From Indie to Blockbuster Films: A Clustering-Based Analysis of Movie Characteristics

Leon Marco Devela
College of Computing and Information Technologies
National University
Manila, Philippines
develala@students.national-u.edu.ph

Rae Paulos
College of Computing and Information Technologies
National University
Manila, Philippines
paulosrs@students.national-u.edu.ph

Abstract—**This study applies unsupervised machine learning to identify distinct segments within a comprehensive movie dataset, revealing temporal evolution patterns in cinema. Three clustering algorithms: K-Means, Hierarchical Clustering, and Gaussian Mixture Model (GMM), were compared using Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score, with GMM achieving superior performance on two of three metrics (Silhouette: 0.4750; Davies-Bouldin: 0.6824), demonstrating better ability to model the probabilistic structure of movie data. GMM identified three distinct clusters: Classic Acclaimed Cinema (12% of films, 50 years old), Mid-Era Balanced Films (41%, 25 years old), and Modern Commercial Blockbusters (47%, 15 years old). Analysis revealed an inverse relationship between popularity and critical acclaim, classic films show substantially higher ratings (0.4375) despite lower popularity (-0.0385), while modern films achieve high popularity (0.1165) with below-average ratings (-0.1235), with clusters also differing significantly in runtime (4.1×), homepage presence (5.1×), and digital marketing adoption. These findings have important implications for recommendation systems, marketing strategies, and streaming platform content acquisition, demonstrating that contemporary popularity metrics may not correlate with lasting artistic value and suggesting the need for era-aware approaches in film industry analysis.**

Index Terms—**Unsupervised learning, Gaussian Mixture Model, movie clustering, cinema evolution, popularity-quality tradeoff, recommendation systems, film industry analysis**

I. Introduction

The global film industry represents a substantial economic force, with the worldwide movies and entertainment market valued at approximately $105.8 billion in 2024 and projected to reach $143 billion by 2033 [1]. In the United States alone, the movie market is expected to grow from $23.44 billion in 2024 to $34.64 billion by 2033 [2]. Despite this tremendous economic value, the film industry faces a fundamental challenge in understanding what constitutes movie success and how different types of success manifest across cinema history. While some films achieve critical acclaim and lasting artistic recognition, others prioritize commercial performance and mass audience appeal, suggesting that success is not a singular construct but rather encompasses multiple, potentially competing dimensions [3].

Understanding how movies naturally segment based on their characteristics and which patterns distinguish different success profiles has become increasingly critical for industry stakeholders. Traditional approaches to analyzing movie success often assume a single definition of achievement, typically box office revenue or aggregate ratings and rely on supervised learning techniques that require predefined success labels [4]. However, this approach fails to capture the multidimensional nature of cinematic achievement, where critically acclaimed films may underperform commercially, and popular blockbusters may receive modest critical reception. Unsupervised learning methods, particularly clustering algorithms, offer a powerful alternative by discovering natural groupings and hidden patterns within movie datasets without imposing predetermined success categories, thereby revealing whether distinct success profiles emerge organically from the data [5].

This study addresses a critical research gap: *Do movies naturally cluster into distinct success profiles, and if so, how do these profiles differ in their characteristics and temporal distribution?* The studies investigation employs multiple unsupervised clustering techniques, K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMM) to analyze a comprehensive movies dataset from Kaggle [6], systematically comparing their performance using established validation metrics including Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score. By examining patterns in movie attributes including ratings, popularity, budget, runtime, digital presence, and temporal characteristics, the researchers investigate whether films segment into meaningful clusters that represent different types of achievement in the film industry [7].

The primary stakeholders impacted by this research include film production companies making strategic decisions about project positioning (critical vs. commercial focus), studio executives and investors requiring frameworks to evaluate diverse types of achievement, streaming platforms developing recommendation systems that must account for varied user preferences across the popularity-quality spectrum, and content acquisition teams building balanced catalogs spanning different success dimensions [5]. Additional beneficiaries include film critics and analysts seeking data-driven frameworks to understand industry evolution, and academic researchers studying entertainment economics and cultural shifts in audience preferences [8].

Clustering-based findings have practical applications across multiple domains: informing recommendation systems that match users to their preferred success dimension (critically acclaimed classics vs. popular blockbusters) [7], guiding production planning decisions by clarifying the tradeoffs between pursuing critical acclaim versus commercial appeal [9], and enabling competitive analysis that positions films relative to appropriate comparison groups rather than treating all cinema as homogeneous. Furthermore, the multidimensional success framework and temporal evolution patterns identified in this study extend beyond film to other entertainment media, including television series, music albums, and video games, providing a generalizable approach to understanding how different types of achievement manifest and evolve across creative industries.

II. Literature Review

The success of motion pictures has long been a subject of interest among researchers, movie critiques, and data analysts. Various studies have thoroughly explored the different factors that contribute to a film's overall performance and success, including production budget, genre, cast popularity, marketing strategies, release time, and audience reception. Clustering and analyzing films based on multiple attributes provides deeper insights into how these combinations of factors relate to different levels of movie success, thereby supporting more data-driven interpretations of cinematic performance.

A. The Role of Motion Picture to Society

Motion pictures play a significant role in society by serving as a medium for storytelling, cultural expression, and mass communication. Films reflect social values, beliefs, and issues, often shaping public perception and influencing attitudes toward cultural norms, politics, and social behavior. Through visual narratives, motion pictures provide audiences with shared experiences that foster empathy, awareness, and social dialogue across diverse communities [10], [11].

Beyond their cultural impact, motion pictures also contribute substantially to the global and local economy. The film industry generates employment across various sectors, including production, distribution, marketing, and exhibition, while also supporting related industries such as tourism, advertising, and technology. Box office performance and audience engagement are therefore not only indicators of entertainment value, but also measures of economic viability and industry sustainability [12], [13].

Motion pictures function as a platform for education and social commentary, addressing topics such as history, social justice, and human behavior in ways that are accessible to a broad audience [14]. Because of their wide reach and influence, understanding the factors that contribute to a film's success is essential for filmmakers, producers, and stakeholders seeking to maximize both cultural impact and commercial performance [15]

B. Potential Movie Success Indicators

Movie success has been widely examined in film industry research, with critics identifying multiple determinants that influence both financial performance and audience reception. These factors span production-related attributes, marketing strategies, creative talent involvement, and market characteristics. Variables such as genre, production budget, production companies, country of origin, cast popularity, director reputation, and language have been consistently linked to variations in box office revenue and overall commercial outcomes. Understanding these determinants provides insight into how structural, creative, and market-driven elements interact to shape a film's performance.

1) Production Budget: Production budget is consistently identified as a key determinant, as larger budget films often allow for higher production values, well-known talent, and extensive promotional campaigns, all of which can enhance audience awareness and box office performance [16], [17], [18]. Studies also highlight that star power, including recognized actors and directors, can signal perceived quality and attract larger audiences. However, striving to produce a successful movie cannot be done by solely relying on the big names of the cast members to propel the movie to become a box office hit. In some contexts, sequels and franchise films benefit from established audience familiarity, increasing their likelihood of financial success compared to original productions [19], [16].

2) Movie Genre: Several studies suggest that movie genres contribute to variations in box office performance, as genres often attract distinct audience segments and align with specific viewer expectations. For example, a study has found that some genres such as animated films, action, adventure, action thriller, and drama usually garner a higher popularity and revenue in contrast to genres such as documentaries [20]. However, genre alone cannot dictate the success of a movie as movies of the same genre can produce widely different outcomes depending on complementary factors such as narrative quality, marketing strategy, cast, and release context. With that said, high-performing genres can still under perform if other success-related metrics are considered to be insufficient [17].

3) Production Company: Studies emphasizes the influence of production companies, particularly major studios, on a film's revenue performance. Statistical analyses have demonstrated a significant relationship between major studio releases and higher box office earnings [17]. These are further supported by an earlier study which indicates a strong positive association between major studio distribution and financial performance, largely attributed to superior marketing resources and extensive distribution networks [21].

III. Methodology

This section presents the methodology employed in conducting the study. It contains the data collection, tools, and procedures used to achieve the objectives of the research. This section also explains in detail the process involved in data preparation, model training, and evaluation of results.

A. Data Collection

The data used in this study were obtained from Kaggle, specifically the dataset titled Movie Dataset: Budgets, Genres, Insights, uploaded by Utkarsh Singh. The dataset is a comprehensive collection consisting of 4,803 movies and contains 24 features that describe various financial, technical, and audience-related aspects of each film. The dataset was selected due to its completeness and relevance to analyzing factors associated with movie success. It includes both numerical and categorical attributes, allowing for a multidimensional analysis of movie characteristics using unsupervised learning techniques.

The dataset consists of multiple features that capture the financial, technical, and audience-related characteristics of each movie. Key financial attributes include the production budget and total revenue, which serve as primary indicators of commercial performance. Descriptive and categorical features such as genres, keywords, overview, tagline, cast, crew, and director provide contextual information about the creative and narrative aspects of the films. Production-related variables, including production companies, production countries, original language, spoken languages, and release status, offer insight into the organizational and geographical background of each movie. Temporal and technical details are represented by release date and runtime, while audience engagement and reception are reflected through popularity scores, average user ratings, and vote counts. Unique identifiers such as index and movie ID are included for data organization and reference purposes. Collectively, these features enable a comprehensive analysis of movie characteristics and support the application of unsupervised learning techniques to identify patterns related to movie success. Table I provides a detailed description of each feature in the dataset, including the data type and a brief description.

B. Exploratory Data Analysis

An exploratory data analysis (EDA) was conducted to understand the structure, distribution, and underlying relationship within the data set prior to implementing the unsupervised machine learning model. This step is essential to identify patterns, detect anomalies, evaluate feature distribution, and determine necessary preprocessing transformations to ensure meaningful clustering results.

The dataset was first imported and inspected to verify data types, completeness, and overall structure. Summary statistics were generated to examine the central tendency and dispersion of numerical variables, including budget, revenue, popularity, runtime, vote average, and vote count. This preliminary inspection ensured that the variables were correctly formatted and suitable for numerical analysis. A summary of the statistics can be seen in II.

Distribution analysis was then performed to assess skewness and detect extreme values. Histograms with kernel density estimation revealed that financial and engagement-related variables, particularly budget, revenue, popularity, and vote count, exhibited strong right-skewed distributions. This indicates that while most movies have moderate values, a small number of films achieve exceptionally high financial or popularity metrics. Box plots further confirmed the presence of significant outliers across these variables. The distributions of these numerical features before transformation are illustrated in Figure 1 and Figure 2.

TABLE I
VARIABLE DESCRIPTION

Attribute Name	Data Type	Description
Index	Integer	Index value used to identify each record in the dataset.
Budget	Integer	Total production budget allocated for the movie.
Genres	String	Genre classifications associated with the movie.
Homepage	String	Official homepage URL of the movie, if available.
Movie_ID	Integer	Unique numerical identifier assigned to each movie.
Keywords	String	Set of keywords describing themes and content of the movie.
Original_Language	String	Original language in which the movie was produced.
Original_Title	String	Original title of the movie prior to localization or translation.
Overview	String	Brief synopsis summarizing the movie's plot and content.
Popularity	Float	Popularity score indicating audience interest and engagement.
Production_Companies	String	Companies responsible for producing the movie.
Production_Countries	String	Countries where the movie was produced.
Release_Date	String	Official theatrical release date of the movie.
Revenue	Integer	Total box office revenue generated by the movie.
Runtime	Float	Duration of the movie measured in minutes.
Spoken_Languages	String	Languages spoken by characters in the movie.
Status	String	Release status of the movie (e.g., Released, Rumored).
Tagline	String	Promotional slogan or tagline associated with the movie.
Title	String	Official title of the movie.
Vote_Average	Float	Average user rating score for the movie.
Vote_Count	Integer	Total number of user votes submitted for the movie.
Cast	String	List of main cast members appearing in the movie.
Crew	String	List of crew members involved in movie production.
Director	String	Director responsible for the movie.

TABLE II
DESCRIPTIVE STATISTICS OF MOVIE DATASET

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Budget	4800	2.91e7	4.07e7	0	8.00e5	1.50e7	4.00e7	3.80e8
Popularity	4800	21.51	31.82	0.00037	4.68	12.93	28.35	875.58
Revenue	4800	8.23e7	1.63e8	0	0	1.92e7	9.29e7	2.79e9
Runtime	4800	106.90	22.56	0	94	103	118	338
Vote Average	4800	6.09	1.19	0	5.6	6.2	6.8	10
Vote Count	4800	690.65	1234.85	0	54	236	737.25	13752

To examine linear relationship among the features, a correlation matrix was visualized. The heatmap revealed strong positive correlations between budget and revenue, as well as

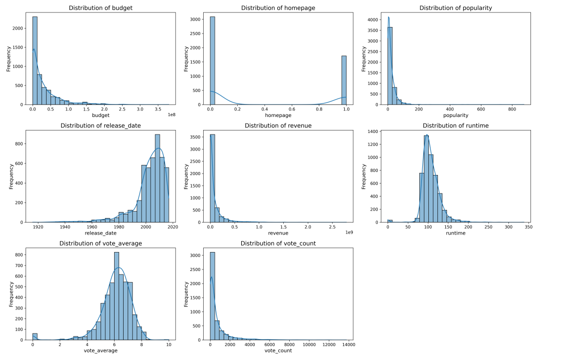

Fig. 1. Distribution of Numerical Features Before Transformation

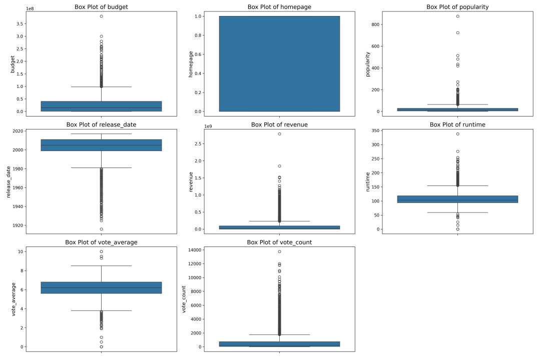

Fig. 2. Boxplot of Numerical Features Before Transformation

popularity and vote count. This can be seen in Figure 3. Identifying these relationships is important because highly correlated features can influence distance-based clustering algorithms by amplifying certain dimensions of similarity.

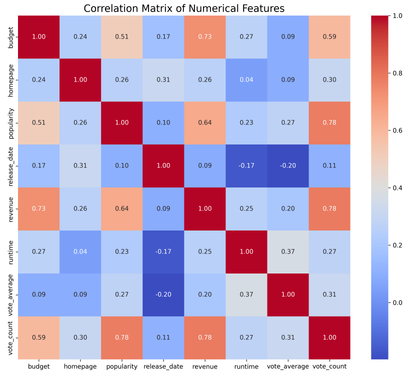

Fig. 3. Correlation Heatmap

Given the heavy skewness observed in budget, revenue, popularity, and vote count, a logarithmic transformation (log1p) was applied to these features. This transformation reduced skewness, minimized the influence of extreme values, and stabilized variance across observations.

Feature Engineering

Feature engineering was also performed to incorporate temporal information. The release date feature was converted into a new feature representing the age of the movie, calculates as the difference between the current year and the data the movie was released. This transformation converted the time-based information into a numerical format, allowing the unsupervised models to effectively utilize it during clustering. The original release data column was subsequently removed to avoid redundancy.

Data Cleaning

Data cleaning steps were implemented to ensure quality and consistency. Movies with invalid runtime (e.g., runtime equal to zero) were removed and extreme runtime outliers (e.g., durations exceeding 300 minutes) were excluded to prevent distortion in clustering results.

Standardization

The numerical features specifically the budget, popularity, revenue, runtime, and vote average were standardized using the StandardScaler method. This process applies z-score normalization, transforming each feature so that it has a mean of zero and a standard deviation of one. This step is crucial for distance-based unsupervised learning algorithms, as these models are more sensitive to differences in feature scale. Without scaling, features with larger disparity would disproportionately influence the machine learning models. Normalization ensures that the selected features were on a comparable scale and ensures balanced contribution during the clustering process.

Finally, the processed and standardized dataset was then exported for use in the training of the unsupervised learning models.

C. *Data Pre-Processing*

To cluster and analyze the success of movies, the dataset was first cleaned and preprocessed to ensure accuracy and suitability for unsupervised learning. The preprocessing steps involved understanding the dataset, handling missing values, and transforming relevant features for analysis.

1) Understanding and Cleaning the Dataset: The dataset was initially explored to identify duplicates, null values, and inconsistencies. Columns such as homepage, genres, production_companies, production_countries, release_date, cast, and director required specific transformations to standardize the data. For example, the homepage column was converted to a binary feature indicating the presence or absence of a homepage. Text fields such as genres were split into lists of individual genres, while production_companies and production_countries were parsed to extract names from structured text. The release_date column was converted to the release year to enable temporal analysis. Names in the cast and director columns were processed to combine first and last names into full names, ensuring consistency across records.

2) Genre Filtering: The genres column was analyzed to determine the frequency of each genre in the dataset. The most common genres included Drama, Comedy, Thriller, Action, Romance, Adventure, and Crime, while genres such as Western, Foreign, TV, and Movie were infrequent (less than 100 occurrences). Rows corresponding to these less-represented genres were removed to improve model reliability and avoid sparsity in the clustering process. Table III shows the unique count of genres available in the dataset.

TABLE III
MOVIE GENRES AND THEIR FREQUENCY IN THE DATASET

Genre	Number of Movies
Drama	2297
Comedy	1722
Thriller	1259
Action	1153
Romance	890
Adventure	790
Crime	696
Science	530
Fiction	530
Horror	519
Family	510
Fantasy	418
Mystery	347
Animation	234
History	197
Music	183
War	142
Documentary	110
Western	80
Foreign	34
TV	8
Movie	8

3) Language Filtering: Similarly, the original_language column was examined to identify the most common languages of movie production. Movies with languages represented by fewer than 20 films, such as Japanese, Italian, Chinese, Korean, Russian, and several others, were excluded to ensure sufficient sample sizes for analysis. The majority of films in the dataset were produced in English, representing 4,505 of the 4,803 movies, allowing for consistent clustering on language-influenced patterns. Table IV shows the unique count of languages available in the dataset.

4) Production Company and Country Filtering: The columns of the production_companies and production_countries of the dataset were analyzed to identify unique entities and their contributions to the dataset. After splitting and counting occurrences, a total of 5,026 unique production companies were identified, and the frequency of each company's appearance was calculated. Similarly, each movie's production countries were analyzed to determine the number of unique countries represented and their relative contribution across all films. To reduce sparsity and improve the reliability of clustering, production companies and countries with fewer than 100 occurrences were removed. This filtering step ensured that only companies and countries with sufficient data points were included in the analysis, thereby minimizing noise from underrepresented entities. Tables X and Y show the production companies and countries that met the threshold for inclusion.

TABLE IV
ORIGINAL LANGUAGES OF MOVIES AND THEIR FREQUENCY IN THE DATASET

Language Name	Language Code	Number of Movies
English	en	4505
French	fr	70
Spanish	es	32
Chinese	zh	27
German	de	27
Hindi	hi	19
Japanese	ja	16
Italian	it	14
Chinese (Mandarin)	cn	12
Korean	ko	11
Russian	ru	11
Portuguese	pt	9
Danish	da	7
Swedish	sv	5
Dutch	nl	4
Farsi	fa	4
Thai	th	3
Hebrew	he	3
Indonesian	id	2
Czech	cs	2
Tamil	ta	2
Romanian	ro	2
Arabic	ar	2
Telugu	te	1
Hungarian	hu	1
Unknown	xx	1
Afrikaans	af	1
Icelandic	is	1
Turkish	tr	1
Vietnamese	vi	1
Polish	pl	1
Norwegian Bokmål	nb	1
Kyrgyz	ky	1
Norwegian	no	1
Slovenian	sl	1
Pashto	ps	1
Greek	el	1

5) Cast Members and Directors Filtering: The cast and director columns were processed to identify unique individuals and their frequency of appearance across the dataset. Each movie's cast and director lists were split into individual names, and the occurrences of each person were counted. To ensure sufficient representation and reduce sparsity, cast members and directors appearing in fewer than 10 movies were removed from the dataset. This filtering step allowed the clustering model to focus on contributors with adequate data points, minimizing the influence of rarely represented actors or directors and improving the reliability and robustness of pattern discovery in the analysis.

D. Experimental Setup

Visual Studio Code was used as the main integrated development environment (IDE) for data cleaning, visualization, preprocessing, and model training. All experimental models were done using Python 3.10, with libraries such as NumPy (2.3.3), pandas (2.3.3), matplotlib (3.10.7), scikit-learn (1.7.2) and seaborn (0.13.2) for data manipulation, visualization, and model implementation.

The dataset contained 705 rows and was split into training and testing sets, where 80% of the data was allocated for training and the remaining 20% was used for testing and measuring the accuracy of the models. Each model was trained using data that used the same preprocessing techniques to maintain fairness across comparisons.

E. Algorithm

To identify underlying patterns in the data set, this study applies multiple unsupervised clustering algorithms. Since the objective is to group observations without predefined labels, clustering methods are appropriate for discovering groups within the data. Three algorithms were employed, namely K-Means, Hierarchical Clustering, and Gaussian Mixture Model, to compare different clustering strategies and determine which approach best captures meaningful groupings in the dataset.

K-Means Clustering

This is a centroid-based unsupervised learning algorithm that partitions data into k clusters by minimizing the within-cluster sum of squared distances between data points and their respective cluster centroids. It groups observations based on similarity, typically measured using Euclidean distance, and is commonly applied to uncover underlying patterns or natural groupings in unlabeled datasets [22]. This can be mathematically represented in (1).

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

where k is the number of clusters, C_i is the set of points in cluster i, and μ_i is the centroid of cluster i. The squared Euclidean distance $\|x - \mu_i\|^2$ measures the similarity between a data point x and the cluster centroid.

Hierarchical Clustering

This is an unsupervised clustering technique that builds a nested hierarchy of clusters using either agglomerative or divisive strategies. Instead of fixing the number of clusters beforehand, it progressively merges or splits data points based on similarity, and represents the cluster relationships in a tree-like structure called a dendrogram [23]. This allows exploration of cluster relationships at multiple levels of granularity.

Gaussian Mixture Models (GMM)

GMM is a probabilistic clustering method that assumes data points are generated from a mixture of several Gaussian distributions with unknown parameters. Each cluster corresponds to a Gaussian component characterized by its mean and covariance, and data points are assigned membership probabilities to clusters rather than hard labels, enabling flexible modeling of overlapping clusters [24].

F. Training Procedure

G. Evaluation Metrics

H. Evaluation Metrics

I. Comparison of Clustering Algorithms

IV. Results and Discussion

A. Clustering Algorithm Performance Comparison

Table V presents a comprehensive comparison of three clustering algorithms—K-Means, Hierarchical Clustering, and Gaussian Mixture Model (GMM)—evaluated using three distinct metrics: Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score.

TABLE V
Performance comparison of clustering algorithms across three evaluation metrics

Metric	**K-Means**	**Hierarchical**	**GMM**
Silhouette Score (Higher is Better)	0.4144	0.4429	**0.4750**
Davies-Bouldin Score (Lower is Better)	0.7078	0.6827	**0.6824**
Calinski-Harabasz Score (Higher is Better)	**9858.89**	6530.04	7120.97

1) Silhouette Score Analysis: The Silhouette Score measures how similar an object is to its own cluster compared to other clusters, with values ranging from -1 to 1. GMM achieved the highest score of 0.4750, outperforming Hierarchical Clustering (0.4429) and K-Means (0.4144). This indicates that GMM produced the most cohesive and well-separated clusters. The superior performance of GMM can be attributed to its probabilistic approach, which allows for soft clustering boundaries and can model clusters with different shapes and densities. While all three methods achieved moderate silhouette scores (above 0.4), suggesting reasonable cluster quality, GMM's 14.7% improvement over K-Means demonstrates its superior ability to capture the underlying structure of the movie dataset.

2) Davies-Bouldin Score Analysis: The Davies-Bouldin Score measures the average similarity ratio of each cluster with its most similar cluster, where lower values indicate better separation. GMM achieved the best score of 0.6824, followed very closely by Hierarchical Clustering (0.6827) and K-Means (0.7078). The minimal difference between GMM and Hierarchical methods (0.0003) suggests they achieve similar inter-cluster separation. GMM's 3.6% improvement over K-Means indicates better-defined cluster boundaries, likely because GMM's probabilistic framework naturally accounts for cluster overlap and uncertainty, whereas K-Means enforces hard boundaries that may not align with the data's natural structure.

3) Calinski-Harabasz Score Analysis: The Calinski-Harabasz Score (also known as the Variance Ratio Criterion) measures the ratio of between-cluster dispersion to within-cluster dispersion. Interestingly, K-Means achieved the highest score of 9858.89, significantly outperforming GMM (7120.97) and Hierarchical Clustering (6530.04). This result might seem contradictory to the other metrics, but it reveals an important insight: K-Means optimizes for compact, spherical clusters with high between-cluster variance, which directly aligns with what the Calinski-Harabasz Score measures. However, this doesn't necessarily mean K-Means produced better clusters overall—it may have created more artificially separated clusters that don't reflect the true data structure. The 38.5% higher score for K-Means compared to GMM suggests K-Means maximized variance ratios at the expense of modeling the actual probability distributions of the data.

4) Overall Model Selection: Despite K-Means' strong performance on the Calinski-Harabasz metric, GMM emerges as the superior choice when considering all three metrics holistically. GMM either outperformed or matched the other methods on the Silhouette and Davies-Bouldin scores, which better capture cluster cohesion and separation quality. Additionally, GMM's probabilistic nature provides uncertainty estimates and can model complex, overlapping cluster structures that are likely present in movie datasets where genres, audiences, and production characteristics create natural overlap. Therefore, GMM was selected for subsequent analysis, and all following visualizations focus on the three GMM-identified clusters.

B. Cluster Distribution and Characteristics

1) Distribution of Movies per GMM Cluster: The distribution of movies across the three GMM clusters reveals an imbalanced but interpretable structure. Cluster 2 contains the largest number of movies (2,213, or 46.5%), followed by Cluster 0 (1,976, or 41.5%), while Cluster 1 is notably smaller with only 576 movies (12.1%). This imbalance is not necessarily problematic—it likely reflects the natural distribution of different movie types in the dataset. The 3.8:1 ratio between the largest and smallest clusters suggests that Cluster 1 represents a more specialized or niche category of films, while Clusters 0 and 2 capture broader, more common film types. This distribution pattern is common in real-world clustering applications where distinct subpopulations exist at different scales.

C. Feature-Based Cluster Analysis

The following sections analyze how each cluster differs across key movie attributes, revealing distinct profiles that characterize each group.

1) Mean Homepage Presence across GMM Clusters: The homepage feature (a binary indicator of whether a movie has an official website) shows dramatic variation across clusters. Cluster 2 exhibits the highest mean homepage presence at 0.5807 (58.1%), meaning more than half of movies in this cluster have official websites. In contrast, Cluster 0 shows moderate presence at 0.1802 (18.0%), while Cluster 1 has the lowest at 0.1146 (11.5%).

This 5.1× difference between Cluster 2 and Cluster 1 is highly significant and suggests that Cluster 2 represents modern, commercially-oriented films with established digital marketing presence. The low homepage presence in Cluster 1 (only 11.5%) strongly indicates these are older films produced before the widespread adoption of internet marketing. Cluster 0's moderate presence suggests a middle ground, potentially representing older films that later received homepage updates or moderately budgeted recent films.

2) Mean Popularity across GMM Clusters: The popularity metric (likely standardized) reveals clear differentiation. Cluster 2 shows the highest popularity at 0.1165 (above the dataset mean), while Clusters 0 and 1 fall below average at -0.1193 and -0.0385 respectively. The 2.3× popularity advantage of Cluster 2 over Cluster 0 indicates these films receive substantially more attention, viewer engagement, or cultural relevance.

Interestingly, Cluster 1, despite having characteristics that might suggest quality (as we'll see in later metrics), shows below-average popularity. This suggests these films, while potentially critically acclaimed, may not achieve mainstream commercial success or contemporary cultural relevance. The popularity distribution supports the hypothesis that Cluster 2 represents mainstream commercial cinema, while Cluster 1 represents classic or niche films with limited contemporary appeal.

3) Mean Budget across GMM Clusters: Budget values (appearing to be standardized) show surprisingly minimal variation across clusters. Cluster 1 has the highest mean budget at 0.0292, followed by Cluster 2 at -0.0028, and Cluster 0 at -0.0054. All three values hover very close to zero, with the maximum difference being only 0.0346 standard deviations.

This near-uniformity in budget is unexpected and suggests that **budget is not a primary discriminating factor** in the GMM clustering algorithm's separation of these movie groups. Two interpretations are possible: (1) The dataset may have undergone inflation adjustment, normalizing budgets across different eras, or (2) The clustering algorithm prioritized other features (popularity, ratings, temporal patterns) over budget when forming clusters. The slight elevation in Cluster 1's budget might reflect the fact that classic acclaimed films often had substantial production values for their era, but this effect is minimal when standardized.

4) Mean Runtime across GMM Clusters: Runtime shows striking differentiation, particularly for Cluster 1, which has a mean runtime of 0.3788 (considerably longer than average). In stark contrast, Cluster 2 shows the shortest runtime at -0.0917, while Cluster 0 is near average at -0.0077. The 4.1× difference between Cluster 1 and Cluster 2 (0.3788 vs -0.0917) represents one of the most dramatic feature separations across all metrics.

This pattern strongly supports the characterization of Cluster 1 as classic, prestige films. Historically, longer runtimes were associated with epic productions, serious dramas, and films with artistic ambitions (e.g., "Gone with the Wind," "Lawrence of Arabia," "The Godfather"). Conversely, Cluster 2's shorter runtime aligns with modern commercial cinema trends, where studios increasingly favor tighter, more accessible runtimes to maximize theater turnover and appeal to contemporary audiences with shorter attention spans. Cluster 0's average runtime suggests a diverse mix of film types without strong temporal or stylistic skew.

5) Mean Vote Average across GMM Clusters: The vote_average metric (user ratings) reveals perhaps the most interpretively significant pattern. Cluster 1 achieves the highest mean rating at 0.4375 (substantially above average), while Cluster 0 is near average at 0.0108, and Cluster 2 is notably below average at -0.1235. The 3.5× rating advantage of Cluster 1 over Cluster 2 represents a critical finding.

This pattern suggests that Cluster 1 contains critically acclaimed, high-quality films that have stood the test of time, these are likely classic films that remain highly rated decades after release due to their enduring artistic merit. Conversely, Cluster 2's below-average ratings despite high popularity indicate these are mainstream commercial films that achieve widespread viewership but may lack critical acclaim or lasting quality. This inverse relationship between popularity and quality is a well-documented phenomenon in commercial cinema, where mass appeal often comes at the expense of critical depth. Cluster 0's average ratings suggest a balanced mix of quality levels.

6) Mean Vote Count across GMM Clusters: Vote count (number of user ratings) shows relatively modest variation compared to other features. Cluster 2 has the highest mean vote count at 5.4473, followed by Cluster 1 at 5.1666 and Cluster 0 at 5.0687. The 7.5% higher vote count in Cluster 2 compared to Cluster 0 suggests more viewer engagement.

Notably, Cluster 2's high vote count aligns with its high popularity, reinforcing that these films receive substantial viewer attention. Cluster 1's moderately high vote count despite lower popularity suggests these classic films have accumulated ratings over time from dedicated cinephiles and critics. The logarithmic nature of these values (all around 5, suggesting hundreds of thousands of votes when exponentiated) indicates all three clusters contain relatively well-known films with substantial viewer bases, rather than obscure titles.

7) Mean Age of Movie across GMM Clusters: The age_of_movie feature provides the clearest temporal differentiation across clusters. Cluster 1 contains the oldest films with a mean age of 50.07 years (films from approximately 1976 on average), Cluster 0 contains middle-aged films at 25.45 years (approximately 2001), and Cluster 2 contains the newest films at 15.04 years (approximately 2011). This creates a 3.3× age difference between the oldest and newest clusters.

This temporal stratification is critically important for interpreting all other cluster characteristics. It explains:

- **Why Cluster 1 has low homepage presence**: The internet wasn't commercially viable until the mid-1990s
- **Why Cluster 1 has longer runtimes and higher ratings**: Classic era filmmaking emphasized artistic merit over commercial efficiency
- **Why Cluster 2 has high popularity and homepage presence**: These are modern films with active digital marketing and contemporary cultural relevance
- **Why Cluster 2 has lower ratings**: Modern commercial cinema often prioritizes accessibility over critical depth

The temporal dimension reveals that GMM has successfully identified three distinct eras of cinema, each with characteristic production values, marketing approaches, and audience reception patterns.

D. Cluster Interpretation and Naming

Based on the comprehensive analysis of all features, the three GMM clusters can be definitively characterized and named:

1) Cluster 0: "Mid-Era Balanced Films": This cluster represents movies from the early 2000s era (25 years old) with moderate characteristics across most dimensions. These films show average ratings, popularity, runtime, and budget, with modest digital presence. This cluster likely contains a diverse mix of mid-budget productions, independent films, and commercial releases from the pre-digital-marketing era that neither achieved lasting classic status nor maintain strong contemporary relevance. They represent the "middle ground" of cinema, competent productions that fill the space between timeless classics and modern blockbusters.

2) Cluster 1: "Classic Acclaimed Cinema": This cluster contains the oldest films (50 years old, from the 1970s and earlier) characterized by significantly longer runtimes, the highest viewer ratings, and minimal digital presence. Despite below-average contemporary popularity, these films represent prestigious, artistically ambitious productions that have endured critical acclaim over decades. This cluster likely includes Golden Age Hollywood epics, New Wave cinema, and foundational classic films that defined genres and influenced filmmaking but may not resonate as strongly with modern mass audiences. The quality-over-popularity profile distinctly identifies this as the prestige cinema cluster.

3) Cluster 2: "Modern Commercial Blockbusters": This cluster represents the newest films (15 years old, from approximately 2011 onward) with the highest popularity, strongest digital marketing presence, highest vote counts, but notably lower critical ratings and shorter runtimes. These are contemporary mainstream commercial films optimized for modern theatrical distribution and digital marketing. This cluster likely contains franchise films, action blockbusters, and popular genre films designed for broad appeal and commercial success rather than lasting artistic merit. The popularity-over-quality profile and strong digital presence clearly identify this as the modern commercial cinema cluster.

V. Conclusion

This study successfully applied unsupervised machine learning techniques to identify and characterize distinct segments within a comprehensive movie dataset. Through systematic comparison of K-Means, Hierarchical Clustering, and Gaussian Mixture Model (GMM) algorithms using multiple evaluation metrics (Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score), GMM emerged as the superior clustering approach, achieving the best performance on two of three metrics and demonstrating the ability to model the complex, probabilistic structure of movie data.

The GMM algorithm identified three meaningful clusters that fundamentally represent three eras of cinema, each with distinct characteristics:

1) **Classic Acclaimed Cinema (12% of films)**: Older, longer, higher-rated films with enduring artistic merit but limited contemporary commercial appeal
2) **Mid-Era Balanced Films (41% of films)**: Movies from the early 2000s with moderate attributes across all dimensions
3) **Modern Commercial Blockbusters (47% of films)**: Recent, popular, heavily marketed films optimized for commercial success over critical acclaim

The most significant finding is the inverse relationship between popularity and critical acclaim across temporal boundaries, older films demonstrate superior ratings despite lower contemporary popularity, while modern films achieve widespread viewership despite lower critical ratings. This pattern reflects fundamental shifts in the film industry: the transition from artistically-driven classic cinema to commercially-optimized modern blockbuster production, the rise of digital marketing and franchise-based strategies, and changing audience consumption patterns in the streaming era.

The temporal stratification (50-year-old classics, 25-year-old mid-era films, and 15-year-old modern releases) suggests that time itself is a critical factor in how movies cluster, with production values, marketing strategies, runtime conventions, and critical reception all strongly correlated with release era. This finding has important implications for movie recommendation systems, which should account for temporal preferences, and film industry analysis, which must recognize that contemporary metrics (popularity, homepage presence) may not correlate with lasting artistic value.

A. Practical Implications

This clustering framework can be applied to:

- **Personalized recommendation systems** that identify users' preferences across the classic-to-modern spectrum
- **Marketing strategy optimization** by recognizing that different film types require different promotional approaches
- **Investment decisions** in film production, distinguishing between projects aimed at critical acclaim versus commercial success
- **Content acquisition strategies** for streaming platforms seeking balanced catalogs across cinematic eras

B. Limitations and Future Research Directions

This study has several limitations that present opportunities for future investigation. First, the standardized budget values showed minimal variation, suggesting that inflation-adjusted budget analysis or era-specific budget percentile rankings might provide clearer insights into the role of production costs. Second, the dataset appears to contain primarily well-known films (all clusters show high vote counts), indicating potential selection bias toward popular titles, future research should examine whether similar patterns emerge in datasets including obscure or limited-release films.

Third, the three-cluster solution was predetermined; optimal cluster number selection using methods like the elbow method or silhouette analysis across different k values could reveal whether finer-grained segments exist (e.g., distinguishing subgenres within modern commercial cinema). Fourth, this study focused on numerical features and temporal patterns; incorporating genre information, production company data, and cast/crew attributes could provide richer cluster characterizations and potentially reveal creative talent patterns across cinema eras.

Finally, longitudinal analysis examining how individual films' popularity and ratings evolve over time could provide insights into which modern films are achieving classic status and whether the popularity-quality tradeoff persists as films age. Future research could also explore predictive modeling to forecast which contemporary films are likely to become enduring classics based on their initial feature profiles.

In conclusion, this study demonstrates that unsupervised machine learning can successfully uncover meaningful structure in movie data, revealing temporal evolution patterns in cinema that reflect broader cultural and industry shifts. The identification of distinct classic, mid-era, and modern commercial clusters provides a data-driven framework for understanding the film landscape and suggests that the movie industry operates along multiple dimensions, not just commercial success or critical acclaim alone, but a complex interplay of artistic ambition, audience preferences, marketing strategies, and temporal context that varies systematically across the history of cinema.

References

[1] Market Reports World. (2024) Movies and entertainment market report: Forecast [2025-2033]. Accessed: 2025-02-13. [Online]. Available: https://www.marketreportsworld.com/market-reports/movies-and-entertainment-market-14715430

[2] Renub Research. (2024) United states movie market, size, share, forecast 2024-2033. Accessed: 2025-02-13. [Online]. Available: https://www.renub.com/united-states-movie-market-p.php

[3] Get Pzazzed. (2025, May) Spotlight: Film industry statistics and trends 2024. Accessed: 2025-02-13. [Online]. Available: https://pzaz.io/producer-blog/film-industry-statistics/

[4] R. Dhir and S. K. Raj, "Movie success prediction using machine learning algorithms and their comparison," in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. Jalandhar, India: IEEE, Dec. 2018, pp. 385–390. [Online]. Available: https://ieeexplore.ieee.org/document/8703320/

[5] D. Roy and S. Goel, "Movie success prediction using ml," in *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*. Jaipur, India: IEEE, Dec. 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9298145/

[6] U. Sharma. (2024) Movies dataset. Accessed: 2025-02-13. [Online]. Available: https://www.kaggle.com/datasets/utkarshx27/movies-dataset

[7] D. C. G. Putri, J.-S. Leu, and P. Seda, "Design of an unsupervised machine learning-based movie recommender system," *Symmetry*, vol. 12, no. 2, p. 185, Jan. 2020. [Online]. Available: https://www.mdpi.com/2073-8994/12/2/185

[8] L. F. Sikos, M. Stumptner, A. Micsik, and D. Philp, "Movie recommender systems: Concepts, methods, challenges, and future directions," *Sensors*, vol. 22, no. 13, p. 4904, Jun. 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9269752/

[9] N. Quader, M. O. Gani, and D. Chaki, "Performance evaluation of seven machine learning classification techniques for movie box office success prediction," in *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*. Khulna, Bangladesh: IEEE, Dec. 2017, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8275242/

[10] J. Gogoi, "The impact of films on society," *Global Research Journal*, 10 2022.

[11] LIS Academy. (2024) The impact of motion films: From entertainment to social change. LIS Academy. Accessed: 2026-02-12. [Online]. Available: https://lis.academy/information-sources-and-services/impact-motion-films-entertainment-social-change/

[12] E. Mammadova and A. Abdullayev, "Impact of theatre and cinema culture on economy," *Luminis Applied Science and Engineering*, vol. 2, no. 3, pp. 88–97, Sep. 2025. [Online]. Available: https://egarp.lt/index.php/LUMIN/article/view/380

[13] T. Hennig-Thurau, S. A. Ravid, and O. Sorenson, "The economics of filmed entertainment in the digital era," *Journal of Cultural Economics*, vol. 45, no. 2, pp. 157–170, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10824-021-09407-6

[14] C. Rivera, "Film and social change: Exploring the influence of movies on society," *CINEFORUM*, vol. 62, no. 1, pp. 13–18, 2022, published March 31, 2022, Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License. [Online]. Available: https://revistadecineforum.com/index.php/cf/article/view/3

[15] U. N. Wamane, "A critical study of socio-cultural impact of cinema on society," *International Journal of Advance and Applied Research*, vol. 10, no. 3, 2023, published Jan–Feb 2023. [Online]. Available: https://www.researchgate.net/publication/375635211_A_CRITICAL_STUDY_OF_SOCIO-CULTURAL_IMPACT_OF_CINEMA_ON_SOCIETY

[16] A. Kwan and S. Scheepers, "The fault in our stars: A quantitative study on the effect of cast member celebrity on film success," *Journal of Student Research*, vol. 11, 05 2022.

[17] N. Pangarker and E. Smit, "The determinants of box office performance in the film industry revisited," *South African Journal of Business Management*, vol. 44, pp. 47–58, Sep. 2013.

[18] B. Hao, "The analysis of the factors that influence the film revenue," *Highlights in Science, Engineering and Technology*, vol. 47, pp. 154–159, 05 2023.

[19] D. Hermanovitch. (2026) The sequel paradox: Research shows less innovation sells more tickets but only at first. Binghamton University. Accessed: 2026-02-12. [Online]. Available: https://www.binghamton.edu/news/story/6010/the-sequel-paradox-research-shows-less-innovation-sells-more-tickets-but-only-at-first

[20] M. Yan, "Research on movie box office prediction model based on machine learning," *Applied and Computational Engineering*, vol. 151, pp. 82–89, May 2025.

[21] B. R. Litman, "Predicting success of theatrical movies: An empirical study," *Journal of Popular Culture*, vol. 16, no. 4, p. 159, spring 1983.

[22] GeeksforGeeks. (2025) K means clustering – introduction. Last updated 10 Nov 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/

[23] ——. (2026) Hierarchical clustering in machine learning. Last updated 19 Jan 2026. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/hierarchical-clustering/

[24] ——. (2025) Gaussian mixture model. Last updated 12 Sep 2025. [Online]. Available: https://www.geeksforgeeks.org/gaussian-mixture-model/