Video Game Sales
Exploratory Analysis

Caitlin Sizemore, csizemore@bellarmine.edu
Eli Dalton, edalton2@bellarmine.edu

I. INTRODUCTION

This dataset is a list of video games. Video games were chosen that had sales greater than 100,000 copies. This dataset was found on kaggle. We chose this dataset because we thought it would be interesting to work with, it was organized well, and it had a lot of data entries with a reasonable number of variables to work with.

II. DATA SET DESCRIPTION

This dataset contains 16,600 samples with initially 11 columns with various data types. These columns are: Rank, Name, Platform, Year, Genre, Publisher, North America Sales, European Union Sales, Japan Sales, Other Sales, and Global Sales. We dropped the Rank column deeming it unnecessary. Also we limited our data exploration to the top 50 video games in order to better visualize the data. A complete listing is shown in **Table 1**.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Name	Nominal/object	0%
Platform	Nominal/object	0%
Year	Interval/float64	0%
Genre	Nominal/object	0%
Publisher	Nominal/object	0%
NA_Sales	Ratio/float64	0%
EU_Sales	Ratio/float64	0%
JP_Sales	Ratio/float64	0%
Other_Sales	Ratio/float64	0%
Global_Sales	Ratio/float64	0%

III. Data Set Summary Statistics

In this summary statistics section, we show the basic statistical measurements for the numerical columns. We are also showing what useful statistical information that we can from the categorical variables. Also, for the Year variable we filled in the meaningful statistical information, but marked the rest N/S as some of the standard statistical measurements do not provide useful information for dates.

Table 2: Summary Statistics for Video Game Sales

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Year	50	N/A	N/A	1984	2002	2006	2010	2015
NA_Sales	50	9.83	7.16	2.55	5.80	8.33	10.58	41.49
EU_Sales	50	5.69	4.37	0.01	3.46	4.52	6.84	29.02
JP_Sales	50	2.81	2.37	0.06	0.47	2.80	4.18	10.22
Other_Sales	50	1.89	2.00	0.23	0.78	1.30	2.13	10.57
Global_Sales	50	20.22	11.58	11.33	13.47	16.00	23.01	82.74

There should be a table for **EACH** categorical variable.

Table 3: Proportions for Name

Category	*Frequency*	*Proportion (%)*
Grand Theft Auto V	3	6%
Call of Duty: Black Ops II	2	4%
Call of Duty: Modern Warfare 3	2	4%
Call of Duty: Black Ops	2	4%

Wii Sports	1	2%
Pokémon Black/Pokémon White	1	2%
Brain Age 2: More Training in Minutes a Day	1	2%
Gran Turismo 3: A-Spec	1	2%
Pokémon Yellow: Special Pikachu Edition	1	2%
Pokémon X/Pokémon Y	1	2%
Call of Duty: Black Ops 3	1	2%
Call of Duty: Modern Warfare 2	1	2%
Grand Theft Auto: Vice City	1	2%
Grand Theft Auto III	1	2%
Super Smash Bros. Brawl	1	2%
Animal Crossing: Wild World	1	2%
Mario Kart 7	1	2%
Halo 3	1	2%
Pokémon Heartgold/Pokémon SoulSilver	1	2%
Super Mario 64	1	2%
Gran Turismo 4	1	2%
Super Mario Galaxy	1	2%
Pokémon Ruby/Pokémon Sapphire	1	2%
Super Mario Bros. 3	1	2%
Super Mario Bros.	1	2%
Nintendogs	1	2%
Mario Kart Wii	1	2%
Wii Sports Resort	1	2%
Pokémon Red/Pokémon Blue	1	2%
Tetris	1	2%
New Super Mario Bros.	1	2%
Wii Play	1	2%
New Super Mario Bros. Wii	1	2%
Duck Hunt	1	2%
Mario Kart DS	1	2%
Super Mario Land	1	2%
Pokémon Gold/Pokémon Silver	1	2%
Wii Fit	1	2%
Wii Fit Plus	1	2%
Kinect Adventures!	1	2%
Grand Theft Auto: San Andreas	1	2%
Super Mario World	1	2%
Brain Age: Train Your Brain in Minutes a Day	1	2%
Pokémon Diamon/Pokémon Pearl	1	2%
Pokémon Omega Ruby/Pokémon Alpha Sapphire	1	2%

Table 4: Proportions for Platform

Category	*Frequency*	*Proportion (%)*
Wii	9	18%
DS	9	18%
X360	7	14%
GB	5	10%
PS2	5	10%
PS3	4	8%
NES	3	6%
3DS	3	6%
PS4	2	4%
SNES	1	2%

GBA	1	2%
N64	1	2%

Table 5: Proportions for Publisher

Category	*Frequency*	*Proportion (%)*
Nintendo	32	64%
Activision	8	16%
Take-Two Interactive	6	12%
Microsoft Game Studios	2	4%
Sony Computer Entertainment	2	4%

Table 6: Correlation Table/Tables

	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Year	1.00	-0.44	0.20	-0.35	0.2	-0.23
NA_Sales	-0.44	1.00	0.52	0.15	0.25	0.89
EU_Sales	0.20	0.52	1.00	0.19	0.40	0.81
JP_Sales	-0.35	0.15	0.19	1.00	-0.18	0.34
Other_Sales	0.20	0.25	0.40	-0.18	1.00	0.44
Global_Sales	-0.23	0.89	0.81	0.34	0.44	1.00

IV. DATA SET GRAPHICAL EXPLORATION

In this section we performed a graphical exploration of our dataset. This included a variety of different graphs from which we are able to learn more information from the data.

A. Distributions

In Figure 1 our distribution shows that the majority of the video games in our dataset were put out between 2005 and 2010. There is also an interesting point between 1990 and 1995 where no video games made it big enough to breach the top 50.

B. ScatterPlots / Pairwise Plots (continuous variables)

In Figure 5 we have our pair-plot of the data. Where each variable is plotted against itself we have distributions and the rest of scatter plots. Overall, there is not a lot of correlation, but there is a little with some which makes sense when sales are concerned because when sales are up in one country or region they are likely up in other countries too. In Figure 17 we have a scatter plot of global sales plotted by platform with a legend indicating the publisher, and Figure 18 shows the same except with North America sales. These 2 plots are interesting because they highlight what publisher's games were on which platforms. For example, the Microsoft Game Studios' games were only on the X-box 360.

C. *Barcharts (categorical variables)*

We used multiple bar charts to help graph the categorical data. In Figure 2 we learn about how many video games of the top 50 were in each genre. In Figure 3 we see how many video games were on each platform. In Figure 4 we see how many video games each publisher produced, learning that Nintendo was by far the most prolific among the top 50 video games by sales. With Figure 10 we can easily see the global sales of each publisher, discovering that Nintendo had the most. In Figure 11 we have global sales plotted by platform with Wii having the highest. Figures 12 and 13 are the same as the last two except that they are for North America sales instead of the whole world which shows an interesting difference. Microsoft Game Studios sneaks ahead of Nintendo in North America.

D. *Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.*

To help learn more about the data we plotted a few other graphs as well. In Figure 6 we created a line graph of North America sales by year and genre; then in Figure 7 we graphed North America sales against global sales. In Figures 8 and 9 we graphed the same as the last two except by platform instead of genre. In Figures 7 and 9 we see the expected correlation between global and North America sales. Another line plot, Figure 14, shows North America sales vs global sales by the game names which is very useful for seeing what specific games sold more in North America verses the rest of the world. Finally, figures 15 and 16 are scatter plots of global sales by platform and North America sales by platform respectively. These plots help illustrate some of the statistical measures laid our earlier in the analysis.

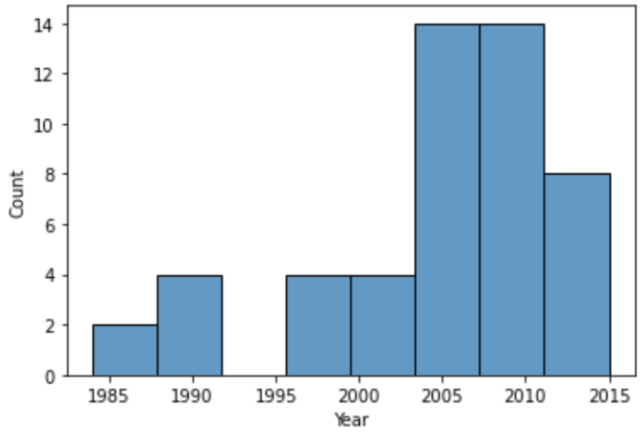

Figure 1: Distribution of video games by year

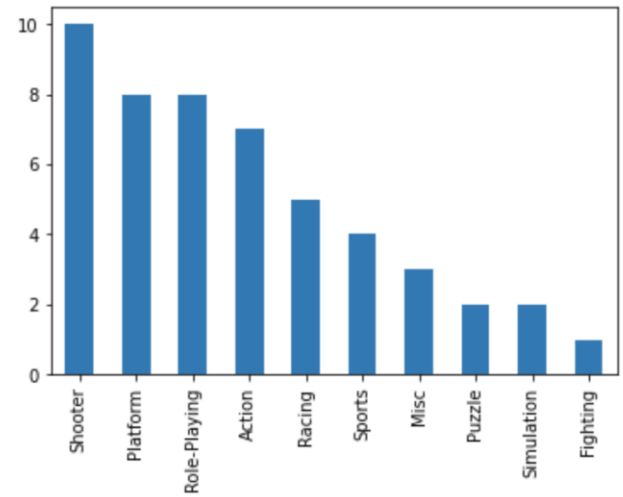

Figure 2: Bar chart of video game genre

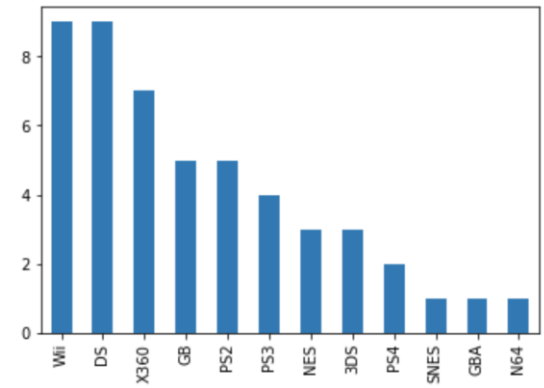

Figure 3: Bar chart of platform

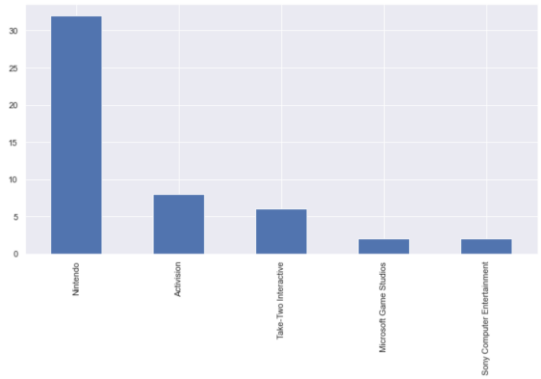

Figure 4: Bar chart of publisher

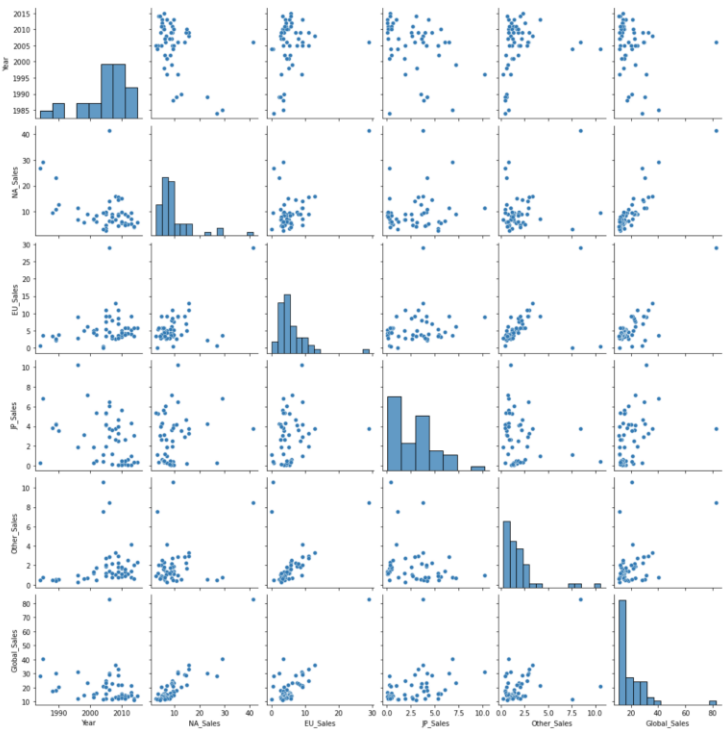

Figure 5: Pair-plot of data

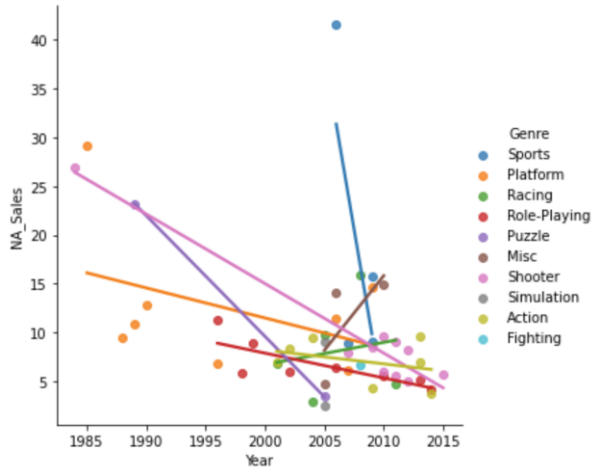

Figure 6: North America sales by year and genre

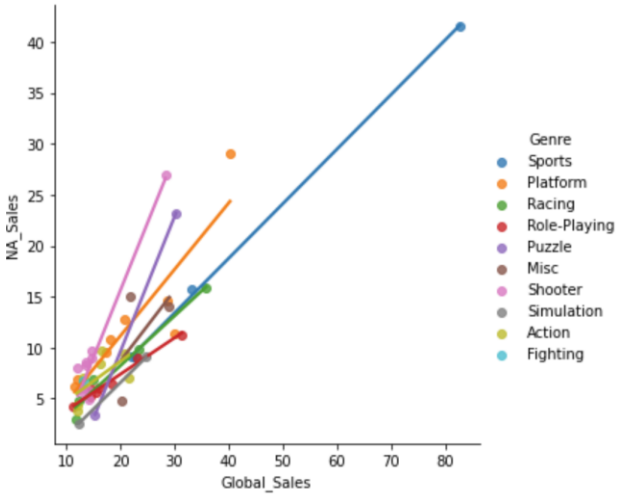

Figure 7: North America vs global sales by genre

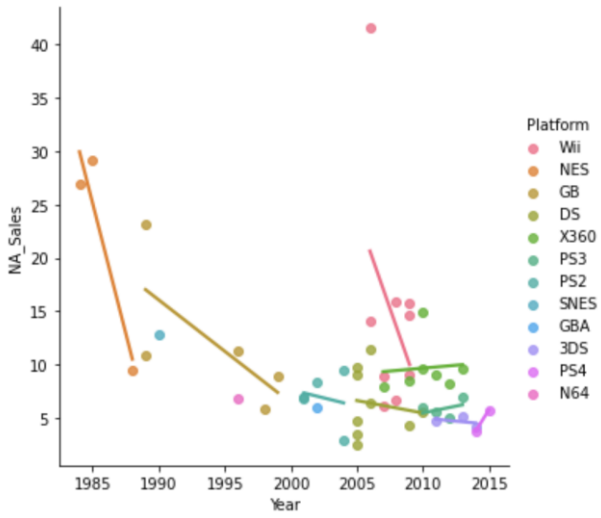

Figure 8: North America sales by year and platform

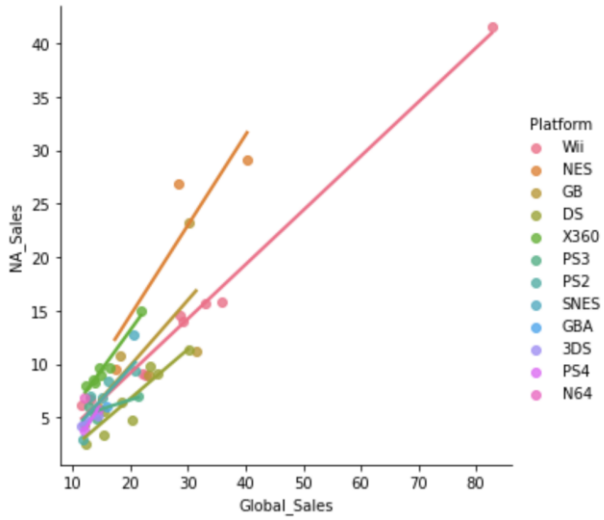

Figure 9: North America sales vs global sales by platform

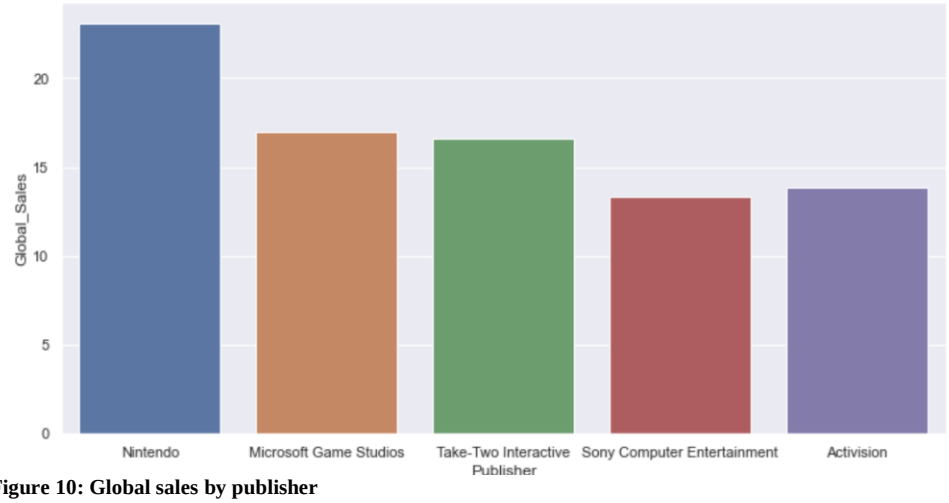

Figure 10: Global sales by publisher

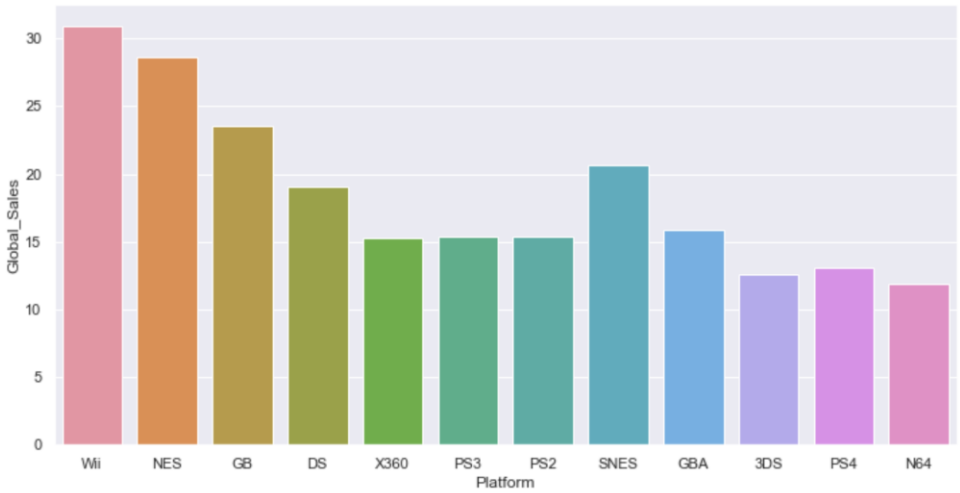

Figure 11: Global sales by platform

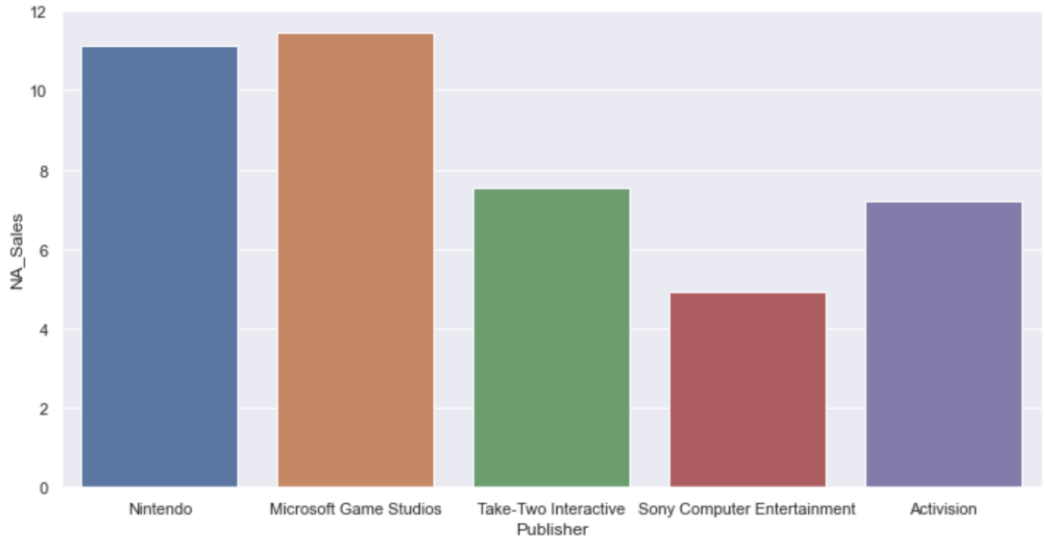

Figure 12: North America sales by publisher

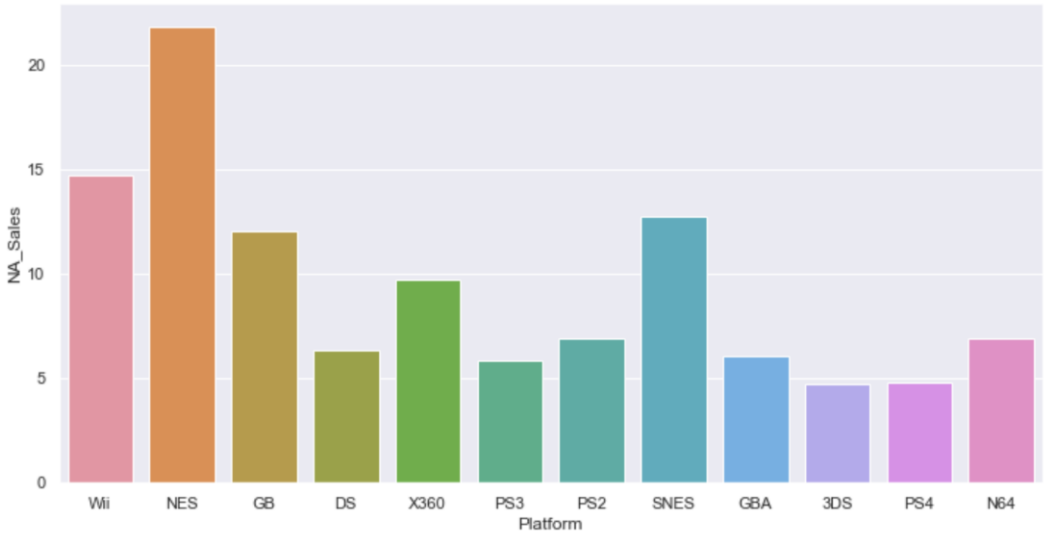

Figure 13: North America sales by platform

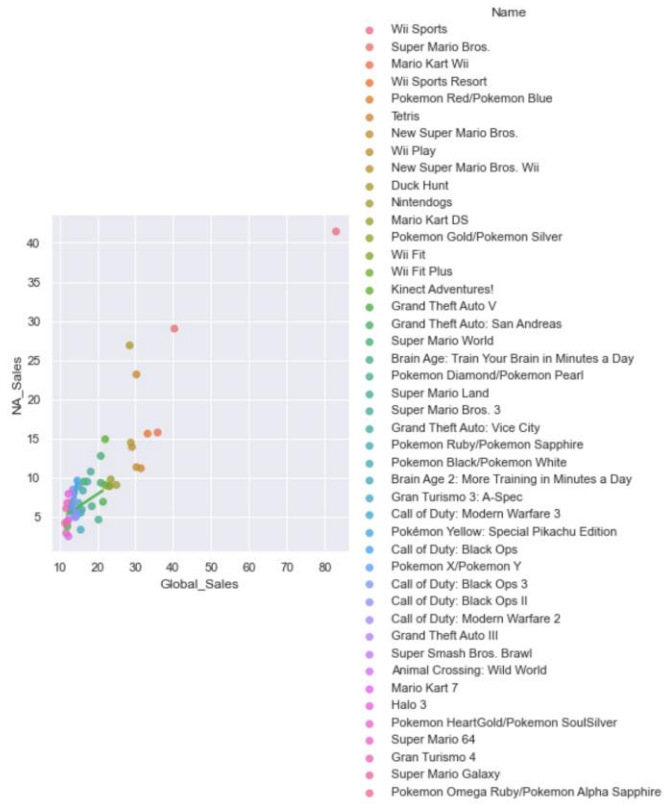

Figure 14: North America sales vs Global sales by video game name

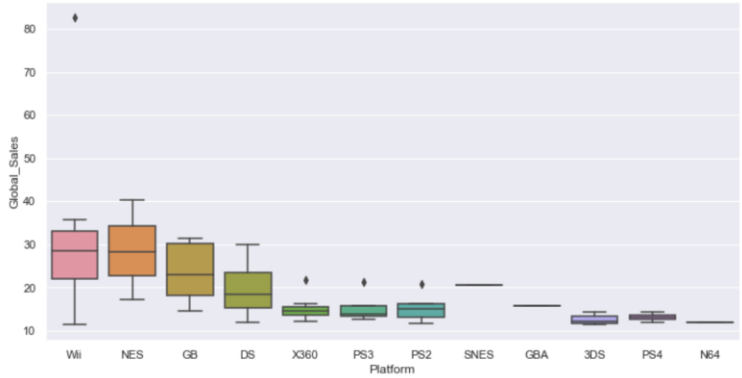

Figure 15: Box plot of global sales by platform

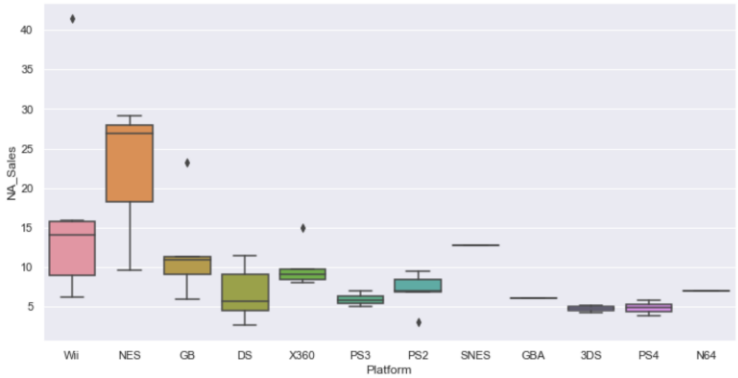

Figure 16: Box plot of North America sales by platform

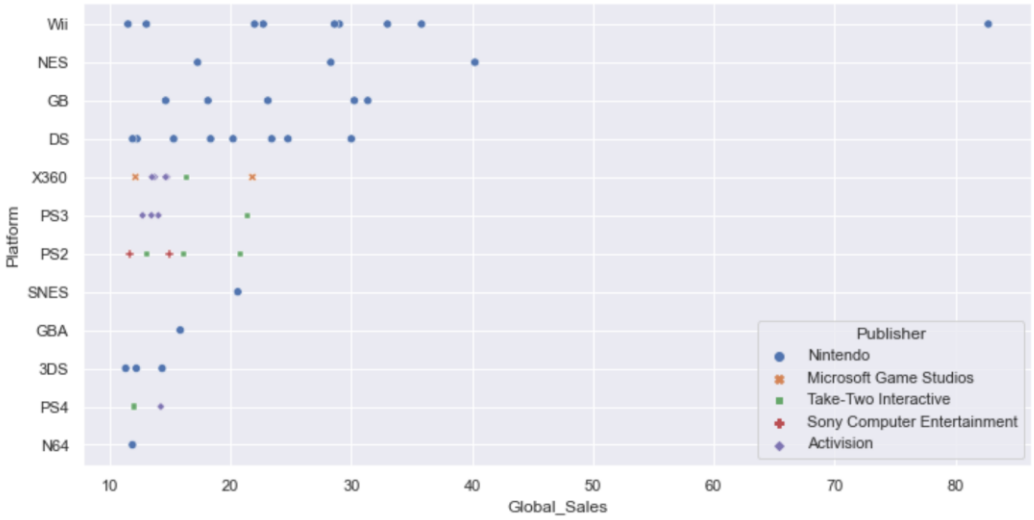

Figure 17: Scatter plot of global sales by platform and by publisher

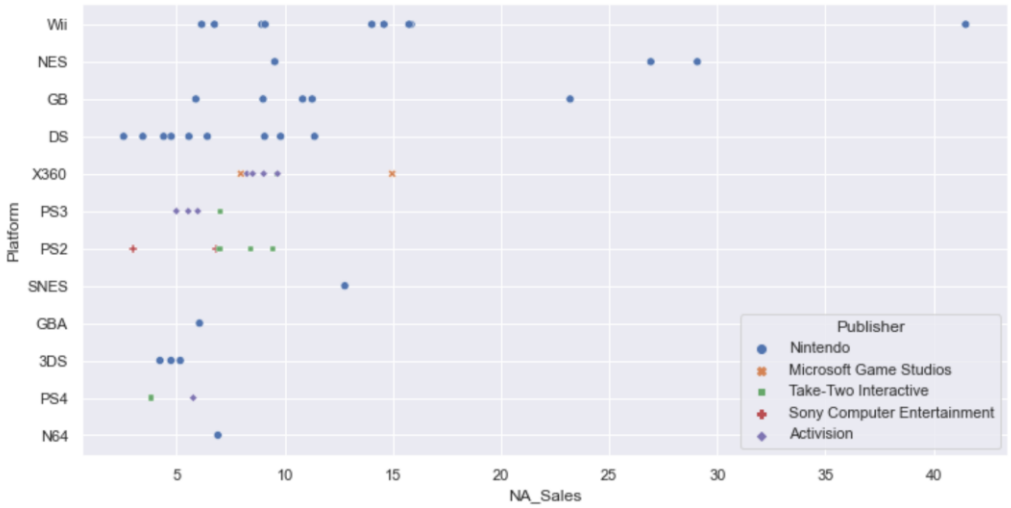

Figure 18: Scatter plot of North America sales by platform and by publisher

V. SUMMARY OF FINDINGS

This video game sales dataset was a very interesting dataset to explore. Through our statistical and graphical explorations of the data we were able to learn of lot of cool information. During the timeframe covered by this data, we learned that North America had the highest sales on average compared to the other regions used in this dataset. We saw that Nintendo was the most prolific creator of top selling video games during this period as well. Through our distribution graph we learned that the majority of the top 50 selling video games were released between 2005 and 2010. We learned all of this information and much more from our exploratory analysis.