USE imdb;
/* Now that you have imported the data sets, let's explore some of the tables.
To begin with, it is beneficial to know the shape of the tables and whether any column has null values.
Further in this segment, you will take a look at 'movies' and 'genre' tables.*/
Segment 1:
Q1. Find the total number of rows in each table of the schema?
Type your code below:
Q2. Which columns in the movie table have null values?
Type your code below:
SELECT
SUM(CASE WHEN id IS NULL THEN 1 ELSE 0 END) AS id,

 $\ensuremath{\mathsf{SUM}}\xspace(\ensuremath{\mathsf{CASE}}\xspace$  WHEN title IS NULL THEN 1 ELSE 0 END) AS title,

SUM(CASE WHEN year IS NULL THEN 1 ELSE 0 END) AS year,

SUM(CASE WHEN date\_published IS NULL THEN 1 ELSE 0 END) AS date\_published,

SUM(CASE WHEN duration IS NULL THEN 1 ELSE 0 END) AS duration,

SUM(CASE WHEN country IS NULL THEN 1 ELSE 0 END) AS country,

SUM(CASE WHEN worlwide\_gross\_income IS NULL THEN 1 ELSE 0 END) AS worlwide\_gross\_income,

SUM(CASE WHEN languages IS NULL THEN 1 ELSE 0 END) AS languages,

SUM(CASE WHEN production\_company IS NULL THEN 1 ELSE 0 END) AS production\_company FROM movie;

- -- Now as you can see four columns of the movie table has null values. Let's look at the at the movies released each year.
- -- Q3. Find the total number of movies released each year? How does the trend look month wise? (Output expected)

+	+	+		
	2019	1		١
	2018	1	•	

Output format for the second part of the question:

SELECT

EXTRACT(MONTH FROM date\_published) AS month,
COUNT(\*) AS total\_movies

FROM movie

GROUP BY EXTRACT(MONTH FROM date\_published)

ORDER BY month;

-- Type your code below:

1	<b> </b>	<b>'</b>				
1	month_num		numbe	er_of_movies		
+	·+					
1	1		1	134	1	
1	2		1	231	I	
1			1			
+	+	+ *	<b>'</b> /			

<sup>/\*</sup>The highest number of movies is produced in the month of March.

So, now that you have understood the month-wise trend of movies, let's take a look at the other details in the movies table.

We know USA and India produces huge number of movies each year. Lets find the number of movies produced by USA or India for the last year.\*/

Q4. Ho	w many movies	were produced	in the USA or	India in the	year 2019??
--------	---------------	---------------	---------------	--------------	-------------

-- Type your code below:

Select country, year, count(\*) as no\_of\_movie from movie where year=2019 and country in('USA', 'India') group by country, year

India 2019 295

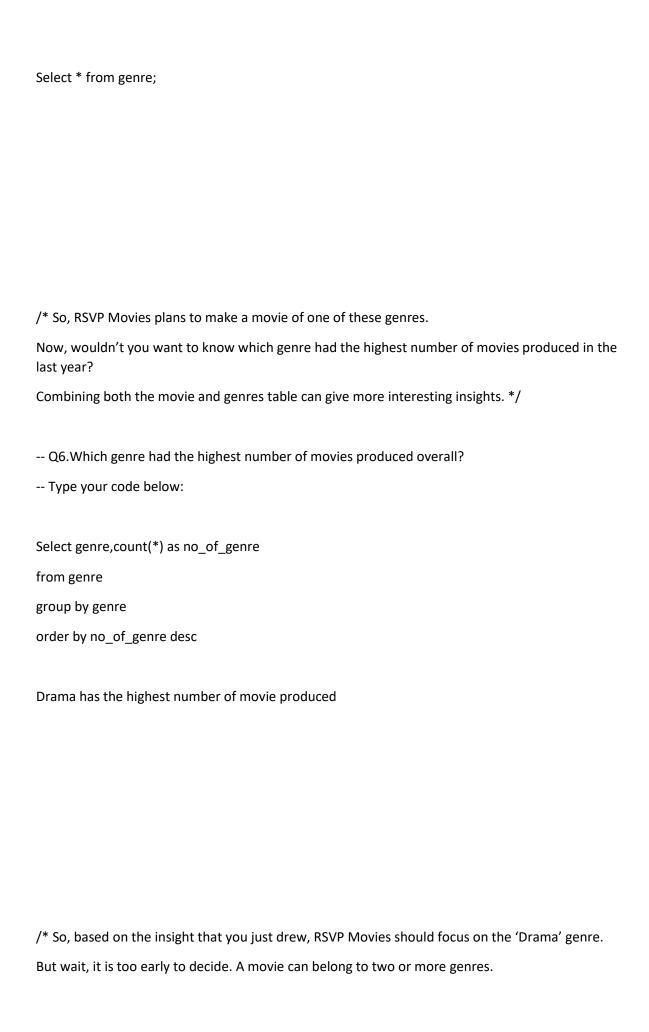
USA 2019 592

/\* USA and India produced more than a thousand movies(you know the exact number!) in the year 2019.

Exploring table Genre would be fun!!

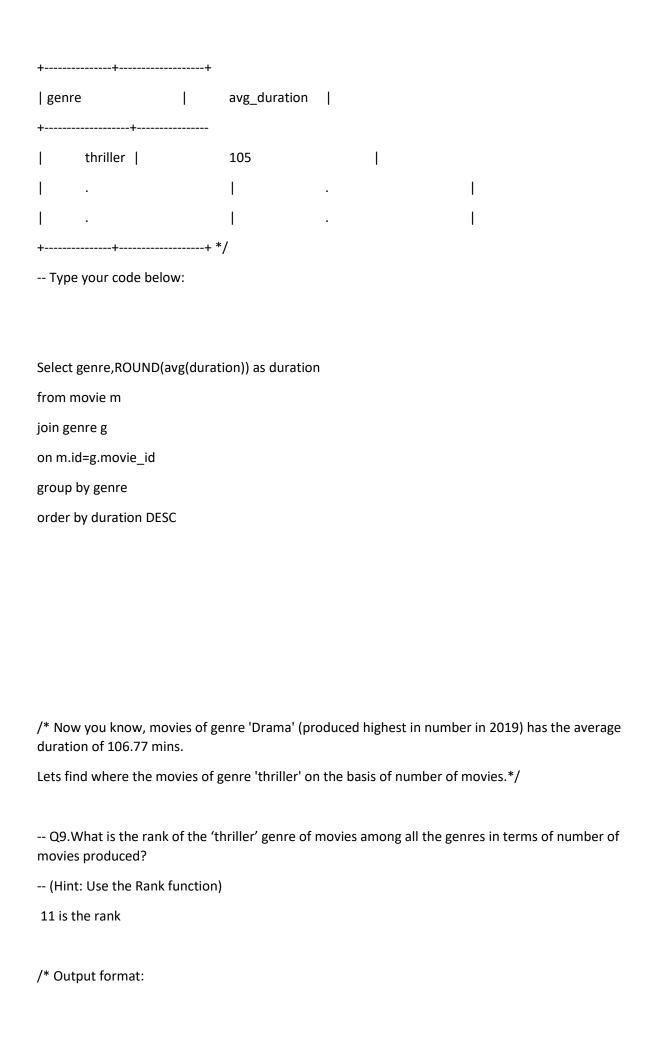
Let's find out the different genres in the dataset.\*/

- -- Q5. Find the unique list of the genres present in the data set?
- -- Type your code below:



-- Q7. How many movies belong to only one genre? -- Type your code below: SELECT genre, COUNT(\*) AS single\_genre\_count FROM ( SELECT movie\_id FROM genre GROUP BY movie\_id HAVING COUNT(genre) = 1 ) AS single\_genre\_movie; /\* There are more than three thousand movies which has only one genre associated with them. So, this figure appears significant. Now, let's find out the possible duration of RSVP Movies' next project.\*/ -- Q8. What is the average duration of movies in each genre? -- (Note: The same movie can belong to multiple genres.) /\* Output format:

So, let's find out the count of movies that belong to only one genre.\*/ -3289



т		т			
genre	I	movie_count	1	genre_rank	
+	+	+			
drama	2312		I	2	ļ
+	+	+*/			
Type your code b	elow:				
Select genre,					
RANK() over (order	by no_of_genre) a	s RANK			
from					
(Select genre,count	(*) as no_of_genre	9			
from genre					
group by genre)					
as Genre_counts					
order by rank					

/\*Thriller movies is in top 3 among all genres in terms of number of movies
In the previous segment, you analysed the movies and genres tables.
In this segment, you will analyse the ratings table as well.

To start with lets get the min and max values of different columns in the table\*/

movie_id column?  /* Output format: +	
++++	
min_avg_rating   max_avg_rating   min_total_vote   min_median_rating   min_total_vote   min_median_rating	tes   max_total_votes
++	8   177

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

/\* So, the minimum and maximum values in each column of the ratings table are in the expected range.

This implies there are no outliers in the table.

-- Segment 2:

Now, let's find out the top 10 movies based on average rating.\*/

-- Q11. Which are the top 10 movies based on average rating? /\* Output format: | title avg\_rating | movie\_rank | | Fan 9.6 5 +----+\*/ -- Type your code below: -- It's ok if RANK() or DENSE\_RANK() is used too Select title,avg\_rating,rank FROM( Select m.title,r.avg\_rating, DENSE\_RANK() over (order by r.avg\_rating DESC) as rank from movie m join ratings r on m.id=r.movie\_id) as movie\_ratings where rank<=10 ORDER BY rank asc

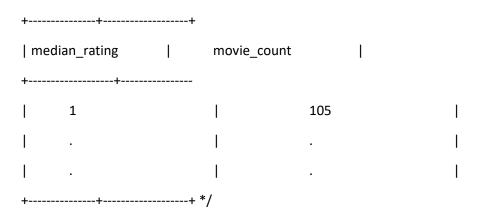
/\* Do you find you favourite movie FAN in the top 10 movies with an average rating of 9.6? If not, please check your code again!!

So, now that you know the top 10 movies, do you think character actors and filler actors can be from these movies?

Summarising the ratings table based on the movie counts by median rating can give an excellent insight.\*/

-- Q12. Summarise the ratings table based on the movie counts by median ratings.

/\* Output format:



- -- Type your code below:
- -- Order by is good to have

select median\_rating,count(\*) as movie\_counts
from ratings
group by median\_rating,movie\_id
order by movie\_counts

/* Movies with a median r	ating of 7 is hig	ghest in number.		
Now, let's find out the pro	duction house	with which RSVP N	Movies can partner fo	r its next project.*/
Q13. Which production  /* Output format:  +			mber of hit movies (a	verage rating > 8)??
production_company mo	ovie_count	prod_comp	pany_rank	
The Archers   +	1	I	1	I
Type your code below:	·	,		
SELECT production_compa	any,			
COUNT(*) AS hit_movi	ie_count			
FROM movie m				

-- It's ok if RANK() or DENSE\_RANK() is used too

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

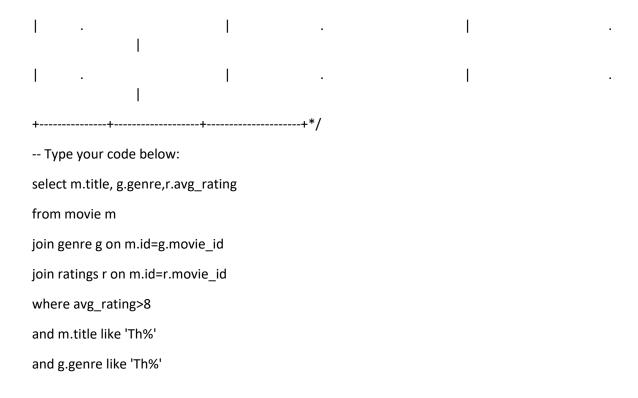
GROUP BY production\_company

ORDER BY hit\_movie\_count DESC

WHERE r.avg\_rating > 8

-- Answer can be Dream Warrior Pictures or National Theatre Live or both

	4. How many m O votes?	ovies rel	eased in	each genre di	uring Marcl	h 2017 in	the USA h	ad more th	an
/* Oı	utput format:								
+		+	-						
ger	nre		movie	_count	I				
+	thriller		105		I				
	•		I				1		
I			I				I		
+	+	<del>-</del>	· */						
Тур	oe your code be	low:							
Let	s try to analyse	with a u	nique pro	blem statem	ent.				
Q1 8?	5. Find movies o	of each g	enre that	start with th	e word 'Th	e' and wh	nich have a	ın average r	ating >
/* Oı	utput format:								
+	+	+		+					
title	e	I		avg_rating	I		genre	1	
+		+		+					
The	eeran	1		8.3		I	7	Thriller	
1			I						
	l								



- -- You should also try your hand at median rating and check whether the 'median rating' column gives any significant insights.
- -- Q16. Of the movies released between 1 April 2018 and 1 April 2019, how many were given a median rating of 8?
- -- Type your code below:

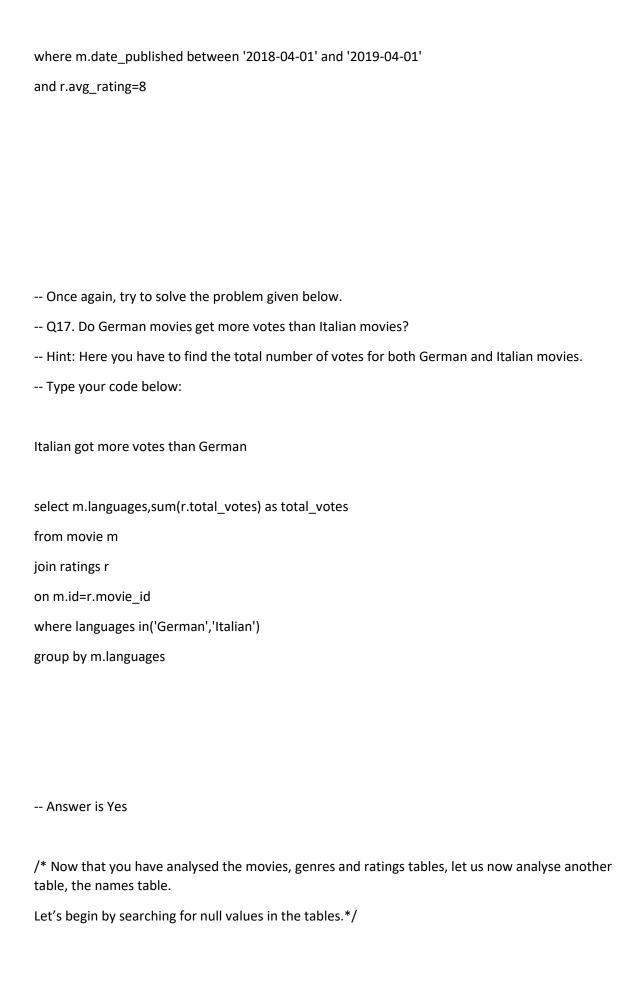
20

Select count(\*)

from movie m

join ratings r

on m.id=r.movie\_id



## -- Q18. Which columns in the names table have null values?? /\*Hint: You can find null values for individual columns or follow below output format +-----+ | name\_nulls | height\_nulls |date\_of\_birth\_nulls |known\_for\_movies\_nulls| +-----+ 0 | 123 | 1234

-- Type your code below:

12345

+-----+\*/

-- Segment 3:

## select

sum(case when id is null then 1 else 0 end) as id,
sum(case when name is null then 1 else 0 end) as name,
sum(case when height is null then 1 else 0 end) as height,
sum(case when date\_of\_birth is null then 1 else 0 end)as date\_birth,
sum(case when known\_for\_movies is null then 1 else 0 end)as known\_for\_movies
from names;

/\* There are no Null value in the column 'name'.

The director is the most important person in a movie crew.

Let's find out the top three directors in the top three genres who can be hired by RSVP Movies.\*/

```
-- Q19. Who are the top three directors in the top three genres whose movies have an average rating
> 8?
-- (Hint: The top three genres would have the most number of movies with an average rating > 8.)
/* Output format:
+----+
| director_name | movie_count
+-----
|James Mangold |
                           +----+*/
-- Type your code below:
Select d.name_id,g.genre,AVG(r.avg_rating) as director_avg_rating
      from director_mapping d
      join genre g on d.movie_id=g.movie_id
      join ratings r on d.movie_id=r.movie_id
      where g.genre in(SELECT
 g.genre
 FROM genre g
 JOIN ratings r ON g.movie_id = r.movie_id
 WHERE r.avg_rating > 8
 GROUP BY g.genre
 LIMIT 3)
GROUP BY g.genre, d.name_id
ORDER BY g.genre, director_avg_rating
LIMIT 3;
```

/\* James Mangold can be hired as the director for RSVP's next project. Do you remeber his movies, 'Logan' and 'The Wolverine'. Now, let's find out the top two actors.\*/ -- Q20. Who are the top two actors whose movies have a median rating >= 8? /\* Output format: +----+ actor\_name movie\_count +-----|Christain Bale | 10 1 +----+\*/ -- Type your code below: Select name,name\_id,avg\_median\_rating,row\_num from (Select n.name,ro.name\_id,AVG(r.median\_rating) as avg\_median\_rating, ROW\_NUMBER() OVER (ORDER BY AVG(r.median\_rating) DESC) AS row\_num from names n join role\_mapping ro on n.id=ro.name\_id join ratings r on r.movie\_id=ro.movie\_id where r.median\_rating>=8

```
group by n.name,ro.name_id
) as ranked_actors
where row_num<=2
I dint find Mohanlal in the list
/* Have you find your favourite actor 'Mohanlal' in the list. If no, please check your code again.
RSVP Movies plans to partner with other global production houses.
Let's find out the top three production houses in the world.*/
-- Q21. Which are the top three production houses based on the number of votes received by their
movies?
/* Output format:
+-----+
|production_company|vote_count
                                            prod_comp_rank
+-----+
                                                 | The Archers
                               830
                                                               1
+----+*/
-- Type your code below:
Select m.production_company,count(distinct m.id),sum(r.total_votes) as Total_votes
from movie m
join ratings r
on m.id=r.movie_id
group by production_company
order by Total_votes desc
```

/\*Yes Marvel Studios rules the movie world.

So, these are the top three production houses based on the number of votes received by the movies they have produced.

Since RSVP Movies is based out of Mumbai, India also wants to woo its local audience.

RSVP Movies also wants to hire a few Indian actors for its upcoming project to give a regional feel. Let's find who these actors could be.\*/

- -- Q22. Rank actors with movies released in India based on their average ratings. Which actor is at the top of the list?
- -- Note: The actor should have acted in at least five Indian movies.
- -- (Hint: You should use the weighted average based on votes. If the ratings clash, then the total number of votes should act as the tie breaker.)

## /\* Output format:

+	+	+		+			+	·+	-		
acto	r_name   actor_avg_rat	total_v ing	otes  actor_ra	ank	1	mo	vie_cou	nt	I		
+		+		+			+	+	•		
I	Yogi Babu I	I	1	1	3455	I		11	1	8.42	
1			Ţ	•				I			I
	•		 		•						
						•	1	l	•		١

```
-- Type your code below:
WITH ActorMovieCounts AS (
  SELECT
    ro.name_id,
    COUNT(ro.movie_id) AS movie_count
  FROM
    role_mapping ro
  GROUP BY
    ro.name\_id
  HAVING
    COUNT(ro.movie_id) > 5
)
SELECT
  n.name,
  m.country,
  r.avg_rating,
  ro.category,
  amc.movie_count
FROM
  role_mapping ro
JOIN
  names n ON ro.name_id = n.id
JOIN
  ratings r ON ro.movie_id = r.movie_id
JOIN
  movie m ON r.movie_id = m.id
JOIN
```

ActorMovieCounts amc ON ro.name\_id = amc.name\_id WHERE m.country = 'India' AND r.avg\_rating >= 8 AND ro.category='actor' **ORDER BY** r.avg\_rating DESC; "nm0001375" -- Top actor is Vijay Sethupathi -- Q23. Find out the top five actresses in Hindi movies released in India based on their average ratings? -- Note: The actresses should have acted in at least three Indian movies. -- (Hint: You should use the weighted average based on votes. If the ratings clash, then the total number of votes should act as the tie breaker.) /\* Output format: actress\_name | total\_votes | movie\_count actress\_avg\_rating | actress\_rank | | Tabu | | 1 3455 | 11 | 8.42

```
-- Type your code below:
With ActressMoviecount as(
       select ro.name_id,ro.category,count(ro.movie_id) as movie_count
       from role_mapping ro
       group by name_id,ro.category
       having count(ro.movie_id)>=3
       and ro.category='actress'
)
Select m.country,m.languages,n.name,r.avg_rating,amc.movie_count,
ROW_NUMBER() OVER (ORDER BY r.avg_rating DESC) AS rank
from movie m
join ratings r
on m.id=r.movie_id
join role_mapping ro
on ro.movie_id=r.movie_id
join names n
on n.id=ro.name_id
join ActressMoviecount amc
on amc.name_id=ro.name_id
where m.country = 'India'
  and m.languages='Hindi'
order by r.avg_rating DESC
limit 3;
```

```
/* Taapsee Pannu tops with average rating 7.74.
Now let us divide all the thriller movies in the following categories and find out their numbers.*/
/* Q24. 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
-- Type your code below:
Select m.title,g.movie_id,genre,r.avg_rating,
        case
            when r.avg_rating > 8 then 'Superhit movies'
                       when r.avg_rating between 7 and 8 then 'Hit movies'
                       when r.avg_rating between 5 and 7 then 'One time watch movies'
                       else 'Flop_movie'
        end as movie_category
from genre g
join ratings r
on g.movie_id=r.movie_id
join movie m
on m.id=r.movie_id
where genre='Thriller'
order by r.avg_rating desc
```

Now, you will pe segment.*/	rform some	tasks that will give	e you a broa	der unders	tanding of th	e data in
Segment 4:						
Q25. What is t	he genre-wi	se running total ar	nd moving av	verage of t	he average m	ovie dura
•		he output table in	the question	n.)		
/* Output forma		++		+		
genre	1	avg_duration	running_	total_dura	tion moving_	_avg_dur
+		++	145	+	106.2	I
	,	l I	·		1	•
		 			I	
		1			I	
1 .						

Round is good	I to hav	e and n	ot a must h	nave; S	Same th	ing a	ıppli	es to sorting		
Let us find top	5 mov	vies of ea	ach year w	ith top	o 3 genr	es.				
Q26. Which a									the top thr	ee genres?
/* Output forma				+				+	_	
genre  worldwide_gro		1	year			I		movie_name		
+		+		+-				++	-	
comedy \$103244842		1	1		2017	1		indian	1	
1			Ļ				ı	1		1
I			1		•		I	1		1
· I			 				1	I		I
+t	le belov	w:						<del>+</del> +	*/	
WITH TopGenre	s AS (									
SELECT										

```
g.genre,
    SUM(CAST(REGEXP_REPLACE(m.worlwide_gross_income, '[^\d.]', ", 'g') AS NUMERIC)) AS
total_gross_income
  FROM
    genre g
  JOIN
    movie m ON m.id = g.movie_id
  WHERE
    m.worlwide_gross_income IS NOT NULL
  GROUP BY
    g.genre
  ORDER BY
    total_gross_income DESC
  LIMIT 3
),
TopMovies AS (
  SELECT
    m.title,
    EXTRACT(year FROM m.date_published) AS year_published,
    CAST(REGEXP_REPLACE(m.worlwide_gross_income, '[^\d.]', '', 'g') AS NUMERIC) AS
worlwide_gross_income,
    g.genre,
    ROW_NUMBER() OVER (
      PARTITION BY EXTRACT(year FROM m.date_published), g.genre
      ORDER BY CAST(REGEXP_REPLACE(m.worlwide_gross_income, '[^\d.]', ", 'g') AS NUMERIC)
DESC
    ) AS rank_within_year_genre
  FROM
    movie m
  JOIN
    genre g ON m.id = g.movie_id
  WHERE
```

```
g.genre IN (SELECT genre FROM TopGenres)

AND m.worlwide_gross_income IS NOT NULL
)

SELECT

title,

year_published,

genre,

worlwide_gross_income,

rank_within_year_genre

FROM

TopMovies

WHERE

rank_within_year_genre <= 5

ORDER BY

year_published, genre, rank_within_year_genre;
```

- -- Finally, let's find out the names of the top two production houses that have produced the highest number of hits among multilingual movies.
- -- Q27. Which are the top two production houses that have produced the highest number of hits (median rating >= 8) among multilingual movies?

```
-- Type your code below:
with Mutilingual as
        (select production_company,median_rating
from movie m
join ratings r
on m.id=r.movie_id
where production_company is not null
and median_rating>=8
and length(languages)-length(replace(languages,',',''))>0
order by median_rating desc
),
ProductionHits as
(
       select production_company,count(*) as hit_count
       from Mutilingual
       group by production_company
)
Select production_company,hit_count
from ProductionHits
order by hit_count desc
```

limit 2

- -- Multilingual is the important piece in the above question. It was created using POSITION(',' IN languages)>0 logic
- -- If there is a comma, that means the movie is of more than one language

-- Q28. Who are the top 3 actresses based on number of Super Hit movies (average rating >8) in drama genre?

```
/* Output format:
+-----+
| actress_name | total_votes | movie_count
|actress_avg_rating |actress_rank |
  Laura Dern |
| 1
                       1016 | 1 |
                                              9.60
-- Type your code below:
```

```
with Tophit as(
       select name_id,
       count(*) as hitcount
       from ratings r
  join role_mapping ro
  on r.movie_id=ro.movie_id
       join genre g
       on g.movie_id=ro.movie_id
       where avg_rating>8
```

```
and category like 'actress'
and genre like 'Drama'
group by name_id
order by hitcount
),
TopActress as(
       select n.name,t.hitcount
       from names n
       join Tophit t
       on n.id=t.name_id
)
Select name, hit count
from TopActress
order by hitcount desc
limit 3
/* Q29. Get the following details for top 9 directors (based on number of movies)
Director id
Name
Number of movies
Average inter movie duration in days
Average movie ratings
Total votes
Min rating
Max rating
total movie durations
```

avg_rating			total_votes   min_rating				per_of_movies   avg_inter_movie_days     max_rating   total_duration						
 + nm1777967	+	'	A.L. Vi	·				,	5	-1			
	77	ı	613	5.65	I I	1 1754	I	3.7	J	I	6.		
	I		I	ı			I	 		1			
·	' 		I	ı		·	I	 		1			
·	l I		I	I			ı			1			
	ı'		I	I			'			1			
	l I		ı	I	I		l			l	٠		
·	I		'	I	Ī		I			I	•		
	I I		l	I	·		I	  -  -		I			
	I		I	ı	1		I	  -  -		1			
·	' 		I	' 		·	I	'    -  -		I			

<sup>--</sup> Type you code below:

```
WITH DirectorMovies AS (
  SELECT
    dm.name_id,
    m.id AS movie_id,
    m.date_published,
    LAG(m.date_published) OVER (PARTITION BY dm.name_id ORDER BY m.date_published)
       AS prev_date_published
  FROM director_mapping dm
  JOIN movie m ON dm.movie_id = m.id
),
InterMovieDurations AS (
  SELECT
    name_id,
    EXTRACT(DAY FROM (date_published - prev_date_published)) AS inter_movie_duration
  FROM DirectorMovies
  WHERE prev_date_published IS NOT NULL
),
DirectorStats AS (
  SELECT
    dm.name_id,
    COUNT(dm.movie_id) AS movie_count,
    AVG(imd.inter_movie_duration) AS avg_inter_movie_duration,
    AVG(r.avg_rating) AS avg_movie_rating,
    SUM(r.total_votes) AS total_votes,
    MIN(r.avg_rating) AS min_rating,
    MAX(r.avg_rating) AS max_rating,
    SUM(m.duration) AS total_movie_duration
  FROM director_mapping dm
  JOIN movie m ON dm.movie_id = m.id
  JOIN ratings r ON m.id = r.movie_id
```

```
LEFT JOIN InterMovieDurations imd ON dm.director_id = imd.director_id
  GROUP BY dm.director_id
  ORDER BY movie_count DESC
  LIMIT 9
)
SELECT
  ds.director_id,
  n.name,
  ds.movie_count,
  ds.avg_inter_movie_duration,
  ds.avg_movie_rating,
  ds.total_votes,
  ds.min_rating,
  ds.max_rating,
  ds.total\_movie\_duration
FROM DirectorStats ds
JOIN names n ON ds.director_id = n.id;
```