

Video game sales analysis

BIG DATA ANALYSIS
JADE MAGBANUA

Contents

Understanding the data	2
Data Pre-processing	3
Data Cleansing:	3
Data Queries	4
Data Visualisation	5
Additional data graphs.....	6
Power BI	7
Statistical Inaccuracy.....	10
Conclusion.....	12

Understanding the data

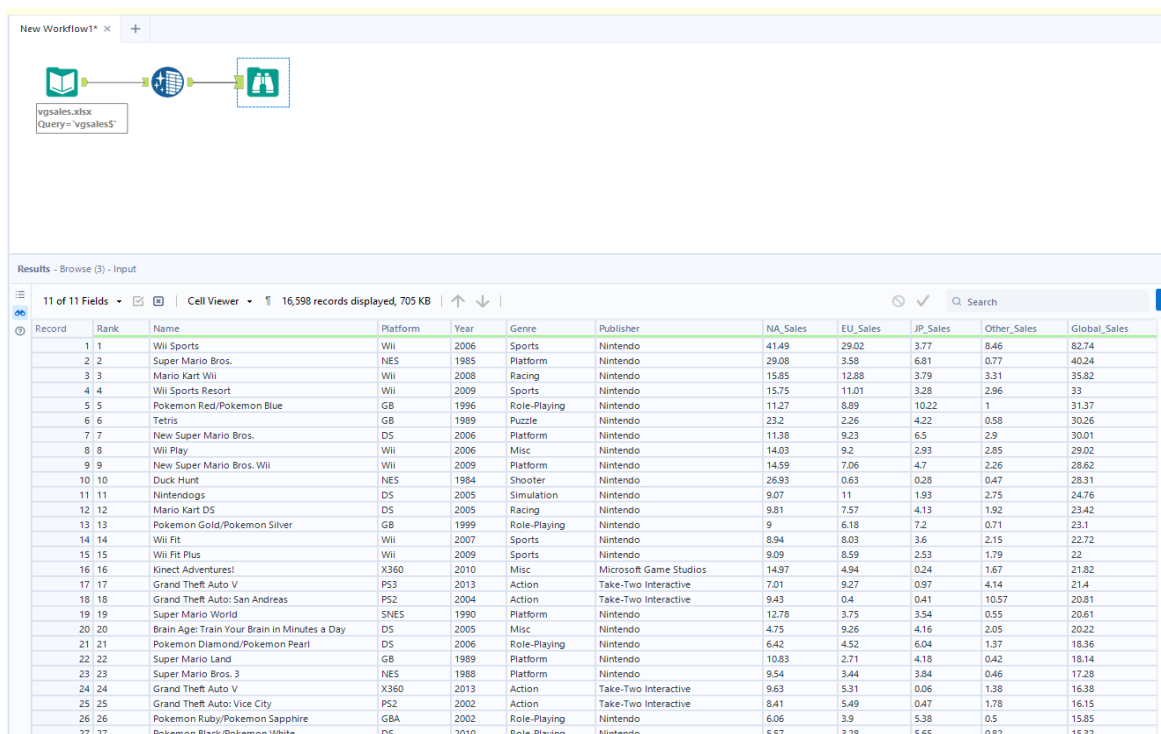
- The data that is going to be used in this report came from [Kaggle.com](https://www.kaggle.com). The dataset that I chose for this report is the sales data for video game sales.
- Before doing any pre-analysis, we have to understand the significant attributes or data types that we're going to deal with. This dataset contains 11 attributes that will remain relevant for this report. These data types are:
 - Rank - Ranking of overall sales, integer
 - Name - The games name
 - Platform - Platform of the games release (i.e. PC, PS4, etc.), object
 - Year - Year of the game's release, float
 - Genre - Genre of the game, object
 - Publisher - Publisher of the game, object
 - NA_Sales - Sales in North America (in millions), float
 - EU_Sales - Sales in Europe (in millions), float
 - JP_Sales - Sales in Japan (in millions), float
 - Other_Sales - Sales in the rest of the world (in millions), float
 - Global_Sales - Total worldwide sales, float
- Ranks help compare and analyse data points' relative position or importance, identify top or bottom values, find percentiles, detect outliers, and perform statistical analyses.
- Names serve as labels or identifiers for specific data elements, helping to organize, analyse, and understand the information contained in datasets or within a program.
- Each platform has unique technical demands, performance considerations, and user interface design principles. Moreover, they may offer specific features, online services, and multiplayer capabilities that impact gameplay experiences and social interactions within games.
- Year is the unit of time measurement, but in this context, year is the release dates of the games.
- Genres are categories that group games by their mechanics, objectives, and themes. Some popular genres include action, adventure, role-playing, strategy, sports, and simulation. Gamers use these genres to know what kind of experience they can expect from a particular game.
- Video game publishers manage the creation, production, and distribution of games. They may give money, help with marketing, and support game developers. They also handle getting the games to consumers, whether it's in a physical store or online.
- Sales in general terms is Persuading customers to purchase products or services in exchange for payment, driving business growth.

Data Pre-processing

- The BI tools that I used for this analysis are
 - Python
 - Alteryx
 - Power BI
 - MongoDB
- The data processing that I will be doing are in this order:
 - Data Cleansing
 - Data Import
 - NoSQL Queries
 - Data modelling
 - Data visualisation

Data Cleansing:

- The data cleansing process is one of the essential process before doing an in-depth analysis for the dataset. We have to exclude outliers and anomalies such as null values to avoid statistical errors. Using alteryx which is a BI (Business Intelligence) tool to modify and analyse data.



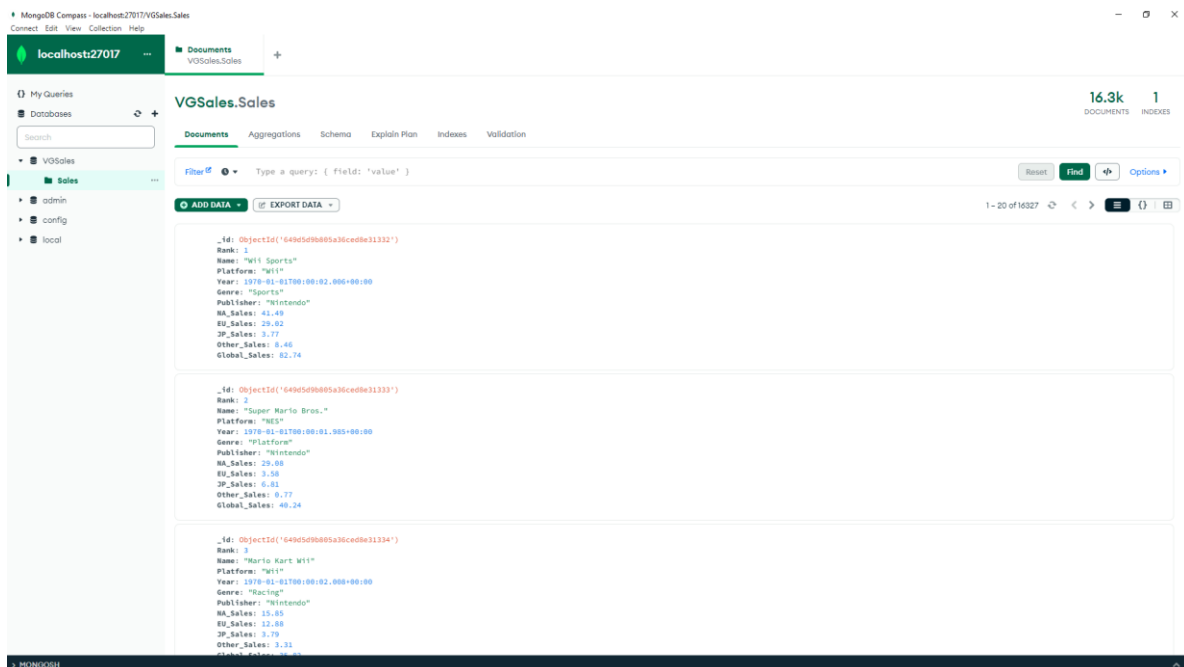
The screenshot shows the Alteryx interface. At the top, a workflow is visible with a data source icon labeled 'vgsales.xlsx' and a query 'Query = vgsales\$'. Below the workflow, the 'Results - Browse (3) - Input' section displays a table with 11 fields. The table contains 27 records of video game data, including columns for Rank, Name, Platform, Year, Genre, Publisher, and various sales figures (NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales).

Record	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
5	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
6	6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
7	7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
8	8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
9	9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
10	10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
11	11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76
12	12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
13	13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1
14	14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
15	15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22
16	16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.67	21.82
17	17	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.14	21.4
18	18	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.4	0.41	10.57	20.81
19	19	Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61
20	20	Brain Age: Train Your Brain in Minutes a Day	DS	2005	Misc	Nintendo	4.75	9.26	4.16	2.05	20.22
21	21	Pokemon Diamond/Pokemon Pearl	DS	2006	Role-Playing	Nintendo	6.42	4.52	6.04	1.37	18.36
22	22	Super Mario Land	GB	1989	Platform	Nintendo	10.83	2.71	4.18	0.42	18.14
23	23	Super Mario Bros. 3	NES	1988	Platform	Nintendo	9.54	3.44	3.84	0.46	17.28
24	24	Grand Theft Auto V	X360	2013	Action	Take-Two Interactive	9.63	5.31	0.06	1.38	16.38
25	25	Grand Theft Auto: Vice City	PS2	2002	Action	Take-Two Interactive	8.41	5.49	0.47	1.78	16.15
26	26	Pokemon Ruby/Pokemon Sapphire	GBA	2002	Role-Playing	Nintendo	6.06	3.9	5.38	0.5	15.85
27	27	Pokemon Black/Pokemon White	DS	2010	Role-Playing	Nintendo	5.57	3.28	5.65	0.82	15.32

- This process did not take a lot of time since the dataset does not have too much anomalies or outliers that can disturb the dataset.

- The process was executed in this order:
 1. Using the input data tool to import the csv file and giving the system the time to pre-analyse the data for outliers and anomalies.
 2. Afterwards, I also used the data cleansing tool to change and remove some null values from one of the tables in the dataset.
 3. For the final process I applied the browse tool to summarise the cleansed dataset. I consider the data to be clean based on the screenshot above.

Data Queries

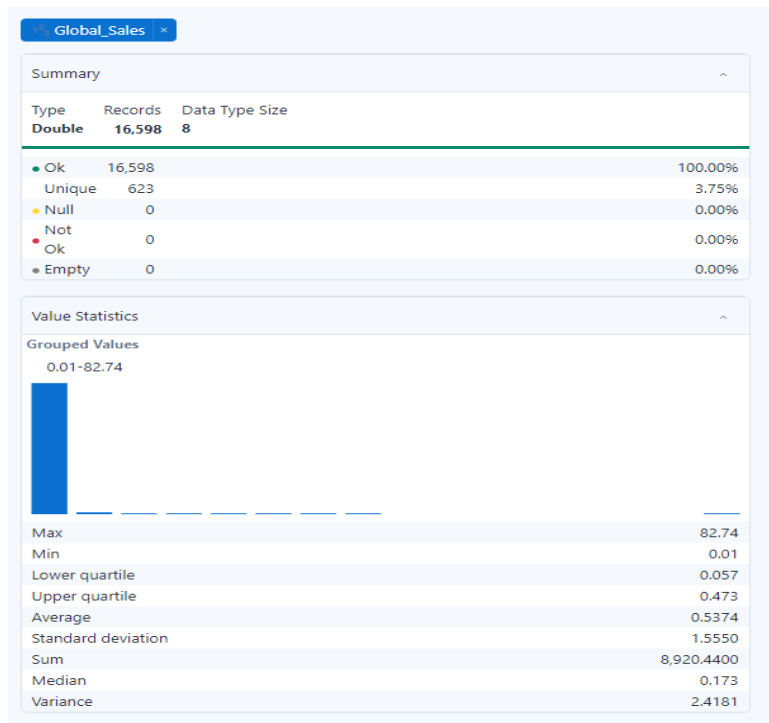


- I only manage to do 1 query that can help me gain insights from the data. On this phase, I used mongodb to create a data warehouse and create nosql queries. Unfortunately my compass DB is not cooperating with my system, meaning that I could not make any data warehouse as well as model.

1. Create a query that will show the most successful publisher.
 - a. `db.Publisher.aggregate`
`([{ $match: { Global_Sales: { $gt: 20 } } }, { $group: { _id: "$Publisher", total: { $sum: "$Global_Sales" } } }, { $sort: { total: -1 } }])`

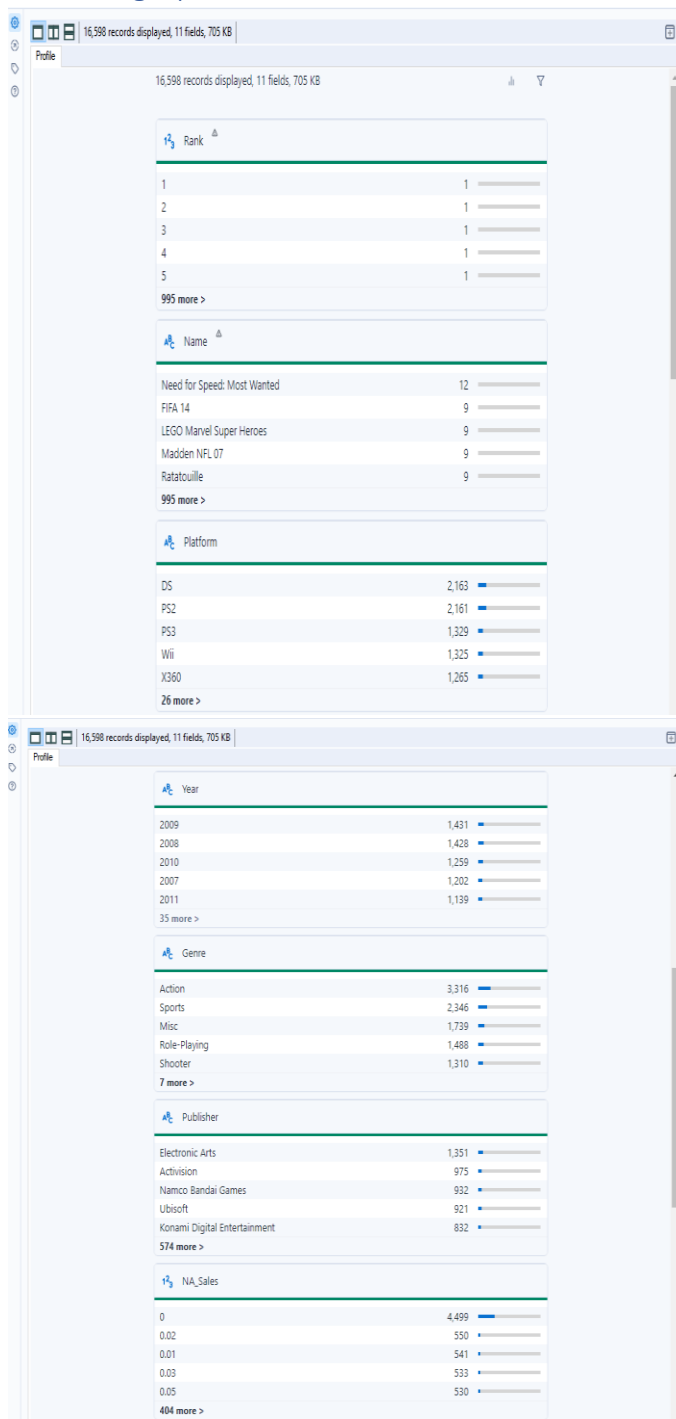
Data Visualisation

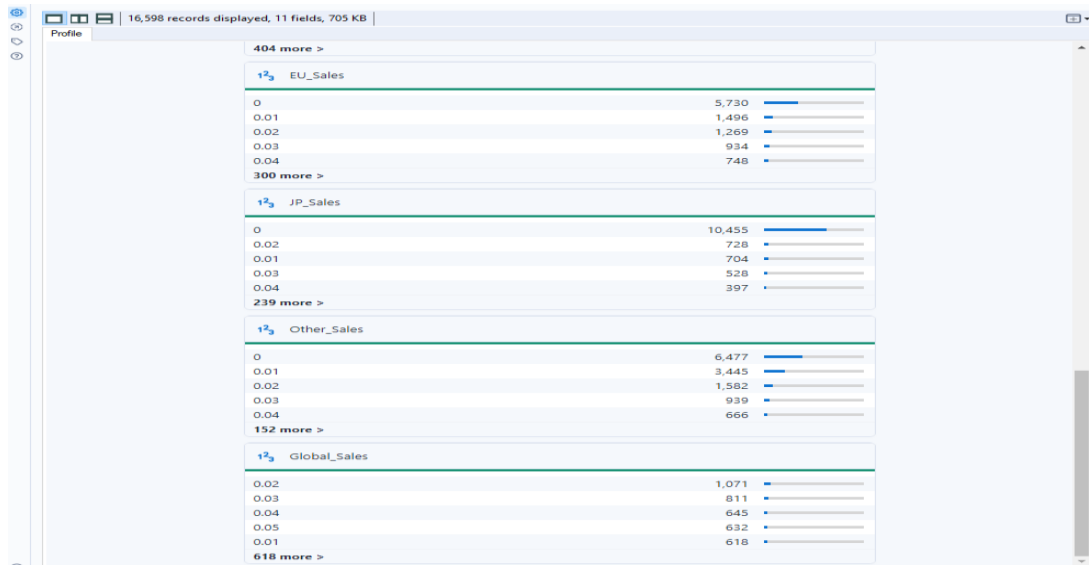
- Data visualisation is a very important process for this report, since we cannot just conclude an analysis with just numbers and letters, we have to use visuals and different models to justify the outcome of the analysis.



-
- With the help of Alteryx, a summarised data can be shown easily and can potentially be used for the analysis phase.
 - Maxed value in global sales - 82.74 – (millions)
 - Min value in global sales - 0.01 – (millions) – 10,000 sales
 - Upper quartile - 0.473
 - Lower quartile - 0.057
 - SD - 1.5550
 - Sum - 8.920.4400
- This summarised data can be used to project profitable sales in the video game industry. We can also see that there are set backs in terms of sales based on the min value for the global sales but to get an in-depth insight we have to use power BI for more visualisations.

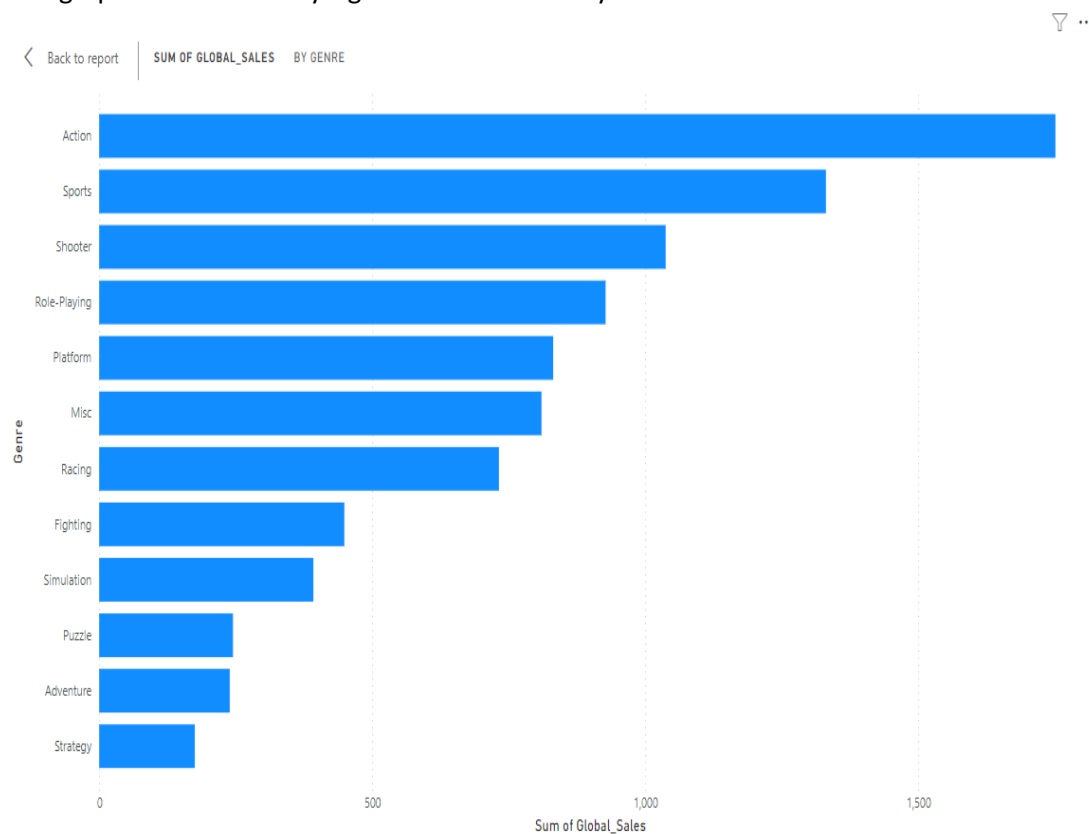
Additional data graphs





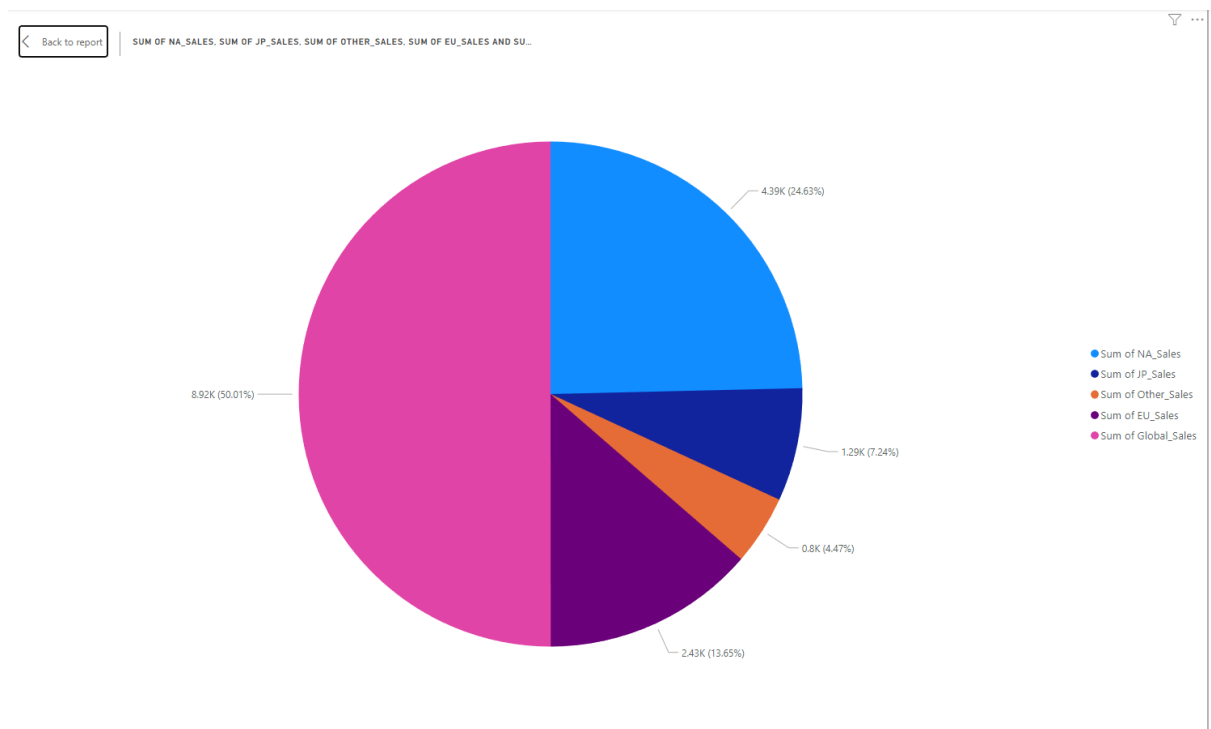
Power BI

- Using this dataset, I managed to use different BI tools for the analysis phase.
- The graphs below are very significant for the analysis of the data.

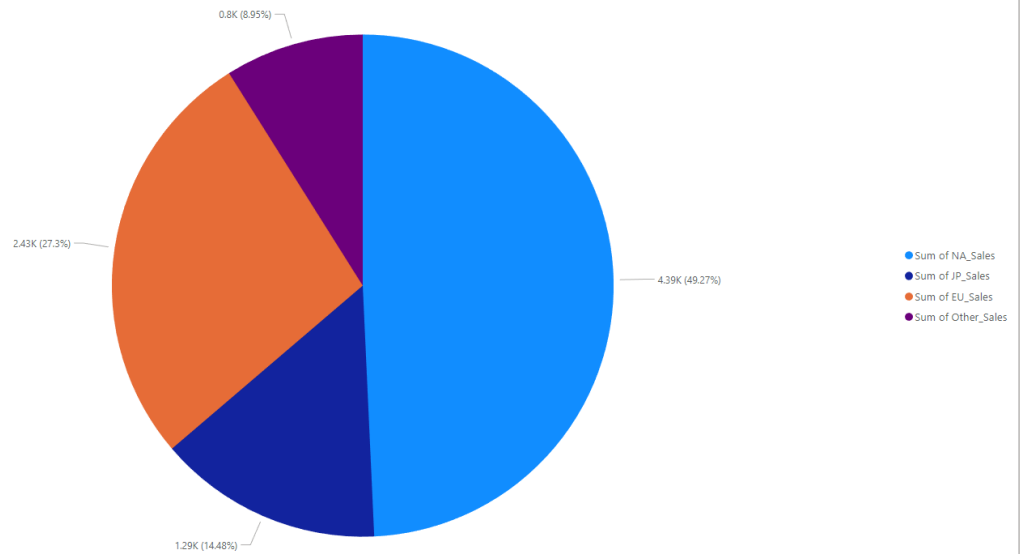


- Based on the graph above, we first have to understand on what this graph is telling us. The data has been organised based on genre as the y-value and global sales as the x-value.

- We can tell from the graph that Action genre has a total 1,751 (millions) in sales which is the highest sales of video game genre.
- And the lowest global sales in video games based on genre is strategy which has a 175.12 (millions) in sales.
- We can say that the genre of video games has a significant part of the sales, most of the people prefer more action games rather than strategy games. We have to consider that the users have different aspects and preferences with their genre choices. And we can tell based on this example.



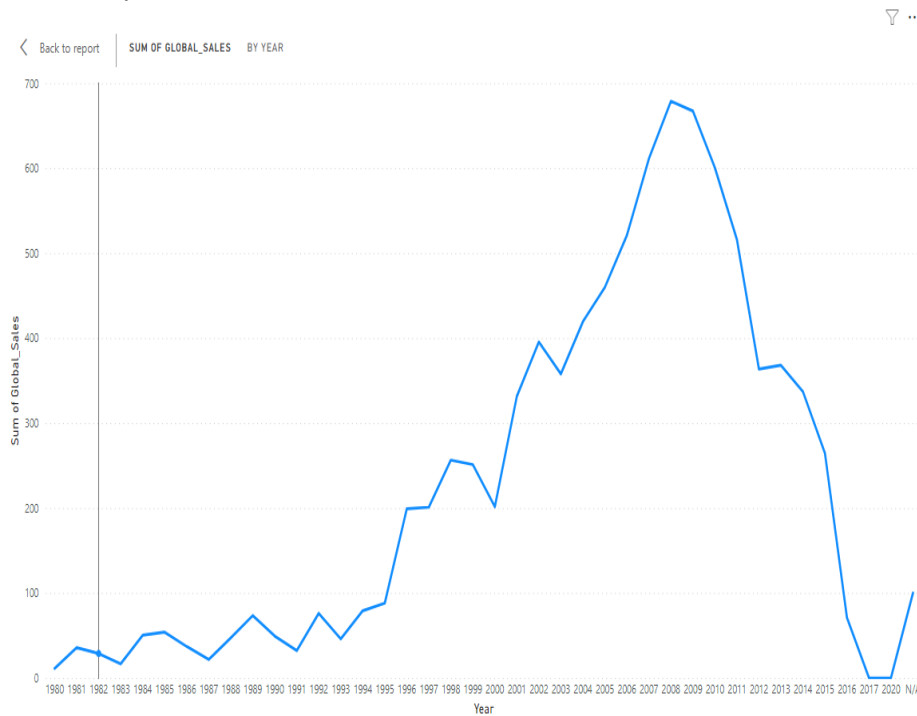
- This pie chart tells that the statistical portions of sales compared based on regions this also includes the global sales, but to accurately create a good analysis for the pie chart we have to exclude the global sales.



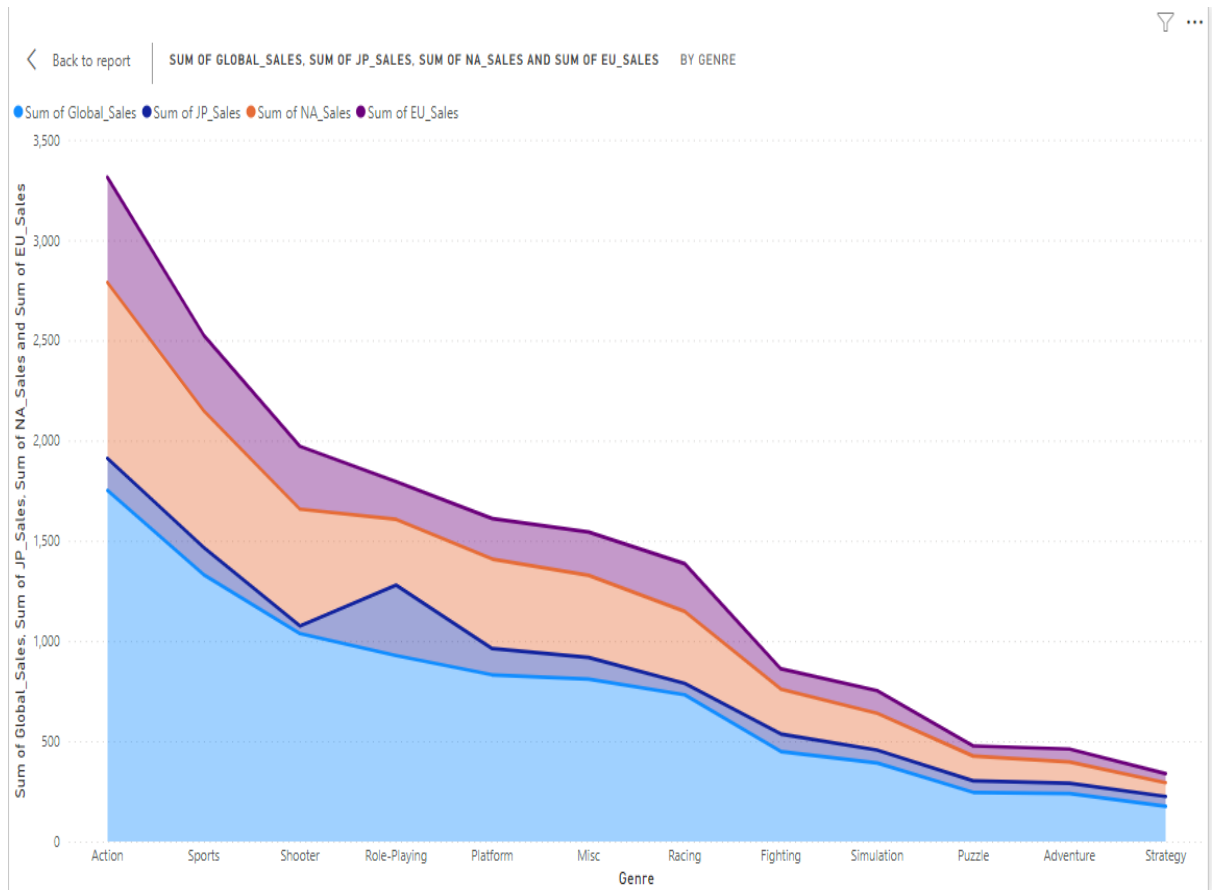
-
- This data is an accurate representation of the video game sales that is based on region.
- We can see that NA sales has a high stakes rating sale of 49.27% sales data compared to JP, EU and Other countries.
- 27.3% sales from EU, 14.48% from Japan and 9.95% from other countries.
- We can see a potential opportunity to raise the production rate of video games in NA to expect more profitable income. Targeting NA for sales may be helpful for the sale's profitable income.

Statistical Inaccuracy

- Doing more research regarding the video game sales data, I found out that the sales for this industry is increasing yearly. But when I ran a line chart for the data set I saw a statistical inaccuracy that this data dictates.

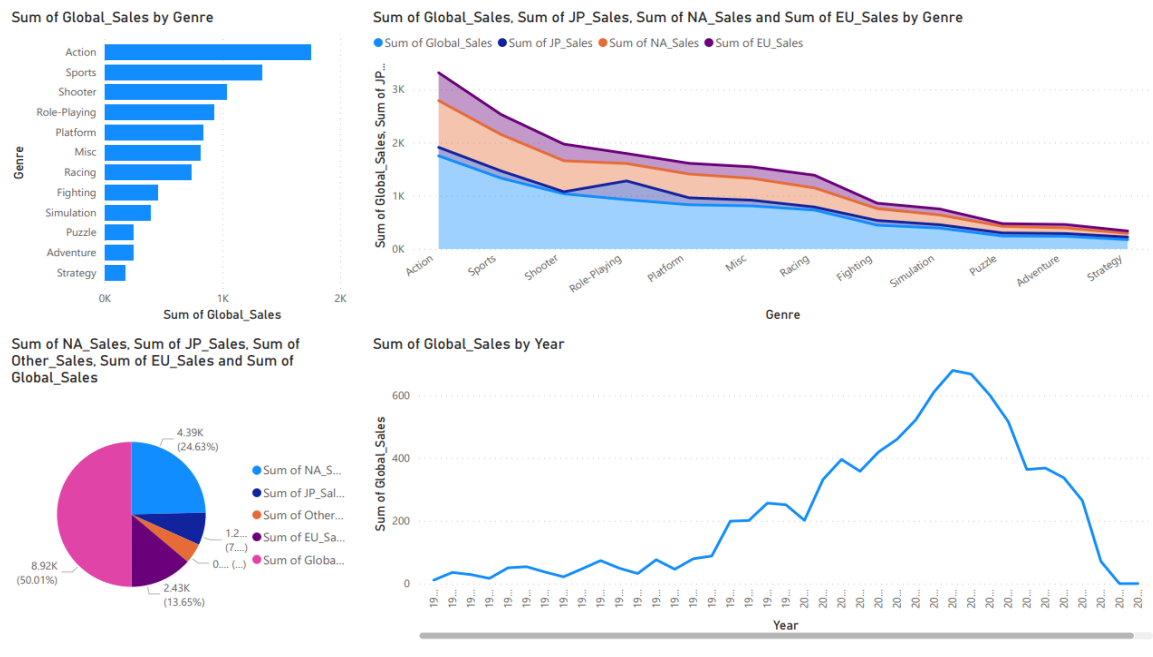


- This data shows a drop of sales between the year 2009 and 2020. This statistical error have to be considered based on the fact that most of the research online, represents an increase of an estimated 35% rate from the year 2022 to 2027. This predictive analysis is cited from statista.com. ("Global video game revenue 2027," 2022).
- Summarising this inaccuracy this shows that from 2009 sales to 2020 has a total of 667.01 (millions) decrease in sales.



- This data shows the regional comparison sales between the chosen countries and as well the total or global sales. We can see that NA sales has a really large area of sales compared to the other countries.
- But in terms of genre for the role-playing video games, we can observe that japan is quite interested in that kind of genre. Meaning that we can increase role-playing games in that region or develop games that are likely to be played for that certain region.

Conclusion



- The data that I have gathered creates a different view in terms of the gaming media industry. Based on the data that I have analysed, I can conclude that investing in this type of industry can be very beneficial. This data shows a good insight regarding the genres of games that people are interested in.
- In terms of the business side of the gaming industry, we can tell that doing more research for this particular sale can convince people to invest in the video game industry. Meaning that we can create more line for the industry to be more productive in sales while maintaining the code of conduct of the company.

REFERENCES:

50 YEARS OF GAMING HISTORY, BY REVENUE STREAM (1970-2020). (2022, SEPTEMBER 2). VISUAL CAPITALIST. <https://www.visualcapitalist.com/50-years-gaming-history-revenue-stream/>

GLOBAL VIDEO GAME REVENUE 2027. (2022, NOVEMBER 11). STATISTA. <https://www.statista.com/statistics/1344668/revenue-video-game-worldwide/>

US VIDEO GAME SALES WERE SOMEHOW EVEN HIGHER IN Q2 THAN LAST YEAR. (2021, JULY 22). GAMESPOT. <https://www.gamespot.com/articles/us-video-game-sales-were-somehow-even-higher-in-q2-than-last-year/1100-6494299/>