SI 618 Project 1 Report: Movie Characteristics and Bechdel Test

Lea Wei

October 24, 2020

1 Motivation

It's not a secret that there exists an inherent gender bias in the movie business. Female actors usually have less income than male actors. There are fewer female protagonists in movies, and female characters are usually lack of serious development and depth compared to their male counterparts. How to quantify this gender bias is one of the topics of interest. Inspired by an article by FiveThirtyEight about the relationship between female prominence in movies, evaluated by the Bechdel test, and movie budget and box office, I decided to explore this topic further in this project. Specifically, I was wondering what kind of movie will pass the Bechdel test. Thus, in this project, I examined and discussed the relationship between a set of movie characteristics with passing the Bechdel test, including release decade, country of production, movie genre, crew gender, IMDb rating, budget, domestic and international box office and return of investment (ROI).

* Bechdel test: promoted by cartoonist Alison Bechdel, this test is a measure of women representation in movies. There are three criteria: 1) there must be at least two named female characters in the movie who 2) talk to each other about 3) something besides men.

2 Data Sources

In this project, I used 5 datasets (files). One dataset is retrieved from an online API and is given in JSON format. The other four are in CSV format.

2.1 Bechdel movie dataset from BechdelTest.com

The API locates at http://bechdeltest.com/api/v1/doc%23getAllMovies. I retrieved the entire dataset with information of 8076 movies released from 1880 to 2020. There are 5 columns in this dataset:

- **imdbid**: the IMDb ID of the movie
- **title**: movie title
- **year**: release year (according to IMDb)
- **rating**: how many Bechdel test criteria the movie satisfies. 0: no two women; 1: the women don't talk to each other; 2: they talk about men; 3: passes the test
- **id**: a unique BechdelTest.com ID

2.2 Boxofficemojo dataset from Kaggle

The URL link is https://www.kaggle.com/igorkirko/wwwboxofficemojocom-movies-with-budget-listed?select=Mojo_budget_update.csv. I used the **Mojo_budget_update.csv**. There are 3239 movies (entries of data) released in the U.S. between 1990 and April 2020 in this dataset. The movies are not necessarily U.S made. However, among all 26 columns, the important variables for this project are:

- **movie_id**: IMDb ID of the movie
- **title**: movie title
- **budget**: movie budget in US dollars
- **domestic**: domestic box office in US dollars
- **international**: international box office in US dollars
- **worldwide**: worldwide box office in US dollars

2.3 IMDb movies extensive dataset from Kaggle

The URL link is https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset. There are 5 CSV files in this dataset. I used two of them: **IMDb movies.csv** and **IMDb ratings.csv**.

There are 85,855 entries of data in the **IMDB movies.csv** including movie information from 1894 to 2020 and I retrieved all of them. Among the 22 columns, the important columns are:

- **imdb_title_id**: IMDb ID of the movie
- **title**: movie title
- **year**: release year
- **genre**: movie genre

- **country**: country of production
- **language**: movie language
- **director**: director name
- **writer**: writer name
- **avg_vote**: total weighted average IMDb rating

The **IMDb ratings.csv** contains 85,855 entries of data and 49 columns. The information of interest is:

- **imdb_title_id**: IMDb ID of the movie
- **males_allages_avg_vote**: average rating of all male users
- **females_allages_avg_vote**: average rating of all female users

2.4 CPI data from the Bureau of Labor Statistics at the U.S. Department of Labor

This dataset is available at https://data.bls.gov/pdq/SurveyOutputServlet and contains month and annual CPI information for all urban consumers from 1913 to 2020. The data is not seasonally adjusted. I retrieved the month and annual CPI figures from 1950 to 2020 for later use. The dataset was originally in xlsx format, and I saved it as CSV format.

* If you encounter difficulty in finding the dataset using the link above, please follow the steps: go to Consumer Price Index (CPI) Databases -> click "One Screen" button in the "All Urban Consumers (Current Series) database" -> in the dropdown menu, select "US city average", "All items" and "Not seasonally adjusted", click "Get Data" -> select "1950" as start year and "2020" as end year, click "include annual averages" and "GO"

3 Data Manipulation

In this section, I will display how I broke my project into pieces and what tool(s) I used in each stage. Below is an overview of the workflow.

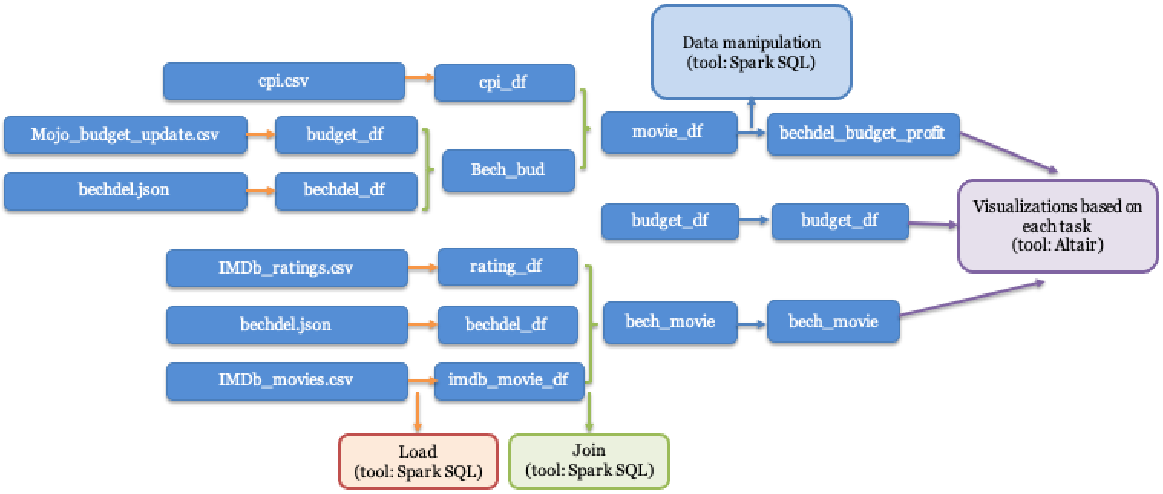


Figure 1: Workflow of the project

3.1 Load and join the datasets

I intended to complete all data cleaning and preprocessing with Spark SQL, so my first step of data manipulation was loading and joining the datasets. After loading the 5 datasets with `sqlContext.read.csv()` and `sqlContext.read.json()`, I conducted some data cleaning to prepare my data frames for joining. I first removed the "tt" at the beginning of the "movie_id" column in Mojo_budget_update data frame (budget_df), and the "imdb_title_id" columns in IMDb_ratings and IMDb_movies data frames (rating_df, imdb_movie_df). I generated new columns "imdb_id" with the clean version of ID in all three data frames above. After that, I converted the type of "budget", "domestic", "international" and "worldwide" columns in budget_df from string to float using `cast()`. After registering the data frames as tables, I created two new data frames by joining the data frames on "imdb_id". Below are some details about the two newly created data frames:

- **bech_bud**: Using Spark SQL language, I joined budget_df and bechdel_df on "imdb_id" and ordered by year. This data frame contains information on movie budget, domestic and international box office, title, and the Bechdel test rating
- **bech_movie**: I joined bechdel_df, imdb_movie_df and rating_df on "imdb_id" and obtained bech_movie data frame, which includes information about movie title, release year, production country, genre, director, writer, total weighted IMDb rating, IMDb rating of male and female voters, and the Bechdel test rating

Now I would like to give more details in how I manipulated these two data frames.

3.2 Bech_bud data frame

This data frame contains 2139 movies released between 1990 and 2020. I first adjusted the budget and box office of each movie for inflation. I used the year 2019 as base year. The formula of adjusting for inflation is

$$adjusted\ value = \frac{unadjusted\ value * current\ cpi}{old\ cpi}$$

Therefore,

$$adjusted\ budget = \frac{unadjusted\ budget * cpi\ 2019}{cpi\ of\ the\ release\ year}$$

$$adjusted\ box\ office = \frac{unadjusted\ box\ office * cpi\ 2019}{cpi\ of\ the\ release\ year}$$

Thus, in order to obtain the CPI information for each release year, I joined the bech_bud data frame with cpi_df on the common column "year" and created a new data frame movie_df. I calculated the adjusted values for budget and domestic and international box office using the formula above and generated three new columns "adj_budget", "adj_domgross" and "adj_intgross" correspondingly.

Budget and box office are absolute values. A movie with extremely large box office isn't necessarily profitable, as its budget may also be very large. In order to better compare the profitability of the movies, I calculated the return on investment for each movie as well. Return on investment of movies is calculated as follows:

$$ROI = box\ office/budget$$

I got two new columns "dom_roi" and "int_roi" representing the ROI in domestic and international markets.

Another small adjustment I made is that I changed the numbers in Bechdel rating to the actual test results. More specifically, I changed "0" to "Pass 0:Fewer than two women", "1" to "Pass 1:Women don't talk to each other", "2" to "Pass 2:Women only talk about men" and "3" to "Pass 3:Passes Bechdel Test". In addition, I did the same change to **bechdel_df** as well.

After the above manipulation steps, I output the new data frame "**bechdel_budget_profit**" and "**bechdel**" as CSV files for later visualization purpose.

3.3 bech_movie data frame

Another data frame that I created after joining is bech_movie. This data frame contains 7683 entries of data, i.e. 7683 movies from 1880 to 2020. Some major manipulations I made are as below:

Since a movie can have multiple genres, countries of production, languages, directors and writers, I **`split`** the corresponding columns on "," and **`explode`**d the columns into individual rows. Therefore, I achieved per genre, country, language, director and writer per movie per row. And I later used the SQL function **`trim()`** to get rid of any potential spaces in my strings.

I created a binary column "binary_pass" which equals 1 if the movie passes the Bechdel test and 0 otherwise. I also changed the numbers in Bechdel rating to the actual test results similar to what I did in 3.2.

After the series of preprocessing, I output the new data frame "**bech_movie**" for later use.

I will further introduce other manipulation steps, especially how I conducted multi-dimensional reduction to answer my exploring questions in the next section, analysis and visualization.

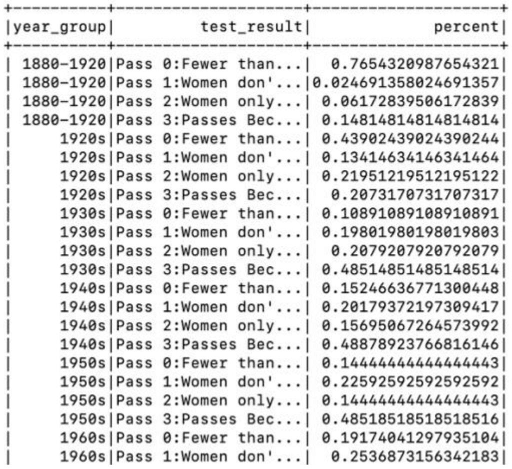
4 Analysis and Visualization

4.1 Percentage of movies passing the Bechdel test over decades: since 1880

4.1.1 Calculation and MapReduce

This is a warm-up question that can be answered with data solely from the **dechdel_df**. The reason why I included it here is because it is a good starting point that can give some insights into how the percentage of movies passing the Bechdel test changes over decades.

Using the dechdel_df, I first defined a function `categorizer(year)` to create 11 year-groups: 1880-1920, 1920s, 1930s...2010s. I combined the data from 1880 to 1920 because of the lack of enough data in each decade. Then I used the Spark SQL function `udf()` to specify the function as user-defined. I created a new column "year_group" based on the release year of the movie in my Bechdel data frame (dechdel_df). Using `GROUP BY year_group, test_result`, I calculated the number of movies satisfied each test criterion in each year group, as well as the total number of movies in each year group. After that, I calculated the percentage of movies passing each test in each year group. For more code details please refer to **si618_project1_zhuoqunw.py**. Below is part of the table generated for visualization purpose:



```
+----------+--------------------+--------------------+
|year_group|         test_result|             percent|
+----------+--------------------+--------------------+
| 1880-1920|Pass 0:Fewer than...|  0.7654320987654321|
| 1880-1920|Pass 1:Women don'...|0.024691358024691357|
| 1880-1920|Pass 2:Women only...| 0.06172839506172839|
| 1880-1920|Pass 3:Passes Bec...| 0.14814814814814814|
|     1920s|Pass 0:Fewer than...| 0.43902439024390244|
|     1920s|Pass 1:Women don'...| 0.13414634146341464|
|     1920s|Pass 2:Women only...| 0.21951219512195122|
|     1920s|Pass 3:Passes Bec...|  0.2073170731707317|
|     1930s|Pass 0:Fewer than...| 0.10891089108910891|
|     1930s|Pass 1:Women don'...| 0.19801980198019803|
|     1930s|Pass 2:Women only...|  0.2079207920792079|
|     1930s|Pass 3:Passes Bec...| 0.48514851485148514|
|     1940s|Pass 0:Fewer than...| 0.15246636771300448|
|     1940s|Pass 1:Women don'...| 0.20179372197309417|
|     1940s|Pass 2:Women only...| 0.15695067264573992|
|     1940s|Pass 3:Passes Bec...| 0.48878923766816146|
|     1950s|Pass 0:Fewer than...| 0.14444444444444443|
|     1950s|Pass 1:Women don'...| 0.22592592592592592|
|     1950s|Pass 2:Women only...| 0.14444444444444443|
|     1950s|Pass 3:Passes Bec...| 0.48518518518518516|
|     1960s|Pass 0:Fewer than...| 0.19174041297935104|
|     1960s|Pass 1:Women don'...|  0.2536873156342183|
```

Table 1: percentage of movies passing each test in each year group

4.1.2 Analysis in findings

I visualized the table using a stacked bar chart, with percent being the x axis and year group being the y axis. As shown in Figure2, overall, there is an increasing trend in the percentage of movies passing the Bechdel test (represented by the green bars) over decades. In the most recent decade, about 64% of movies in the Bechdel dataset passed the test. Although this number might be larger than in reality due to selection bias, i.e. people who contribute to the Bechdel dataset may prefer watching movies with female prominence, it can at least reflect the trend that there is more representation of women in movies. Additionally, the percentage of movies without two named women (represented by the blue bars in Figure 1) shrinks significantly over decades, which can further support our finding.

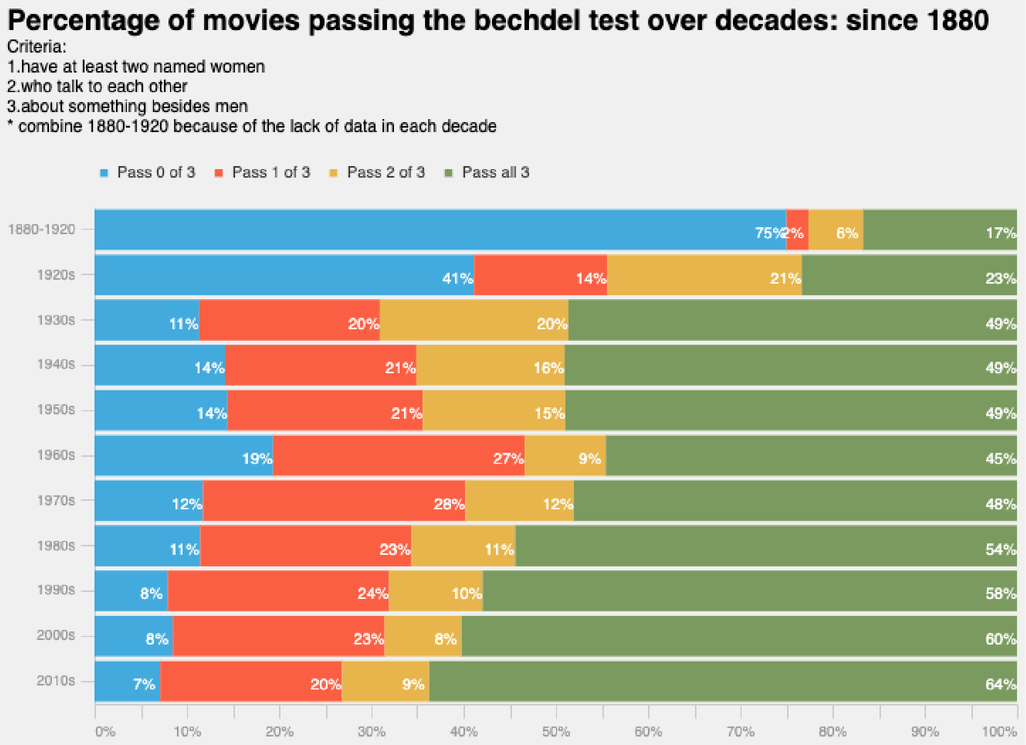
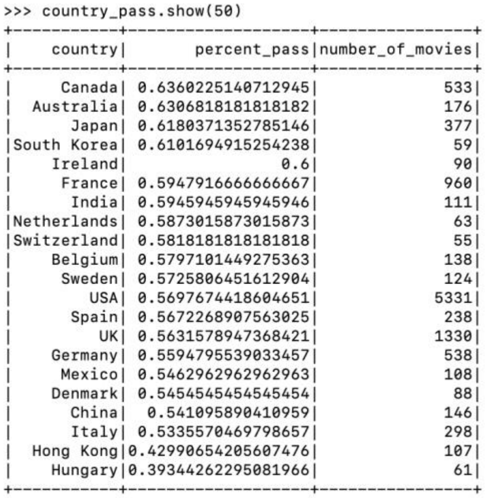


Figure 2: Percentage of movies passing Bechdel test (since 1880)

4.2 Percentage of movies passing the Bechdel test by country of production

4.2.1 Calculation and MapReduce

For this question, I wanted to find the percentage of movies passing the test by country of production. I used the bech_movie data frame (**bech_movie_df**) to answer this question. I first **`select`** "country", "binary_pass" and "imdbid" columns from bech_movie_df and drop the duplicates by **`.drop_duplicates().`** Next, I took the average of "binary_pass" and **`group by`** country to get the average passing rate for each country. To better illustrate the result, I used **`order by percent_pass`** to sort the data. I also limited the countries to be countries with no less than 50 movies using **`having number_of_movies > 50`**. Below is the table illustrating the result.



```
>>> country_pass.show(50)
+-----------+-------------------+----------------+
|    country|       percent_pass|number_of_movies|
+-----------+-------------------+----------------+
|     Canada| 0.6360225140712945|             533|
|  Australia| 0.6306818181818182|             176|
|      Japan| 0.6180371352785146|             377|
|South Korea| 0.6101694915254238|              59|
|    Ireland|                0.6|              90|
|     France| 0.5947916666666667|             960|
|      India| 0.5945945945945946|             111|
|Netherlands| 0.5873015873015873|              63|
|Switzerland| 0.5818181818181818|              55|
|    Belgium| 0.5797101449275363|             138|
|     Sweden| 0.5725806451612904|             124|
|        USA| 0.5697674418604651|            5331|
|      Spain| 0.5672268907563025|             238|
|         UK| 0.5631578947368421|            1330|
|    Germany| 0.5594795539033457|             538|
|     Mexico| 0.5462962962962963|             108|
|    Denmark| 0.5454545454545454|              88|
|      China|  0.541095890410959|             146|
|      Italy| 0.5335570469798657|             298|
|  Hong Kong|0.42990654205607476|             107|
|    Hungary|0.39344262295081966|              61|
+-----------+-------------------+----------------+
```

Table 2: percentage of movies passing each test by country of production

4.2.2 Analysis in findings

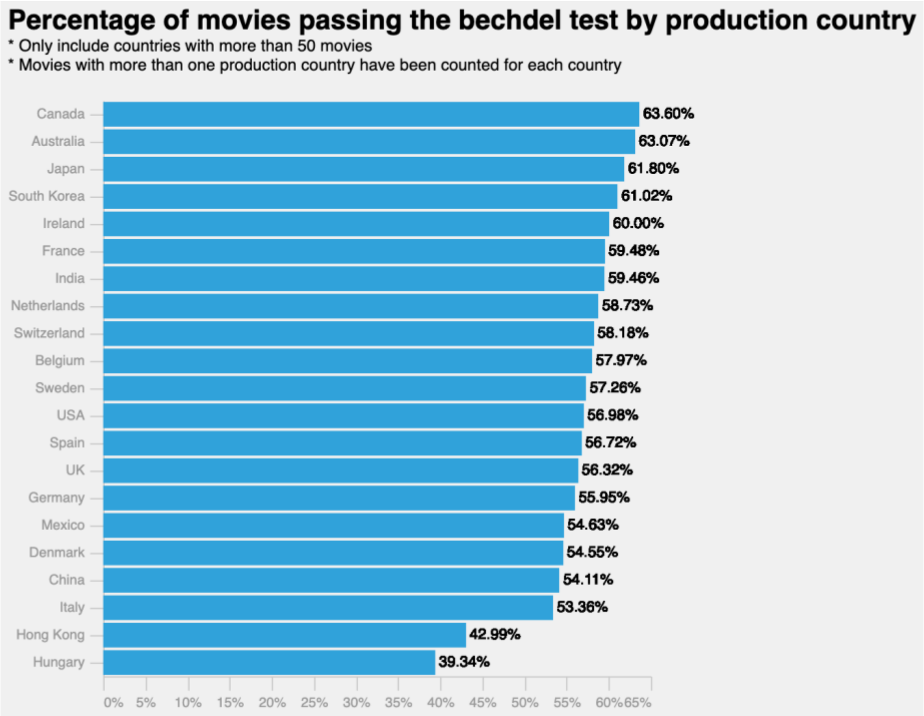


Figure 3: Percentage of movies passing Bechdel test by country of production

As shown in the Figure 3, movies produced by Canada have the highest passing rate, yet movies made by Hungary have the lowest passing rate. There is no clear evidence to show that movies from which part of the world have relatively higher percentage of passing the test (A possible trend would be, for instance, European movies have higher passing rate, yet Asian movies are less likely to pass). This lack of trend may due to multiple reasons. One of the reasons is that, the datasets I was using are from American websites. Thus, more American/English movies are included in the datasets than movies from other parts of the world. Fewer movies may cause bias in the percentages as the movies included may happen to belong to the genres that are more likely to pass the test. As shown in Table 2, the number of movies produced by different countries are quite different. Therefore, further analyses are needed in order to obtain a more accurate conclusion.

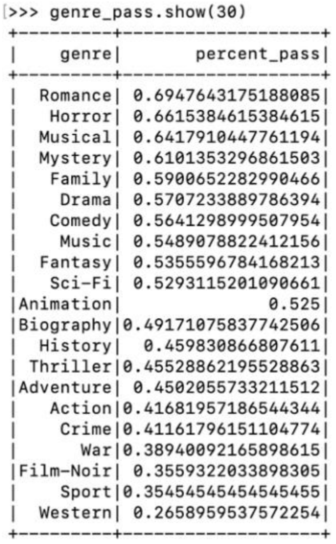
4.3 Percentage of movies passing the Bechdel test by movie genre

4.3.1 Calculation and MapReduce

In this question, I focused on the relationship between movie genre and passing the Bechdel test. Similar to the manipulation procedure in 4.2, I **`selected`** "genre" and "avg(binary_pass)" from the bech_movie table, and **`group by`** genre. Table 3 shows the result in the descending order.

4.3.2 Analysis in findings

As shown in the table and visualization below, movies of genres such as romance, horror, musical, family and drama have higher passing rate, while action, crime, war, sport and western movies are less likely to pass the test. This result makes immediate sense. For example, it's not hard to imagine a mother and her teenage daughter discuss about school in a family movie. It is also common to see female friends talk about some secrets between girls in a drama. On the other hand, action or western movies usually develop the story around a male hero (try to recall 007 series and Mission: Impossible series). There are not as many female characters in such movies. Even if there are some female characters, their roles are usually built for the purpose of supporting the male character without much serious development.



```
>>> genre_pass.show(30)
+---------+-------------------+
|    genre|       percent_pass|
+---------+-------------------+
|  Romance| 0.6947643175188085|
|   Horror| 0.6615384615384615|
|  Musical| 0.6417910447761194|
|  Mystery| 0.6101353296861503|
|   Family| 0.5900652282990466|
|    Drama| 0.5707233889786394|
|   Comedy| 0.5641298999507954|
|    Music| 0.5489078822412156|
|  Fantasy| 0.5355596784168213|
|   Sci-Fi| 0.5293115201090661|
|Animation|              0.525|
|Biography|0.49171075837742506|
|  History|  0.459830866807611|
| Thriller|0.45528862195528863|
|Adventure| 0.4502055733211512|
|   Action|0.41681957186544344|
|    Crime|0.41161796151104774|
|      War|0.38940092165898615|
|Film-Noir| 0.3559322033898305|
|    Sport|0.35454545454545455|
|  Western| 0.2658959537572254|
+---------+-------------------+
```

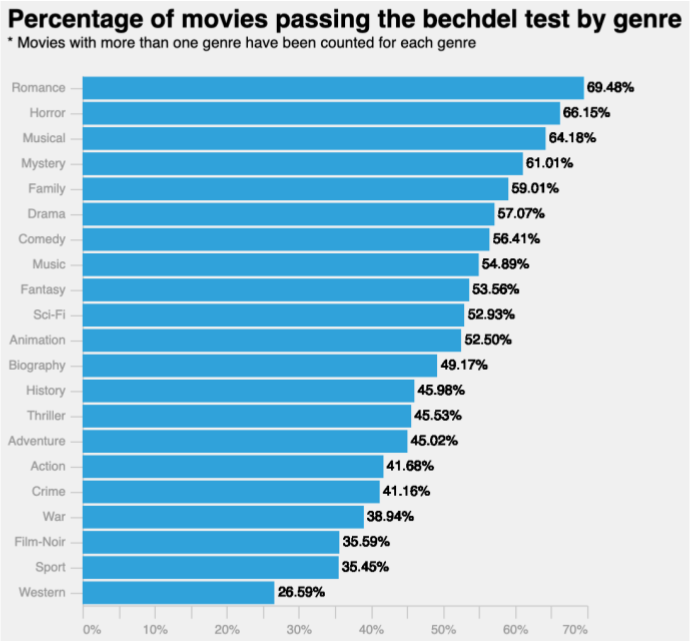
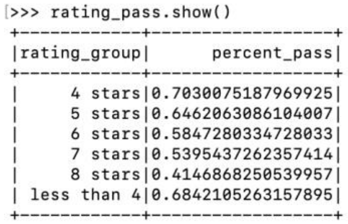


Table 3 & Figure 4: Percentage of movies passing the Bechdel test by movie genre

4.4 Percentage of movies passing the Bechdel test by IMDb ratings

4.4.1 Calculation and MapReduce

Another interesting question to ask would be what is the relationship between IMDb rating and passing the Bechdel test? To answer this, I defined a user-defined function star_bin() using **`udf()`** to create 6 IMDb rating groups: "less than 4 star", "4 star", "5 star", "6 star", "7 star", "8 star". I excluded "9 star" group because of the lack of data in that group. Then I took the average of "binary_pass" in each rating group using **`group by`** in Spark SQL. Below is what I got:



```
>>> rating_pass.show()
+------------+------------------+
|rating_group|      percent_pass|
+------------+------------------+
|     4 stars|0.7030075187969925|
|     5 stars|0.6462063086104007|
|     6 stars|0.5847280334728033|
|     7 stars|0.5395437262357414|
|     8 stars|0.4146868250539957|
| less than 4|0.6842105263157895|
+------------+------------------+
```

Table 4: Percentage of movies passing the Bechdel test by IMDb rating

4.4.2 Analysis in findings

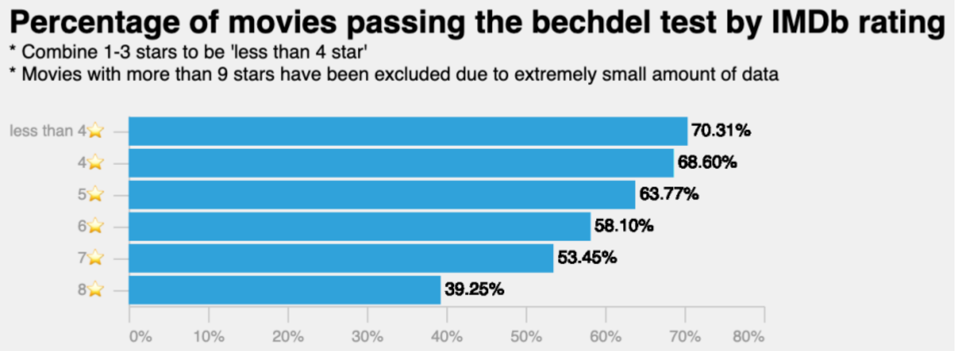


Figure 5.1: Percentage of movies passing the Bechdel test by IMDb rating

There is an overall trend that the higher the IMDb rating, the lower the percentage of movies passing the Bechdel test. There might be multiple reasons why this is the case. First possible reason is that, in general, movies passing the Bechdel test have lower quality than movies failing the test. Lower quality may come from budget limitation. As I will discuss later in the report, movies passing the Bechdel test have lower budget than other movies in general. Without enough investment, movies are less likely to be of high quality. Another possible reason is that, general public prefers movies centered around male characters, which may again be explained partly by the fact that such movies have higher budget and thus, higher quality.

Another interesting finding is that, this negative relationship between IMDb rating and the percentage passing the Bechdel test doesn't differ between male and female voters, as shown in Figure 5.2.

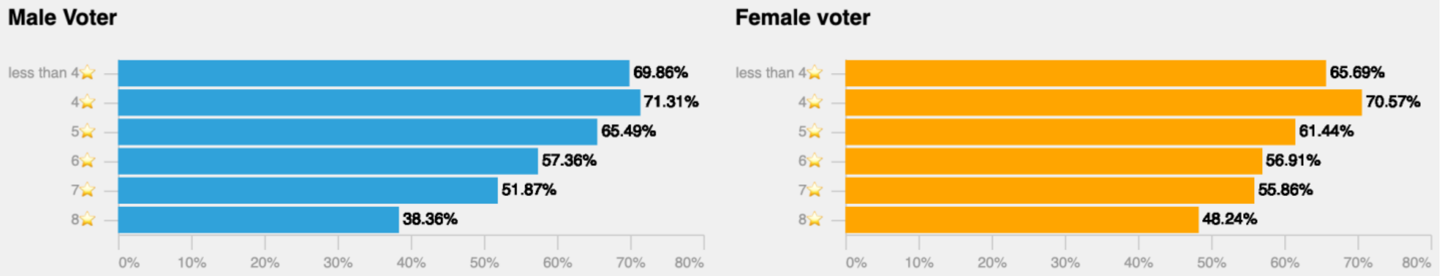
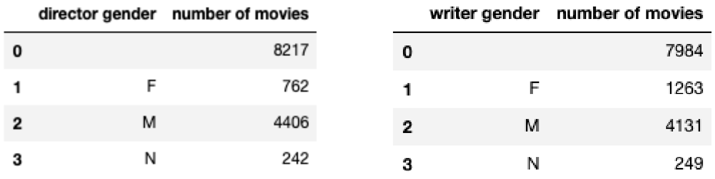


Figure 5.2: Percentage of movies passing the Bechdel test by IMDb rating (male vs. female voter)

4.5 Percentage of movies passing the Bechdel test by crew gender

4.5.1 Calculation and MapReduce

How does crew gender, for example, directors and writers' gender, affects the percentage of movies passing the test? I used the bech_movie_df and a package "genderperformr" in Python to answer this question. I didn't find reliable sources of crew gender that fit my existing datasets well, so I decided to use the Python package "genderperformr" to predict the gender based on first names. Since this package was not available in Pyspark, I predicted and generated the director and writer gender columns on my local computer. There are 4 possible outcomes according to genderperformr document: "M" for male, "F" for female, "N" for neutral and " " for other[1]. I only included "M" and "F" in my analysis for simplicity. Below are two tables showing the number of movies by each predicted gender of directors and writers. If there are multiple directors/writers in a movie, I counted that movie for each director/writer.



	director gender	number of movies
0		8217
1	F	762
2	M	4406
3	N	242

	writer gender	number of movies
0		7984
1	F	1263
2	M	4131
3	N	249

Table 5: Number of movies by crew gender

4.5.2 Analysis in findings

[1] Zijwang. (2018, October 22). Zijwang/genderperformr. Retrieved October 24, 2020, from https://github.com/zijwang/genderperformr/blob/master/LICENSE.txt

Not surprisingly, the percentage of movies passing the Bechdel test by female directors/writers is far larger than that percentage by male directors/writers. Female crew members, especially those that are deeply involved in the creation of the stories such as directors and writers, are more likely to focus on the development of female characters, compared to male crew members.

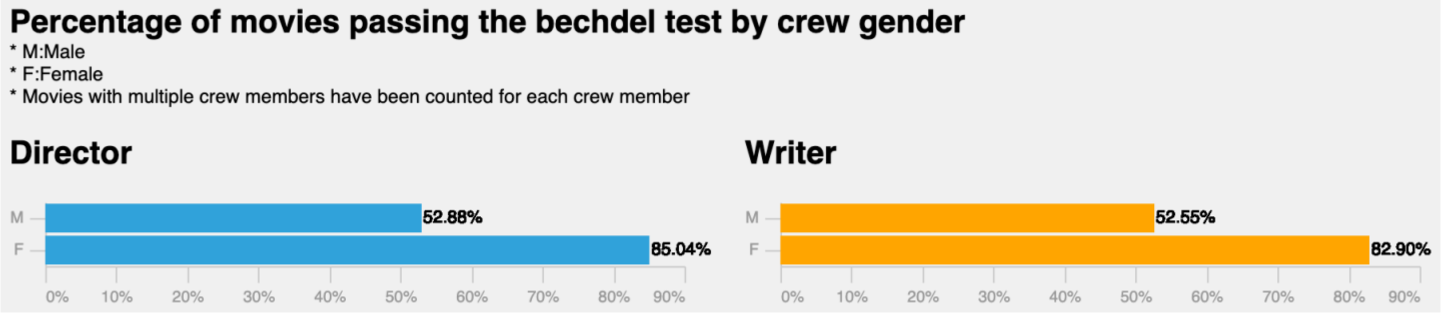


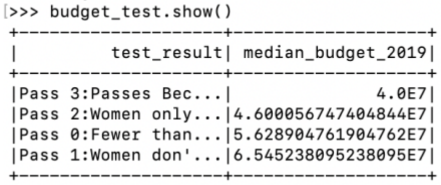
Figure 6: Percentage of movies passing the Bechdel test by crew gender

For the next few questions, I used the **bech_budget_profit** data frame. Again, I only analyzed 2139 movies from 1990 to 2020 due to the limitation of budget and box office data before 1990.

4.6 Median budget and the Bechdel test (since 1990)[2]

4.6.1 Calculation and MapReduce

I selected "test_result" and median adjusted budget from the table and `group by test_result`. There was no direct function returning the median in Spark SQL, therefore I used `percentile_approx(adj_budget, 0.5)` to obtain the median adjusted budget as median_budget_2019. I chose to use median instead of mean because median is more robust when there are outliers. In order to better illustrate the result, I sorted the data with `order by median_budget_2019`.



```
>>> budget_test.show()
+--------------------+--------------------+
|         test_result| median_budget_2019|
+--------------------+--------------------+
|Pass 3:Passes Bec...|               4.0E7|
|Pass 2:Women only...|4.600056747404844E7|
|Pass 0:Fewer than...|5.628904761904762E7|
|Pass 1:Women don'...|6.545238095238095E7|
+--------------------+--------------------+
```

Table 6: Median budget and the Bechdel test (in 2019 dollar)

4.6.2 Analysis in findings

Below is the visualization of the result. As clearly shown in the plot, movies passing the test (represented by the highlighted orange bar) have the lowest median budget. This is not surprising. Action and war movies usually have higher budget in that the special effects need money as well as the super star playing the leading male character. However, as discussed in section 4.3, movies of the genre action or war are less likely to pass the test compared to other movies. This can be one of the reasons why movies pass the Bechdel test have lower median budget than others.

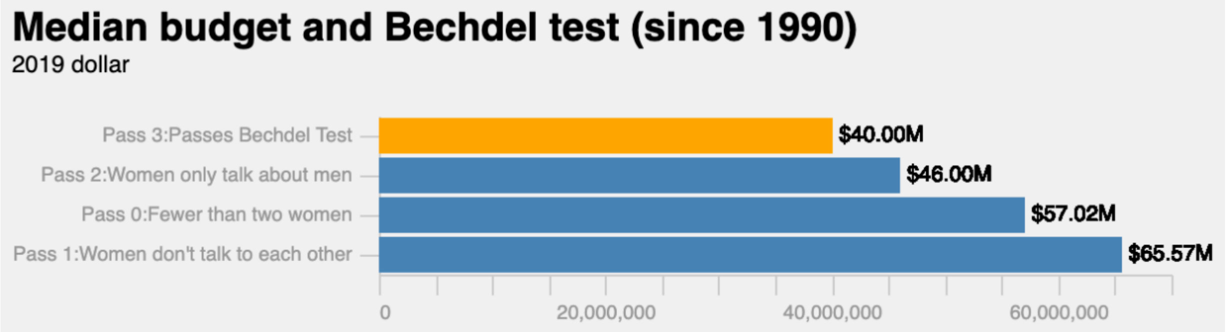


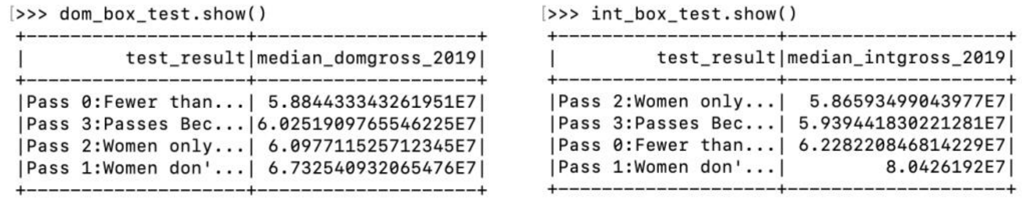
Figure 7: Median budget and the Bechdel test (in 2019 dollar)

4.7 Median box office and the Bechdel test (since 1990)

[2] The ideas for question 4.6-4.8 are inspired by The Dollar-And-Cents Case Against Hollywood's Exclusion of Women from FiveThirtyEight

4.7.1 Calculation and MapReduce

Similar to median budget, I obtained the median box office in both domestic and international markets using **`percentile_approx(, 0.5)`** and **`group by test_result`**.



```
>>> dom_box_test.show()
+--------------------+--------------------+
|         test_result|median_domgross_2019|
+--------------------+--------------------+
|Pass 0:Fewer than...| 5.884433343261951E7|
|Pass 3:Passes Bec...|6.0251909765546225E7|
|Pass 2:Women only...| 6.097711525712345E7|
|Pass 1:Women don'...| 6.732540932065476E7|
+--------------------+--------------------+
```

```
>>> int_box_test.show()
+--------------------+--------------------+
|         test_result|median_intgross_2019|
+--------------------+--------------------+
|Pass 2:Women only...|  5.86593499043977E7|
|Pass 3:Passes Bec...| 5.939441830221281E7|
|Pass 0:Fewer than...| 6.228220846814229E7|
|Pass 1:Women don'...|          8.0426192E7|
+--------------------+--------------------+
```

Table 7: Median box office and the Bechdel test (domestic and international, in 2019 dollar)

4.7.2 Analysis in findings

As shown in Table 7 and Figure 8, the box office of movies passing the Bechdel test is relatively lower compared to other movies both domestically and internationally. However, box office is an absolute value. A movie does not necessarily perform better, i.e. is more profitable, when its box office is extremely large. A more comparable measurement of profitability is the return on investment, which I will discuss in the next section.

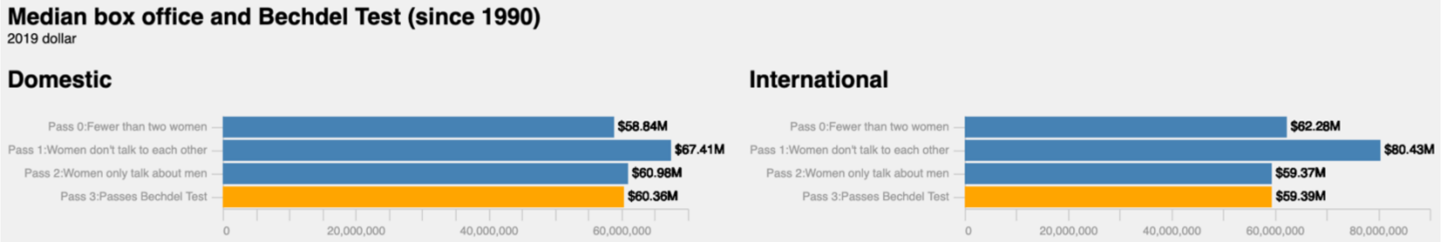
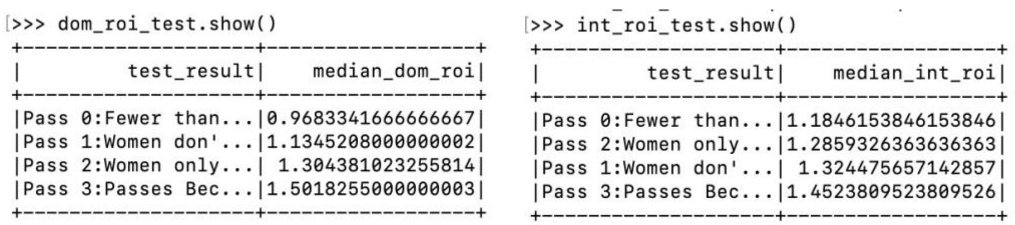


Figure 8: Median box office and the Bechdel test (domestic and international, in 2019 dollar)

4.8 Median ROI and the Bechdel test (since 1990)

4.8.1 Calculation and MapReduce

I got the median ROI for each test result using **`percentile_approx( ,0.5)`** and group by **`test_result`**. The data manipulation procedure is very similar as in sections 4.6 and 4.7, thus I will not discuss in great details here.



```
>>> dom_roi_test.show()
+--------------------+------------------+
|         test_result|    median_dom_roi|
+--------------------+------------------+
|Pass 0:Fewer than...|0.9683341666666667|
|Pass 1:Women don'...|1.1345208000000002|
|Pass 2:Women only...| 1.304381023255814|
|Pass 3:Passes Bec...|1.5018255000000003|
+--------------------+------------------+
```

```
>>> int_roi_test.show()
+--------------------+------------------+
|         test_result|    median_int_roi|
+--------------------+------------------+
|Pass 0:Fewer than...|1.1846153846153846|
|Pass 2:Women only...|1.2859326363636363|
|Pass 1:Women don'...| 1.324475657142857|
|Pass 3:Passes Bec...|1.4523809523809526|
+--------------------+------------------+
```

Table 8: Median ROI and the Bechdel test (domestic and international, in 2019 dollar)

4.8.2 Analysis in findings

Table 8 and Figure 9 suggest that movies passing the Bechdel test (represented by bars highlighted in orange) have higher return on investment rate in both domestic and international market. This is an interesting finding, as "Hollywood believes that international markets don't want to see women in film"[3], and that is one of the reasons why Hollywood is reluctant to produce movies with female leads. However, according to my findings (as well as FiveThirtyEight), movies with female prominence are actually more profitable. That said, to draw more accurate and solid conclusion on this topic, more rigorous statistical tools are needed, such as regression models and hypothesis tests.

[3] Hickey, W. (2014, April 01). The Dollar-And-Cents Case Against Hollywood's Exclusion of Women. Retrieved October 24, 2020, from https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

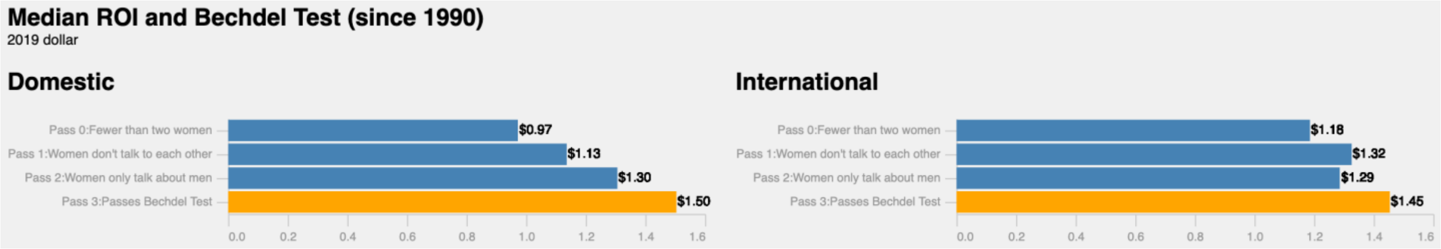


Figure 9: Median ROI and the Bechdel test (domestic and international, in 2019 dollar)

5 Challenge

It is challenging to find the "perfect" datasets that contain reliable information about budget and box office. I was considering using the dataset from the Numbers.com, which is a leading website on budget and box office data. FiveThirtyEight used their data for analysis. However, the datasets from the Numbers are not free. The only free dataset they provide is fairly small, with about 1900 movies released between 2006 and 2018. I needed more data, especially data covering a wider range of time. This is why I chose the Mojo budget dataset from Kaggle which contains information about 3000+ movies between 1990 and 2020.

Another challenge is how to get the crew gender for each movie. I didn't expect gender information was actually hard to find. There are some movie datasets on Kaggle that contain crew gender, but the data is either incomplete, or very messy. Therefore, I decided to predict the gender myself using the package "genderperformr". I would like to thank Professor Eytan Adar, the instructor of SI 649 Info Visualization, for introducing this package to me.

Adjusting for inflation is also not an easy job. There is a package called "cpi" in Python developed by Los Angeles Times Data and Graphics Department that can adjust for inflation automatically for columns in Pandas data frame. However, this package is not available in Pyspark, so I decided to adjust for inflation using the same method as the cpi package myself. That means I had to collect CPI data and calculate the adjusted value manually. But it was worthwhile, as I also learned a lot during the process.

Reference:

[1] Hickey, W. (2014, April 01). The Dollar-And-Cents Case Against Hollywood's Exclusion of Women. Retrieved October 24, 2020, from https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

[2] Cpi package. (n.d.). Retrieved October 24, 2020, from https://github.com/datadesk/cpi

[3] Zijwang. (2018, October 22). Zijwang/genderperformr. Retrieved October 24, 2020, from https://github.com/zijwang/genderperformr/blob/master/LICENSE.txt

[4] Bechdel Movie Test Statistics. (n.d.). Retrieved October 24, 2020, from https://bechdel.hoa.ro/

[5] Bechdel Test. (n.d.). Retrieved October 24, 2020, from https://bechdeltest.com/