Project 2: ETL
IMDB Movie Reviews

Caroline Shah, Nicole Hakkarainen, Devyn Spruell, Ankita Mukhopadhyay
Github
Google Drive

Extract

Dataset 1: IMDB Movie List

- **Source:** IMDB 5000 Movie Dataset (Kaggle)
- **Format:** CSV
- **Shape:** 5043 rows x 28 columns
- **Columns of interest:**
  - movie_imdb_link
  - movie_title
  - genres
  - director_name
  - content_rating
  - imdb_score

Dataset 2: IMDB Movie Reviews

- **Source:** IMDB Spoiler Dataset (Kaggle)
- **Format:** JSON
- **Shape:** 573914 rows x 7 columns
- **Columns:**
  - review_date
  - movie_id
  - user_id
  - is_spoiler
  - review_text
  - rating
  - review_summary

Files were too large to store on Github so they are being stored on Google Drive

Transform

Dataset 1: IMDB Movie List

- Stripped down and renamed the columns to get information of interest
- Dropped duplicate and missing data
- IMDB link column contains a unique movie id within the url that we aimed to merge on
- Used “urlparse” to split the url up and create a new column for the movie id
- Set index to the movie id

Final dataset:

- **Format:** CSV
- **Shape:** 4555 rows x 6 columns

Dataset 2: IMDB Movie Reviews

- Dropped duplicates and rows with missing values from dataset
- Removed unnecessary columns from dataset, e.g., review text, user ids, etc.
- Used groupby function to aggregate data at the movie-level
- Calculated average user rating, number of reviews and number of spoilers

Final dataset:

- **Format:** CSV
- **Shape:** 1573 columns x 4 columns

Load

Table 1: movie_titles

- Consist of unique movie id, title, genres, director, rating, and IMDB score

Table 2: IMDB_data

- Consist of unique movie id, average rating, total reviews, and total spoilers

Final Database:

- Combined on the unique movie id allowing us to see the movie title associated the the various IMDB stats

Why and findings:

- We choose these datasets as a fun way to look at movies when we were done with the project. Specifically looking into if there was any correlation between spoilers and movie rating.
- The Dark Knight was the highest rated movie along with having the most reviews and spoilers (over 500 more than the next closest)
- There wasn't any noticeable correlation between spoilers and rating, but there was a correlation between spoilers and the number of reviews

* final database is hyperlinked above to our google drive