**Data Analysis Report: Video games sales as 22_Dec_2016**

## 1. Introduction

This report presents a structured analysis of the Video games sales, focusing on data cleaning and exploratory data analysis (EDA). The goal is to refine the dataset for accurate insights by identifying patterns, relationships, and anomalies. The main objectives are:

- Cleaning the dataset to improve data quality.
- Handling missing values, duplicate records, and outliers.
- Performing in-depth statistical and visual analysis.

## 2. Data Cleaning

Ensuring clean data is essential for accurate analysis. The following steps were applied:

### 2.1 Handling Missing Values

A thorough check for missing values showed none were present. If any had been found, numerical values would have been replaced with the median, while categorical values would have been imputed using the mode.

### 2.2 Removing Duplicate Records

Duplicate records were checked to avoid redundancy. No duplicate entries were found in the dataset, confirming data uniqueness.

### 2.3 Outlier Detection and Treatment

Outliers were detected using the Interquartile Range (IQR) method, which isolates extreme values in numerical columns. The dataset was found to be free of significant outliers, ensuring data consistency.

### 2.4 Standardization of Categorical Values

To eliminate inconsistencies, categorical values such as product names and brands were standardized by converting them to lowercase and trimming unnecessary spaces. This ensures uniformity in text-based analysis.

## 3. Exploratory Data Analysis (EDA)

### 3.1 Univariate Analysis

Univariate analysis provides insights into individual variables:

- **Summary Statistics:** Measures such as mean, median, variance, and skewness were computed.
- **Frequency Distributions:** Categorical variables were analyzed to identify dominant categories.
- **Visual Representations:** Histograms and box plots were used to visualize numerical distributions and identify patterns.

### 3.2 Bivariate Analysis

Examining relationships between two variables helps in understanding interactions:

- **Correlation Analysis:** A heatmap was used to assess relationships between numerical variables.
- **Scatter Plots:** These were used to detect trends between continuous numerical features.
- **Categorical-Numerical Comparisons:** Box plots and bar plots were used to compare categorical variables against numerical attributes.

### 3.3 Multivariate Analysis

This step extends analysis to interactions among multiple variables:

- **Pair Plots:** These visualized interactions between numerical features.
- **Heatmaps:** These provided a comprehensive view of correlations across multiple variables.
- **Grouped Analysis:** Categorical and numerical variables were analyzed together to reveal trends affecting product performance and pricing.

## 4. Key Findings

- The dataset was clean, with no missing or duplicate values.
- No significant outliers were detected in numerical features.
- Pricing and margin percentages varied across product categories.
- Strong correlations were found between global sales, publisher and name.
- Certain categories exhibited higher shelf life and stock level requirements, which could impact inventory strategies.

## 5. Conclusion

This analysis helped ensure data quality while uncovering valuable insights regarding pricing, stock levels, and category-specific variations. The findings can support better inventory management, pricing optimization, and strategic decision-making.