Data Analysis Project Report

VIDEO GAME SALES ANALYSIS

Renata Golemovic

October 2024

1. Use Case

This project is based on the Video Game Sales dataset from Kaggle, which contains detailed information on sales performance across various platforms, genres, publishers, and regions.

The objective is to analyze and model global video game sales using Python in Google Colab, supported by Power BI for advanced data visualization.

Throughout this chapter, we focus on:

- Loading and exploring the dataset to understand its structure and data distribution.
- Cleaning and preprocessing data by handling missing values, zero entries, and irrelevant columns.
- Analyzing sales trends by platform, genre, and release year to identify key market patterns.
- Visualizing insights with interactive dashboards in Power BI for better interpretation of results.
- Building a predictive Machine Learning model to estimate potential global sales based on historical data and selected features.

By combining analytical tools and machine learning, this project aims to uncover patterns that define success in the world of video games.

2. Data Loading and Preprocessing

In this part of the project, the dataset was imported into Google Colab using the Pandas library.

First, we displayed the shape of the dataset to see how many rows and columns it contains, as well as the names of all available features. This allowed us to get a first impression of the data volume and structure.

Next, we examined the data types of each column to ensure they were correctly recognized (e.g., numerical values as float64 or int64, and categorical values as object). All data types were correctly assigned, so no additional conversion was needed.

We then analyzed the dataset for missing or zero values to identify potential data quality issues. The columns NA_Sales, EU_Sales, JP_Sales, and Other_Sales contained a significant number of zeros, indicating missing or irrelevant sales data for certain regions.

To ensure a cleaner dataset and more accurate results, we decided to focus only on the Global_Sales column, which aggregates total sales across all regions.This approach removed unnecessary noise from the dataset and prepared it for the next analytical stage.

```
Dataset shape (rows, columns):
(16598, 11)

Information about the dataset:
<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          16327 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(1), object(4)
memory usage: 1.4+ MB
None
```

Figure 2.1 Dataset structure and column overview

```
Number of zero values per column:
Rank                0
Name                0
Platform            0
Year                0
Genre               0
Publisher           0
NA_Sales         4499
EU_Sales         5730
JP_Sales        10455
Other_Sales      6477
Global_Sales        0
dtype: int64

Columns containing zero values:
NA_Sales        4499
EU_Sales        5730
JP_Sales       10455
Other_Sales     6477
dtype: int64
```

```
# Keep only relevant columns (for example, Global_Sales)
data = data[['Name', 'Platform', 'Year', 'Genre', 'Publisher', 'Global_Sales']]

# Confirm remaining columns
print("Remaining columns:")
print(data.columns)
```

```
Remaining columns:
Index(['Name', 'Platform', 'Year', 'Genre', 'Publisher', 'Global_Sales'], dtype='object')
```

Figure 2.2 Zero value analysis per column

After cleaning the dataset from missing and zero values, several additional quality checks were performed to ensure the reliability of the data.

First, we examined duplicate entries and found one duplicate row, which was subsequently removed to prevent data redundancy.

Next, we conducted an outlier analysis on the Global_Sales column using a boxplot and the interquartile range (IQR) method. This analysis identified a few extreme values corresponding to highly successful titles such as Wii Sports and Super Mario Bros. Since these represent legitimate market performance rather than data errors, they were considered valid but excluded from further analysis. All subsequent analyses were therefore performed only on games with sales values below the calculated IQR upper limit, ensuring a more balanced and representative dataset.

Finally, we conducted a data consistency validation to confirm the logical integrity of the dataset. No negative or unrealistic values were found in the Global_Sales column, and all Year entries fell within a reasonable range (1980–2025).

Additionally, categorical attributes were verified, showing 12 unique genres, 31 platforms, and 578 publishers, confirming a diverse and well-structured dataset.

With these validation steps completed, the dataset was confirmed to be accurate, consistent, and ready for the upcoming Exploratory Data Analysis (EDA) phase.

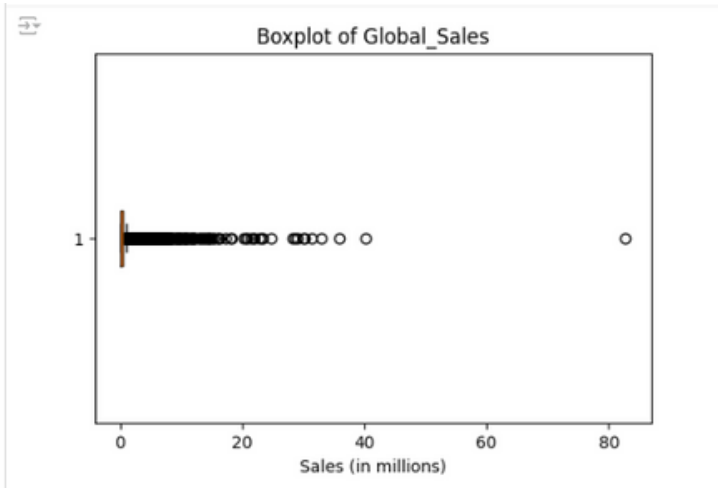
```
# 🔍 1. Check for duplicate records

duplicates = data.duplicated().sum()
print(f"Number of duplicate rows: {duplicates}")

# If any duplicates exist, remove them
if duplicates > 0:
    data.drop_duplicates(inplace=True)
    print("Duplicates removed.")
else:
    print("No duplicate records found.")
```

```
Number of duplicate rows: 1
```

Figure 2.3 Removing Dupliicates




```
[21]
✓ 0s
Q1 = data['Global_Sales'].quantile(0.25)
Q3 = data['Global_Sales'].quantile(0.75)
IQR = Q3 - Q1
upper_limit = Q3 + 1.5 * IQR

print(f"Upper limit for normal values: {upper_limit:.2f}M")

Upper limit for normal values: 1.08M
```

Figure 2.4 Boxplot and IQR-Based Outlier Detection for Global_Sales

```
[14]
✓ 0s
# ✅ 3. Validate data consistency
# Check for negative or unrealistic values
invalid_sales = data[data['Global_Sales'] < 0]
print(f"\nRows with negative sales: {len(invalid_sales)}")

Rows with negative sales: 0
```

```
[15]
✓ 0s
# Check for unrealistic years (before 1980 or after 2025)
invalid_years = data[(data['Year'] < 1980) | (data['Year'] > 2025)]
print(f"Rows with unrealistic Year values: {len(invalid_years)}")

# Check unique values in categorical columns
print("\nUnique Genres:", data['Genre'].nunique())
print("Unique Platforms:", data['Platform'].nunique())
print("Unique Publishers:", data['Publisher'].nunique())

Rows with unrealistic Year values: 0

Unique Genres: 12
Unique Platforms: 31
Unique Publishers: 578
```

Figure 2.5 Consistency Check Results

3. Exploratory Data Analysis (EDA)

In this chapter, we perform an exploratory analysis of the cleaned dataset to identify patterns, trends, and relationships within the video game market.

The goal of this phase is to better understand how different factors, such as platform, genre, publisher, and release year, influence global sales performance.

The analysis includes both descriptive statistics and visual representations to uncover meaningful insights from the data.

We examine the distribution of sales across categories, explore sales evolution over time, and analyze the correlation between numerical variables.

Additionally, we identify the top-performing games, publishers, and platforms, providing a deeper understanding of the market dynamics.

Through this exploration, we aim to reveal the key elements that have shaped the success of video games worldwide and prepare the foundation for predictive modeling in the next stage.

The following analyses are performed within this chapter:

3.1 Descriptive statistics of global sales values

3.2 Distribution of sales across different genres

3.3 Comparison of total sales between gaming platforms

3.4 Analysis of sales performance by publisher

3.5 Examination of global sales trends over time

3.6 Identification of the top 10 best-selling games

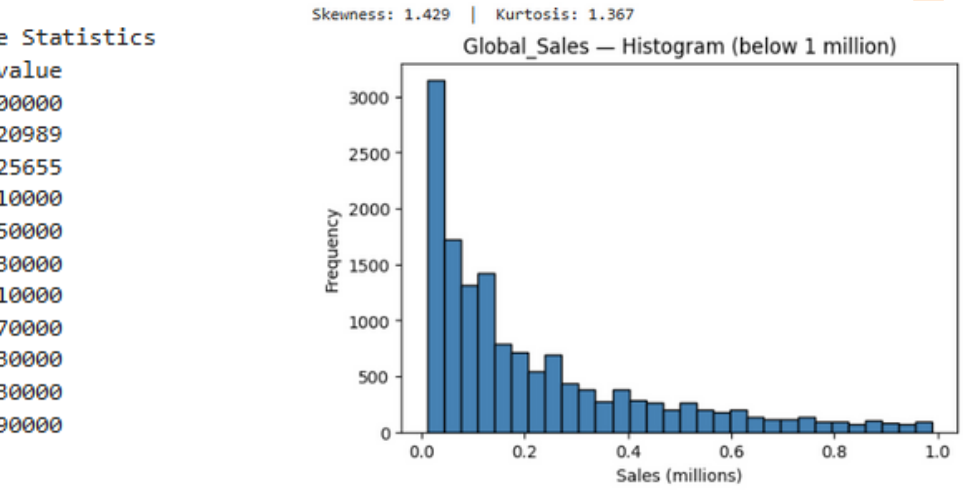
3.7 Correlation analysis between numerical variables

3.8 Summary of key insights derived from the dataset

To avoid distortion from extreme values, the dataset was restricted to games with global sales below 1 million units, as defined by the IQR analysis.

3.1 Descriptive Statistics

```
=== Descriptive Statistics
              value
count  14516.000000
mean       0.220989
std        0.225655
min        0.010000
25%        0.050000
50%        0.130000
75%        0.310000
90%        0.570000
95%        0.730000
99%        0.930000
max        0.990000
```



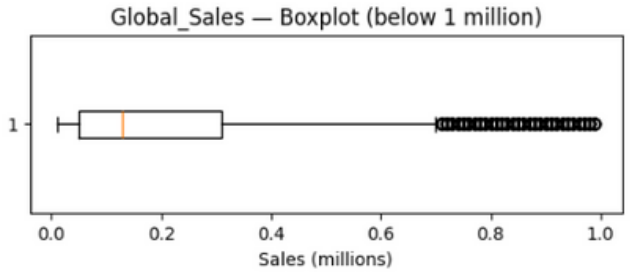


Figure 3.1 Descriptive Statistics

To obtain a more representative view of typical game sales, the dataset was limited to titles with global sales below one million units.The resulting subset includes 14,516 games, with an average global sale of 0.22 million units (≈220,000 copies) and a median of 0.13 million units.

The histogram reveals a highly right-skewed distribution, indicating that the vast majority of games achieve relatively low sales, while only a few approach the upper limit of one million units.

This is further supported by the skewness value of 1.43, confirming a strong positive asymmetry, and a kurtosis of 1.37, suggesting a moderately peaked distribution with several high-performing outliers.

The boxplot also highlights this pattern, showing a dense concentration of values below 0.3 million and a long tail extending toward the upper boundary.

In summary, global game sales display a long-tail distribution, where most games sell modestly, and only a small fraction reach significant commercial success.

3.2 Distribution of Sales Across Different Genres

	Genre	titles	total_sales	mean_sales	median_sales	p90_sales	share_%
1	Action	2890	679.93	0.235270	0.15	0.580	21.20
2	Sports	2043	539.13	0.263891	0.18	0.620	16.81
3	Misc	1566	330.06	0.210766	0.13	0.530	10.29
4	Role-Playing	1285	290.46	0.226039	0.14	0.580	9.05
5	Shooter	1057	266.05	0.251703	0.15	0.640	8.29
6	Racing	1067	245.45	0.230037	0.14	0.590	7.65
7	Fighting	724	175.65	0.242610	0.15	0.607	5.48
8	Platform	691	175.60	0.254124	0.16	0.630	5.47
9	Simulation	774	164.49	0.212519	0.12	0.537	5.13
10	Adventure	1244	146.06	0.117412	0.05	0.300	4.55
11	Strategy	649	108.21	0.166733	0.09	0.432	3.37
12	Puzzle	526	86.78	0.164981	0.09	0.450	2.71

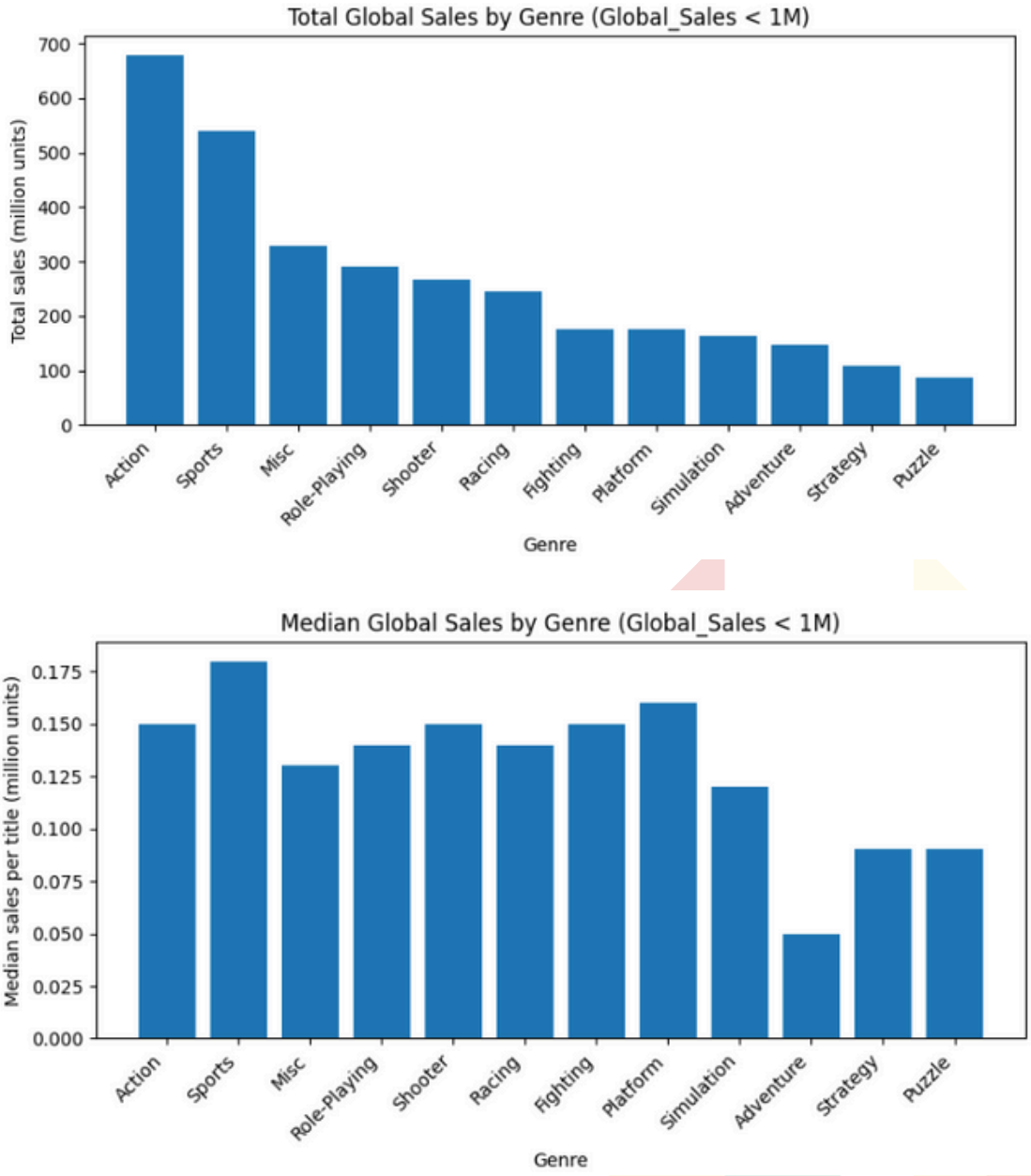


Figure 3.2 Distribution of Sales Across Different Genresistics

To examine market performance by category, global sales below one million units were grouped by genre. The results show that Action and Sports games dominate the market, accounting for roughly 21% and 17% of total sales respectively. These genres also have the highest number of titles, reflecting their broad popularity and frequent releases.

Genres such as Misc, Role-Playing, and Shooter follow, each contributing between 8 – 10% of total sales. While RPG titles are fewer, their average and median sales per game remain relatively strong, suggesting dedicated player communities.

On the other hand, Adventure, Strategy, and Puzzle genres record the lowest sales shares (below 5%), indicating smaller, more niche markets.

The overall distribution highlights a genre-driven concentration of commercial success: mainstream genres like Action and Sports attract the majority of consumer demand, whereas others cater to more specialized audiences.

3.3 Comparision of Total Sales Between Gaming Platforms

	Platform	titles	total_sales	mean_sales	median_sales	share_%
1	PS2	1836	466.23	0.253938	0.180	14.53
2	DS	2019	343.96	0.170362	0.100	10.72
3	PS3	1082	318.58	0.294436	0.200	9.93
4	X360	1029	297.57	0.289184	0.200	9.28
5	Wii	1162	285.91	0.246050	0.160	8.91
6	PS	990	267.00	0.269697	0.180	8.32
7	PSP	1156	182.09	0.157517	0.080	5.68
8	XB	770	164.66	0.213844	0.130	5.13
9	GBA	752	158.56	0.210851	0.140	4.94
10	GC	510	104.56	0.205020	0.130	3.26
11	PC	891	103.95	0.116667	0.040	3.24
12	3DS	462	86.02	0.186190	0.110	2.68
13	N64	268	76.13	0.284067	0.220	2.37
14	PS4	263	58.74	0.223346	0.120	1.83
15	SNES	191	54.66	0.286178	0.260	1.70

	Platform	titles	total_sales	mean_sales	median_sales	share_%
15	SNES	191	54.66	0.286178	0.260	1.70
16	PSV	405	49.77	0.122889	0.060	1.55
17	2600	107	45.61	0.426262	0.390	1.42
18	XOne	170	42.06	0.247412	0.155	1.31
19	WiiU	124	31.46	0.253710	0.150	0.98
20	SAT	170	29.43	0.173118	0.120	0.92
21	GB	43	14.43	0.335581	0.280	0.45
22	NES	24	14.23	0.592917	0.630	0.44
23	DC	46	7.13	0.155000	0.115	0.22
24	GEN	16	1.57	0.098125	0.075	0.05
25	NG	12	1.44	0.120000	0.100	0.04
26	WS	6	1.42	0.236667	0.215	0.04
27	SCD	5	0.37	0.074000	0.060	0.01
28	TG16	2	0.16	0.080000	0.080	0.00
29	3DO	3	0.10	0.033333	0.020	0.00
30	GG	1	0.04	0.040000	0.040	0.00
31	PCFX	1	0.03	0.030000	0.030	0.00

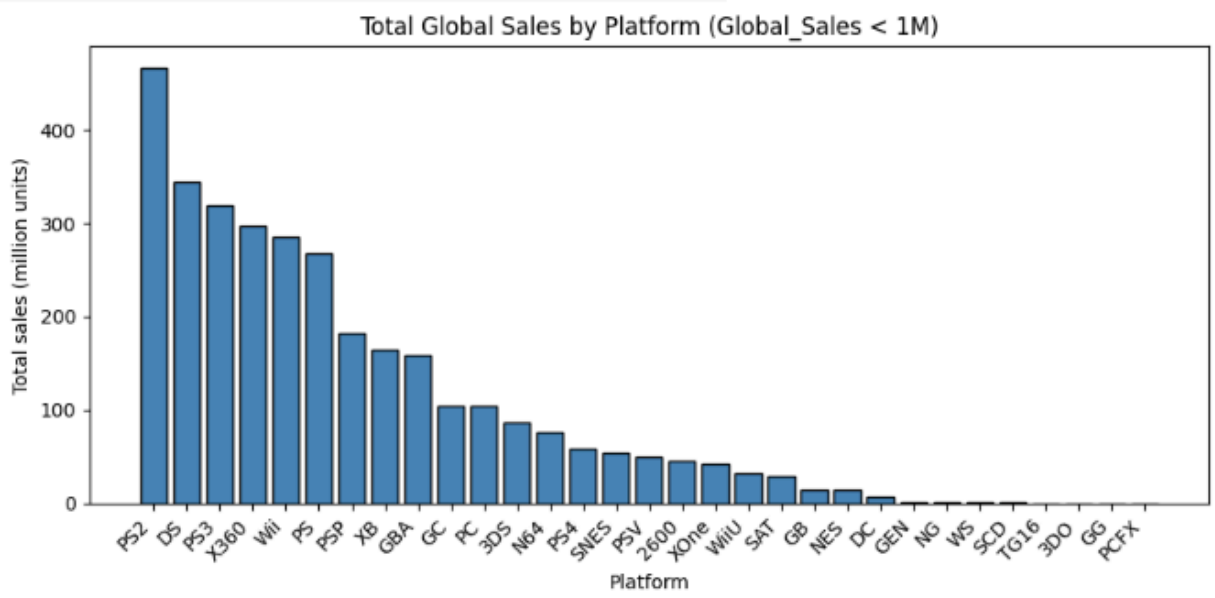


Figure 3.3 Comparision of Total Sales Between Gaming Platforms

The comparison of total sales by platform (for games below one million units) reveals clear differences in market performance across console generations.

The PlayStation 2 (PS2) leads with approximately 468 million units sold, representing 14.5% of total sales in this subset. It is followed by Nintendo DS (10.7%), PlayStation 3 (9.9%), and Xbox 360 (9.6%), confirming the dominance of Sony and Microsoft platforms during the mid-2000s console era.

The Wii and PlayStation (PS1) also maintain strong positions, while older or less common systems such as 3DO, GG, or PC-FX contribute negligibly to total sales.

Median sales per title are relatively consistent across major consoles (≈0.15–0.20 M), suggesting similar sales patterns despite platform differences.

Overall, the analysis shows that PlayStation and Xbox families account for the majority of game sales, while Nintendo maintains a strong but more selective market presence.

3.4 Analysis of Sales Performance by Publisher

```
=== Top 10 Publishers by Total Sales (Global_Sales < 1M, in million units) ===
```

	Publisher	titles	total_sales	mean_sales	median_sales	p90_sales	share_%
1	Electronic Arts	1009	372.28	0.368959	0.32	0.770	11.64
2	Activision	814	238.53	0.293034	0.23	0.640	7.46
3	Ubisoft	804	203.85	0.253545	0.17	0.620	6.37
4	THQ	627	186.06	0.296746	0.22	0.630	5.82
5	Sony Computer Entertainment	534	171.98	0.322060	0.24	0.767	5.38
6	Namco Bandai Games	879	161.73	0.183993	0.11	0.452	5.06
7	Konami Digital Entertainment	770	154.37	0.200481	0.13	0.470	4.83
8	Nintendo	364	132.30	0.363462	0.30	0.740	4.14
9	Sega	565	131.82	0.233310	0.15	0.546	4.12
10	Take-Two Interactive	322	92.69	0.287857	0.20	0.669	2.90

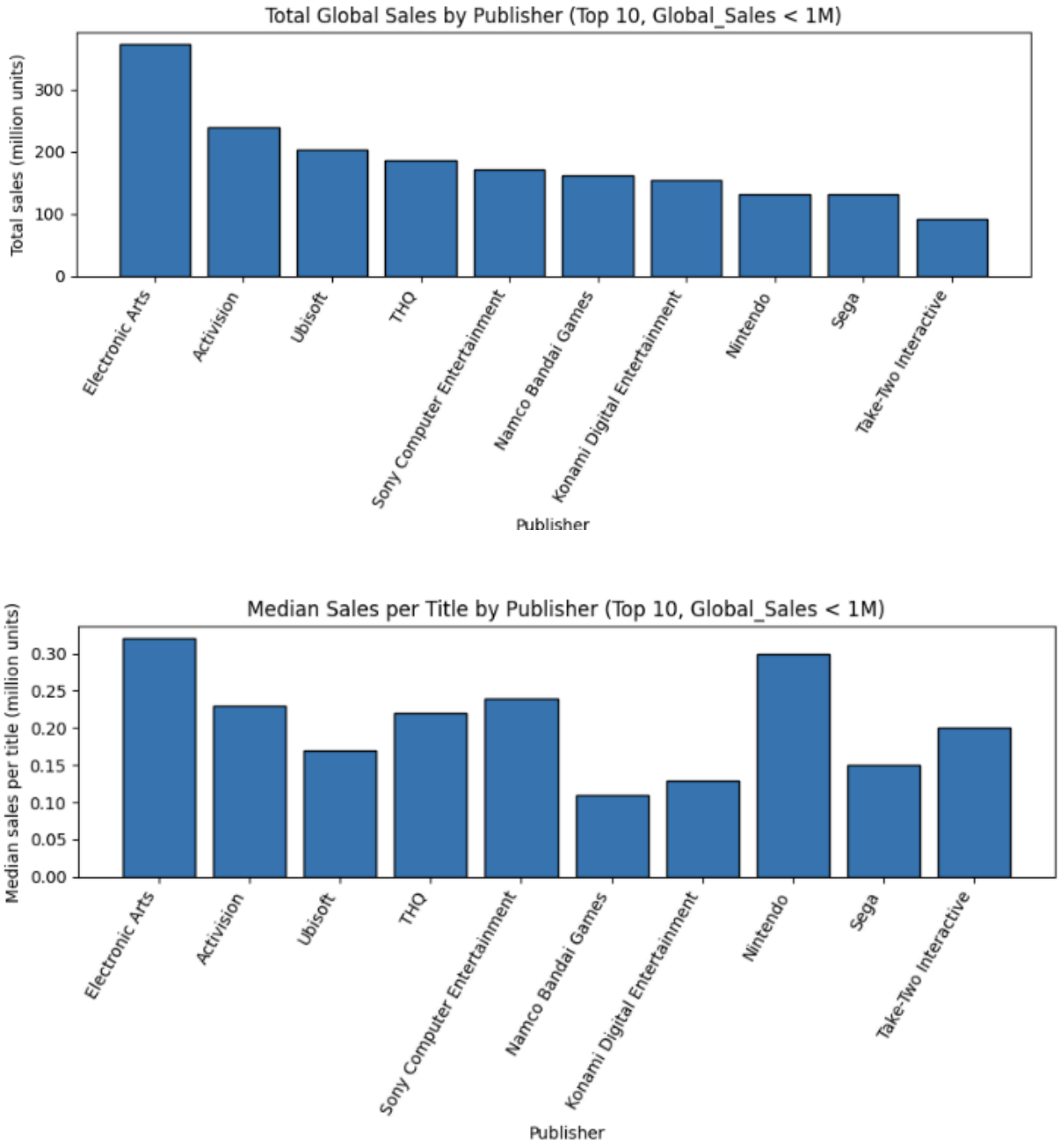


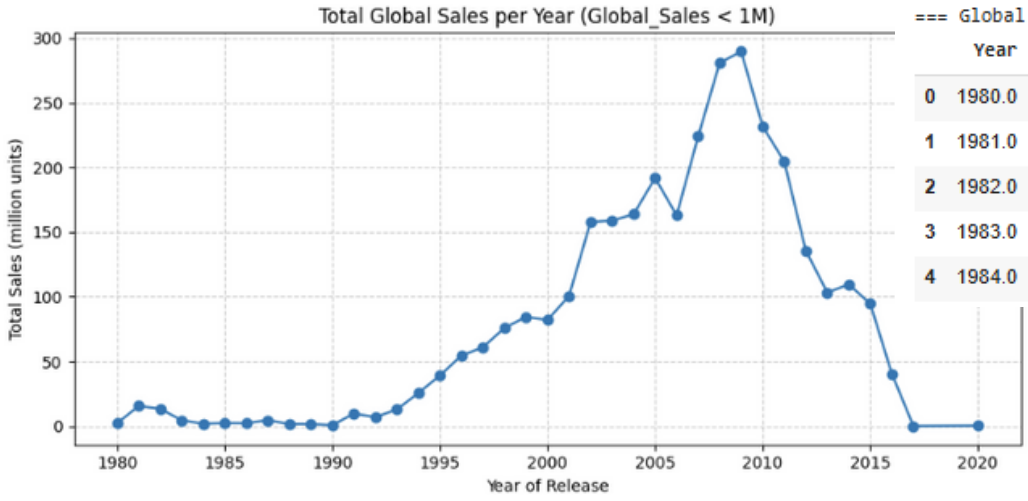
Figure 3.4 Analysis of Sales Performance by Publisher

The analysis of sales distribution across publishers (for titles below one million units) highlights the market dominance of a few key companies. Electronic Arts (EA) leads the ranking with total sales of approximately 372 million units, accounting for over 11% of the dataset. Activision and Ubisoft follow with 7.5% and 6.4%.

Publishers such as THQ, Sony Computer Entertainment, and Namco Bandai Games maintain solid overall performance but with lower average sales per title. In contrast, despite having fewer releases within this subset, achieves one of the highest median sales per title (~0.30 million units), indicating consistently strong performance across its games.

Overall, the results show that a small group of major publishers generates the majority of global sales, while others operate with smaller but stable market shares. This pattern reflects brand strength, franchise continuity, and established player loyalty within the top-performing companies.

3.5 Examination of Global Sales Trends Over Time



=== Global Sales Trends Over Time (Global_Sa

	Year	titles	total_sales	mean_sales
0	1980.0	5	2.11	0.422000
1	1981.0	35	15.48	0.442286
2	1982.0	29	13.34	0.460000
3	1983.0	10	4.52	0.452000
4	1984.0	3	1.78	0.593333

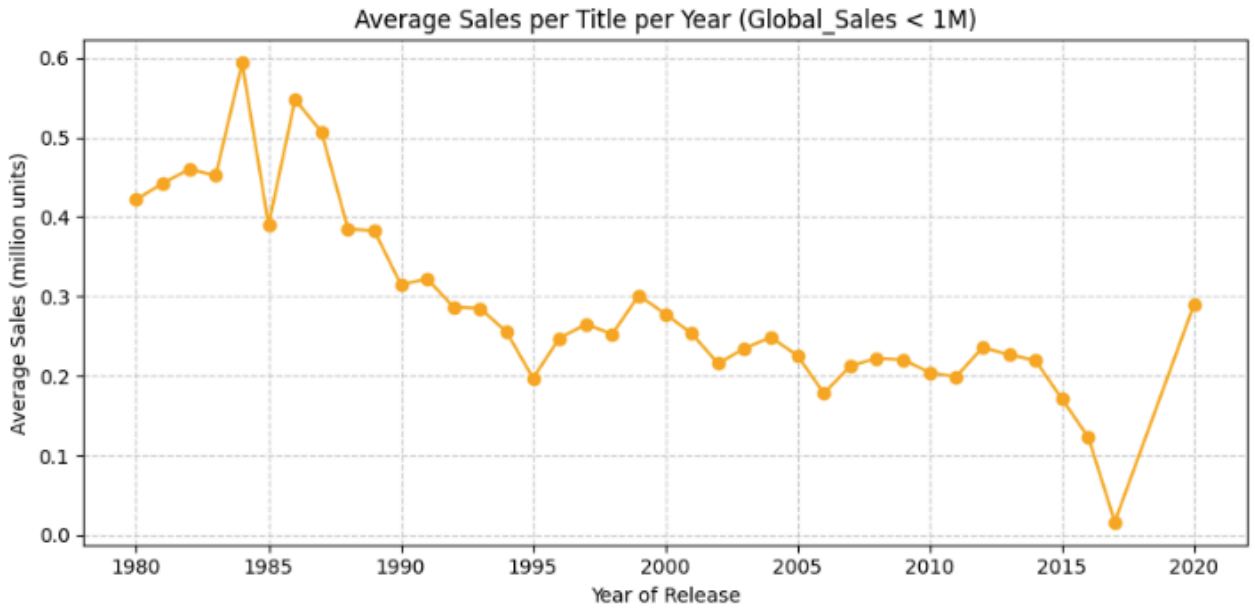


Figure 3.5 Global Sales Trends Over Time

The analysis of annual global sales (limited to titles under one million units) reveals clear temporal patterns in the video game industry. From the early 1980s through the mid-1990s, total sales remained relatively low, reflecting the smaller market size and limited distribution of early gaming platforms.

Starting around 2000, both the number of releases and total sales began to rise sharply, peaking between 2007 and 2010 with annual totals exceeding 250 million units. This period coincides with the commercial success of major consoles such as the PlayStation 3, Xbox 360, and Nintendo Wii.

After 2010, a steady decline is visible, total sales dropped as digital distribution and mobile gaming became dominant, reducing the number of physical game releases captured in the dataset.

Average sales per title show a different pattern: while early years exhibit higher averages due to fewer but stronger titles, the metric gradually decreases over time as the market became more saturated and fragmented.

Overall, the trend indicates a mature and competitive industry, one that reached its peak in the late 2000s and later transitioned toward digital ecosystems and diversified gaming platforms.

3.6 Identification of the top 10 best selling games

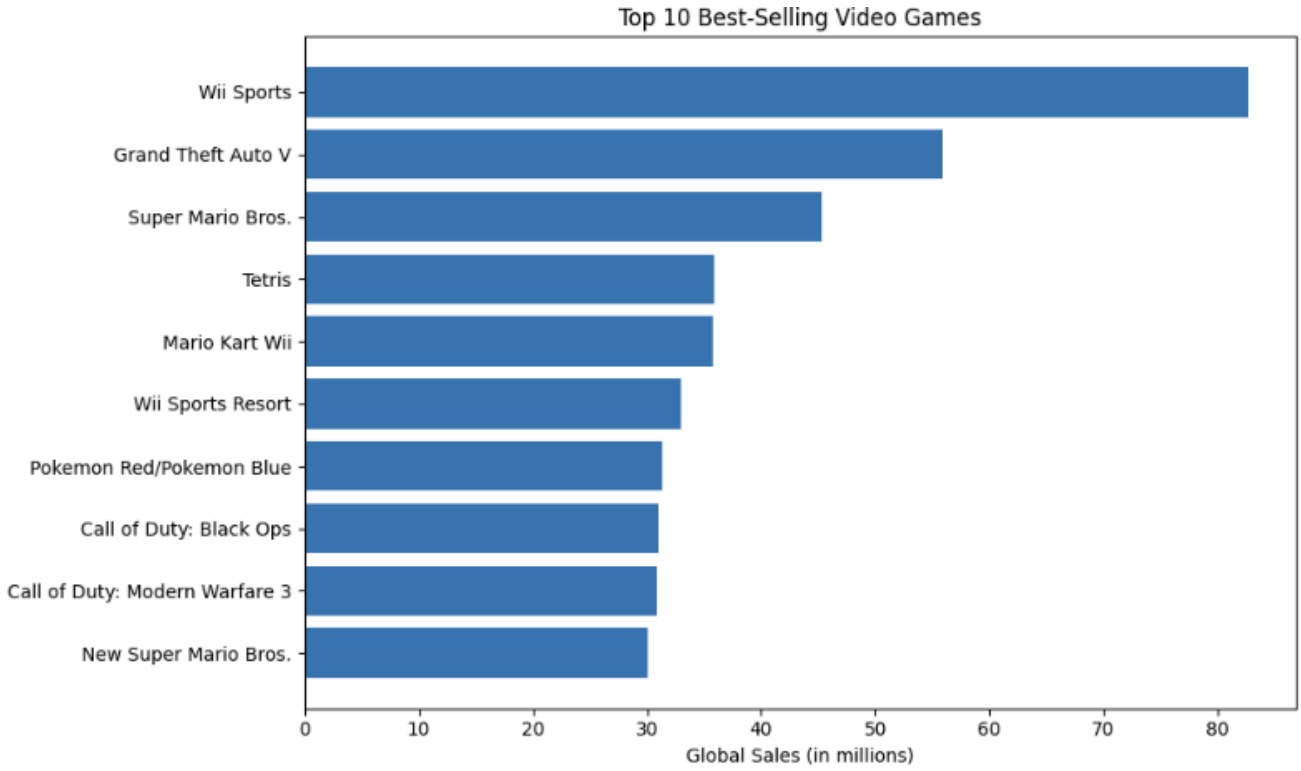


Figure 3.6 Identification of the top 10 best selling games

To identify the most commercially successful video games, the dataset was grouped by game title and sorted by total global sales.

The results show that Wii Sports leads by a large margin with over 80 million units sold worldwide, followed by Grand Theft Auto V and Super Mario Bros.. Most of the top titles belong to globally recognized franchises from Nintendo and Rockstar Games

3.7 Correlation Analysis Between Numerical Variables

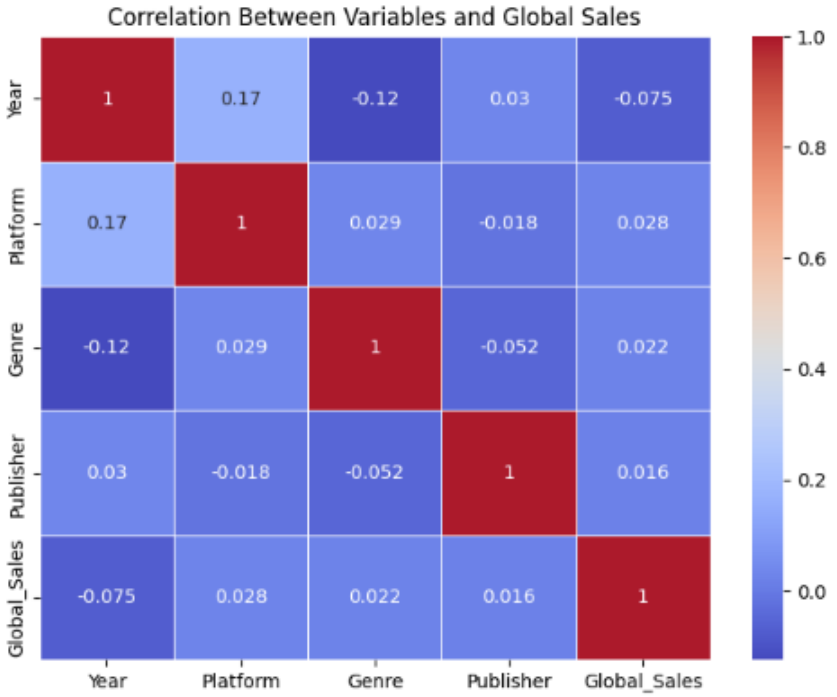


Figure 3.7 Pearson Correlation Matrix Between Variables and Global Sales

To examine potential relationships between numerical and categorical variables, the categorical features (Platform, Genre, Publisher) were first numerically encoded, and a Pearson correlation matrix was calculated.

The results reveal only very weak correlations between Global_Sales and the other variables, indicating the absence of any clear linear dependency.

To further investigate possible non-linear effects, additional analyses using Spearman correlation and pairwise scatter plots were performed. These confirmed that no strong monotonic or non-linear relationships exist either.

Overall, the findings suggest that total global sales are not driven by a single factor, but rather by complex interactions between multiple attributes.

3.8 Summary of Key Insights Derived from the Dataset

The exploratory data analysis provides several key insights into the structure and dynamics of the global video game market.

Most games achieve modest commercial success, with global sales typically below one million units. The distribution of sales is highly right-skewed, confirming that the market follows a long-tail pattern.

Genre-based analysis highlights that Action and Sports games dominate the market, together accounting for nearly 40% of total sales. These categories benefit from mainstream appeal, frequent franchise releases, and cross-platform availability. In contrast, genres such as Adventure, Strategy, and Puzzle remain niche segments with lower market penetration but potentially loyal audiences.

Platform-level analysis demonstrates the historical dominance of Sony's PlayStation and Microsoft's Xbox families, particularly during the 2000s console generation. The PlayStation 2 leads overall, reflecting its extensive lifespan and broad game library. Nintendo maintains strong but more selective market success, supported by exclusive titles and family-oriented genres.

At the publisher level, the market shows significant concentration. A few major companies, Electronic Arts, Activision, and Ubisoft, account for the majority of global sales, driven by established franchises and consistent annual releases. Smaller publishers contribute less overall but often maintain stable performance within specific genres or regions.

The temporal analysis reveals that the industry peaked between 2007 and 2010, coinciding with the commercial success of major seventh-generation consoles. Since then, sales volumes have declined due to the rise of digital distribution, mobile gaming, and platform diversification.

Finally, correlation analyses (Pearson and Spearman) indicate that there are no strong linear or monotonic relationships between single features (e.g., platform, genre, or publisher) and global sales. This confirms that game performance depends on multi-dimensional and non-linear interactions.