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).

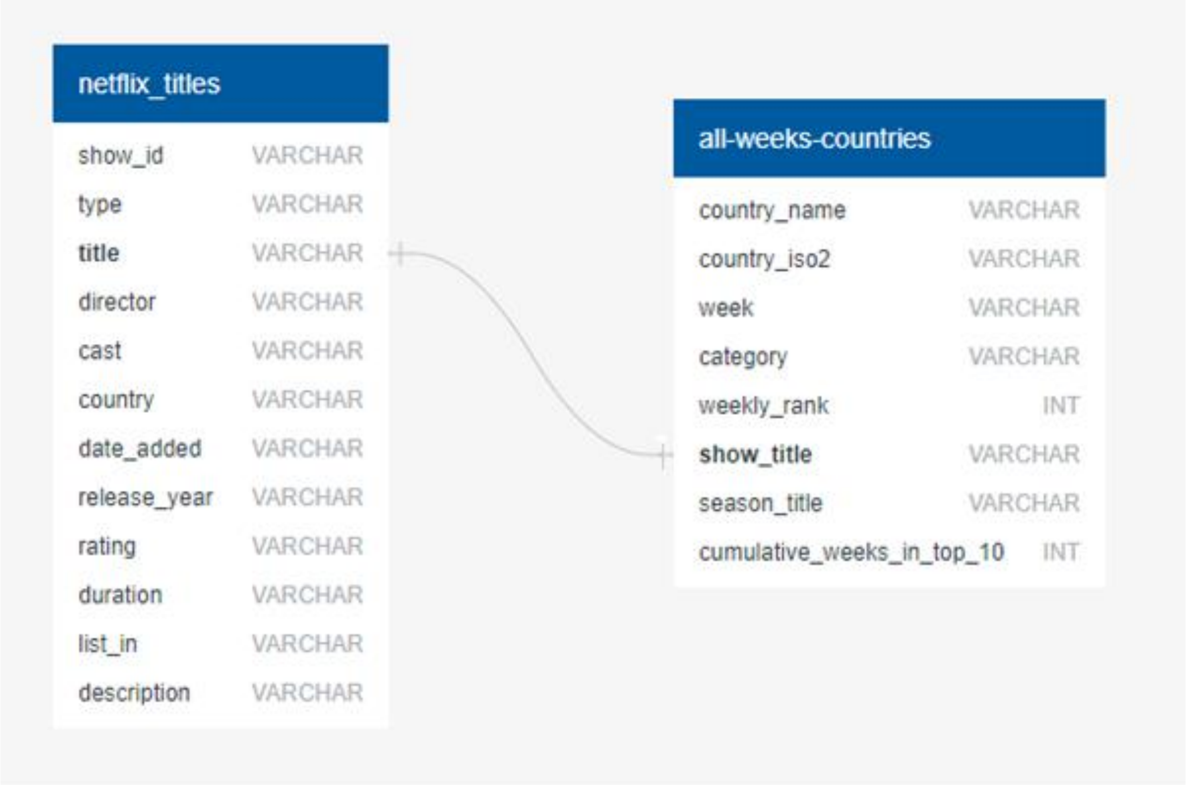
Possible uses:

- 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

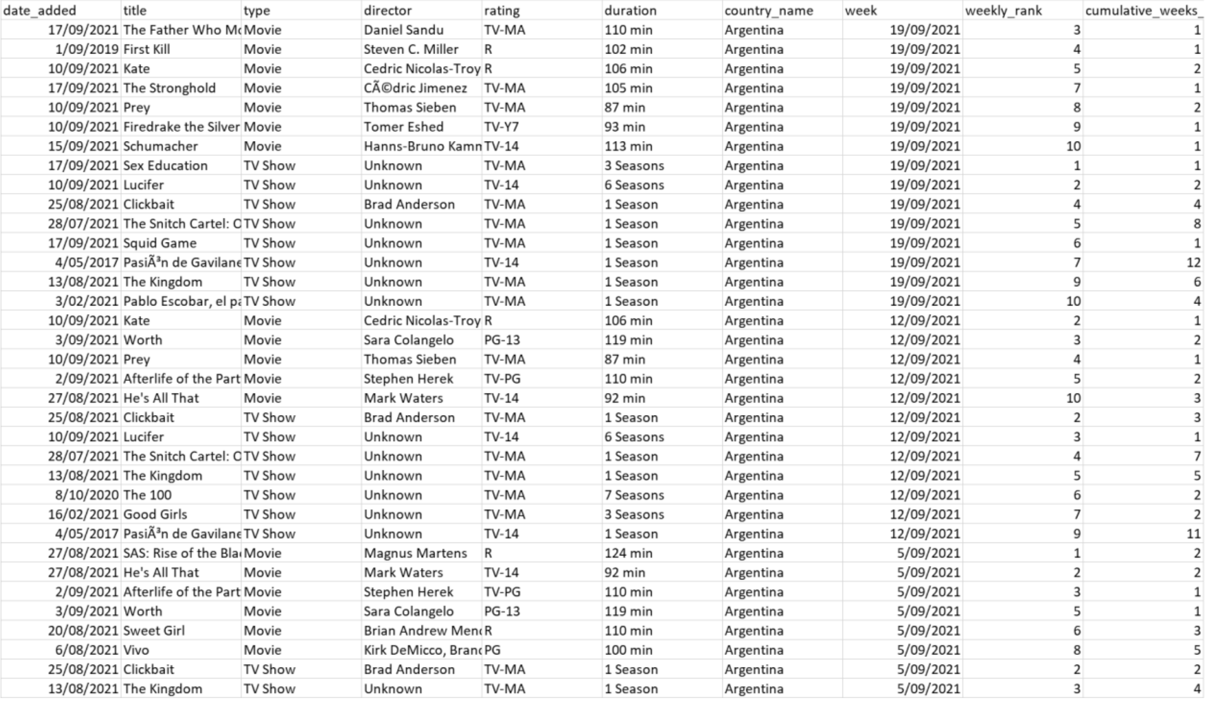
Attachments:



netflix_titles	
show_id	VARCHAR
type	VARCHAR
title	VARCHAR
director	VARCHAR
cast	VARCHAR
country	VARCHAR
date_added	VARCHAR
release_year	VARCHAR
rating	VARCHAR
duration	VARCHAR
list_in	VARCHAR
description	VARCHAR

all-weeks-countries	
country_name	VARCHAR
country_iso2	VARCHAR
week	VARCHAR
category	VARCHAR
weekly_rank	INT
show_title	VARCHAR
season_title	VARCHAR
cumulative_weeks_in_top_10	INT

Figure 1. ERD Diagram



date_added	title	type	director	rating	duration	country_name	week	weekly_rank	cumulative_weeks_
17/09/2021	The Father Who Mo	Movie	Daniel Sandu	TV-MA	110 min	Argentina	19/09/2021	3	1
1/09/2019	First Kill	Movie	Steven C. Miller	R	102 min	Argentina	19/09/2021	4	1
10/09/2021	Kate	Movie	Cedric Nicolas-Troy	R	106 min	Argentina	19/09/2021	5	2
17/09/2021	The Stronghold	Movie	CÃ©dric Jimenez	TV-MA	105 min	Argentina	19/09/2021	7	1
10/09/2021	Prey	Movie	Thomas Sieben	TV-MA	87 min	Argentina	19/09/2021	8	2
10/09/2021	Firedrake the Silver	Movie	Tomer Eshed	TV-Y7	93 min	Argentina	19/09/2021	9	1
15/09/2021	Schumacher	Movie	Hanns-Bruno Kamn	TV-14	113 min	Argentina	19/09/2021	10	1
17/09/2021	Sex Education	TV Show	Unknown	TV-MA	3 Seasons	Argentina	19/09/2021	1	1
10/09/2021	Lucifer	TV Show	Unknown	TV-14	6 Seasons	Argentina	19/09/2021	2	2
25/08/2021	Clickbait	TV Show	Brad Anderson	TV-MA	1 Season	Argentina	19/09/2021	4	4
28/07/2021	The Snitch Cartel: O	TV Show	Unknown	TV-MA	1 Season	Argentina	19/09/2021	5	8
17/09/2021	Squid Game	TV Show	Unknown	TV-MA	1 Season	Argentina	19/09/2021	6	1
4/05/2017	PasiÃ³n de Gavilane	TV Show	Unknown	TV-14	1 Season	Argentina	19/09/2021	7	12
13/08/2021	The Kingdom	TV Show	Unknown	TV-MA	1 Season	Argentina	19/09/2021	9	6
3/02/2021	Pablo Escobar, el pa	TV Show	Unknown	TV-MA	1 Season	Argentina	19/09/2021	10	4
10/09/2021	Kate	Movie	Cedric Nicolas-Troy	R	106 min	Argentina	12/09/2021	2	1
3/09/2021	Worth	Movie	Sara Colangelo	PG-13	119 min	Argentina	12/09/2021	3	2
10/09/2021	Prey	Movie	Thomas Sieben	TV-MA	87 min	Argentina	12/09/2021	4	1
2/09/2021	Afterlife of the Part	Movie	Stephen Herek	TV-PG	110 min	Argentina	12/09/2021	5	2
27/08/2021	He's All That	Movie	Mark Waters	TV-14	92 min	Argentina	12/09/2021	10	3
25/08/2021	Clickbait	TV Show	Brad Anderson	TV-MA	1 Season	Argentina	12/09/2021	2	3
10/09/2021	Lucifer	TV Show	Unknown	TV-14	6 Seasons	Argentina	12/09/2021	3	1
28/07/2021	The Snitch Cartel: O	TV Show	Unknown	TV-MA	1 Season	Argentina	12/09/2021	4	7
13/08/2021	The Kingdom	TV Show	Unknown	TV-MA	1 Season	Argentina	12/09/2021	5	5
8/10/2020	The 100	TV Show	Unknown	TV-MA	7 Seasons	Argentina	12/09/2021	6	2
16/02/2021	Good Girls	TV Show	Unknown	TV-MA	3 Seasons	Argentina	12/09/2021	7	2
4/05/2017	PasiÃ³n de Gavilane	TV Show	Unknown	TV-14	1 Season	Argentina	12/09/2021	9	11
27/08/2021	SAS: Rise of the Bla	Movie	Magnus Martens	R	124 min	Argentina	5/09/2021	1	2
27/08/2021	He's All That	Movie	Mark Waters	TV-14	92 min	Argentina	5/09/2021	2	2
2/09/2021	Afterlife of the Part	Movie	Stephen Herek	TV-PG	110 min	Argentina	5/09/2021	3	1
3/09/2021	Worth	Movie	Sara Colangelo	PG-13	119 min	Argentina	5/09/2021	5	1
20/08/2021	Sweet Girl	Movie	Brian Andrew Menc	R	110 min	Argentina	5/09/2021	6	3
6/08/2021	Vivo	Movie	Kirk DeMicco, Branc	PG	100 min	Argentina	5/09/2021	8	5
25/08/2021	Clickbait	TV Show	Brad Anderson	TV-MA	1 Season	Argentina	5/09/2021	2	2
13/08/2021	The Kingdom	TV Show	Unknown	TV-MA	1 Season	Argentina	5/09/2021	3	4

Figure 2: Sample Database Output