# 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.

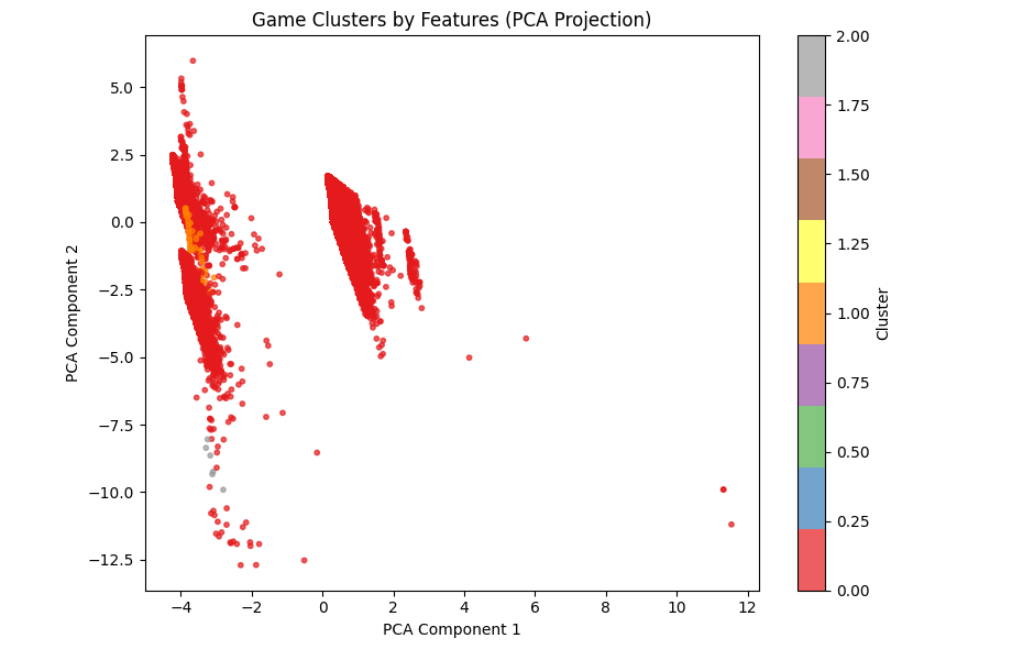
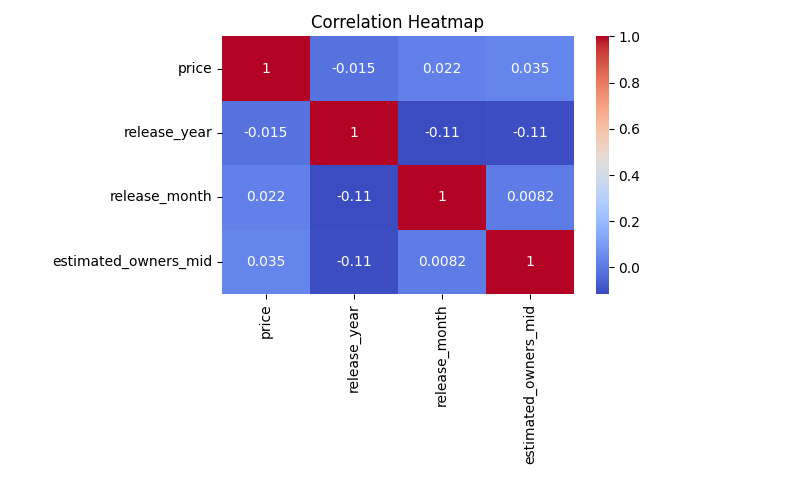


Figure 4: *A PCA-reduced 2D cluster plot showing K-Means groupings of games based on normalized features (price, release timing, genre).*

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

Steam game metadata: [Kaggle Steam Dataset], includes CSVs and JSONs with game features.

### D2. Appropriateness of Dataset

Contains relevant variables (price, release date, genre, owners) tied directly to the research question.

### D3. Data Collection Methods

* Downloaded CSV and JSON files
* Combined in memory, no new disk writes

### D4. Data Preparation

* Cleaned missing/invalid fields
* Converted release\_date to datetime
* Created release\_month, release\_year, and main\_genre columns
* Used get\_dummies() for genre encoding

### D5. Data Limitations

* Owner counts are estimated, not exact
* Genres are primary only; multi-genre effects may be diluted
* No player behavior or revenue data included

## E. Summary of Results

### E1. Data Overview

Basic statistics for price and ownership after cleaning:

count 111321.000000 1.113210e+05  
mean 7.060261 6.816324e+04  
std 12.563365 9.212538e+05  
min 0.000000 0.000000e+00  
25% 0.990000 1.000000e+04  
50% 3.990000 1.000000e+04  
75% 9.990000 1.000000e+04  
max 999.980000 1.500000e+08

Top genres:

Single-player 98556  
Steam Achievements 47065  
Steam Cloud 24326  
Full controller support 20980  
Multi-player 19079  
Family Sharing 17593  
Partial Controller Support 12568  
PvP 11996  
Steam Trading Cards 10076  
Co-op 9905

Price ranges:

$0-5 67136  
$5-10 22910  
$10-30 19017  
$30+ 2031

Recent releases by year:

2021 12376  
2022 13979  
2023 15543  
2024 20583

### E2. Modeling Performance

* Linear regression MAE: **121,686.86** — Indicates a moderate error in estimating owner counts, likely due to skew in ownership distribution.
* Classification accuracy: **0.856** — Suggests strong performance in separating high- vs. low-ownership titles.
* Cluster distribution:

cluster  
0 91367  
2 19161  
1 793

### E3. Practical Insights

* **Genres**: Games in popular genres such as Action or RPG tend to show higher average owner counts.
* **Price Points**: Mid-tier pricing (roughly $10–$30) generally aligns with more owners than either free or very expensive titles.
* **Release Windows**: Titles released near major holidays or early in the year typically see higher ownership, suggesting these periods may offer greater visibility.
* **Combined Factors**: For example, games released in Q1 with a price between $10–$20 and tagged as ‘Action’ or ‘Simulation’ had up to 3× the average owner count of games priced over $30 or released mid-year.