**VIDEO GAME SALES**

**A PROJECT REPORT**

**Submitted by**

***ANIRUDH 1RVU23CSE058***

***ATTILI RAM KISHAN KUMAR 1RVU23CSE089***

***AMOGHA V PRASAD 1RVU23CSE046***

***ACHYUTH ANIL KUMAR 1RVU23CSE019***

*in partial fulfillment for the award of the degree of*

**B.Tech (Hons) – Computer Science & Engineering**

*in* **Data Analysis with Python**

**to Dr. Sahana Prasad**



**School of Computer Science and Engineering**

**RV University**

**RV Vidyaniketan, 8th Mile, Mysuru Road, Bengaluru, Karnataka, India – 562112.**

**December 2023**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **TITLE** |  |
| **1.0** | **INTRODUCTION** | **3** |
|  |  |  |
| **2.0** | **METHODOLOGY** | **4** |
|  |  |  |
| **3.0**  **4.0**  **5.0** | **INFERENCE**  **QUESTIONS AND CODE SNIPPETS**  **DATA STORY TELLING THROUGH IIIP** | **5**  **6-14**  **15** |
|  |  |  |
| **6.0** | **RESULT AND DISCUSSION** | **16** |
|  |  |  |
| **7.0**  **8.0** | **CONCLUSION**  **REFERENCES** | **17**  **18** |

1. **INTRODUCTION**

* In the dynamic realm of the gaming industry, understanding the factors influencing game sales is crucial.
* This project makes use of Python for a comprehensive analysis of a gaming dataset, focusing on genres, publishers, and regional sales.
* Through data visualization and statistical analysis, we aim to uncover trends, preferences, and market dynamics. By tapping into this rich dataset, our goal is to provide valuable insights that empower curious customers and developers alike to make informed decisions, optimize strategies, and contribute to the evolving landscape of interactive entertainment.

|  |  |  |
| --- | --- | --- |
|  |  | **.** |
|  |  |  |
|  |  |  |
|  |  |  |

**2. METHODOLOGY**

.

|  |  |
| --- | --- |
| **Structured Approach** | This project employs a structured methodology, encompassing key stages of data collection, cleaning, and visualization with Python as the primary programming language. |
| **Toolset** | Leveraging Python's versatile ecosystem, we rely on essential data analysis libraries—Pandas, NumPy, Matplotlib, and Seaborn—for seamless data manipulation, visualization, and statistical analysis tailored to our gaming dataset. |
| **Dataset Composition** | Our dataset is carefully selected to cover pivotal aspects of gaming dynamics, including game names, genres, publishers, and regional sales figures, ensuring a comprehensive analysis. |
| **Better Relevance** | By encompassing diverse factors influencing game sales, such as genres, publishers, and regional performance, our dataset's inclusivity enhances the relevance and depth of our analysis. |
| **Data Cleaning** | Ensuring the integrity of our analysis, the data cleaning phase addresses missing values, outliers, and standardizes units, establishing a robust foundation free from potential biases or inaccuracies. |
| **Descriptive Statistics and Visualizations** | To reveal patterns and correlations, we employ descriptive statistics and visualizations tailored to the gaming domain, providing insights into genre preferences, publisher influence, and regional trends. |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**3. INFERENCE**

* Key Factors Analysis: Identify pivotal variables influencing game sales, including genres, publishers, and regional factors. Explore the impact of these factors on overall game performance and market dynamics.
* Temporal and Spatial Trends: Examine temporal trends to discern how game sales fluctuate over time, considering seasonal patterns and evolving consumer preferences. Investigate spatial trends to pinpoint regions where certain genres or platforms thrive or face challenges.
* Correlation Analysis: Uncover correlations between different gaming variables to reveal interdependencies. For instance, understanding how genre popularity correlates with regional sales can provide insights into consumer preferences and market trends.
* Data-Driven Decision - Making: Empower game developers, publishers, and industry stakeholders with actionable insights. Recommendations may include strategic interventions, marketing approaches, or platform-specific optimizations to enhance overall gaming productivity.
* This data-driven approach ensures that the gaming industry's decision-makers are equipped with valuable insights, facilitating informed strategies and policies to adapt to the evolving landscape of interactive entertainment.

**In the course of our project, we proactively pursued a deeper understanding of diverse chart types, including treemaps and sunburst charts, driven by our intrinsic curiosity. Furthermore, we enriched our project by incorporating a meticulously crafted data story presentation, a skill honed through structured learning in our classes.**

**4. Questions and Code Snippets**

**1)Importing necessary libraries and the dataset for our project**

**A screenshot of a computer

Description automatically generated**

**2)Data Preprocessing**

**A screenshot of a computer

Description automatically generated**

Here we find the columns which have missing values.

**df.dropna(inplace=True):** This line drops rows with any missing values from the DataFrame df in place.

**The inplace=**True parameter modifies the original DataFrame rather than creating a new one.

**df.isnull().sum():** After dropping the rows with missing values, this line checks for missing values again using df.isnull().sum(). This will show the count of missing values in each column after the rows with missing values have been removed.

**3)Statistical Analysis.**

**A screenshot of a computer code

Description automatically generated**

* **df.groupby('Genre')['Global\_Sales'].sum():** This part of the code groups the DataFrame (df) by the 'Genre' column. It then calculates the sum of the 'Global\_Sales' column for each genre. This creates a new Series where the index is the unique genres, and the values are the total global sales for each genre.
* **.idxmax():** This method is called on the Series obtained from the previous step. It returns the index (i.e., the genre) where the maximum value occurs. In simple terms, it finds the genre with the highest total global sales.
* **max\_sales\_value = df.groupby('Genre')['Global\_Sales'].sum().max():** This line calculates the maximum total global sales across all genres.
* **Finally, print(max\_sales\_genre, max\_sales\_value):** This prints out the result, showing the genre with the highest global sales and the corresponding total global sales value.

A screenshot of a computer

Description automatically generated

* **`df['Global\_Sales'].idxmax()`:** This part of the code finds the index of the row where the 'Global\_Sales' column has the maximum value. In other words, it identifies the row (or game) with the highest global sales.
* **`df.loc[df['Global\_Sales'].idxmax()]`:** This uses the index obtained in the previous step to locate and retrieve the entire row (game) with the highest global sales.
* **['Year']:** This part of the code extracts the value in the 'Year' column from the row identified in step 2. It represents the year in which the game with the highest global sales was released.
* **rich\_year = df.loc[df['Global\_Sales'].idxmax()]['Year']`:** This line assigns the extracted 'Year' value to the variable `rich\_year`.
* **Similarly, `rich\_year\_value = df.loc[df['Global\_Sales'].idxmax()]['Global\_Sales']`** extracts the global sales value for the game with the highest global sales and assigns it to the variable `rich\_year\_value`.
* **Finally, `print(rich\_year, rich\_year\_value)`:** This line prints out the result, showing the year and the corresponding global sales value for the game with the highest global sales.

A screenshot of a computer code

Description automatically generated

* **`df['Platform'].value\_counts()`:** This part of the code counts the number of occurrences of each unique value in the 'Platform' column. It returns a Series where the index is the unique platforms, and the values are the counts of games for each platform.
* **`.idxmax()`:** This method is called on the Series obtained in step 1. It returns the index (i.e., the platform) where the maximum count occurs. In simple terms, it finds the platform with the highest number of games.
* **`most\_plat = df['Platform'].value\_counts().idxmax()`:** This line assigns the platform with the highest number of games to the variable `most\_plat`.
* **`df[df['Platform'] == most\_plat]`:** This part of the code filters the DataFrame to include only rows where the 'Platform' column matches the platform identified in step 3.
* **`.shape[0]`:** This retrieves the number of rows in the filtered DataFrame, which corresponds to the number of games released on the platform with the highest number of games.
* **`num\_games = df[df['Platform'] == most\_plat].shape[0]`:** This line assigns the number of games for the platform with the highest count to the variable `num\_games`.

**4)Data visualization**

**A graph with colorful bars

Description automatically generated with medium confidence**

* **`plt.figure(figsize=(10, 6))`:** This line creates a new figure with a specified size of 10 inches in width and 6 inches in height.
* **`sns.barplot(x='Genre', y='Global\_Sales', data=df, estimator=sum, ci=None, palette='viridis')`:** This line uses seaborn's `barplot` function to create a bar plot. The 'Genre' column is plotted on the x-axis, the 'Global\_Sales' column (summarized by the sum function) is plotted on the y-axis, and the data is taken from the DataFrame `df`. The `estimator=sum` parameter specifies that the bar heights should represent the sum of global sales for each genre. The `ci=None` parameter disables error bars, and `palette='viridis'` sets the color palette.
* **`plt.title('Genre Distribution of Global Sales')`:** This sets the title of the plot to 'Genre Distribution of Global Sales'.
* **`plt.xlabel('Game Genre')`:** This sets the label for the x-axis to 'Game Genre'.
* **`plt.ylabel('Total Global Sales (in millions)')`:** This sets the label for the y-axis to 'Total Global Sales (in millions)'.
* **`plt.xticks(rotation=45, ha='right')`:** This rotates the x-axis labels by 45 degrees and aligns them to the right for better readability.
* **`plt.show()`:** This displays the plot.

**Conclusion:** The bar graph illustrates that the preeminent genre in terms of sales is action video games, closely followed by sports, while the genre with the least sales is strategy. We have used a bar graph here since it is employed for statistical analysis as it visually and effectively compares sales across discrete genres, providing a clear representation of relative quantities and facilitating quick comprehension for a diverse audience.

A screenshot of a computer

Description automatically generated

* **`plt.figure(figsize=(10, 8))`:** This line creates a new figure with a specified size of 10 inches in width and 8 inches in height.
* **`squarify.plot(sizes=df.groupby('Genre')['Global\_Sales'].sum(), label=df['Genre'].unique(), alpha=0.7, color=sns.color\_palette('pastel'))`:** This line generates the squarify treemap. The `sizes` parameter specifies the sizes of the squares, which are calculated as the sum of global sales for each genre. The `label` parameter sets the labels for each square based on unique genres. The `alpha` parameter controls the transparency of the squares, and `color` sets the color palette for the squares.
* **`plt.title('Treemap of Global Sales by Genre')`:** This sets the title of the treemap.
* **`plt.axis('off')`:** This turns off the axis labels and ticks for a cleaner treemap visualization.
* **`plt.show()`:** This displays the treemap.

**Conclusion:** Global sales data is used to customize a treemap for a demographic audience, emphasizing sports as the top-selling genre worldwide across all platforms, aided by a pastel palette for efficient category differentiation. Treemaps excel in visually presenting hierarchical data structures and depicting proportions and relationships.

A graph with lines and numbers

Description automatically generated

* **`global\_sales\_by\_year = df.groupby('Year')['Global\_Sales'].sum()`:** This line groups the DataFrame `df` by the 'Year' column and calculates the sum of global sales for each year, creating a Series called `global\_sales\_by\_year`.
* **`plt.figure(figsize=(12, 6))`:** This line creates a new figure with a specified size of 12 inches in width and 6 inches in height.
* **`global\_sales\_by\_year.plot(marker='o', color=sns.color\_palette('pastel'**))`: This line uses the `plot` method to create a line plot. The `marker='o'` parameter adds circular markers to the line, and `color=sns.color\_palette('pastel')` sets the color palette for the line.
* **`plt.title('Global Sales Over Years')`:** This sets the title of the plot to 'Global Sales Over Years'.
* **`plt.xlabel('Year')`:** This sets the label for the x-axis to 'Year'.
* **`plt.ylabel('Total Global Sales (in millions)')`:** This sets the label for the y-axis to 'Total Global Sales (in millions)'.
* **`plt.show()`:** This displays the line plot.

**Conclusion:** Utilizing a line chart to analyze video game sales reveals that 2006 emerges as the pinnacle year based on sourced data. This chart effectively communicates annual sales figures, offering a discernible representation of the evolving trends. The choice of a line chart is justified by its aptitude in illustrating the temporal variation in sales across different years, providing a concise and clear depiction that facilitates a straightforward understanding of the data's temporal progression.

A screenshot of a computer

Description automatically generated

* **`top\_publishers = df['Publisher'].value\_counts().head(10)`:** This line calculates the count of games for each publisher and selects the top 10 publishers with the most games.
* **`total\_games = df['Publisher'].count()`:** This line calculates the total number of games in the DataFrame.
* **`market\_share = (top\_publishers / total\_games) \* 100`:** This line calculates the market share for each of the top publishers as a percentage.
* **`market\_share = market\_share.reset\_index()`:** This resets the index of the `market\_share` DataFrame to make it compatible with Plotly.
* **`market\_share.columns = ['Publisher', 'Market Share']`:** This renames the columns of the `market\_share` DataFrame to 'Publisher' and 'Market Share' for better clarity.
* **`fig = px.sunburst(market\_share, path=['Publisher'], values='Market Share', title='Market Share of Top Publishers (Sunburst Chart)', color='Market Share', color\_continuous\_scale='viridis')`:** This line creates a sunburst chart using Plotly Express. The 'Publisher' column is used for the hierarchical path, 'Market Share' is used for the values, and the chart is colored based on the market share. The color scale is set to 'viridis'.
* **`fig.show()`:** This displays the sunburst chart.

**Conclusion:** The sunburst chart provides a lucid depiction of market share differentials among prominent publishers, notably positioning Electronic Arts as the primary entity and designating Take-Two Interactive with the least share. The adjacency of an index enhances interpretability, offering a convenient means to discern trends in market share across publishers within the hierarchical structure, underscoring the efficacy of the sunburst chart in presenting complex nested data.

**DATA STORY TELLING THROUGH IIIP**

**Introduction:**

Embarking on a journey through the realm of video game sales, our analysis employs diverse visualizations to uncover insights and present a comprehensive narrative of the industry's evolution.

**Insights:**

Bar Graph (Genre Sales):

Our visual exploration reveals a robust sales narrative: action and sports genres dominate, while the bar graph serves as a visual compass, offering clarity in understanding genre-based sales disparities.

**Treemap (Global Sales):**

Global sales data, channeled through a treemap, elucidates sports as the top-selling genre universally. A pastel palette enhances category differentiation, aligning with treemaps' prowess in presenting hierarchical data structures and illustrating proportions and relationships.

**Line Chart (Temporal Analysis):**

The line chart singles out 2006 as the pinnacle year in video game sales, effectively communicating annual figures and showcasing evolving trends. Its aptitude in illustrating temporal variations simplifies the understanding of the industry's temporal progression.

**Sunburst Chart (Market Share):**

The sunburst chart provides a lucid depiction of market share among publishers, notably positioning Electronic Arts as primary and designating Take-Two Interactive with the least share. This underscores the efficacy of the sunburst chart in presenting complex nested data.

**Interpretation:**

This amalgamation of insights signifies a symbiotic relationship between varied visualizations—bar graphs illuminate genre hierarchies, treemaps unveil global market dynamics, line charts pinpoint temporal peaks, and sunburst charts dissect market share. Together, these visual tools offer a multifaceted understanding, showcasing the interplay of genres, global preferences, historical milestones, and market dynamics that have shaped the industry.

**Presentation:**

In presenting this cohesive narrative, our analysis demonstrates the potency of varied visualizations—bar graphs, treemaps, line charts, and sunburst charts—as indispensable tools in unraveling the intricate story of video game sales. By seamlessly integrating these visual elements, we illuminate genre dynamics, global preferences, historical landmarks, and market share differentials, providing a comprehensive and accessible portrayal of the video game industry's evolution.

|  |  |
| --- | --- |
|  |  |

**5. RESULT AND DISCUSSION**

1. **Influential Variables in Game Sales:**
   1. Significant variables impacting game sales were identified, with genres, publishers, and regional factors playing crucial roles in determining overall performance.
2. **Temporal Trends in Gaming Dynamics:**
   1. Temporal trends in game sales were analyzed, revealing how consumer preferences and sales patterns evolve over time. Seasonal variations and long-term changes were examined to understand the dynamic nature of the gaming market.
3. **Spatial Dynamics in Gaming Performance:**
   1. Spatial trends were explored to identify regions where specific genres or platforms thrive and regions facing challenges. This spatial analysis offers valuable insights for targeted marketing and regional optimizations.
4. **Correlation Analysis for Strategic Insights:**
   1. Correlation analysis uncovered relationships between various gaming variables. Understanding how genre popularity correlates with regional sales, for example, provides actionable intelligence for strategic decision-making.

**6. CONCLUSION**

In conclusion, the comprehensive analysis of video game sales across diverse platforms provides valuable insights into the dynamic landscape of the gaming industry. The data reveals notable trends, shifts in popularity, and the impact of technological advancements on consumer preferences. The steady rise of digital distribution, coupled with the emergence of new gaming platforms, has significantly influenced sales patterns.

The project highlights the dominance of certain gaming consoles and platforms during specific periods, emphasizing the cyclical nature of the industry. Additionally, the influence of blockbuster titles, gaming genres, and global economic factors on sales trends underscores the complexity of the video game market.

As the gaming industry continues to evolve, stakeholders can leverage these findings to make informed decisions, anticipate market shifts, and adapt strategies to meet the evolving demands of gamers. The analysis not only provides a retrospective view of the past two decades but also serves as a foundation for understanding potential future developments in the ever-expanding world of video game sales.

**REFERENCES**

* **Dataset:** https://www.kaggle.com/datasets/gregorut/videogamesales