DATA ANALYTICS PROJECT 5: MOVIE DATA ANALYSIS WITH MYSQL

NAZIM ATAKAN ERDOGAN

DATA ANALYST INTERN AT OESON

AGENDA

- Introduction and aim of the project
- Part 1: Data analysis and inspection
- Part 2: Exploration of the movies and genres
- Part 3: Exploration of actors, directors, gross income, and ratings.
- Part 4: Analysis and insights,

AIM OF THE PROJECT AND INTRODUCTION

- RSVP Movies is an Indian film production company which has produced many super-hit movies. They have usually released movies for the Indian audience but for their next project, they are planning to release a movie for the global audience in 2025. According to IMDB data they are planning to grow their business and market share.
- The project is segmented into 4 parts, to conceptualize the analysis process better.
- First the insights and strategies will be shown then the data analysis and the related inspection.

ANALYSIS INSIGHTS AND RECOMMENDATIONS; STRATEGIES FOR ENTERING THE FOREIGN MARKET OF THE LOCAL MOVIE PRODUCTION COMPANY

As a local Indian film production company, RSVP wants to enter to new markets. The methods to make data driven decisions should be based on the database given and the insights should be extracted using SQL.

The company should use the data for analyzing the top performing actors, directors, movies, languages, countries and their release dates accordingly.

The analysis should open the door for an insightful report ,that would improve the overall production rates and worldwide gross income.

It will be compared the US and India data also some European countries like Germany and Italy and the movies will be analyzed according to genre, director, top performing actors, total voting, average ratings and production company.

It will be shown the top performing actors, directors with top hit movies and the top performing production companies, which the local company would like to collaborate with.

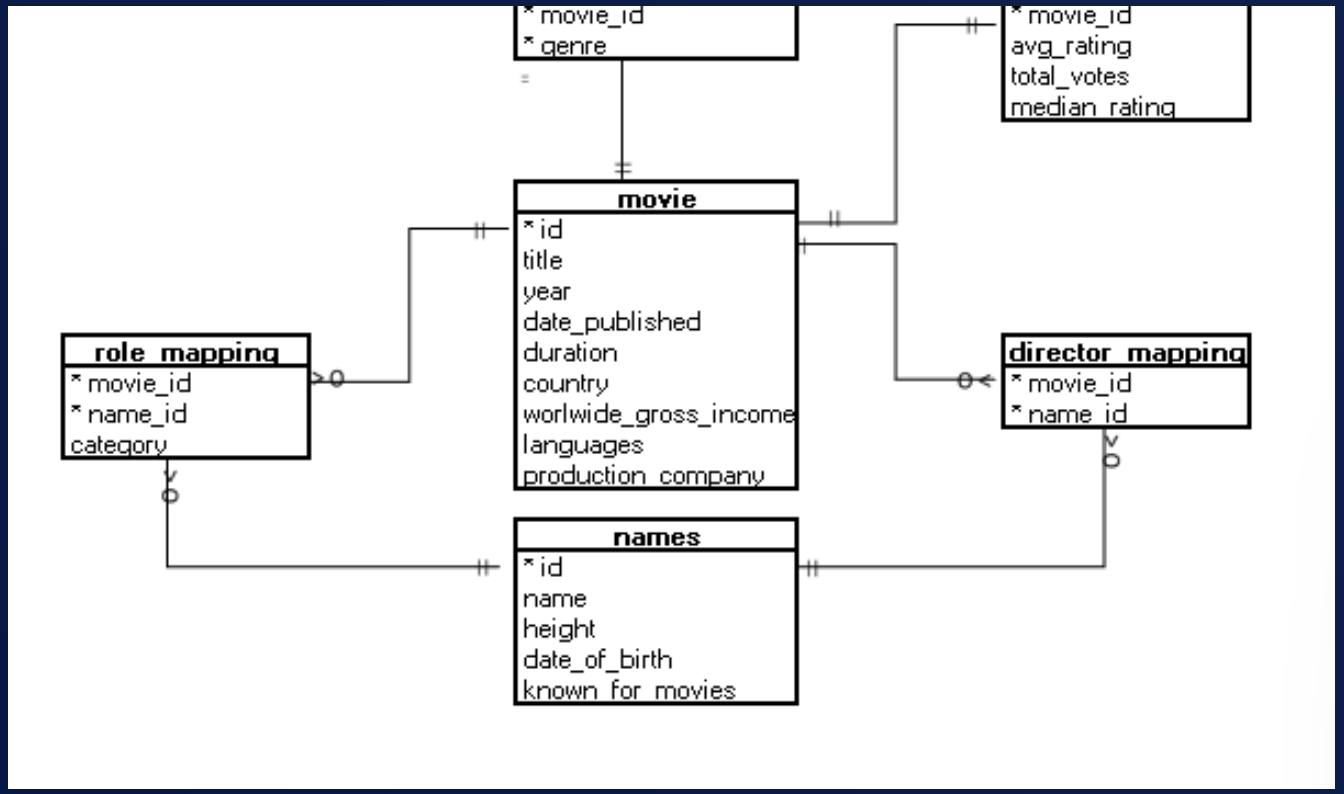
RSVP is a successful company in India which made lots of hit movies and it will be analyzed the key data for entering the foreign market, for example inspecting the gross income of the movies or in which years in what countries were there popular with what range of voting .

OVERVIEW OF THE DATABASE AND ANALYSIS

1. The database used is a model version of IMDB; which involves 6 tables with columns describing actors, directors, production years, release dates, gross incomes, and ids.
2. The entity relationship diagram will be shown
3. The analysis will be performed using MySQL

PART 1 DATA INSPECTION :
ERD – ENTITY RELATIONSHIP DIAGRAM

movie	worlwide_gross_income
movie	languages
movie	production_company
genre	movie_id
genre	genre
director_mappi	movie_id
director_mappi	name_id
role_mapping	movie_id
role_mapping	name_id
role_mapping	category
names	id
names	name
names	height
names	date_of_birth
names	known_for_movies



DATA INSPECTION AND ANALYSIS

genre
Drama
Fantasy
Thriller
Comedy
Horror
Family
Romance

year	country	Total_movies
2017	India	347
2017	USA	853
2018	India	365
2018	USA	815
2019	India	295
2019	USA	592

```
USE imdb;

SELECT 'movie' AS table_name, COUNT(*) AS row_count FROM movie UNION ALL
SELECT 'genres' AS table_name, COUNT(*) AS row_count FROM genres UNION ALL
SELECT 'director_mapping' AS table_name, COUNT(*) AS row_count FROM director_mapping UNION ALL
SELECT 'role_mapping' AS table_name, COUNT(*) AS row_count FROM role_mapping UNION ALL
SELECT 'names' as table_name, count(*) as row_count from  names;

use imdb;
select column_name
from information_schema.columns
where table_name = 'movie'
and table_schema = 'imdb'
and is_nullable = 'yes';
```

Code = total rows per table, the results = unique genres, total movies per year in USA and India, movies per year

table_name	row_count
movie	7997
genres	14662
director_mapping	3867
names	25735
role_mapping	15615

year	movies
2017	3052
2018	2944
2019	2001

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

genre	movie_count
Drama	4285

genre	movie_count	genre_rank
Thriller	1484	3

single_genre
3289

genre	avg_duration
Action	112.8829
Romance	109.5342
Crime	107.0517
Drama	106.7746
Fantasy	105.1404
Comedy	102.6227
Adventure	101.8714

Genre with the highest movies produced = Drama

Thriller is ranked 3rd

Action, Romance, Crime have the most average duration.

```sql
select distinct m.id as movie_id, m.title as movie_title, r.avg_rating
from movie m
join ratings r on m.id = r.movie_id
order by r.avg_rating desc
limit 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

movie_id	movie_title	avg_rating
tt6735740	Love in Kilnerry	10.0
tt10914342	Kirket	10.0
tt9537008	Gini Helida Kathe	9.8
tt10370434	Runam	9.7
tt10867504	Fan	9.6
tt9526826	Android Kunjappan Version 5.25	9.6
tt10869474	Safe	9.5

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

- Top 10 movies based on average rating
- Number of movies grouped by median ratings
- Production company with the highest hit movie rate
- Germany and Italy comparison as an example
- Top studios based on votings for movies
- Top two actors with movies of median rating >8

```sql
select median_rating, count(*) as movie_counts
from ratings
group by median_rating;
```

Result Grid | Filter Rows: | Export: | Wrap C

median_rating	movie_counts
8	1030
7	2257
3	283
6	1975
9	429
2	119
4	479

production_company	hit_movie
Dream Warrior Pictures	3

country	total_votes
Germany	106710
Italy	77965

actor_name	movie_count
Mammootty	8
Mohanlal	5

```sql
with ActorRatings as (
    select
        n.name as actor_name,
        sum(r.total_votes) as total_votes,
        count(r.movie_id) as movie_count,
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

actor_name	total_votes	movie_count	actor_avg_rating	actor_rank
Yogi Babu	8500	11	7.83	1
Ammy Virk	2504	6	7.55	2
Kunchacko Boban	5628	6	7.48	3
Rajkummar Rao	42560	6	7.37	4
Tovino Thomas	11596	8	7.15	5
Mammootty	12613	8	7.04	6
Karamjit Anmol	1970	6	6.91	7

production_company	ranking	movie_count
Star Cinema	1	7
Twentieth Century Fox	2	4

Rank of actors in india,which RSVP would make a deal with.

Top two production companies with multilingual movies, which RSVP would like to cooperate with.

	genre	movie_count
▶	Drama	4285

	single_genre
▶	3289

	genre	movie_count	genre_rank
▶	Thriller	1484	3

```
select year , count(*) as movies from movie group by year order by year;
select year, month(date_published) as month, count(*) as total_movies from movie group by year, month order by year, month;

SELECT year, country,
COUNT(*) AS Total_movies
FROM movie
WHERE country IN ('india', 'USA')
GROUP BY year, country
ORDER BY year, country;

SELECT DISTINCT genre from genres;

select genre, count(*) as movie_count
from genres
group by genre
order by movie_count desc limit 1;
```

year	month	total_movies
2017	1	291
2017	2	228
2017	3	298
2017	4	249
2017	5	205
2017	6	226
2017	7	188

Comparing number of movies produced in India and US as the company wants to open itself to the US market.

Inspection of distinct genres and count of movies per each genre

Results of months and total movies

```sql
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_ratings) AS min_median_rating, MAX(median_ratings) AS max_median_rating
FROM ratings;

select * from ratings order by avg_rating desc limit 10;

select median_rating, count(*) as movie_counts
from ratings
group by median_rating;

select production_company, count(*) as hit_movie from movie
join ratings on movie.id = ratings.movie_id
where avg_rating > 8 and production_company IS NOT NULL group by production_company
order by hit_movie desc limit 1;
```

Minimum and maximum values in each column in rating table, detection of production companies with the most hit movies.

```sql
with GenreCounts as (
    select g.genre, COUNT(*) as movie_count from genre g
    JOIN ratings r ON g.movie_id = r.movie_id
    WHERE r.avg_rating > 8
    GROUP BY g.genre
    ORDER BY movie_count DESC
    LIMIT 3
),
DirectorCounts AS (
    SELECT n.name AS director_name, g.genre, COUNT(*) AS movie_count
    FROM name n
    JOIN movie m ON n.id = m.director_id
    JOIN genre g ON m.id = g.movie_id
    JOIN ratings r ON m.id = r.movie_id
    WHERE r.avg_rating > 8
    GROUP BY director_name, g.genre
)
SELECT dc.director_name, dc.genre, dc.movie_count
FROM DirectorCounts dc
JOIN GenreCounts gc ON dc.genre = gc.genre
ORDER BY dc.movie_count DESC
```

median_rating	movie_counts
8	1030
7	2257
3	283
6	1975
9	429
2	119
4	479

movie_id	avg_rating	total_votes	median_rating
tt6735740	10.0	2360	10
tt10914342	10.0	587	10
tt9537008	9.8	425	10
tt10370434	9.7	133	10
tt10867504	9.6	1010	10
tt9526826	9.6	1176	10
tt10869474	9.5	1017	10

Top three directors in top three genres with movies with avg > 8 *the code median ratings, total votes, avg ratings

```sql
with ActorRatings as (
    select
        n.name as actor_name,
        sum(r.total_votes) as total_votes,
        count(r.movie_id) as movie_count,
        round(sum(r.avg_rating * r.total_votes) / sum(r.total_votes), 2) as actor_avg_rating
    from role_mapping rm
    join names n on rm.name_id = n.id
    join movie m on rm.movie_id = m.id
    join ratings r on m.id = r.movie_id
    where rm.category = 'actor' and m.country = 'India'
    group by n.name having movie_count > 5)

select actor_name, total_votes, movie_count, actor_avg_rating,
rank() over (order by actor_avg_rating desc, total_votes desc) as actor_rank
from ActorRatings;
```

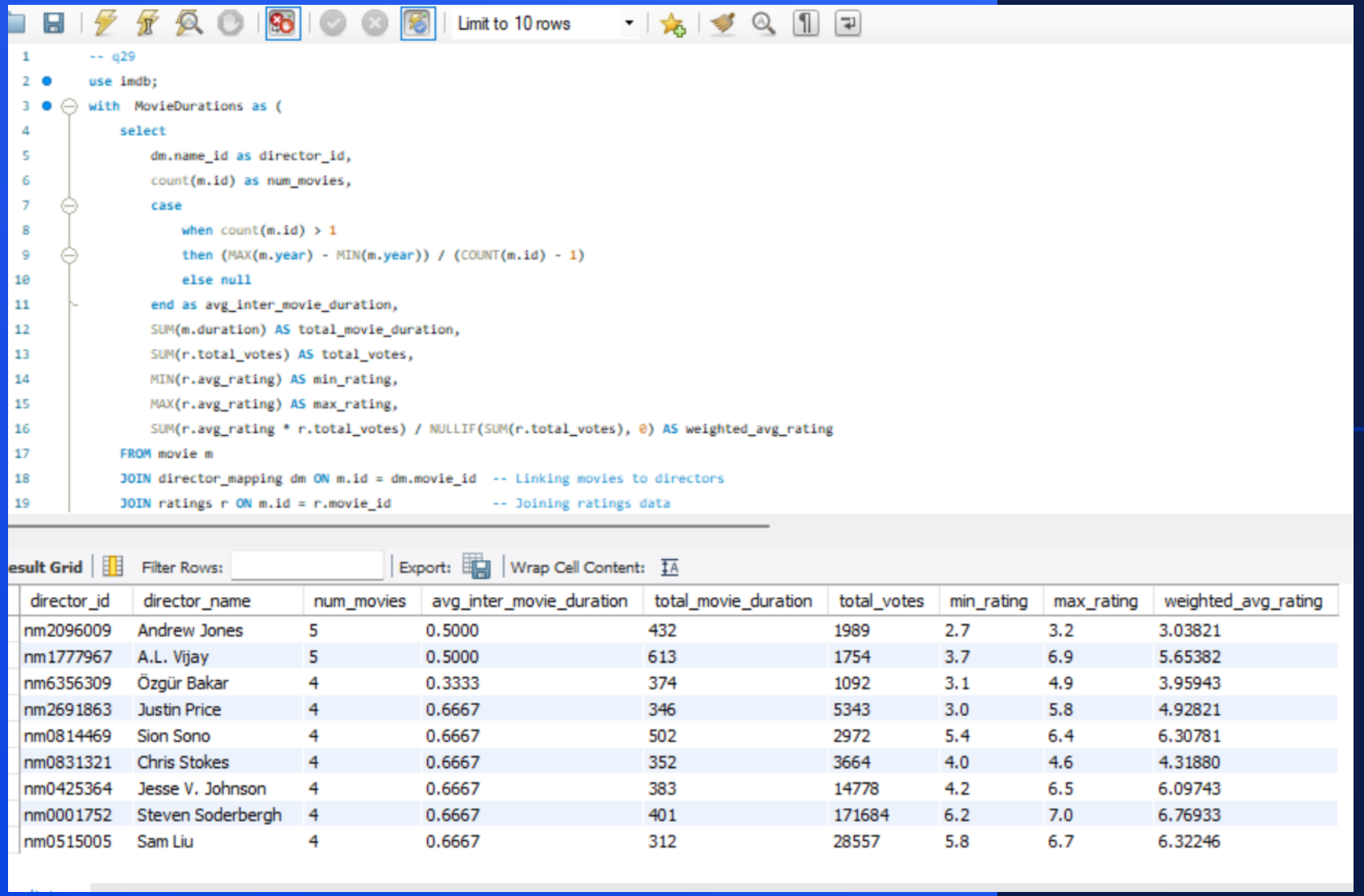
actor_name	total_votes	movie_count	actor_avg_rating	actor_rank
Yogi Babu	8500	11	7.83	1
Ammy Virk	2504	6	7.55	2
Kunchacko Boban	5628	6	7.48	3
Rajkummar Rao	42560	6	7.37	4
Tovino Thomas	11596	8	7.15	5
Mammootty	12613	8	7.04	6
Karamjit Anmol	1970	6	6.91	7

Top actors in India, code of top genres, and top production companies that the local company wants to collaborate with.

```sql
select g.genre,count(m.id) as movie_count
from genres g
join movie m on g.movie_id = m.id
group by g.genre order by movie_count DESC limit 3;
```

actor_name	movie_count
Mammootty	8
Mohanlal	5

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



```sql
-- q29
use imdb;
with  MovieDurations as (
    select
        dm.name_id as director_id,
        count(m.id) as num_movies,
        case
            when count(m.id) > 1
            then (MAX(m.year) - MIN(m.year)) / (COUNT(m.id) - 1)
            else null
        end as avg_inter_movie_duration,
        SUM(m.duration) AS total_movie_duration,
        SUM(r.total_votes) AS total_votes,
        MIN(r.avg_rating) AS min_rating,
        MAX(r.avg_rating) AS max_rating,
        SUM(r.avg_rating * r.total_votes) / NULLIF(SUM(r.total_votes), 0) AS weighted_avg_rating
    FROM movie m
    JOIN director_mapping dm ON m.id = dm.movie_id   -- Linking movies to directors
    JOIN ratings r ON m.id = r.movie_id              -- Joining ratings data
```

Limit to 10 rows

esult Grid | Filter Rows: | Export: | Wrap Cell Content:

director_id	director_name	num_movies	avg_inter_movie_duration	total_movie_duration	total_votes	min_rating	max_rating	weighted_avg_rating
nm2096009	Andrew Jones	5	0.5000	432	1989	2.7	3.2	3.03821
nm1777967	A.L. Vijay	5	0.5000	613	1754	3.7	6.9	5.65382
nm6356309	Özgür Bakar	4	0.3333	374	1092	3.1	4.9	3.95943
nm2691863	Justin Price	4	0.6667	346	5343	3.0	5.8	4.92821
nm0814469	Sion Sono	4	0.6667	502	2972	5.4	6.4	6.30781
nm0831321	Chris Stokes	4	0.6667	352	3664	4.0	4.6	4.31880
nm0425364	Jesse V. Johnson	4	0.6667	383	14778	4.2	6.5	6.09743
nm0001752	Steven Soderbergh	4	0.6667	401	171684	6.2	7.0	6.76933
nm0515005	Sam Liu	4	0.6667	312	28557	5.8	6.7	6.32246

Top 9 directors details.

INSIGHTS AND CONCLUSION

- According to the dataset given,
- The company shall use the top performing actors and cooperate wit the major production companies which produced multilingual movies before, which would be an advantage for the deal.. An analysis based on numeric data have been performed and it has been shown the actors, movies and genres with the highes performance which would show the way to the company in what way they should approach to their next movies. For example they can start with a movie in Hindi language which is the mostly spoken in India, cooperating with an international company which also produces multilingual movies.

THANK YOU

Nazim Atakan Erdogan

Data Analyst Intern @ Oeson

certified in data based marketing and marketing science and strategies.