Anika Hussen

Database Design and Web Implementation March 5, 2018

SQLite Assignment 4 Narrative

https://data.world/popculture/imdb-5000-movie-dataset **OR**

https://github.com/sundeepblue/movie_rating_prediction/blob/master/movie_metadata.csv

J.P. Morgan Vice President and data scientist, Chaun Sun, aggregated data from well-known movie data collection sites including the-numbers.com and imdb.com to create the massive Movie Metadata. It consists of 28 fields for 5043 movies across 100 years. The aggregated records include the Internet Movie Database(IMDb) ratings, film budget, gross and Facebook likes on the movie.

Although this a large collection of data, I found that there were many missing data values, especially when querying for the greatest and lowest rating values. Even if the data represented 66 countries, almost four-fifths of the data represented U.S. productions. The other significant mass of productions also represents the Western regions, including Canada, the U.K., and France. Already, there is skewed cultural representation in film data. When solely querying countries outside of the U.S., surprisingly, I found that among the 15 highest IMDb rated international films, 75% fell into the genres of western, crime and drama. Hollywood and western culture has a massive influence on the creative work around the world. Most films with an IMDb rating of 9.0 or above are western indie films, which in the past decades has risen in popularity from Tarantino and Spike Lee's films to today's Sundance Festival enthusiasm. However, including independent films in this dataset over franchise productions, show that there a move towards personal sharing, but the leading perspectives in film criticism lean heavily towards the western style.

Looking at the monetary values in the dataset, the Western countries, yet, also including Mexico's and Indonesia's outlying profit of over 1600000 and over 2000000, respectively, tend to gain more production profit. While surprisingly many Asian productions like India, Japan, and South Korea face loss. However, this may just be because of the large production cost on Bollywood productions, as they tend to be longer and traditionally include musical numbers and costume. More for the non-western films, there is a lack of data or even replacing monetary value with zero, like of the United Arab Emirates and Cambodia. Perhaps, the foreign data was not readily accessible like the domestic data.

Finding entertainment and film data has been difficult because of private ownership in the entertainment industry. The open source data I find tends to be aggregated from multiple sources like the Movie Metadata set I am working with now. Professional enthusiasts like Chuan Sun, the owner of the open source metadata Github, gather the resources and values from multiple avenues to create a coherent and clean dataset for the public. The data I have worked with in this class so far is not 100% reliable, as it originates from a secondary source, rather than the study themselves. Even if Chaun Sun is a professional in the entertainment field, there is a great possibility that there may have been errors during the transfer of data. From my earlier assignments with the Bechdel Test dataset and my current work diving into the cultural connection in film, I would love to work with data that provided primary contact with the statistics for women and international film. The Movie Metadata gave me insight into the larger film industry. I expected the data to be skewed towards Western culture, with Hollywood's influence, but I was surprised to see a dip in the other entertainment "superpowers" like Bollywood and South Korean entertainment. With all my projects so far, I feel like I am beginning to understand the interactions between entertainment, gender, and culture.