Video Game Sales Exploratory Data Analysis

Project in Data Visualization: Report

Jesse Loi
Ting-Yu Lin
Priyal Sunil Joshi
Erdenetuya Namsrai

December 2024

Table of Contents

List of Figures

Introduction

The most effective way to communicate data insights is through visualization. Visualizing large datasets simplifies complex information, making them easier to comprehend and interpret. To achieve meaningful visualizations, it is essential to thoroughly analyze the data beforehand. This process involves accurately identifying data types, pre-processing the data, and summarizing key elements. Once the data is prepared, the appropriate visualization formats can be selected to best convey the intended insights.

For this project, we have selected the following datasets for analysis and visualization:

1. Video Game Sales
2. Entertainment Software Rating Board (ESRB) Data
3. Economic Data on Gross Domestic Product (GDP)
4. Review Data from IGN and Steam

Using these datasets, we aim to visualize global trends in video games and their regional sales volumes from 1980 to 2020. We will primarily visualize the Video Games dataset, comprising 11 fields and 16,598 records. Additionally, we incorporated the Entertainment Software Rating Board (ESRB) dataset, which contains 34 columns and 1,895 records, as well as the World Bank's economic data on Gross Domestic Product, featuring 16 columns and 267 records. Our IGN dataset contains 15 columns and 11362 records, and our Steam dataset contains 10 columns and 65111 records. Together, these datasets enable us to analyze and illustrate the evolution of the global video game market. All datasets are accessible on Kaggle and other relevant platforms.

- https://github.com/erdenetuya2080/DATA-5310-Data-Visualization/blob/main/vgsales.csv
- https://raw.githubusercontent.com/JesseLoi/Test/refs/heads/main/Video_games_esrb_rating.csv
- https://raw.githubusercontent.com/JesseLoi/Test/refs/heads/main/API_NY.GDP.PCAP.CD_DS2_en_csv_v2_142.csv
- https://www.reddit.com/r/gamedev/comments/x0qs4z/we_gathered_data_about_54000_games_in_steam_and/
- https://www.kaggle.com/datasets/advancedforestry/ign-scores-dataset

To analyze and visualize the evolution of global video game sales trends, we utilized R, a leading data visualization and analysis tool. R was employed to process, summarize, and visualize the

datasets, enabling us to select the most relevant databases and create visualizations that effectively represent the findings. The results were then compiled and presented for further interpretation and insights. We aim for the results presented to be easily comprehensible for video game companies and individuals interested in this market. We hope that businesses operating in this field can leverage these insights at a strategic decision-making level, ultimately driving more informed decisions and enhancing profitability.

Data Methodology

We begin taking our data from its respective Kaggle page (https://www.kaggle.com/datasets/gregorut/videogamesales). Let's check the data to make sure it has integrity. We will first convert the relevant data into numeric entries, such as scores and sales. This conversion process is common to most data clearing. With respect to our particular datasets, there are two immediate issues. First, we must check that the North American (NA) sales, European (EU) sales, Japanese (JP) sales, and sales to all other countries all sum to the global sales, since other sales are stipulated just as the sales in all remaining countries. However, we should realize that, since our data is given in the millions, there may be a rounding error between the countries. In fact, there is a discrepancy of about 0.02 in the data. In other words, around 20,000 copies are not accounted for in the data, but this can be attributed to a fault of how vgchartz displays its data. However, notice that all games needed to have sold at least 100,000 to make it into the dataset, so 20,000 is not such a large discrepancy, especially since only 4 entries have discrepancies that extreme.

Let's move on to the more pressing issue of N/A, or missing, values in the data. The majority of the data has no issues. The only major issue lies in the release year column, which has about 280 missing entries. Year is rather crucial, since our goal is to analyze the growth of the game industry over time. Let's consider some ways to mitigate this problem. First, many games helpfully contain their release year in the title. So, let's use regular expression, which is a technique that sorts patterns in text, to extract 4 consecutive digits (likely the year) from the title. This helps a little bit, but we could do more by extracting two consecutive digits and converting that into the year. For example, a game might be titled '07 to signal a release date of 2007. Likewise, a game with '99 in its title might have been released in 1999. Finally, we should

account for the fact that some games are expressed as digit-k-digit, such as "2K16" representing 2016. After resolving the different columns, we can then convert the appropriate columns to numeric variables (namely the years and the sales) to appropriately plot them. Let's now move onto additional datasets that we will be combining. Still on the topic of video games, we decide to draw data from ESRB rating (https://www.kaggle.com/datasets/imohtn/video-games-rating-by-esrb), IGN rating (https://www.kaggle.com/datasets/advancedforestry/ign-scores-dataset), and Steam reviews (https://www.reddit.com/r/gamedev/comments/x0qs4z/we_gathered_data_about_54000_games_in_steam_and/).

Let's discuss each of these datasets. The Entertainment Software Rating Board gives ratings to games depending on which games would be appropriate for different age groups to help parents choose games for their children. Imagine Games Network (IGN) is a group which provides ratings and reviews for different games, often giving scores for games to assess the quality of the game. Steam, while not specifically a review website, is an online webstore for most games on PC, which does severely limit the data. However, Steam allows users to write reviews for their games, which we can use as an additional metric in our data.

Now, one crucial point in joining this data is which quality of the data to join. Naturally, we would use the game title, and therefore we would combine our different datasets by matching the titles with each other. However, one issue with joining these sets is the risk that the data scrape results in different character encodings. For example, some of the scrapes above include additional quotations or the copyright or trademark symbol, which is exclusive to the Unicode encoding language and may not have been part of our current dataset. Let's resolve this by creating a stripped version of the title using regular expression, which allows users to match patterns of characters together. We downcase the letters and remove all whitespace and punctuation. For example "Call of Duty: Black Ops II" is converted to "callofdutyblackopsii" to match as many copies of it as possible. However, not enough data is being joined still. This is likely because the data may include longer subtitles that other datasets might not include. For the sake of aggressively joining our data, let's also join by the first 12 characters of the data only, which will hopefully get past any failed joins due to subtitles while not being too aggressive in joining (for example, "Super Mario Bros" and "Super Mario Party" should be considered

distinct). We join the data on this shortened title. Let's now consider potential issues in our current dataset.

We notice, upon plotting some of the data, that there is a trend with Steam data that is rather problematic. They have a large number of games with either 0 ratings or an average rating of 0. In the first case, this data is not useful. In the second case, this is a statistical anomaly that we should address. It is almost impossible for a Steam game to have a rating of 0 without having a very small amount of reviews, so let's remove the data with a Steam review score of 0.

Let's now review our data not explicitly related to video game factors, namely economic data on GDP. Let's consider the World Bank's dataset on developing factors of countries. The first thing we should change is that this data does not include a single column for year but instead includes each year as its own column, formatted as XYear. Let's reshape the data and remove the "X" character in front of our year. We are then ready to combine our data. To do so, we group by continent for Europe and North America but by the country of Japan for sales in Japan.

Results

After preprocessing the data, we conducted an analysis using R to derive insights and draw conclusions from our datasets, presenting the results alongside detailed explanations.

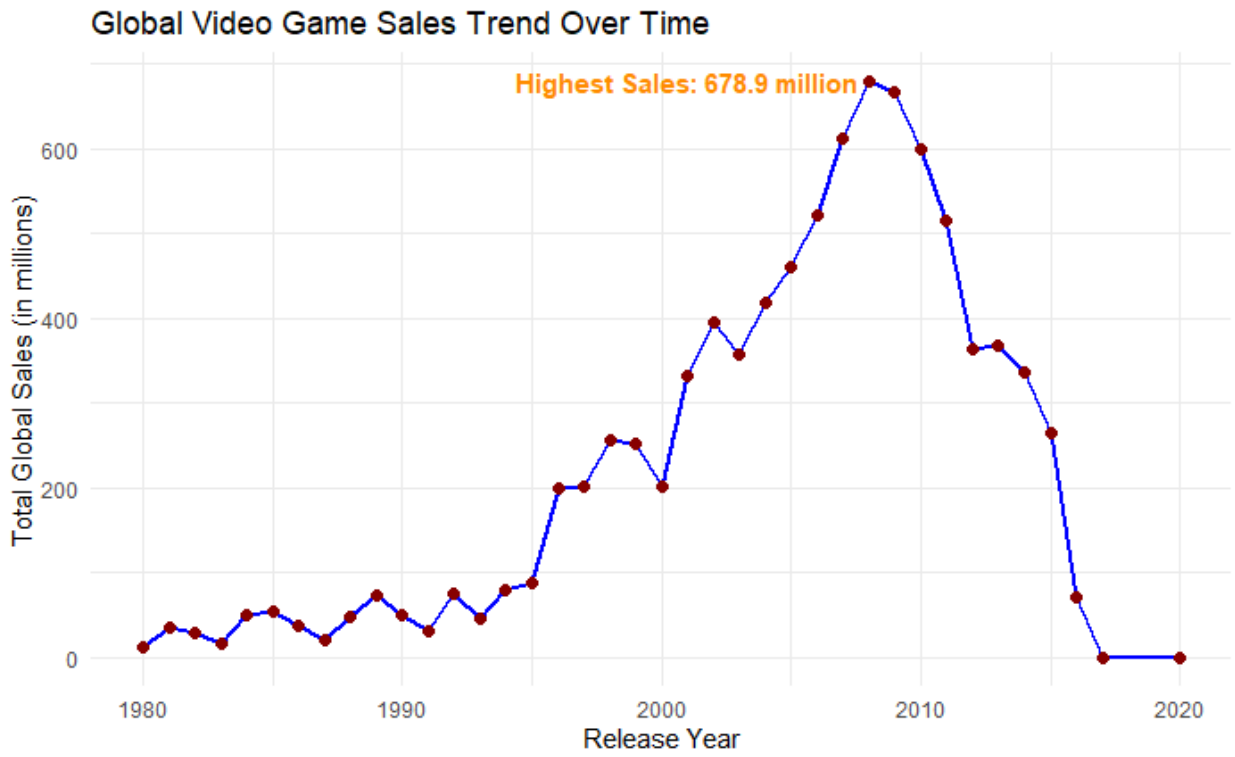

Figure 1. The Line Chart of Global Video Game Sales Trend Over Time

Figure 1 illustrates the trend in global video game sales from 1980 to 2020, highlighting significant changes over time. We observe a peak in the late 2000s, with total global sales reaching 678.9 million units, indicated in orange as the highest point. The steady decline after 2010 likely reflects market saturation, increasing competition from mobile and online gaming, and shifting consumer preferences. This trend emphasizes the evolving dynamics of the gaming industry and the challenges faced by traditional video game platforms in adapting to these changes[1].

[1] Darambazar Amgalan brought to our attention that there were many reports of mobile games on the rise after 2007, with much more optimized phones coming out at that time. His point is crucial for our investigation, and we were not able to thoroughly explore his helpful suggestion in this report.

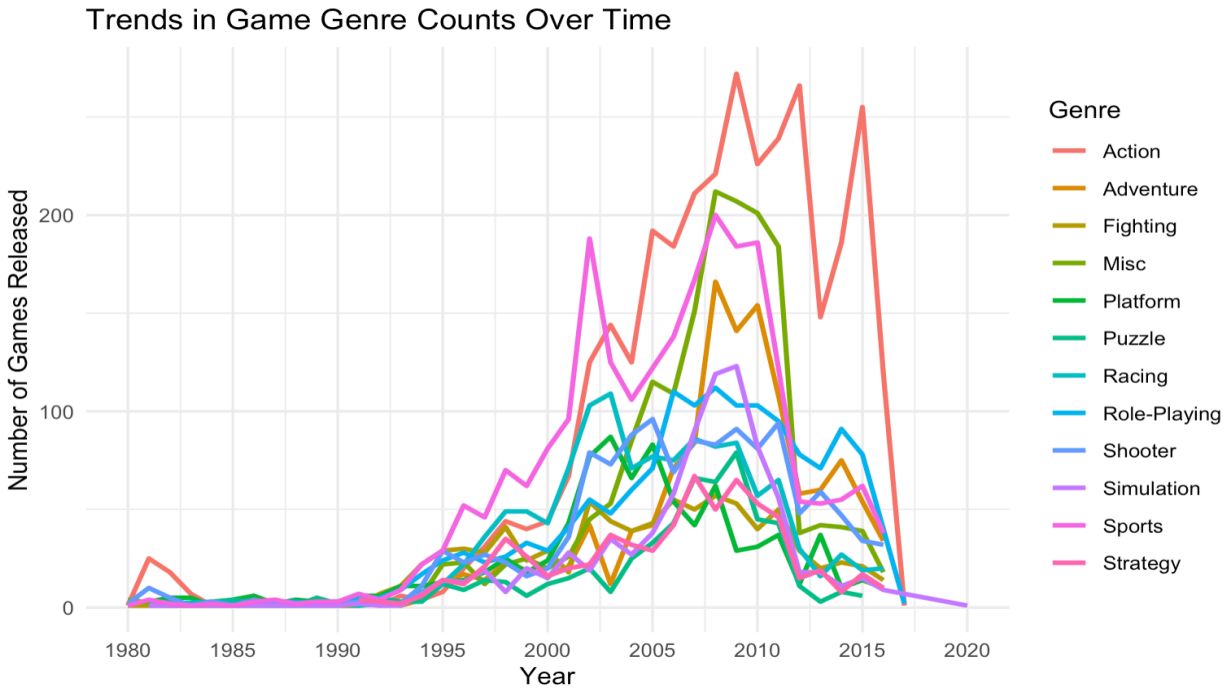

Figure 2. The Line Graphic of Game Genre Trends Over Time

Figure 2 shows the number of games released in each genre per year, allowing us to see how the popularity of different game genres has evolved. From the 1980s to early 2000s, there was a steady increase in the number of game releases across various genres, reflecting industry growth and increased interest in gaming. A notable peak appears around 2008-2010. After 2010, there is a noticeable decline. In the early period shown in the chart, sports games (pink line) had a higher count, indicating their popularity at that stage. However, over time, action games (red line) gradually caught up and eventually surpassed sports games in release numbers[2]. Over time, action games firmly established themselves as the genre with the highest number of releases.

[2] To clarify, it might seem like many popular action games were being released in higher numbers. However, we could also interpret this graph as showing a few extremely popular action games being released as well. As Devin Lim helpfully pointed out, more work needs to be done to observe the correlation between genre and sales after release.

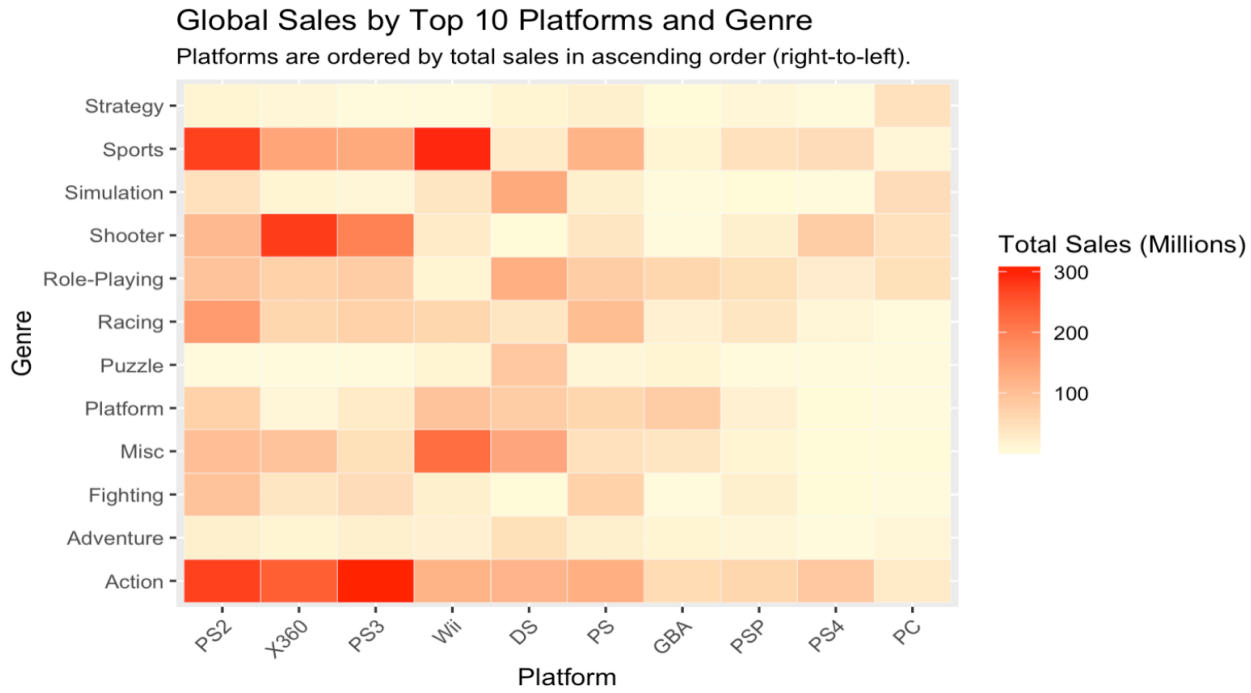

Figure 3. The Heatmap of Global Video Game Sales Across Platforms and Genres

Figure 3 illustrates the global sales of video games across the top 10 platforms and various genres. The intensity of the color represents total sales in millions, with deeper red shades indicating higher sales. The PS2, Xbox 360, and PS3 are the top three platforms in total sales, showing strong performance in genres such as Action, Sports, and Shooter. Action games stand out as the top-performing genre overall, particularly on the PS3 and PS2, while Sports games also demonstrate significant popularity, especially on the Wii and PS2. In contrast, Puzzle, Simulation, and Strategy are all relatively light in shading across the board, which is representative of their relatively low overall sales contributions.

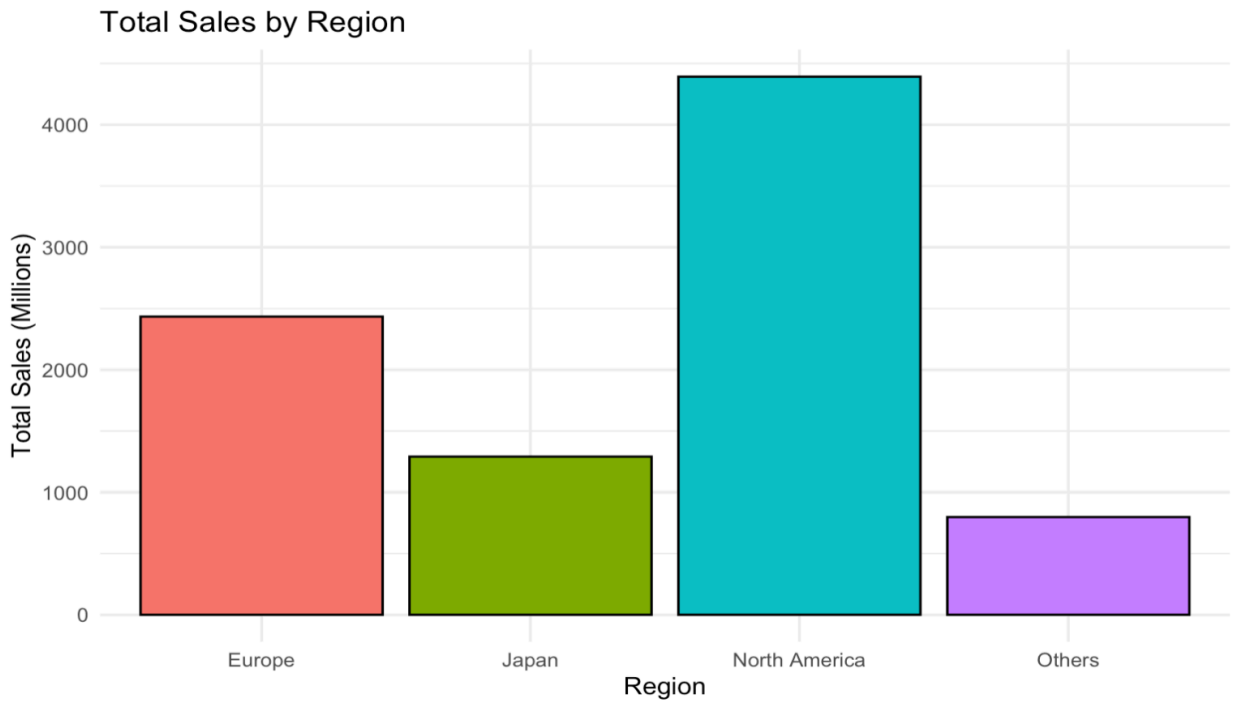

Figure 4. The Bar Plot of Total Video Game Sales by Region

In Figure 4, each bar represents a region, with its height indicating the total video game sales volume. North America clearly leads with the highest total sales, significantly surpassing Europe, Japan, and other regions. Europe ranks second, reflecting its strong presence in the global video game market. Notably, the data for Europe and North America are aggregated from multiple countries, while Japan's data represents a single country. This highlights Japan's remarkable performance and underscores its status as a major video game powerhouse, emphasizing its substantial contribution to the global gaming industry.

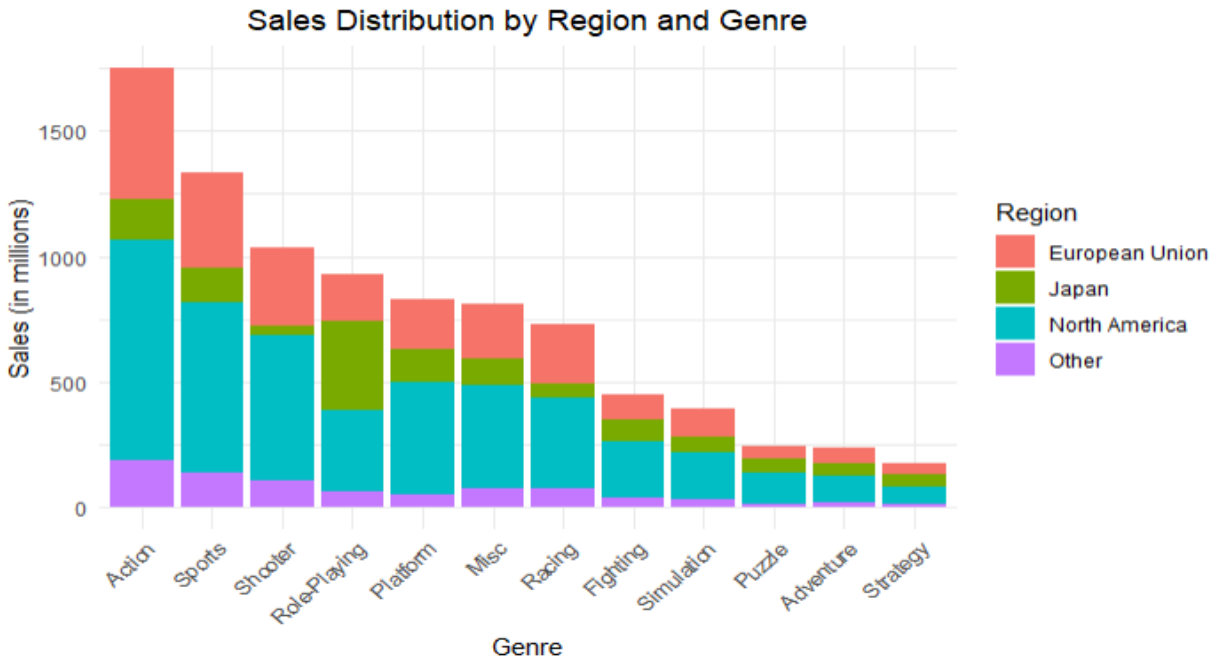

Figure 5. The Stacked Bar Plot of Sales Distribution by Region and Genre

Figure 5 illustrates the sales distribution across various game genres and regions using a stacked bar chart. Each bar represents a genre, with different colors showing the contribution of each region, including North America, Europe, Japan, and others. The chart highlights that the Action genre dominates sales, particularly in the North American region. The Sports and Shooter genres also perform strongly across all regions. In contrast, genres such as Puzzle, Strategy, and Adventure exhibit lower sales overall. North America consistently demonstrates higher sales across most genres, highlighting its significant market potential for game developers.

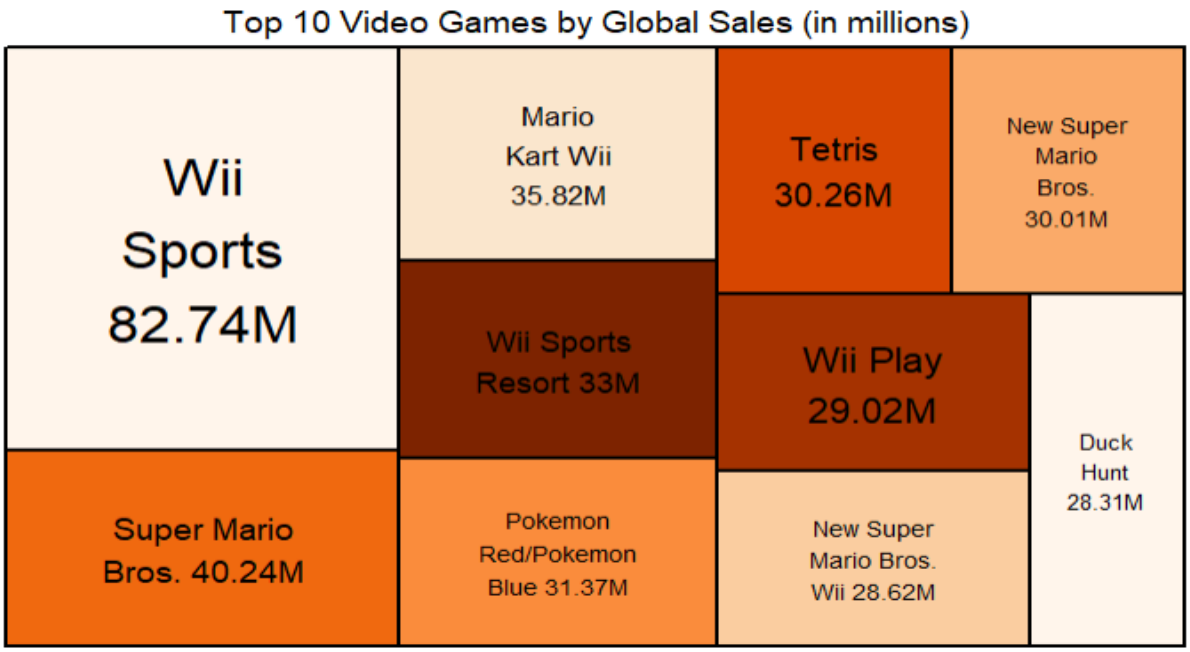

Figure 6. The Treemap Graph of Top 10 Video Games by Global Sales (in Millions)

Figure 6 effectively illustrates the proportions of global sales among the top 10 video games. Each rectangle in the treemap represents a game, with its size corresponding to its global sales volume. The result shows that "Wii Sports" occupies the largest area, representing 82.74 million global sales. "Super Mario Bros." and "Mario Kart Wii" also stand out as leading games in global sales. This visualization underscores the significance of blockbuster games in shaping the video game industry.

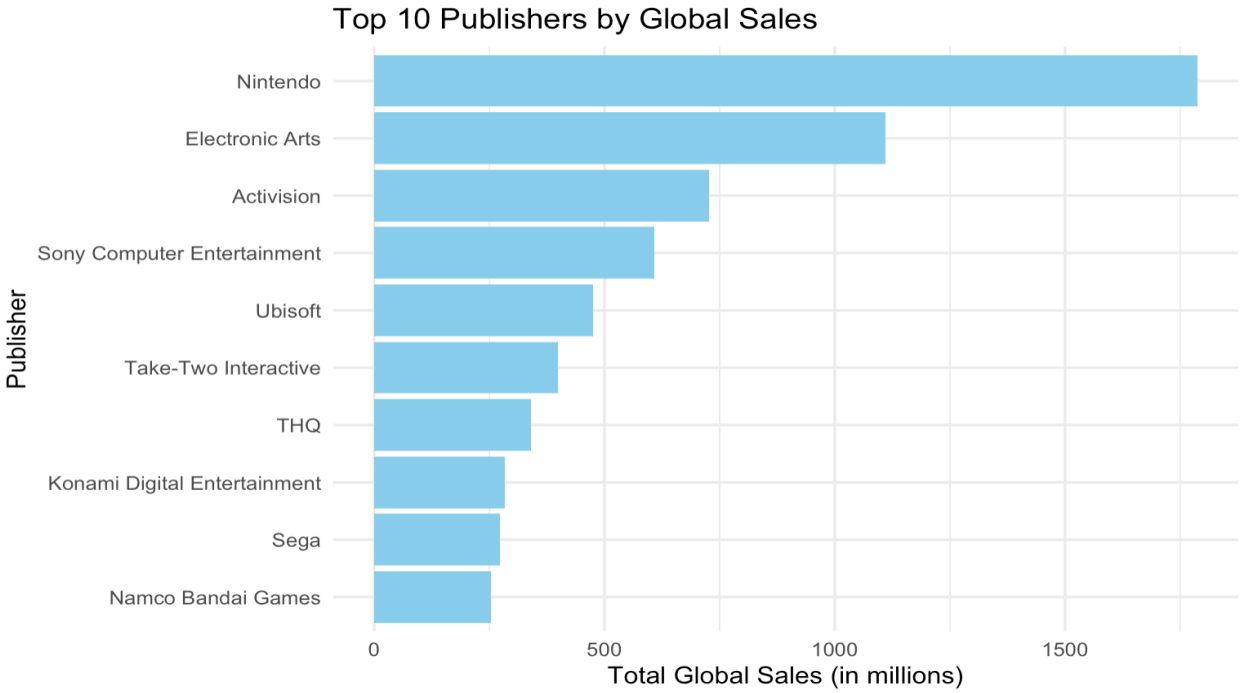

Figure 7. The Bar Chart of Top 10 Publishers by Global Sales

Figure 7 highlights the top 10 video game publishers by global sales, with Nintendo leading by a significant margin, followed by Electronic Arts and Activision. This bar chart emphasizes Nintendo's market dominance and the critical role of major publishers in shaping the global video game industry. Furthermore, many of the top-selling games mentioned in Figure 6, such as Wii Sports and Mario Kart Wii, were developed by Nintendo, reinforcing its influential position in the market.

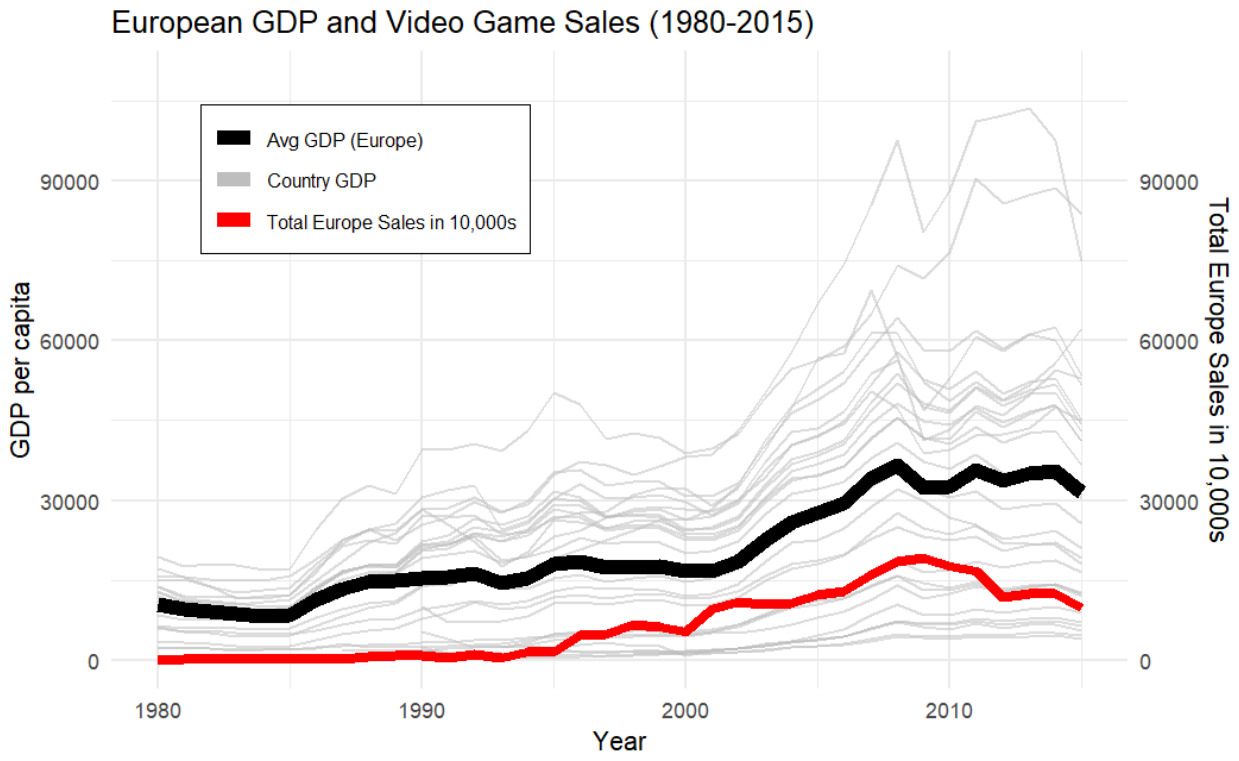

Figure 8. The Line Chart of European GDP and Video Game Sales

Figure 8 presents three key pieces of data. The multiple thin lines represent the GDP per capita of various European countries. To simplify the analysis, we averaged these values into a single thick black line. Additionally, a thick red line represents the total video game sales in Europe, based on the release year of the games. A secondary y-axis is included to reflect the scale of the red line[3]. It is important to note that the y-axis for sales does not represent the sales made in a specific year but rather the total sales of games released in that year.

While this approach differs from measuring annual sales, it is reasonable to assume that most game sales occur within the same year of release, especially for games without significant updates. The chart reveals that both GDP and video game sales experienced growth from the late 1990s until the 2008 financial crisis. However, while GDP began to recover after the crisis, video game sales did not follow the same trend. This divergence could be attributed to businesses exercising caution in an uncertain economic climate, resulting in fewer investments in the video game industry during this period.

[3] To clarify, we will use Europe instead of EU for our legend, since it is vague whether EU stands for Europe or European Union, but for the sake of brevity of the later graphs, we will use abbreviations, since there is no ambiguity with JP or NA.

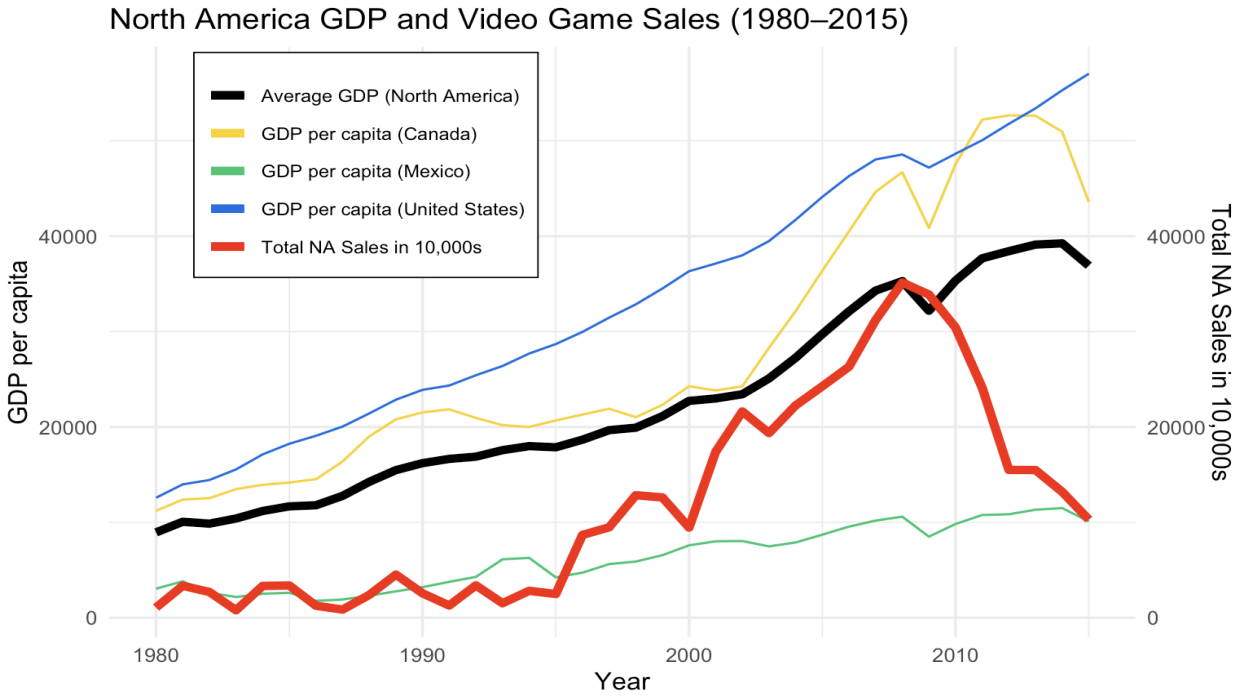

Figure 9. The Line Chart of North America GDP and Video Game Sales

In Figure 9, the thick black line represents the average GDP of the North American region, while the thick red line indicates video game sales in North America. Similar to Figure 8, both trends rose consistently from the late 1990s, peaking before the 2008 financial crisis. However, video game sales did not recover alongside GDP, likely due to businesses avoiding risks in the still recovering economy[4].

[4] As noted in figure 1, another potential explanation for why video game production dropped off after the depression is the rise of mobile games. This explanation does not need to rely on GDP trends. Thanks to Sri Satya Sai Prasanth Siddireddi and Darambazar Amgalan for bringing up mobile games as an explanation and pushing back against the above explanation.

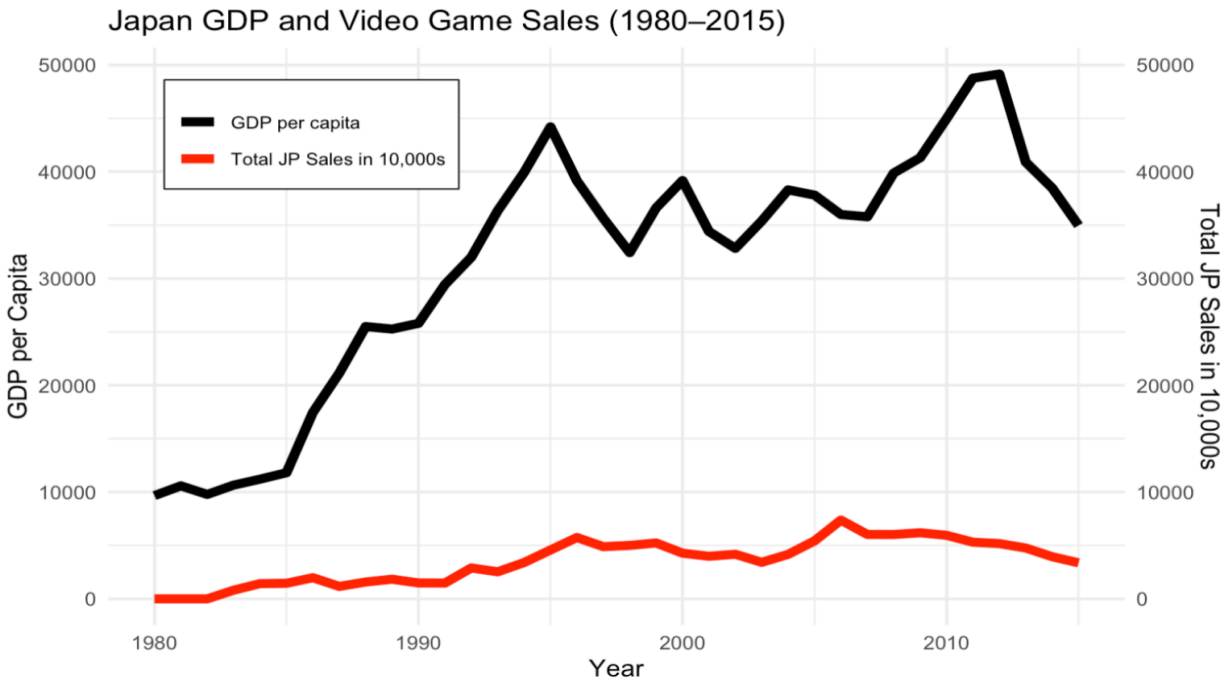

Figure 10. The Line Chart of Japan's GDP and Video Game Sales

In Figure 10, the thick black line represents Japan's average GDP, while the thick red line indicates video game sales in Japan. The chart shows that Japan's GDP experienced fluctuations, but the impact of the 2008 financial crisis was smaller compared to the European and North America regions. This is likely due to Japan's "lost decade", where it faced many financial crises. Additionally, the correlation between video game sales and GDP in Japan appears weak, as video game sales remained relatively stable throughout the period. This stability highlights Japan's prominent and consistent role in the global video game industry.

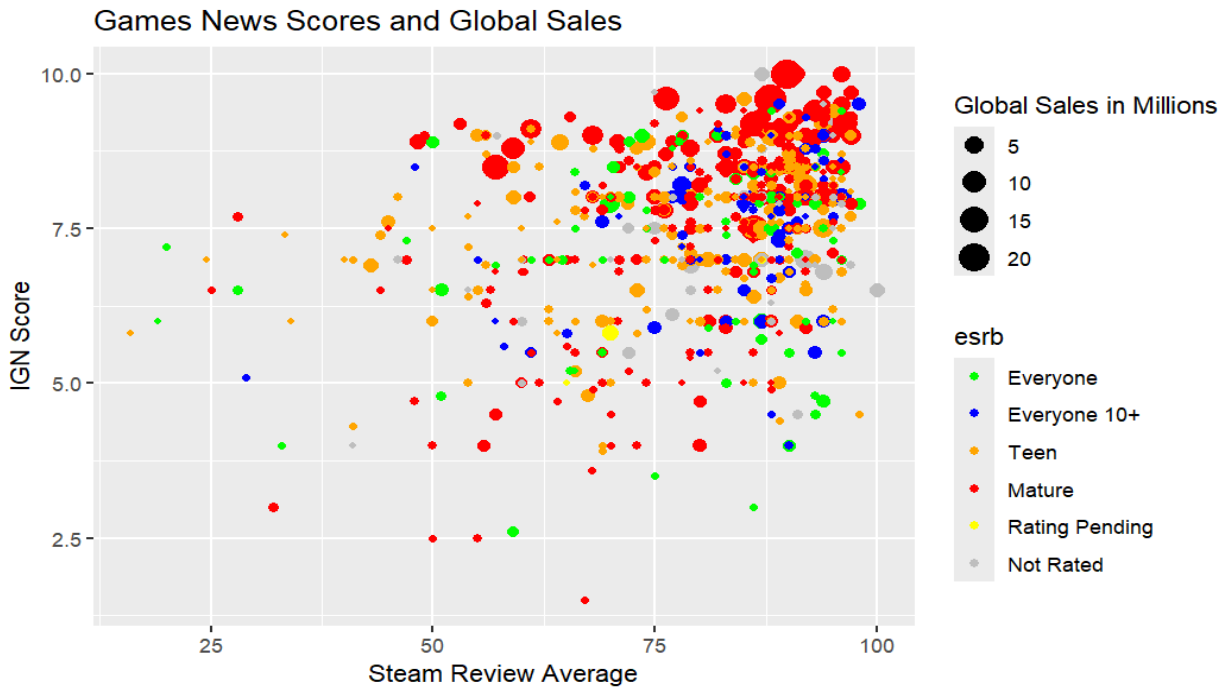

Figure 11. The Scatter Plot Comparing IGN Scores with Steam Reviews

In Figure 11, we set the average of the Steam reviews for a game on the X axis and we set the IGN score given to the game on the y axis. We allow the color of the game to be its ESRB rating and let the size of each of the bots be the global sales in millions for the point. We notice the intuitive trend that, as the scores increase, the number of sales increases as a highly acclaimed game is more likely to sell more copies. One interesting trend we see here is the high variance of Steam reviews even among high selling games. For example, Call of Duty[5] has about 60% Steam rating but still has sold on the high end of the games in the set. On the contrary, it appears that IGN rating is still able to capture many games with high sales. However, we should notice that, once a game rises above a 7.5 IGN rating, a higher score than that does not guarantee higher sales. To the contrary, there are some games with quite a high score which did not sell many copies. This shows that a higher IGN score (above 7.5) might be a necessary condition to sell well but increasing past that threshold does not strictly increase sales.

[5] One will also notice that, even among the low Steam reviews it received, the Call of Duty franchise is still selling especially well. Franchises and sequels tend to do quite well, with Call of Duty and Grand Theft Auto being examples among PC games. If such an analysis could be done with non-PC games, we might also expect a similar trend with sports series like Madden and FIFA and with Nintendo series such as Mario, but that is a further result. Thanks to Garrett Ringler for pointing out this trend of inherited popularity.

Discussion

This analysis highlights the dynamic evolution of the video game industry. It demonstrates how market preferences, regional trends, and economic factors interact to shape sales outcomes. The findings suggest that tailoring strategies to regional preferences and leveraging top-performing game genres and platforms are crucial for driving market success. Additionally, the analysis of stable markets like Japan underscores their potential for sustained growth, even under economic fluctuations.

However, there are some potential sources of misinterpretation in this analysis. For example, the interpretation of game sales data by release year assumes that most sales occur in the year of release. This assumption may not hold for games with long-tail effects or updates that happen. It will potentially lead to biases when correlating GDP with game sales.

The report also faces some challenges. The dataset focuses primarily on traditional video games, potentially overlooking emerging platforms such as mobile gaming or streaming services, which limit the scope of the analysis. Additionally, while the analysis shows a correlation between GDP and sales, establishing causation would require more detailed data on consumer behavior and economic conditions.

To enhance the precision and comprehensiveness of the analysis, several improvements can be made. Incorporating time-series analysis of game sales trends across years could provide a clearer picture of long-term performance and better align with economic trends. Expanding the dataset to include data from mobile gaming platforms and new business models, such as subscription services, would also make the analysis more complete and insightful.

By addressing challenges and expanding the scope of analysis, stakeholders can better predict consumer behavior and market dynamics. That ensures ongoing innovation and competitiveness in the industry.

Conclusion

We have taken steps to identify trends among different platforms, genres, publishers, and geographic regions with respect to sales. Our investigation has let us identify several key insights for game developers and more generally those interested in the video game industry. Namely, we see that the Playstations 2 and 3, the Xbox 360, and the Wii are competitive platforms for those interested in making games, and compatibility with those systems should be considered when the cost is low. Those interested in creating games that cater to a wide audience should consider action and sports games when brainstorming game ideas. Additionally, it may behoove aspiring developers to take some inspiration from Nintendo, Electronic Arts, and Activision, as they have been quite successful. Lastly, though Japan is a single country compared to the larger region, marketers should pay attention to the region given the role of video games in its economy. However, these results should still be contextualized in light of the fast-evolving video game industry. Recall that our data is still missing mobile game data, the burgeoning sub-industry in the broader video game landscape. Our results must be taken with the risk that such platforms and genres may drastically change when accounting for mobile games. Further investigation into mobile games will greatly reward aspiring developers.

References

"Video Game Sales." Kaggle, https://www.kaggle.com/datasets/gregorut/videogamesales/data.

"Video Games Rating By 'ESRB'", Kaggle, https://www.kaggle.com/datasets/imohtn/video-games-rating-by-esrb.

"GDP per Capita (Current US$)", World Bank DataBank, https://databank.worldbank.org/indicator/NY.GDP.PCAP.CD/1ff4a498/Popular-Indicators#

"IGN scores dataset", Kaggle, https://www.kaggle.com/datasets/advancedforestry/ign-scores-dataset.

"Steam Trends 2023", Reddit, https://www.reddit.com/r/gamedev/comments/x0qs4z/we_gathered_data_about_54000_games_in_steam_and/.