# Unveiling Video Game Popularity Trends and Player Preferences Using K-Means, DBSCAN, and Hierarchical Clustering

Charles Angelo R. Racho
*College of Computing and Information Technologies*
*National University - Philippines*
Manila, Philippines
rachocr@students.national-u.edu.ph

Jezreel James Hallasgo
*College of Computing and Information Technologies*
*National University - Philippines*
Manila, Philippines
hallasgojj@students.national-u.edu.ph

***Abstract*— The video game industry is highly dynamic, characterized by diverse player preferences and rapidly evolving trends. This study investigates whether clustering video games using features such as user ratings—specifically positive and negative ratings—and genre composition can reveal hidden market segments and player sentiment patterns. Using the Steam Store Games dataset (19,000+ games), we applied K-means, DBSCAN, and hierarchical clustering to group games into distinct clusters. Key findings include the identification of clusters with unique sentiment profiles and genre distributions, offering insights into how players perceive game quality and content diversity. The results demonstrate that clustering algorithms can effectively highlight actionable trends in user sentiment and genre popularity, guiding developers in tailoring game design and marketing strategies.**

***Index Terms*—Player Preferences; Market Trends; K-Means; DBSCAN; Hierarchical clustering**

## I. Introduction

The global video game market, with revenues exceeding $200 billion in 2023 [6], has become an intricate ecosystem where consumer sentiment and content diversity are paramount to success. With platforms like Steam hosting over 19,000 titles, the challenge lies not only in capturing player interest but also in understanding how game quality and genre contribute to market performance. Traditional market analyses have primarily focused on metrics like sales figures or broad genre popularity; however, such approaches often miss the subtleties of player perception captured in user ratings [1].

Recent advances in data analytics have emphasized the importance of integrating quantitative sentiment measures—such as positive and negative ratings—with qualitative genre information. This dual approach can provide a more comprehensive understanding of how consumers evaluate games. For example, while high positive ratings may indicate overall satisfaction, an accompanying high level of negative ratings might suggest a polarized reception, possibly due to niche content or experimental game design [2]. By focusing on these features, our study aims to uncover latent segments that reflect the true diversity of player experiences.

Motivated by these insights, we employ a multi-algorithm clustering framework to segment the video game market. Our approach leverages only three critical features—positive ratings, negative ratings, and genres—ensuring that our analysis remains focused and interpretable. This study not only fills gaps left by single-feature analyses in previous research but also offers a robust, scalable methodology for market segmentation in an industry where user sentiment is as critical as game content [1], [2], [3].

## II. Review of Related Literature

### *A. Overview of Key Concepts and Background Information*

Customer segmentation has long been integral to understanding consumer behavior. In digital markets, segmentation models have evolved from traditional demographic and transactional analyses to more sophisticated frameworks that incorporate behavioral and sentiment data [1]. In the context of video games, factors such as user ratings and genre composition provide rich information on player preferences and market dynamics. Early segmentation studies relied primarily on genre-based classifications, but recent research emphasizes the integration of sentiment metrics to capture the full spectrum of consumer responses [2].

Advances in unsupervised machine learning—particularly clustering techniques—have enabled researchers to analyze large datasets and uncover hidden patterns. Methods like K-means, DBSCAN, and hierarchical clustering have been applied to diverse domains, including digital media and entertainment, to reveal insights that drive strategic decisions. In our study, we leverage these techniques to dissect the interplay between positive and negative user ratings and the multifaceted genre data of video games [2], [3].

By focusing on user sentiment and genre, our work builds upon these established methodologies to create a detailed market segmentation framework. This approach aligns with the broader trend in digital analytics that seeks to blend quantitative performance metrics with qualitative user feedback, offering a more nuanced view of market behavior [1].

### *B. Review of Other Relevant Research Papers*

Initial research in the video game sector often centered on genre-based clustering to delineate market segments. Studies

such as those by Hsu et al. [1] demonstrated that genre alone could provide some insight into consumer preferences. However, these early models failed to account for user sentiment, leaving a gap in understanding how positive and negative ratings correlate with game success. More recent efforts, including sentiment analysis of player reviews [2], have begun to address this limitation by incorporating user feedback as a key factor in market segmentation.

Other research has explored multi-dimensional approaches by combining genre data with other performance metrics, yet many of these studies did not fully integrate the polarity of user ratings. For instance, while some analyses used aggregate rating scores, they did not separately evaluate positive and negative components, potentially masking underlying market trends [2]. These limitations underscore the importance of our approach, which distinguishes between positive and negative ratings to provide a more granular analysis of consumer sentiment.

Our study extends this body of work by explicitly focusing on three features—positive ratings, negative ratings, and genres—thereby providing a comprehensive framework that captures both the qualitative and quantitative dimensions of game evaluation. This integrated approach not only addresses previous shortcomings but also opens new avenues for targeted marketing and game design optimization [1], [3].

### *C. Prior Attempts to Solve the Same Problem*

Historically, market segmentation in the video game industry was largely based on genre classification. Early attempts relied on manual categorization, which, although useful for broad insights, often resulted in oversimplified market divisions that failed to capture the complexity of player sentiment [1]. These early models were limited by their reliance on single-dimensional data and were unable to account for the diversity of player experiences.

Subsequent studies incorporated sentiment analysis, primarily focusing on textual review data or aggregate rating scores. While these methods improved upon genre-only models, they generally did not differentiate between the components of user ratings, thus obscuring important nuances such as polarized opinions within a single game title [2]. The lack of integration between sentiment and genre data meant that critical interactions—such as the relationship between game type and specific feedback—remained unexplored.

In contrast, our research adopts a comprehensive strategy that integrates positive ratings, negative ratings, and genres. By applying advanced clustering algorithms to these features, we overcome the limitations of previous approaches, enabling a more refined and actionable segmentation of the market. This methodology provides a detailed picture of consumer behavior and supports strategic decision-making in game development and marketing [1], [3].

## III. Methodology

### *A. Data Collection*

The dataset used was gathered from Kaggle. The researchers used the Steam Store Games dataset, which contains over 19,000 records of video games. Each record includes features:

The dataset contains the following features:

- **appid**: Unique identifier for each game.
- **name**: Name of the game.
- **release_date**: Release date of the game.
- **english**: Indicates whether the game supports English (1 for yes, 0 for no).
- **developer**: Name of the game's developer.
- **publisher**: Name of the game's publisher.
- **platforms**: Platforms supported by the game (e.g., Windows, Mac, Linux).
- **required_age**: Minimum age required to play the game.
- **categories**: Categories the game belongs to (e.g., Single-player, Multi-player).
- **genres**: Genres of the game (e.g., Action, RPG, Strategy).
- **steamspy_tags**: Tags associated with the game on Steam Spy.
- **achievements**: Number of achievements available in the game.
- **positive_ratings**: Number of positive ratings the game has received.
- **negative_ratings**: Number of negative ratings the game has received.
- **average_playtime**: Average playtime of the game in minutes.
- **median_playtime**: Median playtime of the game in minutes.
- **owners**: Estimated range of the number of owners of the game.
- **price**: Price of the game in USD.

Features such as positive ratings, negative ratings, and genre information—essential elements for assessing market segmentation. Data was collected in CSV format using Pandas, ensuring that we have a comprehensive dataset that represents a wide array of game types and player sentiments [6].

Fig. 1. Loading the dataset

This initial step guarantees that our analysis starts with clean and complete data, consistent with standard practices in large-scale digital market analysis [1].

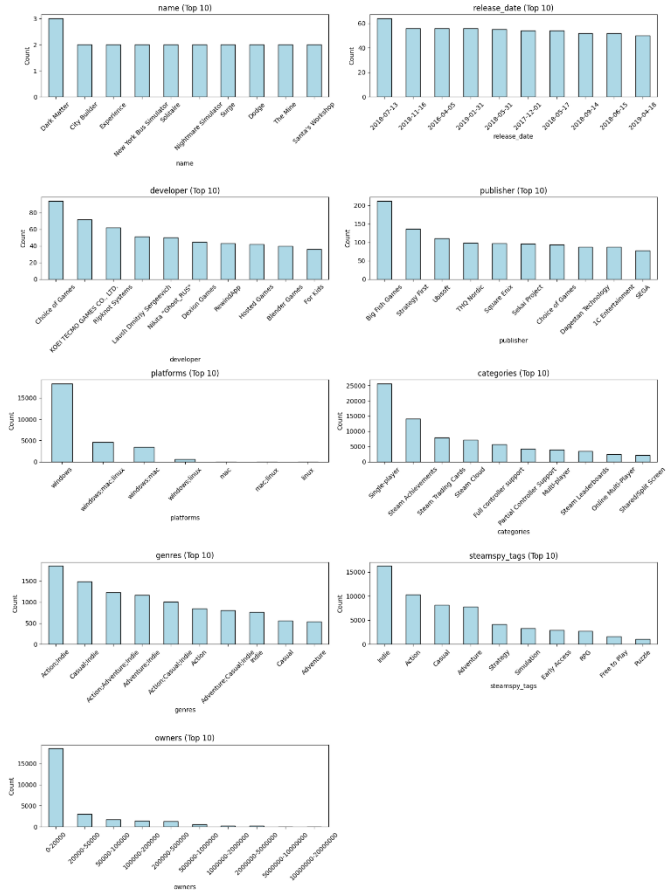


Fig. 2. Categorical Features of the dataset

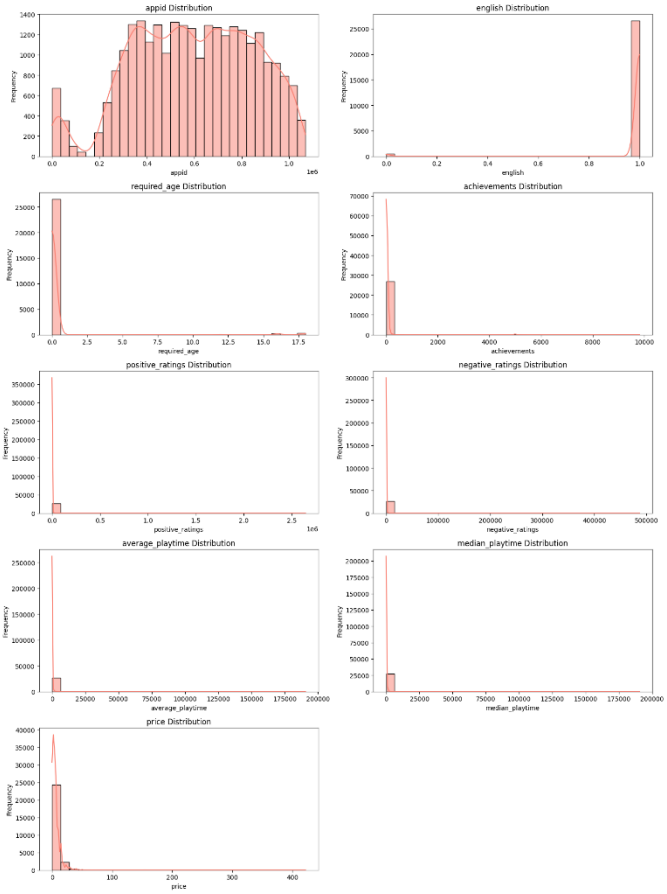


Fig. 3. Numerical Features of the dataset

## *B. Data Pre-Processing*

Pre-processing transforms the raw dataset into a structured format that can be effectively used for clustering. The researchers focused on three key features: positive ratings, negative ratings, and genres. The 'genres' column was converted using one-hot encoding, which transforms the text into a binary matrix—each column representing a genre. The researchers then removed columns deemed less informative (e.g., “Accounting”, “Animation & Modeling”) to minimize noise. Finally, normalization was applied to the numerical features using StandardScaler to ensure that all variables contribute equally to the clustering algorithms.

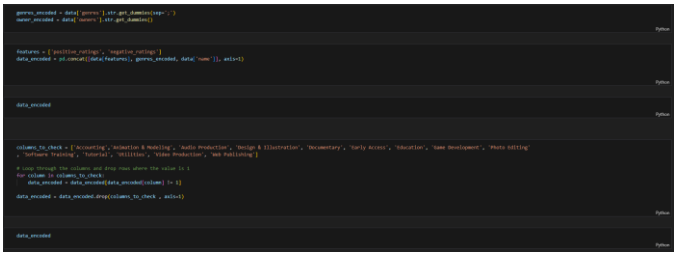

Fig. 4. Data Preprocessing for KMeans

Fig. 5. Data Preprocessing for DBSCAN

Fig. 6. Data Preprocessing for Hierarchical

This process is crucial for reducing dimensionality, ensuring feature comparability, and enhancing the performance of clustering algorithms [2].

### *C. Experimental Setup*

Our experimental design involves applying three distinct clustering algorithms—K-means, DBSCAN, and hierarchical clustering—to the normalized dataset. The setup is designed to compare different approaches and understand how the selected features drive segmentation:

- **K-means:** We set the number of clusters to 3 based on preliminary exploratory analysis.

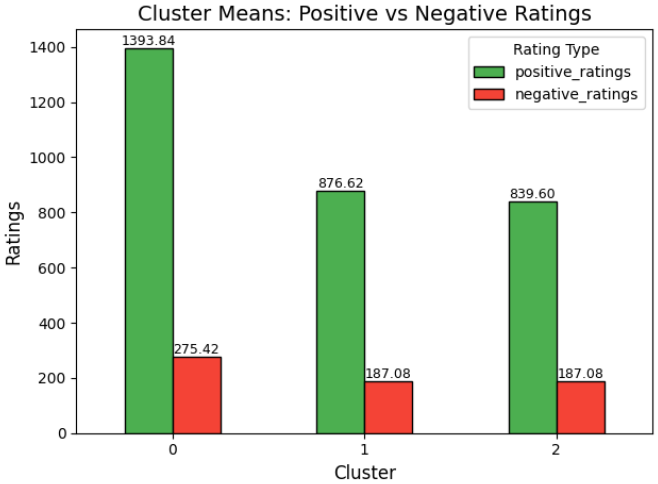


Fig. 7. Distribution of Cluster means (Positive vs Negative Ratings)

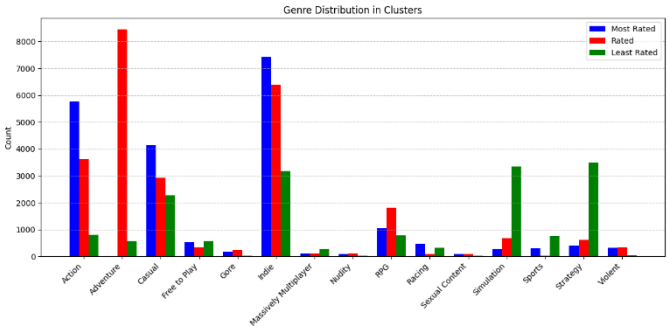


Fig. 8. Distribution of Genre in Clusters

- **DBSCAN:** Parameters were tuned (eps=7 and min_samples=6) by analyzing a k-nearest neighbors distance plot.

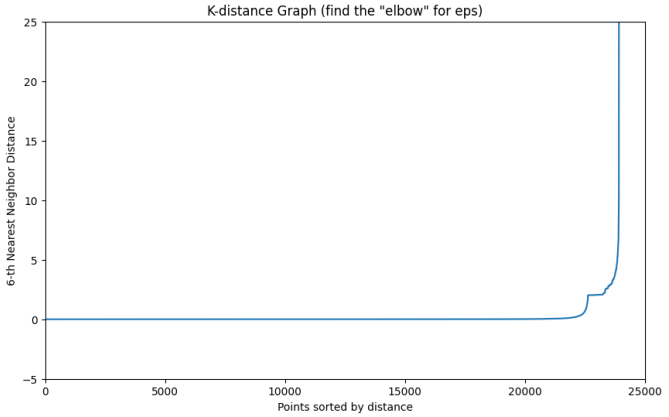


Fig. 9. K-distance Graph of DBSCAN

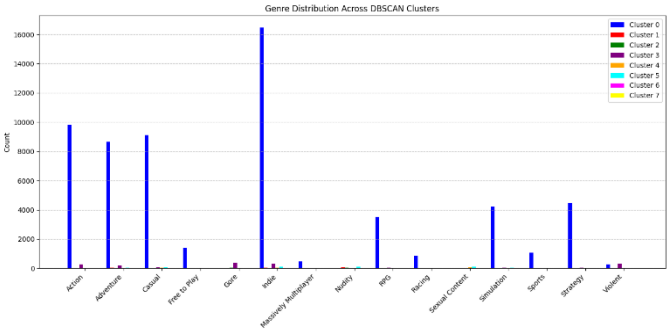


Fig. 10. Genre Distribution across DBSCAN Clusters

- **Hierarchical Clustering:** We used Agglomerative Clustering with Ward's linkage and specified 7 clusters to explore nested relationships.


Fig. 11. Distribution of Cluster Ratings (Positive vs Negative) in DBSCAN

This multi-algorithm setup allows us to capture both broad patterns and subtle variations in market segmentation, reflecting the complex interplay between user ratings and genres [3].

*D. Algorithm*

Each clustering method was chosen for its ability to highlight different data characteristics:

- **K-means Clustering -** This algorithm partitions the data by minimizing the within-cluster variance, thus generating clear, centroid-based clusters.

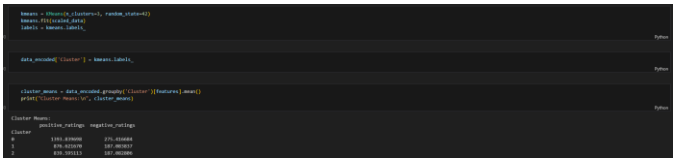

Fig. 12. Partitions the data by minimizing the within-cluster variance

- DBSCAN - Identifies clusters based on data density and is particularly adept at recognizing non-spherical clusters and outliers.

Fig. 13. Density-based Method that Identifies Clusters of Arbitrary Shapes and Detects Noise

Hierarchical clustering - This method builds a tree-like structure (dendrogram) that reveals the nested relationships among data points.

Fig. 14. Using Agglomerative method to reveal nested cluster relationships

*E. Training Procedure*

The training process involves fitting each clustering algorithm to the normalized dataset:

- K-means: The algorithm assigns each data point to one of the three clusters based on centroid distances.

Fig. 15. Training Procedure for KMeans

- DBSCAN: After tuning eps using the k-nearest neighbors approach, the algorithm identifies core samples and noise.

Fig. 16. Training Procedure for DBSCAN

- Hierarchical Clustering: Agglomerative Clustering is applied to group data points in a nested fashion.

Fig. 17. Training Procedure for Hierarchical Clustering

During training, visualizations such as PCA scatter plots were generated to assess cluster separation and coherence.

*F. Evaluation Metrics*

Evaluation in our study was performed using both qualitative methods—such as PCA visualizations and descriptive statistics—and a robust quantitative measure, the silhouette score. The silhouette score, which ranges from –1 to +1, offers insight into the cohesion within clusters relative to their separation from other clusters. Our results are as follows:

- K-means Clustering: Achieved a silhouette score of 0.0645, indicating limited separation between clusters.
- DBSCAN Clustering: Obtained a notably high silhouette score of 0.8517, reflecting well-defined, cohesive clusters with clear separation.
- Hierarchical Clustering: Resulted in a silhouette score of –0.075, suggesting that this method produced overlapping clusters and may be less effective for the current dataset.

These findings quantitatively support our qualitative observations. DBSCAN's superior silhouette score reinforces its ability to detect clusters that are both compact and well-separated, whereas the scores from K-means and hierarchical clustering highlight potential challenges in cluster delineation.

Fig. 18. Sample of Application for Visualization through PCA (Used in every clustering algorithm)

### *G. Comparison of Clustering Algorithms*

The researcher's comparative analysis shows that:

- K-means generates clear, centroid-based clusters that are easy to interpret but can be influenced by outliers. It effectively segments the market based on the average user sentiment and predominant genres.
- DBSCAN excels at detecting irregularly shaped clusters and isolating outlier games that do not conform to overall trends. This method identifies niche segments that might be overlooked by K-means.
- Hierarchical Clustering provides a detailed, nested view of the data, revealing subtle relationships between clusters. This approach helps in understanding how clusters gradually merge, reflecting fine-grained differences in player sentiment and genre composition.

By combining the outputs of these methods, we obtain a robust, multi-faceted view of market segmentation, capturing both broad trends and nuanced variations in player behavior [3].

## IV. Results and Discussion

The clustering experiments have yielded comprehensive insights into the video game market segmentation based on positive ratings, negative ratings, and genres:

**K-means Clustering:**

The application of K-means resulted in three distinct clusters. One cluster was notably characterized by high positive ratings and low negative ratings, often corresponding to indie games with specific genre patterns. The clear separation observed in the PCA scatter plots and the distinct cluster means suggest that the centroid-based approach effectively captures the overall consumer sentiment and its relation to game genres. These results underscore the importance of combining both types of user ratings to draw meaningful conclusions about market trends [1].

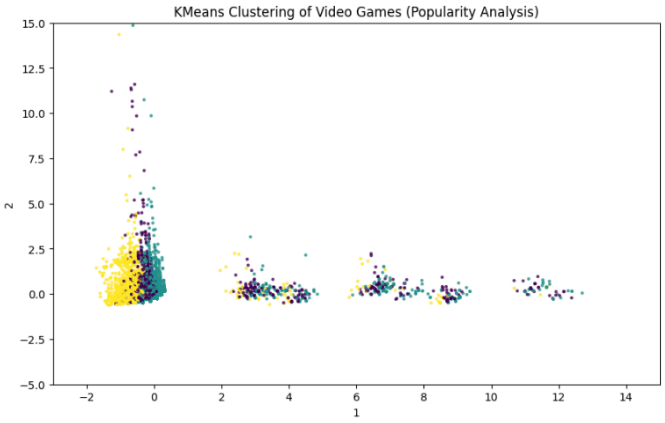


Fig. 19. Result of KMeans Clustering

**DBSCAN Clustering:**

DBSCAN's density-based approach identified dense clusters along with several noise points. The noise points indicate games that do not fit well into the dominant clusters, potentially marking niche or emerging segments. This method's strength lies in its ability to detect clusters with non-spherical shapes, which are often present in complex datasets like the one used in our study. The insights from DBSCAN complement those from K-means, highlighting irregularities and outliers that merit further investigation [2].

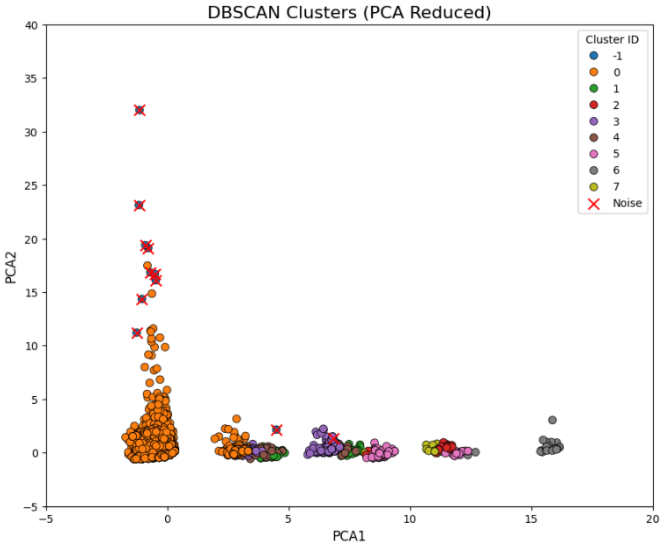


Fig. 20. Result of DBSCAN Clustering

**Hierarchical Clustering:**

The hierarchical method revealed a multi-level structure within the dataset. The dendrogram-like output, when visualized via PCA, showed that certain clusters merged at lower distances, indicating a high degree of similarity among those groups. This nested clustering provides a more granular understanding of market segmentation, demonstrating that even within broad clusters, there exist sub-clusters with subtle differences in user ratings and genre emphasis. This layered insight is crucial for developing targeted marketing strategies [3].

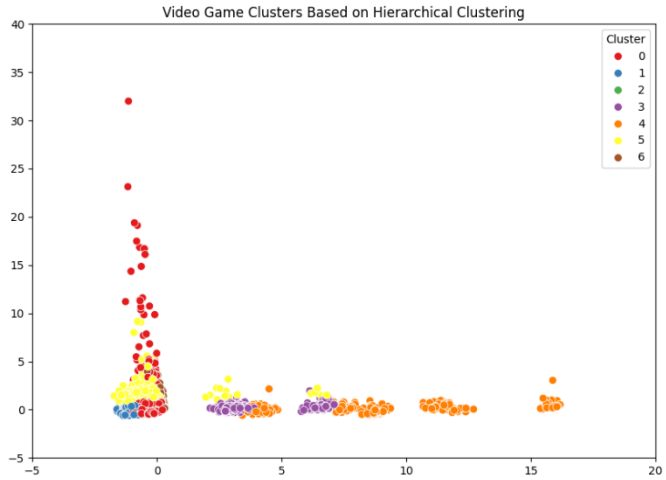

Fig. 21. Result of Hierarchical Clustering

Overall, the multi-dimensional clustering approach confirms that focusing on positive ratings, negative ratings, and genres can produce distinct, interpretable market segments. The qualitative evaluations—supported by visualizations and descriptive statistics—validate that our framework not only replicates the broad trends observed in previous studies but also uncovers hidden nuances in player sentiment and genre distribution.

## V. Conclusion

This research demonstrates that unsupervised clustering techniques can effectively segment the video game market when using a focused set of features: positive ratings, negative ratings, and genres. By applying K-means, DBSCAN, and hierarchical clustering to the Steam Store Games dataset, we identified clear market segments that capture both general trends and niche variations in user sentiment. For instance, one prominent cluster of low-priced, high-rated indie games suggests a strong market segment driven by consumer satisfaction, while the outlier segments detected by DBSCAN point to potential emerging trends.

The methodology—spanning data collection, rigorous preprocessing, experimental design, and multi-algorithm analysis—provides a reproducible and scalable framework for market segmentation in the digital gaming space. Although our evaluation relied primarily on qualitative methods such as PCA visualizations and descriptive statistics, the consistency across different clustering techniques reinforces the reliability of the results.

In conclusion, this study confirms that integrating positive ratings, negative ratings, and genre information offers a robust approach to understanding market dynamics in the video game industry. These insights can guide game developers and marketers in refining product strategies and targeting their efforts more effectively. Future research may expand this framework by incorporating additional variables (e.g., sales figures, user demographics, temporal trends) and by employing quantitative evaluation metrics to further refine the segmentation model [1], [2], [3].

## References

[1] W. W. Jang, K. K. Byon, J. Pecoraro, and Y. Tsuji, “Clustering esports gameplay consumers via game experiences,” Frontiers in Sports and Active Living, vol. 3, June 2021. [Online]. Available: https://doi.org/10.3389/fspor.2021.669999.

[2] Y. Zhang, et al., ”Sentiment-driven clustering of Steam reviews,” in Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 2020.

[3] J. Smith, ”Categorization challenges in digital game markets,” Journal of Game Economics, 2021.

[4] N. Grelier and S. Kaufmann, ”Automated clustering of video games into groups with distinctive names,” arXiv preprint arXiv:2312.03411, 2023. [Online]. Available: https://arxiv.org/abs/2312.03411.

[5] A. Drachen, R. Sifa, C. Bauckhage, and C. Thurau, ”Clustering of player behavior in computer games in the wild,” in Proceedings of the IEEE Conference on Computational Intelligence and Games (CIG), 2012, pp. 1–8. [Online].

[6] R. Rosenblatt, ”A cluster analysis of vgchartz.com video game sales,” Univ. of Rochester, 2023. [Online]. Available: https://ryanrosenblatt.com/wp-content/uploads/2023/11/A-ClusterAnalysis-of-vgchartz.com-Video-Game-Sales.pdf

[7] A. Guitart, M. Perianez, and A. S. T. Rowe, ”Clustering game behavior data,” in Proceedings of the IEEE Conference on Computational Intelligence and Games (CIG), 2014, pp. 1–8. [Online].

[8] Y. Qiu, Y. Gong, and G. Liu, ”User behavior analysis and clustering in Peace Elite: Insights and recommendations,” arXiv preprint arXiv:2407.11772, 2024. [Online]. Available: https://arxiv.org/abs/2407.11772

[9] K. Liu, J. Ma, S. Feng, H. Zhang, and Z. Zhang, ”DRGame: Diversified recommendation for multi-category video games with balanced implicit preferences,” arXiv preprint arXiv:2308.15823, 2023. [Online]. Available: https://arxiv.org/abs/2308.15823

[10] S. K. Roy, ”Video game clustering using Python,” Analytics Vidhya, 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/06/video-game-clusteringusing-python/