ETL PROJECT

Team Members:

- Anumala Thapa
- Sybile Cherenfant

Datasets:

- film-locations-in-san-francisco-1.csv (https://data.world/sanfrancisco/yitu-d5am)
- netflix_titles.csv (https://www.kaggle.com/shivamb/netflix-shows)

Proposal:

- Merge movies and shows from Netflix with filmed locations in San Francisco

Extraction

- Datasets (netflix_titles.csv & film-locations-in-san-francisco1.csv) imported from Kaggle and Data.World as csv files and stored into Pandas DataFrame as netflix_movies and film_locations_df

Transform

- New DataFrame netflix_df and film_locations created from selected columns
    - netflix_df - show_id, title, country, release_year
    - film_locations - Title, Release Year, Locations

Load

- Created tables for the two data frames named netflix and film_locations
    - netflix (show_id INT PRIMARY KEY, title TEXT, country TEXT, release_year TEXT)
    - film_locations (id SERIAL PRIMARY KEY, "Title" TEXT, "Release Year" INT, "Locations" TEXT)
        - Columns in SQL film_locations database named inside quotations to match the columns in film_locations_df
- Connected to relational database
- Ran query to inner join both tables on title to confirm tables were successfully loaded in SQL
- Ran (SELECT *) for both tables in jupyter notebook also

Taking San Francisco as an example, we can see how filming locations can help with tourism. People like to replicate scenes that were shot in specific places which assists in tourism. One of the famous locations in San Francisco, Golden Gate Bridge, has been in numerous movies/shows and has become the icon of San Francisco.

Netflix is one of the most popular streaming services. It has several movies and shows that were shot in San Francisco. Therefore, for our project Netflix movies and San Francisco locations datasets was the most convenient to show how filming locations come to play an important role in tourism and also in the movies itself.