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 or Attachments Figure 2 and Figure 3).
    - 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 or Attachments Figure 4 and Figure 5).
    - For "Netflix Show" data set "Description" and "Where they are found" columns were dropped as they provide minimal usage for future analysis (refer to cell 3 or Attachments Figure 4 and Figure 5).
    - In the "Top 10" data frame Column "Show Title" was renamed as "Title" for easier joining of the data sets (refer to cell 4 or Attachments Figure 2 and Figure 3).
    - In the "Top 10" data frame the season title column was dropped as it is most predominantly populated with NaN (refer to cell 2 or Attachments Figure 2 and Figure 3).
    - 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 or Attachments Figure 2 and Figure 3)
    - 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).
    - NOTE: As there were different countries involved in the datasets some directors with non-english names were not rendered correctly through PGAdmin

- A join was done on the data sets, with the titles used as the primary key (refer to Attachments 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 Attachments Figure 6).

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:

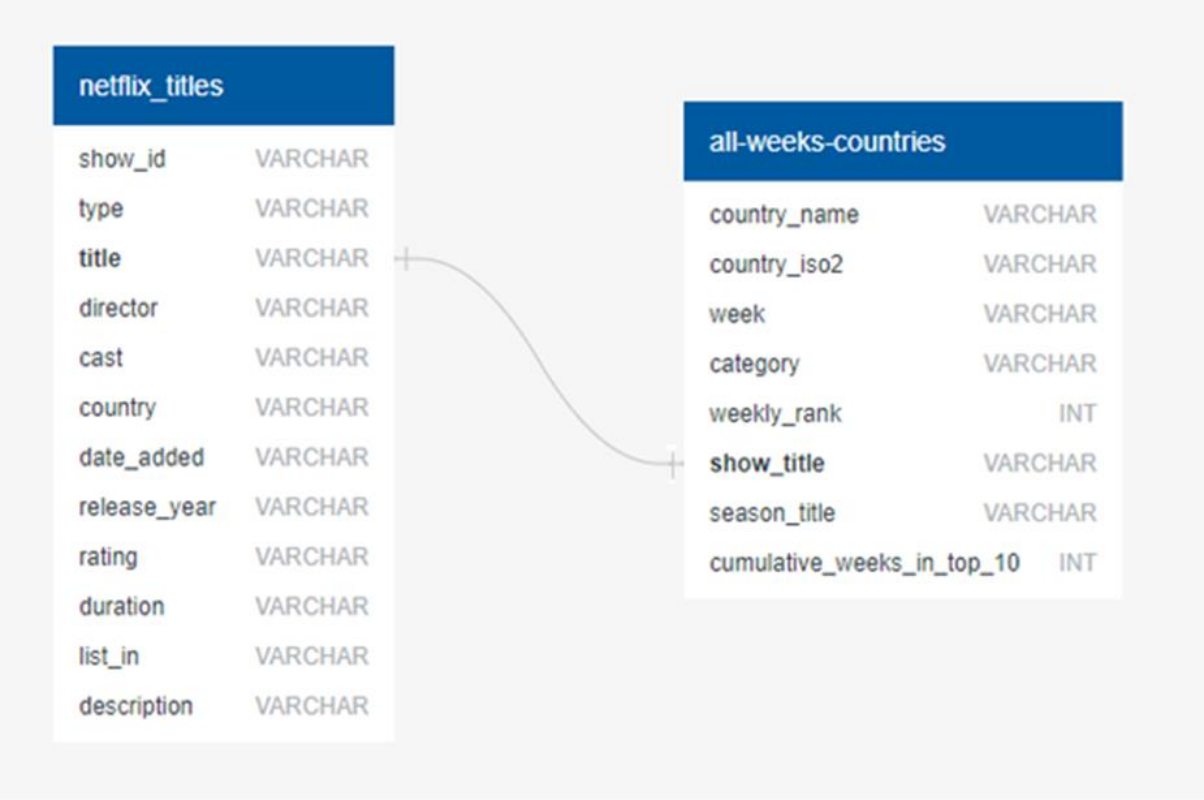

Figure 1. ERD Diagram

	country_name	country_iso2	week	category	weekly_rank	\
0	Argentina	AR	2022-03-06	Films	1	
1	Argentina	AR	2022-03-06	Films	2	
2	Argentina	AR	2022-03-06	Films	3	
3	Argentina	AR	2022-03-06	Films	4	
4	Argentina	AR	2022-03-06	Films	5	

	show_title	season_title	cumulative_weeks_in_top_10
0	Chernobyl 1986	NaN	2
1	The Weekend Away	NaN	1
2	Against The Ice	NaN	1
3	Restless	NaN	2
4	Spider-Man: Into the Spider-Verse	NaN	1

Figure 2. Original “Top 10” dataset pre Transformation

	country_name	country_iso2	week	weekly_rank	title	cumulative_weeks_in_top_10
480	Argentina	AR	2021-09-19	1	The Marksman	1
481	Argentina	AR	2021-09-19	2	The Equalizer 2	2
482	Argentina	AR	2021-09-19	3	The Father Who Moves Mountains	1
483	Argentina	AR	2021-09-19	4	First Kill	1
484	Argentina	AR	2021-09-19	5	Kate	2
...	...	...	...	...	...	...
67655	Vietnam	VN	2021-07-04	6	Reply 1988	1
67656	Vietnam	VN	2021-07-04	7	Nevertheless,	1
67657	Vietnam	VN	2021-07-04	8	Too Hot to Handle	1
67658	Vietnam	VN	2021-07-04	9	Record of Ragnarok	1
67659	Vietnam	VN	2021-07-04	10	Crash Landing on You	1

22560 rows × 6 columns

Figure 3. Transformed "Top 10" Dataset

```
  show_id     type                  title         director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show           Blood & Water              NaN
2      s3  TV Show               Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans              NaN
4      s5  TV Show            Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

          date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                      listed_in  \
0                                  Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                            Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV ...

                                        description
0  As her father nears the end of his life, filmm...
1  After crossing paths at a party, a Cape Town t...
2  To protect his family from a powerful drug lor...
3  Feuds, flirtations and toilet talk go down amo...
4  In a city of coaching centers known to train I...
```

Figure 4. Original "Netflix Show" dataset pre Transformation.

Out[7]:

	type	title	director	date_added	release_year	rating	duration
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	2021-09-25	2020	PG-13	90 min
1	TV Show	Blood & Water	Unknown	2021-09-24	2021	TV-MA	2 Seasons
2	TV Show	Ganglands	Julien Leclercq	2021-09-24	2021	TV-MA	1 Season
3	TV Show	Jailbirds New Orleans	Unknown	2021-09-24	2021	TV-MA	1 Season
4	TV Show	Kota Factory	Unknown	2021-09-24	2021	TV-MA	2 Seasons
...	...	...	...	...	...	...	...
8802	Movie	Zodiac	David Fincher	2019-11-20	2007	R	158 min
8803	TV Show	Zombie Dumb	Unknown	2019-07-01	2018	TV-Y7	2 Seasons
8804	Movie	Zombieland	Ruben Fleischer	2019-11-01	2009	R	88 min
8805	Movie	Zoom	Peter Hewitt	2020-01-11	2006	PG	88 min
8806	Movie	Zubaan	Mozez Singh	2019-03-02	2015	TV-14	111 min

Figure 5. "Netflix Show" dataset post Transformation.

date_added	title	type	director	rating	duration	country_name	week	weekly_rank	cumulative_weeks_
17/09/2021	The Father Who M	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: C	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: C	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 Men	R	110 min	Argentina	5/09/2021	6	3
6/08/2021	Vivo	Movie	Kirk DeMicco, Bran	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 6: Sample Database Output