 ISSN: 3107-6513

A Comprehensive Data-Driven Analysis of the Global Animation Industry Using Integrated Multi-Source Datasets

Shivam Yadav[1], Sandhya Kaprawan[2]
[1]M.S.(Data Analytics),[2]Assistant Professor
University Department of Information Technology, University of Mumbai, Kalina, Maharashtra, India
[1]shivamyadav26037@gmail.com [2]sandhya.kaprawan@udit.mu.ac.in

Abstract:
This study explores the global animation industry using an integrated multi-source dataset combining films, anime, streaming platforms, and box-office data. Exploratory analysis reveals dominant genres, varying studio performance, strong audience engagement in anime despite lower budgets, and the growing impact of OTT platforms. The results highlight the value of multi-source data integration for deeper industry insights.

Keywords - Animation Analytics, Multi-Source Data Integration, Anime Industry, OTT Platforms, Box Office Analysis, Data-Driven Media Studies, Exploratory Data Analysis.

I. INTRODUCTION

Animation has evolved into one of the most influential and dynamic forms of digital media within the contemporary entertainment ecosystem. Traditionally perceived as content primarily targeted at children, animation has undergone a significant transformation and is now recognized as a powerful and versatile storytelling medium that appeals to diverse age groups, cultural backgrounds, and global audiences. This shift has been driven by rapid technological advancements, particularly in computer-generated imagery (CGI), visual effects, and digital rendering techniques, which have substantially enhanced narrative depth, visual realism, and creative expression in animated content.

In parallel, the emergence and rapid expansion of streaming platforms have fundamentally reshaped the production, distribution, and consumption patterns of animated media. Animation is no longer constrained by theatrical releases or regional broadcasting limitations; instead, it is distributed globally through over-the-top (OTT) platforms, enabling instant access to a wide variety of animated films, series, and anime. As a result, the animation industry has experienced accelerated global reach, increased content diversity, and evolving audience engagement behaviors.

Despite its growing economic value and cultural significance, systematic and large-scale analytical research on the animation industry remains relatively limited. Much of the existing literature concentrates on isolated aspects such as individual studios, specific animation genres, or regional markets. While these studies provide valuable localized insights, they often fail to capture the broader industry dynamics and interdependencies between creative quality, audience reception, financial performance, and distribution strategies. Consequently, the current body of research offers fragmented perspectives that do not fully represent the complexity of the global animation ecosystem.

This research identifies a critical gap in existing studies: the absence of a comprehensive, integrated, data-driven analytical framework that simultaneously examines animation films, anime, streaming platform availability, and financial performance indicators. To address this gap, the present study proposes a unified analytical approach that integrates multiple heterogeneous datasets into a single master dataset. The central thesis of this research is that multi-source data integration enables significantly deeper, more reliable, and more holistic insights into global animation industry trends when compared to traditional single-source analytical methods. Animation is not only a technological medium but

also a form of visual art and aesthetic expression, influencing audience perception and engagement [22].

II. LITERATURE REVIEW

Research on animation spans diverse domains, including media studies, psychology, economics, and computer graphics. Early studies emphasized narrative structures and visual storytelling, highlighting how colour, motion, and character design influence audience perception. Other works examined studio-specific success models, particularly focusing on storytelling consistency and technological innovation.

The anime industry has been widely studied as a cultural export phenomenon, with research emphasizing fan-driven ecosystems, narrative complexity, and community-based rating platforms. Meanwhile, studies on OTT platforms demonstrate how digital distribution reshapes audience behavior, enabling global access and personalized content discovery.

However, a common limitation across prior research is reliance on isolated datasets. Few studies attempt large-scale integration across Western animation, anime, financial metrics, and streaming data. This research builds upon existing theories while extending them through a unified data integration and comparative analytics framework.

2.1 A DATASET CONTEXT IN EXISTING LITERATURE

The foundation of this research lies in the integration of multiple heterogeneous datasets representing different segments of the global animation ecosystem. Unlike traditional animation studies that rely on a single database or platform, this research adopts a multi-source data integration approach to capture the complexity and diversity of the animation industry. By combining animation movie datasets, anime repositories, OTT platform records, and box-office financial data, the study constructs a unified analytical framework capable of supporting cross-domain and cross-platform analysis. Prior research has shown that multi-source integration significantly improves the reliability, robustness, and generalizability of analytical insights, particularly in media and entertainment analytics **[19].**

To ensure reliability of the integrated datasets, standard data quality assessment and preprocessing techniques were applied to handle missing values, duplicates, and inconsistencies [20].

The primary datasets used in this study include large-scale animation metadata from TMDb and IMDb-based repositories, anime-specific datasets from community-driven platforms such as My Anime List, streaming platform data from Netflix, and financial performance data from global box-office records **[1], [3], [4], [5].** These datasets collectively represent a broad spectrum of animated content, ranging from theatrical releases and television series to anime and digital-first productions. In addition, studio-specific datasets for major animation producers such as Pixar, Disney, and Studio Ghibli were incorporated to enable detailed studio-level comparisons of quality, popularity, and commercial performance **[6], [7], [8].**

2.2 Dataset Composition and Coverage

The integrated dataset is designed to capture multiple dimensions of the animation industry. It includes content produced across different time periods, geographic regions, and distribution channels. Western animation and Japanese anime are both represented, allowing for meaningful comparative analysis. Furthermore, the inclusion of OTT platform indicators enables the study to examine how digital distribution influences content visibility and audience engagement.

2.3 Dataset Attribute

The unified dataset contains several categories of variables, each serving a specific analytical purpose:

- **Content variables:** title, release year, genre, studio, country of origin
- **Audience variables:** average user rating, number of votes, popularity score
- **Financial variables:** production budget, worldwide box-office revenue, estimated profit
- **Platform variables:** OTT availability, platform type (theatrical, OTT, hybrid)

These variables allow joint analysis of creative, audience-driven, and economic aspects of animated content, which is not possible using isolated datasets.

III. METHODOLOGY

This research adopts a multi-stage analytical methodology specifically designed to handle heterogeneous data sources and enable robust, large-scale comparative analysis of the global animation industry. Given the diversity of data formats, platforms, and metrics involved, a structured methodological framework is essential to ensure analytical consistency and reliability. The overall framework consists of four key stages: data integration, feature engineering, exploratory data analysis (EDA), and comparative evaluation. Similar multi-stage frameworks have been recommended in prior data-driven media and entertainment studies to manage complexity and improve interpretability **[19], [17].**

3.1 Data Integration

Data integration represents the most critical stage of the methodology, as the datasets used in this study originate from multiple independent sources with inconsistent naming conventions, formats, and levels of granularity. To address these challenges, a hybrid record-linkage approach was adopted, combining exact record matching with fuzzy string-matching techniques.

Exact matching was first applied using normalized content titles and release years. Normalization involved converting text to lowercase, removing special characters, and standardizing whitespace to minimize formatting inconsistencies. However, due to variations in naming conventions across platforms such as abbreviations, alternative spellings, and localized titles exact matching alone was insufficient.

To address this limitation, fuzzy string matching was employed to identify probable matches between records with minor textual differences . Fuzzy matching improves linkage accuracy in multi-source datasets by calculating similarity scores between text strings and selecting matches that exceed a defined threshold. This hybrid integration strategy has been shown to reduce data loss and enhance linkage quality in large-scale analytics applications **[19].** The integration of multiple heterogeneous data sources follows established data integration principles widely adopted in analytical research [21].

3.2 Feature Engineering

Following data integration, feature engineering was performed to transform raw variables into analytically meaningful indicators. Feature engineering plays a crucial role in enhancing analytical depth by enabling interpretation beyond surface-level attributes **[17].**

The following derived variables were created:

- Profitability, calculated as the difference between worldwide revenue and production budget, to assess economic efficiency
- Decade category, derived from the release year, to analyze long-term temporal trends
- Content type, categorizing entries as Anime or Western Animation, to support comparative analysis
- Platform indicator, classifying content as OTT-based or non-OTT, to evaluate the influence of streaming platforms

These engineered features enable joint analysis of creative quality, audience reception, financial performance, and distribution strategy, which is not possible using raw data alone.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to uncover underlying patterns, trends, and relationships within the integrated dataset. EDA is widely recognized as a foundational step in data-driven research, particularly in domains involving complex, high-dimensional datasets such as media analytics **[11].**

The EDA process included:

- Descriptive statistical analysis to summarize ratings, budgets, revenues, and vote distributions
- Temporal trend analysis to examine growth patterns across decades
- Genre-wise and studio-wise aggregation to identify dominant content categories

- Visual analysis using charts and plots to reveal relationships between budget, revenue, and ratings

Through EDA, the study identified key structural patterns, such as genre dominance, differences in audience engagement between anime and Western animation, and shifts in content production following the rise of OTT platforms.

3.4 Comparative Framework

The final stage of the methodology involved a comparative analytical framework designed to evaluate the effectiveness of traditional single-dataset analysis against the proposed integrated multi-source approach. Traditional analytical methods typically rely on isolated datasets, limiting their ability to capture cross-domain interactions **[11].**

In this study, insights derived from single-source analyses were compared with those obtained from the integrated master dataset. Performance was evaluated based on:

- Insight richness, measured by the number and diversity of observable relationships
- Analytical accuracy, reflected in consistency across platforms and variables
- Interpretability, assessed by the ability to explain observed trends holistically

This comparative evaluation demonstrates how integrated multi-source analytics provide superior explanatory power and more actionable insights, despite increased preprocessing complexity. Such trade-offs are commonly discussed in large-scale data integration research **[19].**

IV. ASSUMED DATA

Large-scale data-driven studies that rely on publicly available datasets often encounter challenges such as missing values, inconsistent measurement scales, and incomplete metadata. To address these issues and maintain analytical consistency across heterogeneous data sources, this research applies a set of controlled assumptions. The use of assumptions in such contexts is a well-established practice in data analytics and empirical media research, particularly when working with multi-source datasets **[19].**

One of the most significant challenges encountered in this study relates to incomplete financial data. Production budget information was missing for a subset of animated films and anime titles, especially older productions and region-specific releases. To mitigate this limitation, missing budget values were estimated using average budget ranges calculated within the same genre and decade. This approach reduces distortion by accounting for historical and genre-specific production cost patterns rather than applying a single global average **[11].**

Another challenge involved differences in audience rating systems across platforms. Ratings sourced from different databases are often measured on varying scales and reflect distinct user communities. To enable direct comparison, ratings were normalized to a common numerical scale. Normalization ensures that variations in platform-specific scoring systems do not bias comparative analysis and is a standard practice in cross-platform media analytics **[17].**

With respect to distribution channels, OTT availability was modeled as a binary variable, indicating whether a title was available on at least one major streaming platform at the time of data collection. While this simplification does not capture variations in regional availability or exclusivity, it allows for clear evaluation of the general impact of streaming platforms on content visibility and audience engagement **[16].**

Many animated titles are associated with multiple genres, which can complicate genre-based aggregation and trend analysis. To maintain consistency, multi-genre entries were reduced to a primary genre, defined as the first-listed or most dominant genre. This assumption simplifies categorical analysis while preserving the core thematic classification of each title **[11].**

Although these assumptions enable meaningful cross-domain comparisons and reduce analytical noise, they may introduce minor estimation errors or oversimplifications. These limitations are acknowledged and considered when interpreting results. Recognizing and explicitly documenting assumptions enhances the transparency and credibility of data-driven research **[19].**

V. ANALYTICAL COMPARISON

The core contribution of this research lies in its comparative evaluation of traditional single-source analytical methods and the proposed multi-source integrated analytical approach. By systematically contrasting these two methodologies, the study demonstrates how data integration significantly enhances analytical depth, interpretability, and decision-making relevance in the context of the global animation industry. Comparative analysis is particularly important in media analytics, where fragmented data sources often obscure industry-wide patterns **[19].** OTT platforms have significantly transformed content distribution models by enabling global access and reshaping audience viewing behavior [23].

5.1 Traditional Single-Source Analytical Approach

Traditional animation industry studies typically rely on data obtained from a single platform, studio, or content repository. Such approaches are computationally simple and require minimal preprocessing, making them attractive for small-scale or exploratory research. However, their analytical scope is inherently limited. By focusing on a single source, these studies often capture only a narrow segment of the animation ecosystem, such as audience ratings without financial context or box-office performance without platform-level insights **[11].**

A major limitation of the traditional approach is its inability to capture cross-platform and cross-domain relationships. For instance, analyzing audience ratings independently of production budgets and distribution channels may lead to incomplete or misleading conclusions regarding content quality or success. Similarly, studio-specific analyses fail to account for competitive dynamics and comparative performance across the broader industry. As a result, traditional single-source studies provide fragmented insights that lack holistic explanatory power **[11].**

5.2 Proposed Multi-Source Integrated Analytical Approach

In contrast, the proposed multi-source integrated approach combines datasets from animation films, anime repositories, OTT platforms, and box-office records into a unified analytical framework. This integration enables comprehensive cross-studio and cross-platform comparisons that are not possible using isolated datasets (Zhao, 2021).

The integrated framework supports:

- Cross-studio and cross-platform comparisons, allowing evaluation of relative performance across different production houses and distribution channels
- Joint analysis of audience ratings, financial performance, and platform distribution, providing a multidimensional view of success
- Identification of efficiency metrics, such as return on investment, by linking production budgets with audience reception and revenue

Empirical findings derived from the integrated dataset reveal that anime titles frequently achieve higher audience ratings despite substantially lower production budgets, indicating greater efficiency in content creation. Conversely, several high-budget Western animated productions fail to secure proportionally higher ratings, suggesting diminishing returns on excessive production spending. Such insights remain largely invisible in single-source analyses and highlight the value of multi-source integration **[12], [13].**

5.3 Performance Evaluation and Trade-Off Analysis

A comparative evaluation shows that the integrated multi-source approach performs better than traditional single-source analysis across several key aspects. By combining creative, financial, and distribution-related variables, this method provides more complete and meaningful insights and helps in identifying long-term industry trends such as the growing influence of anime and OTT platforms [14]. In addition, the integrated framework offers stronger decision-support for studios, streaming platforms, and policymakers. Although the integration process requires additional preprocessing effort, the improvement in insight quality and analytical depth justifies this complexity [19]. Overall, the results confirm that multi-source analytics are more effective for analyzing complex industries like global animation.

VI. CONCLUSION

This study provides a detailed data-driven examination of the global animation industry using an integrated multi-source dataset. By combining information from animation films, anime databases, OTT platforms, and global box-office records, the research moves beyond the limitations of single-dataset studies and offers a broader view of industry trends. The integrated framework allows creative quality, audience response, financial performance, and distribution strategies to be analyzed together, which is difficult to achieve when data sources are studied in isolation.

The findings reveal several important patterns within the animation industry. Fantasy and adventure genres continue to dominate global animation production, indicating a consistent audience preference for visually rich and immersive storytelling. The analysis also shows that anime often achieves strong audience engagement and positive reception despite operating with comparatively lower production budgets. This suggests that effective storytelling, character development, and serialized narratives play a crucial role in anime's success. In contrast, higher production budgets in Western animation do not always result in better audience ratings, indicating that excessive financial investment does not necessarily guarantee higher appreciation.

The study further highlights the growing influence of OTT platforms on animation consumption and distribution. Streaming services have significantly expanded global access to animated content, accelerated the worldwide reach of anime, and reduced dependence on traditional theatrical release models. Studio-level observations suggest that while Pixar and Studio Ghibli consistently deliver high-quality content that resonates with audiences, Disney continues to lead in commercial performance, reflecting different creative and business strategies adopted by these studios.

6.1 Limitations

Although the study offers valuable insights, certain limitations must be acknowledged. The analysis is based on publicly available datasets, which may contain missing or inconsistent information. In some cases, production budget data had to be estimated, and OTT availability was represented using simplified indicators. These assumptions may introduce minor inaccuracies; however, such challenges are common in large-scale data analytics and were addressed through careful preprocessing and transparent methodological choices.

6.2 Future Work

Future research can further extend this work in multiple directions. Sentiment analysis of audience reviews can be incorporated to better understand viewer opinions and emotional responses. Machine learning or deep learning techniques may also be used to predict ratings, revenue, or overall content success. In addition, the use of real-time streaming data and region-wise audience analysis can help in understanding changing global consumption patterns more accurately. Overall, this study highlights the effectiveness of multi-source data integration for analyzing the global animation industry and provides a solid foundation for future data-driven research in media and entertainment analytics.

REFERENCES

[1] Asaniczka, M. (2024). *52,000 Animation Movie Details Dataset*. Kaggle.
[2] Abhishm, K. (2023). *20K Animated Movies and TV Shows Dataset*. Kaggle.
[3] Hernan4444. (2021). *MyAnimeList Anime Database*. GitHub.
[4] Chirumamilla, B. (2025). *Netflix Movies and TV Shows Dataset*. Kaggle.
[5] Aditya126. (2024). *Movies Box Office Dataset (2000–2024)*. Kaggle.
[6] Prasert, K. (2020a). *Pixar Movies Dataset*. Kaggle.
[7] Prasert, K. (2020b). *Disney Movies Dataset*. Kaggle.
[8] Ehallmar, E. (2021). *Studio Ghibli Films Dataset*. Kaggle.
[9] Kim, H. (2019). Visual psychology in animation: How design and color influence audience perception. *Journal of Media Psychology*, 12(3), 45–60.

 ISSN: **3107-6513**

[10] Price, D. (2018). The art of Pixar storytelling: Narrative depth in modern animation. *Animation Studies Review*, 15(2), 78–95.
[11] Wells, P. (2016). *Understanding Animation*. Routledge.
[12] Napier, S. J. (2018). *Anime from Akira to Howl's Moving Castle*. Palgrave Macmillan.
[13] Cavallaro, D. (2015). *Anime and the Visual Arts*. McFarland & Company.
[14] Statista Research Department. (2023). *Global animation industry market size and forecast*. Statista.
[15] Motion Picture Association. (2022). *Theme Report: Global Streaming and Animation Trends*. MPA.
[16] Netflix Media Center. (2023). *Global animation and anime consumption trends*. Netflix.
[17] Deloitte. (2022). *Digital media trends: Animation and OTT consumption*. Deloitte Insights.
[18] Box Office Mojo. (2024). *Animated film box office performance*.
[19] Wang, H., Li, Y., Zhang, Y., and Zhou, X. (2020). *A survey on data integration: Challenges, techniques, and applications.* IEEE Access, Vol. 8, pp. 213012–213031.
[20] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 1–52.
[21] Lenzerini, M. (2002). Data integration: A theoretical perspective. *Proceedings of PODS*, 233–246.
[22] Furniss, M. (2014). *Art in motion: Animation aesthetics*. Indiana University Press.
[23] Lotz, A. D. (2017). *Portals: A treatise on internet-distributed television*.
Michigan Publishing.