DATA MINING PROJECT:

VIDEO GAME SALES

OVERVIEW:

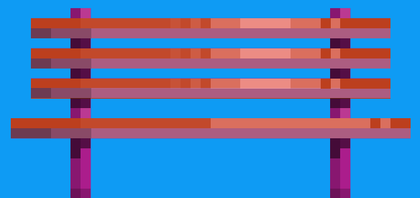
We looked at a 16,598 observation in the Video Game Sales data set.

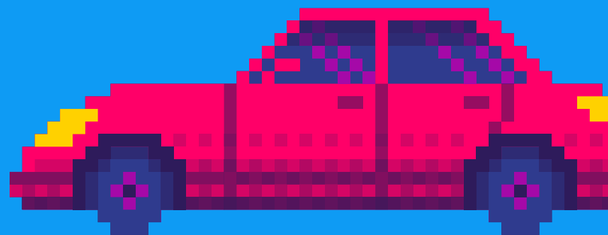
01 Steps to Cleaning Data

02 Inner-Join into one table

03 Visualizations & Queries

04 Conclusions





STEPS TO CLEAN DATA

01

Initial Dataset Overview:

- 16,598 observations of video games released between 1980 and 2020, all with sales exceeding 100,000 copies.

02

Filter Data by Year:

03

Handle Missing Values:

- Verified there were no missing or null values.
- Deleted rows containing "NA" entries only when necessary.

NOTES

- We did not have to deal with missing or null values but we did deleting when necessary to remove the applicable “NA”
- We did not rename column headers as they were already very clear
- We did not rescale it (the figures for sales are in millions)

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

INNER JOIN INTO ONE TABLE

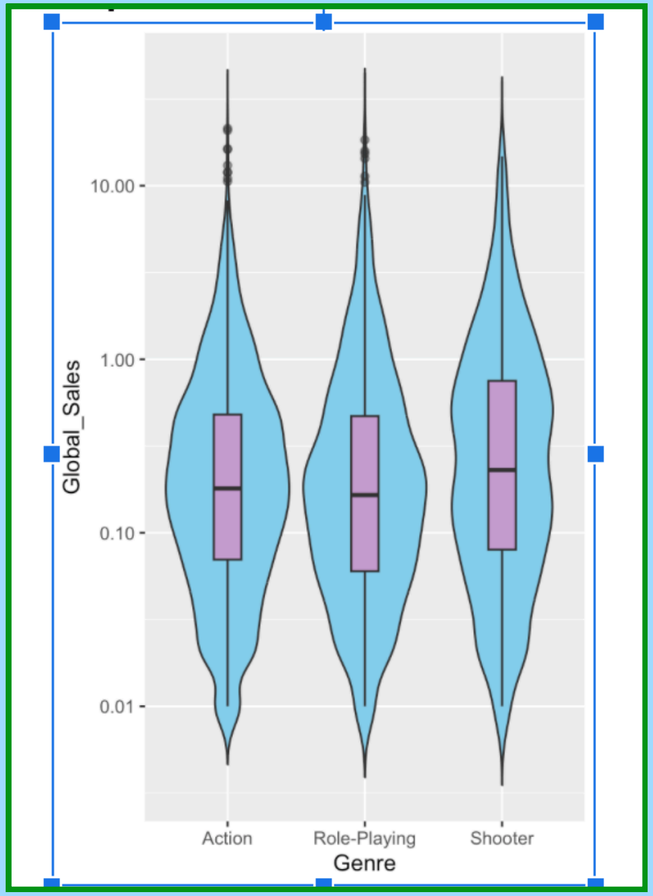
```
#split table and join again
Vg_info <- post %>%
  select(Rank, Name, Year, Platform, Genre, Publisher)
vg_sales <- post %>%
  select(Rank,NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales)
#join the tables using inner join on Rank
table <- 'SELECT *
FROM Vg_info AS t1
INNER JOIN vg_sales AS t2
ON t1.Rank = t2.Rank'
sqldf(table)
```

	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. Wii	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

>


VISUALIZATIONS
& QUERIES

FIGURE 1 & 2




```
query1 <- "
  SELECT Genre, Year, SUM(Global_Sales) AS TotalGlobalSales
  FROM vgsales
  WHERE Year BETWEEN 2000 AND 2017
    AND Genre IN ('Action', 'Role-Playing', 'Shooter')
  GROUP BY Genre, Year
  ORDER BY TotalGlobalSales DESC
  LIMIT 5;
  "
result <- sqldf(query1)
print(result)
```

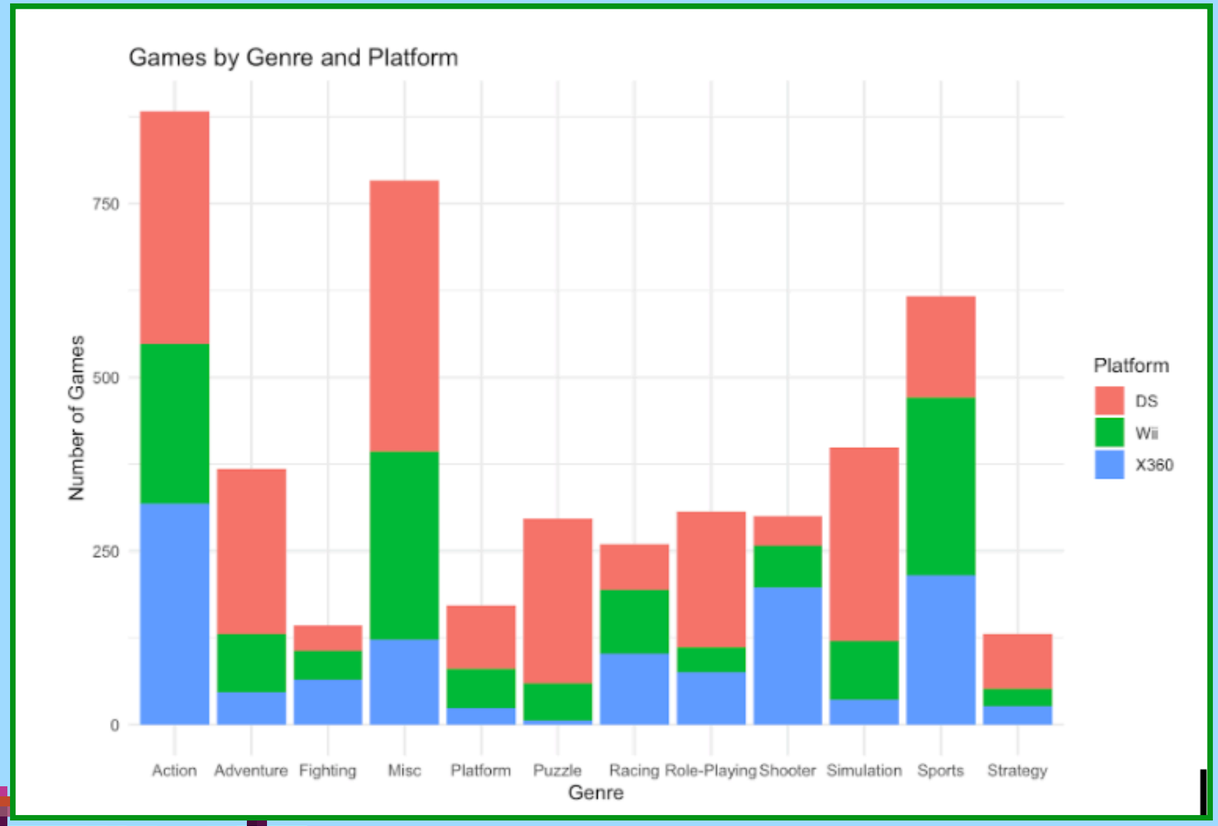
	Genre	Year	TotalGlobalSales
1	Action	2009	139.36
2	Action	2008	136.39
3	Action	2013	125.22
4	Action	2012	122.04
5	Action	2011	118.96
6	Action	2010	117.64
7	Action	2007	106.50
8	Shooter	2011	99.36
9	Action	2014	99.02
10	Action	2002	86.77

WHAT DOES THIS TELL US?

Action dominates Global sales particularly in 2009, with 139.36 million in sales.

FIGURE 3

```
  mutate(Percentage = Count /sum(Count) * 100)  #mutate=create column(percentage)
ggplot(modern_games, aes(x = Genre, y = Count, fill = Platform)) +
  geom_bar(stat = "identity") +
  labs(title = "Games by Genre and Platform",
       x = "Genre",
       y = "Number of Games") +
  theme_minimal()
```



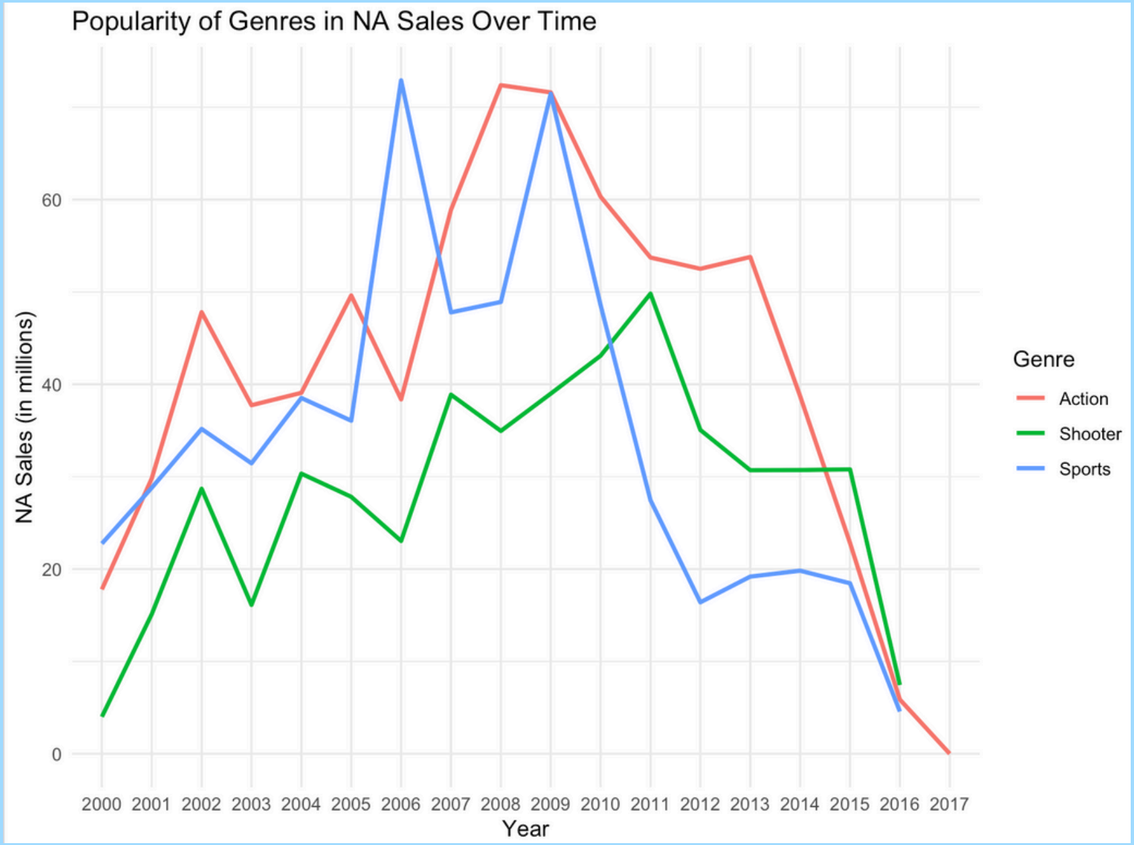
WHAT DOES THIS TELL US?

-DS has the largest library across most genre

-Xbox360 shows a focus on Shooter games

-Sports and Simulation) are strongly represented on platforms like the Wii, likely due to its interactive controls.

FIGURE 4




```r
filtered_data <- vgsales %>%
  filter(Year >= 2000 & Year <= 2017, Genre %in% c("Action", "Sports", "Shooter"))

grouped_data <- filtered_data %>%
  group_by(Genre, Year) %>%
  summarize(TotalNASales = sum(NA_Sales, na.rm = TRUE))

ggplot(grouped_data, aes(x = Year, y = TotalNASales, color = Genre, group = Genre)) +
  geom_line(size = 1) +  # Ensure lines are plotted
  labs(title = "Popularity of Genres in NA Sales Over Time",
       x = "Year",
       y = "NA Sales (in millions)") +
  theme_minimal()
```

WHAT DOES THIS TELL US?

Action games

-dominate, with a peak in 2009

-Sales drop sharply after 2011

Sports

-Sports games experienced steady growth in North American sales from 2000 to 2008

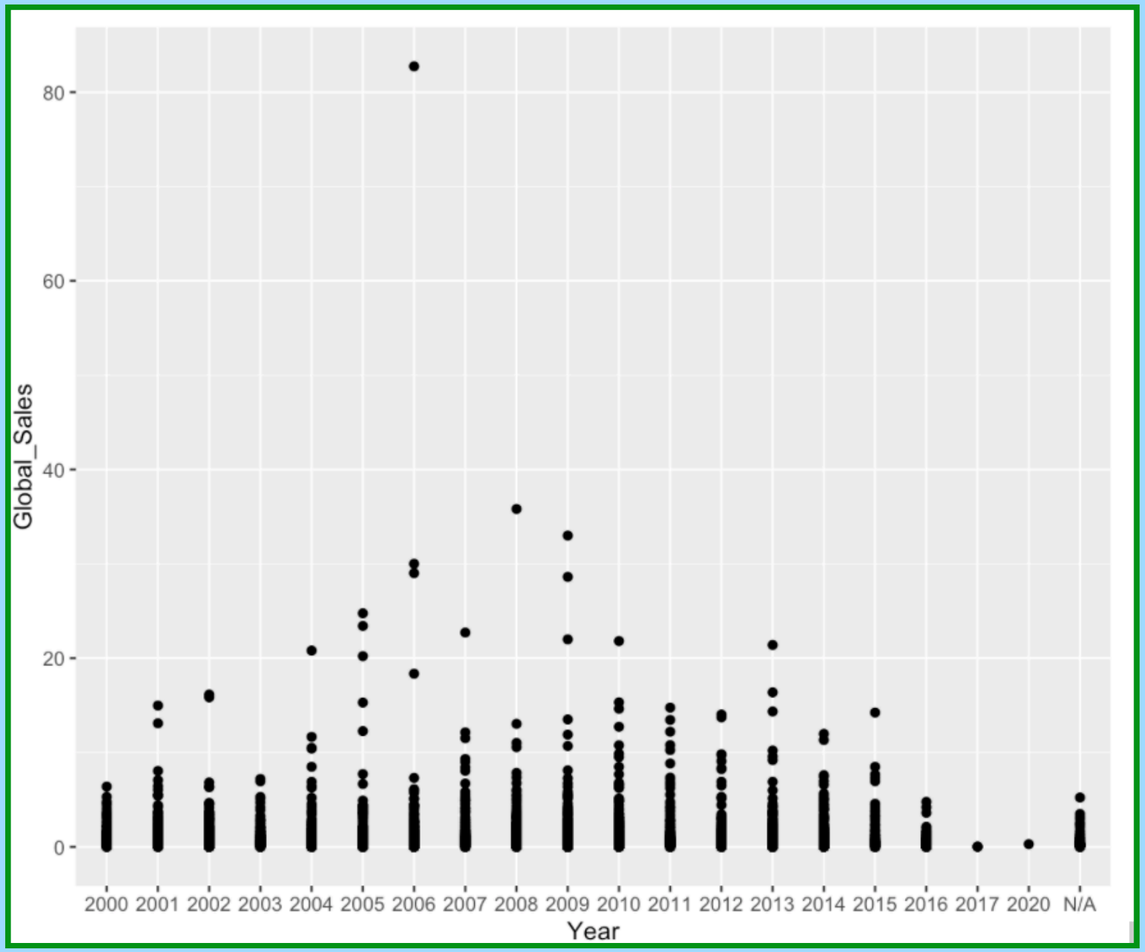
-Post-2010, sales dropped significantly

Shooter games

-steady growth from 2000, peaking around 2010 surpassing 60 million in revenue

-Shooter games experience a decline after 2011

FIGURE 5 & 6



WHAT DOES THIS TELL US?

-Action leads global sales with 1,751 million units, followed by Sports (1,330 million) and Shooter (1,037 million)
-Outliers in dataset represent high selling games

```
> query2 <- "
+   SELECT  Genre,  SUM(Global_Sales) AS TotalGlobalSales
+   FROM vgsales
+   GROUP BY  Genre
+   ORDER BY TotalGlobalSales DESC
+   LIMIT 5;
+ "
> result <- sqldf(query2)
> print(result)
```

	Genre	TotalGlobalSales
1	Action	1751.18
2	Sports	1330.93
3	Shooter	1037.37
4	Role-Playing	927.37
5	Platform	831.37

FIGURE 7 & 8

	Year	Global_Sales
1	2000	6.39
2	2000	5.30
3	2000	4.73
4	2000	4.68
5	2000	4.47
6	2000	4.05
7	2000	3.71
8	2000	3.58
9	2000	3.52
10	2000	3.39

```
query3 <- "
  SELECT  Year,  Global_Sales
  FROM  vgsales
  WHERE Year >= 2000
  ORDER BY Year ASC
  LIMIT 10;
"
result <- sqldf(query3)
print(result)
```

WHAT DOES THIS TELL US?

The highest global sales figure for a single game in 2000 was 6.39 million units, with the 10th highest being 3.39 million units.

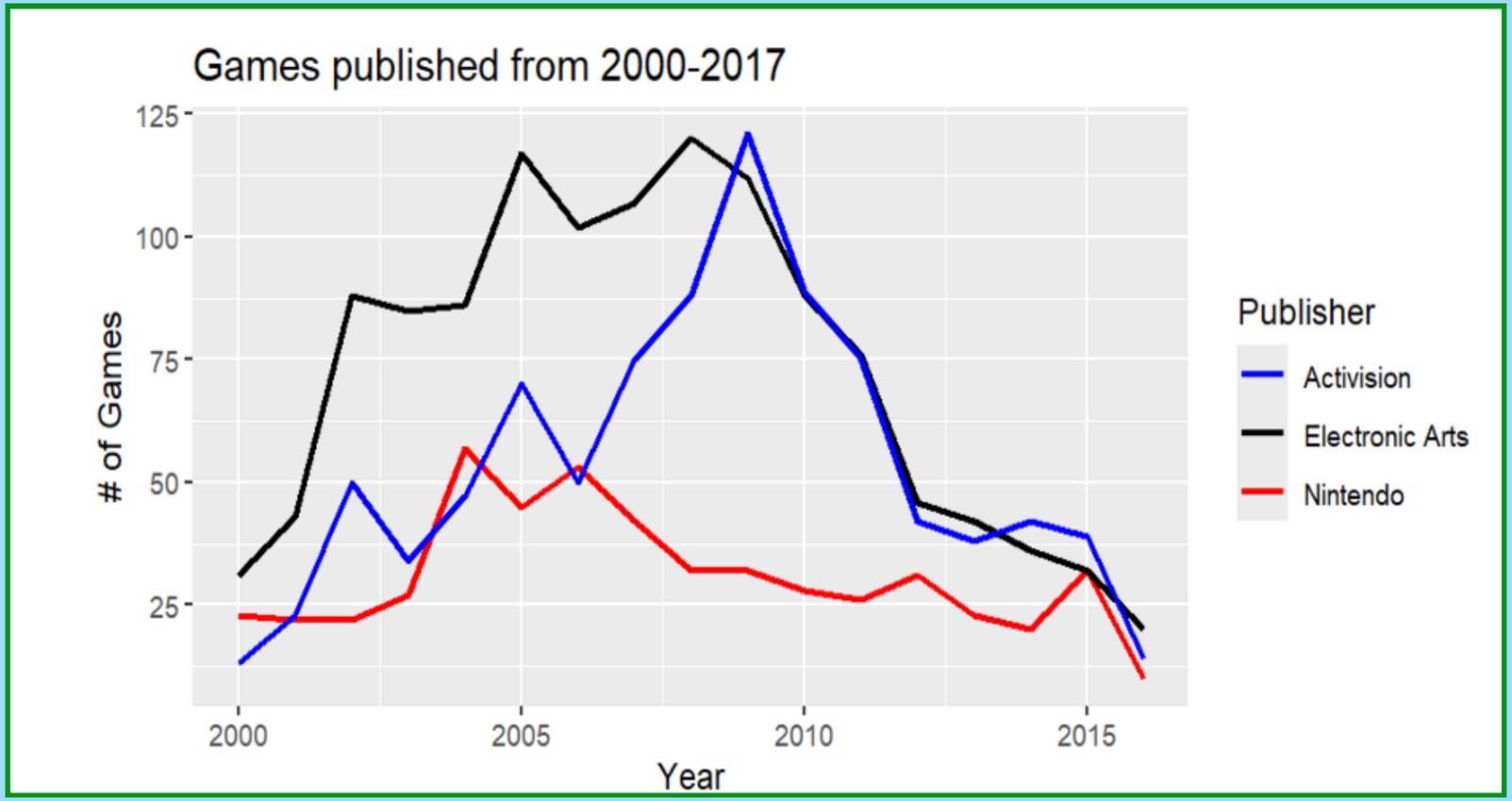
	Name	Global_Sales
1	Grand Theft Auto V	21.40
2	Grand Theft Auto: San Andreas	20.81
3	Grand Theft Auto V	16.38
4	Grand Theft Auto: Vice City	16.15
5	Grand Theft Auto III	13.10
6	Grand Theft Auto V	11.98
7	Pokemon HeartGold/Pokemon SoulSilver	11.90
8	Grand Theft Auto IV	11.02
9	Grand Theft Auto IV	10.57
10	FIFA Soccer 13	8.24

```
query4 <- "
SELECT Name, Global_Sales
FROM vgsales
WHERE Genre = 'Action'
ORDER BY Global_Sales DESC
LIMIT 10;
"
result <-sqldf(query4)
print(result)
```

WHAT DOES THIS TELL US?

-7 out of 10 games belong to the GTA series
-FIFA Soccer represents an outlier in this list

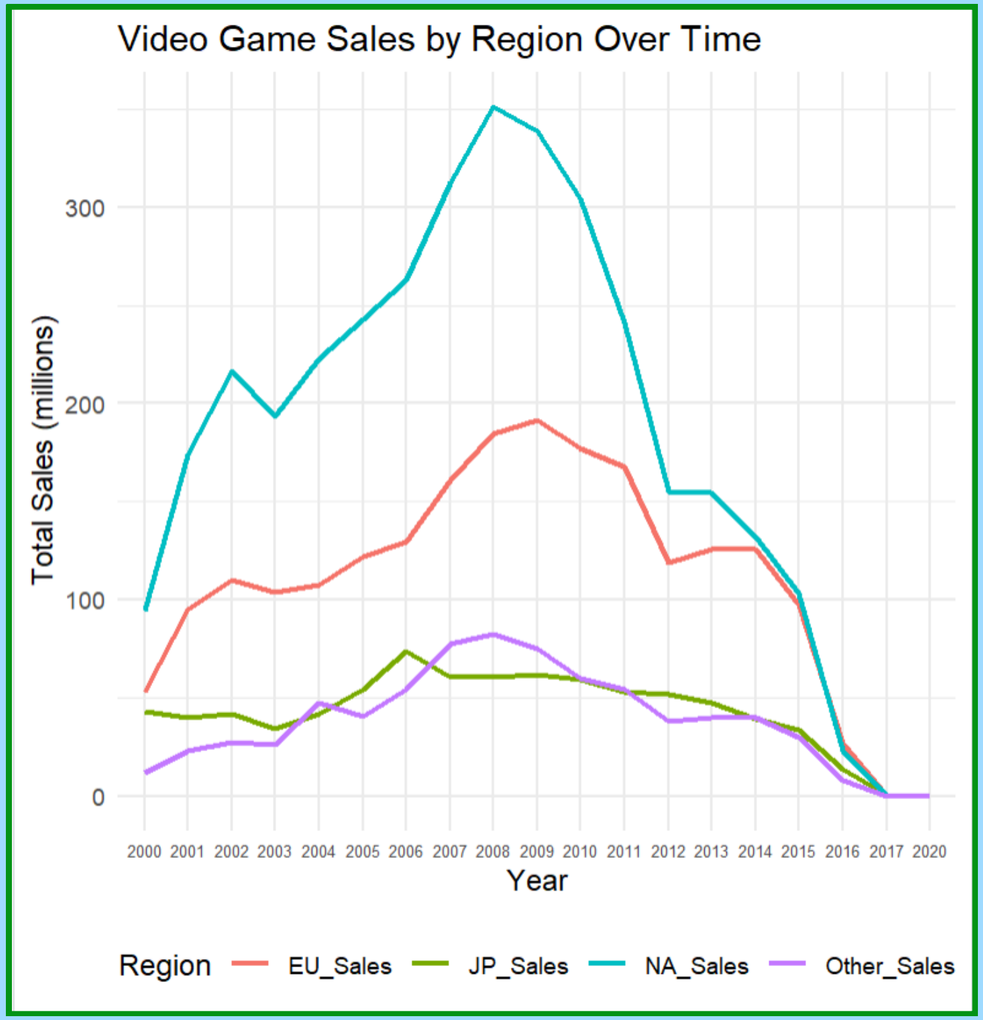
FIGURE 9 & 10




```
+ 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
```

```
num_pub <- 'SELECT Publisher, COUNT (*) AS Games_Published
FROM post
GROUP BY Publisher
ORDER BY Games_Published DESC
LIMIT 10'
sqldf(num_pub)
```

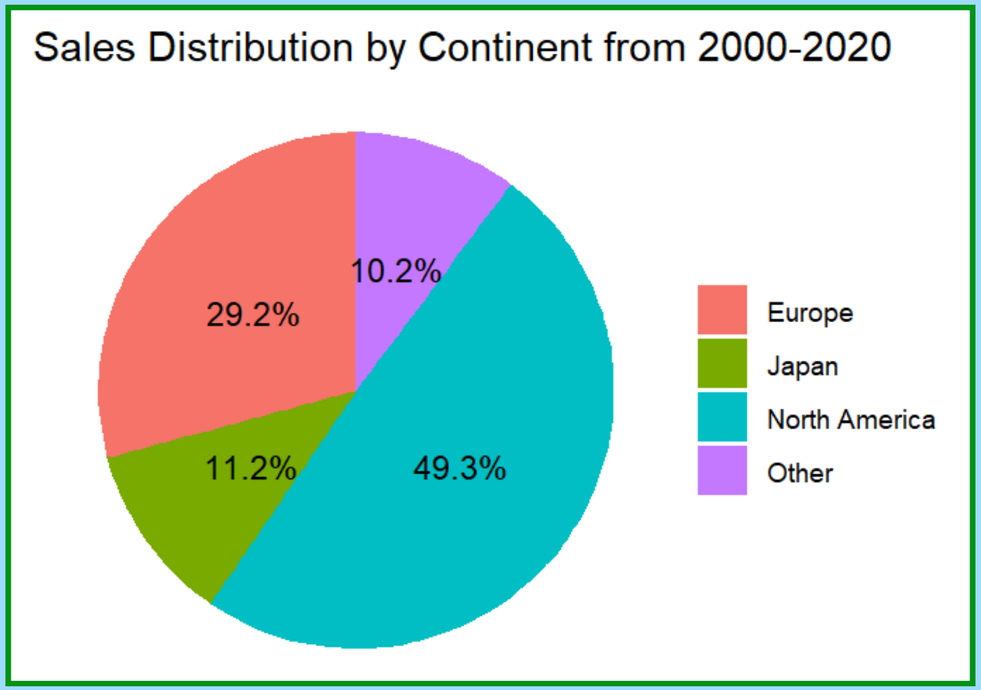
FIGURE 11




```
continent_totalsales<- "SELECT CASE
WHEN Continent = 'NA_Sales' THEN 'North America'
WHEN Continent = 'EU_Sales' THEN 'Europe'
WHEN Continent = 'JP_Sales' THEN 'Japan'
WHEN Continent = 'Other_Sales' THEN 'Other'
ELSE Continent
END AS Continent,
SUM(Sales) AS Total_Sales
FROM continent_sales
GROUP BY Continent
ORDER BY Total_Sales DESC;"
sqldf(continent_totalsales)
```

```
       Continent Total_Sales
1 North America     3581.18
2         Europe     2120.06
3          Japan      816.20
4          Other      743.20
>
```

FIGURE 12 & 13

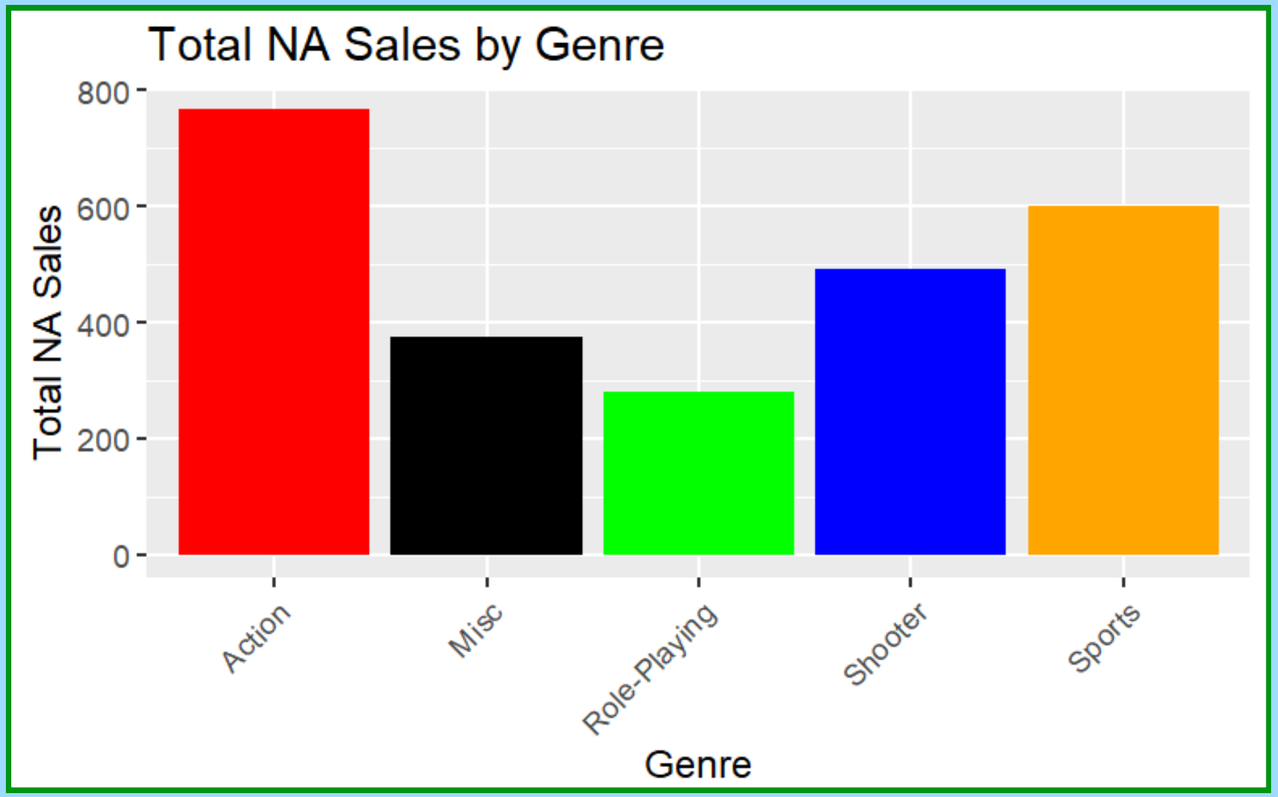



```
90
91  #sql each contentage percentage make-up of global sales
92  Query33<- 'SELECT Continent, percentage
93  FROM continent_sales
94  ORDER BY percentage DESC'
95  sqldf(Query33)
96
```

WHAT DOES THIS TELL US?

```
     Continent percentage
1     NA_Sales   49.32320
2     EU_Sales   29.19935
3     JP_Sales   11.24143
4 Other_Sales   10.23601
>
```

FIGURE 14

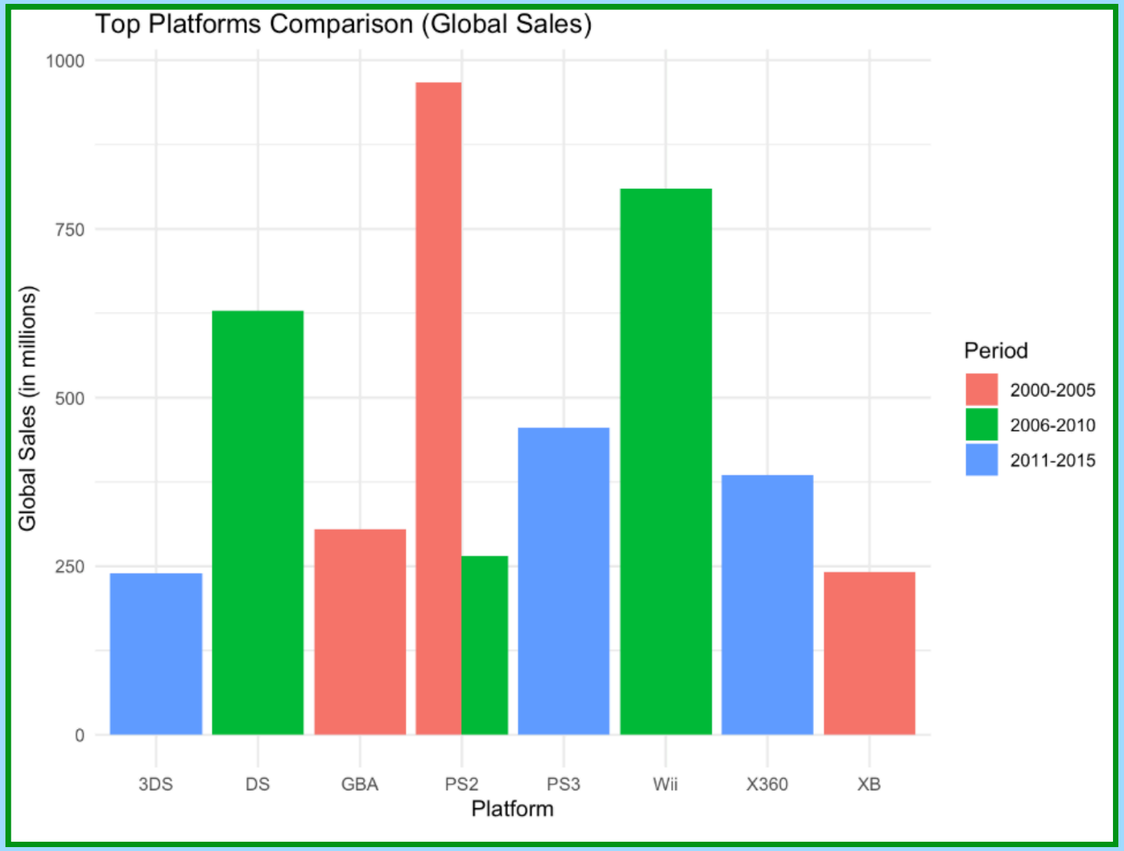


WHAT DOES THIS TELL US?

Top Genres

1. Action
2. Sports
3. Shooter
4. Misc
5. Role-Playing

FIGURE 15



WHAT DOES THIS TELL US?

Top 3 Performers in each time frame.

- 00-05 (PS2)
- 06-10 (WII)
- 11-15 (PS3)

Sales over all is trending down.

```
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 16 - 18

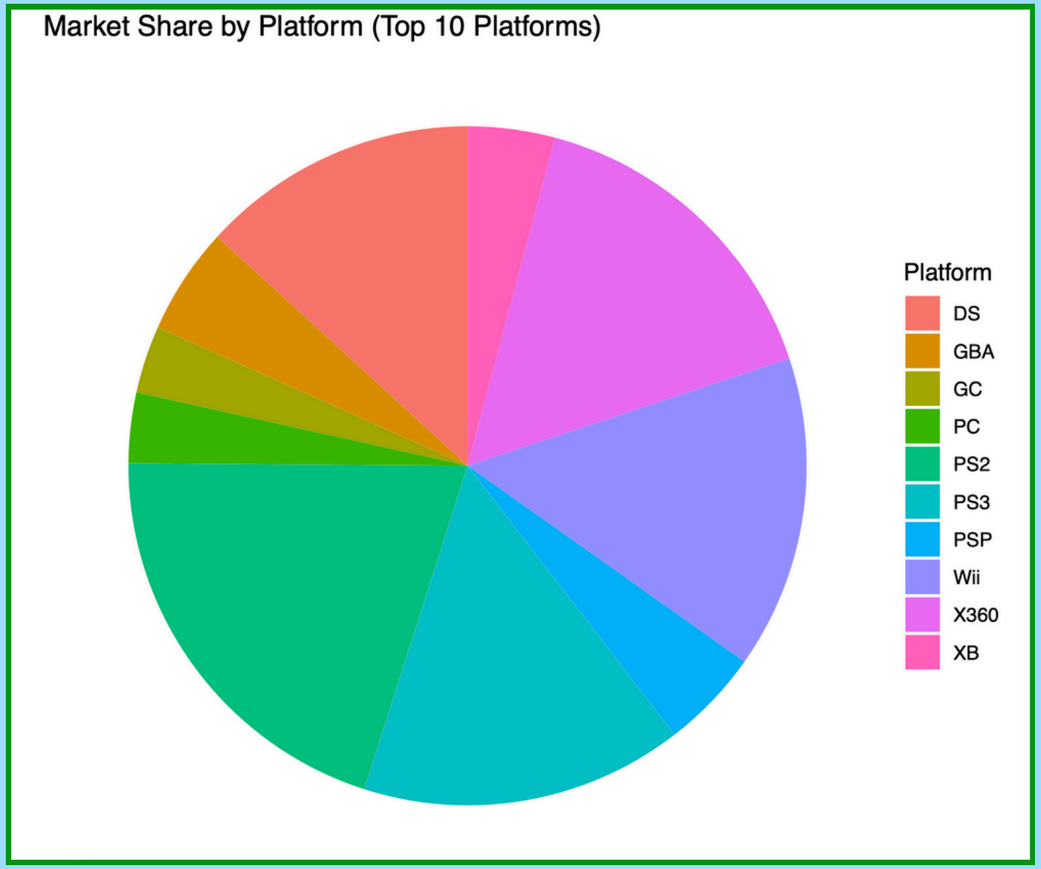
```
> sqldf(Q1.1)
  Platform Total_Sales    Period
1      PS2      967.66 2000-2005
2      GBA      304.78 2000-2005
3       XB      241.21 2000-2005
```

```
> sqldf(Q1.2)
  Platform Total_Sales    Period
1      Wii      809.28 2006-2010
2       DS      628.37 2006-2010
3     X360      575.38 2006-2010
```

```
> sqldf(Q1.3)
  Platform Total_Sales    Period
1      PS3      455.43 2011-2015
2     X360      385.08 2011-2015
3      3DS      239.68 2011-2015
```

Why is this trend of Global Sales decreasing?

FIGURE 13

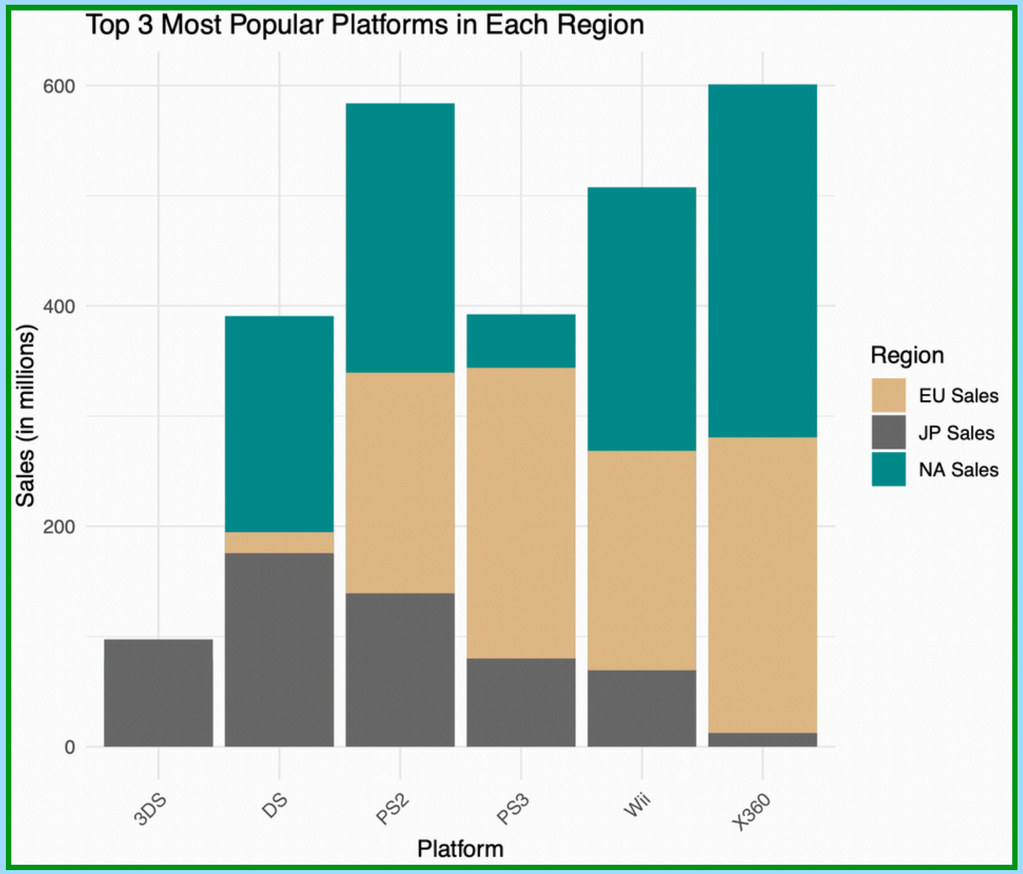


WHAT DOES THIS TELL US?

- Here we can see that the market share is dominated by 5 main platforms.
- Some of these platforms offer very similar experiences XBOX360 and PS2 & PS3 it could come down to personal preference.

```
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



WHAT DOES THIS TELL US?

- Here, we can see that within the top three platforms in each location, NA dominated EU and JP.
- Aswell the leading Platform varies depending on location.

```
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" = "darkcyan", "EU Sales" = "burlywood", "JP Sales" = "dimgray")) +
  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 - 23

```
> sqldf(Q2.1)
   Platform NA_Sales
1      X360   601.05
2       PS2   583.84
3       Wii   507.71
4       PS3   392.26
5        DS   390.71
6       GBA   187.54
7        XB   186.69
8        GC   133.46
9       PSP   108.99
10      PS4    96.80
```

```
> sqldf(Q2.2)
   Platform EU_Sales
1       PS3   343.71
2       PS2   339.29
3      X360   280.58
4       Wii   268.38
5        DS   194.65
6       PS4   123.70
7        PC   120.65
8       GBA    75.25
9       PSP    68.25
10       XB    60.95
```

```
> sqldf(Q2.3)
   Platform JP_Sales
1        DS   175.55
2       PS2   139.20
3       3DS    97.35
4       PS3    79.99
5       PSP    76.79
6       Wii    69.35
7       GBA    47.33
8        GC    21.58
9       PSV    20.96
10       PS    20.14
```

FIGURE 24

WHAT DOES THIS TELL US?

- With this graph, I was able to directly compare the top 10 platform sales in EU and NA which had more sales than JP
- NA Sales dominated in almost every category

```
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

WHAT DOES THIS TELL US?

```
> 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
```

- NA Sales were highest for Xbox 360 closely followed by PS2 however the sales dramatically began to decrease for other platforms

CONCLUSION

01 Action genre leads the industry in terms of global sales

02 XBOX360, PS2 and DS are global dominating platforms

03 North America consumes the most amount of video games

04 How does this affect the Industry?


THANK
YOU