Final Project

Jose De Leon

Northeastern University

College of Professional Studies

ALY6010: Probability Theory and Introductory Statistics

Professor: Alireza Samea

March 30, 2025

Contents

Introduction

This report presents a comprehensive statistical analysis of the global video game market using a dataset from VGChartz containing over 64,000 game records. The project was conducted as part of the ALY6010 course at Northeastern University, with the goal of applying inferential statistical techniques, including t-tests and simple linear regression, to identify patterns and predictors of commercial success in the video game industry. After an initial exploratory data analysis (EDA) that revealed genre dominance, regional sales disparities, and skewed critic ratings, we formulated six research questions related to game reviews, genres, regions, release timing, and platform availability.

Each question was approached using appropriate hypothesis tests or regression models, grounded in academic methodology (Ross, 2020; Samea, 2025). Assumptions of normality, independence, and homoscedasticity were carefully checked, and interpretation of results considered both statistical significance and real-world relevance. This structured approach allowed us to explore the relationship between game characteristics and their total sales, offering insights for game developers, publishers, and marketers.

Summary of Initial EDA

The dataset used in this project originates from VGChartz and includes 64,016 video game records. The data includes game title, platform, genre, publisher, developer, release date, critic scores, and global and regional sales. The dataset was cleaned by removing irrelevant columns (like images), replacing missing numeric values with column means, and converting date formats. Duplicates were removed to ensure independence of observations, and missing dates were filled with either the earliest date or the system date.

From the EDA, we identified several key patterns:

- **North America leads** in total game sales, followed by Europe/Africa (PAL) and Japan.
- **Shooter and Action genres** are among the most represented and top-selling categories.
- **Critic scores** average around 7.22 and are right-skewed, with most ratings clustered around 7.5.
- **Sales distributions** across all regions and genres are heavily right-skewed, indicating a few games dominate the market.

These observations raised questions about the influence of critic scores, genre, platform availability, and region on a game's commercial success.

Research Questions Explored

Based on the EDA findings, we formulated the following questions for inferential analysis:

1. **Is the average critic score significantly different from the Metacritic global average (7.021)?**
   This question was prompted by the observation that most critic scores in the dataset cluster around 7.5, noticeably higher than the global average of 7.021 reported by Metacritic.
2. **Do Shooter games generate significantly higher total sales than Action games?**
   Shooter and Action were the top two genres in representation and popularity. We wanted to determine if their commercial success also differed significantly.
3. **Are North American sales significantly higher than those in Europe and Africa combined (PAL region)?**
   North America appeared to dominate in sales, so we tested if this difference was statistically significant.
4. **Is there a significant linear relationship between Critic Score and Total Sales?**
   This correlation test and regression model explores whether critic ratings help predict game sales.

5. **Does the Release Year predict Total Sales?**
   We hypothesized that more recent games may have different sales performance due to changing market trends. We used simple linear regression to test this relationship.
6. **Does the number of platforms a game is released on predict Total Sales?**
   We explore whether broader platform availability leads to higher commercial success using simple linear regression.

Hypotheses and Testing Methods

Each question was approached using hypothesis testing. Here's a breakdown of the null and alternative hypotheses and the methods used:

Question 1: Does the average critic score of video games significantly differ from 7.021, as reported by Metacritic (Giant Bomb, n.d.)?

- **H_0 (Null Hypothesis):** The mean critic score of our dataset is equal to 7.021.
  H_0: μ=7.021
- **H_1 (Alternative Hypothesis):** The mean critic score of our dataset is different from 7.021.
  H_1: $\mu \neq 7.021$
- **Test Used**: One-sample t-test (two-tailed).
- **Rationale**: We compared the sample mean of critic scores from our dataset to the population mean reported by Metacritic. Since the population standard deviation was unknown, we selected the t-test instead of the z-test. This analysis was conducted to assess whether our dataset accurately represented the broader video game population.

Question 2: Do shooter games have significantly higher average total sales than action games?

- **H_0 (Null Hypothesis):** The mean total sales for shooter games are equal to those for action games
  H_0: $\mu\, Shooter = \mu\, Action$
- **H_1 (Alternative Hypothesis):** The mean total sales for shooter games are higher than those for action games.
  H_1: $\mu\, Shooter > \mu\, Action$
- **Test Used**: One-tailed two-sample t-test
- **Rationale**: We compared two independent groups, Shooter and Action games, to determine whether Shooter games generated higher average sales. These genres represent some of the most commercially successful categories in the industry (Statista, 2023), with iconic franchises like Call of Duty and Grand Theft Auto. This one-tailed test was used to assess whether Shooter games significantly outperformed Action games in total sales. A conservative degrees of freedom estimate was applied due to manual calculation, making the test more robust against violations of normality and variance assumptions. We chose to manually compute the t-statistic and confidence interval using the formula from course

material (Samea, 2025) to reinforce our understanding of the underlying statistical process and to validate results independently of automated functions. This comparison provided insights into which genre performs better commercially and could inform future development or publishing strategies.

Question 3: Are North American video game sales significantly higher than European/African sales on average?

- **H_0 (Null Hypothesis):** The mean total sales in North America are equal to those in Europe/Africa.
  H_0: $\mu\ North\ America = \mu\ Europe/Africa$
- **H_1 (Alternative Hypothesis):** The mean total sales in North America are higher than those in Europe/Africa.
  H_1: $\mu\ North\ America > \mu\ Europe/Africa$
- **Test Used**: One-tailed two-sample t-test
- **Rationale**: We manually computed the t-statistic and confidence interval for this hypothesis using the standard formula provided in the course material (Samea, 2025), instead of using the built-in t.test() function in R. This decision was made to reinforce our understanding of the underlying statistical process and to validate results independently of automated tools. The primary difference between the manual approach and t.test() lies in the calculation of the degrees of freedom and one of the confidence intervals. While t.test() by default applies Welch's method, which adjusts degrees of freedom based on unequal variances, our analysis used a conservative approach by applying the smaller sample size minus one (n - 1) as the degrees of freedom. This approach provided a more robust estimate, especially in the presence of potential violations of normality and variance assumptions. Including the manual calculation also promotes transparency and reproducibility, as encouraged in the feedback.

Question 4: Is there a significant linear relationship between Critic Score and Total Sales?

- **H_0 (Null Hypothesis):** There is no linear relationship between critic score and total sales.
  H_0: $\beta_1 = 0$
- **H_1 (Alternative Hypothesis):** There is a significant linear relationship.
  H_1: $\beta_1 \neq 0$
- **Test Used**: Pearson correlation and simple linear regression
- **Rationale**: This test examines whether critic score, a measure of perceived game quality, can predict total sales. In business and marketing analytics, identifying predictive variables is essential for forecasting performance. By using Pearson correlation, we first assess the strength and direction of the linear relationship between critic score and total sales. Then, we apply simple linear regression to determine whether changes in critic score significantly explain variation in total sales.

  The regression coefficient β 1 quantifies the expected change in total sales for a one-unit increase in critic score. If β 1 is statistically different from zero, it suggests a meaningful

predictive relationship. This analysis helps determine whether critic scores, often seen by consumers before purchase, are associated with actual commercial success.

The model follows standard linear regression assumptions and uses the t-distribution for inference, as outlined by Ross (2020) and Samea (2025). This approach allows for statistical testing of whether the slope of the regression line is significantly different from zero, indicating predictive power.

Question 5: Does the Release Year predict Total Sales?

- **H_0 (Null Hypothesis):** There is no linear relationship between release year and total sales.
  H_0: $\beta_1 = 0$
- **H_1 (Alternative Hypothesis)**: Release year significantly predicts total sales.
  H_1: $\beta_1 \neq 0$
- **Test Used:** Simple linear regression
- **Rationale:** This analysis explores whether the release year of a video game is a significant predictor of its total sales using simple linear regression. The goal is to assess whether there's a measurable trend over time, for example, whether newer games tend to sell more (or less) than older ones. This kind of relationship could reflect evolving market dynamics, technological advancements, changing consumer preferences, or the growth of the gaming industry as a whole.
  In the regression model:
  - Release Year (extracted from release_date) is treated as a continuous independent variable (X).
  - Total Sales is the dependent variable (Y).

  The regression coefficient β_1 quantifies the expected change in total sales for each additional year. If β_1 is significantly different from zero, it indicates that the release year has predictive power, either positively or negatively, on sales performance.

  This test uses the t-distribution for inference, in line with standard linear regression assumptions (Ross, 2020; Samea, 2025). By examining the statistical significance of the slope, we determine whether time (as measured by release year) is meaningfully associated with commercial outcomes in the video game industry.

Question 6: Does the number of platforms a game is released on predict Total Sales?

- **H_0 (Null Hypothesis):** There is no linear relationship between the number of platforms and total sales.
  H_0: $\beta_1 = 0$
- **H_1 (Alternative Hypothesis):** The number of platforms significantly predicts total sales.
  H_1: $\beta_1 \neq 0$
- **Test Used:** Simple linear regression
- **Rationale:**This analysis investigates whether releasing a game on multiple platforms leads to greater commercial success. The number of platforms serves as the independent variable (X), and total sales is the dependent variable (Y). Broader platform availability may

increase a game's exposure to different player bases and improve commercial performance, which this analysis aims to evaluate.

We use simple linear regression to test whether this variable significantly predicts sales. If the slope coefficient β_1 is statistically different from zero, it suggests that the number of platforms has a measurable impact on sales. The test follows the assumptions of linear regression and uses the t-distribution for hypothesis testing, as outlined in Ross (2020) and Samea (2025). This provides insight into how platform strategy may influence revenue generation in the gaming industry.

Hypothesis Testing Process

Test 1

Steps followed and the test selected

1. **Select the Test**: A one-sample t-test was chosen because we are comparing the sample mean of critic scores against a known population estimate of 7.021, as reported by Metacritic (Giant Bomb, n.d.).
2. **Performed the One-Sample T-Test:** To determine if the sample mean is significantly different from 7.021.
3. **Code in R:** creating the code to calculate the p-value, T-statistic and confidence interval. The t.test() function was used to perform a one-sample analysis for this hypothesis.
4. **Interpreted the Results:** based on the p-value, T-statistics and the confidence interval.

Assumptions evaluated

Independence: Each observation represents an independent video game release because all duplicates were eliminated and to note that sales data is recorded separately for each region.
Normality: In Figure 1 (Appendix), we observe the distribution of critic scores, which does not follow a normal distribution, as it exhibits clustering and skewness. However, according to the Central Limit Theorem (CLT), for large samples (n>30), the sampling distribution of the mean is approximately normal, even if the original data is not (Ross, 2020). This ensures that the non-normality of critic scores does not significantly impact our hypothesis test, allowing us to proceed with a t-test. Given that our dataset consists of 63,791 observations, the large sample size further reinforces the robustness of the t-test. Additionally, in Figure 7 (Appendix), the Q-Q plot for critic scores further confirms this non-normality, as it highlights the clustering of values around certain scores and deviations in the distribution.

The results

One-Sample T-Test (Critic Score)

Metric	Value
Alpha	0.05
T-critical	±1.96
T-Statistic	107.03
Degrees of Freedom (df)	63790

P-Value	< 2.2e-16
95% Confidence Interval Lower	7.216682
95% Confidence Interval Upper	7.223982
Sample Mean	7.220332

Table 1: One sample t-test for critic score

Based on the p-value which is significantly smaller than the value of alpha (0.05) and that the t-statistic is bigger than the t-critical we proceed to reject the Null hypothesis. This means that the sample mean is significantly different from 7.021. To complement this rejection, the given mean score of Metacritic 7.021 is not in the confidence interval [7.216682, 7.223982], which confirms that with 95% confidence, we reject the null hypothesis and conclude that the true average critic score is significantly different from 7.021.

Test 2

Steps followed and the test selected

1. **Select the Test**: A two-sample t-test (one-tailed) was chosen to compare Shooter vs Action total mean sales. Since we are evaluating that shooter games might have higher sales, we use a one-tailed test.
2. **Performed the Two-Sample T-Test:** To compare the mean sales of shooter against action games. This will help us know if the hypothesis null is rejected or not.
3. **Code in R:** Creating the code to calculate the p-value, T-statistic and confidence interval. For this hypothesis, we performed the two-sample analysis using the formula to calculate the t-statistic and confidence intervals instead of the t.test() function. It is important to note that the manual calculation and the t.test() function differ only in how they determine the degrees of freedom and one of the confidence intervals. By default, t.test() uses Welch's t-test to calculate the degrees of freedom, whereas in this analysis, we used the method of taking the smallest sample size minus one (n−1).
   This manual approach aligns with the course methodology (Samea, 2025) and helps reinforce conceptual understanding while making the test more transparent and robust to potential assumption violations.
   The formula for manual calculation of the t statistic and the confidence intervals can be found in the appendix as equation 1 and equation 2.
4. **Interpreted the Results:** based on the p-value, T-statistics and the confidence interval.

Assumptions evaluated

Independence: Each game has independent total sales. Since all the duplicates were eliminated, this makes each game is uniquely represented.

Normality: In Figures 2 and 3 (Appendix), the histograms reveal that action sales and shooter sales do not follow a normal distribution, as they exhibit right-skewness with a large number of low-sales games and a few high-sales outliers. However, due to the Central Limit Theorem (CLT) and the large sample size (n>30), the t-test remains robust, even if the individual data points are not normally distributed, as noted by Ross (2020). The t-test assumes normality, but for large samples, CLT

ensures that the sampling distribution of the mean approximates normality, making the test valid. Additionally, in Figures 8 and 9 (Appendix), the Q-Q plots for action and shooter sales further confirm this non-normality, as they show significant deviations from the theoretical normal line, particularly in the upper quantiles.

The results

Two sample t-test for shooter and action games

Metric	Value
T-statistic	5.528977
T-critical	1.645136
Degrees of Freedom (df)	5395
p-value	1.69E-08
95% Confidence Interval	[0.05080999, 0.09385468]
Mean of Shooter Sales	0.4376774
Mean of Action Sales	0.365345

Table 2: Two sample t-test for shooter and action games

Based on the p-value which is significantly smaller than the value of alpha and that the t-statistic is far bigger than the t-critical we proceed to reject the Null hypothesis. This means the difference in sales between Shooter and Action games is statistically significant. To complement this rejection, the confidence interval [0.0508, 0.0938], is entirely positive, this confirms that the difference is statistically significant. We can conclude that, we are 95% confident that the true difference in mean sales between Shooter and Action games is between 0.0508 and 0.0938 units.
In addition, figure 4 from the appendix, we can see how the mean of both video game genres compares. We can see that both genres have significant differences in mean sales.

Test 3

Steps followed and the test selected

1. **Select the Test:** Two-sample t-test (one-tailed) was chosen because we are comparing the average sales between two independent regional markets. We are assuming North America mean sales might be significantly higher than their counterparts in Europe and African mean sales. Since we are evaluating that shooter games might have higher sales, we use a one-tailed test.
2. **Performed the Two-Sample T-Test:** To compare the mean sales of North America with the mean sales of Europe and Africa combined.
3. **Code in R:** Creating the code to calculate the p-value, T-statistic and confidence interval. For this hypothesis, we performed the two-sample analysis using the formula to calculate the t-statistic and confidence intervals instead of the t.test() function. It is important to note that the manual calculation and the t.test() function differ only in how they determine the degrees of freedom and one of the confidence intervals. By default, t.test() uses Welch's t-

test to calculate the degrees of freedom, whereas in this analysis, we used the method of taking the smallest sample size minus one (n−1).
This manual calculation was chosen to mirror the formulas taught in class (Samea, 2025), promote a deeper grasp of the underlying logic, and enhance the robustness of the test when data conditions are not ideal.
The formula for manual calculation for the t statistic and the confidence intervals can be found in the appendix as equation 1 and equation 2.

4. **Interpreted the Results:** based on the p-value, T-statistics and the confidence interval.

Assumptions checked

Independence: Each game has for each specific region has independent total sales. Since all the duplicates were eliminated, this makes each game is uniquely represented.

Normality: In Figures 5 and 6 (Appendix), we see that sales in North America and Europe/Africa do not follow a normal distribution. However, due to the Central Limit Theorem (CLT) and the large sample size (n>30), we can ensure the robustness of the t-test. The t-test generally assumes normality, but for large sample sizes (n>30), CLT ensures that the test remains valid even if the data is not normally distributed as noted by Ross (2020). The correlation between sales in both regions was calculated, yielding r=0.6660649, indicating a moderate to strong positive correlation between North American and PAL (Europe/Africa) sales. This means that games that sell well in North America tend to also sell well in Europe & Africa. However, correlation does not imply causation, nor does it indicate dependency between sales in both regions. Each game's sales in different regions remain separate observations, so we can still use the t-test. Additionally, in Figures 10 and 11 (Appendix), the Q-Q plots confirm once again that North American and Europe/Africa sales do not follow a normal distribution.

The results

Two sample t-test for North American and Europe/African sales

Metric	Value
T-statistic	103.301
T-critical	1.644878
Degrees of Freedom (df)	63790
P-Value	0
95% Confidence Interval	[0.1134602, 0.1171319]
Mean of North America Sales	0.2647974
Mean of Europe/Africa Sales	0.1495014

Table 3: Two sample t-test for North American and Europe/African sales

Based on the p-value which is smaller than the value of alpha and that the t-statistic is a very high value compared to the t-critical, so we proceed to reject the Null hypothesis. This means that there is strong statistical evidence that North America sales are significantly higher than Europe/Africa sales. To complement this rejection, the confidence interval [0.1135, 0.1171], is entirely positive, this means that, at a 95% confidence level, the true difference in means is between 0.1135 and 0.1171.

Test 4

Steps followed and the test selected

1. **Select the Test**: A simple linear regression was chosen to analyze the relationship between critic score and total sales.
2. **Performed the Linear Regression**: The goal was to determine whether critic score significantly predicts total sales. This was done using the lm() function in R, modeling total sales as a function of critic score.
3. **Code in R**: We created the regression model and calculated the slope coefficient, p-value, and R-squared value. A scatter plot with a regression line was also created to visualize the relationship.
   Additionally, a residuals vs. fitted values plot was generated to visually check for potential issues like heteroscedasticity or non-linearity.
4. **Interpreted the Results**: We interpreted the regression output based on the slope estimate, the associated p-value for the coefficient, and the residual distribution. A significant p-value indicates that critic score has predictive power over total sales. The regression model follows standard linear regression assumptions and uses the t-distribution for inference, as outlined by **Ross (2020)** and **Samea (2025)**. This ensures the validity of hypothesis testing on the regression slope.

Assumptions checked

Independence: Each game has for each specific region has independent total sales. Since all the duplicates were eliminated, this makes each game is uniquely represented.

***Homoscedasticity*:** To check for homoscedasticity in the simple linear regression model, we generated a residuals vs. fitted values plot (Figure 12 in the appendix). The plot shows a fan-shaped pattern where residuals spread out more as fitted values increase. This suggests heteroscedasticity, indicating that the variance of the residuals is not constant across the range of predicted values—thus violating a core assumption of linear regression (Ross, 2020; Samea, 2025). While the model may still provide a general sense of direction in the relationship, this violation suggests that standard errors and significance tests should be interpreted with caution.

The results

Regression Coefficients: Effect of Critic Score on Total Sales

Predictor	Estimate	Standard Error	t value	p-value	Significance
(Intercept)	-0.75509	0.02641	-28.59	<2e-16	***
critic_score	0.15296	0.00365	41.9	<2e-16	***

Table 4: Regression Coefficients: Effect of Critic Score on Total Sales

Model Summary for Simple Linear Regression: Critic Score Predicting Total Sales

Metric	Value

Residual Std. Error	0.4337
Degrees of Freedom (Error)	63789
Multiple R-squared	0.02679
Adjusted R-squared	0.02677
F-statistic	1756
DF (Regression, Residual)	1, 63789
p-value	< 2.2e-16

Table 5: Model Summary for Simple Linear Regression: Critic Score Predicting Total Sales

Analyzing table 4 and 5, the regression coefficient for critic_score is positive and statistically significant, indicating that as critic score increases, total sales are expected to increase as well. Specifically, for each 1-point increase in critic score, total sales increase by approximately 0.153 million units on average. Therefore, we reject the null hypothesis and conclude that critic score is a statistically significant predictor of total sales.

Despite the statistical significance, the R-squared value is low (2.7%), suggesting that critic score explains only a small portion of the variance in total sales. This means other factors likely play a much larger role in determining game sales.
Even though the model shows statistical significance, the low explanatory power (low R^2) and presence of heteroscedasticity suggest we should interpret the results with caution. The critic score does have a positive impact on sales, but its influence is small in practical terms.

Scatterplot of Critic Score vs. Total Sales with Fitted Regression Line

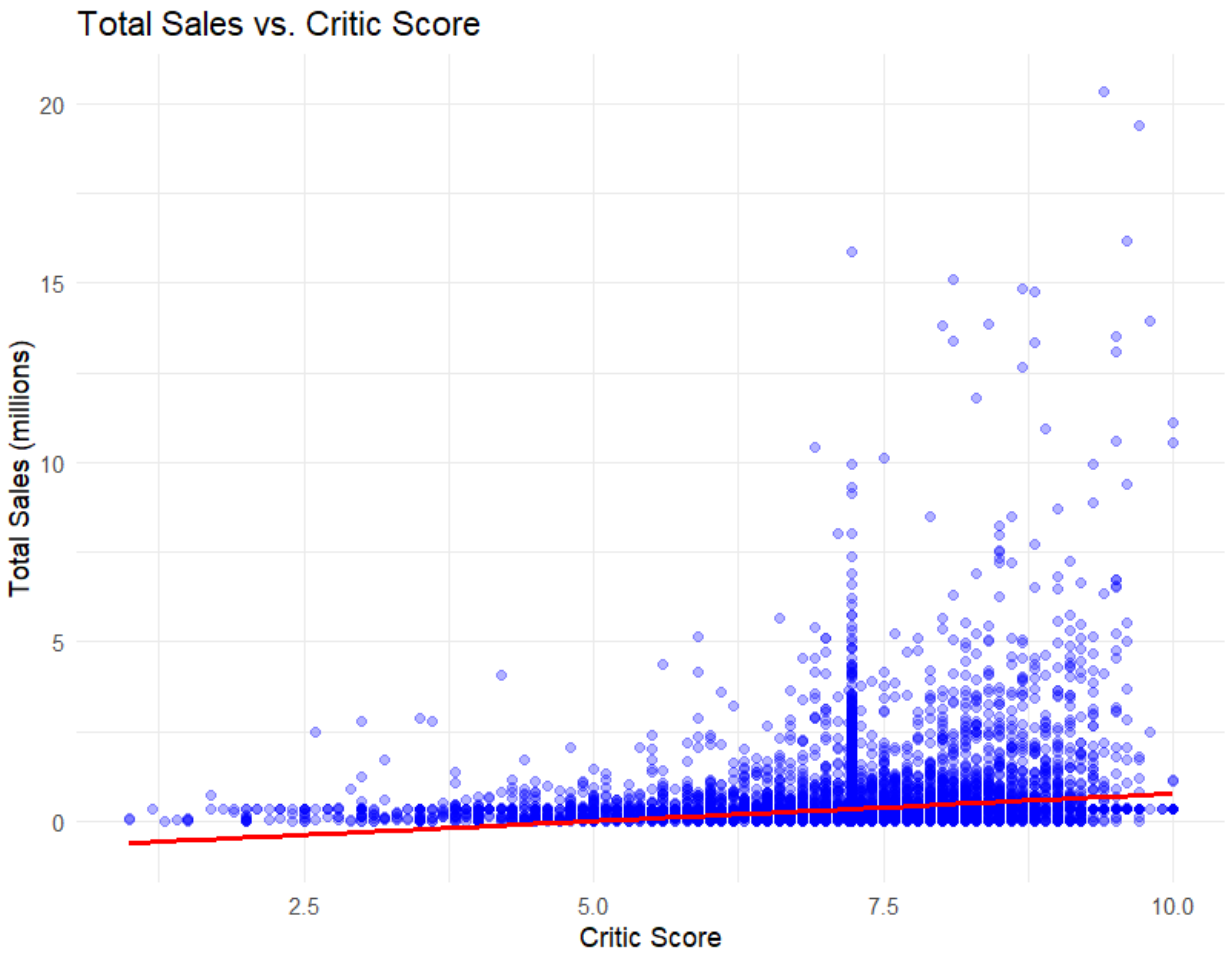

Figure 13: Scatterplot of Critic Score vs. Total Sales with Fitted Regression Line

The scatter plot shows the relationship between critic scores and total sales for video games in the dataset. The fitted regression line (in red) suggests a slight positive linear trend, indicating that higher critic scores are generally associated with higher total sales. However, the relationship is weak. The plot reveals significant variability in sales at nearly every critic score level, with many games achieving modest sales despite strong reviews, and only a few high-scoring titles reaching blockbuster status. Most data points are clustered in the 7.0 to 8.0 score range, aligning with earlier observations about critic score distribution. Additionally, several high-sales outliers pull the regression line upward, but the overall spread indicates that critic score alone is not a strong predictor of total sales. This visual evidence supports the statistical findings, showing a significant but practically weak relationship between the two variables.

Test 5

Steps followed and the test selected

1. **Select the Test**: A simple linear regression was chosen to examine the relationship between a video game's release year and its total sales.

2. **Performed the Linear Regression**: We modeled total sales as a function of release year using the lm() function in R. This allows us to determine whether the year a game was released significantly predicts how much it sold.
3. **Code in R**: The model was constructed and summarized to extract the slope coefficient, R-squared value, and p-value. We also created:
    - A scatter plot with a fitted regression line to visualize the trend.
    - A residuals vs. fitted values plot to check for linearity and homoscedasticity.
4. **Interpreted the Results**: We examined the regression output and diagnostic plots. The regression inference followed the assumptions and procedures taught by **Ross (2020)** and **Samea (2025)**, using the t-distribution to evaluate whether the slope significantly differs from zero.

Assumptions checked

Independence: Each game has for each specific region has independent total sales. Since all the duplicates were eliminated, this makes each game is uniquely represented.

Homoscedasticity: To assess homoscedasticity in the regression model for Hypothesis 5, we generated a residuals vs. fitted values plot (Figure 14 in the appendix). The residuals appear relatively evenly scattered around the horizontal axis, with no pronounced fan shape or increasing spread. This indicates that the variance of residuals remains roughly constant, and the assumption of homoscedasticity is reasonably satisfied (**Ross, 2020; Samea, 2025**). This supports the reliability of the regression inference, even if the predictor is ultimately not significant.

The results

Regression Coefficients: Effect of Release Year on Total Sales

Predictor	Estimate	Standard Error	t value	p-value	Significance
(Intercept)	0.76835	0.25378	3.028	0.00247	**
Release Year	-0.00021	0.00013	-1.651	0.09868	.

Table 7: Regression Coefficients: Effect of Release Year on Total Sales

Model Summary for Simple Linear Regression: Release Year Predicting Total Sales

Metric	Value
Residual Standard Error	0.4396
Degrees of Freedom (Error)	63789

Multiple R-squared	4.28E-05
Adjusted R-squared	2.71E-05
F-statistic	2.727
DF (Regression, Residual)	1, 63789
p-value	0.09868

Table 8: Model Summary for Simple Linear Regression - Release Year Predicting Total Sales

Analyzing the regression output in Tables 7 and 8, the coefficient for *Release Year* is slightly negative (-0.00021), suggesting that more recent games tend to have marginally lower total sales. However, this relationship is not statistically significant, as the p-value for the slope is 0.09868, greater than the 0.05 significance threshold. Consequently, we fail to reject the null hypothesis, meaning that *Release Year* does not significantly predict total sales in this dataset. Furthermore, the R-squared value is extremely low (0.00004), indicating that *Release Year* explains less than 0.01% of the variation in sales, essentially no predictive power.

Regarding the assumptions of the linear regression model, the residuals vs. fitted values plot for this hypothesis does not show strong signs of heteroscedasticity. Residuals are relatively evenly scattered without a pronounced fan shape or funneling, suggesting that the variance of residuals remains relatively constant across predicted values. This supports the assumption of homoscedasticity and the reliability of the inference procedure. However, despite this assumption being reasonably satisfied, the overall conclusion remains that *Release Year* is not a meaningful standalone predictor of total sales in this dataset.

Scatterplot of Total Sales vs. Release Year with Fitted Regression Line

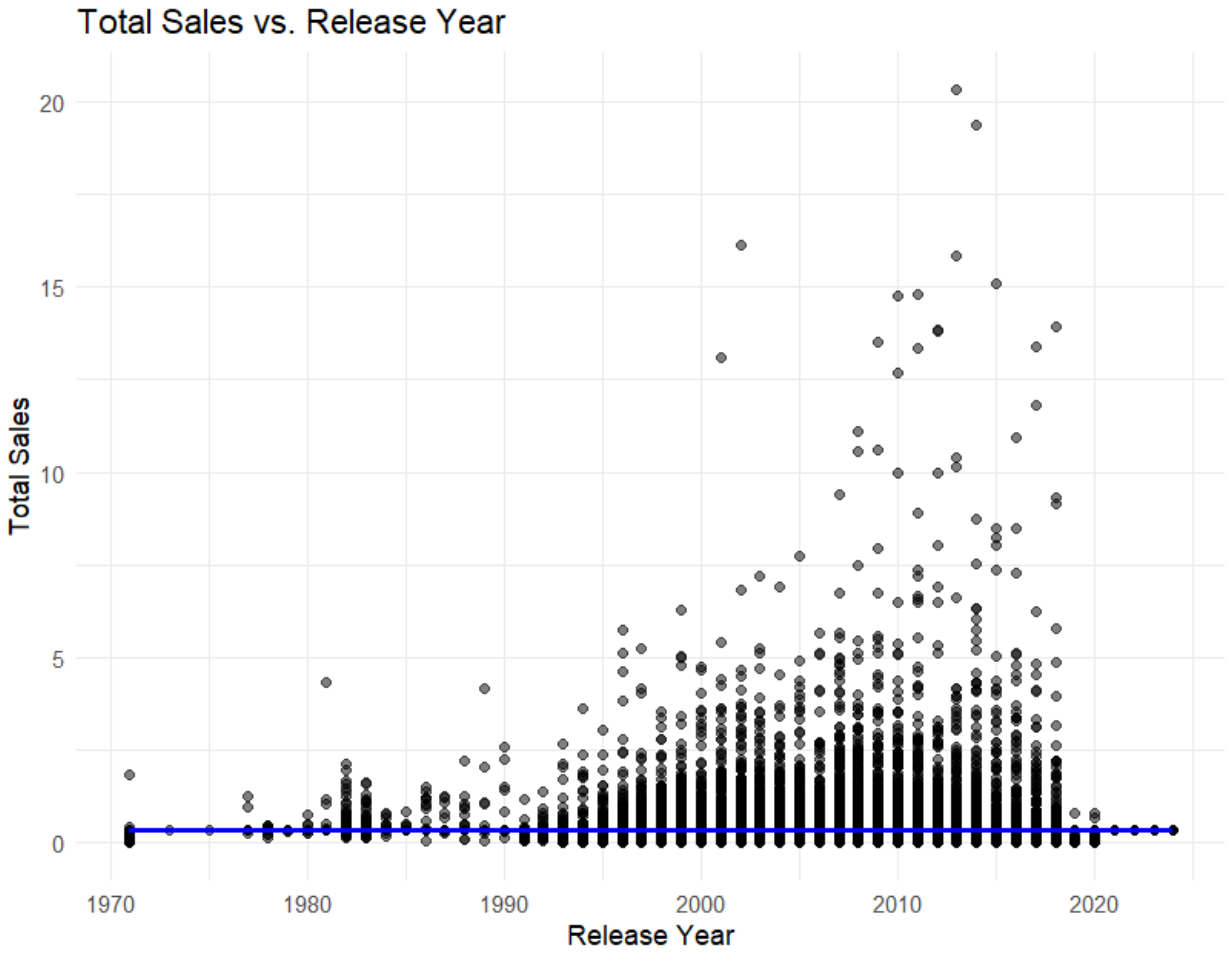

Figure 15: Scatterplot of Total Sales vs. Release Year with Fitted Regression Line

Figure 15, displays the relationship between the release year of a video game and its total sales. The scatterplot includes a fitted regression line in blue, which appears nearly flat, indicating a weak or negligible linear relationship between the two variables. While sales appear to increase slightly for games released between 1980 and 2010, there is a sharp decline after 2015. This likely reflects incomplete data for newer games that have not accumulated lifetime sales yet or are underrepresented in the dataset.

Overall, the distribution is heavily right-skewed, with a large cluster of low-selling games in nearly every year and only a few high-selling outliers. The lack of a strong upward or downward slope in the regression line visually supports the statistical findings, which showed that release year is not a statistically significant predictor of total sales. This scatterplot confirms the conclusion that time alone does not explain commercial success in the video game market.

Test 6

Steps followed and the test selected

1. **Select the Test**: A simple linear regression was chosen to examine whether the number of platforms a video game is released on significantly predicts its total sales.
2. **Performed the Linear Regression**: Using the lm() function in R, we modeled total sales as a function of the number of unique platforms per game. This allows us to evaluate if broader platform availability is associated with higher commercial performance.
3. **Code in R**: We grouped the dataset by game title and calculated how many platforms each game was released on. Then we ran a regression model and created:
    - A scatterplot with a fitted regression line.
    - A residuals vs. fitted values plot to assess homoscedasticity and potential model violations.
4. **Interpreted the Results**: We assessed the coefficient, p-value, and R-squared value. The analysis followed the inference framework based on linear regression assumptions and hypothesis testing procedures discussed in Ross (2020) and Samea (2025).

Assumptions checked

Independence: Each game has for each specific region has independent total sales. Since all the duplicates were eliminated, this makes each game is uniquely represented.

Homoscedasticity: The residuals vs. fitted values plot (Figure 16 in the appendix) shows a relatively consistent vertical spread of residuals, without a strong funnel shape or trend in residual variance. Despite some clustering at lower values and slight tapering at higher fitted values, the pattern appears generally even. We therefore conclude that homoscedasticity is reasonably satisfied, validating the regression results (Ross, 2020; Samea, 2025).

The results

Regression Coefficients: Effect of Platform Count on Total Sales

Predictor	Estimate	Standard Error	t value	p-value	Significance
(Intercept)	0.27679	0.00277	100.01	<2e-16	***
Platform Count	0.02803	0.00084	33.51	<2e-16	***

Table 10: Regression Coefficients: Effect of Platform Count on Total Sales

Model Summary for Simple Linear Regression: Platform Count Predicting Total Sales

Metric	Value
Residual Standard Error	0.4358

Degrees of Freedom (Error)	63789
Multiple R-squared	0.0173
Adjusted R-squared	0.01728
F-statistic	1123
DF (Regression, Residual)	1, 63789
p-value	< 2.2e-16

Table 11: Model Summary for Simple Linear Regression: Platform Count Predicting Total Sales

Analyzing the regression output in Tables 9 and 10, the coefficient for Platform Count is positive and statistically significant ($p < 2.2e-16$), indicating that the number of platforms a game is released on significantly predicts its total sales. Specifically, for each additional platform a game is available on, total sales are expected to increase by approximately 0.028 million units. Given the extremely low p-value, we reject the null hypothesis and conclude that platform count is a statistically significant predictor of total sales.

However, while the relationship is statistically significant, the model's practical explanatory power is modest. The R-squared value is only 0.0173, suggesting that platform count explains just 1.7% of the variance in total sales. This implies that although broader platform availability is associated with higher sales, it accounts for only a small portion of overall sales performance. Many other factors—such as genre, marketing, game quality, or release timing—likely play more substantial roles in driving commercial success.

Scatterplot of Total Sales vs. Number of Platforms with Fitted Regression Line

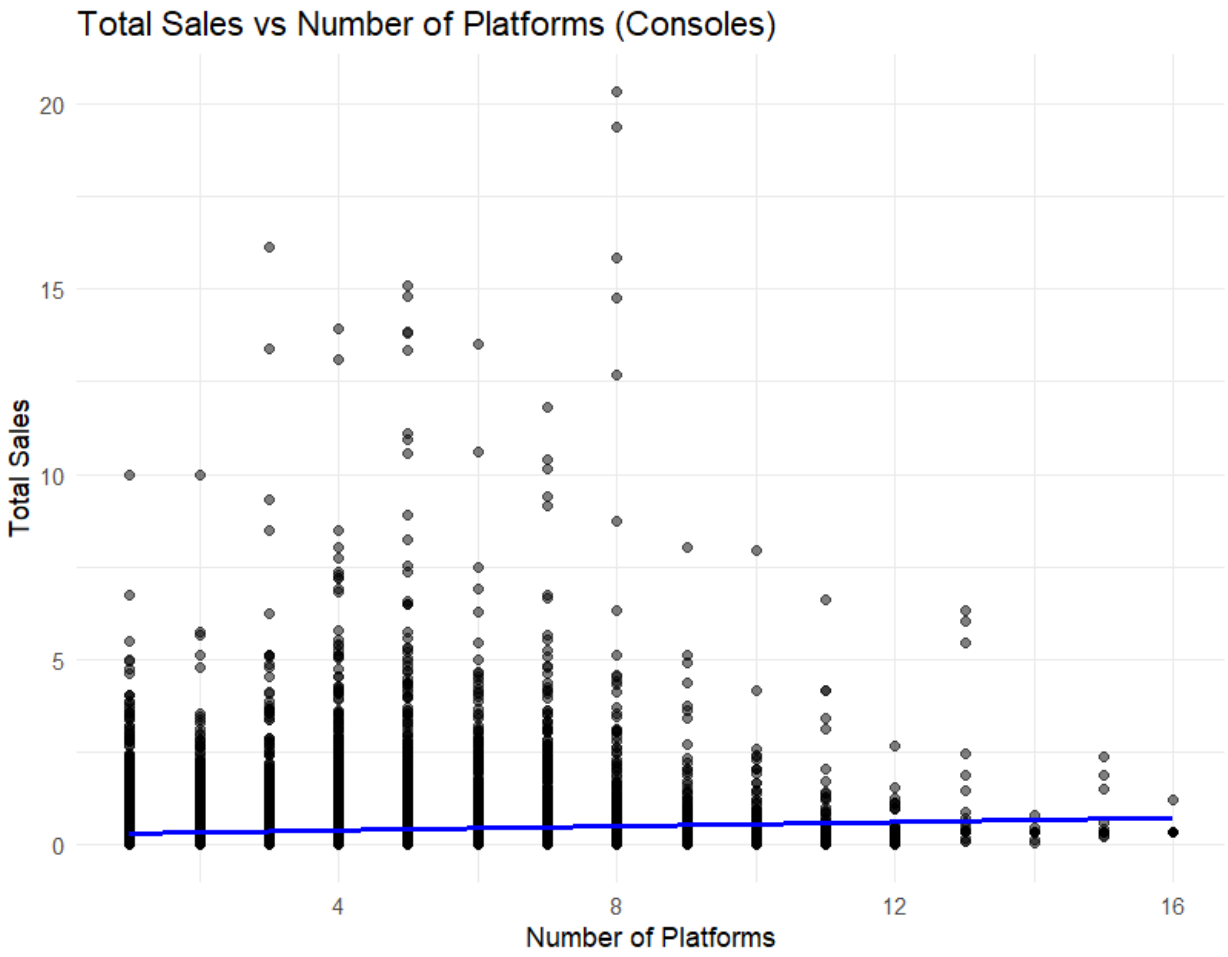

Figure 17: Scatterplot of Total Sales vs. Number of Platforms with Fitted Regression Line

The scatterplot illustrates the relationship between the number of platforms a video game is released on and its total sales. Each point represents an individual game, showing how sales vary across different platform counts. A fitted regression line (in blue) indicates a slight positive linear trend, suggesting that games released on more platforms tend to achieve higher total sales.

However, this trend is weak and the data exhibits substantial variability. There is a large cluster of games with low sales across all platform counts, and a few high-selling outliers particularly in the mid-range of 3 to 9 platforms. Beyond 10 platforms, the number of games decreases sharply, and sales remain modest or even drop. This pattern implies that while broader platform availability can increase a game's reach and potential audience, it is not a guaranteed path to commercial success. Other factors such as game quality, genre, and marketing likely play larger roles. Overall, the plot visually supports the statistical conclusion that the number of platforms has a significant, but limited, impact on total sales.

Results and Interpretation

Hypothesis 1

The results of the one-sample t-test indicate that the average critic score in our dataset (7.2203) is significantly higher than the global Metacritic average (7.021). The test produced a very high t-statistic (107.03), and the p-value (< 2.2e-16) confirms that this difference is statistically significant. The 95% confidence interval ([7.2167, 7.2240]) does not include 7.021, reinforcing that the true mean critic score in our dataset differs from the global estimate.

This finding suggests that games in our dataset tend to have slightly higher ratings compared to the general population of video games. There could be several reasons for this: our dataset might include more high-profile games, which naturally receive better reviews, or there may be platform or genre biases affecting critic scores. While the difference is statistically significant, it is relatively small in practical terms. This could indicate selection bias, meaning that the dataset does not perfectly represent all games reviewed on Metacritic. Future research could explore whether specific genres or platforms tend to receive higher scores and whether regional influences affect critic ratings.

Hypothesis 2

The results of the two-sample t-test confirm that Shooter games have significantly higher total sales than Action games. The mean Shooter sales (0.4377) were notably higher than Action sales (0.3653), and the t-statistic (5.529) is well above the critical threshold, with a p-value of 1.69E-08, making the result highly statistically significant. The confidence interval ([0.0508, 0.0939]) is entirely positive, confirming that this difference is real and not due to random chance.

This suggests that Shooter games tend to be more commercially successful than Action games. A few possible explanations for this trend include the strong franchise loyalty associated with Shooter games (e.g., Call of Duty, Battlefield), higher player engagement in multiplayer shooters, and stronger esports or online gaming communities, which drive sustained sales. In contrast, Action games are a broader category, including many niche or indie titles that may not perform as well commercially. The results indicate that developers and publishers may find Shooter games to be a more profitable genre on average, but further analysis could explore whether specific publishers or platforms contribute disproportionately to this trend.

Hypothesis 3

The two-sample t-test comparing North American sales to Europe/Africa (PAL) sales confirms that North America has significantly higher video game sales. The mean North America sales (0.2648) are substantially higher than Europe/Africa sales (0.1495), and the t-statistic (103.301) is extraordinarily large. The p-value of 0, zero, means that this difference is highly statistically significant. The confidence interval ([0.1135, 0.1171]) is entirely positive, which confirms that the true mean sales difference between these regions is at least 0.1135 units and at most 0.1171 units.

This result suggests that North America is the dominant market for video game sales compared to Europe/Africa. Several factors could contribute to this: higher console ownership rates, stronger gaming culture, and more aggressive marketing strategies in North America. Additionally, certain

console brands (like Xbox) are more popular in North America, which may drive higher sales in this region. While Europe & Africa (PAL) sales are still significant, they lag North America, reinforcing the idea that the North American market is more lucrative for game developers and publishers. Future research could investigate whether certain game genres or publishers contribute more significantly to this regional sales gap.

Hypothesis 4

The results of the simple linear regression analysis suggest that there is a statistically significant relationship between critic scores and total sales of video games. The p-value for the slope coefficient is less than 2e-16, providing strong evidence against the null hypothesis and supporting the conclusion that critic scores do have a measurable effect on sales. Specifically, each one-point increase in critic score is associated with an estimated increase of approximately 0.153 million units in total sales.
Therefore, we reject the null hypothesis and conclude that critic score is a statistically significant predictor of total sales.

However, despite this statistical significance, the practical impact of the relationship is limited. The R-squared value is just 0.02679, meaning that critic score explains only about 2.7% of the variance in total sales. This low explanatory power suggests that other factors likely play a much more substantial role in driving commercial success.

The scatter plot(Figure 13) supports this conclusion, showing a weak upward trend with a wide spread of sales values across all critic score levels. High-sales outliers do exist, but the majority of titles, even those with strong reviews, exhibit modest sales figures. Furthermore, the residuals vs. fitted values plot indicates heteroscedasticity, where residual variance increases with higher fitted values. This violation of the constant variance assumption (homoscedasticity) implies that standard errors may be biased, and that inference should be made with caution.

Overall, while critic scores do show a statistically significant relationship with total sales, they are not a strong standalone predictor of a game's commercial performance. This underscores the importance of incorporating other relevant predictors, such as marketing budget, platform availability, release timing, or franchise recognition, in future models to better understand and forecast video game sales.

Hypothesis 5

The results of the simple linear regression analysis indicate that release year is not a statistically significant predictor of total video game sales. The p-value for the slope coefficient (0.09868) exceeds the conventional 0.05 significance threshold, meaning we fail to reject the null hypothesis. Although the coefficient is slightly negative, suggesting that newer games may sell marginally less on average, this effect is not statistically meaningful. Furthermore, the R-squared value is extremely low (0.0000428), indicating that release year explains virtually none of the variability in total sales, less than 0.01%.

Visual evidence supports these findings. The scatterplot (Figure 15) shows a weak or flat trend line, with little indication of a strong upward or downward trajectory. While there appears to be a slight rise in sales between 1980 and 2010, followed by a sharp decline post-2015, this drop may reflect

incomplete data for newer titles that haven't had enough time to accumulate sales. The residuals vs. fitted values plot (Figure 14) supports the assumption of homoscedasticity, as there is no clear funnel shape or increasing spread of residuals. This means the variance of errors remains relatively constant, validating the inference procedure from a statistical perspective. However, despite meeting this key assumption, the overall conclusion remains: release year, on its own, does not offer any meaningful predictive power regarding the commercial success of a video game.

Hypothesis 6

The results of the simple linear regression analysis indicate that the number of platforms a game is released on is a statistically significant predictor of total video game sales. The slope coefficient for Platform Count is 0.02803, with a p-value less than 2.2e-16, strongly rejecting the null hypothesis. This means that for each additional platform a game is available on, its total sales are expected to increase by approximately 0.028 million units, or 28,000 copies, on average. This finding supports the idea that multi-platform releases can help expand a game's reach and boost its commercial performance.

However, the practical explanatory power of this model is limited. The R-squared value is only 0.0173, indicating that platform count explains just 1.7% of the variance in total sales. This suggests that while broader availability does have a measurable effect, it is far from the dominant factor influencing sales outcomes. Many other elements, such as game genre, marketing budgets, critical reception, or brand recognition, are likely to play a much greater role in determining commercial success. The residuals vs. fitted values plot (Figure 16) supports the assumption of homoscedasticity, with a reasonably uniform spread of residuals across fitted values. This validates the inference drawn from the model, but the modest R-squared reminds us to interpret these results in the broader context of other important predictors.

Conclusion

The analysis revealed several statistically significant findings. First, the average critic score in the dataset was significantly higher than the global average, suggesting potential dataset bias toward more prominent titles. Shooter games outperformed Action games in total sales, and North American sales were significantly higher than those in the PAL region, reaffirming the region's market dominance. Critic score and platform count were statistically significant predictors of total sales, although both had limited explanatory power, with low R^2 values. Interestingly, release year did not significantly predict sales, highlighting that timing alone does not guarantee commercial success.

Across all tests, we ensured that assumptions were met or that violations were clearly acknowledged, particularly in the presence of heteroscedasticity or skewed distributions. The results underscore the complexity of the video game market, where no single factor dominates sales outcomes. Success appears to be multifactorial, driven by combinations of genre, marketing, critical reception, regional demand, and platform strategy.

Real world implications

The findings from this report have meaningful implications for stakeholders in the gaming industry:

- **Developers and Publishers** can use genre insights to prioritize shooter or high-engagement genres when aiming for commercial success, especially in Western markets.
- **Marketing Teams** can leverage critic scores as part of their promotional strategy, even if the direct sales impact is modest, as positive reviews still correlate with better performance.
- **Platform Strategy** decisions can be guided by the evidence that broader availability contributes to higher sales—supporting multi-platform releases when feasible.
- **Regional Sales Insights** reinforce the importance of tailoring strategies to dominant markets such as North America, which consistently outperforms others in terms of revenue potential.
- **Data Analysts and Researchers** can build on this foundational work by incorporating additional variables such as user ratings, marketing spend, franchise history, or online engagement to improve sales prediction models.

Ultimately, this analysis demonstrates how statistical tools can be applied to guide data-driven decisions in a competitive, billion-dollar industry where understanding the drivers of success is critical to sustainability and growth.

Reference

1. Giant Bomb. (n.d.). *Metacritic game data analysis (too many pics?)* General Discussion. Retrieved March 19, 2025, from https://www.giantbomb.com/forums/general-discussion-30/metacritic-game-data-analysis-too-many-pics-1495423/
2. Samea, A. (2025a). *ALY 6010: Probability theory and introductory statistics* [Course PDF]. Department of College of Professional Studies, Northeastern University.
3. OpenAI. (2025, March 18). *ChatGPT response to research query on how to calculate correlation between two variables in R Studio*. OpenAI. https://www.openai.com/chatgpt
4. Ross, S. M. (2020). *Introductory statistics* (5th ed.). Academic Press.
5. Statista. (2023). *Most popular video game genres worldwide in 2023*. https://www.statista.com
6. Samea, A. (2025b). *Lesson 5.4: Linear regression hypothesis testing* [Course webpage]. In *ALY 6010: Probability theory and introductory statistics*. Northeastern University. https://northeastern.instructure.com/courses/200605/pages/lesson-5-4-linear-regression-hypothesis-testing?module_item_id=11680918

Appendix

Two-Sample t-Statistic Formula

$$t = ((X_1 - X_2) - (\mu_1 - \mu_2)) / sqrt((s_1^2 / n_1) + (s_2^2 / n_2))$$

Equation 1: Two-sample t-statistic formula

Where:

$\bar{X}_1$ and $\bar{X}_2$ are the sample means.

μ_1 and μ_2 are the means of the populations.

s_1 and s_2 are the sample standard deviations.

n_1 and n_2 are the sample sizes.

Equation 1 tell us how to calculate manually how to calculate the t-statistic for two samples, where:

Confidence Interval Formula for Two-Sample T-Test

$$(X_1 - X_2) \pm t(\alpha) * sqrt((s_1^2 / n_1) + (s_2^2 / n_2))$$

Equation 2: Confidence interval formula for two-sample t-test

Where:

$\bar{X}_1$ and $\bar{X}_2$ are the sample means.

t(alpha) is the t-critical based on the value of alpha.

s_1 and s_2 are the sample standard deviations.

n_1 and n_2 are the sample sizes.

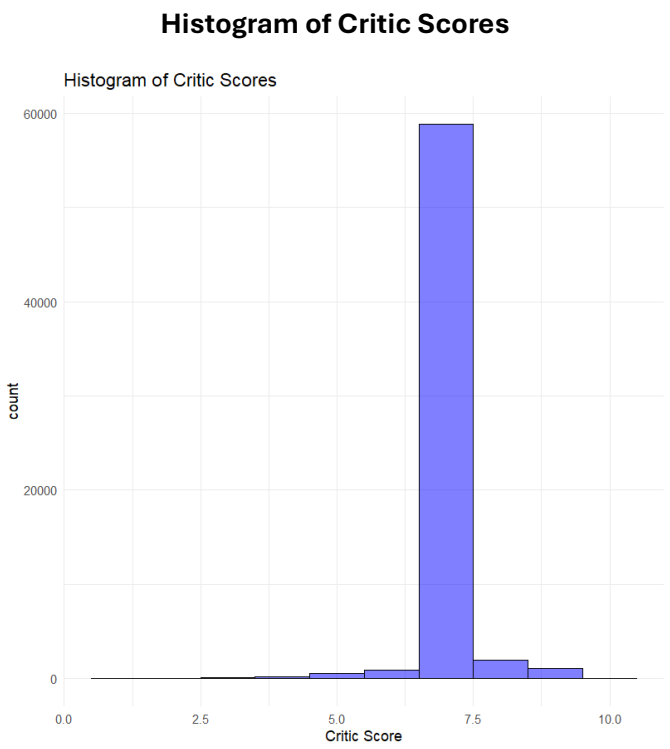

Figure 1: Histogram of critic score

Figure 1 shows the distribution of critic scores, which is highly skewed and does not follow a normal distribution. Most of the scores are concentrated around **7.5**, with very few observations at other values. Despite this non-normality, the **Central Limit Theorem (CLT)** ensures that for large sample sizes, the sampling distribution of the mean remains approximately normal. Therefore, the non-normality of critic scores does not significantly impact the validity of our hypothesis test.

Histogram of Shooter total sales

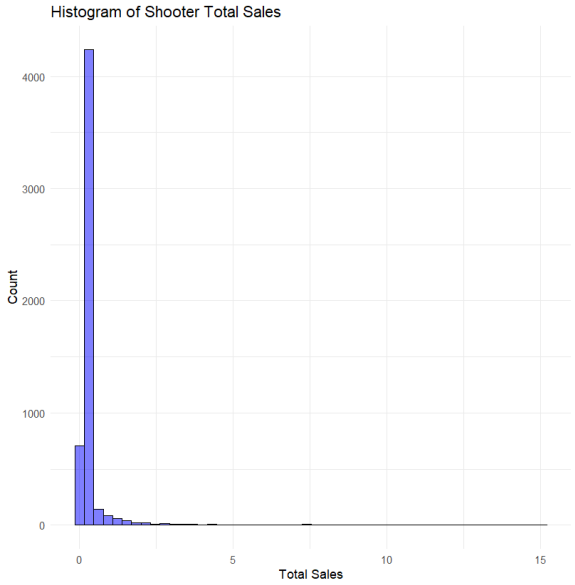

Figure 2: Histogram of Shooter sales

Figure 2 displays the sales total distribution of shooter games. The distribution is non-normal, supporting the use of the t-test under the CLT assumption.

Hisotgram of Action total sales

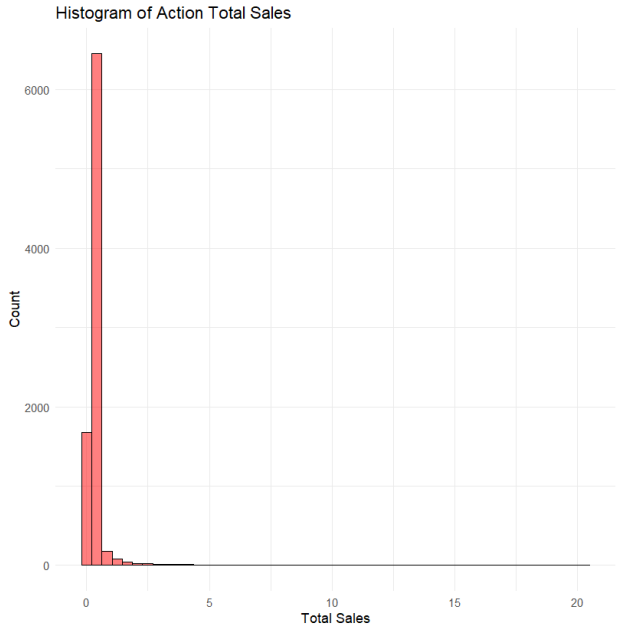

Figure 3: Histogram of Action sales

Figure 3, displays the sales distribution for Action video games. The deviate from normality, indicating that the data is not perfectly normally distributed. However, due to the Central Limit Theorem (CLT), the t-test remains robust for large sample sizes.

Comparison of mean sales between Shooter and Action games

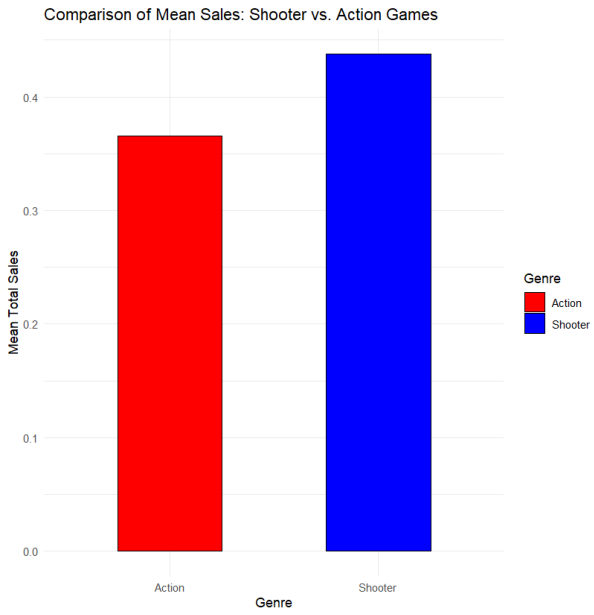

Figure 4: Comparison of Mean Sales Between Shooter and Action Games

Figure 4 visually compares the mean sales of Shooter and Action games. The chart shows that Shooter games have higher average sales than Action games. This aligns with the t-test results, which indicate a statistically significant difference in mean sales between the two genres.

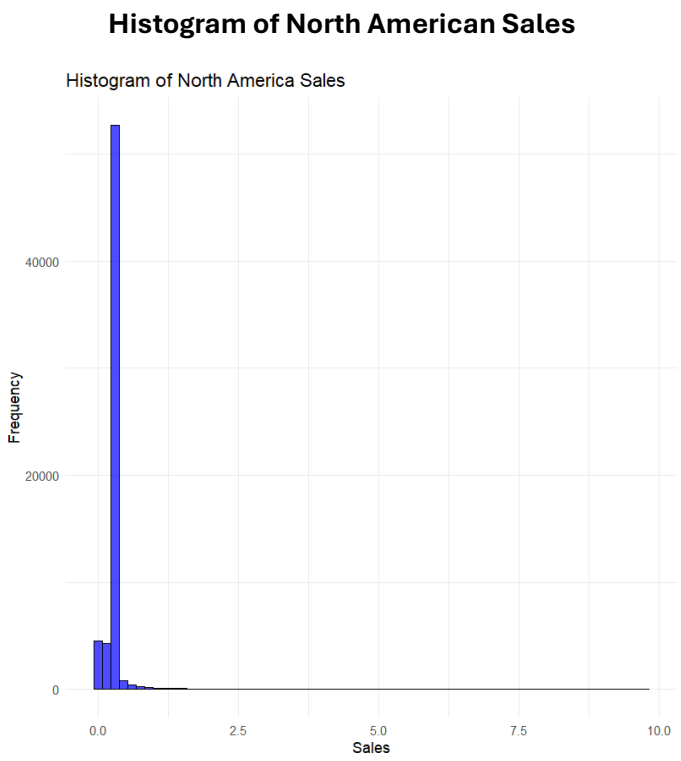

Figure 5: Histogram of North American sales

This histogram represents the distribution of video game sales in North America. The data appears right-skewed, suggesting that sales do not follow a normal distribution. However, due to the Central Limit Theorem, we can still apply the t-test for large sample sizes.

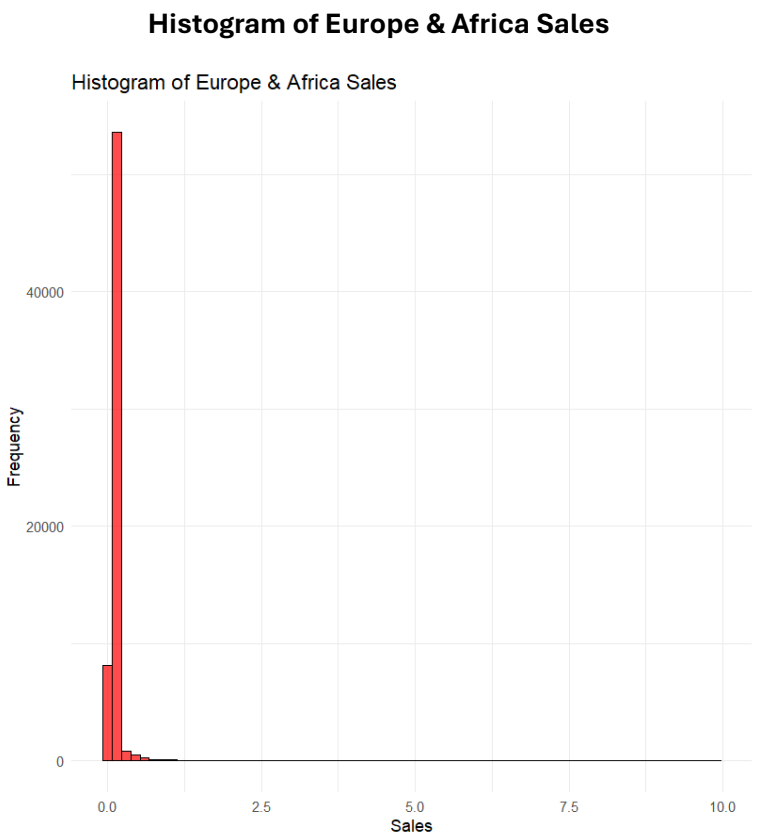

Figure 6: Histogram of Europe & African sales

This figure displays the sales distribution for the combined Europe and Africa regions. Similar to North America, the distribution is non-normal, supporting the use of the t-test under the CLT assumption.

Q-Q plot for critic scores

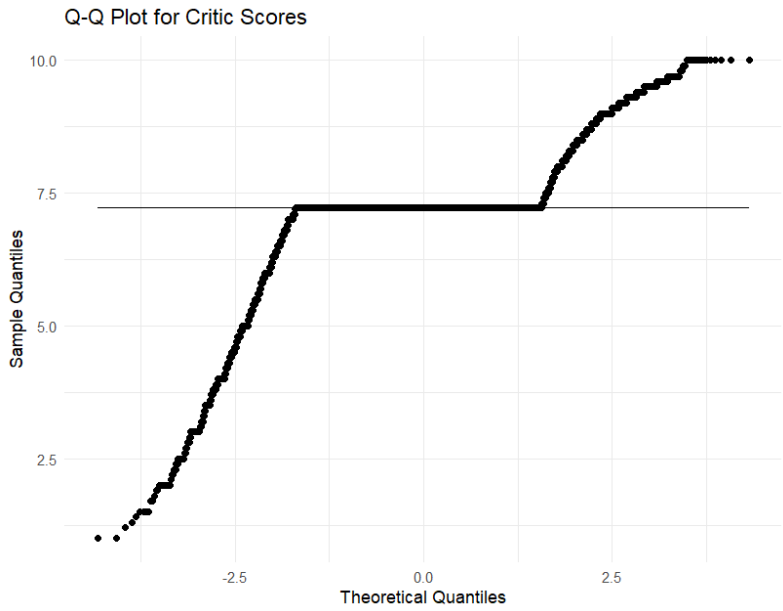

Figure 7: Q-Q plot for critic scores

The Q-Q plot for critic scores in Hypothesis 1 shows noticeable deviations from normality, with a flat middle section around 7.5 and curved tails, indicating clustering and skewness. However, given our large sample size, the Central Limit Theorem (CLT) ensures that the sampling distribution of the mean remains approximately normal, allowing us to proceed with the one-sample t-test. While normality is not strictly met, the t-test remains valid due to the robustness provided by the large dataset. This non-normality suggests that critic scores may be influenced by rating conventions or common score distributions in game reviews.

Q-Q plot for shooter sales

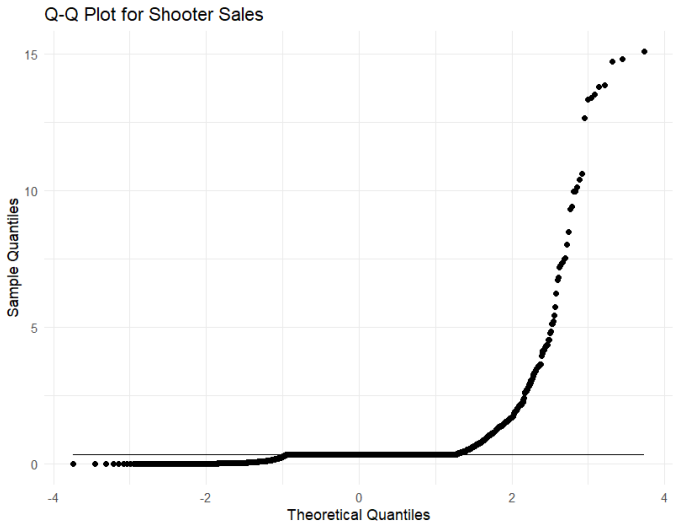

Figure 8: Q-Q plot for shooter sales

The Q-Q plot for shooter sales in Hypothesis 2 reveals a strong deviation from normality, with a flat lower section and an extreme upward curve in the upper quantiles, suggesting heavy right-skewness and the presence of outliers. This indicates that most shooter games have low sales, while a few titles achieve very high sales, creating a long-tailed distribution. However, given our large sample size, the Central Limit Theorem (CLT) ensures that the sampling distribution of the mean is approximately normal, making the two-sample t-test still valid. The skewness suggests that a small number of highly successful shooter games drive the overall sales average, which could be important for interpreting the results.

Q-Q plot for action sales

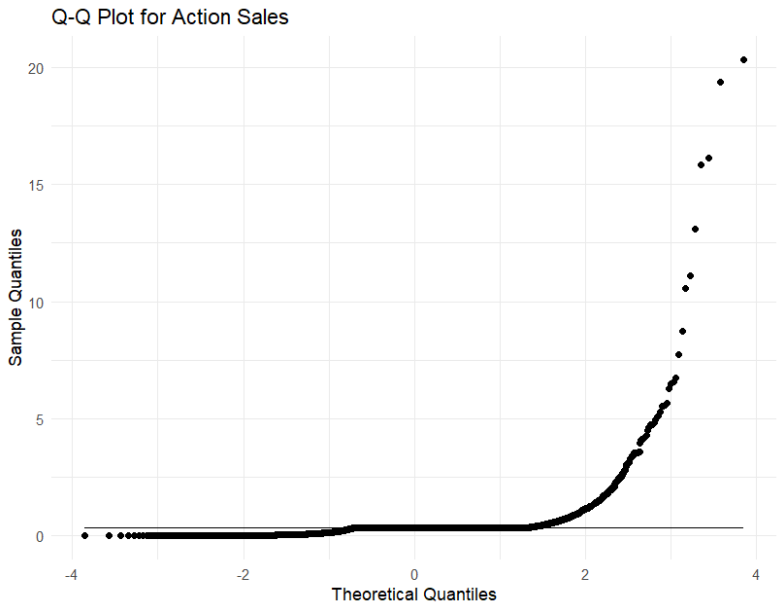

Figure 9: Q-Q plot for action sales

The Q-Q plot for action sales in Hypothesis 2 shows a pattern similar to shooter sales, with a flat lower section and a sharp upward curve in the upper quantiles, indicating strong right-skewness and the presence of outliers. This suggests that most action games have low sales, while a few titles achieve exceptionally high sales, creating a long-tailed distribution. However, given our large sample size, the Central Limit Theorem (CLT) ensures that the sampling distribution of the mean remains approximately normal, making the two-sample t-test still valid. The heavy skewness suggests that a small number of highly successful action games disproportionately contribute to the overall sales average, which is important to consider when interpreting the statistical comparison between shooter and action sales.

Q-Q plot for North America sales

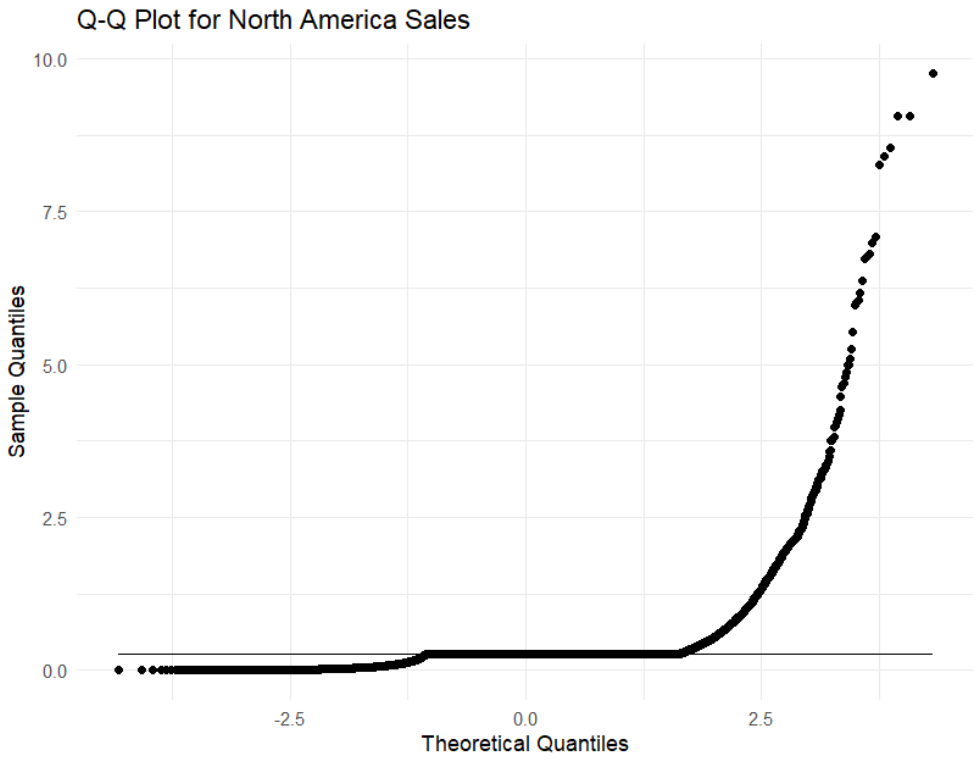

Figure 10: Q-Q plot for North America sales

The Q-Q plot for North America sales in Hypothesis 3 shows a right-skewed distribution, with most games having low sales while a few achieve very high sales, creating a long-tailed pattern. The lower portion of the plot remains flat, indicating a large number of low-sales games, while the upper tail curves sharply upward, suggesting the presence of outliers. Despite this deviation from normality, the Central Limit Theorem (CLT) ensures that the sampling distribution of the mean is approximately normal due to the large sample size. This allows us to proceed with the two-sample t-test, though the skewness suggests that a few highly successful games disproportionately influence the overall sales average.

Q-Q plot for Europe & Africa sales

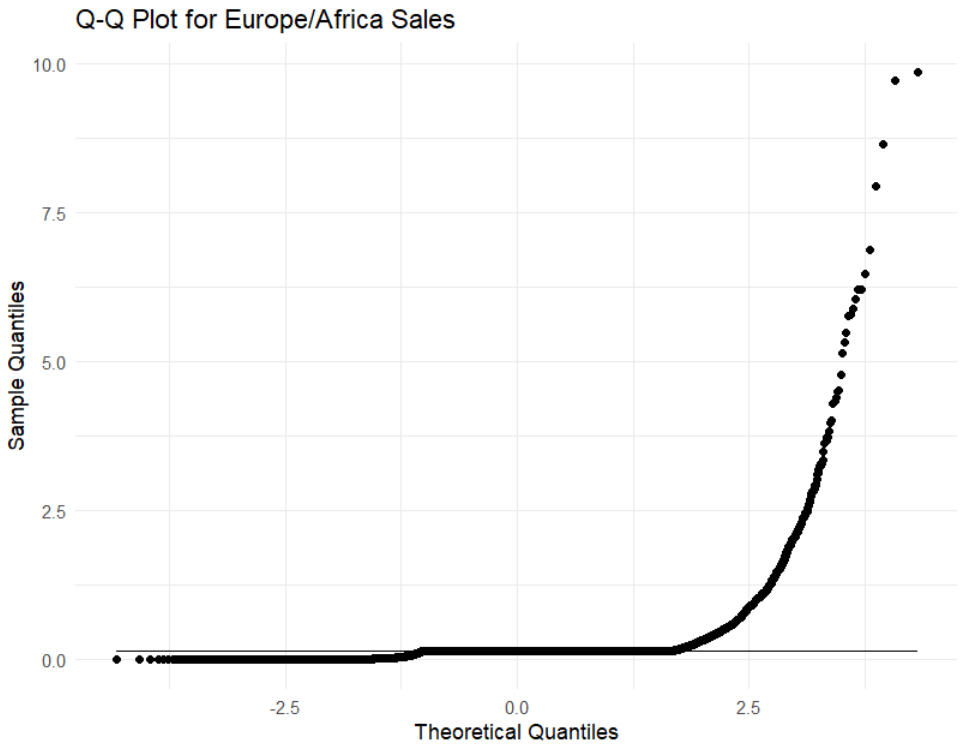

Figure 11: Q-Q plot for Europe & Africa sales

The Q-Q plot for Europe/Africa sales in Hypothesis 3 exhibits a right-skewed distribution, similar to North America sales. The lower portion of the plot remains flat, indicating that most games have low sales, while the upper tail curves sharply upward, suggesting the presence of outliers and highly successful games. This pattern implies that only a few top-selling games contribute significantly to total sales. Despite this deviation from normality, the Central Limit Theorem (CLT) ensures that the sampling distribution of the mean is approximately normal due to the large sample size, making the two-sample t-test valid. However, the skewness indicates that a small number of high-selling games may disproportionately impact the average sales figures in this region.

Residuals vs. Fitted Plot for Regression Model Predicting Total Sales from Critic Score

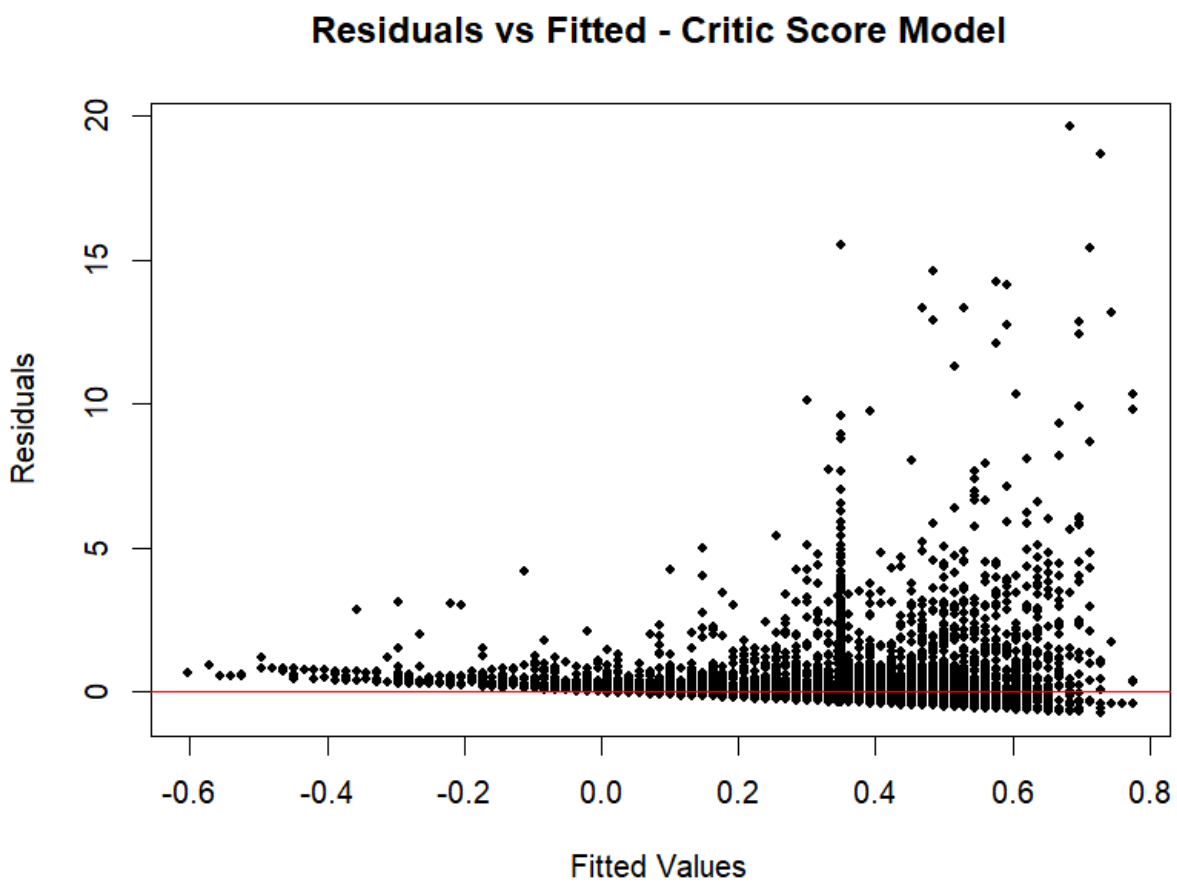

Figure 12: Residuals vs. Fitted Plot for Regression Model Predicting Total Sales from Critic Score

The residuals vs. fitted plot for the Critic Score model (Hypothesis 4) reveals a distinct funnel shape, indicating heteroscedasticity. This means that the variability of residuals increases with higher predicted sales values. While this violates the assumption of constant variance in linear regression, the model remains statistically significant due to the large sample size. Nonetheless, results should be interpreted with caution, acknowledging this limitation.

Summary of Residuals for Regression Model: Critic Score vs. Total Sales

Statistic	Value
Min	-0.7186
1Q	-0.0002
Median	-0.0002
3Q	-0.0002
Max	19.6373

Table 6: Summary of Residuals for Regression Model: Critic Score vs. Total Sales

A residuals summary table was generated to describe the spread and center of the model's prediction errors. The median, 1st quartile, and 3rd quartile residuals were all very close to zero (-0.0002), indicating that the model generally predicts sales accurately for most games. However, the maximum residual of 19.6373 shows the presence of at least one substantial outlier, a game whose actual sales far exceeded the model's prediction based solely on critic score. This highlights the limitations of using critic score alone to explain extreme commercial success in certain cases.

Residuals vs. Fitted Values – Release Year Model

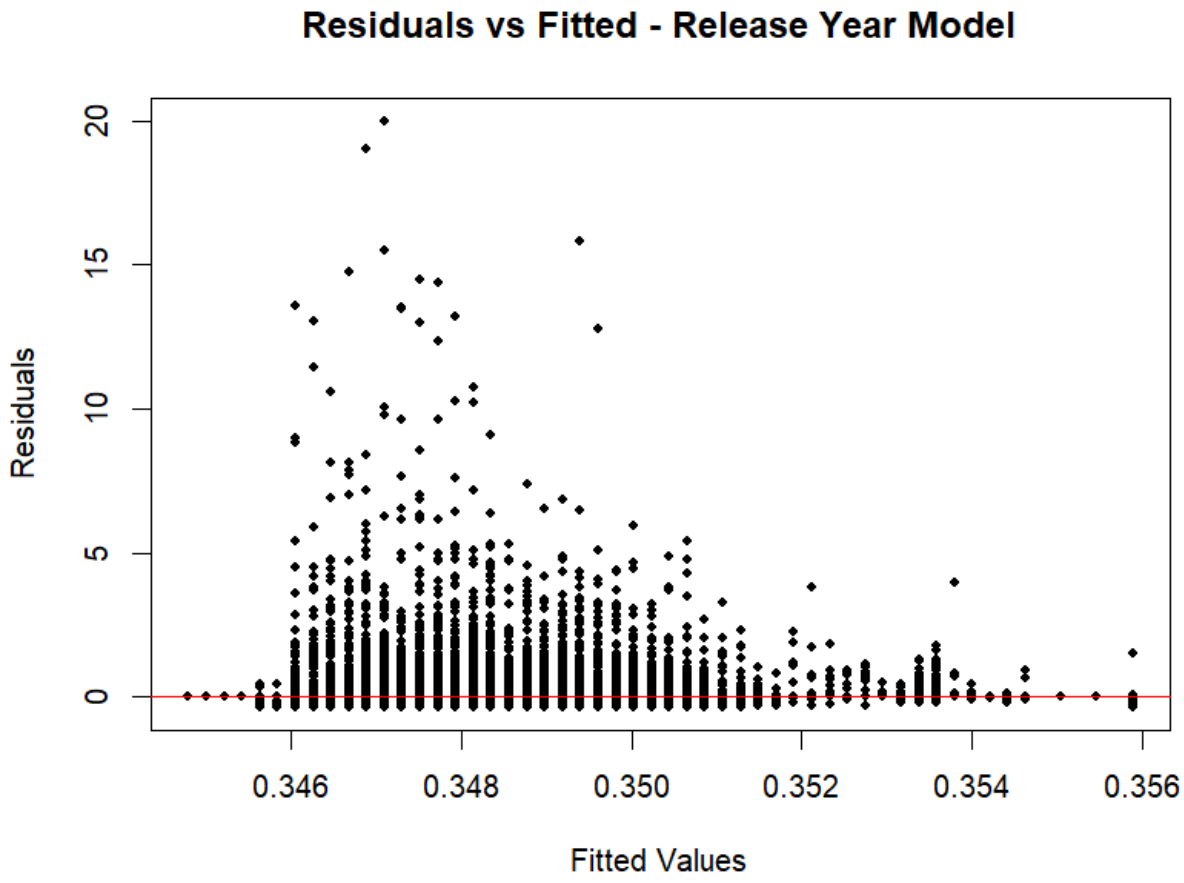

Figure 14: Residuals vs. Fitted Values – Release Year Model

This plot shows the residuals from the regression model of total sales predicted by release year. The residuals appear relatively evenly scattered around the horizontal line at zero, without any distinct curvature or strong pattern. There is no clear indication of a funnel shape or systematic increase in spread as fitted values grow, which suggests that the variance of the residuals remains fairly constant across predicted values. While some mild heteroscedasticity may be present in the lower fitted values, it does not appear severe enough to invalidate the model assumptions. Overall, the plot supports the use of simple linear regression for this analysis.

Summary of Residuals for Regression Model: Release Year vs. Total Sales

Statistic	Value
Min	-0.3559
1Q	-0.0068
Median	-0.0007
3Q	0.0017
Max	19.9729

Table 9: Summary of Residuals for Regression Model: Release Year vs. Total Sales

The residual summary table for Hypothesis 6 reveals that the majority of residuals are tightly clustered around zero, with the first quartile at -0.0068, the median at -0.0007, and the third quartile at 0.0017. This indicates that most of the model's prediction errors are relatively small and centered around the true values. However, the residual range extends from a minimum of -0.3559 to a maximum of 19.9729, suggesting the presence of a few extreme outliers where the model significantly underestimates total sales. These outliers likely correspond to exceptionally high-selling games that the simple linear regression cannot accurately predict. Overall, this residual distribution supports the conclusion that, while the model is statistically significant and reasonably stable for most observations, its practical predictive power is modest and should be interpreted with caution.

Residuals vs. Fitted Plot for Regression Model Predicting Total Sales from Number of Platforms

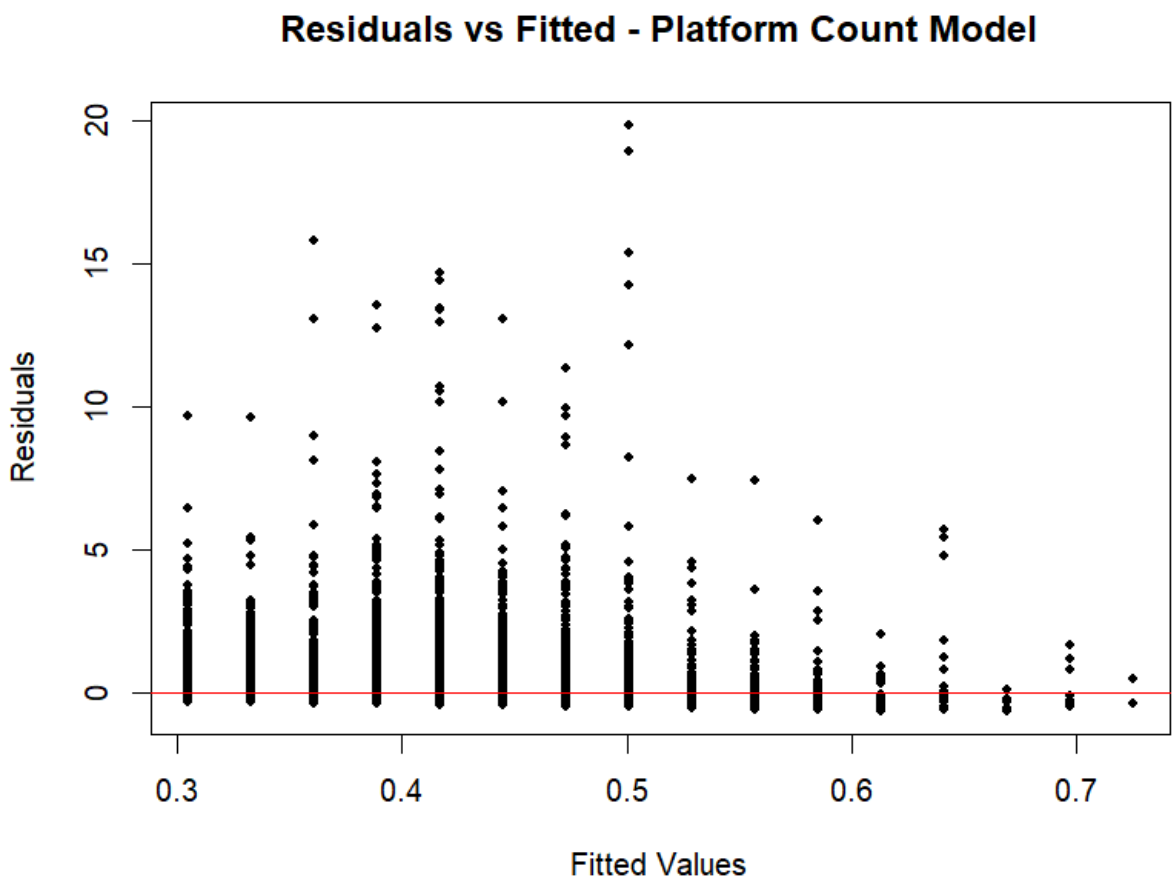

Figure 16: Residuals vs. Fitted Plot for Regression Model Predicting Total Sales from Number of Platforms

Figure 16 shows the residuals plotted against the fitted values for the regression model predicting total sales based on the number of platforms a game was released on. The residuals are generally centered around zero and maintain a fairly constant spread across the range of fitted values. While there is some mild asymmetry, with more variability in residuals at lower predicted values and a slight narrowing as fitted values increase, this does not suggest severe heteroscedasticity. Most data points are tightly grouped, especially at the lower end of the sales spectrum, which is consistent with the broader dataset's right-skewed sales distribution. Overall, the plot suggests that the variance of residuals is acceptably stable across the predicted values, providing visual support that the homoscedasticity assumption is not meaningfully violated.

Summary of Residuals for Regression Model: Platform Count vs. Total Sales

Statistic	Value
Min	-0.6292
1Q	-0.0959
Median	0.0163
3Q	0.0443
Max	19.819

Table 12: Summary of Residuals for Regression Model: Platform Count vs. Total Sales

The residual summary for Hypothesis 6 shows that most prediction errors are small, with the first and third quartiles (−0.0959 and 0.0443) tightly clustered around the median (0.0163). However, the large maximum residual (19.819) indicates the presence of a few high-selling outliers that the model fails to predict accurately. This reinforces the earlier conclusion that while platform count is statistically significant, its predictive power is limited due to substantial variability not captured by the model.