TITLE

1, 1, AND 1

1. Introduction

In recent years, the film industry has undergone significant transformations due to changing audience preferences, technological advancements, and shifts in global box office trends. Understanding consumer preferences in films has become crucial for achieving both profitability and widespread popularity. This article focuses on analyzing the key elements of successful movies to identify strategies for creating a commercially successful film.

One of the key methods employed in this analysis is text frequency analysis. By examining word frequency across various genres, we can identify common elements that resonate with audiences and correlate these patterns with commercial success or critical acclaim. In this study, the team utilized both the TMDb and OMDb APIs to retrieve high-rated movies and their corresponding narratives.

To effectively analyze the collected numeric data, we plan to employ analysis and visualization. EDA is crucial for identifying data distribution patterns, outliers, and potential biases within the dataset. This helps us assess the development of the global film industry, ensuring our analysis is grounded in data-driven insights. Meanwhile, visualizing production countries using both static and interactive choropleth maps might provide an intuitive way to present geographical patterns in film production, helping to highlight regional trends and production dominance.

Nevertheless, several potential challenges may arise that could impact the effectiveness of our methods. Data inconsistencies, such as inconsistent naming conventions and ids between different APIs, might complicate the merging and visualization of datasets. Missing or incomplete data in key attributes like `production_countries`, `budget`, or `revenue` could limit the scope of analysis. Additionally, the dataset may exhibit significant imbalances, with certain countries or genres dominating the sample, potentially skewing the results.

2. Data Acquisition

This project utilizes data from The Movie Database (TMDb) API and OMDb API for their comprehensive movie information. TMDb is chosen over IMDb for its refined rating system, which emphasizes dedicated film enthusiasts' opinions, resulting in more selective and objective ratings. TMDb provides key details while the OMDb API supplements the dataset with movie plot summaries. Together, these sources offer a robust foundation for exploring movie trends, genre-specific insights, and narrative patterns.

2.1. Data Retrieval for Top-Rated 500 Movies.

Focusing on the top-rated 500 movies offers a broad yet manageable dataset that highlights films with consistently high audience and critic acclaim. To gather information on the top 500 highest-rated movies, the TMDb API is used to extract relevant details such as `id`, `title`, `vote_count` and `rating`. Additional requests are made to obtain key attributes including `genre_ids`, `budget`, `revenue`, `production_countries`, and `imdb_ids`, which are crucial for subsequent analysis and visualization.

2.2. Data Retrieval for Top-Rated 200 Movies by Genre.

In addition to an overall analysis, exploring the top-rated 200 movies by genre provides deeper insights into genre-specific patterns. Different genres often reflect distinct cultural preferences, production styles, and regional strengths. Therefore, data on top-rated movies across 17 different genres (including Action, Adventure, Animation, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Music, Mystery, Romance, Sci-Fi, Thriller, and War) is collected, applying criteria such as

1.
1.
1.

minimum vote thresholds of 300 to ensure data quality. Notably, the dataset includes only 147 Documentary movies that meet the criteria, while TV Movies and Western are excluded due to their niche content and limited global presence.

For each genre, multiple requests are performed to gather a representative sample of well-received movies.

2.3. **Data Retrieval for Movie Plots.**

After collecting key details about top-rated movies, our project also incorporates movie plot summaries for the top-rated 200 movies by genre to enhance the dataset with narrative content based on the IMDb id provided in TMDB API. To achieve this, the OMDb API is utilized, which offers detailed movie information based on `imdb_ids` obtained previously. These plots are collected to support Natural Language Processing (NLP) tasks in the subsequent analysis phase.

3. Data Analysis

3.1. **Text Frequency Analysis.**

3.1.1. *Data Cleaning and Text processing.*

Due to the inconsistency of `imdb_ids` between OMDb and TMDb, some movie plots were unavailable when retrieving data. To address this issue, the team used movie titles as a reference to find the correct plots when `imdb_id` mismatches occurred. However, inconsistencies in movie titles across different APIs were also encountered. For these cases and missing data, the team excluded the affected narratives from the analysis to maintain data accuracy.

For data preparation, the team uses Python's Natural Language Toolkit (NLTK) to remove common words from the plots. These words, identified through NLTK's stop-word list, are excluded to focus on meaningful content. Additional preprocessing steps include removing punctuation, converting text to lowercase, normalizing spacing, and eliminating empty strings to ensure consistency.

In total, the team has processed plots for 3,400 movies, comprising 86,635 words.

3.1.2. *Plots Frequency Analysis by Genres.*

Genre	1st Word (Count)	2nd Word (Count)	3rd Word (Count)	4th Word (Count)	5th Word (Count)
Action	must (27)	new (20)	world (19)	two (18)	save (16)
Adventure	world (27)	young (23)	must (22)	new (21)	two (15)
Animation	young (36)	new (26)	girl (23)	world (22)	boy (21)
Comedy	love (22)	new (21)	two (21)	young (20)	life (19)
Crime	two (24)	young (24)	life (23)	murder (21)	police (20)
Documentary	documentary (34)	world (24)	life (17)	look (16)	film (13)
Drama	life (38)	young (26)	love (19)	boy (19)	war (18)
Family	young (34)	new (28)	family (23)	boy (23)	world (21)
Fantasy	young (36)	new (28)	life (26)	girl (21)	world (20)
History	war (36)	world (29)	story (27)	young (23)	life (20)
Horror	young (22)	family (20)	new (20)	man (17)	woman (17)
Music	life (38)	music (37)	singer (21)	band (20)	new (19)
Mystery	murder (32)	man (22)	wife (22)	woman (21)	young (19)
Romance	love (55)	young (36)	two (28)	woman (28)	life (27)
Sci-Fi	must (28)	world (25)	earth (21)	new (19)	time (18)
Thriller	two (23)	murder (19)	one (18)	must (17)	young (17)
War	war (94)	world (51)	ii (36)	german (33)	young (25)

Genre	6th Word (Count)	7th Word (Count)	8th Word (Count)	9th Word (Count)	10th Word (Count)
Action	one (15)	help (14)	team (14)	city (13)	family (13)
Adventure	team (14)	save (14)	find (13)	family (13)	story (13)
Animation	two (17)	life (16)	save (16)	family (15)	must (15)
Comedy	man (16)	friends (16)	must (16)	family (14)	get (11)
Crime	crime (19)	one (19)	man (17)	must (15)	family (13)
Documentary	footage (13)	one (11)	story (11)	filmmaker (9)	years (7)
Drama	two (17)	family (15)	story (14)	son (13)	new (13)
Family	girl (20)	must (20)	life (19)	father (17)	friends (17)
Fantasy	family (19)	must (16)	boy (15)	time (15)	man (14)
History	ii (15)	two (14)	man (13)	german (13)	british (13)
Horror	two (15)	mother (14)	house (13)	life (12)	killer (12)
Music	rock (17)	story (17)	young (16)	love (16)	musical (16)
Mystery	two (18)	find (18)	detective (18)	family (18)	mysterious (18)
Romance	man (23)	girl (19)	one (19)	finds (14)	relationship (14)
Sci-Fi	find (16)	life (16)	future (15)	help (15)	finds (14)
Thriller	police (16)	woman (15)	life (15)	help (13)	find (13)
War	two (23)	army (22)	story (22)	soldiers (21)	american (18)

FIGURE 1. Top 10 Words For Each Movie Genre

In the **Action** genre, words such as "must," "new," and "world" predominate, underscoring high-stakes narratives and urgent missions. Terms like "save" and "one" reflect the genre's emphasis on individual heroism, with protagonists often embarking on quests to protect or rescue others. Similarly, the **Adventure** genre shares thematic elements, frequently featuring words like "world," "young," and "must," which suggest grand explorations, personal growth, and the overcoming of challenges. Both genres also highlight the significance of teamwork, with the term "team" appearing regularly, indicating the centrality of alliances in their plot structures.

The **Animation** genre places a strong emphasis on youth, as evidenced by the frequent occurrence of terms such as "young," "girl," and "boy." The repeated appearance of words like "world" and "life" suggests a thematic affinity for imaginative settings and the exploration of life lessons within the genre. Family films, often a subset of animation, focus on interpersonal relationships and personal development, as reflected in the prominent use of words like "family," "young," "boy," and "girl," highlighting themes of love, connection, and shared experiences. Notably, terms such as "friends," "love," and "journey" also appear with high frequency, underscoring the genre's focus on relationships and emotional growth. These films frequently explore the dynamics of family and close-knit groups, offering narratives that delve into personal journeys and the emotional bonds that shape characters' experiences.

In the **Comedy** genre, terms like "love," "new," and "two" highlight the genre's focus on relationships and humorous situations arising from misunderstandings or unexpected pairings. The prominence

of "life" and "young" suggests that many comedies explore the lighter aspects of life's challenges. The **Romance** genre, by contrast, centers on the theme of love, with "love," "young," and "two" emerging as dominant terms. The repeated occurrence of "woman" and "man" signifies that romantic films frequently examine relationships between individuals, placing particular emphasis on the emotional journey of the characters.

The **Crime** genre is characterized by words such as "murder," "police," and "crime," which are typically associated with stories centered around criminal investigations or law enforcement. The frequent mention of "family" and "detective" indicates that these films often explore the human toll of crime and the quest for justice. The **Mystery** genre shares similar themes, with words like "murder," "man," and "wife" highlighting the genre's focus on investigative narratives. "Detective" and "family" again appear prominently, suggesting that many mystery films revolve around complex investigations that intertwine personal relationships and hidden secrets.

In the **Drama** genre, the term "life" is particularly prominent, emphasizing the genre's exploration of human experiences and emotions. Words such as "young," "love," and "boy" highlight themes of personal struggle, relationships, and character development. Similarly, **History** films focus on significant historical events, with words like "war," "world," and "story" dominating. These terms suggest that historical dramas often delve into global conflicts or pivotal moments, with a focus on individual experiences during times of war and upheaval.

The **Sci-Fi** genre frequently incorporates words like "must," "world," and "earth," indicating futuristic or otherworldly themes involving technology, survival, and exploration. The inclusion of terms such as "time" and "future" further emphasizes the genre's engagement with speculative concepts. In the **Thriller** genre, the prevalence of "murder," "police," and "detective" suggests narratives marked by high-stakes suspense, often revolving around criminal investigations and action-driven plots. The Horror genre, in contrast, relies heavily on terms such as "young," "family," and "man," indicating a focus on ordinary individuals thrust into terrifying situations. The frequent use of words like "killer" and "haunted" reflects the genre's emphasis on danger, fear, and supernatural or psychological terror.

Finally, **War** films are dominated by words such as "war," "world," and "ii," which refer to major historical conflicts, particularly World War II. These films typically explore the impact of war on soldiers, families, and societies, as reflected in the frequent appearances of words like "soldiers," "army," and "battle." The **Documentary** genre stands apart, with terms such as "documentary," "world," and "life" highlighting the genre's focus on real-world events and human experiences. The presence of words like "footage" and "filmmaker" signals the genre's emphasis on factual storytelling, often providing in-depth explorations of specific topics, people, or phenomena.

3.1.3. *Plots Frequency Analysis overall.*

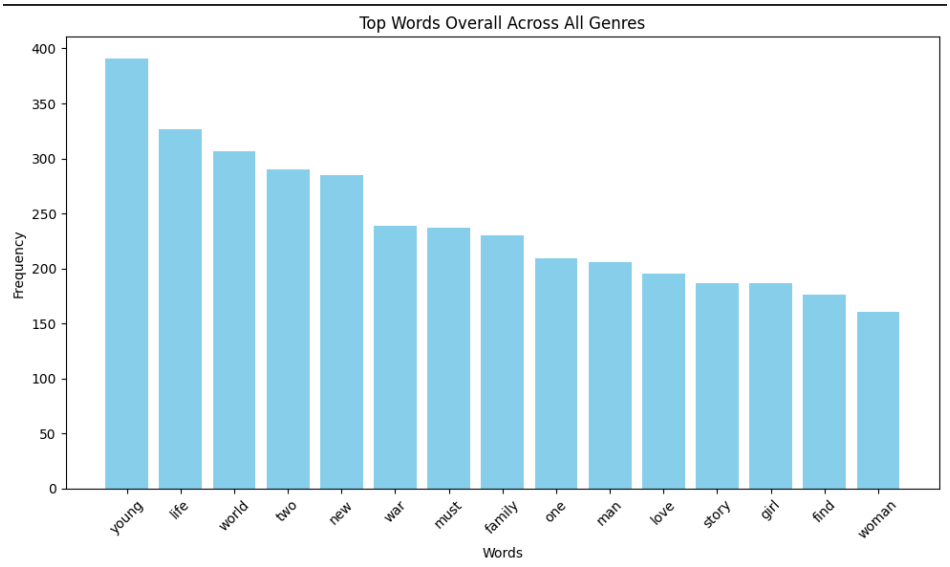

FIGURE 2. Top 15 Words Overall Across All Movie Genres

Overall, the frequent occurrence of words such as "young," "family," "two," "man," and "woman" highlights the central importance of connections with family and romantic partners in films. The term "find" reflects the journey of protagonists discovering the truth of the story and their deeper understanding of life. The word "life" suggests that many films conclude with themes of saving lives or coming to a profound understanding of life's truths. Producers should consider these elements when crafting their films, as they resonate deeply with audiences.

3.2. **Exploratory Data Analysis.**

3.2.1. *Data Cleaning and Processing.*

A crucial aspect of this process involves examining missing and invalid financial values. Many records contain zero or missing values in key financial fields such as `budget` and `revenue`. Since these variables play a fundamental role in profitability calculations, all movies with missing or zero values in either field are removed from the dataset. This step ensures that subsequent `profit` and `ROI` calculations remain valid and do not include unrealistic or misleading results.

Additionally, inconsistencies in country data require attention. Many movies are produced through international collaborations, meaning that multiple countries may be listed as production origins. To standardize this information, country codes are converted into full country names. Certain historical country codes, such as `SU` (Soviet Union), are updated to their modern equivalents, such as `RU` (Russia), to maintain consistency.

For movies produced in multiple countries, the dataset is further refined by splitting entries into separate rows, ensuring that each country is represented individually. This transformation is particularly important for accurately computing average `profit` and `ROI` by country, as it allows each country's contribution to be analyzed independently rather than being grouped into a single, unstructured list.

By implementing these data cleaning and processing steps, the dataset is prepared for further exploratory and predictive analysis, ensuring that all financial metrics and country-based insights are based on a well-structured and reliable foundation.

3.2.2. *Analysis of the Top 10 Regions by Average Profit and ROI.*

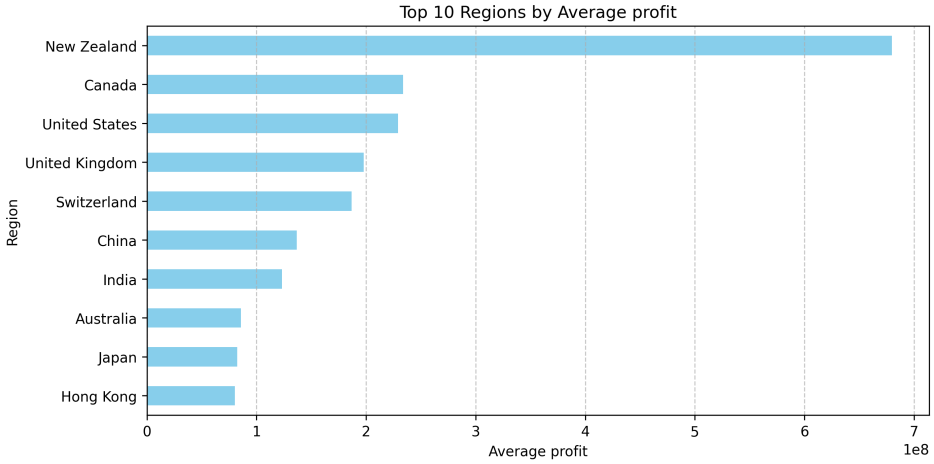

FIGURE 3. Top 10 Regions of Average Profit

Figure 3 presents the top ten regions based on average profit. The results indicate that New Zealand has the highest average profit, significantly surpassing other regions. This pattern suggests that a few exceptionally successful films, such as *The Lord of the Rings* and *The Hobbit* franchises, heavily influence the country's average. Canada and the United States follow, reflecting the dominance of Hollywood and its collaborations with Canadian productions. The United Kingdom and Switzerland also rank high, benefiting from strong domestic and international markets. China and India, despite having massive film industries, exhibit lower average profits, suggesting that while their markets are extensive, they may produce a high volume of mid-range or lower-profit films. The appearance of Australia, Japan, and Hong Kong in the ranking further emphasizes the role of regional industries that occasionally produce highly profitable films.

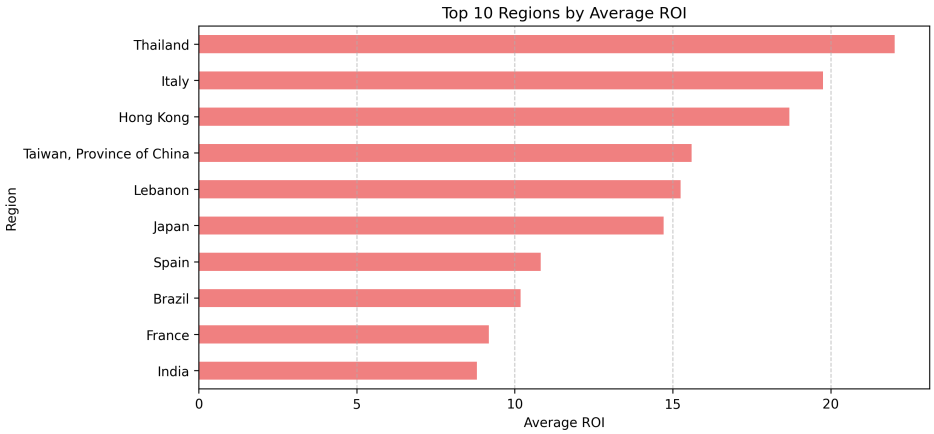

FIGURE 4. Top 10 Regions of Average ROI

Figure 4 illustrates the top ten regions by average Return on Investment (`ROI`). Thailand leads with the highest `ROI`, indicating that Thai films are often produced on relatively low budgets but achieve substantial financial returns. This trend may be due to the popularity of local genres, such as horror and action films, which tend to perform well both domestically and internationally. Italy and Hong Kong follow closely, showing that their film industries generate high returns relative to their investment. Taiwan, Lebanon, and Japan also rank high, highlighting their ability to produce financially efficient films that cater to domestic audiences while gaining niche appeal abroad. Spain, Brazil, and France appear in the ranking, indicating the profitability of European and Latin American cinema. India, despite its massive film output, rounds out the list, reflecting the strong financial returns of certain Indian films despite the presence of many lower-budget productions.

From this analysis, several insights emerge. High `profit` does not necessarily translate into high `ROI`, as seen in the United States and Canada, where large budgets contribute to high absolute `profit` but lower proportional returns. Conversely, countries like Thailand, Taiwan, and Lebanon demonstrate that smaller film industries can achieve significant `ROI` by efficiently managing production costs. New Zealand's dominance in average `profit` underscores how a small number of highly successful films can disproportionately impact national statistics, while the diverse presence of European, Asian, and Latin American countries in the `ROI` ranking highlights the importance of budget efficiency and market strategy in film profitability.

3.2.3. *Analysis of TMDB Ratings vs Profit and ROI.*

TMDB Ratings vs ROI (Colored by Country) suggests that higher `rating` do not always correspond to higher `profit`. Some highly rated films generate substantial `profit`, but there are also low-rated films with significant earnings. A cluster of movies with moderate to high `rating` above 7.0 appears to achieve positive `profit`, suggesting that critically acclaimed movies tend to be financially successful. However, some highly rated films fail to generate large `profit`, indicating that positive reception alone does not guarantee financial success. Additionally, some films with `rating` below 5.0 still manage to generate high `profit`, which implies that certain commercial or franchise-based movies attract audiences and secure strong box office earnings despite poor reviews.

TMDB Ratings vs Profit (Colored by Country) reveals a wide variation in `ROI` across different `rating`. Some low-budget films achieve exceptionally high `ROI`, particularly in the mid-to-low `rating` range. Movies with `rating` between 5.0 and 7.0 tend to have the highest `ROI`, suggesting that a film does not need to be critically acclaimed to be a financially successful investment. This trend aligns with the success of certain genre films, such as horror and comedy, which often have low production costs but significant box office appeal. In contrast, some highly rated films show lower `ROI`, likely due to high production and marketing costs that reduce proportional returns despite strong earnings.

From this analysis, several insights emerge. A film's financial success does not solely depend on its `rating`, as low-rated commercial films still achieve massive box office success. The highest `ROI` often appears in films with moderate `rating`, indicating that moderately received films with controlled budgets tend to be more financially efficient investments. Big-budget productions, even when they generate high `profit`, often exhibit lower `ROI` due to high initial costs and extensive marketing

expenses. This suggests that a movie's profitability depends not only on its `rating` but also on its production budget, genre, and audience appeal.

3.3. **Visualization of Production Countries.**

The visualization part provides clear insights into the geographic distribution of top-rated movies's production. To achieve this, two distinct types of visualization are created: one for the top-rated 500 movies and another for top-rated 200 movies by genre. These visualizations highlight global trends in movie production while offering insights into genre-specific patterns.

3.3.1. *Data Cleaning and Preparation.*

Before visualization, data cleaning and transformation are conducted to ensure accuracy and consistency across multiple data sources.

Since TMDb provides production country data in ISO Alpha-2 format, adjustments are required to ensure compatibility with geographic visualization tools. Like the processing in our previous part, outdated country codes are carefully mapped to modern equivalents. For instance, Soviet Union (`SU`) data is reassigned to Russia (`RU`), while Czechoslovakia (`XC`)'s records are split between the Czech Republic (`CZ`) and Slovakia (`SK`). Additionally, the REST Countries API is used to convert ISO Alpha-2 codes into full country names and ISO Alpha-3 codes to align with mapping libraries.

For the top-rated 500 movies, the count is calculated by iterating through each movie's production countries and incrementing the respective country's total. For the top-rated 200 movies by genre, counts are calculated in a similar manner but are grouped by both country and genre to allow genre-specific visualization.

3.3.2. *Top-Rated 500 Movies.*

The first visualization presents a static choropleth map, shown as **Figure 5** drawn from GeoPandas, that illustrates the number of top-rated movies produced in each country. This map helps identify major contributors to the global film industry and reveals regional disparities in cinematic output.

Since film production data is highly skewed, with certain countries like the United States dominating the count, logarithmic binning is applied to improve visual balance. This technique ensures that smaller contributors remain visible, preventing the visualization from being overwhelmed by a few high-output countries.

Figure 5 uses a red color gradient, where deeper shades indicate a higher number of movies. This result reveals that the United States dominates top-rated film production, reflecting Hollywood's excellence in film production. Western Europe, particularly the United Kingdom, France, and Germany, also shows significant contributions, while Japan and China stand out in East Asia. Brazil emerges as a notable player in South America. In contrast, parts of Africa and Central Asia exhibit minimal representation, indicating limited film industry presence or international recognition.

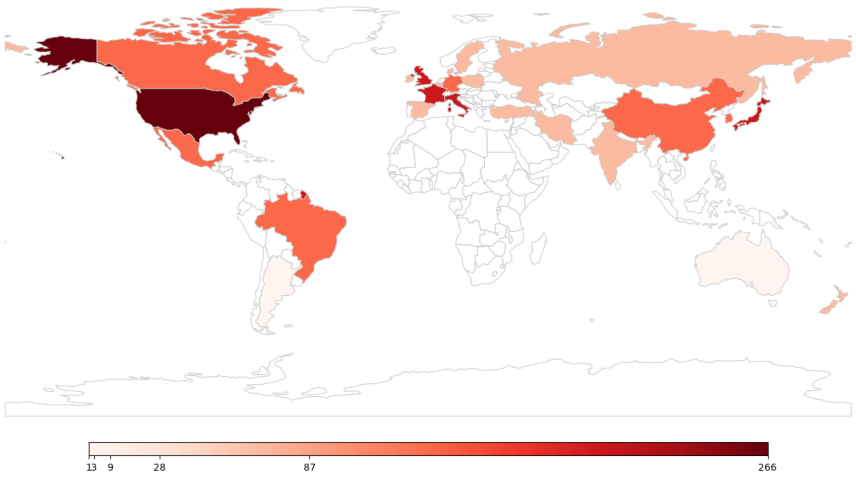

FIGURE 5. Global Distribution of Top-Rated 500 Movies' Production Locations

3.3.3. *Top-Rated 200 Movies by Genre.*

To analyze genre-specific trends, an interactive choropleth map is developed using Folium and is enhanced with geographic data sourced from https://raw.githubusercontent.com/python-visualization/folium/master/examples/data/world-countries.json, ensuring comparatively accurate visualization of country boundaries. This visualization enables users to switch between 17 different genres, observing how film production varies geographically across genres.

The interactive map is available for exploration at the following link: Global Distribution of Top-Rated Movies by Genre. The dataset is processed by aggregating the number of movies in each genre for each country. Each genre is represented by a unique color scheme to enhance visual distinction. To emphasize regions where certain genres are absent or underrepresented, countries with zero movies in the selected genre are shown in black.

We can take the action movies as an example. The United States has the highest number of recognized action films, followed by the United Kingdom and Japan. China, South Korea, and France also show notable contributions. Other regions, such as Canada, New Zealand and Brazil, have moderate representation. Large parts of Eastern Europe, Africa, and South Asia have little to no presence in this genre.

In general, our key observations are that the United States dominates most genres, reflecting its influential film industry, while Europe excels in drama, crime, and historical films with a focus on artistic storytelling. Japan and South Korea stand out in specialized genres like animation, horror, and mystery. China demonstrates significant contributions in action and historical films, showcasing its growing presence in the global movie market. Additionally, countries such as Brazil, India, and New Zealand show notable contributions, highlighting the expanding global film landscape shaped by cultural preferences and industry strengths.

4. Conclusion and Discussion

References

Email address: 1

Email address: 1

Email address: 1