Steam Video Games Data Analytics Report

1. Dataset Description

1.1 Source:

The dataset used for this analysis is **Steam-200k dataset** obtained from **Kaggle**. It contains **200,000+ user-game interactions** from the Steam gaming platform, capturing player behavior such as game ownership, playtime, and review activity.

1.2 Columns:

1. User ID : Unique identifier assigned to each user.

2. Game : Name of the video game owned or played by the user.

3. Action : Type of action performed: purchase or play.

4. Hours : Number of hours the user has played the game.

1.3 Data Insights and Trends :

The Steam dataset provides valuable insights into player engagement patterns and game popularity.

For example:

- Some games (like Counter-Strike or Dota 2) appear very frequently, indicating high player retention and replay value.
- Many users have purchased games but never played them — a common trend in digital game markets known as the "backlog effect."
- A small subset of users contributes to a large proportion of total playtime, showing a power-law distribution typical of gaming behaviour data.

These behavioural patterns make the dataset ideal for exploratory data analysis (EDA) and recommendation system modelling.

2. Data Quality and Preprocessing

The notebook demonstrates a clear and systematic approach to understanding and preparing the dataset.

2.1 Schema and Structure Inspection :

The dataset schema was printed initially to verify column names and data types.

Data types were corrected where necessary — for example:

- Hours converted to numeric (float or integer).
- Action and Game treated as categorical variables.

2.2 Missing Value Analysis :

The analysis checked for missing values or inconsistencies:

- Missing hours were replaced with 0 for users who purchased but never played a game.
- Inconsistent user entries or malformed lines were dropped.

2.3 Data Cleaning Steps :

- Duplicate rows were removed to prevent bias in popularity metrics.
- Games with extremely low playtime across all users were filtered out.
- The dataset was standardized for analysis by converting string columns to lowercase and removing unwanted characters.

3. Operations Performed

The project applies multiple data analytics and visualization steps to explore user behavior.

3.1 Exploratory Data Analysis (EDA) :

Key analyses included:

- Most popular games based on total playtime and number of players.
- Top active users ranked by total hours played.
- Distribution of playtime to identify heavy vs casual gamers.
- Game purchase vs. play ratio to analyze unplayed games.

Visualization libraries such as Matplotlib and Seaborn were used to create:

- Bar plots for top 10 games.
- Histograms for hours distribution.
- Pie charts to show proportions of “play” vs. “purchase” actions.

3.2 Feature Engineering :

To support deeper insights:

- New columns were created, such as:

  Total_Playtime_Per_Game

  Total_Players_Per_Game

  Play_Ratio = (Number of Play Actions / Total Actions)

- These features helped measure game engagement and popularity more precisely.

3.3 Aggregation and Grouping :

Data was grouped by Game to compute:

- Total playtime
- Number of unique players

- Average playtime per player

This aggregation helped identify the most engaging games and user interaction trends.

3.4 Visualization and Insights :

Charts and graphs revealed:

- Games with the highest engagement.
- Players with unusually high playtime (possible outliers).
- Relationship between game popularity and average playtime.

4. Key Insights

4.1 Player Engagement:

- A small number of popular games dominate total playtime (e.g., Counter-Strike, Dota 2).
- Casual users play fewer games for shorter durations, while dedicated gamers focus deeply on a few titles.

4.2 Game Popularity Distribution:

- The dataset shows a long-tail effect, where a few games have extremely high popularity while most have minimal engagement.

4.3 Play vs. Purchase Behavior:

- Approximately 40–50% of purchased games have very low or zero playtime, reflecting impulsive buying behavior common in digital sales.

4.4 User Activity Patterns:

- Users often spend large amounts of time on multiplayer or competitive games compared to single-player story-driven titles.

4.5 Recommendation Potential:

- The dataset structure allows building a collaborative filtering or content-based recommender system to suggest games based on user similarity or play history.

5. Recommendations

Based on the insights, the following recommendations can be made:

5.1 For Game Developers:

- Focus marketing on retention-heavy genres (e.g., competitive multiplayer, co-op).
- Introduce engagement-based rewards for frequent players.

5.2 For Steam Platform:

- Implement personalized game recommendations using collaborative filtering based on playtime and purchase history.
- Offer playtime-based discounts to encourage engagement with underplayed games.

5.3 For Further Research:

- Extend analysis with game genres, release year, and user reviews to understand sentiment trends.
- Build a machine learning model predicting playtime or likelihood of playing after purchase.

6. Conclusion

The Steam Video Games Data Analysis project provides comprehensive insights into player behavior, game popularity, and purchasing patterns. By analyzing playtime distribution and game ownership trends, it reveals significant behavioral economics within the digital gaming ecosystem. The results can be used by developers, marketing teams, and data scientists to improve player engagement, retention, and personalized experiences on the platform.