Analyzing Video Games through Sales and Ratings Data

By Derick Ka Lok Kan and Hinson Cheuk Hin Kwan

Background

The video game industry has grown immensely in popularity over the past few decades, becoming a significant part of global entertainment culture. With advancements in technology and the rise of various gaming platforms, video games have evolved into complex, interactive experiences enjoyed by millions of players worldwide. The gaming industry not only entertains but also drives substantial economic value, with revenue streams coming from game sales, in-game purchases, and esports events.

IGN (Imagine Games Network) is a prominent entertainment website that focuses on video games, films, television shows, and other media. It is widely recognized for its comprehensive game reviews, news, and features. IGN's reviews and ratings play a crucial role in shaping public opinion and consumer decisions in the gaming community, making it a key influencer in the industry.

Motivation

The disparity between professional reviews and public opinion has become a notable phenomenon in the gaming industry. A prominent example of this is "The Last of Us Part II," which received high praise from critics but faced a significant backlash from a portion of the gaming community. IGN's review praised the game for its storytelling and character development, awarding it a high score. However, public opinion was divided, with many players expressing dissatisfaction with the game's direction and narrative choices. The Last of Us 2 sales even drop by 80% in its second week. This phenomenon underscores the importance of understanding the correlation between critic reviews, public opinion, and sales performance in the gaming industry.

This disparity motivated us to delve deeper into the relationships and patterns within the gaming industry, including the roles of critics, publishers, and players. By analyzing these factors, we aim to uncover insights into how different aspects of the gaming ecosystem interact and influence each other.

Similar Study

An insightful analysis on Medium by Chinmay Tuw, titled "Analyzing Video Game Sales Since 1980" provides an overview of trends and patterns in the video game industry's sales. This study highlights the top publisher selling genres, historical top game sales, and how different platforms have fared over the years in terms of total. While this analysis offers valuable insights, we believe there is potential to delve much deeper into the data.

Goal

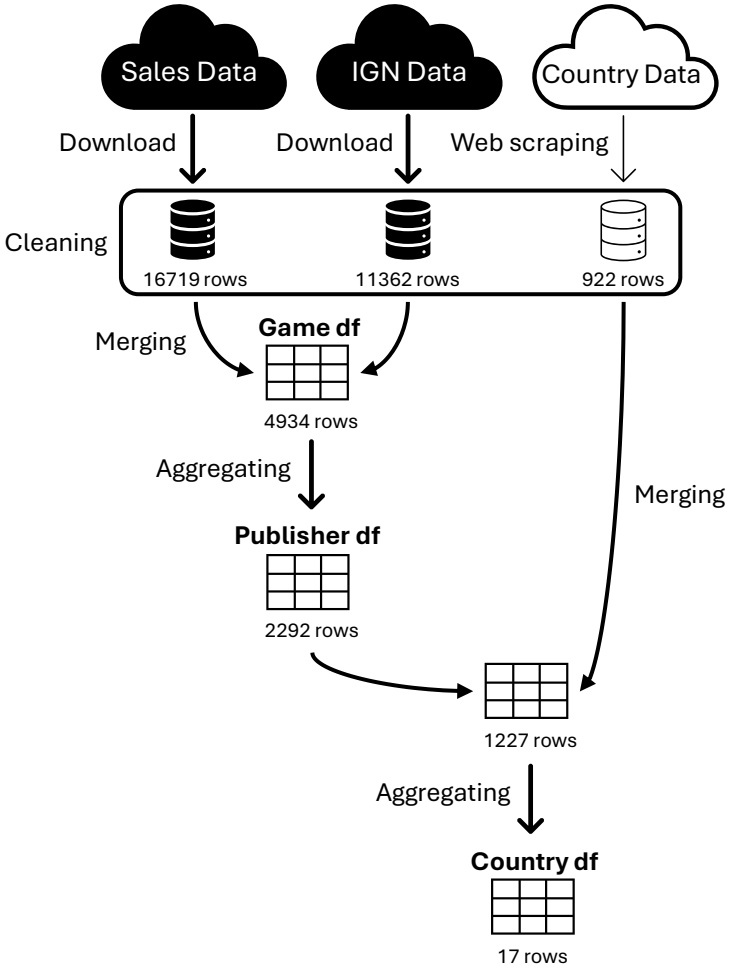
The primary goals of this analysis are:

- To explore the correlation and distribution of sales and scores across various games.
- To identify patterns and trends associated with different publishers.
- To examine how the country of origin of a publisher influences game sales and scores.

By addressing these goals, we aim to provide a more nuanced understanding of the dynamics within the video game industry

Data Source

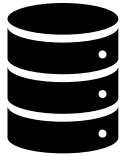
	Game Sales Dataset	IGN Scores Dataset	Publisher's Country List
Obtained from	Kaggle	Kaggle	web scraping Wikipedia using BeautifulSoup
Time Coverage	1980-2020	1976 - 2022	NA
Key Attributes:			
Name of Video Game	✓	✓	
Year of Release	✓	✓	
Publisher	✓	✓	✓
Genre of game	✓	✓	
Sales in different regions (NA, EU, JP, Global)	✓		
The rating given by users	✓		
The rating given by professional critics	✓		
The rating given by IGN		✓	
The gaming platform (e.g. PS4, Xbox)	✓	✓	
Country of publisher			✓



Data Manipulation:
Feature Selection and data cleaning

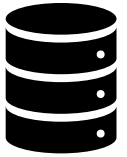
IGN dataset	Sale dataset
Dropping Irrelevant or duplicate Columns • Removing irrelevant columns along datasets to streamline the datasets	
Handling Missing Values • Dropping entire row with missing value (if any) on the critical feature **Game Name** and **IGN Score**	• Dropping entire row with missing value (if any) on the critical feature **Game Name** and **Sales** data
Data Type Conversion • Converting columns which representing lists data as strings to actual list data types, helps in handling columns that are stored as stringified lists. • Adjusting Data Formats by converting game released date from Actual Date to Year format	• Applying numeric conversions to 'User_Score' and 'Critic_Score', including handling 'tbd' entries as NaN to ensure the column contains only numeric values. • The Score in the Sale dataset was re-scaled down to a 10-point scale, standardizing score metrics across datasets.
Cleaning and Standardizing Text Data • Non-alphabetic characters (except bracket or parenthesis) are removed from game names to ensure uniformity and prevent mismatches due to formatting errors. • Double spaces in text are replaced with single spaces to maintain consistency. • Converted the text in the game name columns in datasets to lowercase and stripped trailing spaces to avoid discrepancies.	

IGN Data



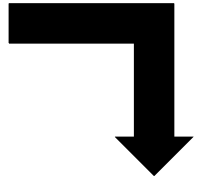
Shape: (11362, 13)

Sale Data

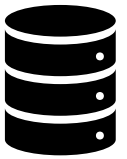


Shape: (16719, 16)

Data Cleaning

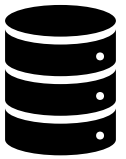


Cleaned IGN Data



Shape: (11319, 7)

Cleaned Sale Data



Shape: (16717, 10)

Data Manipulation Continued: Normalization

Game name Normalization

Since some games are published on multiple platform or as remake in different Year, stating with **[bracket]** or **(parenthesis)**. It causes the duplication on the same game.

We “normalizing” on game titles by removing extraneous text enclosed in brackets or parentheses, which allows for more accurate aggregation and analysis of data related to each game as a singular entity.

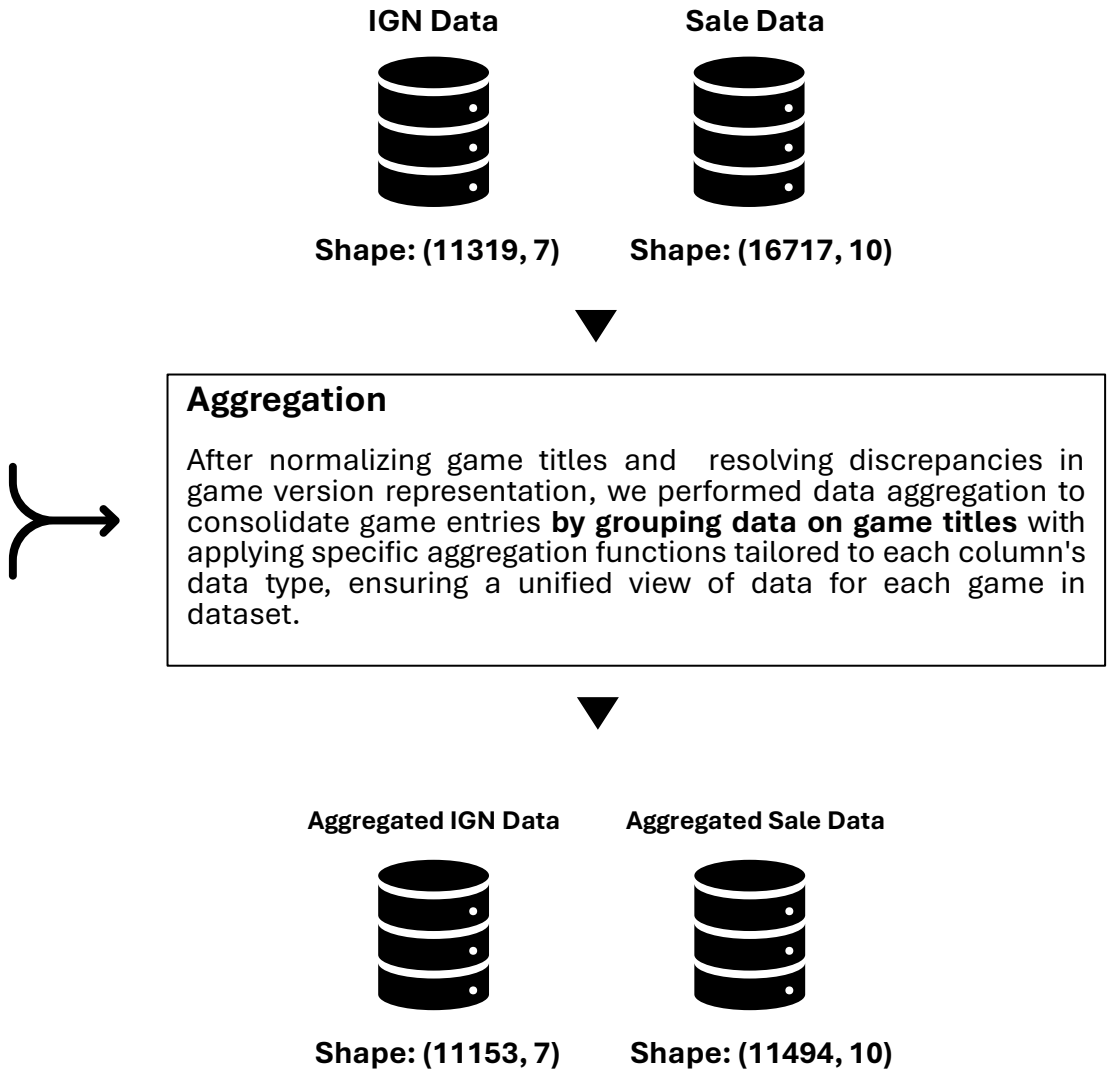
spiderman 2 ~~[gba]~~	→	spiderman 2

Mapping game with alternative version

The other issue is the discrepancy in game version representation across datasets.

In IGN dataset, games with different versions are listed as separate entries, indicating that each version is treated as a distinct game. In contrast, the Sales dataset consolidates versions of the same game into a single entry, using a **slash(/)** to denote the combined versions.
To unite the way of expression on game with alternative version, we map separate game entries with alternative versions in the IGN dataset to their corresponding combined format in the sales dataset.

pokemon blue version	→	pokemon red/pokemon blue
pokemon red version		pokemon red/pokemon blue



Data Manipulation Continued: Joining dataset

Merging Datasets

After taking adjustment on game names in previous step, we then performed an **inner join** between the datasets based on the mapped game name to consolidate game-related data through the pandas **.merge()** method. This operation was performed to create an intersection of both datasets where the game names match.

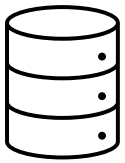
Verification

We implemented **TF-IDF Vectorization** on those game names from both datasets that did not successfully merge but are highly similar, to identify instances where game names from the two datasets might not have merged due to slight variations in spelling or formatting:

lemmings 3d	3d lemmings
metal slug	metal slug 7
metal slug 2	metal slug 7

And after reviewing those unmatched data, we found some of them required manual review and adjustment. Considering the workload, efficiency and the benefit gain from this action, we decided end up here and move on .

Finally, we arrange column retention and cleaning on merged dataset.



Merged dataset

Shape: (4934, 14)

Datasets Preparation on different scope

Aggregated by publisher

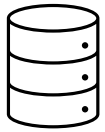
To take different level of analysis for understanding the gaming market dynamics based on publisher , we

- Explodes the Publishers column, enabling aggregation by individual publishers.
- Aggregates remaining data by publisher and release year.

Analysis by Country level

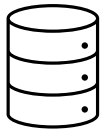
Preparing data that taking country level, we

- Fetches additional data about publishers from **Publisher‘s Country List** dataset, mapping each publisher to a country of origin.
- Merges this geographical data with the aggregated game data, adding a new dimension of analysis by publisher's location.
- Further aggregates data by country, providing a higher-level view of the gaming industry’s metrics, such as average scores and sales, grouped by country.



Publisher dataset

Shape: (2292, 10)



Country dataset

Shape: (17, 9)

Analysis:

Correlation of IGN score and global sales

Considering IGN was created in 1996, and the limited number of games data before 1996, we grouped the game data from before 1996 into one category for creating a more robust and statistical analysis (figure 1b). From 1996 onward, the data is segmented into 5-year intervals to observe trends over time.

From the plots, we can tell:

Consistent Positive Correlation

The correlation coefficients (denoted as r) across different time periods range with **weak positive correlation**. This suggests that while there is some association between game scores and sales, it's not particularly strong but this is consistent across all time periods as a positive proportion.

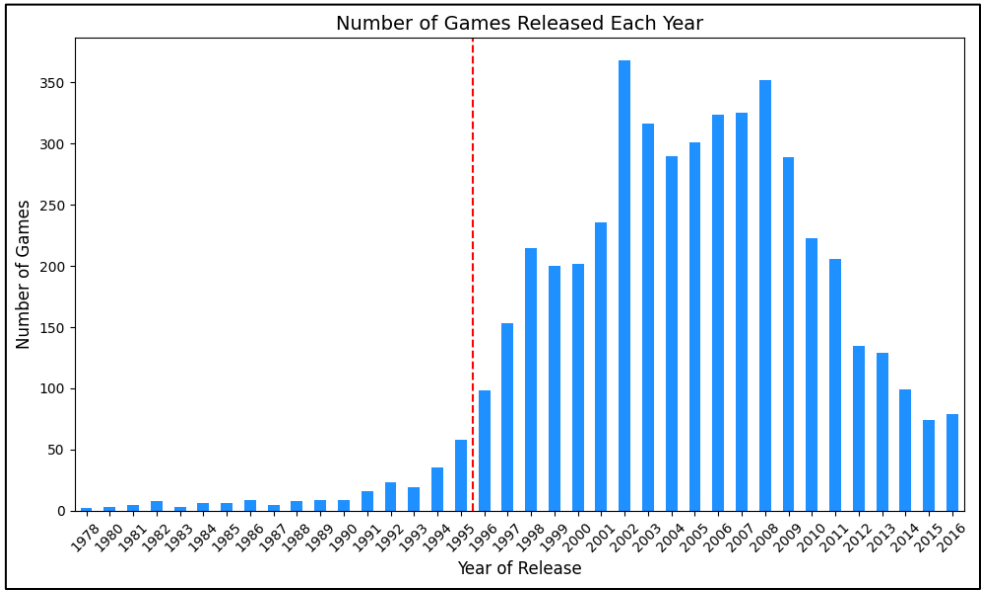
Correlation Strength Over Time

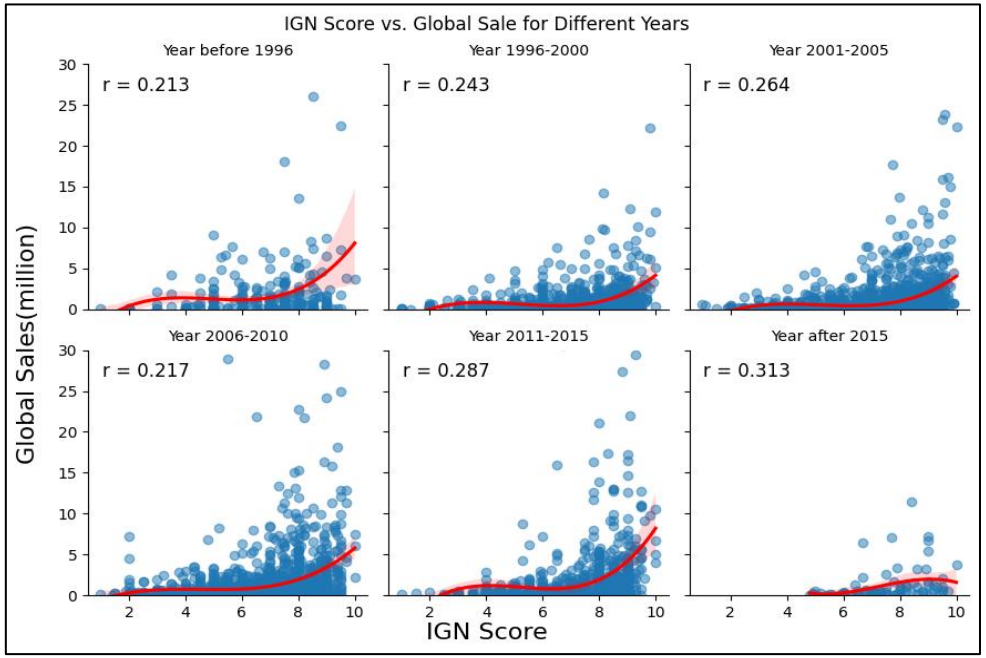
From the years before 2006, the correlation is modest but keeping an increasing on the coefficient continuously, which indicated the impact of critical reviews on sales became slightly more pronounced.

But in period 2006-2010, the correlation drops again among periods, it emphasizes some games that despite lower scores, achieved significant sales, which might be presented by outliers influencing the data.

And for the years after 2011, there is a noticeable increase in correlation, likely indicating a consolidation of market trends where highly-rated games are strongly aligned.

The correlation between IGN scores and global sales keeping a positive Correlation and exhibits modest fluctuations across different periods. Although the correlation dropped slightly in 2006-2010, it stabilizes with a stronger relationship from 2011 onwards and peaking in latest years in dataset. **This trend suggests an increasing in correlation of game reviews on sales by years.**





Score trending and uncertainty analysis

Chart Overview

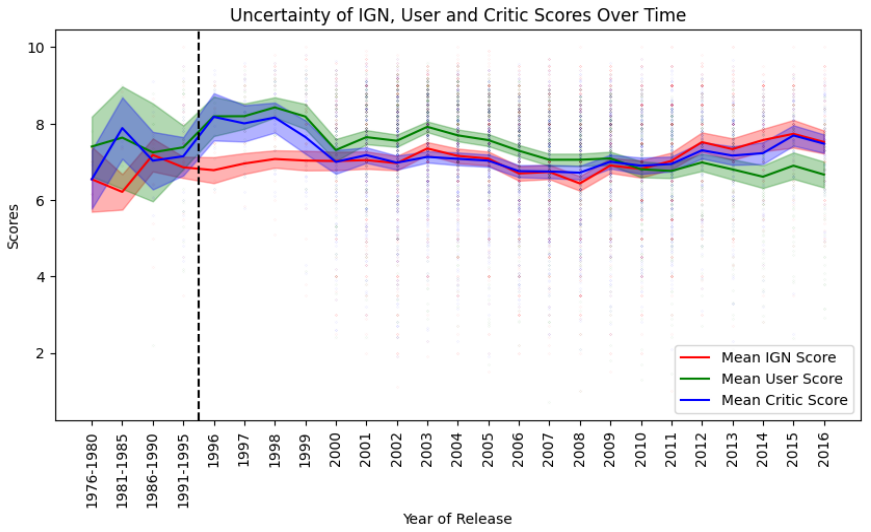
The visualization presented is a comprehensive analysis of the uncertainty of IGN, User, and Critic scores on video games over time. The chart displays the mean scores for each category, with the shaded areas representing the 95% confidence intervals and points illustrating the distribution of scores for each year. Due to the limited data availability before 1996, the periods before 1996 were grouped into 5-year intervals to enhance visualization clarity.

Public vs. Critic Perception

Despite the challenges posed by limited data before 1996, the overall differences between the three types of scores (IGN, User, and Critic) are **minimal**, about 1 point of score. However, a notable trend is observed where user scores exhibit a **slight downward** trajectory, whereas IGN and Critic scores show a **slight upward** trend. The intersection around the year **2010 indicates a shift** where public ratings began to be generally lower than those of IGN and other critics. This suggests a divergence in perception between the public and professional reviewers starting around 2010. It questions about the factors influencing public perception versus professional reviews. Potential reasons for this divergence could include changes in the gaming industry, differences in expectations between the public and critics, or the impact of social media and online reviews.

Stability and Variability of Scores

The period between 2000 and 2010 shows **remarkable stability** in scores, as indicated by relatively narrow confidence intervals. However, post-2010, the scores become **less stable**, with wider confidence intervals. This increase in variability could reflect a period of rapid change in the industry, with the introduction of new gaming platforms, genres, and shifts in consumer preferences. The pre-1996 period is characterized by high instability, likely due to the sparse data available during those years. An important observation is the **positive correlation between the amount of data and the variation in scores**. This is evident from the discrete point distribution, where years with more data points show more stable scores, while years with fewer data points exhibit greater variability. This correlation suggests that the observed variations might be influenced by the volume of available data rather than solely reflecting the true variability in game scores. This is a crucial consideration for us when interpreting the uncertainty analysis.



Game Genre Performance on Sales and Scores

Correlation Overview

From an analysis by game genre, despite expectations that higher scores might correlate with higher sales, the plot **does not distinctly show this relationship** across genres. Each genre exhibits varied patterns, with some high-scoring games not necessarily translating to high sales, and vice versa.

This suggests that factors beyond individual game ratings, including genre-specific market, publisher loyalty or brand recognition, play significant roles in determining sales outcomes.

Genre-specific Performance

Indicating the width of the confidence intervals for both sales and scores showing that the **performance of games can significantly vary within genres**. Some genres, such as First Person or Compilation, contain a wide confidence interval for sales or scores. This suggests significant variability within these genres, which could be attributed to the diverse types of games grouped under a single genre label.

This result indicate some genres' capacity to include both high-performing blockbusters and lower-tier games, which can vary significantly in quality and market reception.

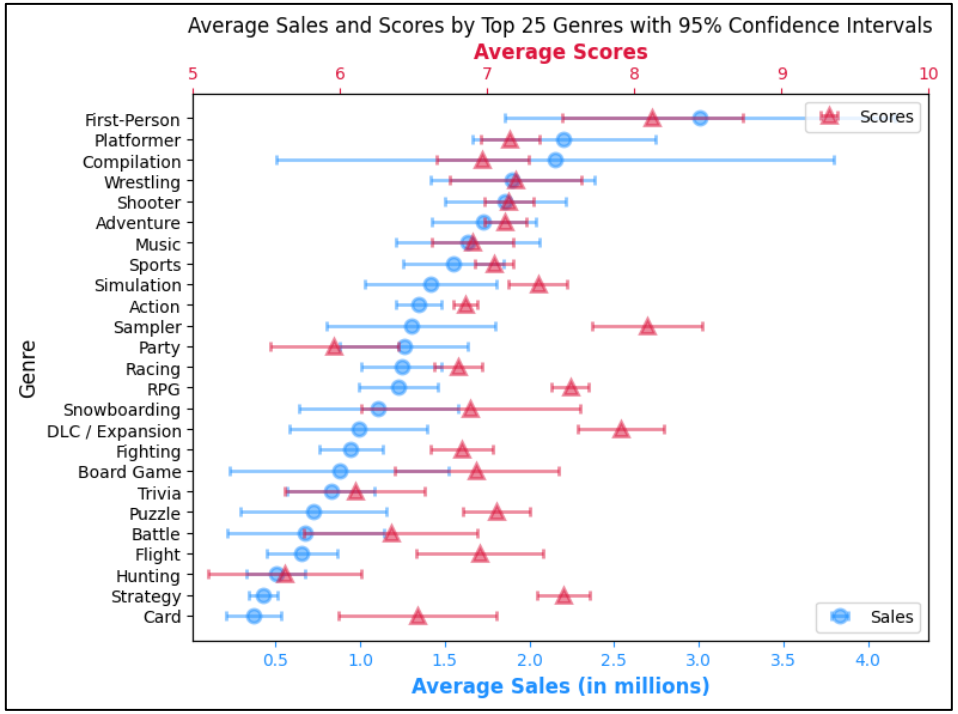
Market niche games

At the same time, there Genres with narrower audience appeal, such as Puzzle and Trivia, tend to have lower sales volumes but can achieve moderate to high scores, suggesting that niche games can be well-received critically despite lower sales figures. Therefore, understanding the dynamics of each genre could be crucial for predicting sales outcomes or market rating.

This analysis illustrates that the correlation between game scores and sales is not straightforward and varies significantly across different genres.

The wide confidence intervals observed in certain genres highlight the diversity within those categories, also noticed the niche genres in the market.

This complex interplay highlights the need for a multifaceted approach in game development and marketing strategies to optimize both score rating and commercial success.



*Sorted by average sales for enhanced readability

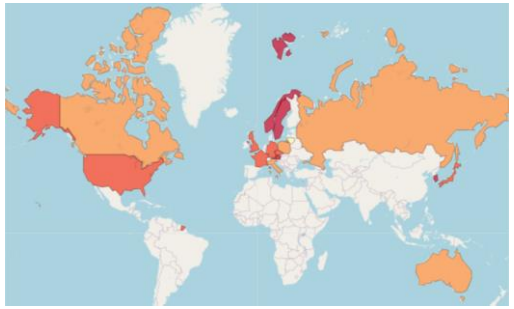
Geographic Heatmap Analysis of Video Game Scores and Sales

Chart Overview

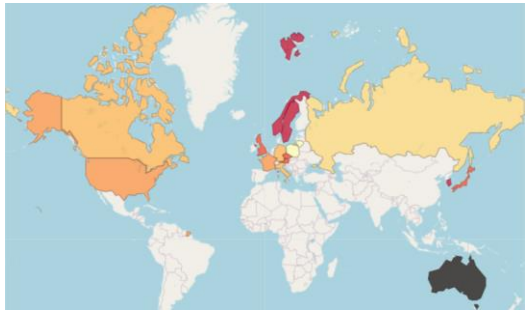
The visualization consists of two sets of geographic heatmaps. The three heatmaps on the left-hand side represent the distribution of scores from IGN, other critics, and users. The three heatmaps on the right-hand side depict the sales distribution of publishers for North America, Europe, and Japan. In these heatmaps, red indicates the highest amount, yellow indicates the smallest amount, and black indicates no data.

Score Distribution

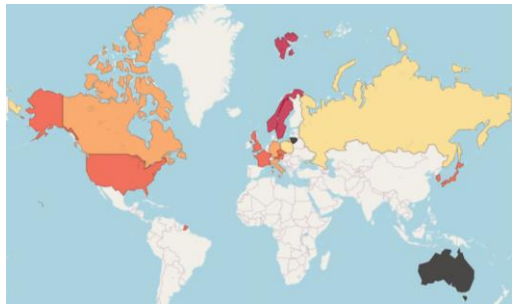
For the score distribution, the patterns reveal that publishers from Europe and Japan receive consistent ratings across IGN, other critics, and user reviews, indicating a general consensus on game quality from these regions. This consistency suggests that games from **Europe and Japan are perceived to maintain a high standard**, leading to similar evaluations from both professional critics and general users. IGN and user scores are notably **more positive towards North American** publishers, suggesting regional favorability or higher appreciation for games from these countries. Conversely, critic and user scores tend to be **more negative towards Russian** publishers, which could be due to regional biases, perceived quality differences, or historical context influencing the reception of Russian games.



IGN score



Critic score



User score

North America sales

Europe sales

Japan sales

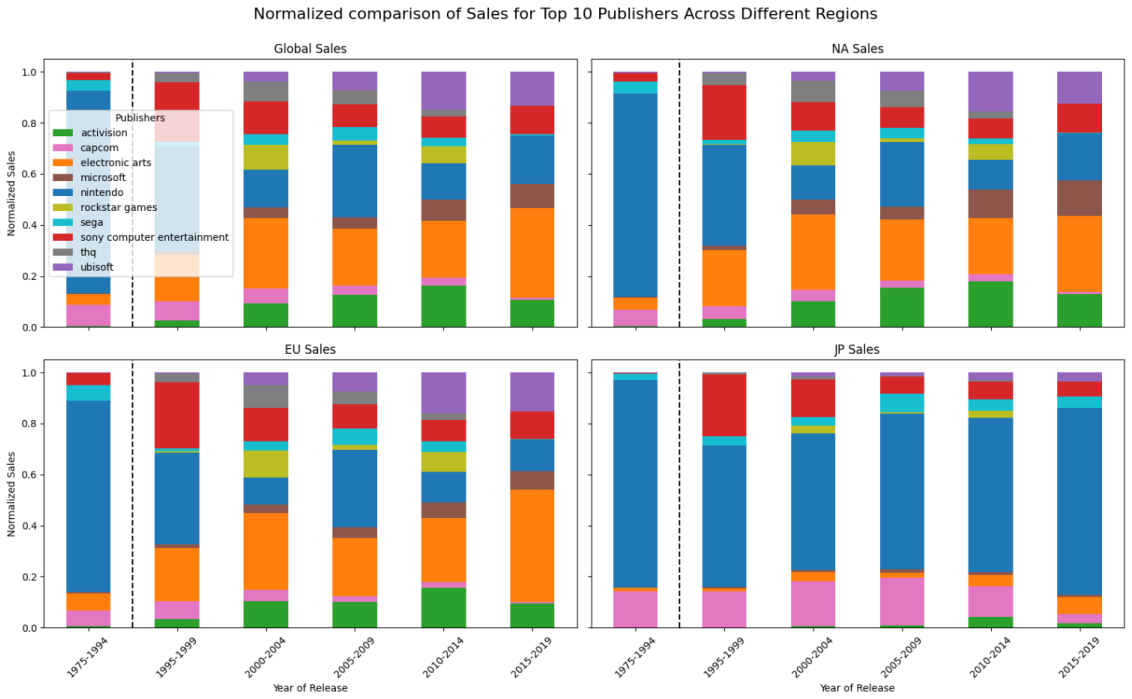
Sales Distribution

Each sales heatmap represent the sales distribution in each region, while the map colors indicate the country of the video game publishers. A significant finding is the **concentrated dominance of Japanese publishers in the Japan** sales map, illustrating a strong preference for local games likely driven by cultural affinity and language barriers, especially in earlier years when foreign games lacked Japanese language support. In contrast, while North American and European sales show a concentration in publishers from France, they still exhibit a **more even distribution among international publishers** compared to Japan. This even distribution suggests that North American and European gamers are more open to games from various countries, reflecting a globalized market with diverse preferences, unlike the localized preference seen in Japan.

Sales proportion of top 10 publishers across different regions

Chart Overview

This visualization presents a normalized comparison of sales for the top 10 publishers across different regions. The sales data are cumulative, and direct comparison over release dates is not feasible due to the inherent advantage earlier releases have in accumulating sales. To address this, the sales data are grouped and normalized within periods of release years, allowing for comparison of the sales proportion of the top publishers within each period. The vertical line in the charts indicates a period with potential data insufficiency, marking the data before this line as possibly unreliable.



General Trends

The normalized sales data reveal regional preferences and market dynamics among the top video game publishers. In Japan, the sales focus predominantly on Japanese publishers such as Sony, Sega, Nintendo, and Capcom. These publishers also contribute significantly to sales in other regions, but their dominance is less pronounced compared to Japan. This finding aligns with the earlier heatmap analysis, highlighting the **strong local preference for Japanese publishers** in their home market.

The sales figures in Europe and North America closely resemble the global sales distribution. This similarity suggests two key implications:

1. The preferences for publishers are **similar among NA and EU markets**, indicating a shared taste in video game publishers.
2. The NA and EU markets are substantial enough to significantly influence the global market trends, reflecting their **large consumer base and market size**.

Publisher-Specific Trends

- **Nintendo:** Despite the unreliable data period, Nintendo shows significant growth in the **Japanese market**. However, its market share varies across time in other regions.
- **Capcom, Sega, and Sony:** These publishers exhibit a **decreasing trend** in market share across all regions. This decline suggests that their dominance is waning, potentially due to increased competition or shifts in market preferences.
- **Microsoft and Electronic Arts:** These publishers demonstrate an **increasing trend** in market share across all regions. Their growth indicates a rising preference for their games, potentially driven by successful franchises or effective market strategies.

The trends indicate that Japanese publishers are experiencing a decreasing market share globally. This shift suggests that **foreign publishers are gradually replacing Japanese publishers**, reflecting a broader trend of globalization and changing market dynamics in the video game industry.

References & Statement of Work

1. IGN. (2020). The Last of Us Part 2 Review. Retrieved from https://www.ign.com/articles/the-last-of-us-part-2-review
2. Miska, B. (2020). Review: 'The Last of Us Part II' Takes Us on an Uneven, Overlong Journey Filled with Bloodshed & Horror. Bloody Disgusting. Retrieved from https://bloody-disgusting.com/reviews/3621556/review-last-us-part-ii-takes-us-uneven-overlong-journey-filled-bloodshed-horror/
3. Republic World. (2020). The Last Of Us 2 Sales Drop 80 Percent In Its Second Week. Retrieved from https://www.republicworld.com/tech/gaming/the-last-of-us-2-sales-drop-80-percent/?amp=1
4. Tuw, C. (2020). Analyzing Video Game Sales Since 1980. Medium. Retrieved from https://chinmaytuw.medium.com/analyzing-video-game-sales-since-1980-42d23ea7b59f

Data Sources:

1. Video Game Sales Dataset. (2022). Kaggle. Retrieved from https://www.kaggle.com/datasets/ibriiee/video-games-sales-dataset-2022-updated-extra-feat
2. IGN Scores Dataset. (2022). Kaggle. Retrieved from https://www.kaggle.com/datasets/advancedforestry/ign-scores-dataset
3. Wikipedia. List of video game publishers. Retrieved from https://en.wikipedia.org/wiki/List_of_video_game_publishers
4. Geometry data for country names to support geographic visualizations https://www.naturalearthdata.com/downloads/110m-cultural-vectors/

	Derick Ka Lok Kan	Hinson Cheuk Hin Kwan
Exploring dataset	✓	✓
Data cleaning	✓	✓
Data merging	✓	✓
Data aggregation	✓	✓
Correlation of IGN score and global sales		✓
Score trending and uncertainty analysis	✓	
Game Genre Performance on Sales and Scores		✓
Geographic Heatmap Analysis of Video Game Scores and Sales	✓	
Sales proportion of top 10 publishers across different regions	✓	
Code review	✓	✓
Report design	✓	✓