



# Video Game Sales Analysis: Exploring Industry Trends and Market Dynamics

A data-driven exploration of the global video game market using Kaggle's comprehensive sales dataset

By Preston Blake

# Dataset Overview & Research Motivation

## Data Source & Scope

This analysis uses the Video Game Sales dataset from Kaggle, which contains comprehensive sales data across multiple gaming platforms and genres worldwide. The dataset provides a rich foundation for understanding market trends, consumer preferences, and the relationship between critical reception and commercial success.

## Key Research Questions

- Do sports games represent more than 10% of the gaming market?
- How do review scores correlate with global sales performance?



□ **Why This Matters:** Understanding these relationships helps predict market trends and informs strategic decisions in game development and publishing.

# Key Findings: Hypothesis Testing & Predictive Modeling

## Sports Genre Hypothesis Test

Testing whether sports games exceed 10% market share:

**Null Hypothesis:**  $H_0: p = .10$ ,

**Alternative Hypothesis:**  $H_1: p > .10$ ,

**Significance Level:** 0.05

**Random Sample:** Action (18), Racing (15), Sports (11), Misc. (9), Platform (8), Simulation (7), Shooter (7), Strategy (6), Adventure (6), Role-Playing (6), Fighting (4), Puzzle (3)

**Test Statistic:** 0.333333

**P-value:** 0.3694413

Given this hypothesis test, we don't have enough evidence to prove that the proportion of sports games exceeds 10%.

## Review Score vs. Global Sales Model

Results:

This linear model has the line  $Y = 77.68996 + 0.54194X_1$ , where  $X_1$  represents the number of global sales in millions, and  $Y$  represents the review score.

- The coefficient of determination ( $R^2$ ): 0.03308
- P-values Global Sales:  $1.2 \times 10^{-15}$
- Intercept:  $2 \times 10^{-16}$

Given all this information from the graphs below and the summary, this wouldn't be a good model to find the review score because of these reasons:

- The coefficient of determination is very low
- The scatterplot isn't randomly scattered
- Histogram of residuals isn't normal, rather it's left-skewed

## Analysis Visualizations

