Philipps Universität Marburg

Data Integration

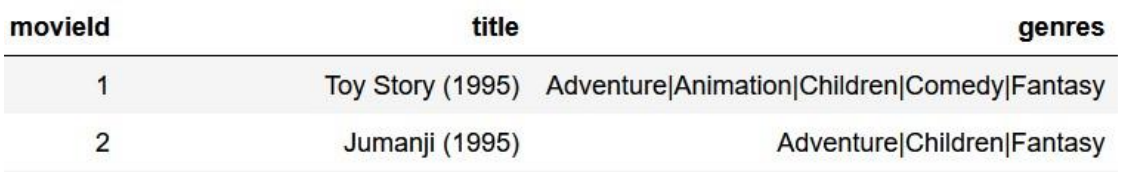
urban-octo-spoon
Step1: Preparation

Isabella Lisica

Nureldien Gebril

Eyad Hamed

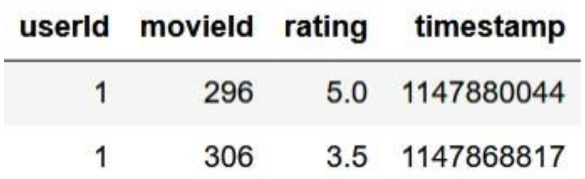
1:Documentation for source datasets: Movies



movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy

- **62.423 unique entries**
- **3 Attributes:** movieId, title, genres
- **Source:** GroupLens
- **Link:** https://grouplens.org/datasets/movielens/25m/
- **Other information:** the (year) when the movie was released has to be removed from the „title“ column in order to be able to connect the table with other tables.

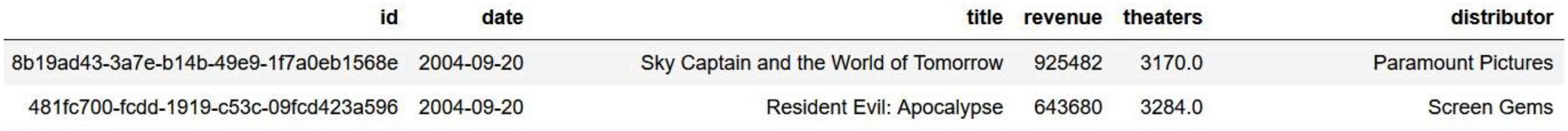
1:Documentation for source datasets: Rating



userId	movieId	rating	timestamp
1	296	5.0	1147880044
1	306	3.5	1147868817

- **25000095 unique entries**
- **4 Attributes:** userId, movieId, rating, timestamp
- **Source:** GroupLens
- **Link:** https://grouplens.org/datasets/movielens/25m/

1:Documentation for source datasets: Revenue per day



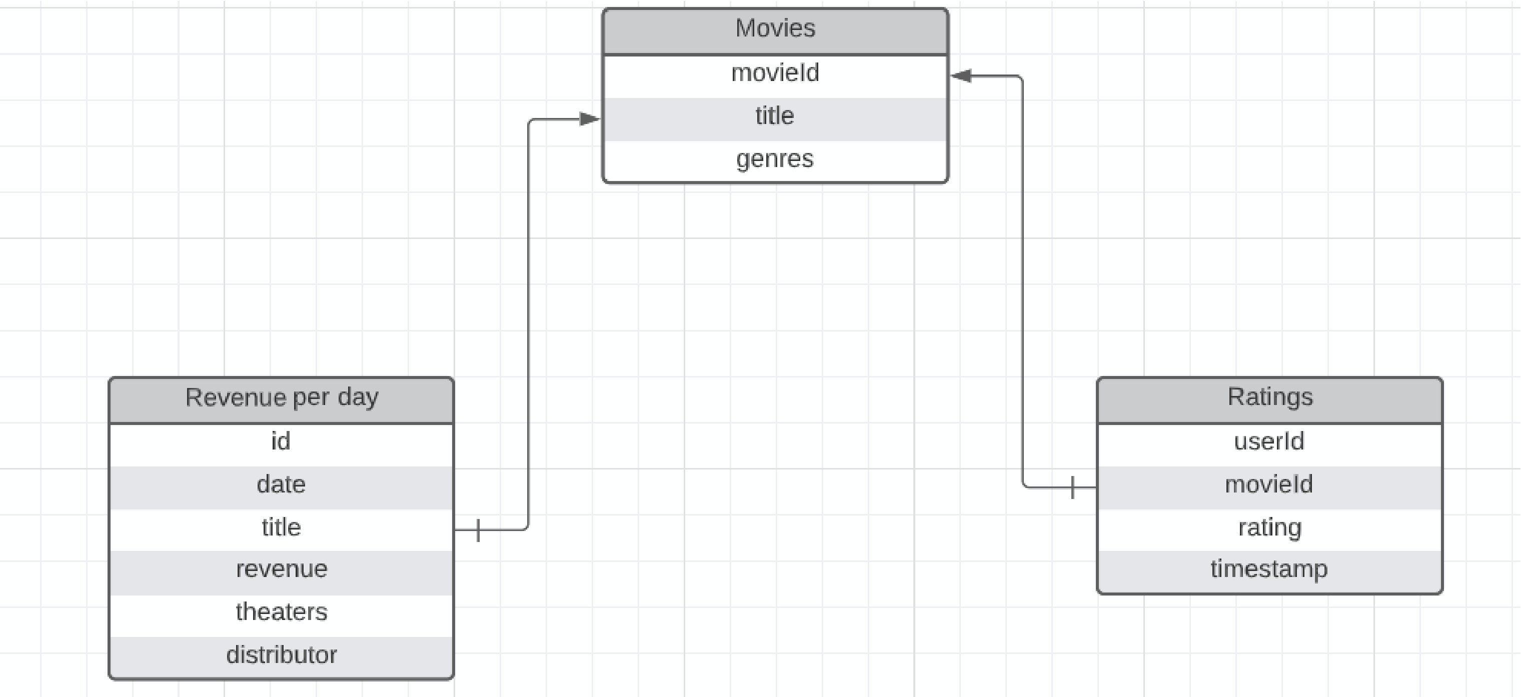
id	date	title	revenue	theaters	distributor
8b19ad43-3a7e-b14b-49e9-1f7a0eb1568e	2004-09-20	Sky Captain and the World of Tomorrow	925482	3170.0	Paramount Pictures
481fc700-fcdd-1919-c53c-09fcd423a596	2004-09-20	Resident Evil: Apocalypse	643680	3284.0	Screen Gems

- **339536 unique entries**
- **6 Attributes:** id, date, title, revenue, theaters, distributor
- **Source:** Boxofficemojo
- **Link:** https://github.com/tjwaterman99/boxofficemojo-scraper

Showcases

1. We aim to provide a general overview of filem data
2. We want to extract ratings for specific movies
3. We want to determine if there is a correlation between a movie‘s rating and it‘s revenue

Integrated schema draft



Movies
- movieId
- title
- genres

Revenue per day
- id
- date
- title
- revenue
- theaters
- distributor

Ratings
- userId
- movieId
- rating
- timestamp