Analysis of Movies Collection in MongoDB

Group 5

Priyadarshani Dash 055033

Divyank Harjani 055010

---

1. Project Information

This project involves performing an analysis of the Movies dataset using MongoDB. The dataset is stored in JSON format, which is well-suited for MongoDB's document-based architecture. MongoDB Compass is used to perform queries and CRUD operations and Atlas is used to make visualizations for analysis.

2. Data Description

The **movies** collection from the Mflix dataset contains **23,149 documents** and occupies around **1.2 GB** of storage in MongoDB Atlas. Each document represents a single movie and contains various attributes related to movie details, such as the title, genre, cast, and ratings. The data is structured in a semi-structured format (BSON) and consists of nested fields and arrays, making it suitable for flexible and complex queries. Each document in the collection represents a single movie and typically includes the following fields:

1. **_id** (ObjectId):
    - Unique identifier for each movie document.
    - Example: "5a934e000102030405000000"

2. **title** (String):
    - Name of the movie.
    - Example: "Inception"

3. **year** (Number):
    - Year of release of the movie.
    - Example: 2010

4. **genres** (Array of Strings):
    - Categories or genres that the movie belongs to.
    - Example: ["Action", "Sci-Fi", "Thriller"]

5. **cast** (Array of Strings):
    - List of actors featured in the movie.
    - Example: ["Leonardo DiCaprio", "Joseph Gordon-Levitt", "Elliot Page"]

6. **directors** (Array of Strings):
    - Names of the directors of the movie.
    - Example: ["Christopher Nolan"]

7. **writers** (Array of Strings):
    - Names of the writers or screenwriters of the movie.
    - Example: ["Christopher Nolan", "Jonathan Nolan"]

8. **languages** (Array of Strings):
    - Languages in which the movie is available.
    - Example: ["English", "Japanese", "French"]

9. **countries** (Array of Strings):
    - Countries where the movie was produced.
    - Example: ["USA", "UK"]

10. **released** (Date):
    - Release date of the movie.
    - Example: ISODate("2010-07-16T00:00:00Z")

11. **runtime** (Number):
    - Duration of the movie in minutes.
    - Example: 148

12. **plot** (String):
    - A brief summary of the movie's storyline.
    - Example: "A thief who steals corporate secrets through dream-sharing technology is given a chance to erase his criminal record."

13. **fullplot** (String):
    - Detailed plot summary.

14. **imdb** (Object):
    - Contains IMDb-specific information:
        - **rating** (Number): IMDb rating of the movie. Example: 8.8
        - **votes** (Number): Number of votes received. Example: 2000000
        - **id** (Number): IMDb ID of the movie. Example: 1375666

15. **tomatoes** (Object):
    - Contains movie ratings and reviews from Rotten Tomatoes:
        - **viewer** (Object): Viewer rating and number of reviews.
            - **rating** (Number): Viewer rating. Example: 4.2
            - **numReviews** (Number): Number of viewer reviews. Example: 5000
        - **critic** (Object): Critic rating and number of reviews.
        - **fresh** (Number): Number of fresh ratings.
        - **rotten** (Number): Number of rotten ratings.
        - **lastUpdated** (Date): Date when ratings were last updated.

16. **type** (String):
    - Type of media (usually "movie").
    - Example: "movie"

---

3. Project Objectives

- To create a structured dashboard summarizing key insights from the *movies* collection of the Mflix dataset.
- To explore the relationships between movie attributes, such as genre, rating, and release year, through data visualization.
- To implement efficient CRUD operations on the dataset using MongoDB commands.
- Togenerate meaningful reports that assistin data-driven decision-making and movie trend analysis.

---

4. Queries (Compass)

Create

I. Insert a New Movie

```
> use sample_mflix
< switched to db sample_mflix
> db["movies"].insertOne({
    title: "Temp Inception",
    year: 2025,
    genres: ["Action", "Sci-Fi"],
    cast: ["Leonardo DiCaprio", "Joseph Gordon-Levitt"],
    imdb: { rating: 9.0, votes: 1200000 },
    isTemporary: true  // Marking as temporary
  })
< {
    acknowledged: true,
    insertedId: ObjectId('67d69d7b67f202db884fb472')
  }
```

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
```

II. Find Movies with an IMDb Rating Above 8.5

```
> db.movies.find({ "imdb.rating": { $gt: 8.5 } }, { title: 1, "imdb.rating": 1 })
< {
    _id: ObjectId('573a1391f29313caabcd9600'),
    title: 'City Lights',
    imdb: {
      rating: 8.6
    }
  }
  {
    _id: ObjectId('573a1392f29313caabcdae3d'),
    imdb: {
      rating: 8.6
    },
    title: 'Modern Times'
  }
  {
    _id: ObjectId('573a1393f29313caabcdc810'),
    title: 'Casablanca',
    imdb: {
      rating: 8.6
    }
  }
  {
    _id: ObjectId('573a1393f29313caabcdd7d8'),
    title: "It's a Wonderful Life",
    imdb: {
      rating: 8.6
    }
```

III. Count Movies in the "Action" Genre

```
> db.movies.countDocuments({ genres: "Action" })
< 2383
Atlas atlas-o9pchr-shard-0 [primary] sample_mflix>
```

Update

I. Update IMDb Rating of a Temporary Movie

```
> db.movies.updateOne(
    { title: "Temp Inception", isTemporary: true },
    { $set: { "imdb.rating": 9.3 } }
  )
< {
    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
  }
```

III. Update Multiple Temporary Movies' Year

```
> db.movies.updateMany(
    { isTemporary: true },
    { $set: { year: 2025 } }
  )
< {
    acknowledged: true,
    insertedId: null,
    matchedCount: 3,
    modifiedCount: 2,
    upsertedCount: 0
  }
```

Delete

I. Delete a Specific Temporary Movie

```
> db.movies.deleteOne({ title: "Temp Inception", isTemporary: true })
< {
    acknowledged: true,
    deletedCount: 1
  }
```

II. Delete All Temporary Movies

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

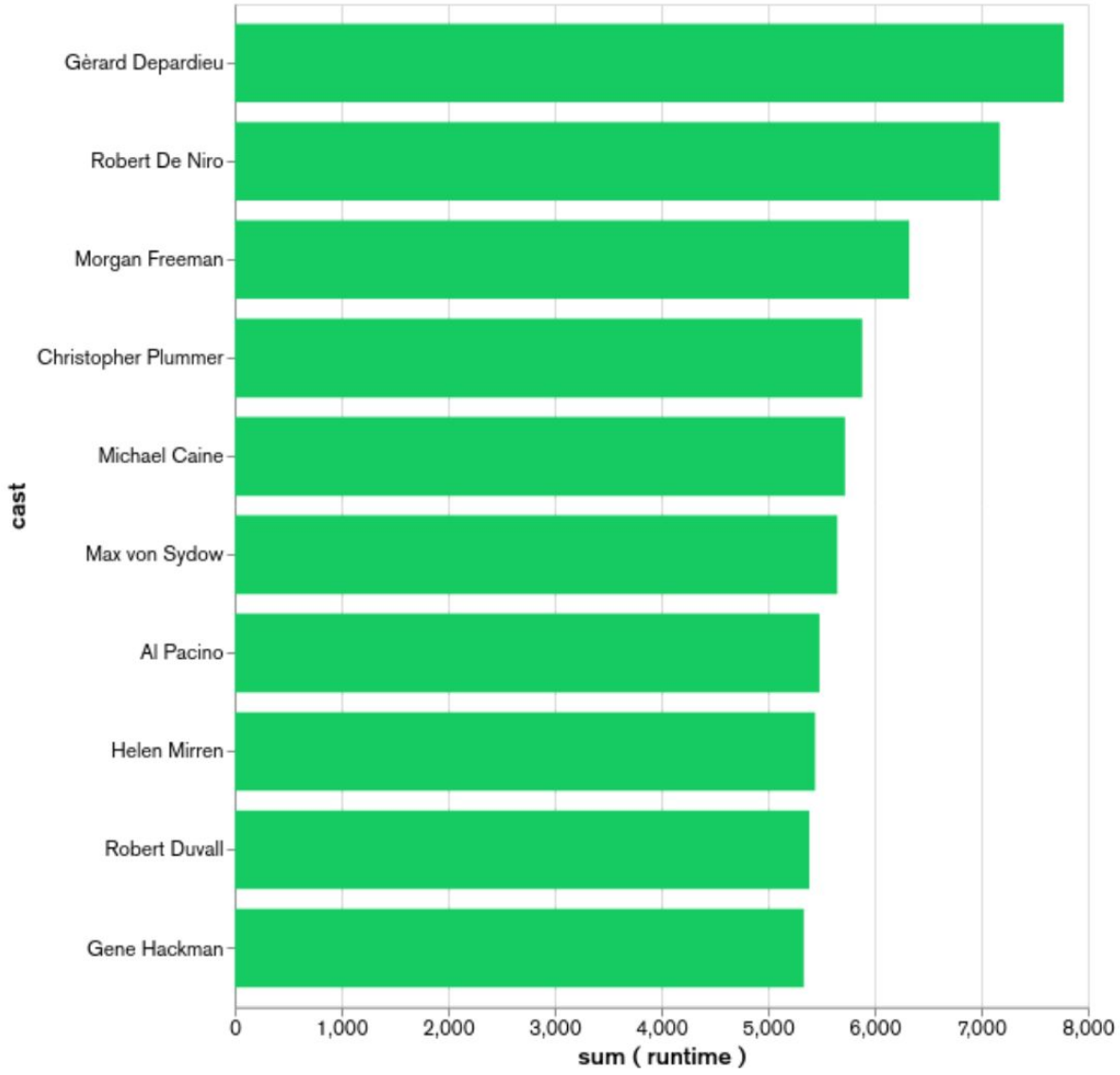
---

5. Problem Statement

The movie industry struggles with analyzing **large, unstructured data** to predict trends and optimize content. Traditional methods fail to provide **clear insights into genre popularity, ratings, and production trends**, impacting data-driven decisions.

6. Dashboard

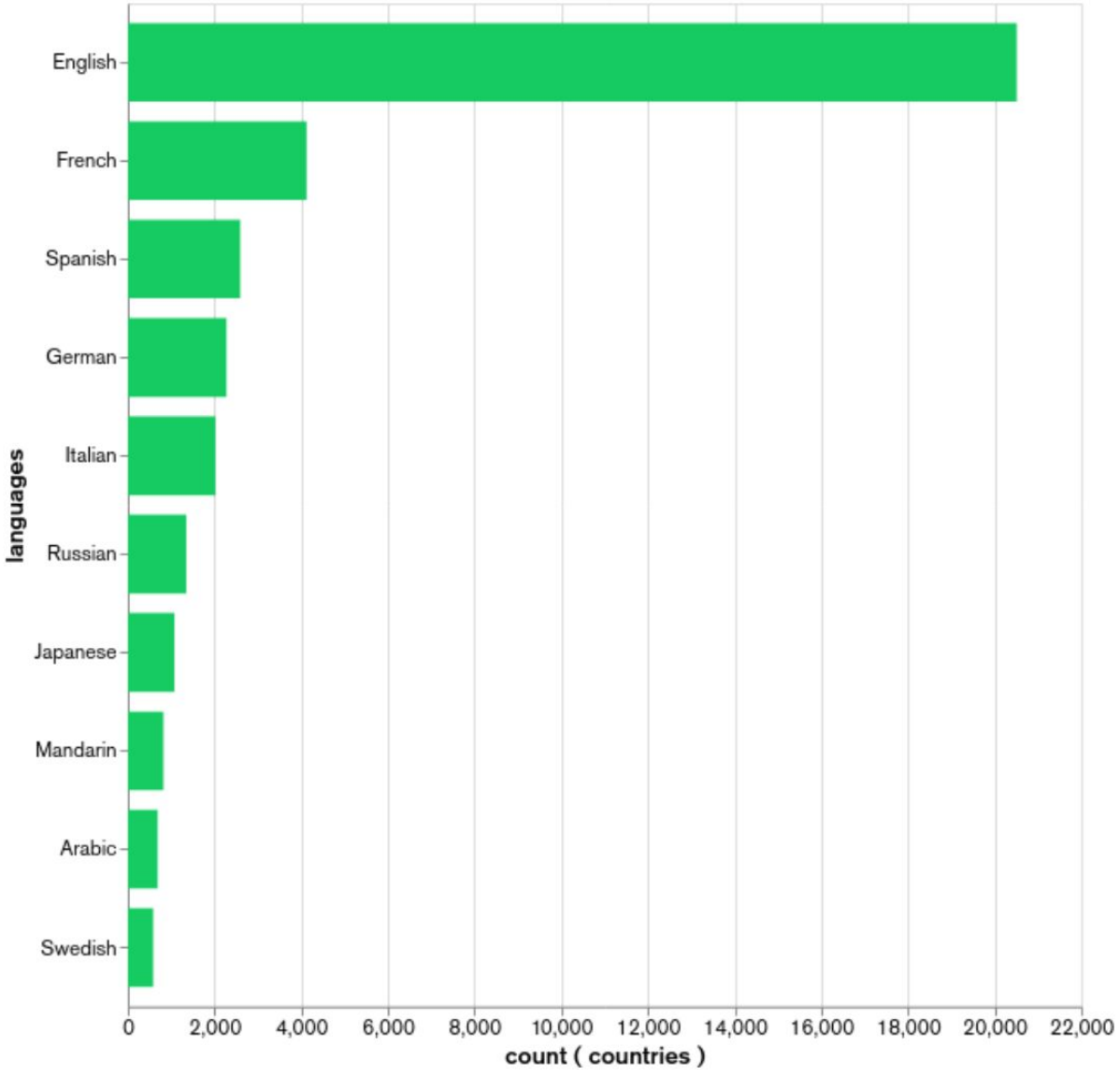
The MongoDB Atlas Dashboard is created to visualize key metrics from the Movies dataset.



II. Top 10 Movies by IMDb Rating

Problem Statement:

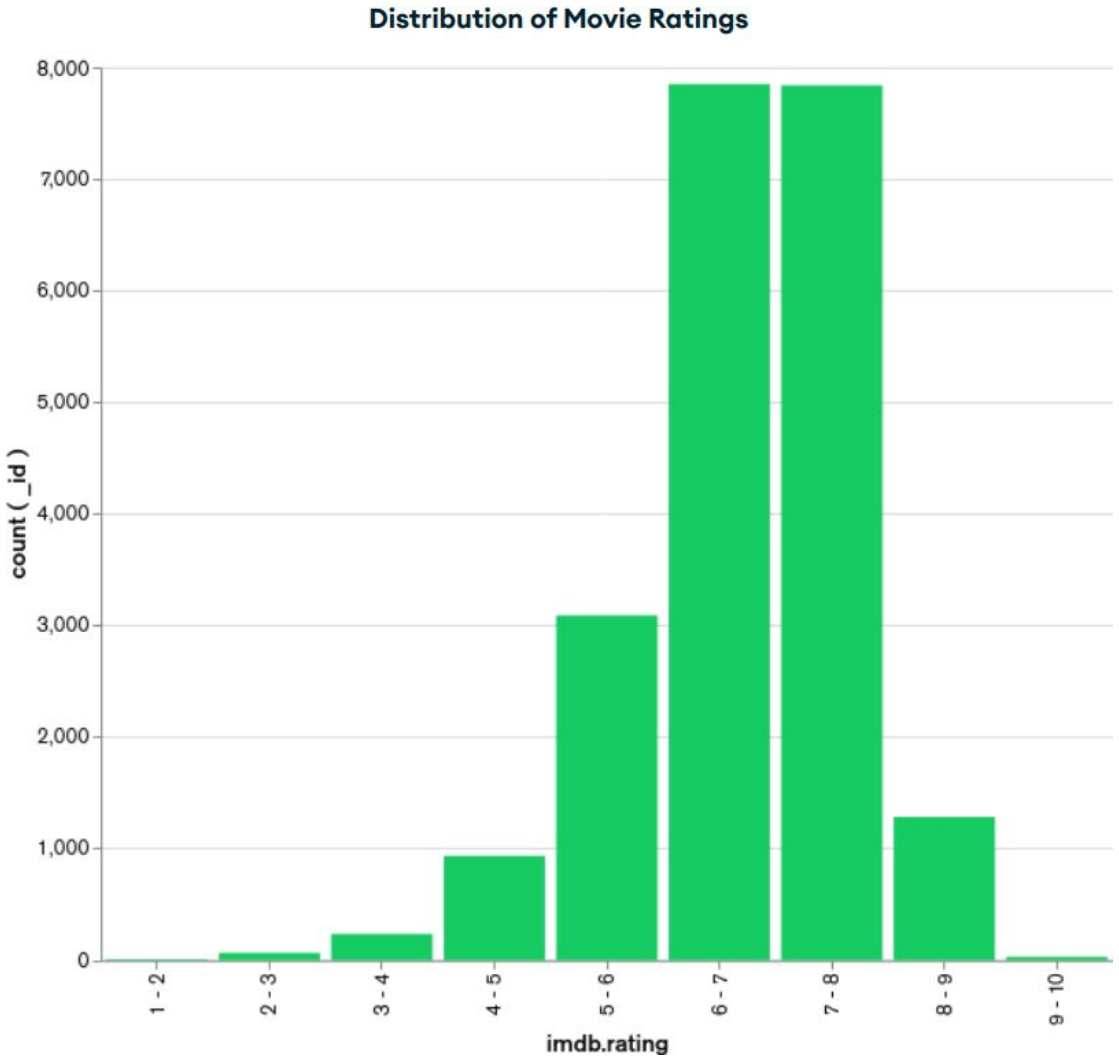
Identify the top 10 highest-rated movies based on IMDb ratings to showcase critically acclaimed films.



III. Distribution of Movie Ratings

Problem Statement:

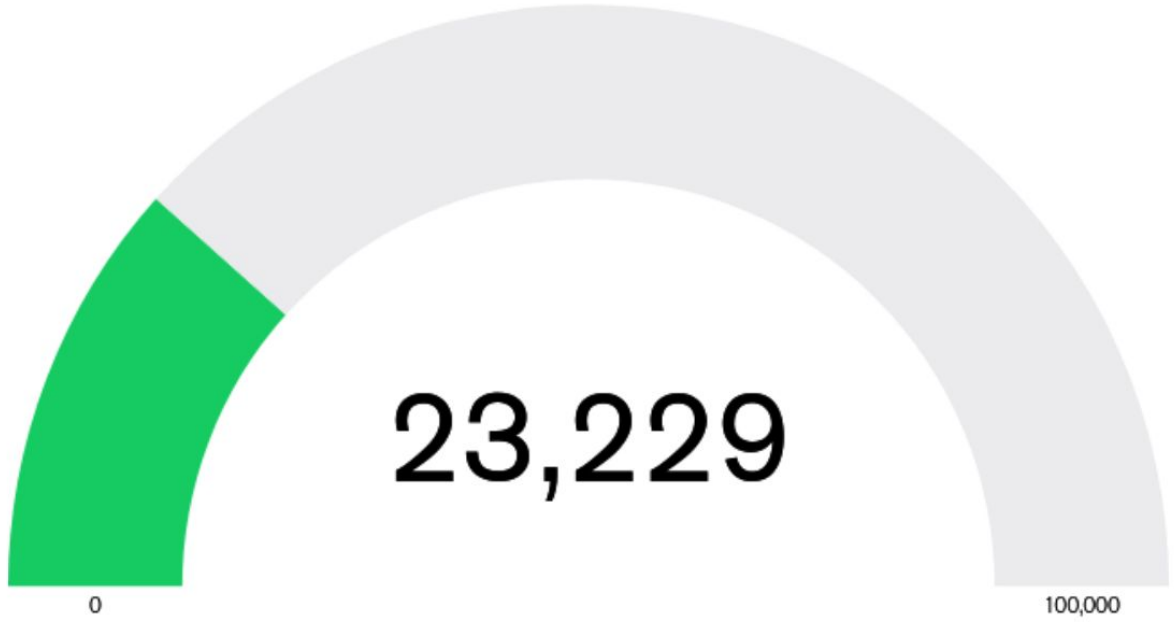
Visualize how IMDb ratings are distributed among movies to understand the general quality and rating patterns.



IV. Movies Released Per Year

Problem Statement:

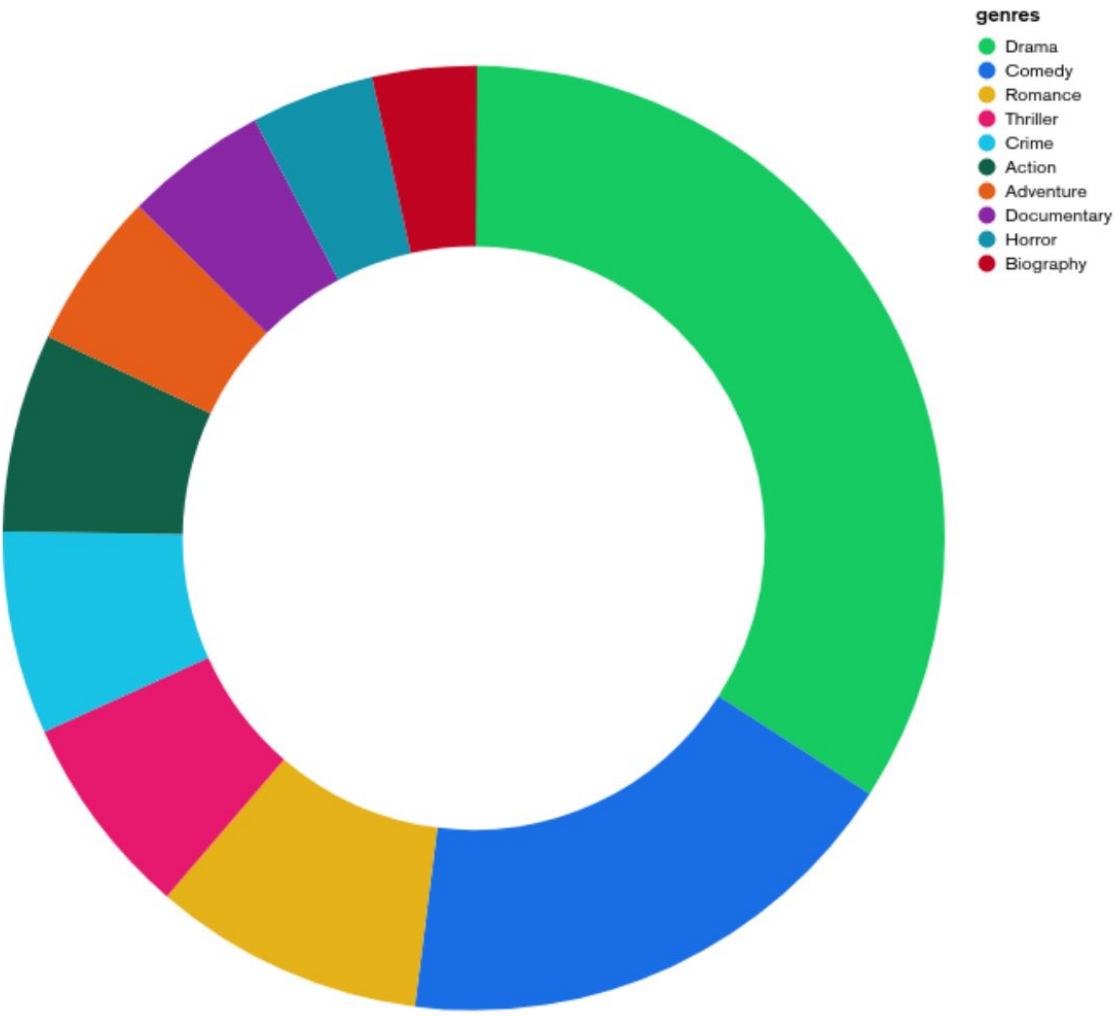
Analyze the trend in the number of movies released per year to observe changes in movie production volume over time.



V. Genre-wise Average Rating

Problem Statement:

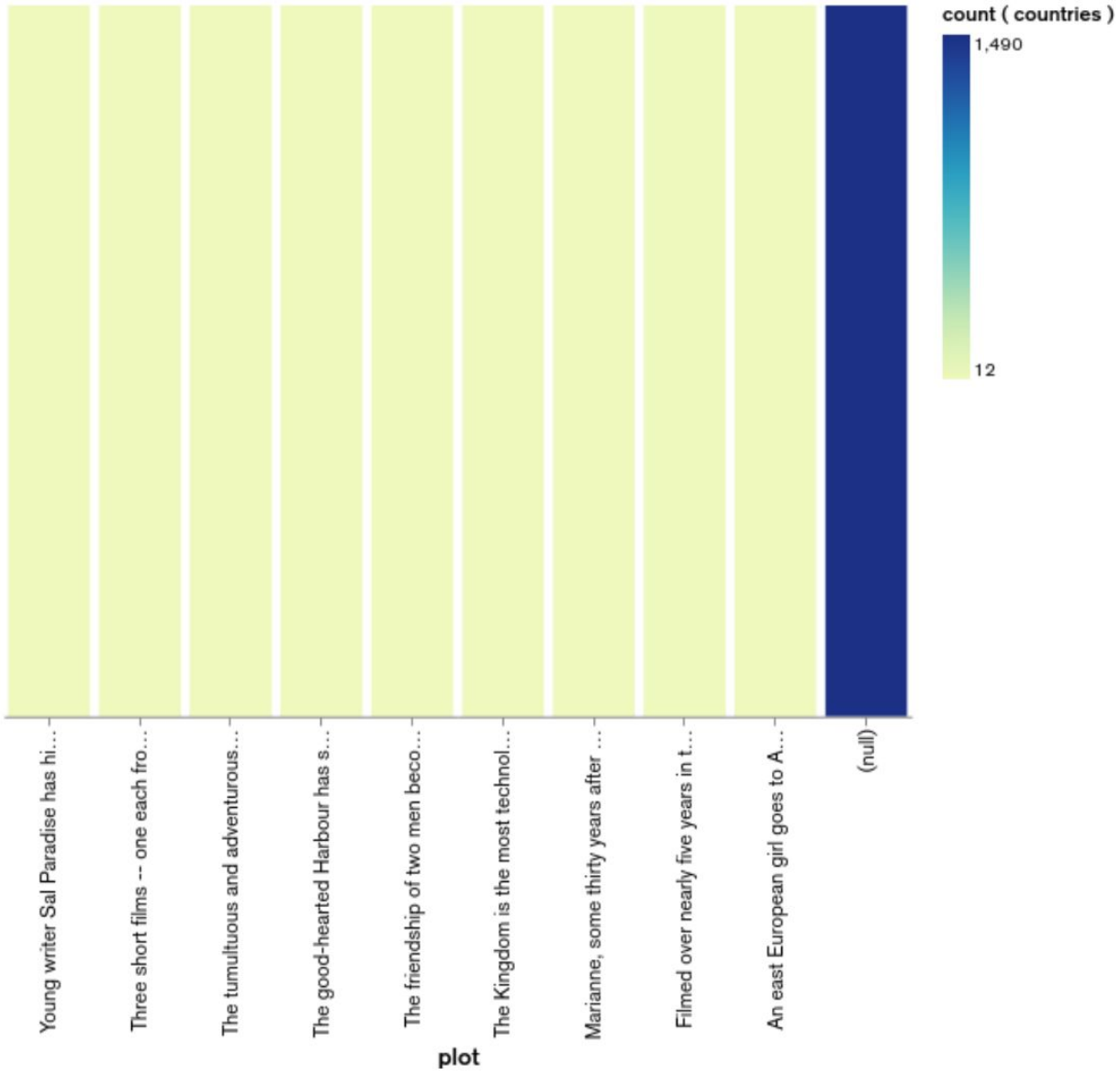
Examine the average IMDb rating for each genre to identify genres that consistently receive high or low ratings.



VI. Most Common Languages

Problem Statement:

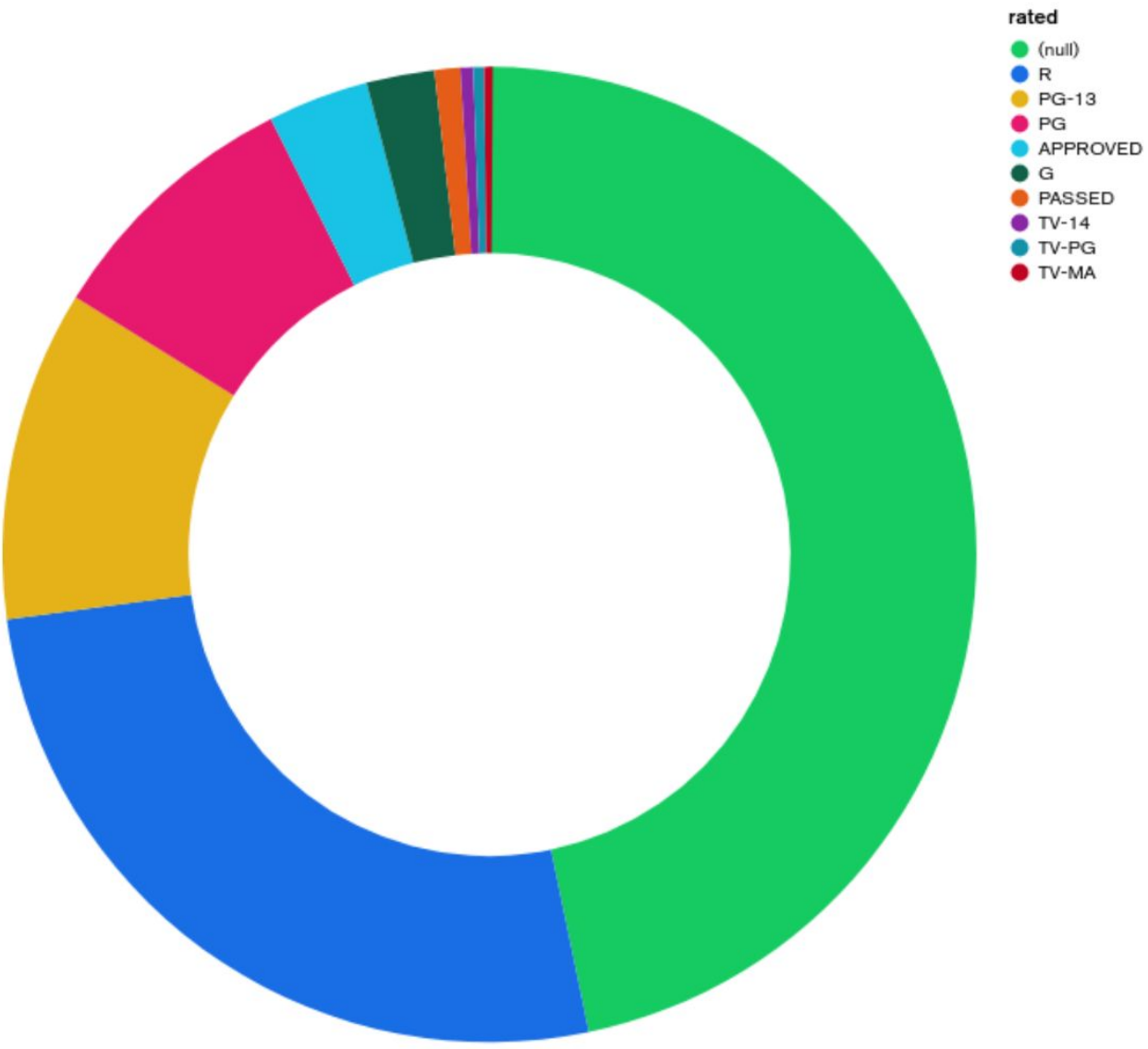
Identify the most frequently used languages in movies to understand linguistic diversity in the dataset.



VII. Comparison of IMDb and Rotten Tomatoes Ratings

Problem Statement:

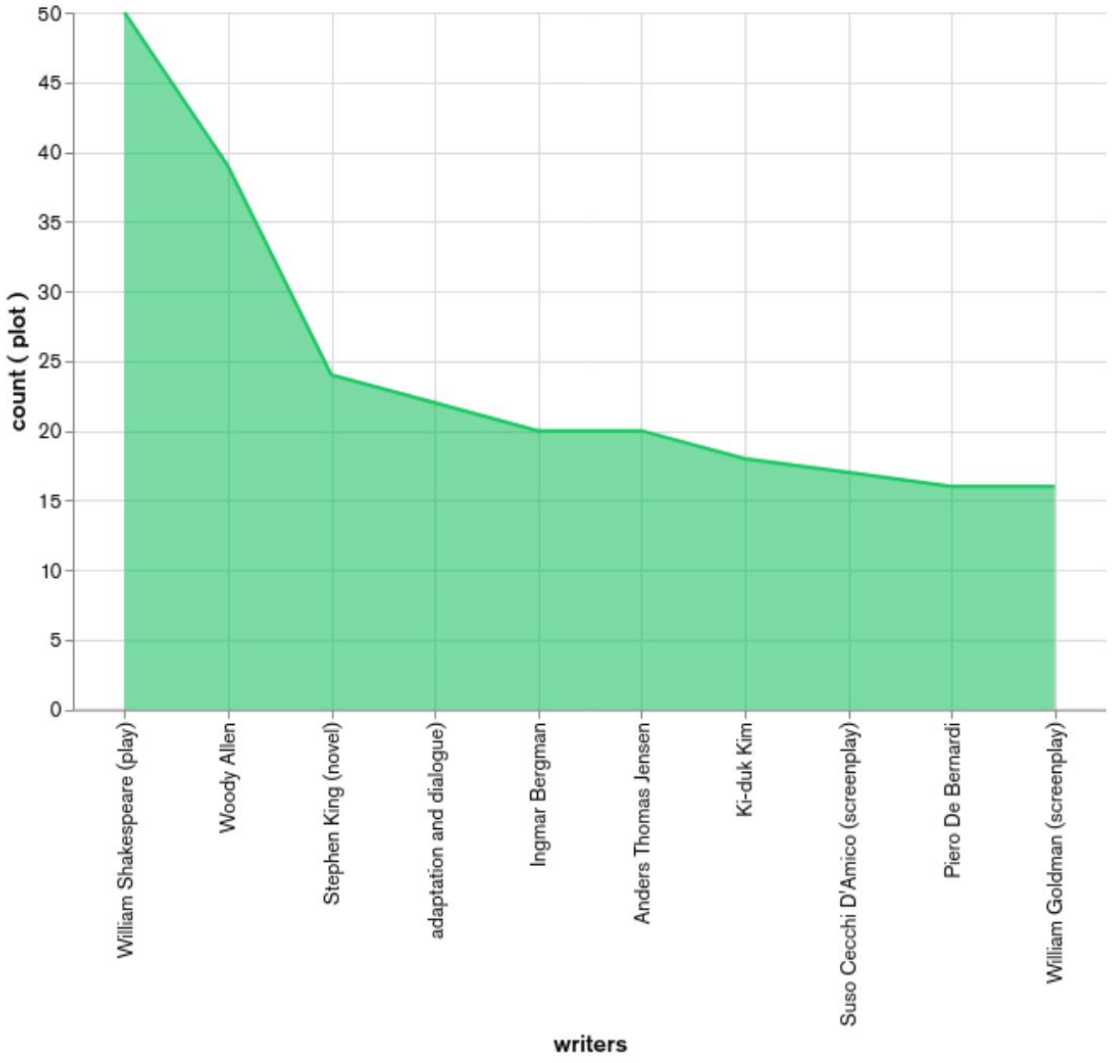
Compare IMDb and Rotten Tomatoes ratings to observe correlations and differences between audience and critic opinions.



VIII. Genre vs. Country Popularity

Problem Statement:

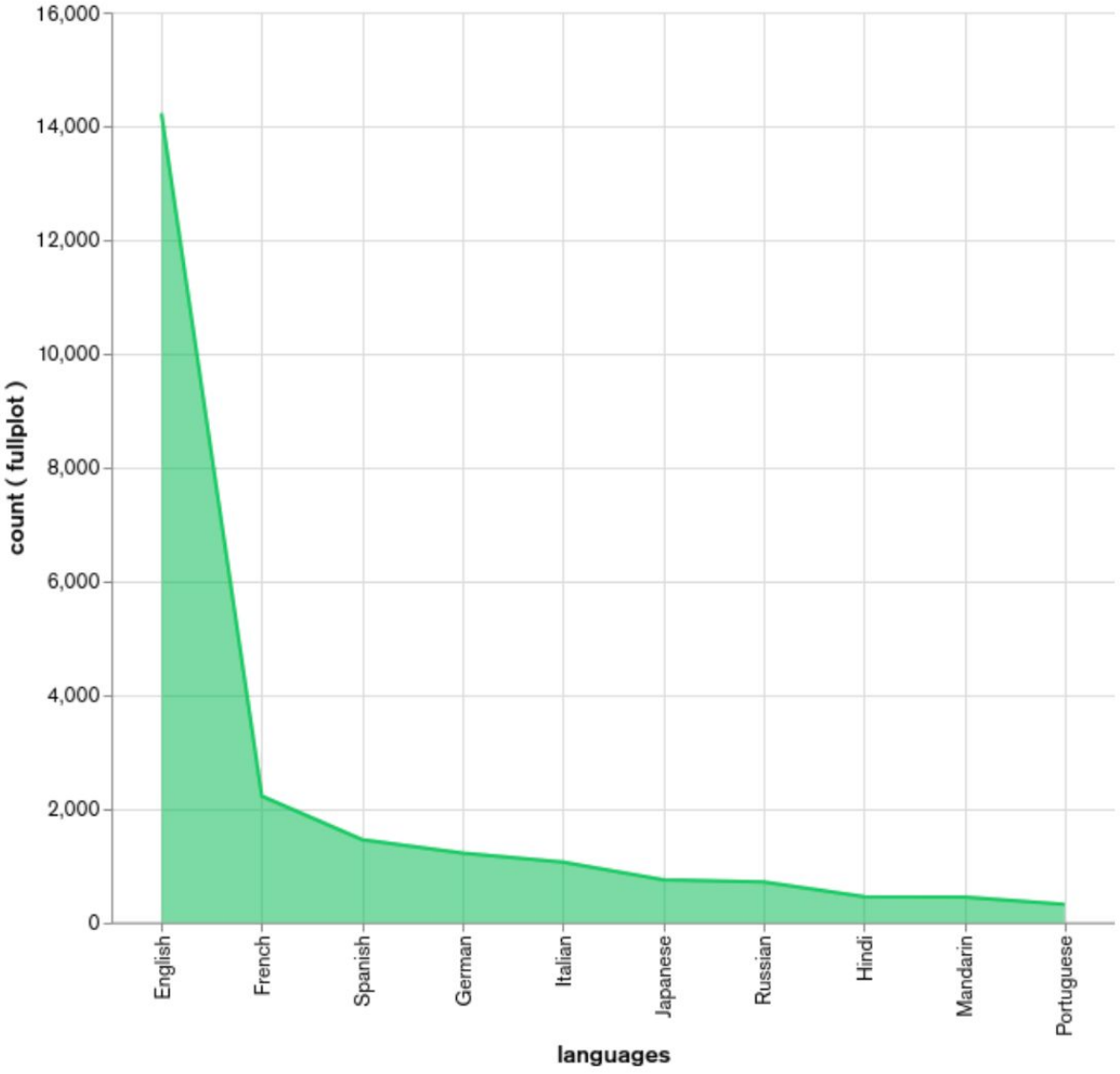
Analyze how genre preferences vary across different countries to identify regional preferences and trends.



IX. Average Awards Won in Each Genre

Problem Statement:

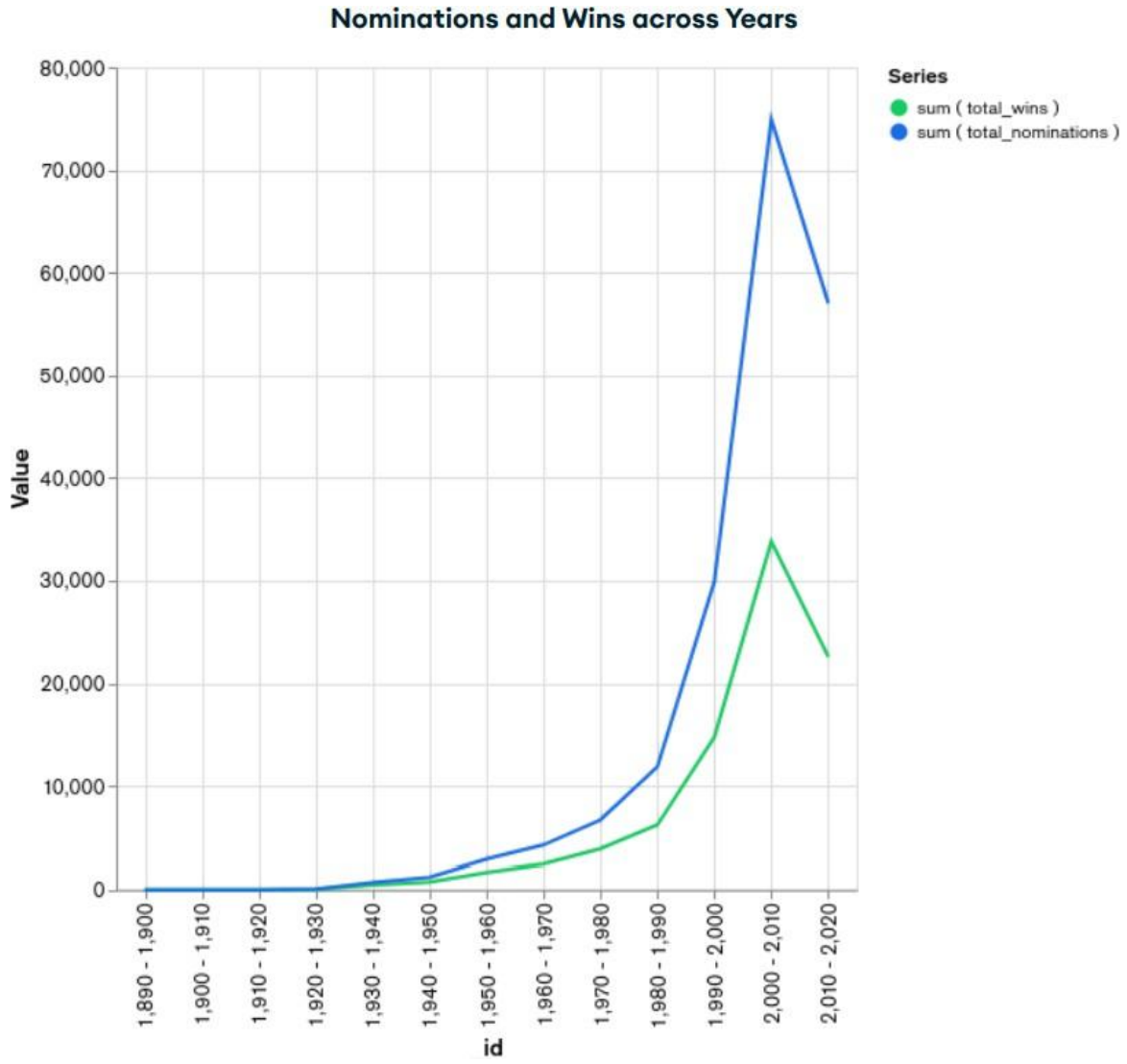
Calculate the average number of awards won by movies in each genre to identify which genres tend to receive more recognition.



X. Nominations and Wins Across Years

Problem Statement:

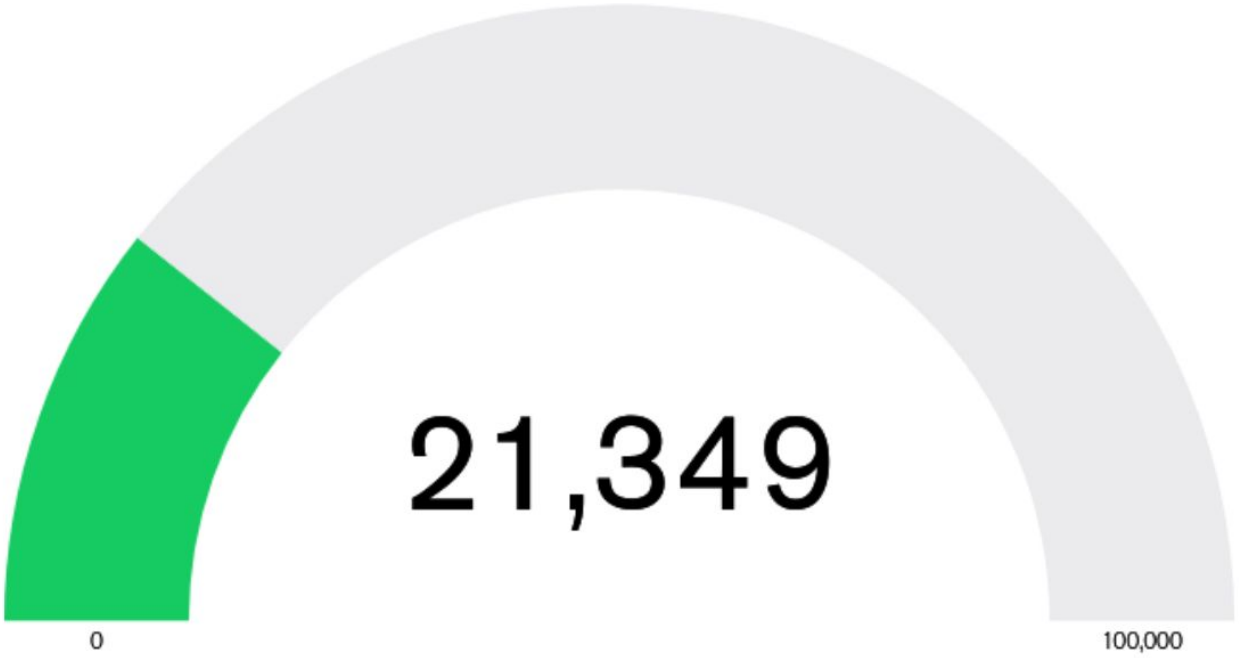
Analyze the trend of award nominations and wins over the years to observe the changing recognition patterns in the movie industry.



XI. Total Movies (Number Card)

Problem Statement:

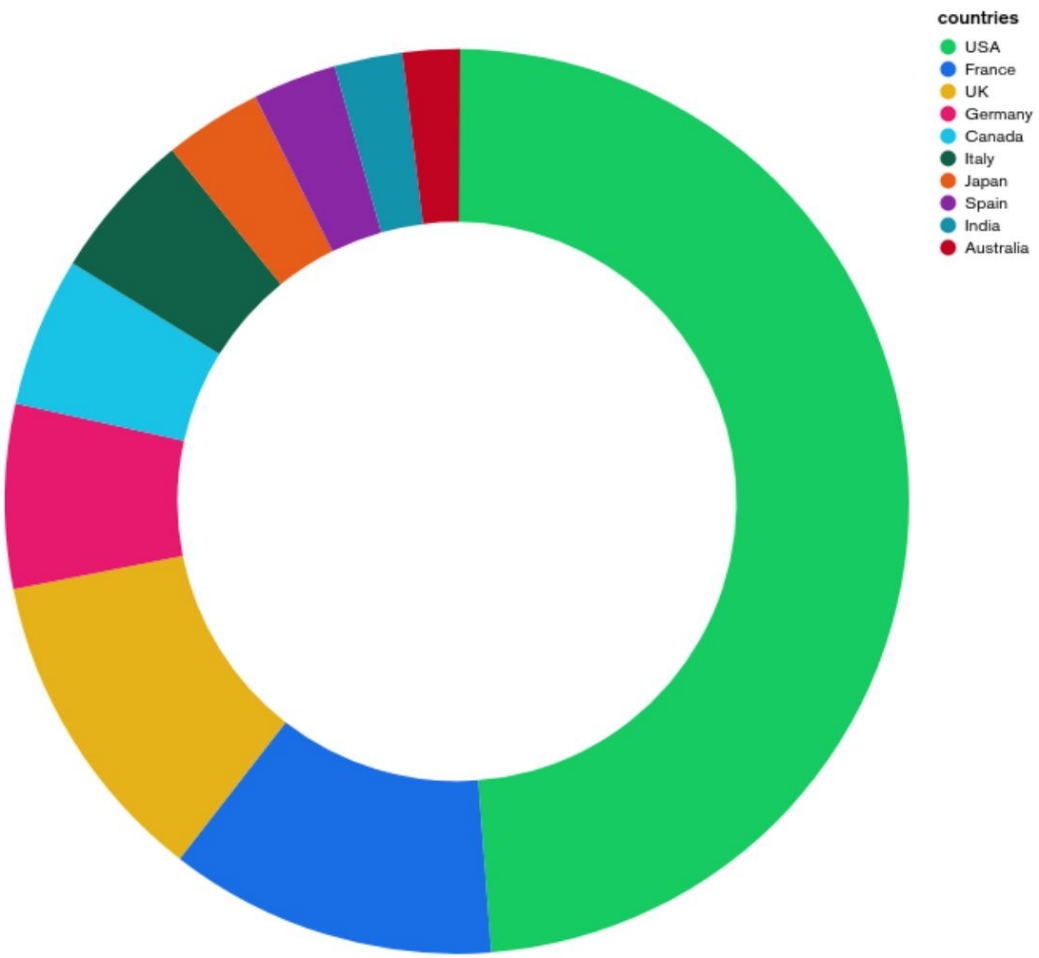
Display the total number of movies in the dataset as an overview metric.



XII. Average IMDb Rating for Action Genre

Problem Statement:

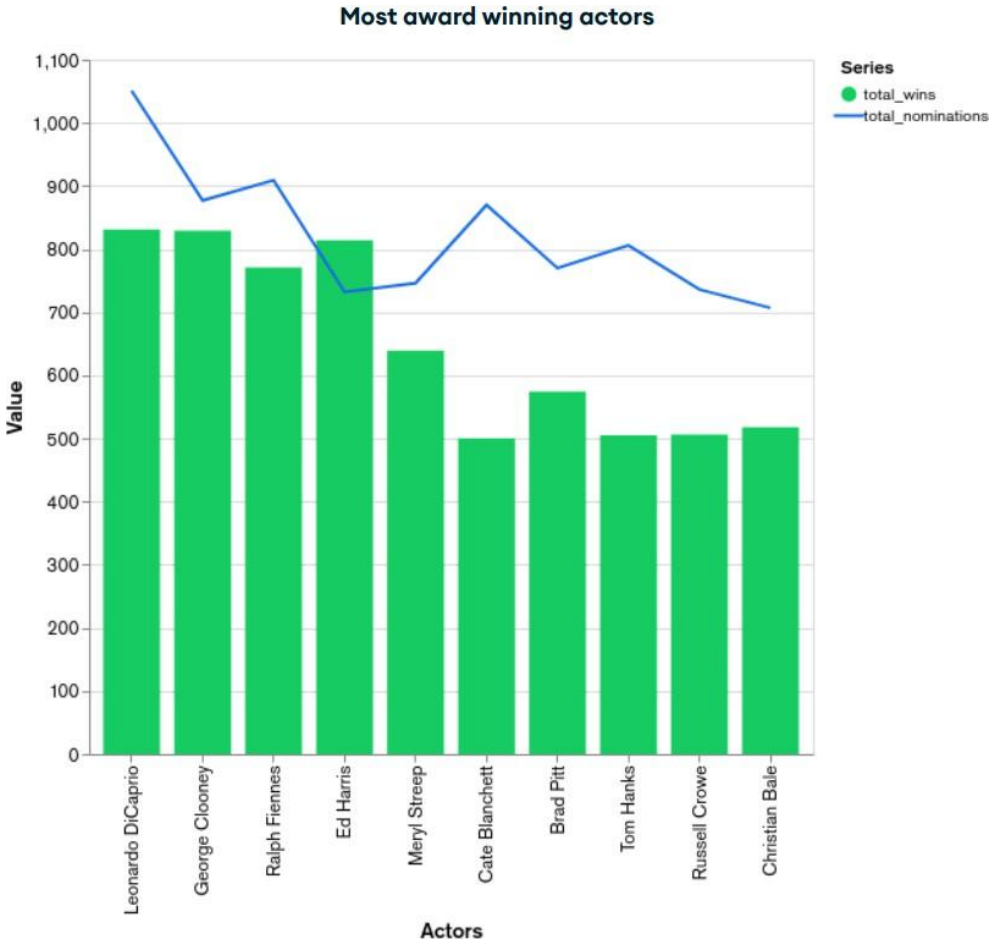
Display the average IMDb rating for movies in the Action genre to assess the general audience reception for action films.



XIII. Most Award-Winning Actors

Problem Statement:

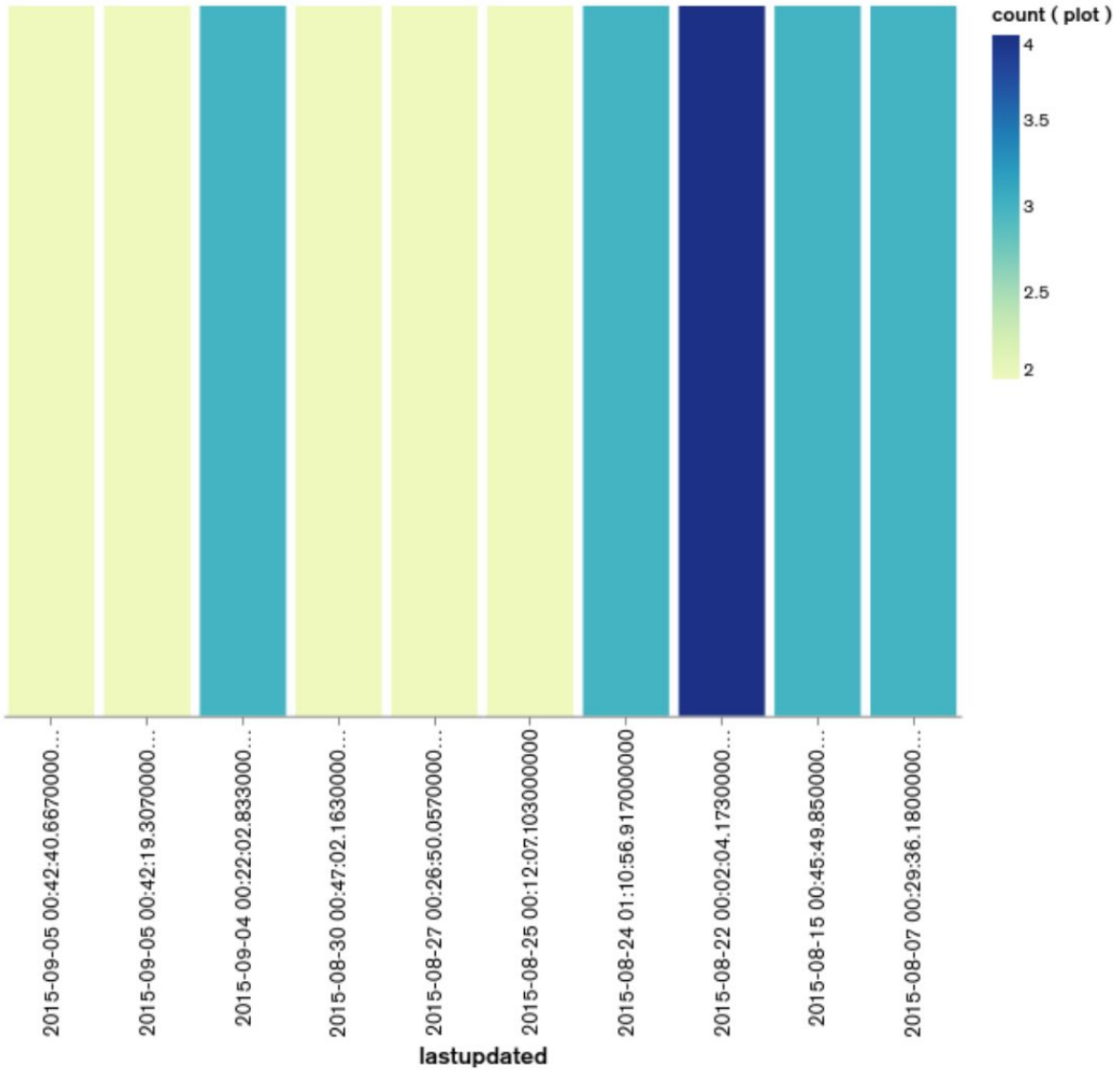
Identify the actors with the most awards to highlight successful and acclaimed performers.



XIV. Ratings Across Decades and Genres

Problem Statement:

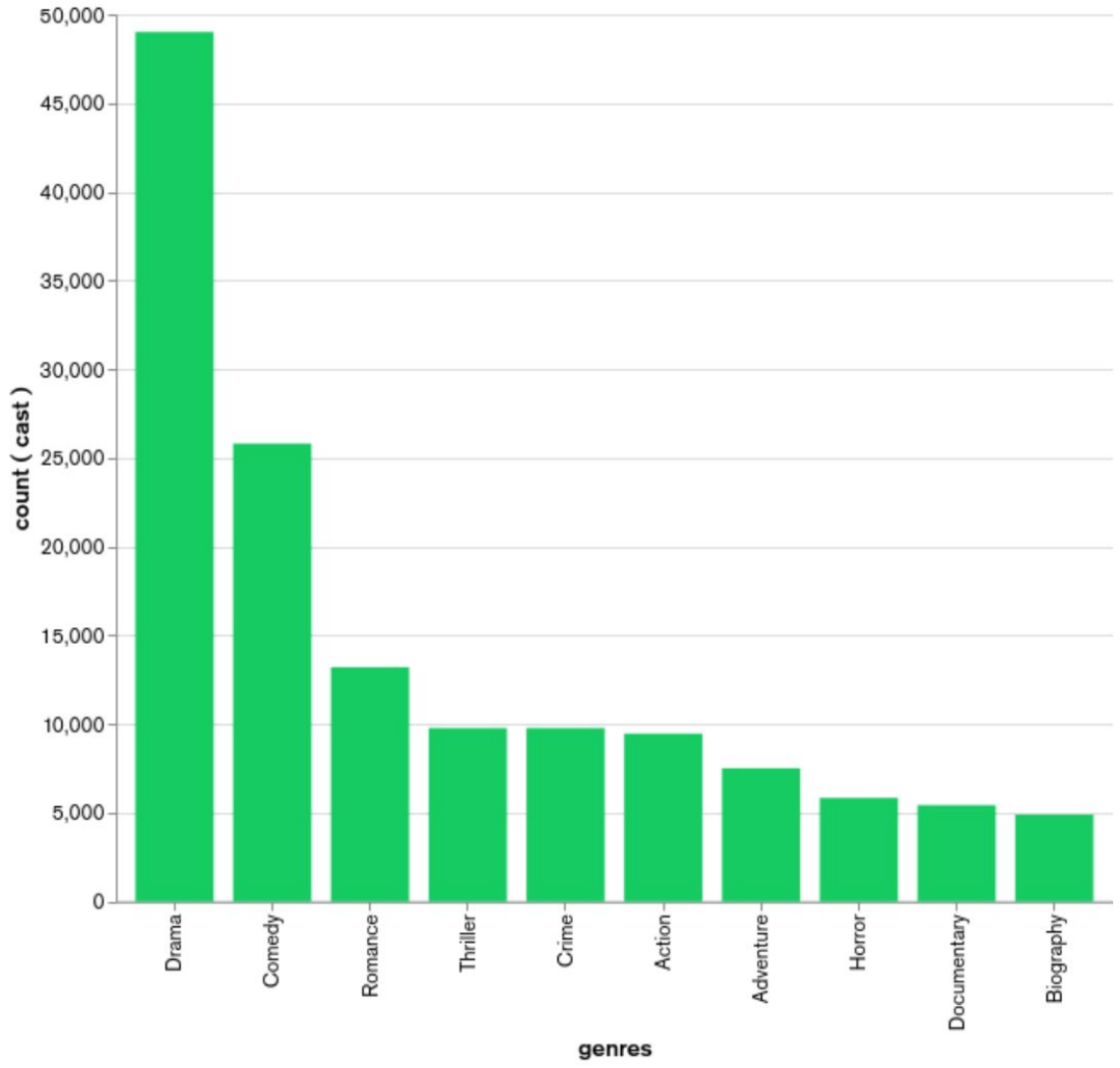
Visualize how movie ratings have changed over the decades, categorized by genre, to identify trends in audience reception over time.



XV. Directors with highest no of movies

Problem Statement:

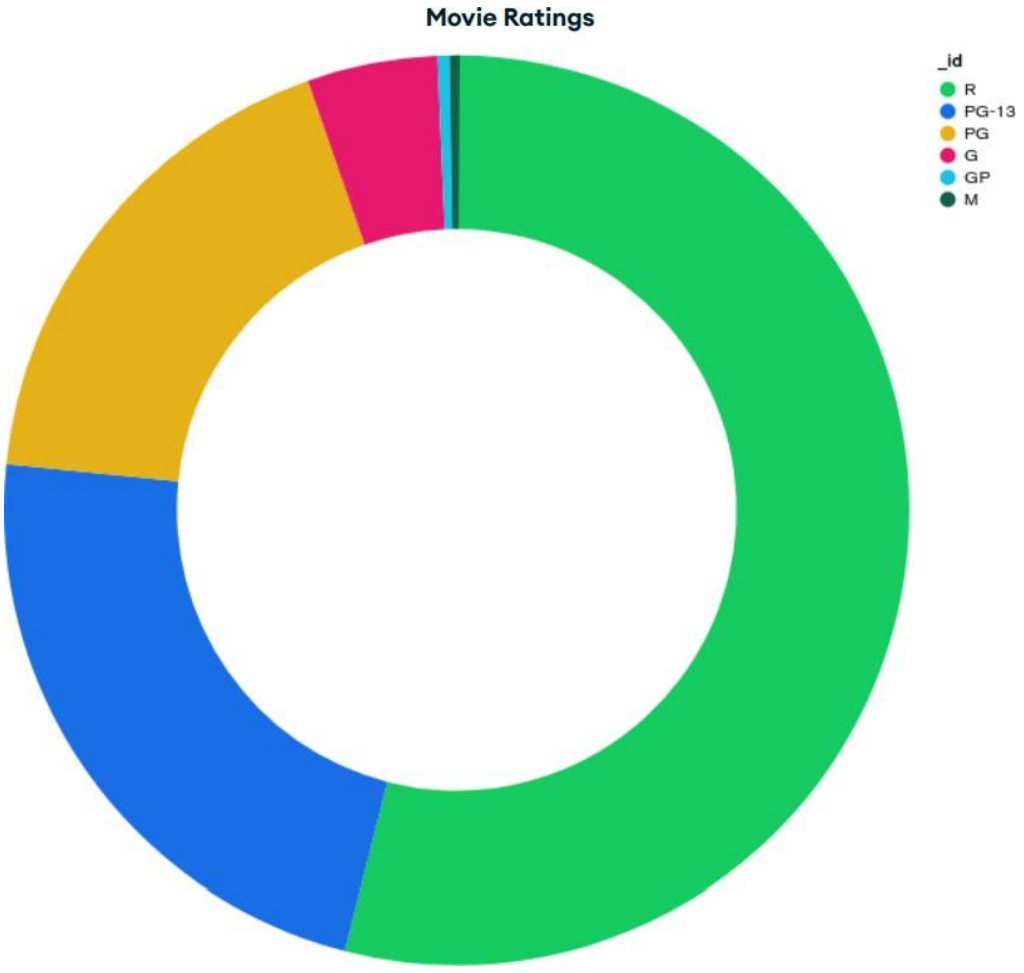
Analyze which directors have highest no of movies and what genre movies do they make



XVI. Movie Ratings by genre

Problem Statement:

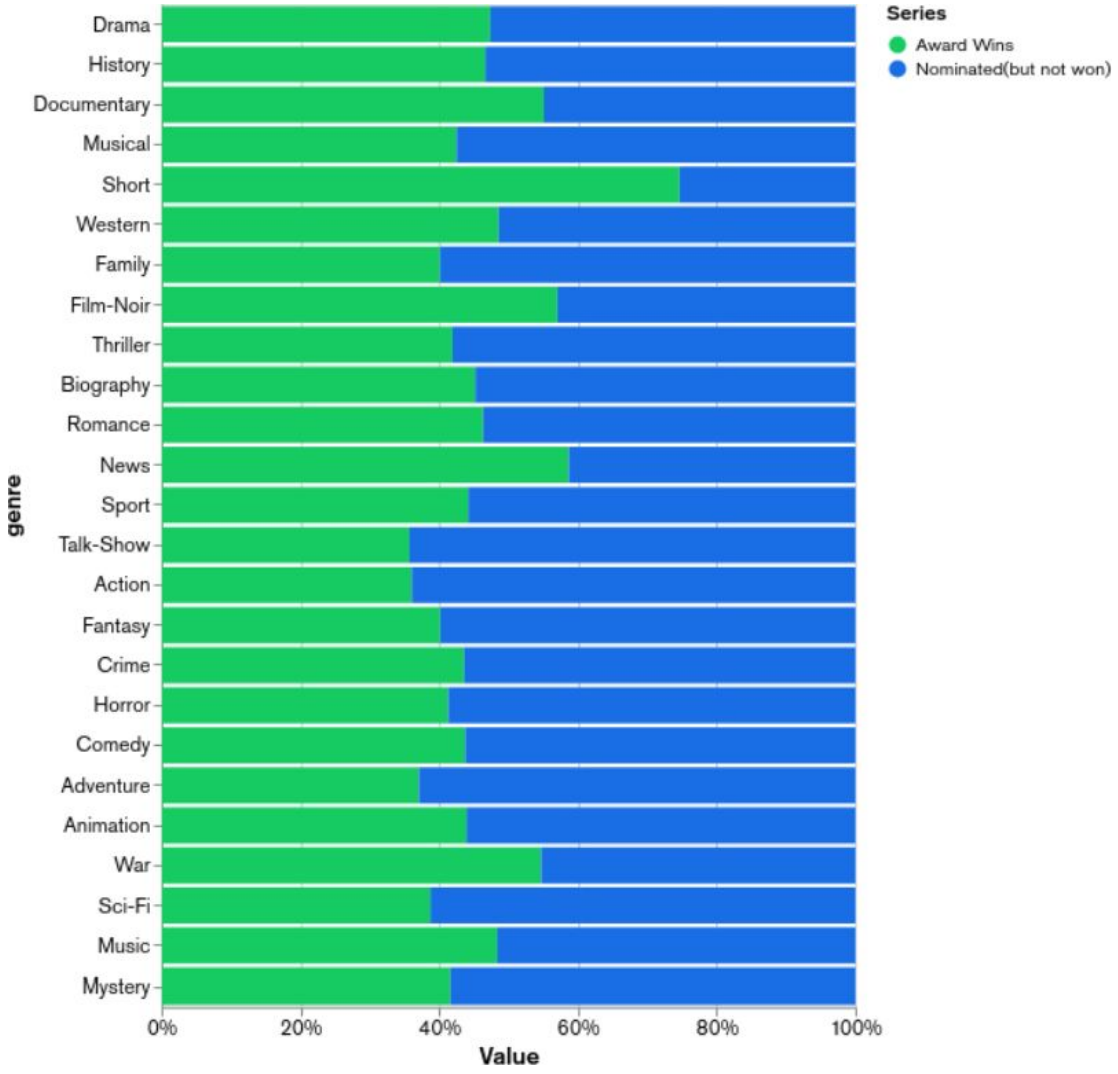
Provide a summary of movie ratings to give a quick snapshot of the overall quality of movies in the dataset.



XVII. Ratio of award wins to nominations

Problem Statement:

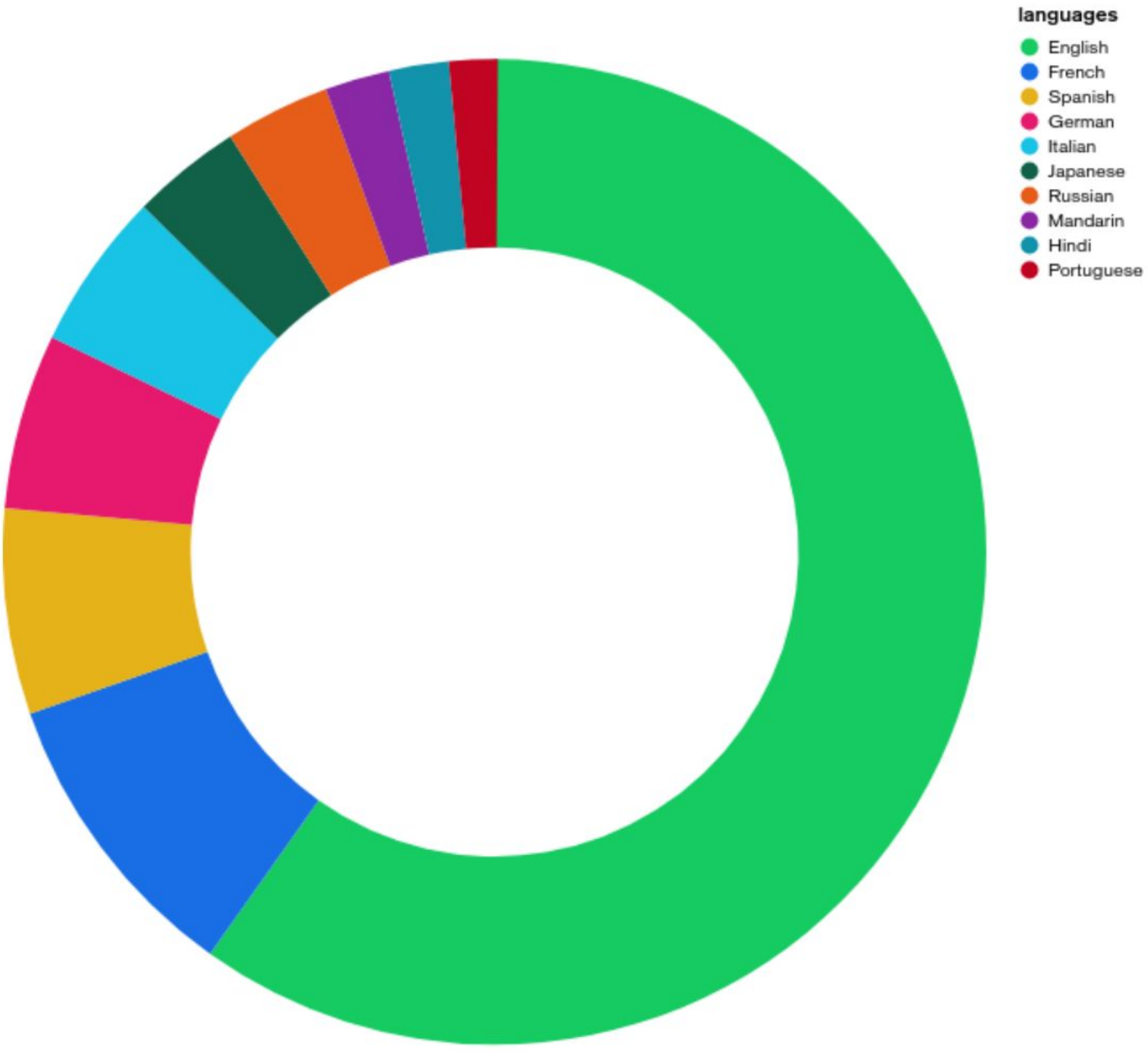
Analyze in each genre how much % of movies win awards and how many only get nominated but don't win.



XVIII. Directors with Highest Metacritic Rating

Problem Statement:

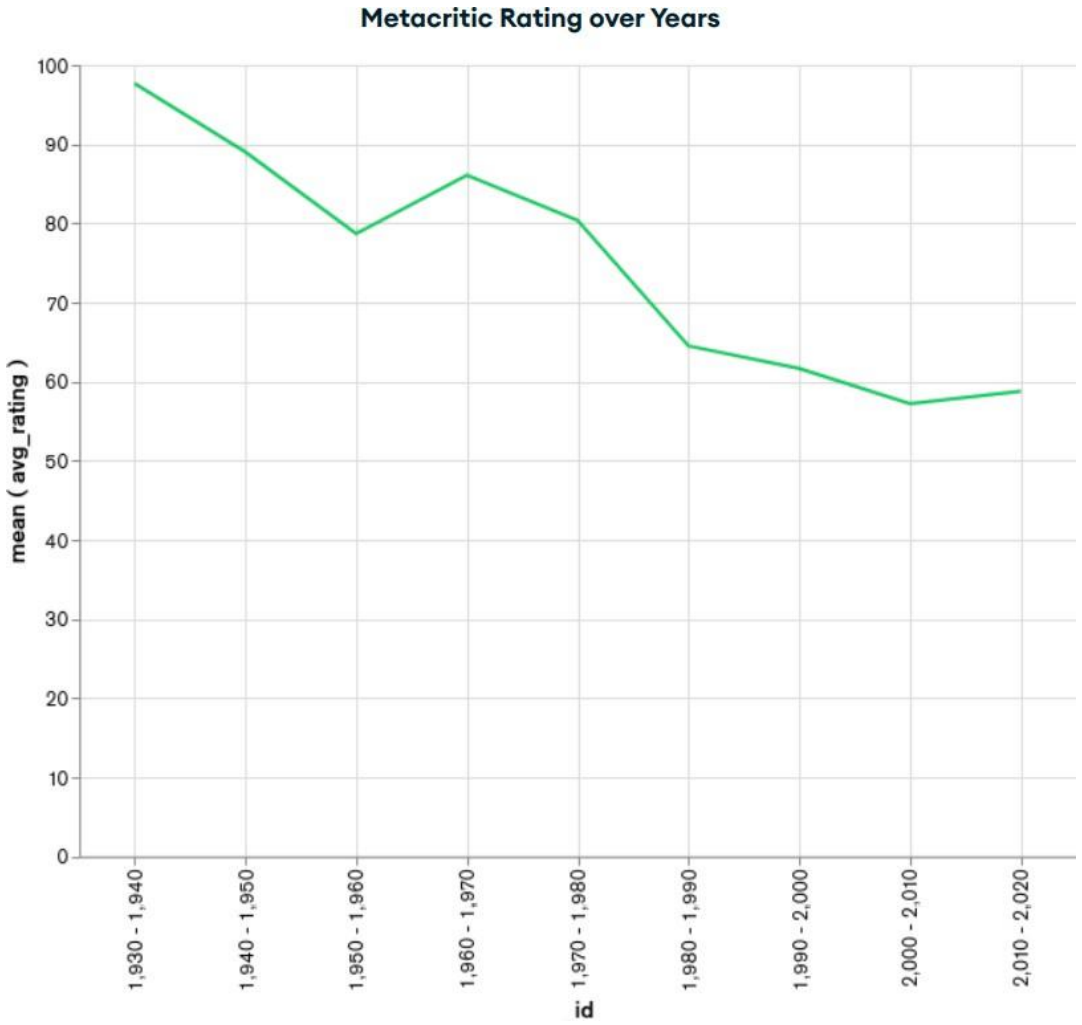
Identify directors whose movies have received the highest Metacritic ratings to showcase critically acclaimed filmmakers.



XIX. Metacritic Rating Over Years

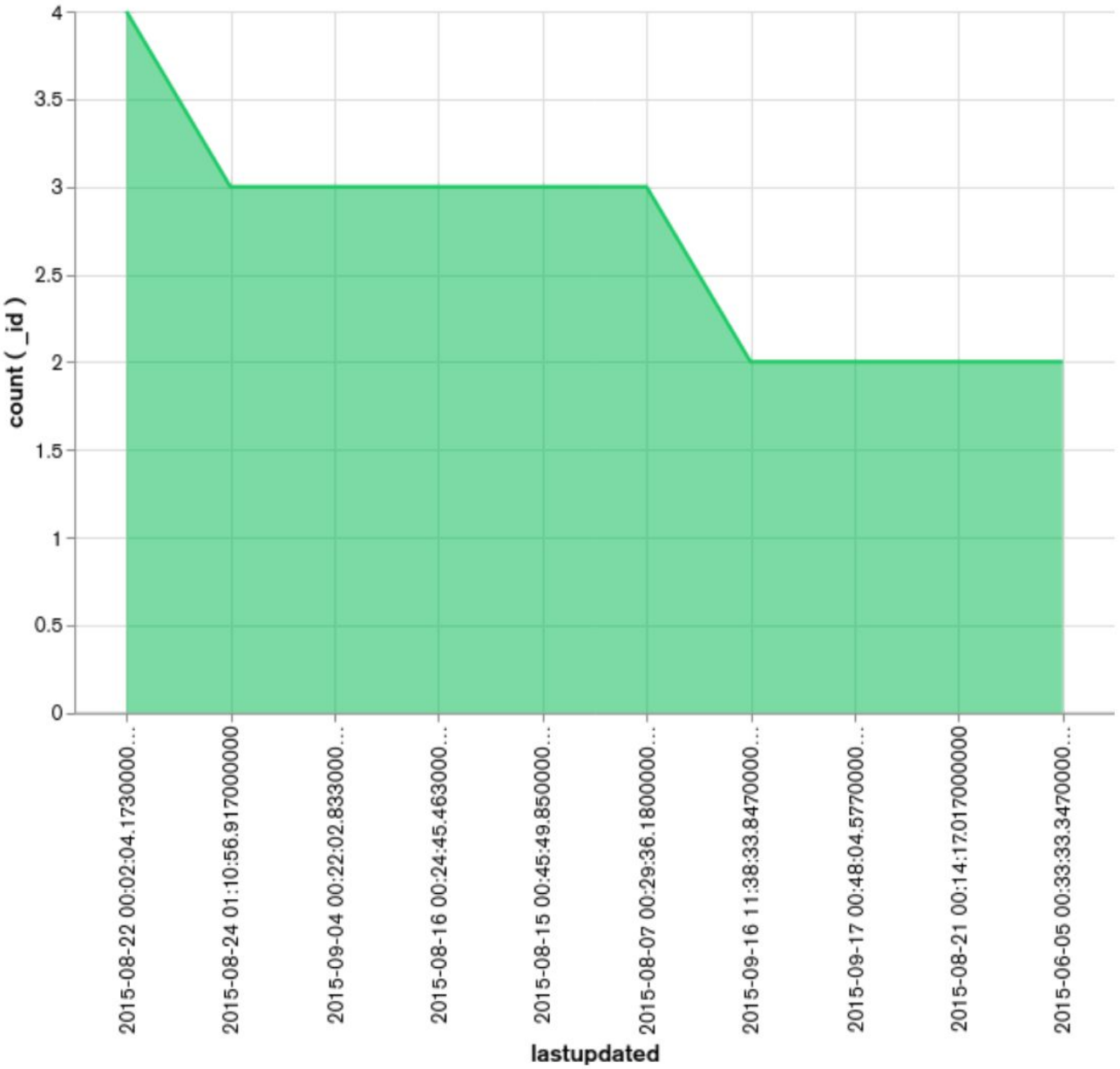
Problem Statement:

Analyze how Metacritic ratings have evolved over the years to identify trends in critical reception.



XX. Average Metacritic in Different Genres

Problem Statement:
Compare the average Metacritic rating across different genres to identify which genres receive the most critical acclaim.



1. Movie Distribution by Genre

a. Audiences favor Drama, Action, and Comedy, making them the most widely produced genres.

b. Sci-Fi and Thriller movies often receive higher ratings, suggesting a niche but dedicated audience.

2. IMDB Ratings Analysis

c. Most movies have IMDb ratings between 6.0 and 8.0, with very few scoring above 8.5.
d. High-rated movies often belong to Drama and Thriller genres, indicating a preference for well-crafted narratives.
e. Movies with IMDb ratings above 8.0 tend to have longer audience engagement.

3. Yearly Trends in Movie Production

f. A significant increase in movie production from 2000 onward suggests the impact of digital transformation and streaming platforms.
g. The 1980s and 1990s had steady movie releases, but growth accelerated in the 21st century.

4. Language Distribution

h. English is the dominant language, but there is a noticeable presence of Spanish, French, and Hindi-language films.
i. Multilingual films have gained popularity, reflecting the globalization of the film industry.

5. Runtime Analysis

j. The majority of movies have a runtime between 90 to 150 minutes, with very few exceeding 3 hours.
k. Short films and documentaries tend to have significantly lower runtimes.

6. Award-Winning Movies & Ratings Correlation

l. Movies with higher IMDb ratings are more likely to have won or been nominated for awards.
m. There is a strong correlation between high critic scores on Rotten Tomatoes and award recognition.

5. Impact of Streaming Services & Future Outlook

a. The rapid increase in movies post-2000 suggests a shift towards streaming platforms and digital content creation.
b. Emerging genres like Documentary and Biographical films may see more growth in the coming years.

6. Director Analysis:

a. Directors with the highest number of movies tend to focus on specific genres, showing their specialization or preference.
b. Directors with consistently high Metacritic ratings often produce movies with strong critical acclaim, regardless of genre.

7. Critic and Viewer Ratings Comparison:

a. There is often a gap between critic ratings (Metacritic) and viewer ratings (IMDb), suggesting that critics and audiences may have differing opinions on movie quality.
b. Genres like **Drama** and **Thriller** tend to receive higher critic ratings, while **Comedy** and **Action** often have mixed responses.
c. **Note:** Some genres like **Film Noir** have exceptionally high average Metacritic ratings, but the number of movies under them is very low, which can inflate the average.

8. Award Analysis:

a. The ratio of award wins to nominations varies significantly between genres. Drama and Thriller genres tend to have higher win percentages compared to Comedy or Action.
b. Many highly rated movies are not necessarily award winners, indicating that critical acclaim does not always translate to formal recognition.

9. Language and Regional Preferences:

a. English dominates the movie dataset, but Spanish, French, and Hindi movies have shown a notable presence, reflecting the impact of globalization on the film industry.
b. Multilingual movies are gaining traction, hinting at a growing audience preference for culturally diverse content.

---

8. Managerial Recommendations

1. Strategic Content Planning:

- Focus on **Drama, Thriller, and Sci-Fi genres** as they tend to attract dedicated audiences and receive higher ratings.
- Consider balancing **high-budget action movies** with **critically acclaimed dramas** to maintain both revenue and quality perception.

2. Targeted Marketing Campaigns:

- Highlight **award-winning and critically acclaimed movies** when promoting content to niche audiences.
- For movies with average ratings but high revenue, leverage star power and visual appeal to draw audiences.

3. Platform and Regional Strategy:

- Invest in producing **multilingual content** to cater to diverse audiences and expand global reach.
- Explore **regional preferences** when distributing movies on streaming platforms to maximize viewership.

4. Future Genre Focus:

- Encourage production of **Documentary and Biographical films**, as they are emerging as popular genres in the streaming era.
- Utilize insights from **viewer-critic discrepancies** to better position movies that are more likely to appeal to general audiences.