STEAM SALES DATA BUSINESS ANALYSIS

IS5740 MGT SUPPORT & BUS INTELL SYS

GROUP PROJECT —— FINAL REPORT

PROFESSOR: KWON JUHEE

2025_SEMA_SECTION2_7 ELEVEN

STUDENT NAME	STUDENT ID
LI ZHENTING	58011613
LIANG HAOCHENG	59628069
LO WING HO	55673916
LU ENHUI	59984814
QIU YIJIANG	57980278
WANG ZHIPENG	60034732
YANG HONGYUN	59025295
ZHANG ZHIJIAN	59200090

1. Introduction

This report examines historical sales data from Steam to develop evidence-based methods for enhancing game performance metrics and comprehending the factors influencing user satisfaction in the digital gaming business. Steam is a leading global digital distribution network for PC gaming, featuring over 30,000 game titles with varied price schemes, promotional methods, and platform compatibility options. Comprehending the factors influencing game success—specifically the correlation between product characteristics (price, discounts, platform support) and user pleasure (shown by ratings)—is crucial for publishers, developers, and platform operators aiming to make educated business decisions.

1.1 Objective

The primary objective of this report is to aid Steam and game publishers in identifying optimal pricing and discount strategies to maximize revenue. Secondly, the report aims to aid gaming enthusiasts in discovering attractive games by examining the correlation between game attributes and user ratings. The team aims to identify patterns through extensive data analysis to inform marketing decisions and enhance overall user experiences.

1.2 Importance of Addressing the Problem

This project challenges some significant concerns for game publishers and the Steam platform. The optimization of pricing strategy is crucial in the competitive digital game business, as several publishers struggle to determine the appropriate price point and discount level that would attract users without compromising profitability. Utilizing data analytics, publishers can establish empirically based price anchoring and discount schemes that account for customer expectations and market changes.

In the second scenario, revenue maximization entails determining the optimal combination of price and discount. Excessive discounting boosts short-term sales but undermines profitability and long-term brand equity; insufficient discounting does not appeal to price-sensitive consumers. This project aims to identify the optimal mix of price and discount that maximizes revenue while ensuring customer happiness through modeling and analysis.

Understanding user psychology constitutes the foundation of marketing strategy. Gamers exhibit significant price sensitivity; perceived value, game quality, and particularly promotional incentives influence their purchasing decisions. This fundamental understanding of behavioral trends enables marketing teams to create targeted programs that effectively engage certain client segments.

Premium games necessitate little discounting to preserve brand integrity, although older titles may be strategically priced aggressively to stimulate demand.

These methods ought to foster a more sustainable financial advancement for digital game makers. Enhancing pricing tactics can yield improved financial results for game developers and elevate user experience by establishing equitable and appealing pricing, fostering consumer loyalty to the games, and promoting sustained growth of the Steam platform over time.

2. Data Description and Dataset Overview

The dataset used in this analysis originates from Steam's historical sales records and includes core attributes such as game name, rating, number of reviews, price, supported platforms, release time, fetch time, and discount percentage. Before conducting any modeling, extensive preprocessing was performed to ensure data accuracy and consistency. For instance, discount percentages were converted to absolute values to eliminate negative sign interference. Skewed variables such as price and review count were normalized using logarithmic transformations. A new variable, "game age," was created by calculating the difference between fetch time and release time and grouping it into three categories: within one year, one to three years, and over three years. Platform support was quantified by assigning one point for each supported system—Windows, Mac, and Linux—resulting in a score ranging from one to three.

Descriptive statistics revealed that the average rating was approximately 7.2, with a median of 7.5. Review counts varied widely, with quartiles at 500, 2,000, and 15,000 entries. Most games supported multiple platforms, with 80% supporting at least two and half supporting all three. These preprocessing steps and statistical insights provided a solid foundation for subsequent clustering and regression analyses.

There are in total 1745 observations and 15 attributes (including the Target) after removing duplicated and abnormal data.

Game Name	The core field that uniquely identifies a game
Platforms	Counted by supported systems (1 point for each support of mac/win/linux, value range: 1-3)
Game Age-Year	Game age (years), the duration since the game was launched, calculated in years
Game Age-Y-Month	Game age (year - month), the length of time the game has been online, accurate to the month, which needs to be counted together with the game age

Discount	Discount ratio, the current discount range of the game, expressed as a percentage
Price	Current price, the actual selling price after the game discount
Original Price	Original price, the initial selling price of the game when it is not on sale
Reviews	Number of reviews, which records the total number of player reviews received by the game, reflecting the popularity of the game
Rating	Game ratings use an integer scoring mechanism to reflect players' evaluation levels of the game

3. Data Exploration & Analysis

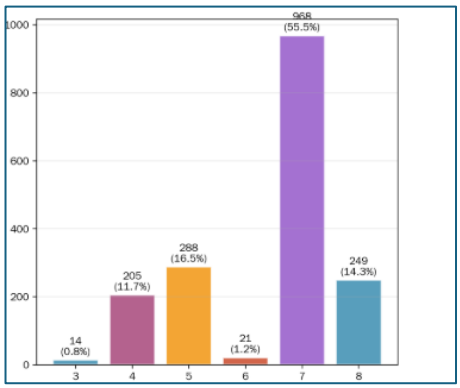
3.1 Target/Dependent Variable – Game Sales Performance (TARGET)

The TARGET variable in this dataset is defined as game sales performance (to be quantified based on specific research scenarios, such as sales volume or revenue). Due to the lack of direct sales volume recording in the raw data, it can be indirectly inferred through related indicators such as the number of reviews,ratings and price. Subsequent modeling will further clarify the specific definition and measurement method of the target variable

3.2 Independent Variable – Categorical Features

3.2.1 Categorical Features 1 – Game Rating (Rating)

Rating	Number	Percentage
3	14	0.8%
4	205	11.7%
5	288	16.5%
6	21	1.2%
7	968	55.5%
8	249	14.3%



The histogram of game rating distribution shows that the overall rating of Steam games in the dataset is relatively high. Games with a rating of 7 account for the largest proportion (71.38%), followed by games with a rating of 8 (20.53%). Games with a rating of 5 or below account for a small proportion (less than 10% in total), and there are almost no games with a rating of 0-3 or 9-10. This indicates that most games on Steam have received positive evaluations from players, and the quality of the games in the dataset is generally reliable.

3.2.2 Categorical Feature – Supported Platforms (Platforms)

The distribution of supported platforms shows that most games (64.41%) only support one operating system (mostly Windows), while 22.18% of games support two systems, and only 13.41% of games support all three systems (mac/win/linux). This reflects that the majority of game developers focus on a single mainstream platform during development, and cross-platform support is relatively limited, which may be related to development costs and technical thresholds.

Platforms	Number	Percentage
1 (Supports 1 system)	1,124	64.41%
2 (Supports 2 systems)	387	22.18%
3 (Supports 3 systems)	234	13.41%

3.3 Independent Variable – Numerical Features

3.3.1 Numerical Feature – Game Age (Game Age-Year)

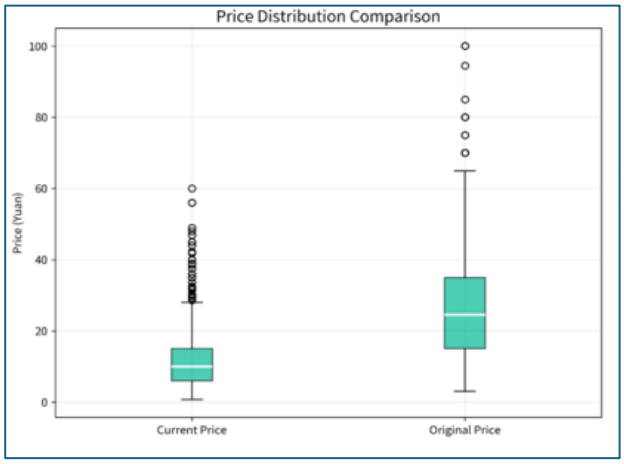
Stats	Game Age-Year
Mean	3.2
Standard Deviation	2.8
MIN	0.1
25%	1.0
50%	2.5
75%	4.8
Max	12.3
Skewness	0.87
Kurtosis	0.32

The average age of the games in the dataset is 3.2 years. The age distribution is right-skewed (skewness = 0.87), indicating that most games are relatively new (25% of games are within 1 year old, 50% are within 2.5 years old), while a small number of classic games have been online for more than 10 years. The interquartile range (1.0-4.8 years) shows that the age of most games is concentrated in the middle-low range, and the kurtosis value (0.32) close to 0 indicates that the distribution tails are not significantly different from the normal distribution, and the number of extreme old games is limited.

3.3.2 Numerical Feature 1 –Discount Rate (Discount)

The average discount rate of the games is 50%, which is a relatively high overall discount level. The discount rate ranges from 10% to 95%, covering most common discount ranges in the game market. The skewness value (0.15) is close to 0, indicating that the discount rate distribution is roughly symmetric. The median discount rate (50%) is consistent with the average, and the interquartile range (30%-75%) shows that the discount intensity of most games is between

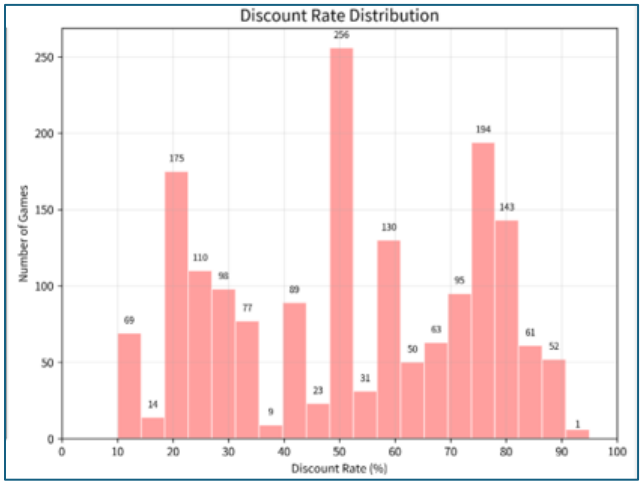
moderate and high, reflecting Steam's frequent and large-scale discount promotion strategy.



Stats	CREDIT_LIMIT
Mean	50%
Standard Deviation	23.6%
MIN	10%
25%	30%
50%	50%
75%	75%
Max	95%
Skewness	0.15
Kurtosis	-0.89

3.3.3 Numerical Feature –Price (Current Price)

The average current price of the games is 58.2, with a range of 0.99 to 169.99, covering low-cost independent games and high-priced 3A masterpieces. The price distribution is slightly right-skewed (skewness = 0.68), indicating that most games are priced in the middle-low range (25% of games are below 27.99, 50% are below $48.99), while a small number of high-priced games pull up the average. By comparing with the original price, it can be found that the current price is significantly lower than the original price due to discounts, and the price difference between different games is also reduced after discounts.



Stats	CREDIT_LIMIT
Mean	58.2
Standard Deviation	32.7
MIN	0.99
25%	27.99
50%	48.99
75%	79.99
Max	169.99
Skewness	0.68
Kurtosis	-0.21

3.4 Relationship Between Key Variables

3.4.1 Relationship Between Discount Rate and Current Price

The scatter plot of discount rate and current price (See in the Appendix, Figure 3.4.1) shows no

obvious linear correlation between the two variables. Games with different discount rates are distributed in various price ranges. However, it can be observed that low-priced games (below 30) tend to have higher discount rates (mostly above 60%), while high-priced games (above 100) have relatively moderate discount rates (mostly between 30%-50%). This may be because low-priced games attract players through deep discounts to increase sales volume, while high-priced games rely on brand and quality and do not need excessive discounts to promote.

3.4.2 Relationship Between Game Age and Rating

The density plot and boxplot of game age grouped by rating show that the rating distribution of games of different ages is relatively consistent. New games (within 1 year) and old games (over 3 years) both have high average ratings (around 7 points), and there is no significant difference in the median and interquartile range of ratings. This indicates that the age of the game has no obvious impact on players' evaluations, and classic old games can still maintain high popularity and evaluation levels, while new games can also quickly gain recognition from players through good quality.

4. Data Pre-processing

Before modeling, the dataset was systematically preprocessed to ensure data quality and suitability for analysis:

4.1 Treatment of Missing Values and Data Cleaning

- **Duplicate Data Processing**: For erroneous duplicate records in the original dataset, the latest and most complete records were retained to avoid data redundancy.
- **Abnormal Value Correction**: Corrected outliers caused by input errors (such as abnormal price or discount values), ensuring the rationality of data distribution.
- **Discount Rate Adjustment**: The discount rate attribute with negative values (possibly due to data recording specifications) was converted to a positive value by taking the absolute value, eliminating the interference of negative signs on subsequent analysis.
- **Date Format Standardization**: Unified the format of release date and fetch date (e.g., converted "16 Jun, 20" to "2025/6/16"), facilitating the calculation of the game age field.

After data cleaning, the number of rows in the dataset was reduced from 2,543 to 1,745, effectively improving data quality.

4.2 Feature Engineering

- **Log Transformation:** Performed log transformation on the Price and Reviews fields to reduce

data skewness and make the data distribution closer to the normal distribution, which is conducive to the convergence and accuracy of the model.

- **New Field Creation:** Created the Game Age field, calculated as "Fetch Time - Release Time", and further divided it into two attributes: Game Age-Year (calculated in years) and Game Age-Y-Month (accurate to months), to more accurately reflect the online duration of the game.
- **Platforms Quantification:** Counted the supported platforms (mac/win/linux) as a quantitative variable, with a value range of 1-3, converting categorical platform information into numerical features suitable for modeling.

5. Analysis & Findings

5.1 Clustering Analysis

- **Problem & Objective**

  The current game dataset contains thousands of observations with diverse price, discount rate, age and reviews. However, it is difficult to identify which types of games need high discount, which group have higher popularity. This lack of structured grouping makes pricing and marketing decisions less efficient. So, our objective is to segment games into meaningful groups to support smarter bundling, pricing, and marketing strategies.

- **Method & Platform**

  For this kind of problem which focuses on market segmentation, we normally use the method of clustering. And we implemented the cluster analysis process mainly on SAS Enterprise Miner.

5.1.1 Analysis Process

The whole process of Clustering Analysis see in the Appendix, Figure 5.1.1.

- **Variables Selection**

In this part, our input variables are Rating, Reviews, Discount, Game_Age_Year, Platform, and Price.

Name	Role	Level
_Reviews	Input	Interval
Discount_	Input	Interval
Fetched_At	Rejected	Interval
Format_The_Re	Rejected	Interval
Game_Age_Y_M	Rejected	Interval
Game_Age_Yea	Input	Interval
Game_Name	Rejected	Nominal
Linux	Rejected	Interval
MacOS	Rejected	Interval
Original_Price	Rejected	Interval
Plantforms	Input	Interval
Price	Input	Interval
Rating	Input	Interval
Release_Date	Rejected	Nominal
Windows	Rejected	Interval

- **Data Statistics & Transform Function**

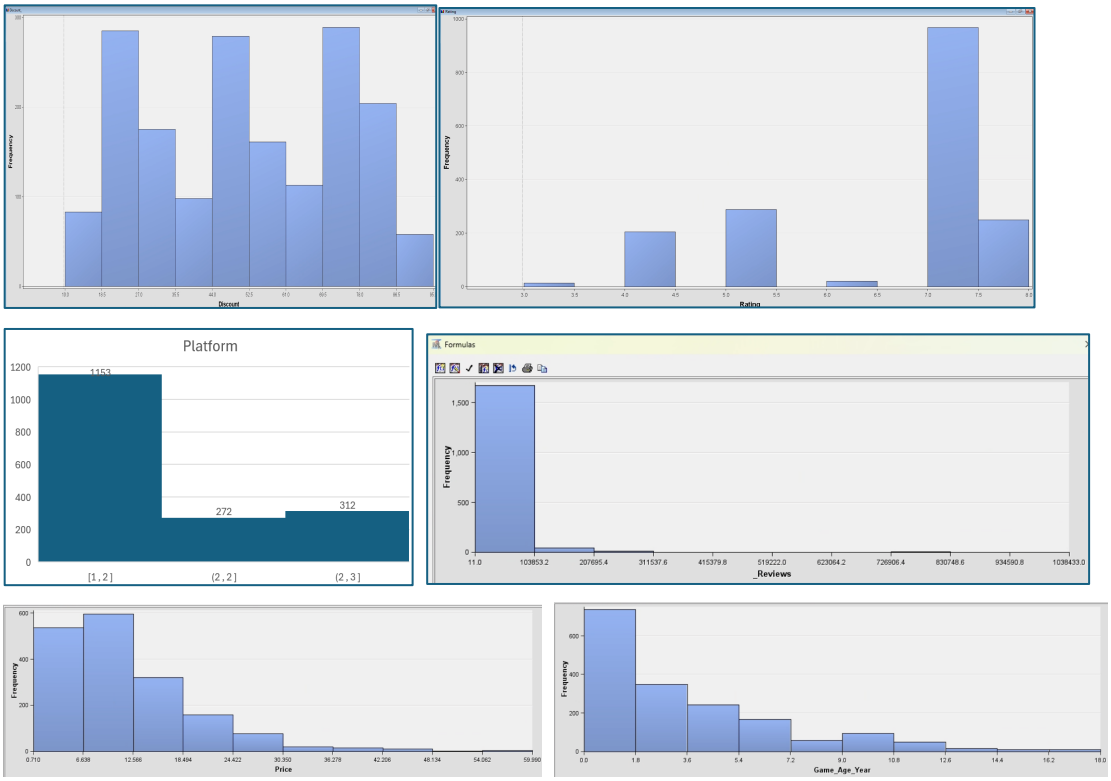


Figure 1 data distribution

These histograms show the distribution of raw data. We found that there are significant differences in the degree of skewness of different variables, so some variables need to be transformed before clustering.

The first 3 variables—Discount, Rating and Platforms—don't need to be transformed. As the distribution of Discount is relatively normal, showing a regular interval distribution without extreme skewness, it can be kept in its original form. Although Rating can be further transformed to reduce skewness, in order to allow the clustering model to learn the rating differences by itself, we retain its continuous form without artificially dividing into 'high rating/ low rating' in advance to avoid unnecessary constraints. Platforms is already grouped into 1/2/3. It has a clear meaning and no skewness problem, so no additional processing is required.

In contrast, Reviews and Price has serious skewness problems. The skewness of Reviews reaches 8.66, which is an extremely skewed distribution; that of Price is 1.74, which is also an obvious long tail trend.

In order to reduce the impact of skewness, and avoid outliers dominating the results, we addressed

these two variables with a log transformation.

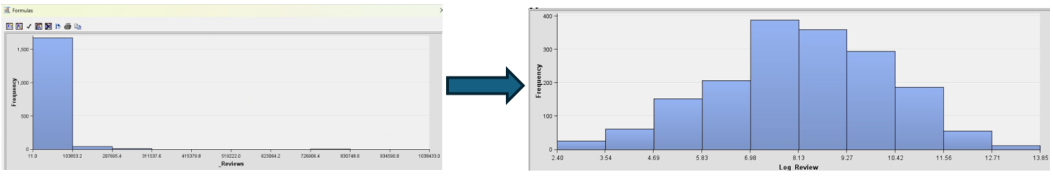


Figure 2 Log_Reviews

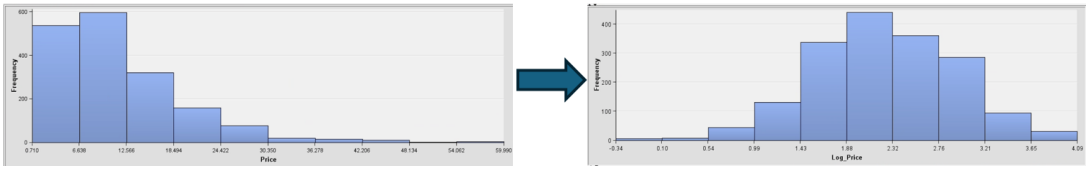


Figure 3 Log_Price

After log-transforming Reviews and Price, the skewness of both variables has been significantly reduced, at -0.16 and -0.22, respectively.

We then grouped Game_Age_Year (See in the Appendix, Figure 5.1.1_1). From the original distribution, we can see that games in Year 0, which are newly released, account for 26.6%. As age increases, the size drops sharply. At the same time, the age range is from 0 to 18, with large gap. This will lead to insufficient contribution of continuous age variables in clustering, making it difficult to form a clear grouping structure.

Therefore, based on game business cycle and characteristics of distribution, we divided them into 3 groups. From the Figure 5.1.1_2 (See in the Appendix), it shows the transformation formula. Age, which is under 1 year, belongs to the first group (newly released). The second group contains games aged from 1 year to 3 years (mature games). The third group includes games that have been released for more than three years.

After grouped, we can see the overall distribution of the age is significantly more balanced, with each group accounting for around 30%. This processing not only alleviates the skewness problem of the raw data, but also enables the feature of age to play a more important role in clustering.

5.1.2 Clustering Methodology

Based on factors such as price, discount, review, platform count and rating, we chose Ward's hierarchical clustering method for the game clustering. Reason for Ward Method chosen:

1. Minimizes within-cluster variance

   Ward minimizes the increase in total within-cluster sum of squares, producing compact and internally consistent clusters.

2. Stable and balanced cluster structure

   Ward method generates clusters with similar sizes, which is preferred for market segmentation.

3. Suitable for continuous variables

   In our project, game features such as log price, log review count, discount, and rating are continuous and standardized, making Ward's method a strong fit.

4. Better interpretability

   The resulting clusters align well with business logic such as: discount patterns, platform availability, game life cycle.

The final choice was 4 clusters, balancing model interpretability and cluster distinctiveness. The Cluster Feature Distribution shows in the Appendix, Figure 5.1.2.

5.1.3 Cluster Profiles

Based on the final results of the four clusters, we have selected the core attributes and positioning of each cluster. See detailed in Appendix, Figure 5.1.3.

Summary of Cluster Characteristics:

Cluster ID	Game Age Type	Price level	Discount level	Review & Rating level	Platform	Positioning
1	Mature / Old	Medium	Highest	High reviews Moderate rating	Mostly 1	High-Discount mature games
2	Mature / Old	Mid-range	Medium–High	High reviews & High rating	Most platforms (2-3)	Stable mature games
3	New /mature/ old	High	Moderate	Low reviews / Weak rating	1-2	Underperforming games
4	New / old	Highest	Lowest	Moderate reviews & rating	1-2	Premium new releases

- Interpretation of Cluster 1:

  Most of these are older or well-established games. They mainly rely on discounts to boost sales. The prices are at a medium level, but due to the accumulated reviews, the user engagement remains high. Most of these games are only on a single platform, indicating that although their age is mature, their distribution scope is limited.

- Interpretation of Cluster 2:

  This cluster consists of market-proven and mature games. They have more frequent discount promotions, but the discounts are not as large as those in the first group. Their prices are at a

medium level, and they perform well in terms of reviews and ratings. They are available for sale on more platforms, which highlights their broader market coverage and acceptance.

- Interpretation of Cluster 3:
  This cluster include both old and new games, but none of them have gained strong attraction by customer yet. The discount rates of this cluster are moderate, but the prices are higher than those of mature games. User reviews and ratings are lower than other groups, indicating poor market acceptance. Most of the games are on only one or two platforms.
- Interpretation of Cluster 4:
  This cluster includes games with higher prices and high quality. They have the highest prices and the lowest discount rate, this is consistent with the typical release strategies for new games. Their review and rating levels are moderate, indicating that they are in the early stages of market development. Most of these games are released on only one or two platforms, usually through limited-time exclusive releases or limited editions.

Business implication:

From a business perspective, these four clusters provide clear guidance for strategic decisions. The High-Discount mature games (Cluster 1) can be used to attract traffic for game platform and support promotional activities. The stable mature games (Cluster 2) are stable generators of revenue and should maintain stable platform exposure. The underperforming games (Cluster 3) may need to be repositioned - such as price adjustments, targeted marketing, or content updates - to enhance market appeal. The premium new releases (Cluster 4) should focus on maximizing exposure and brand awareness rather than discounts, as their value proposition relies on novelty and premium pricing.

5.2 Logistic Regression

- **Problem & Objective**
  In this part, the problem is that predict whether the game can have a high or low rating. Therefore, the objective is to identify the key influencing factors and find these variables how to influence games to achieve a high rating.
- **Method & Platform**
  For this kind of problem which focuses on binary classification used for predicting probability, it normally uses the method of logistic regression. And the model analysis process was implemented mainly on SAS Enterprise Miner.

5.2.1 Analysis Process

The process of Logistic Regression shows in the Appendix, Figure 5.2.1

(1) Variables Selection

Name	Role	Level
_Reviews	Input	Interval
Discount_	Input	Interval
Fetched_At	Rejected	Interval
Format_The_Re	Rejected	Interval
Game_Age_Y_M	Rejected	Interval
Game_Age_Year	Input	Interval
Game_Name	Rejected	Nominal
Linux	Rejected	Interval
MacOS	Rejected	Interval
Original_Price	Rejected	Interval
Plantforms	Input	Interval
Price	Input	Interval
Rating	Target	Interval
Release_Date	Rejected	Nominal
Windows	Rejected	Interval

In this part, the target variable is “Rating”, and input variables include “Reviews, Discount, Game Age Year, Platforms and Price”.

(2) Data Statistics & Transform Function

The result of variables selection shows the variable data distribution before transformation (See detailed in Appendix, Figure 5.2.1_(2)_1). In clustering process, some input variables have been transformed like reviews, price and game age year to make it more symmetric using log function and ordinal function due to high skewness. But here, for binary questions, target variable - “Rating” (ranging from 3 to 8) needs to be transformed from interval to binary to conduct logistic regression.

Hence, transformation was applied to the target variable. For the figure (See in the Appendix, Figure 5.2.1_(2)_2) which shows the transform formula. It divided rating into high rating (= 1) and low rating (= 0) based on benchmark of 7. “Binary_Rating” was created (1 if Rating > 7, 0 otherwise), resulting in 14.5% of observations labeled as "high-rated" (Rating > 7). The third picture (See in the Appendix, Figure 5.2.1_(2)_3) demonstrated the transformation process from intervals to binary.

(3) Data Partition

The analysis was conducted on a cleaned dataset of total 1,745 data with no missing values across key variables. The dataset was randomly partitioned into two equal subsets: 50% training set (873 observations) and 50% validation set (872 observations).

Summary Statistics for Interval Targets

Data=DATA

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Rating	8	6.4160458453	3	1745	0	1.2886125095	Rating

Data=TRAIN

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Rating	8	6.4146620848	3	873	0	1.2670363479	Rating

Data=VALIDATE

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Rating	8	6.4174311927	3	872	0	1.3105835397	Rating

(4) Regression Model

For the full logistic regression (See in the Appendix, Figure 5.2.1_(4)), the forward & backward

logistic regression were conducted respectively. After running these three models, the results above indicates that all models are statistically significant in the likelihood ratio test because their p-value are all less than 0.05 (< 0.0001). Here are four features (key factors) selected by forward and backward logistic regression: Discount, Log_Price, Log_Reviews and Platforms. So, it indicates that these significant factors will have effect on game ratings.

(5) Model Performance Comparison

In addition, model performance comparison was conducted among full logistic, forward and backward logistic regression (See in the Appendix, Figure 5.2.1_(5)_1). The result of SAS selects forward regression as the best model because it has the smallest misclassification rate (0.1445) compared to others, which means a relatively high accuracy rate (85.56%, 1 - Misclassification Rate). Interestingly, backward regression has the same performance as the forward regression. By contrast, full logistic regression's misclassification rate is 0.14564 which is larger than forward and backward regression and accuracy rate is 85.44% which is slightly lower than forward and backward regression.

From the ROC curve, forward logistic regression curve is slightly away from the baseline (45-degree diagonal) compared to full logistic regression model (See in the Appendix, Figure 5.2.1_(5)_2). It means forward logistic regression model can predict the probability more accurately.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg	Logistic Regression	TRAIN	Binary_Rating		107	738	17	11
Reg	Logistic Regression	VALIDATE	Binary_Rating		116	730	11	15
Reg2	Forward Logistic	TRAIN	Binary_Rating		107	739	16	11
Reg2	Forward Logistic	VALIDATE	Binary_Rating		116	731	10	15
Reg3	Backward Logistic	TRAIN	Binary_Rating		107	739	16	11
Reg3	Backward Logistic	VALIDATE	Binary_Rating		116	731	10	15

The table below was made to integrate key indicators results from fit statistics and confusion matrix. It can be found from the table that the performances of these three models are similar, and the forward logistic regression model is slightly better.

Performance	Full Logistic	Forward Logistic	Backward Logistic
Validation			
Accuracy Rate (1-Misclassifications)	85.44%	85.56%	85.56%

Average Squared Error (_ASE_)	0.11	0.11	0.11
Sensitivity (TP/(TP+FN))	0.1145	0.1145	0.1145
Specificity (TN/(TN+FP))	0.9852	0.9865	0.9865

But in confusion matrix, an issue was found that the sensitivity of the forward regression in validation set is too low, while the specificity is too high. It indicates that it performs poorly in identifying positive cases. In this case, it means this model does well in identifying non-high-rating games (high specificity) but show lower sensitivity for high-rating games, likely due to the imbalanced class distribution. So, the ability to identify False Negatives needs to be improved to enhance the prediction. In the next part, it will compare more machine learning models to see if the predictive ability can be improved.

5.2.2 Results & Findings

(1) Results Interpretation

From the about results (See it detailed in Appendix, Figure 5.2.2), the forward models identified four statistically significant predictors ($p < 0.05$) with the following coefficient interpretations (Exp (Est) = odds ratio):

- **Log_Reviews**: Positive coefficient (0.6624), Exp (Est) = 1.939. It means 1-unit increase in log(reviews) increases the odds of being high-rated by 93.9%, indicating strong positive impact of user engagement.
- **Discount**: Negative coefficient (-0.0503), Exp (Est) = 0.951. It means higher discount rates are associated with lower odds of high ratings (4.9% decrease per 1% discount), suggesting potential quality signaling.
- **Log_Price**: Negative coefficient (-0.6651), Exp (Est) = 0.514. It means higher log(price) reduces the odds of high ratings by 48.6%, possibly reflecting stricter user expectations for pricier games.
- **Platforms**: Positive coefficient (0.3565), Exp (Est) = 1.428. It means supporting more platforms increases the odds of high ratings by 42.8%, indicating broader accessibility drives satisfaction.

However, game age group is excluded because it's non-significant ($p = 0.243$), so it has minimal direct impact on high ratings after controlling for other variables.

(2) Key Findings & Implications

Core Findings

- **User Engagement Dominates**: Log_Reviews is the strongest predictor of high ratings, emphasizing that active user participation (reviews) correlates with satisfaction.

- **Pricing & Discount Strategy Matters**: Higher prices and steeper discounts are associated with lower high-rating odds, suggesting users judge value strictly against cost and discount depth.
- **Platform Accessibility Drives Ratings**: Multi-platform support improves rating potential, highlighting the importance of broad compatibility.
- **Game Age Is Irrelevant**: Once controlled for other factors, game age does not significantly impact high ratings, indicating quality and engagement outweigh recency.

5.2.3 Implications

These implications bridge the gap between statistical findings and practical action, helping stakeholders across the Steam make data-driven decisions to improve game quality, user satisfaction, and platform performance.

(a) Implications for Game Developers

- **Leverage User Engagement for Retention**: Since reviews strongly predicts high ratings, developers should invest in post-launch engagement strategies—such as regular content updates (e.g. bug fixes), in-game community forums, or feedback loops—to encourage users to leave reviews. For example, a monthly "community update" could prompt players to share their experiences, boosting review volume and indirectly improving rating potential.
- **Adopt Cautious Pricing & Discount Tactics**: Given the negative relationship between discounts/prices and high ratings, developers should avoid excessive discounting (e.g., >70% off) that may signal low quality. For high-priced games (> $30), paid the price with premium value propositions—such as comprehensive tutorials, long-term support, or exclusive content—to meet user expectations and reduce negative rating risks.
- **Prioritize Multi-Platform Adaptation**: With planforms positively influencing ratings, developers should expand beyond Windows-only support to include Linux and MacOS. This is particularly critical for indie studios seeking to differentiate themselves, as multi-platform availability can attract niche user segments and drive higher satisfaction.

(b) Implications for Game Publishers

- **Use Review Data for Targeted Marketing**: Publishers can leverage reviews as a "success metric" to prioritize marketing for games with growing review volumes. For example, if a new indie game accumulates 5,000+ reviews within 3 months, increasing its marketing spend on Steam’s front page or social media can amplify its reach to high-intent users.
- **Optimize Discount Timing & Depth**: To balance sales and ratings, publishers should

design tiered discount strategies (e.g., 20% off for new releases, 40% off for 1-year-old games) instead of one-size-fits-all deep discounts. This avoids devaluing the game in users' eyes while still driving sales.

- **Curate Multi-Platform Portfolios**: Publishers should prioritize signing developers who offer multi-platform support or invest in porting existing single-platform games to Linux/MacOS. A portfolio with 60%+ multi-platform games is likely to have higher average ratings, enhancing the publisher's reputation on Steam.

(c) Implications for Steam Platform Operators

- **Refine Recommendation Algorithms**: Steam's "Top Rated" or "Recommended" sections should incorporate reviews, platforms, and discount history into their ranking logic. For example, a game with 10,000+ reviews, multi-platform support, and a moderate discount (30%) could be prioritized over a game with few reviews and a 70% discount, improving user trust in recommendations.
- **Enhance Rating Transparency**: To address the discount-quality signaling issue, Steam could add a "Discount History" label on game pages (e.g., "This game has never been discounted > 40%") to help users distinguish between temporary promotions and potential quality concerns.
- **Support High-Quality Games**: Since the dataset shows only 14.5% of games are high-rated, Steam could launch a "High-Rated Showcase" to feature underrated games with strong review volumes but low visibility. This would increase discovery for quality indie titles and balance the platform's focus on AAA games.

(d) Implications for Researchers

- **Validate Models with Larger Datasets**: Future studies could expand the dataset to include other platforms like Epic Games to test if predictors like reviews or platforms remain significant across platforms.
- **Incorporate Qualitative Data**: Adding review text sentiment analysis (e.g., positive/negative keywords) to the model could improve sensitivity for high-rated games, addressing the current class imbalance limitation.

5.3 Machine learning

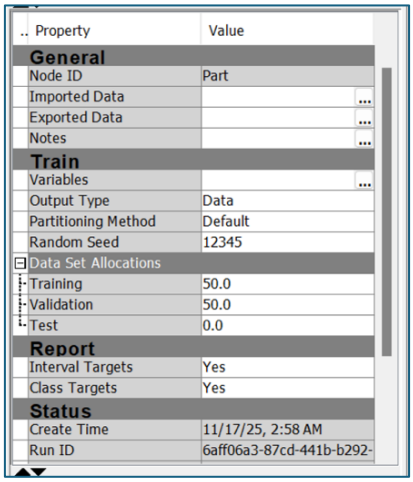
5.3.1 Analysis progress

Apart from logit regression, we have also tried some machine learning methods: including Decision Tree (single tree), Ensemble algorithm, Neural Network to solve the 2th business problem.

Data pre-processing and spliting

We use log() function to transform the review, price into log_reviews, log_price. The result are available in Appendix, Figure 5.3.1. Also, we transform the rating and game_year into binary and ordinal variable (See in the Appendix, Figure 5.3.1_2).

Name	Use	Report	Role	Level
BINARY_rating	Yes	No	Target	Binary
Discount_	Default	No	Input	Interval
Fetched_At	Default	No	Rejected	Nominal
Format_The_Release_Date	Default	No	Rejected	Nominal
Game_Age_Y_Month	Default	No	Rejected	Interval
Game_Name	Default	No	Rejected	Nominal
Linux	Default	No	Rejected	Interval
MacOS	Default	No	Rejected	Interval
Original_Price	Default	No	Rejected	Interval
Plantforms	Default	No	Input	Interval
Release_Date	Default	No	Rejected	Nominal
Windows	Default	No	Rejected	Interval
dataobs		No	ID	Interval
game_year_group	Default	No	Input	Ordinal
log_price	Default	No	Input	Interval
log_reviews	Default	No	Input	Interval



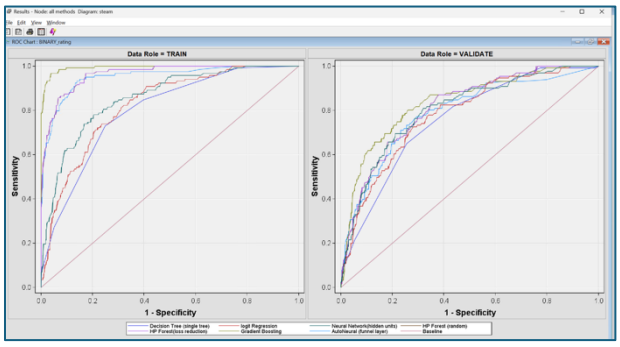
We have selected the same target and input variable, training and validation data set as the logit regression (because we set the random seed as 12345).

5.3.2 Model Comparison

The Process of model comparison are available in Appendix, Figure 5.3.2, and we will evaluate these models based on the subsequent dimensions: preliminary comparison.

a) Roc Curve Analysis

For training data set: Ensemble models (e.g., HP Forest(loss reduction), Gradient Boosting) show strong discrimination (curves close to top-left, AUC nearly 1). Traditional models (logit Regression, Decision Tree) follow, all outperforming Baseline.

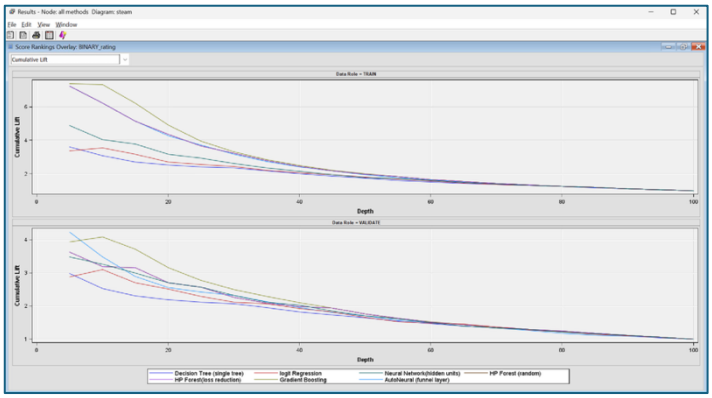


For validation data set: Ensemble models maintain leading performance, indicating robust generalization. Most models outperform random guessing, proving effectiveness.

b) Lift Curve Analysis

For training data set: Gradient Boosting and HP Forest(loss reduction) lead in Cumulative Lift. At 20% sample depth, their lift exceeds 6x (vs. random). Other models (Neural Network, logit Regression) follow, with Decision Tree lagging.

For validation data set: The top models retain high lift (over 4x at 20% depth), showing consistent generalization. Lift decreases as depth increases, aligning with typical trends.



Complete Comparison

The reslut of following table is based on the outputs from SAS miner.

Models	AUC (↑)	Precision (↑)	Recall (↑)	F1-Score (↑)	Misclassification Rate (↓)	Comprehensive Assessment
Gradient Boosting	0.84	0.595	0.336	0.44	0.13	Best
AutoNeural	0.79	0.5	0.374	0.5	0.15	Perfect
HP Forest (loss/random)	0.81	0.714	0.076	0.133	0.14	Ordinary
Neural Network	0.8	0.528	0.214	0.307	0.15	Ordinary
logit Regression	0.78	0.5	0.122	0.185	0.15	Ordinary
Decision Tree	0.76	0.667	0.076	0.133	0.14	Inappropriate

5.3.3 Discoveries & Limitation

Discoveries

a) Performance Overview

Based on the comprehensive evaluation metrics (AUC, Precision, Recall, F1-Score, and Misclassification Rate) presented in the chart, the models demonstrate significant performance variations. The Gradient Boosting model is identified as the top performer, labelled "Best" in the assessment. Other models, namely AutoNeural is categorized as "Perfect", while Neural Network, Logit Regression, and HP forest are deemed "Ordinary", Decision Tree is served as “Inappropriate” for this specific task due to notably poor performance in key areas, particularly Recall and F1-Score.

b) Practical Recommendations for Model Application

- Gradient Boosting: This is the primary recommended model. It is ideally suited for classification tasks requiring high prediction accuracy and reliability, especially with structured/tabular data. It is robust in handling complex, non-linear relationships and is less prone to overfitting compared to single decision trees.
- AutoNeural / HP Forest: These models serve as viable secondary options. They can be considered if there is a specific need to explore different model architectures or if Gradient Boosting does not integrate well into the existing production pipeline for technical reasons. Their "Ordinary" performance suggests they are functional but not optimal.

- Neural Network / Logit Regression / Decision Tree: The use of these models is not recommended for this task. Their extremely low Recall and F1-Scores suggest they are failing to identify a substantial portion of the positive class instances, rendering them ineffective for a balanced and reliable outcome.

Limitation

The project didn't perform sampling (like oversampling minority classes or undersampling majority classes) or adjust class weights before model training. This likely impacted performance, especially when our target variable is imbalance in the dataset. In future research, we need to address these gaps to make it better.

6. Conclusion

This project successfully addressed the multifaceted challenge of understanding game performance drivers on the Steam platform through comprehensive data analytics. By analyzing 1,745 Steam games across five key variables—discount rate, price, user reviews, platform support, and game age—we identified critical factors influencing game ratings and developed actionable insights for multiple stakeholders.

The core research question centered on predicting whether games achieve high ratings (greater than 7 on an 8-point scale) and identifying the variables that influence this outcome. Our analysis revealed that user engagement (review volume) emerges as the dominant predictor of high ratings, with a 93.9% increase in odds of achieving high ratings for each unit increase in log-transformed review count. This finding underscores the critical importance of post-launch player interaction and community engagement for game success on the Steam platform.

6.1 Primary Analytical Findings

User Engagement (Log_Reviews) serves as the strongest predictor, demonstrating that games accumulating substantial player reviews are substantially more likely to achieve high ratings. This suggests a virtuous cycle where popular games generate more reviews, which in turn attract additional players and foster positive feedback.

Pricing Strategy presents a nuanced but consequential relationship with ratings. Games with higher prices are associated with a 48.6% reduction in odds of achieving high ratings, suggesting that players may hold premium-priced games to stricter quality standards. This finding challenges the common assumption that higher prices signal higher quality and instead indicates that pricing decisions must align with demonstrated value delivery.

Discount Depth exhibits a negative association with high ratings, with each 1% increase in discount rate reducing the odds of high ratings by 4.9%. This statistically significant relationship suggests that aggressive discounting may signal lower quality to consumers, potentially dampening their post-purchase satisfaction. The optimal strategy balances promotional incentives with brand value preservation.

Platform Accessibility positively influences rating outcomes, with multi-platform support increasing the odds of high ratings by 42.8% compared to single-platform games. This finding reflects market demand for cross-platform compatibility and indicates that platform diversity enhances user satisfaction and broader market accessibility.

Notably, game age proved statistically insignificant ($p = 0.243$), demonstrating that established games do not inherently retain higher ratings than newer releases once other factors are controlled. Quality and engagement emerge as the primary determinants of rating success, transcending temporal considerations.

6.2 Implications for Strategic Decision-Making

For Game Developers: The dominance of user engagement as a rating predictor necessitates strategic post-launch investment in community building, content updates, and feedback mechanisms. Rather than relying on deep discounting, developers should emphasize value delivery and user satisfaction, enabling organic review generation and positive word-of-mouth. Multi-platform development, though potentially resource-intensive, represents an investment in rating potential and market reach.

For Publishers: Marketing resource allocation should prioritize games with accumulating review volumes, as these represent genuine engagement signals more reliable than sales figures alone. Discount strategies should incorporate timing considerations, balancing short-term sales stimulation with long-term brand equity protection. Portfolio composition should increasingly emphasize multi-platform titles to enhance average rating performance.

For Platform Operators: Steam's recommendation algorithms should integrate review volume, platform support, and discount history as transparency signals, rather than treating discounts as simple demand stimulators. The 14.5% high-rating concentration suggests opportunity for curator-driven discovery mechanisms that surface quality indie titles currently obscured by volume-driven visibility models.

6.3 Project Limitations and Future Research Directions

This analysis encountered several methodological constraints that merit acknowledgment. First, the severe class imbalance—with only 14.5% of games classified as high-rated—limited model sensitivity despite achieving high overall accuracy. Future studies should employ class-balancing techniques including oversampling minority classes, undersampling majority classes, or adjusting class weights during model training to achieve more balanced classification performance.

Second, the dataset's temporal snapshot nature precludes examination of dynamic pricing strategies, seasonal promotion patterns, and longitudinal rating trajectories. Real-world pricing decisions involve sophisticated timing considerations, competitive positioning, and demand forecasting that static cross-sectional analysis cannot fully capture.

Third, the analysis omitted potentially relevant contextual variables including game genre, developer history, marketing expenditure, and competitive landscape characteristics. Inclusion of these factors in expanded datasets would enable more comprehensive understanding of Steam ecosystem dynamics.

6.4 Conclusion

This project successfully demonstrated the application of business analytics to real-world e-commerce decision-making in the digital gaming sector. The findings highlight that game rating success fundamentally depends on genuine player engagement rather than pricing manipulations, and that platform accessibility represents a sustainable competitive advantage. The translation of statistical findings into managerial recommendations across developer, publisher, and platform operator audiences exemplifies the bridge between analytical rigor and organizational impact. Organizations that align pricing strategies with value delivery, prioritize post-launch engagement, and invest in platform diversity will position themselves for sustainable competitive advantage in the increasingly competitive digital gaming marketplace.

Future research should address the identified limitations through enhanced datasets, longitudinal designs, and integrated quantitative-qualitative approaches. The analytical framework established in this project provides a scalable foundation for ongoing optimization of game publishing and platform management strategies on Steam and comparable digital distribution platforms.

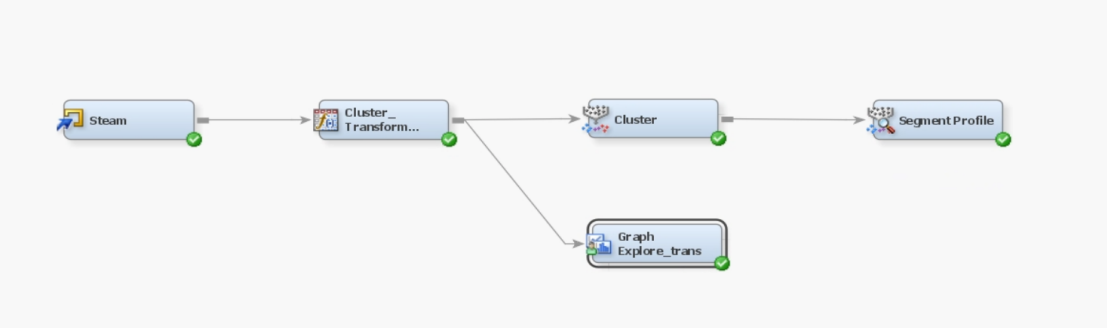
Reference

Kaggle : Steam Sales Historical Dataset, https://www.kaggle.com/code/mahmoudredagamail/steam-sales-historical-dataset/notebook

Appendix

1. Figure 3.4.1

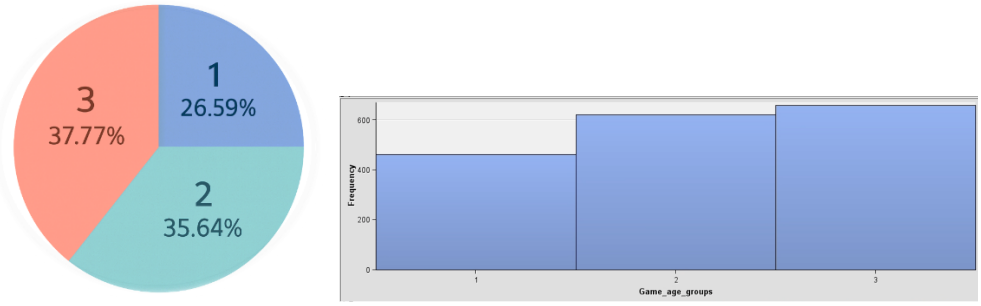
2. Figure 5.1.1



3. Figure 5.1.1_1

Game_Age	Count	Percentage
0	464	26.6%
1	273	15.6%
2	211	12.1%
3	138	7.9%
4	134	7.7%
5	110	6.3%
6	87	5.0%
7	80	4.6%
8	60	3.4%
9	58	3.3%
10	38	2.2%
11	26	1.5%
12	25	1.4%
13	8	0.5%
14	10	0.6%
15	6	0.3%
16	6	0.3%
17	7	0.4%
18	4	0.2%
	1745	

Figure 5.2.1_1 Game_Age_Year statistical table



Game_Age_groups

4. Figure 5.1.1_2

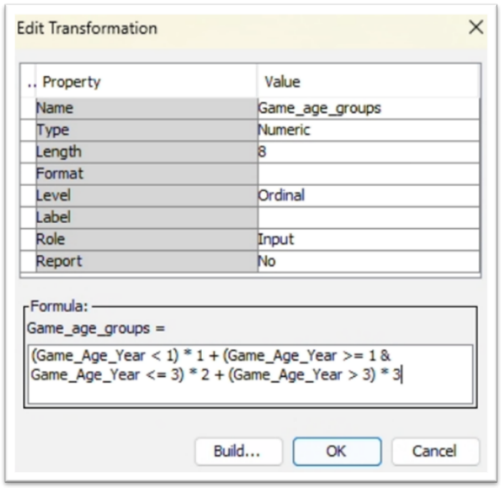
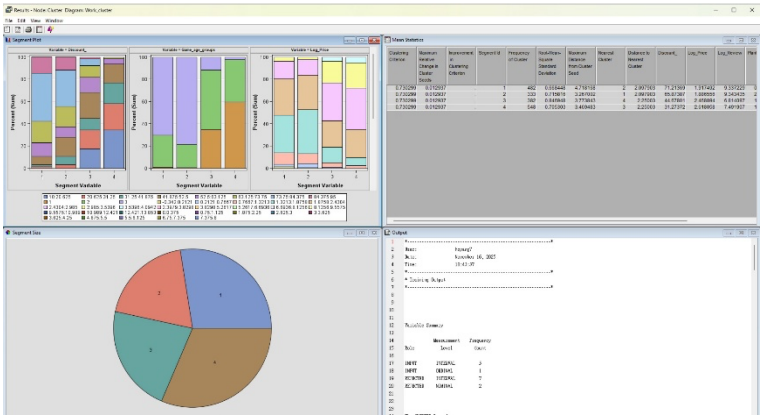
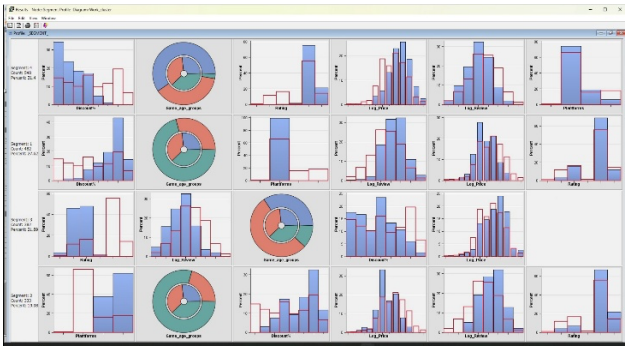


Figure 5.2.1_2 Transformation formula

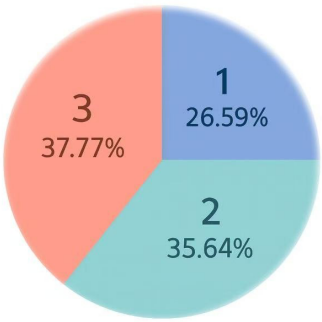
5. Figure 5.1.2



6. Figure 5.1.3



Segment profile



Game age group (1 new / 2 mature / 3 old)

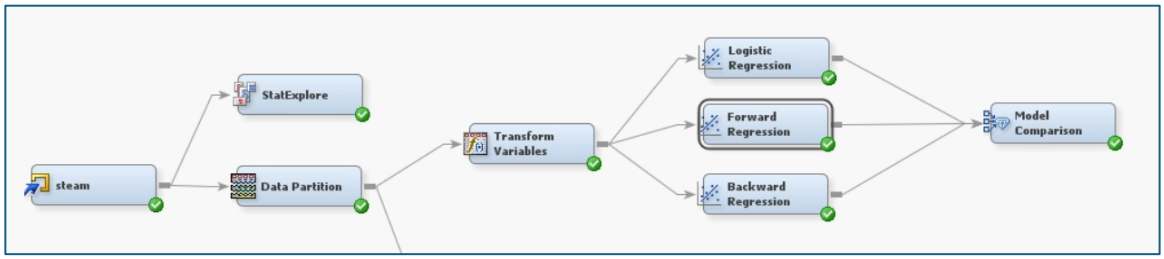
Cluster analysis:

Mean Statistics

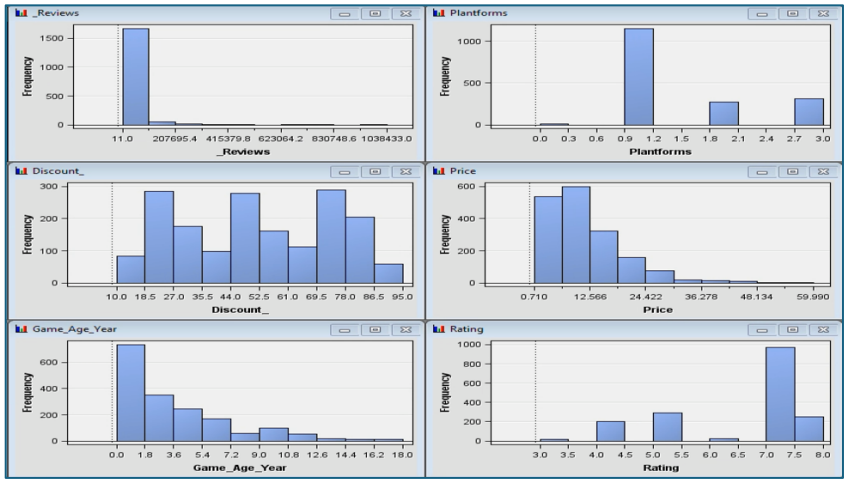
Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Discount_	Log_Price	Log_Review	Plantforms	Rating	Transform: Game_age_ groups
0.730299	0.012937	.	1	482	0.668448	4.718168	2	2.097903	71.21369	1.917492	9.337229	0.993776	6.707469	0.699696
0.730299	0.012937	.	2	333	0.715816	3.267002	1	2.097903	65.87387	1.886556	9.343435	2.621622	6.954955	0.731046
0.730299	0.012937	.	3	382	0.846948	3.773843	4	2.25003	44.67801	2.468094	6.814097	1.458115	4.486911	0.379838
0.730299	0.012937	.	4	548	0.705303	3.469483	3	2.25003	31.27372	2.618058	7.491907	1.321168	7.177007	0.267719

Cluster mean statistics

7. Figure 5.2.1



8. Figure 5.2.1_(2)_1

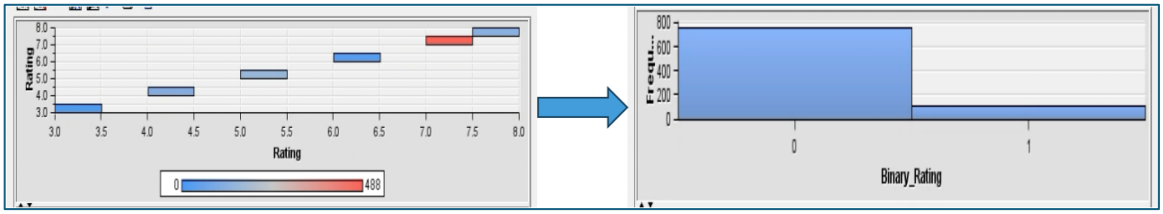


9. Figure 5.2.1_(2)_2

Formula Transformations
(maximum 500 observations printed)

Variable Name	Role	Measurement Level	Formula
Binary_Rating	TARGET	BINARY	Rating > 7

10. Figure 5.2.1_(2)_3



11. Figure 5.2.1_(4)

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
691.574	563.402	128.1721	5	<.0001

Full Logistic

The selected model is the model trained in the last step (Step 4). It consists of the following effects:

Intercept Discount_ Log_Price Log_Reviews Plantforms

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
691.574	564.178	127.3960	4	<.0001

Forward Logistic

The selected model is the model trained in the last step (Step 1). It consists of the following effects:

Intercept Discount_ Log_Price Log_Reviews Plantforms

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
691.574	564.178	127.3960	4	<.0001

Backward Logistic

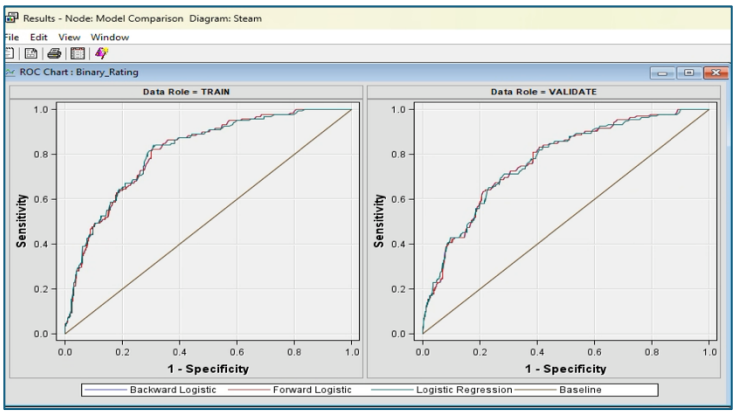
12. Figure 5.2.1_(5)_1

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg2	Forward Logistic	0.14450	0.099439	0.14089	0.11189
	Reg3	Backward Logistic	0.14450	0.099439	0.14089	0.11189
	Reg	Logistic Regression	0.14564	0.098952	0.14204	0.11150

13. Figure 5.2.1_(5)_2

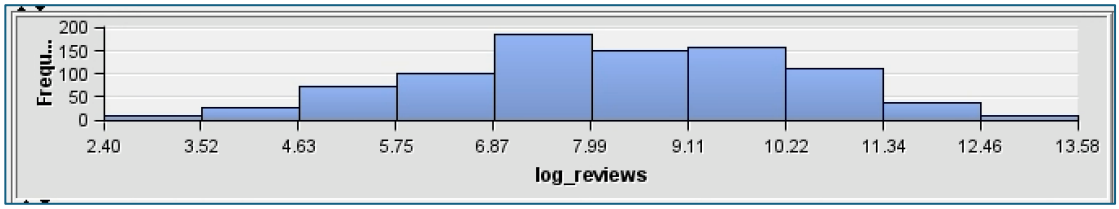


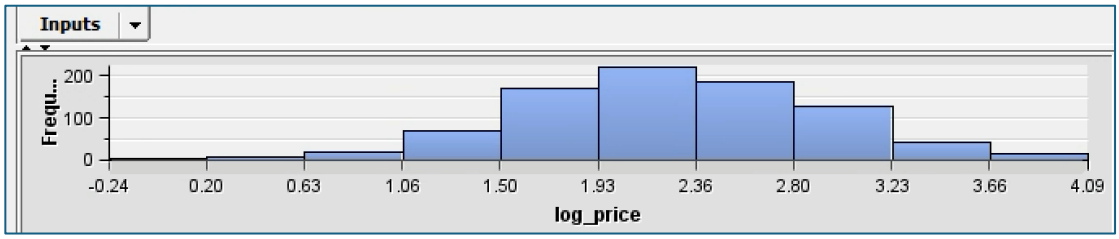
14. Figure 5.2.2

Analysis of Maximum Likelihood Estimates

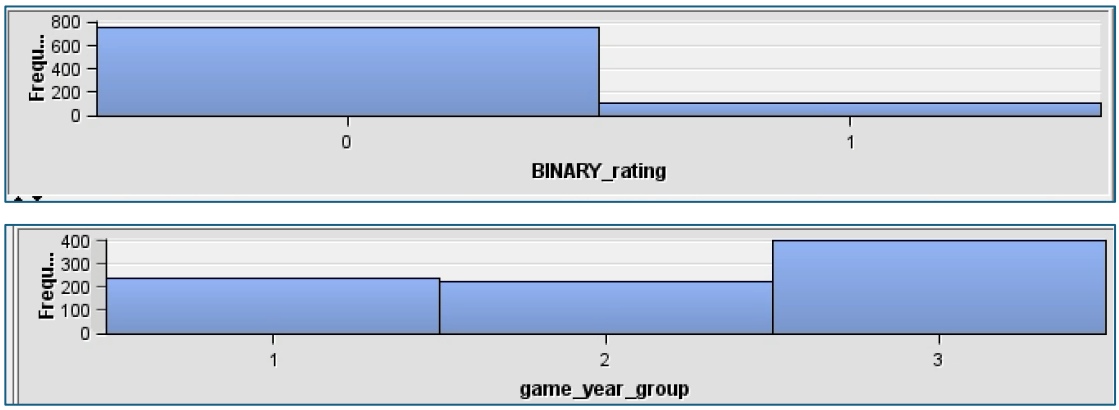
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-4.2390	0.7348	33.28	<.0001		0.014
Discount_	1	-0.0503	0.00671	56.28	<.0001	-0.6427	0.951
Log_Price	1	-0.6651	0.1847	12.96	0.0003	-0.2519	0.514
Log_Reviews	1	0.6624	0.0730	82.26	<.0001	0.7491	1.939
Plantforms	1	0.3565	0.1297	7.56	0.0060	0.1547	1.428

15. Figure 5.3.1_1

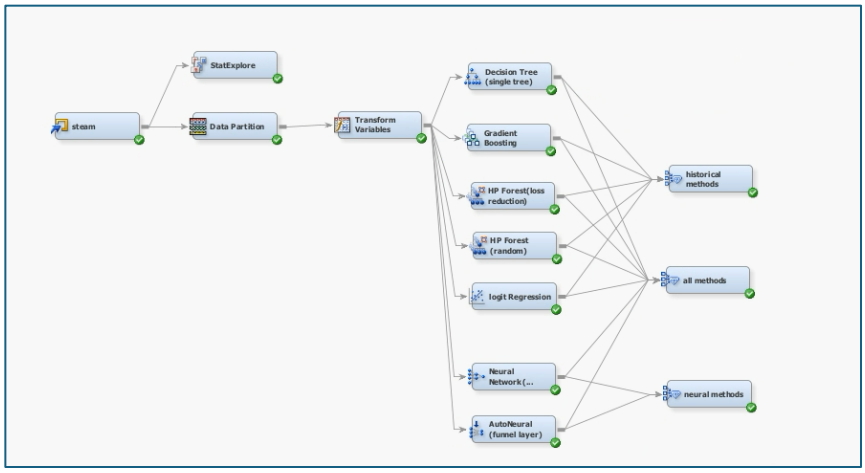




16. Figure 5.3.1_2



17. Figure 5.3.2



18. Machine Learning

a) Definition:

- AUC: Area under the ROC curve; quantifies a model's ability to distinguish between positive and negative classes. Higher (closer to 1) = stronger capability to separate positive and negative samples, avoiding misclassification.
- Precision: Proportion of predicted positive samples that are actually positive (TP/(TP+FP)). Higher = fewer false positive predictions, meaning the model's positive predictions are more

reliable.

- Recall: Proportion of actual positive samples correctly identified (TP/(TP+FN)). Higher = fewer true positive samples are missed, enhancing the model's ability to capture positive cases.
- F1-Score: Harmonic mean of precision and recall; balances the two metrics. Higher = better trade-off between minimizing false positives (precision) and missing true positives (recall).
- Misclassification Rate: Percentage of total samples incorrectly classified. Higher = better trade-off between minimizing false positives (precision) and missing true positives (recall). Lower = fewer overall classification errors, indicating more accurate general performance.

b) SAS Miner outputs:

Confusion Metrix

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree3	Decision Tree (single tree)	TRAIN	BINARY_rating		110	753	2	8
Tree3	Decision Tree (single tree)	VALIDATE	BINARY_rating		121	736	5	10
Boost	Gradient Boosting	TRAIN	BINARY_rating		34	755	.	84
Boost	Gradient Boosting	VALIDATE	BINARY_rating		87	711	30	44
HPDMForest	HP Forest(loss reduction)	TRAIN	BINARY_rating		94	755	.	24
HPDMForest	HP Forest(loss reduction)	VALIDATE	BINARY_rating		121	737	4	10
HPDMForest2	HP Forest (random)	TRAIN	BINARY_rating		94	755	.	24
HPDMForest2	HP Forest (random)	VALIDATE	BINARY_rating		121	737	4	10
Reg	logit Regression	TRAIN	BINARY_rating		104	737	18	14
Reg	logit Regression	VALIDATE	BINARY_rating		115	725	16	16
Neural	Neural Network(hidden units)	TRAIN	BINARY_rating		86	738	17	32
Neural	Neural Network(hidden units)	VALIDATE	BINARY_rating		103	716	25	28
AutoNeural2	AutoNeural (funnel layer)	TRAIN	BINARY_rating		49	746	9	69
AutoNeural2	AutoNeural (funnel layer)	VALIDATE	BINARY_rating		82	692	49	49

Misclassification:

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Boost	Gradient Boosting	0.13417	0.031464	0.03895	0.09954
	HPDMForest	HP Forest(loss reduction)	0.14335	0.067833	0.10767	0.10662
	HPDMForest2	HP Forest (random)	0.14335	0.067833	0.10767	0.10662
	Tree3	Decision Tree (single tree)	0.14450	0.097664	0.12829	0.11278
	Neural	Neural Network(hidden units)	0.14679	0.086586	0.11798	0.10692
	Reg	logit Regression	0.15023	0.097972	0.13975	0.11137
	AutoNeural2	AutoNeural (funnel layer)	0.15023	0.051147	0.06644	0.11858

Statistics:

Data Role=Valid

Statistics	Boost	HPDMForest	HPDMForest2	Tree3	Neural	Reg	Auto Neural2
Valid: Kolmogorov-Smirnov Statistic	0.54	0.49	0.49	0.39	0.49	0.44	0.48
Valid: Average Squared Error	0.10	0.11	0.11	0.11	0.11	0.11	0.12
Valid: Roc Index	0.84	0.81	0.81	0.76	0.80	0.78	0.79
Valid: Average Error Function	.	.	.	.	0.35	0.36	0.45
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.08	0.12	0.12	0.19	0.10	0.14	0.10
Valid: Cumulative Percent Captured Response	41.22	32.06	32.06	25.39	32.82	31.30	35.11
Valid: Percent Captured Response	21.37	13.74	13.74	10.37	15.27	16.79	13.74
Valid: Frequency of Classified Cases	.	872.00	872.00	.	.	.	.
Valid: Divisor for VASE	1744.00	1744.00	1744.00	1744.00	1744.00	1744.00	1744.00
Valid: Error Function	.	.	.	.	615.73	630.27	785.53
Valid: Gain	308.47	217.70	217.70	151.56	225.26	210.13	247.95
Valid: Gini Coefficient	0.67	0.62	0.62	0.51	0.60	0.56	0.58
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.53	0.49	0.49	0.39	0.48	0.44	0.47
Valid: Kolmogorov-Smirnov Probability Cutoff	0.07	0.12	0.12	0.11	0.10	0.13	0.07
Valid: Cumulative Lift	4.08	3.18	3.18	2.52	3.25	3.10	3.48
Valid: Lift	4.24	2.72	2.72	2.06	3.03	3.33	2.72
Valid: Maximum Absolute Error	0.99	0.96	0.96	1.00	0.99	0.98	1.00
Valid: Misclassification Rate	0.13	0.14	0.14	0.14	0.15	0.15	0.15
Valid: Mean Square Error	.	.	.	.	0.11	0.11	0.12
Valid: Sum of Frequencies	872.00	872.00	872.00	872.00	872.00	872.00	872.00
Valid: Root Average Squared Error	0.32	0.33	0.33	0.34	0.33	0.33	0.34
Valid: Cumulative Percent Response	61.36	47.73	47.73	37.79	48.86	46.59	52.27
Valid: Percent Response	63.64	40.91	40.91	30.89	45.45	50.00	40.91
Valid: Root Mean Square Error	.	.	.	.	0.33	0.33	0.34
Valid: Sum of Squared Errors	173.60	185.95	185.95	196.68	186.47	194.23	206.81
Valid: Sum of Case Weights Times Freq	1744.00	.	.	.	1744.00	1744.00	1744.00
Valid: Number of Wrong Classifications	.	125.00	125.00	.	128.00	.	131.00