***Abstract—*This report is an in-depth analysis of player retention rates within video games utilizing charting data from Steam’s database. The goal of this project is to establish which genres are most effective in holding the attention of players over a long period of time.**

***Keywords—Standard Deviation, Visualization, Statistics, Variability***

1. INTRODUCTION

Video games are currently one of the most popular forms of entertainment, and their popularity only continues to grow. As the gaming industry grows, so does player demand and expectations. This begs an essential question; what do the players want? Understanding the dynamics of player retention and engagement has become a critical area of study for developers and marketers within the industry. This project aims to explore the impact of game genres on player retention and engagement with a specific focus on the PC market. Utilizing data provided by Steam, an industry leading platform for the PC gaming market, this project will answer six essential questions; What game genres have the highest retention rates,which game genres have the highest player retention rates, how do in-game microtransactions impact player engagement in free to play vs. paid games, how does player activity fluctuate across top genres in the last 24 hours when compared to all time activity peaks, what are the correlations between game monetization models and average playtime, and which genres in the top 100 Steam games show the most diverse player peaks over time.

1. DATASETS

This project utilizes datasets provided by steamdb.info for all time and historical data, the official Steam Application Programming Interface(API) for both real time data and data from the past 24 hours, and howlongtobeat.com for player retention and play time stats. These sources are all reliable and directly connected to player data within Steam. The size of the specific datasets used is based on how many games player data was collected from. Some sets use the top 50 Steam games to collect data, while others use the top 100 games in order to collect it. Others only use the top 5 games within a specific niche, such as free to play, but collect all player data from the time of the games release to when the data was collected. Some of the dataset use comma-separated value(CSV) files as they are available for download from sources such as steamdb.info, while other datasets retrieved real time data from the API using HTTP. In order to answer certain questions, datasets such as Top 5 Free Games and Top 5 Paid Games, and All Time Player Peaks and 24 Hour Peaks were merged. The Top 5 Free Games and Top 5 Paid Games datasets were merged in order to answer the question of whether game monetization models and average playtime have a correlation. The All Time Player Peaks and 24 Hour Peak datasets were merged in order to answer the question of how player activity fluctuates across the top genres when comparing all-time peaks to activity within the past 24 hours. Each dataset was utilized in the most effective way possible in order to achieve the most useful results and properly answer each question.

1. METHODOLOGY

This project was programmed within Google’s Colab coding environment. Colab is a very intuitive tool for programming, and has a striking resemblance to coding environments such as JupyterLab. Methods used across this project are as follows:

1. *Request: HTTP For Humans*

In order to retrieve the data from certain datasets that were not able to provide CSV files, Request: HTTP For Humans was utilized. This tool allows a user to send HTTP/1.1 requests with ease, and allows for the retrieval of real-time data directly from the Steam API.

1. *MatPlotLib*

For data visualization purposes, matplotlib was utilized. Matplotlib is a tool that allows data taken from, for example, a CSV file, to be displayed in a highly customizable and visually appealing graph format.

1. *Descriptive Statistics*

In order to identify trends and differences in player activity across genres, descriptive statistics and aggregation were used. Groupby was used in Pandas in order to aggregate data by genre and compute averages for player counts over two separate time periods. Difference calculation was used to subtract the averages for those two periods and calculate activity fluctuation. This was a purely statistical approach to answering this question and effectively communicates the relevant information.

1. *T-Test*

In order to statistically validate whether the observed differences in average playtime between free to play and paid games are truly significant, hypothesis testing, AKA a T-Test was used. Ttest\_ind from scipy.stats was utilized in order to test whether the means of free to play and paid games are significantly different, and descriptive statistics were used to calculate the mean playtime for each monetization model in order to compare the two.

1. *Standard Deviation*

In order to identify variability in player peaks for each game genre and rank genres based on that variability, standard deviation was used. To calculate the standard deviation , std() within Pandas was utilized. This measures the dispersion of peak player counts within each genre. The genres were then sorted by variability, meaning that genres were ranked based on their calculated standard deviation in order to identify which genres have the most and least diverse player counts. All of these methods collectively provide both appealing visuals and valuable statistics that were used to effectively answer each question.

1. Results

The results of this project give proper and important answers to each question posed, and have crucial implications on development and marketing strategies that can potentially be more effective within the industry. The results for each question are as follows:

1. *What game genres have the highest retention rates?*

In order to answer this question, visualizations of genre counts of the top 100 games on Steam, and the top 15 genres by average playtime were used. Of the top 100 games on Steam, 63 of those belong to the action genre. While this is a very broad genre, it nonetheless has heavy implications on the kind of games that players are most interested in playing. Of the top 15 genres by average playtime, the battle arena genre easily had the highest average at 344 hours. The battle arena genre can be best described as a sub-genre of the action genre, therefore these two statistics support the conclusion that the genre with the highest retention rate is the action genre.

1. *Which game genres have the highest player retention rates?*

In order to answer this question, visualization of the Top 15 genres by average playtime was also used. Reiterating on the discoveries made in answering the first question, the battle arena genre has by far the highest player retention rate. While the other genres behind it have a relatively consistent average playtime with little difference between them, the battle arena genre has a roughly 7 times greater average playtime amongst players than any genre beneath it. For developers this means that if they want to keep their player base coming back, online multiplayer battle arena games are the most effective way to do so.

1. *How do in-game microtransactions impact player engagement in Free-to- Play vs. Paid Games?*

In order to answer this question, visualizations of both the top 5 free-to-play and the Top 5 paid games were utilized. These visualizations displayed the full player activity history of each game from the day it was released to the time the data was recorded. According to the visualizations, significantly more players are consistently playing free games that include microtransactions as opposed to paid games. This has heavy implications on both potential marketing strategies for developers in the future, and on the place that microtransactions have within the gaming industry.

1. *How does player activity fluctuate across top genres in the last 24 hours when compared to all time activity peaks?*

This question was utilized by merging the datasets of all time peak player activity and player activity within the last 24 hours, and then comparing the average peak players within those two timeframes. The results of this indicate a massive difference between all time peak player activity and activity within the past 24 hours across all genres. The genre with the highest all time peak players is the battle royale genre at 3,257,248 players. The peak players for that genre in the past 24 hours is 750,493. The difference between those metrics is 2,506,755 players. These results indicate that the battle royale genre has suffered a major dip in popularity, although it still maintains a healthy playerbase, regardless. Data such as this can help developers figure out which genres players are most interested in playing at the moment, and create more games in that genre as a result.

1. *What are the correlations between game monetization models and average playtime?*

The answer to this question was obtained by combining the datasets of free to play and paid games in order to calculate the standard deviation of average playtime among all of these games. According to the results, There is not a significant difference in average playtime between free to play and paid games. This gives developers more freedom in creating the kinds of games that they want to create, and indicates that the industry is currently healthy and balanced.

1. *Which genres in the top 100 Steam games show the most diverse player peaks over time?*

In order to answer this question, the dataset of the genre count of the top 100 Steam games by genre was used. After calculating the standard deviation of peak player variability, the results indicate that the most diverse genre is the RPG genre, and the least diverse genre is the strategy genre. This indicates to developers that games such as RPGs are more likely to consistently peak than other genres, and allows for a studio to decide what kind of game to make far easier as a result of using this data to take advantage of what is safe to develop.

1. DISCUSSION

While each question was effectively answered and the data provided by the programs created is incredibly useful to developers, the only thing that one cannot account for in the gaming industry is how unpredictable it can be at times. While certain data provided in this project may indicate a trend, there is no way to know for sure whether a game that a developer creates will succeed. The only thing that can be done, which was done effectively within these models, is predict what will do well and proceed cautiously optimistic.

1. CONCLUSION

In closing, the results of this project provide much insight on the video games industry while also offering practical use cases within the actual industry. Developers can look to this information in order to both better understand the industry that they are operating within, and plan better for any future projects they may be working on.