Example Project Report

Video Game Sales and Metacritic Scores

1. Introduction

According to PEW Research[1], 96% of adults between 18-29 occasionally consult reviews before making a purchase. But does this mean that higher reviews mean more sales? I am also curious whether there has been an increase in the production of games in recent years and whether it is an increase in quality games or if review scores decrease with the oversaturation of video game releases.

In this project, I plan to use Metacritic's user and critic review scores to see if there is a correlation between the scores of a game and how much it sells using sales data that I obtained through VGChartz.

2. Data

This project uses two primary sources of data: Metacritic's[2] pages on review scores of different games and VGChartz[3] data on video game sales.

2.1 Review Scores

I collected data from Metacritic, which contained user and critic review scores for each game.

Metacritic's data was housed on 186 different pages. I wrote a crawling script to collect all the data across these 186 pages. I collected the game title, summary, platform, user score, meta score, and release date for 18,526 games. Some of these games were the same game that was released on different platforms, which is essential to consider when I begin merging the data. This constructed a data frame which is contained in the Metacritic.csv file in the project folder. The web crawling script that I wrote is contained in the metacritic.R file in the project folder.

The data didn't need much cleaning; however, to make the platform match with the VGChartz data, I created a dictionary to convert Metacritic's platform data to match VGChartz. This cleaning was done in the file combining_vgs.R

2.2 Video Game Sales

The website VGChartz provides sales data on the number of units sold.

The information was contained in 1249 pages. Due to the extensive nature of the data, I decided to break the crawling into six different CSV files, all contained in the VGChartz folder.

I collected the game, platform, release date, publisher, and sales data from PAL, Japan, North America, and total sales for 62,398 games. Like the Metacritic data, some of these games were released on multiple platforms, which each count as a unique observation in the data. Another interesting thing to note is that series were counted in this data frame alongside their individual counterparts. For example,

---

[1] https://www.pewresearch.org/internet/2016/12/19/online-reviews/
[2] https://www.metacritic.com/
[3] https://www.vgchartz.com/

Super Mario is counted as a series, but each entry in the series is also contained in the dataset. This was handled in the cleaning of the data.

In cleaning the data, I merged all six files into one complete file by adding all data frames to a list and running a for loop over them to row bind them into a complete data frame. I also removed unnecessary columns, trimmed the white space on the title, and got only the distinct values of the data set. The NA values in VGChartz were written as strings, which I handled to create NA values when I read the CSV into the script. I also removed the series by filtering them out in the data cleaning. This cleaning was done in the file combining_vgs.R

2.3 Combining Reviews and Sales

With both data sets including the same game title multiple times based on the platform, I had to merge based on the game title and platform. Both data sets are also vastly different in the number of observations. I decided to create two separate data sets in merging. One data set is the intersection of the data and only contains data that appears in both data sets. The other data set, called the union, contains all rows from both data sets. All this work was done in the Combining_VGs.R file in the project folder. I then imported both data frames into the Cleaning_VGs.R file in the project folder for further cleaning.

Since the differentiator between both data frames is their number of rows and most of my cleaning was regarding the columns, I created a function to run over both data sets to clean them. This cleaning included trimming the white space on the summary and publisher, turning the sales data into numeric, and adding the platform, publisher, and title as factors. I also turned all the release dates into a "Month, Year" format and combined the release date from VGChartz and Metacritic into one column. I then dropped all sales data except North America and total, the extra release date column, and the index columns. I exported these to VGSales_Union.csv and VGSales_Intersect.csv, respectively. The union had 68,288 observations, and the intersect had 12,288 observations. A description of each variable is contained in Table 1.

Table 1 Data Dictionary

Column	Type	Source	Description
game_title	Text	both	Title of the game
platform	Text	Both	Abbreviation of the video game console that the game is on
publisher	Text	VGChartz	The publisher of the video game
na_sales	Numeric	VGChartz	The number of units sold in North America
total_sales	Numeric	VGChartz	The total number of units sold worldwide
summary	Text	Metacritic	A description of the video game
release_date	Text/Date?	Both	The Month and year that the game was released
meta_score	Numeric	Metacritic	The review score as decided on by game critics
user_score	Numeric	Metacritic	The review score as decided on by the users of Metacritic

3. Analysis

3.1 Video Game Ratings and Sales

Do games with higher review scores sell more? I wanted to find out and navigate the relationship between the different types of sales and review scores. I started by trying to calculate correlation coefficients to explore the relationship between variables. There was a correlation coefficient of .954 between the total sales and North American Sales. This is evidence of an extremely strong correlation between these two variables. North American sales make up 48.659% of the total sales, which is a significant reason behind this strong correlation. However, the correlation between sales and review scores was very weak. The correlation coefficient between total sales and meta score was .248, and it was .102 between total sales and user score. While there was a moderate correlation between user score and meta score at .524, it is nowhere near as strong as it was between the two sales variables.

I initially thought that this might be due to the difference in rating scales between the two groups of people; however, I created a boxplot of the two, using ggplot[4], to prove that this was not the case quickly. This plot is shown in Figure 1. While there might be slight differences in the interquartile range, the medians of the two variables are the same. However, if you take the difference between their means, you get a value of 1.67, which is somewhat significant. A histogram that compares the two variables, such as the one in Figure 2, gives us a good insight into the reason behind the difference in means. We can see that the meta scores have much higher counts of high reviews, while the user score has a greater number of low reviews. This may be part of the reason why the meta score is has a slightly better correlation with sales than the user score.

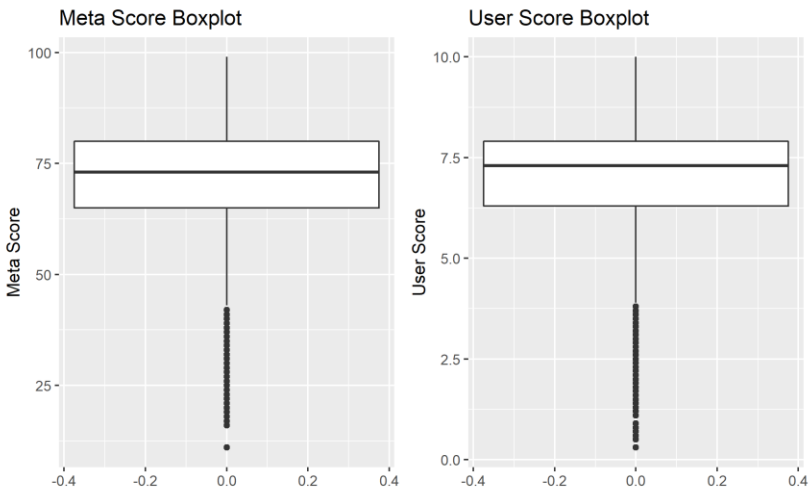


Figure 1 Review Scores Boxplots

[4] https://cran.r-project.org/web/packages/ggplot2/

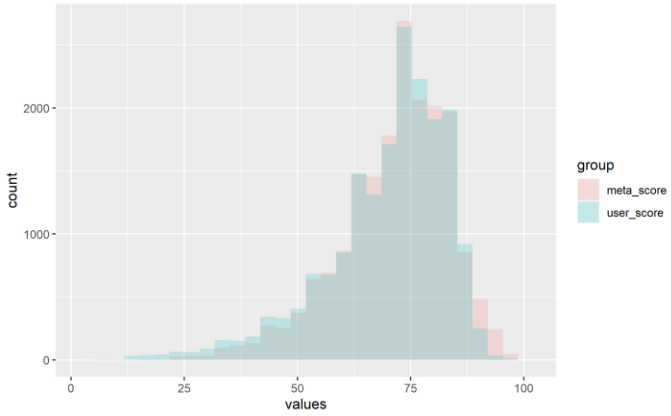


Figure 2 Review Score Histograms

The meta score is also a better predictor of sales. While the correlation isn't extremely strong, there is some basis for the reasoning behind this, as critics usually publish their review scores in the days before a game is released. Many companies send out review copies to critics in hopes that the reviews will boost sales[5]. Increasing excitement around the game before the release date benefits both the video game company and the critic. However, this low correlation appears to show that there isn't much benefit in doing this. While I initially thought this might be due to the large number of games with little sales, recalculating the correlation coefficient between meta score and total sales gave me even less of a correlation at .065. While high review scores may help a company's reputation, they do not correlate well with more sales.

3.2 Video Game Publishers

Which Video Game publishers have the highest sales? To find this out, I created a summary data frame, publisher_df, which was grouped by the publisher. In this data frame, I included the ratings' average, the sales' sum, and the number of games each publisher has released. To find the publishers with the highest sales, I took a slice_max of the data frame based on the total sales. The result appears in Table 2. The information is also saved in the publisher_sales.csv file in the project folder.

Table 2 Slice Max Summary

	publisher	total_sales	na_sales	user_score	meta_score	number_of_games
1	Nintendo	1915460000	859150000	7.86772727272727	77.3954545454545	1410
2	Activision	727710000	428300000	6.80404858299595	70.9635627530364	1567
3	Electronic Arts	659150000	329130000	6.94391891891892	73.8192567567568	1591
4	Sony Computer	557090000	239880000	7.54783950617284	74.9845679012346	1365

[5] https://eriktwice.com/en/2021/09/24/what-are-review-copies/#:~:text=Review%20copies%20are%20what%20their,it%20to%20pen%20a%20review.

	Entertainment					
5	Ubisoft	500080000	260830000	7.00683012259194	70.7302977232925	1646
6	EA Sports	497620000	276250000	6.72665036674817	79.1075794621027	800
7	THQ	339470000	208390000	7.26092715231788	69.0794701986755	1092
8	Sega	291080000	119560000	7.35538116591928	72.2959641255605	2191
9	Capcom	274690000	104710000	7.58064516129032	74.7507331378299	1062
10	Rockstar Games	263430000	125600000	7.75555555555556	82.9506172839506	175

Since I noticed that many publishers had released many games, I calculated the correlation between the total sales and the number of games a publisher had released. The coefficient of correlation between these two variables was .715, which demonstrated evidence of a strong correlation between the number of games released and the total sales of a publisher. Since there was such a strong correlation, I thought creating a scatterplot between the two variables would be useful to analyze the data further. This scatterplot can be seen in Figure 3, which is shown below.

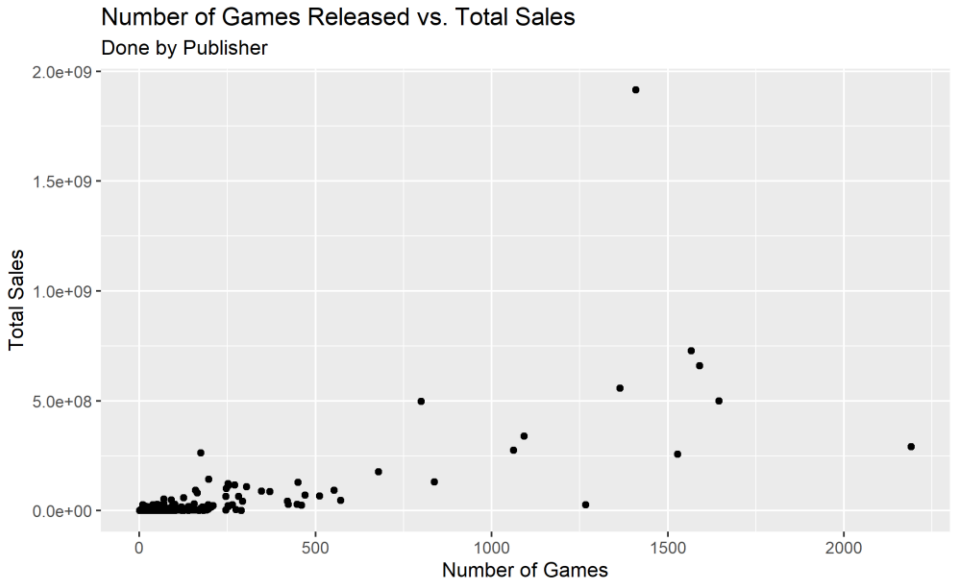


Figure 3 Scatterplot of the number of games by a publisher and their total sales

However, this correlation is not the case between the number of games released and the publisher's average meta score. The correlation coefficient between these two variables is -.013, which is evidence of no correlation. With the strong correlation between the number of games and total sales, we can predict that a company's average meta score doesn't significantly impact its total sales. We can compare the average meta score of 71.145 to the values in Table 2 to prove that meta scores and sales aren't

strongly correlated, as 30% of the observations in Table 2 are below the average. This can also be seen in Figure 4, a scatterplot of the number of games released and the meta score.

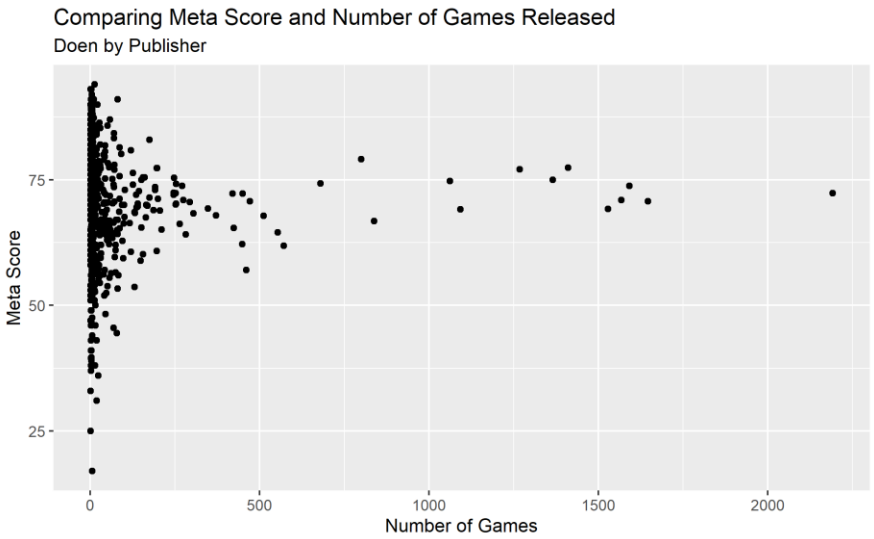


Figure 4 Scatterplot of the number of games by a publisher and meta score

3.3 Release Date

How have the sales, number of games released, and review scores changed over the years? To start, I created a data frame called by_month, which grouped the games by month and calculated the average review scores, the sum of sales, and the number of games. I decided to turn these into a date format, making dates easier to work with on scatterplots. I did this by setting the column as a date with the day value being the first of the month it was released.

To start, I decided to see if there is an increase in the number of games being developed over time. Finding the correlation coefficient between the release date and the number of games released is .533, which is evidence of some correlation, but it is not especially strong. I decided to create a plot of the number of games released in relation to the date. I also decided to compare the total sales of games over time, using the patchwork package[6], to see if there was a clear pattern between the number of games released and total sales. This scatterplot can be seen in Figure 5.

[6] https://cran.r-project.org/web/packages/patchwork/

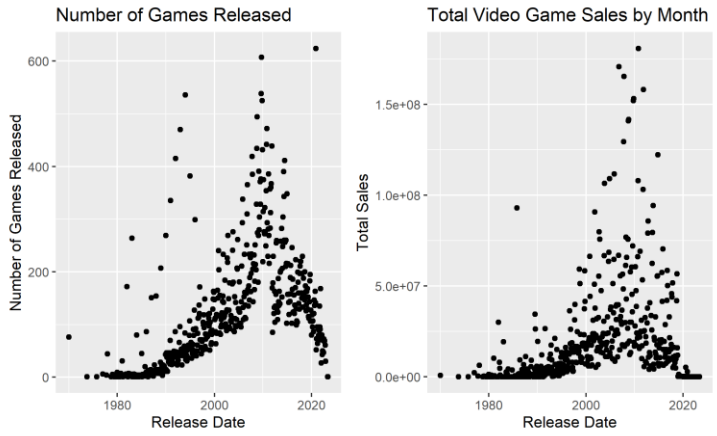
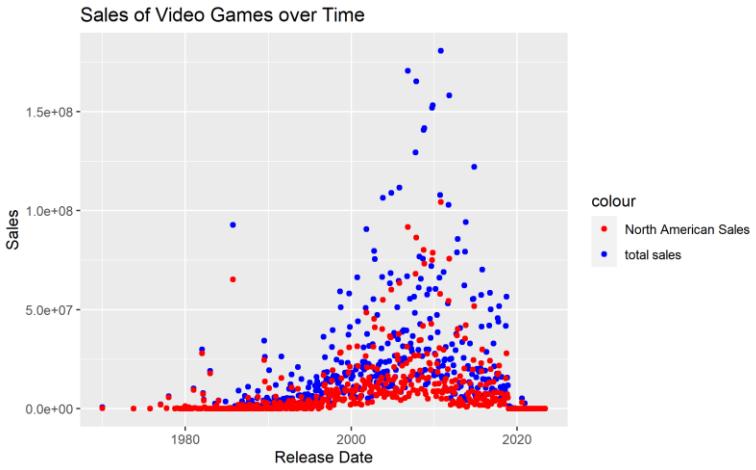


Figure 5 Number of Games Released and Total Sales by Month

The point at the end should be ignored, as it includes two games yet to be released. The outlier of games released in 2020 should also be ignored, as many games were misclassified to this date by the websites. Figure 5 reveals a large peak in the number of games released around 2010 before it decreases again. However, the sales of games did not see the drastic spike that was seen in the games released. However, the purchase of games released after 2010 remained somewhat consistent, unlike the number of games released in that same period. The increase in games around the late 2000s is likely due to the release of new video game consoles that reached many mainstream markets. Many video game consoles could also play Blu-ray discs and had multimedia features[7], which increased their popularity.

To further dive into sales, I wanted to see how the passage of time has affected sales globally versus in North America. The correlation coefficient between total sales and the release date is .273, much lower than for the number of games released and the release date. This also holds for the correlation between North American sales and the release date, which has a correlation coefficient of .229. Figure 6 shows a graph comparing the sales to the release date.



[7] https://ultimatepopculture.fandom.com/wiki/2000s_in_video_gaming

Figure 6 Total Sales by Release Date

You can see that the trend of frequent outliers appears to be much greater for total sales than it is for North American sales. Taking a slice_max of both North American and Total sales shows that this pattern is likely due to increased sales during the holiday season, as the top ten sales months for both variables are in October and November. Many game publishers have caught on to the increase in sales during this time, as many of the months with the most game releases also fall within this timeframe. While there was an increase in sales throughout the 2000s, the same cannot be said about review scores.

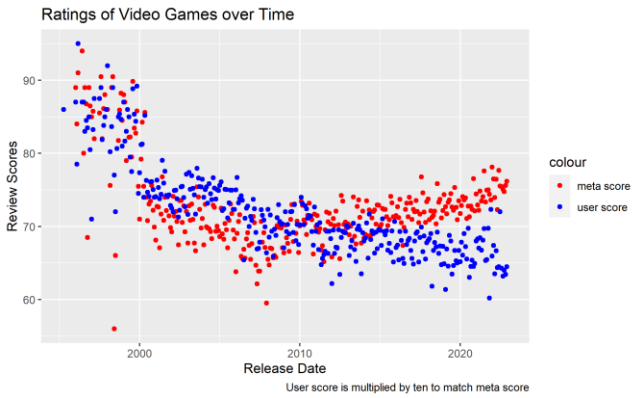


Figure 7 Review Scores by Release Date

Figure 7 shows the interesting development of review scores over time. The review scores of games see a sharp decrease beginning in 2000. However, the meta scores decreased much more drastically than the user scores during this time. While the meta score began to increase again around 2010, this is different from user scores, which continuously decreased throughout this time. This is also displayed in the correlation coefficients of the two variables. The meta score has a correlation coefficient of -.325 with the release, while the user score has a correlation coefficient of -.837 with the release date. The difference in how the direction of review scores after 2010 can partially explain the strong correlation value between the meta score and the total sales of a game.

4. Conclusion

In this project, I analyzed three aspects of video game sales: the publisher, the release date, and how a game's sales and review scores influence each other. Using summary tables and plots, I was able to show the limited correlation between review scores and sales, how publishers that release more games do not have higher review scores, the boom in the number of games released in the 2000s, and how the lack of sales to match plummeted the games released in the following years, and the decline of review scores over time. This project has several limitations, including the lack of review scores for games released before the mid-90s, the improper release dates of some games, and the lack of information about the video game consoles. Future work on this project could include finding more sources to create a more complete data set, analyzing the purchase of different video game consoles to see if the spikes in video game releases coincide with more people buying video game consoles, and analyzing whether more time in between game releases for a publisher really means more a better game (higher sales and review scores).