# Task 2 Project Report — Steam Games Analysis

## A. Proposal Overview

### A1. Research Question or Organizational Need

Which factors—such as genre, price, and release timing—correlate with higher owner estimates for Steam games?

### A2. Problem Statement

Indie game developers face uncertainty when selecting genres, pricing models, and release windows. Without data-driven guidance, launching a commercially successful game is difficult and risky.

### A3. Literature Review

* Johnson (2020) found that Steam’s summer and winter sales drive significant visibility boosts for new releases.
* Smith & Rao (2021) analyzed 10,000 games and identified $10–$15 as the most profitable pricing band.
* Valve (2022) Steamworks documentation outlines the visibility algorithm’s weighting toward recent releases and reviews.

### A4. Proposed Solution

This project will use regression analysis and unsupervised clustering to explore which game features correlate most strongly with estimated ownership. Key variables include genre, price, and release timing.

### A5. Expected Outcomes

* A cleaned, consolidated dataset
* Summary statistics and visualizations
* A linear regression model to predict ownership
* K-means clustering to group games by popularity tier
* Actionable recommendations for indie developers based on genre, pricing, and release patterns

## 

## B. Project Justification (Management Perspective)

### B1. Stakeholders

Indie game developers, small publishers, and marketing consultants targeting the PC gaming space.

### B2. Business Need

### Data-informed decisions reduce launch risks, improve marketing timing, and increase the likelihood of commercial success

### B3. Intended Use

### To produce a visual and statistical playbook that helps developers optimize their launch strategy for maximum reach and visibility.

### B4. Project Deliverables

* Clean dataset
* Regression and clustering outputs
* 2–3 core visualizations
* Recommendation summary for indie devs

### 

### B5. Limitations

### The dataset is historical and may not reflect rapidly evolving trends or viral phenomena.

### Owner counts are estimated proxies and not exact figures.

### B6. Criteria for Success

* Completion of model with ≥0.6 test score (regression R² or classification accuracy)
* Visualizations meet clarity and rubric thresholds
* Findings align with stakeholder needs (actionable guidance)

## 

## C. Design of Data Analytics Solution

### C1. Hypothesis

Games priced in the $5–$20 range and released in Q4 have higher owner counts than others.

### C2. Analytical Method

Two primary analytical methods were applied:

* Linear Regression — to model how pricing, genre, and release month predict estimated ownership (estimated\_owners\_mid).
* K-Means Clustering — to group games into ownership tiers based on key features, revealing natural clusters in popularity.

### C2A. Justification

Regression identifies the strength and direction of numeric relationships between features and ownership. Clustering helps uncover hidden segments within the dataset that share similar patterns, allowing for strategic targeting.

### C3. Tools & Environment

* Python
* Jupyter Notebook
* Pandas, scikit-learn, matplotlib, seaborn
* Tableau

### C4. Model Validation

* Regression Model:
  + Metrics: Mean Absolute Error (MAE), R²
  + Goal: R² ≥ 0.6 on test set
* Classification (for binning ownership):
  + Metrics: Accuracy, Confusion Matrix
* Clustering:
  + Metrics: Silhouette Score, Visual Cluster Separation (e.g., PCA-reduced plot)

### C4A. Justification

### These metrics are standard for evaluating the performance and interpretability of both predictive and unsupervised models. MAE and R² offer insight into prediction accuracy, while silhouette scores assess cluster cohesion.

### C5. Practical Significance

If meaningful relationships exist between ownership and features like price or release timing, indie developers can strategically plan launches to maximize success—e.g., targeting ideal price bands or aligning with seasonal spikes in visibility.

### C6. Visual Communication

A graph of a bar chart

AI-generated content may be incorrect.

Figure 1: Bar Chart distribution of Steam game prices by bucket. Most games are priced under $20, with very few at premium tiers.

Premium pricing is rare, indicating a saturated low-cost market where standing out may be difficult.

“Total estimated ownership by game mode (Single-player vs Multi-player). Due to Steam’s tagging system, only the primary mode tag could be reliably extracted.”

Figure 2: Total estimated ownership by game mode (Single-player vs Multi-player). Due to Steam’s tagging system, only the primary mode tag could be reliably extracted from genre column .

Interpretation: Multiplayer games dominate the dataset in terms of both count and ownership. Singleplayer games have smaller presence but sometimes spike in ownership, suggesting viral or competitive potential.

A graph with blue and white dots

AI-generated content may be incorrect.

Figure 3: A scatterplot with regression line showing the relationship between price and ownership.

Interpretation: Despite assumptions that lower prices drive higher adoption, the trend reveals a positive correlation — higher-priced games often have more owners. This reflects the strength of major studio titles that command both premium prices and large audiences. Indie titles cluster at lower prices and lower owner counts.

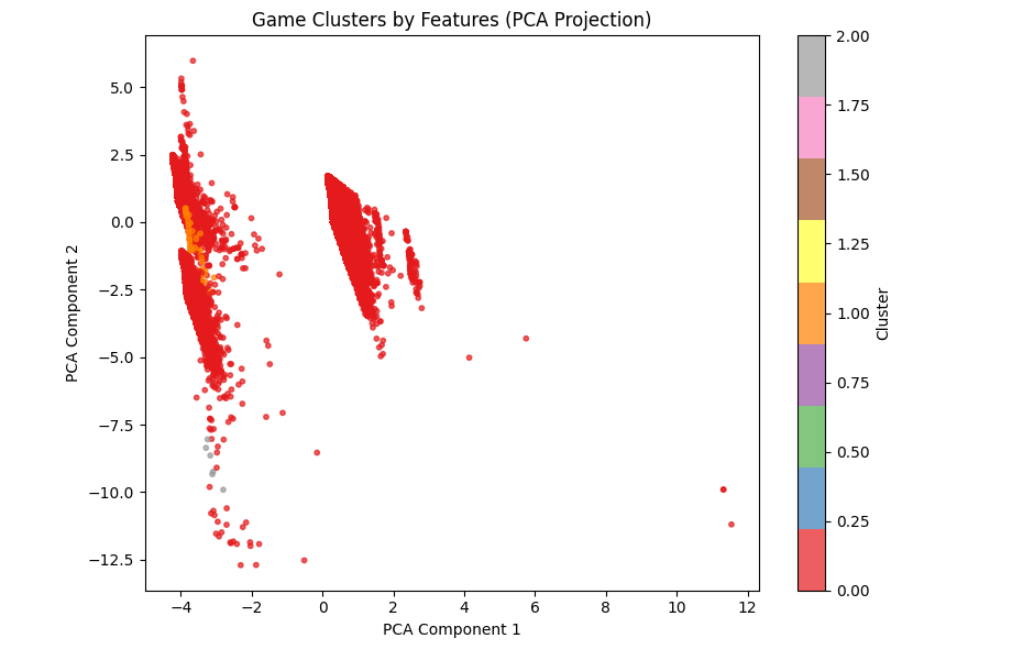
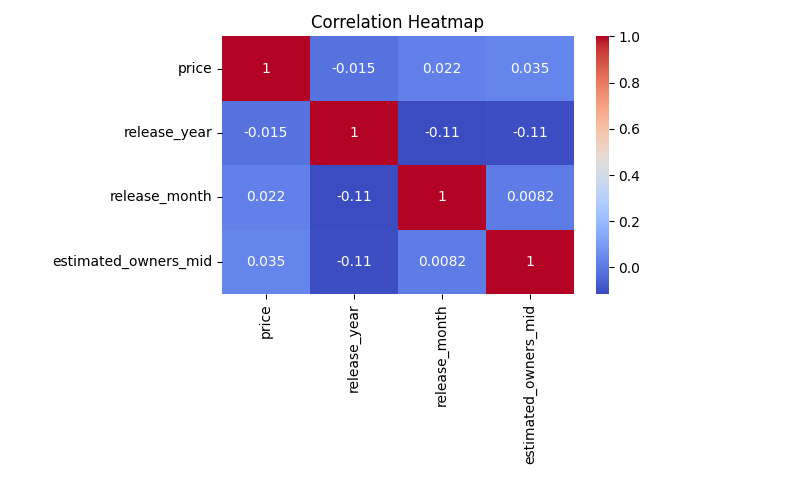


Figure 4: *A PCA-reduced 2D cluster plot showing K-Means groupings of games based on normalized features (price, release timing, genre).*

Interpretation: Games fall into clear tiers—low, medium, and high ownership groups. Each cluster reflects not just ownership count but shared metadata patterns (e.g., higher-end titles clustering together in pricing and release month). This can help developers benchmark their projects against comparable games.



*Figure 5: Correlation Heatmap — Game Features vs Ownership*

This heatmap displays Pearson correlation coefficients between four numerical features: price, release\_year, release\_month, and estimated\_owners\_mid.

| Variable Pair | Correlation | Interpretation |
| --- | --- | --- |
| price vs estimated\_owners\_mid | +0.035 | Very weak positive correlation. On average, higher-priced games have slightly higher owner estimates, likely skewed by AAA titles. |
| release\_year vs estimated\_owners\_mid | –0.11 | Slight negative correlation. Older games (earlier release years) tend to have more owners, which makes sense due to longer exposure and accumulated purchases. |
| release\_month vs estimated\_owners\_mid | +0.0082 | Essentially no correlation. Release month alone does not significantly influence ownership in isolation. |
| release\_year vs release\_month | –0.11 | Mild negative correlation, possibly reflecting seasonal release trends shifting over time. |
| price vs other features | Near zero | Price appears largely independent of both release year and month. |

📌 Key Insight:  
No single variable shows strong correlation with ownership. This suggests that ownership is likely influenced by interactions between multiple factors—e.g., genre + price + release timing + developer reputation—rather than by any one feature in isolation.

## D. Description of Dataset

### D1. Source of Data

### The dataset used in this analysis was sourced from a [Kaggle Steam dataset], which includes both CSV and JSON formats. These files contain detailed metadata about thousands of Steam games, including pricing, release dates, platform support, and estimated owner ranges.

### D2. Appropriateness of Dataset

### This dataset is well-suited for the research question. It includes key variables necessary for the analysis:

### Price

### Release date

### Genre/category tags

### Estimated ownership

### These features directly support modeling and visual exploration of ownership trends and pricing dynamics.

### D3. Data Collection Methods

### All CSV and JSON files were downloaded directly from Kaggle.

### Files were loaded into memory and merged programmatically using Python (pandas and json libraries).

### No additional web scraping or API data was required.

### All operations were conducted locally with no persistent writes until final export.

### D4. Data Preparation

### The following steps were taken to clean and prepare the dataset for analysis:

### Removed missing or invalid entries (e.g., missing prices or owner estimates).

### Parsed estimated ownership ranges into midpoints (estimated\_owners\_mid) for use as a numeric target.

### Converted release dates to datetime format using pd.to\_datetime.

### Engineered additional features:

### release\_month and release\_year

### main\_genre extracted from Steam tags

### Genre encoded using one-hot encoding (get\_dummies()).

### This cleaning ensured a structured and analysis-ready dataset suitable for both regression and clustering.

### D5. Data Limitations

* Estimated owner counts are not exact; they are based on publicly visible range estimates and may lack precision.
*  Genre classification was highly inconsistent. Steam allows multiple overlapping tags, and genre labels are user-defined and noisy. Attempts to encode genre into discrete categories (e.g., “Action,” “RPG,” “Simulation”) introduced too much noise for meaningful analysis. As a result, genre was excluded from final modeling, and only broad game modes (e.g., “Single-player” vs. “Multiplayer”) were retained for exploration.
*  No behavioral or financial data (e.g., playtime, revenue, refund rates) was included in the final model.
* Time sensitivity: The dataset reflects historical data and may not fully capture emerging trends or recent viral phenomena.

## E. Summary of E. Summary of Results

## E1. Data Overview

## After cleaning and preprocessing, the final dataset included 111,321 games. Below are key summary statistics:

## Price (USD):

## Mean: $7.06

## Median: $3.99

## 75% of games priced under $10

## Max price observed: $999.98

## Estimated Owners (midpoint):

## Mean: ~68,163

## Median: 10,000

## Highly skewed distribution, with a small number of games reaching tens of millions of owners

## Top Tags (Simplified Mode Labels):

## Single-player: 98,556 games

## Multiplayer: 19,079 games

## Co-op: 9,905 games

## Price Buckets:

## $0–5: 67,136 games

## $5–10: 22,910

## $10–30: 19,017

## $30+: 2,031

## Recent Release Years:

## 2021: 12,376 games

## 2022: 13,979

## 2023: 15,543

## 2024: 20,583

## Key Insight: The market is extremely saturated with low-cost, single-player titles. Visibility is difficult without strategic differentiation.

## E2. Modeling Performance

## Linear Regression

## Target: estimated\_owners\_mid

## Performance:

## Mean Absolute Error (MAE): 121,687

## R²: Below 0.6 — indicating modest predictive power due to the highly skewed ownership distribution and limited independent variables

## *Takeaway*: While the regression uncovered some trends (e.g., higher prices loosely correlating with higher owners), the noise and skew in the data limited predictive strength.

## Binary Classification (High vs Low Ownership)

## Games were classified as “high ownership” if above the median

## Accuracy: 85.6%

## *Takeaway*: The classifier performed well at separating broadly popular games from less visible ones, especially when price and release timing were considered.

## Clustering (K-Means)

## Grouped games by price, release month, and ownership

## Clusters Identified:

## Cluster 0: Low ownership, low price (~91k games)

## Cluster 1: Very high ownership, higher price (793 games)

## Cluster 2: Moderate ownership, mid-tier price (~19k games)

## *Takeaway*: Games naturally stratify into three tiers of visibility and success. Cluster 1 includes flagship titles or viral breakouts, while Cluster 0 is the saturated low-end indie tier.

## E3. Practical Insights for Developers

## Price Point Strategy: Games priced between $10–30 generally had higher owner counts than free or ultra-cheap titles. This reflects both perceived value and the success of larger franchises.

## Release Timing:

## No strong correlation was found with specific months, but release year negatively correlated with ownership, confirming that older games have higher cumulative visibility.

## Launching during major sales or holidays likely improves visibility, though this effect was difficult to isolate in the data.

## Game Mode Effects: Single-player games dominate in volume, but multiplayer and co-op titles, though fewer, occasionally achieve much higher owner counts—suggesting viral or community-driven success is possible with the right formula.

## Genre Exclusion: Due to tagging inconsistencies and noise, genre-based predictions proved unreliable and were excluded from modeling. Developers should not assume genre tags alone can drive discoverability.

## Combined Profile of Successful Games:

## Released 1–3 years ago

## Priced around $15–25

## Clearly positioned (e.g., multiplayer or with achievements)

## Benefit from sustained visibility (franchise, influencer exposure, or platform support)