Project Report
Big Data Management Analytics

Analysis of the Movies Collection in MongoDB

Submitted to: Prof. Amarnath Mitra
FORE School of Management, New Delhi

Submitted by: Jagriti (055016)
Harsh Jain (055018)

B-18, Qutub Institutional Area
New Delhi – 110016

Table of Contents

1. Project Overview

This project focuses on analyzing a movie dataset stored in MongoDB. The dataset, in JSON format, aligns well with MongoDB's document-oriented structure. MongoDB Compass is utilized for querying and performing CRUD operations, while MongoDB Atlas is used for data visualization.

2. Dataset Description

The **Movies** collection, sourced from the Mflix dataset, comprises **23,149 documents** and occupies approximately **1.2 GB** of storage in MongoDB Atlas. Each document corresponds to a single movie and contains multiple attributes such as title, genre, cast, and ratings. The dataset follows a **semi-structured (BSON) format** with nested fields and arrays, making it well-suited for complex and dynamic queries.

Each movie document generally includes the following attributes:

(i) _id (ObjectId) – A unique identifier for each movie. *(e.g., "5a934e000102030405000000")*

(ii) title (String) – The movie's name. *(e.g., "Inception")*

(iii) year (Number) – The release year. *(e.g., 2010)*

(iv) genres (Array of Strings) – The categories the movie belongs to. *(e.g., ["Action", "Sci-Fi", "Thriller"])*

(v) cast (Array of Strings) – List of actors. *(e.g., ["Leonardo DiCaprio", "Joseph Gordon-Levitt", "Elliot Page"])*

(vi) directors (Array of Strings) – Names of directors. *(e.g., ["Christopher Nolan"])*

(vii) writers (Array of Strings) – Screenwriters of the movie. *(e.g., ["Christopher Nolan", "Jonathan Nolan"])*

(viii) languages (Array of Strings) – Available languages. *(e.g., ["English", "Japanese", "French"])*

(xi) countries (Array of Strings) – Production countries. *(e.g., ["USA", "UK"])*

(x) released (Date) – Official release date. *(e.g., ISODate("2010-07-16T00:00:00Z"))*

(xi) runtime (Number) – Duration in minutes. *(e.g., 148)*

(xii) plot (String) – A brief storyline summary. *(e.g., "A thief who steals corporate secrets through dream-sharing technology is given a chance to erase his criminal record.")*

(xiii) fullplot (String) – A detailed synopsis.

(xiv) imdb (Object) – IMDb-specific details:

- **rating (Number)** – IMDb rating. *(e.g., 8.8)*
- **votes (Number)** – Total votes received. *(e.g., 2,000,000)*
- **id (Number)** – IMDb ID. *(e.g., 1375666)*

(xv) tomatoes (Object) – Rotten Tomatoes ratings and reviews:

- **viewer (Object)** – Viewer rating and review count.
- **rating (Number)** – Viewer rating. *(e.g., 4.2)*
- **numReviews (Number)** – Total viewer reviews. *(e.g., 5,000)*

- **critic (Object)** – Critic reviews.
- **fresh (Number)** – Number of positive reviews.
- **rotten (Number)** – Number of negative reviews.
- **lastUpdated (Date)** – Last rating update.

(xvi) type (String) – Media type, typically "movie". *(e.g., "movie")*

3. Project Goals

- Develop a structured **dashboard** that highlights key insights from the Movies collection.
- Investigate relationships between different attributes such as **genre, ratings, and release year** using visual analytics.
- Perform **CRUD operations** efficiently using MongoDB queries.
- Generate **data-driven reports** to support movie trend analysis and decision-making.

4. Queries

(A) CRUD Operations

Create

(i) Insert a new movie:

```
> use sample_mflix
< switched to db sample_mflix
> db.movies.insertOne({
    "title": "New Movie",
    "year": 2025,
    "genres": ["Action", "Adventure"],
    "cast": ["Actor A", "Actor B"],
    "runtime": 120,
    "imdb": { "rating": 7.5, "votes": 1000 }
  })
< {
    acknowledged: true,
    insertedId: ObjectId('67d71d04563dbe35525c437c')
  }
```

(ii) Insert multiple temporary movies:

```
> db.movies.insertMany([
    { title: "Temp Matrix", year: 2021, genres: ["Action", "Sci-Fi"], isTemporary: true },
    { title: "Temp Avatar", year: 2022, genres: ["Adventure", "Fantasy"], isTemporary: true }
  ])
< {
    acknowledged: true,
    insertedIds: {
      '0': ObjectId('67d69dbf67f202db884fb473'),
      '1': ObjectId('67d69dbf67f202db884fb474')
    }
  }
```

Retrieve

(i) Find all temporary movies:

```
> db.movies.find({ isTemporary: true }).pretty()
< {
    _id: ObjectId('67d69d7b67f202db884fb472'),
    title: 'Temp Inception',
    year: 2025,
    genres: [
      'Action',
      'Sci-Fi'
    ],
    cast: [
      'Leonardo DiCaprio',
      'Joseph Gordon-Levitt'
    ],
    imdb: {
      rating: 9,
      votes: 1200000
    },
    isTemporary: true
  }
  {
    _id: ObjectId('67d69dbf67f202db884fb473'),
    title: 'Temp Matrix',
    year: 2021,
    genres: [
      'Action',
      'Sci-Fi'
    ],
    isTemporary: true
```

Update

(i) Update the IMDb rating of a specific movie:

```
> db.movies.updateOne(
    { "title": "New Movie" },
    { $set: { "imdb.rating": 8.0 } }
  )
< {
    acknowledged: true,
    insertedId: null,
    matchedCount: 1,
    modifiedCount: 1,
    upsertedCount: 0
  }
```

(ii) Add a new genre to a temporary movie :

```
> db.movies.updateOne(
    { title: "Temp Inception", isTemporary: true },
    { $addToSet: { genres: "Thriller" } }
  )
< {
    acknowledged: true,
    insertedId: null,
    matchedCount: 1,
    modifiedCount: 1,
    upsertedCount: 0
  }
```

Delete

(i) Remove a movie from the database:

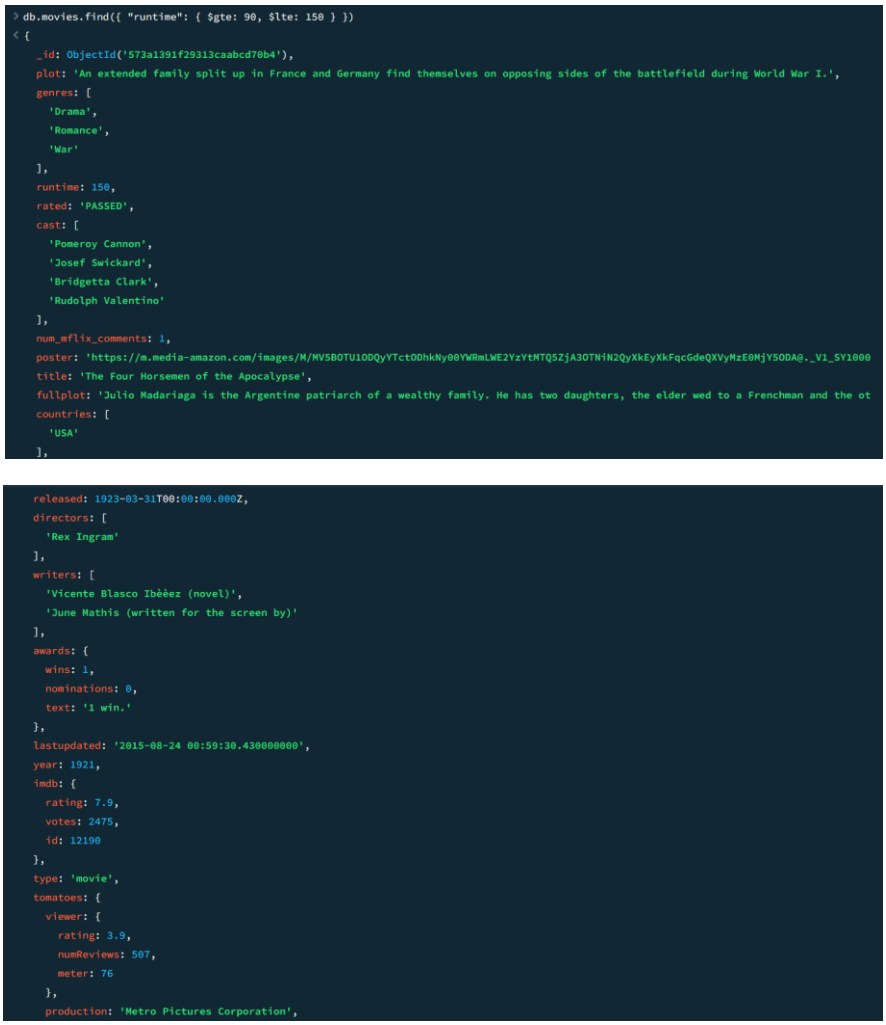
```
> db.movies.deleteOne({ "title": "New Movie" })
< {
    acknowledged: true,
    deletedCount: 1
  }
```

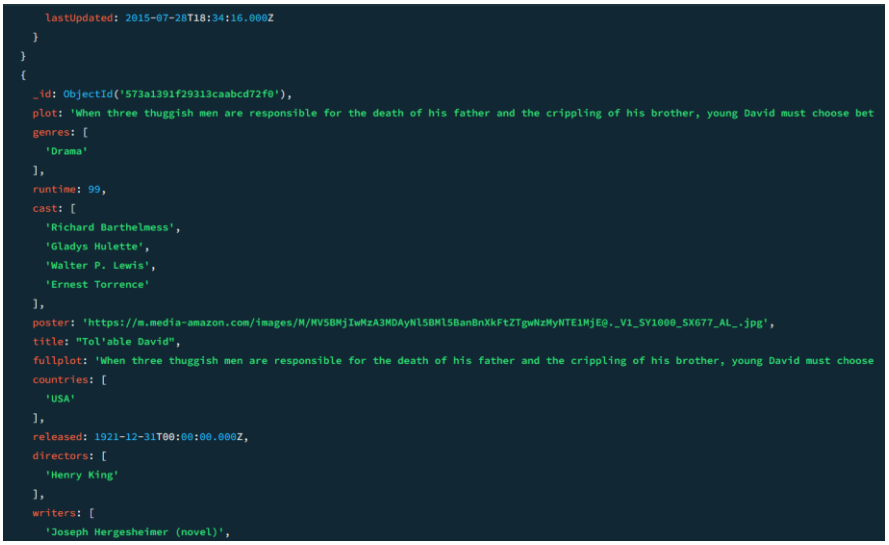
(ii) Delete all temporary movies:

```
> db.movies.deleteMany({ isTemporary: true })
< {
    acknowledged: true,
    deletedCount: 2
  }
```

(B) Filter Queries

(i) Find Movies with an IMDb Rating Greater Than 8:





```
    lastUpdated: 2015-07-28T18:34:16.000Z
  }
}
{
  _id: ObjectId('573a1391f29313caabcd72f0'),
  plot: 'When three thuggish men are responsible for the death of his father and the crippling of his brother, young David must choose bet
  genres: [
    'Drama'
  ],
  runtime: 99,
  cast: [
    'Richard Barthelmess',
    'Gladys Hulette',
    'Walter P. Lewis',
    'Ernest Torrence'
  ],
  poster: 'https://m.media-amazon.com/images/M/MV5BNjIwMzA3MDAyNl5BMl5BanBnXkFtZTgwNzMyNTE1MjE@._V1_SY1000_SX677_AL_.jpg',
  title: "Tol'able David",
  fullplot: 'When three thuggish men are responsible for the death of his father and the crippling of his brother, young David must choose
  countries: [
    'USA'
  ],
  released: 1921-12-31T00:00:00.000Z,
  directors: [
    'Henry King'
  ],
  writers: [
    'Joseph Hergesheimer (novel)',
```

```
    'Edmund Goulding',
    'Henry King'
  ],
  awards: {
    wins: 2,
    nominations: 0,
    text: '2 wins.'
  },
  lastupdated: '2015-08-23 01:12:08.943000000',
  year: 1921,
  imdb: {
    rating: 8.1,
    votes: 1455,
    id: 12763
  },
  type: 'movie',
  tomatoes: {
    viewer: {
      rating: 3.3,
      numReviews: 249,
      meter: 70
    },
    dvd: 1999-03-23T00:00:00.000Z,
    production: 'Universum Film A.G. (UFA)',
    lastUpdated: 2015-08-26T18:24:46.000Z
  },
  num_mflix_comments: 0
}
```

```
{
  _id: ObjectId('573a1391f29313caabcd7472'),
  plot: 'A con artist masquerades a Russian nobility and attempts to seduce the wife of an American diplomat.',
  genres: [
    'Drama'
  ],
  runtime: 117,
  cast: [
    'Rudolph Christians',
    'Miss DuPont',
    'Maude George',
    'Mae Busch'
  ],
  num_mflix_comments: 0,
  poster: 'https://m.media-amazon.com/images/M/MV5BMTk2NDkxMTY1Nl5BMl5BanBnXkFtZTgwNDI1NDU5MTE@._V1_SY1000_SX677_AL_.jpg',
  title: 'Foolish Wives',
  fullplot: '"Count" Karanzim, a Don Juan is with his cousins in Monte Carlo, living from faked money and the money he gets from rich ladi
  languages: [
    'English'
  ],
  released: 1922-01-11T00:00:00.000Z,
  directors: [
    'Erich von Stroheim'
  ],
  writers: [
    'Erich von Stroheim (story)',
    'Marian Ainslee (titles)',
    'Walter Anthony (titles)'
```

```
  ],
  awards: {
    wins: 1,
    nominations: 0,
    text: '1 win.'
  },
  lastupdated: '2015-09-05 00:00:37.817000000',
  year: 1922,
  imdb: {
    rating: 7.3,
    votes: 1777,
    id: 13140
  },
  countries: [
    'USA'
  ],
  type: 'movie',
  tomatoes: {
    viewer: {
      rating: 3.7,
      numReviews: 1079,
      meter: 77
    },
    dvd: 2000-09-19T00:00:00.000Z,
    critic: {
      rating: 9,
      numReviews: 9,
      meter: 89
```

```
    },
    lastUpdated: 2015-09-15T17:02:32.000Z,
    rotten: 1,
    production: 'Universal Pictures',
    fresh: 8
  }
}
{
  _id: ObjectId('573a1391f29313caabcd7626'),
  plot: 'A nobleman becomes the vigilante Robin Hood who protects the oppressed English people from the tyrannical Prince John.',
  genres: [
    'Adventure',
    'Romance',
    'Family'
  ],
  runtime: 143,
  cast: [
    'Wallace Beery',
    'Sam De Grasse',
    'Enid Bennett',
    'Paul Dickey'
  ],
  num_mflix_comments: 0,
  poster: 'https://m.media-amazon.com/images/M/MV5BYzRmMWIyNDEtYTRmYS00Y2FlLWJhOGUtYWZmNTI1YzZjOTc2L2ltYWdlL2ltYWdlXkEyXkFqcGdeQXVyMjUxODE
  title: 'Robin Hood',
  fullplot: "Amid big-budget medieval pageantry, King Richard goes on the Crusades leaving his brother Prince John as regent, who promptly
  languages: [
    'English'
```

```
  ],
  released: 1922-10-18T00:00:00.000Z,
  directors: [
    'Allan Dwan'
  ],
  writers: [
    'Douglas Fairbanks (story)'
  ],
  awards: {
    wins: 1,
    nominations: 0,
    text: '1 win.'
  },
  lastupdated: '2015-08-11 00:29:16.047000000',
  year: 1922,
  imdb: {
    rating: 7.7,
    votes: 1460,
    id: 13556
  },
  countries: [
    'USA'
  ],
  type: 'movie',
  tomatoes: {
    viewer: {
      rating: 3.6,
      numReviews: 659,
```

(C) Aggregation Queries

(i) Count Total Movies in the Database:

```
> db.movies.aggregate([{ $count: "total_movies" }])
< {
    total_movies: 21349
  }
```

(ii) Average IMDb Rating by Genre:

```
> db.movies.aggregate([
    { $unwind: "$genres" },
    { $group: { _id: "$genres", avg_rating: { $avg: "$imdb.rating" } } }
  ])
< {
    _id: 'Romance',
    avg_rating: 6.6564272782136396
  }
  {
    _id: 'Comedy',
    avg_rating: 6.450214658088344
  }
  {
    _id: 'Biography',
    avg_rating: 7.087984189723319
  }
  {
    _id: 'Adventure',
    avg_rating: 6.493680884676145
  }
  {
    _id: 'History',
    avg_rating: 7.1696100917431185
  }
  {
    _id: 'Documentary',
    avg_rating: 7.365679824561403
  }
```

```
{
  _id: 'Animation',
  avg_rating: 6.89669603524229
}
{
  _id: 'Talk-Show',
  avg_rating: 7
}
{
  _id: 'News',
  avg_rating: 7.252272727272728
}
{
  _id: 'Film-Noir',
  avg_rating: 7.397402597402598
}
{
  _id: 'Action',
  avg_rating: 6.347098402018503
}
{
  _id: 'Horror',
  avg_rating: 5.784709897610922
}
{
  _id: 'Mystery',
  avg_rating: 6.527425044091711
}
```

(iii) Count number of movies per year:

(iv) Find the most popular actor (appeared in most movies):

```
> db.movies.aggregate([
    { $unwind: "$cast" },
    { $group: { _id: "$cast", total_movies: { $sum: 1 } } },
    { $sort: { total_movies: -1 } },
    { $limit: 1 }
  ])
< {
    _id: 'Gèrard Depardieu',
    total_movies: 67
  }
```

(v) Top 3 users who commented the most:

```
> db.comments.aggregate([
    { $group: { _id: "$email", total_comments: { $sum: 1 } } },
    { $sort: { total_comments: -1 } },
    { $limit: 3 }
  ])
< {
    _id: 'roger_ashton-griffiths@gameofthron.es',
    total_comments: 277
  }
  {
    _id: 'ron_donachie@gameofthron.es',
    total_comments: 260
  }
  {
    _id: 'jonathan_pryce@gameofthron.es',
    total_comments: 260
  }
```

(vi) Find 20 most common genres:

```
> db.movies.aggregate([
    { $unwind: "$genres" },
    { $group: { _id: "$genres", count: { $sum: 1 } } },
    { $sort: { count: -1 } },
    { $limit: 5 }
  ])
< {
    _id: 'Drama',
    count: 12385
  }
  {
    _id: 'Comedy',
    count: 6532
  }
  {
    _id: 'Romance',
    count: 3318
  }
  {
    _id: 'Crime',
    count: 2457
  }
  {
    _id: 'Thriller',
    count: 2454
  }
```

5. Problem Statement

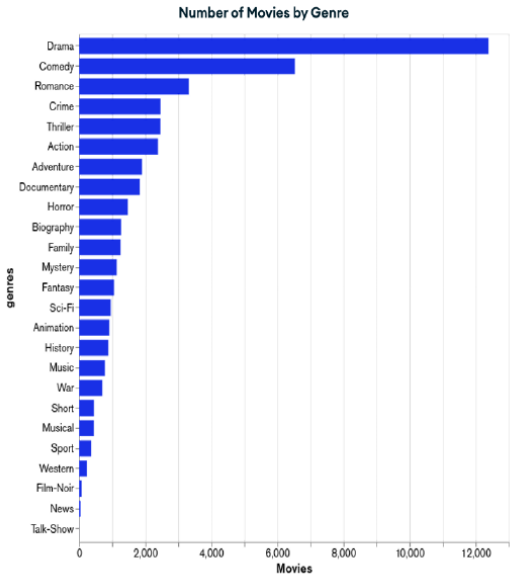
- The movie industry faces challenges in analyzing vast, unstructured datasets to forecast trends and optimize content creation. Conventional analytical methods often fall short in delivering clear insights into genre popularity, audience ratings, and production patterns, limiting data-driven decision-making.

6. Dashboard

A **MongoDB Atlas Dashboard** has been developed to visually represent key metrics from the Movies dataset.

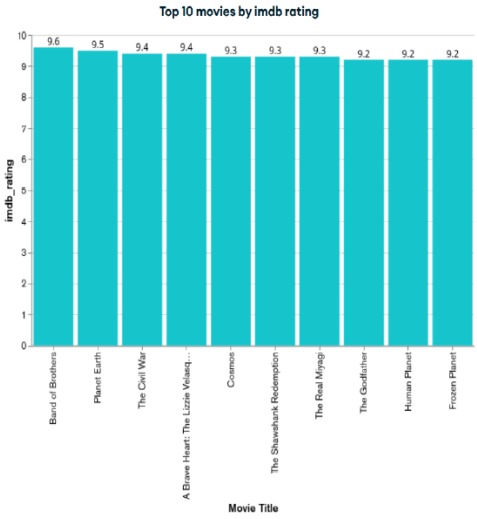
(i) Number of Movies by Genre

- **Objective:** Examine the distribution of movies across various genres to determine which genres are the most and least popular.



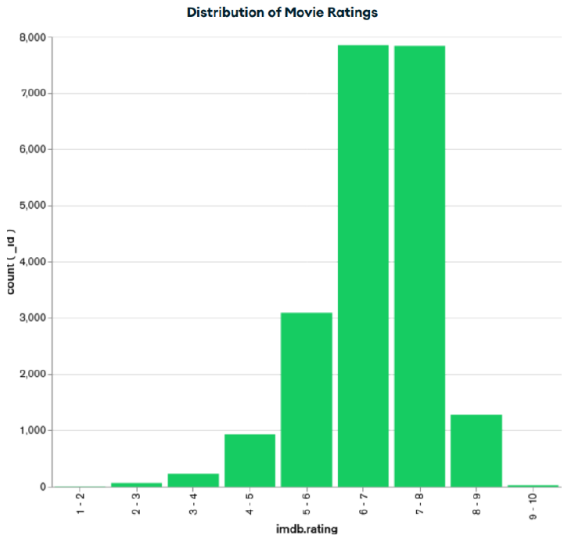
(ii) Top 10 Movies by IMDb Rating

- **Objective:** Determine the **top 10 highest-rated movies** based on IMDb ratings to highlight critically acclaimed films.



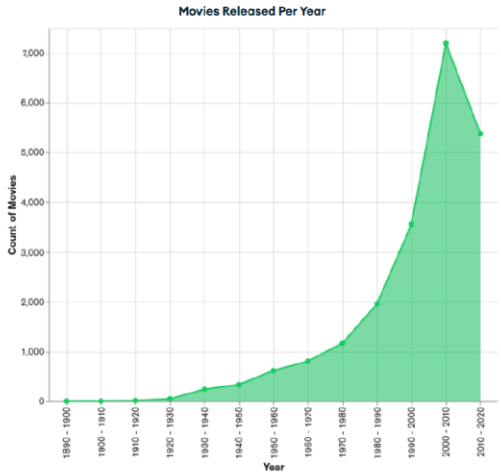
(iii) Distribution of Movie Ratings

- **Objective:** Visualize the distribution of IMDb ratings across movies to analyze overall quality trends and rating patterns.



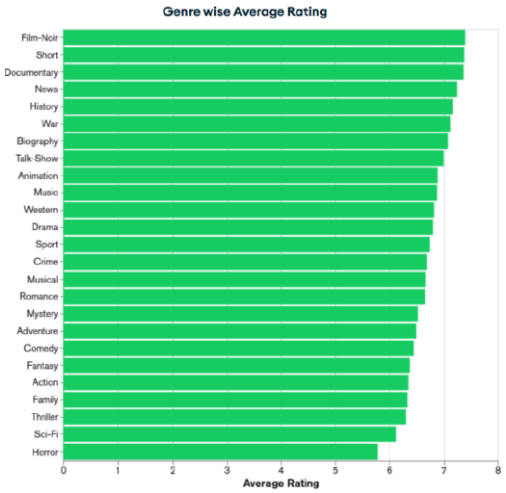
(iv) Movies Released Per Year

- **Objective:** Examine the trend in movie releases over the years to identify fluctuations in production volume and industry growth patterns.



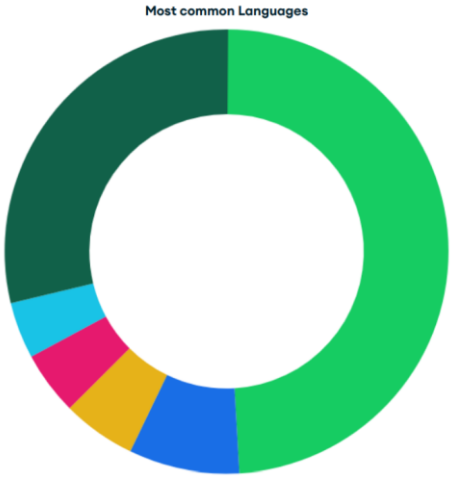
(v) Genre-wise Average Rating

- **Objective:** Analyze the **average IMDb rating** for each genre to determine which genres consistently receive higher or lower audience ratings.



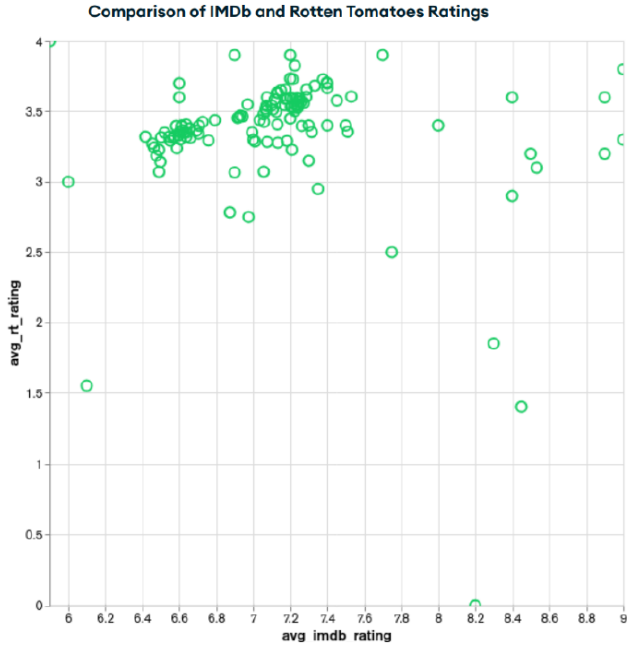
(vi) Most Common Languages

- **Objective:** Determine the most frequently used languages in movies to assess linguistic diversity within the dataset.



(vii) Comparison of IMDb and Rotten Tomatoes Ratings

- **Objective:** Analyze the relationship between **IMDb and Rotten Tomatoes ratings** to identify correlations and differences between audience and critic reviews.

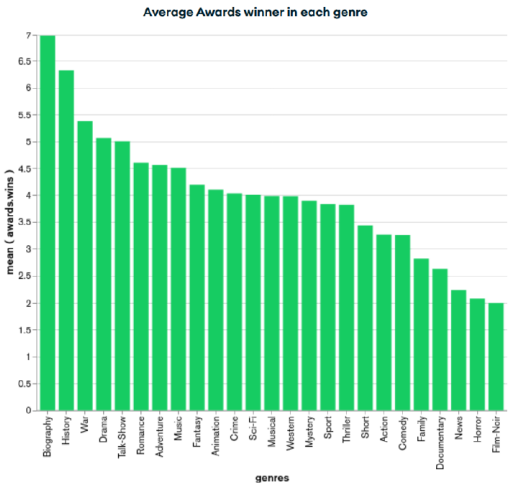


(viii) Genre vs. Country Popularity

- **Objective:** Examine how **genre preferences differ across countries** to uncover regional trends and audience preferences.

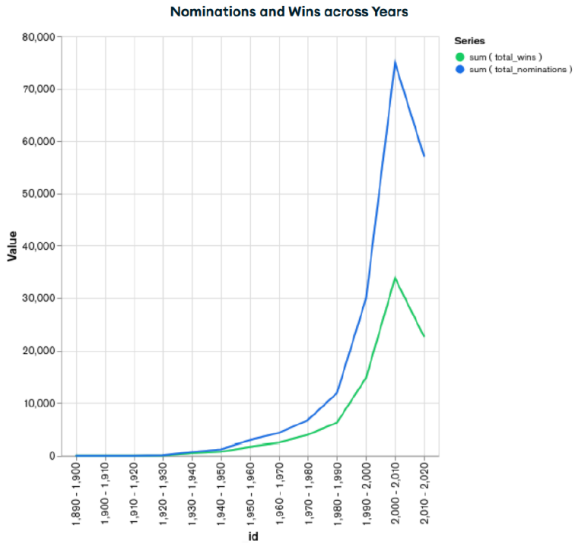
(ix) Average Awards Won in Each Genre

- **Objective:** Determine the **average number of awards** won by movies in each genre to identify which genres receive the most industry recognition.



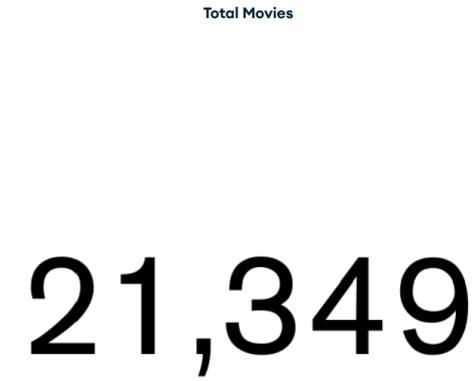
(x) Nominations and Wins Across Years

- **Objective:** Examine trends in **award nominations and wins** over time to understand shifting recognition patterns in the movie industry.



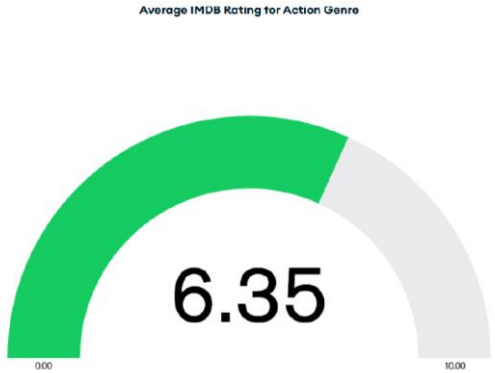
(xi) Total Movies (Number Card)

- **Objective:** Present the **total number of movies** in the dataset as a key overview metric.



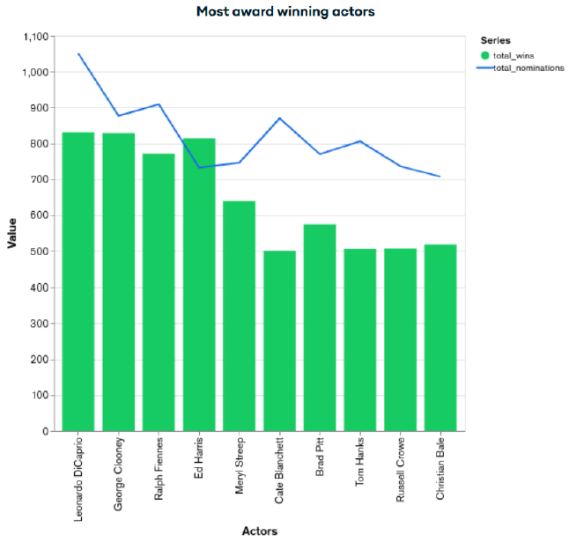
(xii) Average IMDb Rating for Action Genre

- **Objective:** Calculate the average IMDb rating for movies in the Action genre to evaluate the overall audience response to action films.



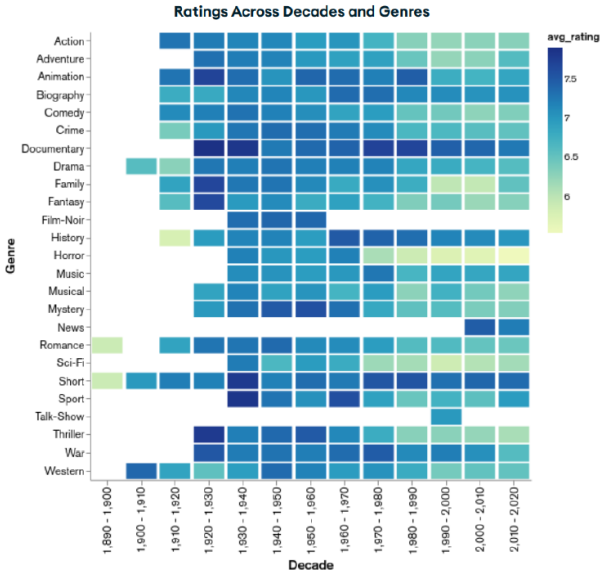
(xiii) Most Award-Winning Actors

- **Objective:** Identify the actors with the highest number of awards to spotlight the most successful and celebrated performers.



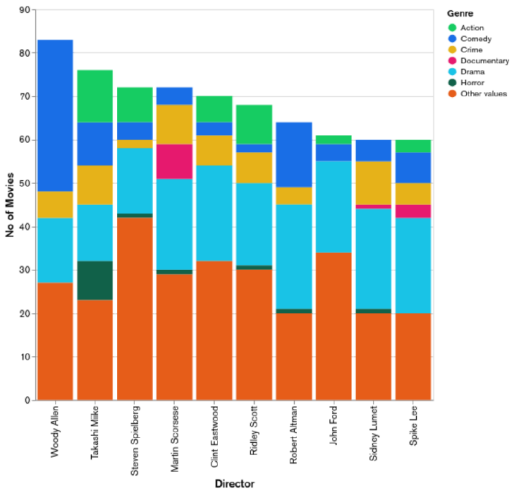
(xiv) Ratings Across Decades and Genres

- **Objective:** Illustrate the changes in movie ratings over the decades, categorized by genre, to uncover trends in audience reception over time.



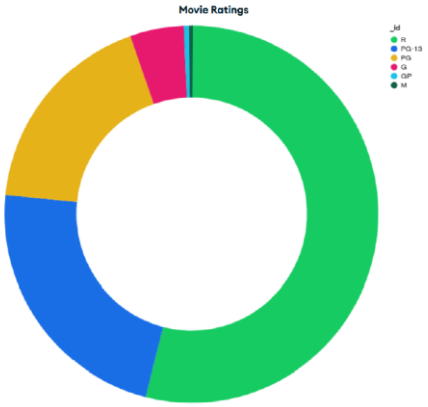
(xv) Directors with the Highest Number of Movies

- **Objective:** Analyze which directors have made the most movies and identify the genres they predominantly work in.



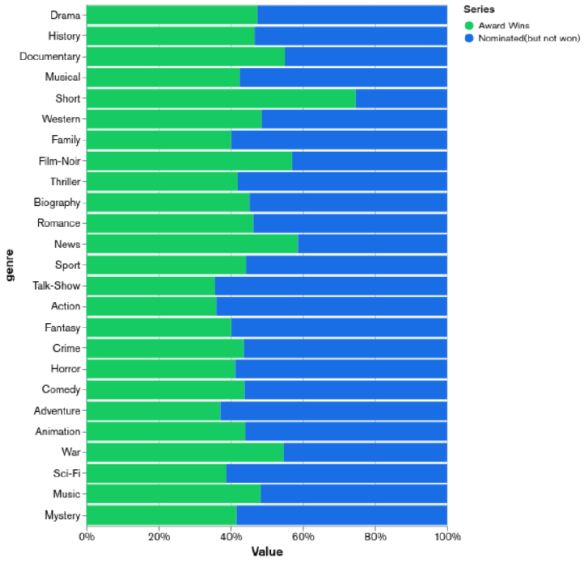
(xvi) Movie Ratings by Genre

- **Objective:** Summarize movie ratings by genre to provide a quick overview of the overall quality of films within each genre in the dataset.



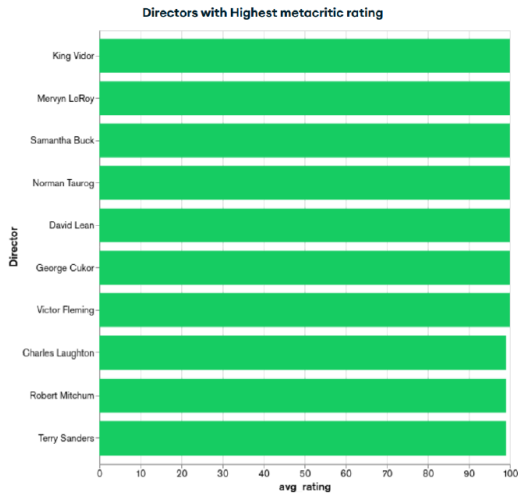
(xvii) Ratio of Award Wins to Nominations

- **Objective:** Analyze the percentage of movies in each genre that win awards versus those that are only nominated but do not win.



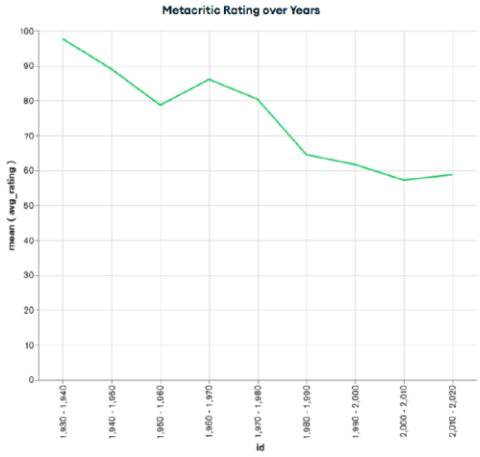
(xviii) Directors with Highest Metacritic Rating

- **Objective:** Identify directors whose films have earned the highest Metacritic ratings to highlight critically acclaimed filmmakers.



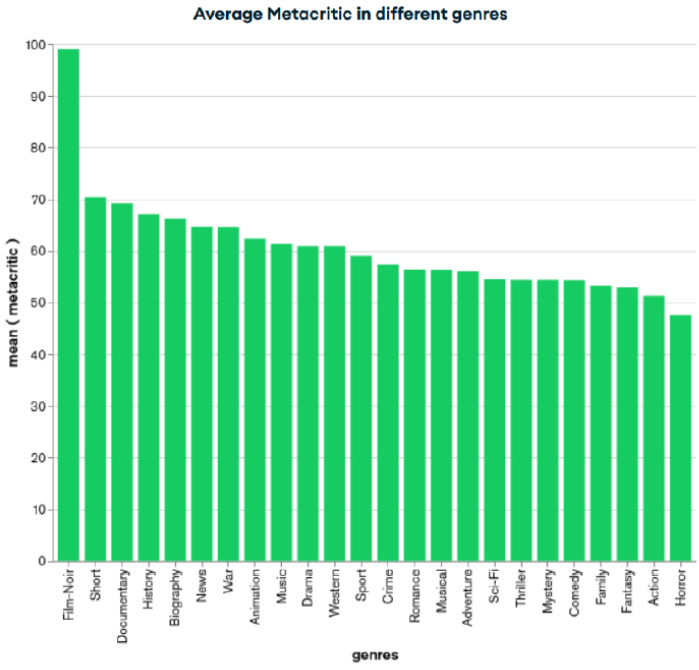
(xix) Metacritic Rating Over Years

- **Objective:** Analyze the evolution of Metacritic ratings over the years to identify trends in critical reception.



(xx) Average Metacritic Rating in Different Genres

- **Objective:** Compare the average Metacritic rating across various genres to determine which genres receive the most critical acclaim.



7. Observations and Findings

(i) Movie Distribution by Genre

- Drama, Action, and Comedy are the most produced genres, reflecting strong audience interest.
- Sci-Fi and Thriller films tend to receive higher ratings, suggesting a more niche but loyal following.

(ii) IMDb Ratings Analysis

- Most movies have IMDb ratings ranging between 6.0 and 8.0, with only a few scoring above 8.5.
- High-rated films are predominantly in the Drama and Thriller genres, highlighting a preference for well-crafted stories.
- Movies with IMDb ratings above 8.0 generally see longer audience engagement.

(iii) Yearly Trends in Movie Production

a. Movie production saw a significant rise after 2000, indicating the influence of digital transformation and the rise of streaming platforms.
b. The 1980s and 1990s saw steady production, but growth surged in the 21st century.

(iv) Language Distribution

a. English-language films dominate, with noticeable contributions from Spanish, French, and Hindi films.
b. Multilingual films are growing in popularity, reflecting the global expansion of the film industry.

(v) Runtime Analysis

a. The majority of movies have runtimes between 90 and 150 minutes, with very few exceeding 3 hours.
b. Short films and documentaries typically have much shorter runtimes.

(vi) Award-Winning Movies & Ratings Correlation

a. Higher IMDb ratings often correlate with winning or being nominated for awards.
b. A strong relationship exists between high critic scores on Rotten Tomatoes and award recognition.

(vii) Impact of Streaming Services & Future Outlook

a. The surge in movies post-2000 highlights the shift towards streaming platforms and digital content creation.
b. Emerging genres, such as Documentary and Biographical films, are likely to grow in popularity in the coming years.

(viii) Director Analysis

a. Directors with the most films often specialize in specific genres, indicating a preference or expertise in those areas.
b. Directors with consistently high Metacritic ratings tend to create films with strong critical acclaim, regardless of genre.

(xi) Critic and Viewer Ratings Comparison

a. A noticeable gap exists between critic ratings (Metacritic) and viewer ratings (IMDb), suggesting differing opinions on movie quality.
b. Drama and Thriller genres typically receive higher critic ratings, while Comedy and Action films have more varied responses.
c. Genres like Film Noir, with high average Metacritic ratings, may have inflated averages due to a smaller number of films.

(x) Award Analysis

a. The ratio of award wins to nominations differs across genres, with Drama and Thriller films having higher win percentages than Comedy or Action.

b. Not all highly rated films win awards, showing that critical success doesn't always translate to formal recognition.

(xii) Language and Regional Preferences

a. English-language films dominate, but there is a growing presence of Spanish, French, and Hindi films, reflecting the global reach of the industry.

b. Multilingual films are becoming more popular, indicating a rising audience interest in diverse cultural content.

8. Managerial Recommendations

(i) Strategic Content Planning

- Prioritize Drama, Thriller, and Sci-Fi genres, as they appeal to dedicated audiences and often receive higher ratings.
- Strike a balance between high-budget action films and critically acclaimed dramas to sustain both revenue and quality.

(ii) Targeted Marketing Campaigns

- Focus marketing efforts on award-winning and critically acclaimed films to attract niche audiences.
- For movies with average ratings but strong revenue potential, emphasize star power and visual appeal to draw in viewers.

(iii) Platform and Regional Strategy

- Invest in producing multilingual content to cater to diverse global audiences.
- Tailor regional distribution strategies on streaming platforms to optimize viewership.

(iv) Future Genre Focus

- Encourage the production of Documentary and Biographical films, as these genres are gaining traction in the streaming era.
- Use insights from the discrepancies between critic and viewer ratings to better target movies to general audiences.