# Final Report

1. Name, Team Members

Project Name:
Analyzing PlayerActivity and Recommendation Efficiency of Top Steam Games
Group members:
Jiahang Liu (student id: 7071108622)
Renjie Huang(students id: 1626544075)

2. Short Description

In this project, we will analyze how recommendation behavior and price relate to each other for more than 50 of the top-selling games on Steam. We will collect a list of popular titles from the Steam Store “Top Sellers” search pages and then use the Steam Storefront API to retrieve each game’s total number of recommendations, free-to-play status, price, and primary genre. Based on these data, we will compare recommendation patterns between free and paid games, study how price is associated with total recommendations, and explore whether certain genres show systematically higher or lower recommendation support relative to their cost. We also plan to use a simple price-normalized efficiency metric (recommendations per dollar) to highlight games that appear unusually strong or weak in terms of community support given their price.

3. Data

Sources: We will use two main data sources:

Steam Store – Top Sellers search pages From the Steam Store search results for Top Sellers https://store.steampowered.com/search/?filter=topsellers&page=1, 2, 3, …, we will scrape the list of globally top-selling games using Python requests and BeautifulSoup. For each game entry, we will collect the Steam application ID (appid), the game name, and the raw price text displayed on the search page.

Steam Storefront API – App details for each appid For each appid obtained from the Top Sellers pages above, we will query the Steam Storefront API at https://store.steampowered.com/api/appdetails?appids={appid} using requests to retrieve structured JSON data. From this response we will extract the total number of recommendations (recommendations.total), whether the game is free-to-play (is_free), the numeric price (price_overview.final and currency), and the primary genre (the first entry in genres).

Number of Data Samples: We plan to collect data for at least 60–100 games by scraping the first several Top Sellers pages, ensuring the dataset includes more than 50 of the most popular titles on Steam. This larger sample size will make our project more concrete and prevent the analysis from suffering from a small or insufficient dataset.

4. Data Cleaning, Analysis & Visualization

After scraping several "Top Sellers" pages from the Steam Store, we first build a raw table of appids, game names, and raw price text. After that we query the Steam Storefront API for each appid to obtain structured JSON data, including total recommendations, free-to-play status, price, currency, and game genre. These two sources are merged on appid to form a single dataset.

For data cleaning, we remove entries that do not have recommendation information (such as hardware, tools, or incomplete records), drop duplicate appids, and restrict the dataset to games priced in USD or free (currency missing). Recommendation counts are converted to integers. Missing price values are treated as zero, representing free games, and missing genre values are labeled as "Unknown". We also create a free_or_paid indicator to distinguish free-to-play titles from paid games, and define a simple price-normalized efficiency metric, which is price_efficiency = recommendations / price + 1(we add 1 since some of the game is free which will make the efficiency to be 0).

**Analysis & Visualization** After cleaning, we retain 173 game entries, including both free-to-play and paid titles. We examine the distributions of recommendations, price, and price efficiency, and observe strong right-skewness in recommendations and efficiency: most games receive modest engagement, while a small number of titles account for extreme values. In our implementation, we compute: price_efficiency = recommendations / (price + 1), which avoids division-by-zero for free games and enables consistent comparisons across all games.

We then use several plots to characterize relationships and group differences. A scatter plot of price vs. recommendations for paid games shows substantial dispersion at nearly all price levels, suggesting that price alone is not a strong predictor of recommendation volume. A Free vs. Paid boxplot indicates that paid games have a higher median recommendation count, while free-to-play games exhibit larger upper-tail outliers, consistent with a small set of highly popular free titles. We also compare price efficiency across top genres and find that most genres cluster at low efficiency values, but some (e.g., Action) show wider variability and extreme outliers.

To evaluate the role of price more directly, we plot efficiency distributions across price ranges and observe that lower-price tiers contain more high-efficiency games, while higher-price tiers concentrate at lower efficiency values. Finally, a genre-level price vs. efficiency plot with a fitted trend line shows a generally negative relationship, indicating that higher prices are not typically matched by proportionally higher recommendation gains.

**Hypothesis / Premise & Expected Conclusions** We expect free-to-play games to have more total recommendations than paid games because they are easier to access. However, when we look at recommendations relative to price, some paid games may show higher price efficiency than free titles, while others may appear over- or under-priced given their recommendation counts. We also expect recommendation levels and price efficiency to differ across genres (for example, RPG and co-operative games versus competitive shooters). Our analysis will test these patterns using the cleaned dataset and the planned visualizations.

**Conclusions**
Our premise that free-to-play games obtain more recommendations is only partially supported: free games dominate the extreme upper tail, but paid games show higher typical (median) recommendations. Overall, efficiency patterns imply that lower-priced games more often achieve stronger recommendation impact per dollar, with genre-specific variation.

5. Changes from Original Proposal

In our original proposal, we planned to use SteamCharts Top Games plus the Steam Storefront API. We wanted to collect peak-player counts from SteamCharts, combine them with total recommendations from the API, and study “recommendation efficiency = recommendations / peak players”.

When we started coding, we discovered that SteamCharts could not be scraped reliably with requests and BeautifulSoup. The HTML returned to our script did not contain the game table (because of JavaScript and anti-scraping protection), so we were not able to get peak-player data for enough games.

To solve this, we changed the data source for the game list to the Steam Store “Top Sellers” search pages, which can be scraped normally. We still use the Steam Storefront API, but our focus is now on recommendations, price, free-to-play status, and genre instead of player counts. We also introduced a simple price-based efficiency metric (recommendations per dollar) and compare free vs. paid games and different genres.

This change makes the data collection stable and allow us to obtain more categories, making our analysis less limited and more diverse.

## 6. Mention of Future Work

If additional time and resources were available, we would extend this work in three directions. First, we would incorporate player-activity measures to better match the original goal of defining recommendation efficiency relative to game popularity. Second, we would account for discounts and historical price variation by collecting time-stamped price data, since Steam pricing is highly dynamic and may influence recommendation growth. Finally, we would strengthen inference using regression-based modeling with genre controls to quantify relationships more formally and reduce sensitivity to extreme outliers.