Video Game Success

Hypothesis:

The goal of this project is to analyze and predict the factors influencing the success of video games using machine learning models. This analysis not only sheds light on consumer behavior and market trends but also explores the socio-historical and economic impacts of the gaming industry.

Data:

Data was collected from top online gaming databases including IGDB, MobyGames, and GiantBomb, resulting in a dataset of over 1000 entries featuring game scores, release dates, genres, developer information, and ratings. Individually they had more data points but we combined using the name of the game as the primary key. We used three regression models, linear, ridge, and random forest, to predict game success, measured by review scores. These models were chosen for their ability to handle the complexity and non-linear relationships within the data. The Random Forest model performed best, with an R-squared score of 0.302, suggesting it could explain about 30.2% of the variance in game success from the features analyzed.

Hypothesis Findings:

Hypothesis 1: Whether games that are multiplayer or single player are rated higher
Support for Hypothesis 1: Two-sample t-test was used. The p-value resulted at about 0.46, much above the 0.05 standard threshold for rejecting the null hypothesis.This means that we currently accept the null hypothesis, which states that multiplayer and single player games have no significant difference in ratings.

T-statistic: -0.7399646416199741
P-value: 0.4596644633094552

Hypothesis 2: If shooter games (represented by Genre1 value of 5) score higher on average than games from other genres.
Support for Hypothesis 2: A two-sample t-test compared the average scores of games identified as Shooters with those that are not. The p-value came to about 0.764, even higher than the first tested hypothesis. Meaning that shooter and non-shooter games have no significant difference in ratings

T-statistic: -0.3000267166266354
P-value: 0.7642949242767471

Hypothesis 3 : Whether games with a high number of reviews have different average scores compared to games with fewer reviews

Support for Hypothesis 3: The dataset was divided into two groups based on the median number of reviews: games with reviews greater than the median are categorized as 'high review' games, and those with reviews less than or equal to the median as 'low review' games. A two-sample t-test was used to compare the average scores between these two groups. The p value in this case was remarkably low: 1.12e-10. This means we must indelibly reject the null hypothesis: our distributions of games with few reviews versus games with many reviews differ greatly. This suggests that in most cases, people are more likely to leave reviews if they have positive comments for the game.

T-statistic: 6.5212444238160066
P-value: 1.1217226791417664e-10

Limitations: The current models have lower accuracy on unfamiliar datasets, highlighting the need for further optimization. If we could've incorporated data like user engagement or money made from the game it would've been a more helpful analysis.

	Hypothesis 1	Hypothesis 2	Hypothesis 3
T-statistic	-0.740	-0.30	6.521
P-value	0.460	0.764	1.122

ML RESULTS:

Random forest regression performs similarly to linear and ridge regression in our dataset, but it typically yields better predictive performance, especially for our complex non-linear dataset. The higher R-squared (R2) score of 0.302 suggests a better fit of the model to the data, implying that approximately 30.2% of the variance in video game success can be explained by the features included in the model.

Linear Regression Results:
Mean Absolute Error (MAE): 0.12533708040836805
Mean Squared Error (MSE): 0.026580229246759
Root Mean Squared Error (RMSE): 0.16303444190341806
R-squared (R2) Score: 0.17164011001753887
Cross-Validation RMSE Scores: [0.17485508 0.1694373 0.16994755 0.15698742 0.15697071]

Improved Random Forest Regression Results:
Mean Absolute Error (MAE): 0.11466551100932643
Mean Squared Error (MSE): 0.022392962635392533
Root Mean Squared Error (RMSE): 0.14964278343907045
R-squared (R2) Score: 0.3021342331990269