Leticia Salazar

Alyssia Smith

Eilish Boyd

Project: GoodRead Book Analysis

Description: We will be using two data sets on popular books and movies to analyze how many books listed were turned into movies.

**EXTRACT**

The first source we used was from Kaggle.com. We used a dataset from GoodReads on popular books that contains metadata for each listed book (goodreads ID, book title, author, rating, etc.). The metadata in this csv file had been extracted from goodreads archive containing 10,000 XLM files. For our project, we referenced the data in columns ‘book ID’, ‘title’ and ‘author name’.

**books.csv --** https://www.kaggle.com/zygmunt/goodbooks-10k#books.csv

Referenced the book\_ID, title and author name

The second dataset we used was also from Kaggle.com. It listed popular movies and information regarding production and release of each movie. We only referenced the movie title from this data because we wanted to see which books were turned into movies.

**tmdb\_5000\_credits.csv --** https://www.kaggle.com/tmdb/tmdb-movie-metadata

**TRANSFORM**

Transformation needed: Movies

Started by renaming columns from their original titles. Dropped all columns **except** ‘title’, ‘release date’ and ‘original language’. We created a bar graph to analyze the number of movies originally released in each language. The data showed that, with the exception of a few movies, the majority of movies in this data set were originally released in English. This demonstrated that this column did not serve as a differentiating variable among entries and would not be relevant to analysis. We re-ordered the columns in the dataframe and then located the cells with null values. From there, we replaced null values with the correct information that was missing.

Transformation needed: Books

Started by renaming columns in the dataframe. From this, we were able to identify which columns were not needed for our analysis. Columns that were not needed were dropped from the dataframe.

After removing irrelevant data, we re-ordered the columns so that two specific columns (‘Title’ and ‘Authors’) were the first two columns in the dataframe. Then, we sorted the title column in ascending order. The last step in cleaning the data was locating null values and filling those cells in with “Unknown”.

**LOAD**

The final production database is relational because the GoodReads and tmdb movies datasets can be represented through tables and compared to one another based on the information they share. The GoodReads and the tmdb datasets were compared by their title column.

Production database:

The final tables and collections used in production database are the popular movies table and the GoodReads books table. We first deleted columns that were not needed from the csv files and then re-imported them into SQL. Then we joined both tables by movie title and original book title. We created a data dictionary with data from both books and movies datasets and put them together into one.



