

Executive Summary

This paper analyzes a comprehensive dataset of 16,598 records, detailing video game sales from 1980 to 2020. The Cross-Sectional pool dataset includes 11,493 unique game sales, providing insights into game names, release years, genres, platforms, and sales figures across various regions, including North America (NA), Europe (EU), Japan (JP), and globally. After cleaning our data and running our analysis we found that Action games dominate global sales particularly in North America. Sports and Shooter games however followed in second and third leading genres of games. We found that North America is the largest consumer of video games in comparison to Europe and Japan while PS2, Xbox 360 and DS were the highest performing platforms globally. Each of us focused on a particular portion of the dataset where Katie worked on publishers, finding that North America leads in video game consumption. Thatcher explored platforms that showed PS2, Xbox 360 and DS as the most dominant platforms and Alissa examined genres, identifying action games as the top-selling genre globally and in North America.

Our combined findings suggested that video game companies should focus on creating titles for the PS5 and Xbox One, as well as other Sony and Microsoft systems, and aim for the North American market in order to increase sales. For gaming companies in the future, this focus could maximize sales and streamline efforts to create new games and platforms.

Cleaning Data

Our dataset, [df](#), had over 16598 observations collected from 1980-2020, of games that sold more than 100,000 copies. We omitted any games released before 2000, as many of those games and/or the consoles they are played on no longer exist. We then saved it as “post” which is the database we will be basing the rest of our analysis on. `post<- video.game %>% select(Name, Year, Platform, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales) %>% filter(Year>=2000)`. We looked at the column headers, and since they were all very clear we did not have to rename any of them. All sales figures were in terms of millions so there was no need to rescale them either. Our dataset had no missing variables but some values were listed as “NA”. These NA values were sparse throughout our dataset so we chose to do the deletion when necessary approach, in order to minimize the amount of data lost.

```
library(tidyverse)
library(sqldf)
df<-read.csv('https://raw.githubusercontent.com/katieluong33/BUAN314-Project/refs/heads/main/vgsales.csv?token=GHSAT0AAAAACZU5YMT6F3')
post<- df %>%
  select(Rank,Name, Year, Platform, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales) %>%
  filter(Year>=2000)
```



Data		
post	14624 obs. of 10 variables	
vgsales	16598 obs. of 11 variables	

Table Merge

Our dataset was split into two tables with a common key of “Rank” in both. The first table “Vg_info” contained the rank, name, year, platform, genre, and publisher info of each game, and the second table “vg_sales” had the rank, sales in North America, Europe, Japan, other countries, and global

sales. Since “Rank” was the common key in both of the tables, we used that to join the two tables together. Prior to doing that we checked to make sure there were no duplicate rank values in either table. **sum(duplicated(vg_sales\$Rank)), sum(duplicated(Vg_info\$Rank))**. The end product is Figure 1, which displays the info and sales of each game as one table.

```
#split table and join again
Vg_info <- post %>%
  select(Rank, Name, Year, Platform, Genre, Publisher)
vg_sales <- post %>%
  select(Rank, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales)
#check for duplicates
sum(duplicated(vg_sales$Rank))
sum(duplicated(Vg_info$Rank))

#join the tables using inner join on Rank
table <- 'SELECT *
FROM Vg_info AS t1
INNER JOIN vg_sales AS t2
ON t1.Rank = t2.Rank'
sqldf(table)
```

Figure 1

	Rank	Name	Year	Platform	Genre	Publisher	Rank	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	1	Wii Sports	2006	Wii	Sports	Nintendo	1	41.49	29.02	3.77	8.46	82.74
2	3	Mario Kart Wii	2008	Wii	Racing	Nintendo	3	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	2009	Wii	Sports	Nintendo	4	15.75	11.01	3.28	2.96	33.00
4	7	New Super Mario Bros.	2006	DS	Platform	Nintendo	7	11.38	9.23	6.50	2.90	30.01
5	8	Wii Play	2006	Wii	Misc	Nintendo	8	14.03	9.20	2.93	2.85	29.02
6	9	New Super Mario Bros.	2009	Wii	Platform	Nintendo	9	14.59	7.06	4.70	2.26	28.62
7	11	Nintendogs	2005	DS	Simulation	Nintendo	11	9.07	11.00	1.93	2.75	24.76
8	12	Mario Kart DS	2005	DS	Racing	Nintendo	12	9.81	7.57	4.13	1.92	23.42
9	14	Wii Fit	2007	Wii	Sports	Nintendo	14	8.94	8.03	3.60	2.15	22.72
10	15	Wii Fit Plus	2009	Wii	Sports	Nintendo	15	9.09	8.59	2.53	1.79	22.00

Analysis

Figure 2 & 3:

We created a violin plot that compares 3 different genres(Action, Role-Playing, Shooter) against the Global_Sales.It combines the features of a box plot to show quartiles and median with a density plot, which provides a visual summary of the data distribution for each genre.To go along with this we used SQL to aggregate and rank the data by total Global Sales for the top-performing years to give us additional insights.The query highlights that the Action genre dominates the top global sales, with peaks in 2009 (139.36 million) and 2008(136.39 million). We can attribute this peak to the release of games like Grand Theft Auto IV and Call of Duty:Modern Warfare 2.This aligns with the violin plot's shape for Action, where there is a noticeable density of higher global sales.

Action games are consistently high-performing, both in top sales and overall distribution.

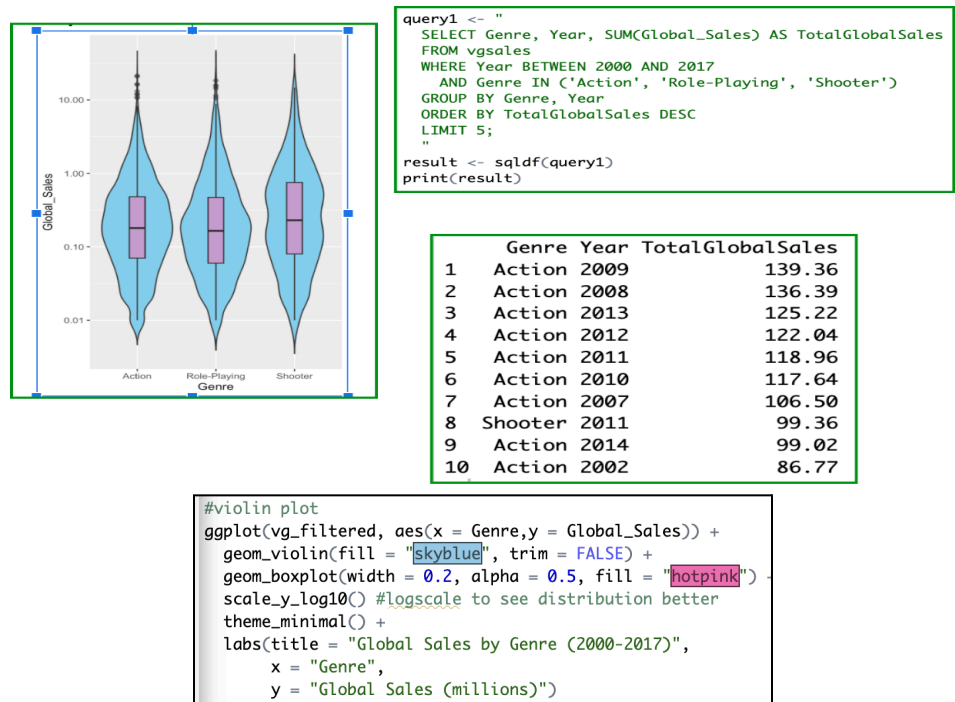


Figure 4:

The data show the number of games distributed across genres and platforms (DS, Wii, X360), with certain genres, such as "Action" and "Sports," having significantly more games compared to others like "Puzzle" and "Fighting." An obvious relationship emerges where "Action" and "Sports" dominate across all platforms, while other genres have fewer games overall. We used colors to represent platforms—red for DS, green for Wii, and blue for X360—distinguishing the data, making it easy to compare the distribution visually.

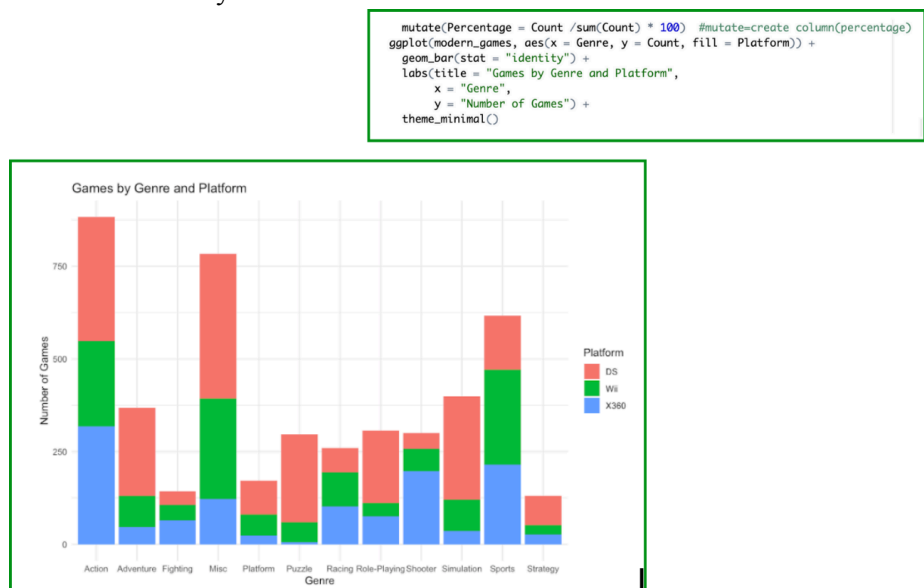


Figure 5:

The data show NA Sales distributed across three genres: Action, Shooter, and Sports. Action games saw the highest sales, peaking at over 70 million units around 2008–2010, while Sports briefly surpassed Shooter in the mid-2000s before both began to decline. Shooter sales peaked at nearly 60 million units in 2010, showing strong growth before a sharp drop. The trends are clear, with all genres

experiencing significant declines after 2010, likely due to market saturation or shifting player interests. The plot uses distinct colors to differentiate genres without smoothers or facet wrappers, ensuring the data trends are easy to analyze. The graph clearly highlights each genre's trajectory, such as Action's dominance around 2008–2010 and the sharp declines after 2010, providing a clear way to analyze long-term trends and relationships.

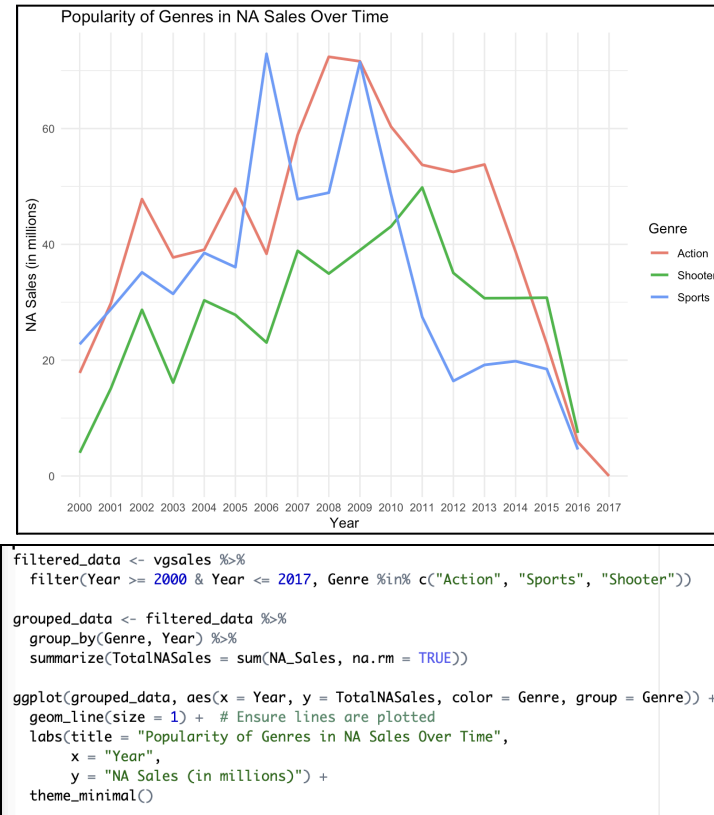


Figure 6 & 7:

The scatter plot shows the Global Sales of video games across years, with a peak around 2006-2009 where several games reached significant sales figures, followed by a gradual decline in the later years. The distribution is concentrated below 20 million, with a few outliers exceeding 80 million, indicating a small number of extremely successful game titles. The points clearly depict trends, using a line plot and adjusting the x-axis limits (seen in code) to focus on 2000-2010 further shows the peak years and overall trend during the most active period. The scatter plot shows the spread of Global Sales over time but does not reveal which genres dominate overall. Using SQL, we summarized total Global Sales by genre, highlighting Action as the leader (1751.18M), followed by Sports and Shooter.

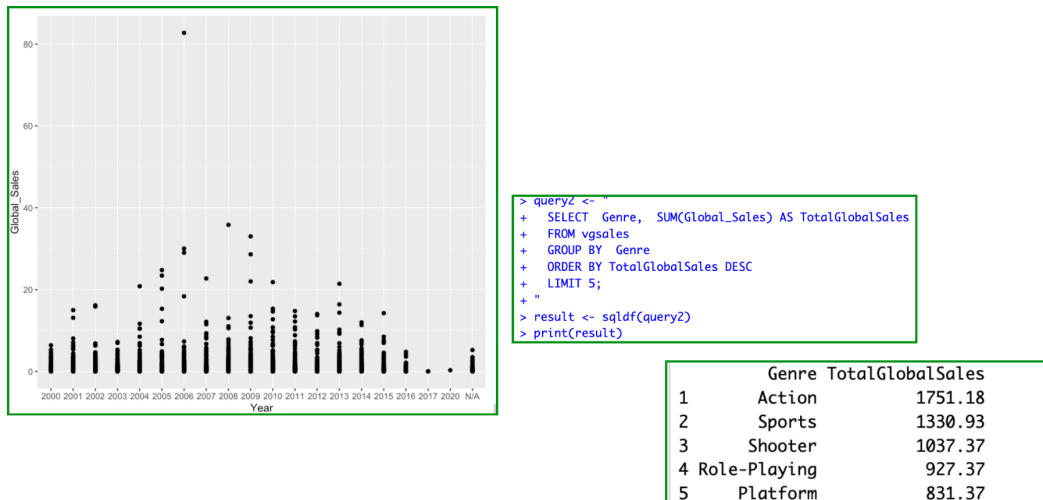


Figure 8 & 9:

Query 8 focuses on listing the top Global Sales values for the year 2000 onward, ordered chronologically, that shows early contributors to the market's success. This query helps identify patterns in sales over time. Query 9 goes deeper into the Action genre, revealing the top 10 games by Global Sales as seen below. Notably, titles from the Grand Theft Auto series dominate as the top 6, confirming the Action genre's outsized contribution to overall sales. By focusing on a specific genre, this query provides a clearer understanding of which games and franchises are driving its success, showing why Action leads in total sales and Global Sales.

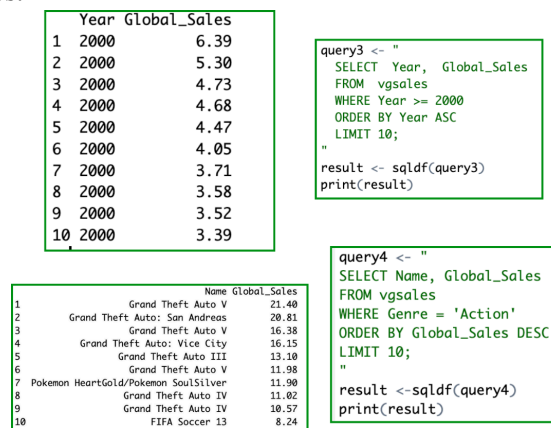


Figure 10:

```

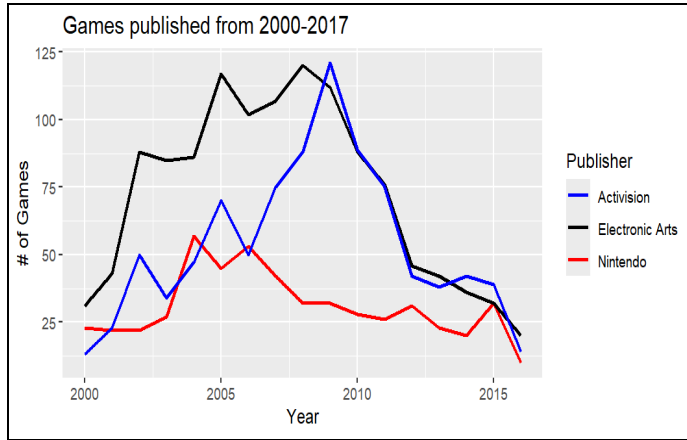
#SQL for number of games by publisher
num_pub <- 'SELECT Publisher, COUNT(*) AS Games_Published
FROM post
GROUP BY Publisher
ORDER BY Games_Published DESC
LIMIT 10'
sqldf(num_pub)

```

	Publisher	Games_Published
1	Electronic Arts	1243
2	Activision	919
3	Ubisoft	902
4	Namco Bandai Games	844
5	Konami Digital Entertainment	707
6	THQ	691
7	Nintendo	532
8	Sony Computer Entertainment	528
9	Sega	523
10	Take-Two Interactive	400

First we ran an sql query from the post df to see the top ten publishers who released the most games over the 20 year period. We sorted it in DESC order to get the publishers with the most games rather than least. We noticed that Electronic Arts released the most by over 300 games.

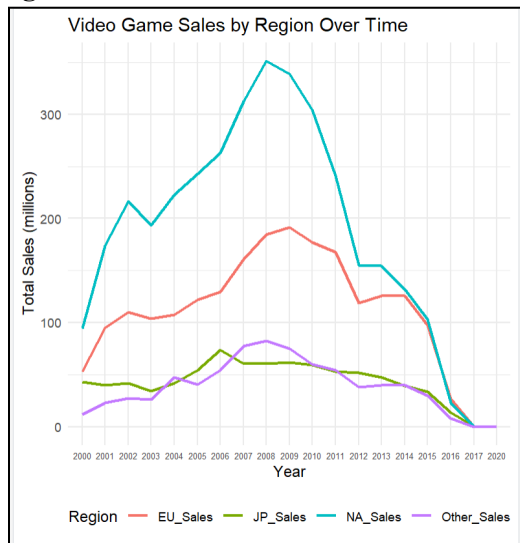
Figure 11:



```
#Top publishers from 2000-2017 bar graph
ggplot() +
  geom_line(data = ninfreq2, aes(x = Year, y = frequency, color = "Nintendo"), size = 1) + # Line for ninfreq2
  geom_line(data = Eafreq2, aes(x = Year, y = frequency, color = "Electronic Arts"), size = 1) + # Line for Eafreq2
  geom_line(data = Actfreq2, aes(x = Year, y = frequency, color = "Activision"), size = 1) +
  scale_color_manual(values = c("Nintendo" = "red", "Electronic Arts" = "black", "Activision" = "blue")) +
  labs(title = "Games published from 2000-2017",
       x = "Year",
       y = "# of Games",
       color = "Publisher") # Title for the legend
```

After running the sql query to find the top publishers, we wanted to see how many games were being released every year. The line graph shows the number of games published by these three publishers. We chose Electronic Arts and Activision because they published the most games. We also chose to include Nintendo because of its popularity within our generation, thanks to gaming consoles like the DS and Switch. We can see that both Activision and EA released the largest amount of video games in the 2008-2009 years, and all three decreased the number of games published significantly after 2015. Please note the # of games scale is 100.

Figure 12:

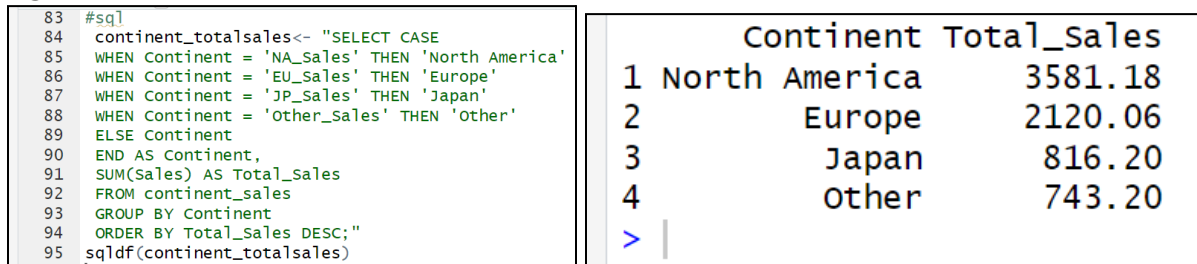


```
ggplot(sales_long, aes(x = Year, y = Sales, color = Region, group = Region)) +
  geom_line(size = 1) + # Line plot
  labs(
    title = "Video Game Sales by Region Over Time",
    x = "Year",
    y = "Total Sales (millions)",
    color = "Region"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(size = 6))
```

Figure 12 shows a breakdown over the years of sales in each continent. The line graph displays the sales in millions, grouped by continent. Europe, Japan, and other countries spiked from 2006 to 2011. Sales in North America dramatically started rising in 2003 until 2008. It shows that North American video game sales have always been more than any other continent. Sales in Japan and other continents are very

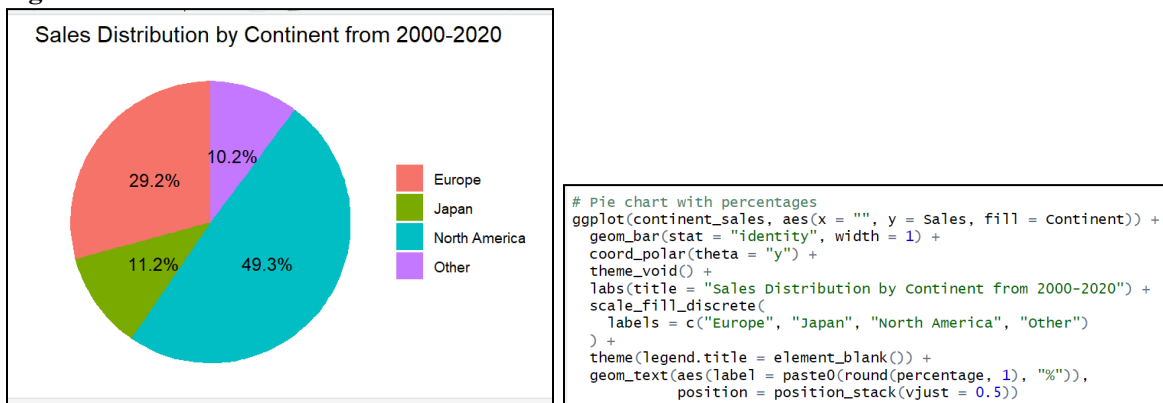
similar and EU sales are significantly higher than Japan but lower than NA. This may be explained by the recession of 2008, when Americans' struggled to cover basic needs let alone entertainment expenses. Before making the graph, it was necessary to create a dataframe with only the sales data and then summing it to get total sales. It was also necessary to remove the last observation because it contained the info for any gems whose year was listed as NA. `continentsales_byyear <- continentsales_byyear[-20,]`.

Figure 13:



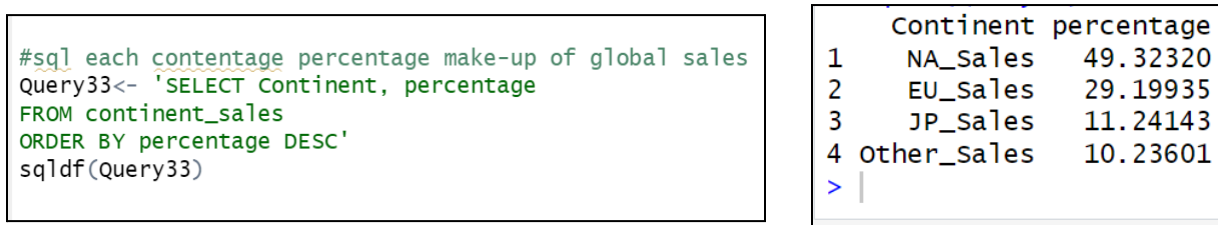
Sales for the entire 20 year period was calculated using an sql query, with North America in the top spot with 3581. But since the sales are in millions this is actually 3,581,180,000. Sales in North America were greater than Europe by over 1,000,000.

Figure 14:



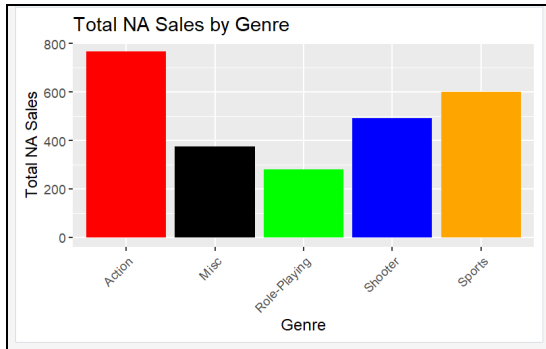
The pie chart breaks down the sales distribution around the globe but by percentages. From this we can see that North America makes up almost half of the video game sales of the entire globe. This is interesting because Europe makes up 9.3% of the world's population compared to North America's 7.66% (["Population by Continent 2024"](#)).

Figure 15



The sql query shows a more detailed breakdown of percent of sales by continent and ranks them in descending order.

Figure 16:

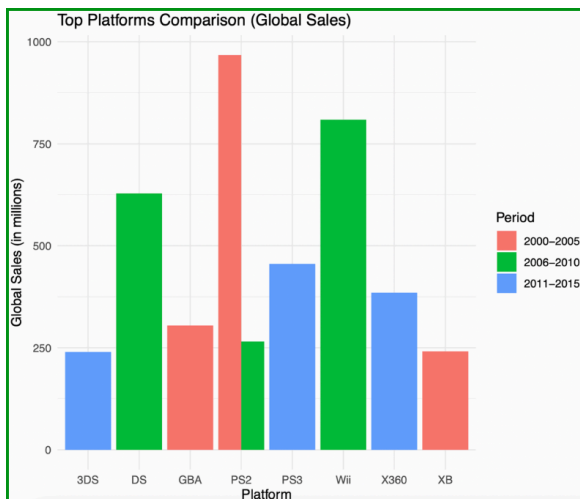


```
# bar plot
ggplot(genre_na, aes(x = Genre, y = total_na_sales, fill = color_map)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Total NA Sales by Genre",
       x = "Genre",
       y = "Total NA Sales") +
  scale_fill_identity() #color from map
```

Shows the top 5 genres that were most popular in North America and how many games were sold in that category over 20 years. Data displayed here reinforces earlier conclusions that the Action genre is the most popular of all.

Figure 17:

Visualization #17 compares the global sales of various gaming platforms across three periods of time: 2000-2005, 2006-2010, and 2011-2015. The y-axis represents global sales in millions, while the x-axis lists the platforms. Each platform's sales are depicted with bars for each period, allowing for a comparison of sales performance over time. In terms of data preparation, there were no missing or null values to address, as the data was already cleaned. Column headers did not require renaming, however regrouping of the data into the correct time frame was necessary. Additionally, there was no need to rescale or transform any variables for this graph. The data was straightforward and ready for visualization without further modifications.



```
ggplot(plot_data, aes(x = Platform, y = Global_Sales, fill = Period)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Top Platforms Comparison (Global Sales)",
       x = "Platform", y = "Global Sales (in millions)") +
  theme_minimal()
```

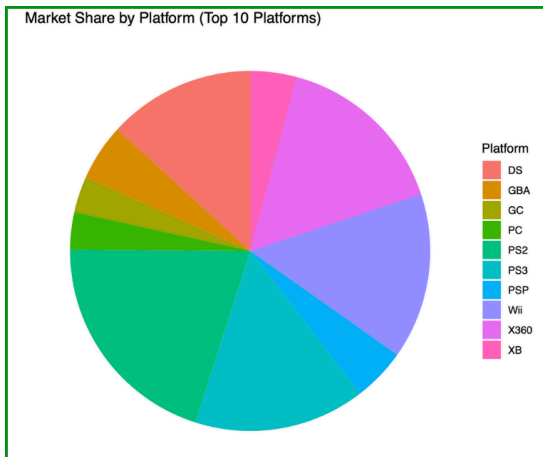
Figure 18

To build off this visualization I thought it was necessary to also incorporate an SQL. Within this I separated it into three different ones allowing for a clearer understanding of what was happening during each time frame. Additionally I feel that this clears up any confusion about PS2 being included in two different time frames.

> sqldf(Q1.1)				> sqldf(Q1.2)				> sqldf(Q1.3)			
Platform	Total_Sales	Period		Platform	Total_Sales	Period		Platform	Total_Sales	Period	
1	PS2	967.66	2000-2005	1	Wii	809.28	2006-2010	1	PS3	455.43	2011-2015
2	GBA	304.78	2000-2005	2	DS	628.37	2006-2010	2	X360	385.08	2011-2015
3	XB	241.21	2000-2005	3	X360	575.38	2006-2010	3	3DS	239.68	2011-2015

Figure 19:

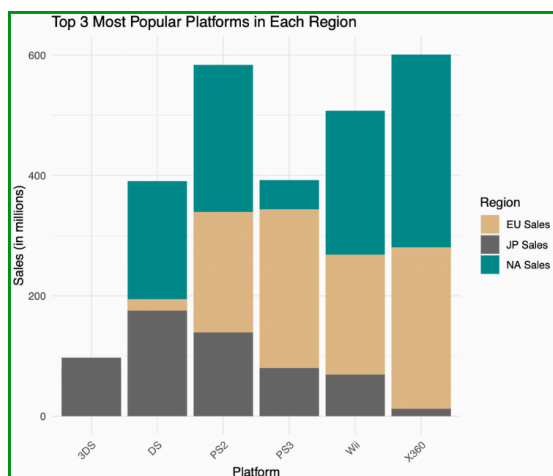
Visualization #19 illustrates the market share of the top 10 gaming platforms using a pie chart. Each platform's market share is represented by a slice, with the size of each slice corresponding to the platform's share of the market—larger slices indicate larger market shares. In this instance, there were no missing values to address, as they were already cleaned from the original dataset. I focused on calculating the market share and identifying the top 10 platforms to facilitate easy graphing and interpretation of the data. No transformation or modification of variables was necessary for this graph.



```
ggplot(market_share, aes(x = "", y = Share, fill = Platform)) +  
  geom_bar(stat = "identity", width = 1) +  
  coord_polar("y") +  
  labs(title = "Market Share by Platform (Top 10 Platforms)") +  
  theme_void()
```

Figure 20

Visualization #20 the data was already clean and complete, so no missing or null values needed addressing. Column headers were appropriately named, and no re-coding or transformation of variables was necessary. The data was summarized to identify the top three platforms in North America, Europe, and Japan. Initial exploratory visualizations confirmed the distribution of sales across platforms and regions. The final bar chart, with different colors representing sales in each region, effectively highlights the top platforms and allows for easy comparison of their popularity across North America, Europe, and Japan.



```
ggplot(top_platforms, aes(x = Platform)) +  
  geom_bar(aes(y = NA_Sales, fill = "NA Sales"), stat = "identity") +  
  geom_bar(aes(y = EU_Sales, fill = "EU Sales"), stat = "identity") +  
  geom_bar(aes(y = JP_Sales, fill = "JP Sales"), stat = "identity") +  
  scale_fill_manual(values = c("NA Sales" = "#1f77b4", "EU Sales" = "#ff7f0e", "JP Sales" = "#2ca02c")) +  
  labs(title = "Top 3 Most Popular Platforms in Each Region",  
       x = "Platform",  
       y = "Sales (in millions)",  
       fill = "Region") +  
  theme_minimal() +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

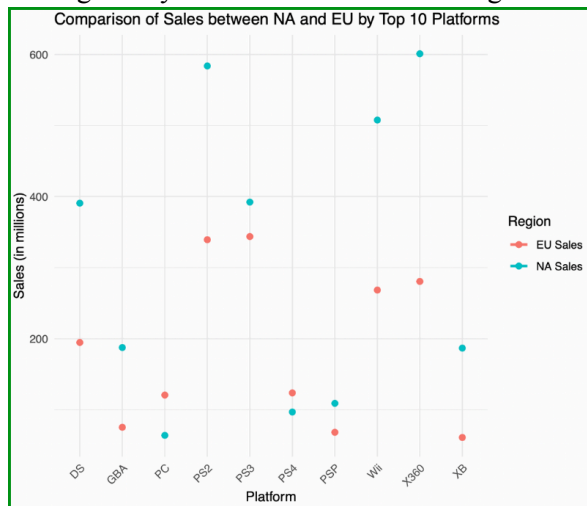
Figure 21

For this visualization an SQL was used to see on a deeper level some of the differences in sales amongst the different regions. For instance this allowed us to see that Wii, a top three selling platform within North America, did not make the top 3 in either of the other regions. Wii didn't even make the top 5 within Japan which further enforces how preferences change from region to region. Something we have found interesting throughout our work.

> sqldf(Q2.1)		> sqldf(Q2.2)		> sqldf(Q2.3)	
Platform	NA_Sales	Platform	EU_Sales	Platform	JP_Sales
1	X360 601.05	1	PS3 343.71	1	DS 175.55
2	PS2 583.84	2	PS2 339.29	2	PS2 139.20
3	Wii 507.71	3	X360 280.58	3	3DS 97.35
4	PS3 392.26	4	Wii 268.38	4	PS3 79.99
5	DS 390.71	5	DS 194.65	5	PSP 76.79
6	GBA 187.54	6	PS4 123.70	6	Wii 69.35
7	XB 186.69	7	PC 120.65	7	GBA 47.33
8	GC 133.46	8	GBA 75.25	8	GC 21.58
9	PSP 108.99	9	PSP 68.25	9	PSV 20.96
10	PS4 96.80	10	XB 60.95	10	PS 20.14

Figure 22

Visualization #22 shows the distribution of sales figures across the top 10 highest selling gaming platforms (DS, GBA, PC, PS2, PS3, PS4, PSP, Wii, X360, XB) for both North America (NA) and Europe (EU) regions. We did this to further investigate what we had seen in graph number 3 where Europe and North America outperformed Japan. The graph highlights the comparison of sales between these two regions for amongst their top 10 platforms, allowing us to observe how each platform performs in different markets. The strength of the relationships between the variables can be seen by comparing the height of the dots compared to one another for each platform. For instance, the DS platform has high sales in both regions however North America significantly is out performing Europe. The graph does not use a smoother, as it is not applicable for categorical comparison, and facet wrappers were also not used. The use of different colors to distinguish between NA and EU sales is appropriate for visual comparison, making it easy to differentiate the sales figures for each region.



```
ggplot(platform_sales, aes(x = Platform)) +
  geom_line(aes(y = NA_Sales, color = "NA Sales"), size = 1) +
  geom_line(aes(y = EU_Sales, color = "EU Sales"), size = 1) +
  geom_point(aes(y = NA_Sales, color = "NA Sales"), size = 2) +
  geom_point(aes(y = EU_Sales, color = "EU Sales"), size = 2) +
  labs(title = "Comparison of Sales between NA and EU by Top 10 Platforms",
       x = "Platform",
       y = "Sales (in millions)",
       color = "Region") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 23

For this final SQL it builds off of the query we ran regarding top 10 sales amongst the respective regions. However in this case it really emphasizes the differences between North America and Europe while also incorporating total sales as well.

> sqldf(Q3)				
	Platform	NA_Sales	EU_Sales	Total_Sales
1	PS2	583.84	339.29	923.13
2	X360	601.05	280.58	881.63
3	Wii	507.71	268.38	776.09
4	PS3	392.26	343.71	735.97
5	DS	390.71	194.65	585.36
6	GBA	187.54	75.25	262.79
7	XB	186.69	60.95	247.64
8	PS4	96.80	123.70	220.50
9	PC	63.91	120.65	184.56
10	PSP	108.99	68.25	177.24

Conclusion

Each member of our team chose to highlight a different area of the data set in the hopes that we would be able to combine our individual findings to draw a group conclusion. Our data consisted of four different sales categories. Our findings mostly highlighted our area of focus and its relation to sales, whether that be sales in Europe (EU), North America (NA), Japan (JP), or sales on a global level. With this in mind, Katie focused on publishers and their relation to overall sales, generally finding that North America consumes far more video games compared to Europe and Japan. Alissa highlighted the variable genre and its relation to overall sales as well as the individual regions. She found that action games lead the industry in terms of global sales as well as in North America. Action games were followed by sports and shooter genres as the second and third leading genres. Thatcher, our third member, explored the relation between different platforms and sales across different regions as well as globally. His findings indicated that PS2, Xbox 360, and DS were the top three platforms dominating globally. As we each came back to the drawing board with our individual conclusions, we wanted to address the question of “How does this affect the gaming industry?” We found that video game companies should focus on creating games compatible with Sony and Microsoft consoles, such as PS2 and Xbox 360, which in today's age are PS5 and Xbox One. These games should largely be marketed to the North American region, as it is the clear sales leader. We hope that this information might, in the future, allow for efforts to be streamlined and sales to be increased for various gaming companies.