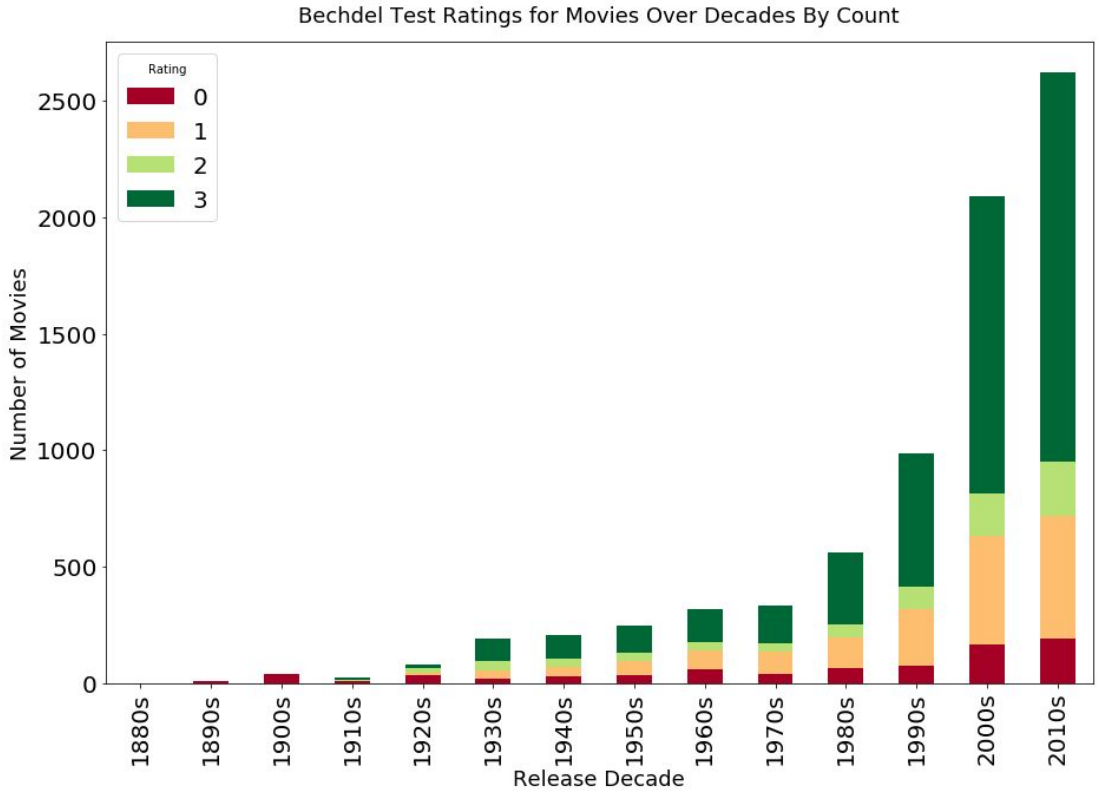
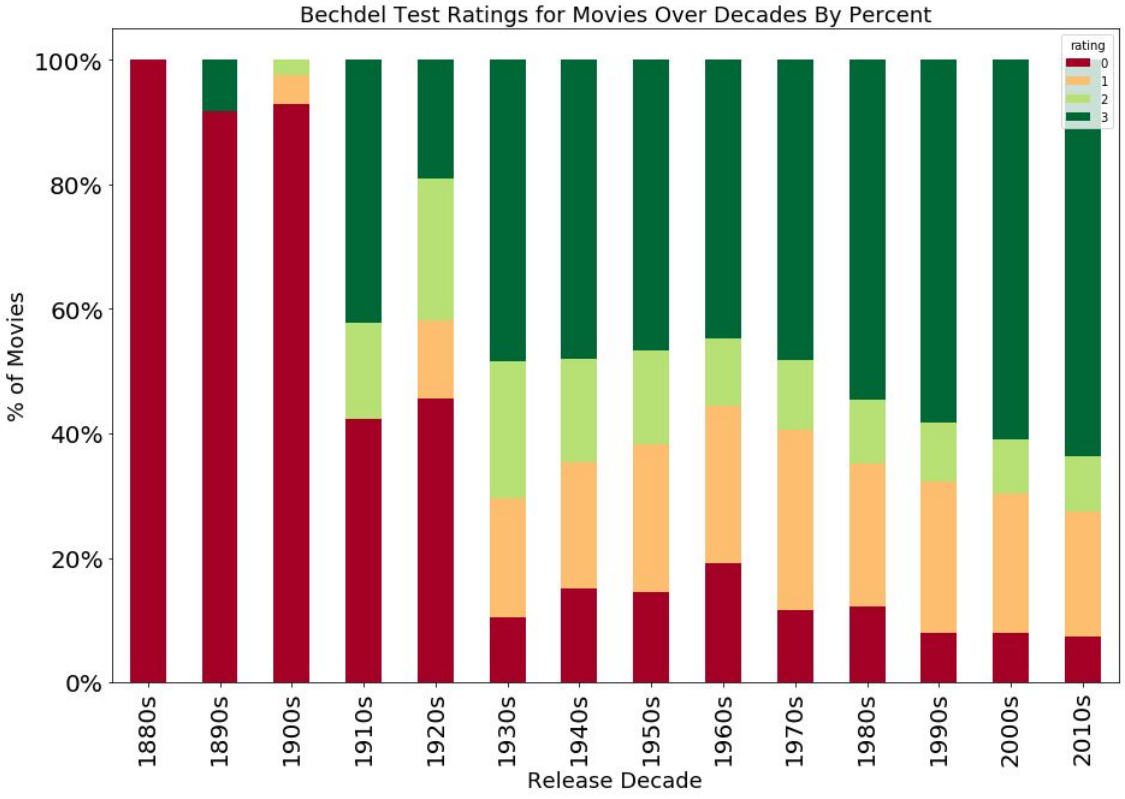


Fig. 13: This percentage breakdown of Bechdel Test ratings for movies binned by release decade allows us to compare the pass/fail rate for each decade.



From the percentage breakdown above, starting in the 1930s, we can see that the number of movies that passed the Bechdel Test eclipsed the number of movies that failed, which coincides with the rise of the Women's Movement in the 1920s. This trend continues in the decades after the 1930s. In the decades following, a little less than 50% of movies pass the Bechdel Test until the 1980s where it increased to 55%. In the current decade, 64% of movies released so far passed the Bechdel Test.

We also noted that in the decades after the 1930s, of the movies that do not pass the Bechdel Test, more movies have a 1 rating (has at least two named female characters) than a 2 rating (those characters talk to each other).

Fig. 14: Table comparing proportions of movies by release decade for IMDB and Bechdel Test data

Release Decade	IMDB Movie Count	IMDB % of Total Movies	Bechdel Test Movie Count	Bechdel Test % Total Movies
1890s	57	0.01%	12	0.16%
1900s	332	0.05%	42	0.54%
1910s	15615	2.53%	26	0.34%
1920s	19460	3.15%	79	1.02%
1930s	19210	3.11%	190	2.46%
1940s	15068	2.44%	206	2.67%
1950s	26355	4.27%	248	3.21%
1960s	41960	6.79%	319	4.14%
1970s	48929	7.92%	336	4.36%
1980s	54748	8.87%	561	7.27%
1990s	61458	9.95%	989	12.82%
2000s	110488	17.89%	2087	27.05%
2010s	155367	25.16%	2619	33.95%
No Release Year	48515	7.86%	0	0.00%
Totals	617562	-	7714	-

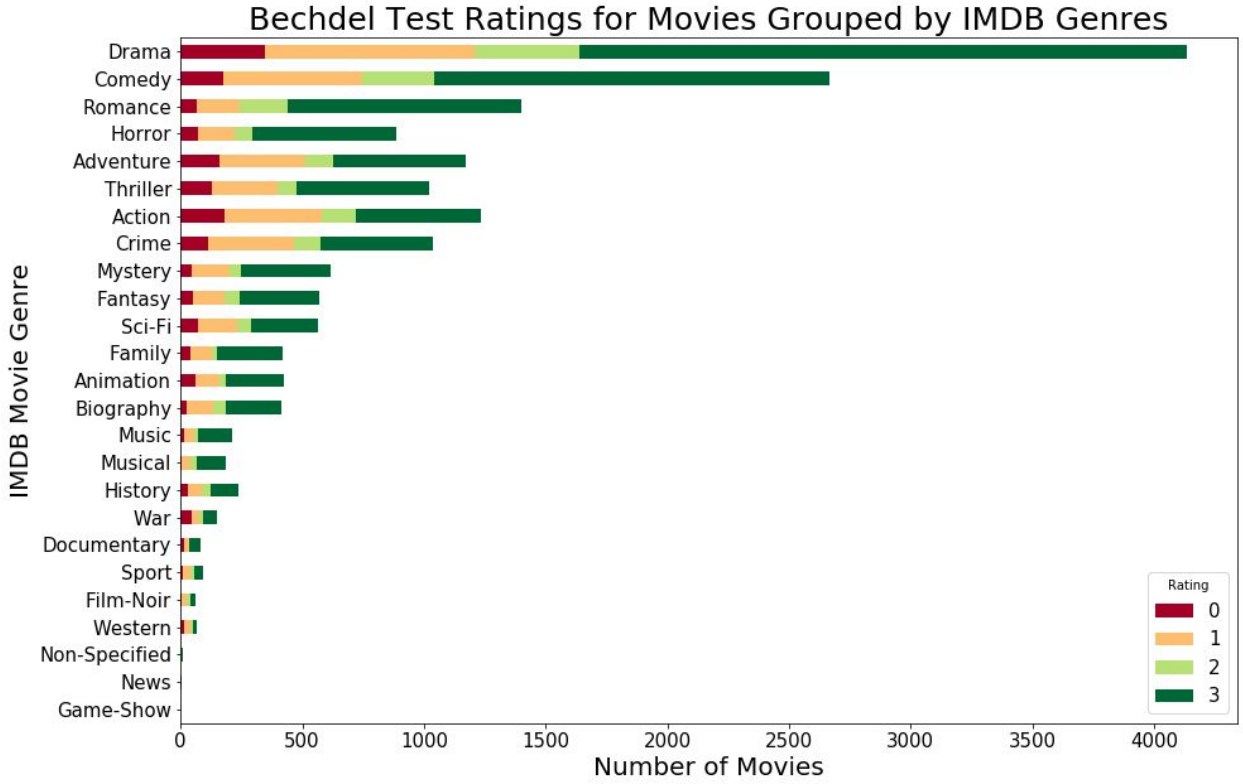
As shown from the Bechdel Test ratings breakdown using raw counts, the last two decades have far more ratings than previous decades. This could be due to more films being released during that time period, but it's likely that the bigger factor is nature of user-generated data as seen in the table above. For example, movies released in in the 2010s make up 34% of the Bechdel Test dataset compared to 25% of the IMDB dataset. It could be that more Bechdel Test users/raters have seen recently released films versus older films, which skews the data. Ideally, even if we couldn't get all the ratings of movies released (617,562 movies according to IMDB at the time of analysis), it would be best to have a random sample of Bechdel Test ratings that is more proportionate to the number of films released in each decade.

6.6 What genres have the most movies that pass the Bechdel Test?

In the raw file title.basics.tsv, every movies has a list of 1 or more genres in the genres column. For the analysis of this question, a pandas dataframe was created from reading the title.basics.tsv raw file, and a second pandas dataframe was created from the Bechdel test data extracted from this source - https://bechdeltest.com/api/v1/doc . The dataframes were merged using the imdb_id, and only movies with a bechdel rating were included in the merged dataframe. The movies were further filtered to include title type of movies and TV movies, and Adult films were excluded.

As discussed in 5.3, the method of analysis we decided upon was to count every movie once for each of the genres listed in the genres column, which means that movies containing multiple genres are counted multiple times. This is a limitation with the genres list, as no weights are assigned to genres that are listed for each particular film. To gather counts by genre, the unique genre set was first created, which shows 25 distinct genres. A value of 'NA' in the genre column was then grouped as 'Non-Specified'. The data was transformed using the looped vectorization of a function that created a column for each genre, then entered a value of 1 or 0 in for each row in the column depending on if the film's genre list contained a particular genre. The resulting dataframe was then grouped by by Bechdel scores, thereby counting the number of films per genre grouped by each Bechdel test score. The plot, sorted by greatest number of films that score a 3 on the Bechdel test is below:

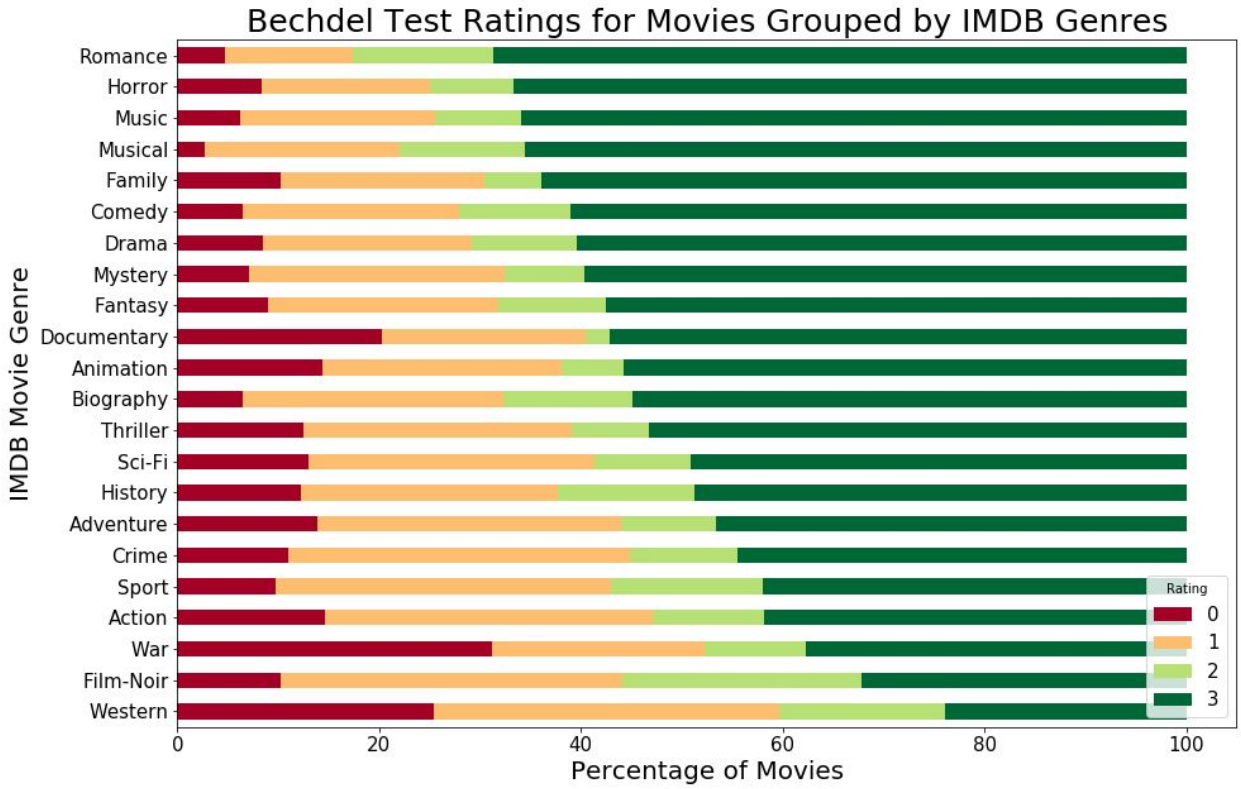
Fig. 15: Movie counts grouped genres, stacked by Bechdel test scores, sorted in descending order of greatest number of movies with Bechdel Rating 3



This figure shows, perhaps as expected, that the genres with the most movies scoring a 3 for the Bechdel test are Drama, Comedy, and Romance. The genres Horror, Adventure, Thriller, Action and Crime all have similar movie raw counts for passing the Bechdel test. To

analyze this data further, since raw movie counts make comparison a little difficult, the data was transformed so that a percentage component bar chart of the same data could be created:

Fig. 16: Percentage of movies with each Bechdel score, grouped by Genres, sorted in descending order of highest percentage of Bechdel 3 ratings.



First, it is important to note that the genres for this chart, unlike the raw counts bar chart, lists the genres in descending order of percentage of movies that score a 3. The data for three genres - Game-Show, News and Non-Specified - were removed for the percentage bar chart creation because the low amount of data in these genres appeared misleading on the percentage bar plot. For instance, Game-Show genre had only one film count with a Bechdel score of 3, and would have show 100% on the percent bar chart and making it the most likely Bechdel passing genre, which is unrealistic.

In this plot, Romance leads as the genre with the highest percentage of Bechdel test passing films, followed by Horror and Music genres. Drama genre, which had the highest count in the previous bar chart, is 7th in highest percentages of Bechdel passing films. For the genres starting from Thriller and going down the list to Western, the percentage of films that pass the Bechdel test is less than 50%. This means that for these lower ranked genres, there are more movies that fail the Bechdel test than movies that pass the Bechdel test. The genres that we are not surprised to see in the lower ranks would be War, Action and Adventure. Since War films are historically dominated by a male cast, the War genre, as expected, has the highest percentage of films with a Bechdel score of 0, meaning 1 or 0 women were cast in these films. Many of these lower ranked genres like Adventure, Action and Film Noir actually do cast multiple female characters, leading to a lower percentage of films scoring a Bechdel 0 rating. However, they likely fail the Bechdel test because while the female characters exist for many of these films, it is less likely that the female characters will talk to each other.

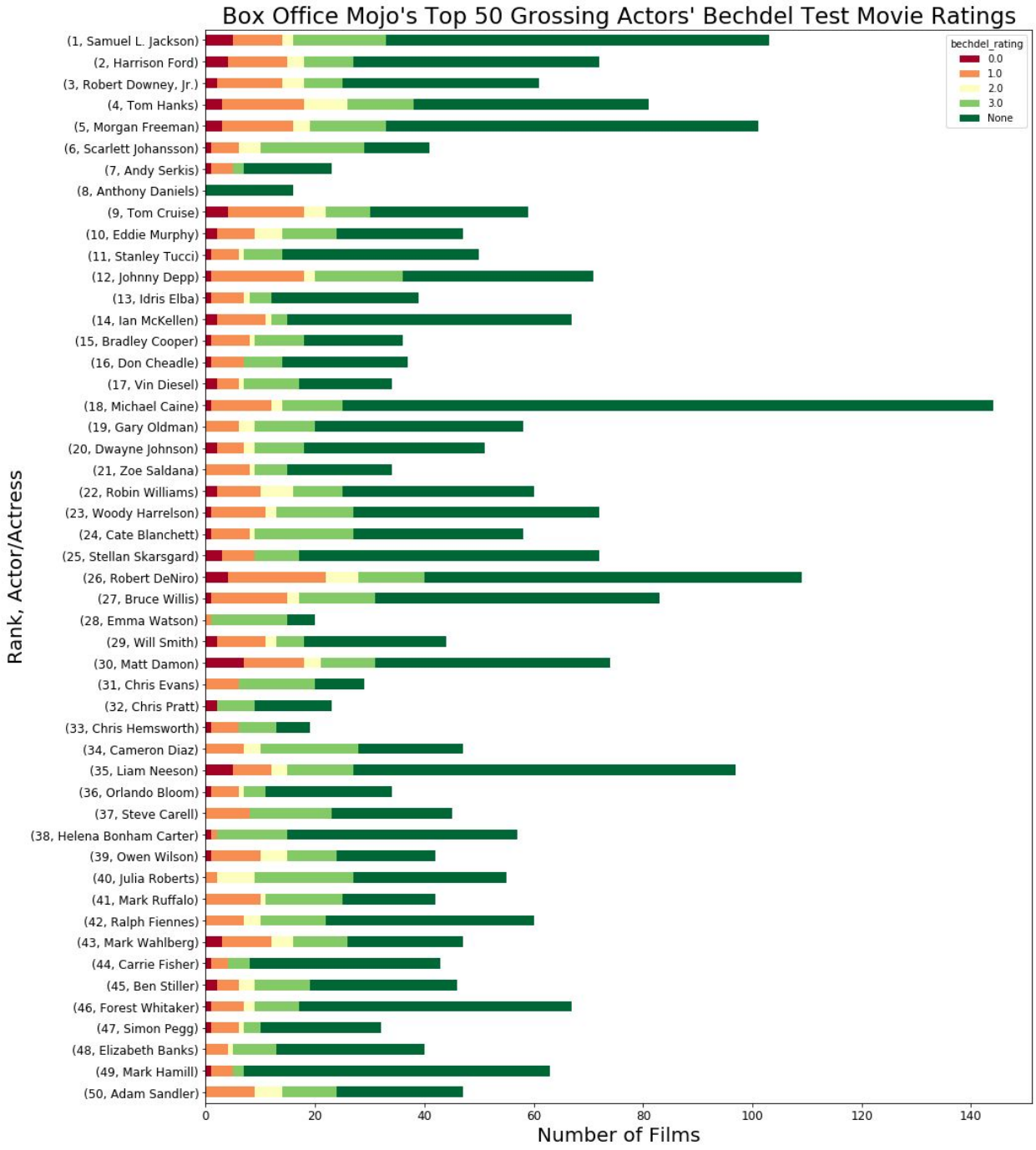
For almost all of the genres with the exception of Romance, the percentage of films that score a 1 is much higher than the percentage of films that score a 2. This would mean that of the movies that fail the Bechdel test, there are usually more movies that cast multiple female characters than movies where the female characters will interact. This phenomenon is pretty much independent of genre, except the Romance genre.

Furthermore, if we look at all genres for Bechdel scores of 2 and 3 movies, we see that the percentage movies that score a 2 is much lower than the percentage of movies that score a 3. This would mean that of movies where female characters do interact, there are usually more movies where the 2 female characters will talk about a topic other than a man, than movies where female characters exclusively talk about a man. This phenomenon also appears to be independent of genre.

6.7 Are highest grossing actors and actresses in movies that pass the Bechdel Test?

The list of top 50 highest grossing actors and actress were scraped from this site - https://www.boxofficemojo.com/people/?view=Actor&sort=sumgross&p=.htm. The actors are ranked by the total grossing amount of all of the films they appeared in. Using the actors' names, we retrieved the nconst from the IMDB file name.basics.tsv for these actors. Using the filtered IMDB file title.basics.tsv (filtered by SQL to include only actors), we then retrieved all of the imdb_ids for films that starred the top 50 highest grossing actors and actress using the nconst id for each actor. This combined data was read into a pandas dataframe. A second pandas dataframe was created that merged the data for title.basics.tsv with the Bechdel test data from our source https://bechdeltest.com/api/v1/doc. These two dataframes were again merge to create one dataframe of all of the top 50 highest grossing actors and actress, their films, and the Bechdel test rating for some of their films, if applicable. Grouping by actors, the number of films stacked by Bechdel rating appears in the barchart below:

Fig. 17: Number of movies starring each Actor/Actress, stacked by Bechdel test rating.



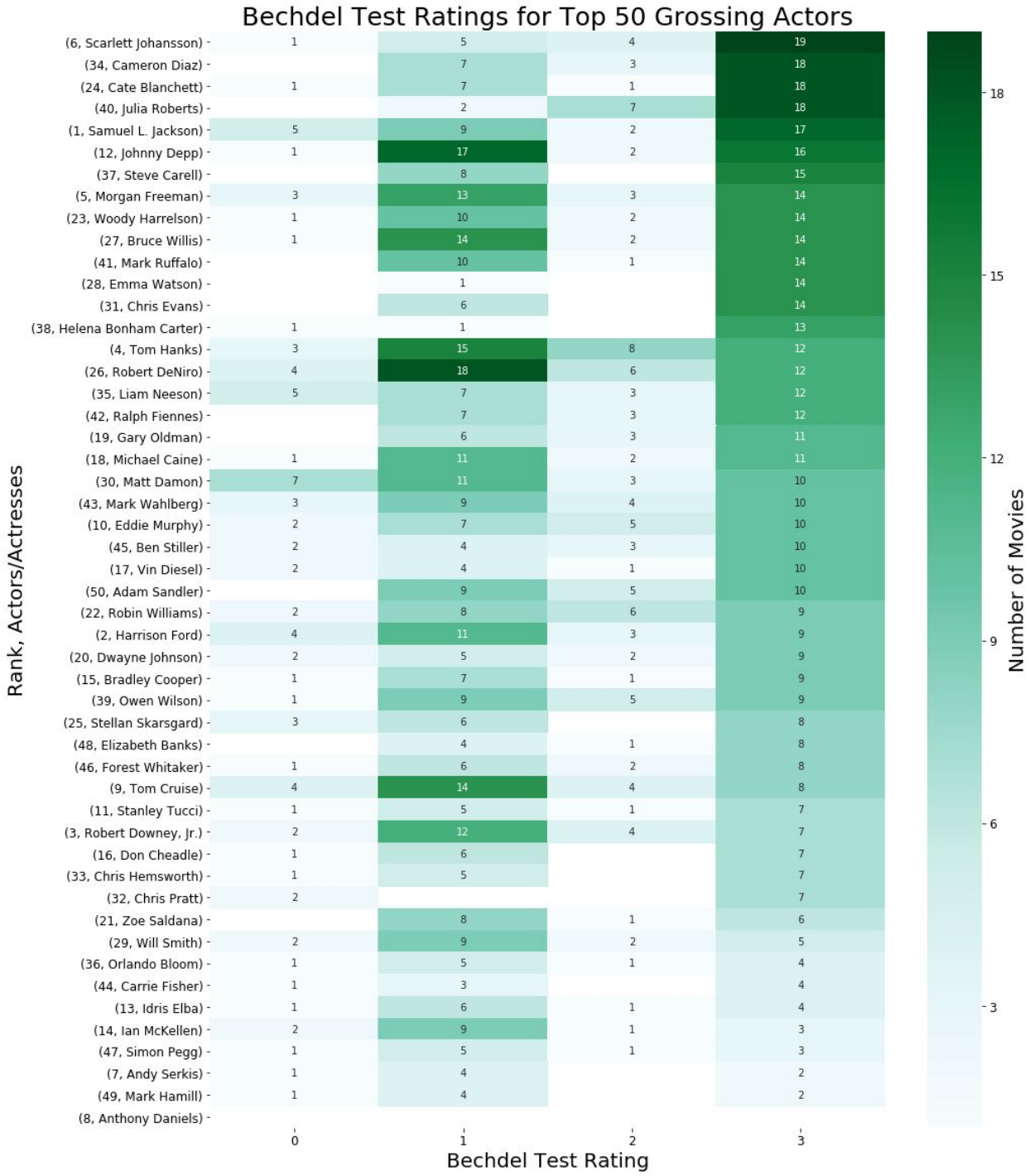
Here, movies are counted once per actor or actress, as long as the actor or actress appears in the film. If a movie stars multiple actors and actress from the top 50 highest grossing, then the movie is counted multiple times in this bar chart. The logic is similar to what was used for the genres analysis, since no weight is given to actors for their participation a film. The number of films that pass the Bechdel test appears to have no correlation with the actor/actress ranking, as each actor or actress likely has different personal criteria for participation in movies.

Actually the lack of Bechdel test data for all of the actors are quite alarming as seen by the stack bar that categorizes movies for each actor with Bechdel rating of ‘None’ meaning that

their movies from IMDB were not rated by the Bechdel data source site. If we summarize the percentage of films for each actor or actress that has no Bechdel scores, the average is 59.6% with a standard deviation of 15.2% and an interquartile range of 49.5% to 68.6%. Therefore, approximately 75% of the actors/actresses in the top 50 highest grossing list have more than 50% of their movies not rated by Bechdel test. Therefore, with the data we do have it is very difficult to find an accurate comparison of which actors/actresses prefer to or have starred in the highest number of bechdel test passing films.

Even so, we wanted to look more closely at the movies for each actor that have been rated by the Bechdel test data. Since the number of movies that do not have bechdel test scores are so overwhelming, they were removed to plot the following heat map, which shows the number of films for each actor grouped by Bechdel scores, sorted vertically in descending order:

Fig. 18: Heatmap of movies starring each Actor/Actress, grouped by Bechdel test ratings.



Here, it is further confirmed that there is no correlation between actors/actress rank and the number of films that pass the Bechdel test. We see that the top people to start in the highest number of Bechdel rating 3 films are all women. However, the rest of the list ranked by the number of Bechdel 3 movies is neither gender based nor grossing rank based. For many actors that have low number of movies that pass the bechdel test, they also have low number of films that are Bechdel rated. The exceptions to this are Tom Cruise, Stanley Tucci, Ian

McKellen, who have nearly twice as many movies with a Bechdel score of 1 than movies with a Bechdel score of 3.

Another phenomenon we see in this heatmap is the same one cited in our analysis of question 6.6 - that there are more movies that score a 1 on the Bechdel test for each actor/actress than movies that score a 2. The only exception to this is Julia Roberts' films, but most actresses (and actors) are not exempt from this rule. This would mean that for many of the actress in the top 50 highest grossing actors list, their films that *fail* the Bechdel test do so because they did not interact with another female character. However, for almost all of the films in this plot, there are more movies that score a 3 for the Bechdel test than a 2, for each actor/actress. Again this speaks to our previous observation, that in the sample of movies where female characters do interact with each other, it is more likely for 2 women to discuss a topic that does not pertain to a man than to only talk about a man.

6.8 Are the current highest paid actresses' movies passing the Bechdel Test?

Given that the highest grossing actors list was primarily male, we wanted to probe further to see how many of the movies that the highest paid actresses of 2017 are in pass the Bechdel Test. Being among the highest paid actresses in the film industry means that they're likely to be offered a variety of roles, specifically in movies that have high box office earning potential and become popular with audiences.

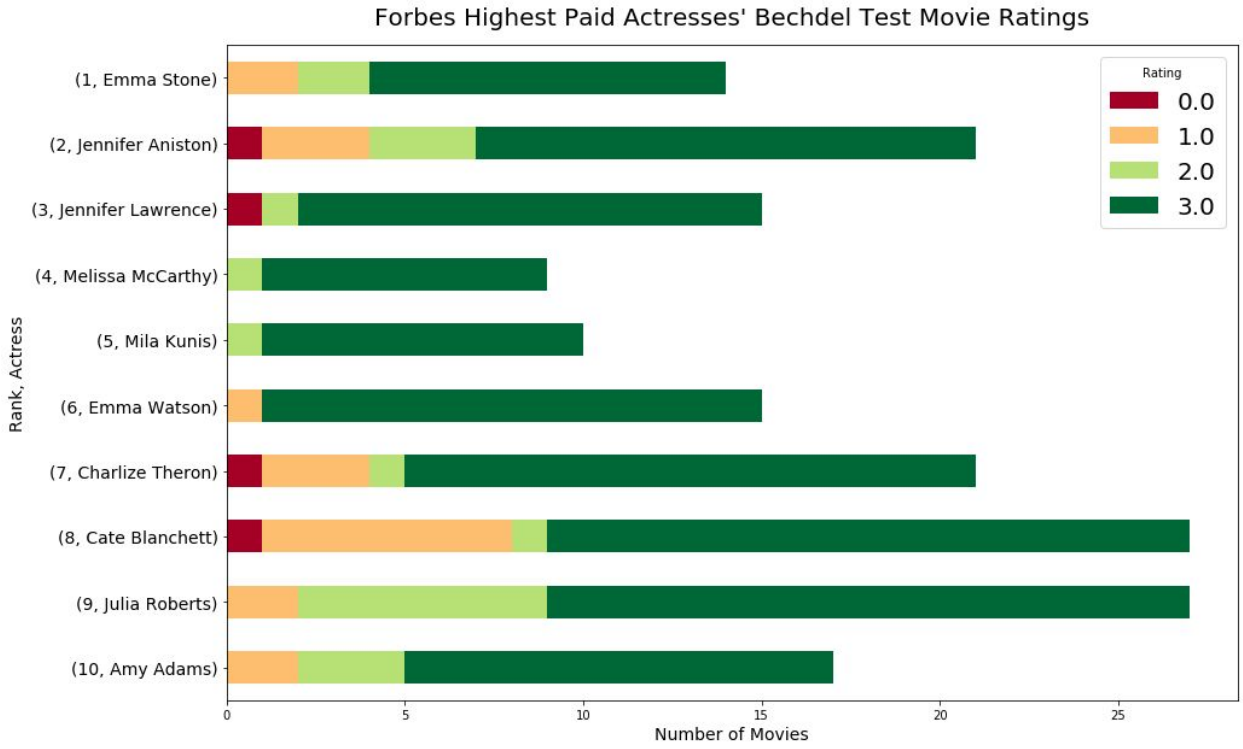
This analysis utilizes the Bechdel Test data, a subset of IMDB title.basics that contains all the movies the actresses have ever been in, IMDB name.basics, and the Forbes list of highest paid actresses in 2017.

Prior to analyzing the Bechdel Test ratings for their movies, we did a comparison to see how many movies were linked to them in the IMDB datasets versus how many of their movies we had Bechdel Test ratings for by joining. We found that we only had ratings for 47% of all the actresses' movies, which affects our ability to draw a conclusion from the analysis. Because of this, we ended up creating two groupby dataframes: one with NaN values and another with NaN values pre-filled with 'None'. The latter dataframe was created because when we did group by rating, NaN values were not included. Despite the missing data, we are still able to gain some insight based on the movies that do have ratings.

Fig. 19: Forbes Top 10 Highest Paid Actresses in 2017

primaryname	Rank	Salary_Millions
Emma Stone	1	$26
Jennifer Aniston	2	$25.50
Jennifer Lawrence	3	$24
Melissa McCarthy	4	$18
Mila Kunis	5	$15.50
Emma Watson	6	$14
Charlize Theron	7	$14
Cate Blanchett	8	$12
Julia Roberts	9	$12
Amy Adams	10	$11.50

Fig. 20: Raw count breakdown of each actress's rated movies by Bechdel Test Score



The chart above shows the breakdown of how many of their movies pass and fail the Bechdel Test. In general, all the highest paid actresses have more movies that pass the Bechdel Test versus fail. 75% out of the 176 movies we have ratings for passed the Bechdel Test with a score of 3. Jennifer Aniston, Cate Blanchett, and Julia Roberts entered the industry earlier and thus, they have been in more films so this could explain why their pass percentage (67%) is slightly lower than that of other actresses. Only Jennifer Aniston, Jennifer Lawrence, Charlize Theron, and Cate Blanchett have been in movies that outright fail the Bechdel Test by not having at least two named female characters. Our analysis may imply that these actresses aren't often in films as the token female character since so few of their films have a rating of 1.

Unfortunately, as mentioned previously, the amount of movies that don't have ratings affects our analysis of the highest paid actresses' films and it makes it difficult to draw a conclusion. To approach it from a different angle, we utilized the 'knownfortitles' column in the IMDB names.basics table, which contains 4 movies that the individual is known for. It's unclear how IMDB determines these movies from the documentation. This column has the 4 movies in an array according to their imdb identifier. To get to these values, we converted the column into a list to extract each individual movie. After this process was completed, we turned the list into a dataframe. When joined with the Bechdel Test dataset, 2 out of these 40 movies were missing ratings. 36 out of the 38 movies (95%) passed the Bechdel Test, which is meaningful because these are their most well known, popular movies.

7. Conclusion

Our analysis of these datasets included several key findings. Examining the number of movies that passed and failed the Bechdel Test over time showed that, indeed, the number of movies that pass has increased. Of the movies that fail, more of them have at least two named female characters (1 rating), but they do not talk or interact with each other. In the latter three decades, over half of the films release pass the Bechdel Test.

Top genres by raw counts are Drama, Comedy and Romance. Runtime does not appear to have any effect on the percentage of movies that pass the Bechdel Test. Percentage of Bechdel test passing films does not appear to increase as IMDB ratings increase, if anything the opposite is true for movies with IMDB ratings between 4.5 and 9. If we take the number of IMDB votes into consideration when examining the IMDB ratings, it appears that as the number of votes increases, the amount of Bechdel test data for films decreases and any correlation between Bechdel score and IMDB rating becomes increasing less likely to detect.

When looking at the number of of movies that pass the Bechdel test for the top 50 highest grossing actors according to Box Office Mojo, there appears to be no correlation between actor ranking and Bechdel passing films. It is important to keep in mind that on average, nearly 60% of the top 50 actors' films do not have any Bechdel test data. Given that the highest grossing actors is primarily male, we also looked at the top 10 highest paid actresses in 2017. With the amount of movie ratings available, we found 75% of their films passed the Bechdel Test.

Other key findings include our analysis of the movies that Forbes' list of highest paid actresses' were featured in and the movies that were listed in the Box Office Mojo's worldwide gross ranking list. From the analysis of movies that Forbes' list of highest paid actresses' were featured in, we were able to learn that 75% of the movies passed the Bechdel Test. From the Bechdel Test analysis of Box Office Mojo's worldwide gross ranking, we were able to detect a possible correlation between the amount of money that a movie generates and the passing result of the Bechdel Test for the same movie.

To highlight some limitations, we have to note that once you pass the first 2 rules of the Bechdel test, it is very likely that the movie will pass the third rule of the Bechdel Test. This is due to the fact that conversation scope of the two female characters will likely involve topics other than man and the fact that this criteria is based off of interpretation by any individual. Additionally, we have to acknowledge that Bechdel Test is crowdsourced and does not cover every single movies that have been produced to date.

8. Future Directions & Recommendations

Bechdel Test website is a great way for film enthusiasts to contribute to an evolving database. More standardization can improve the reliability of the data. For example, they could restructure it so people can vote for what Bechdel Test score they think a movie has and reconcile disputes that way. Additionally, doing Bechdel Test ratings in a more “academic”, controlled sense may be more beneficial and reliable. We could have people trained in the Bechdel Test and watch assigned movies. This can help remove user/rater bias and allow us to do better sampling of all released films.

As mentioned, the Bechdel Test is brilliant in its simplicity, but this also contributes to its flaws. From our analysis, we found numerous films that passed the test, such as *Pulp Fiction*, but remain questionable in terms of quality, nuanced portrayals of female characters. Other proposed alternative tests include specifically looking at meaningful portrayals of African American or Latina women, how female characters factor into the narrative, or how many one-scene roles belong to women in a movie.

Finally, outside of on-screen characters and portrayals, we also believe that it is important to consider examining diversity in all aspects of the film industry. Movie production requires not only actors and actresses, but also staff members and executives behind the scenes. For example, it is well known that women are underrepresented as directors and studio executives. This can have a great influence on the quality of movies and by extension, what kind of female characters we end up seeing.

9. References

Liao, S. (2017, December 22). We can do better than The Bechdel Test. Retrieved from https://www.theverge.com/2017/12/22/16807424/alternatives-bechdel-test-bad-moms-lena-waithe

10. Link to Data

https://berkeley.app.box.com/folder/51754956175