IMDB DATA ANALYSIS PROJECT

Answering essential business questions with SQL

INTRODUCTION

In the realm of film and cinema, data speaks volumes. With an industry that produces thousands of films worldwide each year, there's a wealth of information waiting to be discovered. This project aims to delve into this vast ocean of data, using the IMDb dataset as our guide.

The IMDb dataset is a rich source of information, containing various tables that provide detailed data about movies, their directors, actors, genres, and ratings. This project leverages SQL, a powerful tool for managing and manipulating structured data, to answer advanced business questions.

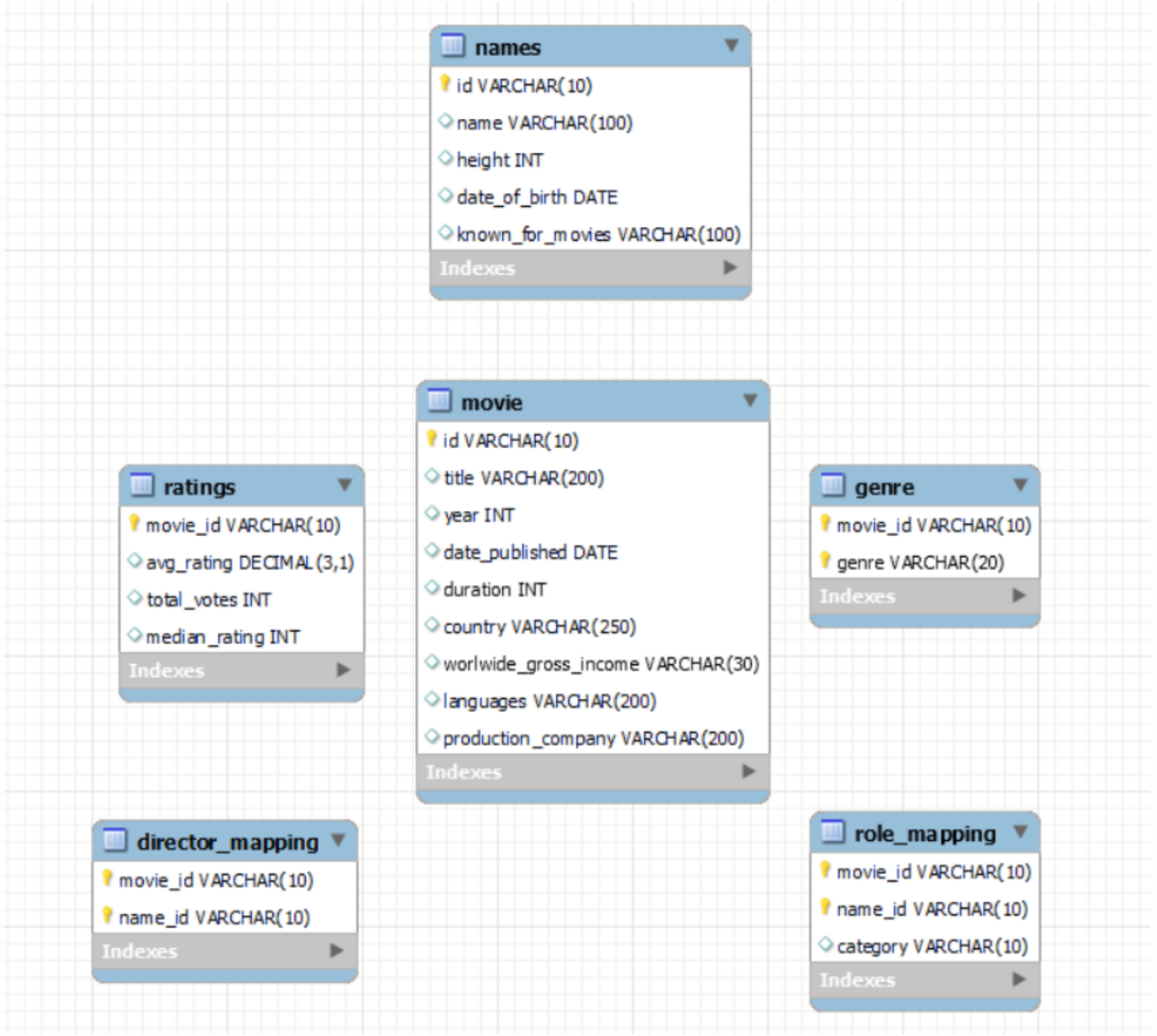
DATASET

- **Director Mapping**: This table maps each movie to its director(s), enabling us to explore patterns and trends among different directors' works.
- **Genre**: This table categorizes all movies into genres, providing a basis for analyzing the popularity and ratings of different genres.
- N**ames**: This table contains the names of actors and directors, establishing a link between the people behind the movies and the movies themselves.
- R**atings**: This table provides the ratings of movies, offering insights into audience reception and preferences.
- **Role mapping**: This table specifies who played the actor and actress in each movie, allowing us to examine the impact of specific actors on a movie's success.

PROJECT GOALS

- **Answer Complex Business Questions:** These could range from understanding the factors that contribute to a movie's success, to assessing the performance of actors/acresses.
- **Uncover Hidden Insights:** With the power of SQL and data analysis, we aim to uncover hidden patterns and trends that aren't immediately obvious, providing a deeper understanding of the movie industry.
- **Storytelling with Data:** This project is not just about data analysis, but also about storytelling with data. It's about uncovering the stories hidden in the numbers, and bringing those stories to life.

Entity Relationship Diagram



Q1. Find the total number of rows in each table of the schema?

Query

```
SELECT table_name, table_rows
FROM INFORMATION_SCHEMA.TABLES
WHERE TABLE_SCHEMA = 'imdb';
```

Result

TABLE_NAME	TABLE_ROWS
director_mapping	3867
genre	14662
movie	6946
names	24531
ratings	7927
role_mapping	14837

Remarks

Names table has the highest number of rows and director mapping has the lowest number of rows.

Q2. Which month had the highest number of movies released?

Query

```
SELECT
MONTHNAME(date_published)
month_name,COUNT(*)
number_of_movies
FROM movie
GROUP BY 1
ORDER BY 2 DESC;
```

Result

month_name	number_of_movies
March	824
September	809
January	804
October	801
April	680
August	678
February	640

Remarks

The highest number of movies were produced in the month of March.

Q3. How many movies were produced in the USA or India in the year 2019?

Query

```
SELECT COUNT(*) AS num_movies
FROM movie
WHERE (country LIKE "%USA%" OR
country LIKE "%INDIA%") AND year
= 2019;
```

Result

num_movies
1059

Remarks

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

Q4. Which genre had the highest number of movies produced overall?

Query

```
SELECT genre,COUNT(*)
num_movies
FROM movie m
LEFT JOIN genre g ON m.id =
g.movie_id
GROUP BY 1
ORDER BY 2 DESC;
```

Result

genre	num_movies
Drama	4285
Comedy	2412
Thriller	1484
Action	1289
Horror	1208

Remarks

‘Drama’ genre had the highest number of movies produced overall

Q5. How many movies belong to only one genre?

Query

```
SELECT COUNT(id)
num_movies_with_one_genre
FROM movie
WHERE id in
(SELECT movie_id FROM genre
GROUP BY movie_id HAVING
COUNT(*) = 1);
```

Result

num_movies_with_one_genre
3289

Remarks

There are more than three thousand movies which have only one genre associated with them.

Q6.What is the average duration of movies in each genre?

Query

```
SELECT genre,AVG(duration)
avg_duration
FROM movie m
JOIN genre g ON m.id = g.movie_id
GROUP BY 1
ORDER BY 2 DESC;
```

Result

genre	avg_duration
Action	112.8829
Romance	109.5342
Crime	107.0517
Drama	106.7746
Fantasy	105.1404

Remarks

Movies of genre 'Drama' (produced highest in number in 2019) have an average duration of 106.77 mins.

Q7.What is the rank of the ‘thriller’ genre in terms of number of movies produced?

Query

```
SELECT *,RANK() OVER(ORDER BY
num_movies DESC) genre_rank
FROM
 (SELECT genre,COUNT(*)
num_movies
 FROM movie m
 JOIN genre g ON m.id = g.movie_id
 GROUP BY 1) t1;
```

Result

genre	num_movies	genre_rank
Drama	4285	1
Comedy	2412	2
Thriller	1484	3
Action	1289	4
Horror	1208	5

Remarks

Thriller in top 3 among all genres in terms of number of movies

Q8. Find the min and max values in each column of the ratings table except the movie_id column.

Query

```
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;
```

Result

MIN_AVG_RATING	MAX_AVG_RATING	MIN_TOTAL_VOTES	MAX_TOTAL_VOTES	MIN_MEDIAN_RATING	MAX_MEDIAN_RATING
1.0	10.0	100	725138	1	10

Remarks

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

Q9. Which are the top 10 movies based on average rating?

Query

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

Result

title	avg_rating	movie_rank
Kirket	10.0	1
Love in Kilnerry	10.0	1
Gini Helida Kathe	9.8	3
Runam	9.7	4
Fan	9.6	5

Remarks

Kriket and Love in Kilnerry have the highest average rating

Q10. Summarise the ratings table based on the movie counts by median ratings

Query

```
SELECT median_rating,COUNT(*)
movie_count
FROM ratings
GROUP BY 1
ORDER BY 2 DESC;
```

Result

median_rating	movie_count
7	2257
6	1975
8	1030
5	985
4	479

Remarks

Movies with a median rating of 7 are highest in number.

Q11. Which production house has produced the most number of hit movies (average rating > 8)?

Query

```
WITH cte AS(
SELECT production_company,COUNT(*)
movie_count
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE avg_rating>8 AND
production_company IS NOT NULL
GROUP BY 1
ORDER BY 2 DESC)
SELECT *,RANK() OVER(ORDER BY
movie_count DESC) prod_company_rank
FROM cte;
```

Result

production_company	movie_count	prod_company_rank
Dream Warrior Pictures	3	1
National Theatre Live	3	1
Lietuvos Kinostudija	2	3
Swadharm Entertainment	2	3
Panorama Studios	2	3

Remarks

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

Q12. How many movies in each genre during March 2017 in the USA had more than 1,000 votes?

Query

```
SELECT genre,COUNT(*) movie_count
FROM movie m
JOIN ratings r ON m.id = r.movie_id
JOIN genre g ON m.id = g.movie_id
WHERE MONTH(m.date_published) = 3
AND year = 2017 AND country like
"%USA%" AND total_votes>1000
GROUP BY 1
ORDER BY 2 DESC;
```

Result

genre	movie_count
Drama	24
Comedy	9
Action	8
Thriller	8
Sci-Fi	7

Remarks

Drama had the most number of such movies.

Q13. Find movies of each genre that start with the word ‘The’ and which have an average rating > 8?

Query

```
SELECT title,avg_rating,genre
FROM movie m
JOIN genre g ON m.id = g.movie_id
JOIN ratings r ON m.id = r.movie_id
WHERE avg_rating>8 AND title REGEXP
'^The'
ORDER BY 2 DESC;
```

Result

title	avg_rating	genre
The Brighton Miracle	9.5	Drama
The Colour of Darkness	9.1	Drama
The Blue Elephant 2	8.8	Drama

Remarks

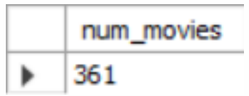
3 movies satisfy the condition.

Q14. Of the movies released between 1/04/2018 and 1/04/2019, how many had a median rating of 8?

Query

```
SELECT COUNT(*)
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE date_published BETWEEN
'2018,04,01' AND '2019,04,01' AND
median_rating = 8;
```

Result



	num_movies
▶	361

Remarks

361 movies satisfy the conditions.

Q15. Do German movies get more votes than Italian movies?

Query

```
SELECT 'Germany' AS
country,SUM(total_votes) total_votes
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE m.country like "%Germany%"
UNION ALL
SELECT 'Italy',SUM(total_votes)
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE m.country like "%Italy%";
```

Result

country	total_votes
German movies	2026223
Italian movies	703024

Remarks

Yes, German movies get more votesthan Italian movies.

Q16. Which columns in the names table have null values?

Query

```
SELECT
 (SELECT COUNT(*) FROM names WHERE
name IS NULL) name_nulls,
 (SELECT COUNT(*) FROM names WHERE
height IS NULL) height_nulls,
 (SELECT COUNT(*) FROM names WHERE
date_of_birth IS NULL)
date_of_birth_nulls,
 (SELECT COUNT(*) FROM names WHERE
known_for_movies IS NULL)
known_for_movies_nulls;
```

Result

name_nulls	height_nulls	date_of_birth_nulls	known_for_movies_nulls
0	17335	13431	15226

Remarks

Height, date_of_birth, known_for_movies columns contain NULLS

There are no Null value in the column 'name'.

Q17. Who are the top three directors in the top three genres whose movies have an average rating > 8?

Query

```
WITH top_3_genres AS(SELECT genre
FROM movie m
JOIN ratings r ON m.id = r.movie_id
JOIN genre g ON m.id = g.movie_id
WHERE avg_rating>8
GROUP BY 1
ORDER BY COUNT(*) DESC
LIMIT 3)

SELECT n.name top_3_directors,COUNT(m.id) num_movies
FROM movie m
JOIN genre g ON m.id = g.movie_id
JOIN director_mapping d ON m.id = d.movie_id
JOIN ratings r ON m.id = r.movie_id
JOIN names n ON d.name_id = n.id
WHERE genre IN (SELECT * FROM top_3_genres) AND
avg_rating>8
GROUP BY 1
ORDER BY 2 DESC
LIMIT 3;
```

Result

	top_3_directors	num_movies
▶	James Mangold	4
	Joe Russo	3
	Anthony Russo	3

Remarks

James Mangold , Joe Russo and Anthony Russo are top three directors in the top three genres whose movies have an average rating > 8

Q18. Who are the top two actors whose movies have a median rating >= 8?

Query

```
SELECT n.name,COUNT(*) num_movies
FROM movie m
JOIN ratings r ON m.id = r.movie_id
JOIN role_mapping rm ON  m.id = rm.movie_id
JOIN names n ON rm.name_id = n.id
WHERE median_rating>=8
GROUP BY 1
ORDER BY 2 DESC
LIMIT 2 ;
```

Result

name	num_movies
Mammootty	8
Mohanlal	5

Remarks

Top 2 actors are Mammootty and Mohanlal.

Q19. Which are the top three production houses based on the no.of votes received by their movies?

Query

```
SELECT
production_company,SUM(total_votes)
sum_total_votes
FROM movie m
JOIN ratings r ON m.id = r.movie_id
GROUP BY 1
ORDER BY 2 DESC
LIMIT 3;
```

Result

production_company	sum_total_votes
Marvel Studios	2656967
Twentieth Century Fox	2411163
Warner Bros.	2396057

Remarks

Top three production houses based on the number of votes received by their movies are Marvel Studios, Twentieth Century Fox and Warner Bros.

Q20. Rank actors with movies released in India based on their avg. ratings. Which actor is at the top?

Query

```
WITH indian_actors_ratings AS
 (SELECT n.name,Round(Sum(avg_rating *
total_votes) / Sum(total_votes), 2) AS
 actor_avg_rating
 FROM movie m
 JOIN ratings r ON m.id = r.movie_id
 JOIN role_mapping rm ON m.id = rm.movie_id
 JOIN names n ON rm.name_id = n.id
 WHERE m.country like "%India%" AND
rm.category = "actor"
 GROUP BY 1
 HAVING COUNT(m.id)>=5)

SELECT *,RANK() OVER(ORDER BY
actor_avg_rating DESC) actor_rank
FROM indian_actors_ratings;
```

Result

name	actor_avg_rating	actor_rank
Vijay Sethupathi	8.42	1
Fahadh Faasil	7.99	2
Yogi Babu	7.83	3
Joju George	7.58	4
Ammy Virk	7.55	5

Remarks

Top actor is Vijay Sethupathi followed by Fahadh Faasil and Yogi Babu.

Q21. Find out the top five actresses in Hindi movies released in India based on their average ratings?

Query

```
WITH indian_actresses_ratings AS
 (SELECT n.name,Round(Sum(avg_rating *
total_votes) / Sum(total_votes), 2) AS
actress_avg_rating
 FROM movie m
 JOIN ratings r ON m.id = r.movie_id
 JOIN role_mapping rm ON m.id = rm.movie_id
 JOIN names n ON rm.name_id = n.id
 WHERE m.country like "%India%" AND
languages like "%hindi%" AND rm.category =
"actress"
 GROUP BY 1
 HAVING COUNT(m.id)>=3)

SELECT *,RANK() OVER(ORDER BY
actress_avg_rating DESC) actress_rank
FROM indian_actresses_ratings;
```

Result

name	actress_avg_rating	actress_rank
Taapsee Pannu	7.74	1
Kriti Sanon	7.05	2
Divya Dutta	6.88	3
Shraddha Kapoor	6.63	4
Kriti Kharbanda	4.80	5

Remarks

Top five actresses in Hindi movies released in India based on their average ratings are Taapsee Pannu, Kriti Sanon, Divya Dutta, Shraddha Kapoor, Kriti Kharbanda
Taapsee Pannu tops with average rating 7.74.

/* Q22. 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

Q22. Select thriller movies as per avg rating and classify them in the following category:

Query

```
SELECT title,avg_rating,CASE WHEN
avg_rating > 8 THEN "Superhit movie"
 WHEN avg_rating between 7 and 8 THEN "Hit
movie"
 WHEN avg_rating between 5 and 7 THEN
"One-time-watch movie"
 WHEN avg_rating < 5 THEN "Flop movie"
 END AS avg_rating_category
FROM movie m
JOIN genre g ON m.id = g.movie_id
JOIN ratings r ON m.id = r.movie_id
WHERE genre = "thriller";
```

Result

title	avg_rating	avg_rating_category
Der müde Tod	7.7	Hit movie
Fahrenheit 451	4.9	Flop movie
Pet Sematary	5.8	One-time-watch movie
Dukun	6.9	One-time-watch movie
Back Roads	7.0	Hit movie

Remarks

Q23. What is the genre-wise running total and moving avg. of 10 rows based on avg. movie duration?

Query

```
SELECT
 genre,AVG(duration) avg_duration,
 SUM(AVG(duration)) OVER (ORDER BY genre
ROWS BETWEEN UNBOUNDED PRECEDING
AND CURRENT ROW) running_total,
 AVG(AVG(duration)) OVER (ORDER BY genre
ROWS BETWEEN 10 PRECEDING AND
CURRENT ROW) moving_average
FROM movie m
JOIN genre g ON m.id = g.movie_id
GROUP BY 1
ORDER BY genre;
```

Result

genre	avg_duration	running_total	moving_average
Action	112.8829	112.8829	112.88290000
Adventure	101.8714	214.7543	107.37715000
Comedy	102.6227	317.3770	105.79233333
Crime	107.0517	424.4287	106.10717500
Drama	106.7746	531.2033	106.24066000

Remarks

Q24. Which are the five highest-grossing movies of each year that belong to the top three genres?

Query

```
CREATE TEMPORARY TABLE new_movie AS(
 SELECT
 *,CASE WHEN LEFT(worlwide_gross_income,1) = "$" THEN
SUBSTRING_INDEX(worlwide_gross_income," ",-1)
 WHEN LEFT(worlwide_gross_income,1) = "I" THEN
SUBSTRING_INDEX(worlwide_gross_income," ",-1)/80 END AS wg_income
 FROM movie);
ALTER TABLE new_movie
MODIFY COLUMN wg_income BIGINT;
WITH top_3_genres AS
 (SELECT genre
 FROM movie m
 JOIN genre g ON m.id = g.movie_id
 GROUP BY 1
 ORDER BY COUNT(*) DESC
 LIMIT 3),
ranked_by_year AS
 (SELECT *, RANK() OVER(PARTITION BY genre,year ORDER BY wg_income DESC)
year_rank
 FROM new_movie m
 JOIN genre g ON m.id = g.movie_id
 WHERE genre IN(SELECT * FROM top_3_genres))
SELECT genre,year,title AS movie_name,wg_income,year_rank AS movie_rank
FROM ranked_by_year
WHERE year_rank<=5
ORDER BY 1;
```

Result

genre	year	movie_name	wg_income	movie_rank
Comedy	2017	Despicable Me 3	1034799409	1
Comedy	2017	Jumanji: Welcome to the Jungle	962102237	2
Comedy	2017	Guardians of the Galaxy Vol. 2	863756051	3
Comedy	2017	Thor: Ragnarok	853977126	4
Comedy	2017	Sing	634151679	5

Remarks

-- Q25. Find the top 2 production houses that have the highest no. of hits among multilingual movies.

Note: Highest number of hits means median rating>=8

Query

```
WITH producers_movie_count AS
(SELECT production_company,COUNT(m.id)
movie_count
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE median_rating>=8 AND
production_company IS NOT NULL
AND LOCATE(",",languages) >0
GROUP BY 1)
SELECT *,RANK() OVER(ORDER BY
movie_count DESC) prod_comp_rank
FROM producers_movie_count;
```

Result

production_company	movie_count	prod_comp_rank
Star Cinema	7	1
Twentieth Century Fox	4	2
Columbia Pictures	3	3
Ave Fenix Pictures	3	3
Viva Films	3	3

Remarks

Star Cinema and Twentieth Century Fox are the top two production houses that have produced the highest number of hits among multilingual movies.

Q26. Who are the top 3 actresses based on no. of Super Hit movies (avg rating >8) in drama?

Query

```
WITH actress_summary AS
(
        SELECT     n.NAME AS actress_name,
                SUM(total_votes) AS total_votes,
                Count(r.movie_id)                    AS
movie_count,

Round(Sum(avg_rating*total_votes)/Sum(total_vote
s),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 g.movie_id = m.id
        WHERE      category = 'ACTRESS'
```

Result

actress_name	total_votes	movie_count	actress_avg_rating	actress_rank
Parvathy Thiruvothu	4974	2	8.25	1
Susan Brown	656	2	8.94	1
Amanda Lawrence	656	2	8.94	1

Remarks

Top 3 actresses based on number of Super Hit movies are Parvathy Thiruvothu, Susan Brown and Amanda Lawrence

Q27. 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

Q27. Get the following details for top 9 directors (based on number of movies)

Query

```
WITH cte AS
 (SELECT
n.id,n.name,duration,avg_rating,total_votes,date_published,LEAD(date_publish
ed) OVER(PARTITION BY n.id ORDER BY date_published ASC)
next_date_published
    FROM movie m
    JOIN director_mapping dm ON m.id = dm.movie_id
    JOIN names n ON dm.name_id = n.id
    JOIN ratings r ON m.id = r.movie_id
)
SELECT
 id director_id,
    name director_name,
    COUNT(*) number_of_movies,
 ROUND(AVG(DATEDIFF(next_date_published,date_published)),2)
avg_inter_movie_days,
 round(AVG(avg_rating),2) avg_rating,
       SUM(total_votes) total_votes,
    MIN(avg_rating) min_rating,
    MAX(avg_rating) max_rating,
    SUM(duration) total_movie_duration

FROM cte
GROUP BY 1
ORDER BY 3 DESC
LIMIT 9;
```

Result

director_id	director_name	number_of_movies	avg_inter_movie_days	avg_rating	total_votes	min_rating	max_rating	total_movie_duration
nm2096009	Andrew Jones	5	190.75	3.02	1989	2.7	3.2	432
nm1777967	A.L. Vijay	5	176.75	5.42	1754	3.7	6.9	613
nm0814469	Sion Sono	4	331.00	6.03	2972	5.4	6.4	502
nm0831321	Chris Stokes	4	198.33	4.33	3664	4.0	4.6	352
nm0515005	Sam Liu	4	260.33	6.23	28557	5.8	6.7	312

Remarks