**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Data Visualisation Techniques* |
| **Assessment Title:** | *CA1\_DataViz\_HDip* |
| **Lecturer Name:** | David McQuaid |
| **Student Full Name:** | Federico Ariton |
| **Student Number:** | sba22090 |
| **Assessment Due Date:** | *13/03/2024* |
| **Date of Submission:** | *05/04/2024* |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Analysis and Visualization of Video Game Sales

# Introduction

This report delves into a detailed analysis of the video game industry using the dataset. Our goal is to extract key insights to inform the upcoming Winter season sales strategy for a retail company. We focus on identifying top-selling games, exploring correlations between regional sales, analyzing popular genres, comparing sales trends over time, and understanding the role of game developers. Additionally, we investigate the relationship between critic scores and global sales. Each analysis is complemented by carefully crafted visualizations, providing clear and actionable insights for strategic decision-making in the dynamic video game market.

# TOP 5 GAMES BY GLOBAL SALES

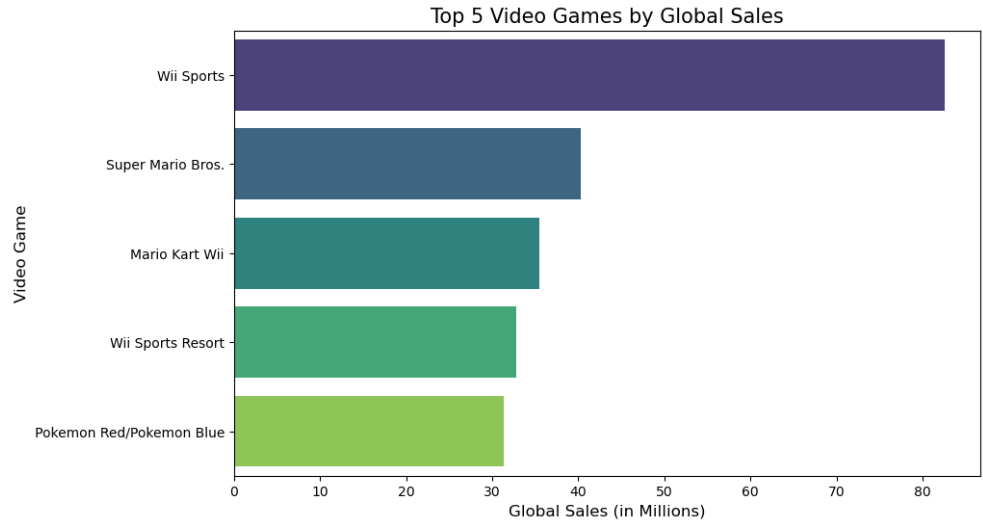


Figure 1 Bar chart of Top 5 Games by global sales

For this data, a bar chart(Figure 1) is an effective visualization tool because it allows for easy comparison of sales figures across different games.

The bar chart illustrates the top 5 video games by global sales. The horizontal layout and the 'viridis' color palette make it visually appealing and easy to interpret. Each bar represents a game, with the length proportional to its global sales, facilitating immediate comparison among the games.

The bar chart above displays the top 5 video games by global sales. Now, let's discuss the various aspects of this visualization:

Data Engineering:

Selection: The 'Global\_Sales' and 'Name' columns were selected for this analysis. 'Global\_Sales' provides the quantitative metric needed to rank the games, and 'Name' offers the qualitative aspect to identify the games.

Sorting: The dataset was sorted in descending order based on 'Global\_Sales' to identify the top 5 games.

Visualization Design:

Chart Type: A horizontal bar chart was chosen for clarity and ease of reading the game names.

Palette: The 'viridis' color palette offers a pleasing aesthetic and clear differentiation between bars.

Grid: A horizontal grid with a dashed style and mild alpha transparency improves readability without cluttering the visualization.

Font and Size: Care was taken to ensure text elements like titles, axis labels, and ticks are legible and appropriately sized.

Layout: Tight layout ensures that the chart fits well within the display area, making efficient use of space.

This visualization effectively highlights the top-selling games, which is crucial for the retail company's sales strategy, as it identifies popular games that might drive sales.

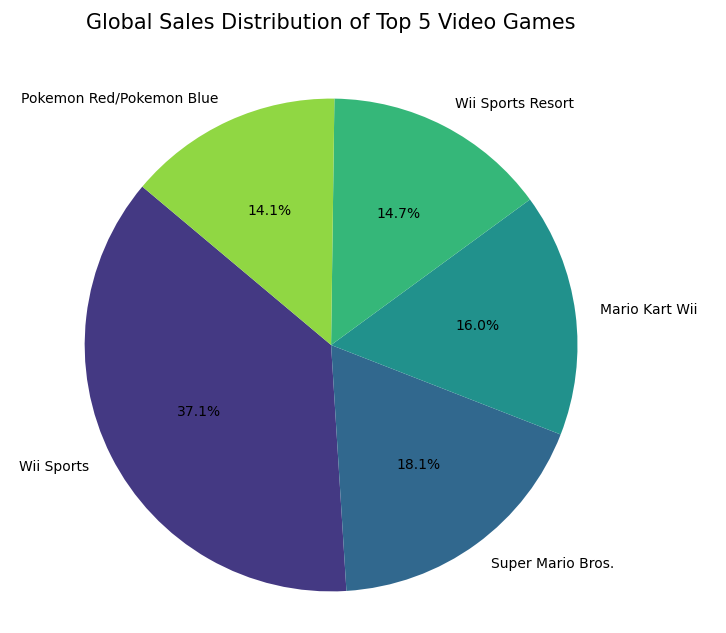


Figure 2 Pie chart Distribution of Top 5 Video Games

The pie chart offers(Figure 2) a different perspective compared to the bar chart, emphasizing the proportion of sales each game contributes to the total of these top games.

Each slice of the pie represents a game, and the size of the slice is proportional to its share of total sales among these top games. The use of a viridis color palette maintains consistency with the previous visualization, while clearly differentiating each game. Percentage values are included in each slice for precise understanding.

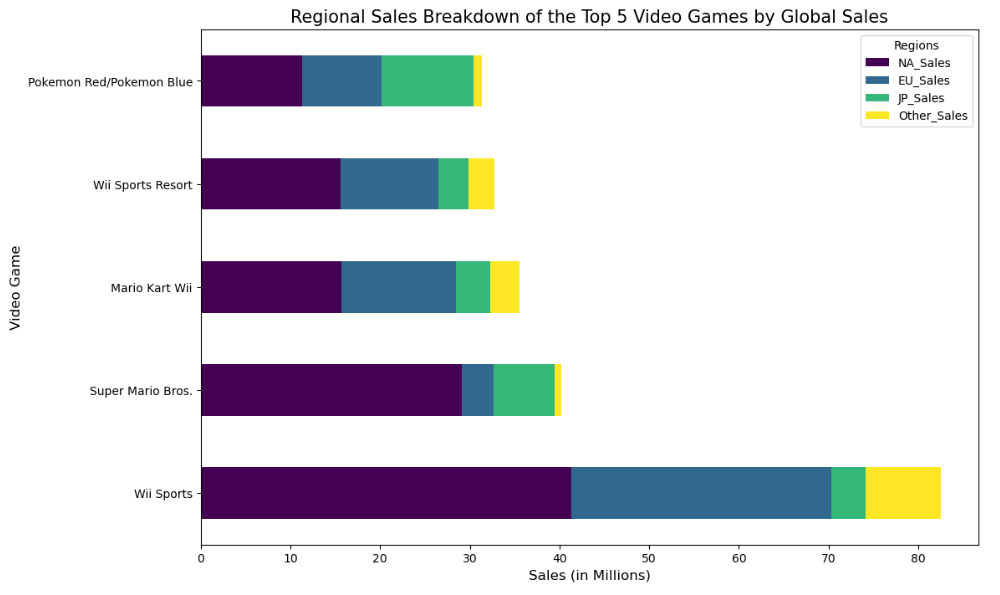


Figure 3 Stacked bar chart of the Top 5 video Games by global Sales

The stacked bar (Atlassian, n.d.) provides an insight into how the sales of each game are distributed across different regions: North America (NA), Europe (EU), Japan (JP), and Other regions.

Each horizontal bar represents a game, and the length of the colored segments within each bar indicates the sales in each region. The use of different colors for each region, derived from the viridis colormap, ensures clear differentiation and readability.

This visualization is particularly useful for understanding the geographic market strengths of each game. For example, a game with a significant portion of its bar in a specific color indicates strong sales in that region. This information can help the retail company tailor its marketing and stocking strategies to regional preferences, potentially focusing more on games that are popular in specific regions or exploring why certain games have broader or more limited geographic appeal. ​​

# CORRELATION BETWEEN NA SALES AND JP SALES

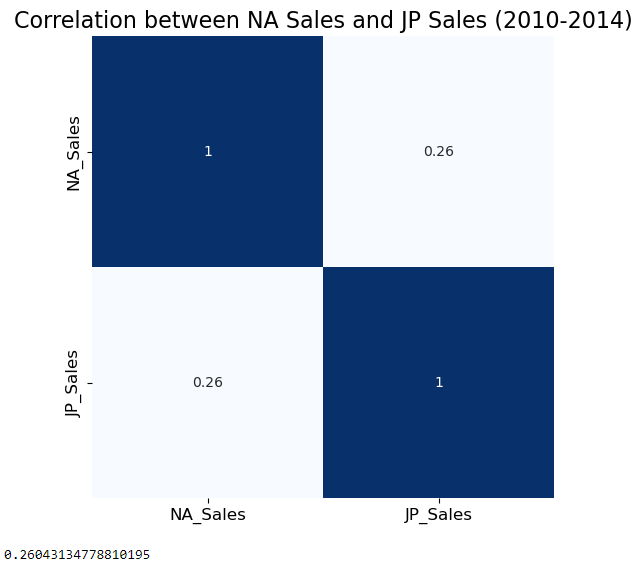


Figure 4 Heatmap correlation between NA Sales and JP Sales (2010-2014)

The correlation between NA\_Sales and JP\_Sales for the years 2010-2014 is 0.26, indicating a weak positive correlation. This means that while there is some relationship between sales in these two regions, it's not strong.

Visualization Design:

A heatmap(figure 4) is used to visualize the correlation matrix. This format is intuitive for representing correlations, with color intensity indicating the strength of the relationship.

The colormap 'coolwarm' provides a clear distinction between positive (warm colors) and negative (cool colors) correlations.

The annotation displays the exact correlation values, making the heatmap more informative.

Data Engineering:

The dataset was filtered to include only the years of interest (2010-2014).

No missing values in NA\_Sales and JP\_Sales, ensuring the correlation calculation is based on complete data.

This analysis helps to understand the sales dynamics between North America and Japan during this period, which can be vital for regional sales strategies. The weak correlation suggests that sales trends in North America may not be a strong indicator of trends in Japan, implying that distinct regional strategies might be necessary

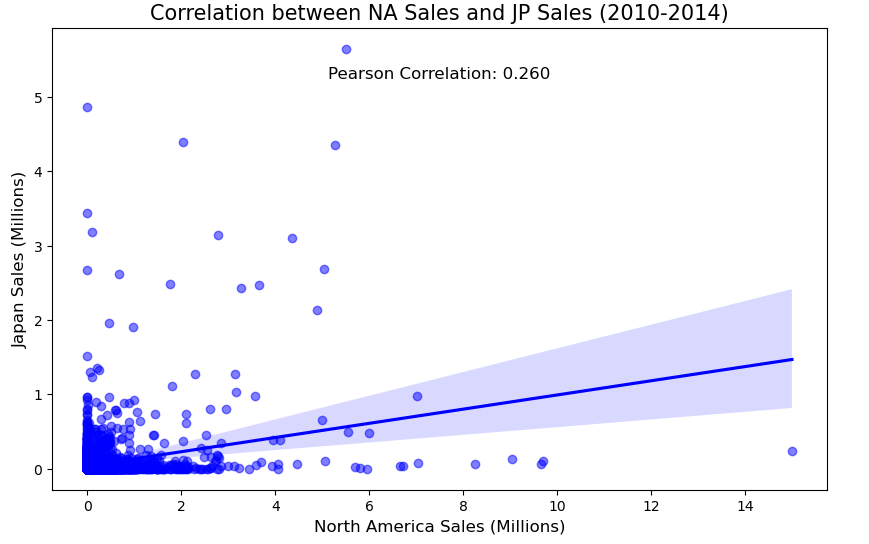


Figure 5 Scatter Plot implementing the Pearson correlation

Visualization: Enhanced Scatter Plot with Pearson (https://www.facebook.com/jason.brownlee.39, 2019)

This scatter plot, enhanced with a regression line and annotated with the Pearson correlation coefficient, provides a comprehensive visual representation of the relationship between North American (NA) Sales and Japanese (JP) Sales for the years 2010-2014.

Scatter Plot: Each point represents a game, with its position showing its sales in NA and JP. The spread of points indicates the nature of the correlation.

Regression Line: The blue line depicts the best fit linear trend through the data, visually illustrating the direction and slope of the relationship.

Correlation Annotation: The Pearson correlation coefficient is annotated on the plot, quantifying the strength of the linear relationship (approximately 0.260), confirming a positive but weak correlation.

Aesthetic Choices: The choice of blue color for points and the line, along with the transparency of points, strikes a balance between aesthetics and clarity.

This visualization effectively combines statistical information with graphical representation, making it easier to interpret the nature and strength of the relationship between the two sales regions.

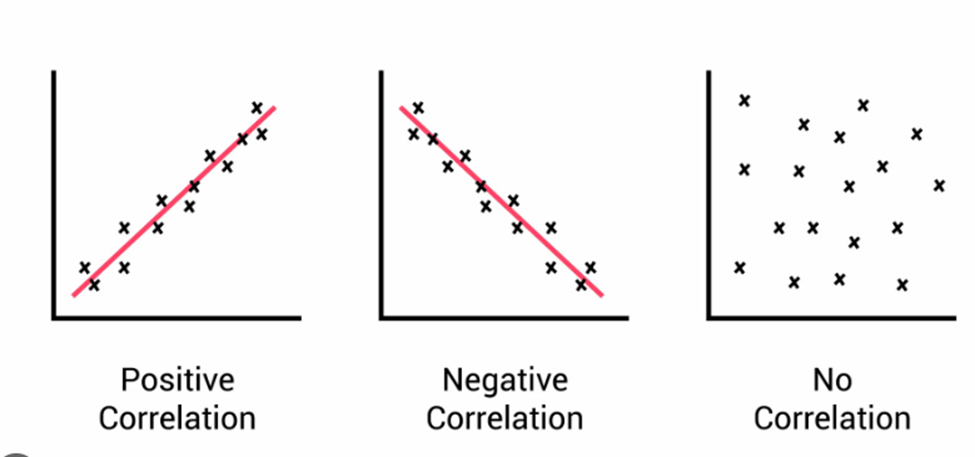


Figure 6 3 types of correlation (Mcleod, 2023)

The image with the 3 types of correlation(figure 6) was introduced to simplify the explanation of relationships in scatter plots. From our analysis, we observe that there is a positive correlation, although it is not as strong as the one depicted in the example of positive correlation.

# DISTRIBUTION OF THE MOST POPULAR 4 GAMES GENRES

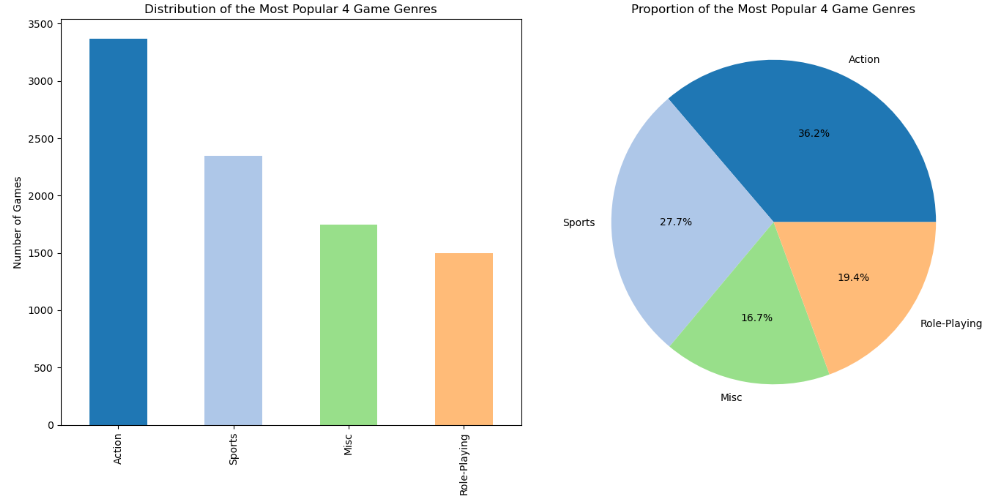


Figure 7 Bar chart and pie chart of the distribution the Most popular 4 games genres

To obtain the number of games per genre and their sales, methods like value\_counts() and groupby().sum() were utilized in Python, exploiting pandas' efficiency in handling such data.

Top Genre Selection: The head() method was applied to the aggregated results to focus on the top 4 genres. This selection process is grounded in presenting the most significant categories for concise and impactful visualization.

We combinate a bar chart and pie chart of the distribution of the Most popular 4 games genres (Figure 7)

Bar Chart (Number of Games):

Visualization Choice: A bar chart offers a clear comparison of discrete quantities. It's ideal for showing the count of items in each category.

Pie Chart (Sales Proportion):

Visualization Choice: A pie chart provides an intuitive representation of parts of a whole, making it a natural choice for displaying sales distribution.

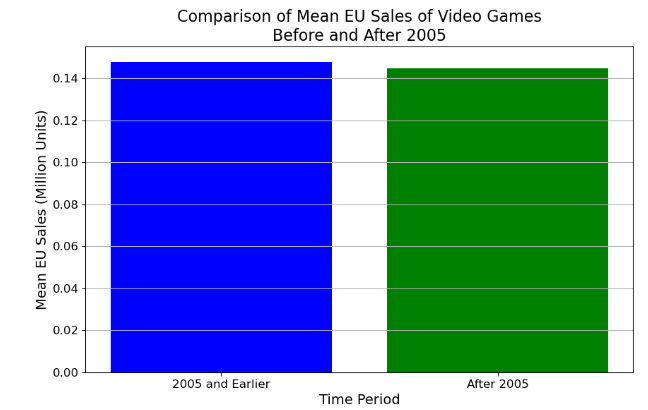
Color: Specific colors are assigned to each genre to maintain visual consistency across both charts, facilitating comparison.

Titles: Both charts are labeled with clear titles that directly inform the viewer of what the data represents.

Font, Text and size: The default font and text size were selected for readability, and the text is positioned to avoid clutter and ensure clarity, the charts are sized to be easily discernible while allowing for side-by-side presentation.

These design choices and methods collectively aim to create an informative and visually coherent presentation of the data.

# OLDER VS NEWER GAMES



Initial Analysis of Mean EU Sales:

We started by calculating the mean European sales for games released in 2005 and earlier.

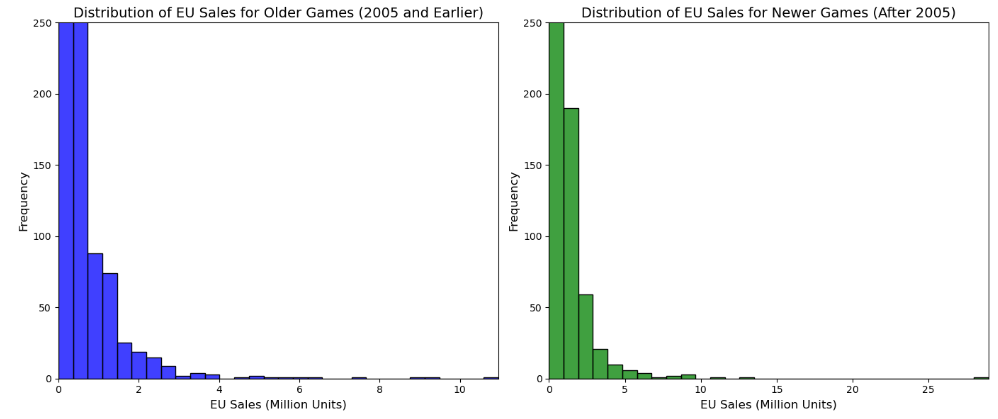
An annotated number visualization was created to present this single statistic, emphasizing clarity and simplicity.

Comparison with Newer Games:

Next, we calculated the mean EU sales for games released after 2005 for comparison.

A bar chart was used to visually compare the mean sales of older games (2005 and earlier) and newer games (after 2005).

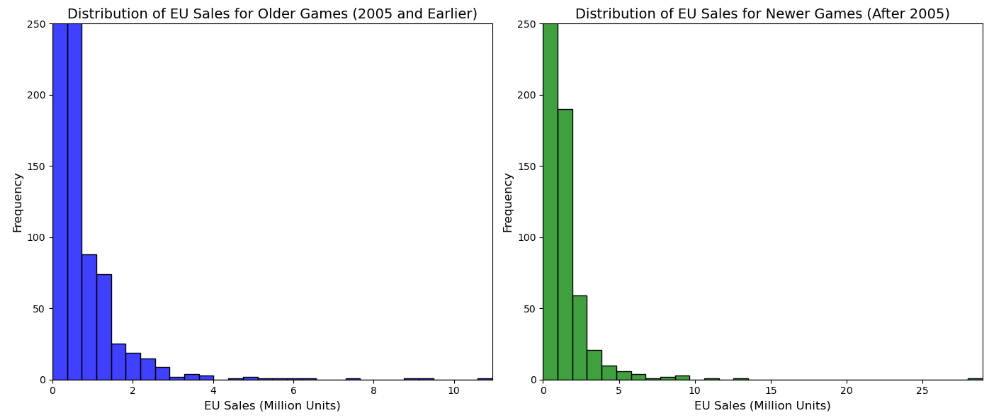
We found that older games had a slightly higher mean EU sales than newer games, though the difference was marginal.



Exploring Sales Distributions with Histograms:

We created histograms to explore the distribution of EU sales for both older and newer games.

These histograms showed that the sales for both time periods were heavily skewed towards lower sales figures.



Applying Log Transformation:

To address the skewness in the sales data, we applied a log transformation to the EU sales of both older and newer games.

New histograms were created for the log-transformed data.

The transformed data revealed more symmetrical distributions, providing a clearer view of the sales patterns.

Analysis Interpretation:

The log transformation indicates that the sales distributions for both older and newer games follow a more normal-like distribution, which is often the case in real-world data.

This more symmetrical distribution can be useful for various statistical analyses and can provide insights for the retail company in terms of sales patterns and tendencies in the European market. It highlights that, despite the skew towards lower sales in the raw data, there is a typical range of sales that most games achieve, which is more evident after transformation.

# THE 3 MOST COMMON DEVELOPERS

Initially, we identified a significant number of missing values in the 'Developer' column(Figure8). This insight was crucial as it impacted the reliability of our analysis(Figure 9).



Figure 8 Missing values in Developer

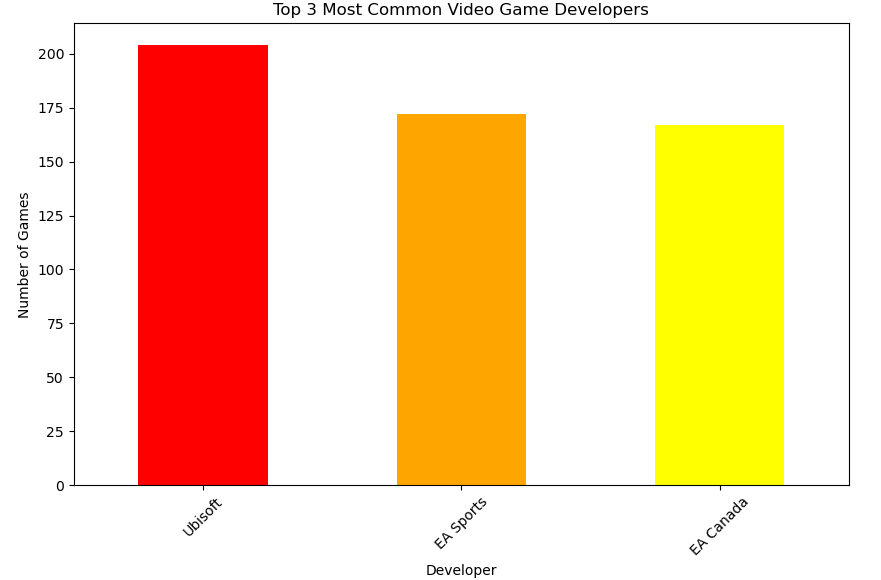


Figure 9 Top 3 most Common Video Games Developers before handling the missed values

To address the missing values, we implemented a method that imputed missing 'Developer' data based on the most common developer associated with each 'Publisher'.

This was achieved through grouping by 'Publisher', finding the mode (most common value) for each group, and using this information to fill missing 'Developer' values.

This step dramatically reduced the number of missing values(figure 10), making our dataset more complete and hence more reliable for analysis.



Figure 10 Missing values in Developer

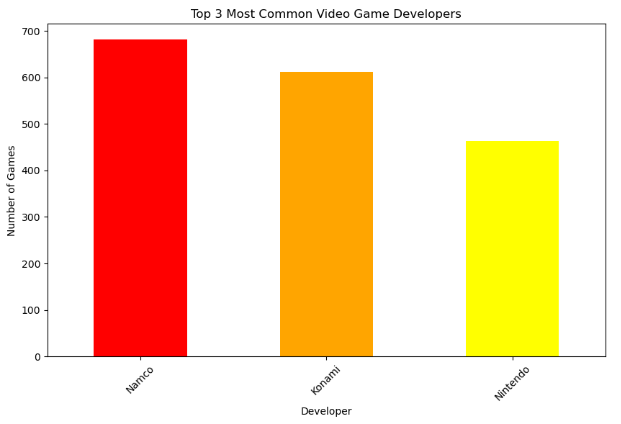


Figure 11 Top 3 most Common Video Games Developers after handling the missed values

With a more complete dataset, we calculated the frequency of each developer, and the result is different to the previous one, instead of Ubisoft, EA Sports and EA Canada, now is Namco, Konami and Nintendo.

To visually present our findings, we created a bar chart, the top three developers were displayed using vertical bars, each colored distinctively (red, orange, and yellow) the color choice creates a warm color gradient across the bars, which could symbolically represent varying levels of activity or output from these developers

# Reference

Atlassian (n.d.). Stacked Bar Charts: A Detailed Breakdown. [online] Atlassian. Available at: <https://www.atlassian.com/data/charts/stacked-bar-chart-complete-guide#:~:text=The%20stacked%20bar%20chart%20(aka>.

https://www.facebook.com/jason.brownlee.39 (2019). How to Calculate Correlation Between Variables in Python. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/>.

Mcleod, S. (2023). Correlation | Simply Psychology. [online] Simplypsychology.org. Available at: https://www.simplypsychology.org/correlation.html.