Anika Hussen

Database Design and Web Implementation

February 21, 2019

Assignment 2: Bechdel Test Dataset Scrubbing Summary

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

My chosen dataset explores the relationship between movie finances, women representation, and their Bechdel scores. The Bechdel-Wallace test measures the female representation in movies on the basis of whether there are two female characters that speak to each other about a topic other than a man. Since I both plan to enter the entertainment industry and I have been passionately involved in the movement for female representation, analyzing this dataset is relevant for my future career. In the discussion around the Dollar and Cents case, advocating for equal pay between genders, an argument point is that there is a positive correlation between female inclusion in film and the median gross compared to the films that failed in the Bechdel test. This dataset with data on detailed grosses, budgets and Bechdel test descriptors will be helpful in finding correlations between the statistics. In fact, this data is a fair and complete representation of the top Hollywood produced films from the year 1970 to 2013. The Bechdel test dataset can also help find patterns in what Bechdel requirements films are often failing to carry out and further encourage leaders in the film industry to strive for the inclusion.

Carolee Mitchell, who, at the time, was a Customer and Academic Relationship Manager at data.world, extracted the data from open source GitHub “FiveThirtyEight” study, along with Brian Keegan’s replicated study, and published the dataset on the public data.world platform. Brian Keegan is a computational social scientist and professor at the University of Colorado Boulder’s Department of Information Science. Although I got the dataset from a public open source platform, I believe that with Keegan’s and Mitchell’s expertise and citations show that the

data is accountable. Tracing the data back to field studies would present the data as correct to my knowledge. Regardless, biases may occur depending on what categories the publishers chose to leave out of the dataset. For example, data on the film crew or certain movies that did not make it to the top produced list could be helpful in understanding what factors contribute to the lack of female representation.

Since the rows in my .csv file in Excel Sheet are organized and coherent, it was straightforward to create graphs. I worked to store numeric and textual data with their data type. The graphs showed an accurate representation of the difference between my numeric values like domestic and International grosses. I removed any “N/A” values altogether, so the integer values would not calculate or display with textual values. When there was a repeated descriptor between the test(changed to "classifier") and clean_test, I changed the "classifier" term to “none” because it is essential to know that there were no extra notes in the Bechdel test reasoning. The “none” matches the textual data in the “classifier” column, allowing me to create frequency graphs. The frequency of “classifier” terms shows me the dominant and least dominant reasons why films fail the Bechdel test. I also removed the “decade code” and the “code” columns, since the decade could be decoded from the “year” data and the “code” can be created by concatenating the “year” value with its corresponding “binary” value. I also changed the HTML entities like “&” and “'” into the readable characters. I replaced my header titles to correspond to my new data(ie. “test” to “classifier”) and to distinguish between time span (ie. “budget_intial”, “budget_2013$”).

To test my data on Excel Spreadsheet, I used the COUNTIF function to find that the frequency of “PASS” films is 803 and “FAIL” films is 991; almost representing a 50:50 split. I also used a bar graph comparing total film grosses, finding an outlier for Star Wars, reaching about a high

$4.3 billion. A significant correlation to this outlier is that Star Wars failed at the Bechdel test. Continuing with the frequency analysis, I found the greatest reason for failing the Bechdel test was not because of the lack of women in films, but because their conversation topics revolved around the male characters. Although there are more movies that fail the Bechdel test, the amount of Bechdel test approved films continue to increase to the point that in the 2010s, Bechdel test approved movies rose up to 409, which is close to the 430 failed movies. However, crossing equivalency is just a threshold to pass towards equal representation. The dataset can be used to study and visualize the relationship between the movie industry and female representation.