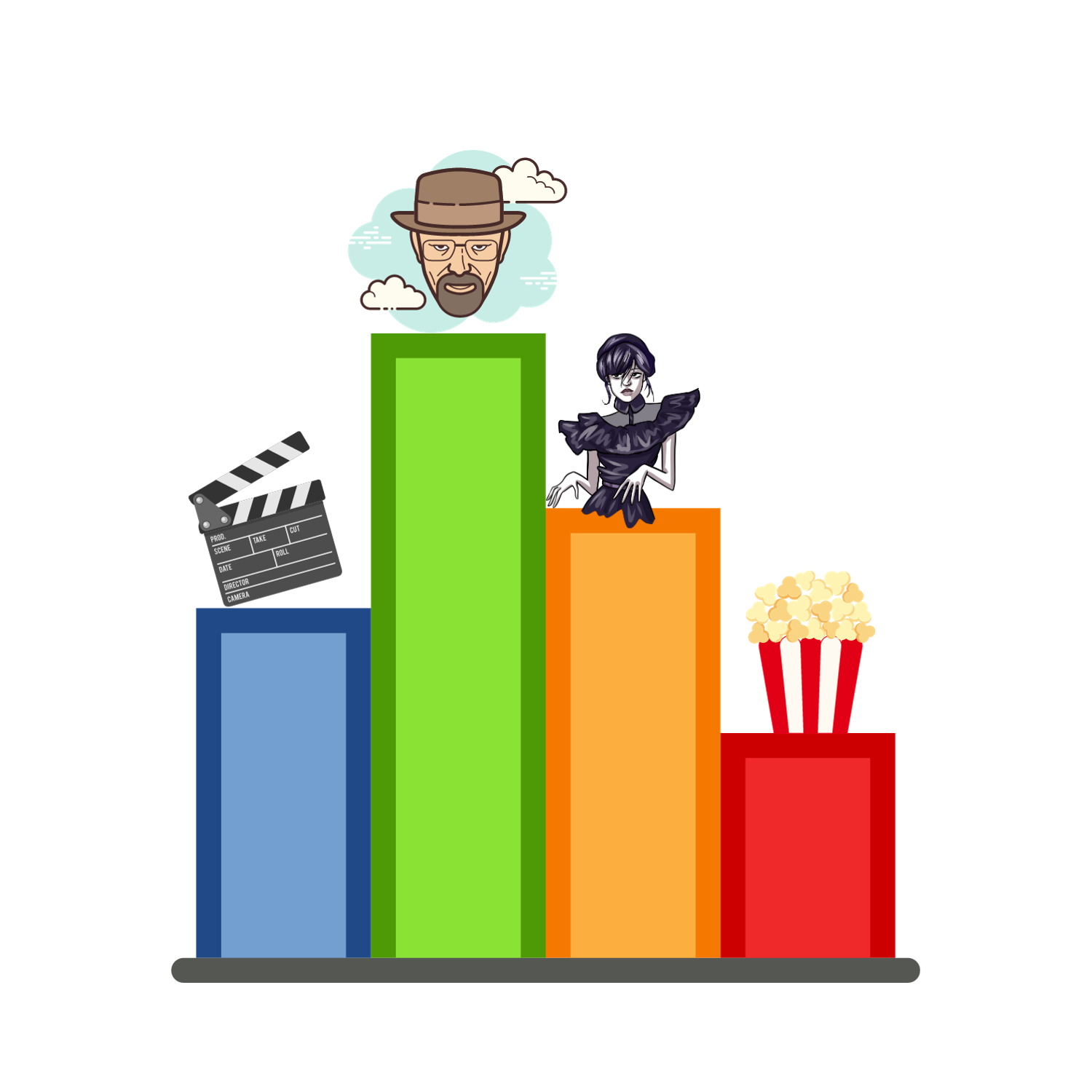
*Visualization of data*

*Final Project*



Submitted by :

* Ali Shaer 307844555
* Raphael Damouny 322465097
* Muhamad Haj Ali 206539991

Table of Contents

[1. Introduction 3](#_Toc146144602)

[2. Data Distribution 5](#_Toc146144603)

[Statistics 5](#_Toc146144604)

[Meta data 6](#_Toc146144605)

[Size and Structure 6](#_Toc146144606)

[Primary Columns 7](#_Toc146144607)

[4. Plans and Ideas 7](#_Toc146144608)

[Initial Ideas 7](#_Toc146144609)

[Sketches/Conceptual Designs 7](#_Toc146144610)

[5. Our solution 9](#_Toc146144611)

[World Map with Bubble Chart (Scatter Plot) 9](#_Toc146144612)

[Area Chart: Movies/Shows Count by Average Rating 10](#_Toc146144613)

[Clickable Drilldown: Average Rating by Genres to Actors 12](#_Toc146144614)

[Heatmap: Average Rating by Genre and Year (2001-2022) 13](#_Toc146144615)

[Count of Titles by Release Year and Type 14](#_Toc146144616)

[Count of Titles by Release Year and Type 15](#_Toc146144617)

[6. Preprocessing 16](#_Toc146144618)

[Handling Null Values 16](#_Toc146144619)

[Handling Titles with Multiple Production Countries 16](#_Toc146144620)

[Geographical Coordinates Mapping 16](#_Toc146144621)

[‘One-Hot’ Encoding of Genres 16](#_Toc146144622)

[Combining IMDB and TMDB Score 16](#_Toc146144623)

[7. Challenges 17](#_Toc146144624)

[Challenge 1: Extracting and Editing CSV Data in JavaScript 17](#_Toc146144625)

[Challenge 2: Designing an Engaging Data Story and Dashboard 17](#_Toc146144626)

[Challenge 3: Working with String Variables (production\_countries and genres) 17](#_Toc146144627)

[8. Story 18](#_Toc146144628)

[9. Conclusion 20](#_Toc146144629)

1. Introduction

Background:

The cinematic universe, replete with its rich tapestry of stories, genres, and characters, has always been a reflection of our society, mirroring our aspirations, fears, and fantasies. With the dawn of digital transformation, movies and TV shows are no longer confined to theaters or television screens; they are now a click away, accessible anytime, anywhere.

Objective:

Amid this vast ocean of content, our project seeks to provide a compass - a guide to navigate the world of movies and TV shows. We aim to uncover the underlying patterns of this industry, identify the prevailing trends, spotlight the stalwarts, and discover the unsung heroes.

Target Audience:

* Movie Buffs: Who wish to explore beyond the mainstream, diving into lesser-known genres or tracing the journey of their favorite actors.
* Film Industry Stakeholders: Producers, directors, and writers can glean insights into what resonates with the audience, helping them tailor their future endeavors.
* Streaming Platforms: In a race to curate a compelling content library, insights from our project can offer platforms a competitive edge.
* Academicians and Researchers: Those studying media and entertainment can benefit from a structured and visual representation of the industry's evolution.

Significance:

In an era of information overload, curated insights can be the difference between discovery and obscurity. Whether it's a filmmaker wondering which genre to explore next or a viewer deciding what to watch on a Friday evening, understanding industry trends and preferences can be pivotal. Through this project, we aim to bridge the gap between content creators and consumers, fostering informed decisions and enriched experiences.

Scope:

Harnessing the power of data, our project delves into an extensive dataset encompassing a myriad of titles, their genres, and the creative minds behind them. Our visualizations promise an immersive journey, unraveling the multifaceted world of films and TV shows.

1. Data Distribution

Statistics

1. **General Statistics:**

* Total Titles: 5,850
* Number of Movies: 3,744
* Number of TV Shows: 2,106

1. **Yearly Distribution (Last 5 years):**

* 2018: 773 titles
* 2019: 836 titles
* 2020: 814 titles
* 2021: 787 titles
* 2022: 371 titles

1. **Top 5 Genres:**

* Drama: 2,968 titles
* Comedy: 2,325 titles
* Thriller: 1,228 titles
* Action: 1,157 titles
* Romance: 971 titles

1. **Top 5 Production Countries:**

* US: 2,323 titles
* India (IN): 622 titles
* Great Britain (GB): 404 titles
* Japan (JP): 287 titles
* France (FR): 248 titles

1. **Average Runtime:**

* Movies: Approximately 98.21 minutes
* TV Shows: Approximately 38.98 minutes

1. **Average Ratings:**

* IMDB: Approximately 6.51
* TMDB: Approximately 6.83

1. **Age Certification Distribution:**

* TV-MA: 883 titles
* R: 556 titles
* TV-14: 474 titles
* PG-13: 451 titles
* PG: 233 titles
* TV-PG: 188 titles
* G: 124 titles
* TV-Y7: 120 titles
* TV-Y: 107 titles
* TV-G: 79 titles
* NC-17: 16 titles

Meta data

**Titles Dataset:**

1. **id:** Unique identifier for each title (movie or TV show).
2. **title:** Name of the movie or TV show.
3. **type:** Categorizes the title as either a movie or a TV show.
4. **description:** A brief summary or synopsis of the title.
5. **release\_year:** The year the title was released.
6. **age\_certification:** The rating or certification indicating the age group the title is suitable for.
7. **runtime:** Duration of the title in minutes.
8. **genres:** Genres associated with the title, such as drama, action, or comedy.
9. **production\_countries:** Countries where the title was produced.
10. **seasons:** For TV shows, indicates the number of seasons.
11. **imdb\_id:** The title's identifier on the IMDB platform.
12. **imdb\_score:** The average rating of the title on IMDB.
13. **imdb\_votes:** The number of users who rated the title on IMDB.
14. **tmdb\_popularity:** A score indicating the title's popularity on TMDB.
15. **tmdb\_score:** The average rating of the title on TMDB.

**Credits Dataset:**

1. **person\_id:** Unique identifier for each person (e.g., actor or director).
2. **id:** Corresponds to the unique identifier of a title in the Titles dataset.
3. **name:** Name of the person.
4. **character:** For actors, indicates the character they portrayed in the title.
5. **role:** Role of the person in relation to the title, such as ACTOR or DIRECTOR.
6. Dataset Description

The datasets used in this project primarily revolve around movies and TV shows, focusing on their attributes, ratings, and associated personnel.

**Source:** The data was sourced from [Kaggle](https://www.kaggle.com/), a popular platform for data science and machine learning. Kaggle hosts a myriad of datasets, competitions, and kernels (notebooks) that allow users to explore, analyze, and generate insights from diverse datasets.

[🔗 Link to dataset](https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies)

**Purpose:** The dataset was curated and shared on Kaggle by the user [victorsoeiro](https://www.kaggle.com/victorsoeiro) to provide a comprehensive overview of movies and TV shows available on Netflix. Kaggle users often utilize such datasets for exploratory data analysis, machine learning tasks, or visualization challenges. In this context, the dataset offers insights into the diverse range of titles Netflix offers, their attributes, ratings, and the associated personnel, making it a valuable resource for researchers, analysts, and enthusiasts wanting to delve into the world of cinematic entertainment.

Size and Structure

**Titles Dataset:** Contains information about 5,850 titles, spread across 15 columns, detailing attributes like genre, release year, ratings, and more.

**Credits Dataset:** Comprises details related to the personnel associated with the titles, spanning 5 columns, and capturing roles like actors and directors for the respective titles.

Primary Columns

**title and type:** Name of the movie or show and its categorization respectively.

**release\_year:** Indicates the year the title was released, useful for temporal trends.

**genres:** Genres associated with the title, facilitating genre-based analyses.

**imdb\_score and tmdb\_score:** Ratings from IMDB and TMDB, respectively, indicating audience reception.

**production\_countries:** Offers insights into the geographical distribution of cinematic production.

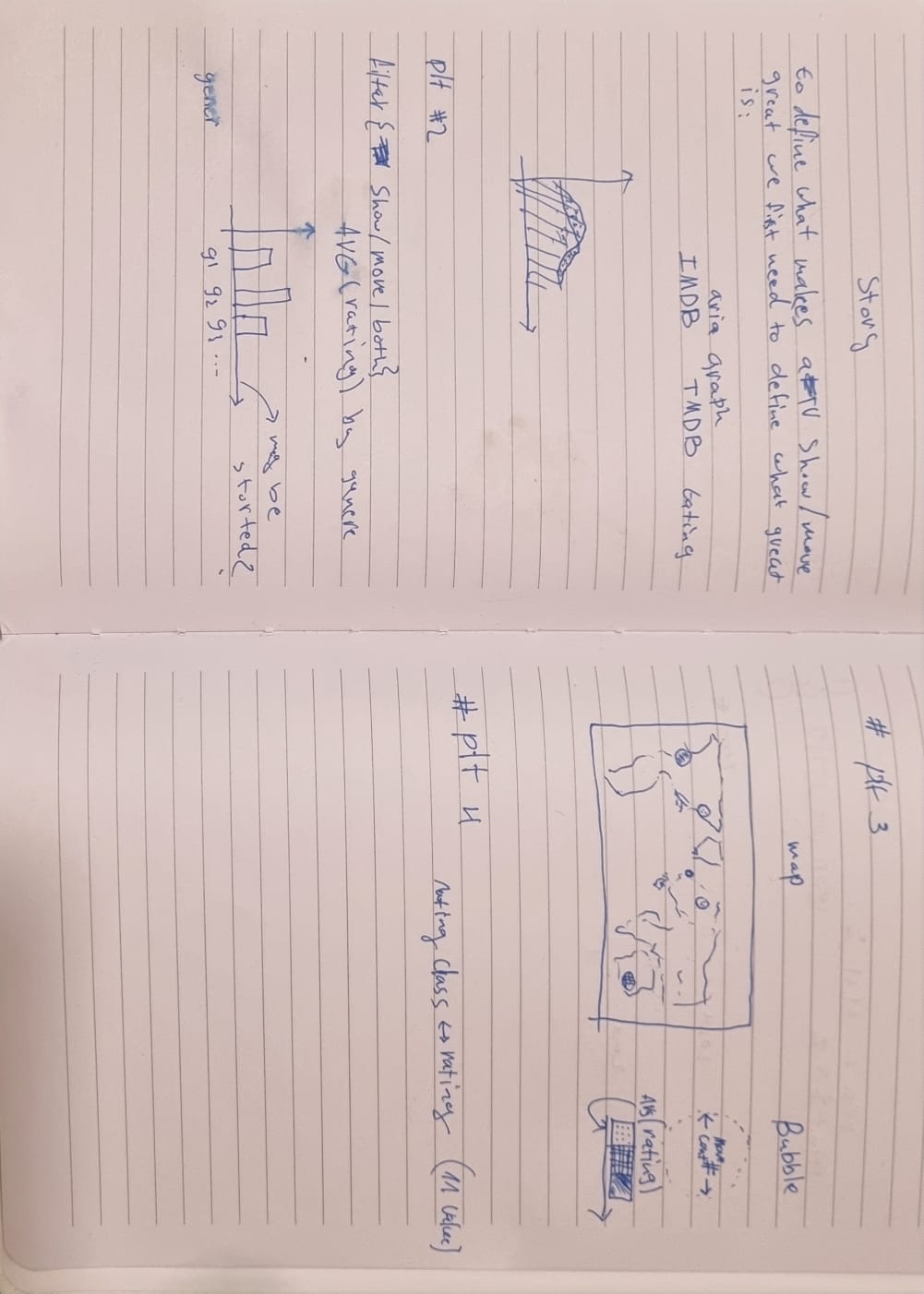
**role (in Credits Dataset):** Highlights the role of the person, such as actor or director, in relation to the title.

1. Plans and Ideas

Initial Ideas

* **Trends Over Time:** Visualize how the number of movies and TV shows has evolved over the years. Is there a growth trend? Are movies or shows becoming more prevalent?
* **Genre Popularity:** Which genres are the most common? Has the popularity of genres changed over the years?
* **Ratings Analysis:** How do ratings (IMDB & TMDB) distribute across titles? Are certain genres more highly rated than others?
* **Geographical Insights:** Which countries produce the most content? Is there a correlation between production countries and ratings or genres?
* **Actor Analysis:** Which actors appear the most? Is there a link between certain actors and higher ratings?

Sketches/Conceptual Designs

* **Rating Distributions with Area Graphs**: To discern the quality benchmarks of movies and TV shows, visualizing the rating distributions from both IMDB and TMDB becomes pivotal. An area graph offers a continuous and intuitive representation, setting the stage for deeper inquiries into what constitutes well-received content.
* **Genre Insights via Bar Charts:** Every genre has its unique appeal, but which ones truly captivate audiences? A bar chart delineating the average ratings across genres can shed light on this, revealing the genres that consistently hit the mark in delivering engaging content.
* **World Cinema Landscape through a Bubble Map:** Cinema is a global affair, and understanding its geographical nuances is enlightening. Bubbles placed over countries, sized by their content output and perhaps colored by average ratings, can paint a comprehensive picture of global cinematic prowess.
* **Dissecting Genres and Age Appropriateness with Bar Charts:** Beyond just genres, how does age certification play into the mix? A bar chart, equipped with an interactive time filter, can offer insights into the alignment (or misalignment) between genres, their typical ratings, and their intended age groups, unveiling trends that might evolve over years.

**A paper with writing on it

Description automatically generated**

Decision-making Process

**Objective Clarification:** Our anchor was discerning what makes a movie or TV show stand out. This objective steered the course of the project.

**Initial Data Dive:** A first-pass exploration highlighted vital insights, notably around rating distributions and genre trends.

**Primary Visualization Choices:**

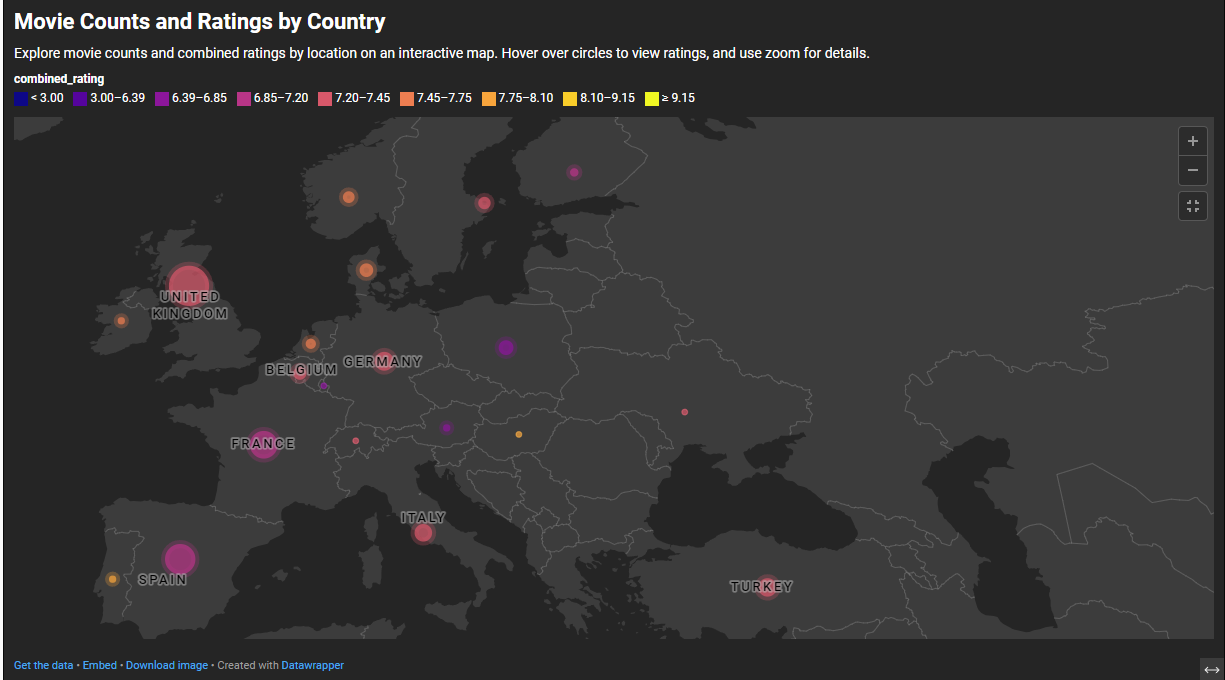
* **Area Graph:** Adopted for its aptness in portraying continuous ratings data.
* **Bar Charts:** Suited for categorical insights, such as genre-based ratings.
* **Bubble Map:** Envisaged to capture the global breadth of cinematic production.

**Adaptable Approach:** While the above served as our foundational visualizations, we championed a fluid strategy. Rather than being boxed into a set plan, we let the data's narrative shape our direction, promoting spontaneous exploration and insights.

**Tool Selection:** Given the project's multifaceted requirements, we gravitated towards [Highcharts](https://www.highcharts.com/). Its versatility and dynamic visualization capabilities made it an ideal choice for our endeavor.

1. Our solution

World Map with Bubble Chart (Scatter Plot)

****

**Description**: This visualization plots countries on a world map using their geographical coordinates. The size of the bubble represents the number of movies produced by each country, while the color represents the combined average rating from both IMDB and TMDB. The visualization aims to provide an overview of the global distribution of movie production and the associated ratings.

**Why We Used It**: The world map with bubble chart offers a spatial representation, allowing viewers to instantly grasp the concentration of movie production globally. By associating the size of the bubble with the number of movies and the color with the ratings, the plot efficiently conveys two different dimensions of the data in a single view. This aids in understanding both the quantity and quality (as indicated by ratings) of movies produced by different countries.

**Pros**:

* **Spatial** **Context**: Provides a clear geographical representation of where movies are being produced.
* **Multi-dimensional**: Simultaneously conveys information about the quantity (number of movies) and quality (ratings) in one visualization.
* **Intuitive**: The size and color-coding are intuitively understood, making the visualization accessible even to those unfamiliar with the data.
* **Comparative**: Easily compare the movie production and ratings of different countries at a glance.

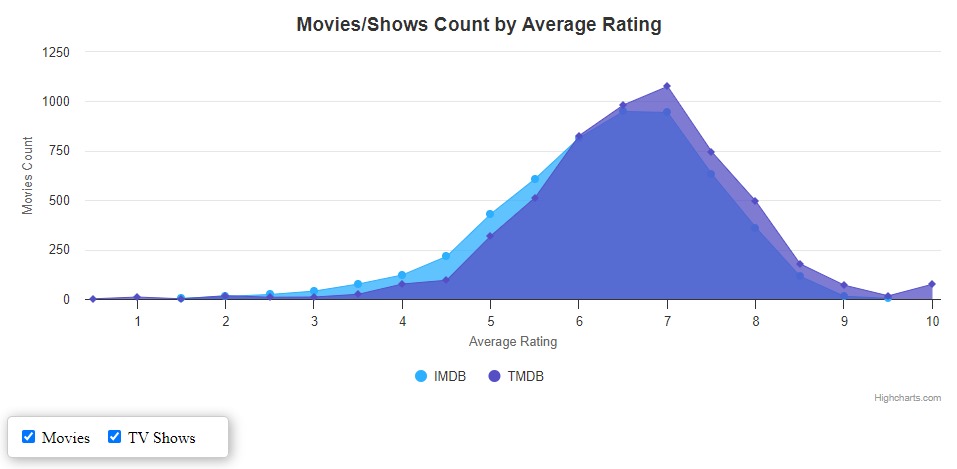
**Cons**:

* **Overlapping** **Bubbles**: In regions with many countries in close proximity, bubbles might overlap, potentially obscuring information.
* **Color** **Ambiguity**: Depending on the color scale used, it might be challenging to differentiate between closely rated countries.
* **Limited** **Depth**: While it provides an overview, the visualization doesn't dive deep into specifics, such as genres, top movies, etc.

**Interactions**:

* **Hovering**: Hovering over a bubble displays the specific combined average rating for that country, offering more detailed insight without cluttering the main view.

## Area Chart: Movies/Shows Count by Average Rating



**Description**: This area chart displays the count of movies and TV shows grouped by their average ratings. It differentiates between ratings sourced from IMDB and TMDB using two distinct areas, one for each platform. The horizontal axis represents the average rating, while the vertical axis indicates the count of movies or shows with that rating.

**Why We Used It**: The primary goal of this visualization is to understand the distribution of ratings for movies and TV shows. By plotting counts against average ratings, we can discern patterns, such as whether most titles are highly rated or if there's a skew towards lower ratings. Differentiating between IMDB and TMDB provides a comparative view between the two platforms.

**Pros**:

* **Distribution Insight**: Quickly grasp the distribution of ratings across movies and TV shows.
* **Comparative View**: By differentiating between IMDB and TMDB, we can compare rating tendencies between the two platforms.
* **Interactivity**: The ability to filter by movie type (Movies, TV Shows, or both) allows for more focused insights.
* **Visual Continuity**: The continuous nature of area charts makes it easy to spot trends or anomalies in the data.

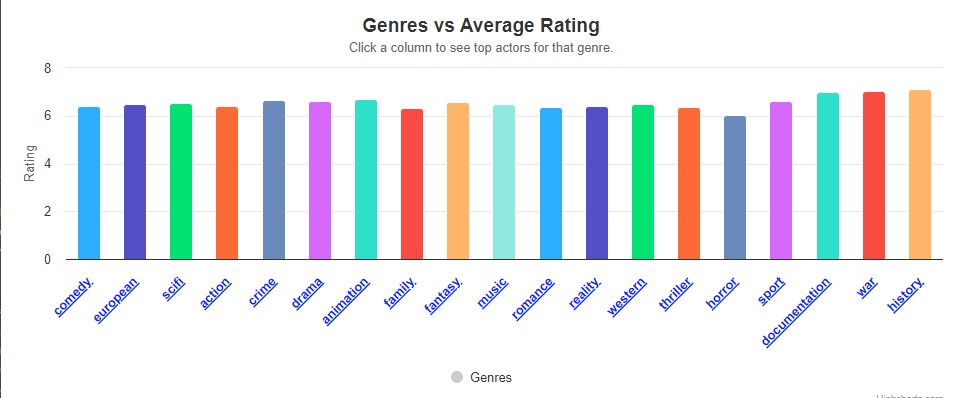
**Cons**:

* **Overlapping Areas**: If the areas for IMDB and TMDB overlap significantly, it might be challenging to distinguish the exact counts for each platform.
* **Perception Bias**: Area charts can sometimes give more visual weight to certain sections, potentially leading to misinterpretation.
* **Granularity**: Depending on the bin size for average ratings, some nuances might be lost.

**Interactions**:

* **Filtering**: Users can select whether they want to view data for Movies, TV Shows, or both, allowing for a tailored visualization experience.

## Clickable Drilldown: Average Rating by Genres to Actors



A graph with orange and white text

Description automatically generated

**Description**: The initial visualization is a bar plot that showcases the average rating grouped by genres. The x-axis represents the various genres, while the y-axis denotes the average rating. Upon clicking a particular genre (bar), the visualization drills down to a second bar plot. This subsequent plot displays actors within the selected genre, with the y-axis representing the average rating of movies that the actor has been in.

**Why We Used It**: The drilldown approach offers a two-tiered insight. Initially, it provides an overview of how genres fare in terms of ratings. Once a genre is selected, it delves deeper into the key actors within that genre and their respective ratings. This hierarchical insight can help identify not just high-performing genres but also standout actors within them.

**Pros**:

* **Layered Insight**: The drilldown mechanism provides both a broad overview and detailed insights in a single interactive visualization.
* **User Engagement**: Interactive visualizations like drilldowns tend to engage users more, offering them control over what they want to explore.
* **Space Efficiency**: Instead of having two separate plots, the drilldown effectively utilizes space by presenting information hierarchically.
* **Focused Analysis**: By exploring actors within a specific genre, users get a more focused perspective, free from the noise of other genres.

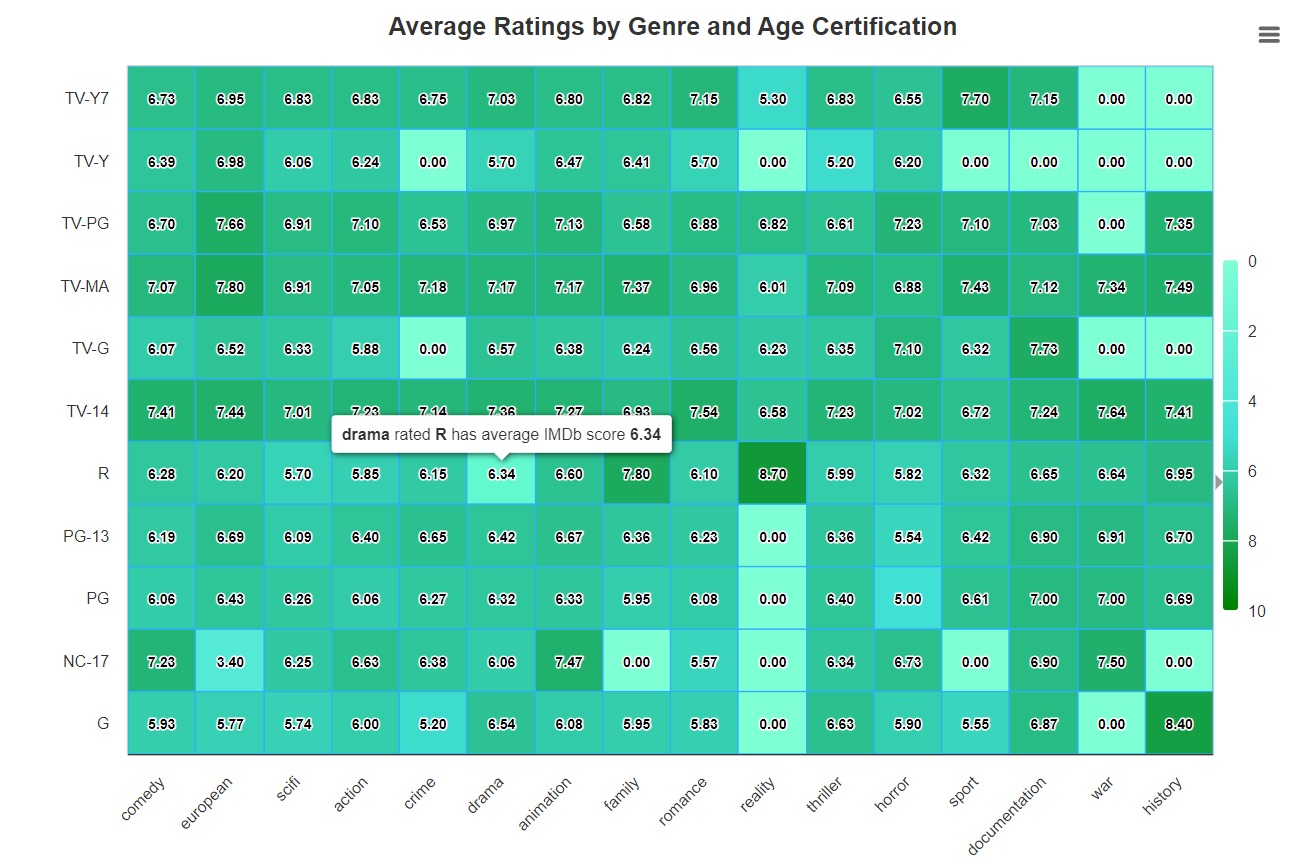
**Cons**:

* **Complexity**: Drilldowns, while informative, can be slightly complex for some users, especially those unfamiliar with interactive visualizations.
* **Data Overload**: If a genre has a vast number of actors, the second plot might be overwhelmed with data, making it less intuitive.
* **Dependence**: The second visualization is dependent on the user's interaction with the first, which might lead to some users missing out on the detailed insights.

**Interactions**:

* **Click**: Clicking on a genre (bar) in the first plot takes the user to the second plot showcasing actors within that genre.
* **Filtering**: Users can adjust a slider to determine the number of top actors they want to view in the second plot, offering a tailored visualization experience.

## Heatmap: Average Rating by Genre and Year (2001-2022)



**Description**: This heatmap provides a visual representation of the average ratings of movies and TV shows by genre and release year. The x-axis represents the different genres, while the y-axis encompasses the years from 2001 to 2022. Each cell's color intensity corresponds to the average rating for that genre in a specific year. Darker hues indicate higher ratings, while lighter ones denote lower ratings.

**Why We Used It**: The heatmap is designed to allow viewers to quickly identify the highest-rated genres for any given year. By juxtaposing genres and years, it offers a snapshot of the evolving preferences and quality of content across two decades. It aids in discerning patterns, such as the rise and fall of genres or consistently high-performing genres over the years.

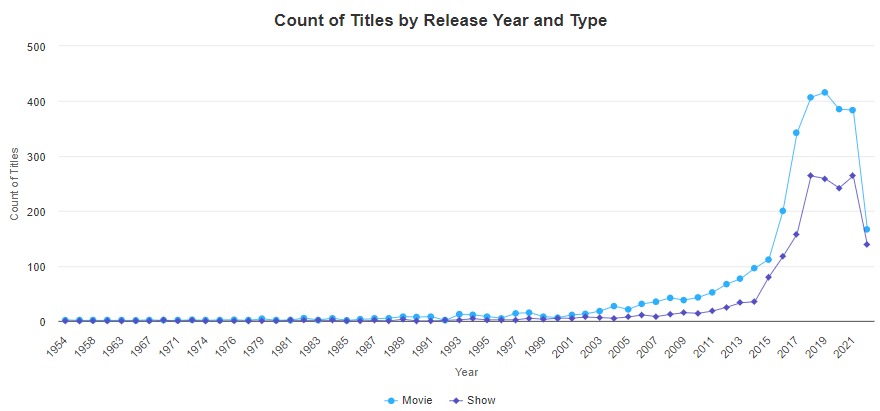
**Pros**:

* **Quick Insights**: The color-coded nature of heatmaps allows for rapid assimilation of trends and patterns.
* **Compact & Multi-dimensional** : Provides a vast amount of data in a condensed visual format, making it easier to compare and contrast, and Simultaneously conveys information about time (years), categories (genres), and metrics (ratings).
* **Dynamic Trends**: Helps in understanding how genre preferences have evolved over the years based on ratings.

**Cons**:

* **Overwhelming**: The dense information might be overwhelming for some viewers, especially if many genres are represented.
* **Quantitative Precision**: While heatmaps are great for spotting trends, they might not be the best for extracting precise quantitative values.
* **Objective**: With this heatmap, a viewer can quickly determine which genres were the most acclaimed in any given year between 2001 and 2022. For instance, if someone is looking to explore movies from 2010, the heatmap will guide them towards genres that were particularly well-received that year.

## Count of Titles by Release Year and Type



**Visualization: Count of Titles by Release Year and Type - CHANGE SCALE**

**Description**: This visualization is a line graph that depicts the count of titles categorized into two types: movies and TV shows, over a range of release years. The x-axis represents the release years, while the y-axis shows the count of titles. There are two lines in the graph, one for movies and another for TV shows, each tracing the variation in title counts over time.

**Why We Used It**: This visualization method is employed to showcase how the count of titles varies for movies and TV shows across different release years. It allows for a straightforward comparison of the trends in title production for both types over time.

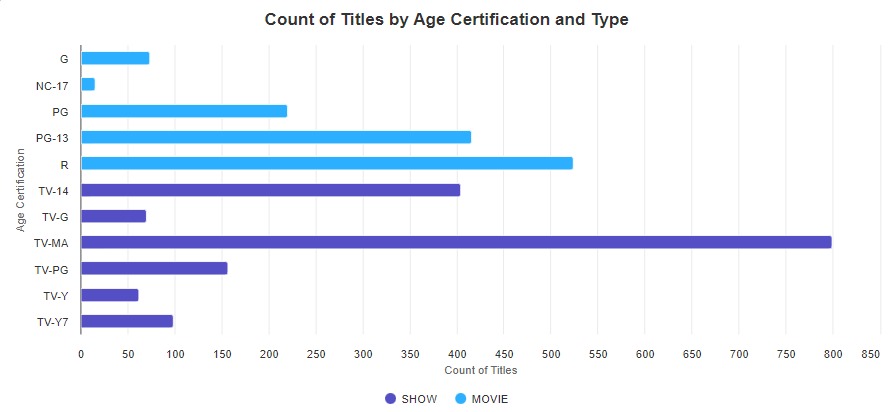
**Pros**:

* **Clear Comparison**: The use of two distinct lines makes it easy to compare and contrast the trends in title counts for movies and TV shows.
* **Temporal Insights**: Users can quickly identify which years had higher production of one type over the other, aiding in temporal analysis.
* **Simplicity**: This visualization is easy to understand and doesn't require user interactions like drilldowns.

**Cons**:

* **Limited Detail**: While it offers a clear comparison, this visualization may lack the depth of insights provided by interactive drilldowns.
* **Potential Overlapping**: If the lines for movies and TV shows overlap frequently, it might be challenging to distinguish them.

## Count of Titles by Release Year and Type



**Description**: This horizontal bar plot provides a visual representation of the count of titles categorized by age certification, with distinct bars for each certification. The x-axis represents the count of titles, while the y-axis displays the age certifications. Each age certification bar is divided into two colors: one representing TV shows and the other representing movies, allowing for a comparison of title counts within each certification for these two types of content.

**Why We Used It**: This visualization method is employed to showcase how the count of titles varies for different age certifications and distinguishes between TV shows and movies. It offers a clear comparison of content types within each certification category.

**Pros**:

* **Comparative Analysis**: Users can easily compare the count of TV shows and movies within each age certification, gaining insights into content distribution.
* **Categorical Insight**: The horizontal bars present a categorical view of age certifications, aiding in understanding content diversity.

**Cons**:

* **Limited Detail**: This visualization may not provide granular details about individual titles or their characteristics.
* **Size Variation**: The length of bars may vary significantly depending on the count, potentially making it challenging to compare very different counts.

# Preprocessing

## Handling Null Values

* **Objective**: Ensure data integrity and reliability for robust analysis.
* **Procedure**:
  + Identified columns and rows with null values.
  + Removed rows with null values in key columns to maintain dataset consistency.  
    columns: ['imdb\_score','title','tmdb\_score','imdb\_score']

## Handling Titles with Multiple Production Countries

* **Objective**: Simplify analysis by focusing on titles with a clear origin.
* **Procedure**:
  + Identified approximately 700 movies/shows that were produced in more than one country.
  + Opted to drop titles with multiple production countries due to their relatively small number, which could introduce analytical uncertainties.
  + This was done exclusively to the dataset used in the map visualization

## Geographical Coordinates Mapping

* **Objective**: Enable geographical visualization of movie origins.
* **Procedure**:
  + Utilized the 'geopy' library to convert the production country of each title into corresponding latitude and longitude coordinates.
  + Ensured that this conversion was performed after addressing the titles with multiple production countries to guarantee accurate mapping.

## ‘One-Hot’ Encoding of Genres

* **Objective**: Facilitate genre-based analyses without ambiguity.
* **Procedure**:
  + Recognized that some titles are associated with multiple genres.
  + Applied one-hot encoding to segregate and represent each genre distinctly.
  + Retained titles associated with multiple genres, as dropping them would lead to a significant loss of data, especially when genres are pivotal to the analysis.

## Combining IMDB and TMDB Score

* **Objective**: Offer a consolidated view of audience reception across two major platforms.
* **Procedure**:
  + Introduced a new column, **combined\_rating**.
  + Computed this column by taking the simple average of the IMDB and TMDB ratings for each title.

## Challenges

## Challenge 1: Extracting and Editing CSV Data in JavaScript

* Description: The team faced difficulty in efficiently extracting and manipulating data from a CSV file using JavaScript. Python was the preferred language, but the project required JavaScript for visualization.
* Solution: To address this challenge, some data was preprocessed and converted into static variables in Python before being incorporated into the JavaScript-based visualizations.

## Challenge 2: Designing an Engaging Data Story and Dashboard

* Description: Creating a compelling data story and designing an effective dashboard that provides a user-friendly and engaging experience was a significant challenge.
* Solution: To overcome this challenge, the team worked on developing a well-structured data story with clear narratives and interactive visualizations to offer a comfortable user experience. User testing and feedback were utilized to refine the dashboard.

## Challenge 3: Working with String Variables (production\_countries and genres)

* Description: Managing columns like 'production\_countries' and 'genres,' which represent lists in string format, presented challenges in data manipulation and visualization.
* Solution: Over time, the team devised creative solutions to handle these string variables based on the specific visualizations they intended to create. These solutions likely involved data preprocessing, parsing, and aggregation to make the data suitable for visualization.

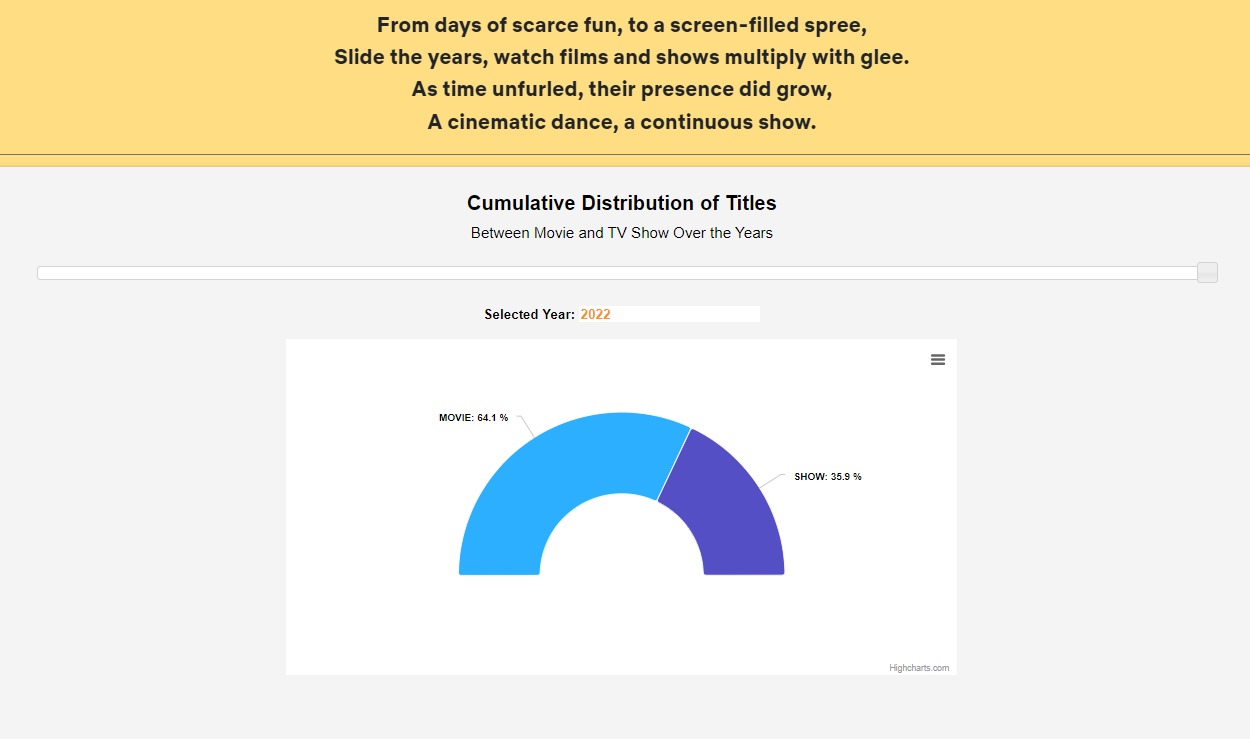
1. Story

From days of scarce fun, to a screen-filled spree,

Slide the years, watch films and shows multiply with glee.

As time unfurled, their presence did grow,

A cinematic dance, a continuous show.

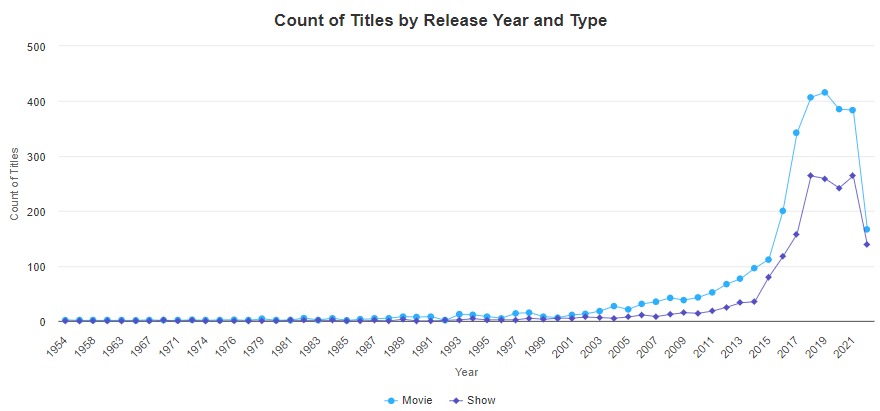


As the years went by, the line soared and dipped,

With movies and shows, the graph tightly gripped.

But during Corona's reign, a slight pause we did see,

Yet entertainment rebounded, like a phoenix, wild and free.

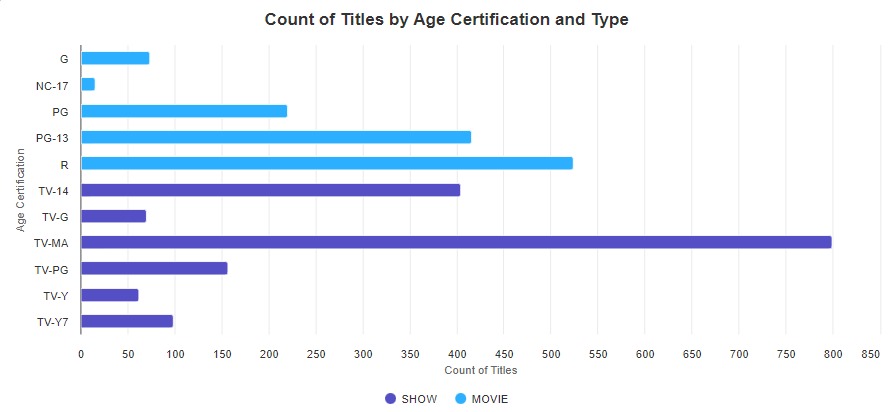


Through highs and through lows, the lines twist and turn,

For shows and for movies, ratings make them churn.

From family-friendly to those for mature minds,

Watch the trends shift, as each viewer finds.

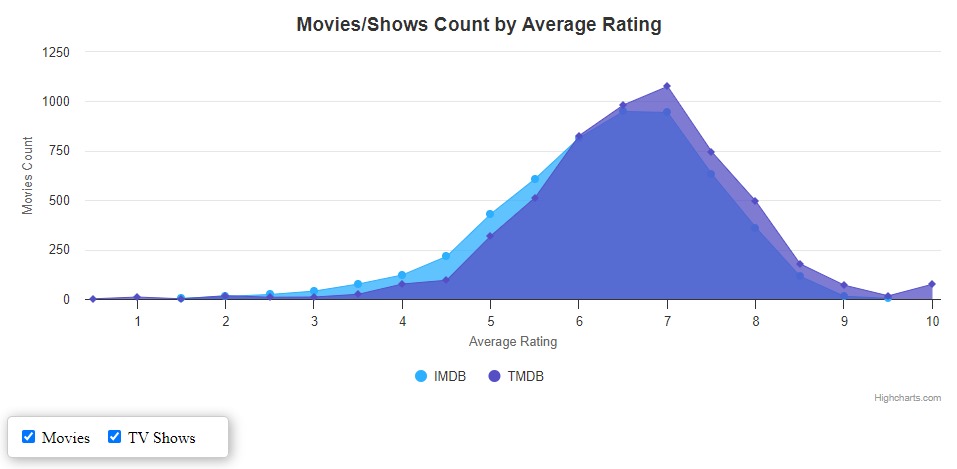


From IMDB to TMDB, the verdicts are in,

With a mode around seven, most films surely win.

Worry not of their worth, for the scales tip so high,

Quality's in abundance, under the critic's sky.

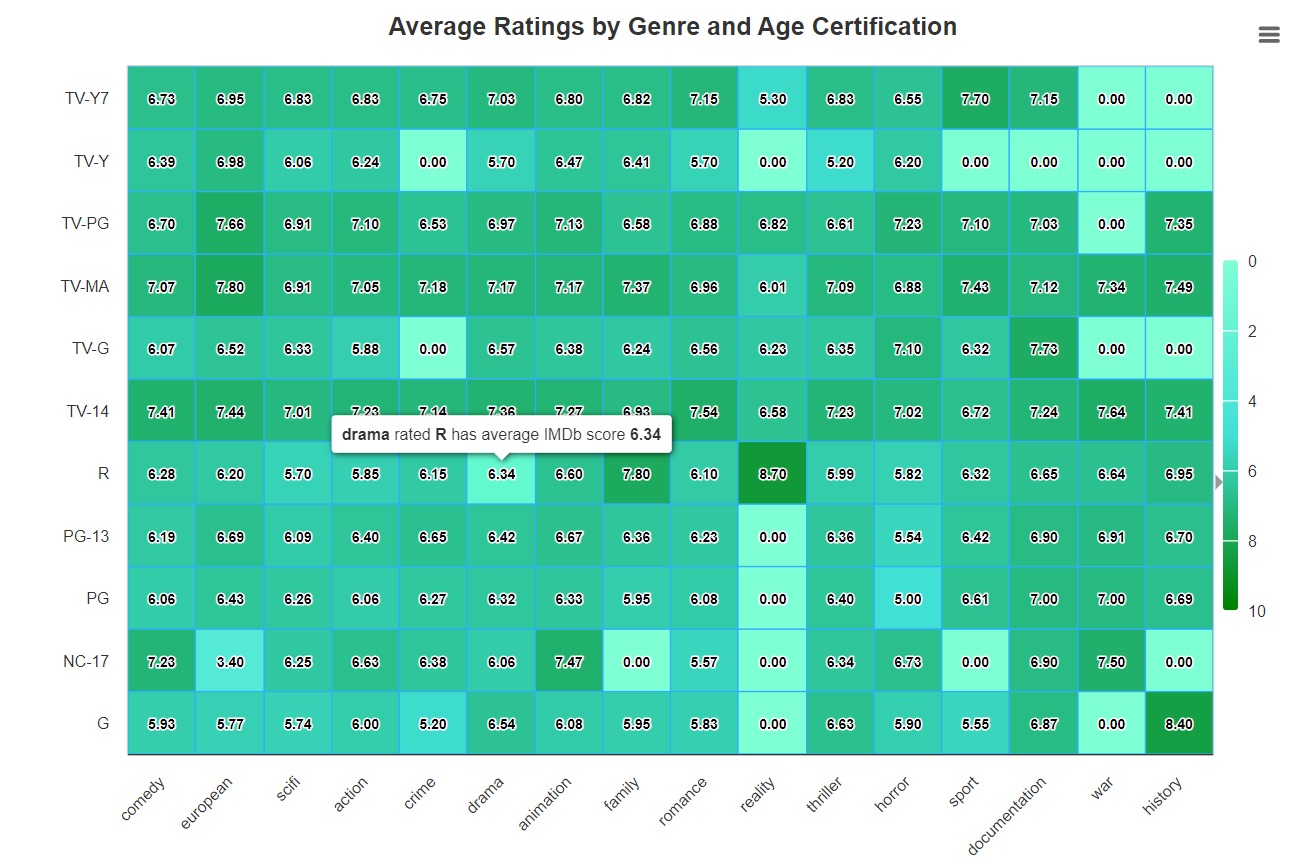


We've not forgotten genres, in this scoring spree,

On the heatmap they shine, for all to see.

Now all that remains, for viewers employed,

Is to choose your delight, and be overjoyed.



1. Conclusion

In our journey through the cinematic universe, we embarked on a quest to uncover what truly defines a "good" movie. Through our analysis of IMDb and TMDb ratings, we've unearthed a rather intriguing pattern: movies, irrespective of their genre, seem to be predominantly well-received. This observation challenges the age-old debate of genre supremacy, suggesting that perhaps the love for cinema goes beyond specific tastes. Furthermore, the predominance of high ratings, characterized by a left-tailed distribution, underscores the richness of quality available in the movie world. It's heartening to see that most filmmakers receive a nod of approval from viewers and critics alike. So, while individual preferences might diverge, it's evident that the cinematic world offers something delightful for everyone.