Shreya Bakashi

Evan Mobley

Xiaoran Li

Darren Raymond

ETL Project

Our team collectively decided that we were going to use data sources from Kaggle. Using our newfound ETL skills we thought we could best tell a story using movies and TV shows from the following databases:

* Disney Plus Shows
* Netflix Titles
* IMDb Ratings
* IMDb Movies

To interpret the data and tell the story correctly we needed to read in the CSV files that we selected from Kaggle. Then we reduced each data frame with the relevant columns that we thought would best provide adequate information. Cleaning the data also involved renaming of the columns and dropping any duplicates that our code might have caught.

The first mental roadblock we came across was how to combine two columns with the same category of information and the dropping of null cells. We researched and found out that within the ***.loc*** we could combine it with a ***.isnull*** to locate the column to merge on while simultaneously dropping null values. We started with over 8,000 rows and our finished product has about 3,000 rows.

Our final production will be loaded into a relational database using three different tables. Our Title table will be our main table that includes a rating ID column and year ID column that has a relationship to the release year table and IMDB rating table. Then, using query’s, the user can build out a table finding out information on each show/movie in the database. These final tables were chosen because we thought that the most important piece of data were the titles of the movies and TV shows. Based on that, then the user would want to look up information to make an informed decision on whether they would like to watch this movie. The final database will be loaded as a SQL file within pgAdmin.