ETL Project Report - Group 10

William Pryor, Jeremy Jackson, Sukhyun Kim, Yan Kong

Objectives: In our ETL project we are

1. Merging tables we found on ratings (P, PG, PG-13, R, etc.) in order to compare movies and TV shows
2. Compare the number of movies and TV-shows released by release years
3. Find the meaning of the ratings of TV-shows and movies
4. In 2018 Netflix released a report that showed that the number of TV shows on Netflix has nearly tripled since 2010. We are assuming that people watch more TV-shows than movies and tv-shows have better ratings than movies.

ETL (Extract, Transform, Load)

Extract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

- We have used two datasets from the sources below.
    - https://www.kaggle.com/shivamb/netflix-shows → netflix_title.csv
    - https://data.world/chasewillden/netflix-shows → netflix_stuff.csv
        - We converted the file into csv from xlsx to use it with Python on Jupyter Notebook

Transform: what data cleaning or transformation was required.

- Data Cleaning - used Pandas library
    - Merged two tables using outer join to get the maximum amount of data to start from.
        - There were rows that could be replaced or filled.
    - Filled in null values with matching data, for better results.
        - There was some data that has rating and user rating score but didn't have type (TV Show or Movie). To keep the data with user rating score, we assigned type based on rating.
    - Dropped duplicate values for a cleaner data set.
    - Also, dropped values with no user rating score - which was important for our analysis.
    - Renamed columns to make the column names cleaner
    - Reset the index numbers to double check the length of data

Load: the final database, tables/collections, and why this was chosen.

- Database

- *We chose to use relational-database(SQL), because*
  - Data such as: Title, Director, Casts, and Release Year cannot be changed. These are unique values.
  - Netflix collects ratings from users, and the data collected has to be for the specific movie/TV shows.
  - It is easier to search a specific movie/TV show with SQL structure.
  - Data integrity and consistency are more maintainable because of SQL requirements for ACID compliance.

Limitations

- There was a significant number of null values in some columns. We would have had better comparison results if the dataset was more complete.
- While we cleaned the data, we needed to drop, replace, or assign values. There could be some errors due to the reassignment.
- We had data with incorrect values. TV shows labeled as movie's and vice versa.
- One of our base assumptions was that the rating starts with 'TV' would be TV shows and the ones without would be Movies. But, there could be some cases that may not always be true.