Netflix Meets IMDB

8th December 2022

Report

About Our Data

Our original data was sourced from Kaggle.

The first dataset (https://www.kaggle.com/datasets/gorochu/complete-imdb-movies-dataset) is a csv file detailing data about IMDB ratings of various movies.

The second dataset (https://www.kaggle.com/datasets/shivamb/netflix-shows) is a json detailing information about various netflix movies. This dataset did not include user rating information.

About Our Transformation

We first took our data into jupyter notebook where we were able to look at it more closely. We decided which columns were relevant to our line of inquiry, and removed the columns that were not useful. We removed the commas in the IMDB dataframe votes column in order to change the data into integer format. This allowed us to sort this column.

From there, we took our data frames over to SQL to clean them up even more. We merged the two tables into a new dataframe using the movie titles.

In order to decipher which movies were most popular within specific years, we sorted the data based on (1) the release year, and (2) the user rating.

We're now able to query the table by year, and see our movie titles ordered by highest rating.

Description

Our final database is a relational SQL database. In the future, we could improve our data by cleaning up the null values.