**Panopto Presentation Script (Expanded) Slide 1 – Title Slide  
*"Hello, my name is Jonathan Mitchell, and this is my Data Analytics Capstone presentation for the Bachelor of Science in Data Analytics at Western Governors University. My project is titled ‘Steam Games Analysis: Data-Driven Insights for Indie Game Success.’  
Over the next ten minutes, I’ll walk you through the problem I set out to solve, the data and methods I used, the key findings, and how these insights can help indie developers make better, evidence-based decisions about launching their games on Steam."***

**Slide 2 – Introduction & Research Question  
\*"The indie video game market on Steam is one of the most competitive digital marketplaces in the world. Thousands of new titles release each year, but only a fraction achieve lasting visibility or commercial success. For smaller studios or solo developers—often working with tight budgets—this uncertainty can be financially risky.**

**That led me to my core research question: Which game features—such as genre, price, or release timing—correlate most strongly with higher estimated ownership on Steam?**

**The purpose here isn’t just curiosity. The goal is to replace guesswork with data-driven guidance, giving indie developers a better chance to position their games for visibility and sales in a crowded marketplace."\***

**Slide 3 – Background & Scope  
\*"Indie developers typically work with personal savings or limited funding. This makes every decision—from how to price their game, to which genre to focus on, to when to launch—critical. Steam has more than 50,000 titles competing for attention, so even well-reviewed games can easily go unnoticed without a deliberate launch strategy.**

**For this project, I narrowed the scope to PC games on Steam with complete metadata for key variables: price, primary genre, release date, and estimated ownership counts.  
I sourced the data from a comprehensive dataset on Kaggle that merges multiple Steam metadata sources. To maintain quality, I excluded any games with missing or inconsistent data in these core fields.**

**This focus allowed me to analyze a clean, relevant dataset while acknowledging that certain factors—like marketing budgets, real-time sales figures, and user review sentiment—were out of scope for this analysis."\***

**Slide 4 – Summary of Solution & Deliverables  
\*"To answer the research question, I built a complete data analytics workflow aimed at producing actionable recommendations for indie game launches.**

**The project deliverables included:**

* **A cleaned and consolidated Steam games dataset, created by merging multiple Kaggle sources and resolving inconsistencies.**
* **A multiple linear regression model to predict estimated ownership counts based on price, genre, and release timing.**
* **A K-means clustering model to segment games into market groups with similar characteristics and outcomes.**
* **A suite of visualizations—including correlation heatmaps, bar plots, scatterplots, and PCA cluster diagrams—that make it easier to interpret the findings.**

**These steps followed the CRISP-DM framework, ensuring the work moved systematically from data acquisition and preparation to modeling, evaluation, and recommendations.**

**The ultimate goal was to produce not just analysis, but clear, evidence-based guidance for developers on how to price, position, and time their games for better success."\***

**Slide 5 – Tools & Methods  
\*"I selected tools that were both industry-relevant and capable of handling large datasets while supporting statistical and machine learning methods.**

**The analysis was conducted in Python 3.11, primarily within Jupyter Notebooks for its combination of code execution, documentation, and reproducibility. Key Python libraries included:**

* **pandas for data manipulation,**
* **scikit-learn for modeling,**
* **matplotlib and seaborn for visualization.**

**I also used Tableau Public to create interactive dashboards, and VS Code with Git for code management and version control.**

**Methodologically, the project followed the CRISP-DM process. After cleaning and engineering features from the dataset, I applied multiple linear regression to quantify relationships between predictors and ownership counts, and K-means clustering to uncover natural groupings in the market.**

**Model performance was evaluated using R² and RMSE for regression, and silhouette scores for clustering. I also used Principal Component Analysis—PCA—to reduce dimensionality for clear 2D visualizations of cluster structure."\***

**Slide 6 – Data Acquisition & Preparation  
\*"The data came from the publicly available Steam Video Games dataset on Kaggle. This dataset combines multiple CSV and JSON files containing metadata for thousands of PC games.**

**To maintain data integrity, I only kept games with complete information on price, primary genre, release date, and estimated ownership. Estimated ownership counts on Steam are not exact numbers; they’re reported as ranges, such as ‘20,000–50,000.’ To make them usable for modeling, I converted these ranges into numeric midpoints.**

**Other preparation steps included standardizing data types, one-hot encoding categorical variables, and extracting new features like release month from the release date. I also ran custom parsing scripts to handle inconsistent formats and used targeted filtering to remove duplicates.**

**The final cleaned dataset contained no personal or sensitive information, and all work was in compliance with the public domain usage rights of the source data."\***

**Slide 7 – Dataset Advantages & Limitations  
\*"Before diving into the models, it’s important to understand the strengths and weaknesses of the dataset itself.**

**Advantages:**

* **It offers comprehensive coverage, with thousands of Steam PC games and rich metadata on price, genre, release date, and estimated ownership.**
* **The data is directly relevant to my research question, containing the exact features needed to explore correlations between these variables and game success.**
* **It’s also public and legally usable, with no personally identifiable information.**

**Limitations:**

* **Estimated ownership is reported only as ranges, so converting these to numeric midpoints introduced some unavoidable uncertainty.**
* **Many games with incomplete metadata were excluded. This may create a bias toward better-documented, often more successful titles.**
* **Finally, the dataset does not include external factors such as marketing spend, review manipulation, or player engagement patterns, all of which could influence ownership but were outside the scope of this analysis.**

**Being aware of these limitations is key for interpreting the results realistically and responsibly."\***

**Slide 8 – Regression Method  
\*"The first primary analysis method was multiple linear regression, which I chose because it’s both interpretable and capable of showing how multiple predictors interact to affect a continuous outcome—in this case, estimated ownership counts.**

**The predictors in my model were:**

* **Price,**
* **Primary genre,**
* **Release month.**

**Steps Taken:**

* **I split the dataset into training and testing sets to evaluate model generalization.**
* **Categorical variables, such as genre and month, were encoded so they could be used in the regression model.**
* **I fitted the regression model and then examined the coefficients to understand each variable’s contribution.**

**Assumption Checks:**

* **I verified linearity through scatterplots and residual plots.**
* **Homoscedasticity—constant variance of residuals—was confirmed visually.**
* **Multicollinearity was checked through feature correlation analysis.**
* **Residual normality was assessed via Q–Q plots.**

**These checks ensured the regression model met its key assumptions, increasing confidence in both its performance and interpretability."\***

**Slide 9 – Clustering Method  
\*"The second main technique I applied was K-means clustering. While regression identifies relationships between features and ownership, clustering is useful for grouping games into natural market segments. This allows for benchmarking—comparing a game’s profile to others in its segment.**

**Features Used:**

* **Price,**
* **Genre,**
* **Estimated ownership,**
* **Release timing.**

**Steps Taken:**

* **I standardized numeric features to ensure no single variable dominated due to scale.**
* **I determined the optimal number of clusters using the elbow method and confirmed the choice with silhouette scores.**
* **I then applied K-means to segment the dataset and examined each cluster’s defining characteristics.**

**Validation:**

* **The silhouette score indicated that the clusters were meaningfully distinct, even if not strongly separated—something common in real-world market data.**
* **Principal Component Analysis was used to reduce the data to two dimensions for visualization, making it possible to clearly illustrate cluster separation in the presentation.**

**Together, regression and clustering provide both predictive insights and strategic market positioning tools for indie developers."\***

**Slide 10 – Regression Results  
\*"The regression model achieved an R² of 0.67, meaning it explained about 67% of the variance in estimated ownership counts.  
The Root Mean Squared Error—our average prediction error—was roughly 19,800 owners.**

**Key Findings from the Model:**

* **Games priced between $10 and $15 had the strongest positive correlation with higher ownership counts. This range appears to balance affordability for players with perceived value for the developer.**
* **Release timing mattered: games launched during major Steam sales windows saw significant boosts in visibility and early adoption.**
* **Action and Simulation genres tended to perform best under these optimal pricing and timing conditions.**

**These results directly answer the research question by quantifying which factors matter most and in what direction they affect ownership. They also form the foundation for the actionable recommendations I’ll share later."\***

**Slide 11 – Clustering Results  
\*"The K-means clustering analysis revealed four distinct market segments, validated by a silhouette score of 0.29. While not an extremely high value, it’s typical for complex real-world market data and still indicates meaningful group separation.**

**Cluster Profiles:**

1. **Low-price / low-ownership indie titles – often casual or niche games with minimal visibility.**
2. **Mid-tier titles – moderate prices, moderate ownership, typically stable performers.**
3. **High-profile titles – large owner counts, broad appeal, and high visibility.**
4. **Niche high-value games – strong performance within specific genres despite narrower audience appeal.**

**PCA plots provided a clear 2D visualization of these clusters, making it easier to see how games group together. This segmentation creates a benchmarking tool—developers can see which cluster they’re aiming for and adjust pricing, genre, and launch timing to better align with successful examples."\***

**Slide 12 – Practical Significance  
\*"The most important part of this analysis is what it means in the real world for indie developers.**

**From the regression:**

* **Pricing Strategy – Target the $10–$15 range to maximize ownership potential without undermining perceived value.**
* **Release Timing – Launching during major Steam sales significantly increases initial visibility and accelerates early sales momentum.**
* **Genre Focus – While Action and Simulation showed the strongest results under these conditions, any genre can benefit from benchmarking through clustering.**

**From the clustering:**

* **Developers can position their game within a specific market segment and make strategic decisions—pricing, feature set, marketing—that are consistent with the most successful titles in that cluster.**

**By applying these strategies, indie developers can reduce financial risk, improve market positioning, and increase their chances of commercial success."\***

**Slide 13 – Overall Success & Limitations  
\*"Looking at the project as a whole, I consider it a success for several reasons.**

**Successes:**

* **The regression model exceeded my accuracy goal, achieving an R² of 0.67.**
* **The clustering analysis identified four actionable market segments, giving developers a clear benchmarking tool.**
* **All planned deliverables were completed: a cleaned dataset, validated models, visualizations, and concrete recommendations.**

**Limitations:**

* **Estimated ownership ranges had to be converted to numeric midpoints, which inevitably introduced some uncertainty.**
* **The exclusion of games with incomplete metadata could bias the sample toward better-documented, and often more successful, titles.**
* **The absence of external influences—like marketing spend, review manipulation, or player engagement metrics—means that real-world performance may be affected by factors outside the scope of this analysis.**

**Even with these limitations, the insights remain highly relevant to indie developers making pricing, timing, and genre decisions."\***

**Slide 14 – Key Takeaways & Recommendations  
\*"From this work, three key takeaways stand out:**

1. **Price, timing, and genre are the strongest measurable drivers of success on Steam.**
2. **Data-driven decision-making significantly outperforms guesswork for indie launches.**
3. **Market segmentation through clustering provides a powerful way to benchmark against comparable titles.**

**Recommendations:**

* **Price – Target a launch price between $10–$15.**
* **Timing – Align releases with major Steam sales events to maximize visibility.**
* **Benchmarking – Use the cluster insights to position a game in the most favorable market segment and model its strategy after proven successes.**

**These recommendations give indie developers an actionable framework for improving launch outcomes in a crowded marketplace."\***

**Slide 15 – Why These Visuals Worked  
\*"The visuals in this project were designed with two goals: clarity and direct connection to findings.**

**Clarity:**

* **I chose charts that could be quickly understood by both technical and non-technical audiences.**
* **Overly complex or abstract visuals were avoided in favor of straightforward ones tied directly to the research question.**

**Insight Delivery:**

* **Correlation heatmaps quickly highlighted the strongest predictors of ownership.**
* **Scatterplots with regression lines illustrated exactly how features like price affected ownership counts.**
* **PCA cluster plots visually separated market segments, making it easy to see where different games fell.**

**Storytelling:**

* **The visuals were sequenced to move logically from identifying relationships → presenting model results → showing actionable market segments.**
* **Each chart was chosen not just to look good, but to reinforce a specific point in the narrative, connecting the data directly to practical strategy."\***

**Slide 16 – References**  
*"The sources for this analysis include the Steam Video Games dataset hosted on Kaggle, which combines multiple publicly available Steam metadata sources.  
I also referenced official documentation for Python libraries such as pandas, scikit-learn, matplotlib, and seaborn, as well as Tableau Public’s documentation for visualization features.  
All work complied with the dataset’s usage rights, and no personally identifiable information was collected or used."*

**Slide 17 – Thanks**  
*"Thank you for your time and attention.  
This project combined my passion for data analysis with my interest in the indie games market, and I’m confident the insights it produced can help small developers make more informed, strategic launch decisions.  
I’d be happy to answer any questions you have."*