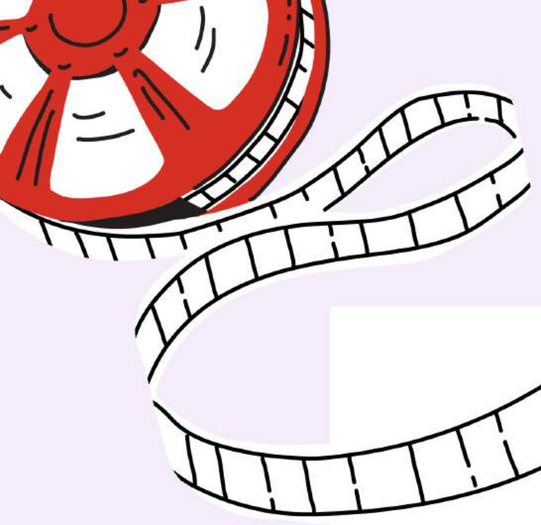
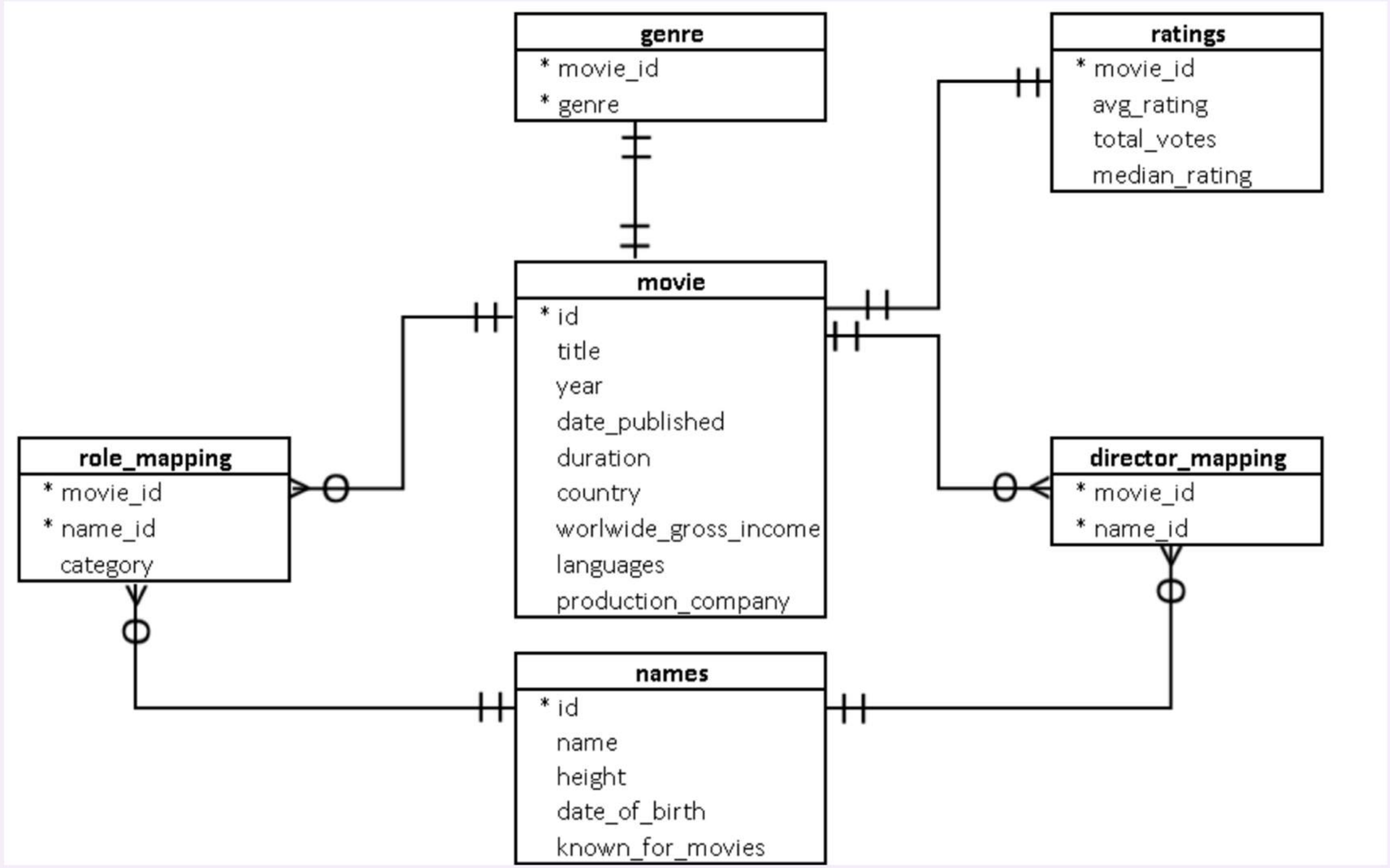


Sheelu Singh

Data-Driven Insights for RSVP Movies Using SQL

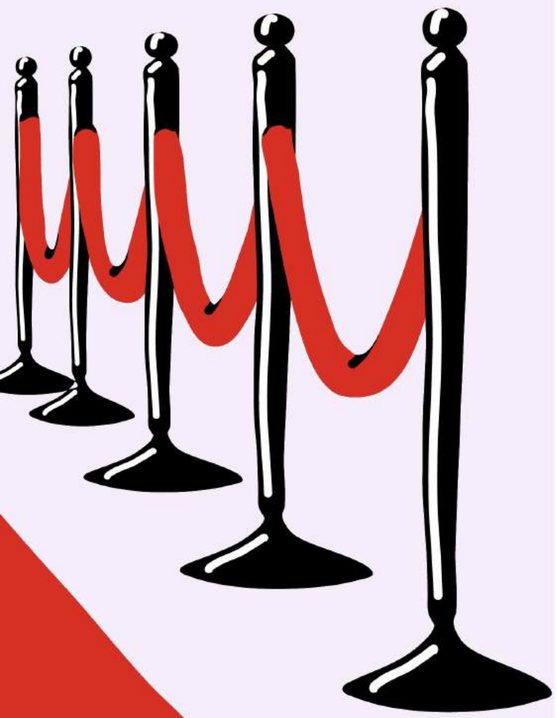
Sheelu Singh

Database Setup (ERD Diagram)



Sheelu Singh

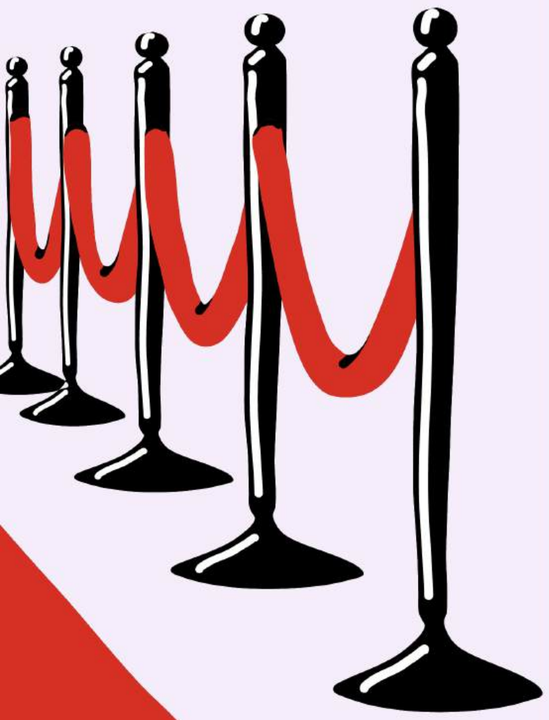
Topics covered in this project:



- Database Creation and Table Design (Database concepts, constraints)
- Data Retrieval and Filtering (Basics SQL)
- Aggregations and Grouping
- Joining Tables (Left Join, Inner Join)
- Window Functions(Row Number, Rank, Dense Rank, Lead, Lag)
- Date and Time Functions
- Conditional Logic
- Subqueries
- Ranking and Ordering
- Data Segmentation

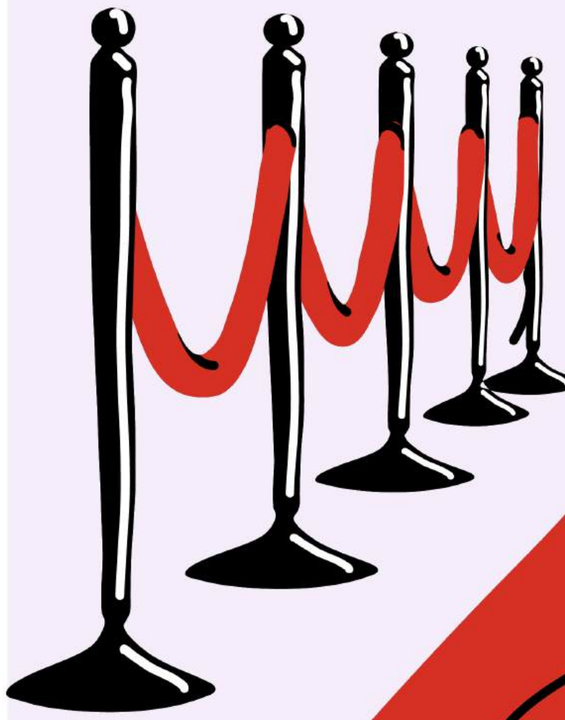
Sheelu Singh

Purpose of the project:



Writing 20 complex SQL queries showcasing problems solving and decision-making skills!

- Target audience segmentation
- Preferred genres and languages
- Optimal release periods
- Key success metrics for global hits



Sheelu Singh

Let's Start!

Sheelu Singh

Business Problem 1

Find the total number of rows in each table of the schema?

Solution

```sql
SELECT COUNT(m.id) AS rows_in_movies,
       COUNT(n.id) AS rows_in_names,
       COUNT(rm.movie_id) AS rows_in_role_mapping,
       COUNT(g.movie_id) AS rows_in_genre,
       COUNT(r.movie_id) AS rows_in_ratings
FROM movie AS m
JOIN genre AS g
     ON m.id = g.movie_id
JOIN ratings AS r
     ON m.id = r.movie_id
JOIN director_mapping AS d
     ON m.id = d.movie_id
JOIN role_mapping AS rm
     ON m.id = rm.movie_id
JOIN names AS n
     ON n.id = rm.name_id;
```

Sheelu Singh

Business Problem 2

Which columns in the movie table have null values?

Solution

```sql
SELECT
    SUM(CASE WHEN title IS NULL THEN 1 ELSE 0 END) AS null_titles,
    SUM(CASE WHEN YEAR IS NULL THEN 1 ELSE 0 END) AS null_year,
    SUM(CASE WHEN date_published IS NULL THEN 1 ELSE 0 END) AS null_date_published,
    SUM(CASE WHEN duration IS NULL THEN 1 ELSE 0 END) AS null_duration,
    SUM(CASE WHEN worlwide_gross_income IS NULL THEN 1 ELSE 0 END) AS null_worldwide_gross_income,
    SUM(CASE WHEN languages IS NULL THEN 1 ELSE 0 END) AS null_language,
    SUM(CASE WHEN production_company IS NULL THEN 1 ELSE 0 END) null_production_comp
FROM movie;
```

Sheelu Singh

Business Problem 3

Find the total number of movies released each year? How does the trend look month wise?

Solution :

```
SELECT
  m.year,
  count(m.title) AS number_of_movies
FROM
  movie m
GROUP BY
  m.year
ORDER BY
  m.year;
SELECT
  month(m.date_published) AS month_num,
  count(m.title) AS number_of_movies
FROM
  movie m
GROUP BY
  MONTH(m.date_published)
ORDER BY
  MONTH(m.date_published);
```

Business Problem 4

How many movies were produced in the USA or India in the year 2019?

Solution

```sql
SELECT COUNT(id) as movies_in_2019
FROM movie
WHERE country = 'USA' OR country = 'India'
      AND year = '2019' ;
```

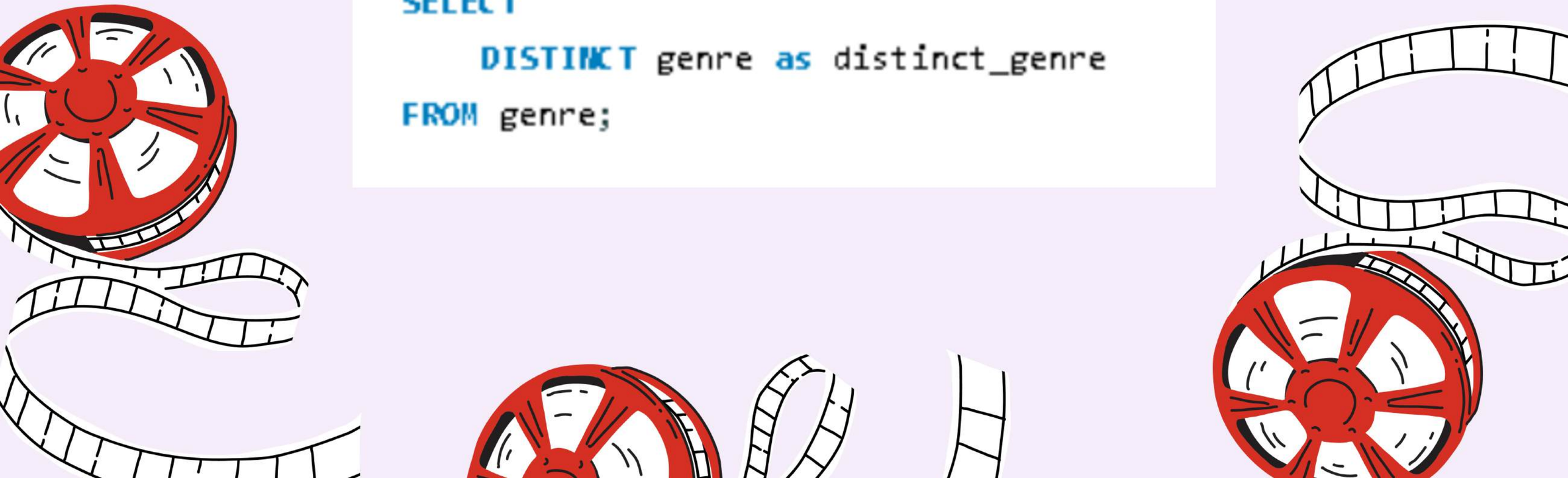
Sheelu Singh

Business Problem 5

Find the unique list of the genres present in the data set?

Solution

```
SELECT
    DISTINCT genre as distinct_genre
FROM genre;
```



Sheelu Singh

Business Problem 6

Which genre had the highest number of movies produced overall?

Solution

```sql
SELECT g.genre
FROM genre g
JOIN movie m ON g.movie_id = m.id
GROUP BY g.genre
ORDER BY COUNT(m.id) desc
LIMIT 1;
```

Business Problem 7

How many movies belong to only one genre?

Solution

```
SELECT COUNT(movie_id) as movie_count,
       g.genre
FROM genre AS g
GROUP BY g.genre;
```

Sheelu Singh

Business Problem 8

What is the average duration of movies in each genre?

Solution

```sql
SELECT
    g.genre,
    ROUND(AVG(m.duration),2) AS avg_duration
FROM genre AS g
JOIN movie AS m
    ON g.movie_id = m.id
GROUP BY g.genre;
```

Sheelu Singh

Business Problem 9

What is the rank of the ‘thriller’ genre of movies among all the genres in terms of number of movies produced?

Solution :

```sql
WITH GenreCounts AS (
    SELECT g.genre,
           COUNT(m.id) AS movie_count
    FROM genre g
    JOIN movie m ON g.movie_id = m.id
    GROUP BY g.genre
),
RankedGenres AS (
    SELECT genre,
           movie_count,
           RANK() OVER (ORDER BY movie_count DESC) AS genre_rank
    FROM GenreCounts
)
SELECT genre,
       movie_count,
       genre_rank
FROM RankedGenres
WHERE genre = 'thriller';
```

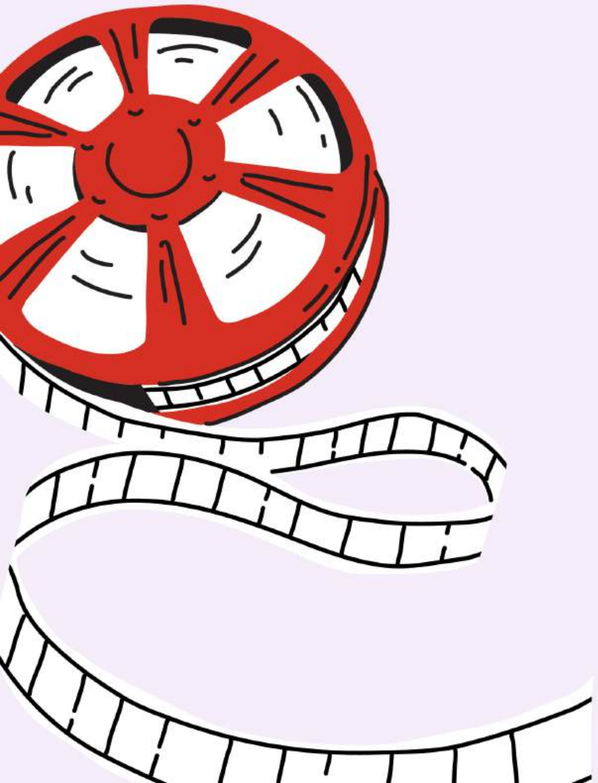
Sheelu Singh

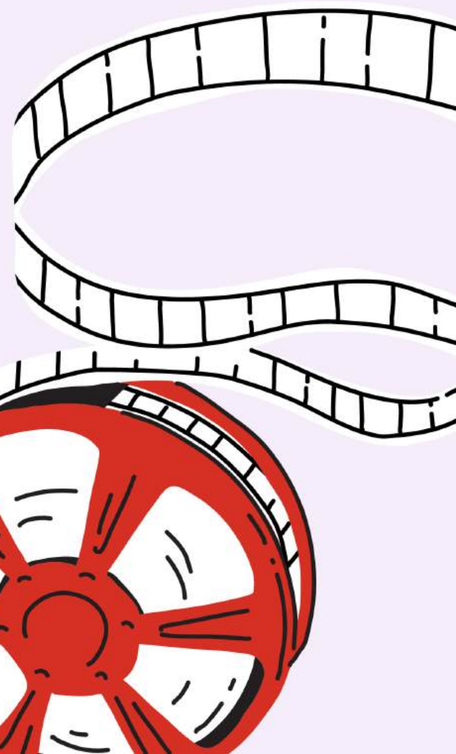
Business Problem 10

Find the minimum and maximum values in each column of the ratings table except the movie_id column?

Solution

```sql
SELECT MIN(avg_rating) AS min_avg_rating,
       MAX(avg_rating) AS max_avg_rating,
       MIN(total_votes) AS min_total_votes,
       MAX(total_votes) AS max_total_votes,
       MIN(median_rating) AS min_median_rating,
       MAX(median_rating) AS max_median_rating
FROM ratings;
```





Sheelu Singh

Business Problem 11

Which are the top 10 movies based on average rating?

Solution

```sql
SELECT
    m.title,
    avg_rating,
    RANK() OVER(ORDER BY r.avg_rating DESC) AS movie_rank
FROM movie as m
JOIN ratings as r
on m.id = r.movie_id
LIMIT 10;
```

Sheelu Singh

Business Problem 12

Summarize the ratings table based on the movie counts by median ratings

Solution

```sql
SELECT r.median_rating,
       COUNT(m.title) AS movie_count
FROM movie m
JOIN ratings r ON m.id = r.movie_id
GROUP BY r.median_rating
ORDER BY r.median_rating;
```

Sheelu Singh

Business Problem 13

Which production house has produced the most number of hit movies (average rating > 8)?

Solution

```sql
SELECT m.production_company, count(m.id) AS movie_count,
       RANK() OVER(ORDER BY count(m.id) DESC) AS prod_company_rank
FROM movie AS m
JOIN ratings AS r
     ON m.id = r.movie_id
WHERE r.avg_rating > 8 AND m.production_company IS NOT NULL
GROUP BY m.production_company
LIMIT 1;
```

Sheelu Singh

Business Problem 14

How many movies released in each genre during March 2017 in the USA had more than 1,000 votes?

Solution

```sql
SELECT g.genre,
    COUNT(m.id) as movie_count
    FROM genre AS g
JOIN movie AS m
ON m.id = g.movie_id
JOIN ratings AS r
ON m.id = r.movie_id
WHERE total_votes > 1000  AND
      m.country = "USA"
GROUP BY genre
ORDER BY movie_count DESC;
```

Sheelu Singh

Business Problem 15

Find movies of each genre that start with the word ‘The’ and which have an average rating > 8?

Solution

```sql
SELECT m.title, round(AVG(r.avg_rating),2) AS avg_rating, g.genre
FROM movie m
JOIN ratings r ON m.id = r.movie_id
JOIN genre g ON m.id = g.movie_id
WHERE m.title LIKE 'The%' AND r.avg_rating > 8
GROUP BY m.title, g.genre;

-- Trying with the median rating
SELECT m.title, round(AVG(r.median_rating),2) AS median_avg_rating, g.genre
FROM movie m
JOIN ratings r ON m.id = r.movie_id
JOIN genre g ON m.id = g.movie_id
WHERE m.title LIKE 'The%' AND r.median_rating > 8
GROUP BY m.title, g.genre;
```

Sheelu Singh

Business Problem 16

Of the movies released between 1 April 2018 and 1 April 2019, how many were given a median rating of 8?

Solution

```sql
SELECT
    COUNT(movie_id) AS number_of_movies
FROM
    ratings AS r
        JOIN
    movie AS m ON r.movie_id = m.id
WHERE
    median_rating = 8 AND m.date_published > '2018-04-01'
       AND m.date_published <'2019-04-01';
```

Sheelu Singh

Business Problem 17

Do German movies get more votes than Italian movies?

Solution

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

Sheelu Singh

Business Problem 18

Which columns in the names table have null values?

Solution

```
WITH nameNulls AS(
  SELECT
    count(*) AS name_nulls
  FROM
    names
  WHERE
    name IS NULL
),
heightNulls AS (
  SELECT
    count(*) AS height_nulls
  FROM
    names
  WHERE
    height IS NULL
),
DOBnulls AS (
  SELECT
    count(*) AS date_of_birth_nulls
  FROM
    names
  WHERE
    date_of_birth IS NULL
),
KFmoviesnulls AS (
  SELECT
    count(*) AS known_for_movies_nulls
  FROM
    names
  WHERE
    known_for_movies IS NULL
)
SELECT
  (
    SELECT
      name_nulls
    FROM
      nameNulls
  ) AS name_nulls,
  (
    SELECT
      height_nulls
    FROM
      heightNulls
  ) AS height_nulls,
  (
    SELECT
      date_of_birth_nulls
    FROM
      DOBnulls
  ) AS date_of_birth_nulls,
  (
    SELECT
      known_for_movies_nulls
    FROM
      KFmoviesnulls
  ) AS known_for_movies_nulls;
```

Sheelu Singh

Business Problem 19

Who are the top three directors in the top three genres whose movies have an average rating > 8?

Solution

```sql
SELECT n.name AS director_name,
       COUNT(dm.name_id) AS movie_count
FROM director_mapping AS dm
JOIN names AS n
     ON dm.name_id = n.id
JOIN ratings
     USING (movie_id)
WHERE avg_rating > 8
GROUP BY n.name
ORDER BY COUNT(dm.name_id) DESC
LIMIT 3;
```

Sheelu Singh

Business Problem 20

Who are the top two actors whose movies have a median rating >= 8?

Solution

```sql
SELECT
    n.name AS actor_name,
    COUNT(m.id) as movie_count
FROM
    names as n
JOIN director_mapping AS dm
ON n.id = dm.name_id
JOIN movie AS m
ON m.id = dm.movie_id
JOIN ratings AS r
ON m.id = r.movie_id
WHERE avg_rating >=8
GROUP BY n.name
ORDER BY movie_count DESC;
```

Sheelu Singh

Business Problem 21

Which are the top three production houses based on the number of votes received by their movies?

Solution

```sql
SELECT *, ROW_NUMBER() OVER () AS prod_comp_rank from (SELECT m.production_company AS production_company,
       SUM(r.total_votes) AS vote_count
       FROM  movie m
JOIN ratings r ON m.id = r.movie_id
GROUP BY m.production_company
ORDER BY vote_count desc) as a
LIMIT 3;
```

Sheelu Singh

Business Problem 22

Rank actors with movies released in India based on their average ratings. Which actor is at the top of the list?

Solution

```sql
SELECT n.name AS actor_name,
       SUM(r.total_votes) AS total_votes,
       COUNT(m.id) AS movie_count,
       ROUND(SUM(r.avg_rating * r.total_votes) /SUM(r.total_votes),2) AS actress_avg_rating,
       RANK() OVER(ORDER BY AVG(r.avg_rating) DESC, SUM(r.total_votes) DESC) AS actor_rank
FROM names AS n
JOIN role_mapping AS rm ON n.id = rm.name_id
JOIN movie AS m ON rm.movie_id = m.id
JOIN ratings AS r ON r.movie_id = m.id
GROUP BY n.name
HAVING COUNT(m.id) >= 5
ORDER BY actress_avg_rating DESC, total_votes DESC;
```

Business Problem 23

Find out the top five actresses in Hindi movies released in India based on their average ratings?

Solution

```sql
WITH hm as
(
SELECT *
FROM movie AS m
WHERE m.languages LIKE "%Hindi%"
AND country LIKE "%India%"
),
act_stats AS
(
SELECT
 n.name AS actress_name,
 SUM(r.total_votes) AS total_votes,
 count(*) AS movie_count,
 ROUND(SUM(r.avg_rating * r.total_votes) /SUM(r.total_votes),2) AS actress_avg_rating
FROM hm
 JOIN role_mapping AS rm
 ON hm.id = rm.movie_id
 JOIN names AS n
 ON rm.name_id = n.id
 JOIN ratings AS r
 ON hm.id = r.movie_id
WHERE rm.category = "actress"
GROUP BY n.name
HAVING movie_count >=3
)
SELECT *,
 DENSE_RANK() OVER(ORDER BY actress_avg_rating DESC, total_votes DESC) AS actress_rank
  FROM act_stats
LIMIT 5;
```

Sheelu Singh

Business Problem 24

Select thriller movies as per avg rating and classify them in the following category:

Rating > 8: Superhit movies

Rating between 7 and 8: Hit movies

Rating between 5 and 7: One-time-watch movies

Rating < 5: Flop movies

Solution :

```sql
SELECT
    m.title,
    CASE
        WHEN AVG(r.avg_rating) > 8 THEN 'Superhit movies'
        WHEN AVG(r.avg_rating) BETWEEN 7 AND 8 THEN 'Hit movies'
        WHEN AVG(r.avg_rating) BETWEEN 5 AND 7 THEN 'One-time-watch movies'
        ELSE 'Flop movies'
    END AS classification
FROM movie m
JOIN ratings r ON m.id = r.movie_id
JOIN genre g ON m.id = g.movie_id
WHERE g.genre = 'Thriller'
GROUP BY m.id, m.title
ORDER BY avg_rating DESC;
```

Sheelu Singh

Business Problem 25

What is the genre-wise running total and moving average of the average movie duration?

Solution :

```sql
WITH genre_avg_duration AS (
    SELECT
        g.genre,
        ROUND(AVG(m.duration), 2) AS avg_duration
    FROM
        movie AS m
    INNER JOIN
        genre AS g ON m.id = g.movie_id
    GROUP BY
        g.genre
    )
    SELECT
        genre,
        avg_duration,
        SUM(avg_duration) OVER (ORDER BY genre ROWS UNBOUNDED PRECEDING) AS running_total_duration,
        AVG(avg_duration) OVER (ORDER BY genre ROWS BETWEEN 10 PRECEDING AND CURRENT ROW) AS moving_avg_duration
    FROM
        genre_avg_duration
    ORDER BY
        genre;
```

Sheelu Singh

Business Problem 26

Which are the five highest-grossing movies of each year that belong to the top three genres?

Solution

```
WITH genre_count as (
  SELECT
    genre,
    COUNT(*) AS movie_count
  FROM
    genre
  Group by
    genre
  order by
    movie_count desc
  LIMIT
    3
), top_gross_movie as(
  SELECT
    g.genre,
    m.year,
    m.title as movie_name,
    m.worlwide_gross_income,
    ROW_NUMBER() OVER(
      PARTITION BY g.genre,
      m.year
      ORDER BY
        m.worlwide_gross_income DESC
    ) as movie_rank
  FROM
    movie as m
    JOIN genre as g ON m.id = g.movie_id
  WHERE
    g.genre IN (
      SELECT
        genre
      from
        genre_count
    )
)
SELECT
  genre,
  year,
  movie_name,
  worlwide_gross_income,
  movie_rank
FROM
  top_gross_movie
WHERE
  movie_rank <= 5
ORDER BY
  genre,
  year,
  movie_rank;
```

Sheelu Singh

Business Problem 27

Which are the top two production houses that have produced the highest number of hits (median rating >= 8) among multilingual movies?

Solution :

```sql
SELECT
  m.production_company,
  count(m.title) AS movie_count,
  RANK() OVER(
    ORDER BY
      count(m.title) DESC
  ) AS prod_comp_rank
FROM
  movie m
  JOIN ratings r ON m.id = r.movie_id
WHERE
  POSITION(',' IN m.languages)> 0
  AND r.median_rating >= 8
  AND production_company IS NOT NULL
GROUP BY
  m.production_company
ORDER BY
  movie_count DESC
LIMIT 2;
```

Sheelu Singh

Business Problem 28

Who are the top 3 actresses based on number of Super Hit movies (average rating >8) in drama genre?

Solution

```sql
WITH actress_summary AS (
    SELECT
        n.NAME AS actress_name,
        SUM(r.total_votes) AS total_votes,
        COUNT(r.movie_id) AS movie_count,
        ROUND(SUM(r.avg_rating * r.total_votes) / SUM(r.total_votes), 2) AS actress_avg_rating
    FROM
        movie AS m
    INNER JOIN
        ratings AS r ON m.id = r.movie_id
    INNER JOIN
        role_mapping AS rm ON m.id = rm.movie_id
    INNER JOIN
        names AS n ON rm.name_id = n.id
    INNER JOIN
        genre AS g ON m.id = g.movie_id
    WHERE
        rm.category = 'ACTRESS'
        AND r.avg_rating > 8
        AND g.genre = 'Drama'
    GROUP BY
        n.NAME
)
SELECT
    actress_name,
    total_votes,
    movie_count,
    actress_avg_rating,
    RANK() OVER (ORDER BY movie_count DESC) AS actress_rank
FROM
    actress_summary
ORDER BY
    actress_rank
LIMIT 3;
```

Sheelu Singh

Business Problem 29

Get the details for top 9 directors (based on number of movies)

Solution

```
WITH rankings AS  (SELECT dm.name_id,
    n.name,
    m.date_published,
    ROW_NUMBER() OVER (PARTITION BY n.name ORDER BY m.date_published) AS local_ranking,
    ROW_NUMBER() OVER (PARTITION BY n.name ORDER BY m.date_published) -1 AS local_ranking_lag1
FROM director_mapping as dm
JOIN movie as m
ON dm.movie_id = m.id

JOIN names as n
on dm.name_id = n.id

JOIN ratings as r
ON m.id = r.movie_id
),
avg_movie_days as (SELECT
    ra.name_id as director_id,
    ra.name as director_name,
  ROUND(AVG(datediff(rb.date_published, ra.date_published)),2) as avg_inter_movie_days
    from rankings as ra
    JOIN rankings as rb
    ON ra.name_id = rb.name_id AND
    ra.name = rb.name AND
    ra.local_ranking = rb.local_ranking_lag1
    GROUP BY ra.name_id, ra.name
    )
SELECT
    dm.name_id,
    n.name as director_name,
    AVG(r.avg_rating) as avg_rating,
    SUM(r.total_votes) as total_votes,
    MIN(r.avg_rating) as min_rating,
    MAX(r.avg_rating) as max_rating,
    COUNT(*) as movie_count,
    amd.avg_inter_movie_days,
    SUM(m.duration) as total_duration
FROM director_mapping as dm
JOIN movie as m
ON dm.movie_id = m.id
JOIN avg_movie_days as amd
ON dm.name_id = amd.director_id
```

```
JOIN names as n
on dm.name_id = n.id

JOIN ratings as r
ON m.id = r.movie_id
group by dm.name_id, n.name
ORDER BY movie_count desc
LIMIT 9;
```

Sheetu Singh

Conclusion

RSVP Movies' global expansion strategy should focus on drama and romance genres, which have shown high audience love. Collaborating with top directors like Joe and Anthony Russo for drama or James Mangold for sci-fi could increase the project's profile. Casting well-performing actresses such as Emma Stone, Rooney Mara, and actors as Kin Wah Chew, or Chris Hemsworth would likely increase market appeal. To penetrate the global market effectively, RSVP Movies should consider partnering with established studios like Marvel Studios, Twentieth Century, or Warner Bros. Additionally, the company should explore the potential benefits and terms of partnering with a major studios in other countries to enhance their global distribution capabilities and market presence.

Sheelu Singh

Thank You!