**ETL Project: Netflix Ratings and Production**

Group Members:  Franklyn De La Cruz, Mihee Park, Shae Stringer-Jones,

Maria Valdes, Walt Zhao

DataBaes are looking at the correlation between IMDB ratings and title genres in Netflix. We followed ETL to clean our data so it can be analyzed.

**Extract**: Pulled CSV files using Python/Pandas from the below websites:

* <https://www.kaggle.com/shivamb/netflix-shows>
* <https://www.kaggle.com/rajatkumar30/netflix-movieseries-rating>

**Transform**:

To transform our data we started by analyzing the above CSV files and narrowing the information. We filtered and dropped unwanted columns and null values from the netflix-shows data to narrow our dataset to eight columns: show\_id, type, title, country, data\_added, release\_year, rating, and listed\_in.  The updated data set is named cleaned\_main\_data.  We used the same process for netflix-movie-ratings and kept: title and imdb\_rating.  The updated data set is named cleaned\_rating\_data.  Once the data was narrowed and cleaned we merged the two datasets using the pd.merge function in Pandas.  Our final clean dataframe is named merged\_df.

After our final dataframe was created we did some analysis to see if our merged dataframe yielded any new insights into the data.  We were able to see how many Netflix titles were created in different countries.  We also looked at the most produced and title count by genres.  These observations were displayed in bar graph form to enable the easy translation of the data.  Finally, we looked at the distribution of average user ratings.  This analysis is presented with histograms.

We discovered that the majority of Netflix titles are produced in the United States.  The average movie rating is 6.802 and the most produced genre is action & adventure.  The genre with the highest viewer rating is music & musicals.

**Load**:

Our final dataframe is housed in a relational database created in pgAdmin titled DataBaes. We chose a relational database to support the table and row structure of our dataframe.  This final database includes information about Netflix titles and also reflects viewer reviews from IMDb.

**Appendix**:







