ETL Project
Members:
Crisharon Beale
Julie Pyle
Ben Stork
Robert Payne

ETL Project Proposal:

Purpose:
Create a movie information data frame in SQL that will allow users to quickly and conveniently do the following when looking for a movie:

1) Compare movie ratings by release year. What were the top 5 rated movies based on the combined data sets (ratings on Rotten Tomatoes, IMDb, and possibly user reviews)?
2) Compare movie ratings by genre. What were the top 5 rated movies based on the combined data sets (ratings on Rotten Tomatoes, IMDb, and possibly user reviews)?
3) Target age group of movies (Top 5). This will allow users to get movies that target their specific age group and interest users of the same age.
4) Inform the consumer where to stream a movie.

Work Breakdown:

- **Crisharon**
    - Database management
    - Created several schemas
    - Developed code for the connection between python and postgres
    - Created tables in python & postgres
    - Created general queries for the questions others may use for there analysis
- **Robert**
    - GitHub - Robert managed the github and all contributions
    - Cleaning data - cleaned data to prepare for Postgres import, focusing on the movie genre classification
- **Julie**
    - HTML - Created all HTML code for the final presentation and webpages
    - ER Diagram - Created ER diagram for all other work to be based on
- **Ben**
    - Importing CSV files as well as combining the data sets.
    - Combining code from all group members at the end.
    - Cleaning data - cleaned data to combine the 2 main datasets based on movie “Title” and “Year”

Current Datasources:
https://www.kaggle.com/rounakbanik/the-movies-dataset

https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

Issues:

- How to combine data sources. If completed off of movie titles these are not unique. Can we do it off of Movie Title or another attribute?
    - This was solved by combining based on movie title as well as movie release year
- Genre column in one of the datasets has multiple entries. WIll need to create a separate column for all Genres and have a boolean response if that movie is part of said genre.
    - Robert separated the genre's into separate columns that then had a boolean response for the movie on each genre column
- Is there any other information that is grouped and we are unaware?
    - Genre's were all combined and needed to be separated out before being pushed to postgres