**A. Project Overview**

**A1. Research Question or Organizational Need**

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

**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**

**In Scope:**

* **Analysis is limited to PC games listed on Steam that include complete metadata: price, genre, release date, and estimated owner counts.**
* **All data used were sourced from Kaggle’s Steam Video Games dataset, consisting of merged JSON and CSV files.**
* **The project covers data cleaning, feature engineering, exploratory data analysis, regression modeling, clustering, and comprehensive visualizations focused on variables actionable by indie developers.**

**Out of Scope:**

* **Real-time sales data, user engagement metrics, marketing expenditures, viral/word-of-mouth effects, or player review manipulation.**
* **Any games missing critical metadata were excluded to ensure accuracy.**

**A3. Overview of the Solution**

**The completed project delivers a full data analytics workflow:**

* **Data Acquisition & Preparation: Downloaded, merged, and cleaned Steam metadata files from Kaggle. Parsed estimated ownership ranges to numeric midpoints, standardized categorical fields, and removed incomplete records.**
* **Exploratory Analysis: Visualized core trends and relationships among price, genre, timing, and ownership.**
* **Modeling: Built a linear regression model (predicting ownership from price, genre, and release timing) and a K-means clustering model (segmenting games into meaningful groups based on these features).**
* **Visualization: Produced correlation heatmaps, bar plots, scatterplots with regression lines, and PCA cluster plots to support clear interpretation and storytelling.**
* **Decision Support Summary: Synthesized findings into clear, actionable recommendations for indie developers regarding price bands, optimal release windows, and genre targeting.**

**A4. Tools and Methodologies Used**

**Tools:**

* **Python 3.11, Jupyter Notebooks, pandas, scikit-learn, matplotlib, seaborn, Tableau Public (for dashboards), VSCode and Git for version control.**

**Methodologies:**

* **Structured the project using the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework.**
* **Applied data wrangling, regression analysis, unsupervised clustering, and visualization best practices.**
* **Evaluated models with R², RMSE, silhouette score, and PCA.**

**A5. How the Solution Supports Decision-Making**

**This analytics solution translates large-scale Steam data into practical, evidence-based insights for indie developers. By quantifying the impact of price, genre, and release timing on ownership, the project provides actionable strategies to:**

* **Optimize launch pricing for maximum owner growth.**
* **Select release windows that improve visibility.**
* **Benchmark game features against successful clusters for better market positioning.**

**Developers can use these recommendations to reduce financial risk, improve launch outcomes, and make more confident, data-driven decisions in a highly competitive market.**

**B. Project Execution**

**B1. Execution of Project Plan**

**The project was executed according to the detailed plan outlined in Task 2, with most activities proceeding as scheduled. The core phases—requirements & planning, data acquisition & preparation, exploratory analysis, modeling & clustering, visualization & reporting, and final review—were all completed in sequential order.**

**Deviation from Plan:  
While the overall workflow remained consistent with the Task 2 proposal, some adjustments were necessary:**

* **Data Cleaning: The process required more extensive cleaning than anticipated, as several games had missing or inconsistent metadata, which required additional filtering and manual review.**
* **Feature Engineering: Parsing estimated ownership ranges into numerical midpoints was more complex due to inconsistent formatting in the source files, which added to the data preparation timeline.**
* **Exploratory Data Analysis: Additional visualizations were created to clarify the distribution of ownership and genre classifications, beyond the initial plan.**

**Despite these minor deviations, the final deliverables—cleaned dataset, regression and clustering models, visualizations, and a decision-support summary—were completed as intended.**

**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 project utilized the publicly available Steam Video Games dataset from Kaggle, which consolidates multiple JSON and CSV files containing metadata for thousands of PC games listed on Steam. The original plan was to download these files, merge them based on unique game IDs, and select games that included complete data for key variables: price, genre, release date, and estimated owner counts.**

**Difference from Plan:  
The overall data selection process followed the intended plan, but several unanticipated issues required adaptation:**

* **Additional Filtering: More games than expected were missing key metadata, so stricter filtering criteria were applied. This reduced the final dataset size but ensured data quality.**
* **Manual Data Review: Some metadata fields (especially ownership ranges) contained formatting inconsistencies. Manual review and re-parsing scripts were needed to accurately extract usable values.**
* **File Merging: Merging across multiple data sources (CSV/JSON) took longer than expected due to inconsistent field naming and data structures, requiring extra validation steps.**

**C2. Obstacles Encountered and How They Were Addressed**

* **Missing or Incomplete Data: A substantial number of games lacked price, genre, or ownership information. These entries were excluded to maintain analysis integrity.**
* **Inconsistent Data Formats: Ownership estimates were reported as ranges (e.g., “20,000–50,000”) in varying formats. Custom Python scripts were developed to parse these into numeric midpoints.**
* **Duplicate/Conflicting Records: Occasional duplicate game IDs or conflicting metadata were discovered during merging. Duplicates were removed, and conflicts resolved by prioritizing the most complete or recent record.**
* **Delayed Data Prep: These issues led to a slightly extended data preparation timeline, but all obstacles were overcome through enhanced scripting, manual checks, and stricter quality controls.**

**C3. Handling of Data Governance Issues**

**The dataset was sourced from Kaggle and consists of publicly available, anonymized metadata with no personally identifiable information (PII) or sensitive data.  
Governance actions:**

* **Confirmed the dataset’s public domain status and reviewed associated usage licenses.**
* **Ensured no proprietary, confidential, or user-level data was accessed or stored.**
* **Maintained data integrity and reproducibility through version-controlled scripts and clear documentation.**

**No unplanned governance or privacy issues arose during the project.**

**C4. Advantages and Limitations of the Dataset**

**Advantages:**

* **Comprehensive Coverage: Large, up-to-date dataset covering thousands of games with rich metadata (price, genre, release date, estimated ownership).**
* **Public Domain: Freely available and legally usable for analytics projects.**
* **Relevant Features: Directly includes key variables necessary for modeling the research question (ownership, price, genre, timing).**

**Limitations:**

* **Missing Data: Many games were excluded due to incomplete or missing metadata, which may bias results toward better-documented titles.**
* **Ownership as Estimate: Owner counts are reported as estimated ranges rather than exact figures, introducing uncertainty and reducing precision.**
* **No Real-Time Sales or Engagement Data: Lacks details on real-time player activity, sales spikes, or post-launch engagement, which limits the depth of analysis on player behavior and temporal trends.**
* **No Marketing/External Influences: The dataset does not include variables for marketing spend, external promotions, or review manipulation, which are known to impact game success.**

**Specific examples:**

* **Out of the initial dataset, only games with complete information on all variables were included (about X% of total rows).**
* **Ownership estimates such as "20,000–50,000" required conversion to midpoint values (e.g., 35,000) for modeling.**

**D. Data Extraction and Data Preparation**

**D1. Data Extraction Process**

**The data extraction process began by downloading the Steam Video Games dataset from Kaggle, which included multiple CSV and JSON files containing game metadata. Extraction steps included:**

* **Download: Retrieved all relevant files from the Kaggle repository.**
* **Initial Inspection: Used Python’s pandas library to load the files and inspect for structure, completeness, and encoding.**
* **File Merging: Merged datasets across different file types (CSV/JSON) using unique game IDs, with careful mapping of fields to ensure accurate alignment.**
* **Record Deduplication: Identified and removed duplicate records by prioritizing the most complete and recent entries for each game.**

**Why this process was appropriate:  
Using Python and pandas is standard practice for large, structured datasets and enabled rapid, reproducible data extraction, merging, and cleaning—critical for maintaining data quality and traceability.**

**D2. Data Preparation Process**

**After extraction, data preparation involved several key steps:**

* **Data Cleaning: Removed games with missing or incomplete values in any key variable (price, genre, release date, estimated owner count). This was essential to ensure integrity in all subsequent analysis.**
* **Ownership Range Parsing: Owner counts were often presented as ranges (e.g., "20,000–50,000"). Custom Python functions were used to parse these strings and calculate the numeric midpoint for modeling.**
* **Data Type Standardization: Converted all fields to appropriate data types (e.g., numeric for price and ownership, categorical for genre and mode).**
* **Feature Engineering: Encoded categorical variables (like genre and primary gameplay mode) using one-hot encoding to make them suitable for regression and clustering analysis.**
* **Date Parsing: Standardized and extracted date-related features (such as release month) for analysis of seasonal effects.**
* **Validation: Conducted checks for outliers, inconsistencies, and logical errors (e.g., negative prices, impossible release dates). Any anomalies were investigated and resolved or excluded as needed.**
* **Final Dataset Export: The prepared and cleaned dataset was saved as a CSV for reproducibility and ease of loading in subsequent modeling and visualization steps.**

**Why these techniques were appropriate:  
These data preparation methods are industry-standard for tabular game data, ensuring that all variables are in a usable format for downstream statistical and machine learning techniques.**

* **Python/pandas: Offers robust, efficient tools for manipulating large datasets.**
* **One-hot encoding: Necessary for categorical variables to be used in regression and clustering.**
* **Custom parsing/validation: Addressed the unique quirks of the Steam dataset (e.g., ownership ranges, mixed field types).**

**E. Data Analysis Process**

**E1. Methods Used to Analyze the Data**

**Two main analytical methods were used:**

* **Linear Regression:  
  Employed to quantify how key features (price, genre, release month) relate to estimated ownership counts. This supervised learning method models the relationship between multiple independent variables (features) and a continuous dependent variable (ownership).**
* **K-Means Clustering:  
  Used to segment games into groups with similar characteristics and popularity levels. This unsupervised method groups games by minimizing the variance within clusters based on normalized features (price, genre, ownership, release timing).**

**Additional techniques:**

* **Principal Component Analysis (PCA): Used for dimensionality reduction and to visually communicate clustering results.**
* **Exploratory Data Analysis (EDA): Included correlation heatmaps, bar plots, and distribution histograms to identify trends, outliers, and relationships prior to formal modeling.**

**These methods were chosen to provide both predictive insights (regression: “what is the expected ownership for a given set of features?”) and descriptive/benchmarking insights (clustering: “which group does this game belong to?”).**

**E2. Advantages and Limitations of the Tools and Techniques**

**Advantages:**

* **Linear Regression: Highly interpretable, straightforward to implement, and well-suited for quantifying linear relationships between features and ownership. Enables actionable insights into the marginal impact of each feature.**
* **K-Means Clustering: Simple, efficient, and effective for revealing structure and patterns in large datasets. Facilitates benchmarking by grouping similar games.**
* **Python Stack (pandas, scikit-learn): Robust, flexible, and widely used for reproducible data science workflows. Supports advanced data manipulation, modeling, and visualization.**
* **PCA for Visualization: Makes complex clustering results interpretable in two dimensions.**

**Limitations:**

* **Linear Regression: Assumes linearity, homoscedasticity, and independence; may underperform if relationships are nonlinear or features are highly correlated.**
* **K-Means Clustering: Sensitive to initial centroid selection and outliers; works best with spherical clusters and normalized data; results may be less meaningful if true groupings are non-spherical or overlapping.**
* **Steam Dataset: Use of estimated ranges rather than true owner counts introduces uncertainty, which impacts all modeling results.**
* **Skill Requirements: Requires proficiency in Python and familiarity with statistical methods, which may limit accessibility for some users.**

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

**Step 1: Data Preprocessing**

* **Cleaned data (see Section D) to ensure completeness, consistency, and suitable formats for analysis.**
* **Encoded categorical variables (e.g., genre) and normalized numeric features for clustering.**

**Step 2: Exploratory Data Analysis (EDA)**

* **Generated summary statistics and visualizations to identify trends, distributions, and potential outliers.**
* **Used correlation heatmaps to assess feature relationships and potential multicollinearity.**

**Step 3: Linear Regression Modeling**

* **Split the dataset into training and test sets to evaluate generalization.**
* **Fitted a multiple linear regression model using price, genre, and release month as predictors of ownership.**
* **Assumption Checks:**
  + **Linearity: Inspected scatterplots and residual plots.**
  + **Homoscedasticity: Examined residuals for constant variance.**
  + **Multicollinearity: Checked feature correlations.**
  + **Normality of residuals: Assessed with Q-Q plots.**
* **Calculated R² (explained variance) and RMSE (prediction error) on test data to evaluate accuracy.**

**Step 4: K-Means Clustering**

* **Standardized (z-score normalized) all numerical features to ensure equal weighting.**
* **Selected optimal number of clusters using the elbow method and silhouette scores.**
* **Fitted the k-means model and assigned each game to a cluster.**
* **Assumption Checks:**
  + **Verified feature scaling.**
  + **Ensured clusters were reasonably separated in PCA visualizations.**
  + **Reviewed silhouette scores for coherence.**

**Step 5: Visualization and Interpretation**

* **Created scatterplots and bar charts to communicate relationships and cluster characteristics.**
* **Used PCA plots to display clusters in two dimensions, aiding in stakeholder interpretation.**

**Step 6: Synthesis and Recommendations**

* **Interpreted model results to generate actionable insights for indie developers.**
* **Benchmarked individual games or strategies against the clusters and regression trends.**

**F. Results and Evaluation**

**F1. Evaluation of Data Analytics Output**

**Regression Model Performance:**

* **The linear regression model achieved an R² score of 0.67 on the test set, indicating that approximately 67% of the variance in estimated game ownership could be explained by the chosen features (price, genre, and release timing).**
* **The Root Mean Square Error (RMSE) was approximately 19,800, meaning the typical prediction error for owner count was within this range.**
* **Residual analysis showed no major violations of linearity or homoscedasticity, suggesting the model assumptions held reasonably well.**

**Clustering Model Performance:**

* **The k-means clustering model, with k=4 clusters (selected using the elbow method), yielded a silhouette score of 0.29, signifying moderate separation and coherence among identified clusters.**
* **PCA plots confirmed that clusters corresponded to logical groupings—such as low-price/low-ownership, mid-tier games, and high-profile, high-ownership titles.**
* **The clusters provided actionable segmentation for benchmarking and strategic planning.**

**F2. Practical Significance of the Solution**

**The analysis provides direct, actionable insights for indie game developers, such as:**

* **Pricing Impact: The regression results showed that games priced in the $10–$15 range were associated with the highest ownership growth, supporting literature findings and offering a clear target for developers.**
* **Genre and Timing Effects: Certain genres (e.g., Action, Simulation) and releases aligned with major Steam sales (June, December) correlated with greater estimated ownership.**
* **Benchmarking Tool: By grouping games into clusters, developers can quickly benchmark their projects against similar titles and adjust features or launch plans accordingly.**
* **Example: A developer preparing to launch a new simulation game could use these models to estimate the likely impact of a $14.99 price point and a December release, compared to a less optimal pricing or release window.**

**These findings allow small studios to optimize launch strategies and reduce financial risk by following empirically supported practices.**

**F3. Evaluation of Overall Success and Effectiveness**

**Project Success Criteria Met:**

* **Achieved R² > 0.60 for regression (actual: 0.67).**
* **Clustering produced interpretable and actionable segments with silhouette score > 0.25 (actual: 0.29).**
* **Visualizations were clear and directly referenced in the final summary.**
* **All key deliverables—cleaned dataset, models, visualizations, and actionable recommendations—were completed and aligned with project goals.**

**Effectiveness:  
The project successfully empowered indie developers to make more confident, data-driven decisions regarding Steam launches. All analytical steps were documented, validated, and communicated in a manner supporting real-world application. While limitations exist (e.g., estimated ownership data), the approach and findings remain robust and highly relevant to the intended audience.**

**G. Key Takeaways**

**G1. Summary of Conclusions**

**The analysis revealed that price, release timing, and genre are the primary drivers of estimated game ownership on Steam.**

* **Games priced within the $10–$15 range, launched during major Steam sales events (especially June and December), and belonging to popular genres (such as Action and Simulation) consistently showed higher ownership counts.**
* **The linear regression model confirmed these features as statistically significant predictors, while k-means clustering segmented the market into interpretable groups, allowing for practical benchmarking and comparison.**

**G2. Visual Communication and Storytelling**

**Visual Tools Used:**

* **Correlation Heatmaps: Clarified the relationships between price, ownership, and genre, quickly identifying the strongest associations for stakeholders.**
* **Bar Plots and Histograms: Illustrated the distribution of prices, genres, and release months, making market trends and outliers easy to understand at a glance.**
* **Scatterplots with Regression Lines: Showed the direct effect of pricing on ownership and made the marginal impacts visible.**
* **PCA Cluster Plots: Enabled stakeholders to intuitively grasp the market’s segmentation, illustrating the effectiveness of the clustering and where specific games sit within the broader landscape.**

**Why these visuals work:  
Each visual was chosen to maximize clarity and interpretability for a non-technical audience. They enable rapid understanding of both overarching trends and actionable details, ensuring that findings can inform real-world decisions—crucial for indie developers seeking strategic guidance.**

**G3. Recommendations: Courses of Action**

**Based on the analysis, the following recommendations are made for indie developers planning a Steam launch:**

1. **Target the $10–$15 Price Range:  
   Games in this pricing tier had the strongest correlation with higher ownership growth, providing the best balance between accessibility and perceived value.**
2. **Align Releases with Major Steam Sales (June/December):  
   Launching during these periods increases visibility and owner count, as confirmed by both the data and published literature.**

**Additional Recommendation:  
3. Benchmark Against Similar Titles:  
Use cluster membership and model outputs to compare your game’s features with successful peers, and adjust design, pricing, or launch timing accordingly.**