DS 203 - Big Data Essentials

Final Project Report

Title: Analyzing Global Gaming Profiles and Player Behavior Across Platforms Using PySpark

Group 4 Team Members: Alemu Nenko & Daniel Beltran

Date: April 2025

Abstract

This project investigates trends in gaming behavior across three major platforms: Steam, PlayStation, and Xbox. By analyzing structured datasets using PySpark, we extracted insights into player achievements, game genres, pricing, and user behavior. Our goal was to evaluate platform-specific player engagement patterns and their potential social or economic implications.

1. Problem Definition

The global gaming industry is diverse and expanding rapidly. Game developers, marketers, and platform managers seek to understand patterns in player behavior, popular genres, and user preferences across platforms. This study addresses the following key questions:

- How many games do users typically own or play per platform?
- Are there user segments with extremely high or low engagement?
- What are the pricing trends and regional influences?

Challenges include inconsistent schema formats across platforms, missing metadata (especially in Xbox), and computational overhead when processing window functions and large joins. Addressing these issues was key to ensuring valid comparisons and deriving meaningful insights.

2. Dataset Overview

We used structured gaming datasets from Kaggle representing three platforms:

- **Steam:** Game reviews, purchases, achievements, and player metadata
- **PlayStation:** Trophy/achievement data, player-country info, and pricing
- **Xbox:** Achievement points, player libraries, and pricing per region

Each platform's dataset was stored in multiple CSV files categorized into tables such as players, games, purchased, achievements, and prices. The total size exceeded 1GB, requiring PySpark for scalable analysis.

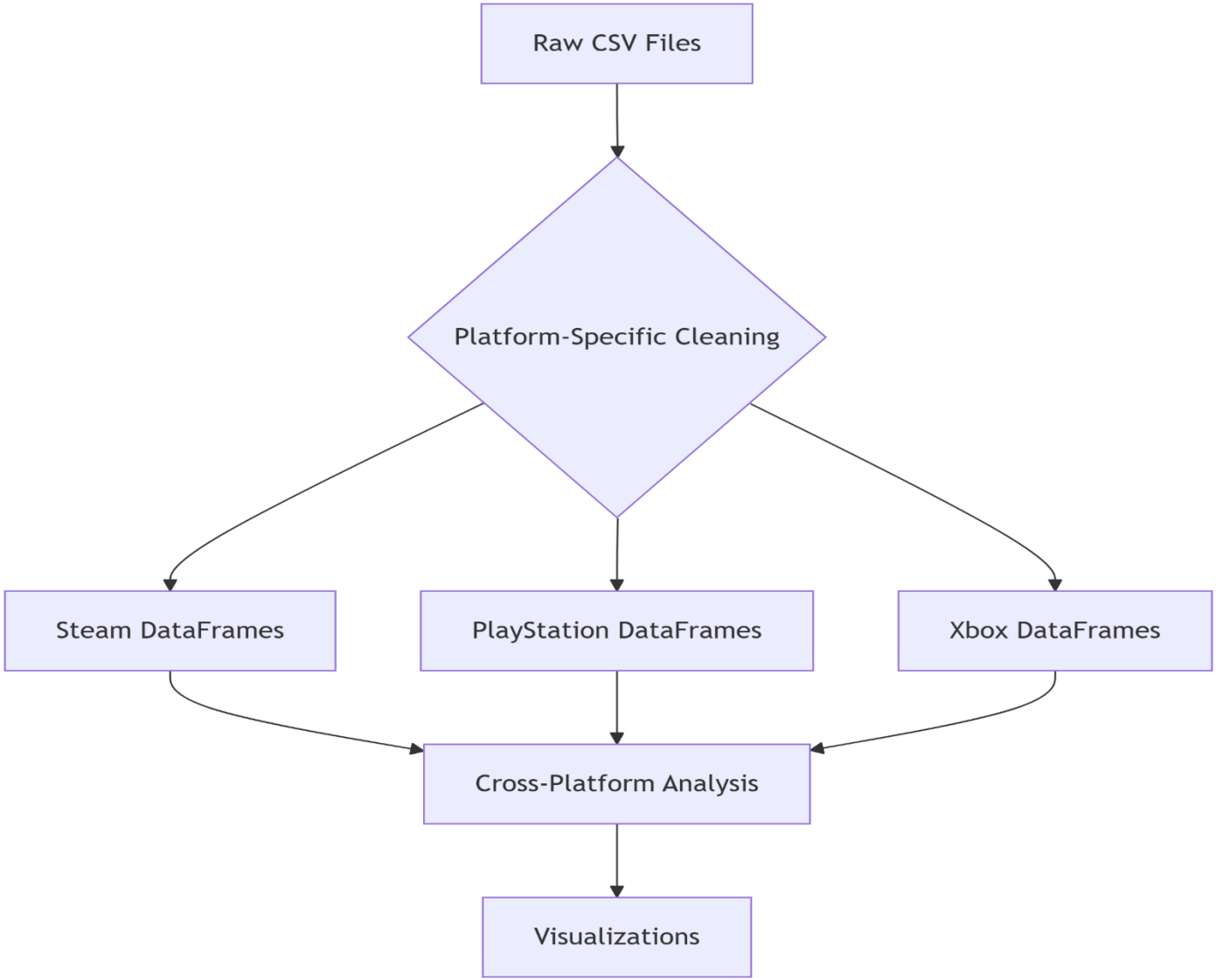
Source: Gaming Profiles 2025 Dataset – Steam, PlayStation, Xbox

3. Methodology

We followed a PySpark-based ETL pipeline:

1) Imported CSVs as Spark DataFrames using spark.read.csv() and/or path.
2) Explored and validated schemas with .printSchema() to understand the structure of each dataset
3) Cleaned missing or null values using .fillna() and removed unnecessary columns
4) Created new derived columns such as num_games using functions like split(), size(), and .withColumn() to measure player engagement
5) Performed joins across tables (e.g., players and purchases) to integrate player behaviors with game and pricing metadata
6) Applied aggregations to compute average and maximum games owned
7) Used .groupBy() and .agg() to analyze regional trends and compute summary statistics
8) Applied window functions (e.g., rank, avg().over(...)) for segmented ranking and group-based analysis
9) Used .partitionBy() to handle performance issues with large window operations

See below, the Visual Workflow:


Raw CSV Files
Platform-Specific Cleaning
Steam DataFrames
PlayStation DataFrames
Xbox DataFrames
Cross-Platform Analysis
Visualizations

Core analytical operations included:

A. **Filtering** – We filtered datasets to isolate specific player or game profiles. For instance, we selected all Brazilian users or filtered action games priced over $50.
B. **Column Creation** – A new column num_games was created using the split and size functions to count games in a user's library.
C. **Aggregations** – Average and maximum games per player were calculated using .agg() to uncover engagement trends.
D. **Grouping** – Players were grouped by country (Steam, PS) or synthetic ID group (Xbox) to analyze regional or clustered behavior.
E. **Sorting** – We used .orderBy() to list top players by library size or top games by price.
F. **Joins** – We joined multiple tables (players, games, purchases) to link individual player behavior with broader game or pricing metadata.
G. **Window Functions** – Applied ranking with .rank().over(...) to evaluate top players within a group.
H. **Aggregate Window Functions** – Used .avg().over(...) to compute average number of games per region or group.

4. Key Findings

Metric	Steam	PlayStation	Xbox	Interpretation
Avg. Games/Player	239.85	233.86	272.63	Xbox leads; possibly due to Game Pass and bundled libraries
Max Games/Player	32,463	13,540	9,018	Steam dominates due to its open marketplace and account age
Top Country (by player)	US, Brazil	US, Brazil	N/A	US leads, Brazil high on all platforms (Xbox lacks location data)
Genre Diversity	High	High	High	All platforms show wide genre representation
Group Avg (regional/syn.)	~165–240	~119–234	~271	Shows regional variation (Steam, PS) vs synthetic grouping (Xbox)
Top Outliers (Players)	32k+ games	13k+ games	9k+ games	Indicates collector or bot-like behavior

5. Findings & Implications

This analysis uncovers several social and market insights. Socially, regions such as the USA and Brazil show strong gaming engagement, indicating gaming's cultural importance. Large game libraries suggest niche behaviors like streaming or review-based consumption.

From a market perspective, Xbox users may present strong monetization opportunities due to high average engagement. Steam’s extensive long-tail of game ownership aligns with freemium

and indie distribution models. PlayStation's international reach supports geo-targeted marketing, especially in Brazil and the UK.

Key recommendations include leveraging long-tail strategies for Steam, geo-targeting on PlayStation, and creating premium engagement opportunities on Xbox. Additional pricing insights suggest players are willing to spend over $100 on premium titles, especially in Xbox's ecosystem.

5.1. Visual comparison

Figure 1: Average Games per Player by Platform

Description:
This bar chart compares the average number of games owned by players on each platform. Xbox leads with an average of **272.63 games per player**, followed by Steam and PlayStation. This may reflect the bundling of games through Xbox Game Pass or platform-based library incentives.

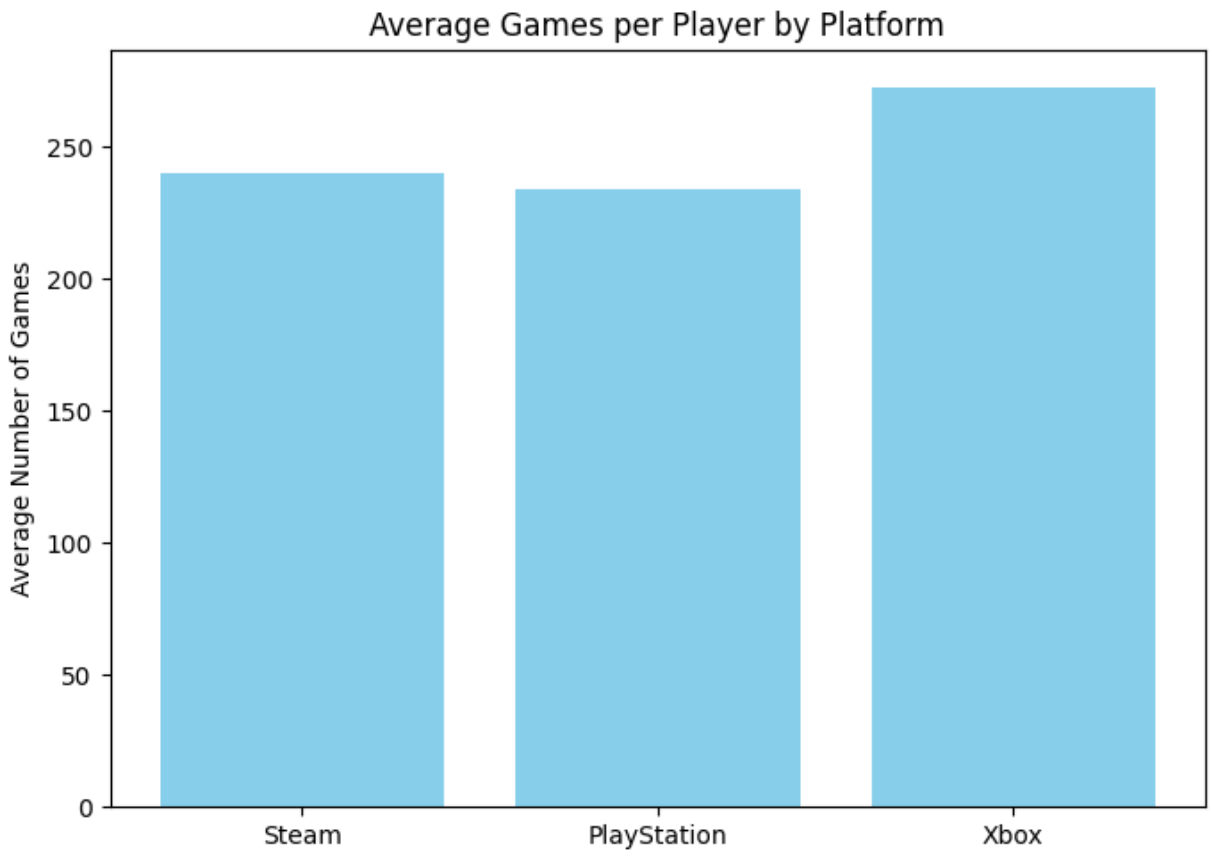
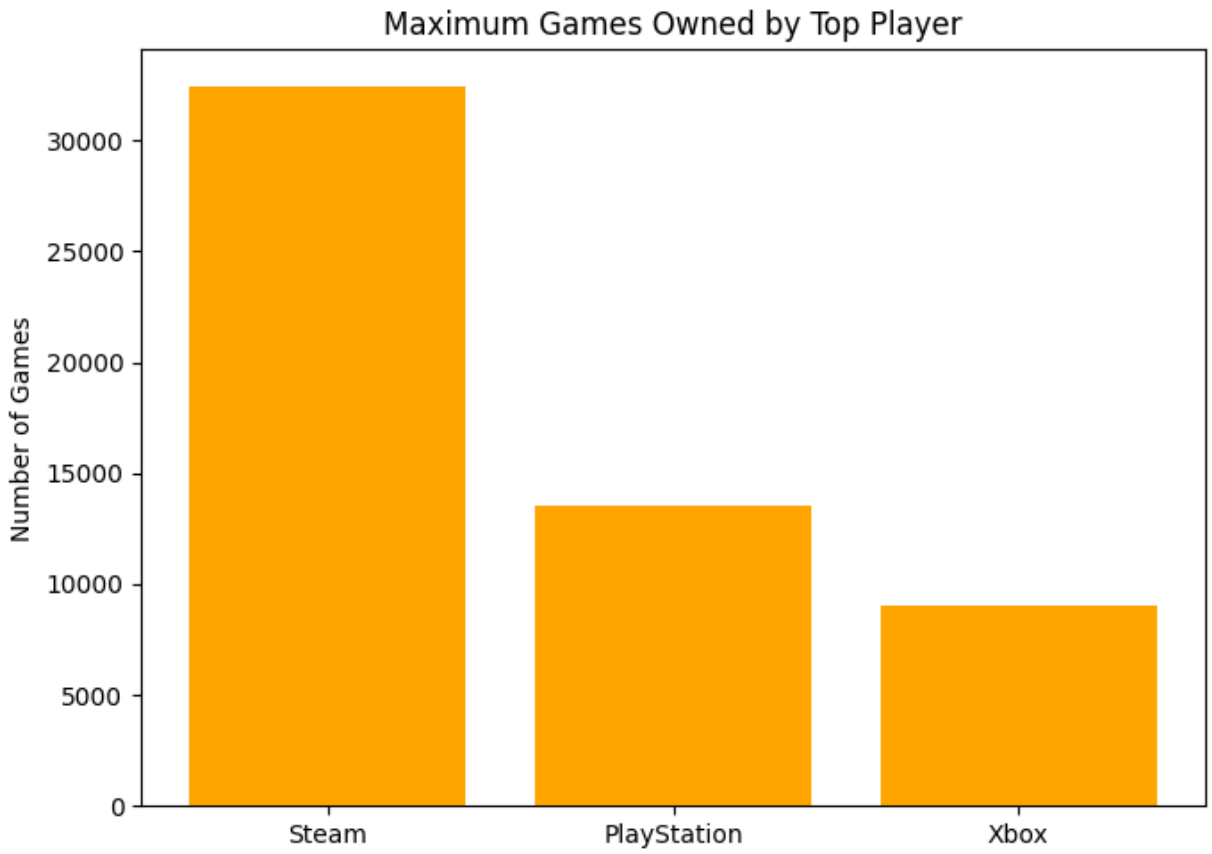


Figure 2: Maximum Games Owned by Top Player

Description:
This chart illustrates the top outlier on each platform. Steam shows an extraordinary collector with **32,463 games**, far exceeding the highest counts on PlayStation (13,540) and Xbox (9,018). Steam's open marketplace and older account history likely contribute to this result.



6. Challenges & Solutions

Problem	Solution
Missing metadata in Xbox	Used synthetic groupings (ID modulo)
Null values in genres/language	Applied .fillna() with inferred replacements
Schema inconsistencies	Normalized with .withColumn() and renaming
Heavy window ops	Used .partitionBy() for scalable ranking

During the analysis, we encountered several key obstacles. First, the Xbox dataset lacked country metadata, which limited regional segmentation. To address this, we introduced a

synthetic grouping strategy using playerid % 3 to analyze behavioral clusters. Another challenge was the prevalence of null values, especially in language and genre metadata fields. These were handled using .fillna() and, when necessary, by dropping incomplete records.

Schema inconsistencies across datasets also posed difficulties, particularly in aligning column names and data types. We resolved these with PySpark's .withColumn() and consistent renaming conventions. Finally, large-scale window functions initially caused performance bottlenecks. We overcame this by partitioning our data intelligently using .partitionBy(), significantly improving execution speed.

7. Conclusion

This project successfully applied PySpark to analyze large-scale, multi-platform gaming data. We achieved the core objectives of the project: importing and transforming large datasets, generating meaningful aggregations, and comparing platform-specific behaviors. Our analyses confirmed key differences in engagement levels, user base distributions, and genre preferences across Steam, Xbox, and PlayStation.

However, there were certain limitations. Xbox lacked country metadata, which restricted our ability to perform regional comparisons. Additionally, across all platforms, some games had incomplete metadata—particularly genres and supported languages—which could limit the depth of genre-specific insights. While we applied cleaning techniques to mitigate this, future work should consider improving the data quality at the source or applying more advanced imputation techniques.

For future improvements, integrating additional dimensions such as playtime, in-game achievements, and player demographics would help distinguish between casual collectors and active users. Also, sentiment analysis of reviews and trend detection using release timelines could offer further marketing or development value. Overall, the findings of this project provide a foundational model for deeper behavioral analytics in the gaming ecosystem.