# Investigating the Impact of Game Consoles on Game Sales

Lab 2 - ZAP: Zane, Addy, Pavan

## Introduction

Out of 8 billion people worldwide, 3.26 billion play video games, highlighting the magnitude of video game reach. US gamers spent $55.5 billion on games in 2022.[1] Evidently, video games are a lucrative industry and ZAP (Zane, Addy, Pavan) Studio has created a next-hit game and is now determining the optimal console for its launch based on game sales. Our board and investors are interested in the financial success of the game and would like us to conduct this study to identify the ideal console.

This study examines how consoles, entertainment software rating Board (ESRB) ratings, and maximum number of allowable players influence game sales. Applying a set of regression models, ZAP Studio identifies the ideal console to launch our game and estimates the sales that results from it. Our research question is:

- How do game consoles impact the sales of video games?
- (Sub-question) Does Nintendo DS, the console with the most number of games, have the most positive impact on game sales?
- (Sub-question) What impact do other game features have on game sales?

## Data and Methodology

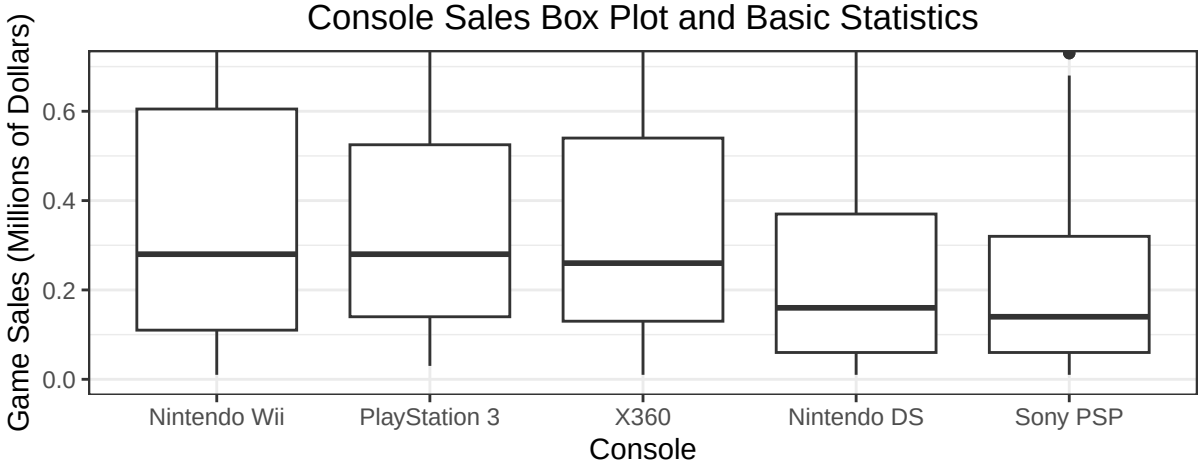
The data in this study comes from the CORGIS (Collection of Really Great, Interesting, Situated) Dataset Project.[2] The data was published by Austin Cory Bart, originally collected by Dr. Joe Cox. It was created from various sources, including crowd-sourced data on "How Long to Beat." It contains information on dollar sales and characteristics of over a thousand video games released from 2004 to 2010. Each row in the data represents a game and the same game can appear multiple times under different consoles. ZAP performed exploratory data analysis (EDA) on the entire dataset then conducted model building on 70% of the data. The remaining 30% of the data was used to generate the statistics in this report.

For the exploratory data analysis (EDA), we focused on key variables, including console, maximum allowable players, ESRB rating, and sales. Sales served as the success metric, while console represented the primary variable influencing sales. We selected maximum players and ESRB ratings as "other game features" from the dataset due to their dissimilarity in data type and characteristics, avoiding variables like online availability, which is console-related. We checked unique fields and nulls to verify if any rows should be dropped. We conducted a correlation matrix between numerical variables to check their relationship, which revealed a small correlation (0.04) between sales and max players. The EDA covered 1212 observations spanning 7 years. Recognizing the video game industry's dynamic nature and the difficulty in developing a causal model with time-series data, we treated the observations as cross-sectional, disregarding temporal ordering in our model due to changing trends.

To operationalize the variables, we initially assessed the distribution of game sales, noting its significant skewness. To address this, a log transformation was applied. As shown blow, a boxplot visualization of sales across consoles showcased the sales range diversity and a summary table detailing count, median, mean, and standard deviation of sales provided insights into each console's performance. It's interesting to note that the two hand-held consoles, Nintendo DS and PSP, had relatively low despite having relatively high count of sales. Additionally, largely due to the effects

[1] Marko Dimitrievski: "33 Evolutionary Gaming Statistics of 2023." Truelist (2023).
[2] Austin Cory Bart: "Video Games CSV File." CORGIS (2017).

of outliers, consoles with the highest IQR for sales don't necessarily have the highest SD. Based on these, we were interested in first examining the console with the most number of games, Nintendo DS.




| Release.Console | Count of Sales | Median Sales Dollars | Mean Sales Dollars | SD Sales Dollars |
|---|---|---|---|---|
| Nintendo DS | 317 | $ 160,000 | $ 508,927 | $ 1,239,082 |
| Nintendo Wii | 187 | $ 280,000 | $ 731,123 | $ 1,700,940 |
| PlayStation 3 | 179 | $ 280,000 | $ 452,291 | $ 557,210 |
| Sony PSP | 242 | $ 140,000 | $ 262,769 | $ 368,659 |
| X360 | 287 | $ 260,000 | $ 582,683 | $ 932,657 |

The initial linear regression model, focused solely on Nintendo DS, revealed a non-significant impact on game sales, with a decrease of 0.15123 units (p-value: 0.129). The model's adjusted R^2 (0.0015) indicated minimal explanatory power for the variance in log sales. Subsequently, a second model encompassing all consoles exhibited statistically significant p-values for each type, notably highlighting X360 with the highest estimated change in log sales. However, the adjusted R^2 remained modest at 0.04421, suggesting approximately 4.4% variability in log sales. A third model, introducing additional game features (ESRB rating and max players), demonstrated overall statistical significance, though individual predictors varied in significance. X360 continued to show the highest estimated change in log sales compared to Nintendo DS (0.322 estimate, p-value: 0.01). Notably, ESRB ratings and maximum number of players, excluding rating T (Teen), exhibited significance, resulting in a slight improvement in adjusted R^2 to 0.0567.

## Results

After constructing the three models on the training data, we subsequently applied them to the testing data. By examining the p-value associated with the F-statistic in the Analysis of Variance (ANOVA), we found a low p-value (2.476e-06) for model 2. This suggests that incorporating all consoles significantly enhanced the model fit compared to the baseline model, which only featured Nintendo DS. However, the p-value associated with including consoles, ratings, and max players in model 3 exceeded 0.05, indicating that the additional features of ratings and max players do not bring a significant improvement in fit compared to model 2.

The regression equation for best performing model (model 2):

$$\widehat{logsales} = \beta_0 - 1.97303 + \beta_1 \cdot (Wii) + \beta_2 \cdot (PS3) + \beta_3 \cdot (PSP) + \beta_4 \cdot (XBOX360) + \mathbf{Z}\gamma$$

Table 1: Estimated Regressions

| | Output Variable: Game Sales | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Nintendo DS | $-0.561^{***}$ | | |
| | (0.135) | | |
| Nintendo Wii | | $0.888^{***}$ | $0.914^{***}$ |
| | | (0.178) | (0.185) |
| Playstation 3 | | $0.776^{***}$ | $0.826^{***}$ |
| | | (0.195) | (0.207) |
| Sony PSP | | $-0.025$ | $-0.036$ |
| | | (0.169) | (0.171) |
| XBOX 360 | | $0.712^{***}$ | $0.730^{***}$ |
| | | (0.164) | (0.173) |
| ESRB M Rating | | | $-0.216$ |
| | | | (0.155) |
| ESRB T Rating | | | $-0.096$ |
| | | | (0.131) |
| Maximum Number of Players | | | $-0.023$ |
| | | | (0.053) |
| Constant | $-1.412^{***}$ | $-1.973^{***}$ | $-1.866^{***}$ |
| | (0.070) | (0.111) | (0.144) |
| Observations | 364 | 364 | 364 |
| $R^2$ | 0.046 | 0.119 | 0.125 |
| Adjusted $R^2$ | 0.043 | 0.109 | 0.107 |
| Residual Std. Error | 1.138 (df = 362) | 1.098 (df = 359) | 1.099 (df = 356) |
| F Statistic | $17.262^{***}$ (df = 1; 362) | $12.141^{***}$ (df = 4; 359) | $7.232^{***}$ (df = 7; 356) |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

The table presents the outcomes of the three models, emphasizing the significance of the console coefficient across all models, notwithstanding the non-significance of Sony PSP. Model 1 indicates that Nintendo decreases game sales (coefficient = -0.561), challenging the initial hypothesis that the console with the most games leads to the highest sales increase. In Model 2, controlling for other variables, Nintendo Wii, PlayStation 3, and XBOX 360 boost sales, with Nintendo Wii showing the highest increase (coefficient = 0.888). Model 3 reveals no statistical significance for ESRB ratings and the maximum number of players. Although Model 3 has a slightly higher R^2 value, suggesting a greater proportion of variance in sales, Model 2 has a higher F-statistic, indicating a better overall fit. Using Model 2's results, we interpret that, holding other factors constant, having games released on Nintendo Wii is associated with an 89% increase in log game sales.

While our statistical analysis reveals higher coefficients for Nintendo Wii, it's crucial to consider factors beyond statistical analysis for understanding practical significance. An ideal console should meet necessary

controls, interface requirements, and provide optimal gaming experiences. Additionally, the console must meet hardware specifications to support the game's computing needs. Moreover, understanding why certain games outsell others requires analyzing gameplay, story, graphics, and other elements, considering factors like consumer economics, gaming trends, and external media influence. In conclusion, besides the preferred console, the substance of a video game, encompassing various features, contributes to an enriched gaming experience and influences market behavior.

## Limitations

In evaluating our model, we considered several assumptions and acknowledged certain limitations. Firstly, the Independent and Identically Distributed (IID) assumption holds given the dataset's extensive coverage of over a thousand video games released across multiple years. While unique events, such as the 2008 financial market crash, might impact the gaming industry at large, the dataset's size and duration allow us to reasonably assume IID. Independence assumes one game does not influence the characteristics of another, and though some games are sequels, the wide possibility of game types and players limit the ability to influence each other, supporting the independence assumption. We explored the existence of a unique Best Linear Predictor (BLP). While we couldn't check finite mean and variance for categorical predictors, we assessed collinearity using variance inflation factors, finding low values (~1) for each predictor. This suggests no evidence to reject the assumption of a unique BLP. The homoskedastic errors assumption was examined through scale location and histogram of residuals plots. The scale location indicated roughly constant residuals spread across features and the histogram should roughly normal distribution of residuals, supporting our assumption of homoskedasticity. We were then able to proceed building our causal model.

Beyond model assumptions, there are additional limitations. We would have preferred to analyze game profit alongside sales for a more comprehensive model, but that was not possible due to data unavailability. The 70/30 training/testing split, while not ideal, was chosen due to data size and time constraints. We would have preferred bootstrapping or cross validation. For structural limitations, omitted variables, such as price, game parent holding company, genre, could bias estimates. Model evolution highlighted the importance of including all consoles for accuracy. Model 1's omission of consoles besides Nintendo DS likely led to an overestimation of its impact. Finally, regarding model results, the adjusted R^2 is not close to 1, and although model 2 yielded the best F-stat, we are aware that it is not by much. We believe that this is due to our small dataset and our key variable being categorical. Despite these limitations, this study was helpful in deriving insights.

## Conclusion

This research study conducted by ZAP Studio provides valuable insights into the video games, particularly the influence of console on game sales. The study's main findings highlight that console choice affects game sales. Among the consoles examined, the Nintendo Wii emerged as the most influential in boosting sales, as indicated by its highest coefficient in the regression models. This finding allows us to reject the initial hypothesis that the console with the most games (Nintendo DS) would lead to the highest sales. The analysis also revealed that while ESRB ratings and the max players are important factors, their impact on sales was not statistically significant.