Authors : M.Haris Iqbal

Ensuring Data Privacy in Video Game Sales Analysis Using SMPC Approaches

Introduction

In an era where data privacy is crucial, this project explores the application of Secure Multi-Party Computation (SMPC) techniques in analyzing video game sales data. Our objective is to analyze and gain insights without compromising the confidentiality of the data. By implementing techniques like Paillier Encryption, Differential Privacy (DP), and Advanced Encryption Standard (AES), we aim to showcase a privacy-preserving approach to data analysis.

Methodology

Our methodology is centered around three key parts, each addressed using a different SMPC approach:

1. Paillier Encryption for Average Sales Analysis:
    - To determine the best average sales per decade (1980-2020), we employed Paillier Encryption, a homomorphic encryption technique that ensures the confidentiality of individual game sales data.
    - This approach was chosen for its ability to perform encrypted calculations, providing the average sales figures without revealing individual data points.

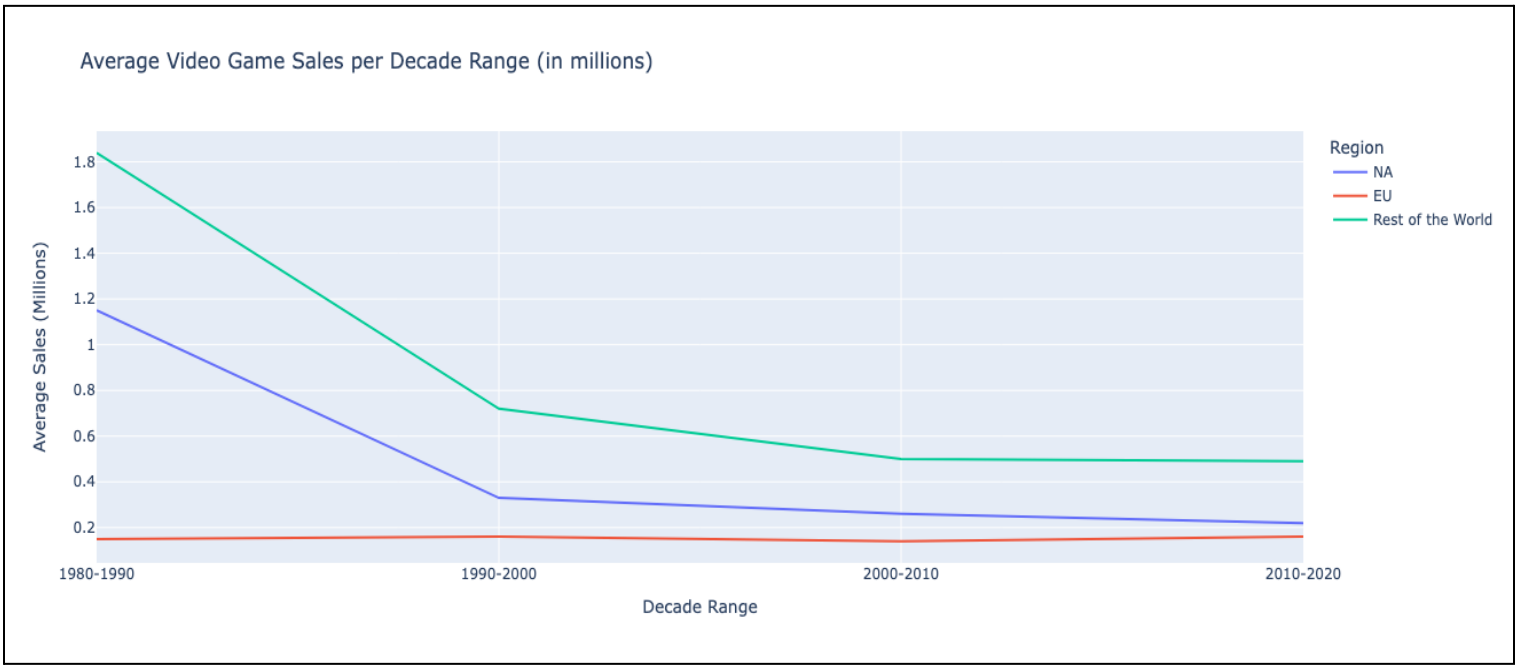
2. Differential Privacy for Ranking Games:
    - For ranking the top 5 favorite games in each region, Differential Privacy was implemented by adding Laplace noise to the sales data.
    - The choice of DP stems from its efficacy in balancing privacy with data utility, ensuring dataset integrity while providing meaningful insights.

3. AES Encryption for Categorical Data Analysis:
    - To identify the most popular platform, publisher, and genre, we utilized AES encryption to secure categorical data.
    - AES was selected for its robust encryption capabilities, ensuring that sensitive categorical information remains confidential throughout the analysis process.

Analysis and Results

Part 1: Paillier Encryption for Average Sales Analysis

The primary objective of this analysis is to understand the trends in video game sales across different decades and regions. By doing so, we can discern which markets have been most lucrative and how consumer preferences or market dynamics may have shifted over time.



The graph depicts a clear trend of declining average sales in North America (NA) and Europe (EU) from the 1980s to the 2010s. In contrast, the Rest of the World (ROW) shows a more stable trend over the same period.

- 1980-1990: This decade shows the highest average sales in NA and ROW, likely indicative of the video game boom and the entrance of major players into the market.
- 1990-2000: There is a noticeable decrease in average sales in NA and EU. This could be due to market saturation or the advent of new forms of entertainment.
- 2000-2010: The downward trend continues in all three regions. This may reflect the increased competition from mobile gaming and other digital platforms.
- 2010-2020: The decline in average sales in NA and EU seems to stabilize, while ROW shows a slight uptick, suggesting emerging markets gaining significance in the gaming industry.

Paillier Encryption was utilized to compute these averages securely, ensuring that the individual sales data of games remained confidential throughout the analysis.

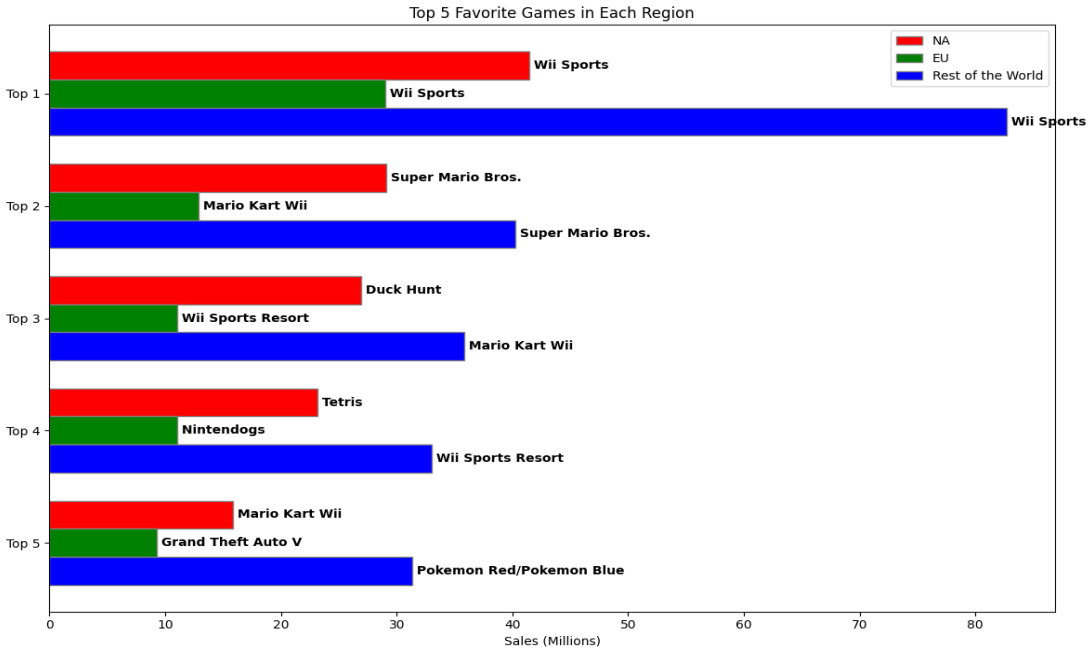
Paillier is additive homomorphic, meaning that it allows for the encrypted sum of the sales data. This property is particularly useful when calculating averages without decrypting the individual data points. By encrypting the sales figures, we ensured that the privacy of the dataset was maintained, complying with any data protection regulations and ethical considerations. Despite the encryption, Paillier allows for efficient computation of encrypted data, which is essential when dealing with large datasets. Although the data is encrypted, the final decrypted result is accurate, as homomorphic encryption does not alter the underlying values during computation.

While there are other cryptographic techniques available, Paillier was chosen over other alternative SMPC or Fully Homomorphic Encryption due to its simplicity and efficiency for the specific task of averaging. Other methods could have introduced unnecessary complexity or computational overhead without offering additional benefits for this particular analysis.

Part 2: Differential Privacy for Ranking Games

In this part, we rank the top 5 favorite games in each region based on their sales figures. To protect the privacy of individual sales data, we apply DP techniques by adding Laplace noise to the sales figures. This ensures the confidentiality of the dataset, adhering to privacy-preserving measures.

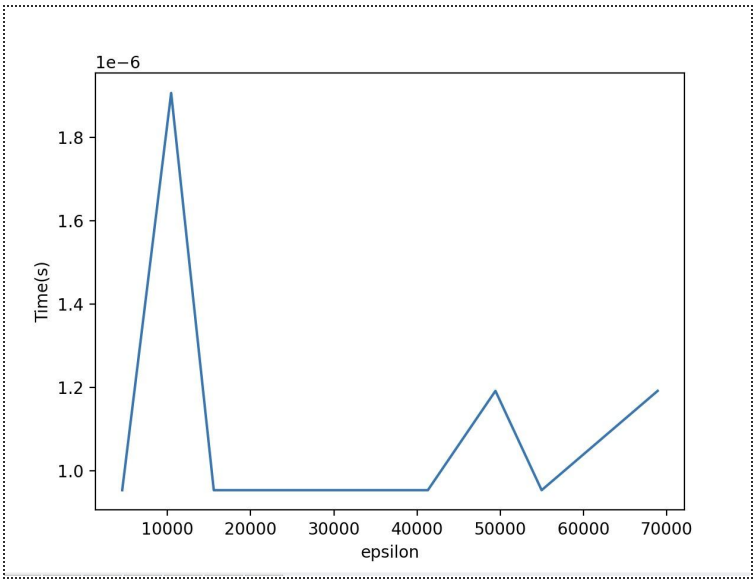
I. First, we plot a bar graph showing the top 5 favorite games in each region without adding any privacy measure to it.



In this graph, Wii Sports, Super Mario Bros, Duck Hun, Tetris, and Mario Kart Wii are the most popular games in the NA region. In EU, Wii Sport is interestingly also ranked first in the top 5, followed by Mario Kart Wii, Wii Sports Resort, Nintendogs, and Grand Theft Auto V. For ROW, Wii Sports is still the most popular game , followed by Super Mario Bros, Mario Kart Wii, Wii Sports Resort, and Pokemon Red/Pokemon Blue.

This conventional method of data analysis allows for direct insights into the sales performance of video games. We leverage this transparent analysis to establish a baseline of game popularity, which we visually represent through the bar graph above. This benchmark serves as a comparative foundation to understand the effects of subsequent privacy measures, such as the introduction of DP, on the accuracy and integrity of our data analysis moving forward.

II. To apply DP on the data, we needed to determine the optimum epsilon value (ε) to add noise. Initially, we set the privacy budget at an epsilon value of 1, prioritizing maximum privacy. To justify this decision, we conducted an analysis to understand the computational implications of applying DP with a wide range of ε values. Specifically, we observed how the computation time varies when ε ranges from 1 to nearly 100,000.



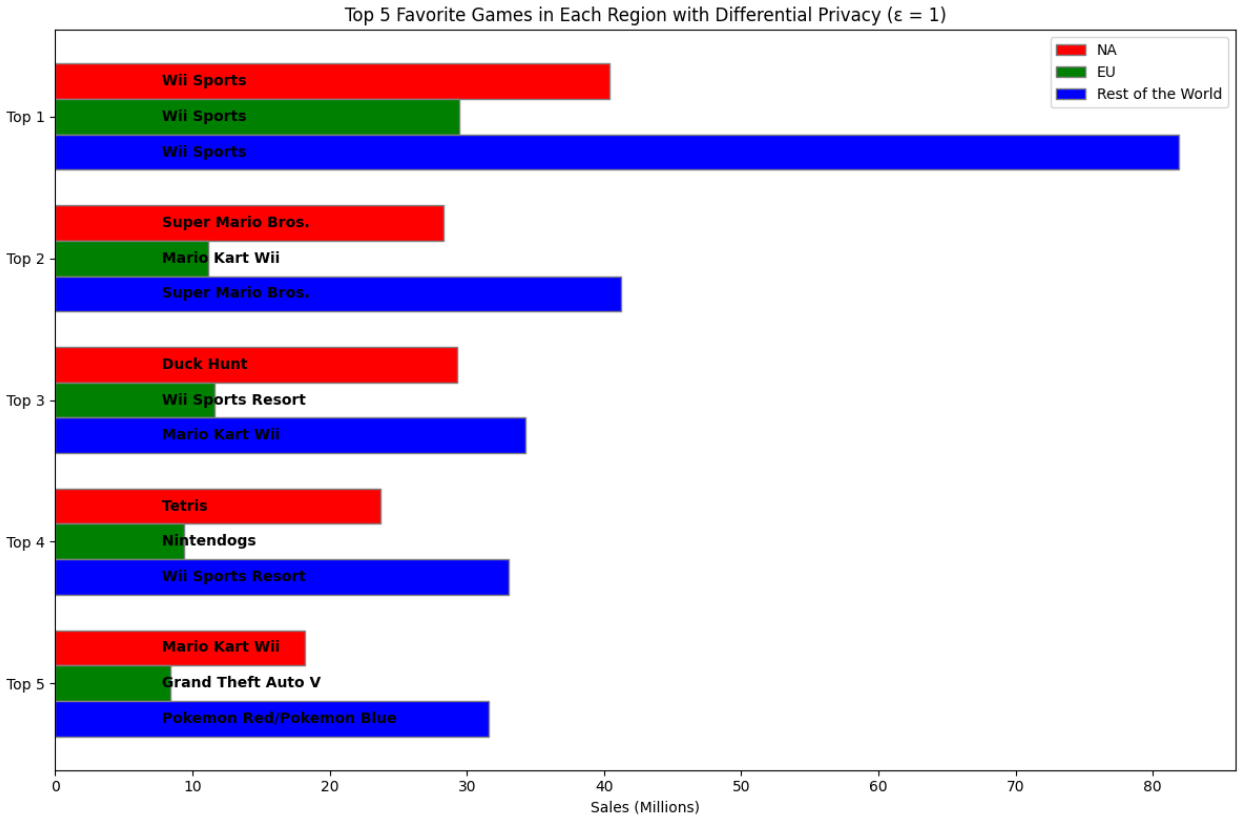
*We only considered NA sales just to show the impact of ε values on time.

The graph above illustrates the relationship between ε values and computation time. It's noticeable that as the value of ε increases, the computation time fluctuates but generally displays a slight decreasing trend, indicating that higher ε values, which correspond to less noise, do not necessarily increase computation time significantly.

This result is expected because a larger ε value in DP implies a weaker guarantee of privacy. This, in turn, leads to faster computation, as the algorithm needs to perform fewer calculations.

This supports our decision to choose an ε of 1 for our DP implementation in the project, as it suggests that a strong level of privacy can be provided without a substantial increase in computation time. Additionally, the chosen ε value ensures a practical trade-off between data utility and privacy, allowing us to perform the analysis efficiently while still maintaining a rigorous standard of privacy.

III. In the visualization below, we have applied DP with an ε of 1, which offers a balance between privacy protection and data utility. While this introduces some uncertainty into the sales figures, potentially affecting the precise ranking order of the top games, it provides a significant level of privacy assurance.



Despite the addition of noise due to DP, the game titles remain the same as in the non-private ranking, indicating that the noise added was not significant enough to alter the order of the top games dramatically.

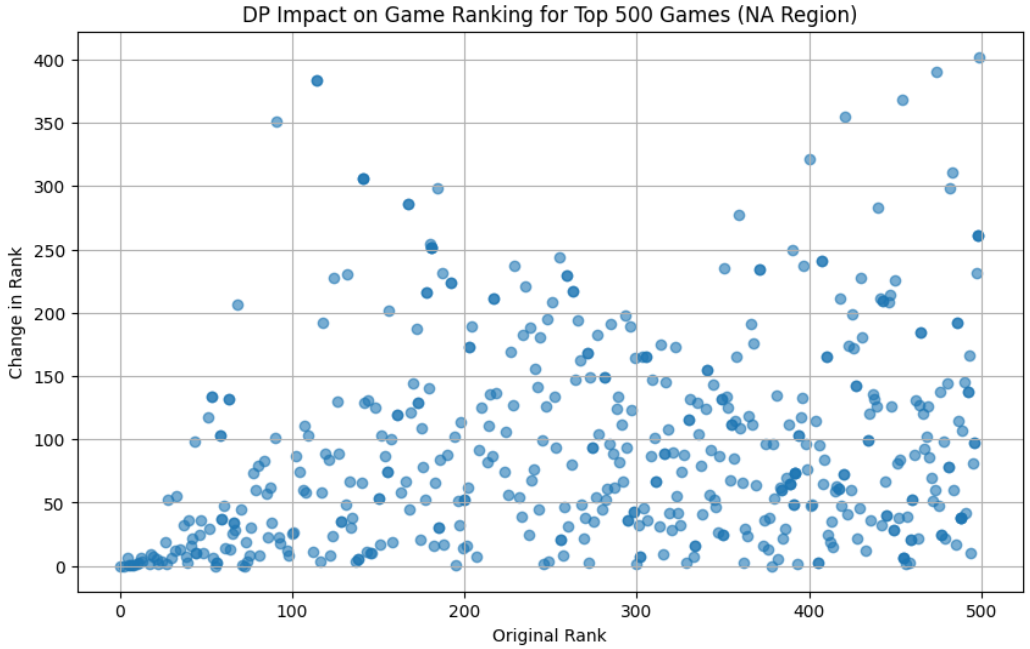
Wii Sports consistently maintains the top position in all regions, indicating a uniform preference robust enough to withstand the noise introduced by DP. While the top rankings seem stable, there are minor variations in the lower rankings, particularly in the ROW. These variations could be due to close sales figures and the impact of noise addition.

This graph also demonstrates that individual game sales privacy is protected by preventing precise determination of any single game's sales figures. The overall integrity of the dataset is preserved, enabling meaningful analysis that offers insights into regional gaming preferences. This shows that it's possible to protect individual data points while still retaining the data's utility for analysis.

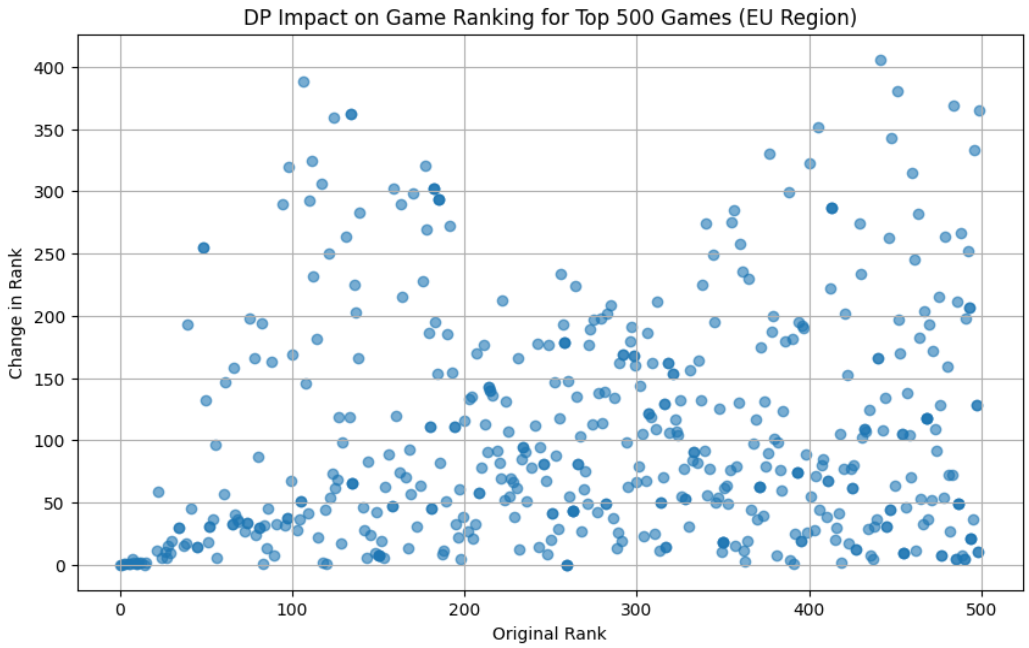
IV. To further extend our analysis beyond the top 5 games, we did an accuracy analysis of DP's impact on game rankings for each region. For the sake of clear visualization and to effectively demonstrate the impact of differential privacy, we have limited our analysis to the top 500 games. This focused approach allows us to more easily observe the changes

in rankings due to noise addition and avoids overwhelming the graph with too many data points, which can make patterns and insights difficult to discern.

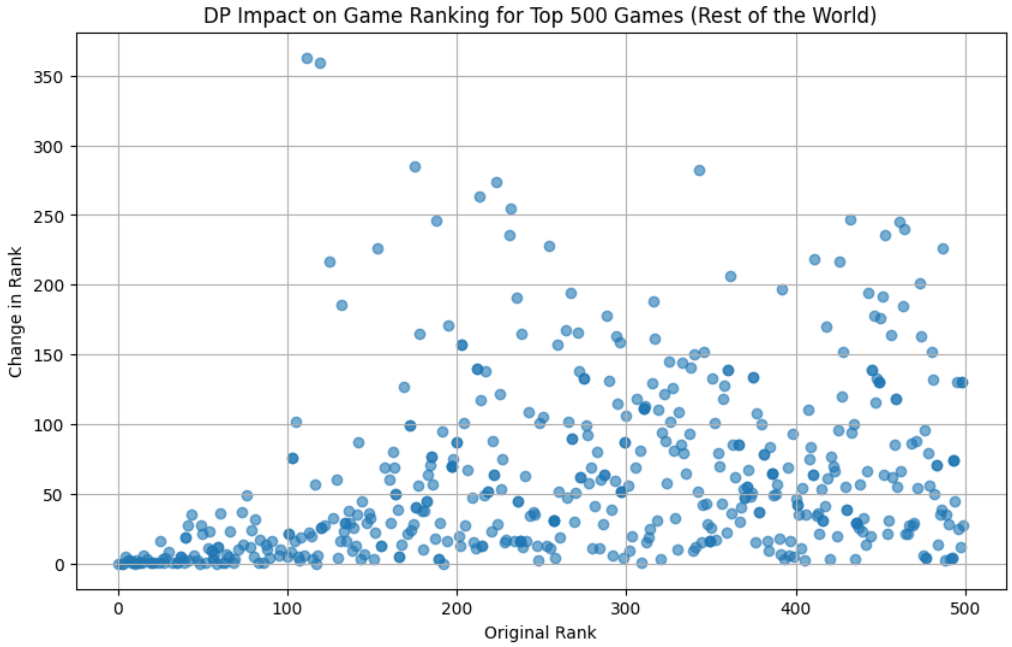
The scatter plots below show the change in ranking for each of these top 500 games. The *x*-axis represents the original ranking based on actual sales, while the *y*-axis indicates how much each game's rank has changed after the application of noise. A greater change in rank suggests a more significant impact of the noise on the game's sales figure, highlighting the trade-off between privacy and data accuracy.



The NA plot shows that most games' rankings do not significantly change, with the majority of points clustered near the zero line of rank change. There are a few outliers with rank changes over 100 positions, suggesting that for certain games, the DP noise significantly altered their sales figures.



Similar to the NA region, the EU plot also has most games with minimal rank changes. The spread of rank changes in the EU region seems slightly more pronounced, with more data points experiencing a rank change greater than 100.



The ROW plot appears to have a denser cluster of games near the lower rank change, implying that the DP noise had less impact on rank changes in this dataset. There are still some significant rank changes but fewer than in the NA and EU regions.

The top games seem to maintain their positions well despite the noise, which may indicate that their lead in sales is substantial enough to withstand the perturbations introduced by DP. Some games are more sensitive to the introduction of noise, which can significantly change their rankings. This could affect market analyses and decision-making based on these rankings.The difference in DP's impact across regions suggests that market characteristics can influence how DP affects data. This insight is crucial for businesses that may rely on such data for strategic decisions.

Part 3: AES Encryption for Categorical Data Analysis

In this part, we uncovered some interesting insights while ensuring the confidentiality of key categorical information. To achieve this, we employed Advanced Encryption Standard (AES) encryption to protect the details of game platforms, genres, and publishers.

Unlike numerical data, string values like platform names, genres, and publishers cannot be directly handled through methods like Paillier or DP, which are better suited for numerical computations. AES allows us to encrypt these strings into indecipherable ciphertexts, ensuring that the actual names remain anonymous throughout the analysis.

AES ensures that a given string value will always produce the same ciphertext when encrypted with the same key, allowing for reliable aggregation and comparison of encrypted data. We only decrypt the final result to reveal the most popular categories.

Process:

- Each platform name, publisher, and genre in the dataset was encrypted using AES. This step transformed readable strings into secure ciphertexts, ensuring the confidentiality of the data.
- After identifying the ciphertexts associated with the highest sales, the final step was to decrypt these ciphertexts to reveal the most popular platform, publisher, and genre.
- The results revealed that "PS2" was the most popular platform, "Nintendo" the most popular publisher, and "Action" the most favored genre. These insights are crucial for understanding market trends, consumer preferences, and guiding strategic decisions in the gaming industry.

Encrypting category names provided privacy protection for the categorical data. It is a unique approach when the names themselves are sensitive or when revealing the most popular categories could lead to privacy concerns. We ensured that decryption is performed only for the final step to reveal the most popular categories, maintaining the confidentiality of the dataset throughout the analysis process.

It's crucial to note that although the category names were encrypted, the actual sales data was not. The aggregation occurred in plaintext, which is suitable for this scenario since the sales data doesn't require anonymization; the priority is protecting the identity of the categories.

This approach exemplifies how cryptographic techniques can be leveraged to ensure data privacy in various analytical contexts. It demonstrates a balance between the protection of sensitive information and the extraction of valuable insights, which is a growing necessity in the era of big data and heightened privacy concerns.