Exploring Global Video Game Sales

About the Project

In this project, you will explore and analyze a dataset of global video game sales to uncover trends and insights in the gaming industry.

The dataset contains information on video games released across various platforms, genres, and publishers, along with their sales in different global regions.

Your task is to perform a complete data wrangling and exploratory data analysis (EDA) workflow.

Part I: Data Cleaning and Preparation

1. Load the dataset and inspect its structure.
2. Check for missing or inconsistent values.
3. Handle missing data appropriately; decide whether to impute, drop, or leave as is.
4. Ensure columns have appropriate data types.
5. Check for duplicates and remove them if necessary.

Part II: EDA

Descriptive statistics

1. Compute key descriptive statistics for numerical variables.
2. Identify the top 10 best-selling games globally.

Trends over time

1. Analyze the number of games released per year.
2. Plot global sales over time — are sales increasing or decreasing?

Genre analysis

1. Which genres are most common?
2. Which genres generate the most global sales?

Platform analysis

1. What are the most popular platforms by number of games and total sales?
2. How has platform popularity changed over time?

Regional analysis

1. Compare preferences between North America, Europe, and Japan.
2. Are there genres that dominate in one region but not others?

Publisher insights

1. Identify the top 10 publishers by total global sales.
2. Is there a "hit publisher" that dominates certain genres?

Market Dominance Analysis

Create a "Market Power Index" for each publisher that considers multiple factors:

1. Calculate each publisher's market share within each genre (their sales / total genre sales)
2. Identify if a publisher is a "genre leader" (has >30% market share in that genre).
3. For each publisher, compute:
    a. Genre Diversity Score: How many different genres they publish in
    b. Regional Consistency Score: Standard deviation of their sales proportions across regions (lower = more consistent)
    c. Hit Rate: Percentage of their games that sold above the median of their genre.
4. Create a final Market Power Index using a custom formula combining these metrics.
5. Rank Publishers and identify different publisher archetypes (e.g., "Genre Specialists" vs "Diversified Giants")

Regional Preference Divergence

Identify games and genres that have dramatically different appeal across regions:

1. For each game, calculate a Regional Divergence Score:
    a. Normalize each region's sales as a proportion of that game's total sales.
    b. Compare each region's proportion to the global average proportion for that region.
    c. Use a custom function to calculate divergence (e.g., coefficient of variation or entropy-based measure)
2. Create genre-region affinity scores:
    a. For each genre, calculate what percentage of its total sales come from each region.

    b. Compare this to the overall regional distribution across all genres.
    c. Identify “over-indexed” and “under-indexed” genre-region combinations.
3. Identify temporal shifts in regional preferences:
    a. Group by year and region
    b. Calculate each region’s favorite genre per year
    c. Track how regional preferences have converged or diverged over time.
    d. Apply a custom function to calculi “preference stability index” for each region.
4. Find cultural outliers:
    a. Games that sold exceptionally well in one region (>50% of sales) but poorly in others (<10% each)
    b. Calculate the “cultural specificity” score for each game.

Part III: Visualization and Storytelling

Use matplotlib or seaborn to visualize your findings.

The plots should:

- Have clear titles, axis labels and legends
- Use appropriate plot types
- Highlight key insights discovered during the EDA.

Example visuals:

- Global sales trend over years
- Bar chart of top-selling genres
- Heatmap comparing regional sales by genre
- Publisher vs total global sales
- Platform evolution over years

Deliverables

A jupyter notebook containing:

- Clean, well-commented code,
- Explanatory text cells for each section
- Visualizations with insights.

A short conclusion summarizing:

- The key insights you discovered
- What data wrangling challenges you faced

- How visualization helped you better understand the dataset