Movies Data Analysis using SQL

By
Ng Hoi Yee

Entity Relationship Diagram - ERD

- helps to visualize how data is connected to depicts relationships

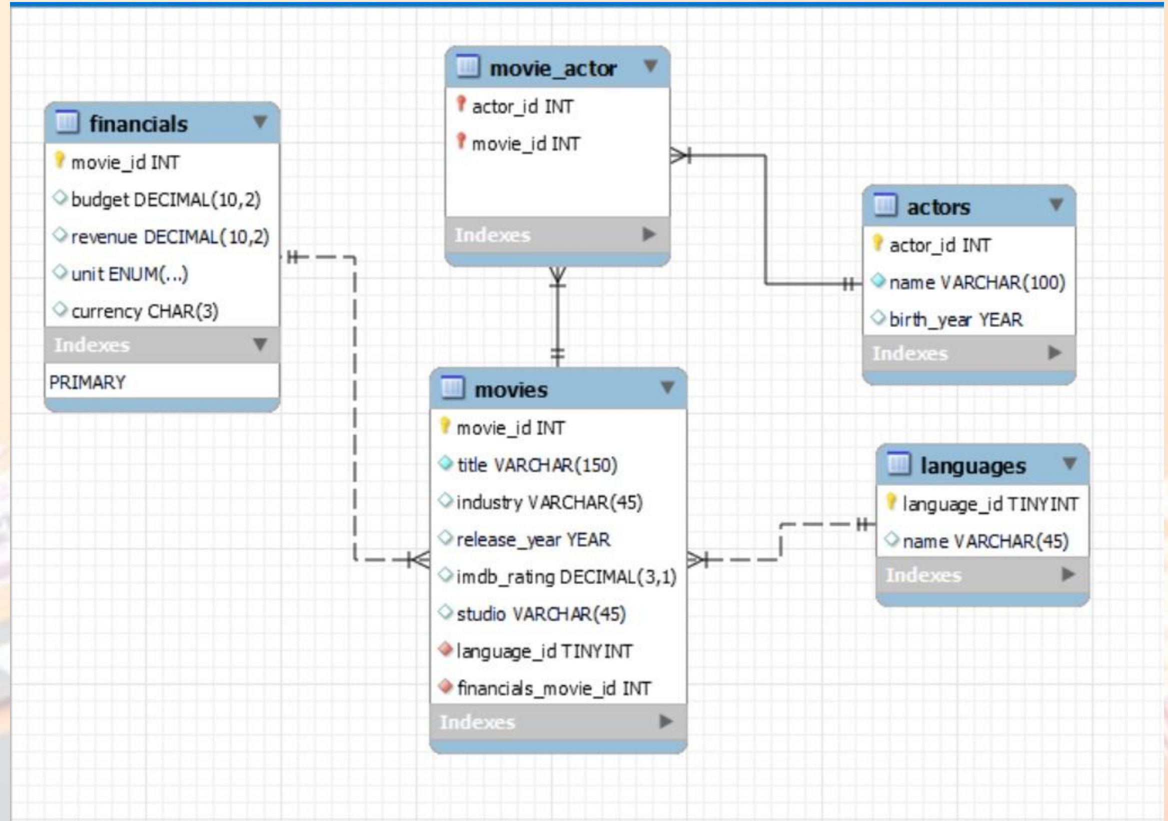
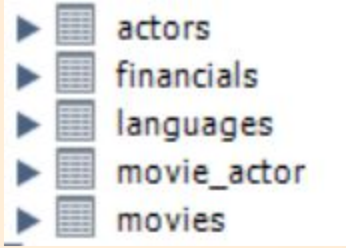
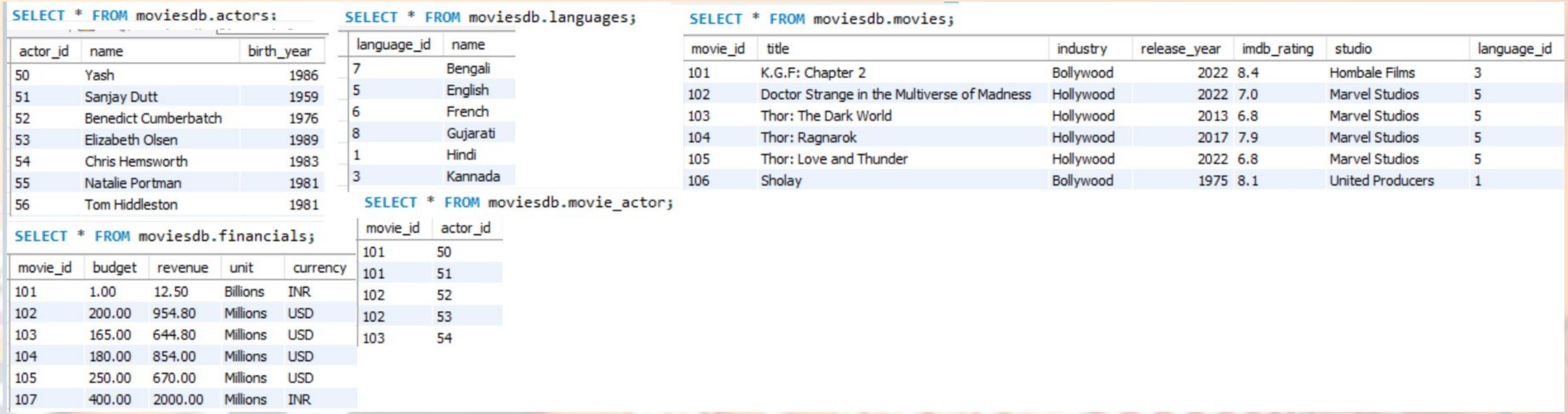


Table overview



- actors
- financials
- languages
- movie_actor
- movies

```
SELECT * FROM moviesdb.actors;
SELECT * FROM moviesdb.financials;
SELECT * FROM moviesdb.languages;
SELECT * FROM moviesdb.movie_actor;
SELECT * FROM moviesdb.movies;
```



```
SELECT * FROM moviesdb.actors;
```

actor_id	name	birth_year
50	Yash	1986
51	Sanjay Dutt	1959
52	Benedict Cumberbatch	1976
53	Elizabeth Olsen	1989
54	Chris Hemsworth	1983
55	Natalie Portman	1981
56	Tom Hiddleston	1981

```
SELECT * FROM moviesdb.financials;
```

movie_id	budget	revenue	unit	currency
101	1.00	12.50	Billions	INR
102	200.00	954.80	Millions	USD
103	165.00	644.80	Millions	USD
104	180.00	854.00	Millions	USD
105	250.00	670.00	Millions	USD
107	400.00	2000.00	Millions	INR

```
SELECT * FROM moviesdb.languages;
```

language_id	name
7	Bengali
5	English
6	French
8	Gujarati
1	Hindi
3	Kannada

```
SELECT * FROM moviesdb.movie_actor;
```

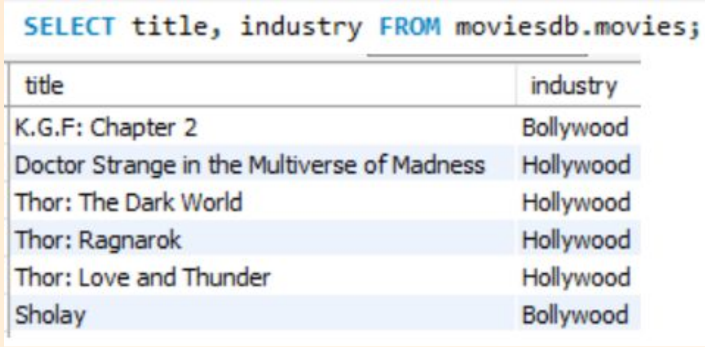
movie_id	actor_id
101	50
101	51
102	52
102	53
103	54

```
SELECT * FROM moviesdb.movies;
```

movie_id	title	industry	release_year	imdb_rating	studio	language_id
101	K.G.F: Chapter 2	Bollywood	2022	8.4	Hombale Films	3
102	Doctor Strange in the Multiverse of Madness	Hollywood	2022	7.0	Marvel Studios	5
103	Thor: The Dark World	Hollywood	2013	6.8	Marvel Studios	5
104	Thor: Ragnarok	Hollywood	2017	7.9	Marvel Studios	5
105	Thor: Love and Thunder	Hollywood	2022	6.8	Marvel Studios	5
106	Sholay	Bollywood	1975	8.1	United Producers	1

By Ng Hoi Yee

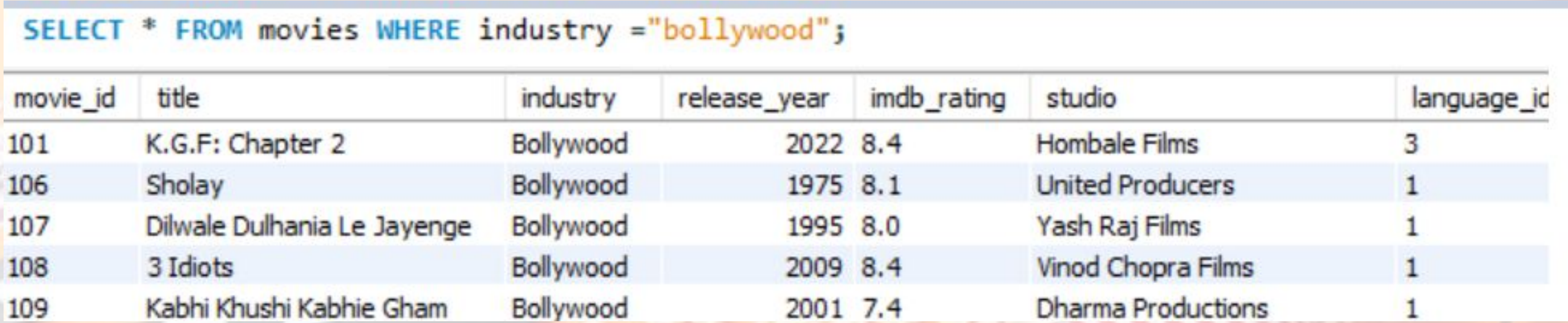
- get title and industry



```sql
SELECT title, industry FROM moviesdb.movies;
```

title	industry
K.G.F: Chapter 2	Bollywood
Doctor Strange in the Multiverse of Madness	Hollywood
Thor: The Dark World	Hollywood
Thor: Ragnarok	Hollywood
Thor: Love and Thunder	Hollywood
Sholay	Bollywood

- get only bollywood titles

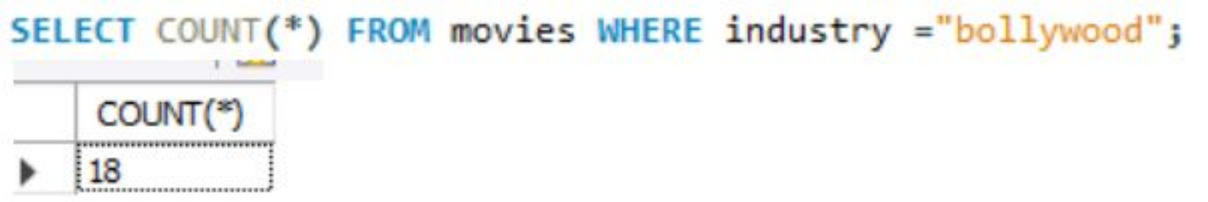


```sql
SELECT * FROM movies WHERE industry ="bollywood";
```

movie_id	title	industry	release_year	imdb_rating	studio	language_id
101	K.G.F: Chapter 2	Bollywood	2022	8.4	Hombale Films	3
106	Sholay	Bollywood	1975	8.1	United Producers	1
107	Dilwale Dulhania Le Jayenge	Bollywood	1995	8.0	Yash Raj Films	1
108	3 Idiots	Bollywood	2009	8.4	Vinod Chopra Films	1
109	Kabhi Khushi Kabhie Gham	Bollywood	2001	7.4	Dharma Productions	1

Movies Data Analysis using SQL

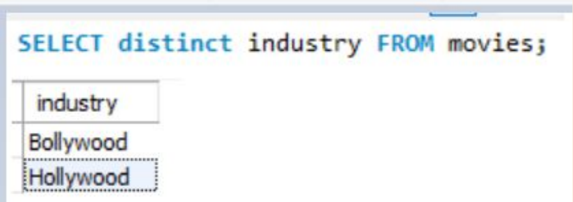
- get number of Bollywood titles



```sql
SELECT COUNT(*) FROM movies WHERE industry ="bollywood";
```

COUNT(*)
18

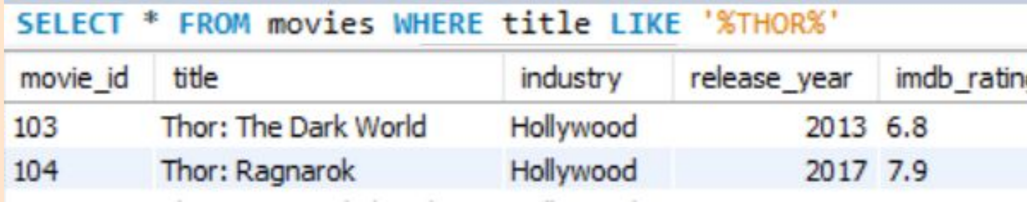
- what are the industries?



```sql
SELECT distinct industry FROM movies;
```

industry
Bollywood
Hollywood

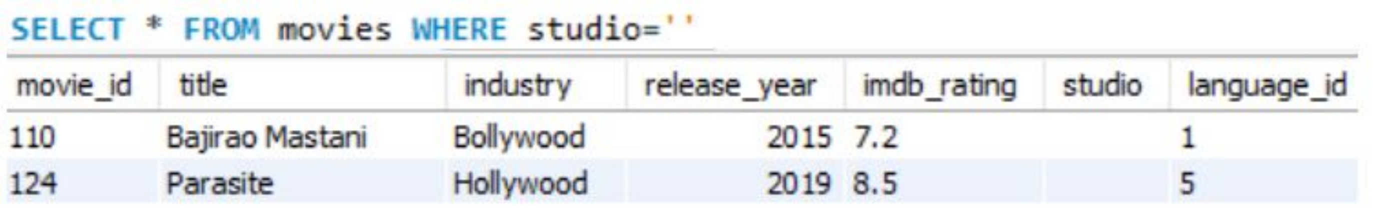
- get the titles with ‘THOR’ in them



```sql
SELECT * FROM movies WHERE title LIKE '%THOR%'
```

movie_id	title	industry	release_year	imdb_ratin
103	Thor: The Dark World	Hollywood	2013	6.8
104	Thor: Ragnarok	Hollywood	2017	7.9

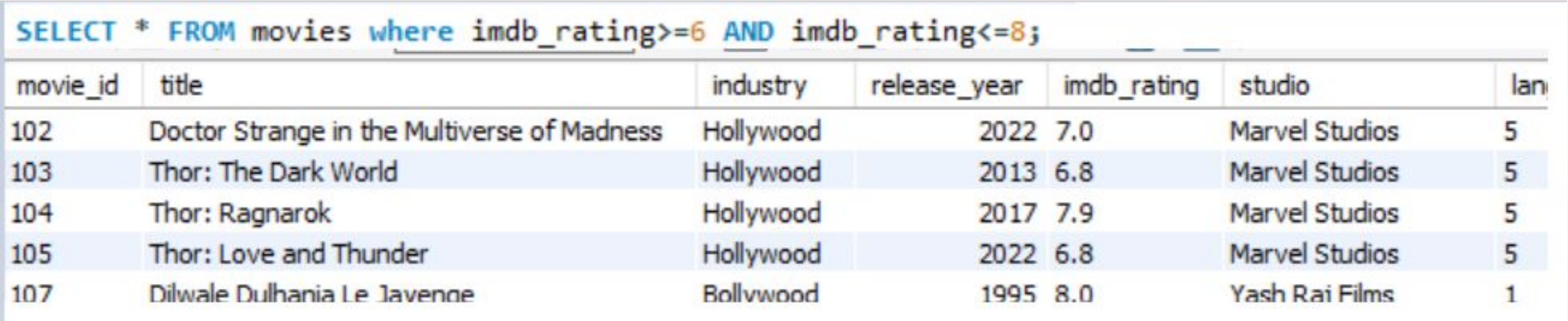
- get rows where studio is blank



```sql
SELECT * FROM movies WHERE studio=''
```

movie_id	title	industry	release_year	imdb_rating	studio	language_id
110	Bajirao Mastani	Bollywood	2015	7.2		1
124	Parasite	Hollywood	2019	8.5		5

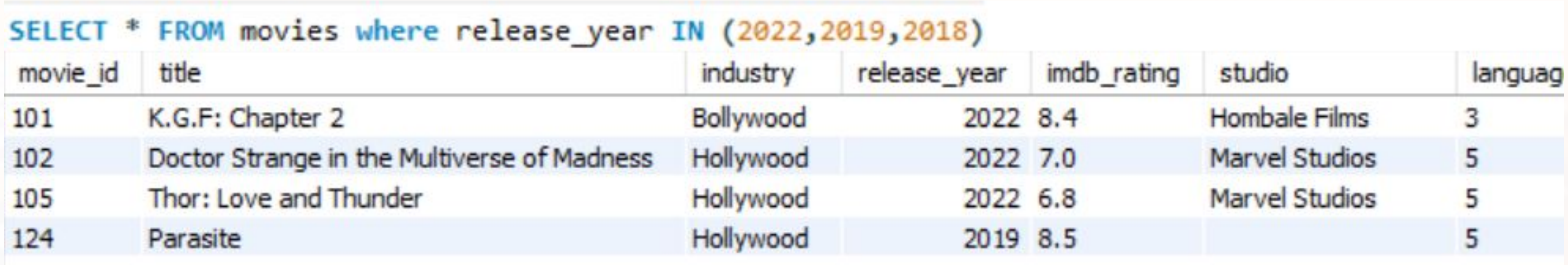
- get titles where imdb_rating is between 6 and 8 including 6 and 8



```
SELECT * FROM movies where imdb_rating>=6 AND imdb_rating<=8;
```

movie_id	title	industry	release_year	imdb_rating	studio	lan
102	Doctor Strange in the Multiverse of Madness	Hollywood	2022	7.0	Marvel Studios	5
103	Thor: The Dark World	Hollywood	2013	6.8	Marvel Studios	5
104	Thor: Ragnarok	Hollywood	2017	7.9	Marvel Studios	5
105	Thor: Love and Thunder	Hollywood	2022	6.8	Marvel Studios	5
107	Dilwale Dulhania Le Jayenge	Bollywood	1995	8.0	Yash Raj Films	1

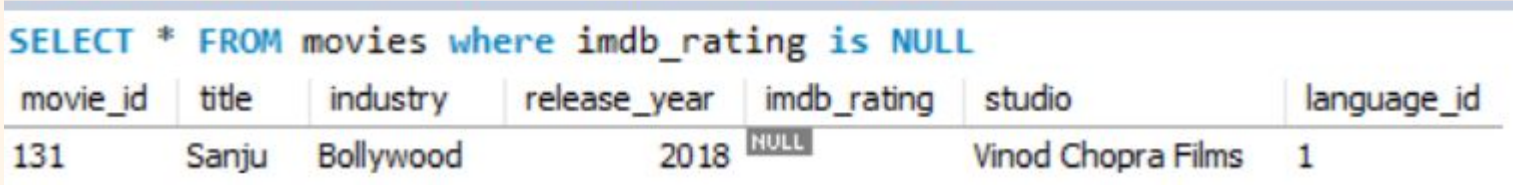
- get titles where release year is 2018 or 2019 or 2022



```
SELECT * FROM movies where release_year IN (2022,2019,2018)
```

movie_id	title	industry	release_year	imdb_rating	studio	languag
101	K.G.F: Chapter 2	Bollywood	2022	8.4	Hombale Films	3
102	Doctor Strange in the Multiverse of Madness	Hollywood	2022	7.0	Marvel Studios	5
105	Thor: Love and Thunder	Hollywood	2022	6.8	Marvel Studios	5
124	Parasite	Hollywood	2019	8.5		5

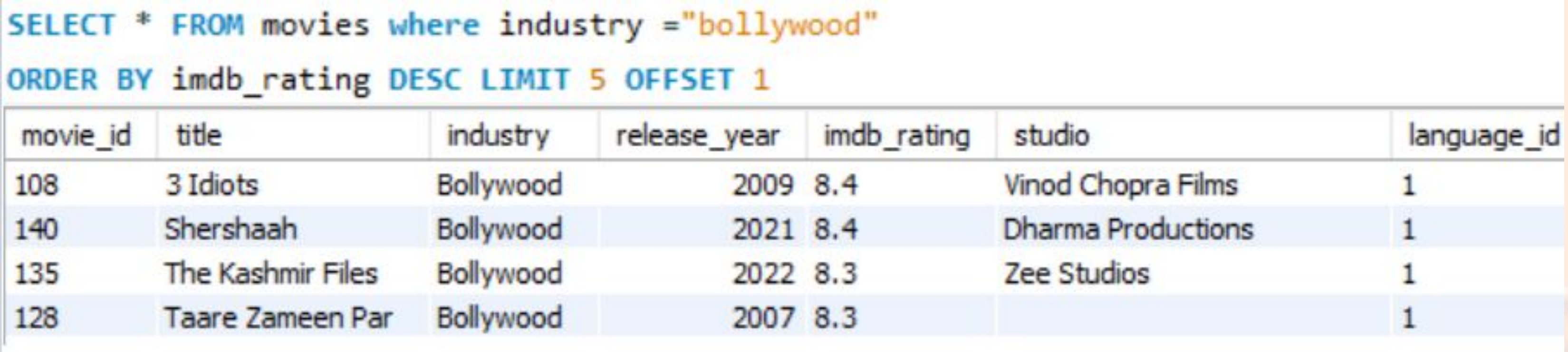
- get titles where imdb_rating is null



```sql
SELECT * FROM movies where imdb_rating is NULL
```

movie_id	title	industry	release_year	imdb_rating	studio	language_id
131	Sanju	Bollywood	2018	NULL	Vinod Chopra Films	1

- get only Bollywood titles where imdb_rating is highest from 2nd onwards and only show 5 titles

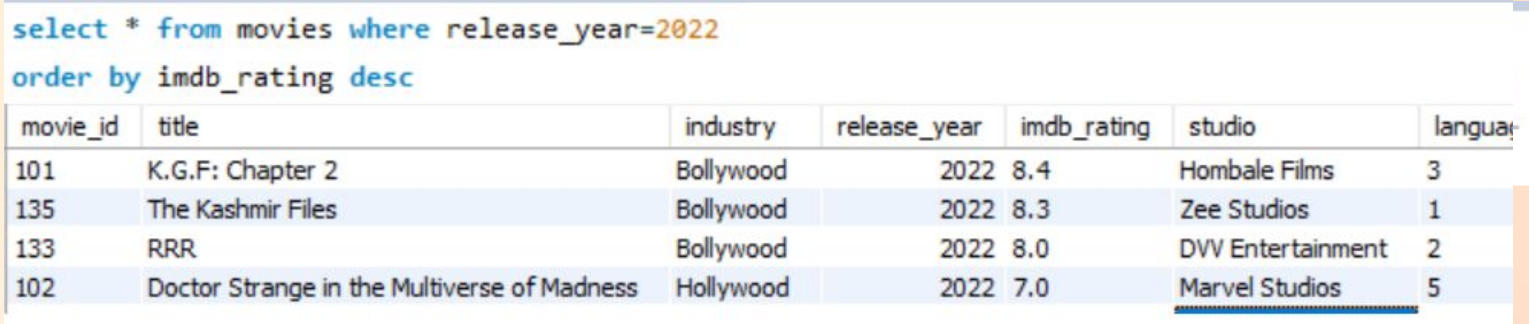


```sql
SELECT * FROM movies where industry ="bollywood"
ORDER BY imdb_rating DESC LIMIT 5 OFFSET 1
```

movie_id	title	industry	release_year	imdb_rating	studio	language_id
108	3 Idiots	Bollywood	2009	8.4	Vinod Chopra Films	1
140	Shershaah	Bollywood	2021	8.4	Dharma Productions	1
135	The Kashmir Files	Bollywood	2022	8.3	Zee Studios	1
128	Taare Zameen Par	Bollywood	2007	8.3		1

Retrieve Data Using Numeric Query

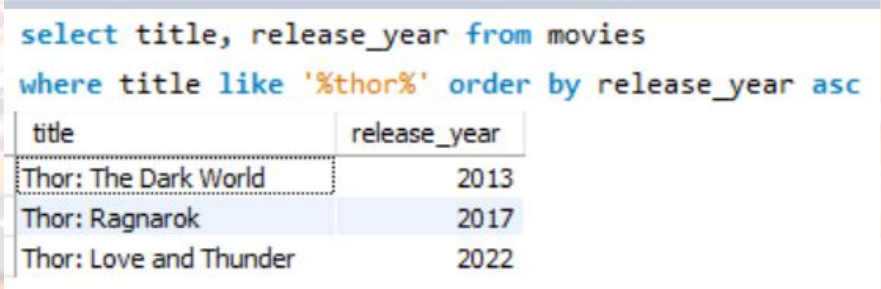
- get titles where release_year=2022 ordered by the highest imdb_rating



```
select * from movies where release_year=2022
order by imdb_rating desc
```

movie_id	title	industry	release_year	imdb_rating	studio	langua
101	K.G.F: Chapter 2	Bollywood	2022	8.4	Hombale Films	3
135	The Kashmir Files	Bollywood	2022	8.3	Zee Studios	1
133	RRR	Bollywood	2022	8.0	DVV Entertainment	2
102	Doctor Strange in the Multiverse of Madness	Hollywood	2022	7.0	Marvel Studios	5

- get titles with the ‘thor’ and ordered by their release year

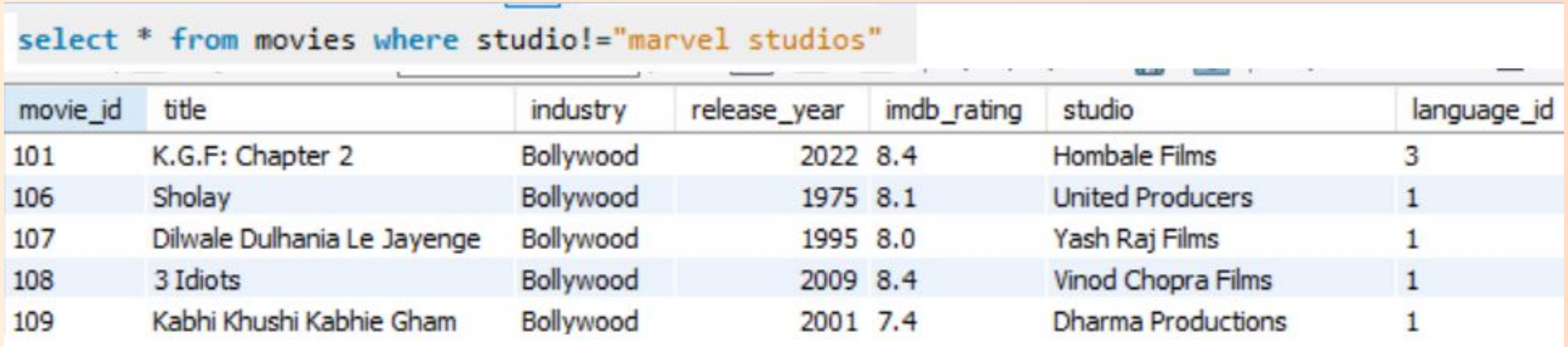


```
select title, release_year from movies
where title like '%thor%' order by release_year asc
```

title	release_year
Thor: The Dark World	2013
Thor: Ragnarok	2017
Thor: Love and Thunder	2022

Retrieve Data Using Numeric Query

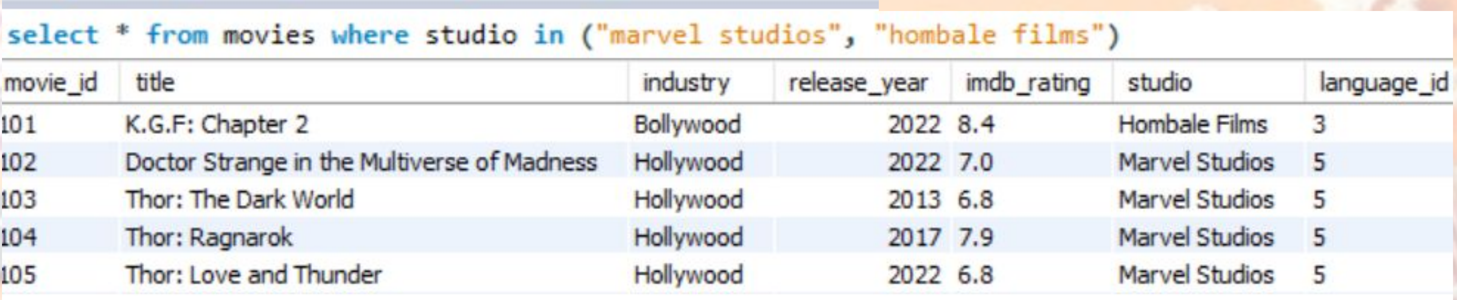
- get movies that are not from marvel studios



```
select * from movies where studio!="marvel studios"
```

movie_id	title	industry	release_year	imdb_rating	studio	language_id
101	K.G.F: Chapter 2	Bollywood	2022	8.4	Hombale Films	3
106	Sholay	Bollywood	1975	8.1	United Producers	1
107	Dilwale Dulhania Le Jayenge	Bollywood	1995	8.0	Yash Raj Films	1
108	3 Idiots	Bollywood	2009	8.4	Vinod Chopra Films	1
109	Kabhi Khushi Kabhie Gham	Bollywood	2001	7.4	Dharma Productions	1

- get titles that are by marvel studios and hombale films

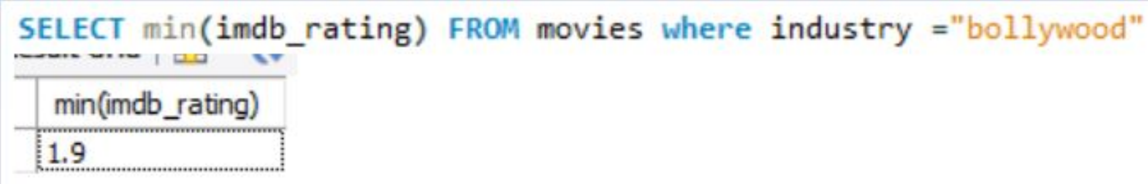


```
select * from movies where studio in ("marvel studios", "hombale films")
```

movie_id	title	industry	release_year	imdb_rating	studio	language_id
101	K.G.F: Chapter 2	Bollywood	2022	8.4	Hombale Films	3
102	Doctor Strange in the Multiverse of Madness	Hollywood	2022	7.0	Marvel Studios	5
103	Thor: The Dark World	Hollywood	2013	6.8	Marvel Studios	5
104	Thor: Ragnarok	Hollywood	2017	7.9	Marvel Studios	5
105	Thor: Love and Thunder	Hollywood	2022	6.8	Marvel Studios	5

Summary Analytics

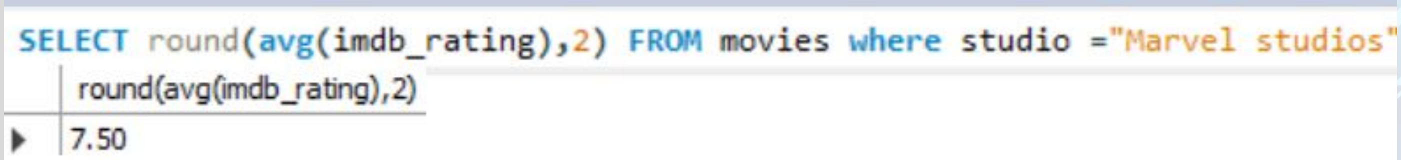
- get the lowest rating of all bollywood movies



```sql
SELECT min(imdb_rating) FROM movies where industry ="bollywood"
```

```
min(imdb_rating)
1.9
```

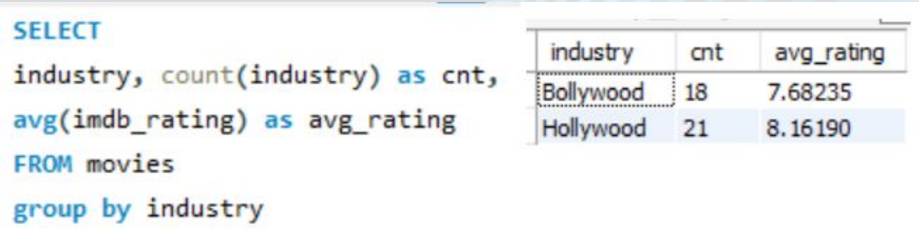
- get the average rating of movies from Marvel studios and round it to 2 decimal places



```sql
SELECT round(avg(imdb_rating),2) FROM movies where studio ="Marvel studios"
```

```
round(avg(imdb_rating),2)
7.50
```

- get the industry and their number of movies and the average rating of these movies

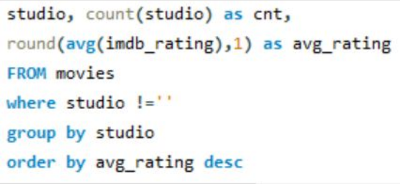


```sql
SELECT
industry, count(industry) as cnt,
avg(imdb_rating) as avg_rating
FROM movies
group by industry
```

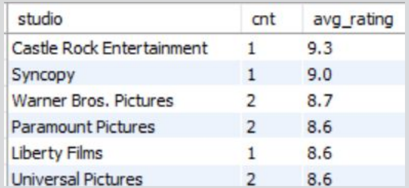
```
industry    cnt   avg_rating
Bollywood   18    7.68235
Hollywood   21    8.16190
```

Summary Analytics

- get all the studios, their number of movies and the average rating

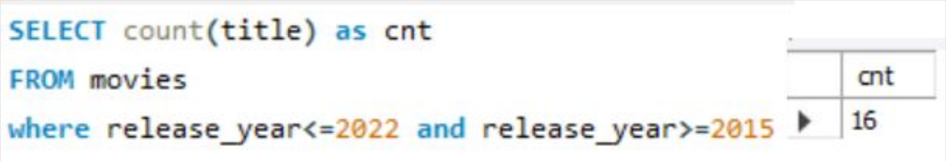


```
studio, count(studio) as cnt,
round(avg(imdb_rating),1) as avg_rating
FROM movies
where studio !=''
group by studio
order by avg_rating desc
```



studio	cnt	avg_rating
Castle Rock Entertainment	1	9.3
Syncopy	1	9.0
Warner Bros. Pictures	2	8.7
Paramount Pictures	2	8.6
Liberty Films	1	8.6
Universal Pictures	2	8.6

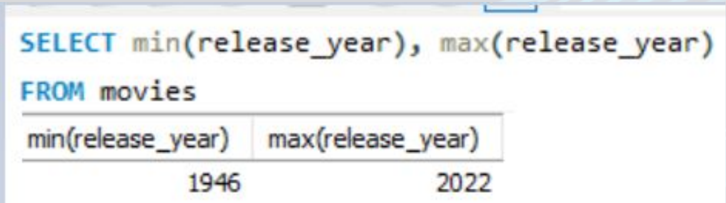
- How many movies were released between 2015 and 2022



```
SELECT count(title) as cnt
FROM movies
where release_year<=2022 and release_year>=2015
```

	cnt
▶	16

- get the max and min movie release year

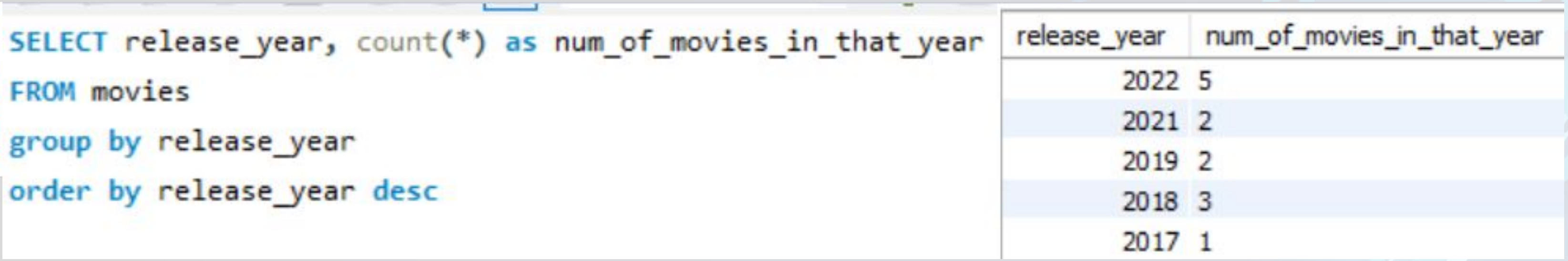


```
SELECT min(release_year), max(release_year)
FROM movies
```

min(release_year)	max(release_year)
1946	2022

Summary Analytics

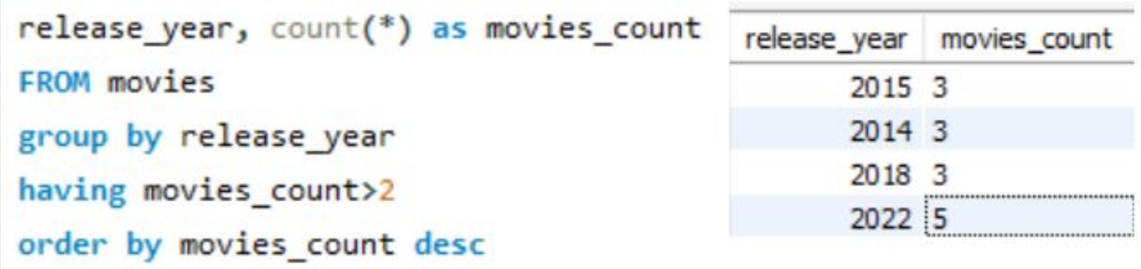
- get year and how many movies were released in that year starting with the latest year



```sql
SELECT release_year, count(*) as num_of_movies_in_that_year
FROM movies
group by release_year
order by release_year desc
```

release_year	num_of_movies_in_that_year
2022	5
2021	2
2019	2
2018	3
2017	1

- get all the years where more than 2 movies were released

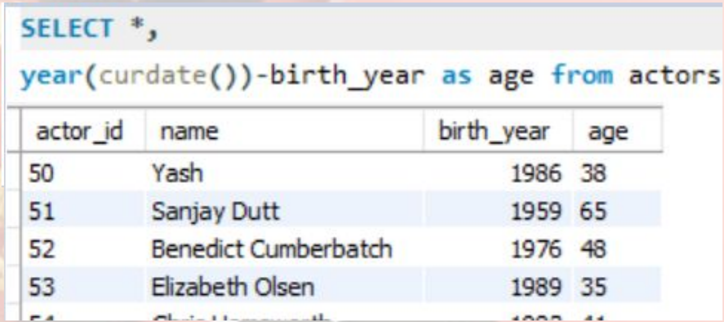


```sql
release_year, count(*) as movies_count
FROM movies
group by release_year
having movies_count>2
order by movies_count desc
```

release_year	movies_count
2015	3
2014	3
2018	3
2022	5

Calculated Columns

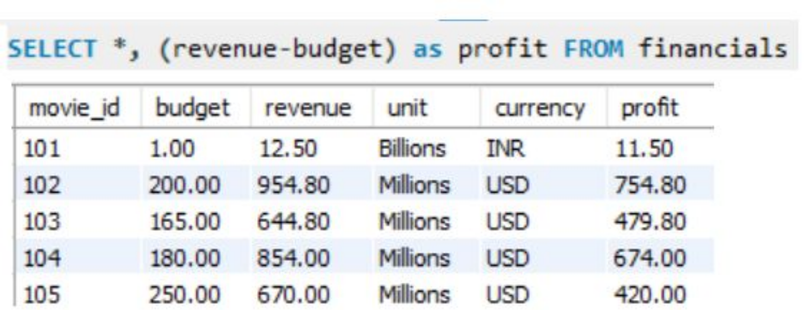
- get all the ages of the actors



```sql
SELECT *,
year(curdate())-birth_year as age from actors
```

actor_id	name	birth_year	age
50	Yash	1986	38
51	Sanjay Dutt	1959	65
52	Benedict Cumberbatch	1976	48
53	Elizabeth Olsen	1989	35

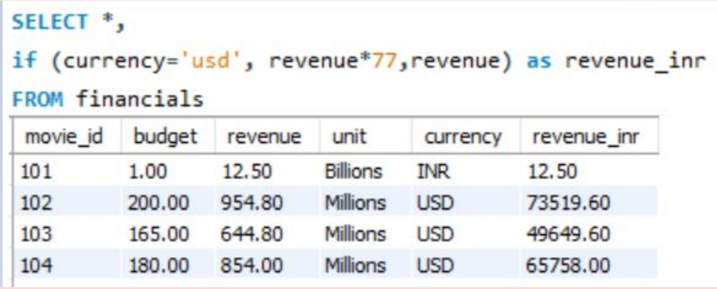
- get profit (revenue-budget) from financials table



```sql
SELECT *, (revenue-budget) as profit FROM financials
```

movie_id	budget	revenue	unit	currency	profit
101	1.00	12.50	Billions	INR	11.50
102	200.00	954.80	Millions	USD	754.80
103	165.00	644.80	Millions	USD	479.80
104	180.00	854.00	Millions	USD	674.00
105	250.00	670.00	Millions	USD	420.00

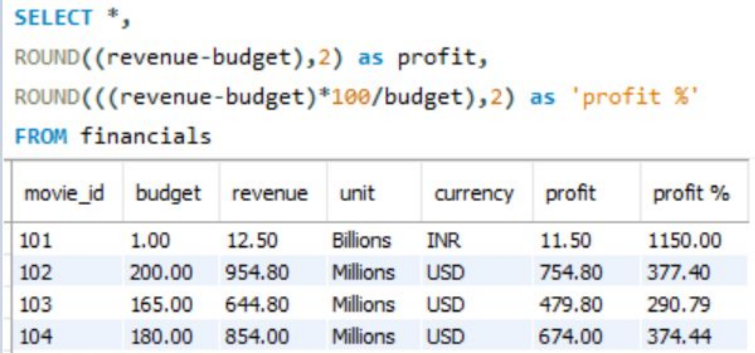
- make a column, revenue_inr, to convert all revenue to INR



```sql
SELECT *,
if (currency='usd', revenue*77,revenue) as revenue_inr
FROM financials
```

movie_id	budget	revenue	unit	currency	revenue_inr
101	1.00	12.50	Billions	INR	12.50
102	200.00	954.80	Millions	USD	73519.60
103	165.00	644.80	Millions	USD	49649.60
104	180.00	854.00	Millions	USD	65758.00

- profit % for all the movies

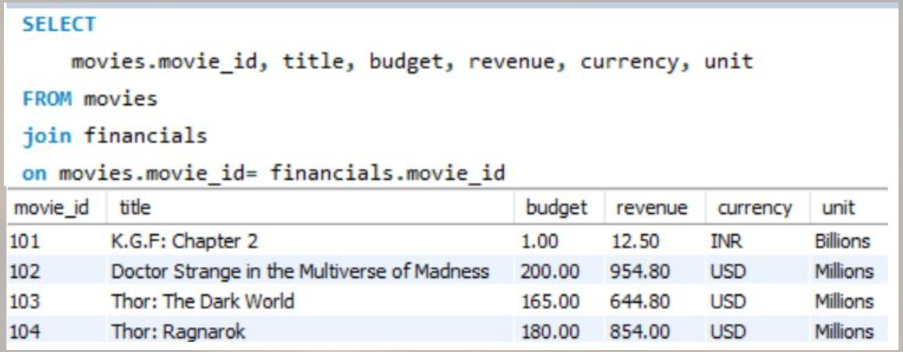


```sql
SELECT *,
ROUND((revenue-budget),2) as profit,
ROUND(((revenue-budget)*100/budget),2) as 'profit %'
FROM financials
```

movie_id	budget	revenue	unit	currency	profit	profit %
101	1.00	12.50	Billions	INR	11.50	1150.00
102	200.00	954.80	Millions	USD	754.80	377.40
103	165.00	644.80	Millions	USD	479.80	290.79
104	180.00	854.00	Millions	USD	674.00	374.44

SQL Joins

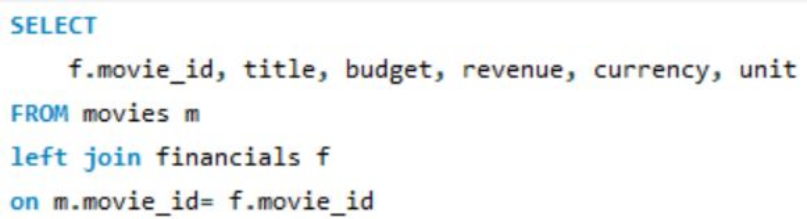
- join movies tables and financials table using movie_id



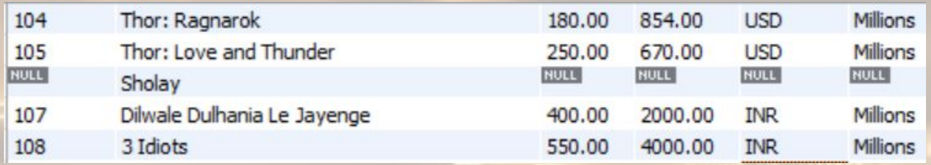
```
SELECT
    movies.movie_id, title, budget, revenue, currency, unit
FROM movies
join financials
on movies.movie_id= financials.movie_id
```

movie_id	title	budget	revenue	currency	unit
101	K.G.F: Chapter 2	1.00	12.50	INR	Billions
102	Doctor Strange in the Multiverse of Madness	200.00	954.80	USD	Millions
103	Thor: The Dark World	165.00	644.80	USD	Millions
104	Thor: Ragnarok	180.00	854.00	USD	Millions

- left join movies tables and financials table using movie_id
- left join show all movies but not the financials details and those shows as null.



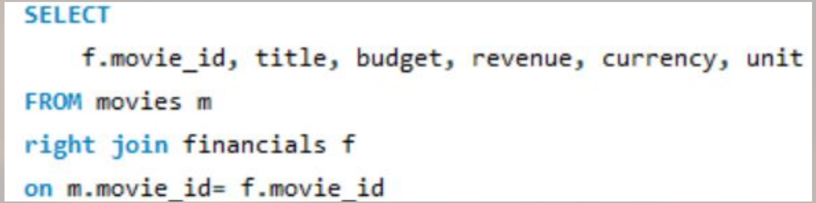
```
SELECT
    f.movie_id, title, budget, revenue, currency, unit
FROM movies m
left join financials f
on m.movie_id= f.movie_id
```



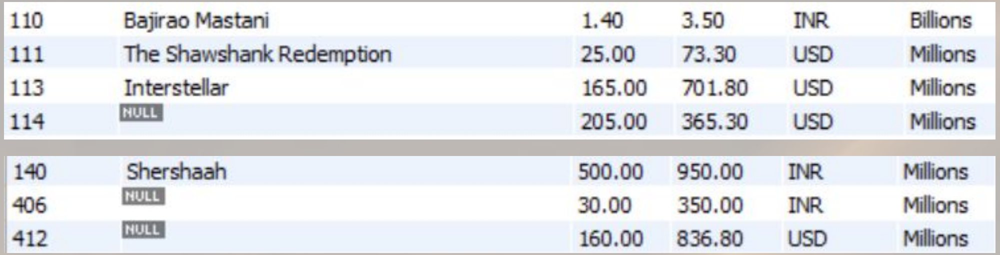
104	Thor: Ragnarok	180.00	854.00	USD	Millions
105	Thor: Love and Thunder	250.00	670.00	USD	Millions
NULL	Sholay	NULL	NULL	NULL	NULL
107	Dilwale Dulhania Le Jayenge	400.00	2000.00	INR	Millions
108	3 Idiots	550.00	4000.00	INR	Millions

SQL Joins

- right join movies tables and financials table using movie_id
- right join show all financials but not the movies and those shows as null.

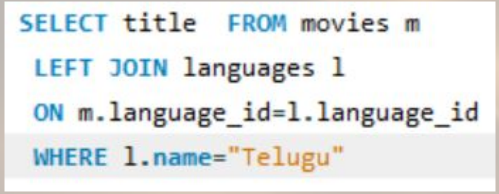


```
SELECT
    f.movie_id, title, budget, revenue, currency, unit
FROM movies m
right join financials f
on m.movie_id= f.movie_id
```

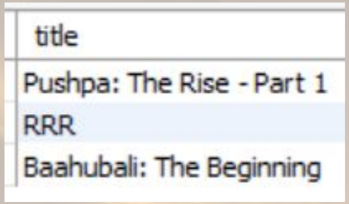


110	Bajirao Mastani	1.40	3.50	INR	Billions
111	The Shawshank Redemption	25.00	73.30	USD	Millions
113	Interstellar	165.00	701.80	USD	Millions
114	NULL	205.00	365.30	USD	Millions
140	Shershaah	500.00	950.00	INR	Millions
406	NULL	30.00	350.00	INR	Millions
412	NULL	160.00	836.80	USD	Millions

- get all Telugu movie names

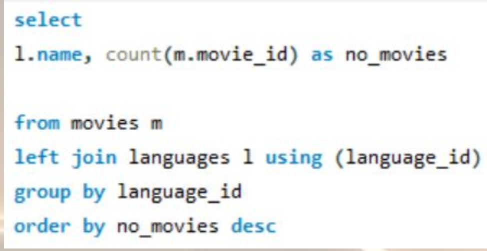


```
SELECT title  FROM movies m
 LEFT JOIN languages l
 ON m.language_id=l.language_id
 WHERE l.name="Telugu"
```

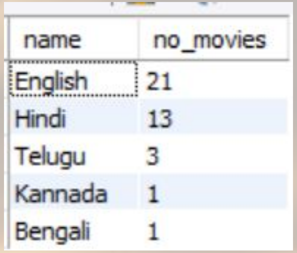


title
Pushpa: The Rise - Part 1
RRR
Baahubali: The Beginning

- get number of movies for each language



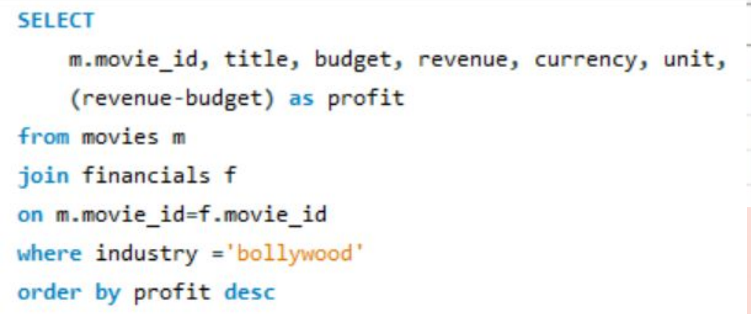
```
select
l.name, count(m.movie_id) as no_movies

from movies m
left join languages l using (language_id)
group by language_id
order by no_movies desc
```



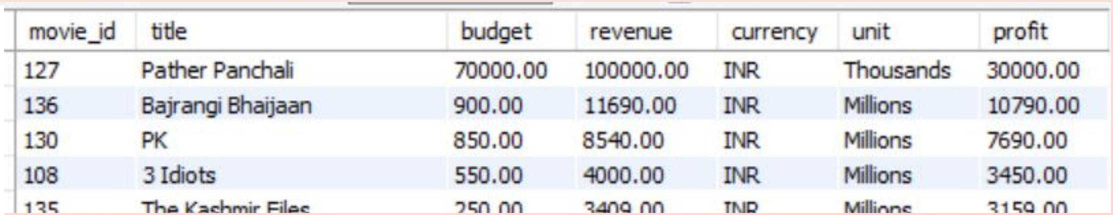
name	no_movies
English	21
Hindi	13
Telugu	3
Kannada	1
Bengali	1

Analytics on Tables

- get bollywood movies and ordered based on amount of profit made

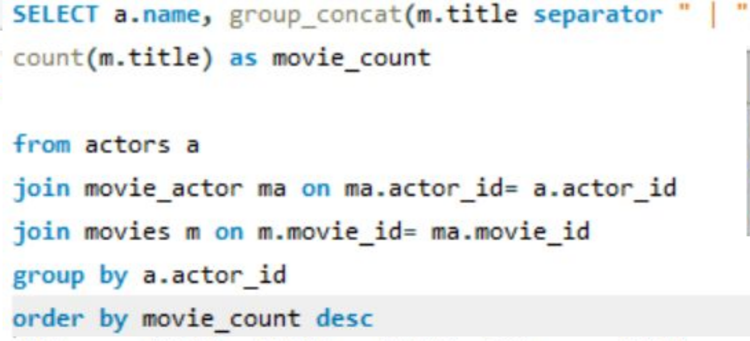


```
SELECT
    m.movie_id, title, budget, revenue, currency, unit,
    (revenue-budget) as profit
from movies m
join financials f
on m.movie_id=f.movie_id
where industry ='bollywood'
order by profit desc
```

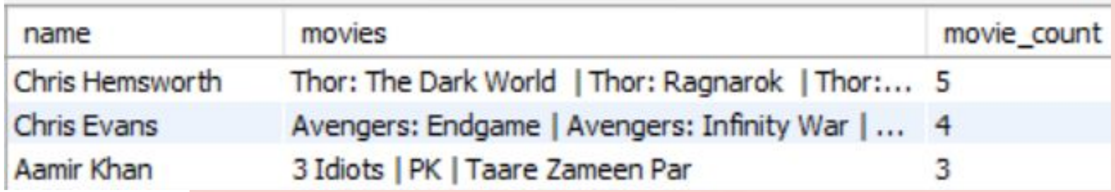


movie_id	title	budget	revenue	currency	unit	profit
127	Pather Panchali	70000.00	100000.00	INR	Thousands	30000.00
136	Bajrangi Bhaijaan	900.00	11690.00	INR	Millions	10790.00
130	PK	850.00	8540.00	INR	Millions	7690.00
108	3 Idiots	550.00	4000.00	INR	Millions	3450.00
135	The Kashmir Files	250.00	3409.00	INR	Millions	3159.00

- get all titles and the number of movies with the same actor in them



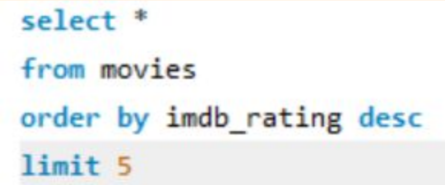
```
SELECT a.name, group_concat(m.title separator " | ") as movies,
count(m.title) as movie_count

from actors a
join movie_actor ma on ma.actor_id= a.actor_id
join movies m on m.movie_id= ma.movie_id
group by a.actor_id
order by movie_count desc
```



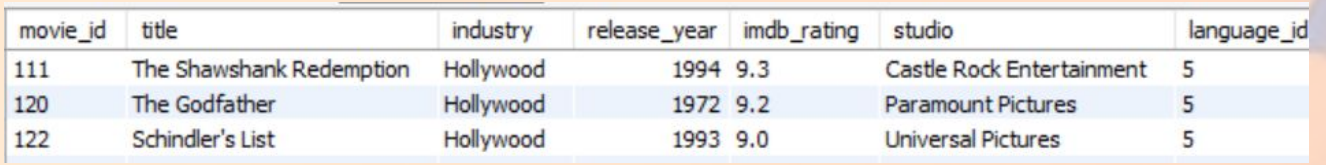
name	movies	movie_count
Chris Hemsworth	Thor: The Dark World | Thor: Ragnarok | Thor:...	5
Chris Evans	Avengers: Endgame | Avengers: Infinity War | ...	4
Aamir Khan	3 Idiots | PK | Taare Zameen Par	3

Subqueries

- get the top 5 most popular titles based on rating

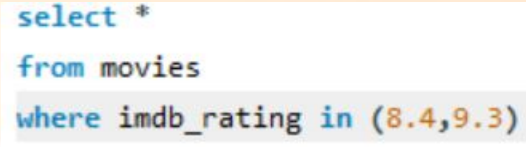


```
select *
from movies
order by imdb_rating desc
limit 5
```

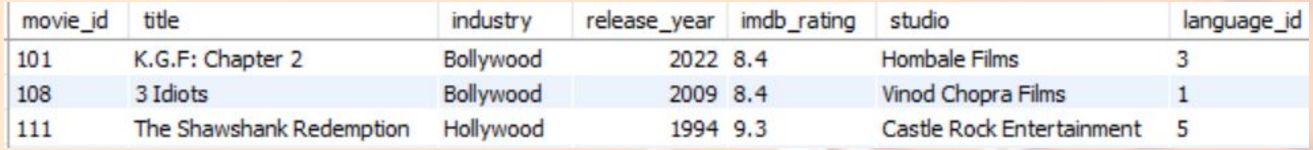


movie_id	title	industry	release_year	imdb_rating	studio	language_id
111	The Shawshank Redemption	Hollywood	1994	9.3	Castle Rock Entertainment	5
120	The Godfather	Hollywood	1972	9.2	Paramount Pictures	5
122	Schindler's List	Hollywood	1993	9.0	Universal Pictures	5

- get the titles with rating 8.4 and 9.3

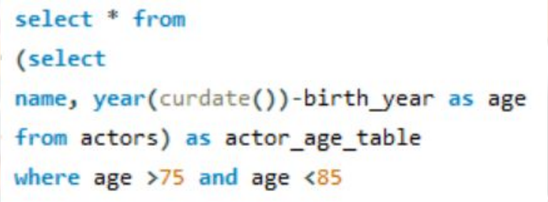


```
select *
from movies
where imdb_rating in (8.4,9.3)
```

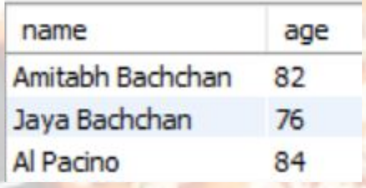


movie_id	title	industry	release_year	imdb_rating	studio	language_id
101	K.G.F: Chapter 2	Bollywood	2022	8.4	Hombale Films	3
108	3 Idiots	Bollywood	2009	8.4	Vinod Chopra Films	1
111	The Shawshank Redemption	Hollywood	1994	9.3	Castle Rock Entertainment	5

- get the actors whose age fall between 75 and 85



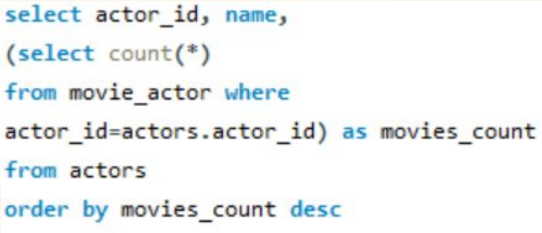
```
select * from
(select
name, year(curdate())-birth_year as age
from actors) as actor_age_table
where age >75 and age <85
```



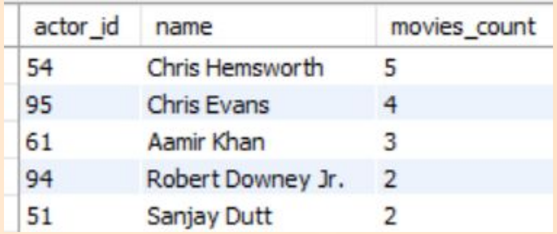
name	age
Amitabh Bachchan	82
Jaya Bachchan	76
Al Pacino	84

Subqueries

- get number of movies each actor is in

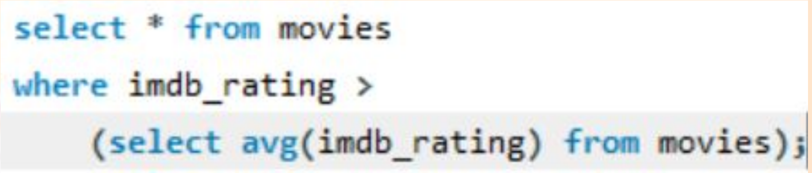


```
select actor_id, name,
(select count(*)
from movie_actor where
actor_id=actors.actor_id) as movies_count
from actors
order by movies_count desc
```

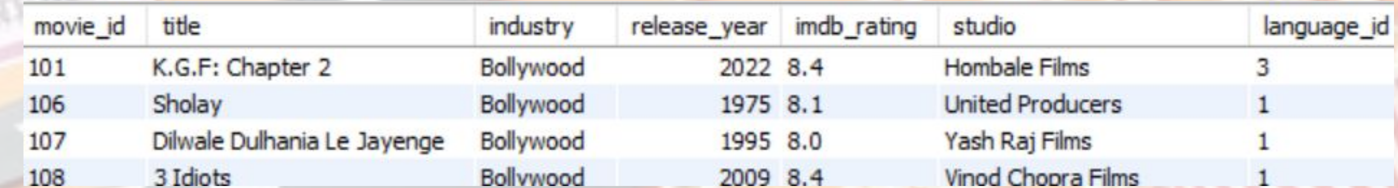


actor_id	name	movies_count
54	Chris Hemsworth	5
95	Chris Evans	4
61	Aamir Khan	3
94	Robert Downey Jr.	2
51	Sanjay Dutt	2

- get all the rows from movies table whose imdb_rating is higher than the average rating



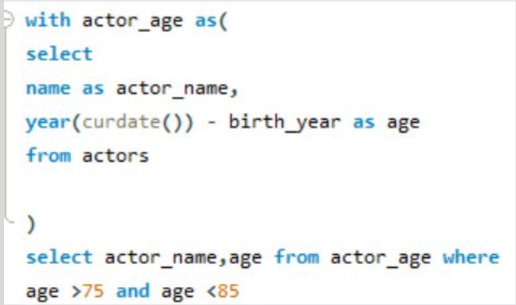
```
select * from movies
where imdb_rating >
    (select avg(imdb_rating) from movies);
```



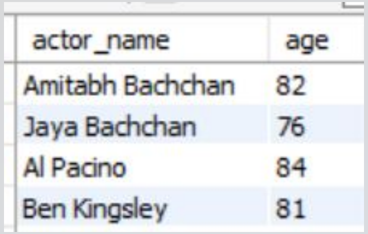
movie_id	title	industry	release_year	imdb_rating	studio	language_id
101	K.G.F: Chapter 2	Bollywood	2022	8.4	Hombale Films	3
106	Sholay	Bollywood	1975	8.1	United Producers	1
107	Dilwale Dulhania Le Jayenge	Bollywood	1995	8.0	Yash Raj Films	1
108	3 Idiots	Bollywood	2009	8.4	Vinod Chopra Films	1

Common Table Expression (CTE)

- get actor name and age whose age >75 and age <85

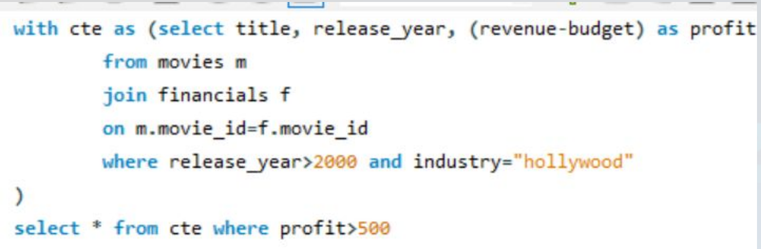


```
with actor_age as(
select
name as actor_name,
year(curdate()) - birth_year as age
from actors

)
select actor_name,age from actor_age where
age >75 and age <85
```

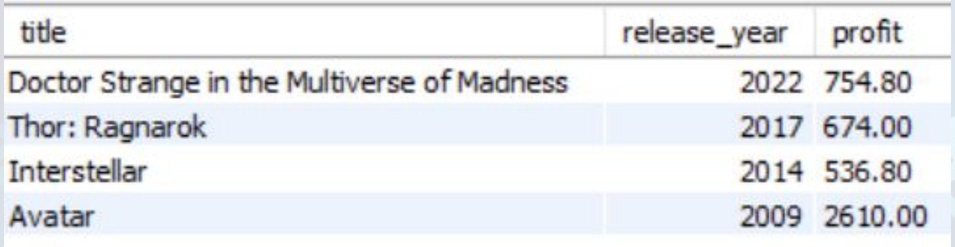


actor_name	age
Amitabh Bachchan	82
Jaya Bachchan	76
Al Pacino	84
Ben Kingsley	81

- get all hollywood movies released after year 2000 that made more than 500 millions $ profit



```
with cte as (select title, release_year, (revenue-budget) as profit
        from movies m
        join financials f
        on m.movie_id=f.movie_id
        where release_year>2000 and industry="hollywood"
)
select * from cte where profit>500
```



title	release_year	profit
Doctor Strange in the Multiverse of Madness	2022	754.80
Thor: Ragnarok	2017	674.00
Interstellar	2014	536.80
Avatar	2009	2610.00

Thank you