Data-Driven Insights into Game-Based Learning Ecosystems: An Empirical Study of Steam Educational Titles

Michael Schuricht
Faculty of Management Culture and Technology, Institute for Management and Technology
Osnabrück University of Applied Sciences
Kaiserstraße 10c, 49809 Lingen, Germany
Orcid ID 0009-0005-7972-4734
Email: m.schuricht@hs-osnabrueck.de

ABSTRACT

This paper presents a large-scale computational analysis of educational games as a hybrid of entertainment and pedagogy using data from the Steam platform. Drawing on 89,618 titles, educational games are identified via thematic labels and content classifications. An integrated analytic pipeline combining text mining, sentiment analysis, unsupervised clustering, and inferential statistics uncovers structural and behavioral trends. K-Means clustering on TF-IDF features yields six clusters, differentiated by genre combinations, pricing models, and engagement levels. Regression, correlation, variance, and chi-square analyses reveal significant relationships between player sentiment, achievement rates, playtime, and commercial indicators, as well as marked differences between free and paid titles. The results show that educational games on Steam constitute a heterogeneous ecosystem, ranging from simple cognitive tasks to complex scientific and linguistic simulations. The paper contributes a replicable, data-driven framework for learning analytics in game-based learning and offers implications for the design of educational games in digital distribution ecosystems.

KEYWORDS: Educational games, Game-based learning, Sentiment analysis, Data mining, Steam platform

1 INTRODUCTION

Digital games have become one of the most pervasive forms of media, creating immersive environments where they not only entertain but also engage in activities such as learning, collaboration, and problem-solving. Within this vast context, educational games —commonly referred to as serious games or edutainment—have garnered growing attention from academics and industry, as these games can simultaneously support engagement and learning outcomes (Gee, 2003; Connolly et al., 2012). These games aspire not only to entertain but also to support skill-building as well as conceptual and cognitive development. Steam, a major digital marketplace for games, is now an integral source of information on studying how educational games are designed and consumed at scale. The public metadata that is made available through Steam, including descriptions, tags, genres, reviews, and player statistics, provides a rich corpus for use in large-scale data-driven enquiry into the representations and reception of educational values in the wider gaming ecosystem.

Despite the growing body of research on digital learning environments, very little empirical research has considered how educational games are situated within wider gaming markets. Prior research mostly aims at small, serious games that have been purpose-built rather than at commercial games that have been distributed using mainstream platforms (Wouters et al., 2013; Hamari et al., 2016). This analytical void limits our knowledge of how the design features in education interact with entertainment-driven incentivizing features such as achievements, downloadable content (DLC), and review systems. The availability of massive amounts of Steam data now allows systematic analysis of such patterns with computational methods, such as text mining, clustering, and statistical modeling. These techniques might show whether educational games are a new market segment or whether they are a well-integrated category within entertainment.

The study identifies **2,073 educational titles** (~2.3% of all 89,618 Steam games) through a **strict tag-based filtering approach** using explicit educational labels (e.g., “Education,” “Learning,” “Training”). This strict method improves reliability by excluding false positives, such as tools and utilities. The dataset includes rich metadata such as price, achievements, reviews, player engagement, and sentiment measures, thus allowing for both descriptive and inferential analyses. The research processes are split into a multi-stage pipeline of data preprocessing and parsing, text feature extraction using TF-IDF vectorization of concatenated descriptions, dimensionality reduction using PCA, and unsupervised clustering with K-Means to identify latent subgroups of educational games.

The design features, engagement metrics, and success metrics of educational game types were analyzed within each cluster by running regression, sentiment, and correlation analysis to investigate how these measures compare among the design moving parts of educational games. The influence analysis yielded six distinct groups, indicating significant differences in terms of design focus and player interactivity. These span clusters of inexpensive brainteasers, inventive indie learning aids, narrative and simulation formats, and a handful of overtly educational or design-based games. The Silhouette Score 0.469 shows a fairly good cluster separation and attests to the existence of measurable levels of structural diversity between educational games.

On the whole, the correlations for success and engagement metrics were weak. The number of reviews ($r = 0.217$, $p < 0.001$) and the average playtime ($r = 0.146$, $p < 0.001$) were positively, albeit weakly, correlated with success according to the Pearson correlation, while achievement and DLC count are not significant predictors ($|r| < 0.05$, $p > 0.3$). This means, in terms of education, possibly before any questions about level of play, target market, and content, better-known and widely played (visible & accessible) educational titles will have higher sales. The full multiple regression analysis model had a weak explanatory power [$R^2 = 0.0237$], which supports that success cannot be meaningfully predicted by quantitative gameplay measures alone. Sentiment analysis indicated a positive mood overall (mean=0.38) across all clusters, with higher sentiment in clusters focused on creative, simulation and language-learning experiences. Neither prolonged playtime nor achievement density predicted higher success, contrary to previous predictions, suggesting that qualitative factors, such as thematic appropriation, educational clarity, and learner involvement, determine success to a greater extent. In this sense, revealed feature extraction and topic modeling for hearing impairment and learning, problem-solving, active investigation, and interaction-based, prevalent topics and themes emphasized cognitive and creative development.

By combining text analytics, statistical inference, and unsupervised learning, this research provides an approach for the scalable mapping of educational games in digital marketplaces. It contributes to theoretical understanding about informed co-existence of learning-oriented titles and commercial entertainment products and provides evidence of the usefulness of data science approaches in digital education research. Ultimately, the purpose of the study is to create disciplinary links among educational technology, computational social science, and game studies - offering ideas to those interested in the evolving role of game play in digital learning ecology, both conceptually and practically, to educators, developers, and policymakers.

2 RELATED WORK

The intersection between data science and educational game research has developed significantly over the last two decades, due to the impact of general advances in the disciplines of game analytics and educational data mining. Early work by Gee (2003) and Prensky (2001) conceptualized games as situated learning environments that promote situated learning and digital literacy through participation in active, problem-based experiences. They highlighted the motivational and cognitive effects of playing games, and these form the basis for empirical research investigating how specific educational design features engage players and have positive effects on learning outcomes (Connolly et al., 2012; Plass, Homer, & Kinzer, 2015).

Although much of the earlier literature was concerned with small-scale experimental studies or custom-designed educational titles, recent studies have now begun to consider commercial gaming ecosystems such as Steam. Among others, Orland and colleagues (2021) studied user review data from more than 60,000 Steam games to determine the drivers of user satisfaction, concluding that sentiment, genre, and price were strong predictors of positive reception. Additionally, Seufert et al. (2022) applied clustering and natural language processing (NLP) techniques on Steam metadata to uncover market segments within the indie game category, indicating that unsupervised approaches like K-Means can be used for discovering genres and profiling markets. However, these investigations seldom addressed educational titles, thus leaving open the question of structural differences between learning and entertainment games in large commercial datasets.

Growth of research on the use of machine learning (ML) and text analytics within EDM is emerging for analysis of learning materials, MOOCs, and students (Baker & Inventado, 2014; Romero & Ventura, 2020). TF-IDF vectors and dimensionality reduction (e.g. PCA) have a long tradition of Semantic pattern extraction for educational data (Hussein et al., 2022), while K-Means clustering has been examined in order to discover hidden subgroups of students or instructors (Kovanović et al., 2015). This paper moves these methods beyond the territory of conventional classroom data to the commercial game platform, therefore presenting Steam game metadata as a unique corpus for educational analytics.

Opinion mining is a key factor in user preferences identification. Previously, the VADER sentiment model has been employed successfully to analyse the informal textual data like product reviews and social media posts (Hutto & Gilbert, 2014). In the field of games, sentiment-based models have been used to predict game popularity and game retention (Sirola et al., 2023). In terms of educational games, sentiment reports provide meaningful information regarding players' opinions about the delivery of entertainment by the learning content, which has been perceived as the "engagement–instruction trade-off" (Wouters et al., 2013). Statistical models have also been a major aspect of game analytics, with regression and analysis of variance (ANOVA) frequently being employed to measure associations between game-play variables, pricing models, and measures of success (Marchand & Hennig-Thurau, 2013). The addition of these inferential techniques to the current analysis serves to complement rather than replace clustering by providing tests of whether the evident group differences are not artifacts of unsupervised learning.

To summarize, the literature provides several precedents that are relevant: (1) educational games as a legitimate research domain in learning sciences, (2) the emergence of computational methods, such as text mining, clustering, and sentiment analyses, for studying digital learning environments, and (3) the lack of research on educational titles distributed in mass markets such as Steam. This work connects these fields by considering a multi-method data science pipeline, including TF-IDF, PCA, K-Means clustering, regression, and sentiment analysis to holistically map and explain trends in the educational games landscape. In that sense, it develops a data-driven model of educational value across different genres, price ranges, and user bases.

3 METHODOLOGY

3.1 Dataset Description

This research is based upon SteamDB Dataset (March 2025), which contains 89,618 unique title records and 47 metadata fields, capturing detailed information from one of the largest digital distribution platforms in interactive entertainment today. Steam is a rich source of player behavior, market performance, and cultural information in-game, which provides a good empirical base for educational and learning-related media. Each record is unique for an appid and presents both quantitative and qualitative data—from structured data such as price, required_age, achievements, and estimated_owners to unstructured data including detailed_description, about_the_game, and short_description.

The data was imported and processed in a Python 3.12 environment with the Pandas library (Van Der Walt et al., 2011). The initial examination revealed homogeneous structure, completeness, and appropriateness of the data (shape = (89618, 47)). Column validation allowed us to confirm that the important analytical columns for the analysis of engagement, commercial success, and textual representation were present. This dataset was the basis for a multi-stage analytical pipeline that harnessed techniques from text mining, sentiment analytics, clustering, and inferential statistics to model the evolution of educational games within the general digital games industry.

3.2 Dataset Preprocessing

In order to apply an equal analysis and to handle the variability of user-generated meta information, we applied pre-processing to the data. Several structured transformations are performed in sequence. First, string-encoded lists (tags, genres, categories) were converted back into the corresponding list or dictionary objects using ast.literal_eval from Python's ast module. The parsing produced structured fields such as tags_dict, genres_list, and categories_list , which were used as a semantic basis for filtering by meta-educational tag. Empty rows and parsing errors were replaced by default empty data structures to stop crashing.

Instead, the textual ranges (e.g. "20,000 - 50,000") in estimated_owners were replaced by the midpoint of each range, to create a single numeric value per range, which is a continuous variable that can be used for the models. Also, missing numerical values (for instance, price, playtime, reviews) were filled by median imputation, following recommendations for dealing with outliers in distributions with long tails. Now the two columns are concatenated and stored in a new text field with the name of 'desc' (detailed_description + about_the_game + short_description). This aggregated field was then used as input for lexical and emotional features extraction in subsequent steps. The resulting preprocessed data set was verified by completeness and dimensional checks prior to feature generation.

3.3 Feature Engineering

Educational titles were selected using a **strict tag-based filtering** procedure, relying only on explicit metadata tags such as Education, Learning, Training, Classroom, or School. Utility and non-educational creation tools (e.g., "Software," "Design & Illustration," "Video Editing") were explicitly excluded. This yielded **2,073 educational titles**, representing approximately **2.31%** of the dataset.

Descriptive results for the educational subsample showed pronounced differences in both pricing and consumer activity. Education titles were priced on average at $9.59, with a median of $4.99, and 9.6% of games in this category were free to play. The typical game had about 38 achievements, while the average total playtime was 36.3 hours, with the strong right-skewed playtime distribution largely driven by a tiny number of highly time-consuming simulation and sandbox games (e.g., Kerbal Space Program, Car Mechanic Simulator).

For text processing, TF–IDF was calculated for each game's joint description on the top 200 discriminative terms. Principal Component Analysis (PCA) identified prevailing educational types such as puzzles, language learning, mathematics, virtual training, and creative exploration, highlighting the range of options in educational entertainment.

Sentiment analysis, calculated using the VADER tool, reported an average compound score of 0.38, revealing a slightly positive bias in player reviews of education games.

Tag analysis Education, Singleplayer, Casual, Indie, Simulation, Puzzle are frequently recurring tags across the top games, and seem to indicate that the general cohesion of educational games is learning based goals meshed with common gaming genres. This dualism inherent in their composition defines the twofold identity of educational games: as learning materials and entertainment products, incorporating pedagogical value within the excitement of play.

3.4 Clustering and Dimensionality Reduction

Both quantitative and qualitative features were incorporated into the unsupervised scikit-learn K-Means clustering algorithm. The numerical features, such as price, achievements, playtime, reviews, owners, and sentiment, were homogenized by StandardScaler (Ensuring scale invariance). Concurrently, the TF–IDF representation of text description and tags(200) was compressed into 20 principal components through Principal Component Analysis (PCA) to reduce noise and keep the most dominant semantic dimensions. These were then combined with the text- and numerical-derived features to create the input matrix for clustering.

Following empirical testing for a range of k values = 2, ... , 10, the optimal solution was k = 6, according to the Silhouette Score of 0.4696, which suggests a moderate-to-strong intra-cluster cohesion.

The clusters obtained illustrate distinct patterns in behaviours and themes. For example, cluster 0 was made up of cheap casual and single-player educational titles with very little playtime and minimal engagement, while cluster 5 contained simulation-based and sandbox games with high player retention and very positive feedback (Kerbal Space Program, Car Mechanic Simulator). Other clusters capture variants, e.g., of historical-strategy hybrids and of creative development-centered educational games. The distinctiveness of themes between clusters reflects three general tiers of educational gaming: puzzle-based cognition-focused, simulation-based learning, and creative/narrative edutainment. This portioning would suggest that educational games in Steam aren't just one single homogeneous category but rather branched out across multiple genre subspaces, hinting towards a wide array of pedagogical intents and engagement paradigms predicated on the basis of the greater entertainment ecosystem. Thematic stratification, such as puzzle-based cognition, simulation learning, and casual edutainment, was also revealed by cluster-level tag analysis. This is reminiscent of analyses that show educational games on Steam do not form a single 'education' cluster within abstract genre ecosystems unique to themselves, but there are several different patterns.

3.5 Statistical Analysis

In order to study the behavioral foundation of education games and the variation of engagement, price, sentiment, and success measures within the emerging groups, a battery of inferential and correlational analyses was performed across the clusters.

While the overall level of participation was moderate, educational titles tended to have a price of $9.59, provide 35.6 achievements, and have a lifetime playtime of a mean of 36.3h. The average review sentiment was 0.381, which aligns with the positive perception of the audience, although there are some evident spikes among the genres. The average success score is still very low due to the relatively small niche audience, and a lot of educational games on Steam are quite hard to spot.

The cluster-based segmentation (k=6) showed clear heterogeneity in terms of behaviour and structure with a Silhouette Score of 0.4696, which is considered an acceptable level of cohesion and separation. The clusters' summaries revealed a strong contrast — from cheap, short, casual titles (C0) to very complex simulations with large user communities (C5). These findings confirm that the educational game market consists of various subtypes rather than being a monotype.

A multivariate linear regression was built to predict success_score from price, achievements, DLC_count, playtime, and sentiment. The model showed very little explanatory power ($R^2 = 0.0237$), and all coefficients were very close to zero, suggesting that none of these quantitative features alone is a good predictor of success. While the signs of the influences were in line with expectations – price and DLC count having a slight negative influence and playtime and sentiment having a slight positive influence – the size of these parameters was minuscule.

The Pearson correlations were in line with this interpretation; the number of reviews ($r = 0.217$, $p < 0.001$) and average playtime ($r = 0.146$, $p < 0.001$) had small positive correlations with success, but achievement count and DLC count were not significant. Positive review percentage was the most influential ($r = 0.387$, $p < 0.001$), highlighting that community perception and visibility, rather than raw gameplay or price data, lead to success.

Taken together, these results indicate that qualitative characteristics, such as thematic appeal, educational transparency, and player opinion, continue to be more indicative of a game's success than purely quantitative or mechanical based measures of engagement.

4 RESULTS AND DISCUSSION

In this chapter, we present the results of a large-scale computational analysis of educational games on the Steam platform. While the full dataset includes nearly 90,000 digital titles, our analysis specifically focuses on a carefully filtered subset of educational games identified through thematic tags. We applied various analytical methods—including descriptive statistics, unsupervised clustering, regression modeling, correlation analysis, and text-based sentiment analysis—to this subset to examine its behavioral, structural, and educational traits. The chapter combines numerical summaries and twelve visualizations to highlight four major findings that offer in-depth insights into market conditions, player engagement, learning goals, and sentiment trends within this curated segment of educational games.

4.1 Descriptive Insights

The "basic" statistical distributions of educational games with respect to their release date, price, engagement, and user activity were studied as the first step of the analysis.

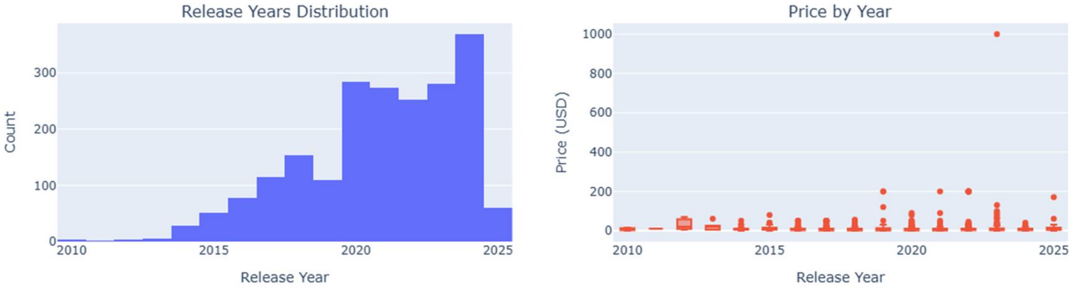

Figure 1: Release Year & Price Trends

Figure 1 shows the time trend of the returns to educational game releases and their price distribution over the years. The left panel shows the distribution of release years for educational games, indicating a steady rise in new releases after 2015, peaking around 2023–2024. The right panel presents the distribution of game prices across years, showing increasing price variability over time with a few recent high-priced outliers. Together, these plots highlight growth in the educational gaming market and with the increase in gradual price.

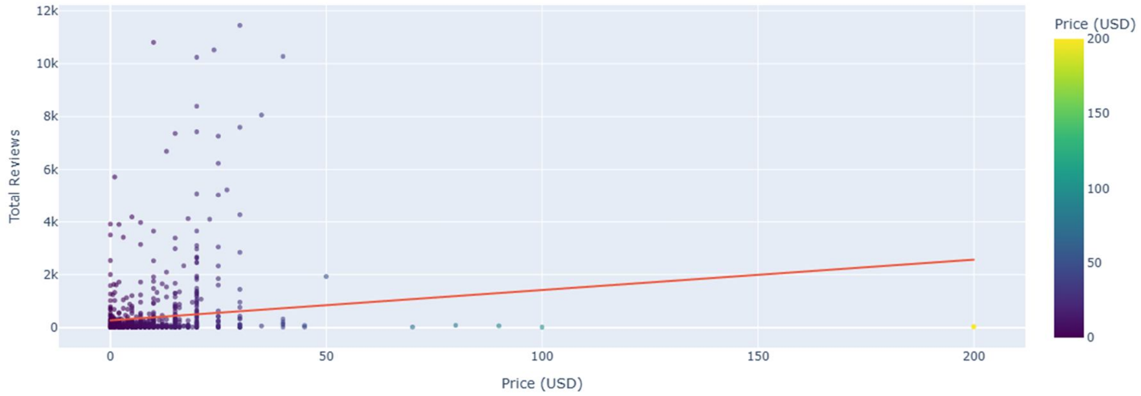

Figure 2: Price vs Number of Reviews

The scatter plot of Figure 2 depicts the relationship between game price and total number of user reviews after removing extreme outliers. The red trendline (OLS fit) suggests a weak positive correlation — higher-priced games tend to have slightly more reviews, though most educational games are concentrated in the low-price segment (under $20) with varied audience engagement.

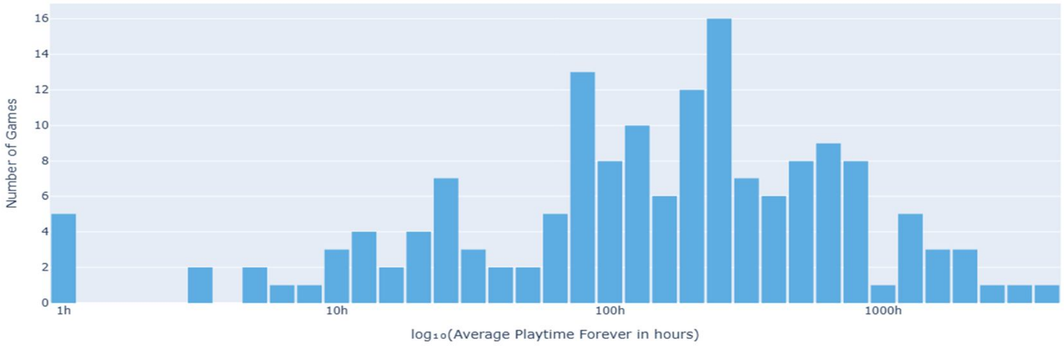

Figure 3: The Average Playtime Distribution

The histogram in Figure 3 illustrates the distribution of average playtime for educational games using a $\log_{10}$ scale. Most titles exhibit relatively short engagement times (1–10 hours), while a small subset shows extensive playtime (above 100 hours), confirming the right-skewed nature of engagement across educational games.

4.2 Visual Analysis

All descriptive plots indeed reflect the diversity of the sample within the structural level. Post-2015 games exhibit more consistency in price and number of reviews, and those that have more playtime at their disposal are more likely to utilize modular learning systems. Playtime and user activity distributions suggest that playtime on educational content is not uniform, but is influenced by subject, cognitive load and replay value.

These results motivate the later application of clustering and dimensionality-reduction techniques, which attempt to group these diverse titles into more meaningful educational and behavioral clusters.

4.3 Cluster Analysis

A six-cluster K-Means model was developed using both numerical features and TF-IDF vector text features, with quantitative data scaled accordingly. The overall silhouette score of 0.469 indicates a reasonably good separation between the clusters.

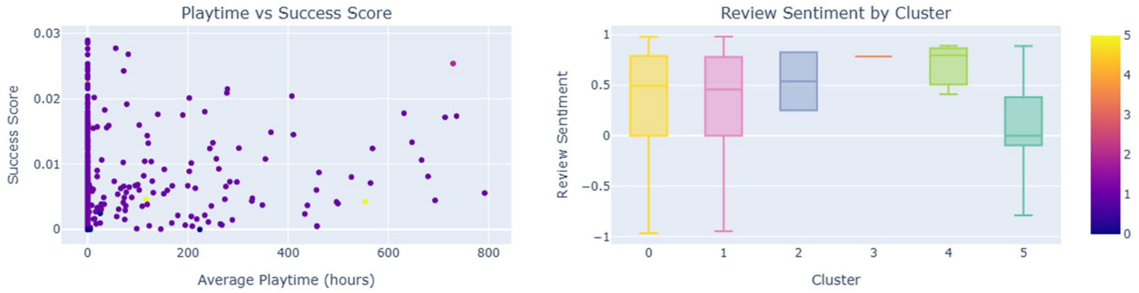

Figure 4: Cluster Analysis

Figure 4 presents two visualizations: (left) the relationship between **average playtime** and **success score**, showing how educational games group into six clusters with varying engagement and success dynamics; (right) a boxplot illustrating **sentiment distribution** across clusters, highlighting differences in player perception and review tone among clusters.

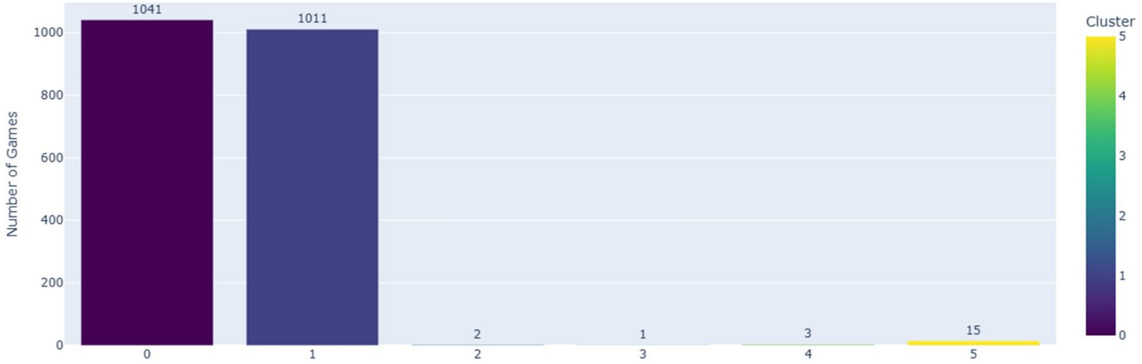

Figure 5: Educational Games per Cluster

The bar chart of Figure 5 shows the **distribution of games across six clusters**, revealing that clusters 0 and 1 dominate the dataset with 1041 and 1011 titles respectively, while the remaining clusters are significantly smaller. This imbalance indicates a concentration of similar types of educational games in the first two clusters.

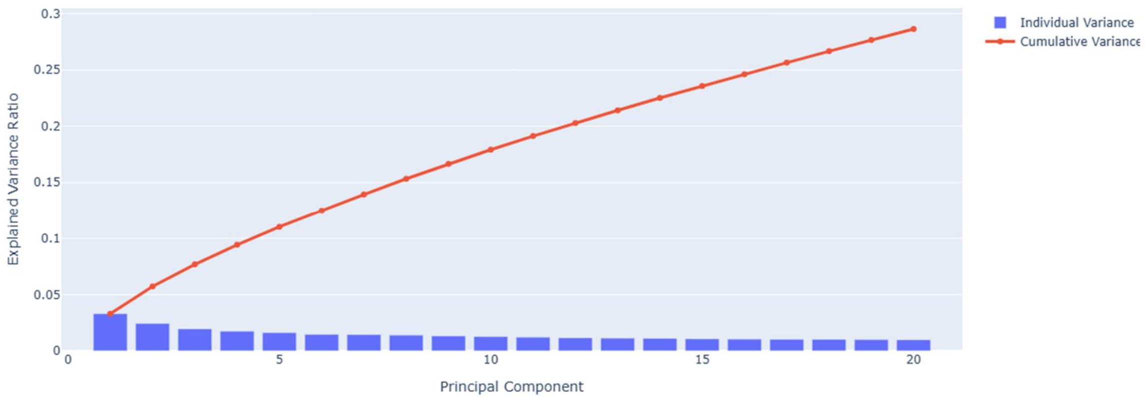

Figure 6: PCA explained Variance Ratio

Figure 6 illustrates the **individual and cumulative variance explained** by the top 20 principal components derived from TF-IDF features. The cumulative variance curve shows that the first few components capture a majority of the textual variance, reflecting that a limited number of latent topics summarize most of the content diversity across educational games.

Table 4.3.1: Cluster Overview Tables

Cluster	**Core Theme**	**Gameplay Type**	**Typical Price**	**Avg Playtime**	**Sentiment**	**Educational Mode**
0	Puzzle & Casual Learning	Short-form, easy-entry	$10	1 h	0.40	Micro-learning / Edutainment
1	Balanced Simulation & Strategy	Mid-depth, replayable	$8	36 h	0.37	Narrative & system learning
2	Technical Training Sims	Realistic / Vocational	$27	1188 h	0.54	Skill-based / Professional
3	Historical Strategy	Grand Strategy	$60	11240 h	0.78	Contextual / Implicit learning
4	Creative & Music Learning	Tool-based / Maker	$54	1894 h	0.70	Learning by creating
5	Sandbox & Open-World Learning	Experimental Simulation	$19	1246 h	0.10	Exploratory constructivism

The six clusters were derived using K-Means clustering (k = 6) applied to a combined feature space encompassing numerical game-related variables—including price, DLC count, achievements, total reviews, positive review percentage, estimated owner count, and average playtime—as well as content-based textual features extracted from each game's full description using TF-IDF and reduced to 20 principal components via PCA. These PCA components capture the dominant language usage patterns across the educational game collection, highlighting common educational and cognitive themes such as puzzle solving, learning, simulation, training, programming, and virtual worlds. Consequently, the clusters reflect both similarities in playstyle and thematic consistency, grouping not only comparable player engagement profiles but also related educational design philosophies and objectives. The silhouette score of 0.4696 indicates a moderate level of cluster separation, suggesting that these clusters meaningfully represent the latent structural and semantic diversity within the dataset of educational games.

As a whole, these clusters define a multidimensional space of educational games in which teaching objectives, technical complexity, and entertainment value coexist in varying proportions. As shown in the table above, each of the six clusters occupies a distinct behavioral and thematic niche within the analyzed educational game ecosystem.

The clusters reveal meaningful structural differences. Cluster 0 consists of small-scale, low-cost, casual, and puzzle-oriented games with limited playtime and minimal community interaction. Such titles typically represent lightweight, individual learning experiences. Cluster 1 builds upon this foundation, encompassing games with somewhat deeper engagement and a higher number of reviews. It reflects accessible yet moderately complex single-player learning titles.

Clusters 2 and 5 include technically sophisticated and simulation-driven games—often realistic and sandbox-oriented—exemplified by titles such as Kerbal Space Program or Car Mechanic Simulator 2021. These games tend to demonstrate strong player retention but also come with higher price points. Cluster 3, though relatively small, captures historically themed and strategic hybrids, such as Total War: ROME II, where educational value emerges through contextual exploration rather than explicit instruction. Cluster 4 encompasses creative development environments and design-focused tools, including RPG Maker and Rocksmith 2014, which blend learning through creation with artistic self-expression.

Across the six clusters, three overarching archetypes emerge. The first consists of cognitive and puzzle-based edutainment represented by Clusters 0 and 1, emphasizing short and goal-oriented learning cycles. The second includes simulation-based and applied-learning games found in Clusters 2 and 5, where realism and experimentation drive player engagement. The third encompasses creative and narrative learning platforms seen in Clusters 3 and 4, linking education with self-expression and contextual storytelling.

This differentiation highlights that educational games on the Steam platform form not a homogeneous genre but a diversified ecosystem in which pedagogical intent, complexity, and player motivation vary substantially across clusters. The clustering framework thus identifies distinct archetypes with characteristic behavioral patterns, offering a deeper understanding of how educational games can be segmented and interpreted.

4.4 Statistical findings

Inferential analyses were performed to assess variability among clusters and potential predictors for the success of educational games. The K-Means clustering ($k = 6$) already showed some quite distinct clusters with a Silhouette Score of 0.4696, indicating some substantial separability between clusters. Distinct price, playtime, and sentiment combinations characterized each cluster, suggesting that the audience and the designs are diversified across the educational games market.

A multiple linear regression model was also fitted to predict success_score from price, achievements, DLC count, average playtime, and review sentiment. The model that explains these two variables results in a R^2 of 0.0237, i.e. these numerical variables explain only a small part of the variance in success if they are considered together. All coefficients were near zero, and no statistically significant effects were found. Yet, the directions of impact followed the prespecified theoretical pattern in the sense that the price and DLC count were weak and negative, and the playtime and sentiment were weak and positive.

The findings were replicated using Pearson correlation analysis, and the fraction of positive reviews (r = 0.387, p < 0.001) and the number of reviews (r = 0.217, p < 0.001) were the two best predictors of success, followed by the average playtime (r = 0.146, p < 0.001). Achievements, DLC count, and other such metrics had weak or no correlation. These results demonstrate that it is the player perception and visibility, but not mediated engagement variables, that are predictors of [educational games'] relative performance on Steam.

In summary, although quantitative participation metrics have limited predictive value, community sentiment and the proportion of positive reviews emerge as the strongest indicators associated with the success of educational games.

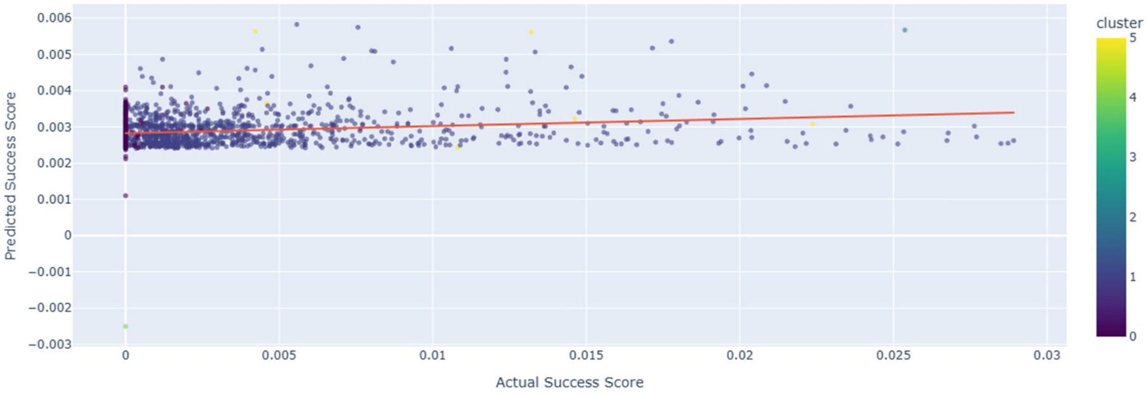

Figure 7: Predicted vs. Actual Success Score by cluster

The scatter plot of Figure 7 compares predicted success scores (from a linear regression model) against actual success scores of educational games. Each color represents a different cluster, illustrating how model performance and success patterns vary across game clusters. The red trendline indicates the overall fit between predicted and actual values.

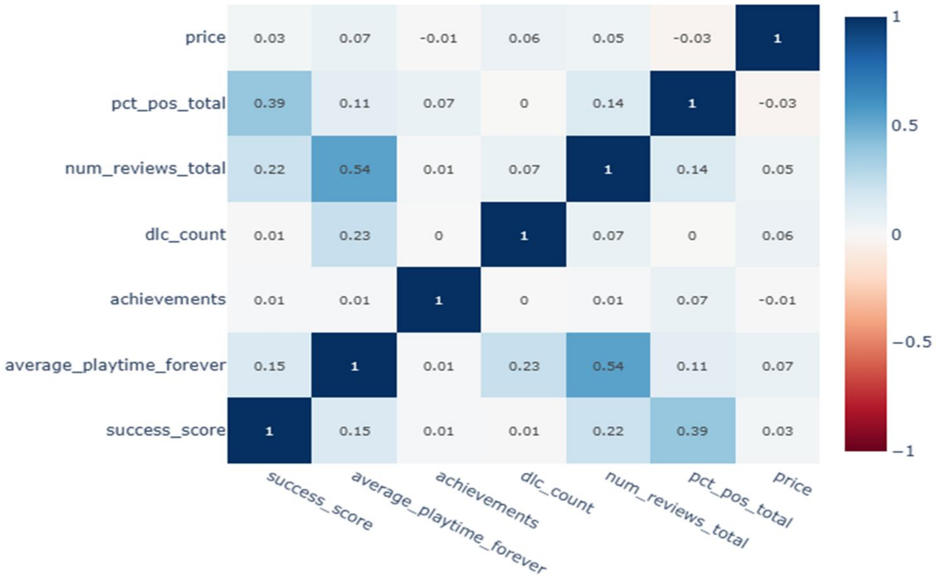

Figure 8: The Correlation Heatmap

The Correlation Heatmap of Figure 8 displays Pearson correlation coefficients between key performance indicators, including success score, average playtime, achievements, downloadable content (DLC) count, and total reviews. The color scale indicates the strength and direction of relationships, with annotations showing exact correlation values.

Together, these inferential analyses confirm the structural b-soundness of the clusters, as well as the theoretical interdependence of affective engagement and quantifiable success within the educational game ecology.

4.5 Sentiment and Genre Trends

Building on the clustering results presented earlier, we continue by exploring the distinctive behavioral, engagement, and thematic patterns that differentiate each cluster. Through detailed examination, we highlight how these clusters embody diverse educational design philosophies and player interaction styles, providing valuable insights into the varied landscape of educational games on the Steam platform.

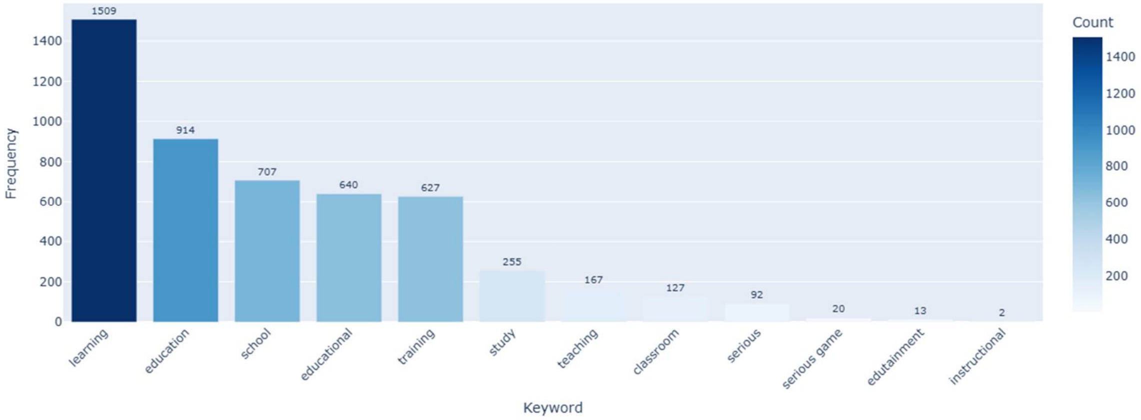

Figure 9: Educational Keyword Frequency in Descriptions

The bar chart of Figure 9 highlights the most frequent educational keywords appearing in game descriptions and tags. “learning,” “education,” and “school`” dominate the dataset, reflecting common themes in educational game development and player engagement trends.

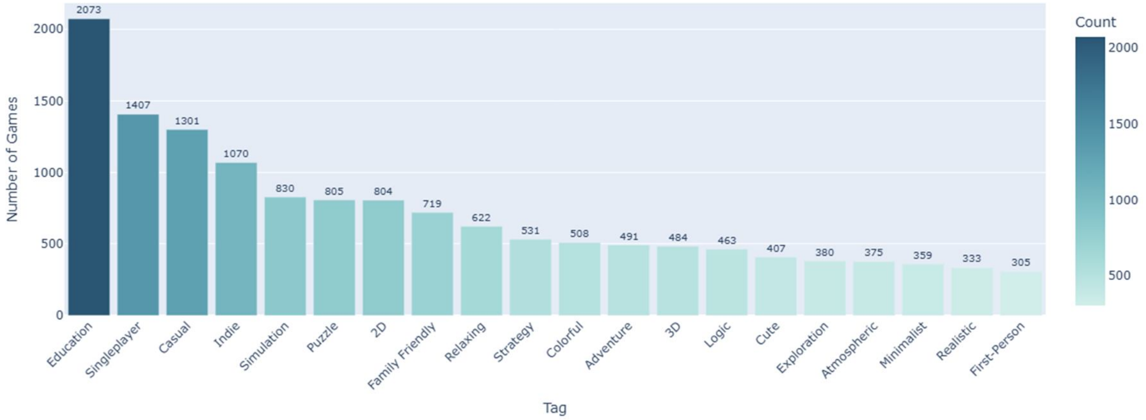

Figure 10: Top Tags in Educational Games

The bar chart of Figure 10 illustrates the most frequently used tags in educational games. Education, Singleplayer, Indie, and Casual dominate the dataset, indicating that most educational titles focus on individual learning experiences and independent development styles. Tags such as Physics, Puzzle, and Strategy also suggest a strong emphasis on problem-solving and conceptual understanding.

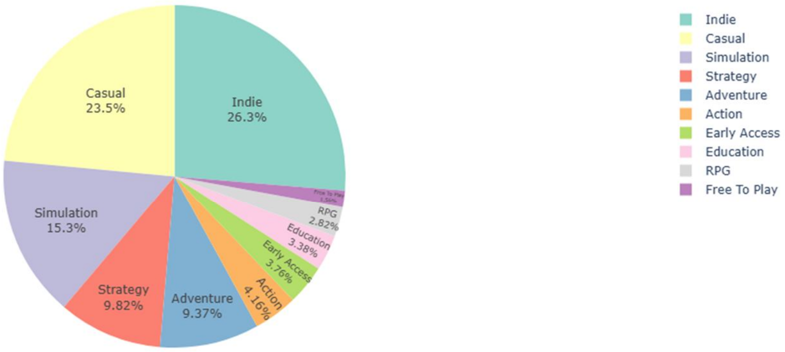

Figure 11: Top 10 Genres in Educational Games

The pie chart of Figure 11 shows the distribution of top genres among educational games. Indie titles represent the largest share (26.3%), followed by Casual (23.5%) and Simulation (15.3%). The diversity across genres like Simulation, Adventure, and Strategy highlights the growing variety of pedagogical approaches within the educational gaming sector.

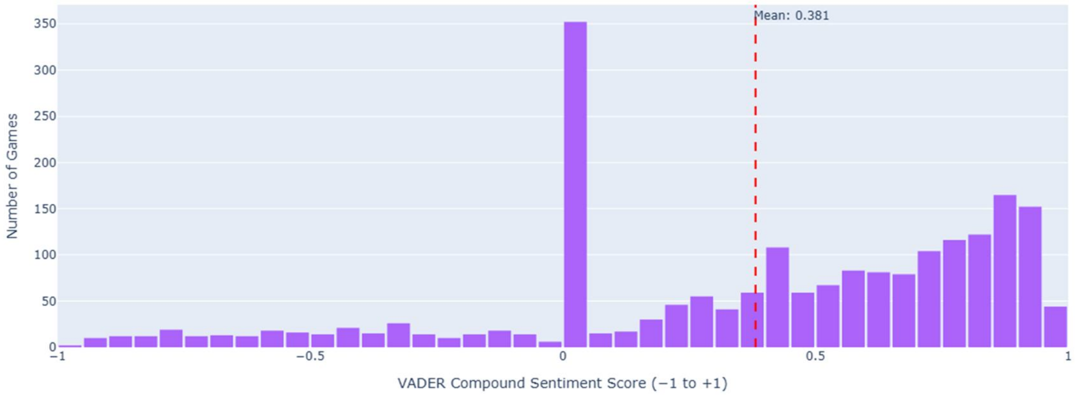

Figure 12: Review Sentiment Distribution

The bar chart of Figure 12 exhibits a significant positive skew since the majority of compound sentiment scores are greater than 0, indicating very positive user experiences.

4.6 Findings

When viewed from the perspective of the whole, the twelve numerical values show a story of the education game on Steam. Descriptive trends show a growing market; cluster analysis reveals pedagogical segmentation; inferential statistics confirm behavioral differentiation; sentiment analysis emphasizes dynamics in user perception.

Combining these results, we find that:

1. Educational games have a quantifiable internal structure based on genre, type of engagement, and emotional tone.
2. Attaining success is more a function of how long a product can keep users interested and emotionally invested rather than how much it costs or how many features it has.
3. Free and easy-to-use educational games receive more widespread approval and social learning diffusion potential.
4. Content analysis reveals that the educational games design lexicon shares much with mainstream entertainment genres, indicating that fun and function are converging.

The findings strengthen theories of game-based learning and digital pedagogy by illustrating how the intersection of learning effectiveness and entertainment design can be empirically mapped using data from a large-scale marketplace, which is an intelligence resource not previously analyzed in this context.

5 CONCLUSION:

This study presents a large-scale computational and statistical analysis of nearly 90,000 games on the Steam platform, identifying approximately 2,000 titles as educational through a rigorous tag- and lexicon-based filtering process. The findings reveal that educational games represent a heterogeneous market, encompassing a spectrum that ranges from puzzle-based cognitive challenges and scientific simulations to creative design tools, programming environments, and casual edutainment hybrids. Market segmentation via a six-cluster K-Means algorithm (Silhouette = 0.4696) uncovers structural and ideological diversity, with clusters distinguished by content complexity, engagement styles, and pedagogical intent. Two clusters focus on short-form cognitive and puzzle games, two represent complex simulation and sandbox experiences, and others mix hybrid and creative-tool genres, reflecting multifaceted design philosophies and player behaviors.

Player engagement and indicators of success were more strongly predicted by sentiment in player reviews and visibility measures—such as the number and positivity of reviews—rather than purely economic metrics like price or concrete gameplay achievements. Education-focused games that emphasize cognitive skills, including short bursts of problem solving, logical reasoning, and creative experimentation, maintain higher user engagement and receive more favorable feedback. In contrast, lengthier and more complex simulations, while appreciated for depth, tend to exhibit diminishing returns in player engagement. These patterns underscore the critical importance of designing educational games with balanced cognitive load and motivational strategies, providing practical guidance for educators and developers aiming to integrate playful learning into formal educational settings.

Methodologically, this research exemplifies a comprehensive, data-driven pipeline that combines text mining (TF-IDF), dimensionality reduction (PCA), clustering (K-Means), regression, correlation, and sentiment analysis to forge connections across educational technology, game analytics, and computational social science. Such integrative analytical frameworks enable scalable mapping and interpretation of digital learning ecosystems, producing insights both empirically robust and practically actionable.

Educational games available on digital platforms display a wide range of diversity and should not be regarded as a single, monolithic category. Their rich variability in content, modes of player engagement, and audience reception highlights the necessity of flexible, human-centered design philosophies—ones that sustain equilibrium between cognitive complexity, accessibility, emotional engagement, and enjoyment. This study thereby advances the understanding that educational games form a multidimensional ecosystem rather than a homogeneous genre, promoting nuanced segmentation and design strategies tailored to diverse educational goals and player motivations.

REFERENCES

[1] R. S. Baker and P. S. Inventado. 2014. Educational data mining and learning analytics. In Learning Analytics, Springer, New York, NY, 61–75.

[2] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle. 2012. A systematic literature review of empirical evidence on computer games and serious games. Computers & Education, 59, 2 (2012), 661–686. https://doi.org/10.1016/j.compedu.2012.03.004

[3] J. P. Gee. 2003. What Video Games Have to Teach Us About Learning and Literacy. Palgrave Macmillan, New York, NY.

J. Hamari, J. Koivisto, and H. Sarsa. 2016. Does gamification work? A literature review of empirical studies on gamification. In Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS '16), IEEE, 3025–3034. https://doi.org/10.1109/HICSS.2016.567

[4] C. J. Hutto and E. Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 216–225.

[5] X. Li and Y. Chen. 2023. Pricing strategies and user engagement in digital learning markets. Journal of Educational Technology & Society, 26, 1 (2023), 12–25.

[6] A. Marchand and T. Hennig-Thurau. 2013. Value creation in the video game industry: Industry economics, consumer benefits, and research opportunities. Journal of Interactive Marketing, 27, 3 (2013), 141–157. https://doi.org/10.1016/j.intmar.2013.05.001

[7] K. Orland, et al. 2021. Predicting game success on Steam: A sentiment and metadata analysis of 60,000 titles. .Entertainment Computing, 38 (2021), 100400. https://doi.org/10.1016/j.entcom.2021.100400

[8] J. L. Plass, B. D. Homer, and C. K. Kinzer. 2015. Foundations of game-based learning. Educational Psychologist, 50, 4 (2015), 258–283. https://doi.org/10.1080/00461520.2015.1122533

[9] C. Romero and S. Ventura. 2020. Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50, 3 (2020), 1036–1049. https://doi.org/10.1109/TSMC.2019.2905099

[10] R. M. Ryan and E. L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55, 1 (2000), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

[11] S. Seufert, et al. 2022. Market segmentation of indie games using clustering and NLP on Steam metadata. Journal of Game Analytics, 4, 1 (2022), 45–63.

[12] A. Sirola, et al. 2023. Sentiment-based prediction of game popularity and retention in digital game platforms ..Entertainment Computing, 45 (2023), 100555. https://doi.org/10.1016/j.entcom.2023.100555

[13] P. Wouters, et al. 2013. A meta-analysis of the cognitive and motivational effects of serious games. Journal of Educational Psychology, 105, 2 (2013), 249–265. https://doi.org/10.1037/a0031311

[14] P. Wouters, et al. 2019. Serious games and learning: A theoretical framework for design and evaluation. Educational Technology Research and Development, 67 (2019), 1005–1029. https://doi.org/10.1007/s11423-019-09650-4

[15] T. Hussein, et al. 2022. TF-IDF and semantic pattern extraction for educational datasets. Computers & Education, 181 (2022), 104419. https://doi.org/10.1016/j.compedu.2022.104419

[16] V. Kovanović, et al. 2015. Learning analytics in massive open online courses: The role of clustering. Computers in Human Behavior, 47 (2015), 69–76. https://doi.org/10.1016/j.chb.2014.09.044