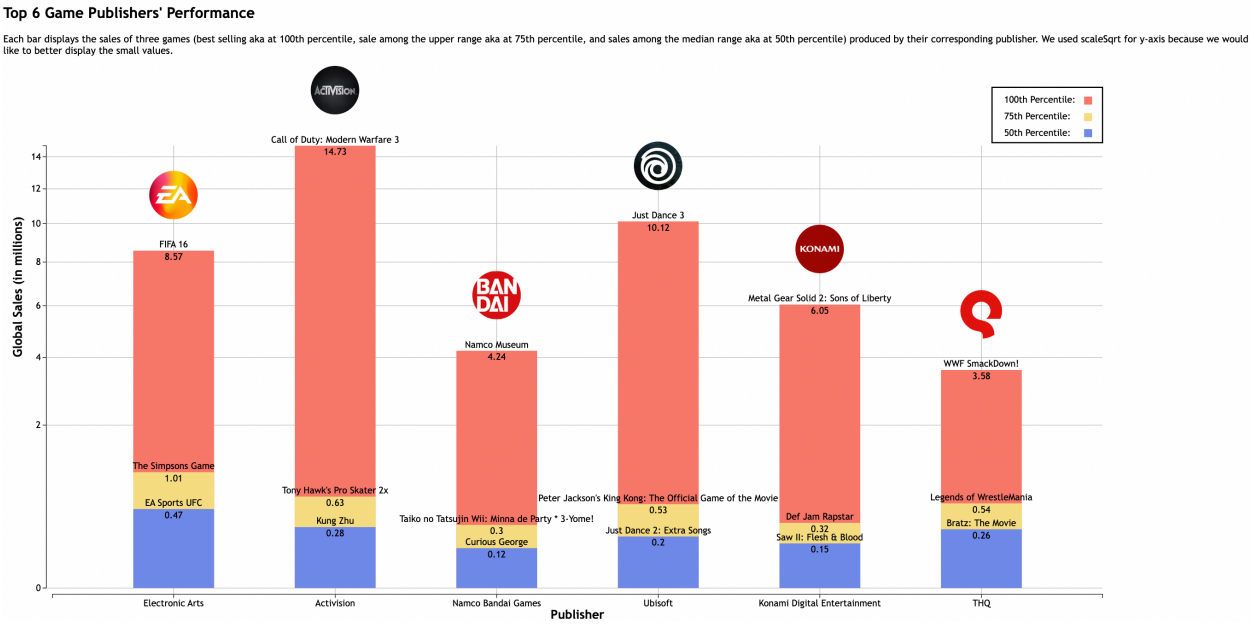
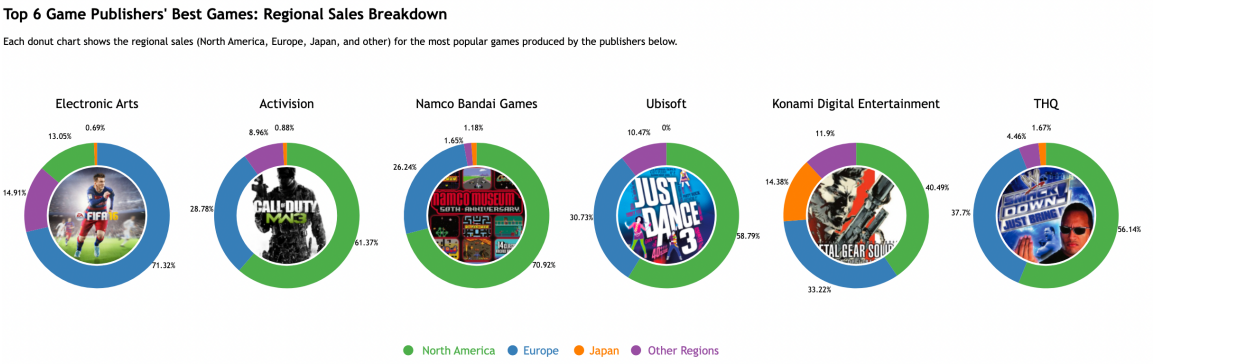
INFO 5100 Project 1 Report

George Gu (lg576), Jerray Wu (sw994), Xiaohan Wang (xw642)

Part 1. Visualization





Part 2. Data

For our project, we used the “Video Games Sales Dataset” from Kaggle (link). This data set consists of video games sales data as of December 22, 2016, alongside some game ratings data scraped from VzCharts. The columns “Name,” “Platform,” “Year_of_Release,” “Genre,” “Publisher,” “NA_Sales,” “EU_Sales,” “JP_Sales,” “Other_Sales,” and “Global_Sales” represent the sales data, whereas the remaining columns “Critic_Score,” “Critic_Count,” “User_Score,”

"User_Count," "Developer," and "Rating" refer to the game ratings data. Given the nature of combining two separate sources of data, we can see the original data set have a lot of empty cells on the game ratings side. Hence, we decided to focus on the sales part.

We discovered that the data set includes games from as many as 583 publishers, with over hundreds of games alone for the more famous publishers such as Nintendo, Sony Computer Entertainment, Activision, etc. As it seems unrealistic to plot all of them out, we chose to compare and contrast the top six publishers with the most published games by looking at their global sales from games on the 50^{th}, 75^{th} and 100^{th} (best seller) percentile for each publisher. Unlike the traditional way of calculating the mean, taking a closer look at specific games allowed us to zoom in and explore the individual stories behind these games. Then we compared sales for each publisher's most popular games from North America, Europe, Japan, and other parts of the world to find out if there was any interest within specific regions for each of the selected games.

Part 3. Design Rationale

In order to evaluate the overall performance of a publisher, we needed to take a closer look at some of its most famous games, games in the upper-level range, and games with global sales that fell at the median. We wanted to also consider the least popular games but figured out that this number was 0.01 for all six publishers; therefore, it did not make much sense to base our analysis on them. We believed that these three games could, to some extent, represent the publisher's overall performance, especially since the six companies we chose to plot are all well-known, large publishers with hundreds of games. Bar charts are easiest to present categorial data. In order to zoom in and plot the three games for each bar, we decided to go with overlapping bars with the best seller in the back, followed by the 75^{th} percentile one above, and the median

sales up front. We also decided to color code the bars to represent the category that the games belong to. Finally, we added some legends and labels to make the visualization more informative. In short, the marks used in the stacked bar chart are colored rectangles. The visual channels are varying the vertical aligned length (i.e., global sales), horizontal aligned position (i.e., publisher), vertical aligned length and position (i.e., different games for the same publisher), and color hue of rectangles.

Since the best sellers' global sale numbers differ most between publishers, we decided to zoom in to compare and contrast regional sales to identify any trend or popularity among certain populations for a specific game. To achieve this, we used donut charts to display regional sales by percentages, as the total sales for each game is different. Besides, a donut chart also allows us to compare the percentage for a game in a region to that of another game in that same region. In this case, it would be less intuitive to plot the sales by number. To sum up, the marks used in the pie chart are colored sectors. The visual channels used are varying the area (quantitative attribute) and color hue (categorial attribute) of sectors. Overall, the pie charts serve as an extension to the overlapping bar chart above and enable us to extract more insights as we zoom in further.

Part 4. The Story

The bar chart presents the top six publishers with the largest sales of games in the world, including Electronic Arts, Activision, Namco Bandai Games, Ubisoft, Konami Digital Entertainment, and THQ. Among those publishers, three of them are from North America; among which, the top two are both from the United States. Two are from Japan, and one is from Europe, which shows the dominance of the United States in global gaming industry.

We originally wanted to select the games with the highest sales volume, the lowest sales volume and the median sales volume produced by each publisher to plot games in the overlapping bar chart, but we found that the sales volume of the games with the lowest sales volume was too small compared with the games with the highest sales volume, so that it could not be effectively displayed on stacked bar chart. Therefore, we chose games with median sales and 75th percentile sales to show the sales gap of different games of each publisher, and also analyze the game categories and customer groups that the publisher is good at producing according to the types of games with the highest sales. For example, the game with the highest sales volume for Electronic Arts is FIFA 16. This publisher is a major producer of sports games. Many famous sports games and action games are produced by this publisher.

The figures below the bar chart are donut charts, which show the regional distribution of sales of the games with the highest sales volume produced by each of the six manufacturers. Among them, green represents sales in North America, blue represents sales in Europe, orange represents sales in Japan, and purple represents sales in other regions. Through our analysis, we found that the sales volume in Japan is generally the lowest, which may be caused by the small population of Japan compared with the other regions. In addition, we found that the regional distribution of sales volume had little relationship with the location of the publisher. The top six manufacturers mainly based their markets in North America, followed by Europe. We can see that video games are very popular in these two regions.

In addition, because our team members are all from China, we found that Chinese users are more inclined to play web games or multiplayer online competitive games than North America and Europe after analyzing the types of games that Chinese netizens love. One reason may be that Chinese users are not as active in buying switch or PS5 as those from Europe and

North America. Most Chinese users will choose to play games on PCs instead of consoles. Besides, we can infer that other countries with relatively low sales on console games have similar reasons as China, which explains why the donut charts display a one-sided trend in the western world. For example, the per capita GDPs for these countries are not as high as that of European and North American countries, so people there do not want to spend money purchasing console devices and games, which leads to console games being unpopular in these countries.

Part 5. Team Contributions

Name	Contributions
George Gu	· Tested several types of visualizations and finalized the appropriate types we used for our product. · Parts of coding, including processing data and stacked bar chart. · Project Report Write-Up (Part 2 Data, Part 3 Design Rationale, Part 5 Team Contributions).
Jerray Wu	· Came up with the visualization ideas and proposed demos of those ideas using Figma. · Majority of coding, including processing data, filtering, stacked bar chart, and pie chart.
Xiaohan Wang	· Examined the video games dataset and analyzed the elements we can use for our final project as well as their correlations. · Parts of coding, including pie chart. · Project Report Write-Up (Part 4 The Story).

We spent about two days processing the data set and three days doing the visualization. The data processing actually took us the most time, because we needed to spend time filtering and analyzing what goes into our visualization and what should be left out. The visualization part was manageable. It was relatively straightforward to make the plots as long as we had the data ready to go, despite we spent an entire day modifying the looks and CSS to make it look more polished.