### **Problem Statement:**

NV Games plans to launch their first game within the next year. To ensure a successful entry into the market, they are eager to understand what elements gamers value most and,therefore, lead to high sales. This strategy is designed to avoid the common pitfall in the gaming industry, where approximately 83% of games fail to sustain sales beyond their initial launch period. For the game to be considered a financial success and cover its development and maintenance costs, it must generate a minimum revenue of $500,000 within the first year.

The gaming industry is highly competitive, with only a few winners where player satisfaction directly impacts a game's commercial success. To pinpoint what gamers truly value, I will analyze hundreds of games and their features, such as rating scores, developers, pricing, and more, to determine which of these characteristics have the strongest correlation with global sales.

### **Approach:**

#### **Data Wrangling:**

After merging two datasets, which initially contained 8,501 rows and 54 columns, I cleaned the data by removing null values and irrelevant columns, such as website URLs and Metacritic URLs. The dataset was then refined to 458 rows and 12 columns. I corrected data types, such as converting the Release Date column to a datetime format. I also excluded rows that lacked meaningful data, such as those with a Metacritic Score of 0 or Player Score marked as ‘tbd.’

#### **Exploratory Data Analysis (EDA):**

With the cleaned dataset, I explored the relationships between features and the target variable, Global Sales. A histogram revealed six outliers within Global Sales exceeding 5 million copies. Removing these outliers adjusted the mean from 1 million to a more accurate 679K. I then created a pair plot, correlation table, and heat map to identify relationships between features and the target variable. The feature most strongly correlated with the target was NA Sales, as expected, while Metacritic Score had the strongest non-sales correlation, with a correlation coefficient of 0.41.

I further examined how these variables interacted with non-numeric features like Developer, Publisher, Platform, Genre, and Rating. This analysis revealed that the majority of the highest selling games were rated 'M', belonged to the 'Action' genre, and had the highest sales on the X360 and PS3 platforms.

#### **Preprocessing and Training Data:**

I created dummy variables for categorical columns and reduced their number by setting a threshold of 0.01 to minimize noise and variation. The numerical columns were scaled for machine learning modeling. The features, X, included all columns except Global Sales and GameID, while y represented the target variable, Global Sales. I then split the data into training and testing sets using an 80/20 ratio.

#### **Modeling:**

With data cleaning, wrangling, EDA, and preprocessing complete, I proceeded to build predictive machine learning models. I experimented with three models to identify the best fit based on minimizing Mean Squared Error (MSE) while avoiding overfitting.

1. **Linear Regression:**
   * Initial MSE: 1.715×10221.715 \times 10^{22}1.715×1022, which was extremely high due to the dataset's complexity.
   * Cross-validation further increased the MSE to 8.367×10228.367 \times 10^{22}8.367×1022, indicating overfitting on the initial subset of data.
2. **Random Forest:**
   * Using 100 n\_estimators, the MSE was significantly reduced to 0.0585.
   * Applying Grid Search for hyperparameter optimization with 5-fold cross validation resulted in an optimal max depth of 10, a minimum sample size of 5, leading to a MSE of 0.0972. The increased MSE indicated potential overfitting in the initial model.
3. **Gradient Boosting Regressor:**
   * This model initially produced the lowest MSE of 0.0516.
   * Applying Grid Search for hyperparameter optimization with 5-fold cross validation resulted in an optimal max depth of 4, a minimum sample size of 10, a learning rate of .05, leading to a MSE of 0.0898. The increased MSE indicated potential overfitting in the initial model.

### **Conclusion:**

The Gradient Boosting Regressor model, combined with Grid Search, emerged as the most reliable model, striking a balance between avoiding overfitting and maintaining predictive accuracy. Although it didn't have the absolute lowest MSE, the Grid Search enhanced the model's robustness and versatility, making it more capable of handling unseen data compared to the other models.

### **Ideas for Further Research:**

* Analyze patterns in video game reviews, focusing on keywords and phrases to gain insights into what players enjoy and dislike.

### **3 Recommendations for Clients:**

1. **Predict Video Game Sales:** Utilize the model to forecast sales based on game features, aiding in strategic planning.
2. **Implement Dynamic Pricing:** Adjust pricing based on predicted sales trends, reducing prices for lower-selling games and maintaining or increasing prices for higher-selling ones.
3. **Optimize Inventory Management:** Use sales predictions to manage inventory levels, avoiding understocking or overstocking issues.