

CIND820: Project Design

Project Name:

Student: Eddie Li

Student #: 500820042

Supervisor: Ceni Babaoglu

Date of Submission: February 21, 2026

Contents

1	Abstract	3
2	Literature Review	3
2.1	Factors that Affect Video Game Sales	3
2.1.1	Video Game Characteristics	3
2.1.2	Reputation	4
2.1.3	Platform	4
2.2	Predictive Modeling and Historical Methods	5
2.3	Research Gaps	6
3	Data Description	6
4	Project Approach and Visualization	8
5	GitHub Repo	10
	References	11

1 Abstract

Video game sales are influenced by multiple factors, including platform, genre, consumer and reviews, and publisher reputation [1]. While prior studies identify these as important factors of determining sales, many of these studies examine them individually rather than analyzing them in tandem [2]. In addition, earlier predictive modeling attempts relied on linear regression techniques which often did not fully capture the nonlinear relationships among the determining factors [3]. To combat that, this proposed study will instead be using predictive analytics with nonlinear regression modeling to evaluate how platform, genre, reviews, and reputation affect global video game sales. The dataset that will be used in this proposed study is a historical dataset obtained from Kaggle which contains platform, genre, publisher, review scores, and regional sales data from 1983 to 2012 [4].

2 Literature Review

Research on video game sales has progressed from descriptive analyses of market characteristics to more advanced empirical and data-driven approaches that examine how multiple factors interact to influence commercial success. Prior studies identify a range of determinants, including platform choice, genre, consumer and critic reviews, corporate reputation, branding strategies, and predictive modeling techniques. Collectively, this literature provides a strong foundation for applying data analytics methods to forecast sales performance in the video game industry.

2.1 Factors that Affect Video Game Sales

2.1.1 Video Game Characteristics

Video game characteristic factors include design characteristics such as genre, themes, and non-design characteristics that could influence consumer preferences. A study on U.S sales data relates genre as a major factor of sales performance, with certain genres

more likely to have high sales than others [1]. Release timing and the current market also influence demand. Launching a game during a time of high competition or unfavorable market conditions can limit its potential sales [1]. Another study documents industry trends indicating a bias for franchise titles, specific themes, and unique gameplay [5]. Collectively, these studies indicate that video game sales are largely determined by video game characteristics which appeal to specific audiences.

2.1.2 Reputation

Reputation factors include game reviews, company reputation, and franchise recognition. All of these are major factors of video game sales. Consumer and critic reviews function as a form of quality control that reduce the amount of misinformation before purchasing [6]. It is important to note that a study showed that consumer reviews exert a stronger influence on sales for less popular games, indicating that reputation is more important when franchise familiarity is low [7]. Furthermore, another study shows that user and critic reviews affect sales in a nonlinear way which depends on how extreme the reviews are. Company and franchise reputation are also key determining factors in video game sales since publisher reputation significantly increases the chances at commercial success. Evidence was found that strong company reputation enhances video game sales, especially when reinforced through strategic branding and collaboration [8]. Together, these studies show that stronger a reputation is associated with higher video game sales.

2.1.3 Platform

Platform factors refer to the influence the gaming platform has on video game releases. Previous research shows that sales for the same title can differ a lot across platforms due to differences in their audiences [9][10]. The sales differences among platforms shows that platform choice is an important decision for developers and publishers. This pattern is also seen on the global market, with the most popular platform choice differencing depending on geography [10]. Platform competition also involves the choice between expanding the current audience or securing exclusive franchises, which can affect sales by

changing platform/company reputation and user access [11]. These factors highlight the role of platform selection in determining commercial success.

2.2 Predictive Modeling and Historical Methods

Previous research shows the importance of preparing and combining data before making predictions. It was shown that predictive accuracy improves when different types of data are properly processed beforehand [6]. While the study was in the healthcare field, the same logic can be applied to video game sales prediction. The main ideas include: handling missing values, normalizing variables, and analyzing multiple characteristics simultaneously [6]. Proper preprocessing ensures that the data is ready for meaningful analysis and reliable modeling.

On top of preprocessing, analysis of past industry data also plays an important role in predictive modeling. It has been shown that studying historical patterns helps identify relationships between factors, in this case: platform, genre, and sales performance [5]. This supports the selection of determining factors for regression models. As a result, exploratory data analysis helps connect historical data to known prediction methods [5].

Predictive analytic techniques like regression, are useful in prior research that models video game sales. By comparing linear and nonlinear regression models, it was found that nonlinear approaches consistently outperformed linear models in predicting sales [12]. This previous research shows that video game sales depend on complex relationships between factors. Making simple linear model less reliable than a regression model [12].

Regression remains one of the most common methods for predicting video game sales. Previous studies show that regression models can estimate potential sales and measure how important different factors are. In addition, model performance is evaluated with numerical metrics to ensure predictions are applicable to new data [12]. All these studies support a proposed methodology of data preprocessing, analysis, regression modeling, and performance evaluation.

2.3 Research Gaps

Despite the existing research on video game sales, several gaps remain in the literature. The first gap exists because many studies examine determining factors individually and separately rather than together [2]. Prior work frequently concentrates on specific factors like the ones mentioned previously. These works do not account for their combined impact on video game sales [2]. As a result, the combined impact among major factors of video game sales remain insufficiently explored.

The second gap exists because earlier work often relies on traditional statistical techniques such as linear regression. This may not accurately account for nonlinear effects between determining factors [12][3]. For example, studies found that user and critic reviews affect sales in nonlinear ways and that nonlinear regression outperforms linear models for complex factors [12][3]. Due to the fact that these effects are rarely accounted for in predictive models, a nonlinear model that incorporates all major factors in a nonlinear model has not been fully explored.

This project will examine how platform, genre, reviews, and reputation together affect sales using nonlinear models. By combining exploratory analysis with predictive modeling, it aims to apply theoretical research to practical decisions in the video game industry.

3 Data Description

The raw dataset used in this study contains historical information on video game titles with the variables and descriptions from Kaggle [4]:

- Rank - The rank of the video game based on global sales volume. (Numerical)
- Game Title - The name of the video game. (String)
- Platform - The platform on which the game is available, such as PC, PS4, Xbox One, etc. (Categorical)
- Year - The year in which the game was released. (Date)

- Genre - The genre of the game, such as action, adventure, racing, etc. (Categorical)
- Publisher - The company that published the game. (String)
- North America - The number of units sold in North America, in millions. (Numerical)
- Europe - The number of units sold in Europe, in millions. (Numerical)
- Japan - The number of units sold in Japan, in millions. (Numerical)
- Rest of World - The number of units sold in the rest of the world, excluding North America, Europe, and Japan, in millions. (Numerical)
- Global - The total number of units sold worldwide, in millions. (Numerical)
- Review - The review score of the game, on a scale of 1 to 10. (Numerical)

Descriptive statistics was calculated to summarize the main characteristics of the dataset using Python. The functions *shape*, *dtypes*, *isnull*, and *describe* were used to calculate basic information about the dataset.

```
(1907, 13)
index      int64
Rank       int64
Game Title  object
Platform    object
Year        float64
Genre       object
Publisher    object
North America float64
Europe      float64
Japan       float64
Rest of World float64
Global      float64
Review      float64
dtype: object
```

	Year	Genre	Publisher	North America	Europe
count	1878.000000	1907	1905	1907.000000	1907.000000
unique	NaN	12	94	NaN	NaN
top	NaN	Sports	Electronic Arts	NaN	NaN
freq	NaN	308	341	NaN	NaN
mean	2003.766773	NaN	NaN	1.258789	0.706675
std	5.895369	NaN	NaN	1.956560	1.148904
min	1983.000000	NaN	NaN	0.000000	0.000000
25%	2000.000000	NaN	NaN	0.510000	0.230000
50%	2005.000000	NaN	NaN	0.810000	0.440000
75%	2008.000000	NaN	NaN	1.375000	0.810000
max	2012.000000	NaN	NaN	40.430000	28.390000

```

              Japan  Rest of World  Global
count  1907.000000  1907.000000  1907.000000
unique    NaN          NaN          NaN
top        NaN          NaN          NaN
freq        NaN          NaN          NaN
mean    0.317493    0.206471    2.489240
std     0.724945    0.343093    3.563159
min      0.000000    0.000000    0.830000
25%     0.000000    0.060000    1.110000
50%     0.020000    0.130000    1.530000
75%     0.300000    0.220000    2.540000
max      7.200000    8.540000    81.120000

Review index
count  1907.0000  1907.0000
unique    NaN      NaN
top        NaN      NaN
freq        NaN      NaN
mean    953.0000  954.0000
std    550.6478  550.6478
min      0.0000    1.0000
25%    476.5000  477.5000
50%    953.0000  954.0000
75%   1429.5000  1430.5000
max   1906.0000  1907.0000

Game Title Platform \
count  1907      1907
unique    NaN      22
top        NaN      NaN
freq        NaN      NaN
mean      NaN      NaN
std      NaN      NaN
min      NaN      NaN
25%      NaN      NaN
50%      NaN      NaN
75%      NaN      NaN
max      NaN      NaN

LEGO Batman: The Videogame  PS2
6                          372
NaN                        NaN
0.724945                   0.343093
0.000000                   0.000000
0.000000                   0.060000
0.020000                   0.130000
0.300000                   0.220000
7.200000                   8.540000

79.038977  Genre
10.616899  Publisher
30.500000  North America
74.000000  Europe
81.000000  Japan
86.230000  Rest of World
97.000000  Global
Review
dtype: int64
```

Figure 1: Raw Data Characteristics

The data preprocessing steps and results are taken are as follows:

- Remove columns that contain a missing "Year" value

- Convert year from float64 to int64
- Convert object data types to string
- Round the float64 columns to 2 decimal places

(1878, 13)									
index	int64				Genre	Publisher	North America	Europe	Japan
Rank	int64				count	1878	1878	1878.000000	1878.000000
Game Title	string				unique	12	94	NaN	NaN
Platform	string				top	Sports	Electronic Arts	NaN	NaN
Year	int64				freq	303	338	NaN	NaN
Genre	string				mean	NaN	NaN	1.262109	0.710517
Publisher	string				std	NaN	NaN	1.967797	0.729295
North America	float64				min	NaN	NaN	0.000000	0.000000
Europe	float64				25%	NaN	NaN	0.510000	0.230000
Japan	float64				50%	NaN	NaN	0.820000	0.440000
Rest of World	float64				75%	NaN	NaN	1.377500	0.820000
Global	float64				max	NaN	NaN	40.430000	28.390000
Review	float64								
dtype: object					Rest of World	Global	Review	index	0
					count	1878.000000	1878.000000	Rank	0
					unique	NaN	NaN	Game Title	0
					top	NaN	NaN	Platform	0
					freq	NaN	NaN	Year	0
					mean	0.207055	2.500532	Genre	0
					std	0.345141	3.587041	Publisher	0
					min	0.000000	0.830000	North America	0
					25%	0.060000	1.110000	Europe	0
					50%	0.130000	1.530000	Japan	0
					75%	0.220000	2.547500	Rest of World	0
					max	8.540000	81.120000	Global	0
					Review				
					0				
					dtype: int64				

modeling by performing exploratory analysis. This highlights which factors are could be important for regression modeling and shows how platform, genre, and sales are related [5].

Following preprocessing and exploratory analysis, regression techniques are applied to estimate global sales and to measure the importance of different factors. Model performance will be evaluated using quantitative metrics to ensure predictions generalize to future data rather than applying only to the historical data the model was built on [12]. These steps of preprocessing, descriptive statistics, exploratory data analysis, factor selection, regression modeling, and model evaluation forms the entire methodology of which sales predictions will be made.

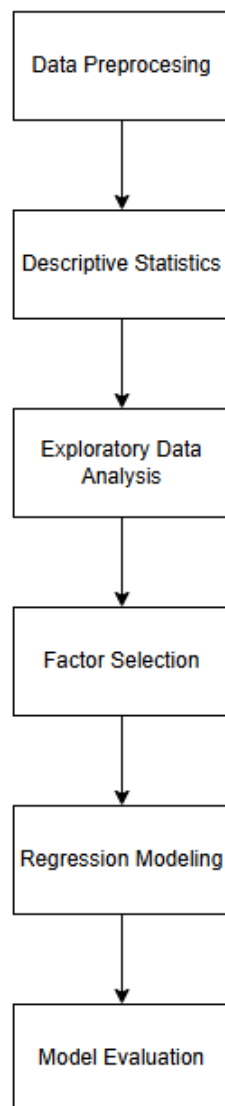


Figure 3: Data Approach Visualization

5 GitHub Repo

<https://github.com/eddie-li-git/CIND820>

References

- [1] J. Cox, “What makes a blockbuster video game? an empirical analysis of US sales data,” en, *MDE Manage. Decis. Econ.*, vol. 35, no. 3, pp. 189–198, Apr. 2014.
- [2] J. Li, Y. Zheng, H. Hu, J. Lu, and C. Zhan, “Predicting video game sales based on machine learning and hybrid feature selection method,” in *2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Chengdu, China: IEEE, Nov. 2021.
- [3] T. Kraemer, W. H. Weiger, and S. Heidenreich, “Do all stars shine the same? investigating the nonlinear effects of user and critic reviews on video game sales,” en, *J. Bus. Res.*, vol. 188, no. 115034, p. 115 034, Feb. 2025.
- [4] A. Bramwell, *Checking your browser - reCAPTCHA*, en, <https://www.kaggle.com/datasets/thedevastator/global-video-game-sales-and-reviews>, Accessed: 2026-2-12.
- [5] H. Cheng, “Video game sales trends and stats,” *BCP Business & Management*, vol. 25, pp. 222–225, Aug. 2022.
- [6] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, “Combining structured and unstructured data for predictive models: A deep learning approach,” en, *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 280, Oct. 2020.
- [7] F. Zhu and X. (Zhang, “Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics,” en, *J. Mark.*, vol. 74, no. 2, pp. 133–148, Mar. 2010.
- [8] H. S. Choi and E. W. Ayaburi, “Effect of corporate reputation and cooperative branding strategy on digital products sales: Evidence from digital video game industry,” en, *Int. J. Inf. Manage.*, vol. 86, no. 102975, p. 102 975, Feb. 2026.
- [9] J. Babb and N. Terry, “Comparing video game sales by gaming platform,” *Journal of the Southwestern Society of Economists*, vol. 40, pp. 25–46, 2013, Accessed: 2026-01-20.

- [10] J. Babb, N. Terry, and K. Dana, “The impact of platform on global video game sales,” *Int. Bus. Econ. Res. J. (IBER)*, vol. 12, no. 10, p. 1273, Sep. 2013.
- [11] C. Cennamo and J. Santalo, “Platform competition: Strategic trade-offs in platform markets,” en, *Strategic Manage. J.*, vol. 34, no. 11, pp. 1331–1350, Nov. 2013.
- [12] C.-W. Chu and G. P. Zhang, “A comparative study of linear and nonlinear models for aggregate retail sales forecasting,” en, *Int. J. Prod. Econ.*, vol. 86, no. 3, pp. 217–231, Dec. 2003.