MVP Data Engineering

**Summary**

### **Objectives:**

This section comprises the questions to be solved with the data analysis.

### **Data Collecting:**

This section shows where, how and why the data was chosen.

### **ETL:**

This section breaks down how the data extraction, transformation, and loading (ETL) were performed.

### **Analysis:**

This section depicts how the analysis of the data was conducted..

### Quality of Data:

This section shows how the data was filtered and enhanced.

### **Fulfilling Objectives:**

This section shows if the questions were answered.

### Metadata:

This section presents metadata about the dataset.

### **Conclusion:**

This section provides an inference on the obtained and unobtained responses.

### **Self-Critique:**

This section provides a self-evaluation of this project.

**Objectives**

In this section we will talk a little bit about the problem we want to solve.

1. Is it better for a company to focus solely on one genre and type of game, or is diversification the key to success?
2. Are the “best” games the most sold?
3. Why did Nintendo insist on portable consoles, and is there an indication that the future hybrid console would work?
4. Is the PC really the most profitable gaming platform?
5. Is there a correlation between Nintendo's success with its games and its disregard for Western public opinion?

**With these questions in mind, let’s dive into the plan.**

We will begin by searching for data related to the video game industry and sales, aiming to find a dataset as close to the year 2017 as possible. Once the dataset is located, we will need to extract it and verify its integrity to ensure that the data is accurate and complete.

**After extracting and validating the data,** the next steps will involve transforming it to prepare it for analysis. This includes cleaning the data, addressing any issues, and making necessary adjustments.

**Following the data transformation,** we will upload the data to Databricks Community Edition, where we will model and enhance it for further analysis.

**Once the data is prepared,** we will publish it and download it as a CSV file for use in a BI tool. We will then select an appropriate BI tool for analyzing the data and creating visualizations.

**This approach will allow us to effectively address the research questions** and achieve the objectives set for this project.

**Data Collecting**

Where:

The data was collected from Kaggle, and this is the specific dataset used:  
[Video Games Sales Dataset 2022 Updated Extra Features](https://www.kaggle.com/datasets/ibriiee/video-games-sales-dataset-2022-updated-extra-feat).

**How:**

To collect the data, the process was very simple. I downloaded the zip file from the Kaggle website and unzipped it. After that, I checked to ensure the data was not corrupted. Once verified, I uploaded the dataset to Databricks Community Edition on the cluster I had previously created for this project.

**Why:**

This dataset was chosen because it comprises a large volume of data, most of which is of good quality. Additionally, it is the most up-to-date data I found for free. It has everything needed to answer my questions and fulfill my objectives.

It is important to have both user and critic counts and scores. Additionally, it was crucial that the dataset includes data for different platforms for the same game.

**ETL**

**Extraction:**

Just as previously explained, the extraction was done by finding the dataset and then down*loading* its zip file from the Kaggle website. After that, I unzipped the file, checked its integrity, and uploaded it to Databricks as a Table.

Transformation:

In this step, after extracting my data, I used Databricks Community Edition with SQL. First, I set the first row as the header of the table. After that, I identified and deleted the rows that had null values in any column. Without going into further details, which will be covered in the Quality of Data section, I also curated the data with queries to answer the questions outlined in the Objectives section.

Loading:

After *transforming* my data, I downloaded each table with the labeled queries as CSV files. Then, I uploaded these CSV files to Tableau Public for analyzing and creating visualizations.

Analysis

All the analyses were done in Tableau Public. You can view them here using this link:

[Tableau Public Dashboards](https://public.tableau.com/app/profile/lucas.maestrelli/vizzes)

Analysis on Video Game Genres

For the first analysis, I filtered for the top 10 companies and then used a text function to show the total sales in millions of dollars by genre.

The second view was created with a pie chart that divides the total sales by genre.

Is the “Best” Game the most successful?

Here, we have three sheets that filter the top 100 games according to User Score, Critic Score, and Global Sales. Each table shows the games ranked from best to worst, with the sheets displayed side by side for easy comparison.

Console or Portable?

In this analysis was used the pie chart to show the sales of the different Nintendo Platforms, we filtered the global sales and the Japanese sales

PC is the best gaming platform?

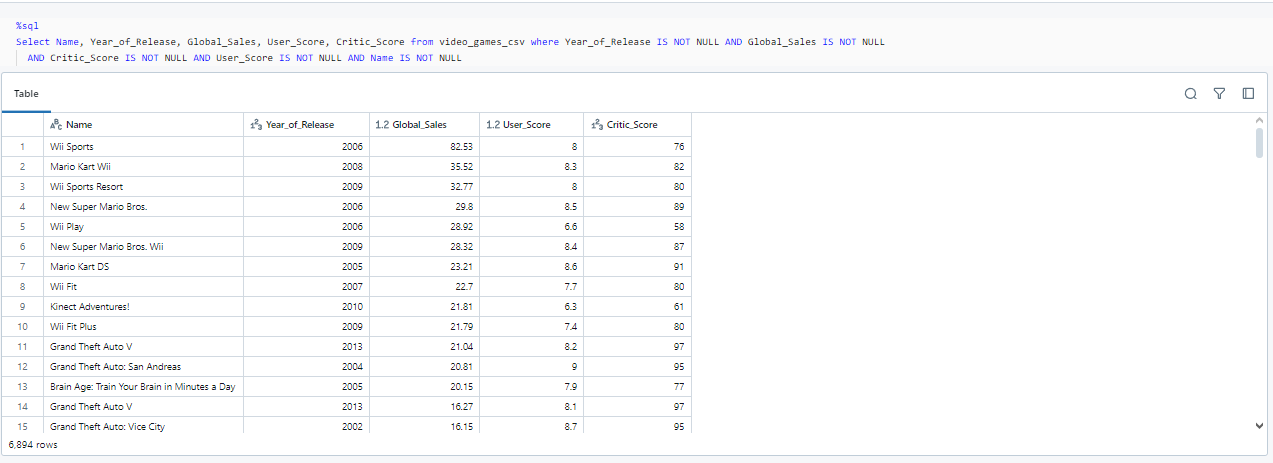
In this view, the relevant platforms since 2000 were filtered to include only those that have not yet stopped launching new games.

Nintendo Sales Perspective

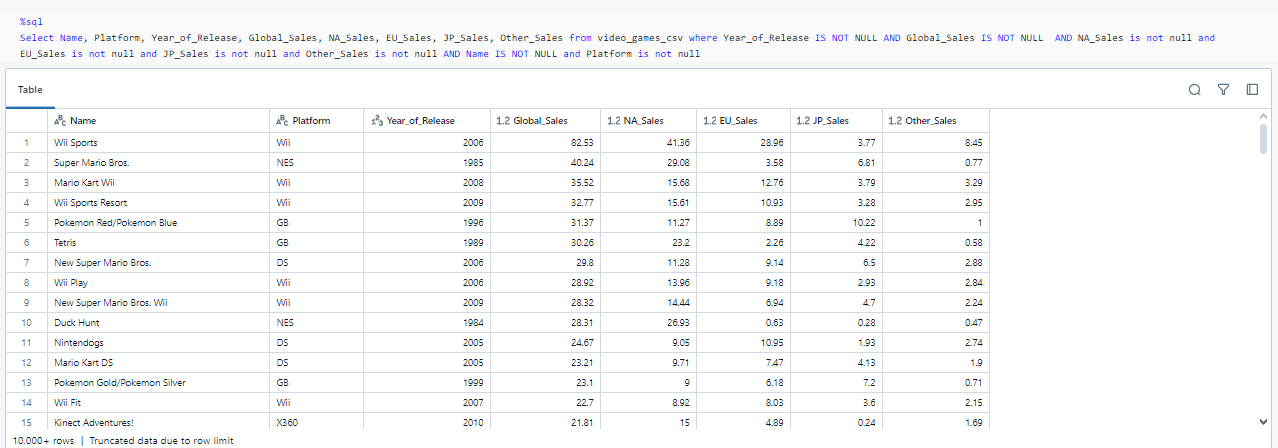
Here, we have three sheets that filter the top 50 Nintendo games by NA Sales, JP Sales, and Global Sales. Each table ranks the games from the highest to the lowest sales, with the sheets displayed side by side for easy comparison.

Quality of Data

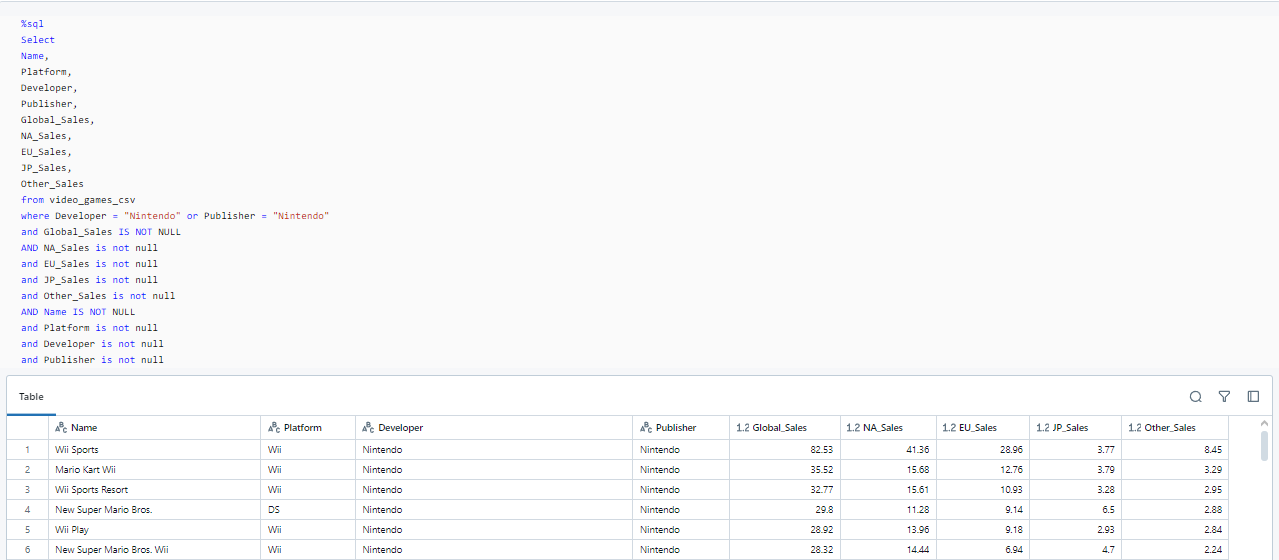
The first thing that was done to the dataset was to remove all null values from the table. After that, numerous queries were made in Databricks Community Edition using SQL for each part of the analysis to help answer the questions we have. You can find those queries and their final products here: [Notebooks](https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1411143408865230/4308825286209158/4579165488638907/latest.html)



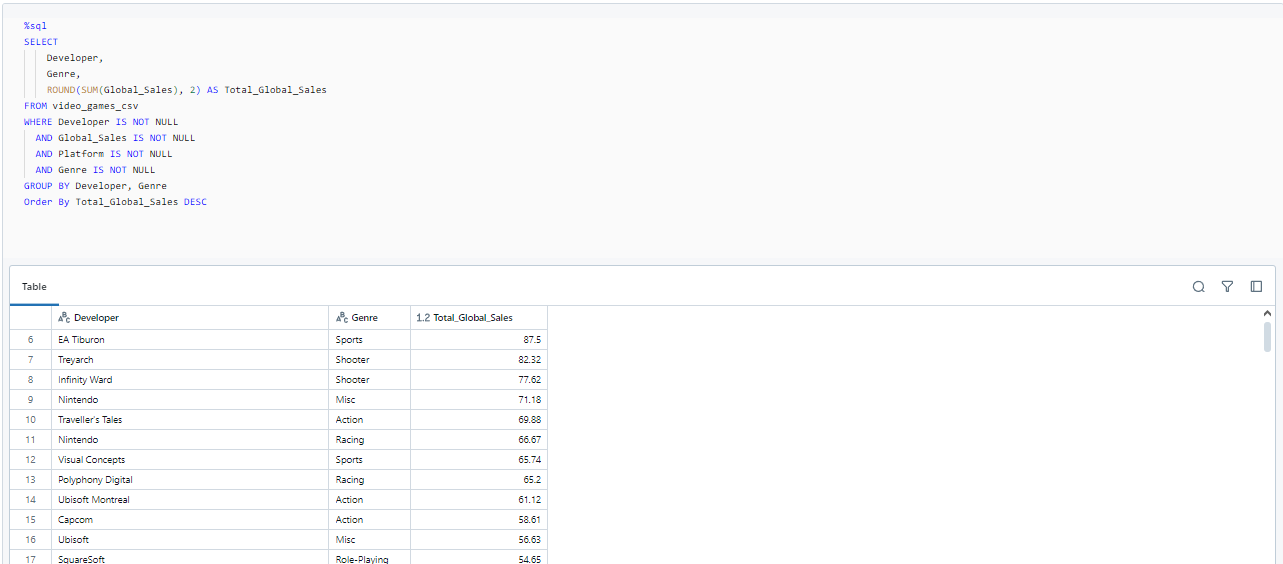
*This analysis was created to compare the user score, critic score, and global sales, which will be used to help answer question number 2.*



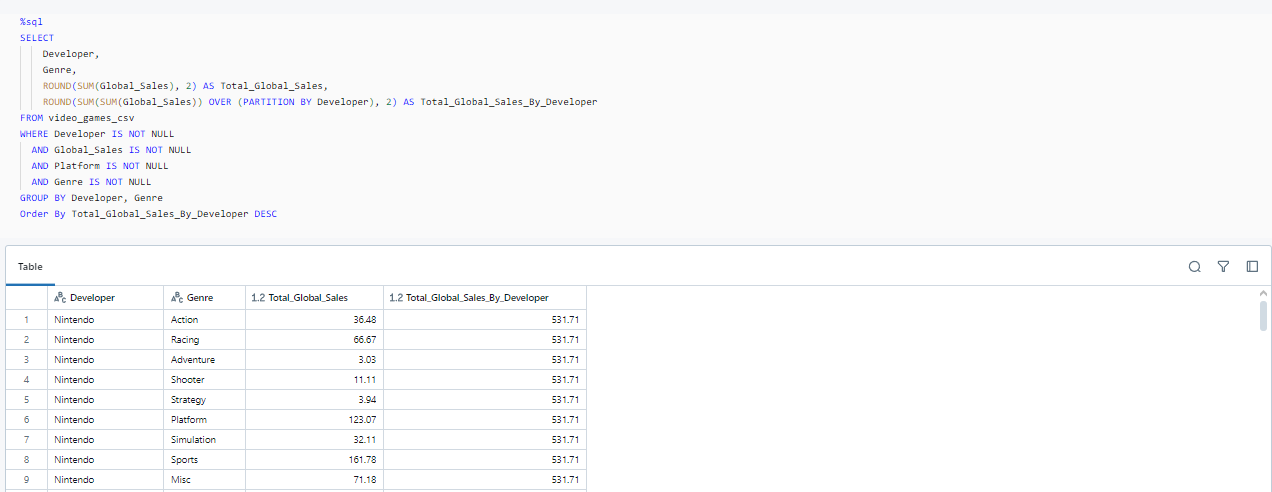
*This query was made to help with the third question, using only the year and sales data.*



*This query was made to further analyze the questions we had about Nintendo.*



*This query aims to identify the best-selling platforms and those with the most potential.*

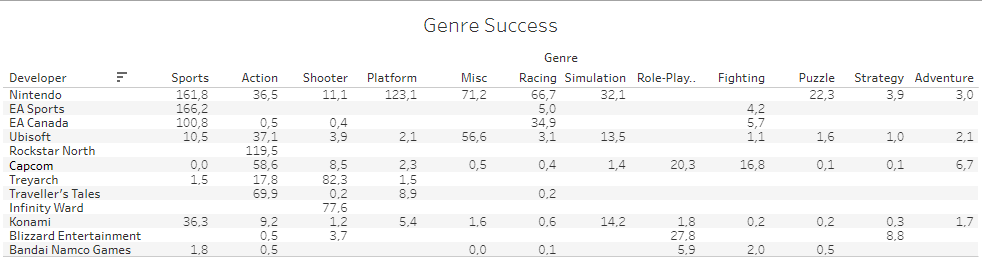


*The goal of this query is to address the question regarding video game genres.*

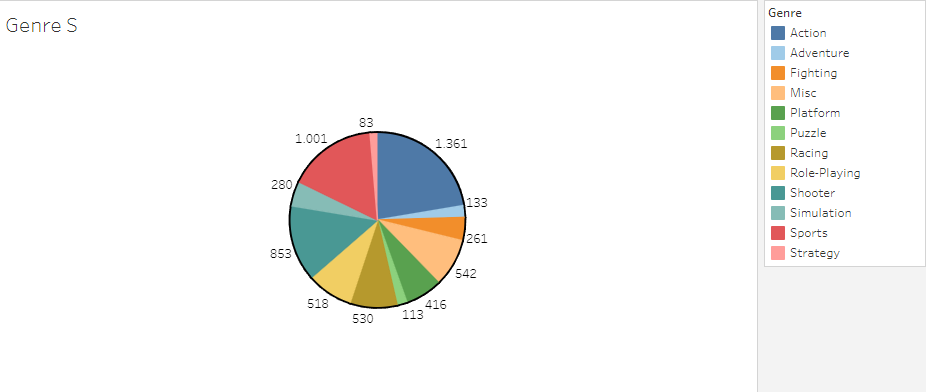
Fulfilling Objectives

Here we will show the analyses done via Tableau:

**Is it better for a company to focus solely on one genre and type of game, or is diversification the key to success?**



In this dashboard, we analyzed the top 10 companies in sales and observed that the truly giant companies today are experimenting and stepping out of their "comfort" zone by publishing games in different genres. There are a few exceptions, such as Rockstar and Infinity Ward, which remain specialized in a single genre.



We also observed that the best-selling games are in the Action, Sports, and Shooter genres. All of the top 10 companies, except those specialized in a single genre, have at least one game in two of these genres (EA Sports and EA Canada are considered the same company).

**Are the “best” games the most sold?**



Here we can see the comparison between the best-ranked games in User Score, Critic Score, and Total Sales. It’s interesting to note that while many franchises repeat, the lists are mostly different.

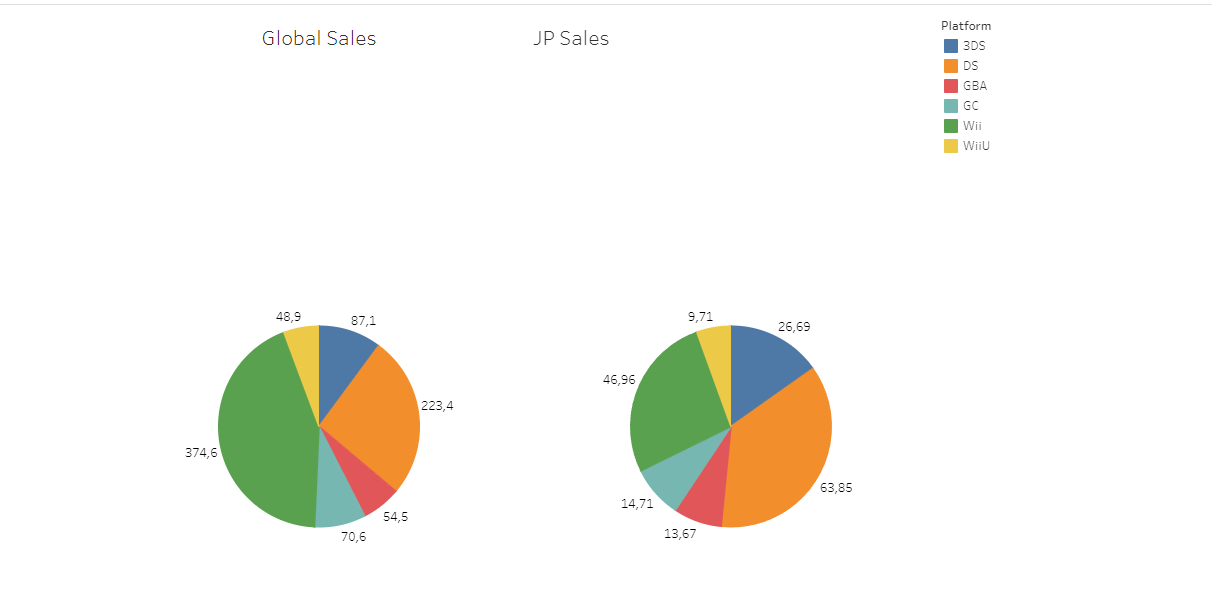
We can assume that critics may have a different opinion when they are not scoring professionally. Additionally, we can infer that sales are driven by marketing or games that are more media-centric.

At the very least, we can assume that when critics give a score, they have a lot of criteria and considerations to account for, and they review a wide variety of games and genres.

We have a lot of assumptions and analyses to make regarding the User Score versus Sales rates. For example, one might say that the games with the best scores are played by fewer people. This makes sense when considering that critics have to review "every" game.

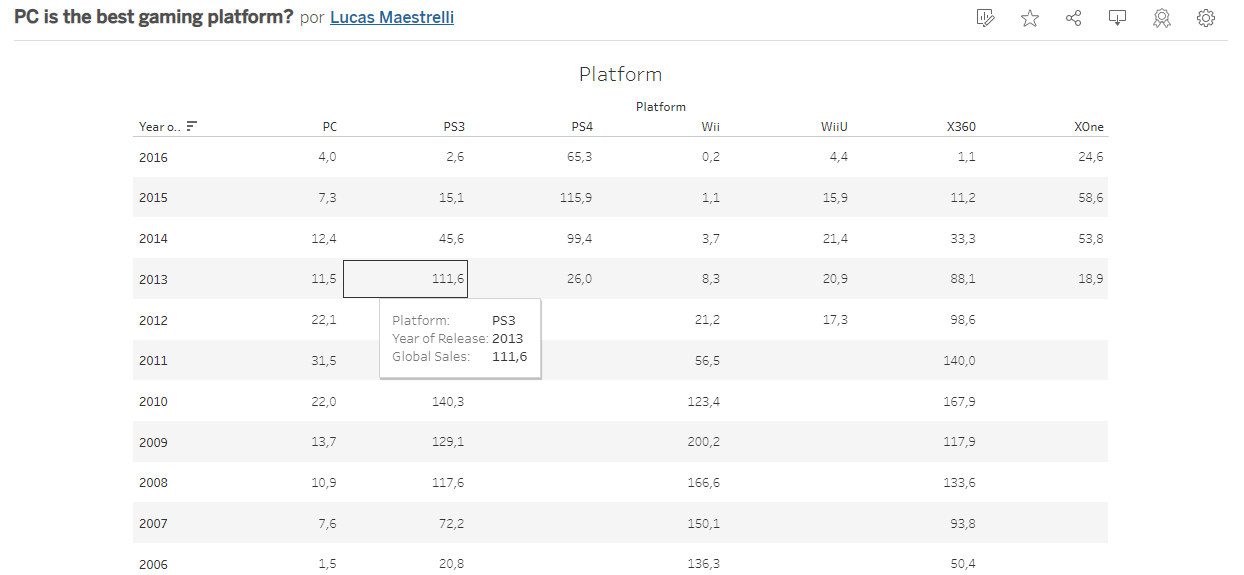
When looking at the lists as a whole, the global sales and critic scores have more similarities. This could be because people consider the critics' reviews before buying. However, they ultimately have a different view of the game and likely give a different score.

**Why did Nintendo insist on portable consoles, and is there an indication that the future hybrid console would work?**



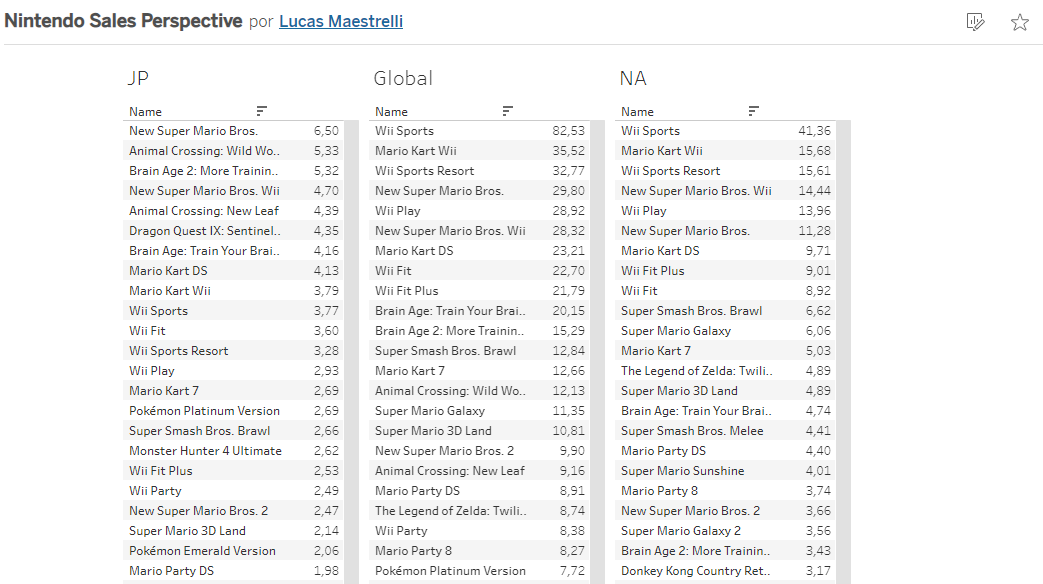
The data here shows in detail how Nintendo’s strategy is very balanced. They have developed strong positions and specific niches in both home consoles and portable gaming. Over the years, they have launched numerous video game platforms, and the data clearly demonstrates their success in both areas. It is no surprise that, a year after this data, that marks its release the Nintendo Switch—the first hybrid console ever—has been breaking records into 2024, even with its comparatively weaker hardware performance compared to other competitors in the same generation.

**Is the PC really the most profitable gaming platform?**

****

This analysis is complex, and the data we have may not be enough to make a definitive conclusion, but let’s investigate it. At the beginning of the 21st century, consoles were very strong but were either moving towards their successors or facing a decline. Despite the fact that the PS4 and Xbox One were leading over PCs in 2015/16, the drastic fall in sales in 2016 might indicate a forthcoming change. This decline could be due to 2016 being a weaker year for the video game industry or it might reflect the beginning of mobile gaming’s rise in popularity with games like Clash Royale and Pokémon GO.

**Is there a correlation between Nintendo's success with its games and its disregard for Western public opinion?**

****

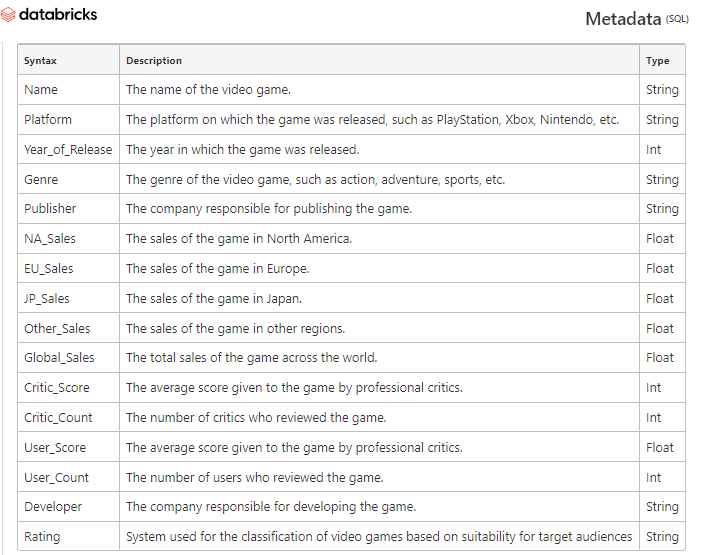
*This question may seem specific and a bit unusual at first, but let’s delve into it.*

For those who have observed the video game industry for a while, there are two key things to note about Nintendo. Firstly, Nintendo is a major player in this industry. They were among the first to venture into video games and are undoubtedly one of the most significant companies in the field. One might question how a small nation like Japan, with historical challenges from WWII, could stand toe-to-toe with larger and more developed economies. This topic could be explored in greater depth but let’s leave it for another time, also it is clear that Nintendo’s success is due to their thoughtful investments, perseverance, and creativity.

Secondly, Nintendo has always created games primarily for the Japanese market. Their games are designed to cater to Japanese tastes, and it took a long time for these games to be made available in other regions. They have also held special events and produced exclusive content for the Japanese audience, even making some games easier for international players as a form of playful mockery.

The data reveals that despite their focus on the Japanese market, a significant portion of Nintendo’s revenue comes from international sales. We can see that NA Sales are very closely aligned with Global Sales, while Japanese consumers tend to favor different types of games.

Metadata



[Metadata in Databricks notebook](https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1411143408865230/2586525567832147/4579165488638907/latest.html)

Conclusion

It appears that we were able to answer nearly all of the questions posed. While there are some aspects that remain too abstract to address definitively, I can say that my initial perceptions of the topic were both confirmed and challenged by the data.

One clear takeaway is that without data, it would have been impossible to explore these questions, make informed predictions, and present findings as clearly as we did. This experience highlights the significance of following a structured analytical process and underscores how essential it is to have high-quality data and effective visualizations for successful analysis.

Self-Critique

### Evaluation Criteria

**Objective (1.0 pt):** The objective of the work must be very well detailed; it should be a comprehensive planning of the project, clearly and objectively outlining the problem to be solved and the business questions to be answered. The quality of this description will be evaluated.

**Data Collection (0.5 pt):** The documentation of the data collection process and the persistence of the datasets on the cloud platform will be assessed.

**Modeling (2.0 pt):** The quality of the data modeling (1.0 pt) and the documentation of the Data Catalog (1.0 pt) will be evaluated.

**Data Loading (1.0 pt):** The quality of the data loading documentation will be assessed, along with the correctness and persistence of the data on the cloud platform after the loading process.

**Analysis (3.0 pt):** The evaluation will focus on the quality of the data analysis (1.0 pt) and the correctness of the solution to the problem, including a well-analyzed discussion of the obtained answers (2.0 pt).

**Self-Evaluation (0.5 pt):** The student’s self-evaluation regarding the achievement of the objectives outlined at the beginning of the project will be assessed.

**Presentation (2.0 pt):** This will be evaluated based on the care and overall quality of the work, assessed in a subjective manner.

### My evaluation

**Objective (1.0 pt):** I am satisfied that the goals of this project were fully met, from the data collection phase through to the final analysis and the visualizations produced.

**Data Collection (0.4 pt):**  I believe that, despite using just one dataset, I successfully met the objectives related to the collection and storage of the data utilized for this project.

**Modeling (1.5 pt):** I believe even though the modeling was really simple and the documentation followed it It was enough to show what I had learned from the classes and to test and improve my abilities with databricks and SQL.

**Data Loading (1.0 pt):** The raw data was already quite consistent and accurate. Thanks to this, with just a few simple assessments, adjustments, and corrections, the data’s quality and integrity were maintained at a high level.

**Analysis (3.0 pt):** The analysis was the key strength of this project. Through this data, I developed solutions, corrected misconceptions, validated previous knowledge, and was able to theorize, make predictions, and compare findings with current data.

**Presentation (1.2 pt):** The presentation was simple and routine, yet it featured clear and well-executed visualizations.

### Conclusion:

I had a difficult time finding suitable datasets and searched most of the suggested sites, but it was quite challenging. Once I found the dataset, the next major challenge was familiarizing myself with Databricks. While using the notebook itself wasn’t hard, setting up the cluster and uploading the table were initial obstacles. After learning that the cluster would be terminated after 60 minutes of inactivity, I planned my work more efficiently to avoid long pauses and potential loss of progress. I experimented with three different BI tools—Looker Studio, Power BI, and Tableau—but ultimately chose Tableau for its speed, intuitiveness, and the ability to create and publish dashboards on Tableau Public. With all of that in mind I decided to keep it simple and focus on the analysis. For future projects, my main change will be to plan more thoroughly before starting and to be more cautious when working with a new tool for the first time.

***Lucas Maestrelli***