Project Proposal: Super Video Game Analysis

Feiran Ji, Chenxi Ge

Basic Info

Project title: Super Video Game Analysis
Team members: Feiran Ji, Chenxi Ge
Email: fji3@usfca.edu, cge4@dons.usfca.edu
Github repository: https://github.com/feiran-kyna-ji/video-game

Background and Motivation

Both member of our group like games. We enjoy novel games with astonishing effects, but also appreciate classical ones that never fade in time. Feiran's practicum is at Ubisoft, a lead game publisher. This also motivated our team, because after doing the practicum in a video game company's marketing department, we see a real potential in the market.
Thus, we decided to do some analysis on the video game industry. We want to know what are the important factors that decide how many copies of a game will be sold, and we also want to see if there are any temporal trend in game genres, platforms, users as well as producers and publishers who made the game. It will be fun for gamers to know more about the industry, and what's more, it might be useful for industry players to see some interesting trends that they have not considered about.

Project Objectives

The objectives of this project include:

- Allow our users to have an overview of how gaming industry evolved over time from 1980 to 2015.
- Allow our users to have an idea of how different company (Sony, Nintendo, and Microsoft)'s consoles/platforms changed, and which one is the winner.
- Provide our users insights about different genres, for example, which ones are popular in a certain region.
- Allow our users to learn about 100 best-selling games in the history.

Data

The dataset has information of more than 16,700 video games, including the metadata of the game, the sales data and scores data.

The metadata of the game and sales data is scraped from vgchartz by Gregory Smith. The scraping script can be found at https://github.com/GregorUT/vgchartzScrape.
The metadata of the game includes the platform, release year, genre, developer, and publisher. The sales data includes the breakdown of leading gaming markets: North America (NA), Europe (EU), Japan (JP) as well as other countries (Other).

The scores data is scraped from Metacritic by Rush Kirubi, and the script can be found at https://github.com/wtamu-cisresearch/scraper.

The scores data of a game on Metacritic consists of two parts, the Critic part, and the User part. Each part has its own scores and the count of people who rated. However, as Metacritic only covers a subset of the platforms, roughly half of the games don't have scores from Metacritic.

Data Processing

We do not expect to do substantial data cleanup, as the data is already scraped and structured using the corresponding scripts. We expect to use all information from the dataset.

From our exploratory analysis, we find that several features have missing values. For features that have few missing values (Name, Year_of_Release, Genre, Publisher), we decide to drop the records with missing values, as they are only 2% of the total records.

For features that have lots of missing values, we treat the missing value as a separate group: for example, we treat all games without a developer record as from developer 'Null', we treat games without a rating as rated 'Null'. For numerical features like score or count, we either put games with missing values in it's own bin, or discard these records for analysis (e.g., if we want to build a machine learning model).

Table 1. Features with at least 1 missing values

Feature	**Count of missing values**
Name	2
Year_of_Release	269
Genre	2
Publisher	54
Developer	6623
Rating	6769
Critic_Score	8582
Critic_Count	8582
User_Score	9129
User_Count	9129

Visualization Design

Figure 1 shows information about the global game market. It has two components, one main graph, and one-time series plot, and these two graphs are linked.

In the main graph, 4 bar plot lies on the 4 major game sale regions: North America (NA), Europe (EU), Japan (JP) and other markets which lies on the Indian ocean. Each plot shows the sales breakdown by game genre in each region, thus give us a good understanding of the user preference in each region.
The time series plot lies under the main graph. It shows the game sales of each region across different years. The user can drag on the time series plot to specify the year to be shown in the main graph.
We choose the bar plot to show the breakdown of genre within each region, as the genres are categorical, and are want to compare the total sales for each.

Figure 2 shows the game sales over different years by each genre. We choose the line chart as the x-axis year is a continuous feature, and we use different colors to denote the different genres.

Figure 3 shows the game sales over different years by rating. We choose to use a stacked bar plot as we are more interested to learn the percentage change.

Figure 4 shows the score, the release year and the genre of top 100 games (by sales). We use the size of each observation to show the total sales of each game, use the color to distinguish the genre. We believe this bubble chart is a great way to show data with multiple dimensions.

Figure 5 shows a heatmap of games on different platforms over time. We think the heatmap is an intuitive way of showing values that change in time.

Figure 6 is a small multiple of bar plots. We show the comparison of genre preference, rating result, market cap and sales on 4 major platforms. A small multiple is the best way for us to compare multiple values of the same group.

Must-Have Features
We are going to use Plotly and Dash to create the website, including click, hover, zooming, and panning, focus+context. We will enable these features in each of our figures. These basic features will support all our objectives because users will have the option to study more about graph if they want.
We will also include filtering as one of our features because we believe it will be useful in figure 1 and 2. There are too many genres in the dataset, and thus we want to let users have the option to choose which genres they want to see. This will be beneficial in providing our users' insights about different genres(objective 3).

Optional Features
Optional features including:

- Linked views: We may add a linked view on figure 1. Figure one is a world map and time series of sales in different regions. Thus, we want to add a feature so that when users choose a certain year on the time series plot, the map will show

the distribution of sales on that year. This will let users see how gaming industry evolved from 1980 to 2015 (objective 1) more clearly.

- Brushing: Brushing might be useful for figure 6, because we have 4 bar plots on major platforms. We may want the users to choose one platform and highlights that platform in other visualizations. This will allow users to see comparison of different consoles more clearly(objective 2).

Project Schedule

Week Of	Deadlines	Responsibilities
4/16	4/19 Website 4/21 Alpha Release	Feiran: set up web application on Flask (website). Chenxi: complete jupyter notebook for data acquisition (alpha release).
4/23	4/26 Beta Release	Feiran: put the plots into web application. Chenxi: plot 2 visualizations in plotly.
4/30	NA	Feiran: plot 4 visualizations and put into web application. Chenxi: plot 4 visualizations and put into web application.
5/7	5/8 or 5/10 Presentation	Integrate all visualizations and prepare for the presentation. Debug the web application and add optional features.
5/14	5/16 Final Submission	Submit.

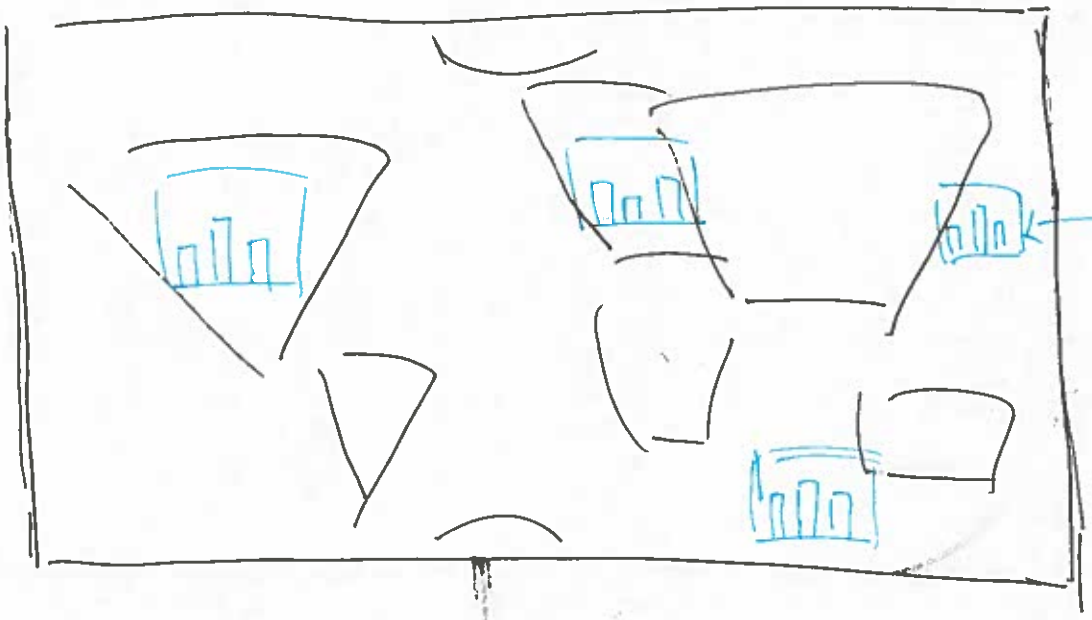
Super Video Game Andysis

< Here is some introduction of project ——————

——————

——————>

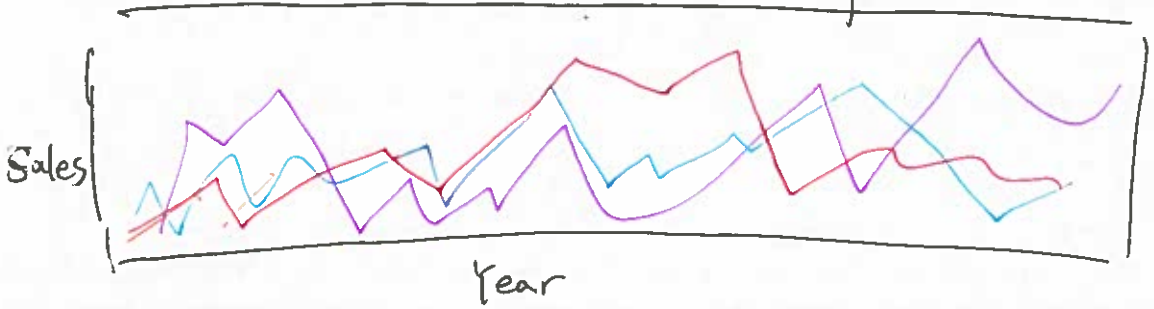
Global Game Market



some plots indicate game sales of different genres in 4 areas NA, JP, EU, others

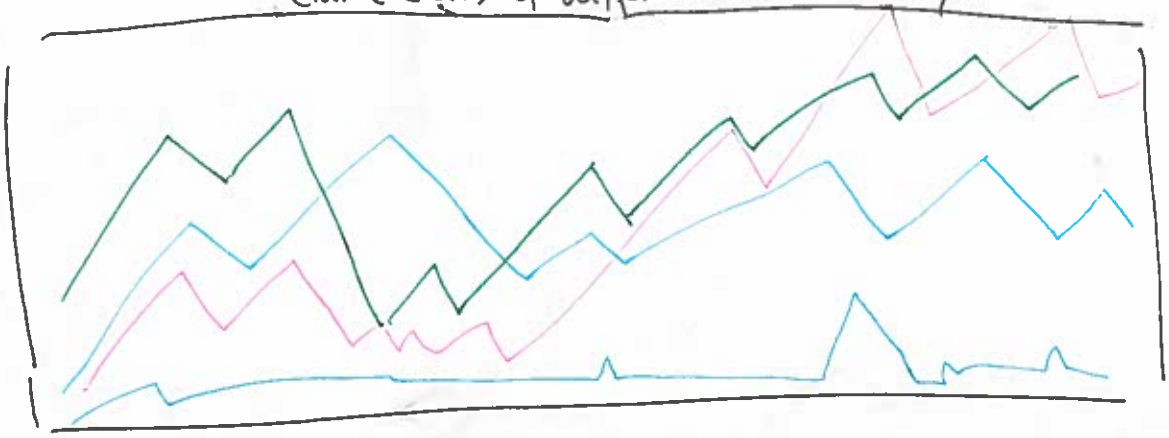
< Narratives >

Game Sales in different regions



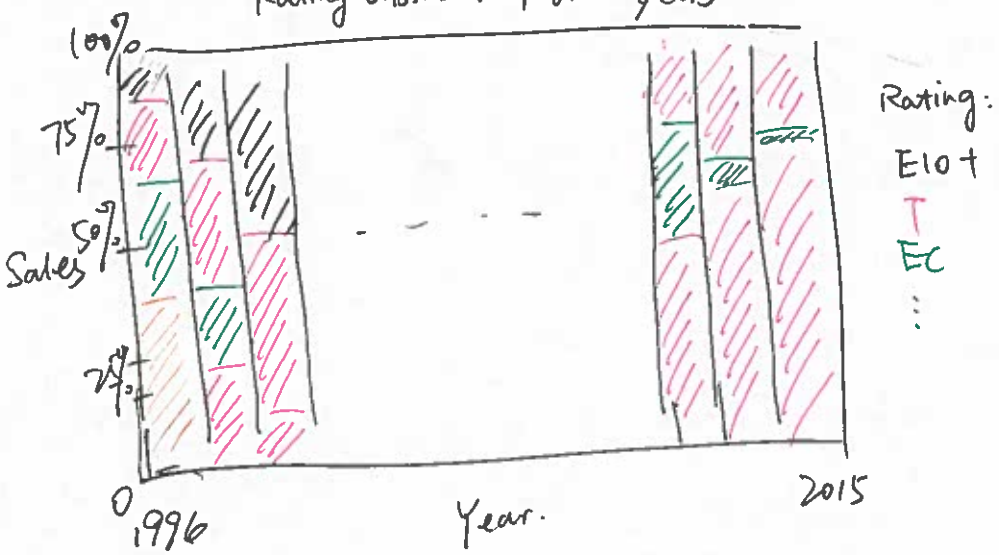
< narrative ——————>

Game Sales of different Genre over year



< narrative ——————>

Rating distribution over years



< Narratives ——

——————

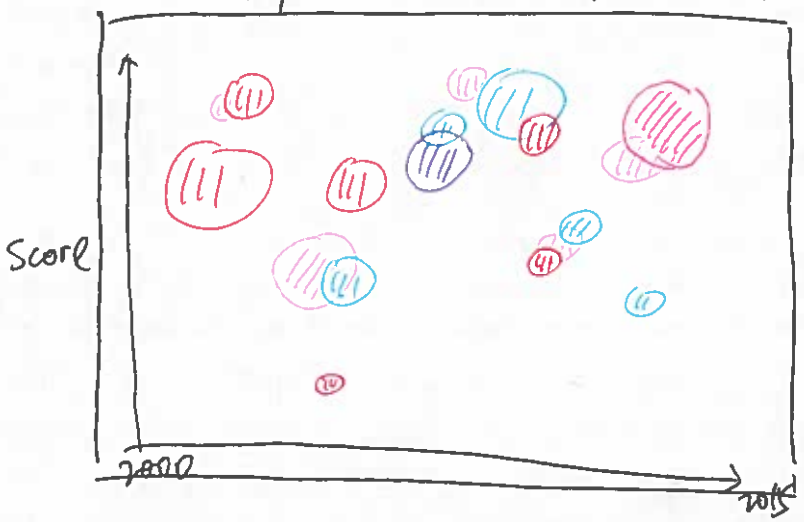
——————

——————

——————

——————>

Top 100 Games (By total Sale) Score trend



< Narratives

——————

——————

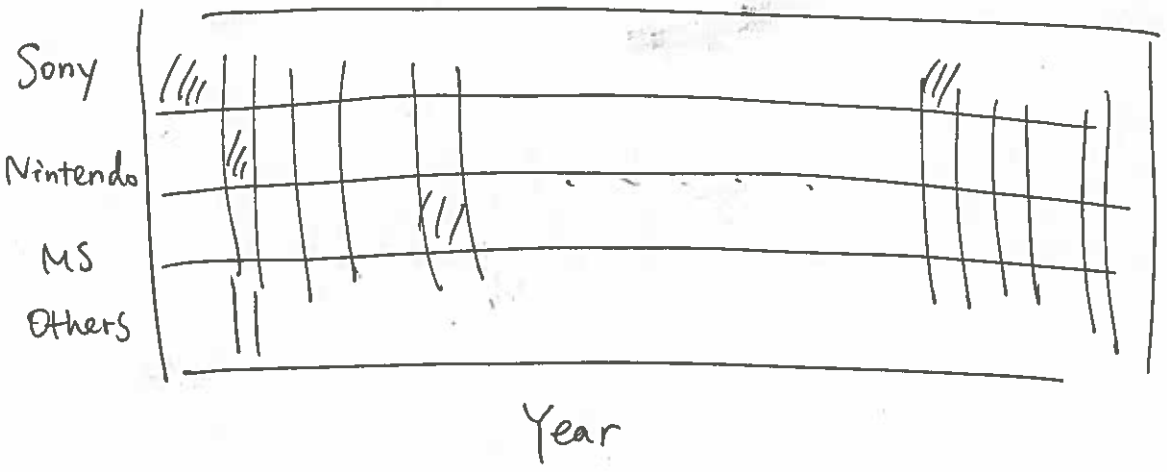
——————

——————

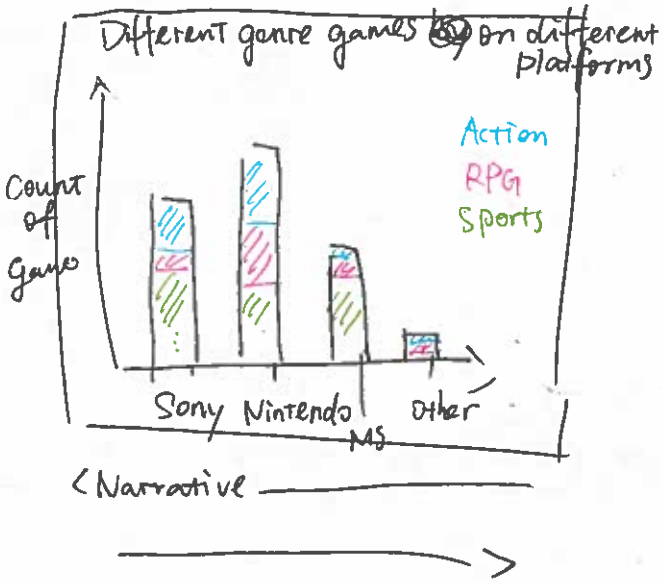
——————

——————>

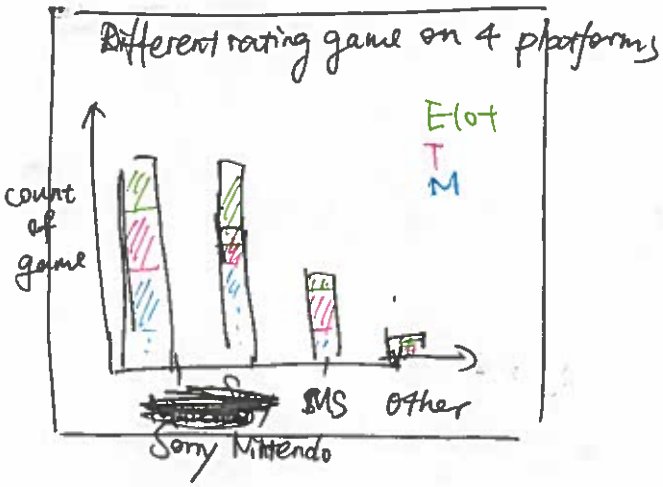
Heatmap of platform time series.



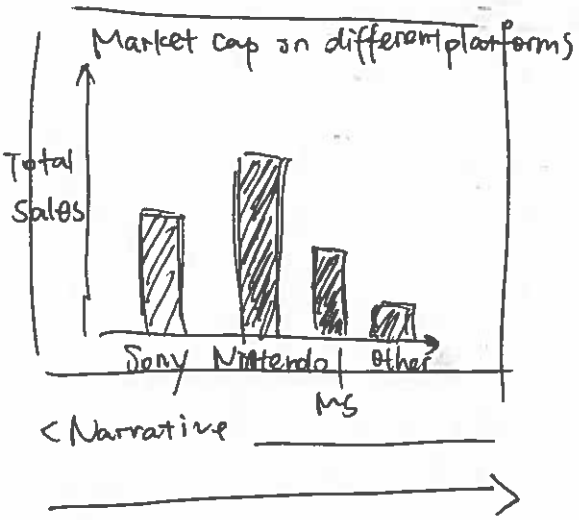
< Narrative ——————————————— >



<Narrative ——————————— >



<Narrative ——————————— >



<Narrative ——————————— >

<Narrative ——————————— >

< Conclusion ——————————— >