“All Press is Good Press?” : Analysis on the Relation Between Subreddit Reviews and Steam Player Counts

Team Members: James Clark (JAC692), Alexander Hertadi (AFH78), Katherine O’Conner (KSO25)

Introduction:

There is a famous saying “all press is good press”. Is this actually true? The aim of our project is to analyze the association between the sentiment and count of subreddit reviews of a few specific video games and their Steam player counts. We will attempt to measure video game sales via proxy through Steam player counts. There are two overarching goals:

1. We want to explore how the play-style design of the game (single-release, single release and DLC content, single-release and updates) correlates to reddit traffic and steam player counts of the games over time.
2. We want to explore how negative vs. positive subreddit reviews correlate with Steam player count

With these goals in mind, we chose three games that each represented the play-style designs described above. We also opted for games released in 2016 as the Reddit torrent data is much smaller and more feasible to work with from that time frame. Also, since these games were released so long ago, we are better able to capture play trends and internet traffic patterns of the games over their most relevant timelines. We chose *Dark Souls III, No Man’s Sky, and Stardew Valley*. From our own background knowledge, *No Man’s Sky* is infamous for its negative review upon release, yet was still wildly popular.

We expect to find that single release video games will have an initial surge in sales/player count and taper off with time, whereas video games that get patches or additional downloadable content will experience more stable sales/player count. We expect this same trend for the count of subreddit reviews. We expect no strong relationship between time of release and sentiment of reviews.

Generally, we expect the number of Reddit reviews to be the most important predictor of Steam player counts. However, we predict that more positive sentiment reviews will be associated with more stable Steam player counts, whereas more negative sentiment reviews will be associated with sporadic steam player counts and an initial spike upon release.

Data:

Our dataset consists of comments and posts from Reddit for the year 2016, specifically for the games Dark Souls III, No Man’s Sky, and Stardew Valley. The specific subreddits we will pull data from are:

- r/DarkSouls3
- r/NoMansSkyTheGame
- r/StardewValley

Our data is sourced from an academic torrents website and because of its size, is stored in .zst format. Our first step in EDA will be to download, decompress and extract all data relating to the above 3 subreddits.

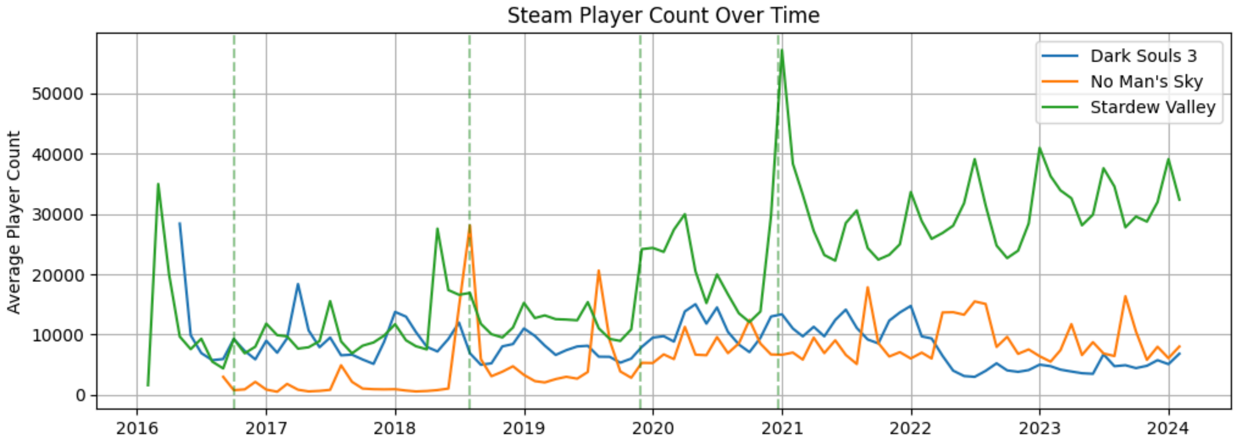
The data in question is very detailed and contains information about:

- Subreddit of the post
- Engagement with the post (upvotes and downvotes)
- Text of post/comment
- Date posted

There are a large number of fields available, but for our analysis we will focus on the above. A full list of features available can be found in the info.md files here.

Note: Because the process of unpacking reddit data is quite time consuming and memory intensive, our exploratory data analysis is quite minimal. If the process of unpacking and cleaning the big messy reddit data becomes too infeasible, we will switch to Google trends data of positive and negative sentiment phrases relating to the game as a backup.

The other datasets were very simply downloaded from SteamCharts. We will be focusing on the month and average number of players to approximate sales and overall popularity of the game.



The dotted lines represent major game updates.

Process:

Upon finding the appropriate dataset, the data will be filtered to the desired subreddits associated with the video games of interest, and posts/comments will be categorized to be positive or negative. The resulting count and ratio of positive to negative post/comments will be compared to the sales/ player count from the steam database.