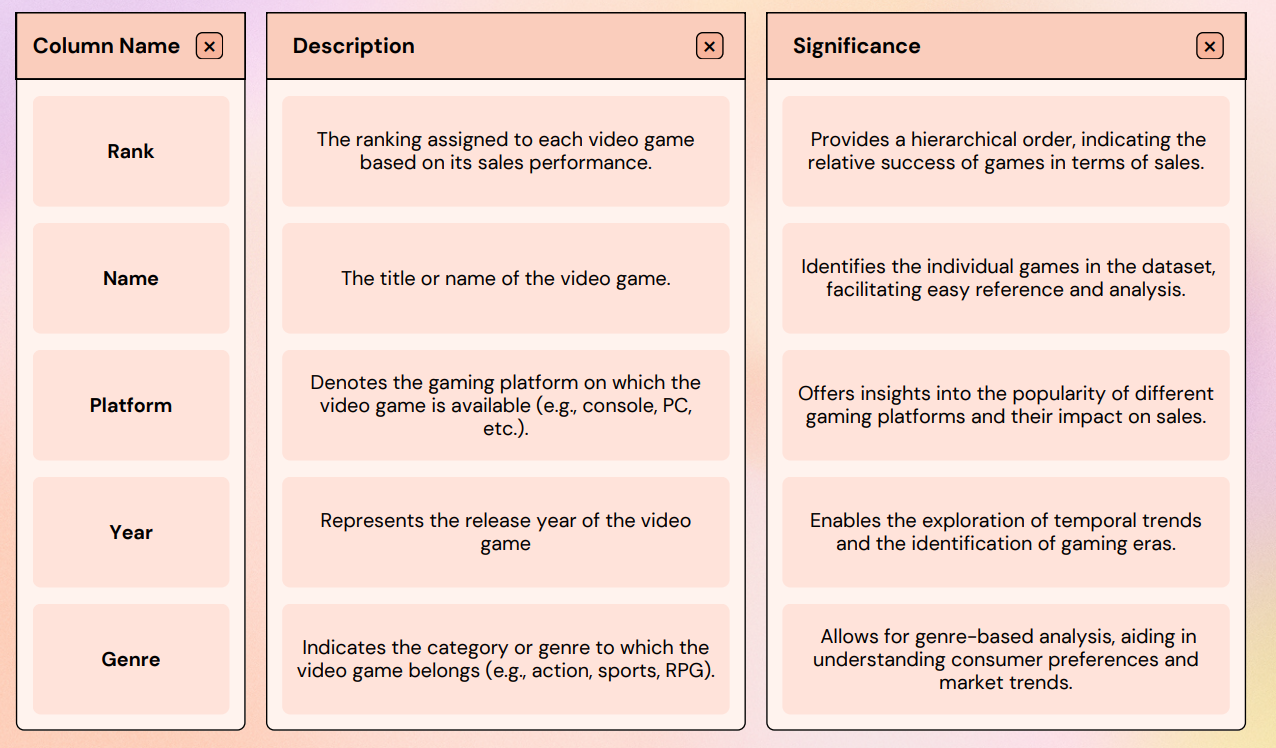
**Video Games Sales – Data Analysis**

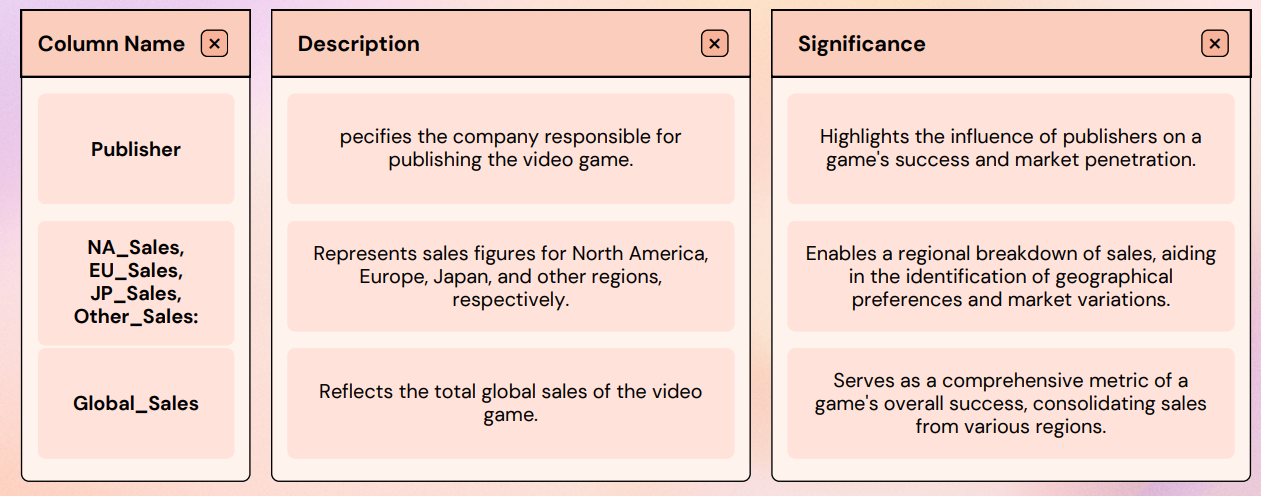
**Introduction:**

The dataset under consideration comprises a comprehensive list of video games with sales exceeding 100,000 copies, generated through the web scraping of vgchartz.com. This dataset encompasses key attributes that provide valuable insights into the video game industry. The columns include:

* Rank: The ranking of each video game in terms of sales.
* Name: The title of the video game.
* Platform: The gaming platform on which the game is available.
* Year: The year the video game was released.
* Genre: The category or genre of the video game.
* Publisher: The company responsible for publishing the game.
* NA\_Sales: Sales figures in North America.
* EU\_Sales: Sales figures in Europe.
* JP\_Sales: Sales figures in Japan.
* Other\_Sales: Sales figures in other regions.
* Global\_Sales: Total global sales.

Exploring this dataset is essential to unraveling trends, patterns, and influential factors within the video game industry. Understanding the sales performance of video games provides valuable insights for developers, publishers, and stakeholders. By delving into this data, we can identify successful genres, platforms, and geographical preferences, aiding in strategic decision-making and potentially uncovering factors contributing to the industry's evolution over time. Through exploratory data analysis (EDA), we aim to derive meaningful conclusions and inform future directions in the dynamic landscape of video game sales.

**Data Set Overview:**

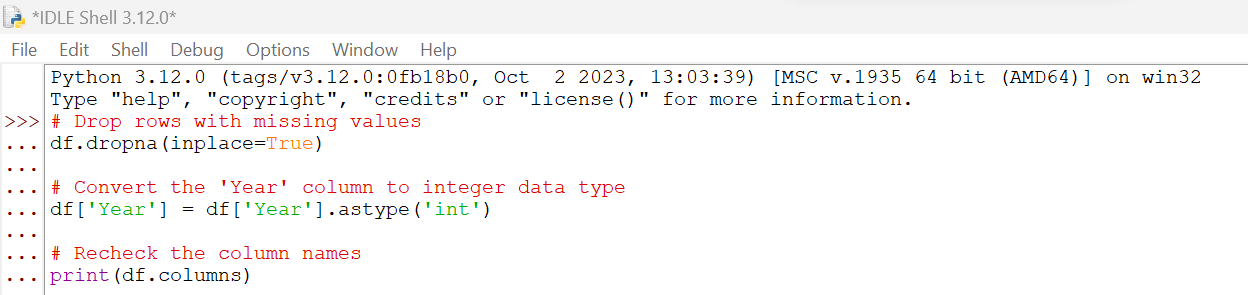
**Data Composition(before cleaning):**

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute: | Type: | Missing Values: | Feature Importance: |
| Rank | Int64 | 0 | Low |
| Name | Object | 0 | High |
| Platform | Object | 0 | High |
| Year | Float64 | 271 | High |
| Genre | Object | 0 | High |
| Publisher | Object | 58 | High |
| NA\_Sales | Float64 | 0 | High |
| EU\_Sales | Float64 | 0 | High |
| JP\_Sales | Float64 | 0 | High |
| Other\_Sales | Float64 | 0 | High |
| Global\_Sales | Float64 | 0 | High |

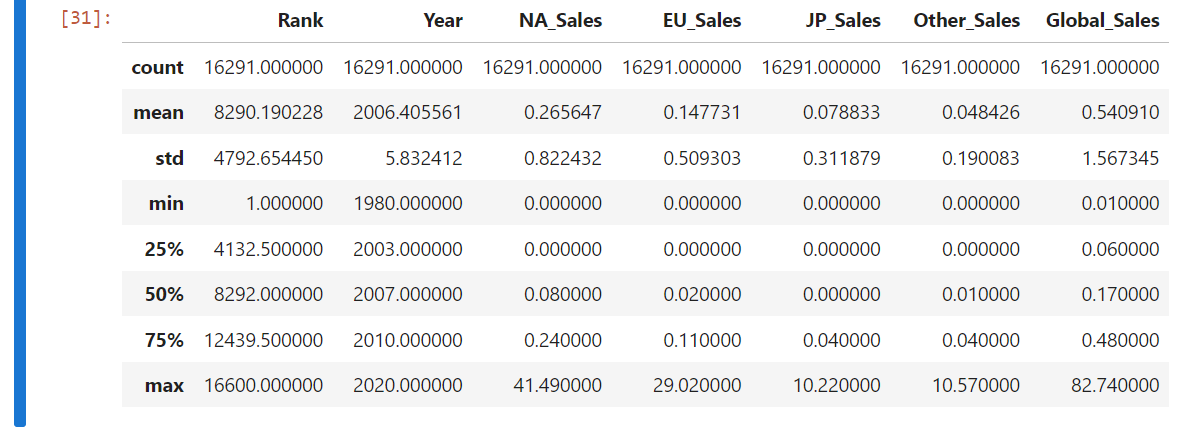
**Data Cleaning:**

* **Analyzing and Changing Data Types**

The initial part of the code displays the shape of the dataset, the first few rows, general information about the dataset, and the count of missing values in each column. This step is crucial to understanding the structure of the dataset and identifying potential data type issues and missing values.



* **Dropping Rows with Missing Values:**
  + **Strategy:** The code uses **df.dropna(inplace=True)** to remove rows with missing values. This strategy is chosen when the number of missing values is relatively small compared to the dataset size, and removing them doesn't significantly impact the analysis.
  + **Reasoning:** This approach ensures that we retain a substantial amount of data for analysis while removing instances where critical information is missing.
* **Converting 'Year' to Integer Data Type:**
  + **Strategy:** The code uses **df['Year'] = df['Year'].astype('int')** to convert the 'Year' column to integer data type. This is done to facilitate numerical analysis and visualization based on the release years of video games.
  + **Reasoning:** Treating the 'Year' column as an integer allows for easier manipulation and interpretation in the context of temporal trends and analysis.
* **Rechecking Column Names:**
  + **Strategy:** The code includes **print(df.columns)** to recheck the column names after data cleaning.
  + **Reasoning:** It ensures that the column names are consistent and accurate for subsequent analysis steps.

**Descriptive Statistics:**

* **Observations:**
* Rank: Ranges from 1 to 16,600, indicating the ranking of games based on sales.
* Year: Spans from 1980 to 2020, reflecting the release years of video games.
* **Sales Figures (in millions):**
* NA\_Sales: Range from 0 to 41.49 million.
* EU\_Sales: Range from 0 to 29.02 million.
* JP\_Sales: Range from 0 to 10.22 million.
* Other\_Sales: Range from 0 to 10.57 million.
* Global\_Sales: Range from 0.01 to 82.74 million.

**Top Genre and Count:**

* Identify the most popular genres based on the number of games.
* Type: Univariate Analysis – Genre Distribution

**Number of Games Released by Year:**

* Explore how the number of game releases has evolved over different years.
* Univariate Analysis - Temporal Trends.

**Total Sales by Region:**

* Understand the distribution of sales across North America (NA), Europe (EU), Japan (JP), and Other regions.
* Univariate Analysis - Regional Sales Distribution.

**Highest Global Sales by Year:**

* Identify the games with the highest global sales in each year.
* Type: Univariate Analysis - Temporal Trends.

**Top 10 Publishers by Global Sales:**

* Analyze the impact of different publishers on global sales, highlighting the top performers.
* Type: Univariate Analysis - Publisher Impact.

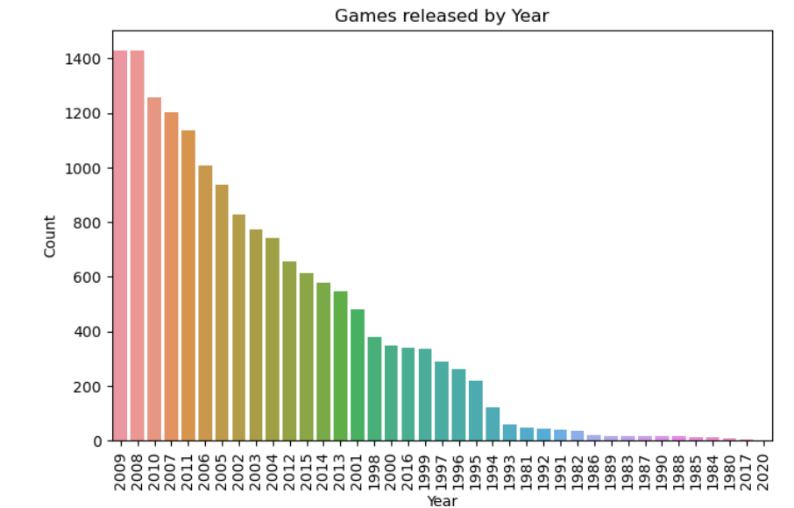
**Video Sales Comparison across EU, Japan and NA:**

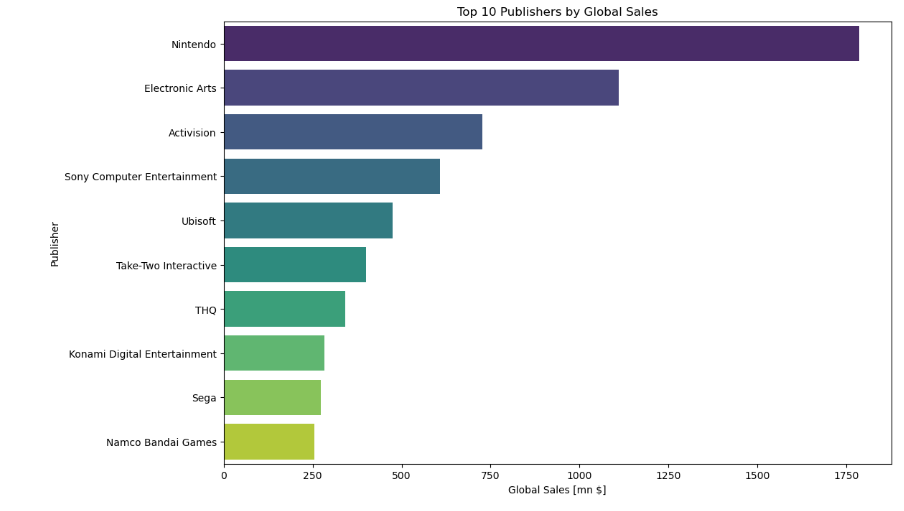
* Compare and contrast sales figures across Europe (EU), Japan (JP), and North America (NA).
* Type: Bivariate Analysis - Regional Sales Comparison.

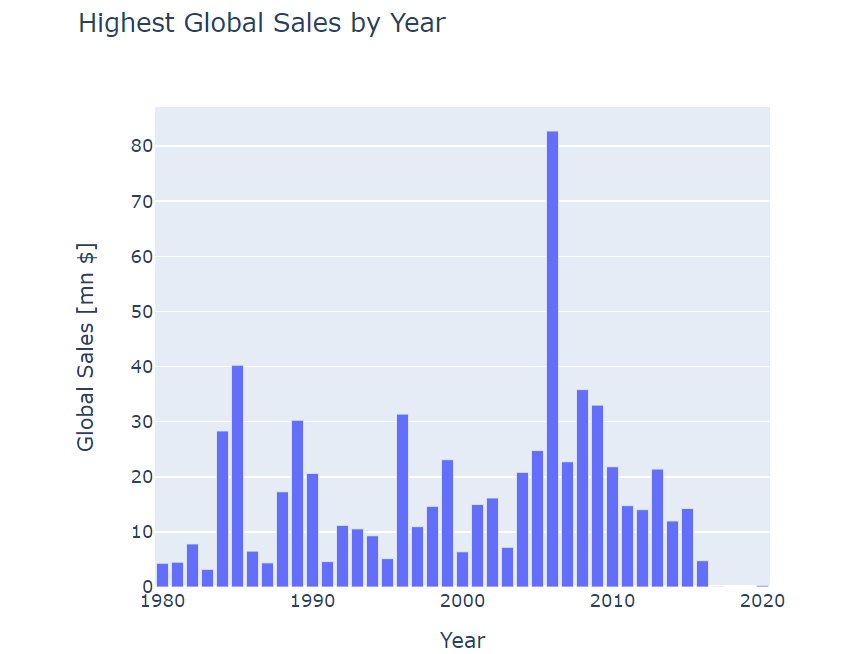
**Correlation Matrix of Numerical features:**

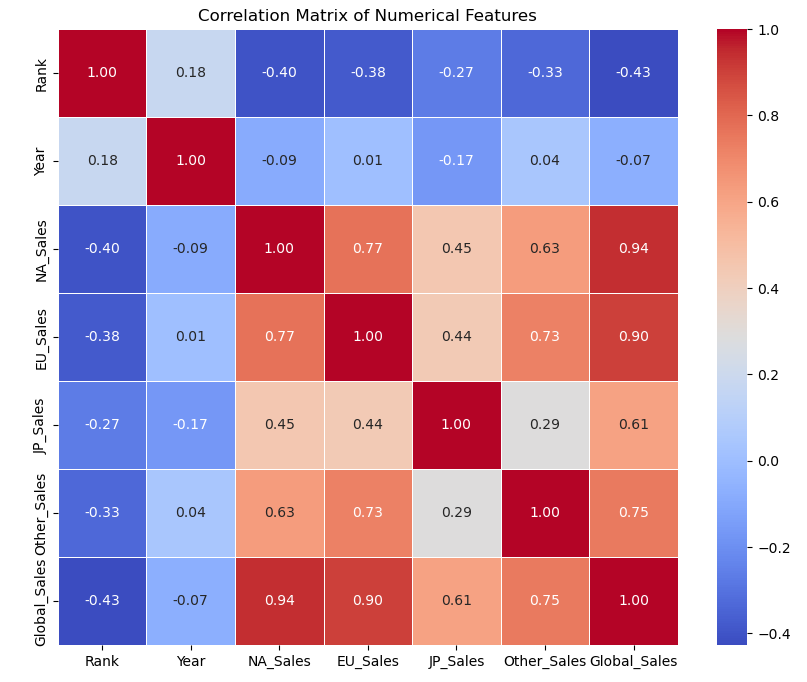
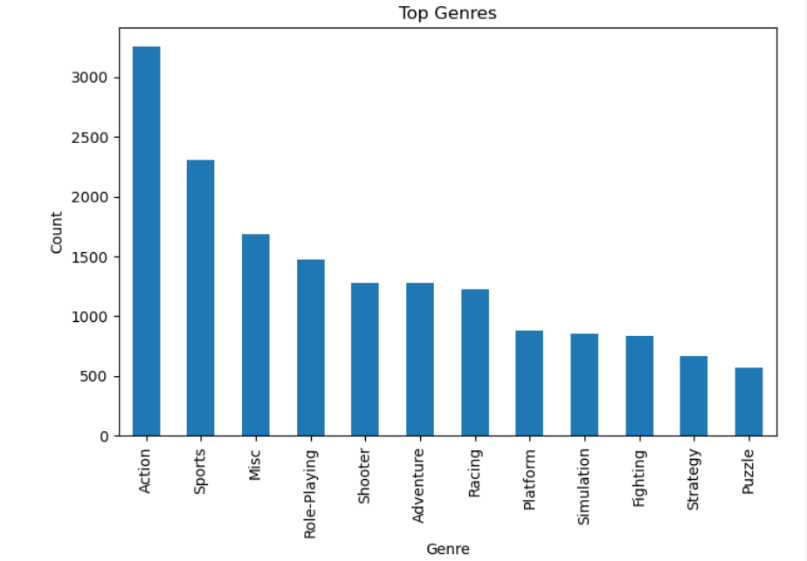
* Investigate relationships between numerical features, exploring correlations between sales figures.
* Type: Bivariate Analysis - Numerical Features Correlation.

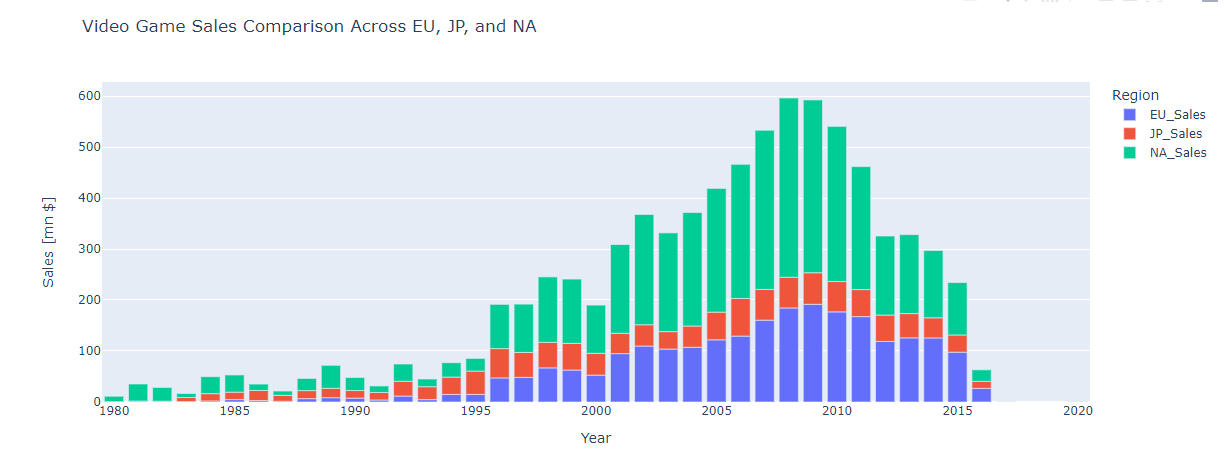
A graph with different colored rectangles

Description automatically generated

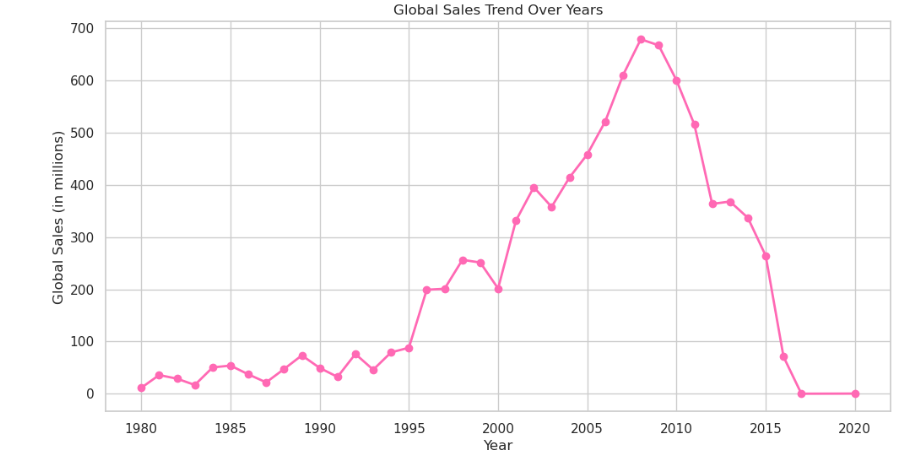






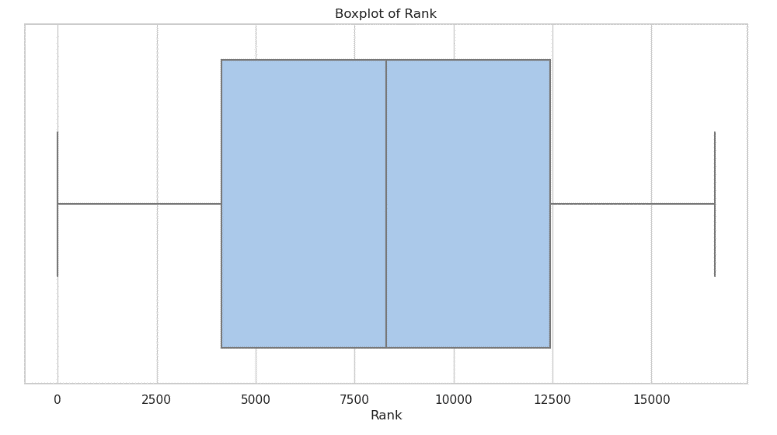


**Univariate Analysis:**

****

* The chart above shows the trend line of global sales, which peaked in around 2007 and 2008. The dataset did not include data for 2015 onwards, so the line is flat at 0. Overall, global sales rose with dips in some years, 2007 being the golden year for the highest global sales of video games. The decline in sales can be due to piracy of video games, competition from other entertainment platforms like social media, transition to new console generations, etc.
* Since Rank is acting like a serial number in the data set, it does not provide valuable insights for analysis. However, if we plot a distribution graph and box plot for it, its distribution is normal. QQ plots have been added at the end of the dataset to further solidify the evidence of normality in the columns of Year and Rank, where a straight red line is an indication of normality. The skewness is towards the right as shown by the whiskers(right).

A graph with numbers and lines

Description automatically generated

* The box plot of the year shows that data is skewed towards the left side. According to the QQ plot (Python notebook), the red line indicates the normality of the data.

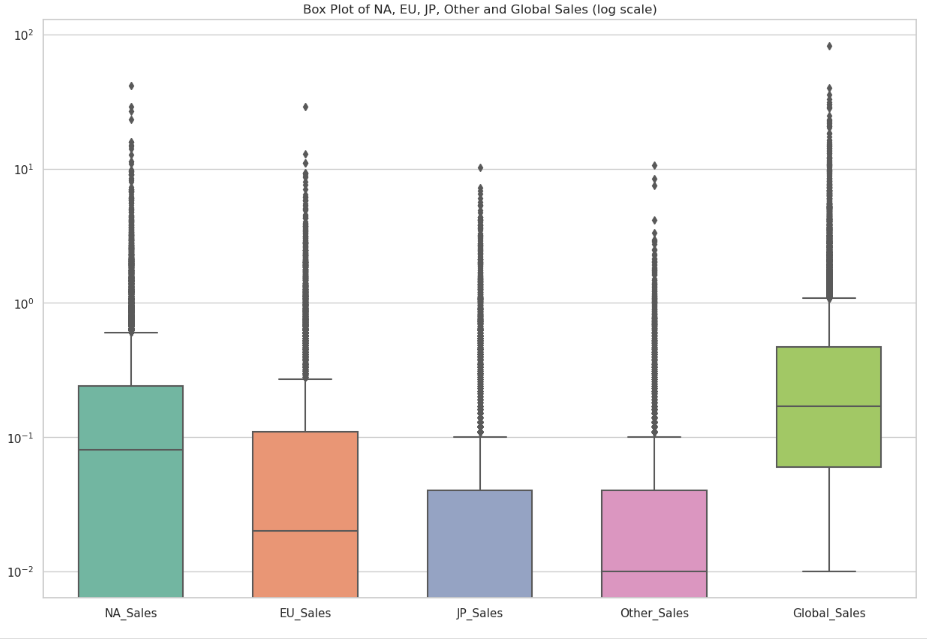
A graph showing a blue rectangle

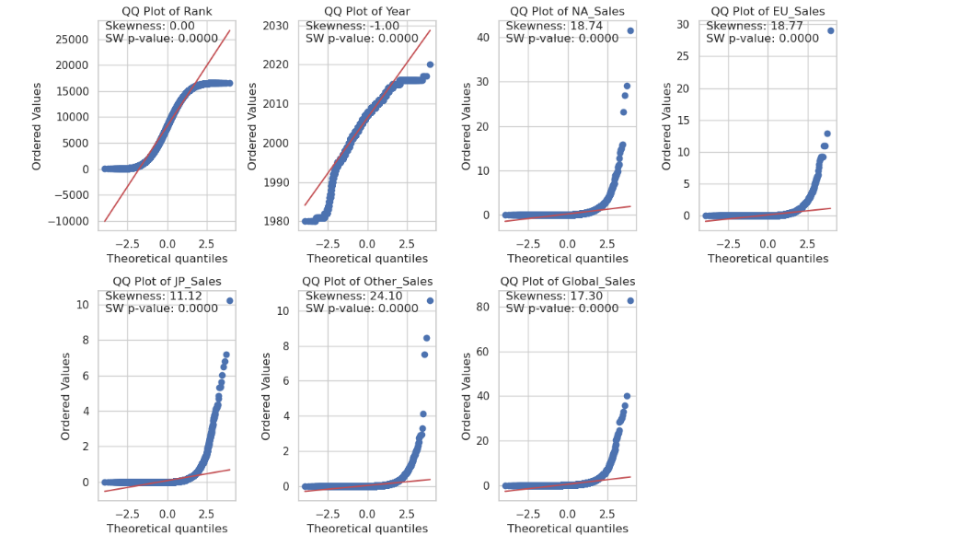
Description automatically generatedA graph with orange lines

Description automatically generated

* The histograms of Sales by region does not show noticeable differences between the columns, so we can use QQ plots and box plots to figure out the distribution and skewness. No straight red line on the QQ plots show absence of normality in NA, EU, JP, Other and Global Sales.
* NA\_Sales : Left skewed
* JP\_Sales: Right skewed
* EU\_Sales: Left skewed
* Other\_Sales: Left skewed
* A graph of sales in regions

  Description automatically generatedGlobal\_Sales: Right skewed(indicated by the whisker extending above, since the boxes do not show a noticeable difference in deviation towards the left or right.)

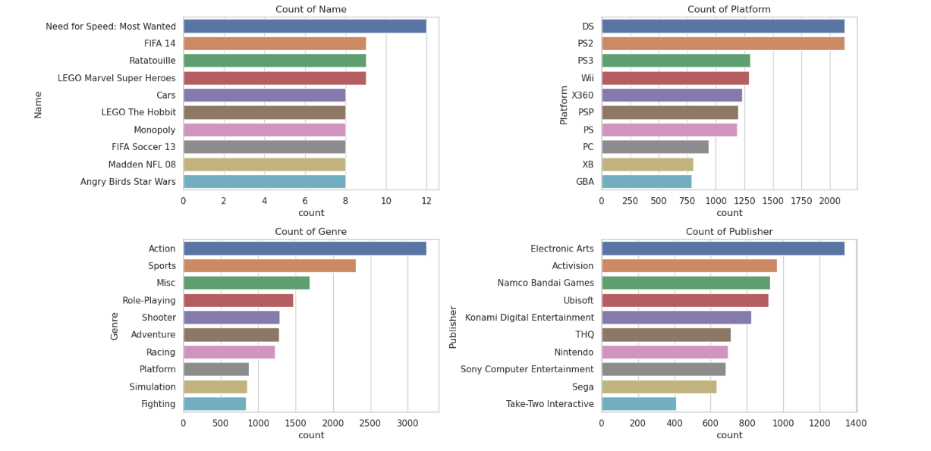




* "p-value may not be accurate for N > 5000," is a common warning when using the Shapiro-Wilk normality test with large sample sizes. It indicates that the p-value might not be accurate or meaningful for very large datasets. To resolve this warning, I considered alternative approaches for assessing normality, especially for large samples, like QQ plots as shown above.

**Charts for Categorical Variables:**

A group of pie charts

Description automatically generated

* Need for Speed: Most Wanted was the most popular video game and Action is the most prevalent genre.
* Electronic Arts is the most popular Publisher and DS is the most popular platform.

**A screenshot of a graph

Description automatically generatedBivariate Analysis:**

* NA\_Sales is an independent variable as plotted on the x-axis. Global\_Sales takes into account the cumulative sales of NA, EU, JP, and other regions, thus it is a dependent variable. The scatter plot shows a positive correlation between the two variables and outliers are detected after a sales of $1 mn and close to $0.01 mn.

*NA Sales as an Influential Factor:*

The positive correlation emphasizes the importance of NA Sales as a significant contributing factor to overall Global Sales. Understanding and effectively managing sales strategies in the North American region could have a direct impact on a game's global success.

*Threshold Concentration:*

The concentration of data around specific sales values, such as 1 million and 0.1 million dollars, highlights the importance of recognizing these thresholds in sales planning. Games that reach or exceed these values may share common characteristics or market conditions that contribute to their success.

In conclusion, the bivariate analysis of the scatter plot provides insights into the relationship between NA Sales and Global Sales. Recognizing the positive correlation and concentration of data at specific sales values can inform strategic decisions for game developers and publishers, emphasizing the importance of tailoring marketing and distribution strategies to maximize sales potential, particularly in the North American market.

**Overall Domain Analysis:**

The business process involves managing the sales and distribution of video games across various platforms, genres, and regions.

*Sales Performance Analytics:*

* Analyzing sales figures (NA, EU, JP, Global) to understand the performance of video games in different markets.
* Identifying popular genres and platforms to optimize development and marketing strategies.
* Assessing the influence of different publishers on global sales to make informed partnership decisions.
* Exploring sales trends over time to adapt strategies based on seasonal variations.
* Understanding regional preferences to tailor marketing and distribution strategies.

*Strategies:*

* Insights from analytics can guide marketing efforts, allowing for targeted campaigns based on popular genres, platforms, or regions.
* Efficient allocation of resources by focusing on genres or platforms with higher sales potential, reducing risks associated with less profitable segments.
* Identifying untapped markets or potential for growth in specific regions to guide expansion strategies.
* Choosing publishers with a track record of success in specific genres or regions to enhance the chances of global success.

*Strategic Game Development:*

- Informing decisions on game development by aligning with popular genres and platforms.

*Global Distribution Strategies:*

- Supporting decisions related to global distribution based on historical sales trends.

- Facilitating decisions on publisher partnerships through an understanding of their impact on sales.

The analytics derived from the video game sales dataset contribute to data-driven decision-making, optimizing various aspects of the sales and distribution process. It allows stakeholders in the video game industry to stay competitive, adapt to market trends, and make informed choices to maximize sales and global reach.