**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) **IMDb.csv**

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

-Rotten tomatoes dataset from Data.world (Michelle) **rotten\_tomatoes.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 7.0 for IMDB, and below 4.0 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 created the engine. We then used pandas to load the converted dataframes into the database. From there, you could do various queries to merge the tables by title, genre, year released, etc.

When we tried to merge the three datasets by title to find what movies they all had in common as ‘good movies’, we found that nothing came up which was surprising as we could see common titles with good ratings by searching in Excel. We then try merging two at a time and found that Rotten Tomatoes and MovieLens would merge by title, but not IMDb with either one. We discover that it was because there was a whitespace at the end of each title in the finished Rotten Tomatoes and MovieLens dataframes after we extracted the years they were released in the title column and put them in a new column. We then ran lines of code to take out the whitespace at the end of each title in the Rotten Tomatoes and MovieLens dataframes.

After running the codes to extract the whitespaces at the end of the titles, we were successfully able to merge all three dataframes together to a combined dataframe. Interestingly enough, the combined dataframe only came out to 12 movies (movies that were rated >= 7.0 in IMDb, >= 70% in Rotten Tomatoes, and >= 4.0 in MovieLens). We exported the combined dataset to a new csv. We would like to think of this combined dataframe as our “Ultimate Top 12 Movies to Watch” based on IMDb, Rotten Tomatoes, and MovieLens ratings.