ETL Project

By: David Lu, Lizeth Orozco, Stephanie Tang

**Extraction**

We used 4 datasets from the public platform Kaggle to compare ratings from the online streaming platforms Disney+ and Netflix. All four datasets were provided in a CSV format.

* Disney+ Shows
* Netflix Shows
* IMDb Movies
* IMDb Ratings

**Transformation**

To clean up the datasets, we used Pandas functions in Jupyter Notebook to load the CSVs. We reviewed the datasets to determine relevant information. We dropped unnecessary columns and duplicate rows using the dropna and drop\_duplicates functions. Renaming of columns was also performed. In order to only obtain movies in our dataset, we filtered using the loc method.

A screenshot of a cell phone

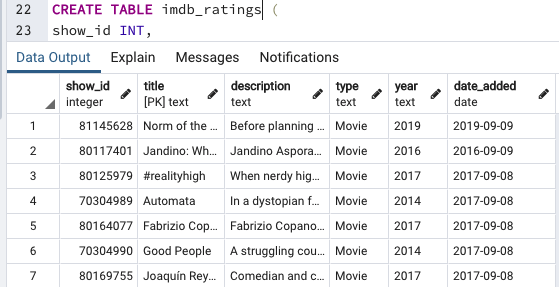
Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Load**

The newly transformed datasets were then loaded into the Postgres database. The datasets were chosen to compare movie titles and ratings across different streaming services.



These datasets could further be used in order to search movie titles by year or based on highest rating. We decided to use a right join on the titles to analyze our data. By joining on the title, we could see the ratings for both streaming services. We can also join on different columns such as on rating or year for a different analysis.

