

Process Book

Matthieu Masouyé, Robin Plumey

June 2022

1 Introduction

Video games are still widely considered to be a new form of entertainment, despite existing for more than 50 years, and only recently have they been popularized to a larger extent.

For instance, in 2020, the video game market exceeded 300 billion dollars in revenue, up from 8 billions in 2000, and 2.8 billions in 1980.

It grew from its humble beginning in a way few markets did, and we want to convey that with our visualizations. We are also interested to see how well received games were by players, and how that shaped this market.

2 Problem Statement

In this project, we want to explore the different relationships between a video game's sales success and its reception by players and critics, and how that evolved over time.

Are the user ratings correlated to the sales success of a video game ? How did the market grow over time ? Which companies came and went as the years passed ?

We want to show how it evolved, what was successful throughout the years in terms of sales and critical reception, and who were and still are the big players in the video game market.

3 Dataset

Our work focuses on two datasets [Video Game Sales — Kaggle](#) and [Metacritic Ratings — Kaggle](#).

The first one, [Video Game Sales — Kaggle](#), comes from a scrape of the [vgchartz](#) website. It contains 16 598 records with the name, platform, publisher, year, genre and the related sales in Europe, North America, Japan and others.

The second, [Metacritic Ratings — Kaggle](#), includes all games from [Metacritic](#) at the time of publication. It contains records for 8831 games, scrapped from the website, with details about the score assigned by critics and users, platform, genre, developer, and link to the review of the game.

4 Data Pre-Processing

In the end, not a lot of pre-processing was needed on the dataset themselves; The first one on video game sales was already very clean, and while the second one on scores had a couple of problematic columns, it did not pose problem as we simply didn't use them (for instance "Genre" or "Type").

Merging them was a bit more complex: we decided to augment the first dataset with the attributes of the second. To do so, we compared each game name of the second, smaller dataset with the names in the first one using Jaro distance; we would then assign the most likely name to the second dataset's game name, and then merge it with an outer merge on the first dataset. The end result is the first dataset with the added columns of the second one, which are NaN if the game was not in the second dataset.

We also generated a json file from the merge dataset and image from the web for our second visualization. To get the images, we used [wikipedia](#) for the platform and publisher and [logos.fandom.com](#) for the games logo and we found them by code and searching in the html the right image.

5 Architecture

We kept the code used to clean and merge the dataset in main, as well as the other python and notebook files used to generate the visualization for the "Home" page, and gathering the images for the "Circle" page.

The website hosting our visualization runs on github pages, and so the code for this is on the branch Page of our repository. Here we can find all the necessary files for the website to run.

6 Home Page

6.1 Graphs

The graphs for the home page were done using the Plotly Python library. We chose this library for several reasons:

- We knew we wanted a slider to navigate through time, and Plotly allows use to animate a graph in such a way natively
- Plotly allows to export an interactive graph directly to an html file, making it easy to integrate it into our webpage, instead of for instance having to setup a way for our page and python script to communicate while the page is loading
- This last feature also allows us to precompute the needed graphs, thus making the page load faster as it has no computation to do on the dataset

The graphs are done mainly through correctly setting up the Pandas dataframe before giving it as an argument to a Plotly Bar graph: Thus the first obstacle

was implementing a sliding window of five years in the dataframe to smooth out irregularities and better represent a video game's lifespan (as otherwise a game would only appear on the year of it's release). We solved this problem by creating a new dataframe, and appending each groups of five years with a different id, such that we could easily group by that id in the Plotly graph. While not very efficient on resources, as it multiplies the size of the original dataframe by the size of the window, we can get away with this solution as we have a relatively small dataframe to begin with (a few thousands of entries). Then, it was a matter of sorting this dataframe, grouping by the element of interest (Games, Publisher or Platform), and taking the top 10 for each sliding window. It's also worth noting that these graphs were made on the subset of games present in both dataset, as both the sales and scores numbers were needed for each entry in the graph.

Plotly allows us to configure the graph with a lot of tools, for instance we configured the color of the bar to represent an additional information, the Meta or User score: we were able to define custom colors [picked to be colorblind friendly](#). Finally, we configured a small popup when hovering on a bar to show more detailed numerical informations, for instance when grouping by publisher it show in this popup how many games the publisher released in the selected timeframe.

Once we had graphs we were satisfied with, we could export and integrate them to the main page.

6.2 Page Presentation

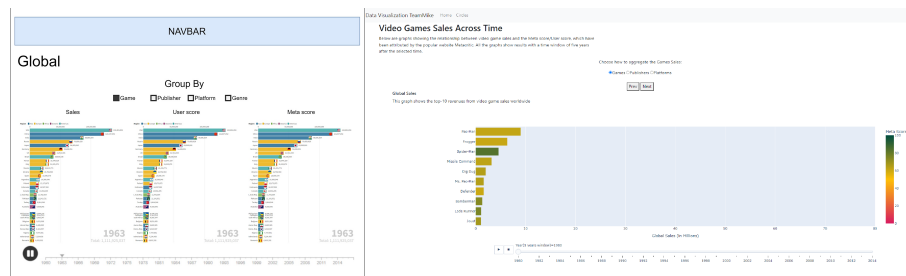


Figure 1: Original sketch vs final page

We stayed close to the original sketch in the end, as we can see in Figure 1. We kept the radio buttons to switch between grouping by Game name, Publisher and Platform. The Plotly graphs also closely resemble the original idea, with a slider on the bottom to choose the timeframe.

However, we wanted to show the graphs in a more compact way; on the original sketch, they are side by side, which wasn't possible with the Plotly graphs as they would be squished way too much for lisibility. Moreover, while having multiple graphs controlled by the same slider is technically feasible, it is

very involved deep in the inner workings of Plotly animation, so we opted for something simpler.

What we implemented is instead two buttons which animates the three graphs so that they slide in and out of the screen, one after the other, using javascript, CSS and JQuery. This way, it provides an easy and smooth navigation between the graphs, and it's easier to understand what each graph shows, as the short description slides with them. If they were side-by-side, or one after the other, it would be more annoying for the user to navigate the webpage and link each description to it's graph.

Finally, we put a link to the original sources of the datasets at the bottom of the page.

7 Circles Page

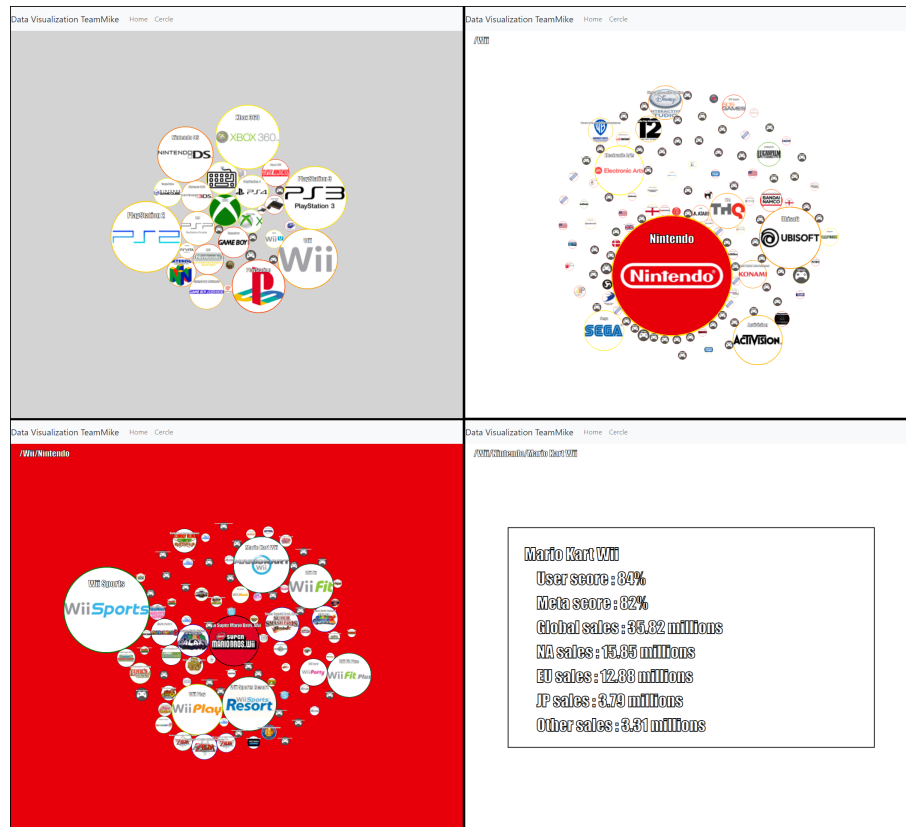


Figure 2: Circles page

This visualization was done with D3. We chose this technology to have a

dynamic animation and also because we wanted to be able to navigate in the different bubbles and see all the different games.

The bubbles are created from a json file containing all the useful information such as the number of sales, the image and score.

The number of sales is represented by the size of the bubbles and the score is used to draw a colored circle according to it. To navigate in this visualization, you just have to click on the bubbles you want to visit and then it will display the internal elements.

In Figure 2 you can see the different steps by clicking on the bubble.

8 Conclusion

Overall, we think we managed to produce an interesting visualization both for the user and for our purpose. While it would benefit from further polish to refine the style and visual appearance of the website, all the visualizations work as intended, are smooth, and provide the user with the experience we envisioned at the start. Users will be able to explore these visualizations which will provide visual and interactive information about the actors in the video game industry and their ratings.

9 Peer Assessment

As we were only two people, work distribution was facilitated: For the first two milestones, we both worked on it together, writing the reports and doing sketches for the visualizations. We cleaned and did the initial graphs on one dataset each, and overall work was distributed very evenly.

For the final Milestone, we shared the workload as such:

- Robin worked on implementing the "Circles" page and visualizations, as well as the script to scrape the images corresponding to each bubble
- Matthieu merged both datasets for the two visualizations, and did the "Home" page and its visualizations

We both contributed to this progress book using Overleaf, and finally the video cast was done by Matthieu and edited by Robin.