Video Game Sales

By: Matthew Cooper

This project is focusing on a dataset containing information on video game sales from 1980 to 2020. I wanted to do this project in Excel because I did not have much prior exposure to Pivot Charts and formulas, and I really wanted to try to become more comfortable with them.

Dataset Details:

Dataset source: https://www.kaggle.com/datasets/gregorut/videogamesales

Column Names:

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (i.e. PC,PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales

Number of data rows: 16,598

The dataset was generated by a scrape of vgchartz.com by the Kaggle user GregorySmith https://www.kaggle.com/gregorut/datasets

The Project:

After opening the file, I noticed the data was already organized by rank. We'll take a quick look at the top ten games with the overall sales. Interestingly enough, all 10 games are from the publisherNintendo.

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31

What I'd like to look at in this dataset are the sales by platform/publisher/genre by year. Since we already looked at the ranking, the column can be removed.

The games in this dataset are listed individually, for each platform they were released on. So, many game titles appear more than once depending on the number of platforms the game had been released on.

[Prototype 2]	X360	2012
[Prototype 2]	PS3	2012
[Prototype 2]	PC	2012
[Prototype]	X360	2009
[Prototype]	PS3	2009
007 Racing	PS	2000
007: Quantum of Solace	X360	2008
007: Quantum of Solace	PS3	2008
007: Quantum of Solace	Wii	2008
007: Quantum of Solace	PS2	2008
007: Quantum of Solace	DS	2008
007: Quantum of Solace	PC	2008
007: The World is not Enough	N64	2000
007: The World is not Enough	PS	2000

First, I wanted to do some error checking. I looked at #N/A, N/A, Blanks, #DIV/0!, #NAME?, and #REF! The only problems found were N/A text entries in the Year and Publisher columns.

Error Checking

	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Formulas Used
#N/A	0	0	0	0	0	0	0	0	0	0	COUNTIF(sheet!range, "#N/A")
N/A	0	0	271	0	58	0	0	0	0	0	COUNTIF(sheet!range, "N/A")
Blanks	0	0	0	0	0	0	0	0	0	0	COUNTBLANK(sheet!range)
Errors	0	0	0	0	0	0	0	0	0	0	SUMPRODUCT(--ISERROR(sheet!range))

I want to get a count of unique values for Name, Platform, Year, Genre, and Publisher. The "Name" column was not cooperating with the functions used. Attempts were made to try other functions, but nothing worked. After some searching, I found a method using PivotTables to find unique values and a total count.

Counts of Unique Values

Name	Platform	Year	Genre	Publisher	Formulas Used
#DIV/0!	31	40	12	579	SUMPRODUCT((data<>"")/COUNTIF(data,data &""))

11493 Unique entries in 'Name' column

Let's take a closer look at the N/A entries in the 'Year' and 'Publisher' columns. Looking back at the error counts, we see that the 'Year' column contains 271 N/A's, and the 'Publisher' column contains 58 N/A's. I'm curious to see if any overlap. Filtering the data by N/A in both columns shows 22 results.

Since I am planning on using the 'Year' and 'Publisher' columns as a major part of my visualizations, it would be best to remove all of the data rows that contain N/A. A total of 307 rows will be removed.

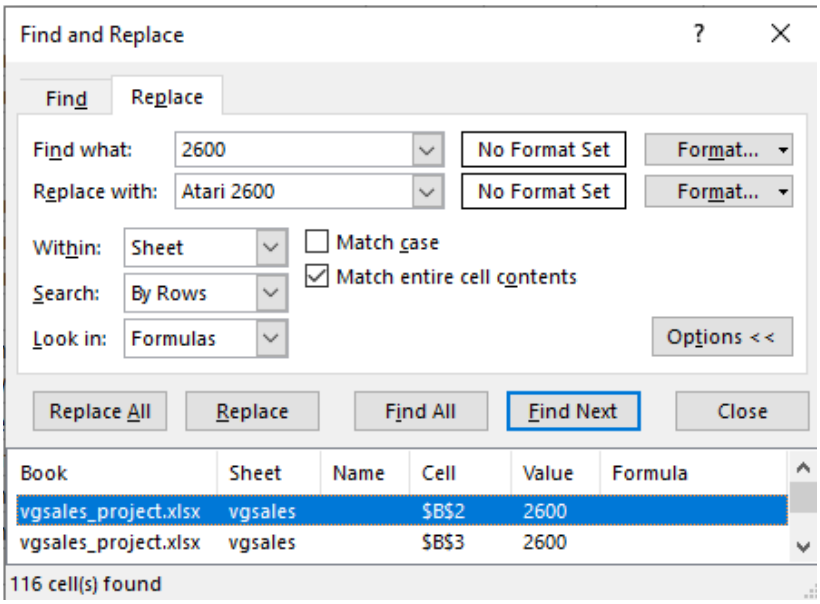
The error checking formulas updated as I removed the rows, and now show no errors remaining.

Error Checking

	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Formulas Used
#N/A	0	0	0	0	0	0	0	0	0	0	COUNTIF(sheet!range, "#N/A")
N/A	0	0	0	0	0	0	0	0	0	0	COUNTIF(sheet!range, "N/A")
Blanks	0	0	0	0	0	0	0	0	0	0	COUNTBLANK(sheet!range)
Errors	0	0	0	0	0	0	0	0	0	0	SUMPRODUCT(--ISERROR(sheet!range))

Returning to the main dataset, there are a few error items to look into. In the 'Platform' column the entry '2600' isn't entirely clear to what platform it is (especially to people who may not be familiar with the video game industry or it's history).

In this case '2600' refers to the Atari 2600 gaming system. So a "Find & Replace" will be used to correct the entries. Since it is the earliest gaming console, sorting by 'Year' will aid in viewing the correction.



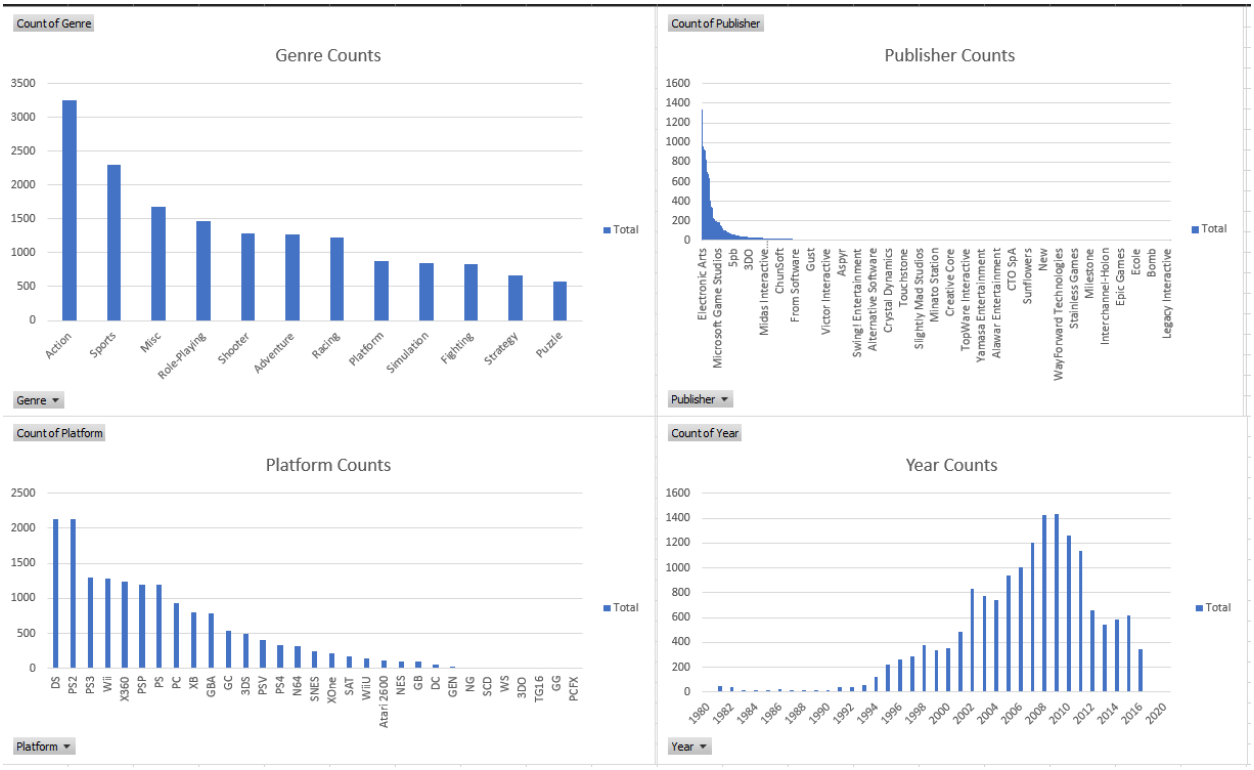
After scrolling through the first couple of years (about 150 rows) another problem appeared. There is a data entry for a game released on the Nintendo DS console in the year 1985. This is incorrect because the DS did not exist until 2004. An attempt was made to locate the correct release year, but was unable to find it. The row will be removed.

I am unsure how to properly determine similar problems. One option I tried was to create a Pivot Table showing 'Platform' with the MIN and MAX of 'Year'. This, at least, allows me to look at the date ranges to see if there are any anomalies. The date ranges look reasonable to the respective platform, so I don't believe there are any other 'Year'/'Platform' issues.

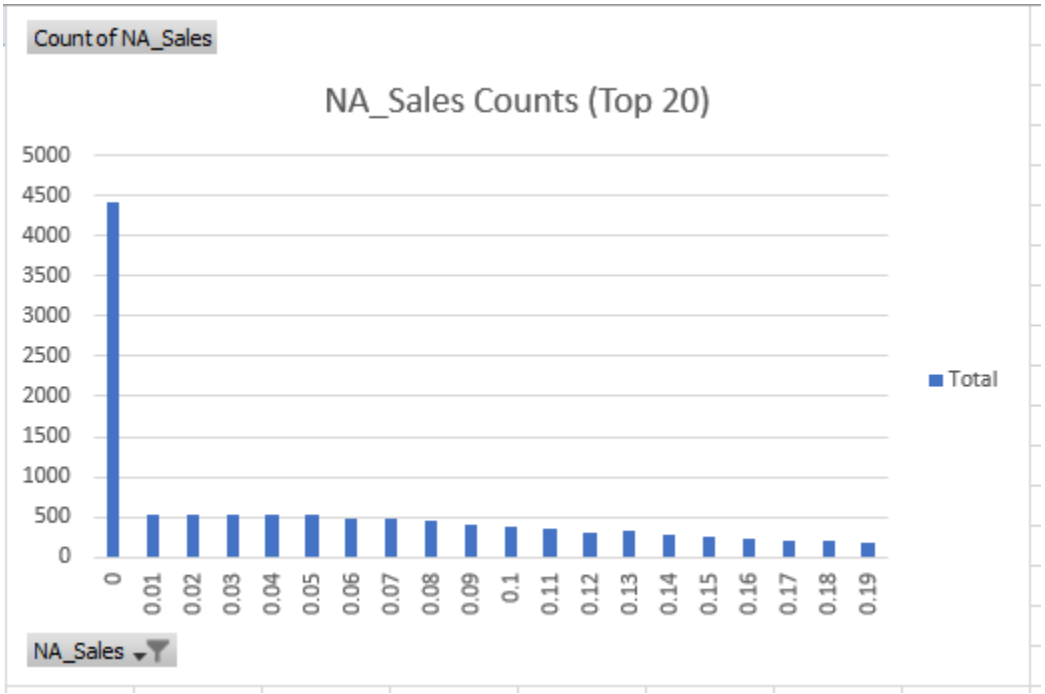
Row Labels	Min of Year	Max of Year
3DO	1994	1995
3DS	2011	2016
Atari 2600	1980	1989
DC	1998	2008
DS	2004	2020
GB	1988	2001
GBA	2000	2007
GC	2001	2007
GEN	1990	1994
GG	1992	1992
N64	1996	2002
NES	1983	1994
NG	1993	1996
PC	1985	2016
PCFX	1996	1996
PS	1994	2003
PS2	2000	2011
PS3	2006	2016
PS4	2013	2017
PSP	2004	2015
PSV	2011	2017
SAT	1994	1999
SCD	1993	1994
SNES	1990	1999
TG16	1995	1995
Wii	2006	2015
WiiU	2012	2016
WS	1999	2001
X360	2005	2016
XB	2000	2008
XOne	2013	2016
Grand Total	**1980**	**2020**

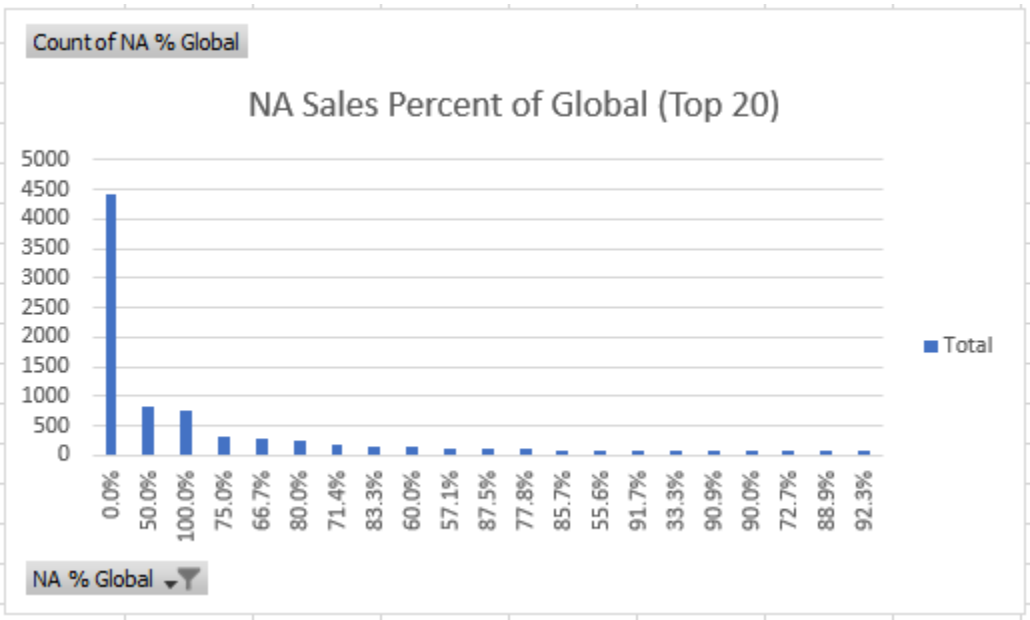
At this point, I'll do some exploratory visualizations to better understand the data. Before that, however, I will remove the JP/EU/Other sales columns in order to focus on the NA_sales. I kept the 'Global_Sales' column in order to create a new column of the percent of NA_Sales of Global_Sales.

I first did graphs of the counts of the Year/Genre/Platform/Publisher.
The distributions are pretty interesting.



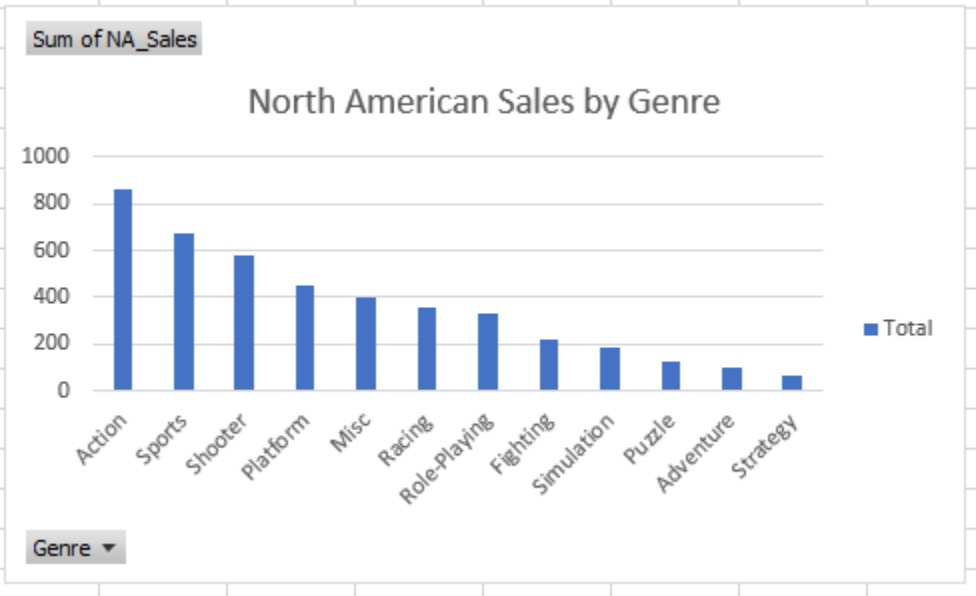
Next, I graphed the NA_Sales and the Percent of NA_Sales of Global_Sales. Becuase of the extreme skewness of both graphs, I limited the results to the top 20 to allow for any useful readability.

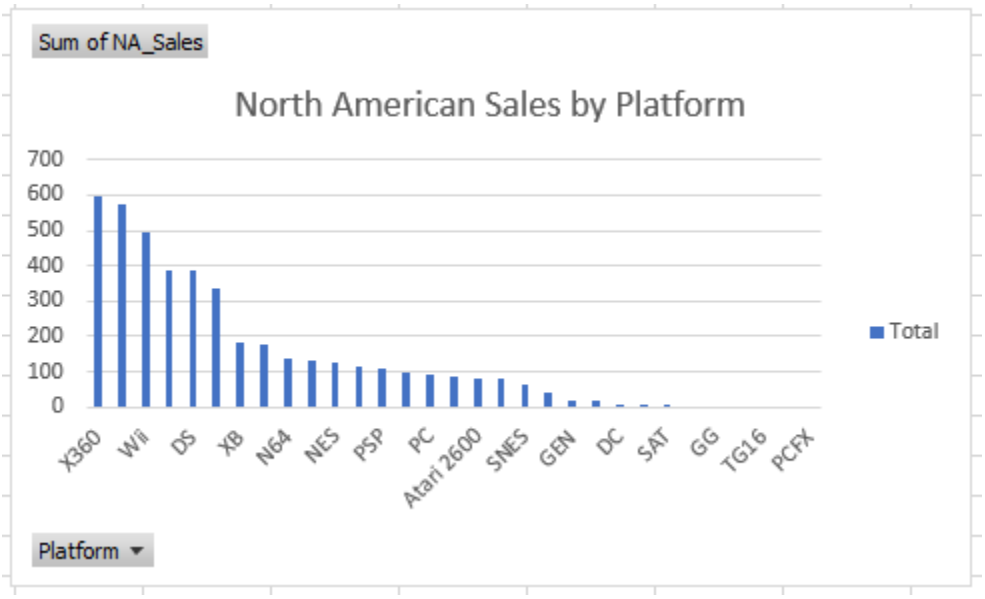




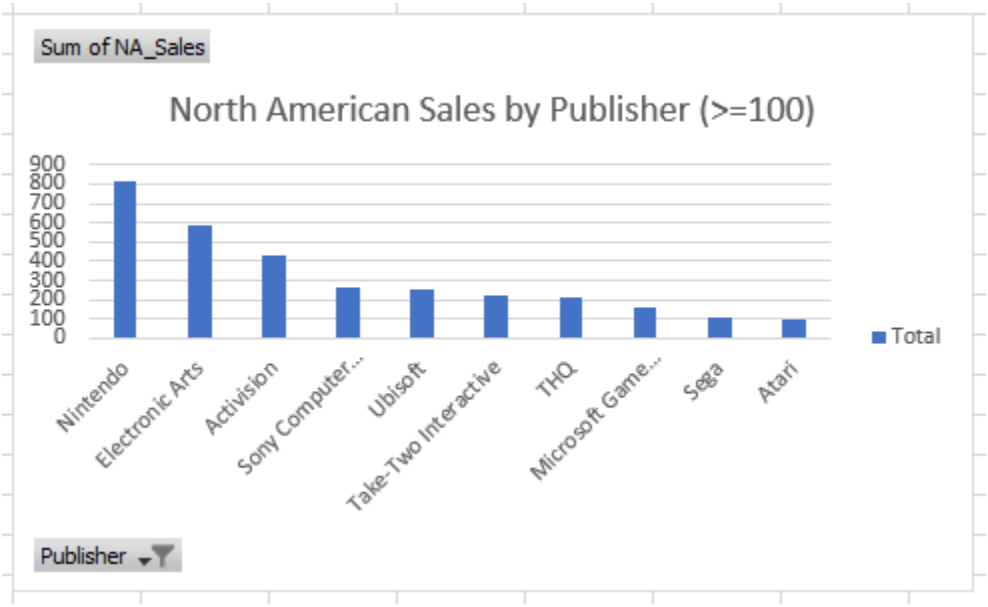
Main Visualizations

Specifically, I'm interested in seeing the Total North American Sales (in $amt) by Genre and by Platform. In the date range covered by this data, I'm very interested to see the highest selling in both areas.





I also wanted to look at the total sales by publisher. Due to the number of publishers in this data (575) the resulting graph was limited to the sales >=100.



I'm interested in seeing which genre by publisher had the most sales. Again, due to the number of publishers, I will limit the results for easier readability.

The final set of Visualizations I'd like to do is to look year-by-year. To do this, I chose to set it up as a dashboard. The previous limit of >=100 on the previous two graphs needed to be changed to 'Top 10' in order to be able to display the data by year. The static image below will be replaced by a video.