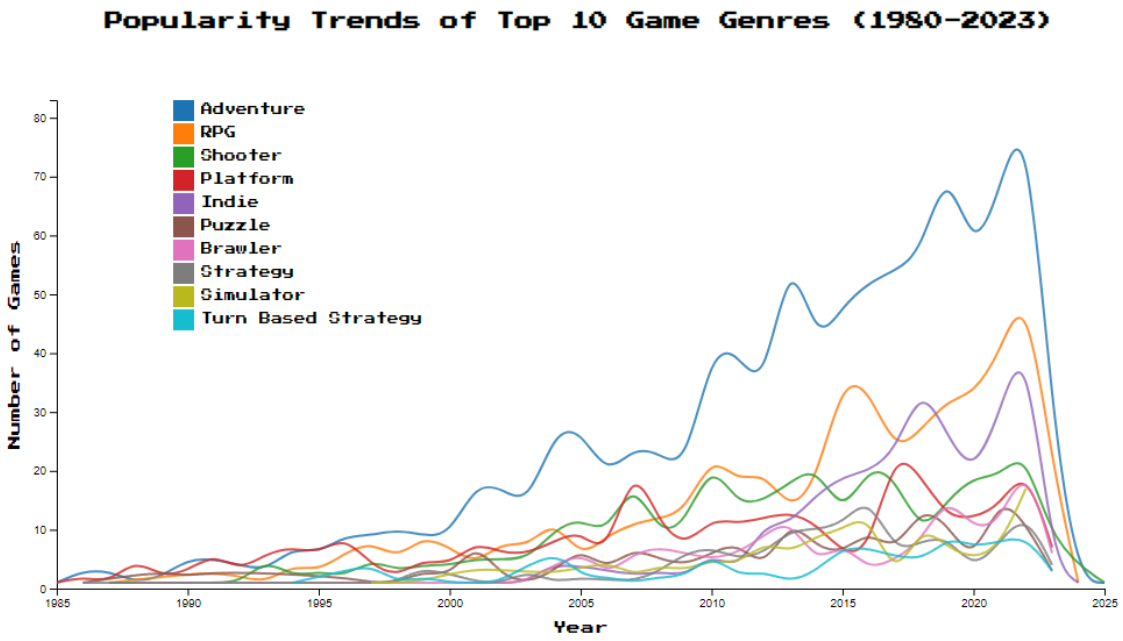
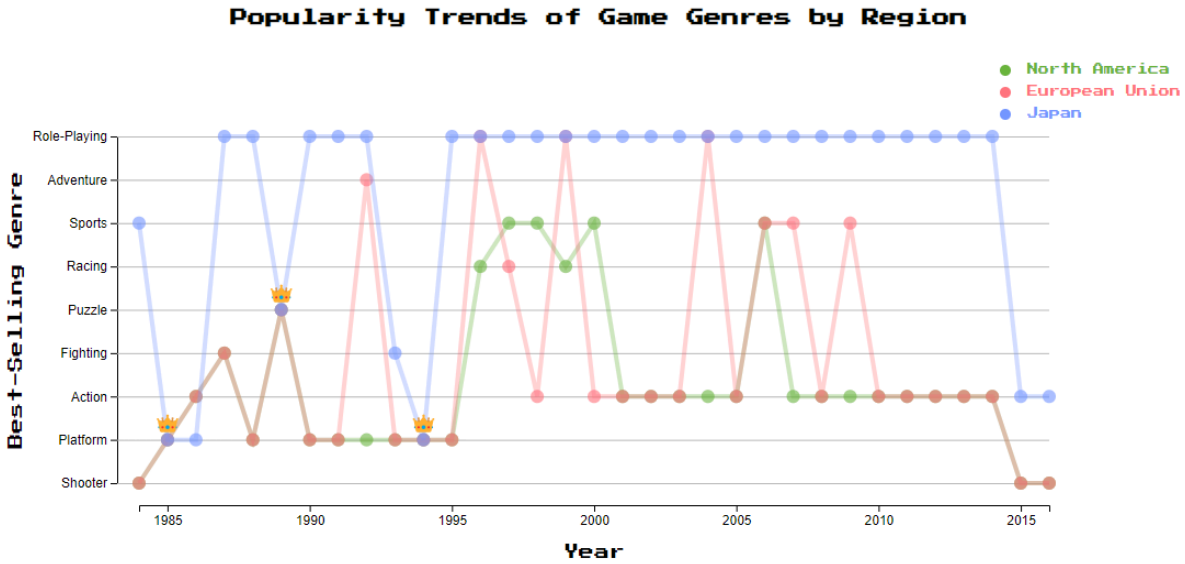
INFO 5100 Project 1 Report

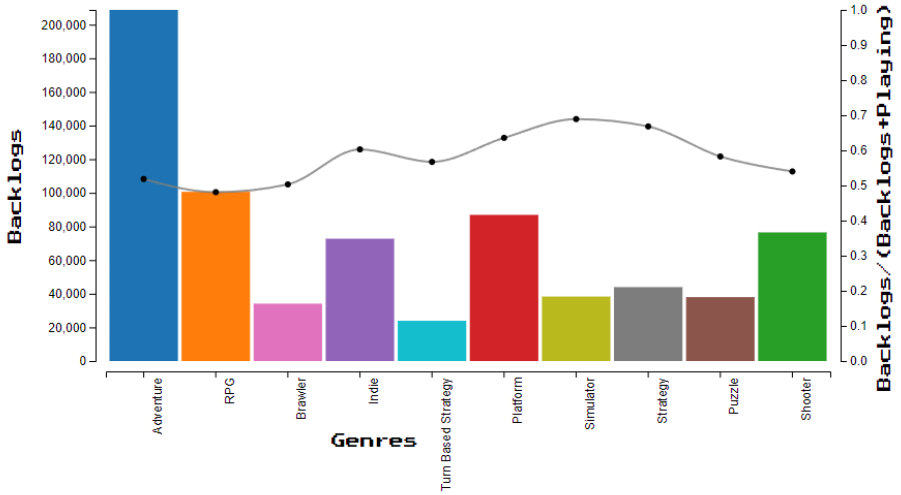
Group Members: Sasha Rabeno, Jade Wang, Menghan Xu, Zhiqi Chen

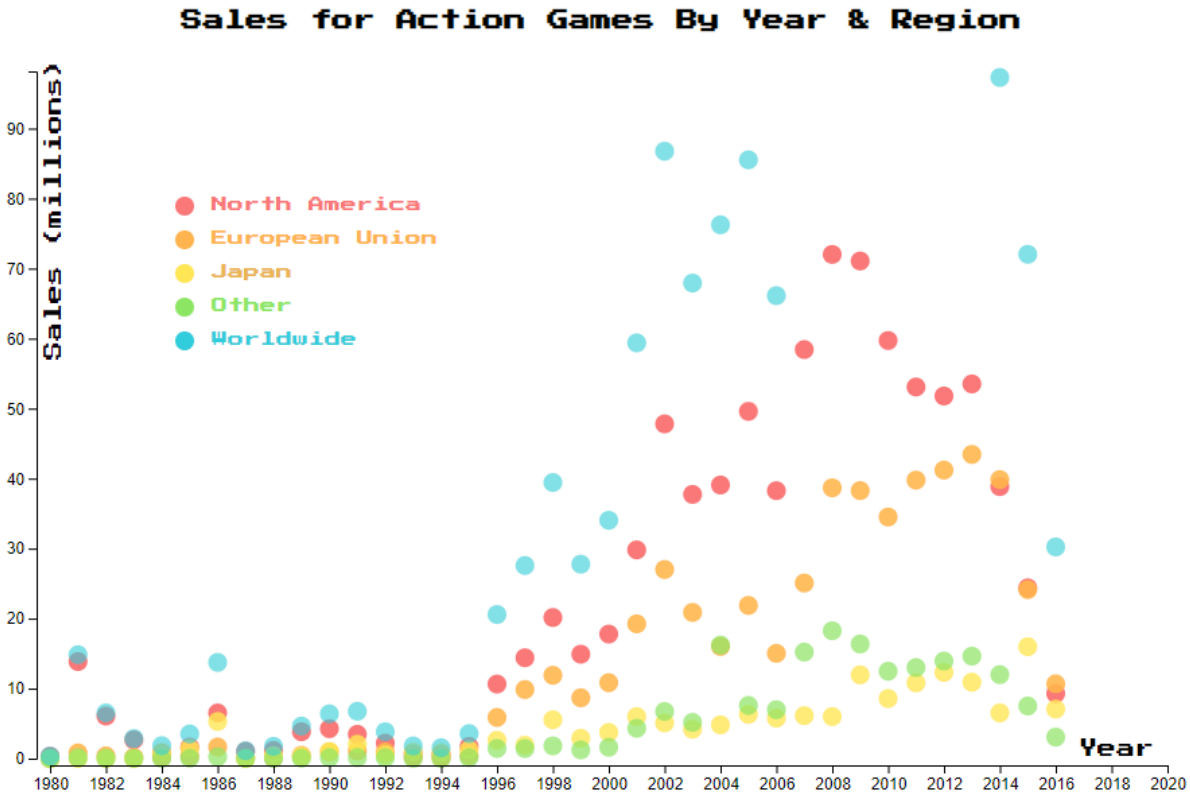
Screenshots of Visualizations





What kinds of games get backlogged (purchased but not played) the most?





Data Description

Where did we get the datasets?

The datasets we used in the project were from Popular Video Games 1980 - 2023 🎮 and Video Game Sales by Genre. Both were from Kaggle.

In our repository, **games_detail.csv** (dataset 1) is Popular Video Games 1980 - 2023 🎮, and **games_sales_year_genre_region.csv** (dataset 2) is the third CSV file in Video Game Sales by Genre.

Variables

Dataset 1: **Menghan, Zhiqi**
The first dataset is "Popular Video Games 1980 - 2023 🎮", which provides detailed information about video games released over more than four decades. The main variables included in this dataset are:
Title: The name of the video game. This variable is used to identify individual games in the dataset.
Release Date: The date when the game was first launched. This variable allows us to analyze trends over time, including the frequency of game releases across different years.
Team: The development team or publisher behind the game. This field may contain multiple entries if several companies or studios were involved. It helps in understanding the collaboration patterns and the influence of different game developers in the industry.
Rating: The user rating for the game, typically on a scale (e.g., 0-5). This variable provides insights into the reception and perceived quality of the game.
Times Listed: The number of times the game has been listed or mentioned across various platforms or lists. This variable serves as an indicator of the game's visibility and popularity.
Number of Reviews: The total count of user reviews for the game. reflects the level of engagement and interest among players.
Genres: The genres associated with the game, such as "RPG," "Adventure," or "Shooter." A game can belong to multiple genres, which allows for a more comprehensive analysis of genre trends over time.
Plays/Playing: The number of players who have played or currently playing the game. This metric serves as an indicator of the game's reach and popularity.
Backlogs: The number of players who have added the game to their "backlog," indicating an intention to play it in the future.
Wishlist: The number of players who have added the game to their wishlist. This variable can be seen as a measure of potential future interest in the game.

Dataset 2: **Jade, Sasha**
The second dataset has the following variables:
Release_year: the year when games were released
Genre: The genres associated with the game. In this dataset, the genre naming is a bit different from the first dataset. For example, “RPG” in the first dataset is named “Role-Playing”.

Region: the region where sales occurred, including “North America”, “European Union”, “Japan”, “Other”, “Worldwide”
Sales: total game sales in millions. This is used as an indicator of genre popularity.

How did you clean the dataset?

Zhiqi:
Step1-Data Preprocessing:
Removed records with missing or incomplete information for critical fields (Title, Release date, Genres) to ensure data quality. Also converted the Release date to a standard date format and extracted the Release year. Records with invalid dates were excluded.

Step2-Splitting the Genres Column:
I realized that the Genres field often contained multiple genres for a single game. These were split into separate rows so that each genre would be treated as a unique entry.

Step3- Aggregating data to create **game_genre_trends.csv**:
The new dataset was grouped by Release Year and Genres, counting the number of games released for each genre annually. The aggregated results were saved in a new file named game_genre_trends.csv with columns: Release Year, Genre, and Count.

step4 -Extracting the Top 10 Game Genres:
Filtering for top genres to make the graph more clean, calculated the total number of releases for each genre across all years and selected the top 10 genres based on their total counts. And retained only the data for these top 10 genres in the cleaned dataset for focused analysis.

Jade:
(All data filtering is done within the script tag, no additional CSV file is created)
Step 1 – Data Filtering: The dataset was filtered to include only records with release years between 1984 and 2016. This is because Japan had zero sales for all genres until 1983, and after 2016 a lot of genres are missing data points.

Step 2 – Identifying Popular Genres: For each year and region, the dataset was scanned to find the most popular game genre based on sales.

Menghan:
Step 1 - Splitting the genres column
Similar to what Zhiqi did, I traverse all the rows and split the genres with the .replace() function if a row has multiple genres. And then I add up all the Backlogs and Playing with the same genre. If there does not exist Backlogs or Playing, I add 0. Then I get a map with all the genres as keys and Backlogs and Playings as values.

Step 2 - Calculating radio(Backlogs/(Backlogs + Playings)

For each genre, I compute Backlogs/(Backlogs + Playings) and store it. I simply use a foreach to do that..

Sasha: There wasn't much cleaning I had to do; I removed data points from after the year 2016 (since the data got much spottier) by reducing their opacity on the graph to 0.

Criteria used for data selection

We chose the two datasets that we did because we all have a shared interest in video games. Additionally, because these datasets weren't too complicated, there wasn't much cleaning or other work to do before we could start working with the data.

Overview of design rationale

Zhiqi (1st visualization):

Visualization Design:
Line chart: I used a line chart to illustrate trends over time, each line representing a different genre.For timeline vs trend I think line chart is the best. It is easy to track changes and observe long-term patterns in genre popularity.
Focus on the Top 10: To maintain clarity, I decide to limit the visualization to the top 10 genres based on total game releases.This shows the most significant trends while avoiding messy overcrowding.
Timeline: Although the dataset only goes up to 2023, we chose to display X-axis labels every five years (e.g., 2015, 2020, 2025) for a cleaner visual design. Extending the X-axis to 2025 not only keeps the labels consistent but also leaves room for potential future trends, allowing readers to better visualize the trajectory in recent years and consider possible future developments.
Smoother the curve: I found the pointy graph is visually abusive so applied d3.curveBasis for smooth curves,now its visual appeal and making trends easier to follow without being distracted by minor fluctuations.

Interactive Features:
Color and Tooltips: Each genre was assigned a color for easy differentiation, and tooltips were added to display genre names on hover.
Legend Interaction: I found its abit to track the lines so now clicking on a legend entry brought that line into focus.

Jade (2nd visualization):

In the visualization, each mark is represented by a dot, with the position and color serving as channels. I initially considered using a stacked bar chart but ultimately chose a line chart for better clarity and comparison between regions.

Stacked Bar Chart Concept:
In a stacked bar chart, each bar would represent one year, divided into three colored rectangles—each color corresponding to a region. The height of each rectangle would reflect the sales amount of the top-selling genre for that region, with the genre name displayed on the rectangle.

Line Chart (Final Choice):
In the line chart, I used three different colors of dots—each color representing one region. The dots' x position corresponds to the year, while the y position represents the top-selling genre for that year in that region. I connected the dots with lines to show trends for each region across time. To highlight the years when all three regions had the same top-selling genre, I added a crown emoji on top of the overlapping dots for that year.

I chose the line chart because it allows for a direct comparison of trends across regions, which other kinds of charts don't provide. While the y-axis represents a categorical variable (genres) rather than a numerical one, the title and axis labels should help users quickly understand the visualization.

Menghan (3rd visualization):

I drew two kinds of charts and combined them into one figure. One is the bar chart and the other is the line chart. The reason why I do this is because I assume backlogs cannot represent which game gets backlogged the most. In some circumstances, some games are really popular, so many people will buy and try them. Then it is likely that the number of people who don't play them is higher. So I also compute the radio to see what will happen.
Bar Chart: x-axis represents the different genres. The reason why I choose 10 top genres is because Zhiqi does so and we want our charts to be consistent. I also make sure that the color of each bar corresponding to the genre matches Zhiqi's. The y-axis represents the number of backlogs. The higher the rectangle, the more backlogs. It is clear to see which one gets backlogged the most. And I can easily compare different genres' backlogs.

Line Chart: The x-axis still represents the different genres, but the y-axis represents the radio this time.I draw a line so that I can easily compare different genres' radio.

I draw these two charts in one figure. And then I can compare the radio and backlogs for different genres and think about which index better reflects which kind of game is more likely to be "abandoned" by people. I can also compare and analyze the relative positions of genres under different indexes.

Sasha (4th visualization):

For my visualization, I made sure the opacities for the dots drawn later on are lower so that you can see all the dots that are stacked on top of each other. However, this does make things a little harder to see for lower sales values. Scales-wise, keeping things in a linear scale made the most sense (and was easiest mentally to deal with). I think drawing a trendline for each region would make the graph easier to read, but this was something I struggled to figure out how to do d3-wise. I alleviated this somewhat by reducing the height of my chart.

Color-wise, I experimented with a couple different color palettes, and found the pastel rainbow one that I settled on to be the least awkward. I did try to match the colors of my regions with Jade's graph (the second visualization), but I struggled to find additional colors that didn't look bad/overlap weirdly.

Each mark represents the # of sales (in millions), for a given year, in one of the five regions given by the dataset (North America, Europe, Japan, Other, and Worldwide). Color is used to distinguish between regions. The "worldwide" dots (in blue) are a sum of the regions underneath them, representing all sales globally for a specific year; thus, the blue dots are always on top.

The story

Zhiqi (1st visualization):

The trends in the top 10 game genres from 1980 to 2023 tell a story about how gaming has evolved over the decades.

- Evolving Popularity: Adventure and RPG genres have gained steady momentum, especially since the 2000s. I can tell that the players' growing appetite for immersive, story-rich experiences as game technology advanced.
  Also the rise of Indie games in the 2010s shows how smaller studios have found their place, challenging the dominance of big-budget titles and bringing fresh ideas to the scene.
- Comebacks and Fluctuations: Platform games thrived in the '80s and '90s but dipped as gaming tastes shifted towards more complex genres. Interestingly, they made a comeback in the 2010s, likely fueled by nostalgia and remakes of beloved classics. Shooters experienced ups and downs, often spiking with the release of blockbuster titles, showing how certain games can shape a genre's popularity.
- Reflecting Broader Trends: The growth of Simulator and Strategy games points to a broader trend towards realism and thoughtful gameplay. This suggests a maturing audience looking for varied gaming experiences.

What Surprised us? We didn't expect Adventure and RPG to grow so consistently over the decades, this is a deep, ongoing love for narrative-driven gaming. The revival of Platform games was also a pleasant surprise, I can see that classic gameplay still resonates.

Jade (2nd visualization):

My visualization explores how the popularity of game genres (measured by sales) changed from 1984 to 2016 across different regions. I initially expected some overlap between North America and the European Union because the two have relatively similar cultures, but I was surprised to find that for more than half of the time, their trends are nearly identical. This alignment is particularly strong before 1990 and after 2010, suggesting that players from both regions tend to have similar tastes in games.

Japan, on the other hand, had far fewer overlapping dots with the two other regions than I anticipated. Specifically, Japanese players have a strong preference for RPGs, with RPGs being the best-selling genre in Japan for nearly 80% of the time between 1984 and 2016. This strong contrast highlights the significant difference in gaming preferences between Western and East Asian societies.

Menghan (3rd visualization):

I find out that the adventure game is the kind of game that has the most backlogs. However, I look at Zhiqi's chart, and I know that actually it's the most popular game for more than 20 years. So it is popular and also gets abandoned the most. This conclusion sounds contradictory. I think the reason is that because there are many people who play adventure games, then it's likely that more people will stop playing it and let them backlog. So I look at the line chart, and find that the simulator game is the kind of game that has the highest abandon radio. In Zhiqi's chart, I can see that not many people play simulator games. So I think it makes sense. The highest abandon rate means the simulator game is the kind of game that get backlogged the most.

Sasha (4th visualization):

My graph looks at sales trends globally for the best-selling genre: action games. I'm particularly surprised that North America and EU sales eclipse those from Japan so much, especially given how many video games come out of Japan (I play a lot of Nintendo games).

The graph does make it pretty easy to see that, especially as time goes on, North America dominates in action game sales (and that sales go up with time for *every* region, suggesting unsurprisingly that action games are popular everywhere).

Team member contributions

1. **Zhiqi**: Clean the dataset for Popular Video Games 1980 - 2023 🎮(1 hour), Test with different visualizations format(2 hours), Created the first visualization and continue to revise it(6 hours), Browsing and testing the font and other visual element for better effect(2 hours)
2. **Jade**: Tried stacked bar chart for the second visualization(4 hr). Created the second visualization(4 hr). Merged all visualizations into one index.html (2 hr). Edited the style for index.html, including margins, font sizes, helping my teammate fix their legend, etc (1 hr). Found & Applied the font family (0.5 hr). The part that took the most time was actually my first attempt, which was creating the stacked bar chart. There was not much documentation online on how to implement it, and also we didn't learn it in class, so I was struggling a lot during that.
3. **Menghan**: looking at the dataset and thinking of how I can answer the question with charts. Decide which kind of charts I would like to use(1h). Clean the data and create a new json data(0.5h). Draw the bar chart and manipulate the color, margins and text(3h). Draw the line chart, make sure it corresponds to the bar chart and revise my code to make my chart look better and correspond to my teammates' charts.(4h). The most difficult part for me is manipulating the colors, margins, and positions of text. I have to correspond to Zhiqi's chart, which costs me a lot of time to revise my original chart. And then I also spent a lot of time on drawing the line chart. At first, it was a messy line, and I spent some time thinking of what's wrong.
4. **Sasha**: Spent time looking through the data to see which genres sold the most, as well as which unique genres and regions there were in the first place (1 hr). Experimented with different colors and layouts (3-4 hours). The hardest/most time-consuming part was all the little tweaks: making sure margins fit everything in correctly, aligning text (and the legend) properly, etc..