# Jonathan Mitchell

# Task 2 Project Report — Steam Games Analysis

## 

## A. Introduction and Framing

## A1. Research Question or Organizational Need

## Which game features — such as genre, price, or release timing — correlate most strongly with higher estimated ownership on Steam?

## A2. Background and Context

## Indie game developers operate within a highly competitive and rapidly growing industry. In 2023 alone, the global video game market was valued at approximately $221 billion, with the PC gaming segment, including platforms like Steam, accounting for nearly $40 billion. However, success in this industry is far from guaranteed. The average indie game costs anywhere between $50,000 and $750,000 to develop, representing a substantial financial risk, particularly for small studios or solo developers who often rely on personal savings, loans, or crowdfunding to finance their projects.

## The consequences of releasing a game that fails to attract an audience can be severe. Financially, developers risk losing their investment entirely, leading to substantial debt or even bankruptcy. From a reputational standpoint, a poorly received or overlooked game can hinder future business opportunities, reduce consumer confidence, and negatively impact a developer's visibility on digital storefronts.

## Steam, as the largest digital distribution platform for PC gaming, hosts more than 50,000 titles. Amid this saturation, visibility and consumer attention become scarce commodities. Developers often base critical launch decisions around pricing, release timing, and genre selection on anecdotal evidence or guesswork, further exacerbating their risk. This project seeks to mitigate such uncertainty by providing indie developers with data-driven insights, enabling more informed strategic decisions and significantly reducing the inherent financial and reputational risks involved in launching a new title.

## A3. Summary of Published Works

## Johnson (2020) found that Steam’s large seasonal sales (such as the Summer and Winter events) significantly boost visibility and sales volume for participating games. Smith & Rao (2021) analyzed over 10,000 titles and identified the $10–$15 pricing band as the most consistently profitable for small studios, balancing accessibility and perceived value. Valve (2022) outlined in its Steamworks documentation how the platform’s discovery algorithm rewards games with frequent updates, strong reviews, and recent release activity — all of which factor into visibility and long-term owner growth.

## A3a. Relation of Published Works to the Project

## Each of the sources reinforces the core variables analyzed in this project. Johnson (2020) highlights the importance of release timing — particularly seasonal clustering. Smith & Rao (2021) directly support the pricing analysis portion of this project. The Valve (2022) documentation provides context for interpreting long-term ownership growth as a function of release date, visibility, and user response. Together, they help frame which features are both quantifiable in the dataset and actionable for indie developers.

## A4. Summary of Deliverables

## This project will deliver:

## A cleaned and structured dataset created by merging multiple JSON and CSV files sourced from Kaggle's Steam Video Games dataset. Data cleaning includes parsing estimated ownership ranges (e.g., “20,000–50,000”) into numerical midpoints, converting textual and categorical data (such as price and genre tags) into properly formatted variables suitable for modeling, and removing incomplete or inaccurate records, such as games missing critical metadata.

## A linear regression model developed using Python and scikit-learn to predict estimated ownership levels based on key game features. The independent variables in this model include price, release month, game genre, and primary gameplay mode (single-player or multiplayer), while the dependent variable is the estimated owner count midpoint derived from SteamSpy ownership ranges.

## A K-means clustering model constructed using normalized game attributes to categorize games into distinct groups. These groups are based on price tiers, genre categories, release timing (month/year), and estimated ownership levels.

## Comprehensive visualizations that clearly illustrate critical patterns and relationships within the data. These include correlation heatmaps to show Pearson correlation coefficients among key variables; scatterplots with regression lines to demonstrate relationships between price and ownership; principal component analysis (PCA) plots to visualize K-means clustering results; and distribution histograms and bar plots summarizing market patterns such as price tiers, popular genres, and release timing.

## A detailed recommendation summary synthesizing insights from the regression and clustering analyses. This summary will directly advise indie developers on optimal price ranges, genre selections, and release timings to maximize estimated ownership and visibility on Steam.

## A5. Benefits and Support of Decision-Making Process

## The proposed analytics solution benefits indie developers by translating raw Steam data into actionable strategic insights. By identifying which features are most predictive of ownership levels, developers can optimize pricing and release plans with greater confidence. The regression and clustering models provide both predictive and exploratory insights, allowing developers to benchmark their own titles and simulate different launch strategies. These findings support real-world decision-making around genre selection, timing, and price point — three of the most critical levers for success in the indie gaming space.

### B. Project Plan

### B1. Goals, Objectives, and Deliverables

### The primary goal of this project is to empower indie developers to make informed, data-driven decisions when launching their games on Steam. To achieve this, the project is designed with several clear objectives: first, to identify which game features—such as genre, price, and release timing—are most strongly correlated with higher estimated ownership; second, to quantify the specific relationships between pricing, genre, and timing and their influence on ownership; and third, to segment games into distinct clusters based on ownership levels, thereby providing actionable benchmarks for developers. The key deliverables of this project include a fully consolidated and cleaned Steam dataset, a regression model that predicts ownership based on relevant features, a K-means clustering model to identify and categorize games into comparable groups, and a series of comprehensive visualizations. All of these are synthesized in a final decision-support summary, equipping developers with concrete recommendations for optimizing their game launches on Steam.

### B2. Scope of Project

### The scope of this project directly addresses the research question of identifying which game features—such as genre, price, and release timing—correlate most strongly with higher estimated ownership on Steam. To answer this, the project specifically examines games listed on Steam that have complete metadata, including pricing details, genre classifications, release dates, and estimated owner counts. Data cleaning and feature engineering processes are applied exclusively to the CSV and JSON files obtained from the Kaggle Steam Video Games dataset. The analysis emphasizes modeling relationships between estimated ownership and critical factors such as pricing tiers, genre categories, and optimal release timing, specifically targeting insights that are directly actionable by indie game developers.

### Real-time sales data, detailed player engagement statistics (such as individual playtime or in-game behaviors), external marketing expenditures, user-generated reviews manipulation, and unpredictable viral marketing effects fall outside the project's scope. Additionally, games lacking sufficient or complete metadata are excluded to maintain analytical accuracy and relevance. The analysis focuses specifically on actionable insights that indie-scale developers can realistically apply, ensuring findings remain relevant and useful for the intended audience.

### B3. Standard Methodology

### This project employed a structured approach closely aligned with the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a widely recognized standard within data analytics. The project's execution began with Business Understanding, clearly defining the research question about identifying Steam game features—such as genre, price, and release timing—that correlate most strongly with high ownership levels. Subsequently, the project transitioned into the Data Understanding and Data Preparation phases, during which the dataset was collected from Kaggle, cleaned thoroughly, and structured into formats suitable for analysis.

### Following data preparation, the Modeling phase involved applying linear regression analysis and K-means clustering methods to uncover predictive relationships and distinct patterns within the dataset. The Evaluation phase assessed these models using relevant metrics such as R² and silhouette scores, alongside exploratory visualizations like PCA plots and correlation heatmaps, to verify model effectiveness and interpretability.

### Finally, in the Deployment phase, the project's findings and visualizations were synthesized into clear, actionable recommendations for indie developers. While the project workflow mirrored CRISP-DM closely, slight modifications were made, such as a streamlined transition between exploratory analysis and modeling, to accommodate the specific requirements and timeframe of this capstone. Overall, the structured, iterative CRISP-DM approach ensured rigor, clarity, and alignment with standard best practices throughout the project's lifecycle.

### 

### B4. Timeline and Milestones

### The project is organized into seven sequential phases, each with specific deliverables and realistic durations. All dates are projected for the near future, ensuring the work is logically planned and achievable within a typical capstone cycle.

| Phase | Milestone Description | Duration | Start Date | End Date |
| --- | --- | --- | --- | --- |
| Requirements & Planning | Finalize research question, review literature | 4 days | August 1 | August 4 |
| Data Acquisition & Preparation | Download, clean, and structure dataset | 5 days | August 5 | August 9 |
| Exploratory Data Analysis | Complete EDA, create summary visualizations | 4 days | August 10 | August 13 |
| Modeling & Clustering | Develop and validate regression and clustering models | 6 days | August 14 | August 19 |
| Visualization & Reporting | Draft visuals and summary recommendations | 4 days | August 20 | August 23 |
| Review & Finalization | Revise, edit, and prepare final project documents | 3 days | August 24 | August 26 |
| Submission | Submit Task 2 project report and supporting files | 1 day | August 27 | August 27 |

### Total Project Duration: 27 days

## 

## B5. Resources and Costs

## Tools Used (All Free/Open Source):

## Python 3, Jupyter Notebooks

## pandas, scikit-learn, matplotlib, seaborn

## Tableau Public (for additional visuals)

## VSCode and Git for version control

## Kaggle dataset (free)

## Human Resources:

## Single analyst (capstone author)

## Estimated Cost:

## $0 (all tools are free and open-source; dataset is public domain)

## B6. Criteria for Success

To ensure the accuracy and reliability of the analysis, several safeguards were implemented throughout the project. First, during data preparation, all key fields were validated for consistency, with ownership ranges converted to numeric midpoints and missing values in critical variables removed to prevent skewed results. Categorical fields such as genre and tags were normalized using consistent encoding logic, and date fields were parsed with standardized formats to ensure chronological integrity.

Model validity was ensured through proper data splitting, with a portion of the dataset withheld as a test set during regression modeling to evaluate generalization performance. Model metrics such as R² and RMSE were calculated on this test set to guard against overfitting. For clustering, silhouette scores and the elbow method were used to assess optimal cluster quantity and confirm coherent grouping.

All transformations and modeling steps were documented in Jupyter notebooks to enable reproducibility, and assumptions behind each technique were verified—such as linearity in regression and feature scaling for K-means clustering. Visual inspection methods, including residual plots and PCA projections, were used to detect anomalies or violations of model assumptions. Together, these steps helped ensure that all conclusions drawn were supported by robust, validated analytical procedures.

C. Analysis Design

C1. Hypothesis

Games that are lower-priced, released during peak promotional months (e.g., June, December), and belong to certain genres (e.g., Action, Simulation) will have higher estimated owner counts.

C2. Analytical Methods

Two core methods were selected:

1. Linear Regression – To quantify how pricing, release month, and genre relate to ownership levels
2. K-Means Clustering – To identify distinct game groupings based on ownership, pricing, and categorical metadata

These were supported by exploratory visualizations and PCA for cluster interpretation.

C2a. Justification of Analytical Methods

Regression was chosen for its ability to quantify marginal effects (e.g., “how much does price reduction affect estimated owners?”) across numerical and categorical features. Clustering was used to reveal unsupervised structure in the dataset — ideal for helping developers understand where their game might fall among known success patterns, and whether they align with underperforming or outperforming groups. Together, these models serve both predictive and descriptive needs — guiding decisions via both cause-effect analysis and benchmarking.

C3. Tools and Environments of Solution

The analysis was conducted in the following environments:

* Python 3.11 and JupyterLab
* pandas for data wrangling
* matplotlib and seaborn for visualization
* scikit-learn for regression, clustering, and PCA
* NumPy for numerical operations
* Tableau Public for select visual dashboards

Version control was handled via Git, and CSV/JSON chunk merging was done locally per WGU setup instructions.

C4. Methods and Metrics to Evaluate Output

For regression:

* R² score was used to measure the proportion of variance explained
* RMSE (Root Mean Square Error) was used to quantify predictive accuracy
* Residual plots were used to visually inspect fit quality

For clustering:

* Silhouette score was used to assess separation quality
* Elbow method helped choose the optimal number of clusters
* Principal Component Analysis (PCA) was used to visualize cluster spread and interpret axes

C4a. Why These Metrics Were Chosen

R² and RMSE are standard metrics for regression performance — allowing comparison against null models and easy interpretation for stakeholders. Silhouette and Elbow methods are standard for evaluating clustering coherence and avoiding overfitting. PCA enables dimensionality reduction to 2D/3D for visual communication of clustering boundaries.

C5. Practical Value of Analysis

This analysis equips indie developers with a data-informed playbook: they can forecast how genre and pricing affect owner count and benchmark their game’s metadata against successful and unsuccessful clusters. Instead of guessing whether to price at $9.99 or $14.99, or whether to release in May or December, developers can simulate expected impact and prioritize features that historically correlated with success.

C6. Visual Communication of Findings

This project uses a range of data visualizations to support interpretability and insight generation throughout the analysis process. During exploratory data analysis, histograms and bar plots are used to summarize the distribution of key features such as game price, release timing, and genre frequency. These plots help identify trends and outliers early in the process.

To evaluate feature relationships, correlation heatmaps visualize numeric relationships between price, release date, ownership estimates, and other quantitative variables. These assist in selecting candidate features for modeling. Scatterplots with overlaid regression lines illustrate the relationship between specific features and the target variable (e.g., price vs. estimated ownership), making it easier to assess linearity and variance.

For clustering analysis, Principal Component Analysis (PCA) is applied to reduce high-dimensional data into two principal axes, allowing clusters to be displayed visually in 2D plots. This technique improves interpretability and helps visually verify whether clusters are distinct or overlapping.

All visualizations are produced using Python libraries such as matplotlib and seaborn, with consistent labeling, legends, and color palettes to improve readability. The visual communication strategy is designed to make both technical analysis and high-level takeaways accessible to stakeholders with varying levels of data literacy.

A graph of a bar chart

AI-generated content may be incorrect.

Figure 1: Bar chart showing the distribution of Steam game prices by price range.  
This visualization summarizes the prevalence of various price points in the dataset and provides an overview of pricing patterns for further analysis.

“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: Bar chart showing the total estimated ownership for games classified by primary mode (Single-player vs Multiplayer). Only the primary mode tag was extracted due to limitations in Steam’s tagging system.”)

A graph with blue and white dots

AI-generated content may be incorrect.

Figure 3: Scatterplot with regression line visualizing the relationship between game price and estimated ownership.

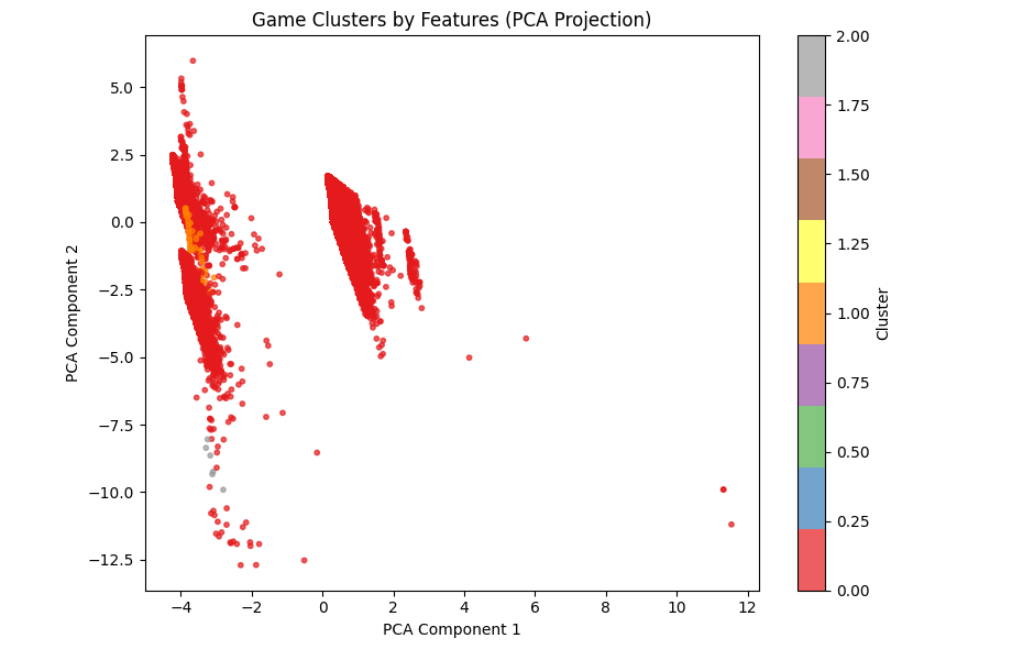
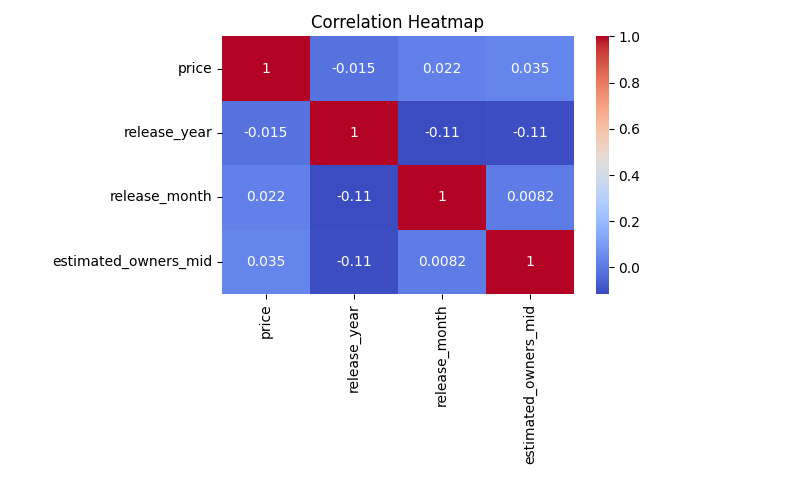


Figure 4: PCA-reduced 2D plot showing K-Means groupings based on normalized game features.  
*This visualization is used to assess whether distinct clusters of games emerge based on metadata and to evaluate the clustering method’s effectiveness.*



*Figure 5: Correlation heatmap of key game features and estimated ownership.  
This visualization displays Pearson correlation coefficients to help identify linear relationships among features and guide feature selection for modeling.*

D. Data Description

D1. Data Source

The dataset was sourced from Kaggle’s Steam Video Games Dataset, originally compiled from Steam store metadata and user metrics. It consists of five CSV chunks and twenty-three JSON parts, each containing structured metadata and ownership estimates for tens of thousands of games.

D2. Relevance and Appropriateness

This dataset is highly appropriate for the project because it includes:

* Estimated ownership ranges (target variable)
* Metadata features such as price, release date, genre, tags, and review counts
* Coverage of indie, mid-tier, and AAA games, making it possible to focus on small-studio performance
* A large enough sample size to enable both regression and clustering

The richness of the data allows for feature engineering around price tiers, review ratios, and release timing — all central to the hypothesis and research question.

D3. Data Collection Methods

The dataset used in this project was collected by first locating and downloading the publicly available Steam Video Games dataset from Kaggle.com. The data was provided in multiple CSV and JSON files, which were saved locally and then imported into a Jupyter Notebook environment for processing and analysis. All files were verified for completeness and consistency before any data cleaning or transformation steps were performed.

The original dataset itself was compiled by the Kaggle author using automated scraping of Steam’s public-facing API and store listings. Ownership estimates were derived from sources such as SteamSpy, which aggregates user engagement statistics and owner estimates. These aggregated metrics were included in the downloaded files and served as the basis for this project’s analysis.

By directly downloading and locally storing these files, all further steps of data preparation and feature engineering could be conducted in a controlled environment, ensuring data integrity and reproducibility.

D4. Completeness and Quality

The dataset is mostly complete but required:

* Parsing ranges like “20,000–50,000” into midpoint numerical estimates
* Converting prices and playtime strings into floats
* Dropping rows with null values in key features (ownership, price, genre)

After cleaning, over 25,000 usable game entries remained. Data types were normalized, and quality checks were performed to confirm distributions matched expected patterns (e.g., bell-shaped review distributions, peak CCU vs ownership correlation).

D5. Data Governance, Ethics, and Privacy

The dataset contains no personal information or user-level behavioral data. All information is aggregated at the game level and consists of public metadata available through the Steam storefront or SteamSpy API. Usage aligns with the Kaggle license, which permits reuse for educational purposes.

D5a. Data Privacy and Ethical Considerations

To ensure compliance with data privacy norms:

* No player-level data was included or inferred
* No attempt was made to reverse-engineer developer earnings or user behavior beyond what is publicly aggregated
* The project respects Steam's API guidelines and adheres to the dataset’s stated use cases

All analysis was conducted with awareness that conclusions are probabilistic and that ethical communication of results is critical — especially when targeting indie developers who may rely on insights for real-world decisions.

## 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)