-- segment 1: --Database - Tables, Columns, Relationships

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

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

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

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

ANSWER 1) There are six table in databases i.e movies, genre, director\_mapping, role\_mapping, names,ratings.

In a relational database, data is organized into tables, which consist of rows and columns.

The relationships between these tables are established through keys and constraints.

### Connections Between Tables

* **movies** is the central table.
* **genre**, **director\_mapping**, **role\_mapping**, and **ratings** tables all reference movies table via movie\_id.
* **director\_mapping** and **role\_mapping** tables also reference names table via name\_id.

These relationships create a relational database structure where the movies table is connected to other tables through foreign keys.

movies <-- genre

<-- ratings

<-- director\_mapping --> names

<-- role\_mapping --> names

Common Types of Relationships

1-One-to-One: A single row in one table is linked to a single row in another table.

2.One-to-Many: A single row in one table is linked to multiple rows in another table.

3.Many-to-Many: Multiple rows in one table are linked to multiple rows in another table through a junction table.

QUESTION 2)-- Find the total number of rows in each table of schema

ANSWER 2)

select COUNT(\*) as genre\_count from genre;

select COUNT(\*) as name\_count from names;

select COUNT(\*) as project\_count from project;

select COUNT(\*) as rating\_count from rating;

select COUNT(\*) as role\_mapping\_count from role\_mapping;

select COUNT(\*) as movie\_count from movie;

QUESTION 3)-- identify which column have null value in movies table

ANSWER 3)

SELECT

SUM(CASE WHEN year IS NULL OR TRIM(year) = '' THEN 1 ELSE 0 END) AS year\_null,

SUM(CASE WHEN title IS NULL OR TRIM(title) = '' THEN 1 ELSE 0 END) AS title\_null,

SUM(CASE WHEN country IS NULL OR TRIM(country) = '' THEN 1 ELSE 0 END) AS country\_null,

SUM(CASE WHEN languages IS NULL OR TRIM(languages) = '' THEN 1 ELSE 0 END) AS language\_null,

SUM(CASE WHEN production\_company IS NULL OR TRIM(production\_company) = '' THEN 1 ELSE 0 END) AS production\_null,

SUM(CASE WHEN worlwide\_gross\_income IS NULL OR TRIM(worlwide\_gross\_income) = '' THEN 1 ELSE 0 END) AS worlwide\_gross\_income\_null,

SUM(CASE WHEN date\_published IS NULL OR TRIM(date\_published) = '' THEN 1 ELSE 0 END) AS date\_published\_null

FROM

movies;

segment 2: --MOVIE RELEASE

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

ANSWER 1)

select year,substr(date\_published,4,2) as month\_count,count(title) as number\_of\_movie\_released

from movies

group by year ,substr(date\_published,4,2)

order by year ,month\_count ;

or

SELECT

SUBSTR(date\_published, 7, 4) AS year,

SUBSTR(date\_published, 4, 2) AS month,

COUNT(title) AS number\_of\_movies\_released

FROM

movies

GROUP BY

SUBSTR(date\_published, 7, 4),

SUBSTR(date\_published, 4, 2)

ORDER BY

year,

month;

QUESTION 2) --calculate the number of movies produced in the USA OR INDIA in the year 2019

ANSWER 2)

select count(\*) as movie\_produced\_in\_2019

from movies

where country in ('USA','India')and year in(2019);

--or

select count(id) as movie\_produced\_in\_2019

from movies

where (country like'%USA%' or country like '%India%') and year =2019;

segment 3 :-- Production statistics and genre analysis

--Retrieve the unique list of genre present in the database

--identify genre with highest number of movie produced overall

--Determine the count of movie that belong to only one genre

--calculate the average duration of the movie in each genre

--Find the rank of the thriller genre among all genres in terms of number of movies produced

question 1 - --Retrieve the unique list of genre present in the dataset.

ANSWER 1)

select Distinct genre from genre;

--OR

select genre from genre

group by genre;

question 2) identify the genre with highest number of movie produced overall

ANSWER 2)

select genre,count(id) as highest\_number\_of\_movie

from genre g

join movies m

on m.id=g.movie\_id

group by genre

order by count(id) desc

limit 1;

question 3) --Determine the count of movie that belong to only one genre

select count(movie\_id) as movies\_with\_one\_genre from

(

select movie\_id,count(genre) AS genre\_count

from genre

group by movie\_id

) A

WHERE A.genre\_count =1;

ASSIGNMENT-- --Determine the movie that belong to only one genre

SELECT

m.title,count(genre) as genre

FROM

movies m

JOIN

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

GROUP BY

m.id, m.title

HAVING

COUNT(g.genre) = 1;

Or

select movie\_id,count(genre)

from genre

group by movie\_id

having count(genre)=1;

question 4) --calculate the average duration of the movie in each genre

select genre,avg(duration)as avg\_duration

from movies m

join genre g

on m.id=g.movie\_id

group by genre

order by avg\_duration desc;

question 5) --Find the rank of the thriller genre among all genres in terms of number of movies produced

SELECT

genre,

RANK() OVER (ORDER BY movie\_count DESC) AS genre\_rank

FROM (

SELECT

g.genre,

COUNT(m.id) AS movie\_count

FROM

movies m

JOIN

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

GROUP BY

g.genre

) AS genre\_counts

WHERE

genre = 'Thriller';

segment 4: --Rating analysis and crew members

-- Retrieve the minimum and maximum value in each column of the rating table(except movie\_id)

--Identify the top 10 movies based on the average rating

-- Summarise the Rating table based on movie count by median rating.

--identify the production house that has produced the most number of hit movies (average rating>0),

-- Determin the number of movies released in each genre during March 2017 in the USA with more than 1000 votes

-- Retrieve movie of each genre starting with the word 'The' and having and average rating >8

question 1)-- Retrieve the minimum and maximum value in each column of the rating table(except movie\_id)

answer 1)

SELECT

MAX(avg\_rating) AS maximum\_average\_rating,

MIN(avg\_rating) AS minimum\_average\_rating,

MAX(total\_votes) AS maximum\_total\_votes,

MIN(total\_votes) AS minimum\_total\_votes,

MAX(median\_rating) AS maximum\_median\_rating,

MIN(median\_rating) AS minimum\_median\_rating

FROM ratings;

question 2) --Identify the top 10 movies based on the average rating

answer 2)

select sum(avg\_rating),title

from ratings r

join movies m

on r.movie\_id=m.id

group by title

order by sum(avg\_rating) desc

limit 10;

or

WITH top\_movies AS (

SELECT

avg\_rating,

m.title,

RANK() OVER (ORDER BY avg\_rating DESC) AS movie\_rank

FROM

movies m

LEFT JOIN

ratings r ON r.movie\_id = m.id

)

SELECT

\*

FROM

top\_movies

WHERE

movie\_rank <= 10;

QUESTION 3) -- Summarise the Rating table based on movie count by median rating.

ANSWER 3)

SELECT

median\_rating,

COUNT(\*) AS movie\_count

FROM

ratings

GROUP BY

median\_rating

ORDER BY

median\_rating;

QUESTION 4) --identify the production house that has produced the most number of hit movies (average rating>8),

ANSWER--4)

SELECT production\_company,count(id)as movie\_count ,avg\_rating

from movies m

left join ratings r

on m.id=r.movie\_id

where production\_company is not null and avg\_rating > 8

group by production\_company ,avg\_rating

order by count(id) desc,avg\_rating desc ;

or

SELECT

production\_company,

COUNT(m.id) AS movie\_count,

AVG(r.avg\_rating) AS average\_rating

FROM

movies m

LEFT JOIN

ratings r ON m.id = r.movie\_id

WHERE

production\_company IS NOT NULL

AND r.avg\_rating > 8

GROUP BY

production\_company

ORDER BY

movie\_count DESC;

QUESTION 5) --Determine the number of movies released in each genre during March 2017 in the USA with more than 1000 votes

ANSWER 5--)

SELECT g.genre,count(id)

from movies m

left join genre g

on m.id=g.movie\_id

left join ratings r

on m.id=r.movie\_id

where year='2017' and lower(country) like '%usa%' AND total\_votes>1000

group by genre;

QUESTION 6) -- Retrieve movie of each genre starting with the word 'The' and having and average rating >8

answer 6)

SELECT

g.genre,

m.title,

r.avg\_rating

FROM

movies 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;

­

SEGMENT 5: -- CREW ANALYSIS

--

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

--2) Determin the top 3 directors in the top three genre with movies having an average rating> 8

--3) Find the top two actors whose movies have a median rating>=8

--4) Identify the top 3 production houses based on the number of votes received by their movies

--5) Rank actors based on their average ratings in Indian movies released in India

--6) Identify the Top five actress IN Hindi movies released in India based on their average rating

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

select

sum(case when id='' then 1 else 0 end) as Null\_for\_id,

sum(case when name='' then 1 else 0 end) as Null\_for\_Name,

sum(case when date\_of\_birth='' then 1 else 0 end) as Null\_for\_DOB,

sum(case when known\_for\_movies='' then 1 else 0 end) as known\_for\_movies,

sum(case when height='' then 1 else 0 end) as Null\_for\_height

from names

QUESTION 2)--Determine the top 3 directors in the top three genre with movies having an average rating> 8

answer 2)

SELECT g.genre ,n.name,d.name\_id,count(\*)as movie\_count

from director\_mapping d

LEFT JOIN names n ON d.name\_id = n.id

LEFT JOIN movies m ON d.movie\_id = m.id

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

LEFT JOIN ratings r ON m.id = r.movie\_id

where avg\_rating > 8

group by g.genre,n.name,d.name\_id

order by count(\*) desc

limit 3;

QUESTION 3) Find the top two actors whose movies have a median rating>=8

answer 3)

SELECT ro.name\_id,COUNT(r.movie\_id) AS movie\_count

FROM role\_mapping ro

LEFT JOIN ratings r

ON ro.movie\_id=r.movie\_id

WHERE ro.category = 'actor' AND r.median\_rating > 8

GROUP BY ro.name\_id ,r.median\_rating

order by r.median\_rating desc

limit 2;

or

select r.name\_id,count(m.id) as movies ,s.median\_rating

from movies m

left join role\_mapping r

on m.id=r.movie\_id

left join ratings s

on m.id=s.movie\_id

where r.category ='actor' and s.median\_rating>8

group by r.name\_id,s.median\_rating

order by s.median\_rating desc

limit 2;

QUESTION 4) --Identify the top 3 production houses based on the number of votes received by their movies

answer 4)

SELECT DISTINCT PRODUCTION\_COMPANY as Production\_House ,SUM(TOTAL\_VOTES)AS VOTES

FROM Movies M

LEFT JOIN RATINGS R

ON R.MOVIE\_ID=M.ID

GROUP BY PRODUCTION\_COMPANY

ORDER BY VOTES DESC

LIMIT 3;

QUESTION 5)-- Rank actors based on their average rating in Indian movies released in India

ANSWER 5)

­

WITH ActorAvgRating AS (

SELECT

DISTINCT R.NAME\_ID,

AVG(S.AVG\_RATING) AS AVERAGE,

RANK() OVER (ORDER BY AVG(S.AVG\_RATING) DESC) AS RATING\_RANK

FROM

movies M

LEFT JOIN

ROLE\_MAPPING R ON M.ID = R.MOVIE\_ID

LEFT JOIN

RATINGS S ON M.ID = S.MOVIE\_ID

WHERE

M.COUNTRY = 'INDIA' AND R.CATEGORY = 'actor'

GROUP BY

R.NAME\_ID

)

SELECT

NAME\_ID,

AVERAGE,

RATING\_RANK

FROM

ActorAvgRating;

­

QUESTION 6)

ANSWER 6)--Identify the Top five actress IN Hindi movies released in India based on their average rating

SELECT DISTINCT R.NAME\_ID, AVG(S.AVG\_RATING) AS AVERAGE

FROM movies M

LEFT JOIN ROLE\_MAPPING R ON M.ID = R.MOVIE\_ID

LEFT JOIN RATINGS S ON M.ID = S.MOVIE\_ID

WHERE M.COUNTRY = 'INDIA' AND R.CATEGORY='actress' AND M.LANGUAGES='Hindi'

GROUP BY R.NAME\_ID

ORDER BY AVERAGE DESC

limit 5;

or

WITH ActressAvgRating AS (

SELECT

DISTINCT R.NAME\_ID,

AVG(S.AVG\_RATING) AS AVERAGE,

RANK() OVER (ORDER BY AVG(S.AVG\_RATING) DESC) AS RATING\_RANK

FROM

movies M

LEFT JOIN

ROLE\_MAPPING R ON M.ID = R.MOVIE\_ID

LEFT JOIN

RATINGS S ON M.ID = S.MOVIE\_ID

WHERE

M.COUNTRY = 'INDIA'

AND R.CATEGORY = 'actress'

AND M.LANGUAGES = 'Hindi'

GROUP BY

R.NAME\_ID

)

SELECT

NAME\_ID,

AVERAGE

FROM

ActressAvgRating

WHERE

RATING\_RANK <= 5;

Segment 6: --Broader Understanding of Data

-- classify thriller movies based on average ratings into different catagories.

-- Analysis the genre-wise running total and moving average of the average movie duration

-- identify the five highest grossing movies of each year that belong to top three GENRE

-- Determin the Top two Production house that have produced highest number of hit among multilingual movies

-- Identify the top three actoress based on the number of Super hit movies (average rating > 8) in the drama genre

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

QUESTION 1) -- classify thriller movies based on average rating into different catagories.

ANSWER 1)

Here is the classification of thriller movies based on their average ratings:

* Excellent (avg\_rating ≥ 8.0)
* Good (7.0 ≤ avg\_rating < 8.0)
* Average (6.0 ≤ avg\_rating < 7.0)
* Below Average (avg\_rating < 6.0)

SELECT m.id,r.AVG\_RATING ,

CASE

WHEN r.avg\_rating >= 8.0 THEN 'Hit'

WHEN r.avg\_rating >= 6.0 THEN 'Average'

ELSE 'Flop' end as Movie\_catagory

FROM movies M

LEFT JOIN GENRE G ON M.ID=G.MOVIE\_ID

LEFT JOIN RATINGS R ON M.ID=R.MOVIE\_ID

WHERE G.GENRE='Thriller' ;

QUESTION 2) -- Analysis the genre-wise running total and moving average of the average movie duration

ANSWER 2)

The movies sheet contains information about individual movies, including their duration. The genre sheet maps movies to their genres. To analyze the genre-wise running total and moving average of the average movie duration, we need to:

1. Merge the movies and genre dataframes on the movie IDs.
2. Calculate the average duration for each genre.
3. Compute the running total and moving average for each genre's average movie duration.

Here is the analysis of the genre-wise running total and moving average of the average movie duration:

| Genre | Average Duration | Running Total | Moving Average |
| --- | --- | --- | --- |
| Action | 112.88 | 112.88 | 112.88 |
| Adventure | 101.87 | 214.75 | 107.38 |
| Comedy | 102.62 | 317.38 | 105.79 |
| Crime | 107.05 | 424.43 | 103.85 |
| Drama | 106.77 | 531.20 | 105.48 |
| Family | 100.97 | 632.17 | 104.93 |
| Fantasy | 105.14 | 737.31 | 104.29 |
| Horror | 92.72 | 830.03 | 99.61 |
| Mystery | 101.80 | 931.83 | 99.89 |
| Others | 100.16 | 1031.99 | 98.23 |
| Romance | 109.53 | 1141.53 | 103.83 |
| Sci-Fi | 97.94 | 1239.47 | 102.55 |
| Thriller | 101.58 | 1341.05 | 103.02 |

* Running Total: It represents the cumulative sum of the average movie durations up to the current genre.
* Moving Average: It is calculated using a window size of 3, giving the average of the current genre and the two preceding ones.

SELECT

m.id,

g.genre,

m.duration,

SUM(m.duration) OVER (PARTITION BY g.genre ORDER BY m.id) AS duration\_sum,

AVG(m.duration) OVER (PARTITION BY g.genre ORDER BY m.id) AS moving\_average

FROM

movies m

LEFT JOIN

genre g ON m.id = g.movie\_id;

QUESTION 3) --identify the five highest grossing movies of each year that belong to top three GENRE

ANSWER3)

WITH RankedMovies AS (

SELECT

m.id,

m.title,

m.year,

g.genre,

m.worlwide\_gross\_income,

RANK() OVER (PARTITION BY m.year, g.genre ORDER BY m.worlwide\_gross\_income DESC) AS ranking

FROM

movies m

LEFT JOIN

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

)

SELECT

id,

title,

year,

genre,

worlwide\_gross\_income

FROM

RankedMovies

WHERE

ranking <= 5;

QUESTION 4) --Determine the Top two Production house that have produced highest number of hits among multilingual movies.

ANSWER 4)

SELECT production\_company,languages, count(id) as Total\_movies

FROM movies

WHERE languages like '%,%'

GROUP BY production\_company,languages

order by total\_movies desc

limit 2 ;

or

WITH HitMovies AS (

SELECT

m.production\_company,

COUNT(\*) AS hit\_count

FROM

movies m

INNER JOIN

ratings r ON m.id = r.movie\_id

WHERE

m.languages <> 'English'

AND r.avg\_rating >= 8.0

GROUP BY

m.production\_company

)

SELECT

production\_company,

hit\_count

FROM

HitMovies

ORDER BY

hit\_count DESC

LIMIT 2;

QUESTION 5)--Identify the top three actoress based on the number of Super hit movies (average rating>8) IN THE DRAMA GENRE

ANSWER 5)

WITH SuperHitActresses AS (

SELECT

n.id AS actress\_id,

n.name AS actress\_name,

COUNT(\*) AS super\_hit\_count

FROM

names n

INNER JOIN

role\_mapping rm ON n.id = rm.name\_id

INNER JOIN

movies m ON rm.movie\_id = m.id

INNER JOIN

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

INNER JOIN

ratings r ON m.id = r.movie\_id

WHERE

g.genre = 'Drama'

AND r.avg\_rating > 8.0

AND rm.category = 'actress'

GROUP BY

n.id, n.name

)

SELECT

actress\_id,

actress\_name,

super\_hit\_count

FROM

SuperHitActresses

ORDER BY

super\_hit\_count DESC

LIMIT 3;

or

SELECT id,g.genre,count( m.id) as movie\_produced

FROM movies m

left join ratings r on m.id=r.movie\_id

left join role\_mapping ro on m.id=ro.movie\_id

left join genre g on m.id=g.movie\_id

where ro.category='actress' and r.avg\_rating>8 and g.genre='Drama'

group by id,g.genre order by movie\_produced desc

limit 3;

QUESTION 6) -- Retrieve the details of top nine directors bases on the number of movies,including average inter movie duration,rating,more

ANSWER 6)

select d.name\_id as director\_id,n.name as director\_name,count(m.id)as num\_Movies\_produced,

avg(m.duration)as average\_duration,avg(r.avg\_rating) from movies m

left join genre g on m.id=g.movie\_id

left join director\_mapping d on m.id=d.movie\_id

left join ratings r on m.id=r.movie\_id

LEFT join names n on d.name\_id=n.id

where d.name\_id is not null

group by d.name\_id,n.name

order by num\_Movies\_produced desc;

or

WITH DirectorStats AS (

SELECT

d.name\_id AS director\_id,

COUNT(\*) AS movie\_count,

AVG(m.duration) AS avg\_inter\_movie\_duration,

AVG(r.avg\_rating) AS avg\_rating,

ROW\_NUMBER() OVER (ORDER BY COUNT(\*) DESC) AS director\_rank

FROM

director\_mapping d

JOIN

movies m ON d.movie\_id = m.id

JOIN

ratings r ON m.id = r.movie\_id

GROUP BY

d.name\_id

)

SELECT

ds.director\_id,

n.name AS director\_name,

ds.movie\_count,

ds.avg\_inter\_movie\_duration,

ds.avg\_rating

FROM

DirectorStats ds

JOIN

names n ON ds.director\_id = n.id

WHERE

ds.director\_rank <= 9;

Segment 7: Recommendations

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

### Answer ) – Recommendations

Based on the analysis, here are the recommendations for the types of content Bolly movies should focus on producing:

1. Increase Production in Highly Rated Genres:
   1. Drama: Although it's already the most produced genre, dramas have a relatively high average rating, suggesting that they are well-received by audiences. Continued production in this genre is advisable.
   2. Romance and Crime: These genres have high average ratings (5.99 and 5.93, respectively) and could benefit from increased production.
2. Explore Hybrid Genres:
   1. Romantic Dramas: Combining the high production volume and success of dramas with the high average ratings of romance could yield popular films.
   2. Crime Thrillers: Merging elements of crime and thriller genres can attract audiences who enjoy suspenseful and engaging stories.
3. Leverage Popular Genres with Room for Improvement:
   1. Comedy and Action: These genres are popular but have average ratings below 6. Improving the quality and storytelling within these genres could enhance their overall reception.
4. Expand into Underserved High-Potential Genres:
   1. Adventure and Mystery: These genres have moderate production volumes and decent average ratings. Increasing production with a focus on quality can attract a wider audience.
5. Target Niche Markets:
   1. Sci-Fi and Horror: Despite lower average ratings, these genres have dedicated fan bases. Investing in high-quality special effects and compelling storylines can elevate these genres' appeal.
6. Quality Over Quantity:
   1. Focus on improving the script, direction, and production quality of movies in the most produced genres (Drama, Comedy, Thriller) to enhance audience satisfaction and ratings.

By aligning content production with these insights, Bolly movies can maximize audience engagement and satisfaction, leading to greater success in the film industry.