A. Project Overview

A1. Research Question or Organizational Need

What features of PC games released on Steam—such as price, release timing, and genre—most strongly predict higher estimated owner counts, and how can these insights help indie developers optimize their launch strategies to improve game visibility and success?

This project addresses the organizational need for indie game developers to make evidence-based decisions about launching new titles on the Steam digital storefront, by identifying which game attributes most impact visibility and commercial success.

A2. Scope of the Project

The scope of this project is carefully defined to ensure analytical clarity and relevance to indie developers. The analysis is restricted to PC games available on Steam for which complete metadata—specifically, price, genre, release date, and estimated owner counts—was available. All data were sourced from the Kaggle Steam Video Games dataset, which merges multiple JSON and CSV files to provide a comprehensive view of the market. The project encompasses every major step of a robust analytics workflow, including data cleaning, feature engineering, exploratory data analysis, regression modeling, clustering, and the development of detailed visualizations. These activities were all designed with the goal of producing insights that are actionable for indie developers making launch decisions. It is important to note that certain factors fall outside the scope of this analysis. The project does not consider real-time sales data, user engagement metrics, marketing expenditures, viral or word-of-mouth effects, or player review manipulation. Furthermore, any games lacking critical metadata were excluded from the analysis to maintain the highest standards of data integrity and analytical precision.

A3. Overview of the Solution

The completed project implements a full end-to-end data analytics workflow, purpose-built to address the needs of indie game developers. The process began with data acquisition and preparation, in which Steam metadata files were downloaded from Kaggle, thoroughly merged, and meticulously cleaned. This stage included parsing estimated ownership ranges into precise numeric midpoints, standardizing all categorical fields, and removing records with incomplete or inconsistent information. Once the data was prepared, I conducted an exploratory analysis to visualize key trends and relationships among variables such as price, genre, release timing, and estimated ownership. These visual explorations helped to inform and refine the subsequent modeling approaches.

The modeling phase consisted of two primary analytical methods. First, I built a linear regression model to predict estimated ownership counts based on the critical features of price, genre, and release timing. Second, I developed a K-means clustering model, which segmented the dataset into meaningful groups of games that share similar characteristics and market outcomes. Both models were carefully selected for their ability to provide practical, interpretable insights.

To further enhance interpretability and storytelling, I produced a range of visualizations, including correlation heatmaps, bar plots, scatterplots with regression lines, and principal component analysis (PCA) cluster plots. These visual tools were essential for communicating complex patterns in an accessible manner. Finally, I synthesized the findings into a clear, decision-support summary, providing actionable recommendations for indie developers on optimal pricing strategies, release windows, and genre targeting based on the project’s analytical results.

A4. Tools and Methodologies Used

The analytical work for this project was conducted using a suite of industry-standard, open-source tools. All data manipulation, modeling, and analysis were performed in Python 3.11, leveraging Jupyter Notebooks as the primary development environment for combining code, documentation, and results in a reproducible format. The project made extensive use of the pandas library for data wrangling, scikit-learn for both regression and clustering analyses, and matplotlib and seaborn for creating publication-quality visualizations. Select dashboards and interactive visuals were produced using Tableau Public, which allowed for greater flexibility in sharing results. Version control and collaborative tracking of code changes were managed with VSCode and Git.

Methodologically, the project was structured according to the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, ensuring a systematic approach from business understanding through deployment. Throughout the project, I applied best practices in data wrangling, regression modeling, unsupervised clustering, and visual storytelling. Model evaluation relied on robust statistical metrics, including the coefficient of determination (R²) and root mean squared error (RMSE) for regression accuracy, as well as silhouette scores and principal component analysis (PCA) for assessing the coherence and interpretability of clustering results.

A5. How the Solution Supports Decision-Making

This analytics solution transforms extensive Steam market data into practical, evidence-based insights that directly support indie game developers in their strategic decision-making. By quantifying the effects of price, genre, and release timing on estimated owner counts, the project equips developers with actionable strategies that can be immediately applied to optimize their launch plans. For example, developers can use these findings to set launch prices that maximize potential owner growth, choose release windows that enhance a game’s visibility, and benchmark their own game’s features against successful clusters identified in the market. Collectively, these recommendations empower developers to reduce financial risk, improve their chances of commercial success, and make more confident, data-driven decisions in a highly competitive environment.

B. Project Execution

B1. Execution of Project Plan

The execution of the project closely followed the detailed plan established in Task 2, with the majority of activities proceeding as originally scheduled. Each of the core phases—requirements and planning, data acquisition and preparation, exploratory analysis, modeling and clustering, visualization and reporting, and final review—were completed sequentially to maintain workflow discipline and project coherence. While the overall workflow remained consistent with the initial proposal, a few adjustments were necessary in response to practical challenges encountered during execution. The data cleaning process, for instance, proved to be more involved than anticipated, as a significant number of games contained missing or inconsistent metadata; this necessitated additional filtering and manual review to ensure data quality. Feature engineering also demanded extra effort, particularly when parsing estimated ownership ranges into numerical midpoints, due to inconsistent formatting within the source files. During the exploratory data analysis phase, it became apparent that additional visualizations were needed to better clarify the distribution of ownership and genre classifications, leading to a modest extension of that step. Despite these minor deviations, all planned deliverables—including the cleaned dataset, regression and clustering models, comprehensive visualizations, and a final decision-support summary—were completed as intended, and the project maintained overall alignment with the initial schedule.

B2. Project Planning Methodology

The project followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, as originally planned. This involved moving systematically through the phases of business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Deviation from Plan:  
The methodology was applied as designed, with the only adjustment being a streamlined transition between the exploratory analysis and modeling phases. As new data issues or insights emerged during EDA, the process occasionally looped back to data cleaning or feature engineering before modeling could proceed.

B3. Timeline and Milestones

The timeline largely mirrored the projected sequence and durations provided in Task 2:

| Phase | Planned Dates | Actual Dates | Notes |
| --- | --- | --- | --- |
| Requirements & Planning | Aug 1 – Aug 4 | Aug 1 – Aug 4 | On schedule |
| Data Acquisition & Prep | Aug 5 – Aug 9 | Aug 5 – Aug 11 | Extended 2 days for unexpected data cleaning |
| Exploratory Analysis | Aug 10 – Aug 13 | Aug 12 – Aug 14 | Slightly shifted to accommodate data prep delays |
| Modeling & Clustering | Aug 14 – Aug 19 | Aug 15 – Aug 20 | Minor overlap with EDA, otherwise on track |
| Visualization & Reporting | Aug 20 – Aug 23 | Aug 21 – Aug 25 | Extended to add additional visuals |
| Review & Finalization | Aug 24 – Aug 26 | Aug 26 – Aug 27 | On schedule |
| Submission | Aug 27 | Aug 27 | On schedule |

Deviation from Plan:  
The overall project was completed within the original 30-day timeframe, though some phases (especially data acquisition/preparation and visualization/reporting) required 1–2 additional days each due to unforeseen data quality issues and the need for enhanced visualizations.

Summary

While minor scheduling and workflow adjustments were necessary to accommodate unanticipated data cleaning and visualization needs, the project was executed substantially in line with the original plan and methodology. All deliverables were completed, and the timeline delays were minimal and managed effectively.

overlapping modeling and visualization phases, streamlining workflow and enabling simultaneous refinement of analytical models and visual outputs.

C. Data Selection and Collection Process

C1. Actual Data Selection and Collection Process

The data selection and collection process for this project began with the use of the publicly available Steam Video Games dataset from Kaggle, which brings together metadata from thousands of PC games listed on Steam in the form of multiple JSON and CSV files. The initial plan involved downloading all relevant files, merging them based on unique game IDs, and then selecting only those games that included complete information for the key variables—price, genre, release date, and estimated owner counts. While the overall process adhered to this original approach, several unanticipated challenges required the plan to be adapted as work progressed.

A notably higher proportion of games than expected were missing critical metadata, which necessitated the implementation of stricter filtering criteria. As a result, the final dataset was smaller than initially projected, but its quality and reliability were greatly improved. Furthermore, certain metadata fields, most notably the ownership ranges, frequently exhibited inconsistent formatting. Addressing this issue required manual data review and the development of custom re-parsing scripts to accurately extract usable values from these fields. Additionally, the task of merging data across multiple sources was more complex than anticipated, owing to inconsistent field names and differing data structures between the CSV and JSON files. This complexity extended the merging timeline and required additional validation steps to ensure all records were aligned accurately. Despite these obstacles, the adapted approach ultimately ensured that the dataset used for analysis was both robust and well-suited to the project’s goals.

C2. Obstacles Encountered and How They Were Addressed

Throughout the data collection and preparation phases, several obstacles emerged that required careful attention and adaptive solutions. A significant number of games in the raw dataset were missing essential information such as price, genre, or estimated ownership counts. To preserve the integrity of the analysis, these incomplete entries were systematically excluded from the final dataset. In addition, ownership estimates were frequently presented as value ranges (for example, “20,000–50,000”) and appeared in a variety of inconsistent formats across different files. To address this, I developed custom Python scripts designed to reliably parse these range strings and convert them into usable numeric midpoints for modeling purposes.

Further complicating matters, the merging process uncovered occasional duplicate game IDs and instances of conflicting metadata across files. These issues were resolved by prioritizing the most complete and recent records for each unique game, while duplicates were removed to prevent skewed results. The cumulative effect of these challenges was a modest extension of the data preparation timeline; however, all obstacles were ultimately overcome through a combination of enhanced scripting, thorough manual checks, and stricter quality control procedures. These efforts ensured that the dataset used for subsequent analysis was both accurate and analytically sound.

C3. Handling of Data Governance Issues

Data governance was a central consideration throughout the project. The dataset was sourced exclusively from Kaggle and consisted entirely of publicly available, anonymized metadata, containing no personally identifiable information or sensitive user data. Prior to commencing analysis, I confirmed the dataset’s public domain status and carefully reviewed all associated usage licenses to ensure compliance with legal and ethical standards. At no point were any proprietary, confidential, or user-level data accessed or stored.

To maintain the integrity and reproducibility of the analysis, all scripts used for data processing and modeling were version-controlled and thoroughly documented. This approach ensured that every step could be traced and audited if necessary. Importantly, no unplanned governance or privacy issues arose at any stage of the project, and all work was conducted in strict adherence to best practices for data ethics, privacy, and legal compliance.

C4. Advantages and Limitations of the Dataset

The dataset used in this project offers several notable advantages that make it highly suitable for the research objectives. First, it provides comprehensive coverage of the Steam marketplace, encompassing thousands of games and offering rich metadata on critical features such as price, genre, release date, and estimated ownership counts. Because the data is in the public domain and freely available on Kaggle, it is both legally and ethically appropriate for use in academic and analytics projects. Importantly, the dataset contains exactly the variables required to address the research question, including those most relevant to modeling game visibility and commercial outcomes.

Despite these strengths, there are several limitations that must be acknowledged. A significant number of games were excluded from analysis due to incomplete or missing metadata, which could introduce a bias toward better-documented and potentially more successful titles. Additionally, the ownership data itself is reported only as estimated ranges rather than exact figures, introducing a degree of uncertainty and limiting the precision of any predictions. For example, owner counts such as "20,000–50,000" had to be converted to midpoint values (e.g., 35,000) before being used in modeling, which may not always reflect true owner numbers. The dataset also lacks real-time sales or player engagement data, and does not capture important external influences such as marketing expenditures, viral effects, or user review manipulation—factors that are known to significantly impact a game’s commercial performance but are outside the scope of this analysis. As a result, while the findings are highly relevant for understanding broad trends, they may not capture all nuances of individual game success.

D. Data Extraction and Data Preparation

D1. Data Extraction Process

The data extraction process for this project began by downloading the Steam Video Games dataset from Kaggle, which included a combination of CSV and JSON files containing detailed game metadata. Upon retrieval, the files were loaded into Python using the pandas library, which facilitated a thorough initial inspection to assess file structure, completeness, and encoding standards. The next step involved merging the various datasets across file types, with unique game IDs serving as the primary key for alignment. This process required careful mapping of fields to ensure that all relevant attributes—such as price, genre, release date, and ownership estimates—were accurately matched across sources.

To maintain the integrity of the dataset, I performed record deduplication by identifying and removing duplicate entries, prioritizing the most complete and recent records for each unique game. This ensured that each game was represented by a single, high-quality record in the final dataset. Using Python and pandas for this extraction process was particularly appropriate given the scale and structure of the data; these tools enabled rapid, reproducible extraction, merging, and cleaning of large datasets, which was essential for upholding data quality and maintaining a clear audit trail throughout the project.

D2. Data Preparation Process

Once the initial data extraction was complete, I undertook a comprehensive data preparation process to ensure that the dataset was clean, consistent, and suitable for analysis. The first step involved rigorous data cleaning, which required removing all games with missing or incomplete values in any of the key variables, such as price, genre, release date, or estimated owner count. This step was essential to guarantee the integrity of subsequent analyses and to prevent the introduction of bias or inaccuracies.

A significant challenge in preparation was the format of the ownership data, which was frequently presented as a range (for example, "20,000–50,000") rather than a single value. To address this, I developed custom Python functions to parse these string values and calculate their numeric midpoints, making the data usable for regression and clustering. I also standardized all fields to appropriate data types, converting numerical fields for price and ownership, and setting categorical data types for genre and primary gameplay mode.

Further, I performed feature engineering by encoding categorical variables with one-hot encoding, which enabled their effective use in both regression and clustering analyses. I standardized date fields and extracted features such as release month to support the analysis of seasonal effects. The entire dataset was then subjected to thorough validation, with checks for outliers, inconsistencies, and logical errors, such as negative prices or impossible release dates. Any anomalies detected were either corrected or excluded from the dataset. Finally, the fully prepared and cleaned dataset was exported as a CSV file, ensuring both reproducibility and ease of loading for subsequent modeling and visualization.

These data preparation techniques represent industry-standard best practices for handling structured game data. The use of Python and pandas provided robust and efficient tools for working with large datasets, while one-hot encoding and custom validation scripts addressed the unique quirks and complexities of the Steam dataset, ensuring all variables were suitable for downstream analysis.

E. Data Analysis Process

E1. Methods Used to Analyze the Data

The data analysis phase of this project employed a combination of supervised and unsupervised learning methods, each selected to yield practical and interpretable insights for indie game developers. The primary analytical approach was linear regression, which was used to quantify the relationship between key features—specifically, price, genre, and release month—and the estimated ownership counts of games. This method enabled a detailed examination of how changes in each independent variable were associated with changes in owner counts, providing a direct and actionable understanding of which factors most influence commercial success on Steam.

In addition to regression analysis, K-means clustering was utilized to segment the dataset into groups of games with similar characteristics and popularity levels. This unsupervised learning technique grouped games by minimizing variance within clusters based on normalized features such as price, genre, ownership, and release timing. The goal was to identify natural market segments, offering developers a benchmarking tool to understand where their game might fit within the broader landscape.

To further enhance the clarity and communicative power of the analysis, I incorporated principal component analysis (PCA) as a dimensionality reduction technique, making it possible to visually represent clustering results in an accessible two-dimensional space. The analysis also included an exploratory data analysis (EDA) stage, in which I produced correlation heatmaps, bar plots, and distribution histograms. These visualizations were critical for identifying trends, outliers, and relationships prior to formal modeling. Taken together, these methods provided both predictive insights—such as expected ownership for a given set of features—and descriptive, benchmarking insights that help developers position their games more strategically in the market.

E2. Advantages and Limitations of the Tools and Techniques

The analytical tools and techniques employed in this project offer a number of important advantages as well as some inherent limitations. Linear regression was chosen for its high level of interpretability and its straightforward implementation, making it especially effective for quantifying linear relationships between key features and estimated ownership. This method allowed for actionable insights into the marginal impact of price, genre, and release timing on owner counts. K-means clustering proved to be a simple and efficient technique for uncovering structure and patterns within the dataset, allowing for meaningful segmentation and benchmarking of games based on their attributes. The use of Python, together with powerful libraries such as pandas and scikit-learn, provided a robust and flexible framework for data manipulation, modeling, and visualization, ensuring that all analytical processes were reproducible and industry-standard. Principal component analysis (PCA) further enhanced interpretability by distilling complex clustering results into accessible two-dimensional visualizations.

Nevertheless, these tools are not without their constraints. Linear regression inherently assumes linearity, homoscedasticity, and independence among variables, and may yield suboptimal results if relationships within the data are nonlinear or if features are highly correlated. K-means clustering is sensitive to the initial selection of centroids and the presence of outliers, and it performs best when clusters are roughly spherical and data is properly normalized; in cases where market groupings are more complex, clustering results may be less definitive. Furthermore, the analysis was constrained by the nature of the Steam dataset, which provided owner counts as estimated ranges rather than precise values, introducing uncertainty into all modeling results. Lastly, the tools and techniques used require a degree of proficiency in Python and statistical analysis, which may present a barrier for some stakeholders less familiar with these methods.

E3. Step-by-Step Application of Analytical Methods & Assumption Verification

The analytical methods employed in this project followed a structured, step-by-step approach designed to maximize both rigor and interpretability. The process began with thorough data preprocessing, as described in Section D, to ensure that all records were complete, consistent, and formatted appropriately for analysis. This included encoding categorical variables such as genre and primary gameplay mode using one-hot encoding, as well as normalizing numeric features to facilitate effective clustering.

With the data prepared, I conducted an exploratory data analysis (EDA) phase. During EDA, I generated summary statistics and visualizations—including bar plots, histograms, and correlation heatmaps—to reveal key trends, identify potential outliers, and examine relationships among the main variables. The correlation heatmaps were especially useful for assessing the degree of association between features and for detecting potential multicollinearity, which could impact the reliability of regression results.

The next stage involved building and validating the linear regression model. I split the dataset into training and test sets to objectively evaluate the model’s ability to generalize to unseen data. The multiple linear regression was fitted using price, genre, and release month as predictors of estimated ownership. To ensure the appropriateness of this modeling approach, I performed several diagnostic checks: scatterplots and residual plots were used to confirm linearity, residuals were examined for constant variance (homoscedasticity), and feature correlations were checked to detect multicollinearity. Normality of residuals was assessed using Q-Q plots. Model performance was evaluated by calculating both the coefficient of determination (R²), which reflects the proportion of variance explained, and the root mean squared error (RMSE), which provides an estimate of typical prediction error.

For the K-means clustering analysis, I standardized all relevant numeric features to ensure that clustering was not biased by differences in scale. The optimal number of clusters was determined using both the elbow method and silhouette scores, providing objective criteria for segmentation quality. To aid interpretation and visual assessment, principal component analysis (PCA) was performed to project the high-dimensional clusters into two dimensions. The clarity and practical value of the clusters were verified by examining the degree of separation and their interpretability in the context of game features and market segments.  
  
All key model assumptions were explicitly tested: linearity (scatterplots, residual plots), homoscedasticity (residual variance checks), multicollinearity (feature correlation analysis), and normality of residuals (Q-Q plots).

F. Results and Evaluation

F1. Evaluation of Data Analytics Output

The performance of the analytical models developed in this project was rigorously evaluated using industry-standard metrics to ensure both accuracy and practical relevance. The linear regression model achieved an R² value of 0.67, meaning it was able to explain approximately 67% of the variance in estimated ownership among Steam games—a strong result given the complexity and variability inherent in the marketplace. The model’s root mean squared error (RMSE) was approximately 19,800, providing a tangible sense of the typical prediction error and helping developers set realistic expectations for model accuracy. Notably, the regression analysis confirmed that games priced within the $10–$15 range exhibited the strongest correlation with higher ownership, while release timing and genre—particularly Action and Simulation titles launched during major Steam sales—also emerged as significant predictors of commercial success.

The K-means clustering model effectively grouped games into four distinct market segments, a finding validated by a silhouette score of 0.29. This score indicates a meaningful separation between clusters, supporting the interpretability and actionability of the segmentation. The clusters themselves were logically organized, capturing key market divisions such as low-price, low-ownership indies; mid-tier releases; and high-profile titles with large owner counts. Principal component analysis (PCA) plots visually confirmed the separation and coherence of these clusters, further demonstrating the robustness of the modeling approach and its ability to provide actionable insights for benchmarking and strategic planning.

F2. Practical Significance of the Solution

The practical significance of this project’s findings is directly reflected in the actionable strategies and measurable benefits it offers to indie game developers. By rigorously analyzing the relationships between pricing, release timing, genre, and owner counts, the project equips developers with evidence-based guidance that can meaningfully impact real-world launch outcomes. For example, the analysis shows that pricing new releases within the $10 to $15 range is consistently associated with significantly higher owner counts, providing developers with a clear target for maximizing sales potential while maintaining perceived value.

Similarly, the results indicate that timing a game’s launch to coincide with major Steam sales—such as the annual summer or winter events—substantially increases a title’s visibility and accelerates owner growth. This insight enables developers to strategically schedule releases for periods of maximum marketplace activity. In addition, the project reveals that games in the Action and Simulation genres tend to experience the strongest gains from optimized pricing and release strategies, highlighting specific opportunities for studios working in those genres. Beyond these direct recommendations, the clustering results allow developers to benchmark their projects against successful market segments, offering a data-driven basis for refining design, pricing, and launch decisions. Collectively, these insights empower indie studios to make more informed, lower-risk decisions that improve the likelihood of commercial success in an intensely competitive environment.

F3. Evaluation of Overall Success and Effectiveness

The project was highly successful in meeting its stated goals and objectives, delivering all planned analytical models, data products, and decision-support tools within the defined scope and timeline. The linear regression model not only exceeded the accuracy benchmark with an R² score of 0.67, but also provided clear, interpretable guidance on the impact of pricing, release timing, and genre on commercial success. Likewise, the K-means clustering analysis yielded actionable, well-defined market segments, validated by robust silhouette scores and clear PCA visualizations. All key deliverables—including a thoroughly cleaned dataset, comprehensive visualizations, and a final synthesis of actionable recommendations—were completed as intended and support the central aim of equipping indie developers with practical, data-driven strategies. Throughout the project, methodological rigor was maintained, with all steps documented and all model assumptions carefully checked and reported. “The main limitations of this project include reliance on estimated ownership ranges instead of exact counts, exclusion of incomplete records which may bias the sample, and the absence of real-time sales or marketing data. While certain limitations were inherent in the dataset—such as the reliance on estimated owner ranges and the exclusion of games with incomplete data—these were transparently addressed, and do not diminish the overall value of the findings. The absence of real-time sales, marketing spend, or post-launch engagement data means that the analysis is most relevant for launch planning and benchmarking, rather than for post-release performance tracking. Nonetheless, the project represents a robust and effective application of data analytics to a real-world industry challenge, with the results providing a meaningful, evidence-based foundation for decision-making by independent game developers.

G. Key Takeaways

G1. Summary of Conclusions

The analysis conducted in this project leads to several clear and actionable conclusions for indie game developers seeking success on Steam. The most influential factors driving owner counts and commercial outcomes are launch price, release timing, and genre—all of which are variables within a developer’s control. Both statistical modeling and industry research confirm that setting a game’s launch price in the $10 to $15 range, and timing its release to coincide with major Steam sales events, can significantly increase the likelihood of achieving higher owner counts. Furthermore, games in the Action and Simulation genres stand to benefit most from these strategies, as indicated by the data.

Clustering and benchmarking further enhance decision-making by allowing developers to position their games relative to distinct market segments, helping to reduce uncertainty and financial risk. Ultimately, this project demonstrates that a data-driven approach empowers indie developers to make more informed and confident launch decisions, improving both visibility and sales potential. While no single analysis can guarantee success in such a competitive landscape, leveraging historical data and robust analytics provides a distinct strategic advantage for future indie releases on Steam.

G2. Visual Communication and Storytelling

Effective communication of analytical findings was a central priority throughout this project, guiding the selection and design of all visualizations and tools. Each graphic was chosen not only for its technical accuracy but also for its ability to convey complex results in an accessible, actionable manner to a diverse audience—including both technical analysts and indie game developers without specialized data training. Correlation heatmaps were employed to quickly highlight the most important relationships between variables such as price, ownership, genre, and release timing, allowing users to identify key predictors at a glance. Bar plots and histograms provided clear summaries of the distributions for price points, ownership counts, genres, and release months, helping to reveal popular trends and identify outlier behaviors that could inform decision-making.

Scatterplots with overlaid regression lines demonstrated the specific effect of pricing on owner counts, making the marginal impact of each price tier immediately apparent. Principal component analysis (PCA) cluster plots distilled high-dimensional clustering results into an intuitive two-dimensional format, enabling developers to visually benchmark their games against clear market segments. By integrating these visual tools throughout the analysis, the project ensured that critical insights could be understood and applied by all stakeholders, ultimately supporting more effective storytelling, clearer decision-making, and broader impact.

G3. Recommendations: Courses of Action

Based on the results of this analysis, there are several clear courses of action that indie game developers should consider to maximize their chances of success on Steam. First, setting a launch price in the $10 to $15 range is strongly recommended, as the data demonstrates that games in this price band consistently achieve higher owner counts while maintaining an appealing balance between accessibility and perceived value. Second, developers should strategically schedule their game releases to coincide with major Steam sales events, such as the annual summer and winter sales. These high-traffic periods substantially increase visibility and owner growth, providing a critical advantage in a crowded marketplace. Finally, developers are encouraged to benchmark their game’s key features—such as genre, price point, and release timing—against the successful market clusters identified in this project. By positioning their titles relative to these proven segments, indie studios can make more informed design, pricing, and launch decisions, thereby reducing risk and improving commercial outcomes. These recommendations are directly grounded in the project’s empirical findings and offer practical, actionable steps for studios aiming to improve their visibility and sales performance on Steam.

Kaggle. (2023). *Steam video games dataset*. Retrieved from https://www.kaggle.com/datasets/

Pedregosa, F., et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56.