Anika Hussen

Database Design and Web Implementation

February 21, 2019

Assignment 3: Bechdel Test Dataset Query Narrative

Bechdel Dataset URL: https://data.world/carolee/women-in-movies

Querying my dataset did not raise any major inconsistencies or errors, but I did find a subtle data placement error. I had initially had my scrubbing program without the CSV module because there did not seem like a problem with extra comma delimiters. However, when test querying data from the "budget_2013$" column, I started to get values like "2008FAIL". This "year_binary" format was in the column "code", which I had removed with my Python scrubbing program. I realized that some of this data had shifted to the left on the rows where the title values had internal commas and delimited by quotes (ie. "Synecdoche, New York"). I edited my Python scrubbing program using the CSV module, eliminating the problem caused by the internal title commas. I also found minor errors like spelling errors in header values, which messed with indexes of the computational values. My queries, then, had their expected results. Pandas already show my unknown values or empty spaces with the notation NaN, so I did not have to account for them separately. However, I did document the "sum" of null values for each column to give me a sense of how accurate my data is. I also queried for the number of rows that did not have null values in the integral numerical columns, labeling them as more sufficient data.

Although my dataset is large enough to make close to accurate conclusions, it would be useful to find the accurate values for the missing data for the domestic and international grosses. The economics and finances surrounding the female representation discussion in the movie

industry help me make conclusions like the most grossing films tend to fail the Bechdel test. From the sample of the 10 highest grossing films, 70% failed the Bechdel test. The *Star Wars* gross is an outlier, which makes sense since the franchise has a large fanbase. For the 5 films with the largest budgets represented an almost equal "PASS" or "FAIL" Bechdel test result. I expected that the budget and gross values statistics would be skewed to the right because movie production is usually expensive. However, I was not expecting as many Bechdel "FAILS" for the most grossing films because representation has become a key part of film industry success. For further research, more recent (full 2010s film data) would help me figure out how female representation has changed in the film industry, especially since there has been an all-time high increase in entertainment award-winning women.

This dataset allows me to see the representation between film financing, Bechdel test requirements, and female representation but does not touch upon the rest of the factors in the film industry. I would love to look into datasets on the film crew and studio makeups, including information of directors, producers, writers, co-actors etc. I would also love to look into the film crew's gender, age, and race, expertise makeup. To support the datasets, I would love to delve into research on the state of Dollar-and-Cents Case in relation to Hollywood women representation, award-winning female and the films they are in. It would also be interesting to look into women in the independent filmmaking industry and compare to Hollywood women statistics. This way I can get closer to figuring out whether or not female representation is a larger film community problem or if it is an exclusively Hollywood/standard industry issue. Understanding the entertainment representation state will prepare me later target the issues from the right avenues.