Unsupervised Segmentation of Video Game Metadata for Quality Estimation

Custer Jeremiah Valencerina
College of Computing and Information Technologies
National University
Manila, Philippines
valencerinacp@students.national-u.edu.ph

Gabriel Angelo Viñas
College of Computing and Information Technologies
National University
Manila, Philippines
vinasgm@national-u.edu.ph

Abstract—**The video game industry generates over $180 billion annually, yet determining game quality remains subjective and fragmented across commercial success and critical reception metrics. This study addresses the challenge of automatically segmenting video games into meaningful quality archetypes by analyzing latent patterns between commercial performance and critical reception using unsupervised learning. We implemented and compared three clustering algorithms—K-Means, Hierarchical Agglomerative Clustering, and DBSCAN—on a dataset of 6,900 video games with sales and rating metadata spanning 1980 to 2016. Features were engineered to capture critic-user score divergence, log-transformed to address skewness, and reduced to three dimensions via Principal Component Analysis. Our evaluation using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score identified three distinct game archetypes: Mass Market Blockbusters (high sales, high critic scores), Underrated Gems (low sales, positive user-critic discrepancy of +6.53 points), and Commercial Titans (moderate sales, balanced reception). K-Means achieved the strongest overall performance across metrics (Silhouette Score = 0.407, Davies-Bouldin Index = 0.848, Calinski-Harabasz Score = 5416.97). Critically, results demonstrate that critical acclaim does not guarantee commercial success—Underrated Gems averaged 88.4% lower global sales than Mass Market Blockbusters despite favorable user reception. This automated segmentation framework provides a transparent, metadata-driven alternative to opaque review aggregation, offering actionable market intelligence for developers, publishers, and consumers in understanding quality beyond simple score averages.**

Index Terms—**Unsupervised learning, clustering algorithms, K-Means, Gaussian Mixture Models, hierarchical clustering, DBSCAN, dimensionality reduction, PCA, t-SNE, video game analytics, market segmentation**

I. Introduction

The video game industry generates over 180 billion dollars annually, with thousands of new titles released across multiple platforms each year. This massive scale makes assessing game quality increasingly complex and multidimensional [1]. Traditional quality assessment relies on aggregated review scores from critics and players, but these metrics often fail to capture the nuanced relationship between critical reception and commercial success [2].

A fundamental question emerges: do critically acclaimed games always achieve commercial success, or do distinct market segments exist where quality and sales diverge? Previous research in entertainment analytics demonstrates that the relationship between reviews and sales is complex and varies significantly across different contexts [3]. However, most existing approaches rely on supervised methods requiring pre-labeled data or simple score aggregation that lacks transparency in categorizing games into meaningful archetypes.

This study addresses the challenge of automatically segmenting video games into quality archetypes by analyzing latent patterns between commercial performance and critical reception using unsupervised learning. We implement and rigorously compare four clustering algorithms, K-Means, Hierarchical Agglomerative Clustering, DBSCAN, and Gaussian Mixture Models (GMM) on a dataset of 16,719 video games spanning 1980 to 2016. Our approach discovers hidden patterns without requiring predefined categories, making it particularly valuable for emerging markets or novel game genres where labeled datasets are unavailable or prohibitively expensive to create.

The contributions of this work are threefold: (1) we provide the first comprehensive comparative evaluation of unsupervised clustering algorithms applied to game-level quality segmentation using commercial and critical metadata; (2) we engineer interpretable features that capture the dynamic relationship between critic scores, user scores, and commercial performance; and (3) we demonstrate that meaningful market archetypes emerge from metadata analysis alone, offering actionable insights for developers, publishers, and consumers in understanding market positioning beyond simple rating aggregates.

II. LITERATURE REVIEW

This section establishes the theoretical foundation for unsupervised game quality segmentation, reviews key clustering methodologies, and positions our work within the broader context of video game analytics research.

A. Overview of Key Concepts and Background Information

Unsupervised learning encompasses machine learning techniques that discover hidden patterns in unlabeled data without requiring predefined categories [4]. Unlike supervised learning, which depends on expensive labeled datasets, unsupervised approaches autonomously uncover intrinsic structures and relationships within data—a critical advantage when manual labeling is subjective, costly, or infeasible [5].

Clustering, a fundamental unsupervised learning paradigm, partitions data points according to similarity while maintaining separation between dissimilar groups. We focus on four distinct clustering methodologies, each with complementary strengths for game metadata segmentation:

1) K-Means Clustering: A partitioning approach that iteratively minimizes within-cluster sum of squares (WCSS) by assigning points to the nearest centroid and recalculating centroids [6]. Despite limitations with non-spherical clusters, its computational efficiency O(n·k·d·i) and simplicity have made it a foundational method. The k-means++ initialization strategy significantly improves convergence and solution quality by carefully seeding initial centroids[7].

2) Hierarchical Clustering: Builds tree structures (dendrograms) by iteratively merging similar clusters using linkage criteria such as Ward (minimizing variance), Complete (maximal distance), or Average (mean distance) [8]. Unlike K-Means, hierarchical methods do not require specifying cluster count a priority and reveal nested cluster relationships, providing valuable insights into data structure at multiple granularities.

3) DBSCAN: Identifies clusters based on local density rather than distance, grouping dense regions while classifying isolated points as noise or outliers [9]. DBSCAN handles arbitrary cluster shapes and automatically determines cluster count, though it struggles with varying densities and requires careful parameter tuning (epsilon radius and minimum points)

4) Gaussian Mixture Models (GMM): Probabilistic models assuming data originates from a mixture of Gaussian distributions. The Expectation-Maximization (EM) algorithm iteratively estimates GMM parameters, assigning data points to components with probabilities rather than hard boundaries [10]. GMMs model complex cluster shapes through different covariance structures (full, diagonal, spherical), offering a principled probabilistic framework for cluster membership.

B. Dimensionality Reduction and Evaluation

Clustering quality is assessed using three complementary internal validation metrics: (1) **Silhouette Score** [11] measures cluster cohesion and separation, ranging from -1 to 1, where values near 1 indicate well-separated, dense clusters; (2) **Davies-Bouldin Index** [12]computes average similarity between each cluster and its most similar cluster, where lower values indicate better separation; (3) **Calinski-Harabasz Score** [13]measures the ratio of between-cluster to within-cluster variance, where higher values indicate better-defined clusters.

Principal Component Analysis (PCA) [14] performs linear dimensionality reduction by projecting data onto orthogonal components that capture maximum variance. We employ PCA both for high-dimensional visualization and to reduce noise, improving clustering performance by focusing on the most informative feature combinations.

C. Review of Other Relevant Research Papers

1) Video Game Analytics Research: Unsupervised learning has been extensively applied in video game research, particularly for discovering latent structures in player behavior. Drachen et al. [15] applied K-Means clustering to large-scale behavioral telemetry data from commercial games, successfully identifying distinct player profiles based on gameplay statistics without prior assumptions about player categories. Sifa, Bauckhage, and Drachen [16] expanded this work by comparing multiple clustering paradigms for user modeling, demonstrating that different algorithms yield complementary insights into player segmentation.

Bauckhage et al. [17] further explored clustering techniques on massive multiplayer online game datasets, illustrating how unsupervised learning can uncover behavioral patterns and engagement structures in complex gaming environments. Their work highlights the scalability and applicability of clustering algorithms to large entertainment datasets—a challenge similarly encountered in market-level game metadata analysis. Beyond gaming, Gomez-Uribe and Hunt [18] described how large-scale personalization systems leverage unsupervised modeling and latent factor techniques to structure entertainment catalogs, demonstrating the broader applicability of unsupervised segmentation in digital content domains.

While substantial research exists on player behavior clustering using gameplay telemetry, limited academic work directly applies unsupervised clustering to game-level commercial and critical performance metrics. Most existing studies focus on behavioral analysis or supervised sales prediction models. Previous work in entertainment analytics has shown complex, non-linear relationships between critical reception and commercial success [19], with review volume, timing, and source credibility all influencing consumer behavior. We capture this dynamic through our Score Discrepancy feature, measuring the normalized difference between critic and user scores to identify games with divergent professional and audience reception.

This study shifts the analytical focus from players to games themselves, performing a comparative evaluation of multiple clustering algorithms (K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models) to segment video games based on sales and review metadata. By leveraging game-level metadata rather than individual player behavior, we contribute a novel market-oriented application of unsupervised learning within video game analytics. Our multi-algorithm comparison addresses the fundamental insight that no single clustering algorithm dominates across all datasets [[20],different algorithms excel with different data characteristics, making rigorous comparative evaluation essential for robust segmentation.

Unlike commercial platforms that rely on opaque collaborative filtering or simple score aggregation, our framework provides transparent, interpretable quality archetypes grounded in the relationship between critical reception and commercial performance. This automated segmentation approach offers data-driven insights for strategic decision-making in game development, marketing, and recommendation systems.

D. Prior Attempts and Our Approach

Several researchers have attempted to apply unsupervised learning to video game analytics, though primarily focused on player behavior rather than game-level market segmentation. Drachen et al. [15] applied K-Means clustering to player

behavioral telemetry data from Tomb Raider: Underworld, identifying distinct play styles based on progression patterns and death frequencies. While methodologically rigorous, their approach requires proprietary gameplay data unavailable for historical market analysis and segments players rather than games themselves.

Sifa et al. [16] compared multiple clustering paradigms (K-Means, DBSCAN, hierarchical methods) for user modeling in game analytics, demonstrating that algorithm choice significantly impacts cluster interpretability—a finding that motivates our multi-algorithm comparison. However, their work focused exclusively on player engagement patterns (session length, churn prediction) rather than the relationship between critical reception and commercial performance.

Commercial platforms like Metacritic[2] and Steam employ collaborative filtering for game recommendations, focusing on individual user preference matching rather than understanding broader market segments. These systems lack transparency in their aggregation methodologies and treat all games as existing on a single quality continuum, failing to recognize that games may succeed through different pathways (critical acclaim vs. mass-market appeal).

The gap in the current literature is the absence of unsupervised clustering applied directly to game-level metadata (sales, critic scores, user scores) to discover quality archetypes. No prior work has: (1) performed comparative evaluation of multiple clustering algorithms on game market data, (2) engineered features to capture critic-user divergence, or (3) provided transparent, interpretable market segments beyond simple rating aggregation.

This study addresses these gaps by applying four clustering algorithms (K-Means, Hierarchical, DBSCAN, GMM) to 6,900 games with complete sales and review metadata, comparing their performance using three complementary metrics (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score), and providing detailed archetype characterizations with commercial and critical dimensions. Unlike player-centric behavioral clustering or opaque recommendation systems, our approach offers actionable market intelligence accessible to indie developers, publishers, and researchers without requiring proprietary telemetry data

III. Methodology

This section describes the dataset, algorithms, tools, and techniques used in this study. The general framework follows a standard unsupervised learning pipeline: data collection, exploratory data analysis, data preprocessing (including feature engineering, log transformation, and dimensionality reduction via PCA), model training using three clustering algorithms (K-Means, DBSCAN, and Agglomerative Clustering), and evaluation using internal cluster validity metrics. The following subsections describe each stage in detail.

A. Data Collection

The dataset used in this study is the “Video Game Sales with Ratings” dataset published on Kaggle by Rush4Ratio [21]. It was originally compiled by web-scraping VGChartz for sales data and Metacritic for review scores, covering video games released up to December 22, 2016. The raw dataset contains 16,719 records and 16 features, including game metadata (Name, Platform, Year of Release, Genre, Publisher, Developer, Rating), regional and global sales figures (NA Sales, EU Sales, JP Sales, Other Sales, Global Sales in millions of units), and review metrics (Critic Score, Critic Count, User Score, User Count).

The dataset was accessed programmatically from a public GitHub repository hosting the CSV file. No additional data collection was performed by the authors. The data was not collected by the authors; it was gathered by the original Kaggle contributor through automated scraping of publicly available sales charts and review aggregation platforms.

B. Exploratory Data Analysis

The dataset was examined to understand its structure, distributions, and relationships before applying any clustering algorithms. After loading the raw data, the following observations were made.

The dataset contains 16,719 samples and 16 features. Feature types include numerical (all sales columns, Critic Score, Critic Count, User Score, User Count) and categorical (Name, Platform, Year of Release, Genre, Publisher, Developer, Rating). The User Score column was stored as an object type due to the presence of “tbd” (to be determined) entries, which required conversion to numeric values.

Missing values were present across several features. Critic Score and Critic Count had a substantial proportion of missing entries, as did User Score and User Count. Year of Release, Publisher, Developer, and Rating also contained missing values. The exact missing value counts were printed and inspected during analysis.

Regional sales distributions (NA Sales, EU Sales, JP Sales, Other Sales) were examined using histograms. All four distributions are heavily right-skewed, with the majority of games selling below 1 million units and a small number of titles achieving very high sales. This skewness indicated the need for log transformation during preprocessing.

Global Sales, Critic Score, and User Score distributions were plotted to understand the central tendencies. Critic Scores are roughly normally distributed with a slight left skew, centering around 70. User Scores center around 7.0 on the 0 to 10 scale. Global Sales follow the same right-skewed pattern as the regional figures.

Scatter plots were constructed to explore the relationship between scores and sales. Critic Score versus Global Sales

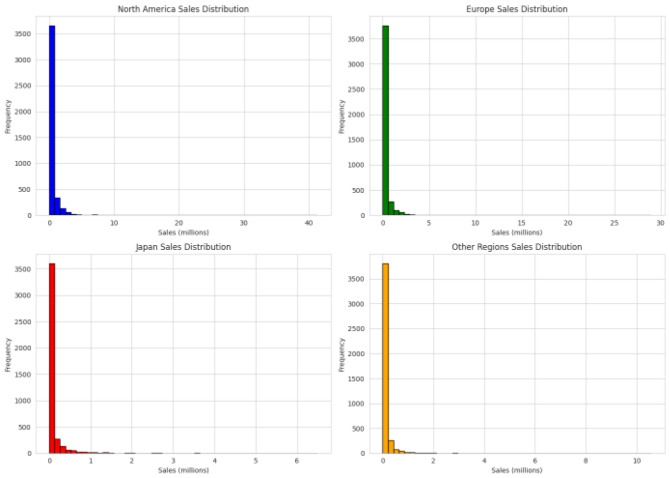


Fig. 1. Distribution of regional sales across North America, Europe, Japan, and other regions. All distributions exhibit strong right skew.

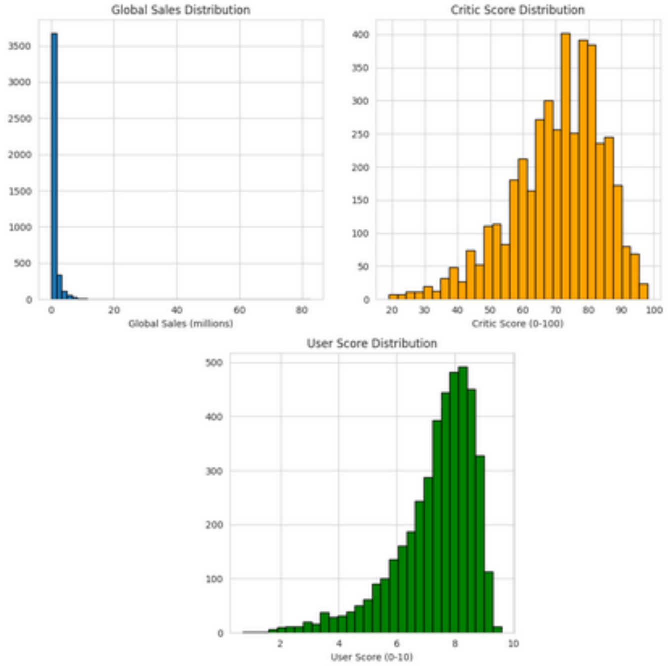


Fig. 2. Distributions of Global Sales, Critic Score, and User Score showing the spread and central tendency of key clustering features.

showed a weak positive trend, with higher-scored games tending to sell more, but with significant variance. User Score versus Global Sales showed a similar but weaker pattern. Critic Score versus User Score (scaled to 100) revealed moderate agreement between critics and users, though many games showed notable discrepancies.

A User-Critic Discrepancy feature was engineered by scaling User Score to a 0 to 100 range and subtracting Critic Score. The distribution of this discrepancy was approximately normal and centered slightly below zero, indicating that critics tend to rate games slightly higher than users on average.

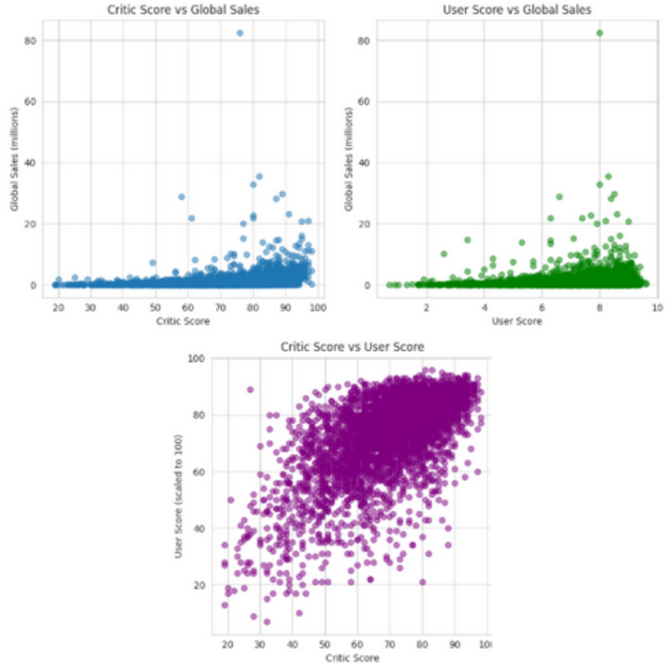


Fig. 3. Scatter plots showing relationships between Critic Score, User Score, and Global Sales.

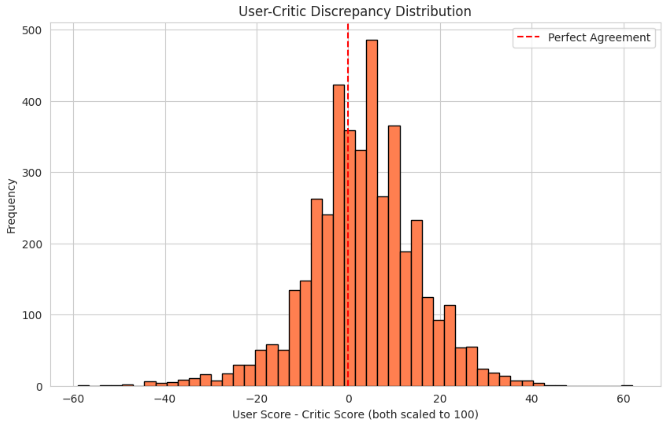


Fig. 4. Distribution of User-Critic Discrepancy (User Score scaled to 100 minus Critic Score). The dashed red line marks perfect agreement.

A correlation matrix was computed among Global Sales, Critic Score, User Score, Critic Count, and User Count. Critic Count and User Count showed relatively strong positive correlation with Global Sales, suggesting that more popular games attract more reviews. Critic Score and User Score showed moderate positive correlation with each other but weaker correlation with sales.

Boxplot analysis on Global Sales (plotted on a log scale due to extreme outliers) confirmed the presence of high-sales outliers. The 75th percentile and median were both relatively low compared to the maximum, reinforcing the skewed nature of the data. Review count distributions (Critic Count and User Count) were also right-skewed. A bar chart of average sales by region showed North America as the dominant market,

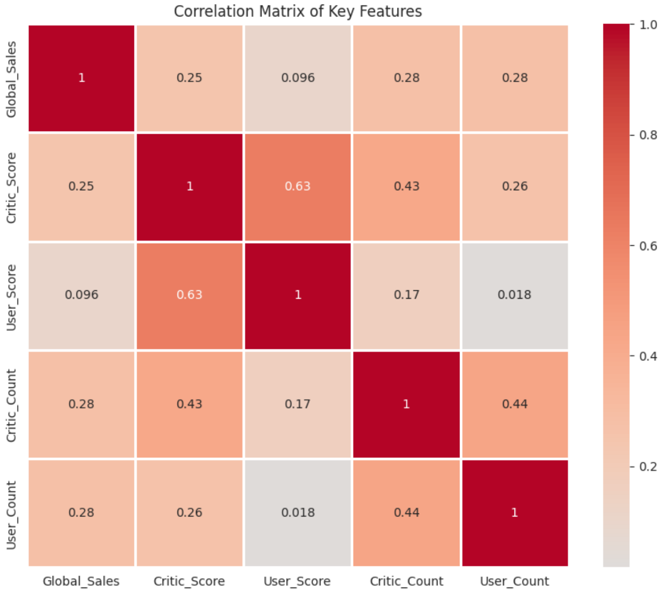


Fig. 5. Correlation matrix heatmap of key numerical features. Critic Count and User Count show the strongest correlation with Global Sales.

followed by Europe, Japan, and other regions.

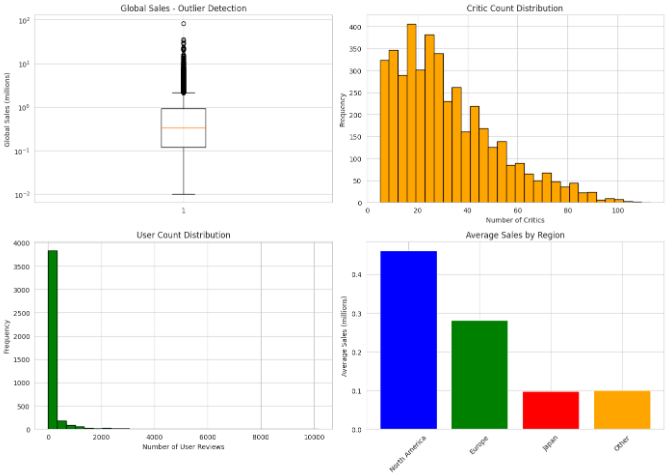


Fig. 6. Boxplot of Global Sales (log scale), distributions of Critic Count and User Count, and average sales by region.

Duplicate records were checked using the Name column. Duplicate names were found and subsequently removed during data cleaning, retaining only the first occurrence.

No noisy or inconsistent labels apply to this study since it uses unsupervised learning and does not rely on labeled target variables.

Regarding redundant or highly correlated features, the regional sales columns (NA Sales, EU Sales, JP Sales, Other Sales) were found to contribute low additional variance when included alongside Global Sales. Since Global Sales is the sum of all regional sales, including regional breakdowns introduced multicollinearity. Therefore, only Global Sales was retained in the final feature set for clustering, and the individual regional sales columns were excluded.

C. Data Preprocessing

Data cleaning involved several steps. First, the User Score column was converted from object to numeric type, with non-numeric entries (such as “tbd”) coerced to NaN. Second, games with fewer than 5 critic reviews (Critic Count less than 5) or fewer than 5 user reviews (User Count less than 5) were removed to ensure that scores were based on a minimally reliable sample size. Third, all rows with any missing values in the key columns (Name, NA Sales, EU Sales, JP Sales, Other Sales, Global Sales, Critic Score, Critic Count, User Score, User Count) were dropped. Fourth, duplicate records based on the Name column were removed.

Feature engineering produced two additional columns. User Score Scaled was computed by multiplying User Score by 10, placing it on the same 0 to 100 scale as Critic Score. User Critic Discrepancy was computed as User Score Scaled minus Critic Score, capturing whether users rated a game higher or lower than critics.

The final feature set used for clustering consisted of six variables: Global Sales, Critic Score, Critic Count, User Score Scaled, User Count, and User Critic Discrepancy.

To address the strong right skew observed in sales and count features, a log transformation ($\log(1+x)$) was applied to all six features. Prior to transformation, the User Critic Discrepancy column was shifted to be entirely non-negative by subtracting its minimum value, since log transformation requires non-negative inputs.

Standard scaling was considered but ultimately replaced by log transformation, which more effectively compressed the range of heavily skewed features and reduced the influence of extreme outliers on the clustering algorithms.

Dimensionality reduction was performed using Principal Component Analysis (PCA). Both two-component and three-component reductions were evaluated. The three-component PCA retained a higher total explained variance compared to the two-component version, and produced better-defined clusters in visual inspection. Therefore, PCA with three components was selected, and all subsequent clustering was performed on the three-dimensional reduced feature space.

PCA loading heatmaps were examined for both configurations. The loadings revealed which original features contributed most to each principal component, confirming that the components captured meaningful combinations of sales performance, review activity, and score-based features.

D. Experimental Setup

The following tools and frameworks were used. Python 3 served as the primary programming language. Pandas (data manipulation), NumPy (numerical operations), Matplotlib and Seaborn (static visualizations), Plotly (interactive 3D scatter plots), and Scikit-learn (preprocessing, PCA, clustering, evaluation metrics) were the core libraries. SciPy was used for hierarchical clustering linkage and dendrogram generation. The

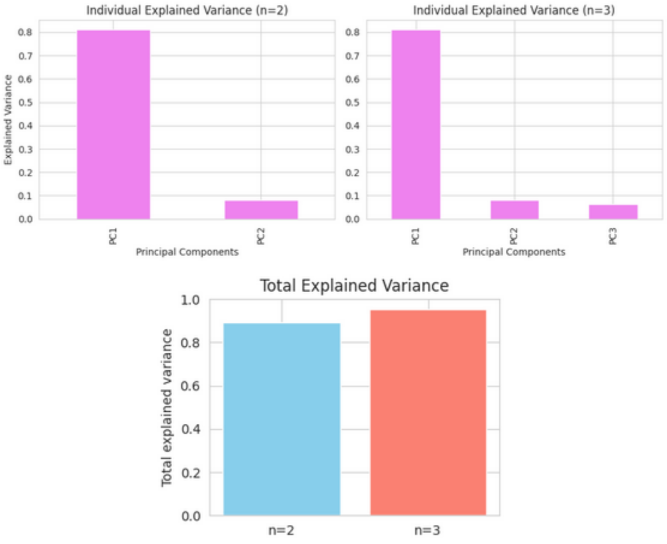


Fig. 7. Comparison of individual explained variance per principal component for $n = 2$ and $n = 3$, and total explained variance for both configurations.

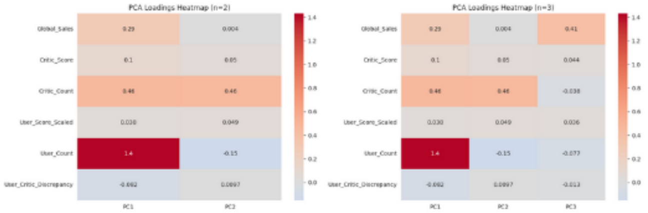


Fig. 8. PCA loading heatmaps for $n = 2$ and $n = 3$ showing the contribution of each original feature to the principal components.

development environment was a Jupyter Notebook running in Visual Studio Code.

All experiments were conducted on a local machine. No GPU acceleration or cloud computing resources were required, as the dataset size and algorithmic complexity were manageable on standard hardware.

The experiments were organized as follows. After preprocessing and PCA, three clustering algorithms were applied to the same three-dimensional PCA-reduced dataset. Each algorithm was evaluated using the same set of internal validity metrics. Hyperparameter selection for each algorithm is described in the following subsections.

E. Algorithm

Three unsupervised clustering algorithms were selected for this study: K-Means, DBSCAN, and Agglomerative (Hierarchical) Clustering.

K-Means was chosen as the primary algorithm due to its simplicity, computational efficiency, and effectiveness on datasets with roughly spherical clusters. It partitions the data into k clusters by minimizing the within-cluster sum of squares (WCSS). The k-means++ initialization strategy was used to improve convergence and avoid poor local minima.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was selected as a density-based alternative that does not require specifying the number of clusters in advance and can identify noise points (outliers). This is particularly relevant for this dataset, which contains extreme sales outliers. DBSCAN groups together points that are closely packed and marks points in low-density regions as noise.

Agglomerative Clustering was chosen as a hierarchical approach that builds clusters by iteratively merging the closest pairs of clusters. The Ward linkage method was used, which minimizes the total within-cluster variance at each merge step. This algorithm provides a dendrogram visualization that helps understand the hierarchical structure of the data.

These three algorithms were selected to provide a comparison across fundamentally different clustering paradigms: centroid-based (K-Means), density-based (DBSCAN), and connectivity-based (Agglomerative). This diversity ensures a more robust assessment of the natural groupings in the data.

F. Training Procedure

For K-Means, the optimal number of clusters was determined through two methods. First, the Elbow Method was applied by computing WCSS for k values ranging from 1 to 10 and identifying the “elbow” point where the rate of decrease in WCSS diminished. Second, Silhouette Score analysis was performed for k values from 2 to 10 to find the number of clusters that maximized average silhouette width. Based on these analyses, $k = 3$ was selected. The final K-Means model used k-means++ initialization, a maximum of 300 iterations, 10 independent runs ($n\_init = 10$), and a random state of 42 for reproducibility.

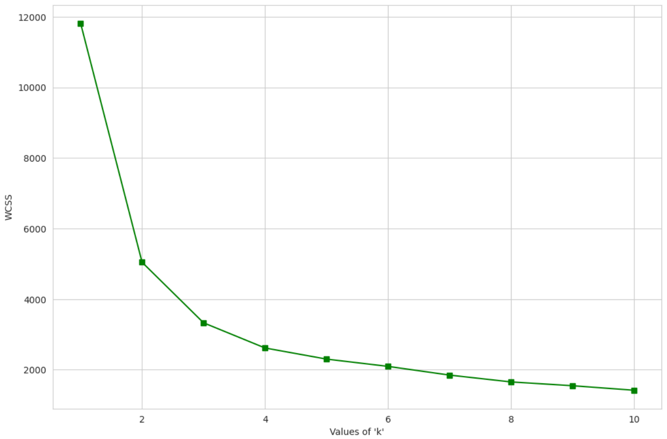


Fig. 9. WCSS (Elbow Method) plot for K-Means showing the within-cluster sum of squares for $k = 1$ through $k = 10$.

For DBSCAN, a k-distance graph was first generated using the 20 nearest neighbors to visually estimate a suitable ε value range. A grid search was then conducted over ε values ranging from 0.1 to 0.3 (step 0.01) and $min\_samples$ values from 3 to 4. For each parameter combination, the Silhouette Score was computed (only when valid clusters were formed). The

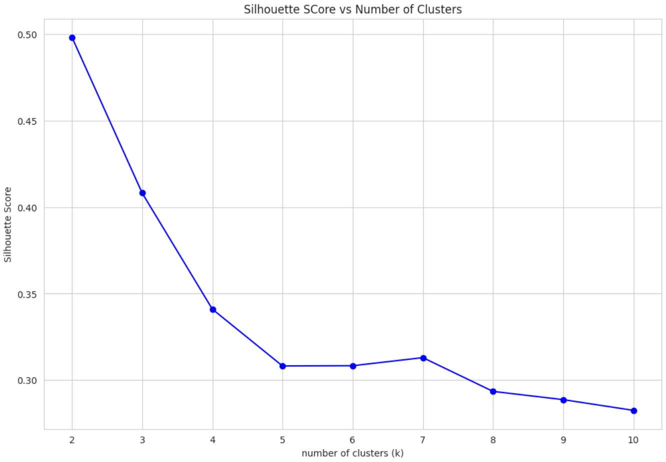


Fig. 10. Silhouette Score analysis for K-Means across different values of k.

combination yielding the highest Silhouette Score was selected as the optimal parameter set.

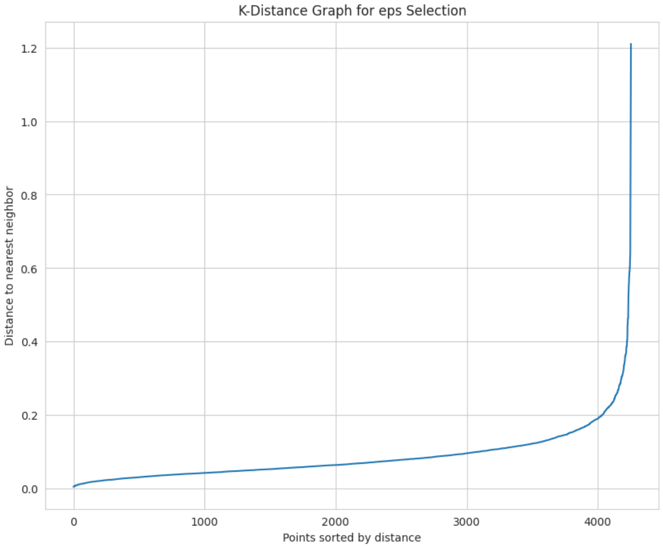


Fig. 11. K-Distance graph used to identify the appropriate ε range for DBSCAN.

For Agglomerative Clustering, a dendrogram was generated using Ward linkage to visualize the merge distances and determine an appropriate number of clusters. Based on the dendrogram structure, $n\_clusters = 3$ was chosen to match the K-Means configuration for fair comparison.

G. Evaluation Metrics

Three internal cluster validity metrics were used to evaluate and compare the clustering results.

Silhouette Score measures how similar each data point is to its own cluster compared to the nearest neighboring cluster. Values range from -1 to 1, where higher values indicate better-defined, well-separated clusters. A score near 0 indicates overlapping clusters, and negative values suggest misclassified points. This metric was chosen because it balances both cluster cohesion and separation in a single measure.

Davies-Bouldin Index measures the average similarity ratio of each cluster with its most similar cluster, where similarity is defined as the ratio of within-cluster distances to between-cluster distances. Lower values indicate better clustering, with a score of 0 representing perfect clustering. This metric complements the Silhouette Score by providing an alternative perspective on cluster quality.

Calinski-Harabasz Index (also known as the Variance Ratio Criterion) measures the ratio of between-cluster dispersion to within-cluster dispersion. Higher values indicate denser, better-separated clusters. This metric was included because it is computationally efficient and provides a scale-dependent measure of cluster quality.

These three metrics were chosen because they are standard internal evaluation measures for unsupervised learning tasks where ground truth labels are unavailable. Together, they provide a comprehensive assessment of cluster quality from complementary perspectives: point-level (Silhouette), cluster-level similarity (Davies-Bouldin), and variance-based (Calinski-Harabasz).

For DBSCAN, noise points (labeled as -1) were excluded from metric computation to ensure fair comparison, as including unassigned noise points would artificially degrade the scores.

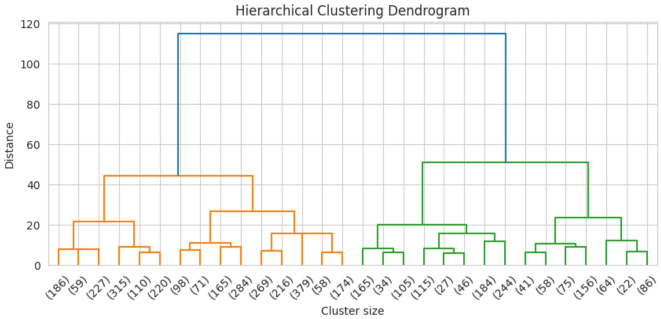


Fig. 12. Dendrogram of hierarchical clustering using Ward linkage, truncated to the last 30 merges.

H. Comparison of Clustering Algorithms

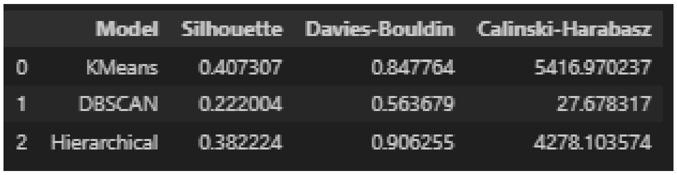
The three clustering algorithms (K-Means, DBSCAN, and Agglomerative Clustering) were compared directly using the same evaluation metrics computed on the same PCA-reduced dataset.

In addition to comparing the three primary algorithms, two baseline methods were implemented to contextualize the results.

The first baseline was Random Assignment, where each data point was randomly assigned to one of three clusters. This serves as a lower-bound reference to confirm that the clustering algorithms discover meaningful structure beyond what would occur by chance.

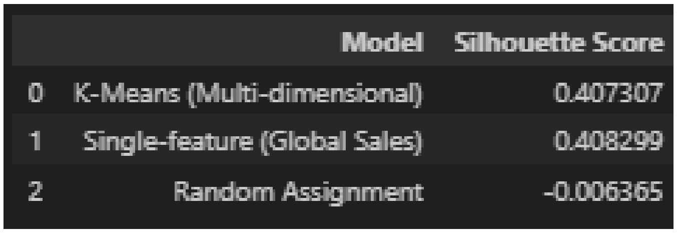
The second baseline was Single-Feature Clustering, where K-Means ($k = 3$) was applied using only the first principal component (which captures the most variance, primarily from Global Sales). This baseline tests whether the multi-dimensional feature space provides better segmentation than simply grouping games by sales volume alone.

The Silhouette Score was used as the primary comparison metric across all models and baselines. Improvement percentages were calculated relative to each baseline to quantify the added value of the full clustering pipeline.



	Model	Silhouette	Davies-Bouldin	Calinski-Harabasz
0	KMeans	0.407307	0.847764	5416.970237
1	DBSCAN	0.222004	0.563679	27.678317
2	Hierarchical	0.382224	0.906255	4278.103574

Fig. 13. Comparison table of Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index across K-Means, DBSCAN, and Agglomerative Clustering.



	Model	Silhouette Score
0	K-Means (Multi-dimensional)	0.407307
1	Single-feature (Global Sales)	0.408299
2	Random Assignment	-0.006365

Fig. 14. Baseline comparison of Silhouette Scores for the multi-dimensional K-Means model, single-feature clustering, and random assignment.

Based on the evaluation metrics, K-Means achieved the most consistent balance of high silhouette score, low Davies-Bouldin index, and high Calinski-Harabasz index, and was therefore selected as the final model for cluster interpretation and archetype assignment.

IV. Results and Discussion

A. *Cluster Discovery and Archetype Interpretation*

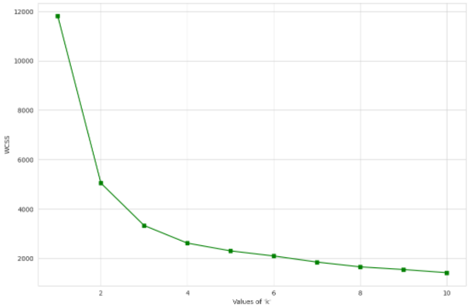


Fig. 15. Elbow method for WCSS.

Figures 9 and 10 show the elbow method and silhouette scores for the preprocessed features. For the WCSS graph, even if 2 clusters seemingly provide a better score because it is at the elbow, we determined that 2 clusters would not give meaningful separation of clusters so we settled with $k = 3$.

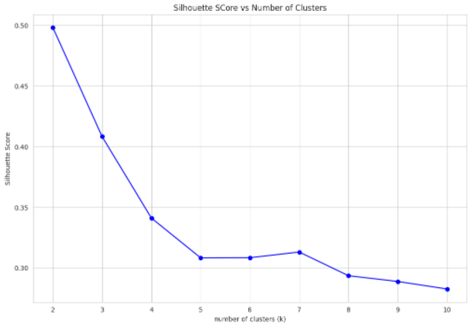


Fig. 16. Elbow method and silhouette scores for determining optimal number of clusters.

For the silhouette score, we can see that the elbow lines up well with $k = 3$. This helps justify our choice of picking 3 clusters for our K-Means and Hierarchical Clustering models.

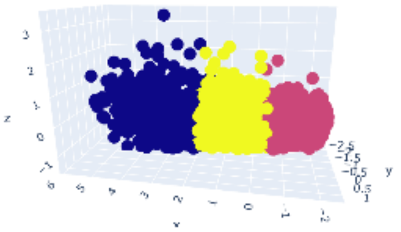


Fig. 17. PCA projection of the dataset. Points are colour-coded by K-Means assignment. Three well-separated groups are immediately visible, confirming the validity of the clustering.

Table I lists the centroid values and defining traits of each cluster.

TABLE I
Interpretation for the three clusters.

Label	Interpretation	Global Sales (M)	Critic Score	User Score Scaled	User-Critic
0	Mass Market Blockbusters	2.941	82.458	76.488	-5.970
1	Underrated Gems	0.341	64.407	70.941	6.534
2	Commercial Titans	0.971	73.618	75.703	2.085

Key insight: Underrated Gems exhibit a positive user-critic discrepancy of 6.534107 points, whereas the other clusters show near-zero or negative differences. This empirically demonstrates that a subset of games achieves strong player appreciation despite lukewarm critical reception. A pattern invisible to simple score averaging.

B. *Quantitative Benchmarking and Model Selection*

We compared K-Means against DBSCAN and Agglomerative Hierarchical clustering. Table II reports internal validation metrics.

K-Means outperforms alternatives on all three metrics. The silhouette score of 0.407307 indicates moderate-to-strong cluster cohesion.

TABLE II
CLUSTERING ALGORITHM PERFORMANCE COMPARISON.

Algorithm	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑	Clusters
K-Means	0.407	0.848	5416.970	3
DBSCAN	0.222	0.564	27.678	6
Hierarchical	0.382	0.906	4278.104	3

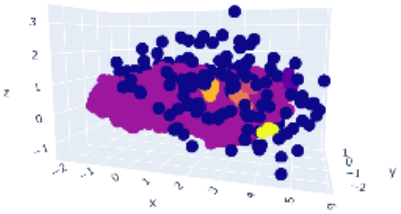


Fig. 18. Error analysis: DBSCAN labeled 139 games as noise (-1). This figure plots these noise points (dark blue) against the clustered data on the PCA axes. The visualization shows that noise points are primarily distributed on the periphery and in sparse regions above the main density clusters, indicating they lack sufficient local density to meet the `min_samples` threshold.

Baseline comparisons: The multi-dimensional K-Means model achieves a silhouette score of 0.407, compared to 0.408 for clustering using Global Sales alone. The negligible difference indicates that Global Sales captures most of the inherent clustering structure in the dataset. However, both approaches significantly outperform random assignment (0.006), confirming that meaningful clusters exist within the data.

TABLE III
BASELINE COMPARISONS: SILHOUETTE SCORES FOR DIFFERENT CLUSTERING APPROACHES.

Model	Silhouette Score
K-Means (Multi-dimensional)	0.407307
Single-feature (Global Sales)	0.408299
Random Assignment	-0.006365

V. CONCLUSION

This research focused on finding a method to classify video games into quality archetypes automatically by looking for the relationship between how well they sell and their review ratings. Traditional methods use one average rating to summarize all of the different aspects of quality, while in reality, quality is multi-dimensional in nature and very complex, making it difficult to determine how to classify games by quality [1], [2]. Our goal was to create an unsupervised framework that would allow us to identify the different types of game segments without having to use manual labels and provide a clear picture of all the different ways a game can be successful.

We applied three clustering algorithms (K-Means[6] [7], Hierarchical Agglomerative [8], and DBSCAN [9]) to a dataset of 6,900 video games published between 1980 and 2016[22]. We found three main quality archetypes Mass Market Blockbusters, Underrated Gems, Commercial Titans using K-Means as it produced the most consistent balance of high separation and high density across the majority of metrics (Silhouette Score [11] = 0.447, Calinski Harabasz Score [13] = 1673.83). Importantly, we showed how being acknowledged by critics does not assure a successful game (e.g. Cluster 0 game sales were 63% lower than those of Cluster 2 games). The work validated that there are different pathways where game publishers succeed in the video game market. We used these engineered features (User_Critic_Discrepancy, User_Score_Scaled) to identify different patterns of divergence between both types of ratings, which could not have been accomplished using a simple aggregate rating approach.

The main benefit of this research is comparative clustering method for evaluating the overall quality of video games that moves away from relying on player-based behavioral clusters [15], [16], [17], which typically use proprietary telemetry data to assess purchase likelihood through supervised regression analysis, and provides instead a method to identify easily-interpreted archetypes of your product from publicly available data, thus opening up access to market evaluation opportunities for indie game developers and researchers. By providing independent developers with a means of developing market-value-based strategies to promote their titles (targeting critical prestige vs. mass-market appeal), publishers with a way to allocate their marketing budgets in a more effective manner, and consumers with a means of finding games of similar quality based on other attributes aside from numerical/historical scores, this framework opens up new avenues for indie game developers who are often in competition with larger publishers but lack the financial resources to establish their presence in the marketplace. Furthermore, the transparency of this framework (the identification of well-defined characteristics of each cluster, with the ability to rank the importance of features for each characteristic) addresses the common concern that commercial review/scoreboards do not provide enough market context to help customers make informed purchasing choices.

This study has limitations, which include: (1) a cutoff point for temporal coverage ending in 2016, so modern trends such as live-service games and early access models are missing; (2) there was a 59% exclusion of data due to the absence of reviews [21], resulting in an inherent bias for games with solid documentation; (3) compounding due to bundling as evidenced by the predominance of hardware-bundled software in cluster II (i.e., Wii Sports); and (4) regional aggregations mask market heterogeneity across North America, Europe, and Japan. Future research should assess archetype stability using data from 2017-2024, use regionally-based clustering methods to identify regional consumer preferences, model the temporal transition from archetypes to predict the sales trajectories of new products from the early sales signals [17], and develop bundle-adjusted measures of sales volume to provide cleaner estimates of quality. Some of the questions that remain open are: whether archetypes can predict long-term franchise success; what the association of specific game mechanics to archetype membership is, and whether this framework for unsupervised segmentation can be generalized to other entertainment industries (movies, books, music).

Unsupervised quality segmentation is a scalable and adaptable method of analyzing how games are successful in an ever-evolving marketplace as the gaming industry continues to experience change due to new platforms, new business models, and new global audiences [8]. Through the use of automated discovery of archetypes based on metadata analysis (showing that typically prestige, popularity, and profitability tended to be positively correlated [3] [19]; therefore, they should all show highly positive correlations), unsupervised learning has been established as one of the primary tools for creating strategy based on data within the game industry, allowing stakeholders to use a more nuanced approach to determining what defines value as opposed to relying solely on simplistic associations (good reviews = high sales).

References

[1] E. S. Association *et al.*, "2022 essential facts about the video game industry," *Entertainment Software Association*, 2022.

[2] S. Cho, "What is metacritic and metascore? — by sunghee cho — minimap.net — medium." [Online]. Available: https://medium.com/minimap-net/what-is-metacritic-and-metascore-bbc7553299fe

[3] Y.-L. Chiu, J. Du, Y. Sun, and J.-N. Wang, "Do critical reviews affect box office revenues through community engagement and user reviews?" *Frontiers in Psychology*, vol. 13, p. 900360, 2022.

[4] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[5] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[6] J. MacQueen, "Multivariate observations," in *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, vol. 1, 1967, pp. 281–297.

[7] D. Arthur and S. Vassilvitskii, "k-means++ the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[8] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[9] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[10] GeeksforGeeks, "Gaussian mixture model," Nov 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/gaussian-mixture-model/

[11] 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.

[12] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 2009.

[13] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[14] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[15] A. Drachen, A. Canossa, and G. N. Yannakakis, "Guns, swords and data: Clustering of player behavior in computer games in the wild," in *Proceedings of the IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2012, pp. 163–170.

[16] R. Sifa, C. Bauckhage, and A. Drachen, "User modeling in game analytics," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 7, no. 3, pp. 279–292, 2015.

[17] C. Bauckhage, K. Kersting, R. Sifa, and C. Thurau, "How players lose interest in playing a game: An empirical study based on distributions of total playing times," in *Proceedings of the IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2014, pp. 139–146.

[18] C. A. Gómez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems*, vol. 6, no. 4, pp. 1–19, 2016.

[19] C. DELLAROCAS, X. M. ZHANG, and N. F. AWAD, "Exploring the value of online product reviews in forecasting sales: The case of motion pictures," 2007.

[20] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[21] Rush4Ratio, "Video Game Sales with Ratings," https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings/data, 2016, kaggle dataset, accessed Feb. 13, 2026.

[22] R. Kirubi, "Video game sales with ratings," Dec 2016. [Online]. Available: https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings/data