

# 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
    - film\_locations columns renamed to match the table in the database

## 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)
- 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 has assisted in tourism.

One of the most famous location in San Francisco, Golden Gate Bridge, has been showcased in numerous movies/shows and has become the icon of San Francisco.

Netflix is one of the most popular streaming services. Because of its affordability and feasibility people can watch several movies and shows that were shot in San Francisco. Therefore, for our project Netflix movies and San Francisco locations datasets were the most convenient to show how filming location come to play an important role in tourism and in the movies itself.