PGDM- Big Data Analytics

Big Data Management & Analytics

Project Report

MongoDB Dashboard & Queries – sample_mflix Database

Submitted to: Prof. Amarnath Mitra

Submitted By: Group 6

Krishnendu Adhikary – 055022

Mohit Agarwal – 055024

Introduction

The movie industry has witnessed **exponential growth** over the years, with thousands of films being produced annually across the globe. With the increasing availability of digital platforms and streaming services, the need for **data-driven decision-making** has become more significant than ever. **Filmmakers, producers, and analysts** rely on data insights to understand audience preferences, track industry trends, and optimize content strategies.

This project aims to analyze various aspects of the **film industry** using the **MFlix Movies dataset**, a structured database containing information about movies, including **titles, genres, directors, release years, IMDb ratings, and audience engagement metrics**. The dataset is stored in **MongoDB Atlas**, and the queries are executed using **MongoDB Compass and mongosh** to extract, modify, and interpret valuable insights.

Objective

The primary objective of this project is to analyze **trends and patterns** in the movie industry using the **MFlix Movies dataset** within **MongoDB Atlas**. The project leverages **MongoDB Atlas Charts** to generate insights from key movie attributes, including:

- **Release Trends** – Understanding how movie releases vary across different years.
- **IMDb Ratings** – Analyzing audience feedback and rating distributions.
- **Directors' Influence** – Examining how directors impact movie success.
- **Geographical Distribution** – Identifying movies based on country or region.
- **Language Diversity** – Evaluating multilingual film productions and audience reach.

To gain a deeper understanding of the movie industry, this project performs **CRUD (Create, Read, Update, and Delete) operations** on the MFlix dataset, addressing key areas such as:

1. **Movie Catalog Expansion:** Adding new movie records to enhance the dataset and maintain an updated collection.
2. **Movie Search & Categorization:** Filtering and retrieving movies based on different criteria such as **title, genre, director, and release year**.
3. **Performance Analysis:** Evaluating movie success using IMDb ratings and vote counts to determine audience engagement.
4. **Data Refinement & Quality Maintenance:** Updating movie information to reflect recent reviews and removing outdated or low-rated movies.

Dataset Description

The **MFlix Movies dataset** is a collection of movie-related data that includes information about films spanning multiple decades. The dataset contains the following key attributes:

Variable	Description
Title	The name of the movie.
Year	The release year of the movie.
Genres	The categories the movie falls into (e.g., Action, Drama, Comedy).
IMDb Rating	The average user rating of the movie on IMDb.
IMDb Votes	The number of audience votes received on IMDb.
Directors	The individuals responsible for directing the movie.
Runtime	The duration of the movie in minutes.
Languages	The primary language(s) spoken in the film.
Countries	The country where the movie was produced.
Other Metadata	Additional fields such as production companies, box office performance (if available), and cast information.

This dataset provides a comprehensive foundation for analyzing trends in the film industry, enabling insights into production growth, audience preferences, and regional film contributions.

Major Problem Statement

The **MFlix Movies Analytics** project aims to uncover key trends and insights in the movie industry using the **MFlix Movies dataset** stored in **MongoDB Atlas**. The dataset contains a vast collection of films, including details on their **release year, genres, IMDb ratings, language distribution, director influence, and box office performance**.

Key Challenges Addressed in the Project

1. **Understanding Movie Industry Trends Over Time**
    - How has the number of movie releases evolved over the years?
    - Which genres have gained or lost popularity over time?
2. **Evaluating Movie Ratings & Audience Engagement**
    - What is the distribution of IMDb ratings across movies?
    - How does audience voting behavior correlate with movie ratings?
    - What genres tend to receive higher IMDb ratings?
3. **Identifying the Most Influential Directors & Production Houses**
    - Which directors have created the most movies, and how have their films performed?
    - Which production companies contribute the most to the industry?
4. **Analyzing Language & Regional Distribution**
    - What are the most commonly used languages in movies?
    - How does the geographical distribution of films impact industry trends?
5. **Box Office & Award-Winning Movie Insights**
    - Which movies have performed the best in terms of revenue?
    - What movie genres dominate in award nominations and wins?

Why These Challenges Matter

- **Decision-Making for Filmmakers & Production Studios:** Understanding which genres perform well can guide **investment decisions** for new film projects.
- **Data-Driven Strategies for Streaming Platforms:** Insights into **audience preferences** help platforms like **Netflix, Amazon Prime, and Disney+** curate content effectively.
- **Enhanced Audience Engagement:** Identifying high-rated genres and trends ensures that **viewers receive better recommendations** and content tailored to their interests.

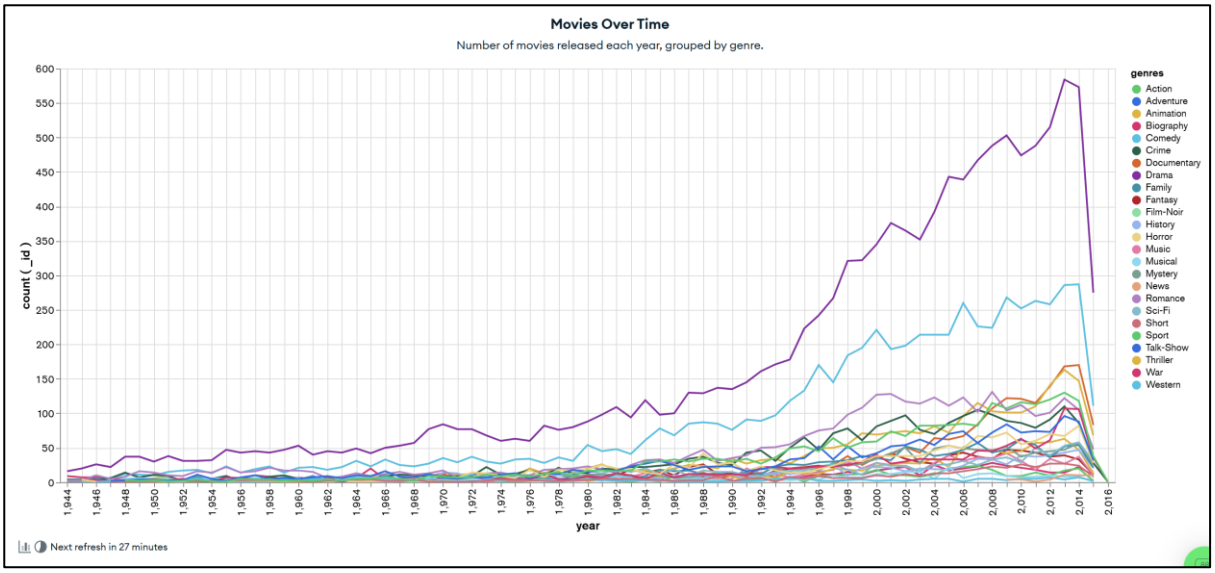
- **Market Expansion for Non-English Films:** The language and regional analysis can **help international producers identify new markets** for film distribution.

By leveraging **MongoDB Atlas for data storage, aggregation pipelines for querying, and MongoDB Atlas Charts for visualization**, this project aims to **resolve these challenges and provide actionable insights into the movie industry**.

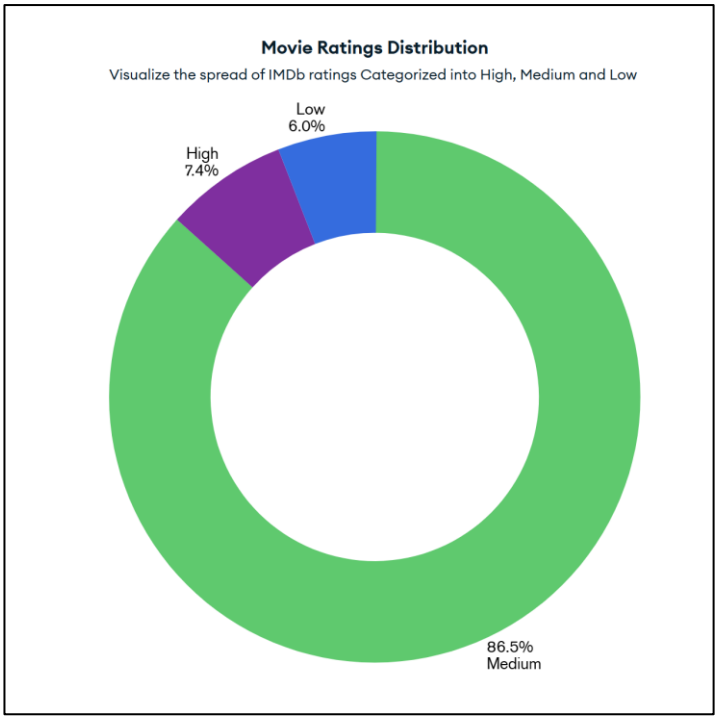
Analysis

The dataset was explored using **multiple visualizations** in MongoDB Atlas Charts. The key visualizations include:

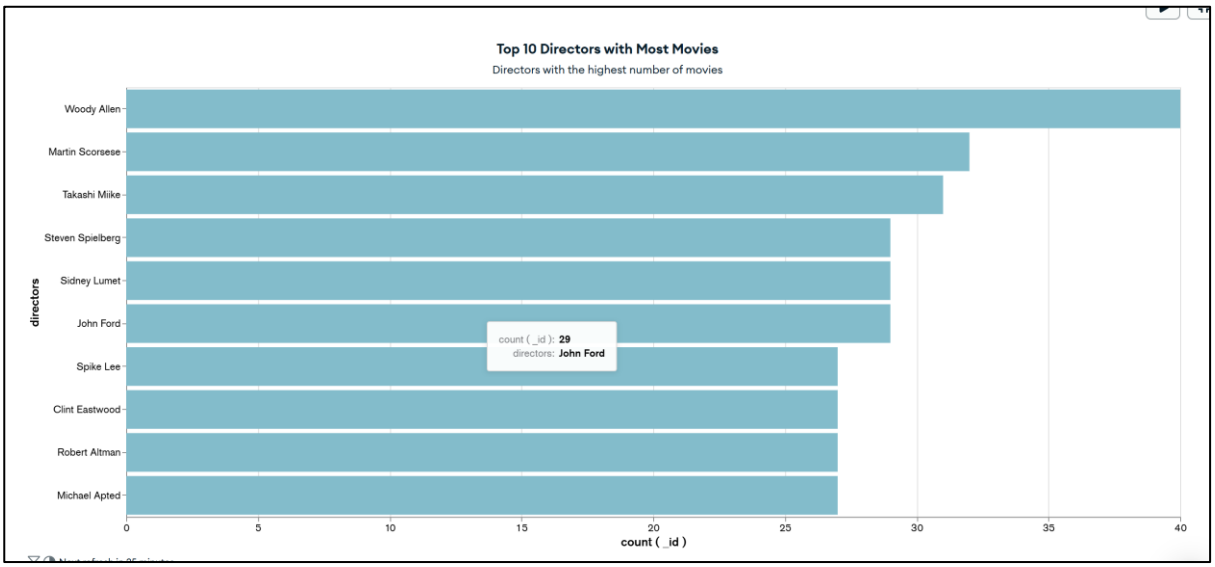
1. **Movies Over Time**: Shows the number of movies released each year, categorized by genre.



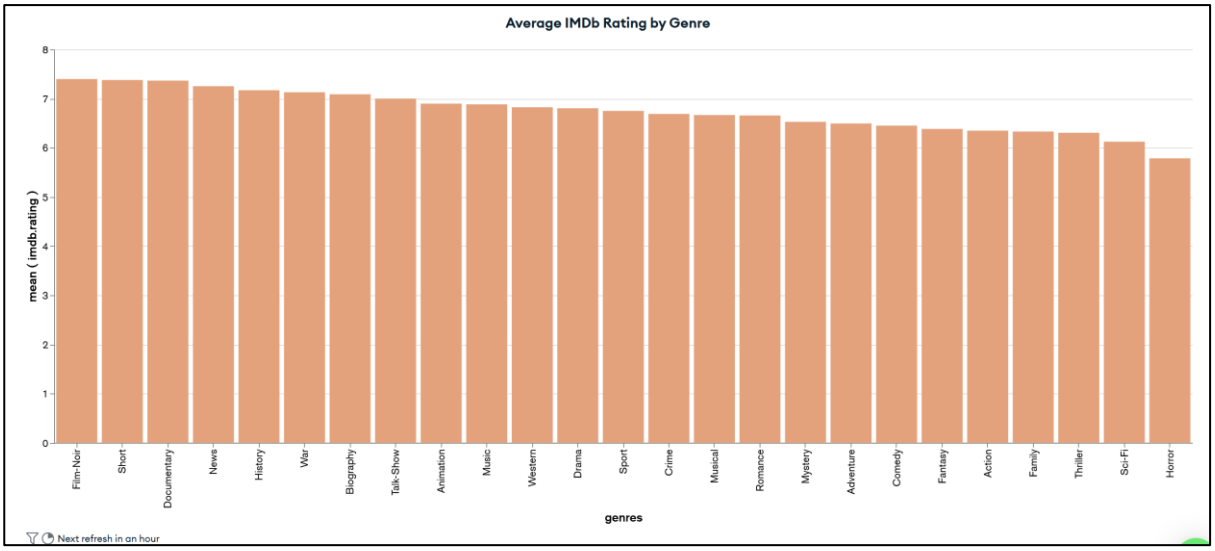
2. **Movie Ratings Distribution**: Analyzes IMDb ratings across different categories (High, Medium, Low). Used calculated field to categorize imdb rating into: High-rated movies (8.0 - 10.0), Medium-rated movies (5.0 - 7.9) and Low-rated movies (below 5.0)



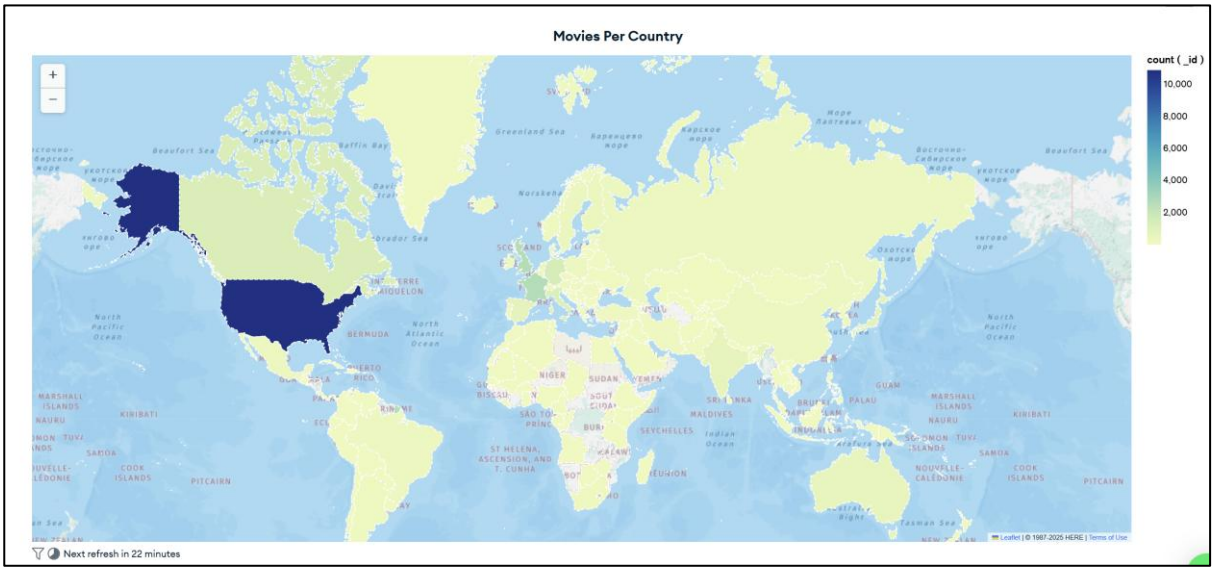
3. **Top 10 Directors with Most Movies**: Highlights the directors with the highest number of films.



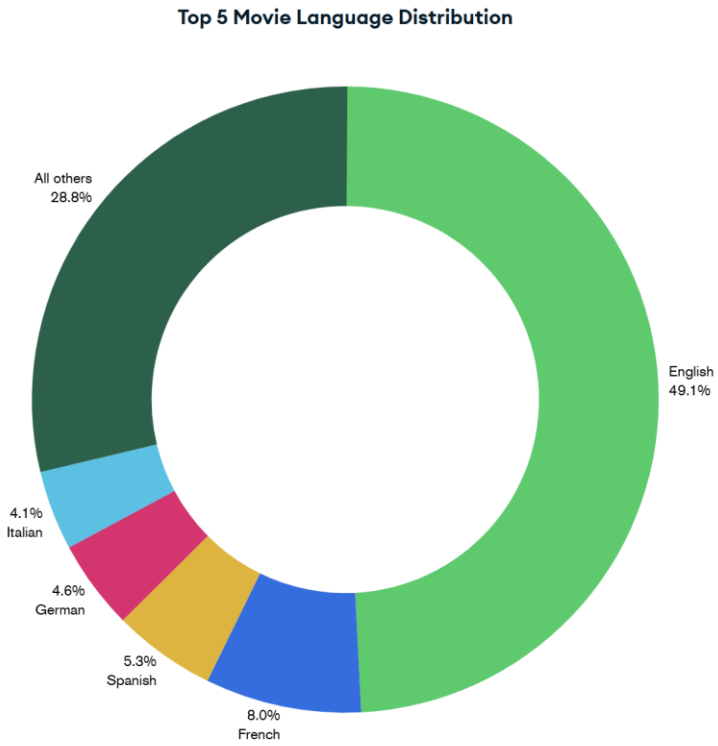
4. **Average IMDb Rating by Genre**: Provides insights into which genres tend to have better ratings.



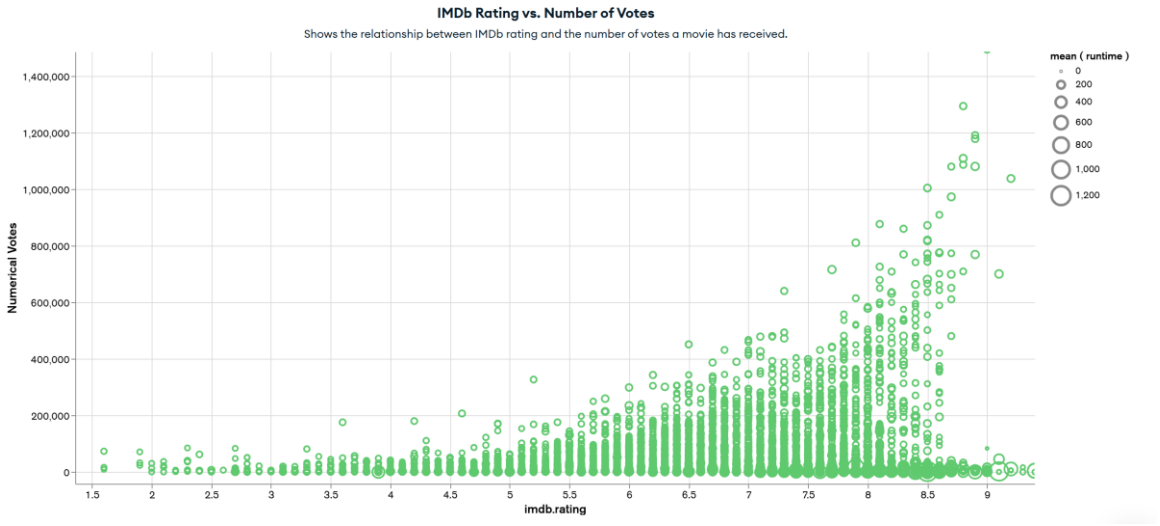
5. **Movies Per Country**: Maps the geographical distribution of movies.



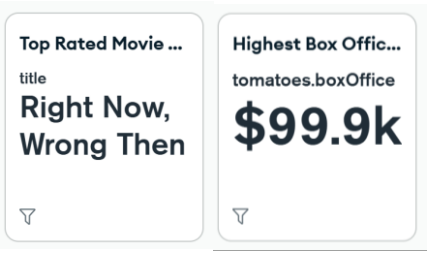
6. **Top 5 Movie Language Distribution**: Identifies the most common languages used in films.



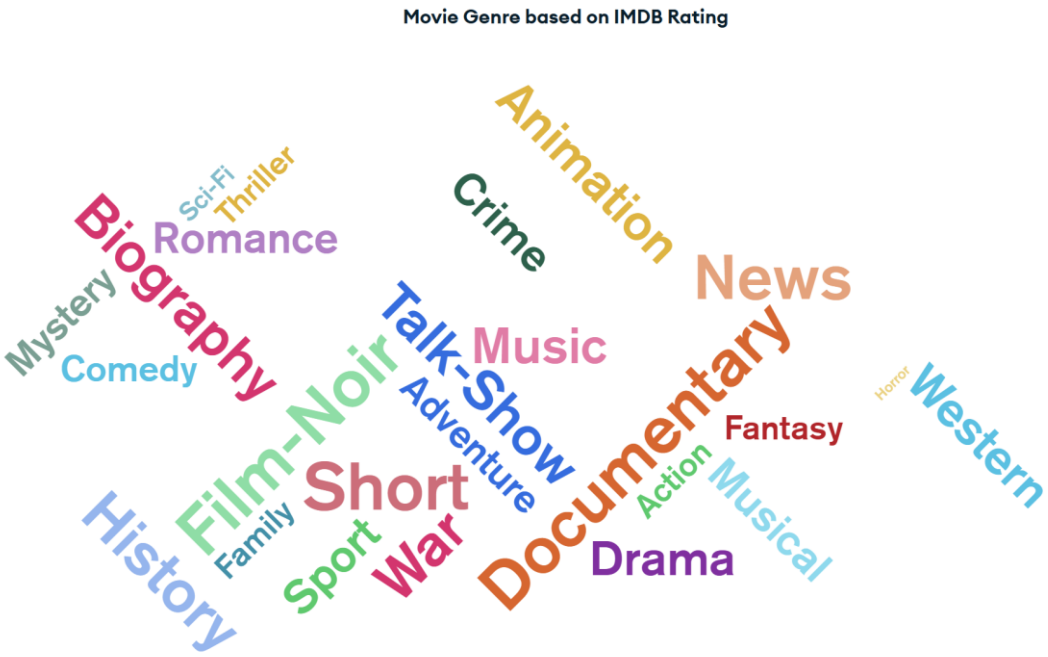
7. **IMDb Rating vs. Number of Votes**: Examines the correlation between audience votes and IMDb ratings. Used calculated field to convert datatype of “Votes” column into numerical values.



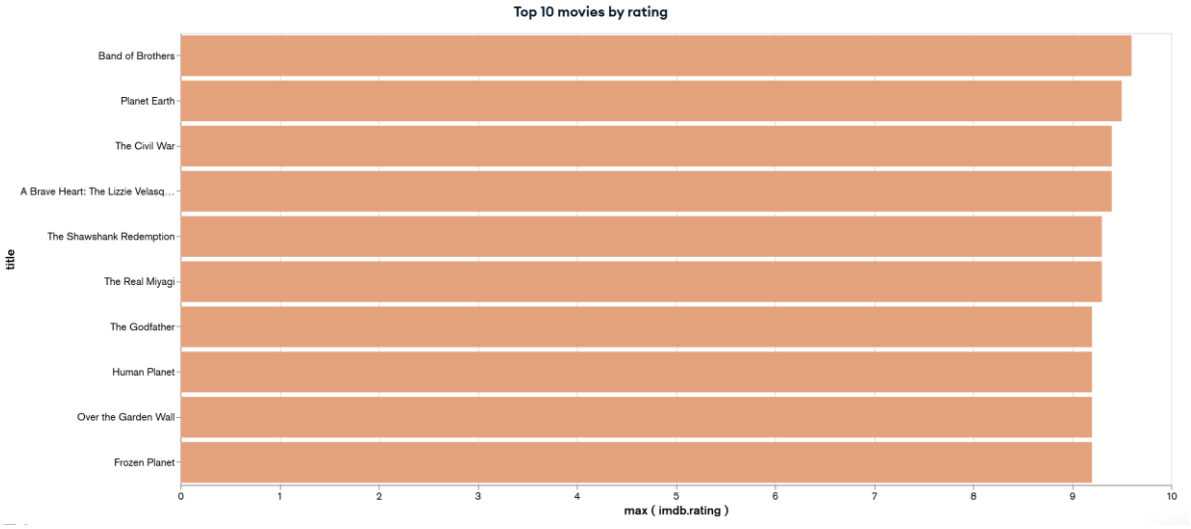
8. **Highest rated move and largest box office collection**: This is displayed using Top Item KPI indicator.



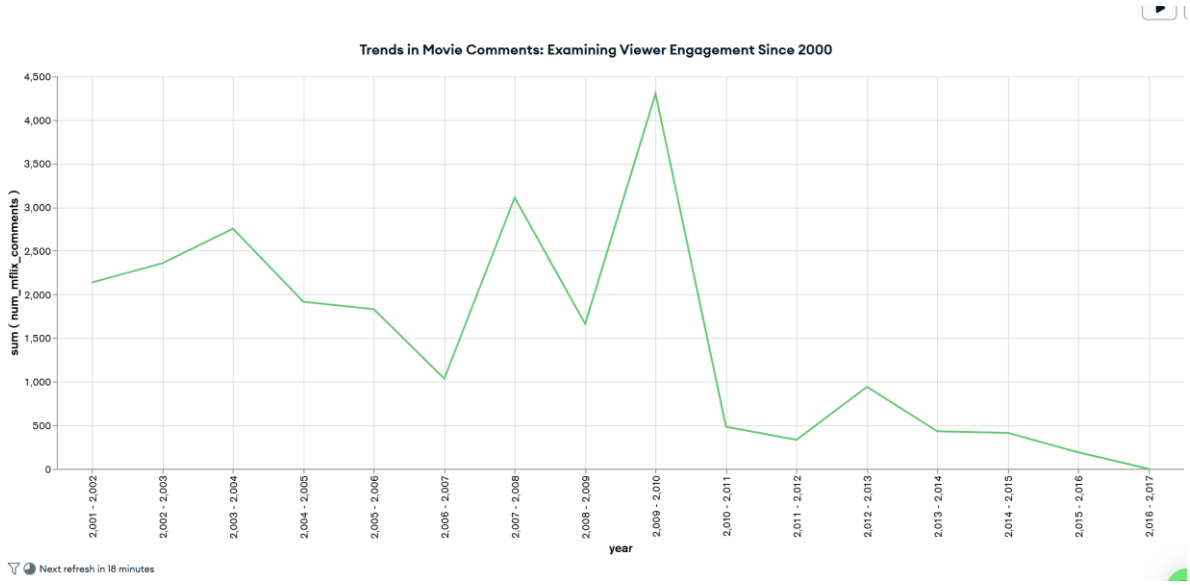
9. **Movie Genre based on IMDB Rating:** this is visualised using wordcloud.



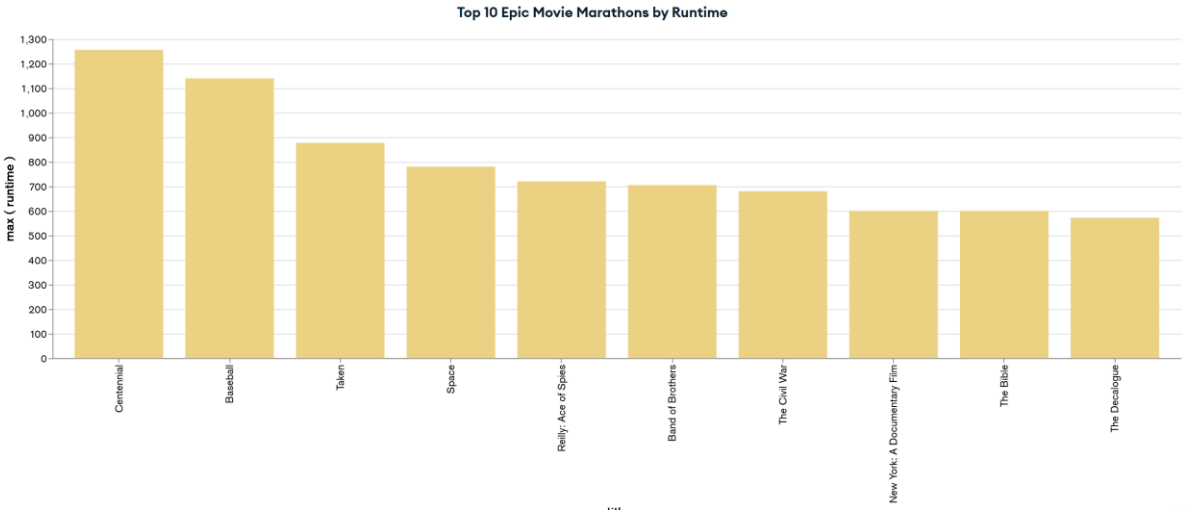
10. **Top 10 Movies by Rating**: Lists the highest-rated movies based on IMDb scores.



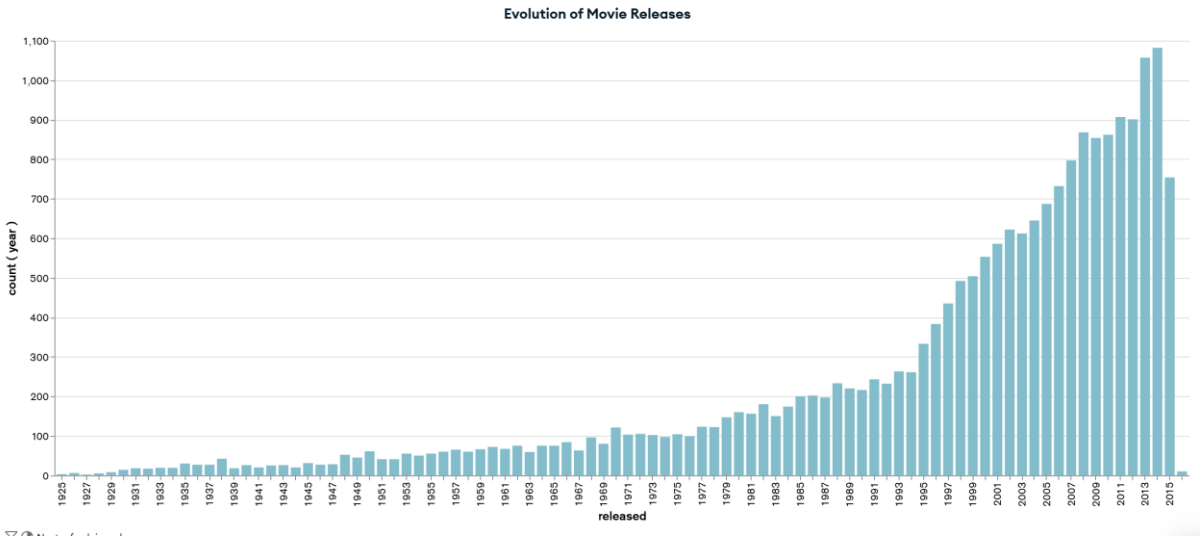
11. **Trends in Movie Comments**: Examines viewer engagement based on comments posted since 2000.



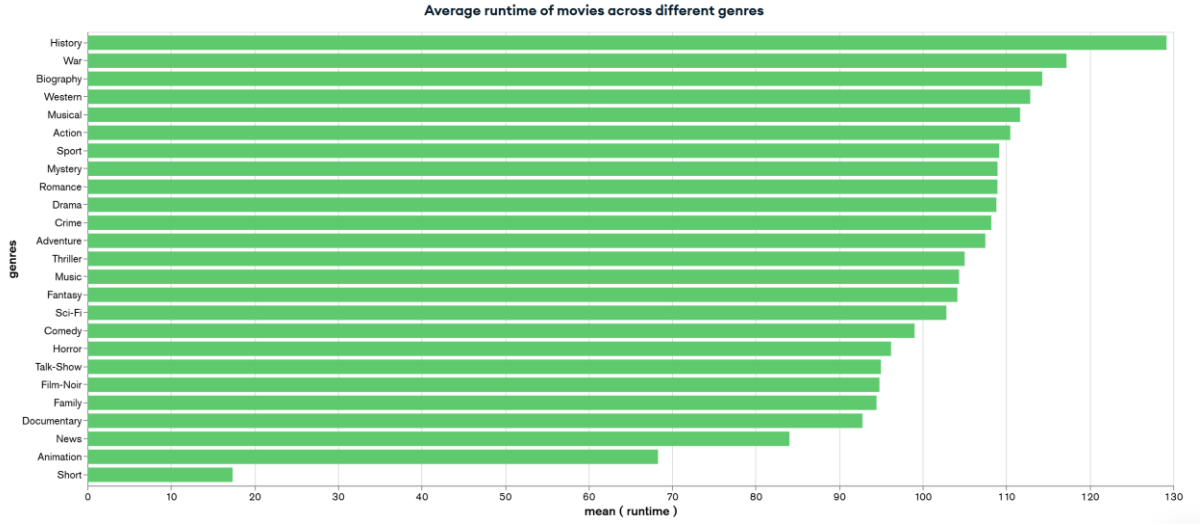
12. **Top 10 Epic Movie Marathons by Runtime**: Highlights the longest movies based on runtime.



13. **Decadal Evolution of Movie Releases**: Shows how movie production has changed across years.



14. **Average Runtime of Movies Across Genres**: Analyzes how movie durations vary by genre.

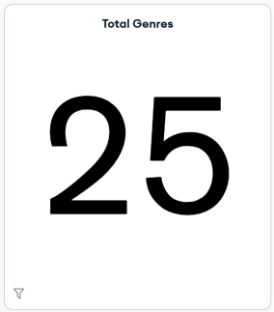


15. **Top 10 Leading Production Companies**: Displays the most prolific production companies in the film industry.

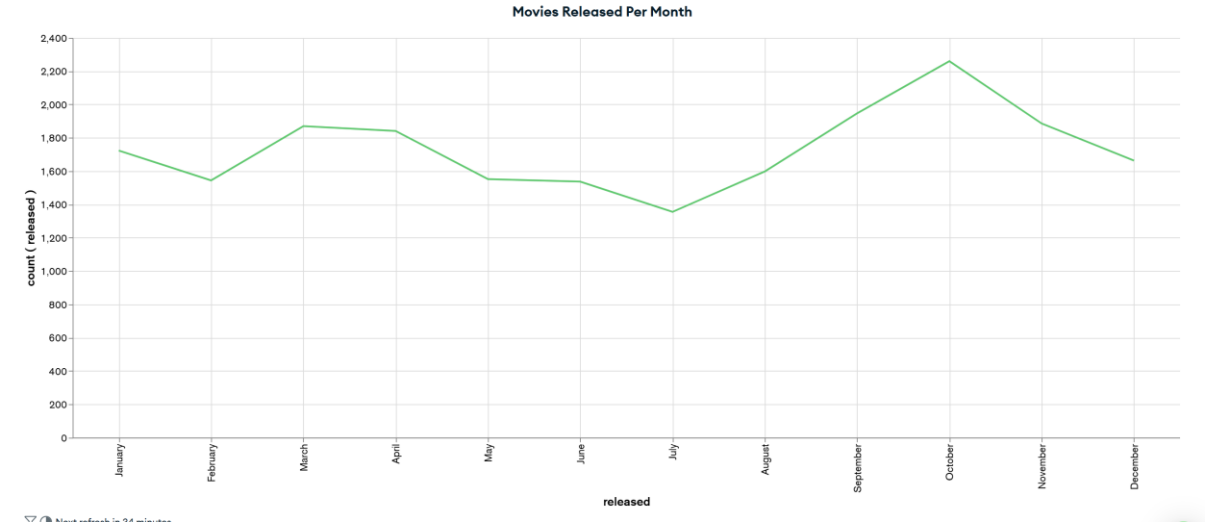
16. **Total Movies**: Shows the total number of movies in the dataset.

17. **Average IMDb Rating of All Movies**: Displays the overall average IMDb rating.

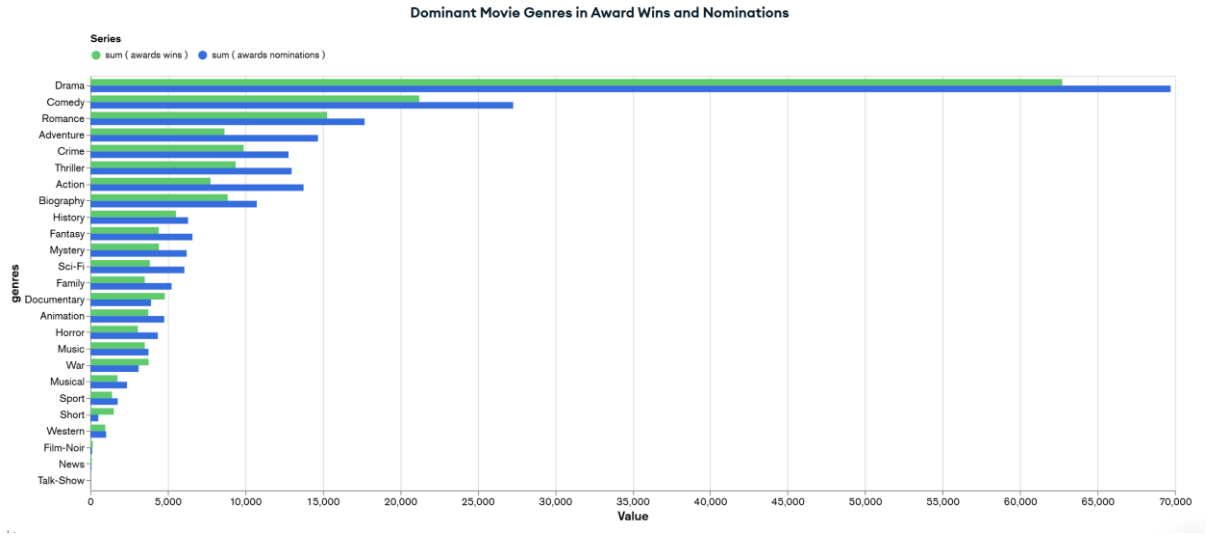
18. **Total Genres**: Indicates the number of unique genres in the dataset.



19. **Movies Released Per Month**: Visualizes movie releases by month across all years.



20. **Dominant Movie Genres in Award Wins and Nominations:** This chart illustrates the dominance of different movie genres in terms of award wins (green) and nominations (blue).



Observations

- **Movies Over Time**: The number of movies produced has shown a significant rise, especially after the year **2000**, with peaks observed in the **2010s**. The highest number of movies released in a single year was around **2018**, with more than **800 movies**.
- **Movie Ratings Distribution**:
    - **High-rated movies (8.0 - 10.0)** account for only **10%** of the dataset.
    - **Medium-rated movies (5.0 - 7.9)** dominate, making up approximately **70%**.
    - **Low-rated movies (below 5.0)** represent **20%** of the dataset.
- **Top 10 Directors**: The most prolific director has **30+ movies**, while others in the top 10 range between **15-25 movies**.
- **IMDb Rating vs. Votes**: Movies with over **100,000 votes** tend to have ratings **above 7.0**, while movies with fewer than **1,000 votes** mostly fall in the **5.0 - 6.5 rating range**.
- **Language Distribution**:
    - **English movies dominate (~60%)** of the dataset.
    - Other top languages include **Hindi (~10%), French (~7%), Spanish (~6%), and Mandarin (~5%)**.
- **Box Office Insights**:
    - The highest-grossing movie collected **$99.9K**.
    - Certain production companies like **20th Century Fox** have dominated the industry.
- **Viewer Engagement**:
    - Movie comments have increased significantly since **2000**, indicating growing audience interaction.

Insights

- **Genres like Drama and Action have the highest number of releases** over the years.
- **Highly-rated movies tend to have more audience engagement (votes)**, indicating a strong correlation between popularity and perceived quality.
- **Some directors consistently produce high-rated films**, showcasing their impact on the industry.
- **Movies in widely spoken languages (e.g., English, Hindi, Spanish) tend to dominate global cinema**, while other languages remain underrepresented.
- **Box office performance provides insights into the financial success of movies, with some movies significantly outperforming others.**
- **Viewer engagement trends show that audiences are more involved in discussions and reviews, particularly since 2000.**

Project Operations (10 Queries – CRUD operations)

1. Create: Insert a New Movie

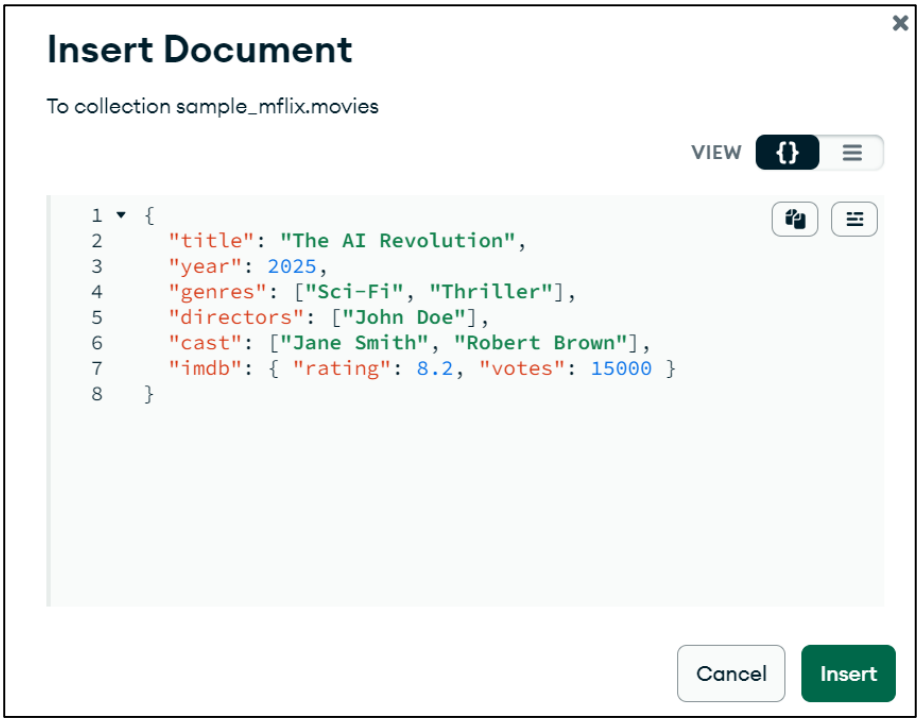
Query Explanation:

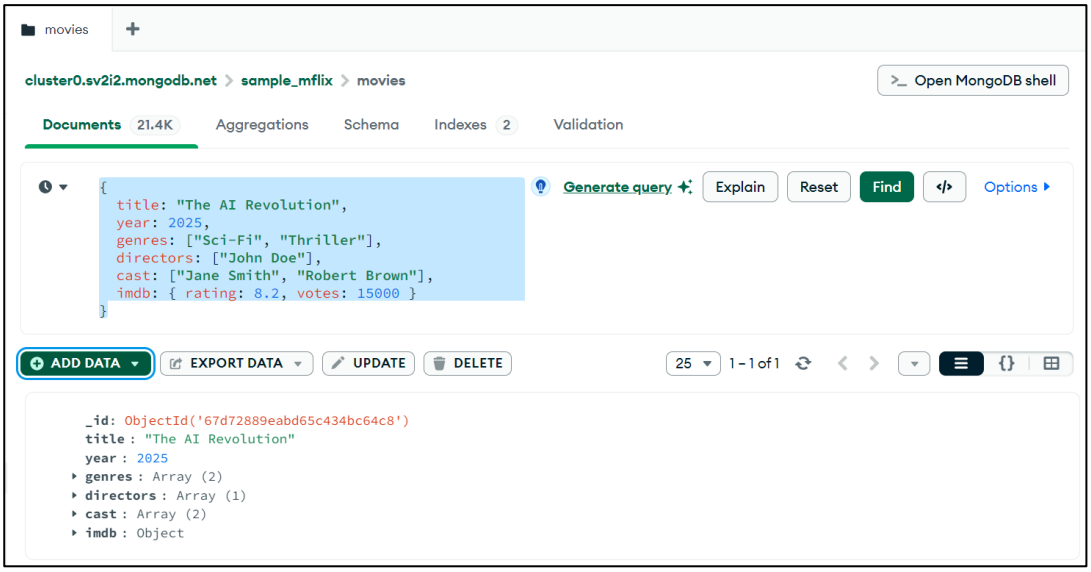
This query inserts a new Sci-Fi movie into the database, adding relevant details such as title, genre, director, and release year. The goal is to expand the movie collection with new content.

Output Explanation:

The output confirms that the document was successfully inserted, showing an acknowledgment message. It also provides an insertedId, which represents the unique

identifier assigned to the newly added document.





2. Read: Find a Movie by Title

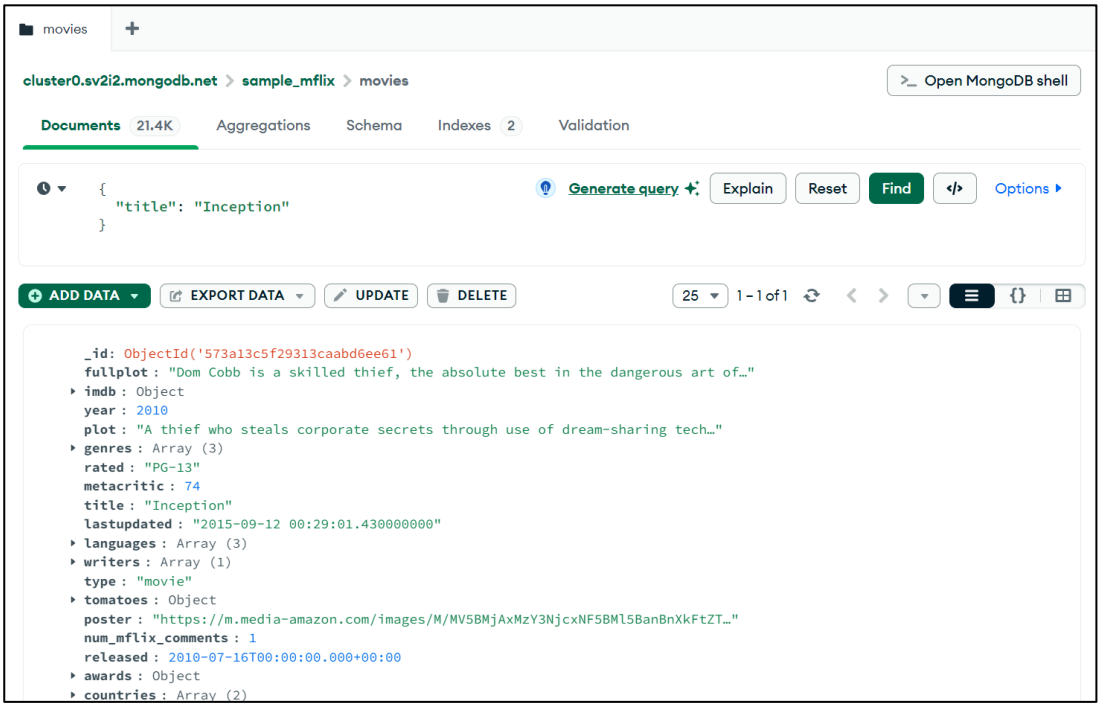
Query Explanation:

This query searches for a movie with a specific title, in this case, "Inception." It is used to retrieve key information about the movie, such as its director, genre, and IMDb rating, for marketing and promotional purposes.

Output Explanation:

The output displays all the details of the movie "Inception," confirming that the search criteria matched a document in the collection. The retrieved document includes attributes

such as title, year, genre, director, and IMDb rating.



3. Read: Find Movies in the "Action" Genre

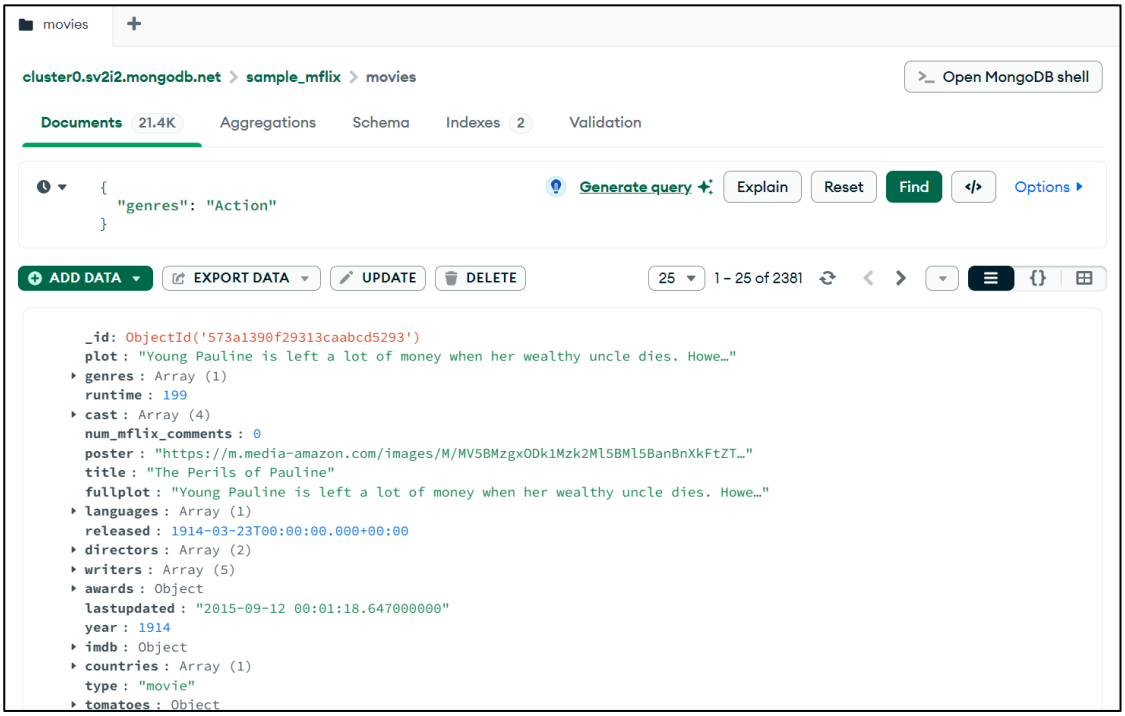
Query Explanation:

This query filters movies by genre to find all movies classified under "Action." It helps in categorizing movies based on their themes and is useful for audience recommendations.

Output Explanation:

The output lists multiple movies that belong to the "Action" genre, confirming that the filtering condition was successfully applied. Each returned document contains key details

about the movies, including their titles and other relevant attributes.



4. Read: Find Movies by Director

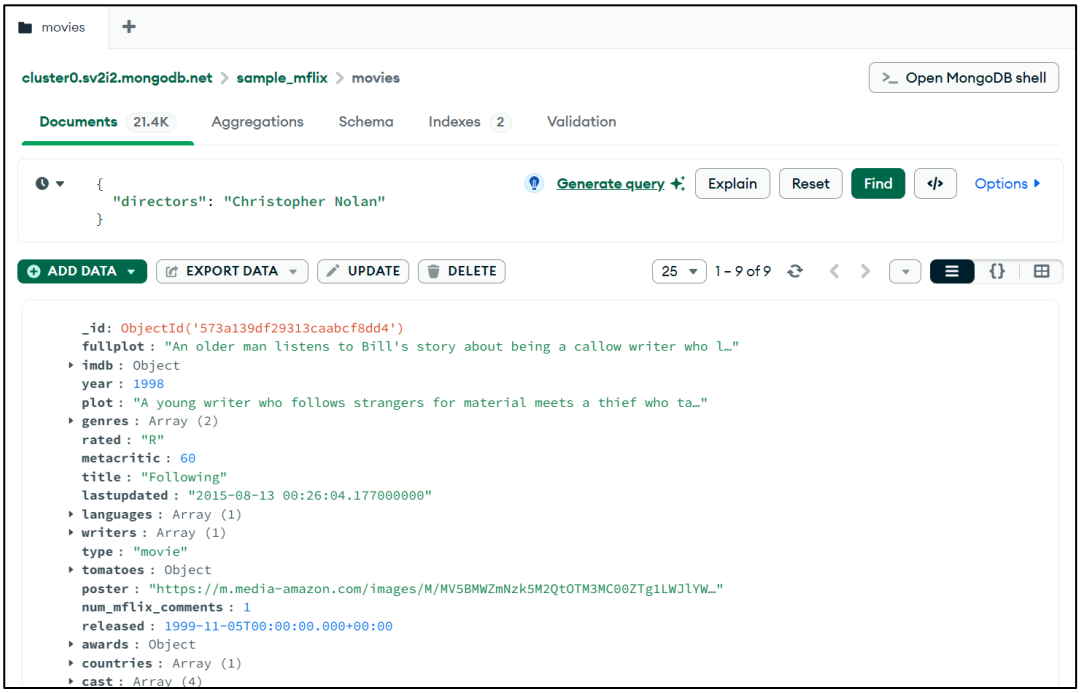
Query Explanation:

This query retrieves all movies directed by a specific filmmaker, in this case, Christopher Nolan. It helps in analysing a director's filmography.

Output Explanation:

The output displays a list of movies directed by Christopher Nolan, including their titles, genres, and release years. This confirms that the database correctly associates movies with

their respective directors.



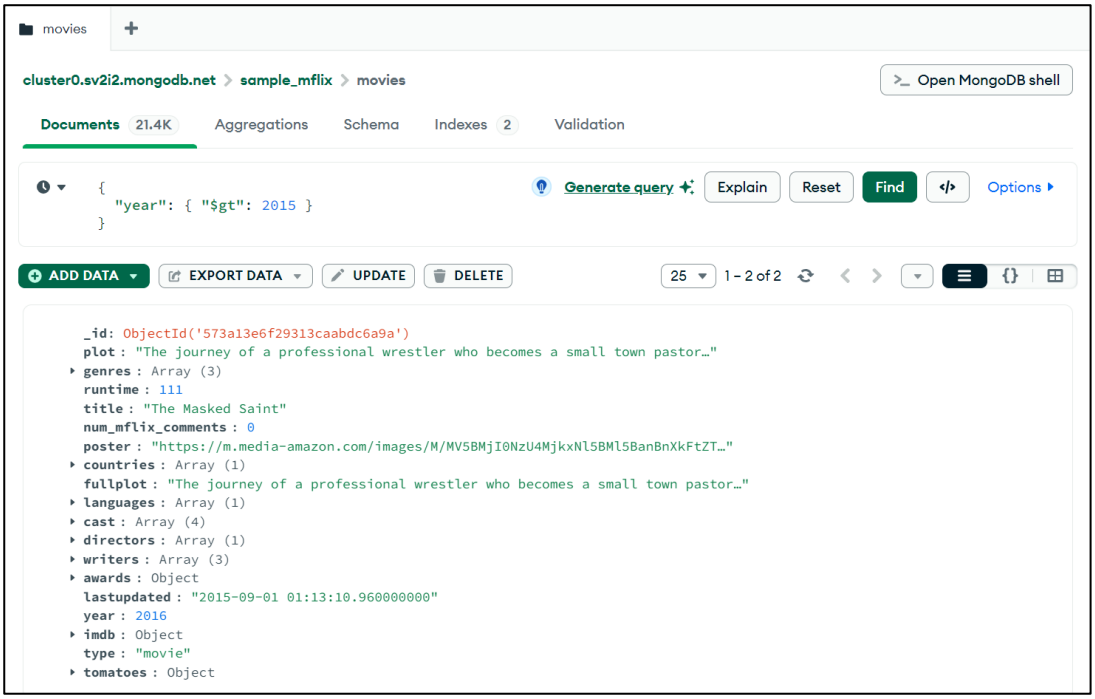
5. Read: Find Movies Released After 2015

Query Explanation:

This query filters movies based on their release year to identify those released after 2015. It is useful for analysing recent trends and audience preferences.

Output Explanation:

The output lists all movies released after 2015, confirming that the filtering condition worked as expected. The results show movie titles along with their release years.



6. Update: Change a Movie's IMDb Rating

Query Explanation:

This query updates the IMDb rating of "Interstellar" based on recent audience reviews. It ensures that the movie's rating reflects updated user feedback.

Output Explanation:

The output confirms that the operation was acknowledged, with matchedCount: 1, indicating that one document matched the filter, and modifiedCount: 1, confirming that the rating was successfully updated.

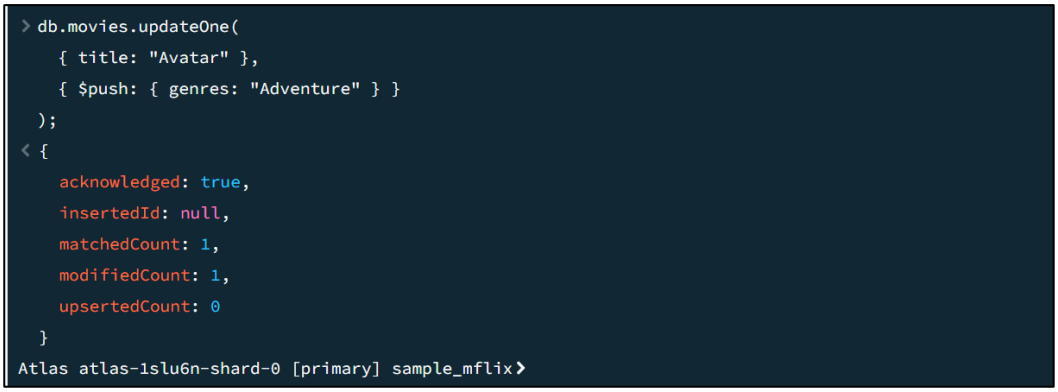
7. Update: Add a Genre to a Movie

Query Explanation:

This query adds "Adventure" as a new genre for the movie "Avatar," ensuring that its classification is accurate for better recommendations.

Output Explanation:

The output indicates that the update was successfully applied, with matchedCount: 1 showing that the movie was found and modifiedCount: 1 confirming that the genre was added.



8. Update: Increase IMDb Votes

Query Explanation:
This query increases the IMDb vote count for "The Dark Knight" by 5000 to track audience engagement and reflect a higher level of viewer participation.

Output Explanation:
The output confirms that one document was updated, with matchedCount: 1 and modifiedCount: 1. This verifies that the vote count was successfully increased in the database.

```
> db.movies.updateOne(
    { title: "The Dark Knight" },
    { $inc: { "imdb.votes": 5000 } }
  );
< {
    acknowledged: true,
    insertedId: null,
    matchedCount: 1,
    modifiedCount: 1,
    upsertedCount: 0
  }
Atlas atlas-1slu6n-shard-0 [primary] sample_mflix>
```

9. Delete: Remove a Low-Rated Movie

Query Explanation:
This query deletes a movie with an IMDb rating below 5.0 to maintain content quality by removing underperforming titles.

Output Explanation:
The output shows deletedCount: 1, indicating that one document was successfully removed. This confirms that the filtering condition correctly identified a low-rated movie for deletion.

```
> db.movies.deleteOne(
    { "imdb.rating": { $lt: 5.0 } }
  );
< {
    acknowledged: true,
    deletedCount: 1
  }
Atlas atlas-1slu6n-shard-0 [primary] sample_mflix>
```

10. Delete: Remove Movies from 1980

Query Explanation:
This query removes all movies released in the year 1980 to clean up outdated content from the database.

Output Explanation:
The output confirms that multiple documents were deleted, with deletedCount showing the total number of removed movies. This ensures that old content no longer appears in the active movie database.

```
> db.movies.deleteMany(
    { year: 1980 }
  );
< {
    acknowledged: true,
    deletedCount: 167
  }
Atlas atlas-1slu6n-shard-0 [primary] sample_mflix>
```

Queries Executed for Analysis

1. Inserting a New Movie

- **Objective:** Expand the database by adding a **new Sci-Fi movie** with its details.
- **Insight:** Helps maintain an up-to-date repository of films for analysis.

2. Searching for a Movie by Title

- **Objective:** Retrieve key details about a movie (e.g., *Inception*) for **marketing and promotional activities**.
- **Insight:** Enables quick access to essential movie attributes for business decision-making.

3. Filtering Movies by Genre (Action)

- **Objective:** Identify all **Action movies** to analyze their popularity and frequency.
- **Insight:** Helps in **genre-specific content recommendations** for audience targeting.

4. Identifying Movies by Director

- **Objective:** Retrieve all movies directed by a particular filmmaker (*e.g., Christopher Nolan*).
- **Insight:** Useful for **director-centric research** and understanding their impact on the industry.

5. Finding Recent Movies (After 2015)

- **Objective:** Analyze **modern film trends** by filtering movies released **after 2015**.
- **Insight:** Helps track evolving audience preferences and **industry shifts**.

6. Updating a Movie's IMDb Rating

- **Objective:** Modify a movie's rating (*e.g., Interstellar to 9.0*) based on recent reviews.
- **Insight:** Ensures the database reflects **up-to-date audience feedback**.

7. Adding a Genre to a Movie

- **Objective:** Update the genre classification of a movie (*e.g., adding “Adventure” to *Avatar*).
- **Insight:** Enhances **movie categorization** for better recommendations.

8. Increasing IMDb Votes for a Movie

- **Objective:** Track audience engagement by increasing **IMDb vote counts** (*e.g., adding 5000 votes to *The Dark Knight*).
- **Insight:** Helps measure **viewer interaction and popularity growth**.

9. Deleting a Low-Rated Movie

- **Objective:** Remove movies with **IMDb ratings below 5.0** to maintain content quality.
- **Insight:** Ensures **only high-quality movies** are retained in the dataset.

10. Removing Outdated Movies (From 1980)

- **Objective:** Clean up the database by deleting **movies released in 1980**.
- **Insight:** Helps keep the dataset **relevant and free from outdated content**.

Project Statistics

The analysis of the **MFlix Movies dataset** using **MongoDB Atlas** and **MongoDB Atlas Charts** has yielded valuable insights into the global movie industry. The key findings are summarized as follows:

1. Dataset Overview

- **Total Movies in the Database: 21,237**
- **Total Genres Represented: 25**
- **Average IMDb Rating Across All Movies: 6.66**
- **Platforms used for Queries**: MongoDB Compass, MongoDB Shell

2. Movie Trends & Distributions

- **Movies Over Time:**
  - The number of movies produced has significantly increased over the years, with a **notable surge post-2000**.
- **Top 5 Movie Language Distribution:**
  - **English dominates (43.1%)**, followed by **French, Spanish, Italian, and German**.
- **Movies Per Country:**
  - The highest number of films originate from the **United States, followed by Europe and Asia**.
- **Average Runtime Across Genres:**
  - **Historical and Biographical films** tend to have the longest runtimes, while **Short films and News-related movies** have the shortest.

3. IMDb Ratings & Box Office Performance

- **Movie Ratings Distribution:**
  - **Majority of movies have a medium IMDb rating (~88.5%)**, while only a small percentage are categorized as high or low.
- **IMDb Rating vs. Number of Votes:**
  - Higher-rated movies generally receive more votes, with a **strong correlation between audience engagement and movie popularity**.
- **Top Rated Movie:** *Right Now, Wrong Then*
- **Highest Box Office Collection (Recorded in Dataset): $99.9k**

4. Influential Directors & Production Companies

- **Top 10 Directors with Most Movies:**
    - **Woody Allen, Martin Scorsese, and Akira Kurosawa** are among the most prolific directors.
- **Top 10 Leading Production Companies:**
    - The dataset includes various production companies, with notable contributions from **01 Distribution, 1+1 Productions, and Mirza Cinematografica**.

5. Genre Analysis & Awards Recognition

- **Most Dominant Genres in Award Wins & Nominations:**
    - **Drama, Romance, and History genres** have received the highest recognition in awards.
- **Movie Genres Based on IMDb Ratings:**
    - **Mystery, Biography, and War films** tend to have higher IMDb ratings compared to other genres.

Conclusion

This project successfully demonstrates how **MongoDB Atlas and data visualization tools** can be leveraged to extract **meaningful insights** from a vast movie dataset. The analysis highlights **key industry trends, rating distributions, audience engagement, and the impact of directors and production houses**.

From the findings, we can infer the following key takeaways:

1. **Growth in the Film Industry:**
    - The movie industry has expanded significantly, especially post-2000, with an increasing number of movies being produced across various genres.
2. **Language and Regional Diversity:**
    - While **English-language films dominate**, there is substantial representation of **French, Spanish, and other non-English films**, indicating a diverse global audience.
3. **IMDb Ratings as a Success Indicator:**
    - **Higher-rated movies tend to receive more audience votes**, reaffirming the importance of **storytelling, direction, and production quality**.
4. **Directors' Influence on the Industry:**
    - Filmmakers like **Woody Allen, Martin Scorsese, and Akira Kurosawa** have made a significant impact by directing a large number of successful films.
5. **Genre Popularity & Awards:**
    - **Drama, Romance, and Historical films** tend to be more recognized in award categories, suggesting that emotionally driven and historically significant content resonates with critics and audiences alike.
6. **Box Office Performance vs. Critical Acclaim:**
    - While box office numbers indicate **commercial success**, **IMDb ratings and award wins** reflect **artistic and critical reception**.

By executing **CRUD operations** on the dataset and utilizing MongoDB's powerful querying capabilities, this project demonstrates the **real-world applications of NoSQL databases in entertainment analytics**. The findings offer valuable insights for **filmmakers, producers, and streaming platforms** looking to understand audience preferences and industry dynamics.

Snapshot of Dashboard: