**Project 2:**

**Project Proposal:**

This project would combine 2 data sources to trend movie data. Data would be pulled from 2 different data sources.  One is a CSV and the other is the OMDB API.

**Data Sources:**

1. The first data source is an OMDB API with extensive movie data.
2. The second data source is an CSV with Top 1000 (rows) movies that is a combination of Rotten Tomato’s and IMDB.

**Objective:**

Download the two data sources to build out a table to focus on actors, directors, ratings & sales made.

Data source 1 (Current – within 1 year)

**OMDB API:**

**About Dataset**

**Context**

The OMDB API is a RESTful web service to obtain movie information, all content and images on the site are contributed and maintained by our users.

**Acknowledgements:**

This data was scraped from OMDB.

**Content**

29 columns of data and each movie has the following attributes:

* budget: the budget of a movie. Some movies don't have this, so it appears as 0
* company: the production company
* country: country of origin
* director: the director
* genre: main genre of the movie.
* gross: revenue of the movie
* name: name of the movie
* rating: rating of the movie (R, PG, etc.)
* released: release date (YYYY-MM-DD)
* runtime: duration of the movie
* score: IMDb user rating
* votes: number of user votes
* star: main actor/actress
* writer: writer of the movie
* year: year of release

Data Source 2 (Current - 10th January 2022)

**Top 1000 Highest Grossing Movies:**

**About Dataset**

**Context:**This dataset contains information about the top 1000 highest grossing Hollywood films. It is up to date as of 10th January 2022.

**Acknowledgements:**This data has been scraped from multiple site and has been added together for performing various data operations. The data has been taken from IMDB, rotten tomatoes and many other sites.

**UPDATE:**The original data contained information about movies and some additional have been added by me. 'None' refers to data points that I could not scrape

**Content:**

There are **1000** movies in the dataset. Each movie has the following attributes:

* Title: Movie Title (Join on name)
* Movie Info: the movie summary
* Distributor: the production company
* Release Date: year of release
* Domestic Sales: US revenue of the movie
* International Sales: International revenue of the movie

**TECHNICAL REPORT:**

**OMDB API:**

The OMDB API was in a JSON format and I pulled the data according to a list of ~1000 movies in the CSV.

**Transformation:**

* Created an account and received the API key
* Pulled all the titles to match my CSV data
* Changed the JSON into a dataframe.pd
* Pulled out the columns I needed and added a “id” name to the index column

**Top 1000 Highest Grossing Movies**

The CSV file I pulled the data according to a list of ~1000 movies in the CSV format.

**Transformation:**

* + Extracted the CSV and added a “id” name to the index column
  + Ran a str.strip to remove spaces in the CSV Title column
  + Ran a str.split on the “Title” column to remove the date and created 2 new columns of “Date” and “Title”.
  + Renamed the 2 new columns “Date” and “Title” from the split function.
  + Renamed the “Sales” columns to remove the (in $)

**Load to Database:**

I added a PDF for the ETL load.

* Connected to the SQL film\_db with the “create engine” function
* Used Pandas to load the CSV converted Dataframe into the database
* Used Pandas to load the JSON converted Dataframe into the database
* Confirmed both were loaded correctly

**SQL:**

Merged the two tables, I have a PDF to demonstrate

**Github Repository:**

Repository has been merged.