ETL Final Report

For our ETL project we used two CSV files from Kaggle titled, “Movies on Netflix, Prime Video, Hulu and Disney+” and “Netflix Movies and TV Shows”. Each data set included various data points on movie and tv show titles found on Netflix and other streaming services.

Initial cleaning of our CSV files included dropping rows with “N/A” values and filtering to titles within the United States.

Next, we merged the two data frames based on “Title” and cleaned the resulting data frame by removing rows with “N/A” values and renaming the columns with proper spacing and capitalization. Our data size was reduced with the merge because of title variances such as : Movie title “9” also seen as “Nine” or “The Iron Giant” also seen as “Iron Giant,The”.

The following columns were considered unnecessary and dropped: show\_id, ID, TV Show/Movie, Unnamed: 0

The following columns were considered duplicate columns and reduced to just one column: Type, Runtime, Country, Age, Year, Directors

In addition, we added a new column with a combined rating which took the average of the IMDB Rating and Rotten Tomatoes Rating values. In order to do this, we first stripped the “%” from Rotten Tomatoes Rating and converted the values to float values. Next, we converted IMDb rating into a percent and combined with Rotten Tomatoes rating for a combined rating.

Finally, we reordered the columns to show all three ratings together and have the movie title as the first column before loading into Postgresql. Our decision to load into Postgresql was based on our familiarity and the type of data we were using.