ADVANCED SQL
REINFORCEMENT PROJECT
IMDB DATASET

Submitted By:

SANJAIKUMAR.S

07/04/2025

DATA ANALYTICS & DATA SCIENCE

February-2025

QUERIES TO BE PERFORMED:

1. *Count the total number of records in each table of the database.*

Query:

```
select count(*)from director_mapping;
select count(*)from genre;
select count(*)from movie;
select count(*)from ratings;
select count(*)from role_mapping;
```

Output:

count(*)
14662

count(*)
7997

count(*)
7997

count(*)
15615

2. *Identify which columns in the movie table contain null values.*

Query:

SELECT * FROM movie WHERE id IS NULL OR title IS NULL OR year IS NULL OR date_published IS NULL OR duration IS NULL OR country IS NULL OR worlwide_gross_income IS NULL OR languages IS NULL OR production_company IS NULL;

Output:

id	title	year	date_published	duration	country	worlwide_gross_income	languages	production_company
tt0069049	The Other Side of the Wind	2018	2018-11-02	122	France, Iran, USA	NULL	English, German	Royal Road Entertainment
tt0071145	Ankur	2019	2019-01-02	131	India	NULL	Hindi	Blaze Film Enterprises
tt0082620	Kiss Daddy Goodbye	2018	2018-11-23	92	USA	NULL	English	Pendragon Film
tt0085953	Mo tai	2019	2019-10-22	84	Hong Kong	NULL	Cantonese	Lo Wei Motion Picture Company
tt0095857	Pestonjee	2019	2019-02-22	125	India	NULL	Hindi	National Film Development Corporation of India ...
tt0097268	Ek Din Achanak	2018	2018-12-30	105	India	NULL	Hindi	National Film Development Corporation

3. *Determine the total number of movies released each year, and analyze how the trend changes month-wise.*

Query:

SELECT YEAR(date_published) AS release_year, MONTH(date_published) AS release_month, COUNT(*) AS total_movies FROM movie WHERE date_published IS NOT NULL GROUP BY release_year, release_month ORDER BY release_year, release_month;

Output:

release_year	release_month	total_movies
2017	1	291
2017	2	228
2017	3	298
2017	4	249
2017	5	205
2017	6	226
2017	7	188
2017	8	246
2017	9	327
2017	10	303
2017	11	276
2017	12	215

release_year	release_month	total_movies
2018	1	302
2018	2	215
2018	3	285
2018	4	247
2018	5	229
2018	6	193
2018	7	167
2018	8	247
2018	9	276
2018	10	324
2018	11	252
2018	12	207

4. *How many movies were produced in either the USA or India in the year 2019?*

Query:

SELECT COUNT(*) AS TotalMovies FROM movie WHERE country IN ('USA') AND year = 2019;

Output:

TotalMovies
592

5. *List the unique genres in the dataset, and count how many movies belong exclusively to one genre.*

Query:

SELECT genre, COUNT(movie_id) AS MovieCount
FROM genre GROUP BY genre HAVING COUNT
(movie_id) ;

Output:

genre	MovieCount
Drama	4285
Fantasy	342
Thriller	1484
Comedy	2412
Horror	1208
Family	302
Romance	906
Adventure	591
Action	1289
Sci-Fi	375
Crime	813
Mystery	555
Others	100

6. *Which genre has the highest total number of movies produced?*

Query:

SELECT genre , COUNT(movie_id) AS TotalMovies FROM genre
GROUP BY genre ORDER BY TotalMovies DESC limit 3;

Output:

genre	TotalMovies
Drama	4285
Comedy	2412
Thriller	1484

7.*Calculate the average movie duration for each genre.*

Query:

SELECT g.genre, AVG(m.duration) AS AvgDuration FROM

movie m JOIN genre g ON m.id = g.movie_id GROUP BY g.genre;

Output:

genre	AvgDuration
Drama	106.7746
Fantasy	105.1404
Thriller	101.5761
Comedy	102.6227
Horror	92.7243
Family	100.9669
Romance	109.5342
Adventure	101.8714
Action	112.8829
Sci-Fi	97.9413
Crime	107.0517
Mystery	101.8000
Others	100.1600

8. *Identify actors or actresses who have appeared in more than three movies with an average rating below 5.*

Query:

SELECT n.name, COUNT(rm.movie_id) AS MovieCount FROM role_mapping rm JOIN names n ON rm.name_id = n.id

JOIN ratings r ON rm.movie_id = r.movie_id WHERE rm.category = 'actor' AND r.avg_rating < 5 GROUP BY n.name HAVING COUNT(rm.movie_id) > 3;

Output:

name	MovieCount
Michael Madsen	4
Tom Sizemore	7
Danny Trejo	4
Eric Roberts	5
Lee Bane	4
Dolph Lundgren	4
Derek Nelson	4

9. *Find the minimum and maximum values for each column in the ratings table, excluding the movie_id column.*

Query:

SELECT

MIN(avg_rating) AS MinRating, MAX(avg_rating) AS MaxRating,

MIN(total_votes) AS MinVotes, MAX(total_votes) AS MaxVotes,

MIN(median_rating) AS MinMedian, MAX(median_rating) AS MaxMedian

FROM ratings;

Output:

MinRating	MaxRating	MinVotes	MaxVotes	MinMedian	MaxMedian
1.0	10.0	100	725138	1	10

10. *Which are the top 10 movies based on their average rating?*

Query:

```
SELECT  m.title, r.avg_rating FROM movie m
JOIN ratings r ON m.id = r.movie_id
ORDER BY r.avg_rating DESC limit 10;
```

Output:

title	avg_rating
Kirket	10.0
Love in Kilnerry	10.0
Gini Helida Kathe	9.8
Runam	9.7
Fan	9.6
Android Kunjappan Version 5.25	9.6
Safe	9.5
The Brighton Miracle	9.5
Yeh Suhaagraat Impossible	9.5
Shibu	9.4

11. *Summarize the ratings table by grouping movies based on their median ratings.*

Query:

```
SELECT median_rating, COUNT(movie_id) AS TotalMovies
FROM ratings GROUP BY median_rating;
```

Output:

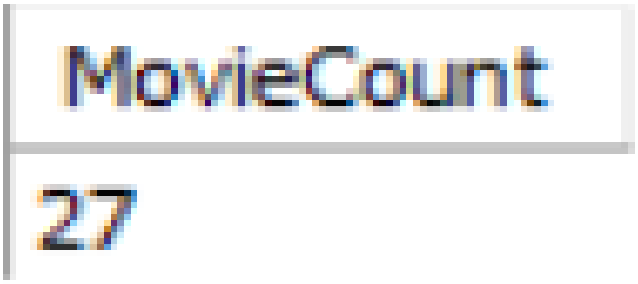
median_rating	TotalMovies
8	1030
7	2257
3	283
6	1975
9	429
2	119
4	479
5	985
10	346
1	94

12. *How many movies, released in March 2017 in the USA within a specific genre, had more than 1,000 votes?*

Query:

```
SELECT COUNT(*) AS MovieCount FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE year = 2017 AND MONTH(date_published) = 3
AND country = 'USA' AND r.total_votes > 1000;
```

Output:



MovieCount
27

13. *Find movies from each genre that begin with the word "The" and have an average rating greater than 8.*

Query:

```
SELECT m.title, g.genre, r.avg_rating FROM movie m
JOIN genre g ON m.id = g.movie_id JOIN ratings r ON m.id = r.movie_id WHERE
m.title LIKE 'The%' AND r.avg_rating > 8;
```

Output:

title	genre	avg_rating
The Blue Elephant 2	Drama	8.8
The Blue Elephant 2	Horror	8.8
The Blue Elephant 2	Mystery	8.8
The Brighton Miracle	Drama	9.5
The Irishman	Crime	8.7
The Irishman	Drama	8.7
The Colour of Darkness	Drama	9.1
Theeran Adhigaaram Ondru	Action	8.3
Theeran Adhigaaram Ondru	Crime	8.3
Theeran Adhigaaram Ondru	Thriller	8.3
The Mystery of Godliness: ...	Drama	8.5
The Gambinos	Crime	8.4
The Gambinos	Drama	8.4
The King and I	Drama	8.2
The King and I	Romance	8.2

14. *Of the movies released between April 1, 2018, and April 1, 2019, how many received a median rating of 8?*

Query:

SELECT COUNT(*) AS MovieCount FROM movie m JOIN ratings r ON m.id = r.movie_id WHERE date_published BETWEEN '2018-04-01' AND '2019-04-01' AND r.median_rating = 8;

Output:

MovieCount
361

15. *Do German movies receive more votes on average than Italian movies?*

Query:

SELECT country, AVG(r.total_votes) AS AvgVotes FROM movie m JOIN ratings r ON m.id = r.movie_id WHERE country IN ('Germany', 'Italy') GROUP BY country;

Output:

country	AvgVotes
Germany	730.8904
Italy	633.8618

16. *Identify the columns in the names table that contain null values.*

Query:

SELECT 'name' AS ColumnName, COUNT(*) AS NullCount FROM names WHERE name IS NULL

UNION ALL

SELECT 'height', COUNT(*) FROM names WHERE height IS NULL

UNION ALL

SELECT 'date_of_birth', COUNT(*) FROM names WHERE date_of_birth IS NULL

UNION ALL

SELECT 'known_for_movies', COUNT(*) FROM names WHERE known_for_movies IS NULL;

Output:

ColumnName	NullCount
name	0
height	17335
date_of_birth	13431
known_for_movies	15226

17. *Who are the top two actors whose movies have a median rating of 8 or higher?*

Query:

```
SELECT n.name, COUNT(rm.movie_id) AS MovieCount
FROM role_mapping rm
JOIN names n ON rm.name_id = n.id
JOIN ratings r ON rm.movie_id = r.movie_id
WHERE r.median_rating >= 8 AND rm.category = 'actor'
GROUP BY n.name
ORDER BY MovieCount DESC LIMIT 2 ;
```

Output:

name	MovieCount
Mammootty	8
Mohanlal	5

18. *Which are the top three production companies based on the total number of votes their movies received?*

Query:

```
SELECT  m.production_company, SUM(r.total_votes) AS TotalVotes
FROM movie m
```

JOIN ratings r ON m.id = r.movie_id

GROUP BY m.production_company

ORDER BY TotalVotes DESC LIMIT 3 ;

Output:

production_company	TotalVotes
Marvel Studios	2656967
Twentieth Century Fox	2411163
Warner Bros.	2396057

19. *How many directors have worked on more than three movies?*

Query:

SELECT dm.name_id, n.name, COUNT(dm.movie_id) AS MovieCount

FROM director_mapping dm

JOIN names n ON dm.name_id = n.id

GROUP BY dm.name_id, n.name

HAVING COUNT(dm.movie_id) > 3;

Output:

name_id	name	MovieCount
nm6356309	Özgür Bakar	4
nm2691863	Justin Price	4
nm0814469	Sion Sono	4
nm0831321	Chris Stokes	4
nm2096009	Andrew Jones	5
nm0425364	Jesse V. Johnson	4
nm0001752	Steven Soderbergh	4
nm0515005	Sam Liu	4
nm1777967	A.L. Vijay	5

20. *Calculate the average height of actors and actresses separately.*

Query:

```
SELECT rm.category, AVG(n.height) AS AvgHeight
FROM role_mapping rm
JOIN names n ON rm.name_id = n.id
WHERE rm.category IN ('actor', 'actress')
GROUP BY rm.category;
```

Output:

category	AvgHeight
actor	162.1818
actress	162.4715

21. *List the 10 oldest movies in the dataset along with their title, country, and director.*

Query:

```
SELECT  m.title, m.country, n.name AS Director
FROM movie m
JOIN director_mapping dm ON m.id = dm.movie_id
JOIN names n ON dm.name_id = n.id
ORDER BY m.year ASC limit 10;
```

Output:

title	country	Director
Critical Eleven	Indonesia	Monty Tiwa
Critical Eleven	Indonesia	Robert Ronny
Deo Te-i-beul	South Korea	Jong-kwan Kim
Far til fire på toppen	Denmark, Norway	Martin Miehe-Renard
Recep Ivedik 5	Turkey	Togan Gökbakar
Brothers in Arms	USA	Caleb J. Phillips
Love Blossoms	Belgium, Canada	Jonathan Wright
Killer Christmas	USA	Tony Shaker
Mify	Russia	Aleksandr Molochnikov
Cheng feng po lang	China	Han Han

22. *List the top 5 movies with the highest total votes, along with their genres.*

Query:

```
SELECT  m.title, g.genre, r.total_votes
FROM movie m
JOIN ratings r ON m.id = r.movie_id
JOIN genre g ON m.id = g.movie_id
ORDER BY r.total_votes DESC limit 5;
```

Output:

title	genre	total_votes
Avengers: Infinity War	Action	725138
Avengers: Infinity War	Adventure	725138
Avengers: Infinity War	Sci-Fi	725138
Avengers: Endgame	Action	602792
Avengers: Endgame	Adventure	602792

23. *Identify the movie with the longest duration, along with its genre and production company.*

Query:

```
SELECT  m.title, g.genre, m.production_company, m.duration
```

```
FROM movie m
JOIN genre g ON m.id = g.movie_id
ORDER BY m.duration DESC limit 3;
```

Output:

title	genre	production_company	duration
La flor	Drama	El Pampero Cine	808
La flor	Fantasy	El Pampero Cine	808
Ang panahon ng halimaw	Drama	Epicmedia	234

24. *Determine the total number of votes for each movie released in 2018.*

Query:

```
SELECT m.title, SUM(r.total_votes) AS TotalVotes
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE year = 2018
GROUP BY m.title;
```

Output:

title	TotalVotes
Ihmisen osa	177
Breath	2155
Overboard	28649
Mortal Engines	85366
The Commuter	89240
A Wrinkle in Time	36925
Alligator X	677
On Chesil Beach	7335
Ready Player One	320181
Been So Long	656
Come Sunday	1654

25. *What is the most common language in which movies were produced?*

Query:

```
SELECT  languages, COUNT(*) AS MovieCount
FROM movie
GROUP BY languages
ORDER BY MovieCount DESC limit 3;
```

Output:

languages	MovieCount
English	3095
Spanish	274
French	260