

# Summary

This dataset (ml-32m) describes 5-star rating and free-text tagging activity from [MovieLens](#), a movie recommendation service. It contains 32000204 ratings and 2000072 tag applications across 87585 movies. These data were created by 200948 users between January 09, 1995 and October 12, 2023. This dataset was generated on October 13, 2023.

Users were selected at random for inclusion. All selected users had rated at least 20 movies. No demographic information is included. Each user is represented by an id, and no other information is provided.

The data are contained in the files `links.csv`, `movies.csv`, `ratings.csv` and `tags.csv`. More details about the contents and use of all these files follows.

This and other GroupLens data sets are publicly available for download at <http://grouplens.org/datasets/>.

# Usage License

Neither the University of Minnesota nor any of the researchers involved can guarantee the correctness of the data, its suitability for any particular purpose, or the validity of results based on the use of the data set. The data set may be used for any research purposes under the following conditions:

- The user may not state or imply any endorsement from the University of Minnesota or the GroupLens Research Group.
- The user must acknowledge the use of the data set in publications resulting from the use of the data set (see below for citation information).
- The user may redistribute the data set, including transformations, so long as it is distributed under these same license conditions.
- The user may not use this information for any commercial or revenue-bearing purposes without first obtaining permission from a faculty member of the GroupLens Research Project at the University of Minnesota.
- The executable software scripts are provided "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of them is with you. Should the program prove defective, you assume the cost of all necessary servicing, repair or correction.

In no event shall the University of Minnesota, its affiliates or employees be liable to you for any damages arising out of the use or inability to use these programs (including but not limited to loss of data or data being rendered inaccurate).

If you have any further questions or comments, please email [grouplens-info@umn.edu](mailto:grouplens-info@umn.edu)

# Citation

To acknowledge use of the dataset in publications, please cite the following paper:

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>

# Further Information About GroupLens

GroupLens is a research group in the Department of Computer Science and Engineering at the University of Minnesota. Since its inception in 1992, GroupLens's research projects have explored a variety of fields including:

- recommender systems
- online communities
- mobile and ubiquitous technologies
- digital libraries
- local geographic information systems

GroupLens Research operates a movie recommender based on collaborative filtering, MovieLens, which is the source of these data. We encourage you to visit <http://movielens.org> to try it out! If you have exciting ideas for experimental work to conduct on MovieLens, send us an email at [grouplens-info@cs.umn.edu](mailto:grouplens-info@cs.umn.edu) - we are always interested in working with external collaborators.

# Content and Use of Files

## Verifying the Dataset Contents

The following files (with the provided [MD5 checksums](#)) should be present in this zip file:

MD5	File
8f033867bcb4e6be8792b21468b4fa6e	links.csv
0df90835c19151f9d819d0822e190797	movies.csv
cf12b74f9ad4b94a011f079e26d4270a	ratings.csv
963bf4fa4de6b8901868fddd3eb54567	tags.csv

We encourage you to verify that the dataset you have on your computer is identical to the ones hosted at [grouplens.org](http://grouplens.org). This is an important step if you downloaded the dataset from a location other than [grouplens.org](http://grouplens.org), or if you wish to publish research results based on analysis of the MovieLens dataset.

To verify the dataset (after unzipping):

```
# on linux
md5sum *; cat checksums.txt
```

```
# on OSX
md5 *; cat checksums.txt
```

# windows users can download a tool from Microsoft (or elsewhere) that verifies MD5 checksums

Check that the two lines of output contain the same hash value.

## Formatting and Encoding

The dataset files are written as [comma-separated values](#) files with a single header row. Columns that contain commas (,) are escaped using double-quotes ("). These files are encoded as UTF-8. If accented characters in movie titles or tag values (e.g. *Misérables, Les* (1995)) display incorrectly, make sure that any program reading the data, such as a text editor, terminal, or script, is configured for UTF-8.

## User Ids

MovieLens users were selected at random for inclusion. Their ids have been anonymized. User ids are consistent between `ratings.csv` and `tags.csv` (i.e., the same id refers to the same user across the two files).

## Movie Ids

Only movies with at least one rating or tag are included in the dataset. These movie ids are consistent with those used on the MovieLens web site (e.g., id 1 corresponds to the URL <https://movielens.org/movies/1>). Movie ids are consistent between `ratings.csv`, `tags.csv`, `movies.csv`, and `links.csv` (i.e., the same id refers to the same movie across these four data files).

## Ratings Data File Structure (ratings.csv)

All ratings are contained in the file `ratings.csv`. Each line of this file after the header row represents one rating of one movie by one user, and has the following format:

```
userId,movieId,rating,timestamp
```

The lines within this file are ordered first by `userId`, then, within user, by `movieId`.

Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars).

Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

## Tags Data File Structure (tags.csv)

All tags are contained in the file `tags.csv`. Each line of this file after the header row

represents one tag applied to one movie by one user, and has the following format:

```
userId,movieId,tag,timestamp
```

The lines within this file are ordered first by `userId`, then, within user, by `movieId`.

Tags are user-generated metadata about movies. Each tag is typically a single word or short phrase. The meaning, value, and purpose of a particular tag is determined by each user.

Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

## Movies Data File Structure (movies.csv)

Movie information is contained in the file `movies.csv`. Each line of this file after the header row represents one movie, and has the following format:

```
movieId,title,genres
```

Movie titles are entered manually or imported from <https://www.themoviedb.org/>, and include the year of release in parentheses. Errors and inconsistencies may exist in these titles.

Genres are a pipe-separated list, and are selected from the following:

- Action
- Adventure
- Animation
- Children's
- Comedy
- Crime
- Documentary
- Drama
- Fantasy
- Film-Noir
- Horror
- Musical
- Mystery
- Romance
- Sci-Fi
- Thriller
- War
- Western
- (no genres listed)

## Links Data File Structure (links.csv)

Identifiers that can be used to link to other sources of movie data are contained in the file `links.csv`. Each line of this file after the header row represents one movie, and has

the following format:

`movieId,imdbId,tmdbId`

`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` 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` 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>.

Use of the resources listed above is subject to the terms of each provider.

## Cross-Validation

Prior versions of the MovieLens dataset included either pre-computed cross-folds or scripts to perform this computation. We no longer bundle either of these features with the dataset, since most modern toolkits provide this as a built-in feature. If you wish to learn about standard approaches to cross-fold computation in the context of recommender systems evaluation, see [LensKit](#) for tools, documentation, and open-source code examples.