Beyond the Store Page: Clustering Steam Games by Player Engagement and Pricing

Antonio Mark Allen B.
College of Computing and Information Technologies
National University Manila
Manila, Philippines
antoniomb1@students.national-u.edu.ph

Mago, Karl Mattheus W.
College of Computing and Information Technologies
National University Manila
Manila, Philippines
magokw@students.national-u.edu.ph

***Abstract*—Steam is one of the largest PC game marketplaces, but its store page does not show how players actually interact with games. Many games have similar prices but very different playtime, review counts, and popularity. In this study, we use unsupervised machine learning to group Steam games by engagement and pricing features. We tested several clustering algorithms, including K-Means, Agglomerative Hierarchical Clustering, and Gaussian Mixture Models (GMM). We used Principal Component Analysis (PCA) to reduce the number of features while keeping at least 90% of the data's variance. Our results show that K-Means gave the clearest and most stable clusters. The final model used a filtered dataset of paid games with non-zero playtime, resulting in three clusters that show different patterns of engagement and market behavior. We also ran an experiment with a dataset that included all games, showing that including games with zero playtime can inflate evaluation scores and make the clusters harder to interpret. These findings help explain how pricing, engagement, and popularity are related among Steam games.**

***Index Terms*—Steam games, clustering, K-Means, PCA, player engagement, pricing, unsupervised learning**

I. Introduction

Steam has tens of thousands of games with varying prices, release dates, and levels of popularity [1]. Store pages display prices and discounts, yet these indicators do not necessarily correspond to actual player engagement. Some low-cost games accumulate extensive playtime and review activity, whereas certain higher-priced titles attract minimal attention [2]. Consequently, it is difficult to construct meaningful groupings of games using only visible store information.

Unsupervised learning provides mechanisms for identifying natural structures in data without predefined labels. In this study, clustering techniques are applied to segment Steam games using variables such as average playtime, review counts, ownership estimates, peak concurrent users, and the proportion of positive feedback. The objective is to uncover groups that reflect real player behavior rather than surface-level characteristics.

A. Problem Statement

Steam games differ a lot in popularity and engagement, even when their prices are similar. There is no easy way to label games by how engaged their players are. This study uses pricing and engagement features to group Steam games into meaningful categories.

B. Objectives

This project aims to preprocess and engineer features from a Steam dataset, use principal component analysis (PCA) to reduce dimensionality, train and compare clustering models such as K-means, hierarchical clustering, and Gaussian Mixture Model, choose the best model using internal evaluation metrics, and interpret the clusters with profiling statistics and visualizations.

C. Scope and Limitations

This study looks at clustering paid Steam games and does not include free-to-play games. The results may be affected by extreme values, missing playtime data, and skewed distributions. While PCA helps reduce dimensionality, the results still depend a lot on the chosen features and preprocessing methods.

II. Review of Related Literature

Game platforms such as Steam generate large datasets reflecting player engagement, pricing dynamics, and review responses across thousands of titles. Prior work in data science and game analytics has demonstrated that features derived from playtime, review volume, and pricing can be effective for understanding performance patterns. For example, Lim analyzed the relationship between hours played per dollar and game quality, showing that simple ratio metrics can reveal differences in value-for-money even when raw playtime and price vary widely across titles [3]. This underscores the importance of feature transformations such as ratios or logarithmic scaling when interpreting behavioral differences among games.

Supervised learning approaches have also been applied to Steam data to predict game "success," typically defined using engagement proxies. Chan utilized SteamSpy metadata, including price, genre, ownership estimates, review scores, and playtime, to train a Random Forest classifier and found that engagement and review-related features were strong predictors of above-median playtime, while price exhibited weaker influence [4]. Although supervised models can provide strong predictive performance, they rely on predefined labels and do not inherently reveal latent structures. This motivates the use of unsupervised clustering techniques to uncover patterns without prior categorization.

Clustering large behavioral datasets introduces challenges related to feature scaling, dimensionality, and validation. Rousseeuw's silhouette method remains a foundational metric for evaluating cluster quality by comparing intra-cluster cohesion with inter-cluster separation, particularly in scenarios where ground-truth labels are unavailable [5]. Within game analytics, Drachen and colleagues highlight additional concerns, including the interpretability of clusters and the risk that numerically strong solutions may still lack practical meaning if features are poorly chosen [6]. These perspectives emphasize the necessity of combining careful feature engineering with multiple validation metrics, such as silhouette, Davies–Bouldin, and Calinski–Harabasz indices.

Another important consideration involves the distributional characteristics of engagement variables. Liang and Yang observed that PC game playtime follows a heavily skewed distribution, where a small subset of titles dominates long durations while the majority receive limited interaction [7]. They showed that logarithmic transformations can mitigate the influence of extreme values and improve modeling stability. This supports methodological choices in clustering, including the preference for ratio-based or log-scaled attributes rather than relying solely on raw magnitudes that may dominate distance computations.

In summary, prior research indicates that (1) engagement and pricing variables contain meaningful signals, (2) unsupervised clustering is suitable for exploratory segmentation without labels, (3) numerical validation must be complemented by interpretability, and (4) appropriate feature transformation is crucial when handling skewed behavioral metrics. These principles guide the methodological decisions in this study, including the use of PCA for dimensionality reduction, comparison across clustering algorithms, and the construction of features designed to yield interpretable groupings.

III. Methodology

A. Dataset

The dataset used in this study was obtained from Kaggle and contains structured information about Steam games. The variables cover pricing, popularity, and player engagement, as summarized in Table I.

TABLE I
Dataset Attributes (Steam Games Dataset)

Attribute	Description
name	Steam game title
price	Current game price (USD)
release_date	Date when the game was released
positive	Number of positive user reviews
negative	Number of negative user reviews
owners	Estimated ownership range
average_playtime_forever	Average total playtime (minutes)
median_playtime_forever	Median total playtime (minutes)
required_age	Minimum required age
dlc_count	Number of DLCs available
metacritic_score	Metacritic rating score (if available)
achievements	Number of achievements
recommendations	Number of Steam recommendations
supported_languages	Supported languages
categories	Steam categories (e.g., Single-player, Co-op)
genres	Steam genres (e.g., Action, RPG)
tags	User-generated Steam tags

This dataset was selected because it combines economic indicators (e.g., price) with engagement-related measures (e.g., reviews, playtime, and concurrent users), making it appropriate for analyzing value and behavioral patterns across games.

B. Data Cleaning and Preprocessing

Several preprocessing steps were performed prior to clustering in order to improve consistency, reduce noise, and ensure stable model behavior.

1) Filtering: The dataset was restricted to paid games. Free-to-play titles often follow different monetization and engagement dynamics, which can distort comparisons when analyzing price-related behavior. Limiting the dataset to paid products improves fairness and interpretability.

2) Handling Missing and Invalid Values: Missing or invalid entries were addressed through removal or imputation depending on the importance of the feature. Records lacking values in critical clustering variables were excluded to avoid instability during PCA and model training.

3) Release Date Transformation: The release date was converted into a numerical variable, *release_age_years*, representing the time elapsed since the game's publication. This allows the model to account for the fact that older titles have had more opportunity to accumulate owners, reviews, and playtime.

4) Feature Standardization: Prior to PCA and clustering, numerical variables were standardized using z-score normalization (mean = 0, standard deviation = 1). This prevents high-magnitude features, such as ownership or review counts, from dominating distance-based calculations.

C. Feature Set Construction

To explore clustering behavior under different assumptions, multiple feature sets were created. However, two datasets were selected for the final analysis because they yielded the most interpretable and defensible results.

1) Main Dataset: X2 Balanced: The primary dataset contains paid games with strictly positive average lifetime playtime. This decision addresses a frequent issue in Steam analytics known as *zero-playtime bias*.

Games without recorded playtime tend to form a trivially separable group, which can artificially inflate clustering evaluation scores. Although this may improve numerical metrics, it does not produce meaningful segmentation among titles that demonstrate real user engagement. Therefore, X2 was adopted as the final modeling dataset because it focuses exclusively on games with observable activity.

2) Supporting Dataset: X3 Ratios (ALL): The supporting dataset includes all paid games and incorporates ratio- and log-based engagement features. These transformations reduce the dominance of raw magnitude differences, such as extremely high playtime or review counts, that might otherwise drive clustering behavior.

X3 was included to complement the analysis by illustrating how the presence of low-engagement and zero-playtime titles influences the overall structure of the clusters.

D. Dimensionality Reduction

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset and improve clustering efficiency. The technique transforms standardized variables into a smaller set of orthogonal components that capture the dominant directions of variation in the data.

Only the minimum number of components required to retain at least 90% of the total variance (PCA90) was preserved. This strategy helps remove redundancy and noise while maintaining most of the information necessary for reliable clustering.

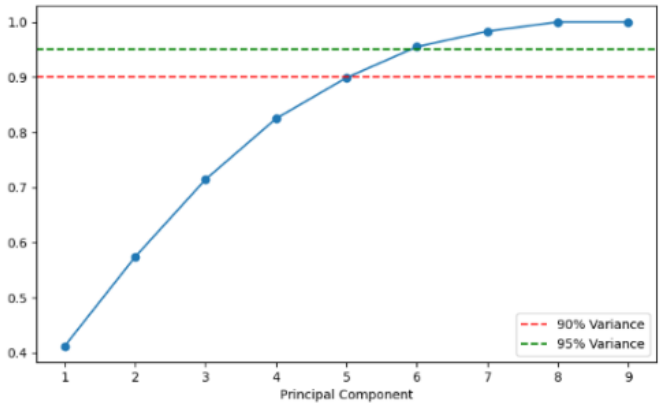

Fig. 1. Explained variance captured by principal components.

E. Clustering Algorithms

Three clustering approaches were evaluated to determine which method produced the most stable and interpretable segmentation of Steam games.

1) K-Means: K-Means served as the primary baseline due to its conceptual simplicity, computational efficiency, and ease of interpretation. The algorithm partitions observations by minimizing within-cluster variance. Multiple values of k were explored to assess sensitivity to the number of clusters.

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

where C_i represents cluster i and μ_i its centroid.

2) Agglomerative Hierarchical Clustering: Agglomerative hierarchical clustering was examined using common linkage strategies. This method was included because it can reveal nested group structures and does not rely on probabilistic distributional assumptions.

3) Gaussian Mixture Models (GMM): Gaussian Mixture Models were evaluated as a probabilistic alternative to K-Means. Unlike hard assignments in K-Means, GMM provides soft cluster memberships and is capable of representing elliptical cluster shapes, offering greater modeling flexibility.

F. Evaluation Metrics

Because the dataset does not contain ground-truth labels, clustering performance was assessed using internal validation measures.

1) Silhouette Score: The silhouette coefficient evaluates how similar an observation is to its assigned cluster compared with other clusters. Higher values indicate better separation and cohesion.

It is defined as:

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

where $a(i)$ represents the average intra-cluster distance and $b(i)$ represents the lowest average inter-cluster distance. The score ranges from -1 to 1, with higher values indicating better-defined clusters.

2) Davies–Bouldin Index: The Davies–Bouldin index quantifies the ratio of within-cluster dispersion to between-cluster separation. Lower values correspond to more distinct and well-separated clusters.

It is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right)$$

where S_i and S_j represent intra-cluster dispersion and M_{ij} represents the distance between cluster centroids. Lower DBI values indicate better clustering quality.

3) Calinski–Harabasz Index: The Calinski–Harabasz score measures the ratio of between-cluster variance to within-cluster variance. Higher values suggest more compact and better-defined groupings.

It is defined as:

$$CH = \frac{\mathrm{Tr}(B_k)/(k-1)}{\mathrm{Tr}(W_k)/(n-k)}$$

where B_k is the between-cluster dispersion matrix and W_k is the within-cluster dispersion matrix. Higher values indicate better-defined cluster separation.

These metrics provided quantitative guidance for comparing model configurations. However, final model selection also incorporated qualitative considerations, including interpretability and the balance of cluster sizes.

IV. Results and Discussion

This section presents the results of the proposed clustering approach for Steam games. The clustering experiments were conducted using multiple feature sets and algorithms; however, the final analysis focused on the most interpretable and defensible results. Two main outputs are discussed:

X2 Balanced as the main clustering model

X3 Ratios as a supporting result, since it highlights how engagement-to-price style features naturally separate high-traction games from the long tail of Steam titles.

A. PCA Dimensionality Reduction Results

Before clustering, Principal Component Analysis (PCA) was applied to reduce feature dimensionality while preserving the majority of information contained in the dataset. The goal of PCA in this study was not to "improve" the dataset artificially, but to provide a stable, low-dimensional representation that makes clustering more efficient and less noisy.

For the datasets used in the final analysis:

- X2 Balanced (playtime ¿ 0) used 6 principal components to retain at least 90% cumulative explained variance.
- X3 Ratios (ALL) also used 6 principal components to retain at least 90% cumulative explained variance.

The PCA results show that the first principal component (PC1) explains roughly 40% of the variance in both datasets, while PC2 contributes approximately 15–16%. This indicates that the dataset contains a strong dominant structure that PCA captures early, making it suitable for clustering.

Dataframe	Principal Component	Cumulative Explained Variance
X2	PC1	41.24%
	PC2	57.41%
	PC3	71.46%
	PC4	82.48%
	PC5	89.90%
	PC6	95.50%
X3	PC1	39.95%
	PC2	56.33%
	PC3	70.24%
	PC4	80.45%
	PC5	89.26%
	PC6	96.34%

Fig. 2. Comparison of Principal Component Explained Variance across chosen Dataframes

Key observation: Since PCA90 only required 6 components out of 9 features, the clustering process can be performed with reduced computational cost while still retaining most of the dataset's information.

B. Selection of the Main Clustering Model

The study evaluated clustering performance across multiple algorithms, including K-Means, Gaussian Mixture Models (GMM), and Hierarchical clustering. The performance of each configuration was measured using standard internal clustering evaluation metrics:

- Silhouette Score (higher is better)
- Davies–Bouldin Index (lower is better)
- Calinski–Harabasz Index (higher is better)

For the final model selection, interpretability and defensibility were prioritized over simply choosing the highest metric value. This is because clustering is unsupervised, and internal metrics alone cannot guarantee that the resulting groups are meaningful for real-world interpretation.

The final main model was chosen as: X2 Balanced (playtime ¿ 0) + PCA90 + KMeans (k = 3)

This configuration was selected because it produced:

- reasonable silhouette performance among tested values
- cluster sizes that were not overly imbalanced
- clusters that were interpretable based on pricing, popularity, and engagement features

Dataframe	*k*	Silhouette	Davies--Bouldin	Calinski--Harabasz
X2	2	0.269	1.422	2741.3
	3	0.273	1.349	2176.3
	4	0.211	1.465	1998.1
	5	0.199	1.350	1850.4
X3	2	0.411	1.277	16883.2
	3	0.267	1.463	14588.3
	4	0.229	1.414	12738.2
	5	0.247	1.270	12164.4

Fig. 3. Model Selection Results

C. Main Clustering Output: X2 (playtime ¿ 0), KMeans k = 3

1) Cluster Size Distribution: The main dataset (X2) contained 6,782 games (paid titles with playtime greater than zero). The KMeans model produced three clusters with the following distribution:

- Cluster 0: 2,133 games (31.45%)
- Cluster 1: 879 games (12.96%)
- Cluster 2: 3,770 games (55.59%)

This distribution indicates that the model does not collapse into one dominant group. Instead, it provides three distinct segments that represent different engagement and pricing behaviors.

TABLE II
X2 MAIN DATASET: CLUSTER SIZES (K-MEANS, $k = 3$)

Cluster	Count	Percentage (%)
0	2133	31.45
1	879	12.96
2	3770	55.59

2) Cluster Profiling Using Mean Statistics: To interpret the clusters, mean values were computed across the main features used in the balanced dataset. The following features were analyzed:

- price
- discount
- release_age_years
- owners_mid
- num_reviews_total
- pct_pos_total
- peak_ccu
- average_playtime_forever

TABLE III
X2 MAIN DATASET: CLUSTER MEANS (K-MEANS, $k = 3$)

Cluster	Price	Discount	Release Age (Years)	Owners (Mid)	Total Reviews	% Positive	Peak CCU	Avg. Playtime
0	23.601	3.421	6.031	1254156.118	20973.847	85.660	1470.802	1655.160
1	4.295	67.311	7.242	207292.378	2018.233	78.297	70.191	1328.688
2	10.275	0.473	7.753	108505.305	552.242	76.202	4.968	284.024

The mean profile reveals strong differences across clusters, especially in popularity-related features (owners_mid, reviews, peak_ccu), and engagement-related features (average_playtime_forever).

3) Cluster Profiling Using Median Statistics: Because Steam data is highly skewed (a small number of games dominate owners and reviews), medians were also computed. Medians are important because they represent the "typical" game in each cluster more accurately than means.

TABLE IV
X2 MAIN DATASET: CLUSTER MEDIANS (K-MEANS, $k = 3$)

Cluster	Price	Discount	Release Age (Years)	Owners (Mid)	Total Reviews	% Positive	Peak CCU	Avg. Playtime
0	19.990	0.0	5.407	350000.0	4733.0	88.0	66.0	674.0
1	2.990	70.0	7.502	75000.0	545.0	81.0	3.0	202.0
2	9.795	0.0	8.437	35000.0	287.0	79.0	1.0	145.5

Fig. 4. X2 MAIN: Mean price (USD) by cluster

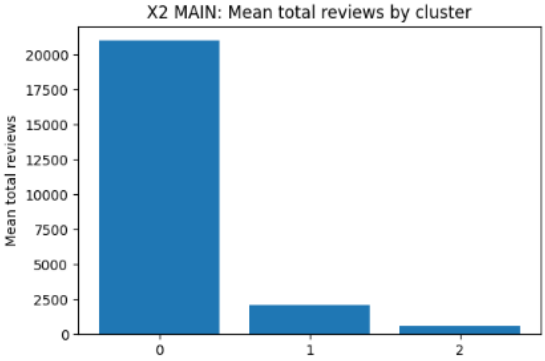

Fig. 5. X2 MAIN: Mean total review by cluster

Notably, the medians confirm that the clusters differ not only due to extreme outliers, but also in their typical pricing and engagement patterns.

D. Interpretation of the X2 Clusters

Based on the profiling tables and the top examples per cluster, the three clusters can be interpreted as follows:

1) Cluster 0 (31.45%): High-Traction, High-Engagement Titles: Cluster 0 contains the most popular and high-engagement games. This cluster has:

- the highest owners_mid
- the highest num_reviews_total
- the highest peak_ccu
- the highest average_playtime_forever
- high pct_pos_total

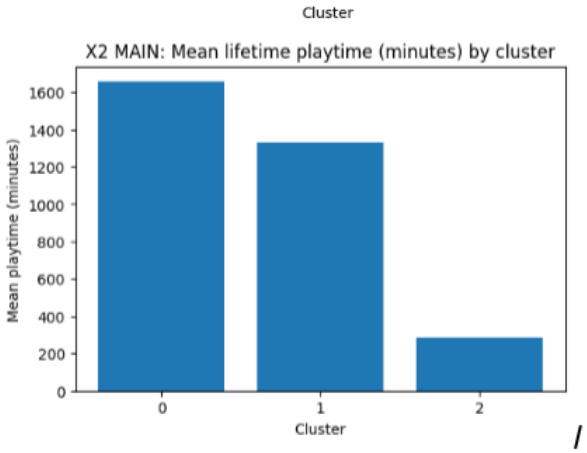

Fig. 6. X2 MAIN: Mean Lifetime playtime (minutes) by cluster

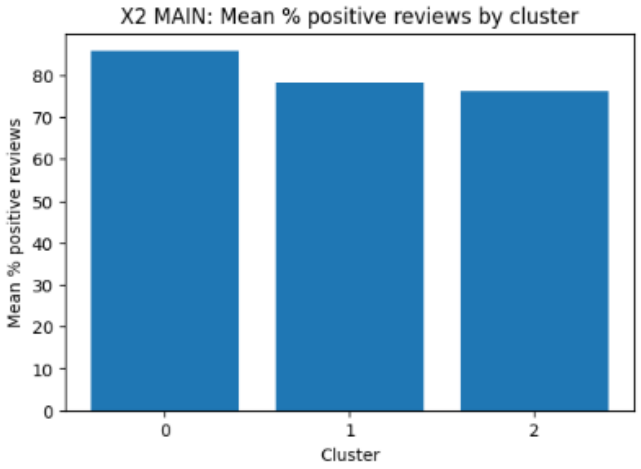

Fig. 7. X2 MAIN: Mean percentage positive reviews by cluster

The median price is also higher than the other clusters, suggesting these are established paid titles that still generate strong engagement. Examples in this cluster include:

- Tom Clancy's Rainbow Six Siege
- Rust
- Garry's Mod
- The Witcher 3
- Elden Ring

These examples confirm that Cluster 0 represents mainstream games with strong player retention and community activity.

2) Cluster 1 (12.96%): Discount-Driven Mid-to-High Engagement Purchases: Cluster 1 is characterized by:

- very low price
- extremely high discount levels (mean discount ∼67%)
- moderate-to-high playtime
- moderate popularity indicators (owners, reviews)

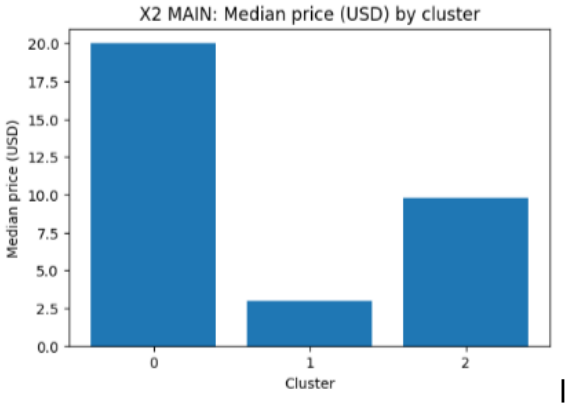

Fig. 8. X2 Main Dataset: Median values by cluster

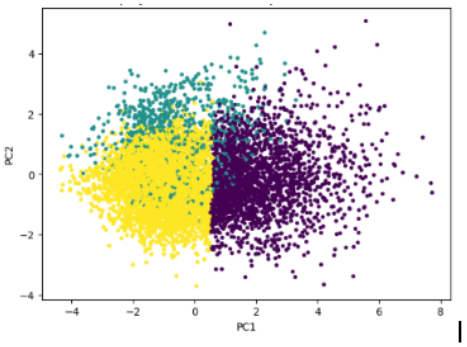

Fig. 9. High-Traction Cluster 0 Examples

This suggests that Cluster 1 represents games that are often purchased during sales and promotions, but still provide meaningful engagement. Examples include:

- Metro Exodus
- Metro 2033 Redux
- Rising Storm 2: Vietnam
- LIMBO

These titles align with the interpretation of "value purchases," where users likely acquire the game during discounts and then spend significant time playing.

3) Cluster 2 (55.59%): Low-Traction, Low-Engagement Long-Tail Games: Cluster 2 is the largest cluster. It contains games with:

- low owners
- low reviews
- extremely low peak_ccu
- low lifetime playtime
- slightly lower pct_pos_total

The median owners_mid and review counts are much smaller, indicating these are not widely played games. Many titles in this cluster appear to be smaller indie games or

niche releases that do not reach large audiences. This cluster represents the "Steam long tail," where most games exist but receive minimal engagement.

E. Supporting Result: X3 Ratios (ALL), KMeans $k = 2$

To support the main findings, the study also examined X3 Ratios (ALL). Unlike X2, this dataset includes all paid games, including those with zero playtime. X3 uses ratio-based and log-scaled engagement features, which emphasize relative engagement differences rather than raw playtime values.

The clustering on X3 produced two clusters:

- Cluster 0: 8,128 games (19.76%)
- Cluster 1: 33,008 games (80.24%)

TABLE V
X3 SUPPORT DATASET: CLUSTER SIZES (K-MEANS, $k = 2$)

Cluster	Count	Percentage (%)
0	8128	19.76
1	33008	80.24

1) Interpretation of X3 Clusters: The X3 result strongly separates Steam games into:

- Cluster 0: high-traction, high-engagement titles
- Cluster 1: the long-tail majority with minimal traction

Cluster 0 contains games with significantly higher owners_mid, reviews, and peak_ccu, while Cluster 1 contains the majority of games with low visibility and minimal engagement. This supporting result is consistent with the expected structure of the Steam market, where a small percentage of games generate most of the player activity.

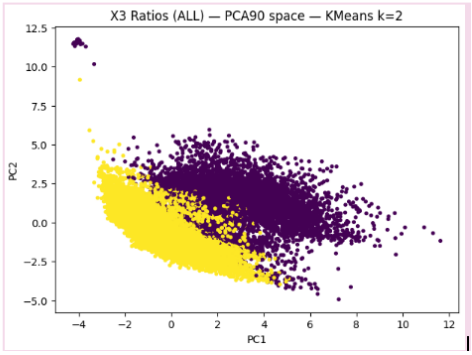

Fig. 10. X3 Support Dataset: Cluster Distribution

2) Why X3 is Used as Supporting and Not Main: X3 includes games with zero average playtime. As observed in earlier experiments, the presence of zero average playtime games can artificially increase silhouette scores by creating an easy separation between "played" and "not played." While this produces clean metric values, it may not represent meaningful segmentation of engaged players. Therefore, X3 is used as supporting evidence for market structure, while X2 remains the main model for interpreting engagement and pricing among active games.

F. Summary of Findings

The clustering results provide a clear segmentation of Steam paid games:

- A smaller group of high-traction games that dominate owners, reviews, and engagement
- A discount-driven cluster that represents value purchases with strong playtime
- A large long-tail cluster that contains most games with low traction and low engagement

These clusters can be used to support insights related to pricing strategy, player engagement, and the Steam marketplace structure.

V. CONCLUSION

Clustering Steam games based on pricing and engagement-related features yields meaningful groupings that effectively summarize marketplace performance. Among K-Means, Hierarchical Clustering, and Gaussian Mixture Models tested on various feature sets, K-Means consistently produced the most interpretable results according to the Silhouette, Davies–Bouldin, and Calinski–Harabasz evaluation metrics.

The final clustering model utilized the X2 dataframe, with principal component analysis (PCA) reducing dimensionality to six components that captured 90% of the variance, and K-Means clustering with $k = 3$. This approach identified three distinct clusters: (1) mainstream games with high engagement and visibility, (2) heavily discounted games exhibiting moderate engagement, and (3) lower-visibility titles characterized by low engagement and limited community activity.

A supplementary experiment employing the X3 dataset, which included all paid games and ratio or log-based features, demonstrated that incorporating games with extremely low or zero playtime alters the cluster structure and diminishes interpretability, despite higher overall clustering scores. These findings indicate that unsupervised learning serves as an effective exploratory tool for identifying broad behavioral segments in large digital game marketplaces, particularly when feature engineering mitigates the influence of extreme engagement values.

REFERENCES

[1] DemandSage, "Steam statistics (2026): Market share, mau and revenue," 2026. [Online]. Available: https://www.demandsage.com/steam-statistics/

[2] S. F. Aulia, Y. A. Gerhana, and E. Nurlatifah, "Game and application purchasing patterns on steam using k-means algorithm," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 3, pp. 117–124, 2021.

[3] Y.-K. Lim, "Some crude, semi-quantitative analyses on game length, price and quality," 2015, theoretical Sandbox Blog. [Online]. Available: https://theoreticalsandbox.wordpress.com/2015/01/23/some-crude-semi-quantitative-analyses-on-game-length-price-and-quality/

[4] K. Chan, "Predicting a steam game's success using supervised learning," 2025, medium.

[5] 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.

[6] A. Drachen, C. Bauckhage, and R. Sifa, "Introducing clustering iii: Challenges and pitfalls," 2025, gameAnalytics Blog. [Online]. Available: https://www.gameanalytics.com/blog/introducing-clustering-iii-challenges-pitfalls

[7] X. Liang and S. Yang, "Pc game play time estimation based on steam data and reviews," Stanford University, Tech. Rep., 2017.