Matthew Belizaire & Christopher Fredrick
Professor Landowski
12/11/2022
Final Project Report

Video Games Sales with Ratings

Introduction

In the 1980s, video games became a mainstay within many societies as they provided a form of entertainment that was easily accessible and could be enjoyed by millions. Since then, it has grown exponentially in terms of overall revenue generated, user base, number of developers, and quantity of titles to play. The video game industry as of 2022 is a multi-billion-dollar industry, and its sales come from both physical and digital copies of the games.

There are many different types of video games, including action, adventure, sports, strategy, and simulation games. The variety and range of the genres make it easy to understand why the medium is so popular. Some of the most popular video game franchises include Super Mario, Pokémon, Call of Duty, and The Legend of Zelda. The video game industry is constantly evolving and innovating, with new technologies and platforms allowing for more immersive and engaging experiences for players. As the industry continues to grow and expand, it is likely that video game sales revenue will continue to increase in the coming years.

The Data

The data being analyzed comes from Kaggle, and is titled Video Game Sales With Ratings. The dataset consists of video game titles with over 100,000 sales and was originally generated from a web scrape of VGChartz. VGChartz is a video game sales tracking website. Additionally, a web scrape from Metacritic was added to the data set. Metacritic is a website that aggregates reviews from verified critics and reports the average score from the critics as the game titles' score. This is known as "Critic_Score" within the dataset. The website also utilizes the same process for user scores, noted as the "User_Score" variable within the data.

In total there are 16 variables and 16719 rows, with 11563 of the titles being unique. The reason there are titles with duplicate names is because the same game sometimes comes out on multiple platforms, and the sales for these are calculated separately. Some of the other variables include "Name," "Platform," "Year_of_Release," "Global_Sales," and more, all of which are self-explanatory in terms of their purpose.

Exploratory Data Analysis

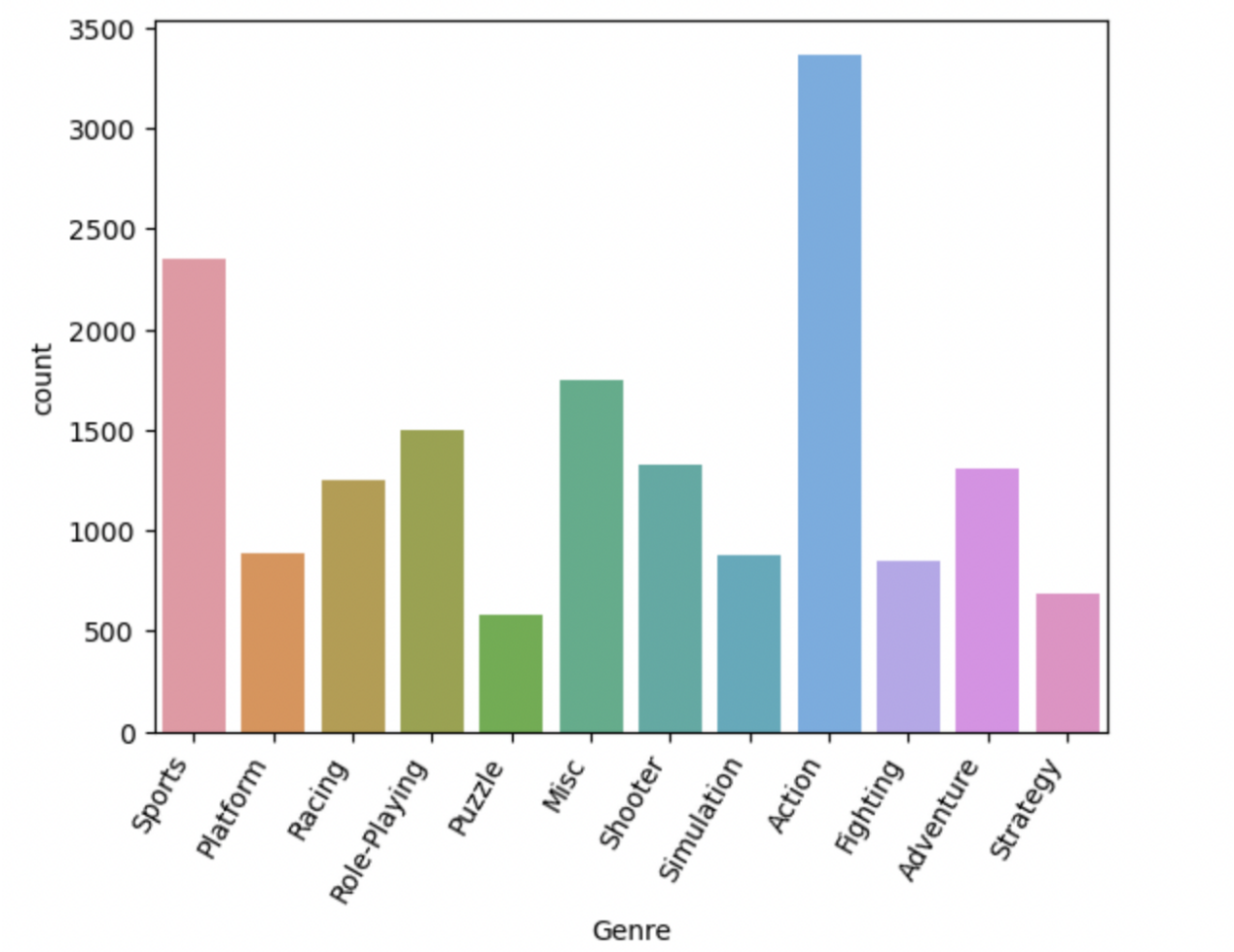
It is important that we understand our data before answering our questions.

```
#Descriptive Statistics
df.describe()
```

	Year_of_Release	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Count
count	16450.000000	16719.000000	16719.000000	16719.000000	16719.000000	16719.000000	8137.000000	8137.000000	7590.000000
mean	2006.487356	0.263330	0.145025	0.077602	0.047332	0.533543	68.967679	26.360821	162.229908
std	5.878995	0.813514	0.503283	0.308818	0.186710	1.547935	13.938165	18.980495	561.282326
min	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000	13.000000	3.000000	4.000000
25%	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000	60.000000	12.000000	10.000000
50%	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000	71.000000	21.000000	24.000000
75%	2010.000000	0.240000	0.110000	0.040000	0.030000	0.470000	79.000000	36.000000	81.000000
max	2020.000000	41.360000	28.960000	10.220000	10.570000	82.530000	98.000000	113.000000	10665.000000

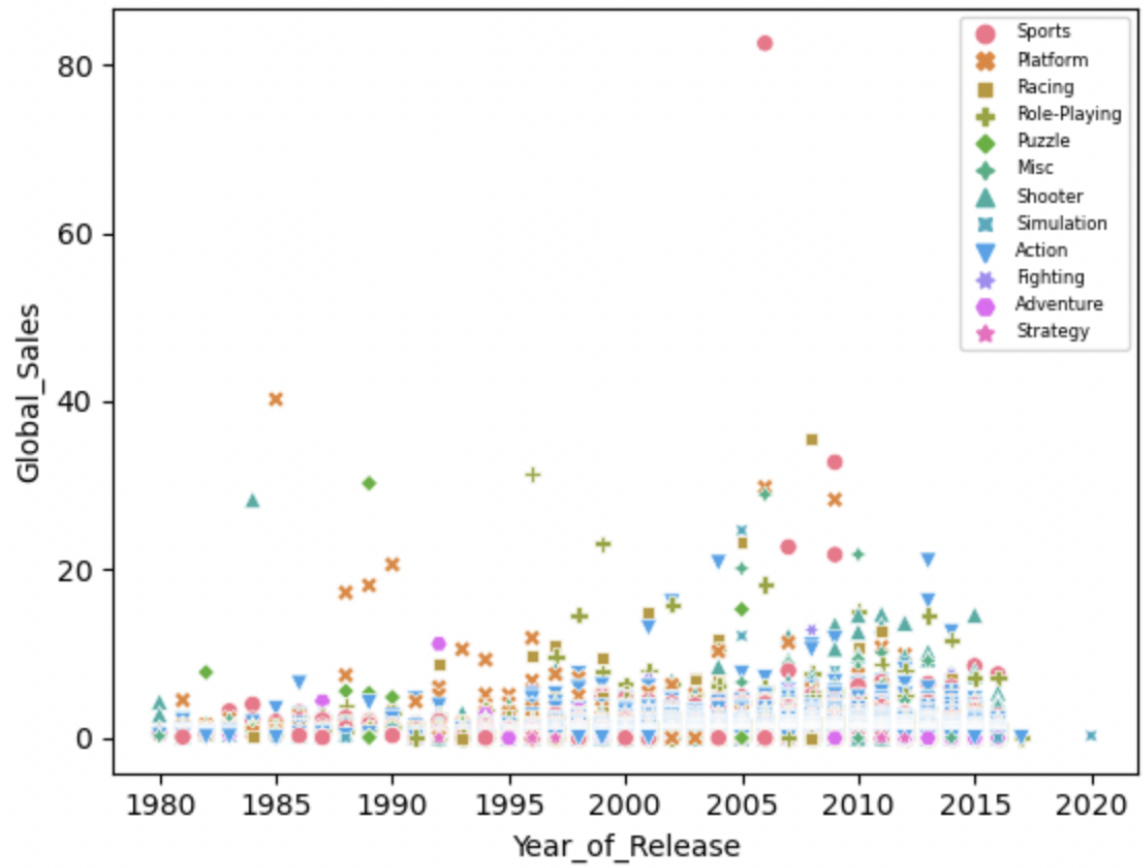
These summary statistics are used to provide a general overview of the data and to identify patterns and trends within the data. It aids in understanding this diverse and in some places complex dataset to make informed decisions based on the results. Something of note here is that the NA_Sales has the highest mean, which can be an early indicator of video games being most popular within North America in comparison to Europe or Japan. The results also show that on average, we can expect the critic score to be composed of an average of about 26 critics, as shown by the mean of Critic_Count. These are just two examples, and other inferences can be made about the data here; but it is important to not be conclusive about anything we see here.

```
#What genre of video games are most popular globally?

#The most popular Genre is Action, followed by Sports, then Misc.
plot1 = sns.countplot(x=df["Genre"])
plot1.set_xticklabels(plot1.get_xticklabels(), rotation=60, ha="right")
plt.show
```



This barplot provides an easily understandable visualization for the distribution of different genres in our data set. From a quick glance, it can be known that Action games are the most common games, followed by Sports, then Misc. This does not mean that they are the highest or best rated games, but simply that they possess the most frequency. With more of anything, the average of something like a User_Score for action game can be lowered in comparison to a genre with a lower volume and higher ratings overall. While this chart provides us with good information, it cannot be conflated with other ideas.

```
#Global sales over time for all video games
plot4 = sns.scatterplot(data = df, x = "Year_of_Release", y = "Global_Sales", hue = "Genre", style = "Genre")

plot4.legend(bbox_to_anchor = (1,1), fontsize = 6)
plt.show()
```



This scatterplot shows the relationship between the Year_of_Release, and Global_Sales. There appears to be little to no correlation between the year a game is released and the sales number for games globally. As video games have steadily gotten more popular since the 1980s, one would presume that the sales of later games would naturally increase with the popularity. A possible deduction from this information can be that while games have gotten more popular over the years, the higher quantity of games does not automatically translate to the quality of them which the sales would be a clear indicator of. Since there are more games overall as well, there are more options to play, thus consumers are spread more thin in comparison to times where only a few popular games stood out among the rest. For example, the outlier “Platform” shape above 1985 can be none other than the original Super Mario Bros., a cultural phenomenon of a game. But let’s confirm this is the case:

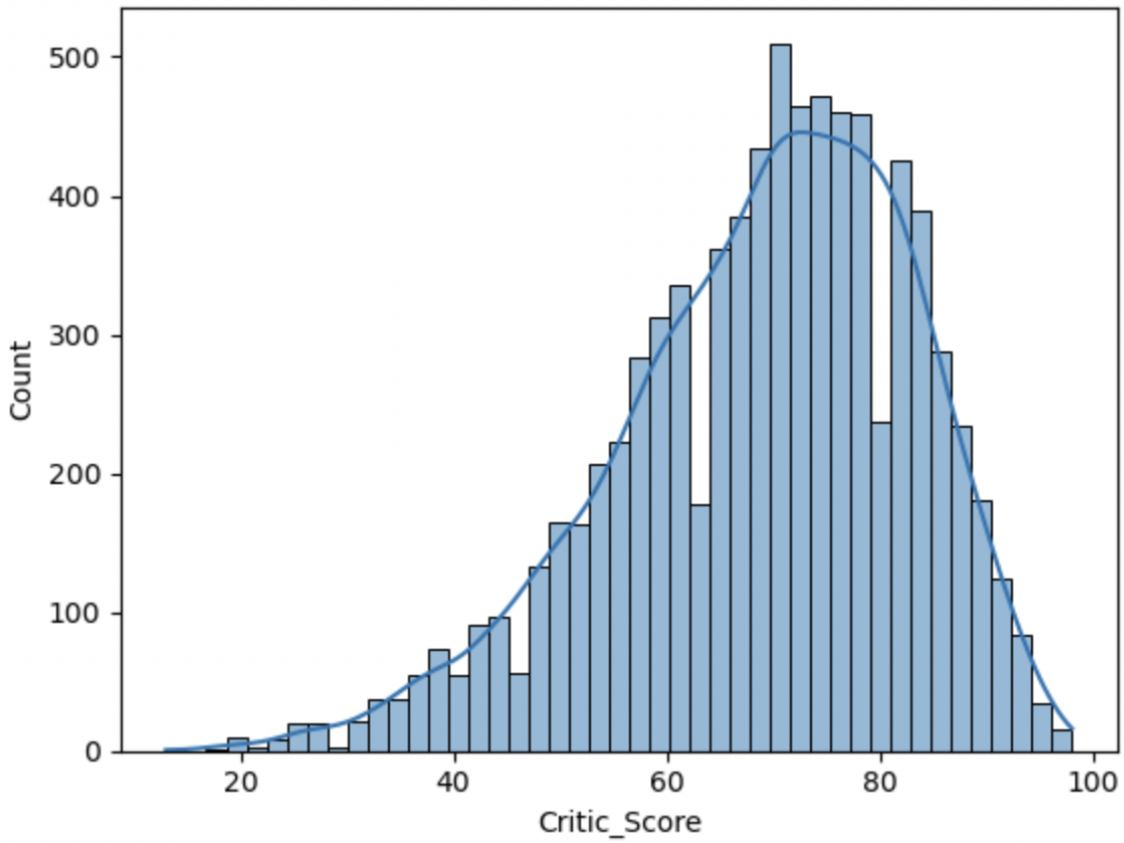
```
df.loc[df['Name'] == 'Super Mario Bros.']
```

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24

It is! Super Mario Bros. came out at a time when there were not as many games, which could have aided in its success since it had a combination in innovation as well as a lack of competition. Today, it would be impossible for the average person to play every exceptional game, thus leading to the sales for said games to never reach the heights of Super Mario Bros. Few games come close though innovation, and one has even surpassed it by packaging the game (Wii Sports) with new Wii consoles which counted for a sale towards the game. But most great games simply will not be purchased by a concentrated number of people because of the sheer amount of options today it seems.

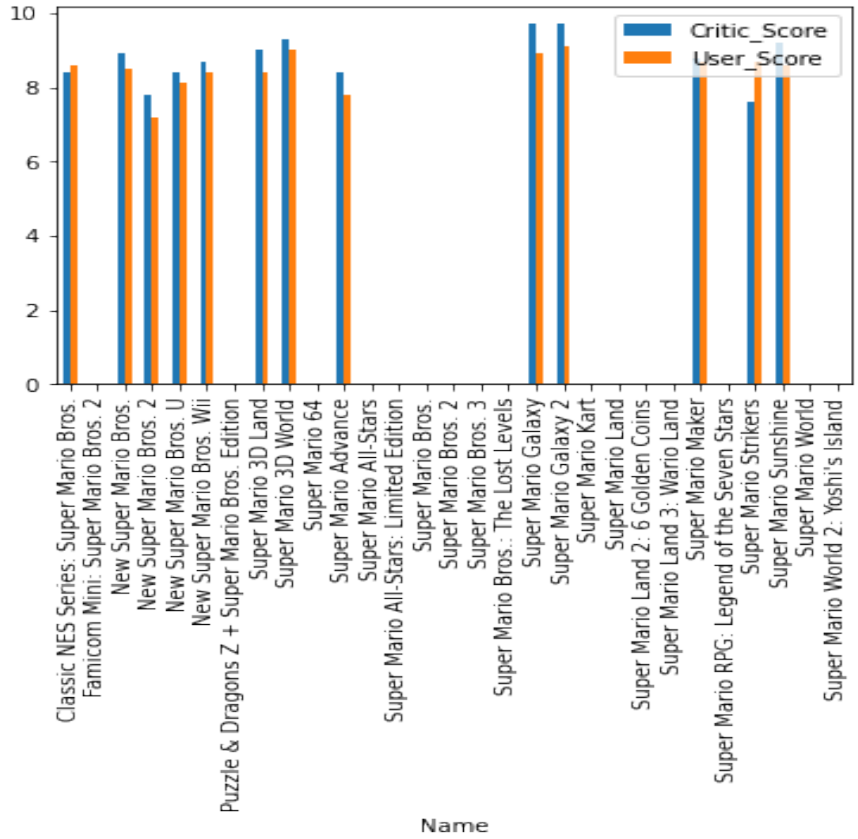
```
#Negative left-skew indicates most of the values are concentrated on the right side of the histogram
sns.histplot(data = df, x = "Critic_Score", kde = True)

#The mean supports the visualization. It appears that games generally rate above average (50).
print(df[['Critic_Score']].mean())
```

```
Critic_Score    68.967679
dtype: float64
```

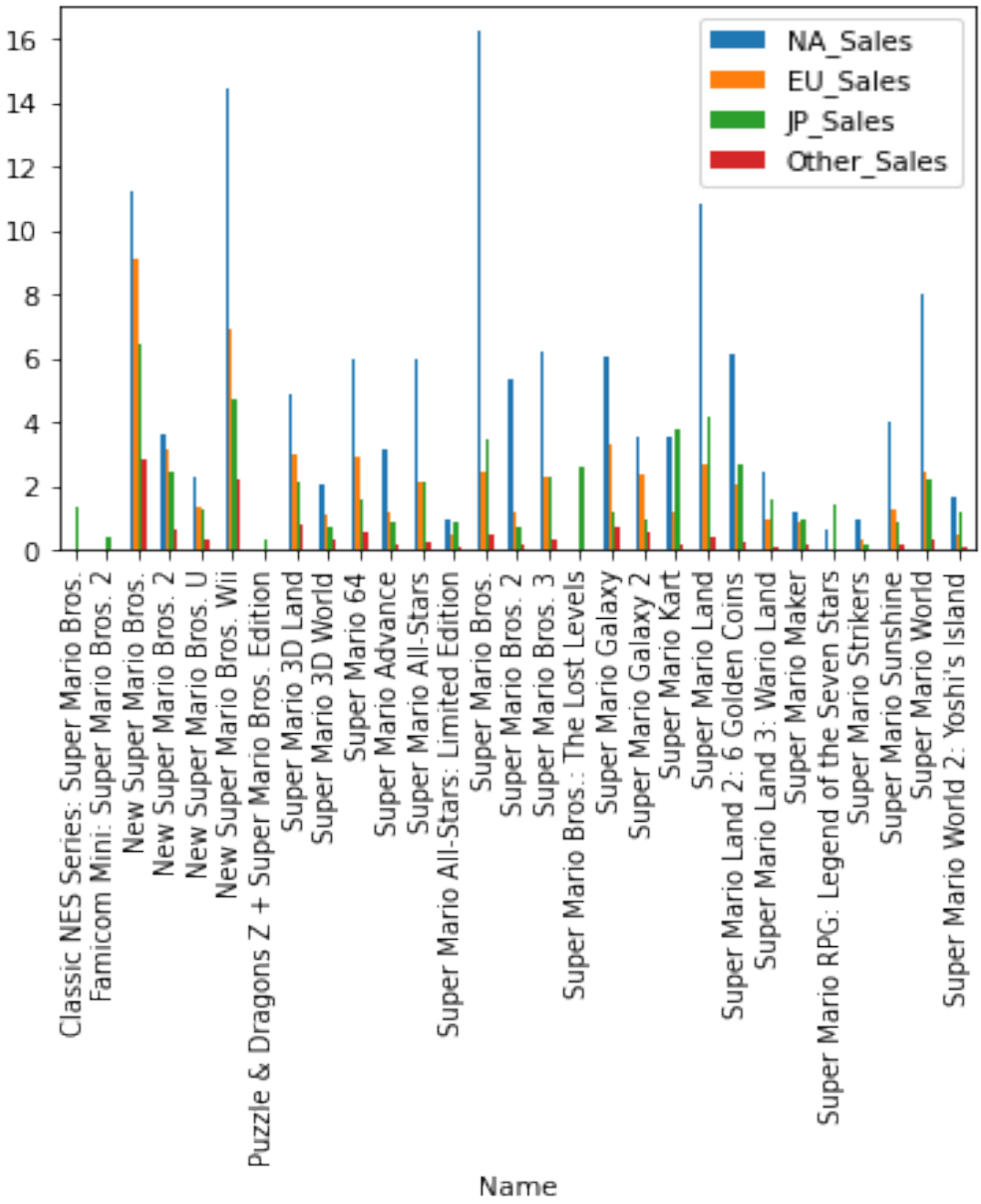


This histogram provides a distribution of Critic_Score separated into bins. The highest bar in this histogram is hovering at around 70, which aligns with the calculated mean of Critic_Score. This is a good indicator that games generally rate above average (50) on a 0 to 100 rating scale. The shape also represents a negative left skew, which means that the majority of the data is concentrated on the right side of the distribution, with a long tail extending to the left. In other words, there are relatively few data points with low values (bad games) and a relatively large number of data points with high values (average and exceptional games). Game developers have done good as a collective throughout the years!

We also decided to take a look at games that were apart of a series and determine which game was the best in terms of sales and scores. The first game was Super Mario.

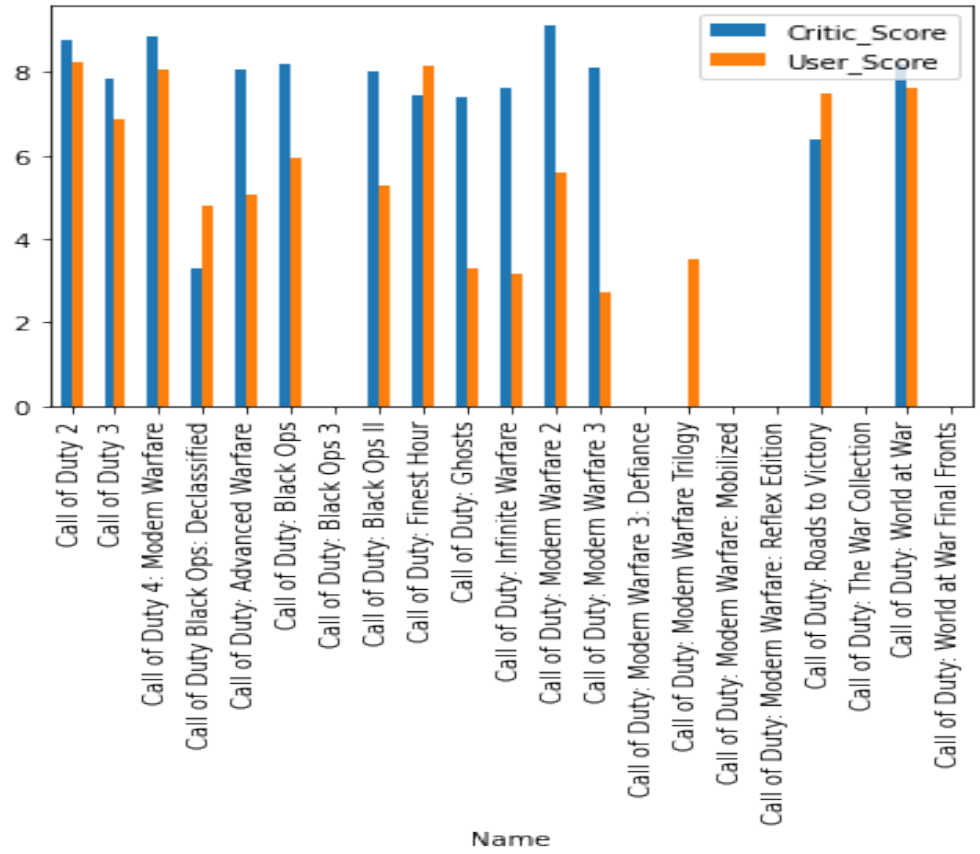


Super Mario Galaxy and Super Mario Galaxy 2 tied for first in critic scores with scores of 9.7. Super Mario Galaxy 2 was first in user scores with a score of 9.1.

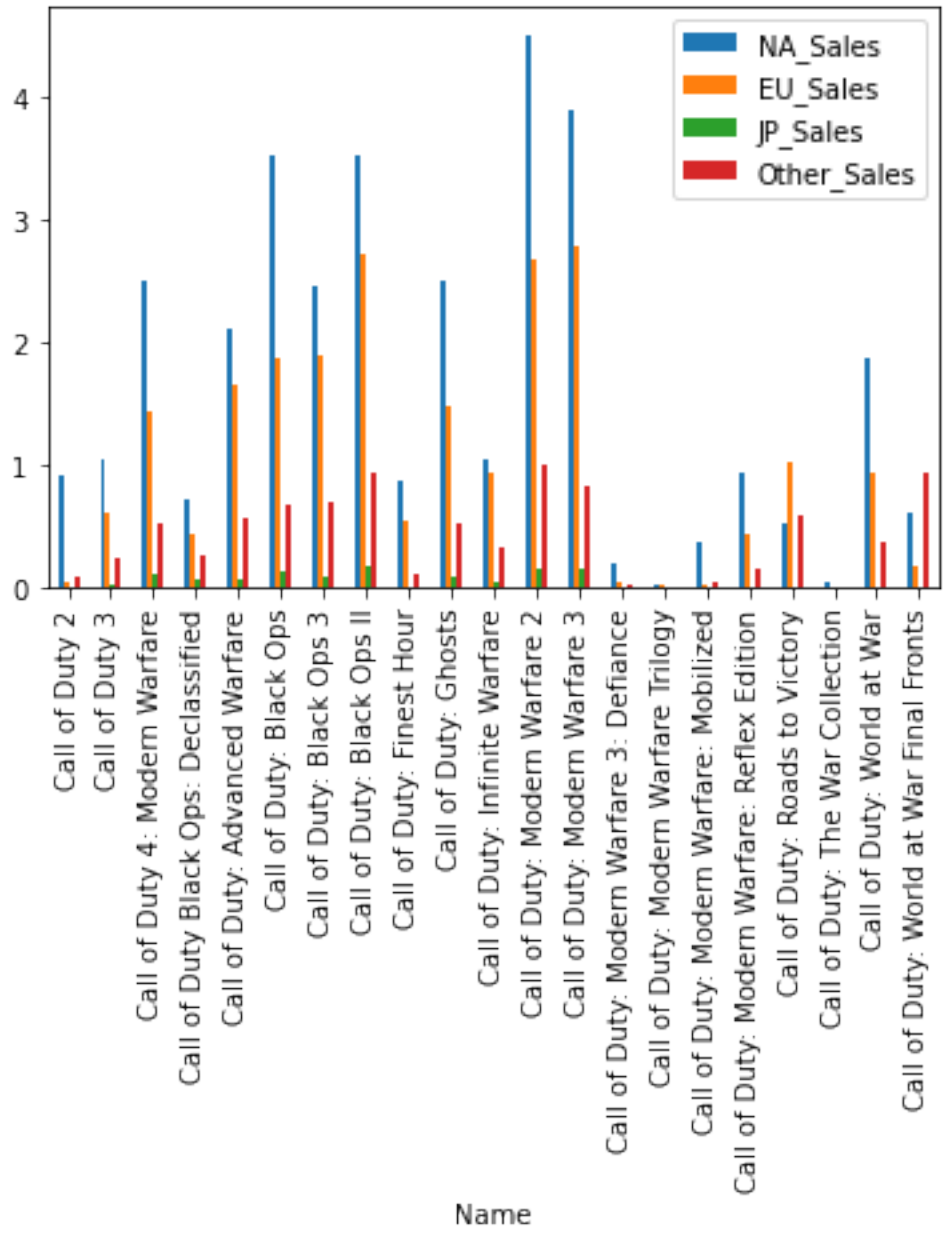


The naming of the games in the dataset can be a little off. The “New Super Mario Bros” should be the same as “Super Mario Bros”. With that being said Super Mario Bros was the best selling Mario game in North America, Europe, and Japan.

The next game we viewed was Call of Duty. Results are below.

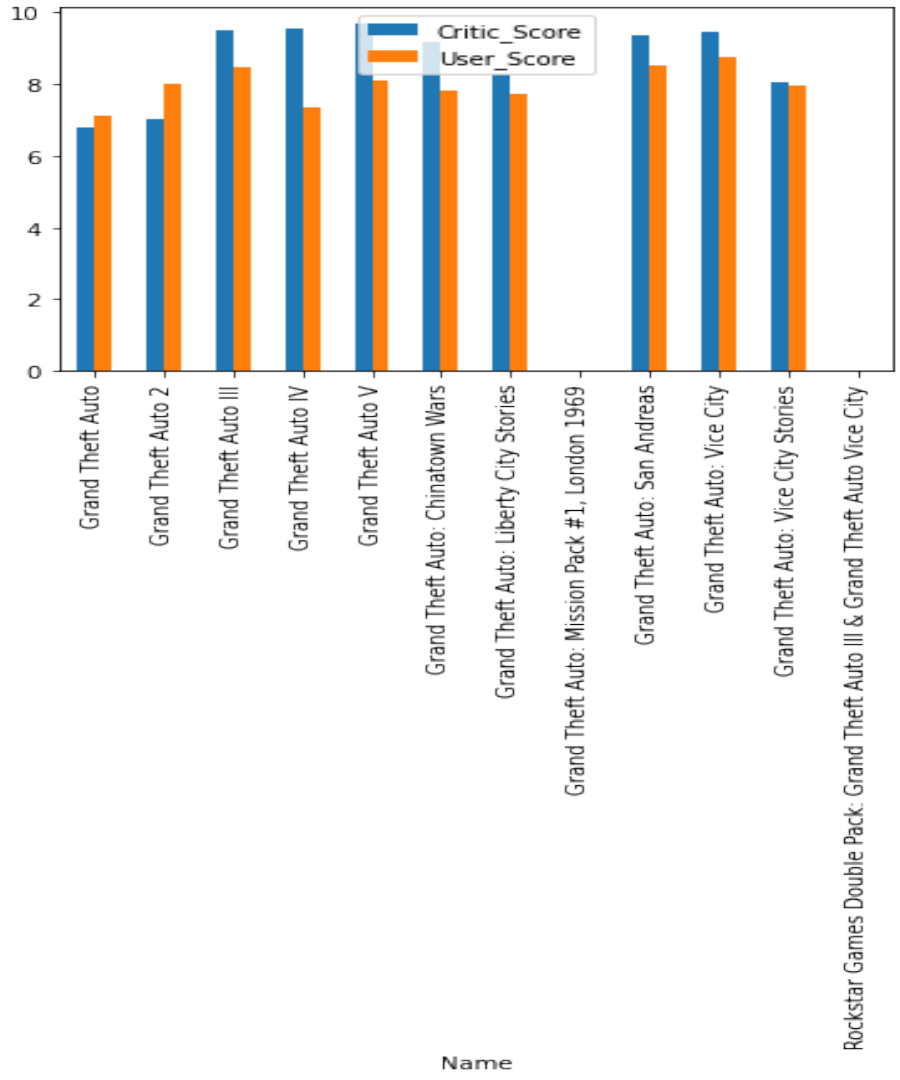


Top Critic Score: Call of Duty: Modern Warfare 2 (9.1)
Top User Score: Call of Duty 2 (8.2)

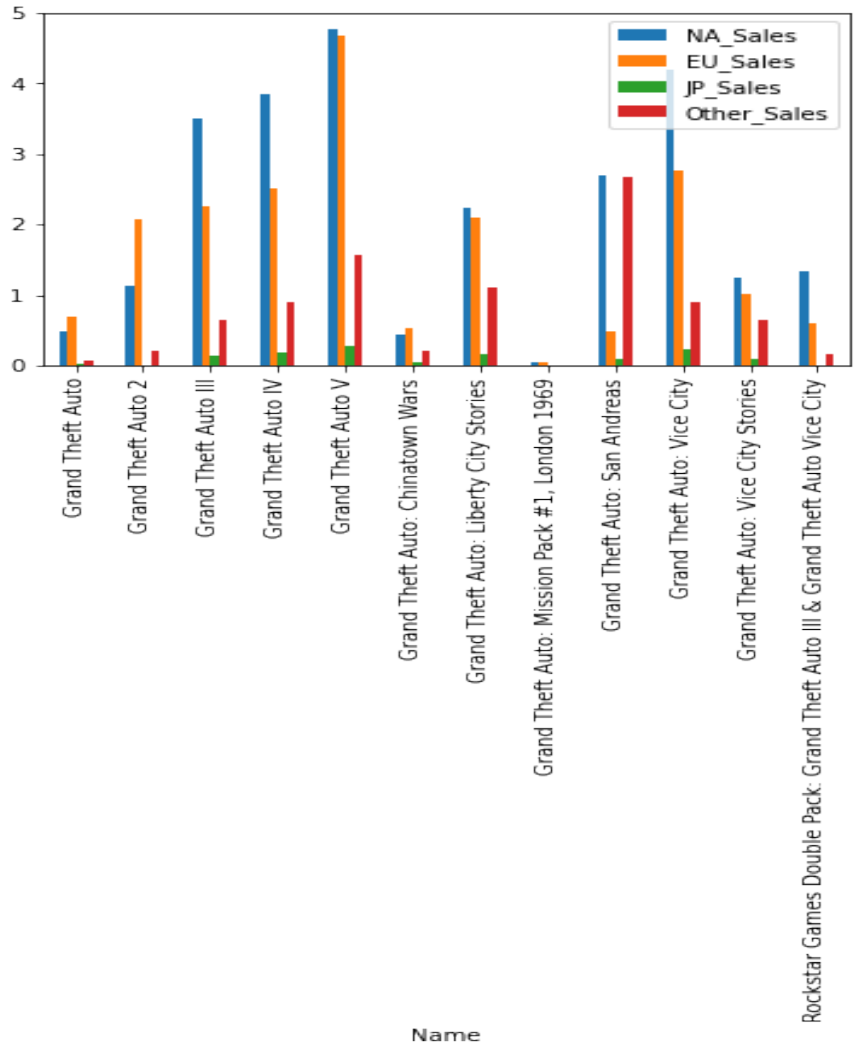


North America: Call of Duty: Modern Warfare 3
Europe: Call of Duty: Modern Warfare 3
Japan: Call Of Duty: Black Ops II

Grand Theft Auto:



Critic Score: Grand Theft Auto V (9.7)
User Score: Grand Theft Auto: Vice City (8.75)



North America: Grand Theft Auto V
Europe: Grand Theft Auto V
Japan: Grand Theft Auto: San Andreas

There are more examples of exploratory data analysis for this dataset within the corresponding Jupyter Notebook for this report.

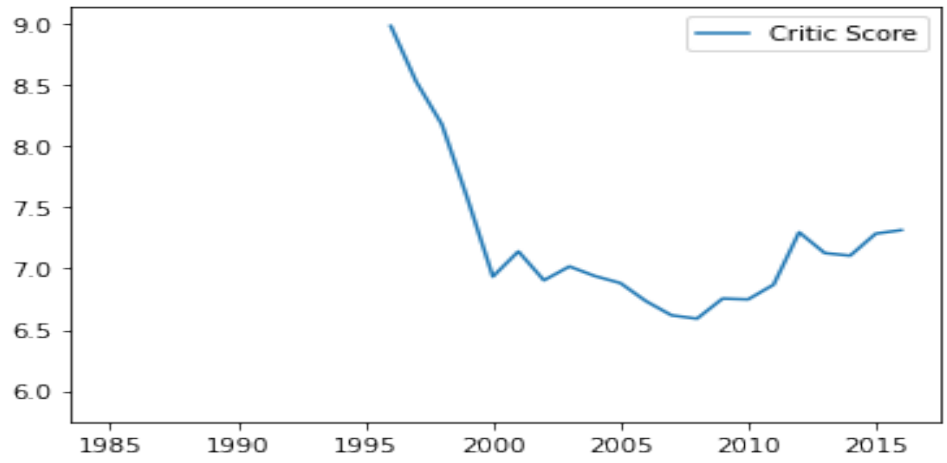
Analysis & Conclusions:

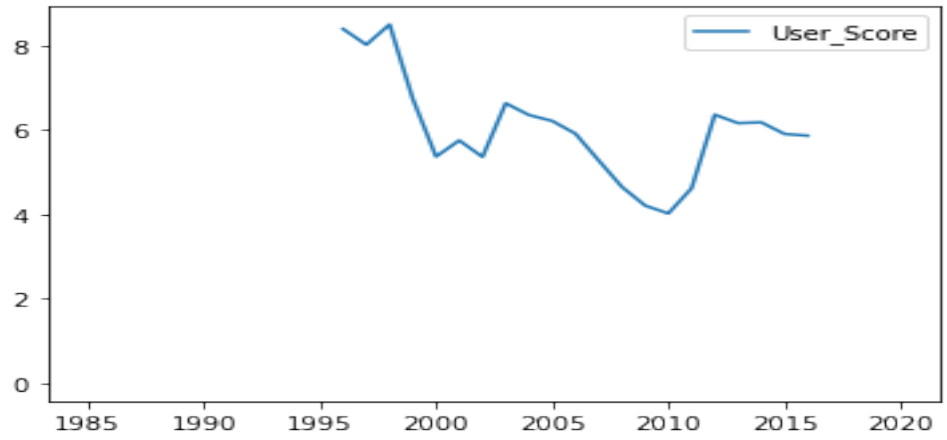
Our analysis involved us grouping the data in different ways to get answers to our questions. Most of our questions involved the trends of specific metrics over a period of time, therefore we created a new data frame that was grouped by a game's year of release and calculated averages for all numeric columns. Below is a snapshot of this new data frame.

	Year_of_Release	Critic_Score	User_Score	Global_Sales	NA_Sales	EU_Sales	JP_Sales	Other_Sales
0	1980.0	NaN	NaN	1.264444	1.176667	0.074444	0.000000	0.013333
1	1981.0	NaN	NaN	0.777609	0.726087	0.042609	0.000000	0.006957
2	1982.0	NaN	NaN	0.801667	0.747778	0.045833	0.000000	0.008611
3	1983.0	NaN	NaN	0.987647	0.456471	0.047059	0.476471	0.008235
4	1984.0	NaN	NaN	3.597143	2.377143	0.150000	1.019286	0.050000
5	1985.0	5.900000	5.800000	3.852857	2.409286	0.338571	1.040000	0.065714
6	1986.0	NaN	NaN	1.765238	0.595238	0.135238	0.943333	0.091905
7	1987.0	NaN	NaN	1.358750	0.528750	0.088125	0.726875	0.012500
8	1988.0	6.400000	2.200000	3.148000	1.591333	0.439333	1.050667	0.066000
9	1989.0	NaN	NaN	4.320588	2.655882	0.496471	1.080000	0.088235
10	1990.0	NaN	NaN	3.086875	1.591250	0.476875	0.930000	0.087500

It is important to note that the critic scores were originally on a scale of 0-100, but for visualization purposes we decided to scale it down to match the user score scale (0-10). We simply divided all of the critic score values by 10.

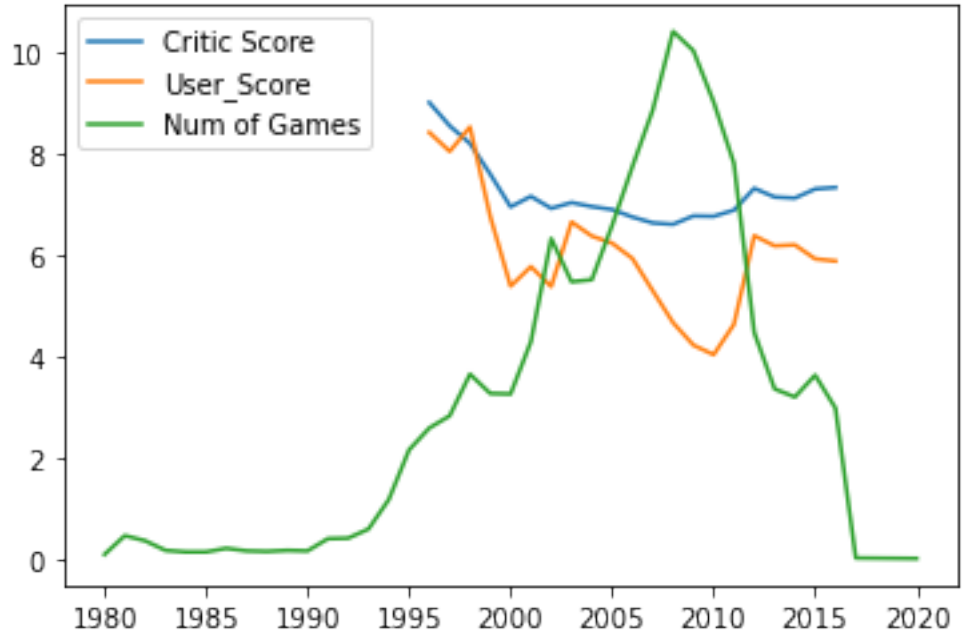
The first question we wanted to answer was, have games gotten better over time? We used the critic score and user score as a measure of satisfaction. The Critic scores were given by Metacritic staff, while user scores were given by Metacritic subscribers. These Metacritic subscribers are mostly common gamers. We visualized this data over time and obtained the results below.





The line plots above have similar shapes. They both have their highest score values around 1998 and see significant declines in the following year or 2. Both groups have local maximums around 2001, 2003, and 2012. Both groups see a decline over the years 2004 – 2009. From these plots we can conclude that games were more satisfactory towards the end of 90s and declined until around 2009. Games then began to improve over the following years and plateaued around 2011-2102. From these plots we can also conclude that users and critics were generally in agreement on their judgement of games.

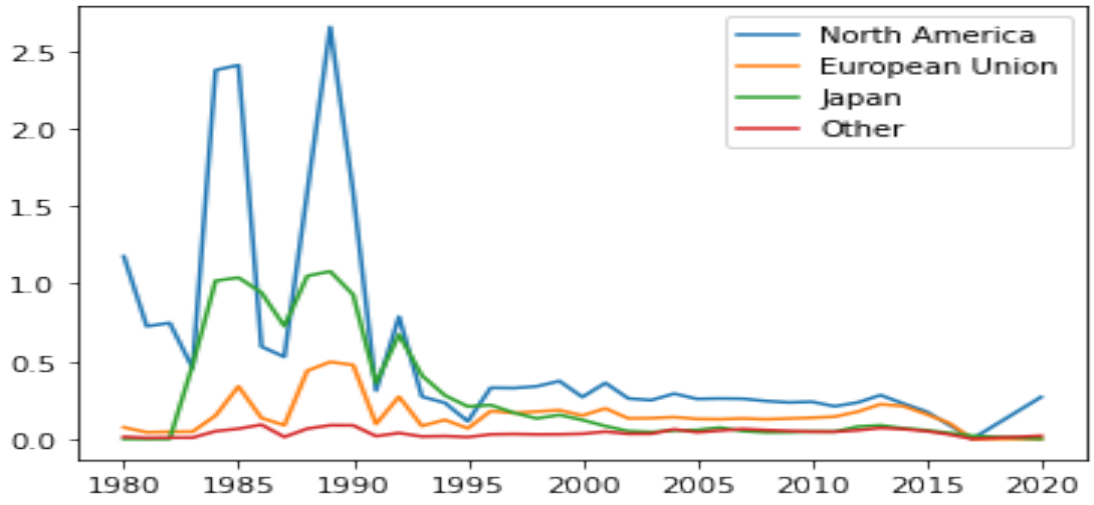
After drawing a few conclusions from the scores data, we wanted to take it a step further and try to determine if there were other data points that could have influence on scores. After comparing a few of our other metrics with critic and user scores we found something quite interesting. The number of games released in a given year appears to be inversely related with critic and user scores. The results are visualized below.



Note: All of the values in Num of Games were divided by 100 for visualization purposes.

It appears that when there are less games available, people tend to enjoy them more and as the number of games increase the satisfaction of games decrease. The graph shows the number of games increasing as user and critic scores decrease. The maximum number of games almost lines up perfectly with the minimum score numbers. This conclusion is fairly intuitive. With video games becoming increasingly popular, the demand for video games increased over time. This caused developers to begin creating more games and it appears that came at the expense of the quality of games.

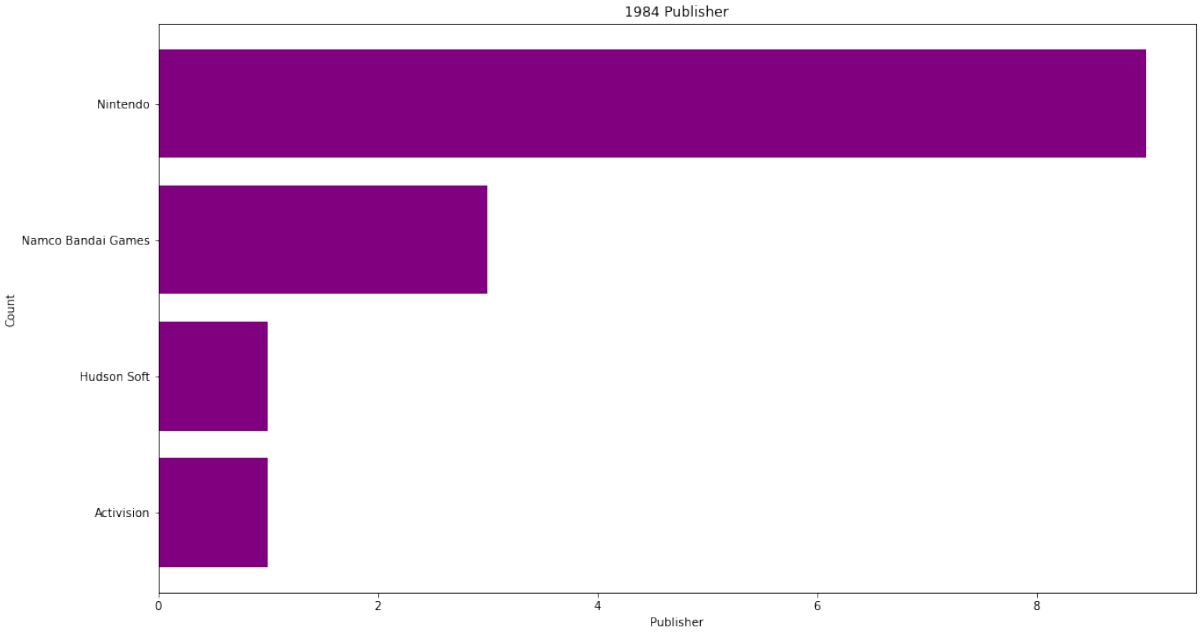
Next, we took a look at video game sales over time. We wanted to determine if sales were typically improving or declining over time. Below are the results.

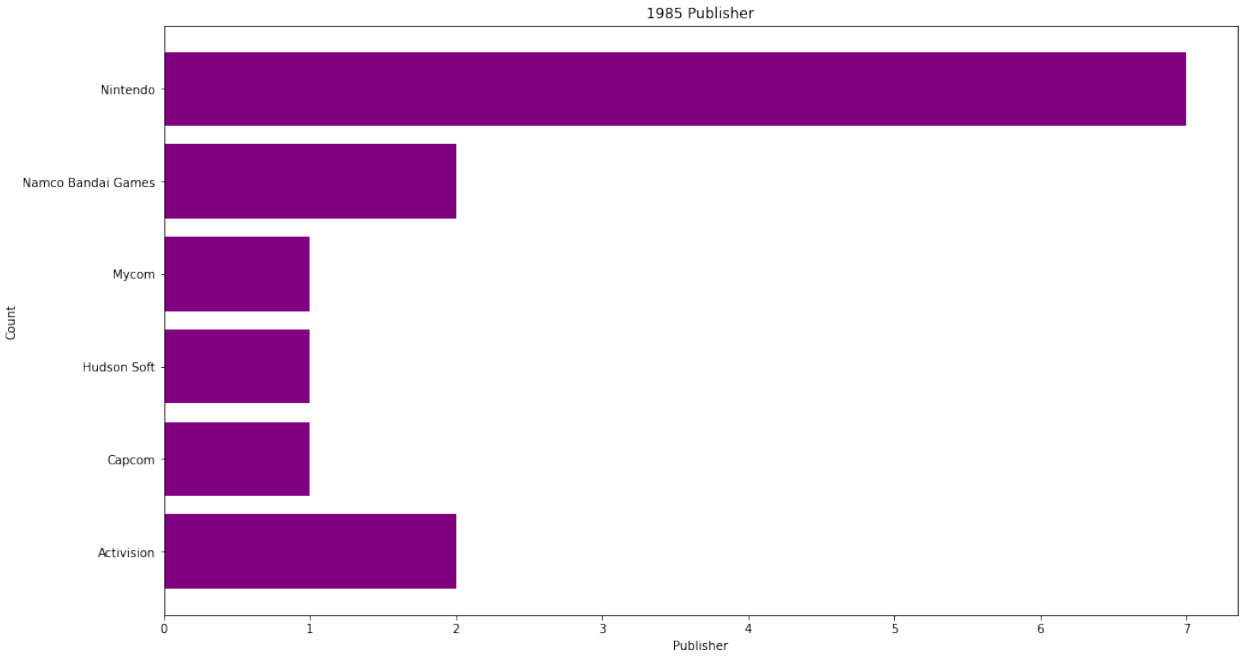


North America, Europe, and Japan all have similar trends in sales. There are huge peaks around years 1984-1985, and 1989. There are dips in years 1986-1987. After 1995 sales numbers flatten out and are more consistent. After viewing this data, we wanted to answer the question: What could have contributed to this similar pattern we see between years 1984 and 1990?

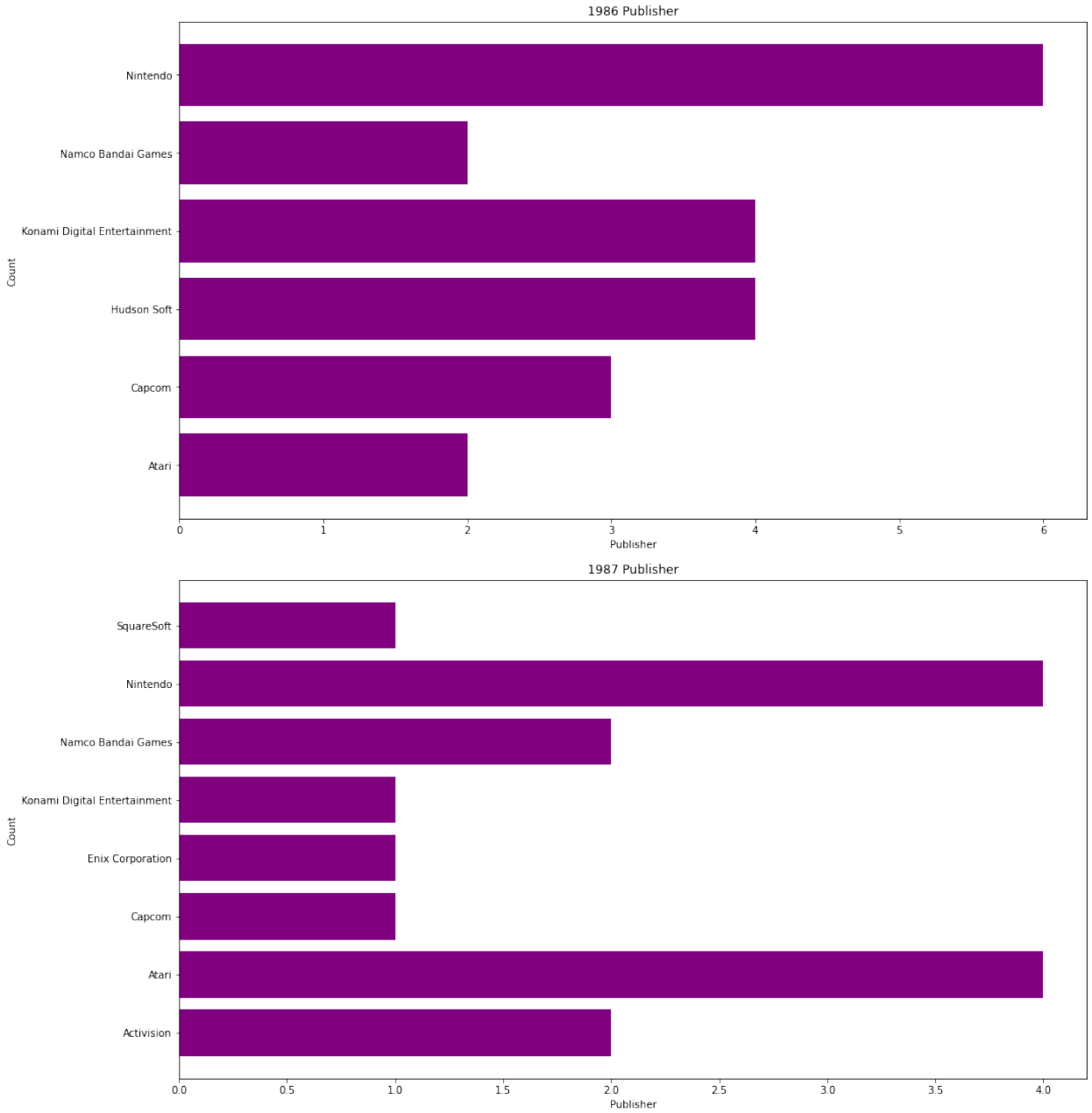
We initially took a similar approach to the previous question and searched for trends in other numeric variables that may have a correlation with this data. We were unlucky in this approach. We then decided to analyze some of the categorical variables over the years in search of a similar pattern. The categorical variables include genre, developer, platform, publisher, and rating. We grouped by the year and calculated the sum of each unique category of a specific variable within that year. For example, in the year 1984 the distribution of genres is shown below with an accompanying bar graph.

After calculating and visualizing the distribution of these categorical variables for each year from 1984-1990 we found that the publisher variable has a similar pattern with sales. In the years 1984 and 1985, Nintendo outproduced other publishers by an average of 7.33 and 5.6 games. These are the years we saw peak numbers in sales.

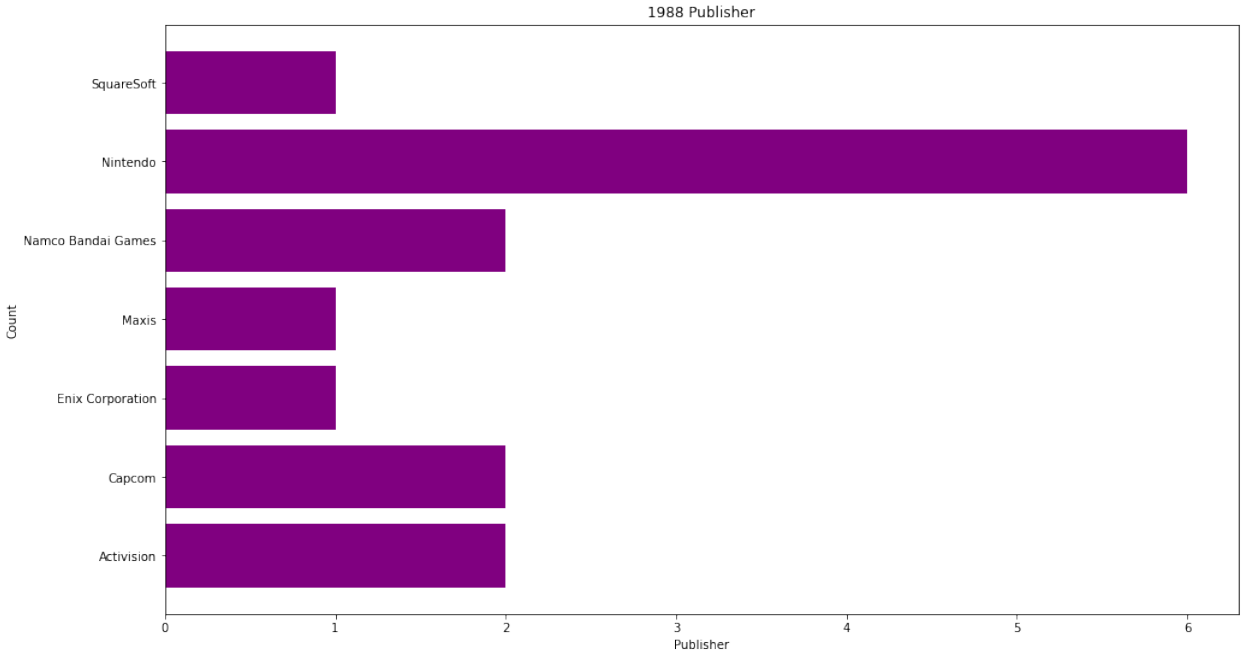




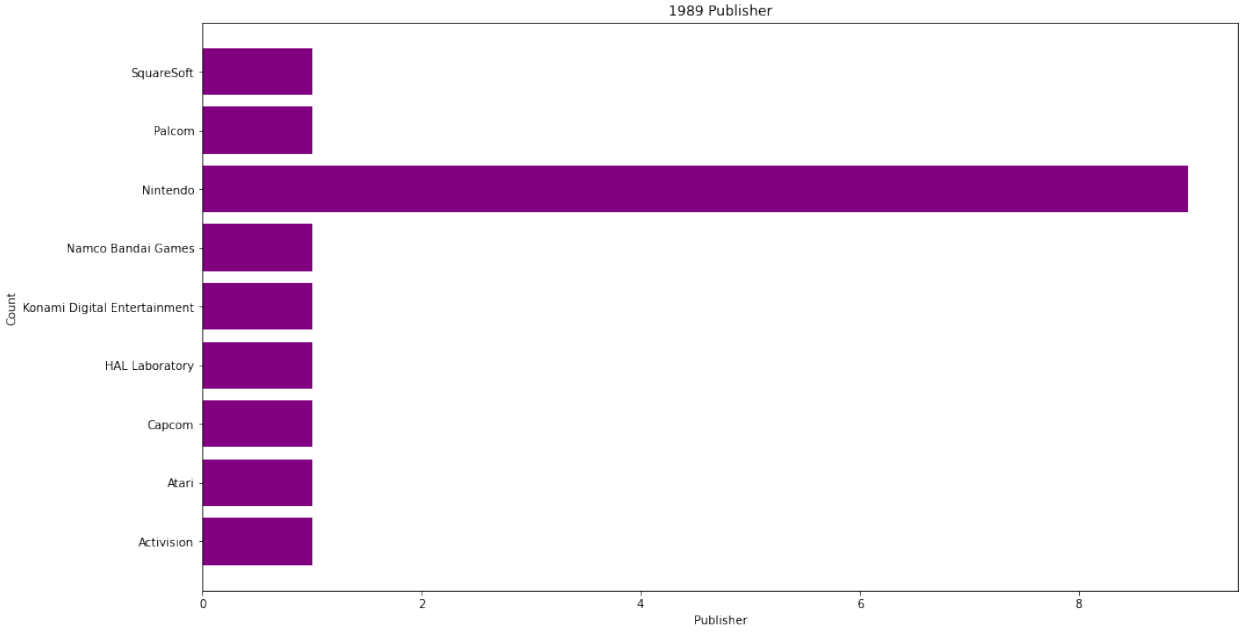
In 1986 and 1987, Nintendo outproduced other publishers by an average of 2.6 and 2.3 games. These are years we saw major dips in sale numbers.



In years 1988 Nintendo outproduced other publishers by 4.5 games. This year saw an increase in sales.



In 1989 sales reached their highest numbers. This year, Nintendo outproduced other publishers by its largest margin, 8 games.



In conclusion, sales numbers tend to mirror Nintendo's level of production dominance over other publishers.

Roles of Group Members:

Matthew Belizaire was responsible for exploratory data analysis. Christopher Fredrick was responsible for analysis and conclusions.