Video Game Sales Analysis

Mirza Faris Bin Mohd Affandi, Muhammad Bin Hanifah Syukri

2025-05-03

Table of contents

1 Introduction

This report explores global video game sales using a curated dataset of published games. We analyze genre popularity, regional sales, platform distribution, rating trends, and statistical relationships to uncover insights that may inform industry strategies and audience preferences.

2 Research Topic: Video Game Sales and Rating Trends on year 2000 - 2016

This research focuses on global video game sales and ratings trends, a topic that is both popular and meaningful to our team as long-time players and analysts of gaming culture. With the rapid evolution of gaming platforms, changing consumer preferences, and the influence of online reviews, understanding what drives game success is more relevant than ever. Our research centers on uncovering how factors such as genre, platform, and region impact global sales, as well as how critic and user ratings relate to commercial performance. Through exploratory data analysis and visualizations, we aim to identify patterns in genre popularity, regional sales differences, and rating correlations. Our goal is to contribute insights that help explain what makes a video game successful and enhance our understanding of trends in the gaming industry.

3 Methodology

We cleaned the dataset, filtered key columns with missing values, converted `user_score` to numeric, and limited platform-level stats to those with adequate sample size (n 30). Data was visualized using `ggplot2`, and tables were styled using `kableExtra`.

4 Research Questions

The first research question we will explore is: how do factors like genre, region, platform, and review scores affect video game sales? We will use different visualizations to help explain these relationships. We are especially interested in seeing which game genres sell the most overall and whether certain regions prefer different types of games. For example, do players in Japan like different genres compared to players in North America?

We also want to find out if critic scores and user scores are related. Do players usually agree with critics, or are there games with low critic scores that still get high user ratings? This will help us understand how important reviews really are to gamers.

Lastly, we will look at how sales are spread across different platforms. Are some platforms better at selling games than others, or do they just have more games released? These questions will guide our project and help us better understand the gaming industry.

5 Provenance Of Our Data

We got our dataset from Kaggle, a website where people share data for projects and analysis. The dataset is called "Video Game Sales 2016" and includes information about video games released between 2000 and 2016. My groupmate and I found and downloaded the dataset together. It was uploaded to Kaggle by another user for people to use in data analysis. Each row in the dataset represents a video game, with details like its name, platform, genre, sales numbers, and review scores from critics and users.

6 FAIR Principles (Findable, Accessible, Interoperable, Reusable):

1) Findable: Our dataset is easily accessible on Kaggle. It comes with a title and description, making it easy for us to search for and identify.
2) Accessible: Anyone can download and use the dataset from Kaggle with no restrictions, which makes it open and easy to access for others.
3) Interoperable: The dataset is in CSV format, a common file type that works with most data tools like Excel, R, and Python. The column names are also clear and consistent.
4) Reusable: The dataset includes important details like sales, genres, platforms, and scores. This allows it to be reused for different kinds of analysis, like sales trends or rating patterns. We also did some cleaning and preprocessing, which makes it easier to reuse.

7 CARE Principles (Collective Benefit, Authority to Control, Responsibility, Ethics):

1) Collective Benefit: This dataset helps the data science community and students like us learn from real-world information. We use it for educational purposes to study patterns in video game sales and ratings.
2) Responsibility: We made sure to handle the data carefully, especially when cleaning it and removing unclear or non-numeric values (like "tbd" in the user score column).
3) Ethics: The dataset doesn't include any personal or sensitive information, so we're not violating anyone's privacy. It only includes public data about video games.

8 Conclusion

Our analysis of video game sales and ratings trends gave us a deeper understanding of what drives a game's success across different regions and platforms. While we were already familiar with the gaming industry, this project allowed us to explore patterns in sales and reviews that we hadn't noticed before. It was especially interesting to see how certain genres perform better in specific regions and how critic and user scores often differ in their evaluation of games. Visualizing these trends helped us uncover new insights, which can be useful for developers, marketers, and even players looking to understand the industry better. Overall, this project highlights the importance of using data to better understand consumer behavior, improve decision-making, and contribute meaningful knowledge to the gaming community.

9 Citation

1) Video Game Sales with Ratings. (2016, December 30). Kaggle. https://www.kaggle.com/datasets/rush4ra game-sales-with-ratings

2) Data Visualization in R Wickham, H. (2021). ggplot2: Elegant Graphics for Data Analysis. https://ggplot2.tidyverse.org/

10 Exploratory Data Analysis

10.1 Top Publishers by Global Sales

Table 1: Top 10 Publishers by Global Sales (in millions)

publisher	total_sales
Nintendo	1788.81
Electronic Arts	1116.96
Activision	731.16
Sony Computer Entertainment	606.48
Ubisoft	471.61
Take-Two Interactive	403.82
THQ	338.44
Konami Digital Entertainment	282.39
Sega	270.35
Namco Bandai Games	254.62

10.2 Average Scores by Platform (n 30)

Table 2: Average Critic and User Scores by Platform (n > 30)

platform	avg_critic	avg_user
PC	75.98	7.04
PS	74.13	7.88
XOne	73.62	6.54
PS4	72.13	6.75
XB	71.52	7.51
GC	71.10	7.60
PS3	70.82	6.79
PSV	70.80	7.49
WiiU	70.67	7.04
GBA	70.50	7.70
PS2	69.51	7.67
X360	69.04	6.79
PSP	68.73	7.26
3DS	67.85	6.94
DS	66.57	7.05
Wii	64.31	6.91

11 Quantitative Analysis

11.1 Figure 1: Global Sales by Genre

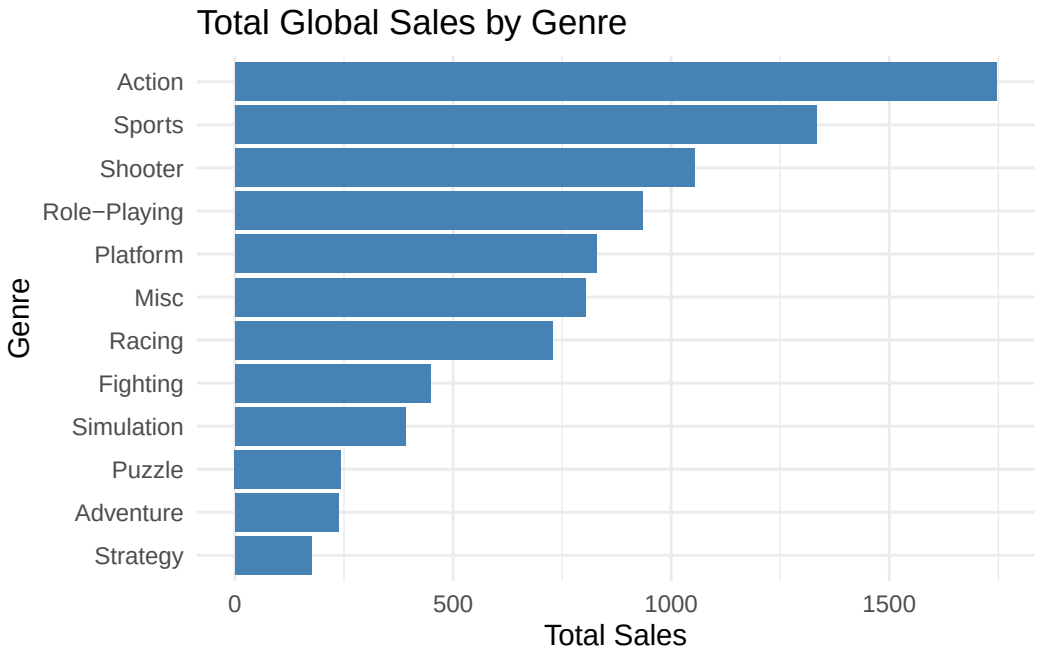

Figure 1: Total Global Sales by Genre

11.2 Figure 2: Sales by Region

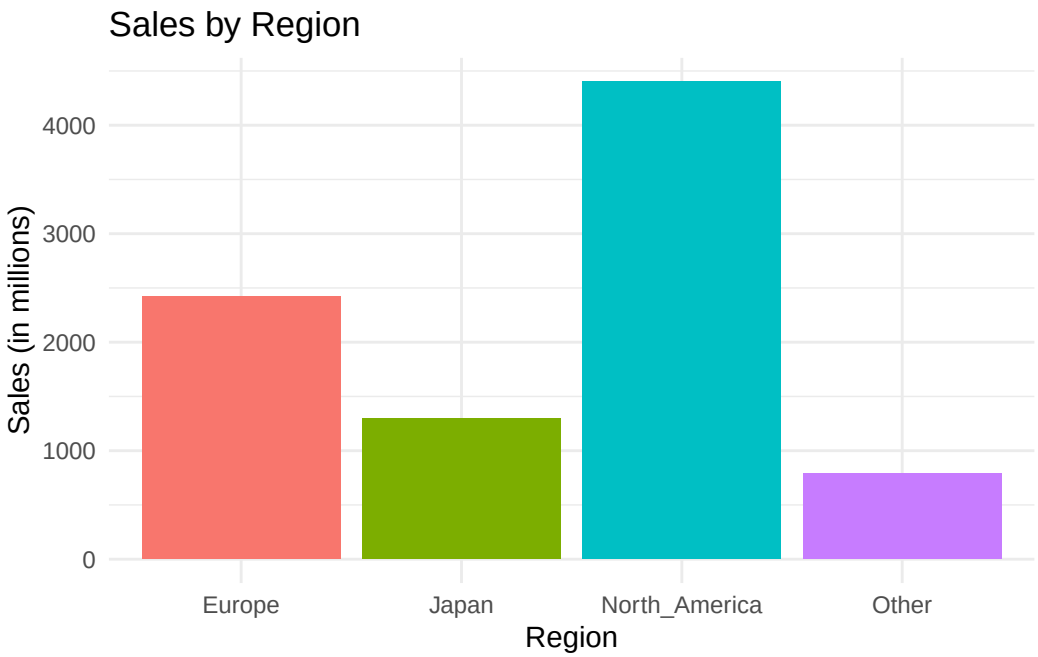

Figure 2: Sales by Region

11.3 Figure 3: Critic Score vs User Score

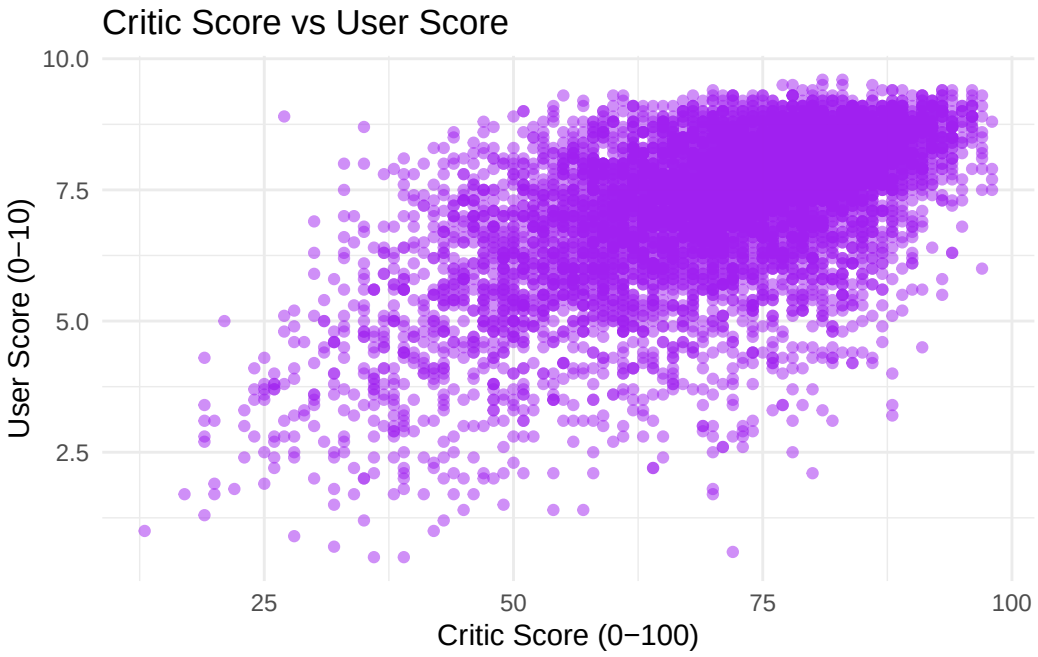

Figure 3: Critic Score vs User Score

11.4 Figure 4: Genre Popularity by Region

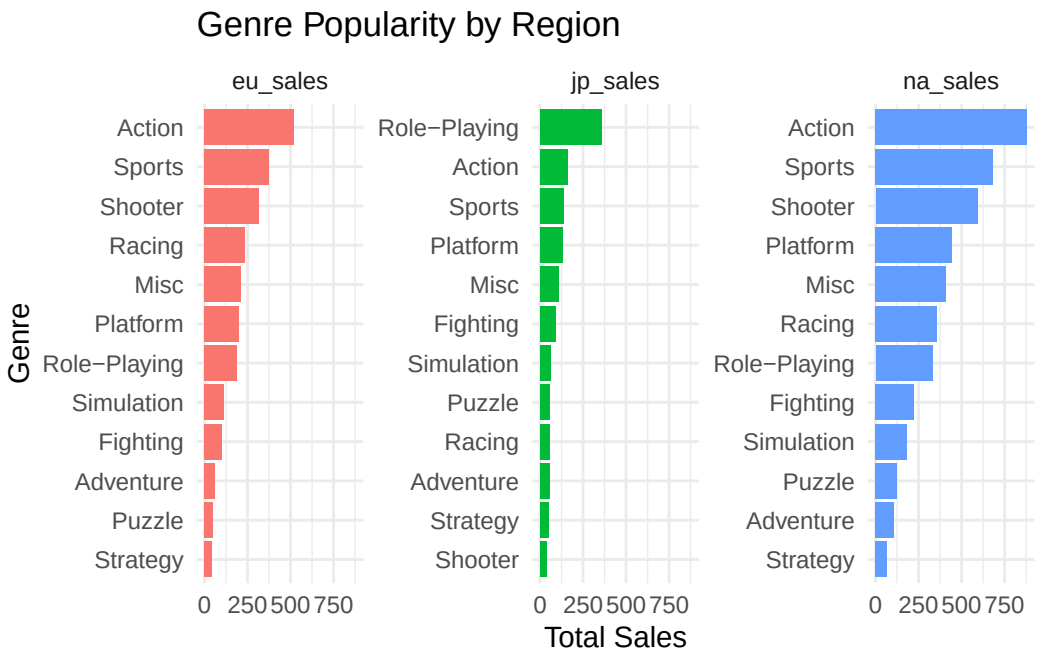

Figure 4: Genre Popularity by Region

12 Regression Analysis

Table 3: Table 3: Linear Regression of Global Sales on Ratings and Platform

term	estimate	std.error	statistic	p.value
(Intercept)	-2.168	0.149	-14.574	0.000
critic_score	0.043	0.002	21.327	0.000
user_score	-0.120	0.020	-6.072	0.000
platformPS2	0.908	0.092	9.893	0.000
platformPS3	0.921	0.097	9.517	0.000
platformX360	0.974	0.095	10.281	0.000
platformXB	0.347	0.106	3.290	0.001
platformOther	0.849	0.080	10.647	0.000

13 Code Appendix

```
# Load libraries
library(tidyverse)
library(janitor)
library(knitr)
library(scales)
library(forcats)
library(tidytext)
library(broom)
library(kableExtra)

# ------------------------------------------------------------------------------
# Load and Clean Data
# ------------------------------------------------------------------------------

# Remove invalid critic scores
games <- games %>% filter(is.na(critic_score) | critic_score > 0)

# ------------------------------------------------------------------------------
# Table 1: Top Publishers by Global Sales
# ------------------------------------------------------------------------------

top_publishers <- games %>%
  group_by(publisher) %>%
  summarise(total_sales = sum(global_sales), .groups = "drop") %>%
  arrange(desc(total_sales)) %>%
  slice_head(n = 10)

# ------------------------------------------------------------------------------
# Table 2: Average Critic and User Scores by Platform
# ------------------------------------------------------------------------------

platform_scores <- games %>%
  filter(!is.na(critic_score), !is.na(user_score)) %>%
  group_by(platform) %>%
  filter(n() >= 30) %>%
  summarise(
    avg_critic = mean(critic_score),
    avg_user = mean(user_score),
    .groups = "drop"
```

```r
  ) %>%
  arrange(desc(avg_critic))

# ------------------------------------------------------------------------------
# Figure 2: Regional Sales Distribution
# ------------------------------------------------------------------------------

region_sales <- games %>%
  summarise(
    North_America = sum(na_sales),
    Europe = sum(eu_sales),
    Japan = sum(jp_sales),
    Other = sum(other_sales)
  )

region_sales_long <- pivot_longer(region_sales, cols = everything(),
                                  names_to = "region", values_to = "sales")

# ------------------------------------------------------------------------------
# Figure 3: Critic Score vs User Score
# ------------------------------------------------------------------------------

critic_vs_user_plot <- games %>%
  filter(!is.na(critic_score), !is.na(user_score)) %>%
  ggplot(aes(x = critic_score, y = user_score)) +
  geom_point(alpha = 0.5, color = "purple") +
  labs(
    title = "Critic Score vs User Score",
    x = "Critic Score (0-100)",
    y = "User Score (0-10)"
  ) +
  theme_minimal()

# ------------------------------------------------------------------------------
# Figure 4: Genre Popularity by Region
# ------------------------------------------------------------------------------

region_genre <- games %>%
  pivot_longer(cols = c(na_sales, eu_sales, jp_sales),
               names_to = "region", values_to = "sales") %>%
  group_by(region, genre) %>%
  summarise(total = sum(sales), .groups = "drop")
```

```r
# ------------------------------------------------------------------------------
# Table 3: Regression Analysis
# ------------------------------------------------------------------------------

reg_data <- games %>%
  filter(!is.na(critic_score), !is.na(user_score)) %>%
  mutate(platform = fct_lump(platform, n = 5))

reg_model <- lm(global_sales ~ critic_score + user_score + platform, data = reg_data)
```