**ETL Netflix Project Report**

This ETL project combines two Netflix data sources from Kaggle into a single PostgreSQL table.

**COLLABORATORS**

This project completed by Jamie Clipperton, Adriana Alcantara, David James & Michele Allen

**EXTRACT**

The type of sources being used are CSV files provided by Kaggle.com. The CSV files include data pertaining a list of movies and series streaming on Netflix as well as main actor and character names. These files are saved in the CSV file folder of the repository and listed below.

|  |  |  |
| --- | --- | --- |
| Source No. | File Name | Link |
| 1 | titles.csv | [List of movies and series streaming on Netflix](https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies?select=credits.csv) |
| 2 | credits.csv | [List of Characters/Actors in movies and series streaming on Netflix](https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies?select=credits.csv) |

**TRANSFORM**

The titles.csv file contains movies and series streaming on Netflix, description of the movies and series as well as imdb ratings. The credits.csv file contains the main actor/character of each movie/series streaming on Netflix. We opted to merge these data sources into a single table and clean the data. The transform includes:

* Used Pandas functions in Jupyter Notebook to load both CSV files.
* Reviewed the files and transformed into two data frames.
* Merged both data frames together based off their ID column and created a new data frame.
* Dropped unnecessary data columns, removed duplicates from the merged data frame.
* Filtered out all movie content and focused on series only.

**LOAD**

Transferring our final output to PostgreSQL was our last step. In PostgreSQL, we created a database and used our final Panda’s new merged data frame (netflix\_df). In Jupyter notebook, the sqlalchemy module was used to connect to the database and append the data frame to PostgreSQL. We stored our clean data set in PostgreSQL and created several queries to display different data based off our new dataset.

Graphical user interface, text, application, email

Description automatically generated

**Top 10 Netflix Series Based off Imdb Votes**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

**Top Series with Name of Main Actor/Character**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

**Series with Most Seasons**

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated