**ETL Project**

Group 12

Xier Zhao, Dan Bruen

Our project goal was to collect data on the US domestic film industry and Indian film industry, known commonly as “Bollywood,” in order to compare and contrast the budgets of US movies to Bollywood movies as well as the critical reception of the two industries’ movies. The data we collected was extracted from kaggle.com after we were unable to find any helpful datasets on dataworld.com. Later on in our project we also scraped the API from OMDB[[1]](#footnote-1).

We began our extraction process by downloading the “IMDB 5000 Movie Dataset” from kaggle[[2]](#footnote-2). We were also able to download a comprehensive Bollywood movie dataset from kaggle that included 10 columns on the Bollywood movies’ title, release year, genre and IMDB ID[[3]](#footnote-3). Cleaning the data and dropping erroneous columns was simple. Our biggest challenge in transforming the datasets, particularly the Bollywood dataset, was extracting the movies’ ‘imdbRating’ based on the films IMDB ID since the ‘imdbRating’ was not provided in the Bollywood csv. The ‘imdbRating’ was crucial because that was our way of determining the critic/audience reception of the Bollywood films themselves. To match the Bollywood movies’ IMDB ID to an ‘imdbRating’ we need to pull the ‘imdbRating’ from the OMDB API using a ‘for’ loop. The ‘for’ loop allowed us to create a new DataFrame that included the Bollywood movies’ IMDB ID, title and rating.

Once we had the Bollywood films’ IMDB ratings, we were able to compare and contrast the reception of US domestic films and Bollywood films. After analyzing the two DataFrames in our notebook, we uploaded the two datasets on to SQL. We chose to use SQL because the data is stored once, which eliminates data duplication for us. With a relational database like SQL we could also change the database without impacting the data or the rest of the database, which is what we were able to do after pulling the API data and importing it into a new DataFrame on our notebook.

Our data analysis clearly showed that Bollywood movies have a lower critical reception than movies within the IMDB 5000 dataset, but it is key to note that Bollywood film budgets are roughly 10% of that of US domestic films[[4]](#footnote-4). Unfortunately, we were unable to extract the precise film budget data for the Bollywood movies because that information is incredibly opaque and not accessible online, but we still believe our analysis shows that Bollywood movies have more “bang for their buck” given the Bollywood movies’ reception relative to those movies’ budgets.

1. <http://www.omdbapi.com/?i=tt3896198&apikey=4b6539f6> [↑](#footnote-ref-1)
2. <https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset> [↑](#footnote-ref-2)
3. <https://www.kaggle.com/mitesh58/bollywood-movie-dataset#BollywoodMovieDetail.csv> [↑](#footnote-ref-3)
4. <http://filmmakersfans.com/real-budget-range-indian-films-check-now/> [↑](#footnote-ref-4)