PROJECT TITLE

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*—**
***Index Terms*—**

I. INTRODUCTION

Over the past decade, the gaming industry has grown rapidly, with more than 180 billion dollars generated globally [1]. And thousands of new games are being launched every year across multiple platforms, this makes the assessment of video games quality very complex and multidimensional [2]. The traditional method used to assess the quality of games has always been the use of aggregated review scores from both players and critics, but these scores often do not accurately represent the significant relationship between favourable reviews and commercial success [3].

This study questions whether critically acclaimed games always sell well, or do distinct market segments exist where quality sales diverge? Previous studies have shown that the relationship between reviews and sales is complex and varies significantly across entertainment industries [4], [5]. Our approach addresses the challenge of finding meaningful game archetypes automatically by analyzing the relationship between commercial performance (measured by global sales) and critical reception (measured by critic and user scores). Rather than using supervised methods that need pre-labeled data, we use an unsupervised approach that finds hidden patterns without needing predefined categories which works very well for new markets or game types where we don't have existing labeled datasets.

The problem is significant as it offers insights of how quality impacts success which benefits consumers who are looking for high quality games, developers who want to enhance their design strategies and identify target market segments to optimize resource allocation, finally, publishers are able to use the information gathered from this study to assess market trends. The proposed framework has implications for use in all aspects of market analysis, recommendation systems, and making strategic decisions related to the gaming industry.

II. LITERATURE REVIEW

This section provides the theoretical and historical context for the dataset and the unsupervised learning models for quality assessment, clustering algorithms, and video game analytics.

A. Overview of Key Concepts and Background Information

Unsupervised learning involves machine learning techniques that discover hidden patterns in unlabeled data [6]. While supervised learning requires labeled datasets, the algorithms applied under unsupervised learning uncover their intrinsic patterns and relationships, which, as mentioned, is extremely useful if you are interested in exploring the data, especially because it is too expensive or subjective to label everything manually [7].

An example of an unsupervised learning algorithm is Clustering, which separates data points according to how similar each one is to other data points and keeps those entries from being clumped together with others that are dissimilar. There are four different and distinct methods for clustering data points, and they all work differently:

1) K-Means Clustering: A partitioning approach proposed by MacQueen in 1967 [8], the K-Means algorithm works iteratively to minimize the WCSS and thus begins with random initialization of points and update using assign the closest cluster centroid to each point and re-calculate centroids. Despite the issues associated with non-spherical clusters, due to its computational efficiency O(n·k·d·i) and simplicity, it has been a popular clustering method. The convergence of k-means that were running was further improved by efficient seeding of initial centroids using the approach proposed by Arthur and Vassilvitskii called k-means++ [9].

2) Hierarchical Clustering: Agglomerative hierarchical approaches [10], [11], construct the tree structures (dendrograms) by repetitively merging together similar clusters. Various linkage methods Ward (minimizing variance), Complete (maximal distance), and Average (mean distance) will result in different cluster shapes. Because of this, hierarchical algorithms do not require prior knowledge about the number of clusters and also offer insights into nested cluster structures which is not available with K-Means.

3) DBSCAN: Density-Based Spatial Clustering of Applications with Noise scouts clusters according to the concept of local density rather than distance. Dense regions of points clump together while isolated points are considered noise/outliers. DBSCAN handles arbitrary cluster shapes and automatically determines cluster count, though it struggles with varying densities and requires careful parameter tuning (epsilon, minimum points) [12], [13].

4) *Gaussian Mixture Models (GMM):*

Gaussian Mixture Models (GMM) are probabilistic models that assume data originates from a mixture of multiple Gaussian distributions. The EM or Expectation-Maximization algorithm estimates the parameters of the GMM iteratively and each iteration assigns data points to the different Gaussians of the model with a probability rather than a hard boundary. GMMs can model complex cluster shapes by using different types of covariance structure such as full diagonal, and spherical [14].

5) *PCA:*

B. Review of Other Relevant Research Papers

1) *RRL:*

2) *RRL:*

3) *RRL:*

C. Prior Attempts to Solve the Same Problem

1) *:*

2) *:*

3) *:*

III. Methodology

A. Overview

IV. Discussion

A. Limitation and Future Work

References

[1] E. Newzoo, "Global games market report," *Amsterdam: gamesindustry. com*, 2017.

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

[3] 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

[4] 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.

[5] G. Cui, H.-K. Lui, and X. Guo, "The effect of online consumer reviews on new product sales," *International Journal of Electronic Commerce*, vol. 17, no. 1, pp. 39–58, 2012.

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

[7] 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.

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

[9] 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.

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

[11] F. Murtagh and P. Contreras, "University of huddersfield repository murtagh, fionn and contreras, pedro," 2017. [Online]. Available: https://eprints.hud.ac.uk/id/eprint/32552/1/DWD_Fmurtagh_31.pdf

[12] 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.

[13] J. SANDER, M. ESTER, H.-P. KRIEGEL, and X. XU, "Density-based clustering in spatial databases: The algorithm gdbscan and its applications," *Data Mining and Knowledge Discovery*, vol. 2, pp. 169–194, 1998.

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