**ETL Project - Group 2**

**Members: Tristan Carlisle, Shreestina Tamrakar & Cheng Tan**

**Project Proposal:**

To create a database of movies and TV shows on Netflix, including information on their respective directors, ratings, creation year, and run times.

**Extract Data Source:**

Data is in the form of two separate CSV flat files downloaded from Kaggle.

“Netflix Shows” (source: <https://www.kaggle.com/shivamb/netflix-shows>)

* The data set provides a list of over 8800 Netflix TV shows and Movies and includes information such as directors and cast.

“Netflix Top 10 - Tv Shows and Films” (source: <https://www.kaggle.com/dhruvildave/netflix-top-10-tv-shows-and-films/version/13>)

* Data provides a list of titles and their rank by country. It does not include any other details (e.g. Directors, Cast etc.)

**Transformations Needed**:

* Data files require cleaning and filtering
  + Netflix shows list has a number of director information missing. These were dropped as most shows in the Netflix Top 10 refer to more recent shows/movies (refer to cell 2).
  + For “Netflix Show” data set cells with “NaN” for directors were replaced with “Unknown” in order to conserve as much information within the data set as possible (refer to cell 6).
  + “Description” and “Where they are found” columns were dropped as they provide minimal usage for future analysis (refer to cell 3).
  + In the “Top 10” data frame Column “Show Title” was renamed as “Title” for easier joining of the data sets (refer to cell 4).
  + In the “Top 10” data frame the season title column was dropped as it is most predominantly populated with NaN (refer to cell 2).
  + In the “Top 10” data frame the “season title” column was also dropped as the “Netflix Show” data set also contains this data and double-ups were avoided.
  + The date was changed from Month (Word) date, Year format to a more usable universal all numerical Year-Month-Date format (refer to cell 5)
  + The “Top 10” was filtered to dates that matched with the “Netflix Show” as shows released post-September 25 2021 would not be found in the “Netflix Show” data set (refer to cell 4).
* A join was done on the data sets, with the titles used as the primary key (refer to Figure 1).

**Load:**

* The type of final production database was loaded using PGAdmin4 as the data sets combined were in a relational format. PGAdmin4 was deemed the most appropriate for this type of data.
* The final tables that will be used for the production database can be seen as per our attachment (refer to Figure 2).

**Possible uses:**

* Used for direct advertising purposes as a single data source.
* Information could be used for the choice of creators for different countries or choice of creators which have broad market appeal when deciding on what projects to green light.
* Popularity factors can also be analysed in terms of length of products, country of origin and classification ratings to assist with future project choices.

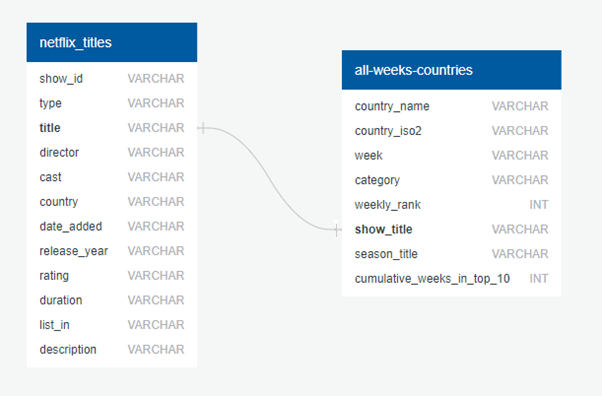
**Related Files:**

Project.ipynb

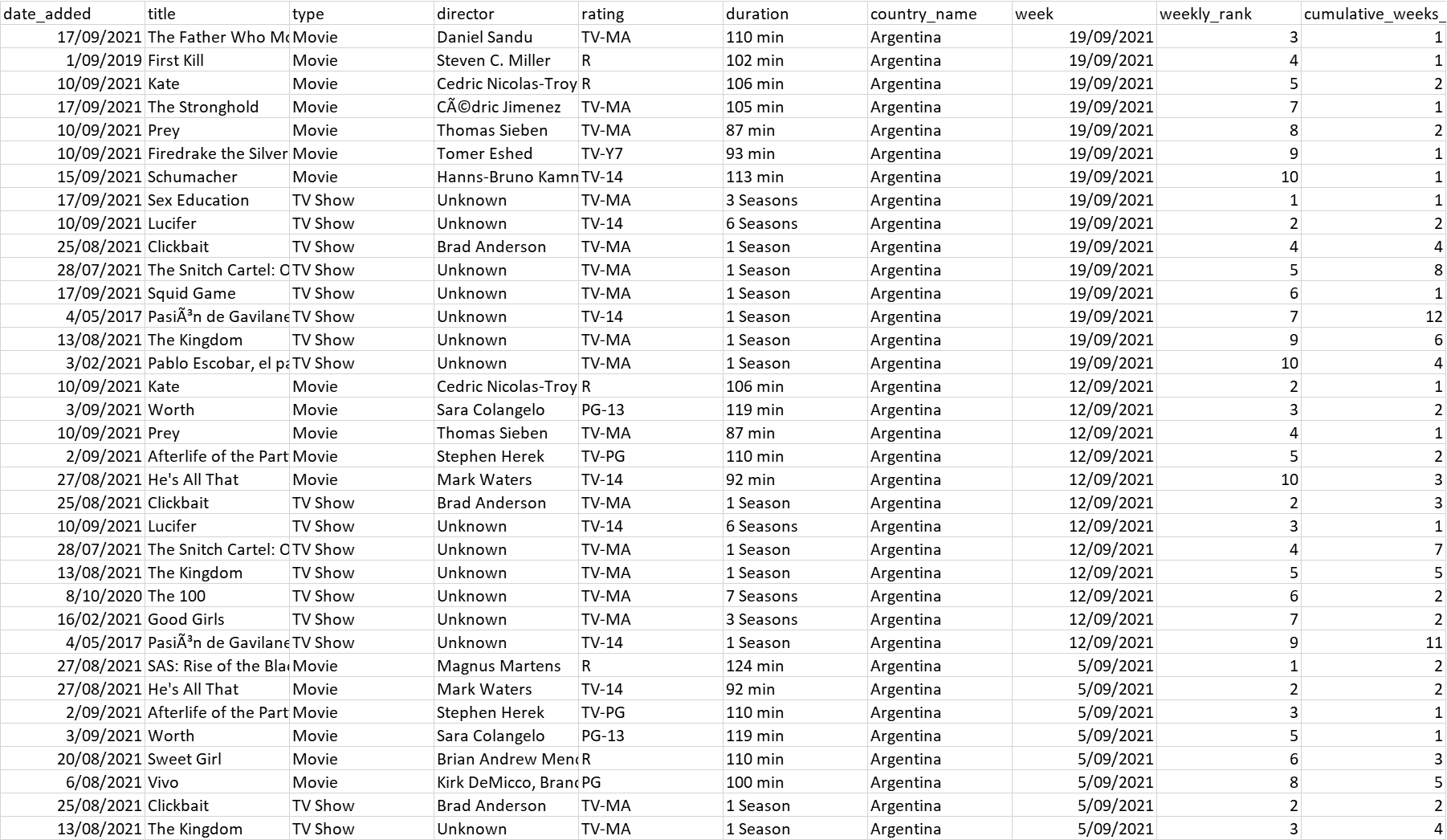
query.sql

schema.sql

**Attachments:**



*Figure 1. ERD Diagram*



*Figure 2: Sample Database Output*