**Project 1: IMDb Movies Analysis using SQL**

Segment 1: Database - Tables, Columns, Relationships

1. What are the different tables in the database and how are they connected to each other in the database?

Answer: There are 6 tables in the database, namely movies, genre, ratings, director\_mapping, names and role\_mapping.

1. Find the total number of rows in each table of the schema

Answer: We can use the count function for each table. For example, the query would look like – SELECT COUNT(\*) FROM movies

This query would give us the count of rows for the movies table. We can run this query with each table in the database to get the total number of rows.

1. Identify which columns in the movie table have null values

Answer: We can use the IS NULL function with each column name in the movie table. The query would look like – SELECT \* from movies WHERE column\_name IS NULL

When we try and execute this query with the column names from the movie table, we can see which column has null values. The column named ‘worldwide\_gross\_revenue’ and ‘production\_company’ from the movie table has null values.

Segment 2: Movie Release Trends

1. Determine the total number of movies released each year and analyse the month-wise trend.

Answer: Query 1 - SELECT year, count(id) as number\_of\_movies from movies

Group by 1

Query 2 - select month(STR\_TO\_DATE(date\_published, '%m/%d/%Y')) as mnth, count(id) as cnt\_mov

from movies

group by mnth

order by mnth

1. Calculate the number of movies produced in the USA or India in the year 2019.

Answer: The query is- SELECT year, country, count(id) as count\_movies from movies

WHERE country IN ("USA","India") and year = 2019

GROUP BY 1,2

Segment 3: Production Statistics and Genre Analysis

1. Retrieve the unique list of genres present in the dataset.

Answer: The query is- SELECT distinct(genre) from genre

1. Identify the genre with the highest number of movies produced overall.

Answer: The query is- select genre, count(movie\_id) from genre

Group by 1

Order by 2 desc

limit 1

1. Determine the count of movies that belong to only one genre.

Answer: The query is- select count(movie\_id) from (select movie\_id, count(\*) as count\_movies from genre

group by 1

having count\_movies = 1) as sc

1. Calculate the average duration of movies in each genre.

Answer: The query is- select g.genre, AVG(m.duration) as Average\_Duration from genre g

JOIN

movies m

ON g.movie\_id = m.id

GROUP BY 1

1. Find the rank of the 'thriller' genre among all genres in terms of the number of movies produced.

Answer: The query is- select \* from (select genre, count(movie\_id), RANK() over (order by count(movie\_id) desc) as rnk from genre

Group by 1) as tbl

WHERE genre = "thriller"

Segment 4: Ratings Analysis and Crew Members

1. Retrieve the minimum and maximum values in each column of the ratings table (except movie\_id).

Answer: The query is- 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

1. Identify the top 10 movies based on average rating.

Answer: The query is- select m.title as movie\_name, r.avg\_rating as average\_rating from movies m join ratings r

ON m.id = r.movie\_id

order by 2 desc

limit 10

1. Summarise the ratings table based on movie counts by median ratings.

Answer: The query is- select count(m.id), r.median\_rating from movies m JOIN ratings r ON m.id = r.movie\_id

GROUP BY 2

ORDER BY 2

1. Identify the production house that has produced the most number of hit movies (average rating > 8).

Answer: The query is- select m.production\_company,count(m.id) from movies m JOIN ratings r ON m.id = r.movie\_id

Where r.avg\_rating > 8 AND m.production\_company IS NOT NULL

GROUP BY 1

ORDER BY 2 desc

LIMIT 1

OFFSET 1

1. Determine the number of movies released in each genre during March 2017 in the USA with more than 1,000 votes.

Answer: The query is- select g.genre,count(m.id) as total\_movies

from movies m

JOIN ratings r

ON m.id = r.movie\_id

JOIN genre g

ON g.movie\_id = m.id

where month(STR\_TO\_DATE(m.date\_published, '%m/%d/%Y')) = 3 AND m.year = 2017 AND r.total\_votes > 1000 AND m.country = 'USA'

group by 1

ORDER BY 2 DESC

1. Retrieve movies of each genre starting with the word 'The' and having an average rating > 8.

Answer: The query is- select g.genre,count(m.id) as total\_movies

from movies m

JOIN ratings r

ON m.id = r.movie\_id

JOIN genre g

ON g.movie\_id = m.id

WHERE m.title LIKE 'The%' AND avg\_rating > 8

group by 1

ORDER BY 2 DESC

Segment 5: Crew Analysis

.

1. Identify the columns in the names table that have null values

Answer: We can use the IS NULL function with each column name in the names table. The query would look like – SELECT \* from names WHERE column\_name IS NULL

When we try and execute this query with the column names from the names table, we can see which column has null values. The columns ‘height’, ‘date\_of\_birth’ and ‘known\_for\_movies’ from the names table have null values.

1. Determine the top three directors in the top three genres with movies having an average rating > 8.

Answer: The query is- SELECT g.genre,n.name as director\_name,r.avg\_rating

from movies m

JOIN ratings r

ON m.id = r.movie\_id

JOIN genre g

ON g.movie\_id = m.id

JOIN director\_mapping dm

ON dm.movie\_id = m.id

JOIN names n

ON dm.name\_id = n.id

WHERE r.avg\_rating > 8

ORDER BY g.genre DESC, r.avg\_rating DESC

LIMIT 3

1. Find the top two actors whose movies have a median rating >= 8.

Answer :

The query is- Select n.name as actor\_name, ROUND(AVG(r.median\_rating),2) as avg\_median\_rating from names n

JOIN role\_mapping rm

ON n.id = rm.name\_id

JOIN ratings r

ON rm.movie\_id = r.movie\_id

GROUP BY 1

HAVING AVG(r.median\_rating) >=8

ORDER BY 2 DESC

LIMIT 2

1. Identify the top three production houses based on the number of votes received by their movies.

Answer- The query is -SELECT m.production\_company as prod\_company, SUM(r.total\_votes) as total\_votes from movies m

JOIN ratings r

ON m.id =r.movie\_id

GROUP BY 1

ORDER BY 2 DESC

LIMIT 3

1. Rank actors based on their average ratings in Indian movies released in India.

Answer- The query is - Select n.name as actor\_name, ROUND(AVG(r.avg\_rating),2) as avg\_rating, DENSE\_RANK() over (ORDER BY AVG(r.avg\_rating) DESC) as Actor\_Rank from names n

JOIN role\_mapping rm

ON n.id = rm.name\_id

JOIN ratings r

ON rm.movie\_id = r.movie\_id

JOIN movies m

ON m.id= r.movie\_id

WHERE m.country = 'India' AND rm.category = 'actor'

GROUP BY 1

ORDER BY 2 DESC

1. Identify the top five actresses in Hindi movies released in India based on their average ratings.

Answer- The query is - SELECT actress\_name, avg\_rating from (Select n.name as actress\_name, ROUND(AVG(r.avg\_rating),2) as avg\_rating, DENSE\_RANK() over (ORDER BY AVG(r.avg\_rating) DESC) as Actor\_Rank from names n

JOIN role\_mapping rm

ON n.id = rm.name\_id

JOIN ratings r

ON rm.movie\_id = r.movie\_id

JOIN movies m

ON m.id= r.movie\_id

WHERE m.country = 'India' AND rm.category = 'actress' AND m.languages LIKE '%Hindi%'

GROUP BY 1

ORDER BY 2 DESC) as Temp

WHERE Actor\_Rank <= 5

Segment 6: Broader Understanding of Data

1. Classify thriller movies based on average ratings into different categories

Answer - The query is - SELECT m.title, r.avg\_rating, CASE WHEN r.avg\_rating > 8 THEN "EXCELLENT"

WHEN r.avg\_rating BETWEEN 6 AND 8 THEN "GOOD" ELSE "AVERAGE AND BELOW" END as Classification

from genre g

JOIN movies m

ON g.movie\_id = m.id

JOIN ratings r

ON r.movie\_id = m.id

WHERE g.genre = "Thriller"

1. analyse the genre-wise running total and moving average of the average movie duration.

Answer - The query is - SELECT g.genre, m.duration, sum(m.duration) over (partition by g.genre order by m.year) as running\_total,

avg(m.duration) over (partition by g.genre order by m.year rows between unbounded preceding and current row) as moving\_average

from genre g

JOIN

movies m

On g.movie\_id = m.id

1. Identify the five highest-grossing movies of each year that belong to the top three genres.

Answer - The query is - WITH top\_three\_genres AS (

SELECT genre, COUNT(\*) AS movie\_count

FROM genre

GROUP BY genre

ORDER BY movie\_count DESC

LIMIT 3

)

highest\_grossing\_movies AS (

SELECT year, m.title, m.worlwide\_gross\_income, g.genre,

ROW\_NUMBER() OVER (PARTITION BY m.year, g.genre ORDER BY m.worlwide\_gross\_income DESC) AS `rank`

FROM movies m

INNER JOIN genre g ON m.id = g.movie\_id

INNER JOIN top\_three\_genres t ON g.genre = t.genre

)

SELECT year, genre, title, worlwide\_gross\_income

FROM highest\_grossing\_movies

WHERE `rank` <= 5

ORDER BY year, genre, `rank`;

1. Determine the top two production houses that have produced the highest number of hits among multilingual movies.

Answer - The query is - select production\_company, count(\*) as movie\_count

from movies m

JOIN ratings r

ON

m.id = r.movie\_id

WHERE r.avg\_rating > 8 and m.production\_company IS NOT NULL

Group by 1

ORDER BY movie\_count DESC

LIMIT 2

offset 1

1. Identify the top three actresses based on the number of Super Hit movies (average rating > 8) in the drama genre.

Answer - The query is - select n.name, count(m.id) as movie\_count

from movies m

JOIN ratings r

ON

m.id = r.movie\_id

JOIN genre g

ON

m.id = g.movie\_id

JOIN role\_mapping rm

ON

m.id = rm.movie\_id

JOIN names n

ON

rm.name\_id = n.id

WHERE rm.category = "Actress" AND r.avg\_rating > 8 AND g.genre = "Drama"

GROUP BY 1

ORDER BY 2 DESC

LIMIT 3

1. Retrieve details for the top nine directors based on the number of movies, including average inter-movie duration, ratings, and more.

Answer - The query is - select dm.name\_id as director\_id, count(m.id) as movie\_count, ROUND(AVG(r.avg\_rating),1) as avg\_rating, SUM(r.total\_votes) as total\_votes

from movies m

JOIN ratings r

ON

m.id = r.movie\_id

JOIN genre g

ON

m.id = g.movie\_id

JOIN director\_mapping dm

ON

m.id = dm.movie\_id

JOIN names n

ON

dm.name\_id = n.id

GROUP BY 1

ORDER BY 2 DESC, 3 DESC, 4 DESC

LIMIT 9

Segment 7: Recommendations

* Based on the analysis, provide recommendations for the types of content Bolly movies should focus on producing.

Answer - Based on the analysis, I recommend focusing on partnerships with Marvel for high-volume production opportunities, leveraging the dominance of the USA as a key production hub, and prioritizing English and multilingual projects to align with the prevalent language trends in the industry.