Proposal: Analyzing Gaming Addiction Trends and Player Engagement

DATA 450 Capstone

Bijay Adhikari

February 13, 2025

1 Introduction

Video game addiction is a growing issue in the gaming industry impacting individuals by excessive playtime and compulsive engagement. Certain game mechanics, monetization strategies, and social features may contribute to prolonged gaming sessions which may lead to potential addiction concerns. Studies suggest that 3-4% of gamers worldwide experience symptoms of gaming disorder, with excessive gaming linked to negative impacts on mental health, productivity, and social well-being (Pontes et al. 2021). As gaming continues to evolve with better immersive experiences, it is important for developers, researchers, and policymakers to understand the factors that contribute to prolonged playtime. This project aims to analyze player engagement trends and addiction risk factors using publicly available gaming data. The primary objectives are to examine how game design elements, monetization models, and game genres influence playtime and to develop a predictive model for addiction potential. The ultimate goal is to improve our understanding of what makes games engaging and potentially addictive, offering insights into responsible game design and its impact on player behavior.

2 Dataset

The dataset for this proect will be obtained primarily from Steam API(Valve 2025) which provides publicly available gaming data. The Steam API allows data related to game metadata, player statistics, pricing information and publicly made user reviews. Most of the required data are easily accessible from their API. The following variables are intended to be used in the analysis:

- game_id – Unique identifier assigned to each game on Steam.
- title – Name of the game as listed on Steam.

- genre – Category of the game (e.g., FPS, Battle Royale, MMORPG).
- multiplayer – Indicator for multiplayer or single-player.
- avg_playtime – Mean number of hours played per user.
- total_playtime – Cumulative hours played across all users.
- player_count – Number of active players at a given time.
- concurrent_players – Peak number of players online simultaneously.
- monthly_active_users – Number of unique players engaging with the game within a month.
- price – Cost of the game at purchase (in USD).
- microtransactions – Indicator for in-game purchases or no purchases.
- review_score – Percentage of positive reviews on Steam.
- addiction_mentions – Number of reviews containing addiction-related keywords.

3 Data Acquisition and Processing

The data for this project will be retrieved using the Steam API. It provides structured JSON data on game metadata, player statistics, pricing, and user reviews. Access to the API requires an API key availble to developers or Steam gamers and it has already been obtained. The data extraction process will involve making API requests to collect information on game details, playtime metrics, monetization models, and user-generated reviews, which will be stored in multiple CSV files, each corresponding to different data categories such as game metadata, player engagement statistics, and other gaming data. However, in order to obtain player engagement data, a random sampling approach will be applied. Since Steam does not provide a direct way to retrieve public user IDs, user IDs will be randomly generated and checked against publicly accessible profiles to extract playtime statistics. The aim is to collect sufficiently large random sample of users so the dataset will allow generalizable insights into engagement trends. Once the data is collected, preprocessing steps will include converting JSON responses into structured tables, handling missing values, normalizing numerical variables, and recoding categorical data using one-hot encoding.

4 Research Questions and Methodology

1. How much time are people spending on gaming per week, and what does it reveal about gaming addiction trends? To answer this, I will filter the dataset for player engagement data, specifically focusing on weekly playtime data and public user profiles. I will compute mean, median, and variance of weekly playtime, categorize users into engagement levels (e.g., low, moderate, and high playtime), and identify outliers that indicate excessive gaming. I will then normalize playtime data across different games to compare engagement levels. Finally, I will visualize the results on a histogram of weekly playtime distributions.

Estimated Time: 3 hours

2. Which game design elements—such as microtransactions, progression systems, and multiplayer vs. single-player structures—are most associated with prolonged playtime and potential addiction? To answer this, I will filter the dataset based on game metadata and monetization features. I will classify games according to presence of microtransactions, multiplayer functionality, and game progression mechanics. I will then compute average playtime for each category and compare playtime across monetization models and game mechanics. ANOVA and t-tests will be used to assess statistical differences in engagement between game categories. The results will be visualized using boxplots comparing playtime distributions for different monetization and design elements.

Estimated Time: 6 Hours

3. Are certain game genres (e.g., MMORPGs, Battle Royales) more prone to fostering addiction, and what common design elements contribute to high engagement? To answer this, I will calculate average playtime per genre and normalize these values to compare engagement trends. Statistical tests (ANOVA, t-tests) will be applied to assess whether certain genres are associated with significantly higher playtime. Additionally, I will use clustering techniques (K-Means or Hierarchical Clustering) to group games based on engagement patterns and uncover common design elements among highly engaging genres. The results will be visualized using a heatmap comparing playtime across different genres and scatter plots displaying engagement clusters.

Estimated Time: 5 Hours

4. Do highly addictive games tend to have higher or lower ratings, and what are the most common sentiments expressed in player reviews? To answer this, I will analyze the relationship between average playtime and review scores by creating a heatmap that plots playtime on one axis and review scores on the other. Next, I will analyze user review text using natural language processing to identify common sentiment patterns among highly engaging games. Then, I will use sentiment analysis to classify reviews as positive, neutral, or negative, and extract addiction-related keywords.

Estimated Time: 5 Hours

5. Are free-to-play games with monetization mechanics more addictive than paid games, and how does their playtime compare? To answer this, I will classify games into free-to-play and paid categories based on pricing data. I will then compare average playtime between these two groups using a violin plot, which will show the distribution of playtime for each pricing model. I will use clustering on playtime and retention data. The clusters will be plotted in a scatter plot with distinct color groups, helping reveal whether free-to-play games tend to fall into high-playtime, high-retention clusters more frequently than paid games.

Estimated Time: 5 Hours

5 Work plan

Week 4 (2/10 - 2/16):

- Data Collection (6 hours)
- Data Tidying (2 hours)

Week 5 (2/17 - 2/23):

- Complete Question 1: Compute playtime distributions and generate histogram (3 hours)
- Complete Question 2: Classify games by monetization and mechanics (6 hours)

Week 6 (2/24 - 3/2):

- Complete Question 3: Genre and engagement analysis (5 hours)
- Complete Question 4: Complete heatmaps and extract and preprocess review scores and text for NLP (4 hours)

Week 7 (3/3 - 3/9):

- Presentation prep and practice (4 hours)
- Complete Question 5: heatmap and clustering

Week 8 (3/10 - 3/16): *Presentations given on Wed-Thu 3/12-3/13.*

- Poster prep (4 hours)
- Presentation peer review (1.5 hours)

Week 9 (3/24 - 3/30): *Poster Draft 1 due Monday morning 3/24 at 9am. Poster Draft 2 due Sunday night 3/30.*

- Peer feedback (2 hours)
- Poster revisions (1.5 hours)
- Finalizing visualizations (2 hours)

Week 10 (3/31 - 4/6): *Final Poster due Sunday 4/6.*

- Peer feedback (1.5 hours)
- Poster revisions (2 hours)
- Refining and finalizing all plots and figures for clarity (3.5 hours)

Week 11 (4/7 - 4/13):

- Begin drafting blog post introduction and methodology sections (3.5 hours)

- Review and interpret key findings from research questions (3.5 hours)

Week 12 (4/14 - 4/20):

- Complete blog post results section and integrate visualizations (4 hours)
- Refine written analysis (3 hours)

Week 13 (4/21 - 4/27): *Blog post draft 1 due Sunday night 4/28.*

- Finalize and proofread blog post draft (4 hours)
- Cross-check blog post, poster, and research findings (3 hours)

Week 14 (4/28 - 5/4):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)

Week 15 (5/5 - 5/8): *Final blog post due Tues 5/7. Blog post read-throughs during final exam slot, Thursday May 8th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)

References

Pontes, H. M., B. Schivinski, C. Sindermann, M. Li, B. Becker, M. Zhou, and C. Montag. 2021. "Gaming Disorder in the ICD-11: A Cross-National Analysis of Its Association with Gaming-Related Behaviors and Psychological Functioning." *The American Journal of Psychiatry* 178 (6): 574–85. https://doi.org/10.1176/appi.ajp.2020.20091191.

Valve. 2025. "Steam API - Game Data and Player Statistics."