MOVIE DATABASE

By Anna Favis, Becky Warren, Jamie de Guzman and Karen Graham

Through a range of data collection techniques including the use of an API, some web scraping as well as using some csv files we collected data relating to movies. We were interested in looking at the movies that are filmed in San Francisco, the highest grossing of these movies, the genres, runtimes and the production company and the directors.

EXTRACTING AND TRANSFORMING THE DATA

Data title: Movies Meta Data
Data source:

- Kaggle - https://www.kaggle.com/rounakbanik/the-movies-dataset#links_small.csv
- Csv file

Process:

- Downloaded data from the above data source file
- Used Jupyter Notebook to read and transform using pandas
- Reviewed files and selected movies_metadata.csv to work with since it contained all the data from all resource files.
- Reduced columns and dropped columns that seemed unnecessary (‘id’, 'adult', 'belongs_to_collection', 'genres', 'homepage', 'overview', 'popularity', 'production_companies', 'production_countries','poster_path', 'spoken_languages', 'status', 'tagline', 'video', 'vote_average', 'vote_count', 'original_title'
- Rearranged the columns for efficiency
- Reviewed data types to ensure consistency and relevance
- Dropped rows with empty cells in the title column (as this was our variable that we decided to use to join across tables)
- Created a schema for the table
- Final Row Count: 45466

Data Title: Top Grossing Movies
Data Source:

- https://www.boxofficemojo.com/alltime/world/?pagenum=1&sort=rank&order=ASC&p=.htm

Process:

- Started off with wiki, but found that it only showed top 50 grossing films
- Moved over to box office mojo, and scraped top 700
- Created schema for db table

Final Row Count: 700

Data title: Movies filmed in San Francisco
Data Source:

- https://data.world/sanfrancisco/yitu-d5am

Processing:

- Downloaded csv file
- Drop duplicate rows
- Reduce table to: Title, Release Year, Locations
- Drop rows without locations

Final row count : 1920

Data title: Runtime of Top Grossing 654 Movies
Data Source:

- https://www.boxofficemojo.com/alltime/world/?pagenum=1&sort=rank&order=ASC&p=.htm
- http://www.omdbapi.com

Processing:

- Looped through list of movies with API
- Create Lists from API calls
- Built dataframe with Movie Title, Awards, Director, Runtime, Genre
- Downloaded csv file

Final Row Count: 654

LOADING INTO FINAL DATABASE

Created schemas for the tables
Created tables
Read in csv files into pandas
Connected to postgreSQL
Uploaded data to tables