

Using SQL

Analysis of MovieLens Dataset

By - Ankit Rai



Description

This SQL project leverages the extensive MovieLens 32M dataset, which captures user interactions with movies through 5-star ratings and free-text tagging. With over 32 million ratings and 2 million tag applications across nearly 88,586 movies, the dataset is a rich resource for exploring user preferences and movie trends.

Dataset Overview

- Ratings: 32,000,204 ratings
- Tags: 2,000,072 tag applications
- Movies: 88,585 movies
- Users: 200,948 users
- Timeframe: Data collected from January 9, 1995, to October 12, 2023

The project will employ SQL techniques such as JOINS, GROUP BY clauses, and aggregate functions to analyze the dataset effectively. The insights derived will inform potential enhancements to movie recommendation systems based on user behaviors and preferences.

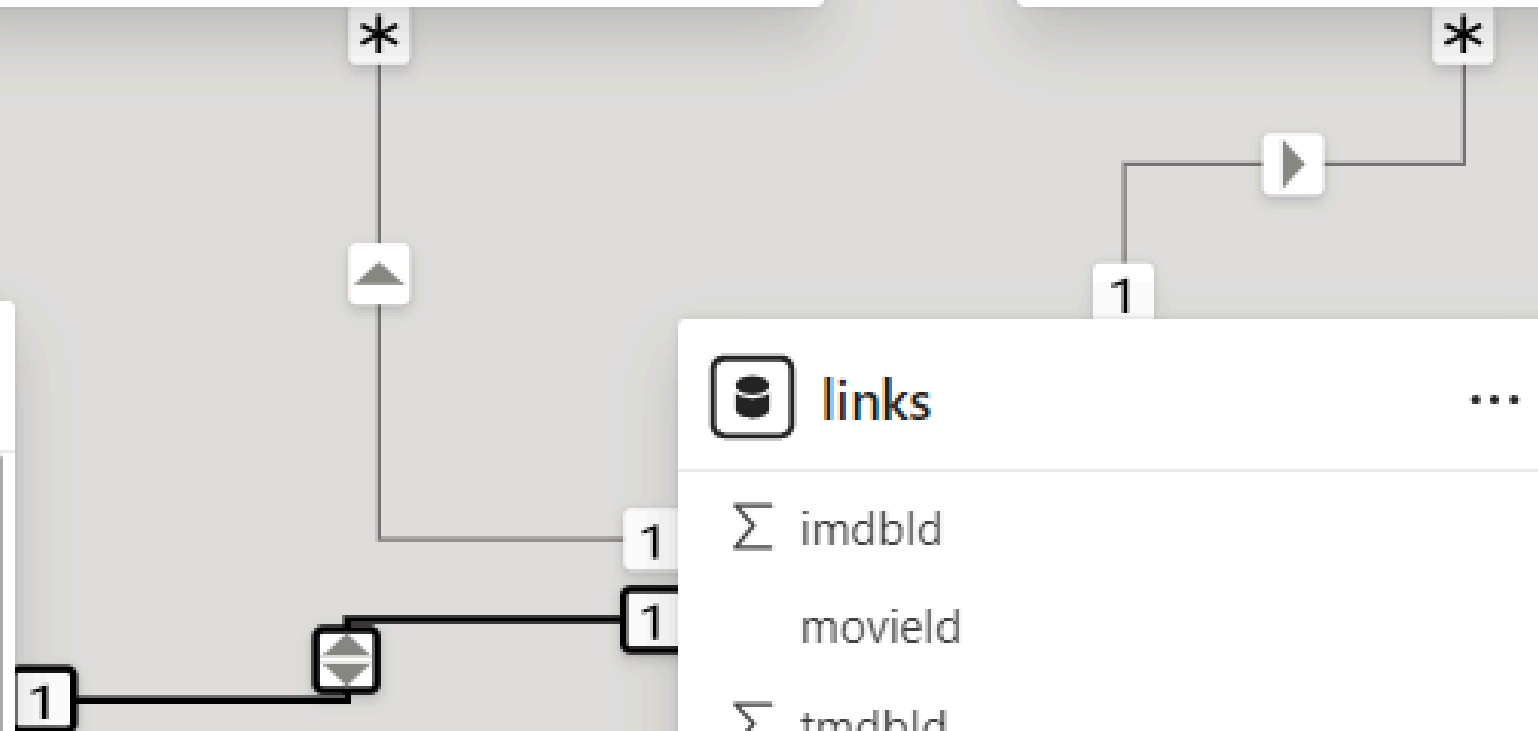
Schemas

movies
genres
movied
Σ Release_year
Title
Collapse ^

ratings
Date
movied
Σ rating
Time
Σ userId
Collapse ^



tags
Date
movied
tag
time
Σ userId
Collapse ^

links
Σ imdbld
movied
Σ tmdbld
Collapse ^



Query to find distribution of ratings (1 to 5)

```
SELECT
    rating,
    COUNT(*) AS rating_count
FROM
    ratings
GROUP BY
    rating
ORDER BY
    rating;
```

	rating double precision 	rating_count bigint 
1	0.5	16804
2	1	30397
3	1.5	16703
4	2	64979
5	2.5	51533
6	3	193637
7	3.5	138851
8	4	279757
9	4.5	101556
10	5	154358



Query to find the top 10 most used tag

```
SELECT
    tag,
    COUNT(*) AS tag_count
FROM
    tags
GROUP BY
    tag
ORDER BY
    tag_count DESC
LIMIT 10;
```

	tag character varying (800) 🔒	tag_count bigint 🔒
1	sci-fi	5089
2	atmospheric	4516
3	action	4243
4	comedy	3850
5	funny	3775
6	based on a book	3569
7	visually appealing	3368
8	surreal	3324
9	twist ending	3098
10	violence	2855

Query to find top 10 user who tagged the most movies

```
SELECT
    userId,
    COUNT(DISTINCT movieId) AS tagged_movies
FROM
    tags
GROUP BY
    userId
ORDER BY
    tagged_movies DESC
LIMIT 10;
```

	userid 	tagged_movies 
1	78213	15769
2	68821	7546
3	17035	2739
4	23053	2522
5	46025	1494
6	28604	1259
7	37008	1234
8	62453	1172
9	47557	1158
10	57737	1150



Query to find top 10 movies based on Average Ratings

```
SELECT r.movieid,m.title
FROM
    movies m
JOIN
    ratings r on m.movieid = r.movieid
GROUP BY r.movieid, m.title
ORDER BY AVG(rating) DESC
LIMIT 10;
```

	movieid bigint	title character varying (500)
1	141064	Uomo e galantuomo
2	122772	Love Birds
3	189393	97% Owned
4	283097	Teen Wolf: The Movie
5	51053	Ringers: Lord of the Fans
6	135350	Toto vs the Four
7	169370	Kathleen Madigan: Bothering Jesus
8	277650	We Are the World: The Story Behind the Song
9	223908	Lung
10	250764	The Girl and the Spider

Query to find the date in which most movies and least movies were tagged

```
SELECT rating_date, total_movies_rated, 'Highest' AS rating_type
FROM (
  SELECT
    DATE(date) AS rating_date,
    COUNT(DISTINCT movieId) AS total_movies_rated
  FROM
    ratings
  GROUP BY
    rating_date
  ORDER BY
    total_movies_rated DESC
  LIMIT 1
) AS highest_rating
UNION ALL
SELECT rating_date, total_movies_rated, 'Lowest' AS rating_type
FROM (
  SELECT
    DATE(date) AS rating_date,
    COUNT(DISTINCT movieId) AS total_movies_rated
  FROM
    ratings
  GROUP BY
    rating_date
  ORDER BY
    total_movies_rated ASC
  LIMIT 1
) AS lowest_rating;
```

	rating_date 	total_movies_rated 	rating_type 
1	2021-04-25	3552	Highest
2	1996-03-24	1	Lowest

Query to find customer who did both Rating & Tags

```
SELECT m.title,r.userid
FROM
    movies m
JOIN
    ratings r ON m.movieid=r.movieid
JOIN
    tags t ON m.movieid=t.movieid
where r.userid=t.userid;
```

	<div>title</div> <div>character varying (500)</div>	<div>userid</div> <div>integer</div>
1	Saturn in Opposition	242
2	Skyfall	58
3	Skyfall	58
4	Skyfall	58
5	Skyfall	58
6	Skyfall	58
7	Skyfall	58
8	Skyfall	58
9	Skyfall	58
10	Skyfall	58
11	Skyfall	58
12	Skyfall	58
13	Skyfall	58
14	Skyfall	58
15	Skyfall	58
16	Skyfall	58
17	Skyfall	58
18	Skyfall	58
19	Skyfall	58
20	Skyfall	58
21	Skyfall	58
22	Skyfall	58
Total rows: 1000 of 52305 Query complete 00:00:00.779 Ln 85, Col 1		

Query to get the highest rated movie by each user

```
SELECT r.userid,m.title, rating
FROM ratings r join movies m on m.movieid=r.movieid
AND rating = (
    SELECT MAX(rating)
    FROM ratings)
group by r.userid,m.title,rating
;
```

	userid integer	title character varying (500)	rating double precision
1	1	12 Angry Men	5
2	1	Airport	5
3	1	Aliens	5
4	1	All About Eve	5
5	1	American History X	5
6	1	Amistad	5
7	1	Apartment	5
8	1	Apocalypse Now	5
9	1	Back to the Future	5
10	1	Best Years of Our Lives	5
11	1	Big	5
12	1	Blade Runner	5
13	1	Boot	5
14	1	Breakfast Club	5
15	1	Citizen Kane	5
16	1	Crimes and Misdemeanors	5
17	1	Dangerous Liaisons	5
18	1	Dark City	5
19	1	Deer Hunter	5
20	1	Doom Generation	5
21	1	Eat Drink Man Woman	5
22	1	Election	5
Total rows: 1000 of 154103 Query complete 00:00:00.525 Ln 105, Col 1			

Query to analyze the distribution of ratings over time

```
SELECT
    Date AS rating_date,
    AVG(rating) AS average_rating,
    COUNT(*) AS total_ratings
FROM
    ratings
GROUP BY
    rating_date
ORDER BY
    rating_date;
```

	rating_date date	average_rating double precision	total_ratings bigint
1	1996-02-28	4.097560975609756	41
2	1996-03-01	3.459119496855346	159
3	1996-03-02	4.166666666666667	18
4	1996-03-03	4	3
5	1996-03-10	4.5	2
6	1996-03-17	4.333333333333333	3
7	1996-03-24	5	1
8	1996-03-26	4.0625	32
9	1996-03-28	4.086956521739131	46
10	1996-03-29	4.323529411764706	34
11	1996-03-30	3.909090909090909	22
12	1996-04-01	3.8974358974358974	39
13	1996-04-02	3.9622641509433962	106
14	1996-04-03	4.357142857142857	14
15	1996-04-04	4.153846153846154	26
16	1996-04-06	4.173913043478261	23
17	1996-04-07	4.027777777777778	36
18	1996-04-08	4.067796610169491	59
19	1996-04-11	4.107142857142857	84
20	1996-04-12	5	1
21	1996-04-13	4.2631578947368425	19
22	1996-04-14	3.8380952380952382	210
Total rows: 1000 of 9594 Query complete 00:00:00.246 Ln 14, Col 16			




Query to find highest rated movies and their genres

```
SELECT
  m.movieId,
  m.title,
  m.genres,
  AVG(r.rating) AS average_rating,
  COUNT(r.rating) AS total_ratings
FROM
  movies m
JOIN
  ratings r ON m.movieId = r.movieId
GROUP BY
  m.movieId
HAVING
  COUNT(r.rating) > 50
ORDER BY
  AVG(r.rating) DESC
LIMIT 10;
```

	movieid [PK] bigint	title character varying (500)	genres character varying (800)	average_rating double precision	total_ratings bigint
1	171011	Planet Earth II	Documentary	4.578947368421052	76
2	159817	Planet Earth	Documentary	4.454954954954955	111
3	170705	Band of Brothers	Action Drama War	4.445652173913044	92
4	318	Shawshank Redemption	Crime Drama	4.421489971346705	3490
5	202439	Parasite	Comedy Drama	4.347043701799486	389
6	858	Godfather	Crime Drama	4.30248557478917	2253
7	50	Usual Suspects	Crime Mystery Thriller	4.298006068487213	2307
8	1203	12 Angry Men	Drama	4.2761394101876675	746
9	195159	Spider-Man: Into the Spider-Verse	Action Adventure Animation Sci-Fi	4.2688953488372094	344
10	1221	Godfather: Part II	Crime Drama	4.26022176022176	1443

Query to find the 10 most active users based on the number of ratings & tags

```
SELECT
  userId,
  COUNT(DISTINCT movieId) AS rated_movies,
  (SELECT COUNT(*) FROM tags t WHERE t.userId = r.userId) AS tagged_movies
FROM
  ratings r
GROUP BY
  userId
ORDER BY
  rated_movies DESC
LIMIT 10;
```

	userid 	rated_movies 	tagged_movies 
1	5029	3893	2424
2	1048	3527	0
3	4392	3525	6
4	6728	3389	0
5	3367	3137	65
6	3408	3125	1
7	6310	2946	0
8	6471	2875	0
9	28	2842	0
10	1668	2591	0

Query to find 10 movies with the most ratings in specific genre

```
SELECT
    m.movieId,
    m.title,
    m.genres,
    COUNT(r.rating) AS total_ratings
FROM
    movies m
JOIN
    ratings r ON m.movieId = r.movieId
WHERE
    m.genres LIKE '%Comedy%'
GROUP BY
    m.movieId
ORDER BY
    total_ratings DESC
LIMIT 10; |
```

	movieid [PK] bigint	title character varying (500)	genres character varying (800)	total_ratings bigint
1	356	Forrest Gump	Comedy Drama Romance War	3380
2	296	Pulp Fiction	Comedy Crime Drama Thriller	3353
3	1	Toy Story	Adventure Animation Children Comedy Fantasy	2328
4	1270	Back to the Future	Adventure Comedy Sci-Fi	2058
5	608	Fargo	Comedy Crime Drama Thriller	1909
6	4306	Shrek	Adventure Animation Children Comedy Fantasy Roman...	1849
7	588	Aladdin	Adventure Animation Children Comedy Musical	1693
8	6539	Pirates of the Caribbean: The Curse of the Black Pea...	Action Adventure Comedy Fantasy	1658
9	1580	Men in Black	Action Comedy Sci-Fi	1643
10	380	True Lies	Action Adventure Comedy Romance Thriller	1612

Query to find the correlation between the number of tags and average ratings

```
SELECT
    m.movieId,
    m.title,
    COUNT(t.tag) AS total_tags,
    AVG(r.rating) AS average_rating
FROM
    movies m
LEFT JOIN
    ratings r ON m.movieId = r.movieId
LEFT JOIN
    tags t ON m.movieId = t.movieId
GROUP BY
    m.movieId
ORDER BY
    total_tags DESC;
```

	movieId [PK] bigint	title character varying (500)	total_tags bigint	average_rating double precision
1	296	Pulp Fiction	11665087	4.22457500745601
2	260	Star Wars: Episode IV - A New Hope	9779271	4.1161668390210275
3	318	Shawshank Redemption	8651710	4.421489971346705
4	2571	Matrix	8270265	4.1732036397866334
5	2959	Fight Club	6978900	4.249418604651162
6	356	Forrest Gump	6097520	4.043343195266273
7	593	Silence of the Lambs	5724745	4.17056305564702
8	79132	Inception	5636043	4.152628892291986
9	2858	American Beauty	3411941	4.093764223941739
10	4226	Memento	3322800	4.1402777777777775
11	4993	Lord of the Rings: The Fellowship of the Ring	3313852	4.088545897644192
12	109487	Interstellar	3282445	4.160354552780016
13	7153	Lord of the Rings: The Return of the King	3161466	4.077816747040772
14	47	Seven	3060982	4.098427073403241
15	589	Terminator 2: Judgment Day	3013116	3.9801381692573403
16	858	Godfather	2883840	4.30248557478917
17	58559	Dark Knight	2783232	4.163330078125
18	541	Blade Runner	2701350	4.10823754789272
19	7361	Eternal Sunshine of the Spotless Mind	2648220	4.075907590759076
20	1196	Star Wars: Episode V - The Empire Strikes Back	2636200	4.1571428571428575
21	527	Schindler's List	2495610	4.2518518518518515
22	480	Jurassic Park	2429805	3.7279843444227008

Total rows: 1000 of 87585 Query complete 00:01:10.657 Ln 146, Col 4

**The
End**