MTH 765P Mini-Project: Top 10 Video Games

Marc Grammersdorf

Registration Number:220891802

Contents

1 Introduction

In this report, the data set used is called, list of best-selling video games (Top 50 Video Games).I have used the Kaggle website to get this data set, which attracted me the most. The data set shows the 50 best-selling video games that were released up to 2020. There were some mistakes in the data set, which needed some pre-processing to be able to make it clear to read and be able to make visualisations correct. This was one reason why I have used this data set, as it made me think of ways of making the data set clearer and be able to visualise in a more efficient way.

1.1 Data Set Information:

Rank: Rank of each video game
Title: Title of the Video game
Sales: Sales of video game
Series: Related release of the video game
Platform(s): The platforms that the games were played in
Initial Release Date: Date that the games were created
Developer(s): video game programmers
Publisher(s): The company that has published the video games

	Rank	Title	Sales	Series	Platform(s)	Initial release date	Developer(s)[a]	Publisher(s)[a]
0	1	Minecraft	238,000,000[b]	Minecraft	Multi-platform[c]	November 18, 2011[d]	Mojang Studios	Xbox Game Studios
1	2	Grand Theft Auto V	170,000,000	Grand Theft Auto	Multi-platform	September 17, 2013	Rockstar North	Rockstar Games
2	3	Tetris (EA)	100,000,000	Tetris	Multi-platform[e]	September 12, 2006	EA Mobile	Electronic Arts
3	4	Wii Sports	82,900,000	Wii	Wii	November 19, 2006	Nintendo EAD	Nintendo
4	5	PUBG: Battlegrounds	75,000,000	PUBG Universe	Multi-platform	December 20, 2017	PUBG Corporation	PUBG Corporation
5	6	Super Mario Bros.	58,000,000	Super Mario	Multi-platform[f]	September 13, 1985	Nintendo R&D4	Nintendo

Figure 1: Part of DataFrame before Pre-processing

Libraries used: Pandas, Matplotlib.pyplot, seaborn, re
Reference of Data Used: https://www.kaggle.com/datasets/devrimtuner/top-100-video-games

2 Pre-processing

For this data set, pre-processing was the major concept of this project. The module pandas was used to pre-process the data set.Further down, I will be explaining some pre-processing that was done to be able to continue in a reliable and convenient way for the visualisations.

2.1 Removing NULL values:

As we can see from the data set, there are some faulty names and values. However, the first thing to do was remove any NULL values from the data set. Then checked, if there is any NULL value found to be extra sure that our code worked. The reason we did this, is because for example,some Game titles were not given any Sales values, so we had to remove them from our data set, as they will not be useful for us and would make our visualisations less accurate.

2.2 Removing Unnecessary Special Characters:

Removed the unnecessary brackets that were found in the data set, for example, "Multi-platform[c]" which is found in the column "Platform(s)". This would help make the visualisations more understandable.Moreover, as there were various faulty brackets in the data set we had to use the function re.compile and do a for loop to be able to delete each bracket were necessary.

2.3 Renaming:

Renamed some columns, to make them easier to read and create new ones were appropriate.For example, further down we will see that we used a new column which we created called, 'Year Released', which was used to find the Average Sales per year for each Publisher.With this column, it made it easier for us to compute and execute a visualisation.

2.4 Data Types:

Had to check if the data types were correct for each column used in the data set. The column "Initial release date", needed to be changed to date-time and the sales should be in integers. This was important, as without it some visualisations would give errors. Some coding was used for the Sales part as the values had commas in them, which first had to be removed and then replace them as integers from objects.(e.g. 238,000,000 to 238000000)

2.5 Faulty Names

There were some faulty names from our data set, that should be the same name and not different.For example, 'Wii U0/0Switch' and 'Wii U / Switch'. There were some more faulty names, but have done the ones that were appropriate for the visualisations below.

3 Visualisation

3.1 Top 10 Video Games

Created a new variable called "top10", which shows us the first 10 rows of our data set. In our case, the first 10 are the top 10 video games.

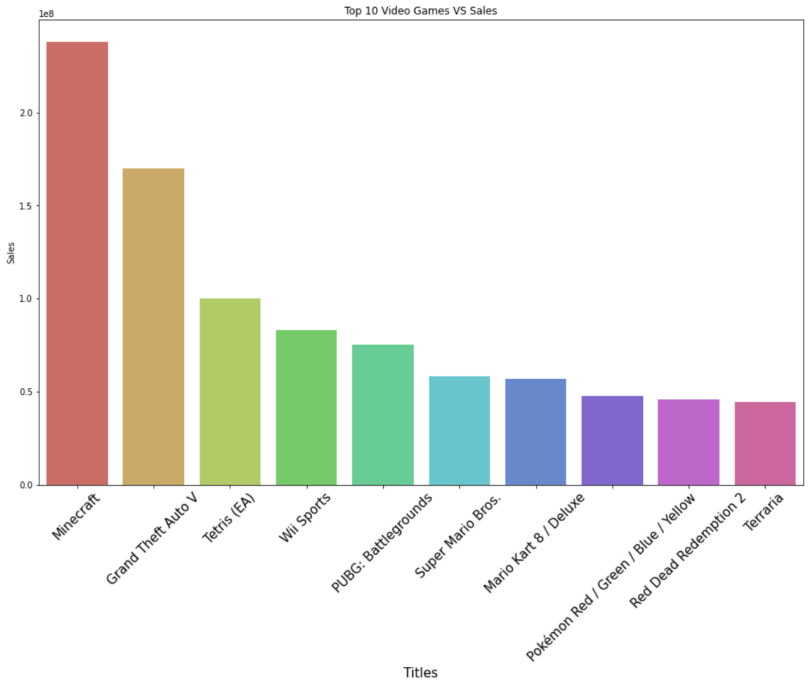

Figure 2: Top 10 Video Games VS Sales

From the figure above, we can see that Minecraft is the leading video game up to 2020 by a big amount of Sales.Next, comes Grand Theft Auto V and the all time classic, Tetris(EA).The bar plot used is in decreasing order, to make it convenient and easy to understand for the reader.

3.2 Number of Times Publishers are found in Data

From this section of our visualisation, we can see the total number of times, Publishers are found in the data set.

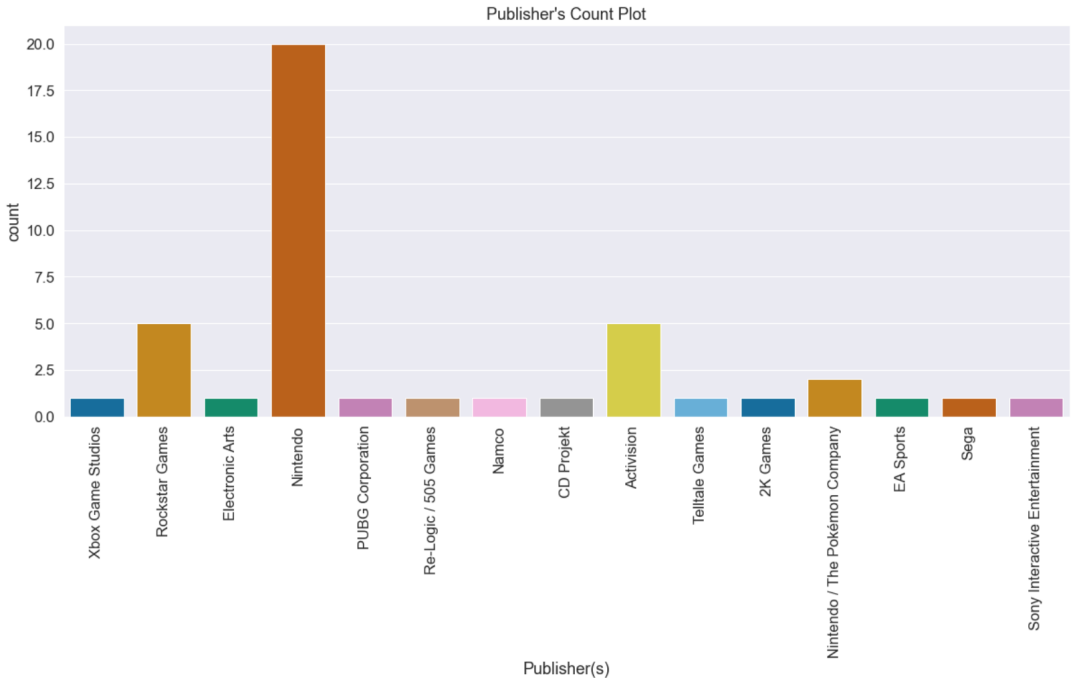

Figure 3: Count for each Publisher in data set

There are three main Publishers found in our data set, precisely, Nintendo,Rockstar Gaming and Activision.This will help us for the visualisations that we created later, as we will be looking precisely to these Publishers to make our visualisations more specific based.As for the other Publishers, it would be hard to create an understandable and reliable viuslisation.

3.3 Sales Per Year

Created new column for only the years that the games have been initially released. With this, we were able to visualise the Sales of each year for the main Publishers, in our case Nintendo, Rockstar Games and Activision.

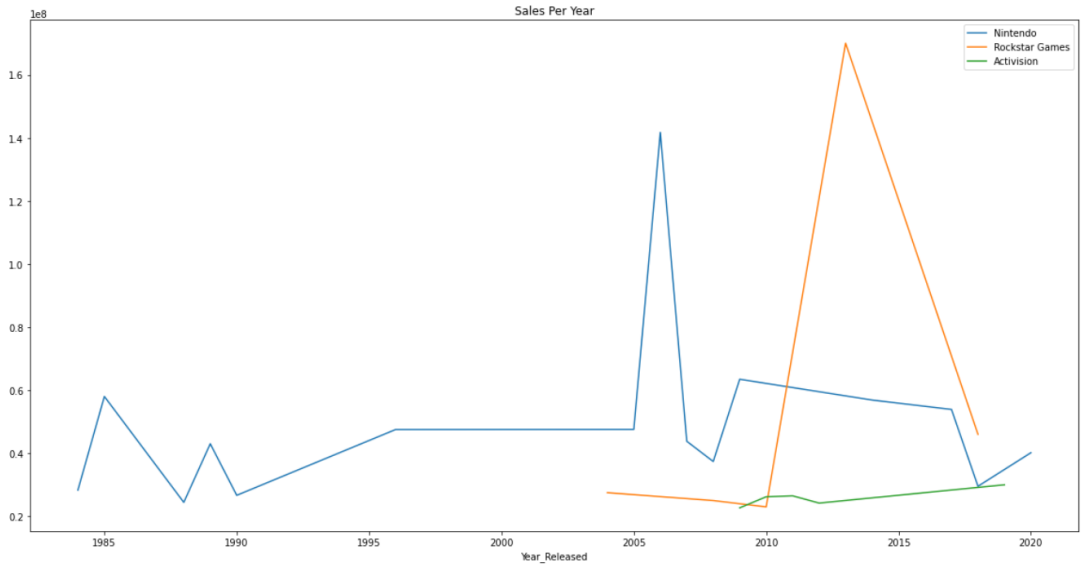

Figure 4: Sales Per Year (Nintendo/Rockstar Games/Activision)

From the figure above we can see various observations.For example, if we take into consideration, Nintendo the largest publishers in the top 50 video games, we can see that it had some up's and down's during the years. One of its greatest years was from around 2005. As for Rockstar Gaming, we can see a huge increase in sales during the years of 2010-2015 and vise versa from 2015 onwards.For Activision, we can see a stable process, as years go by.

3.4 Nintendo's Series Proportion

For our next visualisation, the visualisation material used was more in depth, as we are looking precisely for the Series that includes only the Nintendo publisher.From the figure below, we will be able to see the distributions of the series found as well as the value they hold in it.

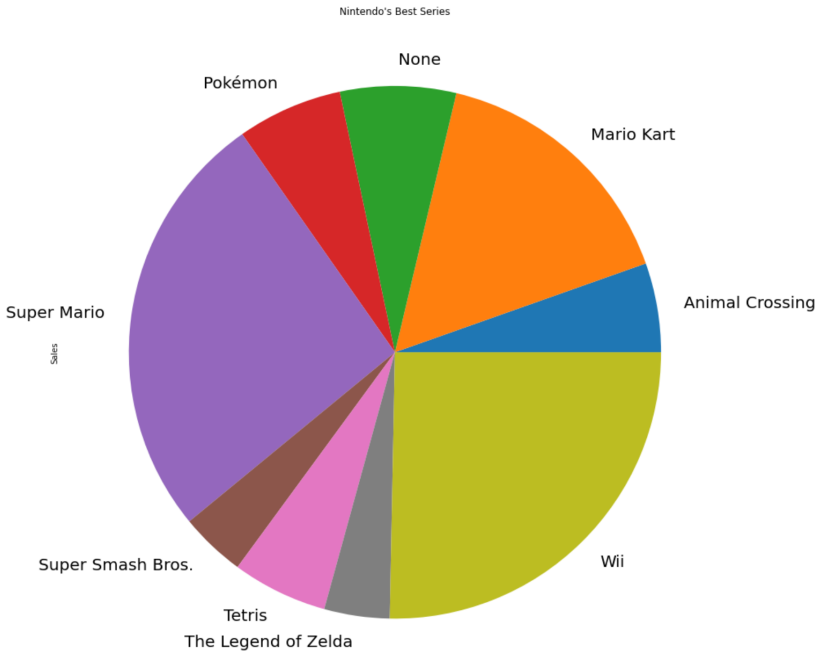

Figure 5: Nintendo's Series Breakdown

We can see from the figure that Nintendo has a range of game series, with the most popular being WII and Super Mario. A pie-chart was used to be able to show each of the game series in percentage form and make it clearer for the viewer to see how each series contributes in Nintendo.

3.5 Nintendo's Platforms

For this subsection, we will be checking to see Nintendo's different kind of platforms and amount of times they appeared in the data set.

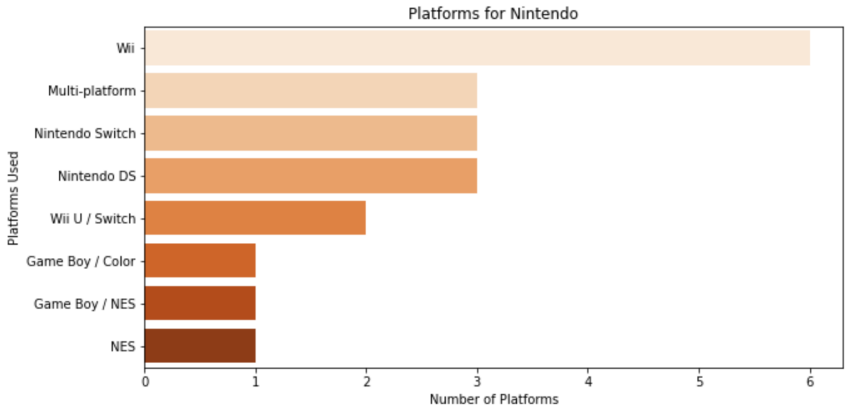

Figure 6: Platforms that were used for Nintendo

As we can observe, from the above figure the Platform WII was found 6 times in the top 50 Video games for Nintendo.This shows us that the WII platform was a great hit for Nintendo and would recommend creating a game from this type of platform. We can also see that, 'Game Boy/Color','Game boy/NES' and 'NES' platforms appeared the least amount of times in our data set, which is one time for each.

4 Conclusion

In conclusion, the analysis of the data set was enjoyable , as I am a gamer myself, and found it very interesting going through the cleaning and pre-processing.Some visualisations used, were not for the data set as a whole, rather more specifically to the one of the main Publisher that we have pointed out from our count plot such as Nintendo.This helped us get a better understanding of the data and were able to come out with some observations regarding Nintendo.For example, from Figure 5 we can say that, creating a new game Series for Super Mario would potentially increase sales for Nintendo.