**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

Index

[Introduction 2](#_Toc163155858)

[TOP 5 GAMES BY GLOBAL SALES 3](#_Toc163155859)

[Data Engineering 3](#_Toc163155860)

[CORRELATION BETWEEN NA SALES AND JP SALES 6](#_Toc163155861)

[Visualization Design 6](#_Toc163155862)

[Data Engineering 6](#_Toc163155863)

[Visualization Design type 2 7](#_Toc163155864)

[DISTRIBUTION OF THE MOST POPULAR 4 GAMES GENRES 9](#_Toc163155865)

[Data Engineering 9](#_Toc163155866)

[Visualization Design: 9](#_Toc163155867)

[Bar Chart (Number of Games) 9](#_Toc163155868)

[Pie Chart (Sales Proportion) 9](#_Toc163155869)

[OLDER VS NEWER GAMES 10](#_Toc163155870)

[Data Engineering 10](#_Toc163155871)

[Visualization Design 12](#_Toc163155872)

[THE 3 MOST COMMON DEVELOPERS 12](#_Toc163155873)

[Data Engineering 12](#_Toc163155874)

[Visualization Design: 14](#_Toc163155875)

[Statistically Relevant Question 14](#_Toc163155876)

[Visualization Design: 15](#_Toc163155877)

[Conclusion 16](#_Toc163155878)

[Reference 16](#_Toc163155879)

# 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 visualizing the 3 most common developers. 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

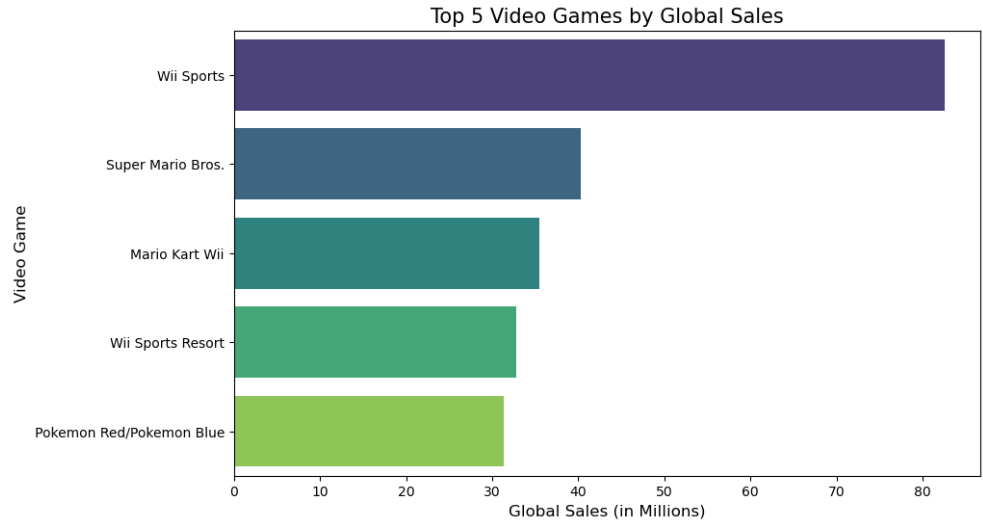


Figure 1 Bar chart of Top 5 Games by global sales

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 was choose from the seaborn library(seaborn, 2013), 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.

Data Engineering:

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

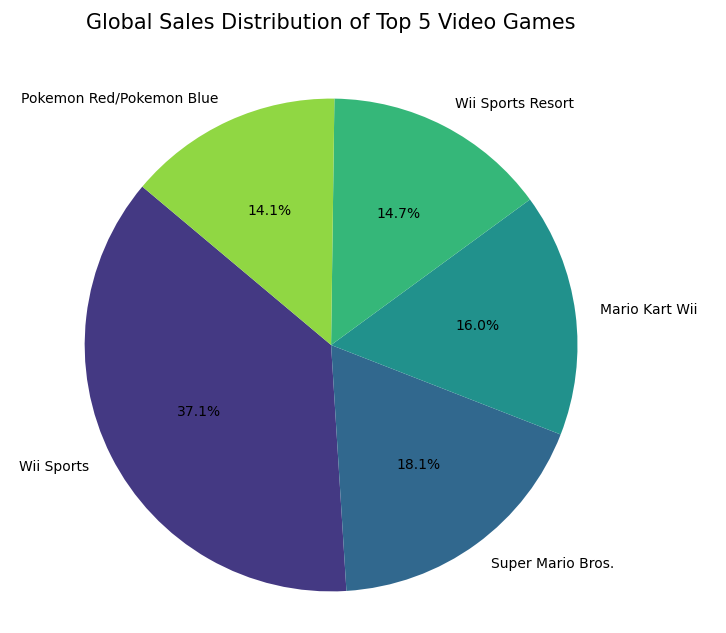


Figure 2 Pie chart Distribution of Top 5 Video Games

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.

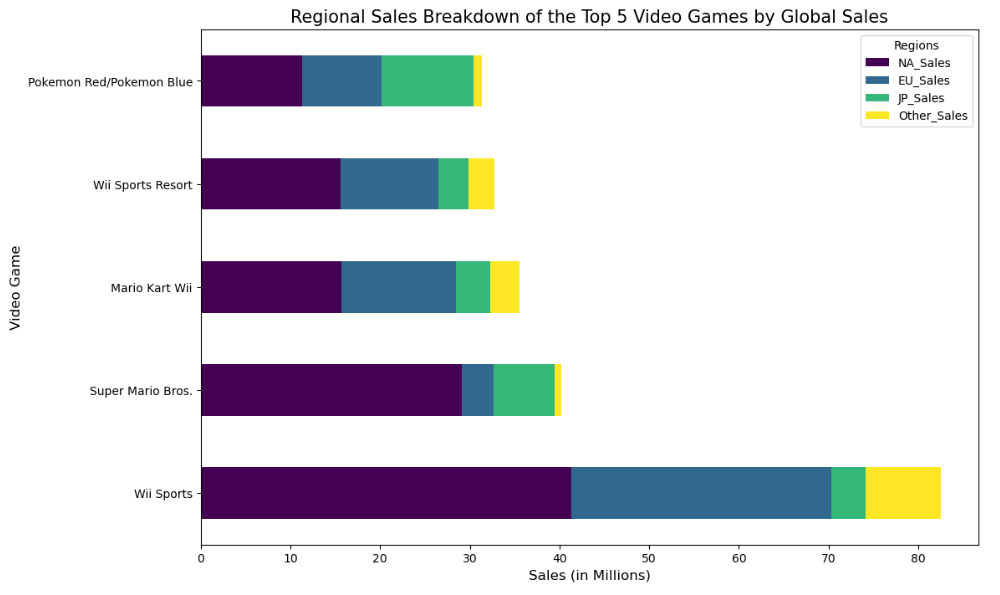


Figure 3 Stacked bar chart of the Top 5 video Games by global Sales

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. The reason that we are visualizing a different way the information can help the 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

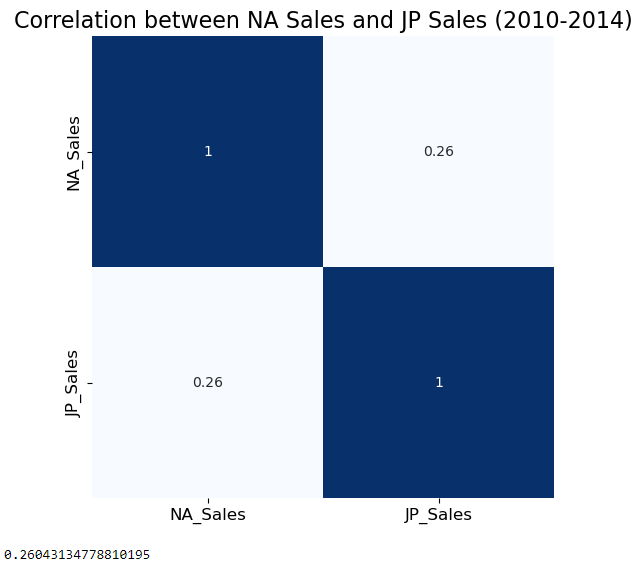


Figure 4 Heatmap correlation between NA Sales and JP Sales (2010-2014)

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,but 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 and negative 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

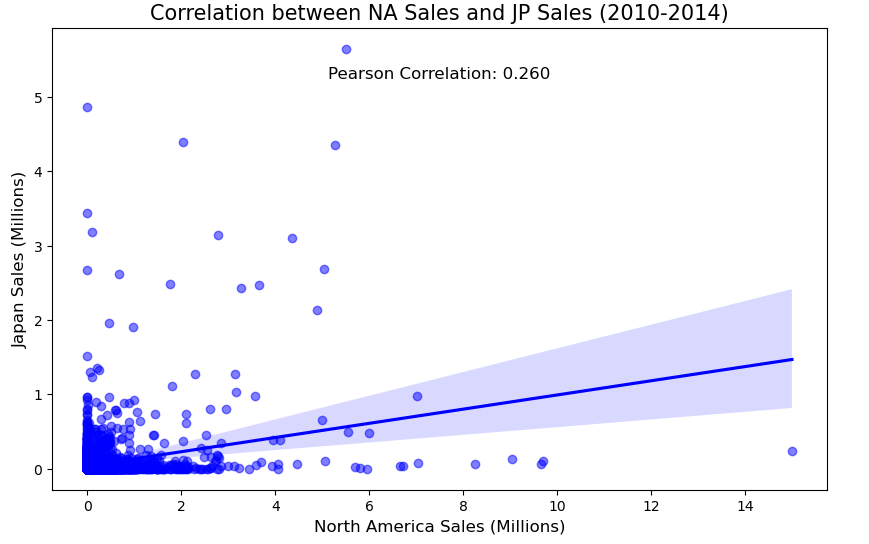


Figure 5 Scatter Plot implementing the Pearson correlation

Visualization Design type 2

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, 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.

The Regression line blue 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.

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.

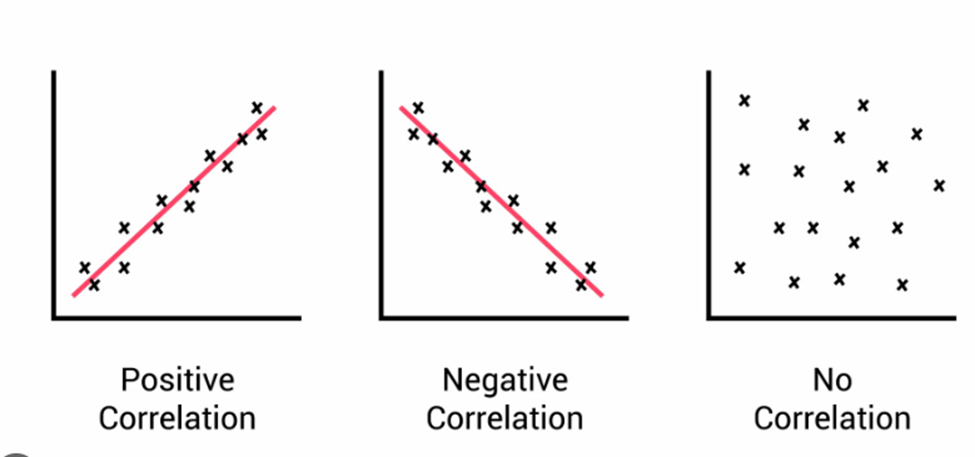


Figure 6 3 types of correlation (Mcleod, 2023)

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

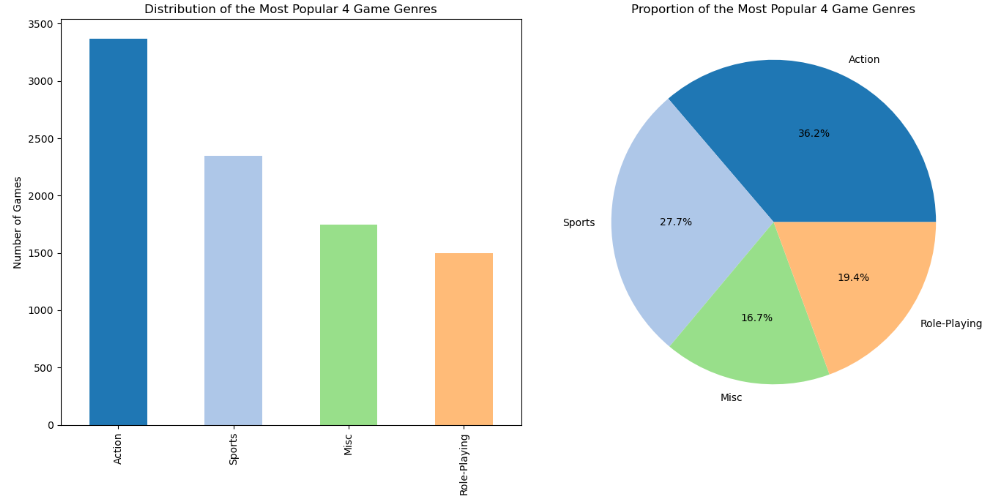


Figure 7 Bar chart and pie chart of the distribution the Most popular 4 games genres

Data Engineering :

The dataset was subjected to a categorical aggregation process, where the volume of video game titles was enumerated based on their respective genres. This was achieved by applying a frequency count across the 'Genre' variable, a method that classifying and quantifying observations in a categorical dataset. This step elucidates the prevalence of each genre within the dataset, offering a foundational understanding of genre distribution.

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

Visualization Design:

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.

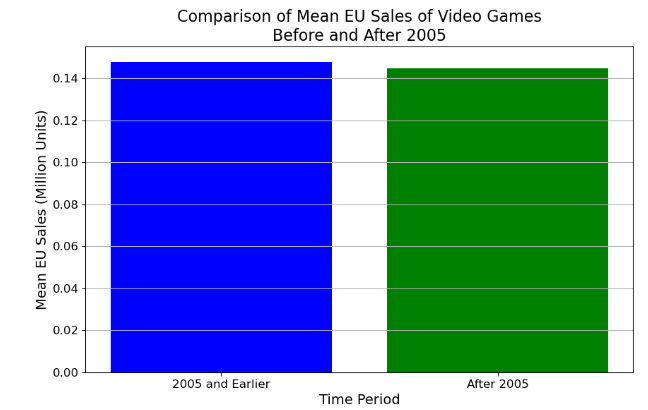
Color: Specific colors are assigned to each genre to maintain visual consistency across both charts, facilitating comparison.

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

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

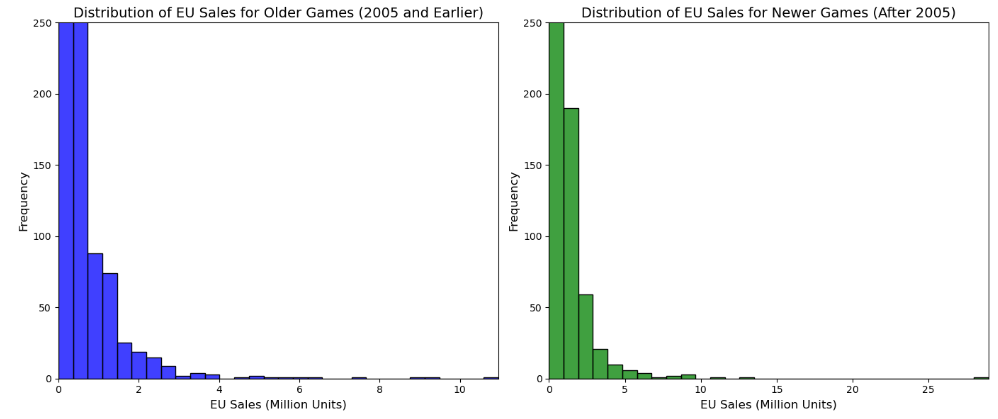


Data Engineering:

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.

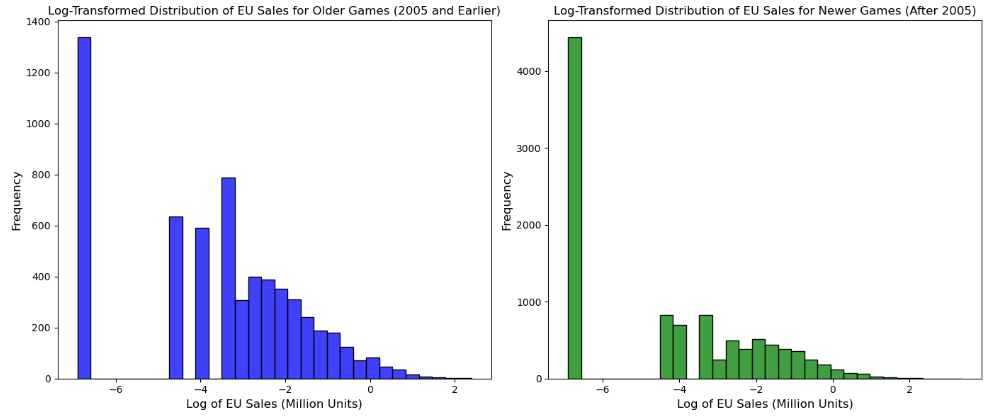
We start calculated the mean EU sales for games released after 2005 for comparison, a bar chart was used to visually compare the mean sales of older games (2005 and earlier) and newer games (after 2005).

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:

The histograms for EU sales of older (2005 and earlier) and newer (after 2005) games are displayed. Both show the distribution of EU sales in millions of units. To provide a clearer picture, especially for sales values that are lower and more common, we are going to applied the log transformation.



Applying Log Transformation:

The transformation provides a clearer view of the sales distribution, especially for smaller values. The log transformation often makes it easier to observe patterns in data with a wide range of values, as is common with sales figures, most video game sales figures are small, but a few games have very high sales. This wide range makes it hard to see patterns, especially among the smaller values.

The log transformation helps by "compressing" the larger values and "stretching" the smaller ones. This makes it easier to see and compare the entire range of 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.

Visualization Design:

Histogram was choose to display the distribution of EU sales. Histograms are great for showing the frequency of data points within certain ranges, two separate plots used to separately highlight the distributions for older and newer games, making it easier to compare the two distributions.

Color decision, blue for older games and green for newer games, maintaining consistency with the bar chart; labels and titles are clearly marked to indicate what each histogram represents, enhancing interpretability.

# THE 3 MOST COMMON DEVELOPERS

Data Engineering

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

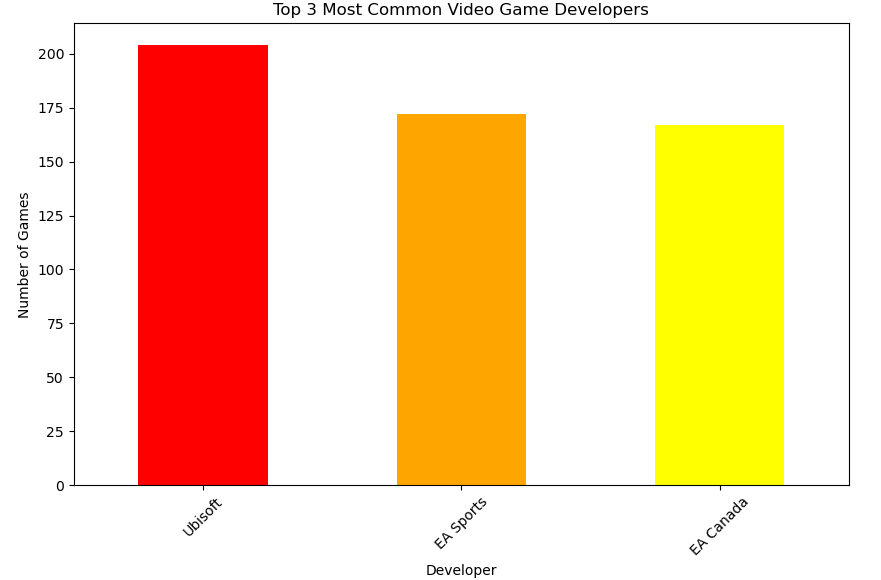


Figure 9 Top 3 most Common Video Games Developers before handling the missed values

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

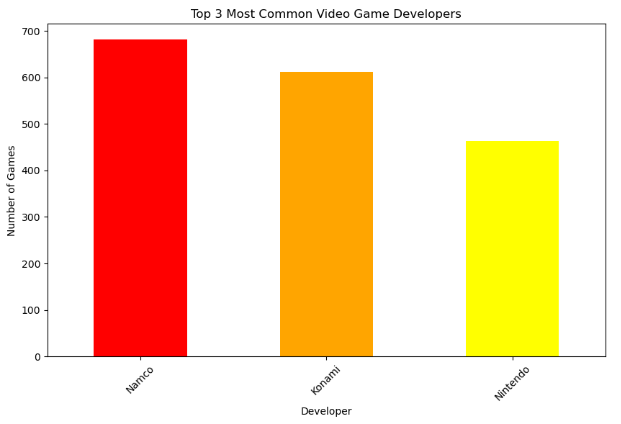


Figure 11 Top 3 most Common Video Games Developers after handling the missed values

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.

## Visualization Design:

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 ,the title as clearly states the purpose of the visualization and the size choose provides enough space for clear visualization while ensuring the bars and labels are easily readable.

# Statistically Relevant Question

"What is the trend in global video game sales over the years, and how have different game genres contributed to these sales?"

This question is vital for understanding the overall market trend and the popularity of genres over time. Such an analysis can provide insights into whether certain genres are gaining or losing popularity, which could significantly influence the company's stocking and marketing strategies for different genres. Additionally, understanding the trend in global sales will help in assessing the overall health and direction of the video game market.

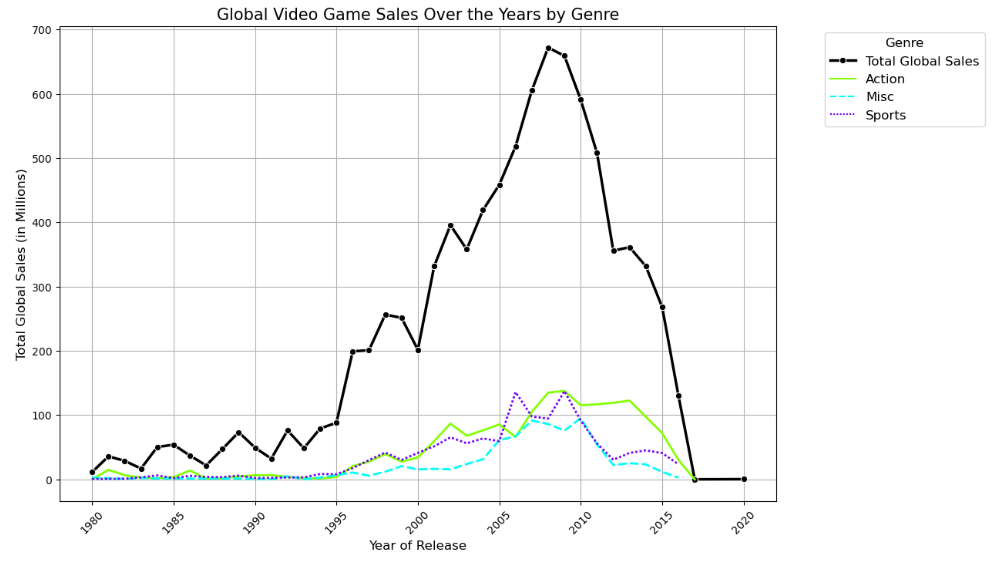


Figure 12 Line Graph of Global Video Games Sales Over the Years by genre

This line graph illustrates the trend in global video game sales over the years, along with the contributions of the top three genres. The rationale for this analysis and visualization is understanding the overall trend in global video game sales and seeing how top genres contribute to these trends informs about changing consumer preferences, which is vital for inventory and promotion planning.

## Visualization Design:

The Line graphs choice is ideal for displaying trends over time. They show the continuity of the data and make it easy to identify trends, peaks, and troughs. This is particularly useful for analyzing yearly sales data, as it helps in observing the progression or decline in sales over the years. Different colors for each genre make it easy to distinguish their respective trends while comparing them against the total global sales (in black).

The choice of a line graph with multiple lines allows for a direct comparison between overall sales trends and genre-specific trends, providing a comprehensive view of the market dynamics.

The use of markers (circles) for the total global sales line and distinct line styles for each genre helps in distinguishing the overall trend from genre-specific trends. This is particularly helpful in a busy graph with multiple lines.

# Conclusion

Our thorough analysis and visualization of the video game dataset effectively combine design principles with data clarity, resulting in visuals that are both informative and user-friendly. These visualizations provide clear insights, supporting strategic decision-making for the company, and highlight key trends in the video game industry. This report serves as a valuable tool for understanding market dynamics and guiding future business strategies.

# Reference

seaborn (2013). Choosing color palettes — seaborn 0.9.0 documentation. [online] Pydata.org. Available at: https://seaborn.pydata.org/tutorial/color\_palettes.html.

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.