BSD 2333 DATA WRANGLING

"Be a Data Wrangler"

2022/2023 SEMESTER II

TITLE: VIDEO GAMES SALES AND RATINGS GROUP NAME: GO GO POWER WRANGLERS !		
MATRIC ID	**NAME**	**SECTION**
SD21031	TAN CHEK CHENG	01G
SD21037	WONG ZI MING	01G
SD21024	LIM JING ROU	01G
SD21061	OOI ZI YING	01G
SD21013	MITRAA A/P KOLANTHAI	01G

Table of Contents

1.0 Synopsis

1.1 Description of the assignment

In this assignment, we will utilize the "Video Game Sales and Ratings" dataset, which contains information about various columns such as Game_Name, Platform, Year_of_Release, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count, and Rating. These columns provide essential data that will be instrumental in uncovering valuable insights related to publishers, platforms, game genres, and user satisfaction.

To maximize the partnership opportunities, we will analyze the Publisher column. By examining the publishers' information, we can identify companies that have a strong track record of successful game releases. Visualizations such as bar charts and treemaps will allow us to showcase the distribution of games across different publishers and platforms. We can also explore the relationship between publishers and their corresponding sales figures, enabling us to identify potential partnership prospects.

To evaluate game performance across genres, we will utilize the Genre column along with the sales data columns (NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global_Sales). Through data visualization techniques such as bar graphs and stacked bar charts, we can analyze the sales performance of different genres in various regions. This analysis will provide insights into which genres are most popular and financially successful, helping us identify lucrative areas for game development and investment.

To understand user satisfaction on different games, we will examine the Critic_Score, Critic_Count, User_Score, and User_Count columns. By comparing critic scores and user scores, we can visualize the alignment or discrepancies between the evaluations made by critics and the perceptions of users. Scatter plots and box plots can be used to depict these relationships and help us understand the factors that contribute to high user satisfaction. Additionally, we can explore the relationship between user satisfaction and sales figures, identifying games that have achieved both critical acclaim and commercial success.

Throughout the analysis, we will employ various visualization techniques to present the data in a visually engaging and informative manner. Visualizations such as bar charts, treemaps, scatter plots, and box plots will allow us to interpret the dataset more effectively and derive meaningful insights. These visual representations will aid in identifying trends, patterns, and relationships, facilitating decision-making processes and providing a comprehensive understanding of the gaming industry.

By leveraging the insights obtained from analyzing the dataset and employing visualization techniques, we aim to achieve the objectives of maximizing partnership opportunities, evaluating game performance across genres, and understanding user satisfaction. The interpretation of these insights will guide strategic decision-making, foster successful collaborations, highlight thriving game genres, and pave the way for the creation of user-centric gaming experiences. Ultimately, this assignment will contribute to a deeper understanding of the gaming industry's dynamics and help stakeholders navigate this ever-evolving landscape.

1.2 Problems to be solved

The review of publishers and platforms, performance evaluation by genres, and user satisfaction analysis posed certain challenges for us.

Determining criteria for selecting prospective partners, analysing performance and market presence, and aligning with project goals are some of the issues in publishers and platform analysis for partnership identification. The video game industry's dynamic nature adds complexity and needs regular monitoring and analysis to facilitate possibilities to cooperate with it. Furthermore, establishing the link between publishers and platforms can be difficult, particularly when many platforms are involved. Continuous investigation is required to comprehend the rapidly changing environment and the influence of market dynamics on publishers and platforms. As can be noticed, our research will concentrate on the interaction between publishers and partners in order to gain the capacity to collaborate with them and enhance video game sales.

The difficulties of genre performance evaluation of video games include selecting appropriate statistical methods, doing comparative analysis, and employing appropriate performance indicators. It is difficult to obtain trustworthy and comprehensive genre-specific sales information. As a result, we must examine the outcomes of genre studies and comparisons while also considering market trends, platform popularity, marketing efforts, and genre-specific preferences. This might help game developers determine the direction and type of video games in the future. To contribute in enhancing performance evaluation by genre, we must prioritise the use of precise data and appropriate statistical techniques to examine and explain performance evaluations of video games by genres.

Comparing critic and user scores, analysing inconsistencies, and finding important factors are all problems encountered in user satisfaction study. Due to linguistic details and personality, conducting sentiment analysis on user evaluations may be difficult. The criteria for user ratings and review systems are not consistent across platforms and websites. The lack of openness in user scores, as well as restricted information on calculating methods and user demographics, complicate matters. For appropriate analysis, we have to normalise the rating scales and synthesize the data from numerous sources. Overall, in user satisfaction analysis, normalising user reviews and analysing inconsistencies between critic and user scores while identifying important elements are critical.

1.3 Questions to be answered

The dataset must provide answers to the following three questions:

1. Which platform and publisher sold the most console video games?
2. Which game genre has the most sales?
3. Which video games on consoles have the greatest user ratings for the platform in question

1.4 Objectives

1. To analyse the publishers and platforms information for maximize the partnerships opportunities.
2. To study the performance evaluation of games by genres with utilizing sales data.
3. To understand the user satisfaction by evaluating the critic scores and user scores.

1.5 Basic description of the data

The Video_Games_Sales_and_Rating dataset includes sales data for various platform video games from 1976 to 2016 in countries such as North America, Europe, Japan, and others. This data set also includes a list of video games that have sold more than 100,000 copies, as well as reviewer and user ratings. For most games, it is a combination of online scrapes from VGChartz and Metacritic, as well as personally inserted year of release values. There are 15 columns containing 17416 data in this dataset.

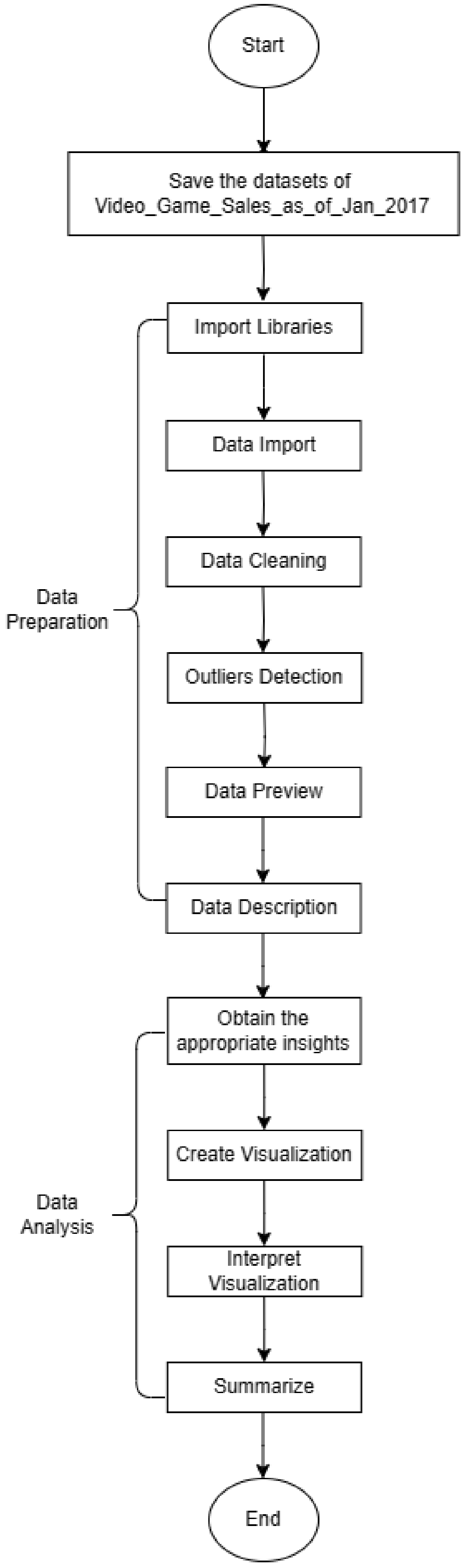
Attributes	Explanation	Data Type
Name	The game's name	String
Platform	Platform of the games release	String
Year_of_Release	Year of the game's release	Date
Genre	Genre of the game	String
Publisher	Publisher of the game	String
NA_Sales	Sales in North America (in millions)	Numeric (float)
EU_Sales	Sales in Europe (in millions)	Numeric (float)
JP_Sales	Sales in Japan (in millions)	Numeric (float)
Other_Sales	Sales in the rest of the world (in millions)	Numeric (float)
Global_Sales	Total worldwide sales (in millions)	Numeric (float)
Critic_score	Aggregate score compiled by Metacritic staff	Numeric (integer)
Critic_count	The number of critics used in coming up with the critic score	Numeric (integer)
User_score	Score by Metacritic's subscribes	Numeric (float)
User_count	Number of users who gave the user score	Numeric (integer)
Rating	The ESRB ratings • E = Everyone • E10+ = Everyone 10+ • T = Teen • M = Mature 17+ • AO = Adults Only 18+ • RP/RP Likely Mature 17+ = Rating Pending	String

2.0 Packages Required

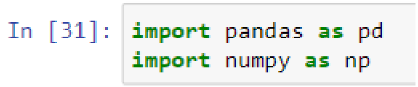
Package/Module	Purpose incompleting the project
pandas	A powerful data manipulation and analysis library. It is used for reading CSV data, such as the 'Video_Game_Sales_as_of_Jan_2017.csv' file in our project, into a DataFrame. The DataFrame is a tabular data structure that enables efficient data handling and analysis operations.
numpy	A fundamental package for scientific computing in Python. It offers support for high-performance, multi-dimensional arrays and a wide range of mathematical operations. In our project, it aids in calculating bar widths, generating angular positions, and converting rotation angles for radial column chart visualisaiton.
seaborn	An advanced statistical data visualization library built on top of Matplotlib. It provides a high-level interface for creating aesthetically pleasing and informative statistical graphics. Our project utilizes Seaborn to produce various visualizations, such as box plots, bar charts, treemaps, and more.
matplotlib.pyplot	A comprehensive plotting library that resembles MATLAB's plotting capabilities. It allows the creation of diverse types of plots, including box plots and bar charts, facilitating data visualization. In our project, it is utilized for chart generation and customization.
altair	A declarative statistical visualization library. It allows us to create interactive and visually pleasing charts by defining visual encodings and interactions using a concise syntax, such as scatter plots that showcase the relationship between critic scores, user scores, and genres.
squarify	A treemap plotting library. It is used to create rectangular treemaps, where the area of each rectangle represents a quantitative value, allowing easy visualization of hierarchical data structures. In our project, it assists in generating visually informative treemaps.

3.0 Data Preparation

3.1 Flow Chart

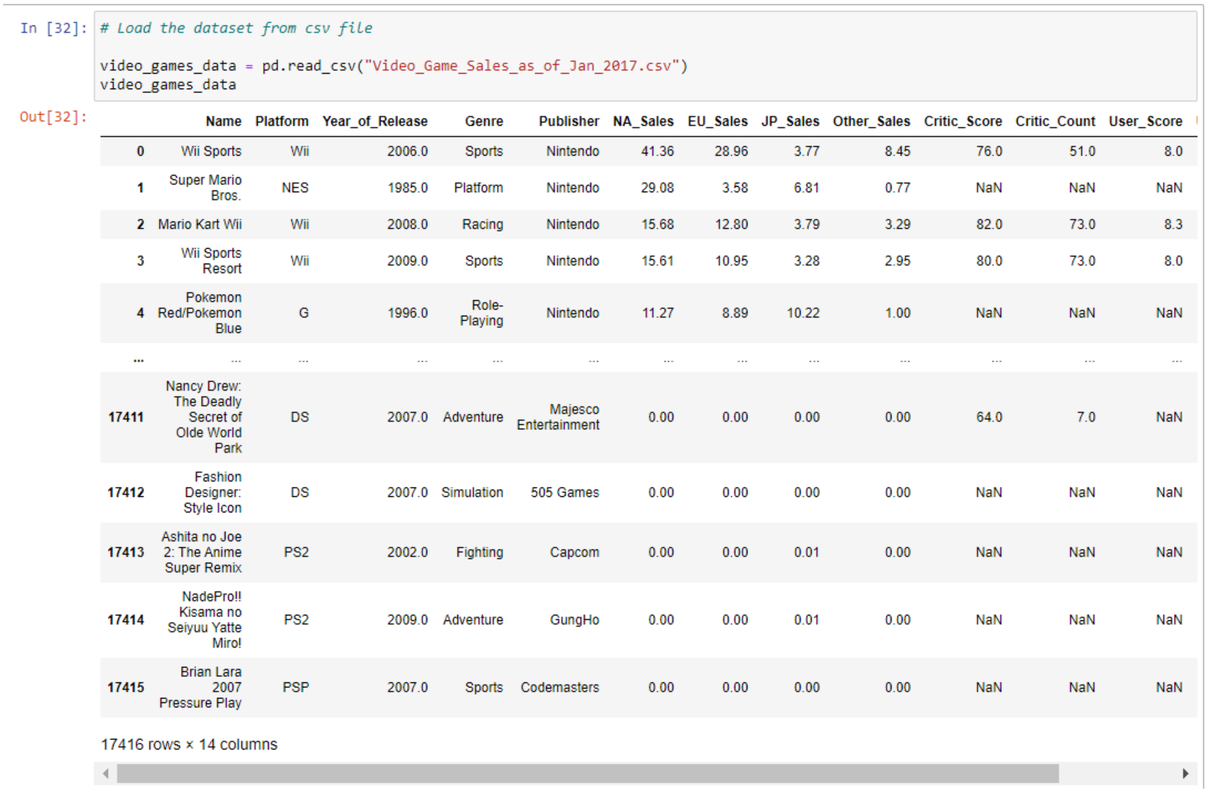


3.2 Data Import



```
In [31]: import pandas as pd
         import numpy as np
```

The provided code snippet is importing two essential libraries in Python, pandas and numpy, which are widely used for data manipulation and analysis tasks. The line "import pandas as pd" imports the pandas library and assigns it the alias "pd." By using an alias, we can refer to the library using a shorter name throughout the code, enhancing readability and reducing typing effort. Similarly, "import numpy as np" imports the numpy library and assigns it the alias "np." By importing these libraries, we gain access to their functionalities and can leverage them for various data processing tasks in our code.



```
In [32]: # Load the dataset from csv file

         video_games_data = pd.read_csv("Video_Game_Sales_as_of_Jan_2017.csv")
         video_games_data
```

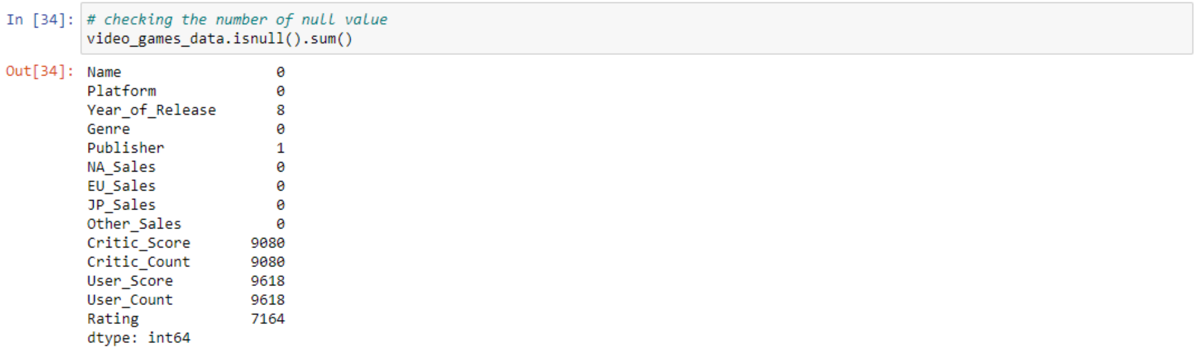
Out[32]:

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Critic_Score	Critic_Count	User_Score
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	76.0	51.0	8.0
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	NaN	NaN	NaN
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.80	3.79	3.29	82.0	73.0	8.3
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.95	3.28	2.95	80.0	73.0	8.0
4	Pokemon Red/Pokemon Blue	G	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
17411	Nancy Drew: The Deadly Secret of Olde World Park	DS	2007.0	Adventure	Majesco Entertainment	0.00	0.00	0.00	0.00	64.0	7.0	NaN
17412	Fashion Designer: Style Icon	DS	2007.0	Simulation	505 Games	0.00	0.00	0.00	0.00	NaN	NaN	NaN
17413	Ashita no Joe 2: The Anime Super Remix	PS2	2002.0	Fighting	Capcom	0.00	0.00	0.01	0.00	NaN	NaN	NaN
17414	NadePro!! Kisama no Seiyuu Yatte Miro!	PS2	2009.0	Adventure	GungHo	0.00	0.00	0.01	0.00	NaN	NaN	NaN
17415	Brian Lara 2007 Pressure Play	PSP	2007.0	Sports	Codemasters	0.00	0.00	0.00	0.00	NaN	NaN	NaN

17416 rows × 14 columns

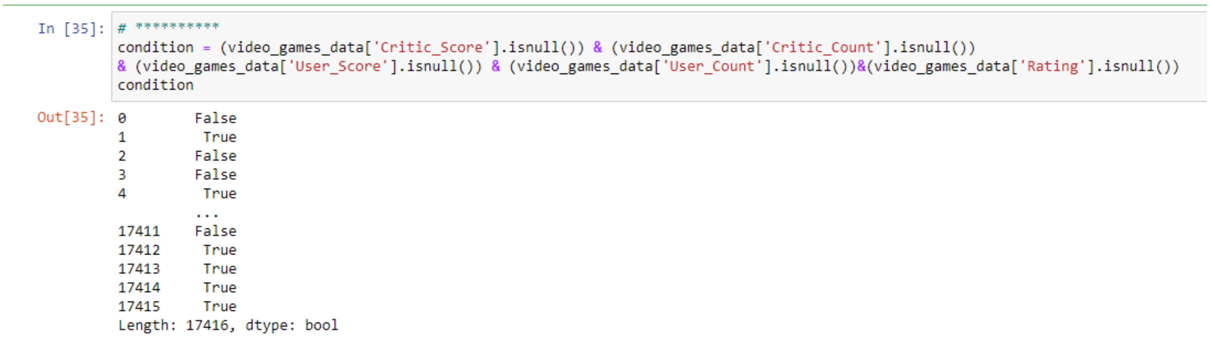
The provided code snippet is loading a dataset of video game sales into a pandas DataFrame using the read_csv function. By executing this code, we can examine the video game sales data stored in the "Video_Game_Sales_as_of_Jan_2017.csv" file, and further analyze and manipulate the data.

3.3 Data Cleaning



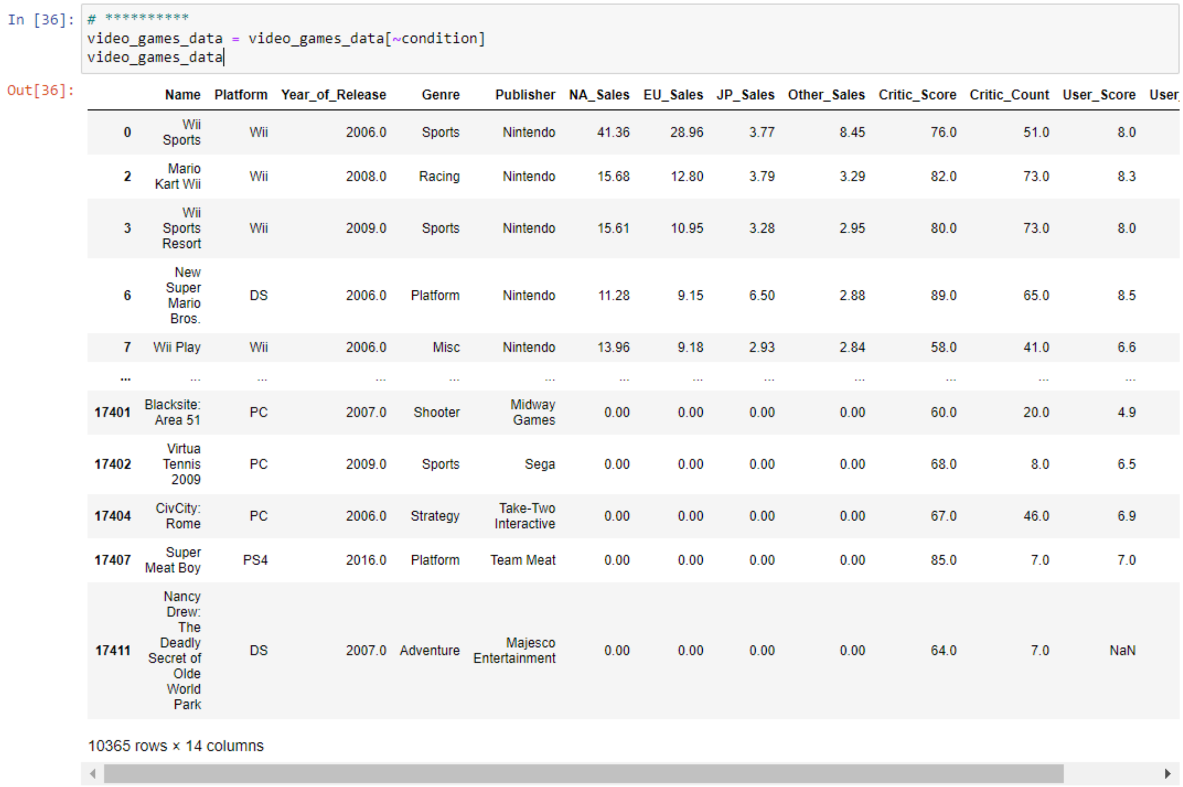
```
In [34]: # checking the number of null value
         video_games_data.isnull().sum()

Out[34]: Name                  0
         Platform              0
         Year_of_Release       8
         Genre                 0
         Publisher             1
         NA_Sales              0
         EU_Sales              0
         JP_Sales              0
         Other_Sales           0
         Critic_Score       9080
         Critic_Count       9080
         User_Score         9618
         User_Count         9618
         Rating             7164
         dtype: int64
```

The code video_games_data.isnull().sum() is used to count the number of missing values in each column of the video_games_data DataFrame. It helps in identifying the amount of missing data present in the dataset.



```
In [35]: # **********
         condition = (video_games_data['Critic_Score'].isnull()) & (video_games_data['Critic_Count'].isnull())
         & (video_games_data['User_Score'].isnull()) & (video_games_data['User_Count'].isnull())&(video_games_data['Rating'].isnull())
         condition

Out[35]: 0         False
         1          True
         2         False
         3         False
         4          True
                   ...
         17411     False
         17412      True
         17413      True
         17414      True
         17415      True
         Length: 17416, dtype: bool
```

In the process of cleaning, we found that these five columns are usually empty together so we apply the code above,The code creates a condition to identify rows in the video_games_data DataFrame that have missing values in these specific five columns. The resulting condition variable will be True for rows that satisfy the condition and have missing values in all specified columns, and False for rows that do not meet this criterion.



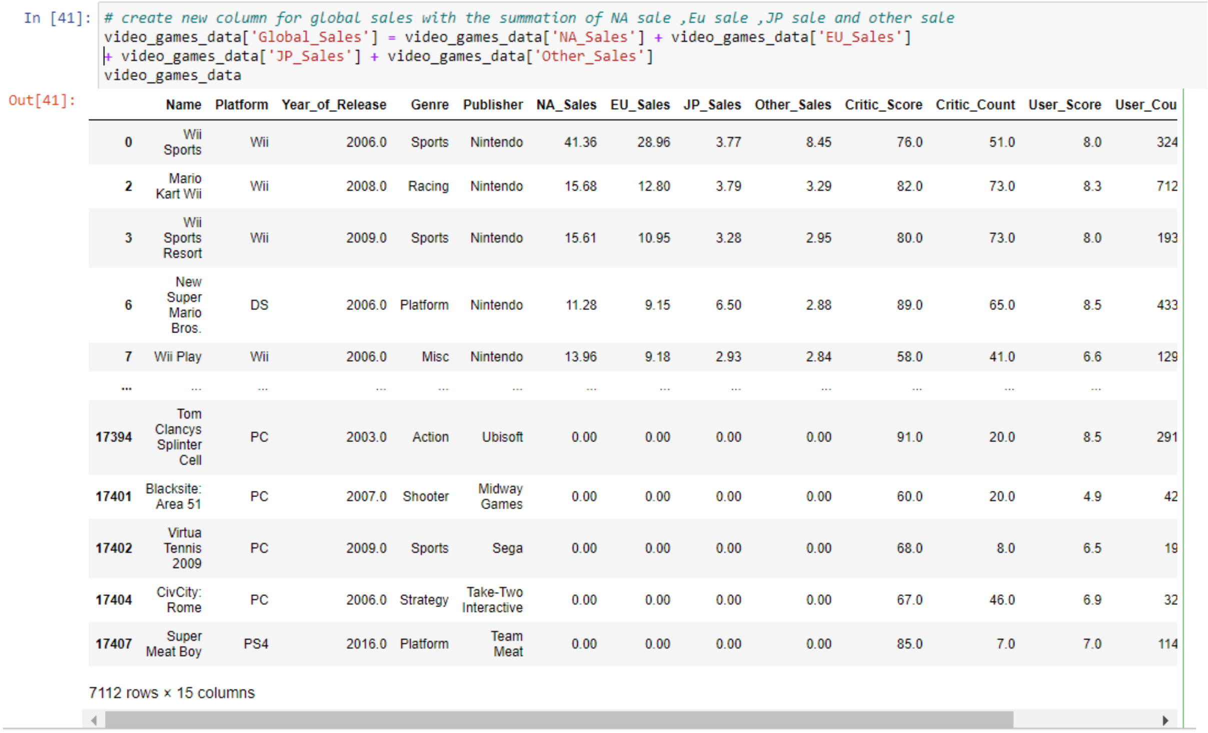
In [36]:
```
# **********
video_games_data = video_games_data[~condition]
video_games_data
```

Out[36]:

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Critic_Score	Critic_Count	User_Score	User
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	76.0	51.0	8.0	
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.80	3.79	3.29	82.0	73.0	8.3	
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.95	3.28	2.95	80.0	73.0	8.0	
6	New Super Mario Bros.	DS	2006.0	Platform	Nintendo	11.28	9.15	6.50	2.88	89.0	65.0	8.5	
7	Wii Play	Wii	2006.0	Misc	Nintendo	13.96	9.18	2.93	2.84	58.0	41.0	6.6	
...	...	...	...	...	...	...	...	...	...	...	...	...	
17401	Blacksite: Area 51	PC	2007.0	Shooter	Midway Games	0.00	0.00	0.00	0.00	60.0	20.0	4.9	
17402	Virtua Tennis 2009	PC	2009.0	Sports	Sega	0.00	0.00	0.00	0.00	68.0	8.0	6.5	
17404	CivCity: Rome	PC	2006.0	Strategy	Take-Two Interactive	0.00	0.00	0.00	0.00	67.0	46.0	6.9	
17407	Super Meat Boy	PS4	2016.0	Platform	Team Meat	0.00	0.00	0.00	0.00	85.0	7.0	7.0	
17411	Nancy Drew: The Deadly Secret of Olde World Park	DS	2007.0	Adventure	Majesco Entertainment	0.00	0.00	0.00	0.00	64.0	7.0	NaN	

10365 rows × 14 columns

The code video_games_data = video_games_data[~condition] filters the video_games_data DataFrame based on the condition defined earlier and assigns the filtered DataFrame back to the video_games_data variable. The "~" operator is used to negate the condition, selecting rows that do not satisfy the condition.



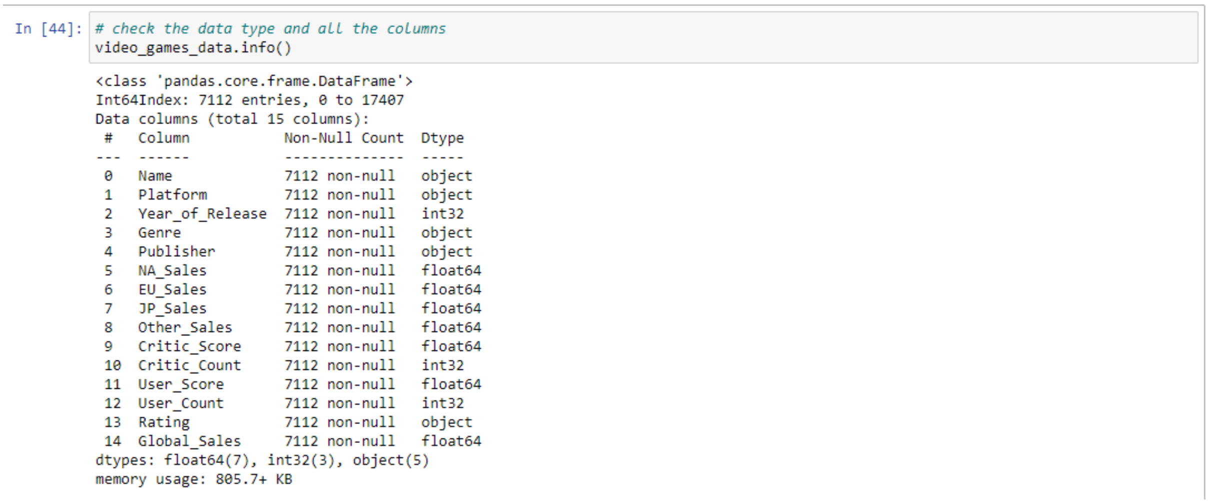
In [41]:
```
# create new column for global sales with the summation of NA sale ,Eu sale ,JP sale and other sale
video_games_data['Global_Sales'] = video_games_data['NA_Sales'] + video_games_data['EU_Sales']
+ video_games_data['JP_Sales'] + video_games_data['Other_Sales']
video_games_data
```

Out[41]:

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Critic_Score	Critic_Count	User_Score	User_Cou
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	76.0	51.0	8.0	324
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.80	3.79	3.29	82.0	73.0	8.3	712
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.95	3.28	2.95	80.0	73.0	8.0	193
6	New Super Mario Bros.	DS	2006.0	Platform	Nintendo	11.28	9.15	6.50	2.88	89.0	65.0	8.5	433
7	Wii Play	Wii	2006.0	Misc	Nintendo	13.96	9.18	2.93	2.84	58.0	41.0	6.6	129
...	...	...	...	...	...	...	...	...	...	...	...	...	
17394	Tom Clancys Splinter Cell	PC	2003.0	Action	Ubisoft	0.00	0.00	0.00	0.00	91.0	20.0	8.5	291
17401	Blacksite: Area 51	PC	2007.0	Shooter	Midway Games	0.00	0.00	0.00	0.00	60.0	20.0	4.9	42
17402	Virtua Tennis 2009	PC	2009.0	Sports	Sega	0.00	0.00	0.00	0.00	68.0	8.0	6.5	19
17404	CivCity: Rome	PC	2006.0	Strategy	Take-Two Interactive	0.00	0.00	0.00	0.00	67.0	46.0	6.9	32
17407	Super Meat Boy	PS4	2016.0	Platform	Team Meat	0.00	0.00	0.00	0.00	85.0	7.0	7.0	114

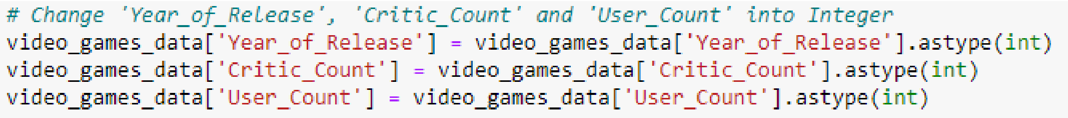
7112 rows × 15 columns

The code calculates the total global sales for different region sales like NA,JP ,EU and other in the video_games_data DataFrame by summing the sales from different regions and assigns the result to a new column named 'Global_Sales'. We need to new column for further analysis.



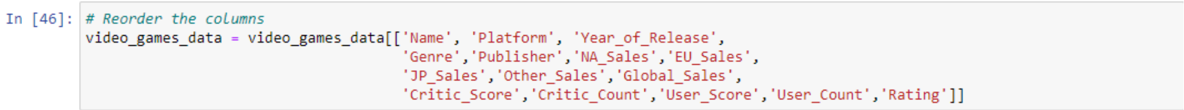
```
In [44]: # check the data type and all the columns
         video_games_data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 7112 entries, 0 to 17407
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Name             7112 non-null   object 
 1   Platform         7112 non-null   object 
 2   Year_of_Release  7112 non-null   int32  
 3   Genre            7112 non-null   object 
 4   Publisher        7112 non-null   object 
 5   NA_Sales         7112 non-null   float64
 6   EU_Sales         7112 non-null   float64
 7   JP_Sales         7112 non-null   float64
 8   Other_Sales      7112 non-null   float64
 9   Critic_Score     7112 non-null   float64
 10  Critic_Count     7112 non-null   int32  
 11  User_Score       7112 non-null   float64
 12  User_Count       7112 non-null   int32  
 13  Rating           7112 non-null   object 
 14  Global_Sales     7112 non-null   float64
dtypes: float64(7), int32(3), object(5)
memory usage: 805.7+ KB
```

The code succinct overview of the structure and information regarding the video_games_data DataFrame, with a particular focus on examining the data types of each attribute. This is crucial as it allows us to identify any data type inconsistencies and make the necessary corrections to ensure the desired data types are applied appropriately. By using this code, we can gain valuable insights into the DataFrame's composition, including the number of rows and columns, column names, and memory usage, thereby facilitating a thorough understanding of the dataset's structure and enabling effective data manipulation and analysis.



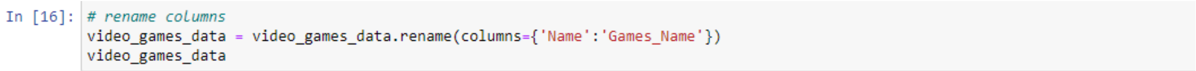
```
# Change 'Year_of_Release', 'Critic_Count' and 'User_Count' into Integer
video_games_data['Year_of_Release'] = video_games_data['Year_of_Release'].astype(int)
video_games_data['Critic_Count'] = video_games_data['Critic_Count'].astype(int)
video_games_data['User_Count'] = video_games_data['User_Count'].astype(int)
```

The code converts the data types of specific columns to integers using the .astype(int) method.



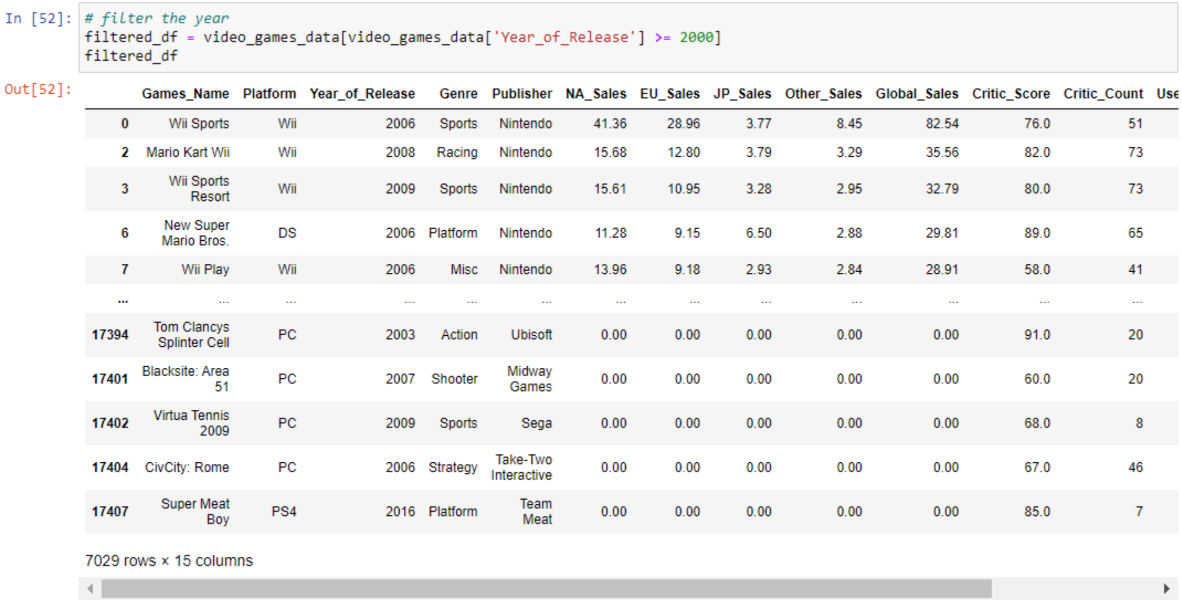
```
In [46]: # Reorder the columns
         video_games_data = video_games_data[['Name', 'Platform', 'Year_of_Release',
                                              'Genre','Publisher','NA_Sales','EU_Sales',
                                              'JP_Sales','Other_Sales','Global_Sales',
                                              'Critic_Score','Critic_Count','User_Score','User_Count','Rating']]
```

To enhance the clarity and comprehensibility of the DataFrame's overview, we reorder the columns in a more organized and intuitive manner. By rearranging the columns, we aim to provide a more coherent representation of the DataFrame that facilitates easier understanding and interpretation of the data.



```
In [16]: # rename columns
         video_games_data = video_games_data.rename(columns={'Name':'Games_Name'})
         video_games_data
```

By executing this code, the 'Name' column in the DataFrame is replaced with 'Games_Name', effectively renaming the column. This can be useful when the original column name is not descriptive enough or when we want to align the column names with a specific naming convention. Renaming columns can enhance the understanding and clarity of the DataFrame, making it easier to work with and interpret the data.



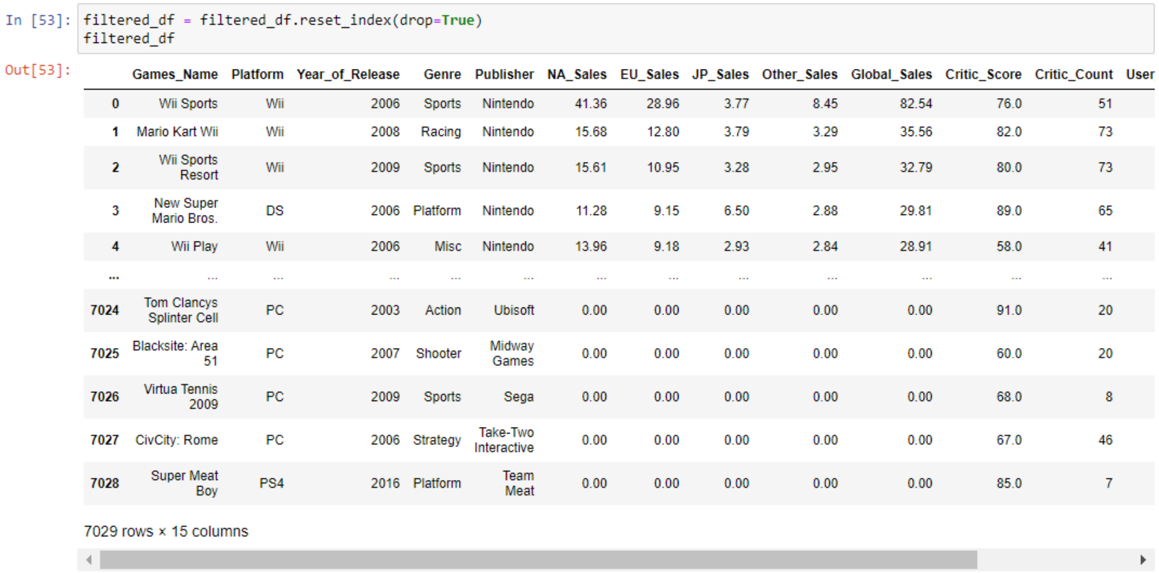
```
In [52]: # filter the year
         filtered_df = video_games_data[video_games_data['Year_of_Release'] >= 2000]
         filtered_df
```

Out[52]:

	Games_Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	Use
0	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.54	76.0	51	
2	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.80	3.79	3.29	35.56	82.0	73	
3	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.95	3.28	2.95	32.79	80.0	73	
6	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.28	9.15	6.50	2.88	29.81	89.0	65	
7	Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.91	58.0	41	
...	...	...	...	...	...	...	...	...	...	...	...	...	
17394	Tom Clancys Splinter Cell	PC	2003	Action	Ubisoft	0.00	0.00	0.00	0.00	0.00	91.0	20	
17401	Blacksite: Area 51	PC	2007	Shooter	Midway Games	0.00	0.00	0.00	0.00	0.00	60.0	20	
17402	Virtua Tennis 2009	PC	2009	Sports	Sega	0.00	0.00	0.00	0.00	0.00	68.0	8	
17404	CivCity: Rome	PC	2006	Strategy	Take-Two Interactive	0.00	0.00	0.00	0.00	0.00	67.0	46	
17407	Super Meat Boy	PS4	2016	Platform	Team Meat	0.00	0.00	0.00	0.00	0.00	85.0	7	

7029 rows × 15 columns

In our case, we are interested in games released from the year 2000 onwards.The resulting filtered_df DataFrame contains only the rows where the 'Year_of_Release' meets the condition, providing a subset of the original dataset that includes games released from the year 2000 onwards.



```
In [53]: filtered_df = filtered_df.reset_index(drop=True)
         filtered_df
```

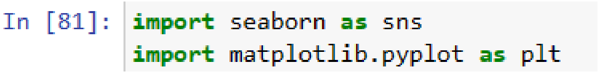
Out[53]:

	Games_Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User
0	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.54	76.0	51	
1	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.80	3.79	3.29	35.56	82.0	73	
2	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.95	3.28	2.95	32.79	80.0	73	
3	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.28	9.15	6.50	2.88	29.81	89.0	65	
4	Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.91	58.0	41	
...	...	...	...	...	...	...	...	...	...	...	...	...	
7024	Tom Clancys Splinter Cell	PC	2003	Action	Ubisoft	0.00	0.00	0.00	0.00	0.00	91.0	20	
7025	Blacksite: Area 51	PC	2007	Shooter	Midway Games	0.00	0.00	0.00	0.00	0.00	60.0	20	
7026	Virtua Tennis 2009	PC	2009	Sports	Sega	0.00	0.00	0.00	0.00	0.00	68.0	8	
7027	CivCity: Rome	PC	2006	Strategy	Take-Two Interactive	0.00	0.00	0.00	0.00	0.00	67.0	46	
7028	Super Meat Boy	PS4	2016	Platform	Team Meat	0.00	0.00	0.00	0.00	0.00	85.0	7	

7029 rows × 15 columns

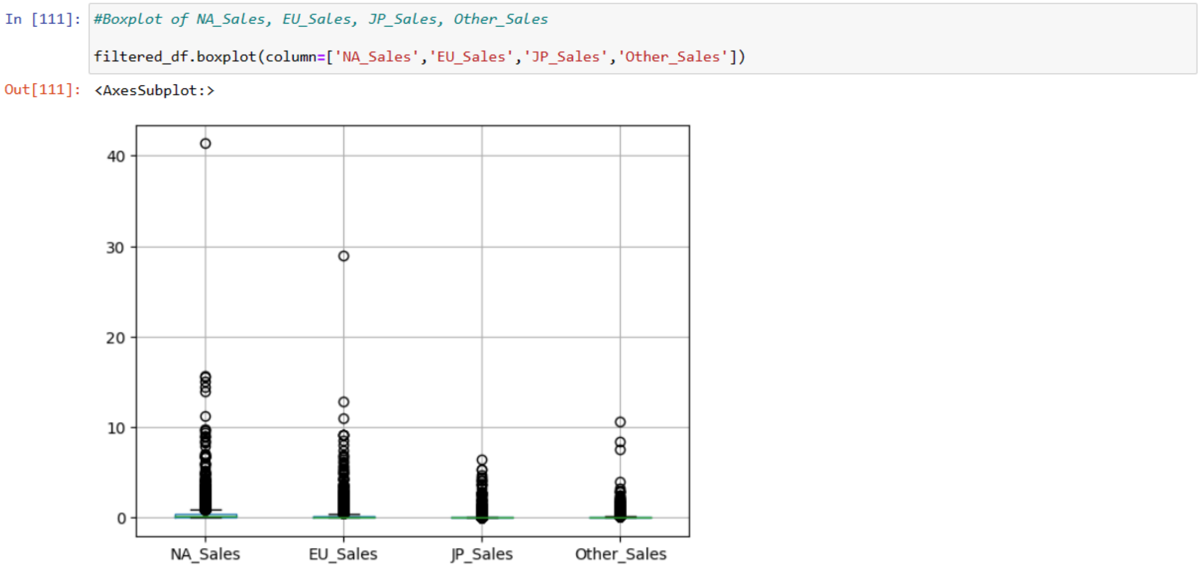
By executing this code, the index of the DataFrame is reset to a new sequential numeric index with starting from 0. This reindexing operation reassigns a new index to each row in the DataFrame, ensuring a consistent and ordered index representation after all the data cleaning process.

Outliers Detection



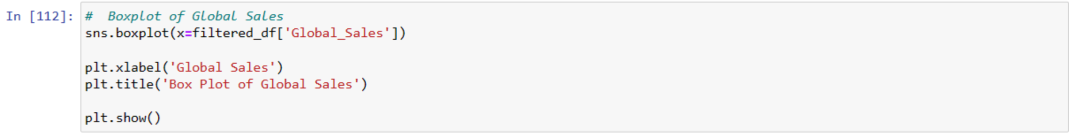
```
In [81]: import seaborn as sns
         import matplotlib.pyplot as plt
```

Furthermore, seaborn and matplotlib.pyplot are imported which are used for data visualization in Python. Seaborn is particularly useful for creating statistical graphics such as scatter plots, box plots, bar plots and heatmaps. It provides a higher-level API, allowing users to create visually appealing plots with fewer lines of code. Matplotlib.pyplot provides a wide range of functions for creating different types of plots and it offers extensive customization options that allow users to control various aspects of the plot. Hence, these libraries allow users to leverage the functionality and capabilities of them to generate visualizations.

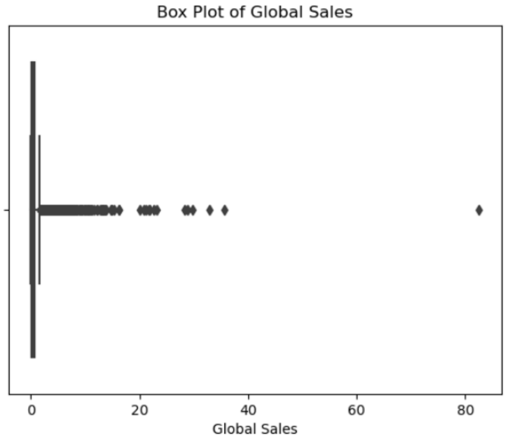


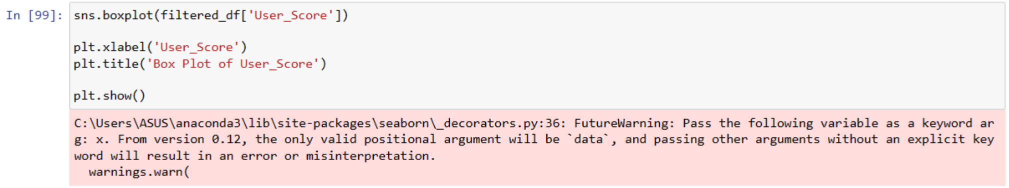
```
In [111]: #Boxplot of NA_Sales, EU_Sales, JP_Sales, Other_Sales

          filtered_df.boxplot(column=['NA_Sales','EU_Sales','JP_Sales','Other_Sales'])

Out[111]: <AxesSubplot:>
```

The code above is used to generate the boxplot for the specified columns NA_Sales, EU_Sales, JP_Sales and Other_Sales in filtered_df dataframe. Each column represents a variable for which the box plot will be created. The boxplots are used to define either data is outliers or not. The output above is shown that there are too many outliers exist and most of them are the higher number of sales.

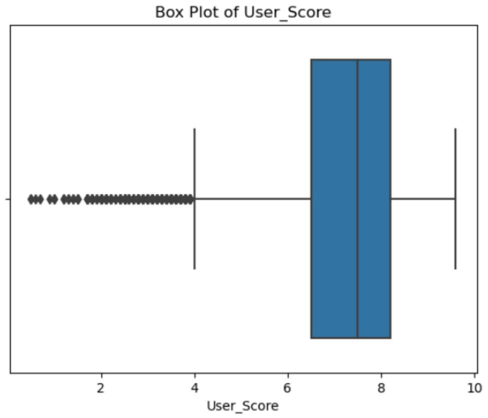


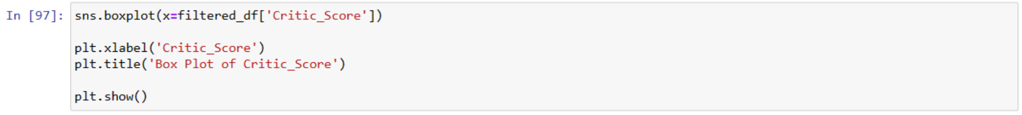
```
In [112]: #  Boxplot of Global Sales
          sns.boxplot(x=filtered_df['Global_Sales'])

          plt.xlabel('Global Sales')
          plt.title('Box Plot of Global Sales')

          plt.show()
```



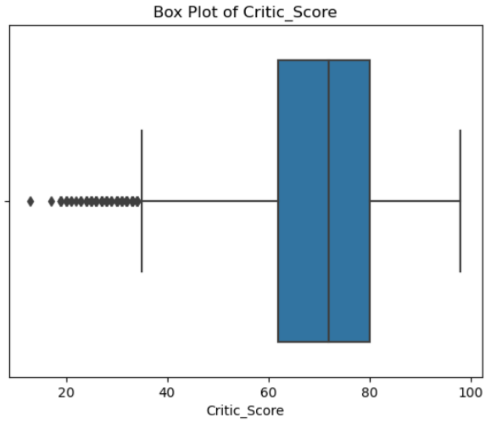


```
In [99]: sns.boxplot(filtered_df['User_Score'])

         plt.xlabel('User_Score')
         plt.title('Box Plot of User_Score')

         plt.show()

C:\Users\ASUS\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword ar
g: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit key
word will result in an error or misinterpretation.
  warnings.warn(
```





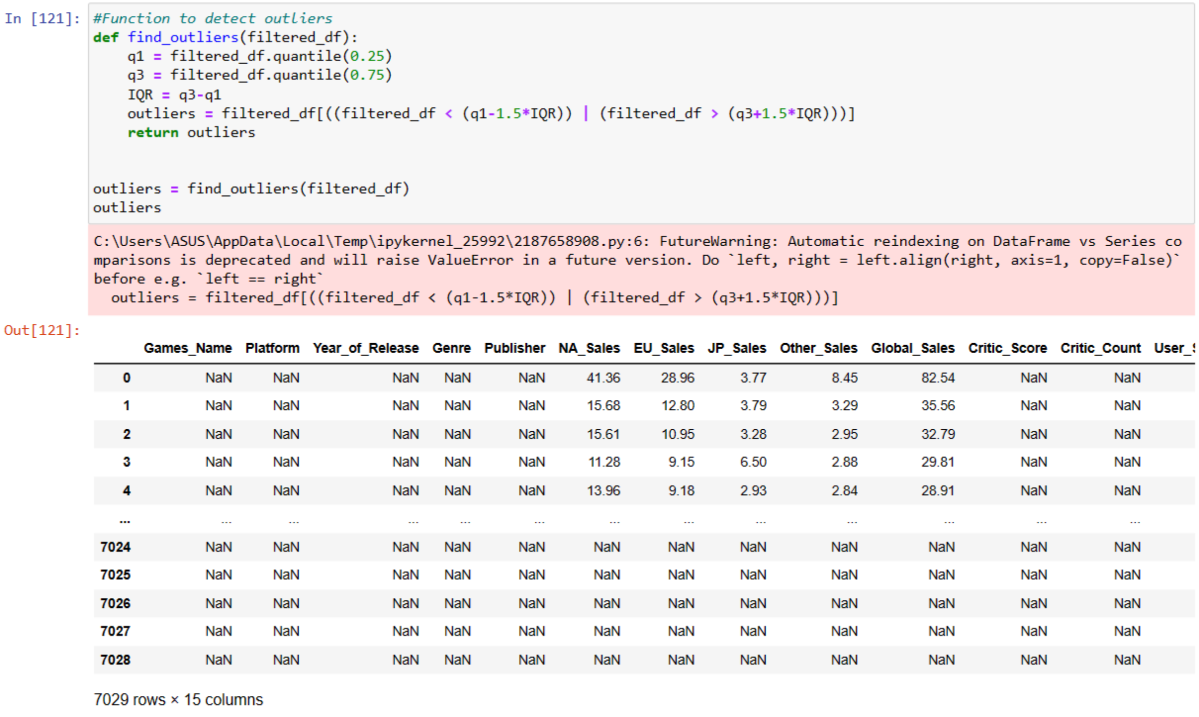
```
In [97]: sns.boxplot(x=filtered_df['Critic_Score'])

         plt.xlabel('Critic_Score')
         plt.title('Box Plot of Critic_Score')

         plt.show()
```



The codes above is used to create boxplots by using Seaborn's boxplot() function. The x parameter specifies the variable to be plotted on the x-axis, which in this case are the

Global_Sales, User Score and Critic_Score columns from the filtered_df dataframe. The boxplots will show the distribution and summary statistics of the columns' data that are mentioned in.

The boxplots are essential for analyzing and interpreting data distribution, identifying outliers and comparing different variables or categories. From the cases above, we generate that there are shown many outliers in the boxplots above, we need to find the outliers that exist in the dataframe.



```
In [121]: #Function to detect outliers
          def find_outliers(filtered_df):
              q1 = filtered_df.quantile(0.25)
              q3 = filtered_df.quantile(0.75)
              IQR = q3-q1
              outliers = filtered_df[((filtered_df < (q1-1.5*IQR)) | (filtered_df > (q3+1.5*IQR)))]
              return outliers


          outliers = find_outliers(filtered_df)
          outliers
```

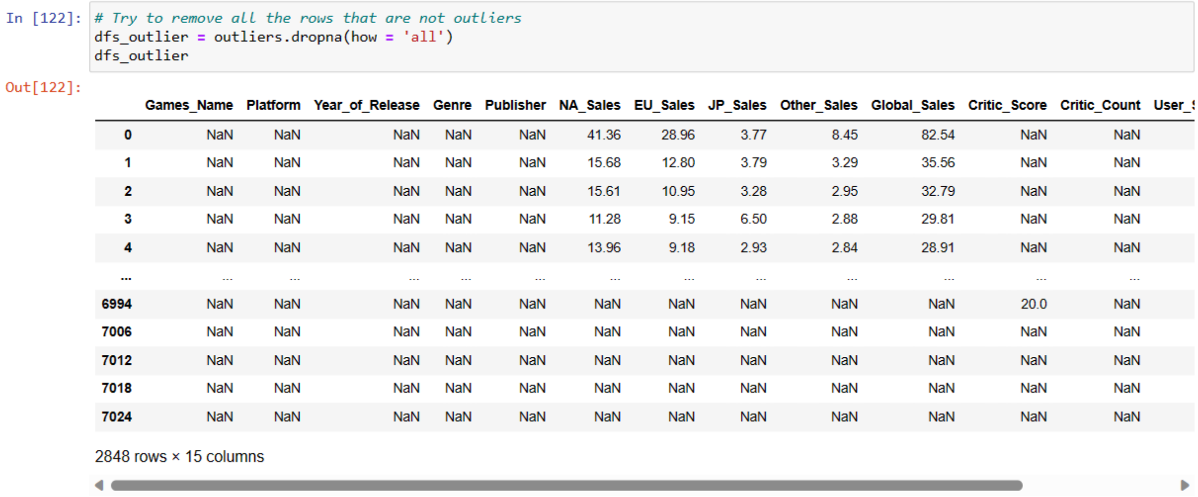
```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_25992\2187658908.py:6: FutureWarning: Automatic reindexing on DataFrame vs Series co
mparisons is deprecated and will raise ValueError in a future version. Do `left, right = left.align(right, axis=1, copy=False)`
before e.g. `left == right`
  outliers = filtered_df[((filtered_df < (q1-1.5*IQR)) | (filtered_df > (q3+1.5*IQR)))]
```

Out[121]:

	Games_Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_
0	NaN	NaN	NaN	NaN	NaN	41.36	28.96	3.77	8.45	82.54	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	15.68	12.80	3.79	3.29	35.56	NaN	NaN	
2	NaN	NaN	NaN	NaN	NaN	15.61	10.95	3.28	2.95	32.79	NaN	NaN	
3	NaN	NaN	NaN	NaN	NaN	11.28	9.15	6.50	2.88	29.81	NaN	NaN	
4	NaN	NaN	NaN	NaN	NaN	13.96	9.18	2.93	2.84	28.91	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	
7024	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7025	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7026	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7027	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7028	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

7029 rows × 15 columns

The code provided is to define a function called find_outliers that takes a dataframe (filtered_df) as input and identifies outliers within the dataset using the interquartile range (IQR) method. The function calculates the IQR by finding the first quartile (Q1) and third quartile (Q3) of the data. Then, it determines the outliers as values below (Q1-1.5*IQR) or above (Q3+1.5*IQR). The function returns a DataFrame containing the identified outliers. By calling this function with a dataframe, we can obtain and display the outliers. From the output above, the numerics are shown that there are outliers value which are needed to remove form the data frame.



In [122]:
```
# Try to remove all the rows that are not outliers
dfs_outlier = outliers.dropna(how = 'all')
dfs_outlier
```

Out[122]:

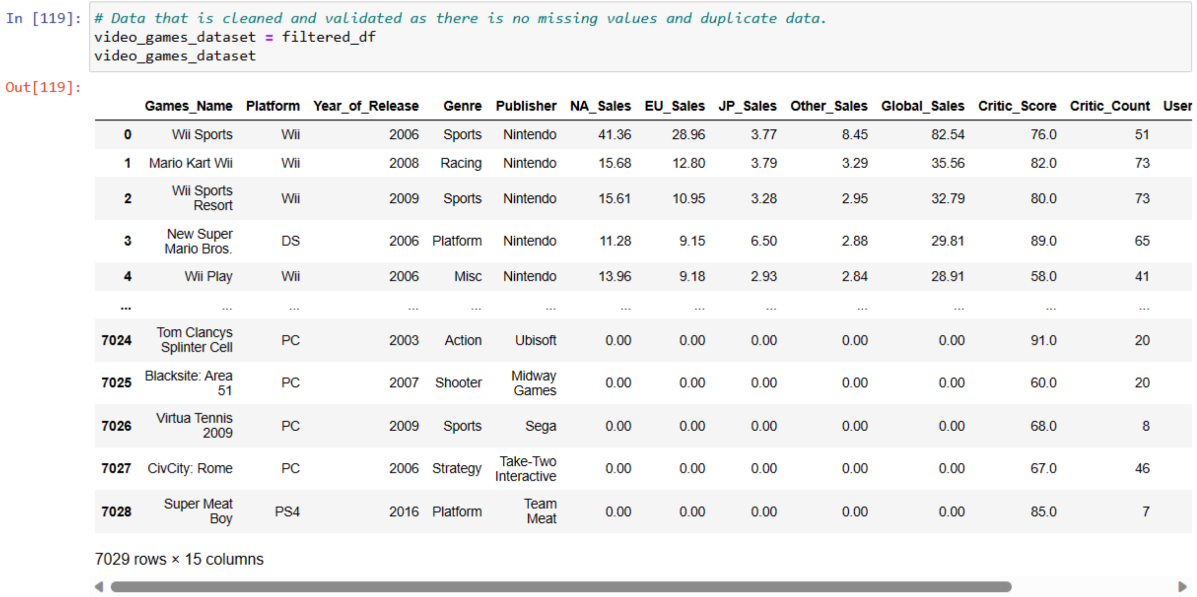
	Games_Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_
0	NaN	NaN	NaN	NaN	NaN	41.36	28.96	3.77	8.45	82.54	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	15.68	12.80	3.79	3.29	35.56	NaN	NaN	
2	NaN	NaN	NaN	NaN	NaN	15.61	10.95	3.28	2.95	32.79	NaN	NaN	
3	NaN	NaN	NaN	NaN	NaN	11.28	9.15	6.50	2.88	29.81	NaN	NaN	
4	NaN	NaN	NaN	NaN	NaN	13.96	9.18	2.93	2.84	28.91	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	
6994	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	20.0	NaN	
7006	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7012	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7018	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7024	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

2848 rows × 15 columns

The provided code aims to remove all the rows from the dataframe's outliers that contain only missing values (NaN) and assign the result to a new dataframe called dfs_outlier. the dropna() function to remove rows from the dataframe outliers that contain all missing values. The parameter how='all' specifies that only rows with all missing values should be dropped. The resulting dataframe with the removed rows is assigned to the variable dfs_outlier. dfs_outlier prints the dataframe dfs_outlier, displaying the remaining rows that contain outliers after removing any rows with all missing values.

Since there are too many outliers in this situation, we will not remove them. Outliers are exceptional. If, for instance, 40.52% of the data is an outlier, this really indicates that there is an interesting pattern in the data that deserves further investigation. The outliers contain all this information such as the mean of NA_Sales, JP_Sales, EU_Sales, Other_Sales, and Global Sales are too low to determine or provide an answer to our inquiries. Therefore, outliers that continue to dominate our data can achieve the goals and address the issues.

3.4 Data Preview



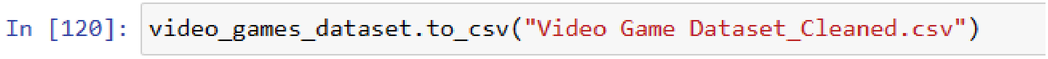
```
In [119]: # Data that is cleaned and validated as there is no missing values and duplicate data.
          video_games_dataset = filtered_df
          video_games_dataset
```

Out[119]:

	Games_Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User
0	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.54	76.0	51	
1	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.80	3.79	3.29	35.56	82.0	73	
2	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.95	3.28	2.95	32.79	80.0	73	
3	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.28	9.15	6.50	2.88	29.81	89.0	65	
4	Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.91	58.0	41	
...	...	...	...	...	...	...	...	...	...	...	...	...	
7024	Tom Clancys Splinter Cell	PC	2003	Action	Ubisoft	0.00	0.00	0.00	0.00	0.00	91.0	20	
7025	Blacksite: Area 51	PC	2007	Shooter	Midway Games	0.00	0.00	0.00	0.00	0.00	60.0	20	
7026	Virtua Tennis 2009	PC	2009	Sports	Sega	0.00	0.00	0.00	0.00	0.00	68.0	8	
7027	CivCity: Rome	PC	2006	Strategy	Take-Two Interactive	0.00	0.00	0.00	0.00	0.00	67.0	46	
7028	Super Meat Boy	PS4	2016	Platform	Team Meat	0.00	0.00	0.00	0.00	0.00	85.0	7	

7029 rows × 15 columns

The code snippet assigns the cleaned and validated data, which is stored in the filtered_df dataframe to a new dataframe called video_games_dataset. This indicates that the video_games_dataset dataframe now contains the cleaned and validated data without any missing values or duplicate entries. By assigning the cleaned data to a new dataframe, it allows for further analysis or processing on the cleaned dataset under the name video_games_dataset.



```
In [120]: video_games_dataset.to_csv("Video Game Dataset_Cleaned.csv")
```

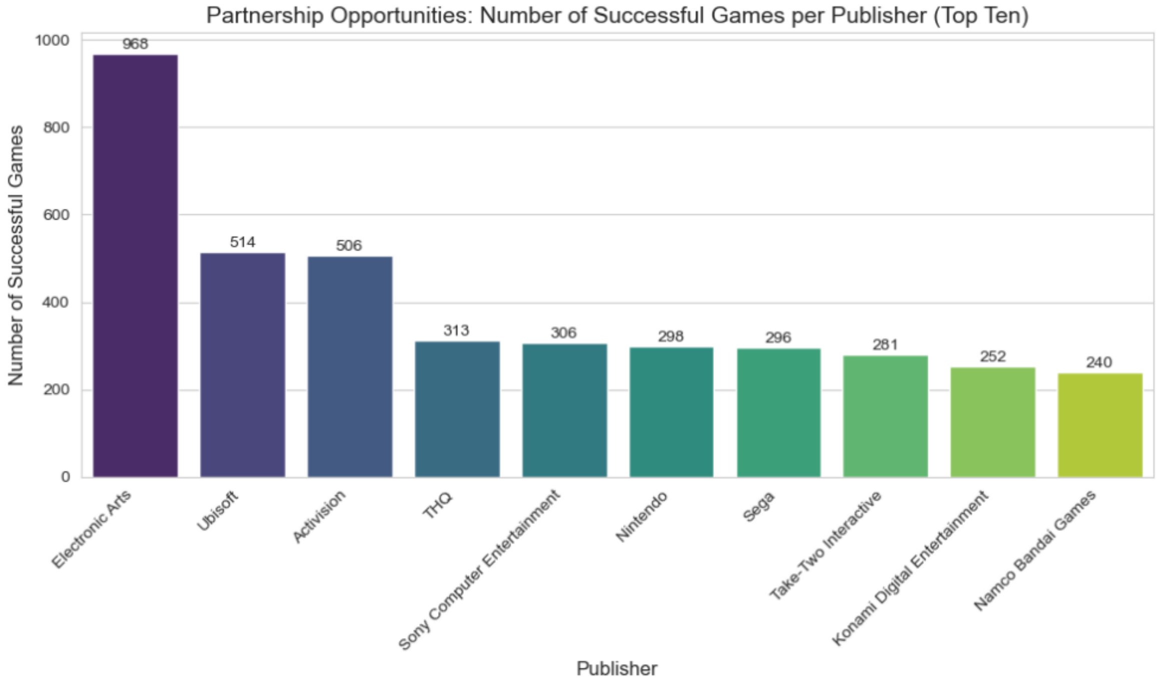
The code snippet saves the video_games_dataset dataframe, which contains the cleaned and validated data to a CSV file named "Video Game Dataset_Cleaned.csv". The to_csv() function is used to convert the dataframe into a CSV file format and save it to the specified file path. This allows for the preservation of the cleaned data in a separate file that can be easily accessed or shared for future use or analysis.

3.5 Data Description

Attribute Name	Data Type	Function on dataframe
Games_Name	object	Name of the game
Platform	object	Gaming platform
Year_of_Release	integer	Year of game release
Genre	object	Genre/category of the game
Publisher	object	Publisher of the game
NA_Sales	float	Sales in North America (in millions)
EU_Sales	float	Sales in Europe (in millions)
JP_Sales	float	Sales in Japan (in millions)
Other_Sales	float	Sales in other regions (in millions)
Global_Sales	float	Total global sales (in millions)
Critic_Score	float	Score given by critics
Critic_Count	integer	Count of critic reviews
User_Score	float	Score given by users
User_Count	integer	Count of user reviews
Rating	object	Content rating of the game

4.0 Data Exploratory Analysis

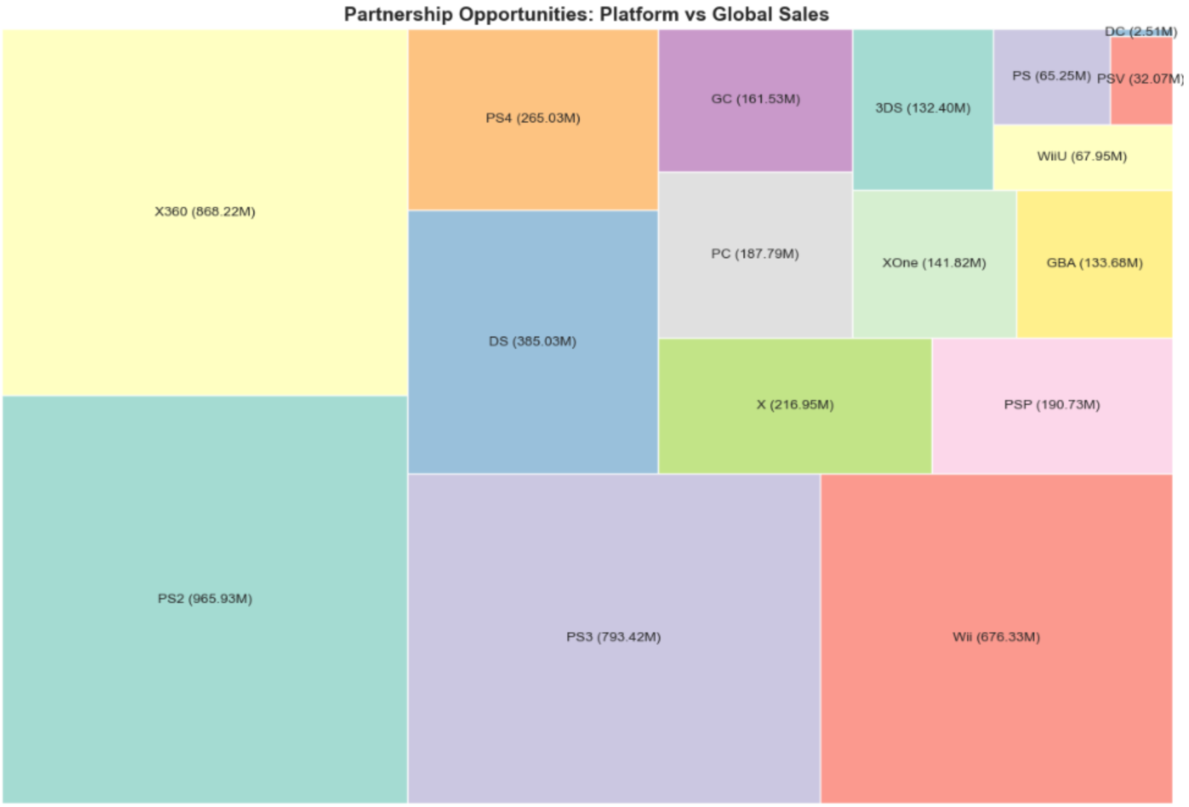
Objective 1: to analyse the publishers and platforms information for identifying the potential partnerships opportunities.



Visualisation of graph:

The provided graph is a bar chart that represents the number of successful games per publisher. Each bar in the chart represents a publisher, and the height of the bar corresponds to the number of successful games released by that publisher. The graph shows the partnership opportunities related to publishers in the gaming industry. It provides insights into the number of successful games released by each publisher, helping us identify potential partnership opportunities based on their track record of success. The top ten publishers with the highest number of successful games are shown in the graph. The top one publisher is Electronic Arts with a total of 968 successful games published, followed by Ubisoft, Activision, THQ, Sony Computer Entertainment ,Nintendo, Sega ,Take-Two Interactive ,Konami Digital Entertainment and Namco BAndai Games recorded the last in the top ten number of successful games published.

By analysing the graph, the bar chart provides a clear visual comparison of the publishers in the top ten, allowing us to easily identify the leading publishers in terms of the number of successful games we have. Through this, we can identify the publishers like Electronic Art that have a strong presence in the market and a proven track record of producing successful games. It may offer promising partnership opportunities as they have demonstrated their ability to develop and publish games that resonate with the audience and users. Publishers with a large number of successful games often possess strong development teams, marketing strategies, and distribution networks. This can be valuable for our company, who may benefit from the expertise and resources of these publishers for their own game development or distribution projects. This can also contribute to building trust and reputation between partner and publisher in the gaming industry at the same time.



Visualisation of graph:

In the provided treemap graph, the rectangles represent different platforms in the gaming industry. The size of each rectangle corresponds to the total global sales achieved by the games released on that platform. By analysing the treemap, we can observe the relative sizes of the platforms in terms of their global sales. The larger the rectangle, the higher the global sales for that particular platform. This provides valuable insights into the market dominance and popularity of each platform. In our graph, the biggest rectangle is PS2 with total global sales of $965.93 million. In contrast, the smallest rectangle is DC, with total global sales of $2.51 million.

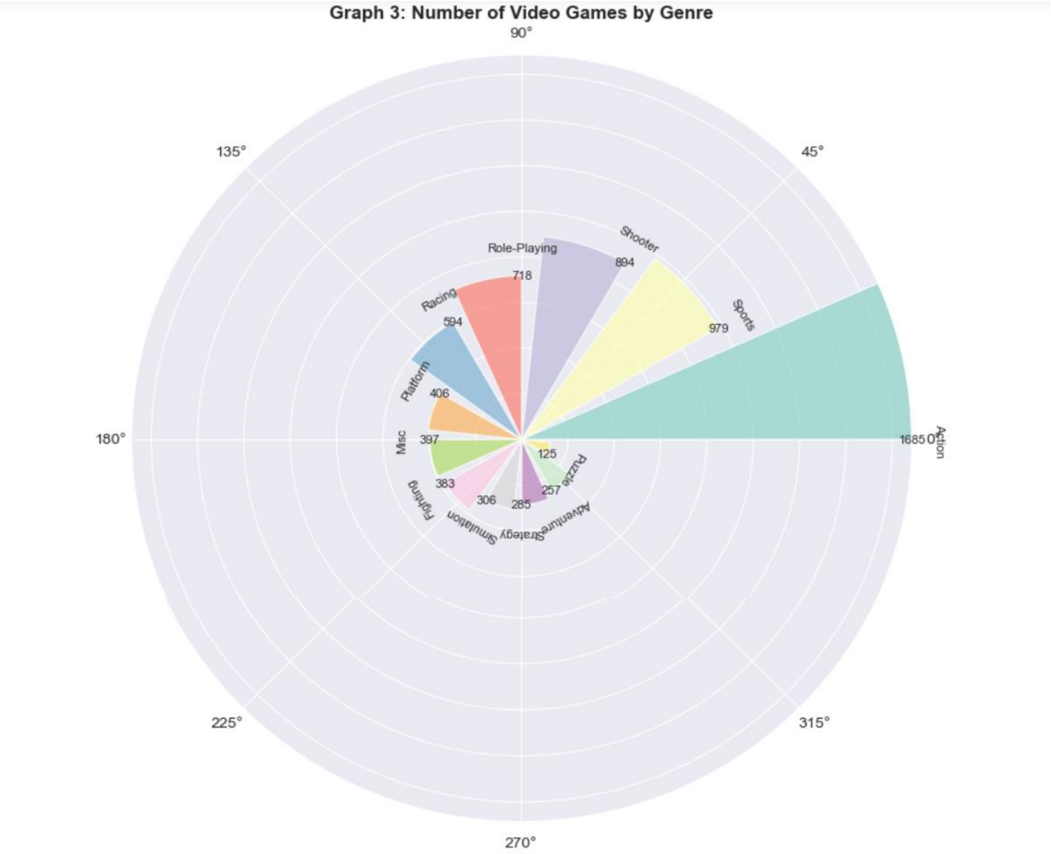
The treemap allows us to quickly identify the platforms with the highest global sales, which can indicate their success and market reach. These platforms, like the PS series, including PS2, PS3 and so on, can be potential partnership opportunities for the company. They have a larger share of the market, and higher global sales may offer a wider audience and greater potential for collaboration or distribution agreements. For further analysis based on the results of the graph, we can extract a few insights on Platform Dominance, Market Share analysis and Competitive Analysis.

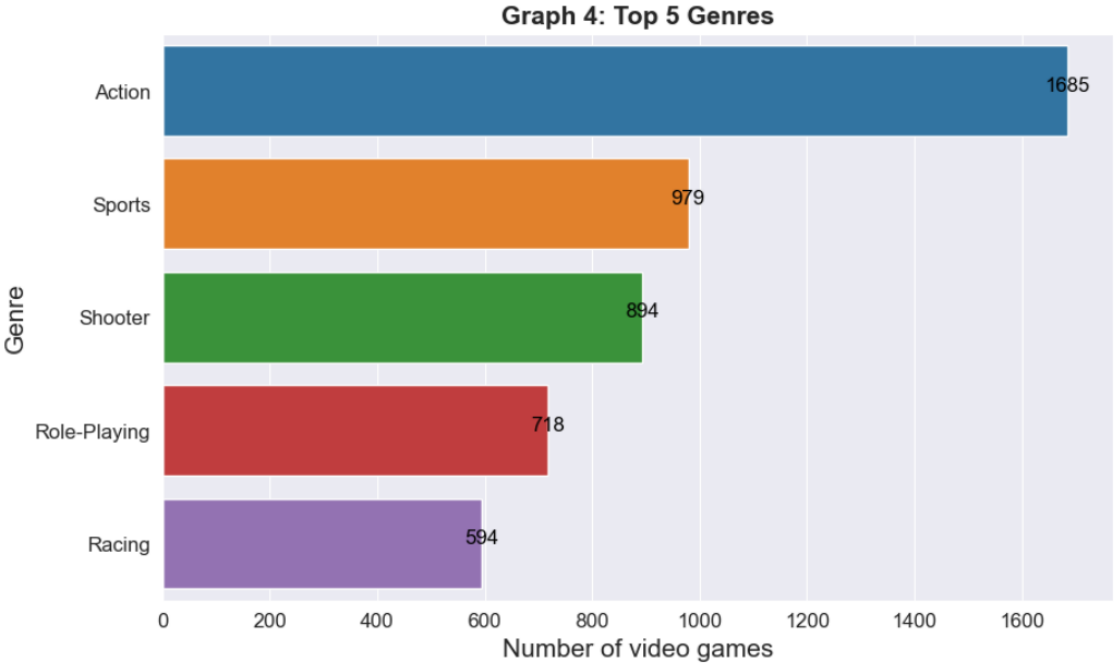
In the consideration of platform dominance, the treemap allows us to compare the sizes of the rectangles representing different platforms. Larger rectangles indicate platforms with higher global sales, suggesting their dominance in the market. These platforms have already established a strong presence and have a significant user base. Partnering with dominant platforms can provide access to a large audience, increasing the visibility and potential success of the company's games.

In the aspect of market share analysis,we can assess the market share of each platform. Platforms with larger rectangles have a larger share of global sales by comparing the sizes of the rectangles, indicating their popularity and user demand. These platforms can offer attractive partnership opportunities as they have a dedicated user base that is more likely to engage with new games and content.

Lastly,the graph enables a competitive analysis of platforms. By comparing the sizes of rectangles on different platforms, the company can identify direct competitors or platforms that target similar audiences. This analysis can help in understanding the competitive landscape and identifying potential partnership opportunities that can provide a competitive edge.

Objective 2: to study the performance evaluation of games by genres with utilizing sales data.





Visualisation of graph:

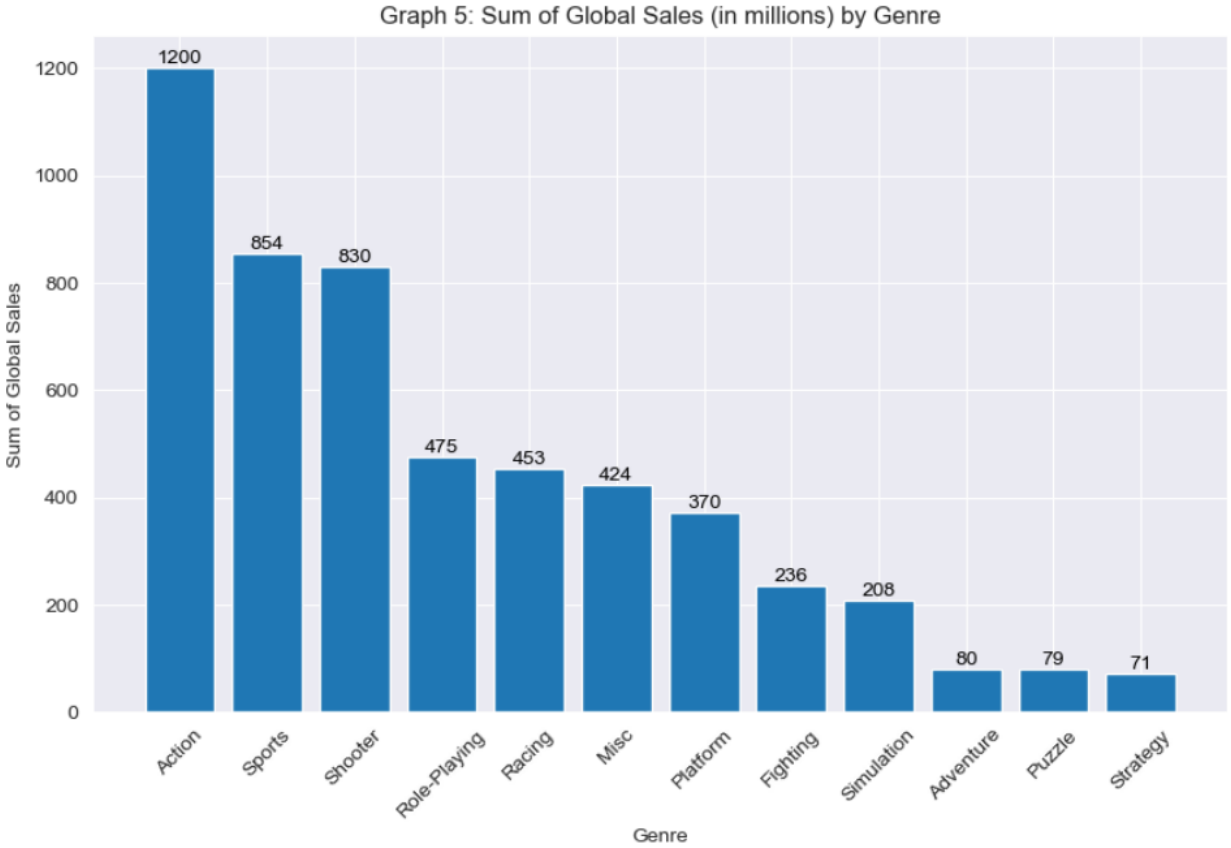
In the radial column chart interpretation, which is graph 3, we can focus on the length of each column radiating outward from the centre of the chart. The length of each column corresponds to the quantity of video games, similar to the height in the bar chart. The longer the column, the larger the number of video games in that genre.

In this graph, the top genre with the highest number of video games remains action games, as indicated by the longest column. (up to 1685 games) This demonstrates that action games have the most significant quantity among all the genres. Following action games, we can observe columns for sports games, shooters, role-playing games (RPG), and racing games, in descending order of the number of video games.

The popularity of action games can still be attributed to their exciting gameplay, as they provide an immersive and intense experience filled with adrenaline and engagement. The fast-paced nature of action games creates a thrilling and immersive environment for users. Additionally, the appeal of skill challenges, which require precise timing, coordination, and strategic thinking, contributes to the popularity of action games. The sense of accomplishment and motivation derived from mastering these skills make action games highly rewarding for users.

However, it's important to acknowledge that gaming preferences vary among individuals. Genres such as adventure, sports, RPGs, and puzzle games also have their own unique appeal. Compared to graph 3, graph 4 makes it straightforward to observe the top 5 game genres, which include action, sports, shooting, role-playing, and racing games. Each genre's popularity is influenced by factors like technological advancements, marketing efforts, and cultural trends.

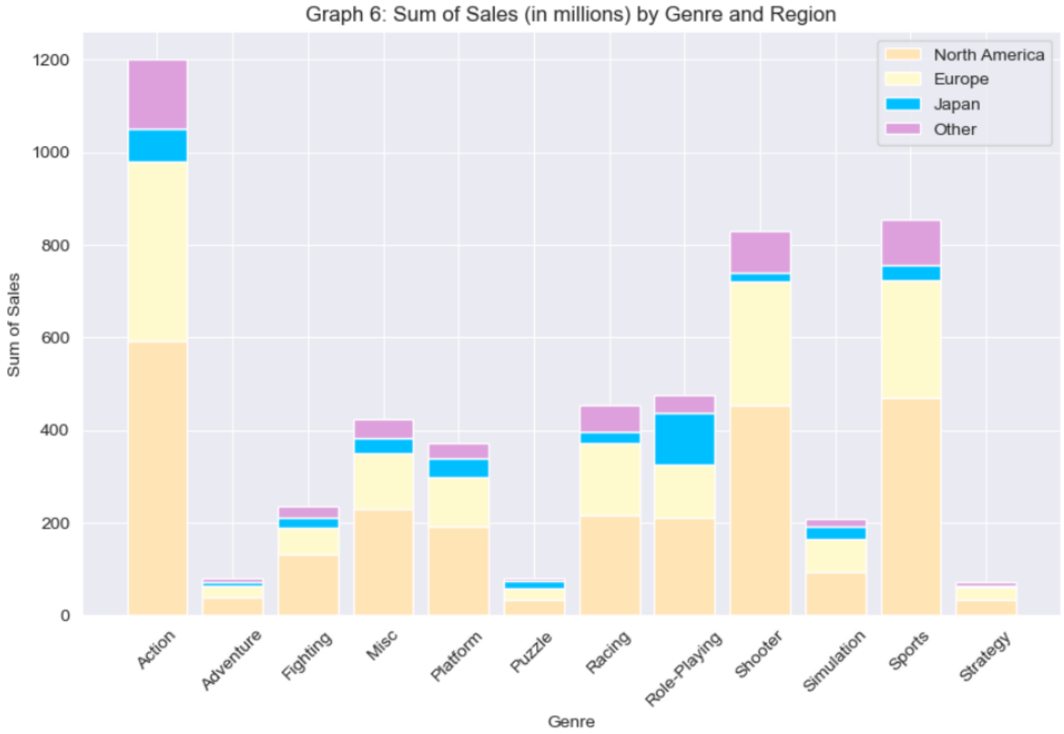
By analysing graphs 3 and 4, we can gain insights into the distribution and prominence of video games across different genres. The lengths of the columns allow for comparisons between genres, identification of the most and least prevalent genres, and assessment of the proportions and patterns within the dataset.



Visualisation of graph:

The total global sales (in millions) are shown in Graph 3 by genre. The y-axis should show the total amount of worldwide sales (in millions), with the x-axis representing various genres (such as action, adventure, and sports). The height of each bar in the graph, which represents a certain genre, corresponds to the total global sales for that genre. The graph displays worldwide video game sales according to gaming genres, which makes it easier for us to evaluate games by genre by using sales data. Through the graph, we can conclude that action games are the most well-liked by users and have the biggest total of worldwide sales when compared to other genres.

The genre with the greatest bar, representing the largest total of worldwide sales with 1200.63 million, is action. The genre with the lowest bar, representing the lowest total of global sales with 71.07 million, is strategy. Action games are quite popular because they frequently appeal to a wide range of age groups and demographics and can be found on a variety of gaming platforms, including consoles, PCs, and mobile devices. Therefore, this game genre may be the top option for game creators that are able to produce action-themed games with the highest game sales worldwide. We may examine the global sales of video games by region, such as Japan, Europe, North America, and other countries, and present the graph in the following utilising the total of video game sales by region and genres to decide which genre will make the most money from video game sales.



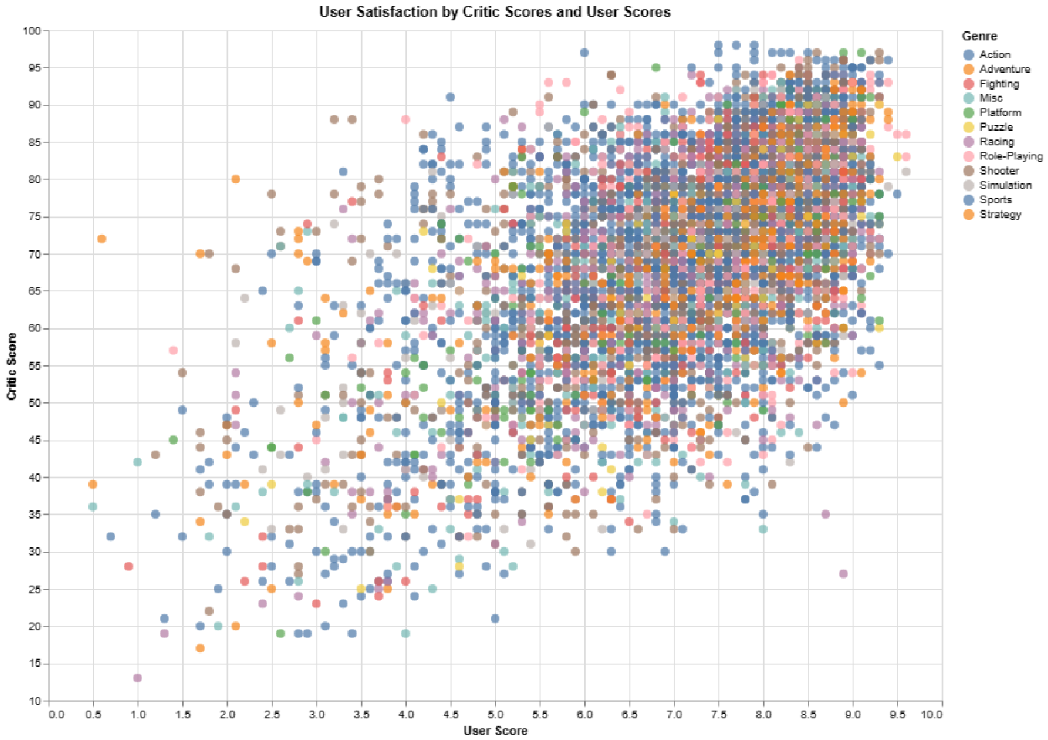
Visualisation of graph:

The total sales (in millions) by genre and location are shown in Graph 4. The various genres should be represented by the x-axis, and the total global sales should be shown by the y-axis. The individual segments within each stacked bar in the graph stand in for various geographical areas (such as North America, Europe, and Japan), while each bar represents a particular genre. The height of each segment implies the contribution of each area to the overall height of the stacked bar, which represents the total amount of worldwide sales for that genre. Given that the action genre games have the largest total worldwide sales in graph 3, we will examine which geographic location has the highest sales in graph 4.

The global sales comprise those from North America (NA_Sales), Europe (EU_Sales), Japan (JP_Sales), and other countries (Other_Sales). The action genres with the highest stacked bars have the largest worldwide sales totals, whereas the startegy genres with the lowest stacked bars have the lowest totals. We are able to observe that in North America, the majority of sales of video games connected to genres occupy the half of total sales in each genre, which suggests that the majority of the sales of video games in North America are related to genres. Graph 4 demonstrates that North American sales, which reached 591.66 million dollars, accounted for fifty percent of the total sales in the action genre. Sales in the other three areas, which included sales in Europe (387.17 million dollars), Japan (72.89 million dollars), and other nations (148.91 million dollars), accounted for the remaining third. In addition, sales of shooter and sports games frequently came from North America, accounting for more than half of the total sales.

North America has a well-established and thriving gaming industry. Major console and video game producers, such as Electronic Arts, Activision Blizzard, Ubisoft, and Microsoft, are based in the region, contributing to the development and release of popular games. This does contribute to an increase in video game sales in North America. Action, sports, and first-person shooters are just a few of the genres that North America has demonstrated a great interest in, contributing to increased sales in those categories as well as showing why these three genres have significantly higher sales than other genres. Therefore, we can conclude that the majority of sales are from North America, which also has the biggest sales and the ability to make more profits than any other area.

Objective 3: to understand the user satisfaction by evaluating the critic scores and user scores.



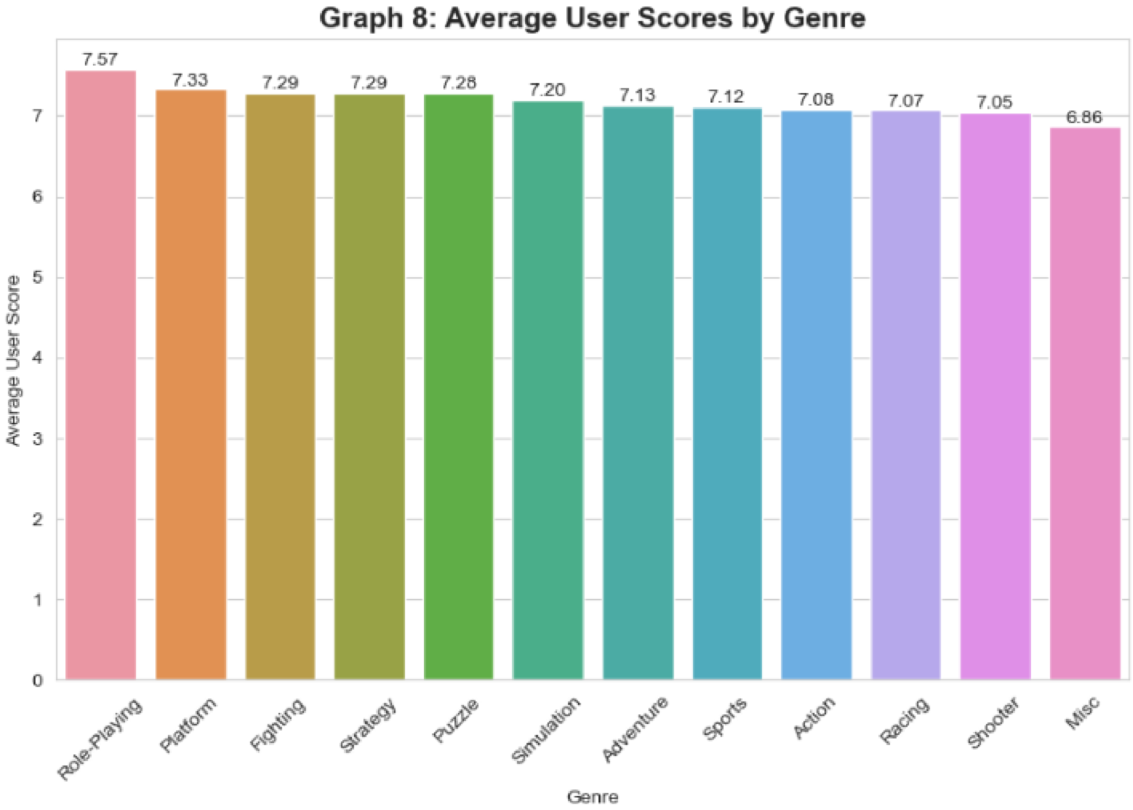
Visualization Interpretation:

The scatter plot reveals a clear trend in the upper-right quadrant, indicating a positive correlation between user satisfaction and critic scores. This means that games with higher critic scores tend to have higher user scores, indicating that users generally express satisfaction with games that receive positive reviews from critics.

For example, we can observe clusters of circles representing Action, Shooting, and Role-playing games, depicted by deep blue, brown, and pink colours, respectively. These clusters are positioned above 6.0 for user scores and 50 for critic scores, indicating that these genres receive high praise from critics and tend to have higher user satisfaction.

However, there are also scattered circles, particularly in genres like Racing and Adventure, which do not follow a clear pattern. These circles are placed below 4.5 for user scores and 45 for critic scores, suggesting a lack of correlation between user satisfaction and critic scores in these games. It implies that user satisfaction for these games can vary independently of critical reception, indicating that factors other than critic reviews, such as personal preferences or gameplay mechanics, may play a significant role in determining user satisfaction.

Overall, the scatter plot provides valuable insights into the relationship between user satisfaction and critic scores in the realm of video games. It demonstrates the overall alignment between critical reception and user sentiment, enabling game developers and enthusiasts to evaluate the influence of reviews on user satisfaction. This information can inform decisions related to game development, marketing strategies, and user engagement, ultimately aiming to enhance user satisfaction and overall gaming experiences.



Visualisation of graph:

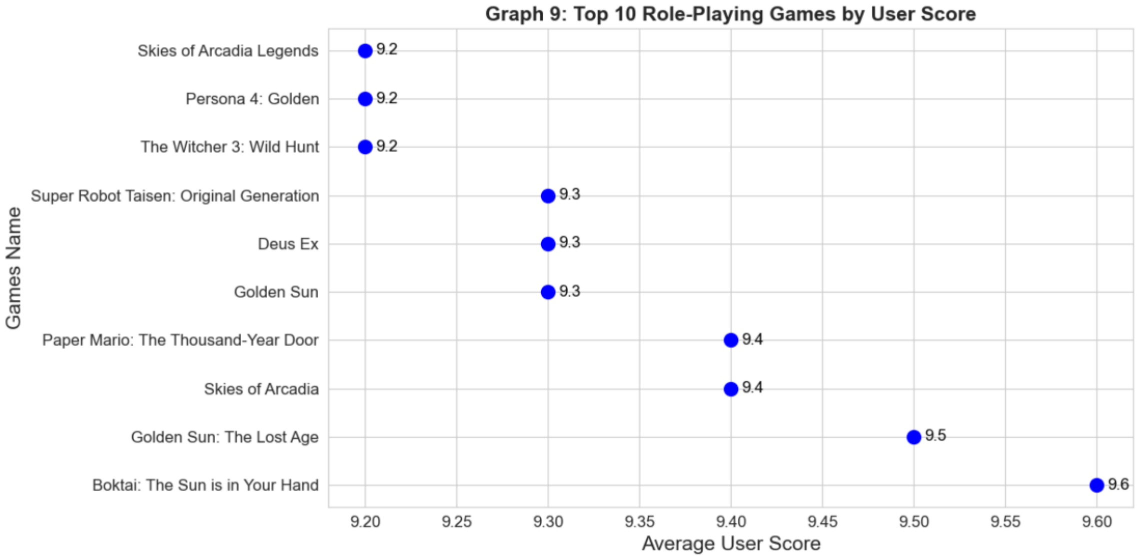
The bar plot displaying the average user scores by genre offers valuable insights into user satisfaction in the gaming industry. It provides a clear measure of user satisfaction for different genres and allows for comparisons between genres. The height of each bar represents the average user score for a specific genre. Taller bars indicate higher average user scores, indicating greater user satisfaction within those genres. Conversely, shorter bars represent lower average user scores, suggesting relatively lower user satisfaction.

In the graph above, we can observe that the overview average user score is between 6 and 8. The Role-Playing genre stands out with the highest average user score of 7.57, indicating a high level of user satisfaction among players of role-playing games. On the other hand, the Misc genre has the lowest average user score of 6.86, suggesting comparatively lower user satisfaction within that genre.

The bar plot allows us to make comparisons across different genres, enabling us to identify genres that consistently receive high or low user satisfaction. This information is valuable for game developers and industry professionals as it helps them understand the preferences and satisfaction levels of users within each genre, allowing for informed decision-making in game development strategies, marketing efforts, and investment priorities.

Additionally, the visual representation of the bar plot makes it easy for stakeholders to interpret and understand the data. The inclusion of value annotations on each bar further enhances clarity by providing precise average user scores for each genre.

In short, the bar plot of average user scores by genre offers a comprehensive view of user satisfaction within the gaming industry. It helps us understand the varying levels of satisfaction across different genres and assists in identifying areas for improvement and potential opportunities for game developers and industry professionals.

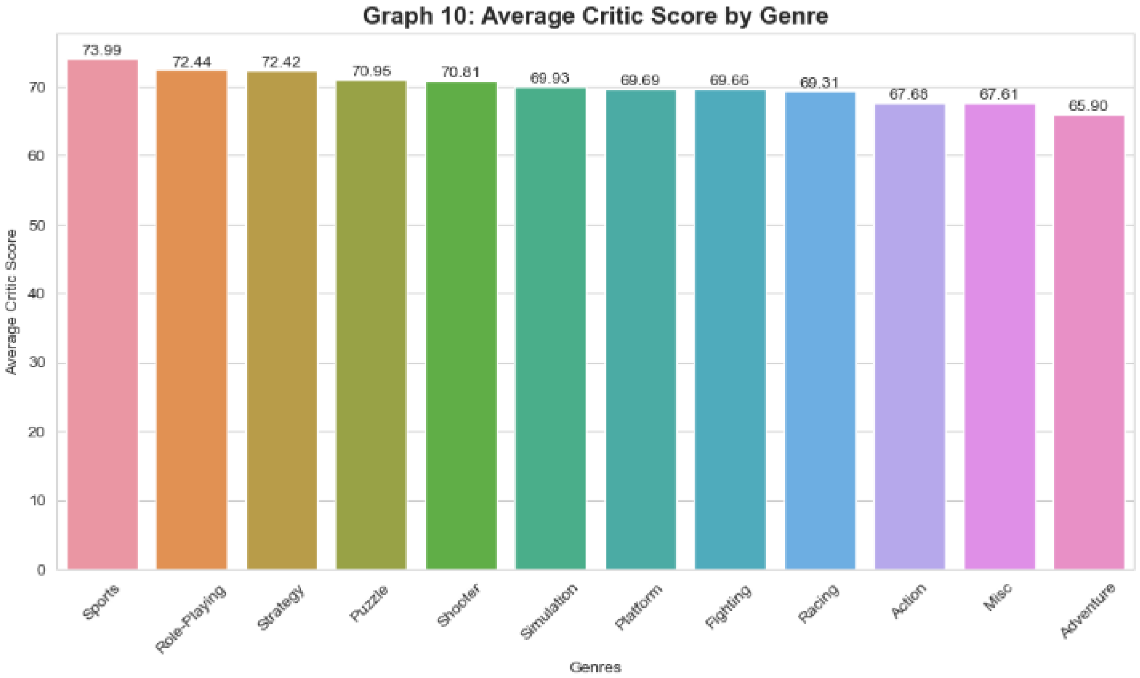


Visualisation of graph:

The visualisation above depicts the top 10 Role-Playing games by users scores provides insights into the impact of reviews on user satisfaction within the role-playing genre of video games. By examining the graph, we can identify the role-playing games that have garnered the highest user satisfaction. These games have achieved overally above 9 from 0 to 10 user scores, indicating that they have received consistently positive ratings from users. Users have expressed high levels of satisfaction with these games, which contributes to their overall positive reputation within the role-playing genre.

The highest user score that is achieved by Role-Playing Games is "Boktai: The Sun is in Your Hand", which achieved 9.6 user scores compared to another Role-Playing games, following the "Golden Sun: The Lost Age", "Skies of Arcadia" and "paper Mario: The Thousand-Year Door", which are accounted to 9.5, 9.4 and 9.4 respectively among the Role-Playing games. Meanwhile, the lowest user score that are achieved in the top 10 Role-Playing games are "Skies of Arcadia Legend", "Persona 4: Golden" and "The Witcher 3: Wild Hunt", which accounted to 9.2 respectively among the top 10 Role-Playing games.

In short, this information helps players and game developers make informed decisions regarding which role-playing games are highly regarded by users and may influence their gaming choices and development strategies. Hence, this visualisation of the top 10 role-playing games by user score highlights the games that have achieved notable user satisfaction within the genre. It provides valuable insights into the preferences and opinions of users, allowing players and industry professionals to identify highly regarded role-playing games and gauge their potential impact on user satisfaction.



Graph 10: Average Critic Score by Genre

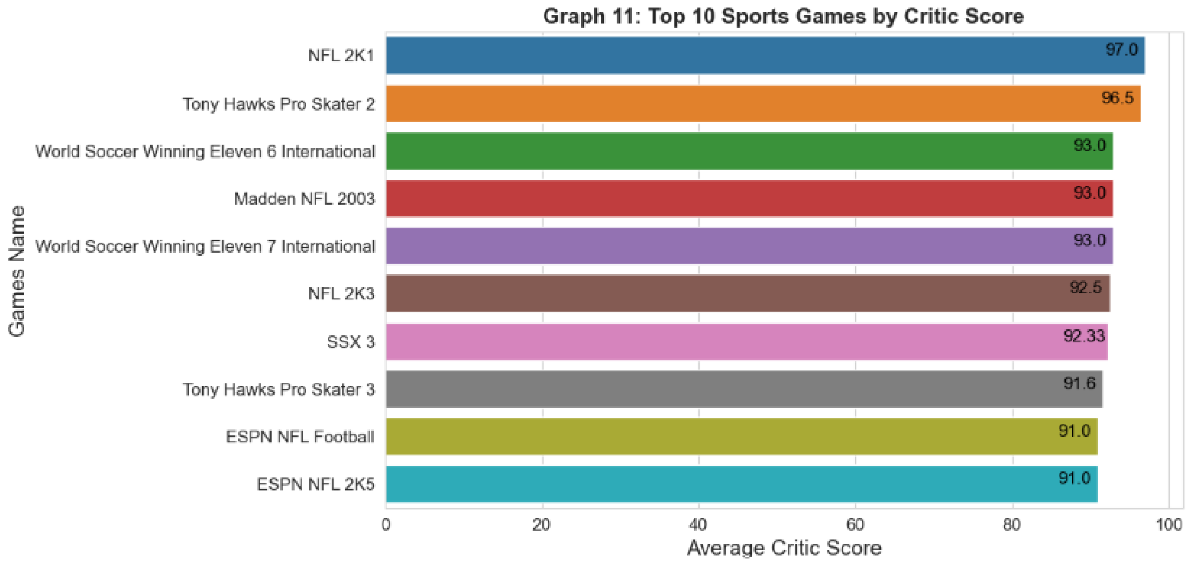
Visualization Interpretation:

The bar plot showcasing the average critic scores by genre provides insights into the impact of critic reviews on user satisfaction within the gaming industry. We can see from the graph that the various genres received overall critic scores ranging from 65 to 75.

For instance, we observe that the Sports genre has the highest average critic score, as indicated by the tallest bar with an average 73.99 critic score compared to other genres. This suggests that Sports games have received significant critical acclaim. This implies that users who enjoy role-playing games are more likely to find satisfaction in this genre, as it consistently receives positive reviews from critics. It highlights the correlation between positive critic reception and user satisfaction within the Sports genre games.

On the other hand, if we notice that the Adventure genre has a lower average critic score, represented by a shorter bar with an average 65.90 critic score among the games' genres, it indicates a relatively lower level of critical acclaim for Adventure games. This suggests that users who are interested in Adventure games might find less overall satisfaction within this genre, as it has received comparatively less positive recognition from critics.

By understanding the relationship between critic scores and user satisfaction within each genre, game developers and industry professionals can gain insights into user preferences and make informed decisions regarding game development, marketing strategies, and investment priorities. It serves as a valuable tool for assessing the impact of critic reviews on user satisfaction and shaping the future of the gaming industry.



Visualization Interpretation:

The bar plot depicting the top 10 sports games by critic score provides insights into the impact of critic reviews on user satisfaction within the sports genre of video games. By examining the lengths of the bars, we can assess the level of critical acclaim associated with each game that has achieved an overall score of 90 to 100. Games with longer bars indicate higher average critic scores, suggesting greater recognition and positive reception from critics. This implies that these sports games have been well-received and highly praised by critics in terms of their quality, gameplay, and overall experience.

In this case, we observe that the highest user score that is achieved by Sports Games is "NFL 2K1", which achieved 97 critic scores compared to another Sports Game. It suggests that this particular sports game has received a high average critic score, indicating a strong critical endorsement. This positive reception can potentially translate into higher user satisfaction, as users often consider critic reviews when making decisions about which games to play. On the other hand, the lowest user score that is achieved in the top 10 sports games are "ESPN NFL Football" and "ESPN NFL 2K5", which accounted for 91 critic scores among the top 10 sports games.

Moreover, we can also see that the NFL series and Soccer series games achieved an average high critic score, which provides valuable insights for both users and game developers. Users who enjoy sports games can use this information as a reference to discover highly regarded titles within the genre, which may increase their likelihood of finding games that align with their preferences and provide a satisfying gameplay experience. Similarly, game developers and industry professionals can analyse the top-performing sports games in terms of critic scores to understand the elements and features that contribute to higher user satisfaction. This information can guide their decision-making process in terms of game development, marketing strategies, and investments in the sports genre.

In short, the bar plot of the top 10 sports games by critic score offers a glimpse into the impact of critic reviews on user satisfaction within the sports genre. It highlights the games that have garnered critical acclaim, indicating their potential to provide enjoyable experiences and high satisfaction levels for users interested in sports games.

5.0 Summary

5.1 Summary of Integer Atttibute

For the "Critic_Score" attribute, it is revealed that the average score for games is 70.08, indicating a moderately positive reception from critics. The scores show a moderate level of peakedness, suggesting that most games received scores close to the mean. The distribution of scores is slightly left-skewed, indicating a larger number of games with lower scores compared to higher scores. The most frequently occurring score is 78, implying that this rating is commonly given by critics. The scores exhibit a wide variation, with a standard deviation of 13.86, indicating significant deviations from the average. The range of scores spans 85 points, from a minimum score of 13 to a maximum score of 98.

Moving on to the "Critic_Count" attribute, the analysis reveals that the average number of critics reviewing a game is approximately 28.85, indicating a moderate level of critical engagement. The median value of 24 suggests that half of the games have been reviewed by 24 or fewer critics, while the other half have been reviewed by more than 24. The mode of 17 indicates that a count of 17 critics is the most common among the games. The standard deviation of 19.25 signifies a considerable dispersion in critic counts, indicating that some games receive significantly more reviews than others. The range of critic counts spans from a minimum of 3 to a maximum of 113, highlighting the wide range of critic engagement for games.

For the "User_Count" attribute, it is found that the mean user count is approximately 172.97, indicating a moderate level of user engagement and feedback. The median value of 27 suggests that half of the games have been rated by 27 or fewer users, while the other half have received ratings from more than 27 users. The mode of 6 indicates that the most common user count among the games is 6. The standard deviation of 584.76 signifies a significant dispersion in user counts, indicating that some games have received a much higher number of ratings compared to others. The range of user counts spans from a minimum of 4 to a maximum of 10,766. The distribution of user counts is highly skewed, with a small number of games receiving an exceptionally large number of user ratings.

Next, the text provides information about sales figures in different regions. The "NA_Sales" attribute represents the sales figures of video games in North America, with an average of 0.38 million copies sold. The distribution is highly skewed, with a significant number of games having low or no sales. The range of sales varies from 0 to a maximum of 41.36 million copies. Similarly, the "EU_Sales" attribute represents sales figures in Europe, with an average of 0.23 million copies sold. The distribution is highly skewed, indicating a significant number of games with low or no sales. The range of sales in Europe varies from 0 to a maximum of 28.96 million copies. The "JP_Sales" attribute represents sales figures in Japan, with an average of 0.06 million copies sold. The majority of games have zero or very low sales, with the range of sales spanning from 0 to a maximum of 6.5 million copies. Lastly, the "Other_Sales" attribute represents sales figures in regions other than North America, Europe, and Japan, with an average of 0.081 million copies sold. The range of sales in other regions extends from 0 to a maximum of 10.57 million copies.

The "Global_Sales" attribute represents the total sales figures of video games worldwide, with an average of 0.752 million copies sold. The range of global sales extends from 0 to a maximum of 82.54 million copies, indicating a wide range of performance among games. The distribution of global sales is heavily skewed, with a concentration of sales towards the lower end.

Lastly, the "User_Score" attribute represents the user ratings or scores assigned to video games. The average user score is 7.162 out of 10, indicating a generally positive reception by users. The distribution of user scores shows some variability, with a mode value of 7.8 and a median value of 7.5. The range of user scores spans from 0.5 to 9.6, with a slightly left-skewed distribution.

Overall, these statistics provide insights into various aspects of video games, including critic and user ratings, critic and user counts, as well as sales figures in different regions. These insights can be valuable for evaluating game performance, understanding market trends, and informing strategic decision-making in the gaming industry.

5.2 Overall Summary

In conclusion, the analysis of the provided visualizations and data has provided valuable insights into the gaming industry. Based on the analysis:

1. Top 3 Platforms: The analysis of successful games per publisher revealed that Electronic Arts, Ubisoft, and Activision are the top three publishers with the highest number of successful games. These publishers have a strong track record of producing successful games and can be potential partnership opportunities.

2. Top 3 Genres: The bar graph showcasing the number of video games by genre identified action games, sports games, and shooters as the top three genres with the highest representation. Action games, in particular, stand out as the most popular genre, offering exciting gameplay and skill challenges that resonate with users.

3. Global Sales: The treemap graph representing global sales by platform indicated that the PS2 platform had the highest global sales, followed by other platforms such as PS3 and Xbox 360. These platforms with higher global sales can be attractive partnership opportunities, providing access to a larger user base and potential collaboration or distribution agreements.

4. Genre Sales: The graph displaying total global sales by genre highlighted that action games generated the highest total worldwide sales compared to other genres. This genre's popularity can be attributed to its exciting gameplay and wide appeal across different age groups and gaming platforms.

5. Geographic Sales: The stacked bar graph depicting total sales by genre and location revealed that North America accounted for the majority of video game sales in each genre, particularly in the action genre. North America has a thriving gaming industry and a significant presence of major publishers, leading to increased sales in the region.

Considering these findings, game developers and companies should take into account the popularity of action games, the potential partnerships with top publishers, and the market dominance of platforms such as PS2 and the North American market. These insights can inform decision-making processes related to game development, marketing strategies, and partnerships, ultimately improving the chances of success in the competitive gaming industry.

6.0 Reference

1. Madhugiri, D. (2021, October 26). How to Build a Treemap in 3 Ways Using Python. Analytics Vidhya. Retrieved September 10, 2022, from https://www.analyticsvidhya.com/blog/2021/10/how-to-build-a-treemap-in-3-ways-using-python/
2. Marchand, A., & Hennig-Thurau, T. (2013). Value creation in the video game industry: Industry economics, consumer benefits, and research opportunities. Direct Marketing Educational Foundation, Inc. Retrieved July 12,2013 from https://www.researchgate.net/publication/255995598_Value_Creation_in_the_Video_Game_Industry_Industry_Economics_Consumer_Benefits_and_Research_Opportunities
3. TM Geethanjali. (2020, May 5). Video Games Sales Analysis: A Data Science Approach, from https://ijcrt.org/papers/IJCRT2005182.pdf
4. Ignocio Chavarria. (2017,March 8). Predicting video game hits with machine learning: Analyze sales data from 8k video games, identify variables most correlated to hits, from https://towardsdatascience.com/predicting-hit-video-games-with-ml-1341bd9b86b0
5. Devin Norris. (2021, Feb 10). A data driven exploration of video games: Sales and scores, from https://medium.com/analytics-vidhya/a-data-driven-exploration-of-video-games-sales-and-scores-3c77f1c6573c
6. Casey Hoffman. (2021, Feb 7). Data trends in video game sales and ratings, form https://nycdatascience.com/blog/student-works/data-trends-in-video-game-sales-and-ratings/
7. Mubarak Ganiyu. (2019, Aug 24). What do critics and gamers say about video games compared to global sales. Do higher amount of sales of video games imply better ratings from users and critics, from https://towardsdatascience.com/what-do-critics-and-gamers-say-about-video-games-compared-to-global-sales-bdf7a395e064
8. Siddharth M, (2021, August 8, 2021). How to create a bar plot in python: A step-by-step guide (Updated 2023). Retrieved April 26, 2023 from https://www.analyticsvidhya.com/blog/2021/08/understanding-bar-plots-in-python-beginners-guide-to-data-visualization/
9. Ryan Browne, (2022, Jul 7). Video game sales set to fall for first time in years as industry braces for recession, from https://www.cnbc.com/2022/07/07/video-game-industry-not-recession-proof-sales-set-to-fall-in-2022.html
10. Andrew Beattie, (2021, Oct 31). How the video game industry is changing, from https://www.investopedia.com/articles/investing/053115/how-video-game-industry-changing.asp