IMDB Movie Analytics: End-to-End Data Analysis Using SQL and Power BI

---

1. Project Overview

This project focuses on analyzing data from the **IMDB Movies Dataset** using **SQL** for data extraction, transformation and analysis, and **Power BI** for interactive visualization. The goal of the project is to extract meaningful insights about movie performance, director information, **revenues**, **popularity trends**, and **industry patterns** using structured queries and rich visual dashboards.

---

2. Dataset Description

2.1 Directors Table

This table contains information about movie directors.

Column Name	Description
Name	Name of the Director
Id	Unique ID of the Director
uid	Unique ID for Movies
Gender	Gender of the Director → *0/2 = Male*, *1 = Female*
Department	Department of the Director

Total rows in Directors table: **2349 directors**.

2.2 Movies Table

This table contains information about movies.

Column Name	Description
ID	Unique ID for Movies
Original Title	Movie name
Budget	Budget of the movie
Popularity	Popularity score
Release Date	Release date

Column Name	Description
Revenue	Revenue collected
Title	Initial title of the movie
Vote Average	Average IMDB rating
Vote Count	Number of votes the movie got
Overview	Movie description
Tagline	Tagline of the movie
UID	Unique ID for movies
Director ID	ID of the director of the movie

Total rows in Movies table: **47 movies**.

---

3. Technology Stack

- **Database**: MySQL Server
- **BI & Analytics:** Power BI Desktop
- **Languages:** SQL (for data processing), DAX (for analysis & measures)
- **Supporting Tools:** MySQL Workbench (database management), MS Excel (initial data review)

---

4. Data Preparation

4.1 Database Creation & Table Setup

- Created a database: IMDB Movies
- Imported Directors.csv & Movies.csv using LOAD DATA LOCAL INFILE
- Verified schema using DESCRIBE and ensured successful loading of all records.

4.2 Data Cleaning & Transformation

- Checked for null values and invalid entries across both tables.
- Ensured integrity of director–movie relationships using director id
- Verified row counts, unique director names, gender codes, and the release year range to confirm data completeness.

4.3 Data Integration

- Created a combined table **movies directors** using an SQL join between **movies** and **directors** on director id.
- This table served as the primary source for advanced SQL analysis and Power BI dashboards.

5. Exploratory Data Analysis (EDA)

A preliminary exploration was conducted to understand data structure and quality:

- Reviewed the structure and row counts of both the **Movies** and **Directors** tables.
- Checked data types and confirmed that all fields were correctly loaded.
- Identified distinct values for key fields such as **director names**, **gender codes**, and **departments**.
- Performed null checks and found only one missing value in the **tagline** column, which was later handled in Power BI.
- Verified the **release year range** to understand the time span of the dataset.
- Observed general distributions of **popularity**, **vote averages**, **votes**, and **revenues**.
- Ensured the correctness of movie–director mapping using the **director id** field.
- Confirmed that the dataset was clean, complete, and ready for SQL-based analysis.

6. SQL Analysis Performed

The following analytical questions were solved using SQL queries on the dataset:

a) Can you get all data about movies?

```
SELECT *
FROM movies_directors;
```

b) How do you get all data about directors?

```
SELECT *
FROM movies_directors;
```

c) Check how many movies are present in IMDB.

```
SELECT count (*) as No_of_movies
FROM movies;
```

d) Find these 3 directors: James Cameron; Luc Besson; John Woo

```
SELECT *
FROM directors
WHERE name IN ('James Cameron', 'Luc Besson', 'John Woo');
```

e) Find all directors with name starting with S.

```
select name
from directors
where name like "s%";
```

f) Count female directors.

```
select count(*) AS No_of_female_directors
from directors
where gender = 1;
```

g) Find the name of the 10th first women directors?

```
SELECT *
FROM directors
WHERE gender = 1
ORDER BY id
limit 1 OFFSET 9;
```

h) What are the 3 most popular movies?

```
select *
from movies
order by popularity
desc limit 3;
```

i) What are the 3 most bankable movies?

```
select *
from movies
order by revenue desc
limit 3;
```

j) What is the most awarded average vote since the January 1st, 2000?

```
SELECT title, vote_average, release_date, vote_count
FROM movies
WHERE release_date >= '2000-01-01'
ORDER BY vote_average DESC, vote_count desc, release_date
limit 5;
```

k) Which movie(s) were directed by Brenda Chapman?

```
select *
from movies_directors
where director_name ="Brenda Chapman";
```

l) Which director made the most movies?

```
SELECT director_name, COUNT(*) AS No_of_Movies
FROM movies_directors
GROUP BY director_name
ORDER BY No_of_Movies DESC
limit 5;
```

m) Which director is the most bankable?

```
select director_name, sum(revenue) as total_revenue
FROM movies_directors
group by director_name
order by total_revenue desc
limit 5;
```

These queries helped understand the relationships and performance metrics across the dataset.

---

7. Power BI Analysis & Visualizations

After SQL analysis, the dataset was imported into **Power BI Desktop** for further transformation, modeling, and interactive visualization.

7.1 Power Query Transformations

- Verified and corrected data types for numeric, text, and date fields.
- Created a **gender label column** (Male / Female / Unknown) based on the gender codes.
- Checked for data quality issues and handled **one null value in the *tagline* column**.
- Ensured all fields were error-free before loading the data into the report view.

7.2 Data Modeling

- **Directors Table** was linked with **Movies Table** using the **ID** column (one-to-many relationship).

7.3 Measures Created in Power BI

Some custom DAX measures created include:

Profit

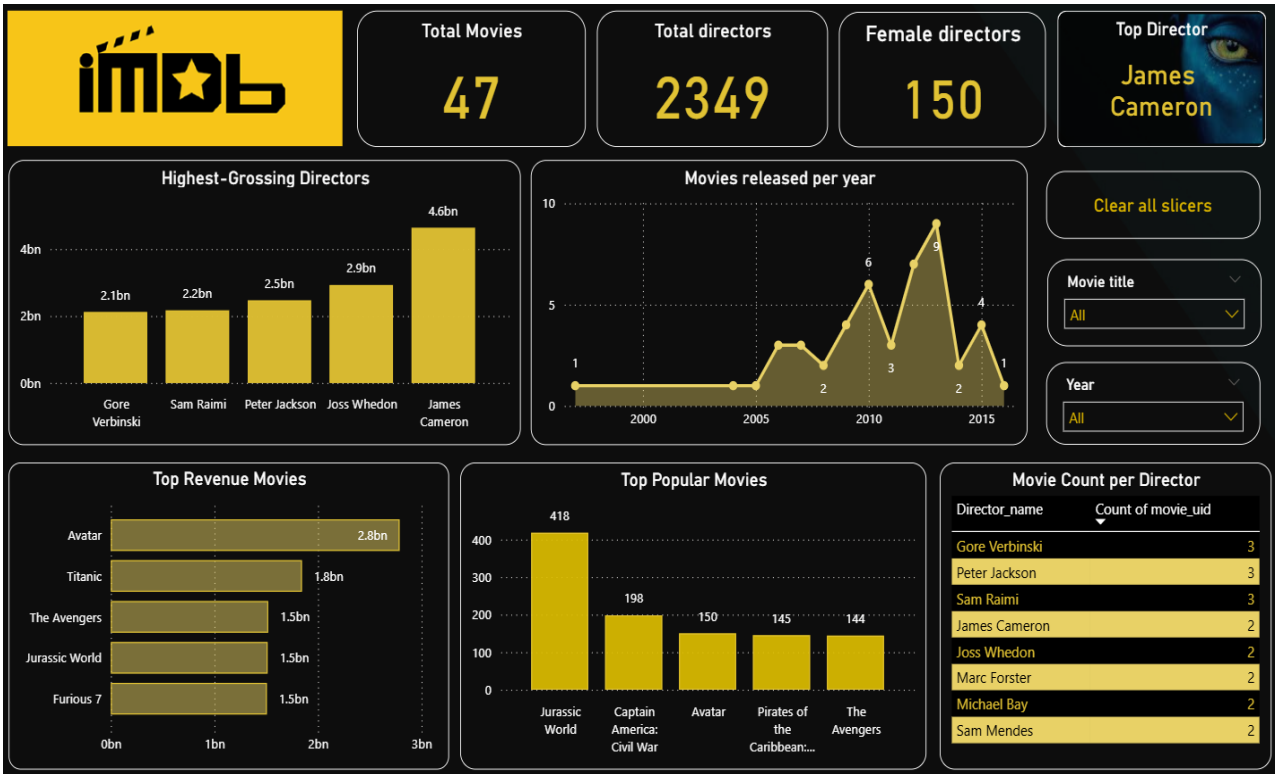
```
Profit = SUM(Movies[Revenue]) - SUM(Movies[Budget])
```

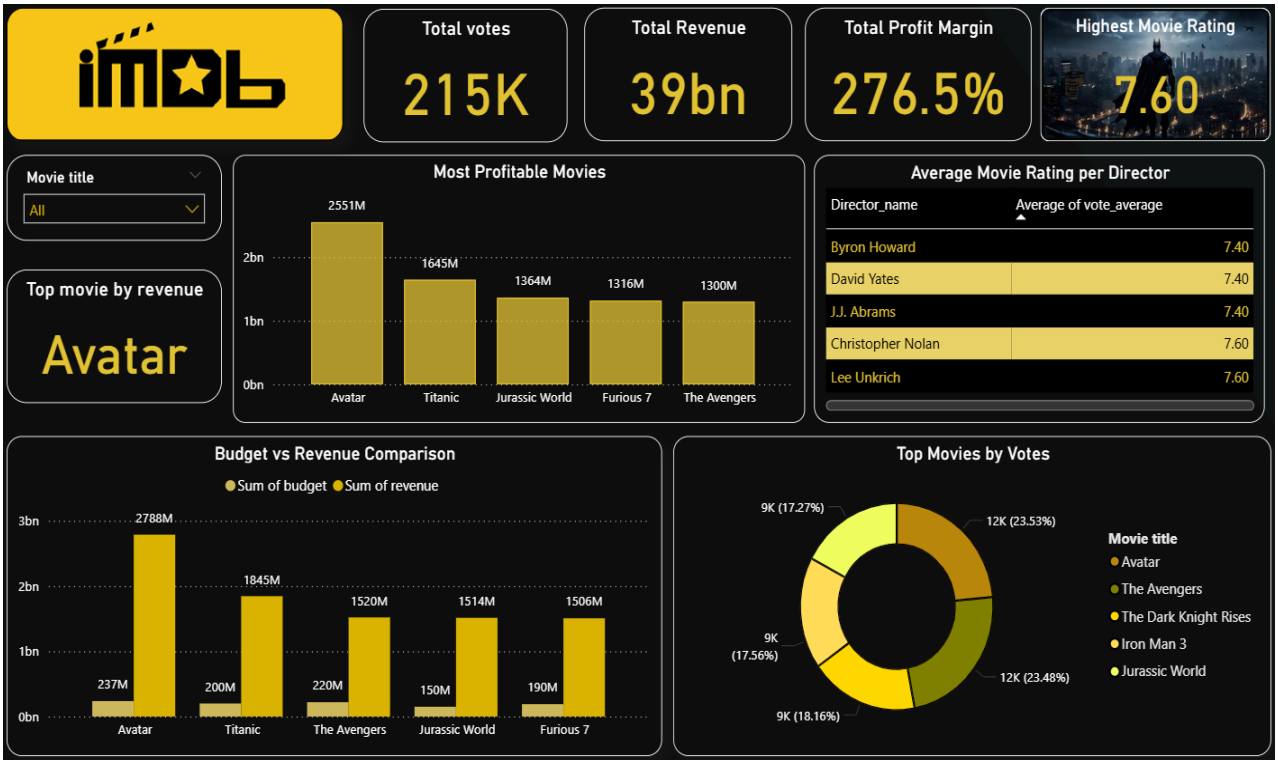
Profit Percentage

```
Profit % =
DIVIDE(
    SUM(Movies[Revenue]) - SUM(Movies[Budget]),
    SUM(Movies[Budget])
)
```

These measures helped analyze business performance of each movie.

7.4. Power BI Dashboard Overview





7.5. Power BI Visuals Used

Below are the visuals used in the IMDB project:

- **Table Visuals:** No of Movies directed by a specific director, Average movie rating per director
- **Cards**: Total Movies, Total Female Directors, Total directors, Total votes, Total revenue, Total profit margin
- **Bar/Column Charts**: Highest grossing directors, Top Revenue Movies, Top popular Movies, Most Profitable movies, Budget vs Revenue Comparison
- **Line Chart**: Movies released per year
- **Donut Chart**: Top Movies by Votes
- **Filters / Slicers:** Movie title, Year, Clear all slicers

---

8. Project Insights

Some key insights derived include:

8.1 🎬 Movie Performance & Audience Behavior

- Popularity and revenue show a strong positive correlation—movies with higher popularity scores tend to generate significantly higher revenue, indicating the commercial impact of audience interest.
- Top-grossing films contribute a major share of total revenue, highlighting a blockbuster-driven industry where a small percentage of movies dominate box office earnings.
- Vote averages remained stable across most movies, suggesting generally consistent audience ratings.
- Movies released after 2000 show stronger audience engagement, with many films achieving high ratings and large vote counts.
- There is a clear correlation between vote count and vote average, indicating that movies with wider audience reach often secure better ratings.

8.2 🎥 Director-Level Insights

- **James Cameron** emerged as the most bankable director with extremely high revenue figures.
- A select group of directors generated the highest cumulative revenues, demonstrating their strong influence on commercial success.

- The number of **female directors is significantly lower**, indicating a skewed gender representation.
- Director productivity varied widely—some delivered multiple films while others contributed only a single title.
- Several directors with fewer movies still achieved high average ratings, proving that output quantity does not always reflect audience approval.

8.3 💰 Financial & Profitability Insights

- Profit analysis revealed that high budgets do not always result in high profitability; some lower-budget films delivered strong profit margins.
- Directors with consistent revenue-generating movies delivered high profitability across their filmography, demonstrating stable box office performance.

8.4 📈 Trend & Distribution Patterns

- Revenue, popularity, and vote metrics displayed wide variability, indicating that only a limited number of movies achieve exceptional commercial and audience success.
- Year-wise trends showed a significant concentration of movie releases and audience activity in the 2010–2016 period, marking a high-growth era.

---

9. Recommendations

Based on the insights from SQL and Power BI, the following recommendations can guide better decisions for the movie industry:

- **Prioritize films with rising popularity**, as popularity strongly correlates with revenue and can be used to predict commercial success.
- **Use vote count and vote average together** to identify movies and directors with strong audience engagement and consistent positive feedback.
- **Support directors with a track record of high revenue or high ratings**, as they significantly influence overall performance.
- **Optimize production budgets**, since higher spending does not necessarily lead to higher profitability.
- **Promote gender diversity among directors**, as the dataset shows a clear underrepresentation of female directors.
- **Continue monitoring year-wise performance trends**, especially during high-activity periods (2010–2017), to refine release strategies.

---

10. Conclusion

This project demonstrated a complete end-to-end analytics workflow by combining SQL for data preparation and querying with Power BI for interactive visualization. The IMDB dataset enabled practical application of core concepts such as data validation, table relationships, SQL joins, DAX measures, and dashboard design. The analysis revealed meaningful insights into movie performance, audience engagement, director productivity, and revenue trends. Overall, the project highlights how structured data processing and business intelligence tools work together to convert raw data into clear, actionable insights that support informed decision-making.

Prepared By:

S. Asmita

T. Sumanjali

Project Team ID: PTID-CDA-OCT-25-800

Project ID: PRSQL-01