Best Streaming Service Analysis: Exploratory Data Analysis and Data Visualization

Dr. R. John Mathew [1]D. Harsha Vardhini[2], K. Hyma Sri[3], A. Durga Maha Lakshmi[4], A. Aditya Subrahmanya Varma[5]
[1]*Professor, DEPT-CSE, DEPT-CSE, Srinivasa Institute of Engineering and Technology*
[2345]*UG Scholar, DEPT-CSE, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-194827-459

Abstract— Due to the competitive ecosystem created by the quick growth of digital streaming services, consumers now have to elect from a variety of platforms with a wide range of features and content libraries. This study compares the main OTT platforms using a data- driven logical frame that blends unsupervised machine literacy, visualization styles, and exploratory data analysis (EDA). To find patterns in happy vacuity, release trends, followership groups, and standing distributions, a dataset comprising movie- related attributes from top services like Netflix, Amazon Prime Video, Hulu, and Disney is examined. To guarantee logical delicacy, the dataset is pre reused using point metamorphosis, statistical confirmation, and cleaning. To probe platform-specific content distribution and standing geste, exploratory analysis is carried out. also, pictures are grouped using K-Means clustering according to parallels in release time, age group, conditions, and platform presence. The figure score is used to assess the quality of the clustering. An interactive web- grounded dashboard and relative visualizations are used to present the logical results. The findings show how statistical analysis and machine literacy ways support well- informed decision- making in the digital streaming ecosystem by offering structured perceptivity into platform specialization and content patterns.

Keywords: Data analytics, Data visualization, Exploratory Data Analysis, OTT platforms, Streaming services.

I. INTRODUCTION

Particularly through the growth of Over-The-Top (OTT) streaming services, the fast development of digital technologies has profoundly changed the entertainment sector. Providing on-demand access to large digital collections spanning several genres, languages, and audience groups, Netflix, Amazon Prime Video, Hulu, and Disney+ have transformed content consumption. Users have trouble choosing the most appropriate service based on content quality, availability, rankings, and audience tastes as competition among streaming providers heats up.

Data analysis has become a vital tool to comprehend customer behaviour and platform trends as digital content grows exponentially. Through exploratory data analysis (EDA), datasets may be systematically examined to expose trends, distributions, and connections between variables. When paired with visualization methods, analytical findings can be more properly understood, therefore enabling stakeholders to gain insightful insights from intricate datasets.

Using a data-driven approach, this study seeks to comparatively evaluate top OTT services. To find content trends and distribution patterns, a dataset of movie-related characteristics including release year, age classification, ratings, and platform availability is examined. Statistical preprocessing methods are used to guarantee data reliability and consistency. Beyond investigative research, unsupervised machine learning methods are used to increase analytical depth. Particularly used to group films according on attribute similarities, K-Means clustering facilitates organized categorization of material on various platforms. Performance indicators are used to assess the clustering results to guarantee relevant grouping.

Comparative visuals and an interactive web-based dashboard show the results of the analysis. This study shows how data analytics can aid educated decision-making inside the competitive digital streaming environment by combining statistical investigation, visual approaches, and machine learning algorithms.

II. LITERATURE REVIEW

Many studies have stressed how critical data analysis is in grasping the nature of digital media platforms. Exploratory Data Analysis has often been used to examine patterns in enormous datasets and find trends affecting user behaviour.

Earlier studies on streaming platforms mostly concentrate on sentiment analysis, recommendations systems, or user interaction research. These investigations look at how users engage with material or how platforms suggest movies.

But somewhat fewer studies use data-driven techniques to examine the platforms themselves. Using dataset-based analysis to grasp platform performance can help to reveal audience targeting, content diversity, and platform strategy.

The current research adds to the body of knowledge by using exploratory data analysis and visual methods to compare OTT platforms at once. Further more combining unsupervised machine learning techniques like K-means clustering offers more in-depth understanding of platforms features and content categorization. This organized analytical approach improves methodical interpretation of streaming data and enables wise decisions in the digital streaming environment.

III. DATASET DESCRIPTION

The Movies_clustered.csv, the dataset utilized in this analysis, includes organized movie-level data acquired from significant OTT channels like Netflix, Hulu, Prime Video, and Disney+. While the columns record descriptive, numerical, and derived analytical characteristics utilized for clustering and exploratory data analysis, each row in the dataset reflects a separate movie. ID, title, release year, age certification category, and Rotten Tomatoes rating are among the features found in the dataset. Originally given in string format (e.g., "98/100"), the Rotten Tomatoes rating is later changed into a numerical variable called RT_Score to allow for quantitative analysis.

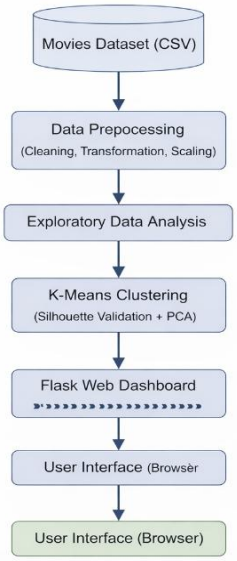
Binary indicator columns—Netflix, Hulu, Prime Video, Disney+—represent platform availability; a 1 means the film is present on the particular platform and a 0 means it is not. This framework enables methodical assessment of cross-platform overlap, content uniqueness, and distribution patterns. Included to help differentiate between exclusive titles and multi-platform titles, a derived feature known as Platform_Count counts the number of platforms on which a given film is accessible.

Through the Age_Num feature, categorical characteristics like Age classification are converted into numerical forms to help machine learning analysis. Among the data preprocessing stages are scaling chosen numerical characteristics, converting ratings into a consistent numerical form, treating inaccurate age labels, and cleaning missing values. Principal Component Analysis (PCA) is also used to reduce dimensionality, which leads to two new components called PCA1 and PCA2. These components are then used to show clusters and make sense of patterns in a smaller space with fewer features.

Also present in the dataset is a Cluster column produced using the K-Means clustering technique. Release year, rating score, age group, and platform presence are a few of the traits used to categorize films. The clustering process helps to find platform-specific content traits as well as segments of content. The dataset generally offers a multi-dimensional analytical base that encourages unsupervised machine learning, comparative platform analysis, and exploratory visualization inside the digital streaming environment.

IV. SYSTEM ARCHITECTURE

The suggested system design processes, analyses, clusters, and visualizes OTT streaming platform data using a modular approach.



Starting with the Data Source Layer, the system reads in a structured CSV dataset with movie features like release year, age group, Rotten Tomatoes rating, and platform availability (Netflix, Hulu, Prime Video, Disney+).

The data is then cleaned and transformed in the Data Preprocessing Layer. To get data ready for clustering, ratings are turned into numbers, age groups are turned into numbers, new features like Platform_Count are added, and standardization is used. Using K-Means clustering, the Analytical Layer organizes movies according to similarities in characteristics including year, rating, age group, and platform presence and conducts exploratory data analysis (EDA). The silhouette score measures the quality of a cluster; PCA helps to improve visual representation.

Using a Flask-based Dashboard Layer, interactive graphs showing summary statistics, comparative analysis, and clustering results are used to present the data. The architecture generally guarantees a planned flow from user-friendly web presentation to data input to machine learning analysis.

IV. METHODOLOGY

Using This study assesses and contrasts top internet streaming services using a methodical, data-driven analytical framework based on exploratory data analysis (EDA), data visualisation, and unsupervised machine learning methods. The approach is meant to guarantee logical data processing, correct statistical analysis, significant clustering, and sensible interpretation of content distribution patterns across channels.

The first step is to gather organized movie-level data on OTT streaming services like Netflix, Hulu, Prime Video, and Disney+, from openly accessible sources. The dataset includes platform availability indicators, movie title, release year, age certification category, and Rotten Tomatoes rating. Every film is shown by binary variables showing whether it is accessible on a given platform. This organized depiction allows for multi-dimensional comparative analysis among streaming platforms.

Data preprocessing and transformation compromise the second stage. Many preprocessing techniques are used since unprocessed data can include mixed data types and conflicting formats, for quantitative comparison, Rotten Tomatoes ratings are translated into uniform numerical form (RT Score). For computer analysis, categories of age certification are converted into numerical values (Age_Num). Generated from the number of platforms hosting each film, a derived feature called Platform_Count helps differentiate exclusive titles from multi-platform material. Feature standardization using z-score normalization is carried out to remove scale variations between features including release year and rating scores:

$$Z = \frac{x - \mu}{\sigma}$$

where X is the original feature value, μ is the mean, and σ is the standard deviation. This transformation ensures equal contribution of all variables during clustering.

After preprocessing, an exploratory data analysis (EDA) is done to find out the dataset's structural features. To analyse platform-wise content distribution, rating behaviour, release year trends, and audience segmentation patterns, statistical methods including descriptive statistics, distribution analysis, and trend analysis are used. EDA enables the discovery of changes in content specialization and platform tactics. K-Means clustering is used as an unsupervised learning method to find secret patterns in the data. Release year, age category, rating score, and platform availability are among the traits used to classify films. K-Means aims to reduce the Within-Cluster Sum of Squares (WCSS), which is defined as:

$$J = \sum_{i=1}^{k} \sum_{i=1}^{k} \sum_{x \in C_i}^{k} |x - \mu_i||^2$$

where k represents the number of clusters, C_i denotes cluster i, and μ_i represents its centroid. The algorithm iteratively assigns movies to the nearest centroid based on Euclidean distance and updates centroid positions until convergence.

To evaluate clustering quality, the Silhouette Score is computed using:

$$S = \frac{max(a,b)}{b - a}$$

where a is the average intra-cluster distance and b is the average nearest-cluster distance. A higher silhouette value indicates well-separated and cohesive clusters.

Principal Component Analysis (PCA) is applied for dimensionality reduction and visualisation since

several connected features are present. PCA projects the feature space into orthogonal components capturing utmost variance.

Visualizing clusters in two-dimensional space helps to increase interpretability by extracting the first two major components and retaining crucial variance information.

Enhancement of interpretability is accomplished by means of visualization methods including bar charts, distribution plots, and cluster scatter plots. Based on content distribution, rating behaviour, and cluster segmentation, these visual tools help to simplify difficult data structures and enable comparative analysis across streaming services.

The web-based dashboard built with Flask finally uses the analytical findings. The dashboard lets you interactively explore streaming platform data by providing organized interpretation, summary statistics, platform-wise comparisons, cluster profiles, and API-driven outputs.

The general study provides observations on content focusing, rating trends, platform crossover, and competitive positioning of streaming services. The approach offers a methodical structure for data-driven evaluation of digital streaming systems.

V. EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION

This Before using clustering algorithms, exploratory data analysis (EDA) is done to grasp the structural and statistical features of the OTT movies dataset. In this work, EDA seeks to find patterns of audience segmentation, cross-platform availability, platform-wise content distribution, rating behaviour, and release trends. Quantitatively interpreting the data requires statistical summaries including count analysis, frequency distribution, median, and mean.

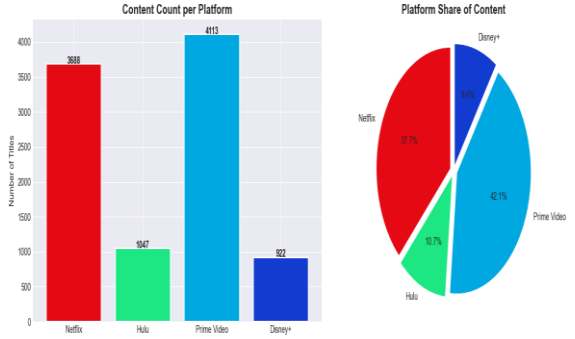


Figure-1: Platform-wise content distribution

Originally, the overall count of films accessible on Netflix, Hulu, Prime Video, and Disney+ is used to evaluate platform-wise content distribution. This study clarifies platform specialization and reveals which platform houses the most content. The comparative distribution exposes variations in streaming services' content concentration.

Rating analysis is also done utilizing the RT_Score, which is the standard Rotten Tomatoes score. The mean rating every platform helps to assess trends in content quality. Distribution charts help one to see whether platforms concentrate on mostly rated films or keep a range of ratings.

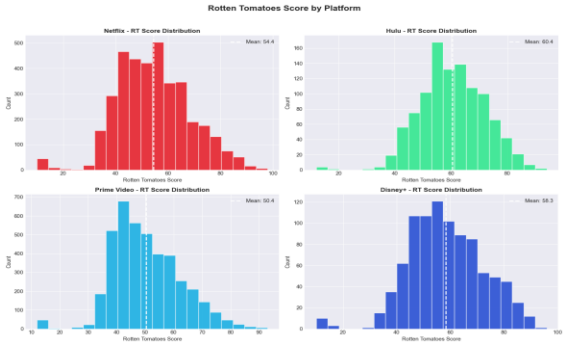


Figure-2: Average RT Rating per Platform

Release year analysis helps us to identify temporal trends in the availability of material. Looking at the minimum, maximum, and average release years helps the research determine if platforms stress current material or keep classic catalogues. Yearly trend visualization shows acquisition and production patterns.

The Age categorization criteria helps with audience segmentation analysis. Understanding the intended audience of each platform helps one to guide frequency distribution of age groups (such as 7+, 13+, 16+, 18+). This makes it easier to tell if a platform mostly targets family material, youth-oriented films, or mature viewers.

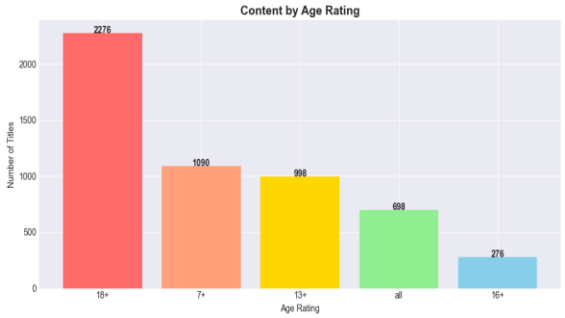


Figure 3: Audience Age Classification Distribution

The Platform_Count tool assesses cross-platform overlap. This study clarifies content sharing patterns across streaming services and separates unique content from multi-platform material so facilitating knowledge of exclusive content.

Many visual approaches help to improve readability. Age distribution analysis and platform content comparison are done using bar charts. Box plots and distribution plots show rating behavior across several platforms. PCA1 and PCA2 scatter plots help to show cluster separation in a lower dimensional space. These graphic depictions help to clarify complicated multi-dimensional relationships and give a clear grasp of the features of streaming platforms.

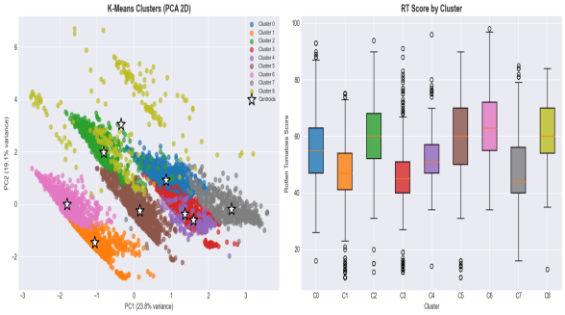


Figure 4: K-Means Clustering Visualization using Principal Component Analysis

Generally speaking, EDA and visualization approaches give organized understanding of audience segmentation, platform differentiation, rating patterns, and content distribution. These observations set the groundwork for following comparative evaluation and clustering study.

VI. COMPARATIVE ANALYSIS OF STREAMING PLATFORMS

The comparative analysis examines differences among major OTT platforms—Netflix, Hulu, Prime Video, and Disney+—in terms of content volume, rating behaviour, release trends, audience segmentation, and platform exclusivity. The objective is to identify structural differences and specialization strategies within the digital streaming ecosystem.

Platform-wise content distribution highlights variations in total movie count, indicating differences in content acquisition strategies and market positioning. Rating analysis using Rotten Tomatoes scores shows variation in average ratings and distribution patterns, reflecting differences in content quality focus and selection standards.

Temporal analysis based on release year reveals that some platforms emphasize recent releases, while others maintain a broader mix of classic and older titles.

Audience segmentation through age classification further indicates strategic targeting, with certain platforms focusing more on family-oriented content and others on mature audiences.

Cross-platform availability analysis using the Platform_Count feature shows that a significant portion of content remains platform-exclusive, strengthening competitive differentiation.

Additionally, K-Means clustering groups movies based on release year, rating, age category, and platform presence, revealing structured content segments across platforms. The silhouette score confirms the statistical validity of these clusters.

Overall, the comparative study demonstrates clear differences in content strategy, audience targeting, and platform specialization, supported by statistical analysis, clustering, and visualization techniques.

VII. CONCLUSION

Using Exploratory Data Analysis (EDA), visualisations, and unsupervised machine learning techniques, this study offers a structured data-driven examination of leading OTT streaming platforms. Analysing movie-level features including release year, age category, rating score, and platform availability, the study offers insightful analysis of content distribution trends and platform differentiation techniques inside the digital streaming scene.

The exploratory research shows clear changes in content volume, rating patterns, audience segmentation, and temporal distribution among channels. These variations point to different strategic decisions made by streaming platforms regarding audience targeting, competitive positioning, and content acquisition. Release trend analysis reveals variations in emphasis toward recent versus archival material, while rating analysis shows that content quality distribution differs across platforms.

Using K-Means clustering helps to improve the analytical framework by categorizing movies according to commonalities. The silhouette score-validated clustering findings show the existence of significant information sections in the data. By reducing multi-dimensional relationships into an understandable two-dimensional depiction, Principal Component Analysis (PCA) helps to properly view and interpret these groups.

Statistical analysis, machine learning approaches, and interactive dashboard visualization together offer a whole comparative evaluation framework for streaming services. The research generally shows how data-driven analytical techniques might help to enable methodical knowledge of platform specialization, content strategy, and competitive dynamics in the fast-changing OTT sector.

VIII. RESULTS AND DISCUSSION

The research shows that there are clear differences between the main OTT streaming platforms when it comes to how much content they have, how good the content is, who they are trying to reach, and how often they release new shows. A platform-based analysis reveals changes in general film availability, which mirror variations in market positioning and content sourcing approaches.

According to ratings analysis based on consistent Rotten Tomatoes ratings, there is moderate fluctuation across platforms. The link between rating and release year shows just a modest correlation, so content quality does not seem to rely much on the year of production. This points to a good balance of newer and older top-performing titles.

The increasing relevance of modern material in streaming libraries is reflected in the fact that temporal distribution shows more movies released in recent years. Further investigation of audience segmentation exposes platform-specific areas of attention, including differences noted in categories for mature material and family-oriented content.

Cross-platform availability evaluation shows that a significant number of films stay platform-exclusive, hence supporting competitive differentiation.

Using release year, rating score, age categorization, and platform availability, K-Means clustering efficiently divides films into several categories. PCA-based visualization shows significant separation among clusters, hence validating the clustering structure; the silhouette score supports this.

Generally, the findings underline that streaming services have unique structural features and content plans. Statistical analysis, clustering algorithms, and visualization together offer a thorough and data-driven approach for comparative assessment of digital streaming services.

REFERENCES

[1] J. W. Tukey, Exploratory Data Analysis, Reading, MA, USA: Addison-Wesley, 1977.

[2] W. McKinney, “Data structures for statistical computing in Python,” in Proc. 9th Python in Science Conf., Austin, TX, USA, 2010, pp. 51–56.

[3] T. Munzner, Visualization Analysis and Design, Boca Raton, FL, USA: CRC Press, 2014.

[4] A. McAfee and E. Brynjolfsson, “Big data: The management revolution,” Harvard Business Review, vol. 90, no. 10, pp. 60–68, Oct. 2012.

[5] C. Chen, “Information visualization and visual analytics,” IEEE Computer Graphics and Applications, vol. 25, no. 5, pp. 12–15, Sept.–Oct. 2005.

[6] S. Few, Now You See It: Simple Visualization Techniques for Quantitative Analysis, Oakland, CA, USA: Analytics Press, 2009.

[7] R. Sharda, D. Delen, and E. Turban, Business Intelligence and Analytics: Systems for Decision Support, 10th ed., Boston, MA, USA: Pearson, 2014.

[8] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Waltham, MA, USA: Morgan Kaufmann, 2011.

[9] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, New York, NY, USA: Springer, 2013.

[10] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, Mastering the Information Age: Solving Problems with Visual Analytics, Goslar, Germany: Eurographics Association, 2010.

[11] P. Chapman et al., “CRISP-DM 1.0: Step-bystep data mining guide,” SPSS Inc., Chicago, IL, USA, Tech. Rep., 2000.

[12] M. Chen, D. Ebert, H. Hagen, R. Laramee, and F. Mansmann, “Data, information, and knowledge in visualization,” IEEE Computer Graphics and

Applications, vol. 29, no. 1, pp. 12–19, Jan.–Feb. 2009.

[13] I. T. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, 2002.

[14] P. J. Rousseeuw, “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis,” Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.