Project 2

This report presents a data analysis of the "Video Game Sales" dataset from Kaggle. The goal is to explore trends in video game sales, identify popular genres, and analyze the performance of different publishers over time.

In [1]:
```
#Import libraries
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import os
```

Setup

We will setup our project using function from source.py script for better organization and readability.

In [2]:
```
from source import (
    loadFile, cleanFile, plotSalesByGenre,
    plotSalesOverTime, plotPublisherComparison,
    plotScatterSalesGlobal, plotScatterSalesPlatform
)

# Ensure matplotlib plots are displayed in the notebook
%matplotlib inline

file_path = "vgsales.csv" # Load the dataset
vgsales_df = loadFile(file_path)

# Display the first 5 rows of the raw data
if vgsales_df is not None:
    print("\nFirst 5 rows of the raw data:")
    display(vgsales_df.head())
    print("\nData information:")
    vgsales_df.info()
```

```
Load successful

First 5 rows of the raw data:
```

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37

```
Data information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16598 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16598 non-null  int64
 1   Name          16598 non-null  object
 2   Platform      16598 non-null  object
 3   Year          16598 non-null  float64
 4   Genre         16598 non-null  object
 5   Publisher     16540 non-null  object
 6   NA_Sales      16598 non-null  float64
 7   EU_Sales      16598 non-null  float64
 8   JP_Sales      16598 non-null  float64
 9   Other_Sales   16598 non-null  float64
 10  Global_Sales  16598 non-null  float64
dtypes: float64(6), int64(2), object(4)
memory usage: 1.4+ MB
```

Cleanup

This section is for cleaning up data - removing missing values or duplicates.

In [3]:
```
# Clean the data
cleaned_vgsales_df = cleanFile(vgsales_df)

# Display the information of the cleaned data
if cleaned_vgsales_df is not None:
    print("\nInformation of the cleaned data:")
    cleaned_vgsales_df.info()

    print("\nDescriptive statistics of numerical columns:")
    display(cleaned_vgsales_df.describe())
```

```
Information of the cleaned data:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16291 entries, 0 to 16290
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16291 non-null  int64
 1   Name          16291 non-null  object
 2   Platform      16291 non-null  object
 3   Year          16291 non-null  int64
 4   Genre         16291 non-null  object
 5   Publisher     16291 non-null  object
 6   NA_Sales      16291 non-null  float64
 7   EU_Sales      16291 non-null  float64
 8   JP_Sales      16291 non-null  float64
 9   Other_Sales   16291 non-null  float64
 10  Global_Sales  16291 non-null  float64
dtypes: float64(5), int64(2), object(4)
memory usage: 1.4+ MB

Descriptive statistics of numerical columns:
```

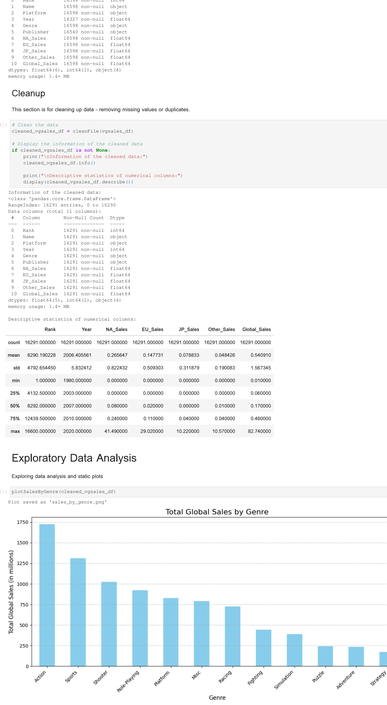
	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000
mean	8290.190228	2006.405561	0.265647	0.147731	0.078833	0.048426	0.540910
std	4792.654450	5.832412	0.822432	0.509303	0.311879	0.190083	1.567345
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4132.500000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8292.000000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12439.500000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.480000
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

Exploratory Data Analysis

Exploring data analysis and static plots

In [4]:
```
plotSalesByGenre(cleaned_vgsales_df)
```

```
Plot saved as 'sales_by_genre.png'
```

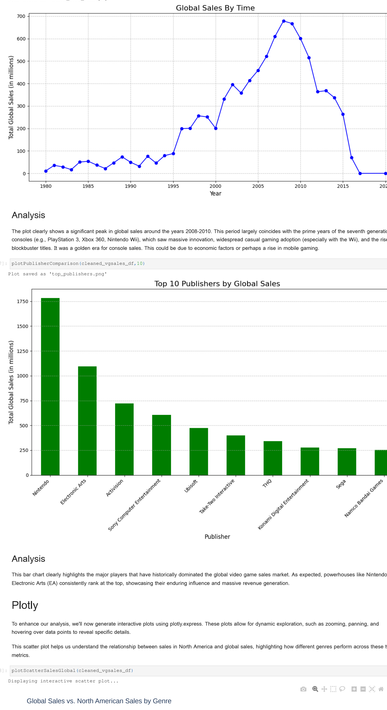


Analysis

We can observe a clear hierarchy, with a few dominant genres at the top (Action, Sports, Shooter, Role-Playing), followed by a moderate tier (Platform, Misc, Racing), and then several genres with considerably lower cumulative sales (Strategy, Puzzle, Adventure). This suggests that consumer preferences are heavily concentrated towards action-oriented and competitive gaming experiences.

In [5]:
```
plotSalesOverTime(cleaned_vgsales_df)
```

```
Plot saved as 'sales_over_time.png'
```

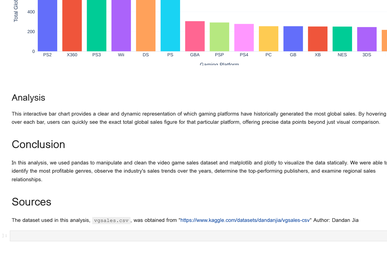


Analysis

The plot clearly shows a significant peak in global sales around the years 2008-2010. This period largely coincides with the prime years of the seventh generation of consoles (e.g., PlayStation 3, Xbox 360, Nintendo Wii), which saw massive innovation, widespread casual gaming adoption (especially with the Wii), and the rise of blockbuster titles. It was a golden era for console sales. This could be due to economic factors or perhaps a rise in mobile gaming.

In [6]:
```
plotPublisherComparison(cleaned_vgsales_df,10)
```

```
Plot saved as 'top_publishers.png'
```



Analysis

This bar chart clearly highlights the major players that have historically dominated the global video game sales market. As expected, powerhouses like Nintendo and Electronic Arts (EA) consistently rank at the top, showcasing their enduring influence and massive revenue generation.

Plotly

To enhance our analysis, we'll now generate interactive plots using plotly.express. These plots allow for dynamic exploration, such as zooming, panning, and hovering over data points to reveal specific details.

This scatter plot helps us understand the relationship between sales in North America and global sales, highlighting how different genres perform across these two metrics.

In [7]:
```
plotScatterSalesGlobal(cleaned_vgsales_df)
```

```
Displaying interactive scatter plot...
```


Analysis

Strong Positive Correlation: Visually, there is a clear strong positive correlation between North American sales (NA_Sales) and Global_Sales. This is expected, as North America is one of the largest video game markets, and games that perform well there typically contribute significantly to their global success. The points generally follow an upward trend from left to right.

The plot implicitly suggests the importance of the North American market. Games with low North American sales rarely achieve very high global sales, reinforcing NA's role as a primary driver of overall market success for many titles.

This interactive bar chart provides an easy way to visualize and explore the top gaming platforms by their total global sales

In [8]:
```
plotScatterSalesPlatform(cleaned_vgsales_df,top_n=15)
```

```
Displaying interactive bar chart for top 15 platforms...
```


Analysis

This interactive bar chart provides a clear and dynamic representation of which gaming platforms have historically generated the most global sales. By hovering over each bar, users can quickly see the exact total global sales figure for that particular platform, offering precise data points beyond just visual comparison.

Conclusion

In this analysis, we used pandas to manipulate and clean the video game sales dataset and matplotlib and plotly to visualize the data statically. We were able to identify the most profitable genres, observe the industry's sales trends over the years, determine the top-performing publishers, and examine regional sales relationships.

Sources

The dataset used in this analysis, `vgsales.csv`, was obtained from "https://www.kaggle.com/datasets/dandanjia/vgsales-csv" Author: Dandan Jia

In [ ]:
```
```