**ETL Project**

Justin, Michelle, and Grant

ETL Project: Movie Rating Data

3 Data Sources: (4 csv’s total)

-IMDb movies extensive dataset from Kaggle (Grant) **movies.csv**

-MovieLens dataset from Kaggle (Justin) **movie.csv** and **rating.csv**

-Rotten tomatoes dataset from Data.world (Michelle) **rotten tomatoes top movies by genre csv**

We extracted the data as CSV’s, but the files were so large it was very difficult to upload to github. We spent a lot of time trying to install git lfs (large file storage) in order to upload the csv’s. We pulled all of the CSV’s into a pandas dataframe once we got past the git lfs issue.

We cleaned and transformed the data separately, and then merged them together by title. We decided to make it more interesting and join tables by titles that show “good movies” from all 3 movie review platforms. We did this by filtering out anything below a rating of 70 for rotten tomatoes, below 8 for IMDB, and below 4 for MovieLens. The Rotten tomatoes dataset had the least amount of data and contained quite a bit of duplicates because they were originally grouped by genre, and some movies had multiple genres.

Two of the datasets had the year in parentheses in the title column in addition to the movie title so we had to split it and keep a “clean title” column and keep the year column separate. We also noticed the order of the title was a little odd in the MovieLens data set. (For ex. it would be the end part of the title, beginning of title.)

We loaded the data to a “production” relational database in postgres using PgAdmin and then joined tables. We connected to postgres with the connection string and create engine, we then used pandas to load csv converted dataframes into the database. The final table that we ended up using all three rating systems for comparison, RottenTomatometer, IMDB\_Rating, MovieLens\_rating.