

Data Table Schema

movie_lense

This set of tables includes ratings from specific users, metadata tags, and genres for movies since 1995. This dataset is sourced from the GroupLens Research Group. Any quotes in the metadata descriptions below are taken from GroupLens' metadata, which can be found in the dataset's README. You are prohibited from redistributing this dataset without first seeking permission from GroupLense.

Size: ~265MB zipped, ~1.2GB unzipped. Source: [GroupLens](#); F. Maxwell Harper and Joseph A. Konstan. 2015. *The MovieLens Datasets: History and Context*. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4: 19:1–19:19. <https://doi.org/10.1145/2827872>

movies. 58098 rows & 3 columns.

Field	Type	Description
movieId	INT	Unique ID assigned to all movies for ease of reference.
title	STRING	The name of the movie, in the form "Name of Movie (<date of movie>)" where <date of movie> is an integer value like "1995".
genres	STRING	All the genres the movie falls into. Multiple genres are pipe (" ") delimited.

ratings. ~28 million rows & 4 columns.

Field	Type	Description
userId	INT	Unique ID to allow tracking of individual reviewers.
movieId	INT	ID assigned to all movies in the movies table.
rating	FLOAT	The rating the userId gave to the movieId. From 0.5 to 5, incrementing by 0.5.
timestamp	INT	Seconds since midnight on January 1, 1970, UTC time.

genome-tags. 1128 rows, 2 columns.

Field	Type	Description
tagId	INT	Unique ID for a specific tag.
itemId	STRING	The name of the tag corresponding to the ID. These tags are reviewer generated and give specific metadata about the movie (e.g. "1920s", or "007 (series)").

tags. ~1 million rows & 4 columns.

Field	Type	Description
userId	INT	Unique ID to allow tracking of individual reviewers.
movieId	INT	ID assigned to all movies in the movies table.
tag	STRING	The specific tag assigned to the movieId by the reviewer with the userId.
timestamp	INT	Seconds since midnight on January 1, 1970, UTC time.

genome-scores. ~15 million rows & 3 columns.

Field	Type	Description
movieId	INT	ID assigned to all movies in the movies table. “The structure is a dense matrix: each movie in the genome has a value for <i>every</i> tag in the genome.”
tagId	INT	ID assigned to all tags in the genome-tags table.
relevance	FLOAT	Created by an ML algorithm. The relevance “encodes how strongly movies exhibit particular properties represented by tags” in a score of 0 to 1. For more detail, see this paper .

links. 58,098 rows & 3 columns.

Field	Type	Description
movieId	INT	“movieId is an identifier for movies used by https://movielens.org . E.g., the movie Toy Story has the link https://movielens.org/movies/1 .”
imdbId	INT	“imdbId is an identifier for movies used by http://www.imdb.com . E.g., the movie Toy Story has the link http://www.imdb.com/title/tt0114709/ .”
tmdbId	INT	“tmdbId is an identifier for movies used by https://www.themoviedb.org . E.g., the movie Toy Story has the link https://www.themoviedb.org/movie/862 .”

the_oscar_award

This table lists all Academy Awards nominations since 1927.

Size: ~900KB unzipped. 10,395 rows & 7 columns. [Source](#)

Field	Type	Description
year_film	INT	The year the film was released.
year_ceremony	INT	The year the ceremony took place.
ceremony	INT	The ceremony's number (e.g. the first ceremony in 1927 is number 1).
category	INT	The name/category of the award in uppercase (e.g. "ACTOR").
name	STRING	The person or entity that received the nomination.
film	STRING	The film that received the nomination.
winner	BOOLEAN	"TRUE" if the nominee won the award, "FALSE" otherwise.

movie_industry

This table lists popular movies from 1986 to 2016, providing industry metadata about the films.

Size: ~950KB unzipped. 6820 rows & 15 columns. [Source](#)

Field	Type	Description
budget	INT	The film's budget (in dollars). Set as 0 if unavailable.
company	STRING	The company that produced the movie.
country	STRING	The country of production.
director	STRING	The name of the film's director.
genre	STRING	The main genre of the film.
gross	INT	The movie's revenue (in dollars).
name	STRING	The name of the film.
rating	STRING	The movie's viewership rating (e.g. R, PG-13, etc.)
released	DATE	Film release date. Format of "YYYY-MM-DD".
runtime	INT	The duration of the movie (in minutes).
score	INT	IMDb user rating.
star	STRING	The lead actor or actress in the film.
votes	INT	The number of votes making up the IMDb score.
writer	STRING	The writer of the movie.
year	INT	The year the film was released.

Other datasets

In addition to the above datasets, you may find the Netflix Prize dataset interesting. [It can be found here](#).